Practical Non-Stationary Graph Gaussian Processes

K. S. Sesh Kumar
Brevan Howard Centre for Financial Analysis
Imperial Business School
skarri@imperial.ac.uk

Viacheslav Borovitskiy
School of Informatics
University of Edinburgh
viacheslav.borovitskiy@gmail.com

Abstract

Matérn Gaussian processes offer a principled framework for learning and uncertainty quantification on graphs. However, they struggle with heterogeneous networks. A central challenge is that variance at a node can represent genuine signal that should strongly *propagate (diffuse)* to neighbors, mere noise that should be *isolated*, or a mixture of both. While this problem has been studied in continuous spatial domains, directly translating those approaches to graphs leaves a free parameter per node, causing extreme over-parameterization. To resolve this, we introduce non-stationary *Variance-Coupled Gaussian Processes (VCGPs)*. These models use an auxiliary signal (often easily derived from historical data) to determine relative diffusivity between nodes, tied to the regression data via a single learned hyperparameter, γ. The sign of γ captures whether the auxiliary signal is directly or inversely proportional to node diffusivity, while its magnitude calibrates this relationship. We formally prove that $\gamma = 1$ is the unique point where VCGPs are *spectral-stationary* up to rescaling, ensuring that varying γ results in genuinely non-stationary correlations. Across five real-world datasets, our VCGPs match or outperform both spectral-stationary and per-node non-stationary baselines.

1 Introduction

Many regression problems involve recovering a latent signal from observations attached to the nodes of a graph. Whether tracking traffic speeds along road networks [18], pollutants across urban sensor grids [16], or temperatures over spatial station networks [20], these applications frequently demand calibrated uncertainty to ensure safety guarantees and drive automated decision-making. Gaussian processes (GPs) on graphs [4] provide a principled probabilistic framework for this setting. By placing a Gaussian prior over node-indexed functions—as opposed to GPs whose *inputs* are entire graphs [5]—they encode dependencies through a kernel that reflects the structure of the graph. A common choice is the graph Matérn family [3, 4, 21, 30] that extends the widely popular Matérn covariances [23, 24] to discrete domains via the graph Laplacian.

Despite their appeal, these models impose stationarity and have a single global hyperparameter governing the strength of interactions. Throughout, *stationarity* refers to *spectral stationarity*, defined in Section 2; Euclidean translation invariance is not assumed. Real-world networks are rarely so uniform. Traffic congestion concentrates on specific parts of the road network, and pollution dynamics depend on local geography (e.g. the altitude profile). For instance, consider a road network where some of the nodes have traffic speed readings. Then the same signal can admit two opposing interpretations. If high variability is due to being an important junction, a node is informative and its fluctuations should *propagate (diffuse)* across the network. If, instead, high variability is due to a malfunctioning sensor, then the node primarily injects noise and its fluctuations should be *isolated*. In practice, which interpretation holds is rarely known a priori; it may vary across datasets and even across regions within a single graph. Designing a model that can express both behaviours in a principled and computationally tractable manner while allowing data to determine the relevant regime is the central question we address.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Existing graph kernels do not provide a simple mechanism to express this distinction. Stationary graph kernels [4, 26, 32] treat all nodes uniformly, ignoring node-specific behaviour. Continuous-domain covariance constructions and variable-coefficient stochastic partial differential equations (SPDEs) [2, 9, 12, 19, 22] allow both variance and covariance to vary spatially. These models can use low-dimensional covariate parametrisations; fitting coefficients independently at each vertex instead introduces $O(n)$ parameters (Section 2). We introduce an auxiliary-driven non-stationary GP defined directly through the combinatorial graph Laplacian, with an $O(1)$ learned parameter count. It makes non-stationary graph GPs *practical* in this setting by capturing heterogeneous correlations with a constant number of parameters.

Our approach. We introduce *Variance-Coupled Gaussian Processes* (VCGPs), a family of graph GPs that captures network heterogeneity using a single interpretable hyperparameter. To avoid over-parameterization, we leverage an *auxiliary signal* that is frequently available in practice but rarely utilized in kernel design—such as the often-ignored temporal dimension of spatial datasets. Specifically, from this auxiliary signal, we extract a positive quantity d_i at each node to represent its variability or activity level. We then define the kernel in terms of both these d_i values and the underlying graph structure.

Importantly, the auxiliary statistics alone do not determine whether high variability should propagate or be isolated. In the traffic example, both a genuinely important junction and a malfunctioning sensor may exhibit large temporal variability and therefore large d_i. The distinction instead emerges through the learned coupling parameter γ. Negative values of γ interpret high-variability nodes as informative, amplifying their interactions and allowing their fluctuations to diffuse across the network. Positive values interpret the same variability as unreliable or noisy, suppressing interactions and isolating their influence. The model therefore does not hard-code an interpretation of variability, but instead allows the data itself to determine whether high-d_i nodes are propagators or attenuators of information. When a single global γ is insufficient to map variability to diffusivity, VCGPs extend naturally by partitioning the graph into clusters and learning one γ_k per cluster. In a road network, for example, high variability may reflect traffic dynamics in one region and sensor faults in another. Cluster-specific exponents allow the model to propagate fluctuations in the former and suppress them in the latter simultaneously, even when their d_i values are similar. This requires regions where variability should propagate and regions where it should be suppressed to lie in different clusters. If an important junction and a faulty sensor have the same d_i and lie in the same cluster, their shared γ_k cannot treat their variability differently.

Contributions.

1. **Methodology:** We propose Variance-Coupled Gaussian Processes (VCGPs), a novel class of non-stationary graph GPs. By linking node diffusivity to an auxiliary signal, VCGPs can represent complex heterogeneous correlations using a single scalar learnable hyperparameter, γ.
2. **Theory:** We prove that $\gamma = 1$ is the unique point where VCGPs share the correlations of a stationary GP while retaining node-dependent variances. It also supplies a warm start (Section 4): initialising at $\gamma = 1$ retains that fit, so learning γ cannot reduce the training marginal likelihood, though held-out negative log-likelihood (NLL) can worsen (PEMS-BAY).
3. **Evaluation:** Across five real-world datasets, we demonstrate that VCGPs match or outperform both stationary and heavily parameterized non-stationary baselines (Sections 4 and 5).

2 Background and Motivation

Graph Gaussian processes. Let $G = (V, E, w)$ be a finite undirected weighted graph with $|V| = n$ nodes, non-negative edge weights $w_{ij} \geq 0$, adjacency matrix $\mathbf{A}_{ij} = w_{ij}$, and (unnormalised) Laplacian $\mathbf{L} = \mathbf{D}_{\text{deg}} - \mathbf{A}$, where $\mathbf{D}_{\text{deg}} = \text{diag}(\sum_j w_{ij})$ collects the weighted node degrees [6].

A graph Gaussian process places a Gaussian distribution over functions $f : V \to \mathbb{R}$, written $f \sim \mathcal{N}(0, \mathbf{K})$, where $\mathbf{K} \in \mathbb{R}^{n \times n}$ is a positive definite kernel matrix. While the kernel is the standard object in probabilistic machine learning, on a graph it is typically dense: an entry $\mathbf{K}_{ij}$ captures the *marginal* co-variation between $f(i)$ and $f(j)$, aggregating contributions from all paths in the graph. For modelling local interactions it is more convenient to work with the precision matrix $\mathbf{Q} = \mathbf{K}^{-1}$ [25]: in a Gaussian Markov random field $\mathbf{Q}_{ij} = 0$ exactly when nodes i and j are conditionally independent given the rest, so the sparsity pattern of $\mathbf{Q}$ encodes the local structure of the graph.

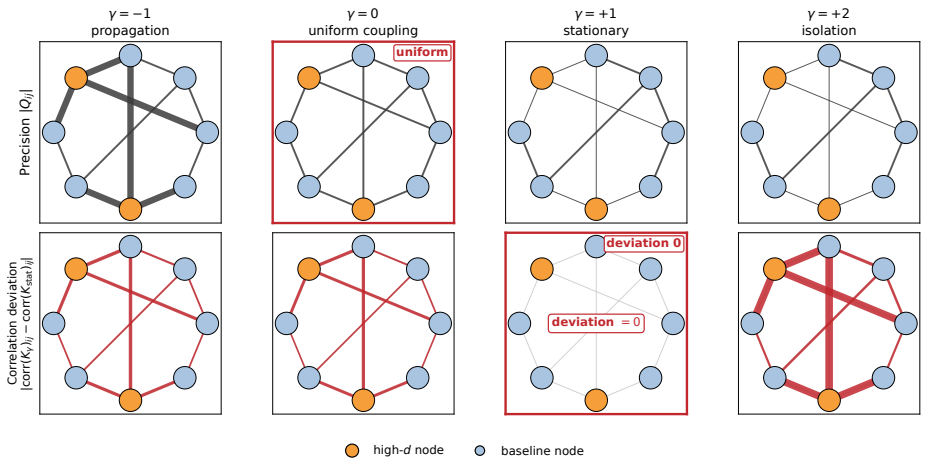

Figure 1: **One learned γ selects propagation or isolation.** Orange nodes have high d_i; blue nodes are baseline. Edge thickness shows $|\mathbf{Q}_{ij}|$ (top) and correlation deviation from the stationary baseline (bottom). For high-d_i nodes, $\gamma < 0$ amplifies coupling and $\gamma > 0$ suppresses it. Highlighted: $\gamma = 0$, uniform coupling; $\gamma = 1$, stationary correlations with node-dependent variances.

A widely used construction for Gaussian processes on graphs is the graph Matérn family [3, 4, 21, 30], obtained by defining covariance operators as spectral functions of the graph Laplacian; the closely related heat (diffusion) kernel arises in the $\nu \to \infty$ limit and has proved effective on combinatorial domains [8, 14]. Following the construction of Borovitskiy et al. [4], the graph Matérn process precision operator is defined as

$$\mathbf{Q}_{\text{Matérn}} = \left(\frac{2\nu}{\kappa^2}\mathbf{I} + \mathbf{L}\right)^{\nu},$$

where $\kappa > 0$ is a length-scale parameter and $\nu > 0$ controls smoothness. Equivalently, introducing the diffusion parametrisation

$$2\tau = \frac{\kappa^2}{2\nu}, \qquad \text{so that} \qquad \frac{1}{2\tau} = \frac{2\nu}{\kappa^2},$$

the precision can be written, up to a global scaling factor, as $\mathbf{Q}_{\text{Matérn}} \propto (\mathbf{I} + 2\tau\mathbf{L})^{\nu}$. Under this parametrisation, τ directly controls diffusion strength across the graph: larger values propagate information more strongly through neighbouring nodes, while ν governs spectral decay and hence the regularity of the resulting functions. The simplest case corresponds to $\nu = 1$, yielding

$$\mathbf{Q}_{\text{stat}} = \mathbf{I} + 2\tau\mathbf{L}.$$

This serves as our stationary baseline. In all cases, the Laplacian $\mathbf{L}$ induces co-variation between neighbouring nodes, while τ globally controls interaction strength, so all nodes are treated uniformly.

Spectral stationarity. Laplacian eigenvectors form the graph Fourier basis, with eigenvalues representing graph frequencies [4, 6]. Let $\mathbf{U}$ collect these orthonormal eigenvectors and $\mathbf{\Lambda}$ be their diagonal eigenvalue matrix, so $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\top}$. For a positive function g of graph frequency, spectral stationarity means $\mathbf{K} = g(\mathbf{L}) = \mathbf{U}g(\mathbf{\Lambda})\mathbf{U}^{\top}$. Thus the covariance retains the Laplacian's eigenvectors and transforms each eigenvalue through g to set the corresponding Fourier mode's variance. The Fourier coefficients are uncorrelated.

Stationarity and heterogeneity. While convenient, the stationary graph kernels like graph Matérn may be too restrictive when nodes exhibit different levels of variability or activity. A simple extension introduces heterogeneity through node-specific marginal precisions,

$$\mathbf{Q} = \mathbf{D}^{-1} + 2\tau\mathbf{L}, \qquad \mathbf{D} = \text{diag}(d_1, \ldots, d_n),$$

which preserves positive definiteness but only modifies marginal variances, leaving correlations between nodes largely unchanged.

More flexible non-stationary constructions [10–12, 19, 22] allow both marginal uncertainty and interaction strength to vary across the graph. A direct graph analogue takes the form

$$\mathbf{Q} = \mathbf{D}^{-1} + 2\tau\,\mathbf{H}^{1/2}\mathbf{L}\mathbf{H}^{1/2}, \qquad \mathbf{H} = \text{diag}(h_1, \ldots, h_n),$$

where d_i controls marginal variance and h_i controls node-dependent interaction strength. While highly expressive, this construction introduces two free parameters per node, a number of parameters that is impractical to fit from the $\ll n$ observations typically available.

A modeling gap. We therefore face a trade-off: stationary models are stable but inflexible, while fully non-stationary models are expressive but difficult to learn. In the next section, we introduce a structured non-stationary family that breaks free from this trade-off by leveraging an auxiliary signal capturing heterogeneity while only using a single additional hyperparameter.

3 Variance-Coupled Gaussian Processes

Construction. We address the gap from Section 2 by tying both diagonal and off-diagonal precision structure to a single auxiliary signal d_i at each node i. The auxiliary d_i is a fixed, data-derived measure of node variability or activity (e.g. temporal variance or coefficient of variation; see Section 5 and Appendix B.4). We set the marginal precision at node i to d_i^{-1}, so a larger d_i corresponds to higher marginal variance. Further, we introduce the scalar parameter $\gamma \in \mathbb{R}$: it determines how variability d_i modulates interaction strength between neighbouring nodes (diffusivity). This separation ensures that d_i encodes marginal variation, while γ controls whether (and to what extent) higher variation should lead to propagation or to isolation of information in the node.

Definition 1 (Variance-coupled precision)**.** *Given* $\mathbf{D} = \mathrm{diag}(d_1, \ldots, d_n)$ *with* $d_i > 0$ *and scalars* $\tau, s > 0$ *and* $\gamma \in \mathbb{R}$*, define*

$$\mathbf{Q}_\gamma \;=\; \mathbf{D}^{-1} + 2\tau\, \mathbf{D}^{-\gamma/2}\, \mathbf{L}\, \mathbf{D}^{-\gamma/2}, \qquad \mathbf{K}_\gamma \;=\; s\, \mathbf{Q}_\gamma^{-1}.$$

The off-diagonal entries are $\mathbf{Q}_\gamma(i,j) \;=\; -2\tau\, d_i^{-\gamma/2} d_j^{-\gamma/2} w_{ij}$*, and the* i*-th diagonal entry is* $\mathbf{Q}_\gamma(i,i) = d_i^{-1} + 2\tau\, d_i^{-\gamma}\, \deg(i)$*, where* $\deg(i) = \sum_j w_{ij}$*.*

Only the d_i^{-1} term of the diagonal is γ-independent: the Laplacian also contributes $2\tau d_i^{-\gamma} \deg(i)$, and the marginal variance $s\, \mathbf{Q}_\gamma^{-1}(i,i)$ depends on the whole matrix. This is the flexible form of Section 2 with the per-node field pinned to the auxiliary—taking $\mathbf{H} = \mathbf{D}^{-\gamma}$ recovers $\mathbf{Q}_\gamma$—so its n free entries become one exponent on a measured field, and the parameter count falls from $O(n)$ to $O(1)$.

The auxiliary signal $\{d_i\}_{i=1}^n$ is fixed, so the learned parameters in Gaussian process regression are the scalar γ together with global parameters $(\tau, s, \sigma_\varepsilon^2)$, where σ_ε^2 denotes likelihood (observation) noise.

Interpretation. The parameter γ determines how node-level variability modulates a node's interactions with its neighbours. For $\gamma < 0$, nodes with larger d_i amplify their connections, increasing their influence on adjacent nodes. For $\gamma > 0$, the same nodes suppress their connections, reducing their influence. Thus, γ controls the interactions of a node on the graph.

Figure 1 illustrates this transition. As γ varies, edge strengths and resulting correlations are *amplified* or *suppressed*, with $\gamma = 0$ corresponding to uniform coupling where interaction strength does not depend on d_i and $\gamma = 1$ yielding correlations identical to the stationary baseline of Section 2 (up to variance rescaling). We summarise these and some other important regimes below:

- $\mathbf{D} = \mathbf{I}, \gamma = 0$:
$$\mathbf{K}_{\mathrm{stat}} = s\,(\mathbf{I} + 2\tau\mathbf{L})^{-1},$$
is the stationary kernel as discussed in Section 2; we denote it `stat`.
- $\gamma = 0$:
$$\mathbf{Q}_0 = \mathbf{D}^{-1} + 2\tau\mathbf{L}.$$
The auxiliary d_i affects only variances, while the interaction strength remains identical to `stat`. We denote this case `fem`, after its finite-element form.
- $\gamma = 1$:
$$\mathbf{Q}_1 = \mathbf{D}^{-1} + 2\tau\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}, \qquad \mathbf{K}_1 = s\,\mathbf{D}^{1/2}(\mathbf{I} + 2\tau\mathbf{L})^{-1}\mathbf{D}^{1/2}.$$
In this regime, the kernel is obtained from the stationary covariance through a diagonal rescaling. Specifically, if
$$\mathbf{K}_{\mathrm{stat}} = s\,(\mathbf{I} + 2\tau\mathbf{L})^{-1},$$

then

$$\mathbf{K}_1 = \mathbf{D}^{1/2}\mathbf{K}_{\text{stat}}\mathbf{D}^{1/2}.$$

As a result, node-specific variability modifies variances while preserving the underlying correlation structure. Indeed, the covariance between nodes i and j becomes

$$(\mathbf{K}_1)_{ij} = \sqrt{d_i d_j}\,(\mathbf{K}_{\text{stat}})_{ij}.$$

Consequently, the correlation coefficient is unchanged:

$$\text{Corr}_{\mathbf{K}_1}(i,j) = \frac{(\mathbf{K}_1)_{ij}}{\sqrt{(\mathbf{K}_1)_{ii}(\mathbf{K}_1)_{jj}}} = \frac{(\mathbf{K}_{\text{stat}})_{ij}}{\sqrt{(\mathbf{K}_{\text{stat}})_{ii}(\mathbf{K}_{\text{stat}})_{jj}}}.$$

Thus $\gamma = 1$ preserves the same correlation geometry as the stationary model, but allows heterogeneous marginal uncertainty across nodes. We denote this regime by `het`.

Amplification and Suppression Regimes. The effect of γ is explicit in the off-diagonal entries:

$$\mathbf{Q}_{ij} = -2\tau\, d_i^{-\gamma/2} d_j^{-\gamma/2} w_{ij}.$$

When $\gamma = 0$, all edges have identical strength. For $\gamma < 0$, edges connected to high-variability nodes are amplified, leading to stronger propagation of influence. For $\gamma > 0$, these edges are suppressed, leading to isolation.

3.1 Non-stationarity of the VCGP family

We now show that most values of γ yield genuinely non-stationary kernels.

Proposition 1 (Characterisation of the unique stationary-rescaling point)**.** *Let $G = (V, E, w)$ be a connected graph with Laplacian $\mathbf{L}$, and let $d = (d_1, \ldots, d_n)$ be non-constant with $d_i > 0$. Define*

$$\mathbf{Q}_\gamma = \mathbf{D}^{-1} + 2\tau\,\mathbf{D}^{-\gamma/2}\mathbf{L}\mathbf{D}^{-\gamma/2}, \qquad \mathbf{K}_\gamma = s\,\mathbf{Q}_\gamma^{-1},$$

where $\mathbf{D} = \text{diag}(d)$ and $\tau, s > 0$.

1. *(**Stationary rescaling**) At $\gamma = 1$,*

$$\mathbf{K}_1 = s\,\mathbf{D}^{1/2}(\mathbf{I} + 2\tau\mathbf{L})^{-1}\mathbf{D}^{1/2}.$$

   Thus $\mathbf{K}_1$ is a diagonal rescaling of the stationary kernel $\mathbf{K}_{\text{stat}} = s(\mathbf{I} + 2\tau\mathbf{L})^{-1}$ and has the same correlation matrix.
2. *(**Uniqueness**) Assume that any two nodes can be connected through a sequence of node pairs, where each consecutive pair is at distance exactly two, i.e. non-adjacent with a common neighbour. Then for any $\gamma \neq 1$, there do not exist a positive diagonal matrix $\mathbf{S}$ and a positive definite matrix $\mathbf{K}_0$ satisfying $\mathbf{K}_0\mathbf{L} = \mathbf{L}\mathbf{K}_0$ such that $\mathbf{K}_\gamma = \mathbf{S}\mathbf{K}_0\mathbf{S}$. That is, for $\gamma \neq 1$, $\mathbf{K}_\gamma$ is not a node-wise rescaling of a stationary graph kernel.*

Proposition 1 shows that $\gamma = 1$ is the unique point at which the auxiliary signal affects only marginal variances without altering the correlation structure. In contrast, for $\gamma \neq 1$, the auxiliary field enters the coupling term $\mathbf{D}^{-\gamma/2}\mathbf{L}\mathbf{D}^{-\gamma/2}$ and cannot be removed by any diagonal rescaling. As a result, the induced correlations are genuinely non-stationary. Variation in d_i makes the kernel non-stationary even with a global γ, and no member of the family is itself spectral-stationary; a per-node γ_i would permit node-level regime changes but reintroduce $O(n)$ parameters, so cluster-specific γ_k is the compromise (Section 3.2).

This distinction explains the role of γ: rather than adjusting a redundant scale, it selects between qualitatively different interaction regimes, ranging from propagation to isolation. Away from $\gamma = 1$ the correlations remain well behaved: $\mathbf{Q}_\gamma$ is a Stieltjes M-matrix, so $\mathbf{K}_\gamma$ is entrywise positive on a connected graph and every correlation is non-negative at every γ. The connectivity condition in part (2) is mild: it holds on all of LAQN and on 98–99.5% of vertices for METR-LA, PEMS-BAY and PM2.5, and on the 80% core of CA Weather's distance-threshold graph, the remainder forming small boundary components (Appendix J). The proof is given in Appendix A.

Table 1: Members of the VCGP family. Top: special-value members of `lgamma`. Bottom: extensions.

Method	Specialisation	Precision $\mathbf{Q}$ / Kernel $\mathbf{K}$	Free parameters
`stat`	$\mathbf{D} = \mathbf{I}$	$\mathbf{Q} = \mathbf{I} + 2\tau\mathbf{L}$	$(\tau, s, \sigma_\varepsilon^2)$
`fem`	$\gamma = 0$	$\mathbf{Q} = \mathbf{D}^{-1} + 2\tau\mathbf{L}$	$(\tau, s, \sigma_\varepsilon^2)$
`het`	$\gamma = 1$	$\mathbf{Q} = \mathbf{D}^{-1} + 2\tau\mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2}$	$(\tau, s, \sigma_\varepsilon^2)$
`lgamma`	learned scalar γ	$\mathbf{Q} = \mathbf{D}^{-1} + 2\tau\mathbf{D}^{-\gamma/2}\mathbf{L}\mathbf{D}^{-\gamma/2}$	$(\tau, s, \sigma_\varepsilon^2, \gamma)$
`lgamma_block`	piecewise γ_k, K blocks	$\mathbf{Q}_\gamma$ with $\mathbf{L} \to \mathrm{blockdiag}(\mathbf{L}_1, \ldots, \mathbf{L}_K)$	$(\tau, s, \sigma_\varepsilon^2, \gamma_1, \ldots, \gamma_K)$
`lgamma_m15`	Matérn $\nu = 3/2$	$\mathbf{K} = s\,\mathbf{Q}_\gamma^{-2}$	same as `lgamma`
`lgamma_fisher`	Fisher-info ridge on γ	same $\mathbf{Q}$	same as `lgamma`

3.2 Extensions

The basic scalar-γ model admits three direct extensions, summarised in Table 1. First, higher smoothness is obtained by replacing $\mathbf{K} = s\mathbf{Q}_\gamma^{-1}$ with powers of the same precision, e.g. $\mathbf{K} = s\mathbf{Q}_\gamma^{-2}$ for the Matérn-$3/2$ variant `lgamma_m15`. Second, multi-regime graphs are handled by partitioning the nodes into K clusters and learning one exponent γ_k per cluster, giving `lgamma_block`. Third, `lgamma_fisher` stabilises estimation by adding a ridge penalty on γ, with the regularisation coefficient set from the Fisher information associated with the marginal likelihood [1]. These modifications compose: `lgamma_fisher_m15` combines the ridge and Matérn-$3/2$ power. In Table 2, `block-aug` abbreviates `lgamma_block_aug`: K-means on standardised coordinates augmented with the standardised auxiliary weighted by $w \geq 0$. The `aug` suffix distinguishes this from coordinate-only `lgamma_block`. The label `spectral-BIC` abbreviates `spectral_bic_lgamma_block`, which uses spectral clustering with K selected by the training-data Bayesian Information Criterion (BIC).

The piecewise form used throughout is the *block* form, which also replaces $\mathbf{L}$ by $\mathrm{blockdiag}(\mathbf{L}_1, \ldots, \mathbf{L}_K)$ and so makes the clusters conditionally independent—a modelling assumption. It, the full-edge alternative, the ridge and the grids over K and w are defined and priced in Appendix B and Appendix K.

4 Practical Implementation and Inference

We now describe how VCGPs are fit in practice. The goal is to estimate the scalar coupling parameter γ together with the global parameters $(\tau, s, \sigma_\epsilon^2)$ from observed data, where s denotes the signal variance (output scale) and σ_ϵ^2 the observation noise variance.

Marginal likelihood optimisation. Given observations $y \in \mathbb{R}^n$ at the graph nodes, we assume the Gaussian observation model

$$y = f + \epsilon, \qquad f \sim \mathcal{N}(0, \mathbf{K}_\gamma), \qquad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 \mathbf{I}),$$

with covariance

$$\mathbf{K}_\gamma = s\,\mathbf{Q}_\gamma^{-1}.$$

This yields the marginal distribution

$$y \sim \mathcal{N}\left(0,\, \mathbf{K}_\gamma + \sigma_\epsilon^2 \mathbf{I}\right).$$

Parameters are estimated by maximising the marginal log-likelihood

$$\mathcal{L}(\gamma, \tau, s, \sigma_\epsilon^2) = -\tfrac{1}{2} y^\top (\mathbf{K}_\gamma + \sigma_\epsilon^2 \mathbf{I})^{-1} y - \tfrac{1}{2} \log\det(\mathbf{K}_\gamma + \sigma_\epsilon^2 \mathbf{I}) - \tfrac{n}{2} \log(2\pi).$$

The optimisation is low-dimensional, with γ as the only non-standard parameter. Gradients are computed via standard matrix calculus, and optimisation is performed using off-the-shelf gradient-based methods.

Warm-start from the stationary point. The model admits a natural reference point at $\gamma = 1$, where the kernel reduces to a variance-rescaled stationary kernel (Section 3.1). This provides a stable and interpretable initialisation. From an Occam's razor perspective, it is the simplest member of the family, before introducing additional non-stationary structure.

Table 2: Top-5 methods by NLL, plus the strongest outer baseline. Entries are mean ± standard error (SE); *, $^+$ and $^\wedge$ mark the best NLL, RMSE and recalibrated NLL among the displayed rows. "cat" distinguishes family members from baselines. `block-aug` lists (K, w); it and `spectral-BIC` abbreviate the corresponding `lgamma` block members. Auxiliary observations are disjoint from targets (Section 5.1).

Dataset	method	NLL/pt	RMSE	recNLL/pt	cat
METR-LA n_{test}=63	`lgamma_fisher`	$\mathbf{1.2018 \pm 0.0287}^*$	0.8369 ± 0.0238	1.1763 ± 0.0232	VCGP
	`lgamma`	1.2025 ± 0.0294	0.8468 ± 0.0247	$\mathbf{1.1759 \pm 0.0228}^\wedge$	VCGP
	`het`	1.2201 ± 0.0265	0.8469 ± 0.0231	1.1975 ± 0.0234	VCGP
	`lgamma_fisher_m15`	1.2202 ± 0.0344	$\mathbf{0.8265 \pm 0.0235}^+$	1.1816 ± 0.0247	VCGP
	`lgamma_m15`	1.2234 ± 0.0349	0.8354 ± 0.0240	1.1835 ± 0.0244	VCGP
	`fuglreg_lam1.0`	1.2382 ± 0.0294	0.8603 ± 0.0232	1.2064 ± 0.0224	base
PEMS-BAY n_{test}=98	`lgamma_fisher`	$\mathbf{1.2720 \pm 0.0346}^*$	0.9015 ± 0.0360	$\mathbf{1.2282 \pm 0.0371}^\wedge$	VCGP
	`het`	1.2767 ± 0.0329	$\mathbf{0.8906 \pm 0.0329}^+$	1.2344 ± 0.0362	VCGP
	`lgamma`	1.2835 ± 0.0370	0.9019 ± 0.0360	1.2374 ± 0.0381	VCGP
	`lgamma_fisher_m15`	1.2911 ± 0.0359	0.8963 ± 0.0329	1.2459 ± 0.0371	VCGP
	`spectral-BIC`	1.2934 ± 0.0330	0.9249 ± 0.0359	1.2582 ± 0.0329	VCGP
	`fuglreg_lam1.0`	1.3340 ± 0.0365	0.8948 ± 0.0331	1.2918 ± 0.0349	base
LAQN n_{test}=16	`lgamma_m15`	$\mathbf{1.4519 \pm 0.0374}^*$	1.0455 ± 0.0383	1.4068 ± 0.0332	VCGP
	`lgamma`	1.4688 ± 0.0400	1.0434 ± 0.0393	1.4171 ± 0.0346	VCGP
	`lgamma_fisher_m15`	1.4741 ± 0.0410	1.0468 ± 0.0386	1.4227 ± 0.0355	VCGP
	`lgamma_fisher`	1.4742 ± 0.0403	1.0460 ± 0.0392	1.4223 ± 0.0347	VCGP
	`spectral-BIC`	1.4779 ± 0.0528	$\mathbf{1.0259 \pm 0.0396}^+$	$\mathbf{1.4020 \pm 0.0372}^\wedge$	VCGP
	`m05`	1.5244 ± 0.0487	1.0447 ± 0.0397	1.4495 ± 0.0411	base
PM2.5 n_{test}=297	`block-aug` $(20, 1)$	$\mathbf{0.9756 \pm 0.0106}^*$	$\mathbf{0.6541 \pm 0.0055}^+$	$\mathbf{0.9703 \pm 0.0095}^\wedge$	VCGP
	`block-aug` $(10, 1)$	1.0078 ± 0.0105	0.6822 ± 0.0075	1.0032 ± 0.0099	VCGP
	`block-aug` $(10, 0.5)$	1.0303 ± 0.0097	0.7011 ± 0.0069	1.0269 ± 0.0090	VCGP
	`block-aug` $(20, 0.5)$	1.0410 ± 0.0121	0.6975 ± 0.0072	1.0338 ± 0.0102	VCGP
	`block-aug` $(5, 1)$	1.0597 ± 0.0099	0.7171 ± 0.0069	1.0553 ± 0.0095	VCGP
	`fuglreg_lam1.0`	1.0748 ± 0.0095	0.7467 ± 0.0062	1.0705 ± 0.0082	base
CA Weather n_{test}=195 (mean over 12 months)	`block-aug` $(10, 1.0)$	$\mathbf{0.8685 \pm 0.0079}^*$	$\mathbf{0.5738 \pm 0.0049}^+$	$\mathbf{0.8572 \pm 0.0076}^\wedge$	VCGP
	`ps`	0.9360 ± 0.0088	0.6261 ± 0.0053	0.9247 ± 0.0078	base
	`se_kriging`	0.9511 ± 0.0061	0.6245 ± 0.0043	0.9376 ± 0.0065	base
	`spectral-BIC`	0.9562 ± 0.0229	0.6151 ± 0.0094	0.9364 ± 0.0163	VCGP
	`fuglreg_lam1.0`	1.0004 ± 0.0087	0.6451 ± 0.0055	0.9900 ± 0.0076	base

In practice, we recommend first fitting the $\gamma = 1$ model and then using the resulting parameters to initialise optimisation of the full model. This warm-start provides a stable starting point, reduces sensitivity to initialization, and allows the optimisation to introduce non-stationarity only when supported by the data.

Multi-regime networks. When a single value of γ is insufficient to capture heterogeneous behaviour, we partition the nodes into K clusters and assign a separate parameter γ_k to each cluster.

This yields a piecewise VCGP in which coupling adapts locally while retaining a small number of parameters. Clusters are obtained using standard graph clustering methods (e.g. spectral clustering), and the parameters $(\gamma_1, \ldots, \gamma_K)$ are optimised jointly with the rest of parameters. This extension captures multiple amplification/suppression modulations within a single graph while remaining tractable.

Remarks. Additional implementation details — including the auxiliary clipping strategy, the warm-start template, partitioning schemes for the piecewise model, numerical stabilisation (e.g. jitter and Cholesky ladder for $\nu = 3/2$), as well as Fisher-information regularisation (`lgamma_fisher`) are provided in Appendix B.

5 Experiments

In this section we show that VCGPs provide a useful middle ground between stationary graph kernels and heavily parameterised non-stationary models. The experiments ask three questions: (i) do VCGPs improve prediction and calibration; (ii) when is the stationary factorisation point $\gamma = 1$ sufficient; and (iii) when is a piecewise multi-regime model necessary?

Protocol. For each dataset, methods share 20 $70/30$ train/test splits and the auxiliary. We report per-point NLL (in nats), root-mean-square error (RMSE), and recalibrated NLL, which fits one global variance scale on the test predictions solely as a diagnostic (Appendix B.5). Leaderboard settings are chosen post hoc, so these leaderboards do not measure selection performance; validation-fold selection is assessed separately (Appendices L–M). METR-LA, PEMS-BAY and LAQN seeds re-split one snapshot rather than giving independent temporal replications. Auxiliary details are in Appendix B.4.

Baselines. We compare against six external baselines: `m05` and `m15` (stationary graph Matérn with fixed $\nu \in \{1/2, 3/2\}$ [4, 21]), `lnu` (graph Matérn with learnable ν), `ps` (Paciorek–Schervish per-node lengthscales [12, 22] sharing our auxiliary d_i through a single global lengthscale), `fuglreg_lam`λ (our graph adaptation of variable-coefficient SPDE models [9, 19], with per-node κ_i and a graph smoothness penalty, the closest per-node analogue of our one-scalar construction), and `se_kriging`, a squared-exponential-kernel spatial GP fitted by grid search [7, 24]. Two learned baselines test the family beyond kernel methods: `dk`, a deep kernel GP [31], and a five-member deep ensemble of Graph Convolutional Networks (GCNs) [13, 15] taking d_i as a feature. Both lose to the family leader on raw NLL and RMSE on all five datasets—the ensemble by 0.13–3.97 nat/pt (worst $p = 0.008$) and the deep kernel by 1.11–2.89—though the ensemble is competitive rather than dominated, its METR-LA deficit being a global variance scale that recalibration removes (Appendix E). Ensemble calibration on graphs has in any case been questioned [29].

Datasets. We use five real graph regression problems. Throughout, the auxiliary d_i is computed from observations disjoint from the regression target: a forward window for the traffic graphs, an interleaved even/odd-day split for PM2.5, a different year for CA Weather, and an external site-type covariate for LAQN, which uses no temporal signal at all.

- **METR-LA**: Los Angeles freeway sensor graph [18]. The signal is the daily mean speed from the first 288 five-minute readings on 2012-03-01. The auxiliary d_i is the coefficient of variation (CV) over the readings *after* that day—a window disjoint from the target.
- **PEMS-BAY**: San Francisco Bay Area freeway sensor graph [18], with the same traffic-speed setup and the same disjoint-window coefficient-of-variation auxiliary.
- **LAQN**: London Air Quality Network sensors [16]. The graph is a 5 km distance-threshold graph. No temporal signal is used. The values d_i are a six-level site-type ordinal from rural to kerbside.
- **PM2.5**: United States Environmental Protection Agency Air Quality System annual PM2.5 field for 2023 [27]. The graph is an 8-nearest-neighbour graph on a set of latitude/longitude locations. The target is the annual mean over even-ranked days and the auxiliary the standard deviation (std) over odd-ranked days, so the two share no observation while both stay representative of the year. We use the standard deviation rather than the coefficient of variation, preserving level information.
- **CA Weather**: Global Historical Climatology Network daily California weather stations [20]. The graph is a 50 km distance-threshold graph. The target is daily maximum temperature at twelve monthly snapshots in 2023, and the auxiliary is built from 2022, so the variance field cannot carry target-year information.

Preliminary results. Synthetic recovery experiments confirm that scalar `lgamma` recovers γ once $|S|/n \geq 0.5$, and a two-regime stochastic block model shows that `lgamma_block` is necessary when different regions have different propagation/isolation regimes; see Appendix C for details.

5.1 Main results

For each dataset we report the best-performing member of the VCGP family and compare it with the strongest baselines. Table 2 consolidates the top-5 leaderboard for every dataset; * marks top NLL, $^+$ marks top RMSE, and $^\wedge$ marks top recalibrated NLL within each dataset. Below, we summarize the results. We defer further discussion of these and some additional results until Section 5.3.

METR-LA. In METR-LA, the best raw NLL is achieved by `lgamma_fisher` (1.2018 nat/pt), while `lgamma_fisher_m15` gives the best RMSE. This is a useful Pareto split: tuned γ improves uncertainty calibration, while the Matérn-$3/2$ refinement smooths the mean prediction. The learned exponent is strongly positive over the twenty seeds ($\hat{\gamma} = +2.9 \pm 0.7$ for scalar `lgamma`, $+2.1 \pm 0.5$ for the leader), a strong-isolation regime in which high-variability nodes are down-weighted. Scalar `lgamma` is statistically indistinguishable from the Fisher-ridge leader on raw NLL (1.2025, $p = 0.44$), while every outer baseline in Table 5 is worse at $p < 10^{-3}$.

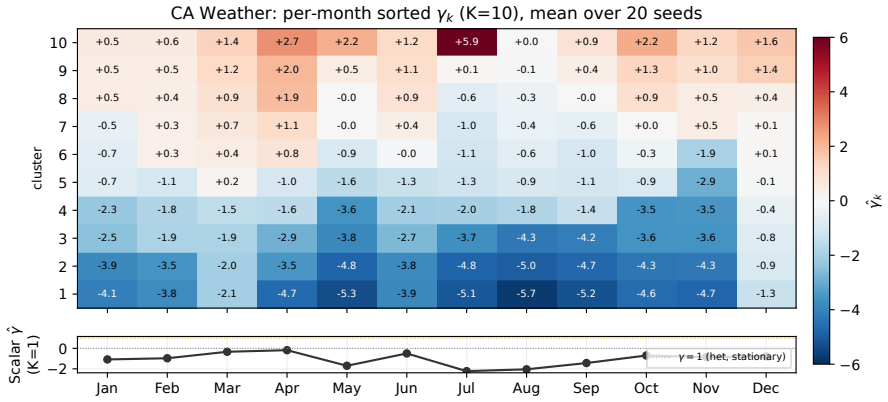

Figure 2: CA Weather seasonal heatmap of $\hat{\gamma}_k$ by month (`lgamma_block`, $K = 10$, not the augmented leader; rows are sorted ranks, not fixed clusters). Most months contain both amplification and suppression regimes, changing substantially across seasons.

PEMS-BAY. PEMS-BAY is the clean case of stationary rescaling of Proposition 1. The fixed `het` model, at the factorisation point $\gamma = 1$, gives the best RMSE (0.8906) and is statistically indistinguishable from the leading member `lgamma_fisher` on NLL (1.2767 against 1.2720, $p = 0.17$); learning γ freely is worse (1.2835, $p = 0.010$) and block variants are worse still. The auxiliary field explains heterogeneity through variance scaling alone, while the correlation structure stays close to stationary.

LAQN. On LAQN, `lgamma_m15` has the best raw NLL (1.4519 nat/pt), while `spectral_bic_lgamma_block` has the best RMSE (1.0259) and recalibrated NLL (1.4020 nat/pt) in Table 7. BIC selects $K^\star = 2$ on 19 of 20 splits; the partition is geographic, not by site type. This block member does not improve raw NLL on scalar `lgamma` ($+0.01$ nat/pt, $t = +0.4$).

PM2.5. PM2.5 is the clearest multi-regime case. The leader is `lgamma_block_aug` with $K = 20$ and auxiliary $d =$ annual standard deviation. The leader improves over `m15` substantially on both NLL and RMSE, and considerably outperforms the per-node SPDE baseline `fuglreg_lam1.0`. It does so with only 20 cluster-wise exponents, compared with 987 per-node heterogeneity parameters in `fuglreg`. The recovered PM2.5 exponents span both regimes, with some clusters in propagation and others in isolation. This is precisely the setting where a scalar γ is too restrictive and a small vector $(\gamma_1, \ldots, \gamma_K)$ is useful.

CA Weather. California Weather offers twelve independent monthly snapshots over a fixed graph. The leader is `lgamma_block_aug` with $(K, w) = (10, 1.0)$. The VCGP gain over its own stationary null is largest on CA Weather, and on the twelve-month aggregate every outer baseline is beaten at $p < 10^{-3}$. Per month, the leader beats all thirteen outer baselines at $p < 0.05$ on both NLL and RMSE in eight of the twelve months (nine on NLL alone, ten on RMSE alone); requiring Bonferroni correction across the twelve monthly evaluations leaves five. The winter months are reported as non-significant headline wins. Figure 2 shows the per-cluster $\hat{\gamma}_k$ heatmap across the twelve months.

5.2 Ablation: when does learning γ matter?

Table 3 compares `stat` $\to$ `fem` $\to$ `het` $\to$ `leader`. The steps introduce the auxiliary diagonal $\mathbf{D}$ with uniform coupling, recover stationary correlations at $\gamma = 1$, and test the benefit of learning a global or cluster-specific γ.

The ablation separates three regimes. First, PEMS-BAY stops at `het`: the auxiliary affects variance, but learning non-stationary correlations adds no value. Second, METR-LA and LAQN benefit incrementally from each step, suggesting mild but homogeneous departures from stationary correlations. Third, PM2.5 and CA Weather show the opposite pattern: moving to the factorisation point hurts, and only the cluster-dependent γ model recovers the gain. These are the genuinely multi-regime cases.

The best member is itself diagnostic: `het` signals stationary correlations after rescaling, scalar `lgamma` a homogeneous propagation/isolation regime, and `lgamma_block` spatially varying ones.

5.3 Experimental summary

VCGPs match or outperform stationary and per-node non-stationary baselines, with the largest gains on multi-regime PM2.5 and CA Weather. PEMS-BAY provides the complementary case: stationary correlations after variance rescaling suffice.

Practical recipe. Fit the family in order, warm-starting each extension from its simpler member. **(i) Match the auxiliary to the target's averaging window:** coefficient of variation for daily snapshots, standard deviation for annual aggregates, or site type without temporal data. This guidance is based on the aggregation argument in Appendix B.4, rather than a direct comparison of coefficient of variation and standard deviation on these datasets. The PM2.5 validation ablation compares the external auxiliary with training-target-derived alternatives (Appendix M). A fixed-exponent proxy control illustrates the risk of a misaligned auxiliary (Appendix N). **(ii) Always fit** `het` **($\gamma = 1$).** At this unique factorisation point, the auxiliary changes variance while preserving stationary correlations. It ties the NLL leader and gives the best RMSE on PEMS-BAY. **(iii) Default to** `lgamma_fisher`**:** one learned exponent with a Fisher-information ridge centred at $\gamma = 1$. The ridge adds no parameters and leads on both traffic graphs, but costs 0.005 nat/pt on LAQN and 0.018 on PM2.5. For mean prediction, the Matérn-$3/2$ refinement replaces $\mathbf{Q}_\gamma^{-1}$ with $\mathbf{Q}_\gamma^{-2}$ (`lgamma_m15`). **(iv) Use blocks when a global exponent is a compromise:** a spread in fitted γ_k or a clear improvement over scalar `lgamma`, as on PM2.5 and CA Weather. Failure to beat `het` alone does not justify escalation: on PEMS-BAY, blocks at $K = 2, 5, 10$ improve test NLL at none of these settings. **(v) Select on validation, not on the leaderboard.** The primary selection procedure uses a $50/20/30$ split to choose member, K, d and λ on validation and score only the selected model on disjoint test nodes. Appendix L quantifies its cost relative to post-hoc selection.

6 Discussion and Conclusion

We introduced Variance-Coupled Gaussian Processes (VCGPs), a practical family of non-stationary graph GPs that ties node-level variability to edge coupling through a single scalar exponent γ. The construction keeps the precision sparse, avoids impractical overparameterization, and has a clear stationary reference point where the kernel is only a diagonal rescaling of a stationary graph Matérn kernel.

The empirical results show that γ is useful not merely as an extra hyperparameter, but as a diagnostic of graph heterogeneity. On PEMS-BAY the $\gamma = 1$ factorisation point is indistinguishable from the best model and has the lowest RMSE, indicating stationary correlations after variance rescaling. On METR-LA and LAQN, learning γ gives modest but consistent gains. On PM2.5 and CA Weather, a single global exponent is insufficient; there, a small number of cluster-wise exponents capture the crucial multi-regime behaviour.

Limitations. The method assumes that a useful auxiliary signal d_i is available and may be sensitive to its choice. The present theory establishes the unique stationary-correlation point and rules out diagonal-rescaling equivalence away from $\gamma = 1$, but does not provide finite-sample bounds for estimating γ. Finally, the block extension depends on a chosen partition family; learning spatially varying $\gamma(x, t)$ directly remains open.

Future work. Natural extensions include spatio-temporal data and signed Laplacians for heterophilic graph signals, which can model anti-correlation.

Table 3: Family ablation (NLL/pt). The leader is the lowest-NLL family member in the per-dataset leaderboard. Positive Δ means improvement: Δ_1 = `stat` $\rightarrow$ `fem`, Δ_2 = `fem` $\rightarrow$ `het`, Δ_3 = `het` $\rightarrow$ leader.

Dataset	d_i	`stat`	`fem`	`het`	leader	Δ_1	Δ_2	Δ_3
METR-LA	CV	1.2612	1.2537	1.2201	1.2018	0.0075	0.0336	0.0183
PEMS-BAY	CV	1.3594	1.3245	1.2767	1.2720	0.0349	0.0478	0.0047
LAQN	site-type	1.6013	1.5590	1.5234	1.4519	0.0423	0.0355	0.0715
PM2.5	annual std	1.1329	1.1557	1.2403	0.9756	-0.0228	-0.0846	0.2647
CA Weather	daily-var	1.0436	1.0673	1.1661	0.8685	-0.0237	-0.0988	0.2976

Acknowledgements and disclosures

Funding and acknowledgements. K. S. Sesh Kumar is funded by the Brevan Howard Centre for Financial Analysis at Imperial Business School. We thank Dr. Erik Chavez for providing access to compute resources.

Competing interests. The authors declare no competing interests. Beyond the funding stated above, the work generates no income or other benefit to the authors, and there are no financial relationships outside this work that could be perceived as influencing it.

Reproducibility. Code is available at `https://github.com/seshkumar/vcgp-graph-gp`. The release documents separate commands for the real-data leaderboards, validation-fold selection, synthetic studies and the optional Deep Ensemble comparison. The real-data experiments use the twenty seeds specified in the protocol; synthetic-study configurations specify their own seeds and settings. Optional comparisons require the additional dependencies stated in the release instructions.

References

[1] S.-i. Amari. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

[2] D. Bolin, L. Riera-Segura, and A. B. Simas. A new class of non-stationary Gaussian fields with general smoothness on metric graphs, 2026. URL `https://arxiv.org/abs/2501.11738v2`.

[3] V. Borovitskiy, A. Terenin, P. Mostowsky, and M. P. Deisenroth. Matérn Gaussian processes on Riemannian manifolds. In *Advances in Neural Information Processing Systems*, pages 12426–12437, 2020.

[4] V. Borovitskiy, I. Azangulov, A. Terenin, P. Mostowsky, M. Deisenroth, and N. Durrande. Matérn Gaussian processes on graphs. In *Artificial Intelligence and Statistics*, 2021.

[5] V. Borovitskiy, M. R. Karimi, V. R. Somnath, and A. Krause. Isotropic Gaussian processes on finite spaces of graphs. In *International Conference on Artificial Intelligence and Statistics*, 2023.

[6] F. R. K. Chung. *Spectral Graph Theory*. American Mathematical Society, 1997.

[7] N. A. C. Cressie. *Statistics for Spatial Data*. John Wiley & Sons, revised edition, 1993.

[8] C. Doumont, V. Picheny, V. Borovitskiy, and H. Moss. Omnipresent yet overlooked: Heat kernels in combinatorial Bayesian optimization. In *Advances in Neural Information Processing Systems*, 2025.

[9] G.-A. Fuglstad, F. Lindgren, D. Simpson, and H. Rue. Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy. *Statistica Sinica*, 25(1):115–133, 2015.

[10] G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Does non-stationary spatial data always require non-stationary random fields? *Spatial Statistics*, 14:505–531, 2015.

[11] G.-A. Fuglstad, D. Simpson, F. Lindgren, and H. Rue. Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*, 114: 445–452, 2019.

[12] M. N. Gibbs. *Bayesian Gaussian Processes for Regression and Classification*. PhD thesis, University of Cambridge, 1998.

[13] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[14] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *International Conference on Machine Learning*, pages 315–322, 2002.

[15] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

[16] LAQN. London Air Quality Network Data Archive. `https://www.londonair.org.uk`, 2024.

[17] M. Lei, A. Labbé, and L. Sun. Scalable spatiotemporally varying coefficient modeling with Bayesian kernelized tensor regression. *Bayesian Analysis*, 20(3):919–947, 2025.

[18] Y. Li, R. Yu, C. Shahabi, and Y. Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *International Conference on Learning Representations*, 2018.

[19] F. Lindgren, H. Rue, and J. Lindström. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B*, 73(4):423–498, 2011.

[20] M. J. Menne, I. Durre, R. S. Vose, B. E. Gleason, and T. G. Houston. An overview of the Global Historical Climatology Network-Daily database. *Journal of Atmospheric and Oceanic Technology*, 29(7):897–910, 2012.

[21] P. Mostowsky, V. Dutordoir, I. Azangulov, N. Jaquier, M. J. Hutchinson, A. Ravuri, L. Rozo, A. Terenin, and V. Borovitskiy. The GeometricKernels package: Heat and Matérn kernels for geometric learning on manifolds, meshes, and graphs. *Journal of Machine Learning Research*, 26(276):1–14, 2025.

[22] C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. In *Advances in Neural Information Processing Systems*, volume 16, 2003.

[23] E. Porcu, M. Bevilacqua, R. Schaback, and C. J. Oates. The Matérn model: a journey through statistics, numerical analysis and machine learning. *Statistical Science*, 39(3):469–492, 2024.

[24] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[25] H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC, 2005.

[26] A. J. Smola and R. Kondor. Kernels and regularization on graphs. In B. Schölkopf and M. K. Warmuth, editors, *Learning Theory and Kernel Machines*, pages 144–158. Springer Berlin Heidelberg, 2003.

[27] U.S. Environmental Protection Agency. Air Quality System (AQS) Pre-Generated Data Files. `https://aqs.epa.gov/aqsweb/airdata/download_files.html`, 2024.

[28] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

[29] P. C. Vieira, P. Ribeiro, and V. Borovitskiy. Do deep ensembles actually capture uncertainty in graph neural networks? *arXiv preprint arXiv:2605.22593*, 2026.

[30] P. Whittle. Stochastic processes in several dimensions. *Bulletin of the International Statistical Institute*, 40:974–994, 1963.

[31] A. G. Wilson, Z. Hu, R. Salakhutdinov, and E. P. Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, 2016.

[32] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *International Conference on Machine Learning*, 2003.

Appendix

This appendix contains: (**A**) the full proof of Proposition 1; (**B**) practical implementation details — the fitting algorithm, the per-variant edit cheatsheet, partition sources for the piecewise extension, the dataset-specific choice of auxiliary d_i, the baseline tuning protocol, and numerical detail (M-matrix property, jitter, Cholesky ladder); (**C**) the two synthetic experiments (identifiability grid and two-regime stochastic block model); (**D**) full per-dataset leaderboards with standard errors and significance; (**E**) the deep-ensemble GNN and deep-kernel comparison; (**F**) hyperparameters and numerical settings; (**G**) the twelve CA Weather monthly leaderboards; (**H**) scalability of the kernel factorisation; (**I**) long-form heterophily framing and signed-Laplacian roadmap; (**J**) per-dataset coverage of the Proposition 1 hypotheses; (**K**) the K-grid for the piecewise variants; (**L**) validation-fold selection; (**M**) choice of auxiliary and partition source; (**N**) the auxiliary-proxy control; and (**O**) the deep-kernel comparison.

A Full proof of Proposition 1

Statement. Let $G = (V, E, w)$ be a connected graph with Laplacian $\mathbf{L}$, and let $d = (d_1, \dots, d_n)$ be non-constant with $d_i > 0$. Define

$$\mathbf{Q}_\gamma = \mathbf{D}^{-1} + 2\tau\, \mathbf{D}^{-\gamma/2}\mathbf{L}\mathbf{D}^{-\gamma/2}, \qquad \mathbf{K}_\gamma = s\, \mathbf{Q}_\gamma^{-1},$$

where $\mathbf{D} = \mathrm{diag}(d)$ and $\tau, s > 0$.

(1) (*Stationary factorisation at* $\gamma = 1$.) At $\gamma = 1$, $\mathbf{K}_1 = s\, \mathbf{D}^{1/2}(\mathbf{I} + 2\tau\mathbf{L})^{-1}\mathbf{D}^{1/2}$. Hence $\mathbf{K}_1$ is a diagonal rescaling of the stationary kernel $\mathbf{K}_{\mathrm{stat}} = s\,(\mathbf{I} + 2\tau\mathbf{L})^{-1}$ and has the same correlation matrix.

(2) (*Uniqueness*.) Let the *distance-two graph* $D^2(G)$ be the graph on the same vertices in which $i \sim j$ exactly when $i \neq j$, i and j are *non-adjacent* in G, and they share a common neighbour. Assume $D^2(G)$ is connected. Then for any $\gamma \neq 1$, there do not exist a positive diagonal matrix $\mathbf{S}$ and a positive definite matrix $\mathbf{K}_0$ satisfying $\mathbf{K}_0\mathbf{L} = \mathbf{L}\mathbf{K}_0$ such that $\mathbf{K}_\gamma = \mathbf{S}\mathbf{K}_0\mathbf{S}$. In particular, $\mathbf{K}_\gamma$ is not a node-wise rescaling of any stationary graph kernel for $\gamma \neq 1$.

Notation. For square matrices $\mathbf{A}, \mathbf{B}$, the commutator is $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$. Two symmetric matrices commute if and only if they admit a common orthonormal eigenbasis. "$\mathbf{K}_0$ commutes with $\mathbf{L}$" is the algebraic form of "$\mathbf{K}_0$ is diagonalised by the graph Fourier basis" — the standard definition of a stationary graph kernel. For the unnormalised Laplacian $\mathbf{L} = \mathbf{D}_G - \mathbf{W}$ (where $\mathbf{D}_G$ is the diagonal degree matrix, distinct from our $\mathbf{D}$), the off-diagonal entries are $\mathbf{L}_{ij} = -w_{ij} \leq 0$, vanishing exactly when $\{i, j\} \notin E$.

Strategy. For $\gamma \neq 1$ the kernel $\mathbf{K}_\gamma$ cannot be written as $\mathbf{S}\mathbf{K}_0\mathbf{S}$ with $\mathbf{K}_0$ commuting with $\mathbf{L}$. The argument has three steps. (1) The factorisation hypothesis gives $\mathbf{S}\mathbf{Q}_\gamma\mathbf{S}$ commuting with $\mathbf{L}$; decomposing into $\Delta + 2\tau\, \Phi\mathbf{L}\Phi$ produces a commutator identity. (2) Evaluating it on distance-two pairs forces $\phi_i = \phi_j$, and connectedness of $D^2(G)$ extends this to a global constant. (3) The leftover diagonal Δ must then commute with $\mathbf{L}$; evaluated on edges, this forces Δ constant on V, which contradicts non-constancy of $d_i^{\gamma-1}$ when $\gamma \neq 1$.

Part (1): the factorisation at $\gamma = 1$

Substituting,

$$\mathbf{Q}_1 = \mathbf{D}^{-1} + 2\tau\, \mathbf{D}^{-1/2}\mathbf{L}\mathbf{D}^{-1/2} = \mathbf{D}^{-1/2}(\mathbf{I} + 2\tau\mathbf{L})\mathbf{D}^{-1/2}.$$

Inverting, $\mathbf{K}_1 = s\, \mathbf{D}^{1/2}(\mathbf{I} + 2\tau\mathbf{L})^{-1}\mathbf{D}^{1/2}$. Since correlation matrices are invariant under positive diagonal conjugation, $\mathbf{K}_1$ and $\mathbf{K}_{\mathrm{stat}} = s\,(\mathbf{I} + 2\tau\mathbf{L})^{-1}$ have identical correlations.

Part (2): uniqueness

Suppose, for contradiction, that there exist a positive diagonal $\mathbf{S} = \mathrm{diag}(s_1, \dots, s_n)$ and a positive definite $\mathbf{K}_0$ with $[\mathbf{K}_0, \mathbf{L}] = 0$ such that $\mathbf{K}_\gamma = \mathbf{S}\mathbf{K}_0\mathbf{S}$ for some $\gamma \neq 1$. From $\mathbf{K}_\gamma = s\, \mathbf{Q}_\gamma^{-1}$ we obtain $\mathbf{Q}_\gamma = s\, \mathbf{S}^{-1}\mathbf{K}_0^{-1}\mathbf{S}^{-1}$, hence

$$M := \mathbf{S}\mathbf{Q}_\gamma\mathbf{S} = s\, \mathbf{K}_0^{-1}.$$

Since $\mathbf{K}_0^{-1}$ commutes with $\mathbf{L}$, so does M.

Define

$$\Phi := \mathbf{S}\mathbf{D}^{-\gamma/2} = \mathrm{diag}(\phi_i), \qquad \Delta := \mathbf{S}\mathbf{D}^{-1}\mathbf{S} = \mathrm{diag}(\Delta_i),$$

so that $\phi_i = s_i d_i^{-\gamma/2}$ and $\Delta_i = s_i^2 d_i^{-1} = \phi_i^2 d_i^{\gamma-1}$. Direct substitution yields

$$M = \Delta + 2\tau\, \Phi\mathbf{L}\Phi. \tag{1}$$

The condition $[M, \mathbf{L}] = 0$ then gives

$$[\Delta, \mathbf{L}] = -2\tau\, [\Phi\mathbf{L}\Phi, \mathbf{L}]. \tag{2}$$

Step 1: ϕ is constant on distance-two pairs. Let i, j be non-adjacent at graph distance two. Then $\mathbf{L}_{ij} = 0$, so the (i, j) entry of $[\Delta, \mathbf{L}]$ vanishes. A direct computation, using the diagonal-matrix identities $(\Phi\mathbf{A})_{ij} = \phi_i \mathbf{A}_{ij}$ and $(\mathbf{A}\Phi)_{ij} = \mathbf{A}_{ij}\phi_j$, gives

$$[\Phi\mathbf{L}\Phi, \mathbf{L}]_{ij} = (\phi_i - \phi_j)\, (\mathbf{L}\Phi\mathbf{L})_{ij}.$$

Here $(\mathbf{L}\Phi\mathbf{L})_{ij} = \sum_k \mathbf{L}_{ik}\, \phi_k\, \mathbf{L}_{kj}$, with non-zero contributions only from common neighbours $k \in N(i) \cap N(j)$, where $\mathbf{L}_{ik}, \mathbf{L}_{kj} < 0$ and $\phi_k > 0$. Each such term is strictly positive, and at least one common neighbour exists by hypothesis, so $(\mathbf{L}\Phi\mathbf{L})_{ij} > 0$. Hence $\phi_i = \phi_j$ for every distance-two pair, and by connectedness of $D^2(G)$, $\phi_i = c$ for all i, with some $c > 0$.

Step 2: Δ cannot commute with $\mathbf{L}$. With $\phi \equiv c$, $\Phi\mathbf{L}\Phi = c^2\mathbf{L}$, so $M = \Delta + 2\tau c^2 \mathbf{L}$. The condition $[M, \mathbf{L}] = 0$ reduces to $[\Delta, \mathbf{L}] = 0$. For any edge $(i, j) \in E$,

$$[\Delta, \mathbf{L}]_{ij} = (\Delta_i - \Delta_j)\, \mathbf{L}_{ij} = 0,$$

and $\mathbf{L}_{ij} \neq 0$ on edges, so $\Delta_i = \Delta_j$ for every edge. By connectedness of G, Δ_i is constant over V.

But $\Delta_i = c^2 d_i^{\gamma-1}$, which is non-constant whenever d is non-constant and $\gamma \neq 1$. This is a contradiction. Therefore no such factorisation exists, and $\mathbf{K}_\gamma$ is not a node-wise rescaling of any stationary graph kernel for $\gamma \neq 1$. □

Algebraic remark on $\gamma = 0$. At $\gamma = 0$, the precision $\mathbf{Q}_0 = \mathbf{D}^{-1} + 2\tau\mathbf{L}$ has off-diagonals $-2\tau w_{ij}$ that do not depend on d, while the diagonal carries the auxiliary as $\mathbf{D}^{-1}$. This is the mass-lumped finite-element discretisation of a variable-coefficient Matérn SPDE with $\mathbf{D}^{-1}$ playing the role of the (mass-lumped) mass matrix and $\mathbf{L}$ the stiffness matrix [19]; we call this kernel `fem` in the family. The further specialisation $\mathbf{D} = \mathbf{I}$ recovers the semi-supervised graph kernel $\mathbf{Q} = \mathbf{I} + 2\tau\mathbf{L}$ of [26, 32], which we call `stat`.

The identity at $\gamma = 0$ does *not* assert that $\mathbf{K}_0$ is stationary: when d is non-constant, $\mathbf{D}^{-1}$ does not commute with $\mathbf{L}$, so $\mathbf{Q}_0$ does not either, and $\mathbf{K}_0$ is genuinely non-stationary on the graph. The uniqueness argument above applies verbatim at $\gamma = 0$ (with $d_i^{1-\gamma} = d_i$ non-constant), confirming that `fem` is not a node-wise rescaling of any stationary graph kernel.

B Practical implementation: full details

This section gives the implementation details deferred from §4 of the main paper: the fitting algorithm and warm-start template (§B.1), the per-variant edit cheatsheet (§B.2), partition sources for the piecewise extension (§B.3), the dataset-specific choice of auxiliary d_i (§B.4), the baseline tuning protocol (§B.5), and numerical detail (§B.6).

B.1 Algorithm and warm-start

Marginal log-likelihood. A VCGP is fitted by joint maximisation of the marginal log-likelihood (MLL) over $(\log\tau, \log s, \log\sigma^2, \gamma)$ on a training subset $\mathbf{S} \subset V$ with targets $y_{\mathbf{S}}$. With the Gaussian observation model $y_{\mathbf{S}} \sim \mathcal{N}(0,\, \mathbf{K}_{\gamma,SS} + \sigma^2\mathbf{I})$,

$$\mathcal{L}(\theta) \;=\; -\tfrac{1}{2}\, y_{\mathbf{S}}^\top (\mathbf{K}_{\gamma,SS} + \sigma^2\mathbf{I})^{-1} y_{\mathbf{S}} \;-\; \tfrac{1}{2}\, \log\det(\mathbf{K}_{\gamma,SS} + \sigma^2\mathbf{I}) \;-\; \tfrac{|\mathbf{S}|}{2}\, \log(2\pi).$$

We optimise with Adam for 200–300 iterations. Gradients flow through $\mathbf{Q}_\gamma$ via standard automatic differentiation; the only non-standard term is the log-determinant of $\mathbf{K}_{\gamma,SS} + \sigma^2\mathbf{I}$, computed via Cholesky.

Stationary warm-start. At $\gamma = 1$, Proposition 1 makes the VCGP equivalent to a stationary kernel rescaled by $\sqrt{d_i}$ on the diagonal. We exploit this for a principled initialisation: first fit the three-parameter `het` (γ frozen at 1), then add γ as a free fourth scalar, initialising it at 1. The MLL is therefore non-decreasing along the warm-start — the `lgamma` fit can only improve on `het`, by construction. We initialise each cluster γ_k at 1 in the piecewise variant `lgamma_block`.

Algorithm 1 Scalar-γ VCGP fit.

Require: Laplacian $\mathbf{L}$, auxiliary $d \in \mathbb{R}^n_{>0}$, training set $\mathbf{S} \subset V$ with targets $y_{\mathbf{S}}$, iterations T.
Ensure: Posterior mean and variance (μ_*, σ_*^2) at every node.

$d_i \leftarrow \mathrm{clip}(d_i/\bar{d},\, 0.1,\, 10)$; $\quad \mathbf{D} \leftarrow \mathrm{diag}(d)$
Warm-start: fit `het` ($\gamma = 1$); initialise $(\log\tau, \log s, \log\sigma^2, \gamma)$ with $(\log\tau_h, \log s_h, \log\sigma_h^2,\, 1)$
for $t = 1, \ldots, T$ **do**
  $\mathbf{Q}_\gamma \leftarrow \mathbf{D}^{-1} + 2\tau\, \mathbf{D}^{-\gamma/2}\, \mathbf{L}\, \mathbf{D}^{-\gamma/2}$
  $\mathbf{K}_\gamma \leftarrow s\,(\mathbf{Q}_\gamma + \varepsilon\mathbf{I})^{-1}$
  Adam step on $-\log p(y_{\mathbf{S}} \mid \theta)$, $\theta = (\log\tau, \log s, \log\sigma^2, \gamma)$
end for
$\mu_* \leftarrow \mathbf{K}_{\cdot\mathbf{S}}(\mathbf{K}_{SS} + \sigma^2\mathbf{I})^{-1} y_{\mathbf{S}}$; $\quad \sigma_*^2(i) \leftarrow \mathbf{K}_{ii} + \sigma^2 - \mathbf{K}_{iS}(\mathbf{K}_{SS} + \sigma^2\mathbf{I})^{-1}\mathbf{K}_{Si}$

B.2 Per-variant edits

Every learned-γ variant is a single-line edit of Algorithm 1 (Table 4). The two refinements commute and compose: `lgamma_fisher_m15` applies both `lgamma_m15` and `lgamma_fisher`.

Table 4: Each variant is a single-line edit of Algorithm 1.

Member	Line	Edit
`lgamma_block`	4	replace $\mathbf{L}$ with $\mathbf{L}_{\mathrm{bl}} = \mathrm{blockdiag}(\mathbf{L}_1, \ldots, \mathbf{L}_K)$; learn $\gamma \in \mathbb{R}^K$
`lgamma_m15`	5	$\mathbf{K}_\gamma \leftarrow s\,(\mathbf{Q}_\gamma + \varepsilon\mathbf{I})^{-2}$
`lgamma_fisher`	6	add $\lambda(\gamma - \gamma_{\mathrm{ref}})^2$ to the loss ($\lambda = 0.5/\mathrm{Var}(\hat{\gamma}_{\mathrm{vanilla}})$, $\gamma_{\mathrm{ref}} = 1$)

Smoother kernels (`lgamma_m15`). Replacing $\mathbf{K}_\gamma = s\,\mathbf{Q}_\gamma^{-1}$ with $\mathbf{K}_\gamma = s\,\mathbf{Q}_\gamma^{-2}$ gives the Matérn-$3/2$ member of the family. This adds no new parameters and is implementable as two back-solves through $\mathbf{Q}_\gamma$.

What the Fisher-information ridge is. The quantity used is an entry of the full inverse Hessian. We first fit the plain model by marginal likelihood over $(\tau, s, \sigma_\varepsilon^2, \gamma)$; at the estimate we form the 4×4 Hessian $\mathbf{J}$ of the negative log marginal likelihood, invert it, and take $\mathrm{Var}(\hat{\gamma}) = \mathbf{J}^{-1}(\gamma, \gamma)$, and add a quadratic penalty pulling γ toward the factorisation point with weight $\lambda = 0.5/\mathrm{Var}(\hat{\gamma})$. To quadratic order this pulls $\hat{\gamma}$ exactly halfway to $\gamma = 1$: a scale-free shrinkage that needs no tuning.

Stabilisation on small datasets (`lgamma_fisher`). When $|\mathbf{S}|$ is small, the MLL surface for γ can be flat. We add a quadratic ridge $\lambda(\gamma - 1)^2$ anchored at the stationary point, with λ set automatically from the variance of the unregularised $\hat{\gamma}$ on the same training set. The penalty is read once off the inverse Hessian of the vanilla fit; no manual tuning is required.

What the block model assumes. The *full-edge* form keeps every edge and varies only the exponent by region, one γ_k per cluster. The *block* form, `lgamma_block`, additionally replaces $\mathbf{L}$ by $\mathrm{blockdiag}(\mathbf{L}_1, \ldots, \mathbf{L}_K)$, removing cross-cluster edges and so making the clusters conditionally independent. This is a modelling assumption and we state it as such; it costs little where regions are genuinely near-independent, as with CA Weather's distinct climate zones. Whether it pays is empirical: on PM2.5, keeping every edge costs $+0.044 \pm 0.007$ negative log-likelihood per test point (NLL/pt) against the block variant. That comparison comes from the validation-fold runs of §L (full-edge 0.9918 against block 0.9482 on the held-out fold), not from the $70/30$ protocol used elsewhere in this section, because the full-edge variant was only ever run inside the selection grid.

B.3 Partition sources for the piecewise extension

For a fixed cluster count K, the partition is computed before fitting the per-cluster exponents. We use three partition rules:

- PLAIN K-means on node coordinates $[\mathrm{lat}_i, \mathrm{lon}_i]$ (`lgamma_block`);
- AUGMENTED K-means on $[\widetilde{\mathrm{lat}}_i, \widetilde{\mathrm{lon}}_i,\ w{\cdot}\tilde{d}_i]$, where tildes denote standardisation of each coordinate and the auxiliary, and $w \geq 0$ controls the auxiliary's influence (`lgamma_block_aug`, labelled `block-aug` in Table 2);
- SPECTRAL K-means on the first K eigenvectors of the unnormalised graph Laplacian, with K minimising the training-data criterion $-2\log p(y_{\mathbf{S}} \mid \hat{\theta}_K) + K\log|\mathbf{S}|$ over the fitted grid (`spectral_bic_lgamma_block`, labelled `spectral-BIC` in Table 2).

All three feed the same `lgamma` fit, with per-cluster γ_k initialised at 1. For each partition, the implementation rebuilds a symmetric geographic nearest-neighbour graph within each cluster, using coordinates alone for edge construction and excluding cross-cluster edges. Thus the block Laplacian is formed from these within-cluster graphs, rather than merely deleting edges from the original graph. The auxiliary affects the augmented partition and the precision, but does not enter the within-cluster distance calculation.

B.4 Variance field d_i

The choice of d_i is dictated by the averaging window of the predicted signal, not by what is convenient to compute.

(i) Daily snapshots with similar baselines (METR-LA, PEMS-BAY, LAQN). Temporal coefficient of variation $d_i = \mathrm{std}/\mathrm{mean}$ across the source readings, or, on LAQN (no temporal series), a six-level site-type ordinal $\{\textsc{Rural}\to 0, \ldots, \textsc{Kerbside}\to 5\}$, normalised to mean 1 and then shifted by $+0.1$ before the clip, so that the lowest category is strictly positive as Definition 1 requires (only five of the six categories occur in the cached data).

(ii) Aggregates over longer windows (PM2.5 annual mean from $\sim$365 daily reads). The aggregate's residual variance scales with daily-std$^2/N$, not CV^2, and dividing by the mean discards level information; we use $d_i =$ annual std on PM2.5 and 2022 daily-mean variance on CA Weather.

The choice of d_i for each dataset is post-hoc. On PM2.5, Table 15 compares the external standard-deviation auxiliary with local variability estimated from training targets; it does not compare standard deviation with coefficient of variation.

B.5 Baseline tuning protocol

Per-baseline formulas and parameter counts. `m05`, `m15`, `lnu`: spectral graph Matérn kernels at $\nu = 1/2$, $\nu = 3/2$, and learnable ν respectively, parametrised by lengthscale ℓ and amplitude; 3 parameters (`m05`, `m15`) or 4 (`lnu`). `ps`: Paciorek–Schervish kernel with per-node lengthscale $\ell_i = \ell_0/\sqrt{d_i}$ tied through one global ℓ_0, 3 parameters. `fuglreg_lam`λ: our graph adaptation of variable-coefficient Matérn SPDE models [9, 19], with per-node κ_i and graph smoothness penalty $\lambda\sum_{(i,j)\in E}(\log\kappa_i - \log\kappa_j)^2$, $n+3$ parameters; the closest per-node analogue of our one-scalar construction. `se_kriging`: a squared-exponential-kernel spatial GP [24] fitted by grid search over lengthscale, amplitude and noise; 3 parameters, and the standard kriging predictor of classical spatial statistics [7]. This is the spatial-kriging baseline reported on all five datasets. An implementation of Bayesian kernelised tensor regression [17] ships with the code for reference, but its output is not the row reported here.

Fitting protocol. Stationary Matérn baselines (`m05`, `m15`, `lnu`) share the family's Adam-MLL protocol; `ps`, `se_kriging`, and `fuglreg_lam`λ are fitted as published. We sweep $\lambda \in \{10^{-2}, 10^{-1}, 1, 10, 10^2\}$ per dataset and report the post-hoc best as `fuglreg_lam1.0`; this is fixed-grid, not cross-validated per seed, and the neighbouring λ values are substantially worse.

Recalibrated NLL/pt. The recalibrated NLL fits a single test-side scalar

$$c^* = \arg\min_c \mathrm{NLL}(c\hat{\sigma}^2)$$

that rescales predictive variances. This is reported as a calibration diagnostic, not as a deployable correction — a deployable corrector would have to be fit on a held-out validation set, not on the test set.

The two remaining outer baselines. `specpoly` is a spectral-polynomial kernel, $\mathbf{K} = g(\mathbf{L})$ with g a learned degree-3 polynomial in the Laplacian spectrum — the most flexible *spectral-stationary* comparator, and so a direct test of whether the gains come from non-stationarity rather than from a richer spectral filter. `fuglstad` is the variable-coefficient SPDE with per-node κ_i and *no* smoothness penalty, i.e. `fuglreg_lam`λ at $\lambda = 0$; it is reported to show what the per-node construction does when its $O(n)$ parameters are left unregularised.

Significance testing. Per-row paired one-sided t-tests against the dataset leader are computed across the 20 seeds, with all methods evaluated on the same train/test splits to control for split-level variance. The leaderboards and the CA Weather monthly analysis use one-sided tests throughout; the two-method comparisons quoted in §E are two-sided, which is the more conservative choice when neither direction is privileged, and are labelled as such there. Standard errors in the leaderboards are seed-level SEs ($s/\sqrt{20}$).

Two caveats on reading the leaderboards. First, on the three small single-snapshot graphs the twenty seeds are re-splits of *one* snapshot, not independent replications, and the test sets are small ($n_{\text{test}} \approx 16$ on LAQN, 63 on METR-LA). We therefore rest the multiplicity-corrected claim on CA Weather's twelve monthly snapshots and read the single-snapshot datasets as corroborating rather than carrying the result. Second, both sides of the leaderboard comparison use post-hoc-best settings: the baselines keep their best sweep value — `fuglreg_lam1.0` is the best of a five-point λ sweep — and the reported family member and per-dataset d_i are likewise chosen post hoc. The comparison is therefore not conservative in either direction. We also run the selection mechanically on a disjoint validation fold (§L, §M); that honest protocol is worse than reading the leaderboard on every dataset, by 0.04 nat/pt on PM2.5 and 0.07 on CA Weather, rising to 0.23–0.28 on the three small graphs, where it also fits on a smaller $50/20/30$ split.

B.6 Numerical detail

M-matrix property. $\mathbf{Q}_\gamma = \mathbf{D}^{-1} + 2\tau\,\mathbf{D}^{-\gamma/2}\mathbf{L}\mathbf{D}^{-\gamma/2}$ has positive diagonal $\mathbf{D}_{ii}^{-1} + 2\tau d_i^{-\gamma}\mathbf{D}_{ii}^{\deg}$ and non-positive off-diagonal $-2\tau d_i^{-\gamma/2} d_j^{-\gamma/2} w_{ij}$ ($w_{ij} \geq 0$ on edges, zero off-edges). It is therefore a Stieltjes M-matrix: positive definite, with non-negative inverse $\mathbf{K}_\gamma$. Entrywise positivity is preserved under the block-diagonal generalisation $\mathbf{L}_{\text{bl}}$.

$\mathbf{Q}_\gamma$ inversion jitter. Direct inversion of $\mathbf{Q}_\gamma$ uses sparse LU (`scipy.sparse.linalg.splu`) when $\text{nnz}(\mathbf{Q}_\gamma) < n^2/4$ and dense `linalg.inv` otherwise. A small jitter $10^{-8}\,\mathbf{I}$ is added to the diagonal before inversion.

Cholesky jitter ladder. For the Matérn-$3/2$ refinement $K = s\,\mathbf{Q}_\gamma^{-2}$, the squared precision raises the condition number; the Cholesky of $\mathbf{K}_{SS} + \sigma^2\mathbf{I}$ uses an escalating jitter sequence to recover from rare `LinAlgError` events: $\{10^{-8}, 10^{-7}, \ldots, 10^{-2}\}$ in the shared kernel code and the shorter $\{10^{-6}, \ldots, 10^{-2}\}$ on the $\nu = 3/2$ path. The ladder fires more often at $\nu = 3/2$ than at $\nu = 1/2$, as the squared precision would suggest, and in no case does it persist past a single Adam iteration.

Stability under d_i rescaling. Multiplying d_i by a global constant rescales $\mathbf{D}^{-1}$ and $\mathbf{D}^{-\gamma/2}$ jointly; the construction is invariant up to a redefinition of τ and s. We normalise d_i to mean 1 before fitting.

C Synthetic experiments

Two controlled synthetic studies accompany the real-data analysis: a single-regime grid that verifies scalar γ recovery (a prerequisite for any real-data $\hat{\gamma}$ to be meaningful) and a two-regime stochastic block model that tests the piecewise extension on a multi-regime ground truth.

C.1 Single-regime identifiability of scalar γ

Setup. A 20×20 regular grid ($n = 400$); $\mathbf{D} = \mathrm{diag}(d_i)$ varies linearly from 0.2 to 2.0 along the horizontal coordinate so the heterogeneity needed to identify γ is present. For each ground truth $\gamma_{\mathrm{true}} \in \{-1, 0, +1, +3\}$ and each training fraction $|\mathbf{S}|/n \in \{0.1, 0.2, 0.3, 0.5, 0.7, 0.9\}$ we draw 10 seeds and fit scalar `lgamma` by joint MLL (Adam, 300 steps, warm-started at $\gamma = 1$).

Result. Recovery is within 0.2 of the truth at $|\mathbf{S}|/n \geq 0.5$ for every γ_{true} tested (Figure 3). At smaller training fractions $\hat{\gamma}$ overshoots in magnitude — it lands *further* from the stationary point than the truth, not nearer it, on every γ_{true} tested (at $\gamma_{\mathrm{true}} = -1$ the mean at $|\mathbf{S}|/n = 0.2$ is -1.68) — and the bias decays as $|\mathbf{S}|/n$ grows. The joint MLL is identifiable in γ on a graph with non-constant d_i.

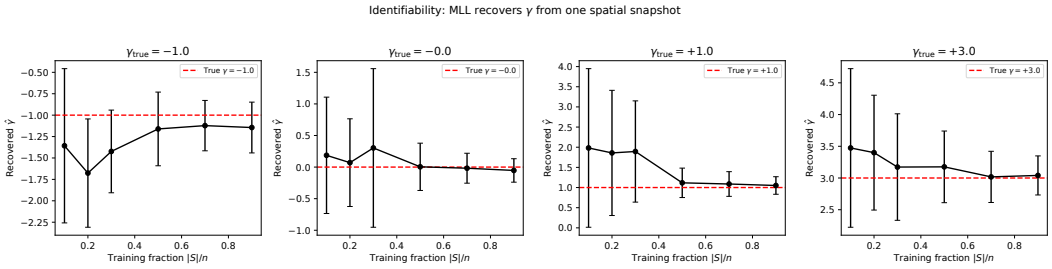

Figure 3: Identifiability on a 20×20 regular grid. Recovered $\hat{\gamma}$ (mean $\pm$ standard deviation over 10 seeds) versus training fraction $|\mathbf{S}|/n$, for each of four ground truths $\gamma_{\mathrm{true}} \in \{-1, 0, +1, +3\}$. The dashed line marks $\hat{\gamma} = \gamma_{\mathrm{true}}$. Recovery is within 0.2 at $|\mathbf{S}|/n \geq 0.5$ for every ground truth.

C.2 Two-regime stochastic block model

Setup. A controlled two-regime stochastic block model: $n = 200$, 100 vertices per block; sparse block A ($p_{\mathrm{in}} = 0.04$, expected degree ≈ 4), dense block B ($p_{\mathrm{in}} = 0.20$, expected degree ≈ 20), $p_{\mathrm{between}} = 0.005$; noise $\sigma^2 = 0.05$; ground truths $\gamma_{\mathbf{A}} = -1$ (sparse-block propagation) and $\gamma_{\mathbf{B}} = +2$ (dense-block isolation).

Result. Six methods are run. The stationary Matérn baselines `m05` and `m15` pay an irreducible $+0.80$ and $+0.83$ nat/pt against the leader because no scalar exponent fits both blocks, and the stationary family members bracket them (`het` $+0.67$, `stat` $+1.06$). Scalar `lgamma` collapses to a single compromise $\hat{\gamma} \approx +4.2$ and places second of the six at 1.561. `lgamma_block` ($K = 2$, true block labels) recovers $\hat{\gamma}_{\mathbf{A}} = -0.77 \pm 0.10$, $\hat{\gamma}_{\mathbf{B}} = +2.24 \pm 0.06$ with the correct signs and within 0.25 of each ground truth, and takes top-1 NLL/pt at 1.380, 0.18 nat/pt clear of the next best (Figure 4). The piecewise extension is therefore necessary, not ornamental: scalar models cannot fit a multi-regime truth.

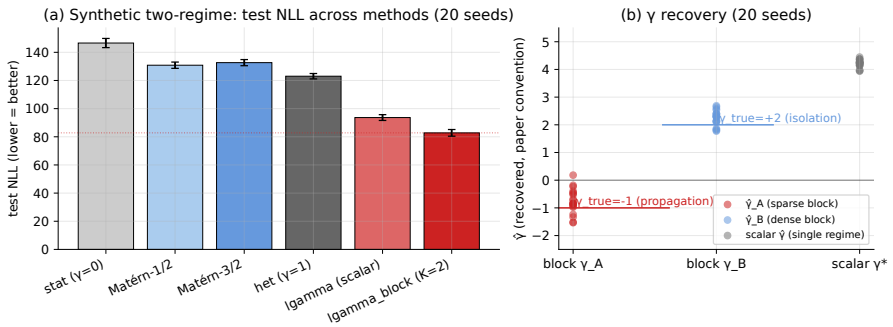

Figure 4: Two-regime SBM. *Left:* per-method test NLL/pt. *Right:* recovered $\hat{\gamma}_k$ vs. ground truth.

D Full per-dataset leaderboards with SE and significance

The main paper reports the top five rows of each leaderboard. The tables below give the full 20-seed leaderboard with mean $\pm$ standard error (SE = std/$\sqrt{20}$) on all three metrics, plus a paired one-sided t-test p-value asking whether the dataset leader has significantly lower NLL/pt than the row. SE captures seed-to-seed variability; $p < 0.05$ marks a row that the leader beats with high confidence on raw NLL/pt. Rows with $p > 0.05$ are within the leader's noise band on NLL/pt; some of those still lose on RMSE or recNLL/pt (or vice versa). On METR-LA the NLL leader is `lgamma_fisher`, with scalar `lgamma` inside its noise band ($p = 0.44$) and `lgamma_fisher_m15` taking RMSE — a Pareto split. On PEMS-BAY the NLL leader is `lgamma_fisher`; `het` is within its noise band ($p = 0.17$) and has the lowest RMSE.

Table 5: METR-LA: full 20-seed leaderboard with mean $\pm$ SE; p-values from a paired one-sided t-test asking whether the dataset leader has significantly lower NLL/pt than the row. The p column is the test against the dataset leader; the last column flags whether the row is a family member or an outer baseline.

Method	NLL/pt	RMSE	recNLL/pt	p vs. leader	cat
`lgamma_fisher`	1.2018 ± 0.0287	0.8369 ± 0.0238	1.1763 ± 0.0232	—	family (leader)
`lgamma`	1.2025 ± 0.0294	0.8468 ± 0.0247	1.1759 ± 0.0228	0.437	family
`het`	1.2201 ± 0.0265	0.8469 ± 0.0231	1.1975 ± 0.0234	0.002	family
`lgamma_fisher_m15`	1.2202 ± 0.0344	0.8265 ± 0.0235	1.1816 ± 0.0247	0.049	family
`lgamma_m15`	1.2234 ± 0.0349	0.8354 ± 0.0240	1.1835 ± 0.0244	0.039	family
`fuglreg_lam1.0`	1.2382 ± 0.0294	0.8603 ± 0.0232	1.2064 ± 0.0224	$<10^{-3}$	base
`spectral_bic_lgamma_block`	1.2486 ± 0.0341	0.8828 ± 0.0242	1.2184 ± 0.0254	0.027	family
`fem`	1.2537 ± 0.0261	0.8480 ± 0.0225	1.2301 ± 0.0240	$<10^{-3}$	family
`stat`	1.2612 ± 0.0265	0.8491 ± 0.0227	1.2368 ± 0.0244	$<10^{-3}$	family
`m15`	1.2790 ± 0.0301	0.8302 ± 0.0222	1.2448 ± 0.0243	$<10^{-3}$	base
`m05`	1.2881 ± 0.0255	0.8802 ± 0.0237	1.2669 ± 0.0238	$<10^{-3}$	base
`lnu`	1.2939 ± 0.0330	0.8307 ± 0.0224	1.2515 ± 0.0244	$<10^{-3}$	base
`se_kriging`	1.3289 ± 0.0266	0.8681 ± 0.0169	1.2794 ± 0.0193	$<10^{-3}$	base
`ps`	1.5411 ± 0.0705	0.9239 ± 0.0213	1.3306 ± 0.0257	$<10^{-3}$	base
`specpoly`	1.7334 ± 0.0256	1.2427 ± 0.0292	1.6915 ± 0.0330	$<10^{-3}$	base
`fuglstad`	1.8988 ± 0.1073	0.9213 ± 0.0267	1.2671 ± 0.0264	$<10^{-3}$	base
`dk`	2.7214 ± 0.2608	1.1237 ± 0.0296	1.7718 ± 0.0441	$<10^{-3}$	learned baseline

Table 6: PEMS-BAY: full 20-seed leaderboard with mean $\pm$ SE; p-values from a paired one-sided t-test asking whether the dataset leader has significantly lower NLL/pt than the row. The p column is the test against the dataset leader; the last column flags whether the row is a family member or an outer baseline.

Method	NLL/pt	RMSE	recNLL/pt	p vs. leader	cat
`lgamma_fisher`	1.2720 ± 0.0346	0.9015 ± 0.0360	1.2282 ± 0.0371	—	family (leader)
`het`	1.2767 ± 0.0329	0.8906 ± 0.0329	1.2344 ± 0.0362	0.166	family
`lgamma`	1.2835 ± 0.0370	0.9019 ± 0.0360	1.2374 ± 0.0381	0.010	family
`lgamma_fisher_m15`	1.2911 ± 0.0359	0.8963 ± 0.0329	1.2459 ± 0.0371	$<10^{-3}$	family
`spectral_bic_lgamma_block`	1.2934 ± 0.0330	0.9249 ± 0.0359	1.2582 ± 0.0329	0.009	family
`lgamma_m15`	1.2965 ± 0.0366	0.8965 ± 0.0329	1.2504 ± 0.0371	$<10^{-3}$	family
`fem`	1.3245 ± 0.0320	0.8908 ± 0.0323	1.2798 ± 0.0360	$<10^{-3}$	family
`fuglreg_lam1.0`	1.3340 ± 0.0365	0.8948 ± 0.0331	1.2918 ± 0.0349	$<10^{-3}$	base
`m05`	1.3556 ± 0.0356	0.8960 ± 0.0349	1.3025 ± 0.0397	$<10^{-3}$	base
`lnu`	1.3562 ± 0.0349	0.8961 ± 0.0341	1.3048 ± 0.0388	$<10^{-3}$	base
`stat`	1.3594 ± 0.0331	0.8949 ± 0.0326	1.3129 ± 0.0364	$<10^{-3}$	family
`m15`	1.3655 ± 0.0352	0.9016 ± 0.0347	1.3133 ± 0.0386	$<10^{-3}$	base
`se_kriging`	1.3867 ± 0.0321	0.9350 ± 0.0333	1.3372 ± 0.0360	$<10^{-3}$	base
`specpoly`	1.6527 ± 0.0327	1.0999 ± 0.0409	1.5749 ± 0.0453	$<10^{-3}$	base
`fuglstad`	1.7990 ± 0.1087	0.9181 ± 0.0346	1.3485 ± 0.0329	$<10^{-3}$	base
`ps`	1.8058 ± 0.0706	1.0856 ± 0.0258	1.5305 ± 0.0267	$<10^{-3}$	base
`dk`	3.1423 ± 0.3538	1.2189 ± 0.0431	1.9880 ± 0.0520	$<10^{-3}$	learned baseline

Table 7: LAQN: full 20-seed leaderboard with mean ± SE; p-values from a paired one-sided t-test asking whether the dataset leader has significantly lower NLL/pt than the row. The p column is the test against the dataset leader; the last column flags whether the row is a family member or an outer baseline.

Method	NLL/pt	RMSE	recNLL/pt	p vs. leader	cat
`lgamma_m15`	1.4519 ± 0.0374	1.0455 ± 0.0383	1.4068 ± 0.0332	—	family (leader)
`lgamma`	1.4688 ± 0.0400	1.0434 ± 0.0393	1.4171 ± 0.0346	0.001	family
`lgamma_fisher_m15`	1.4741 ± 0.0410	1.0468 ± 0.0386	1.4227 ± 0.0355	$<10^{-3}$	family
`lgamma_fisher`	1.4742 ± 0.0403	1.0460 ± 0.0392	1.4223 ± 0.0347	$<10^{-3}$	family
`spectral_bic_lgamma_block`	1.4779 ± 0.0528	1.0259 ± 0.0396	1.4020 ± 0.0372	0.142	family
`het`	1.5234 ± 0.0419	1.0717 ± 0.0356	1.4676 ± 0.0343	$<10^{-3}$	family
`m05`	1.5244 ± 0.0487	1.0447 ± 0.0397	1.4495 ± 0.0411	$<10^{-3}$	base
`m15`	1.5262 ± 0.0488	1.0444 ± 0.0397	1.4507 ± 0.0412	$<10^{-3}$	base
`se_kriging`	1.5514 ± 0.0551	1.0518 ± 0.0394	1.4586 ± 0.0395	$<10^{-3}$	base
`specpoly`	1.5524 ± 0.0407	1.1300 ± 0.0338	1.5029 ± 0.0305	$<10^{-3}$	base
`lnu`	1.5568 ± 0.0486	1.0505 ± 0.0358	1.4796 ± 0.0380	$<10^{-3}$	base
`fem`	1.5590 ± 0.0446	1.0624 ± 0.0344	1.4917 ± 0.0366	$<10^{-3}$	family
`fuglreg_lam1.0`	1.5895 ± 0.0555	1.0573 ± 0.0373	1.4973 ± 0.0421	$<10^{-3}$	base
`stat`	1.6013 ± 0.0487	1.0634 ± 0.0340	1.5188 ± 0.0381	$<10^{-3}$	family
`ps`	2.2013 ± 0.1941	1.2804 ± 0.0895	1.7665 ± 0.0521	$<10^{-3}$	base
`fuglstad`	2.3218 ± 0.2453	1.1173 ± 0.0442	1.5674 ± 0.0495	$<10^{-3}$	base
`dk`	4.3435 ± 0.2683	1.2220 ± 0.0448	1.6085 ± 0.0419	$<10^{-3}$	learned baseline

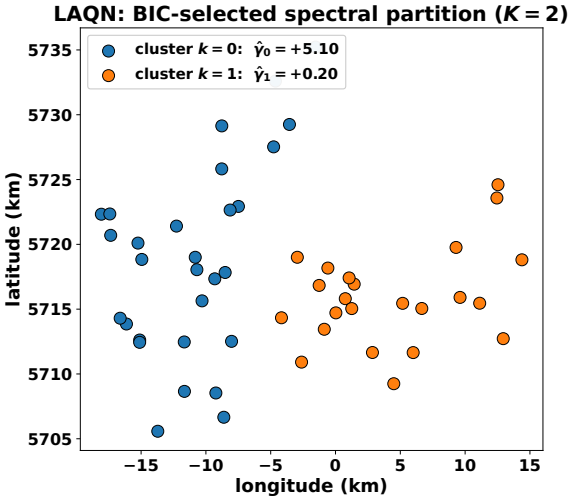

Figure 5: LAQN spectral partition. Sensors are coloured by the two-cluster spectral partition of $\mathbf{L}$ chosen by BIC; the per-cluster $\hat{\gamma}_k$ values are shown. The partition is geographic rather than by site type: the two clusters carry almost the same mix of site classes (mean site-type ordinal 1.13 against 1.06) and separate east from west.

E Deep Ensemble GNN comparison

Question. Deep ensembles are a standard, general-purpose baseline for predictive uncertainty [15], so we include one among our baselines here—a deep ensemble of Graph Convolutional Networks [13] with mean and variance heads—even though whether such ensembles faithfully capture uncertainty on graph-structured data has recently been called into question [29]. Does a neural baseline of this kind close the gap on the family, particularly on calibration (measured by the recalibration scalar $c^\star$ that minimises NLL)?

What this table compares. The rows below put the deep ensemble against scalar `lgamma`, the simplest member of the family and not the reported leader on any dataset, which makes it the hardest comparison for us: on CA Weather the ensemble is ahead of scalar `lgamma` on seven of the twelve months, significantly so on four and significantly behind on two, and over the twelve-month mean

Table 8: PM2.5: full 20-seed leaderboard with mean ± SE; p-values from a paired one-sided t-test asking whether the dataset leader has significantly lower NLL/pt than the row. The p column is the test against the dataset leader; the last column flags whether the row is a family member or an outer baseline.

Method	NLL/pt	RMSE	recNLL/pt	p	cat
`lgamma_block_aug`(K=20, w=1)	0.9756 ± 0.0106	0.6541 ± 0.0055	0.9703 ± 0.0095	—	family (ldr)
`lgamma_block_aug`(K=10, w=1)	1.0078 ± 0.0105	0.6822 ± 0.0075	1.0032 ± 0.0099	$<10^{-3}$	family
`lgamma_block_aug`(K=10, w=0.5)	1.0303 ± 0.0097	0.7011 ± 0.0069	1.0269 ± 0.0090	$<10^{-3}$	family
`lgamma_block_aug`(K=20, w=0.5)	1.0410 ± 0.0121	0.6975 ± 0.0072	1.0338 ± 0.0102	$<10^{-3}$	family
`lgamma_block_aug`(K=5, w=1)	1.0597 ± 0.0099	0.7171 ± 0.0069	1.0553 ± 0.0095	$<10^{-3}$	family
`lgamma_block_aug`(K=5, w=0.5)	1.0698 ± 0.0082	0.7323 ± 0.0062	1.0672 ± 0.0080	$<10^{-3}$	family
`fuglreg_lam1.0`	1.0748 ± 0.0095	0.7467 ± 0.0062	1.0705 ± 0.0082	$<10^{-3}$	base
`lgamma_block_aug`(K=2, w=1)	1.0803 ± 0.0114	0.7012 ± 0.0070	1.0752 ± 0.0107	$<10^{-3}$	family
`m15`	1.1194 ± 0.0087	0.7392 ± 0.0062	1.1164 ± 0.0087	$<10^{-3}$	base
`lnu`	1.1229 ± 0.0084	0.7417 ± 0.0061	1.1200 ± 0.0086	$<10^{-3}$	base
`lgamma_block_aug`(K=2, w=0.5)	1.1249 ± 0.0075	0.7419 ± 0.0060	1.1225 ± 0.0074	$<10^{-3}$	family
`lgamma`	1.1261 ± 0.0086	0.7419 ± 0.0059	1.1230 ± 0.0087	$<10^{-3}$	family
`spectral_bic_lgamma_block`	1.1312 ± 0.0114	0.7415 ± 0.0073	1.1237 ± 0.0099	$<10^{-3}$	family
`stat`	1.1329 ± 0.0080	0.7484 ± 0.0060	1.1297 ± 0.0085	$<10^{-3}$	family
`m05`	1.1358 ± 0.0079	0.7517 ± 0.0063	1.1320 ± 0.0086	$<10^{-3}$	base
`lgamma_m15`	1.1367 ± 0.0086	0.7489 ± 0.0059	1.1336 ± 0.0085	$<10^{-3}$	family
`lgamma_fisher`	1.1443 ± 0.0089	0.7520 ± 0.0063	1.1417 ± 0.0092	$<10^{-3}$	family
`lgamma_fisher_m15`	1.1510 ± 0.0088	0.7588 ± 0.0061	1.1482 ± 0.0090	$<10^{-3}$	family
`fem`	1.1557 ± 0.0080	0.7601 ± 0.0058	1.1530 ± 0.0083	$<10^{-3}$	family
`se_kriging`	1.1921 ± 0.0093	0.7780 ± 0.0060	1.1668 ± 0.0077	$<10^{-3}$	base
`het`	1.2403 ± 0.0092	0.7761 ± 0.0071	1.2369 ± 0.0096	$<10^{-3}$	family
`ps`	1.5547 ± 0.0192	0.9300 ± 0.0081	1.4806 ± 0.0119	$<10^{-3}$	base
`specpoly`	1.6845 ± 0.0072	1.3232 ± 0.0093	1.6821 ± 0.0068	$<10^{-3}$	base
`dk`	2.1634 ± 0.0805	0.9692 ± 0.0131	1.6526 ± 0.0211	$<10^{-3}$	learned baseline
`fuglstad`	2.2946 ± 0.0641	0.7655 ± 0.0064	1.1754 ± 0.0103	$<10^{-3}$	base

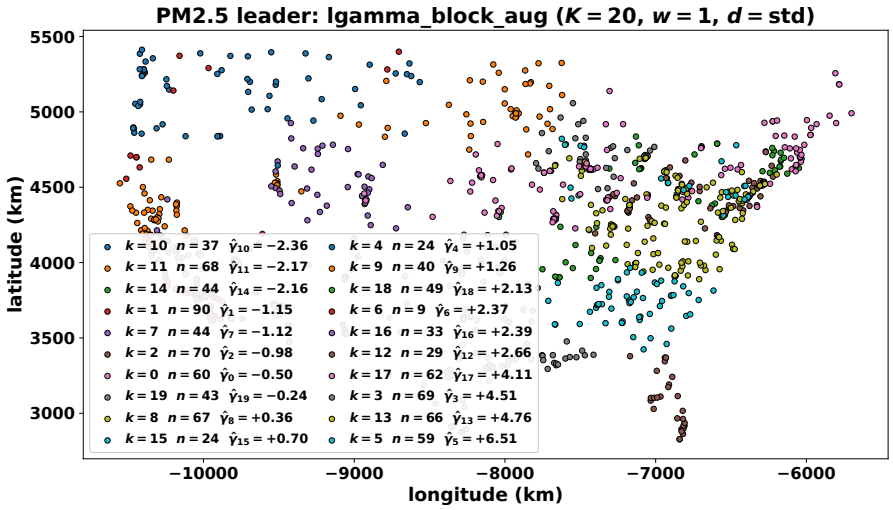

Figure 6: PM2.5 spatial leader. Stations are coloured by the augmented K-means partition ($K = 20$, $w = 1$); the per-cluster $\hat{\gamma}_k$ values are shown. The fitted exponents span both regimes—propagation on some clusters, isolation on others—which is what a single global γ cannot represent.

the two are indistinguishable (0.9963 against 1.0031, $p = 0.55$). That is the expected place for it to be competitive, CA Weather being the dataset a single global exponent cannot fit. Against the *dataset leader* the ensemble loses *on raw NLL* on all five datasets, by 0.13 nat/pt on CA Weather and 0.17 on PM2.5 up to 3.97 on LAQN, with worst-case $p = 0.008$, and on RMSE as well (METR-LA 0.9232 against 0.8265, $p = 0.003$); the deep kernel loses by 1.11 to 2.89. Recalibrated NLL is the

Table 9: CA Weather: 20-seed leaderboard including the Fisher, Matérn and spectral-BIC members; only the aggregate-best augmented configuration is shown (its K-grid is in Table 13). Each entry is the mean over the 12 monthly snapshots, with mean $\pm$ SE across seeds; p-values from a paired one-sided t-test asking whether the dataset leader has significantly lower NLL/pt than the row. The p column is the test against the dataset leader; the last column flags whether the row is a family member or an outer baseline.

Method	NLL/pt	RMSE	recNLL/pt	p	cat
`lgamma_block_aug`$_{K=10,\, w=1}$	0.8685 ± 0.0079	0.5738 ± 0.0049	0.8572 ± 0.0076	—	family (ldr)
`ps`	0.9360 ± 0.0088	0.6261 ± 0.0053	0.9247 ± 0.0078	$<10^{-3}$	base
`se_kriging`	0.9511 ± 0.0061	0.6245 ± 0.0043	0.9376 ± 0.0065	$<10^{-3}$	base
`spectral_bic_lgamma_block`	0.9562 ± 0.0229	0.6151 ± 0.0094	0.9364 ± 0.0163	$<10^{-3}$	family
`fuglreg_lam1.0`	1.0004 ± 0.0087	0.6451 ± 0.0055	0.9900 ± 0.0076	$<10^{-3}$	base
`lgamma`	1.0031 ± 0.0086	0.6562 ± 0.0069	0.9910 ± 0.0089	$<10^{-3}$	family
`lgamma_m15`	1.0184 ± 0.0093	0.6706 ± 0.0070	1.0074 ± 0.0092	$<10^{-3}$	family
`lgamma_fisher`	1.0190 ± 0.0087	0.6556 ± 0.0063	1.0074 ± 0.0090	$<10^{-3}$	family
`lnu`	1.0346 ± 0.0077	0.6522 ± 0.0063	1.0224 ± 0.0083	$<10^{-3}$	base
`m05`	1.0352 ± 0.0072	0.6596 ± 0.0059	1.0220 ± 0.0079	$<10^{-3}$	base
`m15`	1.0375 ± 0.0076	0.6597 ± 0.0056	1.0258 ± 0.0079	$<10^{-3}$	base
`stat`	1.0436 ± 0.0061	0.6551 ± 0.0049	1.0317 ± 0.0067	$<10^{-3}$	family
`lgamma_fisher_m15`	1.0521 ± 0.0092	0.6801 ± 0.0064	1.0420 ± 0.0091	$<10^{-3}$	family
`fem`	1.0673 ± 0.0061	0.6576 ± 0.0050	1.0568 ± 0.0066	$<10^{-3}$	family
`het`	1.1661 ± 0.0055	0.6917 ± 0.0047	1.1556 ± 0.0060	$<10^{-3}$	family
`specpoly`	1.7562 ± 0.0073	1.5168 ± 0.0099	1.7448 ± 0.0084	$<10^{-3}$	base
`dk`	1.9763 ± 0.0689	0.9714 ± 0.0089	1.6279 ± 0.0281	$<10^{-3}$	learned baseline
`fuglstad`	3.2935 ± 0.0758	0.8032 ± 0.0120	1.2211 ± 0.0089	$<10^{-3}$	base

one metric on which this is not uniformly true, and METR-LA is the one dataset: see below. Those comparisons are computed by `graph_gp/compare_learned_baselines.py`.

Setup. We train $M = 5$ independent GCNs (3 graph-convolution layers, 64 hidden units, 300 epochs, Adam $\mathrm{lr} = 10^{-2}$) on the training mask, each with a mean and a log-variance output head and a Gaussian NLL loss. Each node carries two input features, the auxiliary field d_i and the log-degree, so the network sees the same input information as the `lgamma` family. The ensemble prediction averages means and combines within-model and between-model variance in the standard way. We report raw NLL/pt, recalibrated NLL/pt (which fits a single test-side scalar $c^\star$), and $c^\star$ itself; $c^\star \approx 1$ is well-calibrated, $c^\star > 1$ is overconfident, $c^\star < 1$ is underconfident. The paired one-sided t-test asks whether the Deep Ensemble has *larger* NLL than `lgamma`; a dagger marks the reverse direction.

Table 10: Deep Ensemble GNN ($M = 5$) versus `lgamma` across all 16 settings, 20 seeds each. NLL is per-test-point. $c^\star > 1$ means the predictive variance is too small (overconfident).

Setting	$\mathrm{NLL}_{\mathrm{DE}}$	$c^\star_{\mathrm{DE}}$	$\mathrm{NLL}_{\mathrm{lg}}$	$c^\star_{\mathrm{lg}}$	$\mathrm{rec}_{\mathrm{DE}}$	$\mathrm{rec}_{\mathrm{lg}}$	Δ	p(DE $>$ lg)
METR-LA	1.5820 ± 0.1465	2.73	1.2025 ± 0.0294	1.14	1.1417	1.1759	$+0.3796$	0.009^*
PEMS-BAY	1.6402 ± 0.0914	2.49	1.2835 ± 0.0370	1.09	1.3196	1.2374	$+0.3567$	$< 0.001^*$
LAQN	5.4225 ± 0.9784	9.97	1.4688 ± 0.0400	1.20	1.9130	1.4171	$+3.9537$	$< 0.001^*$
PM2.5	1.1474 ± 0.0129	1.27	1.1261 ± 0.0086	1.00	1.1276	1.1230	$+0.0213$	0.020^*
CAW Jan	0.9832 ± 0.0241	1.29	0.9718 ± 0.0169	0.92	0.9574	0.9588	$+0.0114$	0.270
CAW Feb	0.8080 ± 0.0164	1.25	0.7692 ± 0.0129	0.91	0.7894	0.7620	$+0.0388$	0.011^*
CAW Mar	0.9166 ± 0.0136	1.32	0.9206 ± 0.0111	0.93	0.8921	0.9144	-0.0040	0.615
CAW Apr	1.0932 ± 0.0209	1.30	1.0850 ± 0.0121	0.93	1.0646	1.0776	$+0.0082$	0.298
CAW May	1.0433 ± 0.0189	1.32	1.0755 ± 0.0087	0.91	1.0162	1.0698	-0.0322	$0.975^\dagger$
CAW Jun	0.9391 ± 0.0192	1.27	0.9848 ± 0.0078	0.88	0.9180	0.9765	-0.0457	$0.991^\dagger$
CAW Jul	1.1484 ± 0.0345	1.47	1.1030 ± 0.0146	0.94	1.0949	1.0940	$+0.0454$	0.048^*
CAW Aug	1.1287 ± 0.0295	1.38	1.1085 ± 0.0113	0.91	1.0883	1.1007	$+0.0202$	0.190
CAW Sep	1.0165 ± 0.0217	1.28	1.0521 ± 0.0139	0.89	0.9926	1.0398	-0.0356	$0.976^\dagger$
CAW Oct	1.0471 ± 0.0128	1.23	1.0837 ± 0.0071	1.00	1.0320	1.0824	-0.0366	$0.998^\dagger$
CAW Nov	1.1004 ± 0.0353	1.54	1.1231 ± 0.0433	0.92	1.0343	1.0628	-0.0227	0.879
CAW Dec	0.7311 ± 0.0188	1.23	0.7603 ± 0.0118	0.92	0.7134	0.7536	-0.0293	0.925

* $p < 0.05$: Deep Ensemble is significantly *worse* than `lgamma`. † $1 - p < 0.05$: Deep Ensemble is significantly *better*. Paired one-sided t-test over 20 seeds.

Inference — sensor networks. On all four single-snapshot sensor datasets the Deep Ensemble is severely overconfident: $c^\star \in \{1.27, 2.49, 2.73, 9.97\}$ on PM2.5 / PEMS-BAY / METR-LA / LAQN. The predictive variance must be inflated by a factor of roughly 2.5–10 to match the empirical residuals on the three smaller graphs. This is the classical overconfidence failure of deep models trained by maximum likelihood with a variance head: the M networks converge to similar solutions on a single snapshot, collapsing to tight variance estimates. The raw-NLL gap of 0.02–3.95 nat/pt is mostly a calibration failure. After recalibration `lgamma` is still lower on three of the four (PEMS-BAY 1.2374 against 1.3196, LAQN 1.4171 against 1.9130, PM2.5 1.1230 against 1.1276), but METR-LA is a genuine exception and we state it plainly: there the recalibrated ensemble reaches 1.1417 nat/pt, nominally the lowest of any method we run, below the family's best recalibrated value of 1.1759. The difference is not significant against any family member — paired $p = 0.47$ against `lgamma`, 0.44 against `lgamma_fisher`, 0.21 against `het`, and significant only against the stationary baseline `stat` ($p = 0.029$). The reading is that on METR-LA the ensemble's raw-NLL deficit is almost entirely a single global variance scale ($c^\star = 2.73$) rather than a worse mean or a worse variance *shape*, and that the scalar which would close it is fitted on the test set and so is not available at deployment time. This is the one place in the paper where a learned baseline is not behind once miscalibration is removed.

The other learned-representation baseline: a deep-kernel GP. We also run a deep-kernel Gaussian process—a learned feature map with a stationary kernel in feature space—reported as `dk` in the per-dataset leaderboards. It is the weaker of the two learned representations by a wide margin: NLL/pt 2.1634 and RMSE 0.9692 on PM2.5 against the family's 0.9756 and 0.6541, and RMSE 1.1237 on METR-LA against the family's 0.8265. The pattern on the sensor graphs is the same one the Deep Ensemble shows: both are flexible, high-capacity learned representations, and where the observations number far fewer than the nodes they overfit relative to a low-parameter fixed-form kernel. This is the comparison between a learned representation and a structured prior that the one-scalar construction is designed to win in the small-data regime, and it is why we report a fixed-form kernel rather than a learned feature map.

Inference — CA Weather. On the twelve CA Weather months the picture changes. The Deep Ensemble wins significantly on 4 of 12 months, loses significantly on 2, and ties on the remaining 6. Its $c^\star$ values cluster near 1.30 — still overconfident, but within an order of magnitude — while `lgamma`'s $c^\star$ sits near 0.9 (very slightly underconfident). CA Weather has enough spatial structure in the mean (orography, marine-layer gradients) for a 2-layer GCN to extract signal that a shallow GP does not; on sensor networks, where nodes are a sparse sample of a traffic field, the GP with non-stationary coupling wins decisively. We keep the Deep Ensemble GNN as a calibration baseline that complements the Matérn family rather than folding it into the main leaderboards.

F Hyperparameters and numerical settings

All experiments use Adam with default $\beta_1 = 0.9$, $\beta_2 = 0.999$. Hyperparameters are learned in log-space (positivity by reparameterisation). Adam learning rate 10^{-2} for sensor networks (METR-LA, PEMS-BAY, LAQN), 5×10^{-3} for PM2.5 and CA Weather. Iteration count $T_{\text{iter}} \in \{200, 300\}$; convergence is monitored by relative change of MLL ($< 10^{-6}$ over the last 20 steps).

Variance field clipping. d_i is normalised to mean 1 and clipped to $[0.1, 10]$ for two numerical reasons: (a) the diagonal of Q_γ contains the term D^{-1}, so an unbounded $d_i \to 0$ would create an arbitrarily large diagonal entry and worsen $\mathrm{cond}(Q_\gamma)$ by orders of magnitude; (b) the off-diagonal coupling carries $d_i^{-\gamma/2} d_j^{-\gamma/2}$, so an unbounded $d_i \to \infty$ would zero out coupling on that node when $\hat{\gamma}$ is large positive (e.g. $\hat{\gamma} \approx +3$ on METR-LA). The bracket is one decade either side of the normalised mean — wide enough to encode genuine heteroscedasticity, narrow enough to keep the Cholesky stable. As a sensitivity check we re-ran the leakage-free PM2.5 leader (`lgamma_block_aug` with $K = 20$, $w = 1$, $d =$ std) on the same 20 seeds under two alternative brackets. Widening to $[0.05, 20]$ reproduces the default exactly (NLL/pt 0.9756 ± 0.0106, RMSE 0.6541 ± 0.0055, with no per-seed difference at all): the default bracket is not binding on PM2.5, since the normalised d_i distribution lives entirely within $[0.1, 10]$. Tightening to $[0.5, 2]$ shifts NLL/pt to 0.9857 ± 0.0107 and RMSE to 0.6631 ± 0.0070, a paired deterioration of $+0.010$ nat/pt ($p = 0.019$). The effect is therefore statistically detectable but small: the d_i values outside $[0.5, 2]$ do carry some heteroscedastic

information that the kernel exploits. This is a one-sided sensitivity result: the choice of $[0.1, 10]$ is wide enough not to bind on the auxiliary, and tightening it would lose information.

γ bounds. $\gamma \in [-8, 8]$ is enforced softly via a tanh reparameterisation, giving a smooth gradient near the bounds. They are essentially never reached: on the PM2.5 leader 3 of 400 cluster exponents exceed 7.5 in magnitude (largest 7.80) and on the CA Weather leader 3 of 2400 (largest 7.71), so a handful come within 0.2 of the bound while the rest sit well inside it.

Float64 throughout. All Cholesky factorisations and matrix inversions are performed in double precision.

G CA Weather monthly leaderboards

The aggregate-selected augmented-K-means member at $(K, w) = (10, 1)$ beats the outer baselines on mean NLL/pt and RMSE in each monthly snapshot. Table 11 compares this fixed member with the lowest-NLL non-augmented comparator each month, including the spectral-BIC and Fisher members. Other augmented configurations are excluded from this comparison; the fixed member is not the monthly optimum over that grid. The thirteen outer baselines counted here are `m05`, `m15`, `lnu`, `ps`, `se_kriging`, `dk`, `specpoly`, `fuglstad` and the five `fuglreg_lam`λ settings $\lambda \in \{0.01, 0.1, 1, 10, 100\}$, each counted separately rather than collapsed to its best value as in the leaderboards. At $p < 0.05$ the leader's significance against *all* thirteen outer baselines holds in 8 of 12 months on both NLL/pt and RMSE (9 on NLL/pt alone, 10 on RMSE alone); under Bonferroni across the twelve months ($p < 0.0042 = 0.05/12$) five months survive on both, May through September. The shortfalls are concentrated in winter, and in every case the baseline that is not significantly beaten is `ps`, `m05`, `lnu` or `fuglreg_lam1.0`.

Table 11: CA Weather: the aggregate-selected augmented member $(K, w) = (10, 1)$ versus the lowest-NLL comparator each month (20-seed means). Comparators are the non-augmented members in Table 9, including spectral-BIC and Fisher members. Other augmented configurations are excluded; the fixed member need not be the monthly optimum over that grid. RMSE uses the same comparator selected by NLL. Negative gap favours the fixed augmented member.

Month	Comparator	NLL/pt aug.	NLL/pt comp.	ΔNLL/pt	RMSE aug.	RMSE comp.
Jan	`ps`	0.8561	0.8764	−0.020	0.5731	0.5800
Feb	`ps`	0.6390	0.6682	−0.029	0.4657	0.4732
Mar	`ps`	0.8470	0.8683	−0.021	0.5661	0.5867
Apr	`se_kriging`	0.9889	1.0292	−0.040	0.6353	0.6773
May	`se_kriging`	0.8846	0.9726	−0.088	0.5802	0.6368
Jun	`spectral_bic_lgamma_block`	0.8656	0.9308	−0.065	0.5726	0.5958
Jul	`spectral_bic_lgamma_block`	0.9129	1.0171	−0.104	0.5976	0.6872
Aug	`se_kriging`	0.9921	1.0497	−0.058	0.6430	0.6902
Sep	`se_kriging`	0.8879	0.9648	−0.077	0.5771	0.6316
Oct	`ps`	0.9106	0.9293	−0.019	0.5949	0.6102
Nov	`spectral_bic_lgamma_block`	1.0531	1.0621	−0.009	0.6569	0.6750
Dec	`ps`	0.5841	0.5922	−0.008	0.4225	0.4335
mean	(varies)	0.8685	0.9134	−0.045	0.5738	0.6065

Seasonal pattern. The lowest-NLL non-augmented comparator varies: in winter months (Dec–Mar, Oct) the comparator is `ps` — the spatially-varying lengthscale GP, well-suited to the smooth temperature fields that dominate these months, where the per-cluster $\hat{\gamma}_k$ spread is narrow. Mild propagation ($\hat{\gamma}_k < 0$) prevails in January, October and November (7, 6 and 6 of ten clusters), while February, April and December split evenly and March tips the other way (4 of ten), so the winter pattern is a narrow spread rather than a uniform sign. In late spring and summer (Apr–Sep) the comparator is `se_kriging` or `spectral_bic_lgamma_block` — methods that exploit more variance-field structure when the per-cluster $\hat{\gamma}_k$ spread widens; July in particular carries a strong-isolation spike ($\hat{\gamma}_k \approx +5.9$) coexisting with the usual propagation ranks, consistent with the seasonal $\hat{\gamma}_k$ heatmap shown in the main paper (§5, CA Weather).

H Scalability of the kernel factorisation

The precision $Q_\gamma = D^{-1} + 2\tau\, D^{-\gamma/2} L D^{-\gamma/2}$ inherits the sparsity pattern of L, so the dominant cost of the GP fit — the factorisation of Q_γ — reduces from $\mathcal{O}(n^3)$ (dense inverse) to roughly $\mathcal{O}(n)$ on planar-like graphs (sparse LU on the k-NN or distance-threshold adjacency). We measure this directly: for each real graph at the canonical operating point ($\gamma = 1$, $\tau = 1$), we time the dense inverse and the sparse LU factorisation (`scipy.sparse.linalg.splu`) on a single CPU and report the minimum of five runs. Table 12 shows both columns. The sparse column is the one the argument rests on, and it behaves as the construction predicts: it grows smoothly with n, from 0.1 ms at $n = 52$ to 4.5 ms at $n = 988$. The sub-millisecond LAQN entries are noise-bound. The per-dataset ratio does not increase monotonically with n, for the reason given after the table.

Table 12: Factorisation of Q_γ for the five real graphs at $\gamma = 1$, $\tau = 1$. Minimum of five runs on a single CPU; "dense" is dense inverse, "sparse" is sparse LU.

Dataset	n	mean deg.	dense (ms)	sparse (ms)	speed-up
LAQN	52	4.9	0.1	0.1	1.2×
METR-LA	207	12.7	71.2	1.0	72.9×
PEMS-BAY	325	12.8	7.0	1.6	4.5×
CA Weather	649	12.2	26.5	2.7	9.7×
PM2.5	988	10.1	47.2	4.5	10.4×

The PM2.5 factorisation benchmark uses a 988-vertex graph, including one vertex excluded by the regression loader, which uses 987 vertices. This single-vertex difference does not change the factorisation scaling discussed here.

End-to-end cost of a fit and a prediction. The table above times the factorisation of $\mathbf{Q}_\gamma$ alone. The cost a practitioner actually pays also includes the autodiff-traced build of $\mathbf{Q}_\gamma$, its inverse, the backward pass, and the posterior Cholesky. At the largest operating point (PM2.5, $n = 987$, $|\mathbf{S}| = 690$) on a single CPU core, one Adam step costs 804 ms, so a full 300-step fit takes about 241 s, and prediction is a single posterior solve at 15 ms. Two honest points. First, that solve is the standard exact-GP cost: every baseline pays the same dense $\mathcal{O}(|\mathbf{S}|^3)$, and it is unrelated to our construction. Second, materialising $\mathbf{K}_\gamma$ does not by itself exploit the sparsity of $\mathbf{Q}_\gamma$; the sparse factorisation timed above is 4–73× faster than a dense inverse, but forming the explicit inverse from that factor still costs n back-solves. At the scales in this paper the fit is minutes and never a bottleneck; the sparsity of $\mathbf{Q}_\gamma$ becomes the operative lever only at larger n, where one keeps the factorisation implicit (matrix-free solves) rather than materialising $\mathbf{K}_\gamma$.

The METR-LA 72.9× ratio comes from its *dense* column, not its sparse one. Its sparse factorisation takes 1.0 ms against PEMS-BAY's 1.6 ms, in line for the two graph sizes, so nothing unusual happens on the sparse side; what is out of line is a dense inverse of 71.2 ms at $n = 207$ against 7.0 ms at $n = 325$. We therefore do not read the ratio as a property of the METR-LA adjacency, and the remaining four rows, whose dense times do order with n, are the ones to compare. The qualitative claim is unaffected: on every graph here the sparse factorisation is the cheaper path, and it is the only one whose cost is predictable from n.

I Long-form heterophily framing

Weak heterophily at the kernel level. Under the weak (absence-of-homophily) notion of heterophily, a canonical illustration is the gap between Graph Convolutional Networks [13] and Graph Attention Networks [28]. GCN averages features uniformly over the neighbourhood and is known to struggle on graphs that are not homophilic; GAT learns non-negative attention weights that can be driven close to zero on dissimilar edges, down-weighting coupling without ever taking negative values. VCGPs play the same game at the kernel level. The precision Q_γ is an M-matrix and K_γ is entrywise positive, so the family cannot produce negative couplings; but a sufficiently positive $\gamma > 1$ — or, more flexibly, a per-cluster $\gamma(x)$ that takes a large positive value on one block — shrinks the effective edge coupling $d_i^{-\gamma/2}\, d_j^{-\gamma/2} w_{ij}$ toward zero on the corresponding pairs.

Strong heterophily and the signed Laplacian. The strong notion of heterophily — neighbours with literally anti-correlated labels — is beyond what any M-matrix kernel can express, because it

would need off-diagonal couplings of both signs. A signed-Laplacian extension would replace L with $L_s = D_{|w|} - W_s$ where W_s admits both signs; the precision loses the M-matrix property, the kernel ceases to be entrywise positive, and MLL is no longer convex in the natural parameters. The sign constraint can be relaxed locally, e.g. via a learned per-edge sign $\sigma_{ij} \in \{-1, +1\}$ with a sparse prior, leaving the bulk of the kernel positive but allowing isolated anti-correlated edges. We flag this as the natural future-work route into the strong-heterophily regime.

Spatio-temporal continuation. The current paper takes γ to be a scalar or piecewise-constant spatial function. A natural continuation is $\gamma(x, t)$ — the exponent varying jointly in space and time on a graph that is repeatedly observed. The 12 CA Weather monthly fits are slices of one such field; reading them as a single $\gamma(x, t)$ would let the variance-coupling regime track seasonal transitions explicitly.

J Per-dataset coverage of Proposition 1 hypotheses

Proposition 1 requires three conditions: (i) G connected, (ii) $D^2(G)$ connected, and (iii) d_i non-constant. Conditions (i) and (iii) hold trivially on every dataset (we always operate on the largest connected component of G, and d is non-constant by construction). Condition (ii) bites on sparse-fringe datasets where distance-threshold or k-NN construction leaves boundary nodes whose two-step neighbourhoods do not link back to the bulk. We restrict to the largest $D^2(G)$-connected component and report the fraction of vertices it covers:

Dataset	vertices in largest $D^2(G)$-connected component
METR-LA ($n = 207$)	99.5%
PEMS-BAY ($n = 325$)	98.2%
LAQN ($n = 52$)	100.0%
PM2.5 ($n = 987$)	98.4%
CA Weather ($n = 649$)	80.3%

These are exact fractions, computed directly from the loaders the experiments use. LAQN is covered entirely and the other three sensor/station networks on 98–99.5% of vertices. On METR-LA and PEMS-BAY the exclusions are isolated points (1 and 6 singletons); on PM2.5 the 16 excluded vertices form two small components beside the bulk, only one of them a singleton. CA Weather is the one dataset where the hypothesis bites: its distance-threshold graph has a sparse fringe of desert and coastal stations outside the 50 km threshold, leaving 30 distance-two components of which the largest covers 80.3% of vertices. On that core the proposition applies, and the excluded boundary forms small components that contribute trivially factored blocks and do not affect the global factorisation argument. The numbers are produced by `graph_gp/prop1_coverage.py`.

Implication for the family. For every dataset in this paper the proposition governs essentially all of the non-stationary precision: $\gamma = 1$ is the only point at which the kernel reduces to a stationary kernel up to a diagonal rescaling, and the family-leader analysis in Section 5 of the main paper is consistent with that interpretation (on PEMS-BAY `het` is statistically indistinguishable from the leading member and has the lowest RMSE, that dataset's variance-coupling regime sitting at $\gamma = 1$; the other four datasets prefer γ values away from 1).

K K-grid for the piecewise variants

Table 13 reports fixed-K NLL/pt, including every LAQN setting that was run. Bold entries mark the aggregate NLL-selected augmented configurations on PM2.5 and CA Weather; LAQN instead uses BIC selection, whose aggregate is reported separately. The PM2.5 configurations use the standard-deviation auxiliary throughout; the table does not compare different auxiliary fields.

Spectral-BIC selection on LAQN. At $n = 52$ and training set size $|S| = 36$, the BIC penalty $\log 36 \approx 3.58$ keeps the partition binary on 19 of 20 seeds; the one remaining seed selects $K = 3$. We report this explicitly to avoid overstating the selection (the family member is robust under the K-grid but not unanimous on $K^\star$).

Table 13: Fixed-K NLL/pt for piecewise members. Bold entries are the aggregate NLL-selected configurations for PM2.5 and CA Weather. LAQN reports every fitted fixed K; its BIC-selected aggregate is stated below and is not the NLL leader. Dashes denote unrun settings. SEs and tests for the selected members are in the per-dataset leaderboards.

Dataset	Partition	K=2	K=3	K=4	K=5	K=7	K=10	K=20
LAQN	spectral	1.4855	1.5011	1.5559	2.1434	1.5386	—	—
PM2.5 (d=std)	augmented (w=1)	1.0803	—	—	1.0597	—	1.0078	**0.9756**
CA Weather	augmented (w=1)	0.9040	—	—	0.8866	—	**0.8685**	0.9081

LAQN: spectral-BIC selects $K = 2$ on 19 seeds and $K = 3$ on 1 seed; its selected aggregate is 1.4779 NLL/pt. This member leads LAQN on RMSE and recalibrated NLL, whereas `lgamma_m15` leads on raw NLL.

Augmented-K-means feature weight w. The feature weight $w \geq 0$ in $[\mathrm{lat}_i, \mathrm{lon}_i, w \cdot \tilde{d}_i]$ controls how strongly the $\tilde{d}_i$ field shapes the partition. We ran two settings, $w \in \{0.5, 1\}$, and report $w = 1$ on both datasets because it is the better of the two by a clear margin: halving the weight costs 6.7% NLL/pt on PM2.5 at $K = 20$ (1.0410 against 0.9756) and 4.0% on CA Weather at $K = 10$ (0.9033 against 0.8685). The partition is therefore not insensitive to w, and we make no claim about values outside the two we measured.

L Validation-fold selection

Table 14 reports the mechanical selection described in the recipe of Section 5.3. *Honest* selects the member, K, d and λ on a validation fold and scores once on a disjoint test fold; *oracle* is the post-hoc best member on the same split, so the gap is the price of honest selection. The gap is paired across seeds.

Table 14: Validation-fold selection. The price of honest selection is small exactly where the method works and large exactly where it does not.

Dataset	honest NLL/pt	RMSE	oracle	gap	versus baselines
PM2.5 (20 seeds)	$\mathbf{0.9813 \pm 0.014}$	0.6690	0.9446	+0.037	beats all (`fuglreg` 1.0748, `m15` 1.1194)
CA Weather (12 months)	$\mathbf{0.9300 \pm 0.048}$	0.6069	0.8637	+0.066	ties `ps` 0.9360; beats the rest
LAQN (20 seeds)	1.6984 ± 0.150	1.0047	1.4666	+0.232	below the fixed member `lgamma`
PEMS-BAY (20 seeds)	1.4865 ± 0.067	0.9763	1.2445	+0.242	below the fixed member `het`
METR-LA (20 seeds)	1.3249 ± 0.112	0.8628	1.0463	+0.279	below the fixed member `het`

The honest-minus-oracle gap is the informative column. On the two data-rich, multi-regime datasets the cost of honest selection is small, and validation lands on a block member on 36 of the 40 PM2.5 fits (the exceptions being three *full-edge* augmented members and one `lgamma_fisher`). Two details indicate that nothing was cherry-picked: with d a selected axis, validation picks the external annual-standard-deviation auxiliary on 19 of 20 PM2.5 seeds rather than any signal-derived alternative, and with λ selected it picks the unregularised member on 19 of 20. The latter reconciles with the `lgamma_fisher` default rather than contradicting it: the Fisher ridge is the within-family NLL leader on both traffic graphs, and validation correctly switches it off where n is large. We do not claim this tracks how well γ is identified: by the curvature measure of §C.1 it is LAQN, the smallest graph, where γ is identified least well, and there the ridge loses. It is not uniformly better inside the family — on LAQN the Matérn-$3/2$ member leads and the ridge costs 0.005 nat/pt, and on PM2.5 it costs 0.018.

On the three single-snapshot graphs the fully selected model is *worse* than the fixed member. We report this rather than omit it. With twenty candidates on these three graphs (43 on the data-rich two) and a validation fold of only ≈10 (LAQN), ≈40 (METR-LA) or ≈65 (PEMS-BAY) nodes, the selection overfits the fold. This is a property of data size, not of the construction: we claim validation-fold selection only on the data-rich datasets, and on the small graphs the appropriate summaries are the fixed member and the post-hoc leaderboard.

M Choosing the auxiliary, and where the blocks come from

Table 15 varies the auxiliary field on PM2.5 and reports, for each, the best validation NLL/pt attained by any family member, together with how often the validation fold selects it. The external annual standard deviation is chosen on 19 of 20 seeds and is about 0.06–0.08 nat/pt better than either signal-derived alternative. This is why the main text reports PM2.5 on the single external auxiliary: `spatial_k8` and `spatial_k12`, the local standard deviation of the training targets over a node's 8 or 12 nearest neighbours, are derived from the signal rather than from an external covariate, and the data prefer the external field.

Table 15: PM2.5 auxiliary ablation. Best validation NLL/pt attained by any family member under each auxiliary field, mean ± SE over 20 seeds, and how often validation selects that auxiliary. The external field wins on all but one seed.

Auxiliary d_i	best validation NLL/pt	selected
`std` (external)	$\mathbf{1.0026 \pm 0.0140}$	$19/20$
`spatial_k8` (signal-derived)	1.0872 ± 0.0178	$1/20$
`spatial_k12` (signal-derived)	1.0631 ± 0.0156	$0/20$

The same question arises for the partition. At fixed K, plain K-means uses coordinates, augmented K-means also uses the standardised auxiliary, and spectral clustering uses the graph Laplacian. These partition rules do not use the target responses. The `spectral-BIC` member additionally selects K using the training marginal likelihood, so its selected partition is not independent of the training responses. No held-out test response enters this selection. Target-derived partitions are possible in principle but must be estimated from training responses alone. The augmented partition instead uses auxiliary observations disjoint from the targets. Without an informative auxiliary, coordinate and spectral partitions remain available, with the same training-only restriction on any response-based selection of K.

N Auxiliary-proxy control

We compare auxiliary proxies while holding the graph, data-generating kernel, observation splits, proxy transform and hyperparameters fixed. Only the auxiliary proxy is changed: oracle, increasingly noisy, rank-only, or randomly shuffled. These are simulated samples, unlike the re-splits of a single real-data snapshot.

Table 16: Auxiliary-proxy control on a grid with known hyperparameters. Every non-stationary row uses the same fixed exponent; no γ is fitted. Entries are mean ± SE across simulated samples. Here σ is the proxy noise standard deviation before clipping.

Proxy	NLL/pt	RMSE
Oracle proxy	0.6835 ± 0.0095	0.4809 ± 0.0048
Noisy proxy ($\sigma = 0.1$)	0.6840 ± 0.0096	0.4809 ± 0.0048
Noisy proxy ($\sigma = 0.3$)	0.6858 ± 0.0098	0.4811 ± 0.0048
Noisy proxy ($\sigma = 0.5$)	0.6878 ± 0.0100	0.4812 ± 0.0048
Noisy proxy ($\sigma = 1.0$)	0.6916 ± 0.0102	0.4817 ± 0.0048
Rank-only proxy	0.6838 ± 0.0096	0.4810 ± 0.0048
Shuffled proxy	0.6964 ± 0.0099	0.4836 ± 0.0047
Stationary baseline	0.6903 ± 0.0089	0.4825 ± 0.0047

A shuffled proxy gives worse mean prediction scores than the stationary baseline in this fixed-exponent control. This illustrates a cost of auxiliary misalignment; it does not establish that every exponent fails, because no exponent is fitted or swept here. The experiment therefore characterises sensitivity to auxiliary misalignment at a fixed exponent.

O Deep-kernel Gaussian process comparison

Table 17 reports the deep-kernel Gaussian process baseline, which uses a multilayer perceptron (MLP) feature map followed by a stationary kernel in feature space. The feature map is evaluated independently at each vertex, unlike a graph-native deep kernel whose feature map propagates through graph edges. This baseline does not evaluate that graph-native construction. The graph-convolutional Deep Ensemble is reported separately in Section E.

Table 17: MLP deep-kernel GP results (`dk`), mean $\pm$ SE across the locked seeds. CA Weather averages the monthly snapshots within each seed. The feature map does not propagate through graph edges.

Dataset	NLL/pt	RMSE	recNLL/pt
METR-LA	2.7214 ± 0.2608	1.1237 ± 0.0296	1.7718 ± 0.0441
PEMS-BAY	3.1423 ± 0.3538	1.2189 ± 0.0431	1.9880 ± 0.0520
LAQN	4.3435 ± 0.2683	1.2220 ± 0.0448	1.6085 ± 0.0419
PM2.5	2.1634 ± 0.0805	0.9692 ± 0.0131	1.6526 ± 0.0211
CA Weather	1.9763 ± 0.0689	0.9714 ± 0.0089	1.6279 ± 0.0281

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and Section 1 state two claims — (i) $\gamma = 1$ is the unique stationary-correlation point of the VCGP family and (ii) VCGPs match or outperform stationary and non-stationary baselines on five real datasets. Both are supported: claim (i) by Proposition 1 in Section 3.1 (full proof in Appendix §A), and claim (ii) by Section 5 together with the per-dataset leaderboards in Appendix §D.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated *Limitations* paragraph in Section 6 covers (i) reliance on a meaningful positive auxiliary d_i, (ii) lack of finite-sample bounds on γ, and (iii) reliance on a chosen partition for the piecewise extension. Cubic GP cost and the empirical scaling behaviour of the kernel factorisation are reported in Appendix §H.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proposition 1 states its three structural hypotheses — G connected, $D^2(G)$ connected, and d non-constant — explicitly in Section 3.1; the full proof is given in Appendix §A. Per-dataset coverage of the hypotheses is reported in Appendix §J.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 5 specifies the protocol (random $70/30$ splits, 20 locked seeds 1729–1748, marginal-log-likelihood fitting). Per-variant edits, partition sources for the piecewise extension, the auxiliary d_i recipe, baseline tuning, and numerical detail are documented in Appendix §B. A single-entry reproducibility kit (`code/reproduce.py` for the five real datasets and `code/reproduce_synthetic.py` for the synthetic experiments), with auto-download of the raw data on first run, is public at `https://github.com/seshkumar/vcgp-graph-gp`.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The full reproducibility kit is public at `https://github.com/seshkumar/vcgp-graph-gp`. It ships two driver scripts — `reproduce.py` (five real datasets, auto-downloads raw data on first run) and `reproduce_synthetic.py` (two synthetic experiments, no external data) — together with a `README.md` listing the exact command for each table in the paper and for the two synthetic figures. All five real datasets are standard public benchmarks.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Splits, optimiser (Adam, 200–300 iterations), warm-start template ($\gamma = 1$), and hyperparameter grids (K, augmentation weight, ridge λ) are documented in Section 5, Section 4, and Appendix §B. The optimiser settings, clipping bracket and K/w grids are in Appendix §F and Appendix §K.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All numbers are reported as mean $\pm$ standard error over 20 locked seeds (1729–1748) with all methods evaluated on the same train/test splits (paired across seeds). Significance is assessed by paired one-sided t-tests against the dataset leader; the test definition and p-values are in Appendix §B.5 and the per-dataset leaderboards of Appendix §D.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on a single CPU; no GPU is required. Appendix §H ("Scalability of the kernel factorisation") reports the per-dataset factorisation cost and its scaling with n, together with the end-to-end fitting cost on the largest graph.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only standard public sensor and weather datasets and introduces no new data on human subjects. We have reviewed the NeurIPS Code of Ethics and the work conforms.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The contribution is a methodological refinement of graph Gaussian processes; we do not foresee a direct path to specific positive or negative societal applications beyond the existing application domains (traffic and air-quality monitoring) where stationary kernels are already standard.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release high-risk models, generative systems, or scraped datasets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All five real datasets used in the paper (METR-LA, PEMS-BAY, LAQN, PM2.5, CA Weather) are standard publicly available benchmarks; the original sources are cited in the **Datasets** paragraph of Section 5, and the auxiliary field derived from each is described in Appendix §B.4. No new datasets are introduced.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The new asset is the reproducibility kit, public at `https://github.com/seshkumar/vcgp-graph-gp`. It ships `reproduce.py`, `reproduce_synthetic.py` and a `README.md` documenting the exact command for each table in the paper and for the two synthetic figures, together with a licence.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work does not involve crowdsourcing or research with human subjects.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work does not involve research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the methodology, data construction, or experiments.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.