LONG-MEMORY GAUSSIAN PROCESSES FOR STATE-SPACE INFERENCE AND REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The Matérn covariance kernel is a popular choice for Gaussian process (GP) based modeling of temporal dependencies in time-series data. For half-integer smoothness, the Matérn kernel admits a finite-dimensional state-space representation, reducing GP inference from cubic to linear complexity in the time series length. This has made Matérn attractive both for direct sequential modeling and for representing structured latent priors in variational autoencoders. However, Matérn kernels decay exponentially with lag and therefore cannot represent persistent, polynomially decaying long-range dependencies. To address this, we develop the GAMMA covariance kernel family, obtained by gamma-mixing of half-integer Matérn components. The resulting kernels separate local regularity from long-range memory, allowing near-origin smoothness and asymptotic dependence to be controlled independently. For scalable inference, we approximate the continuous mixture using generalized Gauss–Laguerre quadrature, yielding a finite weighted sum of Markovian components and linear inference complexity in time series length. Building on this construction, we further propose GGPVAE, which places GAMMA priors over latent trajectories. We also theoretically prove the convergence of the proposed approximations. Our results show that GAMMA captures long-range dependence, while GAMMA–GL preserves useful persistence over finite horizons and improves long-gap prediction in GGPVAE with scalable inference.

1 INTRODUCTION

Sequential and time-series data arise across a wide range of scientific and machine-learning applications, including probabilistic forecasting, speech, healthcare, and other temporally structured systems (Chung et al., 2015; Rangapuram et al., 2018; Rubanova et al., 2019). Modeling such data requires capturing temporal structure while representing uncertainty about unobserved or future observations (Rangapuram et al., 2018; Krishnan et al., 2017). In latent-variable models for sequential data, this further requires the latent representations themselves to evolve over time rather than being modeled as static objects (Chung et al., 2015; Krishnan et al., 2017; Rubanova et al., 2019). Gaussian processes (GPs) provide a natural framework for this purpose, since they allow us to learn the latent representations from data while only specifying high-level properties of the latent trajectories such as smoothness, characteristic time scales, and temporal dependence via the covariance functions (Rasmussen & Williams, 2006). This flexibility, together with the capability of treatment of spatial and temporal uncertainty, has motivated the use of GP priors in latent-variable and variational autoencoder models (Casale et al., 2018; Fortuin et al., 2020).

Among the widely used GP covariance families, the Matérn class is particularly attractive for temporal modeling because it combines interpretable control of local smoothness with favorable computational structure (Matérn, 1986). For half-integer smoothness parameters, Matérn GPs admit exact finite-dimensional state-space representations, allowing temporal inference to be performed with Kalman filtering and smoothing rather than dense GP operations (Hartikainen & Särkkä, 2010). For fixed state dimension, this reduces the computational complexity from $\mathcal{O}(T^3)$ to $\mathcal{O}(T)$ in the sequence length. This connection has also been exploited in deep sequential models, for example, Zhu et al. (2023) introduced the Markovian Gaussian Process Variational Autoencoder (MGPVAE), which uses state-space GP priors to enable scalable inference over latent temporal trajectories.

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

However, the same Matérn structure that enables efficient Markovian inference also causes weakness in modeling power: its correlations decay exponentially with temporal lag, so sufficiently distant observations become nearly independent (Porcu et al., 2024). This behavior is appropriate when dependence is predominantly local, but it can be inadequate for processes with heterogeneous persistence, where fast transitory effects coexist with much slower adjustments (Schwartz & Smith, 2000). In such settings, a single Matérn time scale fails to represent the full range of temporal dependencies. Therefore, there is a need for covariance models that retain efficient state-space structure while allowing substantially more persistent, long-range behavior.

Several covariance constructions have been proposed to move beyond the short-range dependence of Matérn kernels. Polynomially decaying families, including rational-quadratic and generalized Cauchy covariances, can capture substantially more persistent dependence (Rasmussen & Williams, 2006; Bevilacqua & Faouzi, 2019). For the rational-quadratic kernel, Solin & Särkkä (2014a) developed a scalable state-space approximation by expressing it as a scale mixture of squared-exponential kernels, discretizing the mixing integral using generalized Gauss–Laguerre quadrature, and then approximating each squared-exponential component by a finite-dimensional state-space model. Thus, state-space inference is possible, but requires two approximation layers: quadrature of the rational-quadratic mixture and a separate state-space approximation of each squared-exponential component. This increases the state dimension and can make the long covariance tail harder to capture accurately at moderate approximation orders (Solin & Särkkä, 2014a). More recently, Ma et al. (2023) introduced the confluent hypergeometric (CH) covariance, which provides separate control over local smoothness and polynomial tail decay, but involves confluent hypergeometric function evaluation and likewise does not admit a direct finite-dimensional Markov representation. These limitations motivate covariance constructions that combine persistent, polynomially decaying dependence with a more direct route to finite-dimensional state-space approximation.

Longer temporal dependencies have also been modeled using multiresolution GP and GP state-space architectures (Fox & Dunson, 2012; Longi et al., 2022), as well as explicit long-memory mechanisms that have recently been applied to sparse GP variational autoencoders (Chen et al., 2025). These approaches capture richer temporal structure through multiresolution representations or additional memory mechanisms, but do not retain the same simple finite-dimensional linear-Gaussian Markov structure that enables direct Kalman-based inference for half-integer Matérn GPs. As a consequence, they generally introduce additional architectural and computational overhead relative to inference with a single fixed-dimensional Markov state-space representation.

In this work, we introduce the GAmma-Mixed MAtérn (GAMMA) covariance class, where the latent process is constructed as a continuous mixture of half-integer Matérn components over their rate parameter. This construction admits an infinite-dimensional state-space interpretation, which we approximate using finite-dimensional Markov surrogates derived through generalized Gauss–Laguerre quadrature. Building on this representation, we propose the Gamma Gaussian Process Variational Autoencoder (GGPVAE), which places GAMMA priors over latent trajectories for high-dimensional sequential data with persistent temporal dependencies. For scalable training and inference, we replace the exact GAMMA prior with its quadrature-based state-space surrogate, allowing latent GP dynamics to be handled by Kalman filtering and smoothing, analogously to the Markovian GP framework of Zhu et al. (2023). This allows GGPVAE to approximate the persistent temporal dependence of the GAMMA prior using efficient finite-dimensional state-space inference.

Our main contributions are as follows:

- We develop GAMMA, a Gamma-mixed half-integer Matérn family that allows to independently control local regularity and long-range dependence.
- We derive a moment-preserving Gauss–Laguerre Markov surrogate for scalable inference that approximates GAMMA over finite lags, cf. Theorem 2.
- We prove functional convergence (Theorem 1) of the approximating Markov process towards the original process with GAMMA covariance.
- Building on this construction, we propose GGPVAE, a sequential latent-variable model that equips high-dimensional representation learning with GAMMA-based temporal priors.
- Across forecasting and representation-learning benchmarks, we observe improved performance, with the clearest gains under persistent temporal dependence.

The rest of the article is organized as follows. Section 2 contains background while we introduce GAMMA covariance in Section 3. The Markov approximation is studied in Section 4 and application to GGPVAE in Section 5. Finally Experiments and results are studied in Section 6.

2 BACKGROUND

Let $t_1, \dots, t_T$ denote the observation times and let $Y_i := Y(t_i) \in \mathbb{R}$ denote the corresponding observations. Given $\mathbf{Y}_{1:T} = (Y_1, \dots, Y_T)^\top$, the goal is to characterize the predictive distribution $p(Y_* \mid \mathbf{Y}_{1:T})$ at a new time t_*.

Gaussian Processes. One way to model this problem is through Gaussian process regression, where $Y_i = f(t_i) + \varepsilon_i$, with ε_i Gaussian noise of zero mean and variance σ^2_{noise}, and the latent function is assigned the prior $f \sim \mathcal{GP}(0, k)$. The covariance kernel $k(t, t')$ determines the temporal dependence between observations. Exact GP inference requires operations on a dense $T \times T$ covariance matrix and therefore scales as $\mathcal{O}(T^3)$ (Rasmussen & Williams, 2006).

State-Space Representations of Markovian GP

A GP admitting a finite-dimensional Markov representation can be expressed through a linear stochastic differential equation (SDE) (Särkkä & Solin, 2019),

$$d\mathbf{s}(t) = \mathbf{F}\mathbf{s}(t)\, dt + \mathbf{L}\, d\boldsymbol{\beta}(t), \qquad f(t) = \mathbf{H}\mathbf{s}(t),$$

where $\mathbf{s}(t) \in \mathbb{R}^d$ is the Markov state, $\mathbf{F} \in \mathbb{R}^{d\times d}$ governs its temporal dynamics, $\mathbf{L} \in \mathbb{R}^{d\times e}$ maps the e-dimensional driving Brownian noise $d\boldsymbol{\beta}(t) \in \mathbb{R}^e$ into the state, and $\mathbf{H} \in \mathbb{R}^{1\times d}$ maps the state to the scalar GP value $f(t)$. The covariance structure of the GP is therefore encoded through the state-space parameters rather than through a dense covariance matrix. At observation times $t_1, \dots, t_T$, with $\Delta_i = t_{i+1} - t_i$, the SDE induces the discrete model

$$\mathbf{s}_{i+1} = \mathbf{A}_i \mathbf{s}_i + \mathbf{q}_i, \qquad \mathbf{q}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i), \qquad Y_i = \mathbf{H}\mathbf{s}_i + \varepsilon_i,$$

where $\mathbf{A}_i = \exp(\mathbf{F}\Delta_i)$ propagates the latent state between consecutive observation times and $\mathbf{Q}_i$ is the corresponding process-noise covariance. For a stationary model, $\mathbf{Q}_i = \mathbf{P}_\infty - \mathbf{A}_i \mathbf{P}_\infty \mathbf{A}_i^\top$, where $\mathbf{P}_\infty$ denotes the stationary state covariance. With a conjugate Gaussian likelihood, the posterior over $\mathbf{s}_{1:T}$ can then be computed efficiently using recursive Kalman filtering and smoothing, reducing inference to operations that scale linearly with T for fixed state dimension (Hartikainen & Särkkä, 2010; Särkkä et al., 2013; Särkkä & Solin, 2019).

Variational Autoencoders (VAEs) and Markovian GP Priors

So far, we have considered the predictive distribution $p(Y_* \mid \mathbf{Y}_{1:T})$ using Gaussian process models, where temporal dependence is defined directly through a covariance function. For high-dimensional observations, we instead introduce latent representations $\mathbf{Z}_t \in \mathbb{R}^L$ and use a VAE. The predictive objective remains $p(\mathbf{Y}_* \mid \mathbf{Y}_{1:T})$, but temporal dependence is now represented through the latent trajectory. For a sequence $\mathbf{Y}_{1:T}$, the generative model factorizes as $p_\theta(\mathbf{Y}_{1:T}, \mathbf{Z}_{1:T}) = p(\mathbf{Z}_{1:T}) \prod_{t=1}^T p_\theta(\mathbf{Y}_t \mid \mathbf{Z}_t)$, where the decoder defines the observation model and $p(\mathbf{Z}_{1:T})$ specifies the latent temporal structure. Training maximizes the evidence lower bound (Ranganath et al., 2014),

$$\mathcal{L} = \sum_{t=1}^T \mathbb{E}_{q_\phi(\mathbf{Z}_{1:T}\mid\mathbf{Y}_{1:T})}[\log p_\theta(\mathbf{Y}_t \mid \mathbf{Z}_t)] - \mathrm{KL}(q_\phi(\mathbf{Z}_{1:T} \mid \mathbf{Y}_{1:T}) \,\|\, p(\mathbf{Z}_{1:T}))\,.$$

For sequential data, GPVAEs impose temporal dependence by assigning a GP prior to each latent coordinate (Casale et al., 2018; Fortuin et al., 2020; Jazbec et al., 2021). In particular, $Z^{(\ell)}(\cdot) \sim \mathcal{GP}(0, k_\ell)$ implies $\mathbf{z}^{(\ell)} \sim \mathcal{N}(\mathbf{0}, \mathbf{K}_\ell)$ at the observation times, with $[\mathbf{K}_\ell]_{ij} = k_\ell(t_i, t_j)$ and hence the kernel therefore controls the temporal properties of the learned representation.

MGPVAE (Zhu et al., 2023) applies the Markovian state-space representation introduced in Section 2 to each latent GP coordinate. At the observation times, latent channel ℓ follows

$$\mathbf{s}_{t+1}^{(\ell)} = \mathbf{A}_t^{(\ell)} \mathbf{s}_t^{(\ell)} + \mathbf{q}_t^{(\ell)}, \qquad \mathbf{q}_t^{(\ell)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_t^{(\ell)}), \qquad Z_t^{(\ell)} = \mathbf{H}^{(\ell)} \mathbf{s}_t^{(\ell)}.$$

The latent coordinates form $\mathbf{Z}_t = (Z_t^{(1)}, \ldots, Z_t^{(L)})^\top$, which is passed through the nonlinear VAE decoder. For fixed state dimension, this Markovian formulation avoids dense GP operations and yields complexity linear in the sequence length (Zhu et al., 2023). Although the GP prior is linear Gaussian, the nonlinear decoder makes the observation model non-conjugate, so the exact posterior over the latent states cannot be recovered by standard Kalman recursions. MGPVAE addresses this by introducing encoder-generated Gaussian sites. At each time t, the encoder produces a pseudo-observation and variance for every latent coordinate, $\{\widetilde{Y}_t^{(\ell)}, \widetilde{V}_t^{(\ell)}\}_{\ell=1}^L = \varphi_\phi(\mathbf{Y}_t)$, and defines

$$q_\phi(\mathbf{s}_{1:T}) \propto p(\mathbf{s}_{1:T}) \prod_{\ell=1}^{L} \prod_{t=1}^{T} \mathcal{N}\Big(\widetilde{Y}_t^{(\ell)} \mid \mathbf{H}^{(\ell)} \mathbf{s}_t^{(\ell)}, \widetilde{V}_t^{(\ell)}\Big). \tag{1}$$

The Gaussian sites restore conjugacy with the state-space prior, so the filtering and smoothing marginals of $q_\phi(\mathbf{s}_{1:T})$ can be computed analytically using Kalman filtering and Rauch–Tung–Striebel smoothing. With this Gaussian-site construction, the ELBO can be rearranged as

$$\begin{aligned} \mathcal{L} = \log \mathbb{E}_{p(\mathbf{s}_{1:T})} \left[\prod_{t=1}^{T} \prod_{\ell=1}^{L} \mathcal{N}\Big(\widetilde{Y}_t^{(\ell)} \mid \mathbf{H}^{(\ell)} \mathbf{s}_t^{(\ell)}, \widetilde{V}_t^{(\ell)}\Big) \right] \\ + \sum_{t=1}^{T} \mathbb{E}_{q_\phi(\mathbf{s}_t)} \left[\log p_\theta(\mathbf{Y}_t \mid \mathbf{Z}_t) - \sum_{\ell=1}^{L} \log \mathcal{N}\Big(\widetilde{Y}_t^{(\ell)} \mid \mathbf{H}^{(\ell)} \mathbf{s}_t^{(\ell)}, \widetilde{V}_t^{(\ell)}\Big) \right]. \end{aligned} \tag{2}$$

The resulting linear Gaussian state-space structure makes the Gaussian terms in the ELBO tractable through Kalman filtering and smoothing, while the nonlinear reconstruction term is estimated using Monte Carlo samples from the variational posterior (Zhu et al., 2023).

3 Gamma-Mixed Matérn Covariance

In this section, we derive the proposed GAMMA covariance by mixing half-integer Matérn processes over their rate parameter. The Matérn covariance class (Matérn, 1986) is

$$k_\eta(h \mid \lambda, \sigma^2) := \sigma^2 \, \frac{2^{1-\eta}}{\Gamma(\eta)} \, (\lambda h)^\eta \, K_\eta(\lambda h), \qquad \eta > 0,\ \lambda > 0,\ \sigma^2 > 0,$$

where $K_\eta(\cdot)$ is the modified Bessel function of the second kind. The parameter η controls the mean-square differentiability of the associated Gaussian process, while λ acts as a decay-rate parameter. For half-integer smoothness values, $\eta = p + \frac{1}{2}$, the Matérn family admits a direct stochastic-process interpretation: the corresponding covariance arises as the stationary covariance of a $(p+1)$st-order Ornstein–Uhlenbeck process. A higher-order Ornstein-Uhlenbeck process of order $p+1$ (where $p \in \mathbb{N}_0$) can be defined by the $p+1$-th order stochastic differential equation

$$\left(\frac{d}{dt} + \lambda\right)^{p+1} Y_\lambda(t) = \sigma \dot{W}_\lambda(t).$$

It is well-known that these processes have stationary covariances given by Matérn kernels (Rasmussen & Williams, 2006). Matérn kernel for a half-integer values of $\eta = p + 1/2$ can be written in the form

$$k_p(h \mid \lambda, \sigma^2) = \frac{\sigma^2 \Gamma(p+1)}{\Gamma(2p+1)} e^{-\lambda|h|} \sum_{i=0}^{p} \frac{(p+i)!}{i!(p-i)!} (2\lambda|h|)^{p-i},$$

where $\Gamma(.)$ is the Gamma function. For notational simplicity, we simply write

$$k_p(h \mid \lambda, \sigma^2) := \sigma^2 \, e^{-\lambda h} \, P_p(\lambda h), \qquad h \geq 0,\ \lambda > 0,$$

where $P_p(u) = \sum_{k=0}^{p} a_{p,k} \, u^k$ is a degree-p polynomial with $P_p(0) = a_{p,0} = 1$.

Gamma-mixed OU processes have been used for long-range dependence (Iglói & Terdik, 1999), while supOU provides a broader framework for flexible-memory stationary processes (Barndorff-Nielsen, 2001). Here, we extend this construction to higher-order OU processes corresponding to half-integer Matérn kernels.

$$X(t) = \int_0^\infty Y_\lambda(t) \nu(d\lambda),$$

where each Y_λ for different λ are independent and $\nu(d\lambda)$ is the speed distribution. A formal construction is given in Appendix D. The covariance function of X, denoted by $R_p(h;\nu,\sigma^2)$, is given by the integral of the corresponding Matérn kernels as

$$R_p(h;\nu,\sigma^2) := \int_0^\infty k_p(h \mid \lambda,\sigma^2)\,\nu(d\lambda) = \sigma^2 \sum_{k=0}^{p} a_{p,k} h^k \int_0^\infty e^{-\lambda h}\lambda^k \nu(d\lambda).$$

This framework allows for the simultaneous modeling of local fractal dimensions (via the order p) and long-range dependence via the tail of the measure ν. In the sequel, we choose the mixing measure ν as the Gamma law given by

$$\nu_{\alpha,\beta}(d\lambda) := \frac{\beta^\alpha}{\Gamma(\alpha)}\lambda^{\alpha-1}e^{-\beta\lambda}\,d\lambda, \qquad \alpha>0,\ \beta>0.$$

This yields a covariance, for a fixed p, given by

$$R_{p,\sigma^2,\alpha,\beta}(h) := \int_0^\infty k_p(h \mid \lambda,\sigma^2)\,\nu_{\alpha,\beta}(d\lambda) = \sigma^2 \sum_{k=0}^{p} a_{p,k} h^k \,\frac{\beta^\alpha}{\Gamma(\alpha)}\,\frac{\Gamma(\alpha+k)}{(\beta+h)^{\alpha+k}}. \tag{3}$$

Thus, for half-integer Matérns, the mixed covariance is explicit in terms of Gamma functions and rational powers. It follows that the Matérn order p governs the local behavior near the origin and hence controls the path properties and local behaviour of the process, while memory is controlled by the parameter α describing the tail of the mixing measure ν. In particular, X has long memory if and only if $0<\alpha\le 1$. These facts and other basic properties are discussed in Appendix E.

4 Gauss–Laguerre Markov Approximation of GAMMA

The general perspective of replacing a continuous covariance representation by a structured finite-dimensional state-space approximation is classical in temporal Gaussian-process modeling; see, for example, (Hartikainen & Särkkä, 2010; Särkkä & Solin, 2019). Likewise, quadrature-based finite-mixture approximations have already appeared in nearby kernel settings, including rational quadratic and Matérn scale-mixture approximations (Solin & Särkkä, 2014a; Tronarp et al., 2018). By building on these approaches, we consider a kernel approximation based on Gauss–Laguerre quadrature (see Appendix B for details) given by

$$R^{(J,GL)}_{p,\sigma^2,\alpha,\beta}(h) := \sum_{r=1}^{J} w_r\, k_p(h \mid \lambda_r,\sigma^2), \tag{4}$$

with $w_r = w_r^{GL}/\Gamma(\alpha)$ and $\lambda_r = x_r/\beta$, where $\{x_r, w_r^{\mathrm{GL}}\}_{r=1}^J$ is the J-point generalized Gauss–Laguerre quadrature rule for the weights of $x^{\alpha-1}e^{-x}$. We refer to this finite Gauss–Laguerre approximation of the GAMMA covariance as the GAMMA–GL surrogate and its state-space representation is detailed in Appendix C. This is again a valid covariance as a nonnegative finite mixture of positive definite kernels, and corresponds to the process

$$X_t^{(J,GL)} = \sum_{r=1}^{J} \sqrt{w_r}\, Y_{\lambda_r}(t).$$

We note that the present approach differs from rational quadratic and Mat'ern scale mixtures (Solin & Särkkä, 2014a; Tronarp et al., 2018). For fixed J, GAMMA–GL ultimately has exponential decay and approximates GAMMA's polynomial dependence over finite lags, while preserving low-order moments exactly. We next establish functional convergence.

Theorem 1 (functional convergence in distribution)**.** *For any fixed $T>0$, the process $X^{(J,GL)}$ converges weakly to X as $J\to\infty$ in the space of continuous functions $C([0,T])$.*

The proof is presented in Appendix F. The next theorem shows that GAMMA-GL approximation preserves local regularity properties up to order $2J$.

Theorem 2 (Moment preservation of the GAMMA–GL surrogate)**.** *Fix $p \in \mathbb{N}_0$, and let $R_{p,\sigma^2,\alpha,\beta}(h)$ and $R^{(J,GL)}_{p,\sigma^2,\alpha,\beta}(h)$ be defined by equation 3 and equation 4, respectively. Then there exist coefficients $\{b_{p,n}\}_{n\geq 0}$, depending only on P_p, such that*

$$R_{p,\sigma^2,\alpha,\beta}(h) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n}\, \mathbb{E}[\Lambda^n]\, |h|^n, \qquad R^{(J,GL)}_{p,\sigma^2,\alpha,\beta}(h) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n} \Big(\sum_{r=1}^{J} w_r \lambda_r^n\Big) |h|^n.$$

In particular, $R_{p,\sigma^2,\alpha,\beta}(h) - R^{(J,GL)}_{p,\sigma^2,\alpha,\beta}(h) = O(|h|^{2J}), \qquad h \to 0.$

The proof is given in Appendix G. The significance of Theorem 2 is that the GAMMA–GL is not merely a generic finite-mixture approximation. Because generalized Gauss–Laguerre quadrature is exact for polynomials up to degree $2J-1$, the surrogate matches the first $2J$ moments of the Gamma mixing law. Consequently, it preserves the local properties of covariance through powers $|h|^0, \ldots, |h|^{2J-1}$. This in turn implies that the local path properties are preserved as well provided that $J \geq p+1$, and the error is actually smooth in L^2 sense up to order $2J$.

5 Gamma Gaussian Process Variational Autoencoder

The finite-dimensional Markov approximation in Section 4 enables the GAMMA covariance to be used in sequential latent-variable models. Building on MGPVAE (Section 2), we introduce the Gamma Gaussian Process Variational Autoencoder (GGPVAE), retaining the same encoder, decoder, Gaussian-site construction, and variational objective, while replacing the temporal prior with a GAMMA prior approximated by its GAMMA–GL surrogate.

GAMMA latent prior and Markov realization:

Since GAMMA does not generally admit an exact finite-dimensional Markov representation, For each latent coordinate $\ell = 1, \ldots, L$, we use the J-point generalized Gauss–Laguerre approximation from Section 4 to model the temporal prior. The resulting finite approximation to the latent covariance is then

$$R^{(J,\mathrm{GL})}_{p,\sigma_\ell^2,\alpha_\ell,\beta_\ell}(h) = \sum_{r=1}^{J} w_{\ell r}\, k_p(h \mid \lambda_{\ell r}, \sigma_\ell^2)$$

a weighted sum of J half-integer Matérn components, each admitting a $(p+1)$-dimensional Markov representation. Let $\mathbf{s}^{(\ell,r)}(t) \in \mathbb{R}^{p+1}$ denote the state of component r, with dynamics

$$d\mathbf{s}^{(\ell,r)}(t) = \mathbf{F}^{(\ell,r)}\mathbf{s}^{(\ell,r)}(t)\, dt + \mathbf{L}^{(\ell,r)} d\mathbf{B}^{(\ell,r)}(t), \qquad r = 1, \ldots, J.$$

where $\mathbf{F}^{(\ell,r)}$ corresponds to a half-integer Matérn process with rate $\lambda_{\ell r}$. We collect the component states into $\mathbf{s}^{(\ell,J)}(t) = [(\mathbf{s}^{(\ell,1)}(t))^\top, \ldots, (\mathbf{s}^{(\ell,J)}(t))^\top]^\top \in \mathbb{R}^{J(p+1)}$. Because the components evolve independently, the augmented feedback matrix is block diagonal, $\mathbf{F}^{(\ell,J)} = \mathrm{blockdiag}(\mathbf{F}^{(\ell,1)}, \ldots, \mathbf{F}^{(\ell,J)}) \in \mathbb{R}^{J(p+1)\times J(p+1)}$. The scalar latent coordinate is obtained by

$$Z^{(\ell)}(t) = \sum_{r=1}^{J} \sqrt{w_{\ell r}}\, \mathbf{H}^{(\ell,r)}\mathbf{s}^{(\ell,r)}(t) = \mathbf{H}^{(\ell,J)}_{\mathrm{aug}}\mathbf{s}^{(\ell,J)}(t)$$

where $\mathbf{H}^{(\ell,J)}_{\mathrm{aug}} = [\sqrt{w_{\ell 1}}\mathbf{H}^{(\ell,1)} \cdots \sqrt{w_{\ell J}}\mathbf{H}^{(\ell,J)}] \in \mathbb{R}^{1\times J(p+1)}$. The augmented state therefore gives an exact Markov realization of the GAMMA–GL with dimension $d_J = J(p+1)$ per latent coordinate, and is discretized at observation times as in Section 2. Writing $\Delta_i = t_{i+1} - t_i$,

$$\mathbf{A}_i^{(\ell,J)} = \exp\Big(\mathbf{F}^{(\ell,J)}\Delta_i\Big), \qquad \mathbf{Q}_i^{(\ell,J)} = \mathbf{P}_\infty^{(\ell,J)} - \mathbf{A}_i^{(\ell,J)}\mathbf{P}_\infty^{(\ell,J)}(\mathbf{A}_i^{(\ell,J)})^\top.$$

and $\mathbf{s}_{i+1}^{(\ell,J)} = \mathbf{A}_i^{(\ell,J)}\mathbf{s}_i^{(\ell,J)} + \mathbf{q}_i^{(\ell,J)}$, with $\mathbf{q}_i^{(\ell,J)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i^{(\ell,J)})$.

Variational inference and complexity:

Inference follows the Gaussian-site variational framework of MGPVAE (Zhu et al., 2023). The encoder parameterizes Gaussian pseudo-observation sites as in equation 1, restoring conjugacy with

the augmented GAMMA–GL state-space prior and enabling Kalman filtering and Rauch–Tung–Striebel smoothing. Training then optimizes the corresponding variational objective, analogous to equation 2, jointly over the encoder, decoder, and GAMMA hyperparameters. For fixed J, p, and L, complexity remains linear in sequence length, $\mathcal{O}(LT[J(p+1)]^3)$, while J controls the accuracy–state-dimension trade-off. Thus, for fixed J, p, and L, GGPVAE retains linear complexity in sequence length while providing a finite-dimensional approximation to the persistent dependence of the underlying GAMMA prior.

6 EXPERIMENTS

We evaluate the proposed GAMMA and GAMMA–GL covariances at two complementary levels. First, we study it directly in GP regression and forecasting, testing whether its long-memory structure improves over Matérn covariances and whether the GAMMA–GL preserves the kernel behavior. This isolates the effect of the covariance itself. Second, we use GAMMA–GL as the latent temporal prior in a variational model to assess whether its persistent dependence remains beneficial in learned representations, particularly for prediction beyond the observed history. Our experiment setting is further detailed in Appendix H.

6.1 KERNEL EVALUATION

We compare the exact GAMMA kernel and its approximation GAMMA–GL against matched Matérn baselines to assess both predictive performance and the fidelity of the scalable state-space approximation. Evaluation uses Root Mean Squared Error(RMSE), negative log predictive density (NLPD), 95% predictive-interval coverage (Coverage95), and predictive-interval width (Width95). Results on additional datasets are provided in Appendix I.

Synthetic data:

We generate synthetic long-memory time series from the Gamma-mixture model using a GAMMA–GL approximation, producing trajectories with multiple reversion time scales and slowly decaying temporal dependence. Across multiple seeds and $p \in \{0, 1\}$, we compare Matérn with GAMMA and GAMMA–GL for $J = 12$ under several training fractions. Table 1 shows that the Gamma-based models generally outperform the Matérn baselines, with the strongest gains for $p = 1$: they achieve lower RMSE and NLPD in most settings and predictive interval coverage closer to the nominal 95% level in all cases, while the narrower Matérn intervals tend to under-cover, indicating overconfident uncertainty estimates. In addition, the close agreement between GAMMA–GL and GAMMA supports the effectiveness of the finite-dimensional Markov approximation. More results on the closeness between GAMMA–GL and GAMMA along with full simulation details and additional parameter-recovery results as a function of J are provided in Appendix H.1.1.

Table 1: Comparison of Matérn and GAMMA and GAMMA–GL($J = 12$) for $p \in \{0, 1\}$.

p	Train frac.	Model	RMSE	NLPD	Coverage95	Width95
0	50%	Matérn	**0.983 ± 0.319**	1.899 ± 1.475	0.784 ± 0.220	**2.590 ± 0.581**
		GAMMA	1.015 ± 0.353	**1.724 ± 1.090**	**0.822 ± 0.219**	2.950 ± 0.563
		GAMMA–GL	1.021 ± 0.356	**1.730 ± 1.086**	**0.821 ± 0.220**	2.964 ± 0.565
	70%	Matérn	1.012 ± 0.319	1.593 ± 0.513	0.803 ± 0.142	**2.674 ± 0.416**
		GAMMA	**0.940 ± 0.297**	**1.362 ± 0.386**	**0.921 ± 0.109**	3.000 ± 0.394
		GAMMA–GL	**0.935 ± 0.298**	**1.352 ± 0.382**	**0.923 ± 0.109**	3.018 ± 0.400
1	50%	Matérn	0.932 ± 0.431	2.203 ± 1.600	0.684 ± 0.235	**1.977 ± 0.518**
		GAMMA	**0.830 ± 0.375**	**1.335 ± 0.688**	**0.873 ± 0.152**	2.572 ± 0.581
		GAMMA–GL	**0.834 ± 0.370**	**1.342 ± 0.680**	**0.868 ± 0.155**	2.570 ± 0.579
	70%	Matérn	0.840 ± 0.465	1.830 ± 1.640	0.759 ± 0.275	**2.075 ± 0.551**
		GAMMA	**0.735 ± 0.380**	**1.233 ± 0.762**	**0.842 ± 0.204**	2.385 ± 0.476
		GAMMA–GL	**0.740 ± 0.375**	**1.234 ± 0.740**	**0.836 ± 0.203**	2.372 ± 0.467

Real-world time series:

We evaluate Matérn and GAMMA–GL on three real-world settings with distinct temporal dependence. The UCI Energy data provide a densely sampled indoor-temperature series for long-horizon forecasting of persistent sensor dynamics. The U.S. dollar-index series, DTWEXBGS, exhibits persistent cyclical fluctuations that are qualitatively consistent with mean-reverting dynamics over

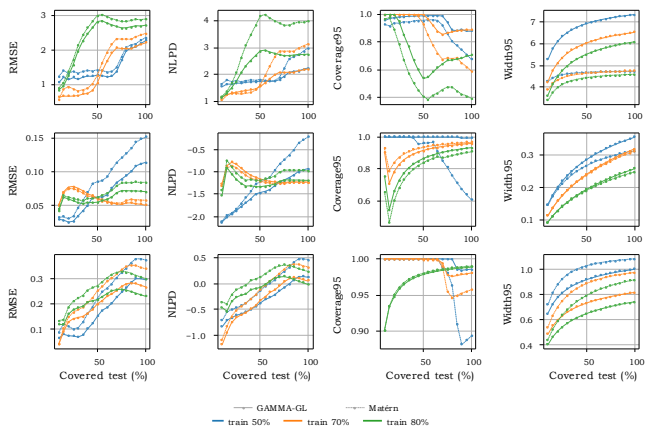


Figure 1: GAMMA–GL vs. Matérn, both for $p = 0$, on UCI, DTWEXBGS, and Coke–Pepsi (top to bottom).

certain horizons, motivating comparison with classical mean-reverting models such as Vasicek (Vasicek, 1977). The Coke–Pepsi pair similarly provides a setting in which a persistent relationship between economically related assets can coexist with temporary spread deviations that revert toward equilibrium, as studied in cointegration and pairs-trading frameworks (Engle & Granger, 1987; Gatev et al., 2006; Elliott et al., 2005). For further information on the datasets and the test settings, see Appendix H.1.2. These datasets therefore assess persistent, multi-scale dependence beyond a single Matérn time scale. We evaluated both $p \in \{0, 1\}$ and report $p = 0$, which performed better for both models, consistent with their short-scale fluctuations. Figure 1 shows that GAMMA–GL generally provides more robust long-horizon predictions than the Matérn model. Farther into the held-out region, it often achieves lower RMSE and NLPD and better Coverage95, particularly for Coke–Pepsi, while the Matérn intervals increasingly under-cover, indicating comparatively overconfident uncertainty estimates.

6.2 Representation Learning

We evaluate GAMMA–GL as a temporal prior for variational representation learning by comparing GGPVAE with MGPVAE on long-memory Rotating MNIST and ERA5. Following Zhu et al. (2023), we focus on these models and complement RMSE and NLPD on Rotating MNIST with content-oriented metrics to test if GAMMA–GL remains beneficial in learned latent spaces.

Long-memory Rotating MNIST: We construct a long-memory Rotating-MNIST benchmark in which a GAMMA process drives the angular velocity, creating controlled long-range temporal dependence. Both GGPVAE and MGPVAE are trained on disjoint splits with 60% randomly withheld frames, and are tested under both contiguous and scattered missingness. We report held-out RMSE, NLPD, rotation-invariant RMSE (RI-RMSE), and content digit accuracy (ContentAcc). All architectural, optimization, masking, validation, and initialization settings are matched so that the intended difference is only the temporal prior. Full details are given in Appendix H.2.1. Figure 2 shows that GGPVAE is most beneficial for contiguous gaps: with 40% scattered missingness the two models are visually similar, whereas for a 40% contiguous gap GGPVAE better preserves digit shape and rotational coherence. Quantitatively, Figure 3 shows that this advantage grows with gap length, becoming consistent from about 30% missingness and particularly pronounced in RI-RMSE and ContentAcc. Scattered-missing results, however, remain broadly comparable because nearby observations are still available, check the Appendix H.2.1.

Spatiotemporal Climate Data: We evaluate GGPVAE on multivariate ERA5 climate observations (Gorelick et al., 2017) to assess temporal forecasting. We use a separable spatiotemporal covariance, $k((r, t), (r', t')) = k_r(r, r')k_t(t, t')$, split the data chronologically into 101 training days, 10 validation days, and 30 test days, using validation NLPD for model selection. Figure 4 reports performance over the held-out 30-day forecasting horizon. GGPVAE achieves lower RMSE throughout the forecasting period and lower NLPD over most horizons. Averaged across the test period, it improves



Figure 2: Posterior-mean reconstructions for (right)scattered missing frames, (Left) contiguous missing block. Red boxes in the input row denote frames withheld from inference.

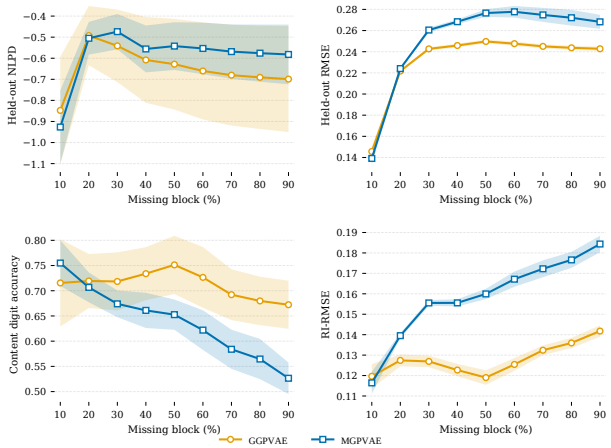


Figure 3: GGPVAE and MGPVAE under contiguous missing blocks comparison across multiple simulations. Held-out here refers to the computation of metrics on the withheld frames.

RMSE by 14.1% and NLPD by 8.9% relative to MGPVAE. At the longest horizons, however, the NLPD gap narrows and slightly favors MGPVAE, while GGPVAE retains lower point error. Thus, GAMMA provides a persistent advantage in point prediction, but its probabilistic advantage diminishes at the longest horizons as forecast uncertainty increasingly dominates the information retained from the observed history.

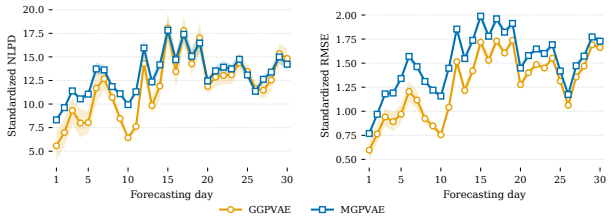


Figure 4: Forecasting performance on ERA5. Averaged over the full horizon, MGPVAE and GGPVAE obtain NLPD values of 13.199 and 12.018, and RMSE values of 1.532 and 1.316, respectively.

7 Conclusion

In this article we have introduced GAMMA covariance for modeling long-range memory and studied its finite-dimensional Markov approximation based on generalized Gauss–Laguerre approximation. Our theoretical results show that this surrogate preserves relevant low-order moments and local covariance structure of the exact GAMMA kernel, providing a principled finite-state approximation. Across the forecasting and GGPVAE experiments, we observe that the Gamma construction is suitable for modeling long memory. The exact GAMMA covariance provides a model of polynomially decaying long-range dependence, while GAMMA–GL provides a finite-dimensional approximation that performs well over the horizons considered in our experiments. More broadly, the results show that persistent covariance structure can be combined with Markovian GP inference for both forecasting and latent representation learning. Future work could for example include exploring adaptive or nonstationary mixing distributions and more direct GGPVAE inference.

AI USE STATEMENT

We used generative AI tools for specific supporting tasks during the research and writing process. The proposed GAMMA prior, its core theoretical formulation, and the main algorithms were developed by the authors.

After developing the GAMMA prior, we used generative AI for literature discovery and brainstorming about possible representation-learning settings in which the prior could be incorporated. This helped identify relevant work and candidate frameworks, including MGPVAE. The authors independently examined the suggested literature and made the final decisions regarding the modeling approach and its adaptation to the proposed prior. Generative AI was also used to assist with drafting parts of the introduction and abstract, language polishing, reference discovery, and suggesting keywords. For the introduction and abstract, the authors determined the scientific content, the intended role of each part, and the order in which the main points should be presented.

For the theoretical component, generative AI was used to improve the presentation and clarity of some proofs and to assist with the corresponding LaTeX code. The mathematical claims and their proofs are by the authors. For implementation, the authors developed the proposed framework and algorithms and started from existing code and resources for the underlying methods. Generative AI was used to assist in adapting and extending this code, and the resulting implementation was reviewed and tested by the authors.

The synthetic-data generation algorithm was designed by the authors, with generative AI assisting its implementation. The authors first analyzed and interpreted the experimental results independently and then used generative AI to check the reasonableness of those interpretations, consider possible alternative explanations, and improve their presentation. All final interpretations and conclusions were determined by the authors.

All AI-assisted work used in the final submission was reviewed by the authors. We take responsibility for the final content of this work, including its text, mathematical arguments, implementation, experimental analyses, data, and other artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the proposed GAMMA covariance construction, its Gauss–Laguerre state-space approximation, and the GGPVAE model in the main text. Complete proofs, dataset preprocessing procedures, experimental protocols, model architectures, and hyperparameter settings are provided in the appendix and supplementary material. Anonymous source code for reproducing the experiments is available at `https://anonymous.4open.science/r/GAMMA-FDF0/`.

REFERENCES

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. Dover Publications, New York, 1970.

O. E. Barndorff-Nielsen. Superposition of Ornstein–Uhlenbeck type processes. *Theory of Probability & Its Applications*, 45(2):175–194, 2001.

O. E. Barndorff-Nielsen, F. E. Benth, and A. E. D. Veraart. *Ambit Stochastics*, volume 88 of *Probability Theory and Stochastic Modelling*. Springer, 2018. doi: 10.1007/978-3-319-94129-5.

G. Benton, W. J. Maddox, and A. G. Wilson. Volatility based kernels and moving average means for accurate forecasting with Gaussian processes. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1798–1816. PMLR, 2022.

M. Bevilacqua and T. Faouzi. Estimation and prediction of Gaussian processes using generalized Cauchy covariance model under fixed domain asymptotics. *Electronic Journal of Statistics*, 13 (2):3025–3048, 2019.

L. M. Candanedo, V. Feldheim, and D. Deramaix. Appliances energy prediction data set. UCI Machine Learning Repository, 2017. URL https://archive.ics.uci.edu/ml/machine-learning-databases/00374/energydata_complete.csv.

F. P. Casale, A. V. Dalca, L. Saglietti, J. Listgarten, and N. Fusi. Gaussian process prior variational autoencoders. In *Advances in Neural Information Processing Systems*, volume 31, pp. 10390–10401, 2018.

J. Chen, W. Mu, Y. Li, and D. Li. On the identifiability and interpretability of Gaussian process models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 70267–70278, 2023.

W. Chen, N. Kiyohara, H. B. H. Zhu, J. Curran-Sebastian, S. Bhatt, and Y. Li. Recurrent memory for online interdomain Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 38, pp. 149482–149519, 2025. doi: 10.52202/085713-4499.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2980–2988, 2015.

N. Cressie. Mission CO_2ntrol: A statistical scientist's role in remote sensing of atmospheric carbon dioxide. *Journal of the American Statistical Association*, 113(521):152–168, 2018. doi: 10.1080/01621459.2017.1419136.

R. J. Elliott, J. Van Der Hoek, and W. P. Malcolm. Pairs trading. *Quantitative Finance*, 5(3): 271–276, 2005.

R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation, and testing. *Econometrica*, 55(2):251–276, 1987.

V. Fortuin, D. Baranchuk, G. Rätsch, and S. Mandt. GP-VAE: Deep probabilistic time series imputation. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1651–1661. PMLR, 2020.

E. B. Fox and D. B. Dunson. Multiresolution Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 25, pp. 746–754, 2012.

E. G. Gatev, W. N. Goetzmann, and K. G. Rouwenhorst. Pairs trading: Performance of a relative-value arbitrage rule. *The Review of Financial Studies*, 19(3):797–827, 2006. doi: 10.1093/rfs/hhj020.

G. H. Golub and J. H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of Computation*, 23:221–230, 1969.

N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore. Google Earth Engine: Planetary-scale geospatial analysis for everyone. *Remote Sensing of Environment*, 202:18–27, 2017.

J. Hartikainen and S. Särkkä. Kalman filtering and smoothing solutions to temporal Gaussian process regression models. In *Proceedings of the 2010 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 379–384. IEEE, 2010. doi: 10.1109/MLSP.2010.5589113.

E. Iglói and G. Terdik. Long-range dependence through gamma-mixed ornstein–uhlenbeck process. *Electronic Journal of Probability*, 4(16):1–33, 1999. doi: 10.1214/EJP.v4-53.

M. Jazbec, M. Ashman, V. Fortuin, M. Pearce, S. Mandt, and G. Rätsch. Scalable Gaussian process variational autoencoders. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3511–3519. PMLR, 2021.

R. G. Krishnan, U. Shalit, and D. Sontag. Structured inference networks for nonlinear state space models. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 2101–2109, 2017. doi: 10.1609/aaai.v31i1.10779.

K. Longi, J. Lindinger, O. Duennbier, M. Kandemir, A. Klami, and B. Rakitsch. Traversing time with multi-resolution Gaussian process state-space models. In *Proceedings of the 4th Annual Learning for Dynamics and Control Conference*, volume 168 of *Proceedings of Machine Learning Research*, pp. 366–377. PMLR, 2022.

P. Ma, A. Bhadra, and M. G. Genton. Beyond Matérn: On a class of interpretable confluent hypergeometric covariance functions. *Journal of the American Statistical Association*, 118(543): 2045–2058, 2023.

B. Matérn. *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 2nd edition, 1986. doi: 10.1007/978-1-4615-7892-5.

H. Nguyen, J. Liu, S. Kulawik, D. Baker, J. Hobbs, A. Braverman, M. Katzfuss, and V. Yadav. OCO-2 gridded bias-corrected XCO_2 and other select fields aggregated as level 4 daily files, version 3, 2022. URL `https://doi.org/10.5067/582L7HTJ343N`.

P. Nummi and L. Viitasaari. Necessary and sufficient conditions for continuity of hypercontractive processes and fields. *Statistics & Probability Letters*, 208:110049, 2024. doi: 10.1016/j.spl.2024.110049.

F. W. J. Olver, D. W. Lozier, R. F. Boisvert, and C. W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, 2010.

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. doi: 10.1214/24-STS923.

R. Ranganath, S. Gerrish, and D. Blei. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 814–822. PMLR, 2014.

S. S. Rangapuram, M. W. Seeger, J. Gasthaus, L. Stella, Y. Wang, and T. Januschowski. Deep state space models for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 31, pp. 7796–7805, 2018.

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

Y. Rubanova, R. T. Q. Chen, and D. Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5321–5331, 2019.

S. Särkkä and A. Solin. *Applied Stochastic Differential Equations*. Cambridge University Press, 2019.

S. Särkkä, A. Solin, and J. Hartikainen. Spatiotemporal learning via infinite-dimensional Bayesian filtering and smoothing: A look at Gaussian process regression through Kalman filtering. *IEEE Signal Processing Magazine*, 30(4):51–61, 2013.

E. Schwartz and J. E. Smith. Short-term variations and long-term dynamics in commodity prices. *Management Science*, 46(7):893–911, 2000. doi: 10.1287/mnsc.46.7.893.12034.

A. Solin and S. Särkkä. Gaussian quadratures for state space approximation of scale mixtures of squared exponential covariance functions. In *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, Piscataway, NJ, 2014a. IEEE. doi: 10.1109/MLSP.2014.6958899.

A. Solin and S. Särkkä. Explicit link between periodic covariance functions and state space models. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pp. 904–912. PMLR, 2014b.

F. Tronarp, T. Karvonen, and S. Särkkä. Mixture representation of the Matérn class with applications in state space approximations and Bayesian quadrature. In *Proceedings of the 2018 IEEE 28th International Workshop on Machine Learning for Signal Processing*. IEEE, 2018. doi: 10.1109/MLSP.2018.8516992.

O. Vasicek. An equilibrium characterization of the term structure. *Journal of Financial Economics*, 5(2):177–188, 1977.

D. Wunch, G. C. Toon, J.-F. L. Blavier, R. A. Washenfelder, J. Notholt, B. J. Connor, D. W. T. Griffith, V. Sherlock, and P. O. Wennberg. The total carbon column observing network. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 369(1943):2087–2112, 2011. doi: 10.1098/rsta.2010.0240.

H. Zhu, C. Balsells-Rodas, and Y. Li. Markovian Gaussian process variational autoencoders. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 42938–42961. PMLR, 2023.

A Gaussian Process Regression

Gaussian processes (Rasmussen & Williams, 2006) provide a flexible nonparametric prior over latent functions. Let $f : \mathbb{R} \to \mathbb{R}$ denote an unknown time-indexed signal. A zero-mean Gaussian process prior is written as

$$f(t) \sim \mathcal{GP}\big(0, k(t, t')\big),$$

where the covariance function $k(t, t')$ determines the smoothness, amplitude, and temporal dependence structure of the process. In the common stationary case, the covariance depends only on the lag $\tau = t - t'$, so that

$$k(t, t') = k(\tau).$$

Let $\mathcal{D} = \{(t_i, y_i)\}_{i=1}^n$ denote a set of observations collected at input locations $t_1, \ldots, t_n$. In the regression setting, the observed output is modeled as a noisy realization of the latent function,

$$y_i = f(t_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}}),$$

where σ^2_{noise} denotes the observation-noise variance. Let $\mathbf{y} = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix}^\top$ be the vector of observations, and define the covariance matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ by $[\mathbf{K}]_{ij} = k(t_i, t_j)$, for $i, j = 1, \ldots, n$. The covariance matrix of the noisy observations is then $\mathbf{K}_y = \mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I}$. For a new input location t_*, the joint distribution of the training observations and the latent test value $f(t_*)$ is Gaussian:

$$\begin{bmatrix} \mathbf{y} \\ f(t_*) \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_y & \mathbf{k}_{f,*} \\ \mathbf{k}_{*,f} & k(t_*, t_*) \end{bmatrix}\right),$$

where $\mathbf{k}_{*,f} = \begin{bmatrix} k(t_*, t_1) & k(t_*, t_2) & \cdots & k(t_*, t_n) \end{bmatrix}$ is the covariance vector between the test location and the training inputs, and $\mathbf{k}_{f,*} = \mathbf{k}_{*,f}^\top$. Conditioning the joint Gaussian distribution on the observed data gives the posterior predictive distribution

$$p\big(f(t_*) \mid \mathbf{y}, \boldsymbol{\theta}\big) = \mathcal{N}\left(\mu_{\text{GP}}(t_*), \sigma^2_{\text{GP}}(t_*)\right).$$

The posterior mean and variance are

$$\mu_{\text{GP}}(t_*) = \mathbf{k}_{*,f}\left(\mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I}\right)^{-1}\mathbf{y}, \tag{5}$$

$$\sigma^2_{\text{GP}}(t_*) = k(t_*, t_*) - \mathbf{k}_{*,f}\left(\mathbf{K} + \sigma^2_{\text{noise}}\mathbf{I}\right)^{-1}\mathbf{k}_{f,*}. \tag{6}$$

The mean in equation 5 provides the posterior estimate of the latent function at t_*, while the variance in equation 6 quantifies the remaining uncertainty after conditioning on the observed data. The direct implementation of Gaussian process regression requires solving linear systems involving the dense matrix $\mathbf{K}_y$. In practice this is typically done using a Cholesky factorization, $\mathbf{K}_y = \mathbf{L}\mathbf{L}^\top$, which has computational complexity $\mathcal{O}(n^3)$.

B Generalized Gauss–Laguerre Quadrature

Generalized Gauss–Laguerre quadrature is a classical Gaussian quadrature rule for weighted integrals on the positive real line (Abramowitz & Stegun, 1970; Olver et al., 2010). In fact, it provides a deterministic rule for approximating integrals whose measure contains a Laguerre-type weight. For a parameter $\gamma > -1$, the quadrature rule (with respect to the measure $x^\gamma e^{-x}dx$) is given by

$$\int_0^\infty f(x)x^\gamma e^{-x}\,dx \approx \sum_{i=1}^n \omega_i f(x_i).$$

The quadrature rule is specified by a set of nodes x_i and positive weights ω_i, $i = 1, \ldots, n$. The nodes are chosen as the zeros of the generalized Laguerre polynomial $L_n^{(\gamma)}(x)$. With this choice, the rule is exact for all polynomials f of degree at most $2n - 1$.

The corresponding quadrature weights can be written as

$$\omega_i = \frac{\Gamma(n + \gamma + 1)\, x_i}{n!\,(n+1)^2\left[L_{n+1}^{(\gamma)}(x_i)\right]^2}, \qquad i = 1, \ldots, n, \tag{7}$$

where $\Gamma(\cdot)$ is the gamma function. Thus, once the zeros $\{x_i\}_{i=1}^n$ of $L_n^{(\gamma)}$ have been computed, the weights are obtained directly from equation 7. This construction is particularly useful when an integral contains a Gamma-type density, since the change of variables can often transform the density into the canonical Laguerre weight $x^\gamma e^{-x}$.

C State-Space Representation of the GAMMA–GL Surrogate

The exact GAMMA covariance is obtained by continuously mixing half-integer Matérn covariances over their rate parameter and given by

$$R_{p,\sigma^2,\alpha,\beta}(h) = \int_0^\infty k_p(h \mid \lambda, \sigma^2)\, \nu_{\alpha,\beta}(d\lambda), \qquad \nu_{\alpha,\beta}(d\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}\, d\lambda.$$

Since the mixing distribution has continuous support, the exact GAMMA process does not in general admit a finite-dimensional Markov representation. Applying generalized Gauss–Laguerre quadrature yields the finite surrogate

$$R_{p,\sigma^2,\alpha,\beta}(h) \approx R^{(J,\mathrm{GL})}_{p,\sigma^2,\alpha,\beta}(h) = \sum_{r=1}^{J} w_r\, k_p(h \mid \lambda_r, \sigma^2), \qquad \lambda_r = \frac{x_r}{\beta}, \qquad w_r = \frac{w_r^{\mathrm{GL}}}{\Gamma(\alpha)},$$

where $\{x_r, w_r^{\mathrm{GL}}\}_{r=1}^J$ are the generalized Gauss–Laguerre nodes and weights associated with the weight $x^{\alpha-1}e^{-x}$. In particular, $w_r > 0$ and $\sum_{r=1}^J w_r = 1$.

For each quadrature node λ_r, the corresponding half-integer Matérn component admits a $(p+1)$-dimensional Markov realization. Let $\mathbf{s}_r(t) \in \mathbb{R}^{p+1}$ denote the state of the rth component, with dynamics

$$d\mathbf{s}_r(t) = \mathbf{F}_r \mathbf{s}_r(t)\, dt + \mathbf{L}_r\, dB_r(t), \qquad f_r(t) = \mathbf{H}_r \mathbf{s}_r(t),$$

where $\mathbf{F}_r = \mathbf{F}(\lambda_r)$ and $B_1, \ldots, B_J$ are mutually independent Brownian motions. Equivalently, each scalar component satisfies a higher-order Ornstein–Uhlenbeck equation of the form $\left(\frac{d}{dt} + \lambda_r\right)^{p+1} f_r(t) = \xi_r(t)$, with white-noise forcing ξ_r chosen so that f_r has covariance $k_p(\cdot \mid \lambda_r, \sigma^2)$.

The finite GAMMA process is then defined by

$$X^{(J,\mathrm{GL})}(t) = \sum_{r=1}^{J} \sqrt{w_r}\, f_r(t) = \sum_{r=1}^{J} \sqrt{w_r}\, \mathbf{H}_r \mathbf{s}_r(t).$$

Because the component processes are independent,

$$\mathrm{Cov}\Big(X^{(J,\mathrm{GL})}(t), X^{(J,\mathrm{GL})}(t')\Big) = \sum_{r=1}^{J} w_r\, \mathrm{Cov}(f_r(t), f_r(t')) = \sum_{r=1}^{J} w_r\, k_p(t-t' \mid \lambda_r, \sigma^2) = R^{(J,\mathrm{GL})}_{p,\sigma^2,\alpha,\beta}(t-t').$$

Thus, the state-space construction realizes exactly the GAMMA–GL surrogate covariance.

To obtain a single Markov system, we stack the component states into $\mathbf{s}^{(J)}(t) = \big[\mathbf{s}_1(t)^\top, \ldots, \mathbf{s}_J(t)^\top\big]^\top \in \mathbb{R}^{J(p+1)}$, and define $\mathbf{F}^{(J)} = \mathrm{blockdiag}(\mathbf{F}_1, \ldots, \mathbf{F}_J)$, $\mathbf{L}^{(J)} = \mathrm{blockdiag}(\mathbf{L}_1, \ldots, \mathbf{L}_J)$, and

$$\mathbf{H}^{(J)}_{\mathrm{aug}} = \big[\sqrt{w_1}\mathbf{H}_1 \quad \sqrt{w_2}\mathbf{H}_2 \quad \cdots \quad \sqrt{w_J}\mathbf{H}_J\big].$$

With $\mathbf{B}^{(J)}(t) = \big[B_1(t), \ldots, B_J(t)\big]^\top$, the augmented state satisfies

$$d\mathbf{s}^{(J)}(t) = \mathbf{F}^{(J)}\mathbf{s}^{(J)}(t)\, dt + \mathbf{L}^{(J)}\, d\mathbf{B}^{(J)}(t), \qquad X^{(J,\mathrm{GL})}(t) = \mathbf{H}^{(J)}_{\mathrm{aug}}\mathbf{s}^{(J)}(t).$$

Now let t_i and t_{i+1} be consecutive observation times, and write $\Delta_i = t_{i+1} - t_i$. For the rth component, the linear SDE solution over $[t_i, t_{i+1}]$ takes the form

$$\mathbf{s}_r(t_{i+1}) = \mathbf{A}_{r,i}\mathbf{s}_r(t_i) + \mathbf{q}_{r,i}, \qquad \mathbf{A}_{r,i} = \exp(\mathbf{F}_r \Delta_i),$$

where

$$\mathbf{q}_{r,i} = \int_{t_i}^{t_{i+1}} \exp(\mathbf{F}_r(t_{i+1} - \tau))\, \mathbf{L}_r\, dB_r(\tau), \qquad \mathbf{q}_{r,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_{r,i}),$$

with covariance

$$\mathbf{Q}_{r,i} = \int_0^{\Delta_i} \exp(\mathbf{F}_r u)\mathbf{L}_r\mathbf{L}_r^\top \exp(\mathbf{F}_r^\top u)\, du.$$

For stationary components one may equivalently write $\mathbf{Q}_{r,i} = \mathbf{P}_{\infty,r} - \mathbf{A}_{r,i}\mathbf{P}_{\infty,r}\mathbf{A}_{r,i}^\top$, where $\mathbf{P}_{\infty,r}$ is the stationary covariance of the rth Matérn state.

Since the Brownian drivers are independent across r, the transition and process-noise covariances of the augmented state are block diagonal:

$$\mathbf{A}_i^{(J)} = \mathrm{blockdiag}(\mathbf{A}_{1,i}, \ldots, \mathbf{A}_{J,i}), \qquad \mathbf{Q}_i^{(J)} = \mathrm{blockdiag}(\mathbf{Q}_{1,i}, \ldots, \mathbf{Q}_{J,i}).$$

Hence the GAMMA–GL surrogate admits the discrete-time Markov form

$$\mathbf{s}_{i+1}^{(J)} = \mathbf{A}_i^{(J)}\mathbf{s}_i^{(J)} + \mathbf{q}_i^{(J)}, \qquad \mathbf{q}_i^{(J)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i^{(J)}).$$

Therefore, although the exact GAMMA process is a continuous mixture and does not generally have a finite-dimensional Markov representation, its J-point GAMMA–GL surrogate is a linear Gaussian state-space model of dimension $J(p+1)$.

For GGPVAE, this construction is applied independently to each latent coordinate ℓ. Writing $\mathbf{s}^{(\ell,J)}(t) = \left[(\mathbf{s}^{(\ell,1)}(t))^\top, \ldots, (\mathbf{s}^{(\ell,J)}(t))^\top\right]^\top$, the latent dynamics take the form

$$\mathbf{s}_{i+1}^{(\ell,J)} = \mathbf{A}_i^{(\ell,J)}\mathbf{s}_i^{(\ell,J)} + \mathbf{q}_i^{(\ell,J)}, \qquad \mathbf{q}_i^{(\ell,J)} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}_i^{(\ell,J)}),$$

and the scalar latent process is recovered through

$$Z_i^{(\ell)} = \mathbf{H}_{\mathrm{aug}}^{(\ell,J)}\mathbf{s}_i^{(\ell,J)}, \qquad \mathbf{H}_{\mathrm{aug}}^{(\ell,J)} = \left[\sqrt{w_{\ell 1}}\mathbf{H}^{(\ell,1)} \quad \cdots \quad \sqrt{w_{\ell J}}\mathbf{H}^{(\ell,J)}\right].$$

Thus, after the GAMMA–GL reduction, each latent GAMMA prior has exactly the linear Gaussian form required for Kalman filtering and smoothing, with the only difference from a standard Markovian GP prior being the enlarged state dimension $J(p+1)$ and the presence of multiple Matérn components with different mean-reversion rates.

D Construction of the process

One way to construct the process X is stochastic integration with respect to a white noise random measure (Barndorff-Nielsen et al., 2018, Chapter 5). Indeed, let $\mathcal{W}$ be a white noise on $\mathbb{R}^2$ with control measure $\nu \otimes \mathrm{Leb}$, where Leb is the Lebesgue measure on $\mathbb{R}$. What this means is that for any measurable $A_1, \ldots, A_n \subset \mathbb{R}^2$ with finite $\nu \otimes \mathrm{Leb}$-measure,

$$\mathcal{W}(A_1), \ldots, \mathcal{W}(A_n)$$

are mutually independent Gaussian random variables with mean zero and variances

$$(\nu \otimes \mathrm{Leb})(A_1), \ldots, (\nu \otimes \mathrm{Leb})(A_n),$$

respectively. The Matérn process on $\mathbb{R}$ can be represented as a moving average process using the one-sided kernel

$$g_{p,\lambda}(t) := c_{p,\lambda} t^{p-1} \exp(-\lambda t), \quad t > 0,$$

where $c_{p,\lambda} > 0$ is a suitable scaling constant. Subsequently, setting

$$X(t) := \sigma \int_0^\infty \int_{-\infty}^t g_{p,\lambda}(t-u)\mathcal{W}(d\lambda, du), \quad t \in \mathbb{R} \tag{8}$$

gives rise to a Gaussian process X with mean zero and covariance

$$\begin{aligned}
\mathbb{E}[X(t)X(t+h)] &= \sigma^2 \int_0^\infty \int_{-\infty}^t g_{p,\lambda}(t-u) g_{p,\lambda}(t+h-u)\, \nu(d\lambda) du \\
&= \int_0^\infty \left(\sigma^2 \int_0^\infty g_{p,\lambda}(u) g_{p,\lambda}(u+h)\, du \right) \nu(d\lambda) \\
&= \int_0^\infty k_p(h \,|\, \lambda, \sigma^2)\, \nu(d\lambda).
\end{aligned}$$

Hence, equation 8 is a mixed moving average representation of the superposition of Matérn processes.

E BASIC PROPERTIES

We begin with the following result providing small and large lag asymptotics for the covariance as well as small and high frequency asymptotics for spectral density. Recall that a spectral density of $R_{p,\sigma^2,\alpha,\beta}$ is defined by

$$f_{p,\sigma^2,\alpha,\beta}(\omega) = \frac{1}{2p+1}\int_{\mathbb{R}} e^{-i\omega h} R_{p,\sigma^2,\alpha,\beta}(|h|)\, dh.$$

Theorem 3. *Let $R_{p,\sigma^2,\alpha,\beta}$ be given by equation 3 and $f_{p,\sigma^2,\alpha,\beta}(\omega)$ its spectral density. Then:*

(i) As $|h| \to 0$,

$$R_{p,\sigma^2,\alpha,\beta}(h) - R_{p,\sigma^2,\alpha,\beta}(0) \sim \begin{cases} -\sigma^2 \dfrac{\alpha}{\beta}|h|, & p = 0, \\ -\sigma^2 \dfrac{\alpha(\alpha+1)}{2(2p-1)\beta^2}|h|^2, & p \geq 1. \end{cases} \tag{9}$$

(ii) As $h \to \infty$,

$$R_{p,\sigma^2,\alpha,\beta}(h) \sim K_{p,\sigma^2,\alpha,\beta}\, h^{-\alpha}, \tag{10}$$

where

$$K_{p,\sigma^2,\alpha,\beta} := \sigma^2 \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{k=0}^{p} a_{p,k}\Gamma(\alpha+k).$$

(iii) If $0 < \alpha < 1$, then, as $\omega \to 0$,

$$f_{p,\sigma^2,\alpha,\beta}(\omega) \sim L_{p,\sigma^2,\alpha,\beta}|\omega|^{\alpha-1}, \tag{11}$$

where

$$L_{p,\sigma^2,\alpha,\beta} := \frac{K_{p,\sigma^2,\alpha,\beta}}{p+\frac{1}{2}}\Gamma(1-\alpha)\sin\Big(\frac{\pi\alpha}{2}\Big).$$

(iv) As $\omega \to \infty$,

$$f_{p,\sigma^2,\alpha,\beta}(\omega) \sim \kappa_p \tau_p(\sigma^2,\alpha,\beta)|\omega|^{-(2p+2)}, \tag{12}$$

where

$$\kappa_p := \frac{2\sqrt{\pi}\,\Gamma(p+1)}{(2p+1)\Gamma\big(p+\frac{1}{2}\big)} \tag{13}$$

and

$$\tau_p(\sigma^2,\alpha,\beta) := \sigma^2\,\mathbb{E}[\Lambda^{2p+1}] = \sigma^2 \frac{\Gamma(\alpha+2p+1)}{\Gamma(\alpha)\beta^{2p+1}}, \qquad \Lambda \sim \Gamma(\alpha,\beta).$$

Proof. **(i)** Recall

$$R_{p,\sigma^2,\alpha,\beta}(h) = \sigma^2 \int_0^\infty e^{-\lambda|h|} P_p(\lambda|h|)\nu_{\alpha,\beta}(d\lambda),$$

where

$$P_p(u) = \sum_{m=0}^{p} a_{p,m}u^m, \qquad P_p(0) = 1.$$

The function $g_p(u) := e^{-u}P_p(u)$ is analytic at the origin and thus can be written as [1]

$$g_p(u) = \sum_{n=0}^{\infty} b_{p,n}u^n.$$

[1] The coefficients $b_{p,n}$ can be computed explicitly and are given by

$$b_{p,n} = \sum_{m=0}^{\min(p,n)} a_{p,m}\frac{(-1)^{n-m}}{(n-m)!}, \qquad n \geq 0.$$

This leads to representations

$$k_p(h \mid \lambda, \sigma^2) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n} \lambda^n |h|^n \tag{14}$$

and

$$R_{p,\sigma^2,\alpha,\beta}(h) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n} \mathbb{E}[\Lambda^n] |h|^n, \qquad \Lambda \sim \Gamma(\alpha, \beta) \tag{15}$$

for small $|h|$. In the latter we have applied dominated convergence theorem in the integral

$$R_{p,\sigma^2,\alpha,\beta}(h) = \int_0^{\infty} k_p(h \mid \lambda, \sigma^2)\, \nu_{\alpha,\beta}(\lambda)\, d\lambda, \qquad \nu_{\alpha,\beta}(\lambda) = \frac{\beta^\alpha}{\Gamma(\alpha)} \lambda^{\alpha-1} e^{-\beta\lambda}$$

which is justified since Λ have all the moments finite. Therefore

$$R_{p,\sigma^2,\alpha,\beta}(h) - R_{p,\sigma^2,\alpha,\beta}(0) = \sigma^2 \sum_{n=1}^{\infty} b_{p,n} \mathbb{E}[\Lambda^n] |h|^n.$$

In particular, using also $\mathbb{E}[\Lambda] = \frac{\alpha}{\beta}$ and $\mathbb{E}[\Lambda^2] = \frac{\alpha(\alpha+1)}{\beta^2}$ this gives

$$R_{p,\sigma^2,\alpha,\beta}(h) - R_{p,\sigma^2,\alpha,\beta}(0) \sim \sigma^2 \frac{\alpha}{\beta} b_{p,1} |h| + \sigma^2 \frac{\alpha(\alpha+1)}{\beta^2} b_{p,2} |h|^2.$$

Now for $p = 0$ we have $P_0(u) = 1$ leading to $g_0(u) = 1 - u + O(u^2)$, i.e. $b_{0,1} = -1$. This proves the case $p = 0$. For $p \geq 1$, we have $a_{p,0} = a_{p,1} = 1$ and

$$a_{p,2} = \frac{p-1}{2p-1}.$$

This leads to

$$\begin{aligned} e^{-u} P_p(u) &= \left(1 - u + \frac{u^2}{2} + O(u^3)\right) \left(1 + u + a_{p,2} u^2 + O(u^3)\right) \\ &= 1 + (a_{p,1} - 1)\, u + \left(a_{p,2} - a_{p,1} + \frac{1}{2}\right) u^2 + O(u^3) \\ &= 1 - \frac{1}{2(2p-1)} u^2 + O(u^2). \end{aligned}$$

This handles case $p \geq 1$ completing the proof of equation 9.

(ii) We have

$$\begin{aligned} R_{p,\sigma^2,\alpha,\beta}(h) &= \sigma^2 \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{k=0}^{p} a_{p,k} \Gamma(\alpha + k) \frac{h^k}{(\beta + h)^{\alpha+k}} \\ &= h^{-\alpha} \sigma^2 \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{k=0}^{p} a_{p,k} \Gamma(\alpha + k) \left(1 + \frac{\beta}{h}\right)^{-(\alpha+k)}. \end{aligned} \tag{16}$$

Since

$$\left(1 + \frac{\beta}{h}\right)^{-(\alpha+k)} \longrightarrow 1, \qquad h \to \infty$$

for all k, we obtain equation 10.

(iii) Since $R_{p,\sigma^2,\alpha,\beta}$ is stationary and even, its spectral density admits the cosine representation

$$f_{p,\sigma^2,\alpha,\beta}(\omega) = \frac{1}{p + \frac{1}{2}} \int_0^{\infty} R_{p,\sigma^2,\alpha,\beta}(h) \cos(\omega h)\, dh.$$

By part *(ii)*, we have

$$R_{p,\sigma^2,\alpha,\beta}(h) \sim K_{p,\sigma^2,\alpha,\beta} h^{-\alpha}, \qquad h \to \infty.$$

Write

$$R_{p,\sigma^2,\alpha,\beta}(h) = K_{p,\sigma^2,\alpha,\beta} h^{-\alpha} + r(h), \qquad r(h) = o(h^{-\alpha}), \qquad h \to \infty.$$

Using equation 16, we get

$$r(h) = h^{-\alpha}\sigma^2 \frac{\beta^\alpha}{\Gamma(\alpha)} \sum_{k=0}^{p} a_{p,k}\Gamma(\alpha+k)\left[\left(1+\frac{\beta}{h}\right)^{-(\alpha+k)} - 1\right].$$

Since

$$\left(1+\frac{\beta}{h}\right)^{-(\alpha+k)} - 1 = O(h^{-1}), \qquad h \to \infty,$$

we have

$$r(h) = O(h^{-\alpha-1}), \qquad h \to \infty.$$

Thus r is integrable at infinity. Also, since $0 < \alpha < 1$, the function $h^{-\alpha}$ is integrable near zero, and $R_{p,\sigma^2,\alpha,\beta}$ is bounded near zero. Hence $r \in L^1(0,\infty)$. Now we decompose

$$\begin{aligned} f_{p,\sigma^2,\alpha,\beta}(\omega) &= \frac{K_{p,\sigma^2,\alpha,\beta}}{p+\frac{1}{2}} \int_0^\infty h^{-\alpha}\cos(\omega h)\, dh + \frac{1}{p+\frac{1}{2}} \int_0^\infty r(h)\cos(\omega h)\, dh \\ &=: I_1(\omega) + I_2(\omega). \end{aligned}$$

For I_1, substituting $u = |\omega|h$ gives

$$I_1(\omega) = \frac{K_{p,\sigma^2,\alpha,\beta}}{p+\frac{1}{2}} |\omega|^{\alpha-1} \int_0^\infty u^{-\alpha}\cos u\, du,$$

where

$$\int_0^\infty u^{-\alpha}\cos u\, du = \Gamma(1-\alpha)\sin\Big(\frac{\pi\alpha}{2}\Big)$$

by the classical cosine-integral identity. Hence

$$I_1(\omega) = \frac{K_{p,\sigma^2,\alpha,\beta}}{p+\frac{1}{2}} \Gamma(1-\alpha)\sin\Big(\frac{\pi\alpha}{2}\Big)\, |\omega|^{\alpha-1}.$$

It remains to handle I_2. Since $r \in L^1(0,\infty)$, $I_2(\omega)$ is bounded and hence $I_2(\omega) = o\big(|\omega|^{\alpha-1}\big)$ as $\omega \to 0$. This proves equation 11.

(iv) The half-integer Matérn factor has spectral density

$$f_p(\omega \mid \lambda, \sigma^2) = \kappa_p \sigma^2 \frac{\lambda^{2p+1}}{(\lambda^2+\omega^2)^{p+1}}, \qquad \omega \in \mathbb{R},$$

where κ_p is defined in equation 13. By linearity of the Fourier transform, this gives

$$\begin{aligned} f_{p,\sigma^2,\alpha,\beta}(\omega) &= \int_0^\infty f_p(\omega \mid \lambda, \sigma^2)\nu_{\alpha,\beta}(d\lambda) = \kappa_p\sigma^2 \int_0^\infty \frac{\lambda^{2p+1}}{(\lambda^2+\omega^2)^{p+1}}\nu_{\alpha,\beta}(d\lambda) \\ &= \kappa_p\sigma^2|\omega|^{-(2p+2)} \int_0^\infty \lambda^{2p+1}\left(1+\frac{\lambda^2}{\omega^2}\right)^{-(p+1)} \nu_{\alpha,\beta}(d\lambda). \end{aligned}$$

Together with dominated convergence theorem this gives

$$|\omega|^{2p+2} f_{p,\sigma^2,\alpha,\beta}(\omega) \longrightarrow \kappa_p\sigma^2 E[\Lambda^{2p+1}], \qquad |\omega| \to \infty.$$

This proves equation 12 and completes the whole proof. □

Remark 1. *By Theorem 3 part (iv), $\tau_p(\sigma^2,\alpha,\beta)$ is the unique microergodic parameter of the original Gamma-mixed half-integer Matérn family at order p. This shows that, for the original Gamma mixing law, the leading fixed-domain channel is governed by the single combination*

$$\tau_p(\sigma^2,\alpha,\beta) = \sigma^2 \frac{\Gamma(\alpha+2p+1)}{\Gamma(\alpha)\beta^{2p+1}}.$$

Thus, although the model is parametrized by (σ^2,α,β), fixed-domain asymptotics retain only this one-dimensional microergodic quantity. In particular, the original Gamma law does not provide separate fixed-domain identifiability of σ^2, α, and β.

The following corollary shows that on short scales, X has path properties similar to the underlying Matérn process. Brownian motion. As such, X preserves local fluctuations of the Matérn class while still having long memory. The proof follows directly from Theorem 3 part (i) together with (Nummi & Viitasaari, 2024, Corollary 2.11). As such, we omit the details.

Corollary 1. *There exists deterministic constants C_p such that for $p = 0$ we have*

$$\limsup_{|t-s|\to 0} \frac{|X_t - X_s|}{|t-s|^{\frac{1}{2}} \left[\log \frac{1}{|t-s|}\right]^{\frac{1}{2}}} \leq C_0.$$

and for $p \geq 1$ we have

$$\limsup_{|t-s|\to 0} \frac{|X_t - X_s|}{|t-s| \left[\log \frac{1}{|t-s|}\right]^{\frac{1}{2}}} \leq C_p.$$

F Proof of Theorem 1

Since both processes X and $X^{(J,GL)}$ are Gaussian, it is sufficient to prove convergence of the finite dimensional distributions together with tightness of the sequence $X^{(J,GL)}$ in J. The convergence of the finite dimensional distributions on the other hand follow from pointwise convergence of the covariance function that holds by design of the Gauss–Laguerre quadrature rule. Hence Theorem 1 follows from the following result providing tightness.

Theorem 4 (Uniform increment bounds and tightness of the GAMMA–GL surrogate)**.** *Fix $p \in \mathbb{N}_0$, $\alpha > 0$, $\beta > 0$, and $\sigma^2 > 0$. Let $X^{(J,GL)} = \{X^{(J,GL)}(t) : t \in \mathbb{R}\}$ be the centered stationary Gaussian process with GAMMA–GL surrogate covariance, $R^{(J,\mathrm{GL})}_{p,\sigma^2,\alpha,\beta}(h)$. Then:*

1. *If $p = 0$, then for every $J \geq 1$ and every $h \in \mathbb{R}$ we have*
$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] \leq 2\sigma^2 \frac{\alpha}{\beta}|h|.$$

2. *If $p \geq 1$, then there exists a constant M_p depending only on p such that for every $J \geq 2$ and every $h \in \mathbb{R}$ we have*
$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] \leq \sigma^2 M_p \frac{\alpha(\alpha+1)}{\beta^2}|h|^2.$$

   define $g_p(u) := e^{-u}P_p(u), M_p := \sup_{u\geq 0}\left|g_p''(u)\right|$. Then $M_p < \infty$, and for every $J \geq 2$ and every $h \in \mathbb{R}$,
$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] \leq \sigma^2 M_p \frac{\alpha(\alpha+1)}{\beta^2}|h|^2.$$

Consequently, for every p and every $T < \infty$ the family $\{X_J : J \geq 1\}$ is tight in $C([0,T])$.

Proof. By stationarity,

$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] = 2\left(R^{(J,\mathrm{GL})}_{p,\sigma^2,\alpha,\beta}(0) - R^{(J,\mathrm{GL})}_{p,\sigma^2,\alpha,\beta}(h)\right).$$

Since $P_p(0) = 1$ and the normalized Gauss–Laguerre weights satisfy $\sum_{r=1}^J w_r = 1$, we obtain

$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] = 2\sigma^2 \sum_{r=1}^J w_r \left[1 - e^{-\lambda_r|h|}P_p(\lambda_r|h|)\right] = 2\sigma^2 \sum_{r=1}^J w_r \left[1 - g_p(u)\right],$$

where we have again used notation $g_p(u) := e^{-u}P_p(u)$. Now for $p = 0$ we have $P_0(u) = 1$ allowing to use $0 \leq 1 - e^{-u} \leq u$ for $u \geq 0$ leading to

$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] \leq 2\sigma^2|h| \sum_{r=1}^J w_r\lambda_r = 2\sigma^2|h|\frac{\alpha}{\beta}.$$

Here we have also used the fact that GAMMA–GL approximation has the moment-matching property. That is, the J-point generalized Gauss–Laguerre rule is exact whenever $m \leq 2J-1$, a standard property of Gaussian quadrature (Golub & Welsch, 1969), that leads to

$$\mathbb{E}[\Lambda^m] = \sum_{r=1}^{J} \frac{w_r^{\mathrm{GL}}}{\Gamma(\alpha)} \left(\frac{x_r}{\beta}\right)^m = \sum_{r=1}^{J} w_r \lambda_r^m, \qquad m = 0, \ldots, 2J-1.$$

Let next $p \geq 1$. As $g_p(0) = 1$ and $g_p'(0) = 0$, Taylor's formula gives

$$g_p(u) = g_p(0) + g_p'(0)u + u^2 \int_0^1 (1-s) g_p''(su)\, ds$$

that allows to bound

$$|1 - g_p(u)| \leq u^2 \int_0^1 (1-s)|g_p''(su)|\, ds \leq u^2 \sup_{x \geq 0} |g_p''(x)| \int_0^1 (1-s)\, ds =: \frac{M_p}{2} u^2.$$

Hence we obtain

$$\mathbb{E}\left[(X_J(t+h) - X_J(t))^2\right] \leq 2\sigma^2 |h| \sum_{r=1}^{J} w_r \frac{M_p}{2} \lambda_r^2 = \sigma^2 M_p \frac{\alpha(\alpha+1)}{\beta^2} |h|^2$$

where we have again used the moment–matching property. This proves the claimed upper bound for $p \geq 1$ as well. Finally, the tightness follows directly from hypercontractivity of Gaussian distribution allowing to handle higher order moments together with Kolmogorov criterion. This completes the whole proof.

□

G Proof of Theorem 2

Proof of Theorem 2. Recall that the finite atomic surrogate covariance is given by

$$R_{p,\sigma^2,\alpha,\beta}^{(J,GL)}(h) = \sum_{r=1}^{J} w_r\, k_p(h \mid \lambda_r)$$

which takes a form

$$R_{p,\sigma^2,\alpha,\beta}^{(J,GL)}(h) = \sum_{r=1}^{J} w_r \left(\sigma^2 \sum_{n=0}^{\infty} b_{p,n} \lambda_r^n |h|^n\right) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n} \left(\sum_{r=1}^{J} w_r \lambda_r^n\right) |h|^n$$

when combined with equation 14. Together with equation 15 this gives us

$$R_{p,\sigma^2,\alpha,\beta}(h) - R_{p,\sigma^2,\alpha,\beta}^{(J,GL)}(h) = \sigma^2 \sum_{n=0}^{\infty} b_{p,n} \left(\mathbb{E}[\Lambda^n] - \sum_{r=1}^{J} w_r \lambda_r^n\right) |h|^n.$$

By moment-matching property, the coefficients of order $j = 0, \ldots, 2J-1$ vanish and we have

$$R_{p,\sigma^2,\alpha,\beta}(h) - R_{p,\sigma^2,\alpha,\beta}^{(J,GL)}(h) = \sigma^2 \sum_{n=2J}^{\infty} b_{p,n} \left(\mathbb{E}[\Lambda^n] - \sum_{r=1}^{J} w_r \lambda_r^n\right) |h|^n.$$

This completes the proof.

For any finite J, GAMMA–GL is a finite sum of Matérn components and therefore ultimately has exponentially decaying covariance at sufficiently large lags. It should thus be interpreted as a finite-dimensional approximation to the polynomially decaying GAMMA covariance over finite lag ranges; the exact long-memory behavior is recovered only in the $J \to \infty$ limit.

□

H DETAILS ON EXPERIMENTS

In this section, we provide additional details on the experimental setup, including descriptions of the datasets, evaluation metrics, and supplementary results. Following the organization of the main paper, we consider two groups of experiments: kernel evaluation and representation learning. We also present additional figures and analyses to provide further insight into the predictive behavior of the proposed Gamma-based models and GGPVAE across the different experimental settings.

H.1 KERNEL EVALUATION

We first provide additional details on the synthetic-data experiments, including the data-generation process, parameter settings, and evaluation protocol. We then present supplementary results that further assess the GAMMA–GL approximation, with particular emphasis on how closely the finite-dimensional surrogate reproduces the behavior of the exact Gamma model across different approximation levels and experimental settings.

We next provide further background on the real-world datasets used for kernel evaluation, together with additional details on preprocessing, train–test construction, and the motivation for each dataset. Supplementary figures and results are also included to further characterize the predictive behavior of the Gamma and Matérn models across different forecasting horizons and data regimes.

We evaluate the models using a train-fraction sweep with cumulative test portions. For each training fraction $r \in (0,1)$, let $n_{\text{train}} = \lfloor rn \rfloor$, $\mathcal{D}_{\text{train}} = \{y_1, \ldots, y_{n_{\text{train}}}\}$, $\mathcal{D}_{\text{test}} = \{y_{n_{\text{train}}+1}, \ldots, y_n\}$. In our experiments, we used training fractions $r \in \{0.5, 0.7, 0.8\}$.

H.1.1 SYNTHETIC DATA

We generate synthetic data from the Gamma-mixture model using its finite GAMMA–GL representation. For each replicate, we set the tail parameter to $\alpha = 0.2$, sufficiently far from the Matérn limit, and the mean-reversion scale to $\bar{\lambda} = 10$. The continuous mixture over reversion rates is approximated using $J = 128$ GAMMA-GL quadrature points. We simulate time series of length $n = 1500$ on an evenly spaced grid with $\Delta t = 0.001$, latent variance $\sigma^2 = 1$, and additive Gaussian observation noise with standard deviation 0.01. This setting produces trajectories with a broad range of reversion time scales and preserves the slow-decay dependence induced by the Gamma mixing distribution. Using multiple random seeds for $p \in \{0, 1\}$, we compare Matérn with both GAMMA and GAMMA–GL surrogate under different training fractions. We evaluate predictive performance using RMSE, NLPD, Coverage95, and Width95. Results for both orders and all training fractions are reported in Table 2.

We also assess how accurately the surrogate recovers the parameters of the GAMMA as the number of surrogate factors J increases. Figure 5 shows that the estimated parameters approach those of the exact model as J increases while Figure 6 demonstrates that as we start to move away from the origin, the difference between the GAMMA and GAMMA-GL first increases and then decreases. As shown in Figure 6, for $p = 1$, the discrepancy between the exact and surrogate kernels is nearly negligible. For a controlled comparison, the surrogate and exact models use identical parameter initializations for each dataset and optimization run.

H.1.2 REAL WORLD TIME-SERIES

In this section, instead of computing each metric only on the full test suffix $\mathcal{D}_{\text{test}}$, we report it on cumulative prefixes of the test set. If the cumulative test fractions are $(q_1, q_2, \ldots, q_m)$, then for each q_j we evaluate on $\mathcal{D}_{\text{test}}^{(q_j)} = \{y_{n_{\text{train}}+1}, \ldots, y_{n_{\text{train}}+\lfloor q_j(n-n_{\text{train}})\rfloor}\}$. We then evaluated performance on cumulative test portions with fractions $q \in \{0.10, 0.15, 0.20, \ldots, 0.95, 1.00\}$. This setup allows us to examine how predictive performance evolves as increasingly longer parts of the forecast horizon are incorporated into the evaluation.

UCI energy dataset. The data used in this study were derived from the Appliances Energy Prediction dataset introduced in (Candanedo et al., 2017) and made publicly available through the UCI Machine Learning Repository. The original dataset contains measurements collected from a low-energy residential building over approximately 4.5 months. The measurements include appliance

Table 2: Comparison of Matérn and Gamma models for $p \in \{0, 1\}$.

p	Train frac.	Model	RMSE	NLPD	Coverage95	Width95
0	50%	Matérn	**0.983 ± 0.319**	1.899 ± 1.475	0.784 ± 0.220	**2.590 ± 0.581**
		Gamma	1.015 ± 0.353	**1.724 ± 1.090**	**0.822 ± 0.219**	2.950 ± 0.563
		GAMMA–GL	1.021 ± 0.356	**1.730 ± 1.086**	**0.821 ± 0.220**	2.964 ± 0.565
	70%	Matérn	1.012 ± 0.319	1.593 ± 0.513	0.803 ± 0.142	**2.674 ± 0.416**
		Gamma	**0.940 ± 0.297**	**1.362 ± 0.386**	**0.921 ± 0.109**	3.000 ± 0.394
		GAMMA–GL	**0.935 ± 0.298**	**1.352 ± 0.382**	**0.923 ± 0.109**	3.018 ± 0.400
	80%	Matérn	0.975 ± 0.361	1.576 ± 0.811	0.854 ± 0.206	**2.678 ± 0.409**
		Gamma	**0.879 ± 0.359**	**1.371 ± 0.730**	**0.910 ± 0.210**	2.852 ± 0.361
		GAMMA–GL	**0.872 ± 0.365**	**1.361 ± 0.738**	**0.909 ± 0.213**	2.858 ± 0.368
1	50%	Matérn	0.932 ± 0.431	2.203 ± 1.600	0.684 ± 0.235	**1.977 ± 0.518**
		Gamma	**0.830 ± 0.375**	**1.335 ± 0.688**	**0.873 ± 0.152**	2.572 ± 0.581
		GAMMA–GL	**0.834 ± 0.370**	**1.342 ± 0.680**	**0.868 ± 0.155**	2.570 ± 0.579
	70%	Matérn	0.840 ± 0.465	1.830 ± 1.640	0.759 ± 0.275	**2.075 ± 0.551**
		Gamma	**0.735 ± 0.380**	**1.233 ± 0.762**	**0.842 ± 0.204**	2.385 ± 0.476
		GAMMA–GL	**0.740 ± 0.375**	**1.234 ± 0.740**	**0.836 ± 0.203**	2.372 ± 0.467
	80%	Matérn	0.751 ± 0.490	1.559 ± 1.569	0.765 ± 0.304	**2.061 ± 0.538**
		Gamma	**0.612 ± 0.388**	**0.981 ± 0.898**	**0.897 ± 0.195**	2.197 ± 0.383
		GAMMA–GL	**0.617 ± 0.381**	**0.983 ± 0.855**	**0.896 ± 0.195**	2.202 ± 0.386

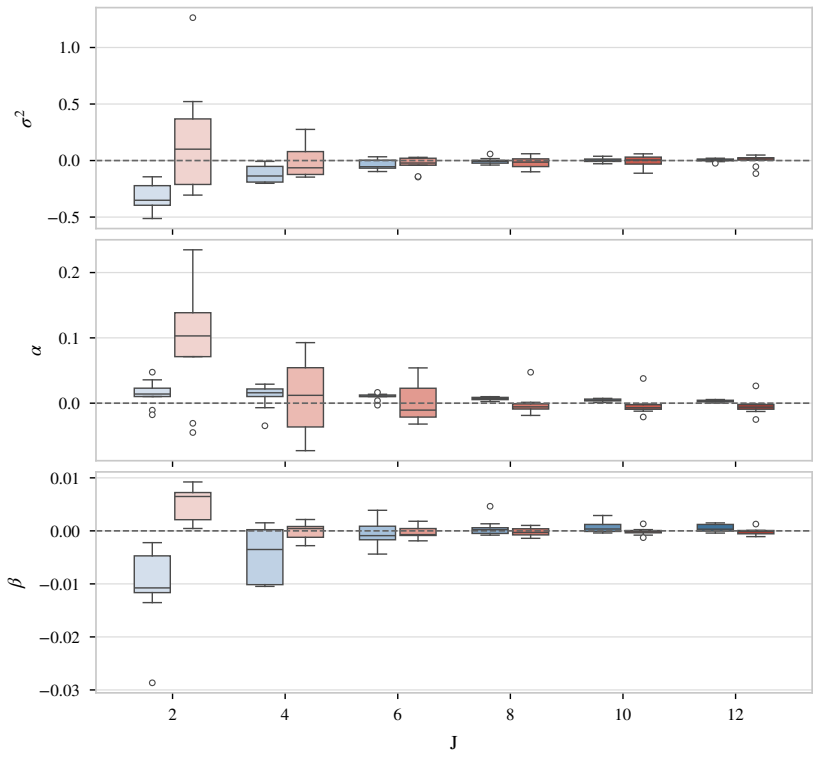


Figure 5: Parameter estimation error across truncation levels J for the Gamma-Laguerre-Markov-GP model. Boxplots show signed differences between surrogate and exact parameter estimates over seeds, with blue for $p = 0$ and red for $p = 1$.

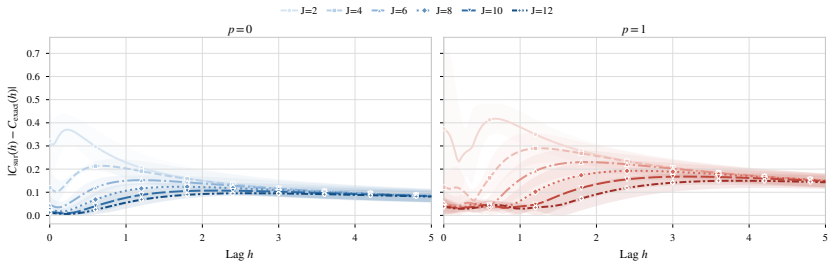


Figure 6: Absolute covariance-function error for the Gamma-Laguerre surrogate across truncation levels J. For each $p \in \{0, 1\}$, the figure shows the mean absolute difference $|C_{\mathrm{sur}}(h) - C_{\mathrm{exact}}(h)|$ over random seeds as a function of lag $h \in [0, 2]$, with shaded bands indicating one standard deviation. Blue curves correspond to $p = 0$ and red curves to $p = 1$.

energy consumption, indoor temperature and relative humidity in different rooms, outdoor weather conditions, and additional environmental variables. The original data consist of 19,735 observations recorded at 10-minute intervals. UCI Appliances Energy Prediction dataset is a widely used real-world regression and time-series benchmark containing approximately four and a half months of measurements recorded at 10-minute intervals. In this work, we use the indoor temperature series T_1. According to the original dataset description, T_1 represents the temperature measured in the kitchen area of the building, expressed in degrees Celsius. Therefore, the simulations presented in this work concern the temporal evolution of indoor kitchen temperature rather than the appliance-energy-consumption variable itself. The simulation data that we use are uniformly sampled with a sampling period of $T_s = 1$ hour. Thus, the processed dataset contains 24 observations per day, resulting in $N = 3290$ samples used in this work.

Paired Assets and Pepsi–Coca-Cola. Cointegration provides the theoretical foundation for pairs trading. Engle and Granger (Engle & Granger, 1987) showed that if two non-stationary $I(1)$ assets, x_t and y_t, are cointegrated, then there exists a coefficient β such that the linear combination

$$z_t = y_t - \beta x_t$$

is stationary. This idea is central to pairs trading, where the spread between two economically related assets is expected to revert toward an equilibrium level (Gatev et al., 2006). Building on this perspective, Elliott et al. (2005) model the spread between paired assets using a mean-reverting Gaussian Markov process and use the resulting dynamics to guide trading decisions. Their approach introduces a state-space formulation for estimating the latent mean-reverting spread and the parameters governing its dynamics, including the strength of mean reversion. These estimates are updated as new observations become available and are used to identify potential trading opportunities in pairs trading.

Economic asset prices may be influenced by forces operating over multiple time scales, with short-term deviations reflecting transient market disturbances and longer-term movements reflecting changes in underlying equilibrium conditions, analogous to the short-term and long-term decomposition proposed by Schwartz & Smith (2000). In our application, we consider PepsiCo and Coca-Cola as a representative pair of economically related firms operating in the same industry. Their prices may be affected by common long-term market and macroeconomic factors while also exhibiting shorter-term idiosyncratic deviations, making the pair a natural example for studying dynamics across multiple time scales. To examine these dynamics empirically, we use daily historical market data retrieved from Yahoo Finance. The Pepsi–Coca-Cola pair consists of PepsiCo, Inc. (PEP) and The Coca-Cola Company (KO), with observations extending from February 1990, to May 2026. The two companies operate in the same consumer-beverage industry and provide a long multi-decade history over which their relative price dynamics can exhibit both short-term deviations and more persistent adjustments. This pair is therefore used in the main experiments to evaluate the ability of the proposed Gamma kernel to model mean-reverting dependence across multiple temporal scales.

DTWEXBGS (Nominal Broad U.S. Dollar Index). We consider the Nominal Broad U.S. Dollar Index (FRED series DTWEXBGS), published by the Board of Governors of the Federal Reserve System and obtained from the Federal Reserve Economic Data (FRED) database from January 2006 to May 2026.

DTWEXBGS is used in this study as a proxy for the broad external value of the U.S. dollar. The index measures the value of the U.S. dollar against a basket of six major currencies: the euro, Japanese yen, British pound, Canadian dollar, Swedish krona, and Swiss franc. The dollar-index series exhibits persistent cyclical fluctuations, with extended periods of appreciation and depreciation as well as repeated movements toward previously visited levels. As a result, The DTWEXBGS series can be modeled using a mean-reverting stochastic process when the objective is to capture deviations of the dollar index from a long-run equilibrium level. We use the logarithm of the DTWEXBGS level. This specification is appropriate when movements in the dollar index are assumed to be partially transitory, so that shocks to X_t gradually dissipate over time and the process tends to move back toward its equilibrium level. (Vasicek, 1977)

We preprocess the data by first removing invalid or missing date and index-value entries and sorting the remaining observations chronologically. The series is then placed on a daily time grid, while dates with no available observation, such as weekends and holidays, are discarded rather than interpolated. We next apply a natural-log transformation to the index values and use the resulting log-index directly as the target variable, without differencing, demeaning, or standardization. Finally, the observation dates are converted into a continuous time variable measured in years from the first sample using 365.25 days per year, so that the actual calendar-time gaps between consecutive observations are retained in the Gaussian process inputs.

H.2 Representation Learning

We first provide additional details on the construction of the long-memory Rotating MNIST dataset and the experimental protocol used for representation learning. We then motivate the use of the complementary RI-RMSE and content accuracy metrics and describe how they are computed, in addition to the RMSE and NLPD measures reported in the main text. Finally, we present the results for the scattered-missingness setting and discuss how they differ from the contiguous-block case, where long-range temporal propagation plays a more prominent role.

H.2.1 Long-memory Rotating MNIST

Data generation. To construct a controlled high-dimensional benchmark with known long-range temporal dependence, we introduce a long-memory variant of Rotating MNIST in which the rotational dynamics, rather than the absolute orientation, are driven by a stochastic long-memory process. For each MNIST digit, we sample a length-T latent trajectory

$$\mathbf{x} = (x_1, \dots, x_T)^\top \sim \mathcal{N}(\mathbf{0}, K_{\mathrm{GAMMA}}),$$

where K_{GAMMA} is the exact GAMMA covariance with $p = 1$. We set $\alpha = 0.5$ and choose β such that $\mathbb{E}[\lambda] = \alpha/\beta = \sqrt{3}/40$.

The latent process modulates the angular velocity according to

$$\omega_t = \omega_0 + s_\omega x_t, \qquad \theta_{t+1} = \theta_t + \omega_t,$$

with $\omega_0 = 7.2^\circ$ per frame and $s_\omega = 1.5^\circ$ per frame. The initial orientation θ_1 is sampled uniformly on the circle. Each frame is generated by rotating the same original MNIST image to orientation θ_t. Importantly, every frame is rendered directly from the original image rather than recursively rotating the previous rasterized frame, thereby avoiding the accumulation of interpolation artifacts. We use sequences of length $T = 100$, corresponding to approximately two complete rotations on average, while the slowly varying angular velocity inherits the polynomially decaying dependence of the underlying GAMMA process.

Training, model selection, and test protocol. We construct disjoint training, validation, and test sets containing $4{,}000$, $1{,}000$, and $1{,}000$ sequences, respectively. Following the missing-frame protocol of MGPVAE (Zhu et al., 2023), during training we randomly withhold 60% of the frames

from each sequence, with the missing locations resampled across epochs. Withheld frames are neither provided to the encoder nor included in the reconstruction likelihood, so both GGPVAE and MGPVAE are trained using only the observed frames.

For model selection, we evaluate checkpoints on the validation set using a single contiguous missing block comprising $10\%, 20\%, \ldots, 90\%$ of the sequence. We select the checkpoint minimizing the equally weighted mean held-out RMSE across these nine gap sizes. At test time, we evaluate both contiguous-block and scattered-missingness regimes over the same missing fractions. For the contiguous-block evaluation, a single consecutive interval is withheld, whereas for the scattered evaluation the same number of frames is removed at dispersed temporal locations. The test data are never used for checkpoint selection.

Reconstruction quality is evaluated only on *held-out* frames, i.e., frames removed from the input and therefore unavailable to the model during inference. The progressive increase in contiguous gap length provides a direct stress test of whether the learned temporal prior can propagate information across increasingly long unobserved intervals when the underlying rotational dynamics exhibit genuine long-range dependence.

Evaluation metrics. We report held-out root mean squared error (RMSE) as a strict pixel-space reconstruction metric and negative log predictive density (NLPD) per held-out pixel to assess probabilistic predictive quality. However, pixel-space metrics penalize errors in the absolute orientation of a digit even when its shape and identity are otherwise reconstructed correctly. This issue becomes particularly important for long contiguous missing intervals, where uncertainty in the latent rotational phase can accumulate. We therefore complement RMSE and NLPD with two content-oriented criteria: rotation-invariant RMSE (RI-RMSE) and content digit accuracy.

For a reconstructed held-out frame $\hat{\mathbf{y}}_t$ and its ground-truth target $\mathbf{y}_t$, we define the rotation-invariant squared reconstruction error as

$$d^2_{\mathrm{rot}}\left(\hat{\mathbf{y}}_t, \mathbf{y}_t\right) = \min_{\phi \in \Phi} \left\| R_\phi(\hat{\mathbf{y}}_t) - \mathbf{y}_t \right\|_2^2 ,$$

where R_ϕ denotes an in-plane image rotation. The corresponding rotation-invariant RMSE is

$$\text{RI-RMSE} = \sqrt{\frac{\sum_{t \in \mathcal{M}} d^2_{\mathrm{rot}}\left(\hat{\mathbf{y}}_t, \mathbf{y}_t\right)}{|\mathcal{M}| HW}},$$

where $\mathcal{M}$ is the set of held-out frames and $H \times W$ is the image resolution. In our implementation, the minimization is approximated by a global search with 15° spacing followed by a local refinement with 2.5° spacing within $\pm 10^\circ$ of the best coarse orientation. RI-RMSE therefore measures reconstruction fidelity after factoring out global orientation and should not be interpreted as an angle-recovery metric.

As a second complementary criterion, we measure preservation of digit identity using an auxiliary convolutional digit classifier. The classifier is trained only on clean Rotating-MNIST training frames and their digit labels, selected using the clean validation split, and then frozen throughout the VAE evaluation. Importantly, the evaluator is not used to train either GGPVAE or MGPVAE and therefore cannot influence their learned representations.

For each reconstructed held-out frame, we evaluate whether the frozen classifier predicts the correct digit identity. To avoid attributing errors of the auxiliary evaluator itself to the generative models, the primary content-accuracy statistic is computed only on frames for which the evaluator correctly classifies the clean target. Let

$$\mathcal{C} = \{ t \in \mathcal{M} : c(\mathbf{y}_t) = \ell_t \} ,$$

where $c(\cdot)$ denotes the frozen evaluator and ℓ_t is the true digit label. We report

$$\text{ContentAcc} = \frac{1}{|\mathcal{C}|} \sum_{t \in \mathcal{C}} \mathbf{1}\left[c(\hat{\mathbf{y}}_t) = \ell_t \right] .$$

Thus, RMSE and NLPD evaluate exact predictive performance, RI-RMSE evaluates image fidelity modulo global rotation, and ContentAcc evaluates preservation of digit identity.

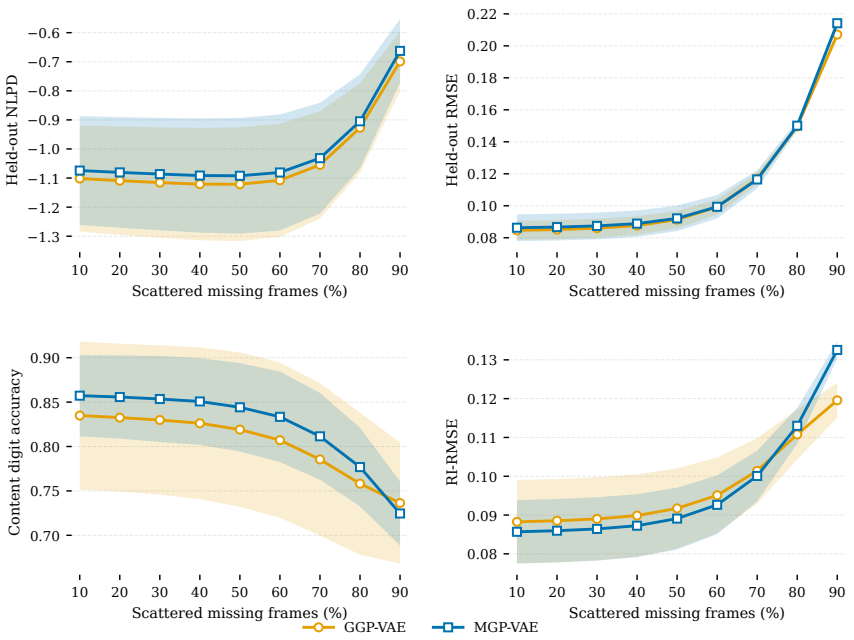


Figure 7: GGPVAE and MGPVAE under scattered missing frames. Curves show mean performance across seeds with ± 1 standard deviation, where nearby observations remain available throughout the sequence.

Contiguous missing blocks. Figure 3 shows a clear separation between the two temporal priors as the missing interval becomes longer. For the shortest 10% block, MGPVAE is slightly better on several metrics, indicating that the GAMMA–GL prior provides little benefit when the interpolation interval is short. The behavior changes once the missing block becomes moderately long. From approximately 30% onward, GGPVAE achieves consistently lower held-out RMSE and better NLPD, and the difference persists as the block grows.

The distinction is substantially stronger under the rotation-insensitive criteria. While strict RMSE remains affected by accumulated rotational phase error, GGPVAE maintains much lower RI-RMSE over long missing blocks. Its content digit accuracy also remains substantially higher, whereas MGPVAE accuracy decreases steadily as the unobserved interval increases. At the largest missing fractions, GGPVAE retains a content accuracy of roughly 0.67, compared with approximately 0.53 for MGPVAE, while its RI-RMSE remains markedly lower. These results indicate that the advantage of the GAMMA prior is not explained only by pixel-level alignment: under long-range interpolation, GGPVAE better preserves the underlying digit structure even when the exact rotational phase is difficult to extrapolate.

Scattered missing frames. The scattered-missing setting in Figure 7 exhibits qualitatively different behavior. Because observed frames remain distributed throughout the sequence, both models have nearby temporal evidence for most missing locations. Consequently, the differences in held-out RMSE and NLPD are much smaller than in the contiguous-block experiment, and both models remain stable until the missing fraction becomes very large.

The content-oriented metrics lead to the same conclusion. MGPVAE achieves somewhat higher content digit accuracy and slightly lower RI-RMSE over much of the low-to-moderate missingness range, while the differences shrink as the missing fraction increases. In particular, RI-RMSE remains close between the two models until the observation density becomes very low. This behavior contrasts with the contiguous-block setting, where a clear advantage for GGPVAE emerges as the temporal distance from observed frames increases.

Taken together, the two masking regimes isolate the role of long-range temporal dependence. Under scattered missingness, local observations are usually available, and a short-memory Matérn prior is

sufficient to interpolate the latent trajectory accurately. In contrast, a long contiguous block requires the model to propagate information across a genuinely long unobserved interval. It is precisely in this regime that the GAMMA prior yields its largest improvement, particularly in RI-RMSE and content preservation. The complementary metrics therefore support the same conclusion as the strict predictive metrics while clarifying an important limitation of pixel-space RMSE: a reconstruction may preserve the correct digit morphology while incurring a large pixel error because its absolute orientation is incorrect.

Implementation details. For Rotating MNIST, both GGPVAE and MGPVAE use the same convolutional encoder and decoder and a 16-dimensional latent representation. The encoder contains two convolutional layers with ReLU activations, while the decoder uses a linear projection followed by transposed-convolution layers. Both models use $p = 1$ temporal components; GGPVAE uses a $J = 4$ GAMMA–GL approximation, initialized with $\alpha = 0.7$. Models are trained with Adam, using learning rate 10^{-3} for the neural-network and observation parameters and 10^{-4} for the temporal-prior parameters, with mini-batches of 40 sequences and gradient clipping. The neural networks are evaluated in single precision, while the state-space filtering and smoothing computations are carried out in double precision. GGPVAE and MGPVAE use identical data splits, masking procedures, network initialization, and model-selection protocols.

H.2.2 Spatiotemporal Climate Data

We compare the Matérn and GAMMA models using a paired evaluation protocol. For each random seed, both models are trained and evaluated on exactly the same spatial locations. Candidate locations are sampled from the Guangdong region, $112 < \text{longitude} < 114$ and $20 < \text{latitude} < 24$, with at most 60 locations initially selected. Locations without a corresponding elevation record are discarded, resulting in a seed-dependent number of retained locations S (e.g., $S = 56$ for seed 0). The same locations are used for training, validation, and testing; hence, the experiments evaluate temporal forecasting at observed spatial locations rather than spatial extrapolation.

The temporal axis is divided into training, validation, and test intervals,

$$\mathcal{T}_{\text{train}} = \{0, \dots, 100\}, \qquad \mathcal{T}_{\text{val}} = \{101, \dots, 110\}, \qquad \mathcal{T}_{\text{test}} = \{111, \dots, 140\}.$$

The model is optimized only on the training interval. Validation consists of an open-loop forecast of $Y_{101:110}$ conditioned on $Y_{0:100}$ and is used solely for checkpoint selection. For final evaluation, the selected model is conditioned on all observations available through $t = 110$ and produces a 30-step open-loop forecast, $p(Y_{111:140} \mid Y_{0:110})$. No test observation is assimilated during forecasting.

For one seed, the predictive mean has shape $\widehat{Y} \in \mathbb{R}^{S \times H \times D}$, with $H = 30$ forecast horizons and $D = 8$ meteorological variables. Let $e_{s,h,d} = \widehat{Y}_{s,h,d} - Y_{s,h,d}$ denote the standardized prediction error. We report overall standardized RMSE and MAE,

$$\text{RMSE} = \sqrt{\frac{1}{SHD} \sum_{s,h,d} e_{s,h,d}^2}, \qquad \text{MAE} = \frac{1}{SHD} \sum_{s,h,d} |e_{s,h,d}|,$$

together with horizon-specific versions obtained by averaging only over locations and variables,

$$\text{RMSE}(h) = \sqrt{\frac{1}{SD} \sum_{s,d} e_{s,h,d}^2}, \qquad \text{MAE}(h) = \frac{1}{SD} \sum_{s,d} |e_{s,h,d}|.$$

We additionally report RMSE after transforming predictions and targets back to their original physical units using scaling statistics estimated from the training data.

Predictive uncertainty is estimated from K Monte Carlo decoder samples $\{\widehat{Y}^{(k)}\}_{k=1}^{K}$, with $K = 20$ in our experiments. The predictive mean and marginal variance are estimated as

$$\widehat{Y}_{s,h,d} = \frac{1}{K} \sum_{k=1}^{K} \widehat{Y}_{s,h,d}^{(k)}, \qquad \widehat{\text{Var}}(Y_{s,h,d}) = \text{Var}_k\left(\widehat{Y}_{s,h,d}^{(k)}\right) + \sigma_{\text{obs}}^2.$$

We evaluate probabilistic performance using the negative log predictive density (NLPD), where the predictive distribution is approximated by an equally weighted Monte Carlo mixture of Gaussian

observation models. The likelihood is evaluated jointly across the $D = 8$ output variables and NLPD is then averaged over spatial locations and forecast horizons. Calibration is assessed using empirical 95% predictive-interval coverage based on $\widehat{Y}_{s,h,d} \pm 1.96\,\widehat{\sigma}_{s,h,d}$.

All metrics are first computed separately for each random seed. For model q and horizon h, the reported curve is the mean across R seeds,

$$\overline{M}_q(h) = \frac{1}{R}\sum_{r=1}^{R} M_{q,r}(h),$$

with uncertainty bands given by the corresponding sample standard deviation across seeds. These bands therefore quantify variability arising from random initialization and seed-dependent spatial location sets, rather than Monte Carlo uncertainty within an individual forecast.

Implementation details. For ERA5, each observation contains eight variables and we use an eight-dimensional latent representation. The encoder and decoder are one-hidden-layer MLPs with 16 hidden units and ReLU activations. The spatial covariance is Matérn-$3/2$, and the MGPVAE temporal baseline also uses a Matérn-$3/2$ prior. GGPVAE replaces the temporal prior with a $J = 4$ GAMMA–GL approximation initialized with $\alpha = 0.7$. All parameters are optimized jointly with Adam using learning rate 10^{-3} and gradient clipping. Training uses the complete training set at each optimization step and 20 Monte Carlo samples for the reconstruction term. Computations are performed in double precision, and model selection is based on validation NLPD before evaluation on the held-out forecasting period.

I Additional Experiments

This section presents additional kernel-evaluation results beyond those discussed in the experimental section. We consider further real-world time series from environmental and financial applications, chosen to provide additional settings in which temporal dependence may occur over multiple time scales. Such behavior is particularly relevant for covariance models constructed from superpositions of Ornstein–Uhlenbeck-type components, which can represent both fast local variation and slower persistent adjustments. These experiments provide further evidence on the predictive behavior of the proposed Gamma kernel and the accuracy of its Markovian state-space approximation across a broader range of temporal dynamics.

I.1 Dataset

Mauna Loa CO_2 dataset. We consider the well-known atmospheric CO_2 dataset consisting of concentration measurements (in parts per million) collected at the Mauna Loa Observatory, Hawaii. This dataset has been widely used as a benchmark for Gaussian process models due to its combination of long-term trend, seasonal patterns, and multi-scale variability (Rasmussen & Williams, 2006).

Satellite XCO_2 dataset. NASA's Orbiting Carbon Observatory-2 (OCO-2) mission provides space-based measurements of atmospheric carbon dioxide to improve understanding of its sources and sinks (Cressie, 2018). Ground-based measurements from the Total Carbon Column Observing Network (TCCON) provide a reference for comparisons with satellite measurements of atmospheric column abundances (Wunch et al., 2011).

We evaluate our method using satellite-derived gridded estimates of column-averaged carbon dioxide dry-air mole fraction, denoted by XCO_2 and expressed in parts per million. Specifically, we use NASA's publicly available `OCO2GriddedXCO2` Version 3 daily files (Nguyen et al., 2022). This Level-4 product is constructed by applying local kriging, also known as optimal interpolation, to daily aggregates of bias-corrected OCO-2 retrievals. Our experiment follows the temporal window, spatial domain, and holdout strategy of the OCO-2 application in (Ma et al., 2023), while using a different data product from the Level-3 product employed in that study.

We use the daily files for June 1–16, 2019. Longitudes are normalized to the interval $[-180^\circ, 180^\circ)$, and missing, non-finite, and designated fill values are removed. At each longitude–latitude grid

location, the available valid XCO_2 values are averaged over the selected 16-day period. We restrict the resulting dataset to longitudes between $140°$W and $50°$W and latitudes between $15°$N and $60°$N, including the boundaries. The processed dataset contains 18,744 distinct grid locations globally, of which 1,933 lie within the study region.

To assess predictive performance, we construct two disjoint test sets. The structured missing-by-design (MBD) test set contains all locations within the study region satisfying

$$-120° < \text{longitude} < -100°.$$

This produces a withheld spatial band containing 508 locations. From the remaining 1,425 locations, we randomly select $\lfloor 0.15 \times 1425 \rfloor = 213$ locations without replacement, using random seed 123, to form the random holdout set, which we label missing-at-random (MAR). The remaining 1,212 locations constitute the training set. The MBD set evaluates prediction across a deliberately withheld spatial band, whereas the MAR set evaluates prediction at randomly withheld locations. In both cases, predictive performance is assessed against held-out values of the gridded Level-4 product.

VIX (Volatility Index). The Cboe Volatility Index (VIX) measures the market's expectation of the 30-day forward-looking volatility of the S&P 500 Index, as implied by SPX option prices, and is widely used as an indicator of financial-market uncertainty and investor risk sentiment. Its temporal dynamics combine abrupt short-term spikes with more persistent periods of elevated or depressed volatility, making it a useful test case for covariance models that aim to represent dependence across multiple time scales. In particular, such behavior is naturally compatible with mixtures of mean-reverting components operating at different reversion rates, which motivates its use for evaluating the proposed GAMMA covariance model. We use the daily VIX closing series from January 1990 to May 2026. The data are preprocessed by removing invalid or missing observations and sorting the remaining values chronologically. Available daily observations are retained without interpolation across weekends or holidays, and the natural logarithm of the VIX level is used directly as the target variable. No differencing, demeaning, or standardization is applied. Finally, the observation dates are converted to continuous time measured in years from the first available sample using 365.25 days per year, so that the Gaussian-process input preserves the actual calendar-time spacing between successive observations.

Wind speed data. We consider wind speed measurements used in (Benton et al., 2022), where Gaussian process models are applied to environmental time series with strong temporal dependence and variability. Wind speed data exhibit both short-term fluctuations and longer-term persistence, making them suitable for evaluating multi-scale covariance structures.

Additional Paired-Asset Datasets. In addition to the Pepsi–Coca-Cola pair considered in the main experiments, we evaluate the proposed Gamma kernel on two further paired-asset datasets. These datasets are also constructed from daily historical market data retrieved from Yahoo Finance. The first additional pair consists of Molson Coors Beverage Company (TAP) and Anheuser-Busch InBev (BUD), covering the period from July 2009 to May 2026. This pair extends the evaluation to the alcoholic-beverage sector, where the two companies operate within closely related product markets and are exposed to common industry, demand, and macroeconomic conditions. Their spread therefore provides an additional setting in which deviations between related asset prices may exhibit mean-reverting behavior while responding to both short-term market shocks and slower economic adjustments. The second additional pair consists of Visa Inc. (V) and Mastercard Incorporated (MA), covering the period from March 2008 to May 2026. These firms form a closely related pair within the digital-payments industry and are exposed to many common forces affecting transaction volumes, consumer spending, financial conditions, and the broader payments ecosystem. The pair therefore provides a distinct financial setting in which highly correlated assets may nevertheless exhibit relative deviations and adjustments over multiple temporal scales.

Together with Pepsi–Coca-Cola, these datasets provide a broader cross-section of economically related assets spanning consumer beverages, alcoholic beverages, and digital payment networks. The additional experiments therefore allow us to assess whether the predictive behavior observed for the Gamma kernel extends beyond a single paired-asset example and remains evident across sectors with different economic drivers and temporal characteristics.

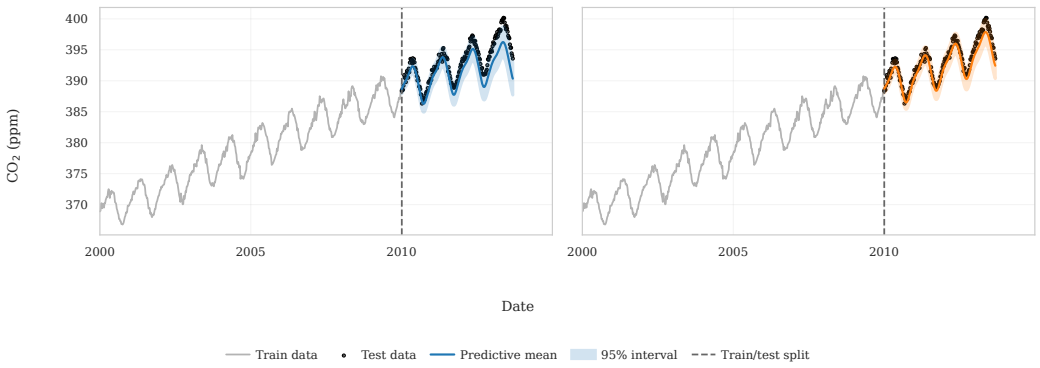


Figure 8: Atmospheric CO_2 concentration measurements from the Mauna Loa dataset ($n = 2227$) together with the 95% predictive region; observations from 1958–2000 are omitted for visual clarity. Following the visualization style of Solin and Särkkä (2014, Fig. 6), the shaded band shows the prediction obtained from the state-space approximation, while the thin solid lines show the corresponding exact GP solution. Left: Model (33) with a Matérn-3/2 component. Right: the same model in which the Matérn-3/2 component is replaced by the GAMMA $p = 1$.

I.2 Results

CO2 Forecasting The CO_2 series contains several qualitatively different sources of dependence: a slow nonstationary-looking increase, a strong annual cycle, and persistent departures from a purely smooth trend. We follow the structure presented by (Solin & Särkkä, 2014b) in which the model is represented through the additive decomposition

$$K(h) = k_1(h) + k_2(h)k_3(h) + k_4(h),$$

where k_1 is a smooth squared-exponential component, k_2k_3 is a damped periodic component, and k_4 is a residual correlated component. Unlike (Solin & Särkkä, 2014b) that assumed k_2 and k_4 as Matérn with $p = 1$, we replace Matérn components with GAMMA kernel while keeping the order $p = 1$. Replacing the Matérn damping and residual terms by Gamma-mixed correlations changes how the likelihood allocates variation across the smooth trend, the seasonal component, and the residual dependence.

The predictive results in Figure 8 shows that the Gamma model yields better long-horizon predictive performance as it maintains persistent dependence in the components that describe seasonal and residual variation. In other words, the Matérn model has a more local dependence structure, and the likelihood may avoid using a dominant smooth component when the remaining Matérn terms cannot adequately correct the resulting long-range covariance mismatches. Thus, the improvement is not simply that the Gamma component is more flexible in isolation; rather, the Gamma replacement leads to a more effective covariance decomposition for the CO_2 series.

CO_2 random-holdout regression Although the proposed kernel is derived from a time-series SDE perspective, it can also be used as a general covariance model for regression over a continuous input domain. To demonstrate this broader applicability, we follow the random train–test evaluation protocol of Chen et al. (2023) on the Mauna Loa CO_2 dataset, in which observations are randomly assigned to training and test sets and predictive performance is evaluated on the held-out observations. Because the split is random rather than chronological, this experiment should be interpreted as a random-holdout regression benchmark rather than a forecasting task. In this setting, the analogue of long-range temporal dependence is slow covariance decay with increasing separation between input locations, allowing observations over a broader portion of the domain to remain informative for prediction. As shown in Figure 9, the GAMMA kernel achieves lower predictive MSE than the corresponding Matérn baselines across the considered training fractions, indicating that its broader dependence structure can also be beneficial in this non-chronological regression setting.

OCO-2 Data Table 3 compares the Matérn and Gamma-mixed kernels on the OCO-2 XCO2 prediction task. The table separates MAR locations, which evaluate interpolation, from MBD locations,

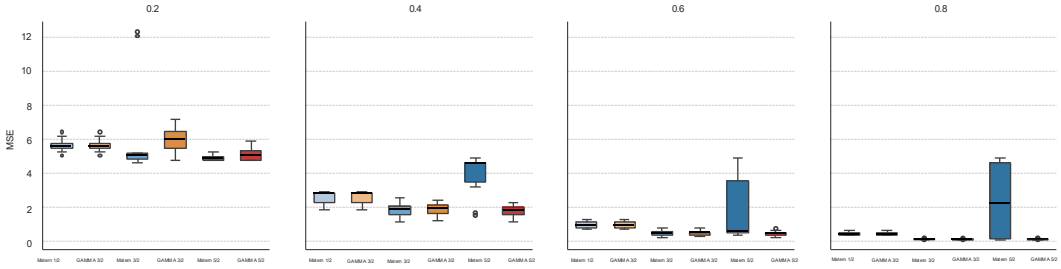


Figure 9: Comparison of Mean Squared Error (MSE) across different kernels.

Table 3: Predictive performance on the OCO-2 XCO2 data. MAR denotes interpolative prediction and MBD denotes extrapolative prediction. Lower RMSE, NLPD, and Width95 are better. For Coverage95, values closer to 0.95 are better. The best value in each column is shown in bold.

		MAR: interpolation				MBD: extrapolation			
Class	Model	RMSE	NLPD	Coverage95	Width95	RMSE	NLPD	Coverage95	Width95
Matérn	$\nu = 0.5$	0.448	0.605	0.934	1.736	0.960	1.363	0.913	3.343
Matérn	$\nu = 1.5$	0.447	0.611	0.915	**1.699**	1.122	1.553	0.874	3.356
GAMMA	$p = 0$	0.449	0.606	**0.939**	1.752	**0.888**	1.289	**0.917**	3.208
GAMMA	$p = 1$	**0.444**	**0.598**	0.911	1.708	0.890	**1.286**	**0.917**	**3.155**

which evaluate extrapolation. As demonstrated by Table 3, for interpolation, the Gamma kernel is competitive with Matérn, but the improvement is small. However, the main difference appears in extrapolation in which Gamma kernel improves extrapolative accuracy while also producing sharper and better calibrated predictive intervals. The matched-smoothness comparison further clarifies the source of the improvement in exploration case. Comparing Matérn $\nu = 1.5$ with Gamma-mixed $p = 1$, the RMSE decreases from 1.122 to 0.890, and the NLPD decreases from 1.553 to 1.286. This supports the intended role of the Gamma mixture: preserving the local half-integer Matérn structure while improving long-range extrapolative behavior through polynomial tail decay.

Other Applications We further compare GAMMA–GL and Matérn models under the $p = 0$ specification, for which the matched Matérn kernel has $\nu = \frac{1}{2}$, on additional environmental and financial time series. Figure 10 reports, from top to bottom, results for the Sioux Falls wind series, the Cboe Volatility Index (VIX), the TAP–BUD asset pair, and the V–MA asset pair. As in Figure 1, the relative advantage of the two covariance models is dataset dependent: neither model dominates uniformly, although GAMMA–GL often provides modest gains when predictive performance benefits from dependence over longer temporal scales. This is consistent with the role of the Gamma mixture, which preserves the local Matérn-$1/2$ behavior while allowing substantially slower covariance decay.

The VIX experiment provides a useful contrasting case. VIX dynamics contain abrupt market-driven spikes together with more persistent periods of elevated volatility. Accordingly, the GAMMA–GL and Matérn results are relatively close, since abrupt shocks are not specifically addressed by changing the covariance tail alone. Nevertheless, GAMMA–GL retains a small advantage in parts of the forecasting horizon, consistent with its ability to exploit the more persistent component of the VIX dynamics. The paired-asset and wind experiments similarly show that the benefit of covariances that can capture memory better is not universal, but becomes most visible when useful dependence persists beyond the single exponential time scale of the matched Matérn model.

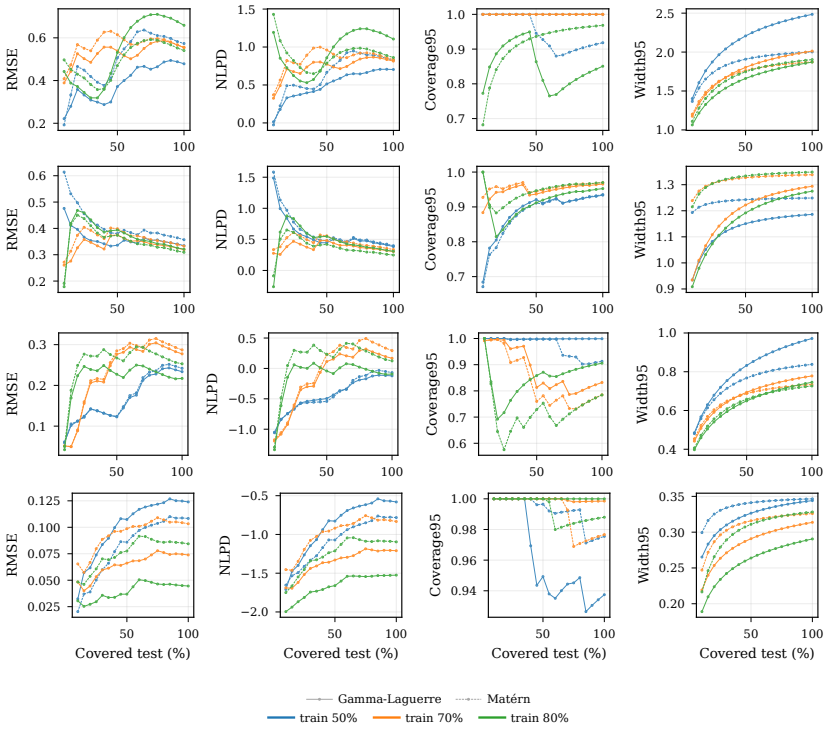


Figure 10: Comparison of GAMMA–GL and Matérn models under the $p = 0$ specification, for which the corresponding Matérn kernel has $\nu = \frac{1}{2}$. From top to bottom, the panels report results for the Sioux Falls wind series, the Cboe Volatility Index (VIX), the TAP–BUD asset pair, and the V–MA asset pair.