FAST NON-STATIONARY GEOSPATIAL MODELLING WITH MULTIRESOLUTION (WAVELET) GAUSSIAN PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

Climate modelling tasks involve assimilating large amounts of geospatial data from different sources, such as simulators and measurements from weather stations and satellites. These sources of data are weighted according to their uncertainty, so good quality uncertainty estimates are essential. Gaussian processes (GPs) offer flexible models with uncertainty estimates, and have a long track record of use in geospatial modelling. However, much of the research effort, including recent work on scalability, is focused on statistically stationary models, which are not suitable for many climatic variables, such as precipitation. Here we propose a novel, scalable, nonstationary GP model based upon discrete wavelets, and evaluate them on toy and real world data.

1 INTRODUCTION

Climate change poses a major challenge in which probabilistic modelling has a lot to offer. The governing physical equations are well-studied, giving a powerful source of prior knowledge, but are typically nonlinear and often chaotic, so that uncertainty in initial conditions, boundary values, or parameters compound rapidly and limit the capability to generalise over both space and time, particularly for resolving on small spatial scales or forecasting. Incorporating data-based modelling methods, the core of modern machine learning techniques, is critical to correct for these mismatches in either model or initialisation.

Reasoning quantitatively about uncertainty is crucial, both for weighting data against prior physical knowledge, and for feeding into downstream applications, for example to optimise sensor placement, or make risk-aware forecasts of renewable energy generation. These methodologies are even more important in regions where measurements are relatively scarce (such as remote mountainous regions), or systems are particularly sensitive to seasonal or climatic variations (such as river systems predominantly fed by glacial melt).

Research in recent years on improving the performance of machine learning models, enhancing their performance at scale, and incorporating prior (physical) knowledge into their training and predictions has been both intensive and fruitful. In particular, spatial or spatiotemperal modelling methods have made rapid recent progress, and there are many performant methods readily available to use, and interest in applying these methods to real-world problems has been growing (Prudden et al., 2021; Lalchand et al., 2022; Ekanayaka et al., 2022; Gahungu et al., 2022; Cuomo et al., 2022; Berlinghieri et al., 2023; Jiang et al., 2024).

Gaussian processes offer a particularly compelling tool for probabilistic regression, with a long history of use in geospatial modelling (Rasmussen & Williams, 2006). Recent work on improving the speed of learning in the conjugate setting (Gaussian measurement noise) has shown considerable potential for scaling up in the low dimensional, geospatial setting (Hensman et al., 2017; Dutordoir et al., 2020; Cunningham et al., 2023; Cheema & Rasmussen, 2023; Wu et al., 2022; Tran et al., 2021; Wilson & Nickisch, 2015).

However, often these methods require the Gaussian process's covariance function to be stationary– that is, invariant to translations. However, many climatic phenomena are highly nonstationary, and

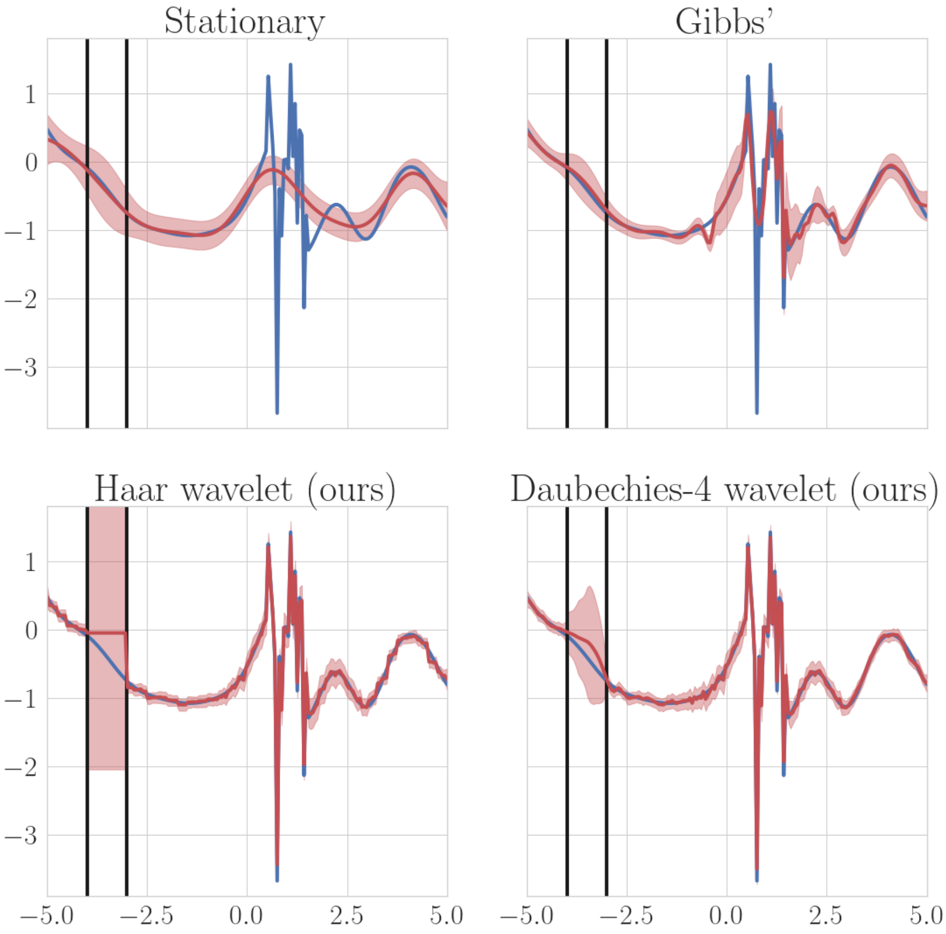

Figure 1: Comparing predictions on 1D toy data. No training data is used within the black lines.

suitable nonstationary covariance functions have not been well-studied (Lalchand et al., 2022; Prudden et al., 2021).

When stationary models are applied to highly non-stationary data, they tend to fail in one of two ways. Either they fit the low frequency trends, and ignore the high frequency parts (Figure 1), or they learn globally low lengthscales to fit the high frequency part, and suffer from rapid reversion to the prior in regions of missing data (Figure 2). Both failure modes lead to problematic poor predictions.

High performing non-stationary GP models exist, but are computationally expensive to train. In this work we propose a GP prior with independent wavelet components. This provides a non-stationary prior, which also has the noteworthy property that fine-scaled behaviour is independent of coarse-scaled behaviour, common in climatic systems (Prudden et al., 2021). Additionally, we provide an efficient learning algorithm comparable to those for stationary covariance functions. The performance approaches that of state of the art non-stationary GPs, but with far better scalability.

2 Gaussian process regression

The conjugate setting for GP regression is the generative model

$$y_n = f(x_n) + \sigma\rho_n \; n \in \{1 : N\} \quad f \sim \mathcal{GP}(0, k) \tag{1}$$

where $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ is a (positive definite, symmetric) covariance function or kernel, $y_n, \rho_n \in \mathbb{R}$, $\sigma \in \mathbb{R}_{\geq 0}$, $x \in \mathbb{R}^D$, and each ρ_n has an independent standard normal distribution.

The posterior $p(f|x_{1:N}, y_{1:N})$ and marginal likelihood $p(y_{1:N})$ are Gaussian and available in closed form (Rasmussen & Williams, 2006). The latter is a principled objective for optimising σ and the kernel's parameters using gradient-based iterative optimisation, but each marginal likelihood evaluation incurs $O(N^3)$ cost due to inversion of the data covariance matrix, which quickly become prohibitive.

Approximate methods such as sparse GP regression (Titsias, 2009; Lázaro-Gredilla & Figueiras-Vidal, 2009) replace the data with a surrogate set of M inducing features, where typically $M \ll N$ without significant loss of performance (Burt et al., 2020b). The training objective is

$$\log \mathcal{N}(y|0,\ K^{*}_{u\mathfrak{f}} K^{-1}_{uu} K_{u\mathfrak{f}} + \sigma^2 I) - \frac{1}{2}\sigma^{-2}\mathrm{tr}(K_{\mathfrak{f}\mathfrak{f}} - K^{*}_{u\mathfrak{f}} K^{-1}_{uu} K_{u\mathfrak{f}}) \tag{2}$$

wherein $K_{\mathfrak{f}\mathfrak{f}}$ is the covariance matrix of f evaluated at the training inputs, K_{uu} is the covariance of the inducing features, and $K_{u\mathfrak{f}}$ is the covariance between the two. By standard manipulations, the expensive log determinant and inverse of the $N \times N$ matrix $(K^{\top}_{u\mathfrak{f}} K^{-1}_{uu} K_{u\mathfrak{f}} + \sigma^2 I)^{-1}$ can be rearranged to those of the $M \times M$ matrices K_{uu} and $(K_{uu} + \sigma^{-2} K_{u\mathfrak{f}} K^{\top}_{u\mathfrak{f}})^{-1}$. Then the dominant cost is $O(NM^2)$ associated with calculating $K_{u\mathfrak{f}} K^{\top}_{u\mathfrak{f}}$.

When N is very large, this $O(NM^2)$ cost per loss function evaluation remains prohibitively expensive. Most strategies for scaling GPs further revolve around mini-batching the data and using a stochastic optimiser (Hensman et al., 2015).

However, recently, authors have proposed precomputable methods, where $K_{u\mathfrak{f}} K^{\top}_{u\mathfrak{f}}$ does not depend on the parameters, and so does not need to be updated during optimisation. This reduces the computational cost to $O(N + M^3)$ in general–where the N is for calculating tr($K_{\mathfrak{f}\mathfrak{f}}$) and the M^3 for the inverse and log determinant (Hensman et al., 2017; Dutordoir et al., 2020; Burt et al., 2020a; Cunningham et al., 2023; Cheema & Rasmussen, 2023). All of these works have been limited to stationary priors, where k is translation invariant, and low dimensional settings. The advantage of these methods is that each optimisation step becomes cheap, without mini-batching, leading to faster optimisation overall.

Non-stationary GPs are less well-explored, but usually involve a hierarchical construction (Heinonen et al., 2016). The state of the art solutions are Gibbs' kernel–using a GP to model spatial variations in the lengthscale of a standard stationary kernel (Gibbs, 1998; Paciorek & Schervish, 2003), and deep GPs–using GPs to warp the input of a standard stationary GP (Damianou & Lawrence, 2013; Salimbeni & Deisenroth, 2017). Both of these are challenging and computationally expensive to train, though we find that Gibbs' kernel in particular performs well for sufficiently small datasets. We seek to provide a more scalable alternative to these methods.

3 Multiresolution (wavelet) Gaussian processes

Discrete wavelets $\{w_{j\ell}\}_{j\in\mathbb{Z},\ell\in\mathbb{Z}}$ form an orthonormal basis of $L^2(\mathbb{R})$. These have been used to preprocess data for GP modelling (Ferkous et al., 2021); here we will use them to construct the GP. Wavelets at scale j have width scaled by 2^{-j}; the second index is a shift by integer multiples of 2^{-j}. The span of coarser wavelets of scale less than ι, say, has an orthonormal basis given by the scaling function $\{v_{\iota\ell}\}_{\ell\in\mathbb{Z}}$ (Mallat, 2009). We can define a suitable covariance function by

$$k(x, x') = \sum_{\ell} \alpha_\ell v_{\iota\ell}(x) v_{\iota\ell}(x') + \sum_{j=\iota}^{\infty} \sum_{\ell} \beta_{j\ell} w_{j\ell}(x) w_{j\ell}(x') \tag{3}$$

which implies that the scale components of f (the inner product $\langle f, w_{j\ell}\rangle$) are independent. Then we define the inducing features by computing the normalised inner product

$$u_{j'\ell} = \langle f, w_{j'+\iota,\ell}\rangle / \beta_{j'+\iota} \tag{4}$$

for $j' > 0$, and replacing w, β with v, α at $j' = 0$. We use a finite range of ℓ which covers the region with training data. Then the elements of $K_{u\mathfrak{f}}$ are calculated by applying the same inner product to k. We focus on $j' > 0$; the case for $j' = 0$ can be calculated with straightforward adaptations. We index the features, and the matrices $K_{u\mathfrak{f}}, K_{uu}$ with the tuple (j', ℓ).

$$[K_{u\mathfrak{f}}]_{(j',\ell),n} = \langle k(\cdot, x_n), w_{j'+\iota,\ell}\rangle / \beta_{j'+\iota} = w_{j'+\iota,\ell}(x_n) \tag{5}$$

Table 1: Precipitation evaluation results: mean and standard error over five runs

TEST		STATIONARY	DEEP GP	GIBBS'	WAVELET–HAAR	WAVELET–DB4
Uniform	RMSE	21.19 (0.87)	58.09 (1.91)	23.52 (1.33)	26.03 (0.72)	22.27 (0.78)
	NLPD	4.484 (0.04)	5.491 (0.04)	4.07 (0.016)	4.60 (0.02)	4.47 (0.03)
Patch	RMSE	21.38 (0.00)	49.64 (0.05)	7.606 (0.135)	18.15 (0.88)	8.287 (0.282)
	NLPD	4.69 (0.000)	5.34 (0.001)	3.734 (0.043)	4.430 (0.025)	4.212 (0.013)

It contains only point evaluations of the wavelet and scaling functions, so does not depend on the parameters. This is the key property which ensures the cost is reduced to $O(N + M^3)$. Similarly, the elements of K_{uu} are calculated by an additional inner product.

$$\begin{aligned} [K_{uu}]_{(j',\ell),(j'',\ell')} &= \langle\langle k(\cdot,\star), w_{j'+\iota,\ell}\rangle/\beta_{j'+\iota}, w_{j''+\iota,\ell'}\rangle/\beta_{j'+\iota} && (6) \\ &= \langle w_{j'+\iota,\ell}, w_{j''+\iota,\ell'}\rangle/\beta_{j''+\iota,\ell'} = \delta_{j'-j'',\ell-\ell'}/\beta_{j'+\iota,\ell} && (7) \end{aligned}$$

Additionally, the matrix to be inverted is sparse, since only wavelets whose support overlap are correlated, similarly to the method of Cunningham et al. (2023).

It remains to select the coefficient α, β. We want to localise fine-scaled behaviour, so that we avoid smoothing out high frequency phenomena (Figure 1) or learning excessively low global lengthscales (Figure 2). We set $\alpha_\ell = 1/2$, and set $\beta_{j\ell}$ to be the sum of a global component and spatially localised (in ℓ) components, all decaying with j at learnable rates; see the appendix for precise expressions. We use two different wavelets: Haar, and the smoother Daubechies-4 (db4).

4 EXPERIMENTAL RESULTS

The failure mode of stationary models is clearest when the data contains regions of low and high frequency data, but a large low frequency region has no training data. We construct these patch tests with toy synthetic data (Figure 1) and with real world, highly non-stationary precipitation data (Figure 2 and Table 1). We also test the performance of uniformly randomly selected test points in the real-world case (Table 1). We do not expect to achieve as strong a performance as Gibbs' kernel, since we are using a much simpler model: we aim to significantly improve on the stationary case, with far better computational scaling than with Gibbs' kernel. Further experimental details are in the appendix.

The toy example of Figure 1 shows how stationary models can end up ignoring high frequency behaviour, whereas non-stationary models can match the variable behaviour. In the smooth region without training data, multiresolution GPs suffer from some prior reversion, but a suitable choice of wavelet considerably improves performance.

On real-world precipitation data (Figure 2 and Table 1), the stationary model suffers from significant prior reversion when in the patch setting, due to learning a low lengthscale. The multiresolution models suffer to a far lesser extent, and are more competitive with the state-of-the-art Gibbs' model.

Notably, the db4 model performs comparably well to the stationary model in the uniform case, but copes much better with the patch test, showing that this model is a significant improvement on the stationary setting.

5 CONCLUSIONS

We have introduces a novel GP model based on discrete wavelets, with an efficient inference algorithm. Empirically, when the wavelet is carefully chosen, this model performs far more robustly than standard stationary covariance functions on non-stationary data widespread in climate modelling applications. However, there is significant scope for further improvements and closing the gap with more expensive non-stationary kernels, for example by changing the form of the coefficients.

REFERENCES

Renato Berlinghieri, Brian L Trippe, David R Burt, Ryan Giordano, Kaushik Srinivasan, Tamay Özgökmen, Junfei Xia, and Tamara Broderick. Gaussian processes at the Helm(holtz): A more fluid model for ocean currents, 2023.

David R Burt, Carl Edward Rasmussen, and Mark van der Wilk. Variational orthogonal features, 2020a. URL `https://arxiv.org/abs/2006.13170`.

David R Burt, Carl Edward Rasmussen, and Mark van der Wilk. Convergence of sparse variational inference in Gaussian processes regression. *Journal of Machine Learning Research (JMLR)*, 21 (131):1–63, 2020b. URL `http://jmlr.org/papers/v21/19-1015.html`.

Talay M Cheema and Carl Edward Rasmussen. Integrated variational fourier features for fast spatial modelling with gaussian processes, 2023.

Harry Jake Cunningham, Daniel Augusto de Souza, So Takao, Mark van der Wilk, and Marc Peter Deisenroth. Actually sparse variational gaussian processes. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10395–10408. PMLR, 25–27 Apr 2023. URL `https://proceedings.mlr.press/v206/cunningham23a.html`.

Salvatore Cuomo, Vincenzo Schiano Di Cola, Fabio Giampaolo, Gianluigi Rozza, Maziar Raissi, and Francesco Piccialli. Scientific Machine Learning Through Physics–Informed Neural Networks: Where we are and What's Next. *Journal of Scientific Computing*, 92(3):88, September 2022. ISSN 0885-7474, 1573-7691. doi: 10.1007/s10915-022-01939-z.

Andreas Damianou and Neil D Lawrence. Deep gaussian processes. In *Artificial intelligence and statistics*, pp. 207–215. PMLR, 2013.

Vincent Dutordoir, Nicolas Durrande, and James Hensman. Sparse Gaussian processes with spherical harmonic features. In *37th International Conference on Machine Learning (ICML)*, 2020.

Ayesha Ekanayaka, Peter Kalmus, Emily Kang, and Amy Braverman. Statistical Downscaling of Sea Surface Temperature Projections with a Multivariate Gaussian Process Model. 2022.

Khaled Ferkous, Farouk Chellali, Abdalah Kouzou, and Belgacem Bekkar. Wavelet-Gaussian process regression model for forecasting daily solar radiation in the Saharan climate. *Clean Energy*, 5(2):316–328, 06 2021. ISSN 2515-4230. doi: 10.1093/ce/zkab012. URL `https://doi.org/10.1093/ce/zkab012`.

Paterne Gahungu, Christopher Lanyon, Mauricio A Álvarez, Engineer Bainomugisha, Michael T Smith, and Richard Wilkinson. Adjoint-aided inference of Gaussian process driven differential equations. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17233–17247. Curran Associates, Inc., 2022.

Mark N Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, University of Cambridge, 1998.

Markus Heinonen, Henrik Mannerström, Juho Rousu, Samuel Kaski, and Harri Lähdesmäki. Nonstationary gaussian process regression with hamiltonian monte carlo. In Arthur Gretton and Christian C Robert (eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pp. 732–740, Cadiz, Spain, 09–11 May 2016. PMLR. URL `https://proceedings.mlr.press/v51/heinonen16.html`.

James Hensman, Alexander G de G Matthews, and Zoubin Ghahramani. Scalable variational Gaussian process classification. In *18th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

James Hensman, Nicolas Durrande, and Arno Solin. Variational Fourier features for Gaussian processes. *Journal of Machine Learning Research*, 2017.

Ziyang Jiang, Tongshu Zheng, Yiling Liu, and David Carlson. Incorporating prior knowledge into neural networks through an implicit composite kernel, 2024.

Vidhi Lalchand, Kenza Tazi, Talay M Cheema, Richard E Turner, and Scott Hosking. Kernel learning for explainable climate science, 2022.

Miguel Lázaro-Gredilla and Aníbal Figueiras-Vidal. Inter-domain Gaussian processes for sparse inference using inducing features. In *26th Conference on Neural Information Processing Systems (NeurIPS)*, 2009.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Elsevier, 2009. ISBN 978-0-12-374370-1. doi: 10.1016/B978-0-12-374370-1.X0001-8.

Christopher Paciorek and Mark Schervish. Nonstationary covariance functions for gaussian process regression. In S Thrun, L Saul, and B Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/326a8c055c0d04f5b06544665d8bb3ea-Paper.pdf.

Rachel Prudden, Niall Robinson, Peter Challenor, and Richard Everson. Stochastic Downscaling to Chaotic Weather Regimes using Spatially Conditioned Gaussian Random Fields with Adaptive Covariance. *Weather and Forecasting*, October 2021. ISSN 0882-8156, 1520-0434. doi: 10.1175/WAF-D-20-0217.1.

Carl Edward Rasmussen and Christopher K I Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. ISBN 978-0-262-18253-9.

Hugh Salimbeni and Marc Deisenroth. Doubly stochastic variational inference for deep Gaussian processes. *Advances in neural information processing systems*, 30, 2017.

Michalis Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2009.

Gia-Lac Tran, Dimitrios Milios, Pietro Michiardi, and Maurizio Filippone. Sparse within sparse Gaussian processes using neighbor information. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10369–10378. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/tran21a.html.

Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *32nd International Conference on Machine Learning (ICML)*, 2015.

Luhuan Wu, Geoff Pleiss, and John P Cunningham. Variational nearest neighbor Gaussian process. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24114–24130. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/wu22h.html.

A Appendix

A.1 Coefficients

For any master pair of scaling function v and wavelet function w, let

$$v_{j\ell}(x) = \sqrt{2^j} v\left(\frac{x - 2^{-j}\ell}{2^{-j}}\right) \tag{8}$$

$$w_{j\ell}(x) = \sqrt{2^j} w\left(\frac{x - 2^{-j}\ell}{2^{-j}}\right) \tag{9}$$

so that larger values of j represent finer scaled features. Let ι be the canonical scale, and define the covariance function according to

$$k(x, x') = \sum_\ell \alpha_\ell v_{\iota\ell}(x) v_{\iota\ell}(x') + \sum_{j=\iota}^{\infty} \sum_\ell \beta_{j\ell} w_{j\ell}(x) w_{j\ell}(x') \tag{10}$$

so that finer scaled behaviour than the canonical scale are independent. The definition of the covariance function is completed by specifying a functional form for α, β. For an almost-stationary version, consider

$$\alpha_\ell = 2^{-1} \tag{11}$$

$$\beta_{j\ell} = \frac{2^{-(1+a_0)}}{2^{1+a_0} - 1} 2^{-(1+a_0)(j-\iota+2)} \tag{12}$$

where $a_0 > 0$ is a parameter. The exponent is designed so that β decays as j increases, and the constant is to normalise so that $\alpha_\ell + \sum_j \beta_{j\ell} = 1$. Note that the sum is convergent for all $a_0 > 0$. If the wavelets have disjoint support (as with the Haar wavelet), then this sum is exactly $k(x, x)$; more generally, the calculation for $k(x, x)$ is more complex, but fixing the value of this sum means that a controls only the coarse-scaledness of the covariance function, and not the variance.

The issue with this first covariance function is that it does not allow for learning localised fine-scaled regions. An alternative form for β is

$$\beta_{j\ell} = \frac{1}{Z_{j\ell}} \left[\frac{2^{1+a_0} - 1}{2^{-(1+a_0)}} 2^{-(1+a_0)(j-\iota+2)} + \sum_{q=1}^{Q} b_q \frac{2^{1+a_0} - 1}{2^{-(1+a_0)}} 2^{-(1+a_0)(j-\iota+2)} 2^{-\left(\frac{\ell 2^{-j} - c_q}{s_q}\right)^2} \right] \tag{13}$$

$$Z_{j\ell} = 2 \left[1 + \sum_{q=1}^{Q} b_q 2^{-\left(\frac{\ell 2^{-j} - c_q}{s_q}\right)^2} \right] \tag{14}$$

which has the capacity to add extra fine scaled terms – if $a_q > a_0$ then we have additional fine scaled behaviour within a few multiples of s_q of c_q. The normalisation terms are chosen so that $\alpha_\ell + \sum_j \beta_{j\ell} \approx 1$, so that the local variance is approximately constant.

To extend to higher dimensions, we construct the kernel and the features by taking a product over dimensions (tensor product).

A.2 Additional experimental details

The toy data is generated by sampling from a GP with a squared exponential kernel with lengthscale 0.2, with the input warped through a piecewise linear function defined by first normalising the range of the inputs to $[0, 1]$, then

$$x' = \begin{cases} xc/a & x < a \\ c + (x - a)(d - c)/(b - a) & a \leq x < b \\ d + (x - b) * (1 - d)/(1 - b) & x \geq b \end{cases} \tag{15}$$

for which we use $a = 0.1, b = 0.9, c = 0.55, d = 0.65$. We apply Gaussian measurement noise with variance 0.05. To generate the train/test split, we generated 180 regularly spaced points, and withheld those in $[-4, -3]$ as test points.

For the stationary model, we used a squared exponential kernel, and used SGPR with as many inducing points as training points, initialised at the training points (which should recover exact regression).

For Gibbs' covariance function, we used the same number of inducing points, and these also were also the points for which we estimated the maximum a posteriori values of the lengthscale process. The log lengthscale was modelled with a GP with squared exponential kernel with GPyTorch default parameters, and with zero mean–that is, the log lengthscale mean is zero, so the lengthscale mean is 1.

For the multiresolution GPs, we used a canonical scale of -1, and one component. We set ℓ at each depth to cover shifts that reached a certain distance from the origin. For the Haar wavelet, we used 10, and for db4 we used 20 (since it is a wider wavelet, so wavelet features centred further away from the data are still correlated with the data). The finest scale we used was $j' = 7$.

For the real world data, we used a similar configuration, with the following differences.

For the stationary model, we used a Matérn-5/2 kernel, and used IFF (Cheema & Rasmussen, 2023) with 5700 frequencies.

For the multiresolution covariance functions, we set the width for db4 by padding the width of the data by 5 on each side; for Haar we just used the width of the data. We used a maximum scale of 5, and used three components. For the Haar wavelet, we tried to improve the mean fit by reducing the canonical scale to -3.

For Gibbs' kernel we set the log lengthscale's prior mean to $\log 0.3$ and its lengthscale to 1.5, based on best performance over multiple possible settings. We used 1000 inducing points.

For the deep GP, we used a standard setup from GPflux, with three layers and 3000 inducing points.

For each model, we tried a number of inducing points or frequencies, or a number of maximum scales with multiresolution kernels, and reported the best result.

Across both sets of experiments, we used Adam with 10000 iterations and a learning rate of 0.01 for Gibbs' and the deep GP. Otherwise we used L-BFGS-B.

The precipitation dataset is modelled precipitation normals in mm in the contiguous United States for the month preceding 1 January 2021. It is publicly available with further documentation at `https://water.weather.gov/precip/download.php`, though note the data at the source is in inches.

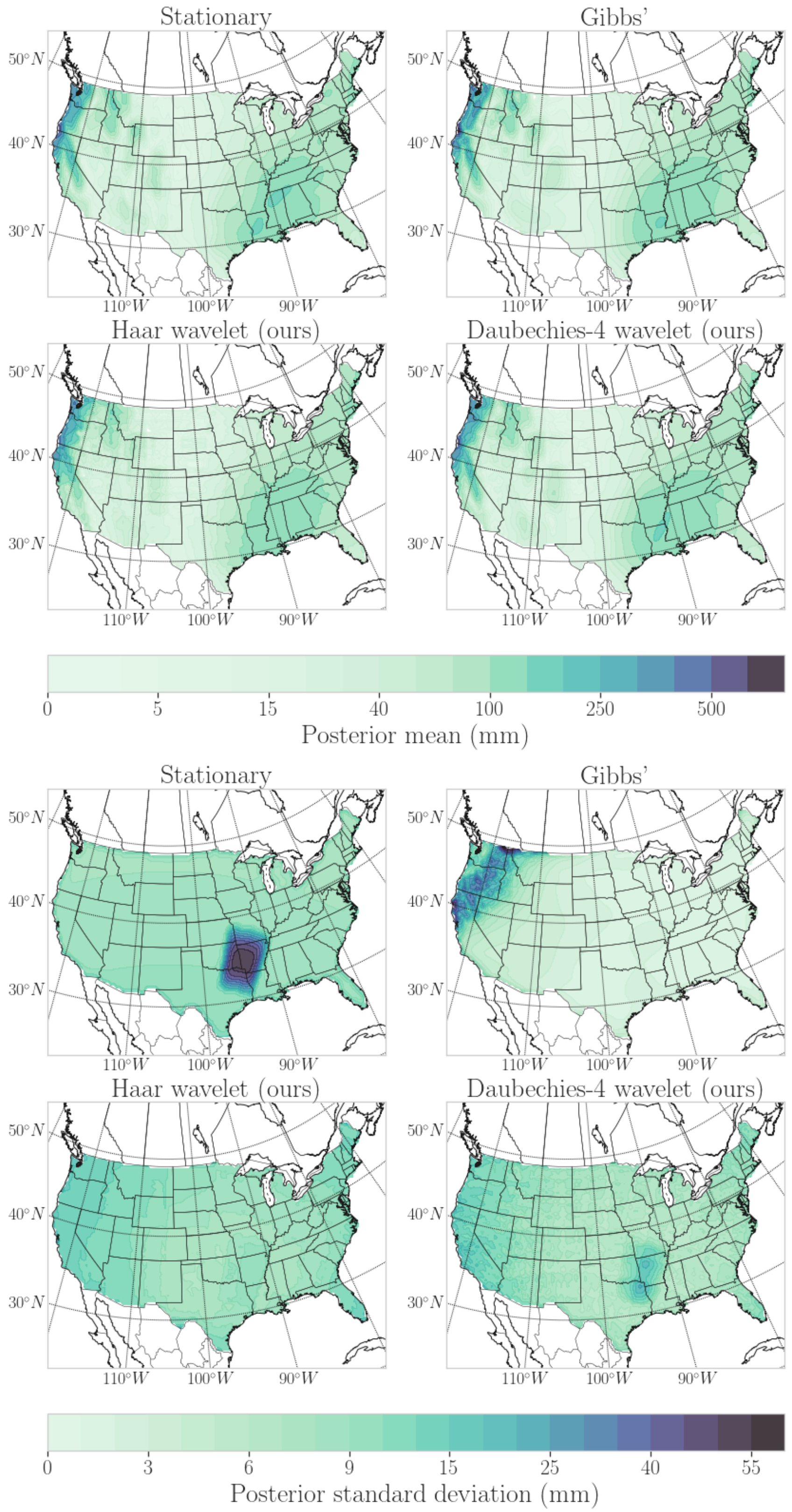

Figure 2: Predictive distribution for the patch test.