Identifying latent climate signals using sparse hierarchical Gaussian processes

Matt Amos*
Maths and Stats
Lancaster University
Lancaster, UK
m.amos1@lancaster.ac.uk

Thomas Pinder*
Maths and Stats
Lancaster University
Lancaster, UK
t.pinder2@lancaster.ac.uk

Paul J. Young
Lancaster Env. Centre
Lancaster University
Lancaster, UK

Abstract

Extracting latent climate signals from multiple climate model simulations is important to estimate future climate change. To tackle this we develop a sparse hierarchical Gaussian process (SHGP), which probabilistically learns a latent distribution from a set of vectors. We use this to predict the latent surface temperature change globally and for central England from an ensemble of climate models, in a scalable manner and with robust uncertainty propagation.

1 Introduction

Climate models simulate the evolving state of the atmosphere, ocean, and land, given certain inputs and according to fundamental physics and chemistry. Individual climate projections, called realisations, from multiple models are routinely combined to estimate the most likely climate projection for given pathways of future emissions. It is a widely used principle that the average of a set of models (climate or statistical) typically produces a better estimate than a single model [19]. It is also possible to approximate the uncertainty of the combined projection from the spread of model realisations. However, which method one should use to average these models to extract the latent climate change signal common between all models, is an open and important question. The most widely used approach is to take a multi-model mean, deriving an estimate of uncertainty from the standard deviation across the models [e.g., 7]. However, this approach does not robustly propagate uncertainties from models and simulations to the final estimate. In order to make important policy relevant decisions we require well characterised uncertainty estimates. Consequently, we develop a Bayesian framework based on a hierarchy of Gaussian processes (GPs).

Gaussian processes A GP is a distribution over functions, defined by a mean function μ[2] and kernel k [18]. Considering functions $f : \mathbb{R}^d \to \mathbb{R}$, we can write

$$p(f(\cdot)) = \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot)) \,. \tag{1}$$

For data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\} = \{\mathbf{x}_n, y_n\}_{n=1}^N$ comprised of inputs $\mathbf{x}_n \in \mathbb{R}^d$ and corresponding outputs $y_n \in \mathbb{R}$, we condition on the function evaluations $\mathbf{f} = [f(\mathbf{x}_n)]_{n=1}^N$ through a factorising likelihood function that we will assume to be Gaussian throughout this work

$$p(\mathbf{y} \,|\, \mathbf{f}) = \prod_{i=n}^{N} \mathcal{N}(y_n \,|\, f_n, \sigma_n^2) \,, \tag{2}$$

*The first two authors contributed equally to this work.

[2]We follow convention and assume this to be zero in this work.

2022 NeurIPS Workshop on Gaussian Processes, Spatiotemporal Modeling, and Decision-making Systems.

where $\sigma_n^2 \in \mathbb{R}_{>0}$ is an observational noise parameter. For conjugate models of this form, evaluating the predictive posterior distribution at a set of test points $\mathbf{X}^\star$ has the analytical form

$$p(f(\mathbf{X}^\star) \,|\, \mathbf{f}, \mathbf{y}) = \mathcal{GP}(\mathbf{K}_{\star\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \sigma_n^2 \mathbf{I}_N)^{-1}\mathbf{y}, \mathbf{K}_{\star\star} - \mathbf{K}_{\star\mathbf{f}}(\mathbf{K}_{\mathbf{ff}} + \sigma_n^2 \mathbf{I}_N)^{-1}\mathbf{K}_{\mathbf{f}\star}) \,, \tag{3}$$

where $\mathbf{K}_{\mathbf{ff}}$ is the matrix formed by evaluating k pairwise on $\mathbf{X}$, and $\mathbf{K}_{\star\star}$ pairwise on $\mathbf{X}^\star$. GPs are widely used within climate modelling [13] as they are capable of producing well characterised uncertainty estimates and, through the choice of kernel function, offer a principled way for experts to posit their prior beliefs into the model. Constructing GP models for climate data with non-trivial supports is an active area of research with recent works considering data defined on a sphere [6], the vertices of a graph [17], and vector fields [10]. However, modelling a set of vectors, such as time-series, in a scalable manner is still an open problem.

In this work we aim to answer the question "*how can we probabilistically infer the latent function from a set of noisy vector-valued realisations*"? To answer this, we compose a set of GPs into a two-layered hierarchical model. In the model's upper layer we learn a GP posterior distribution that represents the latent function. This distribution is then propagated down into the lower layer where we learn a single GP posterior distribution for each observed realisation. To ensure scalable computation, we construct a variational approximation to each GP in the hierarchy by building upon the sparse GP literature. Finally, we use our sparse hierarchical Gaussian process (SHGP) model to infer latent surface temperature changes from an ensemble of climate models.

2 Methodology: From realisations to distributions

We want to model a set of realisations $\mathbf{Y} = \{\mathbf{y}_r\}_{r=1}^R$ where $\mathbf{y}_r$ is a vector in $\mathbb{R}^N$. All $\mathbf{y}_r$ share a latent function $\hat{\mathbf{y}}$ and are defined across the same index set $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^N$. In our examples, these data $\mathcal{D} = \{\mathbf{X}, \mathbf{y}_r\}_{r=1}^R$ are a collection of surface temperature projections across a common temporal range, simulated by multiple climate models, where we wish to extract the latent climate signal.

To model these data we use a Bayesian hierarchical model of GPs

$$g(\mathbf{X}) \sim \mathcal{GP}(\mathbf{0}, k_g(\mathbf{X}, \mathbf{X}')), \tag{4}$$

$$f^{(r)}(\mathbf{X}) \sim \mathcal{GP}(\boldsymbol{\mu}_{g\,|\,\mathbf{y}}, k^{(r)}(\mathbf{X}, \mathbf{X}')) \,, \tag{5}$$

where $g(\mathbf{X})$ is the group-level prior distribution that underpins all $\mathbf{y}_r$. $f^{(r)}(\mathbf{X})$ is the prior distribution for an individual vector and has a mean given by the posterior mean of g that we denote $\boldsymbol{\mu}_{g\,|\,\mathbf{y}} = \mathbb{E}_{p(g\,|\,\mathbf{y})}[g(\mathbf{X})]$. The kernel k_g describes the covariance structure of the latent function and the set of kernels $\{k^{(r)}\}_{r=1}^R$, which have a one-to-one mapping with each $f^{(r)}$, describe the covariance structure of an individual realisation. Different kernel functions may be used for any of the kernel terms. This hierarchical model is linear, which means that for a single vector $f^{(r)}(\mathbf{X}) \sim \mathcal{GP}(\mathbf{0}, k_g(\mathbf{X}, \mathbf{X}') + k^{(r)}(\mathbf{X}, \mathbf{X}'))$, whereas the joint distribution between different time series is described by the covariance function $k_g(\mathbf{X}, \mathbf{X}')$. This model can be extended to multiple hierarchies [e.g., 5] but is computationally limited in its scalability; $\mathcal{O}(N^3)$ for data size as per standard GP regression and $\mathcal{O}(R^2)$ for the number of realisations.

To ameliorate the computational cost associated with fitting models of this form, we approximate each GP in Equations (4)–(5) with a low-rank (*sparse*) GP. To form such an approximation, we introduce a set of *inducing points* $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$ where $M \ll N$. We assume that the inducing points exist on the same support as our observed inputs $\mathbf{X}$, but they need not be a subset of $\mathbf{X}$ [21, 22]. Further, the set of inducing points used is shared across all GPs, as acknowledged by the absence of a subscript on $\mathbf{Z}$.

Our low-rank approximations are themselves GPs and therefore have an associated set of kernel and likelihood parameters. We collectively denote these as $\boldsymbol{\theta}$. Optimising the marginal log-likelihood of this model is intractable, so we instead use variational inference (VI) to form an evidence lower bound (ELBO) [8]. The ELBO lower bounds the marginal log-likelihood and maximising its value is analogous to minimising the Kullback-Leibler divergence from the approximate low-rank processes to the true model. To achieve this, we optimise the model's hyperparameters and the inducing points' values using a first-order gradient-based optimiser [e.g., 2]. The cost of inference now scales linearly in N and quadratically in M, giving a total cost of $\mathcal{O}(NM^2R^2)$. We provide a full derivation of the ELBO in Section A, however, similar results have been derived in the literature for GP regression [22], latent variable GPs [16], and heteroscedastic GP models [11, 20].

3 Experiments and results

A common and important problem in climate science is estimating the most likely climate projection and the associated uncertainty from large climate model ensembles, such as the Coupled Model Intercomparison Project Phase 6 (CMIP6) [4]. CMIP6 is the most recent coordinated effort by climate modelling centres, to simulate the Earth's system and to project the evolving climate. Simulations are produced to a set experimental protocol, where the model *forcings* such as greenhouse gas emissions, are consistent across all models. Rather than producing one simulation, models often produce multiple *realisations* across which the underlying climate signal should be consistent. The realisations themselves will be individually modulated by meteorological and sub-decadal variability. Given that this natural hierarchy exists in the realisations, the SHGP is a suitable model.

In the following experiments, we take realisations from multiple CMIP6 climate models and use the SHGP to model surface temperature projections in order to uncover the latent climate signal. An implementation of our work is provided at `https://github.com/mattramos/SparseHGP` using GPFlow [14] and GPJax [15]. We empirically validate the SHGP in Section B.

Historic daily surface temperature The most simulated CMIP6 scenario is the historical one, spanning the years 1850–2015. This scenario, where models simulate from a range of predetermined fields, provides a useful validation against observations. In this experiment we model daily surface temperatures over central England, extracted from 5 realisations of the HadGEM3-GC3.1 climate model [1], using the SHGP. $\mathbf{Y} = \{\mathbf{y}_r\}_{r=1}^{R=5}$ is the set of HadGEM3-GC3.1 realisations. As these realisations have a strong annual oscillation we augment $\mathbf{t}$ with $\cos(2\pi(\text{day of year})/360)$ and $\sin(2\pi(\text{day of year})/360)$, to aid the SHGP in fitting a seasonal cycle without the need for periodic kernels. Note that 360 is used in the denominator as this climate model runs on a 360 day calendar.

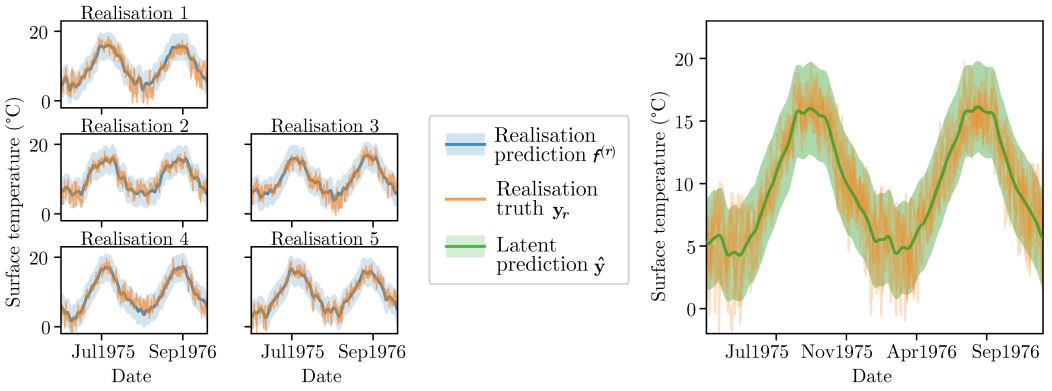

Figure 1: Central England surface temperature as modelled by the SHGP. The five plots on the left show the individual prediction for each of the five realisations, whereas the right plot shows the latent prediction. Shading across all plots denotes 2 standard deviations or approximately a 95 % credible interval.

The SHGP was fit using 1000 inducing points with a second order Matérn kernel [12] for the years 1975–2015 (72000 datapoints). Figure 1 contains a temporal snapshot which shows how the SHGP models both the individual realisations and the latent function underlying the realisations. The latent function (green) captures the overall seasonal oscillation that is present across all the realisations, whilst the individual realisation predictions additionally include the influence of meteorological noise. We can interpret the latent posterior mean as representing the latent surface temperature from the HadGEM3-GC3.1 model, and the variance as encompassing the spread of the individual realisations.

As the SHGP jointly models the realisations and the latent function, we are able to propagate the uncertainty from the realisations to an estimate of surface temperature over central England. Additionally, it enables us to generate samples of surface temperature. Regressing over this large data is only possible because of the implementation of sparse methods.

Future global climate projections We model annual global mean surface temperature (GMST) from 9 CMIP6 models from the SSP2-4.5 (shared socio-economic pathway) scenario. This is a *middle*

of the road scenario of the future, which estimates anthropogenic emissions based on continued growth and little progress towards sustainability goals, with 4.5 Wm^{-2} radiative forcing at the end of the century. We use 3 realisations from each model such that $\mathbf{Y} = \{\mathbf{y}_r\}_{r=1}^{R=27}$ is the set of GMST projections for the years 2016–2100. The index set $\mathbf{t}$ is a vector of the years. The SHGP was fit using 40 inducing points with a second order Matérn kernel.

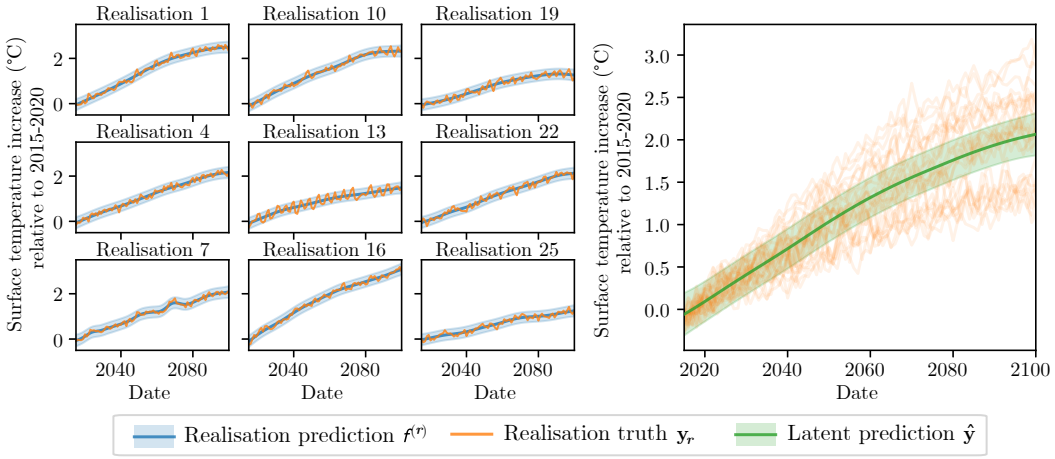

Figure 2: Annual GMST predictions for this century from the SHGP. The 9 plots on the left show predictions and truths for a subset of the 27 realisations from 9 CMIP6 models for the SSP2-4.5 scenario. The right plot shows the latent prediction. Shading across all plots denotes 2 standard deviations or approximately a 95 % credible interval.

Figure 2 shows the predictions for a subset of models and realisations, alongside the prediction for the latent GMST change across 9 CMIP6 models. For this set of models, the predicted GMST increase is 2.1 °C with a credible interval of $[1.8, 2.3]$ °C by the year 2100 relative to the 2015–2020 average. This represents the most likely estimate of GMST change, under the SSP2-4.5 scenario, from the 27 realisations we have used. Unlike a multi-model mean prediction, the SHGP allows for the full propagation of uncertainty across the model ensemble. Extensions could include incorporating a heteroscedastic noise component, to model the non-stationary variance in Figure 2.

4 Conclusion, limitations and further work

We have developed a SHGP that scalably models sets of large climate model simulations for the purpose of extracting and quantifying the latent climate signal, whilst robustly propagating uncertainty. Although our application was estimating the most likely surface temperature change, in practice the SHGP can model a range of data including simulations from other geophysical models, a collection of simultaneously measured time-series, or repeated measurements along a set path.

Limitations Our model scales quadratically in the number of realisations, which for large climate model ensembles of over 100 models will become expensive. This scaling issue is true for any hierarchical GP model, however, as we have implemented a sparse method, computational costs from data size are greatly reduced. Further to this, the SHGP can support batching of the data which will additionally reduce computational cost. A second limitation is that we have assumed that all climate model realisations are equally likely. This flawed assumption [discussed in 9] is widely used throughout climate and Earth science; however, we will tackle this in our upcoming work.

Further work Our implementation features a single hierarchy; that of model realisations to latent climate signal. However, using a multilevel hierarchy would allow the modelling of realisation to climate models to latent climate signal, allowing for inter-model comparison and to aid with disentangling climate modelling complexities. Alongside important climate and modelling applications, the SHGP has a multitude of potential uses across environmental science including, but not limited to, detecting latent air quality from low cost sensor networks, and infilling missing observational data in datasets that share common functionality.

References

[1] Martin B Andrews et al. "Historical simulations with HadGEM3-GC3. 1 for CMIP6". In: *Journal of Advances in Modeling Earth Systems* 12.6 (2020), e2019MS001995 (cit. on p. 3).

[2] Jimmy Ba and Diederik Kingma. "Adam: A Method for Stochastic Optimization". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. 2015. URL: http://arxiv.org/abs/1412.6980 (cit. on p. 2).

[3] Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. "Understanding probabilistic sparse Gaussian process approximations". In: *Advances in neural information processing systems* 29 (2016) (cit. on p. 9).

[4] Veronika Eyring et al. "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization". In: *Geoscientific Model Development* 9 (2016). DOI: 10.5194/gmd-9-1937-2016 (cit. on p. 3).

[5] James Hensman, Neil D Lawrence, and Magnus Rattray. "Hierarchical Bayesian modelling of gene expression time series across irregularly sampled replicates and clusters". In: *BMC bioinformatics* 14.1 (2013), pp. 1–12 (cit. on p. 2).

[6] Michael Hutchinson et al. "Vector-valued Gaussian Processes on Riemannian Manifolds via Gauge Independent Projected Kernels". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 17160–17169 (cit. on p. 2).

[7] IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. 2013 (cit. on p. 1).

[8] Michael Jordan et al. "An Introduction to Variational Methods for Graphical Models". In: *Learning in Graphical Models* 89.1 (1998). Ed. by Michael Jordan, pp. 105–161 (cit. on p. 2).

[9] Reto Knutti. "The end of model democracy?" In: *Climatic Change* 102.3 (2010), pp. 395–404. DOI: 10.1007/s10584-010-9800-2 (cit. on p. 4).

[10] Markus Lange-Hegermann. "Linearly constrained gaussian processes with boundary conditions". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2021, pp. 1090–1098 (cit. on p. 2).

[11] Miguel Lázaro-Gredilla and Michalis K Titsias. "Variational heteroscedastic Gaussian process regression". In: *ICML*. 2011 (cit. on pp. 2, 7).

[12] Bertil Matérn. "Spatial variation : Stochastic models and their application to some problems in forest surveys and other sampling investigations". PhD thesis. Stockholm University, 1960 (cit. on p. 3).

[13] Georges Matheron. "Principles of geostatistics". In: *Economic geology* 58.8 (1963), pp. 1246–1266 (cit. on p. 2).

[14] Alexander G de G Matthews et al. "GPflow: A Gaussian Process Library using TensorFlow." In: *J. Mach. Learn. Res.* 18.40 (2017), pp. 1–6 (cit. on p. 3).

[15] Thomas Pinder and Daniel Dodd. "Gpjax: A Gaussian process framework in JAX". In: *Journal of Open Source Software* 7.75 (2022), p. 4455 (cit. on p. 3).

[16] Thomas Pinder et al. "Gaussian Processes on Hypergraphs". In: *arXiv preprint arXiv:2106.01982* (2021) (cit. on p. 2).

[17] Thomas Pinder et al. "Street-level air pollution modelling with graph Gaussian processes". In: *ICLR: AI for Earth and Space Science* (2022) (cit. on p. 2).

[18] Carl Edward Rasmussen and Christopher K Williams. *Gaussian processes for machine learning*. 3. MIT press Cambridge, MA, 2006 (cit. on p. 1).

[19] Thomas Reichler and Junsu Kim. "How Well Do Coupled Models Simulate Today's Climate?" In: *Bulletin of the American Meteorological Society* 89.3 (2008), pp. 303–312. DOI: 10.1175/BAMS-89-3-303 (cit. on p. 1).

[20] Alan D Saul et al. "Chained gaussian processes". In: *Artificial Intelligence and Statistics*. PMLR. 2016, pp. 1431–1440 (cit. on pp. 2, 7).

[21] Edward Snelson and Zoubin Ghahramani. "Sparse Gaussian processes using pseudo-inputs". In: *Advances in neural information processing systems* 18 (2005) (cit. on p. 2).

[22] Michalis Titsias. "Variational learning of inducing variables in sparse Gaussian processes". In: *Artificial intelligence and statistics*. PMLR. 2009, pp. 567–574 (cit. on pp. 2, 7–9).

[23] R. E. Turner and M. Sahani. “Two problems with variational expectation maximisation for time-series models”. In: *Bayesian Time series models*. Ed. by D. Barber, T. Cemgil, and S. Chiappa. Cambridge University Press, 2011. Chap. 5, pp. 109–130 (cit. on p. 9).

A Variational bound

Let $\mathbf{Z} = \{\mathbf{z}_m\}_{m=1}^M$ be a set of *inducing points* whose corresponding function evaluations are given by $\mathbf{U}_f = \{\mathbf{u}_f^{(r)}\}_{r=1}^R = \{f^{(r)}(\mathbf{Z})\}_{r=1}^R$ and $\mathbf{u}_g = g(\mathbf{Z})$. The function evaluations $\mathbf{U}_f$ and $\mathbf{u}_g$ are termed *inducing variables* as they are themselves random variables assigned the prior distributions of

$$p\left(\mathbf{u}_f^{(r)} \,|\, \mathbf{Z}\right) = \mathcal{N}\left(\mathbf{u}_f^{(r)} \,|\, \mathbf{0}, \mathbf{K}_{\mathbf{uu}}^{f^{(r)}}\right) \tag{6}$$

$$p(\mathbf{u}_g \,|\, \mathbf{Z}) = \mathcal{N}\left(\mathbf{u}_g \,|\, \mathbf{0}, \mathbf{K}_{\mathbf{uu}}^{g}\right) \tag{7}$$

$$\tag{8}$$

where

$$\mathbf{K}_{\mathbf{uu}}^{f^{(r)}} = k_{f^{(r)}}(\mathbf{Z}, \mathbf{Z}) \tag{9}$$

$$\mathbf{K}_{\mathbf{uu}}^{g} = k_g(\mathbf{Z}, \mathbf{Z})\,. \tag{10}$$

Similarly, $\mathbf{K}_{\mathbf{xx}}^{f^{(r)}} = k_{f^{(r)}}(\mathbf{X}, \mathbf{X}')$ and $\mathbf{K}_{\mathbf{xx}}^{g} = k_g(\mathbf{X}, \mathbf{X}')$. When incorporating the inducing variables into our joint prior, we assume that our latent functions are independent

$$p(\mathbf{F}, \mathbf{g} \,|\, \mathbf{U}_f, \mathbf{u}_g) = p(\mathbf{F} \,|\, \mathbf{U}_f) p(\mathbf{g} \,|\, \mathbf{u}_g)\,, \tag{11}$$

where $\mathbf{F}$ is a matrix whose r^{th} column corresponds to the evaluation of $\mathbf{f}^{(r)} = f^{(r)}(\mathbf{X})$ and $\mathbf{g} = g(\mathbf{X})$. By the consistency of a GP, we can be sure that augmenting the GP's joint prior with the inducing variables and marginalising them out will leave no imprint on the true joint distribution.

The marginal log-likelihood of the model presented in Equations (4)–(5) is

$$\log p(\mathbf{Y}) = \sum_{r=1}^{R} \log \int p(\mathbf{y}_r \,|\, \mathbf{f}^{(r)}, \mathbf{g}) p(\mathbf{f}^{(r)}, \mathbf{g} \,|\, \mathbf{u}_f^{(r)}, \mathbf{u}_g) p(\mathbf{u}_f^{(r)}) p(\mathbf{u}_g) \mathrm{d}\mathbf{f}^{(r)} \mathrm{d}\mathbf{g} \mathrm{d}\mathbf{u}_f^{(r)} \mathrm{d}\mathbf{u}_g \tag{12}$$

To enable a tractable model, our approach is to introduce a variational approximation to the model's posterior distribution. By assuming that the latent variables $\mathbf{F}$ and $\mathbf{g}$ factor within the variational posterior distribution, we are then able to enable tractable computations [20]. We will now proceed to derive an ELBO term following the approaches given in Lázaro-Gredilla et al. [11] and Titsias [22] that will allow us to optimise the parameters of our variational approximation such that they offer the best approximation of the true posterior. We begin by defining the augmented joint prior

$$p(\mathbf{F}, \mathbf{g}, \mathbf{U}_f, \mathbf{u}_g \,|\, \mathbf{y}) \approx p(\mathbf{F} \,|\, \mathbf{U}_f) p(\mathbf{g} \,|\, \mathbf{u}_g) q(\mathbf{U}_f) q(\mathbf{u}_g) \tag{13}$$

$$= p(\mathbf{g} \,|\, \mathbf{u}_g) q(\mathbf{u}_g) \prod_{r=1}^{R} p(\mathbf{f}^{(r)} \,|\, \mathbf{u}_f^{(r)}) q(\mathbf{u}_f^{(r)})\,, \tag{14}$$

where $q\left(\mathbf{u}_f^{(r)}\right)$ and $q(\mathbf{u}_g)$ are multivariate Gaussian distributions of the form

$$q\left(\mathbf{u}_f^{(r)}\right) = \mathcal{N}\left(\mathbf{u}_f^{(r)} \,|\, \mathbf{m}_f^{(r)}, \mathbf{\Sigma}_f^{(r)}\right) \tag{15}$$

$$q\left(\mathbf{u}_g\right) = \mathcal{N}\left(\mathbf{u}_g \,|\, \mathbf{m}_g, \mathbf{\Sigma}_g\right)\,. \tag{16}$$

We now seek to obtain a tractable bound on the marginal log-likelihood from Equation (12). To achieve this, we first apply the assumptions of factorisation and latent prior independence to Equation (12) and incorporate Equation (13). Application of Jensen's inequality to the refactored marginal log-likelihood will then yield a tractable lower bound Applying the assumptions of factorisation and independence the marginal log-likelihood from Equation (12) and incorporating Equation (13) givens

$$\log p(\mathbf{Y}) = \sum_{r=1}^{R} \log \int p(\mathbf{y}_r \,|\, \mathbf{f}^{(r)}, \mathbf{g}) p(\mathbf{f}^{(r)} \,|\, \mathbf{u}_f^{(r)}) p(\mathbf{g} \,|\, \mathbf{u}_g) p(\mathbf{u}_f^{(r)}) p(\mathbf{u}_g) \mathrm{d}\mathbf{f}^{(r)} \mathrm{d}\mathbf{g} \mathrm{d}\mathbf{u}_f^{(r)} \mathrm{d}\mathbf{u}_g \tag{17}$$

$$\geq \sum_{r=1}^{R} \mathbb{E}_{q(\mathbf{f})q(\mathbf{g})}\left[\log p(\mathbf{y} \,|\, \mathbf{f}, \mathbf{g})\right] - \mathrm{KL}(q(\mathbf{u}_f^{(r)}) \,||\, p(\mathbf{u}_f^{(r)})) - \mathrm{KL}(q(\mathbf{u}_g) \,||\, p(\mathbf{u}_g))\,. \tag{18}$$

The forms of $q\left(\mathbf{f}^{(r)}\right)$ and $q(\mathbf{g})$ are given by

$$q(\mathbf{f}^{(r)}) = \int p(\mathbf{f}^{(r)} \,|\, \mathbf{u}_f^{(r)}) q\left(\mathbf{u}_f^{(r)}\right) \mathrm{d}\mathbf{u}_f^{(r)} \tag{19}$$

$$= \int \mathcal{N}\left(\mathbf{f}^{(r)} \,|\, \mathbf{K}_{\mathbf{fu}}^{f^{(r)}} \left[\mathbf{K}_{\mathbf{uu}}^{f^{(r)}}\right]^{-1} \mathbf{u}_f^{(r)}, \mathbf{K}_{\mathbf{ff}}^{(r)} - \mathbf{Q}_{\mathbf{ff}}^{(r)}\right) q\left(\mathbf{u}_f^{(r)}\right) \mathrm{d}\mathbf{u}_f^{(r)} \tag{20}$$

$$q(\mathbf{g}) = \int p(\mathbf{g} \,|\, \mathbf{u}_g) q(\mathbf{u}_g) \mathrm{d}\mathbf{u}_g \tag{21}$$

$$= \int \mathcal{N}\left(\mathbf{g} \,|\, \mathbf{K}_{\mathbf{gu}}^{g} \left[\mathbf{K}_{\mathbf{uu}}^{g}\right]^{-1} \mathbf{u}_g, \mathbf{K}_{\mathbf{gg}} - \mathbf{Q}_{\mathbf{gg}}\right) q(\mathbf{u}_g) \mathrm{d}\mathbf{u}_g \,. \tag{22}$$

where $\mathbf{Q}_{\mathbf{ff}}^{(r)} = \mathbf{K}_{\mathbf{fu}}^{f^{(r)}} \left[\mathbf{K}_{\mathbf{uu}}^{f^{(r)}}\right]^{-1} \mathbf{K}_{\mathbf{uf}}^{f^{(r)}}$ and $\mathbf{Q}_{\mathbf{gg}} = \mathbf{K}_{\mathbf{gu}}^{g} \left[\mathbf{K}_{\mathbf{uu}}^{g}\right]^{-1} \mathbf{K}_{\mathbf{ug}}^{g}$.

By virtue of the likelihood function being Gaussian, we are able to analytically evaluate the expectation in Equation (18) using the results outlined in Titsias [22]. Further, both of the Kullback-Leibler divergence (KLD) terms in Equation (18) are being evaluated on a pair of multivariate Gaussian distributions, meaning that they have an analytical form that can easily be computed. As such, when we bring these results together, we are provided with the following ELBO term

$$\mathcal{L}_q^* = \sum_{i=1}^{R} \sum_{j=1}^{R} \Big[\log \mathcal{N}\left(\mathbf{y} \,|\, \mathbf{0}, \sigma^2 \mathbf{I}_n + \mathbf{Q}_{\mathbf{gg}} + \delta_{i,j} \mathbf{Q}_{\mathbf{ff}}^{(i)}\right) - \frac{1}{2\sigma^2} \operatorname{trace}\left(\mathbf{K}_{\mathbf{xx}}^{g} + \mathbf{Q}_{\mathbf{gg}} - \delta_{i,j}\left(\mathbf{K}_{\mathbf{xx}}^{f^{(i)}} + \mathbf{Q}_{\mathbf{ff}}^{(i)}\right)\right) \Big] , \tag{23}$$

where $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise i.e., a Dirac delta function.

B Model validation

Data To empirically validate our model, we simulate four datasets where the latent function in each is a single realisation of a Matérn process. The lengthscale used in the latent functions' draws are 0.1, 0.2, 0.5, and 1. 10 realisations are produced from the latent function where each is corrupted by a homoscedastic, zero-mean Gaussian noise vector with variance of 0.1. The relationship between the latent function and each realisation is further corrupted by shifting each realisation by a constant factor whose value is drawn from a uniform distribution with limits of -1.5 and 1.5. We plot the four datasets used in Figure 3.

Apriori, we should expect our model to perform better on datasets whose latent function was drawn from a Matérn process with a larger lengthscale as this will correspond to a smoother function that will be easier to model, particularly when the ratio of data points to inducing points is high.

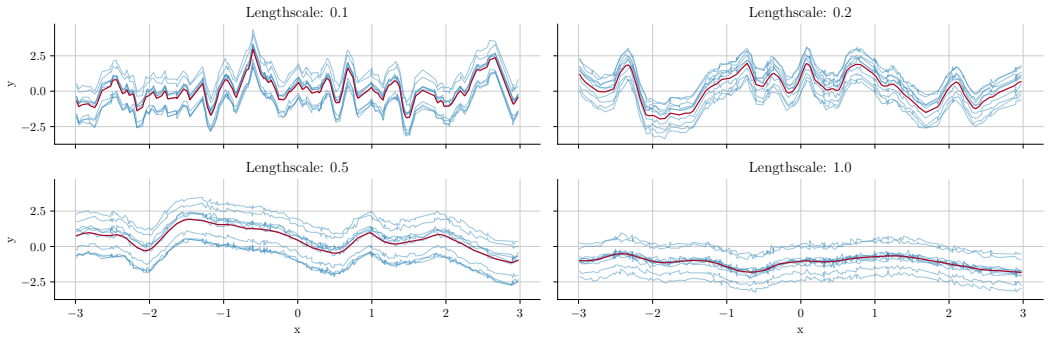

Figure 3: Four simulated datasets used for model validation in Section B. The value of ℓ in each figure denotes the lengthscale used in the latent Matérn function from which each dataset is generated.

Predicitve quality The first question we seek to answer is "*How well can our model recover the true latent function using only a set of realisations?*". To this end, we simulate 10 realisations from the four Matérn processes described above. We further vary the number of data points within a single realisation to be one of 50, 100, or 300. The latent function's lengthscale and the number of data points within a single realisation are our independent variables. To each of the four datasets we fit our model from Section 2 and report the R^2 coefficient to assess how much of the variability within latent function's value is explained by our model. This is our dependent response variable.

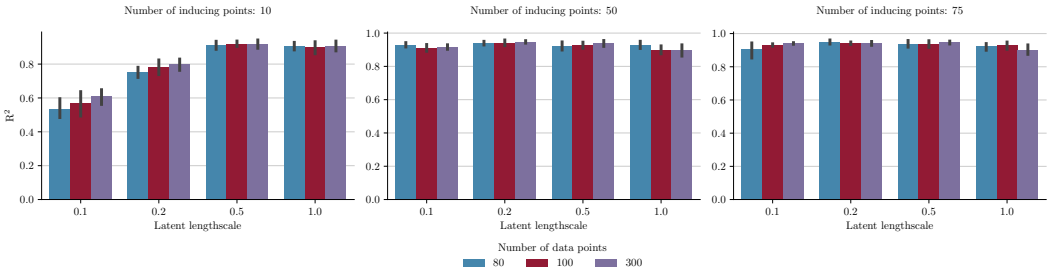

Figure 4: R^2 scores of the our model plotted as a function of latent function's true lengthscale and the number of data points within each realisation.

From Figure 4 we can see that our model is able to consistently model the variance in the latent function to a high fidelity. The only time we see a degradation in performance is when the latent function's lengthscale is 0.1 or 0.2, and we use just 10 inducing points in our model. This is not surprising though as we should never expect to recover such rough functions with so few inducing points as the GP will be forced to smooth out the underlying function's roughness when interpolating from one inducing point to the next.

Posterior calibration We further validate our model by answering the question "*How well calibrated is the uncertainty given by our model when representing the latent function?*". We use the same data generating process as above, however, we now let our independent variables be the true latent function's lengthscale and the number of realisations within a dataset. To test calibration, we query our GP at a set of test locations and report the percentage of times that the true response value fell within our model's 95% credible interval. An over-confident model would report values less that 95%, whilst an under-confident model would report values larger than 95%.

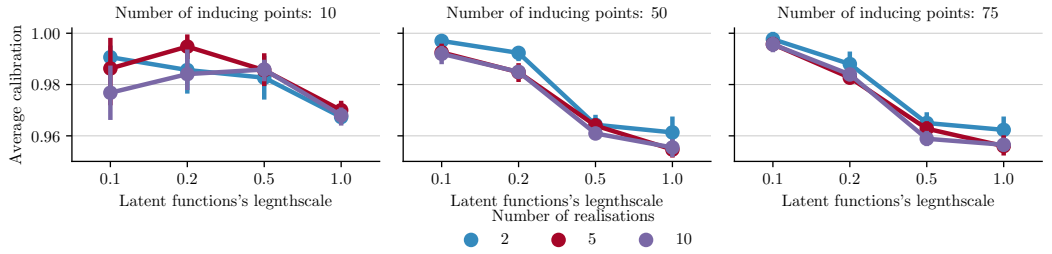

Figure 5: The posterior calibration of our model reported as a function of the number of realisations.

From Figure 5 we can see that the uncertainty estimates of our model are well calibrated. There is a tendency for the posterior distribution of our model to be under-confident in its predictions. However, this is a common artefact of variational GP models of the form given in Titsias [22] due to the different rank of $\mathbf{K}^{g}_{\mathbf{xx}}$ and $\mathbf{K}^{f^{(r)}}_{\mathbf{xx}}$ when compared to $\mathbf{Q}_{\mathbf{gg}}$ and $\mathbf{Q}^{(r)}_{\mathbf{ff}}$ [3, 23]. Further, as the number of inducing points grows, the calibration of the model's posterior improves.