Enhancing Uncertainty Estimation with Deep Gaussian Processes

Oluwole Oyebamiji[1], Dilum Dissanayake[2], Muhammed Cavus[3]
[1]School of Computer Science, University of Birmingham, Birmingham, UK,
[2]School of Geography, Earth and Environmental Sciences, University of Birmingham, UK,
[3]Faculty of Engineering and Environment, Northumbria University, Newcastle Upon Tyne, UK
o.k.oyebamiji@bham.ac.uk

Abstract: Accurately estimating uncertainty in predictive models is crucial for a wide range of applications, from decision-making in landuse modelling to robust forecasting in finance and autonomous systems. Gaussian processes (GPs) offer a solid framework for uncertainty quantification but often struggle with scalability and flexibility when applied to large, high-dimensional datasets. Deep Gaussian processes (DGPs) are a powerful extension of GPs that allow for multi-layer generalisation of GPs, enabling more flexible and expressive modelling of complex data. As the complexity of the model increases, so does the computational cost, which makes it difficult to scale DGP to large-dimensional data. Although variational inference has been used with large datasets, it often produces an overconfident uncertainty estimate because it does not effectively utilise input-dependent function uncertainty. This paper introduces an approach for enhancing uncertainty estimation using the predictive log-likelihood (PLL) objective with DGP model to address these limitations. This relies on a parametric GP regression model designed for a family of predictive distributions and incorporate a modified objective function to restore a full symmetry between various contributions to predictive variance. We evaluate the performance of our methods on several benchmark regressions and large-scale environmental datasets. The results show that the model provides more reliable uncertainty estimates, particularly in regions of sparse data, making them efficient for real-world applications.

Keywords: DGP, Uncertainty Estimation, Variational Inference, Bayesian Methods.

1. Introduction

Uncertainty analysis is essential in various fields, including scientific modelling, engineering, and artificial intelligence, to assess the reliability of predictions and to make appropriate decisions. Uncertainty is inherent in many real-world problems, arising from incomplete knowledge, measurement errors, and model limitations [1]. Understanding and quantifying uncertainty is crucial for decision-making, risk assessment, and model improvement [2]. For example, making effective landuse decisions requires a holistic systems approach to evaluate the trade-offs and benefits associated with each choice in the ecosystem. As a result, large-scale integrated modelling platforms are becoming increasingly popular because they facilitate the simultaneous evaluation of different scenarios along with their environmental, societal, and economic impacts. These modelling techniques are essential for making informed landscape decisions [3, 4]. However, there is a growing need for innovative strategies that support sound landscape decisions in the face of uncertainty, especially concerning the effects of climate change and the efficacy of proposed mitigation measures. Accurately quantifying uncertainty is vital, as decision-makers often face complex choices about which actions to prioritise. Machine learning techniques have been widely applied for uncertainty quantification, utilising probabilistic models, Bayesian inference, and Monte Carlo (MC) methods, among others. These methods

assist in estimating and characterising uncertainty in various types of data and model predictions. In particular, approaches such as Bayesian hierarchical models, Bayesian deep learning, and GPs have seen extensive use in numerous applications, helping to propagate uncertainties across sub-models from one level to the next [5–7]. An alternative approach involves conducting sensitivity analysis, which attributes the uncertainty in a model's outputs to the uncertainty in its inputs [8,9]. This technique facilitates the identification of the critical parameters that drive the majority of the variation in the model's output [4].

Predictive modelling involves using a statistical or machine learning technique to forecast outcomes based on observed data. A Gaussian Process (GP) is a powerful tool in machine learning and statistics used for probabilistic modelling [10]. Their non-parametric form, analytical properties, and ability to model uncertainty are useful in machine learning. However, they are affected with limitations, particularly their significant computation and storage cost. Sparse GPs (SGPs) [11, 12] attempt to address the computational and storage cost. One common approach to SGPs is to use the Nystrom approximation [13, 14]. Several methods have been proposed along this line of research, each with its advantages and limitations. While SGPs and other variational techniques [12,15,16] have addressed the problem of computational costs, GPs are still not suitable for many applications due to their reliance on covariance functions having limited expressiveness. Covariance function is the crucial ingredient in a GP regression, as it encodes our assumptions about the function which we wish to learn.

These functions are most frequently employ fairly straightforward distance metrics. Therefore, in some applications, applying different similarity metrics across various regions of the input space may be necessary. One approach to this challenge would be stacking GPs, similar to the stacking of Perceptrons in a Multi-Layer Perceptron (MLP) [17]. However, arranging GPs so that the output from one layer feeds into the next increases non-linearity and makes the inference analytically intractable. Such methodologies are generally classified as DGPs [18]. DGPs are multi-layer generalisations of GPs that enhance flexibility and expressiveness in modelling complex data. However, as model complexity increases, so does computational cost, making it challenging to apply DGPs to high-dimensional data. Variational inference has been utilised with large datasets, but it frequently produces overly confident uncertainty estimates due to asymmetry in variance contribution and insufficient use of input-dependent function uncertainty [19, 20]. The work of [21] introduced a predictive log likelihood (PLL) objective that restores symmetry between the objective and the predictive posterior, called parametric predictive GP regressor (PPGRR). This paper presents a unified technique to improve uncertainty estimation of DGP models using the PLL objective function and is similar to the work of [22] on Deep Sigma Point Processes (DSPPs).

Our main contributions in this paper are provided below: the DSPP model of [22] used a similar predictive log likelihood objective but applied a Gauss-Hermite quadrature to approximate the intractable expectation, leading to higher computational expenses, particularly with large datasets. In our case, we assume that additional use of quadrature approximation is unnecessary for approximating expectations and that with a sufficient number of MC samples, it converges and provides adequate performance, especially when the aim is uncertainty quantification. We have also demonstrated this through numerous standard benchmark datasets. The Gauss-Hermite quadrature rule can cause an exponential increase in mixture components and subsequent scalability issues as it introduces additional parameters from the learned quadrature rule. Moreover, their empirical comparisons mainly focus on univariate and multivariate regression tasks, potentially limiting the applicability of the results to other task types or more varied datasets. To enhance empirical assessments, we conduct experiments across a broader spectrum of datasets, including those from real-world scenarios that capture spatial and spatiotemporal correlations. We also expand our method to investigate various forms of variational inference for estimating DGP parameters.

2. Background

This section starts with an overview of DGP models, its parameter estimation and predictive distribution. Traditional GPs often struggle to effectively capture the intricate and hierarchical nature of large datasets. To overcome this challenge, DGPs combine several layers of GPs, resulting in more flexible and expressive models. A DGP is formed by layering multiple GPs, where the output from one layer feeds into the next [23]. A DGP consisting of L layers can be mathematically expressed as:

$$\begin{aligned}\mathbf{f}_1(\mathbf{X}) &\sim \mathcal{GP}\left(m_1(\mathbf{X}), k_1(\mathbf{X}, \mathbf{X}')\right),\\ \mathbf{f}_2(\mathbf{f}_1) &\sim \mathcal{GP}\left(m_2(\mathbf{f}_1), k_2(\mathbf{f}_1, \mathbf{f}_1')\right),\\ &\vdots\\ \mathbf{f}_L(\mathbf{f}_{L-1}) &\sim \mathcal{GP}\left(m_L(\mathbf{f}_{L-1}), k_L(\mathbf{f}_{L-1}, \mathbf{f}_{L-1}')\right).\end{aligned}$$

Inference in DGP poses difficulties because of the complexity of the exact posterior distribution. To tackle this issue, various approximate inference methods have been introduced [18, 21, 24]. More information on parameter estimation for a single-layer GP can be found in the next section.

2.1. Single-layer GP

Consider a stochastic function $f : \mathbb{R}^D \to \mathbb{R}$, given a likelihood $p(y \mid f)$ and observations $\mathbf{y} = (y_1, \ldots, y_N)^\mathsf{T}$ at design locations $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)^\mathsf{T}$. A GP prior can be assigned to the function f, modelling all function values as joint Gaussian, characterised by a covariance function $k : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$ and a mean function $m : \mathbb{R}^D \to \mathbb{R}$. Similarly, let us define a variable $\mathbf{u} = f(\mathbf{Z})$, and for a set of $\mathcal{M}$ inducing locations $\mathbf{Z} = (\mathbf{z}_1, \ldots, \mathbf{z}_M)^\top$, and let $\mathbf{f} = f(\mathbf{X})$ denotes the function values at the design points. The expressions $[m(\mathbf{X})]_i = m(\mathbf{x}_i)$ and $[k(\mathbf{X}, \mathbf{Z})]_{ij} = k(\mathbf{x}_i, \mathbf{z}_j)$ represent the joint density $p(\mathbf{f}, \mathbf{u})$, which is Gaussian in nature. The mean of this density is derived from the mean function evaluated at every input $(\mathbf{X}, \mathbf{Z})^\top$, with the corresponding covariance being obtainable as well [18,21]. Given that the joint GP prior $p(f, \mathbf{u}; \mathbf{x}, \mathbf{z}) = p(\mathbf{f} \mid \mathbf{u}; \mathbf{X}, \mathbf{Z}) \times p(\mathbf{u})$, the joint density of $\mathbf{y}, \mathbf{f}$ and $\mathbf{u}$ can be defined as

$$p(\mathbf{y}, \mathbf{f}, \mathbf{u}) = \underbrace{p(\mathbf{f} \mid \mathbf{u}; \mathbf{X}, \mathbf{Z})p(\mathbf{u}; \mathbf{Z})}_{\mathrm{GP}_{\mathrm{prior}}} \underbrace{\prod_{i=1}^{N} p(y_i \mid f_i)}_{\text{likelihood}}.$$

Inference in the joint density presented above can be performed in closed form when the likelihood is Gaussian, although the computational complexity increases cubically with N. We can implement variational inference to approximate the posterior distribution $q(\mathbf{f}, \mathbf{u})$ by minimising the Kullback-Leibler divergence KL[q || p]. This process is equivalent to maximising the lower bound of the marginal likelihood (evidence): $\mathbb{L} = \mathbb{E}_{q(\mathbf{f},\mathbf{u})}\left[\log \frac{p(\mathbf{y},\mathbf{f},\mathbf{u})}{q(\mathbf{f},\mathbf{u})}\right]$, which is given as

$$\mathcal{L}_{\mathrm{svgp}} = \sum_{i=1}^{N} \left\{ \log \mathcal{N}\left(y_i \mid \mu_{\mathbf{f}}\left(\mathbf{x}_i\right), \sigma^2_{\mathrm{obs}}\right) - \frac{\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right)}{2\sigma^2_{\mathrm{obs}}} \right\} - \mathrm{KL}(q(\mathbf{u}) \,||\, p(\mathbf{u})), \tag{2.1}$$

where $\mu_{\mathrm{f}}\left(\mathrm{x}_i\right)$ is the predictive mean function given by $\mu_{\mathrm{f}}\left(\mathbf{x}_i\right) = \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} m$ and $\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right) \equiv \mathrm{Var}\left[\mathrm{f_i} \mid \mathbf{x_i}\right]$ denotes the latent function variance $\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right) = \tilde{\mathbf{K}}_{ii} + \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} \mathbf{S} \mathbf{K}_{MM}^{-1} \mathbf{k}_i$, and $\mathcal{L}_{\mathrm{svgp}}$ which depends on $m, \mathbf{S}, \mathbf{Z}, \sigma_{\mathrm{obs}}$ and the various kernel hyperparameters θ_{ker} can be maximised with gradient methods. Further details is given in Appendix A1.

2.2. Predictive distribution

Given the inducing variable $\mathbf{u}$, the predictive distribution for input $\mathbf{x}^*$ is expressed as $p\left(\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{u}\right) = \mathcal{N}\left(\mathbf{y}^* \mid \mathbf{k}_*^T \mathbf{K}_{MM}^{-1} \mathbf{u}, \tilde{\mathbf{K}}_{**} + \sigma^2_{\mathrm{obs}}\right)$ and by integrating out $\mathbf{u} \sim \mathcal{N}(m, \mathbf{S})$, then yields

$$p\left(\mathbf{y}^* \mid \mathbf{x}^*\right) = \mathcal{N}\left(\mathbf{y}^* \mid \mu_{\mathrm{f}}\left(\mathrm{x}^*\right), \Sigma_{\mathrm{f}}\left(\mathrm{x}^*\right) + \sigma^2_{\mathrm{obs}}\right), \tag{2.2}$$

where $\mu_{\mathrm{f}}\left(\mathrm{x}^*\right)$ is the predictive mean function in (A3) and $\Sigma_{\mathrm{f}}\left(\mathrm{x}^*\right)$ is the latent function variance in (A4). The predictive variance Var[y*/x*] at input x* has two contributions from (2.1) namely the input-dependent $\Sigma_f\left(\mathrm{x}^*\right)$ and input-independent σ^2_{obs}. According to [21], the two contributions appear assymetrically and $\Sigma_f\left(\mathrm{x}^*\right)$ term does not appear in the data fit term that results from the expected log-likelihood $E_{q(u)}[\log p(y_i \mid x_i; u)]$. This frequently results in a disconnect between the training objective and the predictive distribution utilised during testing. To tackle the asymmetrical behaviour of σ^2_{obs} and $\Sigma_{\mathrm{f}}(\mathbf{x}^*)$ in (2.1), a parametric GP regression model can be constructed directly based on the set of predictive distributions outlined in (2.1), which leads to a new objective function rooted in the KL divergence between $p(y \mid \mathbf{x})$ and $p_{\mathrm{data}}(y \mid \mathbf{x})$ such that

$$\mathcal{L}_{\text{ppgpr}} = \mathbb{E}_{\mathrm{p}_{\text{data}}(y,\mathbf{x})} \left[\log p(y \mid \mathbf{x})\right] - \beta_{\text{reg}}\, \mathrm{KL}(q(\mathbf{u}) \,||\, p(\mathbf{u})]. \tag{2.3}$$

The objective in (2.3) (also called, predictive log likelihood (PLL)) can be expanded as:

$$\mathcal{L}_{\text{ppgpr}} = \sum_{i=1}^{N} \log \mathcal{N}(y_i \mid \mu_{\text{f}}(\mathbf{x}_i), \sigma^2_{\text{obs}} + \Sigma_{\text{f}}(\mathbf{x}_i)) - \beta_{\text{reg}} \mathrm{KL}(q(\mathbf{u}) \,||\, p(\mathbf{u})) \tag{2.4}$$

From (2.4), the data fit term can be obtained as $\mathcal{L}_{\text{ppgpr}} \supset -\frac{1}{2}\frac{1}{\sigma^2_{\text{obs}}+\Sigma_{\text{f}}(\mathbf{x}_i)} \left|y_i - \mu_{\text{f}}(\mathbf{x}_i)\right|^2$ and for the (2.1): $\mathcal{L}_{\text{svgp}} \supset -\frac{1}{2}\frac{1}{\sigma^2_{\text{obs}}} \left|y_i - \mu_{\text{f}}(\mathbf{x}_i)\right|^2$. We note that the two objectives, while appearing similar, differ significantly. Specifically, the PPGPR or PLL objective focuses on the predictive distribution, whereas the SVGP objective aims to minimise the KL divergence between the variational distribution and the true posterior. Drawing inspiration from the studies by [18] on DGPs, we note that DGP establishes a prior recursively on vector-valued stochastic functions $F^1, \ldots, F^L$. A prior can be allocated to each function $\boldsymbol{F}^i$, which acts as an independent GP in each dimension. The input locations for these GPs are determined by the noisy corruptions of the function values from the subsequent layer. At layer l, the outputs generated by the GPs are denoted as F^I_d, while the related inputs come from F^{l-1}. It is assumed that the noise across layers follows an i.i.d. Gaussian distribution [1]. As with the single-layer case, we have inducing locations Z^{l-1} and inducing function values $\mathbf{U}^l$ for each dimension. We can define the joint density in a similar way to single-layer GP

$$p(\mathbf{Y}, \{\mathbf{F}^i, \mathbf{U}^l\}_{l=1}^L) = \prod_{i=1}^{N} p(\mathbf{y} \mid \mathbf{f}_i^L) \prod_{l=1}^{L} p(\mathbf{F}^l \mid \mathbf{U}^i; \mathbf{F}^{l-1}, \mathbf{Z}^{l-1}) p(\mathbf{U}^l; \mathbf{Z}^{l-1}),$$

where $F^0 = X$, we can employ a sparse variational inference method to derive the efficient lower bound, which is expressed as $\mathcal{L}_{DGP} = \mathbb{E}_{q(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)} \left[\frac{p(\mathbf{Y}, \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)}{q(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)}\right]$. Through substitution and rearrangement, we derive

$$\mathcal{L}_{\text{DGP}} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbb{F}_i^L)} \mid \log p(\mathbf{y}_i \mid \mathbf{f}_i^L)| - \sum_{l=1}^{L} \mathrm{KL}[q(\mathbf{U}^l) \,||\, p(\mathbf{U}^l; \mathbf{Z}^{l-1})]\,.$$

Further details on derivation is given in Appendix A2.

2.3. A two-layer DGP with predictive log likelihood objective

Consider a basic scenario involving a 2-layer DGP with a single continuous output y, where a GP layer of width W is inputted into the uppermost GP, the joint likelihood for the dataset $(\boldsymbol{y}, \boldsymbol{X})$ can be expressed as

$$p(\mathbf{y}, \mathbf{f}, \mathbf{G} \mid \mathbf{X}) = p\left(\mathbf{y} \mid \mathbf{f}, \sigma^2_{\text{obs}}\right) p(\mathbf{f} \mid \mathbf{G}) p(\mathbf{G} \mid \mathbf{X}), \tag{2.5}$$

where $\mathbf{G}$ represents a matrix of latent function values with dimensions $N \times W$. Specifically, g_{iw} indicates the latent function value of the w^{th} GP in the first layer for the $i^{\textbf{th}}$ input $\mathbf{x}_i$. For $i = 1, \ldots, N$, the vector $\mathbf{g}_i \equiv g_{i,:}$, which has a dimension of W, and acts as the input to the second GP denoted as $\mathbf{f}$. We assume that the prior over $\mathbf{G}$ factorises as $p(\mathbf{G} \mid \mathbf{X}) = \prod_{w=1}^{W} p(\mathbf{g}_w \mid \mathbf{X})$, with each GP governed by its distinct kernel. Due to the intractability of inference in this model, we need to rely on approximate methods. Recall that the predictive distribution at test input x is given by the Gaussian distribution in (2.2). Therefore, as a consequence of (2.5), the predictive distribution of a two-layer DGP is given as

$$\mathbb{E}_{\prod_{w=1}^{W} q(g_{*w} \mid \mathbf{x}_*)} \left[\mathcal{N}\left(y_* \mid \mu_f\left(\mathbf{g}_*\right), \sigma_f\left(\mathbf{g}_*\right)^2 + \sigma^2_{\text{obs}}\right)\right]. \tag{2.6}$$

Given that it is a continuous mixture of Gaussian distributions, and since this expectation is difficult to compute, it needs to be approximated using MC samples, which means representing it as a finite mixture of Normal distributions. According to [22], using the PPGPR objective (2.4) for a direct approximation of (2.6) involves the logarithm of the expectation, making it challenging. To address this problem, the authors suggest substituting the continuous mixture of Gaussians with a finite parametric mixture of Gaussians. This is practically carried out by employing an appropriate quadrature method, which introduces additional computations and parameters to the training. See further details on how to apply the quadratures rules in Appendix A3. We used MC sampling to evaluate the integral in (2.6) instead of a quadrature scheme or learned weights. Approximating expectations with enough samples not only converges but also performs relatively well. We have provided various variational strategies to test these methods in section 4.

[1] Let $k_{\text{noisy}}(\mathbf{x}_i, \mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma_l^2 \delta_{ij}$ where δ is the Kronecker delta, and σ_l^2 is the noise variance between layers.

3. Related works

We discuss popular existing variational inference methods related to our approaches in this section. GPs [14] are widely used Bayesian non-parametric methods. However, their capacity to model complex functions is limited, leading to the creation of DGPs [25]. DGPs build on GPs by layering latent GPs, resulting in models that are more expressive, which introduce hierarchical structure and greater flexibility. However, exact inference in DGPs is challenging due to non-Gaussian posteriors, which requires employing variational inference (VI) [11]. VI aims to optimise the Evidence Lower Bound (ELBO) to approximate the intractable posterior. Although ELBO-based VI is computationally effective, it frequently underrepresents uncertainty in deep models [18]. Stochastic variational inference [26] modifies traditional variational approach for large datasets by maximising the ELBO using stochastic gradient descent. Stochastic variational inference using inducing points was first introduced by [11] to efficiently approximate latent variables. Although this method is scalable, it frequently results in inadequate uncertainty quantification in the presence of deep hierarchies. Salimbeni & Deisenroth [18] build on this technique by integrating MC sampling for both function values and kernel hyperparameters. This enhancement enables more expressive posterior distributions but also raises computational complexity. Conversely, Hernández-Lobato et al. [27] apply expectation propagation (EP), which improves traditional variational methods by minimising Kullback-Leibler (KL) divergence. This technique enhances uncertainty representation, but the need for costly expectation calculations limits its scalability. Knoblauch [28] extended Renyi divergence [29] to robust DGPs using generalised variational inference. The Importance Weighted ELBO (IW-ELBO), proposed by [30], enhances the standard ELBO by incorporating importance weighting. This method yields tighter bounds on marginal likelihood, thereby improving inference accuracy, but entails higher computational complexity. Salimbeni et al. [31] implemented this technique for DGPs with importance-weighted variational inference.

3.1. Uncertainty analysis

We want to clarify that our aim is not to accurately calculate the complex expectation. Rather, we seek to develop range of parametric models that are governed by stable mean and variance functions, allowing for a reliable assessment of uncertainty. In various scientific fields, intricate computer models are often employed to model behaviour of large-scale physical systems. However, the model's response to input changes is not always clear. Therefore, conducting an uncertainty analysis on such models is crucial. [2] Shapley analysis is a well-known method related to sensitivity analysis. Originating from cooperative game theory, it assesses each feature's contribution to a model's predictions. A primary method used in this paper is SHapley Additive exPlanations (SHAP), which effectively clarifies model outputs by calculating the marginal contribution of each feature [33,34]. We provided some results on the Shapley analysis for the IAP2 data in the next section.

4. Experiments

This section outlines the evaluation process for the experiment. We used several datasets from the UCI repository, including Bikes, Elevators, Energy, 3Droad, Protein, Jura, and landuse IAP2 data. The characteristics of each dataset, evaluated across a diverse array of models, are detailed in Table A1 in the Appendix. The data is split in an 80%-20% ratio for training and testing purposes. Results are generated by training models on a designated training dataset and evaluating their performance on a specified testing dataset. [3] The models considered, along with their variants, include DSPP, DGP, DGP* and standard GPs. The model is implemented using GPyTorch and employs a Matern covariance function with independent length scales assigned for each input dimension [16, 37]. The marginal log-likelihood is maximised using the Adam optimiser [38].

[2] Uncertainty quantification focuses on assessing and analysing the uncertainty in model predictions arising from input variability, model structure, and data noise. Sensitivity analysis examines how changes in model inputs affect the outputs and can identify regions within the input variability space that lead to significant and extreme output values. Also, it helps in uncovering interactions among variables and highlights areas where the model could be enhanced for better input values [3, 32]

[3] The experiments are conducted over a total of 400 epochs. Some models used significantly higher computational resources than others in terms of parameter space. For example, IAP2 data takes a little longer to train due to its larger number of dimensions and sample size compared to the Jura dataset, which has just 359 samples.

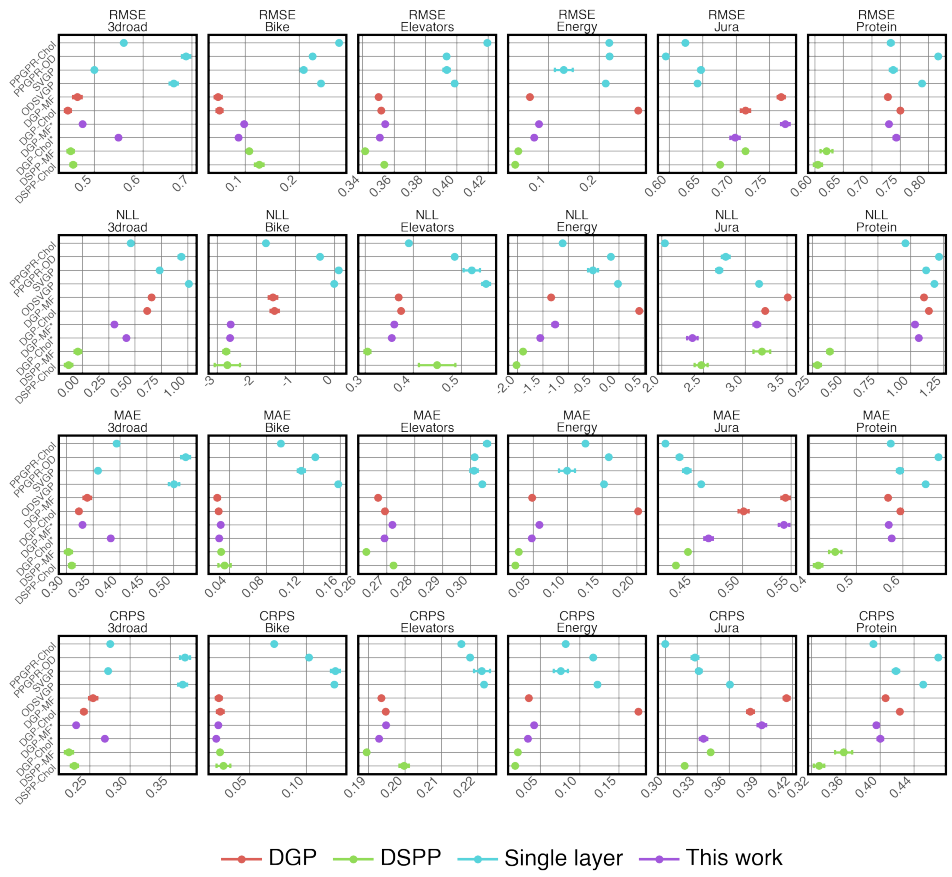

Figure 1. We evaluate the performance of various model variants across six selected benchmark datasets. The metrics used for comparison include Root Mean Square Error (RMSE), Negative Log Likelihood (NLL), Continuous Rank Probability Score (CRPS), and Mean Absolute Error (MAE). In all cases, lower values indicate better performance. The results are averaged over random splits of the training and validation sets. MF stands for mean field, Chol refers to Cholesky, and OD is an abbreviation for Orthogonally Decoupled, as noted in [35]. SVGP denotes the sparse GP techniques discussed in [36].

Both the DGP and DSPP models consist of three layers, using 50 and 8 MC samples and quadrature weights, respectively. Figure 1 presents the performance of these models across various UCI datasets. The 'DGP' model employs the VI ELBO objective, while the 'DGP*' model uses the PLL objective as defined in (2.4), with expectations derived from MC samples. The ODSVGP and SVGP are single-layer GPs that we used as baseline models. The DSPP methods perform better than other approaches, particularly in predicting the 3Droad, Energy, and Protein datasets, as indicated by the RMSE and NLL metrics. In contrast, the models labelled 'this work' (DGP-MF* and DGP-Chol*, where 'MF' denotes the mean field and 'Chol' represents Cholesky approximation) show slightly better results for the Bike dataset based on RMSE and NLL values. For the Elevators and Jura datasets, the performance outcomes are comparable. Our methods (DGP*), along with the DSPP models, perform much better than the VI ELBO-based DGP model [18]. This enhanced performance can be attributed to the implementation of a more effective PLL objective in both instances, which allows for improved uncertainty calibration. The results reveal a distinct divergence regarding the preference for mean field or Cholesky approximation, as performance varies across methods and datasets. Generally, multilayer DGPs outperform single-layer GPs, with the exception of the Jura dataset. We further summarise the results in Table 1 by averaging the metrics across all six selected datasets, highlighting the model variants that perform much better. For both RMSE and NLL metrics, the DSPP model using the full covariance matrix via Cholesky decomposition shows markedly better results, followed by its mean field variant and DGP* models in this study, where lower values are preferred.

We now compare the results for the IAP2 data, as presented in Table 2. The IAP2 dataset consists

Table 1. The results are ranked on average (a lower is better). Results are averaged over the six test datasets.

Model	RMSE	NLL	CRPS	MAE
DSPP-Chol	**0.375**	**-0.280**	**0.187**	**0.249**
DSPP-MF	0.379	-0.136	0.193	0.254
DGP-Chol*	0.407	-0.004	0.209	0.295
DGP-MF*	0.419	0.154	0.213	0.301
DGP-Chol	0.432	0.705	0.240	0.320
DGP-MF	0.404	0.454	0.221	0.299
ODSVGP	0.491	0.980	0.274	0.370
SVGP	0.436	0.776	0.241	0.319
PPGPR-OD	0.490	0.813	0.267	0.371
PPGPR-Chol	0.472	0.163	0.223	0.319

Table 2. We present the RMSE, NLL, CRPS, and MAE for the three-layer DGP and its variants on the IAP2 dataset, where lower values indicate better performance. The results are averaged over random splits of the training and validation sets. Note: MF stands for mean field, Chol for Cholesky, OD for orthogonally decoupled [35], and SVGP refers to the sparse GP techniques discussed in [36].

Model	RMSE	NLL	CRPS	MAE
DSPP-MF	1.083 ± 0.001	-3.658 ± 0.100	0.508 ± 0.000	0.685 ± 0.001
DSPP-Chol	0.891 ± 0.003	**-7.36 ± 0.060**	0.400 ± 0.001	0.562 ± 0.001
DGP-MF*	**0.889 ± 0.003**	6.517 ± 0.180	**0.376 ± 0.002**	**0.536 ± 0.003**
DGP-Chol*	0.922 ± 0.002	6.317 ± 0.167	0.386 ± 0.002	0.548 ± 0.002
DGP-MF	0.925 ± 0.008	20.109 ± 0.188	0.455 ± 0.004	0.563 ± 0.005
DGP-Chol	0.958 ± 0.005	10.275 ± 0.216	0.413 ± 0.002	0.578 ± 0.003

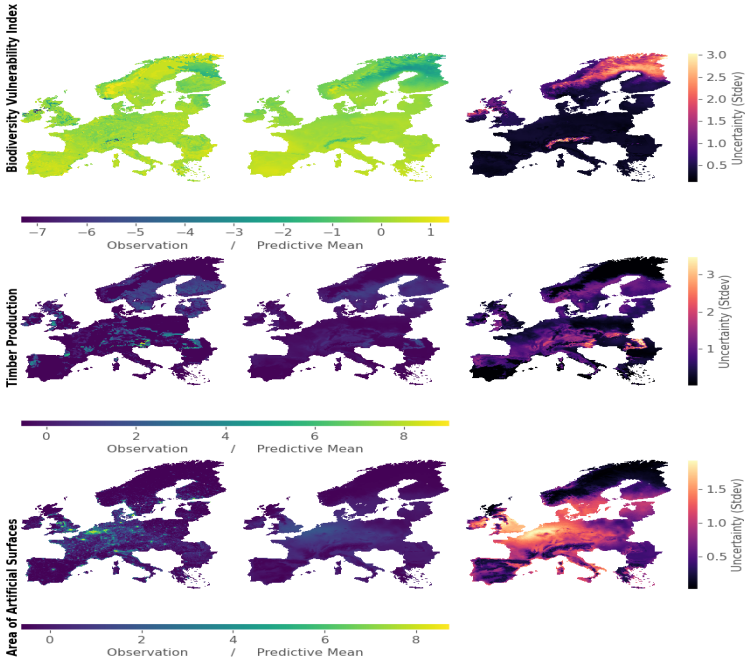

Figure 2. Spatial maps comparing selected model outputs for the SSP3 scenario in 2050 from the IAP2 data of the DGP-MF* model.

of fifteen outputs, and we have calculated the aggregated RMSE, NLL, CRPS, and MAE results. The performance of the various models on this dataset is quite similar according to the RMSE and NLL values. However, the DGP*-based models show slightly better performance than the others, as indicated by their lower metric values, except for NLL. We provide visualisations to demonstrate the performance of the proposed methods for uncertainty quantification. Since all our GP-based models yield very similar results, we report only the outcomes obtained by the DGP-MF* model, as shown in Table 2.

Figure 2 presents a spatial map comparing the IAP2 observations (first column) with DGP* model predictions (middle column) and their associated uncertainty measures (third column) for various outputs. The results correspond to the SSP3 scenario for the 2050s, focusing on the *Biodiversity Vulnerable Index*, *Timber Production*, and *Area of Artificial Surface*. Overall, our model adequately captures the patterns across most EU regions for the three output variables. However, the model's performance varies by region and output. For instance, the uncertainty level is particularly high in Northeastern Europe regarding the *Biodiversity Vulnerable Index*, where the DGP model tends to underpredict. In contrast, predictions are more stable in Central and Southwestern Europe, characterised by relatively low uncertainty levels. For the *Timber Production*, the DGP model demonstrates higher predictability in Northern Europe, showing low uncertainty compared to other regions. A similar trend is observed for the *Area of Artificial Surface*, which is more challenging to predict in Central Europe and the South of England, where uncertainty levels are generally high. Figure A2 in the Appendix provides similar results for other selected outputs. Notably, there is considerable uncertainty (variance) in certain areas, particularly in the eastern regions and some parts of the continent, especially concerning irrigation usage, which requires further investigation. Due to page limitations, additional results from our analysis are provided in Appendix A5. A detailed comparison of the performance across each benchmark dataset is presented in Table A3. Also, Figure A1 shows the performance of six DGP and DPSS models on another 1D regression dataset characterised by heteroskedastic noise. The results of the Shapley analysis, illustrated in Figure A3, highlight the relative contributions of various input variables. This analysis aims to identify the subset of input variables that accounts for the majority of the variability in the output variables. Our findings indicate that average winter and autumn solar radiation, as well as maximum temperatures during these seasons, are the primary contributors to the output variability. In contrast, annual temperature and seasonal precipitation have a relatively minimal impact on the model output. We have employed several climate models under different scenarios to train our DGP model. Our methodology combines various sources of uncertainty, including input parameters, structural uncertainty, and observational error.

5. Conclusion

We have described a new approach for performing uncertainty analysis using the DGP model with a PLL objective. Our experiments span diverse datasets, including those derived from real-world applications, particularly in assessing uncertainty surrounding the IAP2. This platform models the effects of climate change and socio-economic shifts across crucial sectors like agriculture, forestry, water resources, biodiversity, and urban development. We broaden our approach to explore different types of variational inference for estimating DGP parameters. For example, DSPP experiments exclusively utilise mean field covariance approximation. We investigated the effects of applying a full Cholesky decomposition of the covariance function to enhance optimisation characteristics. Also, we assessed the effectiveness of alternative variational methods, including ODGP and SVGP approximation (baseline models). Shapley analysis indicate that average solar radiation in winter and autumn, along with maximum temperatures during these seasons, contribute significantly to the variability observed in IAP2 data. In contrast, annual temperature and seasonal precipitation have a relatively minor impact on model output. We note that the existing DGP model variants that use quadrature approximations instead of MC samples add more parameters and computational complexity to the training. Our experiments showed that approximating expectations with enough samples leads to convergence and performs relatively well for uncertainty quantification. While the DSPP outperforms other methods on many benchmark datasets which is likely due to greater flexibility in approximation, this study also enhances understanding of climate change projections by providing well-calibrated uncertainties. We did not simplify uncertainties into epistemic and aleatory sources but created a common model variant for data analysis. The code to reproduce this analysis is located on GitHub: github.com/wolemi2/UncertaintyWithDGP.

References

[1] E. Smith, "Uncertainty analysis," *Encyclopedia of environmetrics*, vol. 4, pp. 2283–2297, 2002.

[2] A. Saltelli, K. Aleksankina, W. Becker, P. Fennell, F. Ferretti, N. Holst, S. Li, and Q. Wu, "Why so many published sensitivity analyses are false: A systematic review of sensitivity analysis practices," *Environmental modelling & software*, vol. 114, pp. 29–39, 2019.

[3] O. K. Oyebamiji, C. Nemeth, P. A. Harrison, R. W. Dunford, and G. Cojocaru, "Multivariate sensitivity analysis for a large-scale climate impact and adaptation model," *Journal of the Royal Statistical Society Series C: Applied Statistics*, vol. 72, no. 3, pp. 770–808, 2023.

[4] F. Gamboa, P. Gremaud, T. Klein, and A. Lagnoux, "Global sensitivity analysis: A novel generation of mighty estimators based on rank statistics," *Bernoulli*, vol. 28, no. 4, 2022.

[5] C. K. Wikle and A. Zammit-Mangion, "Statistical deep learning for spatial and spatiotemporal data," *Annual Review of Statistics and Its Application*, vol. 10, pp. 247–270, 2023.

[6] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "Leveraging Bayesian deep learning and ensemble methods for uncertainty quantification in image classification: A ranking-based approach," *Heliyon*, vol. 10, no. 2, 2024.

[7] W. Chen and Y. Li, "Calibrating transformers via sparse Gaussian processes," *arXiv preprint arXiv:2303.02444*, 2023.

[8] O. K. Oyebamiji, D. Wilkinson, B. Li, P. G. Jayathilake, P. Zuliani, and T. Curtis, "Bayesian emulation and calibration of an individual-based model of microbial communities," *Journal of Computational Science*, vol. 30, pp. 194–208, 2019.

[9] O. K. Oyebamiji, N. R. Edwards, P. B. Holden, P. H. Garthwaite, S. Schaphoff, and D. Gerten, "Emulating global climate change impacts on crop yields," *Statistical Modelling*, vol. 15, no. 6, pp. 499–525, 2015.

[10] O. Oyebamiji, D. Wilkinson, P. Jayathilake, T. Curtis, S. Rushton, B. Li, and P. Gupta, "Gaussian process emulation of an individual-based model simulation of microbial communities," *Journal of Computational Science*, vol. 22, pp. 69–84, 2017.

[11] M. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," in *Artificial intelligence and statistics*, pp. 567–574, PMLR, 2009.

[12] A. G. d. G. Matthews, J. Hensman, R. Turner, and Z. Ghahramani, "On sparse variational methods and the kullback-leibler divergence between stochastic processes," in *Artificial Intelligence and Statistics*, pp. 231–239, PMLR, 2016.

[13] S. Sun, J. Zhao, and J. Zhu, "A review of Nyström methods for large-scale machine learning," *Information Fusion*, vol. 26, pp. 36–48, 2015.

[14] C. K. Williams and C. E. Rasmussen, "Gaussian processes for regression," 1996.

[15] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," *Journal of the American statistical Association*, vol. 112, no. 518, pp. 859–877, 2017.

[16] J. Gardner, G. Pleiss, K. Q. Weinberger, D. Bindel, and A. G. Wilson, "Gpytorch: Blackbox matrix-matrix Gaussian process inference with gpu acceleration," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[17] M.-C. Popescu, V. E. Balas, L. Perescu-Popescu, and N. Mastorakis, "Multilayer perceptron and neural networks," *WSEAS Transactions on Circuits and Systems*, vol. 8, no. 7, pp. 579–588, 2009.

[18] H. Salimbeni and M. Deisenroth, "Doubly stochastic variational inference for deep Gaussian processes," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] M. Bauer, M. Van der Wilk, and C. E. Rasmussen, "Understanding probabilistic sparse Gaussian process approximations," *Advances in Neural Information Processing Systems*, vol. 29, 2016.

[20] M. Raissi, "Parametric Gaussian process regression for big data," *arXiv preprint arXiv:1704.03144*, 2017.

[21] M. Jankowiak, G. Pleiss, and J. Gardner, "Parametric Gaussian process regressors," in *International Conference on Machine Learning*, pp. 4702–4712, PMLR, 2020.

[22] M. Jankowiak, G. Pleiss, and J. Gardner, "Deep sigma point processes," in *Conference on uncertainty in artificial intelligence*, pp. 789–798, PMLR, 2020.

[23] M. M. Dunlop, M. A. Girolami, A. M. Stuart, and A. L. Teckentrup, "How deep are deep Gaussian processes?," *Journal of Machine Learning Research*, vol. 19, no. 54, pp. 1–46, 2018.

[24] A. Damianou, *Deep Gaussian processes and variational propagation of uncertainty.* PhD thesis, University of Sheffield, 2015.

[25] A. Damianou and N. Lawrence, "Deep Gaussian processes," in *Artificial Intelligence and Statistics*, pp. 207–215, 2013.

[26] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *Journal of Machine Learning Research*, 2013.

[27] J. M. Hernández-Lobato, M. W. Hoffman, and Z. Ghahramani, "Scalable Bayesian optimization using deep gaussian processes," in *International Conference on Machine Learning*, pp. 1575–1584, 2016.

[28] J. Knoblauch, "Robust deep Gaussian processes," *arXiv preprint arXiv:1904.02303*, 2019.

[29] Y. Li and R. E. Turner, "Rényi divergence variational inference," *Advances in neural information processing systems*, vol. 29, 2016.

[30] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015.

[31] H. Salimbeni, V. Dutordoir, J. Hensman, and M. Deisenroth, "Deep Gaussian processes with importance-weighted variational inference," in *International Conference on Machine Learning*, pp. 5589–5598, PMLR, 2019.

[32] J. E. Oakley and A. O'Hagan, "Probabilistic sensitivity analysis of complex models: a Bayesian approach," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 66, no. 3, pp. 751–769, 2004.

[33] Y. Nohara, K. Matsumoto, H. Soejima, and N. Nakashima, "Explanation of machine learning models using shapley additive explanation and application for real data in hospital," *Computer Methods and Programs in Biomedicine*, vol. 214, p. 106584, 2022.

[34] H. Lin, L. Li, Y. Qiang, Y. Zhang, S. Liang, X. Xu, H. Li, and S. Hu, "Sensitivity analysis of slope stability based on extreme gradient boosting and shapley additive explanations: An exploratory study," *Heliyon*, vol. 10, no. 16, 2024.

[35] C.-A. Cheng and B. Boots, "Variational inference for Gaussian process models with linear complexity," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[36] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," in *Conference on Uncertainty in Artificial Intelligence*, pp. 282–290, 2013.

[37] G. Pleiss and J. P. Cunningham, "The limitations of large width in neural networks: A deep Gaussian process perspective," *Advances in Neural Information Processing Systems*, vol. 34, pp. 3349–3363, 2021.

[38] D. P. Kingma, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[39] J. Hensman, N. Fusi, and N. D. Lawrence, "Gaussian processes for big data," *arXiv preprint arXiv:1309.6835*, 2013.

[40] P. Goovaerts, R. Webster, and J.-P. Dubois, "Assessing the risk of soil contamination in the swiss jura using indicator geostatistics," *Environmental and ecological Statistics*, vol. 4, pp. 49–64, 1997.

[41] M. A. Alvarez and N. D. Lawrence, "Computationally efficient convolved multiple output Gaussian processes," *The Journal of Machine Learning Research*, vol. 12, pp. 1459–1500, 2011.

[42] P. A. Harrison, R. W. Dunford, I. P. Holman, and M. D. Rounsevell, "Climate change impact modelling needs to include cross-sectoral interactions," *Nature climate change*, vol. 6, no. 9, p. 885, 2016.

[43] P. A. Harrison, R. W. Dunford, I. P. Holman, G. Cojocaru, M. S. Madsen, P.-Y. Chen, S. Pedde, and D. Sandars, "Differences between low-end and high-end climate change impacts in europe across multiple sectors," *Regional environmental change*, vol. 19, no. 3, pp. 695–709, 2019.

[44] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul, "An introduction to variational methods for graphical models," *Machine learning*, vol. 37, pp. 183–233, 1999.

[45] L. S. Tan, "Analytic natural gradient updates for cholesky factor in Gaussian variational approximation," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, p. qkaf001, 2025.

[46] T. Tanaka, "A theory of mean field approximation," *Advances in Neural Information Processing Systems*, vol. 11, 1998.

[47] W. Han and Y. Yang, "Statistical inference in mean-field variational bayes," *arXiv preprint arXiv:1911.01525*, 2019.

[48] H. Salimbeni, C.-A. Cheng, B. Boots, and M. Deisenroth, "Orthogonally decoupled variational Gaussian processes," *Advances in neural information processing systems*, vol. 31, 2018.

Appendix

A1. Derivation of L_{svgp} objective

Let the variational posterior $q(\mathbf{f}, \mathbf{u}) \equiv p(\mathbf{f} \mid \mathbf{u}; \mathbf{X}, \mathbf{Z}) q(\mathbf{u})$, where $q(\mathbf{u}) = \mathcal{N}(\mathbf{u} \mid \mathbf{m}, \mathbf{S})$. Both terms in the variational posterior are Gaussian, we can analytically marginalise $\mathbf{u}$, which yields

$$q(\mathbf{f} \mid \mathbf{m}, \mathbf{S}; \mathbf{X}, \mathbf{Z}) = \int p(\mathbf{f} \mid \mathbf{u}; \mathbf{X}, \mathbf{Z}) q(\mathbf{u}) d\mathbf{u} = N(\mathbf{f} \mid \mu_f, \Sigma_f) \,.$$

with

$$\begin{aligned} \mu_{\mathbf{f}}(\mathbf{x}_i) &= \alpha(\mathbf{x}_i)^T \mathrm{m} \\ \Sigma_{\mathbf{f}}(\mathbf{x}_i) &= k(\mathbf{x}_i, \mathbf{x}_j) - \alpha(\mathbf{x}_i)^\top (k(\mathbf{Z}, \mathbf{Z}) - \mathbf{S}) \alpha(\mathbf{x}_j) \,, \\ &= \tilde{\mathbf{K}}_{ii} + \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} \mathbf{S} \mathbf{K}_{MM}^{-1} \mathbf{k}_i \end{aligned}$$

where $\alpha(\mathbf{x}_i) = k(\mathbf{Z}, \mathbf{Z})^{-1} k(\mathbf{Z}, \mathbf{x}_i)$, $k_i = k(x_i, \mathbf{Z})$, $K_{MM} = k(\mathbf{Z}, \mathbf{Z})$, $K_{NM} = k(\mathbf{X}, \mathbf{Z})$, and

$$\tilde{\mathbf{K}}_{NN} = K_{NN} - K_{NM} K_{MM}^{-1} K_{MN}.$$

Define

$$\mathcal{L}_{\text{svgp}} = \sum_{i=1}^{N} \mathrm{E}_{q(f_i|.)} \left| \log p(y_i \mid f_i) \right| - \mathrm{KL}[q(\mathbf{u})] \,||\, p(\mathbf{u})|$$

$$
\begin{aligned}
&= \sum_{i=1}^{N} \log \mathcal{N}\left(y_i \mid \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} m, \sigma_{\text{obs}}^2\right) - \frac{1}{2\sigma_{\text{obs}}^2} \operatorname{Tr} \tilde{\mathbf{K}}_{NN} \\
&- \frac{1}{2\sigma_{\text{obs}}^2} \sum_{i=1}^{N} \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} \mathbf{S} \mathbf{K}_{MM}^{-1} \mathbf{k}_i - \mathrm{KL}(q(\mathbf{u}) \,||\, p(\mathbf{u}))
\end{aligned} \tag{A1}
$$

where KL denotes the Kullback-Leibler divergence, (A1) can be rewritten more compactly as

$$
\begin{aligned}
\mathcal{L}_{\text{svgp}} \; = & \sum_{i=1}^{N} \left\{ \log \mathcal{N}\left(y_i \mid \mu_{\mathbf{f}}\left(\mathbf{x}_i\right), \sigma_{\text{obs}}^2\right) - \frac{\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right)}{2\sigma_{\text{obs}}^2} \right\} \\
& - \mathrm{KL}(q(\mathbf{u}) \,||\, p(\mathbf{u}))
\end{aligned} \tag{A2}
$$

where $\mu_{\mathrm{f}}\left(\mathrm{x}_i\right)$ is the predictive mean function given by

$$
\mu_{\mathrm{f}}\left(\mathbf{x}_i\right) = \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} m \tag{A3}
$$

and where $\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right) \equiv \operatorname{Var}\left[\mathrm{f_i} \mid \mathbf{x_i}\right]$ denotes the latent function variance such that

$$
\Sigma_{\mathbf{f}}\left(\mathbf{x}_i\right) = \tilde{\mathbf{K}}_{ii} + \mathbf{k}_i^T \mathbf{K}_{MM}^{-1} \mathbf{S} \mathbf{K}_{MM}^{-1} \mathbf{k}_i \tag{A4}
$$

$\mathcal{L}_{\text{svgp}}$, which depends on $m, \mathbf{S}, \mathbf{Z}, \sigma_{\text{obs}}$ and the various kernel hyperparameters θ_{ker} which can then be maximised with gradient methods. Further details is given in [20, 39].

A2. Derivation of ELBO objective for DGPs

Following from section 2.2, we can utilise sparse variational inference to simplify the correlations within layers while maintaining correlations between them, which preserves an exact model conditioned on U^l. Suppose that the posterior distribution of $\{U^l\}_{l=1}^L$ is factorised between layers. Then the posterior can be expressed as

$$
q(\{\mathbf{F}^i, \mathbf{U}^l\}_{l=1}^L) = \prod_{l=1}^{L} p(\mathbf{F}^i \mid \mathbf{U}^i; \mathbf{F}^{l-1}, \mathbf{Z}^{l-1}) q(\mathbf{U}^l)\,. \tag{A5}
$$

Let $q(\mathbf{U}^l)$ be a Gaussian with mean $\mathbf{m}^l$ and variance $\mathbf{S}^l$ and similar to single-layer GP, we can marginalise the inducing variables from each layer analytically:

$$
q(\{\mathbf{F}^l\}_{l=1}^L) = \prod_{l=1}^{L} q(\mathbf{F}^l \mid \mathbf{m}^l, \mathbf{S}^l, \mathbf{F}^{l-1}, \mathbf{Z}^{l-1}) = \prod_{l=1}^{L} \mathcal{N}(\mathbf{F}^l \mid \mu^l, \mathbf{\Sigma}^l)\,,
$$

where $[\hat{\mu}^l]_i = {\mu_{\mathrm{m}}}^l, \mathbf{z}^{l-1}(\mathbf{f}_i^l)$ and $[\hat{\Sigma}^l]_{ij} = \Sigma_{\mathrm{S}^l, \Sigma^{l-1}}(f_i^l, f_j^l)$. Note: f_i^l is the i^{th} row of F^l.

$$
q(\mathbf{f}_i^L) = \int \prod_{l=1}^{L-1} q(\mathbf{f}_i^l \mid \mathbf{m}^l, \mathbf{S}^l; \mathbf{f}_i^{l-1}, \mathbf{Z}^{l-1}) d\mathbf{f}_i^l\,.
$$

Therefore, we can first sample $\epsilon_i^l \sim N(0, \mathbf{I}_{D^l})$ and then recursively draw the sampled variables $\hat{f}_i^l \sim q(\hat{f}_i^l \mid \mathbf{m}^l, \mathbf{S}^l, \hat{f}_i^{l-1}, Z^{l-1})$ for $l = 1, \cdot\cdot, L-1$ as

$$
\hat{f}_i^l = \mu_{\mathrm{m}^l, \mathbf{Z}^{l-1}}(\hat{f}_i^{l-1}) + \epsilon_i^l \odot \sqrt{\Sigma_{\mathrm{S}^l, \mathbf{Z}^{l-1}}(\hat{f}_i^{l-1}, \hat{f}_i^{l-1})}\,,
$$

$$
\mathcal{L}_{DGP} = \mathbb{E}_{q(\{\mathbf{F}^l, \mathbf{U}^l)_{l=1}^L)} \left[\frac{p(\mathbf{Y}, \prime \{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)}{q(\{\mathbf{F}^l, \mathbf{U}^l\}_{l=1}^L)} \right].
$$

Substitute (2.2) and (A5), and rearrange

$$\mathcal{L}_{DGP} = \sum_{i=1}^{N} \mathbb{E}_{q(\mathbb{F}_i^L)} |\log p(\mathbf{y}_i \mid \mathbf{f}_i^L)| - \sum_{l=1}^{L} \mathrm{KL}[q(\mathbf{U}^l) \,||\, p(\mathbf{U}^l; \mathbf{Z}^{l-1})]\,. \tag{A6}$$

See further details in [18, 20, 21].

A3. Quadrature rules and derivation

Firstly, suppose a finite parametric family of mixture distributions was established, essentially using a sigma point approximation or a quadrature-like integration method for 2.6. If we denote the width of the first GP layer as $W = 2$, we can express this for an input $\mathbf{x}_i$ as follows:

$$p_{\mathrm{dspp}}\left(y_i \mid \mathbf{x}_i\right) = \int d\mathbf{g}_i \mathcal{N}\left(y_i \mid \mu_f\left(\mathbf{g}_i\right), \sigma_f\left(\mathbf{g}_i\right)^2 + \sigma_{\mathrm{obs}}^2\right) \prod_{w=1}^{2} q\left(g_{iw} \mid \mathbf{x}_i\right). \tag{A7}$$

To convert the (A7) above into a finite mixture, a straightforward method uses Gauss-Hermite quadrature. This approximate $q(g_{i1} \mid \boldsymbol{x}_i)$ using an S-component mixture of Dirac delta distributions governed by weights $\boldsymbol{\omega_1}^{(s)}$ and quadrature points $\boldsymbol{\xi}^{(s)}$. $q\left(g_{i1} \mid \mathbf{x}_i\right) = \sum_{s_1=1}^{S} \omega_1^{(s_1)} \delta\left(g_{i1} - \left(\mu_{g_1}\left(\mathbf{x}_i\right) + \xi_1^{(s_1)} \sigma_{g_1}\left(\mathbf{x}_i\right)\right)\right)$. Building on section 2.3, parametric regression models with an objective function that aligns with regularised maximum likelihood estimation can be defined. $\mathcal{L}_{\mathrm{dspp}} = \sum_{i=1}^{N} \log p_{\mathrm{dspp}}\left(y_i \mid \mathbf{x}_i\right) - \beta_{\mathrm{reg}} \sum \mathrm{KL}$ where $\beta_{\mathrm{reg}} > 0$ is an optional regularisation constant and $\sum \mathrm{KL}$ is a sum over KL divergences of inducing variables. This objective relies on σ_{obs} and the parameters $\boldsymbol{m}, \boldsymbol{S}, \boldsymbol{Z}$, along with various kernel hyperparameters for each GP, and can be optimised through stochastic gradient methods. The substitution below was made in the first GP layer with W width.

$$\prod_{w=1}^{W} q\left(g_{iw} \mid \mathbf{x}_i\right) \rightarrow \quad \sum_{s} \omega^{(s)} \prod_{w=1}^{W} \delta\left(g_{iw} - \left(\mu_{g_w}\left(\mathbf{x}_i\right) + \xi_w^{(s_w)} \sigma_{g_w}\left(\mathbf{x}_i\right)\right)\right),$$

where $s = (s_1, \ldots, s_W) \in \{1, \ldots, S\}^W$ is a multi-index. Unlike the Gauss-Hermite quadrature rule, the quadrature points $\left\{\xi_w^{(s_w)}\right\}$ are taken as learnable parameters, resulting in $S \times W$ real parameters in total. For more information, refer to [22].

A4. Dataset used for the analysis

Apart from the standard datasets from the UCI repository namely bikes, elevators, energy, 3Droad and protein, we also use the Jura and IAP2 data. The Jura dataset [40] is a multivariate regression data set comprised of measurements of concentrations of seven heavy metals (cadmium, cobalt, chromium, copper, nickel, lead, and zinc) recorded at 359 locations in the topsoil of a region in the Swiss Jura. Similar to [41], we are interested in predicting the concentrations of metals that are more costly to measure (primary variables) using measurements of metals that are less expensive to sample (secondary variables). In this study, cadmium, copper, and lead are regarded as target variables, while the remaining metals, along with landuse type, rock type, and the coordinates of each location, serve as predictive features.

IAP2 input and output variables

The IMPRESSIONS Integrated Assessment Platform (IAP2) models the impacts of climate and socio-economic change across key sectors (e.g. agriculture, forestry, water, biodiversity, urban development). It integrates climate (RCPs) and socio-economic (SSPs) scenarios to support adaptation decision-making. Operating at a 10×10 arcminute resolution across Europe, IAP2 consists of ten interlinked meta-models, but its computational intensity limits large-scale scenario analysis. It simulates landuse, food production,

carbon storage, water use, and flood risk for three future periods (2020s, 2050s, 2080s) under three emissions (RCPs 2.6, 4.5, 8.5) and four socio-economic scenarios (SSPs 1, 3, 4, 5) [42, 43]. In this paper, we select fifteen output data to demonstrate our techniques. The list of the output data is given in Table A2. The input variables are based on seasonal averages include winter average precipitation, spring average precipitation, summer average precipitation, autumn average precipitation and yearly average precipitation, similarly for radiation, temperature minimum, temperature maximum and absolute mean temperature as well as elevation data.

Table A1. Summary of data used for the analysis. The number of inducing point, learning rate, number of MC samples, number of quadrature sites for DSPP and number of iterations are set at 128, 0.01, 50, 8 and 400 respectively. All DGP and DSPP models have three hidden layers

Dataset	Elevators	Bike	Protein	3droad	Jura	Energy	IAP2
Total samples (N)	12449	13034	34297	326155	359	19735	286452
output dimension (m)	1	1	1	1	3	1	15
Batch size	1024	1024	1024	2048	128	1024	1024

A5. Further model results

This section presents additional results for the model summary for the benchmark dataset given in Table A3, 1D regression dataset with heteroskedastic noise comparing performance of six DGP and DPSS models in Figure A1, the spatial map showing the comparison between the DGP* model for other IAP2 outputs in A2. The Shapley analysis results illustrate the relative contributions of various input variables is also given in Figure A3.

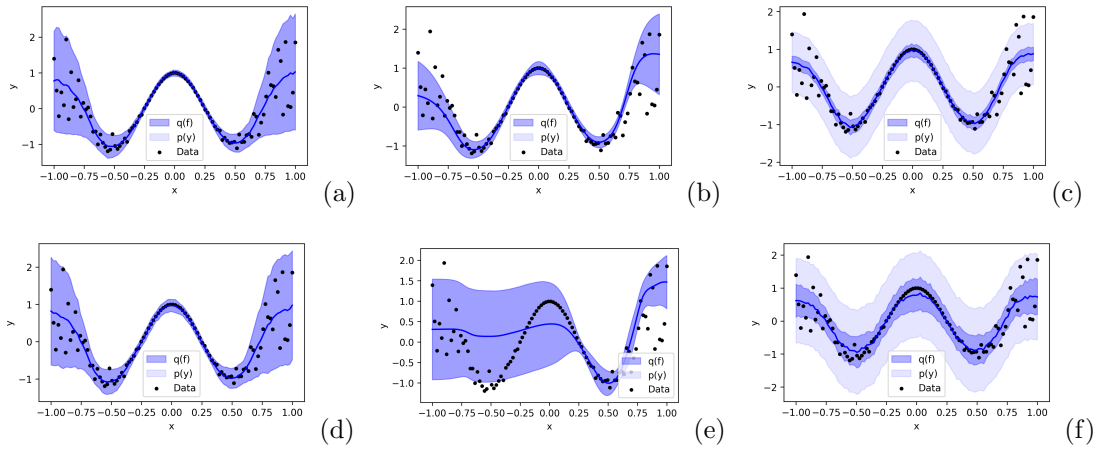

Figure A1. We assessed the performance of six different models applied to a 1D regression dataset that includes heteroskedastic noise. The solid lines represent the mean predictions, while the blue areas indicate the uncertainty bands. The darker blue highlights the input-dependent uncertainty, and the lighter blue represents the observation noise, denoted as σ^2_{obs}. We observed that the observation noise significantly contributed to the overall uncertainty in the DGP models (c) and (f), which were optimised using the VI ELBO objective. Let $x \in \mathbb{R}^{100}$ be linearly spaced in $[0, 1]$, and define: $y_i = \cos(2\pi x_i) + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}\left(0, x_i^3\right)$ for $i = 1, \ldots, 100$. (a) DGP-Chol*, (b) DPSS-CHOL, (c) DGP-Chol, (d) DGP-MF*, (e) DPSS-MF and (f) DGP-MF as defined in the main paper.

A6. Variational strategies

Variational methods encompass techniques designed to estimate posterior distributions within Bayesian inference. VI is a widely used computational strategy for approximating complex probability densities, particularly those involving intractable integrals and multiple latent variables found in intricate Bayesian

Table A2. Selected IAP2 output data for modelling.

Dataset		
Water exploitation index	Landuse intensity index	Intensive grassland
Landuse diversity	Biodiversity vulnerability index	Extensive grassland
Timber production	Area of artificial surfaces	Very extensive grassland
Potential carbon stock	Unmanaged forest	Unmanaged land
Irrigation usage	Arable land	Managed forest

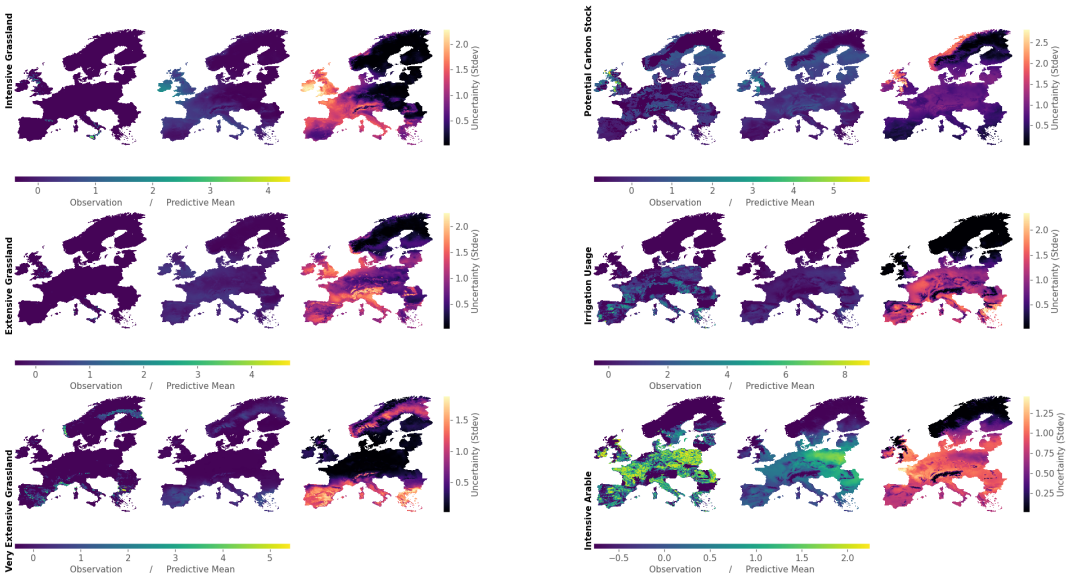

Figure A2. Spatial maps comparing selected model outputs for the SSP3 scenario in 2050, using IAP2 data and the DGP model.

hierarchical models. In this approach, the challenging target distribution is approximated by selecting a member from a specified set of tractable densities that is closest in terms of Kullback-Leibler (KL) divergence [44]. The predictive distribution for approximate GPs is generally expressed as

$$p\left(\mathbf{f}\left(\mathbf{x}^*\right)\right) = \int_{\mathbf{u}} p\left(f\left(\mathbf{x}^*\right) \mid \mathbf{u}\right) q(\mathbf{u}) d\mathbf{u}, \quad q(\mathbf{u}) = \mathcal{N}(\mathbf{m}, \mathbf{S}). \tag{A8}$$

The symbol $\boldsymbol{u}$ denotes the function values at the m inducing points, $\boldsymbol{m} \in \mathbf{R}^m$ and $\boldsymbol{S} \in \mathbf{R}^{m \times m}$ are parameters that can be learned. If m, representing the number of inducing point is large, the number of learnable parameters in $\mathbf{S}$ can become unwieldy. Also, a significant value of m can slow down some of the computations. This section briefly presents several methods for using various variational distributions and strategies to achieve this goal. The standard approach involves using a Cholesky decomposition [45]. The Cholesky variational distribution involves multivariate normal distribution with a full covariance matrix and permits $\mathbf{S}$ to be any positive semidefinite matrix. This is the most general and expressive option for approximate GPs. The variational approximation of the mean field [46] is a specific method within variational inference where the posterior distribution is estimated by assuming that latent variables are independent (diagonal). This assumption simplifies computations while optimising the variational objective, usually represented as ELBO [47].

Table A3. Summary of the Root Mean Square Error (RMSE), Negative Log-likelihood (NLL), Continuous Rank Probaility Score (CPRS) and Mean Absolute Error (MAE) for all models and benchmark dataset. Note: PPGPR-OD and PPGPR-Chol are the variants that use Orthogonally Decoupled and Cholesky full-rank covariance decomposition, respectively. ODSVGP and SVGP are also part of baseline models.

Dataset Model	3droad	Bike	Elevators	Energy	Jura	Protein
RMSE						
DSPP-Chol	0.457 ± 0.006	0.126 ± 0.008	0.354 ± 0.001	**0.034 ± 0.002**	0.676 ± 0.002	**0.605 ± 0.007**
DSPP-MF	**0.452 ± 0.007**	0.107 ± 0.004	**0.342 ± 0.000**	0.041 ± 0.002	0.714 ± 0.001	0.620 ± 0.010
DGP-Chol*	0.55 ± 0.001	0.088 ± 0.003	0.351 ± 0.001	0.072 ± 0.002	0.698 ± 0.007	0.743 ± 0.005
DGP-MF*	0.476 ± 0.004	0.098 ± 0.003	0.354 ± 0.001	0.081 ± 0.004	0.773 ± 0.006	0.73 ± 0.003
DGP-Chol	0.446 ± 0.006	0.053 ± 0.005	0.352 ± 0.001	0.277 ± 0.003	0.714 ± 0.007	0.75 ± 0.003
DGP-MF	0.466 ± 0.009	**0.050 ± 0.006**	0.35 ± 0.001	0.064 ± 0.001	0.767 ± 0.005	0.728 ± 0.004
ODSVGP	0.663 ± 0.008	0.24 ± 0.002	0.399 ± 0.001	0.213 ± 0.004	0.642 ± 0.001	0.788 ± 0.003
SVGP	0.500 ± 0.003	0.208 ± 0.005	0.394 ± 0.002	0.13 ± 0.018	0.647 ± 0.004	0.737 ± 0.006
PPGPR-OD	0.688 ± 0.008	0.224 ± 0.001	0.393 ± 0.001	0.221 ± 0.001	**0.595 ± 0.002**	0.817 ± 0.002
PPGPR-Chol	0.561 ± 0.002	0.273 ± 0.002	0.419 ± 0.002	0.22 ± 0.001	0.624 ± 0.001	0.733 ± 0.005
NLL						
DSPP-Chol	**-0.129 ± 0.036**	-2.742 ± 0.325	0.449 ± 0.038	**-2.021 ± 0.043**	2.472 ± 0.078	**0.289 ± 0.026**
DSPP-MF	-0.042 ± 0.032	**-2.771 ± 0.076**	**0.305 ± 0.006**	-1.892 ± 0.022	3.200 ± 0.100	0.384 ± 0.021
DGP-Chol*	0.418 ± 0.01	-2.673 ± 0.055	0.355 ± 0.002	-1.552 ± 0.022	**2.366 ± 0.063**	1.062 ± 0.008
DGP-MF*	0.304 ± 0.019	-2.657 ± 0.026	0.361 ± 0.002	-1.258 ± 0.055	3.139 ± 0.042	1.034 ± 0.003
DGP-Chol	0.615 ± 0.013	-1.536 ± 0.106	0.375 ± 0.002	0.400 ± 0.009	3.239 ± 0.024	1.139 ± 0.008
DGP-MF	0.657 ± 0.018	-1.575 ± 0.108	0.37 ± 0.002	-1.337 ± 0.023	3.51 ± 0.02	1.102 ± 0.006
ODSVGP	1.01 ± 0.013	-0.009 ± 0.009	0.551 ± 0.008	-0.02 ± 0.026	3.166 ± 0.016	1.181 ± 0.004
SVGP	0.733 ± 0.006	0.107 ± 0.033	0.521 ± 0.017	-0.513 ± 0.099	2.687 ± 0.037	1.118 ± 0.007
PPGPR-OD	0.939 ± 0.026	-0.371 ± 0.029	0.486 ± 0.002	-0.159 ± 0.011	2.766 ± 0.046	1.216 ± 0.002
PPGPR-Chol	0.459 ± 0.006	-1.755 ± 0.023	0.391 ± 0.003	-1.111 ± 0.011	2.035 ± 0.009	0.96 ± 0.008
CRPS						
DSPP-Chol	0.231 ± 0.004	0.027 ± 0.006	0.200 ± 0.001	0.018 ± 0.001	**0.318 ± 0.000**	**0.329 ± 0.006**
DSPP-MF	**0.224 ± 0.005**	0.024 ± 0.001	**0.189 ± 0.001**	0.021 ± 0.001	0.343 ± 0.001	0.358 ± 0.010
DGP-Chol*	0.269 ± 0.001	**0.021 ± 0.001**	0.193 ± 0.000	0.034 ± 0.001	0.336 ± 0.004	0.400 ± 0.003
DGP-MF*	0.233 ± 0.003	0.022 ± 0.001	0.195 ± 0.000	0.042 ± 0.002	0.391 ± 0.004	0.396 ± 0.002
DGP-Chol	0.243 ± 0.003	0.024 ± 0.003	0.195 ± 0.000	0.175 ± 0.002	0.380 ± 0.003	0.423 ± 0.002
DGP-MF	0.254 ± 0.005	0.023 ± 0.003	0.193 ± 0.000	0.035 ± 0.001	0.414 ± 0.003	0.406 ± 0.003
ODSVGP	0.365 ± 0.006	0.125 ± 0.001	0.222 ± 0.001	0.123 ± 0.003	0.361 ± 0.001	0.451 ± 0.002
SVGP	0.273 ± 0.002	0.125 ± 0.004	0.221 ± 0.002	0.076 ± 0.009	0.331 ± 0.003	0.419 ± 0.004
PPGPR-OD	0.368 ± 0.006	0.102 ± 0.001	0.218 ± 0.000	0.117 ± 0.001	0.328 ± 0.003	0.468 ± 0.001
PPGPR-Chol	0.275 ± 0.001	0.072 ± 0.001	0.215 ± 0.001	0.082 ± 0.000	0.300 ± 0.001	0.392 ± 0.003
MAE						
DSPP-Chol	**0.310 ± 0.005**	0.035 ± 0.007	0.272 ± 0.001	**0.024 ± 0.002**	0.432 ± 0.001	**0.419 ± 0.009**
DSPP-MF	0.304 ± 0.007	0.031 ± 0.002	**0.263 ± 0.001**	0.029 ± 0.001	0.444 ± 0.001	0.455 ± 0.013
DGP-Chol*	0.383 ± 0.001	**0.029 ± 0.002**	0.269 ± 0.001	0.048 ± 0.001	**0.435 ± 0.004**	0.576 ± 0.005
DGP-MF*	0.330 ± 0.003	0.030 ± 0.001	0.272 ± 0.001	0.059 ± 0.003	0.543 ± 0.005	0.570 ± 0.004
DGP-Chol	0.324 ± 0.005	0.029 ± 0.003	0.269 ± 0.001	0.202 ± 0.002	0.501 ± 0.005	0.595 ± 0.004
DGP-MF	0.339 ± 0.007	0.030 ± 0.003	0.267 ± 0.001	0.049 ± 0.001	0.544 ± 0.004	0.568 ± 0.004
ODSVGP	0.501 ± 0.010	0.158 ± 0.002	0.304 ± 0.001	0.152 ± 0.003	0.458 ± 0.001	0.649 ± 0.003
SVGP	0.359 ± 0.003	0.118 ± 0.004	0.301 ± 0.001	0.099 ± 0.012	0.443 ± 0.004	0.594 ± 0.006
PPGPR-OD	0.523 ± 0.008	0.133 ± 0.001	0.301 ± 0.001	0.159 ± 0.001	0.435 ± 0.003	0.676 ± 0.002
PPGPR-Chol	0.394 ± 0.002	0.096 ± 0.001	0.306 ± 0.001	0.126 ± 0.001	0.421 ± 0.001	0.574 ± 0.005

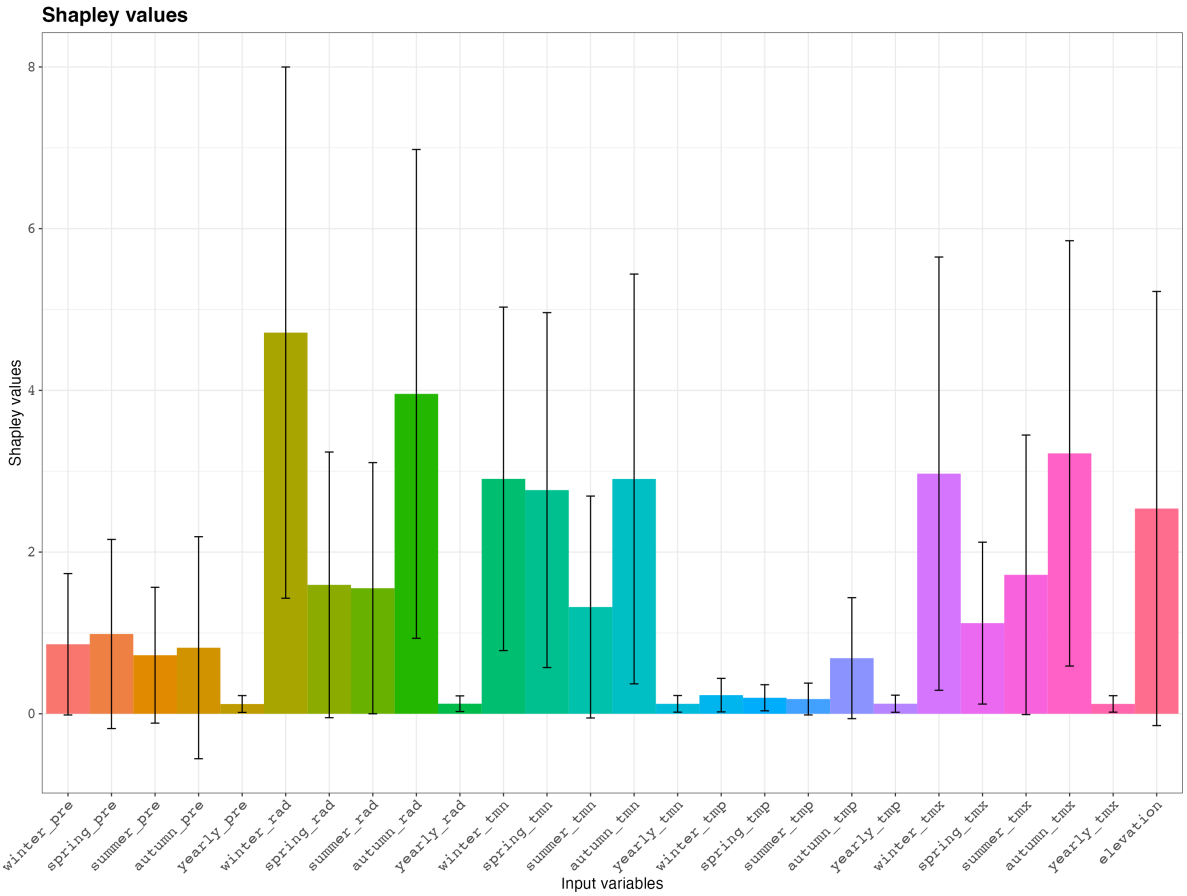

Figure A3. Aggregated Shapley indices for the results of the IAP2 model.

[35] introduce the concept of decoupled GPs. The central idea is to depict the variational distribution within the reproducing kernel Hilbert space (RKHS) derived from the GP's covariance function. This approach utilises a greater number of inducing points for the computationally efficient mean while using fewer for the computationally intensive covariance. This parameterisation separates the variational parameters for the mean (μ) from those for the covariance (Σ), leading to independent computation complexities for both components. Consequently, a larger set of parameters can be employed for the mean to represent more intricate functions. [48] build upon this by introducing orthogonally decoupled variational GPs, a class of variational inference methods designed to enhance the flexibility and efficiency of approximate posterior distributions.