SIAM/ASA J. Uncertainty Quantification
Vol. 12, No. 4, pp. 1085–1112

Hierarchical Shrinkage Gaussian Processes: Applications to Computer Code Emulation and Dynamical System Recovery*

Tao Tang†, Simon Mak‡, and David Dunson†‡

Abstract. In many areas of science and engineering, computer simulations are widely used as proxies for physical experiments, which can be infeasible or unethical. Such simulations are often computationally expensive, and an emulator can be trained to efficiently predict the desired response surface. A widely used emulator is the Gaussian process (GP), which provides a flexible framework for efficient prediction and uncertainty quantification. Standard GPs, however, do not capture structured shrinkage on the underlying response surface, which is present in many applications, particularly in the physical sciences. We thus propose a new hierarchical shrinkage GP (HierGP), which incorporates such structure via cumulative shrinkage priors within a GP framework. We show that the HierGP implicitly embeds the principles of effect hierarchy, heredity, and smoothness widely used for analysis of experiments; such principles allow the HierGP to identify significant structured effects on the response surface with limited data. We propose efficient posterior sampling algorithms for model training and prediction and prove desirable consistency properties for the HierGP. Finally, we demonstrate the improved performance of HierGP over existing models in a suite of numerical experiments and an application to dynamical system recovery.

Key words. computer experiments, dynamical recovery, emulation, Gaussian processes, shrinkage priors, uncertainty quantification

MSC code. 62F15

DOI. 10.1137/23M1550682

1. Introduction.

Scientific computing is playing an increasingly important role in solving modern scientific and engineering problems, particularly with recent breakthroughs in mathematical modeling and computation. Quantities of interest that were difficult or infeasible to observe from physical experiments can now be reliably simulated via computer code. Such *computer experiments* typically involve solving a system of differential equations based on physical models, and have had a wide-reaching impact in many fields, including rocket engine design (Mak et al., 2018), personalized surgical planning (Chen et al., 2021), and universe expansions (Kaufman et al., 2011; Ji et al., 2023). One critical bottleneck, however, is that these virtual experiments can be highly resource-intensive for computation. For example, the full-scale simulation of a single rocket engine injector can require millions of CPU hours

*Received by the editors February 1, 2023; accepted for publication (in revised form) April 26, 2024; published electronically October 3, 2024.
https://doi.org/10.1137/23M1550682
Funding: The work of the first author was supported by the Statistical and Applied Mathematical Sciences Institute. The work of the second author was supported by NSF grants CSSI-2004571, DMS-2210729, DMS-2316012, and DE-SC0024477. The work of the third author was supported by Office of Naval Research (ONR) project N00014-21-1-2510 and European Research Council (ERC) project 856506.
†Department of Mathematics, Duke University, Durham, NC 27708 USA (tt166@duke.edu, dunson@duke.edu).
‡Department of Statistical Science, Duke University, Durham, NC 27708-0251 USA (sm769@duke.edu).

(Yeh et al., 2018), which places heavy demand on computational resources for engine design exploration and optimization.

One solution is to train an *emulator* model that can efficiently predict (or *emulate*) the expensive computer code $f(\boldsymbol{x})$ at different parameters $\boldsymbol{x}$. The idea is to run the computer code at carefully chosen points over the parameter space $\mathcal{X}$ (normalized to $\mathcal{X} = [0,1]^d$) and then use this as training data to fit the emulator model. A popular emulator is the Gaussian process (GP) (see Rasmussen and Williams, 2005)), a flexible Bayesian model for probabilistic predictive modeling. GPs have two key advantages for computer code emulation (Gramacy, 2020): they offer closed-form expressions for prediction and uncertainty quantification of the emulator, and they provide a flexible nonparametric framework for modeling the black-box response surface.

Despite this, standard GPs have several limitations for emulation. One limitation is that, when the training sample size n is small, the highly flexible form of a GP can become more of a vice than a virtue. In particular, with a limited sample size, the GP posterior may distribute probability across an overly broad range of functions, resulting in poor predictive performance and high predictive uncertainty. A promising solution is to integrate prior knowledge on the response surface f (e.g., via domain knowledge from applications) within the GP. This knowledge can take the form of mechanistic models (Wheeler et al., 2014), boundary conditions (Ding et al., 2019), or shape constraints (Golchi et al., 2015; Wang and Berger, 2016), which can greatly improve the predictive performance of a GP with limited data. However, for highly complex systems, such prior knowledge may be difficult to elicit with confidence and too complex to integrate for probabilistic modeling. In lieu of this, alternate models are needed to learn useful structure for prediction in data-limited settings. Relevant existing methods in this vein include GPs with variable selection (Savitsky et al., 2011), graph Laplacian GPs (Dunson et al., 2022), and GPs with embedded manifolds (Seshadri et al., 2019; Li et al., 2023).

A useful clue for physical systems is that its response surfaces, albeit complex, often exhibit *shrinkage,* in that nearly all variation in the surface is dictated by a small number of effects derived from input parameters. This shrinkage is supported by the well-known Buckingham-π theorem (Buckingham, 1914) and extensive literature in experimental physics (see, e.g., Berkooz et al., 1993). Furthermore, such shrinkage is often highly *structured,* satisfying the statistical principles of *effect hierarchy* and *heredity* (Wu and Hamada, 2009; Haris et al., 2016) and the marginality principle (Nelder, 1977); such principles have been widely used for analyzing *limited* data from physical experiments. Here, effect hierarchy presumes that main effects typically have greater influence than interaction effects, and effect heredity (specifically, *strong* effect heredity; more on this later) presumes that such interactions are significant only when its component main effects are significant. This structured shrinkage behavior is not accommodated by existing GPs with variable selection (Savitsky et al., 2011), nor GPs that separate main effects and pairwise interactions (Ferrari and Dunson, 2021). A recent paper (Ding et al., 2020) further noted the principle of *effect smoothness,* which presumes that lower-order (i.e., smoother) basis functions are likely more influential than higher-order basis functions. We aim to carefully embed such principles within the GP, thus providing a flexible and data-driven framework for predictive modeling with limited training data.

We present next the proposed hierarchical shrinkage GP (HierGP), which embeds the aforementioned structured shrinkage principles (effect hierarchy, heredity, and smoothness) via

carefully designed shrinkage priors within a GP framework. The HierGP begins with a basis expansion of a GP and then assigns hierarchical shrinkage priors on basis coefficients to capture the desired structured shrinkage behavior. In particular, we extend the cumulative shrinkage priors proposed in Legramanti et al. (2020) within a GP and show that the resulting HierGP indeed captures the desired properties of effect hierarchy, heredity, and smoothness. We then propose a Gibbs sampler that leverages a data augmentation trick for efficient posterior computation. Under mild conditions on sparsity, we prove posterior contraction results for the HierGP under both fixed and randomly sampled design points. Finally, we demonstrate the effectiveness of the proposed HierGP over existing models in a suite of numerical experiments and an application to dynamical system recovery.

The paper is structured as follows. Section 2 introduces the proposed HierGP. Section 3 presents the data-augmented Gibbs sampler for posterior sampling. Section 4 briefly outlines posterior consistency results for the HierGP. Section 5 reports numerical experiments and an application to dynamical system recovery. Section 6 concludes the paper.

2. Model specification.

In this section, we describe the HierGP and discuss connections with existing GP models. We first review standard GPs and their representation as an infinite basis expansion with random coefficients. We then propose the HierGP as an extension of this basis representation with carefully specified shrinkage priors that capture the desired structured shrinkage properties.

2.1. Gaussian process modeling.

In what follows, we let $f(\boldsymbol{x})$ denote the expensive black-box function to be emulated, where $\boldsymbol{x} \in \mathbb{R}^d$ are its input parameters. A GP (Rasmussen and Williams, 2005; Gramacy, 2020) adopts the following probabilistic prior on $f(\cdot)$:

$$f(\cdot) \sim \mathrm{GP}\{\mu(\cdot), \gamma(\cdot,\cdot)\}. \tag{2.1}$$

Here, $\mu(\cdot)$ is the mean function of the process and $\gamma(\cdot,\cdot)$ its covariance function. A key appeal for GP modeling is that, conditional on observed data $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ from the black-box system, the posterior predictive distribution $[f(\cdot)|f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)]$ remains a GP with closed-form expressions for its posterior mean and covariance. This facilitates prediction and uncertainty quantification via a flexible Bayesian nonparametric model.

In order to integrate the desired structured shrinkage properties, we will employ an alternate representation of the GP as an infinite basis expansion. This relies on the well-known Karhunen–Loève theorem (stated below), which shows that a GP can be represented as an infinite basis expansion of orthonormal functions.

Theorem 2.1 (Karhunen–Loève; Theorem 5.3 of Alexanderian, 2015). *Let $f(\boldsymbol{x}) \sim \mathrm{GP}\{0, \gamma(\cdot,\cdot)\}$ be a zero-mean Gaussian process, with covariance kernel γ continuous on $\mathcal{X} \times \mathcal{X}$ and $\gamma(\boldsymbol{x}, \boldsymbol{x}') \in L^2(\mathcal{X})$. Then there exists an orthonormal basis $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$ of $L^2(\mathcal{X})$ such that*

$$f(\boldsymbol{x}) = \sum_{|\boldsymbol{k}|=1}^{\infty} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{2.2}$$

where $\boldsymbol{k} = (k_1, \ldots, k_d) \in \mathbb{N}_0^d$ is a multi-index, the coefficients $\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^{\infty}$ are independent Gaussian random variables given by $\lambda_{\boldsymbol{k}} = \int_{\mathcal{X}} f(\boldsymbol{x})\phi_{\boldsymbol{k}}(\boldsymbol{x})d\boldsymbol{x}$ and satisfy $\mathbb{E}(\lambda_{\boldsymbol{k}}) = 0$ and $\mathbb{E}(\lambda_{\boldsymbol{j}}\lambda_{\boldsymbol{k}}) = \mathbb{I}\{\boldsymbol{j} = \boldsymbol{k}\}\mathrm{Var}(\lambda_{\boldsymbol{k}})$, where $\mathbb{I}\{\cdot\}$ is the indicator function.

Here, the conditions on kernel γ ensure the GP $f(\cdot)$ is mean-squared continuous, i.e., $f \in L^2(\mathcal{X} \times \Omega)$; see Alexanderian (2015). Equation (2.2) is also known as the Karhunen–Loève expansion, which is widely used in statistics (Wang, 2008) and uncertainty quantification (Xiu, 2010; Ghanem and Spanos, 1991). This expansion can also be viewed as a stochastic extension of the classical Fourier expansion.

As a tangible example, consider the case where $f(\cdot)$ follows the GP in (2.2) with mean function $\mu(\cdot) \equiv 0$ and isotropic squared-exponential covariance function:

$$\gamma(\boldsymbol{x}, \boldsymbol{x}') = \sigma^2 \exp\left\{-\frac{\|\boldsymbol{x} - \boldsymbol{x}'\|_2^2}{2l^2}\right\}. \tag{2.3}$$

Then one can show (see, e.g., Rasmussen and Williams, 2005) that the corresponding Karhunen–Loève decomposition of $f(\cdot)$ takes the form (2.2), with

$$\lambda_{\boldsymbol{k}} = \prod_{m=1}^{d} \left(\sqrt{\frac{2a}{A}} B^{k_m}\right) \xi_{\boldsymbol{k}}, \qquad \phi_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{m=1}^{d} \left(\exp\left\{-(c-a)x_m^2\right\} H_{k_m}(\sqrt{2c}x_m)\right), \tag{2.4}$$

where $\xi_{\boldsymbol{k}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, $H_k(\cdot)$ is the kth order Hermite polynomial (Xiu, 2010), with $a^{-1} = 4\sigma^2$, $b^{-1} = 2l^2$, $c = \sqrt{a^2 + 2ab}$, $A = a + b + c$, and $B = b/A$.

A careful inspection of the prior specification (2.4) shows that it captures the aforementioned principles of *effect smoothness* (Ding et al., 2020) and *effect hierarchy* (Hamada and Wu, 1992). Recall that effect smoothness presumes that lower-order (i.e., smoother) basis functions are likely more influential than higher-order basis functions. This is captured in (2.4): as the index k_m increases (with all other indices in $\boldsymbol{k}$ held constant), the independent Gaussian priors on the coefficients $\lambda_{\boldsymbol{k}}$ have increasingly smaller variances, thus placing less importance on higher-order (i.e., less smooth) basis functions $\phi_{\boldsymbol{k}}(\boldsymbol{x})$. It also captures the principle of effect hierarchy, which presumes that main effects typically have greater influence than interaction effects. To see why, let us fix the dimension as $d = 2$, and consider the two coefficients $\lambda_{(1,0)}$ and $\lambda_{(1,1)}$, where the first represents a main effect and the second represents its corresponding interaction. The prior specification (2.4) assigns greater variance on the first coefficient, thus allowing for greater effect magnitude with high probability a priori. This argument naturally extends for a higher dimension d by comparing, e.g., the two coefficients $\lambda_{(1,0,0,\cdots,0)}$ and $\lambda_{(1,1,0,\cdots,0)}$.

Despite this, the specification in (2.4) also reveals potential drawbacks of the standard GP as a surrogate model. First, it does not capture the desired principle of *effect heredity* (Haris et al., 2016), which presumes interaction effects are significant only when all of its component main effects are significant.[1] This is due to the *independent* prior specification of coefficients for both main effect and interaction terms. Effect heredity is widely used for effective analysis of limited data from physical experiments (see, e.g., Wu and Hamada, 2009; Mak and Wu, 2019a), and the integration of such structure within a GP is thus promising in our data-limited setting. Second, recall that our use of shrinkage aims to reflect the belief that the

[1]Here and throughout the paper, effect heredity refers to the *strong* notion of effect heredity (Haris et al., 2016), as opposed to *weak* effect heredity (Hamada and Wu, 1992), which requires at least one significant main effect.

response surface is dominated by a *small* number of significant effects. The prior specification in (2.4), however, does not capture this belief well. For coefficients $\lambda_{\boldsymbol{k}}$ with $|\boldsymbol{k}| = \sum_{m=1}^{d} k_m$ small, its prior samples from (2.4) will all be relatively large with high probability, resulting in many significant effects present in the surface $f(\cdot)$. Without a careful identification of such underlying dominant effects, the standard GP may thus yield mediocre predictions with high uncertainty, particularly with limited data.

We tackle such limitations via the proposed HierGP, which employs a novel prior specification on the basis representation (2.2). In particular, the HierGP makes use of spike-and-slab priors (Ishwaran and Rao, 2005), widely used in Bayesian variable selection to identify dominant coefficients in $\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^{\infty}$. These priors are then carefully constructed over each term in the expansion to embed the desired principles of effect hierarchy, heredity, and smoothness for structured shrinkage of $f(\cdot)$. In what follows, we first introduce the HierGP in the univariate setting of $d = 1$ and then extend it to the multivariate setting.

2.2. The univariate HierGP model. For ease of exposition, we introduce the HierGP first in the $(d = 1)$-dimensional setting. Suppose the design space is the unit interval $\mathcal{X} = [0, 1]$, and we obtain the training data $\{(x_i, y_i)\}_{i=1}^{n}$, where $x_1, \ldots, x_n \in \mathcal{X}$ are the training input parameters, and $y_1, \ldots, y_n$ are its corresponding outputs. We assume the outputs are obtained from the model:

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \theta^2), \quad i = 1, \ldots, n. \tag{2.5}$$

For the earlier problem of computer code emulation, the error term ϵ_i reduces to zero for deterministic simulators, since observations from f are obtained without noise. For the sake of generality, we will adopt the above noisy model specification for the remainder of the paper and reduce the error term to zero whenever appropriate.

Following (2.2), the HierGP assumes a basis expansion model on the response surface f:

$$f(x) = \sum_{k=1}^{\infty} \lambda_k \phi_k(x). \tag{2.6}$$

Here, $\{\phi_k(x)\}_{k=1}^{\infty}$ is a fixed L^2-orthonormal basis on $\mathcal{X}$ and $\{\lambda_k\}_{k=1}^{\infty}$ its corresponding coefficients. We then adopt the following hierarchical prior specification for the coefficients λ_k:

$$\lambda_k \overset{indep.}{\sim} \mathcal{N}(0, \sigma_k^2), \quad \sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}, \quad k = 1, 2, \ldots. \tag{2.7}$$

The prior $\sigma_k^2 \sim w_k \pi_k + (1 - w_k)\delta_{\sigma_\infty^2}$ is shorthand for the spike-and-slab prior (Ishwaran and Rao, 2005), where π_k is the slab distribution and $\delta_{\sigma_\infty^2}$ (a point mass at σ_∞^2) is the spike distribution. With probability w_k, this prior samples from the slab distribution π_k; otherwise, it samples a point mass at the spike σ_∞^2. The first case can be seen as the coefficient λ_k being *significant,* i.e., influential, for the response surface f, whereas (with a near-zero choice of σ_∞^2) the latter case can be seen as λ_k being *insignificant* for the response surface. This mixture of spike-and-slab priors provides an appealing probabilistic framework for identifying the underlying few dominant effects from data. A common choice for the slab distribution is the Inverse-Gamma prior $\pi_k = \mathrm{IG}(a_\sigma, b_\sigma)$; we later adopt this in section 3 for the proposed Gibbs sampler.

Although from a modeling perspective the choice of $\sigma_\infty^2 = 0$ for the spike distribution can be desirable, it is well known that the corresponding point mass $\delta_{\sigma_\infty^2}$ at 0 may cause computational instability and poor MCMC mixing (Scheipl et al., 2012); we have encountered similar issues in our implementation. A common work-around in the Bayesian variable selection literature (see, e.g., Ročková and George, 2018) is to set σ_∞^2 as a small but nonzero constant, such that σ_∞^2 is much smaller than the mean of the slab distribution $\mathbb{E}(\pi_k)$. Further discussion of the specification of hyperparameters a_σ, b_σ, and σ_∞^2 is provided in section SM5 of the supplementary material.

Of course, prior to data, the probability w_k (for λ_k to be significant) is typically unknown. We thus assign the following *cumulative* priors on $\{w_k\}_{k=1}^\infty$, adapted from the cumulative shrinkage priors in Legramanti et al. (2020) for sparse factor modeling:

$$w_k = \prod_{j=1}^{k}(1-\nu_j), \quad \nu_j \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha_j = \alpha), \quad w_0 = 1, \quad j = 1, 2, \ldots. \tag{2.8}$$

To complete the specification, we assign an independent Inverse-Gamma prior on the noise variance, i.e., $\theta^2 \sim \text{IG}(a_\theta, b_\theta)$. The prior (2.8) provides an appealing *cumulative* property that captures the desired *effect smoothness* principle (Ding et al., 2020), which asserts that lower-order (smoother) effects in f are more important than higher-order (less smooth) effects. To see why, note that with the prior (2.8), the expected probability on the kth coefficient is $\mathbb{E}(w_k) = (\alpha/(1+\alpha))^k$. For smaller indices k (corresponding to smoother effects), this prior favors larger values of w_k and places greater probability on the kth term to be significant; conversely, for larger k, it places less probability on the kth term to be significant. This cumulative property thus nicely reflects effect smoothness (Ding et al., 2020) within the desired spike-and-slab prior specification on the coefficients $\{\lambda_k\}_{k=1}^\infty$.

2.3. The multivariate HierGP model. With this univariate case in hand, we now present the multivariate HierGP model. Suppose the design space is the unit hypercube $\mathcal{X} = [0,1]^d$, and we collect training data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathcal{X}$ are training input parameters and $y_1, \ldots, y_n$ its corresponding outputs. As before, we assume the outputs are obtained from the model

$$y_i = f(\boldsymbol{x}_i) + \epsilon_i, \quad \epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \theta^2), \quad i = 1, \ldots, n, \tag{2.9}$$

where f takes the basis representation form

$$f(\boldsymbol{x}) = \sum_{|\boldsymbol{k}|=1}^{\infty} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \quad \boldsymbol{k} = (k_1, \ldots, k_d), \quad \boldsymbol{k} \in \mathbb{N}_0^d. \tag{2.10}$$

Here, $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^\infty$ is an L^2-orthonormal basis on $\mathcal{X} = [0,1]^d$, and $\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^\infty$ are its coefficients. We then adopt independent spike-and-slab priors on the coefficients:

$$\lambda_{\boldsymbol{k}} \overset{indep.}{\sim} \mathcal{N}(0, \sigma_{\boldsymbol{k}}^2), \quad \sigma_{\boldsymbol{k}}^2 \sim w_{\boldsymbol{k}} \pi_{\boldsymbol{k}} + (1 - w_{\boldsymbol{k}}) \delta_{\sigma_\infty^2}, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{2.11}$$

where $\pi_{\boldsymbol{k}}$ and $\delta_{\sigma_\infty^2}$ are again the slab-and-spike distributions, respectively.

We now extend the earlier prior to the multivariate setting to capture the desired principle of *effect heredity*:

$$w_{\boldsymbol{k}} = \prod_{m=1}^{d} w_{k_m,m}, \quad w_{k_m,m} = \prod_{j=1}^{k_m}(1-\nu_{j,m}), \quad w_{0,m} = 1, \quad \nu_{j,m} \overset{i.i.d.}{\sim} \text{Beta}(1, \alpha_{j,m} = \alpha) \tag{2.12}$$

for $m = 1, \ldots, d$ and $j = 1, 2, \ldots$. Recall that effect heredity (Haris et al., 2016) (or the marginality principle; see McCullagh and Nelder, 1989) presumes an interaction effect is significant only when all of its component variables are also significant. Here, the mth variable is deemed "significant" if its cumulative term $w_{k_m,m}$ is large and "insignificant" if $w_{k_m,m}$ is small, since a larger $w_{k_m,m}$ results in a larger probability that its corresponding coefficient is nonzero and vice versa. To see why the above prior embeds this principle, note that the probability weight $w_{\boldsymbol{k}}$ for each multi-index $\boldsymbol{k}$ is modeled as a *product* of separate cumulative terms $w_{k_m,m}$ over each of the d dimensions. By setting $w_{\boldsymbol{k}}$ as the product form (2.12), it follows that term $\boldsymbol{k}$ is significant (i.e., $w_{\boldsymbol{k}}$ large) only when all of its component variables (represented by its cumulative terms) are significant, which is precisely effect heredity. The prior (2.12) can thus be viewed as a careful specification of the spike-and-slab priors on coefficients $\{\lambda_{\boldsymbol{k}}\}_{|\boldsymbol{k}|=1}^{\infty}$ to reflect the desired effect heredity structure.

In addition to heredity, the prior (2.12) also captures the remaining two desired principles of *effect smoothness* and *hierarchy.* Effect smoothness follows directly from the earlier univariate case. With other indices in $\boldsymbol{k}$ held constant, as the index k_m decreases (i.e., for smoother terms), the employed prior favors larger values of $w_{\boldsymbol{k}}$ and thus places greater probability of the $\boldsymbol{k}$th term to be significant. To see effect hierarchy, let us again fix dimension as $d = 2$ and consider the main effect $\lambda_{\boldsymbol{k}_1}$, $\boldsymbol{k}_1 = (1, 0)$ and its interaction effect $\lambda_{\boldsymbol{k}_2}$, $\boldsymbol{k}_2 = (1, 1)$. By construction, the probability weight $w_{\boldsymbol{k}_1}$ would be strictly greater than $w_{\boldsymbol{k}_2}$, and thus main effects have a greater prior probability of being significant than do interactions. This argument again naturally extends for a higher dimension d by comparing, e.g., the probability weights for indices $\boldsymbol{k}_1 = (1, 0, 0, \ldots, 0)$ and $\boldsymbol{k}_2 = (1, 1, 0, \ldots, 0)$. In this sense, the proposed hierarchical prior (2.12) embeds the principles of effect heredity, hierarchy, and smoothness within the desired spike-and-slab framework.

To highlight the effect of such principles on prior specification, Figure 1 compares the marginal prior densities and correlations of the HierGP with the standard GP, the latter using a squared-exponential kernel with unit length-scale. Figure 1 (left) shows the prior densities on the effects $\lambda_{(1,1)}$ and $\lambda_{(2,2)}$ for both models. As expected, the HierGP priors are noticeably more aggressive in shrinkage towards near-zero values via its spike distribution, which is desirable in identifying the underlying sparse dominant effects. Further, in comparing the marginal priors for $\lambda_{(1,1)}$ and $\lambda_{(2,2)}$, we see that as the order increases in both dimensions, the HierGP assigns greater probability on its spike distribution. This results in a distinctly different prior for $\lambda_{(2,2)}$ compared to the standard GP, as it places increasingly greater weight on shrinking this effect, in accordance with the desired effect smoothness and hierarchy principles. Figure 1 (right) shows, in dimension $d = 2$, the marginal prior covariances between the first $5 \times 5 = 25$ coefficients from the same HierGP; note that the standard GP assigns *independent* priors on all coefficients. To contrast, the HierGP can be seen to impose highly structured correlations between coefficients, following the desired effect hierarchy, heredity, and smoothness principles. As we show later, when the response surface adheres to such principles and is controlled by a few dominant effects, the above structured prior specification from the HierGP can allow for improved predictions with limited data.

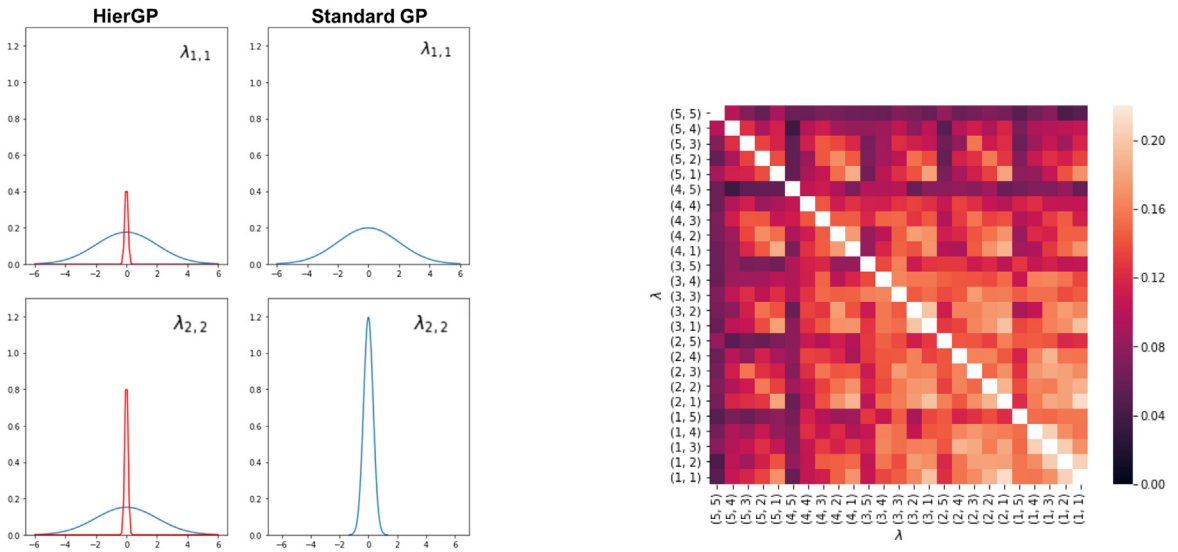

Figure 1. *(Left) Visualizing the prior densities for coefficients $\lambda_{(1,1)}$ and $\lambda_{(2,2)}$ from the HierGP and the standard GP using a squared-exponential kernel with unit length-scale. For the HierGP, the spike distribution is shown in red (scaled by probability $1-w_{\boldsymbol{k}}$) and the slab distribution in blue (scaled by probability $w_{\boldsymbol{k}}$). (Right) Visualizing the marginal prior covariances between different coefficients from the HierGP.*

The performance of the HierGP can of course be sensitive to the choice of the basis functions $\{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{|\boldsymbol{k}|=1}^{\infty}$, and thus a careful specification of this basis is important. In certain physical science applications (e.g., in turbulence dynamics; see Mak et al., 2018), one may be able to elicit guiding prior information on dominant basis functions, and this should of course be integrated within the HierGP. Such an elicited basis is typically of product form from our experience and thus should reflect increasing basis function complexity as each index in $\boldsymbol{k}$ increases. Another instance of this is in our later dynamic systems application (section 5.2), where it is known a priori that the few dominant effects are typically polynomial in the differential equations. For other applications, however, it may be more difficult to elicit an informed choice of basis. In such a setting, one can employ the eigenfunctions of a standard covariance kernel that captures the expected smoothness of the response surface, e.g., the squared-exponential kernel for highly smooth functions or the Matérn kernel for functions with certain desired smoothness.

The hierarchical form (2.12) can be easily modified in several ways to incorporate additional prior information from the modeler. First, in some cases, a modeler may have preference on a choice of basis (which may not be L_2-orthonormal) for which this hierarchical sparsity is expected to hold. This nonorthonormal basis can be directly integrated within this modeling framework and the later posterior sampler; our assumption of orthonormality is, however, needed in later theoretical results. Second, if a modeler expects heavier-tailed effects for coefficients, they can easily adopt heavier-tailed distributions (e.g., the horseshoe priors in Carvalho et al., 2009) within the spike-and-slab prior framework (2.11). One can also integrate heavier tails via a careful specification of the hyperparameter sequence $\alpha_{j,m}$. For example, with $\alpha_{j,m}$ set as an increasing sequence, e.g., $\mathcal{O}(m^2)$, we can place greater weights on latter terms in the expansion and thus impose heavier tail behavior.

2.4. The HierGP$_2$ model. While effect hierarchy and heredity are generally desirable for response surface modeling, there may be scenarios where one does not have a strong prior belief for such structured shrinkage, despite knowing that only a few coefficients $\lambda_{\boldsymbol{k}}$ are likely significant. In this case, one may want to adopt an alternate exchangeable shrinkage prior on the coefficients:

$$\lambda_{\boldsymbol{k}}|\sigma_{\boldsymbol{k}}^2 \sim \mathcal{N}(0, \sigma_{\boldsymbol{k}}^2\tau^2), \quad \sigma_{\boldsymbol{k}} \sim \Psi, \quad \boldsymbol{k} \in \mathbb{N}_0^d, \tag{2.13}$$

where Ψ is a distribution supported on $(0, +\infty)$. Such priors are known as global-local (GL) mixtures, where $\{\sigma_{\boldsymbol{k}}\}_{\boldsymbol{k}}$ are called the *local shrinkage* parameters, and τ the *global shrinkage* parameter. The GL prior provides a flexible framework for Bayesian shrinkage and selection: heavy-tailed distributions for Ψ allow for identification of strong signals, and its concentration around zero provides the desired shrinkage behavior. Examples of GL priors include the horseshoe prior (Carvalho et al., 2010), the Dirichlet–Laplace prior (Bhattacharya et al., 2015), and the generalized double Pareto prior (Armagan et al., 2013). We shall call the basis expansion model (2.10) with GL priors (2.13) the HierGP$_2$ model; in later experiments, the HierGP$_2$ is implemented with Ψ taken as the horseshoe prior from Carvalho et al. (2010). In scenarios where the underlying response surface (while dominated by a few dominant effects) does not adhere strongly to effect hierarchy and/or heredity, we show later that the HierGP$_2$ may also yield improved predictions over existing surrogate models.

3. Posterior sampling.

With the HierGP in hand, we present next an efficient MCMC algorithm for posterior sampling of the response surface $f(\cdot)$ given data. We first present a Gibbs sampler for the univariate HierGP and then extend this to a Gibbs sampler for the multivariate HierGP.

3.1. Gibbs sampling for the univariate HierGP. Suppose we collect training data $\{(x_i, y_i)\}_{i=1}^n$ from model (2.5). Let $\Theta = \{(\lambda_k)_{k=1}^\infty, (v_j)_{j=1}^\infty, \theta^2\}$ be the parameter set for posterior inference. One can write the likelihood function as[2]

$$L(\Theta|\{(x_i, y_i)\}_{i=1}^n) = [\{(x_i, y_i)\}_{i=1}^n|\Theta] = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\theta^2}} e^{-\frac{(y_i - f(x_i))^2}{2\theta^2}}, \tag{3.1}$$

where the response surface f is a function of parameters $(\lambda_k)_{k=1}^\infty$ (see (2.6)). Conditional on data, we then wish to draw samples $\{\Theta_{[b]}\}_{b=1}^B$ from the posterior distribution:

$$[\Theta|\{(x_i, y_i)\}_{i=1}^n] \propto L(\Theta|\{(x_i, y_i)\}_{i=1}^n)\,[\Theta], \tag{3.2}$$

where $[\Theta] = [(v_j)_{j=1}^\infty, (\lambda_k)_{k=1}^\infty, \theta^2]$ follows from the prior model in (2.6) and (2.7) for the univariate HierGP. With samples $\{\Theta_{[b]}\}_{b=1}^B$, posterior predictive samples on $f(\cdot)$ (call this $\{f_{[b]}(\cdot)\}_{b=1}^B$) can then be obtained by plugging $\{\Theta_{[b]}\}_{b=1}^B$ into (2.6).

To sample from (3.2), we make use of a data-augmented Gibbs sampler (Gelman et al., 1995), which leverages closed-form full conditional distributions for efficient posterior sampling. This sampler is similar in spirit to the Gibbs sampler in Legramanti et al. (2020) for factor models but is adapted for the GP setting at hand. As mentioned earlier, we adopt the

[2]Here, $[X]$ denotes the distribution of a random variable X.

Inverse-Gamma prior $\pi_k = \pi = \text{IG}(a_\sigma, b_\sigma)$ for the slab distribution in (2.7). We further employ the following truncation for $f(\cdot)$:

$$f(x) = \sum_{k=1}^{K} \lambda_k \phi_k(x) \tag{3.3}$$

for a sufficiently large choice of truncation index K. Further details are provided in supplementary material section SM6 on an adaptive choice of K, following Legramanti et al. (2020).

We now derive the required full conditional distributions. Let $\tilde{\boldsymbol{X}} = (\phi_1(x_i), \ldots, \phi_K(x_i))_{i=1}^n$ be the design matrix for the data. Further, for $k = 1, \ldots, K$, let z_k be a latent categorical random variable, defined conditionally as $\mathbb{P}(z_k = l|\Theta) = \nu_l w_{l-1}$ for $l = 1, 2, \ldots, K$. It can then be shown that

$$[\sigma_k^2|z_k] \sim \{1 - \mathbb{I}(z_k \leq k)\}\text{IG}(a_\sigma, b_\sigma) + \mathbb{I}(z_k \leq k)\delta_{\sigma_\infty^2}, \tag{3.4}$$

where $\mathbb{I}(\cdot)$ is the indicator function. With this data augmentation trick, one can then derive the full conditional distribution[3] of z_k as

$$[z_k = l|-] \propto \begin{cases} \nu_l w_{l-1} \phi(\lambda_k; 0, \sigma_\infty^2), & l = 1, \ldots, k, \\ \nu_l w_{l-1} t_{2a_\sigma}(\lambda_k; 0, (b_\sigma/a_\sigma)), & l = k+1, \ldots, K, \end{cases} \tag{3.5}$$

where $\phi(\lambda; 0, \sigma_\infty^2)$ and $t_{2a}(\lambda; 0, (b_\sigma/a_\sigma))$ are the densities of the normal and t-distributions (with $2a_\sigma$ degrees-of-freedom) evaluated at λ, respectively.

Consider now the full conditional distributions for the parameter set Θ. Let $\boldsymbol{D} = \text{diag}\{\sigma_k^2\}_{k=1}^K$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$. For the coefficient vector $\boldsymbol{\Lambda} = (\lambda_k)_{k=1}^K$, its full conditional distribution can be shown to be

$$[\boldsymbol{\Lambda}|-] \sim \mathcal{N}(\boldsymbol{V}\theta^{-1}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V}), \quad \boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}. \tag{3.6}$$

Similarly, for parameter v_j, its full conditional distribution becomes

$$[v_j|-] \sim \text{Beta}\left(1 + \sum_{k=1}^{K} \mathbb{I}(z_k = j), \alpha_j + \sum_{k=1}^{K} \mathbb{I}(z_k > j)\right), \quad j = 1, \ldots, K. \tag{3.7}$$

Last, for the noise parameter θ^2, its full conditional distribution can be shown to be

$$[\theta^2|-] \sim \text{IG}\left(a_\theta + \frac{n}{2}, \frac{b_\theta + \boldsymbol{S}^T\boldsymbol{S}}{2}\right), \quad \boldsymbol{S} = \boldsymbol{y} - \Lambda\tilde{\boldsymbol{X}}. \tag{3.8}$$

A complete derivation of these full conditional distributions is provided in section SM4 of the supplementary material.

Algorithm 3.1 presents the detailed steps of the Gibbs sampler, which combines the above full conditional steps (3.5)–(3.8) for sampling the desired posterior distribution $[\Theta|\{(x_i, y_i)\}_{i=1}^n]$. Here, all sampling steps are quite straightforward; for (3.5), we sample from this K-point discrete distribution via inverse transform sampling, with probabilities given by normalizing the weights in (3.5) to sum to one. Further details on the specification of hyperprior parameters can be found in section SM5 of the supplementary material.

[3]For brevity, $[\theta|-]$ denotes the full conditional distribution of parameter θ, conditional on both the data $y_1, \ldots, y_n$ and all parameters in Θ except for the considered parameter θ.

Algorithm 3.1. Gibbs sampling for the univariate HierGP.

Inputs: hyperparameters α, a_σ, b_σ, a_θ, b_θ, σ^2_∞, data $\{(x_i, y_i)\}_{i=1}^n$, number of iterations B.

Set initial parameters $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (v_j^{[0]})_{j=1}^K, \theta^2_{[0]}\}$.
for $b = 1, \ldots, B$ **do**
  for $k = 1, \ldots, K$ **do**
    Sample $z_k^{[b]}$ from the full conditional distribution $[z_k{=}l|-]$ in (3.5) with $\lambda_k{=}\lambda_k^{[b-1]}$. (Step 1)
  for $k = 1, \ldots, K$ **do**
    If $z_k^{[b]} \le k$, let $\sigma_k^2 = \sigma^2_\infty$; else sample $\sigma_k^2 \sim \text{IG}(a_\sigma + 0.5, b_\sigma + 0.5(\lambda_k^{[b-1]})^2)$. (Step 2)
  for $j = 1, \ldots, K$ **do**
    Sample $v_j^{[b]} \sim \text{Beta}(1 + \sum_{k=1}^K \mathbb{I}(z_k^{[b]}{=}j), \alpha_j + \sum_{k=1}^K \mathbb{I}(z_k^{[b]}{>}j))$; update w_j in (2.12).
  Sample $\theta^2_{[b]} \sim \text{IG}(a_\theta + n/2, b_\theta + \boldsymbol{S}^T\boldsymbol{S}/2)$, where $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}_{[b-1]}\tilde{\boldsymbol{X}}$. (Step 3)
  Sample $\boldsymbol{\Lambda}_{[b]} = (\lambda_k^{[b]})_{k \le K} \sim \mathcal{N}(\boldsymbol{V}\theta^{-2}_{[b]}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}_{[b]}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}\{\sigma_k^2\}_{k \le K}$. (Step 4)

3.2. Gibbs sampling for the multivariate HierGP. We now extend the above Gibbs sampler for the multivariate HierGP model,[4] which leverages the full hierarchical shrinkage prior (2.12). Suppose we obtain data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, and let $\Theta = \{(\lambda_{\boldsymbol{k}})_{|\boldsymbol{k}|=1}^\infty, (v_{j,m})_{j=1}^\infty{}_{m=1}^d, \theta^2\}$ be the parameter set. As before, we adopt the Inverse-Gamma prior $\pi_k = \pi = \text{IG}(a_\sigma, b_\sigma)$ for the slab distribution. We employ the following truncation for $f(\cdot)$:

$$f(\boldsymbol{x}) = \sum_{\boldsymbol{k} \le \boldsymbol{K}} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{3.9}$$

where $\boldsymbol{K} = (K_1, \ldots, K_d)$ is the vector of truncation indices for each dimension. Again, these indices $K_1, \ldots, K_d$ should be set sufficiently large; further details on this are provided in supplementary material section SM6 following Legramanti et al. (2020).

We now derive similar closed-form full conditional distributions of Θ for the multivariate HierGP. Let $\tilde{\boldsymbol{X}} = (\phi_{\boldsymbol{k}}(\boldsymbol{x}_i))_{\boldsymbol{k} \le \boldsymbol{K}, i=1,\ldots,n} \in \mathbb{R}^{n \times \|\boldsymbol{K}\|}$ be the design matrix for the data, where $\boldsymbol{k} \in \mathbb{N}_0^d$ is a multi-index that iterates over $\boldsymbol{K}$ and $\|\boldsymbol{K}\| = \prod_{m=1}^d K_m$. Further, let $\boldsymbol{z}_{\boldsymbol{k}}$ be the vector of latent random variables, defined conditionally as

$$\mathbb{P}\{\boldsymbol{z}_{\boldsymbol{k}} = (l_1, \ldots, l_d) | w_{l_1,1}, \ldots, w_{l_d,d}\} = \prod_{m=1}^d v_{l_m,m} w_{l_m - 1,m}, \tag{3.10}$$

where $w_{l_m,m}$ is as defined in (2.12). With this, we can leverage a similar data augmentation trick to derive the full conditional distribution of $\boldsymbol{z}_{\boldsymbol{k}} = [z_{\boldsymbol{k},1}, \ldots, z_{\boldsymbol{k},d}]$ as

[4] We will refer to the multivariate HierGP as simply "the HierGP" from here on.

$$[\boldsymbol{z}_{\boldsymbol{k}} = \boldsymbol{l}|-] \propto \begin{cases} \left(\prod_{m=1}^{d} v_{l_m,m} w_{l_m-1,m}\right) \phi(\lambda_{\boldsymbol{k}}; 0, \sigma_\infty^2); & \text{otherwise,} \\ \left(\prod_{m=1}^{d} v_{l_m,m} w_{l_m-1,m}\right) t_{2a_\sigma}(\lambda_{\boldsymbol{k}}; 0, (b/a)), & l_1 > k_1, \ldots, l_d > k_d, \end{cases} \tag{3.11}$$

by marginalizing out all $\sigma^2_{\boldsymbol{k}}$. Details on this marginalization are provided in section SM4 of the supplementary material.

Let us consider now the full conditional distributions for the parameter set Θ. Let $\boldsymbol{D} = \text{diag}\{\sigma^2_{\boldsymbol{k}}\}_{\boldsymbol{k} \le \boldsymbol{K}}$. For the coefficient vector $\boldsymbol{\Lambda} = (\lambda_{\boldsymbol{k}})_{\boldsymbol{k} \le \boldsymbol{K}}$, its full conditional can be shown to be the same form as (3.6). For parameter $v_{j,m}$, its full conditional distribution follows:

$$[v_{j,m}|-] \sim \text{Beta}\left(1 + \sum_{\boldsymbol{k} \le \boldsymbol{K}} \mathbb{I}(z_{\boldsymbol{k},m} = j), \alpha_j + \sum_{\boldsymbol{k} \le \boldsymbol{K}} \mathbb{I}(z_{\boldsymbol{k},m} > j)\right), \quad j = 1, \ldots, K_m, \quad m = 1, \ldots, d. \tag{3.12}$$

Finally, the full conditional of θ^2 follows the same form as in (3.8). Algorithm 3.2 details the steps in this Gibbs sampling algorithm. Further details on the specification of hyperprior parameters can be found in section SM5. As before, one can adopt a sufficiently large choice of truncation indices $\boldsymbol{K}$ or employ an adaptive setting of $\boldsymbol{K}$; details on the latter are provided in section SM6.

We provide a brief analysis of computational complexity for the Gibbs sampler in Algorithm 3.2. The computation cost for Step 4 can be shown to be $\mathcal{O}(dn\|\boldsymbol{K}\| + (\|\boldsymbol{K}\|)^3)$, since it requires the computation of $\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}$ from n observations and $\|\boldsymbol{K}\|$ bases, and a matrix inversion step for computing $\boldsymbol{V}$. The computational cost for Step 3 is $\mathcal{O}(|\boldsymbol{K}|)$, as we update

Algorithm 3.2. Gibbs sampling for the multivariate HierGP.

Inputs: hyperparameters α, a_σ, b_σ, a_θ, b_θ, σ^2_∞, data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, number of iterations B.

Set initial parameters $\Theta_{[0]} = \{(\lambda^{[0]}_{\boldsymbol{k}})_{\boldsymbol{k} \le \boldsymbol{K}}, (v^{[0]}_{j,m})_{j \le K_m, m \le d}, \theta^2_{[0]}\}$.

for $b = 1, \ldots, B$ **do**

  for $\boldsymbol{k}$ in $(1, \cdots, 1) : (K_1, \ldots, K_d)$ **do**

    Sample $\boldsymbol{z}^{[b]}_{\boldsymbol{k}}$ from $[\boldsymbol{z}_{\boldsymbol{k}} = \boldsymbol{l}|-]$ in (3.11) with $\lambda_{\boldsymbol{k}} = \lambda^{[b-1]}_{\boldsymbol{k}}$. (Step 1)

  for $\boldsymbol{k} = (k_1, \ldots, k_d)$ in $(1, \ldots, 1) : (K_1, \ldots, K_d)$ **do**

    if $\exists m$: $z^{[b]}_{\boldsymbol{k},m} \le k_m$, then let $\sigma^2_{\boldsymbol{k}} = \sigma^2_\infty$; else $\sigma^2_{\boldsymbol{k}} \sim \text{IG}(a_\sigma + 0.5, b_\sigma + 0.5(\lambda^{[b-1]}_{\boldsymbol{k}})^2)$. (Step 2)

  for m in $1, \ldots, d$ **do**

    for j in $1, \ldots, K_m$ **do**

      Sample $v^{[b]}_{j,m} \sim \text{Beta}(1 + \sum_{\boldsymbol{k} \le \boldsymbol{K}} \mathbb{I}(z^{[b]}_{\boldsymbol{k},m} = j), \alpha_j + \sum_{\boldsymbol{k} \le \boldsymbol{K}} \mathbb{I}(z^{[b]}_{\boldsymbol{k},m} > j))$.

  Update $w_{j,m}$ and $w_{\boldsymbol{k}}$ from (2.12).

  Update $\theta^2_{[b]}$ from $\text{IG}(a_\theta + n/2, b_\theta + \boldsymbol{S}^T\boldsymbol{S}/2)$, where $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}_{[b-1]}\tilde{\boldsymbol{X}}$. (Step 3)

  Sample $\boldsymbol{\Lambda}_{[b]} = (\lambda^{[b]}_{\boldsymbol{k}})_{\boldsymbol{k} \le \boldsymbol{K}} \sim \mathcal{N}(\boldsymbol{V}\theta^{-2}_{[b]}\tilde{\boldsymbol{X}}\boldsymbol{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{D}^{-1} + \theta^{-2}_{[b]}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}\{\sigma^2_{\boldsymbol{k}}\}_{\boldsymbol{k} \le \boldsymbol{K}}$. (Step 4)

$O(\|\boldsymbol{K}\|)$ variables and each update costs $O(1)$. The cost of Step 2 is $\mathcal{O}(\|\boldsymbol{K}\|^2)$, since it iterates through $\|\boldsymbol{K}\|$ variables, with each sampling step requiring $\mathcal{O}(\|\boldsymbol{K}\|)$ computation. The cost of Step 1 can similarly be shown to be $O(\|\boldsymbol{K}\|^2)$. Combining this, we have a computational complexity of $O\{dn\|\boldsymbol{K}\| + \|\boldsymbol{K}\|^3\}$ for each iteration of this Gibbs sampler. Note that, with a fixed truncation limit $K_m = K$ for each dimension $m = 1, \ldots, d$, $\|\boldsymbol{K}\|$ reduces to K^d. Thus, as dimension d increases, the computational cost for our Gibbs sampler can grow rapidly; this is unsurprising, as we would require an exponentially growing number of basis functions with fixed truncation limit K.

There are several ways to soften this so-called curse of dimensionality. An easy-to implement solution is to leverage multithreaded and/or distributed processing to parallelize the sampling of the latent vectors $\boldsymbol{z}_{\boldsymbol{k}}$ over $\boldsymbol{k}$, which is the primary computational bottleneck. Such distributed architecture is now commonplace in personal computers and computing clusters, and we have observed significant speed-ups (on the order of tens to hundreds) in exploiting this for our Gibbs sampler, which is highly parallelizable. Another strategy, motivated by the theory of sparse grids (Bungartz and Griebel, 2004), is to use a reduced basis expansion of the form (with equal truncation indices $K_m = K$)

$$f(\boldsymbol{x}) = \sum_{|\boldsymbol{k}|_1 \leq K+d-1} \lambda_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{3.13}$$

where $|\boldsymbol{k}|_1 = \sum_{m=1}^{d} k_m$. Such a "sparse" expansion effectively removes the highest-order terms in accordance with effect hierarchy and smoothness. One can show (Bungartz and Griebel, 2004) that such an approximation can indeed soften the curse of dimensionality for error approximation. This reduced-order basis can be directly used within the HierGP, and decreases the number of basis functions to $\mathcal{O}\{K(\log K)^{d-1}\}$, thus permitting greater scalability in moderate-to-high dimensions.

Finally, we provide some insights on the "data-adaptivity" of the HierGP, i.e., the adaptivity of its embedded effect principles from the data $\boldsymbol{y}$. For standard GP modeling, this adaptivity is typically achieved via inference on its kernel length-scale parameters, which dictate the embedded decay rate for effect smoothness. In the HierGP, there are no such length-scale parameters explicitly specified in its model. Instead, its data-adaptivity arises from the dependence of its prior variance $\mathrm{Var}(\lambda_k)$ on the slab distribution π_k (see (2.11)), which follows an Inverse-Gamma prior. As data are collected, the decay rate of these variance terms will be updated as the slab distribution (and other model parameters) is conditioned on such data, which then allows for adaptivity of the embedded effect principles within the HierGP.

3.3. Gibbs sampling for HierGP$_2$. In scenarios where one expects the presence of a few dominant effects but does not have strong prior belief of effect hierarchy and/or heredity, the alternative HierGP$_2$ model in section 2.4 (which makes use of GL shrinkage priors) may be an appealing alternative. We present next an analogous Gibbs sampler for this alternate model with Ψ taken as the horseshoe priors in Carvalho et al. (2010). Here, the posterior sampling of the parameter set $\Theta = \{(\lambda_k)_{k=1}^{\infty}, \theta^2\}$ reduces to an analogous setting as Bayesian linear regression with horseshoe priors, for which there are existing posterior sampling algorithms. Algorithm 3.3 provides a direct extension of one such sampler—the blocked Metropolis-within-

Algorithm 3.3. Blocked Metropolis-within-Gibbs sampler for the HierGP$_2$.

Inputs: hyperparameters a_θ, b_θ, τ, data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, number of iterations B.

1: Set initial parameters $\Theta_{[0]} = \{(\lambda_k^{[0]})_{k=1}^K, (\theta^{[0]})^2\}$.

2: **for** $b = 1, \ldots, B$ **do**

3: Sample $\boldsymbol{\Lambda}^{[b]} = (\lambda_1^{[b]}, \ldots, \lambda_K^{[b]})$ from $\mathcal{N}(\boldsymbol{V}(\theta^{[b-1]})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}, \boldsymbol{V})$, where $\boldsymbol{V} = (\boldsymbol{D}^{-1} + (\theta^{[b-1]})^{-2}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ and $\boldsymbol{D} = \text{diag}\{\sigma_k^{2\,[b-1]}\}_{k=1}^K$. (Step 1)

4: Sample $\{\sigma_k^{[b]}\}_{k=1}^K$ via Metropolis–Hastings from the density proportional to $\prod_{k=1}^K (1+\sigma_k)^{-1} \exp\{\frac{-(\lambda_k^{[b]})^2\tau\sigma_k}{2(\theta^{[b]})^2}\}$. (Step 2)

5: Update $(\theta^{[b]})^2$ from $\text{IG}(a_\theta + n/2, (b_\theta + \boldsymbol{S}^T\boldsymbol{S})/2)$, where $\boldsymbol{S} = \boldsymbol{y} - \boldsymbol{\Lambda}^{[b]}\tilde{\boldsymbol{X}}$. (Step 3)

Gibbs sampler in Johndrow et al. (2020)—for posterior sampling from the HierGP$_2$. Further details on the specification of hyperprior parameters can be found in section SM5 of the supplementary material.

4. Consistency results for the HierGP. We next present consistency results for the HierGP, which extend existing theory on consistency for high-dimensional Bayesian linear regression, specifically from Song and Liang (2023) and Choi and Schervish (2007). As mentioned before, for the HierGP, we will adopt point masses $\delta_{\sigma^2_{\boldsymbol{k},0}}$ for the spike distributions, where $\{\sigma^2_{\boldsymbol{k}}\}$ is prespecified; this is commonly used in the literature for theoretical analysis. Main theorems are presented next, with assumptions and proofs deferred to Appendix A for brevity.

To prove consistency, we will require several assumptions on the underlying basis representation form of f, its smoothness, and the experimental design of the sample points $\{\boldsymbol{x}_i\}_{i=1}^n$. The technical details of such assumptions are deferred to section SM1.1. We do outline an important assumption (A1) in section SM1.1 on the true function $f_0(\cdot)$ that we wish to predict. Namely, we assume

$$f_0(\boldsymbol{x}) = \sum_{\boldsymbol{k} \in S} \lambda^0_{\boldsymbol{k}} \phi_{\boldsymbol{k}}(\boldsymbol{x}), \tag{4.1}$$

where $S \subset \mathbb{N}_0^d$ is a finite index set. In other words, we presume that the true function f_0 can be represented as a finite sum of the employed basis functions $\phi_{\boldsymbol{k}}(\boldsymbol{x})$; this reflects our belief that f_0 is indeed dictated by a few dominant effects.

Under such assumptions, we can prove the following consistency result for the HierGP under fixed design points.

Theorem 4.1 (consistency of HierGP, fixed design). *Suppose f follows the HierGP model specified in* (2.9)–(2.12) *with corresponding prior measure Π_2^* (with trivial modification made to satisfy assumption* (A5)*; see section* SM1.1*). Let Q_0 be the conditional distribution of the data $\{y_i\}_{i=1}^n$ given fixed design points $\{\boldsymbol{x}_i\}_{i=1}^n$. Suppose assumptions* (A1), (A2), (A4), *and* (A5) *from section* SM1.1 *hold. Then the following hold:*

(a) *For any $\epsilon > 0$, we have*

$$\Pi_2^*\left((f,\theta) \in W_{\epsilon,n}^c | \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n\right) \to 0, \quad [Q_0] \text{ almost surely}, \tag{4.2}$$

where $W_{\epsilon,n}$ is the empirical measure neighborhood around the truth (f_0, θ_0) (see section SM1.2*).*

(b) *For any* $\epsilon > 0$, *we also have*

$$\Pi_2^*(L_\epsilon^c | \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely}, \tag{4.3}$$

where L_ϵ *is the* L_1*-neighborhood around* (f_0, θ_0) *(see section* SM1.2*).*

We can further show consistency of the HierGP under random design points.

Theorem 4.2 (consistency of HierGP, random design). *Suppose* f *follows the HierGP model specified in* (2.9)–(2.12) *with prior measure* Π_1^*. *Let* Q_0 *denote the joint distribution of* $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^\infty$. *Suppose assumptions* (A1), (A2), *and* (A3) *from section* SM1.1 *hold. Then, for random design points drawn from some measure* P_0, *the following hold:*

(a) *For any* $\epsilon > 0$, *we have*

$$\Pi_1^*(U_\epsilon^c | \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely}, \tag{4.4}$$

where U_ϵ *is a* P_0*-neighborhood around* (f_0, θ_0) *(see section* SM1.2*).*

(b) *For any* $\epsilon > 0$, *we also have*

$$\Pi_1^*(H_\epsilon^c | \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n) \to 0, \quad [Q_0] \text{ almost surely}, \tag{4.5}$$

where H_ϵ *is the Hellinger neighborhood around* (f_0, θ_0) *(see section* SM1.2*).*

Theorems 4.1 and 4.2 show that, under regularity conditions, the posterior distribution of the response surface and its corresponding variance parameter indeed converge to the truth under various modes of convergence as sample size n goes to infinity. This establishes posterior consistency of the proposed HierGP and guarantees that the employed structured shrinkage prior indeed provides enough support for predicting the desired class of functions in (4.1). The proofs of both theorems extend results in Choi and Schervish (2007), and its detailed statements are provided in sections SM1 and SM2.

We provide some insight on why only consistency is shown for the HierGP. While there exists a rich literature on contraction rates for standard shrinkage priors, the employed hierarchical shrinkage prior within the HierGP is quite new. Standard analysis tools for high-dimensional Bayesian linear regression (see, e.g., Castillo et al., 2015, Song and Liang, 2023, Jeong and Ghosal, 2021) are thus not directly suitable in this setting, due to the highly structured nature of our shrinkage prior. We thus focus on establishing the consistency of our model in this novel prior setting and defer the more challenging question of contraction rates to future work. Section SM3 provides further consistency and Bernstein–von Mises approximation results on the HierGP_2 model.

5. Numerical experiments.

We now explore the proposed HierGP and HierGP_2 in a suite of numerical experiments. We first investigate their performance for computer code emulation and then demonstrate its effectiveness for the recovery of dynamical systems.

5.1. Computer code emulation.

For our experiments on computer code emulation, we will consider a suite of test functions and compare the proposed models (HierGP and HierGP_2) with several popular and/or related GP-based emulators. This includes the standard GP emulator with Matérn-3/2 kernel (Stein, 1999); the additive GP model in Lu et al. (2022), which builds off of recent work (Duvenaud et al., 2011) on leveraging additive low-dimensional structure; the "least-squares" model, which makes use of a least-squares fit of the data $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$

using a prespecified basis matrix $\tilde{\boldsymbol{X}}$; and the sparse least-squares fit, which makes use of l_1-regularized estimates under the same setup, with penalty parameters tuned via cross-validation. The latter two are akin to surrogate models used in polynomial chaos; see Luthen et al. (2021) for a comprehensive review. Both models provide useful benchmarks for the HierGP in highlighting potential advantages of embedding effect smoothness, hierarchy, and heredity within shrinkage priors on the GP. For a fair comparison, both models make use of the same basis matrix as the HierGP. Finally, we compare with the fixed rank kriging method (see, e.g., Cressie and Johannesson, 2008), a popular approach in spatial statistics that leverages a prespecified truncation of the Karhunen–Loève expansion. This is implemented via the R package `FRK` (Zammit-Mangion and Sainsbury-Dale, 2023).

Consider first the setting where test functions are simulated from the HierGP prior in section 3.2 in $d = 2$ and $d = 3$ dimensions. This function is simulated with parameters $\alpha = 6, a_\sigma = 1, b_\sigma = 1, \sigma_\infty = 0$, a truncation limit of $\boldsymbol{K} = (8, 8)$ and $(4, 4, 4)$ for $d = 2$ and $d = 3$, respectively, and sinusoidal basis $\phi_{\boldsymbol{k}}(\boldsymbol{x}) = \Pi_{m=1}^{d} \sin(2\pi k_m x_m)$. The simulated response surfaces thus are dominated by a few significant effects and capture the presumed effect smoothness, heredity, and hierarchy principles. For model training, we use $n = 70$ uniformly sampled design points. This sample size is higher than the usual rule of thumb $n = 10d$ recommended in Loeppky et al. (2009), which has been noted to be insufficient for more complex functions (Harari et al., 2018), as is the case here. For the HierGP, least-squares, and sparse least-squares fits, we assume the perfectly specified setting where the basis matrix and truncation levels are set to be the same as the simulation model; we will explore a misspecified setting next. This simulation is then replicated 50 times to measure error variability.

Figures 2 and 3 show boxplots of the prediction errors for 400 uniformly sampled testing points for each approach in $d = 2$ and 3 dimensions. We see that the HierGP yields improved predictions over competing models. This is not surprising: when dominant effects in f are structured according to the effect principles, the HierGP should be able to leverage such

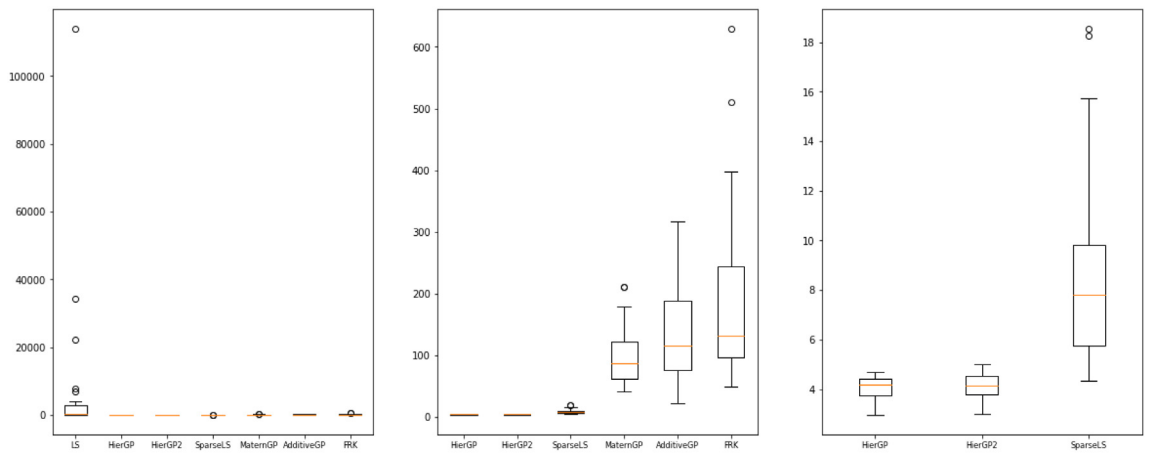

Figure 2. *Boxplots for prediction root-mean-squared errors of the compared methods when f is simulated from the HierGP in $d = 2$ dimensions. From left to right are zoomed-in views of the boxplots. Here, LS denotes the least-squares fit, SparseLS denotes the sparse least-squares fit, and FRK denotes the fixed rank kriging approach.*

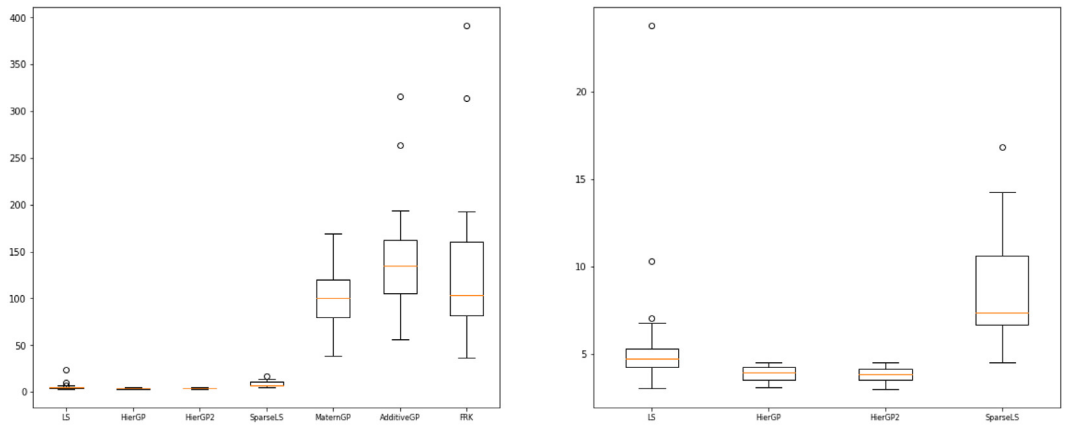

Figure 3. *Boxplots for prediction root-mean-squared errors of the compared methods when f is simulated from the HierGP in $d = 3$ dimensions. From left to right are zoomed-in views of the boxplots. Here, LS denotes the least-squares fit, SparseLS denotes the sparse least-squares fit, and FRK denotes the fixed rank kriging approach.*

Table 1

Empirical coverage rates of the 95% posterior predictive credible intervals and their corresponding average interval widths when the true function f is simulated from the HierGP in d dimensions.

	HierGP	HierGP_2	MatérnGP
Empirical coverage rate ($d = 2$)	100.0%	100.0%	91.0%
Empirical coverage rate ($d = 3$)	97.7%	99.1%	92.2%
Average credible interval width ($d = 2$)	1.721	1.313	4.062
Average credible interval width ($d = 3$)	1.277	1.798	7.250

structure for improved predictions. The HierGP_2 (which adopts a less structured horseshoe prior on basis coefficients) also does quite well, but it has slightly higher errors since it does not integrate full information on effect principles. The standard Matérn GP and the additive GP both yield significantly worse predictive performance; this is intuitive since such models do not account for structured sparsity in f. The two least-squares fits (even with a perfectly specified basis) also yield poor performance, which is expected since such fits also do not identify the sparse number of dominant effects, despite having a perfectly specified basis. Finally, fixed rank kriging also yields worse predictions compared to the HierGP. This is again not too surprising, as the latter uses a predetermined truncation of the Karhunen–Loève expansion, whereas the HierGP leverages a *data-adaptive* and *structured* shrinkage of this expansion to identify important effects in f.

Table 1 compares the uncertainty quantification performance of these models by reporting the empirical coverage rates of 95% posterior predictive credible intervals and corresponding average predictive interval widths. Here, we compared only the two HierGP models with the standard GP, as the other methods are not fully probabilistic. We see that, for both $d = 2$ and $d = 3$, the empirical coverage rates for the HierGP models are noticeably higher than that for the standard GP, which can dip below the nominal 95% rate. Furthermore, the credible interval widths for the HierGP models are significantly smaller than that for the

standard GP. This suggests that, when f has the presumed structure for dominant effects, the proposed models can indeed leverage such structure to provide more *precise* probabilistic predictions with improved coverage over standard GPs, as desired.

Finally, we investigate the computation times of the compared methods in our $d = 2$ experiment (the $d = 3$ experiment yields comparable timing). All computation is performed on a single-core 2.6 GHz Intel Core i7 processor and is measured in seconds. Here, our HierGP and HierGP$_2$ models (which provide the best predictive performance) require 20.1 and 16.3 s, respectively, for model training and prediction. Of the existing GPs, the MatérnGP, fixed rank kriging, and additive GP models require 34.4, 28.6, and 27.3 s, respectively, for the same tasks. Finally, the least-squares and sparse least-squares fit are nearly instantaneous, but as shown earlier, such methods yield worse predictive performance to our methods, as it does not leverage the desired effect principles. Thus, for these experiments, the HierGP models offer improved emulation performance at comparable computing times.

Consider next the emulation of the two synthetic test functions in the literature, the Branin function (Sobester et al., 2008)

$$f(\mathbf{x}) = a(x_2 - bx_1 + cx_1 - r)^2 + s(1-t)\cos(x_1) + s,$$

where $a = 1$, $b = 5.1/(4\pi^2)$, $c = 5/\pi$, $r = 6$, $s = 10$, and $t = 1/\pi$, and the Cheng & Sandu function (Cheng and Sandu, 2010)

$$f(\mathbf{x}) = \cos(x_1 + x_2)\exp(x_1 x_2).$$

These two functions vary in their degree of adherence to the effect heredity and hierarchy principles. Here, the HierGP, least-squares, and sparse least-squares fits make use of the above sinusoidal basis with truncation limit $\boldsymbol{K} = (8, 8)$. This provides a good test for how robust the proposed models are when there are minor violations of the effect principles with respect to the chosen basis functions. As before, $n = 70$ uniformly sampled design points are used for training.

Figure 4 shows boxplots of prediction errors for 400 uniformly sampled testing points for each test function. For the Branin function, the HierGP provides the best predictive performance of all the considered models, with the HierGP$_2$ a close competitor. Upon further inspection, this is not surprising since its functional form suggests some form of effect heredity and hierarchy is present. For the Cheng & Sandu function, we see that the HierGP$_2$ provides the best predictive performance, with the HierGP a close competitor. This can be explained by the more complex interaction structure present in its functional form, which may slightly deviate from the presumed hierarchical structure in the HierGP. Regardless, the above experiments suggest that, when effect smoothness, hierarchy, and heredity are present in f (even with minor violations), the proposed models can indeed identify and integrate such structure for improved predictive performance.

5.2. Recovery of dynamical systems. We now further investigate the HierGP for the problem of dynamical system recovery and prediction, which is widely used in climatology, ecology, and finance (see, e.g., Ghadami and Epureanu, 2022, Luo et al., 2011, Mudelsee, 2019). We first provide a brief review of this problem, following Brunton et al. (2016). We consider here dynamical systems (Guckenheimer and Holmes, 2013) that take the form

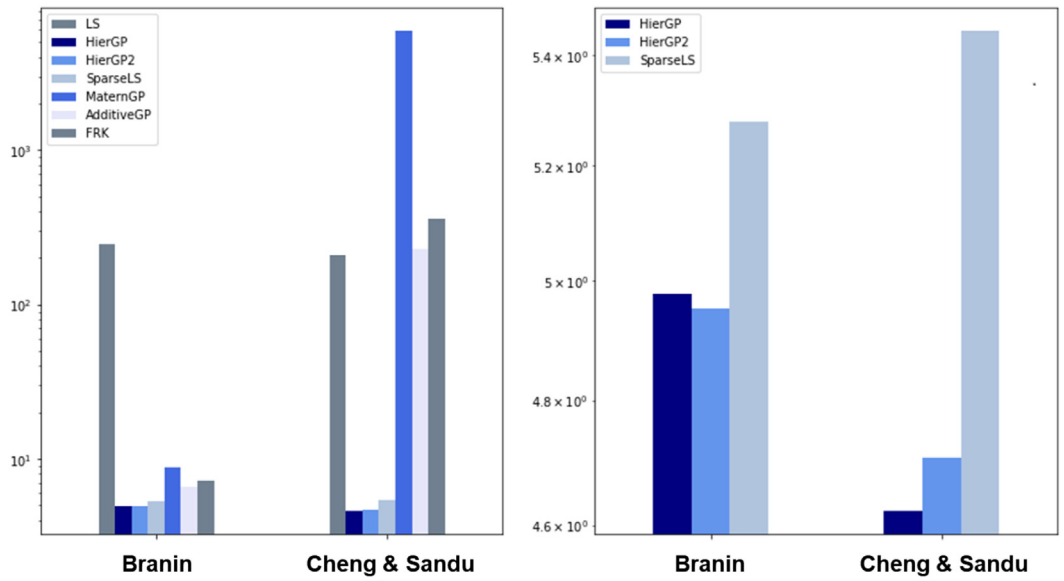

Figure 4. *Barplots of prediction root-mean-squared errors for the compared methods when f is taken as the Branin and Cheng & Sandu functions, respectively. The left figure shows the barplots over all methods, and the right figure shows the zoomed-in barplots for the best-performing methods. Here, LS denotes the least-squares fit, SparseLS denotes the sparse least-squares fit, and FRK denotes the fixed rank kriging approach.*

$$\frac{d}{dt}\boldsymbol{x}(t) := \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t)). \tag{5.1}$$

Here, $\boldsymbol{x}(t) \in \mathbb{R}^q$ denotes the system states at time t, and $\boldsymbol{f}(\boldsymbol{x}(t))$ captures dynamical constraints that govern the equations of motion for such system states. The formulation (5.1) covers a broad range of dynamical systems used in ecology, biology, and other scientific disciplines (Delahunt and Kutz, 2022).

Consider now the setting where data $\{(\boldsymbol{x}(t_i), \dot{\boldsymbol{x}}(t_i)\}_{i=1}^n$ are observed on the system states, where $t_1, \ldots, t_n$ are the sampled time points. We can rearrange this into the following state matrices:

$$\boldsymbol{X} = (\boldsymbol{x}^T(t_1), \ldots, \boldsymbol{x}^T(t_n))^T = \begin{pmatrix} x_1(t_1) & x_2(t_1) & \cdots & x_d(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ x_1(t_n) & x_2(t_n) & \cdots & x_d(t_n) \end{pmatrix} \in \mathbb{R}^{n \times d},$$

$$\dot{\boldsymbol{X}} = (\dot{\boldsymbol{x}}^T(t_1), \ldots, \dot{\boldsymbol{x}}^T(t_n))^T = \begin{pmatrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_d(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_d(t_2) \\ \cdots & \cdots & \ddots & \vdots \\ \dot{x}_1(t_n) & \dot{x}_2(t_n) & \cdots & \dot{x}_d(t_n) \end{pmatrix} \in \mathbb{R}^{n \times d}.$$

With this, we then construct a "library" of candidate functions for recovering the function $\boldsymbol{f}$ in (5.1), thus recovering the underlying system dynamics. Suppose these candidate functions

$\mathcal{F} = \{\phi_{\boldsymbol{k}}(\boldsymbol{x})\}_{\boldsymbol{k}}$ are parametrized by the multi-index $\boldsymbol{k} = (k_1, \ldots, k_d)$, $\boldsymbol{k} \leq \boldsymbol{K}$. Given the sampled time points, this library can be represented by the following model matrix:

$$\Phi(\boldsymbol{X}) = \begin{pmatrix} | & \cdots & | & \cdots & | \\ \Phi(\boldsymbol{X})^{[\boldsymbol{1}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{k}]} & \cdots & \Phi(\boldsymbol{X})^{[\boldsymbol{K}]} \\ | & \cdots & | & \cdots & | \end{pmatrix} \in \mathbb{R}^{n \times \|\boldsymbol{K}\|},$$

where $\|\boldsymbol{K}\| = \prod_{m=1}^{d} K_m$ is the total number of basis functions in $\mathcal{F}$, and $\Phi(\boldsymbol{X})^{[\boldsymbol{k}]}$ is the model matrix consisting of the basis functions in $\mathcal{F}$ with multi-index $\boldsymbol{k}$. The dynamical system (5.1) can then be represented by the following linear system of equations:

$$\dot{\boldsymbol{X}} = \Phi(\boldsymbol{X})\boldsymbol{\Xi}, \tag{5.2}$$

where $\boldsymbol{\Xi} = (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_d) \in \mathbb{R}^{\|\boldsymbol{K}\| \times d}$ is the matrix of coefficients for $\boldsymbol{f}$, and $\boldsymbol{\xi}_m \in \mathbb{R}^{\|\boldsymbol{K}\|}$ is the coefficient vector for the mth component of $\boldsymbol{f}$.

Given data matrices $\dot{\boldsymbol{X}}$ and $\boldsymbol{X}$, the goal of recovering $\boldsymbol{f}$ can be viewed as a regression problem on estimating the coefficient matrix $\boldsymbol{\Xi}$. In a seminal paper, Brunton et al. (2016) argued that, since in many physical systems there are only a few dominant terms that govern the underlying dynamics, the coefficients in $\boldsymbol{\Xi}$ should be estimated in a sparse manner. To achieve this, they proposed a method called Sparse Identification of Nonlinear Dynamics (SINDy), which makes use of compressed sensing algorithms for sparse estimation of $\boldsymbol{\Xi}$, thus allowing for a sparse identification of the system $\boldsymbol{f}$. Since then, there have been further developments on SINDy via sparse regression and deep learning; see Champion et al. (2020) and Both et al. (2021).

A potential limitation with the above SINDy-based methods is that, as mentioned in section 1, dominant effects in physical systems are often structured via the principles of effect hierarchy, heredity, and smoothness (Ding et al., 2020): main effects typically have greater influence than interactions, and interactions are significant only when component main effects are significant. One way to capture such structure is to assign the proposed hierarchical cumulative priors (2.11) and (2.12) independently over each row of the coefficient matrix $\boldsymbol{\Xi}$. With these priors, the resulting linear system (5.2) can be viewed as fitting m independent HierGPs, with basis functions taken from the function library $\mathcal{F}$. The *recovery* of governing equations can thus be performed via posterior sampling of the coefficient matrix $\boldsymbol{\Xi}$ given data $\{(\boldsymbol{x}(t_i), \dot{\boldsymbol{x}}(t_i)\}_{i=1}^{n}$, using the Gibbs sampler in section 3.2. With posterior samples $\{\boldsymbol{\Xi}_1, \ldots, \boldsymbol{\Xi}_B\}$ generated, one can then *predict* and *quantify uncertainty* on the dynamical system via forward solves of (5.1) using each coefficient matrix sample $\boldsymbol{\Xi}_b$, $b = 1, \ldots, B$.

For problems where such structure is present in the governing equations, one would expect that the integration of the effect principles within the HierGP can yield improved dynamical system recovery with greater certainty, particularly using limited data. We explore this below in numerical comparisons with existing methods on two dynamical systems.

5.2.1. 2D cubic equations. Consider the following 2D planar dynamical system (see Brunton et al., 2016 for further details):

$$\begin{aligned} \frac{dx(t)}{dt} &= -ax(t)^3 + by(t)^3, \\ \frac{dy(t)}{dt} &= -bx(t)^3 + ax(t)^3, \end{aligned} \tag{5.3}$$

where a and b are constant parameters. While this system is simple, it has two appealing features that allow interesting comparison of recovery methods. First, its derivative functions capture effect sparsity and hierarchy, as they depend on only two basis functions of relatively low order. Second, one can show for any initial point $(x(0), y(0))$, the system will always converge to a stationary point $(0, 0)$ with quasi-periodic behavior (Brunton et al., 2016), thus providing stability and predictability to numerical solutions. In the following experiment, we set the true parameters as $a = 0.1$ and $b = 2$, with initial condition $(x_0, y_0) = (2, 0)$. We then generate the training data by numerically solving the dynamical system (5.3) and then sampling $n = 500$ observations (with a time step of 0.04) from one trajectory corrupted with i.i.d. Gaussian noise (with variance 0.01). Finally, we set $\boldsymbol{K} = (5, 5)$ for the HierGP.

Figure 5 shows the trajectory of the true dynamical system, as well as the recovered trajectory from a forward simulation of the fitted HierGP model. For the latter, we first performed posterior sampling on model parameters Θ and then used its posterior mean as parameters for a forward solve of the system (5.3). Visually, we see that the HierGP trajectory captures well the desired periodic and asymptotic behavior of the cubic system. Figure 6 shows the corresponding prediction errors of the recovered systems for the HierGP and SINDy in each of the two coordinates. We see that the HierGP indeed yields noticeably improved predictions over SINDy; this suggests that the integration of effect structure (when present) for prior specification indeed allows for improved system recovery. The errors for both methods are relatively small, which is unsurprising since the true dynamical system is quite simple. These errors do grow slightly with time; this is intuitive since recovery errors should propagate in time given estimation errors for dynamical system coefficients.

Figure 7 further explores the uncertainty quantification of the proposed method by showing the forward runs of 50 posterior sample draws for Ξ for the 2D system (5.3) in x- and y-coordinates. The existing SINDy method (Brunton et al., 2016) does not provide such a quantification of uncertainty. We see that the recovered trajectories from the HierGP not only recover the true system well but also does so with relatively high certainty. We do note

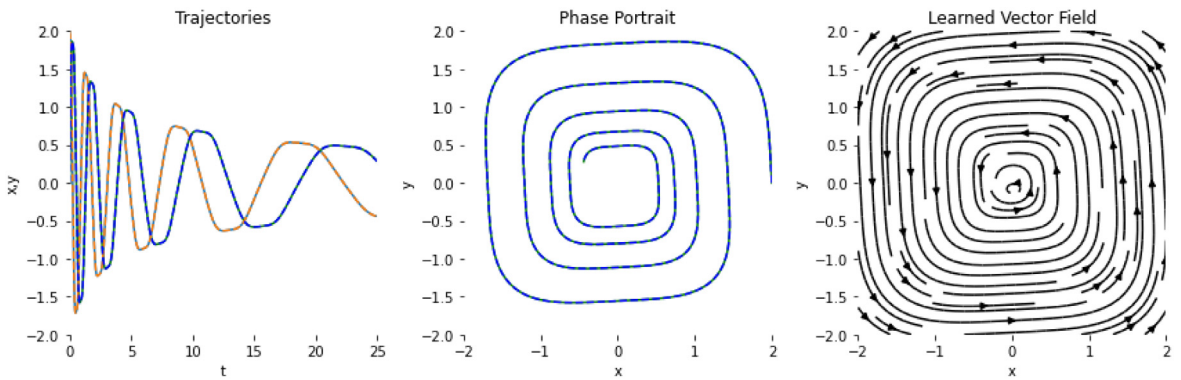

Figure 5. *Visualizing the true and recovered trajectories for the 2D cubic system. [Left] The x-trajectory (orange) and y-trajectory (blue) for the true (solid) and recovered (dashed) systems using the HierGP. [Middle] The 2D trajectories of the true (solid) and recovered (dashed) systems from the HierGP. [Right] The learned vector field from the fitted HierGP model.*

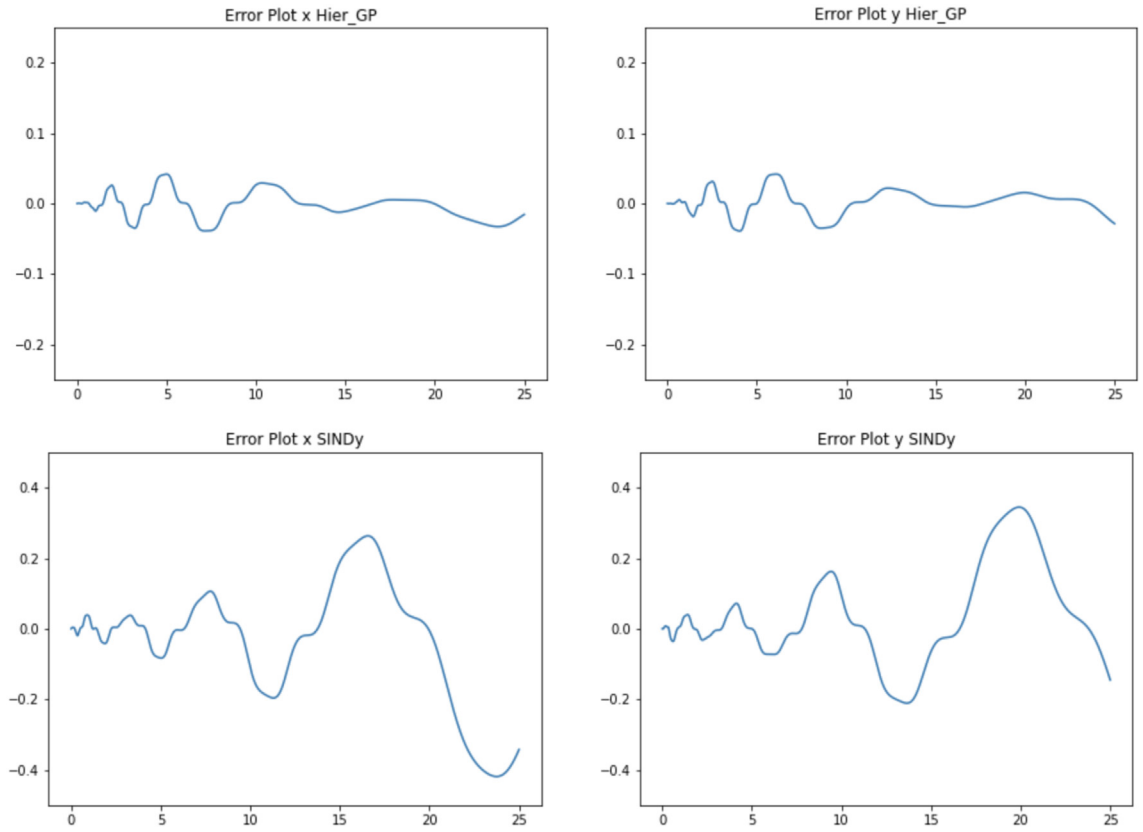

Figure 6. *Prediction errors (in x- and y-coordinates) of the HierGP (top) and SINDy (bottom) for the* 2*D cubic system with* 500 *time steps.*

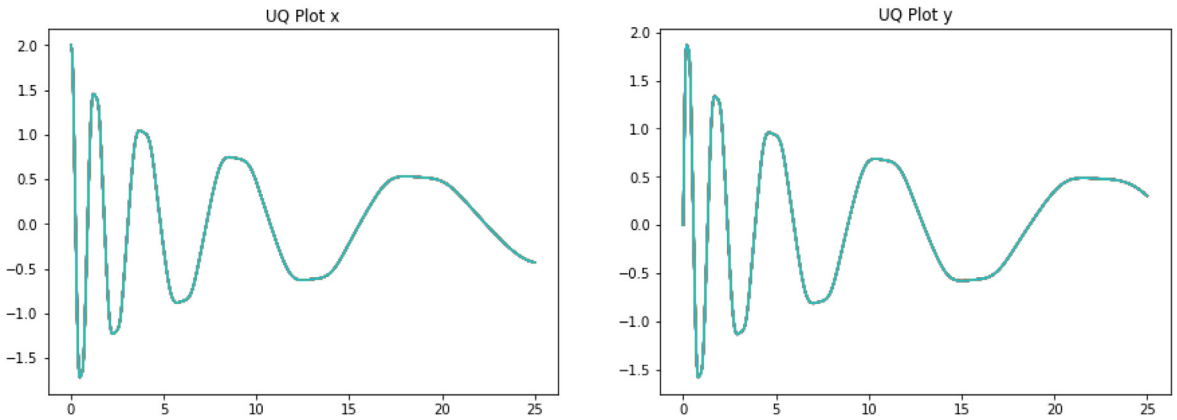

Figure 7. *Visualizing the forward runs of* 50 *posterior sample draws from the HierGP for the* 2*D cubic system in x- and y-coordinates (from left to right).*

that the posterior uncertainties from our model grow gradually in time; this is again not surprising since uncertainties in the recovered system should accumulate over time.

5.2.2. Chaotic Lorenz system. Consider next the following 3D chaotic Lorenz system (Lorenz, 1963), given by

$$\begin{aligned} \frac{dx(t)}{dt} &= \sigma(y(t) - x(t)), \\ \frac{dy(t)}{dt} &= x(t)(\rho - z(t)) - y(t), \\ \frac{dz(t)}{dt} &= x(t)y(t) - \beta z(t), \end{aligned} \tag{5.4}$$

where σ, β, and ρ are constant parameters. Although these equations have rich and chaotic dynamics that evolve on a strange attractor (Brunton et al., 2016), the modeled equations for each derivative are sparse and of relatively low order, thus reflecting the desired effect principles. In particular, the derivative functions are typically influenced by only a few low-order terms that have sparse and hierarchical structure (Brunton et al., 2016). In the following, we set the true parameters as $\sigma = 10$, $\beta = 8/3$, and $\rho = 28$ with initial conditions $(x(0), y(0), z(0)) = (-8, 7, 27)$. As before, the training data are generated by solving the dynamical system (5.4) and then sampling (with time step 0.05) $n = 200$ observations corrupted with Gaussian noise (with variance 0.01) from the resulting solution.

Figure 8 shows the trajectory of the true Lorenz system, as well as the trajectory of the recovered system using the fitted HierGP model with $\boldsymbol{K} = (5, 5, 5)$. Visually, we see that the recovered system captures the desired strange attractor behavior and short-time dynamics: the trajectory moves locally and predictably initially, but more globally and chaotically as time progresses, constrained within a region with complex geometric structure (Lorenz, 1963). Figure 9 shows the corresponding prediction errors of the recovered systems using the HierGP and SINDy in each of the three coordinates. We again see that the HierGP yields noticeably

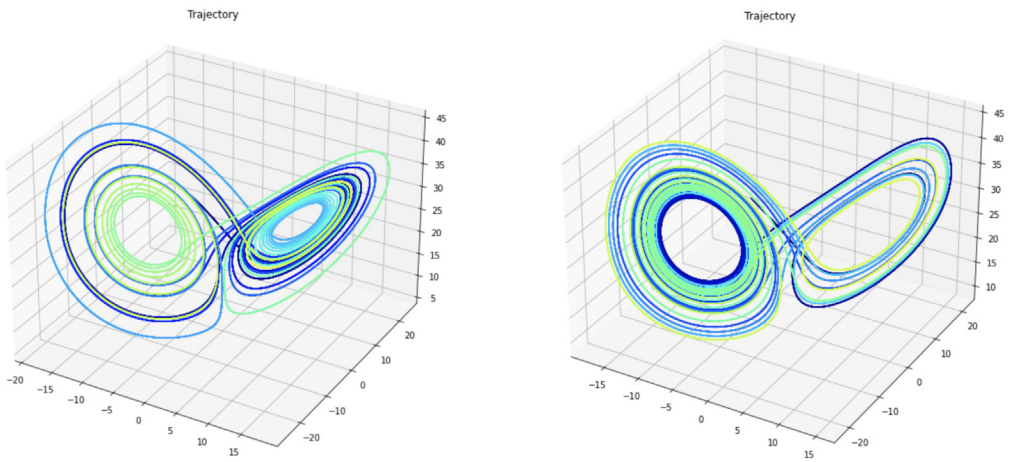

Figure 8. *Visualizing the recovered trajectory (left) of the 3D Lorenz system using the HierGP and the true trajectory of the system (right).*

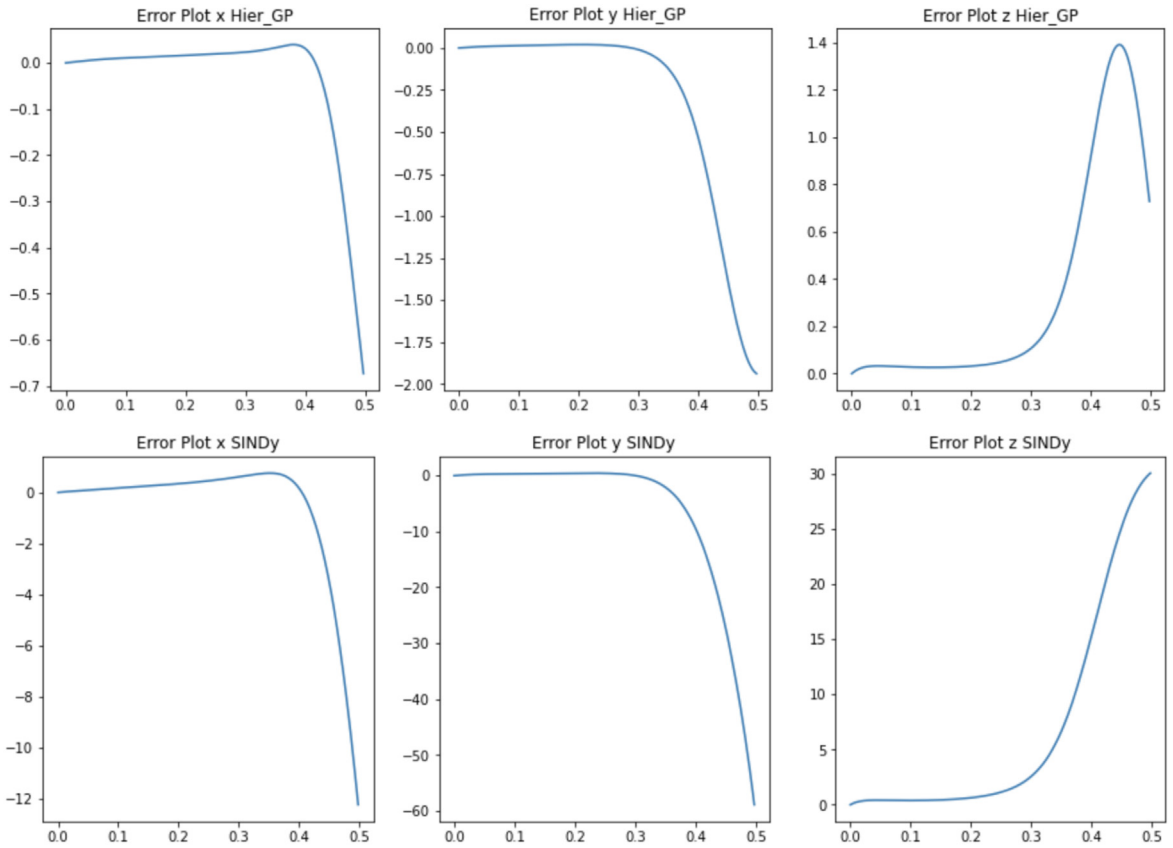

Figure 9. *Prediction errors (in x-, y-, and z-coordinates) of the HierGP (top) and SINDy (bottom) for the 3D Lorenz system.*

improved performance over SINDy, which suggests that the integration of structured shrinkage (when present) can indeed improve system recovery. We note that the errors grow rapidly as time progresses, which suggests that recovery becomes increasingly difficult over time; this is not too surprising given the chaotic nature of this system.

Figure 10 further investigates the uncertainty quantification performance of the HierGP by showing the forward runs of 50 posterior sample draws on Θ for the 3D Lorenz system in x-, y-, and z-coordinates. Initially, we see that the recovered trajectories using the HierGP have low uncertainties, with all sample paths quite close to each other. However, as time progresses, we see much higher posterior uncertainties, with sample paths growing further apart as uncertainty accumulates over time. This is not surprising given the chaotic nature of the system and its error propagation over time and again suggests that long-term prediction of such systems is a challenging problem.

6. Conclusion. We proposed in this work a novel hierarchical shrinkage Gaussian process (HierGP), which embeds the principles of effect hierarchy (Hamada and Wu, 1992), heredity (Haris et al., 2016), and smoothness (Ding et al., 2020) within carefully constructed cumulative shrinkage priors in a Gaussian process. Similar to the use of such principles for classical

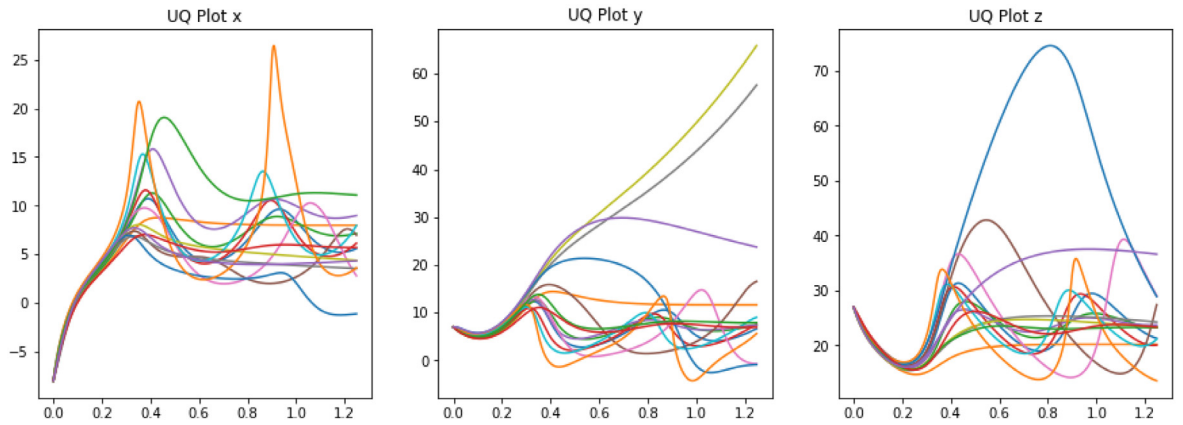

Figure 10. *Visualizing the forward runs of* 50 *posterior sample draws from the HierGP for the* 3*D Lorenz system in x-, y-, and z-coordinates (from left to right).*

analysis of experiments (see, e.g., Wu and Hamada, 2009), the embedding of this structured shrinkage within a Bayesian nonparametric model can allow for improved predictive performance with limited experimental data. We then derived efficient posterior sampling algorithms for model training and prediction and proved desirable consistency results for the HierGP. Numerical experiments confirmed the improved performance of the HierGP over existing models for both computer code emulation and dynamical systems recovery.

Despite promising results, there are many avenues for fruitful future work. One direction is in establishing posterior contraction rates for the HierGP. In the case where f has the presumed effect structure, it would be interesting to explore whether the HierGP rates improve upon standard contraction rates for GPs, which are known to suffer from a curse ofdimensionality (van der Vaart and van Zanten, 2008). Another direction is in exploring cumulative shrinkage priors that can capture a weaker form of effect heredity (see, e.g., Wu and Hamada, 2009; Mak and Wu, 2019b), where interaction effects can be significant when at least one component effect is significant. This can provide a more flexible model in cases where there may be minor violations to the effect principles. Finally, the exploration of the HierGP for factor screening would be of interest; such a direction requires further investigation of variable selection consistency for our model.

REFERENCES

A. Alexanderian (2015), *A Brief Note on the Karhunen-Loeve Expansion*, preprint, arXiv:1509.07526.

A. Armagan, D. B. Dunson, and J. Lee (2013), *Generalized double Pareto shrinkage*, Statist. Sinica, 23, pp. 119–143.

G. Berkooz, P. Holmes, and J. L. Lumley (1993), *The proper orthogonal decomposition in the analysis of turbulent flows*, Annu. Rev. Fluid Mech., 25, pp. 539–575.

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson (2015), *Dirichlet–Laplace priors for optimal shrinkage*, J. Amer. Statist. Assoc., 110, pp. 1479–1490.

G.-J. Both, S. Choudhury, P. Sens, and R. Kusters (2021), *DeepMoD: Deep learning for model discovery in noisy data*, J. Comput. Phys., 428, 109985.

S. L. Brunton, J. L. Proctor, and J. N. Kutz (2016), *Discovering governing equations from data by sparse identification of nonlinear dynamical systems*, Proc. Natl. Acad. Sci. USA, 113, pp. 3932–3937.

E. Buckingham (1914), *On physically similar systems: Illustrations of the use of dimensional equations*, Phys. Rev., 4, pp. 345–376.
H.-J. Bungartz and M. Griebel (2004), *Sparse grids*, Acta Numer., 13, pp. 147–269.
C. M. Carvalho, N. G. Polson, and J. G. Scott (2009), *Handling sparsity via the horseshoe*, in Artificial Intelligence and Statistics, PMLR, pp. 73–80.
C. M. Carvalho, N. G. Polson, and J. G. Scott (2010), *The horseshoe estimator for sparse signals*, Biometrika, 97, pp. 465–480.
I. Castillo, J. Schmidt-Hieber, and A. Van der Vaart (2015), *Bayesian linear regression with sparse priors*, Ann. Statist., 43, pp. 1986–2018.
K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz (2020), *A unified sparse optimization framework to learn parsimonious physics-informed models from data*, IEEE Access, 8, pp. 169259–169271.
J. Chen, S. Mak, V. R. Joseph, and C. Zhang (2021), *Function-on-function kriging, with applications to three-dimensional printing of aortic tissues*, Technometrics, 63, pp. 384–395.
H. Cheng and A. Sandu (2010), *Collocation least-squares polynomial chaos method*, in Proceedings of the 2010 Spring Simulation Multiconference, pp. 1–6.
T. Choi and M. J. Schervish (2007), *On posterior consistency in nonparametric regression problems*, J. Multivariate Anal., 98, pp. 1969–1987.
N. Cressie and G. Johannesson (2008), *Fixed rank kriging for very large spatial data sets*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 70, pp. 209–226.
C. B. Delahunt and J. N. Kutz (2022), *A toolkit for data-driven discovery of governing equations in high-noise regimes*, IEEE Access, 10, pp. 31210–31234.
J. Dick, F. Y. Kuo, and I. H. Sloan (2013), *High-dimensional integration: The quasi-Monte Carlo way*, Acta Numer., 22, pp. 133–288.
L. Ding, S. Mak, and C. F. J. Wu (2019), *BdryGP: A New Gaussian Process Model for Incorporating Boundary Information*, preprint, arXiv:1908.08868.
Y. Ding, F. Hickernell, P. Kritzer, and S. Mak (2020), *Adaptive approximation for multivariate linear problems with inputs lying in a cone,* in Multivariate Algorithms and Information-Based Complexity, De Gruyter, Berlin, pp. 109–145.
D. B. Dunson, H.-T. Wu, and N. Wu (2022), *Graph based Gaussian processes on restricted domains*, J. R. Stat. Soc. Ser. B. Stat. Methodol., 84, pp. 414–439.
D. K. Duvenaud, H. Nickisch, and C. Rasmussen (2011), *Additive Gaussian processes*, in Adv. Neural Inf. Process. Syst. 24.
F. Ferrari and D. B. Dunson (2021), *Bayesian factor analysis for inference on interactions*, J. Amer. Statist. Assoc., 116, pp. 1521–1532.
A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin (1995), *Bayesian Data Analysis*, Chapman and Hall/CRC.
A. Ghadami and B. I. Epureanu (2022), *Data-driven prediction in dynamical systems: Recent developments*, Philos. Trans. Roy. Soc. A, 380, 20210213.
R. G. Ghanem and P. D. Spanos (1991), *Stochastic finite element method: Response statistics*, in Stochastic Finite Elements: A Spectral Approach, Springer, pp. 101–119.
S. Golchi, D. R. Bingham, H. Chipman, and D. A. Campbell (2015), *Monotone emulation of computer experiments*, SIAM/ASA J. Uncertain. Quantif., 3, pp. 370–392, https://doi.org/10.1137/140976741.
R. B. Gramacy (2020), *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*, Chapman and Hall/CRC.
J. Guckenheimer and P. Holmes (2013), *Nonlinear Oscillations, Dynamical Systems, and Bifurcations of Vector Fields*, Appl. Math. Sci. 42, Springer Science & Business Media.
M. Hamada and C. F. J. Wu (1992), *Analysis of designed experiments with complex aliasing*, J. Qual. Technol., 24, pp. 130–137.
O. Harari, D. Bingham, A. Dean, and D. Higdon (2018), *Computer experiments: Prediction accuracy, sample size and model complexity revisited*, Statist. Sinica, pp. 899–919.
A. Haris, D. Witten, and N. Simon (2016), *Convex modeling of interactions with strong heredity*, J. Comput. Graph. Statist., 25, pp. 981–1004.

H. Ishwaran and J. S. Rao (2005), *Spike and slab variable selection: Frequentist and Bayesian strategies*, Ann. Statist., 33, pp. 730–773.
S. Jeong and S. Ghosal (2021), *Unified Bayesian theory of sparse linear regression with nuisance parameters*, Electron. J. Stat., 15, pp. 3040–3111.
Y. Ji, S. Mak, D. Soeder, J. F. Paquet, and S. A. Bass (2024), *A graphical multi-fidelity Gaussian process model, with application to emulation of heavy-ion collisions*, Technometrics, 66, pp. 267–281.
J. Johndrow, P. Orenstein, and A. Bhattacharya (2020), *Scalable approximate MCMC algorithms for the horseshoe prior*, J. Mach. Learn. Res., 21, 73.
C. G. Kaufman, D. Bingham, S. Habib, K. Heitmann, and J. A. Frieman (2011), *Efficient emulators of computer experiments using compactly supported correlation functions, with an application to cosmology*, Ann. Appl. Stat., 5, pp. 2470–2492.
S. Legramanti, D. Durante, and D. B. Dunson (2020), *Bayesian cumulative shrinkage for infinite factorizations*, Biometrika, 107, pp. 745–752.
K. Li, S. Mak, J.-F. Paquet, and S. A. Bass (2023), *Additive Multi-index Gaussian Process Modeling, with Application to Multi-physics Surrogate Modeling of the Quark-Gluon Plasma*, preprint, arXiv:2306.07299.
J. L. Loeppky, J. Sacks, and W. J. Welch (2009), *Choosing the sample size of a computer experiment: A practical guide*, Technometrics, 51, pp. 366–376.
E. N. Lorenz (1963), *Deterministic nonperiodic flow*, J. Atmospheric. Sci., 20, pp. 130–141.
X. Lu, A. Boukouvalas, and J. Hensman (2022), *Additive Gaussian processes revisited*, in International Conference on Machine Learning, PMLR, pp. 14358–14383.
Y. Luo, K. Ogle, C. Tucker, S. Fei, C. Gao, S. LaDeau, J. S. Clark, and D. S. Schimel (2011), *Ecological forecasting and data assimilation in a data-rich era*, Ecol. Appl., 21, pp. 1429–1442.
N. Lüthen, S. Marelli, and B. Sudret (2021), *Sparse polynomial chaos expansions: Literature survey and benchmark*, SIAM/ASA J. Uncertain. Quantif., 9, pp. 593–649, https://doi.org/10.1137/20M1315774.
S. Mak, C.-L. Sung, X. Wang, S.-T. Yeh, Y.-H. Chang, V. R. Joseph, V. Yang, and C. F. J. Wu (2018), *An efficient surrogate model for emulation and physics extraction of large eddy simulations*, J. Amer. Statist. Assoc., 113, pp. 1443–1456.
S. Mak and C. F. J. Wu (2019a), *Analysis-of-marginal-tail-means (ATM): A robust method for discrete black-box optimization*, Technometrics, 61, pp. 545–559.
S. Mak and C. F. J. Wu (2019b), *cmenet: A new method for bi-level variable selection of conditional main effects*, J. Amer. Statist. Assoc., 114, pp. 844–856.
P. McCullagh and J. A. Nelder (1989), *Generalized Linear Models,* Monogr. Statist. Appl. Probab. 37, Chapman & Hall.
M. Mudelsee (2019), *Trend analysis of climate time series: A review of methods*, Earth-Sci. Rev., 190, pp. 310–322.
J. Nelder (1977), *A reformulation of linear models*, J. Roy. Statist. Soc. Ser. A, 140, pp. 48–76.
A. B. Owen (1997), *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25, pp. 1541–1562.
C. E. Rasmussen and C. K. I. Williams (2005), *Gaussian Processes for Machine Learning*, The MIT Press.
V. Ročková and E. I. George (2018), *The spike-and-slab lasso*, J. Amer. Statist. Assoc., 113, pp. 431–444.
T. Savitsky, M. Vannucci, and N. Sha (2011), *Variable selection for nonparametric Gaussian process priors: Models and computational strategies*, Statist. Sci., 26, pp. 130–149.
F. Scheipl, L. Fahrmeir, and T. Kneib (2012), *Spike-and-slab priors for function selection in structured additive regression models*, J. Amer. Statist. Assoc., 107, pp. 1518–1532.
P. Seshadri, S. Yuchi, and G. T. Parks (2019), *Dimension reduction via Gaussian ridge functions*, SIAM/ASA J. Uncertain. Quantif., 7, pp. 1301–1322, https://doi.org/10.1137/18M1168571.
A. Sobester, A. Forrester, and A. Keane (2008), *Engineering Design via Surrogate Modeling: A Practical Guide*, John Wiley & Sons.
Q. Song and F. Liang (2023), *Nearly optimal Bayesian shrinkage for high-dimensional regression*, Sci. China Math., 66, pp. 409–442.
M. L. Stein (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, Springer Science & Business Media.
A. W. van der Vaart and J. H. van Zanten (2008), *Rates of contraction of posterior distributions based on Gaussian process priors*, Ann. Statist., 36, pp. 1435–1463.

L. Wang (2008), *Karhunen-Loève Expansions and Their Applications*, London School of Economics and Political Science.

X. Wang and J. O. Berger (2016), *Estimating shape constrained functions using Gaussian processes*, SIAM/ASA J. Uncertain. Quantif., 4, pp. 1–25, https://doi.org/10.1137/140955033.

M. W. Wheeler, D. B. Dunson, S. P. Pandalai, B. A. Baker, and A. H. Herring (2014), *Mechanistic hierarchical Gaussian processes*, J. Amer. Statist. Assoc., 109, pp. 894–904.

C. F. J. Wu and M. S. Hamada (2009), *Experiments: Planning, Analysis, and Optimization*, John Wiley & Sons.

D. Xiu (2010), *Numerical Methods for Stochastic Computations: A Spectral Method Approach*, Princeton University Press.

S.-T. Yeh, X. Wang, C.-L. Sung, S. Mak, Y.-H. Chang, L. Zhang, C. F. J. Wu, and V. Yang (2018), *Common proper orthogonal decomposition-based spatiotemporal emulator for design exploration*, AIAA J., 56, pp. 2429–2442.

A. Zammit-Mangion and M. Sainsbury-Dale (2023), *FRK: Fixed Rank Kriging*, R package version 2.2.0.