SIAM/ASA J. Uncertainty Quantification
Vol. 5, pp. 598–620

Flexible Correlation Structure for Accurate Prediction and Uncertainty Quantification in Bayesian Gaussian Process Emulation of a Computer Model*

Hao Chen†, Jason L. Loeppky‡, and William J. Welch§

Abstract. Gaussian processes are widely used in the analysis of data from a computer model. Ideally, the analysis will yield accurate predictions with correct coverage probabilities of credible intervals. In this paper, we first review several existing Bayesian implementations in the literature. We show that Bayesian approaches with squared-exponential correlation structure do not always quantify well the uncertainty in prediction. Thus, we propose new Bayesian approaches with power-exponential or Matérn correlation structure, which have more flexibility. Through application examples and a simulation study, we show that the proposed Bayesian methods not only have superior prediction accuracy but are closer to having the correct coverage probability.

Key words. Bayesian predictive distribution, computer experiment, coverage probability, Matérn correlation, power-exponential correlation, squared-exponential correlation

AMS subject classifications. 60G15, 62F15, 62G15, 62M20

DOI. 10.1137/15M1008774

1. Introduction.

Many complex phenomena are expensive or difficult to model through controlled physical experiments. Computer models, also called codes, have therefore become important alternatives for providing insights into such phenomena. A computer model takes vector-valued input $\mathbf{x}$ and produces scalar output $y(\mathbf{x})$. It often has the following features: (1) it is deterministic, that is, rerunning with the same input yields identical output, and (2) it is time-consuming, with one run possibly taking several hours or even days to complete. A computer experiment is a designed set of n runs of the model at n configurations of $\mathbf{x}$.

A primary goal of analysis of a computer experiment is to use the available runs to predict $y(\mathbf{x})$ at untried $\mathbf{x}$ via a fast-running statistical model or emulator. Gaussian processes (GPs) are a popular class of emulators [20]. A GP has mean μ, variance σ^2, and a correlation function $R(\mathbf{x}, \mathbf{x}')$ that quantifies the dependence of the output values $y(\mathbf{x})$ and $y(\mathbf{x}')$ at two input configurations, $\mathbf{x}$ and $\mathbf{x}'$. The mean, μ, is usually a regression function of $\mathbf{x}$ of known form but with unknown values of parameters $\boldsymbol{\beta}$. Popular choices of the correlation function, R, in the literature are the squared-exponential (Gaussian), Matérn, and power-exponential

*Received by the editors February 17, 2015; accepted for publication (in revised form) November 14, 2016; published electronically July 12, 2017.
http://www.siam.org/journals/juq/5/M100877.html
Funding: The research of the second author and the third author was supported by grants RGPIN-2015-03895 and RGPIN-2014-04962, respectively, from the Natural Sciences and Engineering Research Council, Canada.
†Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (hao.chen@stat.ubc.ca).
‡Statistics, University of British Columbia, Kelowna, BC V1V 1V7, Canada (jason.loeppky@ubc.ca).
§Corresponding author. Department of Statistics, University of British Columbia, Vancouver, BC V6T 1Z4, Canada (will@stat.ubc.ca).

functions, which will be described in the next section. All three correlation functions have a vector of parameters, $\boldsymbol{\theta}$, controlling sensitivity of $y(\mathbf{x})$ with respect to each input in $\mathbf{x}$. The Matérn and power-exponential functions have further parameters governing the smoothness of $y(\mathbf{x})$ with respect to each input. The smoothness parameters add flexibility to the types of functions that can be emulated, an important focus of this paper. Conditional on all correlation parameters, the posterior mean of y at an untried input vector is used as the predictor, which can also be viewed as the best linear unbiased predictor that minimizes the mean squared error (MSE) of prediction.

In practice, the GP parameters, $\boldsymbol{\beta}$, σ^2, and the correlation parameters are unknown and hence need to be estimated using the available data. Well-known estimation methods can be broadly classified into two categories based on the underlying paradigm. The first is the empirical Bayes method, where maximum likelihood estimates (MLEs) are used; see, for example, the algorithm of Welch et al. [24]. The second category is Bayesian posterior inference, such as that proposed by Higdon et al. [9] and the treed GP method used by Gramacy and Lee [7]. The Bayesian methods, in principle, take account of parameter-estimation uncertainty and produce better coverage probabilities for credible intervals. However, Chen et al. [5] noted that this is not always true when a GP with a squared-exponential correlation function is used. That work motivated the current paper, which proposes new Bayesian methods with more flexible, power-exponential or Matérn, correlation structures.

Thus, the focus here is quantifying the uncertainty from the statistical emulator of the computer model, including the contribution from parameter estimation. There are other sources of uncertainty, such as the systematic discrepancy between the computer model and physical measurements of the system [11], that are not explicitly considered here. The proposed methods have relevance, however, because modeling the computer-model data is an important component of analyses addressing other objectives.

The rest of the paper is organized as follows. In section 2 the GP model, estimation of its parameters, and prediction methods are reviewed. An example motivating new Bayesian methods is analyzed in section 3. In section 4, we propose Bayesian methods with power-exponential or Matérn correlation structure. These methods are evaluated in section 5 through application examples and a simulation study. Some concluding remarks are made in section 6.

2. Gaussian process model, prediction, and parameter estimation.

2.1. Gaussian process model. Sacks et al. [20] introduced GP models for the analysis of computer experiments by treating the deterministic output function $y(\mathbf{x})$ as a realization of $\mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + Z(\mathbf{x})$. Here, $\mathbf{f}^T(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$ contains k known regression functions, $\boldsymbol{\beta}$ is a vector of regression parameters with unknown values, and $Z(\mathbf{x})$ is a stationary GP with zero mean and unknown variance σ^2. Chen et al. [5] found that the prediction accuracy of a GP with a constant regression function is no worse and sometimes better than when linear trend terms or polynomial terms are included. Therefore, we restrict our attention to GP models with $k = 1$, $f(\mathbf{x}) = 1$, and $\boldsymbol{\beta} = \mu$ for the constant mean of the process. Thus, the GP model simplifies to

$$\mu + Z(\mathbf{x}). \tag{2.1}$$

The correlation structure of the GP model is critical to this approach. Let $\mathbf{x}$ and $\mathbf{x}'$ be

two configurations for the inputs. The correlation between $Z(\mathbf{x})$ and $Z(\mathbf{x}')$ is denoted by the function $R(\mathbf{x}, \mathbf{x}')$. All the correlation functions considered are products of 1-dimensional functions, i.e.,

$$R(\mathbf{x}, \mathbf{x}') = \prod_{j=1}^{d} R_j(h_j),$$

where $h_j = x_j - x_j'$ is a distance for input j. The three popular choices for $R_j(h_j)$ to be investigated in this paper are as follows.

- Squared-exponential (SqExp). The SqExp correlation function, also called Gaussian, is

$$R_j^{\text{SqExp}}(h_j) = \exp\left(-\theta_j h_j^2\right) \quad (j = 1, \ldots, d), \tag{2.2}$$

  with $\theta_j \geq 0$. The sensitivity parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)$ control how fast the correlation decays as the distance between $\mathbf{x}$ and $\mathbf{x}'$ increases. The process is infinitely differentiable with respect to each x_j, i.e., it is extremely smooth, because of the fixed exponent of 2 in the distance metric. SqExp has been used in many applications; see [7] and [11], for instance.
- Power-exponential (PowExp). The PowExp correlation function generalizes the distance metric in SqExp:

$$R_j^{\text{PowExp}}(h_j) = \exp\left(-\theta_j |h_j|^{\alpha_j}\right) \quad (j = 1, \ldots, d), \tag{2.3}$$

  with $\theta_j \geq 0$ and $1 \leq \alpha_j \leq 2$. The introduced parameters, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$, govern smoothness, and hence PowExp is a more flexible family than SqExp.
- Matérn. Another popular choice is the Matérn function. It is often defined as

$$R_j^{\text{Matérn}}(h_j) = \frac{1}{\Gamma(\nu_j) 2^{(\nu_j - 1)}} \left(\frac{2\sqrt{\nu_j}|h_j|}{\tilde{\theta}_j} \right)^{\nu_j} K_{\nu_j} \left(\frac{2\sqrt{\nu_j}|h_j|}{\tilde{\theta}_j} \right),$$

  where Γ is the Gamma function and K_{ν_j} is the modified Bessel function of order ν_j. Here $\tilde{\theta}_j$ is a sensitivity parameter, and ν_j controls smoothness.
  To be consistent with the parameterization of the sensitivity parameters of SqExp and PowExp, we redefine $\tilde{\theta}_j$ as $\theta_j = 2\sqrt{\nu_j}/\tilde{\theta}_j$. Following [5], for simplicity we also consider special cases defined by ν_j. If $\nu_j = \delta_j + 1/2$ with integer $\delta_j \geq 0$, there are δ_j derivatives with respect to x_j. The four special cases considered are

$$R_j^{\text{Matérn}}(h_j) = \begin{cases} \exp(-\theta_j|h_j|) & (\delta_j = 0 \text{ derivatives}), \\ \exp(-\theta_j|h_j|)(\theta_j|h_j| + 1) & (\delta_j = 1 \text{ derivative}), \\ \exp(-\theta_j|h_j|)(\frac{1}{3}(\theta_j|h_j|)^2 + \theta_j|h_j| + 1) & (\delta_j = 2 \text{ derivatives}), \\ \exp(-\theta_j|h_j|^2) & (\delta_j \to \infty \text{ derivatives}). \end{cases} \tag{2.4}$$

  The last case gives SqExp.

In general, $\boldsymbol{\psi}$ will denote the set of correlation parameters to be estimated: $(\theta_1, \ldots, \theta_d)$ for SqExp, $(\theta_1, \ldots, \theta_d, \alpha_1, \ldots, \alpha_d)$ for PowExp, or $(\theta_1, \ldots, \theta_d, \delta_1, \ldots, \delta_d)$ for Matérn.

2.2. Prediction. The GP model leads to a predictive distribution for y at a new input configuration $\mathbf{x}^*$. Suppose the computer model has been run n times at input vectors $\mathbf{x}^{(1)}$, $\ldots$, $\mathbf{x}^{(n)}$ to produce n output values $\mathbf{y} = \left(y(\mathbf{x}^{(1)}), y(\mathbf{x}^{(2)}), \ldots, y(\mathbf{x}^{(n)})\right)^T$. The predictive distribution of $y(\mathbf{x}^*)$ conditional on μ, σ^2, $\boldsymbol{\psi}$, and $\mathbf{y}$ is Gaussian, i.e., $N\left(m(\mathbf{x}^*), v(\mathbf{x}^*)\right)$, where

$$m(\mathbf{x}^*) = \mu + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\mu) \tag{2.5}$$

and

$$v(\mathbf{x}^*) = \sigma^2\left(1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right). \tag{2.6}$$

Here, the $n \times 1$ vector $\mathbf{r}(\mathbf{x}^*)$ is obtained from one of the correlation functions in (2.2), (2.3), or (2.4) with element i given by $R(\mathbf{x}^*, \mathbf{x}^{(i)})$ for $i = 1, \ldots, n$; element (i, j) of the $n \times n$ matrix $\mathbf{R}$ is given by $R(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$ for $i, j = 1, \ldots, n$; and $\mathbf{1}$ is an $n \times 1$ vector with all elements equal to 1.

2.3. Parameter estimation. In practice, the parameters μ, σ^2, and $\boldsymbol{\psi}$ have to be estimated, leading to further variability in the predictive distribution.

2.3.1. Empirical Bayes. Empirical Bayes estimates the correlation parameters $\boldsymbol{\psi}$ by maximum likelihood.

Based on (2.1), the likelihood for μ, σ^2, and $\boldsymbol{\psi}$ is multivariate normal:

$$L(\mathbf{y}|\mu, \sigma^2, \boldsymbol{\psi}) = \frac{1}{(2\pi\sigma^2)^{n/2}|\mathbf{R}|^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{1}\mu)^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\mu)\right), \tag{2.7}$$

where $|\mathbf{R}|$ is the determinant of $\mathbf{R}$. For any fixed $\boldsymbol{\psi}$ and hence $\mathbf{R}$, differentiating with respect to μ and σ^2 gives their MLEs:

$$\hat{\mu} = (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{y} \tag{2.8}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{y} - \mathbf{1}\hat{\mu})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}). \tag{2.9}$$

Note that the MLEs for μ and σ^2 depend on the parameters $\boldsymbol{\psi}$ through $\mathbf{R}$. Plugging $\hat{\mu}$ and $\hat{\sigma}^2$ into (2.7) gives the profile likelihood,

$$\frac{1}{(2\pi\hat{\sigma}^2)^{n/2}|\mathbf{R}|^{1/2}} e^{-n/2}.$$

It needs to be numerically maximized to yield MLEs for $\boldsymbol{\psi}$.

To obtain predictions, the MLEs of μ, σ^2, and $\boldsymbol{\psi}$ are substituted for the true parameter values in (2.5). Extra uncertainty is introduced by use of estimates [1], but only the contribution from $\hat{\mu}$ is easily quantified. The estimated predictive mean is

$$\hat{m}_{\boldsymbol{\psi}}(\mathbf{x}^*) = \hat{\mu} + \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}), \tag{2.10}$$

and the estimate of the predictive variance in (2.6) becomes

$$\hat{v}_{\boldsymbol{\psi}}(\mathbf{x}^*) = \hat{\sigma}^2 \left(1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) + \frac{\left(1 - \mathbf{1}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right)^2}{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}} \right). \tag{2.11}$$

The subscript $\boldsymbol{\psi}$ in the notation for $\hat{m}_{\boldsymbol{\psi}}(\mathbf{x}^*)$ and $\hat{v}_{\boldsymbol{\psi}}(\mathbf{x}^*)$ emphasizes that these quantities depend on $\boldsymbol{\psi}$.

2.3.2. Bayes with a squared-exponential correlation (version K). Two full Bayesian implementations based on the SqExp correlation structure will be described. The first is by Kennedy [10], and hence we call it SqExp-Full (K). It has the following independent prior distributions on the GP parameters:

- an improper uniform distribution (normal with infinite variance) on μ;
- a Jeffreys prior on σ^2, i.e., $\pi(\sigma^2) \propto 1/\sigma^2$; and
- an exponential distribution with rate 0.1 on each correlation parameter θ_j.

By integrating out μ and σ^2, Handcock and Stein [8] gave the marginal posterior distribution of the SqExp correlation parameters $\boldsymbol{\theta}$ as

$$\pi(\boldsymbol{\theta}|\mathbf{y}) \propto \int\int \pi(\mu)\pi(\sigma^2)\pi(\boldsymbol{\theta})L(\mathbf{y}|\mu,\sigma^2,\boldsymbol{\theta})\,d\mu\,d\sigma^2 \propto \frac{\prod_{i=1}^d \pi(\theta_j)}{\left(\hat{\sigma}^2_{\boldsymbol{\theta}}\right)^{\frac{n-1}{2}} |\mathbf{R}|^{1/2}(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{1/2}},$$

where $\pi(\theta_j)$ is the prior distribution of θ_j, $\hat{\sigma}^2_{\boldsymbol{\theta}}$ is given by

$$\hat{\sigma}^2_{\boldsymbol{\theta}} = \frac{1}{n-1}(\mathbf{y} - \mathbf{1}\hat{\mu})^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}), \tag{2.12}$$

and $\hat{\mu}$ is as in (2.8). Except for a change in degrees of freedom, $\hat{\sigma}^2_{\boldsymbol{\theta}}$ in (2.12) is the same as $\hat{\sigma}^2$ in (2.9). The Metropolis–Hastings algorithm is then applied to sample from the posterior distribution of $\boldsymbol{\theta}$. The predictive distribution conditional on $\boldsymbol{\theta}$ is a noncentral t distribution with $n-1$ degrees of freedom [8, 10]. That is,

$$y(\mathbf{x}^*)|\boldsymbol{\theta},\mathbf{y} \sim t(\hat{m}_{\boldsymbol{\theta}}(\mathbf{x}^*), \hat{v}_{\theta}(\mathbf{x}^*)), \tag{2.13}$$

where $\hat{m}_{\boldsymbol{\theta}}(\mathbf{x}^*)$ and $\hat{v}_{\boldsymbol{\theta}}(\mathbf{x}^*)$ are as in (2.10) and (2.11) with $\boldsymbol{\psi} = \boldsymbol{\theta}$.

2.3.3. Bayes with a squared-exponential correlation (version H). A second full Bayesian implementation based on the SqExp correlation function was described by Higdon et al. [9], and we call it SqExp-Full (H). The independent prior distributions on the GP parameters are now the following:

- an improper uniform distribution (normal with infinite variance) on μ;
- an inverse-gamma distribution IG(ϕ_1, ϕ_2) on σ^2 with shape parameter $\phi_1 = 1$ and scale $\phi_2 = 0.0001$; and
- a beta$(1, 0.1)$ distribution on $\rho_i = \exp(-\theta_j/4)$.

A beta prior on ρ_j is equivalent to a log-beta distribution on θ_j.

The parameters μ and σ^2 can be integrated out of the posterior [9], as described further in section 4.5, and a Markov chain Monte Carlo (MCMC) algorithm is applied to sample

the ρ_j. The predictive distribution at an untried input vector is the same as (2.13) but with $2\phi_1 + n - 1$ as the degrees of freedom in the denominator of

$$\hat{\sigma}^2_{\boldsymbol{\theta}} = \frac{1}{2\phi_1 + n - 1}(\mathbf{y} - \mathbf{1}\hat{\mu})^T \mathbf{R}^{-1}(\mathbf{y} - \mathbf{1}\hat{\mu}) \tag{2.14}$$

and in the noncentral t distribution in (2.13).

Chen [4] showed that the different priors on σ^2 in SqExp-Full (K) and SqExp-Full (H) do not substantially affect predictive properties, but the priors for θ_j are an important distinction. Besides these two full Bayesian methods, there exist several other full Bayesian implementations, for example, the treed GP (TGP) method [7]. The aim of the current paper, however, is not to compare existing estimation methods; a comprehensive comparison of Bayesian methods can be found in [4]. Rather, the purpose is to introduce Bayesian implementations with flexible correlation structures.

3. A motivating example: Nilson–Kuusk code. Abt [1] noted that the extra uncertainty from the use of empirical Bayes plug-in MLEs of the correlation parameters can be nontrivial. In principle, a Bayesian method should quantify parameter-estimation uncertainty and thus produce better coverage probabilities of credible intervals. We show in this section, however, that empirical Bayes versus full Bayes may be less important for uncertainty quantification than the correlation-function family.

We consider an ecological computer code that models reflectance for a plant canopy. It was developed by Nilson and Kuusk [15] and analyzed by Bastos and O'Hagan [2] to illustrate diagnostics for GP emulators. Two computer experiments are available for this code with $d = 5$ inputs: the first is a 150-run random Latin hypercube design (LHD), and the second is an independent 100-run random LHD. We randomly sample 100 runs from the 150-run set as the training set to train a GP model. The remaining 50 runs are added to the 100-run set to form a 150-run hold-out (testing) set. Twenty-five different random samples are used so that each method considered below will produce 25 repeat experiments and 25 sets of results. This approach is similar to that of Chen et al. [5], who noted that the variation in results from one design to another can be considerable. Hence, throughout this paper we report results from multiple designs, either by random sampling or by permuting the columns of a base design.

Following along the lines of [5], a Gaussian process with a constant regression term is chosen as the statistical model to emulate the Nilson–Kuusk code. We consider the following four methods, which have different combinations of correlation structure and inference paradigm:

- SqExp-Emp, i.e., SqExp correlation and empirical Bayes;
- SqExp-Full (K), i.e., SqExp correlation and full Bayes with priors from [10];
- SqExp-Full (H), i.e., SqExp correlation and full Bayes with priors from [9]; and
- PowExp-Emp, i.e., PowExp correlation and empirical Bayes.

To measure the prediction accuracy of a method, we use normalized root mean squared error (RMSE) of prediction over the hold-out set ($e_{\text{rmse,ho}}$) and normalized maximum absolute error ($e_{\text{max,ho}}$) [5]:

$$e_{\text{rmse,ho}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{y}(\mathbf{x}_{\text{ho}}^{(i)}) - y(\mathbf{x}_{\text{ho}}^{(i)})\right)}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\bar{y} - y(\mathbf{x}_{\text{ho}}^{(i)})\right)}} \quad \text{and} \quad e_{\text{max,ho}} = \frac{\max|\hat{y}(\mathbf{x}_{\text{ho}}^{(i)}) - y(\mathbf{x}_{\text{ho}}^{(i)})|}{\max|\bar{y} - y(\mathbf{x}_{\text{ho}}^{(i)})|},$$

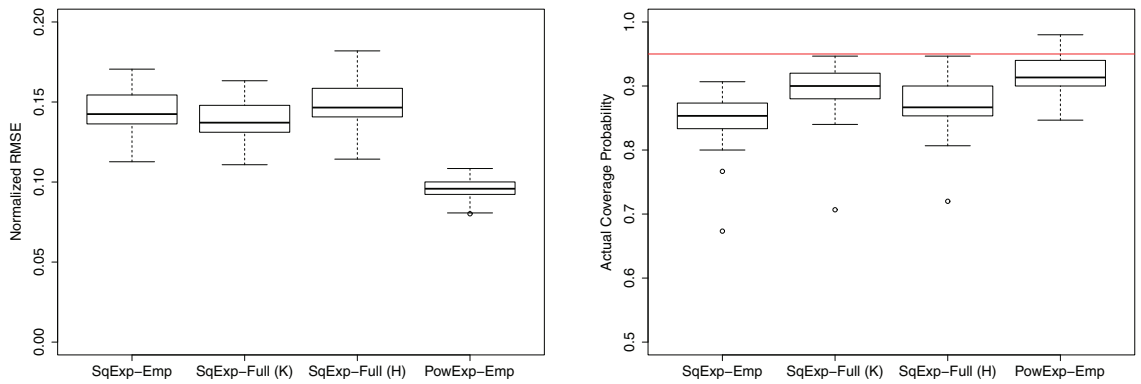

Figure 1. *Normalized RMSE (left panel) and actual coverage probability (right panel) for the Nilson–Kuusk code from four methods: SqExp-Emp, SqExp-Full (K), SqExp-Full (H), and PowExp-Emp. The boxplots show the results from* 25 *random training-test data splits. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

where N is the number of runs in the hold-out set, $\mathbf{x}_{\text{ho}}^{(i)}$ is point i in the hold-out set, $\hat{y}(\mathbf{x}_{\text{ho}}^{(i)})$ is the predicted value, and $\bar{y}$ is the trivial predictor, i.e., the mean value of y in the training set. The normalization in the denominator puts RMSE roughly into $[0, 1]$ whatever the scale of y. A number equal to or greater than 1 indicates that the prediction performance is no better or even worse than the trivial predictor, $\bar{y}$. The normalized maximum absolute error measures the worst case, and the interpretation of its scale is similar.

How well a method quantifies the uncertainty is measured by the actual coverage probability (ACP) [5], the proportion of runs in the hold-out set where the $1-\alpha$ confidence/credible interval contains the true y. We set $\alpha = 0.05$ throughout for nominal (correct) coverage probability of 0.95. A number much less than 0.95 is called undercoverage, which indicates that the method does not fully quantify uncertainty.

Results for the Nilson–Kuusk example are presented in Figure 1. In the left panel the boxplots of normalized RMSE over 25 repeat experiments show that the best prediction accuracy among the four competing methods is from PowExp-Emp; it clearly outperforms all three methods with SqExp structure. The normalized maximum absolute error results in the supplemental material (M100877_01.pdf [local/web 176KB]) follow the same pattern. The ACP results in the right panel of Figure 1 show that the three methods with SqExp structure undercover. PowExp-Emp has the best performance again.

These accuracy and coverage results are likely due to the effect of the fifth input to the Nilson–Kuusk code. Its estimated main effect [22] from a PowExp-Emp analysis accounts for about 90% of the total variance of the predicted output over the 5-dimensional input space. The median of the 25 MLEs of α_5 in the PowExp correlation function is 1.85, which deviates from the fixed power of 2 in SqExp. All of this suggests that full Bayesian methods with SqExp structure need to be generalized to incorporate more flexibility to deal with a computer model like that of Nilson and Kuusk, to achieve better accuracy and better uncertainty quantification.

Unfortunately, PowExp-Emp is not the panacea: it does not always work as well as it

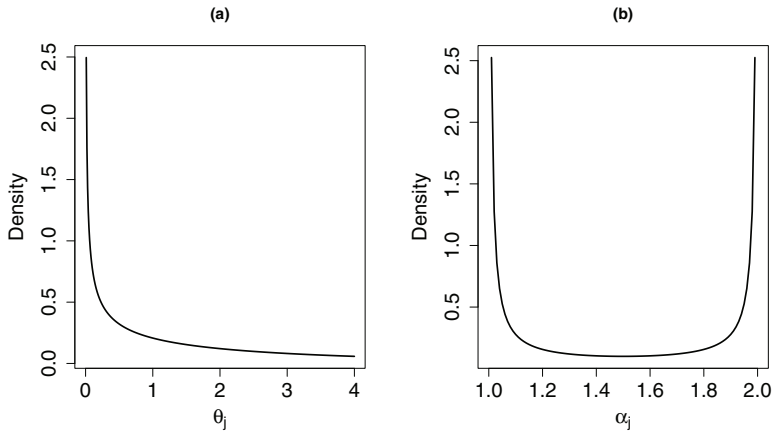

Figure 2. *Prior densities:* (a) *prior on* θ_j *and* (b) *prior on* α_j.

does for the Nilson–Kuusk example. Chen et al. [5] noted that its ACP is much worse than that of the SqExp-Full (K) method for the borehole function, an example in section 5.3. New methods are needed.

4. New Bayesian methods. In this section we outline new Bayesian methods for the PowExp and Matérn correlation functions. We first describe the priors for the parameters treated in a fully Bayesian way; any further correlation parameters for a particular method are estimated by empirical Bayes. We then describe a general approach to implement all methods by outlining the marginal posterior distribution of all correlation parameters estimated by Bayes' rule and the algorithm for sampling them via MCMC.

4.1. Priors for μ, σ^2, and θ. Similar to SqExp-Full (version H), all the new methods treat μ, σ^2, and the sensitivity parameters $\boldsymbol{\theta}$ in a fully Bayesian way. The independent priors are as follows:

- a uniform distribution, $U(-1000, 1000)$, on μ, which approximates an improper normal distribution;
- an inverse-gamma $\mathrm{IG}(\phi_1, \phi_2)$ distribution on σ^2, where $\phi_1 \to 0$ and $\phi_2 \to 0$ (equivalent to the Jeffreys prior); and
- an independent beta$(1, 0.5)$ distribution on $\rho_j = \exp(-\theta_j/4)$ for $j = 1, \ldots, d$.

A beta$(1, 0.5)$ prior on the ρ_j scale is equivalent to assuming a log-beta distribution for θ_j:

$$\pi(\theta_j) = \frac{1}{8} \exp\left(\frac{-\theta_j}{4}\right) \frac{1}{\sqrt{1 - \exp(-\theta_j/4)}} \quad (j = 1, \ldots, d).$$

Figure 2(a) shows that this log-beta prior places heavy weight on small θ_j values. The preference for small values of θ_j reflects a prior belief that the underlying function is predictable, since a small θ_j value means strong correlation between neighboring points in the design space.

When implementing the algorithm, we actually work with $\lambda_j = \ln(\rho_j/(1 - \rho_j))$. After the logistic transformation, λ_j is unconstrained, facilitating proposals in MCMC. The implied

prior density on λ_j is

$$\pi(\lambda_j) = \frac{1}{2}\left(\frac{\exp(-\lambda_j)}{1+\exp(-\lambda_j)}\right)^{-\frac{1}{2}}\left(\frac{\exp(\lambda_j)}{(1+\exp(\lambda_j))^2}\right) \quad (j = 1, \ldots, d). \tag{4.1}$$

The priors above are closely related to the SqExp-Full (H) method of Higdon et al. [9]. In particular, Chen [4] found through simulation that a log-beta prior on θ_j with parameters favoring small θ_j is highly preferred. The improper normal prior on μ and inverse-gamma prior on σ^2 are well-known conjugate priors, which makes it easy to integrate them out when deriving the marginal posterior of the θ_j (described in section 4.5) or, equivalently, the λ_j.

4.2. Power-exponential-hybrid. We explore two approaches to deal with the smoothness parameters, $\boldsymbol{\alpha}$, of PowExp. The first takes the empirical Bayes paradigm for these parameters only, by plugging in the MLEs of the α_j. As the remaining parameters are treated in a fully Bayesian way (see section 4.1), we call this method PowExp-Hybrid.

Spiller et al. [23] recently proposed a Bayesian method in which they fix $\alpha_j = 1.9$ and estimate $\ln(\theta_j)$ by its posterior mean using a reference prior [16]. This approach differs from the PowExp-Hybrid method we propose here in two aspects: (1) we estimate the smoothness parameters instead of fixing them, and (2) we use a different prior distribution on the correlation parameter, θ_j.

4.3. Power-exponential-full. The second new PowExp implementation is fully Bayesian and hence is abbreviated PowExp-Full. In addition to the priors specified in section 4.1, it needs priors on the smoothness parameters α_j.

We actually work on the logistic scale γ_j, where

$$\alpha_j = 1 + \frac{1}{1+\exp(-\gamma_j)}.$$

Like the λ_j for the sensitivity parameters, γ_j is unrestricted, always giving $\alpha_j \in (1, 2)$. In practice, we take an independent uniform prior, $U(-20, 20)$, on γ_j. After a transformation of variables, the implied prior distribution for α_j is

$$\pi(\alpha_j) = \frac{1}{40}\frac{1}{(\alpha_j - 1)(2 - \alpha_j)} \quad (j = 1, \ldots, d), \tag{4.2}$$

as plotted in Figure 2(b). It assigns heavy weight to α_j values less than about 1.2 and greater than about 1.8.

The specification of this prior on α_j (via γ_j) was the result of much trial and error. We tried independent uniform priors on the α_j, which is straightforward, but results were not satisfactory for the borehole example in section 5.3. We also considered independent uniform $U(0, \pi)$ priors on γ_j, where $\alpha_j = 1 + \sin(\gamma_j)$. This implied prior for α_j has heavy weight on values close to the boundary 2, the qualitative feature we were seeking, but it did not work well for the Nilson–Kuusk code. Several other priors were considered, but only the prior in (4.2) from a logistic transformation worked for all the examples in section 5. It is used for all results reported for PowExp-Full.

4.4. Matérn-Hybrid. A Matérn-Hybrid implementation parallels PowExp-Hybrid: the Matérn smoothness parameters δ_j in (2.4) are estimated by their MLEs, with each dimension allowed its own estimate from the four levels of smoothness.

4.5. Marginal posterior distribution of the correlation parameters. For any method, denote by $\boldsymbol{\psi}_B$ the correlation parameters treated in a fully Bayesian way and sampled by MCMC. Any remaining correlation parameters are estimated by empirical Bayes. Thus, for the methods to be compared,

$$\boldsymbol{\psi}_B = \begin{cases} \boldsymbol{\lambda} & \text{(SqExp-Full, with all } \alpha_j \text{ fixed at 2)}, \\ \boldsymbol{\lambda}, \boldsymbol{\gamma} & \text{(PowExp-Full)}, \\ \boldsymbol{\lambda} & \text{(PowExp-Hybrid, with all } \alpha_j \text{ replaced by MLEs)}, \\ \boldsymbol{\lambda} & \text{(Matérn-Hybrid, with all } \delta_j \text{ replaced by MLEs)}. \end{cases}$$

All methods have the priors for μ, σ^2, and the λ_j prescribed in section 4.1. Recall that λ_j and γ_j are logistic transformations of the sensitivity parameter θ_j and the power α_j, respectively.

We want the marginal posterior distribution of $\boldsymbol{\psi}_B$ after integrating out μ and σ^2. With the priors $\pi(\mu) \propto 1$ and $\pi(\sigma^2) = IG(\phi_1, \phi_2)$, as in section 4.1, we have

$$\pi(\boldsymbol{\psi}_B|\mathbf{y}) \propto \frac{\pi(\boldsymbol{\psi}_B)}{|\mathbf{R}|^{1/2}\,(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{1/2}\,(\phi_2 + \frac{n-1}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B})^{(\phi_1+\frac{n-1}{2})}}, \tag{4.3}$$

where $|\mathbf{R}|$ denotes the determinant of $\mathbf{R}$, and $\hat{\sigma}^2_{\boldsymbol{\psi}_B}$ is as in (2.14) but parameterized in terms of $\boldsymbol{\psi}_B$. A derivation may be found in the appendix. The prior $\pi(\boldsymbol{\psi}_B)$ is a product of the independent priors (4.1) for the λ_j and, in the case of PowExp-Full, the independent uniform priors in section 4.3 for the γ_j. For PowExp-Hybrid and Matérn-Hybrid, both sides of (4.3) are conditional on the MLEs of the α_j or δ_j, respectively.

The parameters $\boldsymbol{\psi}_B$ are sampled from the marginal posterior distribution by the algorithm in section 4.6. According to [21], the predictive distribution of $y(\mathbf{x}^*|\mathbf{y}, \boldsymbol{\psi}_B)$ is as in (2.13), but again the degrees of freedom are $2\phi_1 + n - 1$, and the predictive variance changes to

$$\hat{v}_{\boldsymbol{\psi}_B}(\mathbf{x}) = \left(\frac{(n-1)\hat{\sigma}^2_{\boldsymbol{\psi}_B} + 2\phi_2}{2\phi_1 + n - 1}\right)\left(1 - \mathbf{r}^T(\mathbf{x}^*)\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*) + \frac{\left(1 - \mathbf{1}^T\mathbf{R}^{-1}\mathbf{r}(\mathbf{x}^*)\right)^2}{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}\right). \tag{4.4}$$

4.6. Metropolis–Hastings algorithm. Taking the PowExp-Full method for illustration, we briefly describe how Metropolis–Hastings (M-H) is used to sample the posterior distribution of $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$.

At iteration i, the algorithm cycles through the input dimensions, making a proposal for each parameter in turn, i.e., λ_1 followed by γ_1, then λ_2, γ_2, and so on. The details are in Algorithm 1.

For an efficient algorithm, the step length for a proposal is important, and we use the adaptive MCMC algorithm [18]. Let ψ denote the value at iteration i of a parameter under consideration for transition. The candidate value is

$$Q(\psi) = 0.95N(\psi, 2.38^2 s^2) + 0.05N(\psi, 0.1^2), \tag{4.5}$$

Algorithm 1. Implementation of the M-H algorithm.

At iteration i, to update for dimension j, denote the current values of $\boldsymbol{\lambda}$ and $\boldsymbol{\gamma}$ by $\boldsymbol{\lambda}_j^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_{j-1}^{(i)}, \lambda_j^{(i-1)}, \lambda_{j+1}^{(i-1)}, \ldots, \lambda_d^{(i-1)})$ and $\boldsymbol{\gamma}_j^{(i)} = (\gamma_1^{(i)}, \ldots, \gamma_{j-1}^{(i)}, \gamma_j^{(i-1)}, \gamma_{j+1}^{(i-1)}, \ldots, \gamma_d^{(i-1)})$.

1. Use the adaptive proposal $Q(\lambda_j^{(i-1)})$ in (4.5) to generate a candidate λ_j^*.
2. Randomly sample u from $U(0,1)$.
3. Let $\boldsymbol{\lambda}_j^* = (\lambda_1^{(i)}, \ldots, \lambda_{j-1}^{(i)}, \lambda_j^*, \lambda_{j+1}^{(i-1)}, \ldots, \lambda_d^{(i-1)})$. If $\ln(u) < \ln(\pi(\boldsymbol{\lambda}_j^*, \boldsymbol{\gamma}_j^{(i)}|\mathbf{y})) - \ln(\pi(\boldsymbol{\lambda}_j^{(i)}, \boldsymbol{\gamma}_j^{(i)}|\mathbf{y}))$, set $\lambda_j^{(i)} = \lambda_j^*$; otherwise $\lambda_j^{(i)} = \lambda_j^{(i-1)}$. The posterior distribution used here is $\pi(\boldsymbol{\psi}_B|\mathbf{y})$ in (4.3) with $\boldsymbol{\psi}_B = (\boldsymbol{\lambda}, \boldsymbol{\gamma})$. We now have the updated vector $\boldsymbol{\lambda}_{j+1}^{(i)} = (\lambda_1^{(i)}, \ldots, \lambda_{j-1}^{(i)}, \lambda_j^{(i)}, \lambda_{j+1}^{(i-1)}, \ldots \lambda_d^{(i-1)})$.
4. Use the adaptive proposal $Q(\gamma_j^{(i-1)})$ in (4.5) to generate a candidate γ_j^*.
5. Randomly sample another u from $U(0,1)$.
6. Let $\boldsymbol{\gamma}_j^* = (\gamma_1^{(i)}, \ldots, \gamma_{j-1}^{(i)}, \gamma_j^*, \gamma_{j+1}^{(i-1)}, \ldots, \gamma_d^{(i-1)})$. If $\ln(u) < \ln(\pi(\boldsymbol{\lambda}_{j+1}^{(i)}, \boldsymbol{\gamma}_j^*|\mathbf{y})) - \ln(\pi(\boldsymbol{\lambda}_{j+1}^{(i)}, \boldsymbol{\gamma}_j^{(i)}|\mathbf{y}))$, set $\gamma_j^{(i)} = \gamma_j^*$; otherwise $\gamma_j^{(i)} = \gamma_j^{(i-1)}$. We now have the updated vector $\boldsymbol{\gamma}_{j+1}^{(i)} = (\gamma_1^{(i)}, \ldots, \gamma_{j-1}^{(i)}, \gamma_j^{(i)}, \gamma_{j+1}^{(i-1)}, \ldots \lambda_d^{(i-1)})$.

where s^2 is the sample variance of the sampled values of ψ up to iteration i, which adapts as the algorithm iterates. Efficiency is achieved by this mixture proposal for all the examples we consider in section 5 in the sense that almost all acceptance rates are between 0.15 and 0.50 [19]. More details may be found in the supplemental material (M100877_01.pdf [local/web 176KB]).

For any method, at iteration i after all dimensions have been updated, $\boldsymbol{\psi}_B^{(i)}$ has been sampled from the posterior. The sample leads to a conditional predictive mean, $\hat{m}_{\boldsymbol{\psi}_B^{(i)}}(\mathbf{x})$, and conditional predictive variance, $\hat{v}_{\boldsymbol{\psi}_B^{(i)}}(\mathbf{x})$, computed according to (2.10) and (4.4), respectively. The overall predictor is defined as

$$\hat{m}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \hat{m}_{\boldsymbol{\psi}_B^{(i)}}(\mathbf{x}),$$

where M is the number of samples. The predictive variance is obtained through the law of total variance. That is,

$$\hat{v}(\mathbf{x}) = \frac{1}{M}\sum_{i=1}^{M} \hat{v}_{\boldsymbol{\psi}_B^{(i)}}(\mathbf{x}) + \frac{1}{M-1}\sum_{i=1}^{M} \left(\hat{m}_{\boldsymbol{\psi}_B^{(i)}}(\mathbf{x}) - \hat{m}(\mathbf{x})\right)^2.$$

The M-H algorithm is simpler for SqExp-Full, PowExp-Hybrid, and Matérn-Hybrid, because it needs to iterate over the $\boldsymbol{\lambda}$ sensitivity parameters only. The hybrid methods have to find MLEs of all parameters, however, before running the MCMC. The tradeoff in computational requirements is illustrated in section 5.1.

5. Applications and simulation study. We now evaluate the performances of six methods: SqExp-Emp, SqExp-Full, PowExp-Emp, PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid.

(Their names are further abbreviated as Sq-Emp, Sq-Full, Pow-Emp, Pow-Full, Pow-Hyb and Mat-Hyb in subsequent figures for legibility.) As in section 3, normalized RMSE and normalized maximum absolute error are used to assess the prediction accuracy, and the ACP is used to assess the uncertainty quantification. The nominal coverage probability is set to 0.95.

5.1. Nilson–Kuusk code. We revisit the Nilson–Kuusk code with repeat experiments generated as in section 3.

The boxplots of normalized RMSE in the left panel of Figure 3 show that the PowExp and Matérn correlation functions lead to noticeably better prediction accuracy than SqExp. Empirical Bayes versus full Bayes versus hybrid Bayes makes little difference here. The results for normalized maximum absolute error in the supplementary material (M100877_01.pdf [local/web 176KB]) follow a similar pattern.

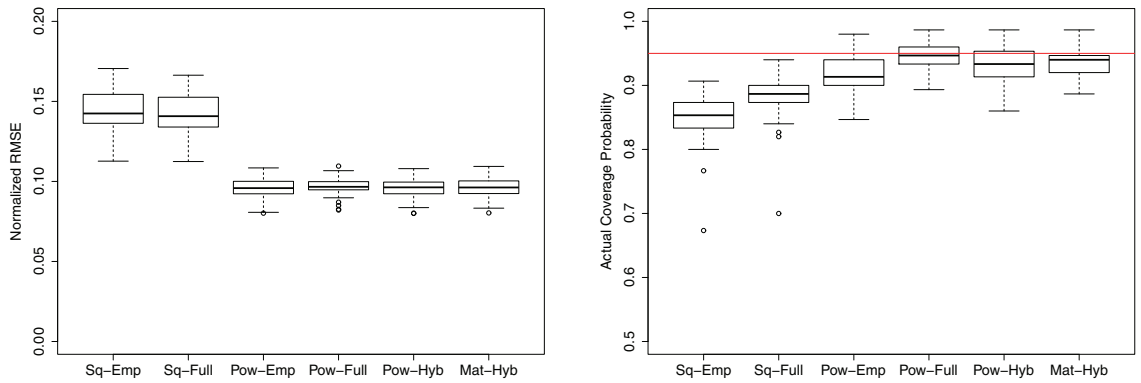

Figure 3. *Normalized RMSE (left panel) and ACP (right panel) for the Nilson–Kuusk code from six methods: SqExp-Emp, SqExp-Full, PowExp-Emp, PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid, abbreviated as Sq-Emp, Sq-Full, Pow-Emp, Pow-Full, Pow-Hyb, and Mat-Hyb, respectively. The boxplots show the results from* 25 *random training-test data splits. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

Sample means of normalized RMSE over the repeat experiments are reported in Table 1 for the four methods giving clearly superior prediction accuracy here: those using either a PowExp or Matérn correlation function. The small differences in sample means and small standard errors of the mean differences also reported in Table 1 indicate that these four methods have similar prediction accuracies on average.

The ACP results in the right panel of Figure 3 also demonstrate the advantage of the PowExp and Matérn correlation functions versus SqExp, with the three new Bayesian methods coming closest to the nominal coverage probability of 0.95. Equipping Bayesian methods with PowExp or Matérn structure is advantageous here in terms of uncertainty quantification relative to SqExp, even if the latter has a fully Bayesian treatment. These findings are confirmed by Table 2, which compares the four methods using PowExp or Matérn structure in terms of $|\text{ACP} - 0.95|$, the absolute deviation of ACP from the nominal coverage probability. PowExp-Emp stands out with the largest mean absolute deviation from nominal coverage, and the observed mean difference of 0.0203 relative to PowExp-Full is large relative to the standard

Table 1

Sample means of the 25 normalized RMSEs from repeat experiments with the Nilson–Kuusk code for four methods. The sample means are in the diagonal entries of the table. An off-diagonal entry shows the difference in sample means between the row method and the column method, with the standard error of the mean difference in parentheses.

	PowExp-Emp	PowExp-Full	PowExp-Hybrid	Matérn-Hybrid
PowExp-Emp	0.0955	−0.0010 (0.0005)	−0.0002 (0.0002)	−0.0003 (0.0008)
PowExp-Full	0.0010 (0.0005)	0.0965	0.0008 (0.0005)	0.0007 (0.0007)
PowExp-Hybrid	0.0002 (0.0002)	−0.0008 (0.0005)	0.0957	−0.0001 (0.0008)
Matérn-Hybrid	0.0003 (0.0008)	−0.0007 (0.0007)	0.0001 (0.0008)	0.0958

Table 2

Sample means of $|ACP - 0.95|$ from 25 repeat experiments with the Nilson–Kuusk code. The sample means of four methods are in the diagonal entries of the table. An off-diagonal entry shows the difference in sample means between the row method and the column method, with the standard error of the mean difference in parentheses.

	PowExp-Emp	PowExp-Full	PowExp-Hybrid	Matérn-Hybrid
PowExp-Emp	0.0388	0.0203 (0.0045)	0.0083 (0.0024)	0.0184 (0.0039)
PowExp-Full	−0.0203 (0.0045)	0.0185	−0.0120 (0.0032)	−0.0019 (0.0027)
PowExp-Hybrid	−0.0083 (0.0024)	0.0120 (0.0032)	0.0305	0.0103 (0.0031)
Matern-Hybrid	−0.0184 (0.0039)	0.0019 (0.0027)	−0.0103 (0.0031)	0.0204

Table 3

Number of likelihood (posterior) evaluations in model fitting, averaged over 25 repeat experiments with the Nilson–Kuusk code.

	Likelihood evaluations		
Method	MLE	MCMC	Total
SqExp-Emp	20,782	0	20,782
SqExp-Full	0	50,000	50,000
PowExp-Emp	39,021	0	39,021
PowExp-Full	0	100,000	100,000
PowExp-Hybrid	39,021	50,000	89,021
Matérn-Hybrid	31,082	50,000	81,082

error of the mean difference. The differences in averages and their standard errors are all fairly small for PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid, indicating that these methods perform about the same, as was seen in Figure 3.

To summarize results for the Nilson–Kuusk code, use of a flexible correlation structure—PowExp or Matérn—leads to nontrivial improvements in prediction accuracy relative to SqExp. Uncertainty quantification is also much more reliable with PowExp or Matérn, with full or hybrid Bayesian implementations adding to the advantage.

Table 3 compares the computational requirements of the five methods for the Nilson–Kuusk application. The computational cost of model fitting dominates prediction, and repeated evaluation of the likelihood is the expensive calculation in both maximum likelihood and MCMC. Hence, Table 3 compares the methods via the number of likelihood evaluations averaged over 25 model fits from the 25 repeat experiments. For a given set of training data

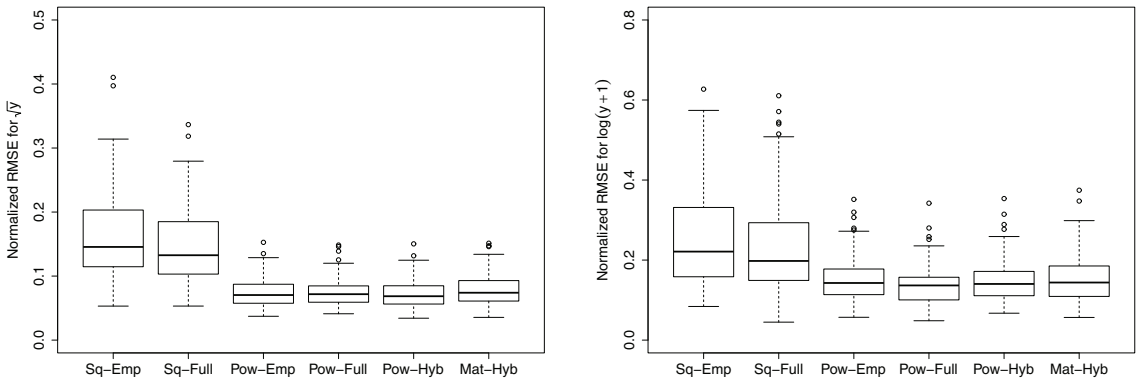

Figure 4. *Normalized RMSE for the volcano code from six methods: SqExp-Emp, SqExp-Full, PowExp-Emp, PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid. Each method has* 100 *RMSE values from* 100 *random training-test data splits. Two transformations of* y *are considered:* $\sqrt{y}$ *(left panel) and* $\log(y+1)$ *(right panel).*

in a repeat experiment, the number of likelihood evaluations for maximum likelihood is accumulated over 25 multistarts of the maximization from different random starting points in the space of correlation parameters. Not surprisingly, the PowExp and Matérn models have more evaluations than SqExp due to the increase in the number of parameters. An MCMC fit involves 10,000 cycles through the correlation parameters, a number providing satisfactory convergence diagnostics (shown in the supplemental material (M100877_01.pdf [local/web 176KB])). Hence, SqExp-Full, PowExp-Hybrid, and Matérn-Hybrid have 50,000 likelihood evaluations during MCMC, while PowExp-Full, with twice as many parameters varying in MCMC, requires 100,000 evaluations. Overall, PowExp-Full requires the most computation here, with PowExp-Hybrid and Matérn-Hybrid providing some computational savings.

5.2. Volcano code. A computer model of pyroclastic flow from a volcano eruption was studied by Bayarri et al. [3] and Chen et al. [5]. The two inputs are initial volume, x_1, and direction, x_2, of the eruption, and the output, y, is the maximum height of the flow at a specific location.

The nonnegative output suggests transformation may be helpful. Bayarri et al. [3] transformed y to $\log(y+1)$. Chen et al. [5] studied both the $\log(y+1)$ and $\sqrt{y}$ transformations, as we do here. A 32-run data set is available. We sample 25 runs out of the 32 runs as a training set and leave the remaining 7 runs as a hold-out set for testing. Sampling is repeated 100 times.

The normalized RMSE results are shown in Figure 4. For both output transformations, methods with a flexible correlation structure—PowExp-Emp, PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid—have higher accuracy than SqExp-Emp and SqExp-Full. Results for normalized maximum absolute error presented in the supplemental material (M100877_01.pdf [local/web 176KB]) show a similar pattern.

Table 4 reports the ACP of each method, obtained as the number of 95% confidence/credible intervals containing the true output value divided by the total number of predictions, 700.

Table 4

ACP for the volcano code, computed as the number of 95% *confidence/credible intervals containing the true output of each method divided by the total number of predictions,* 700. *The nominal coverage is* 0.95.

	Average coverage probability	
Method	$\sqrt{y}$	$\log(y+1)$
SqExp-Emp	0.63	0.71
SqExp-Full	0.78	0.82
PowExp-Emp	0.86	0.85
PowExp-Full	0.92	0.91
PowExp-Hybrid	0.90	0.87
Matérn-Hybrid	0.90	0.88

The ACPs of PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid are closer to 0.95, though they slightly undercover. The SqExp-Emp approach substantially undercovers, with SqExp-Full faring little better. Thus, again the correlation function is more important than the estimation paradigm here, although Bayesian methods, with a full or hybrid implementation, make further improvements.

5.3. Borehole function. The borehole function [14] is an 8-dimensional computer model that has served as a test-bed in many contexts (e.g., [4, 5]). Two base designs, with 80 or 200 runs, are explored; both are approximate maximin LHDs [13] with the maximin criterion adapted to have good space-filling properties in all 2-dimensional projections [25]. The columns of each base design are permuted at random to generate 25 different but equivalent designs [5] and hence 25 repeat experiments. The hold-out set is 10,000 runs generated from a random LHD.

For $n = 80$ the left panel of Figure 5 shows that the prediction accuracies of the six competing methods are similar. The ACP performances of the methods are seen to differ, however, in the right panel of Figure 5. SqExp-Emp and PowExp-Emp show undercoverage, while the four Bayesian implementations have near-nominal coverage probabilities, with PowExp-Full and Matérn-Hybrid the best.

When the sample size is increased to 200, however, the prediction accuracies depicted in the left panel of Figure 6 show methods using PowExp correlation structure are noticeably more accurate than those employing SqExp and slightly better than those using Matérn. A similar pattern is seen for normalized maximum absolute error in the supplemental material (M100877_01.pdf [local/web 176KB]). The ACP results depicted in the right panel of Figure 6 demonstrate that the SqExp correlation structure leads to undercoverage, even with a Bayesian implementation. A possible explanation is that the impact of any inadequacy of the GP statistical model with SqExp in modeling the borehole function increases with sample size. PowExp-Emp also undercovers, though less, while the three Bayesian methods using PowExp or Matérn again have ACP closest to the nominal coverage probability.

5.4. PTW model. Preston, Tonks, and Wallace [17] developed the PTW model to describe the plastic deformation of metals. For our purposes the model contains $d = 11$ input parameters. A base design has its columns permuted at random to generate 25 different but equivalent designs. Two types of base design are considered to investigate the role of design. The hold-out set is 10,000 runs generated from a random LHD.

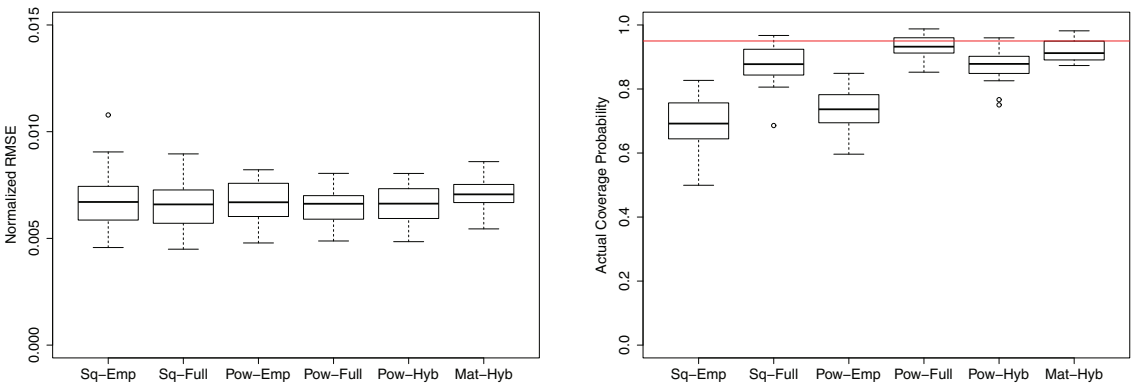

Figure 5. *Normalized RMSE (left panel) and ACP (right panel) for the borehole function and an* 80*-run maximin LHD base design. The boxplots show the results from* 25 *repeat experiments permuting the columns of the base design. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

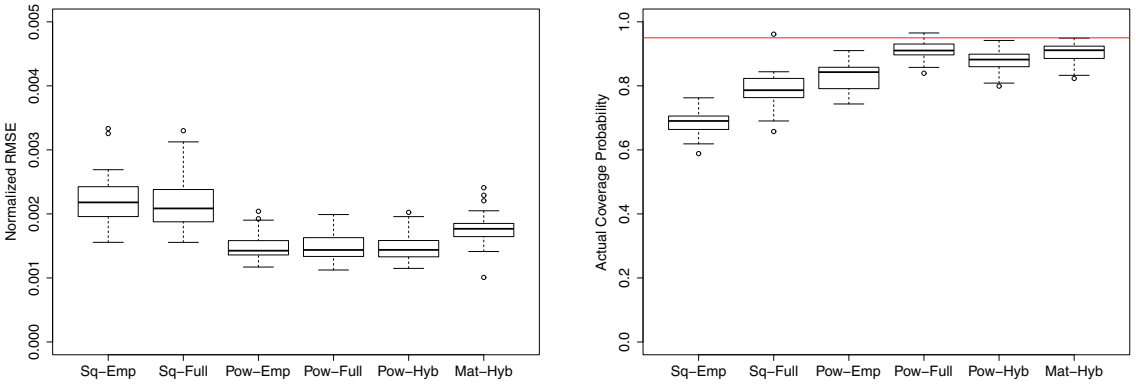

Figure 6. *Normalized RMSE (left panel) and ACP (right panel) for the borehole function and a* 200*-run maximin LHD. The boxplots show the results from* 25 *random training-test data splits. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

The first design is a 110-run maximin LHD, for which results are presented in Figure 7. The normalized RMSE values in the left panel show the four methods employing the Pow-Exp or Matérn correlation functions slightly outperforming both methods with SqExp. The normalized maximum absolute error results in the supplemental material (M100877_01.pdf [local/web 176KB]) show a similar pattern. The right panel of Figure 7 indicates, however, that the four fully Bayesian or hybrid methods have clearly better uncertainty quantification properties than the two empirical Bayes methods, which undercover.

To investigate the effect of design, we repeated the study with a 110-run generalized LHD (GLHD) as the base design. GLHDs were proposed by Dette and Pepelyshev [6] as a form of optimal design claimed to have better prediction accuracy than uniform LHDs. For the PTW code, we took the base maximin LHD design, transformed it to have margins optimal for a

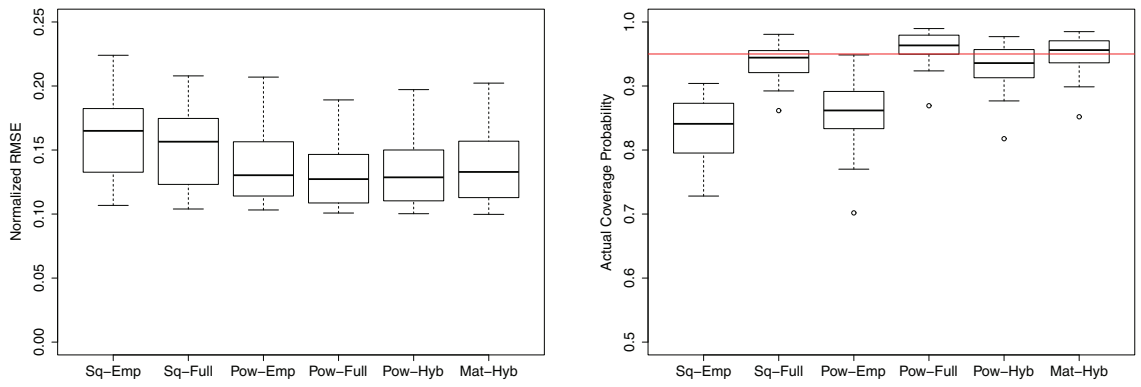

Figure 7. *Normalized RMSE (left panel) and ACP (right panel) for the PTW model and a* 110*-run maximin LHD. The boxplots show the results from* 25 *random repeat experiments permuting the columns of the base design. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

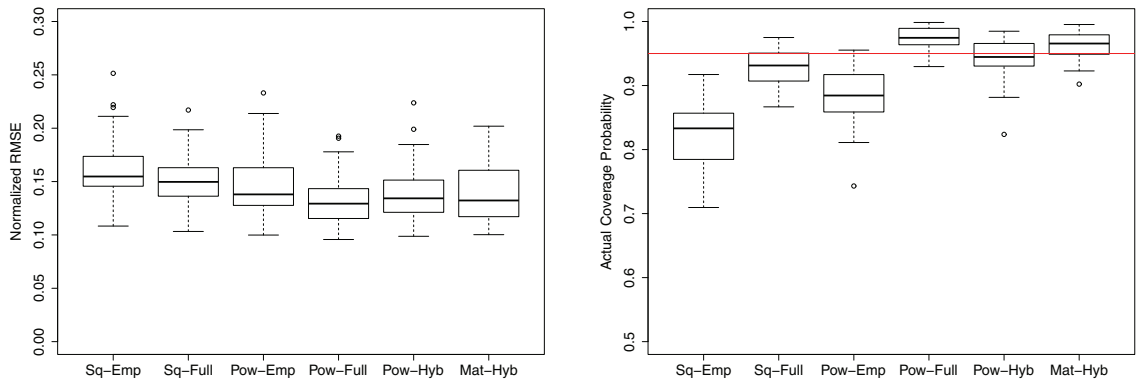

Figure 8. *Normalized RMSE (left panel) and ACP (right panel) for the PTW model and a* 110*-run GLHD. The boxplots show the results from* 25 *random repeat experiments permuting the columns of the base design. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

GLHD, and again randomly permuted the columns 25 times to obtain repeat experiments.

The normalized RMSE results in the left panel of Figure 8 again show prediction accuracy is modestly worse for methods using SqExp correlation relative to PowExp or Matérn. Comparison with the left panel of Figure 7 also indicates no improvement in accuracy from the GLHD design for any method. The normalized maximum absolute error results in the supplemental material (M100877_01.pdf [local/web 176KB]) also show slightly worse accuracy for the methods using SqExp correlation, but some improvement in this metric is evident from use of a GLHD across all methods relative to a maximin LHD. The right panel of Figure 8 demonstrates again the advantage of the fully Bayesian or hybrid methods in terms of ACP relative to the two empirical Bayes methods, following the pattern in the right panel of Figure 7.

Table 5
Values of the θ_j for simulation.

j	1	2	3	4	5	6	7	8	9	10
θ_j	0.271	0.217	0.169	0.127	0.091	0.061	0.037	0.019	0.007	0.001

The accuracy and coverage results for PTW can be summarized as follows. The PowExp and Matérn correlation functions provide slightly more accurate predictions than SqExp, with fully Bayesian estimation versus empirical Bayes less important. The fully Bayesian or hybrid Bayes methods provide much more reliable prediction intervals for inference, however. Comparing the two designs, the GLHD gives better accuracy as measured by maximum absolute error across all estimation methods, but the relative merits of the six methods follow the same patterns.

5.5. Simulations from GPs. The applications so far assess the new Bayesian methods on computer codes, meaning that uncertainty quantification includes model misspecification in representing a code by a statistical GP model. We now consider ideal situations, where functions are realizations of GPs.

The simulation settings are as follows. There are $d = 10$ inputs (results for $d = 5$, not shown, are similar). The true value of μ is 0, and $\sigma^2 = 1$; there is little loss of generality here as the priors proposed in section 4.1 allow much latitude for location and scale. Three true correlation functions are considered: SqExp, PowExp with α_j set to 1.8, and Matérn with all $\delta_j = 1$. Thus, the realized functions have various degrees of smoothness. The true values of the θ_j are given in Table 5; they are a canonical configuration generated by the method of Loeppky, Sacks, and Welch [12] with $\tau = 1$ and $b = 3$. A training sample size of $n = 100$ gives reasonable prediction accuracy for realizations from all three GP families.

Given one of the GP settings for generating data, there are 25 repeat experiments from different random LHD training designs. A set of 100 training points is combined with a 2000-run random LHD for the hold-out set (the same hold-out set is used for all 25 repeats), and a GP realization is sampled from the multivariate normal at the $100 + 2000$ points for the training and hold-out output data. All six of the methods SqExp-Emp, SqExp-Full, PowExp-Emp, PowExp-Full, PowExp-Hybrid, and Matérn-Hybrid are fit to the data, with all relevant parameters estimated. For instance, the θ_j and α_j are estimated for PowExp by empirical Bayes, full Bayes, or hybrid Bayes.

Figure 9 shows the performances of the six methods when data are generated by a GP with SqExp correlation. As SqExp is a special case of PowExp and Matérn, all trained models assume the correct GP family. In the left panel of Figure 9 it is seen that distributions of RMSE over the repeat simulations are practically identical; i.e., there is no overfitting penalty from using PowExp or Matérn with any of the inference methods considered. The ACP results in the right panel of Figure 9 indicate that the empirical Bayes methods SqExp-Emp and PowExp-Emp underestimate prediction uncertainty, whereas the full Bayes or hybrid Bayes methods obtain near-nominal ACP.

With output realized from a GP with PowExp correlation and all $\alpha_j = 1.8$, the RMSE results in the left panel of Figure 10 show that assuming a PowExp or Matérn correlation function, with any of the inference methods, is now clearly much more accurate for prediction

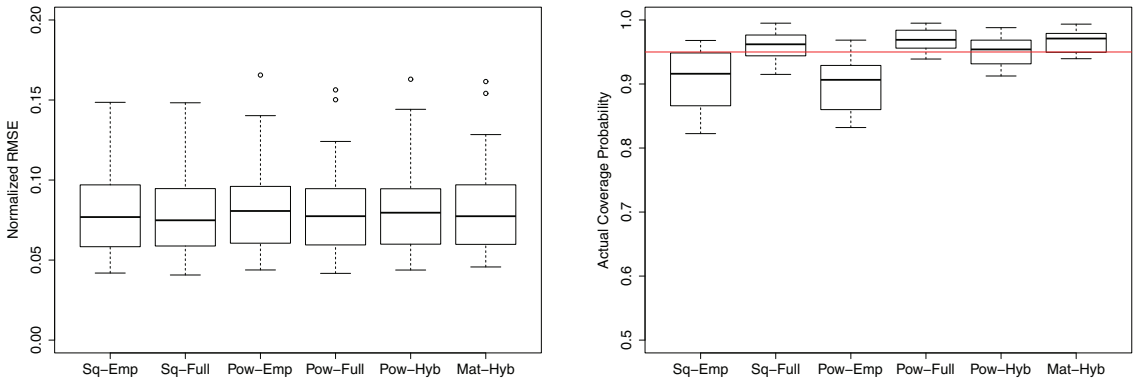

Figure 9. *Normalized RMSE (left panel) and ACP (right panel) with* $d = 10$ *inputs and output simulated from a GP with SqExp correlation. The boxplots show the results from* 25 *random realizations of the GP. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

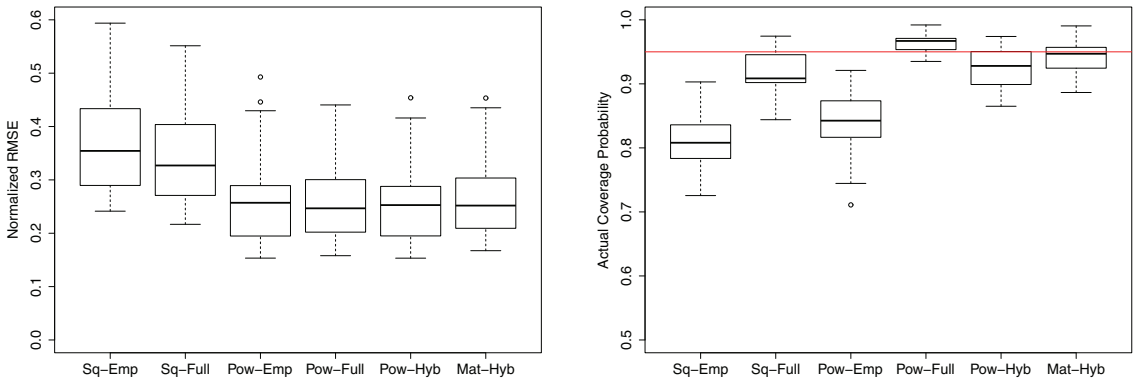

Figure 10. *Normalized RMSE (left panel) and ACP (right panel) with* $d = 10$ *inputs and output simulated from a GP with PowExp correlation and all* $\alpha_j = 1.8$. *The boxplots show the results from* 25 *random realizations of the GP. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

than using the wrong correlation function, SqExp. The ACP results in the right panel of Figure 10 are again driven by the inference method: there is substantial undercoverage for the empirical Bayes methods. SqExp-Full and PowExp-Hybrid slightly undercover, while PowExp-Full and Matérn-Hybrid have near-nominal ACP.

When the output is realized from a GP with Matérn correlation and all $\nu_j = 1$, the left panel of Figure 11 demonstrates that the PowExp and Matérn correlation functions again lead to much smaller RMSEs than does SqExp, regardless of the parameter-estimation method. The right panel of Figure 11 shows severe undercoverage for SqExp-Emp, some undercoverage for SqExp-Full, and modest undercoverage for PowExp-Emp. PowExp-Full leads to overcoverage here, while PowExp-Hybrid and Matérn-Hybrid have near-nominal ACP.

Overall, the simulations suggest that the PowExp and Matérn correlation functions are

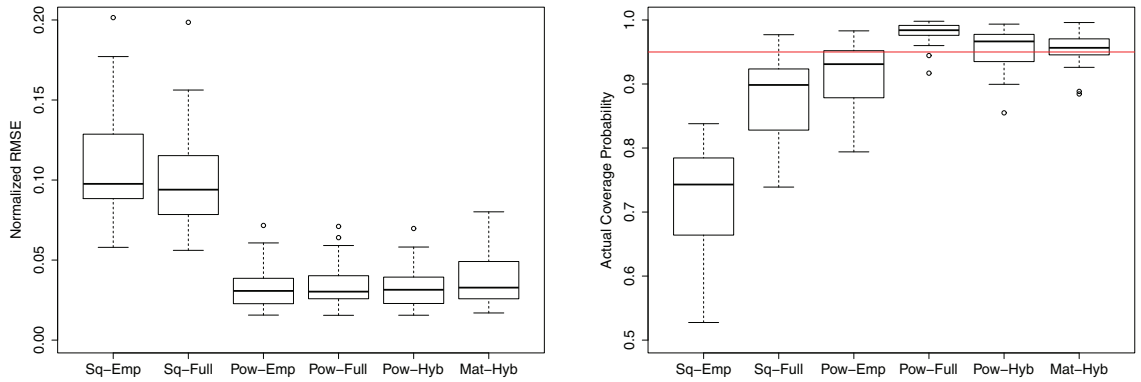

Figure 11. *Normalized RMSE (left panel) and ACP (right panel) with* $d = 10$ *inputs and output simulated from a GP with Matérn correlation and all* $\delta_j = 1$. *The boxplots show the results from* 25 *random realizations of the GP. The horizontal line in the right panel is the nominal coverage probability,* 0.95.

advantageous in terms of prediction accuracy relative to SqExp. When data were generated using SqExp, there was little overfitting penalty in estimating the extra parameters of the PowExp and Matérn families. On the other hand, the more flexible PowExp and Matérn families sometimes led to much more accurate predictions. The normalized maximum absolute error results reported in the supplemental material (M100877_01.pdf [local/web 176KB]) follow similar patterns.

The simulations also point to the empirical Bayes methods SqExp-Emp and PowExp-Emp underestimating uncertainty, sometimes substantially. SqExp-Full also led to smaller-than-nominal ACP values, except when data were generated using an SqExp correlation function. A combination of fully Bayesian or hybrid Bayesian inference with the PowExp or Matérn families provided more reliable inference, however. Such combinations showed modest over- or undercoverage.

6. Conclusions and discussion. Bayesian implementations with SqExp correlation structure were extended to allow PowExp or Matérn structure. To estimate the additional smoothness parameters α_j in PowExp, hybrid and fully Bayesian approaches were proposed and applied. The hybrid method uses MLEs of the α_j, with all other parameters in a GP model handled via Bayes' rule. The fully Bayesian method employs an uninformative uniform prior for a logistic transform of α_j. Similarly, a hybrid approach for the Matérn correlation function uses MLEs for the smoothness parameters δ_j.

Previous work [5] noted that a fully Bayesian treatment tends to have better coverage probability than the empirical Bayes method for applications such as the borehole function, where SqExp is adequate. The previous work also noted that PowExp can give better prediction accuracy for applications such as the Nilson–Kuusk code, where SqExp is less adequate. The applications and simulations considered show that advantages in terms of prediction accuracy and uncertainty quantification tend to go hand in hand when a Bayesian method with flexible correlation function is employed. Moreover, there appears to be little overfitting penalty from employing PowExp or Matérn when the SqExp special case is adequate.

There are some important details of implementation. First, the priors on the sensitivity parameters θ_j based on previous work [9] assume the inputs x_j are scaled to $[0, 1]$. These priors assign heavier weight to small values of the θ_j, reflecting a prior view that the function can be usefully predicted. Second, such priors are particularly appropriate for a constant-mean model. If a regression model with linear trends in the inputs is used, priors giving more weight to moderate values of θ_j [10] may be more appropriate. Finally, we found that MCMC sampling of all correlation parameters, for sensitivity and smoothness, was more efficient with parameters transformed to a logistic scale.

`R` code to implement these methods is available at https://www.stat.ubc.ca/~will/GPflex/GPflex.html.

Appendix. Marginal posterior distribution of the correlation parameters. Here we derive the marginal posterior distribution of $\boldsymbol{\psi}_B$, with the correlation parameters treated in a fully Bayesian way and hence sampled by MCMC. Any remaining correlation parameters are estimated by empirical Bayes, and the expressions below are conditional on their plugged-in MLEs. We also take priors $\pi(\mu) \propto 1$ and $\pi(\sigma^2) = IG(\phi_1, \phi_2)$.

We need to integrate out μ and σ^2 from the joint posterior of μ, σ^2, and $\boldsymbol{\psi}_B$:

$$\begin{aligned}
\pi(\boldsymbol{\psi}_B|\mathbf{y}) &\propto \int\int \pi(\mu)\pi(\sigma^2)\pi(\boldsymbol{\psi}_B)L(\mathbf{y}|\mu,\sigma^2,\boldsymbol{\psi}_B)\,d\mu\,d\sigma^2 \\
&\propto \int\int \pi(\boldsymbol{\psi}_B)\frac{\phi_2^{\phi_1}}{\Gamma(\phi_1)}(\sigma^2)^{-\phi_1-1}\exp\left(-\frac{\phi_2}{\sigma^2}\right) \\
&\qquad \times \frac{1}{(\sigma^2)^{n/2}\,|\mathbf{R}|^{1/2}}\exp\left(-\frac{(\mathbf{y}-\mathbf{1}\mu)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\mu)}{2\sigma^2}\right)d\mu\,d\sigma^2 \\
&= \frac{\pi(\boldsymbol{\psi}_B)}{|\mathbf{R}|^{1/2}}\int \frac{\phi_2^{\phi_1}}{\Gamma(\phi_1)}(\sigma^2)^{-\phi_1-1}\exp\left(-\frac{\phi_2}{\sigma^2}\right) \\
&\qquad \times \frac{1}{(\sigma^2)^{n/2}}\int \exp\left(-\frac{(\mathbf{y}-\mathbf{1}\mu)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\mu)}{2\sigma^2}\right)d\mu\,d\sigma^2. \qquad \text{(A.1)}
\end{aligned}$$

First, we evaluate the integral with respect to μ. Rewrite the quadratic form in the integrand as

$$\begin{aligned}
(\mathbf{y}-\mathbf{1}\mu)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\mu) &= (\mathbf{y}-\mathbf{1}\hat{\mu}-\mathbf{1}(\mu-\hat{\mu}))^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\hat{\mu}-\mathbf{1}(\mu-\hat{\mu})) \\
&= (\mathbf{y}-\mathbf{1}\hat{\mu})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\hat{\mu}) + (\mu-\hat{\mu})\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}(\mu-\hat{\mu}). \qquad \text{(A.2)}
\end{aligned}$$

The cross-product term disappears because

$$(\mathbf{y}-\mathbf{1}\hat{\mu})^T\mathbf{R}^{-1}\mathbf{1} = \mathbf{y}^T\mathbf{R}^{-1}\mathbf{1} - \hat{\mu}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1} = \mathbf{y}^T\mathbf{R}^{-1}\mathbf{1} - (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{-1}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{y}\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1} = 0$$

after substituting $\hat{\mu}$ from (2.8). Further simplification of (A.2) follows by noting that

$$(\mathbf{y}-\mathbf{1}\hat{\mu})^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\hat{\mu}) = (2\phi_1+n-1)\hat{\sigma}^2_{\boldsymbol{\psi}_B}$$

from (2.14). Also note that

$$\int \frac{1}{\sqrt{2\pi}}\left(\frac{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}{\sigma^2}\right)^{1/2}\exp\left(-\frac{(\mu-\hat{\mu})^T\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}(\mu-\hat{\mu})}{2\sigma^2}\right)d\mu$$

is the integral of the probability density of $\mu \sim N(\hat{\mu}, (\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}/\sigma^2)^{-1})$ and hence integrates to 1. Thus, in (A.1),

(A.3)
$$\int \exp\left(-\frac{(\mathbf{y}-\mathbf{1}\mu)^T\mathbf{R}^{-1}(\mathbf{y}-\mathbf{1}\mu)}{2\sigma^2}\right) d\mu \propto \exp\left(-\frac{(2\phi_1+n-1)\hat{\sigma}^2_{\boldsymbol{\psi}_B}}{2\sigma^2}\right)\left(\frac{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}{\sigma^2}\right)^{-1/2}.$$

Substituting (A.3) into (A.1) leaves just the integral with respect to σ^2:

$$\begin{aligned}
&\pi(\boldsymbol{\psi}_B|\mathbf{y}) \\
&\propto \frac{\pi(\boldsymbol{\psi}_B)}{|\mathbf{R}|^{1/2}} \int \frac{\phi_2^{\phi_1}}{\Gamma(\phi_1)}(\sigma^2)^{-\phi_1-1}\exp\left(-\frac{\phi_2}{\sigma^2}\right) \\
&\times \frac{1}{(\sigma^2)^{n/2}}\exp\left(-\frac{(2\phi_1+n-1)\hat{\sigma}^2_{\boldsymbol{\psi}_B}}{2\sigma^2}\right)\left(\frac{\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1}}{\sigma^2}\right)^{-1/2} d\sigma^2 \\
&\propto \frac{\pi(\boldsymbol{\psi}_B)}{|\mathbf{R}|^{1/2}\,(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{1/2}} \int (\sigma^2)^{-(\phi_1+\frac{n-1}{2}+1)}\exp\left(-\frac{\phi_2+\frac{(n-1)}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B}}{\sigma^2}\right) d\sigma^2 \\
&= \frac{\pi(\boldsymbol{\psi}_B)}{|\mathbf{R}|^{1/2}\,(\mathbf{1}^T\mathbf{R}^{-1}\mathbf{1})^{1/2}} \frac{\Gamma(\phi_1+\frac{n-1}{2})}{(\phi_2+\frac{n-1}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B})^{(\phi_1+\frac{n-1}{2})}},
\end{aligned}$$

which is the posterior distribution of $\boldsymbol{\psi}_B$ given in (4.3) up to constants of proportionality. The last line follows since

$$\int \frac{(\phi_2+\frac{n-1}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B})^{(\phi_1+\frac{n-1}{2})}}{\Gamma(\phi_1+\frac{n-1}{2})}(\sigma^2)^{-(\phi_1+\frac{n-1}{2}+1)}\exp\left(-\frac{\phi_2+\frac{n-1}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B}}{\sigma^2}\right) d\sigma^2$$

is the integral of an inverse-gamma probability density for σ^2 with parameters $\phi_1+\frac{n-1}{2}$ and $\phi_2+\frac{n-1}{2}\hat{\sigma}^2_{\boldsymbol{\psi}_B}$.

Acknowledgments. We thank the editor, associate editor, and reviewers for comments that led us to broaden the scope of this paper.

REFERENCES

[1] M. Abt, *Estimating the prediction mean squared error in Gaussian stochastic processes with exponential correlation structure*, Scand. J. Statist., 26 (1999), pp. 563–578, https://doi.org/10.1111/1467-9469.00168.

[2] L. S. Bastos and A. O'Hagan, *Diagnostics for Gaussian process emulators*, Technometrics, 51 (2009), pp. 425–438, https://doi.org/10.1198/TECH.2009.08019.

[3] M. J. Bayarri, J. O. Berger, E. S. Calder, K. Dalbey, S. Lunagomez, A. K. Patra, E. B. Pitman, E. T. Spiller, and R. L. Wolpert, *Using statistical and computer models to quantify volcanic hazards*, Technometrics, 51 (2009), pp. 402–413, https://doi.org/10.1198/TECH.2009.08018.

[4] H. Chen, *Bayesian Prediction and Inference in Analysis of Computer Experiments*, master's thesis, University of British Columbia, Vancouver, Canada, 2013, https://doi.org/10.14288/1.0074133.

[5] H. Chen, J. L. Loeppky, J. Sacks, and W. J. Welch, *Analysis methods for computer experiments: How to assess and what counts?*, Statist. Sci., 31 (2016), pp. 40–60, https://doi.org/10.1214/15-STS531.

[6] H. Dette and A. Pepelyshev, *Generalized Latin hypercube design for computer experiments*, Technometrics, 52 (2010), pp. 421–429, https://doi.org/10.1198/TECH.2010.09157.
[7] R. B. Gramacy and H. K. H. Lee, *Bayesian treed Gaussian process models with an application to computer modeling*, J. Amer. Statist. Assoc., 103 (2008), pp. 1119–1130, https://doi.org/10.1198/016214508000000689.
[8] M. S. Handcock and M. L. Stein, *A Bayesian analysis of kriging*, Technometrics, 35 (1993), pp. 403–410, https://doi.org/10.2307/1270273.
[9] D. Higdon, J. Gattiker, B. Williams, and M. Rightley, *Computer model calibration using high-dimensional output*, J. Amer. Statist. Assoc., 103 (2008), pp. 570–583, https://doi.org/10.1198/016214507000000888.
[10] M. Kennedy, *Description of the Gaussian Process Model Used in GEM-SA*, Technical report, University of Sheffield, Sheffield, UK, 2004; available online at https://www.tonyohagan.co.uk/academic/GEM/.
[11] M. C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, J. R. Stat. Soc. Ser. B Stat. Methodol., 63 (2001), pp. 425–464, https://doi.org/10.1111/1467-9868.00294.
[12] J. L. Loeppky, J. Sacks, and W. J. Welch, *Choosing the sample size of a computer experiment: A practical guide*, Technometrics, 51 (2009), pp. 366–376, https://doi.org/10.1198/TECH.2009.08040.
[13] M. D. Morris and T. J. Mitchell, *Exploratory designs for computational experiments*, J. Statist. Plann. Inference, 43 (1995), pp. 381–402, https://doi.org/10.1016/0378-3758(94)00035-T.
[14] M. D. Morris, T. J. Mitchell, and D. Ylvisaker, *Bayesian design and analysis of computer experiments: Use of derivatives in surface prediction*, Technometrics, 35 (1993), pp. 243–255, https://doi.org/10.2307/1269517.
[15] T. Nilson and A. Kuusk, *A reflectance model for the homogeneous plant canopy and its inversion*, Remote Sens. Environ., 27 (1989), pp. 157–167, https://doi.org/10.1016/0034-4257(89)90015-1.
[16] R. Paulo, *Default priors for Gaussian processes*, Ann. Statist., 33 (2005), pp. 556–582, https://doi.org/10.1214/009053604000001264.
[17] D. L. Preston, D. L. Tonks, and D. C. Wallace, *Model of plastic deformation for extreme loading conditions*, J. Appl. Phys., 93 (2003), pp. 211–220, https://doi.org/10.1063/1.1524706.
[18] G. O. Roberts and J. S. Rosenthal, *Examples of adaptive MCMC*, J. Comput. Graph. Statist., 18 (2009), pp. 349–367, https://doi.org/10.1198/jcgs.2009.06134.
[19] J. S. Rosenthal, *Optimizing and adapting the Metropolis algorithm*, in Statistics in Action: A Canadian Outlook, J. F. Lawless, ed., CRC Press, Boca Raton, FL, 2014, pp. 93–108, https://doi.org/10.1201/b16597-7.
[20] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statist. Sci., 4 (1989), pp. 409–423, https://doi.org/10.1214/ss/1177012413.
[21] T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, Springer, New York, 2003, https://doi.org/10.1007/978-1-4757-3799-8.
[22] M. Schonlau and W. J. Welch, *Screening the input variables to a computer model via analysis of variance and visualization*, in Screening: Methods for Experimentation in Industry, Drug Discovery, and Genetics, A. M. Dean and S. M. Lewis, eds., Springer, New York, 2006, pp. 308–327, https://doi.org/10.1007/0-387-28014-6_14.
[23] E. T. Spiller, M. J. Bayarri, J. O. Berger, E. S. Calder, A. K. Patra, E. B. Pitman, and R. L. Wolpert, *Automating emulator construction for geophysical hazard maps*, SIAM/ASA J. Uncertain. Quantif., 2 (2014), pp. 126–152, https://doi.org/10.1137/120899285.
[24] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, *Screening, predicting, and computer experiments*, Technometrics, 34 (1992), pp. 15–25, https://doi.org/10.2307/1269548.
[25] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, M. D. Morris, and M. Schonlau, *Response to James M. Lucas*, Technometrics, 38 (1996), pp. 199–203, https://doi.org/10.1080/00401706.1996.10484496.