Lightweight Emulators for Multivariate Deterministic Functions

Jonathan Rougier*
Department of Mathematics
University of Bristol

February 8, 2007

Abstract

An emulator is a statistical model of a deterministic function, to be used where the function itself is too expensive to evaluate within-the-loop of an inferential calculation. Typically, emulators are deployed when dealing with complex functions that have large and heterogeneous input and output spaces: environmental models, for example. In this challenging situation we should be sceptical about our statistical models, no matter how sophisticated, and adopt approaches that prioritise interpretative and diagnostic information, and the flexibility to respond. This paper presents one such approach, candidly rejecting the standard Smooth Gaussian Process approach in favour of a fully-Bayesian treatment of multivariate regression which, by permitting sequential updating, allows for very detailed predictive diagnostics. It is argued directly and by illustration that the incoherence of such a treatment (which does not impose continuity on the model outputs) is more than compensated for by the wealth of available information, and the possibilities for generalisation.

*Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.; e-mail J.C.Rougier@bristol.ac.uk.

1 Introduction

An *emulator* is a stochastic representation of a deterministic function f, constructed using function evaluations; the term *surrogate* has also been used, although recent practice has been to reserve this term for fast deterministic approximations of f, e.g., from using a lower-resolution solver. Emulators are useful wherever the function itself is costly to evaluate. A typical example is a solver for a system described by a set of differential equations and equations of state. The input vector x might simply be the initial value of the state vector, but it might also include unknown system parameters. Even an evaluation time of just a few seconds would make this a costly function were it necessary to explore a large input space. Thus emulators are useful in this situation for optimisation calculations designed to choose 'good' values for x, and also for inferential calculations designed to account for uncertainty in the 'true' x. Both of these activities fall under the general heading of *computer experiments*; see, e.g., Sacks *et al.* (1989), Koehler and Owen (1996) and Santner *et al.* (2003). Computer experiments for system inference, particularly within a Bayesian framework, are discussed in Currin *et al.* (1991), Kennedy and O'Hagan (2001), Craig *et al.* (1997, 2001), Goldstein and Rougier (2004, 2006a,b), Higdon *et al.* (2004), and Williams *et al.* (2006).

The challenge with building emulators is to find a parsimonious yet flexible representation for our initial beliefs about f. This is hard when f is multivariate, particularly when the components of $f(x)$ correspond to quantities with quite different characteristics. For example, f might be a computer simulator for weather, where x denotes a particular value for the initial value of the weather state-vector, and $f(x)$ computes the evolution of that state-vector in time. Thus $f(x)$ comprises collections of space- and time-indexed values, where each collection is a different type of quantity: pressure, temperature, humidity, and so on. In the physics these values are all related in complicated ways, and so our initial beliefs about f may be quite vague, possibly not extending much beyond simple transformations of components of $f(x)$ which might be thought to make the resulting marginals more Gaussian.

It should go without saying that where the actions that follow from an analysis based on f will be costly or irreversible, then the emulator for f ought to be carefully constructed. But not all computer experiments are so critical, and even in those that are, an initial cheap emulator can still be informative

both in choosing the evaluation points, and in refining our knowledge about f, in order that we might subsequently construct a more detailed and more accurate emulator. Therefore there is certainly a case for a lightweight emulator that runs out-of-the-box, which is easily tunable to quite vague prior beliefs about f, and which is generalisable should the need arise. This paper presents such an emulator. Section 2 describes its general principles; the notation is given in Table 1. Section 3 considers an implementation for the common situation where all inputs are continuous. Section 4 describes predictive diagnostics. Section 5 compares the lightweight emulator proposed here with its more heavy-weight alternative, the Smooth Gaussian Process. Section 6 provides a simple illustration, including specifying an informative prior. Section 7 describes and illustrates a tractable approach for generalising the emulator, by mixing over priors. Section 8 concludes. An Appendix describes the conjugate analysis used to construct the emulator, and the notation for the various distributions.

2 The function and the emulator

2.1 The function

We write our vector-valued function as

$$f_i(x) \qquad i \in \mathcal{I},\, x \in \mathcal{X}$$

where $\mathcal{X}$ is a set of d-tuples and $\mathcal{I}$ is a discrete set with k components. We term a value x an *input*, i an *index variable*, and $f_i(x)$ an *output*. An alternative notation (e.g., as adopted by Kennedy and O'Hagan, 2001), is to write $f(x; i)$, particularly if i indexes a continuous set, such as spatial locations, or times. In a predictive inference we typically have a well-specified index variables, matching (a) the system observations we want to calibrate against, and (b) the system quantities we wish to predict. In this case, there seems little need to permit the set $\mathcal{I}$ to be continuous. In situations where we would like to optimise over i, however (e.g., to find the best spatial location for a measuring instrument), there may be advantages to the continuous formulation when it is not possible to span the index-space with a high-resolution lattice, usually because $\mathcal{I}$ is too large. Nevertheless, this is simply a notational matter, so far as this paper is concerned. In our notation we drop the index variable

subscript to denote the full collection of outputs at any given input, and write $f(\cdot)$ to refer to the function in full generality.

One fruitful source of functions such as $f(\cdot)$ is the Natural Sciences, where models are constructed to represent complex systems, such as physical or environmental systems, or biological systems. Perhaps the most popular approach is the *compartmental model*, in which the behaviour of the state of the system across a number of different compartments is described by ordinary differential equations. One possible treatment (with a scalar state-vector, for simplicity) is to take i to index the compartment, x to be the values of model parameters, such as coefficients in the equations, and $f_i(x)$ to describe the equilibrium value of the state quantity in compartment i with parameters x. Another treatment is to specify a particular form of dynamic forcing, for example in the boundary conditions, in which case i represents a tuple of compartment index and time (or whatever the derivative is with respect to). In the more general setting of a vector-valued state, or of partial differential equations, i may represent a tuple of type, location and time. For example, in a climate model the value $f_i(x)$ might denote sea-surface temperature in the Azores in the year 2050, with model parameters x.

2.2 Emulating the function

These types of models can be extremely expensive to evaluate, and it is for this reason that we need emulators. An emulator is synomymous with a distribution function

$$\Pi_f(v, v', \ldots; x, x', \ldots) \triangleq \Pr\big(f(x) \leq v, f(x') \leq v', \ldots\big)$$

for any finite collection of xs and vs, where '$\triangleq$' denotes 'defined as'. Emulators are distinguished from other types of functional interpolator, for example a neural network, by providing a probabilistic assessment of uncertainty about $f(\cdot)$. In inferences based on the function, the emulator allows us to account for the uncertainty that arises from having only a limited number of function evaluations, what O'Hagan and Oakley (2004) term *code uncertainty*. For experimental design, emulators allow us to select a set of evaluations that is expected to be highly informative about function behaviour over a range of input values, both starting from scratch, and when augmenting an existing design.

It is usually convenient to express an emulator implicity through the specification of more primitive uncertain quantities. We adopt the general form

$$f(x)^T \equiv g(x)^T B + u(x)^T \tag{1}$$

where $g(\cdot)$ is a q-vector of known regressors, B is a $q \times k$ matrix of uncertain regression coefficients, and $u(\cdot)$ is a k-vector random field, termed the residual.

In our Lightweight Emulator (LWE) we hope to capture most of the variation in $f(\cdot)$ in terms of the fixed-length coefficient matrix B, allowing a relatively simple specification for $u(\cdot)$. We adopt a Bayesian approach, where we impose both structural and distributional restrictions on B and $u(\cdot)$, for tractability. Our LWE emulator can be seen as a fully Bayesian conjugate treatment of multivariate regression, or of multivariate kriging. A brief explanation is given here, with more details in the Appendix. The focus of the paper is not on this treatment *per se*, but on how we can use it when constructing emulators.

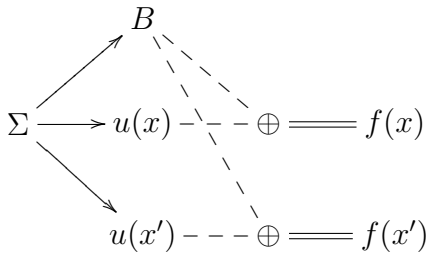
We start by introducing an additional uncertain quantity Σ, a $k \times k$ variance matrix. The *structural* restriction is that Σ separates B, $u(x)$ and $u(x')$ when $x \neq x'$, and consequently we must treat $u(\cdot)$ as a 'nugget', i.e. a process for which

$$u(x) \perp\!\!\!\perp u(x') \mid \Sigma \qquad x \neq x'. \tag{2}$$

This restriction allows us to do sequential updating of $\{B, \Sigma\}$ using model evaluations, because it implies that

$$f(x) \perp\!\!\!\perp f(x') \mid \{B, \Sigma\} \qquad x \neq x'. \tag{3}$$

This can be summarised in the graphical model



(4)

An evaluation of $f(x)$ is an observation on a set of known linear combinations of B and $u(x)$. This observation updates the joint distribution $\{B, \Sigma\}$. This

updated distribution feeds into our prediction of $f(x')$, where $x' \neq x$, through the updated joint distribution $\{B, u(x')\}$. Restricting $u(\cdot)$ to be a nugget is contentious, and is discussed in section 5.

To make the updating of $\{B, \Sigma\}$ tractable we impose further *parametric* restrictions, all of which are specified with respect to our choice of regressors, $g(\cdot)$:

1. The prior conditional distribution for the residual is stationary and Gaussian
$$u(x) \mid \Sigma \sim \mathrm{N}_k\left(\mathbf{0}, \Sigma\right) \tag{5}$$
for all $x \in \mathcal{X}$;

2. The prior conditional distribution of the regression coefficients is Matrix Normal:
$$B \mid \Sigma \sim \mathrm{MN}_{q\times k}\left(M, \Omega, \Sigma\right) \tag{6}$$
where the matrices M $(q \times k)$ and Ω $(q \times q)$ are hyperparameters;

3. The prior marginal distribution of Σ is Inverse Wishart
$$\Sigma \sim \mathrm{IW}_k\left(S, \delta\right) \tag{7}$$
where S $(k \times k)$ and δ (scalar) are hyperparameters.

Taking (6) and (7) together, we say that $\{B, \Sigma\}$ has a Matrix Normal Inverse Wishart (MNIW) distribution, from which it follows that our emulator for $f(x)$ has a multivariate Student-t distribution. Furthermore, this emulator is conjugate in the sense that updating by evaluations of the model just modifies the hyperparameters $\{M, \Omega, S, \delta\}$. Prior or updated, the emulator mean and variance functions, expressed in terms of the hyperparameters, are

$$\mathrm{E}\big(f(x)\big) = M^T g(x) \tag{8a}$$
$$\mathrm{cov}\big(f(x), f(x')\big) = \frac{w(x,x')}{\delta - 2}\, S, \tag{8b}$$

provided that $\delta > 1$ and $\delta > 2$, respectively, where

$$w(x,x') \triangleq g(x)^T \Omega g(x') + \delta(x - x'); \tag{8c}$$

where in (8c) the function $\delta(\cdot)$ is the Dirac delta function (not to be confused

with δ, the hyperparameter of the Invese Wishart, or, below, δ_{ij}, the Kronecker delta function). The notation in this paper is summarised in Table 1.

One notable feature of (8) is that the variance function separates into a part attributable to x and a part attributable to i. This separability is a cornerstone of tractable statistical modelling over products of different types of spaces: the implications have to be fairly brutal, in the light of the huge reduction in complexity that results. In this case we can see that the ratio

$$\frac{\mathrm{cov}\big(f_i(x), f_{i'}(x')\big)}{\mathrm{cov}\big(f_i(x''), f_{i'}(x''')\big)} = \frac{w(x, x')}{w(x'', x''')}$$

is dependent only on the four model-input values, and not on $\{i, i'\}$, the pair of outputs we are selecting. For example, if we fix $i = i'$, $x = x'$ and $x'' = x'''$ and find that the variance of salinity at input x is exactly twice that at x'', then we must conclude that the variance of temperature at x is also exactly twice that at x''. Note that this is a feature of the emulator predictions, but it does not constrain the way in which the emulator updates. That is to say, if we considered a second emulator in which we partitioned the outputs and treated them as block independent, then the predictive distributions within each block would be the same in both emulators. Partitioning the outputs in this way is beneficial only if we also allow our choices for the regressors or for the prior hyperparameters to vary by block.

A discussion of this emulator is deferred until section 5, where it is contrasted with the more standard approach. A simple way to generalise the emulator and 'defeat' the separability is discussed and illustrated in section 7.

2.3 Assessing the hyperparameters

The prior is summarised in terms of initial choices for the hyperparameters $\Psi \triangleq \{M, \Omega, S, \delta\}$. But since the updating is conjugate, the following considerations may also be applied to the updated hyperparameters, to derive summary descriptions of the emulator at any stage of the process. To distinguish the different choices, the prior values of the hyperparameters are labelled Ψ_0.

As described in the Appendix, there exist 'default' choices for Ψ_0 that result in the expected values of the updated B and Σ being the Maximum

Table 1: Notation used in the paper.

Symbol	Size	Definition
	Function quantities	
x	p	Vector of function inputs, index j
$f(x)$	k	Vector of function outputs, index i
$F; X$	$n \times (k+p)$	Ensemble of function evaluations
	Emulator quantities	
$g(x)$	q	Vector of regressors, index r
B	$q \times k$	Matrix of regressor coefficients
$u(x)$	k	Vector of residuals
Σ	$k \times k$	'Column' variance matrix
	Hyperparameters, Ψ	
M	$q \times k$	Mean of B
Ω	$q \times q$	'Row' variance matrix
S	$k \times k$	Scale matrix for Σ
δ	scalar	Degrees of freedom for Σ
Ψ_0	Collection	Prior hyperparameters
	Other	
$\ell_v(\cdot)$		Legendre polynomial of order v
$\delta(\cdot)$		Dirac delta function
δ_{ij}		Kronecker delta function

Likelihood (ML) estimators in multivariate regression, namely

$$\Omega^{-1} = \mathbf{0}_{q\times q}, \quad S = \mathbf{0}_{k\times k}, \quad \delta = 2, \tag{9}$$

for all M, so we might add $M = \mathbf{0}_{q\times k}$. Where there are a very large number of evaluations of the model these default choices may be acceptable, even though $S = \mathbf{0}$ is very unlikely to be a reasonable summary of well-informed judgements about $f(\cdot)$. In general, however, we should consider how to make informative choices based on our (expert's) judgements about the model.

One possibility is to make judgements about the conditional behaviour of $f(x)$, and how it varies for different x. Reasonable choices for Ψ_0 can then be inferred from (8). However, this is a very large collection of assessments, unless it is judged that our prior for $f(\cdot)$ has quite simple structure.

An alternative approach is to treat the input itself as an uncertain quantity. In many applications where $f(\cdot)$ is a model of a complex system, it is natural to think of a 'best' or 'correct' input x^*: that input for which $f(x^*)$ is the best representation of the underlying system (see, e.g., Rougier, 2007, for a discussion of this approach in climate modelling). Tracing uncertainty about x^* through to uncertainty about $f(x^*)$ is known as *uncertainty analysis*, an application where emulators have proved very useful for expensive functions (O'Hagan *et al.*, 1999; Oakley and O'Hagan, 2002). Then judgements about $f(x^*)$ can be used to augment those about $\{f(x); x \in \mathcal{X}\}$. Reasonable choices for Ψ_0 can be inferred in conjunction with a distribution Π_{x^*}.

An effective way of simplifying this process is to use regressors $g(\cdot)$ that are orthonormal with respect to Π_{x^*}, i.e.,

$$\langle g_r, g_{r'} \rangle \triangleq \int g_r(x)\, g_{r'}(x)\, \mathrm{d}\Pi_{x^*}(x) = \delta_{rr'} \tag{10}$$

where $\delta_{rr'}$ is the Kronecker delta, and r will be used to index the regressors. We can take $g_1(x) = 1$. Partitioning M as

$$M = \begin{pmatrix} M_1 \\ M_{2:} \end{pmatrix} \quad \begin{matrix} 1 \times k \\ (q-1) \times k, \end{matrix}$$

we have, for orthonormal regressors,

$$\mathrm{E}\big(f(x^*)\big) = {M_1}^T \tag{11a}$$

$$\mathrm{var}\big(f(x^*)\big) = {M_{2:}}^T M_{2:} + \frac{\mathrm{tr}\,\Omega + 1}{\delta - 2}\, S, \tag{11b}$$

where $x^* \sim \Pi_{x^*}$, which follows directly from (8). Note that orthonormal regressors are unitless, so that the columns of M have the same units as the model-outputs, and Ω is unitless. The prior for M_1 should be set according to our judgements about the aggregate means. We may choose to treat the prior for Ω as a diagonal matrix, because we have no opportunity to discern the impact of its off-diagonal structure in these summaries.

Another way of summarising the hyperparameters for an uncertain x^* is through an 'R^2' value for each output. We can think of R^2 as the squared correlation between the predicted and actual values, or

$$R_i^2 \triangleq \frac{\mathrm{cov}\big(\mu_i(x^*), f_i(x^*)\big)^2}{\mathrm{var}\big(\mu_i(x^*)\big)\ \mathrm{var}\big(f_i(x^*)\big)} \tag{12}$$

for output i, where $\mu(x)$ is the mean function given in (8a). For orthonormal regressors, this simplifies to

$$R_i^2 = \frac{m_i^2}{m_i^2 + \xi_i^2} \qquad \xi_i^2 \triangleq \frac{\mathrm{tr}\,\Omega + 1}{\delta - 2} S_{ii}, \tag{13}$$

where m_i^2 is defined as the ith term on the diagonal of ${M_{2:}}^T M_{2:}$. Obviously if the prior for $M_{2:}$ is $\mathbf{0}$, then the prior R^2 values will also be $\mathbf{0}$. We can extract more information out of the R^2 calculation by focusing on subsets of the regressors, such as the first r according to some ordering; in this case the 'R_i^2' value is $m_{ir}^2/(m_i^2 + \xi_i^2)$, where m_{ir}^2 is the ith term on the diagonal of ${M_{2:r}}^T M_{2:r}$. If these regressors were monomial terms, for example (section 3), it might be interesting to compute the R^2 values for the linear terms, and think of all higher-order terms as being relegated to a generalised (non-stationary) residual.

For Ψ_0, we might hope to be able to assess reasonable values for M and for the lumped parameters ξ_i, $i = 1, \ldots, k$, by considering the unconditional mean and variance, and possibly the R^2 values (if we choose $M_{2:} \neq \mathbf{0}$). To decompose ξ_i into values for the hyperparameters $\mathrm{diag}\,\Omega$, δ and S, we might start with the

cautious choice $\delta = 3$, which would represent minimal prior information with a finite second moment. More generally, δ reflects our confidence in our prior assessment in terms of the number of function evaluations that it might be considered equivalent to. The value $\mathrm{tr}\,\Omega/(\mathrm{tr}\,\Omega + 1)$ represents the proportion of the expected conditional variance that is attributable to uncertainty in the coefficients. If this can be fixed, and δ is chosen, then the diagonal of S can be determined to give appropriate values for $\mathrm{Sd}(f_i(x^*))$. The off-diagonal components of S can be parameterised through a correlation matrix capturing simple structure in the model outputs, such as block independence, or spatio-temporal effects.

All this might seem rather hit-or-miss, both in terms of our ability to provide detailed quantification of our prior judgements about the model, and in terms of how the particular choices for the hyperparameters can be settled on. But this really misses the point: the real issue is the extent to which we do better than the default ML choices for the hyperparameters, as given in (9), if we are willing to make more detailed judgements. At the very least we can take the ML choices as bounds, and investigate whether we feel comfortable moving away from the bounds. We also have the possibility of generalising our choice for Ψ_0, if we judge that to be appropriate (section 7).

3 Continuous inputs

When all of the inputs are continuous, we can make some further suggestions about the choice of regressors. A generalisation of the probability integral tranform can be used to map the domain of $f(\cdot)$ onto the unit cube, on which the transformed Π_{x^*} is independent and uniform (Rosenblatt, 1952). Often, this extra mapping will not to be required, as it is a common judgement to treat the components of x^* as independent, bounded and uniform (possibly after marginal transformations). But note also that in some applications we will discover, during the experiment, that the model does not evaluate for certain choices of inputs. Even if we are prepared to attach zero probability to these 'unsuccessful' input values, we are still left with the problem that the region of successful choices is not well-defined, nor is it easy to summarise. This situation is not treatable using the methods of this section.

We will restrict attention to the case where Π_{x^*} is absolutely continuous, and treat the domain of $f(\cdot)$ as $[0,1]^d$. The main thing this rules out

is 'switches' in x, which sometimes occur in physical models where we can selectively activate subprocesses (see, e.g., the climate model in Murphy *et al.*, 2004). In this case we must think of $f(\cdot)$ as conditional on a given setting of the various switches. If we so choose, we can build the bigger emulator which includes both the continuous and discrete inputs, in which case the following analysis holds for the continuous inputs only, expressed as a function of the discrete ones. Regression-type approaches are natural for these types of emulators: see, for example, Rougier *et al.* (2006).

3.1 Choice of regressors

Where $f(\cdot)$ is continuous and bounded on $[0,1]^d$ we can use a tensor product of Legendre polynomials for $g(\cdot)$, shifted from their usual domain of $[-1,1]$ to $[0,1]$. Each component $g_r(\cdot)$ is a monomial that can be identified by the exponent of each of the d input components, so that, if $d=2$, $(0,0)$ would be the constant, $(1,0)$ would be a linear term in x_1, $(0,2)$ would be a quadratic term in x_2, $(1,1)$ would be a bilinear interaction term between x_1 and x_2, and so on. If r is now generalised to be a d-vector of exponents, $r \in \mathbb{Z}^d$, then an individual monomial can be represented as

$$g_r(x) = \prod_{j=1}^{d} \ell_{r_j}(x_j)$$

where $\ell_v(\cdot)$ is the Legendre polynomial of order v, shifted onto $[0,1]$, and j will be used to index the model-inputs.

This structure generalises immediately to any situation where we construct the regressors as tensor products, in such a way that we can index the individual terms: the r_j do not have to be exponents. One very flexible approach is to augment a small collection of global regressors for each input (e.g., a constant and a linear term) with additional regressors from a given covariance function, which gives rise to a set of approximately orthonormal *prinicpal kriging functions* (see, e.g. Sahu and Mardia, 2005). These will be explored elsewhere; this paper focusses on monomials.

An and Owen (2001) provide a convenient way of specifying collections of

monomials in terms of a triplet $\kappa = (\kappa_1, \kappa_2, \kappa_3)$. The constraints

$$\sum_{j=1}^{d} r_j \leq \kappa_1 \qquad \sum_{j=1}^{d} (1 - \delta_{0r_j}) \leq \kappa_2 \qquad \max_{j \in \{1,\ldots,d\}} r_j \leq \kappa_3 \tag{14}$$

define a collection of monomials. In words, κ_1 is the maximum sum of the exponents, κ_2 is the maximum number of non-zero exponents, and κ_3 is the maximum single exponent. Thus $\kappa = (2, 2, 2)$ would define a collection of constant, linear, quadratic and bilinear interaction monomials: 21 altogether if $d = 5$. It is straightforward to generate the complete set of monomials for any given κ, using recursion. Using this approach, we can easily specify a large collection of monomials for $g(\cdot)$, which we can hand-edit if we require more control.

3.2 Specifying the prior for Ω

We can also use this monomial structure to simplify the task of specifying $\omega \triangleq \operatorname{diag} \Omega$ in the prior (section 2.3). It is natural to follow the structure in the monomial terms, and to take advantage of the orthonormality of the regressors to relate coefficient uncertainty directly to output uncertainty. One possibility is to use the parameterisation

$$\omega_r = \tau_0 \prod_{j=1}^{d} (\tau_j)^{r_j} \qquad \tau \triangleq (\tau_0, \tau_1, \ldots, \tau_d) \gg \mathbf{0}, \tag{15}$$

where we choose tau. The multiplier τ_0 allows for scale differences between the variances of $B^T g(x)$ and $u(x)$, because the d-fold product in (15) is normalised in the sense that its value for the intercept ($r = \mathbf{0}$) is always one.

There are other ways we might have specified ω_r, for example in terms of the values of the $\kappa(r)$-triple inferred from (14). The attraction of (15) is that it allows us to duplicate values for τ_j across similar model-inputs. It often happens in models of physical processes that a subset of the model-inputs corresponds to the same *type* of quantity. In fact when d is large, it is usually because x contains collections of quantities of the same type, often sub-indexed by location and/or time. A model of a hydrocarbon reservoir, for example, will have spatial fields of porosity and permeability as model-inputs (see, e.g., Craig *et al.*, 1997). In this case we may choose to use a common value for τ_j

across all x_j of the same type, leading to a large reduction in the number of hyperparameters required in τ.

3.3 Emulator summaries

We can take advantage of the product structure of each $g_r(\cdot)$ and the orthonormality of the legendre polynomials to derive simple expressions for low-dimensional summaries of $f(x^*)$, expressed in terms of main effects, two-way interactions, and so on; Oakley and O'Hagan (2004) adopt a similar type of approach in their treatment of *sensitivity analysis*. Consider conditioning on some of the components of x^*, say those in the subset $J \subseteq \{1, \ldots, d\}$; for the fully-aggregated case we would have $J = \emptyset$, while for the main effect with respect to $x_j^* = x_j$ we would have $J = \{j\}$. To reduce clutter we write $(\cdot \mid x_J)$ for $(\cdot \mid x_J^* = x_J)$ in what follows.

For the conditional expectation we have, starting from (8a),

$$\mathrm{E}(f(x^*) \mid x_J) = \mathrm{E}\big(\mathrm{E}(f(x^*) \mid x^*) \mid x_J\big) = M^T h(x_J) \tag{16a}$$

where

$$h_r(x_J) \triangleq \mathrm{E}\big(g_r(x^*) \mid x_J\big) = \prod_{j \in J} \ell_{r_j}(x_j) \times \prod_{j \notin J} \delta_{0 r_j}. \tag{16b}$$

In words, the component $h_r(x_J)$ will be zero unless all of the not-conditioned-on inputs are represented as constants, in which case it will be $\prod_J \ell_{r_j}(x_j)$. In the limits $J = \emptyset$ and $J = \{1, \ldots, d\}$ we recover (11a) and (8a), respectively, where we adopt the standard convention that products over the empty set evaluate to 1.

For the conditional variance of the emulator expected value we have

$$\mathrm{var}\big(\mathrm{E}(f(x^*) \mid x^*) \mid x_J\big) = M^T \mathrm{var}\big(g(x^*) \mid x_J\big)\, M \tag{17a}$$

where the variance may be computed from (16b) and the $q \times q$ matrix $H(x_J)$, where

$$H_{rr'}(x_J) \triangleq \mathrm{E}\big(g_r(x^*)\, g_{r'}(x^*) \mid x_J\big) = \prod_{j \in J} \ell_{r_j}(x_j)\, \ell_{r'_j}(x_j) \times \prod_{j \notin J} \delta_{r_j r'_j}. \tag{17b}$$

Where $J = \emptyset$ we will have $H(x_J) = I_q$, the $q \times q$ identity matrix. For the

conditional expectation of the emulator variance we have

$$\mathrm{E}\big(\mathrm{var}(f(x^*) \mid x^*) \mid x_J\big) = \frac{\mathrm{tr}\, H(x_J)\Omega + 1}{\delta - 2}\, S \tag{18}$$

using (8b) and (17b). Adding (17a) and (18) gives the variance matrix $\mathrm{var}\big(f(x^*) \mid x_J\big)$. Again, in the limits $J = \emptyset$ and $J = \{1, \ldots, d\}$ we recover (11b) and (8b). The product structure in each $g_r(\cdot)$ is useful in simplifying the expressions that come *between* these two limiting cases. It is informative to contrast the two sources of variance, since in some outputs the total variance may be dominated by a sensitive mean function, while in others it may be dominated by a large variance function.

4 Diagnostics

This paper advocates a fully-Bayesian approach with proper priors informed by judgements about $f(\cdot)$. Even so, there are two compelling reasons for producing detailed diagnostic information. First, diagnostics are vital in selling the emulator to the expert who constructed $f(\cdot)$, who may be suspicious about a statistical framework that appears to replace the model he spent several months constructing (this is a misconception, but a common and stubborn one). Second, in the interests of tractability, our judgements about $f(\cdot)$ have been structurally constrained and then shoe-horned into a MNIW prior. The adoption of a prior we do not necessarily subscribe to compromises our quantification of Ψ_0, and we would want to perform a detailed sensitivity analysis regarding our choices, and, possibly, generalise those choices somewhat.

The tractability of the LWE cuts both ways. It constrains our judgmements about $f(\cdot)$, but it also ensures that an emulator is very simple to construct for any given ensemble of evaluations. This means that we can use predictive diagnostics to validate our prior judgements, and the structural and parametric restrictions we have imposed. Predictive diagnostics are expensive, but very valuable in situations where we expect our prior judgements to be influential in the posterior. For example, in the illustration in section 6 we have 30 evaluations, but would like to use many more regressors in order to capture the non-linearity and interactions we anticipate from our function. Consequently the ensemble of evaluations will constrain certain linear combinations of the regression coefficients, but will leave other combinations at their prior settings:

exactly what is constrained and what is not will depend in a subtle way on the experimental design (see, e.g., Pistone *et al.*, 2000). A predictive diagnostic such as *leave-one-out* is informative about the model-specification, but also gives us a clear visual picture of the predictive uncertainty that arises in the updated emulator, $n-1$ being close to n.

One limitation of leave-one-out, and its generalisations, is that it is hard to construct simple summary measures across the whole ensemble. A diagnostic which does provide such a summary uses the *prequential* approach (Dawid, 1984; Cowell *et al.*, 1999). This is closely related to *one-step-ahead*, and so it has the weakness that we cannot visualise our updated emulator's predictive uncertainty over a range of x values, since for much of the diagnostic we will be updating with many less than n evaluations. So in practice, both diagnostics are useful.

The prequential ('predictive sequential') approach is based on a sequential analysis using increasing numbers of evaluations from the ensemble. Denote by $\pi_m(\cdot)$ the probability density function for predicting $f(X_m)$ using evaluations $1,\ldots,m-1$ from the ensemble. We can use this density to evaluate the actual outcome F_m in terms of the logarithmic scoring rule

$$S_m \triangleq -\ln \pi_m(F_m) \qquad m=1,\ldots,n. \tag{19}$$

Summing these scores gives

$$\sum_{m=1}^{n} S_m = -\ln \prod_{m=1}^{n} \pi_m(F_m) = -\ln \pi(F; X, \Psi_0) \tag{20}$$

where $\pi(F; X, \Psi_0)$ is the marginal density of the ensemble, showing that this sum is invariant to the ordering of the evaluations in the ensemble. To compute the full set of scores involves having a proper Ψ_0 and then building $n-1$ emulators of increasing size using the ensemble of evaluations. For an improper Ψ_0 e.g., the ML-prior given in (9), some of the evaluations would need to be reserved.

We can calibrate any individual score in terms of its mean and variance under the vaguely-specified null hypothesis

$$\mathrm{H}_0 : \text{The statistical modelling choices are appropriate.}$$

The predictive distribution of the emulator is a multivariate Student-t, denoted $\mathrm{T}_k\,(v; \mu, \Sigma, \delta)$ in the standard parameterisation. Thus

$$S_m = -c + \frac{\delta + k}{2} \ln\bigl(1 + (k/\delta)\, y\bigr), \tag{21a}$$

where c is the logarithm of the Student-t normalising constant and

$$y \triangleq k^{-1}(v - \mu)^{\mathrm{T}} \Sigma^{-1} (v - \mu). \tag{21b}$$

Under the null hypothesis, y has a Fisher $F_{k,\delta}$ distribution; see, for example, Bernardo and Smith (1994, p. 140, p. 123), although note there is a missing k^{-1} in their definition of y.[1] Therefore the mean and variance of S_m under H_0 can easily be computed using a one-dimensional numerical integration. An alternative, when $\delta \gg k$ so that $\ln(1 + (k/\delta)y) \approx (k/\delta)y$, is to use an approximation based on the mean and variance of $F_{k,\delta}$.

For sequential analysis, we can compute the mean and variance of $\sum_{m=1}^{i} S_m$ in exactly the same way, although this would involve the inversion of an $n \times n$ variance matrix. An alternative is to construct a *prequential monitor*,

$$Z_i \triangleq \frac{\sum_{m=1}^{i} S_m - \sum_{m=1}^{i} \mathrm{E}(S_m)}{\sqrt{\sum_{m=1}^{i} \mathrm{var}(S_m)}} \qquad i = 1, \ldots, n. \tag{22}$$

Under fairly broad conditions (Seillier-Moiseiwitsch and David, 1993), this diagnostic has the property that as $i \to \infty$ so the distribution of Z_i tends to the standard Gaussian under H_0. This alternative has the advantage of being 'free', as it reuses quantities we have already computed; it also provides a more precise calibration than standardised deviations from the mean, for large n.

[1] *Proof:* the k-vector v is a multivariate Student-t quantity with mean μ, scale matrix Σ, and degrees of freedom δ: it can be represented as

$$v - \mu = \frac{(x - \mu)}{\sqrt{w/\delta}}$$

where the k-vector x is Gaussian with mean μ and variance Σ, w is χ^2_δ, and x and w are independent. Hence

$$k^{-1}(v - \mu)^{\mathrm{T}} \Sigma^{-1} (v - \mu) = \frac{k^{-1}(x - \mu)^{\mathrm{T}} \Sigma^{-1} (x - \mu)}{w/\delta} = \frac{w'/k}{w/\delta}$$

where w' is χ^2_k and w' and w are independent. This is the characterisation of the $F_{k,\delta}$ distribution. Note that (3) in Mardia *et al.* (1979, p. 43) is incorrect.

Finally, a brief comment on the interpretation of diagnostic information, particularly one-step-ahead and prequential. In the Bayesian approach it is quite acceptable knowingly to hold prior judgements that are different from those in the likelihood. In this situation the early values of the diagnostics will be aberrant, while the likelihood is gradually asserting itself over the prior. Crudely, we might expect this to take at least q evaluations, at which point every linear combination of the regression coefficients has been updated. But during this initialisation we are unable to distinguish between a prior/likelihood conflict and a deeper problem arising from the our structural choices. One solution is to contrast the diagnostics from our Ψ_0 with those that arise when we set Ψ_0 equal to the tuned value

$$\hat{\Psi}_0 \triangleq \underset{\Psi \in \Gamma}{\operatorname{argmax}}\ \pi(F; X, \Psi) \tag{23}$$

for some set Γ of our choosing. If this tuning does not improve the diagnostics then we might reasonably conclude that there are inappropriate structural restrictions in our emulator. An alternative tuning, which may be more robust, is to maximise the 'cross-validation density' in place of the marginal density,

$$\prod_{i=1}^{n} \pi(F_i \mid F_{(i)}; X, \Psi) \tag{24}$$

where $F_{(i)}$ denotes all but the ith evaluation (see, e.g., **?**Gelfand and Dey, 1994). This can be computed directly from the leave-one-out diagnostic.

5 Comparison with standard methods

It is informative at this point to compare our treatment of the emulator with the more standard approach using a Smooth Gaussian Process (SGP) prior for $f(\cdot)$, as exemplified by, for example, Currin *et al.* (1991), Haylock and O'Hagan (1996), and Kennedy and O'Hagan (2001). The first point to note is that, as yet, the SGP has not been effectively generalised to multiple output types, although it can handle mutiple indices (e.g., location) for a given type. Therefore different output types have to be treated independently for any given x. The LWE handles multiple types, within the constraints of the separable covariance function (section 2.2).

Now focus on the case where $f(x)$ is a scalar. In principle, the difference

between the LWE and the SGP is that the former represents the residual as a nugget, while the latter uses a Gaussian process with continuous sample paths (hence, 'smooth'). The SGP requires the specification of a prior variance function

$$\kappa(x, x') \triangleq \mathrm{var}\big(u(x), u(x')\big) ;$$

the LWE requires only the specification of a scalar prior variance. However, it is important to distinguish between what may be done in principle, and what is actually done in practice. With the SGP it appears to be difficult to make an informed choice for the variance function $\kappa(\cdot, \cdot)$; the approach of Craig *et al.* (1997, 2001) is an exception. Thus a parametric form is chosen, and then the hyperparameters, typically controlling variance, roughness (differentiability) and correlation length, are estimated from the ensemble, and plugged in. Typically this choice of variance function follows well-trodden lines: the variance function is restricted to be stationary, and the correlation function to be a product of (univariate) Matérn correlation functions, and the hyperparameters are estimated by maximising the marginal density or the cross-validation density (see, e.g., Santner *et al.*, 2003). This separability across inputs is a very strong structural restriction; it is discussed by O'Hagan (1998).

The estimation of the hyperparameters in the SGP emulator has important implications for the choice of regressors. The residual has a representation in terms the regressors, and so there is an identification problem between the variance function hyperparameters and the regression coefficients. This problem has bedeviled the technique of *universal kriging* in Spatial Statistics (see, e.g., Cressie, 1991, sec. 3.4). One solution, commonly adopted with the SGP, is to limit the number of regressors, often to just a constant (possibly extended to linear terms): this is akin to *ordinary kriging* in Spatial Statistics (Cressie, *op. cit.*); a more sophisticated estimation method for the hyperparameters can also help. Another is to keep the regressors but dispense with the structured residual, as done by the LWE: this is *trend surface prediction*. Another is to adopt a fully Bayesian hyperarchical approach, avoiding the problem of directly quantifying the hyperparameters (see, e.g., Banerjee *et al.*, 2004). However, this typically involves Markov chain Monte Carlo (MCMC), and does not scale well to applications with large numbers of inputs, or where large numbers of emulators are to be constructed, for diagnostic purposes.

Now let's examine the strengths and weaknesses of the nugget residual. The

weakness is easy to spot. The LWE will not, by default, perfectly interpolate the ensemble. If $f(\cdot)$ is a smooth deterministic function then this is simply wrong, because if $f(x)$ is known then we ought to have

$$\lim_{x' \to x} \mathrm{E}\Big(\big[f(x') - f(x) \big]^2 \Big) = 0.$$

O'Hagan (2006) defines perfect interpolation as a key desideratum of emulators. Now the case for the defence. First, if the residual variance is small then this failure to interpolate perfectly will make little difference in practice, when using the emulator to predict the behaviour of the function over a range of possible input values. We can investigate the size of the residual and compare it to $\mathrm{var}\big(f(x)\big)$ for different x, and also to $\mathrm{E}\big(\mathrm{var}(x^*)\big)$, so that we can satisfy ourselves that it *is* small. Where it is not, we have the option of refining our choice of regressors. Second, in many applications we will choose not to model $f(x)$ as a function of all of its inputs, but to focus on a subset. This has been a common theme since the inception of the field of computer experiments (see, e.g., Owen *et al.*, 1989); Craig *et al.* (1996, 1997) refer to the components of this subset as the *active* inputs. Treating $f(\cdot)$ as an explicit function only of the active inputs, there will be unattributable variation from the non-active inputs, which necessitates the inclusion of a nugget such as the LWE residual.

Now to the benefit of the nugget residual. The LWE is fast to fit, and its computational cost scales linearly in n, in the sense that the emulator can be updated sequentially, one evaluation at a time. The cost of the SGP, though, scales as n^3, this being the cost of the Choleski decomposition required to invert an $n \times n$ variance matrix (see, e.g., Golub and Van Loan, 1989, p. 142). With single-type multivariate output, this cost scales as $(kn)^3$, unless an outer product design has been used (scales as n^3), which has its own drawbacks; see Kennedy and O'Hagan (2001) and the discussion in Rougier (2001). As noted by Koehler and Owen (1996, p. 265), the cost of even one inverse will rapidly become prohibitive as the ensemble grows. The challenge of approximating this calculation has been much studied in Spatial Statistics, using both subsets of the ensemble (see, e.g., Besag, 1975; Jones and Vecchia, 1993) or sparse matrix methods (Cornford *et al.*, 2004). This is an extra layer of complexity that is completely absent in the LWE. Another complication with the SGP when n is large is that the variance matrix can be ill-conditioned, in which case one solution is to inflate the diagonal (Higdon *et al.*, 2004): of course this means

that that the emulator no longer interpolates perfectly.

So lack of speed is the drawback of the SGP. This should be measured not in clock-time, though, but in the opportunity to perform detailed diagnostics. There is no sense in which a SGP requires less diagnostic checking than a LWE: in practice both approaches make strong structural choices, and both may be expected to extrapolate badly even in some of the cases where they interpolate well. On this basis, predictive diagnostics are vital for building confidence in the emulator. These predictive diagostics require us to build many emulators. The slowness of the SGP becomes much more acute, and corners may have to be cut, e.g. by not re-estimating the variance function hyperparameters for each different subset of the ensemble. Exactly the same comments apply in situations where we would like to experiment with different transformations of the model-inputs and model-outputs, and different choices of regressors.

To summarise, consider the following question: if you discovered that a LWE did not perform well, would you then be happy to switch to a SGP? I for one would be wary. The same applies the other way around, of course, but the difference is that the LWE costs almost nothing to fit and can be extensively audited, so is the natural candidate to try first.

6 Illustration

This illustration is based on the Atlantic Ocean compartmental model of Zickfeld *et al.* (2004), and uses an ensemble constructed for Goldstein and Rougier (2006b). This comprises $n = 30$ evaluations designed as a maximin latin hypercube over $d = 5$ continuous inputs, with $k = 7$ outputs. This small number of evaluations highlights the role of the prior.

The calculations were performed in the R statistical programming environment (R Development Core Team, 2004).

6.1 Specifying the prior

Section 2.3 discusses how we might specify informative priors for the emulator. Here I outline some simple choices based on my experiences of this model, and of the system it is designed to represent. These choices are genuine attempt to describe my judgements, although the values have been tweaked a little to make the illustration more informative.

Regressors. The inputs are continuous, and I choose to use monomial regressors up to cubics and three-way interactions. In the scheme of An and Owen (2001), this is described by the triplet $\kappa = (3, 3, 3)$; see section 3.1. This gives $q = 56$ regressors in all. (I tried more regressors, including quartics, but it appeared to make no difference.)

Coefficient means. The prior values for the first row of M are set according to my judgements about the unconditional mean of $f(x^*)$, where x^* is uniform on the range of possible values for x, which is $[0, 1]^5$, after transforming the inputs. The first three model-outputs are Atlantic Ocean temperatures in Centigrade, for which I chose 10, 5, 15 (South, North, Tropical). The fourth and fifth are salinity differences in Practical Salinity Units, for which I chose 0 and 0. The sixth is meridional overturning in Sverdrups (Sv, $10^6\,\mathrm{m}^3\,\mathrm{s}^{-1}$), for which I chose 20, and the seventh is critical freshwater forcing in Sv, for which I chose 0.1. This final output is likely to have the most complex relationship with the inputs, as it is the solution of an inverse problem.

I have strong prior judgements about the relationship between the first three inputs and the first three outputs. Let $v_{1:3} \triangleq (v_1, v_2, v_3)$ be atmospheric temperature forcing in the South, North and Tropical compartments. Then, to first order, I judge that

$$\Delta f_{1:3} \propto \Delta v_{1:3}. \tag{25}$$

The inputs $x_{1:3}$ are related to $v_{1:3}$ by

$$x_{1:3} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \\ -1 & 0 & 1 \end{pmatrix} v_{1:3}, \quad \text{or} \quad x_{1:3} \equiv A\, v_{1:3}\,; \tag{26}$$

this was to ensure that the ordering $v_2 \leq v_1 \leq v_3$ was preserved in the choices of $x_{1:3} \in [0, 1]^3$. Therefore, in terms of model-inputs and model-outputs,

$$\Delta f_{1:3} \propto \Delta(A^{-1} x_{1:3}). \tag{27}$$

I chose the proportionality constant to be 3, because the model-inputs are rescaled, but the model-outputs are not. Thus the values $3A^{-1}$ are used in M to relate the linear regressors in the first three model-inputs to the first three model-outputs. All the other prior values of M are set to zero.

Variances. Appropriate values for the variance $\mathrm{var}\big(f(x^*)\big)$ can now be inferred from choices for the lumped parameters ξ_i^2, defined in (13). My choices and the resulting implied values are

ξ_i:	3	3	3	0.2	0.2	5	0.1
$\mathrm{Sd}\big(f_i(x^*)\big)$:	5.2	4.2	6	0.2	0.2	5	0.1
R_i^2 (%):	67	50	75	0	0	0	0

where the difference between the first three outputs (columns) and the others four is due to non-zero values in M.

Finally, I must represent the value of ξ_i^2 in terms of choices for $\mathrm{diag}\,\Omega$, δ, and S, where I choose to treat the prior Ω as diagonal. I choose $\delta = 3$. I expect the residual variance to be quite small, so I choose $\mathrm{tr}\,\Omega = 9$, which implies that $S_{ii} = \xi_i^2/10$. I choose to zero the off-diagonal elements in S. For the actual values of $\mathrm{diag}\,\Omega$, I use the approach in section 3.2, and choose $\tau_j = 0.1$ for $j = 1, 2, 3$, so that the standard deviation of a cubic coefficient is 0.1 that of a linear coefficient: I expect the response to the three temperature inputs to be mainly linear. I choose $\tau_j = 1$ for $j = 4, 5$: I expect the response to these two inputs to be more complicated. It follows that $\tau_0 = 0.75$, to satisfy $\mathrm{tr}\,\Omega = 9$.

Prior and posterior. Samples from this prior emulator are shown in Figure 1, showing how the 7 outputs respond to x_2 (11 values across the range $[0, 1]$), with the other inputs held at their central values. Each of these samples (10 in all) has a (11×7)-dimensional Student-t distribution. Note that the roughness in each sampled path in x_2 is due to the treatment of the residual as a nugget: this roughness would have been smaller had I chosen $\mathrm{tr}\,\Omega$ larger, which I might have done had I chosen to include more regressors.

The prior emulator is updated with the ensemble of 30 evaluations. Samples from the updated emulator are given in Figure 2: this is comparable with Figure 1. My prior judgements about the three temperatures seem to have been well-founded.

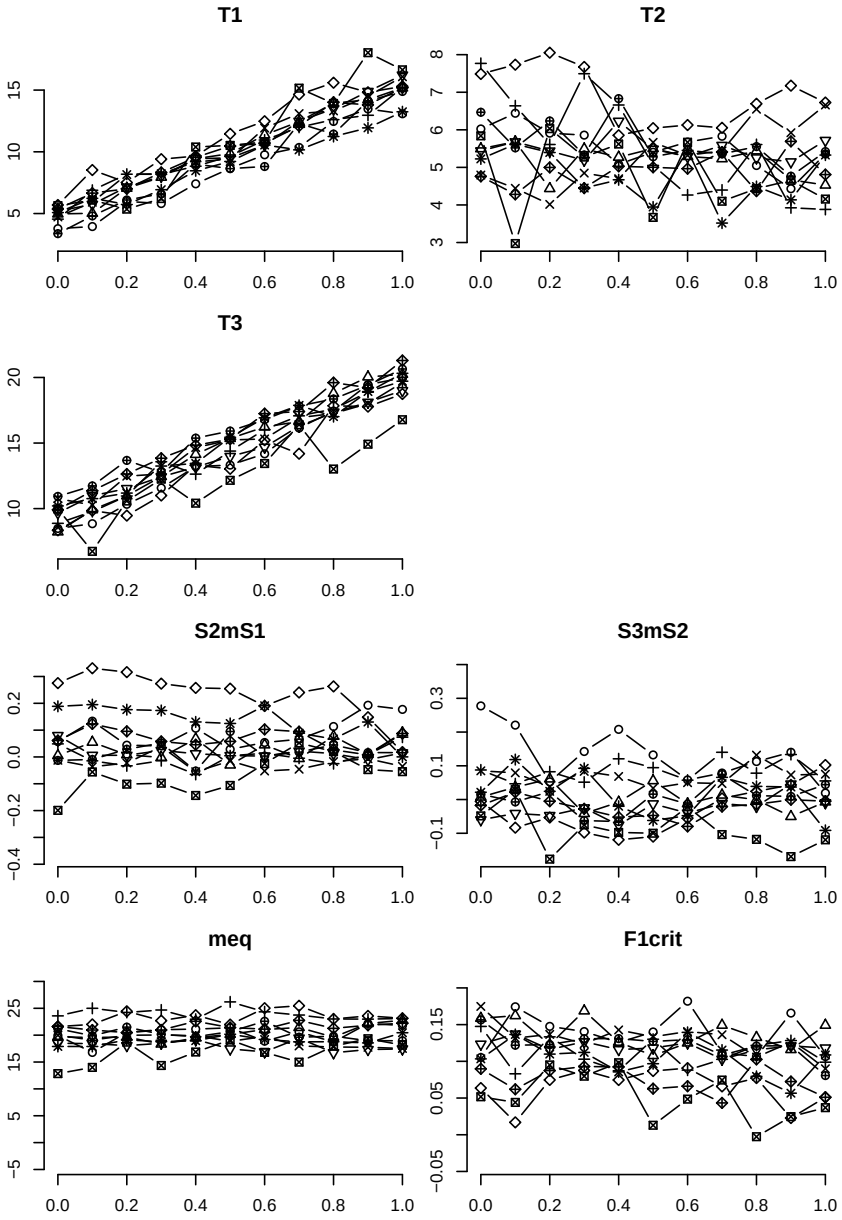


Figure 1: 10 samples from the prior emulator, for f as a function of x_2 with the other inputs set at their central values.

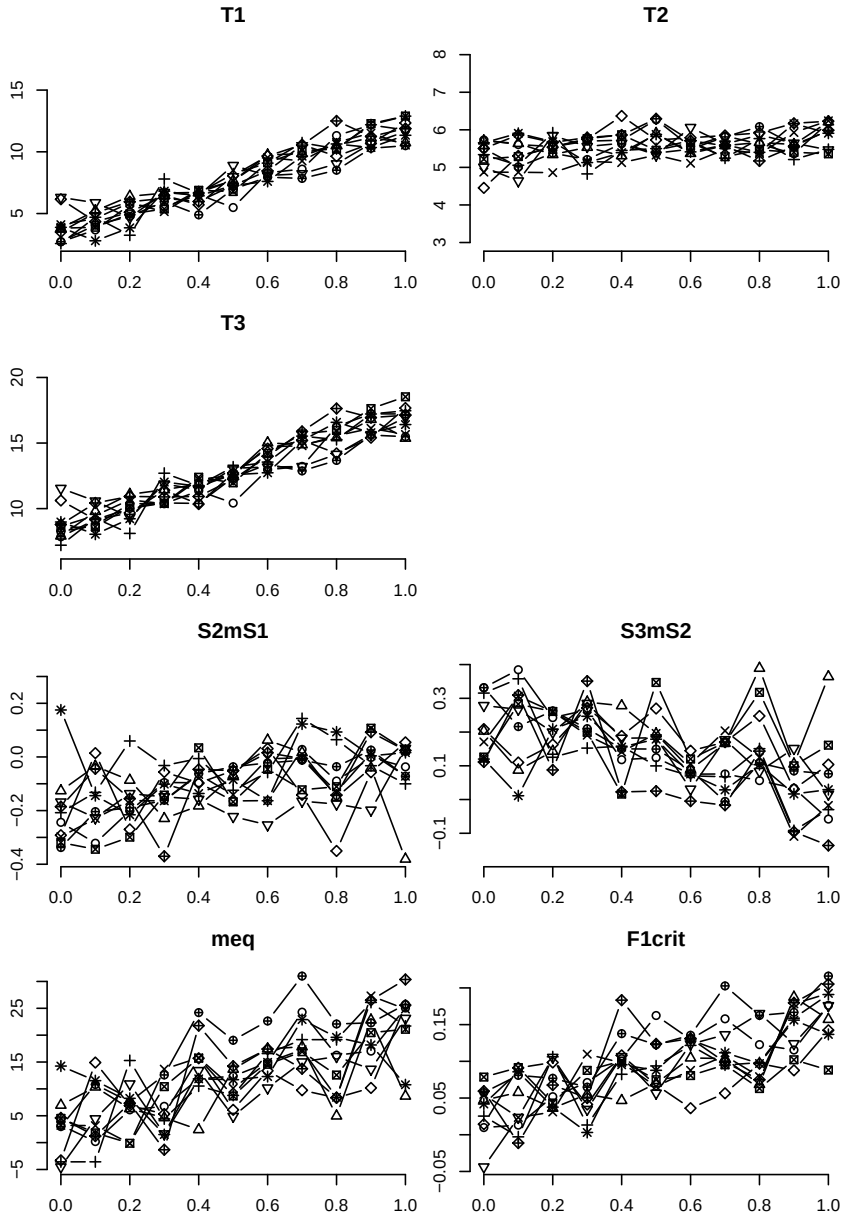


Figure 2: 10 samples from the updated emulator, comparable to those in Figure 1.

6.2 Diagnostics

Section 4 discusses diagnostics for the emulator. Figure 3 shows the leave-one-out prediction errors. This Figure also provides a clear picture of the impact of the separability of the emulator variance function (section 2.2). Treating the ensembles in each of the 30 cases as almost the same, the pattern of relative uncertainty is nearly duplicated across the 7 panels. For example, the ratio of the predictive uncertainty between evaluations 9 and 10 is always about 0.8, no matter what output.

Figure 4 shows the prequential diagnostics, based on one-step-ahead: while these are not terrible, neither are they very encouraging. To investigate the effect of my prior choices, the prior hyperparameters were tuned to maximise the marginal density of the evaluations. This tuning was not unconstrained: the ξ_i^2 quantities were permitted to vary by output type (four multipliers), the τ_j were permitted to vary by input type (three multipliers), and the trace of Ω was allowed to vary (one multiplier). The tuning process reduced the summed prequential scores from 157.0 to 41.1. The effect on the prequential diagnostics is shown in Figure 5; these are now acceptable (evaluation 21 has an extremely low value of x_5). The aberrant prequential diagnostics in Figure 4 would seem to have resulted from a disagreement between my prior and my likelihood, rather than the structural constraints of the LWE. For the remainder of this section, though, I continue with my original prior choices; in practice a reconsideration of my prior in the light of the tuned values might be advisable, possibly including some higher-order x_5 regressors.

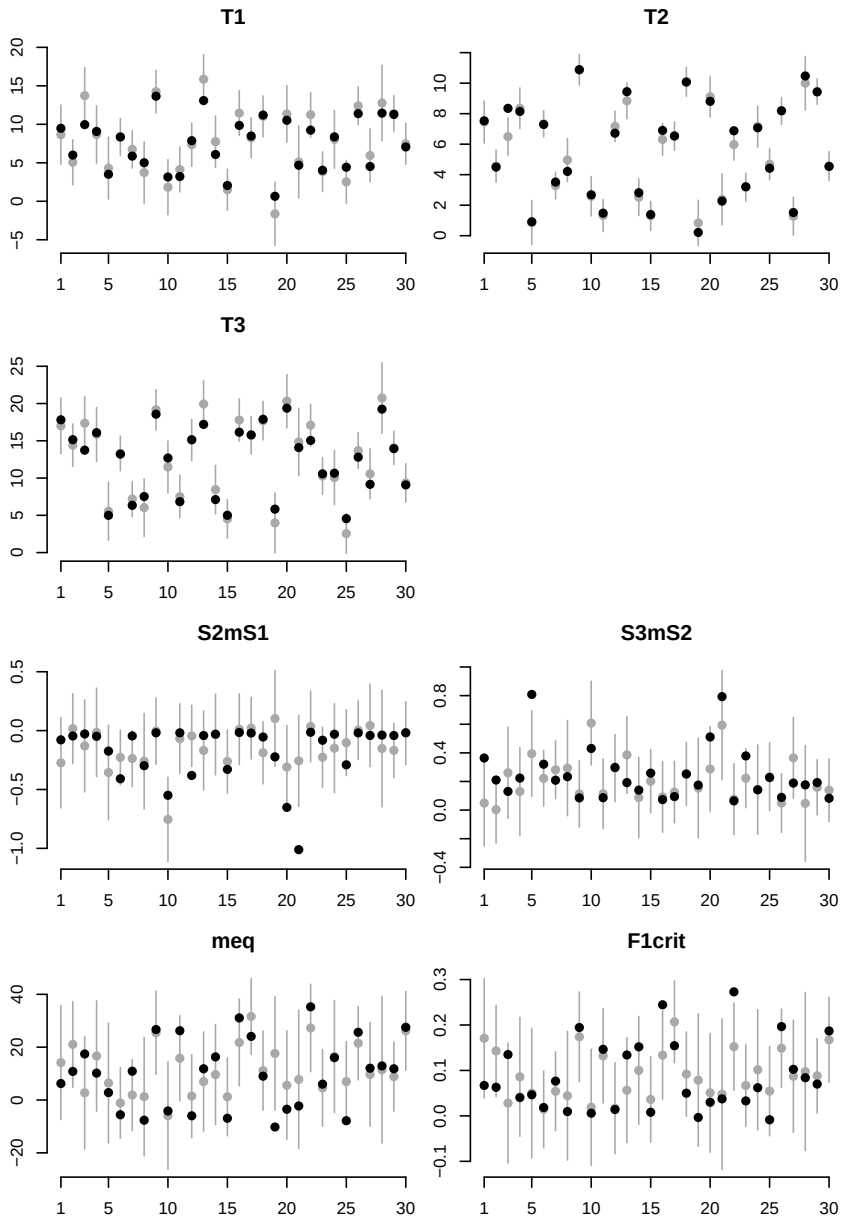


Figure 3: Leave-one-out prediction errors. The horizontal axis shows the evaluations. The grey bars and dots show the predicted 2.5th, 50th, and 97.5th percentiles using all but the indexed evaluation; the black dots show the actual values.

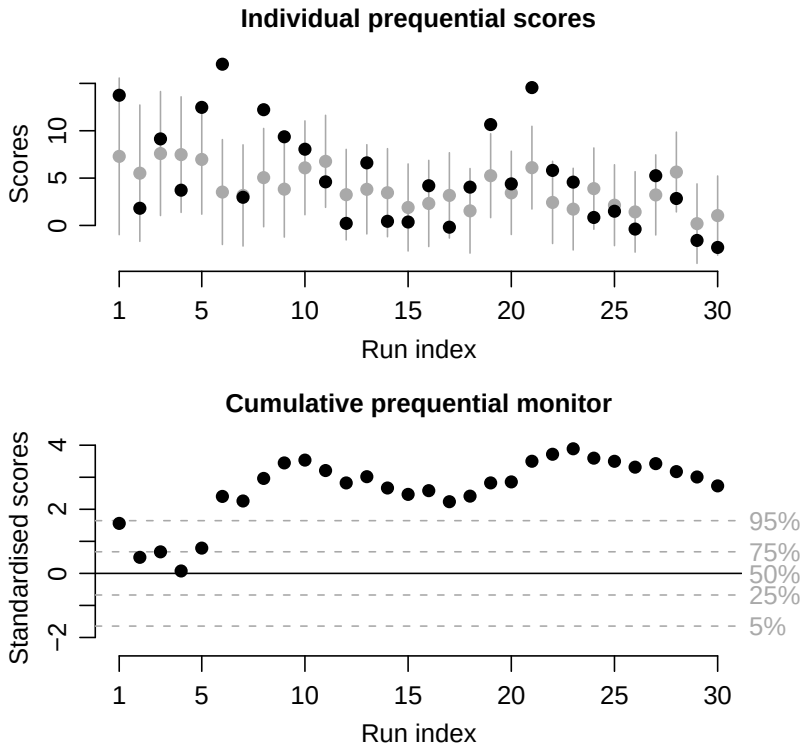


Figure 4: Prequential analysis, based on the one-step-ahead diagnostic. The prequential scores are shown with a bar indicating the mean $\pm$ two standard deviations, derived under the null hypothesis that the emulator is well-specified. The prequential monitor should asymptote (in n) to a standard Gaussian under the null-hypothesis.

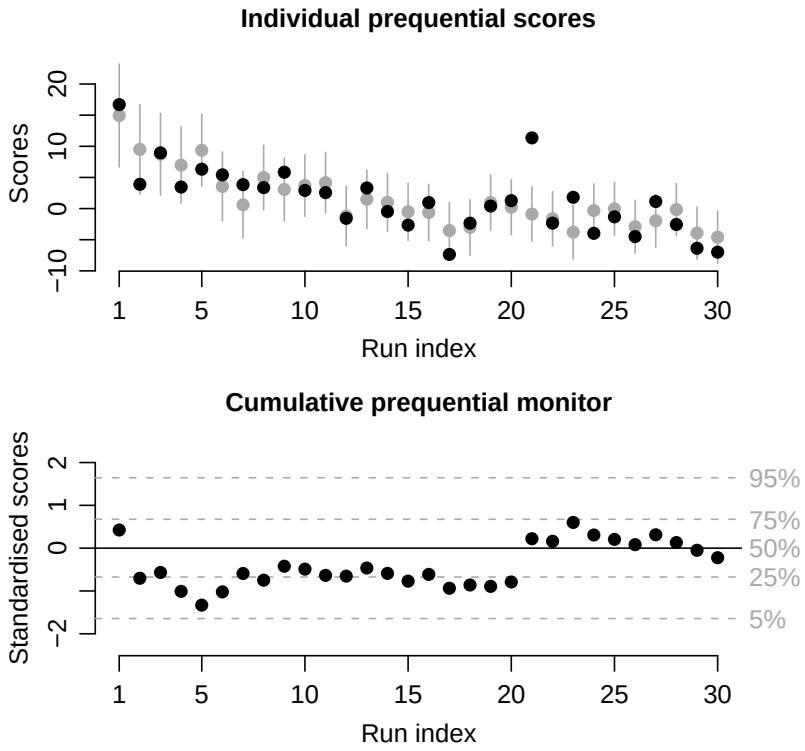


Figure 5: Prequential analysis, after tuning the prior hyperparameters to maximise the marginal density of the ensemble. See Figure 4 for details.

6.3 Summaries

Section 3.3 describes how we can summarise the emulator reponse to a subset of the inputs in terms of the conditional mean and variance, integrating out other inputs according to some distribution x^* as though they were nuisance parameters. I choose x^* to be uniform on $[0,1]^5$. Figure 6 shows the function responses to x_2, represented in this way. Note that the difference between Figure 2 and Figure 6 is that in the former the other four inputs are set to 0.5, while in the latter they are integrated out.

Figure 6 shows a clear difference between the three temperature outputs, `T1`, `T2`, and `T3`, and the other outputs, e,g., `F1crit`, conditionally on x_2. In the temperatures almost all of the variance is attributable to the mean function. In `F1crit` much of the variance is attributable to the variance function. We can infer that the mean functions of the temperatures are highly sensitive to the other four inputs, while the mean function of `F1crit` is not. In fact, we know this to be the case, because the temperatures depend strongly on x_1 and x_3 as well.

Figure 7 shows how x_2 and x_5 interact to determine one of the salinity differences, output 5, with the other three inputs integrated out.

Section 2.3 also discussed numerical summaries, such as 'R^2' values. These are shown in the top panel of Figure 8, plotted against increasing sets of regressors in the grevlex ordering. We can see that the three temperatures are well-explained by linear regressors alone with no interactions; i.e. $f_{1:3}(x)$ is effectively linear in x. The other outputs are poorly explained by linear terms alone; including higher-order terms helps, but not much. The bottom panel shows the absolute sizes of the coefficients, indicating the greater non-linearity of the responses of the non-temperature outputs. Within each degree-block, the ordering favours the non-temperature inputs; the preponderance of sizeable coefficients in the region starting from regressor 22 indicates that these inputs are the more important contributors to the non-linear responses.

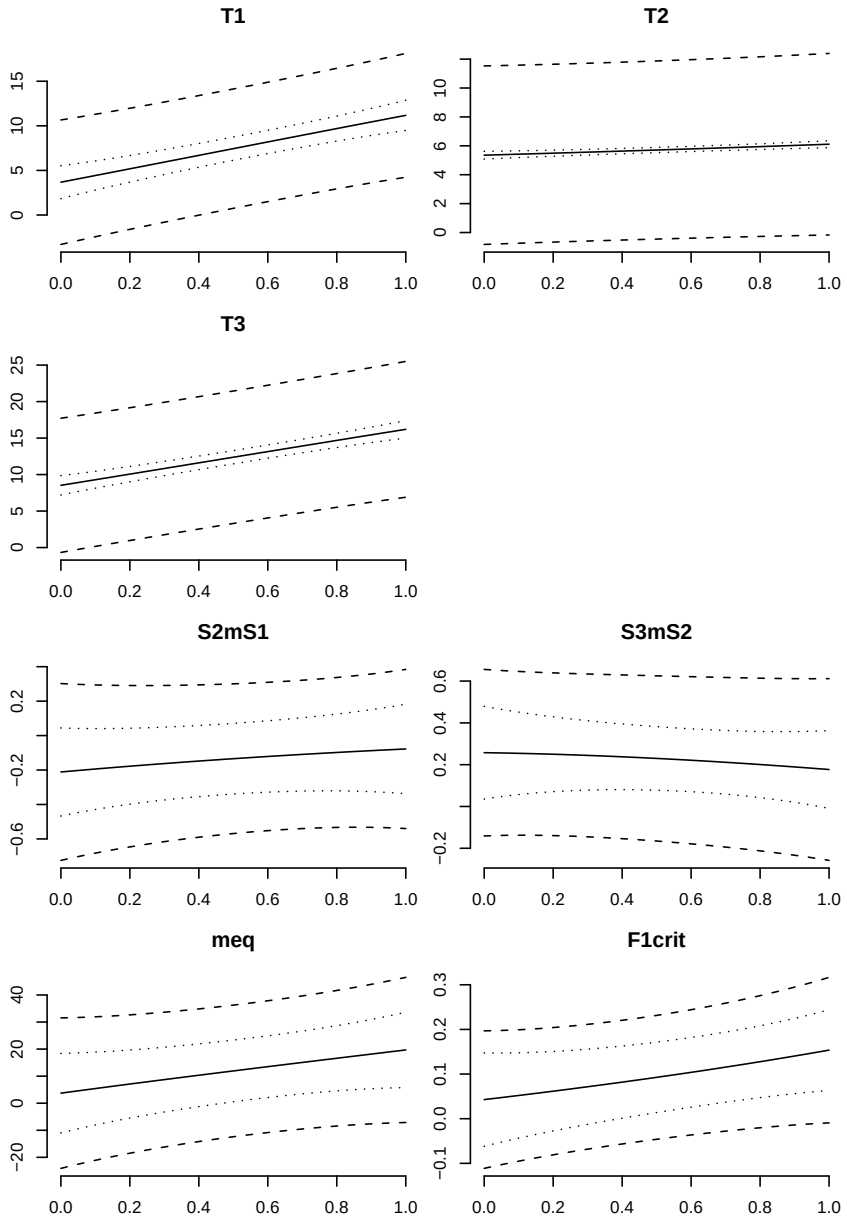


Figure 6: Mean and standard deviation of $f(x^*) \,|\, x_2$, where $x^* \sim U[0,1]^5$. The solid line shows the mean, the dashed interval is ± 2 standard deviations, the dotted interval indicates the proportion of total variance that is attributable to the expectation of the conditional variance.

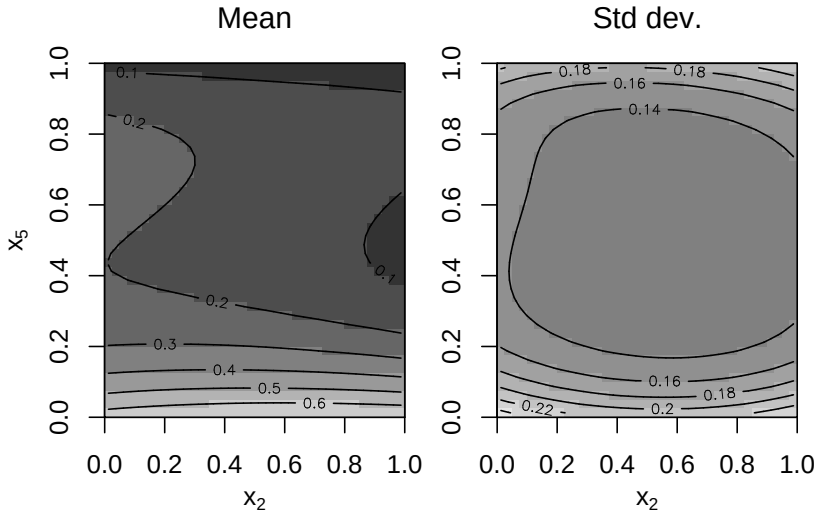


Figure 7: Mean and standard deviation of $f_5(x^*)|\{x_2, x_5\}$, where $x^* \sim U[0,1]^5$.

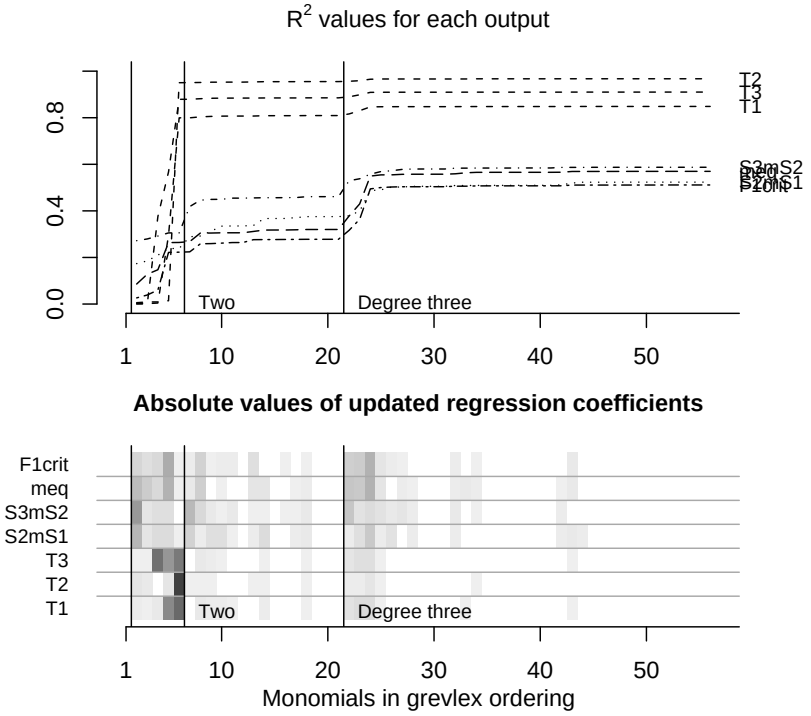


Figure 8: (1) Top panel. 'R^2' values for each of the outputs after updating, shown for increasing sets of monomials in grevlex ordering. (2) Bottom panel. Absolute values (darker signifies larger) of the updated expectations of the regression coefficients (excluding the intercept); same ordering. Each set of coefficients is rescaled to be unitless by dividing through by the range of each output in the ensemble.

7 Mixing over the hyperparameters

We would like to break out from the structural restrictions of the LWE with its MNIW structure to a more general emulator, but without sacrificing tractability. We would also like to account for difficulties in quantifying the prior hyperparameters, Ψ_0. The simplest way to achieve both of these goals is to mix over m different choices of Ψ_0, i.e. to generalise our prior to

$$\begin{aligned} \{B, \Sigma\} \mid \omega = i &\sim \text{MNIW}_k\left(\Psi_0^{(i)}\right) \\ \Pr(\omega = i) &= \lambda_i \qquad i = 1, \ldots, m \end{aligned} \tag{28}$$

where $\lambda \triangleq (\lambda_1, \ldots, \lambda_m)$ is a hyperparameter of probabilities.

The crucial quantity for computing the posterior mixture distribution is the marginal density of the ensemble, $\pi(F; X, \Psi_0)$, which is computed 'for free', along with the prequential diagnostics, see (20). After updating using the ensemble, the posterior emulator is a mixture of the emulators from each of the prior choices for Ψ_0, with the weights themselves updated to $\lambda_i \, \pi\big(F; X, \Psi_0^{(i)}\big)$ and renormalised so that $\sum_{i=1}^m \lambda_i = 1$. Thus with a mixture prior, we have an extended conjugate analysis over the set of hyperparameters $\big\{\Psi^{(1)}, \ldots, \Psi^{(m)}, \lambda\big\}$.

Berger and Berliner (1986) provide a different take on the same type of calculation, in which the prior hyperparameters may be partly or completely tuned using the ensemble. In the simplest case, this involves mixing over two choices of hyperparameters: the elicited prior $\Psi_0^{(1)} = \Psi_0$, and a data-determined prior $\Psi^{(2)} = \hat{\Psi}_0$, given in (23). In the extreme choice of $\lambda_1 = 0$, $\lambda_2 = 1$, and Γ unconstrained, this gives rise to the standard *parametric empirical Bayes* approach (see, e.g., Carlin and Louis, 2000), where the hyperparameters for the prior are determined by maximising the marginal density of the ensemble, and then plugged in. Berger and Berliner themselves suggest using a small value for λ_2, in order to 'robustify' the prior, although they are understandably cautious about whether this genuinely promotes robustness.

The mean and variance with the mixture prior are

$$\mathrm{E}\big(f(x)\big) = \bar{M}^T g(x) \tag{29a}$$

$$\begin{aligned}\mathrm{cov}\big(f(x), f(x')\big) &= \sum_{i=1}^{m} \lambda_i \left[M^{(i)} - \bar{M}\right]^T g(x) g(x')^T \left[M^{(i)} - \bar{M}\right] \\ &\quad + \sum_{i=1}^{m} \lambda_i \, \frac{w^{(i)}(x, x')}{\delta^{(i)} - 2} S^{(i)}\end{aligned} \tag{29b}$$

where

$$\bar{M} \triangleq \sum_{i=1}^{m} \lambda_i \, M^{(i)} \tag{29c}$$

$$w^{(i)}(x, x') \triangleq g(x)^T \Omega^{(i)} g(x') + \mathrm{dirac}(x - x') \tag{29d}$$

(cf. eq. (8)). By inspection of (29b) we can see that a separable variance function is only preserved when mixing if $M^{(i)} = M$ and $S^{(i)} \propto S$, for all i, in which case we have

$$\mathrm{cov}\big(f(x), f(x')\big) = \left[\sum_{i=1}^{m} \lambda_i \, c_i \, \frac{w^{(i)}(x, x')}{\delta^{(i)} - 2}\right] S \tag{30}$$

for some given scalars c_i satisfying $S^{(i)} = c_i \, S$. Therefore a mixture prior gives us an easy way to 'defeat' separability, if we so choose, and the degree to which we defeat it will depend on the range of values we select for $M^{(i)}$, and the way in which we allow the components of $S^{(i)}$ to vary independently of one another. Note, though, that all of this comes to naught if after updating we find that the λ values concentrate onto a single i; but at least in this case we have given non-separability a chance to emerge, should it want to.

For inference and diagnostics, it is incovenient that the mixture prior does not give rise to a readily-available predictive distribution (i.e., one that is already inplemented in standard statistical software). A simple alternative for prediction is to use sampling. Each of the m emulators is built and updated marginally, and then a draw is made from the predictive distribution after first selecting one of the emulators according to the probabilities λ, also updated. This allows us to compute leave-one-out and one-step-ahead diagnostics, with appropriate credible intervals under H_0. For the prequential diagnostics, the statistic S_m can still be calculated easily, as can its mean and variance un-

der H_0.

7.1 Illustration (cont)

As a simple illustration of the effect of a mixture prior, I consider a more general specification for ξ^2, which was defined in (13). In particular, I consider a range of candidate values for $\xi_{4:5}$, the value pertaining to the salinity differences, which I initially set at 0.2. I now use 21 values $\{\alpha_1\xi^2_{4:5}, \ldots, \alpha_{21}\xi^2_{4:5}\}$, where the $\log_4 \alpha_i$ are equally-spaced between -1 and 1, all with equal prior probability. After updating, the modal value for $\xi_{4:5}$ has increased to 0.5, which has posterior probability 0.29.

The leave-one-out diagnostic is shown in Figure 9. Choosing to mix over an element of ξ^2 defeats separability, because the candidate values for $S^{(i)}$ that are inferred from the various choices of $(\xi^2)^{(i)}$ are not proportional. This is evident in the Figure, which if closely inspected shows that the ratio of predictive uncertainties between different pairs of evalutions does vary by output (compare, e.g., evaluations 9 and 10 across the 7 panels). The prequential diagnostics are shown in Figure 10. Not surprisingly, the mixture prior here provides another route to a well-specified posterior emulator. I confess to choosing $\xi_{4:5}$ because the tuning discussed in section 6.2 indicated that my initial choice was low, relative to that inferred from the ensemble.

Note that these calculations took just a few seconds on a standard desktop computer: mixtures over hundreds of candidates would take perhaps a minute.

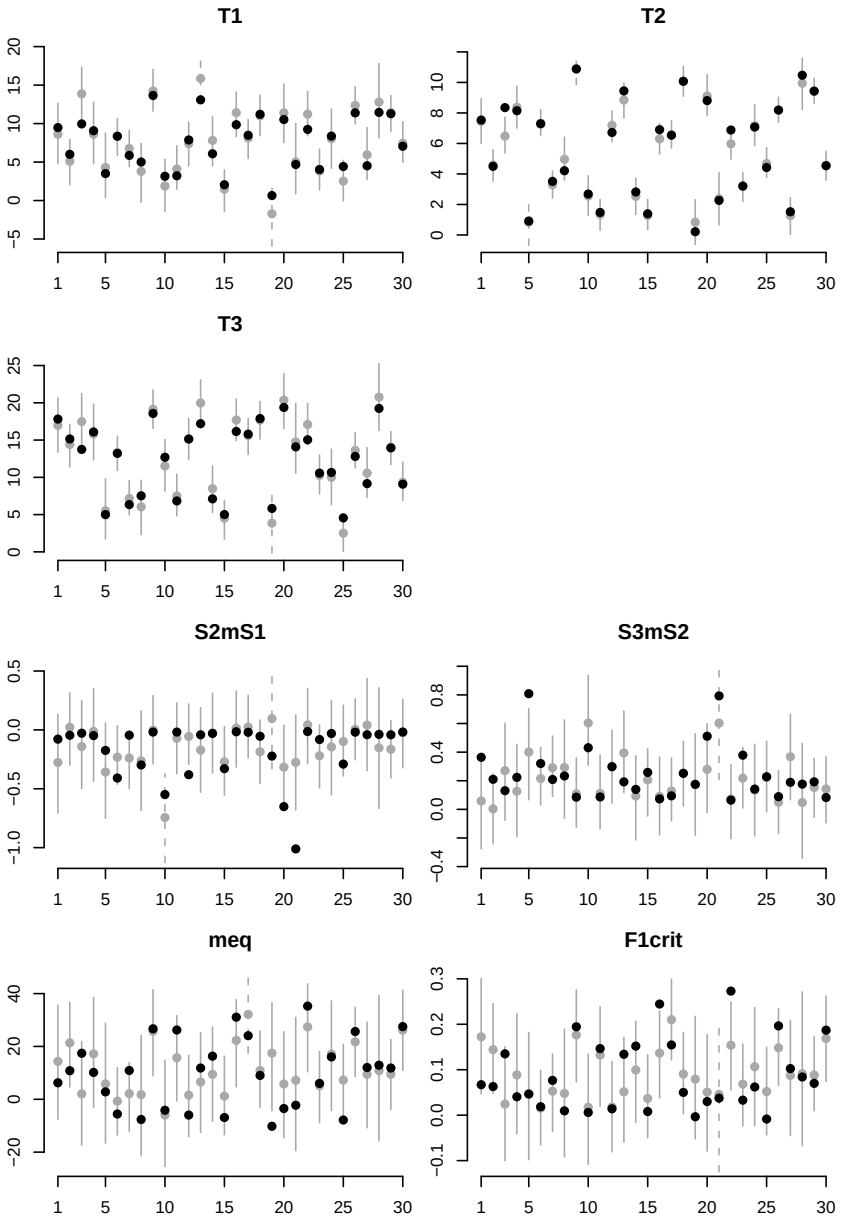


Figure 9: Leave-one-out diagnostic, after mixing over 21 candidate values for $\xi^2_{4:5}$. Shown on the same scale as Figure 3: the dashed lines indicate predictive intervals that exceed the vertical limits.

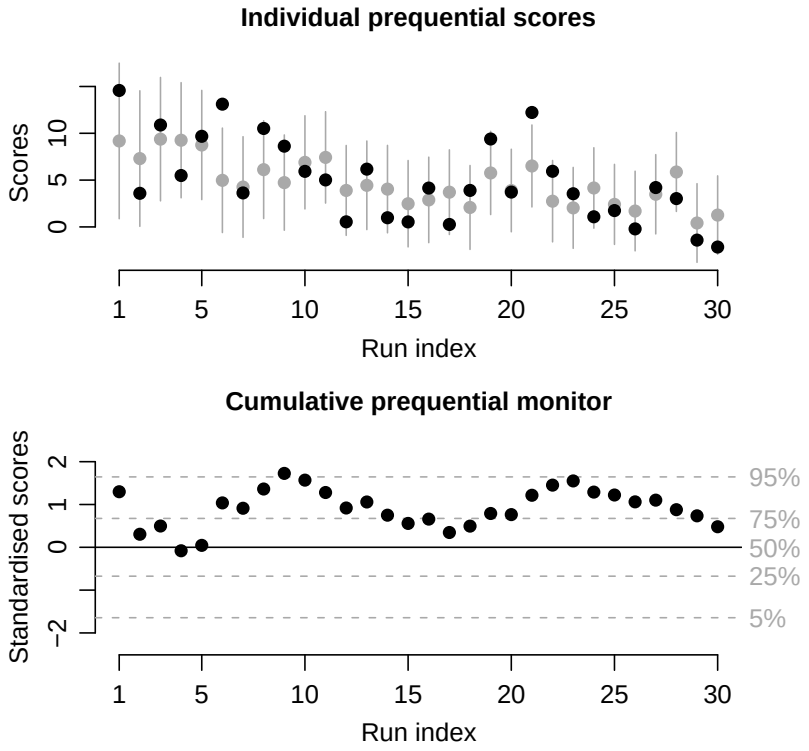


Figure 10: Prequential diagnostics, after mixing over 21 candidate values for $\xi^2_{4:5}$.

8 Conclusion

This paper advocates a fully-Bayesian treatment of multivariate regression as an appropriate framework for building multivariate emulators of complex deterministic functions, such as environmental models. This runs counter to current practice, which favours Smooth Gaussian Processes: I have addressed this directly in section 5. I hope I have also addressed this indirectly, by showing the wealth of interesting features that follow from adopting what I term a 'lightweight' framework. Much of what is proposed is 'standard' statistics, although some care must be taken to work out all the details. Software can largely automate the calculations, leaving the statistician and the model-expert to focus on prior choices, the diagnotic and interpretive information, and walking the fine line between Bayesian purity and aggressive tuning of the model structure (e.g., choice of regressors) and the prior hyperparameters.

If nothing else I hope this paper has raised the bar on emulator diagnostics: these are seldom discussed or displayed in papers that construct emulators. One trusts that they are always pored over during construction; anyhow, I would encourage the model-expert to start off highly sceptical, and to wait to be won-over by the statistician.

A Appendix

1 Notation for distributions

The origin of this paper's conjugate analysis is the remarkable article by Dawid (1981). Some clarification may be helpful, because of notational differences between authors (notably for the Inverse Wishart distribution), because Dawid's derivations are implicit, and because the key distribution, the Matrix Normal Inverse Wishart, has been little used in practice.

This Appendix can also be seen as a description of a fully-Bayesian treatment of Multivariate Regression, generalising the non-informative prior treatment in Box and Tiao (1973, ch. 8).

Inverse Wishart distribution. The $k \times k$ non-singular variance matrix Σ has an Inverse Wishart distribution $\Sigma \sim \mathrm{IW}_k(S, \delta)$ iff

$$\mathrm{p}(\Sigma) \propto |S|^{(\delta+k-1)/2} |\Sigma|^{-(k+\delta/2)} \exp\left\{-(1/2)\,\mathrm{tr}\, S\Psi\right\}, \qquad \Psi \triangleq \Sigma^{-1} \tag{A1}$$

(neglecting proportionate terms not involving S) where '$\triangleq$' denotes 'defined', and 'tr' denotes the trace, under the convention that 'tr' has lower priority than matrix multiplication. It follows that $\mathrm{E}(\Sigma) = S/(\delta - 2)$ for $\delta > 2$. Different authors have used different parameterisations of the IW: all four permutations of S or S^{-1}, and δ or $\nu \triangleq \delta + k - 1$ have been suggested. This reflects the origins of the IW in the Wishart distribution, for which S^{-1} and ν are the more natural parameters. In this paper the IW is the primitive distribution, and S may be referred to as the scale parameter, and δ as the degrees of freedom; the latter term is not consistently applied, with ν also being so-termed.

Normal Inverse Wishart distribution. The k-vector x has a Normal Inverse Wishart distribution $x \sim \mathrm{NIW}_k(\mu, w, S, \delta)$ iff $x \mid \Sigma \sim \mathrm{N}_k(\mu, w\Sigma)$ and $\Sigma \sim \mathrm{IW}_k(S, \delta)$. After integrating out Σ using (A1), the marginal distribution of x has a Student-t distribution:

$$\mathrm{p}(x) \propto \left[1 + \delta^{-1}(x-\mu)^T (wS/\delta)^{-1}(x-\mu)\right]^{-(\delta+k)/2} \tag{A2}$$

or $x \sim \mathrm{T}_k\left(\mu, wS/\delta, \delta\right)$, in the usual parameterisation of the Student-t distribution. It follows that $\mathrm{E}(x) = \mu$ and $\mathrm{var}(x) = wS/(\delta - 2)$, provided that $\delta > 1$ and $\delta > 2$, respectively.

Matrix Normal distribution. This distribution was introduced by Dawid (1981). Let $B \triangleq (b_{ij})$ be a $q \times k$ matrix, and denote by $(\cdot)^V$ the vector created by stacking the columns of a matrix; define $\boldsymbol{b} \triangleq B^V$. Then B has a Matrix Normal distribution $B \sim \mathrm{MN}_{q\times k}(M, \Omega, \Sigma)$ iff $\boldsymbol{b} \sim \mathrm{N}_{qk}(\boldsymbol{m}, \Sigma \otimes \Omega)$, where the matrix M is $q \times k$ and $\boldsymbol{m} \triangleq M^V$, and Ω and Σ are variance matrices with dimensions $q \times q$ and $k \times k$. It is straightforward to show that

$$\mathrm{p}(B) \propto |\Sigma|^{-q/2} \exp\left\{-(1/2)\,\mathrm{tr}(B-M)^T\Omega^{-1}(B-M)\Psi\right\},$$

(neglecting proportionate terms not involving Σ) where $\Psi \triangleq \Sigma^{-1}$, as before.

Proof. Start from the density function of $\boldsymbol{b}$. Outside the exponent we have the scalar $|2\pi\,\Sigma \otimes \Omega|^{-1/2}$, which can be written $(2\pi)^{-qk/2}|\Omega|^{-k/2}|\Sigma|^{-q/2} \propto |\Sigma|^{-q/2}$, as required. The quadratic form inside the exponent has the form (taking $M = \mathbf{0}$, for simplicity)

$$\begin{aligned}\boldsymbol{b}^T(\Sigma \otimes \Omega)^{-1}\boldsymbol{b} &= \boldsymbol{b}^T(\Psi \otimes \Omega^{-1})\boldsymbol{b} \\ &= \boldsymbol{b}^T\left(\Omega^{-1}B\Psi\right)^V \\ &= \mathrm{tr}\, B^T\Omega^{-1}B\Psi\end{aligned}$$

as required, using the general relation $(XYZ)^V = (Z^T \otimes X)Y^V$. □

Matrix Normal Inverse Wishart distribution. This matrix generalisation of the NIW was suggested by Dawid (1981), see also Press (1982, sec. 8.6) and West and Harrison (1997, sec. 16.4). The set of parameters $\{B, \Sigma\}$ has a Matrix Normal Inverse Wishart distribution $\{B, \Sigma\} \sim \mathrm{MNIW}_{q\times k}(M, \Omega, S, \delta)$ iff $B \mid \Sigma \sim \mathrm{MN}_{q\times k}(M, \Omega, \Sigma)$ and $\Sigma \sim \mathrm{IW}_k(S, \delta)$. Note how Ω in the MNIW plays the role of the scalar w in the NIW distribution. For later reference,

$$\mathrm{p}(B, \Sigma) \propto |\Sigma|^{-[k+(\delta+q)/2]} \exp\left\{-(1/2)\,\mathrm{tr}\left[(B-M)^T\Omega^{-1}(B-M) + S\right]\Psi\right\}, \tag{A3}$$

$\Psi \triangleq \Sigma^{-1}$, as before.

2 The emulator

We focus on the marginal emulator, in particular the mean and variance of the k-vector $f(x)$; generalisation to an arbitrary finite set $\{f_i(x), f_{i'}(x'), \ldots\}$

is straightforward. $\{B, \Sigma\}$ has a MNIW distribution, in which the hyper-parameters will depend on the ensemble. For simplicity we take the hyper-parameters in their prior form, so that $\{B, \Sigma\} \sim \text{MNIW}_{q\times k}(M, \Omega, S, \delta)$. If we have updated the emulator using the ensemble $(F; X)$, then we also assume that $x \notin \{X_1, \ldots, X_n\}$, to avoid the trivial case of predicting an output that we know exactly.

Conditional on Σ, the emulator is Gaussian. This follows because $B \mid \Sigma$ and $u(x) \mid \Sigma$ are independent and marginally Gaussian, hence jointly Gaussian, and thus $f(x)^T \mid \Sigma$ is Gaussian, from (1):

$$f(x)^T \mid \Sigma \sim \text{N}_k\left(g(x)^T M, w(x)\Sigma\right) \tag{A4}$$

where $w(x) \triangleq g(x)^T \Omega g(x) + 1$.

Proof. The mean function is straightforward. To derive the expression for $w(x)$, note that

$$\text{var}\big(f(x)^T \mid \Sigma\big) = \text{var}\big(g(x)^T B \mid \Sigma\big) + \text{var}\big(u(x) \mid \Sigma\big),$$

because $B \perp\!\!\!\perp u(x) | \Sigma$. The second term contributes Σ, or the '+1' in $w(x)$. For first term, $g(x)^T B = \left[\big(g(x)^T B\big)^V\right]^T$. Ignoring the outside transpose because $\text{var}(x) = \text{var}(x^T)$,

$$\begin{aligned}
\text{var}(g(x)^T B \mid \Sigma) &= \text{var}\Big(\big(g(x)^T B\big)^V \mid \Sigma\Big) \\
&= \text{var}\big(\big(I_k \otimes g(x)^T\big)\boldsymbol{b} \mid \Sigma\big) \\
&= \big(I_k \otimes g(x)^T\big)\big(\Sigma \otimes \Omega\big)\big(I_k \otimes g(x)\big) \\
&= \Sigma \otimes \left[g(x)^T \Omega g(x)\right] \\
&= \left[g(x)^T \Omega g(x)\right] \Sigma
\end{aligned}$$

as the term in brackets is a scalar. □

As $\Sigma \sim \text{IW}_k(S, \delta)$, so the unconditional distribution of $f(x)$ is Multivariate Student-t:

$$f(x)^T \sim \text{T}_k\left(g(x)^T M, w(x)S/\delta, \delta\right), \tag{A5}$$

and the mean and variance are

$$\mathrm{E}\big(f(x)^T\big) = g(x)^T M \tag{A6a}$$

$$\mathrm{var}\big(f(x)^T\big) = w(x)S/(\delta - 2), \tag{A6b}$$

provided that $\delta > 1$ and $\delta > 2$, respectively.

3 Conjugate analysis

The likehood function for $\{B, \Sigma\}$ using the ensemble $(F; X)$ is

$$\mathrm{Lik}(B, \Sigma) \propto |\Sigma|^{-n/2} \exp\left\{-(1/2)\,\mathrm{tr}\, U^T U \Psi\right\} \tag{A7}$$

where $U \triangleq F - GB$ and $G_{vj} \triangleq g_j(X_v)$, the $n \times q$ matrix of regressor values. This form for the likelihood is a consequence of the treatment of $u(x) \mid \Sigma$ as a Gaussian nugget, or, in more common parlance, as a standard regression-type residual.

The prior for $\{B, \Sigma\}$ is $\mathrm{MNIW}_{q\times k}\,(M, \Omega, S, \delta)$, where $\{M, \Omega, S, \delta\}$ are treated as hyperparameters.

In the updated distribution for $\{B, \Sigma\}$, the scalar becomes proportional to

$$|\Sigma|^{-[k+(n+\delta+q)/2]} \text{ or } |\Sigma|^{-[k+(\hat{\delta}+q)/2]} \tag{A8}$$

where $\hat{\delta} \triangleq \delta + n$. Inside the exponential in the posterior we have, temporarily discounting the factor of $-(1/2)$,

$$\mathrm{tr}\,[(F - GB)^T(F - GB) + (M' - AB)^T(M' - AB) + S]\Psi \tag{A9}$$

where A is defined as the matrix square-root of Ω^{-1}, so that $A^T A \equiv \Omega^{-1}$, and $M' \triangleq AM$. We can complete the square for B on the condition that $G^T G + \Omega^{-1}$ is non-singular, in which case define

$$\hat{\Omega} \triangleq (G^T G + \Omega^{-1})^{-1} \tag{A10a}$$

$$\hat{M} \triangleq \hat{\Omega}(G^T F + \Omega^{-1} M), \tag{A10b}$$

and (A9) becomes

$$\mathrm{tr}\left[(B - \hat{M})^T \hat{\Omega}^{-1} (B - \hat{M}) + \hat{S}\right]\Psi \tag{A11}$$

where

$$\hat{S} \triangleq S - \hat{M}^T\hat{\Omega}^{-1}\hat{M} + F^TF + M^T\Omega^{-1}M. \tag{A12}$$

Comparing (A8) and (A11) with (A3), the updated distribution is

$$B, \Sigma \mid (F; X) \sim \text{MNIW}_{q\times k}\left(\hat{M}, \hat{\Omega}, \hat{S}, \hat{\delta}\right), \tag{A13}$$

provided that $|\hat{\Omega}| > 0$. One simple way to check that these calculations have been implemented correctly is to make sure that they satisfy sequential updating under different orderings of the evaluations.

Comparison to ML estimators. Certain prior choices of the hyperparameters will result in the updated expectation of B and Σ being the same as the Maximum Likelihood (ML) estimators; these choices are stated in eq. (9). The ML estimators are

$$B_{\text{ML}} \triangleq (G^TG)^{-1}G^TF \tag{A14}$$

$$\Sigma_{\text{ML}} \triangleq n^{-1}(U_{\text{ML}})^TU_{\text{ML}} = n^{-1}F^TPF \tag{A15}$$

where $U_{\text{ML}} \triangleq F - GB_{\text{ML}}$, and $P \triangleq I - G(G^TG)^{-1}G^T$, the projection matrix (Mardia *et al.*, 1979, sec. 6.2.1). For $B_{\text{ML}} = \text{E}(B) = \hat{M}$ we require $\Omega^{-1} = \mathbf{0}$ in (A10). Taking $\Omega^{-1} = \mathbf{0}$ and also $S = \mathbf{0}$ in (A12) we have

$$\begin{aligned}\hat{S} &= F^TF - (B_{\text{ML}})^T(G^TG)B_{\text{ML}} \\ &= F^T(I - G(G^TG)^{-1}(G^TG)(G^TG)^{-1}G^T)F \\ &= F^TPF. \end{aligned} \tag{A16}$$

Then for $\Sigma_{\text{ML}} = \text{E}(\Sigma) = \hat{S}/(\hat{\delta} - 2)$ we must also have $\hat{\delta} - 2 = n$, or $\delta = 2$.

References

J. An and A.B. Owen, 2001. Quasi-regression. *Journal of Computing*, **17**, 588–607.

S. Banerjee, B.P. Carlin, and A.E. Gelfand, 2004. *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton, FLorida: Chapman & Hall/CRC.

J. Berger and L.M. Berliner, 1986. Robust Bayes and Empirical Bayes analysis with ϵ-contaminated priors. *Annals of Statistics*, **14**, 461–486.

J.M. Bernardo and A.F.M. Smith, 1994. *Bayesian Theory*. Chichester, UK: Wiley.

J. Besag, 1975. Statistical analysis of non-lattice data. *The Statistician*, **24** (3), 179–195.

G.E.P. Box and G.C. Tiao, 1973. *Bayesian Inference in Statistical Analysis*. Reading, Massachusetts: Addison-Wesley.

B. Carlin and T.A. Louis, 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, Florida: Chapman & Hall/CRC, 2nd edition.

D. Cornford, L. Csató, D.J. Evans, and M. Opper, 2004. Bayesian analysis of the scatterometer wind retrieval inverse problem: Some new approaches. *Journal of the Royal Statistical Society, Series B*, **66**(3), 609–626.

R.G. Cowell, A.P. David, S.L. Lauritzen, and D.J. Spiegelhalter, 1999. *Probabilistic Networks and Expert Systems*. New York: Springer.

P.S. Craig, M. Goldstein, J.C. Rougier, and A.H. Seheult, 2001. Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, **96**, 717–729.

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1996. Bayes linear strategies for matching hydrocarbon reservoir history. In J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*, pages 69–95. Oxford: Clarendon Press.

P.S. Craig, M. Goldstein, A.H. Seheult, and J.A. Smith, 1997. Pressure matching for hydrocarbon reservoirs: A case study in the use of Bayes Linear strategies for large computer experiments. In C. Gatsonis, J.S. Hodges, R.E. Kass, R. McCulloch, P. Rossi, and N.D. Singpurwalla, editors, *Case Studies in Bayesian Statistics III*, pages 37–87. New York: Springer-Verlag. With discussion.

N.A.C. Cressie, 1991. *Statistics for Spatial Data*. New York: John Wiley & Sons.

C. Currin, T.J. Mitchell, M. Morris, and D. Ylvisaker, 1991. Bayesian prediction of deterministic functions, with application to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953–963.

A.P. Dawid, 1981. Some matrix-variate distribution theory: Notational considerations and a Bayesian application. *Biometrika*, **68**(1), 265–274.

A.P. Dawid, 1984. Statistical theory: The prequential approach. *Journal of the Royal Statistical Society, Series A*, **147**(2), 278–290. With discussion, pp. 290–292.

A.E. Gelfand and D. Dey, 1994. Bayesian model choice: Asymptotic and exact calculations. *Journal Royal Statistical Society B*, **56**, 501–514.

M. Goldstein and J.C. Rougier, 2004. Probabilistic formulations for transferring inferences from mathematical models to physical systems. *SIAM Journal on Scientific Computing*, **26**(2), 467–487.

M. Goldstein and J.C. Rougier, 2006a. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*. Forthcoming, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/BLCP.pdf`.

M. Goldstein and J.C. Rougier, 2006b. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*. Forthcoming as a discussion paper, currently available at `http://www.maths.dur.ac.uk/stats/people/jcr/Reify.pdf`.

G.H. Golub and C.F. Van Loan, 1989. *Matrix Computations*. Baltimore: John Hopkins University Press, 2nd edition.

R. Haylock and A. O'Hagan, 1996. On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In J.M. Bernado, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 5*, pages 629–637. Oxford, UK: Oxford University Press.

D. Higdon, M.C. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne, 2004. Combining field data and computer simulations for calibration and prediction. *SIAM Journal on Scientific Computing*, **26**(2), 448–466.

R.H. Jones and A.V. Vecchia, 1993. Fitting continuous ARMA models to unequally spaced data. *Journal of the American Statistical Association*, **88**, 947–954.

M.C. Kennedy and A. O'Hagan, 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, **63**, 425–464. With discussion.

J.R. Koehler and A.B. Owen, 1996. Computer experiments. In S. Ghosh and C.R. Rao, editors, *Handbook of Statistics, 13: Design and Analysis of Experiments*, pages 261–308. North-Holland: Amsterdam.

K.V. Mardia, J.T. Kent, and J.M. Bibby, 1979. *Multivariate Analysis*. London: Harcourt Brace & Co.

J.M. Murphy, D.M.H. Sexton, D.N. Barnett, G.S. Jones, M.J. Webb, M. Collins, and D.A. Stainforth, 2004. Quantification of modelling uncertainties in a large ensemble of climate change simulations. *Nature*, **430**, 768–772.

J.E. Oakley and A. O'Hagan, 2002. Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika*, **89**(4), 769–784.

J.E. Oakley and A. O'Hagan, 2004. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, **66**, 751–769.

A. O'Hagan, 1998. A Markov property for covariance structures. Unpublished, available at `http://www.shef.ac.uk/~st1ao/ps/kron.ps`.

A. O'Hagan, 2006. Bayesian analysis of computer code outputs: A tutorial. forthcoming in Reliability Engineering and System Safety, currently available at `http://www.tonyohagan.co.uk/academic/pdf/BACCO-tutorial.pdf`.

A. O'Hagan, M.C. Kennedy, and J.E. Oakley, 1999. Uncertainty analysis and other inferential tools for complex computer codes. In J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, editors, *Bayesian Statistics 6*, pages 503–519. Oxford University Press. With discussion, pp. 520–524.

A. O'Hagan and J. Oakley, 2004. Probability is perfect, but we can't elicit it perfectly. *Reliability Engineering and System Safety*, **85**, 239–248.

A. Owen, J. Koehler, and S. Sharifzadeh, 1989. Comment on "Design and analysis of computer experiments" by Sacks *et al. Statistical Science*, **4**(4), 429–430.

G. Pistone, E. Riccomagno, and H.P. Wynn, 2000. *Algebraic Statistics: Computational Commutative Algebra in Statistics*. Chapman & Hall/CRC.

S.J. Press, 1982. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*. Malabar, Florida: Robert L. Krieger, 2nd edition.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. ISBN 3-900051-00-3, `http://www.R-project.org`.

M. Rosenblatt, 1952. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, **23**, 470–472.

J.C. Rougier, 2001. Comment on the paper by Kennedy and O'Hagan. *Journal of the Royal Statistical Society, Series B*, **63**, 453.

J.C. Rougier, 2007. Probabilistic inference for future climate using an ensemble of climate model evaluations. forthcoming in Climatic Change.

J.C. Rougier, D.M.H. Sexton, J.M. Murphy, and D. Stainforth, 2006. Emulating the HadSM3 simulator using enembles from different but related experiments. Submitted to the *Journal of Climate*, available at `http://www.maths.dur.ac.uk/stats/people/jcr/hadsm3sens.pdf`.

J. Sacks, W.J. Welch, T.J. Mitchell, and H.P. Wynn, 1989. Design and analysis of computer experiments. *Statistical Science*, **4**(4), 409–423. With discussion, pp. 423–435.

S.K. Sahu and K.V. Mardia. Recent trends in modeling spatio-temporal data. In *Proceedings of the special meeting on Statistics and Environment*, pages 69–83. Società Italiana di Statistica, Università Di Messina, Messina, Italy, 2005.

T.J. Santner, B.J. Williams, and W.I. Notz, 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

F. Seillier-Moiseiwitsch and A.P. David, 1993. On testing the validity of sequential probability forecasts. *Journal of the American Statistical Association*, **88**, 355–359.

M. West and J. Harrison, 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer-Verlag, 2nd edition.

B. Williams, D. Higdon, J. Gatticker, L. Moore, M. McKay, and S. Keller-McNulty, 2006. Combining experimental data and computer simulations, with an application to flyer plate experiments. *Bayesian Analysis*, **1**(4), 765–792.

K. Zickfeld, T. Slawig, and S. Rahmstorf, 2004. A low-order model for the response of the Atlantic thermohaline circulation to climate change. *Ocean Dynamics*, **54**(1), 8–26.