Expert Knowledge and Multivariate Emulation: The Thermosphere–Ionosphere Electrodynamics General Circulation Model (TIE-GCM)

Jonathan ROUGIER

Department of Mathematics
University of Bristol
University Walk
Bristol BS8 1TW, U.K.
(*j.c.rougier@bristol.ac.uk*)

Serge GUILLAS

Department of Statistical Science
University College London
London WC1E 6BT, U.K.

Astrid MAUTE

High Altitude Observatory
The National Center for Atmospheric Research
Boulder, CO 80307-3000

Arthur D. RICHMOND

High Altitude Observatory
The National Center for Atmospheric Research
Boulder, CO 80307-3000

The thermosphere–ionosphere electrodynamics general circulation model (TIE-GCM) of the upper atmosphere has a number of features that are a challenge to standard approaches to emulation, including a long run time, multivariate output, periodicity, and strong constraints on the interrelationship between inputs and outputs. These kinds of features are not unusual in models of complex systems. We show how they can be handled in an emulator and demonstrate the use of the outer product emulator for efficient calculation, with an emphasis on predictive diagnostics for model choice and model validation. We use our emulator to "verify" the underlying computer code and to quantify our qualitative physical understanding.

KEY WORDS: Gaussian process; Outer product emulator; Predictive diagnostics.

1. INTRODUCTION

An emulator is a stochastic representation of a deterministic simulator (typically implemented as computer code) deployed in situations where the simulator is expensive to evaluate. Emulators are a very useful tool for understanding simulators and also in inferences that combine simulator evaluations with system observations to make system predictions. O'Hagan (2006) has provided an introduction to emulators, and more details have been given by Santner, Williams, and Notz (2003, chaps. 3 and 4). Kennedy and O'Hagan (2001), Craig et al. (2001), and Goldstein and Rougier (2006) have described two Bayesian approaches to emulator-based system inference. Oakley and O'Hagan (2002) described the use of emulators for uncertainty analysis, and Rougier and Sexton (2007) contrasted uncertainty analysis for a climate simulator without and with an emulator. Oakley and O'Hagan (2004) described the use of emulators for sensitivity analysis; "screening" for active variables is a variant of this (see, e.g., Linkletter et al. 2006 and references therein). Goldstein and Rougier (2004, 2009) discussed the role of emulators in a general framework for linking model evaluations and system behavior. Sansó, Forest, and Zantedeschi (2008) provided a recent application of emulation in climate prediction, with a discussion that considers some of the foundational and practical issues that arise.

This article develops two themes in parallel: a case study on how to introduce expert knowledge about the simulator into the statistical choices that make up the emulator, along with a "lightweight" approach to multivariate emulation that prioritizes efficient emulators, enabling a detailed analysis of predictive diagnostics. On the latter theme, this is the first work to demonstrate the outer product emulator (OPE) approach (Rougier 2008) to multivariate emulation; other approaches to multivariate emulation have been suggested by Drignei (2006), Conti and O'Hagan (2007), Higdon et al. (2008), and Liu and West (2009). Section 2 describes the standard approach to emulation and the ways in which this can be modified to include expert knowledge. Sections 3 and 4 describe the TIE-GCM simulator and the OPE, respectively. Section 5 describes the choices that we make when building our emulator for the TIE-GCM simulator, as well as ways to produce and use diagnostic information to inform our choices. Section 6 demonstrates the use of the emulator to "verify" the simulator code and better understand the simulator. Section 7 concludes.

TECHNOMETRICS, NOVEMBER 2009, VOL. 51, NO. 4
DOI 10.1198/TECH.2009.07123

2. APPROACHES TO EMULATION

Most emulators (scalar or multivariate) can be understood within the following general framework:

$$f_i(r) = \sum_{j=1}^{v} \beta_j g_j(r, s_i) + \epsilon(r, s_i), \tag{1}$$

where the left side is the ith simulator output at simulator input r (r for "run") and the right side is the sum of a set of regressors with unknown coefficients, together denoted by $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_v)$, and a residual stochastic process, $\epsilon(\cdot)$. The output index, i, is assumed to map to points in some domain, $s_i, \in \mathcal{S}$, which may be continuous or discrete, or a mixture. For example, if $f(\cdot)$ is a climate simulator, then s_i might be a triple of variable type (discrete), location, and time (both notionally continuous). For a scalar emulator, $\mathcal{S}$ is simply an atom, and s_i can be neglected. Often it is easier to write $x \equiv \{r, s\}$, but in multivariate emulators it is important to distinguish between simulator input, r, and the simulator output index, s_i.

The prior emulator is completed by a choice for the distribution of $\theta \triangleq \{\boldsymbol{\beta}, \epsilon(\cdot)\}$, typically by assigning a parametric family to θ and then specifying prior values for the parameters. The updated emulator is then found by conditioning θ on data from simulator evaluations. Finally, (1) is used to infer the joint distribution of simulator evaluations over any collection of (r, s_i) tuples. The standard choice for the distribution of θ is normal inverse gamma, for tractability:

$$\boldsymbol{\beta} \perp\!\!\!\perp \epsilon(\cdot) | \tau, \Psi, \tag{2a}$$

$$\boldsymbol{\beta} | \tau, \Psi \sim \mathrm{N}(m, \tau V), \tag{2b}$$

$$\epsilon | \tau, \Psi \sim \mathrm{GP}(0, \tau \kappa(\cdot)), \tag{2c}$$

$$\tau | \Psi \sim \mathrm{IG}(a, d), \tag{2d}$$

where $\Psi \equiv \{m, V, a, d, \kappa(\cdot)\}$ is the set of hyperparameters. Here N denotes a Gaussian distribution, GP denotes a Gaussian process, $\kappa(\cdot)$ is a covariance function defined on $x \times x$, and IG denotes an inverse gamma distribution. With this choice, $\theta = \{\boldsymbol{\beta}, \tau, \epsilon(\cdot)\}$.

The standard fully probabilistic approach to emulation, as exemplified by Kennedy and O'Hagan (2001) and widely adopted, makes the following choices:

1. For the regressors, $g_j(\cdot)$, a constant and linear terms in each component of x, sometimes just a constant.
2. For the residual covariance function, $\kappa(\cdot)$, the product of squared exponential correlation functions in each component of x, with the correlation length vector λ added to the hyperparameters (see point 4).
3. Vague, often improper choices for $\{m, V, a, d\}$.
4. Residual correlation length vector λ fitted by maximizing the marginal likelihood and plugged in. Other fitting approaches (e.g., restricted maximum likelihood, cross-validation) are advocated as well (see, e.g., Santner, Williams, and Notz 2003, sec. 3.3).

More recently, λ has been moved out of the hyperparameters and into θ, and θ has been updated using Markov chain Monte Carlo (see, e.g., Linkletter et al. 2006; Sansó, Forest, and Zantedeschi 2008). Computationally and practically, this makes a large difference. With λ fixed, the predictive distribution has a closed form (multivariate Student t), but with λ uncertain, the predictive distribution is a mixture of multivariate Student t distributions, and predictions cannot be summarized in terms of parameters, but rather must be presented as a sample.

The approach that we advocate in this work differs from this standard approach. The primary source of this difference is our interest in emulators for *large* simulators, in which the collection of simulator evaluations is too small to span the important regions of simulator input space. Typically this situation arises when the simulator is expensive to evaluate or when the simulator input space is large; climate simulators have both of these characteristics. For large simulators, it is natural to augment the evaluations with expert knowledge, and so we consider ways to incorporate this knowledge be into the emulator.

First, we advocate a careful choice of regressors, typically many more regressors than simply a constant and linear terms. Most statisticians would agree that when detailed prior information is available, we should make informed choices for the regressors, although many believe that this is not necessary when there are plentiful simulator evaluations, because in that case the residual will adapt to the absent regressors. While agreeing with this in principle, we adopt a precautionary attitude in practice. The inclusion of regressors is favorable for extrapolation beyond the convex hull of the simulator inputs, and for large simulators, the convex hull is typically only a small fraction of the total volume. A second, more general reason for preferring carefully chosen regressors is that we typically will make some quite simple and tractable choices for the prior residual, such as separability and isotropy. These choices are unlikely to reflect our judgments, but this will matter less, predictively, if more of the output variability is explained by regressors.

Second, we have a general preference for "rougher" correlation functions, typically from the Matérn class. The squared exponential is tractable, particularly when used in a product over the components of x, but its extreme smoothness is often unrealistic for complex simulators with discrete solvers (subject to fixed-precision numerical errors) and can introduce problems when inverting large variance matrixes. In this choice we are in agreement with Stein (1999, p. 12), who advocated using the Matérn for spatial modeling. (Emulation and spatial modeling are "first cousins.")

Third, we advocate an informed choice for $\{m, V, a, d\}$ based on, for example, simple judgements about the unconditional distribution of $f(\cdot)$, that is, the distribution of $f(x^*)$, where x^* is treated as uncertain with some specified distribution function. This allows us to investigate the behavior of our prior emulator, which, one hopes, will make reasonably sensible predictions and will partially compensate when we have only a small number of evaluations. This point is related to the first point, because using a larger number of regressors entails a greater cost (in terms of predictive uncertainty) of specifying a vague prior. A careful choice of regressors and residual covariance function makes the task of specifying an informative choice for $\{m, V, a, d\}$ much more straightforward, as we demonstrate in Section 5.4.

Finally, we endorse the idea of automating the choice of correlation lengths, because these are hard to elicit, especially conditionally on the choice of regressors—but only in the context

of detailed diagnostic checking. Detailed diagnostic checking is also important for choosing the regressors. Here we depart from current practice by favoring "lightweight" emulators that can be constructed rapidly and have closed-form predictions. More general approaches that mix over candidate models within a sampling framework are theoretically elegant but are often impractical, because they cannot be used to generate predictive diagnostics in a reasonable amount of time. We favor predictive diagnostics, because they assign the task of quality assessment to the domain of the system expert. Simple but powerful predictive diagnostics are readily available in computer experiments, as we discuss later. But these diagnostics require us to repeatedly construct emulators on different subsets of the simulator evaluations, and thus a "quick" emulator is a prerequisite.

3. THE TIE–GCM SIMULATOR

The TIE-GCM simulator (Richmond, Ridley, and Roble 1992) is designed to calculate the coupled dynamics, chemistry, energetics, and electrodynamics of the global thermosphere–ionosphere system at an altitude of about 97–500 km. It has many input parameters to be specified at the lower and upper boundaries, as well as a number of uncertain internal parameters. There also are many output quantities from the TIE-GCM simulator (e.g., densities, winds, airglow emissions, geomagnetic perturbations) that can be compared with observations. For this study, we explore the response of the simulated ionospheric $\mathbf{E} \times \mathbf{B}/B^2$ drift velocity (m/s, where positive is upward), where $\mathbf{E}$ and $\mathbf{B}$ are the electric and geomagnetic fields, to variations in just three inputs: two that help describe atmospheric tides at the TIE-GCM lower boundary and one that constrains the minimum nighttime electron density. The drift varies daily, but also with season, solar cycle, and location of the observation. Averaging over many days for given geophysical conditions is necessary to determine a regular pattern. To avoid confusion, we refer to our particular treatment of TIE-GCM as the *simulator*, reserving the term "model" for "statistical model."

Atmospheric tides are global waves with periods that are harmonics of 24 hours. They are generated at lower atmospheric levels and are modulated by variable background winds as they propagate to the upper atmosphere. They are difficult to define, because observations are limited and the tides vary not only with geographic location, local time, and season, but also in a somewhat irregular manner from one day to the next. Modeling the tidal propagation through the atmosphere, and accurately determining their distribution at the TIE-GCM lower boundary, remain challenging. For this study, we include fixed diurnal (24-hour period) and semidiurnal (12-hour period) migrating (sun-synchronous) tidal components at the TIE-GCM lower boundary, taken from the physical model of Hagan and Forbes (2002a, 2002b), plus an additional variable, tidal forcing [migrating (2, 2) mode], which is known to be important for the electrodynamics (Fesen et al. 2000). The amplitude of the perturbation in the height of a constant-pressure surface at the TIE-GCM lower boundary, $\texttt{AMP} \in [0, 36]$ da m (1 da m is 1 dekameter, or 10 meters), and the local time at which this maximizes, $\texttt{PHZ} \in [0, 12]$ hour, are two of the three inputs that we explore.

At night, the ionospheric electron density below 200 km is small and difficult to measure, but nonetheless it has an important influence on the nighttime electric field. Our third simulator input is the logarithm, base 10, of the minimum nighttime electron density in cm^{-3}, $\texttt{EDN} \in [3, 4]$. All other input parameters in the TIE-GCM simulator are held constant for our experiments. The simulations are done for equinox, at low solar and geomagnetic activity. For each evaluation, the simulator is initially spun up to get a diurnally reproducible state.

The TIE-GCM $\mathbf{E} \times \mathbf{B}/B^2$ drift velocity outputs comprise periodic functions of magnetic local time at numerous sites across the globe. Here we analyze these sites marginally, disregarding shared information that might be available from sites that are proximate. Therefore, the simulator output for each evaluation comprises points on a periodic function of time for some prespecified site. Here we concentrate on the upward drift at the location of the Jicamarca incoherent scatter radar observatory (JRO), Peru, at the geomagnetic equator (11.9°S, 76.0°W geographic). In general, at the geomagnetic equator the upward drift is mostly positive during the day and negative at night. We write the simulator output as the scalar $f_i(r)$, where $r \equiv (\texttt{AMP}, \texttt{PHZ}, \texttt{EDN})$ and $s_i \in \mathcal{S} = [0, 24]$ hours indexes magnetic local time from midnight.

The upward drift for JRO is shown in Figure 1 for the collection of evaluations, generated as a maximin latin hypercube design of 30 evaluations using Euclidean distance on the unit cube (see, e.g., Koehler and Owen 1996). In retrospect, this was not the best choice of design, because it "wasted" evaluations at low values of AMP, where there is little response to either AMP or PHZ (see Section 5.2). Evaluation of TIE-GCM is expensive, taking about 15 minutes of clock time per day of simulated time on a supercomputer. At the time, we were concerned about pressing on with the experiment; we are exposing our shortcomings here as a caution to others! Despite our design, however, our results are very clear. Probably the main consequence of our mistake was an inefficient use of our resources;

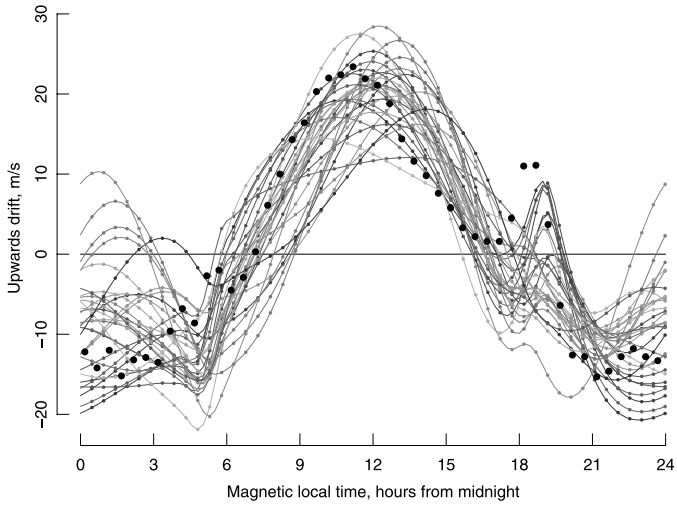

Figure 1. Collection of TIE-GCM evaluations for the JRO site, for a 30-point maximin latin hypercube design with 3 inputs. The small dots represent the actual simulator output, the lines interpolate the dots with a periodic B-spline, and the shading (from dark to light gray) runs from small to large values of EDN. The large dots indicate actual observations (for equinox at low solar activity, Fejer et al. 1991).

with a better design, we could have built a similar emulator with fewer than 30 evaluations. Another option would have been to proceed sequentially (see, e.g., van Beers and Kleijnen 2008).

The main features of the simulator output are a peak around noon and excursions in the early evening. A previous study showed that the tidal forcing at the lower boundary and the nighttime electron density influence these two features (Fesen et al. 2000). The peak in the early evening develops for low electron density in the lower ionosphere (E region), when the relative influence of the upper ionosphere (F region) dominates. The daytime upward drift is influenced mainly by the tidal winds.

4. THE OUTER PRODUCT EMULATOR

Rougier (2008) described a very efficient framework for constructing normal inverse gamma emulators for simulators with multivariate outputs, known as the OPE. An OPE is characterized by the following features:

1. A fixed set of output indexes, $\{s_1, \ldots, s_q\}$, that is invariant to the choice of simulator input, r.
2. A covariance function for the residual that is separable in r and s,
$$\kappa(r, s_i, r', s_{i'}) = \kappa^r(r, r') \times \kappa^s_{ii'}. \tag{3}$$
3. A collection of regressors, $\{g_1(\cdot), \ldots, g_v(\cdot)\}$, made up of the pairwise product of a set of regressors in r, $\mathcal{G}^r \triangleq \{g^r_1(\cdot), \ldots, g^r_{v_r}(\cdot)\}$, and a set of regressors in s, $\mathcal{G}^s \triangleq \{g^s_1(\cdot), \ldots, g^s_{v_s}(\cdot)\}$, thus $v = v_r \times v_s$.

Rougier showed that the construction and use of an OPE is effectively instantaneous even with hundreds of simulator evaluations and hundreds of simulator outputs. All of our calculations in this work were done using `OPE`, a package for constructing and using an OPE available for the R statistical computing environment (R Development Core Team 2004).

A separable covariance function, such as (3), is implied if we treat the residual as the product of two independent processes,

$$\epsilon(r, s_i) = \epsilon^r(r) \times \epsilon^s(s_i), \quad \text{where } \epsilon^r(\cdot) \perp\!\!\!\perp \epsilon^s(\cdot). \tag{4}$$

This is a very useful way to represent a residual with separable covariance, which lends itself to detailed statistical modeling. Standard practice is to go further than (4) and make $\epsilon^r(\cdot)$ itself separable in each of the inputs; we do not impose this additional separability here, as explained in Section 5.2.

The output of the TIE-GCM simulator is recorded at 0.5-hour intervals in universal time and then mapped to magnetic local time. This results in emulator time steps that are site-dependent but invariant to r and roughly 0.5 hour in duration, although some are slightly shorter and some are slightly longer. Using the OPE, we model the 48 simulator outputs directly, without any dimensional reduction and without interpolation onto equally spaced time steps. (We also investigated modeling the simulator outputs after interpolating onto $\{1, 2, \ldots, 24\}$ and found no difference in our conclusions.)

5. STATISTICAL MODELING CHOICES

In this section we describe the process involved in choosing the regressors and the residual covariance function for the emulator, taking into account expert knowledge. The TIE-GCM simulator may be unusual in terms of the strength of the expert knowledge that we have, but it is certainly not unique. The knowledge that we take into account is predicated on the physics of the simulator and would be shared by all well-informed experts.

Note that the regressors in each simulator input and in the simulator output are chosen to be orthonormal with respect to a rectangular weighting function (with one unavoidable exception; see Section 5.4), which substantially simplifies the process of eliciting the hyperparameters $\{m, V, a, d\}$. For this same reason, the covariance functions are correlation functions; that is, they are constructed to have variance 1. This means that τ is the variance of the residual.

5.1 Periodic Simulator Output

The TIE-GCM simulator has a smooth periodic output, so that $f_i(r) = f_{i'}(r)$ for all r, when $s_i = 0$ and $s_{i'} = 24$, and similarly in the first derivative. Therefore, the set of s regressors, $\mathcal{G}^s$, must comprise only smooth periodic functions, and the covariance function, $\kappa^s(\cdot)$, must generate smooth periodic sample paths.

For $\mathcal{G}^s$, it is natural to think of Fourier terms. But from the general nature of the simulator output, it is difficult to intuit just how many terms that we will need. We delegate this to a diagnostic comparison between alternatives. Thus we write

$$\mathcal{G}^s = \{1\} \cup \bigcup_{k=1}^{w} \{\sqrt{2}\sin(2\pi ks/24), \sqrt{2}\cos(2\pi ks/24)\} \quad \text{for some } w \in \{1, \ldots, 6\}, \tag{5}$$

where w, which sets the number of s regressors, remains to be determined, and the $\sqrt{2}$ is for orthonormality with respect to a uniform weighting function on $[0, 24]$.

For the covariance function of the residual $\epsilon^s(\cdot)$ from (4), we use the standard approach for creating periodic sample paths (see, e.g., Yaglom 1987 for the theory and Gneiting 1999 for an application). This involves setting

$$\kappa^s(s, s'; \lambda_s) = \phi\big(2R\sin(\angle(s, s')/2); \lambda_s\big), \qquad s, s' \in [0, 24], \tag{6}$$

where $\phi(\cdot; \lambda_s)$ is some isotropic correlation function with correlation length λ_s, the radius is $R = 24/2\pi$, and $\angle(s, s')$ is the angle (in radians) between s and s'. For $\phi(\cdot; \lambda_s)$, we use a Matérn correlation function with $\nu = 5/2$ degrees of freedom (see, e.g., Rasmussen and Williams 2006, chap. 4), denoted by $\mathrm{Mat}_{5/2}(\cdot)$, and set $\phi(d; \lambda) \triangleq \mathrm{Mat}_{5/2}(d/\lambda)$. This correlation function has reasonably smooth sample paths and is efficient to compute.

5.2 Amplitude and Phase

Recall that in the TIE-GCM simulator, $r = (\mathtt{AMP}, \mathtt{PHZ}, \mathtt{EDN})$. `AMP` and `PHZ` are closely related in the simulator, in the

sense that there can be no PHZ effect when AMP $= 0$, and larger values of AMP increase the impact of PHZ. We can build this property into our emulator by making careful choices for the r regressors in $\mathcal{G}^r$ and in the covariance function $\kappa^r(\cdot)$.

We choose to create our $\mathcal{G}^r$ regressors as products of functions in each of the inputs. For the AMP functions, we use a linear term and a quadratic term,

$$\begin{aligned} \mathtt{AMP}_1 &= \sqrt{3}\overline{\mathtt{AMP}} \qquad \text{and} \\ \mathtt{AMP}_2 &= -3\sqrt{5}\overline{\mathtt{AMP}} + 4\sqrt{5}\overline{\mathtt{AMP}}^2, \end{aligned} \tag{7}$$

where $\overline{\mathtt{AMP}} \triangleq \mathtt{AMP}/36$, that is, AMP scaled to the unit interval. The coefficients in these polynomials are chosen to make the two functions orthonormal with respect to a uniform weighting function on $[0, 36]$. Both of these functions are 0 for AMP $= 0$; thus if we always include an AMP function in a regressor that includes a PHZ function, then the regressors will respect the constraint at AMP $= 0$.

For the covariance function of the residual $\epsilon^r(\cdot)$ from (4), we write

$$\begin{aligned} \epsilon^r(\mathtt{AMP}, \mathtt{PHZ}, \mathtt{EDN}) \equiv \sqrt{\overline{\mathtt{AMP}}}\epsilon_1^r(\mathtt{AMP}, \mathtt{PHZ}, \mathtt{EDN}) \\ + \sqrt{1 - \overline{\mathtt{AMP}}}\epsilon_2^r(\mathtt{AMP}, \mathtt{EDN}), \end{aligned} \tag{8}$$

where $\epsilon_1^r(\cdot) \perp\!\!\!\perp \epsilon_2^r(\cdot)$, so that when AMP $= 0$, there is no contribution from PHZ. If both $\epsilon_1^r(\cdot)$ and $\epsilon_2^r(\cdot)$ have variance 1, then $\epsilon^r(\cdot)$ also has variance 1, as required. This treatment of the residual is an example of how simple knowledge about the simulator can affect the residual covariance function, and in particular how separable residual covariance functions, as are commonly used, can fail to capture this knowledge.

Finally, PHZ itself is a periodic simulator input, in the sense that $f_i(\mathtt{AMP}, 0, \mathtt{EDN}) = f_i(\mathtt{AMP}, 12, \mathtt{EDN})$, for all i, AMP, and EDN. For the PHZ regressor functions, we choose the Fourier terms

$$\begin{aligned} \mathtt{PHZ}_1 &= \sqrt{2}\sin(2\pi\,\mathtt{PHZ}/12) \qquad \text{and} \\ \mathtt{PHZ}_2 &= \sqrt{2}\cos(2\pi\,\mathtt{PHZ}/12), \end{aligned} \tag{9}$$

and for the residual covariance function we write, starting from (8),

$$\epsilon_1^r(\mathtt{AMP}, \mathtt{PHZ}, \mathtt{EDN}) \equiv \epsilon_{11}^r(\mathtt{AMP}, \mathtt{EDN}) \times \epsilon_{12}^r(\mathtt{PHZ}). \tag{10}$$

We use the same approach for the covariance function of $\epsilon_{12}^r(\cdot)$ as we did for $\epsilon^s(\cdot)$, described in Section 5.1.

5.3 Other Choices

To complete the set $\mathcal{G}^r$, we need to specify functions in EDN and then combine the functions for the three simulator inputs together into regressors. For EDN, we use first- and second-order Legendre polynomials shifted onto the interval $[3, 4]$, denoted by $\mathtt{EDN}_1$ and $\mathtt{EDN}_2$. Then our total set of simulator input regressors is

$$\begin{aligned} \mathcal{G}^r = \{1, \mathtt{AMP}_1, \mathtt{AMP}_2, \mathtt{AMP}_1 \times \mathtt{PHZ}_1, \\ \mathtt{AMP}_1 \times \mathtt{PHZ}_2, \mathtt{EDN}_1, \mathtt{EDN}_2\}. \end{aligned} \tag{11}$$

We have chosen a small set of just seven low-degree regressors for the three simulator inputs. Conventional wisdom is that simulator outputs tend to vary quite smoothly and simply with the simulator inputs r. This is in contrast to s, for which the simulator outputs can vary more dramatically, as is the case for TIE-GCM.

For the residuals $\epsilon_{11}^r(\mathtt{AMP}, \mathtt{EDN})$ and $\epsilon_2^r(\mathtt{AMP}, \mathtt{EDN})$, we use a separable covariance function, $\mathrm{Mat}_{5/2}(\cdot)$ in both cases. This leaves us with the choice of three simulator input correlation lengths, $\lambda_{\mathtt{AMP}}$, $\lambda_{\mathtt{PHZ}}$, and $\lambda_{\mathtt{EDN}}$, plus the simulator output correlation length λ_s. For each candidate model [i.e., each w in eq. (5)], we choose the four correlation lengths by maximizing the marginal likelihood. The results are given in Table 1. Note that the estimated correlation length, λ_s, drops as w increases, as expected; the other correlation lengths also change systematically. The final part of Section 6 describes a full Bayes treatment of the correlation lengths.

5.4 Completing the Prior Distribution

Specifying the remaining part of the emulator prior, expressed in terms of the values of the hyperparameters $\{m, V, a, d\}$, is relatively straightforward if the regressors are orthonormal and the covariance function of the residual has variance 1 everywhere. We have arranged for this to be true, with the exception of the regressors $\mathtt{AMP}_1$ and $\mathtt{AMP}_2$ (which were specially chosen to be 0 when AMP $= 0$). These two regressors are not orthogonal to the constant, but we will ignore this, because the effect on the calculation that follows is small.

Here we outline a relatively simple way to choose $\{m, V, a, d\}$, on the basis of broad judgments about the simulator. We consider $f_i(r)$ "averaged" uniformly over i and r, which might be visualized as the evaluations in Figure 1 projected onto the vertical axis (taking the maximin latin hypercube to be approximately uniform in r). To facilitate this, write x^* for $\{i^*, r^*\}$ and $f(x^*)$ for $f_{i^*}(r^*)$, let x^* have a rectangular distribution on the joint space of simulator inputs and simulator

Table 1. Correlation lengths, estimated by maximizing the marginal likelihood, for different candidates for the simulator output regressors, $\mathcal{G}^s$ [see eq. (5)], shown with asymptotic standard errors

	$w=1$		$w=2$		$w=3$		$w=4$		$w=5$		$w=6$	
	$\hat{\lambda}$	SE	$\hat{\lambda}$	SE	$\hat{\lambda}$	SE	$\hat{\lambda}$	SE	$\hat{\lambda}$	SE	$\hat{\lambda}$	SE
AMP	22.86	1.26	19.85	1.63	18.64	1.57	16.83	1.61	15.89	1.56	14.63	1.53
PHZ	1.53	0.06	1.41	0.08	1.39	0.08	1.31	0.08	1.29	0.09	1.23	0.09
EDN	0.50	0.02	0.32	0.02	0.33	0.02	0.29	0.02	0.30	0.02	0.29	0.02
s	1.01	0.02	0.74	0.02	0.72	0.02	0.70	0.02	0.71	0.02	0.72	0.02

output, and consider the mean and variance of $f(x^*)$. For the mean, $E(f(x^*)) = m_1$, the mean of the coefficient on the constant regressor. Because these are unconstrained by our choice for $E(f(x^*))$, we set the other components of m to 0, that is, $m = (E(f(x^*)), 0, \ldots, 0)^T$, which then simplifies the calculation of $\mathrm{Var}(f(x^*)|\tau)$. Because our regressors are orthonormal, it is natural to restrict V to the diagonal matrix $\sigma^2 I$. It then follows that $\mathrm{Var}(f(x^*)|\tau) = \tau(v\sigma^2 + 1)$. Integrating out τ gives $\mathrm{Var}(f(x^*)) = (d/(a-2))(v\sigma^2 + 1)$, providing that $a \geq 3$.

In the NIG emulator, a represents the strength of the prior information, in terms of the equivalent number of evaluations, and we specify this explicitly. This leaves d and σ^2 to be determined. To determine σ^2, we consider our emulator's prior, R^2, the proportion of variance attributable to the regressors,

$$R^2 = \frac{\sigma^2 v}{\sigma^2 v + 1}. \tag{12}$$

The value of d can then be inferred from our choice for $\mathrm{Var}(f(x^*))$.

For the TIE-GCM simulator, we choose $E(f(x^*)) = 0$, and $\mathrm{Var}(f(x^*)) = 15^2$. We also choose $a = 3$; although we are making informative prior judgments about $f(x^*)$, we want them to be replaced rapidly by information from the evaluations. Finally, we choose $R^2 = 0.9$. We might have revised these values had the emulator diagnostics indicated a problem, but, as we show in the next section, the emulator performed well once w was determined.

5.5 Finalizing the Choice of Model

We choose the undetermined parameter w [see eq. (5)] on the basis of visual diagnostics of emulator performance. We use predictive diagnostics, because they are the most relevant for the purpose of our emulator, predicting the simulator output at specific sets of input values. We use two types of diagnostics, both represented visually as samples from the updated prediction, with the true value superimposed:

1. *Leave-one-out* (LOO), in which we predict each evaluation in turn using an emulator constructed from our prior choices and the other evaluations. This shows us how much uncertainty we can expect in our predictions ($n-1$ being close to n) and how this uncertainty varies across the simulator's input-space.
2. *One-step-ahead* (OSA), in which we predict the first evaluation from the prior emulator, predict the second evaluation after updating by just the first, and so on. This shows us how rapidly we learn about the simulator by accumulating evaluations into the emulator. It also is closely linked to the *prequential* diagnostic approach (Dawid 1984; Cowell et al. 1999).

Having a specific ordering in the collection of evaluations is useful for interpreting LOO and affects the results for OSA. We order by the value of EDN, which we consider to have the most complicated impact on the simulator output, particularly at extreme values. In this ordering, every prediction in OSA is an extrapolation from the convex hull of the evaluations used in the emulator, which makes this a stern test.

We inspect LOO and OSA plots for all 30 evaluations, for the 6 candidate values for w. Overall, the hardest prediction to get right seems to be the OSA prediction for job 017. This is shown in Figure 2 for the different values of w. Based on all of the diagnostics, we choose $w = 2$; this also is the value that we judge to give the best emulator in Figure 2, although $w = 3$ is very similar. Figure 3 shows the LOO plot for $w = 2$; it can be seen that our emulator does a very good job of capturing a range of

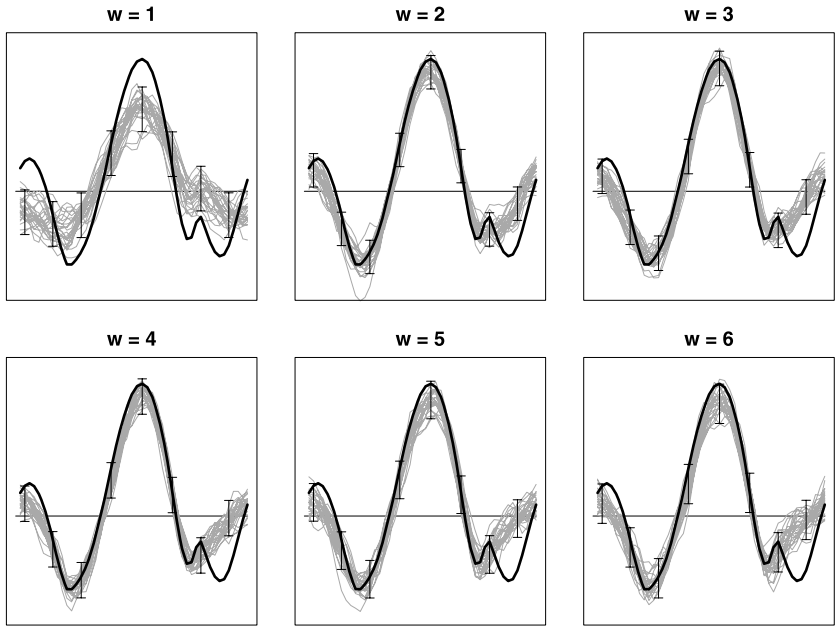

Figure 2. One-step-ahead (OSA) prediction of job 017 (i.e. using only the 15 evaluations with smaller EDN values). In each frame, w represents the set of $\mathcal{G}^s$ regressors [see eq. (5)], the gray lines represent 25 sampled values, the error bars indicate the marginal 95% symmetric credible intervals every six time-steps, and the black line represents the actual simulator output for job 017. See Figure 1 for details of the axes. Overall, $w = 2$ provides the best representation for job 017.

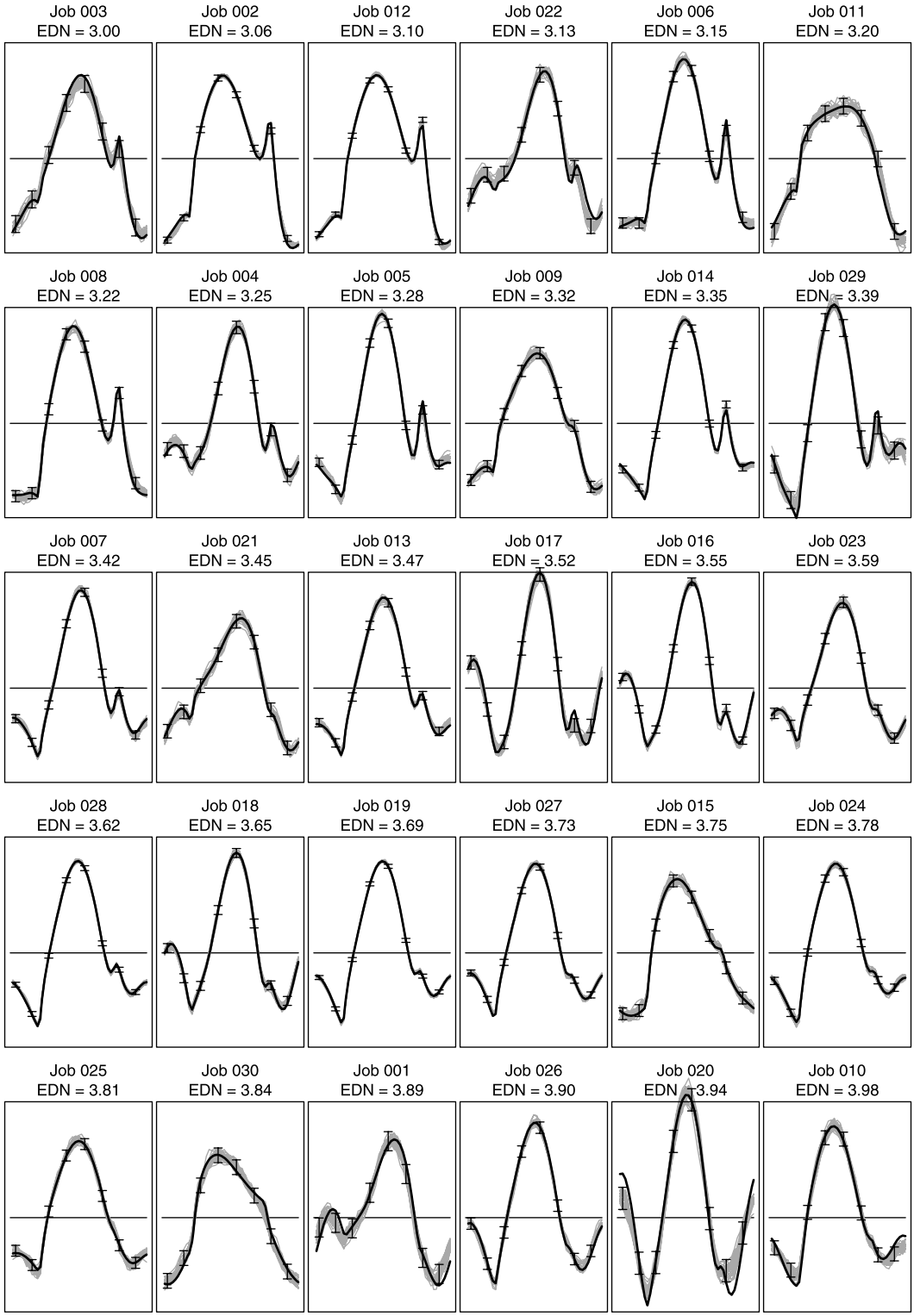

Figure 3. Leave-one-out (LOO) diagnostic plot for $w = 2$, our favoured choice for $\mathcal{G}^s$ [see eq. (5)], ordered by EDN; see the caption to Figure 2 for details.

quite different shapes over the simulator's input space, and that the uncertainties are well calibrated. Note that job 017 is well predicted when information from all of the other 29 evaluations is used.

Overall, how much did it cost to choose our emulator? We had six candidates [i.e. $w \in \{1, \ldots, 6\}$ in eq. (5)], and for each candidate we optimized the residual correlation length and then produced diagnostic plots. Optimizing the correlation lengths required us to build about 120 emulators for each candidate, and the diagnostic plots required about 60. Taken as a one-shot calculation, we had to build and use more than 1,000 emulators. Of course, in practice, we have built and used many times this number during the course of our analysis. This type of approach is possible only with a "lightweight"

emulator like the OPE, for which construction, computing the marginal likelihood, and making predictions are all effectively instantaneous for applications like our TIE-GCM simulator.

6. USING THE EMULATOR TO STUDY THE SIMULATOR

Here we describe one use of our emulator: visualizing the impact of changes in the three simulator inputs. This has two purposes, represented as sequential stages. The first stage is "code verification": does the simulator (as represented by the emulator) have the correct *qualitative* characteristics, as suggested by the physical theory? In the second stage, what are the *quantitative* effects of changing the simulator inputs? In the initial phase of our TIE-GCM experiment, we were able to identify a problem with the simulator code in the first stage, showing in a very direct way how emulators can add value to computer experiments. The data that we use here are from a corrected set of simulator evaluations.

Figure 4 shows a simple layout, with four values of EDN and for each value, a low, medium, and high value for AMP and a low and high value for PHZ. By construction, our emulator should (and does) generate identical sample paths over different values of PHZ when AMP $= 0$. We use the values AMP $\in \{0, 18, 36\}$ da m. We use two values of PHZ that are 180 degrees out of phase. In this case, we expect to see a reversal of the phase effect; we use PHZ $\in \{3, 9\}$ hr.

In all four panels of Figure 4, which shows the mean function for our selected values of the simulator-inputs, we see the qualitative relationship between AMP and PHZ that we anticipate. The two solid lines coincide as required (AMP $= 0$), and larger values of AMP are associated with a stronger response to PHZ. We also can see that the two values of PHZ give outputs that are close to having opposing phases.

Turning to EDN, the relationship revealed here is driven entirely by the evaluations, because our prior for the effect of EDN is neutral. Our main findings are that higher EDN suppresses the evening excursion and increases nighttime drift. Both of these findings are consistent with our qualitative physical understanding. The increase in electron density at night short-circuits the electric field generated in the upper ionosphere (F region), which is responsible for the peak in the early evening, (see, e.g., Eccles 1998). Therefore, the early-evening peak disappears with increased nighttime electron density. Because the electric potential drop along the nighttime equator from dawn to dusk is basically determined by the daytime electrodynamics, with much greater conductivities than at night, nighttime processes have little effect on the integral of the eastward electric field along the nighttime magnetic equator. Therefore, a reduction in the evening upward drift (eastward electric field) must be accompanied by a more positive (less negative) drift at the other hours of the night.

We also see that large values of EDN enhance the effect of AMP and PHZ, especially in the nighttime. This is an interaction between all four components: the three simulator inputs and the output index variable. The greater variability due to the tides during the night with increased E-region nighttime electron density might be due to the fact that the tides are not prop-

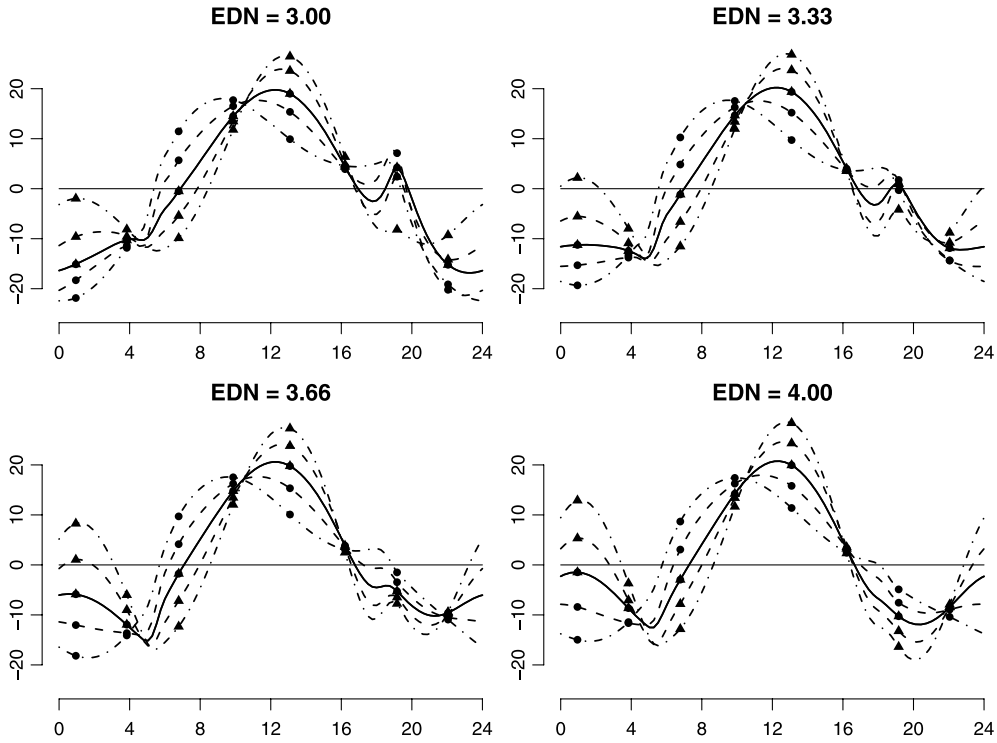

Figure 4. The simulator's response to different values of the three inputs (mean function, interpolated with a periodic B-spline). Line styles denote values of AMP: solid = 0, dashed = 18, dot-dashed = 36. Plotting characters denote values of PHZ: open circle = 3, filled triangle = 9. The two solid lines are coincident, because there is no PHZ effect when AMP = 0. See Figure 1 for details of the axes.

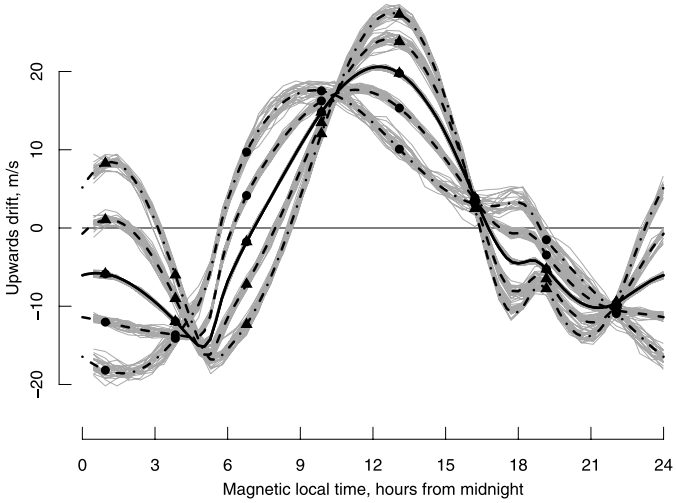

Figure 5. Effect of AMP and PHZ when EDN = 3.66, showing the uncertainty as 25 sampled values behind the mean function. See the caption to Figure 4 for details.

agating up to the F region, but are still reaching the E-region ionosphere. Strengthening the E-region electron density, and therefore the E-region electrodynamics, produces a clearer tidal signal.

The mean function shown in Figure 4 does not tell the whole story. For this, we also need the variance function. Uncertainty is shown in Figure 5, for EDN = 3.66. The uncertainties are much smaller than the signal, and it is clear that the mean function alone does a good job representing the simulator. (This is also the message from Figure 3.)

Sensitivity Assessment

A referee has requested that we provide a full Bayes analysis for comparative purposes, incorporating uncertainty about the correlation lengths of the residual. While this would be prohibitively expensive during the selection of the statistical model, we are happy to oblige with a sensitivity assessment on our favored model [$w = 2$ in eq. (5)], by looking at the effect of replacing the plugged-in values for λ with uncertain values drawn from the posterior distribution. This also gives us an opportunity to show how easy it is to embed the OPE within a hierarchical statistical model. For simplicity, suppose that we are interested in the expected value of the vector $f(r)$ at some specified r. In this case,

$$E[f(r)|F] = E\{E[f(r)|\lambda, F]|F\} = \int \mu(r;\lambda)\pi(\lambda|F)\,\mathrm{d}\lambda, \tag{13}$$

where F is the ensemble of simulator evaluations and $\mu(\cdot)$ is the emulator mean function, which has a closed-form expression because $f(r)|\lambda, F$ has a multivariate Student-t distribution.

We can draw samples from the posterior distribution

$$\pi(\lambda|F) \propto \pi(F|\lambda)\pi(\lambda) \tag{14}$$

using MCMC. Here $\pi(F|\lambda)$ has a closed-form expression because $F|\lambda$ also has a multivariate Student-t distribution. For our prior for λ, we use the product of diffuse Gamma distributions (each with mean equal to the plug-in value and a coefficient of variation equal to 1, i.e., an exponential distribution). Figure 6 shows the marginal prior and posterior distributions for each of

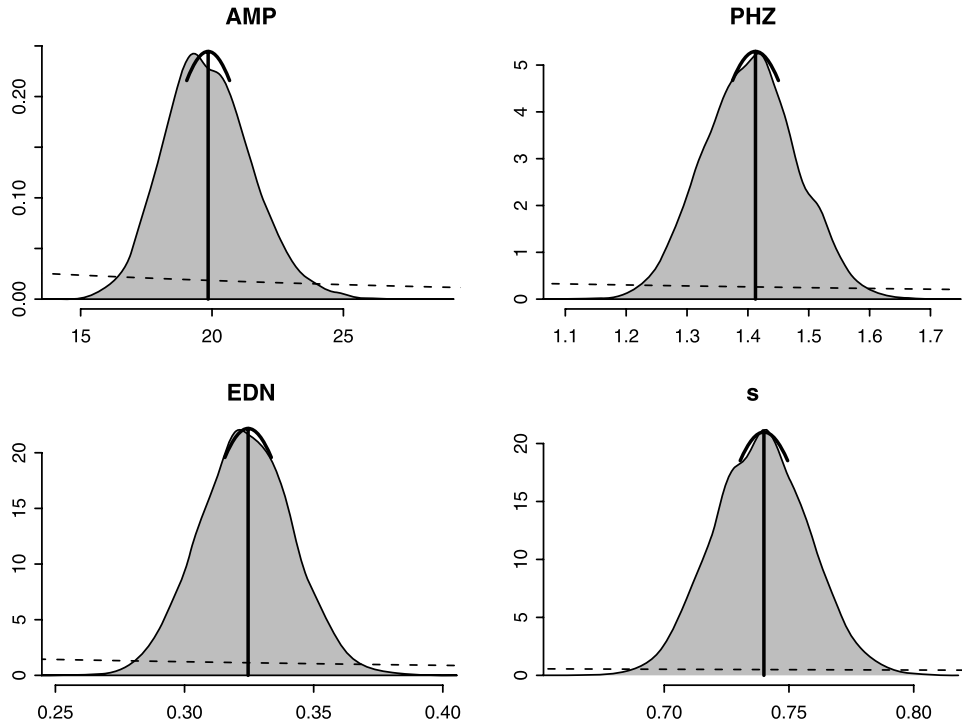

Figure 6. Posterior marginal probability density functions for the four residual correlation lengths. In each panel, the posterior is shown as a gray polygon, the corresponding part of the prior is represented by a dashed line, and the plug-in estimate from Table 1 is shown as an umbrella, ±half an asymptotic standard error.

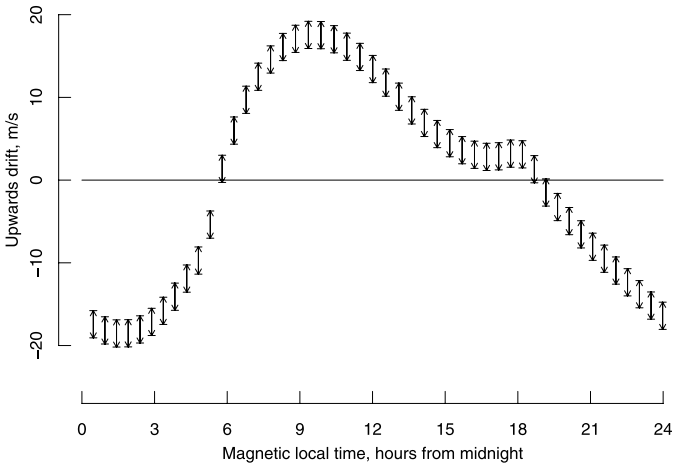

Figure 7. Predicted upward drift at input values AMP = 36, PHZ = 3, and EDN = 3.66, shown as error bars for the mean ± two standard deviations on the simulator time steps. There are two different treatments of the residual correlation lengths λ, plug-in and full Bayes. The plug-in error bars have flat terminals, and the full Bayes error bars have angled terminals. There is no discernible difference between the two sets of error bars.

the four components of λ, along with the plug-in value. This figure was constructed with 100,000 different sampled values for λ, necessitating the construction of 100,000 OPEs; even so, this takes only about 10 minutes on a desktop computer.

Using this sample, we can estimate the mean and variance of $f(r)$ with λ treated as uncertain, and contrast these with the mean and variance of $f(r)$ with λ plugged in. This is shown in Figure 7, for r specified as AMP = 36, PHZ = 3, and EDN = 3.66. There is no discernible difference between the two treatments. Numerically, the predictive standard deviations in the full Bayes case are about 1 percent larger than in the plugged-in case.

7. CONCLUSION

We have described an approach to emulation that goes beyond the standard choices, reflecting our desire to incorporate expert knowledge. This has affected our choice of regressors and of the covariance function for the residual, as well as (albeit to a lesser extent) our specification of the emulator prior distribution. Some of these choices actually affect our predictions only when the number of simulator evaluations is small, but others—most notably, the regressors—are important except in the limiting case, where the number of evaluations is very large. This limiting case is the exception when using simulators of complex physical systems. Such simulators are expensive to construct and to evaluate, and our general attitude is that if the scientific question justifies such expense, then we should not stint on the statistical effort, but should devote resources to eliciting expert knowledge about the simulator and finding ways to incorporate that knowledge into the emulator.

In this work we have used the OPE to model the multivariate output of the TIE-GCM simulator directly. Conditional on its hyperparameters, the OPE has a closed-form predictive distribution. While it thus might be embedded in a hierarchical statistical framework in which we also learn about the hyperparameters, we have chosen a different approach, making model choice (represented here as a choice between different sets of temporal regressors) depend explicitly on expert evaluation of diagnostics. Within each candidate model, we have estimated and plugged in the intractable parameter (the correlation lengths of the residual), to maintain a closed-form prediction. The sensitivity assessment at the end of Section 6 shows that the plug-in and the full Bayes treatments give the same predictions in our application.

We note two points regarding our "lightweight" approach here. First, emulators of complex deterministic functions are very complicated statistical objects, and we *must* have diagnostic validation of our emulator, regardless of how it is constructed. The literature on emulators is noticeably short on detailed diagnostic analysis (Bastos and O'Hagan 2009, is an exception). This is a source of concern, because we know from experience that it is very easy to build a bad emulator and hard to build a good one, if the simulator is complex. Predictive diagnostics are the most powerful, being located in the domain of the system expert and directly related to the purpose of the emulator: predicting the simulator output at untried inputs.

Second, we think that the system expert and the statistician, assisted by powerful visual diagnostics, can do a better job choosing an emulator directly than choosing a joint distribution over the emulator parameters and hyperparameters, as would be required in a hierarchical statistical analysis. Formally, the expert is standing in for the loss function in a decision problem, because he or she has a clear idea of how the emulator is to be used, what aspects of its performance are crucial, and what aspects can be downweighted. It is hard to envisage this information being quantified, but the absence of a loss function in what is clearly a decision problem leaves the statistical inference dangling. We want to put the choice back in, but we prefer to do so by having the system expert and the statistician select their candidate emulator explicitly. Thus we will need to construct thousands of emulators, because each candidate needs to be presented in terms of its predictive diagnostics. A lightweight approach, such as the one we outline here, is then the only option.

ACKNOWLEDGMENTS

This is a substantially revised version of 'Emulating the Thermosphere–Ionosphere Electrodynamics General Circulation Model (TIE-GCM),' by the same authors. This work was started while the first author was a Duke University Fellow, as part of the SAMSI program "Development, Assessment and Utilization of Complex Computer Models," and then as a visitor to the IMAGe group at the National Center for Atmospheric Research (NCAR), Boulder, CO. The authors thank both of these institutions for their support. NCAR is sponsored by the National Science Foundation (NSF). This material is based on work supported by the NSF under Agreement DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the NSF.

[Received August 2007. Revised May 2009.]

REFERENCES

Bastos, L., and O'Hagan, A. (2009), "Diagnostics for Gaussian Process Emulators," *Technometrics*, 51, 425–438.

Conti, S., and O'Hagan, A. (2007), "Bayesian Emulation of Complex Multi-Output and Dynamic Computer Models," Technical Report 569/07, University of Sheffield, Dept. of Probability and Statistics. Available at *http://www.tonyohagan.co.uk/academic/ps/multioutput.ps*.

Cowell, R., David, A., Lauritzen, S., and Spiegelhalter, D. (1999), *Probabilistic Networks and Expert Systems*, New York: Springer.

Craig, P., Goldstein, M., Rougier, J., and Seheult, A. (2001), "Bayesian Forecasting for Complex Systems Using Computer Simulators," *Journal of the American Statistical Association*, 96, 717–729.

Dawid, A. (1984), "Statistical Theory: The Prequential Approach" (with discussion), *Journal of the Royal Statistical Society, Ser. A*, 147 (2), 278–292.

Drignei, D. (2006), "Empirical Bayesian Analysis for High-Dimensional Computer Output," *Technometrics*, 48 (2), 230–240.

Eccles, J. (1998), "Modeling Investigation of the Evening Prereversal Enhancement of the Zonal Electric Field in the Equatorial Ionosphere," *Journal of Geophysical Research*, 103 (26), 709–726.

Fejer, B., de Paula, E., González, S., and Woodman, R. (1991), "Average Vertical and Zonal *F* Region Plasma Drifts Over Jicamarca," *Journal of Geophysical Research*, 96, 13901–13906.

Fesen, C. G., Crowley, G., Roble, R. G., Richmond, A. D., and Fejer, B. G. (2000), "Simulation of the Pre-Reversal Enhancement in the Low Latitude Vertical Ion Drifts," *Geophysical Research Letters*, 27 (13), 1851.

Gneiting, T. (1999), "Correlation Functions for Atmospheric Data Analysis," *Quarterly Journal of the Royal Meteorological Society*, 125, 2449–2464.

Goldstein, M., and Rougier, J. (2004), "Probabilistic Formulations for Transferring Inferences From Mathematical Models to Physical Systems," *SIAM Journal on Scientific Computing*, 26 (2), 467–487.

——— (2006), "Bayes Linear Calibrated Prediction for Complex Systems," *Journal of the American Statistical Association*, 101, 1132–1143.

——— (2009), "Reified Bayesian Modelling and Inference for Physical Systems" (with discussion), *Journal of Statistical Planning and Inference*, 139, 1221–1239.

Hagan, M., and Forbes, J. (2002a), "Migrating and Nonmigrating Diurnal Tides in the Middle and Upper Atmosphere Excited by Troposheric Latent Heat Release," *Journal of Geophysical Research*, 107 (D24), 4754.

——— (2002b), "Migrating and Nonmigrating Semidiurnal Tides in the Middle and Upper Atmosphere Excited by Troposheric Latent Heat Release," *Journal of Geophysical Research*, 108 (A2), 1062.

Higdon, D., Gattiker, J., Williams, B., and Rightley, M. (2008), "Computer Model Calibration Using High Dimensional Output," *Journal of the American Statistical Association*, 103, 570–583.

Kennedy, M., and O'Hagan, A. (2001), "Bayesian Calibration of Computer Models" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 63, 425–464.

Koehler, J., and Owen, A. (1996), "Computer Experiments," in *Handbook of Statistics, 13: Design and Analysis of Experiments*, eds. S. Ghosh and C. Rao, Amsterdam: North-Holland, pp. 261–308.

Linkletter, C., Bingham, D., Hengartner, N., Higdon, D., and Ye, K. (2006), "Variable Selection for Gaussian Process Models in Computer Experiments," *Technometrics*, 48 (4), 478–490.

Liu, F., and West, M. (2009), "A Dynamic Modelling Strategy for Bayesian Computer Model Emulation," *Bayesian Analysis*, 4 (2), 393–412.

Oakley, J., and O'Hagan, A. (2002), "Bayesian Inference for the Uncertainty Distribution of Computer Model Outputs," *Biometrika*, 89 (4), 769–784.

——— (2004), "Probabilistic Sensitivity Analysis of Complex Models: A Bayesian approach," *Journal of the Royal Statistical Society, Ser. B*, 66, 751–769.

O'Hagan, A. (2006), "Bayesian Analysis of Computer Code Outputs: A Tutorial," *Reliability Engineering and System Safety*, 91, 1290–1300.

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

Rasmussen, C., and Williams, C. (2006), *Gaussian Processes for Machine Learning*, Cambridge, MA: MIT Press. Available at *http://www.GaussianProcess.org/gpml/*.

Richmond, A., Ridley, E., and Roble, R. (1992), "A Thermosphere/Ionosphere General Circulation Model With Coupled Electrodynamics," *Geophysical Research Letters*, 19 (6), 601.

Rougier, J. (2008), "Efficient Emulators for Multivariate Deterministic Functions," *Journal of Computational and Graphical Statistics*, 17 (4), 827–843.

Rougier, J., and Sexton, D. (2007), "Inference in Ensemble Experiments," *Philosophical Transactions of the Royal Society, Ser. A*, 365, 2133–2143.

Sansó, B., Forest, C., and Zantedeschi, D. (2008), "Inferring Climate System Properties Using a Computer Model" (with discussion), *Bayesian Analysis*, 3 (1), 1–38.

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

Stein, M. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer Verlag.

van Beers, W., and Kleijnen, J. (2008), "Customized Sequential Designs for Random Simulation Experiments: Kriging Metamodeling and Bootstrapping," *European Journal of Operational Research*, 186 (3), 1099–1113.

Yaglom, A. (1987). *Correlation Theory of Stationary and Related Random Functions*, New York: Springer-Verlag.