

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

Jonathan Rougier*

Department of Mathematics
University of Bristol, UK

Serge Guillas

Department of Statistical Science
University College London, UK

Astrid Maute and Arthur D. Richmond
High Altitude Observatory
National Center for Atmospheric Research
Boulder CO, USA

May 20, 2009

Abstract

The TIE-GCM simulator of the upper atmosphere has a number of features that are a challenge to standard approaches to emulation, such as a long run-time, multivariate output, periodicity, and strong constraints on the inter-relationship 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.

KEYWORDS: OUTER PRODUCT EMULATOR, GAUSSIAN PROCESS, PREDICTIVE DIAGNOSTICS

*Corresponding author: Department of Mathematics, University of Bristol, University Walk, Bristol BS8 1TW, U.K.; email j.c.rougier@bristol.ac.uk. This is a substantially revised version of ‘Emulating the Thermosphere-Ionosphere Electrodynamics General Circulation Model (TIE-GCM)’, by the same authors.

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 in understanding simulators, and also in inferences that combine simulator evaluations with system observations, to make system predictions. O’Hagan (2006) provides an introduction to emulators, with more details in Santner *et al.* (2003). Kennedy and O’Hagan (2001), and Craig *et al.* (2001) and Goldstein and Rougier (2006) describe two Bayesian approaches to emulator-based system inference. Oakley and O’Hagan (2002) describes the use of emulators for uncertainty analysis, and Rougier and Sexton (2007) contrasts uncertainty analysis for a climate simulator without and with an emulator. Oakley and O’Hagan (2004) describes the use of emulators for sensitivity analysis; ‘screening’ for active variables is a variant on this, see, e.g., Linkletter *et al.* (2006) and the references therein. Goldstein and Rougier (2004, 2009) discuss the role of emulators in a general framework for linking model evaluations and system behaviour. Sansó *et al.* (2008) is a recent application of emulation in climate prediction, with a discussion that considers some of the foundational and practical issues that arise.

This paper develops two themes in parallel: a case-study in how to introduce expert knowledge about the simulator into the statistical choices that make up the emulator, and a ‘lightweight’ approach to multivariate emulation which prioritises efficient emulators, enabling a detailed analysis of predictive diagnostics. On the latter theme, this is the first paper to demonstrate the Outer Product Emulator (OPE) approach (Rougier, 2008) to multivariate emulation; other approaches to multivariate emulation are suggested by Drignei (2006), Conti and O’Hagan (2007) and Higdon *et al.* (2008). 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 we make in building our emulator for the TIE-GCM simulator, and ways to produce and use diagnostic information to inform our choices. Section 6 shows the use of the emulator to ‘verify’ the simulator code, and to understand the simulator better. Section 7 concludes.

2 Approaches to emulation

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

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

where the lefthand side is the i th simulator output at simulator input r (r for ‘run’), and the righthand side comprises the sum of a set of regressors with unknown coefficients, and a residual stochastic process. 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 $\theta \triangleq \{\beta, \epsilon(\cdot)\}$, 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:

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

$$\beta \mid \tau, \Psi \sim \text{N}(m, \tau V) \tag{2b}$$

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

$$\tau \mid \Psi \sim \text{IG}(a, d) \tag{2d}$$

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

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 maximising the marginal likelihood and plugged in. Other fitting approaches, e.g. REML or cross-validation, are also advocated (see, e.g., Santner *et al.*, 2003, sec. 3.3).

More recently, λ has been moved out of the hyperparameters and into θ , and θ has been updated using MCMC (see e.g.: Linkletter *et al.*, 2006; Sansó *et al.*, 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 summarised in terms of parameters, but must be presented as a sample.

The approach we advocate in this paper is different to this standard approach. The source of this difference is primarily 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 arises when the simulator is expensive to evaluate, or when the simulator input space is large. Climate simulators, for example, have both of these characteristics. For large simulators, it is natural to augment the evaluations with expert knowledge, and so we consider ways in which this knowledge can be incorporated into the emulator.

First, we advocate a careful choice of regressors, and typically many more regressors than simply a constant and linear terms. Most statisticians would agree that, where detailed prior information is available, we should make informed choices for the regressors, although many believe that it is not necessary when there are plentiful simulator evaluations, since in this 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 favourable 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 wanting carefully-chosen regressors is that we will typically make some quite simple and tractable choices for the prior residual, such as separability and isotropy. These choices are unlikely to reflect our judgements, but this will matter less, predictively, if the prior variance we attribute to the residual is smaller.

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 matrices. In this choice we are in agreement with Stein (1999), who advocates the use of the Matérn for spatial modelling (emulation and spatial modelling are ‘first cousins’).

Third, we advocate an informed choice for $\{m, V, a, d\}$; for example, based on simple judgements about the unconditional distribution of $f(\cdot)$, i.e. the distribution of $f(x^*)$ where x^* is treated as uncertain with some specified distribution function. This allows us to investigate the behaviour of our prior emulator, which one hopes, will make reasonably sensible predictions, and will partially compensate where we only have a small number of evaluations. This point is related to the first, because if we use a larger number of regressors, then there is a larger cost (in terms of predictive uncertainty) to 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 subsection 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 within the context of detailed diagnostic checking*. Detailed diagnostic checking is also important for choosing the regressors. Here we depart from current practice by favouring ‘lightweight’ emulators which are rapid to construct, and have closed-form predictions. More general approaches, which mix over candidate models within a sampling framework, are theoretically elegant. But they are generally impractical, as they cannot be used to generate predictive diagnostics in a reasonable amount of time. We favour predictive

diagnostics, because they locate the task of quality assessment into the domain of the system expert. Simple but powerful predictive diagnostics are readily available in computer experiments, as we show below. But these diagnostics require us repeatedly to construct emulators on different subsets of the simulator evaluations, and hence a ‘quick’ emulator is a prerequisite.

3 The TIE-GCM simulator

The TIE-GCM simulator (Richmond *et al.*, 1992) is designed to calculate the coupled dynamics, chemistry, energetics, and electrodynamics of the global thermosphere-ionosphere system between about 97 km and 500 km altitude. It has many input parameters to be specified at the lower and upper boundaries, as well as a number of uncertain internal parameters. There are also many output quantities from the TIE-GCM simulator (densities, winds, air-glow emissions, geomagnetic perturbations, etc.) 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 upwards), 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 night-time electron density. The drift varies daily, but also with season, solar cycle and location of the observation. Only averaging over many days for given geophysical conditions will give a regular pattern. To avoid confusion, we refer to our particular treatment of TIE-GCM as the *simulator*, reserving the word *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 they are modulated by variable background winds as they propagate to the upper atmosphere. They are difficult to define since 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. Modelling the tidal propagation through the atmosphere, and accurately determining their distribution at the TIE-GCM lower boundary, remains a challenge. 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,b), 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, $\text{AMP} \in [0, 36]$ da m, and the local time at which this maximises, $\text{PHZ} \in [0, 12]$ hr, are two of the three inputs we explore.

At night, the ionospheric electron density below 200 km is small and difficult to measure, but nonetheless has an important influence on the night-time electric field. Our third simulator input is the logarithm, base 10, of the minimum night-time electron density in cm^{-3} , $\text{EDN} \in [3, 4]$. 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 a large collection of sites across the globe. Here we analyse the 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 pre-specified site. We concentrate here 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 during the night. We write the simulator output as the scalar $f_i(r)$, where $r \equiv (\text{AMP}, \text{PHZ}, \text{EDN})$ and $s_i \in \mathcal{S} = [0, 24]$ hrs indexes magnetic local time from midnight. We refer to r as the value of the simulator inputs, and s_i as the simulator output index.

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 subsection 5.2). TIE-GCM is expensive to evaluate, taking about 15 minutes of wall-clock time per day of simulated time, on a super-computer. At the time we were concerned to press on with the experiment: we are exposing our shortcomings here as a caution to others! Despite our design, though, our results are very clear. Probably the main consequence of our mistake was to make an inefficient use of our resources: with a better design we could have built a similar emulator with fewer than thirty 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 night time 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 mainly influenced by the tidal winds.

4 The Outer Product Emulator (OPE)

Rougier (2008) describes a very efficient framework for constructing Normal Inverse Gamma emulators for simulators with multivariate outputs, known as an Outer Product Emulator (OPE). The three features that characterise an

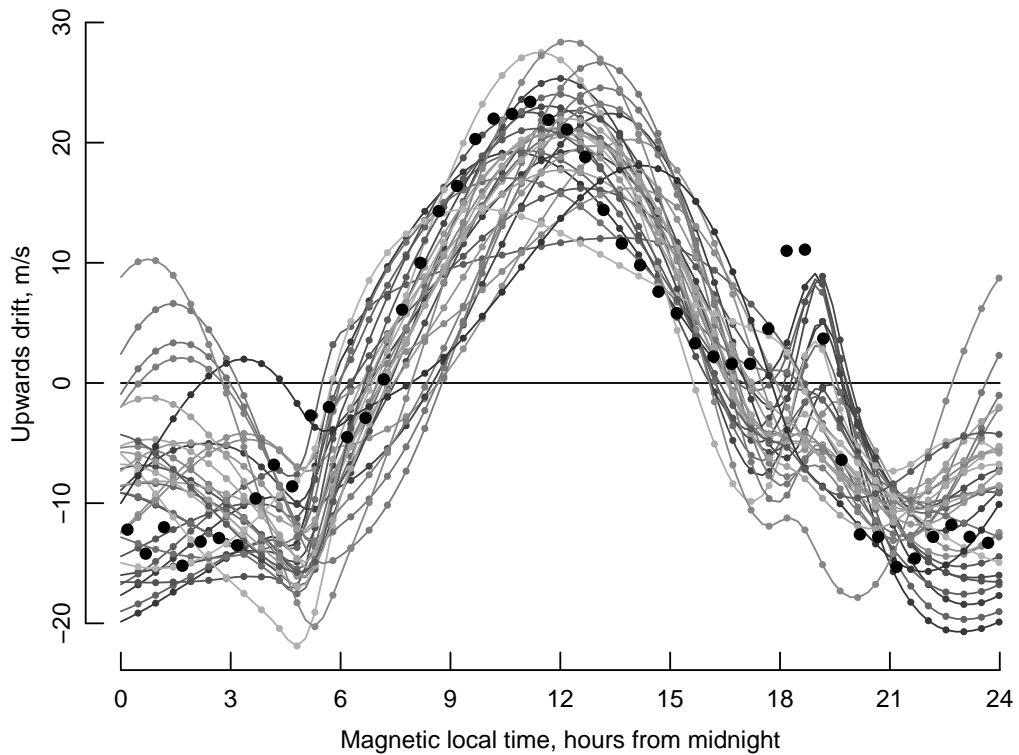


Figure 1: The collection of TIE-GCM evaluations for the Jicamarca (JRO) site, for a 30-point maximin Latin Hypercube design, with three inputs. The small dots represent the actual simulator output, the lines interpolate the dots with a periodic B-spline, and the shade (from dark- to light-grey) runs from small to large values of EDN. The large dots show actual observations (for equinox at low solar activity, Fejer *et al.*, 1991).

OPE are:

1. A fixed set of output-indices $\{s_1, \dots, s_q\}$ that is invariant to the choice of simulator input, r .
2. A covariance function for the residual that is separable in r and s ,

$$\kappa(r, s_i, r', s_{i'}) = \kappa^r(r, r') \times \kappa_{ii'}^s. \quad (3)$$

3. A collection of regressors $\{g_1(\cdot), \dots, g_v(\cdot)\}$ that is made up of the pairwise product of a set of regressors in r , $\mathcal{G}^r \triangleq \{g_1^r(\cdot), \dots, g_{v_r}^r(\cdot)\}$ and a set of regressors in s , $\mathcal{G}^s \triangleq \{g_1^s(\cdot), \dots, g_{v_s}^s(\cdot)\}$, hence $v = v_r \times v_s$.

Rougier shows that the construction and use of an OPE is effectively instantaneous even with hundreds of simulator evaluations and hundreds of simulator outputs. The calculations in this paper were made using `OPE`, a package for constructing and using an OPE, which is available for the statistical computing environment R (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,

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

This is a very useful way to represent a residual with separable covariance, which leads itself to detailed statistical modelling. Standard practice is to go further than (4) and make $\epsilon^r(\cdot)$ itself separable in each of the inputs: we will *not* be imposing this additional separability, as explained in subsection 5.2.

The output of the TIE-GCM simulator is recorded at 0.5 hr 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 hr in duration, although some are slightly shorter, and some slightly longer. Using the OPE, we model the 48 simulator outputs directly, without any dimensional-reduction, and without interpolating onto equally-spaced timesteps. But we also investigated modelling the simulator outputs after interpolating onto $\{1, 2, \dots, 24\}$, and found no difference in our conclusions.

5 Statistical modelling choices

This section illustrates the process of choosing the regressors and the residual covariance function for the emulator, taking account of expert knowledge. The TIE-GCM simulator may be unusual in the strength of the expert knowledge we have, but it is certainly not unique. The knowledge we take account of 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 will be chosen to be orthonormal with respect to a rectangular weighting function (with one unavoidable exception, see subsection 5.4), which substantially simplifies the process of eliciting the hyperparameters $\{m, V, a, d\}$. For the same reason, the covariance functions will be correlation functions, i.e. they

will be constructed to have variance one. 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. However, it is hard to intuit, from the general nature of the simulator output, just how many terms we will need, and so this is something we delegate to a diagnostic comparison between alternatives. Thus we write

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

where w , which sets the number of s -regressors, is yet 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 is to set

$$\kappa^s(s, s'; \lambda_s) = \phi\left(2R \sin(\angle(s, s')/2); \lambda_s\right) \quad s, s' \in [0, 24] \quad (6)$$

where $\phi(\cdot; \lambda_s)$ is some isotropic correlation function with correlation length λ_s , the radius $R = 24/2\pi$ 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, ch. 4), denoted $\text{Mat}_{5/2}(\cdot)$, and set $\phi(d; \lambda) \triangleq \text{Mat}_{5/2}(d/\lambda)$. 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, $r = (\text{AMP}, \text{PHZ}, \text{EDN})$. AMP and PHZ are closely related in our simulator, in the sense that there can be no PHZ effect when AMP = 0, and that larger values of AMP lead to a larger impact from PHZ. We can build this 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 and a quadratic term,

$$\text{AMP}_1 = \sqrt{3} \bar{\text{AMP}} \quad \text{and} \quad \text{AMP}_2 = -3\sqrt{5} \bar{\text{AMP}} + 4\sqrt{5} \bar{\text{AMP}}^2 \quad (7)$$

where $\bar{\text{AMP}} \triangleq \text{AMP}/36$, i.e., 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

zero for $\text{AMP} = 0$, and so if we always include an AMP function in a regressor which includes a PHZ function, then the regressors will respect the constraint at $\text{AMP} = 0$.

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

$$\epsilon^r(\text{AMP}, \text{PHZ}, \text{EDN}) \equiv \sqrt{\overline{\text{AMP}}} \epsilon_1^r(\text{AMP}, \text{PHZ}, \text{EDN}) + \sqrt{1 - \overline{\text{AMP}}} \epsilon_2^r(\text{AMP}, \text{EDN}), \quad (8)$$

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

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

$$\text{PHZ}_1 = \sqrt{2} \sin(2\pi\text{PHZ}/12) \quad \text{and} \quad \text{PHZ}_2 = \sqrt{2} \cos(2\pi\text{PHZ}/12), \quad (9)$$

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

$$\epsilon_1^r(\text{AMP}, \text{PHZ}, \text{EDN}) \equiv \epsilon_{11}^r(\text{AMP}, \text{EDN}) \times \epsilon_{12}^r(\text{PHZ}) \quad (10)$$

and we use the same approach for the covariance function of $\epsilon_{12}^r(\cdot)$ as we did for $\epsilon^s(\cdot)$, described in subsection 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 EDN_1 and EDN_2 . Our total set of simulator input regressors is then

$$\mathcal{G}^r = \{1, \text{AMP}_1, \text{AMP}_2, \text{AMP}_1 \times \text{PHZ}_1, \text{AMP}_1 \times \text{PHZ}_2, \text{EDN}_1, \text{EDN}_2\}. \quad (11)$$

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 $\epsilon_{11}^r(\text{AMP}, \text{EDN})$ and $\epsilon_{12}^r(\text{AMP}, \text{EDN})$ we will use a separable covariance function, $\text{Mat}_{5/2}(\cdot)$ in both cases. This leaves us with the three simulator input correlation lengths to choose, λ_{AMP} , λ_{PHZ} , and λ_{EDN} , plus the simulator output correlation length λ_s . For each candidate model (i.e. for each w in eq. 5) we will choose the four correlation lengths by maximising 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.

	$w = 1$		$w = 2$		$w = 3$		$w = 4$		$w = 5$		$w = 6$	
	$\hat{\lambda}$	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

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

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 one everywhere. We have arranged for this to be true, with the exception of the regressors AMP₁ and AMP₂ (which were specially chosen to be zero when AMP = 0): these two regressors are not orthogonal to the constant, but we will ignore this as the effect on the following calculation is small.

Here we outline a relatively simple way to choose $\{m, V, a, d\}$, on the basis of broad judgements about the simulator. We consider $f_i(r)$ ‘averaged’ uniformly over i and r , which might be visualised 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 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. As they are unconstrained by our choice for $E(f(x^*))$, we set the other components of m to zero, i.e. $m = (E(f(x^*)), 0, \dots, 0)^T$, which then simplifies the calculation of $\text{Var}(f(x^*) | \tau)$. Because our regressors are orthonormal, it is natural to restrict V to the diagonal matrix $\sigma^2 I$. Then it follows that $\text{Var}(f(x^*) | \tau) = \tau(v\sigma^2 + 1)$. Integrating out τ gives $\text{Var}(f(x^*)) = (d/(a - 2))(v\sigma^2 + 1)$, 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,

$$R^2 = \frac{\sigma^2 v}{\sigma^2 v + 1}. \quad (12)$$

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

For the TIE-GCM simulator, we choose $E(f(x^*)) = 0$, and $\text{Var}(f(x^*)) = 15^2$. We choose $a = 3$: although we are making informative prior judgements about $f(x^*)$, we want them to be replaced rapidly by information from the evaluations. Finally, we choose $R^2 = 0.9$. Possibly we would have revised these values if the emulator diagnostics had indicated a problem, but, as the next subsection shows, the emulator performs well once w has been determined.

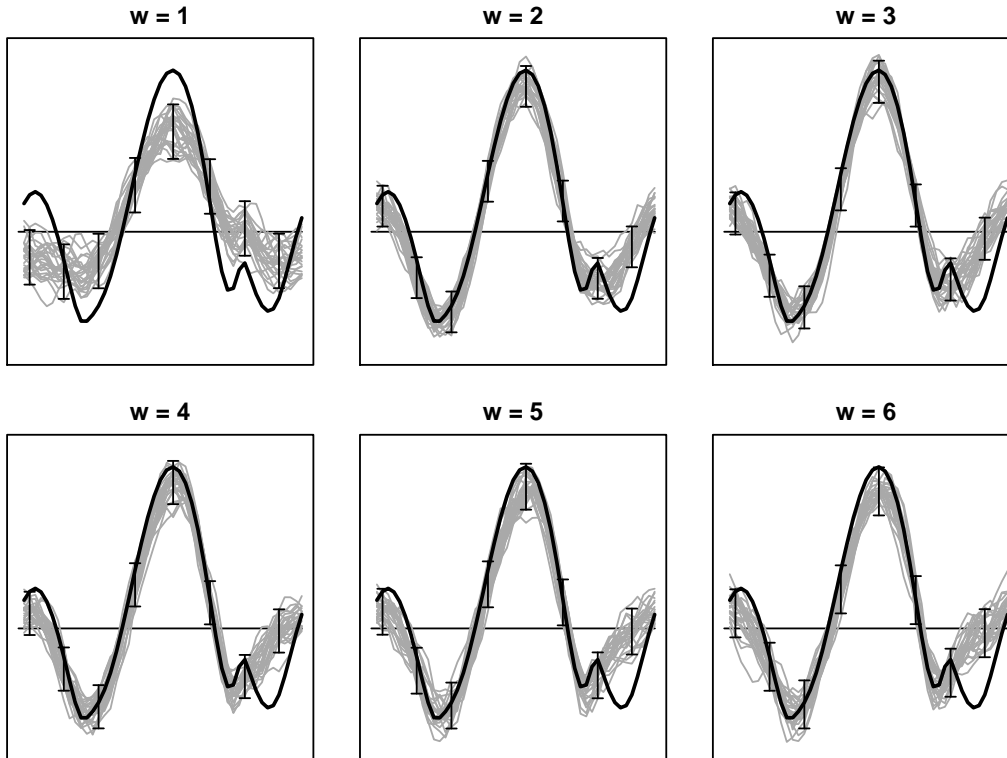


Figure 2: One-step-ahead (OSA) prediction of job 017 (i.e. using only the 15 evaluations with smaller EDN values). In each frame, w shows the set of \mathcal{G}^s regressors (see eq. 5), the grey lines show 25 sampled values, the error bars show the marginal 95% symmetric credible intervals every six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

5.5 Finalising the choice of model

We will choose the undetermined parameter w (see eq. 5) on the basis of visual diagnostics of emulator performance. We use predictive diagnostics, as 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 diagnostic, both represented visually as samples from the updated prediction, with the true value superimposed. First, *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. Second, *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 through accumulating evaluations into the emulator. It is also 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 result for OSA. We order by the value of EDN, as

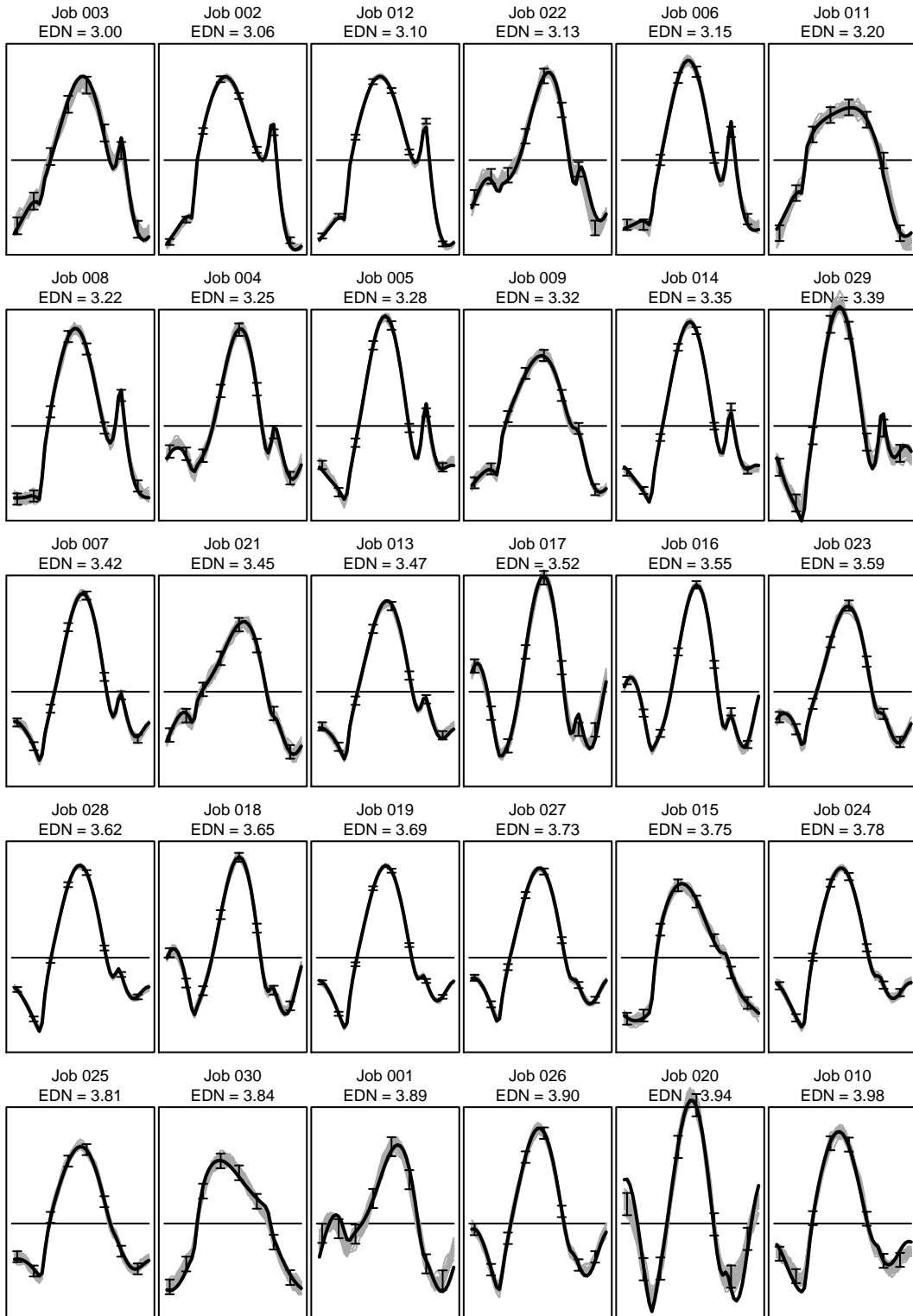


Figure 3: Leave-one-out (LOO) diagnostic plot, for $w = 2$, our favoured choice for \mathcal{G}^s (see eq. 5), ordered by EDN; see the caption to Figure 2 for details.

we judge this 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 thirty evaluations, for the six 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 . On the basis of all of the diagnostics we choose $w = 2$; this is also the value that we judge gives the best emulator in Figure 2, although $w = 3$ is very similar. Figure 3 shows the LOO plot for $w = 2$, and it can be seen that our emulator does a very good job of capturing a range of quite different shapes over the simulator’s input-space, and the uncertainties are well-calibrated. Note that job 017 is well-predicted when using information from all of the other 29 evaluations.

Overall, how much did it cost to choose our emulator? We had six candidates (i.e. $w \in \{1, \dots, 6\}$ in eq. 5), and for each candidate we optimised the residual correlation length, and then produced diagnostic plots. Optimising 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 have had to build and use over 1000 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 only possible 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 show one use of our emulator, visualising 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 we use here are from a corrected set of simulator evaluations.

We show a simple layout in Figure 4, 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 $\text{AMP} = 0$. We use the values $\text{AMP} \in \{0, 18, 36\}$ 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 $\text{PHZ} \in \{3, 9\}$ hr.

In all four panels of Figure 4, which shows the mean function for our selected values of the simulator-inputs, we see exactly the qualitative relation-

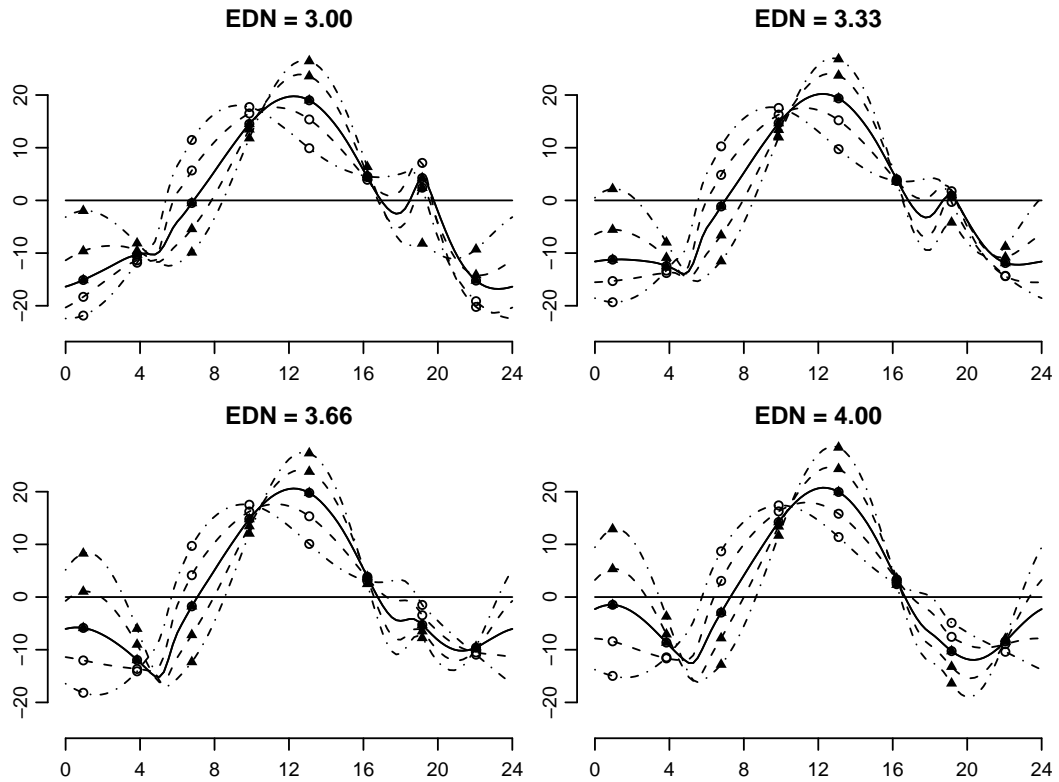


Figure 4: The simulator's response to different values of the three inputs (mean function, interpolated with a periodic B-spline). Line styles denote values of AMP: solid = 0, dashed = 18, dot-dashed = 36. Plotting characters denote values of PHZ: open circle = 3, filled triangle = 9. The two solid lines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

ship between AMP and PHZ that we anticipate. The two solid lines coincide as required ($\text{AMP} = 0$), and larger values of AMP are associated with a stronger response to PHZ. We can also see that the two values of PHZ give outputs that are close to having opposing phases.

Turning to EDN, the relationship uncovered here is entirely driven by the evaluations, as our prior for the effect of EDN was neutral. Our main findings are that higher EDN suppresses the evening excursion, and increases night-time drift. Both these findings are consistent with our qualitative physical understanding. The increase of 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 night-time electron density. Since the electric-potential drop along the night-time equator from dawn to dusk is basically determined by the day-side electrodynamics, where conductivities are much larger than at night, night-time processes have little effect on the integral of the eastward electric field along the night-time 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 night-time; this is an interaction between all four components—the three simulator inputs and the output index variable. The higher variability due to the tides during the night with increased E -region night-time electron density might be due to the fact that the tides are not propagating 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 which we also need the variance function. Uncertainty is shown in Figure 5, for $\text{EDN} = 3.66$. The uncertainties are much smaller than the signal, and it is clear that the mean function alone does a good job of 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 favoured 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. Suppose, for simplicity, that we are interested in the expected value of the vector $f(r)$ at some specified r . In this case

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

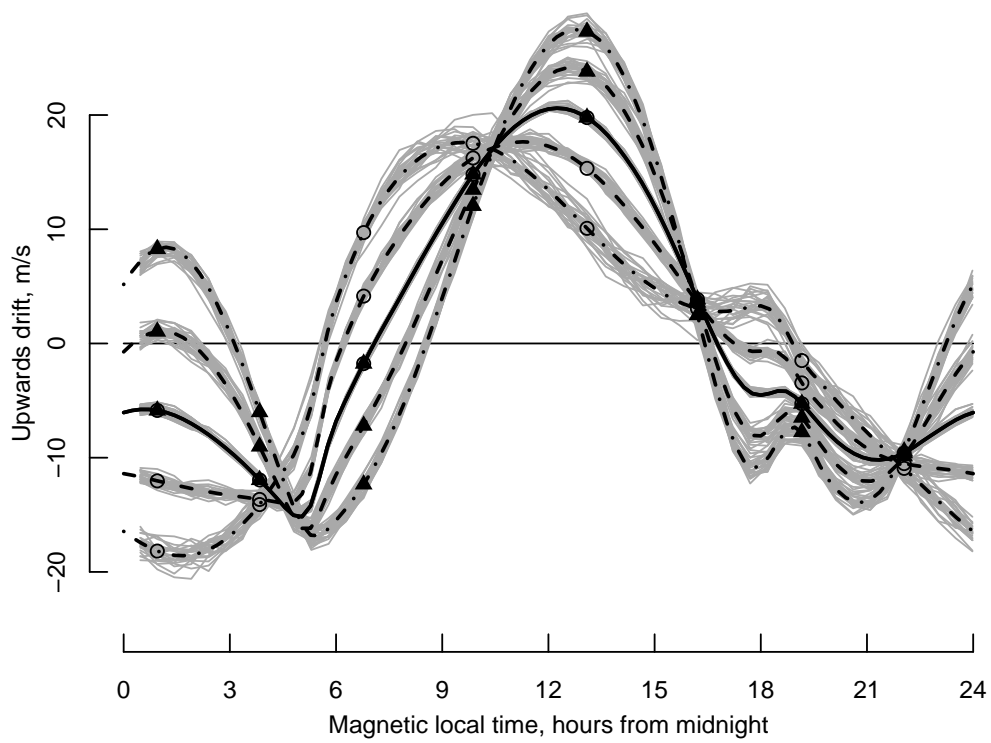


Figure 5: Effect of AMP and PHZ when $EDN = 3.66$, showing the uncertainty as 25 sampled values behind the mean function. See the caption to Figure 4 for details.

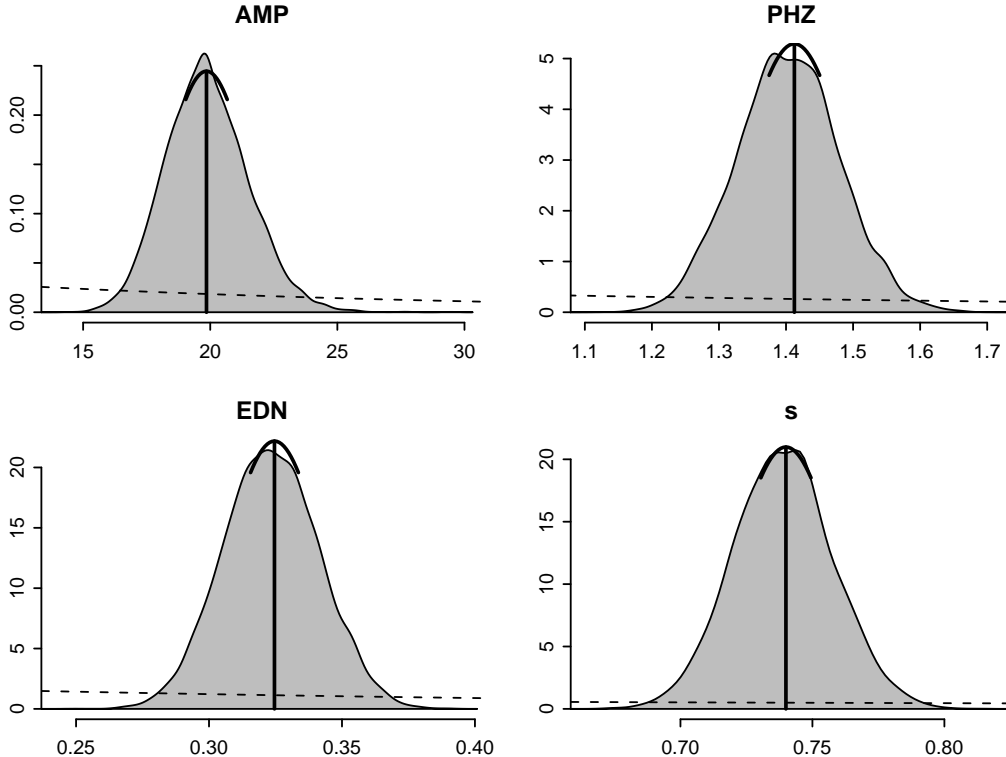


Figure 6: Posterior marginal probability density functions for the four residual correlation lengths. In each panel, the posterior is shown as a grey polygon, the corresponding part of the prior is shown as a dashed line, and the plug-in estimate from Table 1 is shown as an umbrella, \pm half an asymptotic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

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

$$\pi(\lambda | F) \propto \pi(F | \lambda) \pi(\lambda) \quad (14)$$

using MCMC: $\pi(F | \lambda)$ 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 one, i.e. an Exponential distribution). Figure 6 shows the marginal prior and posterior distributions for each of 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 only takes about ten minutes on a laptop 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 $\text{AMP} = 36$, $\text{PHZ} = 3$, and $\text{EDN} = 3.66$. There is no discernible difference between the two treatments. Numerically, the predictive standard deviations in the full Bayes

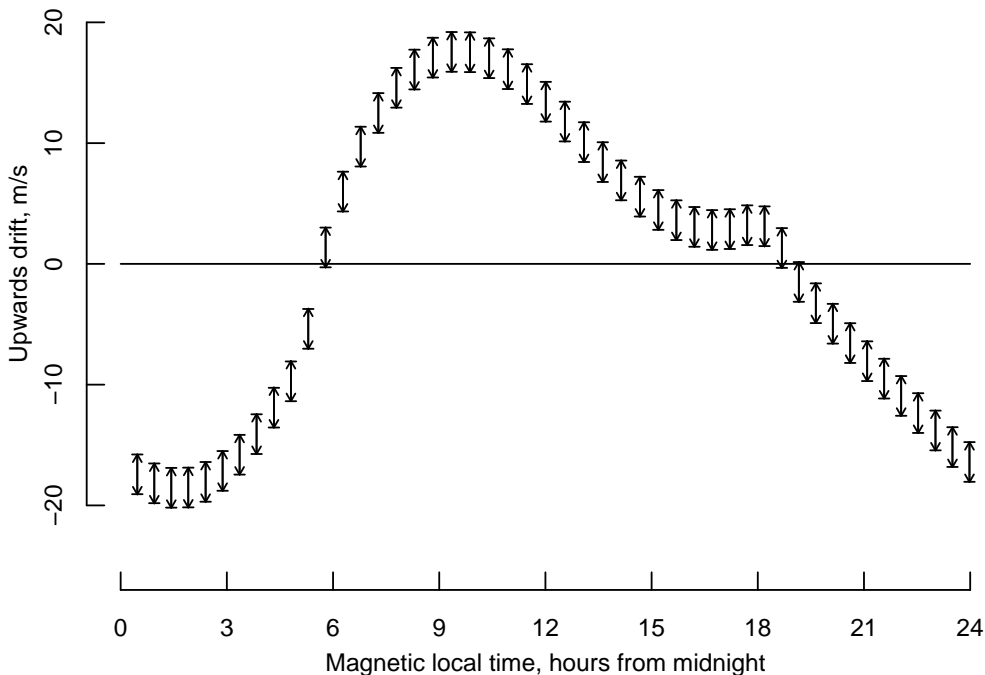


Figure 7: Predicted upwards drift at input values $\text{AMP} = 36$, $\text{PHZ} = 3$, and $\text{EDN} = 3.66$, shown as error bars for the mean \pm two standard deviations, on the simulator time-steps. There are two different treatments of the residual correlation lengths λ , plug-in and full-Bayes. 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.

case are about one 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 impacted on our choice of regressors and of the covariance function for the residual, and also, though to a lesser extent, on our specification of the emulator prior distribution. Some of these choices only really impact on our predictions 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 devote resources to eliciting expert knowledge about the simulator, and finding ways to incorporate that knowledge into the emulator.

In this paper we have used the Outer Product Emulator (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 might therefore be embedded in a hierarchical statistical framework in which we also learn about the hyperparameters, we have chosen a different approach. We have made 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), in order 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.

There are two points to make about our ‘lightweight’ approach. First, emulators of complex deterministic functions are very complicated statistical objects, and we *must* have diagnostic validation of our emulator, no matter how it is constructed. To date, the literature on emulators has been noticeably short on detailed diagnostic analysis (Bastos and O’Hagan, 2008, is an exception). This is concerning, 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 being 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 of choosing an emulator directly, than they can of 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, since he or she has a clear idea of how the emulator is to be used, and 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. Therefore we will need to construct thousands of emulators, since 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.

Acknowledgements

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. We would like to thank both of these institutions for their support.

NCAR is sponsored by the National Science Foundation (NSF). This material was based upon work supported by the NSF under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect

the views of the NSF.

References

- L.S. Bastos and A. O’Hagan. Diagnostics for gaussian process emulators. Technical Report No. 574/07, Department of Probability and Statistics, University of Sheffield, 2008. Currently available at http://mucm.group.shef.ac.uk/Pages/Downloads/Technical_Reports/08-02.pdf.
- S. Conti and A. O’Hagan. Bayesian emulation of complex multi-output and dynamic computer models. Technical Report No. 569/07, Department of Probability and Statistics, University of Sheffield, 2007. Currently available at <http://www.tonyohagan.co.uk/academic/ps/multioutput.ps>.
- 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.
- 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.
- D. Drignei, 2006. Empirical Bayesian analysis for high-dimensional computer output. *Technometrics*, **48**(2), 230–240.
- J.V. Eccles, 1998. Modeling investigation of the evening prereversal enhancement of the zonal electric field in the equatorial ionosphere. *J. Geophys. Res.*, **103**(26), 709–26.
- B.G. Fejer, E.R. de Paula, S.A. González, and R.F. Woodman, 1991. Average vertical and zonal F region plasma drifts over Jicamarca. *Journal of Geophysical Research*, **96**, 13901–13906.
- C. G. Fesen, G. Crowley, R. G. Roble, A. D. Richmond, and B. G. Fejer, 2000. Simulation of the pre-reversal enhancement in the low latitude vertical ion drifts. *Geophys. Res. Let.*, **27**(13), 1851.
- T. Gneiting, 1999. Correlation functions for atmospheric data analysis. *Q. J. R. Meteorol. Soc.*, **125**, 2449–2464.
- 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, 2006. Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association*, **101**, 1132–1143.

- M. Goldstein and J.C. Rougier, 2009. Reified Bayesian modelling and inference for physical systems. *Journal of Statistical Planning and Inference*, **139**, 1221–1239. With discussion.
- M.E. Hagan and J.M. Forbes, 2002a. Migrating and nonmigrating diurnal tides in the middle and upper atmosphere excited by tropospheric latent heat release. *J. Geophys. Res.*, **107**(D24), 4754.
- M.E. Hagan and J.M. Forbes, 2002b. Migrating and nonmigrating semidiurnal tides in the middle and upper atmosphere excited by tropospheric latent heat release. *J. Geophys. Res.*, **108**(A2), 1062.
- D. Higdon, J. Gattiker, B. Williams, and M. Rightley, 2008. Computer model calibration using high dimensional output. *Journal of the American Statistical Association*, **103**, 570–583.
- M.C. Kennedy and A. O’Hagan, 2001. Bayesian calibration of computer models. *Journal of the Royal Statistical Society, Series B*, **63**, 425–450. With discussion, pp. 450–464.
- 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.
- C. Linkletter, D. Bingham, N. Hengartner, D. Higdon, and K.Q. Ye, 2006. Variable selection for Gaussian process models in computer experiments. *Technometrics*, **48**(4), 478–490.
- 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, 2006. Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, **91**, 1290–1300.
- 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/>.
- C.E. Rasmussen and C.K.I. Williams, 2006. *Gaussian Processes for Machine Learning*. MIT Press. Available online at <http://www.GaussianProcess.org/gpml/>.
- A.D. Richmond, E.C. Ridley, and R.G. Roble, 1992. A thermosphere/ionosphere general circulation model with coupled electrodynamics. *Geophys. Res. Lett.*, **19**(6), 601.
- J.C. Rougier, 2008. Efficient emulators for multivariate deterministic functions. *Journal of Computational and Graphical Statistics*, **17**(4), 827–843.

- J.C Rougier and D.M.H. Sexton, 2007. Inference in ensemble experiments. *Philosophical Transactions of the Royal Society, Series A*, **365**, 2133–2143.
- B. Sansó, C. Forest, and D. Zantedeschi, 2008. Inferring climate system properties using a computer model. *Bayesian Analysis*, **3**(1), 1–38. With discussion, pp. 39–62.
- T.J. Santner, B.J. Williams, and W.I. Notz, 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.
- M.L. Stein, 1999. *Interpolation of Spatial Data: Some Theory for Kriging*. New York: Springer Verlag.
- W.C.M. van Beers and J.P.C. Kleijnen, 2008. Customized sequential designs for random simulation experiments: Kriging metamodeling and bootstrapping. *European Journal of Operational Research*, **186**(3), 1099–1113.
- A. Yaglom, 1987. *Correlation theory of stationary and related random functions*. Springer-Verlag.