Emulation and interpretation of high-dimensional climate model outputs

Journal:	*Journal of Applied Statistics*
Manuscript ID:	CJAS-2014-0362.R1
Manuscript Type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Holden, Philip; Open University, Environment, Earth and Ecosystems Edwards, Neil; Open University, Garthwaite, Paul; Open University, Wilkinson, Richard; Nottingham University,
Keywords:	climate modelling, coupled models, emulation, principal components, singular vector decomposition
2010 Mathematics Subject Classification:	46N30

Emulation and interpretation of high-dimensional climate model outputs

Philip B. Holden[1], **Neil R. Edwards**[1], **Paul H. Garthwaite**[2] **and Richard D. Wilkinson** [3]

[1] Earth, Environment and Ecosystems, Open University, UK

[2] Department of Mathematics and Statistics, Open University, UK

[3] School of Mathematical Sciences, University of Nottingham, UK

Acknowledgements: This work was funded under EU FP7 ERMITAGE grant number 265170.

Abstract: Running complex computer models can be expensive in computer time, while learning about the relationships between input and output variables can be difficult. An emulator is a fast approximation to a computationally expensive model that can be used as a surrogate for the model, to quantify uncertainty or to improve process understanding. Here, we examine emulators based on singular value decompositions and use them to emulate global climate and vegetation fields, examining how these fields are affected by changes in the Earth's orbit. The vegetation field may be emulated directly from the orbital variables, but an appealing alternative is to relate it to emulations of the climate fields, which involves high-dimensional input and output. The singular value decompositions radically reduce the dimensionality of the input and output spaces and are shown to clarify the relationships between them. The method could potentially be useful for any complex process with correlated, high-dimensional inputs and/or outputs

Key words: Climate modelling; coupled models; emulation; principal components; singular value decomposition.

Address for correspondence: Philip B. Holden, Environment, Earth and Ecosystems, Open University, Milton Keynes MK7 6JA, UK. E-mail: phlip.holden@open.ac.uk).

1 Introduction

Holden and Edwards [6] demonstrated a methodology for emulating high-dimensional climate outputs as a function of scalar model inputs. Their approach was to decompose the output of a perturbed parameter ensemble of climate model simulations using singular value decomposition and to regress the dimensionally reduced output onto the model input parameters. The methodology was developed for coupling climate models to climate change impact models in the case where the coupling variable from impact to climate model is low dimensional. The method has since been applied to a range of coupling applications [10,12,13]. Here we extend the approach of dimensionally reduced emulation to the case of high dimensional inputs, decomposing both input and output fields and emulating the relationship between the decomposed fields.

There are many classes of problems that would benefit from a statistical model that relates high-dimensional input to high-dimensional simulator output. Such a model may be useful when a simulator is too slow for a particular application or when the dynamics of the connecting process are not known *a priori*. In either case the technique could be applied either for the purposes of dynamical understanding or prediction. Some illustrative examples follow, with a specific focus here on predictive applications in climate science (although we note that potential applications are likely far more general).

1. In most climate coupling problems the two coupled models are required to exchange high-dimensional data in both directions. In the case when one of the models is significantly more expensive than the other, a statistical model (or emulator) of the expensive model would enable couplings that may otherwise be computationally prohibitive.

2. Climate forcing fields are often characterised by complex spatial patterns. Examples include aerosols (which modify both incoming solar radiation and outgoing planetary long wave radiation) and human land use change (which modifies energy and moisture transfer exchange between surface and atmosphere). High-dimensional forcing fields are particularly problematic for climate impact projections [17].

3. Integrated Assessment Models are tools that integrate environmental science and economic models to inform policy making. They are intrinsically defined by high-dimensional (regionally defined) inputs and outputs and so cannot be readily emulated by conventional techniques.

4. Hierarchical emulation techniques attempt to predict the outputs of a high complexity simulator from the outputs ("emergent properties") of a lower complexity simulator. An approach with high-dimensional inputs may be a useful alternative to existing approaches that perform the hierarchy from scalar emergent properties.

5. In the case where high-dimensional simulation data (e.g. climate) is related to high-dimensional observational data that cannot be robustly simulated (e.g. vegetation), the approach may allow improved predictions of future change in the latter.

Here we construct emulators and consider a sixth application: determining statistical relationships between inputs and high-dimensional outputs in order to understand model behaviour and hence gain insight into real world behaviour. In the problem motivating the work reported here, changes in climate impact on some other quantity and changes in that quantity, in turn, impact on the climate. Specifically, we suppose that the latter quantity is the distribution of vegetation over Earth, modelled by a land surface vegetation model, and there is also a climate model. Appropriate climate variables (temperature and precipitation) are passed as fields to the vegetation model. Vegetation-dependent outputs (surface albedos, soil moisture storage capacity and surface roughness) are passed as fields back to the climate model. Climate and local vegetation are thus inextricably linked: vegetation is determined by climate and climate is strongly dependent upon the characteristics of the local vegetation.

Various factors have a marked influence on climate and vegetation, and we here examine the effects of changes in the Earth's orbit. These play an important role in driving climate change on time scales of 10,000 to 100,000 years. They are accepted to be the fundamental drivers of the cyclical "glacial-interglacial" climate observed over the last few million years [2]. A coupled climate-vegetation model was run for a number of simulations for different choices of Earth's orbit. A focus of this paper is on ways to interpret the output from the simulations. Three relationships are of interest: (i) orbit-climate, (ii) orbit-vegetation, and (iii) climate-vegetation. The approach we adopt is to emulate the coupled model using principal components and then examine relationships based on the principal components.

In Section 2 we describe the orbital parameters, the models and the simulation study. In Section 3 we emulate the orbit-climate and orbit-vegetation relationships and explore these relationships. In Section 4 we emulate and examine the climate-vegetation relationships. Sections 3 and 4 yield two means of emulating vegetation, which we shall refer to as one-step and two-step procedures. In the

one-step procedure (Section 3) the vegetation field is related directly to the orbital parameters while in the two-step procedure (Section 4) the climate fields are emulated from the orbital variables during the first step, and in the second step the vegetation field is emulated from the climate fields. Thus the second step involves both high-dimensional input fields and high-dimensional output fields. This second emulation approach, which addresses the case of high-dimensional input, has not previously been reported in the statistical literature. We compare the relationships used in the two procedures in Section 5 and also examine performance of the emulators through cross-validation. In Section 6 we briefly consider the use of Gaussian process models for performing parts of the emulation. An overview and concluding comments are given in Section 7.

2 Models and simulations

The climate model that was used in the simulations is the PLASIM-ENTS model [8]. It comprises the Planet Simulator [4] coupled to the terrestrial carbon model ENTS [19]. The 3D atmospheric dynamics are based on underlying primitive equations (Newton's laws of motion), run here at grid cell (64×32) resolution with ten levels in the vertical dimension. From here we focus on the 2048 grid cells that cover the Earth's surface. Physical processes being modelled include the Sun's radiation and the Earth's thermal radiation, driving 3D motion, the formation of clouds, and convective and large-scale precipitation. The ocean and sea ice are modelled as flux-corrected slabs with no explicit dynamics. Interpolated monthly-averaged ocean heat and sea-ice flux corrections, diagnosed from a simulation with modern-day orbit, are applied. This approximates to fixing the large-scale ocean circulation, but the atmosphere and ocean slab are coupled, so that local orbital-change induced atmosphere-ocean interactions are captured.

In the ENTS model, all vegetation is grouped together as a single quantity. A double-peaked temperature response function is used to capture the different responses of vegetation at low (tropical) latitudes and at high latitudes (towards the poles). Photosynthesis is a function of temperature, soil moisture availability, atmospheric CO_2 concentration and fractional vegetation cover. The simulated vegetation values affect the land surface characteristics (albedo, surface roughness length and moisture bucket capacity) that are needed to determine the climate.

Three variables together describe the configuration of the Earth's orbit around the sun: eccentricity, obliquity and the longitude of the perihelion (the angular position of the Earth in its orbit around

the Sun) at the vernal equinox. They are given a fixed set of values in each simulation and treated as parameters in the climate model. Over the ensemble of simulations, their ranges of values approximately span the values that these orbital variables have taken over the last million years [1].

- Obliquity (X_1) describes the tilt of the Earth with respect to the plane of its orbit. It is this tilt that leads to the seasons: during the period of the year when the northern hemisphere is pointing towards the sun, more incoming solar radiation is received and the days are longer, so the Northern Hemisphere experiences summer. Increased obliquity leads to more pronounced seasonal contrast, especially at high latitudes where the seasons are in general more pronounced. Obliquity was varied between 22° and 25° in the ensemble.

- Eccentricity (X_2) describes the shape of the orbital path. An eccentricity of 0 describes a perfectly circular orbit. Eccentricity was varied between 0 and 0.05 in the ensemble.

- Longitude of the perihelion at the vernal equinox, hereafter referred to as precession X_3, defines one of the two points in the orbit (in spring) when the tilt of the Earth is inclined neither towards nor away from the Sun. Precession was varied across all "solar longitudes" (the angular position of the Earth in its orbit around the Sun). Precession controls where in the orbit the seasons occur and, in conjunction with eccentricity, changes the relative insolation received by the two hemispheres. For instance, if northern summer coincides with the part of the orbit when the Earth is closest to the Sun, northern summers will be warmer than southern summers (when the Earth is most distant from the Sun).

In climate modelling, one common practice when designing an ensemble of simulations is to use a maximin Latin hypercube design in which the variables of interest are varied uniformly over their ranges. This maximizes the minimum distance between design points and ensures the design fills the input space. In the present case there are only three variables of interest (X_1, X_2 and X_3), so the design reduces to a Latin square. An ensemble of 50 simulations was formed by partitioning the range of each variable into 50 intervals of equal length. Taking a point at random from each interval gave 50 'treatment levels' for each variable; a 50×50 Latin square with a 50-level factor was constructed using the maximinLHS function of the *lhs* package in *R* [14].

A PLASIM-ENTS configuration is determined by the settings of many 100's of model parameters. These include switches (which determine the precise numerical schemes applied), physical constants

that are approximately known but vary spatially in the real world (such as the reflectivity of ice) and parameterisations of "sub-gridscale" processes such as cloud formation, which have "tuned" values that are known to result in reasonable model behaviour. All model parameters were set at their defaults (PLASIM Version 6 Revision 4, ENTS parameters [19]). The threshold fractional soil moisture for photosynthesis [8] was set at 0.1.

Each simulation modelled a period of 100 years, starting from a 'dead' planet with no vegetation or rain. Vegetation and climate were coupled at every 45 minute time step, when spatial fields of surface air temperature, precipitation and evaporation were passed from the climate model to the vegetation model, and spatial fields of surface roughness, soil moisture content and albedo were passed from the vegetation model to the climate model.

3 Interpretation of the climate fields and vegetation field

There are several quantities of interest whose relationships we wish to investigate. The independent astronomical forcing variables, denoted x, drive the variation in the other quantities. The outputs of the simulator are three spatial fields resolved onto a grid of 64×32 points on the Earth's surface. They are the annual average surface air temperature, denoted y_1, the annual average precipitation, y_2, and the annual average vegetation carbon density, y_3. We assume that y_1, y_2 and y_3 are functions of x, and additionally that y_3 is a function of y_1 and y_2. Our aim is to build an emulator of the simulator. This is a cheap statistical model approximating the three mappings above.

For each of the 50 simulations there are 2048 data points describing the spatial distributions of the output for each climate field, and for the vegetation field there are 471 data points (only grid cells over ice-free land give data points). Each climate field was used to form a 2048×50 matrix, which we denote by $\mathbf{Y}_1$ for the annual average surface air temperature, and $\mathbf{Y}_2$ for annual averaged precipitation. The 471×50 data matrix for the annual average vegetation carbon density is denoted $\mathbf{Y}_3$. The high dimensionality of these fields makes modelling difficult. However, we can exploit the correlation structure in the spatial fields to produce reduced-rank approximations. We use the singular value decomposition (SVD) of the output matrices, keeping only the most important terms in the decomposition, to reduce the dimension of the problem from 2048 dimensions to fewer than 10 dimensions. We use the SVD as this gives the best low-rank approximation as measured by the Frobenius norm. For $i = 1, 2, 3$, let $\tilde{\mathbf{Y}}_i$

denote the row-centred matrices, so that each row of $\tilde{\mathbf{Y}}_i$ has an average of zero. The SVD of $\tilde{\mathbf{Y}}_i$ is

$$\tilde{\mathbf{Y}}_i = \mathbf{L}_i \mathbf{D}_i \mathbf{R}_i' \tag{1}$$

where $\mathbf{L}_i$ is the matrix of left singular vectors of $\tilde{\mathbf{Y}}_i$, $\mathbf{D}_i$ is the 50×50 diagonal matrix of singular values of $\tilde{\mathbf{Y}}_i$ and $\mathbf{R}_i$ is the 50×50 matrix of right singular vectors. $\mathbf{L}_1$ and $\mathbf{L}_2$ are 2048×50 matrices and $\mathbf{L}_3$ is 471×50. We assume the singular values have been ordered so that $d_{i1} \geq d_{i2} \cdots \geq d_{i50}$, where d_{ij} is the jth diagonal element of $\mathbf{D}_i$ (i=1,2,3).

3.1 One-step emulator

We will be modelling from the orbital variables to columns of $\mathbf{R}_i$. Let $\mathbf{l}_{ij}$ and $\mathbf{r}_{ij}$ denote the jth columns of $\mathbf{L}_i$ and $\mathbf{R}_i$ ($i = 1, 2, 3$; $j = 1, \ldots, 50$), respectively. Then $\mathbf{l}_{ij}$ and $\mathbf{r}_{ij}$ are the jth eigenvectors of $\tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_i'$ and $\tilde{\mathbf{Y}}_i' \tilde{\mathbf{Y}}_i$, respectively. Also, d_{ij}^2 is the jth eigenvalue of both $\tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_i'$ and $\tilde{\mathbf{Y}}_i' \tilde{\mathbf{Y}}_i$. We shall refer to the $\mathbf{l}_{ij}$ ($i = 1, 2, 3$; $j = 1, \ldots, 50$) as principal components; then $\mathbf{r}_{ij}$ is commonly referred to as the *score vector* of $\mathbf{l}_{ij}$. To reduce dimensionality in order to make modelling easier, it is natural to ignore small eigenvalues. Suppose all but the first k eigenvalues and eigenvectors are ignored. Put $\mathbf{L}_{i*} = (\mathbf{l}_{i1}, \ldots, \mathbf{l}_{ik})$, $\mathbf{R}_{i*} = (\mathbf{r}_{i1}, \ldots, \mathbf{r}_{ik})$ and let $\mathbf{D}_{i*}$ be the $k \times k$ diagonal matrix with $d_{i1}, \ldots, d_{ik}$ as its diagonal elements. Then $\tilde{\mathbf{Y}}_i \simeq \mathbf{L}_{i*} \mathbf{D}_{i*} \mathbf{R}_{i*}'$.

To simplify explanation, suppose the temperature field is to be emulated, so $i = 1$. If $\mathbf{R}_{1*}' = (\mathbf{t}_1, \ldots, \mathbf{t}_{50})$, then $\mathbf{L}_{1*} \mathbf{D}_{1*} \mathbf{t}_1$ approximately equals the (centred) temperature values in the 2048 grid cells for the first simulation, $\mathbf{L}_{1*} \mathbf{D}_{1*} \mathbf{t}_2$ those for the second simulation, $\mathbf{L}_{1*} \mathbf{D}_{1*} \mathbf{t}_3$ those for the third simulation, and so on. (If k were set equal to 50, so that no eigenvectors were ignored, then $\mathbf{L}_* \mathbf{D}_* \mathbf{t}_j$ would exactly equal the centred temperature values for the jth simulation.) The key to the one-step emulator is to determine a relationship between an arbitrary score vector $\boldsymbol{\rho}$ and the orbital variables X_1, X_2, X_3, where $\boldsymbol{\rho}$ takes, in the case of temperature, the values $\mathbf{t}_1, \ldots \mathbf{t}_{50}$ in the 50 simulations. Then given a new set of values for these variables, the corresponding value of $\boldsymbol{\rho}$ can be estimated and $\mathbf{L}_{1*} \mathbf{D}_{1*} \boldsymbol{\rho}$ is the emulated 2048×1 vector of (centred) temperatures. This drastically reduces the dimensionality of the estimation problem, only the k-dimensional vector $\boldsymbol{\rho}$ must be estimated as $\mathbf{L}_{1*}$ and $\mathbf{D}_{1*}$ are unchanged. Also, understanding the relationship between the orbital variables and the dominant elements of $\boldsymbol{\rho}$ captures the relationship between these variables and the temperature field.

Let $\boldsymbol{\rho}_1$, $\boldsymbol{\rho}_2$ and $\boldsymbol{\rho}_3$ denote the vector $\boldsymbol{\rho}$ for the temperature, precipitation and vegetation fields,

respectively, and let ρ_{ij} denote the jth component of $\boldsymbol{\rho}_i$. The values taken by ρ_{ij} in the 50 simulations are the elements of $\mathbf{r}_{ij}$. Least squares regression is used to obtain an equation for estimating ρ_{ij} from the three orbital variables. To aid subsequent interpretation, we first normalise each of these variable onto the range -1 to +1. Let $\tilde{X}_1$, $\tilde{X}_2$, and $\tilde{X}_3$ denote the normalised eccentricity, obliquity and longitude variables, respectively, and let $\mathbf{x} = (\tilde{x_1}, \tilde{x_2}, \tilde{x_3})$. To build accurate regression models it is useful to introduce the idea of a feature map. The features of $\mathbf{x}$ denoted $\phi(\mathbf{x})$, are transformations of $\mathbf{x}$ that help us to build accurate emulators. For example, because x_3 is periodic, we find that including $\sin x_3$ and $\cos x_3$ in the set of features leads to large improvements in predictive accuracy of the emulators. Following [7], we also include the linear, quadratic and cross-product terms as explanatory variables in the set of features, $\phi_1 = x_1$, $\phi_2 = x_2$, $\phi_3 = \sin x_3$, $\phi_4 = \cos x_3$. i.e. :

$$E[\rho_{ij} \,|\, (\tilde{\Phi}_1, \tilde{\Phi}_2, \tilde{\Phi}_3, \tilde{\Phi}_4) = (\tilde{\phi}_1, \tilde{\phi}_2, \tilde{\phi}_3, \tilde{\phi}_4)] = \mu_{(ij)} + \sum_{p=1}^{4} \alpha_{(ij)p} \tilde{\phi}_p + \sum_{p=1}^{4} \beta_{(ij)p} \tilde{\phi}_p^2 + \sum_{q=p+1}^{4} \sum_{p=1}^{3} \gamma_{(ij)pq} \tilde{\phi}_p \tilde{\phi}_q, \quad (2)$$

for $i = 1, \ldots, 3$; $j = 1, \ldots, k$. Features of $\mathbf{x}$ are then progressively added and dropped using stepwise selection (using the `stepAIC` function in R [14]), in order to maximise the Akaike Information Criterion. The resulting regression models are then pruned to satisfy the more stringent Bayes Information Criterion. This procedure of first growing the model beyond the BIC constraint and then pruning is an attempt to help avoid local maxima in the stepwise search. Alternative term selection strategies of Lasso [18] and elastic net [20] were not found to offer any improvement over stepwise selection.

In section 6 we discuss using Gaussian process models rather than linear regression. There the selection of features of x is automatic, but at the cost of losing interpretability of the models.

3.2 The main principal components

In this subsection we examine the principal components for the temperature, precipitation and vegetation fields and in Section 3.3 we examine their score vectors. As noted earlier, $d_{i1}^2, \ldots, d_{i50}^2$ are the ordered eigenvalues of both $\tilde{\mathbf{Y}}_i \tilde{\mathbf{Y}}_i'$ and $\tilde{\mathbf{Y}}_i' \tilde{\mathbf{Y}}_i$. Examining these eigenvalues shows that 99% of the variation in temperature (across both grid cells and simulations) is explained by the first ten eigenvectors. The corresponding proportions for precipitation and vegetation carbon are 88% and 96%, respectively. In constructing the emulators, for each field we consider the score vectors of just the first 10 principal components (i.e. we set $k = 10$ for each field). The model is pruned ($k = 6$) in Section 5 where we

evaluate the performance of the emulator as components are progressively added.

The top row of Figure 1 plots the first three principal components of the temperature field ($\mathbf{l}_{11}$, $\mathbf{l}_{12}$ and $\mathbf{l}_{13}$) against geographic location. The top-left plot ($\mathbf{l}_{11}$) shows that the first principal component of temperature varies mostly with latitude, with particularly large (absolute) values in high northern latitudes. The second row of Figure 1 plots the first three principal components of the precipitation field ($\mathbf{l}_{21}$, $\mathbf{l}_{22}$ and $\mathbf{l}_{23}$), and the third row gives those for the vegetation field ($\mathbf{l}_{31}$, $\mathbf{l}_{32}$ and $\mathbf{l}_{33}$). The first component of precipitation is associated with largest values at low latitudes (in contrast to temperature). The first component of vegetation, assumed driven by changes in temperature and precipitation, exhibits significant variability at all latitudes. There are similarities across the three fields – for example, the three fields all show a difference between the northern and southern parts of South America. However, the extent of the similarities is quite limited.

3.3 The main score vectors

For each climate field and the vegetation field, we focus on the score vectors of the first three principal components and examine the regression equations (of the form given by equation (2)) that predict their values from the orbital variables.

To examine the output from a regression, Homma and Saltelli (1996) introduce a *main effect index*. This provides a measure of the variation in ρ_{ij} that is associated with the individual explanatory terms. We describe this index in Appendix 1 and plot the main effect indices for each regression in Figure 2.

The top diagram in Figure 2 relates to temperature. The first principal component of temperature, which explains 86% of the variance in temperature across the ensemble of simulations, has a score that is dominated by (and inversely correlated with) obliquity. Obliquity exerts a strong control on the degree of seasonality, especially at high latitudes, and this response is consistent with the spatial distribution of that component (Figure 1).

The second principal component of temperature, which explains 11% of the ensemble variance, exhibits a more complex relationship, with a score that has dependencies upon all three orbital parameters. However, the strongest predictor of its scores is precession, with significant interactions between precession and the other two orbital parameters. As discussed in Section 2, the interaction between eccentricity and precession exerts a control on the relative strength of seasonality in the two hemispheres.

The phasing of this effect changes over time (the "precession of the equinoxes"). Eccentricity controls the strength of this effect while precession controls the phasing. This dependence on precession and eccentricity, combined with the inter-hemispheric contrast that is apparent in the spatial distribution of the component, suggests that the second principal component is dominantly an expression of this effect. We do not attempt to explain the third principal component, which describes less than 1% of the ensemble variance.

The middle diagram in Figure 2 relates to precipitation. The first principal component explains 35% of the variance in precipitation across the ensemble of simulations. The estimation of scores for the first principal component is, as with temperature, mainly controlled by obliquity, suggesting that this principal component is also dominantly an expression of the strength of obliquity-driven seasonality. This largely results from a strengthening of the SE Asian and West African monsoon systems as obliquity increases. The second principal component of precipitation explains 26% of the ensemble variance and is driven mainly by obliquity and precession. The third principal component explains 13% of the ensemble variance and is driven by precession and the interaction between precession and eccentricity.

The bottom diagram in Figure 2 relates to vegetation. In Section 4 the vegetation field is emulated from the temperature and precipitation fields. Hence, we are also interested in the relationship between the primary vegetation principal components and those of temperature and precipitation, as well as between the vegetation principal components and the orbital variables.

The first principal component of vegetation explains 65% of the ensemble variance. The emulation of this component is dominated by obliquity. The correlation of the 471 data points that comprise the first principal component scores for vegetation with the corresponding data points (those on ice-free land) that comprise the first principal component scores for temperature (+0.63) and precipitation(-0.68), suggests that both climate variables are comparably important in driving the obliquity-driven variability in vegetation. The spatial patterns of the components (Figure 1) suggests that obliquity-driven temperature changes mainly drive high latitude vegetation change whereas obliquity-driven precipitation is, for instance, responsible for vegetation change in South East Asia and eastern USA. The second principal component of vegetation explains 14% of the ensemble variance and is driven by precession and its interaction with eccentricity, suggesting that it is related to the third principal component of precipitation. The similarities in the spatial patterns of these components (Figure 1) reinforces this interpretation. The third principal component of vegetation explains 7% of the ensemble variance and,

like the second component, is also driven by precession and its interaction with eccentricity. Although the main effect indices are similar for the second and third components the functional forms are quite different. The second component is controlled by the sine of precession, whereas the third component is controlled by its cosine.

4 Emulating vegetation from climate input fields

4.1 Two-step emulator

The effect of the orbital variables on vegetation carbon is largely through their effect on temperature and precipitation. Here we first consider emulation of the vegetation field from these climate fields. This involves input and output fields that are both high-dimensional. To reduce dimensionality, we perform the same singular value decompositions as in Section 3, putting $\tilde{\mathbf{Y}}_i = \mathbf{L}_i\mathbf{D}_i\mathbf{R}_i'$ for $i = 1, 2, 3$. To further reduce dimensionality, we then discard score vectors that correspond to small eigenvalues retaining, as in Section 3, only score vectors of the ten largest eigenvalues.

We then relate the matrix score vectors for vegetation ($\mathbf{R}_3$) to those for temperature and precipitation ($\mathbf{R}_1$ and $\mathbf{R}_2$). Specifically, a linear regression is formed for each of the variables ρ_{3j} ($j = 1, \ldots, 10$):

$$E(\rho_{3j}) = a_{(j)} + \sum_{p=1}^{10} b_{(j)p}\rho_{1p} + \sum_{p=1}^{10} c_{(j)p}\rho_{2p}. \tag{3}$$

The regression model containing all 21 terms is then pruned to satisfy the Bayes Information Criterion. As noted earlier, the values taken by ρ_{ij} in the 50 simulations are the elements of the score vector $\mathbf{r}_{ij}$ ($i = 1, 2, 3$; $j = 1, \ldots, 10$). In subsection 4.2 we examine the coefficients of these equations to learn about the relationships between the vegetation carbon field and the climate fields.

The equations are also used in a two-step emulator to relate the orbital parameters to the vegetation field. Given a new set of values for the orbital parameters, X_1, X_2 and X_3, the first step uses the regressions given by equation (2) to estimate the values taken by ρ_{1p} and ρ_{2p} ($p = 1, \ldots, 10$). The second step puts these values into the equations given by (3), which yields an estimate of $\boldsymbol{\rho}_3$ for the given setting of the orbital parameters. The emulation of the vegetation field is then $\mathbf{L}_{3*}\mathbf{D}_{3*}\boldsymbol{\rho}_3$, where $\mathbf{L}_{3*}$ and $\mathbf{D}_{3*}$ are 471×10 and 10×10 matrices.

4.2 Relationships between vegetation and climate fields

Each of the regression models given in equation (3), which comprise from 6 to 10 regression terms, capture a very high proportion of the variation in the dependent variable, especially for the first five score vectors, for each of which R^2 exceeded 90%. These five vectors together explain 92% of the ensemble variance.

The regression coefficients are plotted in Figure 3. (Consideration of main effects indices is unnecessary because no quadratic terms are contained in these emulators, unlike the situation in Section 3). The emulator of the first component of vegetation is dominated by the first component coefficients of temperature and precipitation, consistent with previous inferences. An initially surprising result is that the second element of $\boldsymbol{\rho}_3$ is dominantly a function of the first score vectors of temperature and precipitation. This is unexpected because we know, from Section 3.3, that these score vectors are both strong functions of obliquity, whereas the second score vector for vegetation is a function of eccentricity and precession. The explanation is that the score vectors for temperature exhibit some strong correlations with those for precipitation. (The correlation between the first score vectors is 0.79). It is revealing that if we exclude the first score vectors for temperature and precipitation from the emulator of ρ_{32} (equation (3)), quite different emulator coefficients are apparent although the model fit R^2 (97%) remains very high. The largest terms in the revised emulator of ρ_{32} are ρ_{23} (0.75), ρ_{12} (0.30) and ρ_{13} (-0.30), now consistent with the inferences of Section 3.3, being that the second component of vegetation variability is dominantly controlled by the third component of precipitation variability. Interpretation of these emulators can be less straightforward than was the case when the orbital variables were the explanatory variables (Section 3) because the orbital variables are uncorrelated through the design of the simulation study.

5 Emulator comparison and performance

The one-step and two-step emulators provide two routes to emulate the vegetation output field from orbital variables:

1. Estimate the $\boldsymbol{\rho}$-vector for vegetation ($\boldsymbol{\rho}_3$) directly from the orbital variables and take $\mathbf{L}_{3*}\mathbf{D}_{3*}\boldsymbol{\rho}_3$ as the emulated vegetation field (one-step emulator).

2. Estimate the $\boldsymbol{\rho}$-vector for temperature ($\boldsymbol{\rho}_1$) and $\boldsymbol{\rho}$-vector for precipitation ($\boldsymbol{\rho}_2$) from the orbital

variables. Then estimate $\boldsymbol{\rho}_3$ from $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$, again taking $\mathbf{L}_{3*}\mathbf{D}_{3*}\boldsymbol{\rho}_3$ as the emulated vegetation field (two-step emulator).

The two methods give very similar emulations of the vegetation field. Moreover, the methods give very similar estimators of individual elements of $\boldsymbol{\rho}_3$, especially with those elements that correspond to the larger eigenvalues.

To illustrate this latter point, Figure 4 plots the coefficients of the regression equations that estimate the *first* element of $\boldsymbol{\rho}_3$ (which corresponds to the largest eigenvalue) from the orbital variables. The left-hand (yellow/lighter) bar of each pair correspond to the coefficients given by the one-step emulator (from equation (2)) and the right-hand (red/darker) bars correspond to those given by the two-step emulator. Coefficients for the two-step emulator are obtained by combining the equations that give $\boldsymbol{\rho}_3$ from $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ with the equations that give $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ from the orbital variables. It can be seen the left-hand and right-hand bars in each pair are very similar.

Figure 5 examines the closeness between the two emulators for each of the 10 elements of $\boldsymbol{\rho}_3$. For each element the emulators gave 50 values and the highest (blue) line in Figure 5 plots the correlations between them. The correlation is above 0.94 for each of the first five elements. The gaps in the data for components 7 and 10 are because no 1-step vegetation emulator terms were found to satisfy the BIC requirement, suggesting the vegetation emulator should not include components $k > 6$. However, the high correlations for all components with $k = 6$ results in close agreement between the one-step and two-step emulators in their estimates of $\boldsymbol{\rho}_3$.

The other lines in Figure 5 show the correlations between emulated values and the simulation values given by the full climate-vegetation model. (The jth score vector, $\mathbf{r}_{3j}$, holds the simulation values for the jth element of $\boldsymbol{\rho}_3$.) The correlations are high for the components corresponding to the four largest eigenvalues, but correlations corresponding to most of the smaller eigenvalues are distinctly poorer. This is true of both emulators, though for small eigenvalues the one step emulator (red/middle line) gives slightly highly correlations with the simulated values than the two-step emulator (green/lower line).

Leave-one-out cross validation was used to evaluate the performance of emulators more critically and in slightly greater breadth. Four emulations were examined: the three one-step emulations – from orbital variables to temperature, precipitation and temperature – and the two-step emulation of vegetation. Each of the 50 simulations was omitted in turn and the four emulators built from the

remaining 49 simulations, using the methods described in Sections 3 and 4. The emulators were used to estimate the climate and vegetation fields in the simulation that had been omitted, using the setting of the orbital variables in that simulation. Two such cross-validations given by the two-step emulator with $k = 6$ are illustrated in Figure 6, randomly selected as being the first and last members of the Latin Hypercube ensemble. The upper diagrams map the vegetation levels to be estimated, and the lower diagrams map the estimates given by the two-step emulator in cross-validation. The main features in the upper maps are captured by the emulator.

The sum of squared errors was used to form a quantitative measure of cross-validated model performance. For temperature, for example, $\tilde{\mathbf{Y}}_1$ is the 2048×50 matrix of values given by the PLASIM-ENTS model after centring each row to have a mean of zero. Let $\tilde{y}_{(1)jk}$ denote an element of this matrix and let $\widehat{y}_{(1)jk}$ denote its estimated value when the kth simulation was omitted in building the emulator. Then the sum of squared errors for the emulator is

$$\sum_{j=1}^{2048} \sum_{k=1}^{50} \left(\tilde{y}_{(1)jk} - \widehat{y}_{(1)jk} \right)^2$$

while $\sum_{j=1}^{2048} \sum_{k=1}^{50} (\tilde{y}_{(1)jk})^2$ is the corrected total sum of squares for $\tilde{\mathbf{Y}}_1$ and measures the variablity in each grid cell over the 50 simulations. Hence, the proportion of variation explained by the emulator is

$$\left[\sum_{j=1}^{2048} \sum_{k=1}^{50} (\tilde{y}_{(1)jk})^2 - \sum_{j=1}^{2048} \sum_{k=1}^{50} \left(\tilde{y}_{(1)jk} - \widehat{y}_{(1)jk} \right)^2 \right] \Bigg/ \sum_{j=1}^{2048} \sum_{k=1}^{50} (\tilde{y}_{(1)jk})^2 \, . \tag{4}$$

Figure 7 (top panel) illustrates the cross-validated performances as components are progressively added to the models. This data reinforces the choice to include only components with $k = 6$. Although improvements are modest beyond $k = 4$ they are apparent in all models. For the temperature field, the proportion of variation explained by the one-step emulator was 85.6%. Corresponding figures for precipitation and vegetation were 70.2% and 81.2%. The two-step emulator explained 80.3% of the variation in the vegetation field, very similar to the one-step emulator. Hence, each emulator is a good approximation to the simulator.

We note that the performance of the two-step emulator improved to 81.1% when the second step emulator was also allowed to use x, comparable to the one-step emulator performance of 81.2%. It is unsurprising that the one-step emulator is the better model as the combination of two linear functions (in the two-step emulator) is still linear but with the statistical cost of estimating more parameters. This observation contrasts with the non-linear GP emulators considered in the following Section 6.

The similar performances of the one-step and two-step vegetation emulators suggest that the second step may not be contributing significantly to emulator error. This was tested by projecting the left-out simulated fields of temperature and precipitation onto the relevant principal components and applying the resulting scores to equation (3), enabling us to quantify the proportion of simulated variance explained by the second-step emulator. This is different because the emulated vegetation fields for simulated climate input may be closer to the vegetation simulations than those corresponding to emulated climate input. The result was 87.3%, as compared with 80.3% for the two-step emulation from the orbital variables. Physically, a likely explanation for the difference in accuracy between the two emulation steps is that the climate model contains a representation of chaotic atmospheric dynamics which are intrinsically challenging to predict, while the vegetation model used here has no such explicitly unstable or stochastic elements that might limit its predictability in terms of its climate inputs. Thus although the second step involves high-dimensional inputs and outputs, while the first step involves only high-dimensional output, for this particular combination of models the first step is more challenging to emulate accurately. It should be noted that our vegetation and climate outputs come from a coupled modelling system in which both sets of fields are always present. The second step emulator is therefore primarily a tool for understanding rather than prediction in this particular modelling context. Nevertheless, the results show the viability of using principal component emulation to relate high-dimensional input and output fields of a given process (in this case a climate model). Furthermore, as an approximation to the true relationship between equilibrium climate and vegetation fields in the real world, the second-step emulator could be applied as a predictive tool for future or past situations where only climate projections were available.

Another variation was also considered for the second step of the two-step model. For that step, the explanatory variables are ten temperature score vectors and ten precipitation score vectors. While the ten vectors within each set are orthogonal, the score vectors for temperature are correlated with those for precipitation. This made it a little more difficult to interpret their relationship with vegetation, as noted in Section 4.2. To avoid correlations between the explanatory variables, a possibility that was examined was to combine the temperature and precipitation fields and construct a single set of orthogonal score vectors from the combined field. Performance was very similar when judged by cross-validation. The proportion of variance explained by the 2-step emulation was 79.9%, although more significant degradation was noted when the 2nd step was considered in isolation, reducing from 87.4% to

85.7%. The possibility of over-fitting as a result of the correlations was also considered. To explore this, four alternative models were built, omitting either the temperature (ρ_{11}) or precipitation (ρ_{21}) terms from each of the regression equations for vegetation (ρ_{31} and ρ_{32}). In each case though, performance became marginally poorer.

6 Gaussian process emulation

An alternative to using linear regression is to model the relationship between $\mathbf{x}$ and ρ using Gaussian processes (GPs) [15]. GPs are non-parametric models that are commonly used in the computer experiment literature to build emulators of computer simulators [11,16]. In linear regression the key to predictive accuracy is the selection of the features $\phi(\mathbf{x})$. In GP regression, this choice is less important, and instead, it is the choice of covariance function (or kernel) κ that determines performance. For a given covariance function, a Gaussian process can be interpreted as doing (penalised) linear regression in an infinite dimensional Hilbert space of features. The particular choice of covariance function determines what that basis will be.

The main advantage of using GPs is that less thought needs to be given to choosing a good set of features of $\mathbf{x}$. So for example, using just $\mathbf{x}$ as the input, we can achieve a predictive accuracy of $R^2 = 82.4\%$ for a one-stage emulator (mapping from $\mathbf{x}$ directly to the vegetation field y_3). This compares with an accuracy of 81.2% for the linear regression model found using stepwise regression (on the *features* of $\mathbf{x}$). The GP did not need to be told that x_3 should be transformed into $\sin x_3$ or $\cos x_3$, or that we should include cross terms and quadratic terms. We note that a linear regression model without the trig transformation achieved an accuracy of only 63.6%. It is this 'automatic' selection of features that makes GP regression so popular. However, the use of GPs comes at a cost. The models obtained are no longer interpretable (it is unclear which features of $\mathbf{x}$ are important), they are much more computationally expensive to use ($O(n^3)$ where n is the number of observations, compared to $O(p^3)$ for linear regression where p is the number of parameters), and choosing the covariance function k presents significant difficulties (both the functional form of κ and the hyper-parameters in k are important, and can be difficult to optimise). Because the gain in predictive accuracy from using GPs compared to linear regression is not large, here we prefer to use linear regression to investigate the simulator behaviour as the resulting analysis is easier to understand and interpret.

For comparison, note that the two-step Gaussian process emulator of the vegetation field achieves

an accuracy of 83.1% if the second step emulator uses just the temperature and precipitation fields, and an accuracy of 83.7% if the second step emulator is allowed to also use x. This compares to an accuracy of 80.3% for the two-step linear regression emulator. We note that the two-step GP emulation is more accurate than the one-step GP emulation, showing the potential benefits in predictive performance from our two-step approach. The idea of breaking down the emulation into several stages is comparable to the idea of deep learning [3], which is used in machine learning algorithms to achieve more accurate performance.

Finally, note that it is common practice to combine Gaussian processes with linear regression by using a parametric mean function in the GP, with the aim of benefiting from the strengths of both approaches. Namely, the rigid parametric response captured by the mean function describes the larger scale trend in the simulator output, helping with accuracy when extrapolating outside the range of the data. Whereas the more flexible nonparametric GP describes the departure of the simulator response from the simpler linear regression surface, improving predictive accuracy in regions for which we have data. We tried a variety of parametric mean functions for the GP model. A mean function linear in the three inputs, used in the results presented here, gave slightly improved performance over using a constant or quadratic mean function.

7 Overview and concluding comments

This paper has focused on a specific application but its general approach is potentially useful in many situations where correlated, high-dimensional data are to be emulated. For the one-step emulation procedure, only the output is high-dimensional while the explanatory variables are of low dimension and only weakly correlated. For the two-step procedure, the explanatory variables for the second stage are also high-dimensional. In our example the input and the output variables for the second stage both took values in the same space, namely a grid of spatial locations, but this need not necessarily be the case, the method could be applied to relate high-dimensional inputs and outputs of a wide range of complex models or processes. The second step of the two-step emulation is thus a new and potentially powerful approach, enabling the emulation of very high-dimensional outputs from very high-dimensional inputs. It has many potential applications in climate science, and likely more generally. The method is potentially useful and relevant for predicting or interpreting the input-output response of a process where principal inputs and outputs are both high-dimensional and the connecting process is complex

and potentially nonlinear. This could be either a model that is known in principal but too complex to fully calculate, or some other process that is at least partly deterministic but unknown.

The approach can be summarised as follows, suppose there are two high-dimensional variables $\tilde{\mathbf{Y}}_i$, for $i = 1, 2$ (the extension to three or more, as in our example, is straightforward) and in the second stage we wish to relate $\tilde{\mathbf{Y}}_2$ to $\tilde{\mathbf{Y}}_1$. For the ith high-dimensional variable, the row-centred data matrix $\tilde{\mathbf{Y}}_i$ is expressed as $\tilde{\mathbf{Y}}_i = \mathbf{L}_i\mathbf{D}_i\mathbf{R}_i'$ by using singular-valued decomposition. Singular value decomposition provides a simple method of separating the variation that arises from the explanatory variables (captured by $\mathbf{R}_i$) from the remaining, correlated variation across the dataset, in our example arising from differences in locations (captured by $\mathbf{L}_i$). Each $\mathbf{D}_i$ is a diagonal matrix of eigenvalues. A critical requirement of the emulators is that the $\mathbf{D}_i$ must each contain some eigenvalues of negligible size.

The emulators should prove useful when each $\mathbf{D}_i$ contains only a modest number of non-negligible eigenvalues (our application used ten or fewer). The matrices $\mathbf{L}_{i*}$, $\mathbf{D}_{i*}$ and $\mathbf{R}_{i*}$ are obtained from $\mathbf{L}_i$, $\mathbf{D}_i$ and $\mathbf{R}_i$ by discarding small eigenvalues and their associated eigenvectors. Then $\tilde{\mathbf{Y}}_i \simeq \mathbf{L}_{i*}\mathbf{D}_{i*}\mathbf{R}_{i*}'$ and the only way the explanatory variables influence $\mathbf{Y}_i$ is through their influence on $\mathbf{R}_{i*}$. With the one-step procedure, building an emulator reduces to the task of modelling the relationship between the explanatory variables and rows of $\mathbf{R}_{i*}$. With the two-step procedure, building an emulator reduces to relating the rows of $\mathbf{R}_{2*}$ to the rows of $\mathbf{R}_{1*}$. Condensing the relevant information into the $\mathbf{R}_{i*}$s has the potential to improve emulation, as spurious information has less scope to be influential. A further advantage is that this simplification can aid interpretation of the relationship between input and output fields.

In more detail, for the one step emulator a least-squares regression equation is determined for each score vector (column of $\mathbf{R}_i$) in turn. Each regression equation relates one of these score vectors to the explanatory variables and functions of these variables, such as quadratic and cross-product terms. Some form of variable selection is needed to form a parsimonious model and we favour constructing an overly-large model using AIC and then discarding terms using BIC. To learn more about the main relationships between the input and out terms, we suggest calculating main effect indices (Homma and Saltelli, 1996) for the regression equations of those score vectors that correspond to the largest two or three eigenvalues. A plot of the main effect indices, similar to those given in Figure 2, will generally illuminate which relationships are important.

For the two-step emulator, if $\mathbf{Y}_j$ is the output to be predicted, then each score vector in $\mathbf{R}_{j*}$ is regressed in turn on all the score vectors in the other $\mathbf{R}_{i*}$, or, in our more general case with multiple high-dimensional input fields, on all the score vectors in all the other $\mathbf{R}_{i*}$s, using variable selection (such as AIC and BIC) to form parsimonious models. Plotting the regression coefficients of the equations for the score vectors in $\mathbf{R}_{j*}$ that correspond to, say, the three largest eigenvalues in $\mathbf{D}_{j*}$ will identify the most important relationships between $\mathbf{Y}_j$ and the other $\mathbf{Y}_i$s (c.f. Figure 3.).

As discussed in Section 6, the relationships between inputs and outputs can also be derived using Gaussian Process regression, if the problem size permits and the linear approach proves inadequate, without otherwise altering the structure of the approach.

Acknowledgements

This work was funded under EU FP7 ERMITAGE grant no. 265170.

Appendix 1: Sensitivity analysis

Consider a linear regression model of the form given by equation (2)

$$E(Y) = a + \sum_{i=1}^{d} b_i X_i + \sum_{i=1}^{d} \sum_{j>i}^{d} c_{ij} X_i X_j + \sum_{i=1}^{d} d_i X_i^2, \tag{5}$$

which we fit to the simulator output. In our problem, Y is one of the scores ρ from the singular value decomposition of the temperature, precipitation or vegetation fields, and $\mathbf{X} = (\Phi_1, \Phi_2, \Phi_3, \Phi_4)'$ is the vector of parameter features describing the Earth's orbit.

The aim of variance based sensitivity analysis is to apportion the variance in the output, Y, to the variance in the inputs, $\mathbf{X}$. This will tell us which of the orbital parameters has the largest effect on each of the scores. In order to do this, we need to specify the distribution followed by the X_i, which we set as $X_i \sim U[-1, 1]$ (after rescaling the parameters onto the interval $[-1, 1]$).

There are two primary measures of the sensitivity of Y to the inputs, namely the main effects indices and the total effect indices. The total effect of the uncertainty due to input X_i [9] is defined to be

$$V_{T_i} = Var(Y) - Var(E(Y|X_{[-i]}))$$

where $X_{[-i]}$ is the vector $\mathbf{X}$ with the element X_i removed. The total effect is thus the expected variance remaining about the value of Y after we have learnt all the variables except X_i. It measures the

contribution of X_i to the variance of Y, including variance arising from interaction between X_i and other elements of $\mathbf{X}$. For the model defined by equation (5),

$$Var(Y) = \frac{1}{3}\sum_{i=1}^{d} b_i^2 + \frac{1}{9}\sum_{i=1}^{d}\sum_{j>i}^{d} c_{ij}^2 + \frac{4}{45}\sum_{i=1}^{d} d_i^2$$

where we have used the fact that if $Z \sim U[-1,1]$, then $Var(Z) = 1/3$, $Var(Z^2) = 4/45$ and if Z' is another independent $U[-1,1]$ random variable, then $Var(ZZ') = 1/9$. We can see that

$$E(Y|X_{[-i]}) = a + \sum_{j\neq i} b_j X_j + \sum_{j,k\neq i, j<k} c_{jk} X_j X_k + \sum_{j\neq i} d_j X_j^2$$

and thus

$$Var(E(Y|X_{[-i]})) = \frac{1}{3}\sum_{j\neq i} b_j^2 + \frac{1}{9}\sum_{j,k\neq i,j<k} c_{jk}^2 + \frac{4}{45}\sum_{j\neq i} d_j^2.$$

We usually convert the total effect into the total effect index by dividing by the total variance:

$$S_{T_i} = V_{T_i}/Var(V).$$

Note that $\sum S_{T_i} \geq 1$, as interaction effects are counted multiple times.

The second primary measure of sensitivity is based on the main effects. Following Oakley and O'Hagan (2004), let

$$z_i(X_i) = E(Y|X_i) - E(Y)$$

$$z_{i,j}(\mathbf{X}_{i,j}) = E(Y|\mathbf{X}_{i,j}) - z_i(X_i) - z_j(X_j) - E(Y)$$

and so on, where $z_i(X_i)$ is the main effect of X_i, and $z_{i,j}(\mathbf{X}_{i,j})$ is the first-order interaction effect between X_i and X_j, etc. ($\mathbf{X}_{i,j}$ denotes the vector (X_i, X_j)). The main effects variances are then

$$W_p = Var(z_p(\mathbf{X}_p))$$

where p can be a vector of indices. For a single index

$$W_i = Var(E(Y|X_i)),$$

which is the expected amount by which the uncertainty in Y is reduced if we learn the true value of X_i. We can interpret $W_{i,j}$ as the additional reduction in the variance of Y if we learn X_i and X_j compared

to the sum of the reduction we see when we learn either X_i or X_j alone. For models of the form given by Equation (5),

$$W_i = \frac{1}{3}b_i^2 + \frac{4}{45}d_i^2$$
$$W_{i,j} = \frac{1}{9}c_{ij}^2$$
$$W_p = 0 \text{ if } \dim(p) > 2$$

i.e., third order interaction effects and higher are zero. We again usually convert the main effects variances to the main effects indices by dividing by the total variance:

$$S_p = \frac{W_p}{Var(Y)}$$

Unlike the total effects, the main effects do add to 1 and provide a decomposition of the total variance of Y.

References

1. A. Berger, *Long term variations of caloric insolation resulting from the Earth's orbital elements,* Quaternary Research, 9 (1995), pp. 139–167.

2. W.H. Berger, *On the Milankovitch sensitivity of the Quaternary deep sea record*, Climate of the Past, 9 (2013), pp. 2003–2011.

3. A.C. Damianou and N.D. Lawrence, *Deep gaussian processes*, appearing in Proceedings of the 16th International Conference of Artifical Intelligence and Statistics (AISTATS), preprint (2013), available at http://jmlr.org/proceedings/papers/v31/damianou13a.pdf

4. K. Fraedrich, H. Jansen, E. Kirk, U. Luksch and F. Lunkeit, *The Planet Simulator: Towards a user friendly model,* Meteorologische Zeitschrift, 14 (2005), pp. 299–304.

5. W.D. Gosling and P.B. Holden, *Precessional forcing of tropical vegetation carbon storage*, Journal of Quaternary Science, 26 (2011), pp. 463–467.

6. P.B. Holden and N.R. Edwards, *Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling,* Geophysical Research Letters, 37 (2010), L21707

7. P.B. Holden, N.R. Edwards, S.A. Müller, K.I.C. Oliver, R.M. Death and A. Ridgwell, *Controls on the spatial distribution of* $\delta^{13}C_{DIC}$, Biogeosciences, 10 (2013), pp. 1815–1833.

8. P.B. Holden, N.R. Edwards, P.H. Garthwaite, F. Fraedrich, F. Lunkeit, E. Kirk, M. Labriet, A. Kanudia and F. Babonneau, *PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment*, Geoscientific Model Development, 7 (2014), pp. 433–451.

9. T. Homma and A. Saltelli, *Importance measures in global sensitivity analysis of model output,* Reliability Engineering and System Safety, 52 (1996), pp. 1–17.

10. S.R. Joshi, M. Vielle, F. Babonneau, N.R. Edwards and P.B. Holden, *Physical and economic impacts of sea-level rise: A coupled GIS and CGE analysis under uncertainties,* preprint (2014), submitted to Mitigation and Adaptation Strategies for Global Change.

11. M.C. Kennedy and A. O'Hagan, *Bayesian calibration of computer models*, Journal of the Royal Statistical Society B, 63 (2001), pp. 425–464

12. M. Labriet, S.R. Joshi, F. Babonneau, N.R. Edwards, P.B. Holden, A. Kanudia, R. Loulou and M. Vielle, *Worldwide impacts of climate change on energy for heating and cooling*, preprint (2014), to appear in Mitigation and Adaptation Strategies for Global Change. Available at http://link.springer.com/article/10.1007/s11027-013-9522-7

13. J.-F. Mercure, P. Salas, A. Foley, U. Chewpreecha, P.B. Holden and N.R. Edwards, *The dynamics of technology diffusion and the impacts of climate policy instruments in the decarbonisation of the global electricity sector*, Energy Policy, 73, (2014) pp 686-700

14. *R* Development Core Team, *R: A language and environment for statistical computing.* R foundation for statistical computing: Vienna, 2013. Available for download at http://www.r-project.org/

15. C.A. Rasmussen, *Gaussian processes for machine learning*, Citeseer, 2006.

16. T.J. Santner, B.J. Williams and W.I. Notz, *The design and analysis of computer experiments*, Springer, 2003.

17. C. Tebaldi and J.M. Arblaster, *Pattern scaling: Its strengths and limitations, and an update on the latest model simulations,* Climatic Change, 122 (2014), pp.459–471.

18. R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society B, 58 (1996), pp. 267–288.

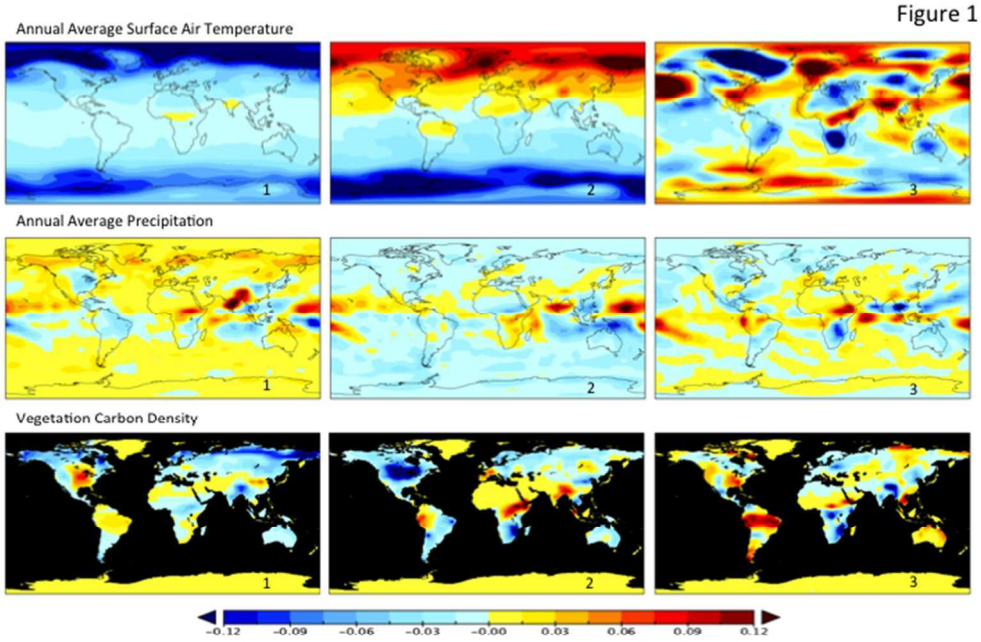
19. M.S. Williamson, T.M. Lenton, J.G. Shepherd and N.R. Edwards, *An efficient numerical terrestrial scheme (ENTS) for Earth system modelling,* Ecological modelling, 198 (2006), pp. 362–374

20. H. Zou and T. Hastie, *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society B, (2005), pp. 301–320

Figure Captions

- Figure 1: The first three components of orbitally-driven change in surface air temperature (top), annual precipitation (centre) and vegetation carbon density (bottom).
- Figure 2: Main effect indices for the first three score emulators of surface air temperature (top), annual precipitation (centre) and vegetation carbon density (bottom).
- Figure 3: Coefficients of the 2nd-step emulators of the first three vegetation scores (Equation 3).
- Figure 4: Comparison between the coefficients of the 1-step emulator of the the first vegetation score with the effective coefficients in the 2-step emulator (see Section 5 for explanation).
- Figure 5: Correlations between actual scores (i.e. decompositions of the simulation data) and emulated scores using the two emulation approaches. Absent data points occur when no terms were found to satisfy the BIC constraint, suggesting that components $k > 6$ are difficult to emulate and should be neglected.
- Figure 6: Comparisons between the simulated vegetation field (top) with the 2-step emulated field (bottom). Simulations were arbitrarily chosen as the first (left) and last (right) members of the Latin Square design.
- Figure 7: Cross-validated performance of the emulators as additional components are added.

Figure 1



254x190mm (72 x 72 DPI)

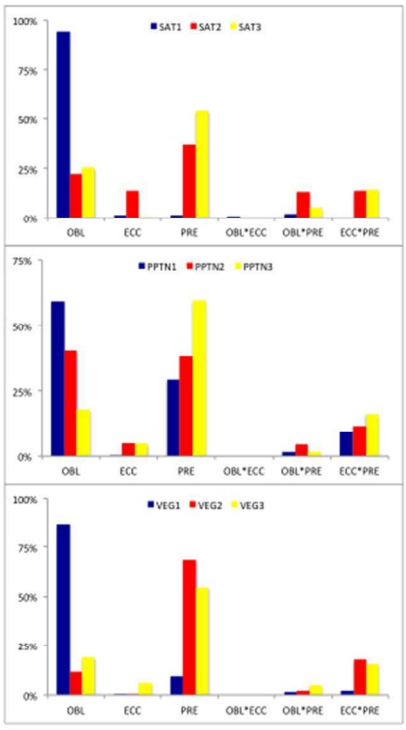
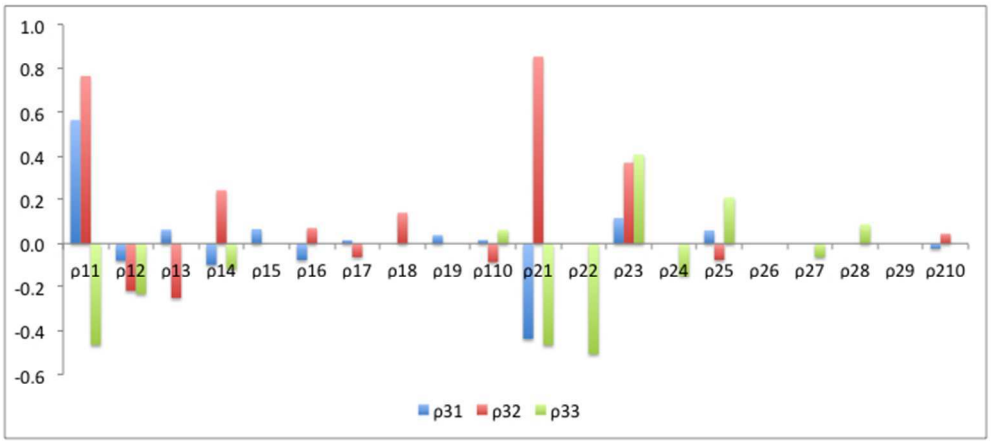


Figure 2

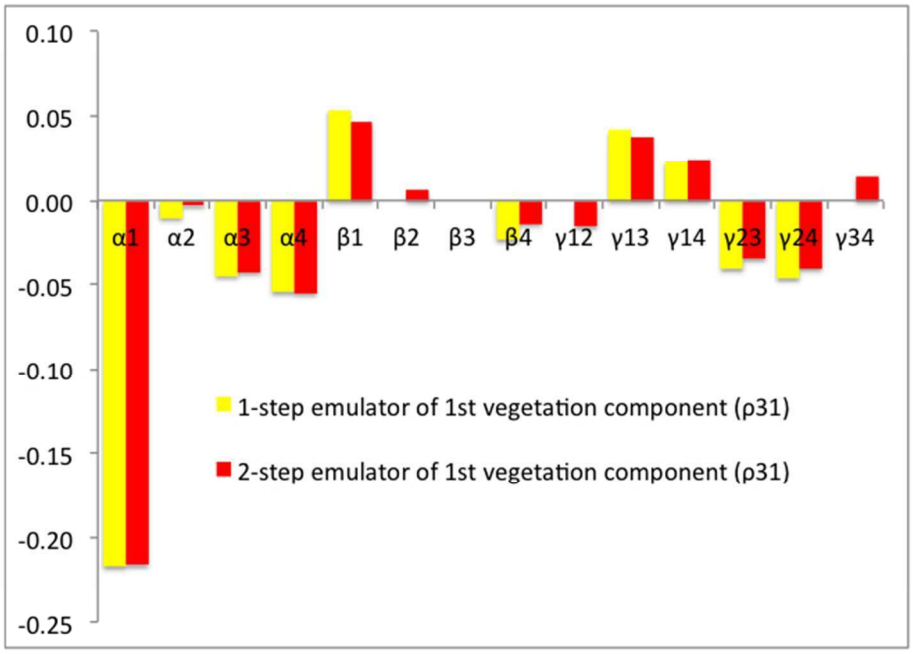
254x190mm (72 x 72 DPI)

Figure 3



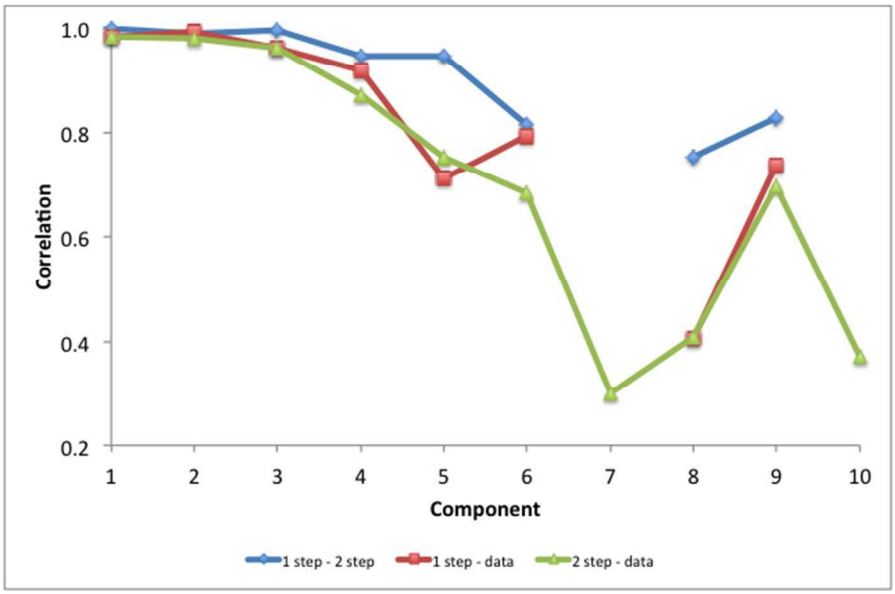
254x190mm (72 x 72 DPI)

Figure 4



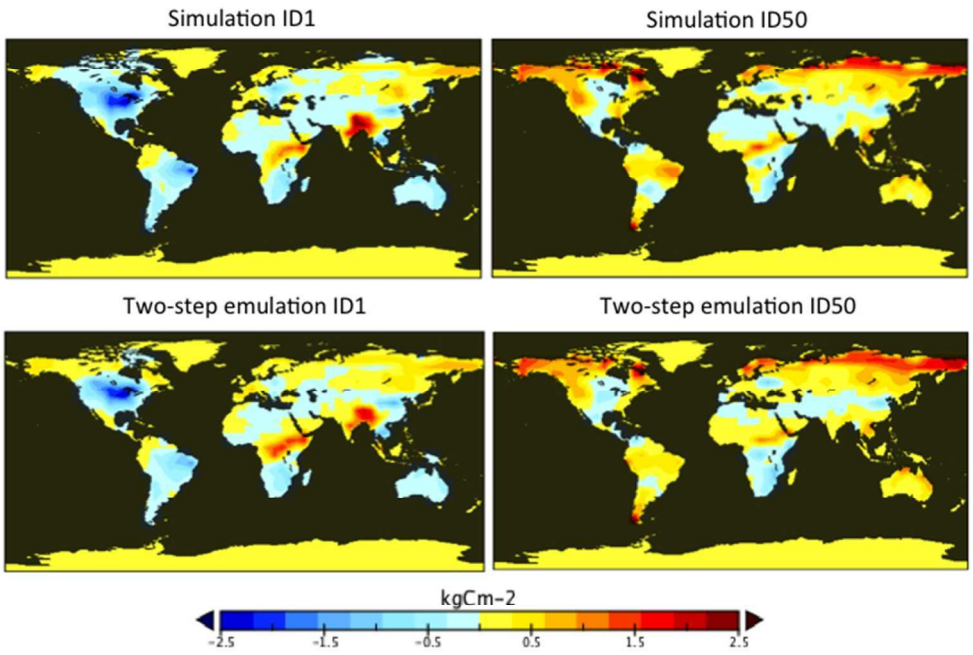
254x190mm (72 x 72 DPI)

Figure 5



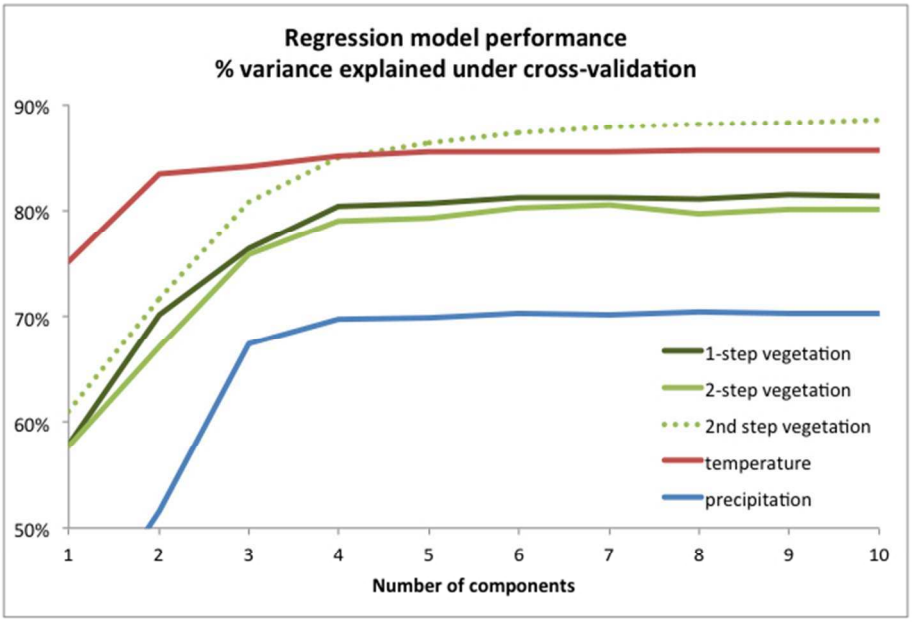
254x190mm (72 x 72 DPI)

Figure 6



254x190mm (72 x 72 DPI)

Figure 7



254x190mm (72 x 72 DPI)