Adv. Stat. Clim. Meteorol. Oceanogr., 2, 17–37, 2016
www.adv-stat-clim-meteorol-oceanogr.net/2/17/2016/
doi:10.5194/ascmo-2-17-2016

ASCMO
Open Access

Building a traceable climate model hierarchy with multi-level emulators

Giang T. Tran[1], Kevin I. C. Oliver[1], András Sóbester[2], David J. J. Toal[2], Philip B. Holden[3], Robert Marsh[1], Peter Challenor[1,4], and Neil R. Edwards[3]

[1]Ocean and Earth Science, National Oceanography Centre Southampton,
University of Southampton, Southampton, UK
[2]Faculty of Engineering and the Environment, University of Southampton, Southampton, UK
[3]Environment, Earth and Ecosystems, Open University, Milton Keynes, UK
[4]College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, UK

Correspondence to: Giang T. Tran (giang.tran@noc.soton.ac.uk)

Received: 1 September 2014 – Revised: 2 March 2016 – Accepted: 25 March 2016 – Published: 18 April 2016

Abstract. To study climate change on multi-millennial timescales or to explore a model's parameter space, efficient models with simplified and parameterised processes are required. However, the reduction in explicitly modelled processes can lead to underestimation of some atmospheric responses that are essential to the understanding of the climate system. While more complex general circulations are available and capable of simulating a more realistic climate, they are too computationally intensive for these purposes. In this work, we propose a multi-level Gaussian emulation technique to efficiently estimate the outputs of steady-state simulations of an expensive atmospheric model in response to changes in boundary forcing. The link between a computationally expensive atmospheric model, PLASIM (Planet Simulator), and a cheaper model, EMBM (energy–moisture balance model), is established through the common boundary condition specified by an ocean model, allowing for information to be propagated from one to the other. This technique allows PLASIM emulators to be built at a low cost. The method is first demonstrated by emulating a scalar summary quantity, the global mean surface air temperature. It is then employed to emulate the dimensionally reduced 2-D surface air temperature field. Even though the two atmospheric models chosen are structurally unrelated, Gaussian process emulators of PLASIM atmospheric variables are successfully constructed using EMBM as a fast approximation. With the extra information gained from the cheap model, the multi-level emulator of PLASIM's 2-D surface air temperature field is built using only one-third the amount of expensive data required by the normal single-level technique. The constructed emulator is shown to capture 93.2 % of the variance across the validation ensemble, with the averaged RMSE of 1.33 °C. Using the method proposed, quantities from PLASIM can be constructed and used to study the effects introduced by PLASIM's atmosphere.

1 Introduction

Complex computer simulations are used in climate research to improve our understanding of the climate system. They are often used to project future changes in global temperature, corresponding to different emission scenarios. Our confidence in these projections is highly dependent on how reliable the simulations are. For example, the study of palaeoclimate offers an insight into the Earth's past climate system and also provides valuable out-of-sample data to validate our simulations. However, this requires running complex simulations on multi-millennial timescales, which is computationally demanding. For most coupled atmosphere–ocean general circulation models (AOGCMs), this is currently not feasible. Other studies such as uncertainty and sensitivity analysis or history matching require a thorough exploration of the input parameter space. The class of fast models, known as Earth system models of intermediate complexity (EMICs) is suitable for these types of studies. Their efficiency is achieved by a combination of lower spatial and/or temporal resolution and the use of simplified parameterisations. However, depending on the nature of the questions asked, these lower fidelity models might be insufficient.

To address this issue, an emulator is often employed to provide a statistical estimation of the expensive model's response without the need to perform a new simulation. Even then, this approach becomes impractical when the models of interest are very computationally intensive. In order to build a reliable emulator, a certain number of simulations is needed to provide the basis upon which the emulator is built. This number can be large, especially when multiple model parameters are varied or when the model's climate response exhibits non-linear behaviours. For a computationally expensive GCM, a sufficient number of simulations are often not affordable. This paper describes an efficient emulation process that utilises the connection between models of different complexities. The idea is to establish a traceable hierarchy, using an emulator for the simple model to construct an emulator of the more complex one (Kennedy and O'Hagan, 2000; Cumming and Goldstein, 2008).

While the high-fidelity (complex) model is computationally expensive, the low-fidelity (simple) model is cheaper to evaluate and can be sampled more finely across the input space, providing extra information where expensive data are sparse. The models forming this hierarchy can be structurally related or structurally unrelated. Models are referred to as structurally related when they are from the same family of code but have different resolutions. These models might have other differences resulting from the change in mesh resolution. Examples of such models are the HadCM3 (Hadley Centre Coupled Model version 3) (Pope et al., 2000) and FAMOUS (Fast Met Office/U.K. Universities Simulator) (Jones et al., 2005) of the MET Office. Multi-level emulation has been employed before to link such models (Forrester et al., 2007; Cumming and Goldstein, 2008; Williamson et al., 2012). Here, our focus is on structurally unrelated atmospheric models, which solve different sets of equations. Since both the cheap and expensive codes model the same physical system, it is reasonable to expect qualitative similarities between the two. This argument is supported by studies showing no systematic difference in model behaviour between EMICs and AOGCMs (Stouffer et al., 2006; Plattner et al., 2008; Zickfeld et al., 2013).

The following work illustrates the use of a method that combines multi-level emulation with a dimensional reduction technique through an example study using GENIE-1, from the Grid ENabled Integrated Earth system modelling framework (GENIE), and PLASIM (Planet Simulator). GENIE-1 and PLASIM are chosen in this case since they are both suitable for Earth system modelling for long timescales, but are structurally different. PLASIM's atmosphere is also substantially more complex and thus, computationally more expensive than GENIE-1's energy–moisture balance model, EMBM, of the atmosphere. EMBM incorporates the vertically integrated energy–moisture balance equations while PLASIM is based on the moist primitive equations representing the conservation of momentum, mass and energy. EMBM, therefore, is not capable of producing air temperature and pressure at different altitude or an interactive cloud and wind field. The hierarchy formed by these two models is exploited using the multi-level technique, allowing us to construct an emulator of PLASIM atmospheric variables at a reduced cost. Specifically, Gaussian process emulators are used to obtain the statistical relationship between the response of the EMBM atmosphere and the PLASIM atmosphere to changes in their boundary conditions (sea surface temperature, long-wave and shortwave radiative forcing). This ability of this relationship to predict behaviour of PLASIM atmosphere, in the absence of feedbacks on other climate system components, is then assessed. The dimensional reduction technique is employed to extend the emulation method for prediction of high-dimensional outputs in addition to scalar summary quantities.

Once constructed, the emulators provide estimates of simulation results, at untried combinations of the inputs, as finely as needed, at a low cost. This enables statistical methods such as history matching (Holden et al., 2010; Edwards et al., 2011) and sensitivity/uncertainty analysis (Rougier et al., 2009). Information from the cheap code can also be used to inform future designs of experiments using the expensive code. Apart from above, the emulators of 2-D surface fields similar to the one constructed here can potentially be used to provide the fields needed for coupling with other climate models or components of climate models.

2 Model configurations

In this study, we utilise the atmospheric component of GENIE-1 (version 2.7.8) (Lenton et al., 2006), an EMIC, as the cheap model. GENIE-1 was originally known as C-GOLDSTEIN in Edwards and Marsh (2005) and has since been modified for incorporation into the GENIE framework (Lenton et al., 2006). It is most recently described in Marsh et al. (2011). GENIE-1 is designed with scalable spatial resolution and high efficiency, suitable for long integrations (10^3 to 10^6 years) to study past climate and large ensembles to explore the uncertain input parameter space (Holden et al., 2010).

The configuration of GENIE-1 employed here couples a single layer EMBM atmosphere to a 3-D frictional geostrophic ocean model with linear drag (GOLDSTEIN) and a thermodynamic, advection–diffusion sea-ice model (GOLDSTEIN sea ice). The ocean component is run at 64 × 32 horizontal resolution and 16 vertical layers. Also incorporated in this configuration is the efficient numerical terrestrial scheme (ENTS) designed for long simulations (Williamson et al., 2006). ENTS represents a hybrid of a simple bucket model with an explicit but simplified carbon cycle. The effect of orography is applied to surface processes in ENTS by applying a constant lapse rate (Holden et al., 2010). Orography, therefore has an effect on the land surface temperature and so indirectly influences the atmosphere. The atmospheric processes such as heat and moisture transport do not interact with the orography.

The parameterisation of atmospheric transport of heat and moisture in EMBM is done by diffusion. Moisture can also be advected by a prescribed monthly climatological wind field. This wind field is fixed and is the same for all simulations in EMBM. The effect of cloud cover on incoming shortwave radiation is captured through a prescribed albedo field, diagnosed from reanalysis data (Lenton et al., 2006). The effect of cloud cover on outgoing long-wave radiation is parameterised as perturbations to the unmodified “clear-skies” outgoing long-wave radiation. Precipitation is assumed to occur whenever the relative humidity is above a certain adjustable threshold.

The atmosphere of PLASIM–ENTS (Holden et al., 2014), driven by boundary conditions specified by GOLDSTEIN ocean and sea ice, is chosen as the expensive model. PLASIM (Fraedrich et al., 2005) consists of an atmospheric GCM of intermediate complexity, which can interact with reduced sub-models of ocean, sea ice and land (Fraedrich et al., 2005). Hereafter, we refer to the atmospheric component of PLASIM–ENTS as simply PLASIM. PLASIM solves the primitive equations for vorticity, divergence, temperature and the logarithm of surface pressure. It includes a hydrological cycle, interactive clouds, and a simple radiation scheme. Coupling between the PLASIM atmosphere and an ocean model other than its own has been used before to study the effects of mountains and ice sheets on ocean circulation (Schmittner et al., 2011). An emulator of PLASIM–ENTS has been employed in a range of integrated assessment modelling couplings with various technico-economic models (Labriet et al., 2015; Mercure et al., 2014).

PLASIM is run at T21 resolution, which corresponds to a triangular truncation applied at wave number 21. It is almost an exact match of GENIE-1’s 64×32 mesh except for negligible differences at the highest latitudes. While EMBM has only one layer, the atmosphere of PLASIM is represented by 10 vertical layers in terrain following σ-coordinates. Most importantly, EMBM uses prescribed wind fields, which means that feedbacks due to changing atmospheric circulation patterns are not captured, while PLASIM’s interactive wind field can change according to the different specified boundary conditions, leading to more diverse climate states. While the cloud albedo in EMBM is prescribed, the cloud albedo in PLASIM is a function of height and area of coverage.

For our study, surface output fields of GENIE-1, namely, sea surface temperature (SST), fractional sea-ice coverage (SIC) and sea-ice thickness (SIH) are used to drive PLASIM. This means that the atmospheric circulation can change according to the underlying sea surface temperature and sea-ice condition but cannot influence the ocean or sea-ice physical state. This constrains PLASIM responses to a certain extent. The atmospheric responses of EMBM and PLASIM to the same set of physically plausible boundary conditions are compared and emulated. The surface air temperature (SAT) from EMBM atmosphere is treated as a fast approximation of PLASIM SAT when multi-level emulation is applied.

3 Ensemble design

3.1 Model parameters

To explore emulator performance in situations where the climate states are very different from modern conditions, an ensemble is designed to fill a large input space; 12 model parameters and one dummy variable are varied, either linearly or logarithmically, over the ranges indicated in Table 1. In this experiment, we are primarily interested in the effects introduced by the dynamical atmosphere of PLASIM and so the parameters were chosen according to their influence on SAT. Parameters with important contributions to SST and the strength of the AMOC, and hence indirectly influencing SAT, are also included. This judgment is based on previous studies using large GENIE-1 ensembles (Lenton et al., 2006; Holden et al., 2010).

The first parameter (ICF) represents the boundary condition of the glacier coverage as well as the corresponding orography at different a snapshot in time extending from the present (0 kyr before present) to the Last Glacial Maximum (LGM) (21 kyr before present) with steps of 1 kyr. Each value of ICF corresponds to a spatial distribution of land ice at a certain period according to the Peltier reconstruction ICE-5G (Peltier, 2004). Both ice area and ice volume are non-linear functions of ICF. Together with ICF, the atmospheric CO_2 concentration (RFC) is varied from 150 ppm to 1400 ppm to include the glacial–interglacial variations, pre-industrial and modern climate as well as future responses to rising greenhouse gas emissions. The upper limit is chosen to include the CO_2-equivalent concentration for all greenhouse gases by 2100 according to the high emission pathway RCP8.5 (Riahi et al., 2011; Meinshausen et al., 2011). The equivalents of these two parameters are also varied accordingly in PLASIM. Other PLASIM parameters are kept at default values, which are listed in Haberkorn et al. (2009).

Mixing and transport in the ocean are controlled by the isopycnal and diapycnal diffusivity parameters (OHD and OVD, respectively), a momentum drag coefficient (ODC) and a wind scaling factor (WSF) (Edwards and Marsh, 2005). These parameters affect the ocean boundary conditions, which are seen by both EMBM and PLASIM directly.

APM is a flux correction responsible for transporting freshwater from the Atlantic to Pacific, affecting deep water sinking in the North Atlantic and hence the strength of the AMOC (Edwards and Marsh, 2005). The uncertain impact of atmospheric transport is captured through atmospheric heat

Table 1. Ten of the chosen parameters, with the exception of ICF and RFC, are taken from an ensemble design used in Holden et al. (2010). The ranges were initially based on those used in the same study. However, adjustments are needed since the model is run at 64 × 32 horizontal resolution here compared to the previously used 36 × 36 mesh. The ranges shown below are obtained after an initial exploratory ensemble. The distribution specifies whether their values (Lin) or the log of their values to base 10 (Log) are used to generate the sampling plan in Sect. 3.2.

	Code	Parameter	Min	Max	Dist.
1	ICF	Ice sheet and orography configuration	0	21	Lin
2	OHD	Ocean isopycnal diffusivity (m^2s^{-1})	300	4000	Log
3	OVD	Ocean diapycnal diffusivity (m^2s^{-1})	5×10^{-6}	2×10^{-4}	Log
4	ODC	Ocean friction coefficient ($days^{-1}$)	0.5	3	Lin
5	WSF	Wind scale coefficient	1	3	Lin
6	AHD	Atmospheric heat diffusivity (m^2s^{-1})	4×10^6	7.0×10^6	Log
7	AMD	Atmospheric moisture diffusivity (m^2s^{-1})	5×10^4	6×10^6	Log
8	APM	Atlantic–Pacific freshwater flux (Sv)	0.032	0.640	Lin
9	RMX	Relative humidity threshold for precipitation	0.6	0.9	Lin
10	OL0	Clear-sky OLR reduction (Wm^{-2})	0	10	Lin
11	OL1	OLR feedback ($Wm^{-2}K^{-1}$)	−0.5	0.5	Lin
12	RFC	CO_2 forcing (ppm)	150	1400	Lin
13	FFX	Dummy variable	0	1	Lin

and moisture diffusivity parameters (AHD and AMD, respectively) (Edwards and Marsh, 2005). OL0 and OL1 modify the outgoing long-wave radiation and are included to allow for uncertainty due to cloud coverage and its dependence on a change in the global average SAT (Thompson and Warren, 1982; Matthews and Caldeira, 2007). RMX is the threshold value of relative humidity for precipitation, capturing the uncertainty in water vapour feedbacks (Lenton et al., 2006). Except for APM, these atmospheric parameters directly control the behaviour of EMBM, but only affect PLASIM indirectly through their influence on the ocean boundary conditions.

In addition to these 12 model parameters, a dummy parameter is included for statistical validation purposes, which will be discussed in more detail in Sect. 4.1.

3.2 Statistical design

First, all input parameters are normalised to [0, 1] from their original ranges in Table 1. An approximate maximin Latin hypercube (MLH) (Morris and Mitchell, 1995) sampling plan is then generated, producing 660 combinations of the 13 chosen parameters to form a GENIE-1 perturbed physics ensemble. The maximin criterion, also known as the Morris–Mitchell criterion, is applied since a randomly generated Latin hypercube does not ensure good space-filling properties, which are desirable to evenly explore the input space. A MLH sample is a Latin hypercube sample that maximises the minimal separation, $\min_{i \neq j} = d(x_i, x_j)$, between pairs of design points, x_i and x_j. Here, the separation is simply the Euclidean distance between the points.

Each member simulation of this ensemble is run for 5000 years to reach a steady state; 600 simulations were completed successfully, producing a large range of climate responses, which are summarised in Table 2. The 60 failures are located at the end of one or more parameter ranges, where numerical instability occurs. Failure is most commonly due to low values of AHD and AMD. Although the design space can be narrowed down to reduce the failure rate, this would also restrict the range of the resulting climate states. Since we wish to build emulators, which can predict a broad range of climate responses without having to extrapolate beyond the designed range, this ensemble design is appropriate.

A second MLH design is generated in the same parameter space, producing 214 successful simulations, for validation purposes. The emulator predictions at these points are compared against the simulated values to assess the performance of the emulators.

For each successful GENIE-1 simulation, surface output fields are extracted and used to force PLASIM for another 35 years. Each sampling plan, therefore, produces two equivalent ensembles of EMBM and PLASIM outputs. The fields used to initiate PLASIM simulations are SST, SIC and SIH as mentioned in Sect. 2. The 600-member ensemble mean and standard deviation of GOLDSTEIN SST and ice area are shown in Fig. 1. The ice coverage plotted is a combination of the fractional sea-ice cover from GOLDSTEIN sea ice (SIC) and the glacier mask described by ICF. The change in elevation corresponding to each glacier mask is applied for both GENIE-1 and PLASIM.

Both ensemble designs are larger than needed in this case. On average, 10 simulations are needed for each parameter being varied. Since 13 parameters are perturbed, a 130-member ensemble would be sufficient. There are several reasons why a 600-member ensemble was used. First, the number of simulations required ultimately depends on the varia-

Table 2. A summary of the simulated climate states from the 600-member ensembles of GENIE-1 with EMBM and PLASIM.

	Min	Max	Mean	SD
PLASIM global mean SAT (°C)	−6.05	23.33	11.25	4.63
EMBM global mean SAT (°C)	−2.62	24.43	12.56	4.25
GENIE-1 global mean SST (°C)	7.77	27.10	17.01	3.24
GENIE-1 maximum strength of the AMOC (Sv)	0.82	36.59	15.31	5.94
GENIE-1 Antarctic sea-ice area ($\times 10^6 \mathrm{km}^2$)	0.00	23.09	1.73	2.60

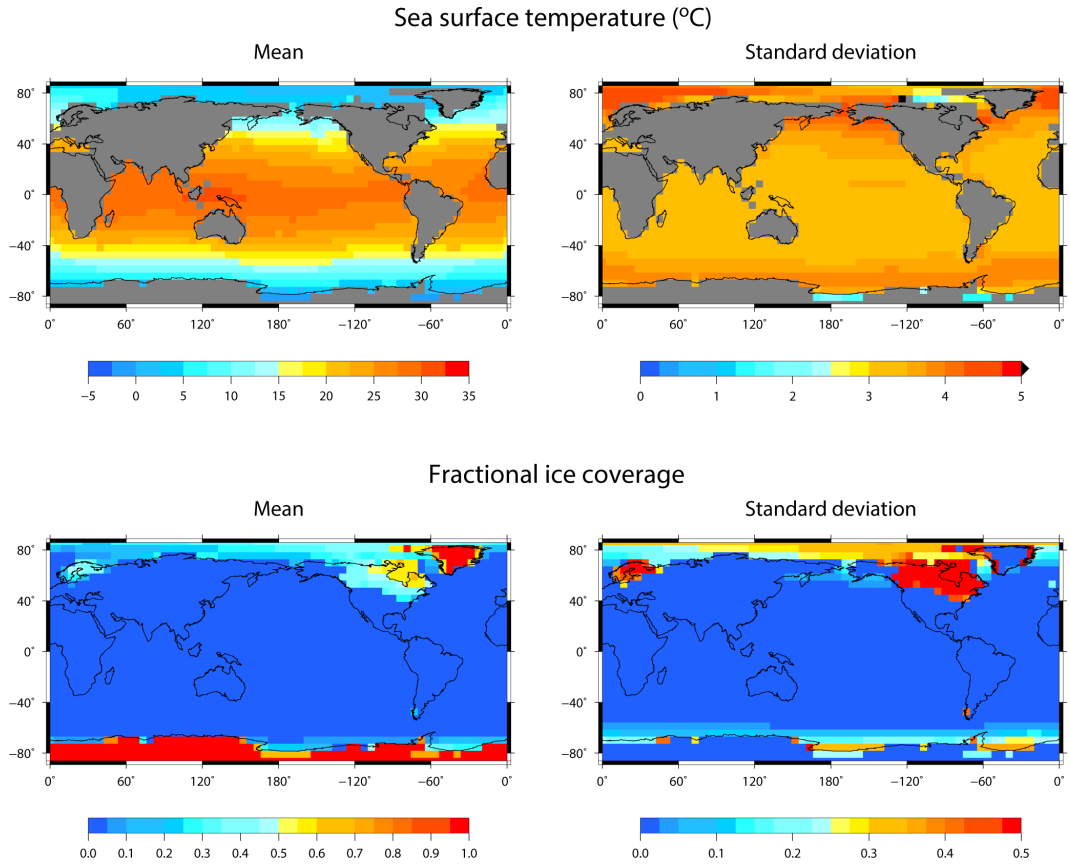


Figure 1. The mean and standard deviation of SST and fractional ice coverage across the 600-member ensemble. The SST and sea-ice coverage are prognostic output of GENIE-1 while the land ice coverage is regridded from Peltier ICE-5G. These fields, among others, are applied as surface boundary conditions to drive PLASIM atmosphere.

tions of the variable of interest within the specified parameter space. If this variable behaves non-linearly and exhibits a bifurcation, more simulations would be required to capture such behaviour accurately. Second, the required number of simulations of cheap and expensive models are unknown. Different combinations of subsets with varying sizes are used and compared in the following section. It is ideal to generate a new design separately for each case but this is highly inefficient and will result in an large incoherent ensemble with low reusability. Therefore, it is preferable to start with a large design from which different subsets can be chosen. These subsets are all subjected to the same maximin criteria mentioned above. The algorithm used is covered in Sect. 4.2. While this ensemble will be more reusable, a subset from it will most likely have a worse space-filling property than an independent MLH design of the same size. This is minimised by starting from a very large ensemble like the one employed here.

4 Statistical emulation

4.1 Gaussian process emulator

In a computer experiment, the model outputs at some combinations of input parameters are considered as observations. An emulator is a statistical surrogate of a model, which is generally much cheaper to evaluate and, once validated, can be used in place of the full model to predict the observation at untried choices of inputs. Our interest focuses on the Gaussian process (GP) emulator, also known as kriging (Rasmussen and Williams, 2006; Forrester et al., 2007), and a multi-level extension to this method, referred to as co-kriging (Kennedy and O'Hagan, 2000; Forrester et al., 2007; Cumming and Goldstein, 2008). The advantage of using the GP emulator is that the curve fits through the known points (training points from model runs at predefined sets of parameters) and an estimated uncertainty is obtained for each emulated point.

To emulate a single summary quantity of the simulation outputs, for example, the global mean SAT, the assumptions made are as follows:

- The model output is a smooth function of its inputs.
- The model can be represented as a GP.
- Each emulator is concerned with a single deterministic scalar output.

The climate model, $f(\cdot)$, is a function of a set of parameters, $\boldsymbol{x} = (x_1, \cdots, x_k)$, where k is the number of perturbed model parameters, which is 13 in this case. This number is commonly referred to as the number of dimensions of the emulator. The function $f(\cdot)$ is distributed as a GP with a mean function $m(\cdot)$ and a covariance function $V(\cdot,\cdot)$. The mean function is given by

$$m(\boldsymbol{x}) = h^T(\boldsymbol{x})\beta, \tag{1}$$

where $h(x)$ is a vector of known regression functions. In the case of traditional kriging, $h^T(\boldsymbol{x}) = \mathbf{1}$, making β the unknown overall mean. A variation of kriging, called universal kriging, uses a linear mean function:

$$h^T(\boldsymbol{x}) = (1, \boldsymbol{x}), \tag{2}$$

where $h^T(\boldsymbol{x})$ is a $(q \times 1)$ vector with $q = k+1$. Then

$$m(\boldsymbol{x}) = \beta_1 + \beta_2 \boldsymbol{x_1} + \cdots + \beta_{k+1}\boldsymbol{x_k}. \tag{3}$$

The coefficients $[\beta_2, \beta_{k+1}]$ now describe the expected trend of the simulator in response to each input.

The covariance function is given by

$$V(\boldsymbol{x},\boldsymbol{x}') = \sigma^2 \boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{x}'), \tag{4}$$

where σ^2 is the variance of the GP and $\boldsymbol{\Psi}(.,.)$ is the assumed correlation function:

$$\boldsymbol{\Psi}(\boldsymbol{x},\boldsymbol{x}') = \exp\left[-\sum_{j=1}^{k} 10^{\theta_j}\left|\boldsymbol{x_j} - \boldsymbol{x'_j}\right|^{p_j}\right]. \tag{5}$$

The function $\boldsymbol{\Psi}$ describes the correlation between pairs of points, which is assumed to be stationary and continuous, that is, it only depends on the distance between the pair of inputs, $(\boldsymbol{x} - \boldsymbol{x}')$. This exponential power form of covariance structure is a popular choice due to its flexibility. Its assumption of stationarity might fail, for example, when there is a bifurcation in the system.

The value of $\boldsymbol{\Psi}$ depends on the correlation parameters p and θ, referred to as hyperparameters. θ is the correlation length parameter, defining how quickly the correlation between the simulator outputs at two input points declines as the distance between them increases. θ indicates the activity of the function in the corresponding dimension. p is a "smoothness" parameter of the correlation function. For simplicity and to reduce computational cost, p is assumed to be the same for all dimensions.

The specified GP is used as a prior for Bayesian inference and is parameterised in terms of the hyperparameters β, σ^2, θ and p. By analytically marginalising β and σ^2, the marginal likelihood of the observed outputs at n training points, $\boldsymbol{y} = [y_1 = f(\boldsymbol{x}_1), \cdots, y_n = f(\boldsymbol{x}_n)]$, given θ and p can then be computed. A more detailed description of the derivations and formulations can be found in Mardia and Marshall (1984). The estimated θ_j in kriging and β_{j+1} in universal kriging indicate the relative activity in the jth corresponding dimension. Very low values of these hyperparameters imply inactive inputs. The dummy parameter, FFX, is included to verify that the emulator is doing a good job at identifying inactive inputs.

Prior beliefs about the model behaviour are combined with observations from training points to produce a posterior distribution for the model. Having obtained estimates for θ and p, the posterior distribution found can be used to make predictions about the model's outputs at unsampled inputs. Full description of the derivation of the posterior distribution as well as distributional assumptions made for $f(\cdot)$, β and σ^2 are available in Kennedy and O'Hagan (2001).

The exponential power form of covariance structure used here is a common choice due to its flexibility. Its assumption on stationary might fail, for example, when there is a bifurcation in the system. The covariance specified, however, provides a weak prior and as more training points are used, it contributes less to the final emulator.

4.2 Multi-level emulator

Co-kriging is an extension to the previously described technique, which is applicable when a fast approximation of the primary simulator is available. In order for this method to work, the primary simulator and its approximation need to fulfil an additional assumption:

- The different levels of code are correlated and contain information about one another.

When only a small number of expensive runs is available, it has been shown that by combining these with cheaper runs from a simplified code, an emulator of the expensive model can be built at a lower cost (Forrester et al., 2007). Potentially, this method can be extended to more code levels (Kennedy and O'Hagan, 2000), including the conceptual "reified" model (Goldstein and Rougier, 2009).

We make a simplification that the expensive and cheap models, f_e and f_c, respectively, can be represented by GP emulators of the same smoothness p. The cheap model is first emulated and then linked to the expensive one using the single multiplier approach:

$$f_\mathrm{e}(\boldsymbol{x}) = \rho f_\mathrm{c}(\boldsymbol{x}) + f_\mathrm{d}(\boldsymbol{x}). \tag{6}$$

The expensive function is modelled as the cheap GP multiplied by a scaling factor ρ, plus a separate GP, f_d, modelling the stochastic residual of the expensive model (Kennedy and O'Hagan, 2000; Forrester et al., 2007). This approximation is chosen for its simplicity as well as the assumption that the main difference between the two models is a matter of scale, rather than changes in the shape or the location of the output. This assumption is made based on the fact that both models share the same ocean component and have the same inputs.

Two sets of training points are required for the construction of a co-kriging emulator, a cheap set $\boldsymbol{y}_\mathrm{c} = f_\mathrm{c}(\boldsymbol{x}_\mathrm{c})$, which finely samples the input space, and a small sparse set $\boldsymbol{y}_\mathrm{e} = f_\mathrm{e}(\boldsymbol{x}_\mathrm{e})$ of expensive points. Let the number of cheap and expensive points be n_c and n_e, respectively.

When the number of PLASIM training points is small, such that a kriging emulator cannot be built with high accuracy, co-kriging employing an additional large number of training points from GENIE-1's EMBM can be used instead. The number of points required depends on the size of the problem as well as the smoothness of the function being emulated. The inputs at which the expensive training set is obtained, $\boldsymbol{x}_\mathrm{e}$, form a subset of the cheap set, $\boldsymbol{x}_\mathrm{c}$. These expensive points are chosen using an exchange algorithm described by Cook and Nachtsheim (1980). A random subset $\boldsymbol{x}_\mathrm{e}$ is selected and the Morris–Mitchell criterion is calculated. The first point $\boldsymbol{x}_\mathrm{e}^{(1)}$ is then exchanged with each of the remaining points in $\boldsymbol{x}_\mathrm{c}$. The exchange that gives the best Morris–Mitchell criterion is chosen. By repeating the same procedure for the remaining points $\boldsymbol{x}_\mathrm{e}^{(2)}, \cdots, \boldsymbol{x}_\mathrm{e}^{(n_\mathrm{e})}$, the "best" subset is obtained.

The covariance matrix for co-kriging, $\boldsymbol{\Psi}_\mathrm{ck}$, can be written in block form as

$$\boldsymbol{\Psi}_\mathrm{ck} = \begin{pmatrix} \sigma_\mathrm{c}^2 A_\mathrm{c}(x_\mathrm{c}) & \rho\sigma_\mathrm{c}^2 A_\mathrm{c}(x_\mathrm{c}, x_\mathrm{e}) \\ \rho\sigma_\mathrm{c}^2 A_\mathrm{c}(x_\mathrm{e}, x_\mathrm{c}) & \rho\sigma_\mathrm{c}^2 A_\mathrm{c}(x_\mathrm{e}) + \sigma_\mathrm{e}^2 A_\mathrm{e}(x_\mathrm{e}) \end{pmatrix}, \tag{7}$$

with $A_\mathrm{c} = \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}'; \theta_\mathrm{c})$ and $A_\mathrm{e} = \boldsymbol{\Psi}(\boldsymbol{x}, \boldsymbol{x}'; \theta_\mathrm{e})$. This covariance matrix encompasses the correlation between cheap points ($A_\mathrm{c}(x_\mathrm{c})$), expensive points ($A_\mathrm{c}(x_\mathrm{e})$ and $A_\mathrm{e}(x_\mathrm{e})$) and the cross-correlation between the cheap and expensive points ($A_\mathrm{c}(x_\mathrm{c}, x_\mathrm{e})$). Details on the formulation and derivation of this equation can be found in Kennedy and O'Hagan (2000) and Forrester et al. (2007).

Both kriging and co-kriging emulators are constructed using readily available software from Forrester et al. (2008).

4.3 Dimensional reduction using principal component analysis

So far, we have only discussed the use of GP emulators for single outputs. This can be a summary quantity such as the strength of the AMOC or the global average SAT (Hankin, 2005). The relevant output is, however, usually a high-dimensional array, containing fields and/or time series of many climate variables (e.g. SST, SAT or precipitation).

Climate variables at different spatial or temporal locations can be emulated independently (Lee et al., 2012). This method, however, requires large computational power and ignores the covariances between outputs close to one another (Rougier, 2007). Other extension techniques using approaches that can capture the correlations between the outputs have been developed (Rougier, 2008; Conti and O'Hagan, 2010). However, these methods are not well suited for high-dimensional output.

In this work, we use principal component analysis (PCA) via singular value decomposition (SVD) to transform the high-dimensional data into a meaningful representation with lower dimensionality. While there are several techniques to accomplish this task, PCA is efficient and has the advantage that the leading components explain the majority of the variance across the ensemble (Holden and Edwards, 2010; Wilkinson, 2010). It is by far the most popular unsupervised linear technique. The mapping from the input parameter space to the reduced dimensional output space, specified by PCA, is the function being emulated instead of the direct input–output relationship. This method has been applied successfully in emulating temporally evolving spatial patterns of climate variables in Challenor et al. (2010), Holden et al. (2013) and Holden et al. (2014).

For each ensemble member, our field of interest, SAT, with dimension 64 × 32 is reshaped as a (2048 × 1) vector. The whole ensemble consisting of n fields is represented by the (2048 × n) matrix $\mathbf{Y}$. Singular value decomposition is then performed on the centred matrix; i.e. the ensemble-averaged vector, μ, is removed:

$$\mathbf{Y} - \mu = \mathbf{LSR}^T, \tag{8}$$

where $\mathbf{L}$ is the (2048 × n) matrix of left singular vectors, also known as the empirical orthogonal functions (EOFs), $\mathbf{S}$ is the ($n \times n$) diagonal matrix of singular values and $\mathbf{r}$ is the ($n \times n$) matrix of right singular vectors, or the component scores. The product $\mathbf{P}$ of the singular values and the component scores is commonly known as the matrix of principal components (PCs):

$$\mathbf{P} = \mathbf{S} \times \mathbf{R}^T. \tag{9}$$

Any of the simulated fields can be constructed as a linear combination of the EOFs, weighted by their respective series of PCs. Each (2048 × 1) column of $\mathbf{Y}$ is an EOF, describing a map or a mode of variation in the ensemble. These are stationary spatial structures that constitute directions of variability with no particular amplitude. The corresponding PC for each of these modes is the $(n \times 1)$ column of $\mathbf{P}$. The nth element of each PC corresponds to the nth simulation from the training ensemble. These PCs provide the sign and the overall amplitude of the EOF corresponding to each simulation. They can, therefore, be considered as scalar functions of the input parameters and can be emulated using kriging or co-kriging. The number of training points, n, become n_c and n_e for the cheap and expensive emulator, respectively.

The EOFs and PCs of EMBM and PLASIM SAT can be obtained by decomposing each set separately. However, we are interested in using EMBM's PCs as the cheap approximation of PLASIM's values; therefore, the SAT fields from both models are projected onto the same orthogonal basis vectors defined by PLASIM's EOFs. This gives a new set of PCs for EMBM's SAT:

$$\mathbf{P}_\mathrm{r} = \mathbf{L}_\mathrm{e}^T \times \mathbf{Y}_\mathrm{c}. \tag{10}$$

In other words, EMBM data ($\mathbf{Y}_\mathrm{c}$) are rotated onto PLASIM's coordinate system ($\mathbf{L}_\mathrm{e}$) and the PCs obtained ($\mathbf{P}_\mathrm{r}$) are the coordinates of EMBM's SAT fields in this new system. For cokriging, the normal PCs are used as expensive training data from PLASIM while the rotated PCs are used as cheap training data from EMBM.

The top (or high order) EOFs explain most of the variance in the data such that the dimension of $\mathbf{Y}$ can be reduced by keeping only the first q components ($q < n$). The elements of the PC vectors are now used as training data instead of the direct climate variable. We assume that these PCs also fulfil the same assumptions made for the climate variables. Emulators are built for the first q PCs, providing an estimation, $\hat{\boldsymbol{P}}$, for an unknown input vector, i.e. the (214×13) input vector of the validation set. They are then used to work out the final prediction of the emulated field:

$$\hat{Y}_{il} = \mu + \sum_{j=1}^{q} \mathbf{L}_{ij} \hat{\boldsymbol{P}}_{jl}^T, \tag{11}$$

where $\hat{Y}_{il}$ is a component of the (2048×214) matrix $\hat{\boldsymbol{Y}}$.

The prediction, $\hat{\boldsymbol{Y}}$, is different from the simulated value of $\mathbf{Y}$ by an error component, which can be decomposed into truncation error and component error. Truncation error is due to dimensional reduction. This is kept low by making sure that enough EOFs are retained to explain most of the variance in the ensemble. Although there is no definite rule on what percent explained would be sufficient, a high value such as 90 % should be satisfactory. EOFs that explain less than 1 % of the total variance are often truncated since the data contained in them are often indistinguishable from random noise. Here, the first 10 EOFs are emulated and added progressively. Validation is performed after each step and only EOFs, which contribute positively to the total variance explained, are kept. Component error is a result of imperfect estimation by the emulator, i.e. an error in estimating the correct hyperparameters. This can be minimised by making sure enough training data are used to ensure the emulator can capture the real trend of the ensemble. The GP emulator also provides an estimate of this error.

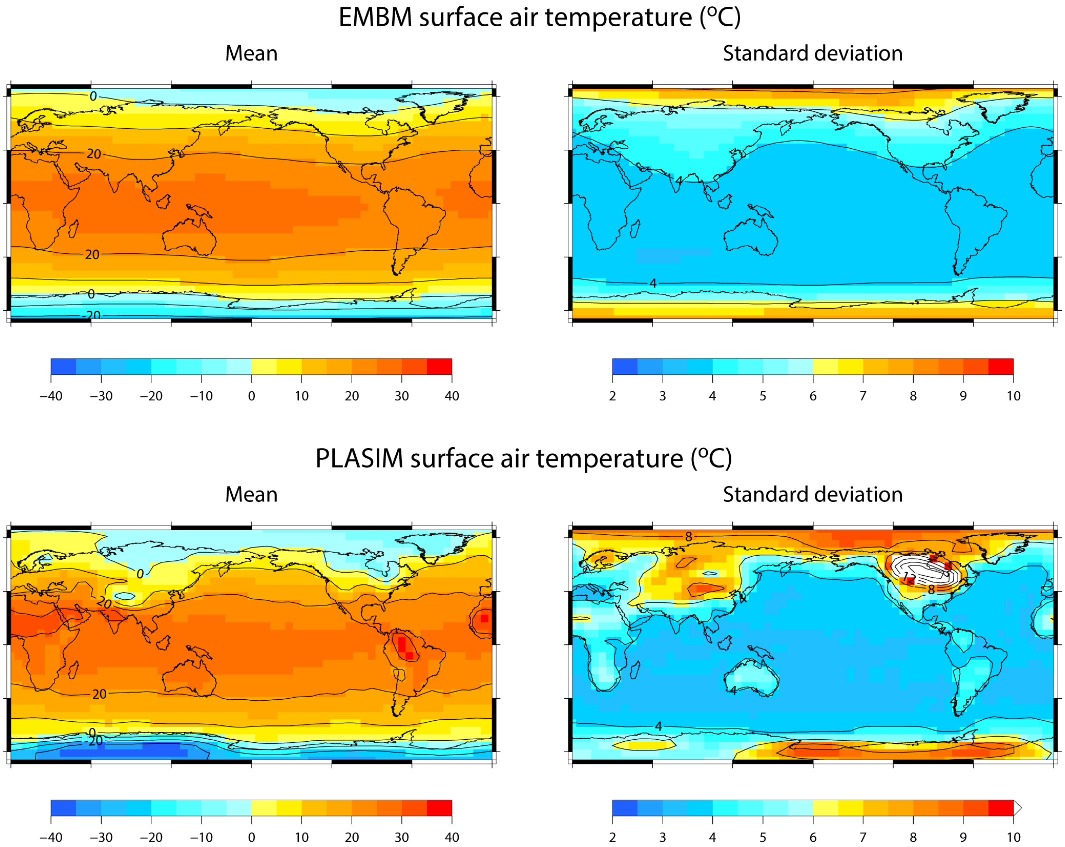
5 Results

5.1 Simulated climates

The EMBM output SATs are averaged over the final year of the 5000-year simulations while PLASIM output fields are averaged over the last 30 years. The ranges of some output variables obtained from the 600-member ensembles of GENIE-1 and PLASIM simulations are summarised in Table 2. The diversity of the output climate states is demonstrated by the large variation in SST, SAT, Antarctic sea-ice area and strength of the AMOC, which is weakened or shut down in some simulations. Because of the large upper limit of atmospheric $\mathrm{CO_2}$ concentration and GENIE-1's general bias towards low Antarctic sea ice, in some simulations, the Southern Ocean appears to be completely ice free. The SAT in PLASIM is lower in general and exhibits a slightly larger variation compared to EMBM's value.

Figure 2 shows the ensemble mean and standard deviation of PLASIM and EMBM SAT. Although similar spatial patterns are seen in both, PLASIM exhibits a larger variation spatially and across the ensemble, especially at high elevation. The comparison between the two models also shows that EMBM climate is much more zonal, with little land–sea difference. This is one of the known weaknesses of the energy–moisture balance model of the atmosphere, which is too diffusive (Lenton et al., 2006). A clearer distinction between the ocean and the continents is modelled in PLASIM as shown in the standard deviation plot of Fig. 2.

The differences seen in Table 2 and Fig. 2 are partly due to the nature of EMBM and PLASIM SAT fields. In contrast to PLASIM, EMBM does not take into account the effect of the elevation when calculating SAT. EMBM's and PLASIM's global mean SAT can be compared to the annual global mean SAT at the 1000 mb pressure surface and the 2 m surface from NCEP-DOE (National Centers for Environmental Prediction – Department of Energy) reanalysis (1979–2013), respectively. The two climatologies have global mean SAT of 8.5 °C (1000 mb) and 6.9 °C (2 m), respectively, correspond to a difference of 2.6 °C. The difference in each PLASIM–EMBM pair ranges from −1.3 to 6.1 °C, with a mean of 1.32 °C. Among simulations with modern glacier configurations and atmospheric $\mathrm{CO_2}$ within 340–400 ppm, the average difference is 1.51 °C, lower than the climatological value by approximately 1 °C. The large difference between the two ensembles can be attributed to the large parameter range and the difference in climate sensitivity. With dynamic wind and interactive cloud, PLASIM is expected to produce a more realistic precipitation pattern, especially over the continents.


EMBM surface air temperature (°C)

Mean

Standard deviation

PLASIM surface air temperature (°C)

Mean

Standard deviation

Figure 2. The mean and standard deviation of SAT across the 600-member ensembles of GENIE-1 and PLASIM. There are white cells on the PLASIM SD plot where the outputs go beyond the plotted range. The largest standard deviation in PLASIM is 17.5 °C. The contours on the mean and SD plots are shown every 10 and 4 °C, respectively.

Interactions between the atmosphere and the ice sheets can also lead to larger variations due to orography or precipitation feedbacks. Their climate sensitivities will be explored later on with the help of the GP emulators constructed.

The resulting SAT from both models are compared against climatology in Fig. 3 using Taylor diagrams. These plots demonstrate the range of output obtained with respect to modern climate. The modern climate states here serve as reference points to better demonstrate the spread of the simulated ensembles as well as their differences. Both the standard deviations (SD) and root mean square differences (RMSD) are normalised (and non-dimensionalised) by dividing them by the SD of the observations. GOLDSTEIN SST from all simulation runs are compared with annual mean SST (1900–2005) from NOAA World Ocean Atlas (Locarnini et al., 2006). The SATs from the single-layer atmosphere EMBM are compared with annual mean surface air temperature over the period from 1979 to 2013 at the 1000 mb pressure surface from NCEP-DOE reanalysis-2 (Kanamitsu et al., 2002). The SATs from PLASIM are compared with the air temperature at 2 m from the same reanalysis. The simulation runs with ice sheet configuration and CO_2 concentration similar to those within the 1979–2013 period, ICF $\in \{0, 1, 2, 3, 4\}$ and 340 ppm $<$ RFC $<$400 ppm, are highlighted in red. A plot showing the difference between the mean surface temperatures over this group of simulations and climatology is included in the Supplement (Fig. S1 in the Supplement).

The simulated pattern of SST correlates well with observation (average correlation coefficient of 0.95), while the majority of the ensemble exhibits smaller spatial variability than climatology (average normalised SD of 0.85). The spread in these modern GOLDSTEIN SST points is due to the large range of the varied GENIE-1 parameters. The standard deviations of SAT are also underestimated in EMBM (average normalised SD of 0.83). PLASIM SAT correlate well with the climatology (average correlation coefficient of 0.97). The spatial variation in PLASIM SAT has a similar mean to EMBM but has a larger range (both ensembles have average normalised SD of 0.83).

5.2 Scalar emulation

An emulator is first constructed for EMBM global mean SAT with a starting number of 30 training points. The coefficients of determination (r^2) and the root mean square er-

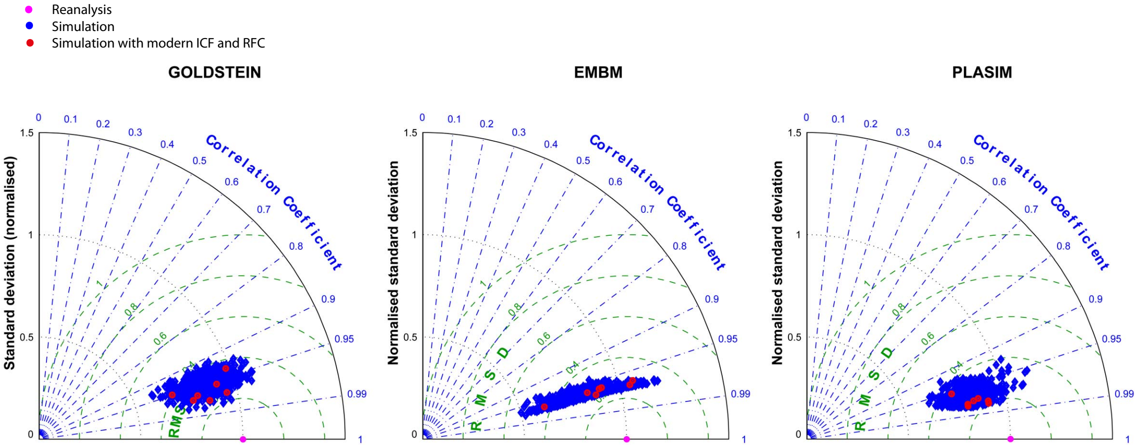


Figure 3. Taylor diagrams showing a comparison between model runs with climatology: GOLDSTEIN SST (left), EMBM SAT (middle) and PLASIM SAT (right). The magenta dots represent reanalysis taken from Locarnini climatology (1900–2005) (Locarnini et al., 2006) (left), NCEP-DOE reanalysis 2 annual mean SAT (1979–2013) at 1000 mb (Kanamitsu et al., 2002) (middle) and NCEP-DOE reanalysis 2 annual mean SAT (1979–2013) at 2 m (Kanamitsu et al., 2002) (right). The points highlighted in red represent runs with ICF $\in \{0,1,2,3,4\}$ and 340ppm $<$ RFC $<$ 400 ppm.

ror (RMSE) between the simulated and emulated validation points (Sect. 3.2) are computed and then used as indications of the validity of the emulator. The coefficient of determination, r^2, is the square of the sample correlation coefficient:

$$r^2 = \left(\frac{\mathrm{cov}(\mathbf{Y}, \widehat{\boldsymbol{Y}})}{\sqrt{\mathrm{var}(\mathbf{Y})\mathrm{var}(\widehat{\boldsymbol{Y}}))}} \right)^2. \tag{12}$$

More training points are gradually added to produce more accurate emulators with decreasing RMSE and increasing r^2. At approximately 200 points ($n_\mathrm{c} = 200$), adding more training data no longer significantly reduces the RMSE value. It is concluded that approximately 200 cheap points are sufficient to capture the variation over the EMBM output space. We then attempt to build co-kriging emulators for global mean SAT in PLASIM using 200 cheap points and additional expensive data points. Again, 30 expensive points are chosen for initial training. It is found that 50 PLASIM points ($n_\mathrm{e} = 50$) are enough to construct a good emulator with RMSE = 0.51 °C and $r^2 = 0.98$.

The number of training points required varies from one emulator to another since it depends strongly on the function being emulated. As the number of parameters increases, the dimension of the emulator also increases and hence more training points are required. Typically an average of 10 points per dimension is assumed. This, however, depends on how non-linear or how "active" the function is. A highly non-linear function might require many more points while a more linear function might not need as many as 10 points per dimension.

Kriging emulators using only expensive points are also constructed to provide comparison between the two techniques. When the same amount of training data is used, co-kriging outperforms kriging. More expensive points are then added to improve the kriging emulator until a similar value of RMSE is obtained. In this case, the kriging emulator using $n_\mathrm{e} = 200$ PLASIM training points gives RMSE = 0.50 °C and $r^2 = 0.98$. Therefore, co-kriging achieves of the same level of accuracy with only 25 % as much expensive data.

A second pair of emulators is produced for the global SAT anomaly from SST (global annual mean SAT minus SST). In this case, the component of the SAT response that is a trivial function of the boundary conditions is removed. Following the procedure described above, a co-kriging emulator using 70 expensive points and 250 cheap points were constructed and compared to a kriging emulator using only 70 expensive points. The RMSE and r^2 are included in Table 4. The co-kriging emulator obtains RMSE = 0.31 °C and $r^2 = 0.95$. This time, a kriging emulator using 100 expensive points gives similar validation result, RMSE = 0.33°C and $r^2 = 0.92$. The co-kriging emulator still manages to utilise meaningful information from EMBM, albeit not as well as in the previous example, and reduces the expensive points needed by approximately 30 %.

For both kriging and co-kriging emulators using the same expensive training points, the emulated global mean SATs at the 214 validation points are plotted against their simulated values (Fig. 4). The corresponding RMSE and r^2 values are shown in Table 3. Tables 3, 4 and Fig. 4 show that the co-kriging emulators reproduce the simulated values more accurately. Tables 3 and 4 also contains the ensemble mean and standard deviation from both co-kriging and kriging emulators, compared with the true values obtained from the simulated ensemble.

While co-kriging outperformed kriging in both cases, multi-level emulation does a much better job at predicting SAT than SAT minus SST. Nevertheless, the r^2 scores be-

Table 3. Validation results for kriging and co-kriging emulators of PLASIM global mean SAT. The co-kriging emulator uses 50 expensive points and 200 cheap points while the kriging emulator here uses the same 50 expensive points.

	Kriging emulator	Co-kriging emulator	Simulated ensemble
RMSE (°C)	0.93	0.51	N/A
r^2	0.94	0.98	N/A
Ensemble mean (°C)	10.96	11.30	11.40
Ensemble SD (°C)	4.89	4.73	4.57

Table 4. Validation results for kriging and co-kriging emulators of PLASIM global mean SAT – SST. The co-kriging emulator uses 70 expensive points and 250 cheap points while the kriging emulator here uses the same 70 expensive points.

	Kriging emulator	Co-kriging emulator	Simulated ensemble
RMSE (°C)	0.42	0.31	N/A
r^2	0.91	0.95	N/A
Ensemble mean (°C)	−5.81	−5.77	−5.72
Ensemble SD (°C)	1.65	1.70	1.50

tween simulated and emulated values from the co-kriging emulators are over 0.90 for both. The standard deviations across the ensembles are slightly overestimated in both emulators. From the figure, the emulated values can be seen to deviate more for larger anomalies.

The uncertainty in the emulator predictions, arising from not having evaluated the model at untried input configurations, is called the “code uncertainty” (O’Hagan, 2006). An advantage of the GP emulator employed is that we can quantify this uncertainty, which is represented as the error bar at each prediction in Fig. 4. The additional information from the cheap training data helps reduce this uncertainty for the co-kriging emulator.

5.3 EOF decomposition

The following analysis attempts to explain the processes and parameters that determine the spatial distributions of SAT in GENIE-1 and PLASIM using PCA. SVD was applied to two $(2048 \times n)$ matrices of EMBM and PLASIM SAT fields, where $n = n_\mathrm{c} = n_\mathrm{e} = 660$. Over 99 % of the variance across the ensemble in these fields can be explained by the top 10 EOFs, as shown in Table 5. This indicates that they are sufficient to generate a good approximation to the simulated responses. As suggested from the emulator for global mean SAT, less than 600 points would be sufficient for the emulators. To ensure that the decomposition is robust, SVD is applied on smaller subsets ($n = 30$ to $n = 250$). The EOFs appear to be qualitatively the same. Only minor quantitative differences are obtained, therefore, the EOFs and PCs are judged as robust and representative of the ensemble behaviour. These subsets are chosen using the same exchange algorithm mentioned in Sect. 4.2 to obtain designs that give the best space-filling Morris–Mitchell criterion (Morris and Mitchell, 1995).

Table 5. Percentage of variance in SAT, explained by the first 10 EOFs for GENIE-1 with EMBM and with PLASIM. The 150-member ensembles are used to obtain these values.

	EMBM	PLASIM
EOF 1	86.33 %	79.53 %
EOF 2	11.27 %	8.62 %
EOF 3	1.55 %	6.85 %
EOF 4	0.47 %	2.61 %
EOF 5	0.10 %	0.43 %
EOF 6	0.07 %	0.57 %
EOF 7	0.05 %	0.30 %
EOF 8	0.03 %	0.21 %
EOF 9	0.03 %	0.16 %
EOF 10	0.03 %	0.07 %
Total	99.93 %	99.35 %

The high percentage of variance explained by the retained EOFs mean that by successfully emulating them, the SAT field of PLASIM can be accurately estimated. For EMBM data to be useful, its EOFs and PCs need to carry meaningful information about PLASIM’s modes. To verify this, an analysis of the EOFs and PCs of the two models are carried out.

The first EOFs of SAT in both models are illustrated in Fig. 5. Their corresponding PCs are emulated as functions of the model parameters using universal kriging (Sect. 4.1). Also shown in this figure are the emulator coefficients, β (as described in Eq. 3), which reflect the relative importance of the parameters in determining each PC. These coefficients are the gradients of the linear mean function fitted to the data. Each coefficient corresponds to a dimension or an input parameter. They are not purely objective measures since their values depend on the ranges over which the parame-

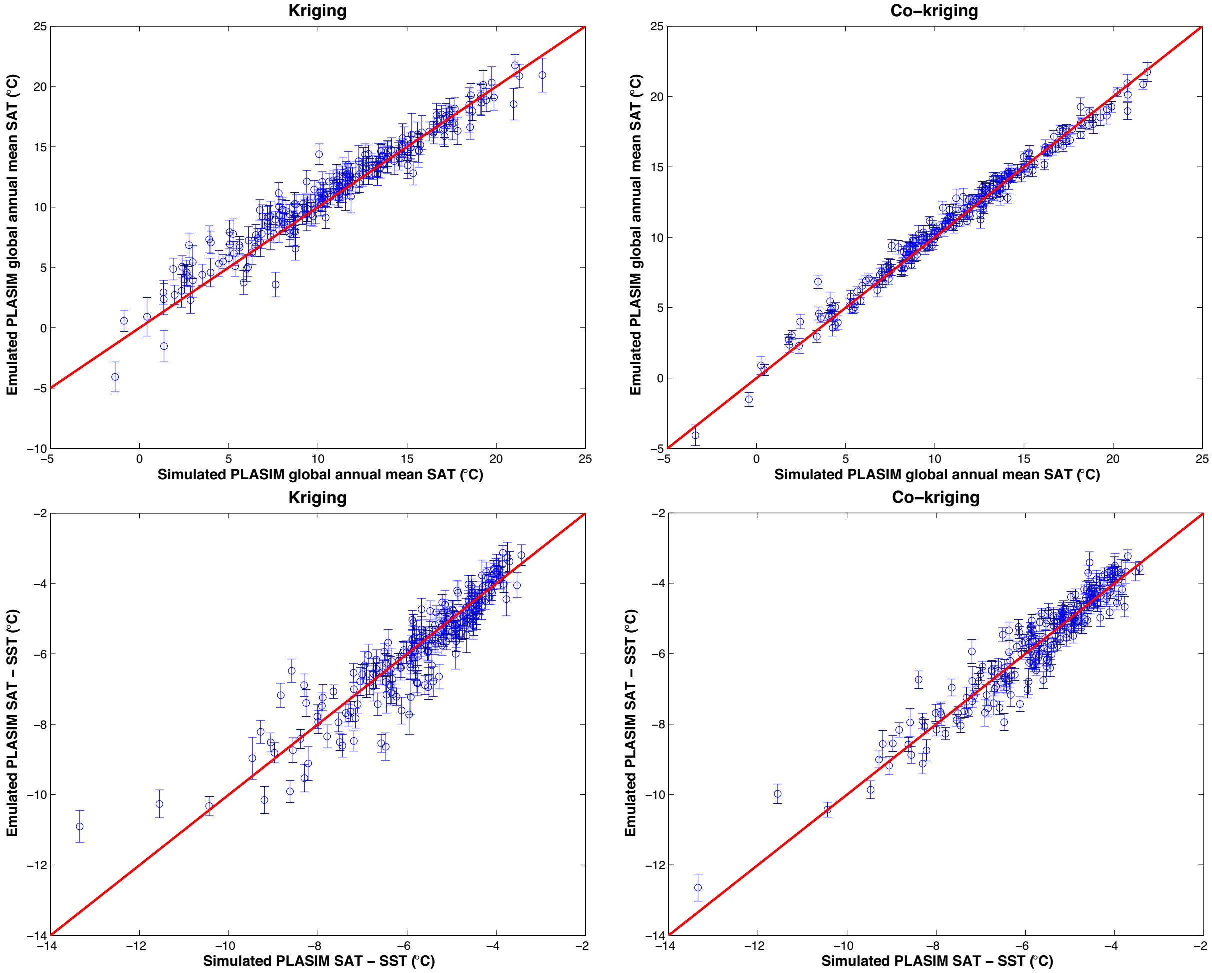


Figure 4. The upper panels show PLASIM simulated global mean SAT at the 214 validation points plotted against their emulated values from both kriging (left) and co-kriging (right) emulators. The error bars indicate a 2 standard deviation interval at each point. The lower panels show the results of the global mean SAT–SST emulators.

ters were varied. Also, the mean function is linear so they do not contain information on the non-linearity of the emulated function. They also inherit uncertainties from imperfect emulation.

The first EOF for both models is of the same sign globally, suggesting a change in the radiation budget due to the greenhouse gas and the albedo effects. The effects due to changing glacier condition and atmospheric CO_2 concentration are accentuated in PLASIM because corresponding changes are taken into account in PLASIM. According to the emulator coefficients, the largest contributions are due to RFC, OL0, RMX and ICF in both PLASIM and EMBM. Large values of ICF result in a lower global mean SAT due to higher albedo. Large values of RFC, OL0 and RMX, on the other hand, have the opposite effect on global mean temperature due to more heat being absorbed by the increased greenhouse gas content in the atmosphere. Hence, ICF has the opposite sign to RFC, OL0 and RMX.

The second EOFs in EMBM and PLASIM exhibit changes of opposite sign at Equator and polar regions, reflecting a redistribution of the heat budget (Fig. 6). The parameters controlling heat diffusivity in the atmosphere (AHD and AMD) play the largest role in this process. While they dominate the signals, there are smaller contributions from the ocean heat diffusivity parameters (OHD and OVD), which have similar but smaller effects compared to AHD and AMD. Other small signals do not necessarily agree with each other; i.e., RFC has opposite signs in the two models.

With emulator coefficients of approximately 0, the dummy variable is correctly identified as an inactive parameter in all cases (Figs. 5 and 6), giving us more confidence in using the coefficients. Any parameter with coefficients of comparable magnitude to FFX is also assumed to be inactive, such as OHD and OVD for EMBM and PLASIM's first EOF.

These EOFs indicate similar modes of variability in GENIE and PLASIM, fulfilling the assumption made for co-

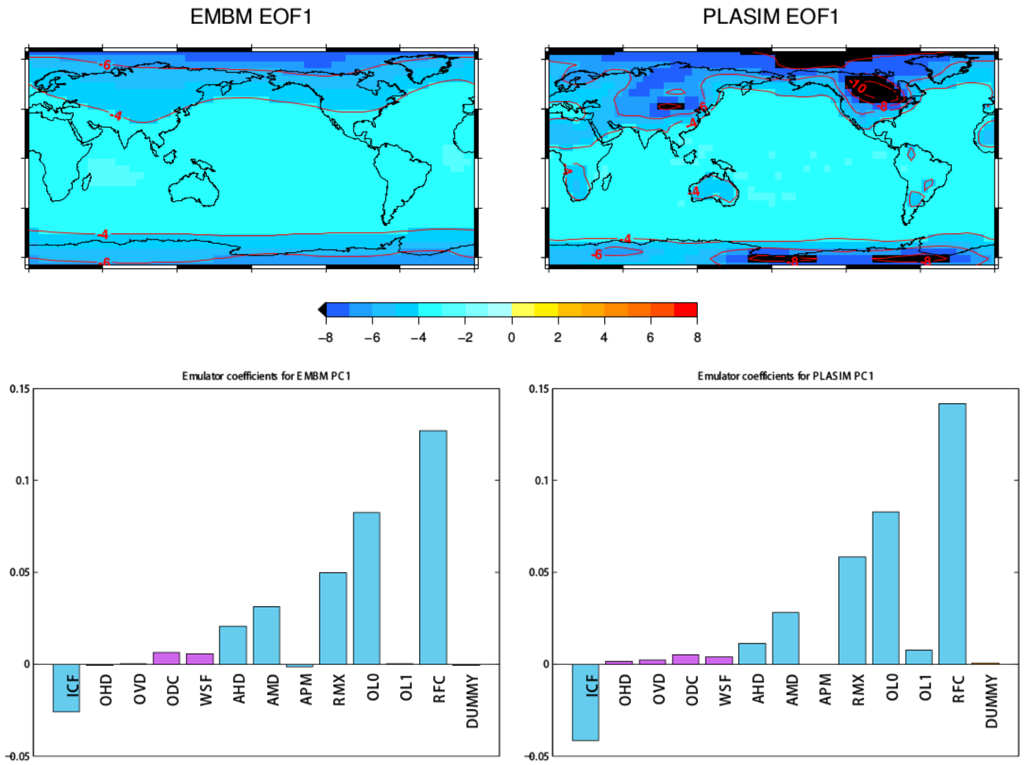


Figure 5. The first EOFs of EMBM and PLASIM SAT (upper) and the universal kriging emulator coefficients of their corresponding PCs (lower). All 600 data points are used to train each of these emulators. The black cells in PLASIM EOF1 indicate values lower than the plotted range. Contours are drawn over both plots at a 2 °C interval.

kriging. The extra training points from EMBM, therefore, are expected to provide inference on PLASIM's behaviour. Each pair of PCs from EMBM and PLASIM form a set of cheap and expensive training data for the corresponding emulator. Even though this is applied to all 10 PCs, according to Table 5, only the first 4 modes contribute significantly to the total variance. Lower-order modes appear indistinguishable to noise. It is difficult to emulate them independently and so it is unlikely that any meaningful relationship between them can be found by co-kriging.

Although all 600 data points are used to train each of these emulators, results obtained from smaller subsets show no systematic differences.

The assumptions made for Eq. (6) are expected to hold in the case of emulating PLASIM's PCs. The emulator coefficients in Figs. 5 and 6 show that the PCs of the two models exhibit similar trends due to the varying input parameters. The difference in the magnitude of the contributions from these parameters should be sufficiently approximated using a scaling constant, ρ, and a stochastic process, f_d. The spatial pattern in PLASIM, however, depends on the EOFs and so different regional responses compared to EMBM can still be emulated using this method.

5.4 Emulation of 2-D output fields

We retained the first 10 EOFs of EMBM and PLASIM SAT, which describe 99.93 and 99.35 % of the simulated ensemble variance, respectively (Table 5). Each individual field can be approximated as a linear combination of these 10 EOFs, scaled by their respective PCs according to Eq. (8). Using this method of dimensional reduction, only 10 emulators or less are needed instead of 2048 emulators if each individual grid point is emulated. Both kriging and co-kriging emulators are then constructed for each of these PCs.

Using the same procedure as described in Sect. 5.2, exploratory exercises show that approximately $n_\mathrm{c} = 150$ training points are needed to obtain a good emulator of the EMBM SAT fields. The cheap data are, therefore, the 150 indices of each of the first 10 rotated PCs of the (2048×150) matrix of GENIE data. It is found that at $n_\mathrm{e} = 50$, we obtain a co-kriging emulator that validates well against simulated values.

Kriging emulators using only the expensive data from PLASIM are also constructed for comparison. Again, co-kriging outperforms kriging when the same 50 expensive training points are used. More expensive points are then added to the kriging emulators and for approximately 150 points, similar RMSE and r^2 are obtained. Therefore,

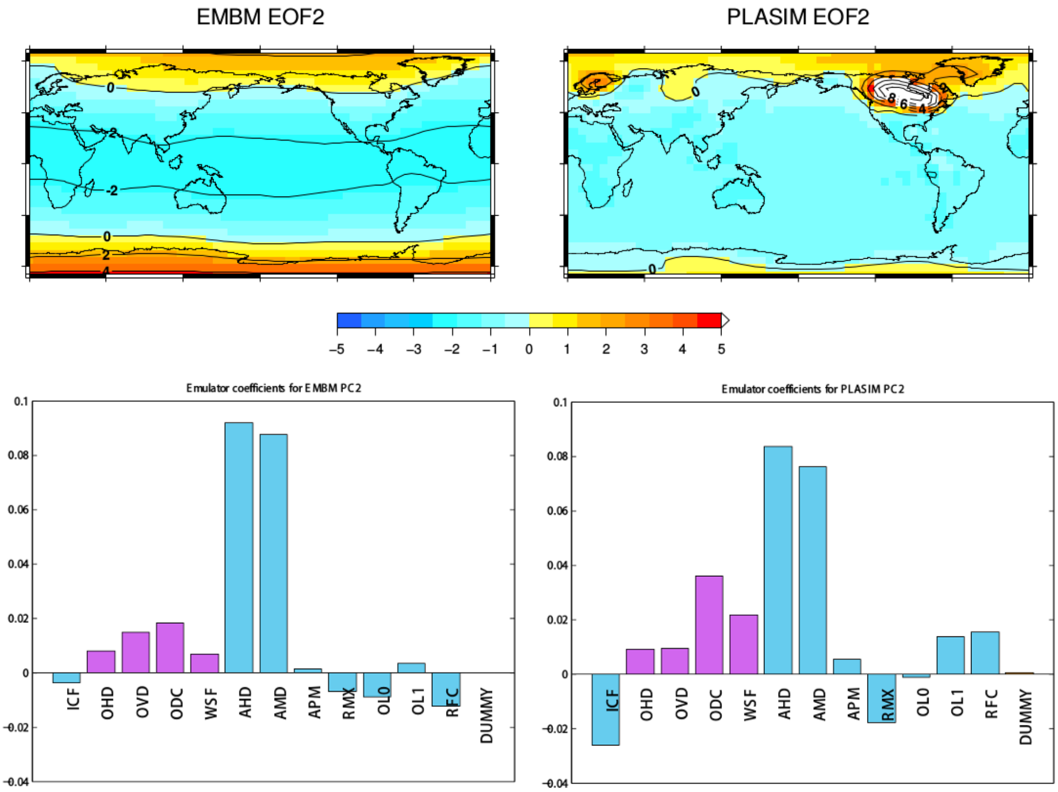


Figure 6. The second EOFs of EMBM and PLASIM SAT (upper) and the universal kriging emulator coefficients of their corresponding PCs (lower). All 600 data points are used to train each of these emulators. The white cells in PLASIM EOF2 indicate values higher than the plotted range. Contours are drawn over both plots at a 2 °C interval.

co-kriging reduces the required expensive training data to one-third of the amount needed when using kriging.

The co-kriging (trained with 50 expensive and 150 cheap points) and kriging (trained with 50 expensive points) are validated using the 214-member validation set. Both the individual PCs and the final reconstructed SAT are validated against true values. First, to test the emulator's ability to reproduce PC values, each emulated PC is validated against those decomposed from the simulated ensemble (Table 6). For the first score, co-kriging emulator validated very well with an r^2 value of 0.97. Lower-order PC coefficients are generally harder to emulate; hence, the value of r^2 decreases down the list. It is possible that they reflect physical processes that are more difficult to represent as simple functions of the input parameters or simply represent stochastic processes. With a low value of r^2, the emulator does little more than adding some random noise, e.g. from the 6th to the 10th PCs, with the exception of the 9th. There are several reasons for this. First, the PCs of EMBM might reflect random noise and so cannot be emulated. Since the cheap emulators are not meaningful, the expensive ones can gain no useful information. Second, PLASIM's PCs might be noise and co-kriging fails to work for the same reason. Finally, the relationships between EMBM and PLASIM PCs might not have been successfully determined. This either means that EMBM did not contain the information on these PLASIM's modes or the emulator fails to determine it. Even though the signal from the 9th mode is very small, it was emulated with some success. Despite the fact that mode 6, 7, 8 and 10 were not emulated successfully, co-kriging still performs either comparably or better than kriging.

The 10 co-kriging emulators of PLASIM PCs are then used to reconstruct the SAT fields at each validation point. To validate the simulated SAT fields, the quality of the individual emulations and the spatial pattern of the emulated field are tested. In order to test the proportion of the total ensemble variance captured by the emulator:

$$V_T = 1 - \sum_{n=1}^{59}\sum_{i=1}^{2048}(S_{n,i} - E_{n,i})^2 \Bigg/ \sum_{n=1}^{59}\sum_{i=1}^{2048}(S_{n,i} - \bar{S}_i)^2, \quad (13)$$

where $S_{n,i}$ is the simulated output at grid cell i in the nth member of the validating ensemble, $E_{n,i}$ is the corresponding emulated output and $\bar{S}_i$ the ensemble mean simulated output at grid cell i. V_T assesses the error in the emulator for each simulation, averaged over the 59 simulations, and measures the degree to which individual simulations can be regarded as accurate. The RMSE values between each emulated and simulated surface fields are calculated and averaged across

Table 6. Validation of each PC emulator using the 59-member ensemble. The correlation coefficients show how well matched the emulated PCs are compared with the simulated values. The co-kriging emulator uses 50 expensive points and 150 cheap points while the kriging emulator here uses the same 50 expensive points.

	Principal component emulator									
	1	2	3	4	5	6	7	8	9	10
Kriging r^2	0.91	0.75	0.84	0.50	0.15	0.04	0.00	0.09	0.24	0.00
Co-kriging r^2	0.97	0.83	0.84	0.64	0.18	0.05	0.02	0.10	0.24	0.00

the whole validation set. V_T and RMSE are used in combination to assess the emulator validity.

Figure 7 demonstrates the effect of each added PC to the value of V_T and RMSE. When only the first emulated component is considered, the co-kriging emulator reproduces 76.2 % of the simulated variance (averaged over all space and all ensemble members), which is close to the 79.5 % variance explained by the first EOF (Table 5). This is also reflected by the high degree of accuracy of the PC 1 emulator (Table 6). The addition of the next four emulated components brings the percentage of simulated variance being captured, V_T, to 93.2 %, close to the total amount of 98.0 % explained by the first five EOFs (Table 5). The average RMSE is 1.33°C, which is approximately 1.7 % of the average spatial variation in temperature or 4.8 % of the average variation across the whole ensemble at each grid point. The last five emulated PCs have a negligible effect on both V_T and RMSE. Among these, only the 9th PC improves the overall result while the others worsen it. For the kriging emulators, the same behaviour is observed but with lower accuracies. The maximum variance explained by the kriging emulators is 85.3 %. Also included in Fig. 7 are lines corresponding to the validation results if the emulators were perfect. These demonstrate the errors introduced by the dimensional reduction process.

Figure 8 shows the emulated and simulated spatial pattern of the ensemble mean and standard deviation. The differences between these emulated and the simulated fields are within 1 °C. Therefore, the ensemble behaviour is well reproduced. There is, however, a slight underestimation of the SD over the Northern America continent where the glacier mask is applied. The 2-D SAT emulator appears to underestimate the ensemble variance by a small amount. The error seen is a combination of the two types of errors introduced in Sect. 4.1. Despite having very different outputs (Fig. 3), the method proposed successfully utilises GENIE-1's EMBM output to aid the construction of PLASIM SAT emulator.

In the work presented here, only annually averaged fields are considered. The generalisation to emulate monthly average fields or seasonal cycles is straight forward. We simply have to replace the current (2048 × 1) annual-averaged maps with a (24576 × 1) map of the 12 monthly averaged fields.

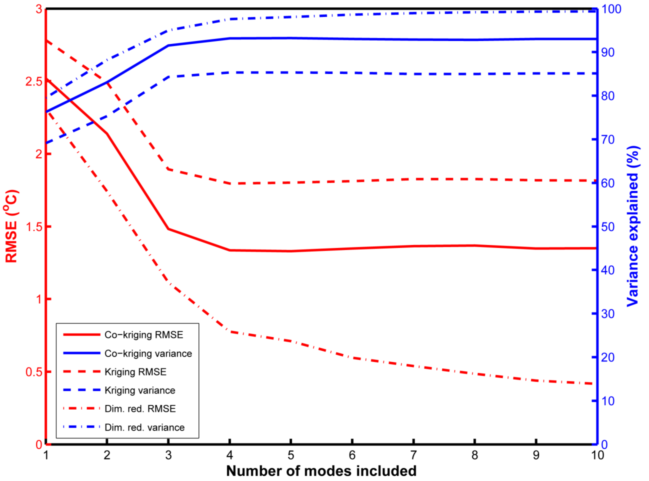


Figure 7. Comparison between kriging (dashed line) and co-kriging (solid line) emulators. The variance explained (blue) when each PC is added is shown together with the RMSE (red) of the corresponding reconstructed validation SAT fields. The dot-dashed lines represents the same values obtained if the emulator were perfect. The deviations of these line from RMSE = 0 °C and $V =$ 100 % are errors introduced by dimensional reduction.

5.5 Relationship with the coupled system

We have demonstrated that information from a cheap atmospheric model (EMBM) can be used to improve predictions of the steady-state behaviour of an expensive atmospheric model (PLASIM) in unsampled parts of parameter-/boundary-forcing space. This behaviour is a function of the boundary conditions on the atmospheric model (SST, longwave and shortwave radiative forcing), as represented in this statistical study by the 13 parameters. This technique has advantages when attempting to understand or project the decoupled response of individual climate system components to their boundary conditions. For example, in the context of impact assessment models, the spatial pattern of changes in SAT and precipitation is often needed to study the impact of climate change on areas such as health, land use and energy production. These spatial temperature and precipitation response patterns are obtained from climate models forced by arbitrary CO_2 concentrations resulting from particular policy decisions. Different statistical emulation techniques have

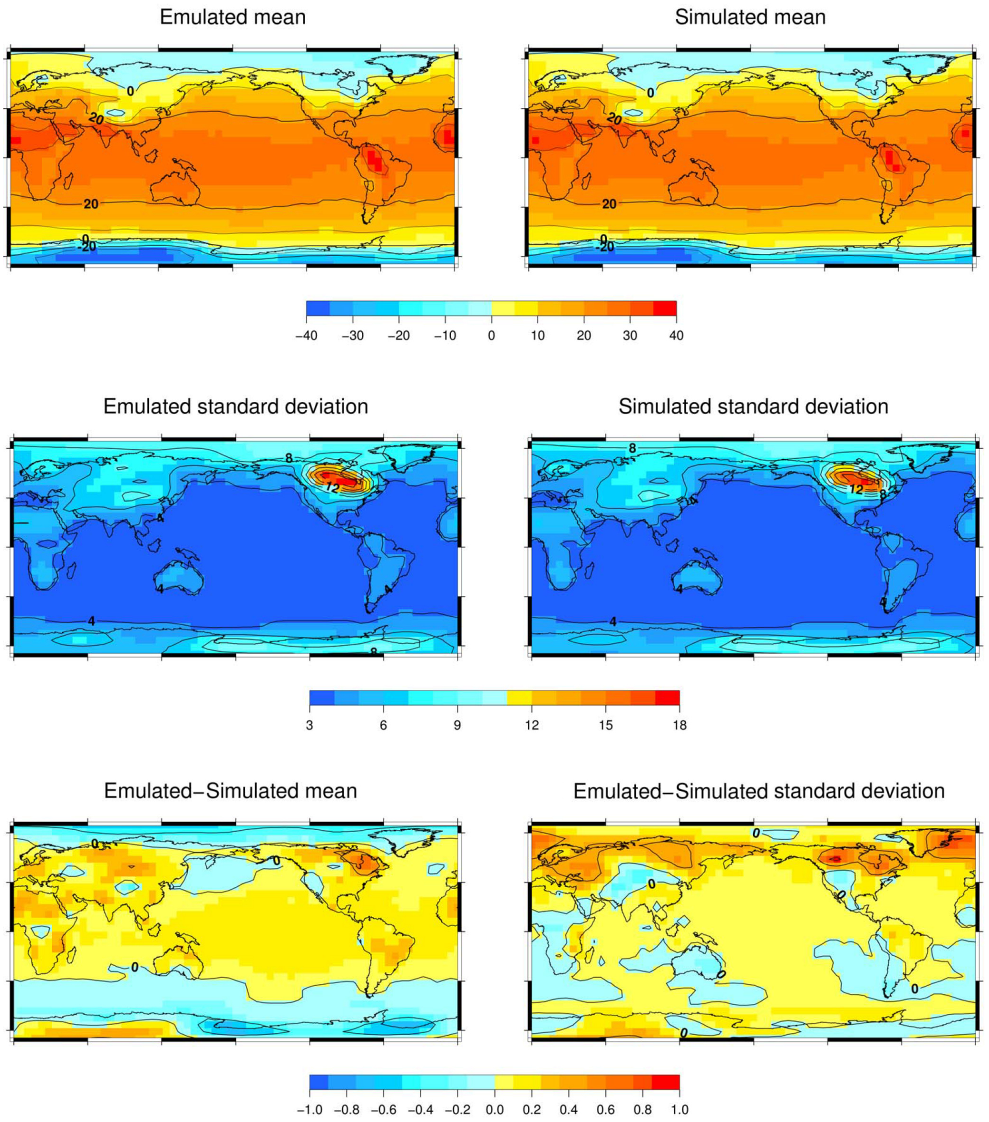


Figure 8. Mean and standard deviation of the emulated (upper and middle left) and simulated (upper and middle right) validating ensembles. The emulated–simulated differences in mean (lower left) and standard deviation (lower right) are also shown.

been employed to reproduce the output of AOGCMs under a large range of forcing scenarios (Holden et al., 2014; Castruccio et al., 2014). Our multi-level emulation technique offers an alternative method to reproduce the key characteristics of an AOGCM using only a small training set, given a larger ensemble of a cheaper model of the same system, covering unsampled CO_2 concentrations. Another example where our technique can be applied is in emulating a carbon cycle model to provide an estimation of the atmospheric CO_2 concentration as a function of a time series of anthropogenic CO_2 emissions and non-CO_2 radiative forcing (Foley et al., 2016). CO_2 concentration from coupled climate–carbon cycle models can be emulated and replace the simple carbon cycle component often used in integrated assessment models.

In reality, changes to the climate system components that are focused on will feed back on other climate system components; i.e., if the present study were extended to the fully coupled system, differences in SAT, wind stress and the hydrological cycle between PLASIM and the EMBM would feed back on SST and sea-ice distribution.

Within this context, we now explore the relationship between the "climate sensitivities" of the EMBM and PLASIM atmospheres, both forced by GENIE–EMBM SSTs as discussed above, before considering how our approach could in future be extended to the fully coupled system. Our 600-member ensemble design generated in Sect. 3.2 is used as

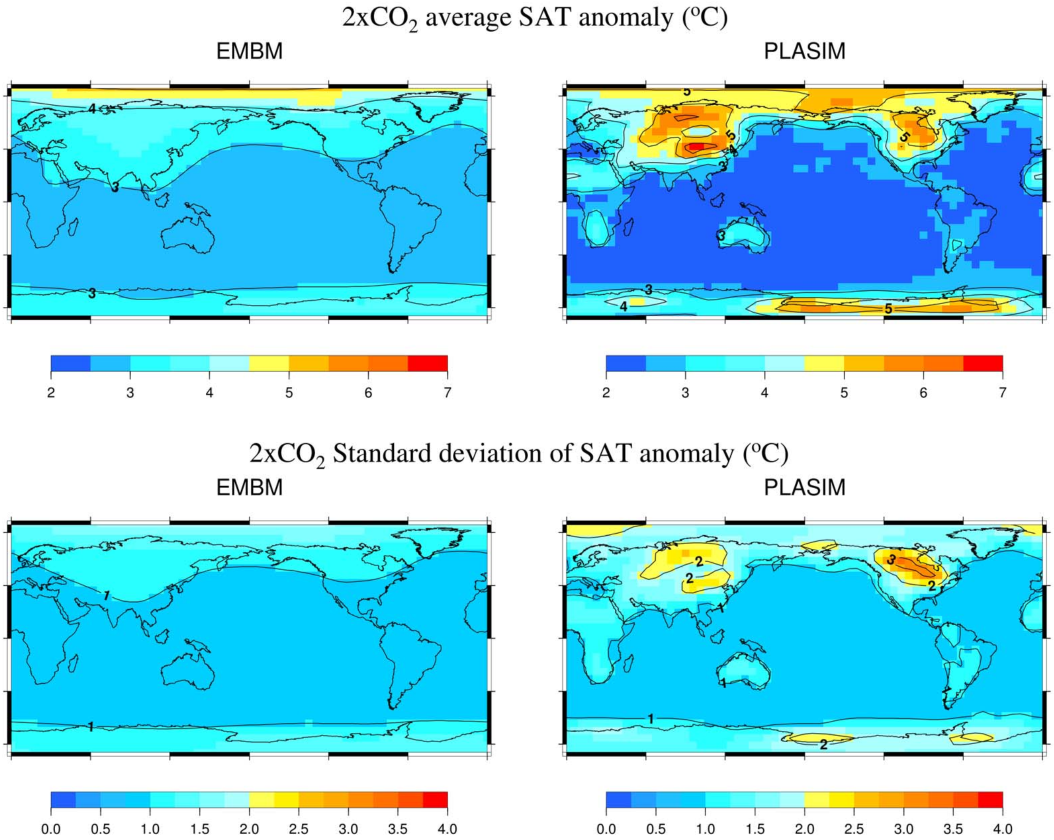

2xCO$_2$ average SAT anomaly (°C)
EMBM
PLASIM
2
3
4
5
6
7
2xCO$_2$ Standard deviation of SAT anomaly (°C)
EMBM
PLASIM
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0

Figure 9. Mean (upper panel) and standard deviation (lower panel) of the SAT anomaly corresponding to a double in atmospheric CO_2 concentration in EMBM and PLASIM.

the basis of two new designs. ICF is fixed at 0 for both sets. Climate sensitivity is defined as the warming response to a doubling of atmospheric CO_2 from the pre-industrial values. Hence, a control set (CTRL) has RFC set to 278 ppm and another set ($2 \times CO_2$) has RFC set to 556 ppm. The emulators constructed in the previous section are used to predict the SAT fields resulting from these two designs. This process can be done within seconds, at almost no additional computational cost.

The average SAT anomalies due to a doubling of atmospheric CO_2 concentration for both models, $2 \times CO_2-$CTRL, are shown in Fig. 9. The area-weighted global mean SAT are used to calculate the probability distribution of climate sensitivity for the two models, shown in the upper panel of Fig. 10. The means of the two distributions are ΔTCO_2 of 2.99 ± 0.91 °C for EMBM and 3.37 ± 0.95°C for PLASIM. Figure 10 shows that the climate sensitivities in the two models have similar distributions with means differing by approximately 0.38 °C. The range is broad due to the parameters varied. PLASIM displays larger changes in temperature over the continent in general and especially over high elevation areas (Fig. 9). Because of this, the average anomaly ΔTCO_2 in a PLASIM simulation is larger than the corresponding value in EMBM. The relationship between the two distributions is approximately linear, as shown in the lower panel of Fig. 10. Since no PLASIM parameter is varied apart from ICF and RFC (which are both held constant in this experiment), PLASIM climate sensitivity is heavily influenced by the GOLDSTEIN surface conditions.

In a hypothetical coupled experiment, it is reasonable to speculate that the generally larger response of SAT to CO_2 in PLASIM than the EMBM would yield a broader range of SSTs in the GOLDSTEIN ocean, amplifying the difference in climate sensitivity between the two models. There are two alternative approaches that could be used to extend the technique described here to this fully coupled system. The first or "direct" approach sees PLASIM fully coupled to GENIE-1's subcomponents, allowing for two-way interaction between the atmosphere and the ocean/sea ice. In this case, the current statistical technique can be applied directly to emulate atmospheric variables from PLASIM as functions of the ocean's parameters, using EMBM as the cheap approximation. How beneficial EMBM's information is in this set-up compared to the result presented in Sect. 5.2 and 5.4 is uncertain without further work. The "indirect" approach involves the coupling of PLASIM's steady-state emulators

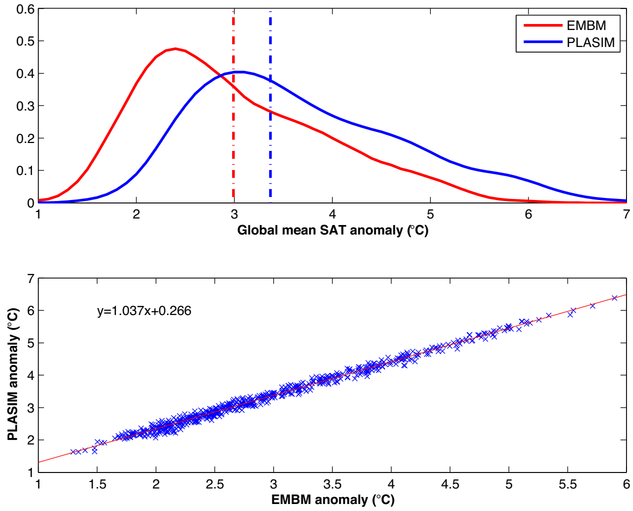


Figure 10. The upper panel shows the probability distributions of EMBM (red) and PLASIM (blue) climate sensitivities. The mean of each distribution is denoted by the dot-dashed line of the same colour. The lower panel shows a plot of PLASIM anomalies against EMBM anomalies. The coefficients of the linear function fitted through the data are included in the figure.

with GOLDSTEIN ocean and sea-ice components. Atmospheric output from PLASIM, such as SAT, precipitation and wind stress can be emulated as a function of the prescribed SST and used, in return, as boundary conditions for the ocean. This framework would be able to capture some processes, which are currently not adequately modelled or not represented at all in EMBM. There are certain implications for when such a framework would be useful since the emulators are built upon a collection of steady-state simulations where only one-way interaction between PLASIM and the ocean component is available. This type of framework would not be suitable in the context of processes such as ENSO (El Niño–Southern Oscillation) in which the atmosphere and ocean vary together on interannual timescales. However, it may be useful when events with much longer timescales, where the atmosphere can regarded as being at equilibrium with the ocean, are considered. While information on chaotic higher-frequency atmospheric variability is lost, extra information from the higher-fidelity atmospheric model is gained without incurring a large computational cost.

6 Summary and conclusions

We have described in this paper the development and evaluation of large ensembles of GENIE-1 and PLASIM simulations for application in statistical emulation.

For this work, we employ the non-parametric fitting method of Gaussian process emulation. Two variations of this well-established method, kriging and universal kriging, are briefly described in Sect. 4.1. Compared to polynomial fitting techniques, such as the one employed by Holden et al. (2014), this approach provides an estimate of the uncertainty introduced by the emulation process, also referred to as “code uncertainty”.

To efficiently extend this method from emulating scalar output to emulating high-dimensional output, e.g. the 2-D SAT fields, principal component analysis is used. This powerful technique decomposes the output surface fields of both EMBM and PLASIM models into orthogonal EOFs, scaled by the respective PCs. The EOFs are, however, statistical modes and direct connection to physical processes cannot always be drawn directly. Emulator coefficients of the PCs corresponding to these modes, however, can provide a link between them and the varying model parameters, allowing for better interpretation of the model behaviour. It also allow us to identify and preserve the correlation between grid cells.

Here, the first five PCA modes are emulated instead of individual grid cell values, reducing the computational cost significantly. Although not explored in this work, the links between different model outputs may also be exploited to allow for further reduction of dimension when emulating multivariate output.

A multi-level emulation technique, co-kriging, is used to build both scalar and high-dimensional output emulators for PLASIM with additional information from EMBM. The constructed co-kriging emulators successfully estimate both the global mean SAT and the 2-D array of SAT fields of PLASIM as functions of the 13 GENIE-1 parameters. Being cheaper to evaluate, EMBM can be used to sample GENIE-1’s parameter space more finely, providing information where PLASIM data are sparse. Despite being structurally unrelated, the link between EMBM and PLASIM is successfully established, resulting in PLASIM emulators being built using a smaller amount of expensive data. The combination of PCA with co-kriging allows us to emulate accurately the spatial pattern of PLASIM SAT despite the model having a different response to EMBM’s. Emulated outputs are validated against simulated values using a separate validation ensemble. Both spatial pattern and magnitude of SAT are well reproduced across the ensemble. Apart from the ensemble mean and standard deviation, individual simulations are also successfully emulated with high accuracy. The emulators, however, show a tendency to underestimate the variance spatially and across the ensemble. This is unavoidable because of the dimensional reduction process. The quantification of the emulator uncertainties are beyond the scope of this paper and should be explored in further studies in order to improve the emulators’ performance.

Here, we have focused only on SAT but this method can be applied to other variables of the atmosphere, such as precipitation (PPTN) or wind fields. In the case of PLASIM, co-kriging emulation of PPTN using GENIE’s PPTN field as a fast approximation is not likely since the description of this field in the two models differed quite significantly. The same goes for other PLASIM quantities, which have no equivalent in EMBM. However, it is possible that other GENIE-1 fields might be more suitable as the fast approximation to PLASIM's PPTN, e.g. SST or elevation. Work has been done in the past using elevation as a fast approximation for PPTN (Hevesi et al., 1992).

This work establishes the technique for emulating the equilibrium response of the model. Compared to available efficient frameworks such as the MIT IGSM-CAM (Massachusetts Institute of Technology – Integrated Global System Model linked with the National Center for Atmospheric Research (NCAR) Community Atmosphere Model) (Monier et al., 2013), a present limitation of this technique is in the scope for two-way coupling (e.g. in the present study the PLASIM atmosphere passively responds to the ocean). However, a future study will show that it is possible to emulate the atmospheric fields (precipitation, surface winds, etc.) that directly influence other model components and use these as boundary conditions. This technique has the limitation that the atmosphere is treated as being in a steady state with the ocean, so that the effect of interannual variability cannot be explicitly represented, but would nevertheless be of value for modelling long-timescale phenomena such as glacial-interglacial cycles.

We have demonstrated that multi-level emulation across structurally unrelated models provides useful information more efficiently than using either model in isolation. Several challenges remain before a coupled model making use of such an emulator can be constructed, and the steady-state vs. transient issue is one of them. The seasonality, which is currently lacking, will also be included by the modification described in Sect. 5.4. PLASIM's parameters, which do not have an equivalent in EMBM, are not yet considered. The current experiment design does not allow for the effect of aerosols, sea ice or vegetation to be studied. It simply attempts to improve the current simulated climate in GENIE-1 by incorporating the dynamic of PLASIM atmosphere. The role of these parameters will likely be explored in future studies.

The advantage of the emulation technique used here is that it does not depend on a fix set of models and can be applied to a wide range of models for different applications. It also provides a useful tool in coupling models of different fidelity and resolutions. The emulators, however, are built for specific applications and so care should be taken to avoid extrapolating beyond the emulated space.

In conclusion, the work presented here demonstrates a concept with applications in not only climate research but extending to a wide range of problems where multi-level computer models are available.

The Supplement related to this article is available online at doi:10.5194/ascmo-2-17-2016-supplement.

Edited by: C. Forest
Reviewed by: four anonymous referees

References

Castruccio, S., McInerney, D. J., Stein, M. L., Crouch, F. L., Jacob, R. L., and Moyer, E. J.: Statistical emulation of climate model projections based on precomputed GCM runs, J. Climate, 27, 1829–1844, 2014.

Challenor, P. G., McNeall, D., and Gattiker, J.: Assessing the probability of rare climate events, in: The Oxford handbook of applied bayesian analysis, edited by O'Hagan, A. and West, M., chap. 16, 403–430, Oxford University Press, New York, 2010.

Conti, S. and O'Hagan, A.: Bayesian emulation of complex multi-output and dynamic computer models, J. Stat. Plan. Infer., 140, 640–651, doi:10.1016/j.jspi.2009.08.006, 2010.

Cook, R. D. and Nachtsheim, C. J.: A comparison of algorithms for constructing exact D-optimal designs, Technometrics, 22, 315–324, 1980.

Cumming, J. and Goldstein, M.: Small Sample Designs for Complex High-Dimensional Models Based on Fast Approximations, Technometrics, 51, 377–388, 2008.

Edwards, N. R. and Marsh, R.: Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model, Clim. Dynam., 24, 415–433, doi:10.1007/s00382-004-0508-8, 2005.

Edwards, N. R., Cameron, D., and Rougier, J.: Precalibrating an intermediate complexity climate model, Clim. Dynam., 37, 1469–1482, doi:10.1007/s00382-010-0921-0, 2011.

Foley, A. M., Holden, P. B., Edwards, N. R., Mercure, J.-F., Salas, P., Pollitt, H., and Chewpreecha, U.: Climate model emulation in an integrated assessment framework: a case study for mitigation policies in the electricity sector, Earth Syst. Dynam., 7, 119–132, doi:10.5194/esd-7-119-2016, 2016.

Forrester, A., Sobester, A., and Kean, A.: Engineering design via surrogate modelling, vol. 1, John Wiley & Sons, Ltd, Chichester, 2008.

Forrester, A. I., Sóbester, A., and Keane, A. J.: Multi-fidelity optimization via surrogate modelling, P. R. Soc. A, 463, 3251–3269, doi:10.1098/rspa.2007.1900, 2007.

Fraedrich, K., Jansen, H., Kirk, E., Luksch, U., and Lunkeit, F.: The Planet Simulator: Towards a user friendly model, Meteorologische Zeitschrift, 14, 299–304, doi:10.1127/0941-2948/2005/0043, 2005.

Goldstein, M. and Rougier, J.: Reified Bayesian modelling and inference for physical systems, J. Stat. Plan. Infer., 139, 1221–1239, doi:10.1016/j.jspi.2008.07.019, 2009.

Haberkorn, K., Sielmann, F., Lunkeit, F., Kirk, E., Schneidereit, A., and Fraedrich, K.: Planet Simulator Climate, Tech. rep., Meteorologisches Institut, Universität Hamburg, 2009.

Hankin, R. K. S.: Analysis of Computer Code Output, J. Stat. Softw., 14, 1–21, doi:10.18637/jss.v014.i16, 2005.

Hevesi, J. A., Flint, A. L., and Istok, J. D.: Precipitation estimation in mountainous terrain using multivariate geostatistics. Part II: Isohyetal maps, J. Appl. Meteorol., 31, 677–688, 1992.

Holden, P. B. and Edwards, N. R.: Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling, Geophys. Res. Lett., 37, L21707, doi:10.1029/2010GL045137, 2010.

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., and Wilkinson, R. D.: A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1, Clim. Dynam., 35, 785–806, doi:10.1007/s00382-009-0630-8, 2010.

Holden, P. B., Edwards, N. R., Müller, S. A., Oliver, K. I. C., Death, R. M., and Ridgwell, A.: Controls on the spatial distribution of oceanic $\delta^{13}C_{DIC}$, Biogeosciences, 10, 1815–1833, doi:10.5194/bg-10-1815-2013, 2013.

Holden, P. B., Edwards, N. R., Garthwaite, P. H., Fraedrich, K., Lunkeit, F., Kirk, E., Labriet, M., Kanudia, A., and Babonneau, F.: PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment, Geosci. Model Dev., 7, 433–451, doi:10.5194/gmd-7-433-2014, 2014.

Jones, C., Gregory, J., Thorpe, R., Cox, P., Murphy, J., Sexton, D., and Valdes, P.: Systematic optimisation and climate simulation of FAMOUS, a fast version of HadCM3, Clim. Dynam., 25, 189–204, doi:10.1007/s00382-005-0027-2, 2005.

Kanamitsu, M., Ebisuzaki, W., Woollen, J., Yang, S.-K., Hnilo, J. J., Fiorino, M., and Potter, G. L.: NCEP–DOE AMIP-II Reanalysis (R-2), B. Am. Meteorol. Soc., 83, 1631–1643, doi:10.1175/BAMS-83-11-1631, 2002.

Kennedy, M. C. and O'Hagan, A.: Predicting the Output from a Complex Computer Code When Fast Approximations Are Available, Biometrika, 87, 1–13, 2000.

Kennedy, M. C. and O'Hagan, A.: Bayesian calibration of computer models, J. Roy. Stat. Soc. B, 63, 425–464, 2001.

Labriet, M., Joshi, S. R., Kanadia, A., Edwards, N. R., and Holden, P. B.: Worldwide impacts of climate change on energy for heating and cooling, Mitigation and Adaption Strategies for Global Change, 20, 1111–1136, doi:10.1007/s11027-013-9522-7, 2015.

Lee, L. A., Carslaw, K. S., Pringle, K. J., and Mann, G. W.: Mapping the uncertainty in global CCN using emulation, Atmos. Chem. Phys., 12, 9739–9751, doi:10.5194/acp-12-9739-2012, 2012.

Lenton, T. M., Williamson, M. S., Edwards, N. R., Marsh, R., Price, a. R., Ridgwell, a. J., Shepherd, J. G., and Cox, S. J.: Millennial timescale carbon cycle and climate change in an efficient Earth system model, Clim. Dynam., 26, 687–711, doi:10.1007/s00382-006-0109-9, 2006.

Locarnini, R. A., Mishonov, A. V., Antonov, J. I., Boyer, T. P., and Garcia, H. E.: Volume 1 : Temperature, in: World ocean atlas 2005, edited by Levitus, S., vol. 1, chap. Volume 1 :, NOAA Atlas NESDIS 61, US Gov. Printing Office, Washington, DC, 2006.

Mardia, K. V. and Marshall, R. J.: Maximum likelihood estimation of models for residual covariance in spatial regression, Biometrika, 71, 135–146, 1984.

Marsh, R., Müller, S. A., Yool, A., and Edwards, N. R.: Incorporation of the C-GOLDSTEIN efficient climate model into the GENIE framework: "eb_go_gs" configurations of GENIE, Geosci. Model Dev., 4, 957–992, doi:10.5194/gmd-4-957-2011, 2011.

Matthews, H. D. and Caldeira, K.: Transient climate-carbon simulations of planetary geoengineering., P. Natl. Acad. Sci. USA, 104, 9949–9954, doi:10.1073/pnas.0700419104, 2007.

Meinshausen, M., Smith, S. J., Calvin, K., Daniel, J. S., Kainuma, M. L. T., Lamarque, J.-F., Matsumoto, K., Montzka, S. A., Raper, S. C. B., Riahi, K., Thomson, A., Velders, G. J. M., and van Vuuren, D. P.: The RCP greenhouse gas concentrations and their extensions from 1765 to 2300, Climatic Change, 109, 213–241, doi:10.1007/s10584-011-0156-z, 2011.

Mercure, J., Pollitt, H., Chewpreecha, U., Salas, P., Foley, A. M., Holden, P. B., and Edwards, N. R.: The dynamics of technology diffusion and the impacts of climate policy instruments in the decarbonisation of the global electricity sector, Energ. Policy, 73, 686–700, doi:10.1016/j.enpol.2014.06.029, 2014.

Monier, E., Scott, J. R., Sokolov, A. P., Forest, C. E., and Schlosser, C. A.: An integrated assessment modeling framework for uncertainty studies in global and regional climate change: the MIT IGSM-CAM (version 1.0), Geosci. Model Dev., 6, 2063–2085, doi:10.5194/gmd-6-2063-2013, 2013.

Morris, M. D. and Mitchell, T. J.: Exploratory designs for computational experiments, J. Stat. Plan. Infer., 43, 381–402, 1995.

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

Peltier, W.: Global glacial isostasy and the surface of the ice-age earth: The ICE-5G (VM2) model and GRACE, Annu. Rev. Earth Pl. Sc., 32, 111–149, doi:10.1146/annurev.earth.32.082503.144359, 2004.

Plattner, G. K., Knutti, R., Joos, F., Stocker, T. F., von Bloh, W., Brovkin, V., Cameron, D., Driesschaert, E., Dutkiewicz, S., Eby, M., Edwards, N. R., Fichefet, T., Hargreaves, J. C., Jones, C. D., Loutre, M. F., Matthews, H. D., Mouchet, A., Muller, S. a., Nawrath, S., Price, A., Sokolov, A., Strassmann, K. M., and Weaver, a. J.: Long-Term Climate Commitments Projected with Climate–Carbon Cycle Models, J. Climate, 21, 2721–2751, doi:10.1175/2007JCLI1905.1, 2008.

Pope, V. D., Gallani, M. L., Rowntree, P. R., and Stratton, R. a.: The impact of new physical parametrizations in the Hadley Centre climate model: HadAM3, Clim. Dynam., 16, 123–146, doi:10.1007/s003820050009, 2000.

Rasmussen, C. E. and Williams, C. K. I.: Gaussian processes for machine learning., vol. 14, The MIT Press, Cambridge, doi:10.1142/S0129065704001899, 2006.

Riahi, K., Rao, S., Krey, V., Cho, C., Chirkov, V., Fischer, G., Kindermann, G., Nakicenovic, N., and Rafaj, P.: RCP 8.5-A scenario of comparatively high greenhouse gas emissions, Climatic Change, 109, 33–57, doi:10.1007/s10584-011-0149-y, 2011.

Rougier, J.: Probabilistic Inference for Future Climate Using an Ensemble of Climate Model Evaluations, Climatic Change, 81, 247–264, doi:10.1007/s10584-006-9156-9, 2007.

Rougier, J.: Efficient Emulators for Multivariate Deterministic Functions, J. Comput. Graph. Stat., 17, 827–843, doi:10.1198/106186008X384032, 2008.

Rougier, J., Sexton, D. M. H., Murphy, J. M., and Stainforth, D.: Analyzing the Climate Sensitivity of the HadSM3 Climate Model Using Ensembles from Different but Related Experiments, J. Climate, 22, 3540–3557, doi:10.1175/2008JCLI2533.1, 2009.

Schmittner, A., Silva, T. A. M., Fraedrich, K., Kirk, E., and Lunkeit, F.: Effects of Mountains and Ice Sheets on Global Ocean Circulation, J. Climate, 24, 2814–2829, 2011.

Stouffer, R. J., Yin, J., and Gregory, J.: Investigating the Causes of the Response of the Thermohaline Circulation to Past and Future Climate Changes, J. Climate, 19, 1365, 2006.

Thompson, S. L. and Warren, S. G.: Parameterization of outgoing infrared radiation derived from detailed radiative calculations, J. Atmos. Sci., 39, 2667–2680, 1982.

Wilkinson, R. D.: Bayesian Calibration of Expensive Multivariate Computer Experiments, in: Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainity, edited by Biegler, L., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B., Marzouk, Y., Tenorio, L., Waanders, B. v. B., and Willcox, K., chap. 10, John Wiley & Sons, Ltd, Chichester, doi:10.1002/9780470685853.ch10, 2010.

Williamson, D., Goldstein, M., and Blaker, A.: Fast linked analyses for scenario-based hierarchies, J. Roy. Stat. Soc. C-App., 61, 665–691, 2012.

Williamson, M., Lenton, T., Shepherd, J., and Edwards, N.: An efficient numerical terrestrial scheme (ENTS) for Earth system modelling, Ecol. Model., 198, 362–374, doi:10.1016/j.ecolmodel.2006.05.027, 2006.

Zickfeld, K., Eby, M., Weaver, A. J., Alexander, K., Crespin, E., Edwards, N. R., Eliseev, A. V., Feulner, G., Fichefet, T., Forest, C. E., Friedlingstein, P., Goosse, H., Holden, P. B., Joos, F., Kawamiya, M., Kicklighter, D., Kienert, H., Matsumoto, K., Mokhov, I. I., Monier, E., Olsen, S. M., Pedersen, J. O. P., Perrette, M., Philippon-Berthier, G., Ridgwell, A., Schlosser, A., Von Deimling, T. S., Shaffer, G., Sokolov, A., Spahni, R., Steinacher, M., Tachiiri, K., Tokos, K. S., Yoshimori, M., Zeng, N., and Zhao, F.: Long-Term climate change commitment and reversibility: An EMIC intercomparison, J. Climate, 26, 5782–5809, 2013.