Noname manuscript No.
(will be inserted by the editor)

Multi-level emulation of complex climate model responses to boundary forcing data

Giang T. Tran · Kevin I. C. Oliver · Philip B. Holden · Neil R. Edwards · András Sóbester · Peter Challenor

Received: date / Accepted: date

Abstract Climate model components involve both high-dimensional input and output fields. It is desirable to efficiently generate spatio-temporal outputs of these models for applications in integrated assessment modelling or to assess the statistical relationship between such sets of inputs and outputs, for example, uncertainty analysis. However, the need for efficiency often compromises the fidelity of output through the use of low complexity models. Here, we develop a technique which combines statistical emulation with a dimensionality reduction technique to emulate a wide range of outputs from an atmospheric general circulation model, PLASIM, as functions of the boundary forcing prescribed by the ocean component of a lower complexity climate model, GENIE-1. Although accurate and detailed spatial information on atmospheric variables such as precipitation and wind speed is well beyond the capability of GENIE-1's energy-moisture balance model of the atmosphere, this study demonstrates that the output of this model is useful in predicting PLASIM's spatio-temporal fields through multi-level emulation. Meaningful information from the fast model, GENIE-1 was extracted by utilising the correlation between variables of the same type in the two models and between

G. T. Tran · K. I. C. Oliver
Ocean and Earth Sciences, University of Southampton, Southampton, SO14 3ZH, UK
E-mail: gtran@geomar.de
Present address of G. T. Tran: GEOMAR Helmholtz Centre for Ocean Research Kiel, Düsternbrooker Weg 20, 24105 Kiel, Germany

P. B. Holden · N. Edwards
Environment, Earth and Ecosystems, The Open University, Milton Keynes, MK7 6AA, UK

A. Sóbester
Engineering and the Environment, University of Southampton, Southampton, SO16 7QF, UK

P. Challenor
College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, EX4 4QE, UK

variables of different types in PLASIM. We present here the construction and validation of several PLASIM variable emulators and discuss their potential use in developing a hybrid model with statistical components.

Keywords Probabilistic prediction · multi-level emulators · model hierarchy · spatio-temporal data · intermediate complexity model

1 Introduction

Climate models describe a set of rules which convert a given forcing into a response. Complex feedbacks in the climate system modify this basic forcing-response relationship. As a result, the model outcomes are uncertain even though the model is deterministic. Due to their computational expense, we often cannot evaluate the model enough times to provide a good statistical analysis of the model's behaviour. In such a situation, statistical emulators, also referred to as surrogate models, are often used to provide estimations of the outputs produced by a model given a specific set of input parameters [48, 50]. A Gaussian process (GP) emulator combines our prior judgments about the model's behaviour with data from simulations to predict some desirable outputs of the climate model. Once constructed and validated, emulator predictions can be obtained at low cost without the need for further evaluation of the climate model. It is worth noting that the emulator is only valid within the input space it was designed for and cannot be used to replace the physical model beyond this space.

Within the designed input space, the emulator's ability to generate a large amount of predictions makes it a valuable tool in many applications that have previously been limited by computational speed, such as: to study the model's behaviour in a large parameter space [23], to perform sensitivity analysis of a variable to certain inputs [42, 6], uncertainty quantification [30], calibration [28, 4] and precalibration/history matching [12, 59]. In situations where the computation cost remains a limitation, multi-level emulation techniques [27, 14] can be used to further reduce the number of simulations needed [56]. This is applicable when a lower complexity model of the same physical system exists.

The majority of the studies mentioned above have focused on problems involving scalar inputs and outputs. However, it is often the case that we are interested in the spatial distribution of the climatic outputs or how they evolve over time. Dimension reduction techniques (most commonly, principal component analysis; PCA) have been introduced to enable emulation of 2-D outputs [40, 21, 58], spatio-temporal outputs [24], time series [7] and multivariate outputs [5].

While much effort has been dedicated to reducing the dimensionality of model outputs, relatively few have attempted to do the same on model inputs such as 2-D boundary/initial conditions, which are often an unquantified source of uncertainty. One example is [25], where the linear decomposition technique, PCA, was employed to reduce the dimensionality of the 2-D input fields (sea surface temperature and precipitation) and the output field (vegetation) of a vegetation model. The relationship between the component scores obtained from decomposing the input and output fields are emulated using a regression model. A statistical model that relates high-dimensional input to high-dimensional output can be useful in producing a fast prediction of the output or to clarify the relationship between the input-output pairs. Here we list some examples in which such emulators could be beneficial:

1. To assess the uncertainty introduced by high-dimensional boundary forcing or initial conditions through a smaller set of latent variables obtained by dimensional reduction. Similarly, to obtain the sensitivity of 2-D surface outputs to the same boundary/initial conditions. This is the case when high-dimensional climate variables from a separate model or observations are used as the boundary forcing condition. Examples include aerosols and human land-use change for climate impact projections or ice-sheet height and extent for simulations of past climate.
2. To improve our understanding of the relationship between high-dimensional climate variables. For example, instead of predicting and assessing model outputs as functions of numerical model parameters which sometimes do not correspond to real physical processes, we can relate a variable of interest (e.g. vegetation) directly to other observable quantities (e.g. temperature or precipitation). A specific change in the vegetation field can now be understood and quantified in terms of different modes of variability in temperature or precipitation.
3. To enable coupling between models or model components which involve the exchange of high-dimensional flux fields. For example, in studies with a focus on the ocean dynamics over long timescales, an ocean general circulation model (OGCM) is often coupled with a simple atmosphere since a fully coupled GCM would require significantly more computing power and time to integrate. By emulating the atmospheric fluxes into the ocean as a function of the input exchange fields from the ocean to the atmosphere, an efficient coupling between the OGCM and emulators of a complex atmospheric model can be obtained. This approach can potentially offer a more realistic representation of the forcing fields than otherwise obtained, without being computationally intensive.

In this work, we bring together multi-level emulation and dimension reduction of the spatio-temporal boundary forcing described by the ocean component of a climate model (GENIE-1) to provide estimations of several 2-D output fields from an AGCM (PLASIM). This builds on the work described in [56], in which the dimensionally reduced surface air temperature from these two models of different complexity was emulated as a function of GENIE-1's scalar input parameters using a multi-level Gaussian process emulation technique called co-kriging. By emulating the output of the lower complexity atmospheric component of GENIE-1 and the relationship between outputs of the two models, an emulator of PLASIM's annual mean surface air temperature (SAT) was obtained using a relatively small number of PLASIM simulations. PCA was employed to reduce the dimensionality of the surface field. Furthermore, combined PCA is introduced to emulate various PLASIM fields in addition to SAT.

The PLASIM variables we emulate are the surface field of specific humidity, precipitation, zonal and meridional wind speed. These variables are chosen because they are more challenging to emulate compared to SAT since these fields, with the exception of humidity, are not as strongly constrained by the provided sea surface temperature (SST). Moreover, the multi-level method introduced in the previous paper [56] is not immediately applicable in this case since the fields we are interested in might not have a direct fast approximation in GENIE-1. Here we consider three different situations: i) Humidity is simulated in GENIE-1 and is likely to be suitable as the fast approximation of PLASIM's humidity. ii) While present in both models, precipitation is poorly represented in GENIE-1 and it's unlikely that GENIE-1's precipitation contains very useful information on PLASIM's field. iii) Unlike these two, surface winds are not simulated in GENIE-1. Therefore, there is no direct fast approximation of PLASIM winds in GENIE-1. In this work, we demonstrate the modifications to our previously established multi-level emulation technique so that it can be applied to all the three cases stated above.

Table 1 provides a summary of all the steps taken in this work. The second column contains a description of each procedure while the third column lists the section in which the procedure is discussed. The bold steps refer to the essential tasks of building our final emulators while the remaining are optional. Step 5 and 6 are conducted to provide comparisons between the single and multi-level emulation techniques. Step 9 and 10 involve the use of an independent ensemble for emulator validation. It is possible to validate using a leave-one-out cross-validation instead hence these steps are only optional.

2 Model descriptions and experiment design

2.1 Models

In this work, we employ two climate models of different complexity. The main focus is an AGCM which is the atmospheric component of the Planet Simulator (PlaSim), developed at the University of Hamburg [16, 15]. PlaSim consists of a fully dynamical 3-D atmosphere based on the Portable University Model of Atmosphere (PUMA), coupled with a 2-D mixed layer ocean. PlaSim and PUMA have previously been employed to study the effect of mountains on the ocean circulation [51], the role of oceanic heat transport and orography on glacial climate [47], the interactions between stationary waves and continental ice sheets [32] and the global energy and entropy budget in a snowball Earth hysteresis [34].

The atmosphere of PlaSim is a coarse resolution AGCM which is based on the moist primitive equations representing conservation of momentum, mass, energy, and moisture on a terrain-following σ-coordinate system. The equa-

Table 1 A summary of the steps taken to construct and validate emulators of PLASIM's atmospheric variables and the sections describing them in this paper. Step 5 and 6 (marked with an asterisk in the section column) are performed to compare the single and multi-level emulator performance and are not explicitly described in this work. The steps in bold are necessary to the statistical technique presented here. Readers who wish to apply the same method should carry out the equivalent of these steps. While validation is also a necessary step, step 9 is not indicated as such since we use independent validation data in this work. It is possible to validate the emulators using a leave-one-out cross-validation instead.

	Procedure	Section
1	**Obtain 200 simulations with fast, fully coupled, Earth system model of intermediate complexity (GENIE-1)**	2.2 and 3.1
2	**Build EOFs of surface ocean conditions and surface air temperature in GENIE-1**	3.2
3	**Build statistical emulator of GENIE-1 atmosphere, relating inputs (CO2, ice sheet configuration, principal components of surface ocean conditions) to output (principal components of surface air-temperature)**	4
4	**Obtain 90 simulations with expensive atmospheric model (PLASIM), forced by CO2, ice-sheet configuration, and surface ocean conditions from a subset of GENIE-1 ensemble**	2.2 and 3.1
5	Kriging Step 1: Build EOFs of atmospheric conditions (humidity, precipitation, winds) in PLASIM	*
6	Kriging Step 2: Build statistical emulator of PLASIM atmosphere, relating inputs (CO2, ice sheet configuration, principal components of surface ocean conditions) to output (principal components of humidity, precipitation, winds)	*
7	**Co-kriging Step 1: Build combined EOFs of atmospheric conditions (SAT and humidity, SAT and precipitation, SAT and zonal winds, SAT and meridional winds)**	6
8	**Co-kriging Step 2: Build statistical emulator of PLASIM atmosphere, making use of output from, and emulator of, GENIE-1 atmosphere**	6
9	Obtain 214 simulations with PLASIM, used to generate independent validation data	2.2
10	**Use of validation data to assess relative performance of kriging and co-kriging**	7

tions are written in spherical coordinates and solved using the spectral transform method. It contains parameterizations of unresolved processes consisting of short and long wave radiation, with the inclusion of the greenhouse gas effect of water vapour, carbon dioxide, and ozone. Other parameterized processes include moist processes with an interactive cloud, boundary layer heat fluxes and diffusion. We run the model at a T21 grid of 64×32 cells, which corresponds approximately to a $5.6^\circ \times 5.6^\circ$ grid, and 10 vertical levels. The interaction with other climate subcomponents is enabled by adding reduced models for ocean, sea-ice and land-surface processes. Here, the simple 2-D mixed layer ocean and sea-ice subcomponents are replaced by the prescribed 2-D boundary forcing fields obtained from a separate model, GENIE-1 (described below). Also, we use a PlaSim implementation in which the land-surface model, ENTS from GENIE is employed instead of the subcomponent provided by PlaSim. This version is referred to as PLASIM-ENTS (efficient numerical terrestrial scheme) and was described in [24]. PlaSim has been employed in the past with prescribed climatological SST [18] or coupled to a different ocean [51]. Emulators of PLASIM-ENTS have been used in an integrated assessment model [13, 29]. We refer to the specific configuration used in this work as PLASIM, where the atmosphere and land component PLASIM-ENTS is used without the remaining subcomponents.

Also used is the Grid-ENabled Integrated Earth system modelling (GENIE) framework [31], an Earth system model of intermediate complexity (EMIC) designed to perform long integrations to investigate glacial climate. The configuration used in this work is commonly known as GENIE-1, which consists of a 2-D energy-moisture balance model (EMBM) of the atmosphere [11], a 3-D frictional geostrophic ocean (GOLDSTEIN; [11]), dynamic-thermodynamic sea ice, and a land surface physics and terrestrial carbon cycle model (ENTS; [60]). The model is on a near-equal increment mesh (equal increments in longitude of $5.625°$ and similar near-equal increments of latitude) of 64×32 grid points. A full description of the model can be found in [38, 11] and references therein. The EMBM of the atmosphere is based on the atmosphere of the UVic Earth system model [57], which includes the representations of the surface exchange of heat and moisture with the ocean, sea ice, and land, horizontal transport of heat and moisture in the atmosphere by diffusion and advection, and precipitation above a relative humidity threshold. The GOLDSTEIN ocean incorporates the surface input of momentum from the surface wind stress and transport of heat and salinity through the combined parameterization for eddy-induced advection and isopycnal mixing. The ocean component has 16 exactly logarithmically-spaced levels with a maximum depth of 5000 m. The sea-ice model, based on the sea ice implemented in the UVic model, includes horizontal transport of sea-ice concentration and thickness, surface exchange of heat and freshwater with the atmosphere and the ocean. ENTS models vegetative and soil carbon densities, assuming a single plant functional type that has a double-peaked temperature response (representing boreal and tropical forests).

GENIE-1 has been designed for model integration on millennial timescales to investigate past climate changes [35], long-term response of the carbon cycle [31] and for large ensemble studies [22, 12]. Compared to PLASIM, the 2-D atmosphere of GENIE-1 is much simpler and does not resolve synoptic activity and hence does not excite multidecadal oscillations in the ocean circulation. As a result, outputs from quasi-steady-state simulations conducted with GENIE-1 have low variability.

The higher complexity model PLASIM simulates better climatology than the EMBM of GENIE-1. However, it lacks dynamic representations of the ocean and sea-ice. A fully coupled PLASIM-GENIE was recently developed by [26] but in this work, PLASIM is driven by prognostic variables supplied by GENIE-1's ocean and sea ice to i) demonstrate the statistical technique and ii) save computational cost as the atmosphere is two orders of magnitude more costly to run than the ocean. The variables used to drive PLASIM atmosphere are the monthly 2D sea surface temperature (SST), sea-ice fractional area (SIC) and sea-ice thickness (SIH). The lack of multidecadal variations in the ocean circulation also influences the variability of PLASIM's atmospheric response. This is not a major concern here since we are focusing only on the mean steady state. In this work, our target is to emulate PLASIM's climate, and thus, the output of this model is treated as reality. The climate from GENIE-1's EMBM, on the other hand, is considered as a 'fast approximation' of that of PLASIM. This implies that due to the lower complexity of the processes represented in EMBM, the resulting climate, while can be obtained faster and required much less computational power, is of lower fidelity compared to that of PLASIM. The following section provides a short description of our experimental design which was provided with more details in [56].

2.2 Experiment design

To build a single level emulator, also known as a kriging emulator, an ensemble of simulations is needed to condition the GP describing the prior judgements. This ensemble is referred to as the training ensemble. Once constructed, the emulator is validated to ensure that it is capable of producing reliable approximations of the climate model output. To do this, a separate validation ensemble is used. For the multi-level emulation technique using both GENIE-1 and PLASIM data, two training and two validating ensembles are produced. First two, a training and a validation, maximin Latin hypercube sampling plans [41] are generated with 12 GENIE-1 model parameters perturbed. Quasi-steady-state simulations of 5000-year length are first performed in GENIE-1. The resulting monthly SST, SIC and SIH are then supplied to PLASIM, driving the atmosphere for another 35 years. Thus, each Latin hypercube sampling plan generates a GENIE-1 and a corresponding PLASIM ensemble.

The 12 model parameters are the inputs of GENIE-1 (Table 2). The outputs we are interested in are the surface air temperature, sea surface temperature and two sea ice fields. The inputs of PLASIM are the surface ocean and sea ice boundary forcing fields prescribed by GENIE-1, together with ice sheet configuration and CO_2 concentration. The PLASIM's variables to be emulated are surface air temperature, surface specific humidity, precipitation rate and the two surface wind components.

In this work, we utilised previously generated ensembles: the training ensembles of 600 members and the validation ensembles of 214 members. The perturbed GENIE-1's parameters and their ranges (Table 2) are based on the previous designs used in [22] with adjustments made due to the use of a different model resolution. A sensitivity test showed that 600 simulations are more than adequate to produce a good emulator. As a result, only small subsets of this initial ensemble are used to construct the emulator. The final number of simulations used are described in Section 3.1. The sensitivity test and the subsampling method used are briefly described in [56]. More detailed information can be found in [55].

Table 2 Ten of the 12 chosen parameters, with the exception of ICF and RFC, are taken from an ensemble design used in [22]. 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 plans.

	Code	Parameter	Min	Max	Dist.
1	ICF	Ice sheet and orography configuration (no unit)	0	21	Lin
2	OHD	Ocean isopycnal diffusivity (m^2 s^{-1})	300	4000	Log
3	OVD	Ocean diapycnal diffusivity (m^2 s^{-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^2 s^{-1})	4×10^6	7.0×10^6	Log
7	AMD	Atmospheric moisture diffusivity (m^2 s^{-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 skies OLR reduction (W m^{-2})	0	10	Lin
11	OL1	OLR feedback (W m^{-2} K^{-1})	-0.5	0.5	Lin
12	RFC	CO_2 forcing (ppm)	150	1400	Lin

The perturbed parameters include a numerical description of the changes in glacial mask and orography over the last 21000 years (ICF) and the atmospheric CO_2 concentration (RFC). ICF represents the boundary condition of the glacier coverage as well as the corresponding orography and albedo at different snapshots 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 [44]. The most prominent features of the land ice configuration are the extent and height of the Laurentide ice sheet over North America and the Fennoscandian over northern Europe. It is worth noting that the glacial configuration and CO_2 concentration are varied independently. Thus, it is possible for a simulation to have, for example, Last Glacial Maximum ice configuration and preindustrial CO_2 concentration at the same time. Other parameters control processes in the atmosphere and the ocean. Only ICF and RFC are also varied in PLASIM. Other PLASIM parameters are fixed at default values [18].

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). APM is a flux correction responsible for transporting fresh water from the Atlantic to the Pacific, affecting deep-water sinking in the North Atlantic and hence the strength of the AMOC. The uncertain impact of atmospheric transport is captured through atmospheric heat and moisture diffusivity parameters (AHD and AMD, respectively) [11]. 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 [54, 39]. RMX is the threshold value of relative humidity for precipitation, capturing the uncertainty in water vapour feedbacks [31].

The atmospheric outputs of each PLASIM simulation are averaged over the last 30 simulation years while the outputs of GENIE-1 are taken from the final model year of the steady-state simulations. As discussed before, due to the simplicity of GENIE-1, there is negligible interannual variability within the model after a 5000-year integration.

3 Statistical method

3.1 Emulator description

The Gaussian process emulation technique is employed in this work to emulate atmospheric variables of PLASIM. Our statistical emulator describes the data as a global linear model with the residuals being modelled by a GP. The standard kriging emulation technique employed here assumes a constant mean function. However, to diagnose and interpret the contribution of each emulator input to the global trend, a single level GP emulation technique with linear mean functions (universal kriging) is also employed. Universal kriging models the simulator response as a sum of a mean response function and a zero-mean GP. The mean response function is a linear combination of regressions whose coefficients can be useful in interpreting the model's behaviour. In our ensemble, the dominating influence from CO_2 concentration on global annual mean SAT is approximately linear over the input space so this linear form is a suitable choice. In cases where no dominant linear trend is found, the universal kriging emulator becomes a standard kriging emulator with a constant mean function. In our previous experience with emulators of PLASIM and GENIE-1, the standard technique slightly outperforms universal kriging, thus it is used to construct the final emulators here. This is perhaps an indication that the linear structure assumed might not hold for the all input-output relationships. Nevertheless, the regression coefficients are useful in interpreting model outputs as shown in Section 4.2

Co-kriging is the multi-level extension to the standard single-level kriging technique, which is applicable when a faster simulator of the same physical system is available. When only a small number of expensive simulations are available, it has been shown that by combining these with a large number of cheaper runs from a simplified code, an emulator of the expensive model can be built at a lower cost [14]. Potentially, this method can be extended to more code levels [27]. The expensive model's output, f_e, is modelled as a GP of its fast approximation, f_c, multiplied by a scaling factor, ρ, plus a separate GP, f_d, describing the stochastic residual of the expensive model [27, 14]. This approach is referred to as the single multiplier approach:

$$f_e(\boldsymbol{x}) = \rho f_c(\boldsymbol{x}) + f_d(\boldsymbol{x}), \tag{1}$$

We use n_e simulations from PLASIM and a larger number of simulations, n_c, from GENIE-1 to train our co-kriging emulators. To identify the relationship between the two models, n_e is chosen to be a subset of n_c, meaning that we have information on both PLASIM's and GENIE-1's behaviours at the same n_e input combinations. In a previous study which used the same ensembles, we determined that $n_c = 200$ and $n_e = 90$ are sufficient to construct a good SAT emulator. Thus, the same number of simulations are used here. All the simulations used are subsampled from the initial 600-member training ensemble. The details on how these numbers were determined can be found in [56] and [55].

A mathematical description of the emulators used can be found in the Appendix. The co-kriging emulators in this work were constructed using the toolbox provided by [14].

3.2 Dimensional reduction using principal component analysis

Many studies have been done on extending from univariate GP emulation to handle multivariate outputs, most commonly through a dimension-reducing technique. There are a large number of linear and non-linear dimensional reduction techniques. These techniques reduce the dimensionality of a data set by embedding this data into a subspace of lower dimensionality. Among these methods, PCA is the most commonly used as this method is quick and has the advantage that the first few components explain the majority of the variance across the ensemble [20, 21, 58, 7]. [7] used PCA to reduce the time dimension of AMOC time series while [25] employed the same technique to reduce both the spatial and temporal dimension of their SAT field, allowing them to emulate the response of the climate system at different time slices. Other methods for reducing the dimensionality of the simulator's inputs prior to emulation are discussed in [36].

Each surface output field is reshaped into an $m \times 1$ vector, where m is the total number of grid points. In the case of a single output from PLASIM, $m = 64 \times 32 = 2048$. By concatenating vectors of all training points together, we form an $m \times n$ matrix $\mathbf{U}$, where n is the number of training points.

A singular value decomposition (SVD) can be applied directly on any $m \times n$ matrix $\mathbf{U}$, giving

$$\mathbf{U} = \mathbf{L}\mathbf{S}\mathbf{R}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{L}$ and $\mathbf{R}$ are the matrices of left and right singular vectors, respectively; $\mathbf{S}$ is the diagonal matrix of singular values. The matrix $\mathbf{L}$ is the matrix of eigenfunctions, referred to as the empirical orthogonal functions (EOFs). The right singular vectors are sometimes referred to as the component scores. The principal components (PCs) can be obtained from these scores as follows:

$$\mathbf{Z} = \mathbf{S}\mathbf{R}^{\mathrm{T}}. \tag{3}$$

The terminology of PCA is not consistent in literature, especially across different research fields. In this work, we adopt the convention of referring to the eigenvectors, which in this case are the spatial patterns, as the EOFs and the coefficients that scale these patterns as the principal components. Any simulated field can be constructed as a linear combination of the EOFs, weighted by their respective series of PCs. Each $m \times 1$ column of $\mathbf{L}$ is an EOF, describing a map or a mode of variation in the ensemble. A simulated field is thus completely described by the set of coefficients of the EOFs, for instance, the training points in the matrix $\mathbf{U}$ are defined by the corresponding $n \times 1$ columns of $\mathbf{Z}$.

We can use PCA via SVD for dimensional reduction of a 2-D data set. The i^{th} grid cell of the j^{th} field from $\mathbf{U}$ can be written as

$$\hat{U}_{ij} = \sum_{k=1}^{q} L_{ik} S_{kk} R_{kj}^{T}, \tag{4}$$

where q is the number of modes retained. When $q = m$, $\hat{U}_{ij} \equiv U_{ij}$, otherwise, $\hat{U}_{ij}$ is an approximation of U_{ij}.

The top few (or low order) EOFs often explain most of the variance in the data such that the dimension of $\mathbf{U}$ can be reduced by keeping only the first q components ($q \ll m$). In this work, the PCs are treated as latent variables, replacing the high-dimensional 2-D field as inputs and outputs of the emulators. The n elements (or indices) of each component score correspond to the n simulations used as training data. Emulators are built for the first q PCs, providing an estimation of $\hat{\mathbf{Z}}$, for any untried input combination. They are then used to reconstruct the final prediction, $\hat{\mathbf{U}}$ of the emulated field.

The prediction, $\hat{\mathbf{U}}$, is different from the simulated value of $\mathbf{U}$ 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 percentage explained would be sufficient, a high value such as 90% for SAT should be satisfactory. This value might be lower for other variables with larger internal variability. The component error is a result of imperfect estimations by the emulator which can be reduced by using more training points. The GP emulator provides an estimate of this error.

4 Dimensional reduction of PLASIM inputs

4.1 Decomposing the surface forcing fields

All the high-dimensional inputs (SST, SIC and SIH) and outputs (SAT, precipitation, surface wind speeds and humidity) concerned in this work are 2-D fields. As seen in the previous section, each surface field contains 64×32 grid cells and can be reshaped into a 2048×1 state vector. Since we are interested in the capturing the seasonal variation, for each variable, there are now 12 vectors corresponding to 12 calendar months for each simulation run.

Following [25], for each atmospheric variable of each ensemble member, a new state vector is constructed by concatenating together all 12 vectors vertically, giving a spatio-temporal field $\mathbf{U} = [\mathbf{U}_1, \cdots, \mathbf{U}_{12}]$, which has the dimension of 24576×1. A training set of n members can be represented as a $24576 \times n$ matrix. PCA is then performed, via SVD, on this new matrix in the same way as before.

The initial conditions driving each PLASIM simulation are provided by GOLDSTEIN's three prognostic variables, SST, SIC and SIH. Therefore, the total dimension of our new input fields is $64 \times 32 \times 12 \times 3 = 73728$. Clearly, without reducing the dimensionality of the input fields, emulation would not be practical. Fortunately, we can utilise the correlation structures in space and/or time within a single variable as well as between variables of different types (e.g. between SST and SIC) to reduce their dimension via PCA to a smaller and more manageable set of inputs. The importance of the inputs, currently spread among the 73728 dimensions, are redistributed within a new set of latent variables in which the first few variables explain the majority of the variation across the ensemble, allowing redundant information to be identified and removed. The PLASIM outputs to be emulated are denoted as as $\mathbf{U}_{\mathrm{p}}$ and the input for emulator are $\mathbf{W}_{\mathrm{g}}^k$, with $k = 1, 2, 3$ corresponding to GENIE-1's SST, SIC and SIH respectively. The subscript g and p denotes GENIE-1 and PLASIM, respectively. Since $\mathbf{W}_{\mathrm{g}}^k$ only exist in oceanic grid cells, the land mask is removed from these fields, leaving the matrix $\mathbf{W}_{\mathrm{g}}^k$ with the dimension of $16332 \times n$. PCA is then applied to each field independently, giving:

$$\mathbf{W}_{\mathrm{g}}^k = \mathbf{L}_{\mathrm{g}}^k \mathbf{S}_{\mathrm{g}}^k {\mathbf{R}_{\mathrm{g}}^k}^{\mathrm{T}}. \tag{5}$$

The columns of $\mathbf{L}_{\mathrm{g}}^k$ are the EOFs of SST, SIC and SIH when $k = 1, 2, 3$ respectively. These are the spatial patterns of different statistical modes identified within these fields, ordered in decreasing importance. The new inputs which we will use for the emulators are the principal components, $\mathbf{Z}_{\mathrm{g}}^k$, with $k = 1, 2, 3$ for SST, SIC and SIH, respectively.

The removal of the land mask does not affect the prediction over land since only the PCs are included in the emulators. These PCs dictate how the EOFs are scaled. As long as the behaviour of the PCs of the atmospheric fields can be described as functions of SST PCs, the emulator will be able to emulate PLASIM's outputs over land.

To better capture the effect of ICF and RFC, which are also varied in PLASIM, these two parameters are also introduced as emulator inputs. Their indirect effects through the GOLDSTEIN ocean is captured through the PCs while their direct effects will be captured from these two parameters. This means that the emulator inputs are not independent of each other.

Since the emulator's inputs become the PCs obtained from the SVD instead of the model parameters, the PCs of a new forcing field need to be obtained before we can make a prediction of PLASIM's output to that field. These input PCs are computed by projecting the new boundary forcing field onto the EOFs obtained from the decomposition of the training ensemble. This is a simple matrix multiplication and can be done relatively efficiently. There is a possibility that at least one of the obtained PCs is outside of the individual training range. Thus, care should be taken that such instances do not go unnoticed.

Table 3 The variance explained by the first 10 modes of SST, SIC and SIH. The 600-member GENIE-1 ensemble was used to obtain these values.

	SST	SIC	SIH
1	85.0	59.9	67.1
2	8.9	12.0	22.5
3	3.4	4.2	4.6
4	0.8	3.4	1.8
5	0.5	2.2	0.9
6	0.3	1.8	0.8
7	0.2	1.3	0.5
8	0.1	1.1	0.4
9	0.1	1.0	0.2
10	0.1	0.8	0.2
Total	99.4	87.7	99.0

By retaining only low-order principal components, the dimension of the input is reduced significantly. The variances explained by the first ten EOFs for these three fields are displayed in Table 3. Evidently, the top EOFs capture the majority of the variance across the ensemble for all three variables. One might decide to keep the first ten EOFs of each field which results in 30 emulator inputs. While this is a significant reduction from three surface fields, each has a dimension of 16332 (after land points are removed), it is still a large number, and further reduction is desirable. A quick inspection shows that the first three modes of SST together explain over 97% of the ensemble variance in SST (Table 3) while each of the remaining modes explains less than 1%. Thus, the fourth to tenth PCs are likely to contain noise which can be truncated. Moreover, sea-ice cover and thickness are both well correlated with SST. The correlation between the first PC pair of SST-SIC and SST-SIH are found to be 0.77 and 0.67, respectively. Therefore, it is likely that they contain redundant information already provided by SST and can be discarded.

To verify this, we construct emulators of the EMBM's 2-D SAT field using various combinations of the dimensionally-reduced inputs. First, emulators using the top 3, 5 and 7 SST PCs as inputs, in addition to ICF and RFC, are built and validated. A comparison of the performance of the three emulators shows that the addition of the fourth and fifth PCs does improve the emulator performance significantly (total variance explained increased by 6.77%). The addition of even higher order PCs, in this case, the sixth and seventh components, do not contribute significantly nor positively to the final emulator. The variances explained by these emulators are shown in Figure 1. Both the variance explained and the RMSE between the emulated and simulated fields are shown in Table 4. While the fourth and fifth modes explain less than 1%

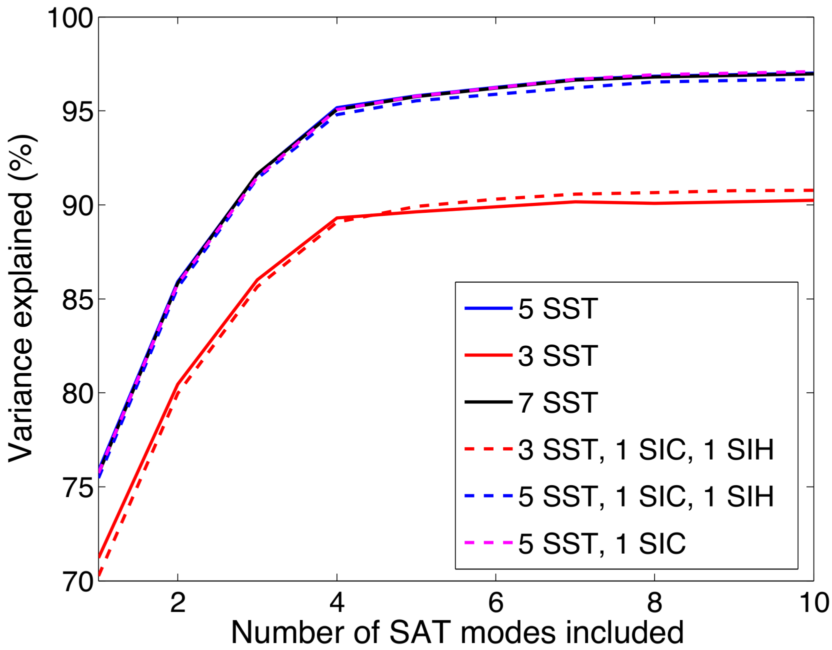

Fig. 1 The variance explained by the SAT emulators using various combinations of inputs. For each emulators, the number of PCs used from each SST, SIC and SIH fields is included in the legend. All the emulators also include ICF and RFC as inputs and use the same amount of training points.

of the total variance, it is possible that despite their small contributions to the global signal, they are important in explaining regional features or features that have a stronger influence to SAT over land.

Table 4 The variance explained by the first 10 EOFs of SST, SIC and SIH. The 600-member GENIE-1 ensemble was used to obtain these values.

Input combination	RMSE (°C)	Variance explained (%)
3 SST	1.59	90.2
5 SST	0.98	97.0
7 SST	0.98	97.0
3 SST, 1 SIC, 1 SIH	1.13	90.8
5 SST, 1 SIC, 1 SIH	0.96	96.7
5 SST, 1 SIC	0.96	97.1

In a similar process, the addition of SIC and SIH PCs is tested. Three different combinations of dimensionally-reduced inputs are used, i) The first three PCs of SST and the first PC of SIC and SIH, ii) The first five PCs of SST and the first PC of SIC and SIH, iii) The first five PCs of SST and the first PC of SIC. The addition of the sea-ice components has little effect on the performance of the emulator. In addition to these PCs, the two parameters ICF and RFC are included in all cases.

Ultimately, we decided to use the first five PCs from SST, which are the first 5 columns of $\mathbf{Z}_{\mathrm{g}}^{1}$, and the two PLASIM parameters as the inputs for all emulators built in this article. The dimensionality of the emulator input has been reduced from three surface fields, each with 16332 grid cells, with two extra parameters down to seven inputs.

4.2 Physical interpretation of the statistical modes

To check whether the statistical modes of SST obtained from PCA correspond to physical processes, we emulate the first three PCs of SST as a function of the original 12 model parameters to study the ensemble climate and its relation to these parameters. This step also aids the interpretation of the PLASIM climate in later sections. Universal kriging is used since it provides the coefficients of the linear part of the emulator, allowing us to compare the relative contribution of each parameter to the overall linear trend. Figure 2 shows the first 3 SST EOFs and their corresponding universal kriging emulator coefficients. These coefficients are the estimated gradients of the linear mean function fitted to the data. The importance of a parameter does not solely depend on the regression coefficients determined here. Each coefficient corresponds to one of the 12 model parameters.

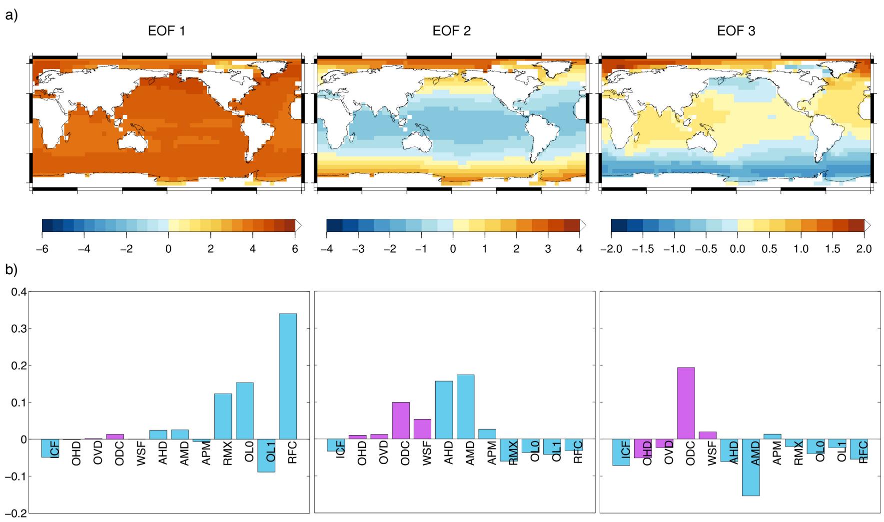

Fig. 2 This figure shows a) The spatial structure of the first 3 EOFs of SST and b) The corresponding emulator coefficients of the universal kriging emulators of these modes. All 600 cheap simulations were used to train these emulators.

The first mode, which explains 85.0% of the total variance across the ensemble, is a radiative forcing mode, dominated by the atmospheric CO_2 concentration, RFC. Other parameters with a significant contribution to this EOF are those which directly affect the energy balance via either albedo or atmospheric greenhouse gas concentration, namely OL0, RMX, OL1 and ICF. A larger atmospheric CO_2 concentration causes the global surface air temperature to rise, which in turn, warms the ocean. Large values of OL0 and RMX have a similar effect since they also increase the effective greenhouse gas concentration in the atmosphere by allowing more water vapour to remain there. The effect of ICF is the opposite sign to RFC since larger values of ICF corresponds to glacier condition closer to the Last Glacial Maximum. Increasing values of ICF are associated with larger continental ice cover and higher orography. The orography mask also matches the glacier mask. Thus, increasing ICF causes a cooling signature globally through the albedo effect and locally due to changes in elevation. The second effect is more apparent when looking at SAT (Figure 5 in [56]). The largest variations in this EOF are seen in regions at high latitudes where sea-ice coverage changes can amplify the surface temperature changes. Areas where sea-ice persists show smaller variations.

The second mode (8.9%) displays a variation in the equator to pole temperature gradient. A smaller gradient, seen in Figure 2 as a warming at high latitudes and cooling in the tropics, is induced mainly by increasing the atmospheric moisture and heat diffusivity, AMD and AHD. As heat and moisture are carried from low to high latitudes more efficiently, the meridional temperature difference is reduced. The inverse ocean drag coefficient, ODC, governs the parameterization of the friction in the ocean and hence, the dissipation of momentum. An increase in ODC reduces the friction in momentum balance. This parameter also has a significant impact on the strength of the meridional ocean circulation in the North Atlantic. WSF is a wind-stress scaling factor, controlling the strength of the wind-driven gyres. The strength of this mode is controlled by two competing groups of parameters. As the above-mentioned parameters with positive coefficients (Fig 1b) increase, the equator to pole temperature gradient decreases due to more efficient distributions of heat in the system. The second group of parameters with negative coefficients have the opposite effect since they affect the net radiation balance. Larger values of these parameters tend to cause polar amplification and hence a larger meridional gradient.

The third mode (3.4%) shows a bipolar pattern with changes of opposite sign in the two hemispheres. A large warming in high Northern latitudes is accompanied by a smaller warming in the tropics and a cooling at high latitudes in the Southern Hemisphere. This mode is dominated by ODC and AMD, parameters which influence the surface density and freshwater forcing in the North Atlantic. Large values of ODC lead to a stronger AMOC and hence more heat being transported northwards in the Atlantic. This can result in a warmer Arctic, most likely associated with less sea-ice coverage, as seen in the spatial pattern of this EOF. Low AMD restricts moisture transport out of low latitudes, enhancing the development of the surface salinity at high latitudes, which leads to a stronger AMOC. The 'bipolar seesaw' pattern observed in North Atlantic vs the Southern Ocean is consistent with the expected role of the AMOC in extracting heat from the South Atlantic ocean and delivers it northward [52]. Also seen is the weaker 'Atlantic-Pacific seesaw', most likely to be associated with a positive feedback between the ocean circulation and the salinity contrasts on an interbasin scale [49].

5 Combined principal component analysis of PLASIM outputs

As mentions in the introduction, PCA was previously used to reduce the dimension of a single high-dimensional variable, the SAT in [56]. The link between two sets of PCs describing EMBM and PLASIM SAT was then determined using co-kriging. Since this convenient relationship is not guaranteed for the variables considered here (i.e., precipitation and wind speeds), we now explore the possibility of 'between variable' dimension reduction. This idea stems from the fact that output variables of different types are not independent, and correlations exist between not only within a spatial field or a timeseries but also across output types. For example, in an EMBM atmosphere, the SAT has a strong effect on both humidity and precipitation. Here, the relationship between the low-dimensional representations of different types of variables is examined. Figure 3 shows the scatter plots of the first 5 PCs of a) the zonal wind, b) the meridional wind, c) specific humidity and d) log-precipitation rate against the first 5 PCs of SAT. The logarithm of precipitation rather than precipitation was used because this improves the linear correlation between the two fields. PC pairs with correlation over 0.5 are annotated in the figure. These are not the correlations between pairs of spatial patterns, but the correlations across the training ensemble between the PCs of the different EOFs. The strong correlations between PCs of similar or the same ranks suggest that the two fields are indeed strongly correlated. The plot also highlights some interesting features in the relationship between fields. For the second PC of zonal wind vs. the first PC of SAT(Figure 3a), two nearly parallel branches can be seen in the scatter plot. This behaviour suggests that a bifurcation exists within our parameter space and the branches represent the two possible regimes. There are also indications of non-linear relationships, for example, between the first pair of PCs of humidity and SAT.

Given the strong correlation between same order PCs of these variables and SAT, 'between variables' dimensional reduction is applied to each pair using combined PCA (also known as combined EOF). This method has previously been employed to emulate a combination of four output time series from a simple climate model by [5]. This is a linear decomposition technique, and thus, nonlinear relationships between variables will be lost. However, given that the variables are related, by combining PLASIM SAT with an additional field such as humidity, we can use the EMBM's SAT as a fast approximation of the combined field in the multi-level emulator. We refer to this as 'cross-

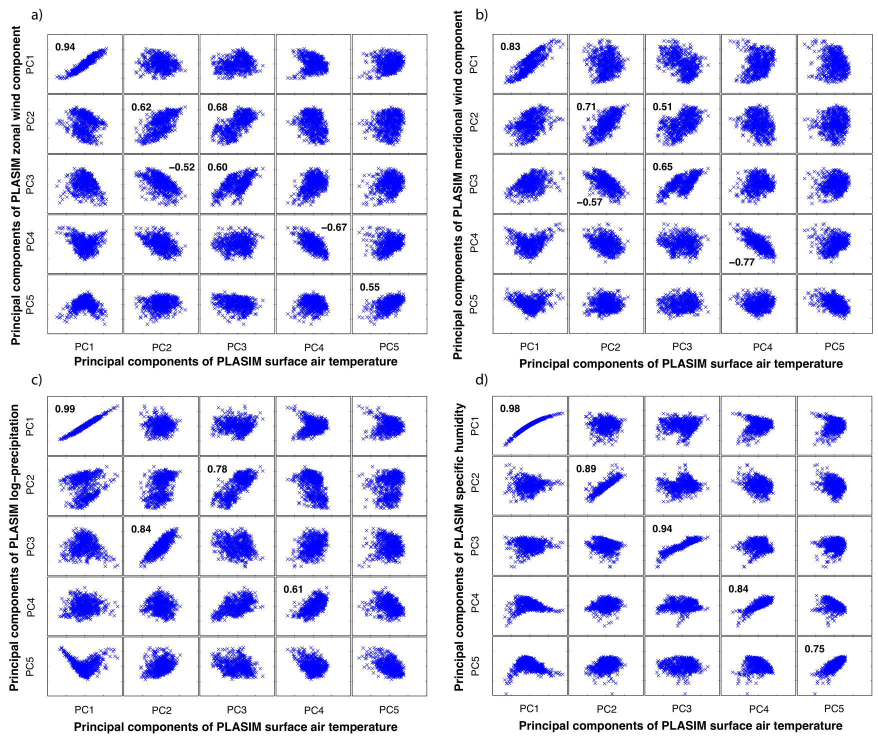

Fig. 3 Correlation between the PCs of SAT and a) zonal wind velocity, b) meridional wind velocity, c) log-precipitation, d) specific humidity. The PCs shown are obtained using the 600-member ensemble of PLASIM simulations.

variable' emulation as information is passed across different types of variable from different models.

While there is no limit on how many fields can be decomposed together, information on each individual field is lost when each individual field is added so for each variable, the SAT is decomposed together with only one other field each time. To do this, we simply concatenate PLASIM SAT and each field of interest together forming a new state vector

$$\mathbf{V}^k = [\mathbf{U}_{\mathrm{p}}^0, \mathbf{U}_{\mathrm{p}}^k], \tag{6}$$

where $\mathbf{U}_{\mathrm{p}}^0$ is the matrix of PLASIM SAT and $\mathbf{U}_{\mathrm{p}}^k$ is the other atmospheric variable; $k = 1, 2, 3$ and 4 for zonal wind, meridional wind, log-precipitation and humidity, respectively. SVD is then applied to decompose the $49152 \times n$ matrix $\mathbf{V}^k$.

Since we are putting together two different quantities with different units, the standardised state vectors are used to avoid having one field dominate the result solely because of its large relative magnitude compared to the other field. For example, the range of SAT is about four times larger than that of wind speeds, and so it is likely that the resulting pattern from non-standardised combined PCA will attribute a higher importance on the SAT. The columns of the standardised state vector are given by

$$\tilde{V}_{kj} = \frac{V_{kj} - \mu_k}{\sigma_k}, \tag{7}$$

where $\mu_k = \frac{1}{n}\sum_{j=1}^{n} V_{kj}$ is the ensemble mean and $\sigma_k = (\frac{1}{n-1}\sum_{j=1}^{n}(V_{kj} - \mu_k)^2)^{1/2}$ is the ensemble standard deviation at the k^{th} grid cell. This transformation makes the data adimensional and treats the two fields equally. It is possible to specify an arbitrary scaling constant to put more emphasis on one variable if desirable. However, we do not attempt this here. The standardisation leaves us with the same result as would be obtained if PCA was performed using an eigendecomposition on correlation matrices instead of a SVD. Therefore, this is often referred to as combined correlation PCA (or combined correlation EOF).

Table 5 The variance explained by the first 10 modes of zonal wind (UWN), meridional wind (VWN), precipitation (PTN) and humidity (HUM). The values listed are obtained using combined PCA of each variable with PLASIM SAT.

	UWN	VWN	PTN	HUM
1	29.47	24.86	27.03	88.76
2	20.46	25.91	14.33	5.48
3	11.65	9.49	9.73	2.05
4	6.07	4.77	7.38	0.60
5	4.96	3.85	5.66	0.80
6	3.52	3.23	3.42	0.34
7	2.19	2.92	3.49	0.32
8	1.88	2.49	2.54	0.08
9	1.38	1.49	1.92	0.15
10	1.09	1.11	0.98	0.01
Total	82.66	80.14	79.44	98.68

This procedure is applied to each of the four PLASIM variables. The variances explained by the first ten modes of $\mathbf{V}^k$ are listed in Table 5. Compared to SAT and humidity, the proportion of variance explained by the first ten PCs is lower for the wind speed components and precipitation rate.

6 Emulating multiple atmospheric outputs

6.1 Co-kriging emulator of combined fields

We now apply co-kriging to the PCs obtained from combined PCA. While PCA and emulation are done on fields containing 12 months, all figures in this section display the annual average fields.

Emulators of the PCs of PLASIM's variables, $\mathbf{V}^k$, are now constructed. Figure 4 shows the data used to construct co-kriging emulators of the PLASIM's

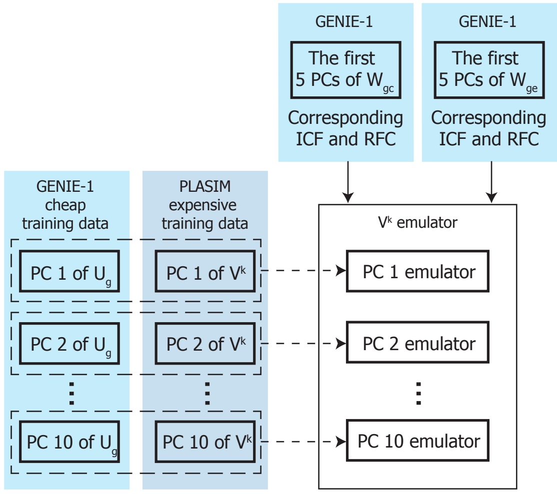

Fig. 4 The emulator of a PLASIM's variable, V^i, consists of 10 co-kriging emulators of the first 10 PCs of V^i. The cheap and expensive training data for each PC emulator are the corresponding PC obtained from U^0 and V^i. These 10 PCs are emulated as functions of the first 5 PCs of GENIE-1's SST, ICF and RFC. The matrices of SST corresponding to the cheap and expensive training data are W and W^1, respectively.

combined field $\mathbf{V}^k$. The SST fields obtained from GENIE-1's training ensemble are formed into matrix $\mathbf{W}_{\mathrm{gc}}$ and $\mathbf{W}_{\mathrm{ge}}$ of size $16332 \times n_c$ and $16332 \times n_e$, respectively. The superscript 1 for SST is dropped since we no longer use sea ice fields. These are the emulator inputs corresponding to the cheap GENIE-1, $\mathbf{U}_g$, and expensive PLASIM, $\mathbf{V}^k$, training data. The first 5 PCs of $\mathbf{W}_{\mathrm{gc}}$ and $\mathbf{W}_{\mathrm{ge}}$ are used together with ICF and RFC as the $\mathbf{V}^k$ emulator's inputs. All emulators constructed here use the same inputs.

The emulator of each field, $\mathbf{V}^k$, is actually a collection of 10 individual PC emulators. Each $\mathbf{V}^k$ PC emulator uses the corresponding PC pairs of $\mathbf{U}_g$ and $\mathbf{V}^k$, whose dimensions are $24576 \times n_c$ and $49152 \times n_e$, as training data. Overall, 40 emulators are constructed to emulate the surface field of PLASIM's zonal wind speed ($k = 1$), meridional wind speed ($k = 2$), log-precipitation ($k = 3$) and specific humidity ($k = 4$).

In the remainder of this section, we analyse the combined fields.

6.2 Surface wind velocity

While each wind speed component is decomposed separately with SAT for the construction of their emulators, both zonal and meridional wind components are decomposed together with SAT here:

$$\mathbf{V}^{\mathrm{wind}} = [\mathbf{U}_{\mathrm{p}}^0, \mathbf{U}_{\mathrm{p}}^1, \mathbf{U}_{\mathrm{p}}^2]^{\mathrm{T}}. \tag{8}$$

This allows a direct comparison between the modes of the two components.

Figure 5 shows the first three modes of SAT, zonal and meridional wind components. The wind components are defined to be positive in the eastward and northward direction for the zonal and meridional components, respectively. Therefore, positive values in the zonal wind EOFs mean a reduction if the mean flow is westward but an increase if the mean flow is eastward. For the meridional EOFs, in the Northern Hemisphere, positive values indicate stronger polewards winds while in the Southern Hemisphere, the opposite is true.

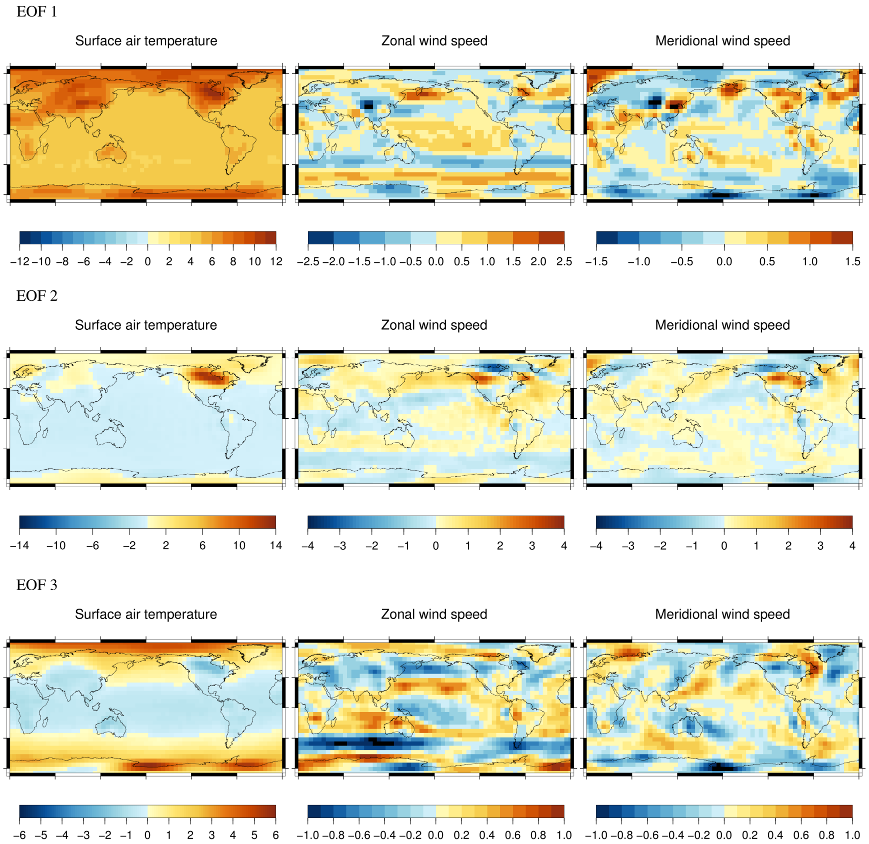

Fig. 5 The first three EOFs of SAT, zonal and meridional wind components obtained from a combined correlation PCA decomposition of $\mathbf{V}^{\text{wind}}$ using outputs from the whole 600-member ensemble.

In the first mode of SAT and wind anomalies, changes are associated with a monopole pattern in temperature with the largest variations over regions at high latitudes (the Arctic and Antarctica) and high elevations (the Tibetan Plateau, North America and Antarctica). The surface zonal wind speed is negatively correlated to SAT. In the warming case, weaker trade winds in both hemispheres are seen in the zonal and meridional wind components, indicating less active Hadley cells. The Westerlies are also slightly weakened and appear to shift poleward. There are also significant decreases in wind velocity over the Arctic and Antarctica in this EOF. Universal kriging coefficients show that this mode is predominantly driven by the second and first EOF of SST, which, in turn, are controlled by the parameters discussed in Section 4 and Figure 2. Therefore, the observed variations in the wind are due mostly to the changes in equator-to-pole temperature gradient and to a lesser extent, due to the global changes in SST. As the SAT at the poles warms up more than at the equator, the global temperature differential decreases, resulting in the observed reduction in wind speed.

The second mode of winds and air temperature is dominated by the Laurentide and the Fennoscandian ice sheets. The changes in elevation and albedo corresponding to different glacial masks lead to the large variation in SAT in these areas, e.g., the strongest wind anomalies are observed in the vicinity of the Laurentide in North America. The surface wind fields are modified due to both thermal and mechanical forcing. The Southern Hemisphere westerlies are slightly weaker while in the Northern Hemisphere, they are disrupted by the continental ice sheets. The third mode displays a strengthening of the westerlies associated with a larger equator-to-pole temperature difference. The anomalies due to both of these modes are weaker in general compared to the first mode. Changes in velocities at low latitudes are relatively small. However, significant local changes are seen due to the presence of the ice sheets.

6.3 Humidity and precipitation

The behaviour seen in surface specific humidity (Figure 6) is relatively straightforward; humidity increases as surface temperature rises and vice versa. The humidity in the tropics appears to be more sensitive to a change in temperature as a result of the Clausius-Clapeyron relation. As the temperature decreases, the atmosphere can hold less water vapour and hence has a lower humidity. This relationship is non-linear and a warm atmosphere can hold a much higher moisture content than a cold atmosphere. Thus, a small change in temperature in the tropics leads to a larger change in humidity than a similar change at high latitudes. This is evident in all three modes. While they have a very distinct spatial pattern compared to SAT, their responses to the changes in boundary forcing conditions, specified by the PCs are the same.

The precipitation pattern in the first EOF (Figure 7) is characterised by a large-scale drying, resulting from the reduced evaporation associated with the global cooling. Largest drying is seen over Antarctica. The desert zones, associated with the downwelling branch of the Hadley cells in both hemispheres experience an increase in precipitation. This is likely to be a result of the weaker Hadley cell associated with this mode of SAT anomaly. We observed stronger zonal and meridional components of the easterlies, which is an indica-

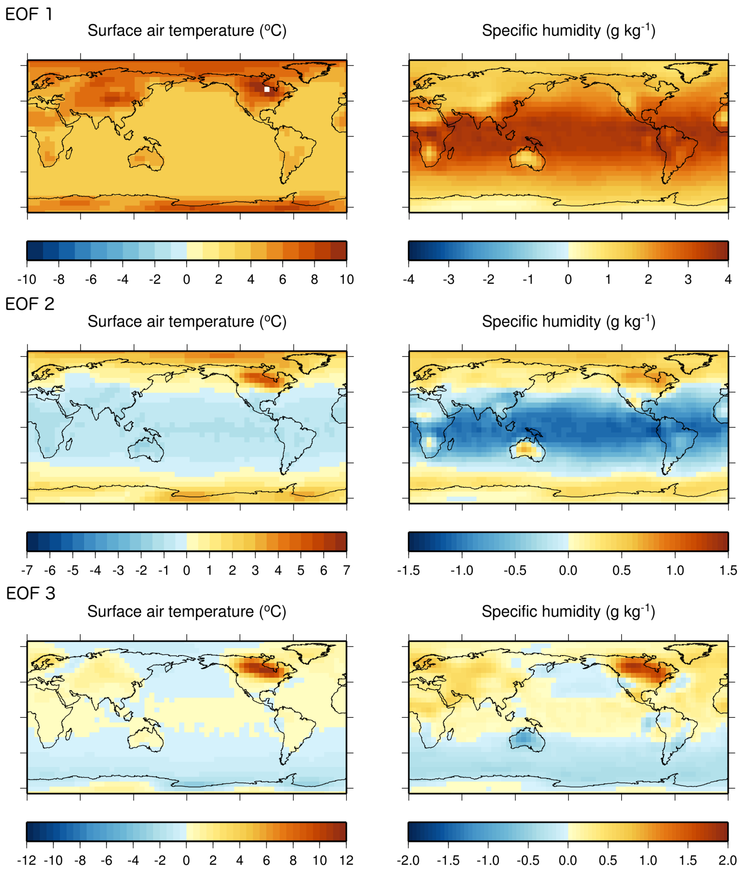

Fig. 6 The first three EOFs of SAT and specific humidity obtained from a combined correlation PCA decomposition of $\mathbf{V}^k$ with $k = 4$, using outputs from the whole 600-member ensemble.

tor of a stronger Hadley cell as SAT increases (Figure 5). A weakened Hadley cell would lead to a wetter dry band and less precipitation over the ITCZ.

In the second mode, we see more rain over the ocean where a positive anomaly in SAT induces more evaporation. Over the Africa and Australia continents, rainfall is negatively correlated to SAT anomaly. Similarly, in the third mode Australia and South Africa appear to become wetter when SAT decreases. In this EOF, the zonal wind anomaly appears to bring more moisture from the ocean into the South African and Australian continents. For both humidity and precipitation, regions with large variations are confined to the tropics and subtropics.

The checkerboard-like patterns seen in the precipitation plots are spurious numerical oscillations, also known as Gibbs oscillations [17]. They are numerical noise associated with the transformation of the truncated spectral

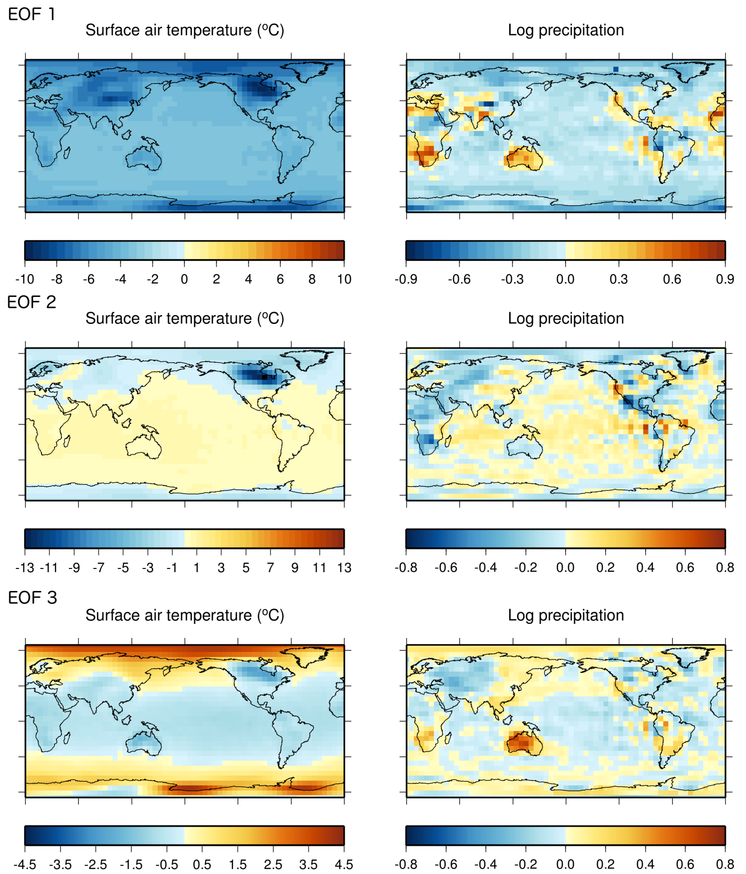

Fig. 7 The first three EOFs of SAT and log-precipitation obtained from a combined correlation PCA decomposition of $\mathbf{V}^k$ with $k = 3$, using outputs from the whole 600-member ensemble.

representation of a field to physical space, often seen in spectral models. The presence of such patterns makes it difficult to distinguish physical variations from noises.

The monthly variations of each set of EOFs show the same global signals to the annual fields discussed here, except for some seasonal features such as stronger signals associating with the winter hemisphere. The seasonal EOFs for precipitation, humidity and the zonal wind components are all included in the Supplementary section (Figure S1-S4).

7 Emulator validation

Once the emulators are constructed, we provide them with the inputs of the validating ensemble (the first 5 SST PCs, ICF and RFC) and receive in re-

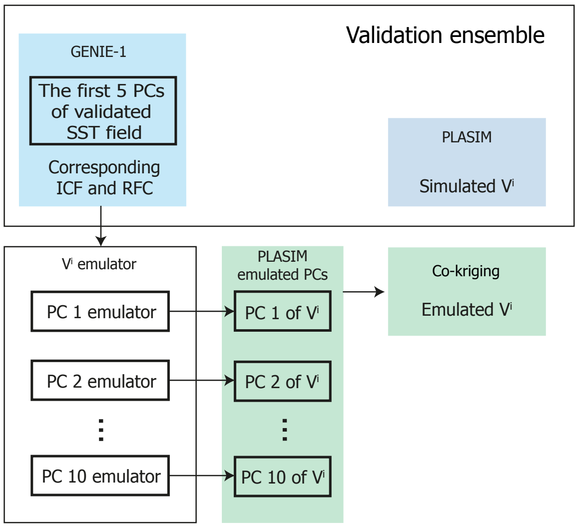

Fig. 8 The $\mathbf{V}^i$ emulator's predictors take the PCs of the validating SST, ICF, and RFC and produce the emulated PCs corresponding to those inputs. The 2-D emulated fields are reconstructed from the emulated PCs and the EOF patterns and are compared to the PLASIM's simulated outputs.

turn, the estimated PCs of precipitation, humidity and wind speed components corresponding to these new input values (Figure 8). The 2-D fields are then reconstructed from the predicted PCs and compared to the simulated values to evaluate how well the emulators do at predicting the ensemble behaviour and how well each simulation is reproduced. The emulators are assessed after each of the 10 PCs is emulated.

Two measures are produced to evaluate the emulator performance across the training ensemble, the average normalised RMSE and the percentage of total variance explained, V_T. For each surface field, the normalised RMSE, $\hat{\epsilon}$ is calculated as

$$\hat{\epsilon}_j = \frac{\epsilon_j}{U_j^{max} - U_j^{min}} \times 100, \tag{9}$$

where ϵ_j is the RMSE between the j^{th} emulated and simulated field and U_j^{max} and U_j^{min} are the maximum and minimum values of the same field. This quantity gives the error as a percentage of the spatial range in the corresponding field. The error plotted in Figure 9 is the normalised RMSE averaged over all 214 validation simulations. The proportion of the total ensemble variance captured by the emulator, V_T is

$$V_T = 1 - \sum_{j=1}^{n_v} \sum_{k=1}^{m\times 12} (S_{j,k} - E_{j,k})^2 \Big/ \sum_{j=1}^{n_v} \sum_{k=1}^{m\times 12} (S_{j,k} - \bar{S}_k)^2, \tag{10}$$

where $S_{j,k}$ is the simulated output at grid cell k in the j^{th} member of the validating ensemble, $E_{j,k}$ is the corresponding emulated output and $\bar{S}_k$ the ensemble mean simulated output at grid cell k. The total number of grid cells is $m \times 12$ since the monthly outputs from all 12 months are used in the validation. The size of the validating ensemble is $n_v = 214$.

The normalised RMSE and the variance explained by the first ten modes of the simulated fields are also shown (dashed line) in this figure. This is the percentage of variance explained that would be achieved by a perfect co-kriging emulator, limited only by information loss due to dimensional reduction of outputs into EOFs and PCs. The departure from this dashed line by the emulator's variance is a result of the component error. The kriging emulator (single level emulator using only PLASIM data) results for the same quantities are also included in the plot. These emulators use the same 90 expensive training points for each PLASIM's variable. They also use covariance PCA as opposed to combined correlation PCA as for co-kriging.

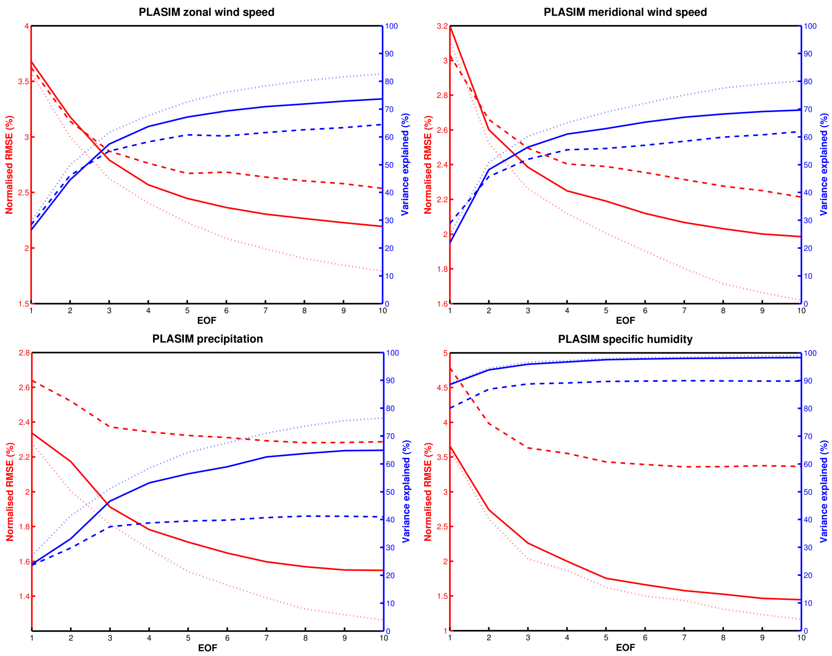

Fig. 9 The normalised RMSE and variance explained by the first 10 emulated EOFs for zonal wind speed, meridional wind speed, specific humidity and precipitation. Co-kriging (solid) is compared against kriging (dashed). The RMSE and the variance that would be achieved by a perfect emulator are also included for each plot (dotted line). Each co-kriging emulator uses 200 cheap and 90 expensive training points while a kriging emulator uses 90 expensive training points.

As more modes are added, emulator performances improve. In all cases, the first three modes are emulated most successfully, based on their high values of

r^2 - the coefficient of determination, and capture over 50% of the total variance while adding further modes improves the emulator by small but not negligible amounts. Starting from the fourth mode, the performance of kriging emulators of zonal wind and precipitation departed significantly from the performance limit (dashed line). Co-kriging emulator of higher order modes continue to contribute significantly after the first three modes.

For the wind speed emulators, 73.7% and 69.6% of the total variance for zonal and meridional components, respectively, are captured using co-kriging. They are approximately 10% lower than the total variance captured by the ten simulated modes but are 9.2% (zonal) and 7.6% (meridional) higher than the kriging emulator results. The final emulated fields for both components have an average normalised RMSE of 2%. For both components, the co-kriging emulators perform comparably or slightly under-perform compared to kriging in the first two modes. The kriging emulators outperform even the perfect co-kriging case here. This is because the first two modes obtained using ordinary PCA of wind speed explain more of the total variance than the combined PCA do. The fact that the co-kriging emulators ultimately outperform kriging ones demonstrate the value of the added information from EMBM's SAT.

The precipitation emulator performs less well in capturing the ensemble variance, explaining 64.9% of the total variance while the ten simulated modes explain 76.5%. Compared to the 41.0% achieved by the kriging emulator, the co-kriging step clearly adds more useful information than it takes away. The average normalised RMSE for precipitation is 1.6%, higher than those of the wind emulators. This RMSE is normalised against the range of the field. When RMSE is normalised against the standard deviation instead (RMSD in Figure 10), precipitation emulator does not outperform the wind emulators. Precipitation tends to exhibit more internal variation than other variables, resulting in a smaller fraction of the total variance being explained by the top EOFs. Among the four variables, precipitation is also least strongly correlated to SAT.

The humidity emulator performs best thanks to its similarity to SAT. 98.2% out of 98.9% of the ensemble variance is captured. The average normalised RMSE is 1.5%. Co-kriging, in this case, manages to capture 8.4% of the total variance that was not emulated by the kriging emulator. It is not surprising that the emulator for humidity performs well considering the high correlation between SAT and the specific humidity.

The expensive emulator's inputs (SST PCs) are members of the cheap input set when co-kriging is used. This means that the PCs corresponding to the 90 expensive training points were a subset of the PCs obtained from the decomposition of 200 SST fields. When kriging is used, however, only 90 SST fields are used in trying to capture the whole ensemble behaviour. As a result, the EOFs and PCs computed in the co-kriging case are more likely to be reliable and robust.

The result shown in Figure 9 are averaged over the 12 calendar months. A breakdown result of the variance explained for each month is included in the Supplementary section (Figure S5).

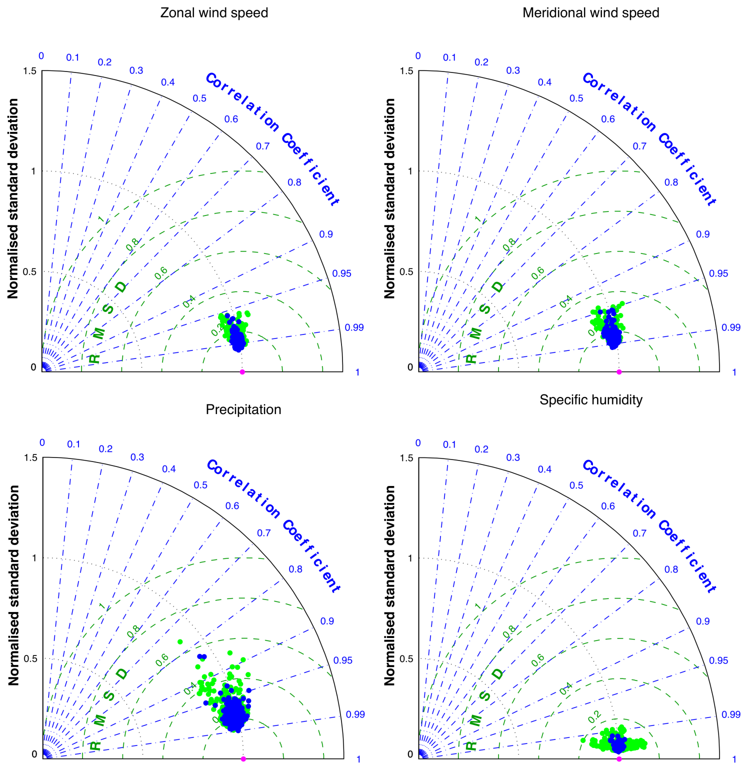

Fig. 10 A comparison of the emulated and simulated values of zonal and meridional wind components, precipitation and specific humidity using Taylor diagrams. The simulated outputs are treated as observations and the emulated values are compared against them in pairs. The blue dots represent emulated values from the co-kriging emulators while the green dots represent those from the kriging emulators.

Taylor diagrams (Figure 10) are used to compare each emulated field to its simulated one. The Taylor diagrams display, at the same time, the correlation coefficient, standard deviation and root mean square difference (RMSD) of the emulated fields with respect to their corresponding simulated field. Both standard deviation and RMSD are normalised by the simulated field's standard deviation. All four emulators can reproduce the surface patterns well, as demonstrated by the high average correlation. For all variables, except precipitation, the minimum correlation is approximately 95% or higher. For precipitation, while the vast majority of the emulated fields are well correlated to the simulated ones, there are two outliers with correlation below 85%. The standard deviations of the emulated ensembles are centred around 0.99 of the original field's values, indicating that the emulators very slightly underestimate the spatial variation in general. The average RMSD appears to be around 15-20% of the original field's standard deviation, except for humidity, where this value is much lower, around 5%. The emulated ensembles tend to cluster together on the diagrams indicating that the performance across the ensemble is mostly consistent.

The two outlying points in the precipitation diagram are associated with very large scores of SST PC1, which depends on RFC, OL0, RMX and OL1. A quick verification shows that the high RMSDs correspond to very large values of both RFC and OL0. While the PC scores for these two points are still within the training range, it is important to note that certain combinations of parameters of the validation simulations can lead to PC scores beyond the training range.

The kriging results are also shown in (Figure 10) for comparison. The collection of kriging points are more scattered with lower correlations, larger spread in standard deviations and lower RMSD. The co-kriging emulator of precipitation shows the largest improvement in terms of RMSD. The kriging emulator of humidity obtains a good correlation but tend to overestimate or underestimate the standard deviation of the spatial field.

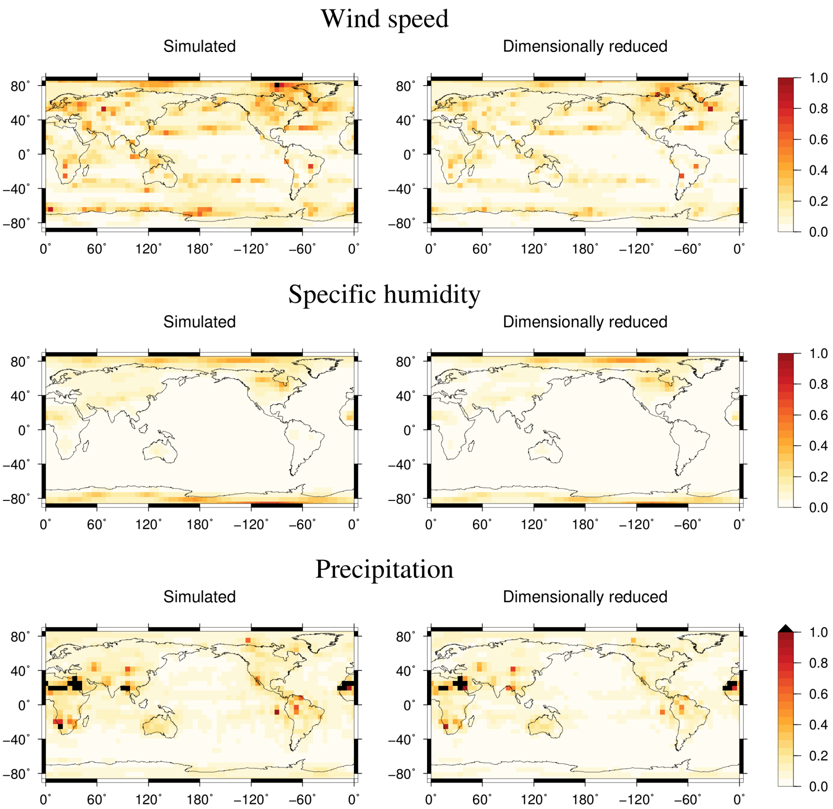

Fig. 11 The distribution of fractional error averaged over the 214-member validation ensemble for wind speed, humidity and precipitation. The total error is shown on the left while the emulator error is shown on the right.

The emulators appear to be doing a good job of reproducing PLASIM's simulated fields in general. Now we want to examine the distribution of errors to identify weak geographical regions in the emulator. The component errors are compared against the original and dimensional reduced fields by evaluating the error at each grid cell as a fraction of the values at that cell with

$$\hat{P}_j = \frac{|\hat{U}_j - U_j|}{U_j}, \tag{11}$$

where $\hat{U}_j$ is the emulated value at the j^{th} grid cell, and U_j is either the simulated or the dimensionally reduced value (the simulated field described by the first ten EOFs only) in the same grid cell. By computing this error using the dimensionally reduced fields, we only look at the error introduced by the emulation process. Using the simulated values gives the total error, a combination of component error and the truncation error. Figure 14 displays the geographical distributions of the total error (left) and component error (right). The errors are calculated for surface wind speed (top), humidity (middle) and log-precipitation rate (bottom). The surface wind speeds are calculated from the emulated zonal and meridional components. The white colour, which indicates small errors, dominates all three plots showing good agreement between the emulated and simulated values. The similarity between the total errors and the component errors suggests that component errors dominate, and truncation errors are less significant. In general, large fractional errors are associated with low values. This is clearest in precipitation, where fractional errors with a magnitude over 1 are seen in the Sahel, where very little rain is observed. The humidity emulator is valid everywhere. The errors are large over the Laurentide but they are still under 50%, and this area is expected to have low humidity. For wind speed, the area with the largest error is also in a location with very low winds to the west of Greenland.

Figure 12 summarises the emulation result by comparing the simulated and emulated mean fields for all four variables. The emulated minus simulated difference is shown in Figure 13. Since this is the difference in the mean fields across the emulated and simulated ensemble, it highlights the systematic error that prevails in all simulations. The emulated and simulated fields are very similar in all four cases. Areas of large errors are seen for the easterlies and Southern Hemisphere westerlies and to the region southeast of Greenland. Precipitation over the maritime continent in the Pacific appears to be underestimated. The largest differences in humidity are seen in the southern hemisphere tropical Pacific. Overall, the magnitude of these differences is small compared to the actual values. The errors in kriging emulators are larger but have similar patterns since the errors depend on the spatial pattern of the truncated modes.

Figure 14 shows the error in the predicted zonal wind in terms of the emulator's estimated standard deviation. For each validation point from the 214-member ensemble, the error at each emulated grid point is normalised by the estimated standard deviation at that point. The errors plotted are

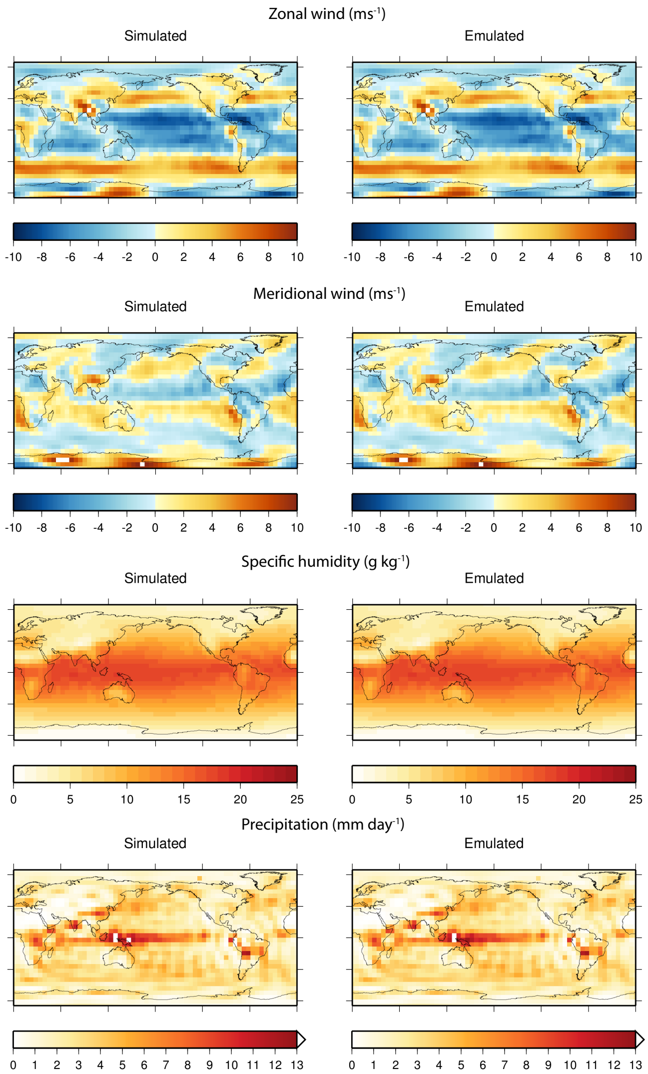

Fig. 12 Comparison between the simulated and emulated ensemble mean of the zonal wind speed, meridional wind speed, humidity and precipitation.

the difference between the emulated value at each grid point and the truncated simulated value (retain the first 10 EOFs only) at the same point. The truncated fields are used to exclude the errors introduced by the dimension reduction. The fraction of grid points that fall within 1, 2, 3, 4 and over 5 standard deviation ranges are shown in the figure. A well-calibrated emulator should have 66-95-99% of the error falls within the 1-2-3 standard deviation range. In our case, while the validation against simulated values showed that the predictions are close, the emulator appears to be overconfident and so the

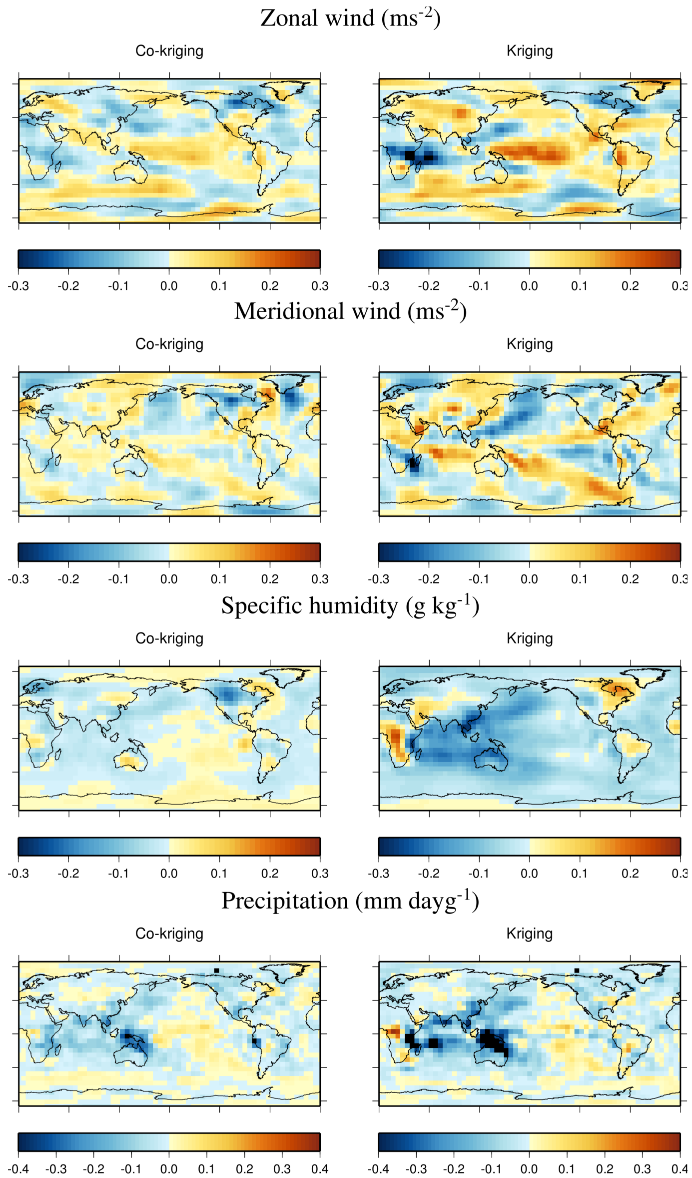

Fig. 13 The difference between the emulated and simulated fields for a) kriging and b) co-kriging emulators.

predicted standard deviation is too small. As a result, less than 55% of the grid points are fall within the 1 standard deviation of the emulated mean. There are several potential causes of this behaviour. It could be due to the assumption made on the mean and covariance structure (see the Appendix). Since we did not impose a specific mean function on the co-kriging emula-

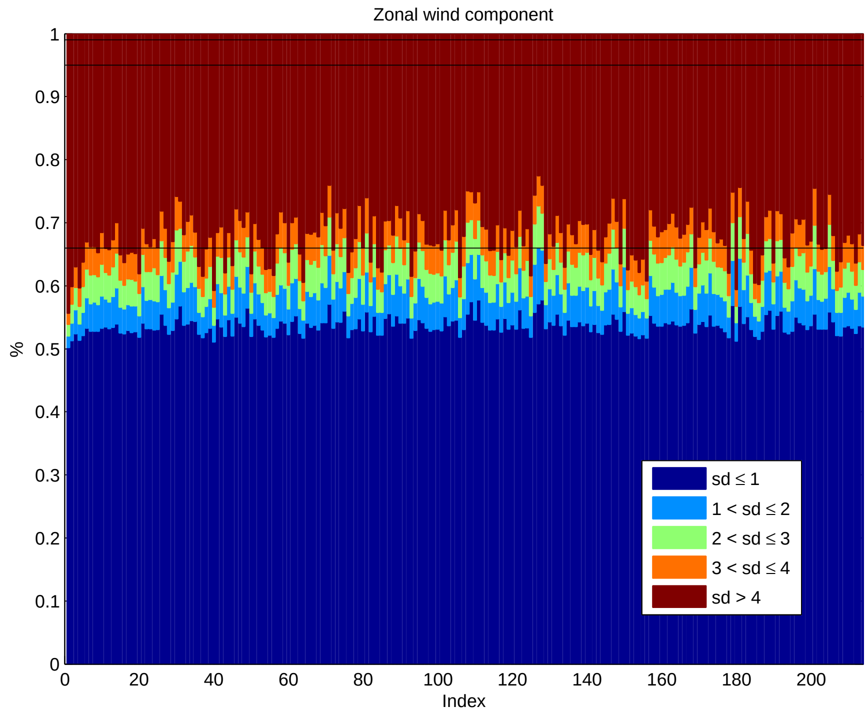

Fig. 14 Evaluation of the emulation error in terms of estimated emulator standard deviation for the zonal wind emulator. Each of the 214 validation points is plotted, showing the fraction of grid points that fall within the 1, 2, 3 and so on standard deviations from the mean. The black horizontal lines indicate the 66th, 95th and 99th inner quantiles of the distribution

tor, the problem could be due to the squared exponential correlation function used. This leads to a smooth covariance structure that can underestimate the credible intervals between training points. A coarser correlation such as the Matern can be used instead. It can also be due to an inappropriate value of the estimated hyperparameters. For example, an underestimation of the variance σ^2 (Eq. A.4 in the Appendix) can also lead to narrow credible intervals. The use of co-kriging might also contribute to this. As information is gained from EMBM, the emulator becomes too confident in regions where cheap training points are available. Further diagnostics are required to identify the root of the problem and to improve the emulator's estimated uncertainty. [3] provides a summary of several diagnostics that can be applied.

8 Summary and conclusions

We have successfully combined several statistical techniques to construct emulators relating PLASIM's high-dimensional atmospheric variables to the 2-D boundary forcing fields provided by GENIE-1's ocean component. Here, we have demonstrated that spatial and temporal correlations exist between two levels of model fidelity and between different variables. Then by employing combined PCA together with co-kriging, PLASIM's surface winds, precipitation and specific humidity can be emulated. Even though the EMBM atmosphere of GENIE-1 cannot simulate a realistic precipitation pattern and does not provide information on the surface wind fields, information on the atmosphere's response to changes in SST is extracted and used to improve the accuracy of the emulators. The emulators constructed are validated and shown to be capable of reproducing all four variables, most successfully for humidity and less so for precipitation. Areas, where systematic errors occur as a result of the linear decomposition and the fact that emulators tend to underestimate ensemble variability, are identified.

Furthermore, based on the work done by [24], PCA was also used to take into account the seasonal cycle by treating the monthly surface fields as a single field. The emulator constructed using this method has a knowledge of how the atmosphere behaves within an average model year as a function of the SST boundary forcing. Another important feature introduced in this study is the dimensional reduction of 2-D input fields following from the work of [25]. In this case, the original input consists of several fields (SST, SIC and SIH) which evolve in time (12 months). An ad hoc screening procedure was then carried out to reduce the number of inputs from three surface fields containing 48996 grid cells to the final number of seven inputs. While traditional model parameters are treated as independent, high dimensional spatial input fields have structures which, if reduced effectively, can be represented by a much smaller number of variables than the dimension of the fields. Since the computational cost of constructing an emulator depends on the number of inputs, due to an increased number of emulator parameters to be optimised, having a manageable number of inputs is desirable. The reduction step can be seen as a screening step, in which redundant information from 'inactive' inputs is removed. This technique has the advantage that while the number of model parameters or spatial resolution increases, it is likely that the number of PCs required would remain relatively low. A 'nugget' can be introduced to the emulators to account for the variability associated with these inactive inputs [5].

The successful emulation relies on the fact that while having distinctive spatial patterns, the first-order response in both models to boundary forcing conditions, described by their PCs, are well correlated and can be linked via the single multiplier formulation of co-kriging. While this connection works particularly well for the first few EOFs, the emulators of high-order PCs perform less well. This might be because the higher order processes in the two models are less correlated. Since the decomposition determines purely statistical modes, they often do not correspond to actual physical processes in the models. It is possible that if the dimensionally-reduced modes of variation represent meaningful behaviours of the system, the correlation between the PCs from the two models can be improved. A possibility which deserves consideration in this case is the rotation of principal components. In cases where the PCs are used purely for the purpose of dimension reduction, unrotated EOFs/PCs are good solutions. However, we are interested in understanding the model behaviour as well as identify meaningful relationships between two levels of model complexity through EOFs/PCs and rotated EOF offers the ability to isolate specific modes of variation. The idea is to transform the EOF to another system of coordinates by applying a rotation matrix which fulfils a specific criterion. There are various rotation criteria in literature as well as methodologies to analyse the data to identify the optimal choice of a solution. Examples of both can be found in [46]. Other possibilities, including the use of factor analysis, unrotated or rotated [43] instead of PCA or the use Canonical Correlation Analysis (CCA) to relate, for example, the wind components or precipitation rate of PLASIM to SAT of PLASIM or directly to SST of GOLDSTEIN. CCA defines coordinate systems such that the correlation between the projections of two different datasets onto these coordinate systems is mutually maximised [1, 19].

The treatment of temporal variation can also benefit from further work. In this work, we followed the approach of [24], which uses PCA to reduce the dimensionality of spatio-temporal data. This method treats the temporal behaviour of each quantity the same way as spatial variations. There exist several alternative approaches to emulate a timeseries output such as those discussed in [9]. The first and also the simplest method is the 'many single-output' emulators which emulate the outputs in time separately. In this situation, we would need to build 12 separate emulators, one for each month. Another possibility is the 'time input' emulator which treats time as another parameter and each month can be considered as a training point. This approach would lead to 12 times as many training points, which increases the computational expense required to build an emulator significantly. Finally, [9] proposed a 'multi-output' emulator which generalised the univariate GP emulator work with a multivariate output. Each of these methods has some advantages over the others. [24] pointed out that the use of spatio-temporal data allows for the possibility of abrupt transitions because spatial fields are not forced to be similar through time. The 'time input' method is more restrictive because it would impose a form in time. The 'many single-output' approach, on the other hand, fits separate emulators to each month and hence no fixed structure in time is assumed. This flexibility can be advantageous, depending on the application. This is less of an advantage in our case since we are not dealing with a long time series but with a seasonal cycle and spatial fields are expected to behave in a known fashion. A 'time input' emulator is a more rigid and computationally intensive special case of the 'multi-output' emulator. These methods, among other multivariate techniques, often deal with either high-dimensional spatial outputs or scalar outputs that evolve in time. Future research on techniques which incorporate both of these factors seems worth investigating, considering its applicability in works involving time-varying high-dimensional boundary conditions.

In conclusion, as a result of the dimensionality reduction, GENIE-1's oceanic fields can be used as inputs enabling the possibility to emulate PLASIM's atmospheric variables as a function of SST directly. This way, fluxes between the atmosphere and the ocean can be emulated, allowing the exchange between the ocean and the statistical emulators. The coupling between a climate model with a statistical emulator of a complex subcomponent can be achieved efficiently using this approach. In our case, the emulated fields can be passed back to GENIE-1, driving its ocean and sea-ice components. Indeed, an interactive coupling between GENIE-1's components with statistical emulators of PLASIM's atmospheric quantities is currently under construction and will be the focus of future studies.

This coupling method can also be applied to models which use climatological records or products from reanalyses as boundary/forcing conditions. This hypothetical model setup shares some similar aspects to the regional hybrid coupled model presented in [2] where a link between the SST from an OGCM to the wind stress field is determined based on the relationship obtained from observational SST and wind stress data. The two observational datasets are decomposed into separate EOFs whose PCs can be linked through a regression, which is then used to provide a prediction of the wind stress response to a new SST field. There exist other regional and global hybrid models which employ different approaches to learning the statistical link between observation fields such as the use of canonical correlation analysis in [1], singular value decomposition of observational data to supply wind stress anomaly seen in [53] and the representation of feedbacks due to the ocean-atmosphere interaction derived from a local deviation of SST through the use of linear regressions by [8].

Compared to directly replacing EMBM with PLASIM, this hybrid coupling strategy has the potential to produce a coupled model that is two orders of magnitude faster and therefore makes a substantial improvement in the range of timescale accessible to the class of coupled model. Here, we assume that the near-surface atmospheric variables are determined by SST, and the atmosphere responds instantaneously to changes in the boundary condition described by the ocean. Compared to the ocean, the atmospheric response time to climate forcing is relatively short due to its lower heat capacity. Thus, on the long timescales considered in palaeoclimate research, the assumption made about the equilibrium response of the atmosphere is acceptable. It is, however, not valid for applications, in which the atmosphere and ocean vary together on interannual timescales (e.g., the ENSO).

A Gaussian process emulator

The climate model, $f(\cdot)$, can be viewed as a function of a set of inputs, $\boldsymbol{x} = [x_1, \cdots, x_d]$, where d is the number of perturbed model parameters. This number is commonly referred to as the number of dimensions of the emulator. The output of each model run is a scalar value y. Supposed we have n simulation runs, providing n realisations $\boldsymbol{y} = [y_1 = f(\boldsymbol{x}_1), \cdots, y_n = f(\boldsymbol{x}_n)]$. These comprise the training set used to train an emulator.

First, the function $f(\cdot)$ is represented by a GP prior described by a mean function $m(\cdot)$ and a covariance function $V(\cdot, \cdot)$

$$f(\cdot)|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta} \sim \mathcal{N}(m(\cdot), V(\cdot, \cdot)). \tag{A.1}$$

This GP is used as a prior for Bayesian inference. The prior does not depend on the training data but specifies the assumptions made about the function of interest. Then, the outputs from a selected number of simulations are incorporated, allowing us to update the prior to the posterior GP. This process is called training the GP model. Following [28], $m(\cdot)$ and $V(\cdot, \cdot)$ are modelled hierarchically, meaning that they are parameterised in terms of hyperparameters. The mean function is given by:

$$m(\boldsymbol{x}) = \boldsymbol{h}^T(\boldsymbol{x})\boldsymbol{\beta}, \tag{A.2}$$

where $\boldsymbol{h}(\boldsymbol{x})$ is a vector of known regression functions of the inputs, describing a class of shapes of the function $f(\cdot)$. $\boldsymbol{\beta}$ is an unknown vector of coefficients. In the case of ordinary kriging, $\boldsymbol{h}(\cdot) = \mathbf{1}$, making $\boldsymbol{\beta}$ the unknown overall mean. A variation of kriging, called universal kriging, uses a linear mean function:

$$\boldsymbol{h}(\cdot) = (\mathbf{1}, \boldsymbol{x}^T), \tag{A.3}$$

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

The covariance function is given by:

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

in which σ^2 is an unknown variance of the GP and $\Psi(\cdot, \cdot)$ is the assumed correlation function:

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

The function Ψ represents 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 power exponential form of covariance structure is a popular choice due to its flexibility.

Both p and θ can be estimated for each dimension. For simplicity and to reduce computational cost, $p = 2$ is assumed for all dimensions. An independent value of θ is obtained for each dimension by maximising the likelihood of $\boldsymbol{y}$.

The specified GP is used as a prior for Bayesian inference and is parameterised in terms of the hyperparameters $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\theta}$ and p. Given that the prior is Gaussian, by analytically marginalising β and σ^2, the marginal likelihood of the observed outputs at n training points, $\boldsymbol{y}$, given θ and p can then be computed (estimated by maximising the likelihood of $\boldsymbol{y}$). A more detailed description of the derivations and formulations can be found in [37].

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 $\boldsymbol{\theta}$ and p, the posterior distribution found can be used to make predictions about the model's outputs at unsampled inputs. The predictive distribution is a Student's t-distribution, with $n-s$ degrees of freedom

$$p(f(\boldsymbol{x})|\mathbf{y},\theta) \sim t_{n-s}(m_1(\boldsymbol{x}), V_1(\boldsymbol{x},\boldsymbol{x'})), \tag{A.6}$$

with

$$m_1(\boldsymbol{x}) = \boldsymbol{h}^T(\boldsymbol{x})\hat{\boldsymbol{\beta}} + \boldsymbol{T}(\boldsymbol{x})\mathbf{A}^{-1}(\boldsymbol{y} - \boldsymbol{H}\hat{\boldsymbol{\beta}}) \tag{A.7}$$

and

$$V_1(\boldsymbol{x},\boldsymbol{x}') = \hat{\sigma}^2[\Psi(\boldsymbol{x},\boldsymbol{x'}) - \boldsymbol{T}(\boldsymbol{x})^T\mathbf{A}^{-1}\boldsymbol{T}(\boldsymbol{x}') + \mathbf{P}(\boldsymbol{x})(\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{P}(\boldsymbol{x'})^T], \tag{A.8}$$

where $\mathbf{H}$ is the regression matrix of the design points, $\mathbf{H} = \boldsymbol{h}(\boldsymbol{x})^T$, and $\mathbf{A}$ is the design points correlation matrix, $\mathbf{A} = \Psi(\boldsymbol{x},\boldsymbol{x'})$; $\boldsymbol{t}(\boldsymbol{x})$ is the correlation vector between $\boldsymbol{x}$ and the training set, i.e. $(\boldsymbol{T}(\boldsymbol{x}))_i = \Psi(\boldsymbol{x},\boldsymbol{x}_i)$ and $\mathbf{P}(\boldsymbol{x}) = \boldsymbol{h}(\mathrm{x})^T - \boldsymbol{T}(\boldsymbol{x})\mathbf{A}^{-1}\mathbf{H}$. The estimated values of σ^2 and $\boldsymbol{\beta}$ are indicated as $\hat{\sigma}^2$ and $\hat{\boldsymbol{\beta}}$, respectively:

$$\hat{\boldsymbol{\beta}} = (\mathbf{H^T}\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{A}^{-1}\boldsymbol{y} \tag{A.9}$$

and

$$\hat{\sigma}^2 = \frac{\boldsymbol{y}^T(\mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{H}(\mathbf{H}^T\mathbf{A}^{-1}\mathbf{H})^{-1}\mathbf{H}^T\mathbf{A}^{-1})\boldsymbol{y}}{n-q-2}. \tag{A.10}$$

A full description of the derivation of the posterior distribution is available in [45].

Co-kriging is an extension to this technique, which is applicable when a fast approximation of the primary simulator is available. For this method to work, the primary simulator and its approximation need to be correlated and contain information about one another.

When only a small number of expensive runs are 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 [14].

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

$$f_e(\boldsymbol{x}) = \rho f_c(\boldsymbol{x}) + f_d(\boldsymbol{x}). \tag{A.11}$$

The right-hand side of the equation consists of a cheap GP, f_c, multiplied by a scaling factor ρ and a separate GP, f_d, modelling the stochastic residual of the expensive model [27, 14]. Together these two terms describe the emulator of the expensive model. This approximation is chosen for its simplicity as well as the assumption that the main difference between the two models is largely a matter of scale. This assumption is made based on the fact that both EMBM and PLASIM are driven by the boundary conditions specified by GENIE-1's ocean. They essentially share similar inputs but have the ability to respond differently.

Two sets of training points are required for the construction of a co-kriging emulator; a cheap set $\boldsymbol{y}_c = f_c(\boldsymbol{x}_c)$, which finely samples the input space, and a small, sparse set $\boldsymbol{y}_e = f_e(\boldsymbol{x}_e)$ of expensive points. When the number of PLASIM training points is small, such that a kriging emulator cannot be built with high accuracy, co-kriging employing a large additional 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. A general rule of thumb for the number of training points for kriging is 10 times the number of parameters [33]. The inputs at which the expensive training set is obtained, $\boldsymbol{x}_e$, is a subset of the cheap set, $\boldsymbol{x}_c$. These expensive points are chosen using an exchange algorithm described by [10].

The covariance matrix for co-kriging, Ψ_{ck}, can be written in block form as

$$\Psi_{ck} = \begin{pmatrix} \sigma_c^2 \mathbf{A}_c(\boldsymbol{x}_c) & \rho\sigma_c^2 \mathbf{A}_c(\boldsymbol{x}_c, \boldsymbol{x}_e) \\ \rho\sigma_c^2 \mathbf{A}_c(\boldsymbol{x}_e, \boldsymbol{x}_c) & \rho\sigma_c^2 \mathbf{A}_c(\boldsymbol{x}_e) + \sigma_e^2 \mathbf{A}_d(\boldsymbol{x}_e) \end{pmatrix}, \tag{A.12}$$

with $\mathbf{A}_c = \Psi(\boldsymbol{x}, \boldsymbol{x'}; \boldsymbol{\theta}_c)$ and $\mathbf{A}_d = \Psi(\boldsymbol{x}, \boldsymbol{x'}; \theta_d)$. This covariance matrix encompasses the correlation between cheap points ($\mathbf{A}_c(\boldsymbol{x}_c)$), expensive points ($\mathbf{A}_c(\boldsymbol{x}_e)$ and $\mathbf{A}_d(\boldsymbol{x}_e)$) and the cross-correlation between the cheap and expensive points ($\mathbf{A}_c(\boldsymbol{x}_c, \boldsymbol{x}_e)$ and $\mathbf{A}_c(\boldsymbol{x}_e, \boldsymbol{x}_c)$). Details on the formulation and derivation of this equation can be found in [27] and [14].

References

1. Barnett TP, Preisendorfer R (1987) Origins and levels of monthly and seasonal forecast skill for United states furface air temperature determined by canonical correlation analysis. Monthly Weather Review 115(9):1825–1850
2. Barnett TP, Latif M, Graham N, Flugel M, Pazan S, White W (1993) ENSO and ENSO-related predictability. Part I: prediction of equatorial Pacific Sea surface temperature with a hybrid coupled ocean-atmosphere model. Journal of Climate 6(8):1545–1566
3. Bastos LS, O'Hagan A (2009) Diagnostics for Gaussian Process Emulators. Technometrics 51(4):425–438
4. Bayarri MJ, Berger JO, Cafeo J, Garcia-Donato G, Liu F, Palomo J, Parthasarathy RJ, Paulo R, Sacks J, Walsh D (2007) Computer model validation with functional output. The Annals of

Statistics 35(5):1874–1906, DOI 10.1214/009053607000000163, URL http://projecteuclid.org/euclid.aos/1194461715

5. Boukouvalas A, Cornford D (2009) Dimension Reduction for Multivariate Emulation. Tech. Rep. November 2009, Aston University, Birmingham
6. Bounceur N, Crucifix M, Wilkinson RD (2015) Global sensitivity analysis of the climate-vegetation system to astronomical forcing: An emulator-based approach. Earth System Dynamics 6(1):205–224
7. Challenor PG, McNeall D, Gattiker J (2010) Assessing the probability of rare climate events. In: O'Hagan A, West M (eds) The Oxford handbook of applied bayesian analysis, Oxford University Press, New York, chap 16, pp 403–430
8. Cimatoribus AA, Drijfhout SS, Dijkstra HA (2012) A global hybrid coupled model based on atmosphere-SST feedbacks. Climate Dynamics 38(3-4):745–760
9. Conti S, O'Hagan A (2010) Bayesian emulation of complex multi-output and dynamic computer models. Journal of Statistical Planning and Inference 140(3):640–651
10. Cook RD, Nachtsheim CJ (1980) A comparison of algorithms for constructing exact D-optimal designs. Technometrics 22:315–324
11. Edwards NR, Marsh R (2005) Uncertainties due to transport-parameter sensitivity in an efficient 3-D ocean-climate model. Climate Dynamics 24(4):415–433
12. Edwards NR, Cameron D, Rougier J (2011) Precalibrating an intermediate complexity climate model. Climate Dynamics 37(7-8):1469–1482
13. Foley AM, Holden PB, Edwards NR, Mercure JF, Salas P, Pollitt H, Chewpreecha U (2016) Climate model emulation in an integrated assessment framework: a case study for mitigation policies in the electricity sector. Earth System Dynamics 7(1):119–132, DOI 10.5194/esd-7-119-2016, URL http://www.earth-syst-dynam.net/7/119/2016/
14. Forrester AI, Sóbester A, Keane AJ (2007) Multi-fidelity optimization via surrogate modelling. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 463(2088):3251–3269
15. Fraedrich K (2012) A suite of user-friendly global climate models: Hysteresis experiments. The European Physical Journal Plus 127(5):53
16. Fraedrich K, Jansen H, Kirk E, Luksch U, Lunkeit F (2005) The Planet Simulator: Towards a user friendly model. Meteorologische Zeitschrift 14(3):299–304
17. Geil KL, Zeng X (2015) Quantitative characterization of spurious numerical oscillations in 48 CMIP5 models. Geophysical Research Letters 42(12):5066–5073, DOI 10.1002/2015GL063931
18. Haberkorn K, Sielmann F, Lunkeit F, Kirk E, Schneidereit A, Fraedrich K (2009) Planet Simulator Climate. Tech. rep., Meteorologisches Institut, Universität Hamburg
19. Hardoon DR, Szedmak S, Shawe-taylor J (2004) Canonical correlation analysis ; An overview with application to learning methods. Neural Computation 16(12):2639–2664

20. Higdon D, Gattiker J, Williams B, Rightley M (2008) Computer Model Calibration Using High-Dimensional Output. Journal of the American Statistical Association 103(482):570–583
21. Holden PB, Edwards NR (2010) Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling. Geophysical Research Letters 37(21)
22. Holden PB, Edwards NR, Oliver KIC, Lenton TM, Wilkinson RD (2010) A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. Climate Dynamics 35(5):785–806
23. Holden PB, Edwards NR, Müller SA, Oliver KIC, Death RM, Ridgwell A (2013) Controls on the spatial distribution of oceanic d13CDIC. Biogeosciences 10(3):1815–1833, DOI 10.5194/bg-10-1815-2013
24. Holden PB, Edwards NR, Garthwaite PH, Fraedrich K, Lunkeit F, Kirk E, Labriet M, Kanudia A, Babonneau F (2014) PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment. Geoscientific Model Development 7(1):433–451
25. Holden PB, Edwards NR, Garthwaite PH, Wilkinson RD (2015) Emulation and interpretation of high-dimensional climate model outputs. Journal of Applied Statistics 42(9):2038–2055, DOI 10.1080/02664763.2015.1016412
26. Holden PB, Edwards NR, Fraedrich K, Kirk E, Lunkeit F, Zhu X (2016) PLASIM GENIE v1.0 : a new intermediate complexity AOGCM. Geoscientific Model Development 9:3347–3361, DOI 10.5194/gmd-9-3347-2016
27. Kennedy MC, O'Hagan A (2000) Predicting the Output from a Complex Computer Code When Fast Approximations Are Available. Biometrika 87(1):1–13
28. Kennedy MC, O'Hagan A (2001) Bayesian calibration of computer models. Journal of the Royal Statistical Society Series B (Statistical Methodology) 63(3):425–464
29. Labriet M, Joshi SR, Vielle M, Holden PB, Edwards NR, Kanudia A, Loulou R, Babonneau F (2015) Worldwide impacts of climate change on energy for heating and cooling. Mitigation and Adaptation Strategies for Global Change 20(7):1111–1136
30. Lee La, Carslaw KS, Pringle KJ, Mann GW (2012) Mapping the uncertainty in global CCN using emulation. Atmospheric Chemistry and Physics 12(20):9739–9751
31. Lenton TM, Williamson MS, Edwards NR, Marsh R, Price aR, Ridgwell aJ, Shepherd JG, Cox SJ (2006) Millennial timescale carbon cycle and climate change in an efficient Earth system model. Climate Dynamics 26(7-8):687–711
32. Liakka J, Nilsson J, Lofverstrom M (2012) Interactions between stationary waves and ice sheets: linear versus nonlinear atmospheric response. Climate Dynamics 38(5-6):1249–1262
33. Loeppky JL, Sacks J, Welch WJ (2009) Choosing the Sample Size of a Computer Experiment: A Practical Guide. Technometrics 51(4):366–376

34. Lucarini V, Fraedrich K, Lunkeit F (2010) Thermodynamic analysis of snowball earth hysteresis experiment: Efficiency, entropy production and irreversibility. Quarterly Journal of the Royal Meteorological Society 136(646):2–11, DOI 10.1002/qj.543, 0905.3669
35. Lunt DJ, Williamson MS, Valdes PJ, Lenton TM, Marsh R (2006) Comparing transient, accelerated, and equilibrium simulations of the last 30 000 years with the GENIE-1 model. Climate of the Past 2(2):221–235
36. Maniyar DM, Cornford D, Boukouvalas A (2007) Dimensionality Reduction in the Emulator Setting. Tech. Rep. October 2007, Neural Computing research group, Aston University, Birmingham
37. Mardia KV, Marshall RJ (1984) Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika 71(1):135–146
38. Marsh R, Müller Sa, Yool A, Edwards NR (2011) Incorporation of the C-GOLDSTEIN efficient climate model into the GENIE framework: "eb_go_gs" configurations of GENIE. Geoscientific Model Development 4(4):957–992
39. Matthews HD, Caldeira K (2007) Transient climate-carbon simulations of planetary geoengineering. Proceedings of the National Academy of Sciences of the United States of America 104(24):9949–54
40. Mcneall DJ (2008) Dimension Reduction in the Bayesian Analysis of a Numerical Climate Model. PhD thesis, University of Southampton
41. Morris MD, Mitchell TJ (1995) Exploratory designs for computational experiments. Journal of Statistical Planning and Inference 43:381–402
42. Oakley JE, O'Hagan A (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66(3):751–769, DOI 10.1111/j.1467-9868.2004.05304.x, URL http://doi.wiley.com/10.1111/j.1467-9868.2004.05304.x
43. Osborne JW, Costello AB (2009) Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. Pan-Pacific Management Review 12.2:131–146
44. Peltier W (2004) Global glacial isostasy and the surface of the ice-age earth: The ICE-5G (VM2) model and GRACE. Annual Review of Earth and Planetary Sciences 32(1):111–149
45. Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press, Cambridge
46. Richman MB (1986) Rotation of Principal Components. Journal of Climatology 6(3):293–335
47. Romanova V, Lohmann G, Grosfeld K, Butzin M (2006) The relative role of oceanic heat transport and orography on glacial climate. Quaternary Science Reviews 25(7-8):832–845
48. Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. Statistical science 4(4):409–423
49. Saenko Oa, Schmittner A, Weaver AJ (2004) The Atlantic Pacific Seesaw. Journal of Climate 17(11):2033–2038, DOI 10.1175/1520-0442(2004)017¡2033:TAS¿2.0.CO;2

50. Santner TJ, Williams BJ, Notz WI (2003) The Design and Analysis of Computer Experiments. Springer, New York
51. Schmittner A, Silva TAM, Fraedrich K, Kirk E, Lunkeit F (2011) Effects of Mountains and Ice Sheets on Global Ocean Circulation. Journal of Climate 24(11):2814–2829
52. Stocker TF, Johnsen SJ (2003) A minimum thermodynamic model for the bipolar seesaw. Paleoceanography 18(4):n/a–n/a, DOI 10.1029/2003PA000920, URL http://doi.wiley.com/10.1029/2003PA000920
53. Syu HH, Neelin DJ, Gutzler D (1995) Seasonal and interannual variability in a Hybrid Coupled GCM. Journal of Climate 8:2121–2143
54. Thompson SL, Warren SG (1982) Parameterization of outgoing infrared radiation derived from detailed radiative calculations. Journal of the Atmospheric Sciences 39:2667–2680
55. Tran GT (2017) Developing a Multi-level Gaussian Process Emulator of an Atmospheric General Circulation Model for Palaeoclimate Modelling. PhD thesis, University of Southampton
56. Tran GT, Oliver KIC, Sobester A, Toal DJJ, Holden PB, Marsh R, Challenor PG, Edwards NR (2016) Building a traceable climate model hierarchy with multi-level emulators. Advances in Statistical Climatology, Meteorology and Oceanography 2(1):1–21
57. Weaver AJ, Eby M, Wiebe EC, Bitz CM, Duffy PB, Ewen TL, Fanning AF, Holland MM, Macfadyen A, Matthews HD, Meissner KJ, Saenko O, Schmittner A, Wang H, Masakazu Y (2001) The UVic earth system climate model : Model description , climatology , and applications to past , present and future climates. Atmosphere-Ocean 39(4):361–428
58. Wilkinson RD (2010) Bayesian Calibration of Expensive Multivariate Computer Experiments. In: Biegler L, Biros G, Ghattas O, Heinkenschloss M, Keyes D, Mallick B, Marzouk Y, Tenorio L, Waanders BvB, Willcox K (eds) Computational Methods for Large-Scale Inverse Problems and Quantification of Uncertainity, John Wiley & Sons, Ltd, Chichester, chap 10
59. Williamson D, Blaker AT, Hampton C, Salter J (2015) Identifying and removing structural biases in climate models with history matching. Climate Dynamics 45(5-6):1299–1324, DOI 10.1007/s00382-014-2378-z, URL http://dx.doi.org/10.1007/s00382-014-2378-z
60. Williamson M, Lenton T, Shepherd J, Edwards N (2006) An efficient numerical terrestrial scheme (ENTS) for Earth system modelling. Ecological Modelling 198(3-4):362–374