PLASIM-ENTSem v1.0: a spatio-temporal emulator of future climate change for impacts assessment

P. B. Holden[1], N. R. Edwards[1], P. H. Garthwaite[2], K. Fraedrich[3], F. Lunkeit[4], E. Kirk[4], M. Labriet[5], A. Kanudia[5], and F. Babonneau[6,7]

[1]Environment, Earth and Ecosystems, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[2]Mathematics and Statistics, The Open University, Walton Hall, Milton Keynes, MK7 6AA, UK
[3]Max Planck Institute of Meteorology, KlimaCampus, Bundesstraße 53, 20146 Hamburg, Germany
[4]Meteorological Institute, University of Hamburg, Bundesstraße 55, 20146 Hamburg, Germany
[5]Eneris Environment Energy Consultants s.l., c/Duque de Alba 15, 3–6, 28012 Madrid, Spain
[6]ORDECSYS, Place de l'Etrier 4, 1224 Chêne-Bougeries, Switzerland
[7]Economics and Environmental Management Laboratory, Swiss Federal Institute of Technology at Lausanne (EPFL), Switzerland

Correspondence to: P. B. Holden (philip.holden@open.ac.uk)

Abstract. Many applications in the evaluation of climate impacts and environmental policy require detailed spatio-temporal projections of future climate. To capture feedbacks from impacted natural or socio-economic systems requires interactive two-way coupling, but this is generally computationally infeasible with even moderately complex general circulation models (GCMs). Dimension reduction using emulation is one solution to this problem, demonstrated here with the GCM PLASIM-ENTS (Planet Simulator coupled with the efficient numerical terrestrial scheme). Our approach generates temporally evolving spatial patterns of climate variables, considering multiple modes of variability in order to capture non-linear feedbacks. The emulator provides a 188-member ensemble of decadally and spatially resolved ($\sim 5°$ resolution) seasonal climate data in response to an arbitrary future CO_2 concentration and non-CO_2 radiative forcing scenario. We present the PLASIM-ENTS coupled model, the construction of its emulator from an ensemble of transient future simulations, an application of the emulator methodology to produce heating and cooling degree-day projections, the validation of the simulator (with respect to empirical data) and the validation of the emulator (with respect to high-complexity models). We also demonstrate the application to estimates of sea-level rise and associated uncertainty.

1 Introduction

This manuscript describes the development of PLASIM-ENTSem (Planet Simulator coupled with the efficient numerical terrestrial scheme), a climate model emulator for use in coupling applications in which the underlying physical climate model would be computationally prohibitive. The emulator generates an ensemble of temporally evolving spatial patterns of climate variables, considering multiple modes of variability in order to capture non-linear feedbacks.

Emulators are widely used tools to inexpensively estimate expensive simulator output. They are generally used to calibrate input parameters or to estimate the uncertainty associated with a prediction (O'Hagan, 2006). The standard emulation approach uses a Gaussian process (Santner et al., 2003), conditioned on simulations at different inputs. This can then be used to predict the model response at a new set of inputs, together with an evaluation of the uncertainty of that prediction. This uncertainty represents an estimate of the error associated with the emulation process and is termed "code uncertainty" (O'Hagan, 2006). Further sources of uncertainty in the prediction are associated not with the emulator, but with the simulator. "Parametric uncertainty" arises from an uncertain knowledge of "best" simulator inputs and, when the simulator is too expensive, can be evaluated from the emulated response over plausible input space. "Structural error"

arises from imperfections in the model and is the irreducible error that remains when the "best" parameter inputs are applied (Rougier, 2007).

Although many emulation applications benefit from Gaussian process emulation, either through the quantification of code uncertainty or through the use of more general non-parametric fitting methods, our interest here is to quantify simulator uncertainty. When parametric uncertainty dominates over code uncertainty, relatively simple regression emulation techniques (that neglect code uncertainty) are adequate to achieve this (Holden et al., 2010; Holden and Edwards, 2010; Edwards et al., 2011). We here evaluate code uncertainty through cross-validation.

The extension of emulation techniques to consider multivariate outputs using approaches that can capture the correlations between the outputs have been developed (Rougier, 2008; Conti and O'Hagan, 2010). However, these approaches are not well suited to the emulation of very high dimensional output, although Rougier (2008) made advances in this regard by factorising the covariance matrix, and data reduction methods may be required to reduce the size of a problem (Wilkinson, 2010). Holden and Edwards (2010) applied this data reduction approach, using singular vector decomposition to project ∼ 1000-dimensional climate model output onto a lower-dimensional space and then emulating the map from the input space to the lower-dimensional output space. We here extend Holden and Edwards (2010) to develop a dimensionally reduced spatio-temporal climate model emulator for applications to impact assessment.

One of the principal obstacles to coupling complex climate and impact models is that the inclusion of feedbacks can induce a multiplier on the overall computational cost that renders the problem intractable. The conventional way to address this intractability is to use pattern scaling (Mitchell et al., 1999). However, replacing the climate model with an emulated version of its input–output response function represents a substantial advance on pattern scaling by retaining the possibility of including non-linear spatio-temporal feedbacks and uncertainty (Holden and Edwards, 2010). In addition to speed, this approach yields two further benefits in the field of integrated assessments. First, the emulation can allow for the construction of gradients of the response function. These may be required, for instance, in an optimisation-based application. (We note that adjoint models can also be used to provide local gradients, but their construction and additional computational cost can present major obstacles and, in addition, local gradients may be non-smooth.) Second, a calibrated statistical emulation, based on an ensemble of plausible simulations (Sect. 3.1), also provides a quantification of uncertainty and modelling errors.

The potential benefits of an approach using emulation have been discussed in some detail in Holden and Edwards (2010). However, several limitations of this earlier emulator have restricted possible coupling applications. Firstly, the emulation was applied to the intermediate-complexity atmosphere–ocean general circulation model (GCM) GENIE-2 (Lenton et al., 2007). The precipitation fields in that model are known to contain structural biases (Annan et al., 2005) and numerical artefacts, making it an unsatisfactory tool for impacts calculations. In order to address this we have developed a new climate model, the Planet Simulator (PLASIM; Fraedrich, 2012; Fraedrich et al., 2005a) coupled with the terrestrial carbon and land-surface model ENTS (Williamson et al., 2006). Secondly, the Holden and Edwards (2010) emulator was limited to spatial predictions at a single time slice. Here a single emulator calculation is used to derive the entire (decadally resolved) temporal history of each output field.

After presenting the new PLASIM-ENTS coupled model in Sect. 2, we describe the ensemble design in Sect. 3. In Sect. 4 we describe the construction and statistical validation of emulators of seasonally resolved temperature, temperature variability and precipitation fields. Heating and cooling degree days are derived (and empirically validated) for impact calculations in Sect. 5. The validation of the emulator with respect to the CMIP5 ensemble is covered in Sect. 6, while conclusions are discussed in Sect. 7.

2   The climate model: PLASIM-ENTS

The climate model used here is the Planet Simulator (PLASIM; Fraedrich, 2012; Fraedrich et al., 2005a), Version 16 Revision 4, with adaptations described below, most notably the incorporation of an alternative representation of terrestrial vegetation. PLASIM is an intermediate-complexity GCM built around the 3-D dynamical atmosphere PUMA (Fraedrich et al., 2005b). We run the model at T21 resolution ($\sim 5.6° \times 5.6°$) with 10 vertical levels.

The Planet Simulator (freely available under http://www.mi.uni-hamburg.de/plasim) is a climate model with a heat-flux-corrected slab ocean. In previous studies the model has been used to analyse the effect of vegetation extremes of a desert world versus green planet (Fraedrich et al., 2005b), the entropy budget and its sensitivity (Fraedrich and Lunkeit, 2008), the global energy and entropy budget in a snowball Earth hysteresis (Lucarini et al., 2010), and the double Intertropical Convergence Zone (ITCZ) dynamics in an aquaplanet setup (Dahms et al., 2011). This model is being employed to reconstruct historic climates (Grosfeld et al., 2007), to determine the younger history of the Andean uplift (Garreaud et al., 2010), to analyse the effect of mountains on the ocean circulation (Schmittner et al., 2011), and to evaluate biogeophysical feedbacks (Dekker et al., 2010). Furthermore, it enables investigations of climates very different from recent Earth conditions as shown in applications for Mars with and without ice (Stenzel et al., 2007), the Neo-proterozoic snowball Earth (Micheels and Montenari, 2008), and the Permian climates (Roscher et al., 2011). For recent

climate-change-related analyses see Bordi et al. (2011a, b, 2013).

The atmospheric dynamics (described in detail in above references) is based on primitive equations formulated for vorticity, divergence, temperature and the logarithm of surface pressure, solved via the spectral transform method. The parameterisations for unresolved processes consist of long- and shortwave radiation. The model takes into account only water vapour, carbon dioxide and ozone as greenhouse gases. The ozone concentration is prescribed according to an analytic ozone vertical distribution. The annual cycle and the latitudinal dependence are introduced. Further parameterisations are included for interactive clouds, moist and dry convection, large-scale precipitation, boundary layer fluxes of latent and sensible heat, and vertical and horizontal diffusion. The land-surface scheme uses five diffusive layers for the temperature and a bucket model for the soil hydrology. The ocean is represented by a mixed-layer (swamp) ocean, which includes a 0-dimensional thermodynamic sea-ice model.

We have adapted the PLASIM land-surface dynamics by the inclusion of the simple land-surface and vegetation model ENTS (Williamson et al., 2006). ENTS represents vegetative and soil carbon through a single plant functional type with photosynthesis a function of temperature, soil moisture availability, atmospheric CO_2 concentration and fractional vegetation cover. A double-peaked temperature response function is used to capture the different responses of tropical and boreal forest. Land-surface characteristics (albedo, surface roughness length and moisture bucket capacity) are diagnosed from the simulated state variables of vegetation and soil carbon densities. We note that land–atmosphere flux parameterisations are unchanged from those in PLASIM as these parameterisations in ENTS (Williamson et al., 2006) were developed for the EMBM module of GENIE. Thus, in PLASIM-ENTS, we only incorporate the ENTS parameterisations for vegetation and soil carbon densities and land-surface characteristics. The motivation for using ENTS in this study, rather than the SIMBA vegetation model incorporated in PLASIM Version 16.4 (Kleidon et al., 2006), is that the ENTS model behaviour has been thoroughly investigated through previous ensemble experiments (Holden et al., 2010, 2013a, b). The coupling was straightforward to implement given the structural similarities between ENTS and the default PLASIM vegetation model.

A number of minor modifications were made to the documented models described above. Firstly, we have introduced two new ENTS parameters, the optimum temperature for photosynthesis and the threshold soil moisture required for photosynthesis.

The optimum temperature parameter T_{adj} was introduced to allow for the uncertain response of photosynthesis to future warming, which is associated with uncertainty in the climate–carbon feedback, especially in the tropics (Matthews et al., 2007). The surface temperature T_a dependencies of photosynthesis in ENTS (Eqs. 19 and 20 of Williamson et al.,

2006) are replaced throughout with $T_a + T_{adj}$, where T_{adj} (an input parameter that is varied across the ensemble described in Sect. 3) acts to shift the photosynthesis diurnal average temperature optima from defaults of ~ 8 and $\sim 31\,°C$.

The variable soil moisture threshold was introduced primarily to allow the simulation of vegetation in semi-arid regions that were not vegetated in the default PLASIM-ENTS coupling. This arises as the "wetness factor" in PLASIM, which acts to linearly scale surface evaporation in order to inhibit evaporation from dry soils, is given by $W_s/0.4W_s^*$ (where W_s is the soil moisture content and W_s^* the bucket capacity), and takes its maximum value of unity when $W_s/W_s^* \geq 0.4$. This leads to drier soils than in the GENIE-ENTS coupling, where the wetness factor is given by $\left(W_s/W_s^*\right)^4$ (Eq. 9 of Williamson et al., 2006). The drier soils inhibit the growth of vegetation in PLASIM-ENTS using the standard ENTS parameterisations. To address this, the functional dependency of photosynthesis on soil moisture in PLASIM-ENTS has been altered to

$$f_2\left(W_s\right) = \left\{\left(W_s/W_s^*\right) - q_{th}\right\}/\left\{0.75 - q_{th}\right\}, \tag{1}$$

where $f_2\left(W_s\right)$ is restricted to values between 0 ($W_s/W_s^* \leq q_{th}$) and 1 ($W_s/W_s^* \geq 0.75$). The expression reduces to the standard ENTS parameterisation (Eq. 17 of Williamson et al., 2006) when the threshold fractional moisture $q_{th} = 0.5$. In the ensembles performed here (see Sect. 3), q_{th} is allowed to vary over the range 0.1 to 0.5.

The extent of modern sea ice is significantly overstated in the default PLASIM model. Sea-ice flux corrections are derived from fixed sea-ice spin-up simulations forced with climatological sea-ice coverage. However, during these spin-ups, the default PLASIM configuration assumes 100 % sea-ice coverage in any grid cell with non-zero climatological ice cover. This then leads to overstated modern-day sea ice in the dynamic flux-corrected mode, and an unreasonably strong sea-ice feedback. The fixed sea-ice configuration has been changed so that sea ice is assumed to be present only when the climatological grid-cell-averaged sea-ice thickness exceeds some threshold, a variable in the ensemble (Sect. 3).

The sea ice was found to be unstable in flux-corrected mode. This is likely a consequence of the fact that sea-ice coverage within a grid cell only takes values of zero or 100 %. Natural variability can lead to the establishment of instantaneous sea-ice coverage across an entire grid cell that may be stabilised due to the local albedo feedback. The result is that the sea-ice extent can drift towards greater coverage leading to an overestimate of modern-day sea-ice coverage. In an attempt to address this, a simple parameterisation was introduced for the latitudinal dependence of ocean albedo, representing the increased albedo of high-latitude ocean and hence the reduced differential between sea ice and ocean albedo. The simple parameterisation applied is

$$\alpha_s = \alpha_{s0} + 0.5\alpha_{s1}\left[1 - \cos\left(2\varphi\right)\right], \tag{2}$$

where the ocean albedo α_s is expressed in terms of latitude ϕ, the albedo at the equator α_{s0} (0.03) and the variable parameter α_{s1}. Although we have not performed the ensemble of simulations that would allow us to quantify the degree to which this may have improved sea-ice stability, the latitudinally dependent ocean albedo parameterisation did not eliminate the sea-ice instability. The parameterisation was retained as a more faithful representation of the latitudinal balance of shortwave radiation. However, it was discovered during review that ocean albedo for the direct beam is redefined during the radiative balance calculation, so the parameterisation in Eq. (2) is only applied to the scattered beam.

The sea-ice drift in PLASIM-ENTS is largely limited to the Southern Ocean, likely because climatological Arctic sea ice is thicker and exhibits less seasonal variability than Antarctic sea ice. The compromise ultimately adopted here was to simulate flux-corrected dynamic sea ice in the Arctic, but fixed sea ice in the Antarctic. Thus the simulations capture the feedback associated with Arctic sea ice (important for Northern Hemisphere impacts) without the bias introduced by the Southern Ocean sea-ice drift.

A number of input/output modifications were also made, being the addition of netCDF output routines, the diagnosis of seasonally averaged land-surface variables and the automated runtime generation of ocean and sea-ice flux corrections. Finally, the radiative-transfer scheme in PLASIM only allows for CO_2. We adapted the model to take two time-varying inputs: equivalent CO_2, the concentration that is equivalent to a given total radiative forcing (for provision to the radiative balance calculation), and actual CO_2 (for input to ENTS for CO_2 fertilisation).

3 Ensemble design

The design procedure is summarised in the flowchart in Fig. 1. The philosophy for the design process has been described in detail elsewhere (Holden et al., 2010). In short, the approach attempts to vary key model parameters over the entire range of plausible input values and to accept parameter combinations which lead to climate states that cannot be un-controversially ruled out as implausible (Edwards et al., 2011). The approach represents an attempt to find plausible parameterisations of the model from all regions of the high-dimensional input space in order to capture the range of possible feedback strengths.

Exploratory ensembles were performed that varied ranges for 22 key model parameters as widely as could be reasonably justified. On the basis of these, the input range for *th2oc* (Table 1) was narrowed to exclude regions where the marginal posterior distribution was never consistent with a plausible simulation (unrealistically high globally averaged precipitation). The resulting input ranges are summarised in Table 1. Twelve of these parameters were chosen to capture uncertainties in atmospheric transport and in the atmospheric
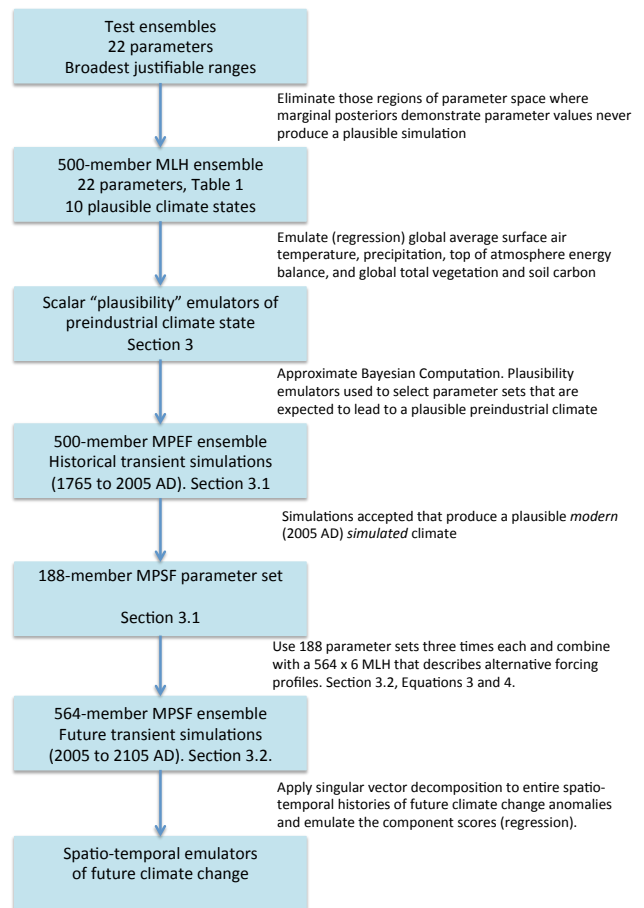

Fig. 1. A flowchart representation of the ensemble design and emulation.

radiative balance. The sea-ice parameter *xmind* was varied to provide a range of modern sea-ice configurations (global annually averaged preindustrial coverage varying between ~ 21 and 27 million km^2.); this variability across the ensemble is designed as a proxy to generate uncertainty in the strength of the sea-ice feedback. Mixed layer thickness was varied to capture uncertain ocean heat capacity and thermal inertia. The three new parameters introduced in Sect. 2, describing ocean albedo and the uncertain vegetation response to temperature and moisture availability, were varied, in addition to five key ENTS parameters, as previously identified in Holden et al. (2013a).

In order to investigate the regions of this 22-dimensional parameter space that produce plausible climate simulations, the parameters were first varied over their maximum plausible ranges (Table 1) to create a 500-member maximin Latin hypercube (MLH) design, using the maximinLHS function of the lhs package in R (R development Core Team, 2013). (We note that a range of 0 to 16 was applied for *nhdiff* in the MLH ensemble, but the range was reduced for the "ABC filtering" (see text in following paragraphs) to be close to the

Table 1. The 22 parameters and their prior ranges.

Module	Parameter	Process	Min	Max	Units
PUMA	vdiff_lamm	Vertical diffusivity	10	500	m
PUMA	nhdiff	Cut-off wave number for horizontal diffusivity	14*	16	
PUMA	tdissd	Horizontal diffusivity of divergence	0.05	10	days
PUMA	tdissz	Horizontal diffusivity of vorticity	0.05	10	days
PUMA	tdisst	Horizontal diffusivity of temperature	0.05	10	days
PUMA	tdissq	Horizontal diffusivity of moisture	0.05	10	days
PLASIM	rhcritmin	Minimum relative critical humidity	0.5	1.0	
PLASIM	gamma	Evaporation of precipitation	0.001	0.05	
PLASIM	tswr1	SW clouds (visible)	0.01	0.5	
PLASIM	tswr2	SW clouds (infrared)	0.01	0.5	
PLASIM	acllwr	LW clouds	0.01	5.0	$m^{-2} g^{-1}$
PLASIM	th2oc	LW water vapour	0.01	0.1	
SEA ICE	xmind	Minimum sea-ice thickness	0.10	0.3	m
OCEAN	albseamax	Ocean albedo (α_s Eq. 2)	0.10	0.3	
OCEAN	dlayer	Ocean slab thickness	300	700	m
ENTS	k14	Photosynthesis CO_2 fertilisation	30	750	ppm
ENTS	tadj	Photosynthesis optimum temperature adjustment	0.0	5.0	K
ENTS	qthresh	Photosynthesis moisture threshold	0.10	0.5	$kgC^{-1} m^2$
ENTS	k17	Fractional vegetation dependence	0.25	1.0	$kgC m^{-2}$
ENTS	k18	Base photosynthesis rate	3.0	7.0	$kgC m^{-2} yr^{-1}$
ENTS	k26	Leaf litter rate	0.075	0.26	yr^{-1}
ENTS	k32	Temperature dependence soil respiration	197	241	K

* Note that range applied for *nhdiff* in the MLH hypercube was 0 to 16; the tabulated range of 14 to 16 was applied during the ABC filtering (i.e. in the MPEF and MPSF ensembles) as low values of *nhdiff* were found to be associated with unrealistic precipitation distributions in the training ensemble.

tuned value of 15; low values were found to produce unrealistic precipitation distributions that, to illustrate, were in general found to lead to dense vegetation throughout Australia and southern Africa, the absence of vegetation in central Amazonia and loss of boreal vegetation in central Eurasia.) The points of this experimental design (corresponding to sets of input parameter values for PLASIM-ENTS) were used to generate a 500-member ensemble of 200 yr spun-up preindustrial simulations (with climatologically prescribed sea surface temperatures and sea-ice coverage). We note that the arbitrary combinations of parameter values in an MLH design do not, in general, produce a reasonable climate state; only 10 of the 500 parameter sets in this MLH ensemble produced a plausible climate state. This necessitates a further step in the statistical design.

The MLH simulations were used to generate scalar emulators for key global model outputs (surface air temperature, precipitation, top-of-atmosphere energy balance, vegetative carbon and soil carbon). The five emulators were built performing stepwise regression including linear, quadratic and cross-terms for all 22 parameters using the stepAIC function (Venables and Ripley, 2002), following the procedure described in Holden et al. (2010).

The parameters that are most strongly constrained by the emulators are illustrated in Fig. 2, the criterion for inclusion here being that a parameter must explain at least 5 % of the

variance in at least one of the five emulators (see Holden el al., 2010, for a similar analysis using this "total effect" sensitivity approach). The SAT and top-of-atmosphere energy balance emulators are controlled dominantly by those parameters that influence the effect of clouds on both shortwave and longwave radiation. The precipitation emulator is dominantly controlled by the parameter that influences the effect of water vapour on outgoing longwave radiation. The vegetation carbon emulator dominantly constrains the photosynthesis and leaf-litter rates. The soil carbon emulator dominantly constrains the photosynthesis rate and the temperature dependence of soil respiration. (It should be noted that, although the 10 parameters in Fig. 2 dominate simulated uncertainty in the preindustrial state, it should not be assumed that other parameters are unimportant for determining the future state. For instance, *rhcritmin*, *dlayer* and *vdiff_lamb* are amongst the most important parameters in determining the magnitude of transient future warming.)

These five emulators were then applied to generate a second 500-member ensemble, the "Modern-Plausible-Emulator-Filtered" MPEF ensemble, using approximate Bayesian computation (ABC; Beaumont et al., 2002), whereby parameters were drawn randomly from their defined input ranges but only accepted when the emulators predict they are consistent with a reasonable modelled preindustrial state. This plausible state is defined as global surface
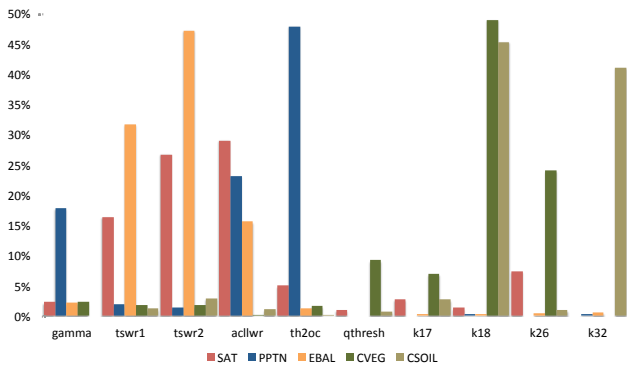
Fig. 2. "Total effects" for the 10 parameters that explain at least 5 % of the variance in at least one of the five plausibility emulators (Sect. 3). The total effect of a parameter is defined as the expectation of the variance that would remain if all other parameters were known; see Holden et al. (2010) for a more formal definition. Total effects are here normalised to 100 % to approximate the percentage contribution of each parameter to the variance of each emulator.

air temperature in the range 11 to 13 °C, precipitation (1025 to 1075 mm yr^{-1}), top-of-atmosphere energy balance (-3 to 3 W m^{-2}), vegetation carbon (450 to 650 GTC) and soil carbon (1000 to 2000 GTC). In this filtering process we specify a narrower range for plausibility than we accept in the simulator itself (Sect. 3.1). The emulator is cheap to run and we can afford this additional computational expense. This is done because we know that the emulators are imperfect and so this step will reduce wastage during the simulation filtering. It also has the advantage of producing more simulations near the centre of the prior, the region that is most plausible. (We note that Gaussian process emulators would be a useful alternative in this selection process, allowing us to rule out only those parameter sets that do not contain the observations within some confidence level.)

In summary, the MPEF parameter set is a Monte Carlo approximation to $\pi(\theta|D_P, S_P)$, where θ is the parameter vector, and D_P and S_P are the observed and simulated preindustrial climate states.

3.1 Historical transients

The 500 MPEF parameter sets were used as inputs for transient simulations with historical radiative forcing from 1765 to 2005 of http://www.pik-potsdam.de/~mmalte/rcps/. Temporally varying, globally averaged radiative forcing was provided, converted into equivalent CO$_2$, together with actual CO$_2$ concentration for vegetation input. Each simulation was initially run to equilibrium with preindustrial forcing, assuming PMIP II sea surface temperature and sea-ice distributions (monthly averages over the whole time period 1870 to 2006, www-pcmdi.llnl.gov/projects/amip). Monthly ocean and sea-ice flux corrections were diagnosed from these equilibrium states. The dynamic flux-corrected mixed-layer

ocean was applied in these historical transients, but sea ice was held fixed globally. The simulations were subjected to four present-day (2005) plausibility tests: global average land temperature (required to be in the range 11.5 to 13.5 °C), global average precipitation (1000 to 1100 mm yr^{-1}), total vegetative carbon (400 to 700 GTC) and total soil carbon (800 to 2200 GTC). Additionally, the top-of-atmosphere energy balance was required to be in approximate equilibrium (-5 to 5 W m^{-2}) in the initialised *preindustrial* state. On the basis of these five tests, 188 simulations were accepted for the future forced experiments, forming the "Modern-Plausible-*Simulator*-Filtered" MPSF parameter set.

In summary, the MPSF parameter set is a Monte Carlo approximation to $\pi(\theta|D_P, S_P, D_M, S_M)$, where D_M and S_M are the observed and simulated modern (2005) climate state.

3.2 Future ensemble

For the future transient simulations, the experimental configuration was changed to model dynamic, flux-corrected Arctic sea ice, but retaining fixed sea ice in the Antarctic (see Sect. 2). The motivation for this came from two exploratory ensembles. The first, with globally fixed sea ice, was found to greatly understate polar amplification, with implications for the evaluation of high-latitude impacts. The second, with globally dynamic sea ice, produced overstated present-day Antarctic sea ice, a cold-biased global temperature and excessive high-latitude southern warming in response to future forcing. The chosen compromise allows us to capture the uncertain response of Arctic sea ice, which is important for the evaluation of high-latitude Northern Hemisphere impacts. Although Antarctic warming is somewhat understated in this experimental setup, a minor issue for the evaluation of societal impacts, the spatial and seasonal distribution of warming otherwise compares favourably to state-of-the-art GCM predictions (Sect. 6).

Future radiative forcing (2005 to 2105) was expressed in terms of CO$_2$ equivalent concentration CO$_{2e}$, with a temporal profile described by a linear decomposition of the first three Chebyshev polynomials:

$$\text{CO}_{2e} = \text{C}_{0e} + 0.5 \left\{ A_{1e}(t+1) + A_{2e}\left(2t^2 - 2\right) \right.$$
$$\left. + A_{3e}\left(4t^3 - 4t\right) \right\}, \tag{3}$$

where C$_{0e}$ is CO$_{2e}$ in 2005 (386.5 ppm), t is time (2005 to 2105) normalised onto the range $(-1, 1)$ and the three coefficients that describe the concentration profile (A_{1e}, A_{2e} and A_{3e}) take values that allow for a wide range of possible future emissions profiles (0 to 1000, -200 to 200 and -100 to 100 ppm, respectively). These ranges encompass the range of representative concentration pathways (RCP) scenarios (Meinshausen et al., 2009; Moss et al., 2010) with a CO$_{2e}$ concentration in 2105 ranging from 387 to 1387 ppm. The motivation for the use of a modified Chebyshev polynomial

is to facilitate emulation, by reducing (and approximating) an arbitrary forcing profile to three input coefficients.

A weakness of the modified Chebyshev approach is that it is designed to operate over a specified time interval and is less well suited to fitting an arbitrary ensemble, such as a multi-model ensemble in which simulations have been performed over differing time intervals. Alternative approaches, including higher-order polynomials, may be more appropriate in such instances. Although the approach appears to violate temporal causality (emulated climate will in general be a function of A_{1e} at all times, whereas A_{1e} is determined by 2105 forcing), no statistical model is constrained by any physical law. Rather the statistical models are trained upon a physical model that is constrained by such laws. The accuracy of the statistical fits is addressed in detail in Sect. 4.3.

The same approach was taken to describe the temporal profile of actual CO_2 (with $C_0 = 380.2$ ppm):

$$CO_2 = C_0 + 0.5 \left\{ A_1 \left(t + 1 \right) + A_2 \left(2t^2 - 2 \right) \right.$$
$$\left. + A_3 \left(4t^3 - 4t \right) \right\}. \tag{4}$$

Each of the 188 MPSF 2005 spun-up states and parameter sets was used three times with different future greenhouse gas concentration profiles. Modified Chebyshev coefficients follow a 564×6 maximin Latin hypercube design. The resulting 564-member ensemble of future simulations provided the training data for the dimensionally reduced emulators described in Sect. 4.

4 The PLASIM-ENTS emulator

4.1 Singular vector decomposition

At this stage, we have an ensemble of 564 transient simulations of future climate, incorporating both parametric uncertainty (22 parameters) and forcing uncertainty (6 modified Chebyshev coefficients). For coupling applications we require an emulator that will generate spatial patterns of climate through time. To achieve this, output data fields at 10 time slices were generated for each ensemble member for each selected output variable. The time slices are decadal averages over the periods (1 January 2005 to 1 January 2015) through to (1 January 2095 to 1 January 2105), expressed as anomalies relative to the baseline period (1 January 1995 to 1 January 2005). We emulate seasonally resolved surface air temperature (SAT), SAT variability and precipitation[1]. These are together appropriate for a range of impacts and can be used, for instance, to derive estimates for maximum/minimum daily temperature in a given season, or seasonal heating/cooling degree days (Sect. 5).

---

[1]Separate emulators are built for each variable in each season (DJF, MAM, JJA and SON)

For each ensemble member, and each variable of interest, the 10 time slices were combined into a single 20 480-element vector where, for instance, the first 2048 elements describe the 64×32 output field over the first averaging period. This vector thus describes the temporal and spatial dependence of a given output (e.g. December-January-February, or DJF; SAT) for its respective ensemble member. These vectors were combined into a $(20\,480 \times 564)$ matrix $\mathbf{Y}$ describing the entire ensemble output of that variable.

An alternative approach would have been to decompose the spatial fields only, collating data from different times to give more observations and introducing time as an input variable to the emulator. We prefer here to decompose spatio-temporal data as this allows for the possibility of abrupt state transitions (i.e. spatial fields are not forced to be similar through time). Furthermore, under our approach we do not need to assume that a linear/quadratic model in time is required; consider the emulator as a function of the third Chebyshev coefficient, which describes the temporal variation of forcing through a combination of cubic and linear terms (Eqs. 3 and 4). We note that a potential advantage of a collated spatial decomposition would be to provide interpolated output for the emulation of sub-decadal time slices; although we assume polynomial time variation of input forcing (here cubic in time), no assumption is made about the time variation of output.

Singular vector decomposition (SVD) was performed on the matrix (Mardia et al., 1979).

$$\mathbf{Y} = \mathbf{LDR}^T, \tag{5}$$

where $\mathbf{L}$ is the $(20\,480 \times 564)$ matrix of left singular vectors ("components"), $\mathbf{D}$ is the 564×564 diagonal matrix of the square roots of the eigenvalues and $\mathbf{R}$ is the 564×564 matrix of right singular vectors ("component scores"). This decomposition produces a series of orthogonal components. The first component is the linear combination of the column vectors of $\mathbf{Y}$ that describes the maximum possible proportion of the variance amongst the column vectors of $\mathbf{Y}$. Each subsequent component satisfies the same constraint, except it is additionally constrained to be orthogonal to the previous components. The SVD thus produces a set of 564 orthogonal components, ordered so that the proportion of the ensemble variance that each explains decreases sequentially. Any one of the 564 simulated fields can be derived as a linear combination of the 564 components.

We have chosen to perform SVD on the non-centred climate change anomaly fields. The more usual approach would have been to first centre the data (i.e. by subtracting the ensemble mean). Although a non-centred decomposition can have advantages (Noy-Meir, 1973), it should be undertaken with care. In a non-centred decomposition, the direction of the first component is from the origin to the centre of the cloud of data. In general, this direction may not have a physical interpretation. Lower-order components will be forced to be orthogonal to the first component, and hence they may

Table 2. Summary description and validation of the DJF (a) temperature, (b) temperature variability and (c) precipitation emulators. For each of the 10 components the table provides the percentage of simulator variance explained (Sect. 4.1), the number of linear and quadratic terms in the component score emulator (Sect. 4.2), the R^2 of the component score emulator fit (Sect. 4.2), the R^2 of the component score emulators under cross-validation strategies A and B (Sect. 4.3.1) and the component score distributions (mean and standard deviation) for the simulated and emulated ensembles (Sect. 4.3.1).

(a) DJF SAT

Cmpnnt	%age variance	Number of linear terms	Number of quad/cross terms	Emulator fitted R^2	Cross-validated R^2 (A)	Cross-validated R^2 (B)	Simulated distribution	Emulated (cross-val) distribution
1	75.1 %	17	10	96 %	95 %	95 %	0.036 ± 0.020	0.037 ± 0.021
2	4.1 %	12	10	79 %	76 %	72 %	-0.003 ± 0.046	-0.000 ± 0.039
3	2.2 %	14	9	30 %	10 %	17 %	0.001 ± 0.042	-0.003 ± 0.024
4	1.8 %	17	10	62 %	51 %	41 %	-0.007 ± 0.044	-0.005 ± 0.034
5	1.1 %	16	10	59 %	45 %	43 %	-0.010 ± 0.039	-0.012 ± 0.032
6	0.8 %	12	8	42 %	21 %	18 %	-0.004 ± 0.038	-0.003 ± 0.031
7	0.7 %	15	10	45 %	32 %	35 %	0.003 ± 0.042	0.005 ± 0.028
8	0.6 %	21	10	41 %	19 %	23 %	-0.004 ± 0.041	-0.002 ± 0.030
9	0.6 %	12	9	30 %	18 %	6 %	0.001 ± 0.042	-0.001 ± 0.026
10	0.5 %	15	9	36 %	9 %	14 %	-0.002 ± 0.039	-0.002 ± 0.031

(b) DJF SAT variability (standard deviation)

Cmpnnt	%age variance	Number of linear terms	Number of quad/cross terms	Emulator fitted R^2	Cross-validated R^2 (A)	Cross-validated R^2 (B)	Simulated distribution	Emulated (cross-val) distribution
1	25.4 %	16	10	86 %	81 %	83 %	-0.037 ± 0.020	-0.037 ± 0.019
2	4.5 %	12	10	40 %	15 %	19 %	-0.003 ± 0.044	0.001 ± 0.029
3	3.8 %	15	9	35 %	16 %	34 %	-0.004 ± 0.043	-0.006 ± 0.025
4	2.8 %	15	10	31 %	6 %	13 %	0.000 ± 0.043	0.001 ± 0.027
5	2.3 %	17	9	33 %	10 %	8 %	-0.002 ± 0.038	-0.004 ± 0.030
6	1.9 %	14	6	33 %	19 %	18 %	-0.005 ± 0.040	-0.008 ± 0.028
7	1.8 %	16	10	24 %	6 %	0 %	0.003 ± 0.043	0.002 ± 0.025
8	1.6 %	14	10	44 %	33 %	16 %	-0.003 ± 0.045	0.000 ± 0.027
9	1.5 %	16	8	33 %	22 %	9 %	-0.000 ± 0.042	-0.002 ± 0.026
10	1.4 %	15	9	40 %	10 %	9 %	-0.005 ± 0.041	-0.004 ± 0.031

(c) DJF precipitation

Cmpnnt	%age variance	Number of linear terms	Number of quad/cross terms	Emulator fitted R^2	Cross-validated R^2 (A)	Cross-validated R^2 (B)	Simulated distribution	Emulated (cross-val) distribution
1	16.8 %	16	10	93 %	92 %	90 %	0.036 ± 0.021	0.036 ± 0.021
2	6.5 %	13	10	70 %	60 %	48 %	-0.005 ± 0.042	-0.006 ± 0.034
3	6.0 %	11	9	74 %	65 %	67 %	0.005 ± 0.042	0.001 ± 0.037
4	3.9 %	12	10	77 %	74 %	65 %	0.001 ± 0.042	0.001 ± 0.027
5	3.3 %	13	10	69 %	56 %	42 %	0.004 ± 0.040	0.006 ± 0.036
6	2.5 %	12	10	50 %	42 %	29 %	0.001 ± 0.043	0.003 ± 0.029
7	2.3 %	18	10	51 %	26 %	21 %	0.005 ± 0.042	0.002 ± 0.029
8	1.9 %	15	10	35 %	16 %	10 %	-0.004 ± 0.039	0.002 ± 0.024
9	1.7 %	13	10	49 %	43 %	22 %	-0.001 ± 0.042	-0.002 ± 0.033
10	1.6 %	14	10	62 %	56 %	25 %	-0.001 ± 0.043	-0.001 ± 0.033

also have no sensible physical interpretation. However, in our case the decomposition of centred and non-centred data is equivalent. The correlation between the first components of the centred and non-centred data is ~ 0.99. We here prefer to decompose the non-centred data so that the emulation of the first component is analogous to conventional pattern scaling,

the ensemble mean-field scaled by an emulation of the global mean change. However, to analyse the variance contributed by each component, it is necessary to first centre the data; the percentage variance contributions presented in Table 2 are derived from a decomposition of the centred data. We note that when the decomposition (Eq. 5) is performed upon

the centred matrix, the components (or left singular vectors) are equivalent to empirical orthogonal functions or principal components.

The physics of the climate system results in spatio-temporal correlations between ensemble members, patterns of change that are a function of the climate model itself rather than of parameter choices. For instance, warming is generally greater over land than it is over ocean, greater over deserts than over forested regions, greater over snow-covered regions than over snow-free regions. As a consequence of these spatial and temporal correlations between ensemble members, it is generally the case that a small subset of the 564 components is sufficient to describe most of the variability across the ensemble. We note that the approach of pattern scaling also utilises these correlations by assuming that a single pattern (equivalent to the first component) can be applied to approximately describe the pattern from any simulation.

In the emulators described here we retain the first 10 components. The contribution of each of these 10 components to the ensemble variance is summarised in Table 2 for the examples of DJF SAT, SAT variability (the standard deviation of daily temperature across the season) and precipitation. The spatial distribution of SAT is well approximated by a single component, describing 75 % of the variance. The ensemble variance of spatio-temporal SAT variability and precipitation are less well explained by the high-order components, and only 25 % and 17 % respectively of the variance in these data sets is explained by the first component. It is worth noting that although the restriction to 10 components limits the percentage of the ensemble variability that can be captured, the approach represents an advance on pattern scaling, as pattern scaling is equivalent to the inclusion of only the first component. It is also worth noting that those outputs that are less completely explained by the first 10 components are the same outputs that are likely to benefit most from going beyond the first-order pattern scaling approach.

4.2 Emulation of components scores

Each individual simulated field can thus be approximated as a linear combination of the first 10 components, scaled by their respective scores. Each set of scores thus consists of a vector of coefficients, representing the projection of each simulation onto the respective component. As each simulated field is a function of the input parameters, so are the coefficients that comprise the scores. So each component score can be viewed, and hence emulated, as a scalar function of the input parameters to the simulator.

For each output field, emulators of the first 10 components were derived as functions of the 22 model parameters and the 6 concentration profile coefficients. These emulators were built in R (R development core team, 2013), using the stepAIC function (Venables and Ripley, 2002). For each emulator, we built a linear model from all 28 parameters, and then allowed the stepwise addition of 10 quadratic and cross

terms. We then successively removed terms according to the Bayesian information criterion (BIC). The choice to allow up to 10 cross and quadratic terms was made for computational simplicity, as they are expensive to fit in a data set of this size. In exploratory analysis, 10 quadratic terms were found sufficient to substantially increase emulator performance while minimising the risk of over-fitting. It is worth noting that in 10 of the 30 emulators (Table 2), between one and four of the quadratic terms were removed during the BIC step.

A more rigorous procedure to determine the optimal number of terms is to progressively add terms and test the resulting model under cross-validation until it ceases to improve significantly. Although a complete treatment would be computationally demanding, DJF SAT and precipitation emulators were constructed allowing the addition of up to 20 terms followed by BIC shrinkage. In both cases the percentage of variance explained by the emulators (V_T in Sect. 4.3.2) was unchanged to the stated significance. Allowing the addition of 100 terms was found to lead to deterioration in both models, demonstrating that the BIC criterion is insufficient to avoid over-fitting. We note that improvements in the emulated variance field (V_v in Sect. 4.3.2) are seen in the over-fitted models, likely because these additional terms (which are not robust under cross-validation) nevertheless contribute to emulator output variance.

The emulator fitted R^2 values are provided in Table 2. The first component S1 emulators all provide a good fit to the simulator ($R^2 \sim 90\%$). This has been found to be the case for all model outputs considered to date. The S1 emulator is equivalent to an emulator of the global average change, scaling the first component to generate an emulation of the spatial distribution, so that high performance of the S1 emulators is thus not surprising. Lower-order component scores are generally harder to emulate, presumably because they reflect physical processes that are more difficult to represent as simple functions of the input parameters and may contain increasing elements of internal variability. The emulator fits generally decrease from $R^2 \sim 90\%$ (S1) down to $\sim 30\%$ (S10).

4.3 Cross-validation

We cross-validate the emulators with two different approaches, quantifying the degree to which the emulator can reproduce the component scores (in Sect. 4.3.1) and the simulated fields (in Sect. 4.3.2). These alternative approaches provide different insights into the strengths and weaknesses of the emulators.

4.3.1 Cross-validating the components scores

A robust measure of emulator performance requires cross-validated statistics. In the first cross-validation approach the emulators are built from the component scores of a subset

of simulations and are applied to estimate the scores for the remaining simulations. Two strategies are applied:

1. The first approach (A) builds the emulators using simulations with all 188 parameter sets, taking two simulations for each (with different Chebyshev coefficients) and cross-validating against the remaining 188 simulations (i.e. in input parameter space that has been sampled by training simulations)

2. The second approach (B) uses only 138 parameter sets (with three simulations for each) and cross-validates against the remaining 150 simulations (50 parameter sets, three simulations for each), i.e. in plausible input parameter space that has not been sampled by training simulations.

The first strategy A mirrors the mode of operation in which the emulator is used in practice, as we never need to apply the emulator to parameterisations other that those of the 188 MPSF ensemble. It tests the future evolution of the known parameter sets. The second strategy B tests the emulator error for plausible inputs in general.

The cross-validated R^2 values between emulated and actual scores for DJF SAT, SAT variability and precipitation are provided in Table 2. The performances of the S1 emulators are not substantially degraded under cross-validation. However, some lower-order emulators, especially those for SAT variability, can perform very poorly under cross-validation, demonstrating that the emulators cannot be assumed to provide a good approximation of the contribution of low-order variability to an individual simulation. An indicative measure of overall emulator performance (how much of the simulated variance we can expect to emulate) is suggested by

$$P \sim \sum_{C=1}^{N} P_C R_C,$$

where P_C is the percentage of the variance explained by component C (Table 2, column 2), R_C is the cross-validated R^2 of the respective score emulator (Table 2, columns 6 and 7) and N is the number of components included (here $N = 10$). Under strategy A, these metrics are 77, 24 and 32 % for SAT, SAT variability and precipitation respectively. Under strategy B they are 77, 25 and 28 %. These estimates are consistent with the analysis in the following Sect. 4.3.2 (cross-validation test a).

It was demonstrated in Holden and Edwards (2010) that under leave-one-out cross-validation, the emulated ensembles of the component scores have similar mean and standard deviation to the simulated ensemble distributions. As the emulated fields are described by linear combinations of the components, the necessary condition for the emulated ensemble to provide a good approximation to the simulated ensemble is that the ensemble distributions of the scores are well reproduced (and that a sufficient number of components

are included to adequately describe the ensemble variability of the underlying fields). We take this approach here, cross-validating the statistics of the emulated ensemble of component scores. All three emulators provide good estimates of the simulated ensemble distributions (Table 2, columns 8 and 9), reproducing the simulated ensemble means very well, but underestimating the simulated ensemble standard deviations by typically ~ 20–30 %.

These data together suggest that the emulated ensemble mean should be a good approximation to the simulated ensemble mean, that the emulated ensemble will understate the variance of the simulated ensemble and finally that individual emulations cannot, in general, be assumed to be accurate. We investigate these conclusions in detail in the following section.

4.3.2 Cross-validating the spatial fields

A more complete cross-validation was performed to investigate the degree to which the fields themselves are well emulated. SVD is applied to the 376 training simulations described in cross-validation strategy A (Sect. 4.3.1) and the component scores are emulated. However, in this validation we extend the analysis by reconstructing the emulated fields for the 188 left-out simulations, i.e. by scaling the 10 components by their emulated scores and summing the scaled components.

We focus upon the emulated fields at 2100 for this comparison as this time slice is expected, in general, to exhibit the greatest climate change. We also consider cross-validations at other time slices in test c.

Four error metrics are applied. These are designed to test the accuracy of: test a, the individual emulations; test b, the spatial pattern of the emulated mean field; test c, the magnitude of global average change; and test d, the spatial pattern of the emulated variance field.

We apply these four tests to six emulators, considering SAT, SAT variability and precipitation in both DJF and JJA. The behaviour of these emulators under cross-validation is shown to be quite different. Furthermore, we apply the tests as components are progressively added to the model in order to investigate the contribution of different components to the predictive power of the emulators. Finally, we apply the tests to emulators built from both centred data and uncentred data; the cross-validated performance is shown to be only weakly sensitive to this choice.

The four tests are as follows:

a. The proportion of the total ensemble variance captured by the emulator:

$$V_T = 1 - \sum_{n=1}^{188}\sum_{k=1}^{2048} \left(S_{n,k} - E_{n,k} \right)^2 \Big/ \sum_{n=1}^{188}\sum_{k=1}^{2048} \left(S_{n,k} - \bar{S}_k \right)^2,$$

where $S_{n,k}$ is the simulated output (2100) at grid cell k in left-out ensemble member n, $E_{n,k}$ the emulated
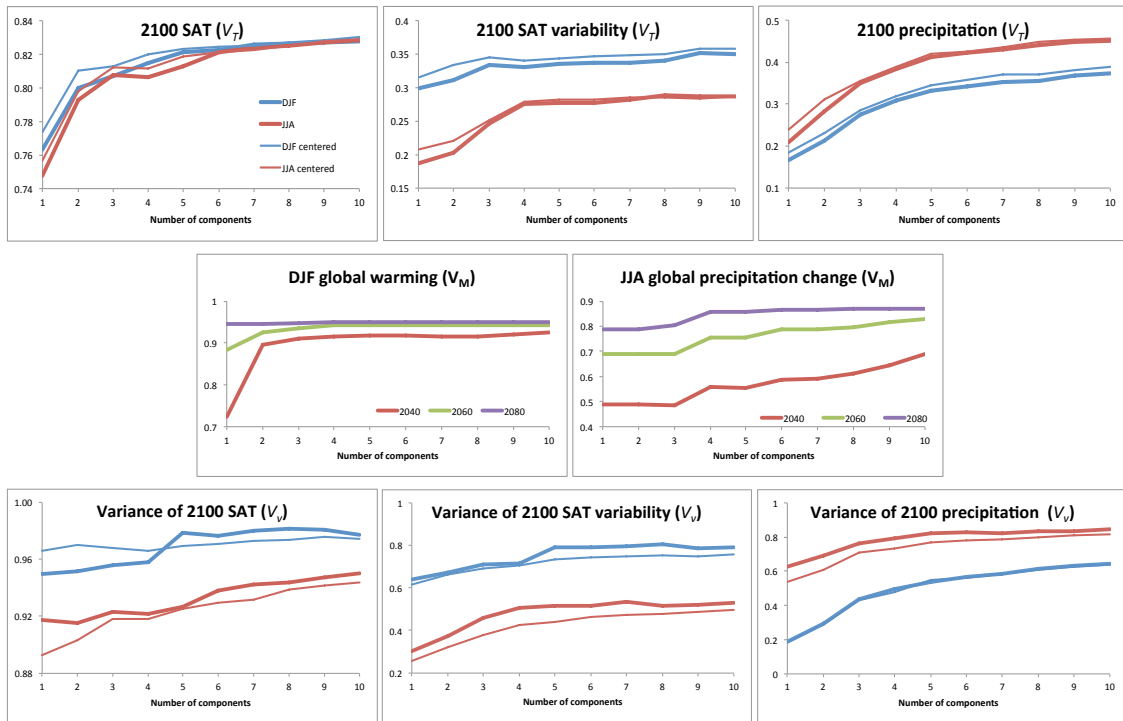
Fig. 3. Emulator errors as components are progressively added (Sect. 4.2). Top panels: V_T, proportion of the total ensemble variance captured by the emulators. Middle panels: V_M, proportion of variance in the spatial mean of the simulated ensemble captured by the emulators. Bottom panels: V_v, proportion of variance in the spatial pattern of the simulated ensemble variance that is captured by the emulators.

output and $\bar{S}_k$ the ensemble mean simulated output at grid cell k. V_T assesses the emulated error for each simulation, averaged over the 188 simulations, and measures the degree to which individual emulations can be regarded as accurate.

Figure 3 (top panel) illustrates the change in V_T as successive components are added to the emulators. The emulation of SAT is relatively accurate; even when only a single component is considered, the model reproduces $\sim 76\%$ of the simulated variance (over all space and all ensemble members) in this time slice. This reflects the high percentage of variance explained by the first component ($\sim 75\%$, over all times) and the high degree of accuracy of the S1 SAT emulator (cross-validated $R^2 \sim 95\%$). Adding additional components improves the model marginally. The emulator explains $\sim 83\%$ of simulated variance when all 10 components are included.

The emulation of SAT variability is much poorer, explaining only $\sim 25\%$ of the ensemble variance when one component is included, and exhibiting little improvement (on this V_T measure) as additional components are added. This reflects the low percentage of variance explained by the first component ($\sim 25\%$) and the difficulty in emulating the scores of

all besides the first component (cross-validated R^2 typically $\sim 10\%$).

The emulation of precipitation is intermediate between these cases. As with SAT variability, but in contrast to SAT, many components are required to explain the ensemble variance. However, in contrast to SAT variability, it is possible to emulate lower-order component scores. Thus, while the inclusion of only one component explains just $\sim 20\%$ of the V_T ensemble variance at 2100 (cf. 17% of variance explained by the first component, over all times), substantial improvements are apparent as more components are added. However, only $\sim 40\%$ of the ensemble variance is explained with 10 components, demonstrating that individual emulations cannot be regarded as accurate.

b. The proportion of variance in the spatial pattern of the simulated ensemble mean captured by the emulator:

$$V_m = 1 - \sum_{k=1}^{2048} \left(\bar{S}_k - \bar{E}_k \right)^2 \bigg/ \sum_{k=1}^{2048} \left(\bar{S}_k - \bar{S} \right)^2,$$

where $\bar{E}_k$ is the ensemble mean emulated output and $\bar{S}$ the ensemble mean spatially averaged simulated output. This measures the degree to which the emulated ensemble-mean field is accurate.

This metric is not graphically illustrated. In every emulator $V_m \approx 99\,\%$, irrespective of the number of components included in the model. This is not surprising as, because we do not centre the data before decomposition, the first component is approximately equivalent to the ensemble mean. If all 10 emulators have unbiased errors (i.e. symmetrically distributed about zero), then the emulated ensemble mean will reproduce the simulated ensemble mean. To a very close approximation this is the case, and *the emulated mean pattern is a highly accurate prediction for all outputs.*

However, some caution should be applied in the interpretation of this metric because of the ensemble design. The second and third forcing coefficients are uniformly distributed about zero and so they are not likely to be represented in the ensemble mean field, so this metric does not demonstrate the degree to which the magnitude of the of change is well described. This is accounted for in the following metric, V_M.

c. The proportion of variance in the spatial mean of the simulated ensemble captured by the emulator:

$$V_M = 1 - \sum_{n=1}^{188} \left(\bar{S}_n - \bar{E}_n\right)^2 \Big/ \sum_{n=1}^{188} \left(\bar{S}_n - \bar{S}\right)^2.$$

In Fig. 3 (central panel) we plot V_M for DJF SAT and JJA precipitation at 2040, 2060 and 2080.

When all 10 components are included, the emulators explains $\sim 95\,\%$ of the variance in globally averaged warming and up to 85 % of the variance in globally averaged precipitation change.

At early times the first component is insufficient to accurately emulate the mean change in either temperature or precipitation. This is unsurprising as the first component is dominated by the linear ramping of forcing (the first forcing coefficient), whereas early times are also expected to be strongly dependent upon the shape of the concentration profile, controlled by the low-order forcing coefficients. The temperature and precipitation emulators are both able to explain more of the ensemble variance of the global average change at later times. This is likely a consequence of the fact that the decomposition is designed to maximise the variance explained, and so is better at describing portions of the data with a high contribution to the overall variance (i.e. here at later times). We note that this effect may also explain the different skills of the emulator in different seasons that are apparent in V_T (Fig. 3, top panels). The best emulation of SAT is during DJF; the dominant uncertainty in the spatial distribution of SAT arises in DJF due to the uncertain response of Arctic sea ice. The best emulation of precipitation is in JJA; the dominant uncertainty in the

spatial distribution of precipitation arises in JJA due to the uncertain response of the SE Asian monsoon. This is a useful consequence of the decomposition approach (maximising variance explained), as it suggests that regions of space and time that experience the greatest change (and hence the greatest impact), and fields that exhibit the greatest variability, are those which are best described by the emulator.

A corollary of this is that for early decades V_M can take negative values. For instance, for the emulation of DJF SAT in the first decade $V_M = -0.02\,\%$ (cf. 0.94 % for the 10th decade). This demonstrates that the code uncertainty dominates over simulation uncertainty when the climate change signal is small. The emulation is nevertheless useful at these early times; the global means of the first decade SAT emulations exhibit a cross-validated correlation of 0.30 with respect to the simulations (cf. 0.97 for the 10th decade). However, this analysis demonstrates that uncertainty is overstated for modest climate change.

d. The proportion of variance in the spatial pattern of the simulated ensemble variance that is captured by the emulator:

$$V_v = 1 - \sum_{k=1}^{2048} \left(\mathrm{var}\,(S_k) - \mathrm{var}(E_k)\right)^2 \Big/$$
$$\sum_{k=1}^{2048} \left(\mathrm{var}\,(S_k) - \bar{\mathrm{var}}(\mathrm{S}_k)\right)^2.$$

This measures the degree to which the emulated ensemble-variance field is accurate. We have already demonstrated (with V_m and V_M together) that the emulated mean field is very accurate. If the emulated variability is also accurate, then together the emulated mean and variance fields provide a meaningful estimate of the spatio-temporal mean and variance (especially useful in those cases when individual emulations are not reliable, as indicated by V_T).

Unsurprisingly, the spatial uncertainty of SAT is very well described by the emulators. This is expected given the high performance of individual emulations for this variable. For both SAT variability and precipitation, the emulated variances with 10 components are satisfactory. Clear improvements in the description of variability are evidenced as additional components are added.

4.3.3 Cross-validation summary

In summary, the emulated mean fields of change are very accurate. Emulations of SAT variability and precipitation underestimate the simulated ensemble variability (although at early times the converse is true, as the emulator error dominates over simulated uncertainty for small climate change).
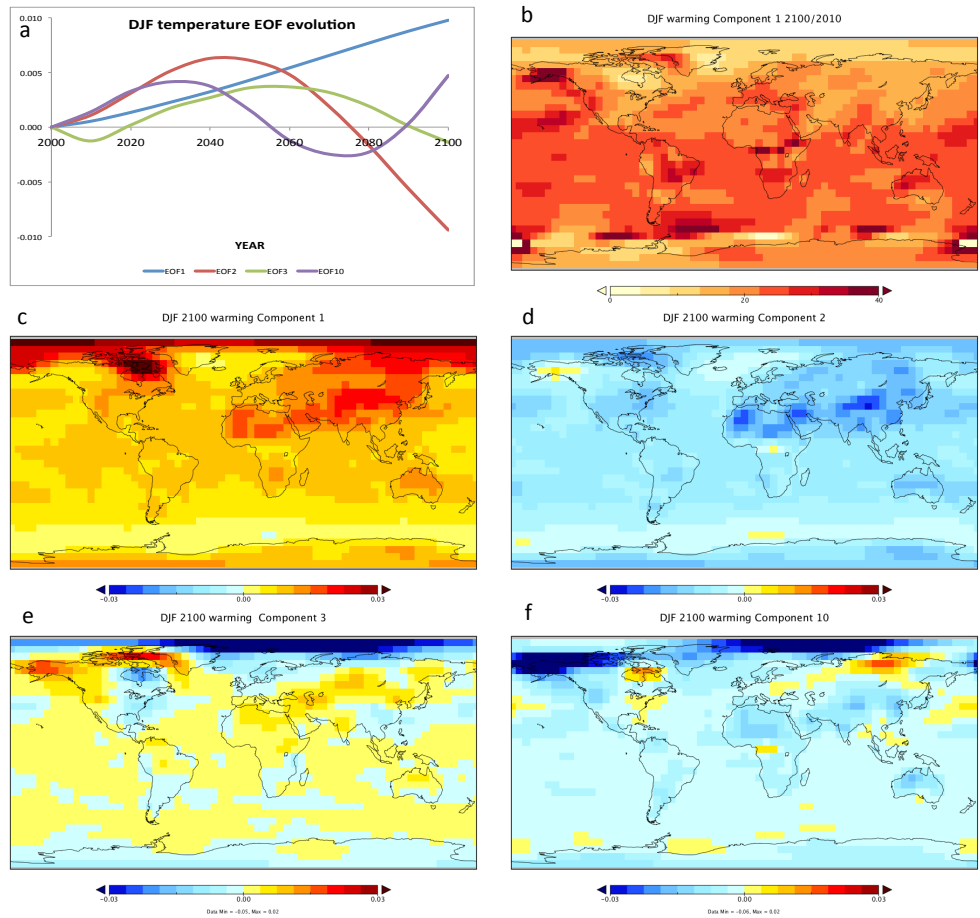
Fig. 4. DJF warming components. **(a)** Temporal evolution of selected spatially averaged components. **(b)** The ratio of the 2100 to 2010 spatial fields of the first component. **(c–f)** Spatial fields of selected components at 2100.

Accuracy cannot be assumed for individual emulations besides SAT. However, we note, given the high cross-validated R^2 of the first component score emulators for all variables, that individual simulations are likely to be at least as well predicted as they would be under pattern scaling. This is especially the case for variables such as precipitation, where low-order components contribute significantly to the variance and can be emulated; i.e. under pattern scaling low-order components are neglected and errors comparable to a single component emulator (see Fig. 3) would be expected.

4.4 Spatio-temporal variability

Each component represents a mode of variability of the ensemble across time, space and between ensemble members. Therefore each component, like each simulation ensemble member, is a 3-D spatio-temporal field that can be averaged over space to give a function only of time. Figure 4a plots the temporal evolution of an illustrative subset of the components of DJF temperature. The first component describes an approximately linear temperature ramp. This describes

the dominant mode of variability across the ensemble and so is expected to display an approximately linear increase over time because the second and third Chebyshev coefficients, which describe deviations from a linear ramp in CO_2 concentration, are both centred on zero in the ensemble design. Unsurprisingly, the S1 emulator is dominated by the first Chebyshev coefficient TC1E, which defines the slope of the linear forcing ramp. Higher-order components are thus required to describe more complex temporal behaviour. The S2 emulator is controlled by all three Chebyshev coefficients, the S3 emulator is mainly controlled by TC2E and TC3E, and the S10 emulator mainly by TC3E.

Figure 4b is a plot of the ratio of the first component at 2100 and 2010 time slices. This figure demonstrates that even when a single component is considered, the spatio-temporal emulation approach captures appreciable temporal evolution of the spatial pattern. We note that a pattern scaling approach would apply the same pattern at all times (i.e. equivalent to a uniform value everywhere in Fig. 4b).

The application of a single spatio-temporal decomposition allows us, in theory at least, to capture abrupt changes in

magnitude (i.e. the temporal evolutions plotted Fig. 4a are not constrained to be a smooth function of time) or in spatial distribution (i.e. the patterns of change are not constrained to be similar in time, as illustrated in Fig. 4b).

Figure 4c–f plot the spatial patterns at 2100 for each of the four components illustrated in Fig. 4a. The second source of non-linear spatio-temporal variability thus arises from the inclusion of the 10 components, which each exhibit distinct warming patterns. We do not attempt to ascribe physical meaning to these patterns. It is well known that caution is required in interpreting components as physically based, although such an approach can be useful (Holden et al., 2013b). It is however worth noting that the first component represents the ensemble averaged warming response, a pattern that is well known and robust across CMIP climate models (cf. Fig. 10.9, Meehl et al., 2007).

5 Baseline heating and cooling degree days

In this section we describe the derivation and validation of baseline (1 January 1995 to 1 January 2005) heating degree days (HDDs) and cooling degree days (CDDs), calculated on the PLASIM-ENTS grid and mapped onto the regions of the TIAM-WORLD model (TIMES integrated assessment model; Loulou and Labriet, 2008). HDDs provide a measure that reflects heating energy demands, calculated relative to some baseline temperature. On a given day, the average temperature is calculated and subtracted from the baseline temperature. If the value is less than or equal to zero, then that day has zero HDDs (no heating requirements). If the value is positive, then that number represents the number of HDDs on that day. The sum of HDDs over a month provides an indication of the total heating requirements for that month. CDDs are directly analogous, but integrate the temperature excess relative to a baseline and provide a measure of the cooling demands for that month. An evaluation of the modern-day distribution of HDDs and CDDs therefore provides a useful validation of the baseline climate simulations, reflecting both spatial and seasonal variability, and furthermore provides a validation of the transformation of the emulated outputs into degree-day data and of the population-weighted mapping of this degree-day data onto the regional level for impacts evaluation. The validation of the emulator itself will be addressed in Sect. 6.

We do not simulate (or emulate) degree days directly but instead derive them from average seasonal temperature and daily variability, defined by the standard deviation of the daily temperature across the season, following the approach of Schoenau and Kehrig (1990). The critical assumption made is that daily temperatures are scattered about the seasonal mean with a normal distribution. Direct calculation of degree-day data from daily variability would be more accurate, but was judged to be overly restrictive as it would prevent recalibration with an altered baseline temperature.
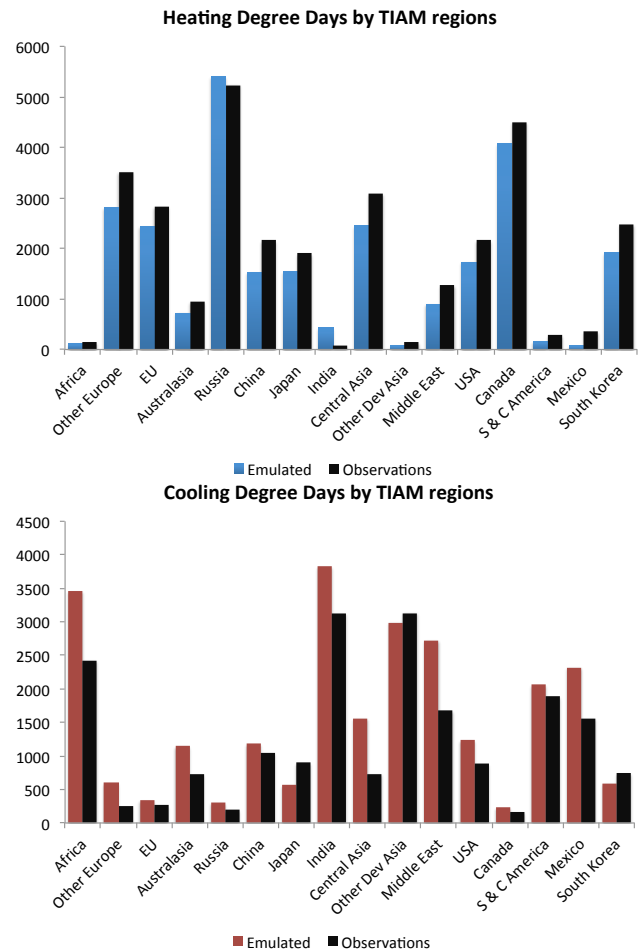

Fig. 5. Baseline heating and cooling degree days. These are derived on the PLASIM-ENTS grid and mapped with population weighting on to the TIAM-WORLD regions (Labriet et al., 2013). Calculated data (coloured bars) are compared against the empirical estimates (black bars) of Baumert and Selman (2003).

Different impact measures may apply different baselines, and baselines may be required to change over time, or be defined differently from region to region.

We calculate HDDs and CDDs at each PLASIM-ENTS grid cell from

$$\text{HDD} = \frac{N}{\sigma\sqrt{2\pi}} \int_{-\infty}^{B_\text{H}} (B_\text{H} - T)\, e^{\left[-(T-\mu)^2/2\sigma^2\right]} \text{d}T \qquad (6)$$

$$\text{CDD} = \frac{N}{\sigma\sqrt{2\pi}} \int_{B_\text{C}}^{\infty} (T - B_\text{C})\, e^{\left[-(T-\mu)^2/2\sigma^2\right]} \text{d}T, \qquad (7)$$

expressed in terms of the number of days in the season N, daily temperature T, HDD baseline temperature B_H, CDD baseline temperature B_C, average daily temperature across
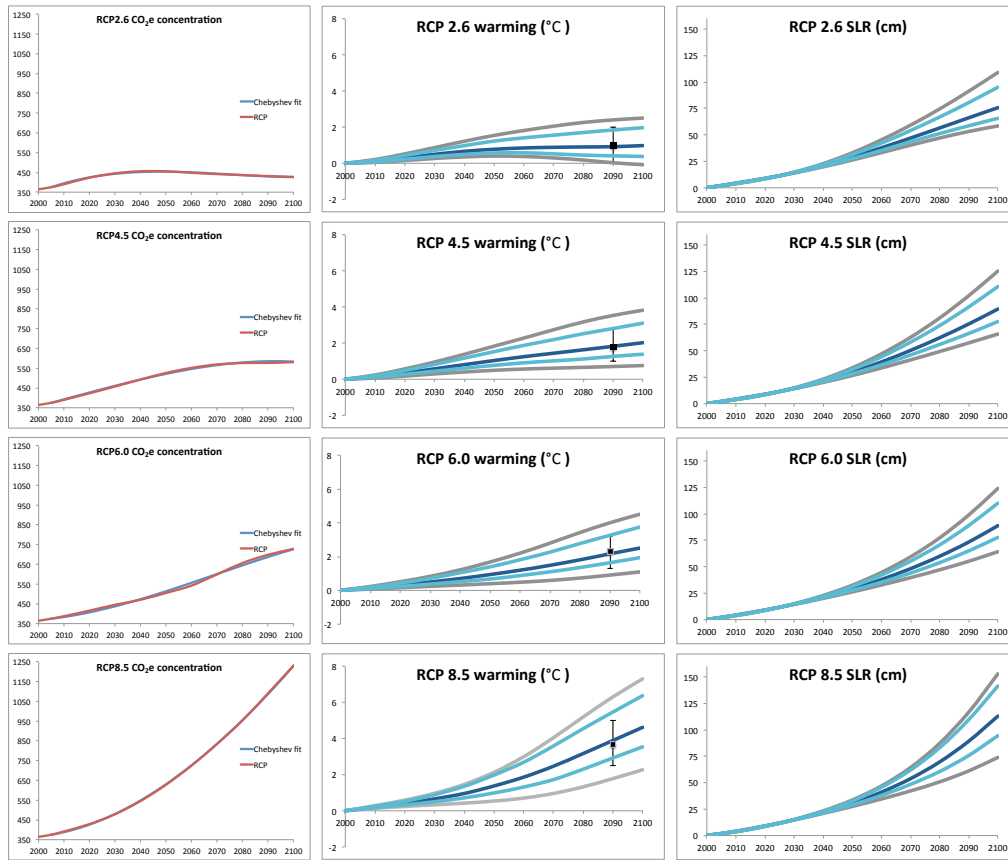
Fig. 6. RCP time series validations. (Left) RCP radiative forcing profiles and Chebyshev polynomial fits. (Centre) Emulated ensemble median warming (dark blue), 5 % and 95 % percentiles (light blue), ensemble end members (grey), compared with CMIP5 (1995 to 2090) projections (Collins et al., 2014) illustrated as the ensemble mean (blue squares) and ensemble range (error bars). (Right) Sea-level projections (median, 5th and 95th percentiles, ensemble end members) derived by applying the emulated warming projections to the empirical form of Rahmstorf et al. (2012).

the season μ and standard deviation of daily temperature across the season σ. The reference temperatures (B_H and B_C) are variable inputs. For the following analysis, $B_H = B_C = 18\,°C$ is applied globally, although it is a straightforward modification to allow the baselines to vary in space or over time.

For input to regionally integrated energy usage calculations in TIAM-WORLD, we derive a population-weighted average over the grid cells that comprise a given region. We apply 2005 population data (CIESIN and CIAT, 2005) at a 0.25° resolution which we integrate up onto the PLASIM-ENTS grid. We note that the low resolution of the climate model inevitably leads to approximations, most notably when highly populated regions near ocean are located in grid cells that are assigned to be ocean in PLASIM-ENTS. We address this here by assigning all grid cells that have a population greater than 500 000 to a TIAM-WORLD region, irrespective of whether or not that cell is assigned to be land or ocean in PLASIM-ENTS. This avoids the potential neglect of densely populated coastal regions, but comes at the expense

of ascribing an oceanic climate to some populated regions, likely understating seasonal variability and future warming.

The seasonally resolved HDDs and CDDs are summed to generate annual data and are compared to observations (Baumert and Selman, 2003) in Fig. 5. PLASIM-ENTS reproduces observational data remarkably well, capturing regional differences and with magnitudes that are generally quite reasonable. The emulator exhibits a warm bias, generally underestimating HDDs and overestimating CDDs. An element of this bias is likely due to recent warming as data sources for the observations are based on long-term averages that in the case of the United States, for instance, includes data that extend back to 1920.

6 Emulating the climate forced by the representative concentration pathways

In order to evaluate the emulator, we consider the warming response to the forcing of the representative concentration pathways (RCPs; Moss et al., 2010). The four RCPs are
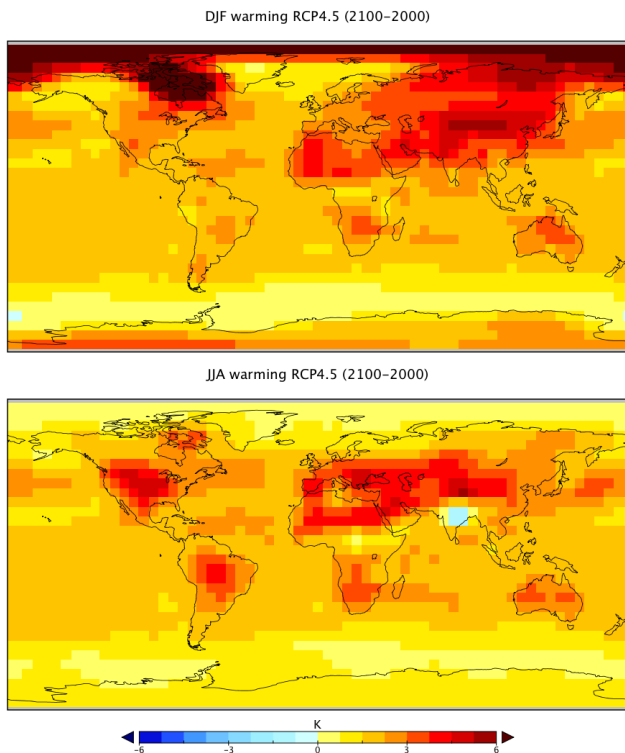
Fig. 7. Emulated ensemble mean warming (2100–2000) in response to RCP4.5 forcing: December-January-February (upper panel) and June-July-August (lower panel).

consistent sets of projections of the components of future radiative forcing, including scenarios of land-use change, aerosol and greenhouse-gas concentrations, designed to serve as inputs for climate models. We cannot force the emulator with the precise RCP temporal profiles, but instead derive Chebyshev fitted pathways to each. These are illustrated in the first column of Fig. 6. In this validation we ascribe the same coefficients to both CO_{2e} and CO_2; CO_2 is only an input to the vegetation in PLASIM-ENTS and hence is of limited importance for temperature. These coefficients are then applied to the emulators of seasonal temperature to generate an ensemble of 188 warming fields, differing through their PLASIM-ENTS parameterisations.

The spatial patterns of emulated ensemble averaged DJF and June-July-August (JJA) warming over the future transient period (2000 to 2100) are plotted in Fig. 7. These compare favourably to the patterns exhibited by CMIP3 AOGCMs (cf. Fig. 10.9, Meehl et al., 2007). The largest differences are apparent in JJA warming. The Southern Ocean JJA warming is weaker than in the CMIP3 ensemble, likely a consequence of the fixed Antarctic sea ice in the training ensembles. South-east Asian JJA warming is also weaker in the emulator than in CMIP3 simulations, in fact displaying a cooling of up to ~ 1.4 K under RCP4.5. This arises due to a strengthening of the south-east Asian monsoon in PLASIM-ENTS that is associated with decreased incoming

shortwave radiation (increased cloud cover) and increased evaporative cooling. Given the neglect of aerosol forcing in PLASIM-ENTS, this JJA cooling in south-east Asia should not be regarded as robust; aerosols are an important forcing of the south-east Asian monsoon through a range of likely competing effects (see e.g. Ganguly et al., 2012). We note that PLASIM-ENTSem spatial fields of precipitation change are provided in Foley et al. (2014).

The temporal development of warming for each RCP is plotted in the second column of Fig. 6. In all scenarios, the median ensemble warming compares favourably with the CMIP5 ensemble (Table 12.2, Collins et al., 2014). The emulated uncertainty is represented by the 5th and 95th confidence intervals of the emulated ensemble, and is compared against the multi-model ranges of Collins et al. (2014). The emulator captures the CMIP5 ranges well. The full range of the emulated ensemble is also plotted to illustrate the emulated extremes.

A final illustration is provided in the third column of Fig. 6. These plots illustrate the sea-level rise and associated uncertainty predicted for each RCP. This approach is currently being applied to address sea-level impacts (Joshi et al., 2014) with the GEMINI integrated assessment model (Bernard and Vielle, 2008). The sea-level estimate is derived from the empirical form of Rahmstorf et al. (2012), which assumes that the rate of sea-level rise depends linearly on both warming and the rate of warming (Vermeer and Rahmstorf, 2009). We do not consider uncertainty in the empirical fit (estimated to be ~ 10 % for RCP4.5) but instead apply the "CW05" (Church and White, 2006) fit throughout. We note that the median emulated sea-level prediction for RCP4.5 (89 cm) is slightly lower than the Rahmstorf et al. (2012) estimate (~ 1 m), despite a slightly greater 2100–2000 warming (2.0 K compared to 1.8 K). This may reflect a somewhat greater thermal inertia in the PLASIM-ENTS ensemble, as also evidenced by the emulated warming under RCP 2.6, which continues to warm through the 21st century in the emulated ensemble median despite the decreasing radiative forcing after 2040.

7 Summary

Building on Holden and Edwards (2010), we have developed an emulator of the spatio-temporal climate response to an arbitrary 21st-century forcing scenario. We apply singular vector decomposition to decompose the modes of variability across a large ensemble of simulations of the intermediate-complexity GCM PLASIM-ENTS. We emulate the high-order components as simple polynomial functions of future forcing and model parameters that we apply to emulate fields of climate change in response to an arbitrary forcing profile. The approach represents an advance on pattern scaling as it allows us to address non-linear spatio-temporal feedbacks and uncertainty.

The approach does not directly quantify emulator error (or "code uncertainty"), but we have indirectly evaluated it through cross-validation. We have demonstrated that for some variables (notably surface air temperature, SAT) individual emulations can be regarded as an accurate approximation to the simulator, whereas for others (notably precipitation) it is only appropriate to regard the emulated ensemble mean and variance as useful predictors. In either case, and cognisant of appropriate limitations, the emulator provides meaningful estimates of the uncertainty associated with climate change projections that can be used to better inform impact assessments.

The emulator reproduces present-day regionally resolved heating and cooling days that are in good agreement with observations. It generates spatial patterns and temporal profiles of seasonally resolved warming and associated uncertainty that are in generally good agreement with the CMIP ensemble of atmosphere–ocean GCMs, making it an appropriate tool for impact assessment. Although not described here, we note that we have also applied the approach to emulate evaporation and fractional cloud cover for application to energy demands (hydroelectric potential) and crop impacts. These will be discussed in future work.

The motivation for this approach is computational speed. A 188-member ensemble of 100 yr PLASIM-ENTS simulations requires ~ 1 CPU year. A 188-member emulated ensemble requires ~ 1 CPU second per output variable. Developing the emulators is not computationally demanding, although extensive cross-validation may be, and the approach is therefore appropriate for application to high-resolution AOGCMs. In this case the limitation is more likely to be the computational expense of running the necessary ensemble of simulations, although hierarchical methods can help address this (Williamson et al., 2012). Most of the computational expense of building the PLASIM-ENTS emulators is fitting the functional forms of the 10 component emulators (~ 3 min when 10 quadratic terms are allowed). The computational demands of this fitting step are independent of the spatio-temporal resolution of the emulator. The decomposition of the $20\,480 \times 564$ matrix considered here requires ~ 40 s.

PLASIM-ENTSem has been applied in a range of IAM couplings. The spatially dependent outputs have been used to investigate regional warming impacts on energy demands for heating and cooling (Labriet et al., 2013). The ability to capture uncertainty has been utilised in a study of the uncertainties associated with the economic damages of sea-level rise (Joshi et al., 2014). In conjunction with an emulator of the carbon cycle (Foley et al., 2014), PLASIM-ENTSem has been applied to investigate uncertain global warming projections in the context of the decarbonisation of the energy sector (Mercure et al., 2014).

References

Annan, J. D., Lunt, D. J., Hargreaves, J. C., and Valdes, P. J.: Parameter estimation in an atmospheric GCM using the Ensemble Kalman Filter, Nonlin. Processes Geophys., 12, 363–371, doi:10.5194/npg-12-363-2005, 2005.

Baumert, K. and Selman, M.: Heating and cooling degree days, World Resource Institute, 2003.

Beaumont, M. A., Zhang, W., and Balding, D. J.: Approximate Bayesian computation in population genetics, Genetics, 162, 2025–2035, 2002.

Bernard, A. and Vielle, M.: GEMINI-E3, a general equilibrium model of international-national interactions between economy, energy and the environment, Computational Management Science, 5, 173–206, doi:10.1007/s10287-007-0047-y, 2008.

Bordi, I., Fraedrich, K., Sutera, A., and Zhu, X.: Transient response to well-mixed greenhouse gas changes, Theor. Appl. Climatol., 109, 245–252, doi:10.1007/s00704-011-0580-z, 2011a.

Bordi, I., Fraedrich, K., Sutera, A., and Zhu, X.: On the climate response to zero ozone, Theor. Appl. Climatol., 109, 253–259, doi:10.1007/s00704-011-0579-5, 2011b.

Bordi, I., Fraedrich, K., Sutera, A., and Zhu, X.: On the effect of decreasing CO_2 concentrations in the atmosphere, Clim. Dynam., 40, 651–662, doi:10.1007/s00382-012-1581-z, 2013.

Church, J. A. and White, N. J.: A 20th century acceleration in global sea-level rise, Geophys. Res. Lett., 33, L01602, doi:10.1029/2005GL024826, 2006.

CIESIN and CIAT (Center for International Earth Science Information Network and Columbia University; and Centro Internacional de Agricultura Tropical): Gridded Population of the World, Version 3 (GPWv3), Palisades, NY: Socioeconomic Data and Applications Center (SEDAC), Columbia University, available at: http://sedac.ciesin.columbia.edu/gpw (last access: 20 March 2012), 2005.

Collins, M., Knutti, R., Arblaster, J. M., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., Gao, X., Gutowski Jr. W. J., Johns, T., Krinner, G., Shongwe, M., Tebaldi, C., Weaver, A. J., and Wehner, M.: Long-term Climate Change: Projections, Commitments and Irreversibility. Climate Change 2013: The Physical Science Basis, Contribution of WG I to the Fifth Assessment Report of the IPCC, edited by: Stocker, T. and Qin, D., Cambridge University Press, Cambridge, UK and New York, USA, in preparation, 2014.

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

Dahms, E., Borth, H., Lunkeit, F., and Fraedrich, K.: ITCZ splitting and the influence of large-scale eddy fields on the Tropical mean state, J. Meteorol. Soc. Jpn., 89, 399–411, doi:10.2151/jmsj.2011-501, 2011.

Dekker, S. C., de Boer, H. J., Brovkin, V., Fraedrich, K., Wassen, M. J., and Rietkerk, M.: Biogeophysical feedbacks trigger

shifts in the modelled vegetation-atmosphere system at multiple scales, Biogeosciences, 7, 1237–1245, doi:10.5194/bg-7-1237-2010, 2010.

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

Foley, A. M., Mercure, J.-F., Salas, P., Holden, P. B., and Edwards, N. R.: Modelling the climate impacts of mitigation policies in the energy sector: An integrated assessment framework using PLASIM-ENTSem, Climatic Change, in review, 2014.

Fraedrich, K.: A suite of user-friendly global climate models: Hysteresis experiments, Eur. Phys. J. Plus, 127, 53, doi:10.1140/epjp/i2012-12053-7, 2012.

Fraedrich, K. and Lunkeit, F.: Diagnosing the entropy budget of a climate model, Tellus A, 60, 921–931, doi:10.1111/j.1600-0870.2008.00338.x, 2008.

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

Fraedrich, K., Kirk, E., Luksch, U., and Lunkeit, F.: The portable university model of the atmosphere (PUMA): Storm track dynamics and low-frequency variability, Meteorol. Z,, 14, 735–745, doi:10.1127/0941-2948/2005/0074, 2005b.

Ganguly, D., Rasch, P. J., Wang, H., and Yoon, J.-H.: Climate response of the South Asian monsoon system to anthropogenic aerosols, J. Geophys. Res., 117, D13209, doi:10.1029/2012JD017508, 2012.

Garreaud, R. D., Molina, A., and Farias, M.: Andean uplift, ocean cooling and Atacama hyperaridity: A climate modeling perspective, Earth Planet. Sc. Lett., 292, 39–50, doi:10.1016/j.epsl.2010.01.017, 2010.

Grosfeld, K., Lohmann, G., Rimbu, N., Fraedrich, K., and Lunkeit, F.: Atmospheric multidecadal variations in the North Atlantic realm: proxy data, observations, and atmospheric circulation model studies, Clim. Past, 3, 39–50, doi:10.5194/cp-3-39-2007, 2007.

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

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

Holden, P. B., Edwards, N. R., Gerten, D., and Schaphoff, S.: A model-based constraint on CO_2 fertilisation, Biogeosciences, 10, 339–355, doi:10.5194/bg-10-339-2013, 2013a.

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

Joshi, S. R., Vielle, M., Babonneau, F., Edwards, N. R., and Holden, P. B.: Physical and economic impacts of sea-level rise: A coupled GIS and CGE analysis under uncertainties, Environ. Res. Econ., in review, 2014.

Kleidon, A., Fraedrich, K., and Low, C.: Multiple steady-states in the terrestrial atmosphere-biosphere system: a result of a discrete vegetation classification?, Biogeosciences, 4, 707–714, doi:10.5194/bg-4-707-2007, 2007.

Labriet, M., Joshi, S. R., Babonneau, F., Edwards, N. R., Holden, P. B., Kanudia, A., Loulou, R., and Vielle, M.: Worldwide impacts of climate change on energy for heating and cooling, Mitig. Adapt. Strateg. Glob. Change, online first, doi:10.1007/s11027-013-9522-7, 2013.

Lenton, T. M., Aksenov, Y., Cox, S. J., Hargreaves, J. C., Marsh, R., Price, A. R., Lunt, D. J., Annan, J. D., Cooper-Chadwick, T., Edwards, N. R., Goswami, S., Livina, V. N., Valdes, P. J., Yool, A., Harris, P. P., Jiao, Z., Payne, A. J., Rutt, I. C., Shepherd, J. G. Williams, G., and Williamson, M. S.: Effects of atmospheric dynamics and ocean resolution on bi-stability of the thermohaline circulation examined using the Grid Enabled Integrated Earth system modelling (GENIE) framework, Clim. Dynam., 29, 591–613, doi:10.1007/s00382-007-0254-9, 2007.

Loulou, R. and Labriet, M.: ETSAP-TIAM: the TIMES integrated assessment model Part I: Model structure, Computational Management Science, 5, 7–40, doi:10.1007/s10287-007-0046-z, 2008/

Lucarini, V., Fraedrich, K., and Lunkeit, F.: Thermodynamic analysis of snowball earth hysteresis experiment: efficiency, entropy production, and irreversibility, Q. J. Roy. Meteorol. Soc., 136, 2–11, doi:10.1002/qj.543, 2010.

Mardia, K. V., Kent, J. T., and Bibby, J. M.: Multivariate analysis, Academic Press, London, 1979.

Matthews, H. D., Eby, M., Ewen, T., Friedlingstein, P., and Hawkins, B. J.: What determines the magnitude of carbon cycle-climate feedbacks? Glob. Biogeochem. Cy., 21, GB2012, doi:10.1029/2006GB002733, 2007.

Meehl, G. A., Stocker, T. F., Collins, W. D., Friedlingstein, P., Gaye, A. T., Gregory, J. M., Kitoh, A., Knutti, R., Murphy, J. M., Noda, A., Raper, S. C. B., Watterson, I. G., Weaver, A. J., and Zhao, Z.-C.: Global Climate Projections, in: Climate Change 2007: The Physical Science Basis. Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, edited by: Solomon, S., Qin, D., Manning, M., Chen, Z., Marquis, M., Averyt, K. B., Tignor, M., and Miller, H. L., Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2007.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., Frame, D., and Allen, M. R.: Greenhouse-gas emission targets for limiting global warming to 2 °C, Nature, 458, 1158–1162, doi:10.1038/nature08017, 2009.

Mercure, J. F., Salas, P., Foley, A., Chewpreecha, U., Holden, P. B., and Edwards, N. R.: The dynamics of technology diffusion and the impacts of climate policy instruments in the decarbonisation of the global electricity sector, Global Environ. Change, in review, 2014.

Micheels, A. and Montenari, M.: A snowball Earth versus a slush-ball Earth: Results from Neoproterozoic climate modeling sensitivity experiments, Geosphere, 4, 401–410, 2008.

Mitchell, J. F. B., Johns, T. C., Eagles, M., Ingram, W. J., and Davis, R. A.: Towards the construction of climate change scenarios, Climatic Change, 41, 547–581, doi:10.1023/A:1005466909820, 1999.

Moss, R. H., Edmonds, J. A., Hibbard, K. A., Manning, M. R., Rose, S. K., van Vuuren, D. P., Carter, T. R., Emore, S., Kainuma, M., Kram, T., Meehl, G. A., Mitchell, J. F. B., Nakicenovic, N., Riahi, K., Smith, S. J., Stouffer, R. J., Thompson, A. M., Weyant, J. P., and Wilbanks, T. J.: The next generation of scenarios for

climate change research and assessment, Nature, 463, 747–756, doi:10.1038/nature08823, 2010.

Noy-Meir, I.: Data transformations in ecological ordination I. Some advantages of non-centering, J. Ecol., 61, 329–241, 1973.

O'Hagan, A.: Bayesian analysis of computer code outputs: A tutorial, Reliability Engineering and System Safety, 91, 1290–1300, doi:10.1016/j.ress.2005.11.025, 2006.

Ramhstorf, S., Perrette, M., and Vermeer, M.: Testing the robustness of semi-empirical sea level projections, Clim. Dynam., 39, 861–875, doi:10.1007/s00382-011-1226-7, 2012.

R Development Core Team: R: A language and environment for statistical computing, R foundation for statistical computing, Vienna, Austria, available at: www.R-project.org (last access: 24 February 2014), 2013.

Roscher, M., Stordal, F., and Svenson, H.: The effect of global warming and global cooling on the distribution of the latest Permian climate zones, Palaeogeogr. Palaeocl., 309, 186–200, doi:10.1016/.j.palaeo.2011.05.042, 2011.

Rougier, J.: Probabilistic inference for future climate using an ensemble of climate model evaluations, Climatic Change, 81, 247–264, doi:10.1007/s10584-006-9156-9, 2007.

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

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

Schmittner, A., Silva, T. A. M., Fraedrich, K., Kirk, E., and Lunkeit, F.: Effects of mountains and ice sheets on global ocean circulation, J. Climate, 24, 2814–2829, doi:10.1175/2010JCLI3982.1, 2011.

Schoenau, G. J. and Kehrig, R. A.: A method for calculating degree-days to any base temperature, Energ. Buildings, 14, 299–302, doi:10.1016/0378-7788(90)90092-W, 1990.

Stenzel, O., Grieger, B., Keller, H. U., Greve, R., Fraedrich, K., and Lunkeit, F.: Coupling Planet Simulator Mars, a general circulation model of the Martian atmosphere, to the ice sheet model SICOPOLIS, Planet. Space Sci., 55, 2087–2096, doi:10.1016/j.pss.2007.09.001, 2007.

Venables, W. N. and Ripley, B. D.: Modern Applied Statistics With S, 4th Edn., Springer, New York, 2002.

Vermeer, M. and Rahmstorf, S.: Global sea level linked to global temperature, Proceedings of the National Academy of Sciences of the USA, 106, 21527–21532, doi:10.1073/pnas.0907765106, 2009.

Wilkinson, R. D.: Bayesian calibration of expensive multivariate computer experiments, in: Large scale inverse problems and the quantification of uncertainty, edited by: Biegler, L. T., Biros, G., Ghattas, O., Heinkenschloss, M., Keyes, D., Mallick, B. K., Tenorio, L., Van Bloemen Waanders, B., and Wilcox, K., Wiley Series in Computational Statistics, 2010.

Williamson, D., Goldstein, M., and Blaker, M.: Fast linked analysis for scenario-based hierarchies, Appl. Statist., 61, 665–691, doi:10.1111/j.1467-9876.2012.01042.x, 2012.

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