JAMES | Journal of Advances in Modeling Earth Systems

RESEARCH ARTICLE
10.1029/2019MS001787

Comparative Assessment of Climate Engineering Scenarios in the Presence of Parametric Uncertainty

Giang T. Tran[1], **Andreas Oschlies**[1], **and David P. Keller**[1]

[1]Marine Biogeochemical Modelling, GEOMAR Helmholtz-Zentrum für Ozeanforschung Kiel, Kiel, Germany

Key Points:
- We demonstrate how uncertainty analysis can become an essential part in comparative studies on climate engineering
- Model outcomes are probability distributions and should be treated as such when communicating the effects and impacts of climate engineering
- The probability of avoiding dangerous threshold is calculated for various climate engineering scenarios

Supporting Information:
- Supporting Information S1

Correspondence to:
G. T. Tran,
gtran@geomar.de

Citation:
Tran, G. T., Oschlies, A., & Keller, D. P. (2020). Comparative assessment of climate engineering scenarios in the presence of parametric uncertainty. *Journal of Advances in Modeling Earth Systems*, *12*, e2019MS001787. https://doi.org/10.1029/2019MS001787

Received 21 JUN 2019
Accepted 27 FEB 2020
Accepted article online 3 MAR 2020

Abstract

Climate engineering (CE) measures are increasingly discussed when dealing with the adverse impacts of climate change. While much research has focused on individual methods, few studies attempt to compare and rank the effectiveness of these measures. Furthermore, model uncertainties are seldom acknowledged and lesser still, estimated when CE scenarios are assessed. In this work, we quantify the variance in outcomes due to poorly constrained model parameters under several idealized CE scenarios. The four scenarios considered are (1) warming under the high emission scenario Representative Concentration Pathway 8.5 without CE applied and the same emission scenario with (2) afforestation, (3) solar radiation management, and (4) artificial ocean alkalinization. By considering the parametric uncertainty in model outputs, we demonstrate the problems with comparing these scenarios using a single parameter setting. Using statistical emulation, we estimate the probability distributions of several model outcomes. Based on such distributions, we suggest an approach to ranking the effectiveness of the scenarios considered according to their probability of avoiding climate thresholds.

Plain Language Summary

Various intervention techniques have been proposed to manipulate the climate system at a large scale to combat the adverse effects of climate change. While many studies have focused on specific techniques, relatively little has been done to compare and rank the effectiveness of these methods within a single model. Furthermore, the uncertainties arising from the use of climate models are seldom acknowledged or assessed. In this work, we analyze the uncertainty in the model's simulated outcomes caused by poorly constrained model settings under four idealized future scenarios, with and without climate interventions. Our results highlight the importance of taking into account the model's uncertainty when analyzing or communicating the simulated outcomes. Moreover, we suggest an approach to ranking the effectiveness of the interventions considered based on goals of societal importance.

1. Introduction

Dramatic reductions in anthropogenic greenhouse gas emissions are required to reduce the risk of dangerous climate change (Allen et al., 2009; IPCC, 2014; Meinshausen et al., 2009). However, the gap between the reductions needed and the national pledges made in the Paris Agreement is large (UNEP, 2017). On top of that, studies have suggested that even if the warming would stop once carbon dioxide emissions are halted (Gillett et al., 2011), elevated temperatures would persist for a long time (Eby et al., 2009; Matthews & Caldeira, 2008; Plattner et al., 2008). As a result, climate engineering (CE), or the deliberate manipulation of the climate system at a large scale, has been gaining attention as a possible means to limit some or all effects of anthropogenic climate change. CE measures can be classed as greenhouse gas removal methods, which act to actively remove greenhouse gases from the atmosphere, or as radiation management methods, which aim to reduce the amount of incoming solar radiation or increase the amount of outgoing longwave radiation. More detailed descriptions of available methods can be found in Vaughan and Lenton (2011), Caldeira et al. (2013), and National Research Council (2015).

If CE should ever become an option that is to be considered seriously, it is crucial to inform society and the decision makers on the potential impacts and side effects of CE measures along with those of other possible mitigation and adaptation scenarios. While many studies have focused on assessing individual CE methods, only a few have attempted to assess them comparatively. Niemeier et al. (2013) and Crook et al. (2015) looked at the responses of temperature and precipitation to different radiation management schemes while Keller et al. (2014) and Sonntag et al. (2018) compared a range of climate indicators under several CE methods, including solar radiation management (SRM) and land- and ocean-based carbon dioxide removal, using Earth system models (ESMs).

ESMs are simplified mathematical representations of the Earth climate system. Due to insufficient computational power and our lack of understanding of actual processes, approximations of the real world have to be made, and thus, models are inevitably imperfect. Model imperfection, together with internal climate variability and the uncertainties in forcing scenarios, means that model projections are likely to be systematically different from the true state of the climate system. Therefore, a model's projection of the impacts of climate change or CE should be accompanied by an assessment of the uncertainty contained in those predictions. As there are a multitude of societal issues competing for limited resources, likelihood statements based on rigorous probabilistic uncertainty assessment of models are crucial for the public and policymakers to make informed decisions on the most pressing matters (Reilly, 2001; Schellnhuber et al., 2006).

Model imperfection can be roughly divided into parametric uncertainty and structural uncertainty. A description of various types of uncertainty in computer models can be found in Kennedy and O'Hagan (2001). In this work, we focus on parametric uncertainty. A computer model, although a simplification of reality, can be very complex with hundreds of input parameters that represent underlying features of the system. In many cases, these parameters are poorly constrained physical constants or artifacts of the simplification of a complex physical process, referred to as parameterization. These free parameters are often estimated from “tuning” exercises; that is, uncertain parameters are varied and “best inputs” are identified based on how close the model outputs are to observation. Since both the models and observations contain errors, there is not a unique set of “best inputs” (Murphy et al., 2004; Stainforth et al., 2005). This lack of knowledge about the “true values” of the parameters propagates through the algorithm and results in uncertainty about the real value of the output.

The objective of this paper is to quantify the variance in the model's outputs as a result of uncertain input parameters in three representative CE scenarios and a no-CE scenario under the high CO_2 emissions Representative Concentration Pathway (RCP8.5) scenario using the University of Victoria Earth system climate model (UVic ESCM) of intermediate complexity. The CE measures were chosen with each focusing on one of the three major components of the Earth climate system: afforestation (AF) as a terrestrial method, SRM as an atmospheric method, and artificial ocean alkalinization (AOA) as an oceanic method. For conceptual simplicity, we restrict our analysis to globally and annually averaged properties. However, the methodology presented here could be extended to deal with higher-dimensional outputs such as time series (Mcneall, 2008; Wilkinson, 2010; Williamson et al., 2012) or high-dimensional spatial output fields (Holden & Edwards, 2010; Lee et al., 2012; Tran et al., 2019).

To quantify parametric uncertainty, we adopt a probabilistic framework utilizing a perturbed parameter ensemble (PPE) and Gaussian process (GP) emulation (O'Hagan, 2006; Santner et al., 2003). The analysis provides not only the mean and confident interval of the desirable outputs but also their estimated probability density functions for each CE scenario. Based on such results, we propose an approach to rank the effectiveness of potential CE scenarios. Throughout the work, we also highlight the importance of uncertainty quantification in comparing the impact of future projections and in communicating modeling results.

In the remainder of this paper, section 2 describes the experiment design, which includes a description of the UVic model, the PPE sampling plan, and the CE measures; section 3 introduces the statistical techniques used and the framework of our analysis; section 4 summarizes the model outputs; and section 5 discusses the results of the uncertainty analysis. Finally, we discuss our findings and provide suggestions for future improvements in section 6.

2. Experiment Design

2.1. UVic ESCM

The model used is the UVic ESCM Version 2.9 (Eby et al., 2013; Weaver et al., 2001), comprising a fully dynamic ocean circulation model (Pacanowski, 1996) coupled to an energy-moisture balance atmosphere (Fanning & Weaver, 1996), a dynamic-thermodynamic sea ice model (Bitz & Lipscomb, 1999), and a land surface and terrestrial vegetation model (Meissner et al., 2003). Ocean biogeochemistry is based on a simple nutrient-phytoplankton-zooplankton-detritus model (Keller et al., 2012; Schmittner et al., 2005). UVic has been extensively used in model intercomparison studies (Eby et al., 2013; Flato et al., 2013; Weaver et al., 2012; Zickfeld et al., 2013), showing comparable responses to other models under the same CO_2 emission forcing.

UVic has been used to investigate the potential of CE measures, individually (Matthews & Caldeira, 2007; Oschlies, Pahlow et al. 2010; Oschlies, Koeve et al. 2010) and comparatively (Keller et al., 2014). UVic is a suitable choice in this study since it includes a prognostic global carbon cycle, allowing us to quantify the carbon cycle response to CE. Being a model of intermediate complexity, UVic is efficient enough to produce PPEs to explore its high-dimensional parameter space. The trade-off is that UVic's atmospheric component is a simple energy-moisture balanced model that does not capture dynamic circulations and the model resolution is coarse (spherical grid resolution of 3.6° by 1.8° and 19 vertical levels in the ocean).

To conduct the uncertainty analysis, we first generated a perturbed ensemble of preindustrial climate. In this ensemble, each simulation was spun up for 5,000 yr under historical atmospheric and astronomical boundary conditions. From 1850 to 2000, the simulations follow historical fossil fuel and land use forcing. From the Year 2000 to 2020, all model runs were under forcings specified by the RCP8.5 scenario. Starting from 2020, the ensemble branches into four future ensembles, all remain under RCP8.5 forcing. The four future ensembles correspond to four scenarios, one without CE and three with CE applied. All simulations end in 2100.

In the AF scenario, soil moisture in North Africa and the Australian Outback are forced to have a constant value of 360 kg/m^2 to simulate irrigation, allowing vegetation to grow and thereby remove CO_2 from the atmosphere. Ocean alkalinization to enhance the oceanic uptake of atmospheric CO_2 was achieved by simulating the addition of 10 Pg/yr of $Ca(OH)_2$ evenly to the surface water between 70°N and 60°S. SRM was performed by reducing the incoming shortwave radiation in such a way that the surface air temperature (SAT) reaches and remains at levels corresponding to an atmospheric CO_2 concentration of 400 ppm. More details on each method and the justification for the applied intensity can be found in Keller et al. (2014) and references therein. Our measures are applied identically to those in Keller et al. (2014) except for SRM where we limit global mean SAT to levels obtained if atmospheric CO_2 concentration was 400 ppm instead of 280 ppm.

Model outputs are numerous, and while some are of particular interest to climate scientists, they might not be the most useful ones for decision makers who need to consider societal objectives. The choice of model outputs or indicators used to assess the CE-induced changes in the Earth system remains an active subject of discussion (Oschlies et al., 2016). In this work, we decided to investigate six outputs chosen from different components of the ESCM: SAT, precipitation, atmospheric CO_2 concentration, vegetation net primary productivity (NPP), ocean oxygen content, and ocean surface Ω aragonite. SAT and precipitation are important indicators of the climate system and also have strong and direct influences on society. Atmospheric CO_2 concentration, vegetation NPP, and ocean oxygen help us understand changes in the carbon cycle due to the contributions of the ocean and land surface. Ocean oxygen highlights the extent of hypoxia, which represents physiological stresses for marine aerobic organisms (Pörtner & Farrell, 2008). Ocean Ω aragonite, or the level of calcium carbonate saturation in seawater, is an indicator of the potential for biotic calcification, for example, of molluscs, crustaceans, and corals. Ocean acidification could lead to undersaturation and dissolution of calcium carbonate in parts of the surface ocean during the 21st century, which might have detrimental effects on marine ecosystems (Orr et al., 2005).

2.2. Uncertain Parameters

The design of a PPE depends on the scenarios being investigated and the outputs of interest. Different CE measures can introduce uncertainties that are specific to the individual method and may be different from the uncertainties related to climate change. Thus, the most sensitive input parameters may differ for the different scenarios. We decided on a small number of model parameters to focus our analysis on since it is too expensive to analyze all of the uncertain parameters simultaneously. Ideally, the selection of parameters should be based on how sensitive the output of interest is to changes in each of the uncertain parameters. However, in the lack of a probabilistic sensitivity analysis (Saltelli et al., 2000) to establish the most sensitive parameters formally, we rely on expert judgment following earlier analyses.

Table 1
The Perturbed Model Parameters, Their Short Names Used in This Paper and Their Prior Ranges

	Parameter	Short name	Range	Distribution	Unit	Reference
1	Meridional moisture diffusivity peak value in Southern Hemisphere	A_diff	0.7×10^6–10.0×10^6	Uniform	m^2/s	Brennan et al. (2012)
2	Equilibrium climate sensitivity scaling	vcsfac	1–8	Uniform	°C	Ross et al. (2012)
3a	Ocean isopycnal tracer mixing coefficient	ahisop	8.0×10^6–3.1×10^7	Uniform	cm^2/s	Ehlert et al. (2017) and MacDougall et al. (2016)
3b	Isopycnal thickness diffusivity	athkdf	8.0×10^6–3.1×10^7	Uniform	cm^2/s	Ehlert et al. (2017)
4	Background diapycnal mixing	kappa_h	0.01–0.4	Log Uniform	cm^2/s	Ross et al. (2012)
5	Vegetation q10	q10_veg	1.5–3.5	Uniform	N/A	Tjoelker et al. (2001) and Atkin et al. (2005)
6	Soil q10	q10_soil	1.5–3.5	Uniform	N/A	Raich and Schlesinger (1992), Lloyd and Taylor (1994), and Meyer et al. (2018);
7	CO_2 fertilization	co2_fert_scal	0.5–1.5	Uniform	N/A	Mengis (2016)
8	CO_2 sensitivity of transpiration	sens_fact	0.5–1.5	Uniform	N/A	Mengis (2016)
9	Detritus sinking speed at surface	wd0	2–20	Uniform	m/day	Berelson (2002) and Armstrong et al. (2009)
10	Maximum potential phytoplankton growth rate at 0°C	abio	0.2–4.0	Uniform	day^{-1}	Quéré et al. (2005), Bissinger et al. (2008), and Sherman et al. (2016)
11a	Phytoplankton half-saturation constant for NO_3 uptake	k1n	0.002–10.000	Uniform	μM	Harrison et al. (1996), Collos et al. (2005), and Schmittner et al. (2008)
11b	Phytoplankton half-saturation constant for PO_4 uptake	k1p	k1n × redptn[a]	Uniform	μM	Schmittner et al. (2008) and Lin et al. (2016)
12	Initial slope of the P-I curve	alpha	0.03–0.90	Uniform	$(Wm^{-2})^{-1}day^{-1}$	Platt et al. (1980)
13	O_2:N ratio	redotn	9.0–12.5	Uniform	N/A	

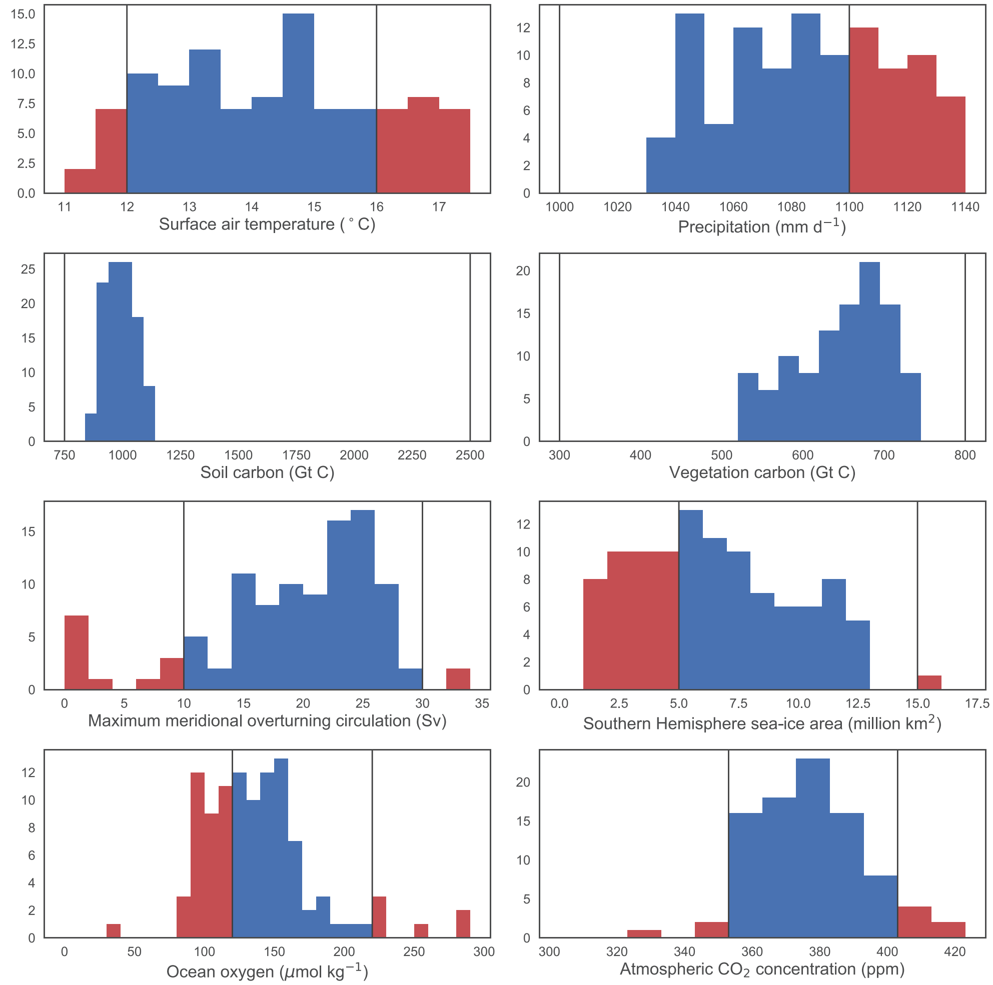

Figure 1. Histograms of the eight outputs used to constrain the parameter space. The acceptable ranges as described in Table 2 are indicated by the vertical solid lines. The rejected and accepted simulations are in red and blue, respectively.

The first group of perturbed parameters are those controlling the physical behavior of the system. Not only do they have direct effects on the climatic response but they also influence ocean biogeochemistry and terrestrial processes. The meridional diffusive moisture transport in the UVic atmosphere is latitude dependent with highest values at midlatitudes to represent actions of transient eddies. The meridional moisture diffusivity scaling parameter (A_diff in Table 1) controls the strength of water vapor transport in the Southern Hemisphere. Altering the moisture transport has a substantial impact on the hydrological cycle and deep water formation (Saenko & Weaver, 2003; Schmittner et al., 2005). To explore the uncertainty in ocean mixing, we varied the Gent-McWilliams parameterization's isopycnal tracer mixing coefficient (ahisop) and isopycnal thickness diffusivity (athkdf) (Gent & Mcwilliams, 1990) and the ocean background thickness diffusivity (κ_h) (Schmittner et al., 2005). The two isopycnal mixing parameters ahisop and athkdf are set to the same value to reduce the number of perturbed parameters. Ocean mixing parameters have been shown to have significant impacts on climate, ocean meridional circulation, ocean heat uptake, and the ocean carbon budget (Ehlert et al., 2017; Goes et al., 2010; Olson et al., 2012; Schmittner et al., 2009; Weaver et al., 2001). Ocean oxygen content and the extent of marine suboxia are also shown to be sensitive to the strength of background vertical mixing (Duteil & Oschlies, 2011). To capture the uncertainty in the temperature-longwave radiation feedback, we perturb a parameter (vcsfac), which was designed to alter the equilibrium climate sensitivity (Zickfeld et al., 2009). The range given in Table 1 for this vcsfac corresponds to the equilibrium climate sensitivity observed when other model parameters are kept at default values. We decided to keep the upper bound of vcsfac at 8°C based on an exploratory ensemble in which all simulations with climate sensitivity above this value were rejected by the metrics displayed in Figure 1.

Beyond uncertainties in the response of the physical system, land carbon-climate feedbacks remain among the largest sources of uncertainty in current ESMs (Arora et al., 2013; Friedlingstein et al., 2014; Todd-Brown et al., 2014). A significant contribution to this uncertainty is the effects of rising temperatures and atmospheric CO_2 on the physiology and growth of vegetation. Another factor contributing to the uncertainty in terrestrial carbon uptake is the variations in respiration (Konings et al., 2019). Here, we vary the temperature sensitivity of soil respiration (q10_soil) and plant photosynthesis via the temperature-dependent maximal rate of carboxylation (q10_veg). The carbon dioxide sensitivity of photosynthesis, that is, the CO_2 fertilization effect, and transpiration are investigated by perturbing co2_fert_scal and sens_fact parameters, respectively (Matthews & Caldeira, 2007; Mengis et al., 2015). These four parameters dictate how climate change affects the biomass production of terrestrial ecosystems and the land's carbon sequestration potential. Furthermore, Fyfe et al. (2013) suggested that biogeochemical processes are as important as radiative processes in influencing the hydrological cycle.

The oceanic contribution to the carbon cycle is also highly uncertain, especially under CE scenarios. We varied six parameters controlling the phytoplankton maximum growth rate (abio), nutrient uptake (k1n, k1p), detritus sinking speed (wd0), the initial slope of the light response curve (alpha), and the O_2:N ratio (Keller et al., 2012; Schmittner et al., 2008). While both the solubility pump and the biological carbon pump are strongly affected by parameters controlling the ocean circulation and physics, only the biological pump is affected directly by the biological parameters. Apart from their vital roles in the distribution and cycling of carbon in the ocean, perturbing the biological parameters also affects the simulated marine ecosystems through the cycling of nutrients such as nitrogen and phosphorus, as well as oxygen-dependent remineralization processes.

While the chosen parameters are expected to have some impacts on the climate and carbon uptake under all scenarios, their roles might vary significantly in each case. It is likely that terrestrial parameters, such as co2_fert_scal and sens_fact, will be more influential under AF, while ocean biogeochemical parameters will have stronger influences under AOA. The specific contributions, however, depend on the complex interactions between all components of the Earth system. The uncertainty analysis let us quantify the uncertainties induced by these parameters for each scenario.

Even though many of the parameters chosen here have been the subject of previous sensitivity studies, probability density functions estimated using observational constraints were only available for the equilibrium climate sensitivity factor and the ocean mixing coefficients (Olson et al., 2012). We did not use these estimates as our prior since we use a newer version of the UVic model and a different mixing scheme. Despite the fact that UVic is an efficient model capable of performing large PPEs, most sensitivity analyses so far were done by perturbing a single parameter to a few higher and lower values than the standard one (Duteil & Oschlies, 2011; Ehlert et al., 2017; Kvale & Meissner, 2017; Mengis et al., 2015) or only perturbing a small number of parameters using a factorial design (Ross et al., 2012). Since little is known quantitatively of the behavior and interactions of these parameters across the whole parameter space, we assume uniform distribution across the defined range for all but one parameter, the background diapycnal mixing which has a log-uniform distribution. The references for the ranges used are provided in Table 1. While it is ideal to associate a joint probability distribution with all the uncertain parameters, in practice, this requires knowledge of the correlation structure between the parameters, which is not available in our case. Thus, we associate a single probability distribution with each parameter as usually done in similar cases (Haylock, 1997).

3. Statistical Framework

The structure of our work is illustrated in Figure 2. Most methods for assessing the effect of uncertain inputs on the output have the same basic structure, that is, sampling from subjective prior distributions that represent the uncertainty on each input, evaluating the output of the model many times, and calculating output sample statistics. The first task of parameter selection was described in section 2.2.

Once the choice of perturbed parameters has been made, a training set, or a set of carefully designed simulations, is generated to inform the GP emulators. Because ESMs are computationally expensive, we want a design that is capable of exploring interactions between parameters and is valid across the whole range of the input parameter space using a minimum amount of simulations. Thus, a Latin hypercube sampling plan (McKay et al., 1979) with the maximin space-filling criteria (Morris & Mitchell, 1995) was employed. We followed the general suggestion from Loeppky et al. (2009) to have 10 simulations for each perturbed

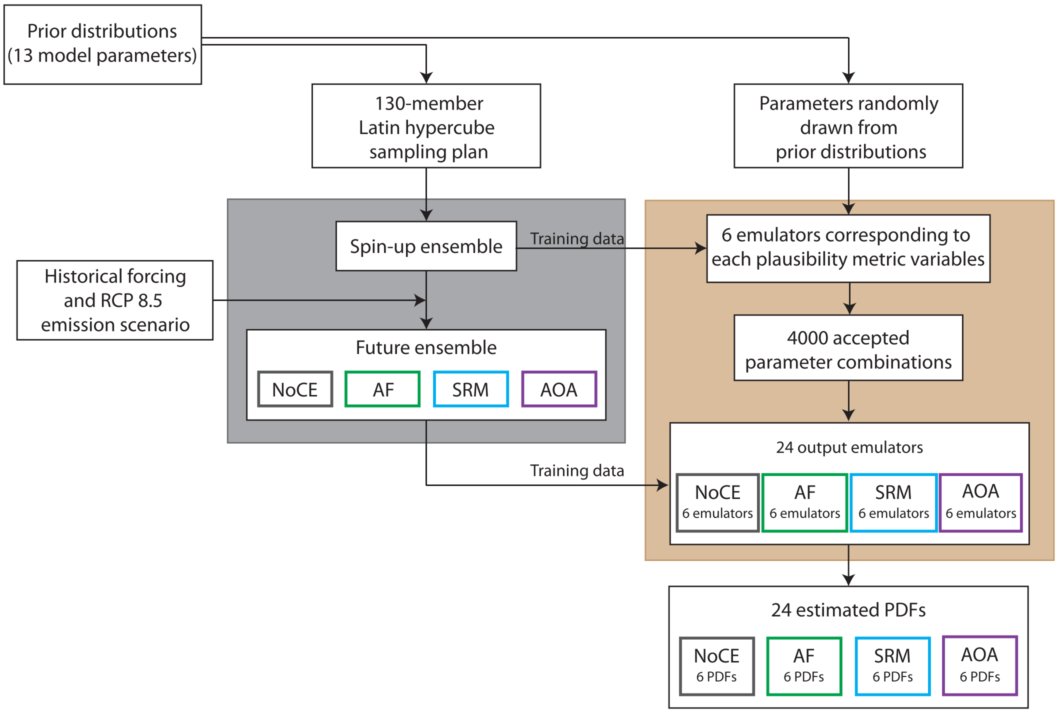

Figure 2. A schematic of the procedure in this study. The gray and brown boxes indicate the simulated and emulated ensemble, respectively. The CE and the control scenarios are color coded. The same color code is used throughout this paper.

parameter, giving us 130 spin-up simulations. In total, we ran 130 spin-up simulations and 130 × 4 future simulations for the four scenarios.

To ensure that model outputs are meaningful and do not include unphysical climate states, we define a set of eight plausibility metrics describing various aspects of the Earth system. While it is a common practice to tune the model by minimizing the misfit between model outputs and observations, there are inherent limitations in treating tuning as an optimization problem (Williamson et al., 2016). Therefore, we do not weight our ensemble members based on how close they are to observations. Our choice of metrics (section 4.1) represents the acceptable range of Earth system variables based on measurement ranges (to take into account observation errors) combined with multimodel ensemble ranges (to encompass the potential structural uncertainty range). We reject parameter choices where the model fails to remain within the acceptable ranges. Only parameter configurations fulfilling all these criteria are used in the uncertainty analysis.

The traditional uncertainty and sensitivity analyses as described in Saltelli et al. (2000) demand a substantial number of model runs. As climate models are computationally expensive, we employ GP emulators as surrogates of the ESM. An emulator is a statistical approximation of the input-output relationship of the model. Once constructed, the emulator can provide estimates of the model outputs at untried input configurations at a low computational cost. More information on GP emulator can be found in Appendix A. Here we utilize the GPy toolbox for Python (GPy, since 2012). GP emulators have been used in the past to conduct uncertainty and sensitivity analyses (Johnson et al., 2015; Lee et al., 2011; Oakley & O'Hagan, 2002, 2004a), calibration (Kennedy & O'Hagan, 2001), and history matching (Craig et al., 2001; Williamson et al., 2013).

In the approach that we followed we consider the model as a function $f(\mathbf{x})$, where $\mathbf{x}$ is a vector of input parameters. In such a probabilistic framework, the input parameters are treated as random variables $\mathbf{X}$. The two important uncertainty measures we are interested in are the uncertainty mean, $M = E[f(\mathbf{X})]$, and the variance due to parametric uncertainty, $V = \mathrm{Var}[f(\mathbf{X})]$. Since we do not know the value of $f(\mathbf{x})$ at all possible points in the plausible parameter space, there remains an uncertainty in the output, termed code uncertainty (Kennedy & O'Hagan, 2001). When estimating $f(\mathbf{.})$ with a GP emulator, we get a probabilistic estimation of $f(\mathbf{.})$ described by a mean function, $m_1(\cdot)$ and a covariance function, $v_1(\cdot,\cdot)$. When using the emulator's predictions instead of simulation outputs, the uncertainty mean, variance due to uncertain parameters, and

Table 2
Seven Preindustrial and One Present-Day Plausibility Metrics

Quantity	Observations	Metric range
Global mean surface air temperature (°C)	~14 (Jones et al., 1999)	12 to 16
Global mean precipitation (mm/yr)	942 to 1,139 (Rudolf & Rubel, 2005)	1,000 to 1,100
Global soil carbon (GtC)	850 to 2,400 (Bondeau et al., 2007)	750 to 2,500
Global vegetation carbon (GtC)	450 to 650 (Bondeau et al., 2007)	300 to 800
Maximum Atlantic overturning (Sv)	~19 (Kanzow et al., 2010)	10 to 30
Southern Hemisphere sea ice area (million km^2)	~7 (Cavalieri et al., 2003)	5 to 15
Global ocean averaged dissolved O_2 (μmol/kg)	~170 (Conkright et al., 2002)	120 to 220
Atmospheric CO_2 in 2005 (ppm)	378 (Keeling et al., 2005)	353 to 403

variance due to code uncertainty become $E*[M]$, $E^*[\text{Var}]$, and $\text{Var}^*[M]$, respectively, with the $E*[.]$ and $\text{Var}^*[.]$ denote the operations with respect to the emulator. These three quantities can be computed using $m_1(\mathbf{x})$ and $v_1(\mathbf{x},\mathbf{x}')$ following Oakley and O'Hagan (2002) and are described in Appendix A.

Each of the emulators constructed is first validated to ensure that they perform sufficiently well. Since we did not have a separate ensemble for validation, leave-one-out cross validation was used. This process consists of leaving out one of the training runs, building the emulator with the remaining runs, and predicting the simulated output of the left out simulation. This process is then repeated for every member of the training ensemble. To measure how close the emulator's predicted means are to the simulated values, we plot the to quantities against each other, calculate the root mean square error and the coefficient of determination, r-squared. The emulator's variances are evaluated by computing the fraction of training points where the emulator correctly predicts the simulated value, within the 66th, 95th, and 99th inner quantiles of the distribution. The validation results are provided in supporting information Text S1.

The uncertainty analysis using GP emulation is performed in two main steps, using the Monte Carlo simulation approach with emulator outputs in place of simulator outputs. In the first step, we constructed emulators of the eight variables corresponding to the plausibility metrics. Two of the metrics turn out to be uninformative in terms of constraining the prior parameter distributions, as all simulator outputs lie within the plausibility range (global vegetation and soil carbon, Figure 1), leaving six emulators in the analysis (section 4.1). Then we apply a rejection sampling method known as approximate Bayesian computation where new parameter combinations are drawn randomly from the prior input distributions but only accepted when the emulators predict they are within all the plausibility ranges. This process resulted in a sample of 4,000 plausible input parameter combinations.

In the second step, four sets of six emulators were constructed for each of the six outputs of interest under four scenarios. These are then used to estimate the model projected outcomes under the future scenarios for each of the 4,000 accepted input parameter sets generated in Step 1. Both steps are shown in Figure 2. We then calculate the uncertainty measures using the emulated outcomes of the 4,000 input settings.

4. Model Outputs

4.1. Spin-Up Ensemble

Figure 1 shows the eight metrics used to constrain the ensemble. The solid vertical lines denote the acceptable ranges listed in Table 2. Model outputs outside of the ranges are indicated in red. The eight metrics defined to reject implausible climate states are global mean SAT, precipitation rate, the strength of the meridional overturning circulation, Antarctic sea ice area, vegetation carbon, soil carbon, ocean oxygen, and atmospheric CO_2 concentration. Seven of the metrics are applied to the preindustrial climate at the end of the spin-up stage, while the atmospheric CO_2 metric is applied to simulations forced by historical forcings until the Year 2005.

By perturbing 13 parameters at the same time, the individual effect from each input and the interactions between them can lead to very diverse climatic outcomes. For example, while most simulations exhibit a vigorous overturning circulation, several simulations show a significant slow down of the meridional overturning circulation (eight simulations have the maximum overturning circulation strength below 5 Sv). Based on a multimodel ensemble of coupled climate models, it is suggested that the preindustrial Atlantic meridional overturning circulation strength ranges from 13.4 to 30.0 Sv (Roberts et al., 2013), while the present-day observations are around 17.2 Sv (McCarthy et al., 2015) or 19 Sv (Kanzow et al., 2010). Thus, we conclude that simulations with very weak (below 10 Sv) or very strong (above 30.0 Sv) preindustrial meridional overturning circulation are unrealistic and should be excluded in the final analysis. Similarly, other metrics are defined to constrain the model outcomes to those that cannot be ruled out by current observational constraints and modeling studies.

Some of the metrics tend to reject very low (sea ice area and ocean oxygen content) or very high values (precipitation rate), while others do not discard any simulations at all (vegetation carbon and soil carbon). On the one hand, this demonstrates the vast range of uncertainty associated with these quantities. On the other hand, this indicates that the prior range of input parameters cannot produce model outputs covering the whole plausibility ranges for the terrestrial carbon budget. While this could be due to structural model errors, that is, the lack of some key physical processes, in our case it is also likely due to the limited number of parameters varied. Among the 13 perturbed parameters, those with the most significant influences on the terrestrial carbon budgets are the CO_2-dependent (sens_fact and co2_fert_scal) and temperature-dependent (the two q10 parameters) parameters. These parameters lead to diverse output ranges under transient forcing with increasing atmospheric CO_2 concentration and rising temperature. However, in spin-up simulations where the atmospheric CO_2 concentration is fixed and is the same across the ensemble, they do not have an influence. Thus, the diverse simulated vegetation and soil carbon only manifest in the transient phase and not in the preindustrial phase used in the metrics. We progress with this limitation in mind and will incorporate this information into future ensemble design.

Since the two terrestrial carbon budget metrics do not constrain the ensemble, they are not considered in the uncertainty analysis. The remaining six metrics are later applied to all emulated outputs, and only parameter sets passing all six constraints are used in the uncertainty analysis.

Overall, 106 out of 130 simulations finished successfully, that is, without numerically induced model crashes. All failed simulations feature very high values of ocean isopycnal diffusivity, ahisop, and isopycnal thickness diffusvity, athkdf. The number of simulations that are consistent with all the metric ranges is 24 or approximately 23% of the successful runs and 18% of the PPE input parameter sets.

4.2. Future Projections

The model outputs under the four transient scenarios extended from the 24 simulations that fulfill all criteria defined in the previous section are shown in Figure 3. The solid lines indicate the mean and the shaded areas denote the range of the 24 simulations for each of the color-coded scenarios. The evolution of the six outputs of interest under the four different scenarios is shown. For SAT, we look at the anomaly with respect to the historical mean between 1986 and 2005.

Here, we compare our RCP8.5 runs (the noCE scenario) with other models. The simulated global mean warming from the reference period (the 1986 to 2005 mean) to the end of this century (the 2081 to 2100 average) is 3.4 ± 0.2 °C, which is comparable to the multimodel average of 3.9 ± 0.9 °C in CMIP5 (Friedlingstein et al., 2014) and the 1.6–4.1 °C range from an Earth system models of intermediate complexity intercomparison study (Zickfeld et al., 2013), albeit with a lower standard deviation. For precipitation, some simulations show a global increase, while others show a decrease. However, the changes throughout the entire transient period are small, ranging from −0.6 to 0.6%/K. This is smaller than the global increase of 1–3%/K seen in most models (Allen & Stainforth, 2002; Held & Soden, 2006). Keller et al. (2014) reported a global decrease of 0.3%/K in UVic precipitation due to a large reduction in terrestrial precipitation. Our ensemble shows that this behavior is parameter dependent and UVic is capable of producing both drier and wetter climates under the RCP8.5 scenario, in agreement with a previous study (Mengis et al., 2015). The simulated atmospheric CO_2 concentration in 2100 is 1,017 ± 35 ppm, which is on the higher end of the CMIP5 range of 985 ± 97 ppm (Friedlingstein et al., 2014). The CMIP5 mean ocean oxygen content in 2090–2099 compared to the 1990–1999 mean is -6.13 ± 0.78 mmol/m^3 or -3.45 ± 0.44%. Our constrained ensemble show a larger change as well as a larger standard deviation of -6.9 ± 2.3 mmol/m^3 or -4.6 ± 1.8%. Overall, our ensemble compares well with the CMIP5 multimodel ensemble.

In comparison with the previous CE comparison study by Keller et al. (2014), apart from the difference in the SRM scenario, we started the historical forcing from 1850 instead of 1765. For all scenarios, we note similar trends in all model outputs. We observe smaller changes in SAT and precipitation by the Year 2100 in the

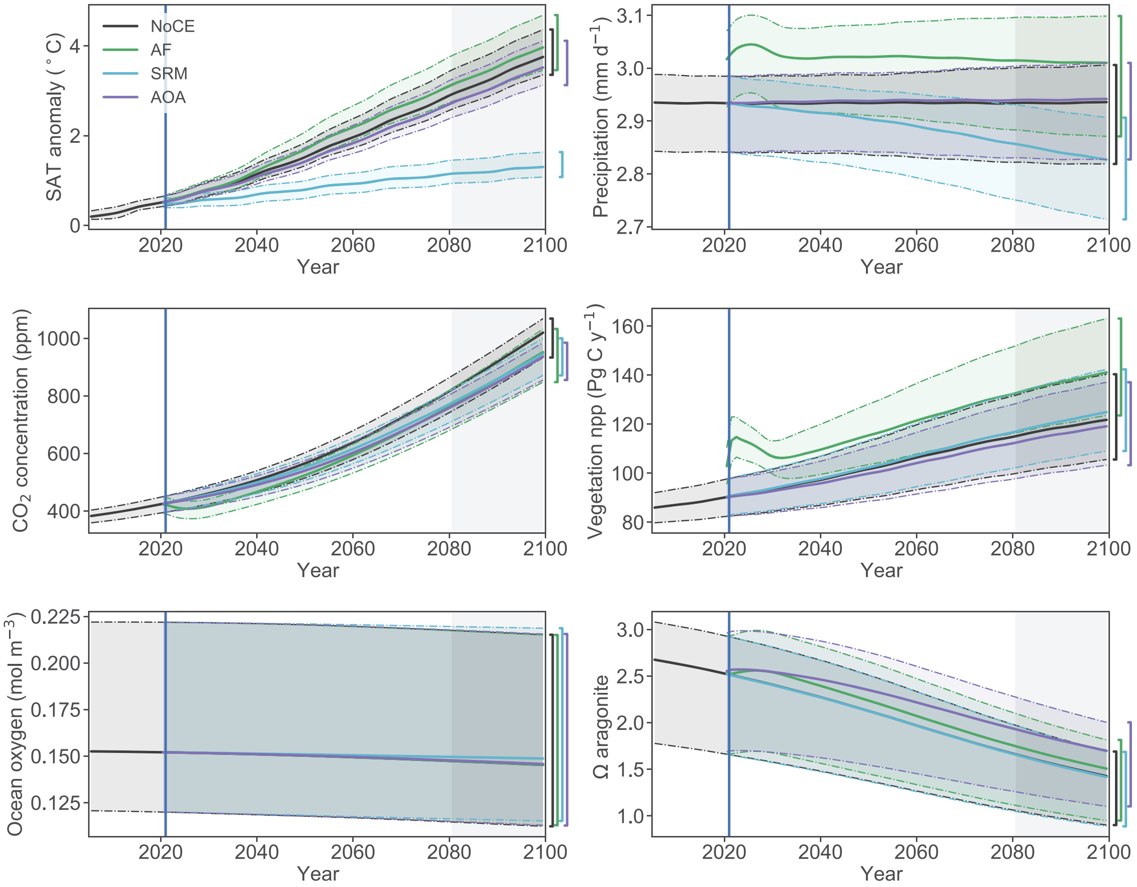

Figure 3. The simulated changes in globally averaged annual SAT (relative to the 1986–2005 average), precipitation, atmospheric CO_2, vegetation NPP, ocean oxygen, and Ω aragonite for model runs which fulfilled all metrics. The CE period starts in 2020 and extends to the end of the simulations in Year 2100. Emulators are constructed for the average outputs over the period from 2081–2100, which is indicated by the vertical gray shaded area in this figure. For each scenario, the solid line shows the ensemble mean and the shaded area shows the range of the remaining 24 model runs. The upper and lower edges of each range are marked with dash-dotted lines of the same color. The ranges in the Year 2100 are shown as brackets on the right side of the plots.

SRM scenario due to the smaller applied reduction in incoming shortwave radiation. For all but vegetation NPP, the results in Keller et al. (2014) fall within the range shown in Figure 3. In the case of vegetation NPP, our ensemble lower bounds are slightly higher than the outputs seen in Keller et al. (2014). This could be a result of the different initial condition of the spin-ups (using atmospheric CO_2 concentration from 1850 instead of 1765). We briefly summarize the main features of the CE results below. For a more detailed discussion, the readers are referred to Keller et al. (2014) and the references therein.

Regarding the surface temperature, AOA leads to a small reduction in warming while AF leads to more warming due to changes in regional surface albedo as the desert is replaced by less reflective vegetation. SRM significantly reduce surface warming by reducing the incoming shortwave radiation. All CE scenarios lead to a sizeable reduction in atmospheric CO_2 compared to when no CE is used. Even though SRM does not directly target atmospheric CO_2, changes in the oceanic and terrestrial carbon sink lead to a lower increase in atmospheric CO_2 concentration. For precipitation, the changes are small under the AOA scenario. SRM leads to a significantly drier climate because of less evapotranspiration and evaporation. Global precipitation under the AF scenario is on average higher than under all other scenarios due to more water being added to the hydrological cycle. The increasing atmospheric CO_2 concentration increases plant productivity due to the fertilization effect. Under AF, terrestrial productivity increases further due to more water from irrigation and rainfall. On average, all scenarios show ocean deoxygenation throughout the 80 years of simulation. However, in some SRM simulation, oxygen content by the end of the century is higher than the historical reference period (1986–2005). Global warming leads to the loss of dissolved O_2 due to O_2 being less soluble in warmer water and because of the resulting increased upper ocean stratification causing a reduction in the O_2 supply to the ocean interior. The cooling effect of SRM works to combat this deoxygenation trend.

Ocean Ω aragonite decreases to a dangerously low level in all scenarios, with average surface Ω aragonite being close to 1.0 in many simulations. The use of AOA appears to alleviate the problem somewhat, while AF has a smaller impact and SRM results in almost no change compared to noCE.

In all cases, we can see that the outcomes of the four scenarios overlap significantly and might not be statistically different. The common practice for comparing CE measures is to compare a single realization for each method, using the same parameter setup. Without properly communicating that these realizations are drawn from different probability distributions, there is a risk of sending the wrong message to the policymakers. The reader might falsely conclude that one measure outperforms others, while in reality, when model uncertainty is considered, the probability distributions of the outcomes of two scenarios may be indistinguishable. We will revisit this point in section 5.

Another feature of the ensemble is that for precipitation and ocean oxygen, the uncertainty range is enormous compared to the magnitude of change in the corresponding quantity. It is crucial to identify the factors contributing the most to the variations in these two outputs if we aim to achieve more meaningful projections.

Due to the small number of simulations available (24 simulations for 13 perturbed parameters), we subsequently employed GP emulation to perform the probabilistic uncertainty quantification shown in the next section. We now construct the emulators of the average outputs (2081–99), except for SAT (2081–99 relative to 1986–2005).

5. Results

The emulator construction is described in section 3 and Appendix A, and validations are detailed in the supporting information. Each emulator was used to calculate the three uncertainty measures described in section 2.

We compare the output variance due to code uncertainty (Var*$[M]$) with the variance due to uncertain parameters ($E * [V]$) to justify the use of the emulators. Table S1 in the supporting information shows the comparison between the two types of variance for all emulated outputs under the four scenarios. For all outputs, the uncertainties due to emulation is small compared to the parametric uncertainty due to perturbed parameters. The only emulator that contributes appreciably to the overall output variance is that of ocean oxygen, with code uncertainty being greater than 1% of parameter uncertainty. The code uncertainties for atmospheric CO_2 emulators are the second largest. Emulator validation shows that for all four CO_2 emulators, there is an excessive amount of simulations (75%) predicted correctly within one standard deviation compared to the expected 68%. However, the percentage of simulations correctly predicted within two or three standard deviations is very close to what we expect. This means that the emulators are slightly overestimating the variance of atmospheric CO_2 concentration and thus overestimating the code uncertainty. Thus, we could treat the estimated code uncertainty as an upper bound of the real value. For the emulators to be considered reliable, the code uncertainties need to be small compared to the parametric uncertainties. Here, since the upper bound of the code uncertainties remain quite small compared to parametric uncertainty, we are satisfied with the emulators.

Figure 4 shows the uncertainties in the six global averaged annual mean outputs. The distributions vary between outputs and scenarios. All distributions are unimodal but display different levels of skewness. Ocean oxygen content and Ω aragonite have the highest skewness. In such cases, by focusing only on the ensemble means, low probability extreme events are not considered despite their substantial potential impacts on human life and the ecosystem. Thus, the likelihood of avoiding a dangerous threshold might be of more interest to society and policymakers.

For SAT, SRM induces a very narrow distribution due to the way this CE method is implemented in UVic. The radiative forcing is not the same for all SRM simulations and also is not constant during the 80-year deployment to adjust SAT to a predefined goal, that is, to return SAT to the level seen when atmospheric CO_2 concentration was 400 ppm. As a result, the SAT is constrained to a narrow range compared to the other methods or when no CE is used.

While the ensemble spread due to parametric uncertainty is large for all scenarios, the distributions of ocean oxygen appear to be insensitive to CE. This could be due to the short duration of the simulations, which does

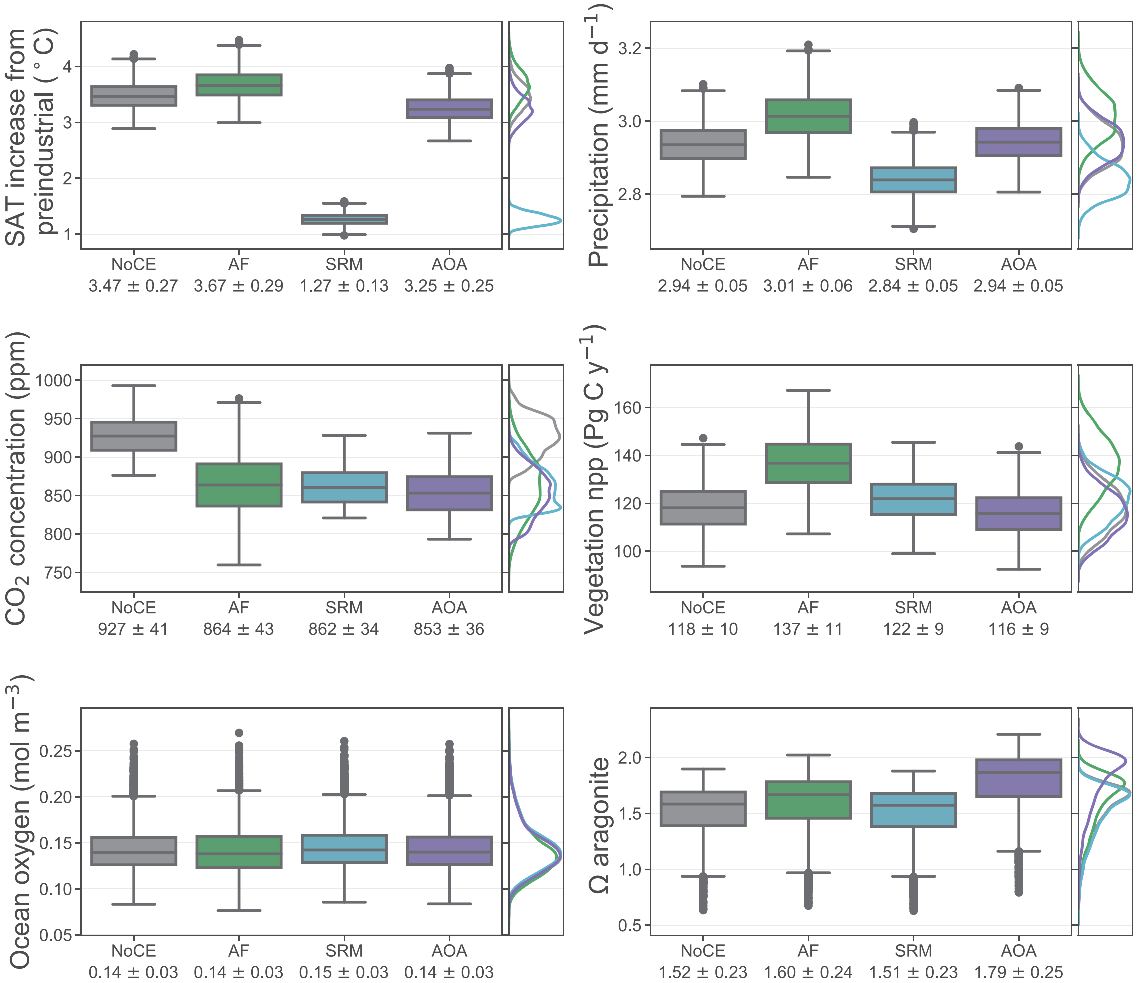

Figure 4. The uncertainty in globally averaged annual SAT (relative to a historical period), precipitation, atmospheric CO_2, vegetation NPP, ocean oxygen, and Ω aragonite for the four scenarios. The quantities are averaged over the last 10 years of the simulations. The boxplot shows the emulated distribution for each quantity. The mean and standard deviations are shown under each boxplot. The right panel of each plot shows the kernel density estimated PDFs (Appendix B).

not allow the ocean to equilibrate. Thus, changes in the ocean might be limited to the surface and have not reached greater depth. Another possible explanation is that the global annual mean quantity masks opposing regional changes. Regional responses of surface oxygen might be more sensitive to the same perturbations. The poorer performance of oxygen emulators compared to others could be a result of this apparent insensitivity. If reducing ocean deoxygenation is a key goal in adopting CE measures, then our result suggests that it is not possible to identify the best method given the large parameter uncertainty.

Figure 4 again highlights the fact that the distributions of the results of the four scenarios overlap and in many cases are unlikely to be statistically different. Comparing the ensemble means does not ensure a reliable assessment of the relative impacts of the CE scenarios. For example, while AOA leads to a lower atmospheric CO_2 CO_2 concentration on average, the ensemble spread is larger than that of the SRM scenario, meaning that under AOA, it is possible to end up with a higher atmospheric CO_2 concentration in 2100 than under SRM. To demonstrate this point further, we compare the ranking of the four scenarios for each of the 4,000 parameter combinations used in the uncertainty analysis. For each parameter combination, the emulators provide the estimated outputs for each scenario. For an output, the scenario producing the minimum value out of the four is ranked 1, while Rank 4 corresponds to the scenario with the maximum output value. We then compute the frequency of each CE scenario achieving each rank. Figure 5 shows that SRM consistently leads to lowest global mean SAT for all parameter combinations, followed by AOA, noCE, and AF. AF leads to a consistently higher air temperature largely due to the changes in surface albedo.

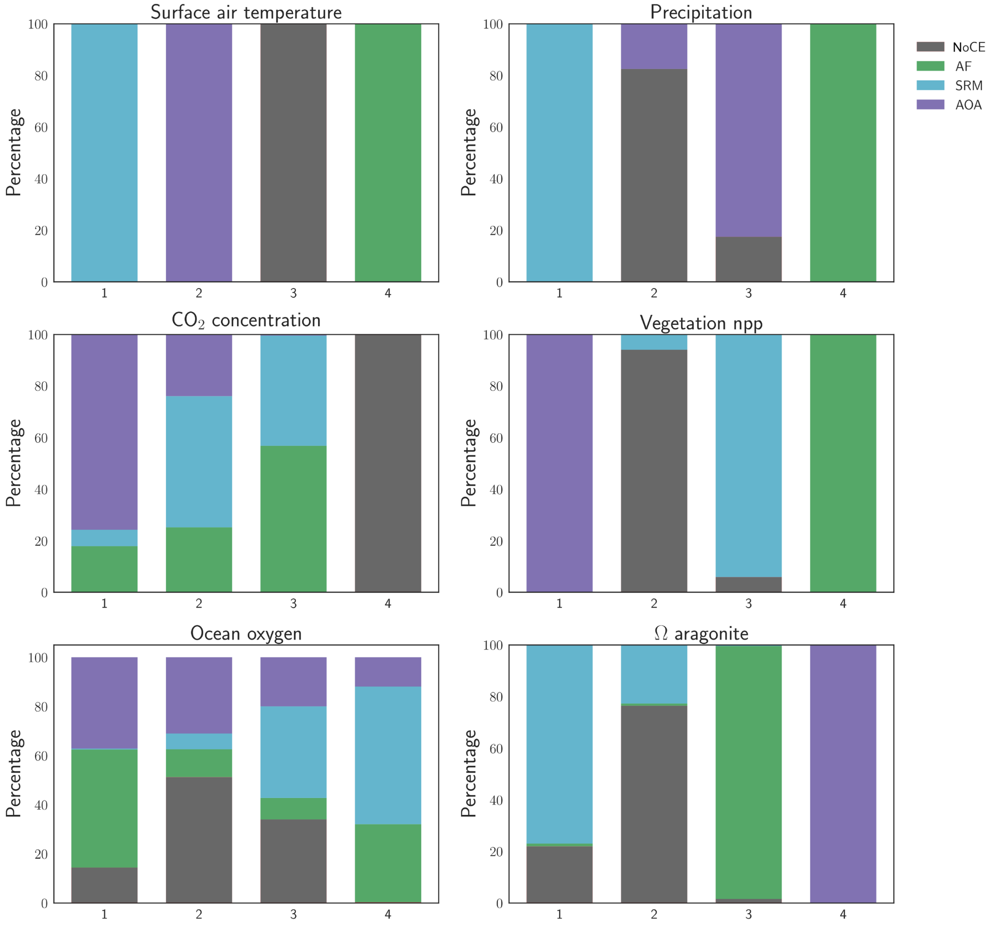

Figure 5. The probability of achieving a certain ranking for each CE measure. For each output of interest, the outcome from the four scenarios are ranked from 1 to 4 with 1 being the lowest.

For the remaining outputs, the ranking depends on the parameter settings. For example, while AOA is the most likely to lead to the lowest atmospheric CO_2 concentration, for over 20% of the parameter settings, SRM or AF can outperform AOA.

The highest precipitation rate under AF is observed consistently throughout the parameters space due to more water being added to the hydrological cycle through irrigation. The lowest precipitation happens when SRM is implemented as found in other studies. AOA is more likely to produce slightly higher precipitation than when no CE measure is used. However, the difference between these two scenarios are often very small (less than 0.01 mm/day). Thus, AOA essentially leads to no meaningful change in precipitation.

All three CE measures reduce the atmospheric CO_2 concentration compared to when none is employed. The ranking appears to depend on the parameter controlling the atmospheric moisture transport and the ocean diffusivity coefficients. The ocean mixing coefficients have a large effect on ocean carbon uptake, while the moisture transport has an effect on both the terrestrial carbon uptake via changes in terrestrial productivity and on the ocean carbon uptake as a result of changes in the ocean circulation.

Ω aragonite experience the lowest decrease when AOA is implemented as expected. AF also manages to keep the saturation states of aragonite higher than when no CE is applied. This is due to a reduction in the ocean carbon inventory in response to a large increase in the terrestrial inventory, with respect to the noCE scenario.

Overall, the ranking of the outputs is often complex and requires further investigation into the model's behavior. For example, the difference in vegetation NPP under the four scenarios is due to a combination of CE's effects on the hydrological cycle, the amount of atmospheric carbon being removed, and the uncertain terrestrial parameters.

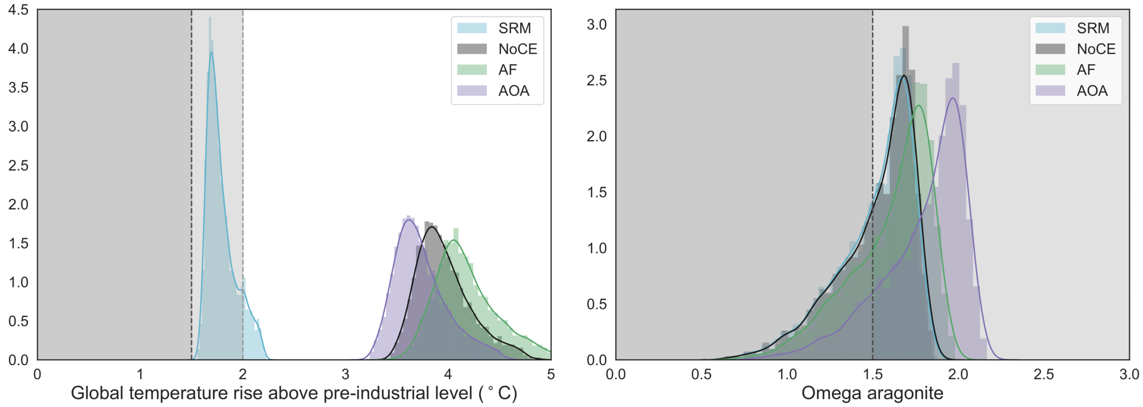

Figure 6. The probability of crossing the surface air temperature (left) and Ω aragonite (right) threshold for the four CE measures. The vertical dashed lines indicate the thresholds.

We propose that a more appropriate way to rank the scenarios would be based on the probability of avoiding a dangerous threshold. For illustrative purpose, we define thresholds for SAT and Ω aragonite (Figure 6). SAT is the most frequently used climate indicator when assessing climate change and CE. The thresholds for SAT are chosen to be 1.5 and 2 °C above the preindustrial value, following to the Paris agreement. The 2 °C threshold is a goal beyond which many believe substantial irreversible climate change would occur. Under the parametric uncertainty considered and the CE scenarios as we defined, it is not possible to reach either temperature goals using AF, AOA, or by noCE. SRM has a 0% and 88.8% chance of achieving the 1.5 and 2 °C goals, respectively. It is not surprising that SRM is the most successful in terms of temperature reduction considering that the method was designed to stabilize SAT to the same level when atmospheric CO_2 concentration was 400 ppm.

For surface Ω aragonite, we set two thresholds of 1.5 and 3.0. Studies have shown that no prominent present-day coral reefs exist in environments with Ω aragonite below 3.0 (Guinotte et al., 2003; Kleypas et al., 1999). Thus, we follow Meissner et al. (2012) and Feng et al. (2016) in choosing this value as a critical threshold for coral habitat. The 1.5 threshold has been used previously to signify water with carbonate chemistry stressful to larvae of shellfish such as oysters (Ekstrom et al., 2015; Gimenez et al., 2018). For Ω aragonite values below 1.5 marine organisms are believed to have trouble forming shells during the first few days of their life (Waldbusser et al., 2015). This significantly lowers their life expectancy and can have a severe effect on the wealth being of marine ecosystems. In all four cases, the surface annual average Ω aragonite stay below the 3.0 threshold. While the means of the distribution for all four cases are above the 1.5 threshold, the probability of crossing this limit is not negligible, that is, 37%, 28.2%, 38.3%, and 13.8% for noCE, AF, SRM, and AOA, respectively.

Other thresholds that might be relevant are the mass balance of Antarctica, the strength of the Atlantic Meridional Overturning Circulation, sea ice area, or sea level rise extent. Detailed investigations on these quantities are beyond the scope of this exploratory study.

6. Conclusions

There are many studies focused on the effectiveness and potential unintended outcomes of existing CE methods conducted in different modeling framework (Schäfer et al., 2015). A study comparing several CE measures in UVic ESCM allows quantitative assessments of their relative impacts and side effects (Keller et al., 2014). We further contribute to such comparison studies by considering the often overlooked parametric uncertainty of the ESM used. Of the three CE scenarios considered here, only SRM could limit the warming to the targeted 2 °C. However, this measure could lead to significant unintended changes in the hydrological cycle and does not tackle the problem of ocean acidification. AOA manages to alleviate ocean acidification but only has a small impact on reducing air temperature. While the simulated changes depend strongly on the modeling framework and the forcing scenarios, our work supports the conclusion drawn from many previous studies that CE is not an alternative to mitigation (Lawrence et al., 2018; The Royal Society, 2009; Vaughan & Lenton, 2011). However, they could be considered as complementary to mitigation efforts to alleviating the adverse effects of climate change on the ecosystem. Since there are many individual

CE measures and it might be desirable to combine any number of these measures with mitigation to form specific portfolios, an assessment of the magnitude and the probability of both intended and unintended consequences of proposed portfolios should be conducted.

In this study, we limit ourselves to estimating the variance in projected outcomes due to poorly constrained parameters in UVic. Using a PPE and GP emulation, we demonstrated the significance of parameter uncertainty in the model projected outcomes; that is, their magnitude can be comparable to multimodel ensemble variance such as in the case of ocean oxygen presented here. Through the use of a PPE, we have demonstrated the importance of considering probability distributions when comparing outcomes from different CE scenarios. The standard deviations and skewness of the estimated probability density functions showed that the use of a single projection or the mean of an ensemble fails to address the likelihood of an outcome and neglects low probability high-risk events. Therefore, a probabilistic uncertainty analysis such as presented here could be more valuable to decision makers who need to weigh the benefits of CE measures against their side effects and their costs.

To summarize our assessment, we ranked the four scenarios considered in terms of their probability of achieving a desirable goal, for example, limiting the mean global warming since preindustrial times to 2 °C or less and maintaining the saturation state of aragonite to above 1.5. The goal of CE or mitigation is unlikely to be a single target but a combination of multiple targets including those specific for individual regions.

The distributions of the uncertainty in the six outputs emulated show that they are different for different outputs/processes. While only uncertainty is investigated in this work, the PPE and the constructed emulators allow probabilistic sensitivity analysis to be conducted, providing further insights into the processes driving the outcomes. Thus, major contributors to uncertainties can be identified, guiding future model development efforts and observational or experimental studies, which could constrain uncertain parameters.

Our methodology demonstrated using UVic could be further develop to tackle uncertainty in more computationally expensive ESMs. The topic of transferring uncertainty information between models of different spatial resolutions and/or timescales is relatively new and has gained more attention in the last few years. There are already several works on extending the emulation method to cope with more computationally expensive models. For example, Sexton et al. (2019) used statistical emulation to investigate the link between model errors that develop in the relatively high resolution the HadGEM3 GA4 model (1.875° longitude × 1.25° latitude) on a short (5-day) and a longer (5-year) timescales and concluded that the short forecast can be used to filter out implausible parameter combinations. This can significantly lighten the computational cost required to explore the parameter space. There are also different approaches such as using multilevel emulation techniques, which relate the "inexpensive" information from a lower complexity model to the "expensive" information from a more complex model. The foundation of such techniques can be found in Kennedy and O'Hagan (2000) or Forrester et al. (2008). Applications of multilevel emulation in climate science are presented in Williamson et al. (2012) and Tran et al. (2019).

This study is currently limited to global annual mean quantities, which cannot guarantee a sensible distribution in space and time and are not very meaningful when regional or extreme events are of interest. Average outputs are also known to hide opposing spatial or temporal effects. Future work is planned to extend the presented statistical technique to address regional climate. There are examples where dimension reduction techniques can be employed to reduce the cost of emulating high-dimensional outputs such as the use of principal components analysis (Holden et al., 2010; Wilkinson, 2010), P-splines (Williamson et al., 2012), and wavelets (Bayarri et al., 2007). It is also possible to compute representative quantities for predefined geographical regions and apply the same methodology here.

The parameter uncertainty discussed in this work is only one of the contributions to the overall uncertainty. The uncertainty in ESM results could originate from structural deficiencies due to our lack of understanding of the Earth system, the inability to model processes due to the limit on temporal and spatial resolution, the inadequate observations used to calibrate the models, and the uncertainty in CE forcings. While we attempted to deal with observation and structural uncertainty by incorporating them into our metrics, this treatment is far from perfect. Technical difficulties remain an obstacle in addressing the uncertainty in ESM projections satisfactorily.

The result obtained in this work, as well as in a previous comparison study (Sonntag et al., 2018), shows that the outcomes depend heavily on the emission scenario and the specific intensity of CE measures used. Here, we are using highly idealized scenarios to assess near-maximum effects. Practical realities of deployment should be taken into account in future comparison works. There is a need to quantify the outcome's sensitivity to difference forcing/scenarios.

Structural uncertainties due to the underlying assumptions of the model are vital to our ability to infer real-world probabilities from model outcomes. Multimodel ensembles such as those from GeoMIP (Kravitz et al., 2011) or CDRMIP (Keller et al., 2018) provide central estimates of potential outcomes under a different modeling frameworks, illuminating the differences and similarities due to different modeling assumptions under common forcings. However, such studies do not provide robust insights into the causes. The use of PPEs such as ours provides information on the distributions, which contain those central estimations and could identify the underlying processes that lead to structural discrepancies. Ultimately, it is desirable to combine information from both multi-model ensemble and perturbed parameter ensemble studies to obtain a better picture of uncertainty in ESM projections.

While the specification of structural uncertainty or model discrepancy is crucial to an uncertainty assessment, this step is not address in this paper due to its complexity. Brynjarsdóttir and O'Hagan (2014) have shown that neglecting structural uncertainty could lead to bias or overconfidence in predictions and the inclusion of genuine prior information about model discrepancy could greatly improve model prediction. The difficulties arise when we have to deal with complex model with a very high dimensional input space and multivariate outputs. Specifying structural error in such cases becomes far less straightforward, especially when considering components which rely less on fundamental physical laws.

There are also issue of identifiability, that is, whether a discrepancy comes from structural or parametric uncertainty. The precipitation in our ensemble exhibits a smaller change and a smaller standard deviation compared to the CMIP5 ensemble. This could mean that our model is incapable of producing the larger range or that we have not included all important parameters. In this case, we believe that model discrepancy certainly plays a role. We have perturbed parameters that control SAT and vegetation response to changing temperature and CO_2. While these are not an exhaustive list of parameters, they are likely to be the most prominent drivers of precipitation. A previous study (Mengis et al., 2015) that perturbed the response of transpiration to CO_2 over a broader range than we considered here showed that UVic can cover the CMIP5 range of precipitation over land but fails to do so for global precipitation change. This is perhaps a consequence of the simple 2-D energy and moisture balance model of the atmosphere.

The model's inability to cover the specified metric ranges for soil and vegetation carbon could be attributed to unaccounted for parametric uncertainty, model discrepancy, or indeed, observational uncertainty. Many processes in the Earth system, such as the biological aspects and the carbon cycle response, are not yet well constrained by observations. The estimations for the soil and vegetation carbon budgets used are very broad due to the wide range in observations and model estimations. There is also a need to identify robust measurements, which could be used to constraint the physiological response of vegetation to rising temperature and CO_2 concentrations.

Given these difficulties, we decided to deal solely with parametric uncertainty to avoid misspecifying structural uncertainty. In future works, multiwave history matching technique (Andrianakis et al., 2015; Williamson et al., 2016) could be used in conjunction with second or third maximum implausibility (Vernon et al., 2010) to guard against wrongly rejecting good parameter sets and provide a more robust assessment of both parametric and structural uncertainty. The work presented will serve as a foundation to further explore the model uncertainties in UVic.

Appendix A: Emulation and Uncertainty Measures

For a model with input $\mathbf{x} = (x_1, \dots \ , x_p)$, where p is the dimension of the input space, a single output y can be treated as the value of a scalar function f evaluated at $\mathbf{x}$. ESMs are expensive simulators; therefore, we approximate f by a cheaper statistical surrogate model. Since $\mathbf{x}$ is uncertain, we can can consider it to be a random vector X with a probability distribution $g(x)$. Consequently, the output $Y = f(X)$ is a random variable. The uncertainty about the Y can be represented using a GP emulator.

$Y(x)$ is multivariate normally distributed such that

$$Y(\mathbf{x}) \sim \mathcal{N}(m_0, V_0) \tag{A1}$$

where m_0 and V_0 are the prior mean and covariance function, respectively. The prior mean of the GP is given by:

$$m_0 = h^T(\mathbf{x})\beta, \tag{A2}$$

where $h(\cdot)$ is a $(s \times 1)$ vector of known regression functions with unknown coefficients β. In this work, $h(x)^T = 1$, making β the unknown overall mean. Other mean functions could be used to express expert belief about the form of $f(\cdot)$.

The prior covariance function is given by:

$$V_0(\mathbf{x}, \mathbf{x}') = \sigma^2 C(\mathbf{x}, \mathbf{x}'), \tag{A3}$$

in which, σ^2 is an unknown variance of the GP and $C(\cdot, \cdot; \theta)$ is a positive-definite correlation function (with unknown correlation length parameters, θ) which decreases as $|x - x'|$ increases. Popular choices of the covariance structure are the squared exponential, Matérn and exponential functions.

The prior does not depend on the training data but specifies the assumptions we've made about the function $f(.)$. Then, the training data from n simulations are incorporated, allowing us to update the prior to the posterior GP. The predictor is taken to be the mean of the posterior process conditional on σ^2, β and θ:

$$m_1(\mathbf{x}) = h^T(\mathbf{x})\hat{\beta} + T(\mathbf{x})A^{-1}(\mathbf{y} - H\hat{\beta}) \tag{A4}$$

The variance of the posterior process is:

$$V_1(\mathbf{x}, \mathbf{x}') = \frac{\hat{\sigma}^2}{n - s - 2}[c(\mathbf{x}, \mathbf{x}') - T(\mathbf{x})A^{-1}T^T(\mathbf{x}') + P(\mathbf{x})(H^T A^{-1} H)^{-1} P^T(\mathbf{x}')], \tag{A5}$$

where H is the regression matrix of the training points, $H = h^T(\mathbf{x})$, and A is the training points correlation matrix, $A = \Psi(\mathbf{x}, \mathbf{x}')$; $T(\mathbf{x})$ is the correlation vector between $\mathbf{x}$ and the training set, that is, $(T(\mathbf{x}))_i = \Psi(\mathbf{x}, \mathbf{x}_i)$ and $P(\mathbf{x}) = h^T(\mathbf{x}) - T(\mathbf{x})A^{-1}H$. The estimated values of σ^2 and β are indicated as $\hat{\sigma}^2$ and $\hat{\beta}$, respectively. A standard non-informative prior is assumed for β.

More detailed descriptions of GP emulation can be found in Haylock (1997) and Rasmussen and Williams (2006).

The goal of uncertainty analysis is then to characterize the distribution of $f(X)$ that is induced by the distribution of X. Since Y is a random variable, any summary of Y is also a random variable. For more on random variables, Sudret (2007). Here we are interested in the mean, M = E[f(X)], and variance, V = Var[f(X)], of the simulator's uncertainty distribution due to uncertainty about input parameters. Another measure of interest is the variance in the mean of the uncertainty distribution due to code uncertainty,

We can now approximate $f(x)$ by $m_1(x)$. Thus, we use E*[.] and Var*[.] to denote the operations of expectation, variance and covariance computed using emulator. The estimates of M and V are now E*[M] and E*[V] respectively, while code uncertainty in these estimates are Var*[M] and Var*[V]. In this work, we do not compute Var*[V].

$$E^*[M] = \int m_1(x)dg(x) \tag{A6}$$

$$Var^*[M] = \int\int V_1(x, x')dg(x)dg(x') \tag{A7}$$

These measures can be evaluated using Monte Carlo computation with pairs of values sampled from $g(x)$.

The mean of V

$$E^*[V] = \left[\int m^*(x)^2 dg(x) - \left\{\int m^*(x)dg(x)\right\}^2\right] + \left[\int v^*(x, x)dg(x) - \int\int v^*(x, x')dg(x)dg(x')\right] \tag{A8}$$

We employ a general approach which make no assumptions about the form of the distribution of the input g(X), correlation function C(X,X'), and regression function h(X). Full derivation can be found in O'hagan 2011. For different special cases, more computationally efficient methods are available (Haylock1997, O'hagan2011).

Appendix B: Kernel Density Estimation

Kernel density estimation (KDE) is a statistical tool, which is used to approximate the probability density function $g(\cdot)$ of a random variable X. Given a sample of independent observations $(x_1, x_2, \dots, x_n)$ from the random variable X, the kernel density estimator is as follows:

$$g^*(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right), \tag{B1}$$

where K represents the nonnegative kernel function, which is controlled by the bandwidth parameter $h > 0$.

Each sample point is replaced by a curve, whose shape is determined by the kernel function, centered at that value. Then, KDE sums over all these curves to compute the value of the density at each point in the support grid. A region with many observations will yield a large value, while regions with few observations will result in a low value. The bandwidth of the KDE controls how “smooth” the resulting curve is. We use a Gaussian kernel but other choices are available. The resulting curve is then normalized so that the area under it is equal to 1.

Acknowledgments
This work is a contribution to the German DFG-funded Priority Program “Climate Engineering: Risks, Challenges, Opportunities?” (SPP 1689). The model data used to generate the table and the figures will be made available online (https://data.geomar.de/).

References

Allen, M. R., Frame, D. J., Huntingford, C., Jones, C. D., Lowe, J. A., Meinshausen, M., & Meinshausen, N. (2009). Warming caused by cumulative carbon emissions towards the trillionth tonne. *Nature*, *458*(7242), 1163–1166.

Allen, M. R., & Stainforth, D. A. (2002). Towards objective probabilistic climate forecasting. *Nature*, *419*(6903), 228.

Andrianakis, I., Vernon, I. R., McCreesh, N., McKinley, T. J., Oakley, J. E., Nsubuga, R. N., et al. (2015). Bayesian history matching of complex infectious disease models using emulation: A tutorial and a case study on HIV in Uganda. *PLoS Computational Biology*, *11*(1), e1003968.

Armstrong, R. A., Peterson, M. L., Lee, C., & Wakeham, S. G. (2009). Settling velocity spectra and the ballast ratio hypothesis. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *56*(18), 1470–1478.

Arora, V. K., Boer, G. J., Friedlingstein, P., Eby, M., Jones, C. D., Christian, J. R., et al. (2013). Carbon-concentration and carbon-climate feedbacks in CMIP5 Earth system models. *Journal of Climate*, *26*(15), 5289–5314.

Atkin, O. K., Bruhn, D., Hurry, V. M., & Tjjoelker, M. G. (2005). Evans review No. 2—The hot and the cold: Unravelling the variable response of plant respiration to temperature. *Functional Plant Biology*, *32*(2), 87–105.

Bayarri, M. J., Berger, J. O., Cafeo, J., Garcia-Donato, G., Liu, F., Palomo, J., et al. (2007). Computer model validation with functional output. *The Annals of Statistics*, *35*(5), 1874–1906.

Berelson, W. M. (2002). Particle settling rates increase with depth in the ocean. *Deep-Sea Research Part II: Topical Studies in Oceanography*, *49*(1-3), 237–251.

Bissinger, J. E., Montagnes, D. J. S., Sharples, J., & Atkinson, D. (2008). Predicting marine phytoplankton maximum growth rates from temperature: Improving on the Eppley curve using quantile regression. *Limnology and Oceanography*, *53*(2), 487–493.

Bitz, C. M., & Lipscomb, W. H. (1999). An energy-conserving thermodynamic model of sea ice. *Journal of Geophysical Research*, *104*(C7), 15,669–15,677. https://doi.org/10.1029/1999JC900100

Bondeau, A., Smith, P. C., Zaehle, S., Schaphoff, S., Lucht, W., Cramer, W., et al. (2007). Modelling the role of agriculture for the 20th century global terrestrial carbon balance. *Global Change Biology*, *13*(3), 679–706.

Brennan, C. E., Weaver, A. J., Eby, M., & Meissner, K. J. (2012). Modelling oxygen isotopes in the University of Victoria Earth system climate model for pre-industrial and Last Glacial Maximum conditions. *Atmosphere-Ocean*, *50*(4), 447–465. https://doi.org/10.1080/07055900.2012.707611

Brynjarsdóttir, J., & O'Hagan, A. (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems*, *30*(11), 114,007.

Caldeira, K., Bala, G., & Cao, L. (2013). The science of geoengineering. *Annual Review of Earth and Planetary Sciences*, *41*(1), 231–256. https://doi.org/10.1146/annurev-earth-042711-105548

Cavalieri, D. J., Parkinson, C. L., & Vinnikov, K. Y. (2003). 30-year satellite record reveals contrasting Arctic and Antarctic decadal sea ice variability. *Geophysical Research Letters*, *30*(18), 1970. https://doi.org/10.1029/2003GL018031

Collos, Y., Vaquer, A., & Souchu, P. (2005). Acclimation of nitrate uptake by phytoplankton to high substrate levels. *Journal of Phycology*, *41*(3), 466–478.

Conkright, M. E., Locarnini, R. A., Garcia, H. E., O'Brien, T. D., Boyer, T. P., Stephens, C., & Antonov, J. I. (2002). World Ocean Atlas 2001: Objective analyses, data statistics, and figures CD-ROM documentation. *National Oceanographic Data Center Internal Report (NOAA Atlas NESDIS)*, *17*(September), 17.

Craig, P. S., Goldstein, M., Rougier, J. C., & Seheult, A. H. (2001). Bayesian forecasting for complex systems using computer simulators. *Journal of the American Statistical Association*, *96*(454), 717–729.

Crook, J. A., Jackson, L. S., Osprey, S. M., & Forster, P. M. (2015). A comparison of temperature and precipitation responses to different Earth radiation management geoengineering schemes. *Journal of Geophysical Research: Atmospheres*, *120*, 9352–9373. https://doi.org/10.1002/2015JD023269

Duteil, O., & Oschlies, A. (2011). Sensitivity of simulated extent and future evolution of marine suboxia to mixing intensity. *Geophysical Research Letters*, *38*, L06607. https://doi.org/10.1029/2011GL046877

Eby, M., Weaver, A. J., Alexander, K., Zickfeld, K., Abe-Ouchi, A., Cimatoribus, A. A., et al. (2013). Historical and idealized climate model experiments: An intercomparison of Earth system models of intermediate complexity. *Climate of the Past*, *9*(3), 1111–1140.

Eby, M., Zickfeld, K., Montenegro, A., Archer, D., Meissner, K. J., & Weaver, A. J. (2009). Lifetime of anthropogenic climate change: Millennial time scales of potential CO_2 and surface temperature perturbations. *Journal of Climate*, *22*(10), 2501–2511.

Ehlert, D., Zickfeld, K., Eby, M., & Gillett, N. (2017). The sensitivity of the proportionality between temperature change and cumulative CO_2 emissions to ocean mixing. *Journal of Climate*, *30*(8), 2921–2935.

Ekstrom, J. A., Suatoni, L., Cooley, S. R., Pendleton, L. H., Waldbusser, G. G., Cinner, J. E., et al. (2015). Vulnerability and adaptation of US shellfisheries to ocean acidification. *Nature Climate Change*, *5*(3), 207–214.

Fanning, A. F., & Weaver, A. J. (1996). An atmospheric energy-moisture balance model: Climatology, interpentadal climate change, and coupling to an ocean general circulation model. *Journal of Geophysical Research*, *101*, 15111.

Feng, E. Y., Keller, D. P., Koeve, W., & Oschlies, A. (2016). Could artificial ocean alkalinization protect tropical coral ecosystems from ocean acidification? *Environmental Research Letters*, *11*(7), 074008.

Flato, G., Marotzke, J., Abiodun, B., Braconnot, P., Chou, S. C., Collins, W., et al. (2013). Evaluation of climate models. In T. F. Stocker, D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschun, et al. (Eds.), *Climate change 2013: The physical science basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change* (pp. 741–866). Cambridge, United Kingdom and New York, NY, USA: Cambridge University Press.

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

Friedlingstein, P., Meinshausen, M., Arora, V. K., Jones, C. D., Anav, A., Liddicoat, S. K., & Knutti, R. (2014). Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks. *Journal of Climate*, *27*(2), 511–526.

Fyfe, J. C., Cole, J. N. S., Arora, V. K., & Scinocca, J. F. (2013). Biogeochemical carbon coupling influences global precipitation in geoengineering experiments. *Geophysical Research Letters*, *40*, 651–655. https://doi.org/10.1002/grl.50166

Gent, P. R., & Mcwilliams, J. C. (1990). Isopycnal mixing in ocean circulation models. *Journal Of Physical Oceanography*, *20*, 150–155. https://dx.doi.org/10.1175/1520-0485(1990)020<0150:IMIOCM>2.0.CO;2

Gillett, N. P., Arora, V. K., Zickfeld, K., Marshall, S. J., & Merryfield, W. J. (2011). Ongoing climate change following a complete cessation of carbon dioxide emissions. *Nature Geoscience*, *4*(2), 83–87.

Gimenez, I., Waldbusser, G. G., & Hales, B. (2018). Ocean acidification stress index for shellfish (OASIS): Linking Pacific oyster larval survival and exposure to variable carbonate chemistry regimes. *Elementa: Science of the Anthropocene*, *6*(1), 51.

Goes, M., Urban, N. M., Tonkonojenkov, R., Haran, M., Schmittner, A., & Keller, K. (2010). What is the skill of ocean tracers in reducing uncertainties about ocean diapycnal mixing and projections of the Atlantic Meridional Overturning Circulation? *Journal of Geophysical Research*, *115*, C12006. https://doi.org/10.1029/2010JC006407

GPy (2012). GPy: A Gaussian process framework in python. http://github.com/SheffieldML/GPy

Guinotte, J. M., Buddemeier, R. W., & Kleypas, J. A. (2003). Future coral reef habitat marginality: Temporal and spatial effects of climate change in the Pacific basin. *Coral Reefs*, *22*(4), 551–558.

Harrison, A. W. G., Harris, L. R., Irwin, B. D., Section, B. O., By, N. S., & Harrison, W. G. (1996). The kinetics of nitrogen utilization in the oceanic mixed layer: Nitrate and ammonium interactions at nanomolar concentrations. *Limnology and Oceanography*, *41*(1), 16–32.

Haylock, R. G. E. (1997). Bayesian inference about outputs of computationally expensive algorithms with uncertainty on the inputs (Ph.D. Thesis), University of Nottingham.

Held, I. M., & Soden, B. J. (2006). Robust responses of the hydrological cycle to global warming. *Journal of Climate*, *19*, 5686–5699.

Holden, P. B., & Edwards, N. R. (2010). Dimensionally reduced emulation of an AOGCM for application to integrated assessment modelling. *Geophysical Research Letters*, *37*, L21707. https://doi.org/10.1029/2010GL045137

Holden, P. B., Edwards, N. R., Oliver, K. I. C., Lenton, T. M., & Wilkinson, R. D. (2010). A probabilistic calibration of climate sensitivity and terrestrial carbon change in GENIE-1. *Climate Dynamics*, *35*(5), 785–806.

IPCC (2014). Climate change 2014: Synthesis report. Contribution of Working Groups I, II and III to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change. Geneva, Switzerland.

Johnson, J. S., Cui, Z., Lee, L. A., Gosling, J. P., Blyth, A. M., & Carslaw, K. S. (2015). Evaluating uncertainty in convective cloud microphysics using statistical emulation. *Journal of Advances in Modeling Earth Systems*, *7*, 162–187. https://doi.org/10.1002/2014MS000383

Jones, P. D., New, M., Parker, D. E., Martin, S., & Rigor, I. G. (1999). Surface air temperature and its changes over past 150 years. *American Geophysical Union*, *37*(2), 173–199.

Kanzow, T., Cunningham, S. A., Johns, W. E., Hirschi, J. J. M., Marotzke, J., Baringer, M. O., et al. (2010). Seasonal variability of the Atlantic meridional overturning circulation at 26.5° N. *Journal of Climate*, *23*(21), 5678–5698.

Keeling, C. D., Piper, S. C., Bacastow, R. B., Wahlen, M., Whorf, T. P., Heimann, M., & Meijer, H. A. (2005). Atmospheric CO_2 and $^{13}CO_2$ exchange with the terrestrial biosphere and oceans from 1978 to 2000: Observations and carbon cycle implications. In *A history of atmospheric CO2 and its effects on plants, animals, and ecosystems* (pp. 83–113). New York: Springer Verlag.

Keller, D. P., Feng, E. Y., & Oschlies, A. (2014). Potential climate engineering effectiveness and side effects during a high carbon dioxide -emission scenario. *Nature communications*, *5*, 3304.

Keller, D. P., Lenton, A., Scott, V., Vaughan, N. E., Bauer, N., Ji, D., et al. (2018). The Carbon Dioxide Removal Model Intercomparison Project (CDRMIP): Rationale and experimental protocol for CMIP6. *Geoscientific Model Development*, *11*(3), 1133–1160.

Keller, D. P., Oschlies, A., & Eby, M. (2012). A new marine ecosystem model for the University of Victoria Earth system climate model. *Geoscientific Model Development*, *5*(5), 1195–1220.

Kennedy, M. C., & O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, *87*(1), 1–13.

Kennedy, M. C., & O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, *63*(3), 425–464.

Kleypas, J. A., McManus, J. W., & Menez, L. A. B. (1999). Environmental limits to coral reef development: Where do we draw the Line? *American zoologist*, *39*, 146–159.

Konings, A. G., Bloom, A. A., Liu, J., Parazoo, N. C., Schimel, D. S., & Bowman, K. W. (2019). Global satellite-driven estimates of heterotrophic respiration. *Biogeosciences*, *16*, 2269–2284.

Kravitz, B., Robock, A., Boucher, O., Schmidt, H., Taylor, K. E., Stenchikov, G., & Schulz, M. (2011). The Geoengineering Model Intercomparison Project (GeoMIP). *Atmospheric Science Letters*, *12*(2), 162–167.

Kvale, K. F., & Meissner, K. J. (2017). Primary production sensitivity to phytoplankton light attenuation parameter increases with transient forcing. *Biogeosciences*, *14*(20), 4767–4780.

Lawrence, M. G., Schäfer, S., Muri, H., Scott, V., Oschlies, A., Vaughan, N. E., et al. (2018). Evaluating climate geoengineering proposals in the context of the Paris Agreement temperature goals. *Nature Communications*, *9*(1), 1–19.

Lee, L. A., Carslaw, K. S., Pringle, K. J., & Mann, G. W. (2012). Mapping the uncertainty in global CCN using emulation. *Atmospheric Chemistry and Physics*, *12*(20), 9739–9751.

Lee, L. A., Carslaw, K. S., Pringle, K. J., Mann, G. W., & Spracklen, D. V. (2011). Emulation of a complex global aerosol model to quantify sensitivity to uncertain parameters. *Atmospheric Chemistry and Physics*, *11*(23), 12,253–12,273.

Lin, S., Litaker, R. W., & Sunda, W. G. (2016). Phosphorus physiological ecology and molecular mechanisms in marine phytoplankton. *Journal of Phycology*, *52*(1), 10–36.

Lloyd, J., & Taylor, J. A. (1994). On the temperature dependence of soil respiration. *Functional Ecology*, *8*(3), 315–323.

Loeppky, J. L., Sacks, J., & Welch, W. J. (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics*, *51*(4), 366–376.

MacDougall, A. H., Swart, N. C., & Knutti, R. (2016). The uncertainty in the transient climate response to cumulative CO2 emissions arising from the uncertainty in physical climate parameters. *Journal of Climate*, *30*, 813–827.

Matthews, H. D., & 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.

Matthews, H. D., & Caldeira, K. (2008). Stabilizing climate requires near-zero emissions. *Geophysical Research Letters*, *35*, L04705. https://doi.org/10.1029/2007GL032388

McCarthy, G. D., Smeed, D. A., Johns, W. E., Frajka-Williams, E., Moat, B. I., Rayner, D., et al. (2015). Measuring the Atlantic Meridional Overturning Circulation at 26°N. *Progress in Oceanography*, *130*, 91–111.

McKay, M. D., Beckman, R. J., & Conover, W. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, *21*(2), 239–245.

Mcneall, D. J. (2008). Dimension reduction in the Bayesian analysis of a numerical climate model (Ph.D. Thesis), University of Southampton, Faculty of Engineering Science and Mathematics, School of Ocean and Earth Science.

Meinshausen, M., Meinshausen, N., Hare, W., Raper, S. C. B., Frieler, K., Knutti, R., et al. (2009). Greenhouse-gas emission targets for limiting global warming to 2 °C. *Nature*, *458*(7242), 1158–1162.

Meissner, K. J., McNeil, B. I., Eby, M., & Wiebe, E. C. (2012). The importance of the terrestrial weathering feedback for multimillennial coral reef habitat recovery. *Global Biogeochemical Cycles*, *26*(3), 1–20.

Meissner, K. J., Weaver, A. J., Matthews, H. D., & Cox, P. M. (2003). The role of land surface dynamics in glacial inception: A study with the UVic Earth System Model. *Climate Dynamics*, *21*(7-8), 515–537.

Mengis, N. (2016). Towards a comprehensive, comparative assessment of climate Engineering schemes metrics, indicators and uncertainties (Ph.D. Thesis), Christian-Albrechts Universität Kiel.

Mengis, N., Keller, D. P., Eby, M., & Oschlies, A. (2015). Uncertainty in the response of transpiration to CO_2 and implications for climate change. *Environmental Research Letters*, *10*(9), 094,001.

Meyer, N., Welp, G., & Amelung, W. (2018). The temperature sensitivity (Q10) of soil respiration: Controlling factors and spatial prediction at regional scale based on environmental soil classes. *Global Biogeochemical Cycles*, *32*, 306–323. https://doi.org/10.1002/2017GB005644

Morris, M. D., & Mitchell, T. J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, *43*, 381–402.

Murphy, J. M., Sexton, DavidM. H., Barnett, D. N., Jones, G. S., Webb, M. J., Collins, M., & Stainforth, D. A. (2004). Quantifying uncertainties in climate change from a large ensemble of circulation change simulations. *Nature*, *430*(August 2004), 768–772.

National Research Council (2015). *Climate intervention: Reflecting sunlight to cool Earth*. Washington, DC: The National Academies Press.

Niemeier, U., Schmidt, H., Alterskjær, K., & Kristjánsson, J. E. (2013). Solar irradiance reduction via climate engineering: Impact of different techniques on the energy balance and the hydrological cycle. *Journal of Geophysical Research: Atmospheres*, *118*, 11,905–11,917. https://doi.org/10.1002/2013JD020445

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliability Engineering and System Safety*, *91*(10-11), 1290–1300.

Oakley, J., & O'Hagan, A. (2002). Bayesian inference for the uncertainty distribution of complex computer codes. *Biometrika*, *89*, 769–784.

Oakley, J. E., & O'Hagan, A. (2004a). Probabilistic sensitivity analysis of complex models: A Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *66*(3), 751–769.

Olson, R., Sriver, R., Goes, M., Urban, N. M., Matthews, H. D., Haran, M., & Keller, K. (2012). A climate sensitivity estimate using Bayesian fusion of instrumental observations and an Earth System model. *Journal of Geophysical Research*, *117*, D04103. https://doi.org/10.1029/2011JD016620

Orr, J. C., Fabry, V. J., Aumont, O., Bopp, L., Doney, S. C., Feely, R. A., et al. (2005). Anthropogenic ocean acidification over the twenty-first century and its impact on calcifying organisms. *Nature*, *437*(September), 681–686.

Oschlies, A., Held, H., Keller, D., Keller, K., Mengis, N., Quaas, M., et al. (2016). Indicators and metrics for the assessment of climate engineering. *Earth's Future*, *5*, 49–58.

Oschlies, A., Koeve, W., Rickels, W., & Rehdanz, K. (2010). Side effects and accounting aspects of hypothetical large-scale Southern Ocean iron fertilization. *Biogeosciences*, *7*(12), 4014–4035.

Oschlies, A., Pahlow, M., Yool, A., & Matear, R. J. (2010). Climate engineering by artificial ocean upwelling: Channelling the sorcerer's apprentice. *Geophysical Research Letters*, *37*, L04701. https://doi.org/10.1029/2009GL041961

Pörtner, H. O., & Farrell, A. P. (2008). Physiology and climate change. *Science*, *322*(October), 690–692.

Pacanowski, R. (1996). MOM 2 documentation user's guide and reference manual, GFDL Ocean Group Technical Report. NOAA/GFDL vol. 3.

Platt, T., Gallegos, C., & Harrison, W. (1980). Photoinhibition of photosynthesis in natural assemblages of marine phytoplankton. *Journal of Marine Research*, *38*, 687–701.

Plattner, G. K., Knutti, R., Joos, F., Stocker, T. F., von Bloh, W., Brovkin, V., et al. (2008). Long-term climate commitments projected with climate-carbon cycle models. *Journal of Climate*, *21*(12), 2721–2751.

Quéré, C. L., Harrison, S. P., Colin Prentice, I., Buitenhuis, E. T., Aumont, O., Bopp, L., et al. (2005). Ecosystem dynamics based on plankton functional types for global ocean biogeochemistry models. *Global Change Biology*, *11*, 2016–2040.

Raich, J. W., & Schlesinger, W. H. (1992). The global carbon dioxide flux in soil respiration and its relationship to vegetation and climate. *Tellus B*, *44*, 81–99.

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Cambridge: The MIT Press.

Reilly, J. (2001). Climate change: Uncertainty and climate change assessments. *Science*, *293*(5529), 430a–433.

Roberts, C. D., Garry, F. K., & Jackson, L. C. (2013). A multimodel study of sea surface temperature and subsurface density fingerprints of the Atlantic meridional overturning circulation. *Journal of Climate*, *26*(22), 9155–9174.

Ross, A., Matthews, D. H., Schmittner, A., & Kothavala, Z. (2012). Assessing the effects of ocean diffusivity and climate sensitivity on the rate of global climate change. *Tellus B*, *64*, 1–10.

Rudolf, B., & Rubel, F. (2005). Global precipitation, (1st ed.). In M. Hantel (Ed.), *Observed global climate, Landolt-Bornstein (numerical data and functional relationships, Group V: Geophysics, volume 6*. Berlin Heidelberg: Springer-Verlag Berlin Heidelberg.

Saenko, O. A., & Weaver, A. J. (2003). Atlantic deep circulation controlled by freshening in the Southern Ocean. *Geophysical Research Letters*, *30*(14), 10–13. https://doi.org/10.1029/2003GL017681

Saltelli, A., Chan, K., & Scott, E. M. (2000). *Sensitivity analysis*. Chichester: Wiley.

Santner, T. J., Williams, B. J., & Notz, W. I. (2003). *The design and analysis of computer experiments*. New York: Springer.

Schäfer, S., Lawrence, M., Stelzer, H., Born, W., Low, S., Aaheim, A., et al. (2015). The European Transdisciplinary Assessment of Climate Engineering (EuTRACE): Removing greenhouse gases from the atmosphere and reflecting sunlight away from Earth. Funded by the European Union's Seventh Framework Programme under Grant Agreement 306993.

Schellnhuber, H. J., Cramer, W., Nakicenovic, N., Wigley, T., & Yohe, G. (Eds.) (2006). *Avoiding dangerous climate change*. Cambridge: Cambridge University Press.

Schmittner, A., Oschlies, A., Giraud, X., Eby, M., & Simmons, H. L. (2005). A global model of the marine ecosystem for long-term simulations: Sensitivity to ocean mixing, buoyancy forcing, particle sinking, and dissolved organic matter cycling. *Global Biogeochemical Cycles*, *19*, 1–17. https://doi.org/10.1029/2004GB002283

Schmittner, A., Oschlies, A., Matthews, H. D., & Galbraith, E. D. (2008). Future changes in climate, ocean circulation, ecosystems, and biogeochemical cycling simulated for a business-as-usual CO_2 emission scenario until year 4000 AD. *Global Biogeochemical Cycles*, *22*(1), 1–21.

Schmittner, A., Urban, N. M., Keller, K., & Matthews, D. (2009). Using tracer observations to reduce the uncertainty of ocean diapycnal mixing and climate-carbon cycle projections. *Global Biogeochemical Cycles*, *23*(4), 1–16.

Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., et al. (2019). Finding plausible and diverse variants of a climate model. Part 1: Establishing the relationship between errors at weather and climate time scales. *Climate Dynamics*, *53*(1-2), 989–1022.

Sherman, E., Moore, J. K., Primeau, F., & Tanouye, D. (2016). Temperature influence on phytoplankton community growth rates. *Global Biogeochemical Cycles*, *30*, 550–559. https://doi.org/10.1002/2015GB005272

Sonntag, S., González, M. F., Ilyina, T., Kracher, D., Nabel, J. E. M. S., Pongratz, J., et al. (2018). Quantifying and comparing effects of climate engineering methods on the Earth system. *Earth's Future*, *6*, 149–168.

Stainforth, D. A., Aina, T., Christensen, C., Collins, M., Faull, N., Frame, D. J., et al. (2005). Uncertainty in predictions of the climate response to rising levels of greenhouse gases. *Nature*, *433*, 403–406.

Sudret, B. (2007). Uncertainty propagation and sensitivity analysis in mechanical models at contributions to structural reliability and stochastic spectral methods (Ph.D. Thesis), Université Blaise Pascal, Clermont-Ferrand, France.

The Royal Society (2009). Geoengineering the climate: Science, governance and uncertainty. September.

Tjoelker, M. G., Oleksyn, J., & Reich, P. B. (2001). Modelling respiration of vegetation: Evidence for a general temperature-dependent Q10. *Global Change Biology*, *7*(2), 223–230.

Todd-Brown, K. E. O., Randerson, J. T., Hopkins, F., Arora, V., Hajima, T., Jones, C., et al. (2014). Changes in soil organic carbon storage predicted by Earth system models during the 21st century. *Biogeosciences*, *11*(8), 2341–2356.

Tran, G. T., Oliver, K. I. C., Holden, P. B., Edwards, N. R., Sóbester, A., & Challenor, P. (2019). Multi-level emulation of complex climate model responses to boundary forcing data. *Climate Dynamics*, *52*(3), 1505–1531.

UNEP (2017). The Emissions Gap Report 2017. United Nations Environment Programme, Nairobi.

Vaughan, N. E., & Lenton, T. M. (2011). A review of climate geoengineering appraisals. *Climate Change*, *3*, 597–615.

Vernon, I., Goldsteiny, M., & Bowerz, R. G. (2010). Galaxy formation: A Bayesian uncertainty analysis. *Bayesian Analysis*, *5*(4), 619–670.

Waldbusser, G. G., Hales, B., Langdon, C. J., Haley, B. A., Schrader, P., Brunner, E. L., et al. (2015). Saturation-state sensitivity of marine bivalve larvae to ocean acidification. *Nature Climate Change*, *5*(3), 273–280.

Weaver, A. J., Eby, M., Alexander, K., Crespin, E., Fichefet, T., Joos, F., et al. (2012). Stability of the Atlantic Meridional Overturning Circulation: A model intercomparison. *Geophysical Research Letters*, *39*, L20709. https://doi.org/10.1029/2012GL053763

Weaver, A. J., Eby, M., Wiebe, E. C., Bitz, C. M., Duffy, P. B., Ewen, T. L., et al. (2001). The UVic Earth system climate model: Model description, climatology, and applications to past, present and future climates. *Atmosphere-Ocean*, *39*(4), 361–428.

Wilkinson, R. D. (2010). Bayesian calibration of expensive multivariate computer experiments. In L. Biegler et al. (Eds.), *Computational methods for large-scale inverse problems and quanti cation of uncertainity* (chap. 10). Chichester: John Wiley & Sons, Ltd.

Williamson, D., Blaker, A. T., & Sinha, B. (2016). Tuning without over-tuning: Pparametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development Discussions*, *0*, 1–41.

Williamson, D., Goldstein, M., Allison, L., Blaker, A., Challenor, P., Jackson, L., & Yamazaki, K. (2013). History matching for exploring and reducing climate model parameter space using observations and a large perturbed physics ensemble. *Climate Dynamics*, *41*(7-8), 1703–1729.

Williamson, D., Goldstein, M., & Blaker, A. (2012). Fast linked analyses for scenario-based hierarchies. *Journal of the Royal Statistical Society Applied Statistics*, *61*(5), 665–691.

Zickfeld, K., Eby, M., Matthews, H. D., & Weaver, A. J. (2009). Setting cumulative emissions targets to reduce the risk of dangerous climate change. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(38), 16,129–16,134.

Zickfeld, K., Eby, M., Weaver, A. J., Alexander, K., Crespin, E., Edwards, N. R., et al. (2013). Long-term climate change commitment and reversibility: An EMIC intercomparison. *Journal of Climate*, *26*(16), 5782–5809.