Improving the efficiency of Calibration in Earth System Modelling with Short-term Simulations

by

Anqi Ma

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Science
in
Geography

Waterloo, Ontario, Canada, 2023

© Anqi Ma 2023

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Calibration is an important process in developing Earth system models to reduce the deviations between model and observations and improve model performance. Current calibration process is quite resource demanding, requiring enormous computing resources to explore model responses to different parameter choices. In this study, we explored a method using short-term simulations for training emulators to predict model responses given parameter changes in calibration to potentially improve its efficiency. We generated two Perturbed Parameter Ensembles (PPEs) using the same model configuration and perturbed parameters values, but differ in the length of simulation. A significant linear relationship between short and long-term simulations in model response to parameters was found through EOF analysis. The contributions of each ensemble member in respond to parameter values were very similar between PPEs ran for short and long-term, which provided a support for using short-term simulations for emulator in calibration. The evaluation of Gaussian Process emulator trained with short-term simulations showed that the emulator trained with short-term simulations could capture major spatial variability in long-term simulations and generally has a good performance in predicting model responses to parameters. Nevertheless, there are more fragmented and chaotic components in model responses in short-term simulations, which may be caused by the larger internal variability in a shorter length of simulation. Such characteristics in short-term simulations may lead to regional biases when assess model responses in regional scales or related to natural variations using emulators trained with short-term simulations. Given that short-term simulations require much less computational time to run, these results implied that it is possible to use short-term simulations to improve the efficiency of calibration in the Earth system modelling, while the potential regional biases should be aware of when assessing model responses on regional bases.

Acknowledgements

I would like to thank all the people who made this thesis possible.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background

Earth system modelling refers to quantitative methods that use numerical models to describe and simulate the behaviour of the Earth's climate system. These models emerged from the mathematical efforts on numerical weather prediction by Richardson (Lynch, 2006) in the early 20th century, and became popular as computer resources advanced. The general circulation models (GCMs) employed a similar technique as numerical weather prediction but extended to global scales to fully capture large-scale atmospheric motions. To better represent the climate system and capture the complex interactions within, models gradually became more comprehensive over time by coupling atmosphere-ocean GCMs with other important systems such as land, hydrology and cryosphere(Edwards, 2011). Earth system models (ESMs) are global climate models with biogeochemical processes, which have been applied in performing climate predictions for future variability of important climate variables such as precipitation and temperature, and assessing climate change under different scenarios to assist decision-making for adaptation and mitigation (Collins et al., 2020).

In an ESM, the Earth system is divided into discrete grid cells, the physical processes represented in ESM are discretized, and a system of partial differential equations representing each process with be solved on the grid at each timestep(Figure 1.1). Because of the

limited computational resource, the resolution of the grid cells must be constrained within a feasible range, and could not be sufficient enough to represent the processes which happen within the grid cells fully (Prodhomme et al., 2016). The limitation of computational resources and resolution is the first challenge of Earth system modelling, and the physical processes in the Earth system were then divided into resolved processes (e.g. large-scale air movements that can be described under current resolution) and unresolved processes (the physical processes that happened within the grid cell and cannot be resolved under given resolution) (Balaji et al., 2022). As a compromise, using parameterization schemes to represent the unresolved processes on sub-grid scales became a common structure of Earth system models.

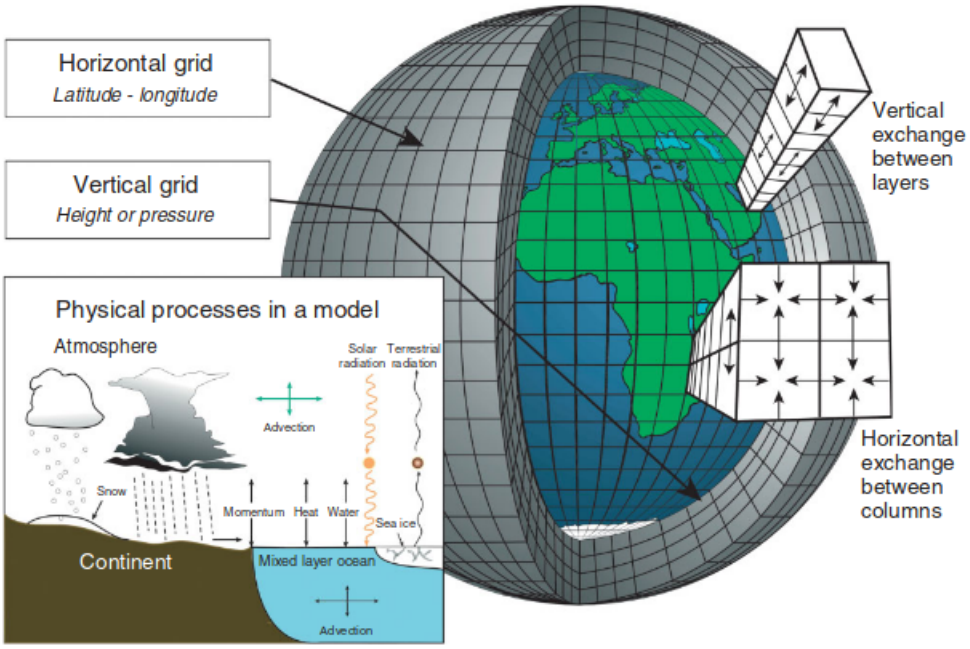

Figure 1.1: Schematic representation of the Cartesian grid structure used in models (Edwards, 2011)

Parameterization often included a set of parameters determined through empirical knowledge to help describe the unresolved processes (McFarlane, 2011). The improvement of parameterization and finding appropriate parameters is of great importance in Earth

system modelling to enhance our understanding of physical processes and their interactions in climate timescale (Krishnamurthy, 2019; Mariotti et al., 2018), which eventually will add to the predictability of climate and advance climate predictions (Alizadeh, 2022).

Table ES4. In which parameterizations do you apply changes? Full results.

	Not used 1	Occasionally 2	Frequently used 3	Total	Average rating (out of 3)
Cloud microphysical processes (e.g., droplet number concentration, conversion rates, fall velocities)	17% (4)	17% (4)	65% (15)	23	2.48
Convection (e.g., entrainment rate, conversion to precipitation)	13% (3)	35% (8)	52% (12)	23	2.39
Cloud fraction [e.g., critical relative humidity threshold or assumptions about probability density functions (PDFs)]	17% (4)	30% (7)	52% (12)	23	2.35
Cloud optical properties (e.g., subgrid inhomogeneity)	39% (9)	48% (11)	13% (3)	23	1.74
Turbulent mixing (e.g., mixing length, explicit top entrainment)	48% (11)	39% (9)	13% (3)	23	1.65
Orographics drag	35% (8)	57% (13)	9% (2)	23	1.74
Ocean physical properties (e.g., mixing, optics)	26% (6)	57% (13)	17% (4)	23	1.91
Soil and runoff properties	52% (12)	43% (10)	4% (1)	23	1.52
Vegetation properties	52% (12)	39% (9)	9% (2)	23	1.57
Snow albedo	22% (5)	70% (16)	9% (2)	23	1.87
Sea ice albedo including meltponds	26% (6)	57% (13)	17% (4)	23	1.91
Sea ice rheology	74% (17)	26% (6)	0% (0)	23	1.26

Figure 1.2: Survey results of in which parameterizations you apply changes for calibration among 23 modelling groups (Hourdin et al., 2017)

In the development of Earth system models, the process of determining uncertain parameters in parameterizations in order to minimize the deviations between observations and model outputs is known as calibration or tuning (Hourdin et al., 2017). Calibration would optimize the approximated parameters in many parameterizations, of which cloud and convection were most important and frequently targeted in calibration for Earth System Models (Figure 1.2). The calibration of ESMs generally first focused on achieving radiation equilibrium at the top of the atmosphere. Cloud and convections are processes that have sufficient effects on radiations at the top of the atmosphere (Ogura et al., 2017). In

ESMs, the cloud feedbacks on radiations mainly showed through 1) rising free-tropospheric clouds; 2) tropical low cloud amount and 3) high-latitude low cloud optical depth (Ceppi et al., 2017; Ceppi & Hartmann, 2015). Yet such processes happen in subgrid-scale so that needed to be parameterized. The representation of cloud feedback (cloud-induced radiation anomalies) in ESMs was found to have important effects on ESMs performance (Zelinka et al., 2020), and calibration on cloud and convection parameters were proven to help improve ESMs performance by reducing uncertainties in climate sensitivity and responses (Tett et al., 2022).

Previous approaches to calibration are rather resources demanding, requiring long-term simulations to make sure the parameters' effects are fully developed (generally at least after 10 years of simulations) and to explore model responses to parameterization choices and make sure the ad-hoc tuned parameters fulfill physical senses (Ogura et al., 2017; Schmidt et al., 2017). For the IPSL-CM6A-LR model alone, the calibration process took more than three years with 15 model versions developed, and thousands of years simulated by hand tuning (Mignot et al., 2021). The computational demands for calibration come from the complexity of the ESMs themselves in terms of getting model responses to a combination of parameter values. In order to simulate the Earth system, ESMs include many parameterized processes (Figure 1.3) with complex (often nonlinear) interactions (Ladyman et al., 2013), which will take a large number of parameters as input (Marrel et al., 2011) and model responses will be time-varying and take a long simulating time to reach stable stages (Zappa et al., 2020). Therefore, when performing calibration for ESMs, one would need to consider a variety of parameter combinations which heavily depend on judgements (Schmidt & Sherwood, 2015), with the expectation that longer runs will be better in optimizing model responses to parameter changes.

The demand for calibration in ESMs, for assessing parameter uncertainty and long-term model responses to parameters values would still remain important as ESMs developed, as there will still be many physical processes (e.g. shallow convection, 3D radiation effects in cloud physics) that must be parameterized even as model resolution increasing (Hourdin et al., 2023). Given this context, there are at least 23 different groups (from survey attendance in Hourdin et al. (2017)) focusing on calibration in different modelling centres across the globe. Reducing computing resources needed to acquire model responses to parameters

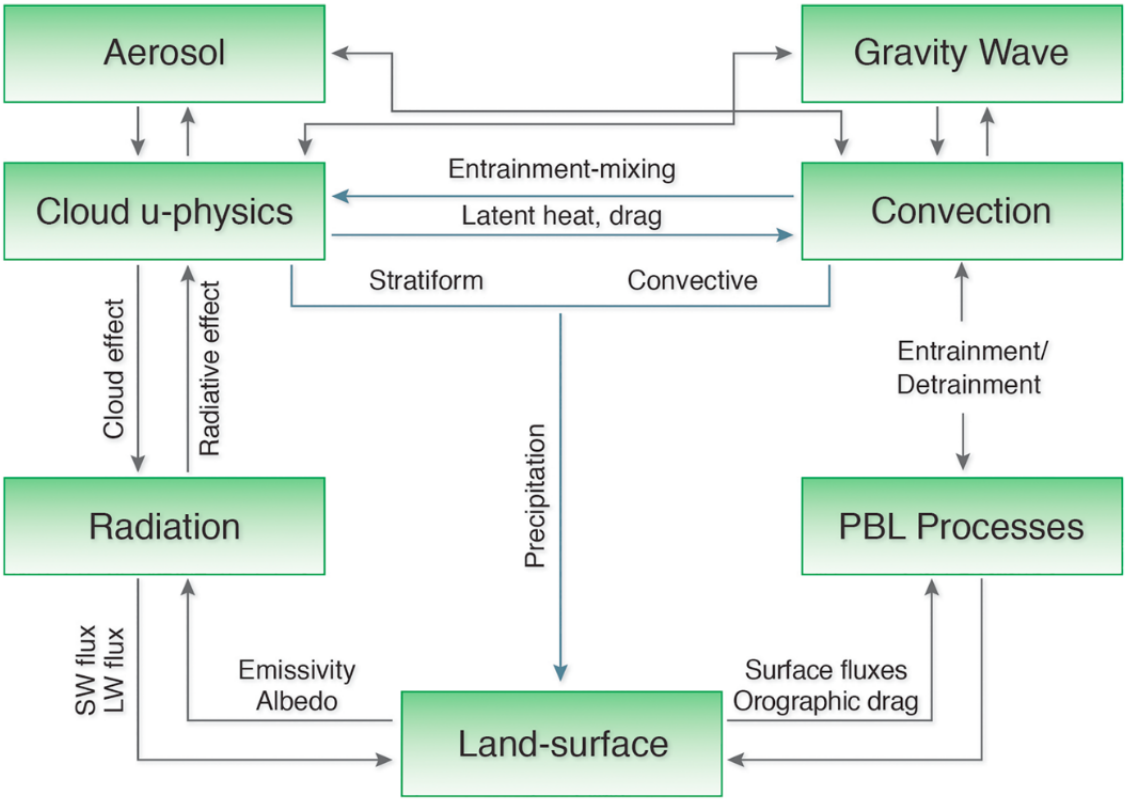

Figure 1.3: Major fast physics processes that are parameterized in ESMs (Liu, 2019).

choices would potentially help improve calibration to a more efficient and effective stage and possibly lead to less uncertainty in climate projections using ESMs. In this study, we attempted to assess current approaches in accelerating calibration in ESMs and seek to potentially improve the efficiency of calibration.

1.2 Accelerating Calibration Process in ESMs Developments

There are generally two kinds of approaches applied in order to help explore parameter space and accelerate the calibration process in the development of ESMs, and they both attempt to replace the resource-demanding long-term model simulations with something faster to get model responses given a combination of parameter values. One way to do this is by replacing long-term model simulations with statistical emulators (Figure 1.4), while the other focusing on acquiring the model responses using short-term simulations (Figure 1.5).

The first approach (Figure 1.4) uses an emulator, i.e. a statistical surrogate model, to predict the model output given input parameter values, and use the emulator to fully explore parameter space and potentially help identify better parameter values (Williamson et al., 2017). By using an emulator, the process of getting model responses to parameter choices has become faster than running large number of long-term simulations using ESMs. This approach has been widely applied in recent calibration efforts (Baker et al., 2022; Dagon et al., 2020; Hourdin et al., 2021; Sexton et al., 2021). In this approach, firstly, there would be a subset of sample parameters being perturbed in the ESM and generating a Perturbed Parameter Ensemble (PPE). This PPE contains information on the ESM responses to a sample of parameters values, which would be used to train an emulator to predict model responses to unseen parameters values. As such, the researcher could use the predicted model responses from the emulator to help make decisions in calibration and potentially make the calibration process more efficient.

The method to construct the emulator was preferred to be nonlinear, in order to help capture the nonlinear responses in the model, of which the Gaussian Process (GP) is the most widely used method in constructing an emulator for calibration in this approach. The reason for favouring GP in constructing emulators in the calibration community are 1) as a nonlinear method, it performed better in capturing potential nonlinear responses in ESMs compared with methods like multi-linear regression (Fletcher et al., 2018). The multi-linear regression method could have good overall performance predicting processes

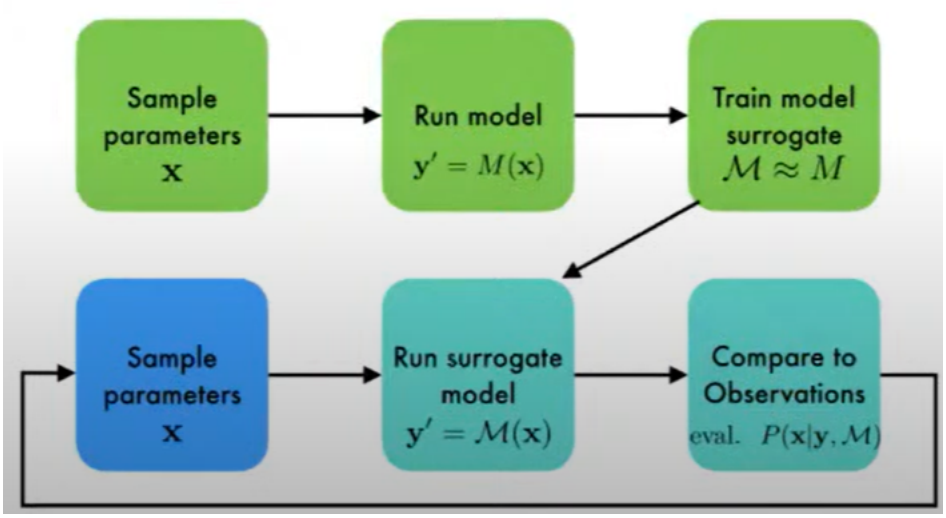

Figure 1.4: Diagram of calibration using surrogate model, credit to Dr. Gavin Schmit from presentation in AI for Good: Climate Informatics: Machine Learning for the study of Climate Change, Sep 21st, 2021.

in the Earth system, but worse than nonlinear methods in validation (Sachindra et al., 2013); and 2) it could not only predict model responses (variables) as the mean but also provide variance value as information to help quantify the plausibility of the prediction which could be useful in selecting parameter values through history matching framework (Salter & Williamson, 2016; Salter et al., 2019). The mean and associated variance values from GP emulator could help quantify the implausibility (I) of model responses given parameters:

$$I = \frac{\parallel E[m_{em}] - m_{obs} \parallel}{\sqrt{\varepsilon_m + Var[m_{em}]}} \tag{1.1}$$

where m_{obs} represents relevant observational data, ε_m is an empirical value decided by the researcher to represent deviation between the model and observations that come from model structural errors, $E[m_{em}]$ stands for the model responses (mean) predicted by GP and $Var[m_{em}]$ represents the associate variance value (Hourdin et al., 2023). The quantitative term of implausibility could make the information from emulator more useful as a metric to exclude parameter values that caused large mismatches with observations in a systematic

and effective way.

By replacing long-term model simulations with an emulator, we are able to assess a larger parameter vector and their induced responses in ESM much faster, given that emulators should require much less computational time and resources than ESMs, and eventually make the calibration of ESMs more efficient. That being said, there are still challenges applying this approach in calibration, which fundamentally affect the performance of the emulator and how large the sample size of PPE is required to train an emulator that could substantially predict model responses to parameter values. The sample PPE used to train emulator are long-term simulations, which ideally, the larger size of sample PPE provided more information and potentially help improve emulator performance. However, the computational resources required for the sample PPE (long-term simulations) were still large. The demanded size of sample PPE could easily increase given a typical calibration process of one parameterization could perturb about 20 parameters (P.-L. Ma et al., 2022) and potentially ask hundreds of ensembles in PPE to make sure parameters space is fully sampled. The demands in sample PPE for training emulators to achieve good performance become one of the challenges in model calibration.

The other approach (Figure 1.5) to accelerate the calibration process focuses on the fast responses in cloud parameterization and uses short simulations to first evaluate and rule out inappropriate parameters. The physics processes whose responses could developed in a few days (instead of years) are called "fast physics", and are often occurred in a relatively small scale which is hard to resolved by coarse resolutions in ESMs (Figure 1.3). This approach took advantage of fast physical processes like cloud and convection which could react to parameter changes in a short timescales (a few days), and the errors developed in the short timescale are related to long-term mean errors climate models (Phillips et al., 2004; Rodwell & Palmer, 2007). These characteristics of fast responses in cloud physics had been used in previous projects to help investigate long-term cloud bias using short-term simulations (H.-Y. Ma et al., 2015; Williams et al., 2013). Recently, the fast responses of cloud and convection in the model had been employed by model centres to filter parameter values in the calibration process (P.-L. Ma et al., 2022; Sexton et al., 2021). In this approach, instead of running long-term model simulations to assess parameter-induced changes, short-term simulations had been used to replace the original long-term

simulations and were compared with observations to select appropriate parameters values. The advantages of using short-term simulations to assess model responses to parameter changes help accelerate calibration massively, since short-term simulations are much faster to run and required fewer computing resources.

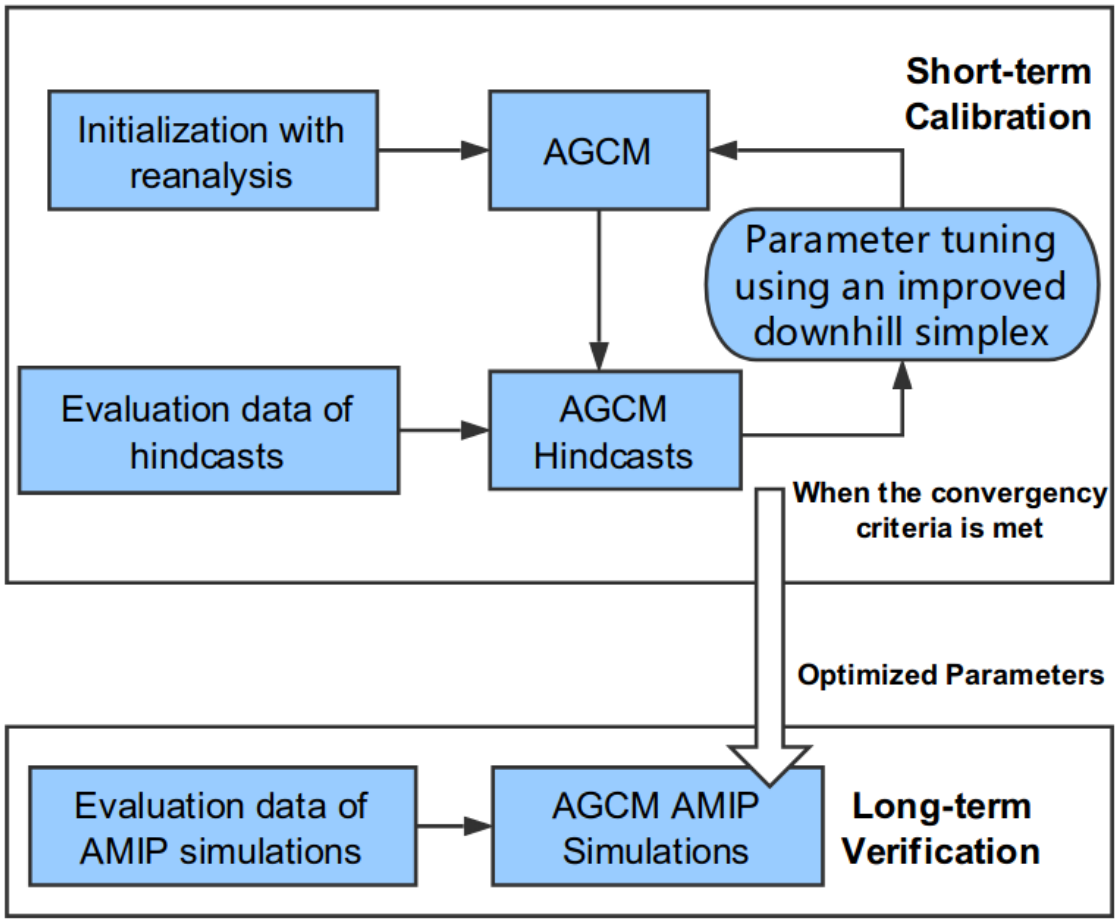

Figure 1.5: Workflow of automatic calibration of parameters using the short-term simulations (T. Zhang et al., 2018), where hindcast represents short-term simulations on the past observational time periods and could be compared with existing observational datasets.

However, apart from that short-term simulations cannot indicate model errors related to long-term climate variability, one of the remaining concerns about using short-term

simulations to replace long-term ones is the relationship between short and long-term responses in the model has not yet been fully investigated. Previous studies proposed the relationship between short and long-term were mostly based on comparing mean values or errors (deviations from observations) in model simulations. Xie et al. (2012) showed the model errors in short and long-term simulations shared similar structure and values, H.-Y. Ma et al. (2014) established the relationship between short and long based on spatial pattern correlations of the model bias and the parameter sensitivity in short and long-term simulations were found quite similar (Wan et al., 2014). Later, in the parameter sensitivity test, the changes in variables values induced by one perturbed parameter is roughly similar in both short and long-term simulations (Qian et al., 2018), and Sexton et al. (2019) found that parameters that were important in long-term simulations were very likely remained important in short-term simulations, thus short-term simulations have great potential to help calibration. Yet, although short-term simulations had been employed in recent calibration efforts to help select parameters, this idea is only supported by similarities in mean values, errors and general parameter sensitivity. Besides, short and long-term simulations are not identical everywhere, especially not in the model errors related to large-scale circulations as showed in H.-Y. Ma et al. (2021) while using short-term simulations to help evaluate cloud and convection processes. Yet there remains no clear explanation of where such difference comes from.

1.3 Motivation

Given the ongoing demand for calibration in order to help improve ESMs performance, and the consistently resources-demanding nature of current calibration processes, we intended to look for methods that could potentially help increase the efficiency of calibration in Earth system modelling development by finding efficient ways to achieve model responses given parameters values. The two main methods used to estimate model responses to parameter changes efficiently in the community both have some shortcomings, either having limited efficiency in getting required PPEs for training emulators or lack of clear understanding of similarities of model responses in short and long timescales. Therefore, in this study, we tried to help improve current approaches to estimate model responses to parameters

in a more efficient way, and help provide further investigation on the relationship between short and long-term simulations.

1.4 Research Objectives

In this study, we use short-term simulations to replace long-term simulations as training data for an emulator in order to predict model responses given unseen parameters values. Since short-term simulations are faster to run compared with long-term simulations, this approach partially solves the time-consuming process to get the required PPE for constructing an emulator. We will also further investigate the relationship between short and long-term simulations and whether the model responses to parameter changes in the short and long term were similar.

The following research objectives will be investigated in this study:

1. Quantify the relationship between short- and long-term simulations and compare model responses to parameter changes in short and long-term timescales
2. Evaluate the performance of the emulators trained with short- and long-term simulations

1.5 Structure

The next chapter will document the details in model simulations and applied methods in this study. The attempts and findings when trying to use short-term simulations to train emulators to predict long-term model responses will be illustrated in Chapter 3. The investigation of the relationship of model responses to parameter changes between short and long-term simulations will be described in Section 3.3.1, and the evaluation of emulators trained with short and long-term simulations will be included in Section 3.3.2. Finally, the research findings and limitations will be discussed in Chapter 4.

Chapter 2

Data and Methods

2.1 Model Simulations

2.1.1 Model Configurations

The model perturbed in this research is the atmospheric component of Community Earth System Model Version 1.2.2 (CESM1) and the Community Climate System Model Version 4 (CCSM4), known as the National Center for Atmospheric Research (NCAR) Community Atmosphere Model Version 4 (CAM4), documented in Gent et al. (2011). The model ensembles were constructed as the Atmospheric Model Intercomparison Project (AMIP, Gates et al. (1999)) runs for CMIP5 protocol. The AMIP run is a standard experimental protocol for the global atmospheric model (Gates et al., 1999). In this model setting, the atmospheric is constrained by a prescribed realistic sea surface temperature and sea ice data from 1979 to the near present, which allowed users to focus on the atmospheric component without the complexity of ocean-atmosphere feedbacks and are widely used in calibrating cloud and convection processes in ESMs. In the calibration of ESMs, model centres would generally focus on achieving global energy equilibrium at the top of the atmosphere, with particular interests in cloud processes and its radiative forcing (Hourdin et al., 2017), and relevant interesting parameters could be deep convection for clouds, humidity threshold for clouds and consumption rate of convective available potential energy (T. Zhang et al.,

2018). All those parameters are related to the formation of clouds at different levels, the stability in the atmosphere and the possibility of convection. As such, the favourite variables for calibration could be shortwave and longwave cloud forcing (directly showed the cloud radiative effects), precipitation and cloud fraction (related to cloud and convection processes) in the calibration of CESM (Danabasoglu et al., 2020; Hannay, n.d.).

In this study, we generated our PPEs following AMIP protocols and perturbed 5 parameters related to cloud and convection in CESM (Table 2.1). The perturbed values of parameters were sampled using Latin Hypercube Sampling to get a near random sampling on multivariate dimensions where values of parameter is evenly distributed across the space. The two PPEs used in this study (long and short-term) have the same model configuration (i.e. same model version, following AMIP protocols and same perturbed parameters values), and only differed in their simulation time (15 years or 5 days) and the initial conditions.

No.	Parameter Name	Min	Default	Max	Description
x1	cldfrc_rhminl	0.80	0.91	0.99	RH threshold for low cloud formation
x2	cldopt_rliqocean	8.4	14.0	19.6	Radius of liquid cloud droplets over ocean
x3	hkconv_cmftau	900	1800	14 440	Timescale of shallow CAPE consumption
x4	cldfrc_rhminh	0.70	0.80	0.90	RH threshold for high cloud formation
x5	zmconv_tau	1800	3600	28 800	Timescale of deep CAPE consumption

Table 2.1: List of parameters that are perturbed, including parameter name, default and range of perturbed values. RH stands for relative humidity; CAPE stands for Convective Available Potential Energy

2.1.2 Initial Conditions

The idea of using short-term simulations comes from Phillips et al. (2004) and Rodwell and Palmer (2007) where using numerical weather prediction to assess climate model parameterizations. The essence of this approach is that if the large-scale state in short-term simulations remains close to the state of which we need to compare (in their case observations, in this study the state in long-term simulations), the main departures from

the large-scale state are mainly due to effects of parameterizations (Sexton et al., 2019). Since the responses of fast physics (Figure 1.3) can develop in a few days and the long-term responses from climate variability have not yet developed, under this design, if initial conditions are close to the large-scale state, the main departures should mostly credit to responses from fast physics. Therefore, a lot of discussions of how to construct appropriate short-term simulations so that the forcing from parameters contributes to the main differences were focused on how to acquire the initial conditions that are close to the large-scale state to be compared with.

Previously, studies focused on assessing model errors compared with observational data, and the initial conditions were generated to be close to observations. At first, people used full numerical weather predictions, with data assimilation so that the initial conditions are close to observations (Phillips et al., 2004; Rodwell & Palmer, 2007). Then Klocke and Rodwell (2014) found that if a data assimilation system is not available, using Transpose-AMIP method (Williams et al., 2013) for initial conditions could also diagnose the model errors in short-term simulations. In the Transpose-AMIP project, the initial conditions for each field in model simulations were either initialized from a climatology state (model's AMIP run) or initialized using a nudge simulation (nudge to observations and reanalysis data) (Boyle et al., 2005). This approach of generating initial conditions was then employed in Cloud-Associated Parameterizations Testbed (H.-Y. Ma et al., 2015) to assess errors in cloud parameterizations using short-term simulations. The detailed protocols and relevant code in the CAPT project could be found in the Program for Climate Model Diagnosis and Intercomparison funded by Lawrence Livermore National Laboratory.

In this study, we do not intend to compare with observations to assess model errors, instead, we would like to focus on the relationship between short and long-term responses in model simulations. Therefore, we have an advantage in that we can actually get the large-scale state in long-term simulations through AMIP run, and use it as initial conditions for short-term simulations. The nudge simulations which are intended to make initial conditions closer to observations are not necessary in this case. The initial conditions for long-term simulations are the default mean climate state (Neale et al., 2013) and the initial conditions for short-term simulations are obtained from corresponding previous one year AMIP run using the same model configurations. The initial conditions for this one year

AMIP run are also default, the one year lead simulation time could be referred to H.-Y. Ma et al. (2021) where their initial conditions for multi-year short-term simulations also obtained from simulations with one-year lead time. The short-term simulations used in this study are not multi-year but with four start dates in the first date of January, April, July and October. So for each start date, its initial conditions were obtained from the AMIP run after at least one year simulation. The goal to generate initial conditions like this for short-term simulations is to make its initial conditions as close to the large-scale state in long-term simulations (the ground truth in this study, the one we compare short-term simulations with) as possible. Under this design, the long-term simulations, long PPE, represent the model responses to parameter values. And short-term simulations, with initial conditions obtained from corresponding AMIP run, similar to the large-scale state in long-term simulations. The main departures from its initial state, would be largely attributed to parameters effects.

The short and long-term simulations were run on the high-performance supercomputer 'Niagara' owned by the University of Toronto and operated by SciNet, and were run in large parallel jobs on multiple CPUs. One node per simulation for short-term simulations, four nodes for long-term simulations, and each node has 40 CPU cores. On Niagara, each physical core becomes two virtual cores, so each node has 80 cores to be run in one job. The computing resources for running long-term simulations took about 7 core years, which was more than 6000 CPU hours. The computing resources for short-term simulations and their initial conditions took only 0.2 core years (less than 2000 CPU hours).

2.2 Empirical Orthogonal Function

We employ Empirical Orthogonal Function (EOF, or principal component) analysis to assess the model responses to parameter changes. EOF analysis is commonly used in climate science research, to assess spatial patterns related to climate variability and how they vary across time (Z. Zhang & Moore, 2015). However, in this study, instead of performing EOF on the time dimension, we perform on the dimension where indexing different model ensemble members in a given PPE. In a PPE, each PPE member only

differs in the perturbed parameter values, and the model outputs from each PPE member represent model responses to given parameters values. Consider a dataset from one of our PPE: Y, as a matrix with the size of $m \times n$, where m represent the dimension where indexing different model ensemble members and $m = 50$, while n represents the stack of every else dimension (i.e. variables, time, latitude and longitude dimensions) in the PPE. First standardized the data and obtained a standardized matrix A. The covariance matrix will then be $R = A^T A$. Solving eigenvalue problem for covariance matrix R:

$$R - \Lambda I = 0 \tag{2.1}$$

$$RV = V\Lambda \tag{2.2}$$

where the eigenvalues are on the diagonal of Λ, each eigenvalue divided by its sum represents the explained variance of the corresponding EOF pattern, and the EOF patterns are the eigenvectors on each column of V. Therefore, by performing EOF on the PPE member dimension, each EOF pattern in our EOF analysis represents the spatial patterns related to the general model responses to parameter changes. Correspondingly, for each EOF pattern, the Principal Component (PC) timeseries, $P = AV$, in this case, represent the contributions of the given EOF pattern in each model ensemble member.

2.3 Internal Variability

In this study, we assess the internal variability in our short and long-term simulations. The internal variability is related to natural fluctuations in the Earth system and was found to be relatively important in shorter timescales (Hawkins & Sutton, 2009). The uncertainty from internal variability is calculated as the mean of the standard deviation across ensemble runs. We calculate the fractional uncertainty represented as the coefficient of variance, i.e. the predicted uncertainty divided by the expected mean change:

$$\frac{\overline{\sigma}_{time}(y)}{\mu(y)} \tag{2.3}$$

where y represents model output variables. Since we don't have multi-year short-term simulations, when calculating internal variability in short-term simulations, the time dimension

consists of the daily mean of the four start dates in January, April, July and October. Correspondingly, for long-term simulations, the monthly mean values for the above months were selected, and its time dimension consist of the monthly mean of the above months in the last 5 years of long-term simulations. In this case, although inevitably consist of seasonal variations in this calculation, the internal variability is still comparable between short and long-term simulations since the seasonal variations in long-term simulations are also included.

2.4 The Gaussian Process Emulator

We used the Gaussian Process Emulator as a statistical model to predict model outputs given parameters values. The emulator was built using previously ran PPEs as training data with the assumption that the ESM responses to parameters were smooth and that each ensemble in the PPE gave some information about model responses to different parameters values.

Consider we have input x as a vector of five, containing perturbed values for the five selected parameters, and y represents the values of model output variables. We first standardized the model outputs and consider $y = f(x)$ where f stands for the function representing our model. Since we intend to focus on model responses given parameters values, the model outputs from our PPEs represent the model responses to parameter values without noise, so $y = f(x)$ without noise term. $f(x)$ could be specified by a mean function matrix $m(x)$ and a covariance function matrix $k(x, x')$ following:

$$m(X) = \mathbb{E}[f(X)] \tag{2.4}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))] \tag{2.5}$$

where X is a matrix of parameters values used for our PPE, and $f(X)$ is the matrix of model outputs from the PPE, corresponding with a vector of parameter values x in X. Here x and x' are both a vector in X. This is a representation of a Gaussian Process, and could also be expressed as:

$$f \backsim \mathcal{GP}(m(x), k(x, x')) \tag{2.6}$$

We called $f \backsim \mathcal{GP}$ the prior. The model PPE paired with corresponding parameters values consisted of our training dataset: D_n, where $D_n = (X_n, Y_n)$ consisting of n samples paired with parameters values and model outputs: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

By using the GP emulator to predict model responses given unseen parameter values, what we would like to calculate is $f(x*)|D_n$, i.e. the posterior, where $x*$ represents unseen parameters values and $f(x*)$ is the model outputs using $x*$ as parameter values. The posterior $f(x*)|D_n$ is a conditional distribution of $f(x*)$ given data D_n. We considered $f(x*)$ and $f(X)$ have a joint multivariate normal distribution, with zero means (since we standardized our data in prepossessing):

$$\begin{bmatrix} f(x*) \\ f(X) \end{bmatrix} \backsim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} k(x*, x*) & k(x*, X) \\ k(X, x*) & k(X, X) \end{bmatrix} \right) \tag{2.7}$$

following basic properties of multivariate Gaussian distributions (Gramacy, 2020) we have:

$$p(f(x*)|x*, X, f(X)) \backsim \mathcal{N}(k(x*, X)k(X, X)^{-1}f(X), k(x*, x*) - k(x*, X)k(X, X)^{-1}k(X, x*)) \tag{2.8}$$

which could also be interpreted as:

$$p(f(x*)|D_n) \backsim \mathcal{N}(m(x*), \sigma^2(x*)) \tag{2.9}$$

$$m(x*) = k(x*, X)k(X, X)^{-1}f(X) \tag{2.10}$$

$$\sigma^2(x*) = k(x*, x*) - k(x*, X)k(X, X)^{-1}k(X, x*) \tag{2.11}$$

where $m(x*)$ is the mean value predicted by the GP emulator using D_n as the training dataset, and $\sigma^2(x*)$ is the variance function representing the uncertainty, the smaller $\sigma^2(x*)$ indicating less uncertainty for the predicted point. And noticing that $\sigma^2(x*) < k(x*, x*)$ so we do learn something from the training dataset D_n and the variance decreased and uncertainty decreased from the original.

Particularly, in this study, we use Squared Exponential Kernel (also known as the Radial Basis Function kernel, RBF) for the covariance function $k(x, x')$. This "kernel trick" means we replace the reoccurring inner products in the covariance matrix with a simple kernel function, corresponding to a large or infinite set of basis functions and makes the

actual computation more convenient than working on the covariance matrix (Rasmussen & Williams, 2006). The RBF kernel is expressed as:

$$k(x, x') = \alpha^2 exp(-\frac{(x - x')^2}{2l^2}) \tag{2.12}$$

where α is a scale factor, representing the range of the data variance away from its mean; and l is lengthscale, representing the scale of correlation between two data points (how far away to data points need to be uncorrelated) in the training dataset.

Therefore, we have in total of four hyperparameters in our GP emulator: mean function m, variance σ^2, and the scale factor α and lengthscale l in RBF kernel. And we optimize these hyperparameters by maximizing log marginal likelihood:

$$\log p(f(X)|X) = -\frac{1}{2} f(X)^T k(X, X)^{-1} f(X) - \frac{1}{2} \log |k(X, X)| - \frac{n}{2} \log 2\pi \tag{2.13}$$

By maximizing log marginal likelihood, we made sure the final hyperparameters in our GP emulator lead to a GP that could best represent the probability distribution of our training dataset (fit to training data).

The above was the statistical explanation of the Gaussian Process emulator used in this study. We assume the predicted model outputs $f(x*)$ and the existed outputs obtained from PPE runs $f(X)$ sharing a joint multivariate distribution and use Gaussian Process to estimate the model outputs given unseen parameter values $x*$. The results includes a posterior mean function (2.10) which represents the estimation of model outputs using GP emulator, and a posterior covariance function (2.11) which represents emulator uncertainty. The 95% confidence integrals for the emulator prediction could be quantified following:

$$m(x*) \pm 1.96 \times \sigma(x*) \tag{2.14}$$

We note that Gaussian Process is not the only statistical methods used for emulator in calibration. There are general three types of methods employed in emulation: Gaussian Process, Neural Network and random forest. Gaussian Process was found to have exceptional predictive power in emulation, compared with random forest (Watson-Parris et al., 2022); yet Gaussian Process and Neural Network seemed to have similar accuracy, and both being applied to construct emulator in calibration. The Gaussian Process, in general,

is more popular for building emulator in calibration. Apart from the history heritage from the Bayesian calibration framework in Kennedy and O'Hagan (2001), the Gaussian Process provides a variance function to help quantify the uncertainty (1.1) and select parameter values. Besides, when testing methods for emulator in calibration cosmology simulations, researcher found that although Gaussian Process and Neural Network both processed similar accuracy, the uncertainty for prediction using Neural Network is substantially large (Figure 2.1). The Gaussian Process emulator produces narrower posteriors, indicating more confidence in its predictions.

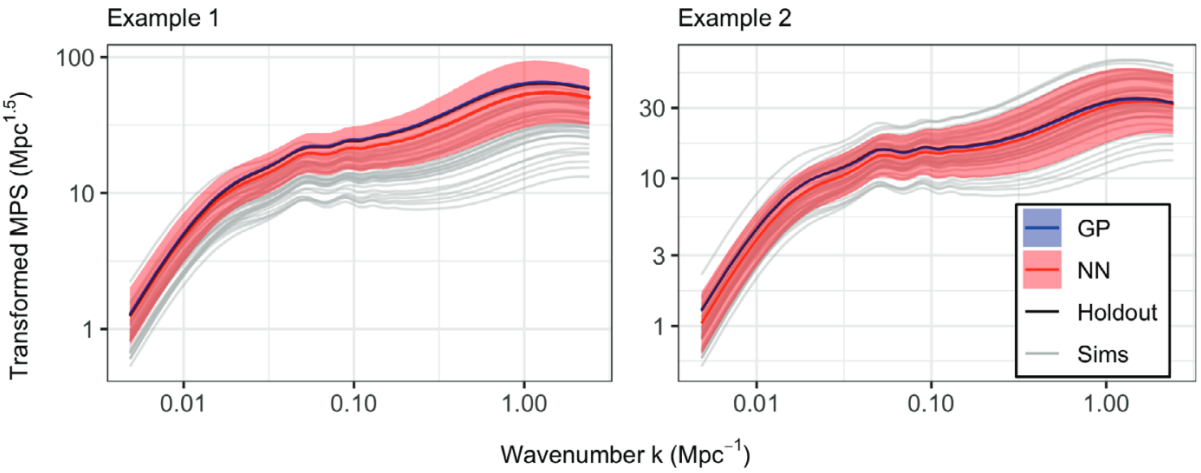

Figure 2.1: Two examples from the Leave-One-Out cross-validation results for the cosmology data set (Myren & Lawrence, 2021). The light gray lines are the sample of the simulated cosmology data containing all 36 samples from the training set. The blue and red shaded areas indicate the 95% confidence intervals for each emulator. The 95% confidence intervals for GP were calculated using its mean and variance function (2.14), while for NN the 95% confidence intervals were calculated following Gal et al. (2017). "Holdout" stands for the selected sample from the training dataset for testing in Leave-One-Out cross-validation, here referred to the "ground truth".

Chapter 3

Using Short Simulations for Calibration in Earth System Models

3.1 Overview

Calibration about parameter values is the key step in the development of Earth System Models (ESMs) and could potentially help quantify parameter uncertainty, improve ESMs performance and eventually better climate projections for human beings (Tett et al., 2022). Recently there is an advocate for the need to quantify parameter uncertainty and reduce the number of scenarios and protocols in CMIP practice (Hourdin et al., 2023) and assessment of Perturbed Parameter Ensembles (PPEs) along with careful calibration processes could potentially help quantify model responses to different parameter values, quantify parameter uncertainty and further to improve the performance of ESMs. Although calibration is and will still be an important process when developing ESMs, this process was often left on hand-tuning parameters in an ad-hoc manner, requiring numerous computing and human resources to find appropriate parameter values to minimize deviations between model and observations (Schmidt et al., 2017).

Previous studies sought to find a more systematic and efficient way to better calibrate the parameters, they either tried to use a statistical emulator to predict model responses

given parameter values (Dagon et al., 2020; Hourdin et al., 2022) or select parameters based on model responses developed within a few days (Sexton et al., 2021; T. Zhang et al., 2018) to help make the process more efficient. The first approach relied on whether the statistical emulator could successfully reproduce model responses given unseen parameter values, which were proven possible and using mean and variance values generated by Gaussian Process could further help quantify and uncertainty that came from the emulator and help researchers make better decisions in filtering parameters (Williamson et al., 2017). The second approach was based on the relationship parameter-induced responses in model simulations between short and long-term timescales (Phillips et al., 2004; Rodwell & Palmer, 2007). The theory is that if using the to-be-compared mean large state as initial conditions, in a rather short simulating timescale, the large-scale state of the model remained very close to the large-scale state, and any departures from it would be resulted from different parameter values (Sexton et al., 2019). For 'fast' physical processes like cloud and convictions, previous studies found that their responses to parameters in a short timescale were similar to the model response in the long-term (Qian et al., 2018; Wan et al., 2014; Xie et al., 2012). Given short-term simulations seemed to have similar responses to parameters as the long-term simulations, and it generally took much less time to run short-term simulations, modelling centres also used this characteristic to filter parameter value in calibration (H.-Y. Ma et al., 2021; Sexton et al., 2021).

Yet there are still remain challenges for these two methods in accelerating the calibration process for ESMs. To achieve emulator that could successfully predict model responses to parameters, there should already be a large size of Perturbed Parameter Ensemble (PPE) which included the information of model responses given different sample parameters values as training dataset. And the computational resource to get such a size of PPE could be a lot, which potentially slowed down the calibration using emulator. On the other hand, although short and long-term simulations were similar given the same parameters to some degree, there remained no direct evidence showing that model responses to parameters in short-term simulations could be a good indicator for that in long-term timescale (Sexton et al., 2019). And when applying short-term simulations to help assess cloud processes in a climate model, there are some differences regarding large-scale circulations found (H.-Y. Ma et al., 2021), yet there was no further explanation of where such differences came from.

Given the need to accelerate the calibration process for the development of ESMs, and current methods to do this either lack of computational efficiency in acquiring model responses to parameters or lack of detailed investigations of whether those applied responses are valid. We would like to further investigate the relationship between short and long-term simulations in model responses to parameters and find a new way to acquire model responses to parameters efficiently by using short-term simulations for training emulators. Our hypotheses in this study are that 1) short-term simulations could be used as a relatively cheaper surrogate in training emulators for calibration; 2) the emulator trained by the short-term simulations could predict the long-term model responses to parameter changes with high fidelity; 3) this method substantially increases the efficiency of training an emulator to be used for calibrating an ESM. The model configurations, the construction of the emulator and evaluation metrics will be described in Section 3.2. The mean state of the short and long-term simulations, spread and responses to parameter changes will be compared in Section 3.3.1, and the potential relationship and differences between short and long-term simulations will also be explored. Section 3.3.2 will evaluate the emulator's performance and assess whether short-term simulations could be helpful in training emulators for calibration. Finally, we will provide a summary and conclusion for our purposed hypothesis in Section 4.

3.2 Method

3.2.1 Perturbed Parameter Ensembles (PPEs)

The model perturbed in this research is the atmospheric component of Community Earth System Model Version 1.2.2 (CESM1) and the Community Climate System Model Version 4 (CCSM4), known as the National Center for Atmospheric Research (NCAR) Community Atmosphere Model Version 4 (CAM4), documented in Gent et al. (2011). The model ensembles were constructed as the Atmospheric Model Intercomparison Project (AMIP, Gates et al. (1999)) runs for CMIP5 protocol, which included atmosphere model CAM4 and the Community Land Model Version 4.0 (CLM4) with a biogeochemistry model that simulates the carbon and nitrogen cycling (Thornton et al., 2007). All model simulations

were performed at 1.9×2.5 degree resolution with 26 vertical levels, which result in a grid size of 96×144 grid cells across the globe in each time step.

There are five perturbed parameters in the long-term simulation. These five parameters are parameters related to clouds and convection in CAM4 (Table 3.1). These parameters were found highly important in sensitivity analysis (Covey et al., 2013) and have been used in previous study to construct PPEs for training emulators (Fletcher et al., 2022) and were found to explain a large fraction of total variance for multivariate skill score for quantifying climate model performance (Fletcher et al., 2018). In long-term simulations, there are 50 combinations of parameter values (x1-x5) selected by Latin Hypercube Sampling (Stein, 1987) to ensure a good distribution of sampling values and potentially help improve the trained emulator performance (Urban & Fricker, 2010). The long-term simulation has a 15-year length of simulation time and the last five years of model outputs (long PPE) will be employed to analyze long-term model responses to parameters.

No.	Parameter Name	Min	Default	Max	Description
x1	cldfrc_rhminl	0.80	0.91	0.99	RH threshold for low cloud formation
x2	cldopt_rliqocean	8.4	14.0	19.6	Radius of liquid cloud droplets over ocean
x3	hkconv_cmftau	900	1800	14 440	Timescale of shallow CAPE consumption
x4	cldfrc_rhminh	0.70	0.80	0.90	RH threshold for high cloud formation
x5	zmconv_tau	1800	3600	28 800	Timescale of deep CAPE consumption

Table 3.1: List of parameters that are perturbed, including parameter name, default and range of perturbed values in CAM4. RH stands for relative humidity; CAPE stands for Convective Available Potential Energy

The short-term simulations were generated using the same model configuration and perturbed parameters values as the long-term simulations above. The short-term simulations only differ from long-term simulations in simulation length (5 days for short-term simulations) and initial conditions. The initial conditions for short-term simulations were created following the workflow in the Cloud-Associated Parameterizations Testbed (CAPT) project (H.-Y. Ma et al., 2013) to generate initial conditions for hindcasts experiments. We noted that in the CAPT documentation, there were two methods for obtaining the necessary variables for generating initial conditions: 1) performing a nudge simulation;

2) obtaining from a previous AMIP run. We used the second method to generate initial conditions for short PPE, obtaining all necessary variables from previous AMIP runs. The reason is that we do not intend to perform a full calibration process in this study and our short-term simulations will only be used to assess the model responses in a short timescale, but not compared with observation. Therefore, making initial conditions for short-term simulations from previous AMIP runs would make sure the initial state of the model remains close to the large-scale state in the long-term simulations, and the main departures in short-term simulations can be taken as responses given different parameter values, which helps compare the model responses in short-term simulations with long-term. Without using nudged simulation to generate initial conditions, we acknowledge that our short-term simulations are technically not hindcast experiments. By obtaining the necessary variables from previous AMIP runs, the initial conditions could be easier generated and potentially save computational resources for additional nudge simulations. There are four start dates in short-term simulations, the first date of January, April, July and October. The short PPE is an average of short-term simulations with four start dates. The average approach is employed to eliminate the potential seasonal differences resulting from different start dates.

We selected four variables from model simulations outputs to assess model responses to parameter changes: shortwave cloud forcing (SWCF), longwave cloud forcing (LWCF), precipitation (PRECT) and total cloud fraction (CLDTOT). These variables were selected to reflect changes related to cloud and convictions and potential impacts on radiation balance due to clouds. SWCF and LWCF are related to the radiation impact of cloud changes. These two variables had been linked with opaque cloud cover and effective cloud albedo to quantify the cloud's impact on atmospheric cycle (Betts et al., 2015; Betts et al., 2013) and are important variables that described cloud effects on Top-of-Atmosphere radiation, both are key values in tunning practices of Earth system model (Schmidt et al., 2017). Precipitation and total cloud fraction were chosen to represent general changes in clouds and convections.

3.2.2 Empirical Orthogonal Function Decomposition

To assess major variance across different model responses to parameter changes, we employed Empirical Orthogonal Function (EOF, or principal component) analysis. In this study, instead of performing EOF analysis on the time dimension, we performed EOF on the dimension which indexing different model ensemble members in the given PPE. Therefore, the EOF patterns obtained could represent the spatial patterns related to the general model responses to parameter changes. Consider a dataset from one of our PPEs as:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{bmatrix} \tag{3.1}$$

where m stands for the dimension of PPE members, in our case $m = 50$; and n has the size of multiplication every other dimension from the PPE, $n = len(var) \times len(time) \times len(lat) \times len(lon)$, $len(*)$ represent the dimension size of each selected dimension. $len(var) = 4$, $len(lat) = 96$, $len(lon) = 144$ and $len(time)$ is 121 for short PPE (hourly data for five days), 60 for long PPE (monthly data for the last five years). We first standardized the original matrix and got a standardized matrix A so that all four variables contribute equally to the analysis. To perform EOF on PPE member dimension (m), we computed the covariance matrix $R = A^T A$ and solve eigenvalue problem:

$$RV = V\Lambda \tag{3.2}$$

where each column in V, with a size of n, is an eigenvector (EOF patterns), representing the corresponding variance for all four variables in the PPE across different ensemble members; and the diagonal of Λ are the eigenvalues (indicating EOFs explained variances after divided by the sum of all eigenvalues). The PCs were computed by projecting standardized matrix A onto EOFs: $P = AV$.

By performing EOF on PPE member dimension, the EOF patterns we obtained would then describe the common variability across model ensemble members shared by the four selected variables and the first EOF pattern will be the major model response to parameter

changes in a given PPE. The relevant PC would represent how each model ensemble member varied to each EOF pattern, and the first PC, in a way, represents the degree to which each ensemble member contributed or varied in responses to major parameter-induced changes.

3.2.3 Emulation

We used Gaussian Processes (GP) emulator in this study to emulate the model response given parameters changes. Gaussian Processes are probabilistic models (Rasmussen & Williams, 2006) which have been widely employed in geoscience to capture the nonlinear relationships among data. In machine learning terms, the process of emulation in this study is a supervised learning where GP will take any finite collection of model outputs given different parameters values as a multivariate normal distribution. The training dataset was set up using the perturbed values of parameters and the selected model outputs variables described in Section 3.2.1. The perturbed values of parameters in Table 3.1 will be considered the input, connected with the selected model outputs variables filed from each model ensemble member respectively. Together, these are constructed into our training dataset for the GP emulator. All the input values and selected output variables were standardized using the mean and standard deviation of training data before feeding into GP emulators. We used open open-source Python library GPFlow (Matthews et al., 2017) for setting the GP emulator used in this research. Fitting a GP includes fitting a mean function and a covariance function. In this study, the GP emulator was set with a constant zero mean prior, and the radial basis function kernel (RBF kernel) was employed for the input. The RBF kernel is universal and commonly used for smooth and continuous responses. There were 4 hyperparameters in our GP emulator: mean value, variance the lengthscale and the scale factor for the RBF kernel. These 4 hyperparameters for the GP emulator were jointly tuned against the training dataset through marginal likelihood maximization using the limited memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization algorithm, employed through the SciPy package (Virtanen et al., 2020).

To investigate whether short-term simulations could help make the calibration process more efficient, two GP emulators were trained in this study using short-term and long-term

responses as training datasets, extracted from short PPE and long PPs correspondingly. The short-term responses are determined as the selected model output variables from short PPEs, averaged across the 5 days of simulation time, while the long-term responses are determined as the selected model output variables averaged across the last 5 years of model simulation time. The short-term and long-term responses, paired with the perturbed parameters' values act as the training dataset for each GP emulator correspondingly. If short-term simulations have a strong relationship with long-term simulations, short-term simulations carry similar model responses to parameters changes as that in long-term simulations. Then in hypothesis, the GP emulator trained with short PPEs should have a similar performance as the emulator trained with long PPEs, thus making short-term simulations good substitutes for long-term simulations, help save computational time running long-term simulations and eventually improve the efficiency of calibration on Earth system models.

The Leave-One-Out method was used to validate our GP emulators, following McNeall et al. (2020). When training each emulator, remove one of the ensemble members from the training PPE successively, and use the rest of the PPE to train the GP emulator. The trained emulator would be employed to predict the model output variables from the removed ensemble member and thus lead to an emulated PPE the same size as the training PPE.

The GP emulator will be first examined by calculating the spatial skill metric (SS):

$$SS = 1 - \frac{MSE(X, Y)}{MSE(\overline{X}, Y)} \tag{3.3}$$

$$MSE(X, Y) = \frac{1}{C \times M \times N} \sum_{c=1}^{C} \sum_{i=1}^{M} \sum_{j=1}^{N} (X_{cij} - Y_{cij})^2 \tag{3.4}$$

where C, M, N represent the number of ensemble members, the number of grid points in latitude and longitude correspondingly. Thus $MSE(X, Y)$ stands for mean-square-error for each selected variable. After normalization, if the prediction by an emulator is identical to model outputs, then $SS = 1$, and if the emulator captures no spatial variability of the model outputs $SS = 0$. In addition, we also employed the Taylor diagram to assess

emulator performance (Taylor, 2001) showing the quantification of correlation, root-mean-square-error and standard deviation between prediction from emulators and model outputs.

3.3 Results

3.3.1 Relationship between short- and long-term simulations

To first investigate the relationship between short and long-term simulations, we compared the global mean values of selected model output variables from short and long PPE (Figure 3.1). Similar to previous studies (H.-Y. Ma et al., 2014; Sexton et al., 2019), all selected variables shared similar global mean values and spread between short and long-term simulations. The global mean values of each variable in each ensemble member showed a clear positive linear relationship. Particularly, for SWCF, short and long-term simulations showed a significant one-to-one linear relationship (with a correlation coefficient of 0.9990), while for other variables, although a significant linear relationship was found, their correlation was less strong compared with SWCF.

In order to assess the model responses given parameters changes in each PPE, we employed an EOF analysis on the PPE member dimension for model simulation data. Given that all variables showed a linear relationship in Figure 3.1, a linear transformation for model simulations data to new coordination (EOF) where its variance across different model ensemble members shall be shown. Figure 3.2 indicating the first EOF pattern for long-term simulations (Figure 3.2 left column (a,c,e,g)) and the first EOF pattern for short-term simulations (Figure 3.2 right column (b,d,f,h)), which have a very similar spatial pattern. The similarity in EOF patterns showed that the perturbed parameters values caused roughly similar spatial variance in both short and long-term simulations. Both EOF patterns indicated a clear similarity over tropical, especially along South America, Atlantic and Africa. However, it is noticeable that the first EOF pattern from short-term simulations shows more fragmented components over high-latitude areas and over the Pacific, indicating the model responses in short-term simulations had large differences compared with long-term simulations in the above-mentioned area. We also compared the

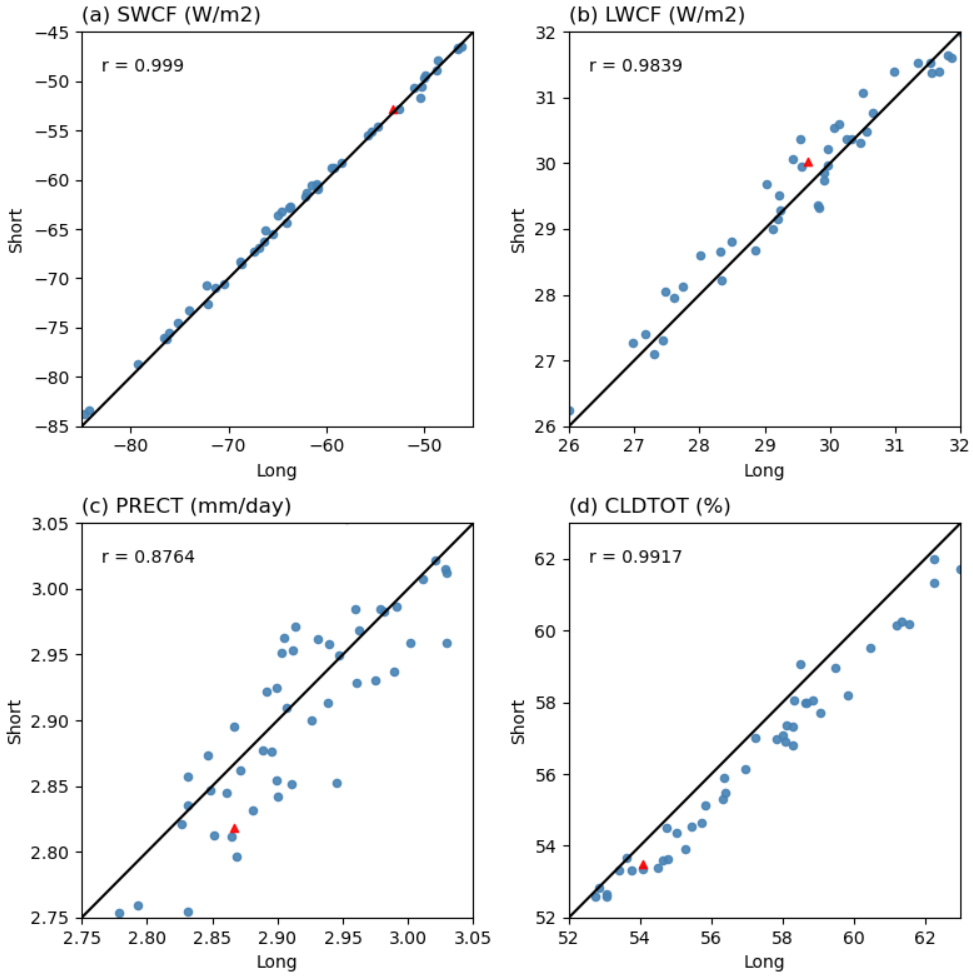

Figure 3.1: Scatter plot of global-mean values of each variable (SWCF, LWCF, PRECT and CLDTOT) in short and long PPEs. Each blue dot denotes the value from each model ensemble member, and the red triangle represents the variable value from the model simulation using the default values of the perturbed parameters. The r value represents the correlation coefficient between variable values from long and short PPEs. The black linear lines are the one-to-one diagonal lines that data would follow if they have a complete one-to-one relationship.

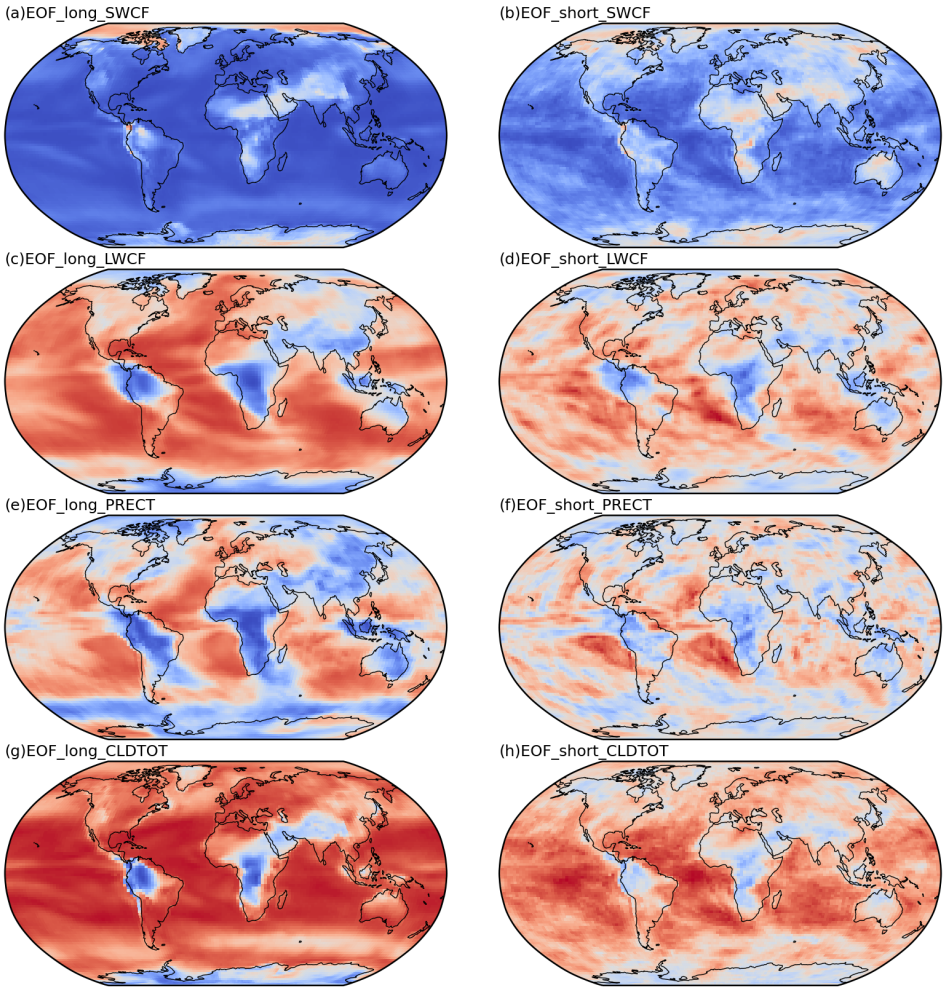

Figure 3.2: The first EOF pattern for model simulations data. The column on the left represents the first EOF pattern for long-term simulations, while the column on the right indicates the first EOF pattern for the short-term simulation. Each EOF pattern was the eigenvector obtained from solving the eigenvalue problem for the covariance matrix of model data, which could be expressed as the covariance between the first principal component (PC1) and the PPE members of the model data for each variable at each grid point.

PCs for the first three EOF patterns in short and long-term simulation, shown as a scatterplot in Figure 3.3. Given that our EOF analysis was performed on the PPE members' dimension, therefore, unlike EOF on the time dimension, our PC values do not have an order like time. Therefore, the PC values were shown as individual points, not following any pre-defined order. Figure 3.3 (a) is a scatterplot showing the relationship between the first PCs from short (y-axis) and long-term (x-axis) simulations. The PC values for each ensemble member had a highly correlated linear relationship between short and long simulations (r=0.9877), indicating that the contributions of each ensemble member toward the major parameter-induced variance are similar, or the model responses given parameters in each ensemble member are similar in short and long-term timescales. However, the explained variance of the first EOF patterns were quite largely different in short and long-term simulations. The first EOF pattern explained 45.48% of the total variance in long PPE, while it only explained 15.82% of the total variance in short PPE. The difference in explained variance indicates that short-term simulations are less aligned along a common pattern/mode of response to parameter variations than the long-term simulations are. The second and third EOF patterns from short and long-term simulations were also calculated and correlations between second EOF patterns from short and long-term simulations were found (Figure 3.3 (b)). A significant correlation (r=0.8766) between PC2 from long and short-term simulations was found, supporting strong similarities of model responses to parameters in the second EOF patterns, while the PC3 did not show a clear relationship. However, the second EOF pattern from short-term simulations explained 3.8% of total variance while the third EOF pattern explained 3.4% and are not uniquely separated based on North's Rule of Thumb for EOF significance (North et al., 1982), which indicated a potential of mixture in signals for lower order PCs from short-term simulations. Viewing these results together with the more fragmented first EOF pattern in short PPE, it seems that short-term simulations have another large source of variances that made the ensembles in short PPE change, and the responses in short-term simulations seem to be chaotic and fragmented.

There should be two different sources of uncertainty in our simulations, one is model uncertainty, caused by parameterization, in our case induced by different parameter values; another is internal variability, caused by natural variations in the Earth system. The

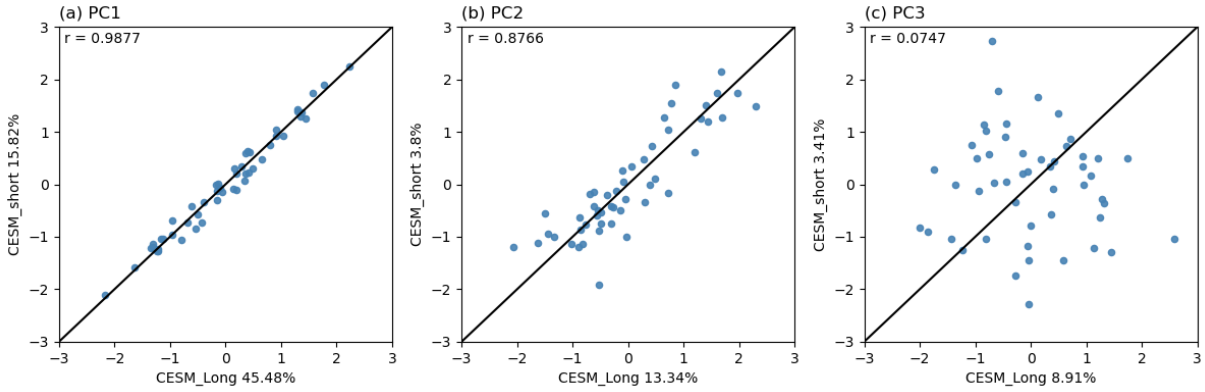

Figure 3.3: Scatter plot of PCs for the EOF patterns from long simulations and short simulations The R-value is the correlation coefficient between the two PCs. The black lines indicate lines following one-to-one linear relationships.

contribution of model responses to parameters in different ensemble members was found to be significantly correlated between short and long-term simulations (Figure 3.3). Therefore, to better assess the source of variance and uncertainty in short and long-term simulations, we calculated the internal variability uncertainty following Hawkins and Sutton (2009). Figure 3.4 showed the uncertainty from internal variability and from the model in short and long-term simulation for each targeted variable. As shown in Figure 3.4 for all four targeted variables, the internal variability in short PPE is much larger than that in the long PPE, with around 10-40% more uncertainty in short PPE compared with long. We noted that our internal variability has included the seasonal variability, since short PPE only had four start dates in one year, we cannot exclude the seasonal variability from our internal variability as we don't have multi-years short-term simulations. However, the calculation of internal variability from short and long-term simulations used the same equation and both included seasonal variability, so the values are still comparable. Similar to previous studies (Hawkins & Sutton, 2011; Schwarzwald & Lenssen, 2022), simulations in shorter timescales, in our case, 5 days, do have larger internal variability compared with our long-term simulations.

When looking at the difference of uncertainty from internal variability in each variable

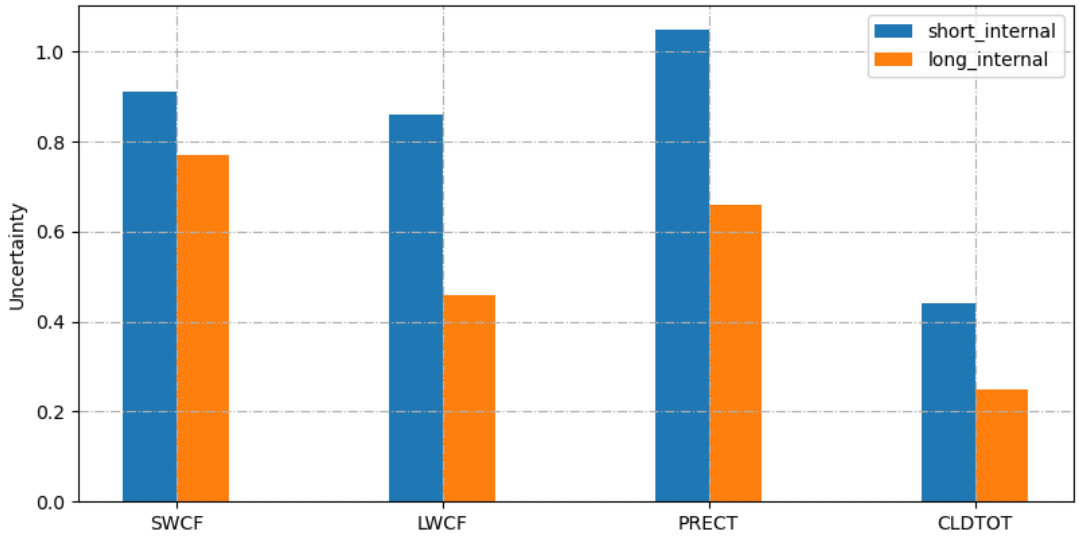

Figure 3.4: Uncertainty from internal variability from short (blue) and long-term (orange) simulations for each targeted variable. The internal variability is calculated as the mean of the standard deviation in each ensemble run. The uncertainty is represented by the coefficient of variance, i.e. the above-mentioned standard deviations divided by the mean values across all ensemble runs.

in Figure 3.4, it can be found that precipitation (PRECT) has the largest deviation of fractional uncertainty (40%) among all four targeted variables, followed by LWCF, CLDTOT and finally SWCF, with only about 10% difference in uncertainty. This order ranked by the difference in internal variability is exactly the same order if we rank the relationship in each variable by its correlation in the scatterplot of global mean values (Figure 3.1). Similarly, in Figure 3.2, the variable whose EOF pattern showed more fragmented features in short PPE and less similar to the EOF pattern from Long PPE, had a larger difference in internal variability here in Figure 3.4. The internal variability in the model is related to the natural climate variations (Schwarzwald & Lenssen, 2022), it can be caused by predictable variations in circulations or caused by natural chaotic noise in the Earth system. The difference in model output variables between short and long-term simulations may be related to different internal variability in short and long timescales. The large internal variability in short-term simulations may be the reason why short-term simulations differ

from the long-term, and the model responses to parameters in the short term have more fragmented features.

To further investigate the difference between short and long-term simulations and their difference related to climate variations (caused by internal variability), we showed the difference maps between short and long-term simulations (Figure 3.5) with a focus on the area with the largest difference value, the tropical Pacific. The short-term simulations modelled larger precipitation over the western Pacific warm pool (Figure 3.5 (a)), which corresponding showed stronger convection regionally, causing more high cloud (Figure 3.5 (d)) and stronger longwave cloud forcing (Figure 3.5 (e)) induced by high clouds. Corresponding, this linked to stronger Walker circulations in the wind fields, where there were stronger eastern trade winds in the lower atmosphere, converged in the western Pacific warm pool (Figure 3.5 (b)), and correspondingly, diverged in the higher atmosphere, led to strong western winds toward west Southern America (Figure 3.5 (c)). The large difference between short and long-term simulations regionally, is related to large-scale circulations, in this case, Walker circulations which are caused by stronger natural climate variations, or larger internal variability (Figure 3.4 (a)).

In conclusion, the relationship between short and long-term simulations: the short and long-term simulations showed a close relationship, the mean values of targeted variables were highly linear correlated, and shared a similar spread among ensemble members. The major spatial patterns of model responses given parameter changes showed as first EOF patterns, were also similar, and its relevant PC had a significant one-to-one linear relationship. Thus, the model responses given parameter changes in short and long-term simulations shared similar patterns and the parameter-induced model changes in each model ensemble were very similar. However, the model responses in short-term simulations had more fragmented features and short-term simulations were relatively more chaotic and seemed to have other sources of variance across model ensemble members. Such differences in short-term simulations were related to the fact that it has larger internal variability given it only simulated for a very short timescale. This additional internal variability would result in the difference in climate variations in PPE, and short-term simulations potentially have other variations across model ensembles, which may affect its representation of model responses given parameters in longer timescales.

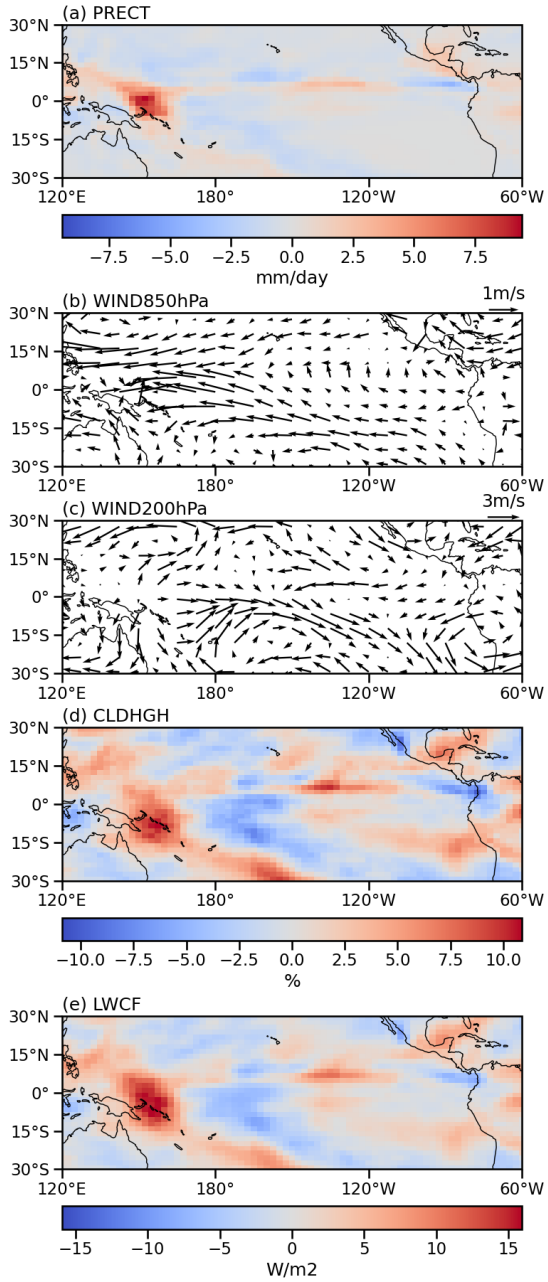

Figure 3.5: Regional difference between short and long-term simulations over tropical Pacific

3.3.2 Comparing emulators trained with short and long ensembles

To investigate the potential usage of short-term simulations in improving the efficiency of calibration using emulators, and whether short-term simulations could replace long-term simulations in representing model responses given parameter changes, we trained two GP emulators in this study using short PPE and long PPE correspondingly.

We first evaluated the overall emulator performance using SS metric (Table 3.2) on the emulated results obtained through the Leave-One-Out validation method. The SS metric was calculated by comparing the predicted results generated by each emulator with the model output from long-term simulations as ground truth. For all four targeted variables, the SS metric for emulators trained with short or long PPE both have a relatively good performance, compared with the multi-linear regression method. The SS metrics for all variables were close to 0.9, indicating the emulators were able to capture major proportions of the spatial variability from long-term simulations. The GP emulator trained with long PPE showed a small advantage in its performance compared with the other GP emulator trained with short PPE. This difference in performance may relate to their training dataset, as short PPE although found to have similar model responses given parameter changes, also has large internal variability across different ensembles, which may potentially affect the emulator performance as the training dataset has more variances that were non linked to parameters changes.

SS	SWCF	LWCF	PRECT	CLDTOT
Linear Regression	0.955	0.924	0.833	0.908
emulator_short	0.955	0.934	0.855	0.915
emulator_long	0.993	0.989	0.979	0.984

Table 3.2: Spatial skill metric (SS) of each variable emulated by emulators trained with short and long ensembles, with multi-linear regression trained with short-term simulations as benchmark comparison.

We also looked into the spatial distribution of selected variables predicted by emulators. Figure 3.6 showed the spatial distribution of selected variables in original long-term model

simulations results averaged over the last 5 years which worked as ground truth. The figure indicated the emulators' performance in reproducing spatial patterns of long-term simulations, and both emulators have successfully captured a majority of the spatial patterns in long-term simulations, as the predicted patterns looked very similar to the original model simulation data (when comparing middle and rightmost sub-figures with the sub-figure in the same row). Nevertheless, there are some large regional differences in results predicted by emulators trained with short PPE compared with long-term simulations, especially in the west tropical Pacific area. The spatial difference between predicted values and long-term simulations maps (Figure 3.7) could illustrate such difference more clearly. In Figure 3.7, when looking at the predicted differences using a GP emulator trained with long PPE (Figure 3.7 (a,c,e,g)), the value of differences were quite small, areas with relatively large differences were the same area with large absolute values and higher variance, so it's perhaps not surprising that those areas have larger differences. On the other hand, when looking at the predicted differences using a GP emulator trained with short PPE (Figure 3.7 (b,d,f,h)), the predicted results showed larger differences globally, and for LWCF, PRECT (Figure 3.7 (d,f)) there was an area with extremely larger differences compared with the rest of the globe. Such an area is the western Pacific warm pool, which also showed large differences when comparing short-term simulations with long-term in Figure 3.5. This indicated that the GP emulator trained with short PPE suffered its performance from its training data, as short-term simulations have larger internal variability, with stronger Walker circulations compared with long-term simulations, the GP emulator trained with short PPE also differs a lot when trying to reproduce the model responses relevant to climate variations or large scale circulations.

To help quantify the performance of emulators through pattern statistics, we employed the Taylor diagram (Taylor, 2001), shown as Figure 3.8. In the Taylor diagram, the data used as a reference was from long-term simulations (REF). In general, the closer each point is to REF, the better the performance of the emulator or the more similar spatial variability in test data. The short-term simulations are very different spatially compared with long-term simulations. Although the correlation coefficients of all four targeted variables, were not too small, most larger than 0.8, showing significant correlations, it suffered from much larger standard deviations within the dataset, potentially related to its larger

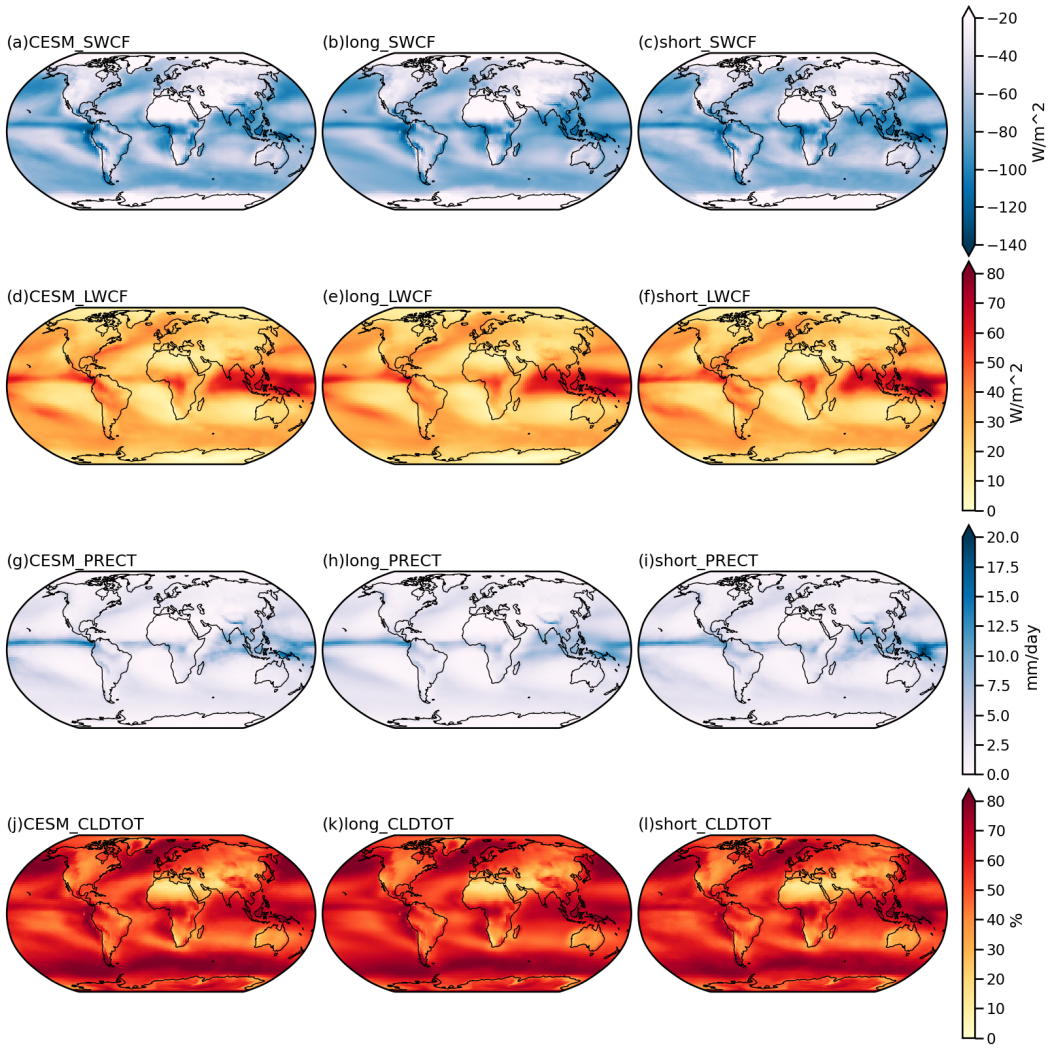

Figure 3.6: One example of the spatial distribution of the emulated results from emulators trained with long and short simulations. The leftmost column shows the spatial patterns in one PPE member for each variable, which worked as ground truth in assessing emulator performance. The middle column is the predicted results using the GP emulator trained with long PPE, and the rightmost column are same but predicted results with the GP emulator trained with short PPE.

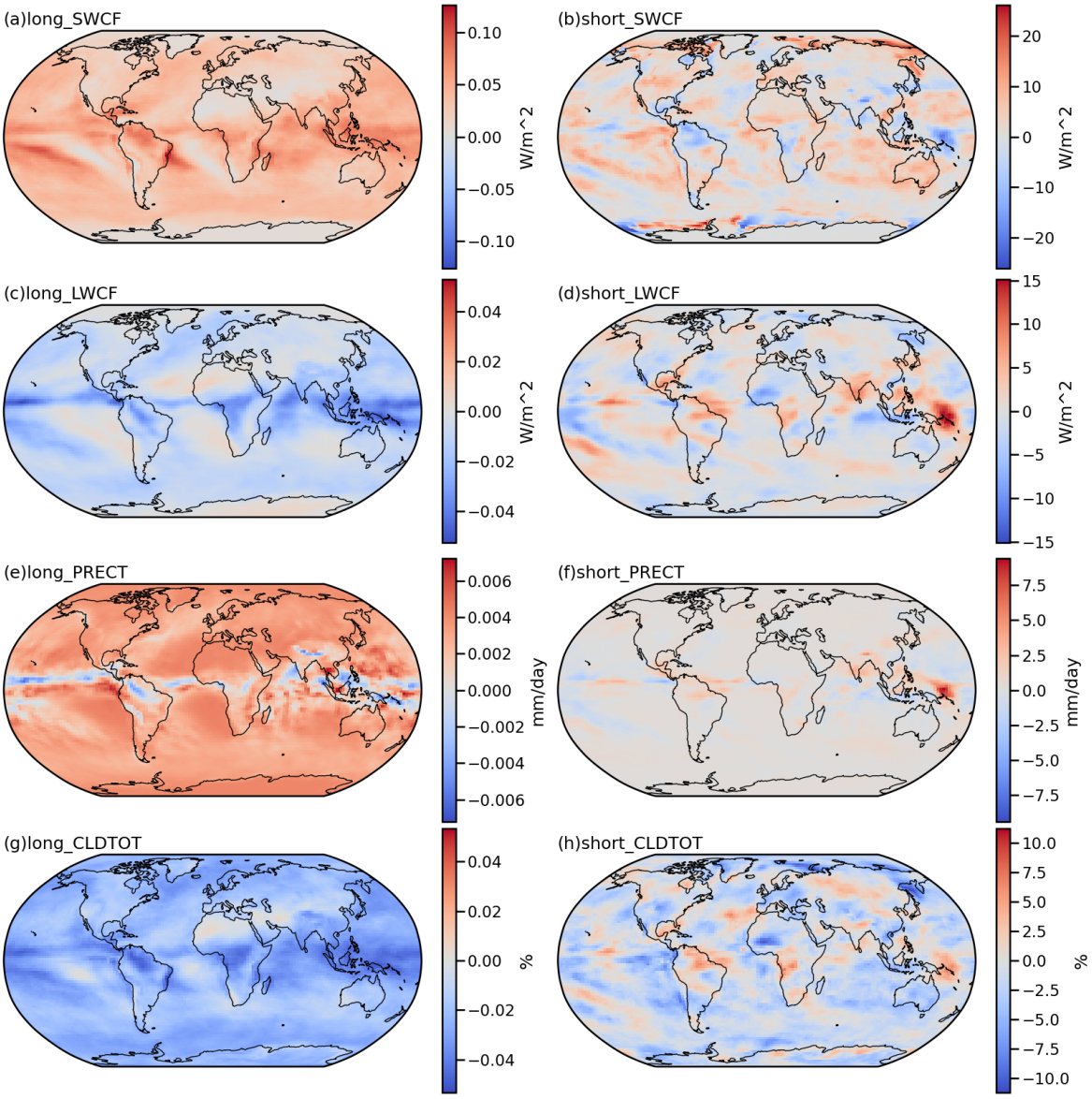

Figure 3.7: Spatial differences of the emulated results from emulators trained with long and short PPE correspondingly. The left column indicated predicted results using GP emulator trained with long PPE, while the right column was the same except using GP emulator trained with short PPE.

internal variability due to a much shorter simulation time. The predicted results using the multi-linear regression method have much better pattern similarity compared with long-term simulations, as it showed a much stronger correlation. Since the majority of model responses given parameters were found to have a kind of linear relationship (Figure 3.1) and could be expressed as EOF patterns on PPE member dimension (Figure 3.2). The multi-linear regression possibly captured the linear responses proportion, which was found to have a strong connection between short and long-term simulation in Figure 3.3. However, it still suffered from a larger standard deviation with predicted data, possibly coming from its training dataset, where short-term simulations have a larger internal variability which potentially affects emulator performance. Similarly, predicted results from the GP emulator trained with short PPE were also found to have larger standard deviations compared with long-term simulations. As a nonlinear method, the GP emulator could slightly outperform multi-linear regression and potentially exclude some misinformation from large internal variability in short PPE. Although it is promising to use short-term simulations to represent model responses for the emulator, the GP emulator trained with long PPE was still the best-performed emulator in reproducing model responses. The predicted results from it are not perfect, with some deviation from long-term simulations, yet remained to be most similar to long-term simulations among all emulators. Therefore, replacing long-term simulations with short when training emulators for calibration could get an emulator that captures the majority of the model responses given parameters, both in the mean state and in global spatial patterns. Therefore, short-term simulations are very promising to use in calibration with emulators and potentially help improve the overall efficiency of this process. However, for assessing regional responses, especially in areas highly relevant to climate variations, short simulations and emulators trained with them are not the best indicators for regional model responses related to climate variability, since short-term simulations have larger internal variability.

3.4 Discussion and Conclusion

In this study, we use short-term simulations as a relatively cheaper substitute for training emulators in the calibration process for Earth system modelling. There were two PPEs

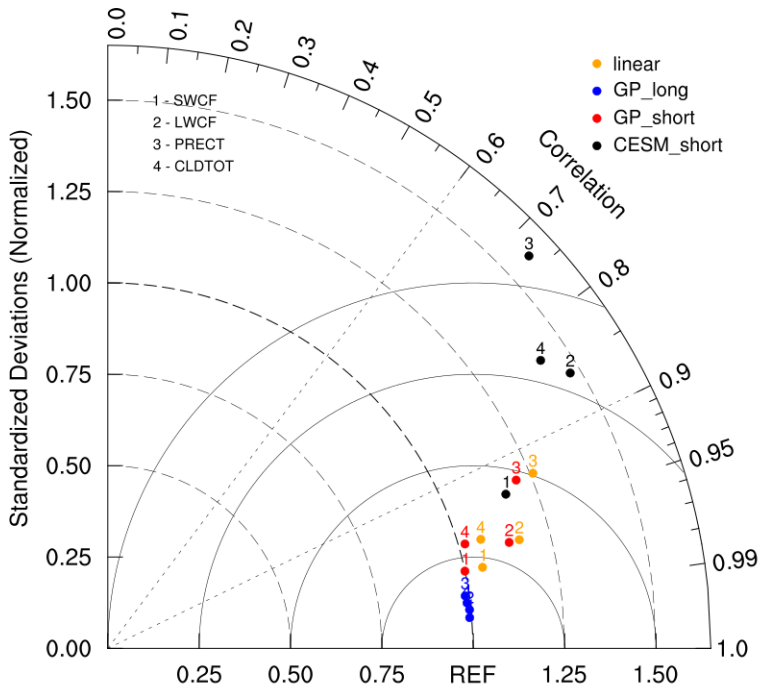

Figure 3.8: Taylor diagram showing pattern statistics metric between short and long PPE, predicted results using multi-linear regression, GP trained with short and long PPE correspondingly. The reference data in this diagram is the long PPE, i.e. long-term simulations data. Each dot represents pattern statistics for each targeted variable in short PPE (black), predicted by multi-linear regression (Yellow), predicted by GP emulator trained with short PPE (red) and predicted by GP emulator trained with long PPE (blue). The number associated with each dot referred to each targeted variable correspondingly, in order from 1 to 4, the targeted variables are SWCF, LWCF, PRECT and CLDTOT.

generated using one of the Earth system models (CESM) with five cloud-related parameters being perturbed. Following previous studies where short-term responses to cloud-related parameter changes were found to be linked with long-term model changes (Qian et al., 2018; Xie et al., 2012), and potentially be used to understand model parameterization errors (H.-Y. Ma et al., 2021) and help first filtering parameter values in calibration (Sexton et al., 2021). We raised the hypotheses that 1) short-term simulations could be used as a surrogate in training emulators for calibration; 2) the emulator trained by the short-term simulations could predict the long-term model responses to parameter changes with high fidelity; 3) this method substantially increases the efficiency of training an emulator to be used for calibrating an ESM.

For the first hypothesis, we investigated the relationship between short and long-term simulations in model responses to parameters changes. We selected four model variables (shortwave cloud forcing, longwave cloud forcing, total cloud fraction and precipitation) that are helpful in calibrating and identifying parameters' impact on clouds and the convection process. The short and long-term simulations were found to have similar mean values and spread across model ensembles for each of our targeted variables, and there was a significant one-to-one linear relationship between global mean values in short and long-term simulations. By performing EOF analysis on PPE member dimension, we acquired the first EOF pattern from short and long-term simulations correspondingly. Such EOF patterns indicated the major spatial patterns of model responses to parameter changes in each ensemble, and they were very similar. Moreover, their PC values showed a significant one-to-one linear relationship (r=0.9877), indicating the contributions or changes in each ensemble member's response to the same parameter values were similar relatively. This provided a piece of good evidence and strong support for us to use short-term simulation as a surrogate to represent long-term model responses and to use as the training dataset for the emulator, given that the model responses to parameters extracted from the data were found to be highly correlated and showed strong linkages.

Although short-term simulations shared similar model responses to parameters, there are certain regional differences between short and long-term simulations which were shown in EOF pattern where EOF pattern obtained from short-term simulations has more fragmented features, especially over the tropical Pacific. Such difference came from the fact

that short-term simulations have a larger internal variability compared with long-term, given its simulation time was much less. The larger internal variability led to stronger climate variations like Walker circulations over the western Pacific warm pool, and could potentially affect its ability to represent model responses for training emulators.

For the second hypothesis, we compared two GP emulators trained on short and long-term simulations correspondingly, to assess whether short-term simulations could replace long-term simulations for emulators to predict model responses to parameter changes. Both emulators have a certain ability to reproduce model responses, capturing the spatial variability in long-term simulation ensembles well. Therefore, it is possible to use short-term simulations for the emulator. Nevertheless, the emulator trained with short PPE suffered from larger internal variability from short-term simulations, its predicted results also processed similar large variance related to climate variability, leading to slightly poorer overall performance and larger standard deviation within the predicted dataset.

For the last hypothesis, given that short-term simulations take much less computational resources to run but have highly similar responses to parameters compared with long-term simulations, and the emulator trained with short-term simulations still has good performance in predicting model responses to parameters, we concluded that it is possible the accelerate the calibration process using short-term simulations for training emulators. The long-term simulations in this study took more than 60,000 CPU hours of computing resources on a high-performance computer platform, while short-term simulations took less than 2,000 CPU hours. Therefore, using short-term simulations to represent model responses to parameters could potentially save 95% of original computing resources and time in calibration. However, given the larger internal variability in short-term simulations, short-term simulations and emulators trained with it both have noticeable regional bias over the tropical Pacific. Using an emulator trained with short-term simulations to help select parameters could potentially be more trustworthy compared with using short-term simulations directly, yet there are still considerable deviations in regional biases and potential users of this method should be cautious when calibrating to minimize regional biases and when assessing natural variations.

Chapter 4

Summary

4.1 Conclusion

In this study, we tried to improve the efficiency of the current calibration process for ESMs. In the calibration for ESMs, the most resources demanded are to explore model responses to parameter values using long-term simulations. The two main methods to accelerate the calibration process either focus on predicting model responses given parameters using emulators or using short-term simulations exploring model responses to parameters. Inspired by those efforts, we use short-term simulations as a relatively cheaper surrogate for training emulators to predict model responses given parameters changes. Since short-term simulations take much fewer resources to run, using them as surrogates to train emulators will help improve the efficiency in calibration. We investigated two objectives in this study: 1) quantify the relationship between short and long-term simulations; 2) evaluate the performance of emulators trained with short-term simulations.

In addition to the similar mean and spread found in previous research, we identified a strong relationship between model responses to parameters between short and long-term simulations by using EOF analysis to decompose the major variance across ensemble members. The first EOF patterns from short and long-term simulations are spatially similar and their PC values have a significant one-to-one linear relationship. These results suggested

that the contributions of changes in each ensemble member give the same parameter values are very similar in short and long-term simulations, and supported methods using short-term simulations as a surrogate for long-term simulations in representing model responses to parameter changes.

The emulator trained with short-term simulations was evaluated against long-term simulations using spatial skill metrics and the Taylor diagram. The results suggested that the emulator trained with short-term simulations can capture the main spatial variability in long-term simulations and show a relatively good performance overall. Given that short-term simulations need much less computing time and resources to run, it is promising to use short-term simulations for the emulator in calibration to improve its efficiency, and using an emulator instead of short-term simulations directly is found to be more robust in estimating model responses in the long-term simulations.

Nevertheless, there are certain differences between short and long-term simulations which are mainly expressed as more fragmented and chaotic features in the short term. Such differences may be related to the larger internal variability in shorter timescales which is related to natural variations in the Earth system. The uncertainty and variance caused by internal variability are inevitable and potentially result in large regional biases when using short-term simulations and the emulator trained with them for calibration.

4.2 Limitations and Future Work

In this study, we did not process a full calibration and did not identify better parameter values by comparing them with observations. Instead, we only focused on whether short-term simulations could be used to represent model responses in long-term simulations, and whether the emulator trained with it has good performance. Therefore we do not have the opportunity to test our method in selecting parameter values and improving model performance.

The short-term simulations used in this study, only have four start dates throughout a year to sample the seasonal variability. This is a much smaller sample size compared

with the previous study (H.-Y. Ma et al., 2021) in evaluating parameterization using short-term simulations. Given that short-term simulations are much faster and cheaper to run, having multi-year short-term simulations with a variety of start dates throughout could potentially be a better sample for model responses in the long-term timescales.

There are certain differences between short and long-term simulations regionally, which may related to the large internal variability in shorter timescales, and potentially cause regional biases when using emulators trained with short-term simulations for calibration. Perhaps a better way to use short-term simulations and an emulator trained with them is as a piece of evidence for exclusion: instead of using the emulator trained with short-term simulations to identify parameters that lead to good model performance, using it to exclude parameters that definitely lead to large deviations between model and observations. By using it as a piece of evidence for exclusion, we can exploit the efficiency of running short-term simulations for the emulator to fast-filtering parameter space while less affected by the potential regional biases from the emulator. Given the large difference in internal variability in short and long-term simulations, I would like to further discuss the potential to distinguish internal variability from model errors in calibration for ESMs. Using short-term simulations to replace the long came from previous calibration efforts assessing short-term simulations' mean and spread. Although there is potential to use short-term simulations for calibration, it also showed another feature in calibration that the evaluation of internal variability in model assessment in calibration is not popular. The targets that were commonly calibrated onto are non-time-varying: energy equilibrium at top-of-atmosphere or observational records. Therefore, using simulations with similar means to replace long-term simulations is possible when only considering non-time-varying errors. However, internal variability is also an important source of uncertainty in assessing model performance (Schwarzwald & Lenssen, 2022), and the mismatch between model and observations, inability to capture regional or extreme events in the model may be due to the chaotic internal variability (Jain et al., 2023). It may be helpful to quantify the potential range of internal variability with an ensemble constructed like Kay et al. (2015) before using PPEs to assess parameter uncertainty in calibration, and potentially distinguish internal variability using statistical method (Barnes et al., 2019) from parameter uncertainty in calibration.

References

Alizadeh, O. (2022). Advances and challenges in climate modeling. *Climatic Change*, *170*(1), 18. https://doi.org/10.1007/s10584-021-03298-4

Baker, E., Harper, A. B., Williamson, D., & Challenor, P. (2022). Emulation of high-resolution land surface models using sparse Gaussian processes with application to JULES [Publisher: Copernicus GmbH]. *Geoscientific Model Development*, *15*(5), 1913–1929. https://doi.org/10.5194/gmd-15-1913-2022

Balaji, V., Couvreux, F., Deshayes, J., Gautrais, J., Hourdin, F., & Rio, C. (2022). Are general circulation models obsolete? [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *119*(47), e2202075119. https://doi.org/10.1073/pnas.2202075119

Barnes, E. A., Hurrell, J. W., Ebert-Uphoff, I., Anderson, C., & Anderson, D. (2019). Viewing Forced Climate Patterns Through an AI Lens. *Geophysical Research Letters*, *46*(22), 13389–13398. https://doi.org/10.1029/2019GL084944

Betts, A. K., Desjardins, R., Beljaars, A. C. M., & Tawfik, A. (2015). Observational study of land-surface-cloud-atmosphere coupling on daily timescales. *Frontiers in Earth Science*, *3*. Retrieved June 28, 2023, from https://www.frontiersin.org/articles/10.3389/feart.2015.00013

Betts, A. K., Desjardins, R., & Worth, D. (2013). Cloud radiative forcing of the diurnal cycle climate of the Canadian Prairies. *Journal of Geophysical Research: Atmospheres*, *118*(16), 8935–8953. https://doi.org/10.1002/jgrd.50593

Boyle, J. S., Williamson, D., Cederwall, R., Fiorino, M., Hnilo, J., Olson, J., Phillips, T., Potter, G., & Xie, S. (2005). Diagnosis of Community Atmospheric Model 2 (CAM2) in numerical weather forecast configuration at Atmospheric Radiation Measurement

sites. *Journal of Geophysical Research: Atmospheres*, *110*(D15). https://doi.org/10.1029/2004JD005042

Ceppi, P., Brient, F., Zelinka, M. D., & Hartmann, D. L. (2017). Cloud feedback mechanisms and their representation in global climate models. *WIREs Climate Change*, *8*(4), e465. https://doi.org/10.1002/wcc.465

Ceppi, P., & Hartmann, D. L. (2015). Connections Between Clouds, Radiation, and Midlatitude Dynamics: A Review. *Current Climate Change Reports*, *1*(2), 94–102. https://doi.org/10.1007/s40641-015-0010-x

Collins, M., Barreiro, M., Frölicher, T., Kang, S. M., Ashok, K., Roxy, M. K., Singh, D., Tedeschi, R. G., Wang, G., Wilcox, L., & Wu, B. (2020). Frontiers in Climate Predictions and Projections. *Frontiers in Climate*, *2*. Retrieved April 6, 2023, from https://www.frontiersin.org/articles/10.3389/fclim.2020.571245

Covey, C., Lucas, D. D., Tannahill, J., Garaizar, X., & Klein, R. (2013). Efficient screening of climate model sensitivity to a large number of perturbed input parameters: PERTURBED INPUT-PARAMETER SENSITIVITY. *Journal of Advances in Modeling Earth Systems*, *5*(3), 598–610. https://doi.org/10.1002/jame.20040

Dagon, K., Sanderson, B. M., Fisher, R. A., & Lawrence, D. M. (2020). A machine learning approach to emulation and biophysical parameter estimation with the Community Land Model, version 5. *Advances in Statistical Climatology, Meteorology and Oceanography*, *6*(2), 223–244. https://doi.org/10.5194/ascmo-6-223-2020

Danabasoglu, G., Lamarque, J.-F., Bacmeister, J., Bailey, D. A., DuVivier, A. K., Edwards, J., Emmons, L. K., Fasullo, J., Garcia, R., Gettelman, A., Hannay, C., Holland, M. M., Large, W. G., Lauritzen, P. H., Lawrence, D. M., Lenaerts, J. T. M., Lindsay, K., Lipscomb, W. H., Mills, M. J., … Strand, W. G. (2020). The Community Earth System Model Version 2 (CESM2). *Journal of Advances in Modeling Earth Systems*, *12*(2), e2019MS001916. https://doi.org/10.1029/2019MS001916

Edwards, P. N. (2011). History of climate modeling. *WIREs Climate Change*, *2*(1), 128–139. https://doi.org/10.1002/wcc.95

Fletcher, C. G., Kravitz, B., & Badawy, B. (2018). Quantifying uncertainty from aerosol and atmospheric parameters and their impact on climate sensitivity. *Atmospheric*

Chemistry and Physics, *18*(23), 17529–17543. https://doi.org/10.5194/acp-18-17529-2018

Fletcher, C. G., McNally, W., Virgin, J. G., & King, F. (2022). Toward efficient calibration of higher-resolution Earth System Models. *Journal of Advances in Modeling Earth Systems*, *n/a*(n/a), e2021MS002836. https://doi.org/10.1029/2021MS002836

Gal, Y., Hron, J., & Kendall, A. (2017). Concrete Dropout. *Advances in Neural Information Processing Systems*, *30*. https://papers.nips.cc/paper_files/paper/2017/hash/84ddfb34126fc3a48ee38d7044e87276-Abstract.html

Gates, W. L., Boyle, J. S., Covey, C., Dease, C. G., Doutriaux, C. M., Drach, R. S., Fiorino, M., Gleckler, P. J., Hnilo, J. J., Marlais, S. M., Phillips, T. J., Potter, G. L., Santer, B. D., Sperber, K. R., Taylor, K. E., & Williams, D. N. (1999). An Overview of the Results of the Atmospheric Model Intercomparison Project (AMIP I). *Bulletin of the American Meteorological Society*, *80*(1), 29–56. https://doi.org/10.1175/1520-0477(1999)080⟨0029:AOOTRO⟩2.0.CO;2

Gent, P. R., Danabasoglu, G., Donner, L. J., Holland, M. M., Hunke, E. C., Jayne, S. R., Lawrence, D. M., Neale, R. B., Rasch, P. J., Vertenstein, M., Worley, P. H., Yang, Z.-L., & Zhang, M. (2011). The Community Climate System Model Version 4. *Journal of Climate*, *24*(19), 4973–4991. https://doi.org/10.1175/2011JCLI4083.1

Gramacy, R. B. (2020). *Surrogates: Gaussian Process Modeling, Design, and Optimization for the Applied Sciences*. Chapman; Hall/CRC. https://doi.org/10.1201/9780367815493

Hannay, C. (n.d.). Assessing and tuning model parameterizations, 49.

Hawkins, E., & Sutton, R. (2009). The Potential to Narrow Uncertainty in Regional Climate Predictions [Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society]. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/10.1175/2009BAMS2607.1

Hawkins, E., & Sutton, R. (2011). The potential to narrow uncertainty in projections of regional precipitation change. *Climate Dynamics*, *37*(1), 407–418. https://doi.org/10.1007/s00382-010-0810-6

Hourdin, F., Ferster, B., Deshaye, J., Mignot, J., Musat, I., & Williamson, D. (2022). Tuning and quantifying parametric uncertainty of climate change simulations, 16.

Hourdin, F., Mauritsen, T., Gettelman, A., Golaz, J.-C., Balaji, V., Duan, Q., Folini, D., Ji, D., Klocke, D., Qian, Y., Rauser, F., Rio, C., Tomassini, L., Watanabe, M., & Williamson, D. (2017). The art and science of climate model tuning. *Bulletin of the American Meteorological Society*. https://doi.org/10.1175/BAMS-D-15-00135.1

Hourdin, F., Ferster, B., Deshayes, J., Mignot, J., Musat, I., & Williamson, D. (2023). Toward machine-assisted tuning avoiding the underestimation of uncertainty in climate change projections [Publisher: American Association for the Advancement of Science]. *Science Advances*, *9*(29), eadf2758. https://doi.org/10.1126/sciadv.adf2758

Hourdin, F., Williamson, D., Rio, C., Couvreux, F., Roehrig, R., Villefranque, N., Musat, I., Fairhead, L., Diallo, F. B., & Volodina, V. (2021). Process-Based Climate Model Development Harnessing Machine Learning: II. Model Calibration From Single Column to Global. *Journal of Advances in Modeling Earth Systems*, *13*(6), e2020MS002225. https://doi.org/10.1029/2020MS002225

Jain, S., Scaife, A. A., Shepherd, T. G., Deser, C., Dunstone, N., Schmidt, G. A., Trenberth, K. E., & Turkington, T. (2023). Importance of internal variability for climate model assessment [Number: 1 Publisher: Nature Publishing Group]. *npj Climate and Atmospheric Science*, *6*(1), 1–7. https://doi.org/10.1038/s41612-023-00389-0

Kay, J. E., Deser, C., Phillips, A., Mai, A., Hannay, C., Strand, G., Arblaster, J. M., Bates, S. C., Danabasoglu, G., Edwards, J., Holland, M., Kushner, P., Lamarque, J.-F., Lawrence, D., Lindsay, K., Middleton, A., Munoz, E., Neale, R., Oleson, K., ... Vertenstein, M. (2015). The Community Earth System Model (CESM) Large Ensemble Project: A Community Resource for Studying Climate Change in the Presence of Internal Climate Variability [Publisher: American Meteorological Society Section: Bulletin of the American Meteorological Society]. *Bulletin of the American Meteorological Society*, *96*(8), 1333–1349. https://doi.org/10.1175/BAMS-D-13-00255.1

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. https://doi.org/10.1111/1467-9868.00294

Klocke, D., & Rodwell, M. J. (2014). A comparison of two numerical weather prediction methods for diagnosing fast-physics errors in climate models. *Quarterly Journal of*

the Royal Meteorological Society, *140*(679), 517–524. https://doi.org/10.1002/qj.2172

Krishnamurthy, V. (2019). Predictability of Weather and Climate. *Earth and Space Science*, *6*(7), 1043–1056. https://doi.org/10.1029/2019EA000586

Ladyman, J., Lambert, J., & Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, *3*(1), 33–67. https://doi.org/10.1007/s13194-012-0056-8

Liu, Y. (2019). Introduction to the Special Section on Fast Physics in Climate Models: Parameterization, Evaluation, and Observation. *Journal of Geophysical Research: Atmospheres*, *124*(15), 8631–8644. https://doi.org/10.1029/2019JD030422

Lynch, P. (2006). *The emergence of numerical weather prediction: Richardson's dream* [OCLC: ocm70399629]. Cambridge University Press.

Ma, H.-Y., Zhou, C., Zhang, Y., Klein, S. A., Zelinka, M. D., Zheng, X., Xie, S., Chen, W.-T., & Wu, C.-M. (2021). A multi-year short-range hindcast experiment with CESM1 for evaluating climate model moist processes from diurnal to interannual timescales. *Geoscientific Model Development*, *14*(1), 73–90. https://doi.org/10.5194/gmd-14-73-2021

Ma, H.-Y., Chuang, C. C., Klein, S. A., Lo, M.-H., Zhang, Y., Xie, S., Zheng, X., Ma, P.-L., Zhang, Y., & Phillips, T. J. (2015). An improved hindcast approach for evaluation and diagnosis of physical processes in global climate models [_eprint: https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1002/2015MS000490]. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1810–1827. https://doi.org/10.1002/2015MS000490

Ma, H.-Y., Xie, S., Boyle, J. S., Klein, S. A., & Zhang, Y. (2013). Metrics and Diagnostics for Precipitation-Related Processes in Climate Model Short-Range Hindcasts. *Journal of Climate*, *26*(5), 1516–1534. https://doi.org/10.1175/JCLI-D-12-00235.1

Ma, H.-Y., Xie, S., Klein, S. A., Williams, K. D., Boyle, J. S., Bony, S., Douville, H., Fermepin, S., Medeiros, B., Tyteca, S., Watanabe, M., & Williamson, D. (2014). On the Correspondence between Mean Forecast Errors and Climate Errors in CMIP5 Models. *Journal of Climate*, *27*(4), 1781–1798. https://doi.org/10.1175/JCLI-D-13-00474.1

Ma, P.-L., Harrop, B. E., Larson, V. E., Neale, R. B., Gettelman, A., Morrison, H., Wang, H., Zhang, K., Klein, S. A., Zelinka, M. D., Zhang, Y., Qian, Y., Yoon, J.-H., Jones, C. R., Huang, M., Tai, S.-L., Singh, B., Bogenschutz, P. A., Zheng, X., ... Leung, L. R. (2022). Better calibration of cloud parameterizations and subgrid effects increases the fidelity of the E3SM Atmosphere Model version 1 [Publisher: Copernicus GmbH]. *Geoscientific Model Development*, *15*(7), 2881–2916. https://doi.org/10.5194/gmd-15-2881-2022

Mariotti, A., Ruti, P. M., & Rixen, M. (2018). Progress in subseasonal to seasonal prediction through a joint weather and climate community effort [Number: 1 Publisher: Nature Publishing Group]. *npj Climate and Atmospheric Science*, *1*(1), 1–4. https://doi.org/10.1038/s41612-018-0014-z

Marrel, A., Iooss, B., Jullien, M., Laurent, B., & Volkova, E. (2011). Global sensitivity analysis for models with spatially dependent outputs. *Environmetrics*, *22*(3), 383–397. https://doi.org/10.1002/env.1071

Matthews, A. G. d. G., Wilk, M. v. d., Nickson, T., Fujii, K., Boukouvalas, A., Le{\'o}n-Villagr{\'a}, P., Ghahramani, Z., & Hensman, J. (2017). GPflow: A Gaussian Process Library using TensorFlow. *Journal of Machine Learning Research*, *18*(40), 1–6. Retrieved June 30, 2023, from http://jmlr.org/papers/v18/16-537.html

McFarlane, N. (2011). Parameterizations: Representing key processes in climate models without resolving them. *WIREs Climate Change*, *2*(4), 482–497. https://doi.org/10.1002/wcc.122

McNeall, D., Williams, J., Betts, R., Booth, B., Challenor, P., Good, P., & Wiltshire, A. (2020). Correcting a bias in a climate model with an augmented emulator [Publisher: Copernicus GmbH]. *Geoscientific Model Development*, *13*(5), 2487–2509. https://doi.org/10.5194/gmd-13-2487-2020

Mignot, J., Hourdin, F., Deshayes, J., Boucher, O., Gastineau, G., Musat, I., Vancoppenolle, M., Servonnat, J., Caubel, A., Chéruy, F., Denvil, S., Dufresne, J.-L., Ethé, C., Fairhead, L., Foujols, M.-A., Grandpeix, J.-Y., Levavasseur, G., Marti, O., Menary, M., ... Silvy, Y. (2021). The Tuning Strategy of IPSL-CM6A-LR. *Journal of Advances in Modeling Earth Systems*, *13*(5), e2020MS002340. https://doi.org/10.1029/2020MS002340

Myren, S., & Lawrence, E. (2021). A comparison of Gaussian processes and neural networks for computer model emulation and calibration. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *14*(6), 606–623. https://doi.org/10.1002/sam.11507

Neale, R. B., Richter, J., Park, S., Lauritzen, P. H., Vavrus, S. J., Rasch, P. J., & Zhang, M. (2013). The Mean Climate of the Community Atmosphere Model (CAM4) in Forced SST and Fully Coupled Experiments [Publisher: American Meteorological Society Section: Journal of Climate]. *Journal of Climate*, *26*(14), 5150–5168. https://doi.org/10.1175/JCLI-D-12-00236.1

North, G. R., Bell, T. L., Cahalan, R. F., & Moeng, F. J. (1982). Sampling Errors in the Estimation of Empirical Orthogonal Functions [Publisher: American Meteorological Society Section: Monthly Weather Review]. *Monthly Weather Review*, *110*(7), 699–706. https://doi.org/10.1175/1520-0493(1982)110⟨0699:SEITEO⟩2.0.CO;2

Ogura, T., Shiogama, H., Watanabe, M., Yoshimori, M., Yokohata, T., Annan, J. D., Hargreaves, J. C., Ushigami, N., Hirota, K., Someya, Y., Kamae, Y., Tatebe, H., & Kimoto, M. (2017). Effectiveness and limitations of parameter tuning in reducing biases of top-of-atmosphere radiation and clouds in MIROC version 5. *Geoscientific Model Development*, *10*(12), 4647–4664. https://doi.org/10.5194/gmd-10-4647-2017

Phillips, T. J., Potter, G. L., Williamson, D. L., Cederwall, R. T., Boyle, J. S., Fiorino, M., Hnilo, J. J., Olson, J. G., Xie, S., & Yio, J. J. (2004). Evaluating Parameterizations in General Circulation Models: Climate Simulation Meets Weather Prediction. *Bulletin of the American Meteorological Society*, *85*(12), 1903–1916. https://doi.org/10.1175/BAMS-85-12-1903

Prodhomme, C., Batté, L., Massonnet, F., Davini, P., Bellprat, O., Guemas, V., & Doblas-Reyes, F. J. (2016). Benefits of Increasing the Model Resolution for the Seasonal Forecast Quality in EC-Earth [Publisher: American Meteorological Society Section: Journal of Climate]. *Journal of Climate*, *29*(24), 9141–9162. https://doi.org/10.1175/JCLI-D-16-0117.1

Qian, Y., Wan, H., Yang, B., Golaz, J.-C., Harrop, B., Hou, Z., Larson, V. E., Leung, L. R., Lin, G., Lin, W., Ma, P.-L., Ma, H.-Y., Rasch, P., Singh, B., Wang, H., Xie,

S., & Zhang, K. (2018). Parametric Sensitivity and Uncertainty Quantification in the Version 1 of E3SM Atmosphere Model Based on Short Perturbed Parameter Ensemble Simulations. *Journal of Geophysical Research: Atmospheres*, *123*(23), 13, 046–13, 073. https://doi.org/10.1029/2018JD028927

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning* [OCLC: ocm61285753]. MIT Press.

Rodwell, M. J., & Palmer, T. N. (2007). Using numerical weather prediction to assess climate models [_eprint: https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.23]. *Quarterly Journal of the Royal Meteorological Society*, *133*(622), 129–146. https://doi.org/10.1002/qj.23

Sachindra, D. A., Huang, F., Barton, A., & Perera, B. J. C. (2013). Least square support vector and multi-linear regression for statistically downscaling general circulation model outputs to catchment streamflows. *International Journal of Climatology*, *33*(5), 1087–1106. https://doi.org/10.1002/joc.3493

Salter, J. M., & Williamson, D. (2016). A comparison of statistical emulation methodologies for multi-wave calibration of environmental models. *Environmetrics*, *27*(8), 507–523. https://doi.org/10.1002/env.2405

Salter, J. M., Williamson, D. B., Scinocca, J., & Kharin, V. (2019). Uncertainty quantification for computer models with spatial output using calibration-optimal bases [arXiv:1801.08184 [stat]]. *Journal of the American Statistical Association*, *114*(528), 1800–1814. https://doi.org/10.1080/01621459.2018.1514306

Schmidt, G. A., Bader, D., Donner, L. J., Elsaesser, G. S., Golaz, J.-C., Hannay, C., Molod, A., Neale, R. B., & Saha, S. (2017). Practice and philosophy of climate model tuning across six US modeling centers. *Geoscientific Model Development*, 3207–3223. https://doi.org/10.5194/gmd-10-3207-2017

Schmidt, G. A., & Sherwood, S. (2015). A practical philosophy of complex climate modelling. *European Journal for Philosophy of Science*, *5*(2), 149–169. https://doi.org/10.1007/s13194-014-0102-9

Schwarzwald, K., & Lenssen, N. (2022). The importance of internal climate variability in climate impact projections [Publisher: Proceedings of the National Academy of

Sciences]. *Proceedings of the National Academy of Sciences*, *119*(42), e2208095119. https://doi.org/10.1073/pnas.2208095119

Sexton, D. M. H., Karmalkar, A. V., Murphy, J. M., Williams, K. D., Boutle, I. A., Morcrette, C. J., Stirling, A. J., & Vosper, S. B. (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), 989–1022. https://doi.org/10.1007/s00382-019-04625-3

Sexton, D. M. H., McSweeney, C. F., Rostron, J. W., Yamazaki, K., Booth, B. B. B., Murphy, J. M., Regayre, L., Johnson, J. S., & Karmalkar, A. V. (2021). A perturbed parameter ensemble of HadGEM3-GC3.05 coupled model projections: Part 1: Selecting the parameter combinations. *Climate Dynamics*, *56*(11-12), 3395–3436. https://doi.org/10.1007/s00382-021-05709-9

Stein, M. (1987). Large Sample Properties of Simulations Using Latin Hypercube Sampling. *Technometrics*, *29*(2), 143–151. https://doi.org/10.1080/00401706.1987.10488205

Taylor, K. E. (2001). Summarizing multiple aspects of model performance in a single diagram. *Journal of Geophysical Research: Atmospheres*, *106*(D7), 7183–7192. https://doi.org/10.1029/2000JD900719

Tett, S. F. B., Gregory, J. M., Freychet, N., Cartis, C., Mineter, M. J., & Roberts, L. (2022). Does Model Calibration Reduce Uncertainty in Climate Projections? *Journal of Climate*, *35*(8), 2585–2602. https://doi.org/10.1175/JCLI-D-21-0434.1

Thornton, P. E., Lamarque, J.-F., Rosenbloom, N. A., & Mahowald, N. M. (2007). Influence of carbon-nitrogen cycle coupling on land model response to CO $_2$ fertilization and climate variability: INFLUENCE OF CARBON-NITROGEN COUPLING. *Global Biogeochemical Cycles*, *21*(4), n/a–n/a. https://doi.org/10.1029/2006GB002868

Urban, N. M., & Fricker, T. E. (2010). A comparison of Latin hypercube and grid ensemble designs for the multivariate emulation of an Earth system model. *Computers & Geosciences*, *36*(6), 746–755. https://doi.org/10.1016/j.cageo.2009.11.004

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... van Mulbregt, P. (2020). SciPy 1.0: Fundamental algorithms for

scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wan, H., Rasch, P. J., Zhang, K., Qian, Y., Yan, H., & Zhao, C. (2014). Short ensembles: An efficient method for discerning climate-relevant sensitivities in atmospheric general circulation models. *Geoscientific Model Development*, *7*(5), 1961–1977. https://doi.org/10.5194/gmd-7-1961-2014

Watson-Parris, D., Rao, Y., Olivié, D., Seland, Ø., Nowack, P., Camps-Valls, G., Stier, P., Bouabid, S., Dewey, M., Fons, E., Gonzalez, J., Harder, P., Jeggle, K., Lenhardt, J., Manshausen, P., Novitasari, M., Ricard, L., & Roesch, C. (2022). ClimateBench v1.0: A Benchmark for Data-Driven Climate Projections. *Journal of Advances in Modeling Earth Systems*, *14*(10), e2021MS002954. https://doi.org/10.1029/2021MS002954

Williams, K. D., Bodas-Salcedo, A., Déqué, M., Fermepin, S., Medeiros, B., Watanabe, M., Jakob, C., Klein, S. A., Senior, C. A., & Williamson, D. L. (2013). The Transpose-AMIP II Experiment and Its Application to the Understanding of Southern Ocean Cloud Biases in Climate Models. *Journal of Climate*, *26*(10), 3258–3274. https://doi.org/10.1175/JCLI-D-12-00429.1

Williamson, D., Blaker, A. T., & Sinha, B. (2017). Tuning without over-tuning: Parametric uncertainty quantification for the NEMO ocean model. *Geoscientific Model Development*, *10*(4), 1789–1816. https://doi.org/10.5194/gmd-10-1789-2017

Xie, S., Ma, H.-Y., Boyle, J. S., Klein, S. A., & Zhang, Y. (2012). On the Correspondence between Short- and Long-Time-Scale Systematic Errors in CAM4/CAM5 for the Year of Tropical Convection. *Journal of Climate*, *25*(22), 7937–7955. https://doi.org/10.1175/JCLI-D-12-00134.1

Zappa, G., Ceppi, P., & Shepherd, T. G. (2020). Time-evolving sea-surface warming patterns modulate the climate change response of subtropical precipitation over land [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *117*(9), 4539–4545. https://doi.org/10.1073/pnas.1911015117

Zelinka, M. D., Myers, T. A., McCoy, D. T., Po-Chedley, S., Caldwell, P. M., Ceppi, P., Klein, S. A., & Taylor, K. E. (2020). Causes of Higher Climate Sensitivity in CMIP6

Models. *Geophysical Research Letters*, *47*(1), e2019GL085782. https://doi.org/10.1029/2019GL085782

Zhang, T., Zhang, M., Lin, Y., Xue, W., Lin, W., Yu, H., He, J., Xin, X., Ma, H.-Y., Xie, S., & Zheng, W. (2018). Automatic tuning of the Community Atmospheric Model CAM5.3 by using short-term hindcasts with an improved downhill simplex optimization method. *Geoscientific Model Development Discussions*, 1–22. https://doi.org/https://doi.org/10.5194/gmd-2018-87

Zhang, Z., & Moore, J. C. (2015). Chapter 6 - Empirical Orthogonal Functions. In Z. Zhang & J. C. Moore (Eds.), *Mathematical and Physical Fundamentals of Climate Change* (pp. 161–197). Elsevier. https://doi.org/10.1016/B978-0-12-800066-3.00006-1