Multimodel Emulation Paper Figures and Rough Draft

RDCEP

Abstract

Impacts assessment for climate change requires simulation of a broad range of future greenhouse gas scenarios, making the direct use of state-of-the-art climate models known as General Circulation Models (GCMs) computationally infeasible. Previous work has shown that statistical emulation is an accurate tool for predicting climate model yearly temperature projections. However, these studies either only showed that they were effective for a single specific model or were applicable only to a limited range of emissions scenarios. Here we show that the emulator is a general purpose tool and thus could replace direct use of GCMs in impacts analysis by showing its accuracy across 23 models in CMIP5 (Coupled Model Intercomparison Project Phase 5), a publicly available archive of runs from state of the art climate models. Furthermore, we find that a simple functional form of the emulator accurately predicts model temperature even with a minimal training set of only 4 precomputed model runs. We also find the emulator is capable of making accurate predictions at the grid scale resolution level and outperforms pattern scaling down from global mean temperature.

I. Introduction

Climate change is commonly accepted to have potentially destructive effects. (cite?) Impacts Assessment Models (IAMs) attempt to quantify these destructive effects, often for the purpose of finding optimal mitigation strategies. (cite examples of impacts assessment groups?) IAMs require simulation of future climate to make their predictions. Many impacts studies require simulation over a broad range of greenhouse gas emissions scenarios to quantify the uncertainty in future emissions. In addition, because many models differ in their projections of future climate these studies often make use of combined projections from multiple models known as multimodel ensembles.

State of the art climate models known as Ocean Atmosphere General Circulation Models (referred to here as GCMs) were developed for the purpose of making projections about future climate based on human emissions scenarios. However, using these models directly is computationally infeasible for many impacts studies as GCMs take on the order of weeks to months to produce their runs. IAMs therefore require an alternative strategy to get projections of future climate. For a detailed summary and comparison of the various approaches we refer the reader to Castruccio et al.[3]

In addition to the discussion of emulation approaches in Castruccio et al., there are two more recent approaches of which we are aware: Caldeira et al. [2] fits sum of exponential functions for temperature to GCMs used in the Coupled Model Intercomparison Project 5 (CMIP5) and is able to reproduce model projections for the abrupt4xCO2 and 1pctCO2 experiments. Foley et al. [10] developed the emulator PLASIM-ENTSem which uses singular vector decomposition to produce regional predictions for the model of intermediate complexity PLASIM-ENTS. However, this approach is limited to that single model.

In this paper we focus on extending the statistical emulation approach of Castruccio et al. The parameters of a simple function relating temperature to a past trajectory of CO_2 are trained on *GCM* temperature output. Once trained, the emulator is able to predict *GCM* temperature for CO_2 scenarios not in its training set. Emulation is extremely computationally cheap (producing regional projections for a single scenario takes on the order of seconds) which allows it to produce climate projections for impacts assessment studies in a reasonable amount of time.

Castruccio et al. showed that this emulator can accurately predict model output for a run not in the training set but only showed it for a single model: the Community Climate System Model 3 (CCSM3)[6]. For emulation to be a effective tool for impacts assessment, it has to meet two goals: (1) work across a wide range of state-of-the-art models, and (2) require a training set that consists only of already generated, publicly available model runs. An emulator that can meet both of these goals is capable of producing the multimodel ensembles desired by impacts assessment studies. Therefore, in this paper we evaluate the accuracy of the emulator for models in the Coupled Model Intercomparison Project 5 (CMIP5), a publicly available archive of runs for many GCMs from many different modeling groups, in order to test if the emulator is accurate across models.

CMIP5 specifies experiments for participating modeling groups to run, described in Taylor 2012[19]. Many of the future projection runs produced from these experiments are in some respects suboptimal for training an emulator. The Representative Concentration Pathway (RCP) experiments, described in more detail in section 2, are the most commonly available across models but 1) have a very smooth increase in CO_2 concentration which may make it difficult to estimate lags in temperature increase 2) are at the centennial scale which is potentially not sufficiently long for temperature to approach its equilibrium value, which may make it difficult to estimate long run climate sensitivity. The experiments that contain more information for training an emulator, such as the extended RCP runs out to 300 years or the abrupt4xCO2

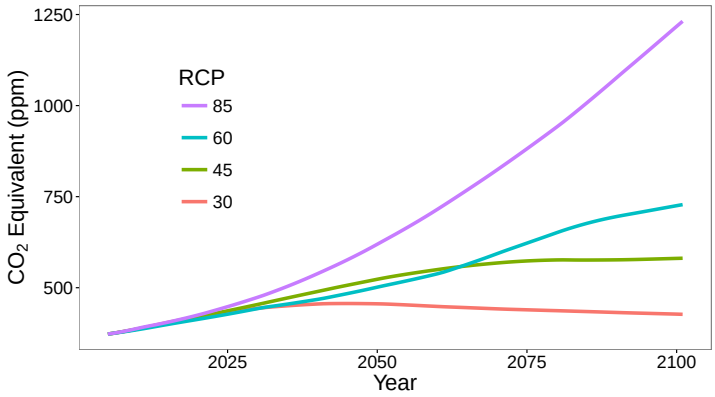

Figure 1: *CO_2 equivalent concentrations for the four Representative Concentration Pathways (RCPs) for the nonhistorical period.*

experiment, are not available across as many models with as many realizations as the standard 100 year RCPs (we discuss realizations and the benefit of having multiple realizations in the next section). We thus focus on evaluating the emulator when trained on only the standard 100 year RCPs.

In section 2, we describe the GCMs and the data used to train the emulator in more detail. In section 3, we present the functional form of the emulator and describe how it differs from the one presented in Castruccio et al. In section 4, we describe metrics for evaluating the accuracy of emulator predictions. In section 5, we present the emulation of 23 GCMs along with the corresponding evaluation metrics. In section 6, we find the minimum size of a training set needed to get an accurate emulation. In section 7, we test if the emulator is capable of reproducing runs that both interpolate and extrapolate from its training set. In section 8, we attempt to emulate at the model grid scale resolution directly and compare emulation to the common technique of pattern scaling from a simple global climate model. In section 9 we summarize our findings and conclude the paper.

II. Data Used

The Coupled Model Intercomparison Project Phase 5 (CMIP5) is a collection of climate model runs for use in the IPCC AR5 report (cite). It includes 61 models developed by 28 groups from which we select 23 models developed by 18 groups. Models were selected if they had at least one realization of at least three of the Representative Concentration Pathways (RCPs) to ensure an adequate training set. The 23 selected models are listed in table 1.

The IPCC specifies a standard set of future CO_2 scenarios known as the Representative Concentration Pathways (RCPs). These scenarios are given as a time series off CO_2 equivalent in the atmosphere from the year 1850-2100. The four RCP scenarios represent a range of possible future climate outcomes, with RCP30 as a low scenario, RCP85 as a business-as-usual scenario, and RCP45 and RCP60 as two middle scenarios. The years 1850-2005 are called the historical period and are the same in all RCP runs.[20]

In addition, several models had multiple realizations of at least three RCP runs available. Each realization of a scenario is run with different initial conditions and the realizations are assumed to be statistically independent from one another. Including multiple realizations of a scenario in the training set allows for a better estimate of the parameter values. The exact benefit is explored further in section 6.

This paper often refers to models by a 3 letter abbreviation. The key for all abbreviations is also listed in table 1, as well as the group that produced the model.

The models with multiple realizations available and the number of available realizations are listed in table 2.

We emulate annual mean temperature at both the model grid resolution and at the regional level. The 60 regions are shown in figure #. The data retrieved from the CMIP5 archive was monthly temperature at surface (TAS) at model grid resolution. This data was averaged to yearly using an unweighted 12 month average. It was then aggregated into 60 geopolitical regions with each pixel weighted by its area, as well as 3 additional "global" regions which consisted of all land pixels, all ocean pixels, or all pixels, respectively. Pixels and Regions are considered to be either entirely ocean or entirely land.

III. Emulator Model

Castruccio et al. described a model for the emulator where each parameter can be loosely interpreted to capture some aspect of the physical climate system that may differ across models. Our model differs slightly from the model presented there and is described below.

Model	Abbreviation	Center or Group
BCC-CSM1.1	BCC	Beijing Climate Center, China Meteorological Administration [24]
BNU-ESM	BNU	College of Global Change and Earth System Science, Beijing Normal University [13]
CANESM2	CAN	Canadian Centre for Climate Modeling and Analysis [4]
CCSM4	CCS	National Center for Atmospheric Research [11]
CESM1(CAM5)	CES	Community Earth System Model Contributors
CNRM-CM5	CNR	Centre National de Recherches Meteorologiques / Centre Europeen de Recherche et Formation Avancee en Calcul Scientifique [21]
CSIRO-Mk3.6.0	CSI	Commonwealth Scientific and Industrial Research Organization in collaboration with Queensland Climate Change Centre of Excellence [5]
FGOALS-g2	FGO	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences and CESS, Tsinghua University [15]
FIO-ESM	FIO	The First Institute of Oceanography, SOA, China [16]
GFDL-ESM2G	GFG	NOAA Geophysical Fluid Dynamics Laboratory [8] [9]
GFDL-ESM2M	GFM	NOAA Geophysical Fluid Dynamics Laboratory [8] [9]
GISS-E2-H	GIH	NASA Goddard Institute for Space Studies [18]
GISS-E2-R	GIR	NASA Goddard Institute for Space Studies [18]
HadGEM2-ES	HAD	Met Office Hadley Centre (additional HadGEM2-ES realizations contributed by Instituto Nacional de Pesquisas Espaciais) [14]
IPSL-CM5A-MR	IPS	Institut Pierre-Simon Laplace [7]
MIROC5	MIR	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology [22]
MIROC-ESM	MIM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies [23]
MIROC-ESM-CHEM	MIC	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies [23]
MPI-ESM-LR	MPL	Max-Planck-Institut fur Meteorologie (Max Planck Institute for Meteorology) [12]
MPI-ESM-MR	MPR	Max-Planck-Institut fur Meteorologie (Max Planck Institute for Meteorology) [12]
MRI-CGCM3	MRI	Meteorological Research Institute [25]
NorESM1-M	NOR	Norwegian Climate Centre [1]

Table 1: *General Circulation Models from the CMIP5 archive used in this study as well as their 3 letter abbreviations used throughout the paper.*

Model	Realizations
CCSM4	6
CESM1-CAM5	3
CSIRO-MK3-6-0	10
FIO-ESM	3
HadGEM2-ES	4
MIROC5	3

Table 2: *CMIP5 models with multiple realizations of all RCPs available.*

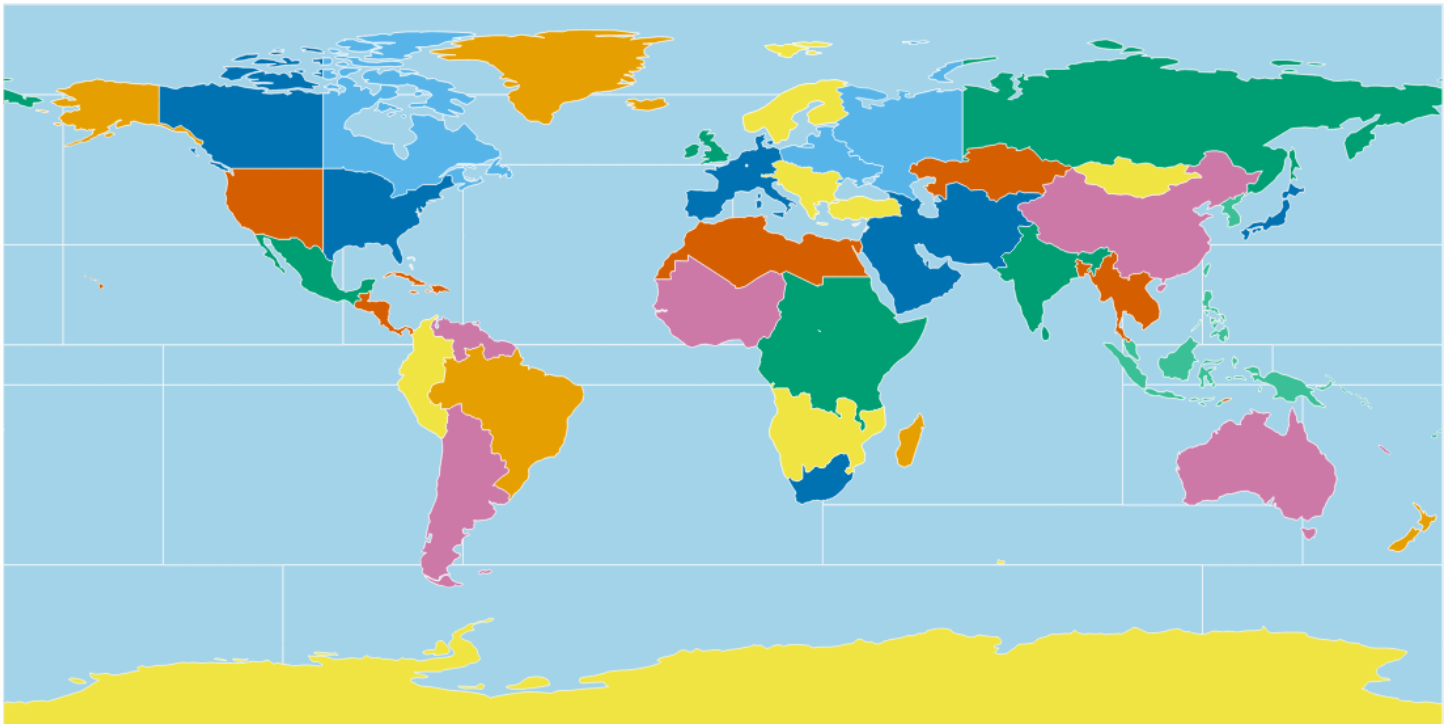

Figure 2: *Map of the 60 regions used in this study. Taken with permission from emulator.rdcep.org*

$$T(t) = \beta_0 + \beta_1(1-\rho)\sum_{k=0}^{n-1}\rho^k \log\left(\frac{CO_2[t-k]}{CO_{2preindustrial}}\right) + \epsilon(t) \tag{1}$$

β_0 is the mean temperature at preindustrial levels of CO_2.

β_1 is the long term temperature response to an increase in the log ratio of CO_2. Equilibrium temperature anomaly is expected to be proportional to the log of the ratio of current CO_2 to preindustrial CO_2.

ρ is the lag term in temperature response. Temperature is dependent on the past CO_2 trajectory and the dependence decreases the further we go back in time. We represent this dependence with an exponentially decaying weighting of previous years. $0 < \rho < 1$ and $\sum_{i=0}^{\inf}\rho^k = 1$.

We capture internannual variability using a simple autoregressive model $\epsilon(t) = \phi\epsilon(t-1) + v(t)$ where $v(t)$ is Gaussian white noise with unknown variance σ^2.

Castruccio et al. contained separate parameters for the short term and long term temperature response. However, the RCP runs used in this paper were sufficiently smooth that there was no difference in the short term and long term response and thus for the emulator model in this paper we used only a single temperature response parameter β_1. It may be possible to distinguish between short and long term effects when training set data contains runs with sudden increases or decreases in CO_2 concentration and distinguishing between the two may give some benefit to the accuracy of the emulation for models in which the mean temperature response is different on different time scales.

β_0 is calculated as the average temperature over the first 20 years of the historical portion of the GCM RCP simulation. The remaining parameters are fit using a fixed β_0 by minimizing the residual sum of squares to all data in the training set.

We will use the following notation for the training set of a particular instance of emulation. The training set "r realizations of (A_B_C)" indicates that the emulator was trained using r realizations each of the model output for the RCP scenarios A, B, and C.

We focus on the ability of the emulator to capture the mean trend and leave further study of the variability terms to future work.

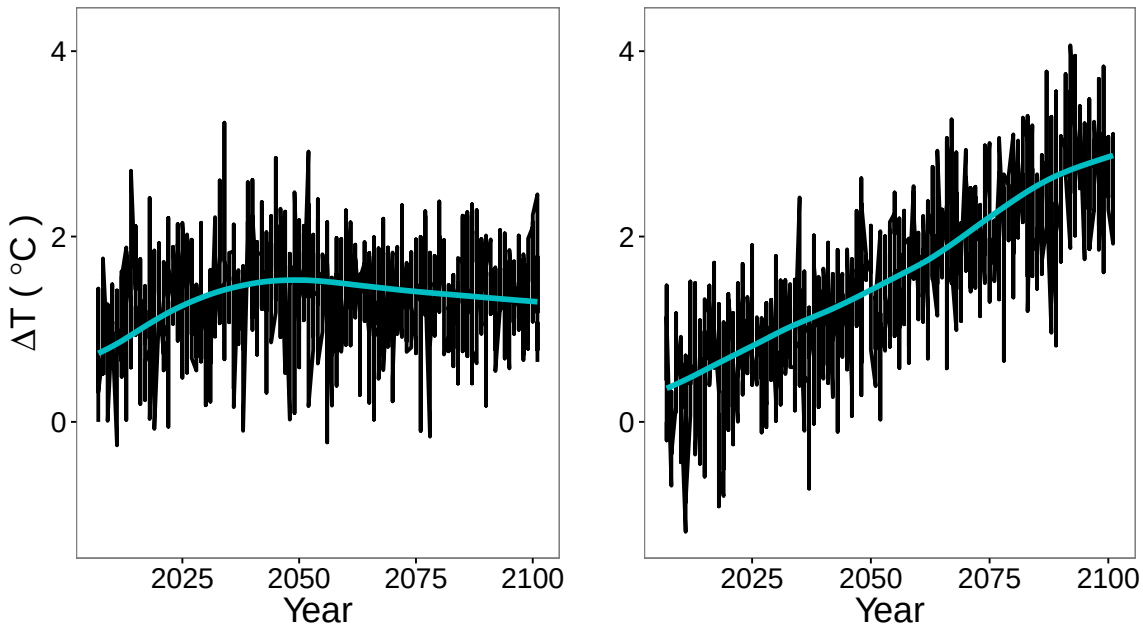

Figure 3: *All 6 realizations of model output (black) versus emulation (blue) of CCSM4 mean temperature for Western United States, trained on 6 realizations of (45_85) and emulating RCP30 (left) and RCP60 (right). Even with a modest training set of two scenarios, the emulator is able to accurately predict model temperature output for any arbitrary CO_2 scenario not in the training set. The I_1 and I_2, values, defined in section #, for the left and right emulation are (1.023, 1.967) and (1.068, 6.615), respectively.*

IV. Evaluating Quality of Emulation

One possible definition of an accurate emulator is one that can predict model yearly mean output temperature for any scenario. Any model run has both a mean and a stochastic component and in this paper we wish only to evaluate the ability of the emulator to predict the mean component. We therefore require a way to estimate the true model mean temperature prediction for a given year for a given scenario. One approach is to use the multiple realizations of model runs for a given scenario as estimates of the mean temperature. We can then evaluate the deviation of the emulator output from the average across realizations of the model run.

Castruccio et al. developed metrics to measure the performance of the emulator when there are multiple realizations of each scenario using the approach described above. The I_1 index measures emulator performance relative to a theoretical perfect emulator that could always reproduce the model mean temperature. The I_2 index measures the trend available to emulate relative to the uncertainty in the model mean. We now describe the two metrics in more detail.

I_1 compares the sum of squared deviations of each realization from the emulated temperature to the sum of squared deviations of each realization to the average across realizations $\bar{T}(t) = \frac{1}{R}\sum_{r=1}^{R} T_r(t)$.

$$I_1 = \frac{\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \hat{T}(t)]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \bar{T}(t)]^2} = \frac{N_1}{O_1} \tag{2}$$

Because we do not know the true mean and use $\bar{T}(t)$ as an estimate of it, we multiply the denominator by a factor of $R/(R-1)$. If the emulator perfectly reproduces the average mean temperature across realizations at each time t, then the numerator and the denominator should be equal and $I_1 = 1$. Thus a value of $I_1 = 1$ means that the emulator performed the best possible given the uncertainty in the true mean temperature. If $I_1 \gg 1$ then the emulator could have more accurately estimated the mean. The value of I_1 may occasionally be below 1 due to random variation in N_1 and O_1.

If the interannual variability of temperature is large compared to the mean trend then I_1 can be close to 1 even if the emulator inaccurately captures this mean trend. (Should I include an explanation of why, or a cartoon showing how the error in emulation is "hidden" by the large noise?). We thus have the I_2 index as a measure of the variation in the data that is attributable to the trend.

$$I_2 = \frac{\frac{n}{n-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \bar{T}_r]^2}{\frac{R}{R-1}\sum_{r=1}^{R}\sum_{t=1}^{n}[T_r(t) - \bar{T}(t)]^2} \tag{3}$$

$$\bar{T}_r = \frac{1}{n}\sum_{t=1}^{n} T_r(t)$$

If the GCM data shows no trend, then the numerator and denominator are unbiased estimates of the same quantity and I_2 should be close to 1. If $I_2 \gg 1$, then the variation in the data is explained more by the increasing temperature trend than the interannual variability. In the case that $I_1 \approx 1$ and $I_2 \gg 1$, there was a large trend and the emulator captures it accurately.

As mentioned, these metrics require that the model have multiple realizations. However, most models in the CMIP5 archive that have any RCP run data available only have one realization of each RCP run available. Evaluating the quality of emulation for these models requires the development of a new metric that only requires a single realization. We have the same goal as before of estimating the model mean temperature for each year. We accomplish this with a boxcar smooth. that is, we treat a window of neighboring years as approximately independent measurements of the center year. The new hypothetical perfect emulator is one that reproduces the windowed average $\hat{E}(t) = \frac{1}{m}\sum_{j=-m}^{m} T(t)$ for every t in $[m+1, n-m]$, where n is the length of the time series and m is the smoothing radius. Two possible problems arise with this technique for estimating the model mean. First, temperature may show some autocorrelation which means that neighboring years are not truly independent. To correct for this, we multiply by a factor $F(m,\phi)$, where m is again the window radius and ϕ is the fitted parameter from the $AR(1)$ component of the emulator model. Second, if the warming trend in temperature is nonlinear then the windowed average will be biased. However, this bias is small for small m.

We compare the sum of squared deviations of the model temperature from the emulated temperature to the sum of squared deviations of model temperature from the windowed average,

$$J_1 = \frac{\sum_{i=m+1}^{n-m}(T(t) - \hat{T}(t))^2}{\sum_{i=m+1}^{n-m}(T(t) - \hat{E}(t))^2} * F(m,\phi) \tag{4}$$

$$\hat{E}(t) = \frac{1}{2m+1}\sum_{j=-m}^{m} T(i+j)$$

$$F(m,\phi) = \sum_{i=m+1}^{n-m}\sum_{i'=m+1}^{n-m}\sum_{j=-m}^{m}\sum_{j'=-m}^{m} a_j a_{j'} \phi^{|i+j-i'-j'|}$$

$$a_k = \begin{cases} \frac{2m}{2m+1} & \text{if } k = 0 \\ \frac{-1}{2m+1} & \text{if } k \neq 0 \end{cases}$$

where $\hat{T}(t)$ is the emulated temperature at time t and $T(t)$ is the model temperature at time t.

We validate the metric J_1 by calculating both the I_1 and J_1 metric for the emulation of a model that has multiple realizations available. If J_1 is a suitable replacement forI_1 when there are not multiple realizations available, then for a given training set and scenario I_1 and J_1 should have identical distributions. We find that the I_1 and J_1 index tend to have similar values and that furthermore the regional distributions look nearly identical.

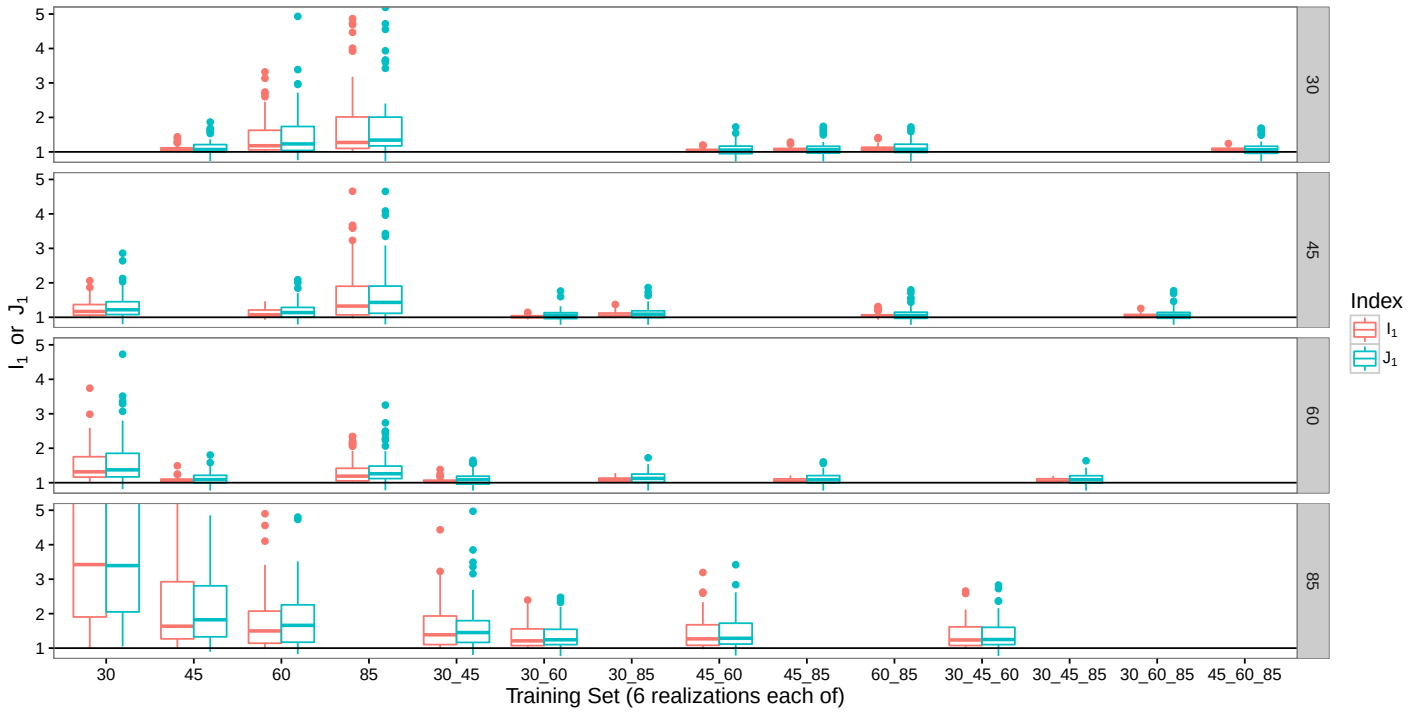

Figure 4: *Comparison of regional distributions of I_1 to J_1 for each training set and scenario for CCSM4. The label at the right of each of the 4 panels shows the emulated scenario. I_1 here is for the emulation with 6 realizations. J_1 here is the average of the J_1's for each individual realization of the 6 and the comparison model scenario. For each training set / scenario pair, J_1 has a similar distribution to I_1, meaning J_1 is near 1 only when I_1 is near 1, indicating that J_1 is a suitable replacement for I_1.*

We can similarly define a new metric J_2, which compares the sum of squared deviations from the mean $\bar{T} = \frac{1}{n-2m+1}\sum_{m+1}^{n-m}(T(t))$ to the same denominator as J_1, the sum of squared deviations from a boxcar smooth.

$$J_2 = \frac{\sum_{i=m+1}^{n-m}(T(t) - \bar{T})}{\sum_{i=m+1}^{n-m}(T(t) - \hat{E}(t))^2} * F(m, \phi) \tag{5}$$

Whenever multiple realizations are available, this paper uses the I_1 and I_2 metrics.

V. Emulation of All CMIP5 Models

We emulated every GCM in the CMIP5 archive for which the following data was available: at least one realization of RCP30, RCP45, and RCP85. We chose this requirement because our results indicate that two scenarios is the minimum training set needed to get an accurate emulation (see section #) and a third scenario is needed for comparison to test the emulator. These three specific RCPs were the most commonly available across all models. We trained the emulator on (30_85), then emulated RCP45.

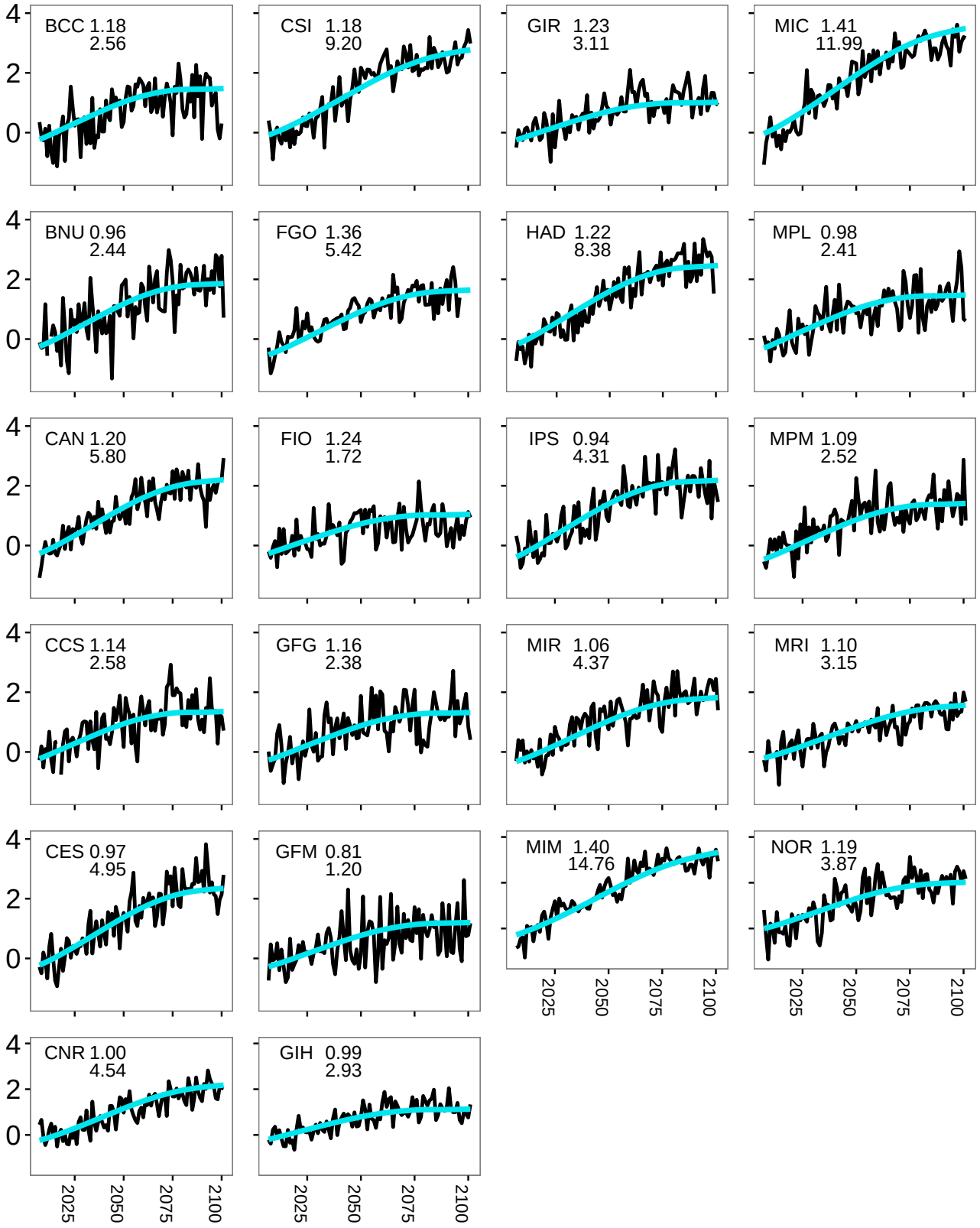

Figure 5: *Emulation of all models for the Western United States region, chosen as a representative example where the emulator performs well. Each plot is labeled by the three letter abbreviation followed by the J_1 index for the emulation. The emulator was trained on one realization of (30_85). The emulated scenario is RCP45. Refer to table 1 for the model name corresponding to each abbreviation. $J_1 \approx 1$ indicates the emulator does as well as it can relative to uncertainty. Note that the quality of emulation for this region, training set, and scenario is generally satisfactory for all tested models.*

Figure # (bouquet) demonstrates the ability of the emulator to accurately predict model mean temperature for all of the 22 models chosen for this study. Most models have a J_1 that is near 1 and the quality of the emulation can be confirmed by

eye. A few models have higher J_1, indicating that the emulator could have performed better, but even for these models (such as MIC with $J_1 = 1.41$ and FGO with $J_1 = 1.40$), the emulator mean temperature prediction is always bounded by the interannual variability of the model time series. The accuracy of emulation demonstrates the efficacy of the emulator as a tool for the prediction of model mean temperature across all CMIP5 models for this particular region and training set. We now turn our attention to the efficacy of the emulator for all of the 60 regions defined for the experiment.

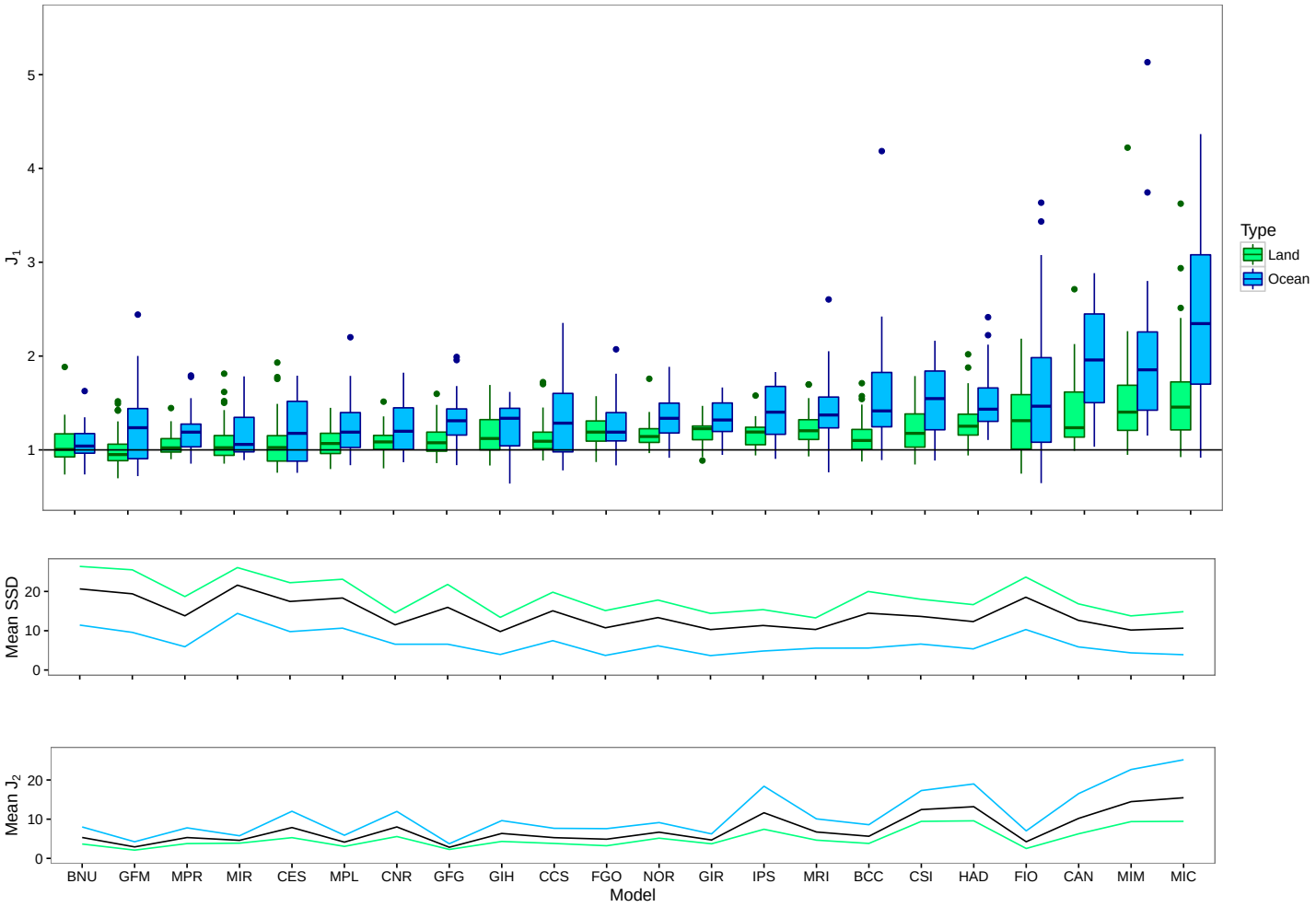

Figure 6: *(Top) Regional J_1 for emulation of all chosen models, separated by land and ocean regions and ordered by mean J_1. Each element in the boxplot represents one region. (Middle) Mean regional sum of squared deviations from a boxcar smooth of radius 2 (the denominator of J_1) for each model. Blue indicates ocean region average, green indicates land region average, and black indicates model region average. (Bottom) Mean regional J_2 for each model. (All) The emulator was trained on one realization of (30_85) and emulating RCP45. The J_1 values for land are generally near 1 which indicates the emulator performed optimally for most models. Note that we can achieve better emulation with more data in the training set (see section #), but for many models this was the maximum amount of data available. The model HadGEM2-AO was excluded for having some regions with extreme outliers, $J_1 > 10$; the regional mean J_1 for this model, including the outliers, was 3.84, and the interquartile range was 3.27. The emulation is close to as good as possible for many regions and for many models but there are also models where the emulation could be improved.*

With the modest training set of one realization of two scenarios, the emulation performs well for many models and many regions but has large room for improvement for others. There are a few things to note:

First, there is a significant land ocean contrast. The J_1 values for land regions are closer to 1 than those for ocean regions. (I'm not sure how to phrase this but) We can also note, however, that the average J_2 value for ocean regions is also nearly double that of the average J_2 value for land regions. This indicates that the land J_1 values may be closer to one because of the higher internannual variability over land. the emulator does not necessarily emulate land regions more accurately than ocean regions. The higher J_1 values for ocean only indicate that for those regions, the emulator has more room to improve than for the land regions, and it has more room to improve because there is a stronger temperature increase signal relative to the noise over oceans.

When the training set is large the emulator accurately captures the behavior of almost all tested models. For the models where the emulator has a large J_1 and thus does not perform as well as it could have, there are two possibilities.

The emulator might be capable of capturing the behavior of the model more accurately, but the chosen training set was insufficient. Adding additional data to the training set in the form of more scenarios or more realizations of existing scenarios can improve the accuracy of the emulator.

Alternatively, the function form of the emulator may be unable to capture the model's temperature response to CO_2. If this is the case, it might give some physical intuition about the model. For example, it could indicate that in the model the temperature increase does not scale linearly with the log ratio of CO_2 and thus no value of the β_1 term could accurately capture model behavior. For another example, it might mean the lag in temperature response is more complex than a single distributed weighted sum and that there might be multiple lags on multiple time scales. Future work with the emulator would involve a closer investigation into the physical properties of models based on the ability of various functional forms of an emulator to capture their behavior.

VI. Effect of Training Set

The emulator performs well even with a small training set. However, it is possible to get more accurate estimates of the parameters by adding additional data to the training set in the form of more scenarios or more realizations of the same scenario. In this section, we attempt to find the smallest amount of data in the training set necessary to get as an accurate emulation as possible (that is, to get $I_1 = 1$). We will do the intermodel comparison from the previous section again with an additional scenario in the training set. Then we will test how I_1 values change as we add additional data to a training set using a single test model, CCSM4, chosen because it was a model with multiple realizations available. In both cases, we find that additional additional data when J_1 is high greatly improves the quality of the emulation, but there are rapidly diminishing returns and adding additional data when $J_1 \approx 1$ does not gain any additional advantage. (This is an intuitive result: when he emulator is already doing almost as well as it theoretically can, adding more data can't make it do much better.)

We now select a subset of models with more available data to get an idea of the intermodel spread of emulator performance when the training set contains more data. For models that have at least 1 realization each of RCP30, RCP45, RCP60, and RCP85, we emulate using a training set of (30_60_85), and get a much more accurate emulation of RCP45 as indicated by the J_1 values being closer to 1.

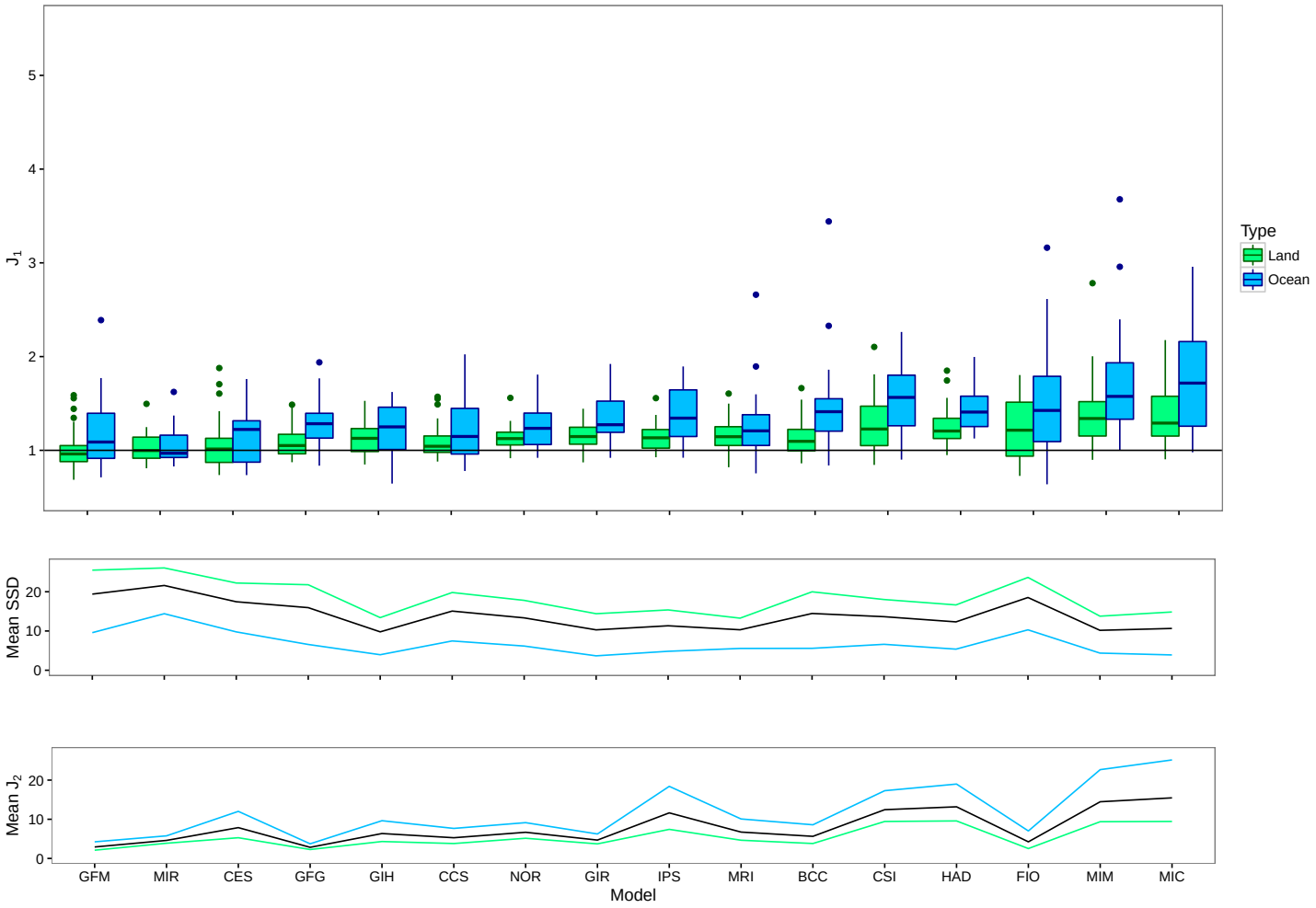

Figure 7: *Same as the previous figure, but with the training set (30_60_85) and emulating RCP45. (Top) Regional J_1 for emulation of all models, separated by land and ocean regions and ordered by mean J_1. Each element in the boxplot represents one region. (Middle) Mean regional sum of squared deviations from a boxcar smooth of radius 2 (the denominator of J_1) for each model. Blue indicates ocean region average, green indicates land region average, and black indicates model region average. (Bottom) Mean regional J_2 for each model. (All) For models where the emulator was already performing well, there is little improvement. For models where $J_1 \gg 1$, there is large improvement with mean J_1 for ocean decreasing by 1 for some models. Again, HadGEM2-AO was excluded, with a regional mean J_1 of 3.94 and a interquartile range of 3.25.*

The below table shows the effect of adding additional data to the training set for the single model CCSM4. The first column is the average I_1 across the three training sets with one scenario {(30), (60), (85)}. The second column is the average across the three training sets with two scenarios {(30_60), (30_85), (60_85)}. The third is for the one training set with three scenarios {(30_60_85)}.

		Scenarios		
		1	2	3
Realizations	2	1.585	1.095	1.090
	3	1.472	1.084	1.082
	4	1.414	1.076	1.074

Table 3: *Mean I_1 across all regions for CCSM4 emulating RCP45. Adding an additional new scenario to the training set improves the quality of emulation more than adding an additional realization. However, there is almost no noticeable benefit from moving to three scenarios.*

Adding additional runs or additional realizations of the same runs to the training set improves the quality of emulation, but quickly reaches a point of diminishing returns. Once $I_1 \approx 1$, there is almost no noticeable effect from adding more data. The quality of emulation also improves more when adding additional scenarios to the training set rather than just more realizations of the same scenario. In general, an adequate training set appears to be 2 realizations of 2 scenarios.

Models with many realizations allow us to better identify the point of diminishing returns on additional realizations. We take the scenario and training set for which the emulator had the highest I_1 and the lowest quality of emulation: (85) emulating RCP30, and measure the improvement in I_1 as we additional realizations to the training set, as shown in Figure #. This illustrates one example of the general trend that additional realizations beyond the 4th have virtually no effect.

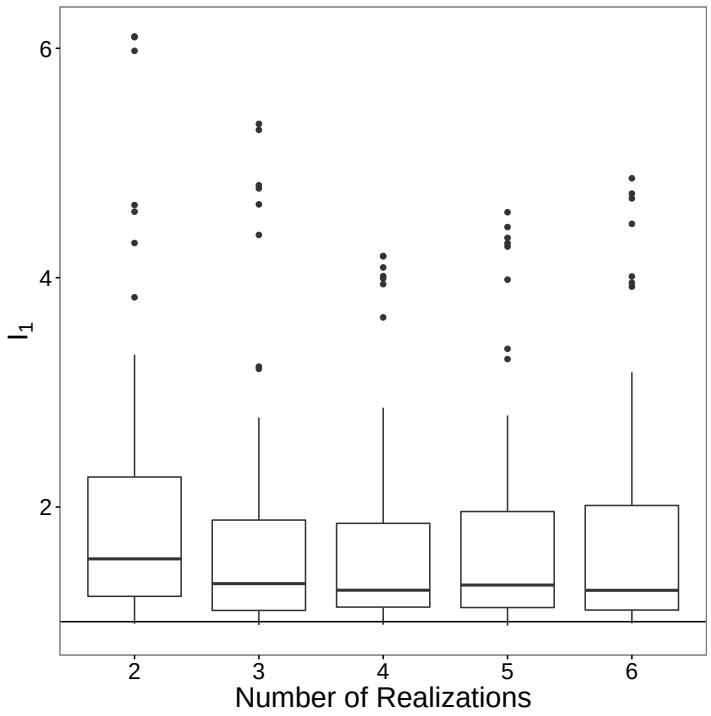

Figure 8: *Effect of adding additional realizations to the training set for CCSM4 trained on RCP85 and run on RCP30, chosen as a representative training set and scenario for which the emulator performed extremely poorly and thus had the potential to be improved by adding additional realizations. Note that adding realizations modestly improves the quality of emulation as measured by I_1 but after 4 realizations there is little effect. The outliers are the same for all realization counts and are: Southern South America, Asian Isles, South Western Pacific, Southern Atlantic, Warm Pool East, Antarctic Ocean Pacific, Antarctic Ocean Indian. The increased I_1 emulation for 5 and 6 realizations is due to random variation in the realizations.*

We can also measure the effect of changing the number and types of scenarios in the training set. Figure # shows that when training on a single scenario, whether the scenario is a slow or fast increase in CO_2 affects the quality of emulation: the emulation is better when the training set scenario is "closer" to the emulated scenario. However, when there are two or more scenarios in the training set, this effect disappears and the type of runs does not seem to affect the quality of emulation.

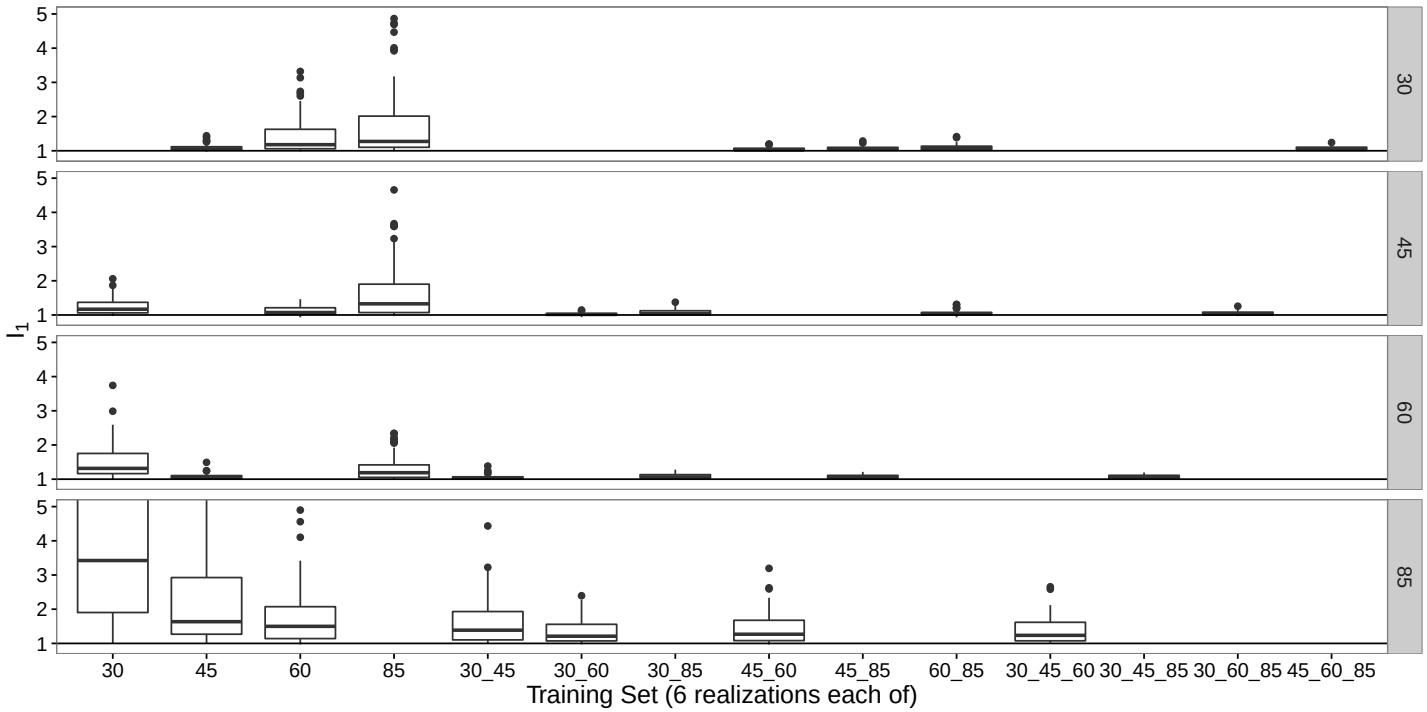

Figure 9: *Emulation of CCSM4 for all possible training sets for all four scenarios, excluding training sets which contained the emulated scenario. Note three things: (1) Adding additional scenarios improves the quality of emulation drastically, even at 6 realizations of each run in the training set. (2) When there is only a single scenario in the training set, the type of scenario matters. (3) 85 is difficult to emulate with any training set for CCSM4.*

VII. Extrapolation with Emulators

The emulated CO_2 scenario can either be in between (interpolation) or outside (extrapolation) runs in the training set. (Maybe a cartoon figure showing this?). We test if the emulator is capable of predicting model mean temperature for runs that extrapolate from the data in the training set as well as runs that interpolate. We find no evidence that extrapolation is more difficult for the emulator than interpolation.

There are only two possible training sets using the 4 RCPs that leave both a extrapolation and interpolation scenario to test with. The first is (30_60), for which 45 is the interpolation scenario and 85 is the extrapolation scenario. The second is (45_85), for which 60 is the interpolation scenario and 30 is the extrapolation scenario. In figure #, we compare the emulation of the interpolation scenario to the emulation of the extrapolation scenario, for each of the two training sets. Each training set contained 3 realizations of each run. We performed this test using 4 of the models with multiple realizations. The models CSIRO-Mk3-6-0 and FIO-ESM also had multiple realizations available but were excluded for having extreme outliers in their I_1 values in all four panels.

For the training set 3 realizations of (45_85), the I_1 values do not consistently increase across models when moving from interpolation to extrapolation. For CCSM4, HadGEM2-ES, and MIROC5, they increase slightly, but for CESM1-CAM5 they decrease drastically. Thus the data does not indicate that the emulator performs worse at extrapolation than interpolation.

For the training set 3 realizations of (30_60), the I_1 values do increase drastically when moving from interpolation to extrapolation. However, 85 is also a "more difficult" scenario to accurately emulate in general (as shown by a previous figure), and thus it is not possible from this data alone to distinguish between whether the increase in I_1 is due to extrapolation being "more difficult" or just 85 being "more difficult". (It might be a good idea to just remove the bottom row from the plot entirely, since all I conclude from it is that I can't conclude anything from it.)

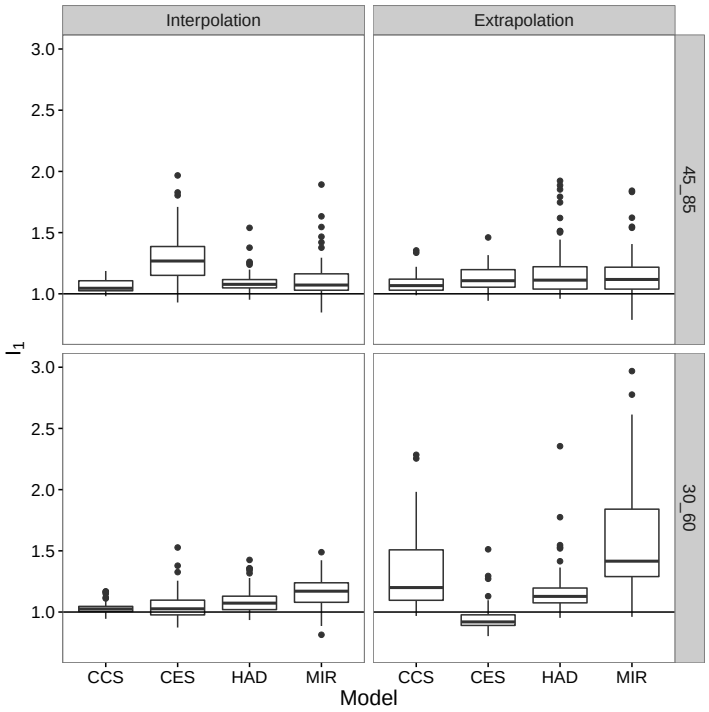

Figure 10: *Left: Interpolation, Right: Extrapolation. Each element in the boxplot is the I_1 for one region of the 63 regions; each panel is the emulation of one scenario for 4 models. The right hand label indicates the training set for panels in that row and each training set contained 3 realizations of those runs. The emulated scenario in each panel is (top left : 60), (top right: 30), (bottom left: 45), (bottom right: 85). Taking into account that 85 is a difficult scenario to emulate accurately, the emulator appears to do as well for interpolation as it does for extrapolation across the four selected models.*

Thus, we find no evidence that extrapolation is a harder problem for the emulator than interpolation.

VIII. Emulating Pixel Level Directly

Previous work with the emulator did not attempt to fit parameters at pixel level, that is, at model grid scale resolution. This may be because Castruccio et al. assumed that the interannual variability of pixel level temperature was too high to accurately estimate the parameter values for emulation, or because the functional form of the emulator in that paper contained a both a long and short term lag and individual pixels would not have had enough data to fit both of those parameters. Castruccio et al. attempted to emulate grid resolution only indirectly though a hybrid technique called regional pattern scaling by fitting parameters and emulating for regions, then downscaling from regional to pixel level using the method of pattern scaling. We now try to emulate grid resolution directly by fitting parameters to the temperature time series of each pixel individually for a representative test model: CCSM4. We then investigate whether this pixel level emulation is accurate as measured by the I_1 and I_2 values for each pixel in the grid. We also verify that this pixel level emulation is better than the alternative strategy of predicting the model mean temperature for a pixel: the hybrid technique presented in Castruccio.

I. Comparison to Regional Emulation

We first compare our ability to emulate individual pixels to our ability to emulate larger regions.

We find that the variability of a single pixel is often not significantly higher than the variability of the aggregated regions. Interestingly, there is a greater interregional variability than between some pixels and their containing regions. (Not sure how to phrase this scientifically, but:) if it was true that pixels were too "noisy" to emulate, then we would also have to reject entire regions as too "noisy" to emulate as many of these regions contain higher variability than many of the pixels. (We might no longer need this figure at all. Now that we have the map of I_2 for all pixels, do we need to "zoom in" on any of them?)

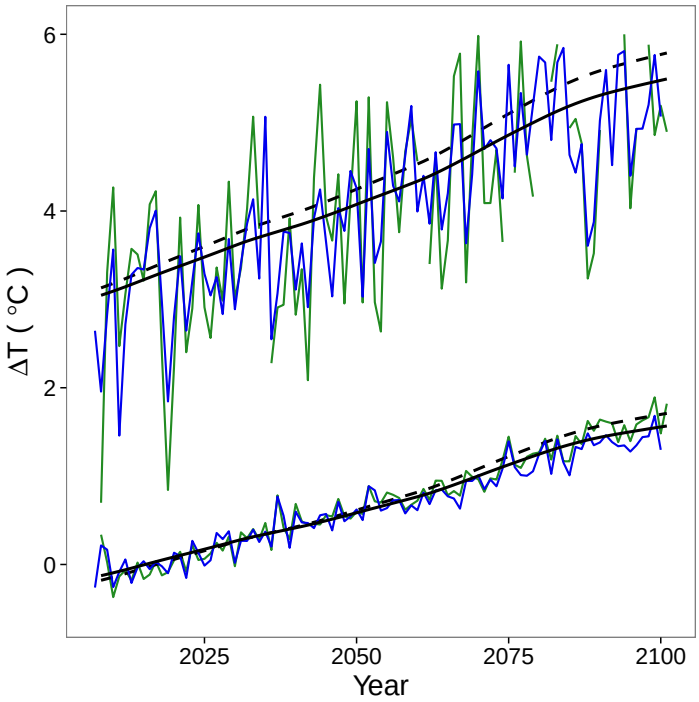

Figure 11: *Temperature Anomaly for RCP60 for a region (Blue) compared to a single representative pixel at the center of that region (Green), for two regions. The bottom lines are Indonesia and a representative pixel in the Indonesia region, the top lines are Western United States and a representative pixel. Solid black is the emulation of the region and dashed black is the emulation of the pixel, both using the training set of 6 realizations of (30_45_85). Western United States was shifted up by 3 degrees for clarity. Note that interannual variability for pixels is similar to that of the containing region; in addition, the difference in interannual variability between the two regions is greater than between a pixel and its containing region. For pixels where I_2 of the region is near I_2 of the pixel, fitting the parameters at the pixel level should not be harder than fitting the parameters to the containing region.*

We emulated every pixel in the CCSM4 grid. The emulator was trained on 6 realizations of (30_45_85), and emulated RCP60.

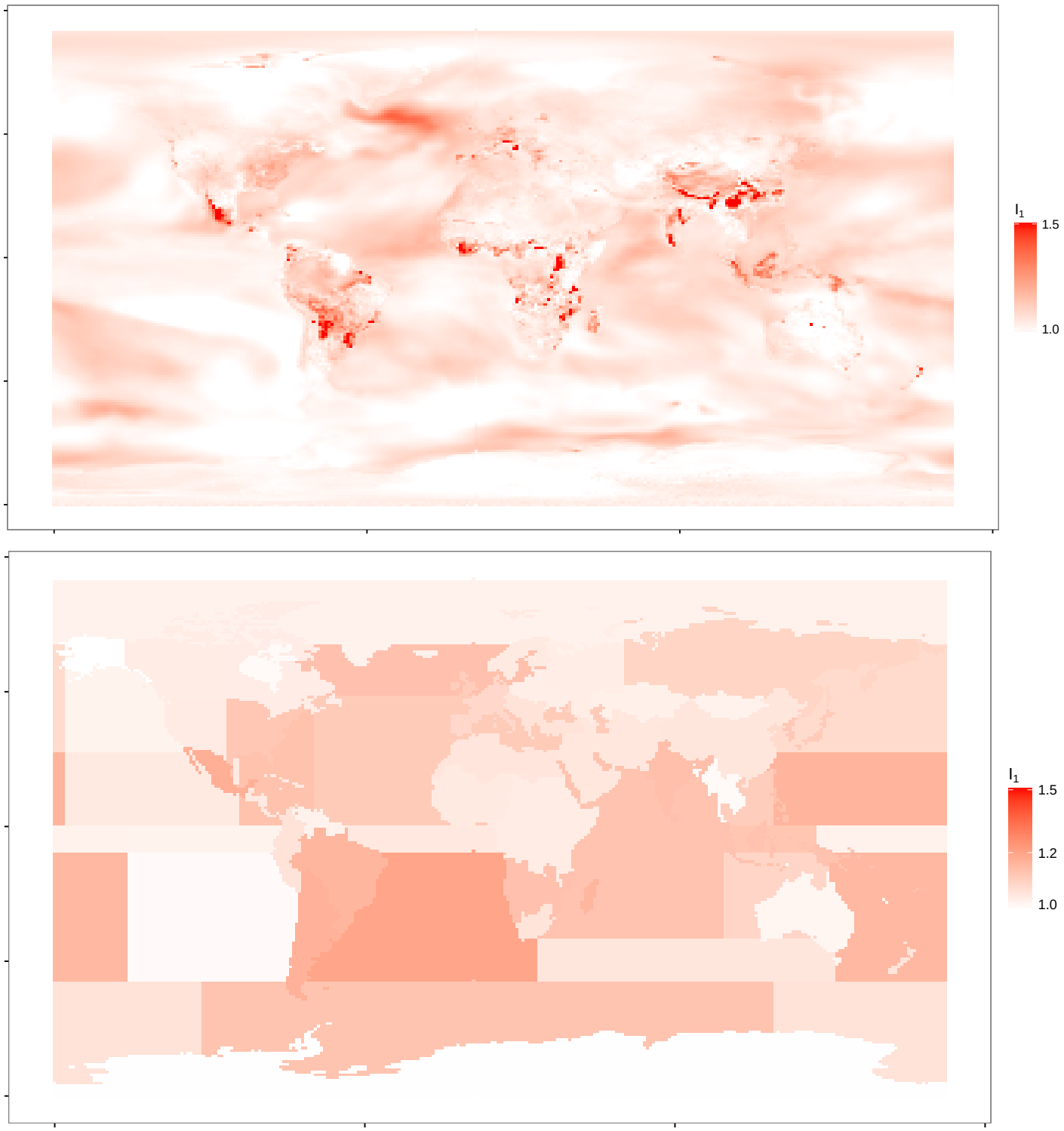

Figure 12: *(top) I_1 for emulation of every pixel, trained on CCSM4 (30_45_85), emulating RCP60. (bottom) I_1 for emulation of regions, same training set and scenario. The scale in both plots is capped so that the darkest red indicates an $I_1 \geq 1.5$ and the lightest white indicates an $I_1 \leq 1.0$. I_1 close to 1 indicates that the emulator accurately captures the mean trend relative to noise. Note that pixel level emulator is generally as good as regional emulation with a few outliers over land. These outliers may require regional emulation to accurately predict the mean trend.*

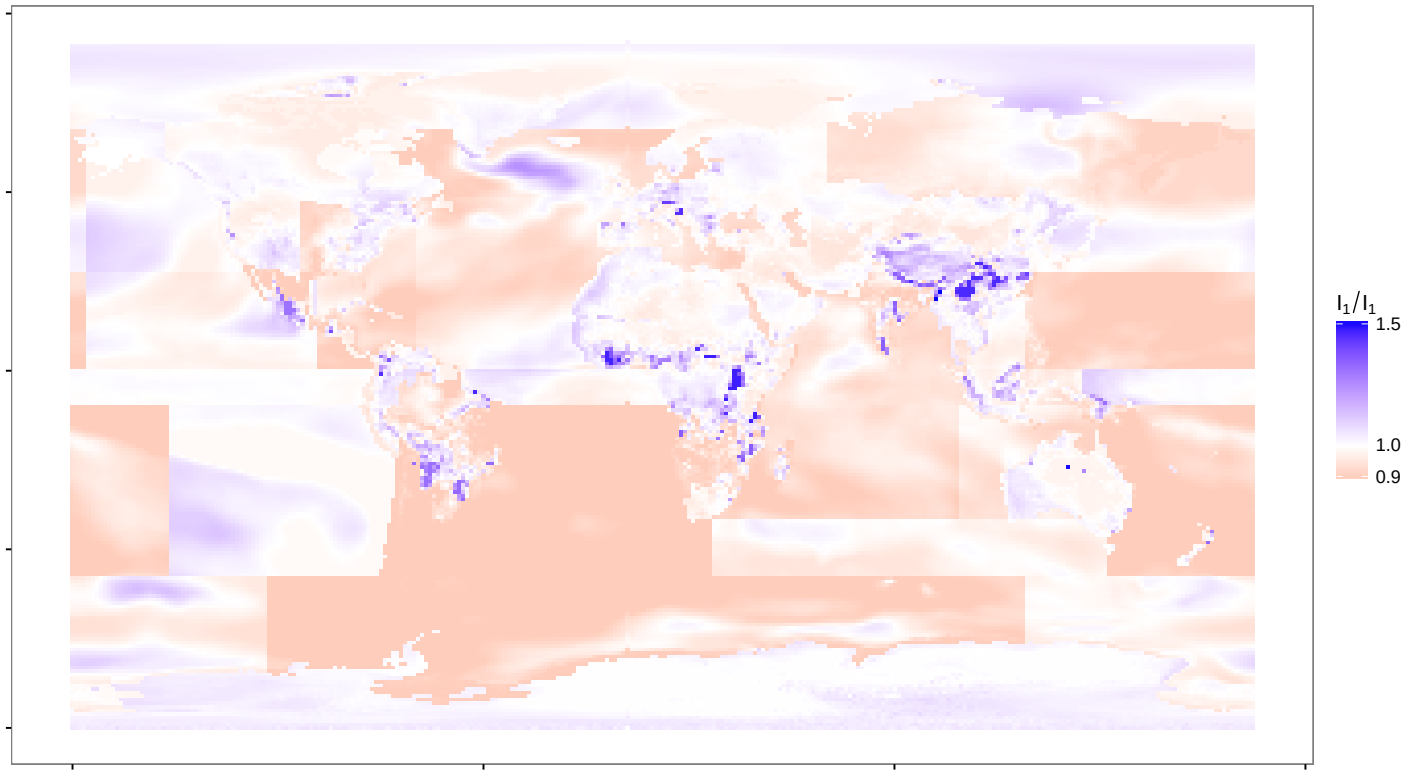

Figure 13: *Ratio of pixel I_1 to region I_1. Values <1 (red) indicate the pixel has a lower I_1 than its corresponding region. Values > 1 (blue) indicate the pixel has a higher I_1 than its corresponding region. The median ratio is 0.98 which suggests emulation at the pixel level is generally as good as emulation at the regional level.*

The I_1 values for pixel level emulation are generally very near 1 which means the emulation is predicting the model mean trend as accurately as possible relative to uncertainty in the model mean temperature. In addition, the emulation for a pixel tends to be as accurate as the emulation for its corresponding region as shown by figure #. Some land pixels are outliers with $I_1 >> 1$ which means that there was significant room for improvement. These pixels tend to be on geographic boundaries. One strategy for emulating at the pixel level might be to choose separately for each pixel between the direct pixel level emulation and the regional emulation to ensure an accurate emulation for all pixels.

As mentioned previously, an I_1 value near 1 only indicates that the emulator performs well when I_2 is large. The I_2 values appear suitably large for the I_1 values to be meaningful.

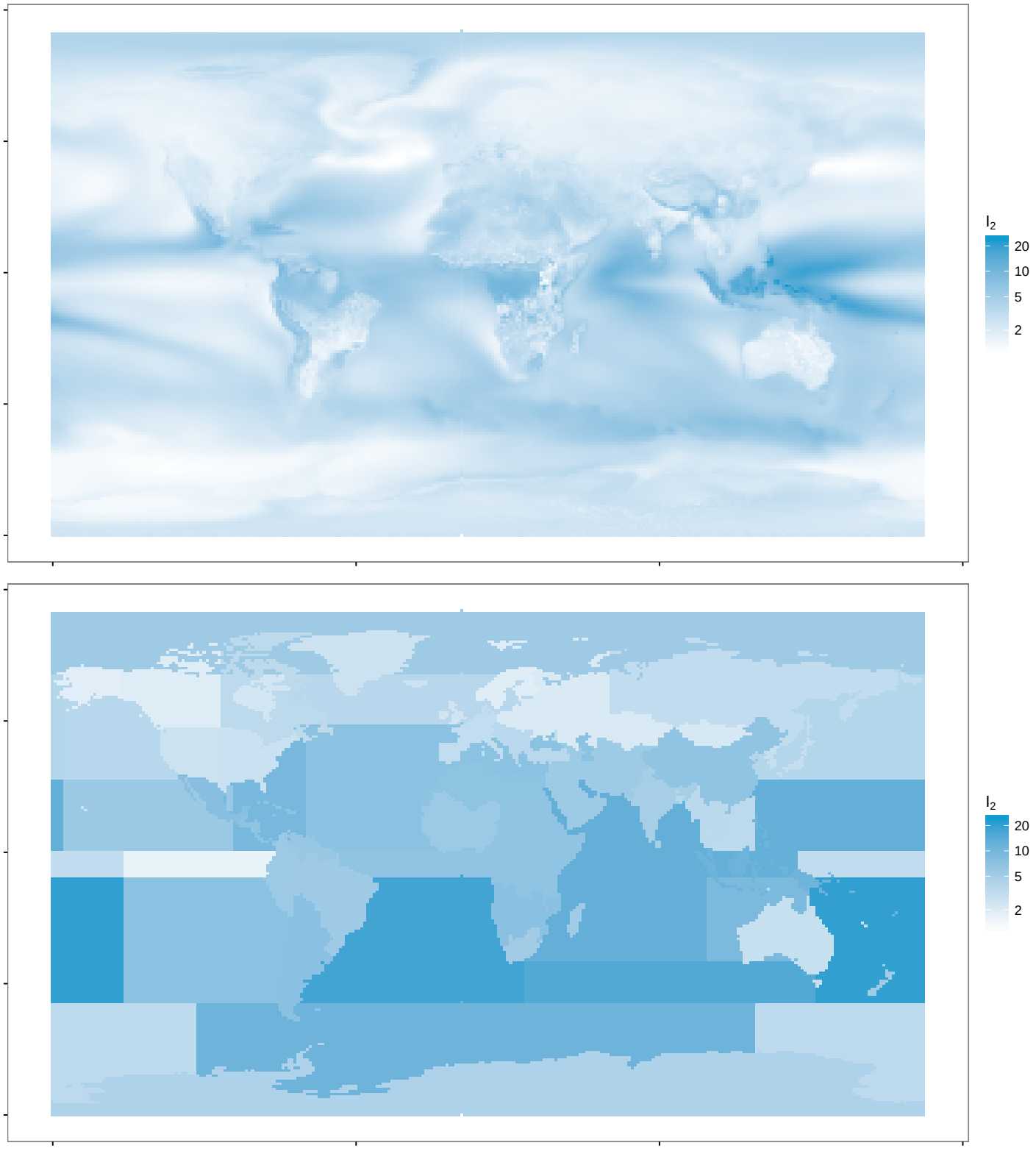

Figure 14: *(top) I_2 for every pixel in the CCSM4 rcp60 scenario. (bottom) I_2 for every region in the same scenario.I_2 » 1 indicates there is a strong signal relative to noise. Values of I_2 at the pixel level appear close to regional I_2 which indicates pixel level data contains a sufficiently high signal relative to noise for the I_1 near 1 to be meaningful.*

We also compare the I_2 values to the I_2 for the corresponding region. The ratios being mostly near 1 (white) indicates that the signal relative to initial condition uncertainty in each pixel is approximately the same as in the regions themselves, suggesting that it is not necessarily "harder" to emulate a pixel than a region.

Note that sharp cutoffs in spatial features in figure 10 indicate places where regions were poorly chosen.

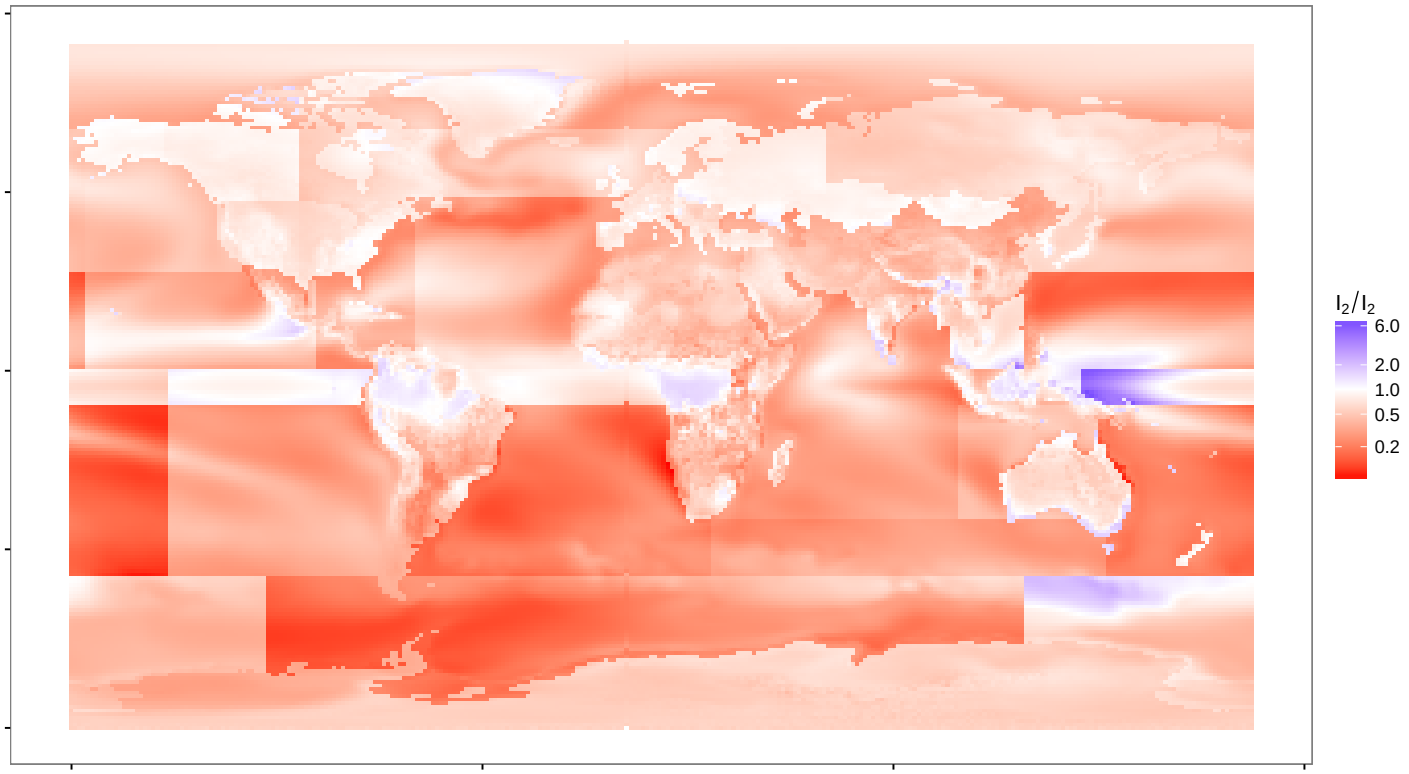

Figure 15: *Ratio of pixel I_2 to regional I_2. Values < 1 (red) indicate the pixel has a lower I_2 than its corresponding region. Values > 1 (blue) indicate the pixel has a higher I_2 and thus the inclusion of the pixel in the region hurts the emulation rather than helps it. Sharp cutoffs indicate places where regions were improperly chosen. The median ratio is 0.47 which suggests pixels contain, on average, half as much signal relative to noise as their corresponding region.*

II. Comparison to Regional Pattern Scaling

Pattern scaling is an alternative method to emulation for projecting future climate for a given emissions trajectory using simulated GCM data. It was first presented in Santer et al. [17] as a way to produce spatial climate projections using a global projection by treating the projection in each region as a scalar multiple of the global projection. A common technique in impacts assessment studies is to start with a simple energy balance model trained on GCM output which produces a global projection and then pattern scale down to the desired regional level. Examples and the limitations of this method are discussed in Castruccio et al.

Castruccio et. al. showed that a hybrid approach where regional temperature is emulated and pixel level temperature is calculated by pattern scaling to the containing region outperforms global mean temperature pattern scaling. Thus, we compare direct pixel level emulation to the hybrid method described in Castruccio. The method treats each pixel temperature anomaly time series as a scalar multiple of the of the anomaly time series of it's containing region and estimates this scalar using linear regression:

1. Fit parameters of emulation equation using a training set *tset*.

2. Emulate regional temperature using these parameters.

3. For each pixel, use the same data in *tset* to fit the parameters $\alpha_{(i,j)}$ and $c_{(i,j)}$ in $T_{(i,j)}(t) - T_{(i,j),preindustrial} = \alpha_{(i,j)} * (T_r(t) - T_{r,preindustrial}) + c_{(i,j)}$ using linear regression, where $T(t)_{(i,j)}$ is the temperature of the pixel at (i,j) in the grid, $T_r(t)$ is the temperature of the containing region, and $\alpha_{(i,j)}$ is the scalar pattern for that pixel.

4. We can then calculate the pattern scaled temperature $\hat{T}_{(i,j),sce}(t)$, for any scenario *sce* by $T_{(i,j),sce}(t) = \alpha_{(i,j)} * T_{r,sce}(t) + T_{(i,j),preindustrial}$.

We performed this method using the training set of all 6 realizations each of RCP30, RCP45, and RCP85 for CCSM4. The preindustrial temperature for each pixel and region was calculated as the average over the entire first realization of the CCSM4 preindustrial run. We then calculated the pattern scaled temperature for scenario RCP60 and then calculated the I_1 values using the pattern scaled temperature as $\hat{T}(t)$ in equation (#).

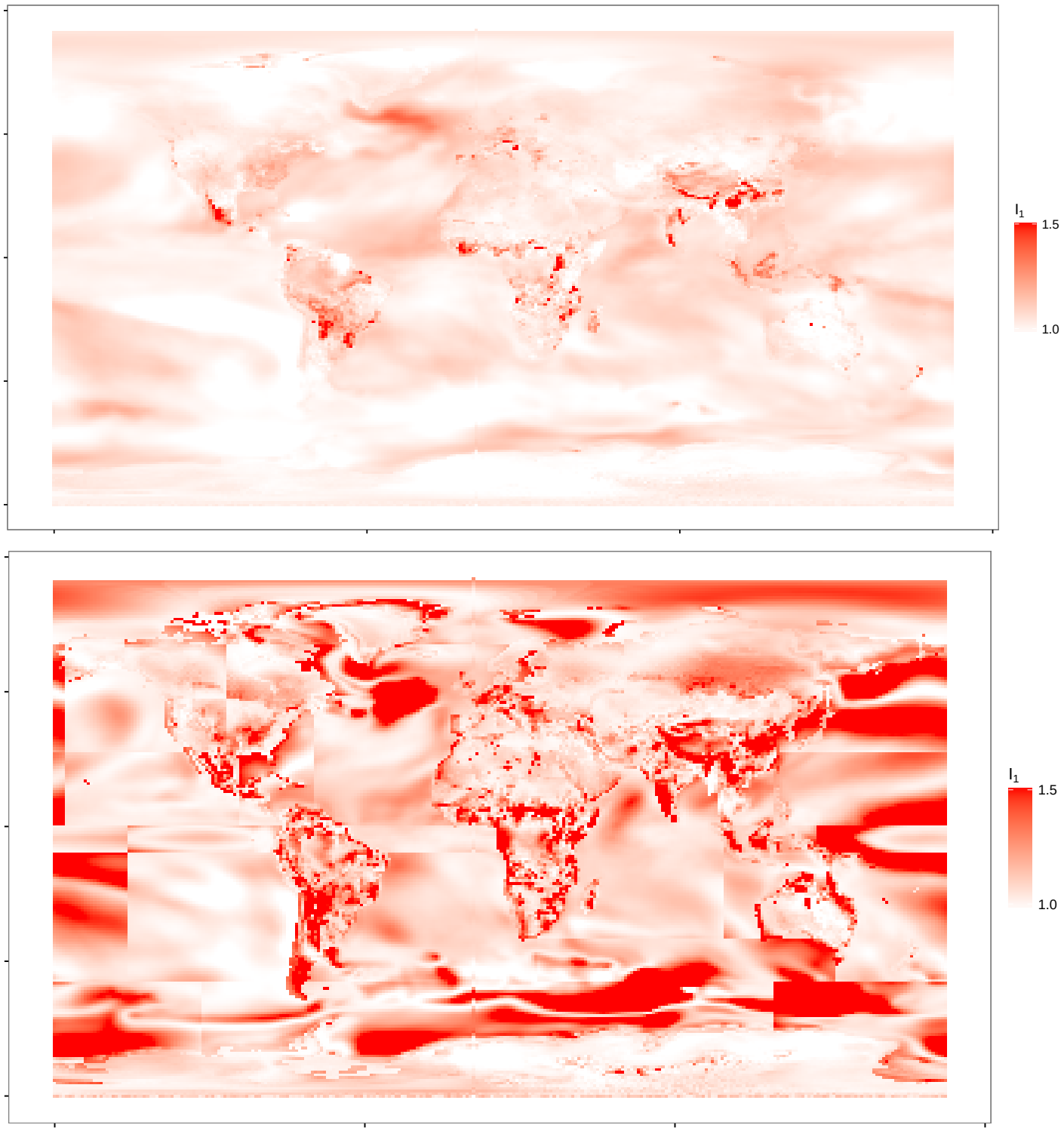

Figure 16: *(top) I_1 for emulation of every pixel, trained on CCSM4 (rcp30, 45, 85), emulating rcp60. This is the same plot as from figure (#). (bottom) I_1 for the pattern scaled pixels, trained on CCSM4 (rcp40, 45, 85), and scaling rcp60. Pattern scaling performs significantly worse than direct emulation. The average I_1 for direct emulation is 1.03, the average for pattern scaling is 1.24 .*

We found that direct pixel level emulation outperforms pattern scaling for nearly all pixels.

III. Summary of Section

Direct pixel level emulation gives values of $I_1 \approx 1$ and $I_{2,pixel} \approx \frac{1}{2} I_{2,region} \gg 1$ which indicates that the emulation is accurate. In addition, direct pixel level emulation outperforms alternative techniques such as pattern scaling to emulated regional temperature, which itself outperforms pattern scaling to global mean temperature. Thus, direct pixel level emulation works.

However, in many cases it is still desirable to aggregate into regions. First, impacts assessment often deals with spatially aggregated economic or political regions. Application of emulation to impacts would require aggregating temperature to those regions. Second, there is an improvement in I_2 values in many cases when using region aggregation.

IX. Conclusions

The emulator is able to accurately reproduce the mean temperature trend of nearly all models tested from the CMIP5 archive using the small number of publicly available representative concentration pathway runs.

Emulation is accurate even with a small training set and there are rapidly diminishing returns to adding more data. There appears to be little to no benefit after the second scenario and after the fourth realization. Adding additional scenarios improved the quality of emulator fit more than adding additional realizations of the same scenarios. A training set of 2 realizations each of 2 scenarios (for a total of 4 model runs) is perfectly adequate to get as accurate of an emulation of mean temperature as possible given uncertainty in the true mean.

We could not find evidence that the emulator performs worse when extrapolating scenarios than when interpolating. However, the types of runs in the training set does affect emulator performance over specific scenarios. For example, a training set of only slow scenarios is not adequate to emulate a fast scenario and vice versa. A training set consisting of multiple realizations of a variety of runs appears able to emulate both fast and slow scenarios accurately.

Pixel level emulation is possible. Interannual variability in a pixel is similar to the interannual variability of the regions. Additionally, the I_1 metric indicates that the emulator is accurate over almost all pixels and the I_2 metric confirms that the I_1 values are meaningful. Pixel level emulation also outperforms the alternative method of pattern scaling from a global projection.

X. acknowledgments

We acknowledge the World Climate Research Programme's Working Group on Coupled Modelling, which is responsible for CMIP, and we thank the climate modeling groups (listed in Table 1 of this paper) for producing and making available their model output. For CMIP the U.S. Department of Energy's Program for Climate Model Diagnosis and Intercomparison provides coordinating support and led development of software infrastructure in partnership with the Global Organization for Earth System Science Portals.

References

[1] M Bentsen, I Bethke, JB Debernard, T Iversen, A Kirkevåg, Ø Seland, H Drange, C Roelandt, et al. The norwegian earth system model, noresm1-m-part 1: Description and basic evaluation. *Geoscientific Model Development Discussions*, 5:2843–2931, 2012.

[2] K Caldeira and N P Myhrvold. Projections of the pace of warming following an abrupt increase in atmospheric carbon dioxide concentration. *Environmental Research Letters*, 8(3):034039, 2013.

[3] Stefano Castruccio, David J. McInerney, Michael L. Stein, Feifei Liu Crouch, Robert L. Jacob, and Elisabeth J. Moyer. Statistical emulation of climate model projections based on precomputed gcm runs*. *Journal of Climate*, 27(5):1829 – 1844, 2014.

[4] P Chylek, J Li, MK Dubey, M Wang, and G Lesins. Observed and model simulated 20th century arctic temperature variability: Canadian earth system model canesm2. *Atmospheric Chemistry and Physics Discussions*, 11(8):22893–22907, 2011.

[5] MA Collier, Stephen J Jeffrey, Leon D Rotstayn, KK Wong, SM Dravitzki, C Moseneder, Carlo Hamalainen, JI Syktus, Ramasamy Suppiah, Joseph Antony, et al. The csiro-mk3. 6.0 atmosphere-ocean gcm: participation in cmip5 and data publication. In *International Congress on Modelling and Simulation–MODSIM*. Citeseer, 2011.

[6] William D Collins, Cecilia M Bitz, Maurice L Blackmon, Gordon B Bonan, Christopher S Bretherton, James A Carton, Ping Chang, et al. The community climate system model version 3 (ccsm3). *Journal of Climate*, 19(11):2122–2143, 2006.

[7] J.-L. Dufresne, M.-A. Foujols, S. Denvil, A. Caubel, O. Marti, O. Aumont, Y. Balkanski, et al. Climate change projections using the ipsl-cm5 earth system model: from cmip3 to cmip5. *Climate Dynamics*, 40(9):2123–2165, 2013.

[8] John P Dunne, Jasmin G John, Alistair J Adcroft, Stephen M Griffies, Robert W Hallberg, Elena Shevliakova, Ronald J Stouffer, et al. Gfdl's esm2 global coupled climate-carbon earth system models. part i: Physical formulation and baseline simulation characteristics. *Journal of Climate*, 25(19):6646–6665, 2012.

[9] John P Dunne, Jasmin G John, Elena Shevliakova, Ronald J Stouffer, John P Krasting, Sergey L Malyshev, PCD Milly, et al. GfdlâĂŹs esm2 global coupled climate–carbon earth system models. part ii: Carbon system formulation and baseline simulation characteristics*. *Journal of Climate*, 26(7):2247–2267, 2013.

[10] A. M. Foley, P. B. Holden, N. R. Edwards, J.-F. Mercure, P. Salas, H. Pollitt, and U. Chewpreecha. Climate model emulation in an integrated assessment framework: a case study for mitigation policies in the electricity sector. *Earth System Dynamics*, 7(1):119–132, 2016.

[11] Peter R. Gent, Gokhan Danabasoglu, Leo J. Donner, Marika M. Holland, Elizabeth C. Hunke, Steve R. Jayne, David M. Lawrence, et al. The community climate system model version 4. *Journal of Climate*, 24:4973–4991, 2011.

[12] Marco A. Giorgetta, Johann Jungclaus, Christian H. Reick, Stephanie Legutke, JÃijrgen Bader, Michael BÃűttinger, Victor Brovkin, et al. Climate and carbon cycle changes from 1850 to 2100 in mpi-esm simulations for the coupled model intercomparison project phase 5. *Journal of Advances in Modeling Earth Systems*, 5(3):572–597, 2013.

[13] D Ji, L Wang, J Feng, Q Wu, H Cheng, Q Zhang, J Yang, et al. Description and basic evaluation of beijing normal university earth system model (bnu-esm) version 1. *Geoscientific Model Development*, 7(5):2039–2064, 2014.

[14] C. D. Jones, J. K. Hughes, N. Bellouin, S. C. Hardiman, G. S. Jones, J. Knight, S. Liddicoat, et al. The hadgem2-es implementation of cmip5 centennial simulations. *Geoscientific Model Development*, 4(3):543–570, 2011.

[15] Lijuan Li, Pengfei Lin, Yongqiang Yu, Bin Wang, Tianjun Zhou, Li Liu, Jiping Liu, et al. The flexible global ocean-atmosphere-land system model, grid-point version 2: Fgoals-g2. *Advances in Atmospheric Sciences*, 30(3):543–560, 2013.

[16] Fangli Qiao, Zhenya Song, Ying Bao, Yajuan Song, Qi Shu, Chuanjiang Huang, and Wei Zhao. Development and evaluation of an earth system model with surface gravity waves. *Journal of Geophysical Research: Oceans*, 118(9):4514–4524, 2013.

[17] B. D. Santer, T. M. L. Wigley, M. E. Schlesinger, and Mitchell J. F. B. Developing climate scenarios from equilibrium GCM results. 1990.

[18] Gavin A. Schmidt, Max Kelley, Larissa Nazarenko, Reto Ruedy, Gary L. Russell, Igor Aleinov, Mike Bauer, et al. Configuration and assessment of the giss modele2 contributions to the cmip5 archive. *Journal of Advances in Modeling Earth Systems*, 6(1):141–184, 2014.

[19] Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of cmip5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[20] Detlef P Van Vuuren, Jae Edmonds, Mikiko Kainuma, Keywan Riahi, Allison Thomson, Kathy Hibbard, George C Hurtt, Tom Kram, Volker Krey, Jean-Francois Lamarque, et al. The representative concentration pathways: an overview. *Climatic change*, 109:5–31, 2011.

[21] A. Voldoire, E. Sanchez-Gomez, D. Salas y Mélia, B. Decharme, C. Cassou, S. Sénési, S. Valcke, I. Beau, A. Alias, M. Chevallier, M. Déqué, J. Deshayes, H. Douville, E. Fernandez, G. Madec, E. Maisonnave, M.-P. Moine, S. Planton, D. Saint-Martin, S. Szopa, S. Tyteca, R. Alkama, S. Belamari, A. Braun, L. Coquart, and F. Chauvin. The cnrm-cm5.1 global climate model: description and basic evaluation. *Climate Dynamics*, 40(9):2091–2121, 2012.

[22] Masahiro Watanabe, Tatsuo Suzuki, Ryouta O'ishi, Yoshiki Kormuo, Shingo Watanabe, Seita Emori, Toshihiko Takemura, et al. Improved climate simulation by MIRCO5: Mean states, variability, and climate sensitivity. *Journal of Climate*, 23:6312–6335, 2010.

[23] S Watanabe, T Hajima, K Sudo, T Nagashima, T Takcmura, H Okajima, T Nozawa, H Kawase, et al. Miroc-esm 2010: model description and basic results of cmip 5-20 c 3 m experiments. *Geoscientific Model Development*, 4(4):845–872, 2011.

[24] Tongwen Wu, Rucong Yu, Fang Zhang, Zaizhi Wang, Min Dong, Lanning Wang, Xia Jin, Deliang Chen, and Laurent Li. The beijing climate center atmospheric general circulation model: description and its performance for the present-day climate. *Climate Dynamics*, 34(1):123–147, 2008.

[25] Seiji Yukimoto, Yukimasa Adachi, Masahiro Hosaka, Tomonori Sakami, Hiromasa Yoshimura, Mikitoshi Hirabara, Taichu Y Tanaka, et al. A new global climate model of the meteorological research institute: Mri-cgcm3âĂŤmodel description and basic performanceâĂŤ. *æřůèśąéŻĘèłŇ. çňň 2 èijŕ*, 90(0):23–64, 2012.