Pattern Scaling of Climate Extreme Indices

Claudia Tebaldi, Alex Armbruster

December 2018

We use the pattern scaling technique to emulate the values of 10 climate extreme indices under low warming scenarios, and evaluate the accuracy of our emulation. We propose a way to quantify the error in emulations in the context of experiments that provide initial condition ensembles. Our measure separates systematic error from internal variability. In particular it allows us to compare the size of the unavoidable internal variability of the true quantities targets of the emulation; the internal variability of the emulated quantities and, importantly, the error in the emulation proper. We can compute this measure and its components at a global, regional or grid-point scale. We demonstrate our method in the context of the emulation of a suite of extreme indices of temperature and precipitation from the ETCCDI list. In particular, we test and compare the performance of simple pattern scaling and time shift methods when approximating low warming scenario outcomes, inspired by the Paris agreement, limiting warming to 1.5C with respect to a pre-industrial baseline, on the basis of model output available under higher emission scenarios, RCP4.5 and RCP8.5.

1 Introduction

Pattern scaling is used to emulate grids of general circulation model (GCM) output under forcing scenarios that have not been run on a GCM. This technique allows a researcher who only has access to model output corresponding to one or two RCP's to explore a full range of possibilities for the future climate, under varying assumptions about future human emissions. This technique may be particularly useful for a researcher who is studying a phenomenon that is not captured well in existing GCM output. For example, there is limited model output covering lower warming scenarios (1.5C and 2.0C), but much interest in these scenarios following the Paris agreement. Also, measures of climate extremes are particularly useful for impact assessment modeling, but there are few full datasets of climate extremes. In these situations, pattern scaling can be used to approximate what GCM output might look like for scenarios that have not been run. Our aim is to evaluate how accurately pattern scaling can emulate low warming scenarios. We make use of ensemble simulations corresponding to four forcing scenarios, RCP8.5, RCP4.5, 2.0C and 1.5C. Both of the lower scenarios are emulated using both of the higher scenarios as starting points. We can then evaluate how the accuracy of our emulation depends on the magnitude of the difference between the starting and the target scenario. The actual model output for 2.0C and 1.5C is taken to be the ground truth, against which our emulations are validated. As described by Deser 2012, the range of predictions given by ensemble simulations can be used to distinguish a signal from noise created by internal variability. In the context of our experiment, we use the multiple ensemble members to decompose the total error (difference between the target value and our predicted value) into a component that comes from the internal variability of the index, and another component that corresponds to the error introduced by pattern scaling.

Tebaldi and Arblaster 2013 established the accuracy of pattern scaling for temperature and precipitation using the variation between models in the CMIP5 ensemble as a benchmark. They found that pattern scaling produces more accurate emulations for temperature than for precipitation, but that the uncertainty in emulations of both is significantly smaller than the disagreement between models in the CMIP5 ensemble. We extend this analysis to ten climate extreme indices, but use the initial condition uncertainty as our benchmark for accuracy, rather than inter-model variability.

2 Methods

We evaluate the accuracy of pattern scaling when used to create emulations for 10 climate extremes defined by ETCCDI[1] and computed based on NCAR-CESM temperature and precipitation projections. These indices are total annual precipitation when daily rainfall exceeds the 95th percentile (r95ptot), monthly maximum consecutive five day precipitation (rx5day), warm spell duration index (wsdi), frost days (fd), annual count of days when precipitation exceeds 10mm (r10mm), average monthly minimum temperature (tnn), average monthly maximum temperature (txx), and average precipitation intensity (sdii), and maximum number of consecutive days with precipitation is less than 1mm (cdd). These indices are helpful for impact researchers who model the effect of extreme climate events on humans systems. The pattern scaling technique would be useful for an impact researcher who only has one scenario of extreme indices available, but wants to compare the impacts of extreme events across multiple scenarios.

2.1 Pattern Scaling

We use two different methods for approximating lower warming scenarios from higher ones. We evaluate two methods for pattern scaling, both of these methods are dependent upon the assumption that local climate can be modeled as a linear function of GAT, with some slope that differs from grid point to grid point. The first, called linear rescaling, also sometimes known as simple pattern scaling, is done by multiplying the value for local climate, at each grid point, by a scaling factor - century GAT change in the target scenario over century GAT change in the starting scenario. For example, if we were to approximate local temperature in scenario y using model output from

[1] http://etccdi.pacificclimate.org/

scenario x, our approximation for local temperature at gridpoint (i,j) using linear rescaling would be

$$\hat{temp}_{i,j}^{y} = temp_{i,j}^{x} \star \frac{\Delta_y^{GAT}}{\Delta_x^{GAT}} \tag{1}$$

where Δ^{GAT} is the difference between the 1995-2015 GAT average, and the 2080-2100 GAT average. When scenario y is the lower warming scenario, this scaling factor will always be less than one.

The second method for pattern scaling is the timeshift approach. If we were to approximate lower warming scenario y using higher warming scenario x, we would find a five year window when GAT from scenario x is very close to the end of century average of scenario y. We would then take the climate in scenario y, averaged over this five year window as our approximation for scenario x.

2.2 Measurement of Error

Our model output contains 10 initial condition ensembles, and we pattern scale for each of these 10 separately, so that we end up comparing 10 emulations against 10 ground truth values. In this setting we are able to distinguish internal variability (uncertainty that is present in the climate model output, and is thus treated as unavoidable) from systemic error, introduced in areas where pattern scaling is not an effective approximation method. If $\mathbf{y}$ is the 10 initial condition ensemble members which correspond to the ground truth, and $\hat{\mathbf{y}}$ is our approximation of these ten values, our error metric is

$$\frac{(\bar{\mathbf{y}} - \bar{\hat{\mathbf{y}}})^2 + var(\hat{\mathbf{y}})}{var(\mathbf{y})} \tag{2}$$

Our initial error metric was the average of all the possible differences between a member of $\hat{\mathbf{y}}$ and $\mathbf{y}$. This simplifies to the sum of the three terms above, meaning that each term is a component of the total error. We rearranged the terms as shown above to avoid over-penalizing a quantity that is uncertain to begin with, or in other words so that our emulation is to blame for any increase in the error.

The 2.0C and 1.5C forcing scenarios are both approximated using both RCP4.5 and RCP8.5, with both the time shift method and simple pattern scaling. Two target scenarios, times two starting scenarios, times two pattern scaling methodologies gives us eight total emulations. We can then see which combinations of starting, target, and emulation method produce the most accurate emulations.

3 Results

Figure 1 displays the emulation error as a function of end of century GAT difference between starting and target scenarios. The warm spell duration index (WSDI) is a clear outlier. Simple pattern scaling performs extremely poorly for WSDI, and the time shift approach performs better than simple pattern scaling but still extremely poorly compared to the other indices. The failure of simple pattern scaling can be explained by the distributional properties of wsdi - it is defined by the sum of days where daily average temperature is above the 90^{th} percentile of daily average temperature at each grid point, in streaks of at least 6. A distribution can be assumed for temperature, and as the center of this distribution shifts to the right, the probability of observing several consecutive values that were extreme with respect to the original distribution increases greater than linearly. Consequently, the central assumption of pattern scaling - that local climate is related linearly to GAT is violated. The time shift approach is more accurate than linear rescaling for WSDI which suggests that the time shift approach may be more resilient to indices that have a non linear relationship with GAT. WSDI is removed from the right panel of figure 1, and from Figures 2 and 3 to allow for a closer look at the emulation performance of the other indices.

Figures 2 and 3 break down the error by component for each index and for each scaling. The size of each bar is computed by aggregating over the entire grid. All of the error components are normalized by the variance between the 10 ensemble members of the ground truth (the blue bar that is of length 1 for each index, the denominator of our error metric). 1 is the natural threshold when using our error metric, as this is the point where the error in the emulation outweighs the uncertainty in the ground truth. The orange dotted line in Figures 1, 2, and 3 corresponds to this threshold.

We find that simple pattern scaling is generally more accurate than the time shift approach. The only index for which this is not the case is WSDI. Comparing Figures 2 and 3 shows that this is generally because the green components of the bars (which represent variance within the emulation) are smaller when the simple pattern scaling approach is used. This is to be expected, as the procedure for simple pattern scaling includes dividing the ten ensemble members by a constant scaling factor. The fact that this decrease in variance does not come at the expense of a greater skew between the means of our emulation and the ground truth means that simple pattern scaling works well.

We find that there is no clear relationship between the distance between the starting and target scenario and the emulation error. It may seem like a reasonable assumption that more similar forcing scenarios can be used to approximate each other more accurately than less similar ones, but this does not seem to be the case. In fact, the approximations of index values in the 1.5C scenario are generally more accurate than approximations of the 2.0C scenario, even though our two starting scenarios are closer to 2.0C. Regardless of the starting scenario (RCP4.5 or RCP8.5) all indices with the exception of WSDI are more accurately emulated for 1.5C than for 2.0C. With RCP4.5 as the starting point, these nine indices (WSDI excluded) have, on average, have an error score than is 0.3 higher in 2.0C and in 1.5C. When RCP8.5, this error difference is less pronounced (0.1), but is still consistently in the opposite direction than one might expect.

4 Figures

Figure 1

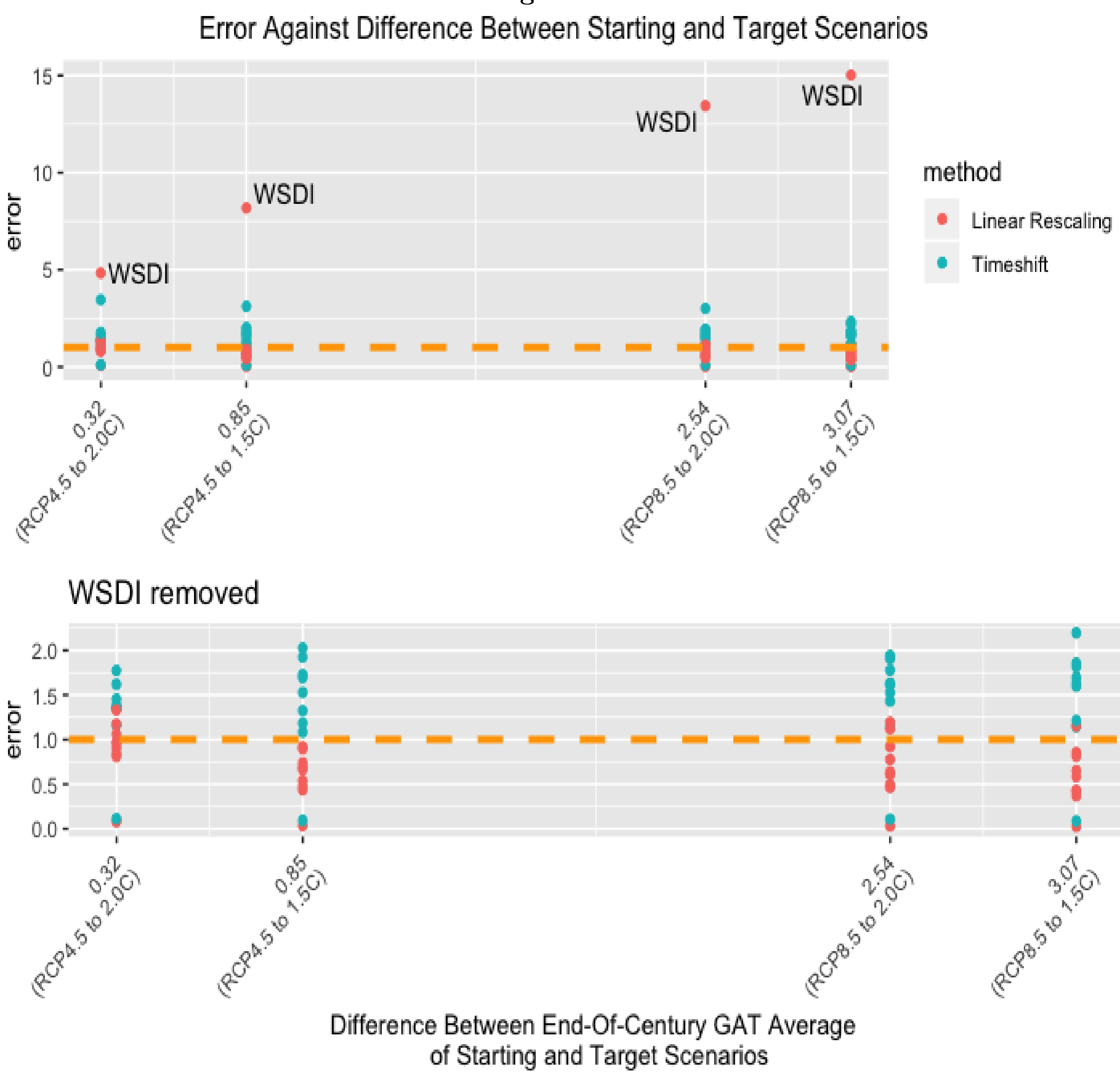
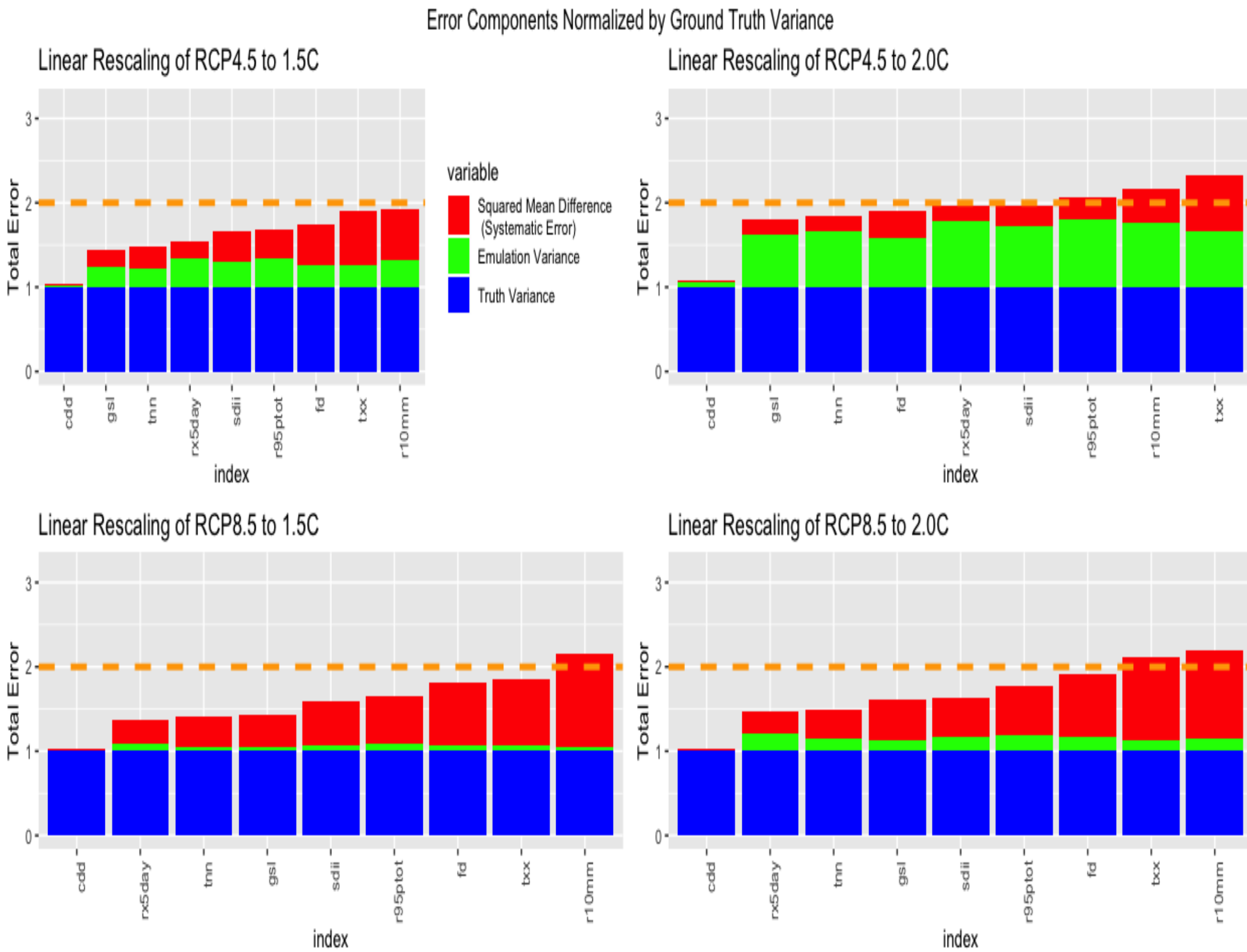


Figure 2



Error Components Normalized by Ground Truth Variance
Linear Rescaling of RCP4.5 to 1.5C
Linear Rescaling of RCP4.5 to 2.0C
Linear Rescaling of RCP8.5 to 1.5C
Linear Rescaling of RCP8.5 to 2.0C
Total Error
index
variable
Squared Mean Difference (Systematic Error)
Emulation Variance
Truth Variance
cdd
gsl
tnn
rx5day
sdii
r95ptot
fd
txx
r10mm

Figure 3

Error Components Normalized by Ground Truth Variance