Quantifying the Independence of Climate Models

Yoni Ackerman

Introduction

Anthropogenic climate change is the most dangerous threat facing global human society. In order to predict spatio-temporal changes in global climate under a variety of scenarios and assumptions, climate researchers use computational models to simulate long-term atmosphere and ocean dynamics. There is hope that these models can provide policy makers with predictive tools for climate change preparedness and mitigation (see, for example, this link). This goal, however, has a statistical impediment: model uncertainty. Much research has gone into understanding and limiting the sources of model uncertainty (Hawkins and Sutton 2009; Brohan et al. 2006; Regier and Stark 2013). Despite this progress, there remains a deep concern regarding the effect of model non-independence on uncertainty quantification

Even between research groups, climates models often share code modules, so their outputs do not represent independent draws from a space of all possible climate climate states. Thus, agreement in model predictions does not grant greater certainty in their consensus results (Larose et al. 2005; Pirtle, Meyer, and Hamilton 2010).

Work has been done to accomodate the issue of model non-independence. Knutti et. al. describe a method to interpolate model characteristics that takes into account their non-independence (Sanderson, Knutti, and Caldwell 2015). By then using model accuracy (relative to historical observations) to inform a prior over the interpolated space, independent samples of climate characteristics can be drawn and used in analyses.

We began our project attempting to extend Knutti's methods by incorporating more variables in their compression procedure. This proved relatively fruitless: it was unclear how adding these variables improved their analysis given the severe limitations on available historical observations. Instead we chose to explore the critical issue underlying these models's utility: their mutual non-independence. In our study of the literature, model non-independence is always assumed, but never tested for and never quantified. We decided to focus our project on: (1) proving to ourselves that the models are not independent, (2) quantifying their degree of non-independence using as much of the model data as possible, and (3) finding a way to visualize the dependence relationships between the models.

Data Summary

We used data from the Coupled Model Intercomparison Project (cmip5) made available by ETH Institute for Atmospheric and Climate Science, as well as observational data gathered according to (Knutti 2010). Each model included some combination of 36 variables (see table in the appendix), produced using a combination of parameters. They were: simulated under some combination of 10 different scenarios; aggregated in daily, monthly, and annual time-scales; and spatially distributed in two available coordinate grids (one provided by the model source, and another interpolated to a 2.5 by 2.5 degree grid by ETH). Furthermore, each model output was in fact a group average of a collection of model runs, known as ensembles, each with its own set of initial conditions. 47 different ensembles were used to generate the ETH data.

To get an idea of the data's temporal structure see figure 1. This plot shows time series data for surface temperature (tas) taken from the ACCESS1-3 model, under the historicalGHG scenario and r1i1p1 ensemble, at longitude-latitude location (126.25, -58.75). These data display the expected seasonal signature which can be seen in the auto-covariance plot in left of figure 2.

Similarly, we can look at auto correlation in the spatial domain. Here again we find strong auto-correlations, shown in the right of figure 2, but this time of the spatial variety.

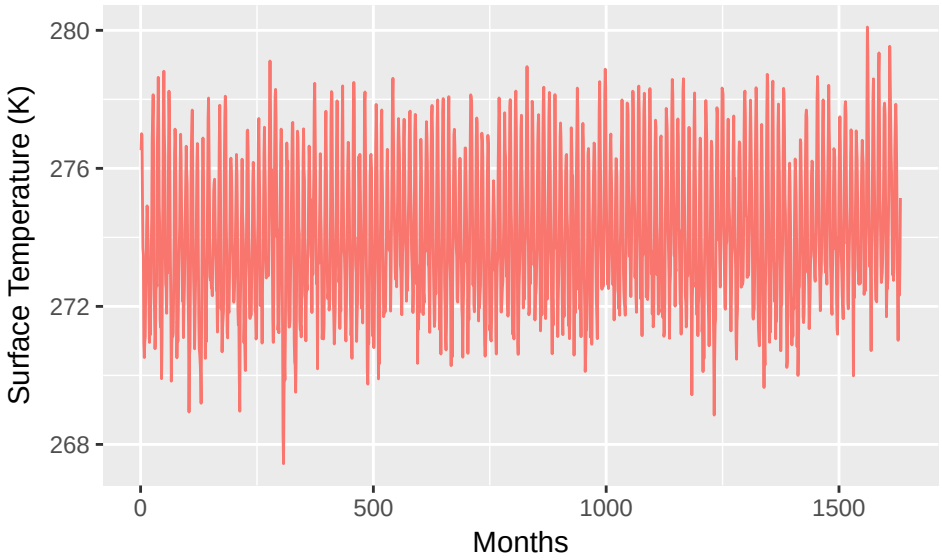

Figure 1: The output from model ACCESS1-3 (Scenario: historicalGHG. Ensemble: r1i1p1) for variable tas at longitude-latitude location (126.25, -58.75) is shown. Although it is hard to tell at the scale shown, the data have a distinct seasonal pattern.

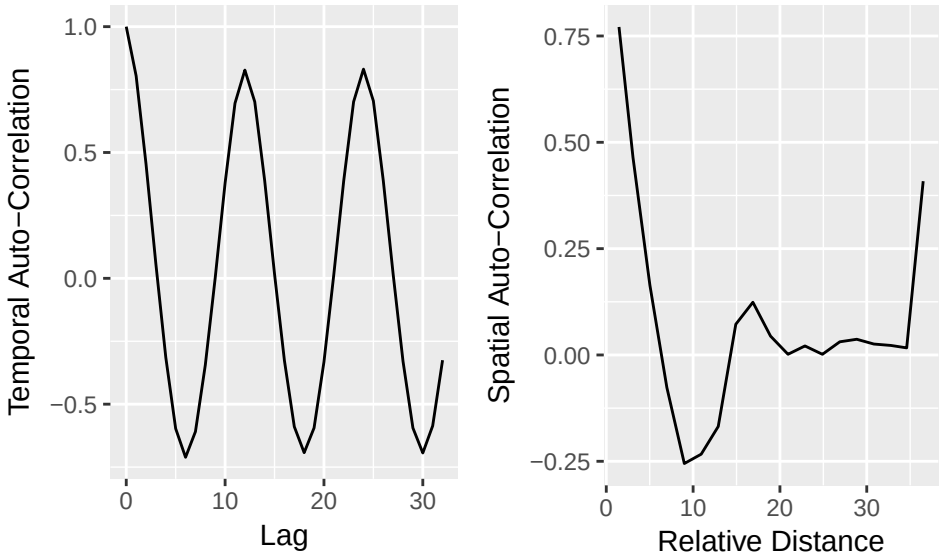

Figure 2: Spatial correlation of tas from model ACCESS1-3 (Scenario: historicalGHG. Ensemble: r1i1p1) 231 months after simulation start is shown on the right. Temporal autocorrelation for tas from the same model/scenario/ensemble are shown on the left. Both demonstrate patterns typical of their respective genres.

Methods

There were three parts to our methods: initial testing, data compression, and final testing.

Initial testing

Our first goal was to establish that the cmip5 models are indeed not independent. A variety of time series methods test for the independence between series (Hong 1996; Koch 2013). We chose, somewhat arbitrarily, to apply the techniques put forward by Hong et. al. The test procedure for testing two time series X_t and Y_t is as follows (paraphrasing from Zhenqi):

- Step 1. The assumption of covariance stationarity implies that X_t, Y_t have an AR(∞) representation. So fit AR models, denoted as $X_t(p)$ and $Y_t(q)$, to X_t and Y_t.
- Step 2. Run ordinary least square regression of X_t on $X_t(p)$ and Y_t on $Y_t(q)$ and obtain their least square residuals: $\hat{u}_t$ and $\hat{v}_t$, respectively.
- Step 3. Calculate residual cross-correlation series. For all $j \in [1-N, N-1]$, where N is the minimal items of series $(X_t(p), Y_t(q))$, define residual cross-correlation as:

$$\hat{\rho}_{uv}(j) = \frac{\hat{R}_{uv}(j)}{\sqrt{\hat{R}_{uu}(0)\hat{R}_{vv}(0)}} \qquad \hat{R}_{uv}(j) = \begin{cases} N^{-1}\sum_{t=j+1}^{N} \hat{u}_t\hat{v}_{t-j}, \text{ if } j \geq 0 \\ N^{-1}\sum_{t=-j+1}^{N} \hat{u}_{t+j}\hat{v}_t, \text{ if } j < 0 \end{cases}$$

$$\hat{R}_{uu}(0) = N^{-1}\sum_{t=1}^{N} \hat{u}_t^2 \qquad \hat{R}_{vv}(0) = N^{-1}\sum_{t=1}^{N} \hat{v}_t^2$$

- Step 4. Choose a kernel function k that is symmetric and continuous at 0 and a smoothing parameter M that together satisfy the following conditions:

$$k : \mathbb{R} \to [-1, 1] \qquad k(0) = 1 \qquad \int_{-\infty}^{\infty} k^2(z)dz < \infty \qquad M = M(N) \to \infty \qquad M/N \to 0$$

- Step 5. Construct test statistic Q:

$$Q = \frac{N\sum_{j=1-N}^{N} k^2(\frac{j}{M})\hat{\rho}_{uv}(j) - S_N(k)}{\sqrt{2D_N(k)}}$$

$$S_N(k) = \sum_{j=1-N}^{N-1} (1 - \frac{|j|}{N})k^2(\frac{j}{M}) \qquad D_N(k) = \sum_{j=2-N}^{N-2} (1 - \frac{|j|}{N})(1 - \frac{|j|+1}{N})k^4(\frac{j}{M})$$

Under the null hypothesis (X_t and Y_t are independent), the test statistic Q is asymptotically standard normal. Furthermore, Q is comprised of a linear combination of all the lagged correlations between X_t and Y_t.

This test requires that both time series be covariance stationary, which the cmip5 data was not. To achieve covariance stationarity, we chose to transform one dimensional time series by taking first differences. After doing so, we used the Augmented Dickey Fuller test to assess whether or not the coverted series met the requirement. No nulls were rejected.

We then conducted Hong et. al.'s test on all the pairs of the 37 models from the cmip5 RCP45 scenario using variables surface temperature (tas) and precipitation (pr). Hong's method requires two time series for the test, limiting how much data we could test at once. We chose the following procedure:

- Step 1. For each model pair, sample 32 spatial points from the model grids. Call these temporal series $X_{k,t}$ and $Y_{k,t}$ for $k \in \{1, ..., 32\}$.
- Step 2. For each k calculate Q using $X_{k,t}$ and $Y_{k,t}$, and the Daneill kernel, then record the p-value as p_k.
- Step 3. For each model pair, calculate $\bar{p} = \frac{1}{32} \sum_k p_k$.

Data compression

The 46 models in cmip5 collection were on average 1.3 gigabytes in size. Thus one of our initial questions was how to go about comparing two models in R, given its memory limitations. Previous work (such as that by Knutti et. al.) solved this problem by compressing each model using principle component analysis. Even compressing in this way, including more than four variables in the compressed result would still have overwhelmed R. With this in mind, we chose to work with distances between models, rather than compressed forms of the models themselves. We first define $M_i(v,t) \in \mathbb{R}^2$ to be the spatial data of model M_i for variable v at time t. We then complete the following steps:

- Step 1. For a single v and t, vectorize each $M_i(v,t)$ into a single row vector $r_i(v,t)$.
- Step 2. Row-bind the $r_i(v,t)$ into a single model matrix $M(v,t)$, then center and scale this new matrix's columns.
- Step 3. Calculate $D(v,t)$ to be the euclidean distances between the rows of $M(v,t)$, i.e. $D(v,t)_{i,j} = ||r_i(v,t) - r_j(v,t)||$.
- Step 4. For each t, computed $D(v_k,t)$ for each variable v_k, and form $D(t) = \sqrt{\sum_k D^2(v_k,t)}$ where the squares are taken element wise.

Steps 5-7: For simplicity, perform classical multidimensional scaling on each $D(t)$:

- Step 5. Let $B = -\frac{1}{2} J D(t) J$ where $J = I - \frac{1}{N} \mathbf{1}\mathbf{1}^T$ and N is the number of models. This step centers the data.
- Step 6. Let Q such that $B = Q^T \Lambda Q$, where Λ are the eigenvalues of B$. This step decomposes the data into it's principle components.
- Step 7. Choose the two columns of Q associated with the largest eigenvalues (call this matrix $\tilde{Q}$) and form $Y(t) = \tilde{Q}\Lambda^{1/2}$. This step selects the two principle components that capture the largest amount of the variation.

The resulting $Y(t)$ has as many rows as there are models but only two columns. Furthermore, the distances between the rows of $Y(t)$ are faithful to the distances of $D(t)$, making $Y(t)$ an ideal tool for visualizing our data. See figures 3 and 4 for a summary of the $Y(t)$ output.

Final testing

We have so far come up with three ways to test for independence using matrix time series $Y(t)$. Unfortunately, they are all still somewhat rough around the edges. Nonetheless, we believe they are steps in the right direction. Below, we describe the three tests, then given more details on the methods used in carrying them out.

Test 1: clustering

This test uses the KNN clustering algorithm to create a test statistic which is binomial under the null (the models are independent). It's reasoning is as follows:

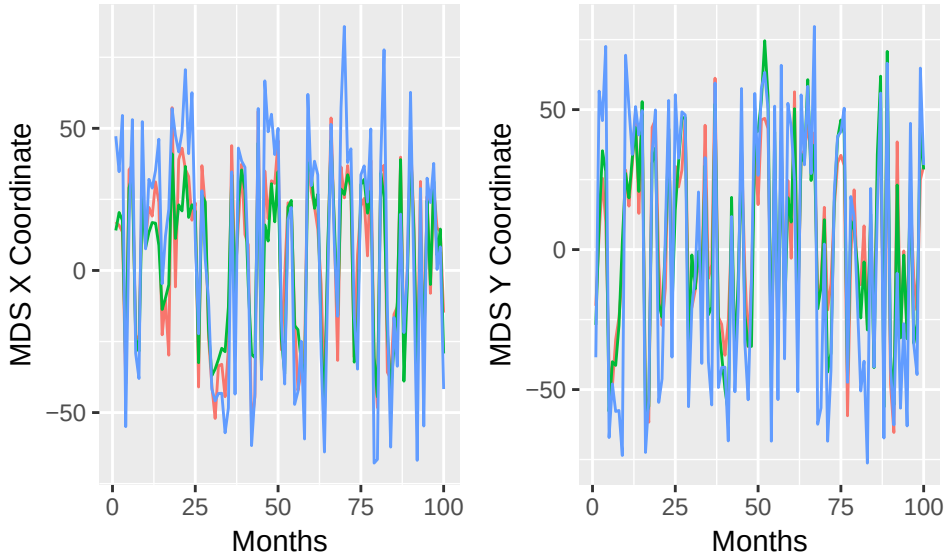

Figure 3: First 100 months of output of the multidimensional scaling compression procedure shown for models ACCESS1-0 (red), ACCESS1-3 (green), CSIRO-Mk3-6-0 (blue). The left plot shows the time series of X coordinates, the right shows Y coordinates.

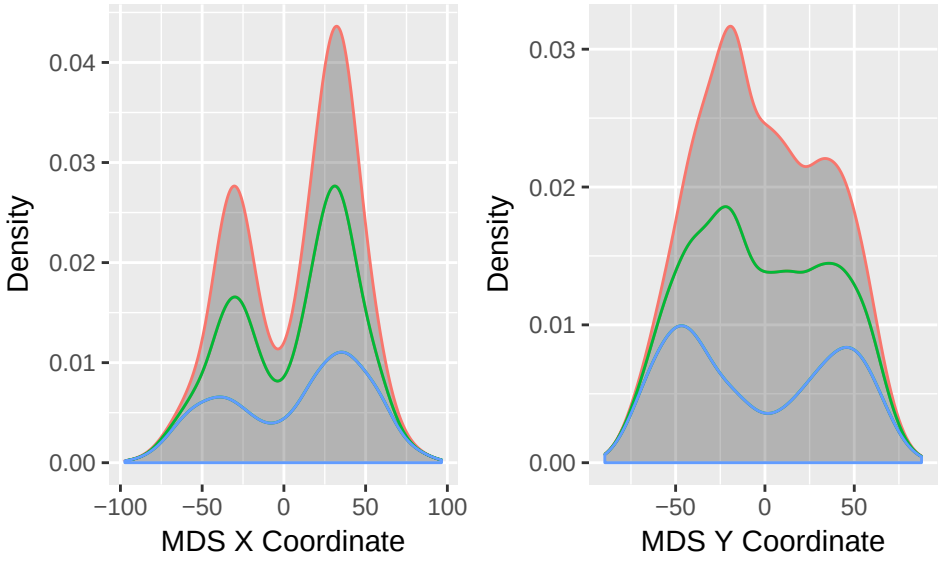

Figure 4: Empirical densities of the output of the multidimensional scaling compression procedure shown for models ACCESS1-0 (red), ACCESS1-3 (green), CSIRO-Mk3-6-0 (blue). The left plot shows the densities of the X coordinates, the right shows densities of the Y coordinates. This is not something we paid much attention to, however we guessed that the bimodal nature of these distributions is tied somehow to the seasonal aspects of the original data.

Due to the indepence assumption under the null, at each timepoint t, the probability that two models of $Y(t)$, $M_i = (x_i, y_i)$ and $M_j = (x_j, y_j)$, are in the same cluster C_n ($n \in \{1, ..., N\}$) can be written as:

$$\sum_n \mathbb{P}(M_i \in C_n \text{ and } M_j \in C_n) = \sum_n \mathbb{P}(M_i \in C_n)\mathbb{P}(M_j \in C_n)$$

Then, since the C_n are exchangeable, we write:

$$\begin{aligned}
\sum_n \mathbb{P}(M_i \in C_n)\mathbb{P}(M_j \in C_n) &= \sum_n \mathbb{P}(M_i \in C_1)\mathbb{P}(M_j \in C_1) \\
&= N\mathbb{P}(M_i \in C_1)\mathbb{P}(M_j \in C_1) \\
&= N\frac{1}{N^2} \\
&= \frac{1}{N}
\end{aligned}$$

Following this calculation, if we set a cluster size N, cluster $Y(t)$ into N clusters at each $t \in \{1, ..., T\}$, and calculate the proportion p_{ij} of times any two models M_i and M_j are in the same cluster, then under the null $T * p_{ij} \sim Binom(T, \frac{1}{N})$ - a hypothesis we can easily test with a two-tailed z-test.

This procedure has one major flaw: how do we choose N. If we choose N too small, then even independent series can appear dependent, too large and the opposite will happen. In fact, we are not entirely sure that their is a correct choice of N: we have not yet carefully gone through the KNN algorithm to double check the probabilistic assumptions we made above.

Test 2: nearest neighbor

This test does away with the KNN portion of Test 1, but requires a new (and undesireable) assumption under the null: that models M_i are IID. Under this assumption and given a model M, we can calculate the probability that any other model M_i will be M's nearest neighbor (roughly):

$$\begin{aligned}
\mathbb{P}(dist(M, M_i) < min_{j \neq i} dist(M, M_j)) &= \mathbb{P}(\sqrt{M - M_i} < min_{j \neq i}\sqrt{M - M_j}) \\
&= \mathbb{P}(M - M_i < min_{j \neq i} M - M_j) \\
&= \mathbb{P}(M_i > max_{j \neq i} M_j) \\
&= \frac{1}{N - 1}
\end{aligned}$$

Thus, choosing a model M_i we calculate its nearest neighbor at every timepoint. We then calculate the proportion p_{ij} of times M_j was M_i's nearest neighbor. We can then test whether the set $(p_{ij})_{j \neq i}$ differs significantly from its expected value under the null $\frac{1}{N-1}$ (where N is the number of models) with, again, a two-tailed z-test. As hinted at before, the null hypothesis for this test is not quite right: if the test result is significant models M_i might be non-independent, non-identically distributed, or neither - we'll reject the null but not know exactly why.

Test 3: binomial correction

This test is in the same vein as Test 1, but attempts to do away with having to cluster. Assuming independence, we can calculate the probability that any two models are in the same quandrant of the plane. This probability is $\frac{1}{4}$, and this procedure has the benefit (unlike in clustering) that extreme-value points will not demand a cluster for themselves. Just as in Test 1, this results in a proportion $T * p_{ij} \sim Binom(T, \frac{1}{4})$, which we can again test using a two-tailed z-test.

Final testing specifications

We performed the compression procedure, Test 1 (using six clusters), and Test 2 (Test 3 on the way) on monthly data from the abrupt4xCO2 scenario, outputted from the r1i1p1 ensemble. There were some constraints to using this subset of the data: not all models had the same run-times (some went 3600 months, others went 1800 months or less) and some models did not contain data for all variables. To addres this, We chose a subset of 23 models (see results) each with a run-time of at least 1800 months and each containing 2D spatial variables: clt, evspsbl, hfls, hfss, pr, psl, rlds, rlus, rlutcs, rlut, rsds, rsdt, rsut, sic, tas, tos (see table in the appendix for details).

Results

Initial testing results

The results in our initial testing provide strong evidence against the null hypothesis (the models are independent). The p-values for each model pair are shown in figure 5. The darker the square corresponding, the more significant the p-values. These results are evidence for what all the papers we encountered assumed: the models are not independent - at least not in tas and pr.

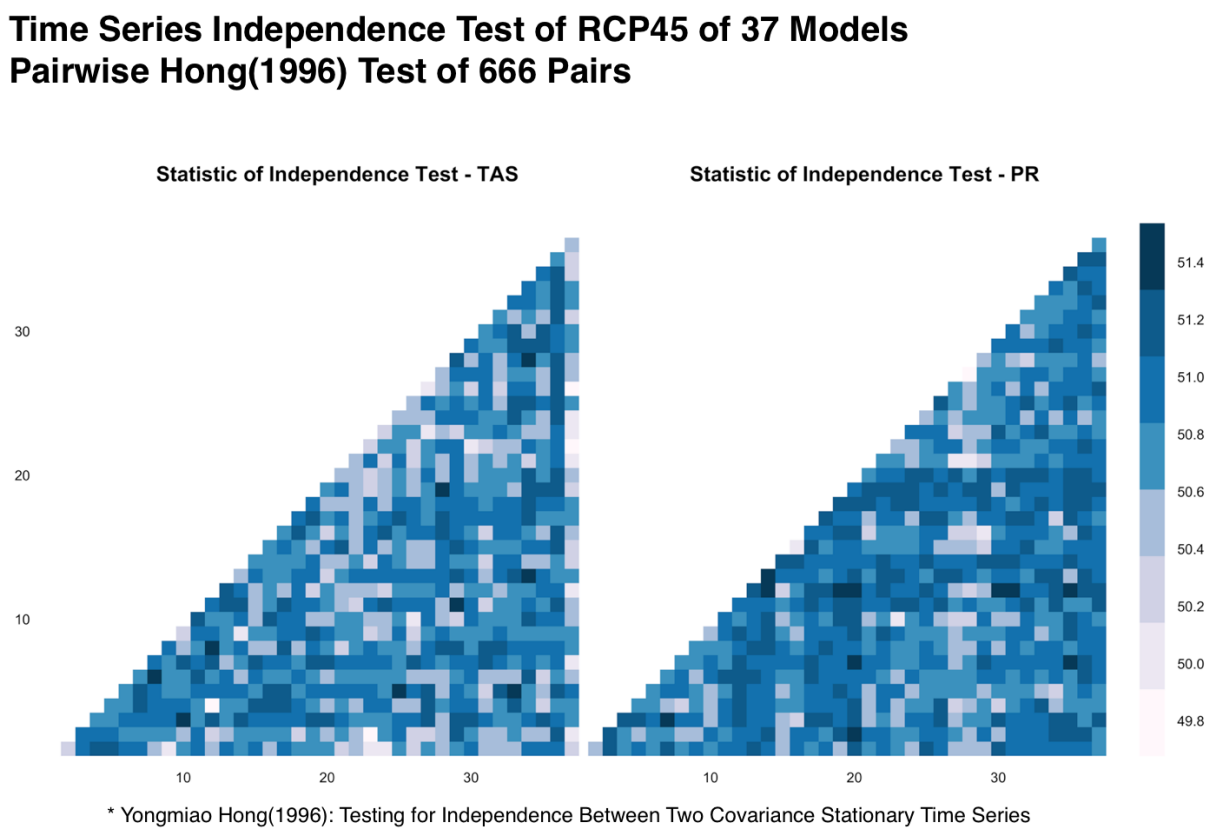

Figure 5: Results of initial testing are shown for two variables, tas and pr. The models are on the two axes. The square at intersection of any row i and column j represents the p-value of the test between the model of row i and the model of column j. Darker squares indicate more significant p-values, however all the results provide strong evidence against the null.

Final testing results

The results of our final testing procedure are best displayed in heatmap form. The results of the two tests are shown in figure 6. Rows represent the base model for the tests and each square in the row indicates whether the model corresponding to that column shared a significant result with the base model (light-blue implying significance). As was expected, in both tests models tend to be associated in ways that would not be expected under the respective null hypotheses. There are some inconsistencies between the two tests. For instance, models bcc-csm1-1 and MIROC5 share a significant result in Test 1, but not in Test 2. One reason for this could be the symmetry enfored by clustering in Test 1. More unsettling, however, is the lack of overlap between non-significant results in the two tests. This is an indication that there is something wrong. Since Test 1 is (we thought) specifically testing for independence and Test 2 is testing for IID, anything that fails Test 1 should also fail Test 2. The fact that this is not the case indicates what we suspected: something is wrong with the assumptions of Test 1, Test 2 or both. Lastly, if we assume Test 1 is in fact testing for independence, then what do we make of the result that the test between models CanESM2 and bcc-csm1-1-m (among other pairs) fails to reject the null? Is this just due to chance, and if so, can we quantify that probability?

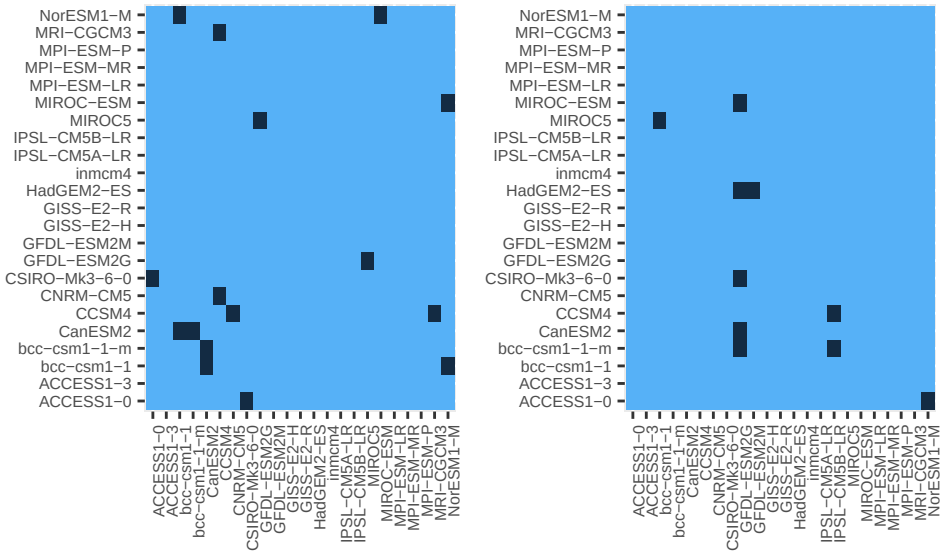

Figure 6: Heat maps for the results. Test 1 is on the left and Test 2 is on the right. A light-blue box indicates a significant result, thus we reject the null hypothesis that the two models in question are independent/IID. A dark blue box indicates the test failed to reject the null hypothesis. Notice that there is no overlap between the non-significant results in the two tests.

Conclusions

We believe we have met, or at least partially met, all three of our goals. In our initial testing, using methods from Hong et. al., we provided evidence that individual tas time series located at spatial points are not independent for any model pair.

In our final testing procedure, we found there to be little doubt that the cmip5 climate models are not independent. In addition we provided two rough tests (one in process) that attempt to quantify the degree of inter-model non-dependency. Instead of getting reliable results, however, we encountered some of the interesting challenges that result when trying to test for independence. There are various ways in which we can improve and extend our methods, the most obvious being an examination of how cluster size affects the results of Test 1. Furthermore, in any test requiring the use of multi-dimensional scaling (such as Tests 2 and

3), we really need to be clear about how our dimension reduction technique may affect the null hypothesis of IID or independence.

Finally, we visualized our results using a heatmap. We also experimented with visualizing the results as a graph with vertices and edges corresponding to models and the results of their mutual tests. In further work, we would like to explore how graph theoretical results could help us to better understand the intermodel model relationships we found.

The most natural extension of our work is to apply it to more model scenarios and ensembles. We have only looked at three scenarios (historicalGHG, RCP45, and abrupt4xCO2), each under a single ensemble (r1i1p1), and we have only run our tests on one of them (abrupt4xCO2). There are 10 scenarios and 47 different ensembles in total - examining how results vary across them will provide us with more information about the models' relationships and the usefulness of our tests.

Appendix

List of Variables

tas	surface air temperature
tasmax	maximum surface air temperature
tasmin	minimum surface air temperature
clt	total cloud cover
evspsbl	evaporation
rsds	downward shortwave radiation at the surface
rsus	upward shortwave radiation at the surface
rsut	upward shortwave radiation at the top of the atmosphere
rtmt	net radiation at top of the model
rlds	downward longwave radiation at the surface
rlut	upward longwave radiation at the top of the atmosphere
rltucs	outgoing clear-sky longwave radiation at the top of the atmosphere
rlus	upward longwave radiation at the surface
hfls	upward latent heat flux at the surface
hfss	upward sensible heat flux at the surface
pr	total precipitation
psl	atmospheric pressure
sos	ocean surface salinity
tos	ocean surface temperature
sic	sea-ice concentration
sit	sea-ice thickness
snd	snow depth
mrro	total runoff
mrros	surface runoff
mrso	total soil moisture content
mrsos	moisture in the upper part of the soil column
gpp	carbon mass flux out of atmosphere due to gross primary production on land
nbp	carbon mass flux out of atmosphere due to net biospheric production on land
npp	carbon mass flux out of atmosphere due to net primary production on land
ra	carbon mass flux into atmosphere due to autotrophic (plant) respiration on land
rh	carbon mass flux into atmosphere due to heterotrophic respiration on land
lai	leaf area index

References

Brohan, Philip, J. J. Kennedy, I. Harris, S. F B Tett, and P. D. Jones. 2006. "Uncertainty Estimates in Regional and Global Observed Temperature Changes: A New Data Set from 1850." *Journal of Geophysical Research Atmospheres* 111 (12): 1–21. doi:10.1029/2005JD006548.

Hawkins, Ed, and Rowan Sutton. 2009. "The Potential to Narrow Uncertainty in Regional Climate Predictions." *Bulletin of the American Meteorological Society* 90 (8): 1095–1107. doi:10.1175/2009BAMS2607.1.

Hong, Y. 1996. "Testing for Independence Between Two Covariance Stationary Time Series." *Biometrika* 83 (3): 615–625.

Knutti, Reto. 2010. "The End of Model Democracy?" *Climatic Change* 102 (3): 395–404. doi:10.1007/s10584-010-9800-2.

Koch, Paul D. 2013. "A Method for the Independence of Two Testing Time Series That Accounts for a Potential Pattern in the Cross-Correlation Function." *Journal of the American Statistical Association* 81 (394): 533–544. doi:10.1080/01621459.1986.10478301.

Larose, Simon, Catherine F Ratelle, Frederic Guay, Marylou Harvey, and Evelyne Drouin. 2005. "in Science and Technology :" *Structure*: 171–192.

Pirtle, Zachary, Ryan Meyer, and Andrew Hamilton. 2010. "What Does It Mean When Climate Models Agree? A Case for Assessing Independence Among General Circulation Models." *Environmental Science and Policy* 13 (5): 351–361. doi:10.1016/j.envsci.2010.04.004. http://dx.doi.org/10.1016/j.envsci.2010.04.004.

Regier, Jeffrey C, and Philip B Stark. 2013. "Mini-Minimax Uncertainty Quantification for Emulators" 3: 1–23. http://arxiv.org/abs/1303.3079v1\protect\T1\textbraceleft/%\protect\T1\textbraceright5Cnpapers2://publication/uuid/4BBCA0BD-7BCA-4C83-8508-DFD7B157603D.

Sanderson, Benjamin M., Reto Knutti, and Peter Caldwell. 2015. "Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties." *Journal of Climate* 28 (13): 5150–5170. doi:10.1175/JCLI-D-14-00361.1.