Individual Project Summary I

Yoni Ackerman

Introduction

Anthropogenic climate change is the most dangerous threat facing human societies and global stability (Union of Concerned Scientists). In order to predict spatio-temporal changes in global climate under a variety of scenarios and assumptions, climate researchers are developing computational models to emulate long-term atmosphere and ocean dynamics. There is hope that these models can provide policy and decision makers with predictive tools for climate change preparedness and mitigation (see, for example, UCS). A statistical hurdle, however, impedes this goal: model uncertainty. Much research has gone into understanding and limiting the sources of model uncertainty (Hawkins and Sutton 2009). Despite this progress, there remains a deeper concern regarding model independence and its effect on uncertainty quantification.

Because the models often share both code modules as well as the biases of their implementors, their outputs do not represent independent draws from a space of "all possible future climate trajectories". Thus, agreement in model predictions does not grant greater certainty (Larose et al. 2005). To skirt this issue, Knutti et al describe a method to build consensus models from combinations of model prediction that takes into account their interdependence. By then placing a prior over all possible consensus models, independent samples of "possible future climates" can be drawn and used in analyses (Knutti 2010).

There are still a number of problems with this approach. For one, it doesn't directly address the original central problem: the models are not independent, so model agreement doesn't imply higher certainty. Beyond the lingering problems with prediction, we argue that there are shortcomings with the methods used in this result. In particular, each model is originally expressed - even before EOF (PCA) - in a severely limited form. Only 6 of the original climate variables are used in model comparison, which appears to be due to a computational constraint (at least, it certainly is on our machines). We would like to find a way to Incorporate more variables into Knutti et al's procedure, and we would like to get a better answer to the question: how non-independent are these models?

Data Summary

We are using Coupled Model Intercomparison Project (CMIP5) data made available by ETH Institute for Atmospheric and Climate Science, as well as observational data gathered according to ((Knutti 2010) Table 1). We have access to the observational dataset, as well as data from a total of 46 models, 36 variables, under 10 different scenarios, combined using 47 different ensembles, available in daily/monthly/annual aggregations, and in two spatial coordinate grids (one provided by the model source, and another interpolated to a 2.5 by 2.5 degree grid).

To get an idea of the data in their most basic, time-series form, see Figure 1. This plot shows time series data for surface temperature taken from the ACCESS1-3 model, under the historicalGHG scenario, resampled with r1i1p1 ensemble, and grided to 2.5 by 2.5 degrees.

These data display the expected seasonal signature which can be seen in each time series' auto-covariance plots (see figure 2). The degree to which this auto-covariance holds suggest that these data are suited for compression - a characteristic exploited by Knutti et al.

In addition we can look at the entire spatial-map at an instance in time for a single model (see figure 3):

Here again we find strong auto-correlation, but this time of the spatial variety (see figure 4).

We can also look at the euclidean distance of all the timeseries to give us an idea of the amount of redundancy in the data (keeping in mind this is for air temperature at surface only):

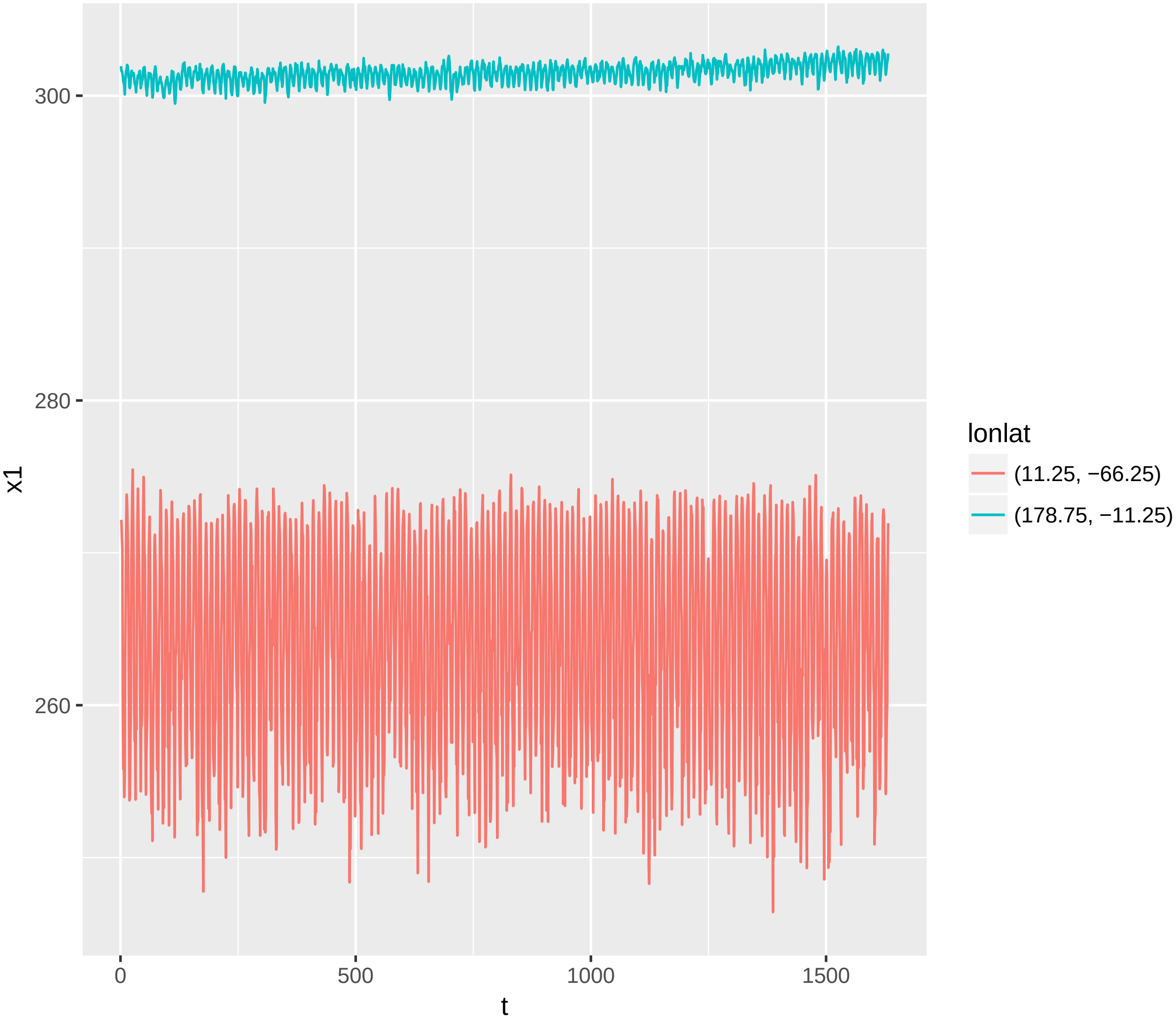

Figure 1: Surface Temperature at (178.75,-11.25) and (11.25, -66.25)

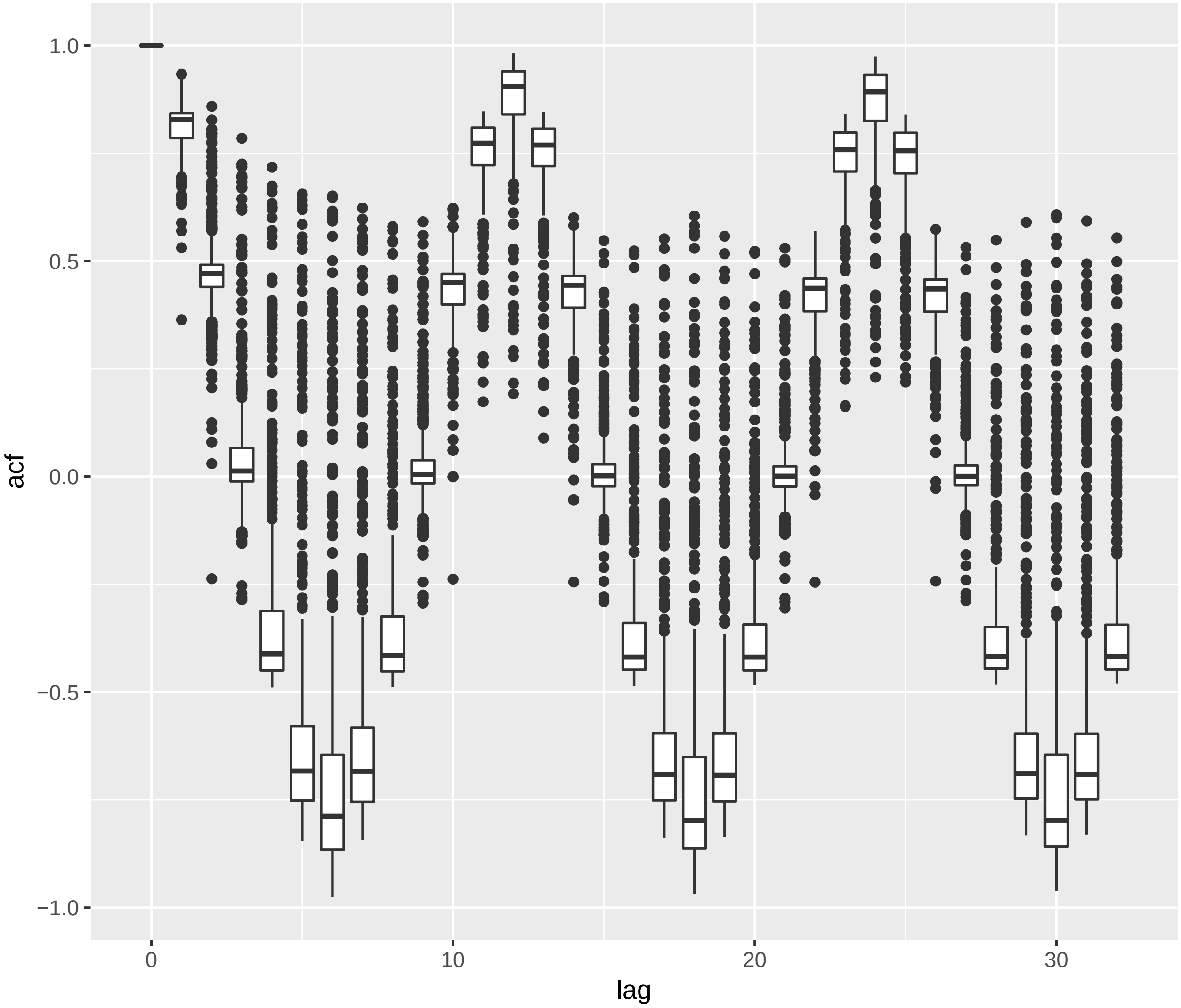

Figure 2: Auto-covariance boxplots over the lag values for the timeseries at a sample of lat/long pairs

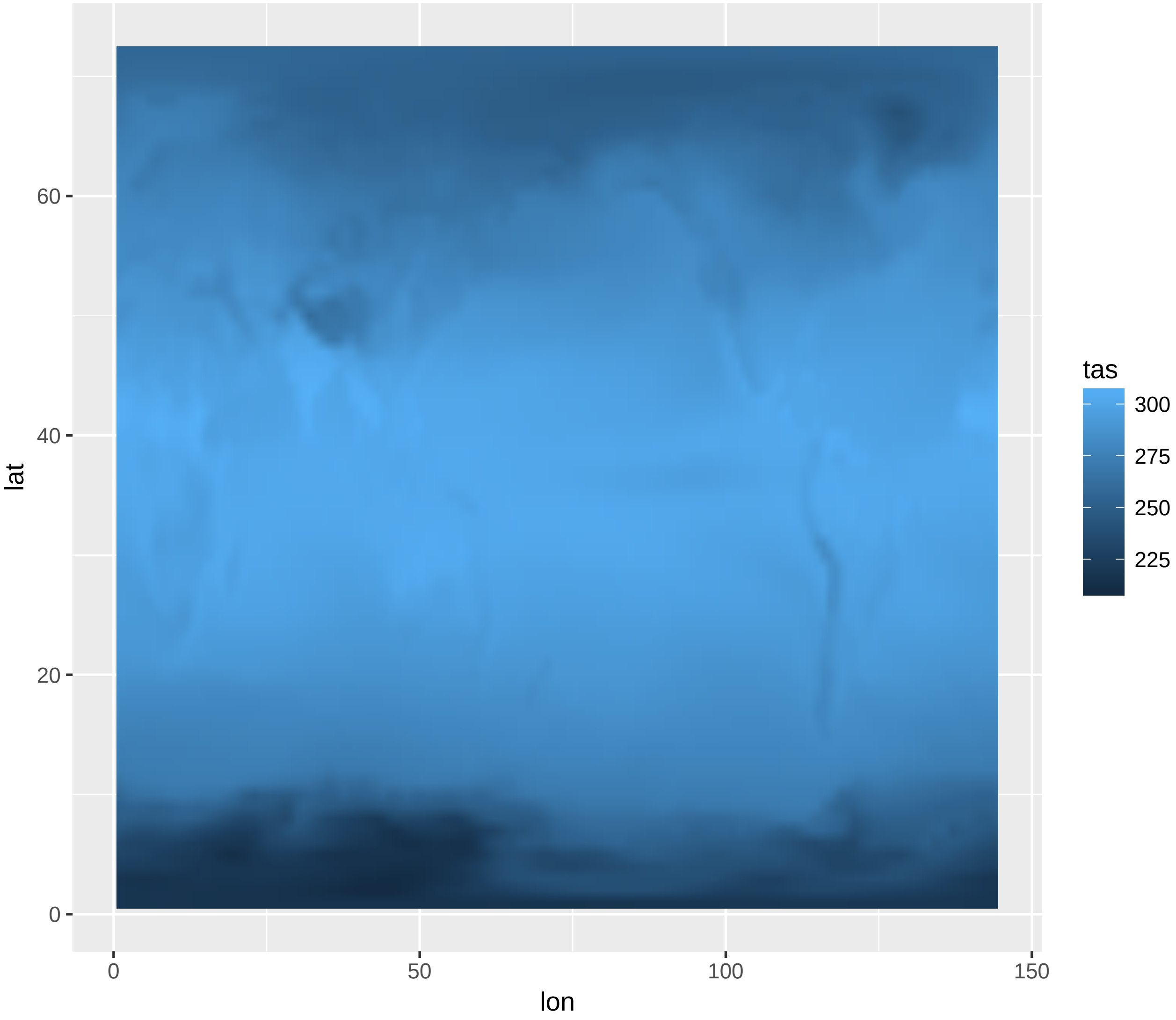

Figure 3: Geo-spatial variation in air surface tempurature (K) 400 months after simulation start

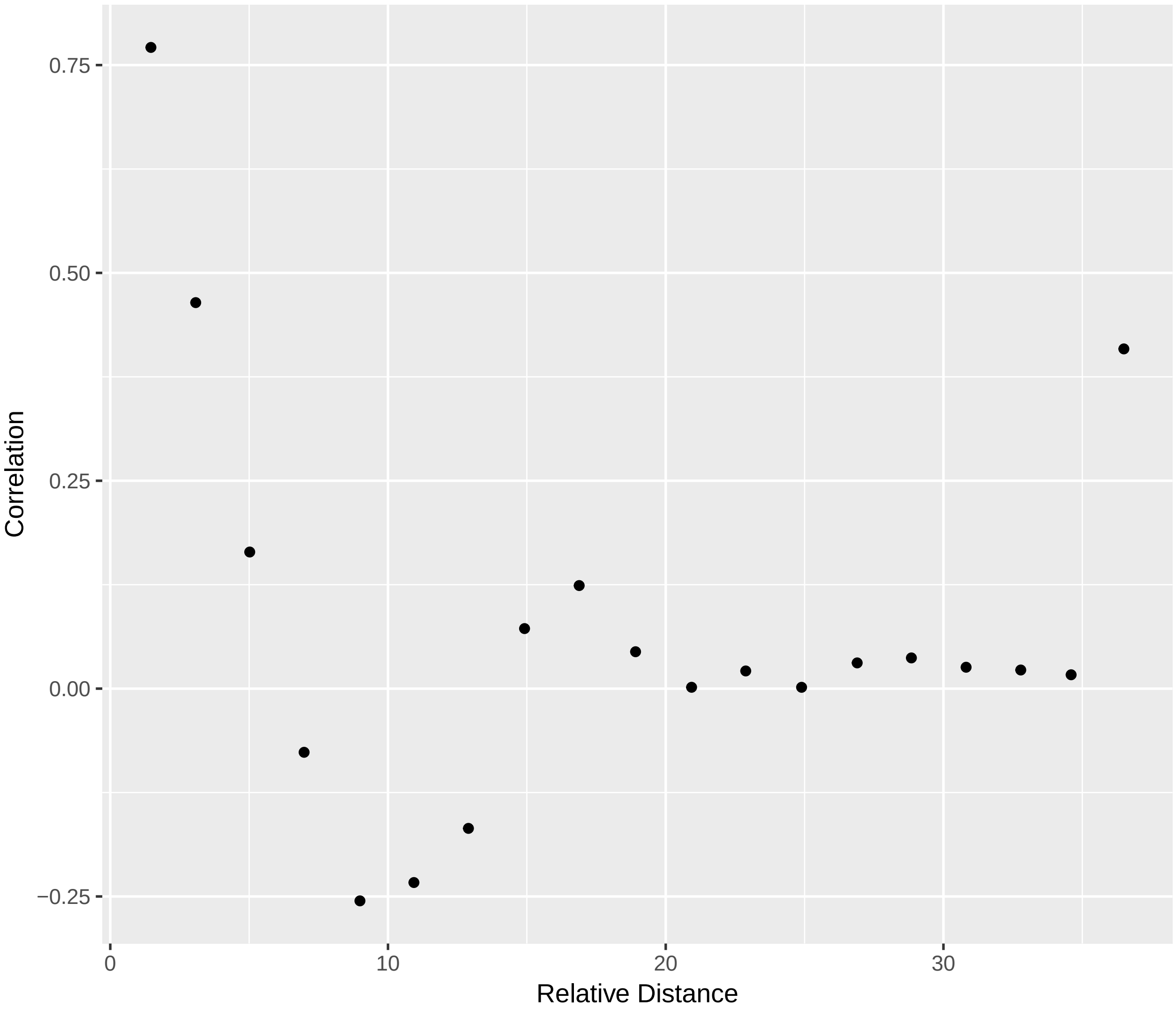

Figure 4: Correlogram for Geo-spatial data 231 months after simulation start

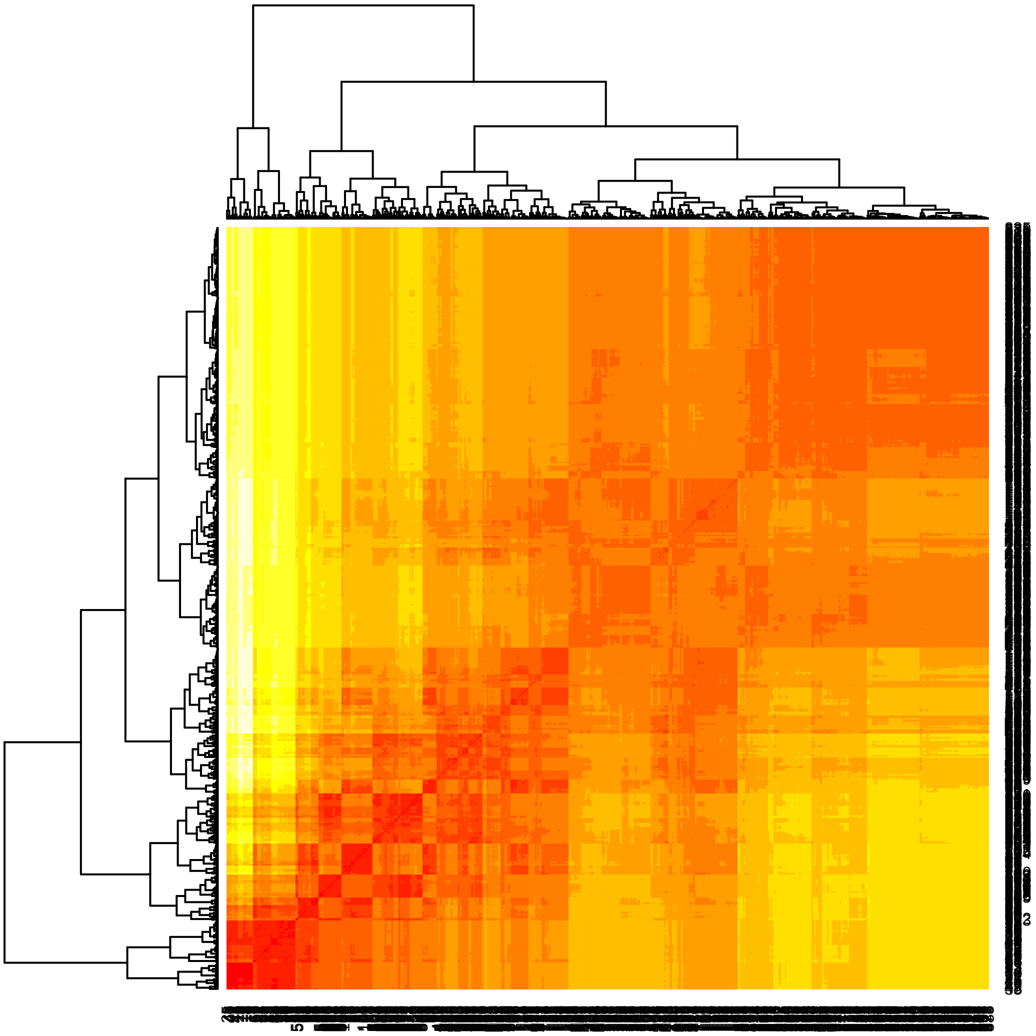

Figure 5: Heat map of spatial similarity of TAS time series

Despite the spatial autocorrelation demostrated here, Knutti et al do not attempt to compress the data in the spatial domain.

Methods

We first attempted to reproduce (roughly) the pca/mds project methods performed in Knutti et al. (Sanderson and Knutti 2012) which involve for each model: normalizing each variable; flattening all data from each variable into a single vector; translating by either the mean of all rows or by the observation row; compressing with PCA; and finally performing an MDS on the compressed model matrix. Because of the quantity of the data, this set of methods limits the number of variables it is feasible to include in the analysis. Our methods attempt to address this.

We attempted to incorporate our knowledge of spatial and temporal auto-correlation. We first flattening the data for each individual variable, and then perform either one (1) or two (2) PCA's:

- (1) we perform PCA dimension reduction to N principle components on the temporal dimension and stretch the results into a single vector representing the compressed data for that model variable.
- (2) we perform a PCA on the spatial dimension, selecting out the first N principle components, and then perform a PCA on the temporal dimension, represented by the transpose of the N principle component. We then select K of these principle components and strech them into a single vector representing the compressed data for that model variable.

We then concatenate the vectors for all model variables involved to form a row representing the entire model in compressed form. With this row of data, we then assemble a model matrix and performing an MDS on the model distances (we don't shift by the observation row, as of now).

Our algorithm for method (2) (omit A4 for method (1)):

- A) for mod^j in J models and for var_i in I variables:
    - 1) $M_{var_i} = mod^j_{.,.,.,var_i}$
    - 2) $T = (flatten(M_{var_i}))^T$
    - 3) Let P such that $T^T T = P^T \Lambda_{spatial} P$, then: $\tilde{M} = ((TP)_{.,1:N})^T$
    - 4) Let Q such that $\tilde{M}^T \tilde{M} = Q^T \Lambda_{time} Q$, then: $\hat{M} = (\tilde{M}Q)_{.,1:K}$
    - 5) $R_{var_i} = [\hat{M}_{1,.}, ..., \hat{M}_{N,.}]$

  - Let $mod^j_{row} = [R_{var_1}, ..., R_{var_I}]$
- B) Let $\mathcal{M} = [(mod^1_{row})^T, ..., (mod^J_{row})^T]^T$
- C) Perform Classical MDS:
  - Let $\Delta = [d^2_{uv}]$ where $d_u v = \sqrt{||mod^u_{row} - mod^v_{row}||}$
  - Let $B = -\frac{1}{2} J \Delta J$ where $J = I - \frac{1}{J}\mathbf{1}\mathbf{1}^T$
  - Let Q^B such that $B = Q^{BT} \Lambda_{dist} Q^B$, then: $\hat{\mathcal{M}} = Q^B_{.,1:2} diag\{\lambda_1, \lambda_2\}^{\frac{1}{2}}$, where λ_1 and λ_2 are the two largest eigenvalues of B and $Q^B_{.,1:2}$ are the corresponding eigenvectors.

for method (1) we chose $N \in \{2, 4\}$ and for method (2) we chose $N = K \in \{5, 10, 20\}$.

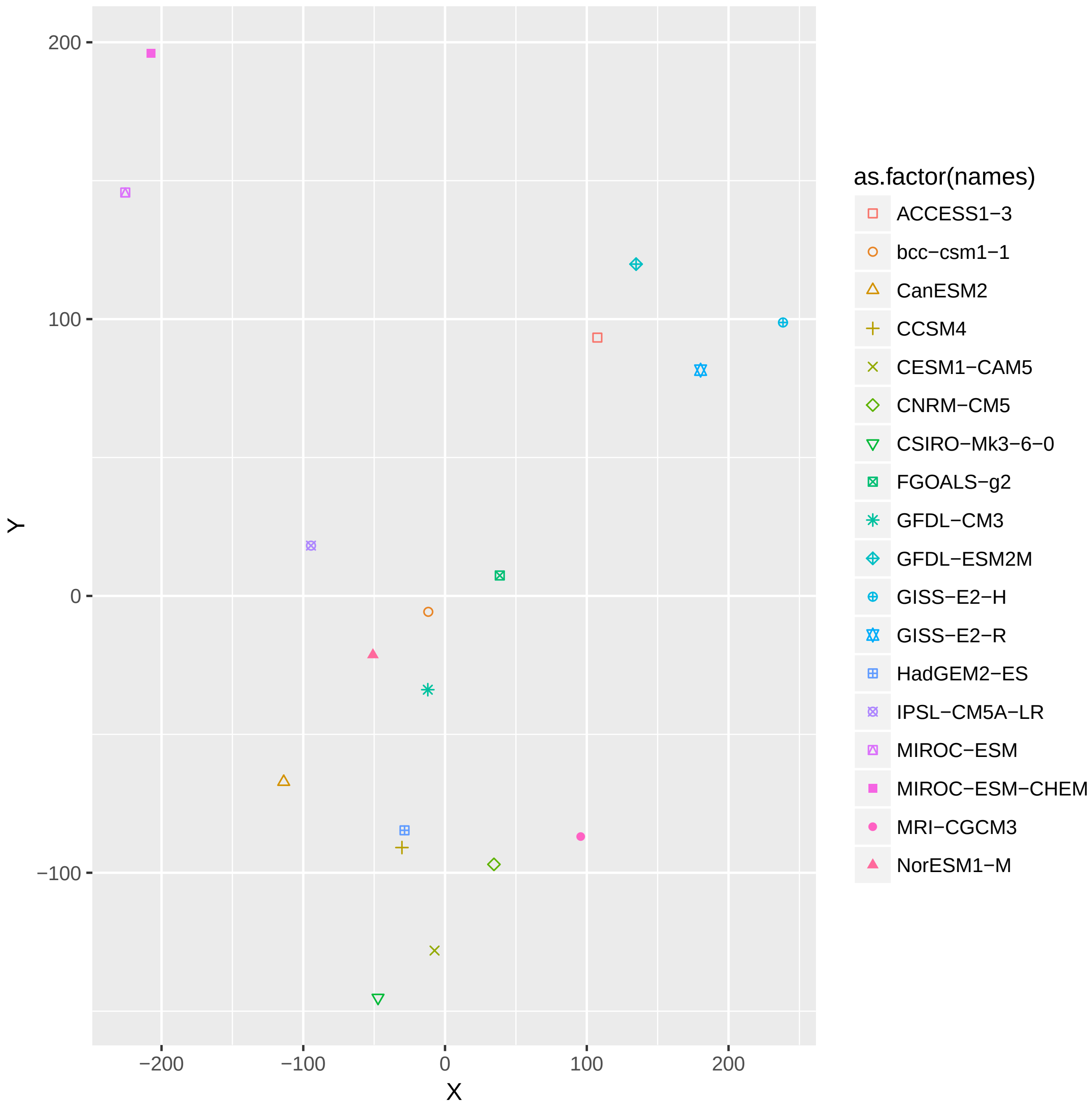

Figure 6: Method (1) with $N = 2$

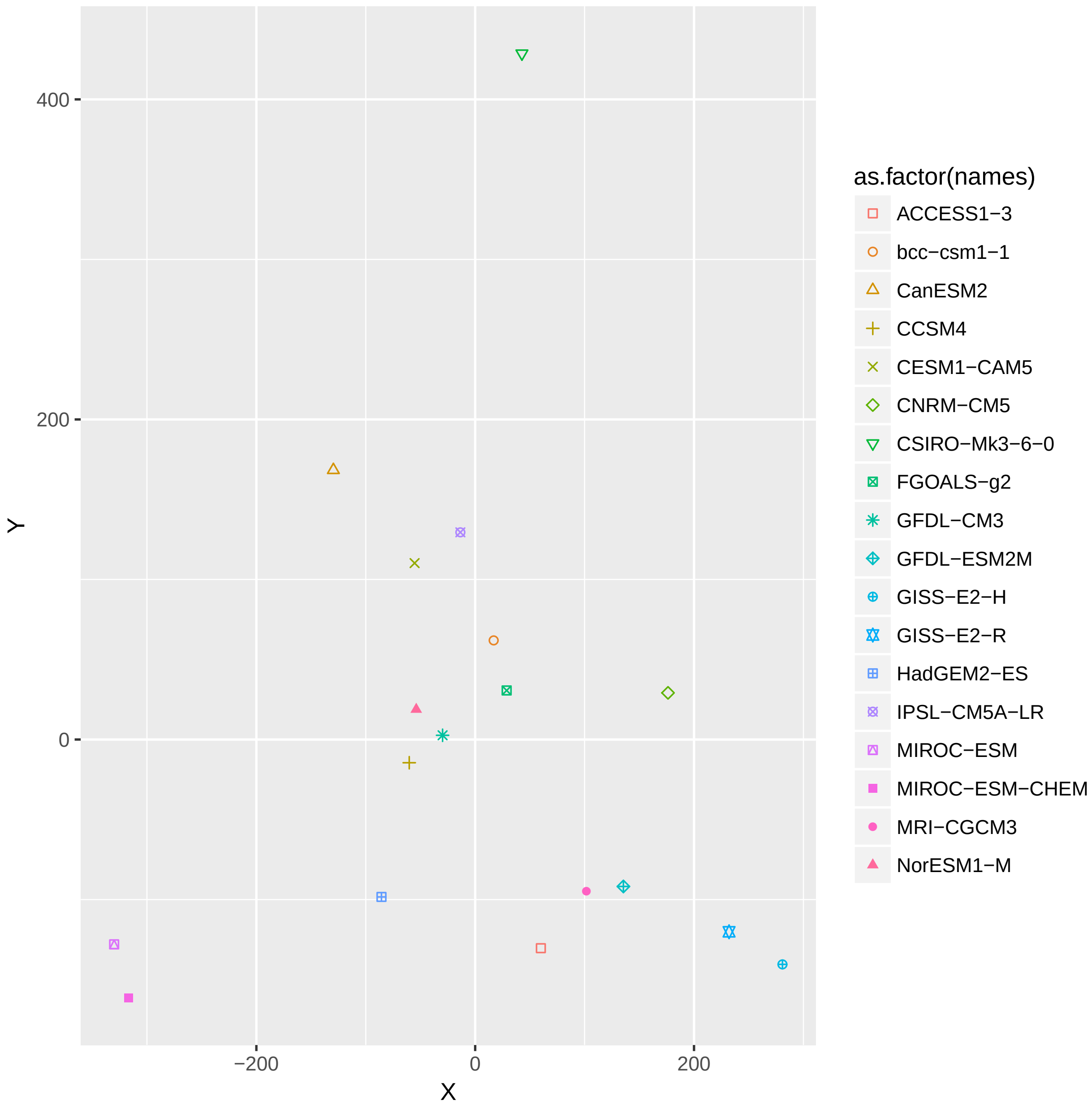

Figure 7: Method (1) with $N = 5$

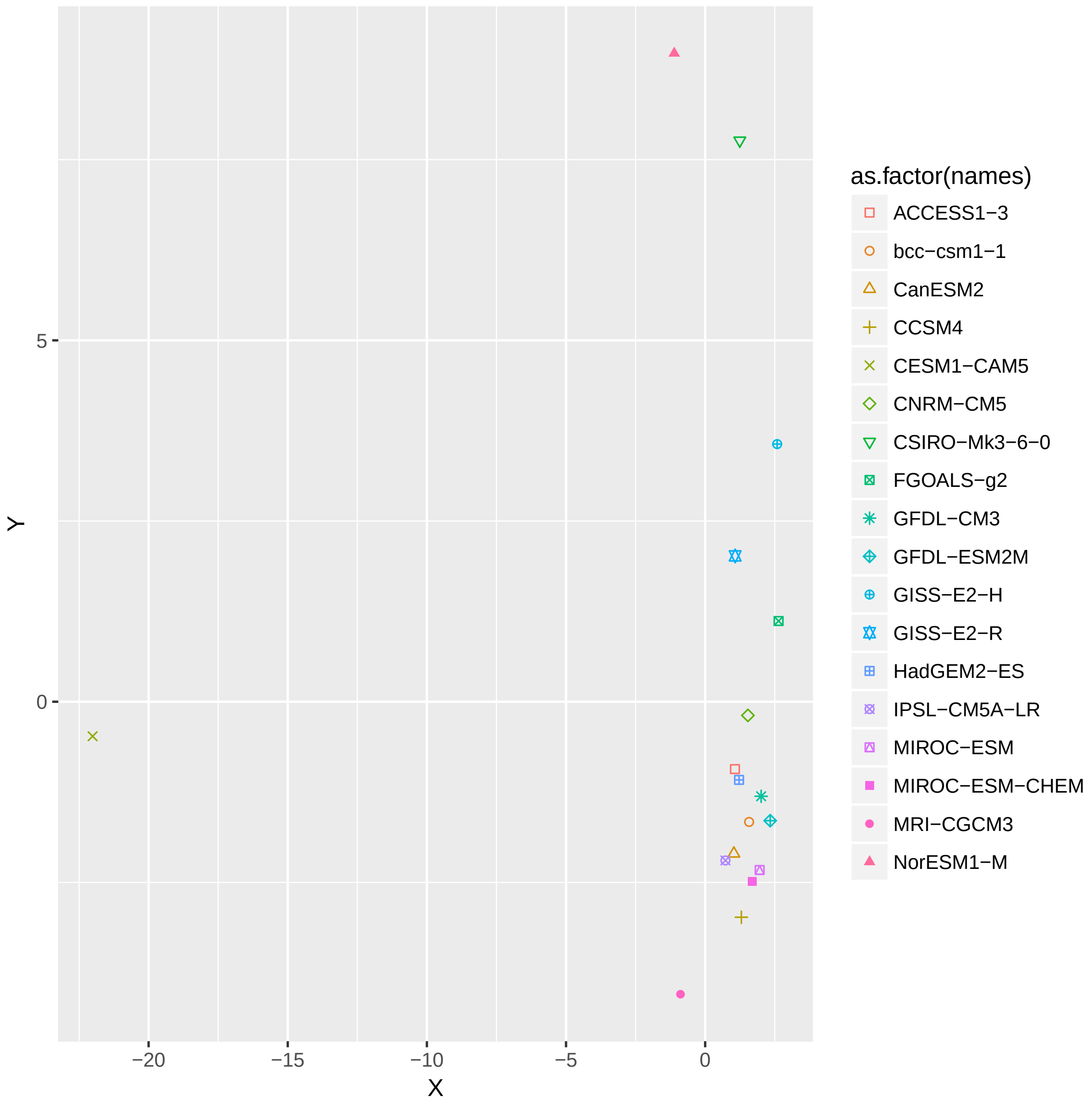

Figure 8: Method (2) with $N = M = 5$

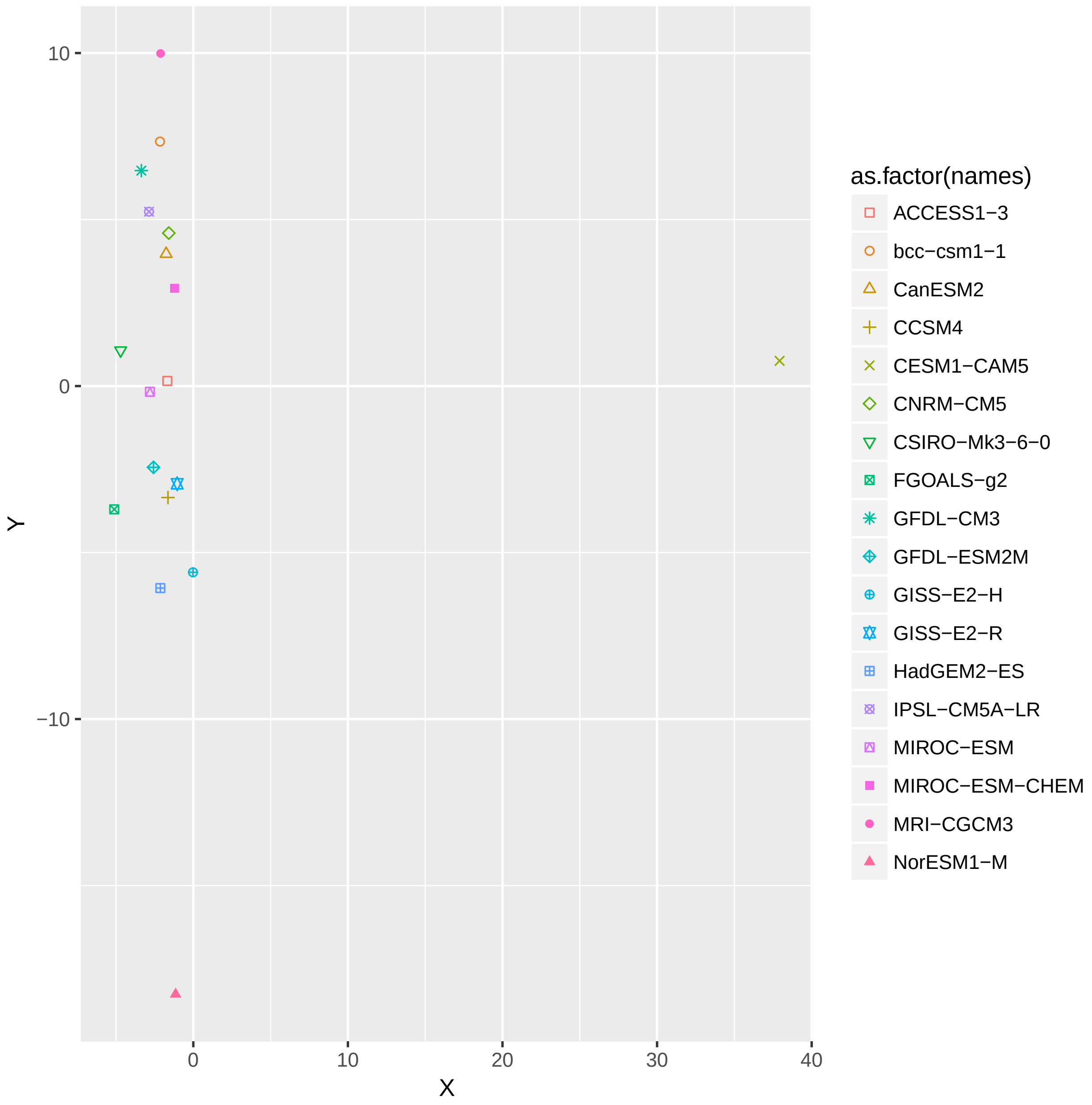

Figure 9: Method (2) with $N = M = 10$

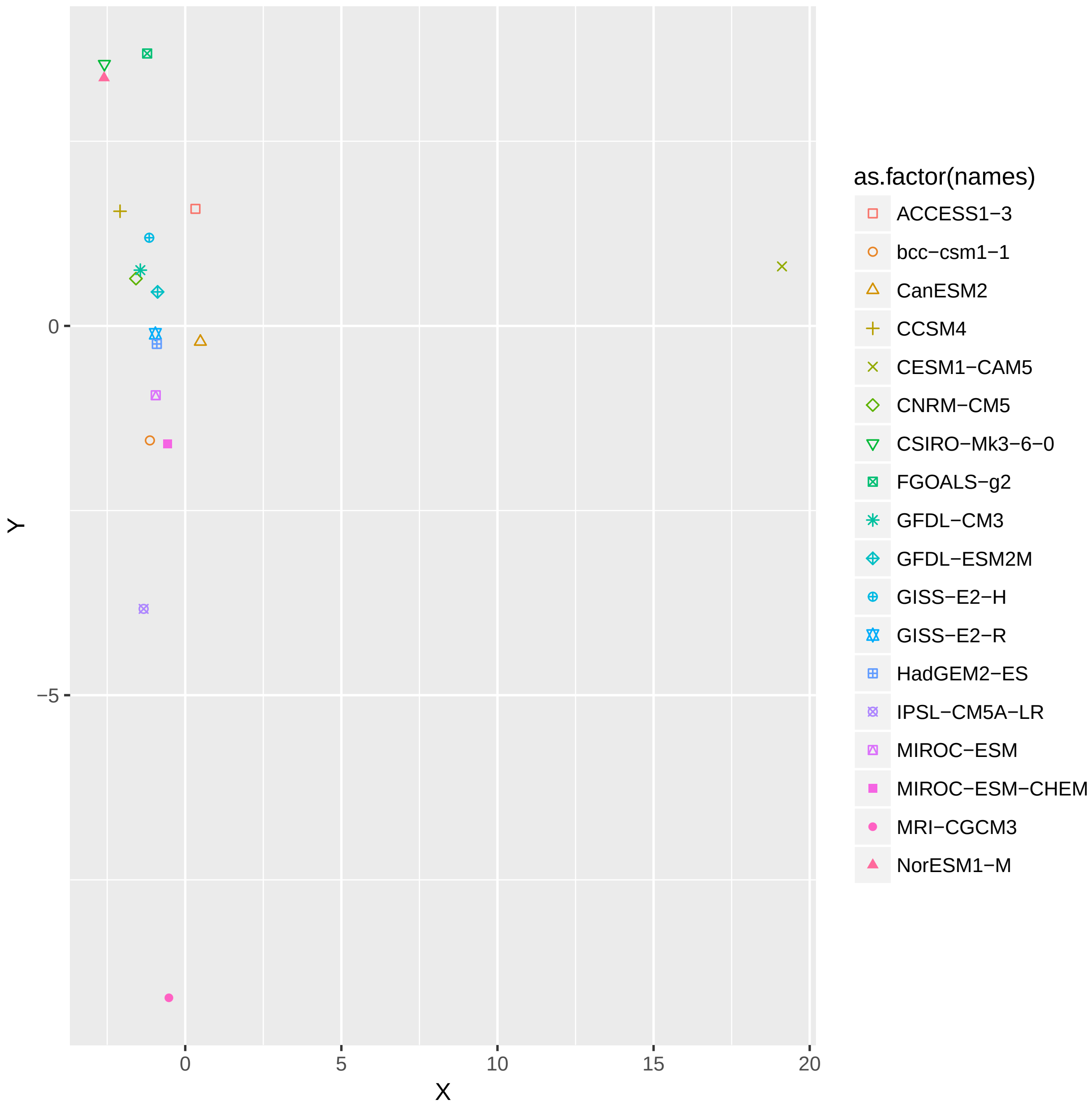

Figure 10: Method (2) with $N = M = 20$

Results

The following plots showcase our results.

There are odd differences between the two methods, though limited differences within methods. Because of the multi-step compression processes, the sources of the between-method differences are hard to trace. Method 1 (Figures 6 and 7) yields results similar to Knutti et. al: we see a cloud of points with some clusters with no egregious outliers. Method 2 (Figures 8, 9, and 10) sees the majority of the points settling along a single axis, however in this case we have less clustering and the presence of outliers.

Conclusions

While it is interesting to see how the degree of compression changes the outputs, the conclusions we can draw are limited. Knutti et. al. include a select group of variables in their analysis and proceed with a heuristic governing how to choose the number of principle components to include in dimension reduction. Roughly, we've shown there needs to be some further justification for such decisions.

The trouble is that nearly all of the work we have done has been exploratory. And with the amount of data we have, and the complexity of the models for which we have output, we could continue with EDA for quite some time. Figuring out what to do is our biggest challenge…

There are multiple directions we could go. The first is to proceed with our initial goals: improve our methods by optimizing N and M and then incorporate all model variables as planned. Should this task prove too trivial, another more technically challenging option exists: in a sense, the most natural way to think about these models is not as matrices in $\mathbb{R}^{n\times n}$ but as multidimenisonal arrays in $\mathbb{R}^{L1\times L2\times T\times V}$ [as in citation]. We can then ask what multidemensional array in $\mathbb{R}^{l1\times l2\times t\times v}$, with $l1 < L1$, $l2 < L2$, $t < T$ and $v < V$, is closest to the original. Doing so would compress all dimensions concurrently, capturing correlations between and within dimensions that our flat PCA methods cannot. The caveats to this direction are the mathematics, the implementation, and the lack of guaranteed improvement in the results (given the effort needed for caveats one and two). Nonetheless, it is something to keep in mind.

Our goals for the break are much more reasonable. We plan to do one final EDA sweep: we want to better understand the variables for each model. In particular, we want to know how correlated they are. We have checked spatial and temporal auto-correlation for each variable, but it might actually be much more useful to understand cross correlations between variables.

The reason for this is that we would ultimately like to statistically test for the independence between these models. Our hope being that the degree to which any tests for indepence should fail could act as an estimate for the models' non-independence. Finding a set of principle components that decently well approximate the many variables involved in each model could give us a way to further compress each model into a single timeseries. We would then be able to use known tests (Hong 1996),(Koch 2013) to determine those timeseries' independence.

References

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.

Sanderson, Benjamin M., and Reto Knutti. 2012. “On the Interpretation of Constrained Climate Model Ensembles.” *Geophysical Research Letters* 39 (16): 1–6. doi:10.1029/2012GL052665.