Global Climate Model Tracking using Geospatial Neighborhoods

Scott McQuade* and **Claire Monteleoni**
Department of Computer Science, The George Washington University
801 22nd Street NW, Washington, DC 20052
mcquade@gwmail.gwu.edu, cmontel@gwu.edu

Abstract

A key problem in climate science is how to combine the predictions of the multi-model ensemble of global climate models. Recent work in machine learning (Monteleoni et al. 2011) showed the promise of an algorithm for online learning with experts for this task. We extend the Tracking Climate Models (TCM) approach to (1) take into account climate model predictions at higher spatial resolutions and (2) to model geospatial neighborhood influence between regions. Our algorithm enables neighborhood influence by modifying the transition dynamics of the Hidden Markov Model used by TCM, allowing the performance of spatial neighbors to influence the temporal switching probabilities for the best expert (climate model) at a given location. In experiments on historical data at a variety of spatial resolutions, our algorithm demonstrates improvements over TCM, when tracking global temperature anomalies.

Introduction

Climate models are complex systems of interacting mathematical models designed by meteorologists, geophysicists, and climate scientists, and run as computer simulations, to predict climate. These General Circulation Models (GCMs) simulate processes in the atmosphere and ocean such as cloud formation, rainfall, wind, ocean currents, radiative transfer through the atmosphere etc., and their simulations are used to make climate forecasts. GCMs designed by different laboratories will often have a high level of variance due to differing assumptions and principals that are used to derive the models.

The Intergovernmental Panel on Climate Change, a panel appointed by the United Nations, is a scientific body informed by climate modeling laboratories around the world, that produced a report on climate change for which it was awarded the Nobel Peace Prize in 2007 (shared by Al Gore). Due to the high variance between the predictions from the different GCMs that inform the IPCC, climate scientists are currently interested in methods to combine the predictions of this "multi-model ensemble," as indicated at the IPCC Expert Meeting on Assessing and Combing Multi-Model Climate Projections in 2010.

Previous work provided techniques to combine the predictions of the multi-model ensemble, at various geographic scales, when considering each geospatial region as an independent problem (Reifen and Toumi 2009; Monteleoni et al. 2011). However since climate patterns can vary significantly and concurrently across the globe, this assumption is unrealistic, and could therefore limit the performance of these previous approaches.

Our contributions in this paper are the following:

1. A richer modeling framework in which the GCM predictions are made at higher geospatial resolutions,
2. Modeling of neighborhood influence among the geospatial regions, using a non-homogeneous Hidden Markov Model,
3. Experimental validation of the effectiveness of these extensions.

Related Work

Machine learning and data mining techniques are starting to be applied to a variety of problems in climate science. Our work relates most directly to the Tracking Climate Models (TCM) approach of (Monteleoni et al. 2011). To our knowledge, our work is the first extension of TCM to model geospatial neighborhood influences.

A number of studies have looked at how a multi-model ensemble of climate models can be used to enhance information over and above the information available from just one model. For example, (Reichler and Kim 2008; Reifen and Toumi 2009) show that the average of the models' output gives a better estimate of the real world than any single model.

There has been recent work on developing and applying more sophisticated ensemble methods (Raftery et al. 2005; Greene, Goddard, and Lall 2006; DelSole 2007; Tippett and Barnston 2008; Peña and van den Dool 2008; Casanova and Ahrens 2009; Smith et al. 2009; Sain and Furrer 2010). For example, (Smith et al. 2009) propose univariate and multi-variate Bayesian approaches to combine the predictions over a variety of locations of a multi-model ensemble, in the batch setting. In the case of regional climate

*Scott McQuade is a PhD student at The George Washington University and an employee of The MITRE Corporation.

Algorithm: Neighborhood-augmented Tracking Climate Models (NTCM)

Input:
Set of geographic sub-regions, indexed by $r \in \{1, \cdots, q\}$ that collectively span the globe, with weights w_r.
Set of climate models, $M_{i,r}$, $i \in \{1, \cdots, n\}$ that output predictions $M_{i,r}(t)$ at each time t and each region r.
Set of $\alpha_j \in [0, 1]$, $j \in \{1, \cdots, m\}$: discretization of α parameter.
β, a parameter for regulating the magnitude of the spatial influence
Initialization:
$\forall j, r,\ p_{1,r}(j) \leftarrow \frac{1}{m}$
$\forall i, j, r,\ p_{1,r,j}(i) \leftarrow \frac{1}{n}$
Upon tth observation:
For each $r \in \{1 \dots q\}$:
Set $P(i|k; \alpha_j, \beta, r, t)$ using Equation (1), $\forall i, k \in \{1, \dots, n\}, j \in \{1, \dots, m\}$.
For each $i \in \{1 \dots n\}$:
Loss[i] $\leftarrow (y_{t,r} - M_{i,r}(t))^2$
For each $j \in \{1 \dots m\}$:
LossPerAlpha[j] $\leftarrow -\log \sum_{i=1}^{n} p_{t,j,r}(i)\, e^{-\text{Loss}[i]}$
$p_{t+1,r}(j) \leftarrow p_{t,r}(j) e^{-\text{LossPerAlpha}[j]}$
For each $i \in \{1 \dots n\}$:
$p_{t+1,j,r}(i) \leftarrow \sum_{k=1}^{n} p_{t,j,r}(k)\, e^{-\text{Loss}[k]}\, P(i|k; \alpha_j, \beta, r, t)$
Normalize $P_{t+1,j,r}$
PredictionPerAlpha[j, r] $\leftarrow \sum_{i=1}^{n} p_{t+1,j,r}(i)\, M_{i,r}(t+1)$
Normalize $P_{t+1,r}$
Prediction[r] $\leftarrow \sum_{j=1}^{m} p_{t+1,r}(j)$ PredictionPerAlpha[j, r]
GlobalPrediction $\leftarrow \sum_{r=1}^{q} w_r$ Prediction[r]

Figure 1: NTCM, our extension of TCM (Monteleoni et al. 2011) to encorporate geospatial neighborhood influence.

$$P(i \mid k; \alpha, \beta, r, t) = \begin{cases} (1-\alpha) & \text{if i=k} \\ \frac{1}{Z}\left[(1-\beta) + \beta \frac{1}{|S(r)|} \sum_{s \in S(r)} P_{\texttt{expert}}(i, t, s)\right] & \text{if i}\neq\text{k} \end{cases} \quad (1)$$

$$\text{where } Z = \frac{1}{\alpha} \sum_{\substack{i \in \{1 \cdots n\} \\ \text{s.t. } i \neq k}} \left[(1-\beta) + \beta \frac{1}{|S(r)|} \sum_{s \in S(r)} P_{\texttt{expert}}(i, t, s)\right]$$

models, (Sain and Furrer 2010) proposes ensemble methods involving multivariate Markov random fields.

Algorithms

The Tracking Climate Models (TCM) algorithm was introduced in (Monteleoni et al. 2011) to dynamically combine the temperature predictions of the multi-model ensemble using Learn-α, an algorithm for online learning with experts under non-stationary observations (Monteleoni and Jaakkola 2003). The algorithm is a hierarchical learner, with updates derived as Bayesian updates of a set of generalized Hidden Markov Models (HMMs), in which the identity of the current best climate model is the hidden variable.

In the present work, we extend the TCM algorithm to operate in a setting in which the global climate models output predictions at higher spatial resolution. We design a variant of the algorithm to take into account regional neighborhood influences when performing updates. Our algorithm differs from TCM in two main ways:

1. The Learn-α algorithm is modified to include influence from a geospatial region's neighbors, in updating the weights over experts (the multi-model ensemble of GCMs' predictions in that geospatial region). The HMM transition dynamics are modified from a time-homogenous transition matrix (as in TCM) to a non-homogenous matrix based on the performance of GCMs in neighboring regions. This influence is parameterized by β; when $\beta = 0$ the new algorithm reduces to Learn-α.
2. Our master algorithm runs multiple instances of this modified Learn-α algorithm simultaneously, each on a different geospatial region, and uses their predictions to make a combined global prediction.

Neighborhood-augmented Tracking Climate Models (NTCM)

In (Monteleoni and Jaakkola 2003) it was shown that a family of algorithms for online learning with experts could be derived as Bayesian updates of the appropriately defined

generalized[1] Hidden Markov Model, where the identity of the current best expert is the hidden variable. This family of algorithms generalizes the Fixed-Share(α) algorithm due to (Herbster and Warmuth 1998) to arbitrary transition dynamics among experts. The Learn-α algorithm (Monteleoni and Jaakkola 2003) hierarchically tracks (as "meta-experts") a set of Fixed-Share(α) algorithms, each with a different value of the parameter α, the modeled temporal switching rate among best experts.

Using the interpretation in (Monteleoni and Jaakkola 2003), our extension to TCM can be viewed as a modification of the transition matrix among experts, of the HMM from which the meta-experts (Fixed-Share(α) algorithms) are derived. This modification uses a non-homogeneous transition matrix, allowing a geospatial region to be influenced by its neighbors.

Figure 1 shows our algorithm, Neighborhood-augmented Tracking Climate Models (NTCM). The algorithm takes a set of q geospatial regions indexed by r, as input, along with weights w_r. These user-specified weights could capture information such as the number of data points in each region or the area of each region. As in TCM, the n global climate models are indexed by i; here the ith model's prediction at time t, $M_{i,r}(t)$ is additionally indexed by r, the geospatial region. Similarly, j indexes Learn-α's meta-experts, Fixed-share(α_j) algorithms each running with a different value of the parameter α, computed via the discretization procedure in (Monteleoni and Jaakkola 2003).

For each region r, the non-homogeneous transition matrix among experts is defined by Equation (1). $S(r)$ is the set of all geographical regions that are spatial neighbors for the region r. This set is determined by the neighborhood scheme, which could be defined using a variety of shapes and sizes. In this work, we define the neighborhood as the four adjacent regions (as described in more detail in the Experiments section), however one could also consider other options such as more circularly-defined neighborhoods, or a scheme that is complex enough to model teleconnections. β is a parameter regulating the magnitude of the spatial influence, $P_{\texttt{expert}}(i,t,s)$ (conditioned over all α values) is the current probability of expert (climate model) i, as determined by the modified Learn-α algorithm for spatial neighbor s, and Z is a normalization factor that ensures that each row of the transition matrix sums to 1 (i.e. the off-diagonal terms of each row sum to α).

The β parameter regulates the influence of the spatial neighbors. A β value of 0 models no spatial influence and corresponds to the normal Learn-α algorithm (with the switching probability α being being shared equally by all experts). Note however that our master algorithm still differs from Learn-α in that we run multiple instances of Learn-α, one per geospatial region, r. A β value of 1 would distribute the switching probability α based solely on the expert probabilities of the spatial neighbors. An interesting direction of future work is to extend the algorithm to simultaneously learn β, analogously to the learning of α.

[1]This HMM generalization allows arbitrary dependence among the observations.

Time Complexity

The total running time of the global Learn-α algorithm with m values of α, a set of n models, over t time iterations, is bounded by $O(tmn^2)$, excluding all data pre-processing and anomaly calculations. Subdividing the globe into q regions adds a factor of q to this bound. The addition of the neighborhood augmentation requires $O(sn^2)$ steps each time a HMM transition matrix is updated, where s is the number of neighbors. Since the matrix would need to be updated for each time iteration, for each geographic region and for each α-value, the overall complexity would be increased by a factor of qs to $O(qstmn^2)$. Additionally, if b separate β values are used, the complexity would further increase by a factor of b. So the total running time of NTCM will be bounded by $O(bqstmn^2)$.

Experiments

We ran experiments with our algorithm on historical data, comparing temperature observations and GCM hindcasts (predictions of the GCMs using historical scenarios). Since the GCMs are based on first principles and not data-driven, it is valid to run them predictively on past data. The hindcast GCM predictions were obtained from the International Panel on Climate Changes (IPCC) Phase 3 Coupled Model Intercomparison Project (CMIP3) archive (CMIP3 2007). Data from the Climate of the 20th Century Experiment (20C3M) was used. Multiple institutions have contributed a number of different models and runs to the 20C3M archive. In this experiment, a single model from each institution, and a single run of that model, was arbitrarily selected, as this is standard practice according to (Monteleoni et al. 2011). Our ensemble size was 13.

We obtained historical temperature anomaly observations from across the globe and during the relevant time period from the NASA GISTEMP archive (NASA GISS). See the Temperature Anomalies section for an explanation of temperature anomalies.

Preprocessing

The different climate model data sets had been created with different parameters, including different initialization dates and spatial resolutions. We performed several preprocessing tasks on the datasets to improve uniformity and consistency. All datasets, including both the models and the observed anomaly data, were temporally truncated to the period occurring between the beginning of 1890 and the end of 1999. The data was then averaged over entire years, forming a consistent time series of 110 points across all data sets. The dataset were also spatially resampled to a geographic grid of 5 degree squares across the globe, creating a consistent spatial grid of dimensions 36x72. The data was spatially resampled using an unweighted mean of all original data points that fell within each 5 degree square. Note that these "degree squares" were defined using the WGS 84 coordinate system, and as such they are neither perfect squares (in terms of their footprint on the Earth) nor of a consistent area. However, the same grid of "squares" was used for all data sets.

The observed anomaly data set contains a number of "missing" data points, where the observed temperature anomaly was not available for a variety reasons. To improve the uniformity of the data, any 5 degree geographic square that had missing data (no observed data points within the 5 degree square) for at least one year in the time series was excluded from the experiment. The majority of the excluded cells are close to the Polar Regions, where limited historical temperature observations are available, particularly for the early part of the dataset.

Temperature Anomalies

Climate scientists often work with *temperature anomalies* as opposed to raw temperatures. A temperature anomaly is the change in temperature at a particular location from the (average) temperature at that same location during a particular benchmark time period. Temperature anomalies are used by climate scientists because they tend to have lower variance when averaged across multiple locations than raw temperatures (as discussed further in (Monteleoni et al. 2011)).

The observed data from NASA GISTEMP had already been converted to temperature anomalies. The benchmark data used to create these anomalies was not provided, however the benchmark period was noted as 1951-1980. The models provided data in absolute temperatures. To convert the absolute data to anomalies, separate benchmarks were calculated for each model by averaging the model data over the 1951-1980 time period.

Another motivating factor for using anomalies based on benchmarks from the respective data set was that the GCMs had been initialized at different times with different parameters. This caused a sizable and consistent offset between different models, around 4 degrees Celsius for some pairs of models. Using anomaly values based on a consistent benchmark period mitigates this constant offset.

Figure 2 shows the observed global temperature anomalies, as well as the anomalies for input data (GCM predictions), over the 1890-2000 time period. Both the observed data and the input data had been spatially averaged over all of the valid 5 degree cells (the included regions).

Results and Analysis

We ran several experiments with the goal of answering the following questions:

- How does the β parameter impact performance? Will the algorithm produce better results as the β parameter increases (i.e. as more geospatial influence is applied)?
- How does the size of the regions impact performance? Is there an optimal region size?

For most of the experiments, we use the set of regional predictions from NTCM to calculate a global prediction. As described in Figure 1, the NTCM global prediction is calculated from a weighted average of all the regional predictions. In these experiments we base the regional weights, w_r, on the number of valid 5-degree cells in each region (in order to properly weight the predictions from regions with mostly excluded data, such as the Polar Regions). These global predictions from NTCM can then be directly compared with the

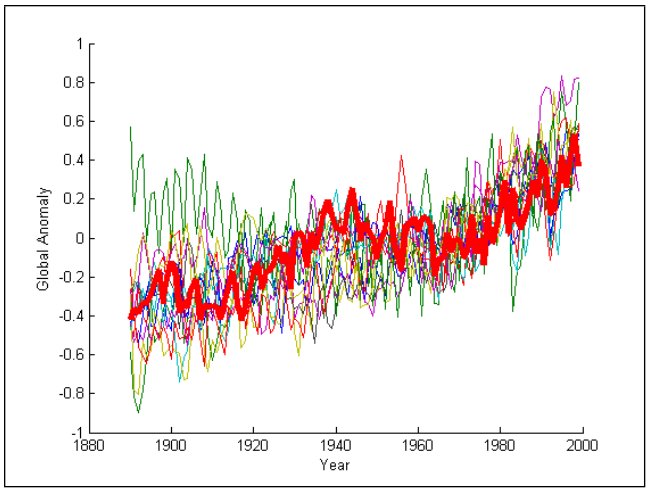

Figure 2: The observed global temperature anomaly (in bold red), and the predicted anomalies from the 13 individual climate models.

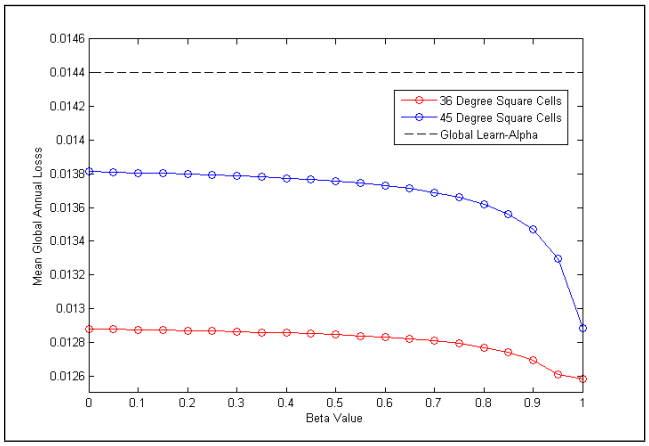

Figure 3: The performance across different β values for 45 degree and 36 degree square cells, compared to the Global Learn-α performance.

global Learn-α method of TCM(Monteleoni et al. 2011) that tracks a single set of models at the global level. We compare the performance using the annual global loss, which is the squared difference between the predicted and observed temperature anomalies.

β Values

Figure 3 shows the mean annual losses over a range of β values for 45 degree and 36 degree square cells. This figure shows a clear trend of decreased loss with increased β values, indicating that increased influence from the spatial neighbors improves the performance. For comparison the Global Learn-α performance is also shown on the graph.

Figure 4 shows how the global anomalies from the 45 degree square regions performed over time versus global Learn-α through a graph of the cumulative annual loss. This graph indicates that for most years, and in particular

	Mean Annual Loss	Variance	Cumulative Annual Loss (1890-2000)
Global Learn-α	0.0144	0.0003	1.5879
45 Degree Squares $\beta = 0$	0.0138	0.0004	1.5194
45 Degree Squares $\beta = 1$	0.0129	0.0003	1.4173

Table 1: Cumulative Annual Losses for 45 degree square cells and Global Learn-α.

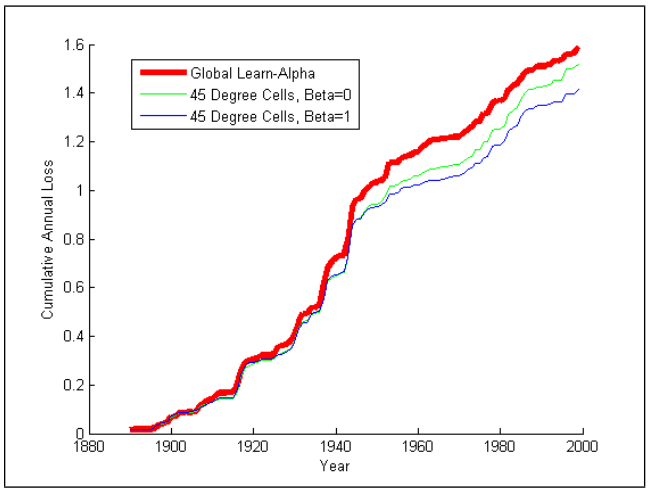

Figure 4: Cumulative annual losses for the 45 degree square cells, compared to the Global Learn-α cumulative loss.

for years later in the time-sequence, the NTCM cumulative global loss was less than that of Learn-α, and with the NTCM results, the $\beta = 1$ loss was less than the $\beta = 0$ loss. The numeric results from this figure are shown in Table 1.

Region Sizes

The experiments described below used a larger set of spatial resolutions to define the regional cells. Again we compare the performance to the TCM algorithm at the global level. The results shown in Table 2.

The data in Table 2 demonstrates two important results at the global level: (1) each of the regional algorithms outperformed all of the traditional global methods, including Learn-α, and (2) for the regional methods the $\beta = 1$ methods outperformed the corresponding $\beta = 0$ method (corresponding to no neighborhood influence) for the same region size.

Figure 5 shows results for NTCM over various region sizes (including a few other region sizes not shown in Table 2). While this graph does not show a clear trend in performance over region size, the best performance for this set of region sizes is achieved with the 36 degree square cells. It may be the case that smaller region sizes enhance the global performance to a point, but as the regions become too small the increased variability within each region begins to diminish the global performance. Another weakness of the smaller regions is that they experience significant distortion near in the Polar Regions as the actual length of a longitudinal degree decreases. We can conclude however that using $\beta = 1$ is superior to $\beta = 0$ for all region sizes tested.

Algorithm	Annual Loss (1890-2000)
Hemispherical Quadrants ($\beta = 1$)	Mean: 0.0131 Variance: 0.0003
Hemispherical Quadrants ($\beta = 0$)	Mean: 0.0141 Variance: 0.0004
45 Degree Square Cells ($\beta = 1$)	Mean: 0.0129 Variance: 0.0003
45 Degree Square Cells ($\beta = 0$)	Mean: 0.0138 Variance: 0.0004
15 Degree Square Cells ($\beta = 1$)	Mean: 0.0137 Variance: 0.0004
15 Degree Square Cells ($\beta = 0$)	Mean: 0.0139 Variance: 0.0004
5 Degree Square Cells ($\beta = 1$)	Mean: 0.0138 Variance: 0.0004
5 Degree Square Cells ($\beta = 0$)	Mean: 0.0140 Variance: 0.0004
Global Learn-α	Mean: 0.0144 Variance: 0.0003
Global Mean Prediction	Mean: 0.0162 Variance: 0.0005
Global Median Prediction	Mean: 0.0164 Variance: 0.0004
Best Global Model	Mean: 0.0159 Variance: 0.0005
Worst Global Model	Mean: 0.1172 Variance: 0.0274

Table 2: Mean Annual Losses from Global Anomalies.

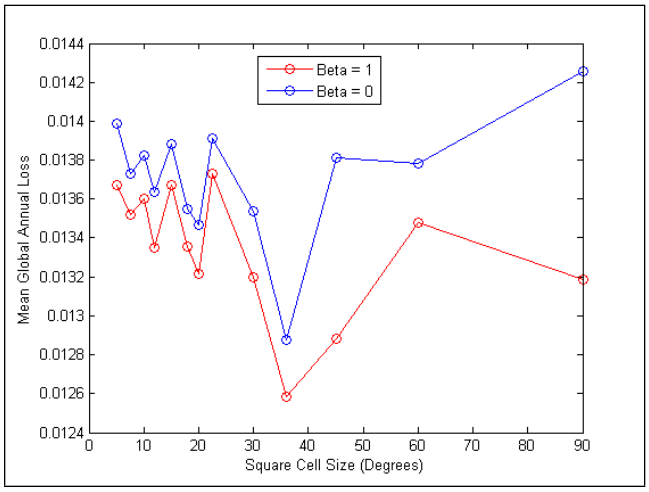

Figure 5: Mean annual global losses for various region sizes.

Region	($\beta = 0$)	($\beta = 0.5$)	($\beta = 1$)
NW Hemisphere	μ: 0.0609 σ^2: 0.0049	μ: 0.0608 σ^2: 0.0049	μ: 0.0579 σ^2: 0.0047
NE Hemisphere	μ: 0.0517 σ^2: 0.0048	μ: 0.0516 σ^2: 0.0048	μ: 0.0511 σ^2: 0.0046
SE Hemisphere	μ: 0.0074 σ^2: 0.0001	μ: 0.0073 σ^2: 0.0001	μ: 0.0075 σ^2: 0.0001
SW Hemisphere	μ: 0.0221 σ^2e: 0.0008	μ: 0.0221 σ^2: 0.0008	μ: 0.0215 σ^2: 0.0007

Table 3: Mean Annual Losses for the Hemispherical Experiment.

	($\beta = 0$)	($\beta = 0.5$)	($\beta = 1$)
Mean over all regions	μ: 0.1359 σ^2:0.1328	μ: 0.1357 σ^2: 0.1327	μ: 0.1347 σ^2:0.1357

Table 4: Mean Annual Losses for the 45-Degree Cell Experiment.

Effect of β on regional results

In addition to examining the effect of β on the global losses (as in the previous experiments), we also ran experiments to determine the effect of β on regional losses (while still incorporating the β-regulated influence between the regions).

Hemispherical Regions

In this experiment we divided the globe into its 4 hemispherical quadrants (NW, NE, SE, SW). The mean annual losses for each of the hemisphere quadrants for each β value are displayed in Table 3.

In three of the four quadrants the β=1 run outperforms the other β values. The exception was the SE hemisphere, where a particularly low mean annual loss and variance was achieved for all β values.

45-Degree Cells

We also ran regional experiments at at higher spatial granularity (45-degree cells). The means of the annual losses over all 32 regions are shown in Table 4. Note that these are regional losses averaged over the globe, rather than global losses from regional predictions averaged over the globe as in the 'Region Sizes' experiment.

The results are of limited statistical significance due to the high variance. The high variance was caused, in part, by the cells around the Polar Regions, where the footprint of each region is decreased, and additionally, significant portions of these regions are excluded due to lack of observed anomalies. Table 5 shows the mean losses when these Polar regions are excluded (so the remaining 16 regions are included) from the calculation, but still contribute to the algorithm's neighborhood augmentation.

When these Polar regions are excluded there is a much smaller variance, and some indication that the $\beta = 1$ case is outperforming the other β values.

	($\beta = 0$)	($\beta = 0.5$)	($\beta = 1$)
Mean over non-polar regions	μ: 0.0380 σ^2:0.0033	μ: 0.0379 σ^2: 0.0033	μ: 0.0368 σ^2:0.0031

Table 5: Mean Annual Losses for the 45-Degree Cell Experiment, Excluding Polar Regions.

Conclusions and Future Work

The NTCM algorithm showed promising signs when incorporating neighborhood influence. The results showed that runs using the full neighborhood influence ($\beta = 1$) outperformed runs without neighborhood influence ($\beta = 0$) for predicting global temperature anomalies. Also at the global level, the NTCM algorithm outperformed previous global algorithms such as the Learn-α method of TCM.

Several aspects of the experimental setup were primitive, and it's possible that improved results could be obtained with modifications in future work:

- The grid system: A very basic system based on WGS 84 coordinates was used in these experiments. In this system the cells become distorted towards the Polar Regions. Improvements could be made to this system so that it more accurately reflects the Earth's surface as well as the distribution of observed temperatures.
- The shape and size of the neighborhoods: Only the immediately adjacent four cells were used as neighbors in these experiments. Future work could examine the effect of using more advanced neighborhood schemes. This could include a weighted scheme that considers a much larger neighborhood, where cells in closer proximity are given more weight, for example, by using a Gaussian weighting function over distance.
- Algorithmic extensions include allowing the weights over regions, w_r, or the spatial influence parameter β, to vary over time, or to be learned online.

Acknowledgments

We thank the anonymous reviewers for their comments used in revision. We acknowledge the modeling groups, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) and the WCRP's Working Group on Coupled Modelling (WGCM) for their roles in making available the WCRP CMIP3 multi-model dataset. Support of this dataset is provided by the Office of Science, U.S. Department of Energy.

References

Casanova, S., and Ahrens, B. 2009. On the weighting of multimodel ensembles in seasonal and short-range weather forecasting. *Mon. Wea. Rev.* 137:3811–3822.

CMIP3. 2007. The World Climate Research Programme's (WCRP's) Coupled Model Intercomparison Project phase 3 (CMIP3) multi-model dataset. `http://www-pcmdi.llnl.gov/ipcc/about_ipcc.php`.

DelSole, T. 2007. A bayesian framework for multimodel regression. *J. Climate* 20:2810–2826.

Greene, A.; Goddard, L.; and Lall, U. 2006. Probabilistic multimodel regional temperature change projections. *J. Climate* 19(17):4326–4343.

Herbster, M., and Warmuth, M. K. 1998. Tracking the best expert. *Machine Learning* 32:151–178.

Monteleoni, C., and Jaakkola, T. 2003. Online learning of non-stationary sequences. In *NIPS '03: Advances in Neural Information Processing Systems 16*.

Monteleoni, C.; Schmidt, G.; Saroha, S.; and Asplund, E. 2011. Tracking climate models. *Statistical Analysis and Data Mining: Special Issue on Best of CIDU* 4(4):72–392.

NASA GISS. GISTEMP. `http://data.giss.nasa.gov/gistemp/`.

Peña, M., and van den Dool, H. 2008. Consolidation of multimodel forecasts by ridge regression: Application to pacific sea surface temperature. *J. Climate* 21:6521–6538.

Raftery, A. E.; Gneiting, T.; Balabdaoui, F.; and Polakowski, M. 2005. Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.* 133:1155–1174.

Reichler, T., and Kim, J. 2008. How well do coupled models simulate today's climate? *Bull. Amer. Meteor. Soc.* 89:303–311.

Reifen, C., and Toumi, R. 2009. Climate projections: Past performance no guarantee of future skill? *Geophys. Res. Lett.* 36.

Sain, S., and Furrer, R. 2010. Combining climate model output via model correlations. *Stochastic Environmental Research and Risk Assessment*.

Smith, R. L.; Tebaldi, C.; Nychka, D.; and Mearns, L. O. 2009. Bayesian modeling of uncertainty in ensembles of climate models. *Journal of the American Statistical Association* 104(485):97–116.

Tippett, M. K., and Barnston, A. G. 2008. Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.* 136:3933–3946.