Challenges and opportunities toward improved data-guided handling of global climate model ensembles for regional climate change assessments

Evan Kodra, University of Tennessee at Knoxville and Oak Ridge National Laboratory
Snigdhansu Chatterjee, University of Minnesota at Twin Cities
Auroop R. Ganguly*, Oak Ridge National Laboratory and University of Tennessee at Knoxville

Abstract

Global climate models (GCMs) are important tools for addressing climate change questions at a global scale. Recent research has attempted to combine outputs from multiple GCMs to quantify uncertainty in regional climate change, which may ultimately inform regional stakeholders and policy-makers. Using a case study, we illustrate a potential path toward improvement in an existing Bayesian formulation, which may ultimately result in more physically meaningful GCM combination and hence add value to decision-making. This area of research is still nascent and may benefit from innovations in the computational data sciences.

1. Introduction

Regional climate change is considered a fundamental gap in climate science (Schiermeier, 2010). Although statistical or dynamical downscaling is the state-of-the-art in regional assessment, both approaches have drawbacks, notably including the fact that they may not resolve differences between multiple GCMs (Lettenmaier, in US EPA 2009). Thus, it may be of interest to explore regional climate change directly from multiple GCMs.

Multi-GCM approaches are now being used more often in attempts to capture a more complete scope of possible future regional climate. Equally-weighted Multimodel Ensemble (MME) averaging has been oft-considered a best practice in global and regional climate reports (IPCC, 2007; Karl et al., 2009) and studies (Pierce et al., 2009; Santer et al., 2009). In addition, researchers in several disciplines including data mining and machine learning (Seni and Elder, 2010) have suggested that model combinations often lead to better prediction across different applications. However, a consideration of recent work by Knutti, (2010), Knutti et al., (2010), and Perkins et al., (2009) may suggest that, in many cases, GCM inclusion/weighting within model ensembles should be based on their ability to simulate key physical processes.

Another recent line of work attempts to find optimal weights of GCMs, although there may be pitfalls in doing so (Weigel et al., 2010). The latest (current generation) approach of this type can be found in Smith et al., (2009), an aspect of which will be the focus of this work. In addition, Monteleoni et al. (2010) developed a tracking approach which updates GCM weights continuously; this appears to outperform equally-weighted MMEs and handle non-stationarity with short lead times well. However, it is unclear whether this approach in its current form can handle long lead times. The simultaneous handling of long lead times and non-stationarity could be a significant advance in the climate community; thus, the line of research deserves further attention.

GCM combinations may be useful for informing concise decision-making tools and in turn motivates the present study of Bayesian GCM combination (specifically, the latest generation exemplified in Smith et al., (2009)). We focus on regional climate assessments and pick one direction as an example where improvements may be possible. Specifically, we suggest that process-based weighting may improve the climate-science relevance and hence the ability of the current approach to generalize.

2. Approach

2.1 Bayesian GCM Combination: Latest Generation

2.1.1 Description of the Statistical Model

Bayesian combinations of GCMs attempt to synthesize multiple GCM-forecasts to produce probabilistic projections of future regional climate. The latest generation in this line of work (the *univariate* statistical model in Smith et al. 2009, which we use here) is a formalization of the Reliability Ensemble Average (REA) concept proposed by Giorgi and Mearns, (2002). The REA sought to weight GCMs based on *bias*, or the difference between past observed and modeled temperature, and *convergence*, or the distance of a given GCM from the future average of all GCMs. These two components together formed the weight, λ_i (or reliability) of model *i*.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

* Correspondence to Auroop R Ganguly: auroop@alum.mit.edu

Unknown parameters of interest in the univariate model include μ (current mean temperature), υ (future mean temperature), λ_i (model i weight or precision (inverse variance)), β (correlation between model i hindcast and model j forecast), and θ (a scaling parameter to allow for a difference in past and future GCM weight). All of the above unknowns are updated via a Gibbs sampler. X_i and Y_i are known data and represent past and future simulated spatio-temporal mean temperatures, respectively, from GCM i. X_0 represents past observed mean temperature, and λ_0 is a measure of variability in past observed temperature; in this work, we set λ_0^{-1} to the sample variance of 30 annually- and spatially-averaged (1970-1999) temperature values.

2.1.2 The Role of Bias and Convergence

One key aspect of the univariate model is the expected value of the weight (reliability) for any given model i:

$$E[\lambda_i] = \frac{a_\lambda + 1}{b_\lambda + 0.5[X_i - \mu]^2 + 0.5\theta[Y_i - \upsilon - \beta(X_i - \mu)]^2}. \quad (1)$$

It can be observed from (1) that the expected weight of model i is a function of "approximate" bias, X_i-μ, and "approximate" convergence, Y_i-(υ+β(X-μ)). Because of the conditional dependence present in the Markov Chain Monte Carlo (MCMC) simulation used, both bias and convergence are no longer as straightforward to interpret as from Giorgi and Mearns, (2002), hence the use of "approximate". In particular, each simulated parameter is dependent on other parameters simulated previously; for instance, a simulated μ is dependent on the previous $\sum\lambda_i$ and λ_0, and the next simulated $\sum\lambda_i$ is dependent on the previous μ. This dependence cycle makes it difficult to distinguish the true relative contributions of bias and convergence, and hence the contribution of observational and GCM quantities, in the MCMC simulation.

2.1.3 Testing the Appropriateness of Bias and Convergence Criteria

Past literature (Knutti et al. 2010) has noted that heavy focus on GCM consensus may be unwise, particularly because it is cannot be assured that the center (mean, median) of the spread of GCMs is necessarily "more correct", especially given long lead times and potential non-stationarity in climate.

For these reasons, it may be of interest to at least empirically ascertain whether and when the univariate model weights information from GCMs more heavily than information from observations and how that may affect posterior distributions of regional temperature change. We wish to make progress toward ultimately testing the hypothesis that too much focus on consensus leads to suboptimal and potentially misleading consequences for estimation of uncertainty in regional climate change.

There may also be a risk in focusing too heavily on skill, unrealistically constraining uncertainty of future climate with respect to a few GCMs that exhibit apparent skill in the past (Knutti et al. 2010). However, we hypothesize that the univariate model in its current form may be better served by integrating process-based weighting schemes which are less trivial than the current bias measure; a simple bias measure may not be adequate for assessing model skill anyway and thus may not inform the univariate model well. We begin to test these hypotheses in the next section.

3. Analysis

3.1 Empirical Assessment of Bias versus Convergence

3.1.1 Regions, Data, and Model Initialization

To empirically assess the affect of bias and convergence on conditions of the weights and their implications from the univariate model, we examine case studies from three of the 21 regions defined in Giorgi and Francisco (1999): Greenland (GRL), the Amazon Basin (AMZ), and West Africa (WAF). X_0 and λ_0 are calculated from surrogate observational NCEP/NCAR-1 reanalysis data (http://www.esrl.noaa.gov/psd/data/gridded/data.ncep.reanalysis.html, Kalnay et al. 1996). All (24) X_i's and Y_i's are obtained using 20th century hindcasts and moderate fossil fuel-emission International Panel on Climate Change (IPCC) Special Report on Emissions Scenarios (SRES) A1B forecasts, respectively, from all 24 Coupled Model Intercomparison Project (CMIP3) GCMs available from the Program for Climate Model Diagnosis and Intercomparison (PCMDI, http://www-pcmdi.llnl.gov/). For past GCM and reanalysis climate we obtain area-weighted spatio-temporal averages from the years 1970-1999, and for future GCM projections, we do the same from 2070-2099.

3.1.2 Results

For each region, we obtain X_0, X_A, Y_A, and Y_B where X_A is the average of GCM hindcasts, Y_A is the average of GCM forecasts, Y_B is the bias-corrected MME temperature forecast $Y_B = (1/M)\sum(\boldsymbol{Y}\text{-}[\boldsymbol{X}\text{-}X_0])$ (where $\boldsymbol{Y}$ and $\boldsymbol{X}$ are vectors of future and past GCM projections), and X_0 is as previously defined. The goal is to ascertain from the MCMC simulation the distance of simulated values υ from Y_A and Y_B; this should yield an initial idea of the importance of convergence versus bias as assigned by the univariate model. Thus, we also compute, for each MCMC iteration s in 1 to S=15,000, $B_s = |\upsilon - Y_B|$ and $C_s = |\upsilon - Y_A|$, and compute B = % of time ($B_s < C_s$) and C = % of time ($C_s < B_s$), and thus a rough estimate of whether bias (B) or convergence (C) is treated as more important in the univariate model. Results in the last two columns of Table 1 suggest that, in general, convergence seems to be favored by the univariate model.

Table 1. Bias versus Convergence in the Univariate Model

REGION	X_0	X_A	Y_A	Y_B	B	C
GRL	263.32	265.66	269.96	267.62	40.91%	59.09%
AMZ	297.27	295.98	300.37	299.07	27.34%	72.66%
WAF	298.36	298.24	301.13	301.25	36.84%	63.16%

However, note that Table 1 does not capture properties and possible effects of individual models. Thus, to take into account the potential importance of individual GCMs, Figure 1 plots GCM weights as proportions ($\lambda_i/\sum\lambda_i$), averaged over the 10,000 iterations, versus their bias and convergence ranks, for all three regions. Note that here (but not previously) $\sum\lambda_i$ includes λ_0, where for each region λ_0 is an observed constant but holds different weight.

Subsequently, Figure 1 plots the values of X_i-X_0 and Y_i-Y_A. From Figure 1, we can visualize whether models with smaller biases or models closer to Y_A seem to be favored in the formation of weights (and ultimately the distribution of υ). The relationship between weights and bias and convergence rankings is, in general, fairly strong as calculated by Kendall's τ, a rank correlation measure. Model weights seem to increase with higher ranks in both bias and convergence, but it is not clear from Figure 1 which criterion is favored by the statistical model; this may depend on the input data and hence the region of study. Additionally, there may be some region-dependent degree of redundant information used when applying both criteria in the univariate model, as Kendall's τ measures between the bias and convergence are 0.30 (GRL), 0.51 (WAF) and 0.15 (AMZ).

From Table 1 and Figure 1, it appears that the posterior distribution of temperature change may be more influenced by the convergence criterion, although bias seems to play an important role as well.

3.1.3 ADEQUACY OF THE CURRENT UNIVARIATE MODEL

Figure 2 displays a posterior distribution of temperature change in GRL, with individual GCM estimates plotted on the x-axis. Several GCMs project higher changes in temperature than most – ipslcm4 (France), ukmohadgem (United Kingdom), mirocmed (Japan), mirochi (Japan) – and several project lower changes than most – ncarccsm3, gisseh, gisser, and gissaom (United States).

The univariate model has effectively assigned an approximate zero probability to the eight models mentioned. However, recent literature suggests that some of these same GCMs (mirocmedres, ukmohadgem, ncarccsm3) may capture important aspects of GRL climate well (Walsh et al., 2008; Stoner et al. 2009). If it is assumed that GCMs with past skill will continue to be skillful in the future, which is a debated but oft-assumed notion (Knutti et al. 2010), then GCM weighting may be informed by measures of past skill.

Assuming (despite the debate) that using past skill in this case is justifiable, note that, for example from Figure 1, mirocmed is ranked poorly as per bias and convergence, and hence in Figure 2 is assigned virtually 0 probability. However, Walsh et al. (2008) notes that *after* a bias correction, mirocmed may be the most skillful GCM in simulating GRL temperature. This may suggest that, if we intend to capture skill well in the statistical model and let it guide the posterior distribution of temperature, *bias alone is not sufficient measure for GCM weighting*. We emphasize however, that this is just one example and that, even in this case, this notion should be tested more rigorously.

The appropriateness of convergence may be more difficult to test and is outside of the scope of this work. However, we note that others have pointed out that consensus may serve to unrealistically constrain uncertainty (e.g. Knutti et al. 2010).

4. Discussion and Conclusion

Questions around the appropriateness of observation-based skills to generalize in the future, especially for long lead times, nonlinear processes and non-stationary conditions, may motivate the integration of more rigorous and physically meaningful process-based evaluation of GCMs into statistical models. Convergence seems to be treated with higher importance by the current statistical model, although this appears to vary per region. We posit that creative hypothesis-driven studies can be designed to test the value of model convergence as a metric.

The improvement in the formation and treatment of skill, process-evaluation and convergence is just one example of potential improvement to this particular univariate model. Further development in statistical (Smith et al. 2009) or machine learning (Monteleoni et al. 2010) approaches may offer pathways toward the eventual integration of process-based GCM weighting, as well as enhanced handling of non-stationarity and long lead times, which would be a significant advance in climate science. These areas deserve attention from the climate, statistical, and machine learning communities.

Acknowledgements

This work was funded by the NSF Expeditions in Computing grant "Understanding Climate Change: A Data Driven Approach", award number 1029166.

References

Giorgi, F., and Mearns, L.O. Calculation of Average, Uncertainty Range, and Reliability of Regional Climate Changes from AOGCM Simulations via the "Reliability

Ensemble Averaging" (REA) Method. *J. Clim*. 15: 1141-1158, 2002.

IPCC. US Climate Change Science Program, AR 4, 2007.

Hagedorn, R.F., Doblas-Reyes, F.J., and Palmer T.N. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *TellusA* 57: 219-233, 2005.

Karl, T. R., Melillo, J. M., and Peterson, T. C., eds. *Global climate change impacts in the United States*. Cambridge University Press, 196 pages, 2009.

Knutti, R. The end of model democracy? *Clim Change* 102: 395-404, 2010.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.A. Challenges in combining projections from multiple climate models. *J. Clim*. 23(10): 2739-2758, 2010.

Monteleoni, C., Schmidt, G., and Saroha, S. Tracking Climate Models. *Conference on Intelligent Data Understanding*, 2010.

Perkins, S.E., Pitman, A.J., and Sisson, S.A. Smaller projected increases in 20-year temperature returns over Australia in skill-selected climate models. *Geophys. Res. Lett.*, 36: L06710, doi:10.1029/2009GL037293, 2009.

Pierce, D.W., Barnett, T.P., Santer, B.D., and Gleckler, P.J. Selecting global climate models for regional climate change studies. *Proc. Natl. Acad. Sci. USA*, 106(21): 8441-8446, 2009

Schiermeier, Q. The real holes in climate science. *Nature*, 463 (7279) 284-287, 2010.

Santer, B.D. et al. Incorporating model quality information in climate change detection and attribution studies. Proc. Natl. Acad. Sci. USA, 106(35): 14778-14783, 2009.

Seni, G., and Elder, J. *Ensemble Methods in Data Mining: Improving Accuracy Through Combining Predictions*. Morgan and Claypool Publishers, 2010.

Smith, R.L., Tebaldi, C., Nychka, D., and Mearns, L.O. Bayesian modeling of uncertainty in ensembles of climate models. Journal of the American Statistical Association, 104(485): 97-116. doi:10.1198/jasa.2009.0007, 2009.

Tebaldi, C., and Knutti, R. The use of the multi-model ensemble in probabilistic climate projections. *Phil. Trans. R. Soc. A* 365(1857): 2053-2075, 2007.

US EPA. Proceedings of the First National Expert and Stakeholder Workshop on Water Infrastructure Sustainability and Adaptation to Climate Change. Publication No. EPA 600/R-09/010, 2009.

Weigel, A.P., Knutti, R., Lininger, M.A., and Appenzeller, C. Risks of Model Weighting in Multimodel Climate Projections. *J. Clim.* 23(15): 4175-4191, 2010.

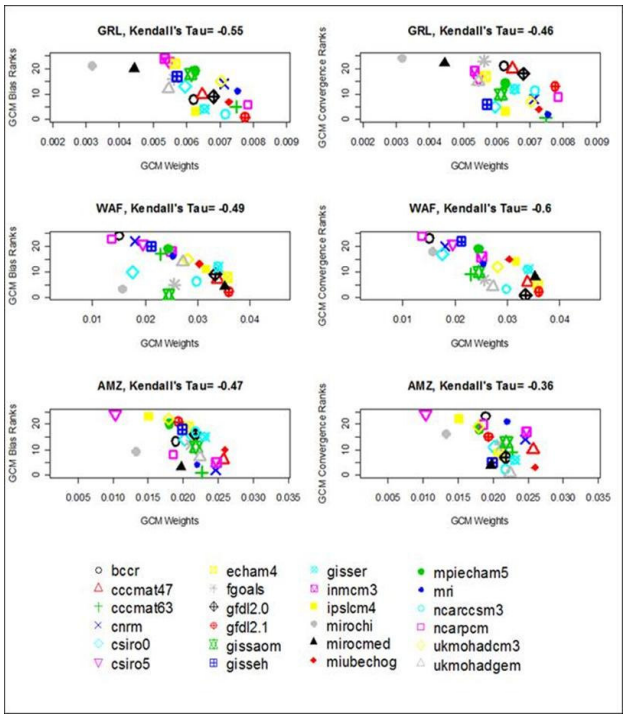

Figure 1. For each region, GCM rankings of the two criteria, bias and convergence, are plotted against average GCM weights as obtained from the Bayesian statistical model. In general, the two criteria seem to be related, perhaps non-linearly, to weights.

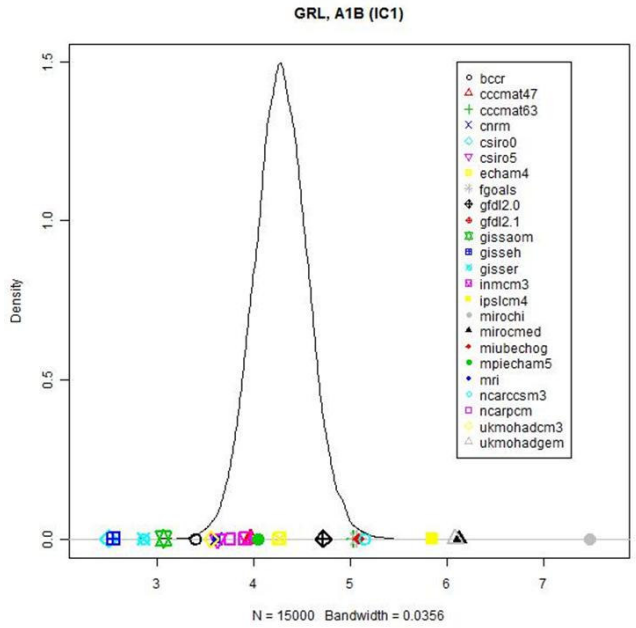

Figure 2. Posterior distribution of temperature change, υ-μ, for GRL. IC1 indicates that a particular set of initial condition GCM runs were used to generate this distribution. Results were found to be insensitive to different sets of initial conditions (not shown).