Geosci. Model Dev. Discuss.,
https://doi.org/10.5194/gmd-2020-253-RC2, 2020

Geoscientific
Model Development
Discussions
Open Access

Interactive comment on "A Markov chain method for weighting climate model ensembles" *by* Max Kulinich et al.

Anonymous Referee #2

Received and published: 10 November 2020

In this study, the authors present a novel method that employs Markov Chains as a means to weight members of global climate model (GCM) ensembles. Using three case studies involving historical simulations of global average temperature, regionally downscaled seasonal temperature, and a regional heat wave heuristic, they compare the performance of three model weighting schemes: simple model averaging (by definition, equal weights), a 'Convex Optimization Ensemble' (COE) method, and their 'Markov Chain Ensemble' (MCE) approach. Standard observational datasets from the recent past (up to ∼120 years) are used for comparison based on RMSE and R2 skill scores.

The proposed approach is interesting, and could be quite useful as a means to weight model ensembles and its simplicity is attractive, while also presenting a less ad-hoc approach than simple model averaging. However, I cannot evaluate the scientific merit of the approach because the model-observation tests are ill-posed in their current form. The main problem is that comparing unfiltered GCM time series to observations is very problematic when applying typical skill scores because the interannual variability will not correspond between CMIP5 models and the observations. So although the underlying trend or evolving signal (assuming there is a signal, such as in global temperatures) of a perfectly performing model should match what is observed, the full time series from the model would not necessarily match the observed time series. This is because in the CMIP5 experiments, the GCMs begin the experiments with different initial conditions, and different model runs within the same model will begin at different points in the same control run before then beginning the perturbation experiment (i.e. including anthropogenic forcings). This makes direct time series comparisons very tricky if not handled carefully. Take a very simple example as shown in the example figure.

Here are three simulated white noise time series (mean = 0, standard deviation = 1) overlaid onto two linear trends (0.1 for the black and blue time series, 0.03 for the red time series). The blue and red time series symbolize what could occur in a multi-member CMIP5 ensemble. Note that even though the blue time series has exactly the same trend as the 'observed' time series, the RMSE is higher than the red time series simply because the inter-annual variation is misaligned (of opposite sign in this case) with the observed anomalies. In contrast, the red time series has a lower RMSE, despite the fact that it does not capture the true forced trend. But the anomalies are aligned perfectly with observations. The red time series would be weighted higher in this case. It is 'right' for the 'wrong' reasons. Similarly the test set up in this manuscript is subject to the same problem. Individual models could exhibit anomalies that are more similar to the observed time series (driving down the skill scores), while the model response to the perturbation is less accurately simulated than other models with higher (by-chance) anomaly errors.

Of course, all the methods that were tested (AVG, COE, MCE) could be similarly biased, in which case perhaps the results hold. But, there is no way to ascertain that in the current test structure.

My other, more minor comments relate to similar issues with the structure of the model evaluation exercise. It is unsurprising that the model RMSE and R2 values were so poor when comparing GCM results to heat wave heuristics based on local weather station data (and again, where the model internal variability would have no reason except by chance to match the observed internal variability). The GCMs were developed at a scale that was never intended to resolve such localized patterns, and of course any annual heat waves that were observed, would only by chance occur in the same years (and be of similar magnitude) in the ensemble members.

To improve the evaluation I suggest the authors revisit the literature to see how others have tackled this problem. More attention should be paid to, for example, efforts by Sanderson et al. (2015) to carefully construct valid comparisons between GCM ensembles and observations (in this case, by focusing the model skill evaluation on climatologies, rather than time series anomalies) while also taking into account model inter-dependence; something which the authors admit they do not account for in their method. Others have addressed the non-initialized climate model/observation comparison problem by comparing long-term trends (e.g. Terando et al. 2012), which removes some of the problems with mis-matched internal variability, and gets closer to an actual forecast verification approach, but does not address other issues such as the robustness and reliability of weighting methods (see discussion in Knutti et al. 2017). It should be possible to construct a rigorous test of their MCE method, but the numerous challenges that have been widely and repeatedly documented in the literature should be acknowledged and addressed.

References:

Knutti, R., J. Sedláček, B. M. Sanderson, R. Lorenz, E. M. Fischer, and V. Eyring,

2017: A climate model projection weighting scheme accounting for performance and interdependence. Geophys. Res. Lett., 44, 1909–1918, doi:10.1002/2016GL072012. http://doi.wiley.com/10.1002/2016GL072012.

Sanderson, B. M., R. Knutti, and P. Caldwell, 2015: Addressing Interdependency in a Multimodel Ensemble by Interpolation of Model Properties. J. Clim., 28, 5150–5170, doi:10.1175/JCLI-D-14-00361.1. http://journals.ametsoc.org/doi/abs/10.1175/JCLI-D-14-00361.1.

Terando, A., K. Keller, and W. E. Easterling, 2012: Probabilistic projections of agro-climate indices in North America. J. Geophys. Res., 117, D08115, doi:10.1029/2012JD017436. http://doi.wiley.com/10.1029/2012JD017436.

Interactive comment on Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2020-253, 2020.

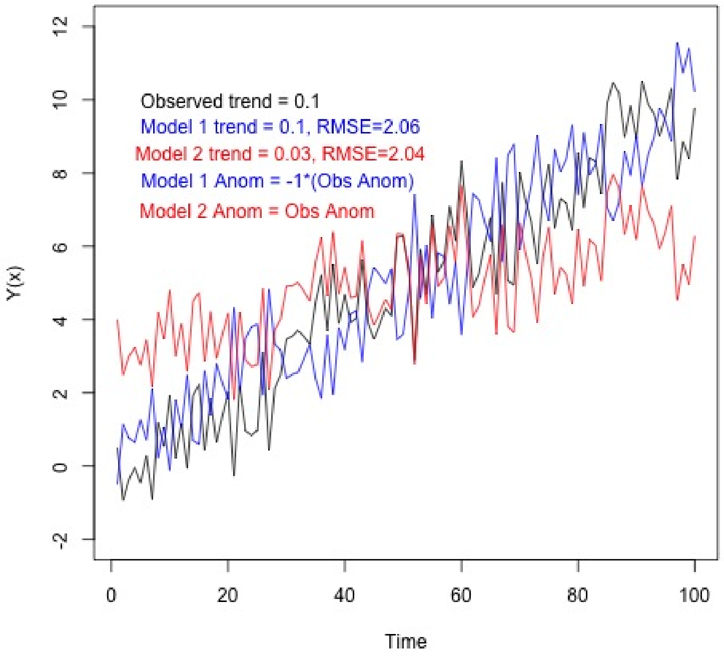

Fig. 1. RMSE Example