Improving the Potential Accuracy and Usability of EURO-CORDEX Estimates of Future Rainfall Climate using Frequentist Model Averaging

Stephen Jewson, Giuliana Barbato, Paola Mercogliano, Jaroslav Mysiak, and Maximiliano Sassi

Response to community comment 1,

Dear Rasmus,

Many thanks for taking the time to read our paper and make helpful comments. I have pasted your comments below, greyed out, and then responded to them in black.

Discussions about *estimation uncertainty* is really timely and important, and I was pleased to come across this paper. One question I have is if this is the same concept that my group has tried to cope with that we call "*the law of small numbers*". We use large multi-model ensembles and downscale them with empirical-statistical donwscaling since we find that RCM-based ensembles tend to be too small, especially since they are not independent and involve many of the same GCM simulations. This is demonstrated in Mezghani A., A. Dobler, R. Benestad, J.E. Haugen, and K.M. Parding (2019), Sub-sampling impact on the climate change signal over Poland based on simulations from statistical and dynamical downscaling, J. Appl. Meteor. Climatol., 0, https://doi.org/10.1175/JAMC-D-18-0179.1.

In the abstract of your paper you say:

Further, an additional bootstrap test revealed an underestimation in the warming rate varying from 0.5° to more than 4°C over Poland that was found to be largely influenced by the selection of few driving GCMs instead of considering the full range of possible climate model outlooks. Furthermore, we found that differences between various combinations of small subsets from the GCM ensemble of opportunities can be as large as the climate change signal.

This certainly sounds like estimation uncertainty, due to the combination of a large spread in the ensemble and a small ensemble size. If it is, then if you applied the methods I describe in the manuscript to each subset, the differences between the subsets would be smaller, because the changes would all be closer to zero.

We also look at ways to evaluate downscaled results from large multi-model ensembles that involve 5 different levels. Two of these look at the ability of the downscaled GCM results reproduce the historical trends and interannual variability (e.g. Benestad, Rasmus; Parding, Kajsa; Isaksen, Ketil, Mezghani, Abdelkader (2016) “Climate change and projections for the Barents region: what is expected to change and what will stay the same?", ERL-102170.R2, DOI: 10.1088/1748-9326/11/5/054017). My question is how such efforts can be combined with SMMA/BMMA to improve our ability to assess the skill of the projections, e.g. for disaster modelling.

Probably the best way to influence the disaster risk modelling community is to publish results (as you have done) that then feed into the big reports (EU, IPCC). My experience has been that the summaries in the big reports are very influential (since many people don't have time to read all the individual papers).

What's the signal-to-noise ratio of the results in your paper (or equivalently, the p-value)? If they are clearly significant, then it wouldn't make sense to apply the methods described in our paper. If they are borderline significant, or not significant, then it might make sense.

The result in that paper that storminess in certain regions might increase is interesting. Is that a robust result across different studies, do you know, and is the change large enough to matter? If yes and yes then that's certainly something that the disaster risk management community should be interested in.

Best regards,

Steve Jewson