Nonlinear Processes in Geophysics Discussions

Nonlin. Processes Geophys. Discuss., referee comment RC2
https://doi.org/10.5194/npg-2021-12-RC2, 2021

Comment on npg-2021-12

Anonymous Referee #2

Referee comment on "Improving the potential accuracy and usability of EURO-CORDEX estimates of future rainfall climate using frequentist model averaging" by Stephen Jewson et al., Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2021-12-RC2, 2021

This manuscript describes two variants of a correction method for the ensemble mean that aims to minimize predictive root mean square error (PRMSE). The methods are then applied to three rainfall variables in the EURO-CORDEX regional climate ensemble.

The correction method consists in adding a scaling factor k to the estimator of the ensemble mean, and minimizing the PRMSE with respect to k. Note that this occurs at the expense of the expected bias, and hence is referred to as a bias-variance trade-off. One of the goals is to provide an alternative for statistical significance testing for a nonzero signal, as the latter gives rise to spatial discontinuities.

General comments:

The manuscript addresses a relevant question and is well-structured. The examples clearly show in which situations the method reduces the PRMSE. Some sentences are a bit long or hard to parse but otherwise it is well-written.

The chosen scaling approach seems quite ad-hoc to me, and this could be improved by a better motivation and/or more context. If I understand it well, the work takes an operational correction method that seems to work well in practice, and tries to motivate it in a more scientific way and apply it in the context of climate change.

Secondly, I do not find the term "Model Averaging" appropriate here. In the broadest sense of the word it can indeed be seen as model averaging (with equal weights) and a fictitious "zero change" model. However the terms "calibration" or "post-processing" are more appropriate for this kind of approach.

The method is proposed as an alternative to significance testing, which introduces spatial discontinuities. This seems like an unfair comparison though, as significance testing provides maps of significant climate change as a qualitative indicator (e.g. shading), but the resulting discontinuous fields are not typically used in impact models; or else please provide references.

In general the manuscript would benefit from a better motivation in the form of a concrete example or impact model for which the reduced PRMSE from this method provides a tangible benefit over using the uncorrected ensemble mean. Perhaps an example from agriculture, urban hydrology, infrastructure planning… ?

Specific comments and questions:

If you want to use the current title please provide strong arguments that the potential accuracy and usability is improved. This depends on your definition of accuracy (since the bias is increased and this might be detrimental for many use cases). Do you use a proper score? How does it affect other known scores such as the Brier score?

L34: Aren't the correlations, present in ensembles, very relevant for catastrophe models? Doesn't one lose much of this information when moving to probabilities?

L98: You perform cross-validation in the ensemble, and elsewhere in the manuscript you assume that any model correlations or biases should be tackled first, before applying the methods. Did you do this for the ensemble and if not, what are the implications for cross-validation?

L116: Did you use the historical experiment for the 1981-2010 baseline? This period includes a few years of the RCP scenarios (the 2005-2010 period), did you always use the corresponding RCP data in the comparison or did you choose the same baseline for both RCPs?

Fig. 1: the image quality of the maps a-c is bad (compression artefacts) and I cannot easily distinguish positive from negative change. It would be useful if 0 change has a clearly identifiable color (e.g. white).

L169: Here you mention model dependence should be addressed. You could perhaps mention briefly how this could be done (cluster analysis, expert knowledge on the models, ...?) and whether you do it for your example.

L173: "the distribution they are sampled from perfectly accounts for their biases". The meaning of this sentence is unclear to me. Do you mean the model errors are on average unbiased? Or the model biases compensate for each other (the biases of the models, aggregated over the ensemble, behave like a statistical error and not a systematic one)?

L180: Any reason why you don't use the unbiased estimator (divide by n-1)?

L200: Could be unclear whether the square is inside the E(...) function, maybe add brackets?

L230-245: This might be clarified by adding an equation for the distribution here.

L246-256: Section 3.3: seems a bit brief to me, again I feel like some equations could clarify this. Do you approximate a marginal distribution (compute an integrate) by sampling. please make this explicit. Why 250 values?

L258-259: I find this sentence a bit strange, didn't you just derive these estimates to minimize PRMSE? Here you just verify your method with a numerical experiment.

L282: how does it behave as a function of the estimated SNR value?

L309: Section 3.5: Move this to earlier in the manuscript

Fig. 4 is very confusing to me, or there was something I didn't understand. Am I correct in understanding that the precipitation reduction is reduced quite a lot (so the drying is less severe) in the southeast of Spain? But it seems like the SNR was quite high there, so why this large change? I think it's unclear partially due to the choice of the color scale. I find it hard to read the SNR map in Fig. 1; the shades of orange/pink are too similar.

L377: Small changes when PMLL is used as a metric: is this also found in literature?

L431 SSMA -> SMMA

Fig 9: Figure labels are not present (a-d)

The y labels don't match the description and this is confusing. What is the "signal" in panel b? Mention what PRMSRE stands for in panel c.

L444: Root mean square size: please clarify, why are the values in the figure smaller than 1?

L514: between scenarios? I guess it's normal that there can be a jump, did you mean something else (between time periods)?

L518: "Falsely identifying a change as being due to climate change": this is related to (erroneous) attribution, which is something else than a false positive (detecting change where there is none expected).

L519: Please show how this is beneficial for risk modelling or add a reference.