Interactive comment on "Post-processing of seasonal predictions – Case studies using the EUROSIP hindcast data base" *by* Emmanuel Roulin and Stéphane Vannitsem

Anonymous Referee #1

Received and published: 21 October 2019

The authors compare several simple post-processing methods for seasonal forecasts of temperature, pressure and precipitation from five models in the EUROSIP ensemble forecasting system. Results are presented for six regions of southern and western Europe, at lead times of between zero and two months. Both deterministic and probabilistic forecast skill is assessed relative to climatology. Results are presented for both individual models and various combinations of the EUROSIP models.

I am unclear what the purpose of this study is. A large amount of effort has been expended to perform a huge number of comparisons. However, the study lacks sufficient structure to provide useful, original, generalizable results regarding either the

best post-processing methods, the best models, useable lead times, useful ensemble sizes or required training period. In its current form this manuscript feels more like an internal technical report than a piece of research of international interest.

The authors are clearly motivated by the two hydrological applications described in Section 2.1, but the inclusion of the other four study areas seems unnecessary and makes the manuscript extremely long. The additional study areas do not appear to be chosen to be representative of difference climate regimes or other systematic differences and no reference is made to their relative positions or conditions in the text.

The introduction mentions a lot of details relevant for post-processing studies, but at the same time is often disjointed and lacks context, jumping between topics in the same paragraph. The authors also specifically mention some previous findings and recommendations which they later ignore, e.g., Page 3, Lines 5-7.

The notation used throughout Section 2 to describe difference post-processing methods is never explained, making it difficult to be certain of what is being proposed.

Most of the results presented are based on in-sample comparisons which would never be used in practice. Why not just show the cross-validated results which are more believable?

The authors often attempt to summarize the results by the number of scenarios where a given post-processing method performed best. This seems potentially problematic and misleading given the small number of regions considered, and the fact the results are pooled across regions, seasons, lead times, skill scores etc.

On Page 14, the authors appear advocate choosing the method of confidence interval calculation based on which one gives most significant results rather than any scientific basis!

Finally, in Section 3.4 we are told that almost all the positive results described over the last eight pages were simply due to warming trends.

In summary, I find no evidence of original results that would be of interest to a wider audience.

C3