Review of ‘Assessment of S2S ensemble extreme precipitation forecasts over Europe’

In this paper, the authors characterise the forecast skill of extreme precipitation in the ECMWF S2S hindcast ensemble, taking some account of seasonality, and considering the impacts of spatial and temporal aggregation on the forecast skill. They measure skill in a deterministic way using the Brier score and using the binary loss index.
They find that extended winter is more predictable than extended summer, and that temporal and spatial aggregation both extend the last skilful day. They also identify several regions where forecast skill is higher: Norway, Western Iberia and the South of France. In general more mountainous and coastal areas show higher skill.

I like the analysis approach, and the temporal and spatial aggregation is well done, as is the bootstrap approach to determining the last skilful day. I have a minor methodological question, but I also see some more substantial issues and so am recommending major revisions.

Firstly, I am a little confused by the choice of the BLI as a metric. A fair amount of time is spent discussing this new metric, but as the authors point out it is not at all novel in meteorology, simply being (1 - the well known Critical Success Index). This raises the question of why not just use the more established CSI from the beginning?

More generally, my main sense is that the analysis, while well done, is quite basic. Skill maps are computed for some variant event definitions using two different seasons and scores, and we are done. As you discuss in your conclusions, there are many interesting questions that arise from this foundation. I think the manuscript could do with answering at least some of them.

Without suggesting you wildly broaden the scope, I suggest answering the following questions:

- How does the skilful day change for different cost-loss ratios (i.e. using different ensemble thresholds)? Understanding this sensitivity has a lot of real-world relevance. To simplify presentation you could average skill over a few boxes of interest, so you had scalar values for each threshold.
- How does the spatial pattern of skill change for compound events? Yes, we can in theory read this from figures 6 and 7, but perhaps some anomaly plots might be helpful here.

These are only ideas; the main point for me is that some extra richness is needed, whether that be a discussion of regional differences, dynamical drivers, sensitivity of the results, a deeper analysis of scale-dependence etc.

On top of that the current comparison of the Brier score and BLI are a bit superficial, and should be discussed in more detail.

Minor comments

- The Github repository, which is supposed to contain the code, is empty
- You define the forecast event thresholds using the forecast data, conditional on both season and lead time, rather than using the observational thresholds to account for bias. But is there a risk here that you make the model seem too good? You are bias correcting with your testing data! If you were to set thresholds with only half your data and then test the skill in the other half, would the skill go down? I would like to see this for at least the unaggregated case.
- Can you make sure figures 5, 7, B1 and D2 all use the same colorbar? Currently it is hard to compare them.

Very Minor comments

- References need formatting to be in parentheses
- ‘Skilfull’→’skilful’ L185
- ‘BLF’→ ‘BLI’ L204