General comments

The paper is generally well written, although I found the sections on the actual postprocessing technique tough to read. Although I have done some postprocessing of precipitation forecasts myself, I find this a difficult topic and I have concentrated on how the resulting forecasts have been verified.

Although the ensemble forecasts have been verified to some extent, I think a more extensive verification could improve the manuscript. If, however, the authors decide otherwise, I would like to see a note added that explains the limitations of the 'as is' verification. This pertains to both the (absence of) conditional verification in terms of CRPSS and bias, as well as to the verification of added value of the Schaake shuffle.

I don't mind revealing identity (Jan Verkade, Deltares, Delft, the Netherlands). I don't need to see a revised version of the manuscript; I trust that the authors will deal with the comments appropriately.

Specific comments

Section 1: Introduction

- P6767, l12 onwards: why single out initial conditions and weather forecasts as sources of uncertainty, and not also mention the models, parameters also?

Section 2: Study catchment and data

- I think it would be worthwhile to say something about response times of the catchment and of subcatchments within Ovens. This would further clarify the rationale for having ensemble precipitation forecasts at a sub-daily time step.

Section 3: Methods

- Throughout the paper, the terms "forecast period" and "lead time" are used. Do they indicate the same thing? If so, maybe good to mention that these are used synonymously, or use a single term only. P6774, l4 onwards is an example of this.
- Section 3.1, p6774, l11 onwards: "The post processed probabilistic forecasts of three hour rainfall accumulations (for lead times of 0–60 h) do not contain appropriate spatial and temporal correlation structures.". I believe this may indeed be true, but we really need to see some evidence of this, either by supplying an argument or through some quantitative measure. Alternatively, you could cite evidence gathered elsewhere.
- You may want to consider moving most of the contents of section 3.3 to an appendix, as these are standard metrics.
- Section 3.3.3: note that here (i.e. in the case of ROC), 'unskilled' has a different meaning than in the section on CRPS.
- Section 3.3.3: "Post processing does not influence forecast discrimination". While this claim has been substantiated with a reference, a brief argument would help the reader understand this.
- Section 3.3.4: I don't think reliability plots for temporally aggregated forecasts say anything about the Schaake shuffle's ability to restore spatial correlations. Please change accordingly.

Section 4: Results

- This section contains both results and conclusions. I recommend renaming it accordingly.
- CRPSS and Bias Score are presented for the full available sample only. While for the full sample, the metrics show an improvement in forecast quality, I would be interested in seeing the results for higher quantiles of the climatological distribution. For example, if one would single out observed precipitation events of 5mm/d and higher only, what would CRPSS and bias be then? I would highly recommend augmenting the analysis this way. Please refer to Brown and Seo (2013) for a good example of how this can be done. If, however, authors decide otherwise, then I would like to see a note about interpretation of the scores in results and/or discussion section, explaining the limitations of a full sample analysis only.
- After reading paragraph 4.2.3, the reader could be forgiven for thinking that the diagonal line in the ROC plot corresponds to the ROC curve of the climatological forecast. This is not the case. I would suggest sharpening the text to distinguish between the ROC curve of a climatological forecast and the diagonal line, which corresponds to a situation in which there is no correlation between a forecast and an observation.
- Similar to my comment above, I would be interested in learning about ROC curves for more than two events only. As it would be impractical to plot many more figures, you could consider not showing ROC curves, but showing values of the AUC instead. Note that while I think this would constitute an improvement, it is not strictly necessary.
- The reliability plots in figures 10 and 11 each contain three points only. Why is that? The plots would be more informative if they would show, for example, observed relative frequencies for forecast probabilities {0, 0.1, 0.2, … 1.0}. Conceivably, this would more evenly distribute samples between the then 11 points on the curve near the now fullest bins.
- In section 4.2.4, it is stated that "For day 2 the forecast probability of a rainfall event of greater than 5mm appears to be unreliable." I agree that it appears to be less reliable than the other forecasts, but when is a forecast reliable and when is it not? To facilitate comparison, you may want to consider looking at the decomposition of Brier's probability score, and then specifically at the reliability component. Again, while I think this would constitute an improvement, it is not strictly necessary.
- While I agree that the Schaake Shuffle aims to restore space – time correlations, I don't think the spatial correlation is tested here. As a matter of fact, I'm not so sure about the temporal correlation, either – I would be interested in seeing verification results of aggregated precipitation before and after application of the Schaake Shuffle. I am guessing that this will show that indeed, temporal correlations have been restored, but until then we simply don't know.
- The Results section does not contain any information on the uncertainty in the verification metrics, except for some comments related to small bin sizes for some of the points on the reliability diagrams. I would like to see a statement on this, preferably in quantitative terms but in any case in qualitative terms.
- I would recommend merging Figures 4 & 5 and 6 & 8 respectively, after connecting the points with lines.

Section 5: Discussion

- The Discussion section is quite good in how it describes merits and limitations of the postprocessing technique. I have no comments to add to that.

- I think the real proof of this method is when the postprocessed precip forecasts are used to force the hydrologic model. This will put both the BJP and the Schaake shuffle to the test. I am much looking forward to seeing the results of that!

Section 6: Conclusions

- This section contains a summary of the manuscript only. I would recommend renaming the title to “Summary”. I would then also rename section 4 to “Results and Conclusions”. You may even consider removing Section 6 altogether, as it contributes little to what’s already stated in the abstract.

Technical corrections

- P6768, l15: “To generalise the approach requires” is grammatically incorrect. Maybe “Generalising the approach requires” would be better.
- The prefix ‘post’ in conjunction with ‘processing’ can be written either with or without a hyphen, but if the latter option is chosen, the words should be joined: postprocessing, not separated.
- I find the use of the term ‘forecast period’ somewhat confusing; why not use ‘lead time’ instead (similar to in Section 2, p6770, l25).
- P6771, l7: consider omitting local time. It doesn’t contribute much.
- P6773, l13: “is used to generate~~d~~ a value”
- P6777, l18: “by a single point”
- P6777, l18: “Here, ROC plots…”
- P6780, l21: “than in the raw forecasts”
- P6781, l4: “than for the events where rainfall is less than ~~the event of rainfall less than~~ 0.2mm.”
- P6784, l5: overfitting
- Fig1 uses a colorscale that I think is somewhat unusual. It is more common to associate blue with lower altitudes, and yellow/red with higher altitudes.
- Figures 2, 9, 10 and 11 use a large margin between plots. I suggest reducing, or removing the margin and labelling top row and upperleft columns only. This will increase legibility of the figures.
- The figures lack legends. Instead, the authors have described the meaning of the lines/dots in the caption. I think the figures would benefit from a proper legend, though.

Bibliography

Brown, J. D. and Seo, D.-J.: Evaluation of a nonparametric post-processor for bias correction and uncertainty estimation of hydrologic predictions, Hydrol. Process., 27(1), 83–105, doi:10.1002/hyp.9263, 2013.