Reply to the RC1

First we would like to thank the RC1, Dr. Jonas Bhend, for his comments and for the time he spent on the paper. You will find our point-to-point response below. It is organized as follows: the initial comments are in blue and the response for each comment in black. For some major comments we answer directly in black after certain sentences. The changes in the manuscript are in red with the line or figure number. If an entire section has been modified or reorganized, the line number is omitted.

General comments:

The introduction to the postprocessing methods (Sections 2 - 4) and the constant switching back and forth between temperature and precipitation is difficult to follow. I suggest to either reorganize the discussion to first introduce the complete temperature processing chain and then the rainfall processing chain, or to better guide the reader through the manuscript. For the latter, a more detailed description of the (similarities in) processing across the two cases in the introduction (L 41ff) would be helpful. In particular, this description should reflect the structure of the remainder of the manuscript to manage reader expectations.
This was a major point of debate between the authors. We have followed your advice and organized Sections 2-4 into one section (Section 2) dealing with the complete temperature processing chain and another section (Section 3) dealing with the complete rainfall processing chain. Moreover, a flowchart for each post-processing procedure has been provided.

The discussion of the operational challenges now in the conclusion should be moved to the discussion section (5). In the conclusion, I instead propose to re-emphasize the benefits and challenges of the postprocessing as being implemented at Meteo France and also try to highlight the aspects of the processing chain that are portable (across parameters, but also to other NHMSs). We agree with this suggestion; the conclusion has been rewritten.

The manuscript would clearly benefit from thorough copy-editing. Some of the errors I have mentioned in the section below (e.g. some of the many articles missing), but my edits in that regard are not complete and I therefore encourage the authors to invest into polishing the language. At the very least, please use a spell checker before submission!
The manuscript will be checked by an English speaker.

Minor comments:

L35: as one of the ok

Section 2.1 and 2.2: why “For”, could just be shortened to “ARPEGE and ARPEGE EPS” ok

L65: throughout the years ok

L85: on the limited area of Figure 1. The associated 16-member EPS, called PEAROME . . . ok

L87: for better readability, I suggest to indicate grid resolution only in km as above. ok

L103: It might be easier to introduce ICA here than in the table ok

Table1, 2: please indicate the target parameter also in the table caption ok

L106: exhibits is confusing. Do you mean that 400 samples are drawn corresponding to the 400 ensemble members you have available for ECC? Please rephrase. ok

L117: Hard to read. Do you mean: The final groups (called “leaves”) contain training observations with similar predictor values? Yes.

L119: “robustless” doesn’t exists nor do I think it applies here. The predictions could be characterized as very robust but are probably prone to overfitting? We have changed this sentence to:

Binary decision trees are prone to unstable predictions insofar as small variations in the learning data can result in the generation of a completely different tree .

Table3: is hard to follow. Could table 3 be made easier by rearranging by member rather than by grid point? ok

L180: post-processing can introduce rain in a grid box that is dry in raw member only if there is a grid point with rain close by in the raw member. ok

L216: For the interpolation of climate data, most of the time only topographic data is available which may play. . . ok

L225: with greatly varying data density ok

L226: Note that the size ok

L236ff: the lines should be joined with + yes

L247: Please specify the impact of model selection. Are generally all predictors used, or is the model usually greatly reduced in complexity due to model selection ? The following sentence has been added. This model selection is influenced by the weather situation, but most often selected variables are the linear projection function of Tj and/or altitude effect – since those two are very well correlated. Distance to sea and PC1 may also be selected quite frequently. PC2 to PC4 selection is much less frequent.

L288: It is not clear to me whether you assess the currently operational spatialization algorithm against COSYPROD (an earlier algorithm) or whether you assess an earlier version of the currently operational spatialization algorithm against COSYPROD (which would predates either of the algorithms).The tested algorithm is the first version of the current operational algorithm – a bit simpler as it does not take into account the linear projection function of Tj. This earlier version is the tester COSYPROD method, which predates our proposed method (both first and current versions). Replaying the full benchmark and adding the linear projection of Tj would have been complex. This benchmark was realized several years ago. Since we added an extra (and valuable) predictor in the full process, we do expect that the conclusions hold for the current operational algorithm – which should even be a bit better than exhibited. We add “current spatialization” to “an earlier version of the algorithm over France” and “which predates both first and current versions”.

Figure 5: please redraw this figure for better readability of the labels. The grid lines, which are now quite dominating, could be reduced in weight and the data lines on the other hand should be increased in weight to stand out. Also the legend only has to be shown in the main panel. This figure has been redrawn and labels rescaled.

Figure 5: Maybe include the number of stations used for evaluation in the figure caption. ok

L298: large set of climatological ok

L300: independent ok

Figure 6: a lot of the figure space is lost to redundant labels. Maybe this could be integrated in one single plot with multiple groups of boxplots? There is a new Figure 6 now. Thank you for the suggestion.

L305ff: please combine the following sentences into one paragraph ok

L307: The temperature field . . .. with the same ok

L316: What is the period for comparison? This sentence has been added: the validation is made by a 2-fold cross-validation on the two years of data (one sample per year).

L317: Due to the high ok

L317: I do not understand the rationale to aggregate the scores across lead times. A lead-time-specific or spatially explicit skill plot would be much more informative. This aggregation relies on a huge amount of data to verify and a number of different configurations tested (more than 920). Moreover, different HCA are chosen for each lead time and so inter-lead-time comparisons are difficult. This paragraph is now more detailed.

L320: If the only conclusion drawn from Figure 12 is that postprocessing improves the forecast, Figure 12 could be dropped and the quantitative results could be mentioned in the text (e.g. PP reduced forecast errors by XX-YY. We removed this figure and mention the improvement in the text: The averaged CRPS between the raw and the post-processed ensemble is improved by approximately 30% (from 0.118 to 0.079).

Figure 7-11: I really like the figures and the intent, but I think in the paper, this cannot be presented at the scale one can actually still recognize details. Therefore, I suggest to rearrange the plots to better illustrate the processing steps and the merits (locationof front). To illustrate the postprocessing steps, I suggest to show a zoom-in: one figure with the 5 subpanels (raw member at 0.1, raw member at 0.01, spatial trend, field of residuals, result) for a small area in the French Alps or wherever you deem suited. This could be complemented with a small multiples plot showing the whole domain where the focus is on the position of the front in the raw and postprocessed members. We agree. Figures 7-11 have been replaced by the set of figures suggested. Thank you for this excellent idea.

Figure 12: What is depicted by the box and whiskers? Is it spatial variability or some bootstrap resampling? Please specify. Also, if it is spatial variability, it may be beneficial to show a map (see comment above). Bootstrap resampling. In any case, Figure 12 has been removed.

L323: Figure 13 focuses. . . ok

L323: please specify what exactly constitutes a rain event. It is now defined in the text. Are the same thresholds
applied to observations and (postprocessed) forecasts? Yes.

L326: frequency bias ok

L326: Please also discuss the tendency of the postprocessed forecast to underpredict rain events at least in some areas? Is this a feature of the short time period used for evaluation and/or the limited predictability? Is a small or considerable fraction of the forecasts underpredicting rain? Is there an interpretable spatial pattern to it (see comment above on figure 12). The categorical performance diagram in Figure 13 is not fully used. You noticed the tendency of the postprocessed forecast to underpredict rain. We realized that have not explained the construction of this Figure and that it is confusing. Like the ROC curve, the curve in the performance diagram is computed for each quantile of the forecast. This latter sentence has been added. We can assume here that the minimum of predictive distributions nearly never forecast rain occurrence, but when it does, a false alarm is never made. The minimum of the raw ensemble detects rain occurrence around 40 times over 100, but when it does this forecast is wrong around 35 times over 100 (1-success ratio). This remark has also been added.

L327: increased by 15 ok

L328: Figure 14 depicts . . . ok

L331: The conclusion that the improvement in forecast value is constant needs better support. With only two thresholds sampled with similar (absolute) improvement in max PSS, I think this is a bit of a stretch. This conclusion has been removed. Also, the absolute improvement is constant, but I assume that for most applications relative improvement would be more relevant. It is not clear to us what you mean by relative improvement. We guess that it is the decomposition of the PSS in Hit Rate and False Alarm Rate. If so, a sentence replaced the (too conclusive) former one by : Most of this improvement is due to the the improvement of the Hit Rate.

Figure 13: The size of the axis labels should be increased for better readability. Also, I would love to see how many cases fall into the bins of the reliability diagram (e.g. the relative contribution via the point size) and in the performance diagram it is not clear where the mass of the distribution is. In case it is many more points than can be visually separated, binning (with bin size corresponding roughly to discernible points in the reliability plot) and point size or some shading with transparency could be used to let the reader focus more strongly on the bulk of the distribution. The Figure 13 has been redone.

Figure 14: Please specify which line is the raw and which the postprocessed forecast. ok

L333-343: spell check! ok

L333ff: this section may profit from reversing the order of arguments. First start with the verification of daily rainfall totals (improvement, but relatively less than for hourly rainfall due to the ‘disappearance’ of timing errors in the raw model forecast). Second, does it also work for extreme events (we don’t have the full verification, but present a case study)? This has been done. Thank you for this (logical) suggestion.

L342: It is not clear to me what is meant by “The bc-ECC method does not solve temporal agregation. As a consequence, it is not surprising that the daily post-processed CRPS is roughly 24 times the averaged hourly one.” Is there a problem with temporal aggregation that needs solving? Aggregation is not the right word, so we have replaced it with “penalties”. Do you intend to say that because the target is hourly precipitation, the effect of postprocessing is not as beneficial as it could be if daily precipitation were the target of postprocessing? Completely. We think that a direct postprocessing of daily precipitation is more effective. It is an intuition that comes from the nature of daily precipitation compared to hourly ones (fewer zeros, smaller variance…). This sentence has been added : Due to the nature of daily, compared to hourly, precipitation distribution (fewer zeros, smaller variance and lighter tail behavior), we believe that a direct post-processing of daily precipitation is more effective if the target variable is daily precipitation.