Nonlin. Processes Geophys. Discuss.,
https://doi.org/10.5194/npg-2019-65-RC1, 2020

Nonlinear Processes in Geophysics
Discussions

Interactive comment on "From research to applications – Examples of operational ensemble post-processing in France using machine learning" *by* Maxime Taillardat and Olivier Mestre

Jonas Bhend (Referee)

jonas.bhend@meteoswiss.ch

Received and published: 6 February 2020

The authors present two applications of machine learning for the operational postprocessing of numerical weather predictions. Their manuscript provides a detailed insight into the full processing chain including state-of-the-art ensemble postprocessing and the many necessary adjustments to fulfill the requirements for high-resolution probabilistic forecasts. Also, the challenges when running a complex ensemble postprocessing in operations are briefly mentioned. The presentation of the results (both text and figures) and their discussion, however, should be reworked for better readability. Therefore, I recommend to accept the manuscript with minor revisions.

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.

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).

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!

Minor comments:

L35: as one of the

Section 2.1 and 2.2: why "For", could just be shortened to "ARPEGE and ARPEGE EPS"

L65: throughout the years

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

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

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

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

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.

L177: Hard to read. Do you mean: The final groups (called “leaves”) contain training observations with similar predictor values?

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?

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

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.

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

L225: with greatly varying data density

L226: Note that the size

L236ff: the lines should be joined with +

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?

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).

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.

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

L298: large set of climatological

L300: independent

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?

L305ff: please combine the following sentences into one paragraph

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

L316: What is the period for comparison?

L317: Due to the high

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.

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

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 (location of 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.

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).

L323: Figure 13 focuses. . .

L323: please specify what exactly constitutes a rain event. Are the same thresholds applied to observations and (postprocessed) forecasts?

L326: frequency bias

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).

L327: increased by 15

L328: Figure 14 depicts . . .

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. Also, the absolute improvement is constant, but I assume that for most applications relative improvement would be more relevant.

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.

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

L333-343: spell check!

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)?

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? 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?

Interactive comment on Nonlin. Processes Geophys. Discuss., https://doi.org/10.5194/npg-2019-65, 2020.