

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 exist 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 predate 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 (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. 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 aggregation. 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.**

Reply to the RC2

First we would like to thank the RC2 for the time and the work he/she spent on the paper. You will find our point-to-point response to his/her comments below. It is organized as follows: the initial comments are in blue and the response to each comment in black. For some major comments, we answer directly in black after certain sentences. The changes in the manuscript will be in red, with the line number or the figure number. When an entire section has been modified or reorganized, the line number is omitted.

Major remarks:

1. For post-processing temperature and precipitation data are used from 2-year periods. However, the paper does not mention which part of the data set is used for training and which part for verification and whether cross-validation has been used or a completely independent verification period is considered as in the verification of daily precipitation amounts (Fig. 16)? Besides, it would also be good to add the verification period in the captions of Figs. 5, 6 and 12-14. The validation is made by a 2-fold cross-validation on the two years of data (one sample per year). **This sentence has been added in the text and in captions.**
2. The hyperparameters of QRF (like the number of trees and the terminal node size) should be added and whether these are optimized for the training period. **This information is now provided at the beginning of the verification section.**
3. I think it is good to add flow diagrams in which all steps involved in the post-processing of temperature and precipitation are displayed. Thank you for this excellent suggestion. **Two flowcharts are now available.**
4. The conclusion section is rather short and there is no discussion paragraph. The current conclusion section mainly focuses on computational aspects, which could be moved to a separate earlier section. Instead it would be good to have a real conclusions and discussion section in which the main conclusions are given and the results are placed in context (in relation to other papers if possible) and in which future work is mentioned. **The former conclusion section is now reshaped as the discussion section and a “real” conclusion section has been added.**
5. Some of the figures are really small and will become much more clear if they are enlarged, notably Figs. 7-9, 11, 13 and 15. Alternatively the legend in some of the figures can be enlarged to be readable. **Most of the figures have been redrawn, and labels rescaled.** Thank you for this suggestion.

Minor remarks:

1. Lines 2 and 18: “misdispersed” and “misdispersion” do not seem to be correct English words. Please replace. These terms have been replaced with “poorly dispersed.”
2. Line 5: Please add something like “and subsequent interpolation to a grid” after “temperature”. ok
3. Line 40: Please introduce the abbreviation EGP. done

4. Lines 44 and 46: Please add references for the two EPS systems. ok
5. Line 47: Please place “(calibrated with rain gauges)” after “data”. ok
6. Line 53: Please replace “adjustments” by “adjusted”. This appears to be a misunderstanding of the sentence. The noun form “adjustments” is correct in this case, as it is the subject of the clause.
7. Line 83: Why is it needed to apply the spatialization algorithm twice? Does ECC not account for that? The spatialization algorithm is applied once. A **flowchart has been added**.
8. Line 89: Please replace “subgrid” by “grid”.ok

9. Line 91: I would say spatial penalties issues are reduced rather than solved.
10. Lines 94-95: Please replace “calibrated (with rain gauges) radar data ANTILOPE” by “radar data set ANTILOPE (calibrated with rain gauges)” and add the resolution of that data set. ok
11. Line 100: Please insert “potential” before “predictors”. ok
12. Line 138: Please provide a reference for the method of moment. ok
13. Line 141: Please insert “forecasting a” before “temperature”. ok
14. Line 159: I would use another title for this section. As the paper has been reorganized, this section has now been omitted.
15. Line 167: Please replace “data” by “observations”. ok
16. Line 171: The use of “natural” is a bit strange here. **This has been replaced by “innate”**.
17. Line 181: Please replace “close” by “close by” and add “fields” after “cover”. ok
18. Line 224: Is one year of data enough? We are limited by (too frequent) model updates here.
19. Lines 237-240: Please correct the equation: “+” instead of “=” and use the “for all” symbol (?) instead of the infinity symbol. ok
20. Line 247: Please delete β 1D the 2nd time is it mentioned and replace the multiple α 1D 's by α 1D , α 2D , α 3D , α 4D .ok
21. Line 248: Please introduce the abbreviation AIC.ok

22. Line 292-293: Please add a reference for the COSYPROD interpolation scheme. COSYPROD is just the name for an internal IDW-like scheme. **IDW references have been provided**.
23. Line 307: It is not necessary to start a new paragraph here. ok

24. Line 329: Please replace “according to” by “for”. ok
25. Line 331 and caption of Fig. 14: I would not say that the improvement is constant. Also noticed by the RC1. **We have removed this statement**.
26. Axis labels of Fig. 13: Please move 1 of the 2 labels to the other side of the figure (top panel) and add “= 1-FAR” after “Success Ratio” (bottom panel). Done.

27. Caption of Fig. 13: Please add that the red curves are for the post-processed forecasts and add the meaning of the different background curves and dotted lines in the bottom panel (respectively CSI and bias). Ok, **the CSI and bias are now in the caption**.
28. Line 334: Please replace “propose” by “show”.ok
29. Line 336: Please insert “24-h” after “Observed” and delete “in the day”.ok
30. Line 340: I would say “slightly improves” instead of “does not deteriorate”. “Slightly improves” would be a bit of a stretch.
31. Line 341: I would replace “Figure 12 for raw CRPS. Indeed,” by “same results as in Figure 12 for the raw CRPS, because”.ok
32. Line 342: The bc-ECC method itself does not reduce time penalties because it does not involve temporal aggregation, but I wonder why time aggregation does not have more or less the same effect on the raw and post-processed precipitation forecasts in terms of CRPS? For raw daily rainfall amounts, an error in the timing of the shower is made up by the lead time one (or several) hour(s) before or after. Post-processed hourly rainfall amounts are independent (each lead time is post-processed separately). A sentence has been added: **Due to**

the nature of daily precipitation distribution compared to hourly ones (fewer zeros, smaller variance and lighter tail behavior), we believe that direct post-processing of daily precipitation is more effective if the target variable is daily precipitation.

33. Line 353: Please choose either allow or enable. ok

34. Line 357: Please add “for bc-ECC” after “3 minutes”.ok

35. Line 362: Please replace “substantial save space” by “to save space substantially”.ok

From research to applications - Examples of operational ensemble post-processing in France using machine learning

Maxime Taillardat^{1,2} and Olivier Mestre^{1,2}

¹Météo-France, Toulouse, France

²CNRM UMR 3589, Toulouse, France

Correspondence: Maxime Taillardat (maxime.taillardat@meteo.fr)

Abstract. Statistical post-processing of ensemble forecasts, from simple linear regressions to more sophisticated techniques, is now a well-known procedure for correcting biased and poorly dispersed ensemble weather predictions. However, practical applications in national weather services are still in their infancy compared to deterministic post-processing. This paper presents two different applications of ensemble post-processing using machine learning at an industrial scale. The first is a station-based post-processing of surface temperature and subsequent interpolation to a grid in a medium-resolution ensemble system. The second is a gridded post-processing of hourly rainfall amounts in a high-resolution ensemble prediction system. The techniques used rely on quantile regression forests (QRF) and ensemble copula coupling (ECC), chosen for their robustness and simplicity of training regardless of the variable subject to calibration.

Moreover, some variants of classical techniques used, such as QRF and ECC, were developed in order to adjust to operational constraints. A forecast anomaly-based QRF is used for temperature for a better prediction of cold and heat waves. A variant of ECC for hourly rainfall was built, accounting for more realistic longer rainfall accumulations. We show that both forecast quality and forecast value are improved compared to the raw ensemble. Finally, comments about model size and computation time are made.

1 Introduction

Ensemble Prediction Systems (EPS) are now well-established tools that enable the uncertainty of Numerical Weather Prediction (NWP) models to be estimated. They can provide a useful complement to deterministic forecasts. As recalled by numerous authors (see e.g. Hagedorn et al., 2012; Baran and Lerch, 2018), ensemble forecasts tend to be biased and underdispersed for surface variables such as temperature, wind speed and rainfall. In order to settle bias and poor dispersion, ensemble forecasts need to be post-processed (Hamill, 2018).

Numerous statistical ensemble post-processing techniques are proposed in the literature and show their benefits in terms of predictive performance. A recent review is available in Vannitsem et al. (2018). However, the deployment of such techniques in operational post-processing suites is still in its infancy compared to deterministic post-processing. A relatively recent review of operational post-processing chains in European National Weather Services (NWS) can be found in Gneiting (2014).

25 NWS data-science teams have investigated the field of ensemble post-processing with different and complementary techniques, according to their computational abilities, NWP models to correct, data policy, and their forecast users and targets, see e.g. Schmeits and Kok (2010); Bremnes (2019); Gascón et al. (2019); Van Schaeybroeck and Vannitsem (2015); Dabernig et al. (2017); Hemri et al. (2016); Scheuerer and Hamill (2018). The transition from calibrated distributions to physically coherent ensemble members has also been examined using the Ensemble Copula Coupling (ECC) technique and its derivations, explained in Ben Bouallègue et al. (2016), or variants of the Shaake Shuffle, presented in Scheuerer et al. (2017).

30

Regarding statistical post-processing for temperatures, a recent non-parametric technique such as Quantile Regression Forests (QRF Taillardat et al., 2016) has shown its efficiency in terms of both global performance and value. Indeed, this method is able to generate any type of distribution because assumptions on the variable subject to calibration are not required. Moreover, this technique selects, by itself, the most useful predictors for performing calibration. Recently, Rasp and Lerch (2018) have used QRF as one of the benchmark post-processing techniques.

35 For trickier variables where the choice of a conditional distribution is less obvious, such as rainfall, van Straaten et al. (2018) have successfully applied QRF for 3-h rainfall accumulations. The QRF approach has recently been diversified, both for parameter estimation (Schlosser et al., 2019) and for a better consideration of theoretical quantiles (Athey et al., 2019). In the same vein, Taillardat et al. (2019) have shown that the adjunction of a flexible parametric distribution, an Extended Pareto Distribution (EGP), built on the QRF outputs (named QRF EGP TAIL) compares favorably with state-of-the-art techniques and provides an added value for heavy 6-h rainfall amounts.

In this paper, we present two examples of deployment of ensemble post-processing in the French NWS operational forecasting chain in order to provide gridded post-processed fields. The two examples are complementary:

- 45
- A station-based calibration using local QRF of surface temperature on western Europe of the ARPEGE global EPS (Descamps et al., 2015), associated with an interpolation step and a classical application of ECC.
 - A grid-based calibration using QRF EGP TAIL of hourly rainfall on France of the high resolution AROME EPS (Bouttier et al., 2016) using radar data (calibrated with rain gauges), with a derivation of the ECC technique developed for our application.

50 We also expose some derivations of QRF, QRF EGP TAIL, and ECC techniques in order to take into account extremes prediction, neighborhood management and weather variable peculiarities.

This paper is organized as follows: Section 2 and 3 are devoted, respectively, to the complete post-processing chain of surface temperature and hourly rainfall, exposed in both flowcharts 1 and 2. For each Section, a first subsection describes the EPS subject to post-processing and its operational configuration. We also describe the predictors involved in post-processing procedures. The second subsection comprises a short explanation of QRF or QRF EGP TAIL technique, particularly their adjustments set up for an operational and robust post-processing. The third subsection introduces the post-processing "after

post-processing" work: for the post-processing of post-processed temperatures, we exhibit the algorithm of interpolation and downscaling of scattered predictive distributions. For rainfall intensities, a variant of the ECC technique is presented. The last subsection describes the evaluation of post-processing techniques through both global predictive performance and/or a day-to-day case study. A discussion and our conclusions are presented in Sections 4 and 5.

2 Surface temperature

We present here the French global NWP model ARPEGE, for temperature calibration.

2.1 ARPEGE and ARPEGE EPS

The ARPEGE NWP model (Courtier et al., 1991) has been in use since 1994. Its 35-member EPS, called PEARP, has been in use since 2004, and a complete description is available in Descamps et al. (2015). These global models have been drastically improved throughout the years and their respective grid scale on western Europe is 5km for ARPEGE and 7.5km for PEARP ; forecasts are made 4 times per day from 0 to 108h every 3h. Calibration is performed on more than 2000 stations across western Europe, see Figure 3 for the localization of these stations on our target grid (called EURW1S100). The gridded data is bilinearly interpolated on the observation locations. The data spans 2 years from September 1st, 2015 to August 31st, 2017. The variables involved in the calibration algorithm are provided in Table 1. Operational calibration is currently performed for 2 initialisations only (6 and 18 UTC). Moreover, predictors from the deterministic ARPEGE model are available up to the lead time 60h (except total surface irradiation predictors, which are available from 60h to 78h every 6h).

We can assume that this dataset is less abundant than in Taillardat et al. (2016). This is mainly due to the number of stations covered and the target grid after interpolation, the kilometric AROME grid on western Europe (EURW1S100), which is composed of more than 4 million grid points. Since the principle of statistical post-processing is to build a statistical model linking observations and NWP outputs, two strategies can be considered: the first one is to build a gridded observation archive on the target grid, using scattered station data and a spatialization technique, and to estimate statistical models for each gridpoint or each group of gridpoints (block-MOS technique, Zamo et al., 2016). But although the block-MOS technique is efficient when dealing with deterministic outputs, preliminary tests (not shown here) are inconclusive regarding post-processing of ensembles. Furthermore, estimating a QRF model for each grid point and lead time is not adapted to the operational use, since it would involve a prohibitive size of constants (around 4 Terabytes in this case) to load and store into memory. The alternative strategy is the following: perform calibration on station data and use a quick spatialization algorithm, very similar in its principle to regression kriging, in order to produce quantiles on the whole grid. The computation of calibrated members involves an ECC phase and the same spatialization algorithm.

2.2 QRF calibration technique

Based on the work of Meinshausen (2006), QRF rely on building random forests from binary decision trees, in our case the classification and regression trees from Breiman et al. (1984). A tree iteratively partitions the training data into two groups. A

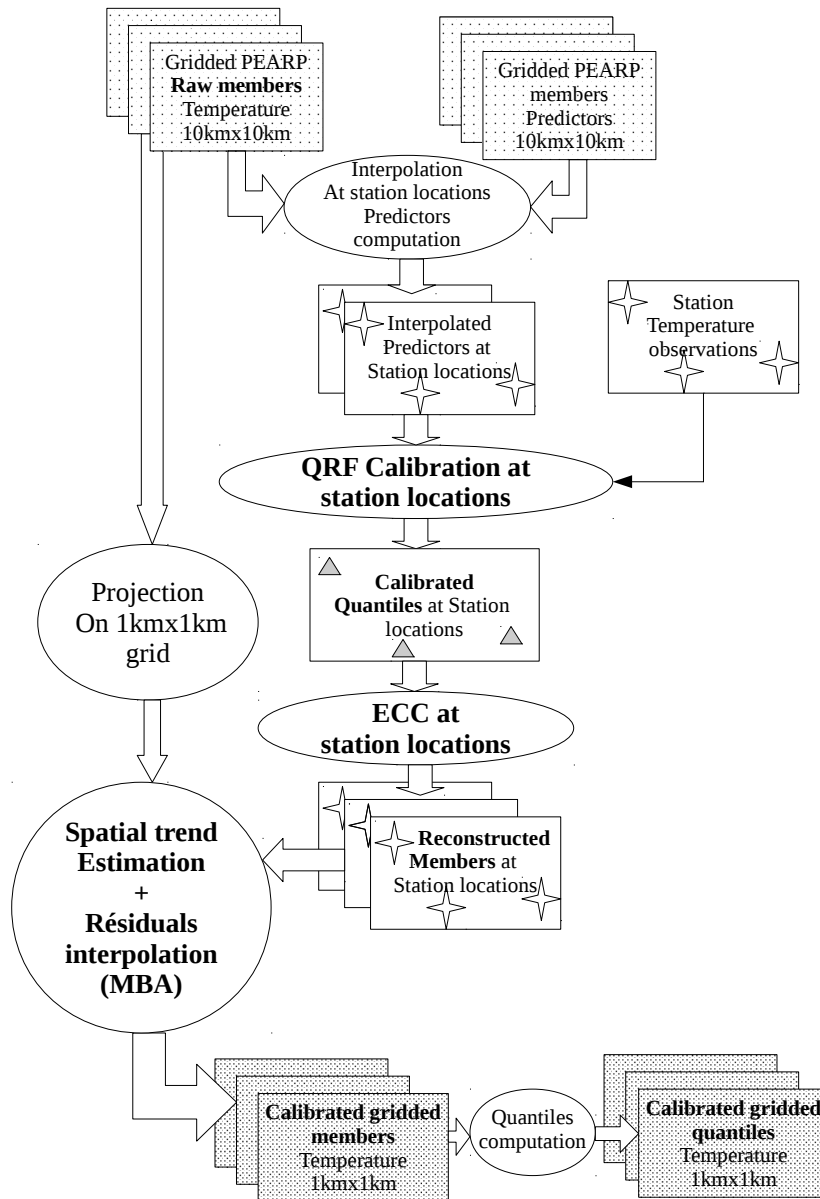


Figure 1. Flowchart of the temperature post-processing chain.

split is made according to thresholds for one of the predictors (or according to some set of factors for qualitative predictors)
 90 and chosen such that the sum of the variance of the two subgroups is minimized. This procedure is repeated until a stopping

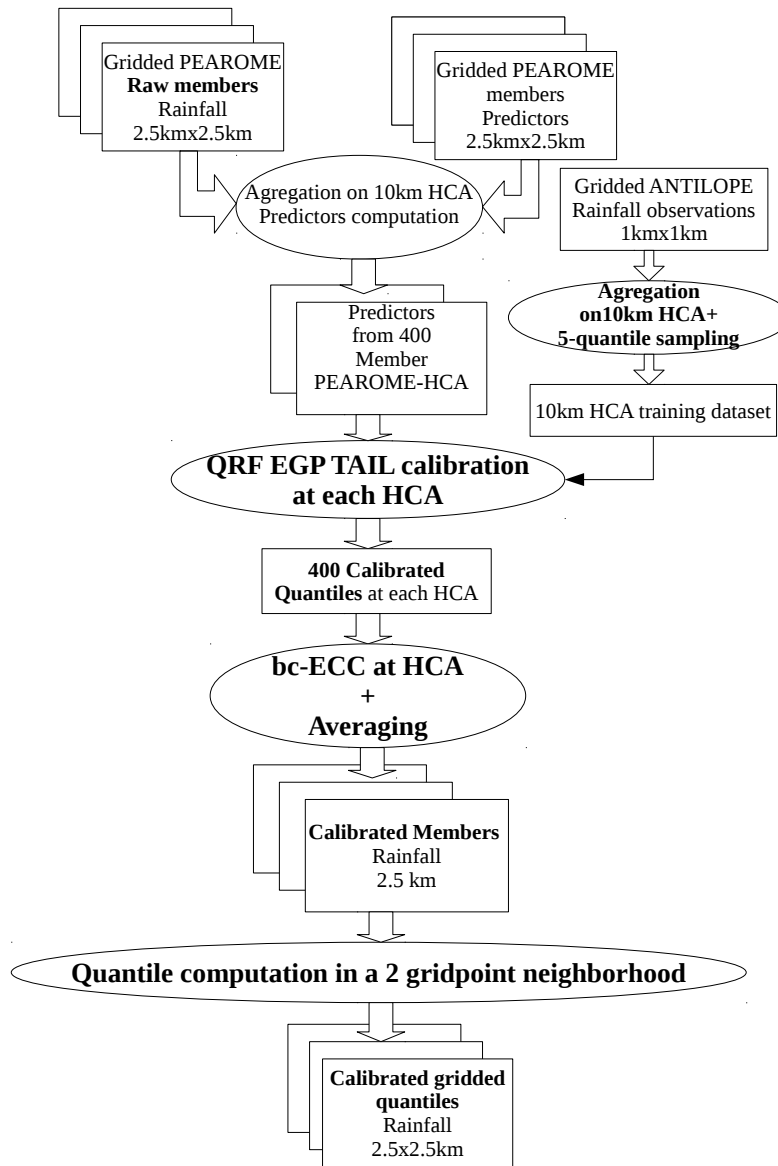


Figure 2. Flowchart of the hourly rainfall post-processing chain.

criterion is reached. The final group (called "leaf") contains training observations with similar predictors values. An example of a tree with 4 leaves is provided at the top of Figure 4.

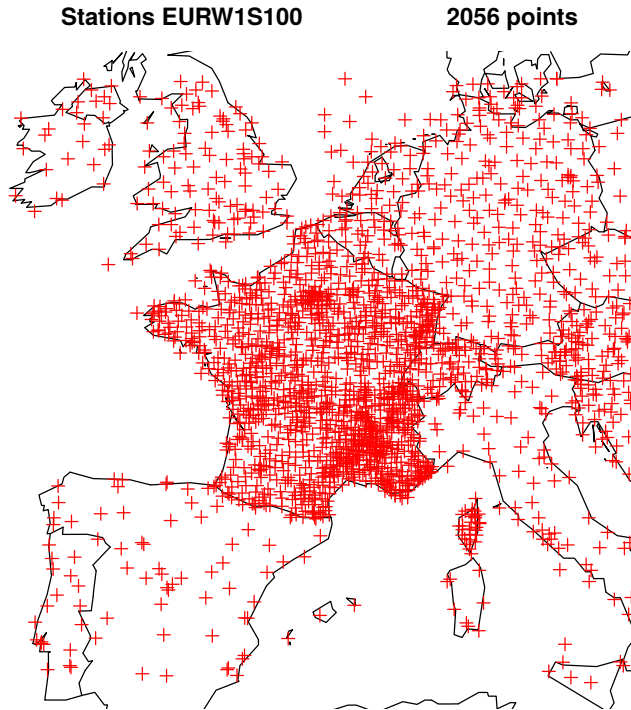


Figure 3. Localization of stations on the target grid.

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. In random forests, Breiman (2001) solves this issue by averaging over many trees elaborated from a bootstrap sample of the training dataset. Moreover, each split is determined on a random subset of the predictors.

When a new set of predictors \mathbf{x} is available (the blue cross in Figure 4), the conditional Cumulative Distribution Function (CDF) is made by the observations Y_i corresponding to the leaves to which the values of \mathbf{x} lead in each tree. The predicted CDF is thus

$$100 \quad \hat{F}(y|\mathbf{x}) = \sum_{i=1}^n \omega_i(\mathbf{x}) \mathbf{1}(\{Y_i \leq y\}), \quad (1)$$

where the weights $\omega_i(\mathbf{x})$ are deduced from the presence of Y_i in a final leaf of each tree when following the path of \mathbf{x} .

See, for example, Taillardat et al. (2016, 2019); Rasp and Lerch (2018); Whan and Schmeits (2018) for detailed explanations and comparisons with other techniques in a post-processing context.

2.3 Operational adjustments for temperature

105 A direct application of the QRF algorithm for forecasting temperature distribution is suboptimal. Indeed, although QRF is able to return weather-related features such as multi-modalities, alternatives scenarios, and skewed distributions, the method cannot

Table 1. Predictors involved in station-based PEARP post-processing. The target variable is surface temperature.

From the ARPEGE model (and up to lead time 60h/78h for irradiation predictors):
surface temperature
vertical gradient of temperature between surface and 100m
surface temperature 3h trend
zonal gradient of surface temperature
meridian gradient of surface temperature
850hPa potential wet-bulb temperature
surface wind speed
surface wind direction (factor)
sea level pressure
mean (on 4 grid point squares) of total cloud cover
mean (on 4 grid point squares) of low level cloud cover
surface relative humidity
accumulated snow depth on ground
3h total surface irradiation in infrared wavelengths
3h total surface irradiation in visible wavelengths
From the PEARP (ARPEGE EPS) model:
mean of surface temperature
median of surface temperature
minimum of surface temperature
maximum of surface temperature
second decile of surface temperature
eighth decile of surface temperature
freezing probability
Others:
month of the year (factor)

110 go beyond the range of the data. In the operational chain, the QRF algorithm is not trained with observations but with the errors between the observation and the ensemble forecast mean. The result of equation 1 is, in this case, the error distribution before translation around the raw ensemble mean. The predictive distributions are now constrained by the range of errors made by the ensemble mean. This anomaly-QRF approach generates better distributions than QRF for the prediction of cold and heat waves, and leads to an improvement of about 7% (not shown here) in the averaged Continuous Ranked Probability Score (CRPS ; Gneiting and Raftery, 2007), thanks to this NWP-dependent variable response.

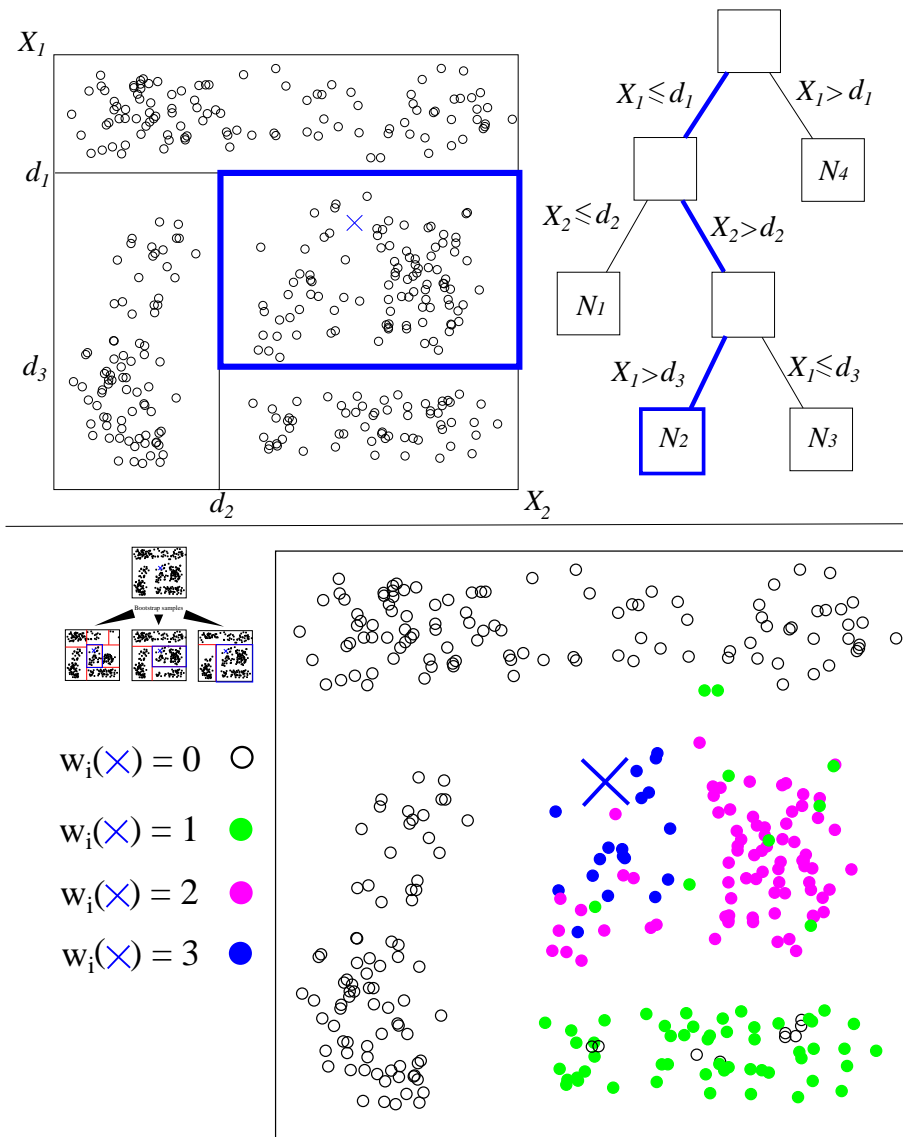


Figure 4. Two-dimensional example of a (top) binary regression tree and (bottom) three-tree forest. A binary decision tree is built from a bootstrap sample of the data at hand. Successive dichotomies (lines splitting the plane) are made according to a criterion based on observations' homogeneity. For a new set of predictors (the blue cross), the path leading to the corresponding observations is followed. The predicted CDF is the aggregation of the results from each tree

2.4 Ensemble Copula Coupling

The Ensemble Copula Coupling method (Scheffzik et al., 2013) provides spatiotemporal joint distributions derived from the raw ensemble structure. Its small computational cost makes it, for us, the preferred way to reorder calibrated marginal distributions,

even if other techniques, such as Schaake Shuffle, have their advantages (Clark et al., 2004). Therefore, we make the assumption that on HCA, the structure of the raw ensemble is temporally and spatially sound. Recently, Ben Bouallègue et al. (2016) and Scheuerer and Hamill (2018) proposed an improvement of the ECC technique using, respectively, past observations and simulations. In the context of hourly quantities in hydrology, Bellier et al. (2018) show that perturbations added to the raw ensemble lead to satisfactory multivariate scenarios.

2.5 Interpolation of scattered post-processed temperature

2.5.1 Principle

The problem at hand is challenging:

- The domain covers a large part of western Europe, from coastal regions to Alpine mountainous regions, subject to various climate conditions (oceanic, Mediterranean, continental, Alpine).
- Data density is very inhomogeneous (from the high density of stations over France to the somewhat dense network over the UK, Germany, and Switzerland and the sparse density over Spain and Italy).
- Interpolation has to be extremely fast, since more than 1824 high resolution spatial fields have to be produced in a very short time.

Common methods used to interpolate meteorological variables include Inverse Distance Weighting (IDW; Zimmerman et al., 1999), and the Thin Plate Splines (TPS; Franke, 1982) - both considered deterministic methods, and kriging (Cressie, 1988), including kriging with external drift to take topography effects into account (Hudson and Wackernagel, 1994). But while IDW suffers from several shortcomings such as cusps, corners, and flat spots at the data points, preliminary tests showed that both TPS and kriging did not satisfy computation time requirements.

Therefore, a new technique has been developed, very similar to "regression-kriging", based on the following principle: at each station location, perform a regression between post-processed temperatures and raw NWP temperatures, using additional gridded predictors as well. The resulting equation is then applied to the whole grid to produce a spatial trend estimation. Regression residuals at station locations are then interpolated. Spatial trend and interpolated residuals are summed to produce the resulting field. Interpolation of residual fields is performed using an automated Multi-level B-splines Analysis (MBA; Lee et al., 1997), an extremely fast and efficient algorithm for the interpolation of scattered data.

2.5.2 Spatial trend estimation

Several studies have investigated the complex relationships between topography and meteorological parameters ; see e. g. Whiteman (2000); Barry (2008). A naive model would be a linear decrease of temperatures with altitude, which is not realistic for temperature at the daily or hourly scale, since the vertical profile may be very different from the profile of free air temper-

ature. An important phenomenon, which has often been studied and subject to model is cold air pooling in valleys with the diurnal cycle. Frei (2014) uses a change-point model to describe non linear behaviour of temperature profiles.

Topographical parameters include altitude, distance to coast and additional parameters computed following the AURELHY method (Bénichou, 1994). The AURELHY method is based on a Principal Component Analysis (PCA) of altitudes. For each
150 point of the target grid, 49 neighboring gridpoint altitudes are selected, forming a vector called a landscape. The matrix of landscapes is processed through a PCA. We determine that this method efficiently summarizes topography, since first principal components can easily be interpreted in terms of peak/depression effect (PC1), northern/southern slope (PC2), eastern/western slopes (PC3) or "saddle effect" (PC4). These AURELHY parameters are presented in Figure 5.

For the interpolation of climate data, most of the time only topographic data is available, which may play the role of ancillary
155 data in estimating the spatial trend. In our case, another important source of information is provided by the NWP temperature field at corresponding lead time for each member. As such, PEARP data may not be directly used, since its resolution is coarser than the target resolution (7.5km rather than 1km). Therefore, PEARP data are projected on the target grid using the following procedure: for each of 7.5km gridpoint, a linear transfer function is estimated through a simple linear regression between each of the 100 AROME temperature data points (available on the 1km resolution grid) and the corresponding ARPEGE data point.
160 **Since this relationship is likely to change over seasons and time of day, these regressions are computed seasonally, and for every hour of the day, using one year of data.** This is a crude but quick way to perform downscaling of PEARP data, as will be shown later.

Since interpolation is to be performed on a very large domain, with greatly varying data density, several regressions are
165 computed on smaller sub-domains denoted by D , whose boundaries are given in Figure 6. Note that the size of domains depends on the stations' spatial density. Further, domains overlap: at their intersection, spatial trends are averaged, and weights add up to one and are a linear function of inverse distance to domain frontier. This simple algorithm is very efficient in eliminating any discontinuity between adjacent domains that might appear otherwise.

For a given basetime b and leadtime t , validity time is denoted as v , and season is denoted as S .

170 We denote alti_i , (resp. $\text{d2s}_i, \text{PC1}_i, \text{PC2}_i, \text{PC3}_i, \text{PC4}_i$) values of altitude (resp. distance to sea, and principal component of elevation 1 to 4) at gridpoint i of the target grid. For every basetime b and leadtime t , let T_k be the calibrated temperature forecast of the k th station point of subdomain D , corresponding to gridpoint i of the target grid ($0.01^\circ \times 0.01^\circ$) and gridpoint j of PEARP $0.1^\circ \times 0.1^\circ$ grid, and T_j be the corresponding raw PEARP temperature forecast (same member, base time and lead time as T_k) at the gridpoint j . Then:

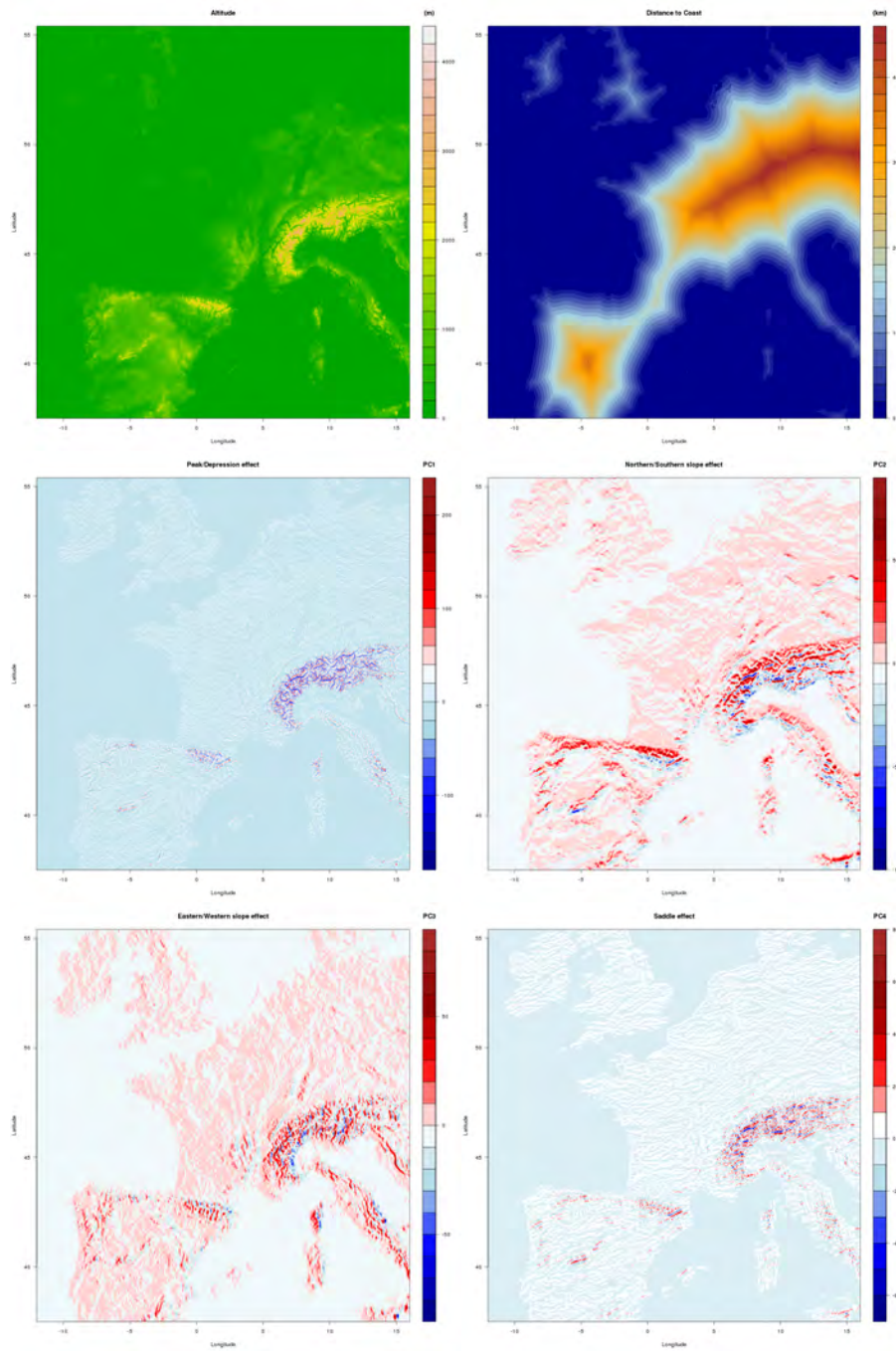


Figure 5. Altitude (upper left panel), distance to sea (upper right panel), and PC1 to PC4 (from middle left panel to lower right panel).

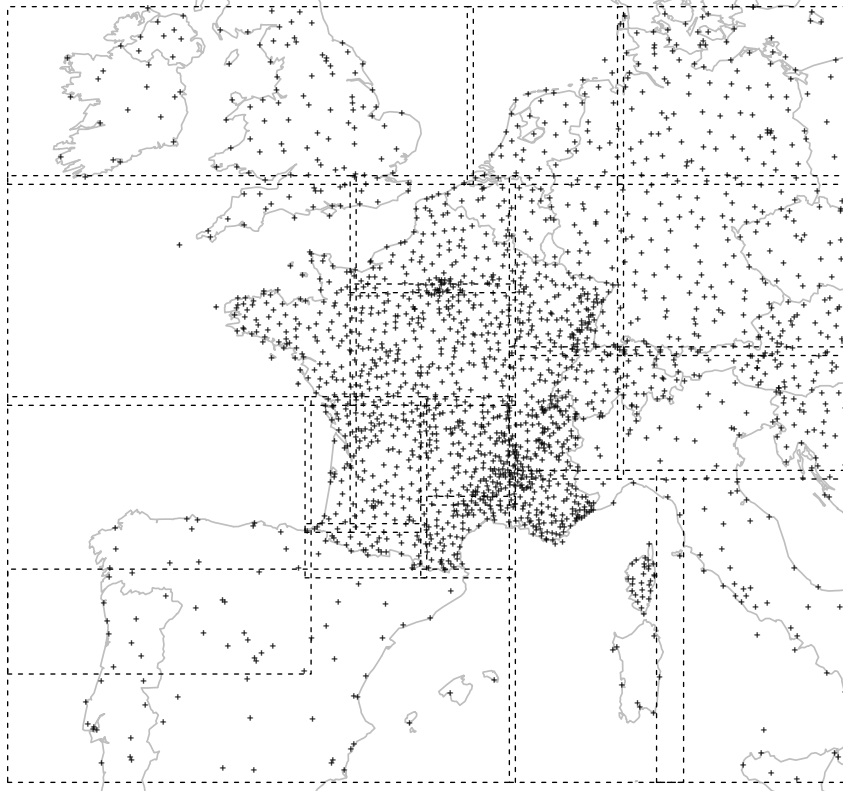


Figure 6. Domains used for spatialisation of post-processed temperatures.

$$175 \quad T_k = \beta_{0_D} + \beta_{1_D}(\gamma_{0_{i,jvS}} + \gamma_{1_{i,jvS}}T_j) \quad (2)$$

$$+ \beta_{2_D}\text{alti}_i + \beta_{3_D}(\text{alti}_i - a*_{D})1\{\text{alti}_i > a*_{D}\} \quad (3)$$

$$+ \beta_{4_D}d2s_i \quad (4)$$

$$+ \alpha_{1_D}PC1_i + \alpha_{2_D}PC2_i + \alpha_{3_D}PC3_i + \alpha_{4_D}PC4_i \quad (5)$$

$$+ \epsilon_k \quad (6)$$

180 The term (2) corresponds to the linear influence of the linear projection function of T_j on target gridpoint i . The term (3) corresponds to the altitude effect, with a possible change in slope of the vertical temperature gradient at altitude $a*_{D}$, the value of which is tested on a grid of ten specified elevations for each domain D . The term (4) is the influence of distance to sea. Term (5) is related to the first four Principal Components of elevation landscapes. The last term ϵ_k is the regression residual.

The distance to sea predictor appears only for domains including seashores. Furthermore, domains containing too few station
185 points, namely Spanish and Italian domains, have only one predictor, which is a linear projection of PEARP temperature data:
 $\gamma_{0,jvS} + \gamma_{1,jvS}T_j$.

The model estimation of parameters $\beta_{0D}, \beta_{1D}, \beta_{2D}, \beta_{3D}, \beta_{4D}, \alpha_{1D}, \alpha_{2D}, \alpha_{3D}, \alpha_{4D}$, and $a*D$ is performed by means of ordinary least squares, with the model selection automatically ensured by an Akaike Information Criterion (AIC) procedure. **This model selection is influenced by the weather situation, but the most often selected variables are the linear projection function of**
190 **T_j and/or the altitude effect – since they are very well correlated. Distance to sea and PCI may also be selected quite frequently. PC2 to PC4 are selected much less frequently.**

2.5.3 Residuals interpolation

We aim to use an exact, automatic and fast interpolation method for residual interpolation. Although TPS and kriging may be computed in an automated way, those methods do not meet our criteria in terms of computation time.

195 While not strictly an exact interpolation method, the MBA algorithm was chosen as it is an extremely fast algorithm. Furthermore, the degree of smoothness and exactness of the method may be precisely controlled, as recalled by Saveliev et al. (2005).

A precise description of this method is beyond the scope of this article. We just briefly recall that the MBA algorithm relies on a uniform bicubic B-Spline surface passing through the set of scattered data to be interpolated. This surface is defined by
200 a control lattice containing weights related to B-spline basis functions, the sum of which allows surface approximation. Since there is a tradeoff between smoothness and accuracy of approximation via B-Splines, MBA takes advantage of a multiresolution algorithm. MBA uses a hierarchy of control lattices, from coarser to finer, to estimate a sequence of B-splines approximations whose sum achieves the expected degree of smoothness and accuracy. Refer to Lee et al. (1997) for a complete description of the algorithm.

205 During testing, we found out that 13 approximations were sufficient to ensure a quasi-exact interpolation (magnitude of error, around 0.0001°C at station locations), for a visual rendering extremely similar to interpolation TPS, at the cost of a small and acceptable computing time. The solution with 12 approximations was discarded, as it was not precise enough (magnitude of error, around 0.3°C at station locations) meaning that interpolation could no longer be considered to be exact. When using 14 approximations, computation time dramatically increased.

210 An important point at the practical level is that the interpolation of residuals is performed only once on the whole grid. We found that undesirable boundary effects could appear at the edges of domains D when residuals were interpolated at each domain D alone.

2.6 Results for temperature post-processing chain

2.6.1 Results of station-wise calibration

215 We present here the results of the post-processing of PEARP temperature in EURW1S100 stations. The hyperparameters for QRF are derived from Taillardat et al. (2016) but with a smaller number of trees (200). The validation is made by a 2-fold cross-validation on the two years of data (one sample per year). For each base and lead time, Figure 7 shows the averaged CRPS in the top panel and the PIT statistic mean and $12\times$ variance in the bottom panels. These statistics represent the bias and dispersion of the rank histograms (Gneiting and Katzfuss, 2014; Taillardat et al., 2016). Subject to probabilistic calibration, the
220 mean of the statistic should be 0.5 and the variance $1/12$, which implies the flatness of rank histograms.

The gain in CRPS is obvious after calibration, whatever the base and lead times. Moreover, the hierarchy among base times is maintained. In both bottom panels, post-processed ensembles are unbiased and well dispersed, contrary to raw ensembles, which exhibit (cold with diurnal cycle) bias and under-dispersion. Nevertheless, we notice that post-processed distributions show a slight under-dispersion at the end of lead times. This is due to the absence of predictors coming from the deterministic
225 ARPEGE model. These predictors do not relate directly to temperature, and thus the addition of weather-related predictors is crucial here for uncertainty accounting. We believe that radiation predictors are most important here, since the presence or absence of these predictors is linked to the "roller coaster" behavior of post-processed PIT dispersion around a 3-day lead time.

2.6.2 Performance of interpolation algorithm

Prior to any use in the spatialization of post-processed PEARP fields, performances of the interpolation method were evaluated
230 for deterministic forecasts.

This paragraph is devoted to the evaluation of an earlier version of the current spatialization algorithm over France, which differs only in the fact that NWP temperature fields are not available in the predictor set for spatial trend estimation. Benchmarking data consists of 100 forecasts. For each date, 20 cross validation samples are randomly generated, removing 40 points from the full set of points. Original forecast values and interpolated forecast values are then compared, and standard scores
235 (bias, Root Mean Square Error, Mean Absolute Error, 0.95 quantile of absolute error) are computed. Scores are then compared to the COSYPROD interpolation scheme, the previous operational interpolation method. COSYPROD is a quick interpolation scheme which predates both the first and current versions of our algorithm, adapted to interpolation at a set of some production points, and derived from the IDW method.

Results show that regardless of the method, bias remains low, but the new spatialization method outperforms COSYPROD
240 in terms of RMSE, MAE, and .95 quantile of absolute error (Figure 8).

Additionally, the described spatialization procedure is already used operationally for the interpolation of deterministic temperature forecasts since May 2018. In this application, its performances were evaluated routinely over a large set of climatological station data, which only measures extreme temperatures and do not provide real time data. Hence, this dataset is discarded from any post-processing, but may serve as an independent dataset for validation. When comparing forecast performances
245 related to this dataset, the increase in root mean square error is around 0.3°C compared to forecast errors estimated at post-

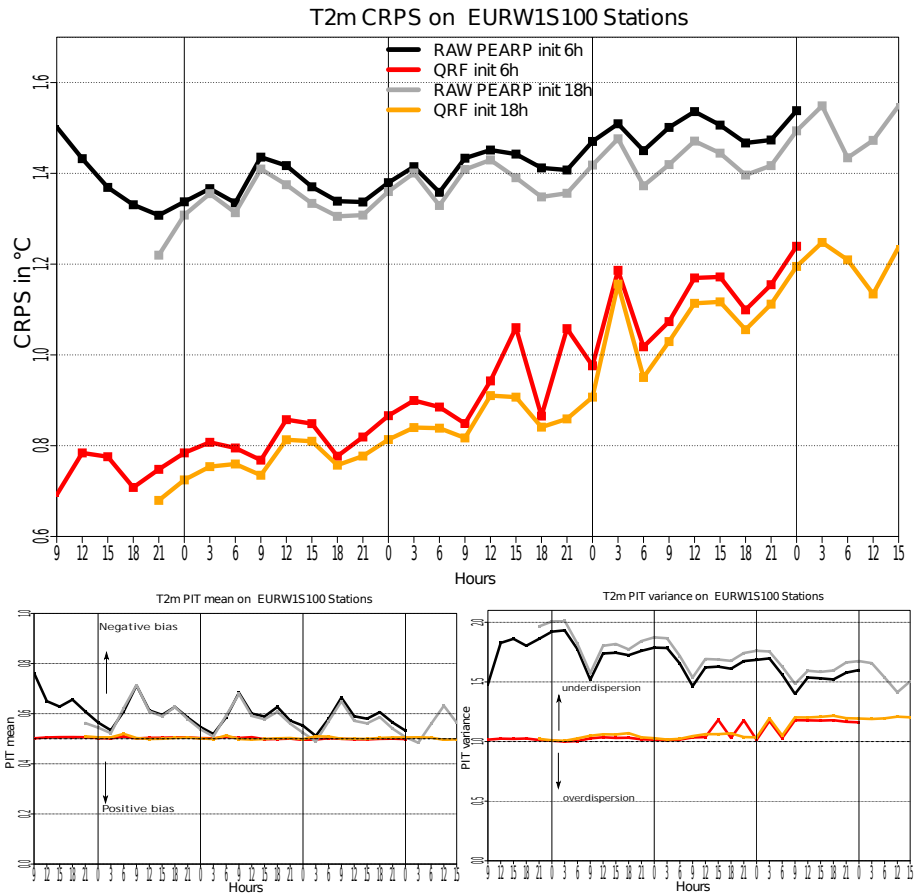


Figure 7. Results of PEARP post-processing of temperature in the 2056 EURW1S100 stations with averages CRPS (top), and mean and variance of PIT statistic, related to rank histograms. The validation is made by a 2-fold cross-validation on the two years of data (one sample per year).

processed station data. Hence, this extra 0.3°C root mean square error may be considered as an error due to the interpolation process. Note that this is much lower than what was estimated during the cross-validation phase: all in all, forecast errors and interpolation errors are not added together, but compensate each other to some extent.

An illustration of the whole procedure is illustrated on PEARP temperatures of base time 10/03/2019, 18UTC, for lead time 42h. The temperature field of raw member 16 is presented in Figure 9, together with the same field projected on the EURW1S100 grid. The estimated spatial trend is also shown, and residuals are interpolated using the MBA procedure with 13 approximation layers can also be found in Figure 9. The resulting field, after calibration, ECC, and spatial interpolation phases is presented in Figure 10. The same process is repeated here for member 6 (Figure 11).

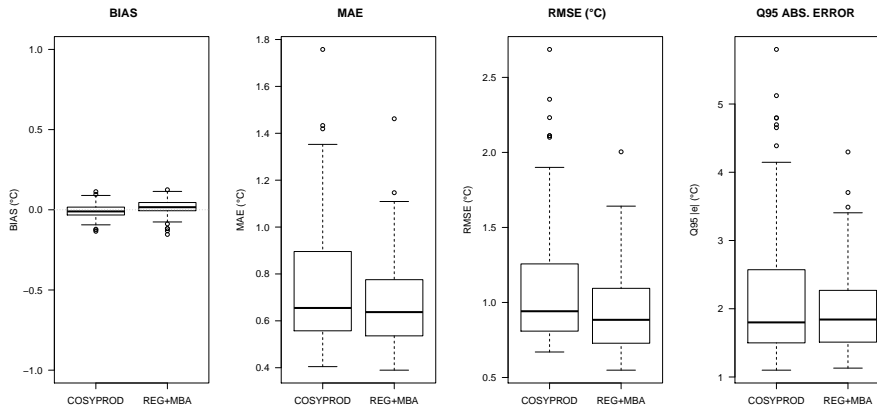


Figure 8. Boxplots of Bias, Mean Absolute Error, Root Mean Square Error and 0.95 Quantile of Absolute Error for COSYPROD (left boxplot) and the new method (right boxplot).

Note that during the full processing, field values were modified during the calibration process. But ECC and interpolation are able to maintain the main features of the original field, i.e. is passage of a front, which is not situated in the same location for both members.

3 Hourly rainfall

We present the high-resolution limited model area NWP model AROME, for the post-processing of hourly rainfall.

3.1 AROME and AROME EPS

The non hydrostatic NWP model AROME (Seity et al., 2011) has been in use since 2007 on the limited area of Figure 3. The associated 16-member EPS, called PEAROME (Bouttier et al., 2016), has been in operational use since the end of 2016. The deterministic model operates on the 1km EURW1S100 grid whereas the PEAROME runs on a 2.5km grid. Forecasts are made 4 times a day from 0 to 54h. Data spans 2 years from December 1st, 2016 to December 31st, 2018. Calibration is not performed on the 2.5km grid, but on a 10km grid. Thus we consider PEAROME here as a $16 \times 5 \times 5 = 400$ -member pseudo-ensemble on a 10km grid. We do this for 3 reasons:

- We reduce spatial penalties issues due to the high resolution of the raw EPS (see e.g. Stein and Stoop, 2019).
- We improve ensemble sampling and, we hope, the quality of predictors.
- We reduce computational costs by a factor 25.

The post-processing is conducted on these 10km "homogeneity calibration area" (HCA) grid points using ANTILOPE (Laurantin, 2008), the 1km-gridded French radar data set calibrated (with rain gauges). Predictors involved in the calibration algo-

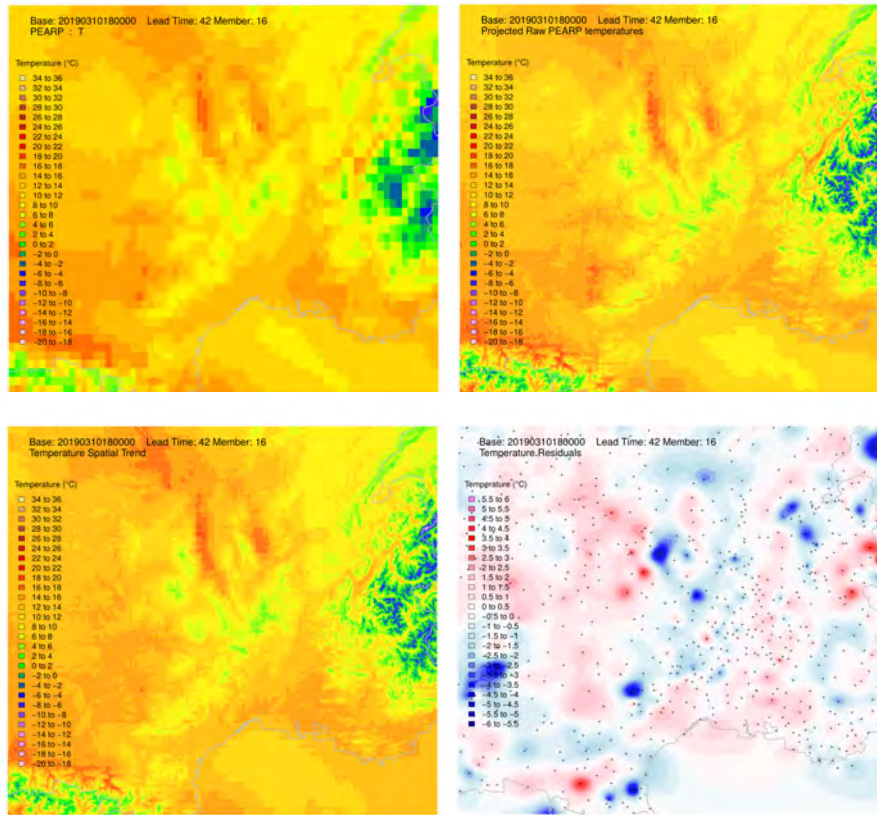


Figure 9. Step-by-step procedure illustrated over the southeast of France: raw member temperatures on 7.5km grid (upper left panel), raw projected temperatures on a 1km grid (upper right panel), spatial trend estimation using regression model on subdomains (lower left panel), field of residuals interpolated using a MBA procedure with 13 layers of approximation (lower right panel).

rithm are listed in Table 2. Note that the temporal penalties due to the high resolution are considered in this choice of predictors. Operational calibration is currently performed for two initialisations only (9 and 21 UTC) and for lead times up to 45h.

The number of predictors is less abundant here than in Taillardat et al. (2019). This number was reduced to 25 due to operational constraints on model size. These predictors were chosen after a variable selection step using VSURF (Genuer et al., 2015) and the R (R Core Team, 2015) package randomForestExplainer (Paluszynska, 2017) among more than 50 potential predictors. A complete description of the variable selection is beyond the scope of the paper. To summarize, the most important variables are, on average, minimal and maximal rainfall intensities. These variables are followed by "synoptic" variables such as wind or humidity at medium-level and potential wet-bulb temperature. ICA is roughly the product of the modified Jefferson index (Peppier, 1988) with a maximum between 950hPa convergence and maximal vertical velocity between 400 and 600hPa. This latter variable and other variables representing the shape of the raw distribution of precipitation are less decisive on

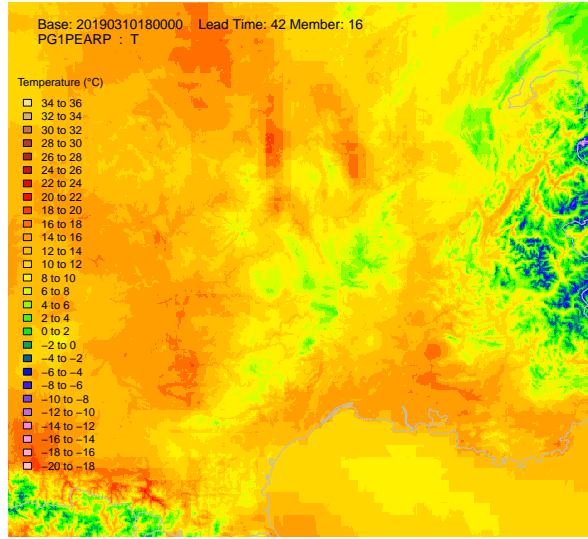


Figure 10. Resulting field over the southeast of France.

average. Variables not retained in the selection procedure are redundant with the main predictors, such as other convection indices, medium-level geopotential, and low-level cloud cover, and surface variables. For each of the 13900 HCA, the quantile regression algorithm gives 400 quantiles, attributed to each member of each grid point of the HCA after a derivation of ECC technique. The value of grid points and members overlapping two or more HCA are averaged.

285 3.2 QRF EGP TAIL calibration technique

Note in equation 1 that the QRF method cannot predict values outside the range of the training observations. For applications focusing on extreme or rare events, it could be a strong limitation if the data depth is small. To circumvent this QRF feature, Taillardat et al. (2019) propose to fit a parametric CDF to the observations in the terminal leaves rather than using the empirical CDF in the equation 1. The parametric CDF chosen for this work is the EGPD3 in Papastathopoulos and Tawn (2013) which is
 290 an extension of the Pareto distribution. Naveau et al. (2016) show the ability of this distribution to represent both low, medium and heavy rainfall and its flexibility. Thus, the QRF EGP TAIL predictive distribution is

$$G(y|\mathbf{x}) = P_0 + (1 - P_0) \left[1 - \left(1 + \frac{\xi y}{\sigma} \right)^{-\frac{1}{\xi}} \right]^\kappa, \quad (7)$$

where P_0 is the probability of no rain in the QRF output: $\widehat{F}(y = 0|\mathbf{x})$. The parameters (κ, σ, ξ) in equation 7 are estimated via a robust method-of-moment (Hosking et al., 1985) estimation.

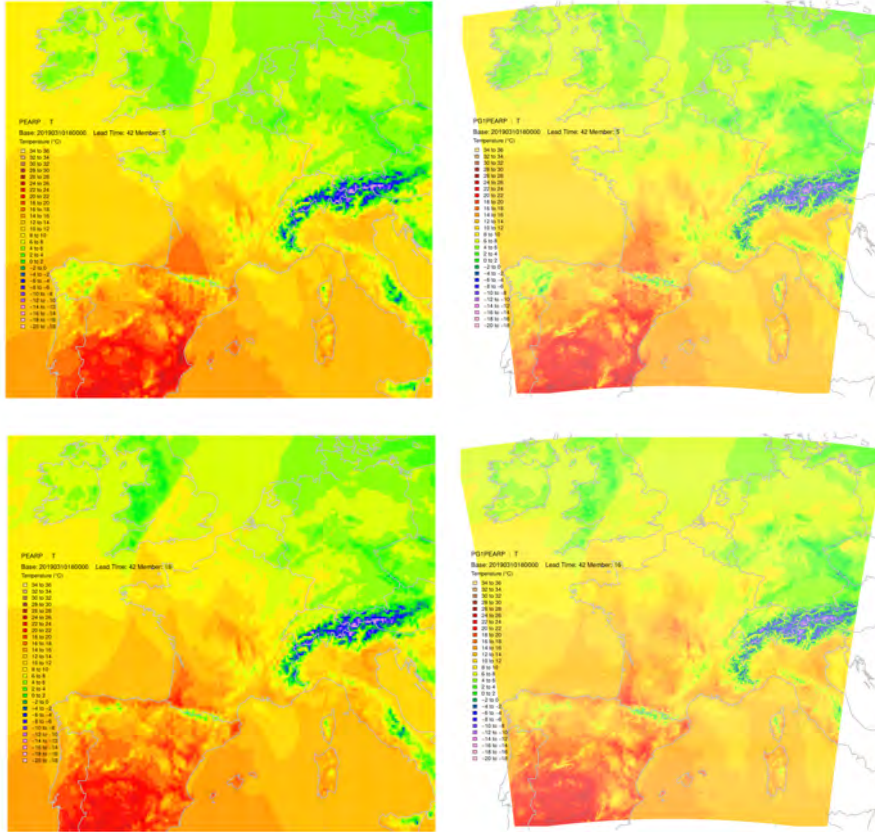


Figure 11. Raw PEARP member 6 temperature field (upper left panel), the same after calibration, ECC and interpolation phase (upper right panel) together with raw (lower left panel) and post-processed temperature field (lower right panel) for member 16. Note that the whole procedure can only be applied to the AROME domain.

295 3.3 Operational adjustments for hourly rainfall

The anomaly-based QRF approach is not employed for hourly rainfall. We believe that the choice of a centering variable is as difficult as choosing a good parametric distribution for predictive distributions. In the case of hourly rainfall, the adjustments are not relative to the method but rather the construction of the training data.

For each HCA, we consider predictors calculated with the 400-member pseudo-ensemble. For each HCA of size $10\text{km} \times 10\text{km}$,
 300 100 ANTILOPE observations are available. We can consider the observation data as coming from a distribution. Practically speaking, instead of having one observation Y_i for each set of predictors, in our case we have $(Y_{i_0}, Y_{i_{25}}, Y_{i_{50}}, Y_{i_{75}}, Y_{i_{100}})$, corresponding to the empirical quantiles of order 0, 0.25, 0.5, 0.75, 1 of ANTILOPE distribution in the HCA. The length of the training sample is inflated by a factor 5, which allows us to take advantage of all the information available instead of upscaling high resolution observation data.

Table 2. Predictors involved in HCA-based PEAROME post-processing. **The target variable is hourly rainfall.**

From the HCA-PEAROME pseudo ensemble:

mean of hourly rainfall
median of hourly rainfall
first decile of hourly rainfall
ninth decile of hourly rainfall
maximum of hourly rainfall
standard deviation of hourly rainfall
probability of rain
probability of rain $> 5mm.h^{-1}$
maximum of hourly rainfall at previous lead time
probability of rain at previous lead time
first decile of maximum radar reflectivity
ninth decile of maximum radar reflectivity
mean of convective available potential energy
mean of 850hPa potential wet-bulb temperature
first decile of 500m relative humidity
ninth decile of 500m relative humidity
first decile of 700hPa relative humidity
ninth decile of 700hPa relative humidity
first decile of total cloud cover
ninth decile of total cloud cover
mean of surface wind gust speed
mean of 700hPa zonal component of wind speed
mean of 700hPa meridian component of wind speed
mean of 700hPa wind speed
mean of ICA (AROME Convection Index)

305 3.4 ECC for rainfall intensities

As already observed by Scheuerer and Hamill (2018); Bellier et al. (2017), ECC has innate issues with an undispersed ensemble and, more precisely, the attribution precipitation to zero raw members (ie. if the calibrated rain probability $\overline{P_0}$ is greater than the raw one $\overline{F_0}$).

310 In our case, 400 values have to be attributed to the 16 members of the 25 grid points of the HCA. The procedure, called bootstrapped-constrained ECC (bc-ECC), is as follows:

- If $\overline{F_0} > \overline{P_0}$, a simple ECC is performed.

- If not, we perform ECC many times (here 250 times per HCA) and average values.
- Then, a raw zero becomes a non-zero only if there is a raw non-zero in a 3 raw grid point neighborhood.

In this case, $b = 250$ and $c = 3$. The Table 3 gives an example of an HCA of 3 grid points and 2 members.

Table 3. Example of bc-ECC ($b = \infty$, $c = 1$) for 2-member (M) ensemble in a 3-grid points (gP) linear HCA.

In the HCA:	gP1M1	gP2M1	gP3M1	gP1M2	gP2M2	gP3M2
Raw values:	2	2	5	0	0	1
HCA Calibrated values:	0	4	5	5	6	7
b-ECC and average:	5.5	5.5	7	2	2	5
Is rain in M in c gP around ?	-	-	-	no	yes	-
Final values:	5.5	5.5	7	0	2	5

315 As a result, in a member, post-processing can introduce rain in a grid box that is dry in a raw member only if there is a grid point with rain close by in the raw member. This approach ensures coherent scenarios between post-processed rainfall fields and raw cloud cover fields, for example.

3.5 General results and day-to-day examples for rainfall

3.5.1 Hourly rainfall calibration

320 Due to the high amount of data to process for evaluation (around 200Gbytes), scores are presented with averaged lead time and for the base time 9UTC only. More precisely, for each lead time evaluation is made on 400 HCAs over 13900. More than 920 sets of hyperparameters for QRF EGP TAIL were tried, and the numbers retained are 1000 for the number of trees, 2 for the predictors to try and 10 for the minimal node size. In order to make the comparison as fair as possible, the predictive distributions are considered on HCAs and the observation is viewed as a distribution (like in the Section 3.2.2). As a consequence, 325 the divergence of the CRPS should be used, but the computation of the CRPS on the observations is equivalent (Salazar et al., 2011; Thorarinsdottir et al., 2013). The validation is made by a 2-fold cross-validation on the two years of data (one sample per year).

The averaged CRPS between the raw and the post-processed ensemble is improved by approximately 30% (from 0.118 to 0.079).

330 Figure 12 focuses on the rain event (more than $0.1mm.h^{-1}$). The top panel shows an ROC curve and a reliability diagram on the same plot. Post-processing improves both the resolution and reliability of predictive distributions for the rain event, overpredicted by the raw ensemble. Overprediction of the raw ensemble is also exhibited in the performance diagram (Roebber, 2009) on the bottom panel. Indeed, there is an asymmetry to the top left corner, where frequency bias is more important. Like the ROC curve, the curve in the performance diagram is computed for each quantile of the forecast. The critical success index

335 is increased by 15%, which means that the ratio of rain events (predicted and/or observed) forecast well is improved by 15%.
Moreover, we can assume here that the minimum (quantile 1/401) of the post-processed distributions nearly never forecasts
rain occurrence, but when it does, a false alarm is never made. The minimum (quantile 1/401) of the raw ensemble detects rain
occurrence around 40 times over 100, but when it does, the forecast is wrong around 35 times over 100 (1-success ratio).

As for increased precipitation, the focus is placed on forecast value. Figure 13 depicts the maximum of the Peirce Skill Score
340 (PSS ; Manzato, 2007) for hourly accumulation thresholds. The maximum of the PSS, which corresponds to the nearest point
of the top left corner in ROC curves, is a good way to summarize forecast value (Taillardat et al., 2019). **More precisely, most
of this improvement is due to the the improvement of the Hit Rate.**

3.5.2 Effects on daily rainfall intensities

Daily rainfall in between raw and post-processed ensembles was compared in the pre-operationnal chain during October 2019.
345 In Figure 14, the CRPS of daily distributions shows that bc-ECC does not deteriorate predictive quality. If we divide by 24, we
do not obtain the same results as for raw hourly CRPS, because time penalties disappear with temporal aggregation of hourly
quantities. The bc-ECC method does not solve temporal penalties. Therefore, it is not surprising that the daily post-processed
CRPS is roughly 24 times the averaged hourly one. Due to the nature of daily precipitation distribution compared to hourly
ones (fewer zeros, smaller variance and lighter tail behavior), **we believe that direct post-processing of daily precipitation is**
350 **more effective if the target variable is daily precipitation.**

We then seek to determine whether calibrated hourly intensities lead to unrealistic or worse daily rainfall intensities than
the raw ensemble. In other words, does the bc-ECC generates coherent scenarios ? First, in Figure 15 we show the comparison
of the predictive quantiles of daily post-processed (after bc-ECC) and raw intensities. The date is 10/22/2019 and related to a
heavy precipitation event in the South of France. 24-h observed accumulations (left of the Figure) reach 300 mm. On the right,
355 the quantiles of order 0.1, 0.5, and 0.9 of the post-processed ensemble (top right) and raw ensemble (bottom right) are presented.
For this event of interest, we see that bc-ECC does not create unrealistic quantities.

4 Discussion

The two applications described in this article (PEARP temperature and PEAROME rainfall post-processing) are extremely
computationally demanding and therefore could not be run on standard workstations within an acceptable timeframe. While
360 codes are implemented on Météo-France's supercomputer, a crucial optimization phase must still be achieved, as two problems
had to be solved during the implementation phase:

- The very large number of high resolution fields required, since for each lead time, not only statistical fields (quantiles,
mean, standard deviation fields), but also calibrated member fields are computed. This was achieved using inexpensive
but efficient methods such as ECC and MBA, and a massive parallelisation of operations, thanks to R High Performance
365 Computing capabilities. The operational code relies on parallel, foreach, DoSNOW, and DoMC packages, that enable
OpenMP multicore and MPI multinodes capabilities. The number of cores used in each node is driven by memory

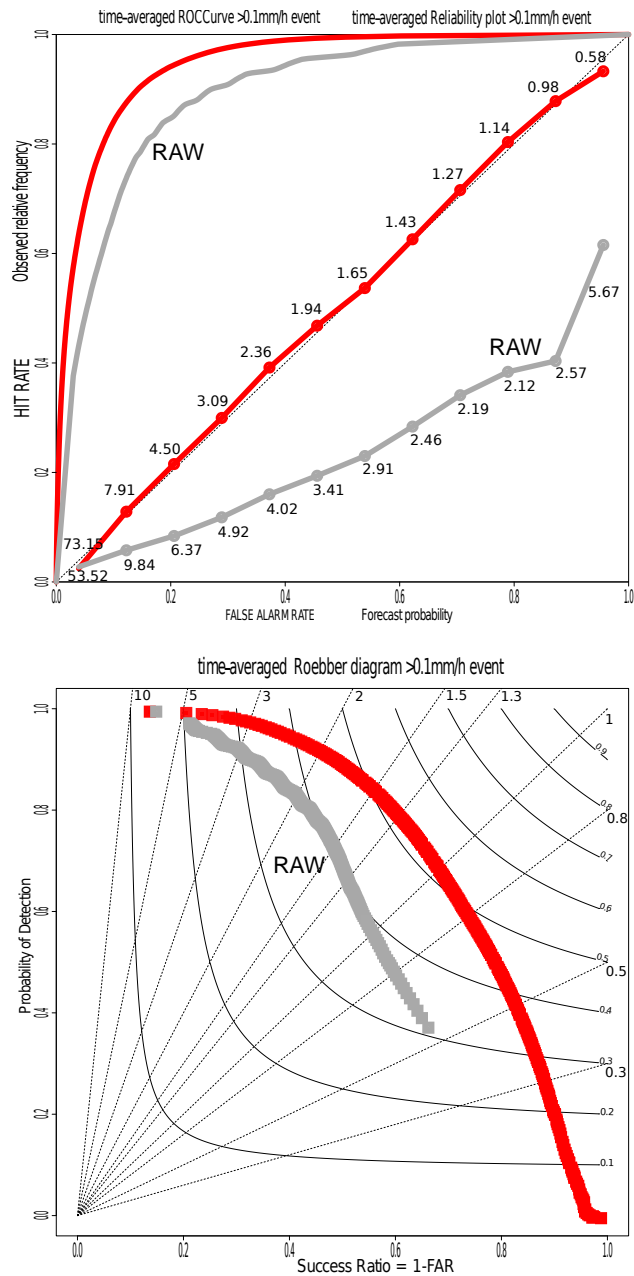


Figure 12. ROC curve and reliability diagram (top) and categorical performance diagram (bottom) for the rain event. In the performance diagram, the background dotted lines represent Frequency Bias Index and the curves represent Critical Success Index. The raw ensemble suffers from overprediction. The validation is made by a 2-fold cross-validation on the two years of data (one sample per year).

occupation of each process. For example, PEARP temperature uses 4 HPC nodes in 25 minutes, (QRF calibration: 64 cores on 4 nodes (16 cores per node) during 10 minutes, ECC phase: 12 cores on one node during 2 minutes, spatialisation

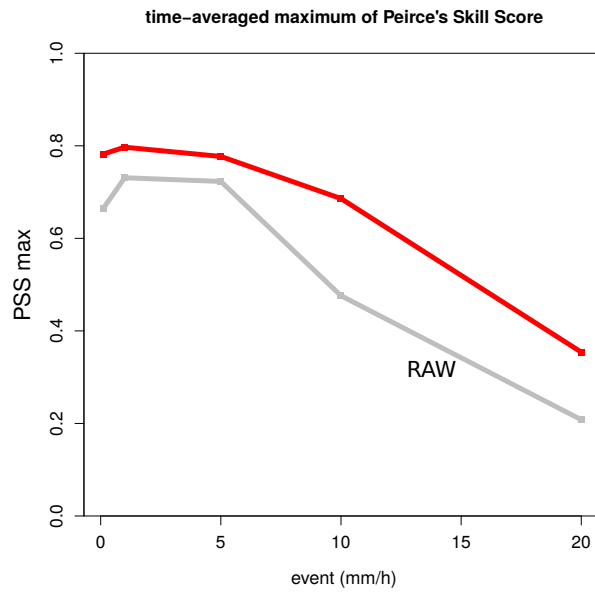


Figure 13. Maximum of the Peirce Skill Score among thresholds ; the improvement is mainly due to the improvement of the Hit Rate. **The validation is made by a 2-fold cross-validation on the two years of data (one sample per year).**

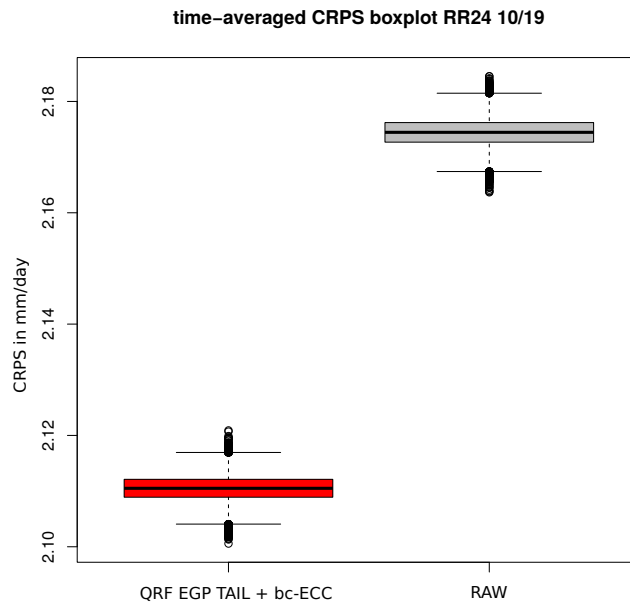


Figure 14. CRPS of daily distributions during October 2019.

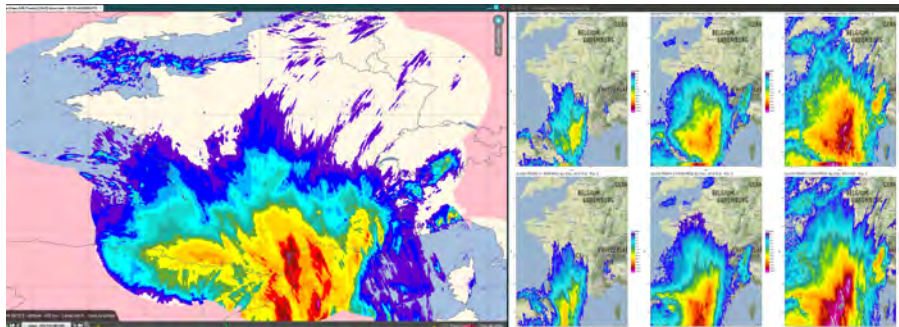


Figure 15. Illustration of a heavy precipitation event. On the left, rainfall accumulated the 10/22/2019, with peaks over 300 mm. On the right, quantiles of order 0.1, 0.5, and 0.9 of post-processed (on top) and raw (on bottom) daily rainfall distributions. (Right maps data ©2019 Google)

phase: 76 cores on 4 nodes during 15 minutes). PEAROME rainfall uses 162 cores on 18 HPC nodes during 22 minutes
 370 for QRF EGP TAIL calibration, and 432 cores on 6 HPC nodes during 3 minutes for bc-ECC.

– The huge size of objects produced by quantile regression forests. For a given base time, PEARP temperature application
 requires around 300 Gbytes of data to be read and loaded into memory, while PEAROME rainfall forests represents
 more than 600Gbytes of data. Reading this huge amount of data in a reasonable time is possible primarily due to the
 Infiniband network implemented in the supercomputer, which features a very high throughput and very low latency in
 375 I/Os operations. Also, stripping R QRF objects from useless features (regarding prediction) allows us to save a substantial
 amount of space.

Those two applications now deliver post-processed fields of higher quality than raw NWP fields, and will be used in the
 future Meteo-France automatic production chain, which is currently in its implementation phase. Post-processed fields are also
 of higher predictive value, and can lead to great benefits for (trained) human forecasters provided that the dialog between NWP
 380 scientists, statisticians and users is strengthened (Fundel et al., 2019).

5 Conclusion

In this article, we show that machine learning techniques allow a very large improvement of probabilistic temperature forecasts
 - a well known result that can also be achieved with simpler methods such as EMOS. But while EMOS outputs follow simple
 and fixed parametric distributions, QRF produces distributions that may preserve the richness of the initial ensemble. Also, a
 385 simple method such as ECC coupled with our spatialisation algorithm is able to restore realistic high resolution temperature
 fields for each member.

Moreover, HCA-based QRF calibration is able to calibrate efficiently a much trickier parameter such as hourly rainfall accumulation - for which the signal for extremes is of special importance and provide realistic rainfall patterns that match initial members.

390 In the context of forecast automation, it is important to identify the end users and their expectations in order to choose a method that balances complexity and efficiency. In the same vein, minimizing an expected score may be less important than reducing big (and costly) mistakes. For example, the European Center for Medium-Range Weather Forecasts (ECMWF) recently added the frequency of large errors in ensemble forecasts of surface temperature as a new headline score (Haiden et al., 2019).

395 Of course, applicability of those methods is not restricted to temperature and rainfall. For any parameter that can be interpolated rather easily (humidity for example), our "temperature scheme" that is, calibration on station locations, ECC and spatialisation may be applied. This approach is much less greedy in terms of computation time and disk storage. In addition, Feldmann et al. (2019) show the benefit of using observations rather than gridded analyses. For other parameters, such as cloud cover, windspeed, adaptations of HCA-based calibration (where the observation can also be viewed as a distribution) would be
400 a better option.

The only limitations of post-processing are the availability of good gridded observations or a sufficiently dense station network, and the existence of relevant predictors produced by NWP. Those conditions may not yet always be fully achieved, for parameters that remain challenging, such as visibility, for example.

Finally, as recalled in the discussion, production of high resolution post-processed fields with such techniques has proven
405 to be extremely demanding in terms of CPU and disk storage. Moving the post-processing chain to supercomputers is a challenging but fruitful investment: the learning phase that could take weeks is now achieved in a few hours. This provides extra possibilities for tuning parameters of powerful or promising statistical methods, as mentioned in Rasp and Lerch (2018), this is unavoidable for a quick operational production. Note that using the operational supercomputer hardly interferes with NWP production: by definition, post-processing comes after NWP runs are completed, and the number of nodes required by
410 post-processing is two orders of magnitude smaller.

Author contributions. MT developed the station-wise post-processing of PEARP and the post-processing of PEAROME with bc-ECC. OM developed algorithms of interpolation of scattered data and ECC for temperatures. OM configured the operational chain for temperature. OM and MT currently configure the operational chain for rainfall. OM made figures for temperature. MT created the figures for rainfall and scores. OM and MT wrote the publication, each rereading the other's part.

415 *Competing interests.* MT is one of the editors of the Special Issue.

Acknowledgements. The authors would thank the team COMPAS/DOP of Météo-France and more particularly Harold Petithomme and Michaël Zamo for their work on R codes. The authors would also thank Denis Ferriol for their help during the set-up of R codes on the supercomputer.

References

- 420 Athey, S., Tibshirani, J., Wager, S., et al.: Generalized random forests, *The Annals of Statistics*, 47, 1148–1178, 2019.
- Baran, S. and Lerch, S.: Combining predictive distributions for the statistical post-processing of ensemble forecasts, *International Journal of Forecasting*, 34, 477–496, 2018.
- Barry, R. G.: *Mountain weather and climate*, 2008.
- Bellier, J., Bontron, G., and Zin, I.: Using meteorological analogues for reordering postprocessed precipitation ensembles in hydrological forecasting, *Water Resources Research*, 53, 10 085–10 107, 2017.
- 425 Bellier, J., Zin, I., and Bontron, G.: Generating Coherent Ensemble Forecasts After Hydrological Postprocessing: Adaptations of ECC-Based Methods, *Water Resources Research*, 54, 5741–5762, 2018.
- Ben Bouallègue, Z., Heppelmann, T., Theis, S. E., and Pinson, P.: Generation of scenarios from calibrated ensemble forecasts with a dual-ensemble copula-coupling approach, *Monthly Weather Review*, 144, 4737–4750, 2016.
- 430 Bénichou, P.: Cartography of statistical pluviometric fields with an automatic allowance for regional topography, in: *Global Precipitations and Climate Change*, pp. 187–199, Springer, 1994.
- Bouttier, F., Raynaud, L., Nuissier, O., and Ménétrier, B.: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX, *Quarterly Journal of the Royal Meteorological Society*, 142, 390–403, 2016.
- Breiman, L.: Random forests, *Machine learning*, 45, 5–32, 2001.
- 435 Breiman, L., Friedman, J., Stone, C. J., and Olshen, R.: *Classification and Regression Trees*, CRC Press, 1984.
- Bremnes, J. B.: Ensemble post-processing using quantile function regression based on neural networks and Bernstein polynomials, *Monthly Weather Review*, In press, <https://doi.org/10.1175/MWR-D-19-0227.1>, <https://doi.org/10.1175/MWR-D-19-0227.1>, 2019.
- Clark, M., Gangopadhyay, S., Hay, L., Rajagopalan, B., and Wilby, R.: The Schaake shuffle: A method for reconstructing space–time variability in forecasted precipitation and temperature fields, *Journal of Hydrometeorology*, 5, 243–262, 2004.
- 440 Courtier, P., Freydier, C., Geleyn, J.-F., Rabier, F., and Rochas, M.: The Arpege project at Meteo France, in: *Seminar on Numerical Methods in Atmospheric Models*, 9-13 September 1991, vol. II, pp. 193–232, ECMWF, ECMWF, Shinfield Park, Reading, <https://www.ecmwf.int/node/8798>, 1991.
- Cressie, N.: Spatial prediction and ordinary kriging, *Mathematical geology*, 20, 405–421, 1988.
- Dabernig, M., Mayr, G. J., Messner, J. W., and Zeileis, A.: Spatial ensemble post-processing with standardized anomalies, *Quarterly Journal of the Royal Meteorological Society*, 143, 909–916, 2017.
- 445 Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P.: PEARP, the Météo-France short-range ensemble prediction system, *Quarterly Journal of the Royal Meteorological Society*, 141, 1671–1685, 2015.
- Feldmann, K., Richardson, D. S., and Gneiting, T.: Grid-Versus Station-Based Postprocessing of Ensemble Temperature Forecasts, *Geophysical Research Letters*, 46, 7744–7751, 2019.
- 450 Franke, R.: Smooth interpolation of scattered data by local thin plate splines, *Computers & mathematics with applications*, 8, 273–281, 1982.
- Frei, C.: Interpolation of temperature in a mountainous region using nonlinear profiles and non-Euclidean distances, *International Journal of Climatology*, 34, 1585–1605, 2014.
- Fundel, V. J., Fleischhut, N., Herzog, S. M., Göber, M., and Hagedorn, R.: Promoting the use of probabilistic weather forecasts through a dialogue between scientists, developers and end-users, *Quarterly Journal of the Royal Meteorological Society*, 2019.

- 455 Gascón, E., Lavers, D., Hamill, T. M., Richardson, D. S., Bouallègue, Z. B., Leutbecher, M., and Pappenberger, F.: Statistical post-processing of dual-resolution ensemble precipitation forecasts across Europe, *Quarterly Journal of the Royal Meteorological Society*, 2019.
- Genuer, R., Poggi, J.-M., and Tuleau-Malot, C.: VSURF: an R package for variable selection using random forests, 2015.
- Gneiting, T.: Calibration of medium-range weather forecasts, *European Centre for Medium-Range Weather Forecasts*, 2014.
- Gneiting, T. and Katzfuss, M.: Probabilistic forecasting, *Annual Review of Statistics and Its Application*, 1, 125–151, 2014.
- 460 Gneiting, T. and Raftery, A. E.: Strictly proper scoring rules, prediction, and estimation, *Journal of the American Statistical Association*, 102, 359–378, 2007.
- Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M., and Palmer, T.: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts, *Quarterly Journal of the Royal Meteorological Society*, 138, 1814–1827, 2012.
- Haiden, T., Janousek, M., Vitart, F., Ferranti, L., and Prates, F.: Evaluation of ECMWF forecasts, including the 2019 upgrade, 465 <https://doi.org/10.21957/mlvapkke>, <https://www.ecmwf.int/node/19277>, 2019.
- Hamill, T. M.: Practical aspects of statistical postprocessing, in: *Statistical Postprocessing of Ensemble Forecasts*, pp. 187–217, Elsevier, 2018.
- Hemri, S., Haiden, T., and Pappenberger, F.: Discrete postprocessing of total cloud cover ensemble forecasts, *Monthly Weather Review*, 144, 2565–2577, 2016.
- 470 Hosking, J. R. M., Wallis, J. R., and Wood, E. F.: Estimation of the generalized extreme-value distribution by the method of probability-weighted moments, *Technometrics*, 27, 251–261, 1985.
- Hudson, G. and Wackernagel, H.: Mapping temperature using kriging with external drift: theory and an example from Scotland, *International journal of Climatology*, 14, 77–91, 1994.
- Laurantin, O.: ANTILOPE: Hourly rainfall analysis merging radar and rain gauge data, in: *Proceedings of the International Symposium on* 475 *Weather Radar and Hydrology*, pp. 2–8, 2008.
- Lee, S., Wolberg, G., and Shin, S. Y.: Scattered data interpolation with multilevel B-splines, *IEEE transactions on visualization and computer graphics*, 3, 228–244, 1997.
- Manzato, A.: A note on the maximum Peirce skill score, *Weather and Forecasting*, 22, 1148–1154, 2007.
- Meinshausen, N.: Quantile regression forests, *Journal of Machine Learning Research*, 7, 983–999, 2006.
- 480 Naveau, P., Huser, R., Ribereau, P., and Hannart, A.: Modeling jointly low, moderate, and heavy rainfall intensities without a threshold selection, *Water Resources Research*, 52, 2753–2769, 2016.
- Paluszynska, A.: Biecek P.randomForestExplainer: Explaining and Visualizing Random Forests in Terms of Variable Importance, R package version 0.9, 2017.
- Papastathopoulos, I. and Tawn, J. A.: Extended generalised Pareto models for tail estimation, *Journal of Statistical Planning and Inference*, 485 143, 131–143, 2013.
- Peppier, R. A.: A review of static stability indices and related thermodynamic parameters, Tech. rep., Illinois State Water Survey, <http://hdl.handle.net/2142/48974>, 1988.
- R Core Team: R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, <https://www.R-project.org/>, 2015.
- 490 Rasp, S. and Lerch, S.: Neural networks for postprocessing ensemble weather forecasts, *Monthly Weather Review*, 146, 3885–3900, 2018.
- Roebber, P. J.: Visualizing multiple measures of forecast quality, *Weather and Forecasting*, 24, 601–608, 2009.

- Salazar, E., Sansó, B., Finley, A. O., Hammerling, D., Steinsland, I., Wang, X., and Delamater, P.: Comparing and blending regional climate model predictions for the American southwest, *Journal of agricultural, biological, and environmental statistics*, 16, 586–605, 2011.
- Saveliev, A. A., Romanov, A. V., and Mukharamova, S. S.: Automated mapping using multilevel B-Splines, *Applied GIS*, 1, 17–01, 2005.
- 495 Schefzik, R., Thorarinsdottir, T. L., Gneiting, T., et al.: Uncertainty quantification in complex simulation models using ensemble copula coupling, *Statistical science*, 28, 616–640, 2013.
- Scheuerer, M. and Hamill, T. M.: Generating calibrated ensembles of physically realistic, high-resolution precipitation forecast fields based on GEFS model output, *Journal of Hydrometeorology*, 19, 1651–1670, 2018.
- Scheuerer, M., Hamill, T. M., Whitin, B., He, M., and Henkel, A.: A method for preferential selection of dates in the S chaake shuffle approach
500 to constructing spatiotemporal forecast fields of temperature and precipitation, *Water Resources Research*, 53, 3029–3046, 2017.
- Schlosser, L., Hothorn, T., Stauffer, R., Zeileis, A., et al.: Distributional regression forests for probabilistic precipitation forecasting in complex terrain, *The Annals of Applied Statistics*, 13, 1564–1589, 2019.
- Schmeits, M. J. and Kok, K. J.: A comparison between raw ensemble output, (modified) Bayesian model averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts, *Monthly Weather Review*, 138, 4199–4211, 2010.
- 505 Seity, Y., Brousseau, P., Malardel, S., Hello, G., Bénard, P., Bouttier, F., Lac, C., and Masson, V.: The AROME-France convective-scale operational model, *Monthly Weather Review*, 139, 976–991, 2011.
- Stein, J. and Stoop, F.: Neighborhood-based contingency tables including errors compensation, *Monthly Weather Review*, 147, 329–344, 2019.
- Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated ensemble forecasts using quantile regression forests and ensemble model
510 output statistics, *Monthly Weather Review*, 144, 2375–2393, 2016.
- Taillardat, M., Fougères, A.-L., Naveau, P., and Mestre, O.: Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting, *Weather and Forecasting*, 34, 617–634, 2019.
- Thorarinsdottir, T. L., Gneiting, T., and Gissibl, N.: Using proper divergence functions to evaluate climate models, *SIAM/ASA Journal on Uncertainty Quantification*, 1, 522–534, 2013.
- 515 Van Schaeybroeck, B. and Vannitsem, S.: Ensemble post-processing using member-by-member approaches: theoretical aspects, *Quarterly Journal of the Royal Meteorological Society*, 141, 807–818, 2015.
- van Straaten, C., Whan, K., and Schmeits, M.: Statistical postprocessing and multivariate structuring of high-resolution ensemble precipitation forecasts, *Journal of Hydrometeorology*, 19, 1815–1833, 2018.
- Vannitsem, S., Wilks, D. S., and Messner, J.: *Statistical postprocessing of ensemble forecasts*, Elsevier, 2018.
- 520 Whan, K. and Schmeits, M.: Comparing area probability forecasts of (extreme) local precipitation using parametric and machine learning statistical postprocessing methods, *Monthly Weather Review*, 146, 3651–3673, 2018.
- Whiteman, C. D.: *Mountain meteorology: fundamentals and applications*, Oxford University Press, 2000.
- Zamo, M., Bel, L., Mestre, O., and Stein, J.: Improved gridded wind speed forecasts by statistical postprocessing of numerical models with block regression, *Weather and Forecasting*, 31, 1929–1945, 2016.
- 525 Zimmerman, D., Pavlik, C., Ruggles, A., and Armstrong, M. P.: An experimental comparison of ordinary and universal kriging and inverse distance weighting, *Mathematical Geology*, 31, 375–390, 1999.