Reviewer #1

RC#1.1. The authors propose a quantile regression forest (QRF)-based postprocessing method for the height of new snow (HN). The results are compared to a recently proposed ensemble model output statistics (EMOS) approach for postprocessing HN forecasts. QRF shows clear improvements over the EMOS model, in particular the inclusion of additional predictor variables seems to be benefitial.

Overall, I found the paper to be interesting, well-written and easy to follow throughout. I only have some minor and technical comments that are outlined below.

We thank the reviewer for this positive feedback and these useful comments.

RC#1.2. I found the description of the forecast distribution in Section 4.3 a bit confusing. Perhaps it would help to specifically clarify here that the resulting forecast is a set of quantiles derived from the observations from the final nodes, and not the empirical distribution in equation (1).

Thank you for this relevant comment. It is true that the empirical distribution given in Eq. 1 of the current manuscript does not correspond to the implementation with the function `quantregForest`. Indeed, it is correct to indicate that the resulting forecast is a set of quantiles derived from the observations from the final nodes, and not the empirical distribution in equation (1). We will correct this paragraph in the revised manuscript to clarify this point:

"*For QRF, the theoretical predictive distribution given a new set of predictors x is the conditional CDF introduced by Meinshausen (2006):*

$$F(y|x) = \sum_{i=1}^{n} w_i(x)1(Y_i \leq y) \qquad \textit{Eq. (1)}$$

where the weights wi(x) are deduced from the presence of Yi in a final leaf of each tree when one follows the path determined by x. In practise, the resulting forecast is a set of quantiles from $F(y|x)$ *using the function* `predict.quantregForest` *of the package* `quantregForest` *in* `R`*. Different quantiles are thus computed for synthetic graphical representations or scores computations.*"

RC#1.3. line 145 ff: Perhaps a few more details should be provided on why the CRPS is computed in this form. The motivation to create more quantile forecasts than raw ensemble members is not entirely clear to me. In the end, you compute an ensemble of quantiles that has many more members than the raw ensemble. While this makes it easier to account for the necessarily finite size of the sample from the forecast distribution and makes the comparison to EMOS (with a continuous forecast distribution) more "fair", doesn't this represent an "unfair" advantage when comparing to the raw ensemble?

We agree that this part is complicated and maybe not sufficiently discussed. When the true forecast CDF is not fully known and represented as an ensemble of values, the CRPS is estimated with some error. Thus, using the CRPS to compare parametric probabilistic forecasts with ensemble forecasts may be misleading due to the unknown error of the estimated CRPS for the ensemble. Here, creating more quantile forecasts aim at reducing the error of the CRPS estimation, and provides a more fair comparison with the EMOS and the CRPS of the raw distribution. The raw ensemble is composed of a limited number of members (it is not a predictive distribution) and the best (the most fair) estimate is obtained using Eq. (3). An artificial augmentation of the raw ensemble is not relevant in that case because the raw ensemble does not contain more information than what is contained in the different members. This is the difference with QRF forecasts which provide a much larger set of different quantiles (usually hundreds of them) but there is no easy way to know precisely the corresponding order. The method proposed in Zamo and Naveau (2018) consists in computing a reasonable number of quantiles with a regular order and to apply an interpolation between these quantiles in order to be as close as possible to the true CRPS value that would be obtained if all predicted quantiles were known. The different CRPS computations can thus be explained by the information contained in the corresponding predicted forecasts.

RC#1.4. line 169 ff: Compared to the description of the QRF model, the description of feature importance is rather short and probably difficult to understand for readers unfamiliar with QRF. Perhaps a few more details and fomulas here might help make this more clear.

The notion of feature importance will be developed in the revised version.

RC#1.5. Section 6, first paragraph: Are the CRPS values computed for the test set, the training set, or another validation period?

Thank you for this comment. These sensitivity tests have been carried out on the test set. It will be specified in the revised manuscript.

RC#1.6. Figure 2: Given that the show forecasts seem to be of particular importance, wouldn't include more summary statistics from that variable further improve results?

This is a point that could be tested. But first, it must be noted that snow rate forecasts are very correlated with HN forecasts simulated by Crocus. Very likely, including more statistics for the snow rate just adds more redundancy in the predictors. The first selection of the predictors for the snow and rain rates was based on our knowledge about the most important variables for the prediction of HN.

To verify this point, an additional experiment was performed. In addition to the approach tested in the manuscript, we test the QRF approach with 8 predictors based on snow and rate forecasts. For both snow and rain rates, we add the mean, the standard deviation, the probability to be non-zero, and the interquartile range of the forecast ensemble. Figure 1 below illustrates the corresponding results in terms of CRPS and CRPSS, similarly to Figure 5 of the manuscript. Clearly, adding more predictors does not change the overall performance of the QRF approach, with the same average performance, and a very small variability according to the stations (see the small range of the CRPSS for the case “QRF+Pred”).

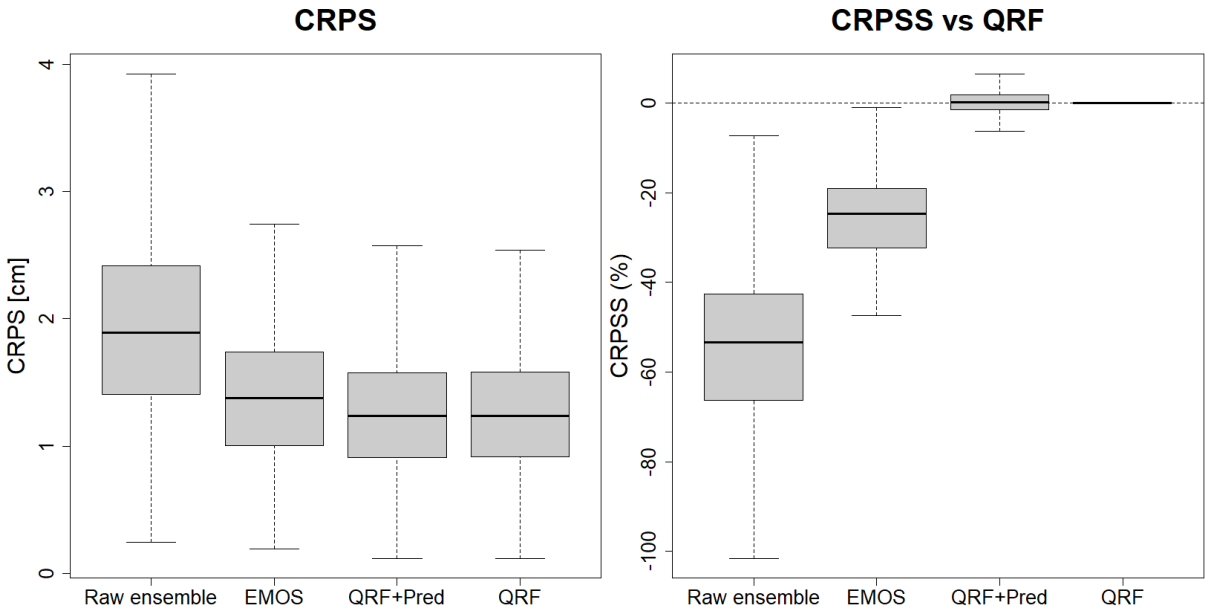

Figure 1. Boxplots of CRPS (left plots) and relative CPRS with QRF with the predictors chosen in the manuscript as a reference (right plots) with the different methods (Raw ensemble, EMOS and QRF+Pred corresponding to the QRF with more snow and rain rates predictors) for all locations, for a 1-day lead time.

RC#1.7. Figure 3: I find the confidence intervals difficult to distinguish due to the overlap and would suggest to split up the plot into three panels for all of the 3 models.

The first version of this plot was considering separate plots but it appears that the comparison becomes difficult. Following a suggestion made by the other reviewer (see comment RC#2.7), we propose to overlay plain colored lines instead, in order to highlight the lower and upper bounds of each approach.

RC#1.8. Overall, the paper in particular demonstrates that the inclusion of additional predictor variables improves performance. This is very much in line with the previous work on QRF and also several other machine learning-based postprocessing methods (for example proposed in Messner et al (2017, https://doi.org/10.1175/MWR-D-16-0088.1), Rasp and Lerch (2018, https://doi.org/10.1175/MWR-D-18-0187.1), Bremnes (2020, https://doi.org/10.1175/MWR-D-19-0227.1), and others). Since EMOS only uses forecasts of the variable of interest as predictor, it would have been more "fair" to compare the the boosting extension of EMOS proposed in the paper by Messner et al. While this is beyond the scope of the paper in the current form, I'd suggest to include this aspect in the discussion as an avenue for future work.

Thank you for this discussion. Indeed, a boosting extension of EMOS could in principle be applied in our context. Recently, Schulz and Lerch (2021) compares a gradient-boosting extension of EMOS (EMOS-GB) to many machine learning methods for postprocessing ensemble forecasts of wind gusts, using a truncated logistic distribution. The performances of EMOS-GB were honorable, but, as you indicate, other machine learning-based postprocessing methods seem promising. The Distributional Regression Network of Rasp and Lerch (2018) and the Bernstein Quantile Network of

Bremnes (2020) often outperform all the other methods in Schulz and Lerch (2021), including QRF. This will be added in the discussion.

Concerning gradient-boosting extension of EMOS, in our knowledge, it has never been applied with a zero-censored distribution (e.g. a Censored Shifted Gamma distribution as in our study). The implementation needs to be developed, which does not seem straightforward and the gain of performances for the prediction of zero values, which is critical in our case, has not been shown.

RC#1.9. line 59: Missing reference?

Thank you. There was indeed an issue with a reference. It will be removed.

RC#1.10. Section 4.2: I'd suggest to consider moving this Section to the beginning of Section 6. In the current form, the meaning of the hyperparmeters mtry and nodesize is not yet explained and will be difficult to understand for readers not familiar with QRF.

Thank you for this relevant comment. As Section 4.2 belongs to methodological aspects and are not results, we will keep this paragraph here, but the revised manuscript will provide the definitions of mtry and nodesize in Sec. 4.2, as we agree that it cannot be understood in the current form.

RC#1.11. Code and data availability: The limitations on availability of the EMOS and NWP model code are clearly explained, but it is unclear to me whether (or where) the QRF code is available.

This will be added. We essentially use available R packages for the QRF method.

References

Bremnes, J. B.. 2020. “Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bernstein Polynomials.” *Monthly Weather Review* 148 (1): 403–14. https://doi.org/10.1175/MWR-D-19-0227.1.

Rasp, S., and S. Lerch. 2018. “Neural Networks for Postprocessing Ensemble Weather Forecasts.” *Monthly Weather Review* 146 (11): 3885–3900. https://doi.org/10.1175/MWR-D-18-0187.1.

Schulz, B., and S. Lerch. 2021. “Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison.” *ArXiv:2106.09512 [Physics, Stat]*, June. http://arxiv.org/abs/2106.09512.