Response to reviewer's comments on

npg-2021-18 v2

”Calibrated ensemble forecasts of the height of new snow using quantile regression forests and Ensemble Model Output Statistics"

19 August, 2021

We thank the Editor and the three reviewers for their positive feedback and their comments on the revised version. These comments are reported in *blue* and in italic font. The additional suggestions have been taken into account. Please see our response to the different comments below.

EC

EC#1: My first comment has actually to do, at least in part, with science. You apparently use the words resolution and sharpness as if they corresponded to different properties of a probabilistic prediction system. They actually correspond to the same property, namely (as you write) the ability to separate a priori the probability classes, or to distinguish a priori between different outcomes (see Broecker, 2014). Please use only one word or, if you use both, say they refer to the same property.

Thank you for this relevant comment. The term “resolution” has been replaced by “sharpness” throughout the revised version, except when it was used to refer to the spatial resolution of NWP models.

EC#2: I add that the ROC curve, shown on your Figure 7, is a diagnostic of that property. It shows the degree to which the system under consideration is able to distinguish a priori between ‘hits’ and ‘false alarms’, i.e. between occurrence or non-occurrence of the considered events. That is exactly sharpness. I suggest you replace the words ‘good’ prediction in the caption of the figure with the words sharp prediction system (that will also remove the uncertainty implied by the quotation marks in ‘good’).

Thank you, this has been done.

EC#3: A number of acronyms are not expanded, at least not the first time they are used (e.g. PEARP-S2M on l. 31). Please check systematically that all acronyms are expanded on their first occurrence, and give appropriate references whenever necessary.

The meaning of PEARP-S2M and the related acronyms are now defined at l. 31-34 of the revised manuscript. Different corrections have been made for other acronyms (CPRSS replaced by CRPS, CI and IC have been replaced by PI for “predictive intervals, definition of CART)”.

EC#4: Figure 6. Was the number of intervals used for building the histograms arbitrary, or did it correspond to anything imbedded from the start in the prediction system. If yes, to what does it correspond (that is not clear to me) ?

The number of intervals was arbitrary, not too small to have a fine description of the distribution of the ranks, but not too high in order to have a sufficient number of items inside each class.

EC#5: . L. 202, … at the end of ~~March~~ the period.

Thank you, this has been corrected.

EC#6: Table 1, l. 7. PR0 Raw probability of HN > 0

This has been corrected.

EC#7: L. 216, The second column … → The right panel …

This has been corrected.

EC#8: I find the caption of Figure 3 somewhat confusing. I suggest … lead time (orange full lines) …, QRF (purple dashes) and EMOS (green points). And next sentence For each of the three prediction systems, the lower and upper curves represent the 10th and 90th percentiles respectively.

The captions of Figures 3 and 4 have been modified. We now refer to “orange plain lines”, “purple dashed lines” and “green dotted lines”.

EC#9: Figure 2. Say more precisely what the vertical coordinate on the figure is.

We now precise that importance is the “sum of squares of the differences between predicted and observed response variables, averaged over all trees obtained with the random permutations”. We have also corrected the related sentence before, as it was not correct to indicate that it was the standard deviation of the response variable.

EC#10: L. 139, … equals 0 (or is equal to 0)

This has been changed to “equals 0”.

EC#11: L. 219 (and Table 2). I presume CI means centiles?

CI was indeed misleading and we now indicate that these ranges are related to predictive intervals (PI).

Reviewer #1

Thank you for your responses and revision of the paper. I only have one remaining very minor comment. In your response, you wrote that "Thank you for this comment. These sensitivity tests have been carried out on the test set. It is now specified (l. 181)." In line 181 of the revised paper, you write that ") average CRPS values for the validation datasets" Which is correct?

Thank you for your comment. The manuscript is correct, the average CRPS values for the sensitivity tests have been computed on the validation datasets, using the leave-one season out cross-validation scheme.

Reviewer #2

Term Height of new snow – I accept the authors argument that this is a standard term in cryosphere science and the addition of reference is helpful. Nevertheless, most readers are not cryosphere specialists and I still think a very short phrase such as "…height of new snow (Fierz et al., 2009) (also commonly known as depth of snow) …." would be helpful to a large proportion of readers.

We appreciate this comment and we have added "also commonly known as depth of fresh snow" after "Fierz et al., 2009", the term "fresh" being important to avoid confusions with the long-term snow depth (height of the snow cover).