



Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics

Guillaume Evin, Matthieu Lafaysse, Maxime Taillardat, Michaël Zamo

► To cite this version:

Guillaume Evin, Matthieu Lafaysse, Maxime Taillardat, Michaël Zamo. Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics. *Nonlinear Processes in Geophysics*, 2021, 28 (3), pp.467-480. 10.5194/npg-28-467-2021 . meteo-03543728

HAL Id: meteo-03543728

<https://meteofrance.hal.science/meteo-03543728>

Submitted on 26 Jan 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License



Calibrated ensemble forecasts of the height of new snow using quantile regression forests and ensemble model output statistics

Guillaume Evin¹, Matthieu Lafaysse², Maxime Taillardat³, and Michaël Zamo³

¹Univ. Grenoble Alpes, INRAE, UR ETGR, Grenoble, France

²Univ. Grenoble Alpes, Université de Toulouse, Météo-France, CNRS, CNRM, Centre d'Etudes de la Neige, 38000 Grenoble, France

³CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France

Correspondence: Guillaume Evin (guillaume.evin@inrae.fr)

Received: 23 April 2021 – Discussion started: 28 April 2021

Revised: 19 August 2021 – Accepted: 23 August 2021 – Published: 16 September 2021

Abstract. Height of new snow (HN) forecasts help to prevent critical failures of infrastructures in mountain areas, e.g. transport networks and ski resorts. The French national meteorological service, Météo-France, operates a probabilistic forecasting system based on ensemble meteorological forecasts and a detailed snowpack model to provide ensembles of HN forecasts. These forecasts are, however, biased and underdispersed. As for many weather variables, post-processing methods can be used to alleviate these drawbacks and obtain meaningful 1 to 4 d HN forecasts. In this paper, we compare the skill of two post-processing methods. The first approach is an ensemble model output statistics (EMOS) method, which can be described as a nonhomogeneous regression with a censored shifted Gamma distribution. The second approach is based on quantile regression forests, using different meteorological and snow predictors. Both approaches are evaluated using a 22 year reforecast. Thanks to a larger number of predictors, the quantile regression forest is shown to be a powerful alternative to EMOS for the post-processing of HN ensemble forecasts. The gain of performance is large in all situations but is particularly marked when raw forecasts completely miss the snow event. This type of situation happens when the rain–snow transition elevation is overestimated by the raw forecasts (rain instead of snow in the raw forecasts) or when there is no precipitation in the forecast. In that case, quantile regression forests improve the predictions using the other weather predictors (wind, temperature, and specific humidity).

1 Introduction

In cold regions (e.g. mountainous areas), the height of new snow (Fierz et al., 2009; also commonly known as the depth of fresh snow) expected for short lead times is critical for many safety issues (e.g. avalanche hazard) and the economical impacts of dysfunctional transport networks (road, airports, and train track viability). National weather services increasingly provide automatic predictions for that purpose, usually relying on numerical weather prediction (NWP) model outputs. Forecasting the height of new snow (HN) is particularly challenging for many reasons. First, the precipitation forecasts in NWP models are biased and underdispersed. Then, HN is strongly dependent on elevation in mountainous areas, and this relationship cannot be perfectly reproduced by the current resolution of NWP models. Finally, several processes affecting snow properties (density, height, and precipitation phase) are either absent or poorly represented in NWP models (e.g. density of falling snow and mechanical compaction during the deposition). In particular, the evolution of the rain–snow limit elevation can greatly differ according to meteorological conditions and is only partly understood (Schneebeli et al., 2013).

Few attempts have been made to post-process ensemble HN forecasts. To the best of our knowledge, Stauffer et al. (2018) and Scheuerer and Hamill (2019) are the first studies to present post-processed ensemble forecasts of HN. They consider direct ensemble NWP outputs as predictors (precipitation and temperature). Nousu et al. (2019) incorporate physical modelling of the snowpack in order to integrate the

high temporal variations in temperature and precipitation intensity during a storm event, which can have highly nonlinear impacts on HN. In addition, Nousu et al. (2019) demonstrate the ability of a nonhomogeneous regression method to improve the ensemble forecasts of HN from the PEARP-S2M ensemble snowpack (ARPEGE – Action de Recherche Petite Echelle Grande Echelle; PEARP – Prévision d’Ensemble ARPEGE; SAFRAN – Système Atmosphérique Fournissant des Renseignements Atmosphériques à la Neige; SURFEX – SURFace EXternalisée; MEPRA – Modèle Expert pour la Prévision du Risque d’Avalanches; S2M – SAFRAN–SURFEX MEPRA). Using a regression method based on the censored shifted Gamma distribution (Scheuerer and Hamill, 2015, 2018), the forecast skill was improved for the majority of the stations from common events to more unusual events. However, as this method only considers a single predictor (the simulated HN itself) at a given point, dry days and rainy days cannot be discriminated as long as all forecast members provide a zero value for HN. This prevents an appropriate correction of some specific NWP errors, such as a systematic error among all simulation members in the rain–snow transition elevation.

In this study, we consider the application of quantile regression forests (QRFs) as an alternative to nonhomogeneous regression methods. This approach has been successfully applied for the post-processing of ensemble forecasts of surface temperature, wind speed (Taillardat et al., 2016), and rainfall (Taillardat et al., 2019). QRFs are often considered as being a non-parametric method since they do not rely on an explicit mathematical relationship between the predictors and the target distribution of the predictand. Furthermore, many predictors can be incorporated without decreasing the forecast skill (Taillardat et al., 2019), which can be particularly interesting in our case when, for example, the raw ensemble only contains zero HN while rainfall forecasts are large. Indeed, in some cases, the PEARP-S2M ensemble snowpack completely misses large snow events (e.g. due to an erroneous rain/snow limit). For some problematic meteorological situations, QRFs can possibly provide a specific correction.

Section 2 summarizes the forecasts and observations data set used in this study. Section 3 provides the details of the ensemble model output statistics (EMOS) method tested in this study, a particular nonhomogeneous regression method already employed in Nousu et al. (2019) and considered here as a benchmark method. Section 4 describes the QRF method. In Sect. 5, we detail the evaluation of the performances of each method. Section 6 presents the results. Finally, Sect. 7 provides a discussion of the results with some future outlooks.

2 Data

In this study, we select 92 stations in the French Alps and Pyrenees based on a minimum availability of observations

Table 1. Set of all available predictors.

| Name | Description |
|---|--|
| CTRL | Control member of raw ensemble of HN |
| MEAN | Mean of raw ensemble of HN |
| MED | Median of raw ensemble of HN |
| Q10 | First decile of raw ensemble of HN |
| Q90 | Ninth decile of raw ensemble of HN |
| PR0 | Raw probability of HN > 0 cm |
| PR1 | Raw probability of HN > 1 cm |
| PR3 | Raw probability of HN > 3 cm |
| PR5 | Raw probability of HN > 5 cm |
| PR10 | Raw probability of HN > 10 cm |
| PR20 | Raw probability of HN > 20 cm |
| SIGMA | Standard deviation of raw ensemble of HN |
| IQR | Interquartile range of raw ensemble of HN |
| Q10, 50, and 90 are the first decile, the median, and ninth decile of the raw ensemble for the following variables: | |
| SNOWR_q10,50,90 | Snow rate ($\text{kg m}^{-2} \text{h}^{-1}$) |
| RAINR_q10,50,90 | Rain rate ($\text{kg m}^{-2} \text{h}^{-1}$) |
| WIND_q10,50,90 | Wind speed (m s^{-1}) |
| TAIR_q10,50,90 | Near-surface air temperature (K) |
| QAIR_q10,50,90 | Near-surface specific humidity (ϕ) |

of 60 % (percentage of missing observations thus varies between 0 % and 40 %, with an average of 18 %). Forecasts and observations are available and reliable for these 92 stations presented in Fig. 1 for 22 winter seasons covering the period 1994–2016, where each winter season starts on 6 December and ends on 30 April of the following year (3218 d in total).

The forecasts are obtained by a chain of ensemble numerical simulations. The 10-member reforecasts of the PEARP ensemble NWP (Descamps et al., 2015; Boisserie et al., 2016) are downscaled by the SAFRAN system (Durand et al., 1999) to obtain a meteorological forcing adjusted in elevation. The Crocus multilayer snowpack model, part of the S2M modelling chain (Vernay et al., 2019), is forced by these forecasts to provide ensemble simulations of HN, accounting for all the main physical processes explaining the variability in HN for a given precipitation amount, namely the dependence of falling snow density on meteorological conditions, the mechanical compaction over time depending on snow weight, the microstructure and wetness of the snow, a possible surface melting, and so on. The forecasts used in this paper are the same as those used by Nousu et al. (2019), who provided more details on the models’ configurations.

Table 1 presents the selected predictors based on the available reforecasts. This selection is derived from studies of Scheuerer and Hamill (2015) and Taillardat et al. (2019) for rainfall. It includes summary statistics and probabilities of the variable to be predicted (rainfall in the previous references transposed into HN in our case). We also consider statistics of other weather variables of the ensemble suspected to add predictability because they might affect the statistical relationship between observed and simulated HN.

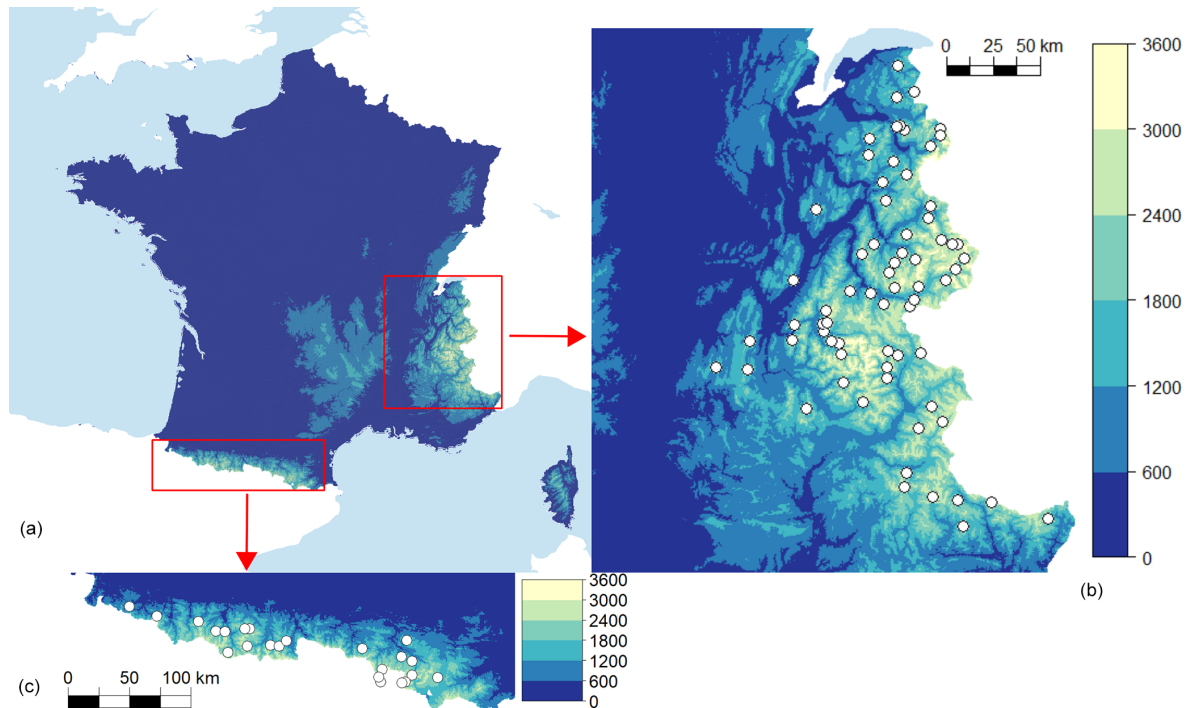


Figure 1. Map of the 92 observation stations (white dots) in the French Alps (b) and Pyrenees (c).

In this paper, for each station, we thus consider $i = 1, \dots, n = 3218$ d with an observed HN Y_i (the response) and a vector of corresponding predictors X_i .

3 Ensemble model output statistics

Among the ensemble model output statistics (EMOS) methods available, non-homogeneous regression approaches are the most common and were originally based on Gaussian regressions, whose mean and variance are linear functions of ensemble statistics (Gneiting et al., 2005; Wilks and Hamill, 2007). Non-homogeneous regression methods can also incorporate climatological properties and additional predictors. For meteorological predictands such as rainfall and snow, however, the high number of zero values motivate the use of a discrete–continuous distribution with a mass of probability at zero. In this study, we use a regression based on the zero-censored shifted Gamma distribution (CSGD; see Scheuerer and Hamill, 2015, 2018; Nousu et al., 2019).

The non-homogeneous regression method applied in this study is similar to the approach presented in Nousu et al. (2019) and is referred to as EMOS hereafter. Further details of this EMOS method are presented in Appendix A. More precisely, we detail how the CSGD is used to represent the predictive distribution of daily HN forecasts, the parameter estimation method, and the related predictive distribution. Please note that expression (A3) is slightly different from expression (4) in Nousu et al. (2019) and strictly corresponds

to Scheuerer and Hamill (2018) (see their expression of σ in Sect. 3a, p. 1653). While this difference is not critical on the performances, expression (A3) avoids scaling issues in parameter β_2 .

4 Quantile regression forest

Compared to the EMOS method, quantile regression forest is expected to incorporate any predictor without degrading the quality of the predictions. Subsets of the space covered by the predictors are created in order to obtain homogeneous groups of observations inside these subsets. If the predictors include many meteorological forecasts, these subsets are expected to describe different meteorological situations. Compared to EMOS, this so-called non-parametric regression does not assume a particular distribution for the predictors or the response, and empirical distributions represent the uncertainty about the prediction.

4.1 Method

The QRF methods presented in this paper are based on the construction of binary decision trees, as proposed by Meinshausen (2006). These decision trees (Classification and Regression Trees or CART for short; Breiman et al., 1984) are built by iteratively splitting each predictor space (\mathcal{D}_0) into two groups (\mathcal{D}_1 and \mathcal{D}_2) according to some threshold. The predictor and the threshold are chosen in order to maximize the homogeneity of the corresponding values of the response

(here the observed HN) in each of the resulting groups, i.e. we want to minimize the sum of variances of the response variable within each group as follows:

$$v(\mathcal{D}_j) = \sum_{Y \in \mathcal{D}_j} [Y - \bar{Y}]^2,$$

where Y and \bar{Y} correspond to the response sample and its mean in \mathcal{D}_j , respectively. The optimal threshold s maximizes the following:

$$\mathcal{H}(\mathcal{D}_1, \mathcal{D}_2) = \max_{s \in \mathcal{T}^*} [v(\mathcal{D}_0) - v(\mathcal{D}_1) - v(\mathcal{D}_2)],$$

where \mathcal{T}^* is a random subset of the predictors in the predictors' space \mathcal{T} . These trees are obtained by bootstrapping the training data, which justifies the name of “random forest” since each split of each tree is built on a random subset of the predictors (Breiman, 2001). The final “leaf” corresponds to the group of predictors at the end of each tree (see Fig. 1 in Taillardat et al., 2019, for an illustration).

4.2 Implementation

The QRFs are obtained using the function `quantregForest` of the package `quantregForest` in R (R Core Team, 2017). The splitting procedure described above can be constrained by different choices, e.g. a minimum number of observations in leaves. In this paper, we grow 1000 trees, which represent a sufficiently large number of trees to span a great variety of meteorological situations without demanding an unbearable computational effort. Different values for the parameters `mtry`, which specify the number of predictors randomly sampled as candidates at each split (usually small, i.e. less than 10) and `nodesize` which defines the minimum number of cases (days) in terminal nodes, have been tried, and the best performances being found for `mtry` = 2 and `nodesize` = 10 (see Sect. 6).

4.3 Predictive distribution

For QRFs, the predictive distribution, given a new set of predictors x , is the conditional cumulative distribution function (CDF) introduced by Meinshausen (2006) as follows:

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

where the weights $w_i(x)$ are deduced from the presence of Y_i 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 $\hat{F}(y|x)$ obtained with the function `predict.quantregForest` from the R package `quantregForest`. Different quantiles are thus computed for synthetic graphical representations or for score calculations.

5 Evaluation

This section details the process applied to assess the performance of the different approaches. Classical evaluation metrics include the continuous ranked probability score (CRPS), which sums up the forecast performance attributes in terms of both reliability and sharpness simultaneously (Murphy and Winkler, 1987; Hersbach, 2000; Candille and Talagrand, 2005). Rank histograms are also a common tool to assess systematic biases and over/under dispersion.

5.1 Cross-validation

For all the experiments in this study, we use a leave-one-season-out cross-validation scheme. For each of the 22 seasons, one season is used as a validation data set while the other 21 seasons are used for training. It first ensures that a robust calibration of the post-processing methods is obtained. It also avoids the evaluation of the performances with a unique validation period that could be atypical (e.g. a very snowy/dry winter season).

5.2 CRPS

The CRPS is one of the most common probabilistic tools for evaluating the ensemble skill in terms of reliability (unbiased probabilities) and sharpness (ability to separate the probability classes). For a given forecast, the CRPS corresponds to the integrated quadratic distance between the cumulative distribution function (CDF) of the ensemble forecast and the CDF of the observation. Commonly, the CRPS is averaged over n days as follows:

$$\overline{\text{CRPS}} = \frac{1}{n} \sum_{i=1}^n \int_{\mathbb{R}} [F_i(y) - H(y - Y_i)]^2 dy, \quad (2)$$

where $F_i(x)$ is the CDF obtained from the ensemble forecasts for day i , Y_i is the corresponding observation, and $H(z)$ is the Heaviside function ($H(z) = 0$ if $z < 0$; $H(z) = 1$ if $z \geq 0$). The CRPS value has the same unit as the evaluated variable and equals zero for a perfect system.

For the EMOS-CSGD model described above, an analytic formulation of the CRPS is available (Scheuerer and Hamill, 2015), and a correct CRPS estimation is directly obtained.

In other cases, a correct evaluation of the CRPS defined in Eq. (2) can be difficult. For example, the raw ensemble does not provide a forecast CDF but only a very limited ensemble of values. In this case, the CRPS is estimated with some error. In this study, we apply the recommendations given by Zamo and Naveau (2018). More specifically, when the forecast CDF is known only through an M -ensemble x_1, \dots, x_M , we apply the following definition to estimate the instantaneous CRPS (i.e. for one ensemble) as follows:

$$\widehat{\text{CRPS}}_{\text{INT}} = \int_{\mathbb{R}} \left[\frac{1}{M} \sum_{m=1}^M H(x - x_m) - H(x - y) \right]^2 dx, \quad (3)$$

where y is the observation corresponding to the forecast ensemble. This expression is evaluated with the function `crpsDecomposition` of the R package `verification`. In the case of the instantaneous CRPS of the raw ensemble forecasts, Eq. (3) is applied directly, while some refinements can be made to improve the estimated CRPS in the case of QRFs, which provide a much larger number of different quantiles (so-called order) than what is available in the raw ensemble. Unfortunately, the set of possible quantiles and their corresponding order cannot be known a priori, which represents an additional difficulty. To evaluate instantaneous CRPS values for QRFs, we thus use the recommendations by Zamo and Naveau (2018), i.e. we use the average $\overline{\text{CRPS}}_{\text{INT}}$ given in Eq. (3) with linearly interpolated regular quantiles between unique quantiles. The so-called regular ensemble of $M = 200$ quantiles $x_{i=1,\dots,M}$ of orders $\tau_{i=1,\dots,M} \in [0; 1]$ is defined as $z_i = F^{-1}(\tau_i)$, for all i , with $\tau_i \in \{\frac{1}{M}, \frac{2}{M}, \dots, \frac{M-1}{M}, \frac{M-0.1}{M}\}$.

5.3 Sharpness

While the CRPS are often used to verify the overall quality of the predictive distributions, it can also be interesting to assess the sharpness of the predictions. Gneiting et al. (2007) propose looking at the width of the predictive intervals for different nominal coverages (e.g. 50 % and 90 %).

5.4 Rank histograms

The reliability of ensemble forecast systems can be assessed using rank histograms (Hamill, 2001). If the predictive distributions obtained with the different post-processing methods are adequate, then the CDF values of the predictive distributions for the observations should be uniformly distributed (so-called probability integral transform – PIT). The flatness of the histogram of these CDF values is a necessary but not a sufficient condition of the system reliability. Systematic biases are detected by strongly asymmetric rank histograms. It is also an indicator of the spread skill, as underdispersion will result in a U shaped rank histogram and overdispersion in a bell-shaped rank histogram. Rank histograms can be computed for the whole forecast data set or stratified according to different classes of average ensemble forecasts (stratifying according to the observations leading to erroneous conclusions; see Bellier et al., 2017). In this study, as proposed in the recent study of Bröcker and Bouallègue (2020), a stratification based on the average of the combination of raw forecasts and the verification observations is used. In total, the following three HN intervals are considered: [0 cm; 10 cm), [10 cm; 30 cm), and [30 cm; ∞). To guarantee a sufficient sample size for rank histograms, they are computed for the whole evaluation data set by considering all dates and stations as being independent.

5.5 ROC curves

Finally, the relative operating characteristic (ROC) curves (Kharin and Zwiers, 2003) can be used to assess the quality of probability forecasts by relating the hit rate (probability of detecting an event which actually occurs) to the corresponding false alarm rate (probability of detecting an event which does not occur).

6 Results

We first discuss the application of the QRF methods with regards to the parameters `mtry` (number of predictors randomly sampled as candidates at each split) and `nodesize` (minimum number of days in terminal nodes). Different values have been tried for both parameters (2, 4, 6, 8, and 10 for `mtry` and 5, 10, 15, and 20 for `nodesize`). For a 1 d lead time, the best (smallest) average CRPS values for the validation data sets are obtained for small values of `mtry` (2 or 4) and high `nodesize` values (15 or 20), with the mean CRPS being minimized for `mtry` = 2 and `nodesize` = 10 (results not shown). However, the range of mean CRPS values is narrow (1.282 and 1.294). We conclude here that the performances obtained with the QRF approach are not very sensitive to the value of the QRF parameters, and `mtry` = 2 and `nodesize` = 10 are retained in the rest of this study.

Figure 2 highlights the most important predictors for the QRF method for different lead times. The importance criteria here is related to the accuracy of the predictions when the predictors are permuted. Random permutations of each predictor variable X_j are applied in order to verify how well the response Y can be predicted with this deterioration. When an important permuted variable X_j and unpermuted predictor variables are used to predict the response, the prediction accuracy, quantified here with the sum of squares of the differences between predicted and observed response variables, decreases substantially. The variable importance is the difference in prediction accuracy before and after permuting X_j and is implemented by the function `importance` of the package `randomForest` in R (see, e.g., Louppe et al., 2013, for further details).

For a 1 d lead time, the Q90 of the forecast snow rate, followed by the Q90 of raw forecasts of HN, are the most important predictors. The most important predictors are directly related to snow quantities, and the role of other meteorological forcings is minor. As the lead time increases, the importance of the snow predictors decreases while the importance of the forecast rain rate becomes larger. In particular, the Q90 of the snow and rain rates are the two most important predictors at a 4 d lead time.

Figure 3 shows the time series of observed HN for a period with large snowfalls, along with raw reforecasts and predictive intervals with a 80 % probability obtained with the different post-processing methods, for the station 5079400 at

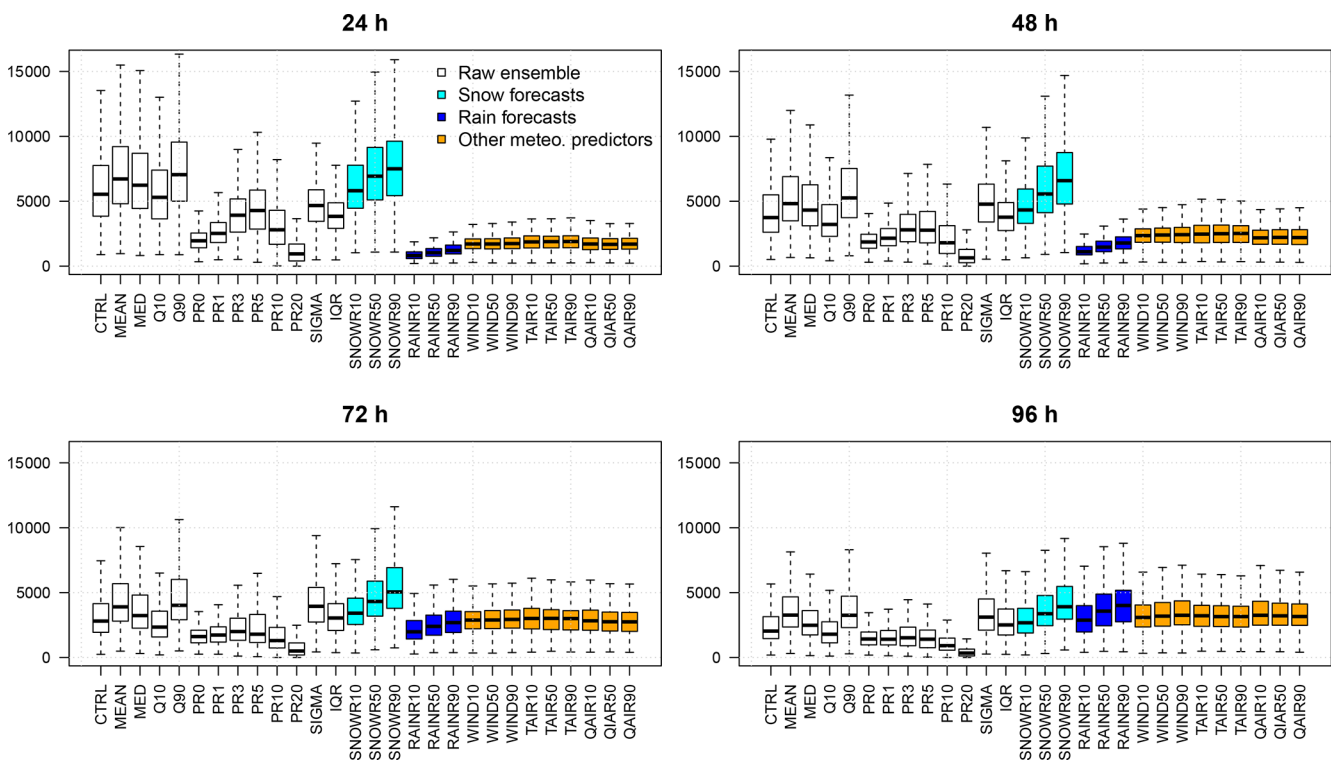


Figure 2. Importance criteria (sum of squares of the differences between predicted and observed response variables, averaged over all trees obtained with the random permutations) of the predictors for different lead times for the QRF method.

Table 2. Mean and standard deviation (SD) of the width of the predictive intervals (PIs; 50 % and 90 % nominal coverages), with the different methods and for all locations and dates, for a 1 d lead time. The width associated to a 50 % probability corresponds to the difference between the 25th and 75th percentiles, and the width for a 90 % probability corresponds to the difference between the 5th and 95th percentiles.

| | Mean PI 50 % | SD PI 50 % | Mean PI 90 % | SD PI 90 % |
|------|--------------|------------|--------------|------------|
| Raw | 0.40 | 1.16 | 0.69 | 1.66 |
| EMOS | 2.24 | 4.03 | 9.52 | 9.86 |
| QRFs | 2.62 | 4.87 | 8.66 | 11.12 |

Le Monétier-les-Bains during the period 26 February 2001–18 March 2001 and for a 1 d lead time. The following observations can be made:

- The raw ensembles generally underestimate the largest observed HN (see, e.g., the period 26 February–5 March). The intervals given by the raw ensembles are thin and underdispersed in comparison to post-processed ensembles.
- Predictive intervals obtained with the post-processing methods are large and look very similar. Observations generally lie within these intervals (with one major exception at the end of the period).
- When the raw reforecasts are all equal to zero, the EMOS method mechanically predicts zero HNs, which is often verified (see, e.g. on 5, 6, and 11 March). How-

ever, EMOS predicts these zero values with a 100 % probability, while QRF predicts small intervals in this example, which avoids failures (i.e. prediction of a zero value with absolute certainty while a positive HN value is observed). In this example, it happens for 2 d, on 7 and 9 March.

Figure 4 shows the time series for station 4193400 at Saint-Paul-sur-Ubaye during the period 29 March–18 April 2012. A large observed HN of 40 cm occurred on 10 April. The EMOS method completely misses this event because no snow was present in the raw forecasts. In this case, QRF predicts a large interval, with a 90th percentile around 20 cm. Looking at the raw forecasts of the meteorological forcings for this day, the 90 % intervals of the snow rate is [0.7, 8.1] and [1.2, 9.4] cm h^{−1} for the rain rate and [2.1, 3.2] °C for the air temperature. High snow/rain rates combined to above-

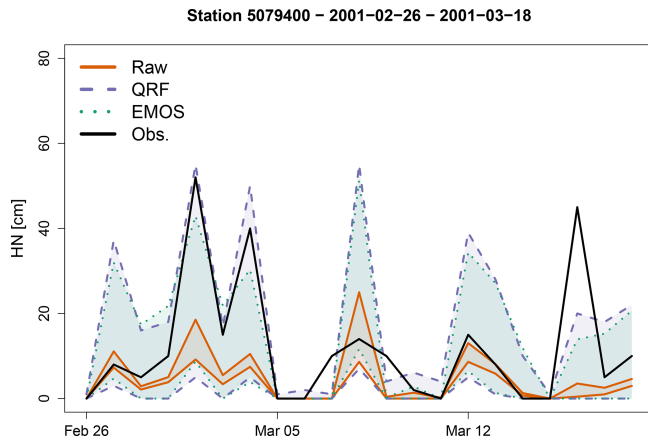


Figure 3. Time series of the raw reforecasts for a 1 d lead time (orange plain lines) and predictive quantiles using QRFs (purple dashed lines) and EMOS (green dotted lines) during March 2001 for the station 5079400. For each of the three prediction systems, the lower and upper curves represent the 10th and 90th percentiles, respectively. The solid black line represents the time series of the HN observations.

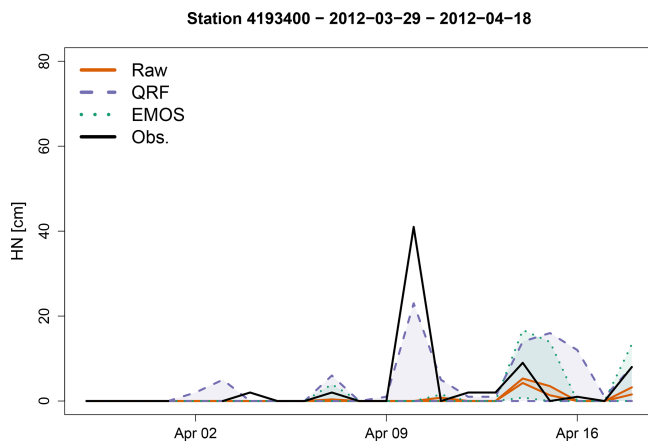


Figure 4. Time series of the raw reforecasts for a 1 d lead time (orange plain lines) and predictive quantiles using QRFs (purple dashed lines) and EMOS (green dotted lines) during April 2012 for the station 4193400. For each of the three prediction systems, the lower and upper curves represent the 10th and 90th percentiles, respectively. The solid black line represents the time series of the HN observations.

zero temperatures led to zero HN forecasts by the snowpack model, while the QRF method exploits these high precipitation rates in order to predict large HN amounts.

Figure 5 shows the 2024 CRPS values averaged over the different winter seasons (92 stations \times 22 winter seasons) obtained with the raw reforecasts, and with EMOS and QRF post-processing methods, for a 1 d lead time (left plots). While EMOS gives a considerable gain of performance, it is still outperformed by the QRF method. The right panel quantifies this improvement as a percentage in terms of rela-

tive CRPS. For most of the stations, EMOS shows a degradation of the performances between 20 % and 30 %, up to 40 % compared to QRF. Results (not shown) are very similar for the other lead times.

Table 2 reports the mean width and the corresponding standard deviation of the predictive intervals (50 % and 90 % nominal coverages) over all locations and dates, for a 1 d lead time, with the different methods. As indicated above and illustrated in Figs. 3 and 4, the predictive intervals obtained with the raw ensembles are a lot thinner than with EMOS and QRF, but they are underdispersed. The sharpness of the post-processed ensembles are very similar, the mean width for a 50 % probability being around 2.5 and 9 cm for a 90 % probability.

Figure 6 shows the rank histograms of HN with the raw forecasts and with EMOS and QRF post-processing methods. As indicated in previous studies (see, e.g., Nousu et al., 2019), raw forecasts are clearly underdispersed, leading to a *U* shape rank histogram, and also usually underestimate large HN values (over-representation of the last class). These defaults are particularly visible for classes of raw ensemble averages above 10 cm (rows 2–3). The rank histogram with the EMOS method is almost perfectly flat for the small ensemble/observation averages ([0, 10) cm). For larger classes of events, it seems that the EMOS predictive distribution is slightly underdispersed. The QRF method shows better performances than EMOS in that regard, with the only limitation being an underestimation of the largest snowfalls (see the last class for HN > 30 cm in the bottom-right plot).

Figure 7 shows the ROC curves for three categories of HN observed values, i.e. all snow events (HN greater than 1 cm; 19 % of the observed cases), “moderate” snow events (HN greater than 10 cm; 5 % of the observed cases), and “rare” snow events (HN greater than 30 cm; 1.4 % of the observed cases). Figure 7a shows that the raw forecast ensemble performs almost as well as post-processed ensembles when all snow events are considered. For this category, the purple curve corresponding to the QRF approach deviates farther away from the no-skill diagonal than the green curve corresponding to the EMOS method, indicating the better skill of the QRF approach. For moderate snow events (Fig. 7b), while QRF and EMOS show similar performances, the ROC curve corresponding to raw ensembles is close to the diagonal and indicates almost no skill. For rare and intense snow events exceeding 30 cm of fresh snow on a single day Fig. 7c shows a slight gain of performance with the QRF approach compared to EMOS.

To investigate further the different behaviours of EMOS and QRF, Fig. 8 shows the relative CRPS value of EMOS versus QRF for all dates and stations with a positive observed HN values (greater than 1 cm) and for different classes based on the predictors. More specifically, we try to investigate the difference in performances according to the presence or not of at least one positive rain/snow rate value among the different members of the ensemble forecast. Cases where there

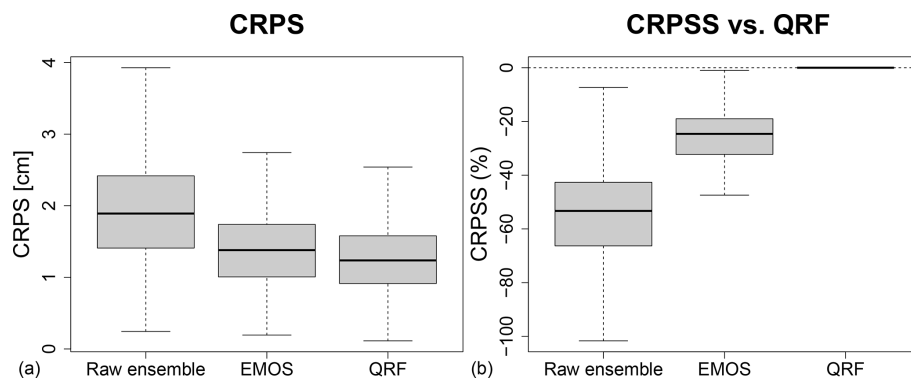


Figure 5. Box plots of CRPS (a) and relative CRPS with QRFs as a reference (b), with the different methods and for all locations, for a 1 d lead time.

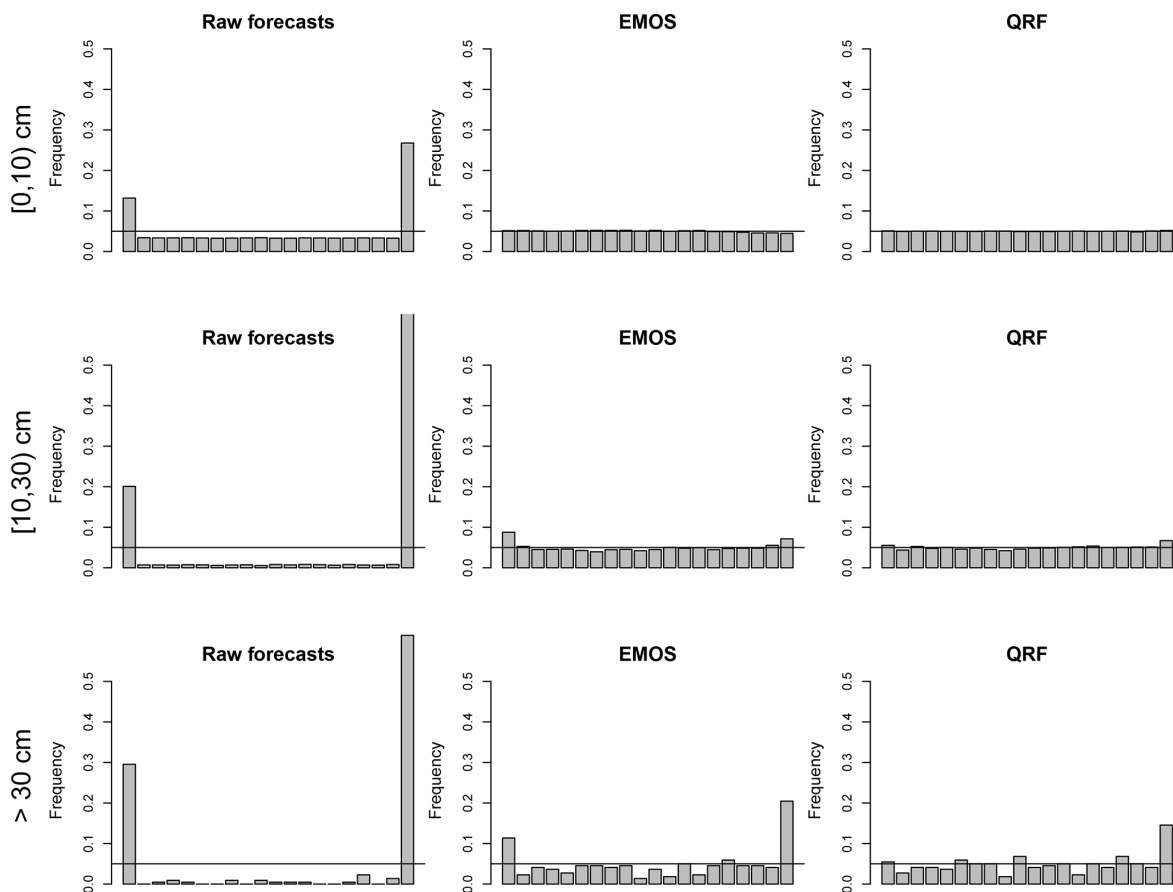


Figure 6. Rank histograms of HN forecasts for three classes of HN ensemble/observation mean, with the different methods, for a 1 d lead time.

is not any rain or snow in the forecasts while positive HN values have been observed represent only 0.4 % of all dates and stations (Fig. 8a). Cases corresponding to precipitation phase errors (at least one member with rain in the forecasts but no snow, while a positive HN has been measured; Fig. 8b) represents 1.2 % of all cases. Obviously, cases with snow in the forecasts and a positive HN are more frequent (8.9 %

and 14.7 % for cases c and d, respectively). Overall, while QRF outperforms EMOS in all cases (as outlined in Fig. 2), we see that the gain of performances is particularly marked for cases (a) and (b), i.e. when there is no snow in the forecasts. These results demonstrate the advantage of the QRF approach in this case, i.e. when other predictors (rain, temperature, etc.) can be exploited to overcome the limitations

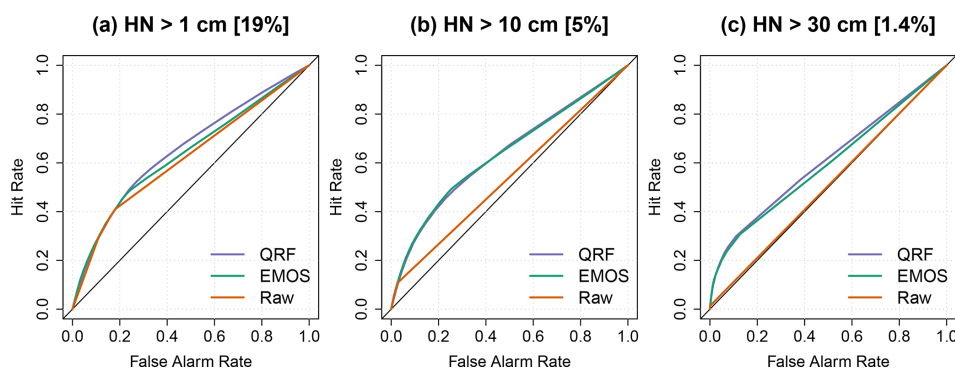


Figure 7. ROC curves for different snow events. (a) HN exceeds 1 cm. (b) HN exceeds 10 cm. (c) HN exceeds 30 cm. Values in brackets indicate the observed frequencies in percent. A sharp prediction system must maximize the hit rate and minimize false alarms.

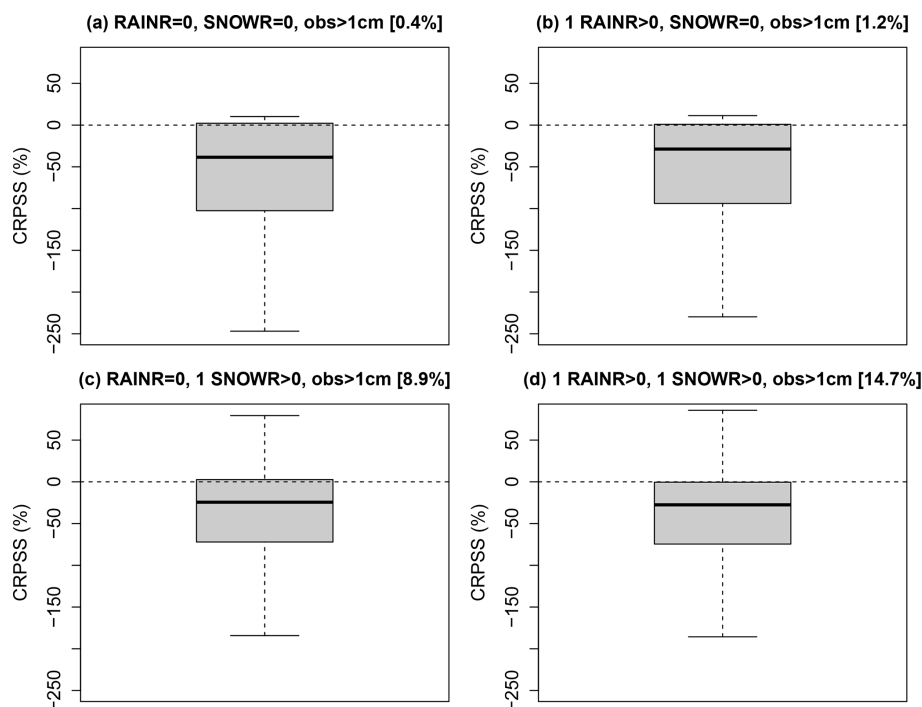


Figure 8. Relative CRPS values of EMOS versus QRFs as percentages, with QRFs as a reference, for all dates and stations and for a 1 d lead time. Only dates with a positive observation greater than 1 cm are selected and for different classes of predictors. (a) All forecast members of rain rate and snow rate equals to zero. (b) At least one member with a positive rain rate value and all forecast members of snow rate equals to zero. (c) All forecast members of rain rate equal to zero and at least one member with a positive snow rate value. (d) At least one member with a positive rain rate value and one member with a positive snow rate values. Values in brackets indicate the corresponding frequencies, in percent, among all the dates.

of the snow forecasts for the prediction of observed HN (see a further discussion in Sect. 7 below).

7 Discussion and outlook

7.1 Comparison of performances between QRF and EMOS approaches

In this paper, we compare the scores of post-processed forecasts of the 24 h height of new snow between two commonly used statistical methods, namely EMOS and QRF. With this data set, the added value of QRF is unambiguous, with a general improvement in CRPS, an improvement in rank dia-

grams for severe snowfall events, and a slight improvement in ROC curves for more common events. The predictors selected by the QRF training clearly suggest that the simulated HN from the Crocus snow cover model is useful but not sufficient to optimize the post-processed forecasts as the meteorological variables forcing the snow cover model are also selected by the algorithm. The added value coming from these meteorological predictors is the most likely explanation of the improvement obtained between QRF and EMOS. This improvement is frequent in various situations, and the physical reason for which the simulated HN does not translate all the predictive power of the meteorological forcings is probably not unique but can be partly explained by the presence of precipitation phase errors.

It must be noticed that the EMOS–CSGD model applied in this study only uses forecasts of the variable of interest as predictors. Different EMOS extensions can include more predictors, in particular the boosting extension (Messner et al., 2017). Schulz and Lerch (2021) compare 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 and other machine-learning-based postprocessing methods are promising. In particular, the distributional regression network (Rasp and Lerch, 2018) and the Bernstein quantile network (Bremnes, 2020) often outperform all the other methods, including QRFs. These recent models need, however, to be adapted to HN forecasts, i.e. using a zero-censored distribution with possibly long tails such as the CSGD.

7.2 Role of precipitation phase errors in the added value of QRFs

The examples selected for illustration suggest that phase errors (or, in other words, errors in the rain–snow transition elevation) is one of the possible explanations for the insufficient predictive power of the simulated HN. Indeed a number of observed snowfall events are simulated with a zero value in terms of HN, sometimes for all members, but with a large precipitation amount. EMOS is not able to consider these days with a large probability of positive HN because they are identical to dry days when considering only this predictor, whereas the other predictors considered by QRF (total precipitation and air temperature) can help to discriminate the days with an error in phase but with forecast precipitation and relatively cold conditions from dry days or warm days. This assumption is difficult to statistically generalize due to the large variety of situations, i.e. errors in precipitation phase often concern only a part of the total duration of a snowfall event and/or a part of the simulation members. Nevertheless, our classification of CRPS, depending on rainfall and snowfall occurrence shows a systematic improvement of CRPS by QRF for the cases where an error in the rain–snow transition elevation is the most obvious (e.g. observed snow-

fall with simulated rainfall but no simulated snowfall during the whole day for all members; Fig. 8b).

The sensitivity of snow cover models to errors in precipitation phase was already illustrated by Jennings and Molotch (2019), with a meteorological forcing built from weather stations. The magnitude of errors is expected to be much higher when the forcing comes from NWP forecasts. The reduction in phase errors in atmospheric modelling is beyond the scope of this paper. However, an improvement in post-processed forecasts might be opened by considering predictors more directly related to this phase issue. In particular, interviews of operational weather forecasters show that expert HN forecasts strongly rely on the 1 °C isothermal level in terms of pseudo-adiabatic wet-bulb potential temperature (θ'_w). Unfortunately, this diagnostic was not available in the PEARP reforecast, but this feedback encourages future reforecast productions to include this additional diagnostic, as the post-processing might be able to more directly account for phase errors with such a predictor. More simply considering the surface wet-bulb temperature is also increasingly done in land surface modelling for phase discrimination (Wang et al., 2019), and it may also be an easier alternative predictor for statistical post-processing, although the information content of the simulated atmospheric column is probably better summarized by the pseudo-adiabatic wet-bulb temperature iso- θ'_w (WMO, 1973). Nevertheless, forecasters also mention that a common limitation of NWP models is their inability to simulate the unusually thick 0 °C isothermal layers encountered in some intense storms (up to 1000 m). The complex interactions between the processes involved in this phenomenon are only partly understood (latent cooling from melting precipitation and evaporation/sublimation, melting distance of snowflakes, adiabatic cooling of rising air, specific topographies, blocked cold air pockets, etc.; Minder et al., 2011; Minder and Kingsmill, 2013). In these specific cases, even the level $\theta'_w = 1$ °C is considered to be a poor predictor of the rain–snow transition elevation. These situations are often the most critical in terms of impacts (wet snow at low elevations affecting the roads and the electrical network), but their very low frequency will remain a severe challenge even with statistical post-processing.

7.3 Limitations for operational perspectives

In order to investigate the potentials of the statistical methods themselves, regardless of the constraints on the available data set, we choose in this paper to calibrate and evaluate the post-processing methods on the same 22-year-long data set with a cross-validation scheme. However, Nousu et al. (2019) illustrate the strong impact of the discrepancies between reforecasts and operational forecasts in the post-processing efficiency. In complementary investigations (not shown), we noted that QRF is even more sensitive to the homogeneity between calibration and application data sets. For instance, the added value of QRF compared to EMOS was completely

lost when using the evaluation data set of (Nousu et al., 2019; operational PEARP-S2M forecasts). Therefore, despite the large added value of QRF compared to EMOS with consistent and homogeneous data sets for calibration and evaluation, its practical implementation in real-time operational forecasting products is still a challenge because reforecasts strictly identical to operational configurations are often not available. Time-adaptive training based on operational systems is an alternative to favour the homogeneity of the data set. Although new theories are emerging to face the challenge of model evolutions (Demaeyer and Vannitsem, 2020), several consistent recent studies show that the length of the calibration period is more critical than the strict homogeneity of data sets to forecast rare events (Lang et al., 2020; Hess, 2020). In the case of HN forecasts from EMOS (Nousu et al., 2019), even a 4-year calibration period was detrimental for the reliability of severe snowfall events compared to a longer heterogeneous reforecast. However, Taillardat and Mestre (2020) manage to successfully implement QRFs in real-time forecasting products of hourly precipitation, using a calibration limited to 2-year operational forecasts, because they adapt the distribution tail with a parametric method. Producing reforecasts that are more homogeneous with operational forecasts is still one of the most promising solutions to improve the forecast probabilities of severe events, but the evolutive skill of NWP systems is strongly linked to the available data to be assimilated and will never be completely removed. Therefore, the robustness of post-processing algorithms for their transfer to operational data set or their efficiency when calibrated with shorter data sets will always remain the most important criteria compared to their theoretical added values with perfect and long data sets. This is, therefore, a major point to consider to transpose the advances of this paper towards operational automatic HN forecasts.

Appendix A: Ensemble model output statistics for post-processing of ensemble forecasts of the daily HN

A1 Zero-censored, censored shifted Gamma regression

Here, the zero-censored, censored shifted Gamma regression distribution (CSGD) is used to represent the predictive distribution of daily HN forecasts and is defined as follows:

$$\tilde{G}_{k,\theta,\delta}(y) = \begin{cases} G_k\left(\frac{y-\delta}{\theta}\right) & \text{for } y \geq 0 \\ 0 & \text{for } y < 0 \end{cases}, \quad (\text{A1})$$

where k , θ , and δ are shape, scale, and shift parameters, respectively, and G_k is the CDF of a standard gamma distribution with unit scale and shape parameter k . The shape parameter k and scale parameter θ are directly related to the mean μ and the standard deviation σ of the gamma distribution through the relations $\mu = k\theta$ and $\sigma^2 = k\theta^2$. Scheuerer and Hamill (2018) propose a non-homogeneous regression based on the CSGD which combines a CSGD representing the climatology of past observations. For a given day, the parameters μ , σ , and δ of the predictive CSGD are related to the climatology and to the raw forecast ensemble with the following expressions (Scheuerer and Hamill, 2018, Sect. 3a):

$$\mu = \frac{\mu_{\text{cl}}}{\alpha_1} \log 1p[\exp m1(\alpha_1)(\alpha_2 + \alpha_3 \text{POP} + \alpha_4 \bar{x})], \quad (\text{A2})$$

$$\sigma = \sigma_{\text{cl}} \left(\beta_1 \sqrt{\frac{\mu}{\mu_{\text{cl}}}} + \beta_2 \text{MD} \right), \quad (\text{A3})$$

$$\delta = \delta_{\text{cl}}, \quad (\text{A4})$$

where $\log 1p(u) = \log(1 + u)$, and $\exp m1(u) = \exp(u) - 1$. The shift parameter δ is fixed at its climatological value δ_{cl} . This regression model only employs the statistical properties of HN ensemble forecasts, summarized by its ensemble mean \bar{x} , the probability of having a positive value POP, and the ensemble mean difference MD (a metric of ensemble spread), as defined by the following equations:

$$\bar{x} = \frac{1}{M} \sum_{m=1}^M x_m, \quad (\text{A5})$$

$$\text{POP} = \frac{1}{M} \mathcal{I}_{x_m > 0}, \quad (\text{A6})$$

$$\text{MD} = \frac{1}{M^2} \sum_{m=1}^M \sum_{m'=1}^M |x_m - x_{m'}|, \quad (\text{A7})$$

with x_m the raw HN forecast of each member m among the M members, and $\mathcal{I}_{x_m > 0} = 1$, if $x_m > 0$ and 0 otherwise.

A2 Parameter estimation

For each station, the six parameters $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \beta_1, \beta_2\}$ in Eqs. (A2)–(A4) are estimated by optimizing the CRPS prediction skill on the training data set. As CRPS can be directly expressed in the case of a CSGD (when $F_i = \tilde{G}_{k,\theta,\delta}$), this score can be easily minimized for this EMOS model. Complete expressions of the CRPS and details about model fitting are given in Scheuerer and Hamill (2015).

A3 Predictive distribution

Using Eqs. (A2)–(A4), parameter estimates $\{\hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\alpha}_4, \hat{\beta}_1, \hat{\beta}_2\}$ obtained on the training data set, and summary statistics \bar{x} , POP, and MD of a new ensemble forecasts, we directly obtain $\hat{\mu}$, $\hat{\sigma}$, and $\hat{\delta}$. This fully specifies the predictive distribution $\tilde{G}_{\hat{k},\hat{\theta},\hat{\delta}}(y)$ of this new ensemble forecast, with $\hat{k} = \hat{\mu}^2 / \hat{\sigma}^2$ and $\hat{\theta} = \hat{\sigma}^2 / \hat{\mu}$.

Code and data availability. The R code used for the application of the EMOS approach is based on different scripts originally developed by Michael Scheuerer (Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, and the NOAA Earth System Research Laboratory, Physical Sciences Division, Boulder, Colorado, USA). The modified version can be provided on request, with the agreement of the original author. The QRF approach has been applied using the R package `randomForest` for training and predictions. The score calculations have been performed using the R package `verification` and R functions developed by Mickaël Zamo. The Crocus snowpack model has been developed as part of the open-source SURFEX project (<http://www.umr-cnrm.fr/surfex/>, CNRM, 2021). The full procedure and documentation with respect to accessing this Git repository can be found at https://opensource.cnrm-game-meteo.fr/projects/snowtools_git/wiki (last access: 22 April 2021). The codes of PEARP and SAFRAN are not currently open source. For reproducibility of results, the PEARP version used in this study is “cy42_peace-op2.18”, and the SAFRAN version is tagged as “re-forecast_2018” in the private SAFRAN Git repository. The raw data of HN forecasts and reforecasts of the PEARP-S2M system can be obtained on request. The HN observations used in this work are public data available at <https://donneespubliques.meteofrance.fr> (Météo-France, 2021).

Author contributions. ML developed and ran the SURFEX/Crocus snowpack simulations forced by PEARP-SAFRAN outputs. GE set up the statistical framework, with the scientific contributions of MT and MZ. GE produced the figures. GE and ML wrote the publication, with contributions from all the authors.

Competing interests. The contact author has declared that neither they nor their co-authors have any competing interests.

Disclaimer. Publisher’s note: Copernicus Publications remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Acknowledgements. The authors would like to thank Bruno Joly, who developed and ran the PEARP reforecast, Matthieu Vernay, who developed and ran the SAFRAN downscaling of the PEARP reforecast and real-time forecasts, and Michael Scheuerer, for providing the initial code of the EMOS-CSGD. CNRM/CEN and INRAE are part of LabEX OSUG@2020 (ANR10 LABX56).

Financial support. This research has been supported by the Horizon 2020 Framework Programme, H2020 European Institute of Innovation and Technology (PROSNOW; grant no. 730203).

Review statement. This paper was edited by Olivier Talagrand and reviewed by Ken Mylne and two anonymous referees.

References

- Bellier, J., Bontron, G., and Zin, I.: Using Meteorological Analogues for Reordering Postprocessed Precipitation Ensembles in Hydrological Forecasting, *Water Resour. Res.*, 53, 10085–10107, <https://doi.org/10.1002/2017WR021245>, 2017.
- Boisserie, M., Decharme, B., Descamps, L., and Arbogast, P.: Land Surface Initialization Strategy for a Global Reforecast Dataset, *Q. J. Roy. Meteor. Soc.*, 142, 880–888, <https://doi.org/10.1002/qj.2688>, 2016.
- Breiman, L.: Random Forests, *Mach. Learn.*, 45, 5–32, <https://doi.org/10.1023/A:1010933404324>, 2001.
- Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A.: Classification and Regression Trees, Chapman and Hall/CRC, Boca Raton, United States, 1984.
- Bremnes, J. B.: Ensemble Postprocessing Using Quantile Function Regression Based on Neural Networks and Bernstein Polynomials, *Mon. Weather Rev.*, 148, 403–414, <https://doi.org/10.1175/MWR-D-19-0227.1>, 2020.
- Bröcker, J. and Bouallégue, Z. B.: Stratified Rank Histograms for Ensemble Forecast Verification under Serial Dependence, *Q. J. Roy. Meteor. Soc.*, 146, 1976–1990, <https://doi.org/10.1002/qj.3778>, 2020.
- Candille, G. and Talagrand, O.: Evaluation of Probabilistic Prediction Systems for a Scalar Variable, *Q. J. Roy. Meteor. Soc.*, 131, 2131–2150, <https://doi.org/10.1256/qj.04.71>, 2005.
- CNRM: SURFEX, available at: <http://www.umr-cnrm.fr/surfex/>, last access: 15 September 2021.
- Demaeyer, J. and Vannitsem, S.: Correcting for model changes in statistical postprocessing – an approach based on response theory, *Nonlin. Processes Geophys.*, 27, 307–327, <https://doi.org/10.5194/npg-27-307-2020>, 2020.
- Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., and Cébron, P.: PEARP, the Météo-France Short-Range Ensemble Prediction System, *Q. J. Roy. Meteor. Soc.*, 141, 1671–1685, <https://doi.org/10.1002/qj.2469>, 2015.
- Durand, Y., Giraud, G., Brun, E., Mérendol, L., and Martin, E.: A Computer-Based System Simulating Snowpack Structures as a Tool for Regional Avalanche Forecasting, *J. Glaciol.*, 45, 469–484, <https://doi.org/10.3189/S0022143000001337>, 1999.
- Fierz, C., Armstrong, R., Durand, Y., Etchevers, P., Greene, E., McClung, D., Nishimura, K., Satyawali, P., and Sokratov, S.: The International Classification for Seasonal Snow on the Ground (UNESCO, IHP (International Hydrological Programme)–VII, Technical Documents in Hydrology, No 83; IACS (International Association of Cryospheric Sciences) Contribution No 1), UNESCO/Division of Water Sciences, Paris, France, 2009.
- Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, *Mon. Weather Rev.*, 133, 1098–1118, <https://doi.org/10.1175/MWR2904.1>, 2005.
- Gneiting, T., Balabdaoui, F., and Raftery, A. E.: Probabilistic Forecasts, Calibration and Sharpness, *J. Roy. Stat. Soc. B*, 69, 243–268, <https://doi.org/10.1111/j.1467-9868.2007.00587.x>, 2007.
- Hamill, T. M.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, *Mon. Weather Rev.*, 129, 550–560, [https://doi.org/10.1175/1520-0493\(2001\)129<0550:IORHFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0550:IORHFV>2.0.CO;2), 2001.
- Hersbach, H.: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction Systems, *Weather*

- Forecast., 15, 559–570, [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2), 2000.
- Hess, R.: Statistical postprocessing of ensemble forecasts for severe weather at Deutscher Wetterdienst, *Nonlin. Processes Geophys.*, 27, 473–487, <https://doi.org/10.5194/npg-27-473-2020>, 2020.
- Jennings, K. S. and Molotch, N. P.: The sensitivity of modeled snow accumulation and melt to precipitation phase methods across a climatic gradient, *Hydrol. Earth Syst. Sci.*, 23, 3765–3786, <https://doi.org/10.5194/hess-23-3765-2019>, 2019.
- Kharin, V. V. and Zwiers, F. W.: On the ROC Score of Probability Forecasts, *J. Climate*, 16, 4145–4150, [https://doi.org/10.1175/1520-0442\(2003\)016<4145:OTRSOP>2.0.CO;2](https://doi.org/10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2), 2003.
- Lang, M. N., Lerch, S., Mayr, G. J., Simon, T., Stauffer, R., and Zeileis, A.: Remember the past: a comparison of time-adaptive training schemes for non-homogeneous regression, *Nonlin. Processes Geophys.*, 27, 23–34, <https://doi.org/10.5194/npg-27-23-2020>, 2020.
- Louppe, G., Wehenkel, L., Sutura, A., and Geurts, P.: Understanding Variable Importances in Forests of Randomized Trees, in: *Advances in Neural Information Processing Systems 26*, edited by: Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., Curran Associates, Inc., Red Hook, NY, United States, 431–439, 2013.
- Meinshausen, N.: Quantile Regression Forests, *J. Mach. Learn. Res.*, 7, 983–999, 2006.
- Messner, J. W., Mayr, G. J., and Zeileis, A.: Nonhomogeneous Boosting for Predictor Selection in Ensemble Postprocessing, *Mon. Weather Rev.*, 145, 137–147, <https://doi.org/10.1175/MWR-D-16-0088.1>, 2017.
- Météo-France: Portail de données publiques de Météo-France, available at: <https://donneespubliques.meteofrance.fr/>, last access: 15 September 2021.
- Minder, J. R. and Kingsmill, D. E.: Mesoscale Variations of the Atmospheric Snow Line over the Northern Sierra Nevada: Multi-year Statistics, Case Study, and Mechanisms, *J. Atmos. Sci.*, 70, 916–938, <https://doi.org/10.1175/JAS-D-12-0194.1>, 2013.
- Minder, J. R., Durran, D. R., and Roe, G. H.: Mesoscale Controls on the Mountainside Snow Line, *J. Atmos. Sci.*, 68, 2107–2127, <https://doi.org/10.1175/JAS-D-10-05006.1>, 2011.
- Murphy, A. H. and Winkler, R. L.: A General Framework for Forecast Verification, *Mon. Weather Rev.*, 115, 1330–1338, [https://doi.org/10.1175/1520-0493\(1987\)115<1330:AGFFV>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1330:AGFFV>2.0.CO;2), 1987.
- Nousu, J.-P., Lafaysse, M., Vernay, M., Bellier, J., Evin, G., and Joly, B.: Statistical post-processing of ensemble forecasts of the height of new snow, *Nonlin. Processes Geophys.*, 26, 339–357, <https://doi.org/10.5194/npg-26-339-2019>, 2019.
- R Core Team: R: A Language and Environment for Statistical Computing, available at: <https://www.R-project.org/> (last access: 15 September 2021), 2017.
- Rasp, S. and Lerch, S.: Neural Networks for Postprocessing Ensemble Weather Forecasts, *Mon. Weather Rev.*, 146, 3885–3900, <https://doi.org/10.1175/MWR-D-18-0187.1>, 2018.
- Scheuerer, M. and Hamill, T. M.: Statistical Postprocessing of Ensemble Precipitation Forecasts by Fitting Censored, Shifted Gamma Distributions, *Mon. Weather Rev.*, 143, 4578–4596, <https://doi.org/10.1175/MWR-D-15-0061.1>, 2015.
- Scheuerer, M. and Hamill, T. M.: Generating Calibrated Ensembles of Physically Realistic, High-Resolution Precipitation Forecast Fields Based on GEFS Model Output, *J. Hydrometeorol.*, 19, 1651–1670, <https://doi.org/10.1175/JHM-D-18-0067.1>, 2018.
- Scheuerer, M. and Hamill, T. M.: Probabilistic Forecasting of Snowfall Amounts Using a Hybrid between a Parametric and an Analog Approach, *Mon. Weather Rev.*, 147, 1047–1064, <https://doi.org/10.1175/MWR-D-18-0273.1>, 2019.
- Schneebeli, M., Dawes, N., Lehning, M., and Berne, A.: High-Resolution Vertical Profiles of X-Band Polarimetric Radar Observables during Snowfall in the Swiss Alps, *J. Appl. Meteorol. Clim.*, 52, 378–394, <https://doi.org/10.1175/JAMC-D-12-015.1>, 2013.
- Schulz, B. and Lerch, S.: Machine Learning Methods for Postprocessing Ensemble Forecasts of Wind Gusts: A Systematic Comparison, *arXiv [preprint]*, arXiv:2106.09512, 2021.
- Stauffer, R., Mayr, G. J., Messner, J. W., and Zeileis, A.: Hourly probabilistic snow forecasts over complex terrain: a hybrid ensemble postprocessing approach, *Adv. Stat. Clim. Meteorol. Oceanogr.*, 4, 65–86, <https://doi.org/10.5194/ascmo-4-65-2018>, 2018.
- Taillardat, M. and Mestre, O.: From research to applications – examples of operational ensemble post-processing in France using machine learning, *Nonlin. Processes Geophys.*, 27, 329–347, <https://doi.org/10.5194/npg-27-329-2020>, 2020.
- Taillardat, M., Mestre, O., Zamo, M., and Naveau, P.: Calibrated Ensemble Forecasts Using Quantile Regression Forests and Ensemble Model Output Statistics, *Mon. Weather Rev.*, 144, 2375–2393, <https://doi.org/10.1175/MWR-D-15-0260.1>, 2016.
- Taillardat, M., Fougères, A.-L., Naveau, P., and Mestre, O.: Forest-Based and Semiparametric Methods for the Postprocessing of Rainfall Ensemble Forecasting, *Weather Forecast.* 34, 617–634, <https://doi.org/10.1175/WAF-D-18-0149.1>, 2019.
- Vernay, M., Lafaysse, M., Hagenmuller, P., Nheili, R., Verfaillie, D., and Morin, S.: The S2M meteorological and snow cover reanalysis in the French mountainous areas (1958–present), *AERIS [data set]*, France, <https://doi.org/10.25326/37>, 2019.
- Wang, Y.-H., Broxton, P., Fang, Y., Behrangi, A., Barlage, M., Zeng, X., and Niu, G.-Y.: A Wet-Bulb Temperature-Based Rain-Snow Partitioning Scheme Improves Snowpack Prediction Over the Drier Western United States, *Geophys. Res. Lett.*, 46, 13825–13835, <https://doi.org/10.1029/2019GL085722>, 2019.
- Wilks, D. S. and Hamill, T. M.: Comparison of Ensemble-MOS Methods Using GFS Reforecasts, *Mon. Weather Rev.*, 135, 2379–2390, <https://doi.org/10.1175/MWR3402.1>, 2007.
- WMO: Compendium of Meteorology – for Use by Class I and II Meteorological Personnel: Volume I, Part 1 – Dynamic Meteorology, Publications of Blue Series, Volume 1 (1955–1984) – Education and Training Programme (2004), Geneva, Switzerland, 1973.
- Zamo, M. and Naveau, P.: Estimation of the Continuous Ranked Probability Score with Limited Information and Applications to Ensemble Weather Forecasts, *Math. Geosci.*, 50, 209–234, <https://doi.org/10.1007/s11004-017-9709-7>, 2018.