Cold Regions Science and Technology 120 (2015) 251–262

Contents lists available at ScienceDirect

Cold Regions Science and Technology

journal homepage: www.elsevier.com/locate/coldregions

Ensemble forecasting of snowpack conditions and avalanche hazard

Matthieu Vernay [1], Matthieu Lafaysse *, Laurent Mérindol, Gérald Giraud, Samuel Morin

Météo-France-CNRS, CNRM-GAME UMR3589, Centre d'Études de la Neige (CEN), Grenoble, France

ARTICLE INFO

Article history:
Received 3 November 2014
Received in revised form 21 April 2015
Accepted 23 April 2015
Available online 28 April 2015

Keywords:
Ensemble forecasting
Avalanche hazard
Snow cover
Predictability
Probabilistic forecast
Dispersion

ABSTRACT

The prediction of avalanche hazard involves an analysis of current snow conditions, the upcoming meteorological conditions and their combined impact on the future state of the snowpack. The SAFRAN–SURFEX/ISBA-Crocus–MEPRA (S2M) chain of numerical models is used by avalanche forecasters in France to estimate present and future avalanche hazard over areas assumed to be meteorologically homogeneous (massifs), primarily as a function of altitude. Until now, the meteorological forecast data provided to S2M comes from the deterministic numerical weather prediction model ARPEGE with a lead-time of 2 days. In this study, we introduce the application of ensemble meteorological forecasting to avalanche hazard forecasting by using the output of an ensemble of 35 ARPEGE predictions to feed S2M and thus provide an ensemble of 35 different predicted snowpack conditions. A posteriori ensemble forecasts were generated and evaluated in the French Alps for the winter 2013–2014 with 4 days lead time, initialized each day at 6 UTC. Forecasts over the Pyrenees during the exceptional winter and spring 2012–2013 were also carried out. Results indicate that accounting for the uncertainty in meteorological forecast significantly improves the skill and the usefulness of the model chain, regardless of the prediction lead time. The predictability of snowpack conditions using the ensemble forecast technique remains good at a 4 day lead time. These results provide the foundation for the development of probabilistic estimates of simulated avalanche hazard levels for operational avalanche hazard forecasting.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The prediction of regional-scale avalanche hazard for the production of avalanche bulletins, for example, involves an analysis of current snow conditions, the upcoming meteorological conditions and their combined impact on the future state of the snowpack. In France, the SAFRAN–SURFEX/ISBA-Crocus–MEPRA (S2M) chain of numerical models (Durand et al., 1999; Lafaysse et al., 2013) is used to provide an objective assessment of the past and future snow conditions including mechanical stability. It explicitly accounts for altitude, slope and aspect within geographical areas assumed to be meteorologically homogeneous (so-called massifs). Météo-France avalanche forecasters use S2M outputs for current and predicted snow conditions in combination with weather forecasts and field information from a dedicated ground observation network. Until now, the meteorological inputs to S2M used for operational forecasting have primarily been based on the output of the deterministic numerical weather prediction (NWP) model ARPEGE operated by Météo-France with a lead-time of 2 days. This setup only provides relevant information for the day after the bulletin has been issued and does not take into account errors originating from the meteorological forecast itself or intra-massif variability of snowpack properties within a given elevation band, incline category and aspect sector.

Ensemble forecasting is increasingly used for meteorology and hydrology applications. Several major meteorological centers such as the European Center for Medium-Range Weather Forecasts (ECMWF, Molteni et al., 1996), the National Centers for Environmental Prediction (NCEP, Toth and Kalnay, 1997) and the Canadian Meteorological Centre (CMC, Pellerin et al., 2003) have developed operational ensemble forecasting systems with global NWP models for medium-range prediction accounting for uncertainties related to synoptic and large scales. These ensemble meteorological forecasts are sometimes used as input of hydrological models for medium-term forecast of river discharges (e.g. Thirel et al., 2008, 2010; Voisin et al., 2011). For several years, ensemble forecasting has also become an essential tool for forecasting high-impact weather events two or three days in advance. Reliable information about uncertainties in the localization and the intensity of these events is crucial for issuing meaningful warnings. This requires to keep increasing the horizontal and vertical resolutions of ensemble global systems and to keep improving their perturbation methods. Ensemble systems based on limited-area NWP models are also increasingly used for that purpose (e.g. Bowler et al., 2008; Frogner et al., 2006; Marsigli et al., 2008). The benefits of multimodel weather forecasts are also currently explored through projets such as the THORPEX Interactive Grand Global Ensemble (TIGGE, Bougeault et al., 2010).

To the best of our knowledge, ensemble forecasting has so far not been applied to avalanche hazard forecasting, despite the fact that the

* Corresponding author.
E-mail address: matthieu.lafaysse@meteo.fr (M. Lafaysse).
[1] Now at Météo-France, DSI/OP, Toulouse, France.

http://dx.doi.org/10.1016/j.coldregions.2015.04.010
0165-232X/© 2015 Elsevier B.V. All rights reserved.

large sensitivity of the snowpack to meteorological conditions and the numerous threshold effects make its prediction challenging, especially in mountain regions. Snowpack modeling and avalanche hazard warning therefore provide an excellent opportunity for the application of ensemble-based prediction techniques. In this study, we introduce an ensemble forecasting system bringing together tools and data operated by Météo-France. The performance of the system is comprehensively assessed by comparing its output with the current deterministic prediction system and the analysis of snow conditions using height of 24-hour new snow (HN24) and a dedicated regional scale natural avalanche hazard index (NHI) as target variables.

2. Models

2.1. SAFRAN SURFEX/ISBA-Crocus MEPRA model chain

The generation of consistent meteorological input data for the numerical snowpack simulation is carried out by the meteorological downscaling and surface analysis tool SAFRAN (French acronym for Système d'Analyse Fournissant des Renseignements Adaptés à la Neige, Durand et al., 1993, 1999). SAFRAN operates at the geographical scale of meteorologically homogeneous mountain ranges (so-called "massifs", Fig. 1) within which meteorological conditions are assumed to depend only on altitude and aspect. For the analysis of meteorological surface fields, the guess used by SAFRAN consists of vertical atmospheric profiles from NWP models. A robust assimilation scheme corrects the initial guess based on ground-based and radiosonde observations as well as remotely-sensed cloudiness. Thus, SAFRAN provides hourly meteorological conditions for each massif for 300 m-spaced elevation bands. Variables covered by SAFRAN do not only include precipitation (rainfall and snowfall rate) and air temperature, but also relative humidity, wind speed, incoming longwave and shortwave radiation. SAFRAN also has a forecast mode, in which it solely uses NWP output (vertical atmospheric profiles and precipitation fields). In this mode, SAFRAN is simply a downscaling tool that converts the NWP model grid to the massifs/altitude bands geometry. SAFRAN outputs feed the detailed snowpack model SURFEX (Surface Externalisée)/ISBA (Interactions between Soil Biosphere and Atmosphere)-Crocus (Vionnet et al., 2012), which computes the energy and mass balance to simulate the evolutions of the physical properties of a multi-layer snowpack and the underlying ground as a function of altitude, slope and aspect within each massif (slopes of 0, 20 and 40° for 8 aspects: North, North-East, East, South-East, South, South-West, West and North-West). The mechanical stability of the snowpack simulated by Crocus is subsequently estimated by the model MEPRA (Modèle Expert d'aide à la Prévision du Risque d'Avalanche, Giraud, 1992). First, the shear resistance/shear stress ratio (Föhn, 1987) is computed for each layer of each simulated snowpack. An expert approach then associates natural and accidental avalanche

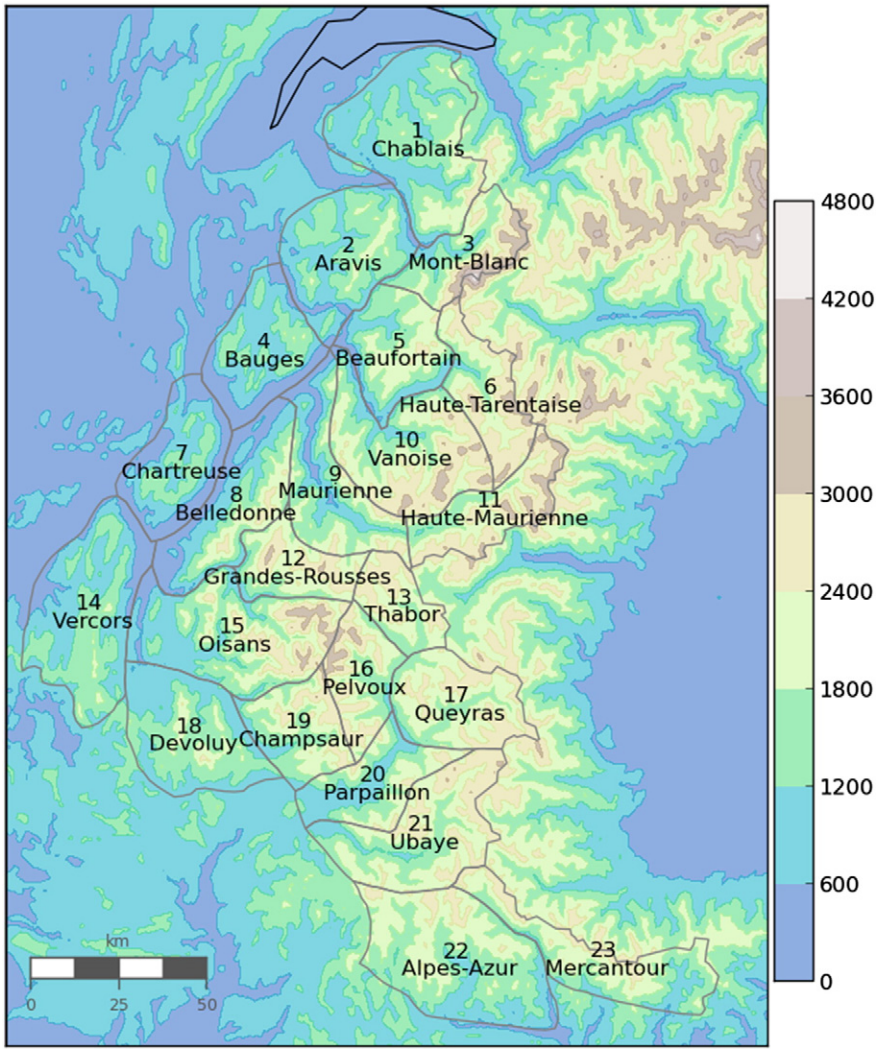

Fig. 1. Map of the 23 French alpine massifs, with altitude in meters.

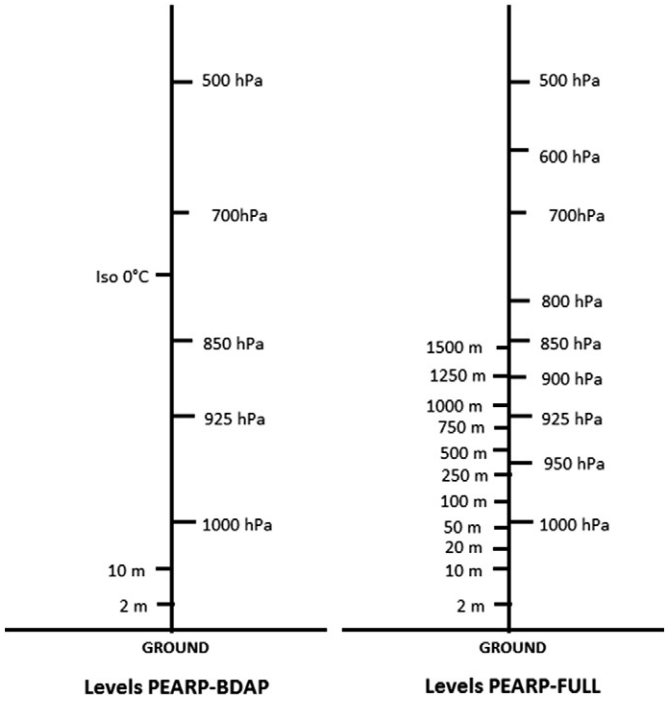

Fig. 2. Atmospheric levels available in PEARP-BDAP and PEARP-FULL database: pressure levels and height levels (height above ARPEGE grid altitude). The deterministic simulations presented in this paper use the same levels as PEARP-BDAP although deterministic operational forecasts use the same levels as PEARP-FULL.

hazard levels to each simulated snowpack profile taking into account the levels of mechanical instability, the thickness of the unstable snow and the temporal evolution of the shear resistance/shear stress ratio. Wet snow conditions are assessed by a separate algorithm. MEPRA also provides an integrated natural avalanche hazard index at the massif scale (NHI) ranging from 0 (lowest level) to 8 (highest level; Martin et al., 2001). This index is a combination of MEPRA natural hazard indexes at different altitudes from 1500 to 3000 m for 40° slopes and 8 aspects. It was designed to be comparable to a massif scale index of observed avalanche activity defined by Giraud et al. (1987).

2.2. PEARP ensemble forecasting

In operational use, SAFRAN (thus S2M) is currently fed by the deterministic global NWP model ARPEGE (Action de Recherche Petite Echelle Grande Echelle, Courtier et al., 1991) operated by Météo-France. In this study, we use the 35 members of the ensemble prediction system based on ARPEGE (PEARP for Prévision d'Ensemble ARPEGE, Descamps et al., 2014) instead. This system is based on a 6-member ensemble assimilation (Berre et al., 2007) combined with the singular vectors perturbations method (e.g. Buizza and Palmer, 1995; Molteni et al., 1996) to provide 35 initial states — including an undisturbed control member — to the NWP model. The 34 other members are randomly run with 10 different parameterization sets of sub-grid processes including different physical schemes for vertical diffusion, shallow convection, deep convection and oceanic fluxes. The PEARP horizontal resolution is larger than ARPEGE deterministic configuration (about 15 km vs. 10 km over France, respectively), but the vertical resolution is nearly the same (65 levels with a top level at 50 km). PEARP covers 4.5 days lead time (108 h). The inputs of SAFRAN are not restricted to the surface variables of the NWP model but also include vertical profiles, which are key to properly represent altitude variations of meteorological conditions in mountainous terrain. Two different types of output of the PEARP are used in this study, which mainly differ in the number of vertical levels available. One could expect that a larger number of vertical levels would lead to better predictions. The first one (named PEARP-FULL) includes all the ARPEGE vertical levels currently used by the deterministic operational S2M chain (9 pressure levels and 10 height levels, Fig. 2). However, PEARP does not provide all of these levels operationally in real-time and their storage is limited to 6 months. The second output type of PEARP (PEARP-BDAP) refers to outputs that are available in real time from the BDAP operational database and could be used as such for a real-time forecasting system. While PEARP-BDAP does not suffer from time restricted data storage, it contains fewer vertical levels (only 5 pressure levels and no height levels except from surface variables). To mitigate against the low number of vertical levels in the PEARP-BDAP output, the height of the 0 °C isotherm predicted by ARPEGE was used to create an additional level with fixed temperature (0 °C) but variable height. The difference between PEARP-FULL and PEARP-BDAP can therefore not simply be reduced to a difference in number of levels. Using these two different types of PEARP outputs allows assessing whether PEARP-BDAP data could still be used operationally even with a different representation of altitude dependency of atmospheric variables than what is currently used in the operational deterministic S2M model

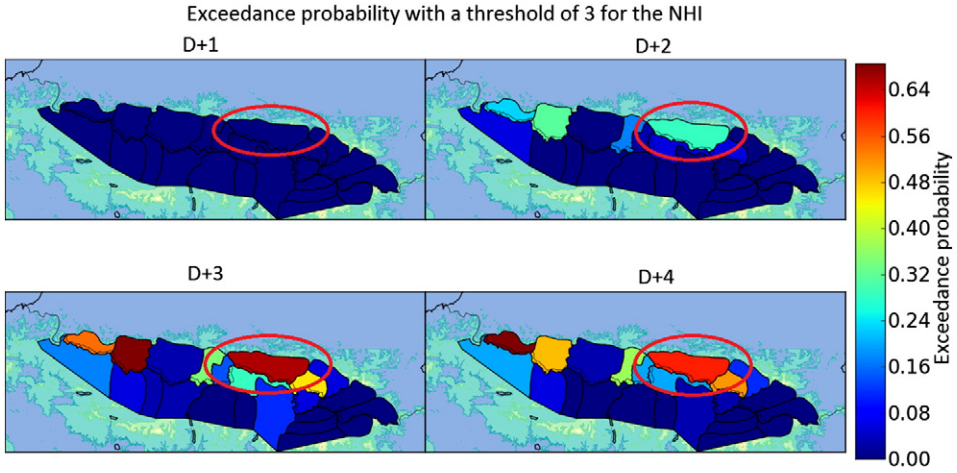

Fig. 3. Map of exceedance probability with a threshold of 3 for the massif level NHI over the Pyrenees. Forecast from the Jan. 13, 2013 at 6 UTC for D + 1 (Jan. 14, 2013 at 06 UTC), D + 2 (Jan. 15, 2013 at 06 UTC), D + 3 (Jan. 16, 2013 at 06 UTC), and D + 4 (Jan. 17, 2013 at 06 UTC). The circled massif is the Couserans massif (details on Figs. 4 and 5).

chain. To compare the PEARP-BDAP–S2M chain with the deterministic ARPEGE–S2M chain without any influence of the atmospheric levels used, we ran the deterministic version of S2M using only the same vertical levels as for PEARP-BDAP and extending the prediction lead time to 4 days instead of using the operational deterministic S2M outputs.

3. Data and evaluation methodology

3.1. Evaluation period and variables

The performance of the new PEARP-BDAP–S2M model chain was assessed and compared to results of the deterministic and the PEARP-FULL–S2M model chains with 4 days lead time. In both cases the reference used was the S2M chain in analysis mode in order to compare similar variables and because the spatial scale of simulated variables do not have observation equivalents. Quantitative evaluations were made using 2760 forecasts resulting from 120 analysis dates spanning the period from Nov. 1, 2013 to Mar. 1, 2014 (one 4-day forecast per day) for the 23 massifs of the Alps. During this winter, the snowpack was built up quickly and early at the beginning of November in the Northern Alps at the highest elevations. In the other massifs and at lower elevations, November snowfalls were less significant. An extended period of high pressure dominated the weather from the end of November to the end of December. Then, a number of snowfall events occurred until the end of February in all massifs and elevations. These snowfall events were never particularly intense, but very frequent, resulting in a total snow depth above the long-term climatological average in all massifs, especially in the Southern Alps. However, the natural avalanche activity was relatively low over all massifs due to a good settling between each snowfall event. There were no widespread persistent weak layers as observed during more accident-prone winters. The number of avalanche fatalities was relatively low with 21 compared to the long-term average of 32 per season.

The variables that were evaluated in this study are the height of 24-hour new snow as defined by Fierz et al. (2009) at 1800 m altitude (HN24) and the NHI described in Section 2.1. The HN24 variable can be easily derived from the S2M chain because the age of snow layers is a tracked variable in the SURFEX/ISBA-Crocus snow model and the dynamical evolution of the number and thicknesses of the numerical snow layers is designed not to aggregate new snowfall with older surface snow with significantly different physical properties (Vionnet et al., 2012). Estimating the uncertainty in the S2M produced HN24 caused by errors in the meteorological analysis and the snow model is difficult because of the lack of a spatialized observed reference. However, since the overall analyzed snow depth is not biased at the Alps scale (Lafaysse et al., 2013), we can expect that HN24 is not highly biased either. The evaluation of NHI is even more difficult because human observations of avalanches are not reliable, especially during bad weather. However, Martin et al. (2001) compared the simulated NHI with an equivalent index of observed avalanche activity derived from the number of observed events over 2 winters (1986–1987 and 1994–1995) and concluded that the simulated index is a good synthesis of the natural snow cover instability and avalanche hazard at the massif scale.

Since the period used for the statistical evaluation does not include any severe natural avalanche events, we complemented our analysis with a more qualitative assessment of the ensemble forecasting system during challening situations that occurred in the Pyrenees during the months of January, February (major avalanche cycle) and June 2013 (significant snowmelt event).

3.2. Scores

The fundamental difference between deterministic and ensemble forecasting in the forecast of a given binary event, is that the deterministic system predicts the occurrence or no occurrence of the event while the ensemble system predicts the occurrence probability of the event. Fundamentally, a probability cannot be validated or invalidated by a single observation. As a consequence, it is not possible to assess the quality of an individual probabilistic forecast, except in some extreme cases, such as when the observation falls well outside the range of the predicted values. The assessment can only be statistical, and performed on a sufficiently large number of realizations of the prediction system.

The scores used for the evaluations in the present study are both deterministic and probabilistic: the Root Mean Square Error (RMSE) is a classical deterministic score to compare the forecasts for N days with reference values r_k for each day k. When applied to ensemble forecasting, it requires the restriction of the ensemble of forecasts to their average m_k.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{k=0}^{N}(m_k - r_k)^2}.$$

The ensemble dispersion D represents the standard deviation of the members relatively to their average:

$$D = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\frac{1}{n}\sum_{i=1}^{n}\left(x_{k_i} - m_k\right)^2}$$

where x_{k_i} is the forecast of member i for day k and n the number of ensemble members. The dispersion is expected to be of the same magnitude as the RMSE if the ensemble correctly captures the forecast uncertainty.

The most common probabilistic score is the Brier Score (Brier, 1950), which describes the ensemble forecast system performance in terms of a given threshold exceedance. The Brier Score is defined by:

$$BS = \frac{1}{N}\sum_{k=1}^{N}(y_k - o_k)^2 \quad \text{with} \quad 0 \le BS \le 1$$

where y_k and o_k are the forecasted probability of the event for the day k, and the corresponding binary observation ($o_k = 0$ or 1) respectively. The Brier Score ranges from 0 to 1, where 0 corresponds to a perfect score. In this study, the chosen exceedance thresholds to define an event are 10 cm for HN24 and 1 for NHI. They are relatively low because the season 2013–2014 exhibited too few intense events to obtain statistically significant samples using higher thresholds.

The Brier Score can been decomposed in three terms, the reliability BS_{rel}, the resolution BS_{res} and the uncertainty BS_{unc} (Murphy, 1973):

$$BS = BS_{\mathrm{rel}} - BS_{\mathrm{res}} + BS_{\mathrm{unc}}.$$

This algebraic decomposition is based on separating all possible forecast probabilities in J classes. A probability Y_j is associated with each class j. N_j is the number of days for which the forecast probability is Y_j. In our study, the most detailed decomposition uses $J = 36$ classes based on the 35 ensemble members. We define the observed occurrence frequency of the event in the class j as:

$$\overline{o_j} = p(o = 1 | Y_j) = \frac{1}{N_j}\sum_{k \in N_j} o_k$$

with $o_k = 1$ if the event is observed for day k, $o_k = 0$ if not. Here, the sum only applies to days with the forecast probability Y_j. We then define the overall occurrence frequency of the event by:

$$\overline{o} = \frac{1}{N}\sum_{k=1}^{N} o_k.$$

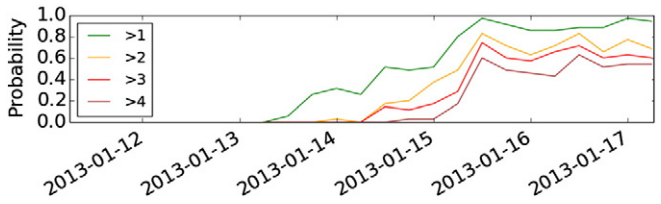

Fig. 4. Time evolution of exceedance probability forecast for different thresholds of NHI for the Couserans massif from the Jan. 13, 2013 at 06 UTC to Jan. 17, 2013 at 06 UTC.

The reliability BS_{rel} represents the ability to predict good probabilities. It compares the predicted probabilities to the observed occurrence frequencies in each class:

$$BS_{\mathrm{rel}} = \frac{1}{N}\sum_{j=1}^{I} N_j \ (Y_j - \overline{o_j})^2.$$

Perfect reliability occurs when for any class an event is predicted N_j times with the probability Y_j and it is observed $Y_j \times N_j$ times within these N_j days. This corresponds to $BS_{rel} = 0$.

While a system that always uses the climatological probability to predict the likelihood of an event would be perfectly reliable under the previous definition (i.e. the forecast probability Y_j is always equal to $\bar{o}$ so $N_j = N$ for a unique class where $Y_j = \overline{o_j} = \bar{o}$ and $N_j = 0$ for other classes), this approach does not provide any practical value. A second necessary property for a meaningful forecast system is its ability to predict different probabilities from one forecast to another. This property can be assessed by the resolution BS_{res} which compares the frequencies of the event in each probability class to the overall occurrence frequency:

$$BS_{\mathrm{res}} = \frac{1}{N}\sum_{j=1}^{I} N_j \ (\overline{o_j} - \bar{o})^2.$$

A purely climatological forecast corresponds to the minimal resolution: $BS_{\mathrm{res}} = 0$ (again because $N_j = N$ for a unique class where $Y_j = \overline{o_j} = \bar{o}$ and $N_j = 0$ for other classes). The maximal resolution corresponds to a perfect deterministic forecast where the forecast probabilities are $Y_1 = 0$ or $Y_{36} = 1$ with $\overline{o_1} = 0$, $\overline{o_{36}} = 1$, $N_1 = (1 - \bar{o})N$, $N_{36} = \bar{o}N$. In such a case, we obtain $BS_{\mathrm{res}} = \bar{o}(1 - \bar{o})$

The uncertainty BS_{unc} in the Brier Score decomposition corresponds to the intrinsic difficulty to forecast the occurrence probability of the event, which is maximal for an event occurring 50% of the time. It is independent of the forecasting system.

$$BS_{\mathrm{unc}} = \bar{o}(1 - \bar{o})$$

BS_{unc} is equal to the maximal value of the resolution in the case of the perfect deterministic forecast.

The Brier Skill Score (BSS) offers the possibility to compare the Brier Score of an ensemble forecast system with the Brier Score of a reference system:

$$BSS = 1 - \frac{BS}{BS_{\mathrm{ref}}}.$$

BSS values can range between −∞ and 1 and are equal to 0 if both systems have the same skill. BSS was used to compare the PEARP-BDAP chain with the deterministic and the PEARP-FULL chains, using respectively the deterministic and the PEARP-FULL chains to compute BS_{ref}. In the case of the deterministic forecast, we assumed that the model only predicts probability values of 0 and 1. This approach is similar to the method used by DeMaria et al. (2009), who computed Brier Skill Scores for comparing the skill of a Monte Carlo probabilistic model with the official deterministic model of the operational forecast centers in terms of wind speed probability in tropical hurricanes in both Atlantic and Pacific oceans. A probabilistic criteria is obviously more favorable towards ensemble systems, but classical deterministic scores reducing ensembles to their average are conversely in favor of deterministic systems, ignoring a large part of the information included in the ensemble.

3.3. Robustness of evaluations

Because of the limited common availability of both PEARP-BDAP and PEARP-FULL archives, the dataset available to evaluate the performance of the ensemble forecast system is far smaller than recommended by Candille and Talagrand, 2005 (order of magnitude of 10^5 days for a 35 members system). One way to address this problem is to make an ergodic hypothesis, which assumes that the predicted variables (random variables) have the same behavior averaged over time (in one point of the domain) as averaged over space. This assumption is commonly made to evaluate ensemble forecasting systems even if not completely verified (e.g., Descamps et al., 2014; Molteni et al., 1996) as it nonetheless provides useful information in lieu of a statistical framework able to deal with such samples in a fully consistent manner. Under this assumption the 120 forecasts of each massif can be considered independent and pooled into one single sample of $120 \times 23 = 2760$ independent forecasts. This pooled sample is still smaller than the recommanded one but comparable to classical sample sizes in other ensemble forecasting studies (e.g., Rousset-Regimbeau, 2007). However, this method of pooling the sample prevents the analysis of any spatial pattern of skills.

To evaluate the robustness of the calculated Brier Scores at both the general and massif scales (i.e. with 2760 or 120 forecasts, respectively), confidence intervals were computed using the Bootstrap method (Efron, 1979). Bootstrapping is a statistical method based on random sampling with replacement, which can be used to derive a measure of accuracy of any statistical estimate based on a finite sample. Considering a sample $x = (x_1, \ldots, x_N)$ of an independent and identically distributed population and an estimator θ of which we want to characterize the

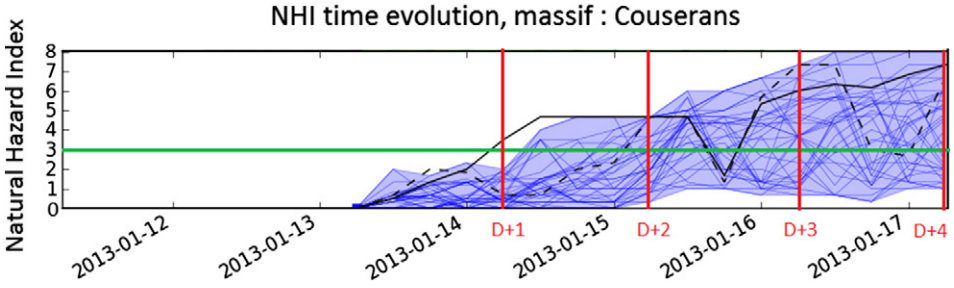

Fig. 5. NHI time evolution forecast for the Couserans massif from the Jan. 13, 2013 at 06 UTC to Jan. 17, 2013 at 06 UTC. Blue lines: 35 members of PEARP-BDAP-S2M; black solid line: S2M analysis (from the Jan. 11, 2013 at 06UTC); black dash line: ARPEGE-S2M deterministic forecast.

confidence interval, the bootstrap technique consists in forming B new so called bootstrap samples $x_1, \ldots, x_N$ of size N from the initial sample using sampling with replacement. The estimator can be computed for each of these samples, which provides a distribution and the confidence interval of θ with the estimator variance:

$$\sigma^2 = \frac{1}{B}\sum_{i=1}^{B}\left(\hat{\theta}_i - \hat{\theta}\right)^2, \text{ with } \hat{\theta} = \frac{1}{B}\sum_{j=i}^{B}\hat{\theta}_j.$$

In this study, we used the algorithm developed by Ferro (2007) with $B = 1000$ bootstrap samples for the unconditional expected Brier Score (i.e., resampling of dates) in the simplest case where the number of members is fixed ($n = 35$). The comparison of the scores $\hat{\theta}_{FULL}$ and $\hat{\theta}_{BDAP}$ obtained for the two forecast systems on the 2760 sample values was done with a Student t-test on the bootstrap samples. If the H_0 hypothesis that the means of the two bootstrap populations are equal is true, then the random variable

$$T = \frac{\left|\hat{\theta}_{FULL} - \hat{\theta}_{BDAP}\right|}{\sqrt{s^2\left(\frac{1}{n_{FULL}} + \frac{1}{n_{BDAP}}\right)}}, \text{ with}$$

$$s^2 = \frac{(n_{FULL}-1)\sigma^2_{FULL} + (n_{BDAP}-1)\sigma^2_{BDAP}}{n_{FULL} + n_{BDAP} - 2}$$

follows a Student's t distribution with $n_{FULL} + n_{BDAP} - 1$ degrees of freedom. In our case, the sample sizes are equal ($n_{FULL} = n_{BDAP} = 2760$). The comparison of the associated p-value with theoretical value from the Student table allows us to accept or reject the H_0 hypothesis.

To evaluate the robustness of the Brier Scores spatial pattern, we tested the significance of the inter-massifs variability by a one-way Analysis Of Variance (ANOVA, Scheffé, 1959), which is a generalization of the Student t-test for more than two samples. The H_0 hypothesis is that the mean of the Brier Score among the Bootstrap samples is equal among all massifs. This hypothesis was tested using a Fischer test.

4. Results and discussion

4.1. Example

An example of a forecast of the PEARP-BDAP-S2M chain is presented in Fig. 3, which shows a map of exceedance probabilities with a threshold of 3 for the NHI over the Pyrenees in January 2013. In this particular case, high probabilities of important avalanche events were forecast for many French massifs, especially at a 3 days lead time. Fig. 4 illustrates the time evolution of exceedance probabilities for different NHI thresholds for the Couserans massif (highlighted by the red circle in Fig. 3). During the day of Jan. 16, 2013, numerous big avalanches indeed occurred, resulting in roads break and important damages in many areas of the Pyrenees. The corresponding time evolution of the 35 members for the Couserans massif is illustrated in Fig. 5. The number of forecast members over a given threshold explains the computed probabilities of Fig. 4. The number of members over the threshold 3 (green line), for example, determines the probabilities reported in Fig. 3 for different prediction lead times. As generally expected for any ensemble forecasting method, the dispersion between members increases with lead time and the deterministic forecast (black dash line) sometimes differs significantly from the ensemble mean.

4.2. Evaluation of the PEARP-BDAP-S2M chain over French Alps

4.2.1. General evaluation on the whole French Alps

The general evaluation of the PEARP-BDAP–S2M chain over the French Alps is summarized in Table 1 for HN24 and in Table 2 for NHI. The scores assessing the quality of the forecasts of the PEARP-BDAP–S2M chain (Brier Score, reliability, resolution, uncertainty, RMSE and dispersion) show that this model chain features promising forecast skills at all lead times. The low Brier Score (between 0.06 and 0.10) indicates a good forecast of threshold exceedance probabilities. However, for both variables, the dispersion is lower than the RMSE which shows an under-dispersion of the system. The deterministic scores such as RMSE are slightly better for the ensemble mean than for the deterministic forecast. From a probabilistic point of view, the ensemble forecast is far better with BSS ranging from 0.15 to 0.25. To better illustrate the quality of the forecast probabilities, all forecasts were divided into eight classes of forecast probabilities and the observed occurrence frequency of the event was computed for each class. The reliability diagrams presented in Figs. 6 and 7 show the relationship between forecast probability and observed frequency for different HN24 and NHI thresholds and lead times. The diagonal corresponds to a perfect ensemble forecast system. For both variables, the probabilities appear slightly underestimated, especially for probabilities below 0.7. In general, however, the forecast probabilities are quite reliable, even at D + 4. Another satisfying behavior is that the forecast probabilities match very well with the observed frequencies when extremely low or high probabilities are forecasted (none of the members or all members predict an event respectively).

Table 1
Summary of the different performance scores computed for the forecast of HN24 (with a threshold of 10 cm for the Brier Score) for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC and for the entire French Alps from Nov. 1, 2013 to Mar. 1, 2014.

Score	D + 1	D + 2	D + 3	D + 4
Brier	0.07	0.07	0.08	0.09
Reliability	0.01	0.01	0.01	0.01
Resolution	0.06	0.06	0.04	0.03
Uncertainty	0.12	0.12	0.12	0.12
BSS	0.21	0.25	0.24	0.23
Dispersion (cm)	1.9	2.2	2.6	3.3
RMSE (cm)	4.0	4.2	4.9	5.1
RMSE deterministic (cm)	4.1	4.3	4.9	5.3

4.2.2. Evaluation on massifs scale

The results are not homogeneous in space as illustrated by the map of Brier Scores for the NHI (Fig. 8). The observed spatial patterns of skill depend on scores and lead time. Brier Scores appear lower in the Eastern central massifs (from Haute-Tarentaise to Ubaye). This behavior may be specific to the particular year as these massifs experienced less snowfall than the others. The bootstrap method showed that the assessment of these scores at the massif scale is less robust than at the scale of the entire French Alps: the average width of confidence intervals at the massif scale is 0.09 in comparison to 0.015 for the entire French Alps. This observation is further illustrated by the different dispersions of Brier Scores shown with boxplots in Fig. 9. Most importantly, however, the observed differences in Brier Scores between massifs are significant (no-overlapping confidence intervals even for contiguous massifs, as illustrated on Figs. 8 and 9). The p-values obtained with the ANOVA

Table 2
Summary of the different performance scores computed for the forecast of NHI (with a threshold of 1/8 for the Brier Score) for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC and for the entire French Alps from Nov. 1, 2013 to Mar. 1, 2014.

Score	D + 1	D + 2	D + 3	D + 4
Brier	0.06	0.08	0.09	0.10
Reliability	0.01	0.01	0.01	0.01
Resolution	0.06	0.05	0.04	0.03
Uncertainty	0.11	0.12	0.12	0.12
BSS	0.15	0.22	0.25	0.24
Dispersion	0.28	0.30	0.35	0.44
RMSE	0.51	0.60	0.67	0.69
RMSE deterministic	0.53	0.63	0.70	0.75

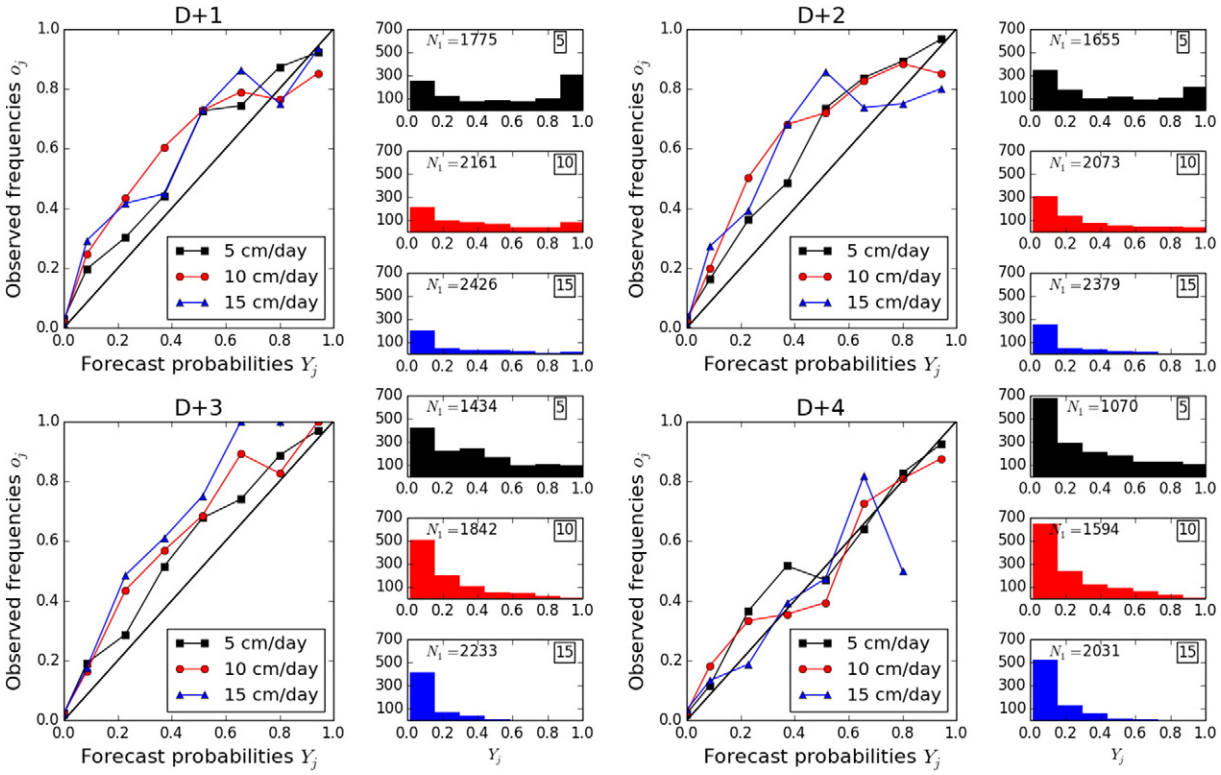

Fig. 6. Reliability diagrams of the forecasted probabilities of HN24 with different exceedance thresholds (black: 5 cm; red: 10 cm; blue: 15 cm) and for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC. The histograms represent the bin sizes for each threshold and lead time. N_1 is the bin size for the probability $Y_1 = 0$. Evaluations for the entire French Alps from Nov. 1, 2013 to Mar. 1, 2014.

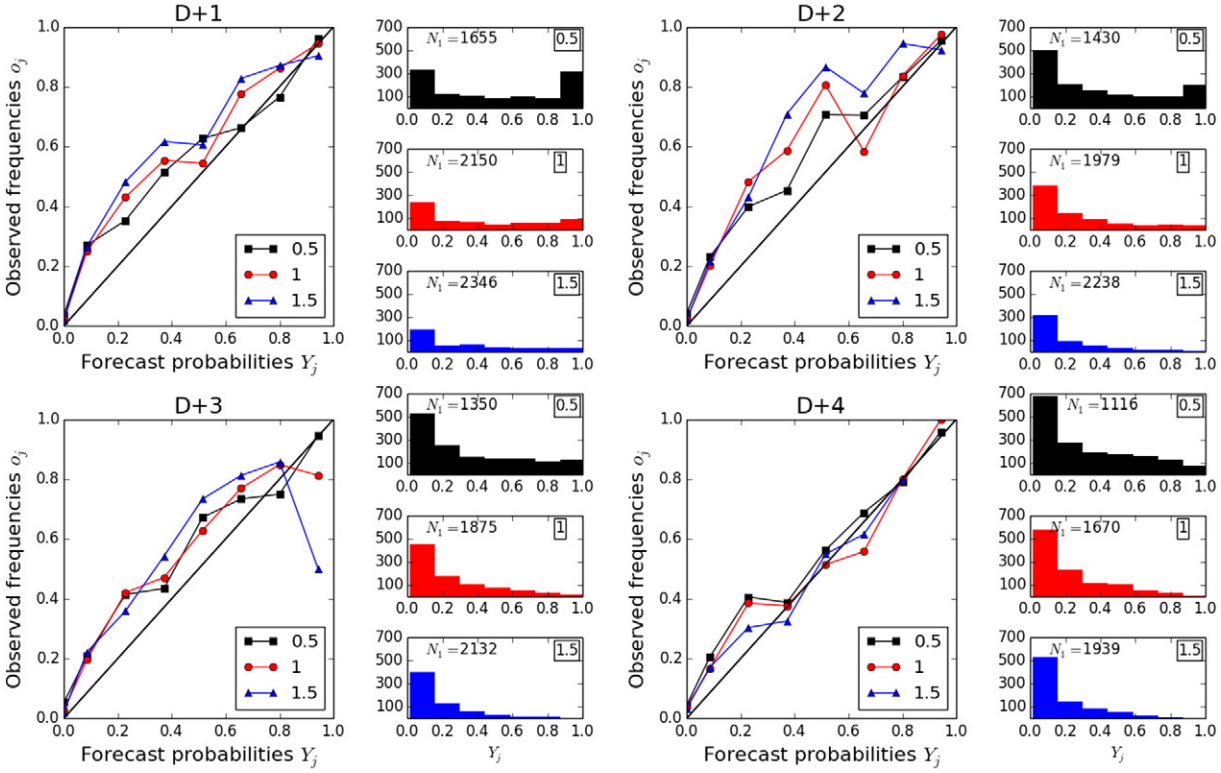

Fig. 7. Reliability diagrams of the forecasted probabilities of NHI with different exceedance thresholds (black: 0.5; red: 1; blue: 1.5) and for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC. The histograms represent the bin sizes for each threshold and lead time. N_1 is the bin size for the probability $Y_1 = 0$. Evaluations for the entire French Alps from Nov. 1, 2013 to Mar. 1, 2014.

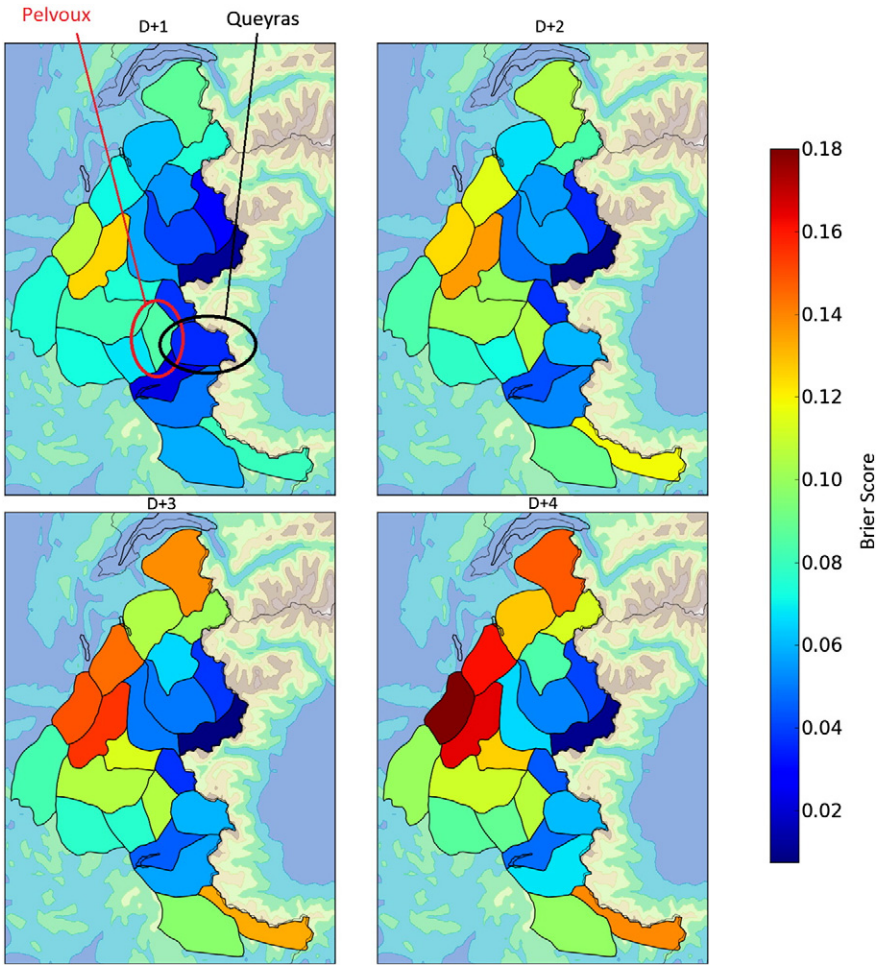

Fig. 8. Brier Score by massif for NHI forecasts with a threshold of 1 for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC. Evaluation from Nov. 1, 2013 to Mar. 1, 2014.

are very low (<0.001 at all lead times), which means that the null hypothesis of equality of means between massifs can be rejected with high confidence, despite the uncertainties around the evaluation sample size. Despite the observed differences among massifs, there are only few massifs where the ensemble chain does not present better skills than the deterministic system regardless of lead time (Fig. 10). These differences are also statistically significant.

The decomposition of the Brier Score by massif is more uncertain than the BS and BSS assessment. The reliability diagrams at the massif scale are highly affected by the sampling (not shown) because the number of forecasts and observations in each probability class is too small.

The results at the massif scale for HN24 are very similar and therefore not presented in this section.

4.3. Sensitivity to the number of available vertical levels

All scores were also computed for the PEARP-FULL–S2M ensemble forecasting system, which incorporates more atmospheric vertical levels. The difference between the Brier Score of both systems ranges from −0.001 to +0.003 for HN24 and NHI, while the Brier Skill Scores of PEARP-FULL–S2M using PEARP-FULL-BDAP as reference are close to zero (Table 3). Brier Scores are much closer between PEARP-FULL–S2M and PEARP-BDAP–S2M than between PEARP-BDAP–S2M and the deterministic ARPEGE–S2M. For HN24, PEARP-FULL–S2M exhibits slightly poorer Brier Scores whereas for NHI the sign of the difference depends on lead time. Even though the observed differences in Brier Scores are statistically significant (Student *t*-test; see Table 3 for p-values), they are of limited relevance since they are very small in comparison to the actual scores (10 to 100 times smaller).

When looking at the differences in Brier Scores within each massif (Fig. 11 for NHI at D + 1), the medians of the bootstrap samples range from −0.005 to 0.01 with a wide spread. At a 95% confidence level, the differences are significantly different from zero for 11 to 17 massifs depending on lead time (Table 4). The variability of these differences among massifs is often lower than the uncertainty of the scores assessment linked to the sampling. Therefore, it would not be meaningful to display a spatial pattern of these differences, although the ANOVA revealed significant differences of Brier scores among massifs for all lead times (all p-values <0.001).

Overall, these results show that the sensitivity of the PEARP–S2M chain to the number of available vertical levels is statistically significant but sufficiently low to conclude that the levels available in the PEARP-BDAP are sufficient to drive S2M.

4.4. Case studies over Pyrenees

The period between Jan. 16 to Jan. 18, 2013 was exceptional in the Pyrenees with large amounts of new snow (60 cm to 1 m) and severe avalanches releasing naturally. Fig. 12 illustrates the forecasts of the PEARP-BDAP–S2M system for snow depth and NHI during this exceptional event for the massifs Aspe-Ossau (western part of Pyrenees) and Couserans (eastern part of Pyrenees). For the Aspe-Ossau massif, the event would have been anticipated by the ensemble forecast system from January 13th, but underestimated. The most pessimistic member always simulated snow depth 20 cm below S2M analysis and only few

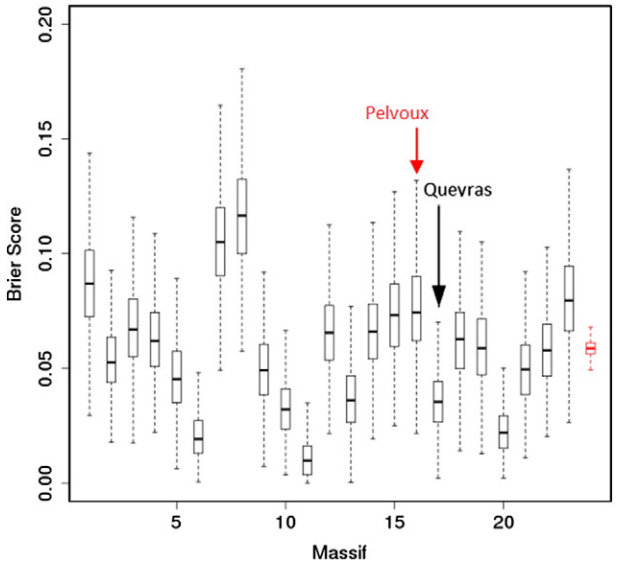

Fig. 9. Dispersion of Brier Score obtained by bootstrap samples of each massif (black boxplots) and of all massifs together (red boxplot). Brier Scores were computed for D + 1 NHI forecasts at 06 UTC with a threshold of 1. Evaluation from Nov. 1, 2013 to Mar. 1, 2014. The x-axis massif indices correspond to the numbers displayed on Fig. 1. The bottom and top of the boxes represent the first and third quartiles, the band inside the box is the median, the ends of the whiskers represent the minimum and maximum among all bootstrap samples.

Table 3
Differences of Brier Scores between PEARP-FULL-S2M and PEARP-BDAP-S2M among massifs for HN24 with the 10 cm threshold and for NHI with the 1/8 threshold (p-values of the Student *t*-test of the significance of this difference in brackets); Brier Skill Scores of PEARP-FULL using PEARP-BDAP as reference. Both assessments are calculated for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC and the entire French Alps from Nov. 1, 2013 to Mar. 1, 2014.

		D + 1	D + 2	D + 3	D + 4
Differences	HN24	−0.001 (<0.001)	−0.0002 (<0.001)	−0.001 (<0.001)	−0.001 (<0.001)
	NHI	0.003 (<0.001)	−0.0006 (0.001)	−0.0007 (<0.001)	0.0004 (0.023)
BSS	HN24	−0.016	−0.030	−0.012	−0.017
	NHI	0.046	−0.008	−0.008	0.005

members forecasted NHI values as high as the analyzed ones. As we do not have any evidence of significant errors in the S2M analysis for this case, it is very likely that both forecasting systems underestimate the real snowfall amount 3 days before the event. The forecasts for the Couserans massif are more satisfactory as the analyzed snow depth is covered by the ensemble dispersion. In some cases (especially at the end of this period), the deterministic model forecasted a significant decrease of NHI while the analyzed NHI remained very high. In the ensemble forecast, the dispersion is very high but some members clearly suggest that the NHI would remain high. This is an example where the probabilistic forecast system is able to provide more relevant information than the deterministic one. The forecast of the probabilistic chain would have obviously included uncertainty, but the deterministic

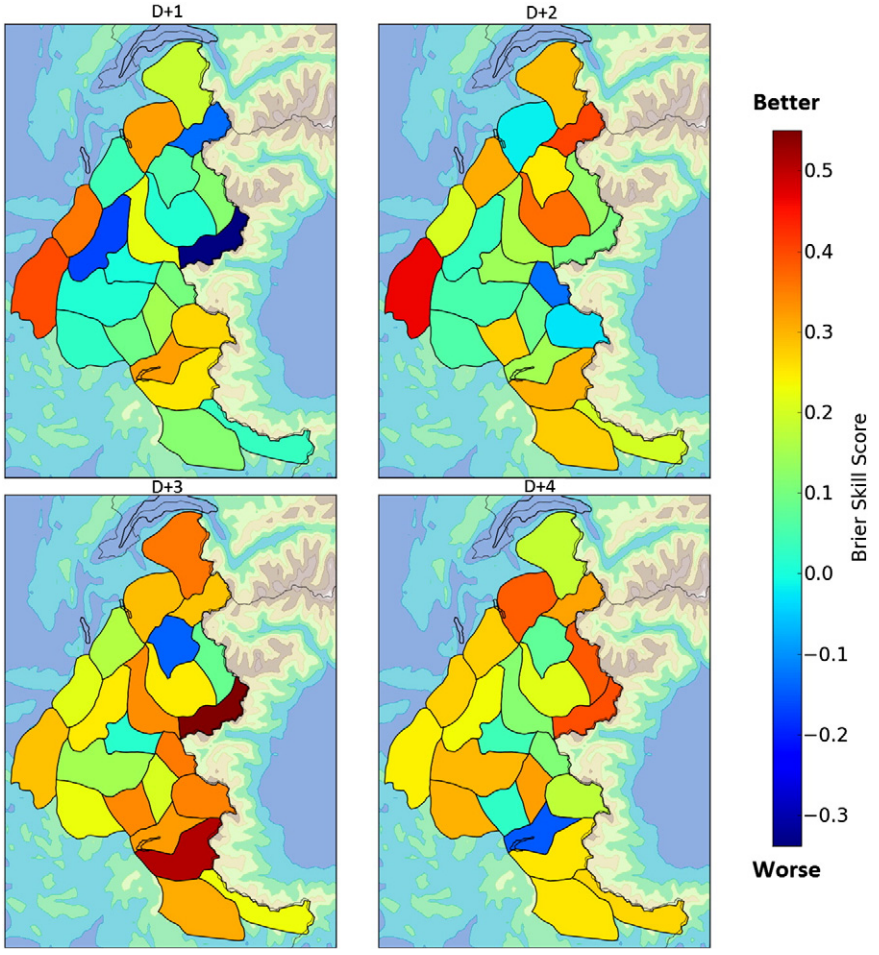

Fig. 10. Brier Skill Score by massif for NHI forecasts with a threshold of 1 and with deterministic forecast as reference, for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC. Evaluation from Nov. 1, 2013 to Mar. 1, 2014.

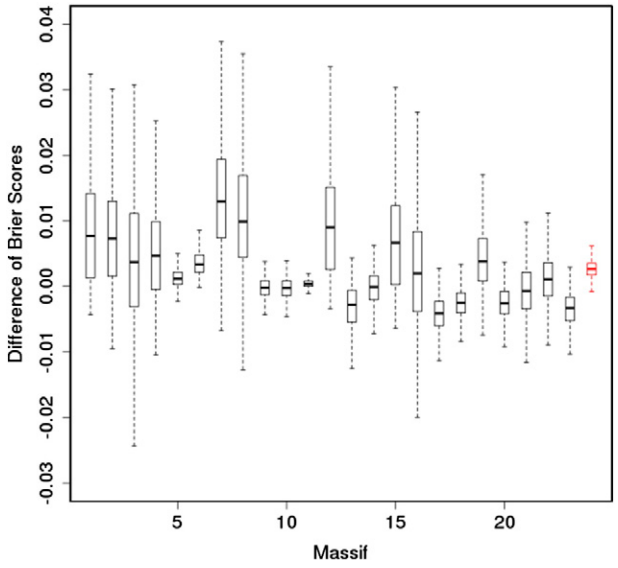

Fig. 11. Dispersion of the difference of Brier Scores between PEARP-BDAP-S2M and PEARP-FULL-S2M obtained by bootstrap samples of each massif (black boxplots) and of all massifs together (red boxplot). Brier Scores were computed for D + 1 NHI forecasts at 06 UTC with a threshold of 1. Evaluation from Nov. 1, 2013 to Mar. 1, 2014. The x-axis massif indices correspond to the numbers displayed on Fig. 1. The bottom and top of the boxes represent the first and third quartiles, the band inside the box is the median, the ends of the whiskers represent the minimum and maximum among all bootstrap samples.

chain would have simply been wrong assuming that the S2M analysis represents the "truth". In other cases, the deterministic forecast is closer to the analysis than all of the ensemble members (such as the forecast for Jan. 15 initialized on Jan. 12 for the Aspe-Ossau massif). If observed frequently, this behavior could be partly explained by the better resolution of the deterministic NWP model, which also explains the differences between the deterministic run and the undisturbed control run of the ensemble. However, the positions of any member of the ensemble forecast (including the control run) as well as the deterministic run are random variables. In our case, the deterministic run theoretically has a probability of 1/35 to exceed the forecasts of all ensemble members.

On Jun. 17 and Jun. 18, 2013, large amounts of precipitation (60 to 160 kg m^{-2}) were observed over the western and central parts of the Pyrenees. Since the conditions prior to the event were promoting snowmelt (high temperatures, strong winds, high solar radiation, cloudy nights preventing snow refreezing) in the context of a snow cover exceptionally important for the season above 2000 m, the precipitation event resulted in exceptional floods. A detailed understanding of the behavior of the snowpack is extremely important for the prediction of these events since snowmelt can significantly contribute to the saturation of a catchment. In this particular case, the damages from the flood event were much larger than expected solely based on precipitation amounts. Fig. 13 shows the time evolution of S2M analyzed snow water equivalent (black line) over the Luchonnais massif at 2400 m. The snowpack lost 250 mm water equivalent within one week. The figure shows that the intense snowmelt would have been anticipated well by the ensemble forecast, even at the farthest lead times. The dispersion between members is much lower than in previous plots, illustrating the better predictability of these type of events in comparison to winter snowfalls.

Table 4
Statistical tests describing the differences of massif scale BS values between PEARP-FULL–S2M and PEARP-BDAP–S2M. Number of massifs where the Student test reject the null hypothesis at the 95% confidence level for HN24 with the 10 cm threshold and for NHI with threshold 1/8, for four prediction lead times (D + 1, D + 2, D + 3, D + 4) at 06 UTC. Evaluation from Nov. 1, 2013 to Mar. 1, 2014.

	D + 1	D + 2	D + 3	D + 4
HN24	17	14	11	11
NHI	16	12	13	13

5. Conclusions and outlook

To the best of our knowledge, this study implemented an ensemble forecasting system dedicated to predict avalanche hazard for the first time. The implementation combines the meteorological ensemble forecasting system recently developed by Météo-France (PEARP) with the new model chain used to predict avalanche hazard (SAFRAN–SURFEX/ISBA-Crocus–MEPRA). The results obtained in various mountain ranges and for various types of meteorological conditions including events with exceptionally high avalanche levels in the Pyrenees indicate that the ensemble forecasting system PEARP–S2M can offer a valuable contribution for the avalanche hazard warning activities due to its probabilistic methodology and increase in prediction lead time. The statistical scores computed at the scale of the entire French Alps are very satisfactory and robust as demonstrated by the Bootstrap method. The spatial pattern of the scores is significant although more affected by sampling. More robust spatial patterns may be obtained with longer evaluation periods. Even if not all the atmospheric levels provided by PEARP are stored in the operational database BDAP, the results show that the reduced number of available levels are sufficient to drive PEARP–S2M. This conclusion allows the evaluation of the PEARP-S2M chain to be extended to other seasons and higher exceedance thresholds in a near future. This is an important step, since the exceedance thresholds investigated in the present study had to be adjusted to account for the number of observed events in the investigated season but are too low to reliably evaluate the forecast skill of high impact avalanche situations.

According to our results, the main limitations of the PEARP–S2M chain are an underdispersion of the ensemble and a relative underestimation of the forecast probabilities of snowfall and avalanche events. One possibility to address these limitations in the product provided to avalanche forecasters is to apply statistical post-processing methods to calibrate the output probabilities. An overview of available state-of-the-art methods can be found in Gneiting (2014). Examples include the Bayesian Model Averaging (BMA) approach developed by Raftery et al. (2005) and the Ensemble Model Output Statistics (EMOS) technique of Gneiting et al. (2005). The EMOS approach, for example, predicts a probability density function for each variable of interest (e.g., Gaussian or Gamma law depending on the variable) whose parameters are computed from the forecast members. For a Gaussian law, the mean is a bias corrected affine function of the ensemble mean and the variance is a dispersion-corrected affine function of the ensemble variance. This type of method could be easily applied to PEARP–S2M forecasts. Furthermore, a non-parametric method is currently being developed to post-process PEARP meteorological forecasts. This method is based on Quantile Regression Forests (Meinshausen, 2006) and does not rely on the strong assumption of a universal law to represent the conditional distribution of errors. It also allows other predictors than ensemble members to be included in the forecast (e.g., month of year if bias depends on the season). Obviously, the calibration of these methods requires data covering significantly longer periods than the four month period covered in the current study. This will be possible in the future using the PEARP-BDAP archive.

Nevertheless, the technical environment and the development approach chosen make it possible to implement this model chain in real time and make its output available to avalanche forecasters, at least experimentally. This tool will help them to include uncertainty in their bulletins and to provide a more objective assessment of medium-

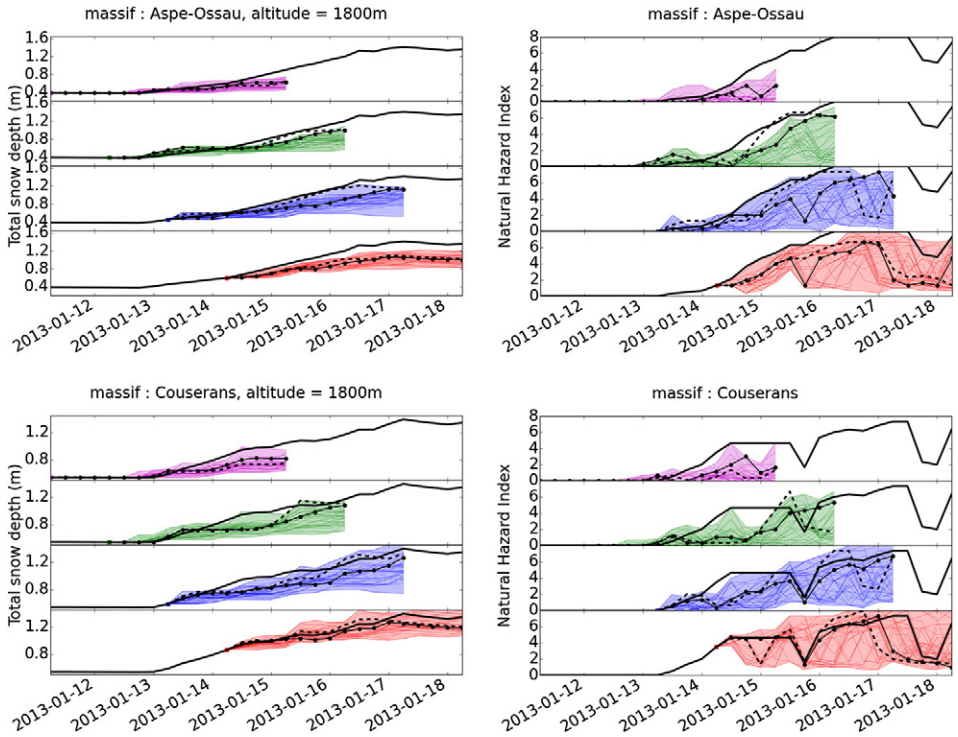

Fig. 12. Total snow depth (left) and NHI (right) time evolution for the Aspe-Ossau (top) and Couserans (bottom) massifs. The 4 forecasts (4 lines) start from the Jan. 11, 2013 at 6 UTC to Jan. 14, 2013 at 06 UTC and cover the four subsequent days respectively, with a 6 hour time resolution. Colored lines: 35 members of PEARP-BDAP-S2M; black solid line: S2M analysis; black dash line: ARPEGE-S2M deterministic forecast; black line with circles: PEARP-BDAP-S2M control run.

range trends. However, they will need to account for the current limitations of the system (underdispersion and relative underestimation of the forecast probabilities) in their interpretation of outputs. They should also consider the forecast probabilities of severe events with caution until the system has been evaluated using longer time periods and higher exceedance thresholds. As illustrated with the snowmelt example, we believe that other applications could benefit from this approach as well. Examples include hydrological prediction in mountainous catchments, meteorological forecast in mountainous regions and additional services yet to develop targeting the broader mountain community (e.g. ski resort managers).

Although the meteorological forecast is expected to be the main source of uncertainty in avalanche hazard forecasting, future developments of ensemble systems should also aim to include other uncertainties, such as uncertainties in initial conditions linked to analysis errors (including errors in observations), uncertainties associated with

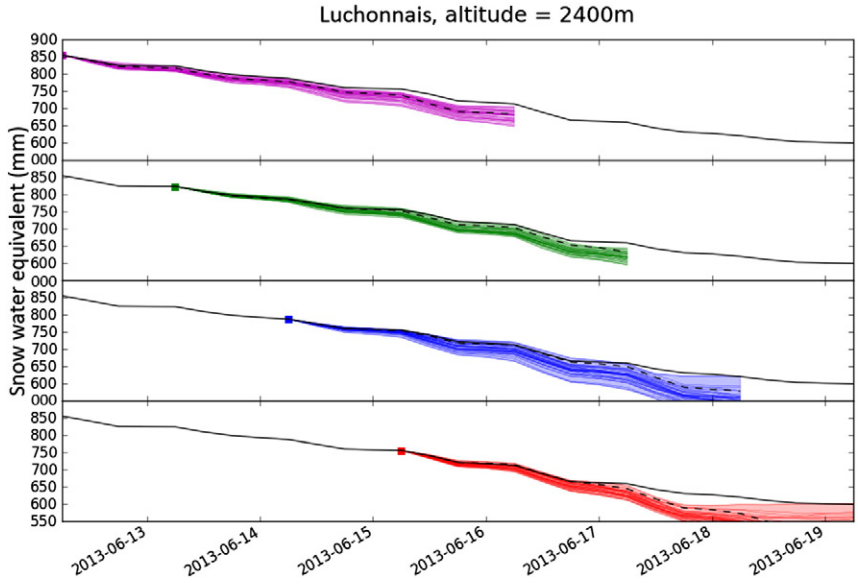

Fig. 13. Total snow water equivalent time evolution for the Luchonnais massif. The 4 forecasts (4 lines) start from the Jun. 12, 2013 at 6 UTC to Jun. 15, 2013 at 06 UTC and cover the four subsequent days respectively. Colored lines: 35 members of PEARP-BDAP-S2M; black solid line: S2M analysis; black dash line: ARPEGE-S2M deterministic forecast.

the snow model errors, and uncertainties associated with the intra-massif spatial variability of snow properties.

Acknowledgements

We thank Pierrick Cébron, Carole Labadie and Pascal Raynaud from CNRM-GAME/GMAP/RECYF team for providing PEARP-FULL data and for their explanations about the PEARP system. We thank Fabienne Rousset-Regimbeau (DCLIM/AVH), Pascal Laveau (ENM), Laure Raynaud (CNRM-GAME/GMAP/RECYF) and Cécile Coléou (CNRM-GAME/CEN) for useful advices and discussions. We also thank ISSW Special Issue Managing Guest Editor Pascal Haegeli and two anonymous reviewers for their careful revision, which helped improve this manuscript. CNRM-GAME/CEN is part of Labex OSUG@2020 (ANR10 LABX56).

References

Berre, L., Pannekoucke, O., Desroziers, G., Stefanescu, S., Chapnik, B., Raynaud, L., 2007. A variational assimilation ensemble and the spatial filtering of its error covariances: increase of sample size by local spatial averaging. ECMWF Workshop on Flow-dependent Aspects of Data Assimilation, Reading, pp. 151–168.

Bougeault, P., Toth, Z., Bishop, C., Brown, B., Burridge, D., Chen, D.H., Ebert, B., Fuentes, M., Hamill, T.M., Mylne, K., Nicolau, J., Paccagnella, T., Park, Y.Y., Parsons, D., Raoult, B., Schuster, D., Dias, P.S., Swinbank, R., Takeuchi, Y., Tennant, W., Wilson, L., Worley, S., 2010. The THORPEX Interactive Grand Global Ensemble. Bull. Am. Meteorol. Soc. 91, 1059–1072. http://dx.doi.org/10.1175/2010BAMS2853.1.

Bowler, N.E., Arribas, A., Mylne, K.R., Robertson, K.B., Beare, S.E., 2008. The MOGREPS short-range ensemble prediction system. Q. J. R. Meteorol. Soc. 134, 703–722. http://dx.doi.org/10.1002/qj.234.

Brier, G.W., 1950. Verification of forecasts expressed in termes of probability. Mon. Weather Rev. 78, 1–3.

Buizza, R., Palmer, T., 1995. The singular-vector structure of the atmospheric global circulation. J. Atmos. Sci. 52, 1434–1456. http://dx.doi.org/10.1175/1520-0469(1995)052<1434:TSVSOT>2.0.CO;2.

Candille, G., Talagrand, O., 2005. Evaluation of probabilistic prediction systems for a scalar variable. Q. J. R. Meteorol. Soc. 131, 2131–2150. http://dx.doi.org/10.1256/qj.04.71.

Courtier, P., Freydier, C., Geleyn, J.F., Rabier, F., Rochas, M., 1991. The arpege project at météo-france. Proceedings of the 1991 ECMWF Seminar. ECMWF, Reading, U.-K, pp. 193–231.

DeMaria, M., Knaff, J.A., Knabb, R., Lauer, C., Sampson, C.R., DeMaria, R.T., 2009. A new method for estimating tropical cyclone wind speed probabilities. Weather Forecast. 24, 1573–1591. http://dx.doi.org/10.1175/2009WAF2222286.1.

Descamps, L., Labadie, C., Joly, A., Bazile, E., Arbogast, P., Cébron, P., 2014. PEARP, the Météo-France short-range ensemble prediction system. Q. J. R. Meteorol. Soc. http://dx.doi.org/10.1002/qj.2469.

Durand, Y., Brun, E., Mérindol, L., Guyomarc'h, G., Lesaffre, B., Martin, E., 1993. A meteorological estimation of relevant parameters for snow models. Ann. Glaciol. 18, 65–71.

Durand, Y., Giraud, G., Brun, E., Mérindol, L., Martin, E., 1999. A computer-based system simulating snowpack structures as a tool for regional avalanche forecasting. J. Glaciol. 45, 469–484.

Efron, B., 1979. 1977 Rietz Lecture — Bootstrap methods, another look at the Jackknife. Ann. Stat. 7, 1–26. http://dx.doi.org/10.1214/aos/1176344552.

Ferro, C.A.T., 2007. Comparing Probabilistic forecasting systems with the brier score. Weather Forecast. 22, 1076–1088. http://dx.doi.org/10.1175/WAF1034.1.

Fierz, C., Armstrong, R.L., Durand, Y., Etchevers, P., Greene, E., McClung, D.M., Nishimura, K., Satyawali, P.K., Sokratov, S.A., 2009. The international classification for seasonal snow on the ground. IHP-VII Technical Documents in Hydrology n 83, IACS Contribution n 1.

Föhn, P., 1987. The stability index and various triggering mechanisms. Int. Assoc. Hydrol. Sci. Publ. 162, 195–214.

Frogner, I.L., Haakenstad, H., Iversen, T., 2006. Limited-area ensemble predictions at the Norwegian Meteorological Institute. Q. J. R. Meteorol. Soc. 132, 2785–2808. http://dx.doi.org/10.1256/qj.04.178.

Giraud, G., 1992. MEPRA: an expert system for avalanche risk forecasting. Proceedings of the International snow science workshop, 4-8 oct 1992, Breckenridge, Colorado, USA, pp. 97–106.

Giraud, G., Lafeuille, J., Pahaut, E., 1987. Evaluation de la qualité de la prévision du risque d'avalanche. Int. Assoc. Hydrol. Sci. Publ. 162, 583–591.

Gneiting, T., 2014. Calibration of Medium-Range Weather Forecasts. Technical Memoranda 719. ECMWF (URL:http://old.ecmwf.int/publications/library/ecpublications/pdf/tm/701-800/tm719.pdf).

Gneiting, T., Raftery, A., Westveld, A., Goldman, T., 2005. Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. Mon. Weather Rev. 133, 1098–1118. http://dx.doi.org/10.1175/MWR2904.1.

Lafaysse, M., Morin, S., Coléou, C., Vernay, M., Serça, D., Besson, F., Willemet, J.M., Giraud, G., Durand, Y., 2013. Toward a new chain of models for avalanche hazard forecasting in french mountain ranges, including low altitude mountains. Proceedings of the International Snow Science Workshop - Grenoble and Chamonix, pp. 162–166.

Marsigli, C., Montani, A., Paccangnella, T., 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. Meteorol. Appl. 15, 125–143. http://dx.doi.org/10.1002/met.65.

Martin, E., Giraud, G., Lejeune, Y., Boudart, G., 2001. Impact of a climate change on avalanche hazard. Ann. Glaciol. 32, 163–167. http://dx.doi.org/10.3189/172756401781819292.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Molteni, F., Buizza, R., Palmer, T., Petroliagis, T., 1996. The ECMWF ensemble prediction system: Methodology and validation. Q. J. R. Meteorol. Soc. 122, 73–119. http://dx.doi.org/10.1002/qj.49712252905.

Murphy, A.H., 1973. A new vector partition of the probability score. J. Appl. Meteorol. 12, 595–600.

Pellerin, G., Lefaivre, L., Houtekamer, P., Girard, C., 2003. Increasing the horizontal resolution of ensemble forecasts at CMC. Nonlinear Process. Geophys. 10, 463–468.

Raftery, A., Gneiting, T., Balabdaoui, F., Polakowski, M., 2005. Using Bayesian model averaging to calibrate forecast ensembles. Mon. Weather Rev. 133, 1155–1174. http://dx.doi.org/10.1175/MWR2906.1.

Rousset-Regimbeau, F., 2007. Modélisation des bilans de surface et des débits sur la France, application à la prévision d'ensembles des débits. (Ph.D. thesis). Université Paul Sabatier, Toulouse III, France.

Scheffé, H., 1959. The Analysis of Variance. John Wiley and Sons, New York.

Thirel, G., Rousset-Regimbeau, F., Martin, E., Habets, F., 2008. On the impact of short-range meteorological forecasts for ensemble streamflow predictions. J. Hydrometeorol. 9, 1301–1317. http://dx.doi.org/10.1175/2008JHM959.1.

Thirel, G., Regimbeau, F., Martin, E., Noilhan, J., Habets, F., 2010. Short- and medium-range hydrological ensemble forecasts over France. Atmos. Sci. Lett. 11, 72–77. http://dx.doi.org/10.1002/asl.254.

Toth, Z., Kalnay, E., 1997. Ensemble forecasting at NCEP and the breeding method. Mon. Weather Rev. 125, 3297–3319. http://dx.doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

Vionnet, V., Brun, E., Morin, S., Boone, A., Martin, E., Faroux, S., Le-Moigne, P., Willemet, J.M., 2012. The detailed snowpack scheme Crocus and its implementation in SURFEX v7.2. Geosci. Model Dev. 5, 773–791. http://dx.doi.org/10.5194/gmd-5-773-2012.

Voisin, N., Pappenberger, F., Lettenmaier, D.P., Buizza, R., Schaake, J.C., 2011. Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River basin. Weather Forecast. 26, 425–446. http://dx.doi.org/10.1175/WAF-D-10-05032.1.