Clustering and selection of boundary conditions for limited-area ensemble prediction

François Bouttier, Laure Raynaud

Clustering and selection of boundary conditions for limited area ensemble prediction

François Bouttier and Laure Raynaud

22 March 2018

affiliation: CNRM, Toulouse University, Météo-France and CNRS, Toulouse, France

corresponding author: François Bouttier, CNRM/GMME/PRECIP Météo-France 42 Av. Coriolis F-31057 Toulouse cedex, France. Email: francois.bouttier@meteo.fr

Orcid identifier: François Bouttier, 0000-0001-6148-4510

Funding information: Météo-France and CNRS.

Cite as: Bouttier, F. and L. Raynaud, 2018: Clustering and selection of boundary conditions for limited area ensemble prediction. *Quart. J. Roy. Meteor. Soc.* **144**:2381-2391. doi:10.1002/qj.3304

Abstract

Limited area ensemble predictions can be sensitive to the specification of lateral boundary conditions, that are often built by subsampling larger ensembles. Using the operational PEARP and AROME-EPS ensembles, we compare several subsampling methods, including random selection, 'representative members' as defined in Molteni (2001), and a new selection method. The tests show that the algorithms used for the clustering and the member selection have a significant impact on the resulting ensembles. Clustering-based methods are shown to outperform random subsampling, mostly (but not only) because they change the ensemble spread. Cluster sizes can be highly variable, which can complicate ensemble interpretation. We present a subsampling algorithm that has little impact on performance scores, but better preserves ensemble spread and produces nearly equally likely members, by limiting cluster size variability.

keywords: atmospheric model, numerical weather prediction, AROME ensemble prediction, subsampling, clustering, representative members

1 Introduction

Limited area ensemble forecasting has become a standard tool for probabilistic numerical weather prediction. Most major meteorological services now run such systems routinely. Typical examples are found in Du et al. (2015), Clark et al. (2012) in the USA, Saito et al (2011) at the Japan Meteorological Agency, Li et al. (2008), Hagelin et al. (2017) at the UK Met Office, Gebhardt et al. (2011) at Deutscher Wetterdienst, and Bouttier et al. (2016) at Météo-France. These systems rely on large scale models to provide boundary conditions for their ensemble members. The variability of boundary conditions is essential for representing large scale uncertainties in limited area predictions beyond a few hours, as demonstrated by several studies including Gebhardt et al. (2011) and Vié et al. (2011). Some ensemble systems introduce boundary condition variability by mixing deterministic forecasts from several global models. For instance, the systems described in Gebhardt et al. (2011) and García-Moya (2011) use a few global deterministic models to drive limited area ensembles. In other meteorological services, global ensemble prediction systems are used.

Limited area ensembles need to provide significant added value over their global counterparts, under computational constraints. Consequently, they are usually configured with higher resolution and lower ensemble size than global systems, which reflects different tradeoffs in terms of computational costs. When the size n of the limited area ensemble is smaller than the size N of its driving ensemble, one has to pick n out of N driving members. This operation is called 'subsampling' in this paper. We intend to investigate the sensitivity of an operational ensemble system to different subsampling strategies.

The ensemble of N large scale members provides an estimate of atmospheric uncertainty for the limited area ensemble. This estimate may not be perfect, but it is designed to be the best available in a real-time forecasting setting. Thus, the limited area ensemble should select n members with a distribution that is as faithful as possible to the full ensemble. This leads to the most common method, called 'random subsampling', in which the n members are picked at random. Some teams have investigated more elaborate sampling methods, in order to optimize the performance of the limited area ensemble. Molteni et al. (2001) have devised a subsampling method for the ECMWF (European Centre for Medium-Range Weather Predictions) ensemble to drive the COSMO-LEPS limited area ensemble, with a focus on predicting high precipitation events over the Alps. This method, called the 'Molteni 2001' subsampling here, has been extensively studied by Marsigli et al. (2001 and 2005), Keil and Craig (2007), and Montani et al. (2011). They have provided convincing proof of its superiority in the COSMO-LEPS context. The Molteni 2001 method has also been independently tested at the Austrian Meteorological Institute (Weidle et al. 2013), Météo-France (Nuissier et al. 2012), and Meteoswiss (A. Walser and S. Westerhuis, personal communication).

Ensemble subsampling has been applied to other weather forecasting problems: clustering was used in Atger (1999) and Brill (2015) to summarize output from large ensembles using a few forecast scenarios as guidance for human interpretation. Branković et al. (2008) applied it to identify inconsistencies between global and limited area ensemble predictions. Kowaleski and Evans (2016) have shown a statistical relationship between cyclone trajectory and structure, using ensemble clustering. Alhamed et al. (2002), Yussouf et al. (2004) and Johnson et al. (2011a, 2011b) applied clustering to demonstrate how the dispersion of multimodel (or multiphysics) ensembles tends to be determined by model physics diversity, which dominates over other representations of forecast uncertainty. Lee et al. (2012, 2016) developed this idea for optimizing the design of multiphysics ensemble systems. Tien et al. (2014) proposed using the Molteni 2001 subsampling to optimize air traffic management operations.

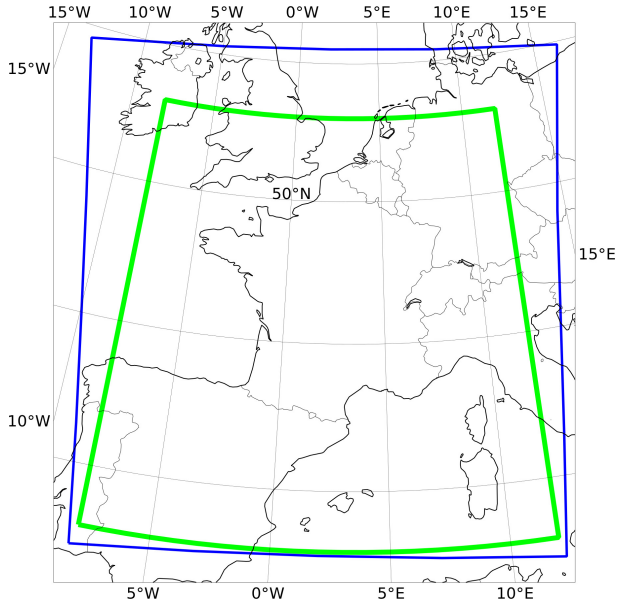

Figure 1: Geographical domains used in this study. Outer domain (thinner blue line): AROME-EPS computational model area. Inner domain (thicker green line): area used to define the clustering distance and the verification diagnostics.

In this paper, we will compare several subsampling methods, including random selection and the Molteni 2001 algorithm, in the context of the AROME-EPS limited area ensemble prediction system. The main question is whether we can identify an optimal method for driving a limited area ensemble. This paper is organized as follows. Section 2 introduces the forecast data, subsampling algorithms, and the verification metrics. Section 3 and 4 compare the algorithms using two different point of views: the intrinsic behaviour of the subsampling algorithms in the large scale ensemble, and their impact on limited area ensemble forecasts. Section 6 presents a summary and discussion.

2 Methodology

2.1 Forecast dataset

The ensemble prediction system used for this study is AROME-EPS, which covers the geographical area depicted in figure 1. AROME-EPS has been in operational production at Météo-France since November 2016. It uses the non-hydrostatic convection-permitting AROME model at an horizontal resolution of 2.5km (Seity et al., 2011), stochastic physics perturbations (Bouttier et al., 2012), initial and surface perturbations (Bouttier et al., 2016), and large scale boundary conditions provided by the global 35-member PEARP ensemble system (Descamps et al., 2015).

In its 2017 operational configuration, AROME-EPS had 12 members driven by a subsampling of the PEARP ensemble, using the algorithm described in Nuissier et al. (2012), which is based on the proposal of Molteni et al. (2001). Raynaud and Bouttier (2018) provide an updated view of the operational performance of the AROME-EPS system.

The PEARP ensemble (Descamps et al. 2014) contains a so-called 'control' member, which is not

perturbed. This control member is not used in the generation of AROME-EPS, because including it would lead to an inhomogenous ensemble (it can be shown that including the control member in the ensemble systematically decreases its spread). Thus, in this work the PEARP control member is ignored and PEARP is regarded as a 34-member ensemble. The generation of PEARP and AROME-EPS perturbations is random, so that their members are exchangeable.

We use a version of the AROME-EPS that is identical to the operational system at Météo-France in 2017, except that it contains 34 members instead of 12. Each AROME-EPS member is driven by a perturbed PEARP member. This setup allows a clean comparison of various subsampling methods, without rerunning a full 12-member AROME-EPS ensemble for each, which would be computationally expensive. In doing so, the centering of the initial condition perturbations uses a fixed 34-member ensemble mean, which is an approximation to the genuine 12-member AROME-EPS system. The centering procedure is explained in Bouttier et al. (2016), who showed that details of the initial perturbations do not matter beyond a few hours of forecast. We restrict our present study to forecast ranges beyond 12 hours, so the centering approximation is not expected to affect the conclusions of this paper.

The forecast dataset comprises 88 ensemble forecasts from PEARP and AROME-EPS. The ensembles are run once per day from 15 February 2015 to 5 April 2015, and from 10 May to 20 June 2015. A few forecasts are missing for technical reasons. We apply each subsampling method as follows:

- each 34-member PEARP ensemble forecast is converted into a 12-member ensemble by applying the selected subsampling method.
- the same subsampling is applied to the corresponding 34-member AROME-EPS ensemble forecast, which produces a 12-member AROME-EPS ensemble as if the subsampling algorithm were used in operations.

2.2 Subsampling methods

The following subsampling methods will be compared. The '**random**' subsampling is a random draw of 12 out of 34 PEARP members. Members are drawn without replacement, which is the usual practice in the community, because it is easy to implement and it ensures that each member is equally likely (each member can only be drawn once). Assuming that ensemble members are equally likely is convenient because it facilitates the design of probabilistic post-processing and verification tools. Drawing a subsample without replacement produces an ensemble that is not exactly distributed like the full ensemble, because this operation introduces correlations between the members, which changes the ensemble spread. Thus, random subsampling cannot be regarded as a perfect method, but it is included in this study as an important reference because of its popularity.

The '**Molteni 2001**' subsampling method closely follows the algorithm proposed by Molteni et al. (2001). It works in three steps:

- the *metric* used to define the distance between two PEARP members is the sum of the point-by-point root mean square difference between fields of (zonal and meridional wind components, geopotential height, relative humidity) at pressure levels (500hPa, 700hPa, 850hPa) and at a forecast range of 30 hours. The summation is carried out on a regular latitude/longitude grid of mesh 0.025 degrees over the inner domain depicted in figure 1. In operational production, more forecast

ranges can be used. In this paper, we only use one range to facilitate the visual interpretation of the algorithmic choices in terms of meteorological fields.

- a *clustering* method is used to partition the 34-member PEARP ensemble into 12 subsets, called clusters. The clusters are designed so that its members are as similar as possible within each cluster, while making different clusters are as dissimilar as possible. Many clustering algorithms exist, with variable definitions of similarity between members and clusters, and various solving strategies. The Molteni subsampling uses the complete linkage agglomerative clustering with the above metric. 'Complete linkage' means that the dissimilarity between two clusters is defined as the maximum distance between any pair of members belonging to the clusters.

- a *representative member (RM) selection* step picks one member per cluster. Again, there are several possible RM selection strategies. The Molteni subsampling selects the RM of each cluster as the member that minimizes the so-called 'representativeness index', which is the ratio between (a) its rms average distance to other members *inside* the same cluster, and (b) its rms average distance to members in the *other* clusters. The practical implications of this RM selection technique will be illustrated below.

The following methods are designed to evaluate the impact of changing the clustering and RM selection algorithms.

The '**Molteni central**' subsampling is a simple modification of the Molteni 2001 method, by which the RM selection now picks the member of each cluster that is closest to the cluster centroid. This is equivalent to setting the (b) term above to a constant. Intuitively, this (b) term tends to favour the selection of extreme members, in clusters that are near the outside of the set of members. There is some compromise, though, between (a) and (b), since (a) favours members close to the center of each cluster. In the following section, a simple low-dimensional test will be performed to clarify their net impact before moving on to full-size meteorological ensembles.

The '**penalized Ward**' method uses a modified Ward's hierarchical clustering method instead of the complete linkage, and the same RM selection algorithm as 'Molteni central'. In the standard Ward's clustering method (Ward, 1963), the dissimilarity $d_{\mathrm{w}}(A, B)$ between two clusters A and B is defined as the variance of their union. According to the clustering literature, Ward's method tends to produce more evenly populated clusters than the complete linkage. Preliminary tests showed, however, that there still was substantial variability in the cluster sizes, so in this paper the clustering algorithm uses a modified cluster dissimilarity measure in order to discourage the formation of large clusters. A nearly even partition of a 34-member ensemble would produce 12 clusters of size close to $34/12 \simeq 2.83$. Forming larger clusters implies that other clusters will be smaller. Intuitively, RMs of bigger clusters carry more forecast probability than RMs of smaller ones. If clusters have similar sizes, then the subsampled ensemble members can be expected to be equally likely. The ability to regard the 12 RMs as nearly equally likely is an advantage for ensemble users. In practice, the 'penalized Ward' clustering algorithm is constructed by defining a new dissimilarity measure between clusters A and B, d_{p}, as

$$d_{\mathrm{p}}(A, B) = d_{\mathrm{w}}(A, B) \mid A \cup B \mid /2 \tag{1}$$

where $\mid A \cup B \mid$ is the size of their union.

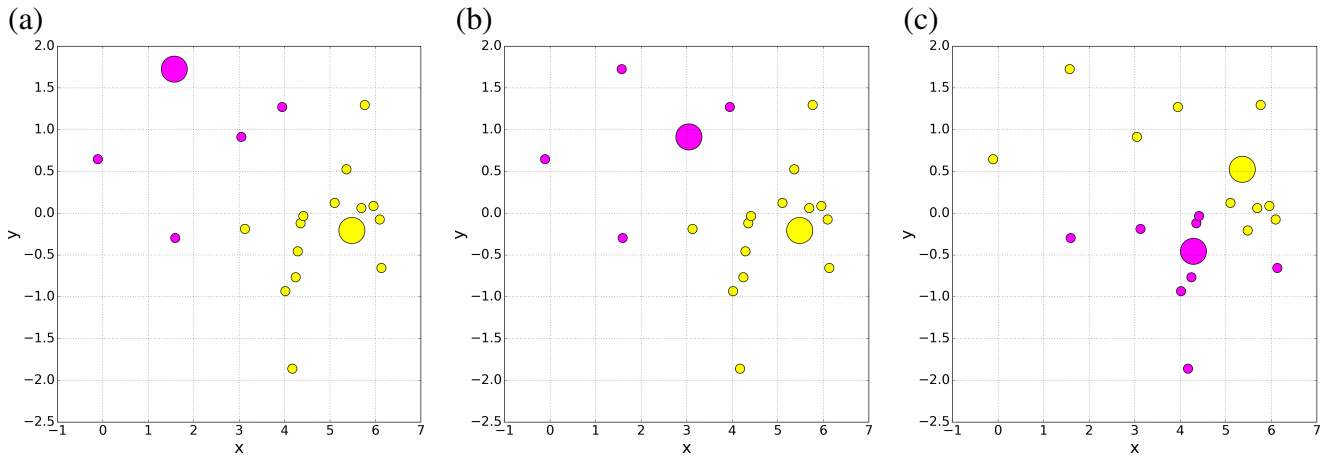

Figure 2: Comparison of three subsampling methods applied to a 20-member two-dimensional ensemble: (a) Molteni 2001 method, (b) Molteni central method, (c) Penalized Ward method. There is one bullet per member. The colours indicate cluster membership. The bigger bullets are the selected representative members.

2.3 A simple illustration of selection algorithms

In this section, we illustrate the impact of changing the selection methods using a simple two-dimensional framework. We randomly draw 20 points at random and subsample this dataset as two members using the 'Molteni 2001', 'Molteni central', and 'penalized Ward' methods. The results are shown in figure 2. The complete linkage clustering creates uneven classes of sizes 5 and 15, whereas the 'penalized Ward' clustering produces nearly equal classes of sizes 9 and 11. In the Molteni 2001 selection method, the RM is often 'pushed' towards the extremes of the dataset, which occurs less often with the centroid selection, particularly in smaller clusters. As a result, the dispersion of the RMs is smaller with the 'penalized Ward' than the 'Molteni 2001' method.

This example has shown the typical impact expected from changing the subsampling algorithm. In other datasets, the impact may be less obvious than in the example shown, because the Molteni 2001 algorithm does not always produce very unequal clusters or extreme RMs. The behaviour of the subsampling algorithms may depend on the dataset dimensionality and number of RMs. In the sequel, we shall demonstrate that these results remain statistically valid in meteorological ensembles.

2.4 Performance measures

In the following sections, the performance of several selection methods will be measured by objective ensemble scores. These scores measure the consistency between probabilistic forecasts implied by the ensemble forecasts and the truth provided by observations. For more details on ensemble verification, the reader is referred to Jolliffe and Stephenson (2011). We use the following scores with respect to observations of two-metre temperature (T2m), ten-metre wind speed (ff10m), and six-hourly accumulation of precipitation (rr6):

- the **spread-skill ratio** measures the consistency between ensemble spread and the root mean square (rms) error of the ensemble mean. A reliable ensemble has a spread-skill ratio of one. Lower values mean that the ensemble is underdispersive according to this measure.

- the **CRPS** (Continuous Ranked Probability Score): this is an integral measure of the consistency between forecast probabilities and observed values. A lower value indicates a better ensemble.
- the **ROCA** (or area under the Relative Operating Characteristic curve): this is a summary measure of success rates of binary forecasts based on the ensemble output. The ROCA is normalized so that zero corresponds to an unskillful (random) forecast, one corresponds to a perfect forecast. Here, the binary forecast events used to compute the ROCA are defined as the exceedance of fixed thresholds of (8K, 2.7ms^{-1}, 10mm) for (T2m, ff10m, rr6), respectively. These thresholds are chosen to provide large amounts of events and non-events from our dataset; the results are not very sensitive to them.

Observations errors are taken into account in all scores (except ROCA) by assuming standard measurements errors of (1K, 0.33ms^{-1}, $0.1 + 0.1rr$) for (T2m, ff10m, rr6), respectively, where rr is the observed precipitation. Score differences are deemed significant if they have the same sign as their average, according to a bootstrap test at the 95% level.

The observations are taken from real-time meteorological networks. The observing stations conform to WMO (World Meteorological Organization) standards for ground-based meteorological measurements. Most are located in mainland France, with a typical spacing of 30km between stations, plus lower density stations in neighbouring countries, and ship or buoy data over sea. At each verification time, there are typically about 1600 temperature observations and 1000 observations of wind and precipitation. The data is screened for gross errors. About 3% of T2m and wind stations are permanently excluded because their departures from AROME forecasts are consistently much higher than the others over the experiment period, in terms of bias or rms. The excluded stations are located in mountainous regions, where the AROME model resolution is insufficient to resolve important local meteorological features. The selected observations are compared with AROME model diagnostics of the corresponding physical quantities on a regular latitude-longitude grid of resolution 0.025 degrees, using a nearest-neighbour interpolation.

3 Impact of member selection methods in the PEARP ensemble

This section investigates the consequences of subsampling in the global PEARP ensemble before looking at them in the AROME-EPS the regional ensemble. Note that no data from the latter is used in the subsampling process.

The first step is the examination of the clustering behaviour. Clustering is not relevant for the 'random' algorithm, where representative members are directly picked from the PEARP ensemble. The 'Molteni 2001' and 'Molteni central' methods have identical clusters, since both use the complete linkage agglomerative algorithm.

Assessing the performance of a clustering algorithm is difficult for two reasons. First, a clustering contains complex information that involves multiple subsets of a 34-member high-dimensional ensemble. Few summary measures are available to assess the quality of a clustering. An example is the silhouette index (Rousseuw, 1987). Such measures can be useful, but their interpretation tends to be highly subjective. Second, even if a satisfactory measure of goodness could be identified for PEARP clustering, it would be hard to relate to the quality of RMs in terms of AROME-EPS ensemble generation. For instance, if an ensemble has many members close to a given meteorological scenario A and a few other

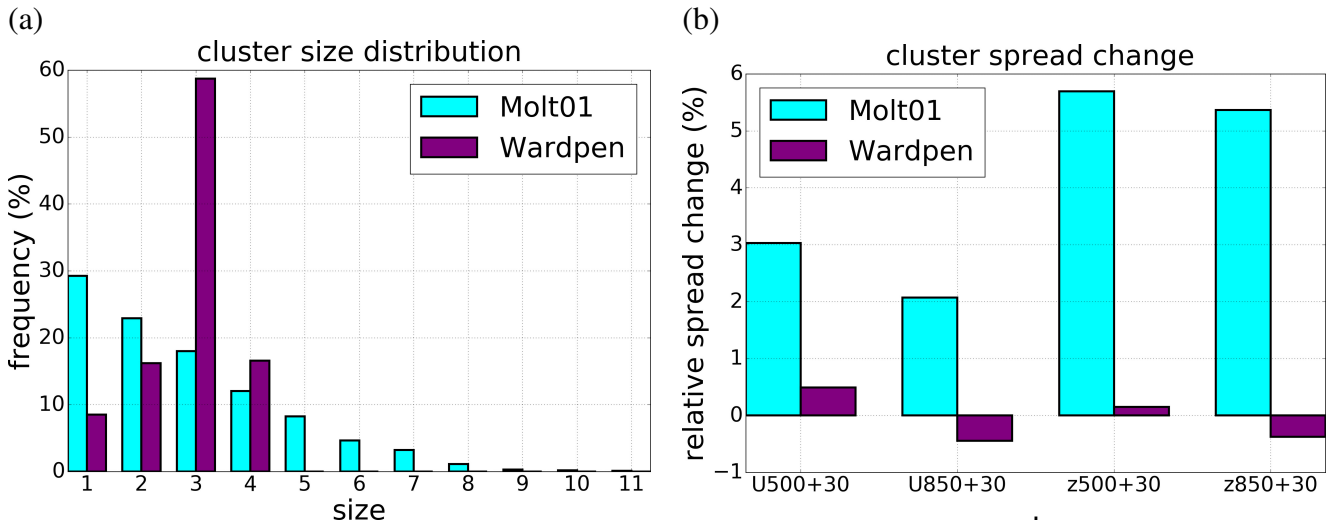

Figure 3: (a) Distribution of the cluster sizes over the dataset, using two clustering algorithms: the complete linkage as used in the 'Molteni 2001' and 'Molteni central' procedures, and the penalized Ward algorithm. (b) Relative changes of the PEARP 12-member subensemble dispersion with respect to the 'random' subsampling procedure: 'Molteni 2001' (blue) and 'penalized Ward' (purple) algorithms. The dispersion is computed using the parameters indicated by the bottom labels: 30-hour forecasts of zonal wind at levels 500 and 850hPa (U500+30 and U850+30, respectively), geopotential at the same levels (z500+30 and z850+30, respectively).

members close to an alternative scenario B, a two-cluster classification could lead to one cluster around A and another around B. A different, yet reasonable classification could involve splitting A into two large clusters A1 and A2, with all members arbitrarily merged into them. From the point of view of identifying local probability maxima (insofar as this concept makes sense with 34 members), the first clustering is better, because it leads to a subsampling that emphasizes differences between A and B. In terms of faithfulness to the full-ensemble probability distribution, the second clustering is better, because it reflects the fact that scenario A is much more likely than B. In order In order to correctly summarize both properties of the full ensemble (i.e. 'A is well separated from B', and 'A is much more likely than B'), more than two clusters would be necessary. This is not possible in operational numerical weather prediction, where the ensemble size is bounded by computational constraints: there is no guarantee that, on any given day, the size of the limited area ensemble will match the number of clusters needed to satisfactorily depict the distribution of all members. In summary, it does not seem possible to identify a best clustering algorithm for limited area ensemble performance, because the conceptual relationship between clustering and subsampling is complex.

In this work we focus on the distribution of cluster sizes, because it has been flagged by human forecasters as an important aspect of ensemble generation. Figure 3a shows the frequency distribution of cluster sizes with the 'complete linkage' and 'penalized Ward' methods. These histograms have been produced by clustering each PEARP ensemble over the 88-day period, and counting how many clusters of each size are produced. The histograms display the relative frequencies of cluster sizes. By construction, the total number of clusters is constant and equal to 12 times the number of ensemble forecasts, so that the average cluster size is constant.

Figure 3a shows that the complete linkage produces highly variable clusters, with a substantial proportion of very small clusters (1 or 2 members). Large clusters (5 members and above) occur in about

10% of ensemble forecasts, where one single large cluster usually coexists with very small ones. These cases were independently flagged by human forecasters during a subjective evaluation of the AROME-EPS ensemble system at Météo-France. From summer 2016 to end 2017 (which includes the dataset used in this paper), forecasters were asked to record if they obtained added value from the ensemble forecasts as a daily guidance tool, beside the inspection of deterministic models from various centres. A survey of their remarks showed that AROME-EPS was often deemed uninformative because the cluster sizes were unbalanced: members associated to small clusters were felt to be too unlikely to represent credible scenarios, and the member associated to the large, main cluster did not account for the intracluster variability. The conclusion is that, even if unbalanced clusters can be justified from a mathematical point of view, they are not optimal for conveying forecast uncertainty, because when they occur, only one member of the resulting ensemble forecast can be interpreted as a likely weather forecast. As expected from its algorithmic design, the penalized Ward clustering exhibits much less dispersion in cluster sizes, with fewer one-member clusters, and no cluster bigger than four over the experimental period.

The next step is to examine how changes to the clustering and member selection affect properties of the selected PEARP subensemble. Figure 3b shows the impact on the PEARP subensemble spread of changing the subsampling method (including both clustering and RM selection). Here, the spread is measured by the domain averaged standard deviation of the subensembles, in the domain used to define the metric. The spread of each method is normalized by the spread of the 'random' subensemble, which is statistically undistinguishable from the spread of the full 34-member PEARP ensemble (taking into account spread biases linked to the varying ensemble size). The 'Molteni 2001' subsampling increases the ensemble spread by 2 to 6%, depending on the parameter, and a bootstrap test shows that these differences are significant at the 95% level. The 'penalized Ward' subsampling exhibits much smaller impacts that depend on the considered parameter; bootstrap testing did not show them to be significant at the 95% level for any of the shown parameter. A bigger dataset might show that there is indeed some significant spread change, but it seems to be of no practical significance.

4 Impact on the limited area ensemble

The impact of changing the subsampling algorithm will now be examined in terms of the performance of the AROME-EPS limited area ensemble, which is the important end product. The previous section has shown an impact on the PEARP upper level fields at the 30 hours forecast range. The relationship between these fields and the AROME-EPS actual weather fields is not a priori obvious, since the modelled scales and forecast ranges are going to be different. One expects AROME-EPS to react in a similar fashion to the PEARP ensemble in terms of spread, but other measures of forecast quality need to be checked as well.

4.1 Spread-skill diagnostics

Figure 4 shows the impact on the reliability of ensemble spread. Changes to the displayed spread-skill ratio are consistent with variations of the absolute ensemble spread (not shown): when the spread-skill ratio increases, the absolute ensemble spread increases, too. All ensembles tested here are underdispersive in terms of temperature and wind, whereas precipitation is nearly correctly dispersed according to the spread-skill ratio. The spread-skill ratio tends to increase with range, with some variations linked to the diurnal cycle.

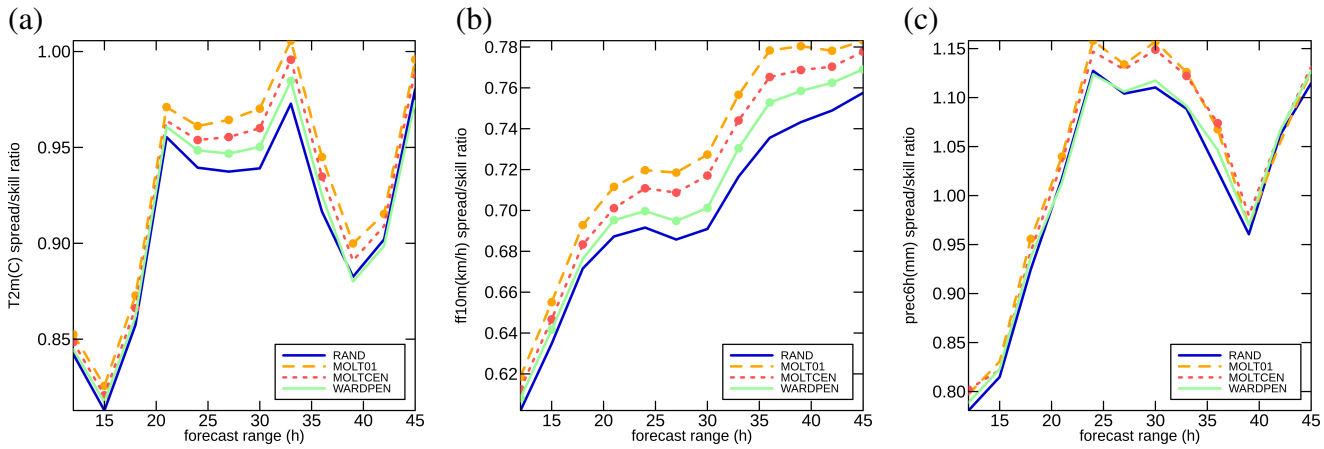

Figure 4: Average spread-skill ratio of the AROME-EPS ensemble as a function of forecast range (hours), using four subsampling methods: random (RAND), 'Molteni 2001' (MOLT01), 'Molteni centered' (MOLTCEN), and 'penalized Ward' (WARDPEN). Forecast error is taken into account so that the ideal average ratio is 1, lower values indicate underdispersiveness. The ratios are computed over the verification area for each range of the ensemble forecasts, then averaged over the 88 ensembles. The bullets indicate values that are significantly different from the random subsampling, according to a bootstrap test at the 95% level.

The gap between the RAND and MOLT01 curves indicates that the 'Molteni 2001' procedure significantly increases the AROME-EPS ensemble spread by a few percents, which is consistent with the impact on the PEARP ensemble seen in the previous section. The impact is largest around the 27 hour range, which is the one used by the member selection (it corresponds to 30 hours in PEARP since the AROME-EPS forecasts are initiated 3 hours after PEARP). The impact remains fairly constant inside the range interval from 24 to 36 hours, it is smaller outside. It suggests that more forecast ranges should be included in the subsampling distance, if one aims to control a wider interval of forecast ranges.

The gap between the 'Molteni 2001' and 'Molteni central' curves (respectively labelled MOLT01 and MOLTCEN) shows that changing the member selection procedure reduces the ensemble spread-skill ratio by about one percent. The 'Molteni central' ratio remains significantly higher than the 'random' one, which shows that in Molteni 2001, both clustering and RM selection contribute to the subensemble spread.

The 'Molteni central' and 'penalized Ward' curves, show that changing the clustering from the complete linkage to the modified Ward's algorithm reduces the ensemble spread, making it closer to the random subsampling. This reduction would be much smaller if we had used the original Ward's algorithm (not shown). The modifications to the spread-skill ratios are significant for temperature and wind, but usually less clear for precipitation which exhibits similar variations with less statistical significance.

In summary, all clustering-based subsampling procedures tested here improve the reliability of the AROME-EPS ensemble spread for temperature and wind, by adding dispersion with respect to a random subsampling. The representative member selection proposed by Molteni (2001) also produces higher spread than a centroid-based selection. Increasing the spread-skill ratio can be a good or a bad thing depending on the user's priority. Nuissier et al. (2012) showed that the Molteni (2001) procedure has attractive properties for the detection of high precipitation events; the previous sections suggest that the Molteni (2001) algorithm sometimes achieves this effect by emphasizing extreme weather scenarios in the subsampling, and generating unevenly-sized clusters, both of which are felt to be undesirable for

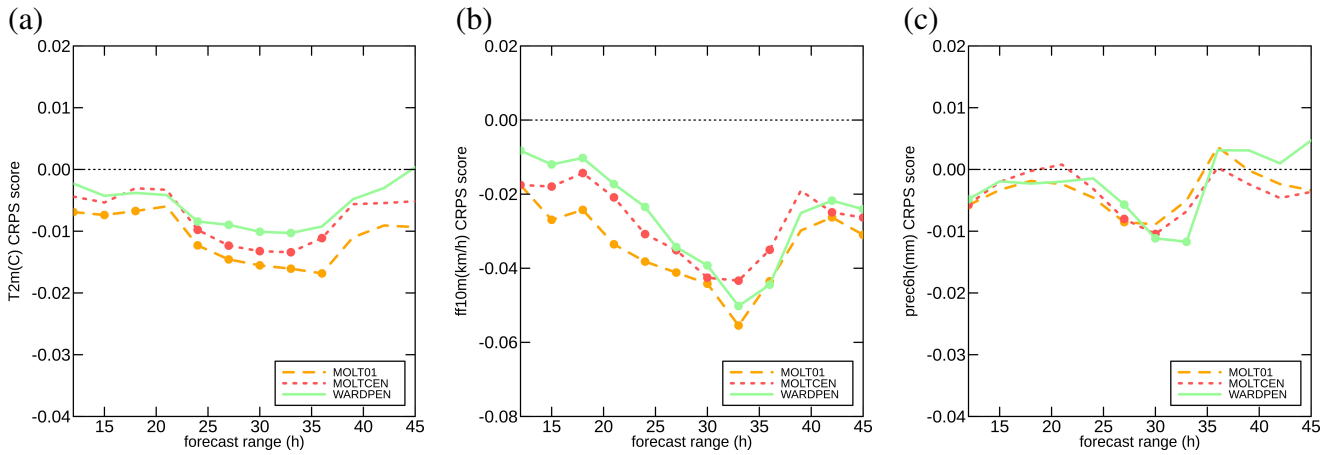

Figure 5: Like figure 4, for the CRPS score difference with respect to the random subsampling. Lower values indicate better ensemble forecast probability distributions.

ensemble use by human forecasters. These issues can be addressed using the proposed modifications of the subsampling procedure, at the price of reduced spread-skill ratios.

4.2 CRPS scores

The CRPS score provides an integral measure of the quality of ensemble forecast. Figure 5 compares the CRPS scores of the AROME-EPS ensemble using different subsampling methods, with the same graphical conventions as in the previous subsection. Generally speaking, the CRPS improves (decreases) when the spread-skill ratio increases. For temperature and wind, an explanation is that forecast probabilities of an underdispersive parameter are usually improved when the ensemble spread increases. The CRPS is also improved for precipitation, even at ranges that are overdispersive according to the spread-skill ratio. In terms of CRPS, the bootstrap test indicates that all clustering-based subsamplings perform significantly better than a random selection. It is important to note that the CRPS is not significantly degraded by replacing Molteni 2001 with the penalized Ward method, despite the associated reduction of spread-skill consistency.

4.3 ROCA score

The ROCA score is another measure of ensemble quality. Figure 6 shows its sensitivity to different subsampling methods, using the same conventions as above. None of the displayed variations between clustering-based subsamplings appear to be significant, but all are better than a random subsampling for temperature and wind. This improvement is significant at forecast ranges close to the one used for clustering (between 24 and 36 hours). There seems to be a similar impact for precipitation, but it was not deemed statistically significant, probably because our dataset is too short: the frequency of the binary event used to compute the ROCA is much smaller for precipitation than for temperature and wind. If the precipitation threshold is lowered, its exceedance frequency increases, but then the ROCA differences become too small to achieve statistical significance. In other words, a dataset with more precipitation events would be needed to robustly conclude whether changing the subsampling method has a significant impact on the ROCA or not.

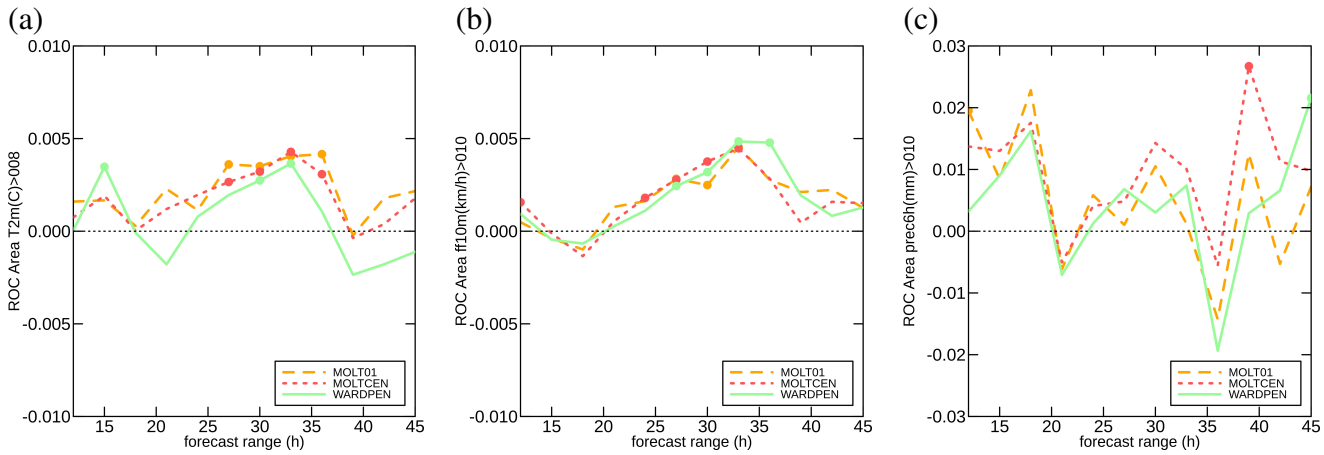

Figure 6: Like figure 4, for the ROCA score difference with respect to the random subsampling. Higher values indicate better forecasts for the binary events indicated in the text.

5 Summary and discussion

The sensitivity of a limited area ensemble prediction system to its large-scale forcing has been investigated by comparing different subsampling methods. Two commonly used approaches are considered: a random subsampling, which is our reference for score computation, and a clustering-based procedure along the lines of Molteni et al. (2001). Modifications of the latter are also considered. The main results are:

- all clustering-based subsampling method increase the ensemble dispersion and improve the forecast quality in terms of the CRPS and ROCA scores. This impact is statistically significant, it extends to a few hours before and after the time range used to compute the subsampling.
- in the Molteni et al. (2001) method, this dispersion increase can be attributed to both clustering and 'representative member' selection algorithms. They appear to have a cumulative effect on enhancing the ensemble spread. Our results are compatible with independent tests presented in Weidle et al. (2013).
- the subsampling method of Molteni et al. (2001) can be altered to facilitate the human interpretation of the ensemble forecast by reducing the occurence of unevenly sized clusters, and avoiding the selection of too extreme forcing members in the subsampling. The 'penalized Ward' subsampling method presented in this paper achieves this goal, while preserving the superiority of the Molteni et al. (2001) method over a random selection in terms of scores.

A four month long subjective evaluation of ensemble forecasts was carried out by experienced forecasters over an independent period at Météo-France, in order to compare the Molteni et al. (2001) and the 'penalized Ward' procedures. They confirmed that the latter does indeed produce a better subsampling, in the subjective sense that it seems more representative of the distribution of synoptic-scale scenarios in the PEARP ensemble.

We propose the following interpretation of our results: clustering techniques are designed to summarize datasets by grouping similar data points (here, weather scenarios) while maximizing intercluster

distances. By using clusters to guide representative member selection, one obtains a subensemble where the distance between members has been maximized. It explains why the clustering-based procedures tend to produce more dispersive ensembles than random subsamples. This effect is probably smaller with the penalized Ward method, because the penalization of large clusters is a constraint that reduces the ability of the clustering to reach its goal (of grouping similar members while maximizing intercluster distance). The spread-enhancing effect of clustering algorithms is amplified if the representative member selection favours outliers as in Molteni et al. (2001), instead of central members of each cluster. The spread enhancement is observed in the PEARP ensemble, and it propagates into the AROME-EPS ensemble through the lateral boundary coupling between both systems. In summary, variations of cluster sizes and spread in both ensembles can be regarded as relatively straightforward consequences of the subsampling algorithms used.

It is more difficult to explain why clustering-based subsamples have a better CRPS and ROCA than random ones. For clearly underdispersive parameters such as AROME-EPS temperature and wind, one can argue that it is linked to an improvement of reliability thanks to the added spread. This interpretation, though, is not convincing for precipitation, which is not underdispersive. It also fails to explain why the clear reduction of spread between the Molteni et al. (2001) and the 'penalized Ward' algorithms does not translate into visible degradations of CRPS and ROCA. We argue that clustering-based subsampling methods improve ensemble-based forecasts for reasons that are deeper than an enhancement of spread: the conversion of ensembles into point probability forecasts is, essentially, a numerical integration of the probability distribution functions (PDF) implied by the members. Thus, subsampling can be regarded as an approximate numerical integration of the PDF of the complete forcing ensemble: for instance, in the dataset used in this paper, we are trying to approximate the 34-member PEARP ensemble PDF by the 12 members used to force the AROME-EPS ensemble. In the CRPS and ROCA scores, point forecast probabilities are evaluated by counting the number of ensemble members below any given threshold, which is equivalent to a Riemann sum of the PDF. This computation is usually more accurate if it is performed on equally spaced points; intuitively, clustering methods attempt to find equally-sized partitions of the full ensemble, whereas the random subsampling selects integration points at random. If this interpretation is correct, one would expect point ensemble forecasts from our clustering-based subsamples to be more evenly distributed than random-based subsamples, among the members of the 34-member ensemble. Preliminary tests (not shown) with our dataset show that it is indeed the case, but a larger dataset is needed to substantiate it with good statistical significance. Thus, the impact of the subsampling algorithms on the precision of probability computations is left for a future study.

The results presented in this paper are based on raw ensemble output. It would be interesting to check whether they hold for dressed and/or calibrated ensemble output. One could argue that ensemble spread could be more easily increased by ensemble calibration than by tweaking the subsampling. A problem with calibration techniques is that they often improve probabilities at points only (i.e. they are univariate). By contrast, subsampling methods process the ensemble as a whole: the subsampled members provide complete scenarios for the evolution of the weather, preserving physical consistency between the modelled atmospheric parameters. Thus, optimising the subsampling step for limited area ensemble forecasts may be important for applications that are sensitive to the space and time consistency of the forecasts, such as human interpretation, or the derivation of multivariate weather information.

The techniques presented here should be useful for all limited area ensemble predictions that are forced by a subset of members in an equiprobable larger scale ensemble. At the time of writing, beside Météo-France that uses a setup close to the one used in this article, this problem is relevant for most European HIRLAM and ALADIN centres (e.g. Wang et al. 2011), at Meteoswiss, and at the Italian

ARPAE, most of which use the ensemble of the European Centre for Medium-Range Weather Forecasts (ECMWF). Limited area ensembles in the USA evolve rapidly, they are usually coupled to the global GEFS ensemble, which suggests similar questions.

This is not the case for the currently planned systems at NCEP (Du et al. 2015), Deutscher Wetterdienst (DWD) and the UK Met Office (Hagelin et al. 2017), where member selection is obviated by setting up the global ensembles with the same number of members as the respective limited area ensembles. Other centres, such as the Spanish Meteorological Service (García-Moya et al. 2011), or DWD until 2017 (Gebhardt et al. 2011) have driven their limited area ensembles with global multimodel ensembles that are clearly not equiprobable (different members are produced using different models). Thus, our approach is probably not relevant to them. Lee (2012) have shown that clustering algorithms can indeed be used to subsample multimodel ensembles, but with a setup that is quite different from ours (in order to manage variability in systematic model errors). We conclude from the above (incomplete) review of operational systems that the issues raised in this paper are relevant for many of them. The potential advantages of optimizing ensemble subsampling need to be weighed against the technical requirements of clustering algorithms, since they imply time-critical access to many fields in all members of the forcing ensemble.

6 Conclusion

We have demonstrated that the selection of forcing members can have a significant impact on limited area ensemble forecasts, in terms of probability scores and usability for human forecasters. The selection techniques based on clustering algorithms outperform the random selection in our tests. They tend to enhance ensemble dispersion. The popular selection method proposed by Molteni et al. (2001) appears to emphasize extreme weather scenarios. This feature is consistent with its historical focus on the detection of high precipitation events, but it can have drawbacks for other uses. In particular, the presence of unequally sized clusters in the selection procedure can be regarded as contradictory with the assumption that the members are equally likely, an assumption that is frequently made in the interpretation of ensemble forecasts.

In order to address this issue, we have developed a new selection procedure of forcing members, based on a modification of Ward' hierarchical clustering. The tests presented in this paper show that it reduces variations in cluster size, and it improves the consistency of a limited area ensemble with its forcing ensemble in terms of spread, while remaining superior to a random selection in terms of probability scores. In summary, we have shown that some weaknesses of limited area ensemble prediction can be addressed by a careful subsampling of the forcing ensemble.

Acknowledgements

This work has been funded by Météo-France and CNRS.

References

[1] Alhamed A, Lakshmivarahan S, Stensrud DJ. 2002. Cluster Analysis of Multimodel Ensemble Data from SAMEX. *Mon. Weather Rev.* 130: 226–256. DOI: 10.1175/1520-0493(2002)130<0226:CAOMED>2.0.CO;2

[2] Atger F. 1999. Tubing: An Alternative to Clustering for the Classification of Ensemble Forecasts. *Weather Forecast.* 14: 741–757. DOI: 10.1175/1520-0434(1999)014<0741:TAATCF>2.0.CO;2

[3] Bouttier F, Vié B, Nuissier O, Raynaud L. 2012. Impact of stochastic physics in a convection-permitting ensemble. *Mon. Weather Rev.* 140: 3706–3721. DOI: 10.1175/mwr-d-12-00031.1

[4] Bouttier F, Raynaud L, Nuissier O, Ménétrier B. 2016. Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Q. J. R. Meteorol. Soc.* 142: 390–403. DOI: 10.1002/qj.2622

[5] Branković C, Matjačić B, Ivatek-Šahdan S, Buizza R. 2008. Downscaling of ECMWF Ensemble Forecasts for Cases of Severe Weather: Ensemble Statistics and Cluster Analysis. *Mon. Weather Rev.* 136: 3323–3342 DOI: 10.1175/2008MWR2322.1

[6] Brill KF, Fracasso AR, Bailey CM. 2015. Applying a Divisive Clustering Algorithm to a Large Ensemble for Medium-Range Forecasting at the Weather Prediction Center. *Weather Forecast.* 30: 873–891. DOI: 10.1175/WAF-D-14-00137.1

[7] Clark AJ, Weiss SJ, Kain JS, Jirak IL, Coniglio M, Melick CJ, Siewert C, Sobash RA, Marsh PT, Dean AR, Xue M, Kong F, Thomas KW, Wang Y, Brewster K, Gao J, Wang X, Du J, Novak DR, Barthold FE, Bodner MJ, Levit JJ, Entwistle CB, Jensen TL, Correia J. 2012. An Overview of the 2010 Hazardous Weather Testbed Experimental Forecast Program Spring Experiment. *Bull. Am. Meteorol. Soc.* 93: 55–74. DOI: 10.1175/BAMS-D-11-00040.1

[8] 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.* 141: 1671–1685. DOI: 10.1002/qj.2469

[9] Du J, DiMego G, Zhou B, Jovic D, Ferrier B, Yang B. 2015. Regional ensemble forecast systems at NCEP. In *Proceedings of the 23rd Conf. on Numerical Weather Prediction and 27th Conf. on Weather Analysis and Forecasting, Chicago, IL, Amer. Meteor. Soc., June 29-July 3, 2015.* Paper 2A.5. Available at http://www.emc.ncep.noaa.gov/mmb/SREF/NWP2015_NCEP_RegionalEnsembles_paper.pdf (accessed on 22 March 2018)

[10] García-Moya, JA, Callado A, Escribà P, Santos C, Santos-Muñoz D, Simarro J. 2011. Predictability of short-range forecasting: a multimodel approach. *Tellus A* 63: 550–563. DOI: 10.1111/j.1600-0870.2010.00506.x

[11] Gebhardt C, Theis SE, Paulat M, Ben Bouallègue Z. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.* 100: 168–177. DOI: 10.1016/j.atmosres.2010.12.008

[12] Hacker J, Ha SY, Snyder C, Berner J, Eckel F, Kuchera E, Pocernich M, Rugg S, Schramm J, Wang X. 2011. The U.S. Air Force Weather Agency's mesoscale ensemble: scientific description and performance results. *Tellus A* 63: 625–641. DOI: 10.1111/j.1600-0870.2010.00497.x

[13] Hagelin S, Son J, Swinbank R, McCabe A, Roberts N, Tennant W. 2017. The Met Office convective-scale ensemble, MOGREPS-UK. *Q. J. R. Meteorol. Soc.* 143: 2846–2861 DOI: 10.1002/qj.3135

[14] Jolliffe IT, Stephenson DB. 2011. Forecast verification: a practitioner's guide in atmospheric science, 2nd edition. *John Wiley and Sons*, 292 pp. DOI: 10.1002/9781119960003.ch1

[15] Johnson A, Wang X, Kong F, Xue M. 2011a. Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 spring experiment. Part I: Development of the object-oriented cluster analysis method for precipitation fields. *Mon. Weather Rev.* 139: 3673–3693. DOI: 10.1175/MWR-D-11-00015.1

[16] Johnson A, Wang X, Kong F, Xue M. 2011b. Hierarchical cluster analysis of a convection-allowing ensemble during the Hazardous Weather Testbed 2009 spring experiment. Part II: ensemble clustering over the whole experiment period. *Mon. Weather Rev.* 139: 3694–3710. DOI: 10.1175/MWR-D-11-00016.1

[17] Keil C, Craig GC 2007. A Displacement-Based Error Measure Applied in a Regional Ensemble Forecasting System. *Mon. Weather Rev.* 135: 3248–3259. DOI: 10.1175/MWR3457.1

[18] Kowaleski AM, Evans JL. 2016. Regression mixture model clustering of multimodel ensemble forecasts of hurricane Sandy: partition characteristics. *Mon. Weather Rev.* 144: 3825–3846. DOI: 10.1175/MWR-D-16-0099.1

[19] Lee JA, Kolczynski WC, McCandless TC, Haupt SE. 2012. An objective methodology for configuring and down-selecting an NWP ensemble for low-level wind prediction. *Mon. Weather Rev.* 140: 2270–2286. DOI: 10.1175/MWR-D-11-00065.1

[20] Lee JA, Haupt SE, Young GS. 2016. Down-selecting numerical weather prediction multi-physics ensembles with hierarchical cluster analysis. *J. Climatol. Wea. Forecast.* 4: 156. DOI: 10.4172/2332-2594.1000156

[21] Li X, Charron M, Spacek L, Candille G. 2008. A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Weather Rev.* 136: 443–462. DOI: 10.1175/2007MWR2109.1

[22] Marsigli C, Montani A, Nerozzi F, Paccagnella T, Tibaldi S, Molteni F, Buizza R. 2001. A strategy for high-resolution ensemble prediction. II: Limited-area experiments in four Alpine flood events. *Q. J. R. Meteorol. Soc.* 127: 2095–2115. DOI: 10.1002/qj.49712757613

[23] Marsigli C, Boccanera F, Montani A, Paccagnella T. 2005. The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlin. Proc. Geophys.* 12: 527–536. DOI: 10.5194/npg-12-527-2005

[24] Molteni F, Buizza R, Marsigli C, Montani A, Nerozzi F, Paccagnella T. 2001. A strategy for high-resolution ensemble prediction. I: Definition of representative members and global-model experiments. *Q. J. R. Meteorol. Soc.* 127: 2069–2094. DOI: 10.1002/qj.49712757612

[25] Montani A, Cesari D, Marsigli C, Paccagnella T. 2011. Seven years of activity in the field of mesoscale ensemble forecasting by the COSMO-LEPS system: main achievements and open challenges. *Tellus A* 63: 605–624. DOI: 10.1111/j.1600-0870.2010.00499.x

[26] Nuissier O, Joly B, Vié B, Ducrocq V. 2012. Uncertainty on lateral boundary conditions in a convection-permitting ensemble: A strategy of selection for Mediterranean heavy precipitation events. *Nat. Hazards Earth Syst. Sci.* 12: 2993–3011. DOI: 10.5194/nhess-12-2993-2012

[27] Raynaud L, Bouttier F. 2018. The impact of horizontal resolution and ensemble size for convective-scale probabilistic forecasts. *Q. J. R. Meteorol. Soc. accepted author manuscript, published online on 9 Sept 2017.* Available at https://www.rmets.org/publications DOI: 10.1002/qj.3159

[28] Rousseuw P. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comp. and Applied Math.* 20: 53-65.

[29] Saito K, Hara M, Kunii M, Seko H, Yamaguchi M. 2011. Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the Meteorological Research Institute for the WWRP Beijing 2008 Olympics Research and Development Project (B08RDP). *Tellus A* 63: 445–467. DOI: 10.1111/j.1600-0870.2010.00509.x

[30] Seity Y, Brousseau P, Malardel S, Hello G, Bénard P, Bouttier F, Lac C, Masson V. 2011. The AROME-France convective scale operational model. *Mon. Weather Rev.* 139: 976–999. DOI: 10.1175/2010MWR3425.1

[31] Tien SL, Taylor C. 2014. Comparing and clustering ensemble forecast members to support strategic planning in air traffic flow management. In *Proceedings of the 2014 Conference of the American Meteorological Society, Atlanta, GA 2-6 Feb 2014.* Available at https://ams.confex.com/ams/94Annual/webprogram/Manuscript/ Paper237042/pr_14-0744.pdf (accessed on 22 Mar 2018)

[32] Vié B, Nuissier O, Ducrocq V. 2011. Cloud-resolving ensemble simulations of Mediterranean heavy precipitating events: uncertainty on initial conditions and lateral boundary conditions. *Mon. Weather Rev.* 139: 403–423. DOI: 10.1175/2010mwr3487.1

[33] Wang Y, Bellus M, Wittmann C, Steinheimer M, Weidle F, Kann A, Ivatek-Sahdan S, Tian W, Ma X, Bazile E. 2011. The Central European limited area ensemble forecasting system: ALADIN-LAEF. *Q. J. R. Meteorol. Soc.* 137: 483–502. DOI: 10.1002/qj.751

[34] Ward JH. 1963. Hierarchical grouping to optimize an objective function *J. Am. Stat. Assoc.* 58: 236–244. DOI: 10.1080/01621459.1963.10500845

[35] Weidle, F, Wang Y, Tian W, Wang T. 2013. Validation of strategies using clustering analysis of ECMWF EPS for initial perturbations in a limited area model ensemble prediction system. *Atmos. Ocean* 51: 284–295. DOI: 10.1080/07055900.2013.802217

[36] Yussouf N, Stensrud DJ, Lakshmivarahan S. 2004. Cluster analysis of multimodel ensemble data over New England. *Mon. Weather Rev.* 132: 2452–2462. DOI: 10.1175/1520-0493(2004)132<2452:CAOMED>2.0.CO;2