Cloud-Resolving Ensemble Simulations of Mediterranean Heavy Precipitating Events: Uncertainty on Initial Conditions and Lateral Boundary Conditions

BENOÎT VIE, OLIVIER NUISSIER, AND VÉRONIQUE DUCROCOQ
CNRM-GAME, Méteo-France/CNRS, Toulouse, France

(Manuscript received 4 May 2010, in final form 28 September 2010)

ABSTRACT

This study assesses the impact of uncertainty on convective-scale initial conditions (ICs) and the uncertainty on lateral boundary conditions (LBCs) in cloud-resolving simulations with the Application of Research to Operations at Mesoscale (AROME) model. Special attention is paid to Mediterranean heavy precipitating events (HPEs). The goal is achieved by comparing high-resolution ensembles generated by different methods. First, an ensemble data assimilation technique has been used for assimilation of perturbed observations to generate different convective-scale ICs. Second, three ensembles used LBCs prescribed by the members of a global short-range ensemble prediction system (EPS). All ensembles obtained were then evaluated over 31- and/or 18-day periods, and on 2 specific case studies of HPEs. The ensembles are underdispersive, but both the probabilistic evaluation of their overall performance and the two case studies confirm that they can provide useful probabilistic information for the HPEs considered.

The uncertainty on convective-scale ICs is shown to have an impact at short range (under 12 h), and it is strongly dependent on the synoptic-scale context. Specifically, given a marked circulation near the area of interest, the imposed LBCs rapidly overwhelm the initial differences, greatly reducing the spread of the ensemble. The uncertainty on LBCs shows an impact at longer range, as the spread in the coupling global ensemble increases, but it also depends on the synoptic-scale conditions and their predictability.

1. Introduction

Heavy precipitating events (HPEs) frequently occur over the western Mediterranean regions, mainly in autumn, producing disastrous flash floods that have been responsible for many casualties and heavy damages in recent decades. Such high-impact events have been well observed and studied over all the western Mediterranean regions (see Buzzi et al.1998; Penarrocha et al. 2002; Ducrocq et al. 2002; Nuissier et al. 2008; Ducrocq et al. 2008, among others). While large amounts of precipitation can accumulate over several days when frontal systems are slowed down and strengthened by coastal mountains, some of these Mediterranean HPEs can be attributed to back-building quasi-stationary mesoscale convective systems (MCSs) staying over the same area for several hours, producing large rainfall totals in a very short time (typically over 200 mm in 6–24 h).

Using nonhydrostatic cloud-resolving models (CRMs) improves the realism of simulated precipitating systems, because of enhanced representation of the water cycle and resolved deep convection. The HPEs considered here depend strongly on both synoptic-scale circulation patterns (Nuissier et al. 2008, 2010, manuscript submitted to Quart. J. Roy. Meteor. Soc.) and convective-scale flow characteristics and processes (Ducrocq et al. 2008; Bresson et al. 2009). The intrinsic predictability of these HPEs is directly related to the variety of physical processes involved in their triggering and evolution. Strong, slowly evolving synoptic-scale conditions (e.g., a large upper-level trough northwest of the region of interest inducing a diffusent southerly flow over the Mediterranean regions) help make these HPEs more predictable. Other mechanisms also increase the predictability, such as the triggering of deep convection by orographic lifting. In contrast, other processes, in particular at the convective scale (e.g., microphysical processes, turbulence, and convective instability) and their complex interactions, involving nonlinearities and threshold effects, reduce that predictability. Precise forecasting of both the location and intensity of the quasi-stationary MCSs involved in
HPEs is thus still very challenging, and yet the hydrological response of the Mediterranean steep coastal watersheds is very sensitive to the precise location of the heaviest precipitation (Chancibault et al. 2006; Vincendon et al. 2010). The hydrological runoff forecasts may be strongly affected by the uncertainty of the atmospheric rainfall forecasts. Walser et al. (2004) have studied the predictability of precipitation in CRMs and suggest that, for convective precipitation, the uncertainty may be so large that runoff forecasts derived from a single deterministic CRM may have no value.

Quantifying the uncertainty associated with cloud-resolving forecasts is therefore a crucial issue. Probabilistic forecasts have been a subject of research since the 1960s. At that time, Epstein (1969) proposed integrating the Liouville equation, which describes the temporal evolution of a probability density function (PDF). Both the imperfect knowledge of the initial PDF and the prohibitive numerical cost of this integration led to the development of ensemble prediction (Leith 1974). Using ensemble prediction systems (EPSs) is now a well-known approach, which has been developed since the 1990s for medium-range synoptic-scale forecasting. Such global operational EPSs generate a set of atmospheric states sampling the probability density function for the initial state with different techniques. Global EPSs were first implemented at the European Centre for Medium-Range Weather Forecasts (ECMWF) using the computation of singular vectors (Molteni et al. 1996), and at the National Centers for Environmental Prediction (NCEP) through the breeding modes technique. The Meteorological Service of Canada used an ensemble data assimilation technique and, more recently, an ensemble Kalman filter to generate the initial conditions (ICs) of their ensemble (Houtekamer et al. 2009). The characteristics and performance of ensembles generated by these different techniques have been well studied and compared (Descamps and Talagrand 2007; Buizza et al. 2005; Hamill et al. 2000).

However, these methods designed to generate a global, large-scale ensemble cannot be easily adapted to limited-area CRMs. Hohenegger and Schär (2007b) found that the tangent-linear approximation for CRMs has no value beyond 7 h. As physical parameterizations at smaller scales show stronger nonlinearities, convective-scale perturbations grow much faster and even impact the large-scale predictability. The sensitivity to ICs (Ducrocq et al. 2002) is also different between parameterized and resolved convection. Moreover, in addition to the uncertainties on ICs and the model errors, cloud-resolving ensembles must also consider the uncertainty due to lateral boundary conditions (LBCs), since they are run over a limited area. Through a spectral approach of the one-way nesting, Laprise (2003) investigated the different resolved scales between global and limited-area models. He showed that long wavelengths cannot be represented in the LAM and thus are only introduced by the LBCs. All of these reasons, as well as the much higher computing time at fine resolution, call for the development of a dedicated ensemble generation method that samples all the uncertainty sources.

Recent studies make use of a wide range of techniques to create ensembles with regional, limited-area, convection-parameterizing models (typically at a 10–20-km resolution). Zhang et al. (2006) applied an ensemble Kalman filter (EnKF) approach to generate different ICs. Stensrud et al. (2000) showed that the impact of model errors (represented by different physical parameterizations) and uncertainty on ICs depends on the large-scale forcing for upward motion. Grimit and Mass (2002) used multimodel large-scale conditions to initialize and drive their ensemble. Marsigli et al. (2001) and Molteni et al. (2001) used a selection of a few representative members from the global ECMWF EPS to drive their regional model. These last three studies use different LBCs and different ICs for each member. Nutter et al. (2004) evidenced the link between coarsely resolved, temporally interpolated LBCs and ensemble underdispersion.

For convection-resolving ensembles, various generation methods have also been explored. Hohenegger et al. (2008) applied a clustering technique on global EPS members to both generate CRM ICs and provide coupling conditions. They showed that the benefit of using a CRM ensemble depended on the synoptic-scale conditions. They also found that the LBCs tended to overwhelm the differences in ICs after about 12 h. Walser et al. (2004) used a shifted initialization technique, with an amplification of the ensemble spread at the beginning of the comparison day. Through four case studies, they suggest that convective activity reduces predictability at scales much larger than the convective cells. Hohenegger and Schär (2007a) compared the shifted initialization technique to two different methods of perturbation of the initial fields: applying a local Gaussian-shaped temperature perturbation in the domain, or adding random perturbations at each point of the 3D temperature field. They concluded that the three ensembles were approximately equivalent and highlighted the same area of limited predictability. They also showed that initial differences can grow and propagate to affect the whole domain in a few hours.

Few works have assessed the behavior of cloud-resolving ensembles in terms of probabilistic forecasts based upon a long evaluation period. During April–June 2007, as part of the National Oceanic and Atmospheric Administration (NOAA) Hazardous Weather Test Bed
Spring Experiment, a real-time convection-resolving (4-km grid mesh) ensemble experiment was conducted (Kong et al. 2007). The ensemble members were generated both through perturbations of initial states from a control member and through different physical parameterizations. Clark et al. (2009) compared a subset of this CRM ensemble to a regional ensemble over a selection of 23 days. They found that the high-resolution ensemble, despite a lower number of members, provided better forecasts.

Although the benefit of high-resolution ensembles is demonstrated in these studies, most of them are based on a limited number of cases. They do not provide much information about the impact of the different sources of uncertainty on CRM forecasts, and often involve subjective tuning of perturbations. The most important points concern the relative impact of each source of uncertainty on the forecast and their respective links with the atmospheric situations. Moreover, the method providing the best sampling of each uncertainty source and/or the best forecast results needs to be assessed. Finally, if a high-resolution EPS has to sample all the uncertainty sources, it is necessary to find their optimal combination through a limited number of members. Roebber et al. (2004) highlighted the advantages and disadvantages of high-resolution modeling and ensemble forecasts, and identified some major remaining challenges. At this time, these scientific issues about the methods for generating a cloud-resolving ensemble remain open questions.

The present work assesses the different impact on CRM forecasts of uncertainties on convective-scale ICs and synoptic-scale LBCs, assuming the perfect-model hypothesis. For that purpose, it sets up a new framework in an operational context. Four distinct ensembles are designed to separately sample these two uncertainty sources, and are evaluated first over a one-month period and then for two specific case studies.

Based upon the operational forecast suite at Météo-France, the high-resolution ensemble experiments benefit from convective-scale data assimilation in a 3-hourly rapid update cycle for the whole month. In our strategy, the impact of uncertainty on synoptic-scale LBCs is represented through three ensembles using LBCs provided by a global EPS. A last ensemble assesses the uncertainty on convective-scale ICs through a convective-scale ensemble data assimilation technique. Instead of modifying the ICs for each member directly, before each data assimilation step, the observations are randomly perturbed according to their observational error. This technique is known to provide a good sample of the analysis error (Berre et al. 2006) and performs well at computing covariance matrices of background errors, but had not yet been used to generate a cloud-resolving EPS.

The numerical experiments set up and the scores used are described in section 2. Section 3 presents a probabilistic evaluation of these ensembles over a one-month period. Section 4 provides a detailed analysis of two specific cases of Mediterranean HPEs. Conclusions are presented in the final section.

2. Description of the ensemble experiments
   a. The cloud-resolving model: AROME

This study was performed with the operational CRM Application of Research to Operations at Mesoscale (AROME) from Météo-France (Bouttier 2007; Seity et al. 2010). The model configuration operational in late 2008 was used. AROME was run at a horizontal resolution of about 2.5 km over a domain mainly covering France. The vertical grid comprised 41 vertical levels. AROME is based on the nonhydrostatic version of the adiabatic equations of the limited-area model Aire Limitee Adaptation Dynamique Developpement International (ALADIN; Benard 2004; Bubnova et al. 1995), using physical parameterizations from the research model Méso-NH (Laforte et al. 1998). The bulk microphysics scheme following Caniaux et al. (1994) governs the prognostic equations of the six water variables: water vapor, cloud water, rainwater, primary ice, graupel, and snow. Shallow convection is parameterized by the eddy diffusivity Kain–Fritsch (EDKF) scheme (Pergaud et al. 2009) and the turbulent scheme follows Cuxart et al. (2000). No deep convection parameterization is used, as it is considered that deep convection is explicitly resolved at this resolution.

AROME uses LBCs interpolated from ALADIN forecasts (about 10-km horizontal resolution over France), with a coupling frequency of 3 h. ALADIN is itself driven by the global model Action de Recherche Petite Echelle Grande Echelle (ARPEGE, about 15-km horizontal resolution over France), also with a coupling frequency of 3 h. ALADIN is used as an intermediate step in the down-scaling of synoptic-scale circulations to prevent too large a gap in resolution between AROME and the applied LBCs. Both ALADIN and AROME have their own data assimilation cycle based on a three-dimensional variational data assimilation (3D-VAR) scheme. A comprehensive description of the ALADIN 3D-VAR data assimilation scheme used in the present study has been given by Fischer et al. (2005) and Montmerle et al. (2007). The AROME 3DVAR assimilation system is based on the ALADIN 3DVAR one, but with different background and observation statistics to cope with the finer resolution of the AROME model better (Yan et al. 2009; Boniface et al. 2009). The observation thinning is also less restrictive for AROME, allowing assimilation
of observations at higher spatial resolutions. The background error covariances are estimated by an ensemble-based method (Berre et al. 2006) applied to the ALADIN and AROME systems, individually. The AROME data assimilation system uses a rapid forward sequential assimilation cycle with a 3-hourly data analysis frequency, whereas the ALADIN one is updated at only a 6-hourly data analysis frequency.

The observations assimilated in the 3D-VAR ALADIN and AROME systems as of October 2008 included those from radio soundings, screen-level stations, wind profilers, GPS, buoys, ships, and aircraft. Assimilated satellite data included horizontal winds from atmospheric motion vectors (AMVs) and the QuickSCAT scatterometers, Advanced Microwave Sounding Unit (AMSU)-A and -B radiances from the NOAA-15, -16, -17, and the Aqua satellites, High-resolution Infrared Sounder (HIRS) radiances from NOAA-17, and clear-air Spinning Enhanced Visible and Infrared Imager (SEVIRI) radiances from the Meteosat-8 satellite. In addition, AROME assimilates Doppler radial winds from the weather radar network over France.

Three ensembles were designed to assess the impact of uncertainty on large-scale LBCs by driving the AROME simulations with the members of the Météo-France global short-range EPS Prévision d’Ensemble ARPEGE (PEARP) (downscaled through ALADIN). One more ensemble was created to assess the impact of uncertainty on convective-scale ICs through an ensemble data assimilation technique. The four ensembles are presented below and their features are briefly summarized in Table 1.

b. The AROME–PEARP ensemble experiments

The PEARP EPS used in this study has 11 members, and considers only errors in ICs. The PEARP system uses a stretched horizontal grid over the globe, allowing a grid spacing of about 23 km over western Europe. The control run is similar to the ARPEGE operational deterministic forecast. The ICs for the other 10 members are obtained by adding perturbations to the control run. The perturbation method blends a breeding technique and calculation of 12-h dry singular vectors. The perturbations are computed to get an optimal spread at a 12-h forecast range. For this study, the PEARP was run 4 times daily at 0000, 0600, 1200, and 1800 UTC. The 24-h forecasts are issued at 1200 UTC, whereas only 6-h forecasts are run at 0600, 1200, and 1800 UTC.

1) AROME–PEARP1

Figure 1 describes the model chain and set-up of the AROME–PEARP1 experiment. The figure shows a 36-h sequence of our experiments. The 24-h forecasts issued at 1200 UTC, and the preceding assimilation cycles are drawn in black. The continuing assimilation cycles after 1200 UTC, which will be used to issue the next-day forecasts, are represented in light gray. The first line shows PEARP forecasts.

For each PEARP member, the following downscaling procedure was applied, resulting in an 11-member AROME ensemble. The downscaling of the PEARP members at the AROME resolution follows the Météo-France NWP suite operating in late 2008, except that the ARPEGE deterministic forecast is replaced by each PEARP member forecast.

First, a 6-hourly data assimilation cycle was carried out with the ALADIN system (middle row of Fig. 1), using the PEARP forecast as LBCs. LBCs were updated every 3 h. For instance, the LBCs used in the 6-h ALADIN forecast run starting at 1800 UTC were the 0-, 3-, and 6-h forecasts arising from the PEARP run starting at 1800 UTC.

Then, the ALADIN forecast from the ALADIN data assimilation cycle was used as LBCs to perform a 3-hourly AROME data assimilation cycle (bottom row of Fig. 1). AROME LBCs were updated every 3 h. The 24-h forecasts were performed at 1200 UTC every day. This ensemble will be called AROME-PEARP1 hereafter.

2) AROME–PEARP2

A second experiment (hereafter AROME–PEARP2) has been performed by removing the assimilation of observations in the ALADIN 3D-VAR scheme from the LBCs downscaling procedure. ALADIN performed only a dynamical downscaling the PEARP forecasts. By doing this, the LBCs used for the AROME runs should be closer to the original PEARP forecast. Since the ALADIN data assimilation reduces the spread between
the ALADIN forecasts used as LBCs for the AROME forecasts, the spread of the AROME–PEARP2 ensemble should grow faster.

3) AROME–PEARP3

Since the PEARP ensemble spread grows over time to become optimum between 12 and 24 h, another experiment (hereafter AROME–PEARP3) was set up using a PEARP run 6 h earlier to provide LBCs to the AROME simulations. For example, instead of using LBCs coming from the 1200 UTC PEARP run for an AROME run starting at 1200 UTC, the 0600 UTC PEARP run was extended up to 30 h and then down-scaled through ALADIN. The spread of the PEARP ensemble thus had more time to grow. However, this meant that the forecast errors also had more time to grow. The benefit of using an earlier PEARP run as LBCs was thus not obvious and was worth examining through the AROME–PEARP3 experiment. As in AROME–PEARP2, ALADIN was run for dynamical downscaling without data assimilation.

c. The AROME–PERTOBS ensemble experiment

To assess the impact of uncertainty on convective-scale ICs, an ensemble data assimilation technique, as described by Berre et al. (2006) and Houtekamer et al. (1996), has been used. The variational assimilation of randomly perturbed observations (in the observational error range) sampled the analysis error. The uncertainty on LBCs was not considered in this experiment. The mesoscale deterministic ALADIN forecast, for which LBCs are provided by the ARPEGE deterministic forecast, provided the same coupling conditions for all members, as for the deterministic AROME forecast.

In this ensemble experiment, 11 parallel AROME data assimilation cycles were run, as shown by Fig. 2. The control member (P0), similar to the operational deterministic forecast, used unperturbed observations through the 3-hourly data assimilation cycle. The 10 perturbed members used randomly perturbed observations following a Monte Carlo approach: each observation $o$ was extracted from the observations database with its

---

**Fig. 1.** The AROME–PEARP1 ensemble experiment. Plain (thick) arrows represent short (long) model runs. Vertical dashed arrows indicate coupling relationships. Each PEARP member provides LBCs to one ALADIN data assimilation cycle, itself providing LBCs to one AROME data assimilation cycle with data analysis every 3 h. Each day, 24-h AROME forecasts are run, starting from the 1200 UTC analysis of each AROME data assimilation cycle. The continuing assimilation cycles after 1200 UTC (independent from the 24-h forecast) are shown in light gray.
observational error $\sigma_o$, a normally distributed random number $r$ was drawn (with 0 expectation, and unit variance), and the perturbed observation was computed as $o + \sigma_o r$. Thus, the assimilation of randomly perturbed observations every 3 h created different ICs for the AROME members. This 11-member ensemble will be called AROME–PERTOBS hereafter.

d. Probabilistic scores

Different measures can be used to evaluate the statistical behavior of EPSs. The scores used in this study are briefly presented below, a more detailed description of probabilistic evaluation of ensemble forecasts can be found in Talagrand et al. (1997) and Toth et al. (2003).

1) RELATIVE OPERATING CHARACTERISTICS

The ROC is a measure of the resolution of the ensemble relative to binary events. For a given probability threshold $s$ between 0 and 1, the probabilistic forecast is transformed into a deterministic one, considering that the event is predicted if the probability is $p \geq s$. A $2 \times 2$ contingency table is built from all forecast–observation pairs (every grid point and date of simulation) and the probability of detection (POD) and false-alarm rate (FAR) are thus computed similarly to the deterministic scores. The ROC curve is then obtained as a plot of POD($s$) versus FAR($s$). The area under the ROC curve measures the quality of the ensemble: the larger the area, the better the resolution of the ensemble.

In this study, the thresholds used were $s = 1/11$, $\ldots$, $s = 10/11$, $s = 1$. With 11-member ensembles, $p \geq s = k/11$ means that at least $k$ members predicted the event.

2) RELIABILITY DIAGRAM

Reliability diagrams were also computed, to assess whether the forecast probabilities fit the observed frequencies. If the computed probabilities are accurate, any given event should happen 30% of the times when it is predicted with a probability of 30%, so the curve of observed frequencies plotted against forecast probability should be close to the diagonal for a good ensemble forecast.

3) RANK HISTOGRAM

The rank histogram is a measure of the statistical consistency of the ensemble forecast (Talagrand et al. 1997). Let us consider a scalar variable $x$. At one time and location, an $N$-member ensemble gives us $N$ forecast values of $x$, which define $N + 1$ intervals ($-\infty < x_1 < x_2 \ldots < x_N < +\infty$). The verification value is in one of these $N + 1$ intervals. By repeating this process for each forecast and observation pair (at each verification point and for every ensemble forecast) and counting the
verification values in each interval, we build the rank histogram.

The flatter the histogram, the better is the ensemble reliability. A U-shaped histogram is the sign of an under-dispersive ensemble (the verification value is too often outside the ensemble extrema), a J-shaped histogram is the sign of a biased and/or underdispersive ensemble.

Building and interpreting a rank histogram with precipitation data is difficult as precipitation values are often null. For global EPSs, rank histograms are often calculated for the higher troposphere (e.g., \(Z_{500}\)). As this study focused on HPEs, low-level convective-scale parameters closely related to the triggering, intensity, and stationarity of HPEs were selected instead (e.g., temperature, humidity at 925 hPa, etc.).

4) BRIER SKILL SCORE

The Brier score (BS) is a measure of the forecast of a binary event, for instance precipitation exceeding a given threshold, and is defined as follows:

\[
BS = \frac{1}{M} \sum_{i=1}^{M} (p_i - o_i)^2,
\]

where \(M\) is the number of forecast–observation pairs used to compute BS, \(p_i\) is the forecast probability for this event to happen, and \(o_i = 1\) if the event happened, \(o_i = 0\) if not. A perfect deterministic forecast has a BS equal to 0, and the worst forecast has a BS equal to 1. The BS obtained by a given forecast can then be compared to the one for a reference forecast by computing the Brier skill score (BSS):

\[
BSS = 1 - \frac{BS}{BS_{ref}}.
\]

A perfect forecast has BSS = 1. In this study, the deterministic AROME forecast was chosen as the reference.

5) INFLUENCE OF THE OBSERVATION ERRORS ON THE SCORES

These probabilistic scores assume that the observations used as the reference are perfect. Some studies have shown that considering the uncertainties on observations can change the computed scores (Candille and Talagrand 2008; Bowler 2006). Among the different methods available to account for observation uncertainty, the perturbed ensemble method was used in this study: scores were computed first from the forecast precipitation fields, and then recomputed after adding a random noise to all the forecast precipitation values. The differences in scores after addition of a Gaussian noise sampling a 10% error in precipitation observations were minimal, and are therefore not shown in this paper.

3. Overall performance of the ensembles

The AROME–PEARP1 and AROME–PERTOBS ensembles were run over one full month from 6 October 2008 to 5 November 2008 inclusive. This time window was chosen because of the occurrence of 4 HPEs during that month (7, 8, 20, and 21 October, and 1–2 November). The period also included different atmospheric conditions, with drier days in the first two weeks and rainy conditions afterward. Important convective activity (identified by precipitation and lightning impacts) occurred on 7, 8, and 20–24 October 2008, and from 30 October to 5 November 2008, mostly over southeastern France and the Mediterranean Sea.

An examination of the AROME–PEARP1 and AROME–PERTOBS ensembles over shorter periods included in the 31-day period showed that their behavior is not significantly different, and the conclusions are unchanged. Therefore, to save computing time, the AROME–PEARP2 and AROME–PEARP3 data assimilation cycles were run only over 18 days, from 0000 UTC 15 October 2008 to 1200 UTC 1 November 2008 inclusive. They thus encompassed three of the four HPEs. These two ensembles were initialized at 0000 UTC 15 October 2008, with the AROME–PEARP1 3-h forecasts.

a. Comparison between AROME–PEARP1 and AROME–PERTOBS

1) PRECIPITATION

The ROC scores computed for 24-h accumulated precipitation forecasts against the observations are shown in Figs. 3a,b. The observations came from the Météo-France mesoscale hourly surface station network over France. Approximately 53,500 observations were used to compute the scores over the 31-day period. For the rain–no-rain threshold (0.5 mm, Fig. 3a), both ensembles perform very well and have similar ROC values, showing that both ensembles have a good resolution. For intermediate thresholds (10 mm, Fig. 3b), the areas under the ROC curves are smaller than for the rain–no-rain threshold. The AROME–PEARP1 ensemble shows a slightly better resolution than the AROME–PERTOBS one.

Higher thresholds (not shown) give ROC values around the diagonal, which indicate that the ensembles have no resolution, still with an advantage for the AROME–PEARP1 ensemble. These results may not be representative, however, due to the diminishing number of cases: about 150 observations over 100 mm and 550 observations over 50 mm.
The reliability diagram (Fig. 3c) computed for precipitation intervals shows curves close to the diagonal indicating good reliability of the two ensembles. Both ensembles are slightly overconfident, the AROME–PERTOBS ensemble being a little more reliable than the AROME–PEARP1 ensemble.

BSSs computed for different thresholds are shown in Table 2. The BSS is positive for both ensembles and every threshold, which indicates that the two ensembles produce valuable probabilistic information when compared to the deterministic forecast. The AROME-PEARP1 ensemble gives better scores throughout the range of precipitation thresholds. It is also worth noting that both ensembles obtain higher scores for the highest thresholds, suggesting that they provide even better information (cf. the deterministic forecast) for the HPEs in which we are interested.

The ROC and the reliability diagram were also computed limiting the validation sample to HPE days. A day was considered as an HPE one if at least one rain gauge observation exceeded 100 mm (24 h). Over the 31-day period, 11 days matched this criterion: 7, 8, 20, 21, 22, 23, 30, and 31 October and 1, 3, and 4 November 2008. Figure 4 shows the precipitation scores for these days. They are very similar to the monthly scores overall. For higher precipitation thresholds, such as 50 mm (Fig. 4b), both ensembles perform a little better than over the 31-day period (not shown).

2) OTHER LOW-LEVEL PARAMETERS

A validation against surface observations for 2-m temperature, 2-m humidity, and 10-m winds was also carried out at different forecast ranges for the whole month. Since all three variables behave similarly, only

<table>
<thead>
<tr>
<th>Threshold (mm)</th>
<th>0.5</th>
<th>2</th>
<th>5</th>
<th>10</th>
<th>20</th>
<th>30</th>
<th>40</th>
<th>50</th>
<th>75</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>AROME–PEARP1</td>
<td>0.32</td>
<td>0.32</td>
<td>0.29</td>
<td>0.28</td>
<td>0.31</td>
<td>0.30</td>
<td>0.31</td>
<td>0.27</td>
<td>0.37</td>
<td>0.40</td>
</tr>
<tr>
<td>AROME–PERTOBS</td>
<td>0.32</td>
<td>0.32</td>
<td>0.28</td>
<td>0.22</td>
<td>0.24</td>
<td>0.26</td>
<td>0.27</td>
<td>0.23</td>
<td>0.30</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 2. The BSSs computed for different precipitation thresholds for both the AROME–PEARP1 and AROME–PERTOBS ensembles. The reference is the deterministic AROME forecast.
results for 10-m winds are shown in the following. Figures 3d,e,f show the ROC scores and reliability diagrams for 10-m wind speed at a 24-h forecast range. The ROC curves for the 2 and 5 m s\(^{-1}\) thresholds (Figs. 3d,e) again show that both ensembles have good resolution. The reliability diagram (Fig. 3f) shows strong overconfidence for high forecast probabilities: events forecast with a probability of 1 have an observed frequency of about 60%. This may be related to the underdispersion of both ensembles for low-level parameters (evidenced by rank histograms, see below). If all members predict a 10-m wind speed between 2 and 3 m s\(^{-1}\), since the ensemble is underdispersive, there is a chance that the observation will fall out of that range.

Rank histograms were computed for mesoscale low-level parameters that are known to have an impact on the triggering and development of Mediterranean HPEs (Ducrocq et al. 2008), such as the 925-hPa wind speed. Considering the few available observations at that height, the 3-hourly high-resolution deterministic AROME analyses were used as reference.

The AROME–PEARP1 ensemble shows marked underdispersion at all forecast ranges (Figs. 5a–d). Only the convective-scale data assimilation performed in
AROME helps it remain close to the observations in the ICs. The histogram degrades during the first 12 h, but is slightly better after 24 h than at a 12-h forecast range. This is consistent with the perturbation used to generate the coupling PEARP ensemble. The singular vectors calculated for the global short-range ensemble PEARP have a maximum perturbation growth over the mid-tropospheric region, and these perturbations need time to propagate downward and impact the lowest levels. Thus, histograms computed at 500 hPa indicate that the ensemble has a much better spread at this height. Moreover, the LBCs propagate gradually inside the AROME domain and have a growing impact with time. The computation of ensemble spread shows that it grows over time, although it is not large enough after 24 h compared to the forecast error (not shown). Another effect contributing to the underdispersion of the ensemble is the data assimilation performed in the ALADIN cycle (as shown below).

Other low-level parameters (e.g., temperature, humidity, and conditional instability) behave like wind speed. The mean sea level pressure as well as the 500-hPa geopotential height have different histograms, exhibiting a bias of the ensemble (not shown).

Rank histograms computed for the AROME–PERTOBS ensemble are shown for ICs and forecast ranges of 3, 6, and 12 h in Figs. 5e–h. The ensemble spread is better over the early hours of simulation compared to the AROME–PEARP1 ensemble, because of the greater differences in ICs. But, as the initial differences are overwhelmed by the LBCs, the ensemble spread vanishes over time.

b. Scores for the AROME–PEARP2 and AROME–PEARP3 experiments

Figure 6 shows rank histograms for all three AROME–PEARP ensembles, for forecast ranges of 3, 6, 12, and 24 h. The benefit of removing the ALADIN data assimilation step from the downscaling procedure is shown by the better histogram for the AROME–PEARP2 ensemble compared to the one of AROME–PEARP1. Performing a dynamical downscaling instead of a data assimilation cycle with ALADIN gives LBCs closer to the PEARP runs and thus a larger spread. The gain is significant at all forecast ranges.

Using an earlier PEARP run to provide LBCs further enhanced the histograms, as shown by the fact that the AROME–PEARP3 ensemble produces the best histogram. Compared to the AROME–PEARP2 ensemble, there is an improvement at all forecast ranges. The benefit is greater at short ranges (up to 12 h). When the PEARP starting time is shifted by 6 h, the gain in spread outweighs the error growth in the meantime. The differences between the rank histograms for the AROME–PEARP2 and AROME–PEARP3 ensembles become smaller after 24 h. All three AROME–PEARP ensembles have very similar ROCs, reliability diagrams, and BSSs (not shown).

c. Impact of the synoptic-scale conditions

The synoptic-scale circulation seems a dominant factor controlling the relative impact of ICs and LBCs. The quality of daily rank histograms was evaluated through a “flatness factor” defined as the ratio between the height of the two side columns and the height of the two middle columns. Correlations between this quality and the average of different atmospheric parameters (such as the 500-hPa wind speed or 24-h accumulated precipitation) over the AROME domain were computed. No significant correlation was found for the AROME–PERTOBS ensemble. The quality of AROME–PEARPs ensemble histograms is well correlated to both the mean 500-hPa wind speed (at a forecast range of 12 or 24 h) and the 24-h accumulated precipitation. Therefore, histograms and probabilistic scores were computed separately for days with a mean 500-hPa wind speed (at a forecast range of 12 or 24 h) and the 24-h accumulated precipitation. Therefore, histograms and probabilistic scores were computed separately for days with a mean 500-hPa wind speed at a 12-h forecast range above (under) the monthly median. Figure 7 shows rank histograms at 12 h for days with strong (weak) synoptic-scale forcing. The AROME–PEARP ensembles perform better for days with a strong synoptic circulation (Fig. 7a).

Probabilistic scores for precipitation for days with a strong synoptic forcing show little differences between
ensembles for low thresholds, and an advantage for the AROME–PEARP1 ensemble for thresholds over 10 mm (not shown). For days with a weak synoptic forcing (which are also days with little precipitation, so the scores may be less significant), the differences are a bit larger, the AROME–PERTOBS ensemble is better for low thresholds, whereas the AROME–PEARP1 ensemble performs better over 10 mm (not shown).

4. Detailed analysis for two HPEs

Following the probabilistic and overall evaluation of the two ensembles, this section presents an analysis and discussion for two of the HPEs that occurred during the 31-day period. They were chosen because they were significantly different, and associated with two different meteorological situations. The first one is a good example of convective precipitation developing in a quasi-stationary frontal system. The second case is an isolated MCS developing in a much weaker synoptic-scale circulation.

a. Description of the cases

1) CASE 1: 1–2 NOVEMBER 2008

Precipitation began progressively in the morning of 1 November 2008, associated with a trough over western France. Convection formed by noon and radar reflectivities showed an organized line of high reflectivities at 1400 UTC, which strengthened over the night with reflectivities as high as 60 dBZ (Fig. 8c). By the end of the night and in the morning of 2 November 2008, convective cells had formed over the sea and the convective system moved southwestward to produce high precipitation over the Aude region (Fig. 8d, the Gard and Aude regions are shown in Fig. 8b). Rainfall amounts for case 1 were high, reaching 365 mm in 24 h (Fig. 8e).

The analysis of low-level conditions (Fig. 8b) clearly shows a very strong low-level jet that brings moist, unstable air to the Massif Central foothills. This south to southeasterly jet at 0000 UTC takes a more easterly component beyond 0900 UTC. The upper-level synoptic-scale conditions are shown in Fig. 8a for 0000 UTC 2 November 2008. An upper-level low is cut off from the polar air mass west of the region, evidenced by the 1.5-PVU surface height. It induces a strong upper-level southwesterly flow over southeastern France, with a growing divergence over time. The airmass instability is thus increased.

2) CASE 2: 20 OCTOBER 2008

Case 2 was a completely different type of HPE. As shown by Fig. 9a, the synoptic-scale context was characterized by much weaker baroclinic activity over southern France, with only a small anomaly in 1.5-PVU surface height over southeastern France. The quasi-stationary MCS is not located over the Massif Central foothills but upstream over the plain (Fig. 9b). A detailed study of this situation (F. Duffourg 2009, personal communication) revealed that the MCS was driven by very finescale mechanisms such as the forming of a low-level cold pool and the triggering of deep convection by small orography features.

At 0600 UTC 20 October 2008, some convective cells formed over the Mediterranean Sea, then intensified, formed a convective line and moved northeastward. At 1300 UTC, a fully developed MCS had become quasi-stationary over the Gard region. The MCS then moved...
southward to reach the coast at 1700 UTC and dissipated by the end of the day. The accumulated observed precipitation (Fig. 9c) showed that the MCS produced more than 150 mm in 6 h.

b. Results

1) CASE 1

Figures 10 and 11 show the 24-h accumulated precipitation for all members of AROME–PEARP1 and AROME–PERTOBS experiments, respectively. There are obvious differences between the AROME–PEARP1 ensemble members, in location as well as in intensity. A local maximum of only 177 mm in 24 h is simulated by member 5, whereas 331 mm is obtained by member 2 (Figs. 10c,f). Some of the members reproduce the rainfall amounts quite well. The AROME–PERTOBS ensemble members show less variability between members, with respect to both the location of precipitation and the maximum of accumulated precipitation. The maximum ranges from 184 mm for member 9 to 232 mm for member 4 (Figs. 11e,j).

Time series of 3-h accumulated precipitation averaged over the southeast of France are shown in Fig. 12 for the observations and the forecast of each member. The AROME–PEARP1 ensemble (Fig. 12a) shows good variability, and several members capture the precipitation peak with good timing and intensity.

For this case, neither the AROME–PEARP2 (not shown) nor the AROME–PEARP3 (Fig. 12b) experiments performed as well as the AROME–PEARP1 ensemble, with an underestimation of precipitation between 6 and 9 h. This case is characterized by a strong upper-level synoptic-scale circulation and large uncertainties on the global forecast. In this context, the spread of the ensembles grows too rapidly for much to be gained by removing the data assimilation in ALADIN or using an earlier PEARP run to provide LBCs. In the meantime,
the errors become much larger in the AROME–PEARP2 and AROME–PEARP3 ensembles compared to the AROME–PEARP1. Thus, for this case, using the latest PEARP run with ALADIN data assimilation yields improvements in the high-resolution forecasts.

The AROME–PERTOBS ensemble (Fig. 12c) performs well during the first 12 h, with precipitation of the different members well distributed around the observations. However, none of the ensemble members are able to simulate the precipitation peak, and, over time, all the members tend to produce very similar precipitation forecasts. After some time (around 9 h for this case) the synoptic-scale conditions that are used as LBCs tend to propagate within the AROME domain and overcome the differences in the initial states, resulting in a decrease in forecast spread.

In the case of a strong synoptic-scale forcing with large uncertainties on the global forecast, the impact of LBCs is predominant. The uncertainty associated with high-resolution forecasts cannot be accounted for by the uncertainty on ICs only.

2) CASE 2

Considering the timing of this event, the 24-h AROME forecasts were run at 0000 UTC instead of 1200 UTC, to encompass the whole event from the triggering to the decay stage.

As was expected for case 2, because of weaker synoptic-scale forcing, the AROME–PERTOBS ensemble performs better than it does for case 1, with a greater variability between members, whereas the AROME–PEARP1 ensemble is not as convincing as for case 1.

Time series of precipitation show that the AROME–PEARP1 ensemble (Fig. 13a) fails to capture the precipitating event. This ensemble produces from 43 mm to 69 mm in 24 h when 164 mm was observed. This behavior is very similar to the AROME deterministic forecast.

For this case, both the AROME–PEARP2 and AROME–PEARP3 ensembles perform much better than AROME–PEARP1. They produce similar time series of precipitation, and some of their members capture the precipitation peak quite well. Time series of precipitation for the AROME–PEARP3 ensemble are shown in Fig. 13b. In this respect, the AROME–PEARP2 ensemble is slightly worse than AROME–PEARP3 (not shown), with fewer members capturing the precipitation peak and a tendency to simulate too much precipitation during the last hours of the forecast. However, the location of precipitation is better represented in some of the AROME–PEARP2 ensemble members.

The benefit of removing the data assimilation in the downscaling procedure is clear for this case. The impact of
Fig. 10. (a)–(l) The 24-h accumulated precipitation (mm), at 1200 UTC 2 Nov 2008, for the 11 AROME–PEARP1 ensemble members. Rain gauge observations are shown in (l).
FIG. 11. As in Fig. 10, but for the AROME–PERTOBS ensemble.
FIG. 12. The 3-h accumulated precipitation, averaged over the domain used for Fig. 10 display (mm): (a) AROME–PEARP1 ensemble, (b) AROME–PEARP3 ensemble, and (c) AROME–PERTOBS ensemble from 1200 UTC 1 Nov 2008 to 1200 UTC 2 Nov 2008. The solid black line is for rain gauge observations, dashed lines and boxes are for the ensemble forecast. The box-and-whisker plot shows the median, lower, and upper quartiles, and the small dots stand for outliers.
Fig. 13. As in Fig. 12, but for case 2 from 0000 UTC 20 Oct 2008 to 0000 UTC 21 Oct 2008.
using an earlier PEARP run is more uncertain, yielding better results on some aspects but not all.

The AROME–PERTOBS ensemble (Fig. 13c), in contrast, has some members that capture the precipitation peak fairly well. Although the time series of precipitation show an underestimation in precipitation forecasts, the maximum precipitation produced by the ensemble members ranges from 62 to 281 mm, to be compared with the observed 164 mm. The benefit of creating this ensemble with finescale perturbations in the ICs was clear for this case, since the P0 run completely missed this event. The precipitating systems simulated by all members are very different. Some members show high rainfall rates produced by very intense convective cells, while other members show larger and weaker systems. This reveals the lower predictability of this system compared to case 1. Case 2 mainly involved finescale and convective processes known to disrupt predictability.

The comparison between case 1 and case 2 indicates that the synoptic-scale conditions, and mainly the strength of the circulation, have an important effect on the behavior of both ensembles. A weaker synoptic-scale circulation leads to an enhanced impact of the differences introduced in the ICs. The initial convective-scale perturbations can last for a longer time.

3) Geographical Analysis of Ensemble Spread

To complete this study, the spatial distribution of ensemble spread was examined. Figure 14 shows, for case 2, the 925-hPa wind speed ensemble spread after 18 h for AROME–PERTOBS and AROME–PEARP2. As already stated, the AROME–PERTOBS ensemble spread diminishes with time. After 18 h, the remaining spread is located exactly on the active regions (MCSs, fronts, etc.), and no information on the predictability remains in the vicinity. On the contrary, the AROME–PEARP2 ensemble spread grows with time, and highlights a more extended area around the active regions. For instance, AROME–PEARP2 samples more uncertainty on the frontal activity entering the domain by the northwest corner. The impact of synoptic-scale conditions and uncertainty is again highlighted here. Both ensembles show a similar spread around the MCS, showing that the uncertainty on this system comes from both the uncertainties on ICs and LBCs. However, the AROME–PEARP2 ensemble spread suggests a larger uncertainty on the low-level jet upstream the HPE.

5. Conclusions

This study assessed the impact of uncertainty on LBCs and convective-scale ICs in cloud-resolving simulations, with a focus on Mediterranean heavy precipitating events (HPEs). Four distinct high-resolution ensembles were created with the AROME cloud resolving model to separately evaluate the uncertainty associated with convective-scale ICs on the one hand and synoptic-scale LBCs on the other.

The impact of uncertainty on synoptic-scale conditions was assessed through the three AROME–PEARPs ensembles, the members of which were driven by the PEARP global, short-range ensemble members. To quantify the uncertainty on convective-scale ICs, an ensemble data assimilation technique was used to create the AROME–PERTOBS ensemble. For that purpose, the observations were randomly perturbed before the assimilation by the AROME 3D-VAR assimilation system, thus generating different ICs for the ensemble members.

The probabilistic evaluation of these two ensembles over a 31-day period gave very promising results for precipitation forecasts. Although they considered the uncertainty sources separately, they already provided valuable probabilistic information compared to the deterministic forecast; they gave comparable scores. This may be because of the accumulation of precipitation over 24 h (there are more differences over the last 12 h). For low-precipitation thresholds, very large areas are concerned, which also contributes to good but similar scores. At a 10-m height, forecasts were found to be very similar, probably because of the importance of surface conditions. Scores computed at 925 hPa show larger differences between the ensembles.

Overall, the AROME–PEARP1 ensemble had a little more resolution whereas the AROME–PERTOBS ensemble was a little better in terms of reliability. For high-precipitation and wind thresholds, the AROME–PEARP1 ensemble slightly outperformed the AROME–PERTOBS one.

However, both ensembles were shown to be strongly underdispersive for low-level fields. The AROME–PEARP1 ensemble spread grew over time but the ensemble was still strongly underdispersive after 24 h. The AROME–PERTOBS ensemble had a better rank histogram during the first hours, but the spread diminished over time due to the use of single LBCs.

The impact of uncertainty on LBCs has been further studied through the AROME–PEARP2 and AROME–PEARP3 ensembles. It was shown that removing the data assimilation from the downscaling procedure of the large-scale ensemble members improved the rank histograms. Indeed, this data assimilation step moves the large-scale ensemble members toward the same set of observations, thereby reducing spread. Using an earlier PEARP run to provide coupling conditions, so that the initial perturbations of the large-scale forecast had more
Fig. 14. The 925-hPa wind speed spread (m s$^{-1}$) at 1800 UTC 21 Oct 2008 (forecasts issued at 0000 UTC 20 Oct 2008 for an 18-h forecast range) for (a) AROME–PERTOBS and (b) AROME–PEARP2.
time to grow, further improved the ensemble spread, mostly at intermediate forecast ranges (around 12 h).

To provide the best sampling of the uncertainty on LBCs, it is necessary to retain the maximum information about that uncertainty in the downscaling procedure. Thus, a better solution would be to use a higher-resolution global EPS so as not to need an intermediate downscaling model in the ensemble configuration. Since some information is lost through the temporal interpolation of LBCs, a higher LBCs update frequency would also improve the ensemble. Another solution would be to have an intermediate downscaling ensemble taking into account all the uncertainty sources instead of the LBCs only, but this involves more research on this aspect and requires longer computing time.

Both the probabilistic evaluation and the results for the two Mediterranean HPEs concur to show that the two uncertainty sources have different impacts on the cloud-resolving simulations. The uncertainty on convective-scale ICs has a stronger impact over the first hours (<12 h) of simulation, before the LBCs overwhelm differences in initial states. The uncertainties on LBCs have a growing impact at a longer range (beyond 12 h).

The results for the two cases of HPE also reveal that the impact of both uncertainty sources depends strongly on the large-scale meteorological situation. When the synoptic-scale dynamics are predominant in driving the HPE (e.g., when a trough interacts with the larger-scale baroclinic westerly flow), the impact of uncertainty on LBCs is by far more important than the impact of uncertainty on convective-scale ICs. When the synoptic-scale circulation is weaker and the HPE is mainly driven by local and mesoscale processes, convective-scale initial perturbations have more impact on the simulated precipitating systems.

The design of a future convective-scale EPS thus has to combine both uncertainty sources. It would be possible to apply the ensemble data assimilation technique to each member of the large-scale EPS, but this approach is far too expensive in computing time since it involves the computation of 11 × 11 members in our configuration. A solution less demanding in computer time is to apply one perturbation of ICs for each PEARP member (i.e., to build an ensemble similar to AROME–PEARP2, but using perturbed observations for the AROME convective-scale data assimilation cycles members for 1–10). This approach is currently under investigation.

Several scientific issues remain to be addressed in future work. The ensemble data assimilation strategy could be refined (i.e., more relevant perturbations of observations, dealing with areas where fewer observations are available, etc.). Since global EPSs have a large number of members, a selection will have to be made of a few members to drive a cloud-resolving ensemble at a reasonable computing cost. This paper has focused on the ICs and LBCs, but model errors remain to be investigated (e.g., physical parameterizations, surface conditions, etc.). Results for these aspects will be reported in forthcoming papers.

Acknowledgments. The authors thank Dr. L. Descamps for his useful insight on ensemble evaluation. The authors also gratefully acknowledge the two anonymous reviewers for their relevant comments and suggestions. This work was carried out in the framework of the MEDUP project funded by ANR “Vulnérabilité Milieux Climat.”

REFERENCES


