Tellus
SERIES A
DYNAMIC
METEOROLOGY
AND OCEANOGRAPHY
PUBLISHED BY THE INTERNATIONAL METEOROLOGICAL INSTITUTE IN STOCKHOLM

Tellus (2011), 63A, 550–563

TELLUS

Predictability of short-range forecasting: a multimodel approach

By JOSE-ANTONIO GARCÍA-MOYA[1], ALFONS CALLADO[2], PAU ESCRIBÀ[2], CARLOS SANTOS[1], DANIEL SANTOS-MUÑOZ[1] and JUAN SIMARRO[3]* [1]*AEMET, C/Leonardo Prieto Castro 8, Ciudad Universitaria, 28071 Madrid, Spain;* [2]*AEMET, Delegación Territorial en Cataluña, C/Arquitecto Sert 1, 08071 Barcelona, Spain;* [3]*AEMET, Delegación Territorial en la Comunidad Valenciana, C/Botánico Cavanilles 3, 46010 Valencia, Spain*

(Manuscript received 11 April 2010; in final form 19 November 2010)

ABSTRACT

Numerical weather prediction (NWP) models (including mesoscale) have limitations when it comes to dealing with severe weather events because extreme weather is highly unpredictable, even in the short range. A probabilistic forecast based on an ensemble of slightly different model runs may help to address this issue. Among other ensemble techniques, Multimodel ensemble prediction systems (EPSs) are proving to be useful for adding probabilistic value to mesoscale deterministic models. A Multimodel Short Range Ensemble Prediction System (SREPS) focused on forecasting the weather up to 72 h has been developed at the Spanish Meteorological Service (AEMET). The system uses five different limited area models (LAMs), namely HIRLAM (HIRLAM Consortium), HRM (DWD), the UM (UKMO), MM5 (PSU/NCAR) and COSMO (COSMO Consortium). These models run with initial and boundary conditions provided by five different global deterministic models, namely IFS (ECMWF), UM (UKMO), GME (DWD), GFS (NCEP) and CMC (MSC). AEMET-SREPS (AE) validation on the large-scale flow, using ECMWF analysis, shows a consistent and slightly underdispersive system. For surface parameters, the system shows high skill forecasting binary events. 24-h precipitation probabilistic forecasts are verified using an up-scaling grid of observations from European high-resolution precipitation networks, and compared with ECMWF-EPS (EC).

1. Introduction

From the dynamical point of view the atmosphere is a chaotic non-linear system. This fact implies that, even if a quasi-perfect numerical weather prediction (NWP) model were initialized with quasi-perfect initial conditions, the forecast would cease to be valid within a finite time interval (Lorenz, 1963). The lack of atmospheric predictability is due to the non-linear amplification, as the forecast period lengthens, of small errors in both the initial conditions and in the NWP models formulation. This intrinsic deficiency in the atmospheric predictability can be found at a wide range of time and space scales, including the mesoscale. A single NWP model being initialized with a single initial condition only provides one forecast of the future atmospheric state, and it has been largely proved that generating several predictions based on slightly different initial conditions and model configurations can improve the forecast (e.g. Hou et al., 2001). The improvement comes with a probabilistic representation of the atmospheric forecasts, which in turn comes usually from an equally likely set of deterministic forecasts or ensemble of forecasts.

A variety of approaches are used to generate an ensemble prediction system (EPS) from deterministic numerical models, most of them sample both initial state and model uncertainties. One of the first proposed techniques to sample initial state uncertainty consists in the use of Monte Carlo methods to construct multiple random initial conditions for feeding models. This technique was proposed by Leith (1974), Hollingsworth (1980) and Mullen and Baumhefner (1989) among others. Hoffman and Kalnay (1983) proposed a time-lagged averaged forecast using forecasts from lagged starting times as members, as an alternative technique, which has led to some expertise (Ebisuzaki and Kalnay, 1991). More recent approaches are based on generating dynamically constrained perturbations. Bred vectors (e.g. Toth and Kalnay, 1993, 1997) and singular vectors (e.g. Buizza and Palmer, 1995, 1997; Hamill et al., 2000) are two of the main methods of introducing perturbations into the subspace of fastest growing errors. Houtekamer et al. (1996) developed the idea of obtaining a better

*Corresponding author.
e-mail: jsimarrog@aemet.es
DOI: 10.1111/j.1600-0870.2010.00506.x

representation of initial conditions by using a set of assimilation cycles and perturbing the observations with random errors in each of them. The Ensemble Transform Kalman Filter (ETKF) technique (Bishop et al., 2001; Wang and Bishop, 2003) provides a framework to assimilate observations and estimate the effect of observations on forecast error covariance. This technique scales the ensemble perturbations according to the observation errors and the computation is done at much lower cost compared to, for example, the singular vectors.

NWP model errors are another main source of weather forecast uncertainty. Generating EPS by altering the model parameterization of subgrid-scale physical processes (e.g. Houtekamer et al., 1996; Andersson et al., 1998; Stensrud et al., 1998) or using stochastic physics methods (Buizza et al., 1999) may help to take into account model errors. The multimodel technique is another way of sampling model uncertainties using different numerical models to generate an EPS. The SAMEX (Hou et al., 2001), UKMO Test of Poor Man's EPS (Arribas et al., 2005) and DEMETER (Palmer et al., 2004) projects have shown that this technique can be useful for both short-range and seasonal forecasting. Boundary conditions, an additional source of uncertainty, must be considered when using limited area models (LAMs). When an EPS is built using a LAM, both lateral boundary conditions (LBCs) and initial conditions give their contribution to the spread and skill of the system (Clark et al., 2008).

Medium-range forecast uncertainties in mid-latitudes are qualitatively related to baroclinic instability (Buizza and Palmer, 1995) whereas short-range forecast uncertainties have more physical processes involved and therefore are more difficult to characterize. Uncertainties can indeed grow critically in the short range at a wide range of spatial scales due to different kind of atmospheric instabilities such as baroclinic, inertial and potential instabilities (Emanuel, 1979; Roebber and Reuter, 2002; Zhang, 2005; Hohenegger and Schär, 2007). Though the first arising EPS were global medium-long-range systems, short-range EPS have shown to be of potential use in early warnings and severe weather events of quick growth. For instance, floods (Bright and Mullen, 2002; Hacker et al., 2003) and wind gales (Leslie and Speer, 1998) or tornadoes (Stensrud and Weiss 2002).

National Centers for Environmental Prediction (NCEP) was the first operational meteorological centre in developing a short-range ensemble, using the multimodel approach with two LAMs and boundary conditions from the NCEP medium-range global ensemble (Tracton et al., 1998; Stensrud et al., 1999). Bred vectors (Toth and Kalnay, 1993) are used to take into account the uncertainty in the initial and boundary conditions, whereas model uncertainty is addressed by using different parameterization schemes.

The United Kingdom Meteorological Office (UKMO) has developed the MOGREPS short-range ensemble (Bowler et al., 2008), which consists of global and regional ensembles with the global ensemble providing initial and boundary conditions to the regional ensemble. The ETKF technique (Bishop et al., 2001) is applied to construct perturbations for the initial conditions (Bowler and Mylne, 2009; Bowler et al., 2009). Model error is represented by applying stochastic perturbations to the model, mainly to the parameterized model physics.

The COSMO-LEPS ensemble (Marsigli et al., 2004, 2008) uses the ECMWF-EPS (EC hereafter) (Molteni et al., 1996; Palmer et al., 1997) to provide a set of different initializations for the high-resolution COSMO model. The 51 EPS members are divided in 16 clusters and one member of each cluster is selected to generate initial conditions for the COSMO model. Model uncertainty is sampled by the use of different convective parameterization schemes.

LAMEPS is a short-range ensemble developed at the Norwegian Meteorological Institute (Frogner et al., 2006). It is based on the use of the High-Resolution Limited Area Model (HIRLAM) with different initial and boundary conditions with a grid spacing of 12 km. The initial and boundary conditions are constructed by the use of a 21-member version of the EC ensemble in which the singular vectors are optimized to maximize the 48 h growth over northwestern Europe (TEPS, Frogner and Iversen, 2001). In the LAMEPS perturbations of the TEPS are added to the high-resolution regional analysis. The whole system is called NORLAMEPS. Model uncertainty is included in two ways: alternating different physical parameterizations and combining LAMEPS and TEPS.

Current areas of research on short-range EPS include the use of high-resolution non-hydrostatic models to describe forecast uncertainty up to a convection-permitting scale, that is 2 or 3 km of horizontal resolution (Wandishin et al., 2008, Wandishin et al., 2010).

This is a non-comprehensive list of short-range forecasting systems, mainly in Europe, and it shows that this is an active area of development, in which different ensemble techniques are used to provide probabilistic insight in the short-range weather forecast.

This paper sets out the results of the Short-Range Ensemble Prediction System (SREPS) developed at the Spanish Meteorological Service (AEMET: Agencia Estatal de Meteorología) using the multimodel multiboundary technique. The system is built using a set of LAMs and a set of deterministic global models supplying the initial and boundary conditions. In this way, both initial conditions and model errors are represented in the EPS. The system is focused on short-range forecast (up to 72 h) and has been developed to help in the forecast of extreme weather events (gales, heavy precipitation and snow storms). For instance, the Spanish Mediterranean region has a meteorological behaviour dominated by the interaction of synoptic flow with small-scale orographic features and the Mediterranean Sea. Such interaction produces mesoscale structures that are difficult to model using global models or even LAM. The relevance of most of these events is related to the value that meteorological

parameters take at the surface, therefore different verification exercises have been performed to validate the system in this context.

AEMET-SREPS (AE hereafter) project began in 2005. In 2006 the system was capable to run every day at 00 UTC the first set of models with different boundary conditions. During 2007 the ensemble reached a mature stage, running 20 members with good performance at a horizontal resolution of $0.25^\circ \times 0.25^\circ$ and 40 vertical levels. During 2008 it was added a second daily run at 12 UTC. In October 2009 it was included the CMC global model, increasing the number of ensemble members up to 25. Currently (July 2010) the system is running daily (00 and 12 UTC) with a forecast range T+0 to T+72 h at 0.25° and 25 members. Verifications shown here cover 2007–2008 and thus correspond to SREPS 00UTC run at 0.25° with 20 members.

In the second section we present the methodology used, and in the subsequent sections we set out the results of the objective verification exercises. In Sections 3 and 4 we introduce the verification strategy and methods followed, showing and discussing results of several measures for different aspects of performance. Conclusions and further results are presented and discussed in Section 5 together with suggestions about on-going development.

2. Methodology

Following the results of the Storm and Mesoscale Ensemble Experiment (SAMEX) performed in the United States in 1998 (Hou et al., 2001) AEMET chose the multimodel multiboundary technique to build a short-range ensemble prediction system, called SREPS, to provide probabilistic forecasts for a wide range of parameters and thresholds. These probabilistic forecasts give complementary information to the operational deterministic NWP model at AEMET and may improve the prediction of severe weather events such as gales, heavy precipitation, snowstorms or heat waves. Severe weather warnings are one of the most important tasks of a modern weather service because of the damage that such events cause.

2.1. Multimodel technique

Operating the AE consists of running five LAMs each of them using initial and boundary fields from five different global deterministic models. The SREPS has 25 members as a result of this combination. The system runs twice a day, at 00 and 12 UTC, with a forecast range of up to 72 h. LAMs are configured to have a horizontal resolution of about 25 km and 40 vertical levels. As not all LAMs use the same map projection (HIRLAM, COSMO, High-Resolution Regional Model [HRM] and UM use rotated latitude-longitude, and MM5 uses Lambert conformal) it is not possible to cover the same integration area at the same horizontal resolution. This is why it was decided to use a large integration area and a smaller common post-process area for the purpose of computing the probabilistic forecasts.

The LAM models used are High-Resolution Limited Area Model (HIRLAM, McDonald and Haugen, 1992; Undén et al., 2002), High-Resolution Regional Model (HRM, Majewski, 1991; Majewski and Schrodin, 1994) from Deutsche Wetterdienst (DWD), Mesoscale Model version 5 from Penn State University and NCAR (MM5, Dudhia, 1993; Grell et al., 1994); Unified Model (UM, Cullen, 1993) from UKMO and COSMO Model (LM, Doms and Schättler, 1997) from the COSMO Consortium. Each model has its own numerical features and has different physics parameterization schemes.

When trying to interface the LAMs with the different global models, the standard pre-process tools available for each LAMs have been used when possible. However, in several cases it was unavoidable to perform an additional vertical interpolation and do some modifications in the codification of GRIded Binary (GRIB) files provided by global models (e.g. change the global model fields from isobaric levels to hybrid levels based on surface pressure, before the standard pre-processing tool is used), thus making necessary the development of additional software. In addition to the technical difficulties, there can be some meteorological impact in those members where two vertical interpolations are done on the global fields instead of one. However, the interpolation method used has been chosen to minimize this potential loss of information. Then, it is also necessary in some cases to calculate new variables from the available set of fields provided by the global model (a typical case is finding the relative humidity from the specific humidity, pressure and temperature). Finally, the standard pre-process tools of each LAM model is applied to produce the initial and boundary conditions, including the necessary horizontal interpolations and adjustments of the fields to the LAM topography.

2.2. Multiboundary and initial conditions

Initial and boundary conditions for LAMs are taken from the forecasts of five different global deterministic models. These global forecasts are initialized 12 h before the start time of the SREPS integration. So, if the global model integration is based on the global HH UTC analysis, the SREPS members are initialized with the meteorological fields of the HH + 12 UTC global model forecast. This strategy may decrease thedeterministic performance of each member because the most recent observations are not used. However, the performance of the SREPS as a tool for probabilistic forecasting is not very affected and the fact that its cycle finishes earlier makes it more useful for the forecasters.

The initial and boundary conditions are obtained from the following: well-known global deterministic models: Integrated Forecast System (IFS, Simmons et al., 1989; Jakob et al., 1999) from European Centre of Medium Range Weather Forecasts (ECMWF); Global Unified Model (UM, Cullen, 1993) from UKMO; Global Forecast System (GFS, Sela, 1980, 1982) from

NCEP; Global Model (GME, Majewski et al., 2002) from DWD and Global Canadian Model (CMC, Côté et al., 1998a,b). All of these include Variational Data Assimilation Schemes (IFS, UM and CMC use 4DVAR and GFS and GME use 3DVAR) and up-to-date physical parameterization schemes. Using these global models as initial and boundary conditions ensures a good initialization of the system.

2.3. Post-process

One of the features of an EPS forecast is the huge amount of information that the system can supply. For this reason, when working in an operational environment the use of post-process methods becomes necessary for reducing this information down to a comprehensive set that can be better used by a human forecaster. The first step in the SREPS post-processing consists in the interpolation of the integration area of each model into a common area of 0.25° horizontal resolution in latitude-longitude. The common area is shown in Fig. 1 and covers most of North-Atlantic Ocean, Europe, Northern Africa and the Mediterranean Sea. Probabilistic forecasts of meteorological parameters are computed from SREPS assuming that each member is equally likely, as it is in other ensemble systems. Most of these products are routinely verified to assess the performance of SREPS.

3. Large-scale flow consistency

To assess the overall consistency of the system on the large scale, it has been carried out a verification of some synoptic fields against ECMWF operational analysis. Different results are shown in Sections 3.1 and 3.3 for mean sea level pressure (MSLP) and in Section 3.2 for geopotential at 500 hPa (Z500). Using objective analysis as reference has the advantage that performance measures cover the whole integration domain (shown in Fig. 1) with the same weight and then no priority is given to land areas where the density of observations is higher. Verification against SYNOP/TEMP observations (not shown here) gives, as expected, worse but qualitatively similar results.

A 21-month verification period from April 2007 to December 2008 has been chosen to fit with the same verification period given in Section 4 for weather parameters (see 4.1 for details). A number of 614 daily forecasts at 00UTC has been selected from out of the whole April 2007 to December 2008 period, because due to usual operational problems on the development stage there is a non-negligible number of days without forecasts. The number of grid points on the domain is $380 \times 164 = 62\,320$. Thus, the total number of realizations for each individual score at any forecast length is $62\,320 \times 614 \sim 3.8\text{e}7$. The forecast range shown covers T+6 to T+72 every 6 h, which means 12 forecast lengths. Different aspects of large-scale flow consistency have been considered: individual member performance (shown in 3.1

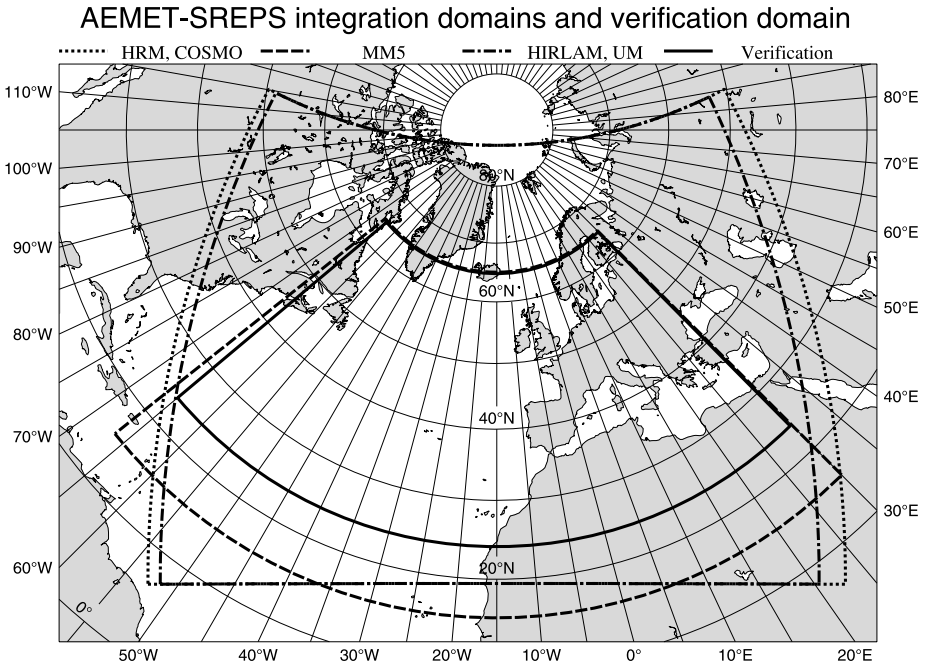

Fig. 1. AE LAM models integration domains: HRM, COSMO (dotted line), MM5 (dashed line), HIRLAM and UM (dash-dot line). The verification domain (solid line) was selected as the maximum possible area defined by a regular 0.25° lat-lon grid, and it covers part of the North Atlantic Ocean, Northern Africa and a big part of Europe.

for MSLP), statistical consistency with the analysis (in 3.2 for Z500) and ensemble spread (in 3.3 for MSLP).

3.1. Individual member performance

Compared with a single-model ensemble, some extra issues arise when a multimodel EPS is being considered. The task of interfacing several LAM models to different global models in an ensemble is intricate and expensive from the technical point of view, and the individual member deterministic quality helps to monitor each model implementation. Moreover, from the validation point of view, it is first necessary to assess this individual member deterministic performance, because every member can be weighted equally in the computation of probabilistic forecasts when they perform similarly. It is also expected that the root mean square error (RMSE) of the ensemble mean is smaller than that of any member (Leith, 1974; Murphy, 1988; Whitaker and Loughe, 1998; Ziehmann, 2000). To measure this, evolutions with forecast length of synoptic variables bias and RMSE (Z500, T500, MSLP) have been computed for each member and also for the ensemble mean. Results are shown for MSLP in Fig. 2. As far as the performance of each individual model is not the main purpose here, there are no labels in Fig. 2 to distinguish between the different members; instead, the ensemble mean is highlighted, as well as those LAMs who are driven with their "normal operating" global model, that is, MM5 and GFS, HIRLAM and ECMWF, COSMO and GME, UM and global UM, HRM and GME (this last aspect has been shown due to its special interest). These results indicate similar performance of the members, with no clear improved quality for the 'normal operating' members, and the ensemble mean shows a lower RMSE than any of the rest. It can therefore be concluded that, as expected, the forecast quality of every member is similar and that the ensemble mean outperforms any member.

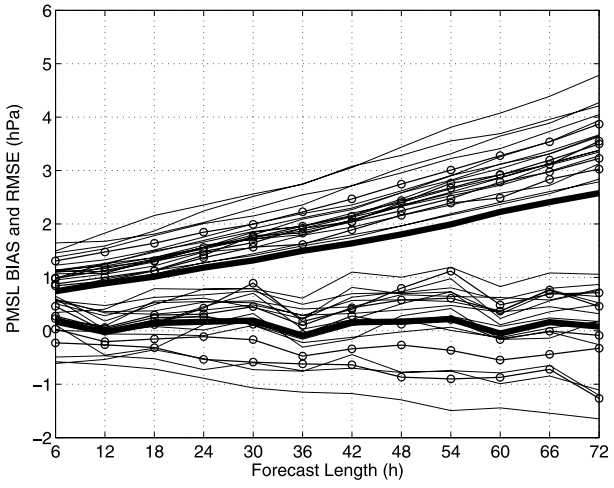

Fig. 2. Evolutions with forecast length of mean sea level pressure (MSLP) bias and RMSE computed for each member (thin lines) and for the ensemble mean (thick line). 'Normal operating' members are highlighted (circles).

3.2. Statistical consistency with the analysis

The first step in the validation of an EPS, as a probabilistic prediction system, is to check its statistical consistency with observations (analysis) in the large-scale flow. The rank histogram (Anderson, 1996; Hamill and Colucci, 1997, 1998; Hamill, 2001; Candille and Talagrand, 2005) can be used to check if the verifying observation is statistically indistinguishable from the set of forecast values (or if any ensemble member, as well as the verifying observation, can be considered equally likely to be the truth), and thus whether the system is statistically consistent with the observations ('reliable' in this context). Such a system must show an approximately flat-shaped rank histogram. The rank histogram corresponding to the mentioned period April 2007 to December 2008 for Z500 at forecast length T+24 (Fig. 3) shows overall consistency and some outliers indicating a possible slight subdispersion typical of current EPS systems. Averaging over 21 months could hide seasonal variability. In Fig. 4 rank histograms corresponding to different seasons are spanned to show seasonal differences: clear subdispersion in winter ('U' shape), flat in spring, clear overdispersion in summer (inverted 'U' shape) and some subdispersion in autumn. Rank histograms for other variables and forecast lengths (not shown) show similar shapes.

3.3. Ensemble spread

The consistency with observations is also related to the ensemble spread. An EPS is expected to sample the uncertainties of NWP models (ensemble spread), as well as to give an explicit and quantitative information about the predictability of the atmosphere (represented by the ensemble mean error compared with the observation). A consistent EPS is expected to show a sort of linear relationship between these two magnitudes: the

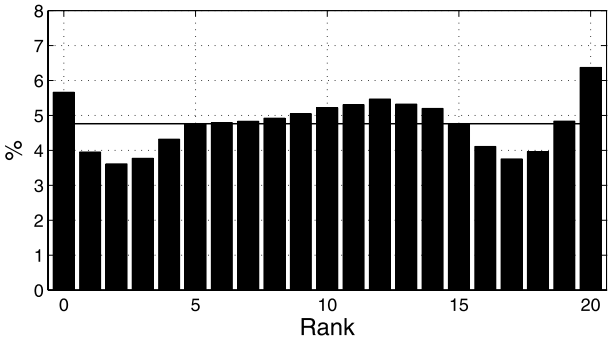

Fig. 3. Rank histogram corresponding to the period April 2007 to December 2008 for 500 hPa geopotential height and forecast length T+24. The analysis of the ECMWF model is taken as observation value.

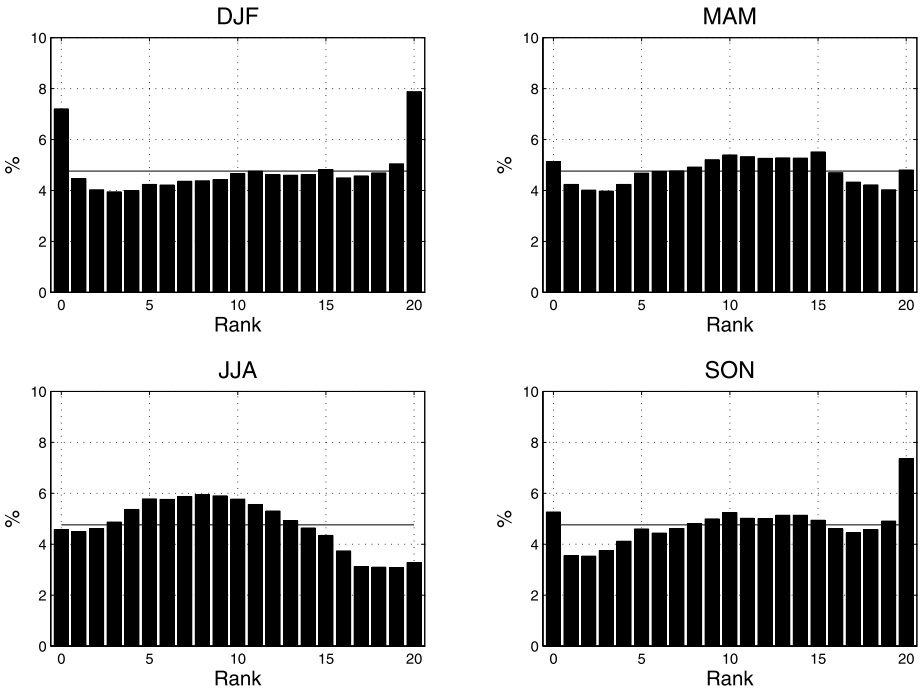

Fig. 4. Rank histograms corresponding to different seasons: December 2007 to February 2008 (DJF), March to May 2008 (MAM), June to August 2008 (JJA) and September to November 2008 (SON).

ensemble spread (as far as there is not a control forecast in the system, the spread is measured here by the standard deviation of the ensemble values with respect to the ensemble mean) and the RMSE of the ensemble mean with respect to the analysis (Buizza and Palmer, 1997; Whitaker and Loughe, 1998).

As far as an AE is composed of a set of LAMs ('multimodel') and global models as initial and boundary conditions ('multiboundaries'), an interesting issue arises: how much of the ensemble spread comes from the different LAMs and how much from the global ones? To try to answer this question, two kind of subensembles have been considered: those composed of one LAM and four global (five LAMs give five possible combinations for strictly 'multiboundaries' subensembles) and those composed of five LAMs and a single global (four globals give four possible combinations for strictly 'multimodel' subensembles). Then, the formulation of the problem can be given as follows: how is the whole system spread compared with that of any of these subensembles? The verification of 'multiboundaries' and 'multimodel' subensembles gives insight to this question. The impact of difference in ensemble size (20 members for the whole system, 4 or 5 members for the different subensembles) has not been addressed in this study though it could add useful information (see Buizza and Palmer 1998; Ferro, 2007; Ferro et al., 2008).

To summarize results, Fig. 5 shows MSLP spread-error diagrams from T+6 to T+72 every 6 h for the whole 'multimodel multiboundaries' system (20 members) and for the five different 'multiboundaries' subensembles. Figure 6 shows the same

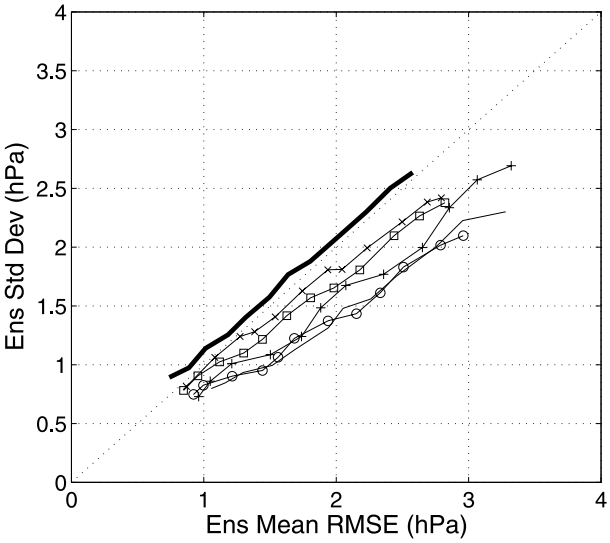

Fig. 5. Mean sea level pressure (MSLP) spread-error diagrams from T+6 to T+72 every 6 h, for the whole 'multimodel multiboundaries' system (thick solid line, 20 members) and for the five different 'multiboundaries' subensembles (thin lines).

information for the whole 'multimodel multiboundaries' system and the four different 'multimodel' subensembles.

As a first result, the system as a whole shows a clearly linear relationship between ensemble spread and RMSE of the ensemble mean, that is, the system is consistent with the observations

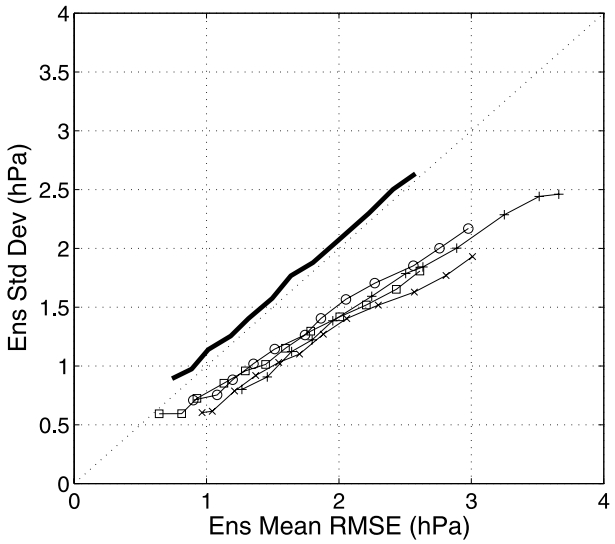

Fig. 6. Mean sea level pressure (MSLP) spread-error diagrams from T+6 to T+72 every 6 h, for the whole 'multimodel multiboundaries' system (thick solid line, 20 members) and for the four different 'multimodel' subensembles (thin lines).

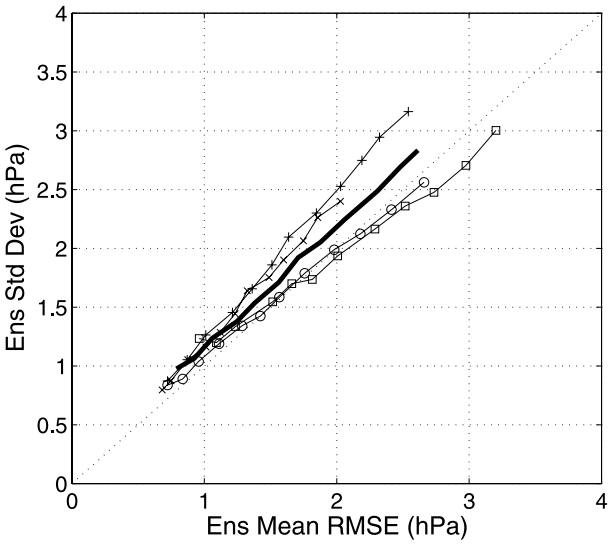

Fig. 7. Mean sea level pressure (MSLP) spread-error diagrams from T+6 to T+72 every 6 h. Each thin line corresponds to a period of 3 months: December 2007 to February 2008 (square), March to May 2008 (cross), June to August 2008 (plus sign) and September to November 2008 (circles). The thick line corresponds to the spread-error diagram for the whole period, December 2007 to November 2008.

(analysis). A second result is that all subensembles show lower spread than the full system. This conclusion has been reached in other ensembles, see for instance the SAMEX project (Hou et al., 2001), UKMO Test of Poor Man's EPS (Arribas et al., 2005) and DEMETER project (Palmer et al., 2004), this last for seasonal prediction ensembles. Finally, as a third result, 'multimodel' subensembles happen to be, as moving to T+72, more underdispersive than the 'multiboundaries' ones. The contribution of the different global models to the spread of the whole system is larger than the contribution of the LAM models. This is a confirmation that the perturbation of the boundaries is an important requirement for LAM ensemble systems like AE. This result needs further research and a comprehensive explanation is beyond the scope of this paper; as an outline the seasonal behaviour is briefly described later. Similar results have been found for other parameters (not shown). It can therefore be concluded that the whole system performs significantly better than any of its subensembles.

To include seasonal behaviour that could be hidden on the overall average, Fig. 7 shows MSLP spread and ensemble mean RMSE (EMRMSE) only for the whole system (20 members) for winter (2007D-2008JF), spring (2008MAM), summer (2008JJA), autumn (2008SON) and the whole 12 months (2007D-2008N) in separate curves. These results give an insight to seasonal differences that are more or less consistent with that for Z500 at T+24 (Section 3.2) but give further details. In winter, the ensemble gets underdispersive as the forecast length grows; in spring the system clearly shows some overdispersion that grows slightly with forecast length (with Z500 T+24 it was not clearly overdispersive) where in T+72 large values of spread and EMRMSE are reached; in summer there's some overdispersion but both spread and EMRMSE reach smaller values than that for spring (this is consistent with the larger predictability expected in summer in contrast with the lower predictability and corresponding larger spread expected in spring); finally, the system shows in autumn the best spread–error relationship, fairly close to the diagonal and reaching values of spread and EMRMSE smaller than that for winter–spring and larger than that for summer. These different seasonal behaviours compensate each other to produce a more balanced relationship spread EMRMSE, a bit overdispersive, in the overall average.

4. Verification of precipitation forecasts using HR networks

A short-range EPS is mainly focused on surface parameters (precipitation, 2 m temperature, 10 m wind), becoming nowadays an important tool for the short-range forecast guidance. Thus, the ensemble response to binary events related to these parameters, over a selected set of thresholds in this operational forecast context, must be assessed. Even though EC is a medium-range forecast system, it has been selected as a high-performance available reference ensemble forecast to compare with AE, selecting some feasible forecast lengths for the comparison.

Detailed results for the 24 h precipitation forecast verification for both AE and EC ensembles are shown here, whereas for 2 m temperature and 10 m wind only a brief summary is given.

4.1. Data sets

Observed precipitation data from High Resolution (HR) networks over Europe have been used for verification. The period chosen covers almost 2 yr, from April 2007 (here the ensemble 00UTC run reaches a mature stage) to December 2008 (observations for 2009 were not still available for this study), thus it comprises 21 months. To avoid (i) the impact of spatial density of observations and (ii) the potential lack of statistical significance due to spatial dependence between close ones (both are typical problems of model interpolation to the observation points), an up-scaling method (Cherubini et al., 2002) has been applied. Precipitation observations have been provided by ECMWF (who collects raw data from different member and cooperating states over Europe) already up-scaled to $0.25^\circ \times 0.25^\circ$ boxes, picking up only those ones in which a minimum significance is guaranteed with at least five observations available inside (which makes the number of daily observations not to be constant). The box average value is used as representative of the precipitation over that box.

AE and EC forecasts for the same period are compared, this latter selected as a high-performance available reference ensemble forecast. EC is a medium-range forecast system, whereas AE is a short-range one, so any comparison should take this difference into account, especially by choosing a suitable range of selected forecast lengths. As climatological networks often record precipitation from 07 to 07 UTC and only 00UTC run was available at the whole period, to assess 24-h accumulated precipitation only T+30 and T+54 rain forecasts (from the system whole forecast range outputs T+0 to T+72 every 6 h) can be used. In this case T+54 has been selected as the only fair possibility to compare EC and AE performance. 24 h accumulated precipitation forecast is computed for the time window 06–06UTC; this 1-h shift between observed and forecast periods is not taken into account in this study. The average number of daily realizations is above 1000, whereas the average number of realizations for 3-month periods is about 76 000. Due to their operational forecast importance, rainfall thresholds 1, 5, 10 and 20 mm have been selected.

4.2. Verification strategy, performance measures

To assess the performance of AE and EC in terms of reliability, resolution and discrimination standard probabilistic verification methods have been followed (Wilks, 1995; Jolliffe and Stephenson, 2003; Candille and Talagrand, 2005; Stensrud and Yussouf, 2007).

First, a set of binary events corresponding to the set of rainfall thresholds 1, 5, 10 and 20 mm has been addressed. As usual, for each event (the rainfall threshold is exceeded or not) the joint distribution of observations and forecasts has been computed, giving to any observation the value 0 or 1 (whether the event occurred or not) and in the case of ensemble forecast taking the probability of occurrence as the number of forecasts exceeding the threshold divided by the ensemble size, that is, considering all members equally likely. $N+1$ probability classes are used for an ensemble size of N, in order to obtain the best possible performance measure (Ziehmann, 2000), which means 52 classes for EC and 21 classes for AE. However, as already mentioned in Section 3.3, the impact of the difference in the ensemble size (in this case 51 members for EC and 20 members for AE) has not been included in this study, though it could be addressed in future works (Buizza and Palmer 1998; Ferro, 2007; Ferro et al., 2008). As this difference can give better performance to EC it should be taken into account. From the joint distribution, computed in the form of contingency tables, a set of several performance measures has been obtained. For brevity, only a detailed description of some aspects is given here.

Brier Skill Score (BSS) and its decomposition in reliability and resolution terms (Jolliffe and Stephenson, 2003, p.147) have been selected as summary measures to show an indication of overall skill, reliability and resolution, respectively (a system with BSS greater than 0 is more skilful than the sample climatology, while BSS = 1 indicates a perfect deterministic forecast). ROC (relative operating characteristic) skill area (RSA), obtained as $2A - 1$, where A is the area under the ROC curve, has been used as a measure of discrimination (a system with RSA greater than 0 shows better discrimination than the sample climatology). The ROC areas have been crudely estimated by straight lines joining the points and computing the areas of the underlying trapezoids (empirical method). The estimation by a parameteric method (Swets, 1988; Wilson, 2000; Jolliffe and Stephenson, 2003) would provide a less-biased estimate of the area under the ROC. The empirical method tends to underestimate this area, especially for events with low base rates (higher thresholds), and would tend to favour larger ensemble sizes.

Discrimination is a measure that can provide complementary performance information to BSS. While BSS is relatively insensitive to extreme events (Gutiérrez et al., 2004), RSA is not. On the other hand, RSA can be insensitive to some kinds of forecast biases (Kharin and Zwiers, 2003). Sample climatology for each 3-month period has been used as reference (Mason, 2004). As a caveat, here it is not given assessment of the uncertainty related to the measures used. Different approaches can be followed to give, for instance, confidence intervals (Jolliffe, 2007; Casati et al., 2008; Mason, 2008) and will be included in further studies. To give a minimum statistical and meteorological consistence, any score has been averaged on 3-month periods and over the whole set of grid boxes (averaging first in time and after in space or vice-versa gave similar results). The corresponding time-series with the mentioned 3-month moving average are shown for BSS and RSA separately for each rain threshold (see Section 4.3). Therefore, the performance evolution of AE and EC along 19 months (3-month moving average), with different

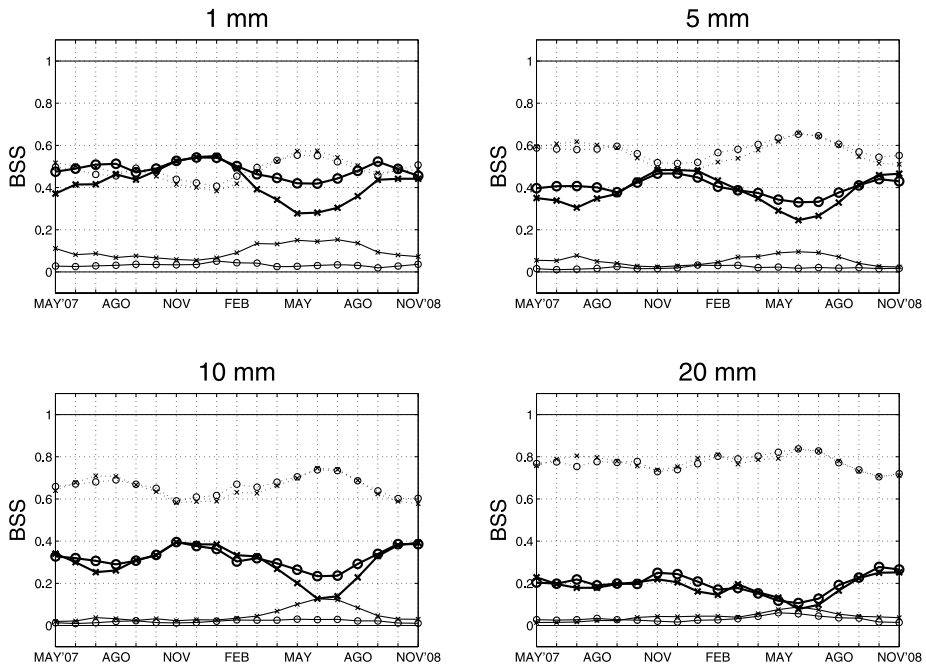

Fig. 8. Time-series of BSS (thick line), BSS reliability (thin lines) and BSS resolution (thin dotted lines) components. Values plotted are 3-month moving average from May 2007 to November 2008 corresponding to binary events of 1, 5, 10 and 20 mm of accumulated precipitation in 24 h given by T+54 forecasts, respectively. Results are shown for AE (circles) and EC (crosses).

aspects of the forecast skill and their seasonal variability can be assessed.

4.3. Summary of verification results for precipitation

Time-series from May 2007 to November 2008 for BSS and its components for rainfall thresholds 1, 5, 10 and 20 mm are shown in Fig. 8. BSS (thick lines) is positively oriented (where 0 indicates no skill with respect to sample climatology and 1 corresponds to a perfect forecast) while reliability (thin dashed lines) and resolution (thin lines) components are negatively oriented (higher scores mean worse performance). As common and general patterns, both systems present their highest skill in winter, skill is degraded as threshold grows and for 20 mm the skill seasonal variability is not so significant. This is consistent with the large-scale precipitation predominance in winter. AE (circles) outperforms EC (crosses) in spring and summer seasons, significantly for 1, 5 and 10 mm, while for 20 mm this clear out-performance happens only in winter. To explain this difference in skill, horizontal resolution plays a main role: AE (0.25°) can better resolve convective systems than EC (0.50°). This skill difference was also expected, but not measured, in autumn: AE convection in autumn is already known to need some improvement. Further detail is given by the BSS reliability and resolution components. AE is clearly more reliable (with a decreasing difference with threshold) especially in spring and summer. On the other hand EC shows better resolution (possibly due to its larger ensemble size) in spring, autumn and winter, except for 20 mm. This resolution is a clear indication of EC forecast quality, as far as resolution cannot be improved under a calibration process. On the other hand, reliability can be improved by calibration at the expense of resolution, and the benefits of calibration are difficult to achieve in an operational context (Atger, 2003). The difference in reliability, higher than that in resolution in spring and summer, gives AE higher BSS in these seasons.

Time-series for the same period and rainfall thresholds are shown for RSA in Fig. 9. RSA is positively oriented, where a value of 0 indicates no better discrimination than the sample climatology, and 1 corresponds to perfect discrimination. For 1, 5 and 10 mm both AE (circles) and EC (crosses) show high values of RSA, slightly better in winter where large-scale precipitation is predominant, and no relevant differences can be found between the two systems. For 20 mm a more clear difference in discrimination is found, AE showing higher RSA in autumn and winter, possibly due to the RSA sensitivity to rare events. The role of horizontal resolution and the relation with predominance of convective activity is not so clear in this case.

In summary, AE shows generally better performance than EC in terms of BSS (while EC shows better resolution, AE is more reliable and gives generally a better overall BSS), with bigger difference in those seasons where convective precipitation is more frequent, whereas when large-scale precipitation is expected to be dominant, no skill difference is observed. These results are consistent with the horizontal resolution of each ensemble, despite the EC advantage in ensemble size. No

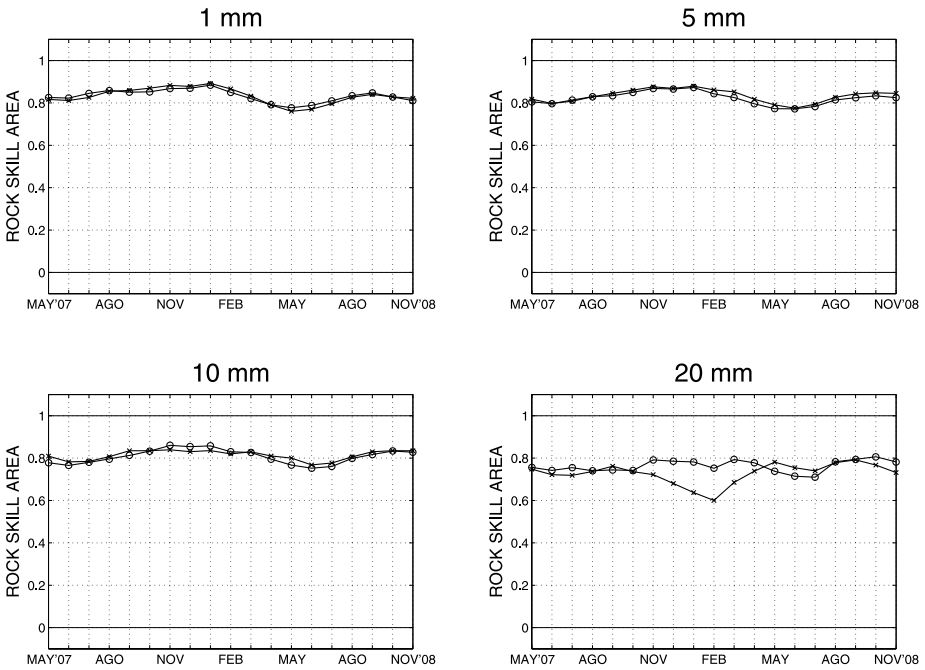

Fig. 9. Time-series of RSA. Values plotted are 3-month moving average from May 2007 to November 2008 corresponding to binary events of 1, 5, 10 and 20 mm of accumulated precipitation in 24 h given by T+54 forecasts, respectively. Results are shown for AE (circles) and EC (crosses).

major differences in discrimination were found according to RSA measures, except for 20 mm threshold where AE showed higher values, possibly due to RSA sensitivity to rare events. If further assessment on the impact of the ensemble size had been done, AE could be expected to give even better performance difference when compared with EC. According to these results, AE improves EC forecasts probably due to its higher horizontal resolution (25–50 km). In January 2010, ECMWF upgraded model resolutions, and EC became to run at 32 km. It is not so clear that AE can now improve the quality of EC in the same way as the results shown here. AEMET plans include an upgrade in resolution into the mesoscale, then new performance assessments should be done.

4.4. Verification of 10 m wind speed and 2 m temperature

To complement the information about validation of weather parameters, in this section a brief summary of 10 m wind speed (10 mWS) verification is given as well as some remarks about 2 m temperature (2 mT). For consistency, the same period has been assessed (April 2007 to December 2008), but now SYNOP stations have been used, and no comparison with other ensembles is provided. The same verification strategy and performance measures than that for precipitation have been used, selecting only a few thresholds for brevity (10 m s^{-1} for 10 mWS). Time-series are depicted with 3-month moving-average from May 2007 to November 2008 for BSS and its components, with RSA in the same graph.

For 10 mWS (Fig. 10) resolution and reliability (both negatively oriented) show the worst values in May–June–July and keep good behaviour along the rest of the period. Thus, the whole BSS (positively oriented) presents local minimums in May–June–July and reaches a minimum value of −0.5 (the system is not more skilful than the sample climatology) in May–June–July 2008. The RSA (as a measure of discrimination) shows the same behaviour, with a minimum value of

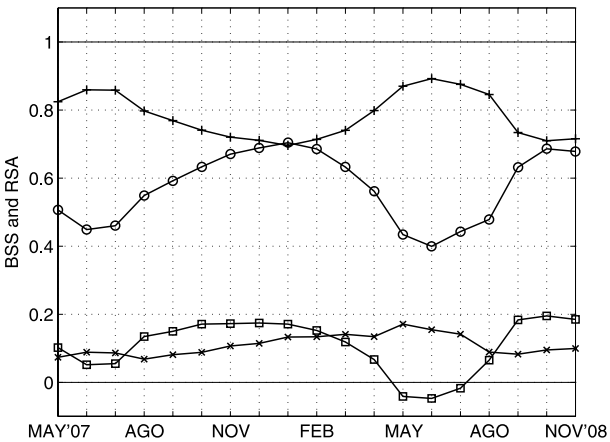

Fig. 10. Time-series of BSS (square), BSS reliability (crosses), BSS resolution (plus sign) components and RSA (circles). Values plotted are 3-month moving average from May 2007 to November 2008, corresponding to binary events of 10 m s^{-1} of wind speed at 10 m given by T+54 forecasts.

0.4 (even here the system is more discriminative than the sample climatology) in June 2008, and values larger than 0.5 almost along the whole period. This is consistent examining the sample climatology: the monthly typical base rates (frequency of occurrence) for the event WS $\geq$ 10 m s^{-1} are about 0.004–0.010; in June–July–August 2008 the base rates were about 0.001–0.002. The sensitivity of BSS (relative) and RSA to rare events is clear in this case. Therefore, the system is generally skilful, always discriminative and provides better information than the sample climatology, except for those months (May, June and July) in which the event shows to be really rare, with base rates under 0.002.

For 2 mT the geographical variability over the verification domain is extremely high and, if the same verification methods were applied, could lead to mix very different sample climatologies and thus producing results of little significance (Hamill and Juras, 2006). For instance, taking the event 2 mT $>$ 0 °C, the monthly base rate for North Europe in winter is far away for that one over North Africa. Separate verifications should be done over smaller subparts of the domain, and then the problem would become the small sample size. More advanced verification methods should be applied in this case that could be object of further studies.

5. Conclusions

A multimodel multiboundary SREPS has been developed at the Spanish Meteorological Service (AEMET). The system consists of five different LAMs using initial and boundary fields from five different deterministic global models. The system is composed of 25 members running twice a day (at 00 and 12 UTC) producing forecasts up to 72 h ahead. The resolution of the models is approximately 25 km in the horizontal and with 40 vertical levels.

The validation of the system on the large-scale flow indicates that AEMET-SREPS (AE) is statistically consistent with ECMWF analysis, with a slight underdispersion typical of state-of-the-art ensembles. The ensemble spread shows a fairly linear correlation with the error of the ensemble mean, better than any of its subensembles. Furthermore, subensembles of LAM models happen to be more underdispersive than subensembles of global models; further research is needed to better understand this behaviour.

For surface parameters (2 m temperature, 10 m wind speed, accumulated precipitation) forecasts, the most important in short-range ensemble forecasting, the system presents high skill (shown only for 10 m wind speed and accumulated precipitation). To illustrate this, assessment of 21 months of 24-h precipitation probabilistic forecasts performance has been shown in detail. Observations from European high-resolution precipitation networks have been used up-scaled to a 25 km grid. AE and ECMWF-EPS (EC) forecasts are compared, selecting an appropriate forecast length of T+54. Rainfall thresholds of 1, 5, 10 and 20 mm have been selected due to their importance in operational forecast. In terms of BSS and its components, AE turns out to be a more reliable system, and despite its worse resolution it shows a higher overall skill, more significant in those seasons where convective activity is higher. Measuring discrimination of both systems with ROC Skill Area (RSA), the only difference was found for 20 mm, where AE discrimination is higher in autumn-winter. AE horizontal resolution (0.25°), compared with that of EC (0.50°) has an important role on this difference in performance, despite its lower ensemble size: convective situations are better described at higher resolution. Therefore, in those seasons where convection is expected there is a significant difference between EC and AE, whereas when large-scale precipitation is predominant no skill difference is found. On the other hand, the role of each ensemble nature is fairly difficult to assess in this context. AE nature is multimodel multiboundaries, while EC is an ensemble built with initial perturbations and stochastic physics. Further research is also needed here, and though some experiences could help to address this issue (e.g. Ziehmann, 2000), a straightforward application to this case is not clear for the authors. Finally, the role of the ensemble size (which could benefit in principle EC with 51 members compared with AE with 20 members) has not been addressed and will be object of further studies.

The different performance assessments done show a system that is statistically consistent with the analysis in the large-scale flow and provides probabilistic forecasts of weather parameters with good reliability, resolution and discrimination. Therefore, the multimodel multiboundaries strategy to build the AE is confirmed as a feasible option to sample initial and model errors. These probabilistic forecasts of surface parameters can help in the forecast guidance as a complementary tool for high-resolution deterministic models. As far as it gives explicit and quantitative information about predictability, as well as several atmospheric scenarios of potential risk, it can be a powerful help in early warnings of severe weather events of great relevance in Spain, for example, Mediterranean heavy rain and floods, Cantabric Sea wind gales and summer heat waves.

6. Acknowledgments

Developing a multimodel ensemble prediction system is a difficult issue to address. The authors are very grateful to many people. First, we would like to acknowledge Dr. Eugenia Kalnay from the University of Maryland for her support at the beginning of the project. We would also like to give specific mention to Jorge Bornemann and Ken Mylne from UKMO, Detlev Majewski and Michael Gertz from DWD, Chiara Marsigli and Ulrich Schättler from the COSMO Consortium, Olivier Talagrand from the Laboratoire de Meterologie Dynamique (ENS Paris) and Anna Ghelli, Martin Leutbecher and the Metview Team from the ECMWF. We would also like to acknowledge the European Meteorological Agencies for allowing us access to their

climate network precipitation data, specifically the agencies in Spain, France, UK, Germany, Portugal, Greece, the Netherlands, Finland, Hungary, Romania, Bulgaria, Slovenia, Slovakia, and the Italian Regions of Emilia-Romagna, Lombardia, Marche, Trento and Friuli-Venezia-Giulia.

References

Andersson, E., Haseler, J., Undén, P., Courtier, P., Kelly, G. and co-authors. 1998. The ECMWF implementation of the three-dimensional variational assimilation (3D-Var). III: experimental results. *Q. J. R. Meteorol. Soc.* **124**, 1831–1860.

Anderson, J. L. 1996. A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **9**, 1518–1530.

Arribas, A., Robertson, K. B. and Mylne, K. R. 2005. Test of poor man's ensemble prediction system. *Mon. Wea. Rev.* **133**, 1825–1839.

Atger, F. 2003. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: consequences for calibration. *Mon. Wea. Rev.* **131**, 1509–1523.

Bishop, C. H., Etherton, B. J. and Majumdar, S. J. 2001. Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Wea. Rev.* **129**, 420–436.

Bowler, N. E. and Mylne, K. R. 2009. Ensemble transform Kalman filter perturbations for a regional ensemble prediction system. *Quart. J. Roy. Met. Soc.* **135**, 757–766.

Bowler, N. E., Arribas, A., Mylne, K. R., Robertson K. B. and Beare S. E. 2008. The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**, 703–722.

Bowler, N. E., Arribas, A., Beare, S. E., Mylne, K. R. and Shutts G. J. 2009. The local ETKF and SKEB: upgrades to the MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **135**, 767–776.

Bright, D. R. and Mullen, S. L. 2002. Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather Forecast.* **17**, 1080–1100.

Buizza, R. and Palmer, T. 1995. The singular vectors structure of the atmospheric general circulation. *J. Atmos Sci.* **52**, 1434–1456.

Buizza, R. and Palmer, T. 1997. Potential forecast skill of ensemble prediction, and spread and skill distributions of the ECMWF ensemble prediction system. *Mon. Wea. Rev.* **125**, 99–119.

Buizza, R. and Palmer T. N. 1998. Impact of ensemble size on ensemble prediction. *Mon. Wea. Rev.* **126**, 2503–2518.

Buizza, R., Miller, M. and Palmer, T. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908.

Candille, G. and Talagrand, O. 2005. Evaluation of probabilistic prediction systems for a scalar variable. *Q. J. R. Meteorol. Soc.* **131**, 2131–2150.

Casati, B., Wilson, L. J., Stephenson, D. B., Nurmi, P., Ghelli, A. and co-authors. 2008. Forecast verification: current status and future directions. *Met. Appl.* **15**, 3–18.

Cherubini, T., Ghelli, A. and Lalaurette, F. 2002. Verification of precipitation forecasts over the Alpine region using a high-density observing network. *Weather Forecast*, **17**, 238–248.

Clark, A. J., Gallus W. A. Jr. and Chen, T. 2008. Contributions of mixed physics versus perturbed initial/lateral boundary conditions to ensemble-based precipitation forecast skill. *Mon. Wea. Rev.* **136**(6), 2140–2156.

Côté, J., Desmarais, J. G., Gravel, S., Méthot, A., Patoine, A. and co-authors. 1998a. The operational CMC–MRB global environmental multiscale (GEM) model. Part II: results. *Mon. Wea. Rev.* **126**, 1397–1418.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M. and co-authors. 1998b. The operational CMC–MRB global environmental multiscale (GEM) model. Part I: design considerations and formulation. *Mon. Wea. Rev.* **126**, 1373–1395.

Cullen, M. J. P. 1993. The unified forecast/climate model. *Meteorol. Mag.* **122**, 81–94.

Doms, G. and Schättler, U. 1997. The nonhydrostatic area model LM (Lokal-Modell) of DWD. Part I: scientific documentation. Deutscher Wetterdienst (DWD), Offenbach. March 1997.

Dudhia, J. 1993. A nonhydrostatic penn state-NCAR mesoscale model: validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Wea. Rev.* **121**, 1493–1523.

Ebisuzaki, W. and Kalnay, E. 1991. Ensemble experiments with new lagged average forecasting scheme. WMO Research Activities in Atmospheric and Oceanic Modeling Rep. 15, 308 pp.

Emanuel, K. A. 1979. Inertial instability and mesoscale convective systems. Part I: linear theory of inertial instability in rotating viscous fluids. *J. Atmos. Sci.* **36**, 2425–2449.

Ferro, C. A. T. 2007. Comparing probabilistic forecasting systems with the Brier score. *Weather Forecast,* **22**, 1076–1089.

Ferro, C. A. T., Richardson, D. S. and Weigel, A. P. 2008. On the effect of ensemble size on the discrete and continuous ranked probability scores. *Met. Appl.* **15**, 19–24.

Frogner, I.-L. and Iversen, T. 2001: Targeted ensemble prediction for northern Europe and parts of the north Atlantic Ocean. *Tellus* **53A**, 35–55.

Frogner, I.-L., Haakenstad, H. and Iversen, T. 2006. Limited-area ensemble predictions at the Norwegian Meteorological Institute. *Q. J. R. Meteorol. Soc.* **132**, 2785–2808.

Grell, G. A., Dudhia, J. and Stauffer, D. R. 1994. A description of the fifth-generation Penn State/NCAR mesoscale model (MM5). NCAR Technical Note, NCAR/TN-398+STR, 117 pp.

Gutiérrez, J. M., Cofiño, A. S., Cano, R. and Rodríguez, M. A. 2004. Clustering methods for statistical downscaling in short-range weather forecasts. *Mon. Wea. Rev.* **132**, 2169–2183.

Hacker, J. P., Krayenhoff, E. S. and Stull, R. B. 2003. Ensemble experiments on numerical weather prediction error and uncertainty for a north Pacific forecast failure. *Weather Forecast,* **18**, 12–31.

Hamill, T. M. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.* **129**, 550–560.

Hamill, T. M. and Colucci, S. J. 1997: Verification of ETA-RSM short-range ensemble forecast. *Mon. Wea. Rev.* **125**, 1322–1327.

Hamill, T. M. and Colucci, S. J. 1998. Evaluation of Eta-RMS ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.* **126**, 711–724.

Hamill T. M. and Juras J. 2006. Measuring forecast skill: is it real or is it the varying climatology?. *Q. J. R. Meteorol. Soc.* **132**, 2905–2923.

Hamill, T. M., Snyder, C. and Morss, R. E. 2000. A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.* **128**, 1835–1851.

Hoffman, R. N. and Kalnay, E. 1983. Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**, 100–118.

Hohenegger, C. and Schär, C. 2007: Atmospheric predictability at synoptic versus cloud-resolving scales. *Bull. Am. Meteor. Soc.* **88**, 1783–1793.

Hollingsworth, A. 1980. An experiment in Monte Carlo forecasting. In: *Proceedings of the Workshop on Stochastic-Dynamic Forecasting*, Reading, United Kingdom, ECMWF, 65–85.

Hou, D., Kalnay, E. and Droegemeier, K. 2001. Objective verification of the SAMEX'98 ensemble forecast. *Mon. Wea. Rev.* **129**, 73–91.

Houtekamer, P. L., Lefaiver, L., Derome, J., Ritchie, H. and Mitchell, H. L. 1996. A system simulation approach to ensemble prediction. *Mon. Wea. Rev.* **124**, 1225–1242.

Jakob, C., Andersson, E., Beljaars, A., Buizza, R., Fisher, M. and co-authors. 1999. The IFS cycle CY21r4 made operational in October 1999. *ECMWF Newsletter,* **87**, 2–9.

Jolliffe, I. T. 2007. Uncertainty and inference for verification measures. *Weather Forecast,* **22**, 637–650.

Jolliffe, I. T. and Stephenson, D. B. 2003. *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, New York.

Kharin, V. V. and Zwiers, F. W. 2003. On the ROC score of probability forecasts. *J. Clim.* **16**, 4145–4150.

Leith, C. E. 1974. Theoretical skill of Monte Carlo forecast. *Mon. Wea. Rev.* **102**, 409–418.

Leslie, L. M. and Speer, M. S. 1998. Short-range ensemble forecasting of explosive Australian east coast cyclogenesis. *Weather Forecast,* **13**, 822–832.

Lorenz, E. N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141.

McDonald, A. and Haugen, J. 1992. A two-time-level, three-dimensional semi-Lagrangian, semi-implicit, limited-area gridpoint model of the primitive equations. *Mon. Wea. Rev.* **120**, 2603–2621.

Majewski, D. 1991. The Europa-Modell of DWD. In: *Proceedings of ECMWF Seminar on Numerical Methods in Atmospheric Science* **2**, 147–191, ECMWF, Reading, UK.

Majewski, D., Liermann, D., Prohl, P., Ritter, B., Buchhold, M. and co-authors. 2002. The operational Global Icosahedral-Hexagonal gridpoint model GME: description and high-resolution tests. *Mon. Wea. Rev.* **130**, 319–338.

Marsigli, C., Montani, A., Nerozzi, F. and Paccagnella, T. 2004. Probabilistic high-resolution forecast of heavy precipitation over Central Europe. *Nat. Hazard. Earth Sys.* **4**, 315–322.

Marsigli, C., Montani, A. and Paccangnella, T. 2008. A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Met. Appl.* **15**, 125–143.

Mason, S. J. 2004. On using "climatology" as a reference strategy in the Brier and ranked probability skill scores. *Mon. Wea. Rev.* **132**, 1891–1895.

Mason, S. 2008. Understanding forecast verification statistics. *Met. Appl.* **15**, 31–40.

Molteni, F., Buizza, R., Palmer, T. N. and Petroliagis, T. 1996. The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–120.

Mullen, S. L. and Baumhefner, D. P. 1989. The impact of initial condition uncertainty on numerical simulations of large-scale explosive cyclogenesis. *Mon. Wea. Rev.* **117**, 2800–2821.

Murphy, J. M. 1988. The impact of ensemble forecasts on predictability. *Q. J. R. Meteorol. Soc.* **114**, 463–493.

Palmer, T. N., Barkmeier, J., Buizza, R. and Petroliagis, T. 1997. The ECMWF ensemble prediction system. *Meteorol. Appl.* **4**, 301–304.

Palmer, T. N., Alessandri, A., Andersen, U., Cantelaube, P., Davey, M. and co-authors. 2004. Development of a European multi-model ensemble system for seasonal to inter-annual prediction (DEMETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872.

Podzun, R., Cress, A., Majewski, D. and Renner, V. 1995. Simulation of European climate with a limited area model. Part II: AGCM boundary conditions. *Contrib. Atmos. Phys.* **68**, 205–226.

Roebber, P. J. and Reuter, G. W. 2002. The sensitivity of precipitation to circulation details. Part II: mesoscale modeling. *Mon. Wea. Rev.* **130**, 3–23.

Sela, J. G. 1980. Spectral modeling at the National Meteorological Center, Mon. *Wea. Rev.* **108**, 1279–1292.

Sela, J. G. 1982. The NMC spectral model, NOAA Technical Report NWS-30, 36 pp.

Simmons, A. J., Burridge, D. M., Jarraud, M., Girard, C. and Wergen, W. 1989. The ECMWF medium-range prediction models: development of the numerical formulations and the impact of increased resolution. *Meteorol. Atmos. Phys.* **40**, 28–60.

Stensrud, D. J. and Weiss, S. J. 2002. Mesoscale model ensemble forecasts of the 3 May 1999 Tornado outbreak. *Weather Forecast,* **17**, 526–543.

Stensrud, D. J. and Yussouf, N. 2007. Reliable probabilistic quantitative precipitation forecasts from a short-range ensemble forecast system. *Weather Forecast,* **22**, 3–17.

Stensrud, D. J., Bao, J.-W. and Warner, T. T. 2000. Using Initial Condition and Model Physics Perturbations in Short-Range Ensemble Simulations of Mesoscale Convective Systems. *Mon. Wea. Rev.* **128**, 2077–2107.

Stensrud, D. J., Brooks, H. E., Du, J., Tracton, M. S. and Rogers, E. 1999. Using ensembles for the short-range forecasting. *Mon. Wea. Rev.* **127**, 433–446.

Swets, J. A. 1988. Measuring the accuracy of diagnostic systems. *Science*, **240**(4857):1285–1289.

Toth, Z. and Kalnay, E. 1993. Ensemble forecasting at NMC: the generation of perturbations. *Bull. Am. Meteor. Soc.* **74**, 2317–2330.

Toth, Z. and Kalnay, E. 1997. Ensemble forecasting at NCEP: the breeding method. *Mon. Wea. Rev.* **125**, 3297–3318.

Tracton, M. S., Du, J., Toth, Z. and Juang, H. 1998. Short-range ensemble forecasting (SREF) at NCEP/ECM. In: *Proceedings of the 12th Conference on Numerical Weather Prediction*, Phoenix, American Meteorological Society, pp. 269–272.

Undén, P., Rontu, L., Järvinen, H., Lynch, P., Calvo, J. and co-authors. 2002. HIRLAM-5 Scientific Documentation. Available from Hirlam-5 Project, c/o Per Undén, SMHI, S-60176, Norrköping, Sweden, 144 pp.

Wandishin, M. S., Stensrud, D. J., Mullen, S. L. and Wicker, L. J. 2008. On the predictability of mesoscale convective systems: two-dimensional simulations. *Weather Forecast,* **23**, 773–785.

Wandishin, M. S., Stensrud, D. J., Mullen, S. L. and Wicker, L. J. 2010. On the predictability of mesoscale convective systems: three-dimensional simulations. *Mon. Wea. Rev.* **138**, 863–885.

Wang, X. and Bishop C. H. 2003. A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* **60**, 1140–1158.

Whitaker, J. S. and Loughe, A. F. 1998. The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.* **126**, 3292–3302.

Wilks, D. S. 1995. *Statistical Methods in Atmospheric Sciences*. Academic Press, San Diego, CA, 467 pp.

Wilson, L. J. 2000. Comments on "Probabilistic Predictions of Precipitation Using the ECMWF Ensemble Prediction System". *Weather Forecast,* **15**, 361–364.

Zhang, F. 2005. Dynamics and structure of mesoscale error covariance of a winter cyclone estimated through short-range ensemble forecasts. *Mon. Wea. Rev.* **133**, 2876–2893.

Ziehmann, C. 2000. Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus* **52A**, 280–299.