Open access • Journal Article • DOI:10.1002/QJ.2622

Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX — Source link

François Bouttier, Laure Raynaud, Olivier Nuissier, Benjamin Ménétrier

Institutions: Centre national de la recherche scientifique

Published on: 01 Aug 2016 - Quarterly Journal of the Royal Meteorological Society (John Wiley & Sons, Ltd)

Topics: Ensemble forecasting and Data assimilation

Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX

François Bouttier, Laure Raynaud, Olivier Nuissier, Benjamin Ménétrier

Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX

François Bouttier, Laure Raynaud, Olivier Nuissier, Benjamin Ménétrier

3 July 2015

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

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

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

Funding information: Météo-France and CNRS.

Cite as: Bouttier, F., L. Raynaud, O. Nuissier and B. Ménétrier, 2015: Sensitivity of the AROME ensemble to initial and surface perturbations during HyMeX. *Quart. J. Roy. Meteor. Soc.*, **142**, 390-403. doi:10.1002/qj.2622

Abstract

The AROME-EPS convection-permitting ensemble prediction system has been evaluated over the HyMeX-SOP1 period. Objective verification scores are computed using dense observing networks prepared for the HyMeX experiment. In probabilistic terms, the AROME-EPS ensemble performs better than the AROME-France deterministic prediction system, and a state-of-the-art ensemble at a lower resolution. The strengths and weaknesses of AROME-EPS are discussed. Here, impact experiments are used to study perturbation schemes for the initial conditions and the model surface. Both have a significant effect on the ensemble performance. The interactions between the perturbations of lateral boundaries, initial conditions and surface perturbations are studied. The consistency between initial and lateral perturbations is found to be unimportant from a meteorological point of view. Ensemble data assimilation is not as effective as a simpler surface perturbation scheme, and it is noted that both approaches could be usefully combined.

keywords: atmospheric model, numerical weather prediction, AROME model, HyMeX observations, ensemble prediction, surface perturbations, ensemble data assimilation.

1 Introduction

Ensemble prediction has become a standard technique for probabilistic numerical weather prediction. It is also an important component of ensemble-based data assimilation systems. After several decades of development with global models, ensemble prediction is now used with limited area models at convection-permitting kilometric horizontal resolutions. Despite their numerical cost, convection-permitting resolutions have been shown to be useful in deterministic prediction systems (e.g. Seity *et al.* , 2011), and similar benefits are beginning to be demonstrated in high resolution ensemble prediction systems.

Advanced convection-permitting ensemble prediction systems include COSMO-DE-EPS (Gebhardt *et al.* , 2008; Montani *et al.* , 2011), MOGREPS-UKV (Migliorini *et al.* , 2011), WRF-based ensembles (Clark *et al.* , 2011; Hacker *et al.* , 2011), and AROME-EPS (Bouttier *et al.* , 2012; Nuissier *et al.* , 2012; Vié *et al.* , 2011), which is used in this paper. Modern ensembles usually include perturbations to the initial conditions (to represent uncertainties on atmospheric analyses) and to the surface conditions (to represent uncertainties on the analysis and modelling of surface variables). The purpose of this work is to objectively quantify the benefits of a high resolution ensemble, and to understand the relative importance of initial and surface perturbations. A key difficulty of high-resolution ensemble verification is the need to use large forecast samples and dense observation datasets, because the models used are typically restricted to a small geographical area: this limits the number of independent weather events that can be sampled on any given day. The feasible length of ensemble prediction experiments is limited by their high computational cost, and dense observation datasets can be difficult to collect routinely. These constraints hamper research on state-of-the-art ensembles by making it difficult to extract robust information from probabilistic scores (Candille and Talagrand, 2005). In the impact experiment considered here, the aim is to decide whether modifications to an ensemble prediction system significantly alter its performance or not. We pursue this objective over a period and area where an exceptionally dense set of observations is available: the Special Observing Period 1 of the HyMeX field experiment, called HyMeX-SOP1 below.

The HyMeX-SOP1 is described in Ducrocq *et al.* (2014). There is substantial geographical overlap between the AROME-EPS ensemble system studied here, and the HyMeX-SOP1 area of interest. Much observation data collected for HyMeX-SOP1 is relevant for the AROME-EPS verification, and most of

it is not available for use during normal operations because of institutional and commercial data policies. As documented below, the HyMeX-SOP1 specific data is primarily located around the Mediterranean sea. It is relatively rich in observations of high precipitation events.

The design of convection-permitting ensembles differs from global ensembles because of the higher resolution and the small geographical extent of the used models. Following the principles of ensemble prediction, the aim is to run forecasts where perturbations have been introduced to represent the major sources of uncertainty in the numerical simulation process. Four types of perturbations are known to be important: the representation of initial condition errors (e.g. Stensrud *et al.* , 2000; Wei *et al.* , 2006; Li *et al.* , 2008; Hacker *et al.* , 2011; Peralta *et al.* 2012), surface errors (e.g. Barthlott and Kalthoff, 2011; Hacker *et al.* , 2011; Lavaysse *et al.* , 2013; Tennant and Beare, 2014), model errors (e.g. Buizza *et al.* , 1999; Schwartz *et al.* , 2010; Berner *et al.* , 2011; Gebhardt *et al.* , 2011; Bouttier *et al.* , 2012) and large scale boundary condition errors (e.g. Gebhardt *et al.* , 2011). Model and large scale boundary condition errors have been studied in several ensemble systems including AROME-EPS ; although their current representation is far from perfect, they are not the focus of this paper, and we shall use the perturbation methods already documented by Bouttier *et al.* (2012) for the model errors, and Nuissier *et al.* (2012) for the large scale boundary conditions.

The purpose of this work is to investigate the sensitivities of AROME-EPS to the design of initial and surface perturbations, in order to identify optimal perturbation strategies. New questions about these perturbations are raised; some have been partly answered in large-scale ensembles, but experience with convection-permitting ensembles remains limited. This study aims at improving an operational forecasting system, so that its focus is on actual weather parameters that are easily compared with conventional observations: screen-level temperature, humidity, wind speed and accumulated precipitation.

It has been shown that convection permitting forecasts can be sensitive to certain surface fields, such as soil moisture (e.g. Barthlott and Kalthoff, 2011). Modern high-resolution atmospheric models often involve many surface parameters and processes. The contribution of each to forecast uncertainties is not always well known, so that it is not clear which should be perturbed in ensemble prediction. They are not necessarily the same as the ones identified in lower resolution ensembles (notably Lavaysse *et al.* , 2013 and Tennant and Beare, 2014) because the scales and forecasts ranges are different here. A goal of this paper is to identify an effective surface perturbation strategy for the AROME-EPS system.

The main question about initial perturbations that is raised here, is the usefulness of an ensemble data assimilation scheme with respect to a cheaper perturbation scheme. We are also interested in clarifying the relationship between the various perturbation schemes: is it important to have consistent perturbations of initial, large scale and surface conditions ? A key concern is the impact of the ensemble data assimilation approach on the surface perturbation strategy outlined above.

The investigations are performed by testing modifications of the AROME-EPS system with respect to a reference version, which is scheduled to enter operational production at Météo-France by 2016. The structure of the paper is as follows. The HyMeX-SOP1 data, AROME-EPS ensemble and verification methodology are explained in section 2. The performance of the reference AROME-EPS ensemble during HyMeX-SOP1 is documented in section 3. Sections 4 and 5 show the impact of changes to the surface and initial perturbation schemes, respectively. The results are summarised and discussed in section 6.

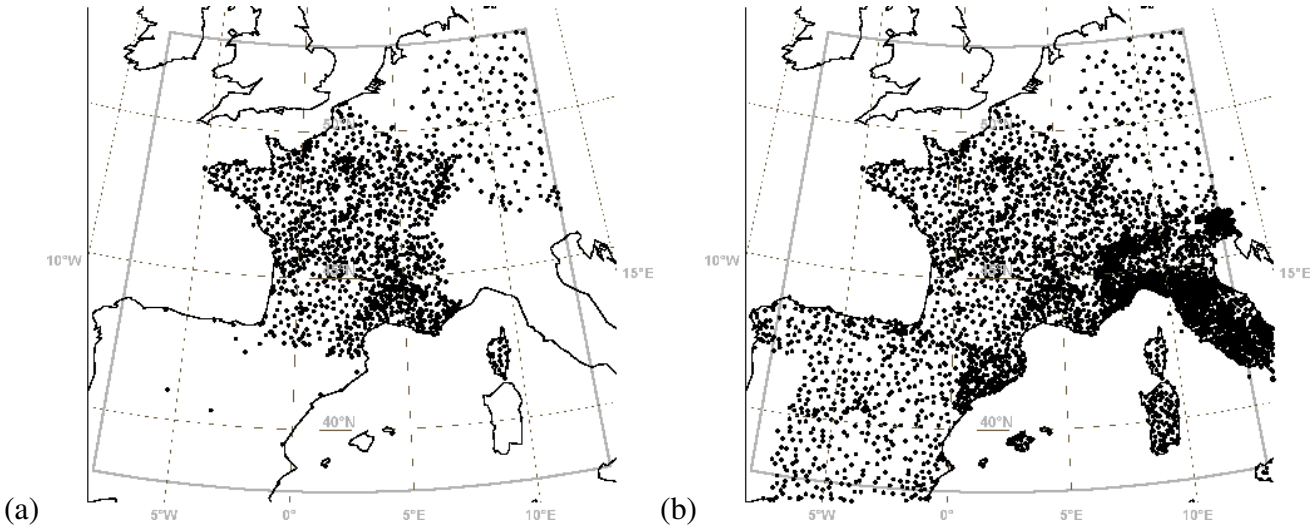

Figure 1: Geographical distribution of stations that provided usable rr3 precipitation data at least 30% of the time during the 75-day period. (a): routinely available data, (b): data obtained from the HyMeX database. The thick grey contour delineates the verification area. The AROME model domain is approximately identical to the plotting area.

2 Experimental design

2.1 The HyMeX-SOP1 verification data

As documented in Ducrocq *et al.* (2014), the HyMeX-SOP1 period lasted from 5 Sept to 5 Nov 2012. Its focus was the study of heavy precipitation events over the Northwestern Mediterranean region, including northeastern Spain, southeastern France and northern Italy including the Alpine range. In this study, we deal with the region covered by the AROME-EPS system, which covers France and most of the HyMeX-SOP1 area. There hardly was any exceptional rain event during the SOP1 proper, but our experiments spans a slightly longer period that includes a dozen heavy Mediterranean rain cases, with daily rainfall accumulations exceeding 100mm. The AROME-EPS system also samples some oceanic and continental weather since its domain includes parts of England and Germany. The AROME-EPS verification and model domains used are represented in Figure 1.

The HyMeX-SOP1 period is an opportunity to access specially collected observations that would not be available at other times. The HyMeX-SOP1 database contains about three times more such data than routinely available to a national weather centre like Météo-France, through improvements in both geographical coverage and observation frequency. Routinely collected data includes dense national networks, and sparser data from WMO networks and reports provided by neighbouring countries through bilateral agreement. In the area of interest, abundant low-level observations were released by participating institutes during HyMeX-SOP1. both geographical coverage and observation frequency. In this study, three-hourly model output will be verified against observations of two metre temperature (T2m), relative humidity (RH2m, converted from dewpoint as necessary), ten metre wind speed (ff10m), and precipitation over three-hour periods (rr3). The average observation densities are shown in Figure.1 for precipitation, which illustrates the benefits of using HyMeX observations. The data distribution is such that scores will be dominated by forecast performance over land in Spain and France. Italy features a high

Table 1: Typical number of usable observations per day, available routinely and using the HyMeX dataset. σ_o: assumed observation standard error. Threshold: values used to define the binary event in the scores, unless otherwise specified.

	N_{obs}/day routine	N_{obs}/day HyMeX	σ_o	threshold
T2m (K)	1700	2800	1	288
RH2m (%)	1400	2400	5	80
ff10m (m/s)	1200	2000	0.33	5.56
rr3 (mm)	1200	4200	0.2+0.2rr3	10
U250 (m/s)	400	400	1.5	20
T250 (K)	400	400	1.3	215

density of precipitation observations, but a density is lower for temperature, humidity and wind speed. The observation counts are nearly constant over the whole period studied in this paper: from 5 Sept. to 18 Nov. 2012, i.e. 75 days.

The observations used have been quality controlled, in order to weed out erroneous and duplicate data. Excessively dense parts of the network have been thinned down. Data has been further filtered by comparing observed values departures with operational AROME-France forecasts, in order to remove stations that are clearly not suitable for model verification. In this process, called data selection, stations with time-averaged rms (root mean square) errors much larger than usual AROME-France error statistics are removed from the verification dataset. The data selection rejects about 5% of observations from each type (T2m, RH2m, ff10m, rr3). They usually are stations in complex mountainous terrain, with obviously large representativeness errors. This selection procedure is not optimal for precipitation, because it tends to remove stations with no representativeness problem, but that have observed intense rainfall events. These events are usually highly local and associated with large forecast errors. It would be interesting to improve the sampling of these events by keeping heavy rain reports that look consistent with independent data (e.g. neighbouring reports or radar reflectivities), but implementing this rather complex procedure has been left for a future study.

The resulting data volumes used for computing ensemble scores are indicated in Table 1. In summary, using HyMeX-SOP1 observations instead of the routine dataset has the following consequences: (a) a two- to three-fold increase in data volumes, (b) high data density over Spain, France, and Italy (the latter for precipitation only), (c) the HyMeX-specific data appears to have good quality because its data selection statistics are similar to routine data, (d) HyMeX data contains many more high rain observations (more than 10mm in 3 hours) than the routinely available data over the SOP1 period. As will be seen in the score studies presented in the following sections, this data helps with the identification of statistically significant changes to high precipitation scores.

2.2 The AROME reference ensemble prediction system

The AROME model and ensemble prediction system have been documented in Seity *et al.* (2011) and Bouttier *et al.* (2012), respectively. A brief summary will be given here. The main novelty here is the introduction of new surface and initial perturbation schemes, as explained below.

Table 2: Main features of the AROME-EPS preoperational and PEARP operational ensembles. The resolutions are quoted for the verification area used in this paper. The same models are used in AROME-EPS and in the AROME-France deterministic system.

	AROME-EPS	PEARP
horizontal resolution	2.5km	15.5km
vertical levels	60	65
lowest model level height	10m	17m
number of members	12	35

AROME is a non hydrostatic, limited area atmospheric numerical modelling system, designed for numerical weather prediction at kilometric scales. Its model and data assimilation software is derived from ECMWF's model, Météo-France's ARPEGE model, and the model of the ALADIN European numerical weather prediction consortium. Most of the AROME physical parametrisations are derived from the French Méso-NH mesoscale research model (Seity *et al.* , 2011). The AROME modelling system is operationally used for weather prediction in most national weather services of the ALADIN and HIRLAM consortia (sometimes under the 'HARMONIE' name), which represents about 20 European and North African countries.

The AROME configuration used in this study is identical to the AROME-France numerical weather prediction system in operations at Météo-France during 2014, with a 2.5km horizontal mesh, 60 vertical levels, a domain size of about 1600×1600km, and a three-hourly 3D-Var data assimilation scheme (Brousseau *et al.* , 2011). The physics used are:

- a radiation scheme borrowed from ECMWF's IFS model with six visible spectral bands, the RRTM scheme for long-wave radiation, and climatological aerosol distributions.
- ICE3, a five-species cloud/precipitation microphysics scheme that manages cloud water, cloud ice, precipitating rain, snow and graupel, with a PDF-based sedimentation scheme.
- a 1-D vertical mixing scheme with prognostic turbulent kinetic energy.
- a sub-grid shallow convection scheme based on an eddy diffusivity mass flux approach.
- SURFEX, a detailed coupled surface model that computes the evolution of independent tiles to model prognostic soil, vegetation, towns, lakes, sea and ice surface; it also contains a multilayer prognostic snow scheme, an interface to the ECOCLIMAP-2 physiographic database, and a 1-D turbulent surface layer scheme to diagnose screen-level variables.

The AROME-EPS ensemble setup used as the reference in this paper is called REF. It is similar to Bouttier *et al.* (2012), except for the surface and initial perturbations. It is very close to the AROME-EPS system that will become operational at Météo-France. The most relevant features are summarized in table 2.

- the ensemble runs start daily at 0000 UTC and comprise 12 members up to 36-hour range. Some experiments only have 6 members.

- lateral and upper boundary conditions (LBCs) are provided by selected members of the PEARP ensemble run started at the same time. PEARP is the French operational global ensemble prediction system, its configuration is documented in Descamps *et al.* (2014), and summarized in table 2. The members used for the AROME-EPS runs are selected using the Nuissier *et al.* (2012) technique: the PEARP 35-member ensemble forecasts are classified by a complete-linkage clustering technique. The clustering distance function is the rms difference between upper-level atmospheric fields, in the AROME verification domain (Figure 1) and over the forecast ranges of the AROME ensemble run. In each PEARP cluster, the member that is closest to its centroid is selected to provide LBCs to an AROME-EPS run.

- atmospheric model errors are represented by the SPPT scheme (stochastic perturbation of physics tendencies) described in Bouttier *et al.* (2012). SPPT perturbations are random numbers that multiply the wind, temperature, and humidity tendencies of the AROME physical parameterisations. The perturbations are correlated in time and space using prescribed functions, and their amplitude is such that the tendency perturbations typically range between 50% and 150% of the parametrised tendencies. The SPPT scheme is not applied near the surface and the top of the model. It has a small, but beneficial impact on the ensemble performance (Bouttier *et al.* 2012).

- surface conditions of the AROME model are perturbed using a new scheme, whereby auto-correlated random modifications are applied to various aspects of the SURFEX surface model. The surface perturbation scheme is documented and studied in section 4.

- initial conditions are perturbed using a simple scheme based on re-scaled PEARP initial perturbations. Section 5 will test possible improvements to this scheme taking into account interactions with the surface perturbations.

In existing ensemble studies, the sensitivity of ensemble performance to each type of perturbation (lateral, model, surface or initial) is often studied in isolation from the others. Here, we investigate their impact in combination with all other perturbation classes, taking into account their mutual interactions.

2.3 Verification procedures

The verification methodology is similar to Bouttier *et al.* (2012). Here, we use the larger HyMeX observation database, and the experiments are longer, to facilitate the identification of statistically robust impacts.

The ensemble performance is measured using objective scores to compare in situ observations with raw (i.e. uncalibrated) ensemble forecasts for the same parameters. Three-hourly observations of T2m, RH2m, ff10m and rr3 are compared with AROME forecasts horizontally interpolated at station location, using a nearest neighbour method, with a ten-minute tolerance on the observation time. T2m, RH2m and ff10m are vertically interpolated from the model vertical levels using a high resolution, one-dimensional prognostic boundary layer model. The assumed observation errors are listed in Table 1.

The following scores will be taken into account when assessing ensemble experiments; several of them have been discussed in Candille and Talagrand (2005):

- rms errors of the member forecasts. Their realism will be measured by the ensemble mean of the member rms errors.

- the spread-to-skill ratio, which is the standard deviation of ensemble members (i.e. the internal ensemble spread) divided by the sum of two terms: the rms error of the ensemble mean and the observation error standard deviation. This ratio should be one in an ensemble with correct spread.

- rank diagrams (sometimes called rank histograms or Talagrand diagrams), which measure the consistency with observations of ensemble bias and spread. The distance between this diagram and a perfect one can be measured (for any ensemble size) by the delta score, or by the outlier frequency. In this paper, both scores are normalized so that they do not depend on ensemble size, and they are equal to one in a perfect ensemble, as recommended in Candille and Talagrand (2005).

- ROC (relative operating characteristic) diagrams, a graphical summary of the decision-making skill of an ensemble with respect to user defined a binary event. The events are defined as the forecast being above a given threshold. The thresholds are given in table 1. The skill of an ensemble can be summarised by the ROC area, which is the scaled area under the ROC curve (Clark *et al.* , 2011).

Other scores have been examined (notably, the ensemble mean score, continuous ranked probability score (CRPS), Brier score, centred control variable score, relative economic value and reliability diagram) but they will not be shown here, because they have been found to convey essentially the same information as is already given below. Some information on these scores can be found in Jolliffe and Stephenson (2003). The event thresholds used for the Brier and ROC scores have been chosen so that they are crossed with a reasonable frequency. It would have been interesting to verify precipitation against a higher threshold than used here, but then some scores would have been difficult to interpret because of high sampling noise. In this work, the focus is on obtaining results that are likely to be reproducible on other periods than the HyMeX-SOP1 experiment.

The evaluation of statistical significance is a key aspect of ensemble evaluation. As explained in Candille *et al.* (2007), some aspects of ensemble prediction performance are very difficult to score objectively, because they can require an unpractically large sample of independent test cases and observations to verify. Here, the following steps have been taken in order to minimise the risk that apparent score variations in our experiments are, in fact, statistical artefacts of the data sample used:

- when preparing the HyMeX data, observations with the largest rms departures from the model have been removed as already explained above. This data selection procedure prevents the scores from being overwhelmed by a few stations with high representativeness errors.

- the aim being the identification of score changes between couples of ensembles, bootstrap statistical significance testing is applied to score differences between experiments: the question asked is whether one ensemble configuration is better or worse than the other.

- scores are computed three-hourly for each of the 75 forecast days. The score differences are independently tested at each forecast range. Thus, the test is insensitive to spatial correlations, but forecast error correlation between successive days are neglected. This is similar to Candille *et al.* (2007), where results suggest that serial correlation of forecasts errors can be neglected at the short ranges (less than 36 hours) considered here.

- a bootstrap technique at the 95% level is used to decide whether the apparent sign of an average score difference between two ensembles is statistically significant: each 75-day score average is recomputed many times using random draws of daily scores, and the average score difference is deemed significant if its sign is not contradicted by more than 5% of the draws.

Each daily score uses thousands of observations, so that the data being tested is quite trustworthy to start with. The main sources of variations in the score distributions are the diurnal cycle, which is taken into account in the analysis of the results, and the large scale weather types. In this paper, all statements made about ensemble scores are supported by the 95%-level bootstrap test in the sense of the above procedure. Thus, they are robust with respect to the conditions experienced during HyMeX-SOP1, but they might not be relevant to different weather conditions.

3 Performance of the pre-operational AROME ensemble

The performance of the AROME-EPS system is examined by comparing it with the PEARP ensemble and the AROME-France deterministic system. PEARP has the potential advantage of having many more members, and the drawback of having a lower model resolution. It is noted that the systematic errors of AROME-EPS and AROME-France can be slightly different from each other, because although the distributions of ensemble perturbations are designed to have zero mean, they can change the average AROME-EPS model behaviour through nonlinear effects.

The ensemble means of the rms scores of the forecasts used in the three systems are shown in Figure 2. AROME-France has the smallest rms errors, followed by AROME-EPS where ensemble perturbations introduce errors, and by PEARP which also suffers from its lower resolution. The main exception is the PEARP precipitation rms error, which is smaller in PEARP than in both AROME systems. All these rms differences are statistically significant according to a bootstrap test at the 95% level. The good precipitation scores of PEARP are caused by the 'double penalty' effect affecting AROME systems: smaller-scale precipitating structures simulated by higher resolution models tend to increase point-based error measures such as the rms, even if these features are physically more realistic than the smoother fields produced by lower resolution models. The double penalty can be avoided to some extent by the use of neighbourhood methods (Ebert, 2008). The following results will show that ensemble prediction is another way to circumvent the double penalty issue, because probabilistic scores are indeed improved by the higher AROME-EPS resolution.

The ensemble spread and biases are illustrated by the rank diagrams in Figure 3. Ranges shorter than nine hours have been discarded because they are affected by some spin-up of ensemble spread in AROME-EPS, as will be shown in section 5. The rank diagrams have no significant dependency with respect to the considered forecast ranges, as illustrated for wind in Figure 5. The diagrams show information about bias, since their slope is linked to the frequencies at which model forecasts are above or below the observed values. Both AROME systems have similar biases, as indicated by the asymmetric appearance of the diagrams. Both AROME-based systems are too cold, too dry, and the wind is too fast. The U shape of the AROME-EPS and PEARP rank diagrams indicates that both ensembles lack spread. This problem is stronger in PEARP than in AROME-EPS for T2m, RH2m and rr3, and the differences are statistically significant in terms of the spread-to-skill ratio (table 3). The main weakness of AROME-EPS is its lack of low-level wind spread. The ensemble spread in the upper atmosphere can be assessed using aircraft observations of wind and temperature; this data is abundant during daytime and reliable at cruise levels, near 250hPa. Both PEARP and AROME-EPS have correct spread at these levels beyond the 9-hour range, according to the ensemble standard deviations (table 3). The PEARP ensemble uses singular vectors to enhance spread at 18-hour range in the AROME area (Descamps *et al.* , 2014). In the upper troposphere, AROME-EPS is strongly constrained by PEARP through its lateral and upper boundary conditions (LBCs), which explains why its spread is similar to PEARP there. At low levels,

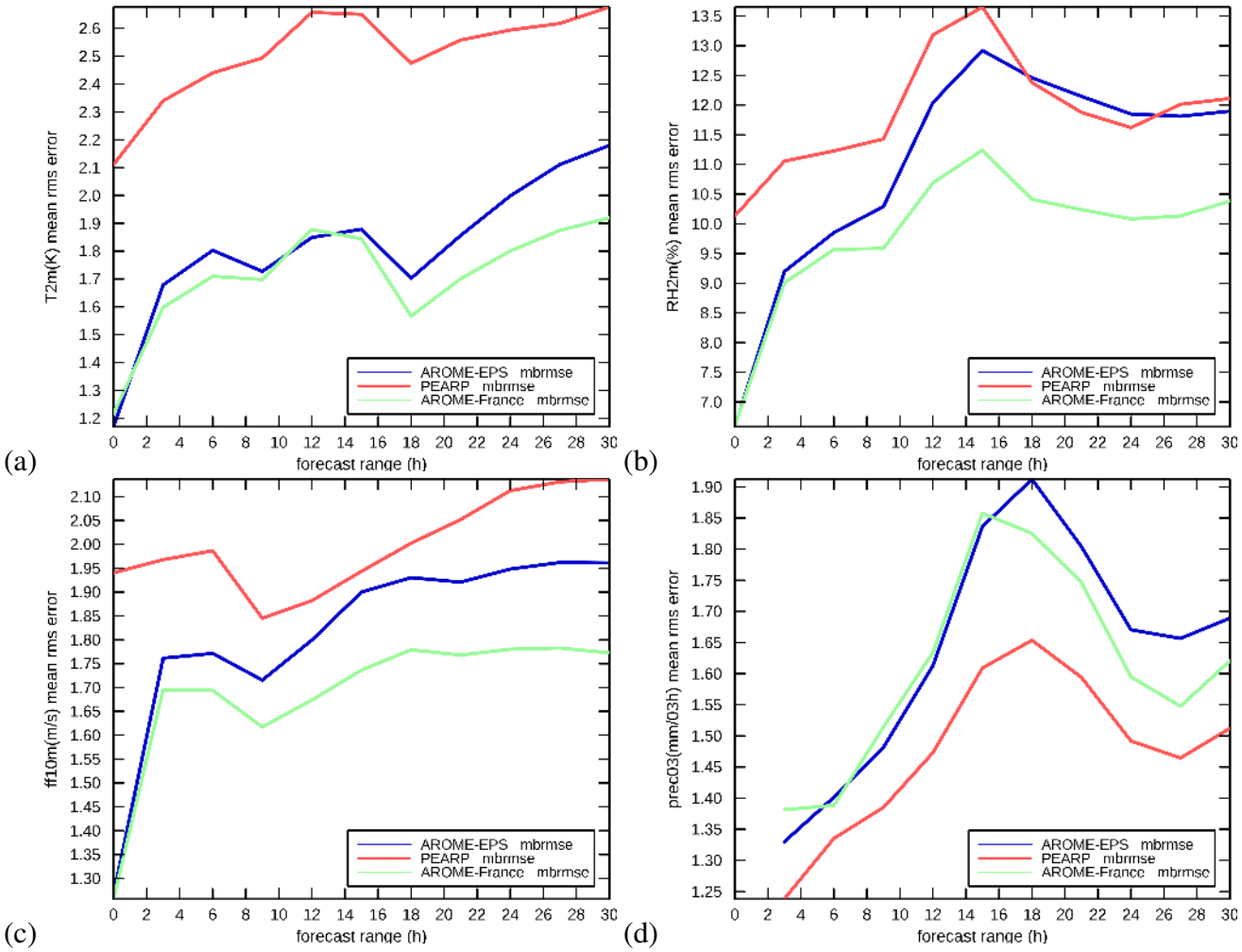

Figure 2: Ensemble means of the member rms forecast error averaged over 75 days, as a function of forecast range. The parameters are (a) T2m, (b) RH2m, (c) ff10m, (d) rr3. There is one curve for each forecasting system: AROME-EPS, PEARP, AROME-France.

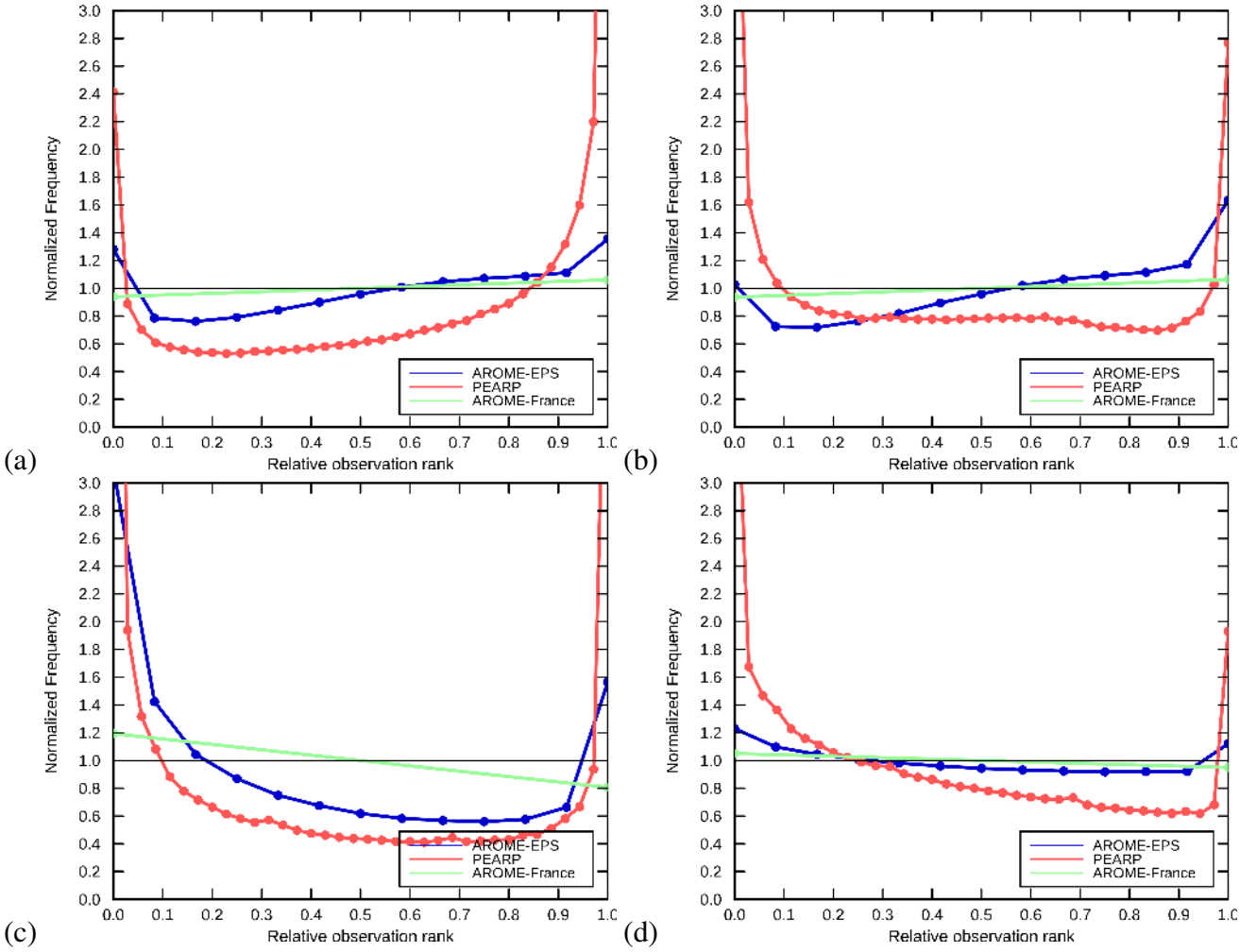

Figure 3: Rank diagrams averaged over 75 days and over forecast ranges 9 to 30 hours. The parameters are (a) T2m, (b) RH2m, (c) ff10m, (d) rr3. There is one curve for each forecasting system: AROME-EPS, PEARP, AROME-France. The dots represent the possible ranks. The abscissa is the rank relative to the maximum possible value for each ensemble, which is one plus the ensemble size N. The ordinate is the observation frequency relative to the ideal value for a perfectly reliable ensemble, which is $1/(N+1)$ (represented by a thin horizontal black line). The left and right endpoint values for PEARP are, respectively: $(2.41, 7.25)$ for T2m, $(4.97, 2.77)$ for RH2m, $(2.77, 1.44)$ for ff10m, $(4.35, 2.16)$ for rr3.

Table 3: Spread/skill ensemble ratios for various experiments, averaged between range 9 and 36 hours over the 75 days. Observation errors are taken into account. There is no change in the ranking of the values with respect to forecast range.

experiment	T2m	RH2m	ff10m	rr3	U250	T250
PEARP	0.50	0.61	0.42	0.58	0.85	0.91
REF	0.87	0.84	0.64	0.97	1.01	0.98
SURF-NONE	0.80	0.74	0.62	0.94	1.01	0.98
SURF-TSWG	0.87	0.84	0.64	0.97	1.01	0.98

the wind profile is more dependent on local influences and model physics, so that AROME model errors are probably responsible for the lack of low-level wind spread.

The probabilistic performance of the ensembles is investigated using the ROC diagrams in Figure 4 (RH2m is not shown because it is similar to T2m). The diagrams are averaged over forecast ranges 9 to 30 hours. They have no significant dependency with respect to the considered forecast ranges, as illustrated for wind in Figure 5. The AROME-EPS skill appears to be better than both AROME-France and PEARP. This would remain true even if the same number of points had been chosen to construct the AROME-EPS and PEARP diagrams. For all four parameters displayed, the AROME-France ROC curve is triangular because it is produced by a deterministic system. Its vertex approximately lies on the AROME-EPS ROC curve, which indicates that both systems have the same quality at a particular operating level. The concavity of the AROME-EPS curve allows it to encompass a larger area than AROME-France, thanks to the ensemble spread. The PEARP curve is even smoother because it is based on more members, but it defines a smaller ROC area for two reasons. First, PEARP lacks model resolution, which limits member skill and lowers the ROC curve. Second, it lacks spread, which causes its ROC curve points to be closer together than the AROME-EPS ones. The PEARP and AROME-France ROC curves intersect each other, which means that a low resolution ensemble can provide better information than a high-resolution deterministic system (given prior knowledge of the optimal operating level). A high-resolution ensemble beats both systems despite its relatively small ensemble size. These findings confirm the result of Clark *et al.* (2009) in an independent framework.

The visual interpretation of the ROC diagrams is confirmed by the ROC area scores, according to which the superiority of AROME-EPS over both AROME-France and PEARP is significant at the 95% level, at all forecast ranges between 9 and 30 hours. This result is robust with respect to changes to the event thresholds (within values that can be reliably tested in the available dataset). The Brier and CRPS scores (not shown) lead to the same conclusions. An exception is weak precipitation, for which AROME-EPS is not clearly better than PEARP for rr3 thresholds less than 3mm. Both systems mainly differ in the success rates of a user who forecasts rain if any ensemble member predicts it, i.e. at the lowest non-trivial ROC operating level. Graphically, this fact is illustrated by the ROC curve points clustering on the left before jumping to the top right point, so that the position of the second highest ROC point has a very strong impact on the ROC area metric. PEARP often predicts widespread weak rain areas, which gives it good detection rate, whereas AROME tends to predict smaller, more intense rain cells. This problem is a form of the double penalty problem mentioned above. Wind speed exhibits this behaviour as well. It could probably be mitigated by dressing the AROME-EPS output with probability density functions that blur small-scale detail (e.g. Hamill and Colucci, 1997; Clark *et al.* , 2011), or by employing verification methods with a built-in spatial tolerance (Ebert, 2008). Unfortunately such methods raise other questions, such as the objectivity of the parameter tuning that they involve. At high precipitation thresholds (Figure

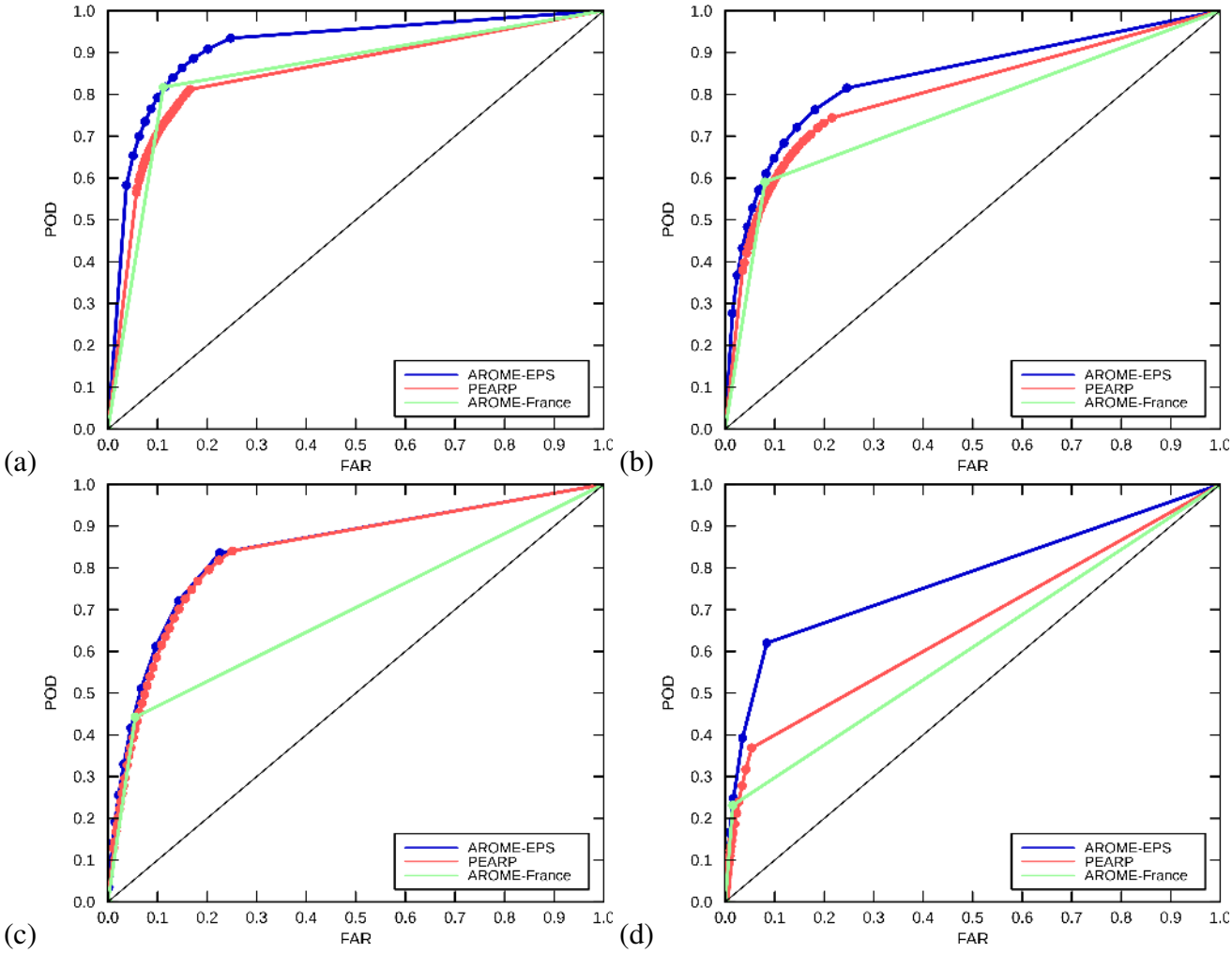

Figure 4: ROC diagrams averaged over 75 days and over forecast ranges 9 to 30 hours. The events considered are (a) T2m>288K, (b) ff10m>5.56m/s, (c) rr3>1mm, (d) rr3>10mm. The abscissa is the false alarm rate (FAR), the ordinate is the probability of detection (POD). There is one curve for each forecasting system: AROME-EPS, PEARP, AROME-France. The thin black line is the $y = x$ diagonal, which is the ROC curve of a system with no skill.

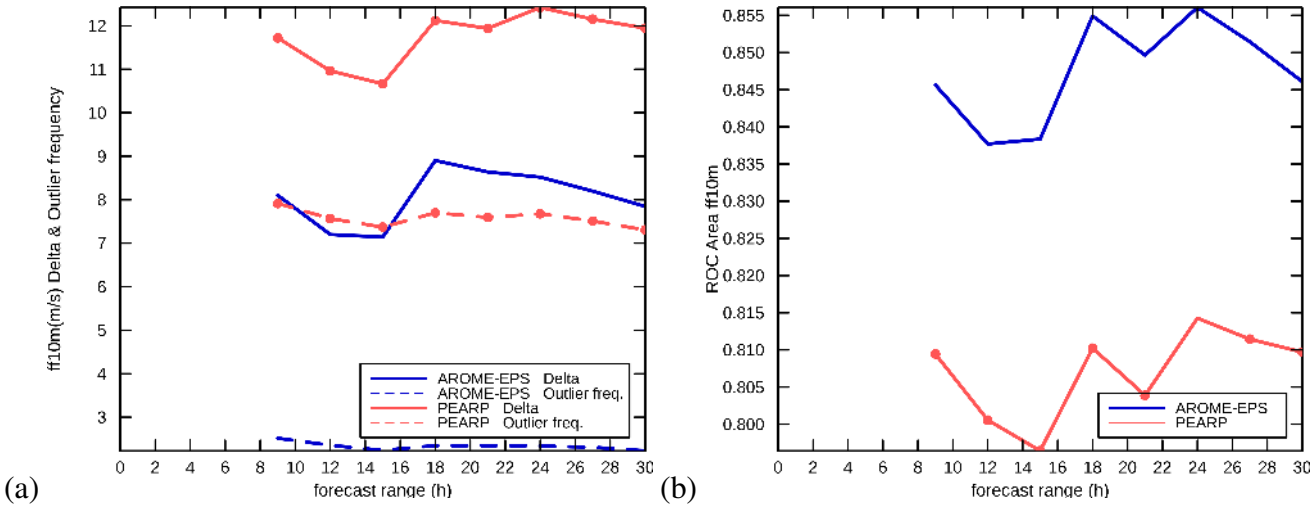

Figure 5: Dependency of the rank and ROC diagrams with respect to forecast range: (a) Delta score and outlier frequency derived from the rank diagram, (b) ROC area for the event ff10m>5.56m/s. There is one curve for each ensemble system: AROME-EPS and PEARP. The values are normalized so that their ideal value is 1.

4d), AROME-EPS is significantly superior to PEARP according to all scores. It shows that, even if a treatment of the double penalty effect could further improve the AROME-EPS scores, such a treatment is not really necessary because the benefit of high resolution is readily visible in the raw ensemble scores.

In conclusion, AROME-EPS produces better forecasts than both its deterministic counterpart AROME-France and the lower-resolution PEARP ensemble, although AROME-EPS has fewer members and is affected by the double penalty problem. AROME-EPS has correct spread in the upper atmosphere, its low-level and precipitation parameters lack spread, and this lack of spread is worst at shorter ranges. These features are shared by the AROME-EPS and PEARP ensembles. AROME-EPS improves a lot upon the deficiencies of PEARP, in particular its low-level spread is much better. Improvements to PEARP would probably be very beneficial to AROME-EPS, because of the strong coupling between both systems. The following sections will demonstrate that some weaknesses of AROME-EPS can be more directly alleviated using surface and initial perturbations.

4 Sensitivity to surface perturbations

4.1 The direct surface perturbation scheme

From here on, the attention (in terms of experiments and verification) will be only focused on AROME-EPS. The direct surface perturbation scheme, used in experiment REF, adds random perturbations to parameters of the SURFEX surface scheme. The perturbations are randomly generated for each member, starting date, and surface field. They are kept constant during each 36-hour forecast, except for prognostic variables (soil temperature and humidity, snow depth) which are freely evolving. Ideally, the surface perturbations should be time dependent, because the corresponding errors may evolve over time. Using static perturbations seems appropriate because only short ranges are considered here. Each surface parameter to perturb is associated to a perturbation pattern generated by filling the model surface grid with

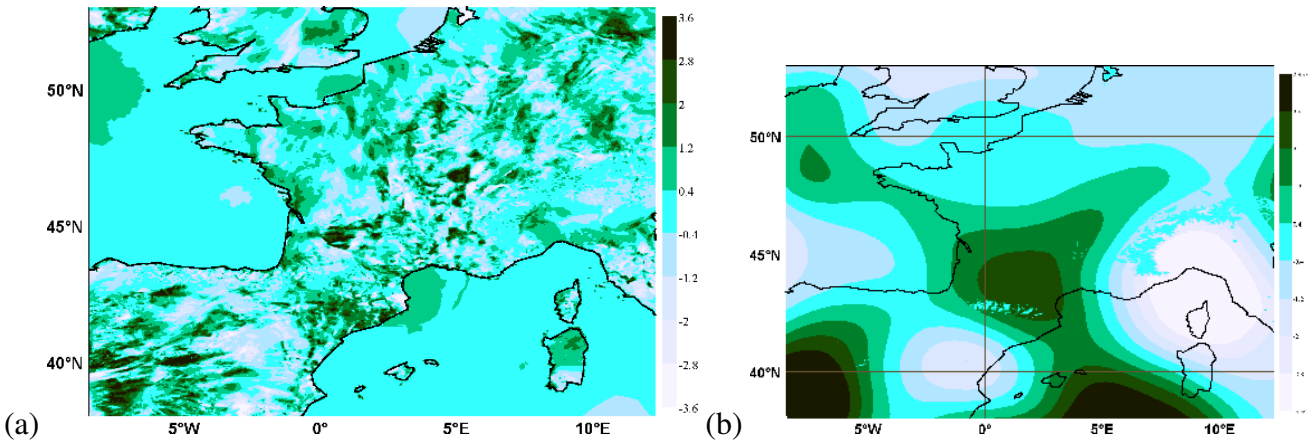

Figure 6: Surface temperature perturbations (K) with respect to the AROME-France analysis on 1 Nov 2012, 0000 UTC. (a) INI-EDA experiment, (b) REF experiment.

Table 4: Parameters affected by the reference direct surface perturbation scheme, with their standard deviation (std.dev) and the perturbation type.

parameter name	std.dev	perturbation type
vegetation index	0.1	multiplicative
vegetation heat coefficient	0.1	multiplicative
leaf area index	0.2	multiplicative
land albedo (all wavelengths)	0.1	multiplicative
land roughness length	0.2	multiplicative
soil/sea surface temperature	1.5K	additive
soil moisture	0.1	multiplicative
snow depth	0.5	multiplicative
sea surface fluxes	0.2	multiplicative

white noise (independent random numbers uniformly distributed between 0 and 1). This noise is spatially smoothed by repeated application of recursive spatial low-pass filters in both grid directions, until a pre-defined correlation length-scale of about 400km is achieved. The filtered pattern exhibits approximately two-dimensional isotropic Gaussian auto-correlations. Each point has a fairly Gaussian distribution of values with spatially constant mean and standard deviation. An optimisation of the auto-correlation length scale could produce better results; it will be attempted in a future study.

The above patterns are re-scaled and clipped with spatially constant values that are tuned for each surface parameter. The tuning is such that the perturbation standard deviations are roughly consistent with the precision at which surface parameters are known. The clipping values are designed to constrain the perturbed surface fields within physically sensible values. Depending on the parameter, the perturbation is applied either as a multiplicative or additive pattern. A multiplicative perturbation by pattern α means that each field value x is replaced by $x(1+\alpha)$, an additive perturbation means that x is replaced by $x+\alpha$. The procedure is applied to the operational AROME-France surface analysis at the starting time of the corresponding AROME-EPS run. The standard deviations and clipping values of the reference configuration are documented in Table 4. An example of perturbation pattern is shown in Figure 6.

The surface fields that are perturbed have been selected according to their physical importance in

determining the surface fluxes, taking into account the known uncertainty in their modelling. Over soil and vegetated surfaces, the vegetation index, heat coefficient, leaf area index, soil moisture and temperature have a recognised influence on the diurnal cycles of low-level temperature and humidity, mainly through their influence on the partitioning of the surface energy budget between latent and sensible heat fluxes. Soil moisture and temperature perturbations are applied to all prognostic soil layers of the SURFEX model in the ISBA soil/vegetation module (Le Moigne *et al.* 2012). Town and lake fields are not perturbed.

The land albedo variability can influence the amount of daytime radiation that enters the surface energy budget. Land roughness length may have an influence on all surface fluxes. Over sea, surface temperature (SST) has a well known impact on sensible and latent heat fluxes; only ice-free points are perturbed. It is difficult to directly perturb the sea roughness length without creating numerical problems, because it is interactively coupled with the sea state and wind speed by complex closure assumptions. Here, a single multiplicative perturbation pattern is applied to sea surface fluxes of sensible heat, latent heat, and momentum at all model time-steps. Snow cover is perturbed in terms of the snow depth only. Perturbing the snow cover extent would make sense, but it is a complex problem: in order to generate a physically perturbed snow cover field, one should take into account the horizontal shape of the field itself as well as other parameters such as orography and urbanisation.

Our choice of surface parameters is supported by the existing literature, in particular Barthlott *et al.* (2011) performed model sensitivity studies. Lavaysse *et al.* (2013) studied the impact of various surface perturbations in a large-scale EPS system. The present study is the first in our knowledge that studies such a comprehensive surface perturbation scheme in a convective-scale ensemble.

4.2 Impact study

A twin experiment measures the impact of the whole surface perturbation scheme by comparing two 75-day sets of AROME-EPS runs: REF is the reference AROME-EPS with the full direct surface perturbation scheme. The perturbed experiment SURF-NONE is identical to REF, but the surface is not perturbed. The ensemble spread statistics are summarised in table 3, and the ROC areas differences are shown in Figure 7. Like most ensemble perturbation schemes, the surface perturbations increase the rms errors of the ensemble members. They improve the probabilistic scores of T2m, RH2m, and, to a lesser extent, ff10m. The improvement is clear when looking at the rank, ROC and reliability diagrams. They are statistically significant at all ranges at the 95% level, in terms of the delta, CRPS, Brier and ROC area scores (not shown). The impact on precipitation scores is weak, with some non significant improvement of spread, outlier frequency, and Brier for light precipitation. Perhaps this effect would be more apparent in the Spring or Summer seasons, because the HyMeX-SOP1 period did not exhibit much deep convection over the plains. Mediterranean convective precipitation events are known to be sensitive to sea surface fluxes (Lebeaupin-Brossier, 2008), but in our experiment this mechanism does not have a conspicuous effect.

A second experiment measures the impact of restricting the direct surface perturbation scheme to SST, soil temperature and humidity. The experiment is called SURF-TSWG, its impact on spread is summarised in table 3. The corresponding ROC curve areas are shown in Figure 7. The restriction of perturbations significantly reduces the spread that the full scheme brings with respect to SURF-NONE, but over 90% of that spread is retained. The corresponding change to the ensemble mean rms errors is such that the spread-to-skill ratio is nearly identical to REF. The probabilistic scores such as the ROC

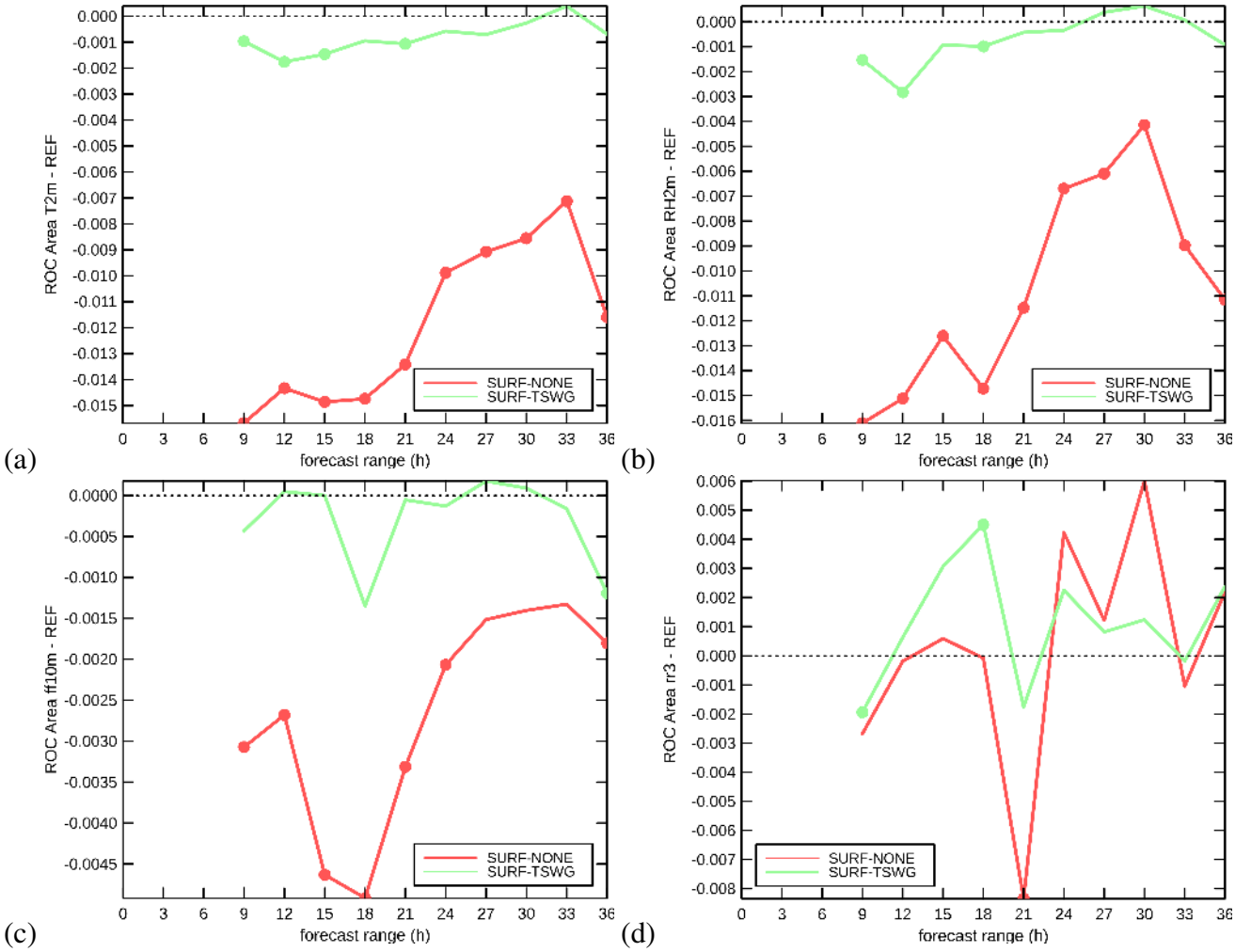

Figure 7: ROC area differences with respect to experiment REF, averaged over 75 days, as a function of forecast range. The events considered are (a) T2m>288K, (b) RH2m>80%, (c) ff10m>5.56m/s, rr3>3mm. The curve are for experiments SURF-NONE and SURF-TSWG. The black horizontal dotted line indicates the $y = 0$ level; data points above this line indicate a performance better than REF, a bullet is plotted where the sign of the deviation from REF is statistically significant.

area are degraded with respect to REF. This degradation is statistically significant, but much smaller than the difference between REF and SURF-NONE. In other words, the main components of the surface perturbation scheme are the SST, soil temperature and humidity, but perturbations to the other parameters do improve the ensemble as well. It has been checked in other, shorter experiments (not shown) that all listed parameters are useful to the surface perturbation scheme, and that the perturbation amplitudes used are approximately optimal.

In conclusion, direct surface perturbations provide a simple but effective way of improving the low-level spread and the probabilistic scores of the ensemble. The impact of the perturbation scheme is strong on temperature and humidity, smaller on wind speed, and almost negligible on precipitation. A better perturbation strategy, not necessarily limited to the surface, is needed for wind and precipitation. Wind speed spread could probably be improved by perturbing the boundary layer turbulence process, because this problem seems localised in the lowest model levels. A more physical approach than SPPT could be implemented, by perturbing the TKE or the mixing length (Hally *et al.* , 2014; Wang *et al.* , 2012). The strategy for increasing the precipitation spread is less clear. Perturbations of the cloud microphysics scheme can yield disappointing results (Hally *et al.* , 2014). One could directly perturb the humidity field, or the intensity of the horizontal diffusion scheme, which has a large impact on the dynamics of convective clouds resolved by the model (Seity *et al.* , 2011).

5 Sensitivity to initial perturbations

5.1 The reference initial perturbation scheme

The initial condition (IC) perturbation scheme used in the reference (REF) experiment adds centred PEARP perturbations to the AROME-France operational analysis. It acts on the atmospheric part of the model:

$$x_i = x_a + m(z_j - \overline{z})$$

where x_i and z_j are the IC field values of AROME-EPS and PEARP members i and j, respectively, x_a is the AROME-France operational analysis, $\overline{z}$ is the mean of the z_j used to prepare the AROME-EPS run. Index j is the PEARP member used to provide LBCs to AROME-EPS member i. Operator m interpolates PEARP perturbations to the AROME model grid, and re-scales them by applying a vertical amplitude modulation. The modulation is such that the PEARP perturbations are untouched in the mid troposphere and lower stratosphere, and they smoothly go to zero near the model surface and top, in order to avoid numerical and physical problems. This formulation avoids most incompatibilities between the PEARP and AROME models, while preserving the large-scale upper-level dynamics that are important for the PEARP singular vectors. The IC perturbation scheme is applied to wind, temperature, and surface pressure. Humidity perturbations are not applied because they have been found to have a detrimental impact on ensemble performance, perhaps because of inconsistencies between the cloud physics of the models.

The perturbations are added to the AROME operational analysis, so that the ensemble uses the latest fine-scale and low-level information available on the AROME grid through the AROME-France analysis. The resulting IC perturbations have an amplitude that is consistent with the AROME analysis error statistics, although their correlation structures probably have a much larger scale, because they stem from a lower resolution system.

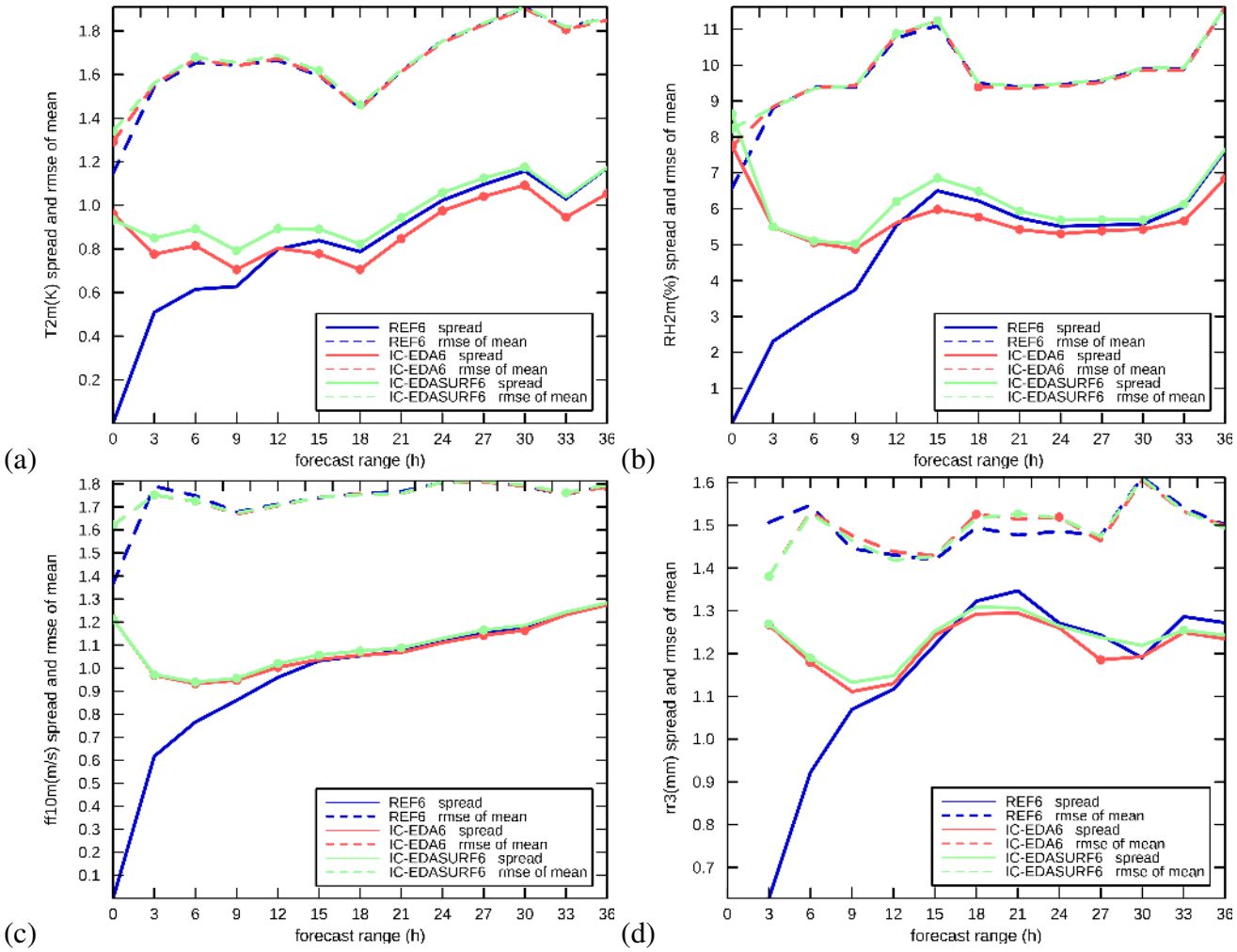

Figure 8: Ensemble spread (solid curves) measured by its standard deviation averaged from 12 Oct. to 18 Nov. 2012, as a function of forecast range, for (a) T2m, (b) RH2m, (c) ff10m and (d) rr3. The dashed lines indicate the rms error of the ensemble mean (rmse of mean). There is one curve for each AROME-EPS experiment: REF6, IC-EDA6, IC-EDASURF6. The bullets indicate the data points for which the apparent departure from REF6 is statistically significant.

Several authors have shown that short-range ensemble performance is much degraded if there is no IC perturbation. A demonstration with AROME-EPS is provided in Vié *et al.* (2011) with an assimilation ensemble. It has been checked that the reference IC scheme used here is much better than no perturbation. Its drawback is that the reference AROME-EPS lacks spread at short ranges, as can be seen on Figure 8 Indeed, subjective evaluation by forecasters has confirmed that the AROME-EPS REF members are too close to each other to provide a useful diversity of weather predictions at ranges less than 9 hours. In the present section, improvements to the reference IC scheme are sought.

5.2 Importance of IC vs LBC consistency

An attractive feature of the IC scheme is its close consistency with PEARP. The j index above selects the PEARP representative member used for AROME-EPS member i according the procedure explained in section 2. Thus, the same PEARP member is used to compute the IC perturbation and LBC forcing of

each AROME-EPS member. The other IC schemes tested below do not exhibit this kind of consistency, so it is worth checking its practical significance. In an experiment called IC-SHUFFLE, the j index of the formula in section 5.1 is shuffled so that no AROME-EPS member uses the same PEARP member for its initial and lateral boundary conditions. The set of selected PEARP members remains the same, so that the IC-SHUFFLE performance is not affected by the quality of the PEARP members used (PEARP uses a multi-physics perturbation scheme, so that model biases and rms errors are different from one member to the next). Thus, experiment IC-SHUFFLE only differs from REF by the loss of consistency between the initial and lateral boundary conditions. The scores (not shown) indicate that there is no measurable impact on the ensemble performance. The shuffling causes a tiny, statistically significant reduction of spread by about 0.05% of the ensemble standard deviation, but no change to the probabilistic scores.

5.3 Comparison with an ensemble data assimilation

In an experiment called IC-EDA6, the reference IC perturbation scheme is replaced by an initialisation by the atmospheric and surface analyses of an AROME ensemble data assimilation system (EDA). The EDA algorithm is described in Brousseau *et al.* (2012) and Vié *et al.* (2011). Its principle is to run an ensemble of independently perturbed data assimilation systems. The result is an ensemble of analyses that samples the uncertainties on the initial state of the AROME ensemble prediction system. The EDA algorithm tested here is a six-member 3D-Var AROME data assimilation system at the same model resolution as AROME-EPS. The EDA perturbations stem from adding random Gaussian noise to the observations in six instances of the AROME-France 3D-Var assimilation, including the surface analyses. This EDA algorithm is similar to the one used in Bouttier *et al.* (2012), with recent improvements to the adaptive ensemble inflation algorithm (Raynaud *et al.* , 2012). EDA is a computationally expensive system, so that experiment IC-EDA6 only contains six members from 12 October to 18 November 2012 (36 days). In order to avoid verification biases linked to varying ensemble size, IC-EDA6 scores will be compared to REF6, which comprises the six corresponding members of the REF experiment; these members use the same lateral boundaries, SPPT and surface perturbations patterns.

The EDA system is related to PEARP in the sense that they both use the same global ensemble data assimilation scheme, AEARP. EDA relies on AEARP to provide its large scale boundary conditions, and PEARP initial conditions use a combination of singular vectors and AEARP perturbations. In IC-EDA6, the ICs and LBCs do not necessarily use the same AEARP members. It has been shown in section 5.2, with a similar framework, that this kind of inconsistency has a negligible impact on the ensemble performance.

The spread and the ROC area scores are shown in Figure 8 and 9, respectively. Short range ensemble spread is larger with EDA than with the reference IC scheme. The spread at range zero is excessive, short lived, and arguably nonphysical: it is caused by a strong inflation of perturbations in the EDA system. After a spin-up period of the order of one hour, the ensemble spread settles to more stable, slowly increasing levels. Figure 8 shows that the IC-EDA6 ensemble spread-to-skill ratio is much closer to one than the reference for ranges between 3 and about 12 hours. At longer ranges, the spread becomes smaller than the reference. The variations of spread are closely related to ensemble performance, leading to significantly improved rank diagrams, CRPS, Brier and ROC area scores until the 6- to 9-hour range. The improvement is particularly spectacular for precipitation: its reliability and ROC diagrams are much improved at all available thresholds. At longer ranges, the spread becomes smaller in IC-EDA6 than in REF6, and the IC-EDA6 scores become neutral or slightly worse than REF6.

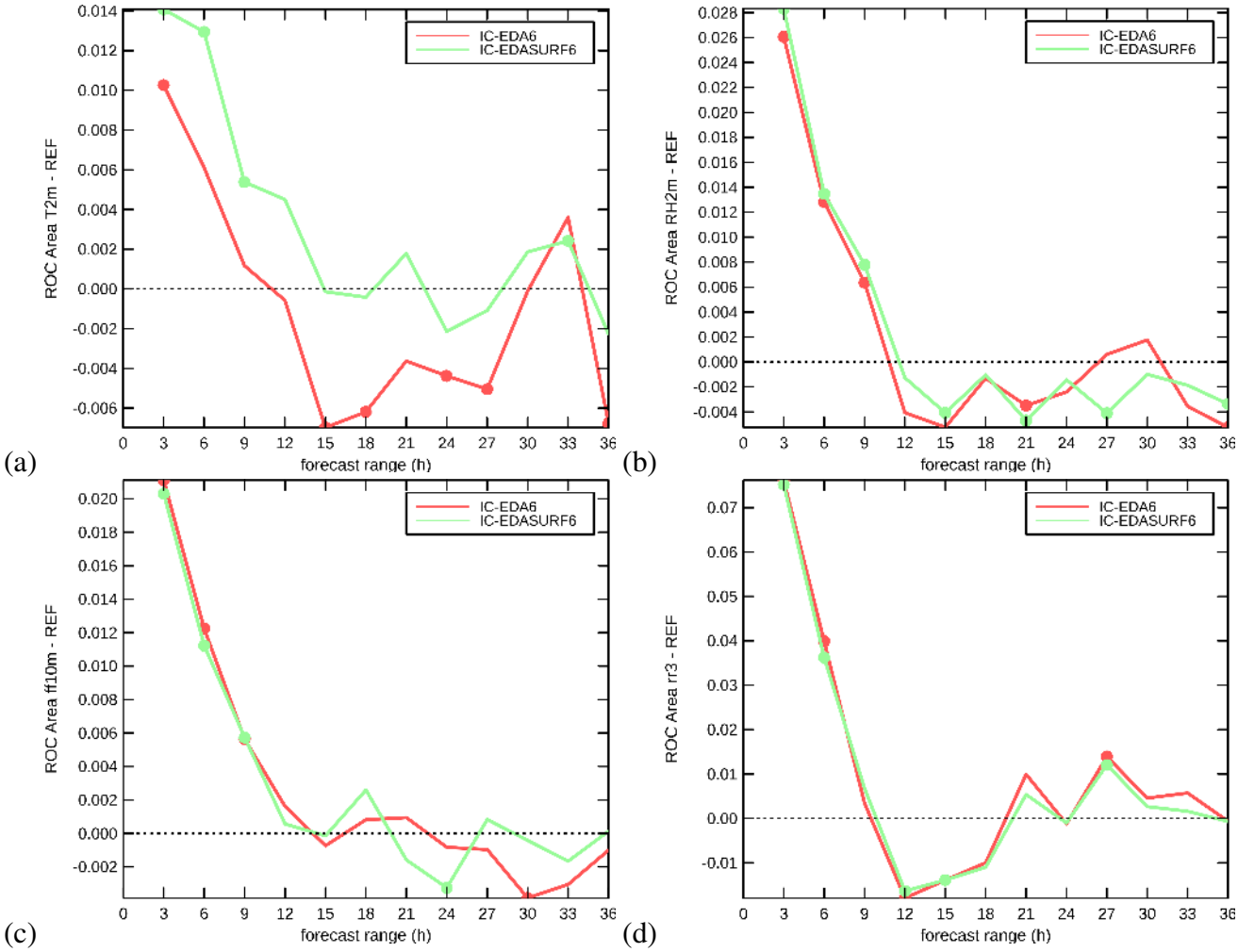

Figure 9: ROC area differences as in Figure 7, this time for experiments IC-EDA6 and IC-EDASURF6, with respect to experiment REF6. The scores are averaged over 12 Oct. to 18 Nov. 2012.

5.4 Interaction between initial and surface perturbations

The IC-EDA6 experiment has shown that ensemble data assimilation improves the short term ensemble scores, but degrades them at longer ranges. A likely explanation is that EDA provides good atmospheric perturbations, but they tend to be advected from the model domain after a few hours, because the AROME only covers a relatively small region. At longer ranges, we hypothesize that the impact of EDA is dominated by slowly-evolving EDA surface perturbations, which are not as good as in REF. To test this hypothesis, an experiment called IC-EDASURF6 has been set up like IC-EDA6, except that the surface conditions are replaced by those from REF6. As expected, the result (Figures 8 and 9) combines the short term benefit of EDA with the longer term impact of the direct surface perturbation scheme. In terms of spread, IC-EDASURF6 is equivalent to or better than REF6 and IC-EDA6 at all ranges. Wind speed is improved for up to 21 hours. In terms of ROC area, IC-EDA6 is clearly better than REF6 up to approximately 12 hours, and then more or less neutral. The improvement of IC-EDASURF6 over IC-EDA6 is clear for T2m at most ranges, limited to short ranges for RH2m, and not significant for ff10m and rr3.

The surface analyses are perturbed in EDA by noise added to observations in the assimilations of SST, soil temperature and soil moisture. The EDA land surface fields are also influenced by atmospheric perturbations that affect surface fluxes in the forecast steps of the data assimilation. An example of EDA perturbation is shown in Figure 6, which shows that the direct surface perturbations tend to have larger amplitudes, and perhaps more effective horizontal structures than EDA. The EDA soil perturbations have the physical attractiveness of consistency with the atmospheric perturbations. This theoretical advantage does not translate into competitive atmospheric scores, but the consistency between soil and atmosphere that is built in the EDA technique could be important for hydrological applications. For instance, flood severity can depend on both precipitation intensity and soil moisture content, which depends on past precipitation. The direct surface perturbation scheme, being independent from the atmospheric state, does not take these interactions into account, whereas EDA does to a certain extent. The conclusion is that, although the IC-EDASURF6 setup is the best according to our probabilistic scores, there may be some value in using the EDA-produced surface fields in a more elaborate perturbation scheme.

Experiment SURF-EDA6 tests a setup that is identical to REF6, except that the surface fields (only) are taken from EDA. The intention is to compare the EDA and direct surface perturbations, without any change to the atmospheric perturbations. The result (not shown) is that the EDA surface is not as good as the reference scheme, but when compared with SURF-NONE it improves the scores in a way that is similar, although weaker, than the direct surface perturbations. It suggests that EDA surface perturbations contain valuable information that should ideally be combined with the direct scheme, in order to deliver even better ensemble forecasts. For instance, one could think of combining both sets of field perturbations with a constraint on the total perturbation amplitude. One could even build a surface-specific perturbation and inflation scheme inside the EDA algorithm. Since optimised surface perturbations improve short-range ensemble forecasts, they could be beneficial to EDA-based ensemble data assimilation schemes as well, by improving the background error covariances (Brousseau *et al.* , 2011).

6 Discussion and conclusions

The AROME-EPS ensemble prediction system has been tested over 75 days during the HyMeX-SOP1 field experiment. Observations from the field experiment have been used to improve the sampling when verifying the ensemble prediction.

The comparison between AROME-EPS, the global PEARP ensemble prediction system, and the AROME-France high-resolution deterministic system, show that AROME-EPS combines the advantages of a high resolution model with those of an ensemble approach, despite its modest ensemble size. At upper atmospheric levels, AROME-EPS is closely coupled with PEARP, and both systems have correct spread. At low levels, AROME-EPS spread is slightly less than optimal. Its probabilistic forecasts are superior to both other systems, particularly for temperature, humidity, and strong precipitation, but not for weak precipitation. The AROME-EPS probabilities of low-level wind speed are hampered by their lack of spread. In summary, the comparison proves that a high resolution ensemble with few members can outperform a lower resolution ensemble with many more members.

A direct surface perturbation scheme is presented that significantly contributes to convective-scale ensemble performance for two-metre temperature, humidity, and (to a lesser extent) wind. The most important fields to perturb are SST, soil moisture and temperature. A comparison with surface perturbations from an ensemble data assimilation scheme shows that the direct surface perturbation scheme performs better than the more sophisticated EDA-based one. Consistency between surface and atmosphere only has a minor impact on the forecast quality. The best ensemble is obtained by combining EDA atmospheric perturbations with the direct surface perturbation scheme. The results suggest that ensemble performance could be improved by a better handling of surface perturbations in the EDA algorithm.

Experimentation with the initial condition setup reveals that initial and lateral boundary conditions do not need to be derived from the same large-scale forecast. One can initialise the AROME-EPS with an ensemble data assimilation (EDA) system that is not consistent with the lateral boundary conditions, and obtain a large improvement in ensemble performance with respect to a simpler initial condition scheme. Using 6 members and a shorter period, the EDA-related improvement is shown to be strong between the 3 and 9-hour forecast ranges. At shorter ranges, there is a detrimental spin-down of the ensemble spread. At longer ranges, the benefit of using EDA vanishes unless the EDA surface perturbations are replaced by a simpler, but more effective, surface perturbation scheme. With this setup, the EDA-based AROME-EPS system beats the reference configuration up to the 12-hour range, and even longer for some parameters. The conclusion is that, despite its high numerical costs, an EDA approach is very effective for short range ensemble performance.

These results shed some light on the interaction between various types of ensemble perturbations. In existing ensemble development studies, the representations of various perturbation sources are often treated independently from one another, which raises questions about their mutual consistency. It is shown here that in a reasonable ensemble framework the consistency between surface, initial and lateral boundary perturbations has no impact in terms of meteorological forecast performance. Nevertheless, the consistency between surface and atmospheric perturbations may be important for hydrological applications, a point that needs to be verified by future studies.

This study has highlighted the benefits of initialising a convective scale, short range ensemble with EDA. The main problem with EDA is its high numerical cost. More cost effective approaches will be tried in the future by coupling AROME-EPS with EDAs at lower resolutions, and with less members than the forecast ensemble.

The AROME-EPS performance is not the same for all weather parameters. Wind speed spread is correct at high levels, too weak at low levels, and wind speed is positively biased against observations. These facts point to weaknesses in the ten-metre wind modelling. Possible solutions include the development of a model error representation in the turbulence scheme, and an increase of vertical model resolution, so that the ten-metre wind can be interpolated directly from the AROME 3-D model grid, rather than

diagnosed though an ad hoc one-dimensional boundary layer model (see Seity *et al.* , 2011).

Precipitation lacks spread in AROME-EPS. This behaviour is not easy to interpret, since the distribution of precipitation intensities is far from symmetric. Lack of spread may be due to non-intuitive reasons, such as an under-prediction of light rain, because most forecast values are equal to zero. Some authors have reported that rain intensity biases can be alleviated using model bias correction and/or ensemble calibration (e.g. Hamill and Colucci, 1997; Ben Bouallègue, 2013). Insufficient precipitation spread may also be caused by a lack of diversity in the forecast precipitation patterns, even if the model itself simulates rain with a perfect climatology. The upper-level model dynamics of AROME-EPS are correctly dispersed according to flight level temperatures and winds, so that the lack of precipitation spread is probably linked to missing perturbations in physical processes related to humidity or convection. One could try to improve this situation by perturbing the turbulence parametrisation, the humidity field, and model dynamics, for instance using the horizontal diffusion operator.

A key objective of AROME-EPS is the prediction of high impact weather events. The HyMeX-SOP1 dataset used here over 75 days was helpful in sampling the overall ensemble performance, but it did not include many intense winds or rain. The results are believed to be an adequate assessment of ensemble performance for average autumn weather in the HyMeX area, but they may not be informative about extreme weather. For instance, the highest rain threshold for which ensemble reliability diagrams can be convincingly constructed here is about 20mm in 3 hours. Catastrophic flooding events in the area are often triggered over small areas by at least twice higher accumulations. Such events are detected by high-resolution ensembles with subjectively good accuracy in published case studies (e.g. Vié *et al.* , 2011; Nuissier *et al.* , 2012; Descamps *et al.* , 2014), but their rarity makes it impossible to objectively measure the ensemble performance over contiguous test periods such as in this paper. The verification of high wind events has the same issues. The study of ensemble performance with respect to high impact weather would necessitate a sampling strategy that is different from the one adopted here. Ideally, one would like to sample at least dozens of extreme events spread out over several years. Given the high numerical costs of convective-scale ensembles, such experiments should only run during relevant weather situations, the selection of which inevitably raises sampling bias issues. This will be the topic of future work.

7 Acronyms

AEARP ensemble data assimilation with the ARPEGE model.

ALADIN an hydrostatic limited-area numerical weather prediction model.

AROME a non-hydrostatic limited-area numerical weather prediction model.

ARPEGE a global numerical weather prediction model used at Météo-France.

COSMO-DE-EPS an ensemble prediction system based on the COSMO model used at the German national weather service.

EDA ensemble data assimilation

HARMONIE a generic name for several numerical weather prediction models used in the HIRLAM group, among which the AROME model.

HIRLAM a group of European national weather services.

HyMeX a field experiment dedicated to hydrology in the Mediterranean area.

HyMeX-SOP1 a special observing period of HyMeX that focused on weather prediction during autumn 2012.

MOGREPS-UKV a non-hydrostatic limited-area numerical weather prediction model used by the UK Met Office.

PEARP a global ensemble prediction system based on the ARPEGE model.

RRTM a radiative transfer parametrisation used in the AROME model.

SPPT a stochastic physics scheme that involves random perturbations of physics tendencies.

SURFEX a surface parametrisation used in the AROME model.

TKE turbulent kinetic energy, a quantitiy used in the parametrisation of subgrid fluxes.

WRF a non-hydrostatic limited-area model system.

Acknowledgements:

This work has been funded by Météo-France and CNRS. It is a contribution to the HyMeX (HYdrological cycle in Mediterranean EXperiment) program of the MISTRALS metaprogram. It received support from the LEFE/MANU/VIPS contract of the Institut National des Sciences de l'Univers. Helpful comments from two anonymous reviewers are acknowledged.

References

[1] Barthlott C, Kalthoff N. 2011. A numerical sensitivity study on the impact of soil moisture on convection-related parameters and convective precipitation over complex terrain. *J. Atmos. Sci.* **68:** 2971–2987. doi:10.1175/jas-d-11-027.1

[2] Ben Bouallègue Z. 2013. Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather Forecast.* **28:** 515–524. doi:10.1175/waf-d-12-00062.1

[3] Berner J, Ha SY, Hacker JP, Fournier A, Snyder C. 2011. Model uncertainty in a mesoscale ensemble prediction system: stochastic versus multiphysics representations. *Mon. Weather Rev.* **139:** 1972–1995. doi:10.1175/2010mwr3595.1

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

[5] Brousseau P, Berre L, Bouttier F, Desroziers G. 2011. Background-error covariances for a convective scale data-assimilation system: Arome-France 3D-Var. *Q. J. R. Meteorol. Soc.* **137:** 409–422. doi:10.1002/qj.750

[6] Buizza R, Miller M, Palmer T. 1999. Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125:** 2887–2908. doi:10.1002/qj.49712556006

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

[8] Candille G, Côté J, Houtekamer PL, Pellerin G. 2007. Verification of an ensemble prediction system against observations. *Q. J. R. Meteorol. Soc.* **135:** 2688–2699. doi:10.1175/mwr3414.1

[9] Clark A, Gallus W, Xue M, Kong F. 2009: A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather Forecast.* **24:** 1121–1140. doi:10.1175/2009WAF2222222.1

[10] Clark A, Kain J, Stensrud D, Xue M, Kong F, Coniglio M, Thomas K, Wang Y, Brewster K, Gao J, Wang X, Weiss S, Du J. 2011. Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Weather Rev.* **139:** 1410–1418. doi:10.1175/2010mwr3624.1

[11] Descamps L, Labadie C, Joly A, Bazile E, Arbogast P, Cébron P. 2014: PEARP, the Météo-France short-range ensemble prediction system. Accepted for publication in *Q. J. R. Meteorol. Soc.* . doi:10.1002/qj.2469

[12] Ducrocq V, Braud I, Davolio S, Ferretti R, Flamant C, Jansa A, Kalthoff N, Richard E, Taupier-Letage I, Ayral PA, Belamari S, Berne A, Borga M, Boudevillain B, Bock O, Boichard JL, Bouin MN, Bousquet O, Bouvier C, Chiggiato J, Cimini D, Corsmeier U, Coppola L, Cocquerez P, Defer E, Delanoë J, Di Girolamo P, Doerenbecher A, Drobinski P, Dufournet Y, Fourrié N, Gourley JJ, Labatut L, Lambert D, Le Coz J, Marzano FS, Molinié G, Montani A, Nord G, Nuret M, Ramage K, Rison W, Roussot O, Said F, Schwarzenboeck A, Testor P, Van Baelen J, Vincendon B, Aran M, Tamayo J. 2014. HyMeX-SOP1: The field campaign dedicated to heavy precipitation and flash flooding in the northwestern Mediterranean. *Bull. Am. Meteorol. Soc.* **95:** 1083–1100. doi:10.1175/bams-d-12-00244.1

[13] Ebert, E. 2008. Fuzzy verification of high resolution gridded forecasts: A review and proposed framework. *Metorol. Appl.* **15:** 51–64. doi:10.1002/met.25

[14] Gebhardt C, Theis SE, Krahe P, Renner V. 2008. Experimental ensemble forecasts of precipitation based on a convection-resolving model. *Atmos. Sci. Lett.* **9:** 67–72. doi:10.1002/asl.177

[15] Gebhardt C, Theis SE, Paulat M, Ben Bouallègue Z. 2011. Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.* **100:** Issues 2-3, May 2011, Pages 168–177. doi:10.1016/j.atmosres.2010.12.008

[16] Hacker J, Ha SY, Snyder C, Berner J, Eckel F, Kuchera E, Pocernich M, Rugg S, Schramm J, Wang X, 2011: The U.S. Air Force Weather Agency's mesoscale ensemble: scientific description and performance results. *Tellus* **63A:** 625–641. doi:10.1175/2010mwr3595.1

[17] Hally A, Richard E, Fresnay S, Lambert D. 2014. Ensemble simulations with perturbed physical parametrizations: Pre-HyMeX case studies. *Q. J. R. Meteorol. Soc.* **140:** 1900–1916. doi:10.1002/qj.2257

[18] Hamill TM, Colucci SJ. 1997. Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125:** 1312–1327. doi:10.1175/1520-0493(1997)125

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

[20] Lavaysse C, Carrera M, Bélair S, Gagnon N, Frenette R, Charron M, Yau MK. 2013. Impact of surface parameter uncertainties within the Canadian regional ensemble prediction system. *Mon. Weather Rev.* **141:** 1506–1526. doi:10.1175/MWR-D-11-00354.1

[21] Lebeaupin-Brossier C, Ducrocq V, Giordani H. 2008. Sensitivity of three Mediterranean heavy rain events to two different sea surface fluxes parametrizations in high-resolution numerical modeling. *J. Geophys. Res.* **113:** D21109. doi:10.1029/2007jd009613

[22] Le Moigne P. 2012. SURFEX scientific documentation. *Available online at* `www.cnrm.meteo.fr/surfex-lab` (accessed 15 Oct. 2014)

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

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

[25] Migliorini S, Dixon M, Bannister R, Ballard S. 2011. Ensemble prediction for nowcasting with a convection-permitting model — I: description of the system and the impact of radar-derived surface precipitation rates. *Tellus* **63A:** 468–496. doi:10.1111/j.1600-0870.2010.00503.x

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

[27] Peralta C, Ben Bouallègue Z, Theis SE, Gebhardt C, Buchhold M. 2012. Accounting for initial condition uncertainties in COSMO-DE-EPS, *J. Geophys. Res.* **117:** D07108, 13pp. doi:10.1029/2011JD016581

[28] Raynaud L., L. Berre, G. Desroziers, 2012: Accounting for model error in the Météo-France ensemble data assimilation system. *Q. J. R. Meteorol. Soc.* **138:** 249–262. doi:10.1002/qj.906

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

[30] Schwartz CS, Kain JS, Weiss SJ, Xue M, Bright DR, Kong F, Thomas KW, Levit JJ, Coniglio MC, Wandishin MS. 2010. Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather Forecast.* **25:** 263–280. doi:10.1175/2009waf2222267.1

[31] Stensrud DJ, Bao JW, Warner TT. 2000. Using initial condition and model physics perturbations in short-range ensemble simulations of mesoscale convective systems. *Mon. Weather Rev.* **128:** 2077–2107. doi:10.1175/1520-0493(2000)128<2077:uicamp>2.0.CO;2

[32] Tennant W, Beare S. 2014. New schemes to perturb sea-surface temperature and soil moisture content in MOGREPS. *Q. J. R. Meteorol. Soc.* **140:** 1150–1160. doi:10.1002/qj.2202

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

[34] Wang Y, Tascu S, Weidle F, Schmeisser K, 2012: Evaluation of the added value of regional ensemble forecasts on global ensemble forecasts. *Weather Forecast.* **27:** 972–987. doi:10.1175/waf-d-11-00102.1

[35] Wei M, Toth Z, Wobus R, Zhu Y, Bishop C, Wang X. 2006. Ensemble Transform Kalman Filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus* **58A:** 28–44. doi:10.1111/j.1600-0870.2006.00159.x