TECHNICAL MEMORANDUM

559

Comparison of a 51-member low-resolution (T_L399L62) ensemble with a 6-member high-resolution (T_L799L91) lagged-forecast ensemble

Roberto Buizza

Research Department

To appear in Mon. Wea.Rev.

March 2008

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

Abstract

The 51-member T_L399L62 ECMWF Ensemble Prediction System (EPS51) is compared with a lagged ensemble system based on the six most recent ECMWF T_L799L91 forecasts (LAG6). The EPS51 and LAG6 systems are compared to two 6-member ensembles with a 'weighted' ensemble-mean, EPS6wEM and LAG6wEM: EPS6wEM includes 6 members of EPS51 and has the ensemble-mean constructed giving optimal weights to its members, while LAG6wEM includes the LAG6 6 members and has the ensemble-mean constructed giving optimal weights to its members. In these weighted ensembles, the optimal weights are based on 50-day forecast error statistics of each individual member (in EPS51 and LAG6 the ensemble mean is constructed giving the same weight to each individual member). The EPS51, LAG6, EPS6wEM and LAG6wEM ensembles are compared for a 7 month period (from 1 April to 30 October 2006, 213 cases) and for two of the most severe storms that hit the Scandinavian countries since 1969.

The study shows that EPS51 has the best-tuned ensemble spread, and provides the best probabilistic forecasts, with differences in predictability between EPS51 and LAG6 or LAG6wEM probabilistic forecasts of geopotential height anomalies of up to 24 hours. In terms of ensemble-mean, EPS51 gives the best forecast from forecast day 4, but before forecast day 4 LAG6wEM provides a slightly better forecast, with differences in predictability smaller than 2 hours up to forecast day 6, and of about 6 hours afterwards. The comparison also shows that a larger ensemble size is more important in the medium-range rather than in the short range.

Overall, these results indicate that if the aim of ensemble prediction is to generate not only a single (most-likely) scenario but also a probabilistic forecast, than the 51-member T_L399L62 ECMWF Ensemble Prediction System (EPS51) has a higher skill than the lagged ensemble system based on the six most recent ECMWF T_L799L91 forecasts (LAG6 or LAG6wEM).

1. The ECMWF forecasting system

Since 1992, ensemble prediction systems have been a component of the operational forecasting suite at many meteorological institutes. The first centers producing global ensembles forecasts were in 1992 ECMWF (*Palmer et al.* 1993, *Molteni et al* 1996) and the U. S. National Center for Environmental Predictions (NCEP, previously NMC, *Tracton & Kalnay* 1993, *Toth & Kalnay* 1997), and in 1995 the Meteorological Service of Canada (MSC, *Houtekamer et al.* 1996). In the past decade, global ensemble prediction systems have been developed and implemented in eight other centers: the US Navy in Monterey (NRL), the Bureau of Meteorology Research Center (BMRC) in Melbourne, the Centro de Previsao de Tempo e Estudos Climatico (CPTEC) in Sao Paulo, the Chinese Meteorological Agency (CMA) in Beijing, the Japanese Meteorological Agency (JMA) in Tokyo, the Korean Meteorological Agency (KMA) in Seoul, and more recently Meteo-France in Toulouse and the UK Met-Office in Exeter (see the WMO CAS/JSC WGNE report of 2005, and *Park et al* 2008 for a very recent comparison of the performance of the global ensemble systems that are taking part in TIGGE, the WMO THORPEX Interactive Grand Global Ensemble project). The operational implementation of these ensemble systems followed the theoretical and experimental work of, among others, *Epstein* (1969), *Gleeson* (1970), *Fleming* (1971a-b), *Leith* (1974) and *Lorenz* (2006), with each system trying a different strategy to simulate the impact of uncertainties in the initial conditions and approximations in atmospheric numerical models, but with all systems aiming to provide users with a more complete information of the possible future atmospheric states than the one that can be constructed using a single forecast.

At ECMWF, the ensemble prediction system underwent two major updates in 2006:

- On the 1st of February, the resolution of the 10-day ensemble forecasts was increased from T_L255L40 to T_L399L62 (spectral triangular truncation T399 with 62 vertical levels); this followed the increase in the resolution of the ECMWF analysis and high-resolution forecast from T_L511L60 to T_L799L91.

- On the 12th of September, the ensemble forecast length was extended from 10 to 15 days, with forecasts run with variable resolution (*Buizza et al* 2007): T_L399L62 up to forecast day 10, and T_L255L62 from day 10 to 15 (dissemination of the 15-day forecasts started on the 28th of November).

In the current system, initial uncertainties are simulated by adding (to the unperturbed analysis) initial perturbations defined by T42L62 singular vectors (*Buizza & Palmer* 1995) with the fastest growth over the first 48-hour of the forecast period; model uncertainties due to physical parameterisation schemes are simulated using a stochastic scheme based on perturbed physical tendencies (*Buizza et al* 1999); and the ensemble unperturbed analysis is generated by interpolating the T_L799L91 analysis to the ensemble T_L399L62 resolution.

The performance of the ECMWF ensemble system has been improving continuously since its implementation in 1992, as documented, for example, by the increase of about 2-days per decade of the skill of probabilistic predictions of 500 hPa geopotential height anomalies (*Buizza* 2006, *Leutbecher & Palmer* 2007), and by the increasing capability of the EPS to provide valuable information often 12 to 24 hours before single, higher-resolution forecasts in cases of extreme events (*Buizza & Hollingsworth* 2002, *Buizza & Chessa* 2002). This work discusses the ensemble performance in predicting the synoptic scale flow (represented by the 500 hPa geopotential height), but from a slightly different point of view than the one followed in the earlier works. More specifically, attention focuses on the following question: ***is the EPS adding value to the existing range of ECMWF single high-resolution forecasts?*** More precisely, is the T_L399L62 EPS performing better than an ensemble system based on lagged (as in *Hoffman & Kalnay* 1983) ECMWF high-resolution (T_L799L91) forecasts?

This issue is addressed by comparing the performance of the 51-member T_L399L62 EPS with two lagged-forecast ensemble systems based on the most recent six T_L799L91 high-resolution forecasts. More precisely, four ensemble systems are compared in this study: the original 51-member EPS (EPS51), a weighted 6-member EPS (EPS6wEM), with its ensemble-mean defined by giving optimally computed weights to the six individual members, the original 6-member lagged ensemble (LAG6) and a weighted 6-member lagged ensemble (LAG6wEM), with its ensemble-mean defined by giving optimally computed weights to the six individual members. The performance of these four ensembles is assessed considering both their average performance during a 7-month period, and a synoptic evaluation of their predictions of two recent storms that affected the Scandinavian countries in January 2005 and October 2006.

After this introduction, section 2 describes in more details the four ensemble systems and the accuracy measures used to assess their performance. Section 3 illustrates the methodology used to compute the ensemble-mean optimal weights. Section 4 discusses some average results valid for the Northern Hemisphere and Europe. Section 5 compares the performance of the two 'original' ensemble systems for two cases of extreme weather. Finally, section 6 summarizes the key results of this work and draws some

considerations on the value of the ECMWF 51-member EPS compared to the lagged-forecast ensemble system.

2. Ensemble systems and accuracy measures

Table 1 summarizes the key characteristics of the four ensemble configurations compared in this study:

- EPS51 is the 51-member ECMWF EPS, constructed using the 'control' forecast defined by the T_L399L62 forecast starting from the unperturbed analysis, and the 50 perturbed members with initial conditions perturbed using singular vectors (*Buizza & Palmer* 1995): this was the system operational during the period under investigation. In this system, the ensemble-mean is defined by giving the same weight (1/51) to the 51 members.

- EPS6wEM is a 6-member ensemble, constructed using the EPS control and 5 randomly-selected perturbed members, and with the ensemble-mean computed giving optimal weights to the 6 forecasts.

- LAG6 is the 6-member ensemble constructed using the 6 most recent, lagged T_L799L91 high-resolution forecasts (i.e. the forecasts started at the initial time, and 12, 24, 36, 48 and 60 hours earlier). In this system, the most recent T_L799L91 forecast (i.e. the one started at the initial time) is the 'control' forecast, and the ensemble-mean is defined by giving the same weight (1/6) to the 6 members. Note that to avoid reducing the ensemble membership after t+192h, for forecast lengths longer than t+192h one or more lagged t+240h forecasts have been persisted beyond 240 hours, as required.

- LAG6wEM is the same as LAG6, but with the ensemble-mean computed giving optimal weights to the 6 lagged, high-resolution forecasts.

Config	Size	Init time	FC times	Resolution	Ensemble-mean weighting
EPS51	51	12UTC of day d	0 to +180h	TL399L61	No
EPS6wEM	6	12UTC of day d	0 to +180h	TL399L61	Yes
LAG6	6	12UTC of day d to 00UTC of day (d-60h)	0 to +240h LAGGED	TL799L91	No
LAG6wEM	6	12UTC of day d to 00UTC of day (d-60h)	0 to +240h LAGGED	TL799L91	Yes

Table 1. Ensemble configurations: name, size, initial time, forecast times, forecast resolution and weighting procedure.

It is worth mentioning that two other lagged systems have been considered: the first one was based on the latest 6 lagged T_L399L62 EPS control forecasts, but since the average skill of the T_L799L91 forecast is better than the skill of the EPS control (see the discussion in section 4.1 and Figs. 3 and 4), its performance is not discussed. The second system was based on the latest 4 (instead of 6) lagged T_L799L91 forecasts: this system includes members with an average smaller root-mean-square-error (*rmse*) than LAG6 (since it includes only forecasts up to 36-hour older), but this has a negative impact on the ensemble spread, which becomes too small, and on the skill of the ensemble-mean and of probabilistic forecasts, which are both smaller than the ones of LAG6. Because of this, its performance is also not discussed.

Forecasts issued from 1 April to 30 Oct 2006 (213 days) of 500 hPa geopotential height (Z500), defined on a 2.5 regular latitude-longitude grid, have been compared over two regions:

- Northern Hemisphere (NH): latitude from 20°N to 80°N, longitude from 0°E to 360°E
- Europe (EU): latitude from 20°N to 80°N, longitude from 20°W to 45°E

No post-processing or calibration has been applied to any forecast field.

The average performance of the four ensemble systems has been assessed considering the level of ensemble spread, based on the comparison of the area-average ensemble standard deviation with the error of the ensemble-mean, the accuracy of single forecasts (control, ensemble-mean and perturbed members), measured using the *rmse* and the anomaly correlation coefficient (ACC), and the accuracy of probabilistic forecasts, measured using the ranked probability score and skill score, the area under the relative operating characteristic curve and the Brier score and skill score. The performance of the four ensemble systems in predicting two storms that affected the Scandinavian countries in January 2005 and in October 2006 have been assessed considering also mean-sea-level-pressure (MSLP) forecasts. For these forecasts, the *rmse*, and the intensity and position errors in predicting the MSLP minima inside a 20-degree region centred on the storm position at verification time have been computed.

3. Weighting methodology

In the weighted ensembles, the weight given to each member is a function of (a) its forecast error relative to the other members, and (b) the percentage of the analysis anomaly that projects onto the forecast anomaly, with anomalies computed with respect to the analysed climate. Forecast quality is assessed by considering one of the most commonly used measures of forecast error, the *rmse*. For consistency, a Euclidean norm is used to normalize vectors and to compute any scalar products between vectors. One of the strengths of this weighting methodology is that it is more general than if the weights were defined to specifically optimize one measure of forecast error, e.g. the *rmse* of the weighted ensemble-mean or the Brier score of the probabilistic prediction of positive geopotential height anomalies (please note that it is beyond the scope of this work to compare combination methods). By contrast, one of its weaknesses is that for results might not be optimal for all regions, variables and verification measures.

Consider the ensemble forecasts started at day *d*. Denote by $f_j(d,t)$ the $+t$ forecast of the j-th members of a 6-member ensemble system started at day *d*, by $a(d+t)$ the verifying analysis, and by c the climate. For each verification area, variable and forecast time $+t$, the weights have computed as follows:

a) Compute the anomaly of the verifying analysis, $[a(d+t)-c]$.

b) Compute the *rmse* of all ensemble members, $\sqrt{<[f_j(d,t)-a(d+t)]^2>}$, for $j=1,N$.

c) Find the *k-th* ensemble member with the smallest *rmse*.

d) Define the first element of the orhonormal basis as the normalized difference of the *k-th* member from the climate: $b_1 = \frac{f_k(d,t) - c}{\mid f_k(d,t) - c \mid}$.

e) Define the second element of the orhonormal basis as the difference of another ensemble member (1 if $k \neq 1$, or 2 if $k = 1$) from the climate orthogonal to the first element of the basis, normalized to have unitary norm: $b_2 = \frac{f_1(d,t) - c - < (f_1(d,t) - c); b_1 > b_1}{\mid f_1(d,t) - c - < (f_1(d,t) - c); b_1 > b_1 \mid}$. Continue until all 6-elements of the ortho-normal basis have been defined, by defining the j-th element of the orthonormal basis as the components of the j-th forecast anomaly orthogonal to the already defined elements of the basis, normalized to have unitary norm.

f) Compute the projection of the analysis anomaly (defined in a) onto each element of the orthonormal basis, $p_j = < (a(d+t) - c); b_j >$.

g) Compute the *j*-th day-d weight $w_j(d,t)$ as the normalize projection, i.e. the projection p_j normalized so that the sum of the daily normalized weights is one, $w_j(d,t) = \frac{p_j}{\sum_{j=1,N} p_j}$.

h) Once the daily weights $w_j(d,t)$ have been computed for the whole training period *D*, compute their average, $w_j(t) = \frac{1}{D} \sum_{d \in D} w_j(d,t)$. The average weight is the one used in the subsequent forecast.

In this procedure, steps (b)-(c) have been introduced to select the best forecast as the first element of the basis; in step (e), the removal of the projection of the j-th ensemble member anomaly onto the previous (j-1)-elements of the basis has been introduced to remove the cross correlation terms; in step (g), the renormalisation is equivalent to computing the relative amount of analysis anomaly projecting along each element of the basis; in step (h), the averaging has been introduced to stabilize the results. Please note that weights have been computed for each verification area (Northern Hemisphere and Europe).

Figure 1 shows the average LAG6wEM weights of the most recent member (i.e. the high-resolution forecast started at day *d*) and of the 12-hour lagged member, averaged for three 50-day training periods, the periods starting on the 5th of February, the 5th of April and the 5th of June 2006. Considering the period starting on the 5th of February (see also Table 2), the weight of the most recent member of LAG6wEM decreases with the forecast time from 1 at forecast day 1 to 0.52 at forecast day 10, while the weight of the d-12h member increases from zero to 0.24 at forecast day 10. Figure 1 shows that there is only a limited sensitivity to the period used to compute the average weights. It is also worth to point out that the weights listed in Table 2 are not too dissimilar from the ones computed by *Simmons* (1995): the differences between these and Simmons' values are most likely due to the fact that the period under investigation is different, that the forecasts used in Simmons have a lower resolutions, and to the fact that 10 instead of only 6 lagged forecasts were used by Simmons.

	Lagged forecasts					
	d	**d-12h**	**d-24h**	**d-36h**	**d-48h**	**d-60h**
T+24h	1.00	0.00	0.00	0.00	0.00	0.00
T+48h	1.00	0.00	0.00	0.00	0.00	0.00
T+72h	0.97	0.02	0.00	0.00	0.01	0.00
T+96h	0.91	0.04	0.01	0.01	0.02	0.01
T=120h	0.85	0.08	0.03	0.02	0.01	0.01
T+144h	0.81	0.12	0.03	0.02	0.01	0.01
T+168h	0.74	0.15	0.04	0.03	0.02	0.02
T+192h	0.69	0.17	0.05	0.03	0.03	0.03
T+216h	0.59	0.19	0.07	0.07	0.05	0.03
T+240h	0.52	0.24	0.09	0.07	0.06	0.02

Table 2: Weights computed for the members of the LAG6wEM ensemble during the 50-day periods starting on the 5th of February 2006. Each column represents one lagged-forecast, from the most recent to the one with the 60-hour lag; each row lists the weights given to the 6 lagged forecasts for a specific forecast step.

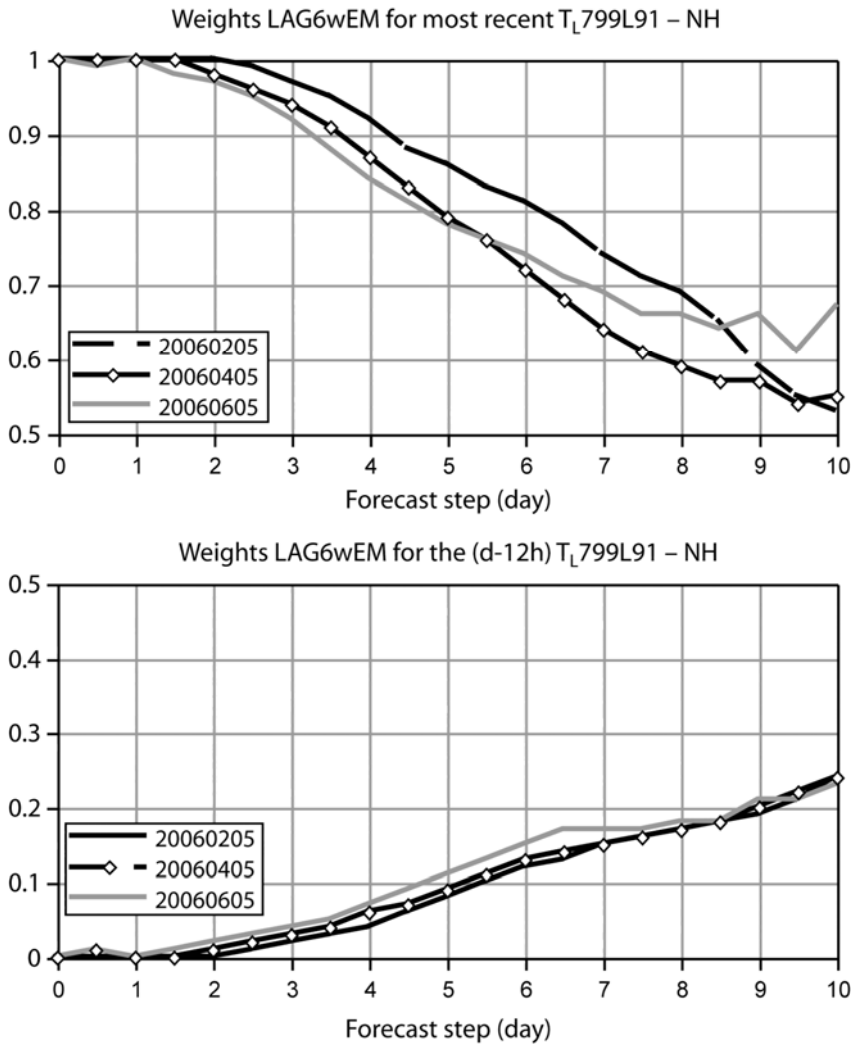

Figure 1: Lagged ensemble system LAG6wEM: average weights computed over three 50-day periods starting on the 5th of February (solid black line), the 5th of April (solid black line with symbols) and the 5th of June (grey line), for the most recent T_L799L91 forecast (top panel) and the d-12h T_L799L91 forecast (bottom panel). Weights refer to the 500 hPa geopotential height over Northern Hemisphere.

Figure 2 shows the EPS6wEM weights valid for the control forecast and a randomly-selected perturbed member (since, on average, the EPS perturbed forecasts have the same skill, their weights are, on average, equal). Compared to the weight of the LAG6wEM most recent member, the weight of the EPS6wEM control forecast decreases more steeply with the forecast time. Results also indicate that the EPS6wEM weights have a weaker sensitivity to the period used to compute them. Considering the weights for the period started on the 5th of February, at forecast day 7 the EPS6wEM control forecast has a weight of 0.41, which is still higher than the weight that should be expected if the control forecast had a skill similar to the perturbed members (~0.17, i.e. 1/6, for a 6-member ensemble system). This result is consistent with the fact that at forecast day 7 the average skill of the control forecast is still higher than the average skill of the perturbed members, as shown in Fig. 3 and discussed later in section 4.1. At forecast day 10 the average weight of to the control forecast (0.23) is close to the one of the ensemble perturbed members. Since the sensitivity of the average weights to the training period has been found to be rather weak, for ease of computation it has been decided to use for whole 7-month period the weights computed for the 50-day period starting on the 5th of February 2006. In practical terms, using a fixed set of weights instead of, e.g., moving-average weights, makes the procedure easier to run and, if used in an operational framework, it avoids the need to access a large amount of data to compute the daily weights.

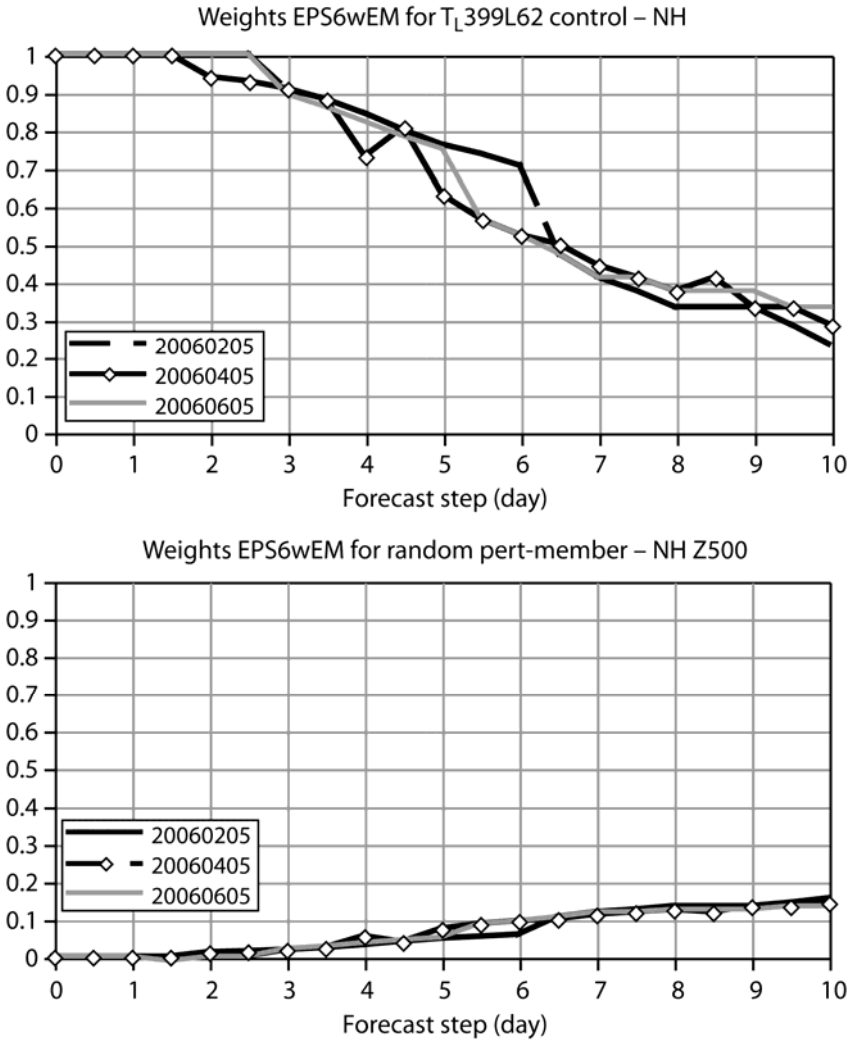

Figure 2: EPS6wEM: average weights computed over three 50-day periods starting on the 5th of February (solid black line), the 5th of April (solid black line with symbols) and the 5th of June (grey line) for the control forecast (top panel) and one (randomly selected) of the five perturbed forecast (bottom panel). Weights refer to the 500 hPa geopotential height over Northern Hemisphere.

It is worth to point out that the use of optimal weights also in the computation of probabilities has been investigated: this was done by adding, for each forecast, a contribution equal to the forecast optimal weight (instead of an equal weight set to 1/N, where N is the total number of ensemble members). In the case of EPS6 (see Fig. 2 for weights), for example, if a perturbed member (control) predicted an event at forecast day 4, the forecast probability would have been increased by 0.05 (0.85) instead of by 1/6. Results (not shown) have indicated that the use of weights in the computation of the probabilities has a negative impact in the early forecast range, when the control forecast has a very high optimal weight. This negative impact is due to the fact that, in the short forecast range, the probability density function of forecast states becomes too narrow if optimal weights are used (in other words the ensemble spread collapses), and this has a negative impact on the accuracy of the probabilistic forecasts. A similar negative impact was detected when weights were applied to LAG6 probabilistic forecasts.

4. 213-day average ensemble performance

In the first part of this section, the error of the control and ensemble-mean forecasts and the ensemble spread are discussed, while in the second part of this section the accuracy of the probabilistic predictions is discussed.

4.1 Error of the control, perturbed-members and ensemble-mean forecast, and ensemble standard deviation

Figures 3 and 4 show the error of the control forecasts, of the single perturbed members and of the ensemble-mean, and the ensemble standard deviation of the four ensembles. The LAG6/LAG6wEM control forecast (these two ensembles have, by construction, the same control forecast, which coincides with the most recent T_L799L91 forecast) have a smaller *rmse* than the EPS51/EPS6wEM control forecast (these two ensembles have, by construction, the same control forecast, which coincides with the EPS T_L399L62 forecast) due to the higher resolution. The comparison of the two *rmse* curves is a measure of the impact of increasing the forecast resolution from T_L399L62 to T_L799L91 on the average skill of a single forecast: results indicate a small difference, e.g. smaller than 6 hours at forecast day 6. Considering the perturbed members, the LAG6/LAG6wEM ensemble members have an average larger error than the EPS51/EPS6wEM members up to forecast day 8: this is due to the fact that the lagged ensembles include forecasts up to 60-hour ‘older’ than the control, while the EPS51/EPS6wEM ensembles use forecasts with the same ‘age’.

It is worth mentioning that if only the most recent 4 (instead of 6) lagged high-resolution forecasts were used, the average error of the lagged members would be similar to the one of EPS51 (not shown). But reducing the membership from 6 to 4 of the lagged ensemble would make the ensemble performance worse in many other aspects: not only the spread, but also the skill of the ensemble-mean and of probabilistic forecasts would be lower (not shown): for this reason, the 4-lagged configuration has been judged sub-optimal and is not discussed.

Figures 3 and 4 show also the ensemble spread measured by the ensemble standard deviations (i.e. the average distance of each single member from the ensemble-mean). Compared to LAG6, EPS51 has a slightly smaller standard deviation up to forecast day 2, and a larger one afterwards. The EPS51 ensemble has the best match between the ensemble-mean error and the ensemble standard deviation, especially in the medium range (say after forecast day 5), as can be seen from Fig. 5 (bottom panels). This closer match between the

ensemble standard deviation and the ensemble-mean error is one of the reasons why EPS51 has a lower ensemble-mean *rmse* and a higher probabilistic skill (see discussion in section 4.2) in the medium-range.

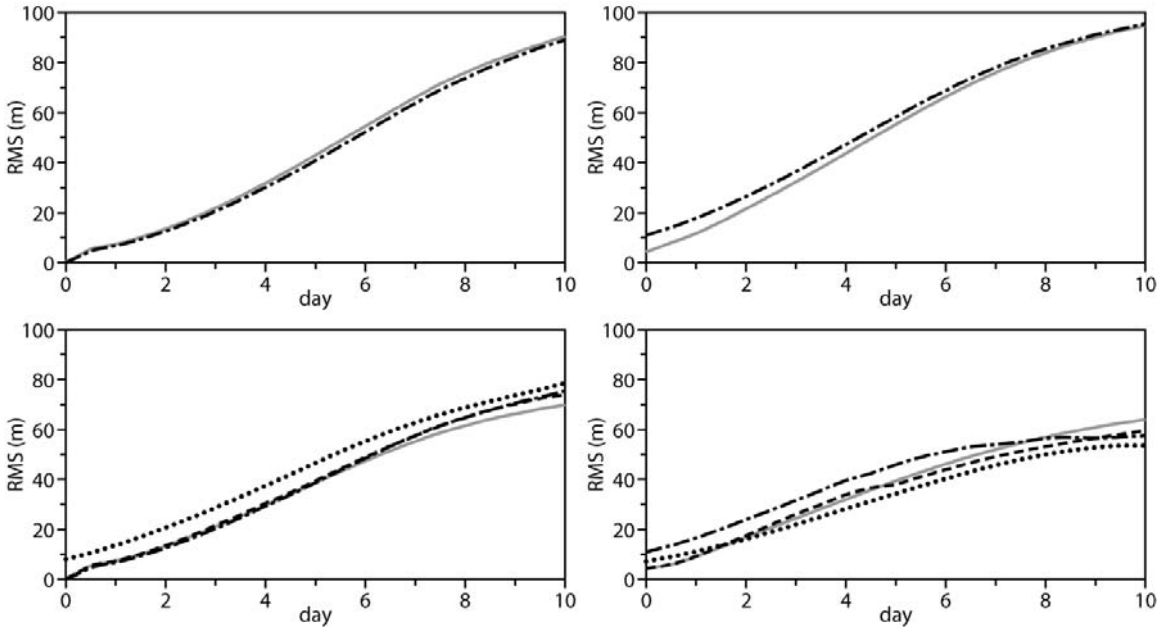

Figure 3: Top panels: 213-case average error of the forecasts given by the control (top-left panel) and of the perturbed members (top-right panel) of EPS51/EPS6wEM (solid grey lines) and LAG6/LAG6wEM (chain-dashed black lines). Bottom panels: as top panels but for the error of the ensemble-mean (bottom-left panel) and ensemble standard deviation (bottom-right panel), of EPS51 (solid grey lines), EPS6wEM (dashed black lines), LAG6 (dotted black lines) and LAG6wEM (chain-dashed black lines). Values refer to the 500 hPa geopotential height over the Northern Hemisphere.

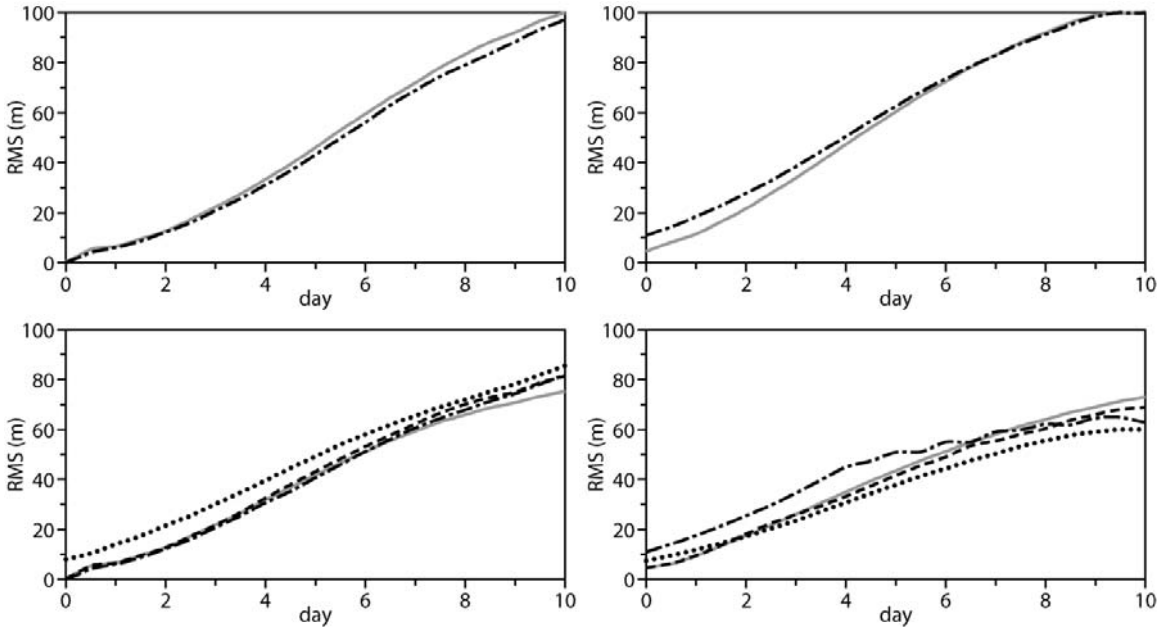

Figure 4: As Figure 3 but for Europe. Top panels: 213-case average error of the forecasts given by the control (top-left panel) and of the perturbed members (top-right panel) of EPS51/EPS6wEM (solid grey lines) and LAG6/LAG6wEM (chain-dashed black lines). Bottom panels: as top panels but for the error of the ensemble-mean (bottom-left panel) and ensemble standard deviation (bottom-right panel), of EPS51 (solid grey lines), EPS6wEM (dashed black lines), LAG6 (dotted black lines) and LAG6wEM (chain-dashed black lines). Values refer to the 500 hPa geopotential height over Europe.

Considering the ensemble-mean forecasts, the LAG6wEM ensemble has the smallest error up to forecast day 4, while EPS51 has the smallest error from forecast day 6. Up to forecast day 4, the EPS6wEM ensemble-mean has a slightly smaller rmse than EPS51: this indicates that giving a higher weight to the control forecast has a small but positive impact on the skill of the ensemble-mean up to this forecast day, and that ensemble-size has a negligible impact on the skill of the ensemble-mean forecast. The fact that after forecast day 6 the EPS51 ensemble-mean has the smallest error, smaller also than EPS6wEM, indicates that having a larger size (51 instead of 6) have a positive impact in this forecast range. The comparison of the errors of the LAG6 and LAG6wEM ensemble-mean forecasts indicate that giving a different weight to the 6 members of the LAG6 ensemble has a very large impact on the error of the ensemble-mean up to forecast day 10 (it should be noted that the impact decreases after forecast day 8 is related to the fact that for some of the lagged forecasts, t+240h forecasts have been persisted beyond forecast day 10). This large difference between the error of LAG6wEM and LAG6 is due to the fact that the lagged ensemble members have very different forecast error statistics, with older members characterized by larger errors.

Figure 5 (top panels) shows the difference between the *rmse* of the ensemble-mean and the control forecast for the four ensembles: note that the LAG6 ensemble-mean has a larger *rmse* than its control up to forecast day 6, confirming that weights are required to properly use lagged forecasts. Figure 5 also shows that the difference between the *rmse* of the ensemble-mean and the control forecast is larger for EPS51, especially after forecast day 6, confirming that having a larger ensemble size improves the ensemble-mean performance especially in the medium range. Since the EPS51 and the EPS6wEM ensembles use the same control forecast, Fig. 5 also confirms that the difference between the *rmse* of the EPS51 and EPS6wEM ensemble-mean forecasts is very small in the early forecast range, while it becomes larger after forecast day 4, with the EPS51 ensemble-mean having a lower *rmse* than EPS6wEM.

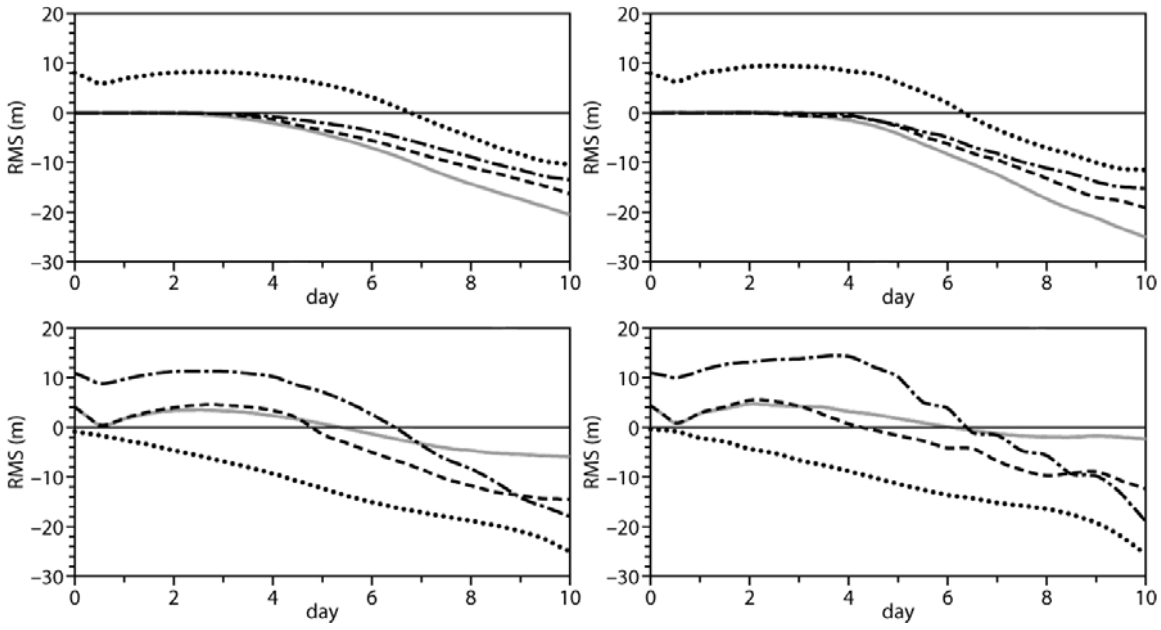

Figure 5. 213-case average difference between the rmse *of the ensemble-mean and the control (top panels, negative values means that the ensemble-mean has a lower* rmse*) and between the ensemble standard deviation and the ensemble-mean* rmse *(bottom panel, negative values means that the ensemble standard deviation is too small) computed for the Northern Hemisphere (left panels) and Europe (right panels) for EPS51 (solid grey lines), EPS6wEM (dashed black lines), LAG6 (dotted black lines) and LAG6wEM (chain-dashed black lines). Values refer to the 500 hPa geopotential height.*

4.2 Accuracy of probabilistic forecasts

The accuracy of probabilistic forecasts has been measured using three measures: the ranked probability score and skill score (RPSS, *Wilks* 1995), which is a measure of the distance between the forecast and the observed distributions, the area under the relative operating curve (ROCA, see e.g. *Swets* 1986 and *Wilks* 1995), which depends on the capability of the probabilistic forecast for a categorical event to discriminate between occurrence and non-occurrence, and the Brier skill score (*Brier* 1956, *Wilks* 1995), which is the equivalent of the root-mean-square error for probabilistic forecasts of a categorical event. The RPSS has been computed using persistence as reference forecast, while the BSS has been computed using the sample climatology as reference forecast.

Figure 6 shows the RPSS for the probabilistic prediction of Z500 anomalies, and the ROCA and the BSS for the probabilistic prediction of Z500 positive anomalies (similar considerations could be draw considering other thresholds) of ensembles EPS51, EPS6wEM and LAG6/LAG6wEM (note that by construction the LAG6 probabilistic forecasts coincide with LAG6wEM ones). Overall, the original EPS51 probabilistic forecasts are the best for all forecast ranges, with differences in predictability reaching 24 hours in the medium-range (say after forecast day 5). The comparison of the EPS51 and the EPS6wEM scores gives a measure of the impact of ensemble size on the accuracy of probabilistic prediction: results confirm again that

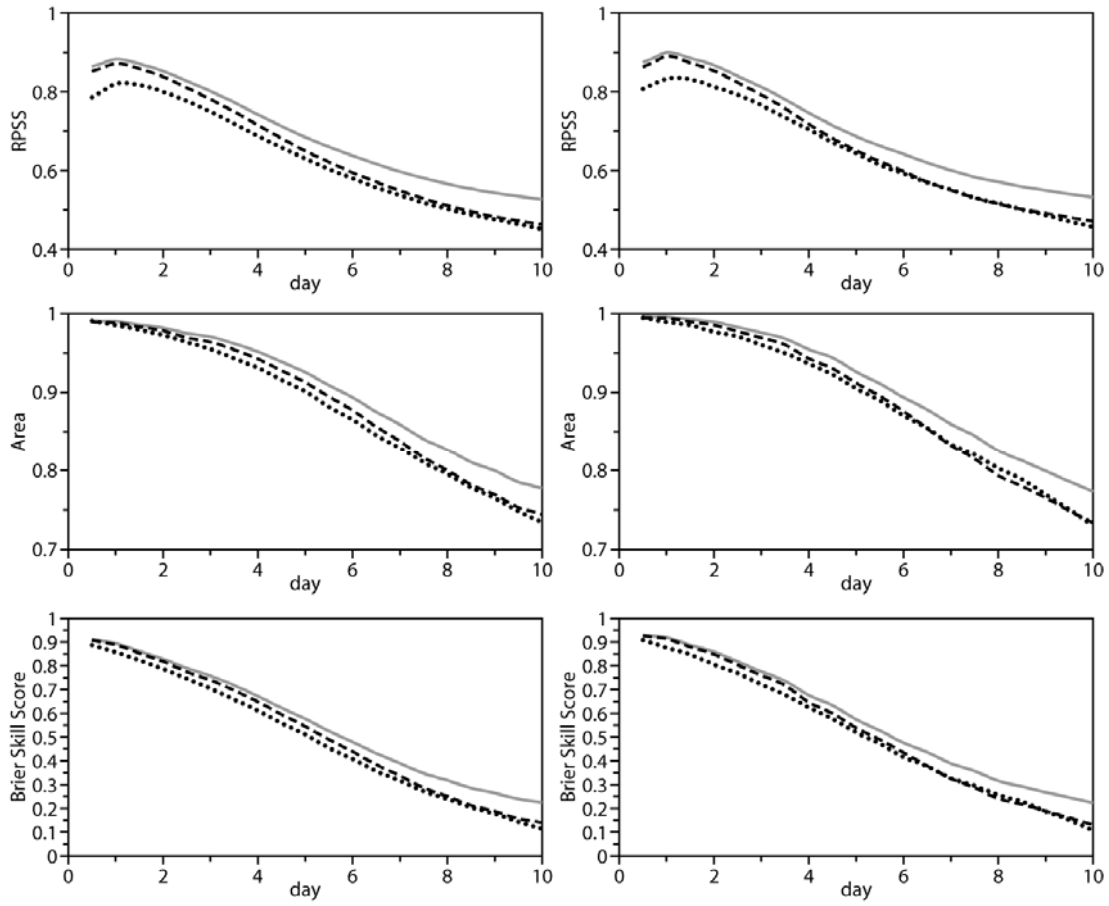

Figure 6: 213-case average ranked probability skill score (top panel), area under the relative operating characteristics curve (middle panel) and Brier skill score (bottom panel) computed for the Northern Hemisphere (left panels) and Europe (right panels) for EPS51 (solid grey lines), EPS6wEM (dashed black lines), LAG6 (dotted black lines) and LAG6wEM (chain-dashed black lines). Values refer to the probabilistic prediction of positive 500 hPa geopotential height.

a larger ensemble size is more important in the medium-range rather than in the short range, with differences in predictability of about 36 hours at forecast day 10. The comparison of the EPS6wEM and LAG6wEM curves indicates that EPS6wEM performs better than LAG6wEM up to forecast day 8 over Northern Hemisphere, and day 6 over Europe. The difference is due to the fact that EPS6wEM has a better tuned spread than LAG6wEM, and it includes members with an average smaller error (see Figs. 3, 4).

It is worth mentioning that if the lagged-ensemble were built using only the latest four (instead of the latest six) lagged high-resolution forecasts, the skill of the probabilistic forecasts would be even smaller than the one of LAG6 (not shown): as mentioned in section 3, this is due to the fact that such a 4-member lagged ensemble system would have a too smaller spread.

5. Synoptic analysis of two Northern European storms

To complete the comparison of the 51-member EPS and the 6-member lagged ensemble system, forecasts from the EPS51 and LAG6 systems for two storms that affected Northern Europe are discussed in this section. The first storm is Gudrum, the strongest storms since 1969 to hit the Scandinavian countries in January 2005. The second storm hit the Scandinavian countries in October 2006, and was declared the strongest storm since 1969 after Gudrum. Since in 2005 the operational ensemble and the high-resolution systems had a lower resolution, ensemble forecasts for Gudrum have been re-run from the 1st to the 7th of January with the current T_L399L62 resolution, and single forecasts have been re-run with a T_L799L91 resolution. The comparison will focus on the t+48 and t+96 hour forecast ranges, to assess whether in this forecast range there is any major difference between the forecasts of two ensemble systems in the case of an extreme event.

5.1 9 January 2005: storm (Gudrum)

Hurricane-strength winds whipped across Sweden and Denmark, leaving at least 11 people dead, seven of them in Sweden, as road and rail traffic was disrupted in the deadliest storm since 1969. The powerful storm left more than 400,000 homes in southern Sweden without power, with 200,000 of them still in the dark on the 10th of January. Damage was estimated at 500 million Swedish kronor (72 million US dollars), according to insurance group Laensfoersaekringar. In Denmark, some 16,500 homes were still without power on the 10th of January, including 12,500 in the Copenhagen region, while damage in the entire country was estimated at one billion Danish kronor (176 million US dollars).

EPS51 and LAG6 mean-sea-level-pressure (MSLP) forecasts valid for 00 UTC of the 9th of January 2005 (the time when the storm intensity peaked) have been compared to assess the difference in quality of two ensemble systems. The T_L799L91 forecasts have been re-run starting from a T_L799L91 analysis (*A. Simmons* is acknowledged for providing access to this analysis data). Since these high-resolution analysis was available only from the 1st of January, LAG6 ensembles could have been generated only up to 120-hour before the verification time. Thus, the EPS51 and the LAG6 ensembles are compared from t+36h to t+120h.

Figure 7 shows the MSLP analysis at 00 UTC of the 9th of January (the time when the storm was positioned over the Baltic Sea), and the corresponding 48-hour forecasts started at 00 UTC of the 7th of January, from the T_L799L91 high-resolution system and from the EPS51 control, ensemble-mean and the 16 members with the smallest *rmse*. Figure 8 shows the verifying analysis and the corresponding 48-hour forecasts from the EPS51 control and ensemble-mean, the T_L799L91 high-resolution forecast, and the LAG6 ensemble-mean

and 5 perturbed members (i.e. the T_L799L91 forecasts started 60, 72, 84, 96 and 120 hours before the verification time). Figure 7 shows that all the 16 EPS51 perturbed members predicted an intense storm, with the best forecast having an *rmse* (computed inside a 40-degree region centred on the observed position) of 2.4 hPa, and 15 perturbed members with an *rmse* smaller than the *rmse* of the EPS51 control forecast (which has an *rmse* of 3.9 hPa). By contrast, only two EPS51 members have an *rmse* smaller than the T_L799L91 forecast (which has an *rmse* of 2.6 hPa). The fact that the T_L799L91 forecast has a ~45% smaller rmse than the EPS51 control forecast confirms the average results discussed before, i.e. that increasing the resolution from T_L399L62 to T_L799L91 has a positive impact on the forecast accuracy. Note that the EPS51 ensemble-mean forecast predicted a weaker system than each of the 16 best EPS51 members, with an *rmse* of 3.7 hPa: this confirms earlier results that the ensemble-mean forecast is not an ideal product to be used to predict an intense, rapid developing system (*Buizza & Hollingsworth* 1999). Figure 8 shows that the LAG6 ensemble-mean has a larger *rmse* than the EPS51 ensemble-mean, and that none of the LAG6 lagged members outperforms the LAG6 most recent forecast.

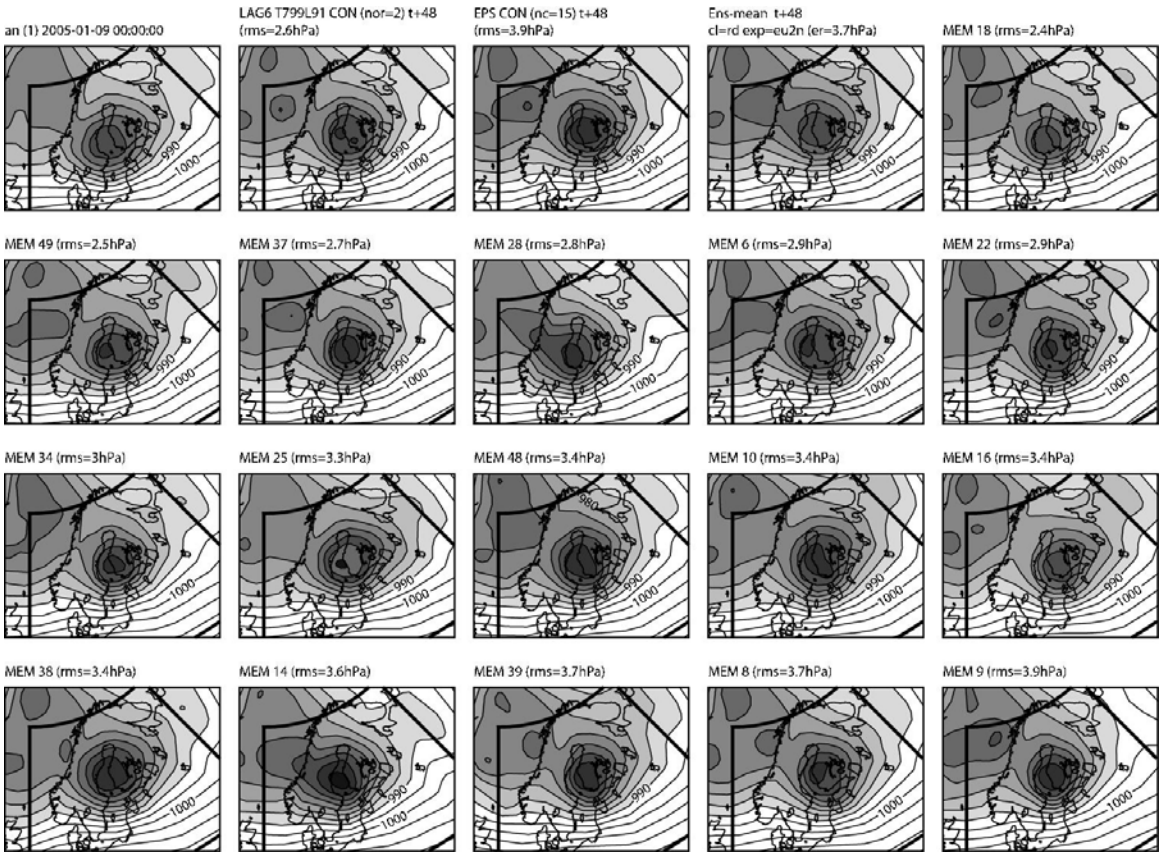

Figure 7: Storm of 00 UTC of 9 January 2005 (Gudrum). First row: mean-sea level pressure analysis at 00UTC of 9 Jan 2005 (1st panel) and corresponding t+48h forecasts started at 00UTC of 7 Jan: T_L799L91 forecast (2nd panel), T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel), and EPS51 member with smallest RMS error inside verification region (5th panel). Other panels: forecasts from 15 other EPS51 members with the lowest rmse *inside the verification area, ranked by* rmse. *MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

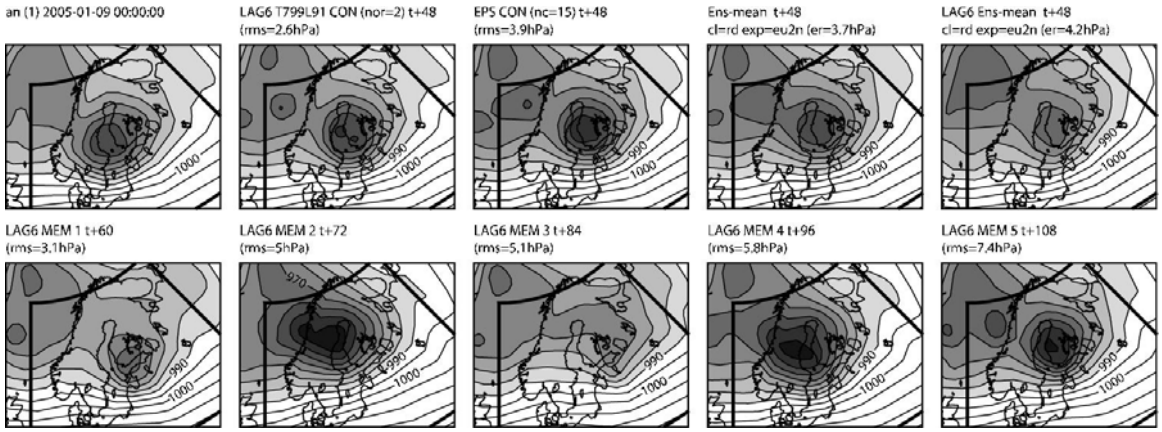

*Figure 8: Storm of 00 UTC of 9 January 2005 (Gudrum). First row: mean-sea level pressure analysis at 00UTC of 9 Jan 2005 (1st panel), t+48h T_L799L91 forecast started at 00UTC of 7 Jan (2nd panel), t+48h T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel) started at 00UTC of 7 Jan, and t+48h LAG6 ensemble-mean forecast (5th panel). Second row: T_L799L91 LAG6 t+60h forecast started at 12UTC of 6 Jan (1st panel), t+72h forecast started at 00UTC of 6 Jan (2nd panel), t+84h forecast started at 12UTC of 5 Jan (3rd panel), t+96h forecast started at 00UTC of 5 Jan (4th panel) and t+108h forecast started at 12UTC of 4 Jan (5th panel). MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

Figures 9 and 10 show the 96-hour EPS51 and LAG6 ensemble forecasts valid for the same verification time (00 UTC of the 9th of January). At this forecast range, the difference between the EPS51 control (with an *rmse* of 8.4 hPa) and the LAG6 most-recent forecast (with an *rmse* of 5.8 hPa) is still very large. The EPS51 96-hour best forecast has an *rmse* of 5.5 hPa, and 18 EPS51 members have an *rmse* smaller than the EPS51 control forecast (which had an *rmse* 8.4 hPa), while only one EPS51 member has an *rmse* smaller than the T_L799L91 forecast. Figure 10 shows that the LAG6 ensemble-mean has a smaller *rmse* than the EPS51 ensemble-mean, and that none of the LAG6 lagged members outperforms the LAG6 most recent forecast.

Figures 7-10 have shown that in the t+48h and the t+96h forecast many EPS51 members outperform the EPS51 control forecasts, with only one or two EPS51 members having an *rmse* smaller than the T_L799L91 forecast, while none of the LAG6 lagged forecasts outperform the T_L799L91 most-recent forecast. To assess whether these results are valid also for other forecast times, the *rmse* of the EPS51 best member and of the ensemble-mean have been compared with the *rmse* of the T_L799L91 high-resolution forecast. The top panel of Fig. 11 shows that only for some forecast steps the EPS51 best member has an *rmse* smaller than the high-resolution forecast, and that the difference between the *rmse* of the ensemble-mean and the high-resolution forecasts is always positive, indicating that the EPS51 ensemble-mean has a higher *rmse* than the T_L799L91 forecast. The bottom panel of Fig. 11 shows that the number of EPS51 perturbed-members with an *rmse* smaller than the EPS51 control forecast varies between 1 and 31, while the number of perturbed members with intensity and position errors smaller than the EPS51 control forecasts is smaller, ranging from 0 (for forecast step t+120h) to 21. Figure 12 compares these EPS51 statistics (expressed now in percentages instead of in number of perturbed members) with the corresponding LAG6 statistics (note that since the LAG6 ensemble includes only 5 perturbed members, its percentages are multiples of 20%). Figure 12 indicates that for this case the percentage of EPS51 members outperforming the EPS51 control is higher than the percentage of LAG6 members outperforming the LAG6 control (i.e. the T_L799L91 most-recent forecast).

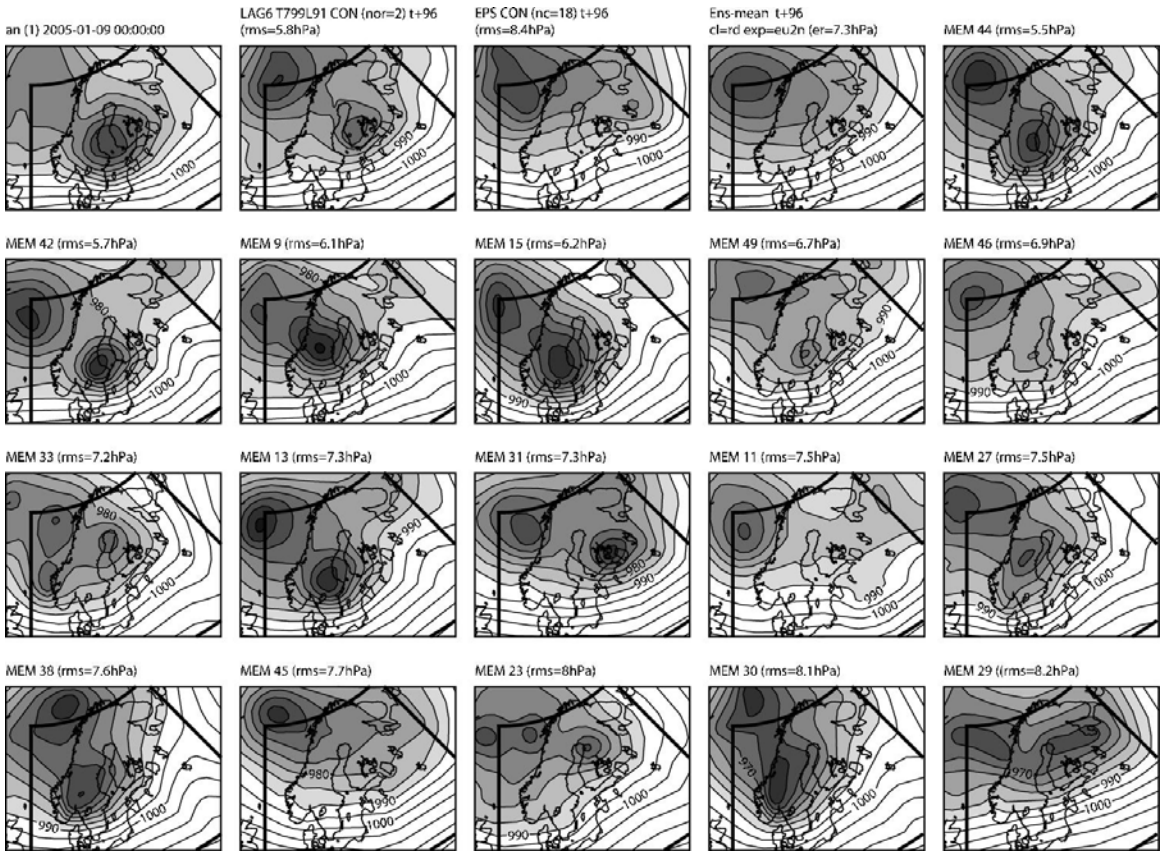

Figure 9: Storm of 00 UTC of 9 January 2005 (Gudrum). First row: mean-sea level pressure analysis at 00UTC of 9 Jan 2005 (1st panel) and corresponding t+96h forecasts started at 00UTC of 5 Jan: T_L799L91 forecast (2nd panel), T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel), and EPS51 member with smallest rmse *inside verification region (5th panel). Other panels: forecasts from 15 other EPS51 members with the lowest* rmse *inside the verification area, ranked by RMS error. MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

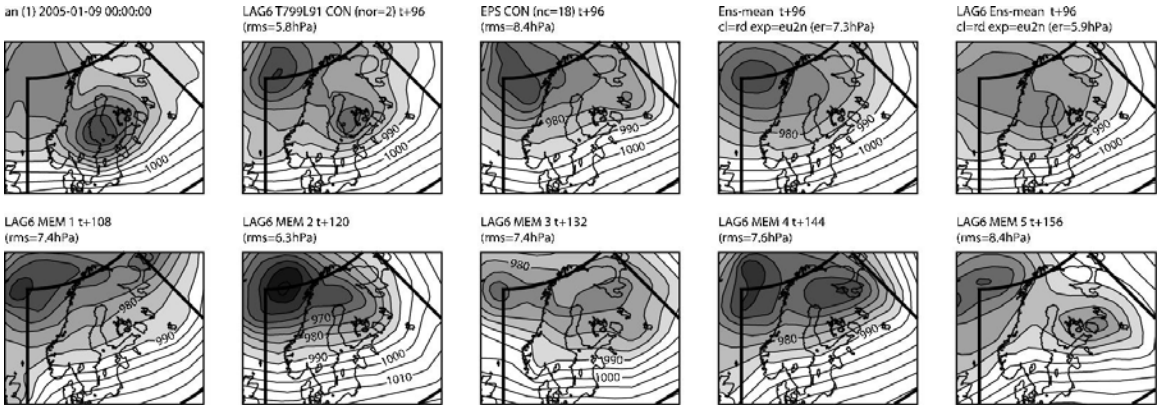

*Figure 10: Storm of 00 UTC of 9 January 2005 (Gudrum). First row: mean-sea level pressure analysis at 00UTC of 9 Jan 2005 (1st panel), t+96h T_L799L91 forecast started at 00UTC of 5 Jan (2nd panel), t+96h T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel) started at 00UTC of 5 Jan, and t+96h LAG6 ensemble-mean forecast (5th panel). Second row: T_L799L91 LAG6 t+108h forecast started at 12UTC of 4 Jan (1st panel), t+120h forecast started at 00UTC of 4 Jan (2nd panel), t+132h forecast started at 12UTC of 3 Jan (3rd panel), t+144h forecast started at 00UTC of 3 Jan (4th panel) and t+156h forecast started at 12UTC of 2 Jan (5th panel). MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

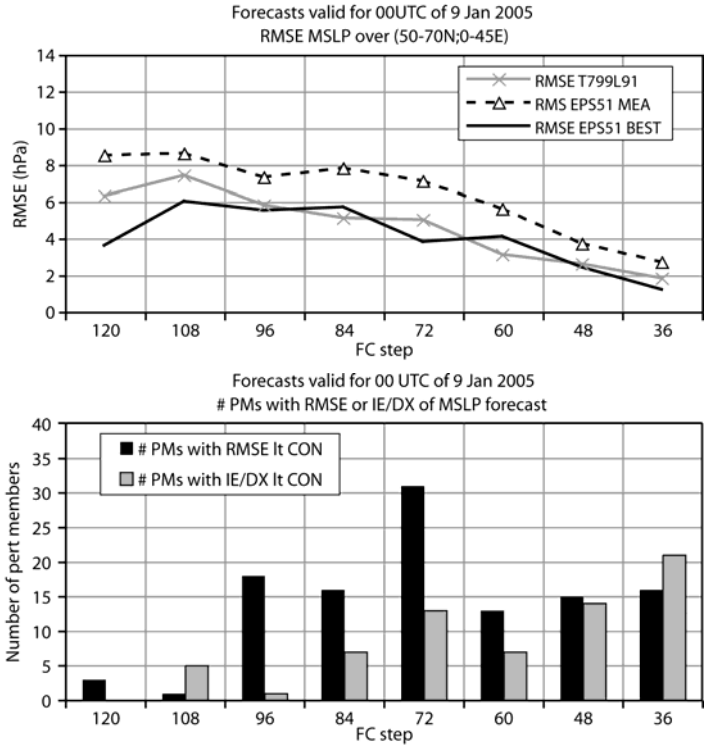

Figure 11: Storm of 00 UTC of 9 January 2005 (Gudrum). Top panel: rmse *of the high-resolution T_L799L91 forecast (grey line with crosses), and of the T_L399L62 EPS51 ensemble-mean (dashed line with diamonds) and best ensemble member (solid black line). Bottom panel: number of EPS51 perturbed members with* rmse *(black bars) and with intensity and position error (grey bars) smaller than the EPS51 control. [*rmse *have been computed inside (50-70N; 0-45E).]*

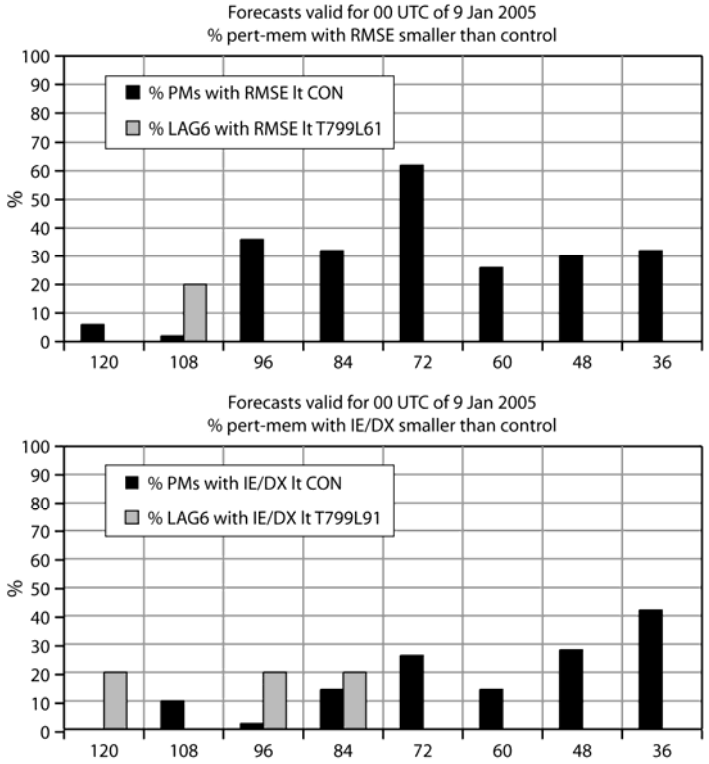

Figure 12: Storm of 00 UTC of 9 January 2005 (Gudrum). Top panel: percentage of EPS51 (black bars) and LAG6 (grey bars) perturbed members with rmse *(top panel) and with intensity and position error (bottom panel) smaller than the corresponding control forecasts (i.e. T_L399L62 for the EPS51 ensemble, and T_L799L91 for the LAG6 ensemble). [*rmse *has been computed inside (50-70N; 0-45E).]*

5.2 27 October 2006 storm

At 12 UTC of the 27th of October 2006, another very severe storm, although slightly less intense than Gudrum, hit Scandinavia, causing again a lot of damages and disruption. As for the 2005 storm, MSLP original forecasts given by the EPS51 and the LAG6 ensembles have been compared.

Figures 13 shows the MSLP analysis at 12 UTC of the 27th of October (the time when the storm was positioned over the Baltic Sea), and the corresponding 48-hour (i.e. of the forecasts started at 12 UTC of the 25th of October) forecasts from the operational high-resolution system (which had a resolution of T_L799L91), and from the EPS51 control, ensemble-mean and 16 perturbed members with the smallest *rmse*. Figure 14 shows the verifying analysis and the corresponding 48-hour forecasts from the operational high-resolution system, the EPS51 control and ensemble-mean, together with the ensemble-mean and the 5 perturbed members of the corresponding LAG6 ensemble (i.e. T_L799L91 forecasts started 60, 72, 84, 96 and 120 hours before the verification time). Figure 13 shows that all the 16 EPS51 perturbed members predicted an intense storm, with the best forecast having an *rmse* (inside a region centred on the observed position) of 1.8 hPa, and 3 perturbed members with an *rmse* smaller than the high-resolution (which has an *rmse* of 2.3 hPa) and the control forecasts (which has an *rmse* of 2.3 hPa). Figure 14 shows that none of the LAG6 lagged forecasts has an *rmse* smaller than the T_L799L91 most recent forecast and that the LAG6 ensemble-mean forecast was worse than the EPS51 ensemble-mean forecast, with an *rmse* of 3.6 hPa compared to 2.5 hPa.

Figures 15 and 16 show the 96-hour EPS51 and LAG6 ensemble forecasts valid for the same verification time (12 UTC of the 27h of January). At this forecast range, the difference between the EPS51 and the lagged ensembles is more evident. The EPS51 96-hour best forecast has an *rmse* of 3.3 hPa, four EPS51 members have an *rmse* smaller than the *rmse* of the TL799L91 most recent forecast and three have an *rmse* smaller than the EPS51 control forecast. By contrast, all the LAG6 ensemble forecasts had an *rmse* larger than the 96-hour high-resolution forecasts, all larger than 7 hPa. Note that both the EPS51 and the LAG6 ensemble-means predict a weaker system, with the EPS51 ensemble-mean characterized by a smaller *rmse* (5.7 compared to 6.2 hPa).

Figures 13-16 have shown that in the t+48h and the t+96h forecast range, few EPS51 members outperform the control and the T_L799L91 high-resolution forecast while none of the LAG6 forecasts outperform the T_L799L91 most recent forecast, that the difference between the EPS51 ensemble-mean and the high-resolution forecasts is rather small. The top panel of Fig. 17 confirms that for all forecast steps the EPS51 best member has an *rmse* smaller than the higher-solution forecast, and that the difference between the *rmse* of the ensemble-mean and the high-resolution forecasts is rather small. The bottom panel of Fig. 17 shows that the number of EPS51 perturbed-members with an *rmse* smaller than the EPS51 control forecast varies between 0 (for forecast step t+60h) and 34, while the number of perturbed members with intensity and position errors smaller than the high-resolution forecasts is smaller, ranging from 0 to 14. Figure 18 compares these statistics (expressed now in percentages instead of in number of perturbed members) for the EPS51 and the LAG6 ensembles. Figure 18 shows that while the first statistics indicate that the EPS51 and the LAG6 ensembles perform very similarly, the second statistics indicate that the EPS51 performs better than the LAG6 ensemble, especially in the short forecast range.

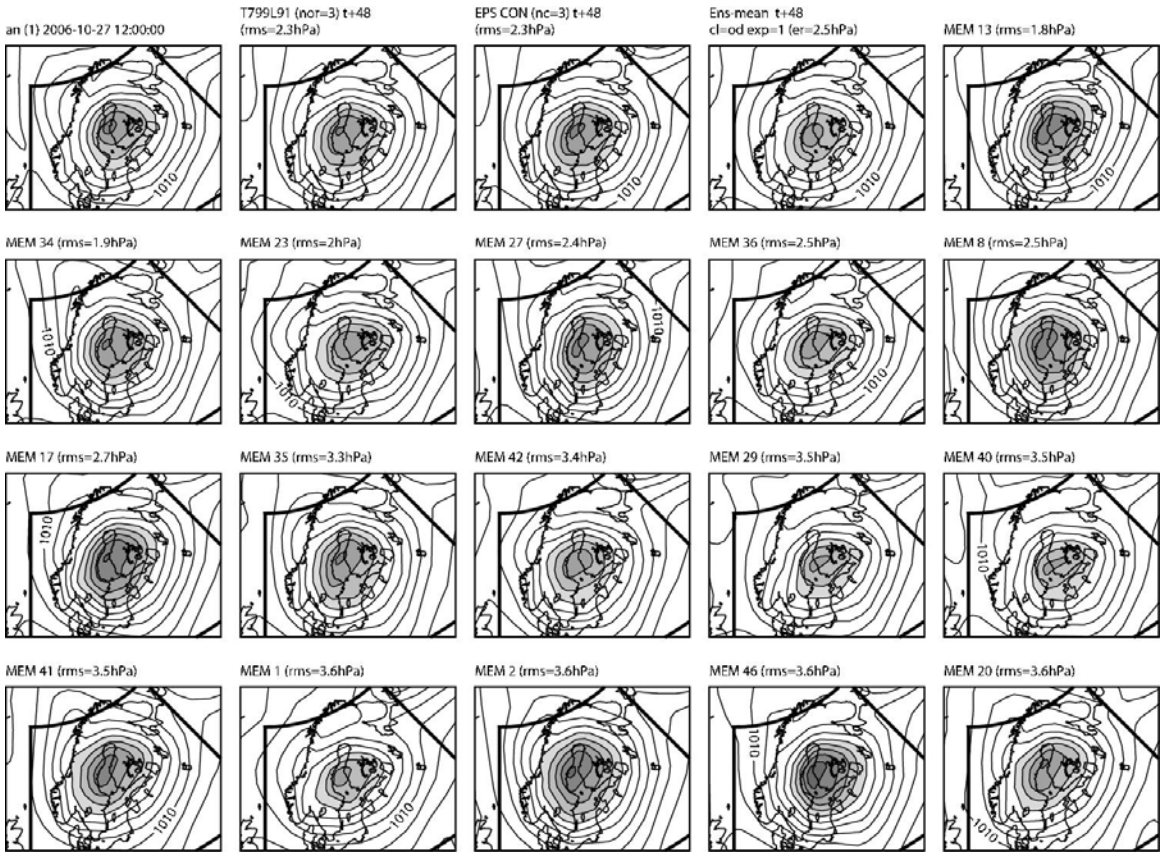

Figure 13: Storm of 12 UTC of 27 October 2006. First row: mean-sea level pressure analysis at 12UTC of 27 Oct 2006 (1st panel) and corresponding t+48h forecasts started at 12UTC of 25 Oct: T_L799L91 forecast (2nd panel), T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel), and EPS51 member with smallest rmse *inside verification region (5th panel). Other panels: forecasts from 15 other EPS51 members with the lowest* rmse *inside the verification area, ranked by RMS error. MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

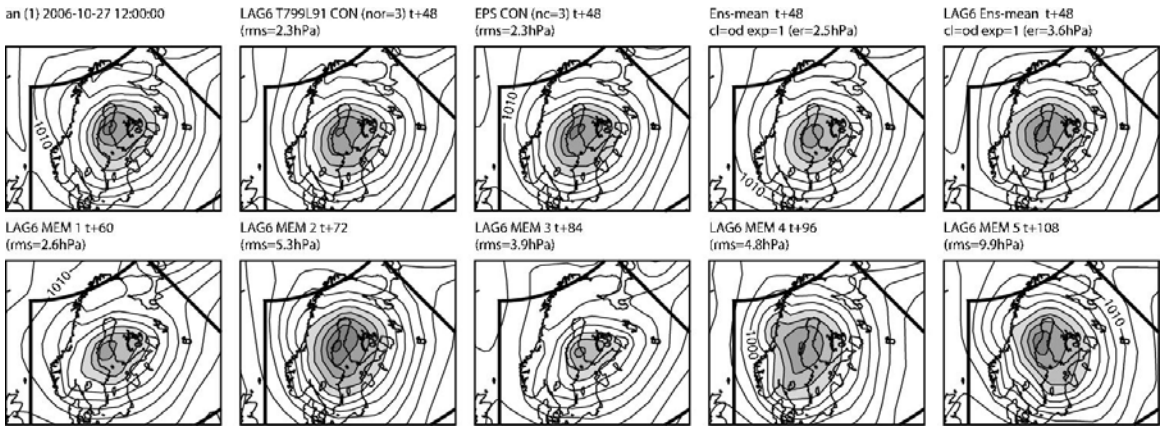

*Figure 14: Storm of 12 UTC of 27 October 2006. First row: mean-sea level pressure analysis at 12UTC of 27 Oct 2006 (1st panel), t+48h T_L799L91 forecast started at 12UTC of 25 Oct (2nd panel), t+48h T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel) started at 12UTC of 25 Oct, and t+48h LAG6 ensemble-mean forecast (5th panel). Second row: LAG6 T_L799L91 t+60h forecast started at 00UTC of 25 Oct (1st panel), t+72h forecast started at 12UTC of 24 Oct (2nd panel), t+84h forecast started at 00UTC of 24 Oct (3rd panel), t+96h forecast started at 12UTC of 23 Oct (4th panel) and t+108h forecast started at 00UTC of 23 Oct (5th panel). MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

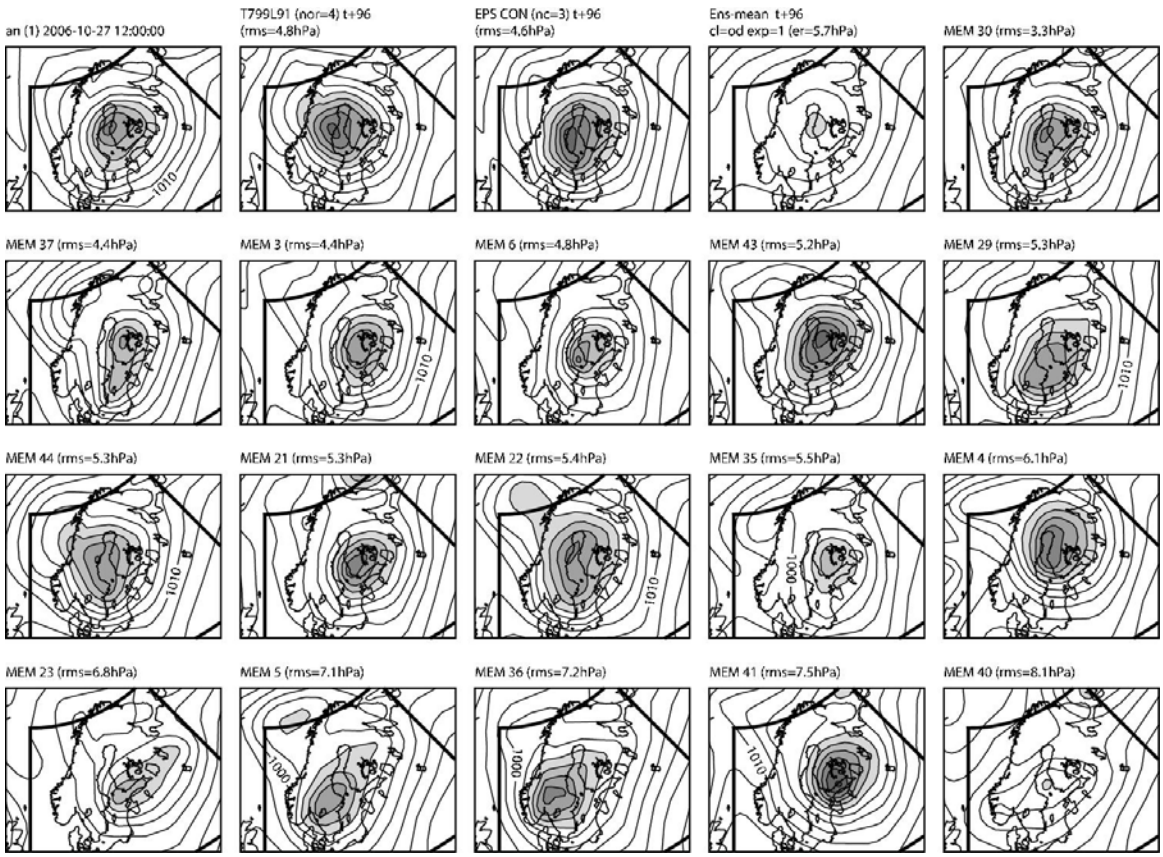

Figure 15: Storm of 12 UTC of 27 October 2006. First row: mean-sea level pressure analysis at 12UTC of 27 Oct 2006 (1st panel) and corresponding t+96h forecasts started at 12UTC of 23 Oct: T_L799L91 forecast (2nd panel), T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel), and EPS51 member with smallest rmse *inside verification region (5th panel). Other panels: forecasts from 15 other EPS51 members with the lowest RMS error inside the verification area, ranked by* rmse*. MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [*rmse *has been computed inside (50-70N; 0-45E).]*

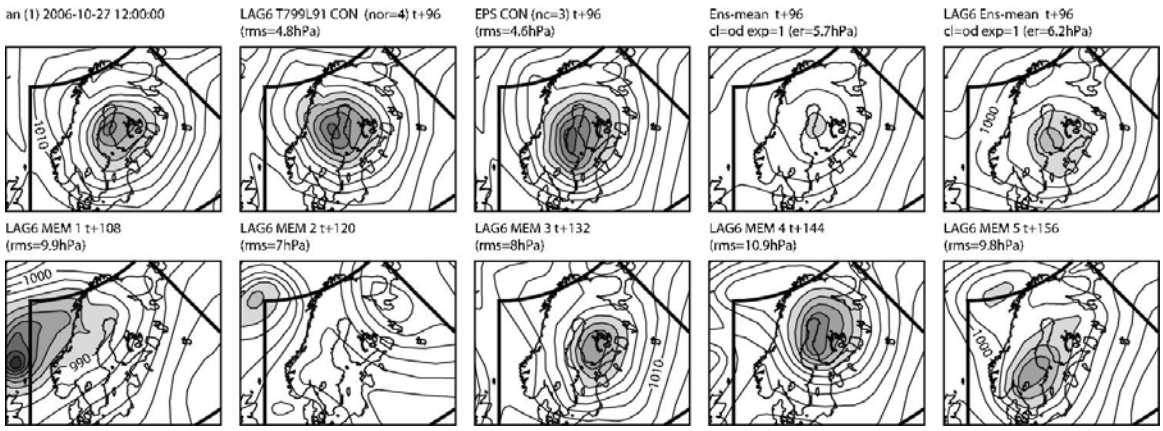

Figure 16: Storm of 12 UTC of 27 October 2006. First row: mean-sea level pressure analysis at 12UTC of 27 Oct 2006 (1st panel), t+96h T_L799L91 forecast started at 12UTC of 23 Oct (2nd panel), t+96h T_L399L62 EPS51 control (3rd panel) and ensemble-mean forecasts (4th panel) started at 12UTC of 23 Oct, and t+96h LAG6 ensemble-mean forecast (5th panel). Second row: LAG6 T_L799L91 t+108 forecast started at 00UTC of 23 Oct (1st panel), t+120h forecast started at 12UTC of 22 Oct (2nd panel), t+132h forecast started at 00UTC of 22 Oct (3rd panel), t+132h forecast started at 12UTC of 21 Oct (4th panel) and t+144h forecast started at 00UTC of 21 Oct (5th panel). MSLP contour interval is 5 hPa, with shading for MSLP values lower than 990 hPa. [rmse has been computed inside (50-70N; 0-45E).]

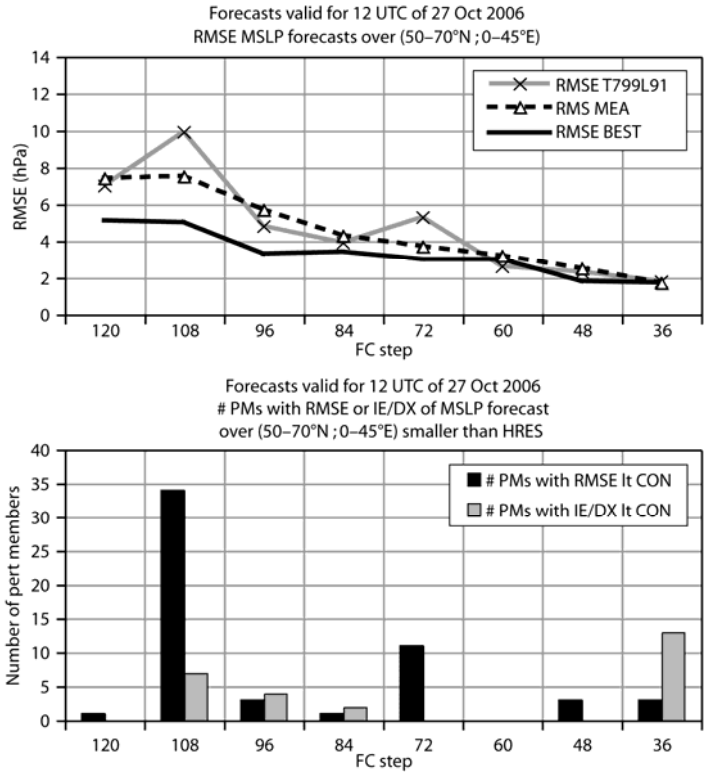

Figure 17: Storm of 12 UTC of 27 October 2006. Top panel: rmse of the high-resolution $T_L799L91$ forecast (grey line with crosses), of the EPS51 $T_L399L62$ ensemble-mean (dashed line with diamonds) and of the EPS51 $T_L399L62$ best ensemble member (solid black line). Bottom panel: number of EPS51 perturbed members with rmse (black bars) and with intensity and position error (grey bars) smaller than the EPS51 control. [rmse has been computed inside (50-70N; 0-45E).]

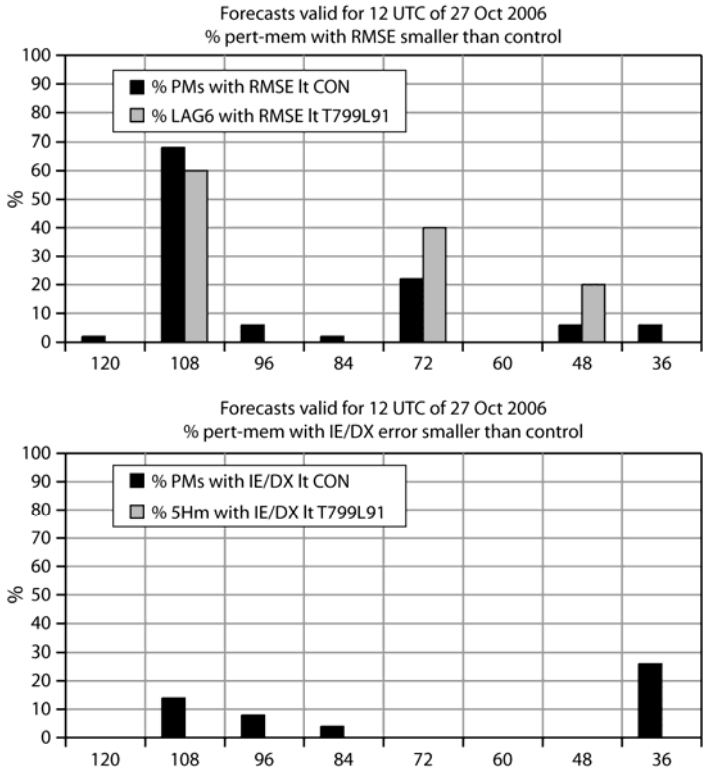

Figure 18Storm of 12 UTC of 27 October 2006. Top panel: percentage of EPS51 (black bars) and LAG6 (grey bars) perturbed members with rmse *(top panel) and with intensity and position error (bottom panel) smaller than the corresponding control forecasts (i.e. $T_L399L62$ for the EPS51 ensemble, and $T_L799L91$ for the LAG6 ensemble). [*rmse *has been computed inside (50-70N; 0-45E).]*

6. Discussion and conclusions

This work compared the performance of four ensemble systems in the prediction of synoptic scales (represented by the 500 hPa geopotential height flow), two of them of the ECMWF lower-resolution EPS forecasts and two of them based on lagged ECMWF high-resolution forecasts:

- EPS51, the 51-member ECMWF EPS, constructed using the 'control' forecast defined by the TL399L62 forecast starting from the unperturbed analysis and the 50 perturbed members with initial conditions perturbed using singular vectors, and the ensemble-mean is defined by giving the same weight (1/51) to the 51 members.

- EPS6wEM, a 6-member ensemble, constructed using the EPS control and 5 randomly-selected perturbed members, and with the ensemble-mean computed giving optimal weights to the 6 forecasts.

- LAG6, the 6-member ensemble constructed using the 6 most recent, lagged TL799L91 high-resolution forecasts (i.e. the forecasts started at the initial time, and 12, 24, 36, 48 and 60 hours earlier). In this system, the most recent TL799L91 forecast is the 'control' forecast, and the ensemble-mean is defined by giving the same weight (1/6) to the 6 members.

- LAG6wEM, constructed using the same members as LAG6, but with the ensemble-mean computed giving optimal weights to the 6 lagged, high-resolution forecasts.

The comparison included the average analysis of their performance for a 7 month period (from 1 April to 30 October 2006, 213 cases), and a synoptic analysis of their forecasts for the two most severe storms that hit the Scandinavian countries since 1969. The average performance has been assessed considering the level of ensemble spread and the accuracy of single (control, ensemble-mean and perturbed members) and probabilistic forecasts, while the single forecasts of the two storms have been assessed considering the mean-sea-level-pressure field.

Weighted ensembles have been generated using a newly introduced methodology, whereby the weight given to each member depends on its forecast quality relative to the other members, and on the average amount of analysis variability (with respect to the climate) projecting onto the phase-space direction identified by it. Forecast quality has been assessed considering one of the most commonly used measures of forecast error, the *rmse*, a choice that makes the methodology more general than if the weights were defined to specifically optimize one measure of forecast error, e.g. the *rmse* of the weighted ensemble-mean or the Brier score of the probabilistic prediction of positive geopotential height anomalies. Results have shown that for the 6-member lagged ensemble LAG6wEM, most of the weight should be given to the he most recent forecast (defined as the control forecast for this ensemble configuration), with its relative weights decreasing from 1 at forecast day 1 to ~0.5 at forecast day 10 (Fig. 1). Similarly, results have indicated that for the 6-member ensemble EPS6wEM, most of the weight should be given to the control forecast, but with a relative weight decreasing more steeply than in the lagged ensemble system, from 1 at forecast day 1, to ~0.3 at forecast day 10 (Fig. 2).

Weights have been used only in the definition of the ensemble-mean and not in the definition of the forecast probability distribution functions, since in this latter case the accuracy of the ensemble probability forecasts in the short forecast range was deteriorated significantly due to the collapse of the ensemble spread.

The comparison of the average performance of the four systems has indicated that:

- ***Ensemble spread*** - The LAG6 ensemble has a larger spread than EPS51 up to forecast day 2, and a smaller spread afterwards. If the ensemble standard deviation is compared with the error of the ensemble-mean, EPS51 has a better tuned ensemble spread than any other system (Figs. 3, 4 and 5).

- ***Control (unperturbed) single forecast*** - The error of the EPS51 control forecast (T_L399L62) is slightly higher than the error of the LAG6 control forecast, which coincides (by construction) with the T_L799L91 high-resolution forecast (Figs. 3 and 4). The difference is a measure of the impact of increasing the resolution from T_L399L62 to TL799L91.

- ***Perturbed forecasts*** - The average error of the LAG6 perturbed members is higher than the average error of the EPS51 members up to about forecast day 7 (Figs. 3 and 4).

- ***Ensemble-mean*** - The LAG6wEM ensemble-mean has a slightly smaller error than EPS6wEM and EPS51 up to forecast day 4, reflecting the fact that the LAG6 T_L799L91 control forecast has a smaller error than the EPS control forecast (Figs. 3, 4 and 5). From forecast day 5, the EPS51 ensemble-mean has the smallest error than EPS6wEM: this can be seen as an indication of the advantage of having 51 instead of 6 members.

- ***Probabilistic forecasts*** - EPS51 probabilistic forecasts have the highest RPSS, ROCA and BSS (Fig. 6). The difference between the skill of EPS51 and EPS6wEM gives a measure of the impact of ensemble size on the accuracy of probabilistic forecasts: results indicate that ensemble size have a larger impact in the medium than in the short forecast range (say after forecast day 5).

Forecasts from the 51-member EPS51 and the 6-member LAG6 ensembles have been compared also for two of the most severe storms that affected Northern Europe in January 2005 and in October 2006. The comparison of MSLP forecasts for these two cases has indicated that the T_L799L91 most recent forecast is more accurate than the EPS51 T_L399L62 control forecast, in agreement with the comparison of the 7-month average *rmse* of the two forecasts. Results have also indicated that many EPS51 members outperform the EPS51 control, but very few of them outperform the T_L799L91 most-recent forecast. If one compares perturbed and control forecasts performed with the same resolution, results show that the percentage of members outperforming its corresponding control forecast is higher for EPS51 than LAG6. This suggests that future increases in the ensemble resolution should lead to more accurate predictions of severe weather systems.

Overall, the 51-member T_L399L62 EPS has a better tuned ensemble spread, provides more skilful probabilistic forecasts than the 6-member T_L799L91 lagged ensemble for the whole forecast range, and the best ensemble-mean forecast in the medium-range. The application of the weighting methodology reduces the error of the ensemble-mean of both the EPS6wEM and the LAG6wEM ensembles, with both of them having a slightly smaller *rmse* than the EPS51 ensemble-mean up to forecast day 4, but has a very small impact on the skill of probabilistic forecasts. Results have also shown that a large membership is more important in the medium- than in the short-range: this raises the issue of whether ensemble systems should be designed to have not only variable resolution (*Buizza et al* 2007) but also variable membership, with more members used in the medium-range. This could be realized, in a variable resolution ensemble framework, by starting at the truncation time not only one but several low-resolution forecasts: research along this line is beyond the scope of this work, but it is encouraged.

It is worth mentioning two limitations of this study. The first one is that attention has been focused on the 500 hPa geopotential height and the mean-sea-level-pressure field. Results might be different if one considers surfaced weather parameters, such as precipitation, low-level temperature or wind. The second one is that results might be different if one considers higher resolution systems, e.g. lagged forecasts with a ~ 5 km resolution compared to ensemble systems with ~25 km resolution. The reader should be aware of these two limitations if different variables are considered, and/or if higher (or lower) resolution ensemble systems are considered. Furthermore, it is worth mentioning another area that could be worth investigating, more specifically whether an ensemble of smaller size, higher-resolution lagged ensembles could outperform the current ECMWF ensemble configuration. Work to investigate these three issues is encouraged.

In conclusion, since the scope of ensemble prediction is not only to predict a single, most likely scenario, but also to provide users with an estimate of forecast uncertainty, expressed in terms of probability forecasts, these results indicate that the 51-member T_L399L62 ECMWF ensemble system (EPS51) is a superior system to an ensemble system constructed using the six most recent T_L799L91 forecasts.

Acknowledgments

Philippe Bougeault, Adrian Simmons and Franco Molteni are acknowledged for useful and insightful discussions, and for their comments to earlier versions of this work. Anabel Bowen and Rob Hine are thanked for their very valuable work to improve the figures' quality.

References

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Buizza, R., 2006: The ECMWF Ensemble Prediction System. Published in *Predictability of Weather and Climate*, Cambridge University Press, 459-489.

Buizza, R., & Chessa, P., 2002: Prediction of the US-storm of 24-26 January 2000 with the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **130**, 1531-1551.

Buizza, R., & Hollingsworth, A., 2002: Storm prediction over Europe using the ECMWF Ensemble Prediction System. *Meteorol. Appl.*, **9**, 1-17.

Buizza, R., & T.N. Palmer, 1995: The singular-vector structure of the atmospheric global circulation. *J. Atmos. Sci.*, **52**, 1434-1456.

Buizza, R., & Palmer, T. N., 1998: Impact of ensemble size on the skill and the potential skill of an ensemble prediction system. *Mon. Wea. Rev.*, **126**, 9, 2503-2518.

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

Buizza, R., D. S. Richardson, & T. N. Palmer, 2003: Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man's ensembles. *Q. J. R. Meteorol. Soc.*, **129**, 1269-1288.

Buizza, R., Bidlot, J.-R., Wedi, N., Fuentes, M., Hamrud, M., Holt, G., & Vitart, F., 2007: The new ECMWF VAREPS. *Q. J. R. Meteorol. Soc.*, **133**, 681-695.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus*, **21**, 739-759.

Fleming, R. J., 1971a: On stochastic dynamic prediction. I: the energetics of uncertainty and the question of closure. *Mon. Wea. Rev.*, **99**, 851-872.

Fleming, R. J., 1971b: On stochastic dynamic prediction. II: predictability and utility. *Mon. Wea. Rev.*, **99**, 927-938.

Gleeson, T. A., 1970: Statistical-dynamical predictions. *J. Appl. Meteorol.*, **9**, 333-344.

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

Houtekamer, P. L., Lefaivre, L., Derome, J., Ritchie, H., & Mitchell, H., 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.*, **124**, 1225-1242.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.*, **102**, 409-418.

Leutbecher, M. and T. N. Palmer, 2007: Ensemble forecasting. *J. Comp. Phys.* (in press, DOI: 10.1016/j.jcp.2007.02.014). Also available as ECMWF Tech. Memo. No. 514, available from ECMWF, SHinfield Park, Reading, RG2-9AX, UK.

Lorenz, E. N., 2006: Predictability - a problem partly solved. In Palmer, T.N. and Hagedorn, R., editors, *Predictability of Weather and Climate*. Cambridge University Press.

Molteni, F., Buizza, R., Palmer, T. N., & Petroliagis, T., 1996: The new ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.*, **122**, 73-119.

Palmer, T. N., F. Molteni, R. Mureau, & R. Buizza, 1993: Ensemble prediction. *ECMWF Seminar Proceedings 'Validation of models over Europe: Vol. I'*, available from ECMWF, Shinfield Park, Reading RG2-9AX, UK.

Park, Y.-Y., Buizza, R., & Leutbecher, M., 2008: TIGGE: preliminary results on comparing and combinaning ensembles. ECMWF Research Department Technical Memorandum No. 548, ECMWF, Shinfield Park, Reading RG2-9AX, UK (also submitted to *Q. J. R. Meteorol. Soc.*).

Simmons, A., 1995: The skill of 500 hPa height forecasts. Proceedings of the ECMWF *Seminar on Predictability*, available from ECMWF, Shinfield Park, Reading RG2-9AX, UK, pp 19-68.

Swets, J. A., 1986: Form of empirical ROCs in discrimination and diagnostic tasks: Implications for theory and measurement of performance. *Psycol. Bull.*, **99**, 181-198.

Toth, Z., & Kalnay, E., 1997: Ensemble Forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, **125**, 3297-3319.

Tracton, M. S., & Kalnay, E., 1993: Operational ensemble prediction at the National Meteorological Center: practical aspects. *Weather and Forecasting*, **8**, 379-398.

Wilks, D. S., 1995: *Statistical methods in the atmospheric sciences*. Academic Press, Inc., San Diego, pp. 467 (ISBN 0-12-751965-3).

WMO CAS/JSC WGNE Report No. 20, 2005: *Report of the 20th session of the CAS/JSC Working Group on Numerical Experimentation*, 11-15 October 2004, UK MetOffice, Exeter (available on-line from the WMO WCRP web page at http://www.wmo.int/pages/prog/wcrp/).