TECHNICAL MEMORANDUM

540

The Ensemble Prediction System - Recent and Ongoing Developments

T N. Palmer, R. Buizza, M. Leutbecher, R. Hagedorn,
T. Jung, M. Rodwell, F. Vitart, J. Berner, E. Hagel[1],
A. Lawrence, F. Pappenberger, Y-Y. Park[2],
L von Bremen[3], I Gilmour[4]

Research Department

October 2007

Paper presented to the 36th Session of the SAC,
8-10 Oct 2007

[1] Hungarian Met Service, Budapest
[2] Korea Meteorological Administration, Seoul
[3] ForWind, Center for Wind Energy Research, Germany
[4] Merrill Lynch, London

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

European Centre for Medium Range Weather Forecasts
Shinfield Park, Reading, Berkshire RG2 9AX, England

The information within this publication is given in good faith and considered to be true, but ECMWF accepts no liability for error, omission and for loss or damage arising from its use.

Abstract

The ECMWF Ensemble Prediction System has been operational for 15 years. In this paper progress in the development and skill of the operational EPS is reviewed, and research in future developments is described, including use of the EPS for quantitative decision making in a variety of application sectors.

1. Introduction

The ECMWF Ensemble Prediction System (EPS) was implemented operationally 15 years ago (Palmer et al, 1993; Molteni et al, 1996) and has undergone many changes since then. In this paper we review progress in the development of the EPS, focussing on specific ongoing research, and consider possible future developments in the coming years.

The scientific basis for the EPS is summarized in Figure 1 - in a nonlinear chaotic system such as the atmosphere, predictability is flow dependent. That is to say, the amplification of inevitable uncertainties during the forecast - both in the initial state and in the prediction model - varies with the flow itself. Conceptually, the EPS is designed to provide the means to forecast such flow-dependent predictability. In practical terms, the value of the EPS is that it allows forecast users to assess quantitatively their risk to weather sensitive events, in the days ahead.

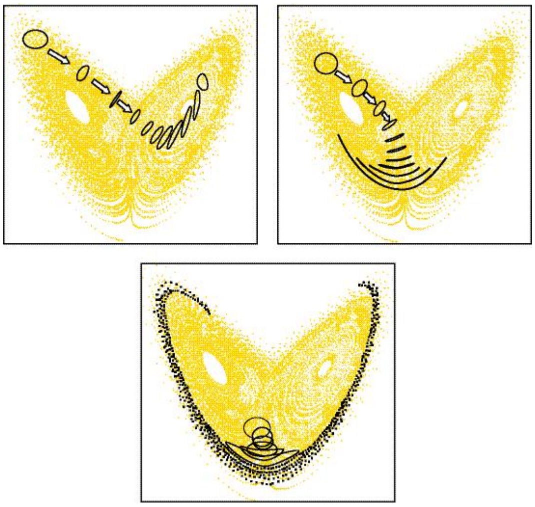


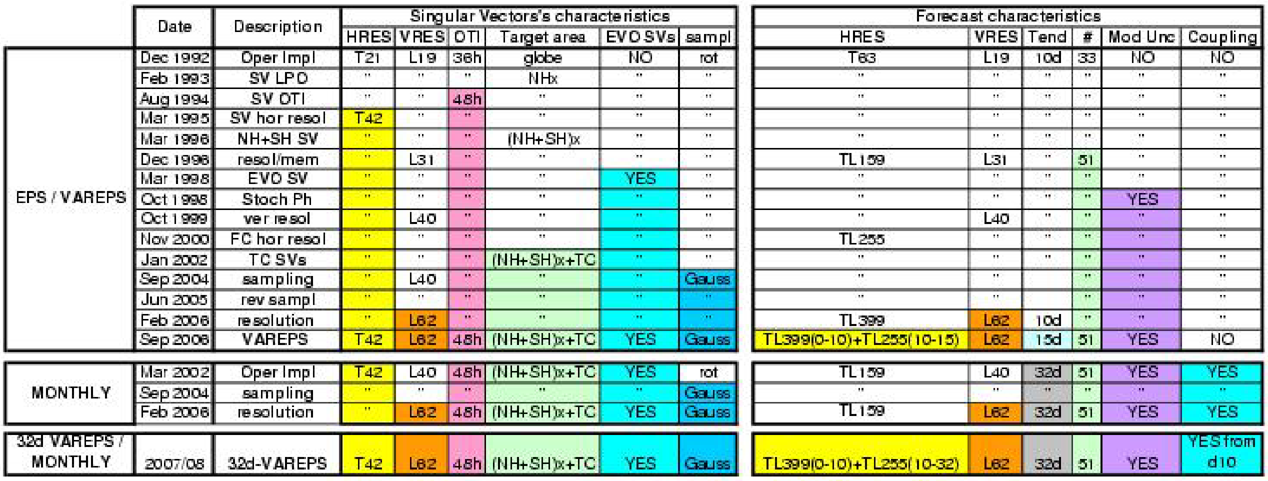
Figure 1: The scientific basis for ensemble prediction illustrated by the prototypical Lorenz model of low-order chaos, showing that in a nonlinear system, predictability is flow dependent. (a) A forecast with high probability, (b) forecast with moderate predictability, (c) forecast with low predictability.

A key element of any ensemble prediction system is its representation of these "inevitable uncertainties". In principle, these uncertainties should be represented by random samplings of known probability distributions of error ("known unknowns"). However, in practice, such probability distributions are not known. Consider, for example, potential sources of initial error. These arise not only from observation errors themselves, but also many other related sources: errors in assuming that observations are not significantly influenced by scales of motion that the forecast model cannot represent (representativity), errors in rejecting observations based on disagreement with neighbours or first-guess forecasts (quality control), errors in the methodology used to assimilate the observations into the model and errors in the forecast model used to propagate information from observations forwards and backwards in time during the assimilation process.

The fact that initial error could not be generated by some spatially uncorrelated white noise, representing a simple estimate of observation error, was well understood by those developing operational ensemble prediction systems in the early 1990s. On the other hand, the fact that there was no theoretical underpinning to construct realistic probability distributions of initial error posed practical problems. If such theoretical underpinning had existed, then initial perturbations of these early ensemble prediction systems would have consisted of random perturbations of the underlying probability distributions and the methodology would have been known by the phrase “Monte Carlo”, used in other areas of physics. The fact that the technique is known by the alternative phrase “Ensemble Forecasting” is testament to the fact that Monte Carlo approaches have never been found to work - they inevitably lead to underdispersive (usually grossly underdispersive) ensembles, due to the presence of “unknown unknowns”.

The philosophy used both at ECMWF and in the US (NMC/NCEP) to overcome this problem of “unknown unknowns” was to perturb the initial state within an estimate of the unstable phase space. In the US this approach was based on estimates of local Lyapunov vectors (Toth and Kalnay, 1993), in Europe the approach was based on estimates of local singular vectors (Molteni and Palmer, 1993: Buizza and Palmer, 1995). Whilst there has been much debate over the years as to which of these estimates is the better, the point to emphasise here is that they were both motivated by the same underlying problem of finding initial perturbations with realistic amplitude that would not produce a grossly underdispersive ensemble. It is interesting to note that these methods form the basis of operational EPSs at NCEP and at ECMWF to the present day.

Major milestones in the development of the operational EPS system since its operational implementation in 1992 are outlined in Table 1. The original EPS comprised 32+1 members, made with a T63L19 model using



	Date	Description	Singular Vectors's characteristics						Forecast characteristics					
			HRES	VRES	OTI	Target area	EVO SVs	sampl	HRES	VRES	Tend	#	Mod Unc	Coupling
EPS / VAREPS	Dec 1992	Oper Impl	T21	L19	36h	globe	NO	rot	T63	L19	10d	33	NO	NO
	Feb 1993	SV LPO	"	"	"	NHx	"	"	"	"	"	"	"	"
	Aug 1994	SV OTI	"	"	48h	"	"	"	"	"	"	"	"	"
	Mar 1995	SV hor resol	T42	"	"	"	"	"	"	"	"	"	"	"
	Mar 1996	NH+SH SV	"	"	"	(NH+SH)x	"	"	"	"	"	"	"	"
	Dec 1996	resol/mem	"	L31	"	"	"	"	TL159	L31	"	51	"	"
	Mar 1998	EVO SV	"	"	"	"	YES	"	"	"	"	"	"	"
	Oct 1998	Stoch Ph	"	"	"	"	"	"	"	"	"	"	YES	"
	Oct 1999	ver resol	"	L40	"	"	"	"	"	L40	"	"	"	"
	Nov 2000	FC hor resol	"	"	"	"	"	"	TL255	"	"	"	"	"
	Jan 2002	TC SVs	"	"	"	(NH+SH)x+TC	"	"	"	"	"	"	"	"
	Sep 2004	sampling	"	L40	"	"	"	Gauss	"	"		"	"	"
	Jun 2005	rev sampl	"	"	"	"	"	"	"	"		"	"	"
	Feb 2006	resolution	"	L62	"	"	"	"	TL399	L62	10d	"	"	"
	Sep 2006	VAREPS	T42	L62	48h	(NH+SH)x+TC	YES	Gauss	TL399(0-10)+TL255(10-15)	L62	15d	51	YES	NO
MONTHLY	Mar 2002	Oper Impl	T42	L40	48h	(NH+SH)x+TC	YES	rot	TL159	L40	32d	51	YES	YES
	Sep 2004	sampling	"	"	"	"	"	Gauss	"	"	"	"	"	"
	Feb 2006	resolution	"	L62	48h	(NH+SH)x+TC	YES	Gauss	TL159	L62	32d	51	YES	YES
32d VAREPS / MONTHLY	2007/08	32d-VAREPS	T42	L62	48h	(NH+SH)x+TC	YES	Gauss	TL399(0-10)+TL255(10-32)	L62	32d	51	YES	YES from d10

Table 1: Development of EPS since operational implementation in 1992, giving information on the types of initial and model perturbations used, the resolution of the forecast model, the number of members, and, recently, the link to the monthly forecast system.

T21 initial singular-vector perturbations only. The current EPS comprises 50+1 members made with a T399L62 model to D+10, T255L62 from D+10 to D+15, using T42 initial and evolved singular vectors and a simple stochastic representation of model uncertainty (Buizza et al,1999).

Verification is an essential component of model development, and the use of a wide range of probabilistic verification tools is imperative when developing probabilistic forecast systems. Thus, the EPS verification scheme has been recently re-designed to enable comprehensive, fast and reliable assessments of forecast performance. The development of this new verification package and related aspects, e.g. the choice of climatology, are described in Section 2. In addition, the evolution of operational EPS scores over the last decade or so is illustrated using this new verification scheme.

In Section 3 we discuss the development of the Variable-Resolution EPS system (VAREPS; Buizza et al, 2007). This has allowed the EPS to be daily extended to 15 days, running the last 5 days at lower resolution. VAREPS will also allow the monthly forecast system to be integrated as part of the EPS, thus developing a unified approach to medium and extended-range prediction at ECMWF. Such unification will allow the reforecast dataset, necessary to allow the monthly forecasts to be corrected for systematic model error, to be applied to calibrate the medium range EPS. The development of calibration techniques and their application to the EPS is discussed in Section 6. The development of the unified medium-range/monthly system is consistent with the strategic framework of the World Climate Research Programme on seamless prediction and in some sense can be seen as “closing the circle with the past”, since the first ever operational ensemble forecast systems were designed for the monthly range (Murphy and Palmer, 1986).

In Section 4 we discuss recent developments in the formulation of initial perturbations for the EPS. Ensemble data assimilation (EDA) is a tool being developed to generate flow-dependent covariance matrices for data assimilation - it may also serve value as a means to generate initial perturbations for the EPS. In Section 4 we compare EDA with singular vector-based perturbations, and discuss progress with attempts to combine EDA and singular vector perturbations. Section 4 also discusses recent work to include in the EPS moist processes in the singular vector computations, and to calculate singular vectors using a Hessian metric consistent with analysis error statistics.

Section 5 includes work on representing model uncertainty. The primary method for representing model uncertainty at ECMWF is the stochastic physics method (Buizza et al, 1999: Palmer, 2001) and recently the method has been developed to include the effects of upscale transfer of energy from sub-grid processes that would otherwise be dissipated by the parametrisation schemes (Shutts, 2005). The second method, which has been shown to have proven value in the context of seasonal forecasting (Palmer et al, 2004), is the multi-model method. The THORPEX TIGGE project provides the means to assess a multi-model EPS and compare with the stochastic physics approach.

To date, EPS products have been completely separate from the high-resolution “deterministic” forecast. Additionally, EPS products currently make no direct use of ensemble integrations from earlier start dates. Ideally the best probabilistic products should combine all possible sources of information. In Section 7 we discuss the potential benefits of “combined prediction systems” which combine these different sources of information in a statistically optimal way.

The ultimate value of a weather forecast system lies in its ability to improve weather related decision-making processes. Thus, it is of importance to consider the end users of EPS forecasts and the way the EPS is integrated into their decision-making processes. In Section 8 we discuss some of the applications to which the EPS is currently being used and will be used in the future.

Over the past 15 years the EPS has undergone major developments, and these improvements will continue to take place in the future. Section 9 discusses the outlook for the development of the EPS over the coming years.

2. Verification

In this section, various aspects of the verification of ensemble forecasts with analyses (Secs. 2.1–2.4) are described. Then the skill of the high-resolution deterministic forecast and the EPS control forecast are compared (Sec. 2.5). Finally, verification of the EPS with station observations of precipitation is presented (Sec. 2.6).

2.1. Methodology

A new package has been developed to verify ensemble forecasts against analyses. This package is now used to monitor the performance of the operational EPS in the medium-range (Section 2.2), to compare ensembles in the TIGGE archive (Section 5), and to verify research experiments. The main differences with the previously used verification are:

- use of a new long-term climatological distribution to define events and skill scores;
- all verification regions identical to the operationally-used regions for verifying deterministic forecasts;
- probabilistic scores also computed for the control forecasts and the high-resolution deterministic forecast.

2.1.1. Climatology

False skill can be obtained if a poor estimate of the climatological distribution is used in the probabilistic verification (Hamill and Juras, 2006). The climate used in the new probabilistic verification is based on ERA-40 analyses, which are expected to provide the most accurate available, consistent, long-term description of the atmosphere. The climatology consists of daily fields of the mean, standard deviation of anomalies, and quantiles of anomalies. In order to obtain good estimates globally, the climate is based on the years 1979–2001 during which satellite data constrain the analysis well in the southern hemisphere (Uppala et al. 2005). For each day of the year, statistics are based on a 61-day window centred on the day of interest. The statistics are computed with weights which are maximum at the window centre and gradually decrease to zero at days. Thus, dates contribute to the climate statistics of one day. These variable weights are superior to constant weights in terms of resolving the annual cycle and in filtering high-frequency sampling uncertainty.

2.1.2. Probabilistic scores

Ensemble verification statistics are computed on a regular 2.5 deg x 2.5 deg lat-lon grid with fields truncated to total wavenumber 63 prior to the spectral to gridpoint transformation. The operational verification of deterministic forecasts uses the same spatial resolution. It is planned to move to a higher spatial resolution, say, T159 and 1.5 deg x 1.5 deg in the future. However, preliminary tests indicate that the probabilistic scores for upper air fields like geopotential and temperature are very similar for the two resolutions (not shown). When the verification statistics are averaged over a region, weights proportional to the cosine of the latitude are used in order to approximate integration on the sphere.

A number of different verification statistics are computed which focus on different aspects of the ensemble prediction system. For binary events, the Brier Score, the Ignorance Score, also known as Logarithmic Score and the area under the Relative Operating Characteristic (ROC) are computed. By default, 7 events are considered: positive anomalies larger than 0, 0.5, 1, 1.5 climatological standard deviations and negative anomalies smaller than –0.5, –1, and –1.5 standard deviations. For multiple categories, the Ranked Probability Skill Score is computed. It is equivalent to the arithmetic mean of the Brier Scores for exceeding thresholds that separate the categories (Candille and Talagrand, 2005). Here, we use 10 climatologically equally likely categories to define the Ranked Probability Score, i.e. the Ranked Probability Skill Score is the arithmetic mean of the Brier Scores for exceeding the 1st, 2nd, …, 9th decile. Ranked Probability Scores for 5 and 20 climatologically equally likely categories have been tested also and are very close to the 10 category score. The probabilistic scores are computed for the ensemble (perturbed forecasts and control with equal weights), for the control forecast, for the high-resolution deterministic forecast and for the climatogical forecast. Skill scores are computed from monthly or seasonal scores using the score of the long-term climatological forecast as reference. In addition, rank-histogram (Talagrand diagram) statistics are computed as well as the standard deviation of the ensemble, i.e. the spread, and the RMS error and ACC of the ensemble mean and the deterministic forecasts.

2.2. Monitoring of EPS performance

In order to monitor EPS performance over the years, the ensemble has been re-scored from 1 June up to date using the new verification package (daily operational runs began on 1 May 1994). Here, a small subset of the vast range of verification statistics will be presented to summarise the overall performance. Figure 2 shows time series of the Ranked Probability Skill Score (RPSS) for temperature at forecast steps 72, 120 and 168 h. The RPSS increases by about 0.2–0.3 both for the ensemble and for the high-resolution deterministic forecast over the considered period. However, the similar changes in RPSS for EPS and deterministic forecasts do not imply similar gains in lead time because RPSS drops faster with forecast lead time for the deterministic forecasts. The asymptotic value ($t \rightarrow \infty$)of RPSS for a reliable EPS is 0 whereas the asymptotic value for an unbiased deterministic forecast is –1.

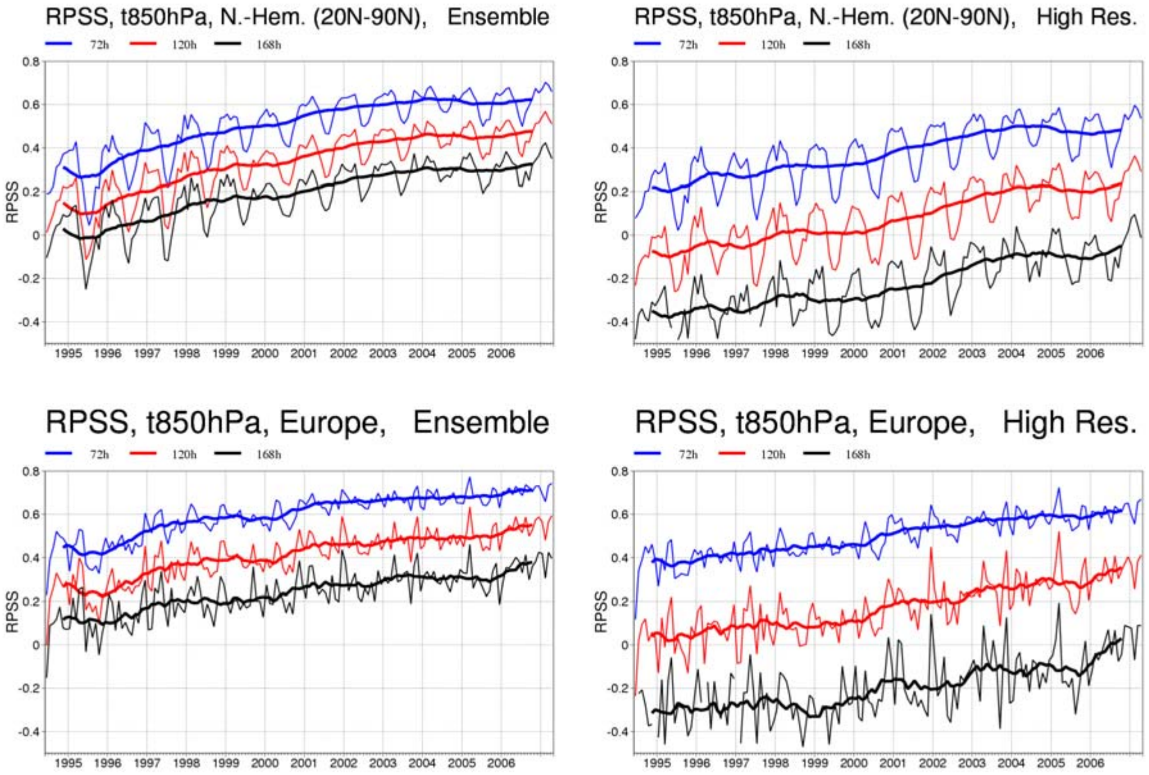


Figure 2: Time series of RPSS for temperature at 850 hPa and forecast lead-times 72, 120, 168 h for the EPS (left panels) and the high-resolution deterministic forecast (right panels). Top row: Northern Hemisphere extra-tropics, bottom row: Europe.

Instead of displaying a time series of scores, one can look at the forecast lead time at which the score reaches a threshold. For deterministic forecasts, it is a tradition to consider the lead time at which the anomaly correlation reaches 0.6. Here, we propose a similar approach for probabilistic scores. The threshold for the probabilistic scores is set to correspond to a certain lead time in a reference year, which we fix to 2006. We set the lead time to be the lead time at which the anomaly correlation of the high-resolution forecast reaches 0.6 in the reference year. This choice implies that the lead time in the reference year and the threshold vary somewhat with parameter and verification region. For the northern hemisphere extra-tropics, the lead times at which the anomaly correlation of the deterministic forecast reaches 0.6 in 2006 are 7.7 d and 7.2 d for 500 hPa geopotential and 850 hPa temperature, respectively. The corresponding thresholds for the RPSS are 0.301 and 0.297. Figure 3 shows RPSS lead-time plots for 500 hPa geopotential and 850 hPa temperature over the Northern Hemisphere extra-tropics and over Europe. Significant increases in skill can be identified at the three resolution increases of the EPS (T_L159 end of 1996, T_L255 end of 2000, T_L399 beginning of 2006). Note, however that the increases in skill at each resolution increase vary depending on parameter and verification region. The EPS gains about 2.5 to 4 days in lead time over the verification period while the deterministic forecasts gain about 1.5 to 2 days in lead time.

Sometimes, e.g. to formulate strategic goals, it is necessary to condense the evolution of the performance of a forecasting system into one single figure due to page limits etc. For these purposes, it is recommended to use the lead time of RPSS of 850 hPa temperature over the Northern Hemisphere extra-tropics for the EPS. While such a condensation of information is desirable for some users, it should be used with caution because the atmosphere is a multivariate system and users of probabilistic forecasts have many different and potentially competing objectives. Thus, when new configurations of the EPS are tested, it is necessary to look at a broader spectrum of verification measures, some of which will be discussed in the next sections.

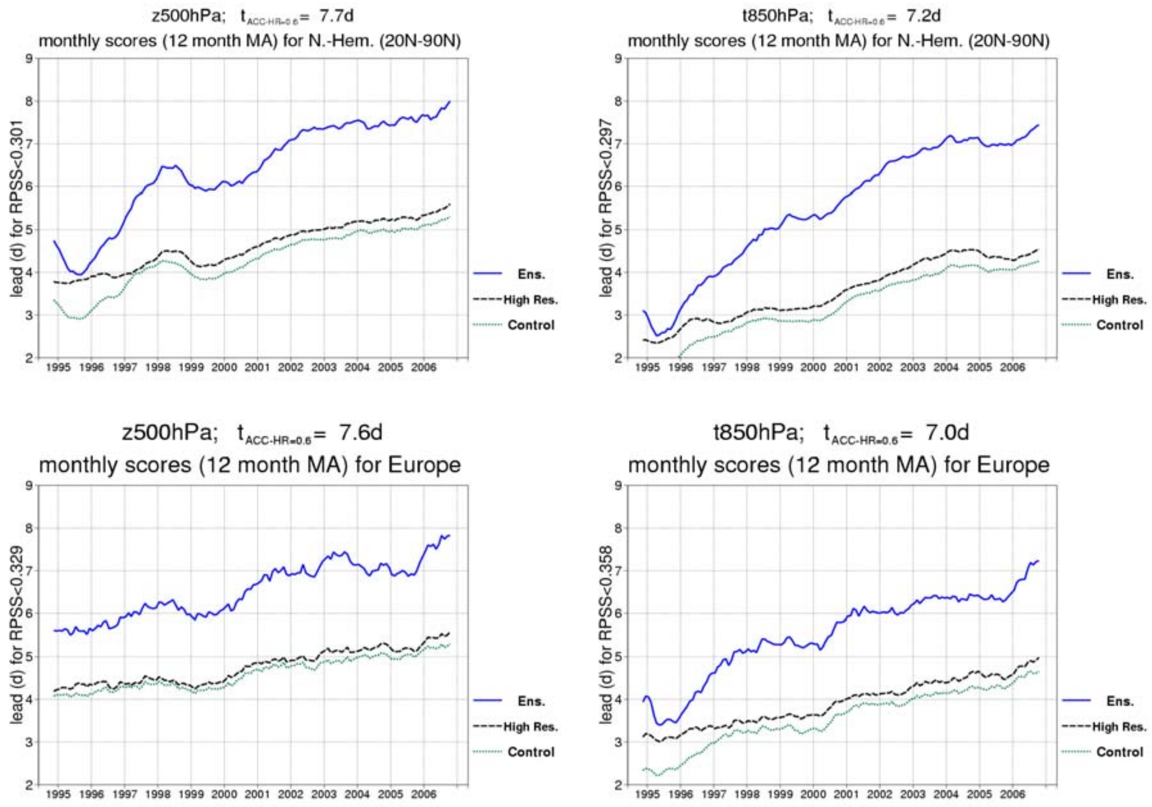


Figure 3: Lead time at which the RPSS of 500 hPa geopotential (left panels) and 850hPa temperature (right panels) reaches a threshold equivalent to the 2006 lead time at which ACC of the high-resolution deterministic forecast reaches 0.6. Top row: Northern Hemisphere extra-tropics, bottom row: Europe.

2.3. Prediction of flow-dependent error bars

An important characteristic of a reliable ensemble prediction system is that the dispersion of the ensemble matches the expected magnitude of the ensemble mean error. As perturbation error growth is flow-dependent, the ensemble dispersion varies with the flow. Leutbecher and Palmer (2007) present statistics which quantify whether the ensemble dispersion is correct. For a given forecast lead-time, all data are stratified by the predicted ensemble standard deviation. Ideally, this analysis is performed for individual grid-points so the sample size is equal to the number of grid-points in the verification region times the number of forecast start dates. The distribution is partioned in *N* equally populated bins which differ by the ensemble dispersion. Then, the ensemble mean RMS error and the ensemble standard deviation are computed for each bin. This diagnostic evaluates both the regional variations of spread, which may or may not be flow-dependent, as well as the purely flow-dependent variations. In order to focus only on the flow-dependent component, we consider a modified statistic in which data is normalised by the spatially varying climatological spread. In principle, re-forecasts could provide the estimate of the long-term climatological spread. Here, we simply use the seasonally averaged spread. Figure 4 shows the spread reliability diagrams of temperature at 850 hPa and geopotential at 500 hPa for the last winter season DJF06/07 over the Northern

Hemisphere mid-latitudes (35N¬–65N) and lead times 48 h and 120h. Generally, the relationship between spread and average ensemble mean RMS error improves with forecast lead time; it is quite close to the perfect relationship at a lead time of 5 days. However, at shorter lead times, say, 48 h the ensemble is significantly over-dispersive for the large-spread bins and under-dispersive for the low-spread bins.

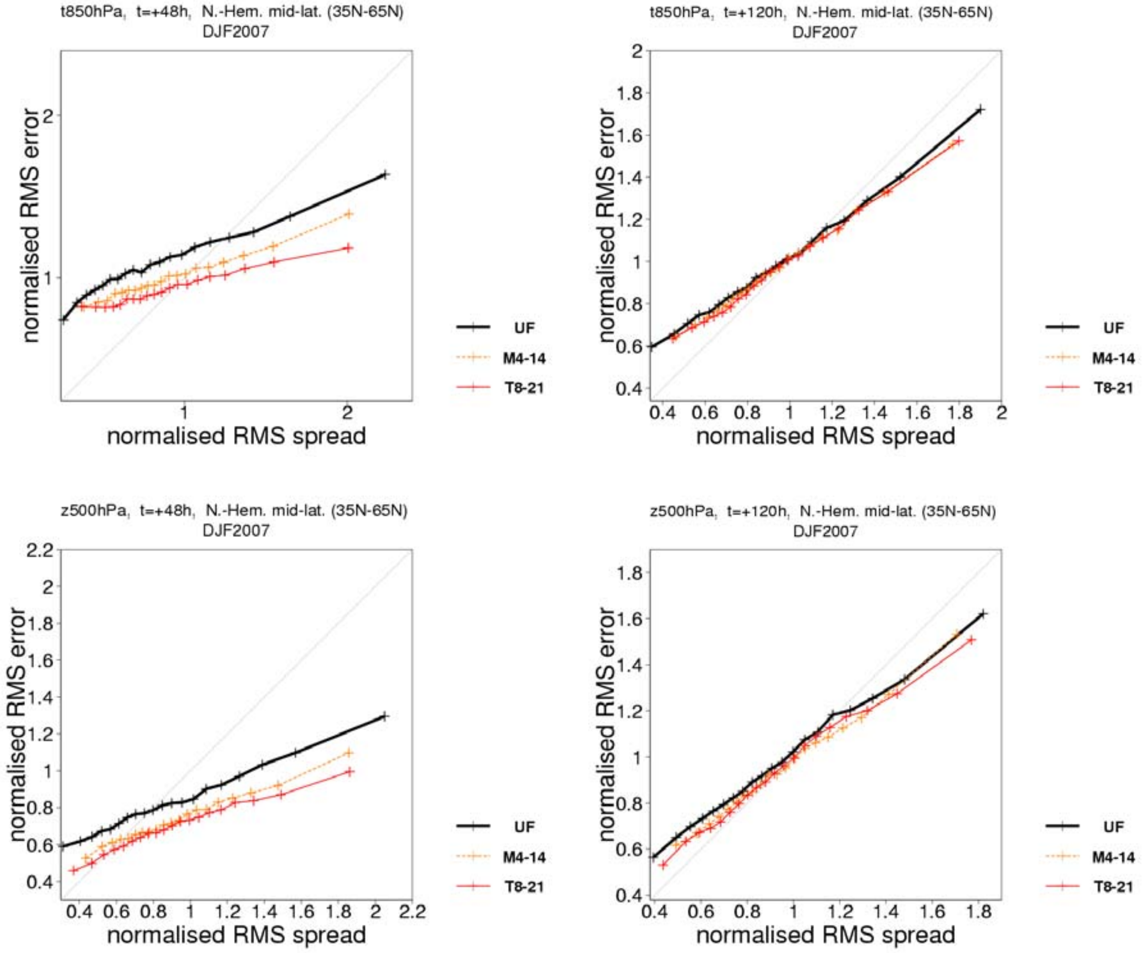


Figure 4: Normalised ensemble standard deviation (spread) versus normalised ensemble mean RMS error. The three lines correspond to the full fields (UF), and "synoptic scales" obtained by keeping only zonal wavenumber 4-14 (M4-14) or total wavenumber 8-21 (T8-21).

As part of the new verification package a scale-dependent probabilistic verification has been developed. The verification is run after spectrally filtering the forecast and verifying analyses. Similarly, climatologies are used which are computed from spectrally filtered ERA-40 analyses. The evolution of probabilistic scores of in the synoptic and subsynoptic scales is discussed by Jung and Leutbecher (2007) for the Northern Hemisphere extra-tropics and the Northern polar cap. Here, we focus on the synoptic scales in the Northern Hemisphere mid-latitudes. Synoptic scales are obtained by retaining only total wavenumbers from 8 to 21 (T8-21) or by retaining only zonal wavenumbers from 4 to 14 (M4-14). Figure 4 shows that the ensemble over-dispersion for the large-spread bins in the short-range is even more pronounced in the synoptic scales. However, at a lead time of 120 h, the synoptic scale ensemble spread is almost perfectly reliable. This spread reliability is maintained at larger lead times (not shown).

2.4. Accounting for analysis uncertainty

In general, most probabilistic verification scores are fairly insensitive to uncertainties of the verifying data, e.g. the analyses. An exception is the frequency of outliers especially in the early forecast range. The frequency of outliers is the relative number of times when the verifying analysis falls outside the range of the ensemble. The frequency of outliers is related to the rank-histogram, which is also referred to as Talagrand-diagram — a useful diagnostic to quantify the statistical reliability of an ensemble. Saetra et al. (2004) proposed a method to account for the uncertainty of the verifying analyses in the probabilistic verification by perturbing the ensemble forecasts with random numbers drawn from a distribution representing the assumed analysis uncertainty. This method has been implemented in the new probabilistic verification software. Figure 5 shows an example of how the percentage of outliers depends on whether or not analysis uncertainty is accounted for. Without accounting for analysis uncertainty, the EPS appears to have an excessive number of outliers particularly in the early forecast range. The frequency of outliers is significantly reduced when analysis uncertainty is accounted for. Work is in progress to obtain more accurate estimates of initial uncertainty.

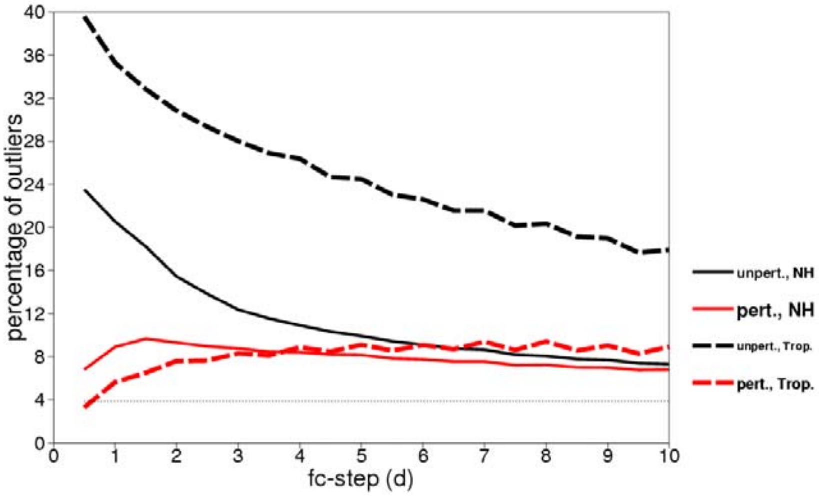


Figure 5: Percentage of outliers for 850 hPa temperature for DJF05/06. The dotted line is the expected number of outliers for a perfectly reliable 51-member ensemble.

2.5. Deterministic versus EPS control forecast

The deterministic and EPS control forecasts differ in their horizontal and vertical resolution. It has been investigated whether/how these differences affect (1) deterministic forecast skill, (2) the level of synoptic activity, which is a crucial measure when it comes to the prediction of severe weather events, and (3) systematic model error. The investigation was carried out for a two months period (February-March) of the years 1997 and 2007 in order to assess how possible differences have changed during the last 10 years.

Differences in deterministic forecast skill for geopotential height fields between the EPS control forecast and the high-resolution deterministic forecast at D+2 (upper row) and D+5 (lower row) are shown in Figure 6 for 1997 (left column) and 2007 (right column). In the most recent system the EPS control forecasts performs clearly worse than the high-resolution deterministic forecast; differences amount to about 10% in the extratropical troposphere. In the Northern Hemisphere polar region of the stratosphere differences in excess of 30% occur, which might be a result of the lower vertical resolution in the upper atmosphere and/or lower upper boundary (5hPa) used in the EPS. Since the late 1990s the gap between the EPS control forecast and the high-resolution deterministic forecast has clearly widened.

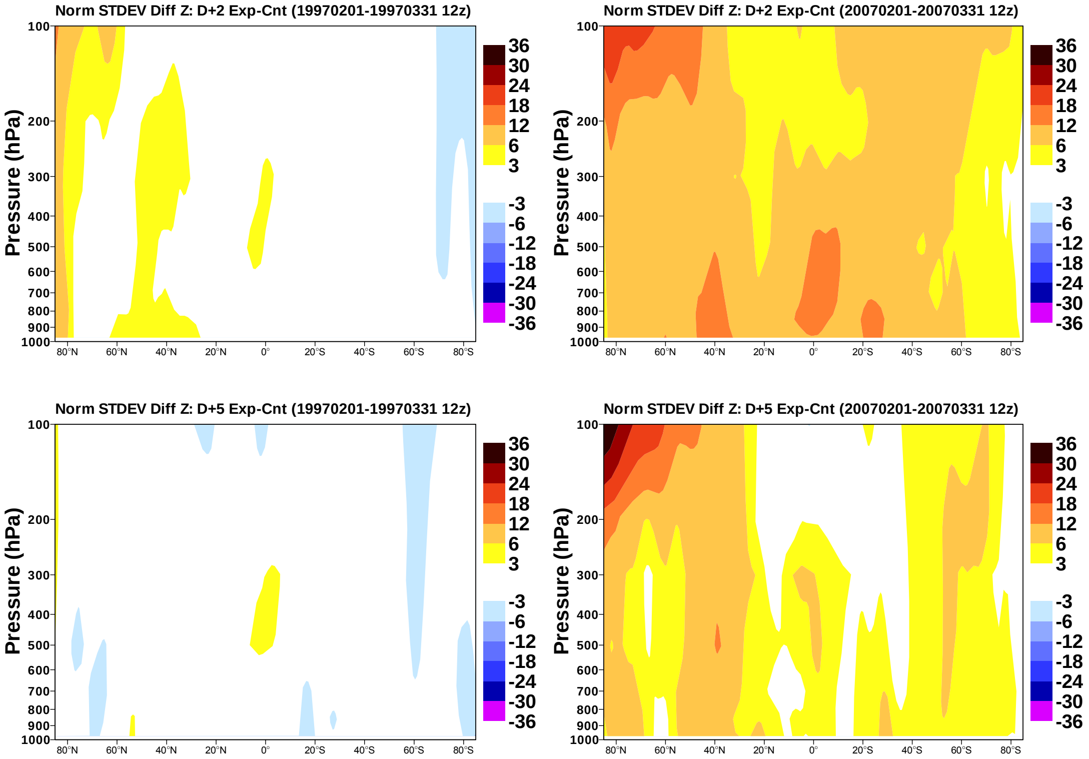


Figure 6: Zonal average of the normalized difference (in %) in the standard deviation of geopotential height forecast error between the EPS control forecast and the high-resolution deterministic forecast. Results are shown for a two-month period (February-March) in (left column) 1997 and (right column) 2007 and (upper row) D+2 and (lower row) D+5 forecast. Red shading indicates that the EPS control forecast performs worse than the high-resolution deterministic forecast.

There has been concern in the past that the lower-resolution control forecast produces lower levels of synoptic variability than the high-resolution deterministic forecast. A detailed analysis reveals that in 1997 the EPS control forecasts used to be less active than the higher-resolution deterministic forecast (not shown). However, for the current system, the EPS control forecast, if anything, seems to be more active in the tropics and the boundary layer of the extratropics. Finally, systematic error characteristics in EPS control forecasts are very similar to those for the high-resolution deterministic system; this finding is consistent with results from a recent study in which the sensitivity of climate of the ECMWF model to resolution has been studied in detail.

2.6. Prediction of precipitation at station locations

From a user's perspective, rainfall is one of the most important aspects of a weather forecast. However, it is also one of the hardest weather variables to predict. Nevertheless, there may still be value for the user in a probabilistic rainfall forecast. An individual user is generally interested in the prediction of rainfall at a point, rather than the average rainfall over a grid-box. Here, we quantify the skill of the EPS in predicting SYNOP station rainfall observations. This quantification is important for benchmarking the skill of limited area models and statistical downscaling techniques and also important from an IFS development perspective. Figure 7a shows seasonal and annual-mean Brier Skill Score (BSS) for the prediction of European SYNOP rainfall greater than 1mm in 24 hours. In all seasons and the annual mean, there is generally positive skill to around day 8. Summer is the poorest predicted season; presumably owing to the smaller-scale convective nature of summer rainfall. Figure 7b shows the same results for a 10mm rainfall event. Generally, the more extreme the event, the harder it is to predict.

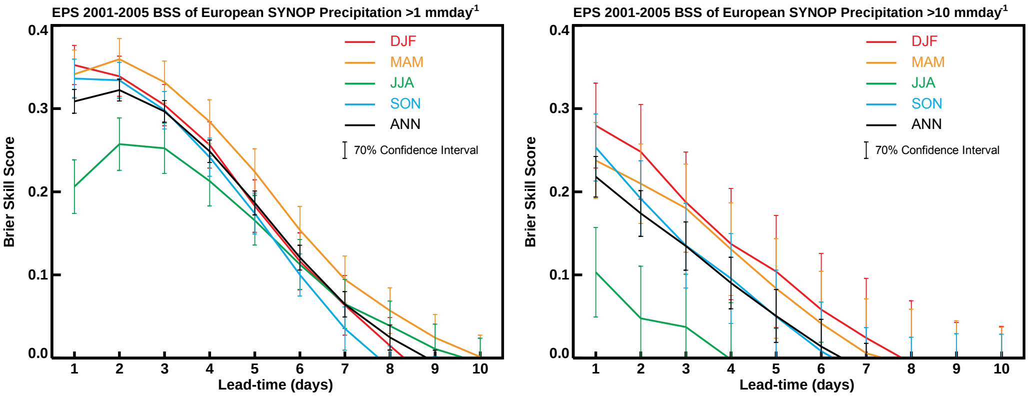


Figure 7: The seasonal means and annual mean of the daily Brier Skill Score (BSS) for the event that 24-hour accumulated precipitation (interpolated to European SYNOP stations) is (a) greater than 1mm and (b) greater than 10mm. Results are based on all operational 12UTC EPS forecasts made between 2001-2005. Bars indicate the 70% confidence interval for the estimated mean values taking autocorrelation into account.

3. The new Variable Resolution Ensemble Prediction System (VAREPS): a unified medium-range and monthly ensemble prediction system

Over the past few years, ECMWF has been working to merge the medium-range and the monthly ensemble prediction systems so that users have access to a unified ensemble prediction system providing them, at each forecast step, with the most accurate ensemble prediction forecast. The merger of these two systems has lead to the development of the new ECMWF VAriable Resolution Ensemble Prediction System (VAREPS), which has been involving three implementation phases. The first two phases were completed in 2006 (Buizza et al 2007), while the third phase is planned for the end of 2007 or the beginning of 2008:

- ***Phase 1*** - On the 1st of February 2006, the resolution of the 10-day operational Ensemble Prediction System (EPS) was increased from T_L255L40(day 0-10) to T_L399L62(day 0–10). The wave model spectral resolution was increased to 30 frequencies and 24 directions respectively without any change to horizontal resolution. To compensate for the fastest growth of the initial perturbations in the higher resolution model, the initial amplitude of the ensemble perturbations was slightly reduced (in terms of 500 hPa geopotential, e.g., by approximately 10%).

- ***Phase 2*** - On the 12th of September 2006, the T_L399L62(day 0–10) was extended to 15-days VAREPS, with a T_L399L62 resolution up to forecast day 10 and a T_L255L62 resolution between forecast day 10 and 15. The wave model resolution was set to 30 frequencies and 24 directions up to forecast day 10, and to 25 frequencies and 12 directions afterwards. Product dissemination from day 10 to day 15 started on 30 November 2006.

- ***Phase 3*** - This phase is planned for the end of 2007 or beginning of 2008, when the 15-day VAREPS will extended to one month, with a T_L255L62 atmospheric resolution and ocean coupling most likely introduced at day 10 (the detailed configuration of this system that will merge the current 15-day VAREPS and the lower resolution, T_L159L62 monthly ensemble system is currently being finalized).

In the first part of this section, the rationale behind VAREPS and the performance of the 15-day VAREPS during winter 2006/07 is discussed, while in the second part of this section the rationale behind the planned merging of the 15-day VAREPS and the monthly systems in a unified, 31-day VAREPS is discussed.

3.1. The 15-day VAREPS

VAREPS is a variable resolution ensemble forecast system where the highest resolution occurs early in the forecast range, when small-scale features show evidence of predictability. This variable resolution approach leads to a more cost-efficient use of the computer resources, with most of them used in the early forecast range to resolve the small but still predicable scales. It is worth noting that this approach to ensemble prediction had been followed at the National Centers for Environmental Prediction (NCEP, Washington) from 2000 to 2006 (see Szunyogh & Toth 2002, and the NCEP web site for more details, http://wwwt.emc.ncep.noaa.gov/gmb/ens_imp_news.html.

Technically, each 15-day VAREPS forecast is performed as a 2-leg time integration:

- *leg-1*: T_L399L62, from day 0 to day 10

- *leg-2*: T_L255L62, from day 9 to day 15

Both forecast legs are run with a coupled wave model (Saetra & Bidlot 2004): the horizontal resolution of the wave model remains unchanged (~110 km) in the two legs; however, *leg-1* runs with 30 frequencies and 24 directions, while the second leg uses 25 frequencies and 12 directions. Each *leg-2* forecast starts from a *leg-1* day-6 forecast, interpolated at the T_L255L62 resolution. High resolution wave spectra are smoothed out to the lower spectral resolution of the second leg. The 24-hour overlap period has been introduced to reduce the impact on the fields that are more sensitive to the truncation from the high to the low resolution (e.g. convective and large scale precipitation). Earlier experimentation indicated that the truncation from T_L399 to T_L255, despite being performed with the same methodology (and software) used at ECMWF to generate a low resolution analysis from a high resolution one (e.g. the TL399 ensemble initial conditions from the T_L799 analysis), can induce changes in some model fields such as divergence and vertical velocity. These changes appear to have a negligible impact on upper-air fields, but affect variables strongly linked with converging and vertical motions. In particular, they have a large impact on the precipitation field. Without the overlap period, the ensemble spread measured in terms of precipitation would decrease systematically at the truncation time, and would re-turn to more correct levels after 12 (extra tropics) to 24 (tropics) hours. The introduction of the overlap period makes the transition from the high to the low resolution smoother for fields such as precipitation, without changing substantially the upper-air fields.

Buizza et al (2007) compared the performance of an earlier version of the VAREPS system with a resolution change at forecast day 7 instead of day 10, with two constant resolution ensembles:

- *T255*: T_L255L40(day 0–13), with a 2700 second time step (this was the EPS configuration operational before 1 February 2006).

- *VAREPS*: T_L399L40(day 0–7) with a 1800 second time step and T_L255L40(day 6–13) with a 2700 second time step (note that both this configuration requires ~3.5 times the computing requirements of the T255 EPS)

- *T319*: T_L319L40(day 0–13) with a 1800 second time step (note that this configuration requires the same amount of computing requirements of the T255 EPS).

Based on results averaged over 111-cases, it was concluded that VAREPS is more skilful than a T255 EPS, with differences statistically significant in the early forecast range. Buizza et al (2007)'s analysis of two cases characterized by severe weather developments indicated that differences can be very large especially in the early forecast range and for the prediction of surface weather variables, such as mean-sea-level-pressure, wind speed, significant wave height and total precipitation. The fact that a larger impact of the resolution increase was detected in the earlier forecast range is consistent with the results found by Szunyogh & Toth (2002), who showed that also for the NCEP ensemble system the impact of increasing the forecast resolution from T62 to T126 in the first 3 forecast days gradually diminishes with forecast time. Average results have also shown that VAREPS forecasts can provide some skilful forecasts of average quantities, such as 850 hPa temperature, beyond forecast day 10. Finally, Buizza et al (2007)'s comparison of VAREPS forecasts with forecasts generated using a constant-resolution T319 EPS, which requires the same amount of computing resources as VAREPS, have indicated that VAREPS provided better forecasts in the early forecast range without losing accuracy in the longer forecast range.

As discussed in the introduction of this section, the 15-day VAREPS was implemented on the 12th of September 2006, with a truncation from a T_L399 to a T_L255 resolution at forecast day 10. Figure 8 shows the winter-average error of the ensemble-mean forecast and the ensemble standard deviation, which is a measure of the ensemble spread, computed for the 500 hPa geopotential and the 850 hPa temperature over Northern Hemisphere for the most recent two winters. For winter 2006/07, VAREPS values have been extended up to 15 days. Results indicate that the current system has a more skilful ensemble-mean than the old one, and a better tuned ensemble spread, indicated by the closer matching between the ensemble-mean error and ensemble standard deviation curves. Note in particular that the ensemble spread is slightly smaller in the early forecast range and slightly larger in the medium-range. Figure 9 shows the winter-average accuracy of the ensemble probabilistic prediction of 500 hPa geopotential and 850 temperature anomalies over Northern Hemisphere for the most recent five winters. Results indicate that the VAREPS probabilistic forecasts during winter 2006/07 achieved the highest level of skill.

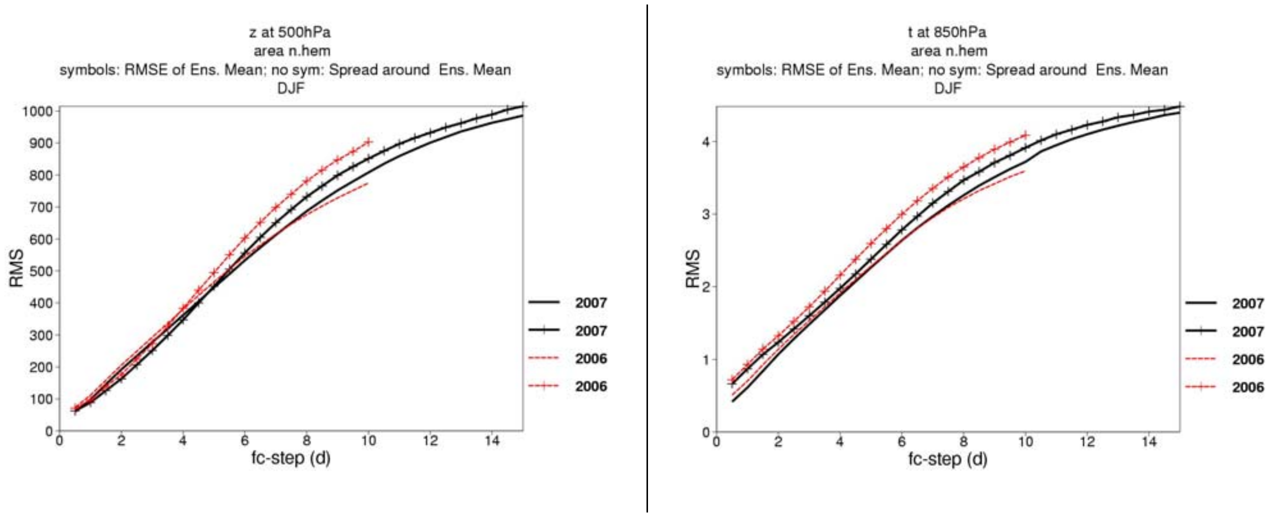


Figure 8: Winter 2006/07 (DJF, black lines) and winter 2005/06 (DJF, red lines) average root-mean-square-error of the ensemble-mean forecast (lines with crosses) and average ensemble standard deviation (full lines), computed for the 500 hPa geopotential (top panel) and the 850 hPa temperature (bottom panel) over Northern Hemisphere.

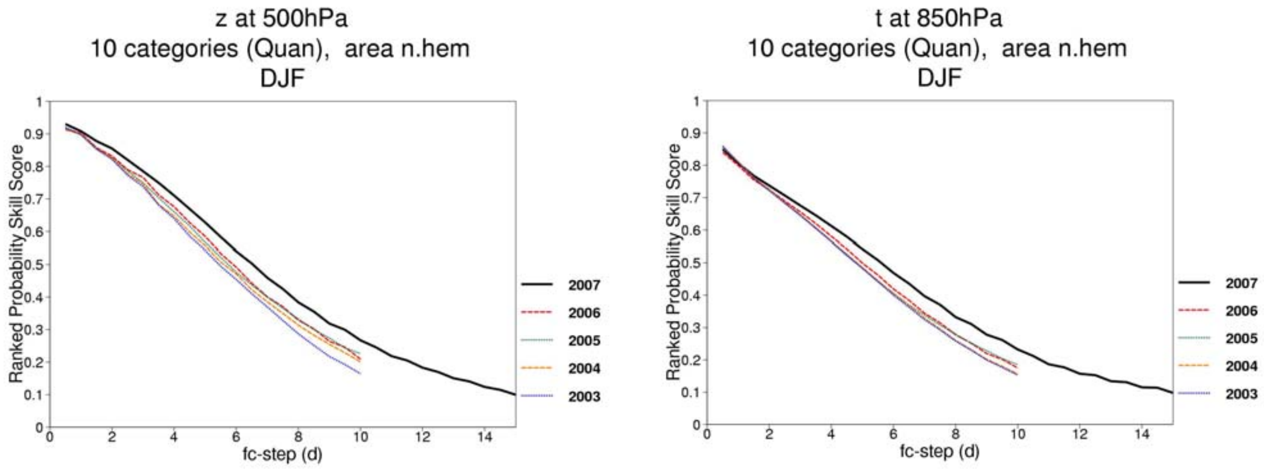


Figure 9: Ranked probability skill score (computed considering 10 categories and a climatological forecast as reference) of 500 hPa geopotential (top panel) and the 850 hPa temperature (bottom panel) over Northern Hemisphere for the latest five winters: 2006/07 (solid black lines), 2005/06 (dashed red lines), 2004/05 (dotted green lines), 2003/04 (dashed orange lines) and 2002/03 (dashed blue lines).

Since the results shown in Figure 8 and Figure 9 refer to different seasons, it is difficult to attribute the improvements to one factor only. Part of the skill improvement is due to the system change from T_L255L40 to VAREPS, as detected in the clean comparison of Buizza et al (2007). This change involved a retuning of the amplitude of the initial amplitude implemented in phase 1, which lead to a better matching between ensemble standard deviation and ensemble-mean error. Moreover, the ensemble benefited from the better quality of the T_L799L91 analysis compared to the old T_L511L61 analysis, and from the changes in the model characteristics associated with the implementations of the new model cycles in phases 1 and 2.

3.2. The planned merged 32-day VAREPS/monthly ensemble system

The current operational version of the monthly forecasting system consists of 51 integrations of 32 days fully-coupled ocean-atmosphere integrations every Thursday. This system runs independently of both the seasonal forecasting system and the 15-days VAREPS. The atmospheric horizontal resolution is T_L159 (about 120 km) with 62 vertical levels during the whole integration. The ocean model is HOPE, from the Max-Plank Institute, with the same resolution as for the seasonal forecast System 3 (about 1.4 degrees in the high latitudes and 29 vertical levels). In order to calibrate this system, a set of re-forecasts is produced the week before the real-time integrations. This set of re-forecasts consists of a 5-members ensemble 32-days integrations starting on the same day and same month as the real-time forecast but over the past 12 years.

Merging VAREPS and monthly forecasting has several advantages:

- Resources are saved by not having to run separately the first 15 days of the monthly forecasting system

- Resources are saved also by having a common set of re-forecasts (see section 6)

- The monthly forecast skill may benefit from the higher resolution in the first 10 days of the forecast

- Users' access/retrieval to the ECMWF ensemble forecasts will be simplified

The current plan consists of extending the VAREPS integrations to 32 days once a week (although higher frequency could be implemented at a later stage) with a resolution of $T_L255L62$ from day 10 to day 32. Therefore, the atmospheric horizontal resolution will be higher than in the current monthly forecasting system (T_L159) during the 32-day integrations. Ideally, the atmospheric model should be coupled from day 0 to 32, but for technical reasons (mostly because the current ocean model is not parallel) the T_L399 atmospheric and ocean models cannot be coupled during the first 10 days. Therefore the coupled ocean-atmosphere integrations will start at day 10. In order to build the ocean initial conditions for the coupled integrations at day 10, the ocean model will be run alone for 10 days forced by the fluxes provided by each VAREPS ensemble member. Coupling from day 0 will be possible when the next ocean system (NEMOVAR) is implemented, in a few years time.

The proposed version with ocean-atmosphere coupling at day 10 should be operational towards the end of 2007/beginning of 2008. This system has been tested in research mode to compare its skill to the skill of the current monthly forecasting system. The comparison has been based on 5-member ensembles run in the proposed VAREPS/monthly configuration, and in the current monthly configuration using the same model cycle from 1st January, 1st April, 1st June and 1st October 1991 to 2003 (52 cases in total). This comparison has indicated that the VAREPS system has slightly higher skill in the extra-tropics than the current monthly forecasting system. In addition this new system has been tested over some past cases. The heat wave in summer 2003 which killed more than 10000 people in Western Europe is a particularly important case for monthly forecasting. A 51-member ensemble of VAREPS/Monthly integrations with IFS cycle 32R1 has been run from 23 July 2003, along with a 5-member ensemble from 23 July 1991 to 2002. The 2-metre temperature anomalies at day 12 to 18 (3-9 August) produced by this forecast are significantly higher than those produced by the monthly forecasting system in 2003 (Figure 10). In order to check if this improvement is due to a change in model physics since 2003 or to the higher horizontal resolution in VAREPS, a re-forecast of this case with the current version of the monthly forecasting system with IFS cycle 32R1 has also been produced (bottom left panel in Figure 10). Results suggest that most of the improvement is due to the higher resolution of VAREPS. This result also suggests that this new system could be useful for early warning of this type of extreme heat wave over Europe.

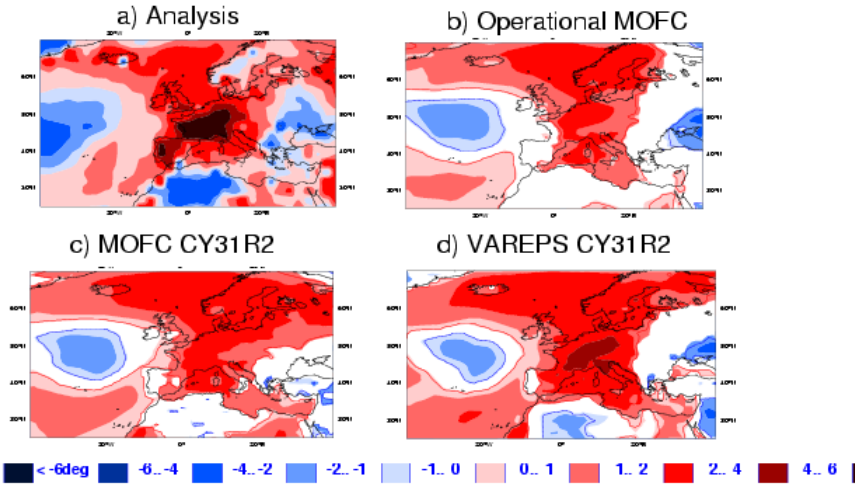


Figure 10: 2-meter temperature anomaly relative to the past 12-year climate over the period 3-9 August 2003. Red colours indicate positive anomalies, and blue colours represent negative anomalies. The top left panel shows the anomaly from the operational analysis/ERA40. The top right panel shows the forecast issued in 2003. The bottom right panel shows a re-forecast using the present version of the monthly forecasting system with IFS cycle 31r2. The bottom right panel shows a re-forecast with VAREPS/Monthly and IFS cycle 31r2 .

One justification for including an interactive ocean below the atmospheric model is the impact of the ocean-atmosphere interaction on the propagation of the Madden Julian Oscillation (MJO) which is a main source of predictability in the Tropics on the monthly time scale. In order to assess the impact of coupling in the proposed configuration of VAREPS /monthly, a 5-member ensemble of 32-day integrations of VAREPS has been performed each day from 15 December 1992 until 31 January 1993, when a strong MJO event took place (same experiment's framework as in Woolnough et al 2006 and Vitart et al 2007). The same experiment was repeated, but with persisted SSTs throughout the integration instead of coupling the atmosphere to the ocean model after day 10. The MJO is diagnosed by projecting each ensemble member into the two leading combined EOFs of velocity potential at 200 hPa, zonal wind at 850 hPa and outgoing long wave radiation computed using ERA40 (see Woolnough et al 2006 or Vitart et al 2007 for more details). The skill of the monthly forecasting system to predict the MJO is then estimated by computing the anomaly correlations of principal components 1 and 2 (PC1 and PC2) predicted by the model with the PC1 and PC2 computed from the analysis (in this case ERA40). Figure 3.4 shows the time evolution of the mean of the correlations obtained with PC1 and PC2 (the individual correlations of PC1 and PC2 look similar). During the first 10 days the scores are identical since the experiment setting is identical in both experiments. However, when the coupled ocean-atmosphere integrations start, at day 10, the VAREPS/monthly integrations display higher correlations with analysis than those obtained with 32-days of persisted SSTs. This shows that the ocean-atmosphere coupling is important at this time range (this conclusion is similar to the one reached by Woolnough et al 2006), and therefore justified in the context of VAREPS.

However, the scores displayed in Figure 11 are lower than those obtained with the current monthly forecasting system. This is due to the fact that the current monthly forecasting system is coupled from day 0 instead of day 10 for VAREPS, and the ocean-atmosphere coupling seems to have more impact on the MJO propagation than an increase in horizontal resolution (Vitart et al 2007). Therefore, VAREPS/Monthly is likely to have less skill than the current monthly forecasting system for the prediction of the MJO, but the skill over the Northern extra-tropics seem to be nevertheless higher than in the current monthly forecasting system because of the increased resolution, as discussed in the previous subsection. As mentioned, a fully coupled VAREPS system is envisaged within the term of the 4-year plan.

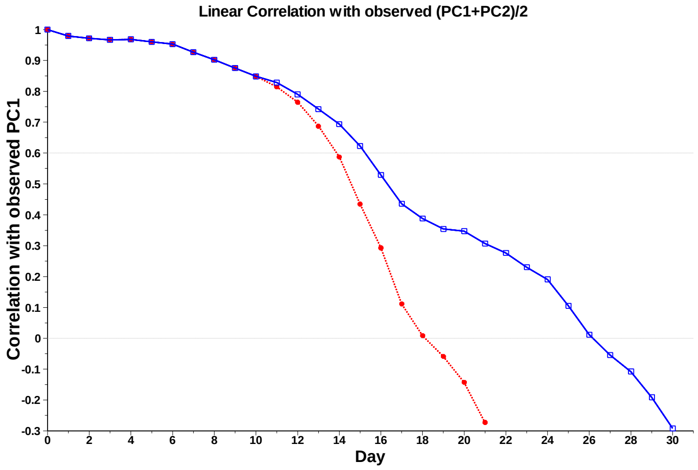


Figure 11: The MJO is diagnosed by projecting each ensemble member into the two leading combined EOFs of velocity potential at 200 hPa, zonal wind at 850 hPa and outgoing long wave radiation computed using ERA40. The skill is estimated by computing the anomaly correlations of principal components 1 and 2 (PC1 and PC2) predicted by the model with the PC1 and PC2 computed from the analysis (in this case ERA40). This figure shows the time evolution of the mean of the correlations obtained with PC1 and PC2 (the individual correlations of PC1 and PC2 look similar): the blue line represents the score obtained with VAREPS/monthly coupled after day 10, while the red line represents the score of VAREPS/monthly uncoupled (persisted SSTs).

4. Representation of initial uncertainties

4.1. Representation of initial uncertainties using ensemble of analyses

Until March 1998, the EPS initial perturbations were computed to sample instabilities growing in the forecast range, and no account was taken of perturbations that had grown during the data assimilation cycle leading up to the generation of the initial conditions (Molteni et al., 1996). To overcome this potential weakness, perturbations constructed using singular vectors growing in the past 2-days and evolved to the ensemble initial time were introduced in March 1998 (Barkmeijer et al. 1999). Initial and evolved singular vectors are computed for the Northern Hemisphere extra-tropics (30°N-90°N) and the Southern Hemisphere extra-tropics (30°S-90°S). For the two extra-tropical regions, the two sets of initial and evolved singular vectors are linearly combined and scaled so that, on average over the Northern Hemisphere, the ensemble root-mean-square (rms) spread matches the control rms error in the medium-range. By contrast, over the tropical region only initial singular vectors localized in regions where tropical depressions have been detected, are used (Barkmeijer et al 2001).

Results, documented in Barkmeijer et al. (1999), indicated that the use of evolved singular vectors has a small positive impact on ensemble skill, and reduces the percentage of times the analysis lays outside the ensemble forecast range, especially in the early forecast range. But despite these improvements, the ensemble spread near the surface and in regions not covered by the singular vectors (e.g. the tropical areas where singular vectors are computed only in few target regions where tropical depressions are detected) is still too small.

A different approach is used at the Meteorological Service of Canada (MSC): instead of using only the fastest growing perturbations, initial uncertainties are simulated by generating an ensemble of analysis, whereby each analysis is defined by data assimilation cycles run randomly perturbing the observations and using different parameterization schemes for some physical processes in each run (Houtekamer et al. 1996). The ensemble of initial states generated by the different data assimilation cycles defines the initial conditions of the Canadian ensemble system (Houtekamer & Mitchell 1998). Following Houtekamer et al (1996), Ensemble Data Assimilation (EDA) experiments have been performed at ECMWF to generate a set of perturbed analyses that could be used to improve the simulation of the initial uncertainties growing during the data assimilation cycle.

10-member ensembles of analyses have been generated using the ECMWF 4D-Var assimilation system with resolution TL255, with inner loop resolutions T_L159/T_L95 and a 12-hour assimilation window. Each analysis has been generated adding random perturbations to each observation, with random numbers sampled from a Gaussian distribution with the mean and the standard deviation defined, respectively, by the observation value and by the observation error standard deviation. To simulate the effect of model uncertainty on the analysis, the new stochastic backscatter scheme (see section 5) has also been used. Work is in progress to introduce a level of correlation in the observations' perturbations which will depend on the observation type (L. Isaksen, 2007, personal communication).

Results obtained for 20 cases (Table B), with initial conditions from the 16th of September to the 30th of October 2006, every other day (this period is limited by the time period covered by the availability of ensembles of analyses), are discussed hereafter. They are based on 51-member ensembles (resolution $T_L255L62$) run in four different configurations:

- EDA: initial perturbations have been generated using only EDA analyses
- SVINI: initial perturbations have been generated using only initial singular vectors (i.e. singular vectors growing into the first 48-hour of the forecast range)
- EDA-SVINI: initial perturbations have been generated using both sets of perturbations
- SVINI-EVO: initial perturbations have been generated using initial and evolved (i.e. singular vectors that grew during the 48-hours leading to the initial time), which is the configuration used in the operational ensemble system

The operational stochastic physic scheme (Buizza et al 1999) has been used in all four ensembles. EDA ensembles have been run using each of the 10 analyses as centre points from where 5 perturbed forecasts were started:

- in the EDA-only ensembles, from each perturbed EDA analysis 5 perturbed forecasts started, and were integrated using different perturbations in the model tendencies (as generated by the stochastic physics)

- in the EDA-SVINI ensembles, each perturbed EDA analysis was further perturbed using singular vectors-based perturbations, so that each perturbed forecasts started from a different initial state; moreover, each perturbed forecast used different perturbations in the model tendencies (as generated by the stochastic physics)

EDA-only ensembles have been run in different configurations to investigate the impact of the stochastic backscatter scheme on the spread of the ensemble of analyses. All these ensembles have also been compared also to the 4-member multi-analysis ensemble (MA EPS), an ensemble compounded of forecasts starting from the Deutscher Wetterdienst, ECMWF, MeteoFrance and UK-Met Office analyses.

The initial amplitude and spatial scale of the EDA initial perturbations are smaller than the ones based on singular vectors: compared to the standard deviation of the MA-EPS (Figure 4.1, top panels), the EDA initial perturbations have smaller scale but comparable local maxima, while the SVINI initial perturbations are more localized and have smaller local maxima. Initially, the average standard deviation of the EDA and the MA ensembles are very similar, both slightly smaller than the one of the SVINI ensemble, but already at about forecast day 2, due to the slow growth of the EDA perturbations, the difference between the standard deviation of the EDA and the SVINI ensemble becomes larger (Figure 12 and Figure 13).

It is interesting to point out that the EDA ensemble shown in Figure 12 was generated using the stochastic back scatter scheme discussed in section 5, with the addition of a linear balance equation used to generate temperature perturbations from vorticity. Results have indicated that the use of this new version of the stochastic back scatter (this change to the back scatter scheme discussed in section 5 was introduced by L. Isaksen) increases the spread of the ensemble of analyses, and brings it almost to the level of the spread detected in the MA-EPS (Figure 13, top panel). Compared to the standard deviation of the SVINI-EVO ensemble, the EDA standard deviation is, on average, almost half, but results indicate that the combination of the EDA and SVINI perturbations brings the ensemble spread very close to the SVINI-EVO one (Figure 13, middle panel). The comparison of the ensemble standard deviation with the error of the ensemble-mean (Figure 13, bottom panel) shows very clearly that the EDA ensemble have a too small ensemble spread,

The impact of replacing the evolved singular vectors with EDA-based perturbations is more evident over the tropics, where, as discussed above, the current ensemble has initial perturbations only in few regions (Figure 14, top panel). In this region, the EDA perturbations generate a larger spread than the singular vectors for the whole forecast range. This has a beneficial effect on the skill of the ensemble-mean forecast and of probabilistic predictions (Figure 14, bottom panel).

Over the extra-tropics, initially the EDA-SVINI ensemble has a slightly smaller spread than the SVINI-EVO ensemble initially, but in the medium-range the reverse is true (Figure 15, top panel). The spread difference is small, and it has a negligible impact on the skill of the ensemble-mean: the net effect is to improve the matching between the ensemble standard deviation and the error of the ensemble-mean. The fact that the EDA-SVINI ensemble has a better calibrated spread has a small but positive impact on the skill of probabilistic forecasts (Figure 15, bottom panel). In the extra-tropics, the severe spread underestimation of the EDA ensemble has a large, negative impact both on the ensemble-mean error and on the skill of the probabilistic predictions from about forecast day 3.

Although it is still too early to draw any statistically significant conclusion, these preliminary results indicate that an ensemble based on EDA-only initial perturbations would be severely under-dispersive and worse than an ensemble based on singular vectors, but that the substitution of the evolved singular vectors with EDA-based perturbations could improve the performance of the ensemble system, especially over regions only partially covered by singular vector initial perturbations, such as regions characterized by slow-growing perturbations which are not sampled by the leading singular vectors, or regions such as the tropics where singular vectors are computed only over sub-regions.

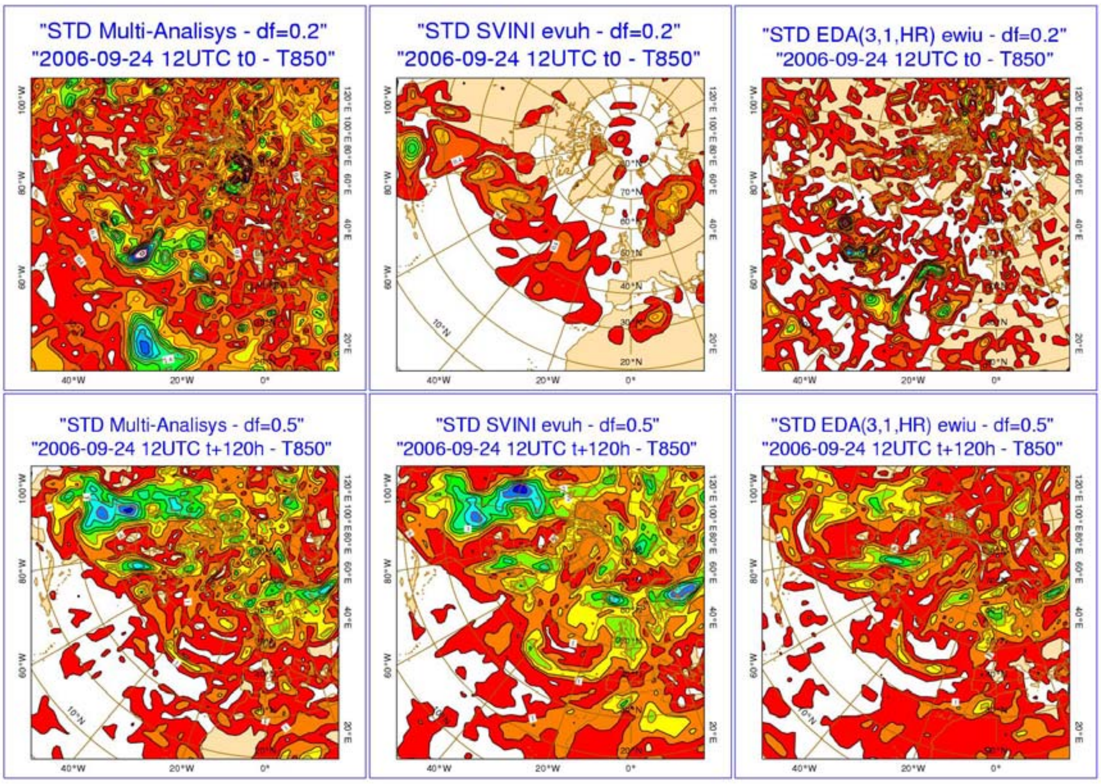

"STD Multi-Analisys - df=0.2"
"2006-09-24 12UTC t0 - T850"
"STD SVINI evuh - df=0.2"
"2006-09-24 12UTC t0 - T850"
"STD EDA(3,1,HR) ewiu - df=0.2"
"2006-09-24 12UTC t0 - T850"
"STD Multi-Analisys - df=0.5"
"2006-09-24 12UTC t+120h - T850"
"STD SVINI evuh - df=0.5"
"2006-09-24 12UTC t+120h - T850"
"STD EDA(3,1,HR) ewiu - df=0.5"
"2006-09-24 12UTC t+120h - T850"
40°W
20°W
0°
100°W
80°W
60°W
120°E 100°E 80°E 60°E
40°E
20°E

Figure 12: Ensemble standard deviation at initial time t0 (top panels) and at t+120 hours (bottom panels) of the multi-analysis ensemble (left panels), the SVINI ensemble (middle panel) and the EDA(3,1,HR) ensemble (right panels). Spread is measured in terms of 850 hPa temperature. Contour interval shading is 0.2 K for t0 and 0.5 K for t+120h.

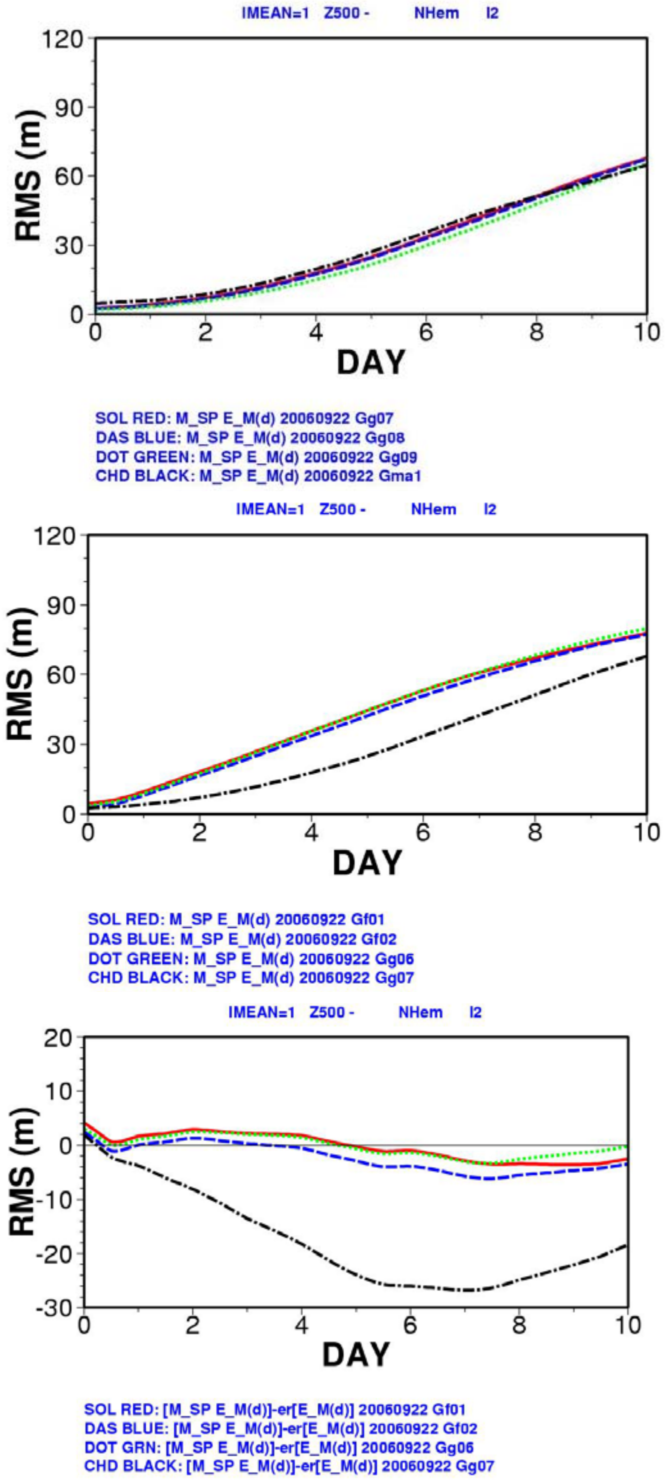
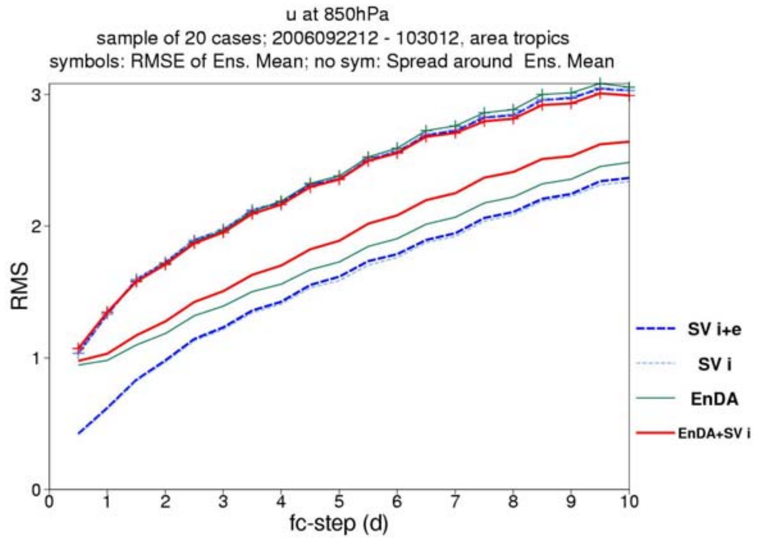


Figure 13: Top panel: standard deviation of the multi-analysis ensemble (chain-dashed black line), and of three EDA ensembles, run without stochastic back scatter (solid red line), with stochastic back scatter (dashed blue line) and with stochastic back scatter with non-linear balance omega equation (dotted green line). Middle panel: standard deviation of the EDA ensemble with stochastic back scatter with non-linear balance omega equation (chain-dashed black line), of the SVINI-EVO ensemble (solid red line), SVINI ensemble (dashed blue line) and EDA-SVINI ensemble (dotted green line). Results refer to the 500 hPa geopotential height over Northern Hemisphere. Bottom panel: as middle panel but for the difference between the ensemble standard deviation and the error of the ensemble-mean (positive/negative values indicate that the ensemble is over/under dispersive . Results refer to the 500 hPa geopotential height over Northern Hemisphere.



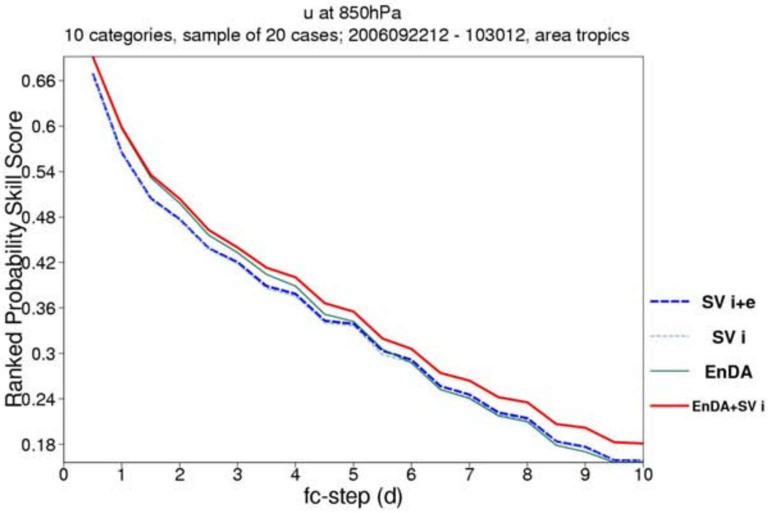
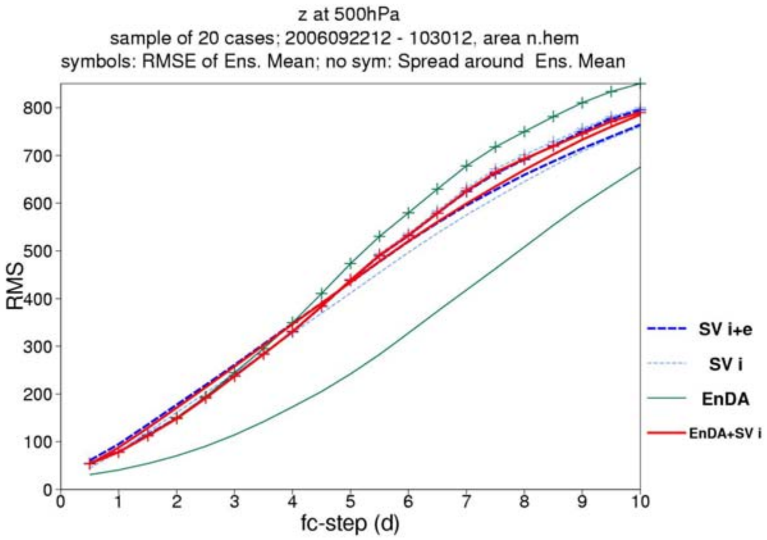


Figure 14: Top panel: standard deviation (lines without symbols) and ensemble-mean error (lines with symbols) of the EDA ensemble with stochastic back scatter with omega equation (solid green lines), of the SVINI-EVO ensemble (dashed blue lines), SVINI ensemble (dotted green lines) and EDA-SVINI ensemble (solid red lines). Bottom panel: as top panel but for the ranked probability skill score computed using a climatological forecast as reference. Results refer to the 850 hPa zonal wind component over the Tropics.



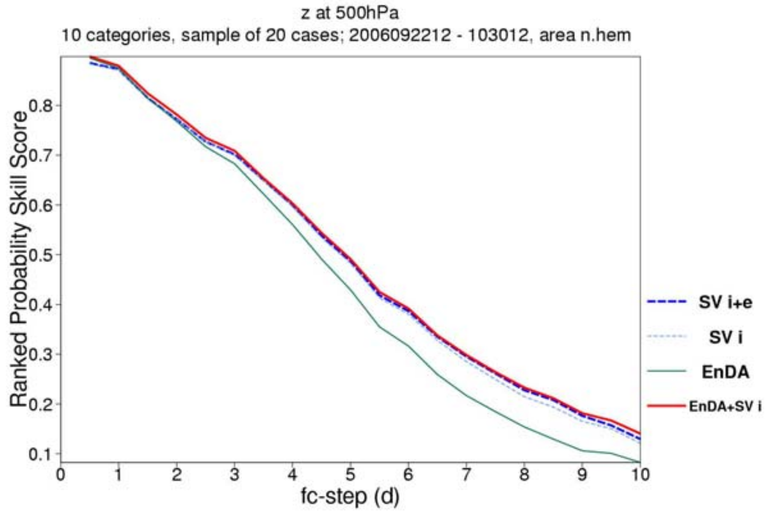


Figure 15: Top panel: standard deviation (lines without symbols) and ensemble-mean RMS error (lines with symbols) of the EDA ensemble with stochastic back scatter with omega equation (solid green lines), of the SVINI-EVO ensemble (dashed blue lines), SVINI ensemble (dotted light blue lines) and EDA-SVINI ensemble (solid red lines). Bottom panel: as top panel but for the ranked probability skill score computed using a climatological forecast as reference. Results refer to the 500 hPa geopotential over the Northern Hemisphere extra-tropics.

4.2. Representation of initial uncertainties using singular vectors

4.2.1. Moist singular vectors

The extra-tropical singular vector computation in the operational EPS uses a tangent-linear model representing perturbation dynamics of the adiabatic part of the model only (apart from vertical mixing and surface drag). Coutinho et al. (2004) studied the impact of using a diabatic tangent-linear model on the structure and growth rate of extra-tropical singular vectors. They used the parameterisations developed by Mahfouf (1999). When the large-scale, i.e. resolved, condensation is represented in the tangent-linear model, the singular vectors become smaller scale and tend to grow faster than their dry counterparts. The faster growth of moist singular vectors is consistent with results obtained by Ehrendorfer et al. (1999) with a Limited Area Model. In order to resolve the smaller scales, it appears appropriate to increase the horizontal resolution in the singular vector computation. Coutinho et al. (2004) also recommended a shortening of the optimisation time from 48 h to 24 h in order to remain in the linear regime. Such a singular vector configuration (moist, shorter optimisation time, higher resolution) has been tested in the EPS showing positive results for the prediction of some intense extra-tropical cyclones at forecast ranges of about 2 days (Walser et al. 2006). However, overall, the shortening of the optimisation time has a slightly detrimental impact on the skill of probabilistic forecasts in the present EPS configuration due to a lack of spread in the medium-range (not shown). The 24-hour singular vectors grow faster than the 48-hour singular vectors during the first day of the forecast; after 48 h, the latter singular vectors yield larger ensemble standard deviation up to the forecast range of 10 days.

Recently, more accurate tangent-linear schemes for large-scale condensation (Tompkins and Janiskova; 2004) and for parameterised convection (Lopez and Moreau; 2005) were implemented in 4D-Var (cycle 32r2). Here, we will refer to these two schemes as the new moist physics and to the schemes developed by Mahfouf (1999) as the old moist physics. Earlier experimentation (up to 2006) with the new moist physics identified cases with spurious singular vector growth in several model cycles (prior to cycle 32r2). The development of tangent-linear versions of parameterisations of physical processes requires replacing some non-differentiable functions which arise for instance through conditional statements by differentiable ones. This procedure is referred to as regularisation. The spurious growth in the singular vector computation could be linked to a lack of regularisation in the nonlinear schemes from which the tangent-linear schemes are derived. Results obtained with cycle 29r2 using partially revised regularisations were consistent with the earlier results obtained with the old moist physics, i.e. neutral impact on the skill of ensemble forecast of using the moist physics in the extra-tropical singular vector computation and a slightly negative impact due to shortening the optimisation time.

More cases with spurious growth associated with the new moist physics were identified in cycle 31r2 and further refinements of the regularisations were developed for the large-scale condensation scheme and the convection scheme. The new moist physics has been tested in the EPS in the tropical singular vector computation for 27 cases in August 2006 (T_L399L62, 50+1 member) in cycle 32r1 using the improved regularisations which went into model cycle 32r2. These experiments indicate a positive impact on the reliability of 5-day tropical cyclone strike probability forecasts (not shown). In general the impact is neutral in most cases but a few cases show a clear positive signal in the strike probability maps. As an example, Figure 16 shows the strike probability forecasts for tropical cyclone Ernesto for 27 August 2006, 00 UTC.

The track predicted by the control forecast (green) and the high-resolution deterministic forecast (black) are too far to the south. In the EPS using the old moist physics (left panel), only a few tracks follow the observed track while in the EPS using the new moist physics (right panel) a larger number of tracks follow closely Ernesto's actual track. This results in higher strike probabilities along the observed track in the configuration using the new moist physics. The singular vector subspaces computed for Ernesto with the old and with the new moist physics have a similarity index of 0.47 indicating some structural changes. In the extra-tropics, the impact of the revised tropical singular vector configuration on spread and probabilistic scores is close to neutral.

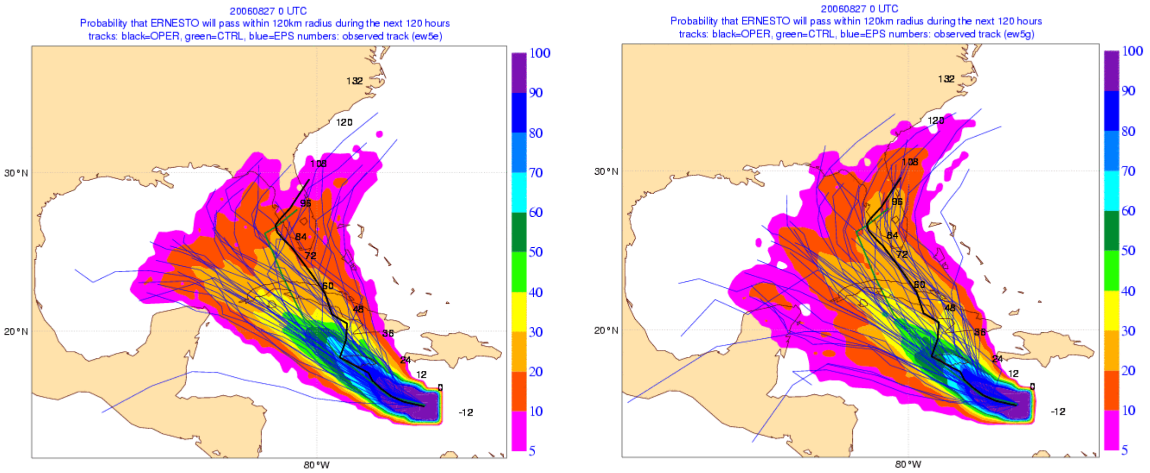

20060827 0 UTC
Probability that ERNESTO will pass within 120km radius during the next 120 hours
tracks: black=OPER, green=CTRL, blue=EPS numbers: observed track (ew5e)
20060827 0 UTC
Probability that ERNESTO will pass within 120km radius during the next 120 hours
tracks: black=OPER, green=CTRL, blue=EPS numbers: observed track (ew5g)
30°N
20°N
80°W
132
120
108
96
84
72
60
48
36
24
12
-12
100
90
80
70
60
50
40
30
20
10
5

Figure 16: Strike probability forecast for tropical cyclone Ernesto using the operational singular vector configuration (left) and the new moist physics (right) for ensemble forecasts started on 27 August 2006, 0 UTC.

4.2.2. Hessian singular vectors

Singular vector structures at initial time can depend on the choice of norm, which can also translate to differences in performance of the ECMWF EPS. Total Energy Singular Vectors (TESVs) provide a simple metric that can be considered as a first approximation of the analysis error covariance matrix (Palmer et al., 1998). Hessian Singular Vectors (Barkmeijer, 1998) can incorporate analysis error statistics directly into the singular vector computation as the Hessian of the 4D-Var cost function can be shown to be equal to an approximation of the inverse of the analysis error covariance matrix (providing the background and observation errors are uncorrelated). Therefore, Hessian Singular Vectors (HSVs) take into account the observation error statistics and also provide consistency with the corresponding data assimilation system.

In two of the following experiments, TESVs and HSVs have been used to define initial perturbations for the ECMWF EPS. A third experiment, denoted HSVinf, incorporates an inflation of the initial perturbation amplitude that renders the average ensemble spread of 500 hPa geopotential (Z500) at day 2 equal to the spread of that from the TESV experiment. The EPS was initiated with singular vector horizontal resolution T42 and forecast horizontal resolution T255 (both with 60 vertical levels) and forecasts were run up to a range of 10 days. Initial perturbations for 50 perturbed forecasts are generated from 25 leading initial singular vectors using a Gaussian sampling technique (Ehrendorfer and Beck, 2003; Leutbecher and Palmer, 2007). They are optimised for the Northern Hemisphere extra-tropics (30o - 90oN) over a 48-hour period. The comparison of ensemble performance is based on a sample of 12 cases (1 July to 24 December 2005) and results were evaluated using standard verification methods to estimate the quality of the forecasts; the root-mean-square (RMS) error, the area under the relative operating characteristic (ROC) curve and the Brier score.

Figure 17(a) shows the average Z500 RMS error of the ensemble mean and spread (around the ensemble mean) for TESV, HSV and HSVinf ensembles over all 12 cases for the Northern Hemisphere. With respect to the ensemble mean error, all three ensembles become under-dispersive after day 4. The HSVinf ensemble has similar spread as the TESV ensemble at day 2, but still significantly less spread than the TESV ensemble in the later forecast ranges. This reduced spread corresponds to a reduced probabilistic forecast skill evident in Figure 17(b). The area under the ROC curve for the TESV, HSV and HSVinf ensembles (for the positive anomalies) shows the TESV ensemble is the most skilful ensemble beyond day 6. However, at short forecast ranges (< 3 days) the HSV ensembles show more forecast skill. Between 3-6 days, all ensembles have equivalent measures of skill. The trend is qualitatively similar for the other events considered (anomalies <-1 and >+1 standard deviation).

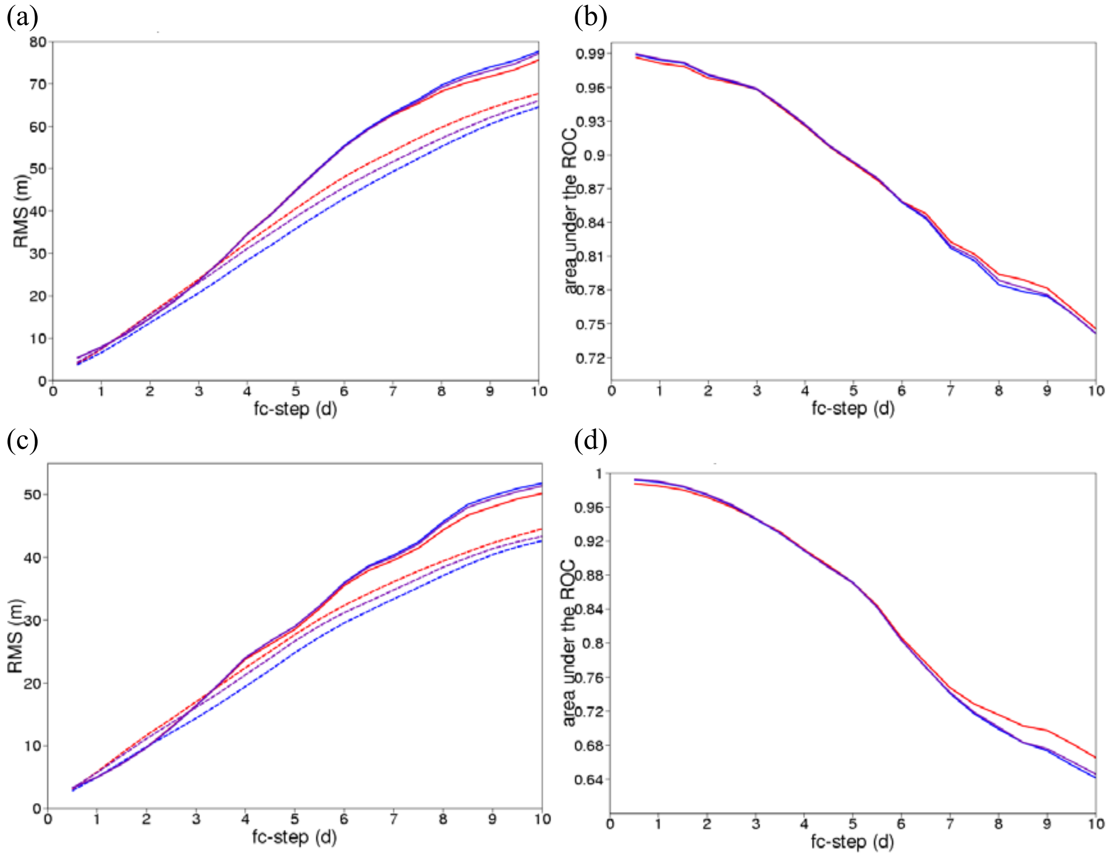


Figure 17: (a) NH (20o - 90oN) RMS error of ensemble mean (solid lines) and spread (dashed lines) of Z500 averaged over a sample of 12 cases for TESV (red), HSV (blue) and HSVinf ensembles (purple) (b) Corresponding area under the ROC curve for positive anomalies. (c) Same as (a), but limited to synoptic-scales (wavenumbers 4-12) (d) Corresponding area under the ROC curve for synoptic-scale positive anomalies.

If synoptic-scale wavenumbers are considered in the verification over the Northern Hemisphere as shown in Figure 17(c), the TESV ensemble displays larger spread than the ensemble mean error up to day 3, particularly for mid-latitudes (not shown). Under these circumstances, the reduced spread exhibited by the HSV ensemble (without the inflation) provides comparatively good spread-skill at short time ranges.

However, the spread is consistently less than the TESV spread over the entire forecast range and the HSV ensemble becomes more under-dispersive (than the ensemble mean spread) beyond day 3. The area under the ROC curve for these synoptic scales is displayed in Figure 17(d) and reveals similar results to that of Fig. 4.6(b), where the skill of the TESV ensemble is greater than that of the HSV (or HSVinf) beyond day 6 (for positive anomalies). Performing the same evaluation using positive anomalies of Brier score, the TESV and both HSV ensembles have equivalent skill up to day 3. Beyond this forecast range the TESV ensembles become marginally more skilful.

From the results and statistical scores presented here, it is evident that the TESVs perform better over the medium-range and hence are more valuable (given the reduced computational cost) in an operational EPS. The under-dispersion of the HSV ensemble in the medium-range cannot sufficiently be addressed by inflating the initial perturbation amplitude. However, as the HSVs seem to perform better than the TESVs at short ranges, this may have significant benefit for observation targeting, particularly as the full Hessian metric also includes up-to-date observational information in its formulation.

5. Representing model uncertainty

5.1. A stochastic spectral backscatter scheme (SPBS) for representing model error in ensemble forecasts

In addition to initial condition error, the probabilistic forecasting system is affected by shortcomings in the formulation of the model itself. One persistent form of error in the ensemble system is that the ensemble spread does not grow at the same rate as the ensemble mean error, leaving it underdispersive in the extended medium-range. To counteract this problem, the initial perturbations are artificially inflated to be somewhat larger than the real analysis error. Stochastic parameterizations are one way to increase the spread, by representing model error along individual ensemble trajectories, which allows one to reduce artificially-inflated initial perturbations without losing ensemble spread in the extended medium range. Thus, the inclusion of a model-error term can lead to a better representation of error growth and a more skilful ensemble.

Factors that contribute to model error are parameterization uncertainty and the misrepresentation of certain physical processes, e.g., the flow over orography and deep convection(Palmer, 2001). Since 1999, parameterization uncertainty is represented by a stochastic diabatic tendency scheme (Buizza et al., 1999), which perturbs the tendencies of the parameterized processes.

Currently, a SPectral stochastic kinetic energy BackScatter scheme (SPBS) is under development, that aims at modeling the effects of unrepresented processes and is run in addition to the stochastic diabatic tendency scheme. The SPBS is based on the notion that a part of the kinetic energy of subgrid-scale processes is not dissipated but cascades upscale and acts as a kinetic energy forcing on the resolved flow (Shutts, 2004, 2005). To model this effect, random streamfunction perturbations with prescribed spectral characteristics are generated in spectral space. To introduce non-locality in space and time, the perturbations are spatially correlated and a first-order-autoregressive process is used for their temporal evolution. To introduce flow-dependence and be consistent with the notion of energy backscatter, the random patterns are subsequently weighted with the total dissipation rate. The total dissipation rate is composed of the numerical dissipation, a contribution from the gravity/mountain drag and the dissipation associated with deep convection.

The scheme was tested in a 50-member ensemble system with a horizontal resolution of T255 and 40 model levels in the model cycle CY31R1. All results shown are for the Northern Hemisphere. The stochastic scheme increases the spread of the ensemble system, so that the initial perturbations, that are chosen artificially large to account for the under-dispersivity at the extended medium range, can be reduced by 15% without reducing the spread at the extended medium-range.

The scheme improves the relationship between spread and ensemble-mean erro (Figure 18) which ideally should grow at the same rate. Due to the reduction in the initial perturbations the scheme reduces the overdispersion of z500 for small forecast times up to day 5 (Figure 18a) and for u850 it increases the spread after day 5, bringing it closer to the curve for the ensemble mean error (Figure 18b).

The impact on a number of skillscores and variables is summarised in Table 2. For example, the scheme improves the Brier skillscore for +1 standard deviation (σ) events for all selected variables, and particularly large improvements are seen for z500 and T850. A positive impact is seen for negative as well as for positive anomalies. The improvements are especially large for skillscores that are sensitive to an underestimation of the ensemble spread, like the ignorance score (Roulston et al., 2002) and ROC area. The impact of the scheme is positive throughout with the exception of T850 where the percentage of outliers (analysis lying outside the range spanned by the ensemble) is larger up to 12h, but smaller for all other forecasting steps.

In particular, the scheme reduces the percentage of outliers of u850 in the Northern Hemisphere at day 2 from 8% to 5% (Figure 18c). The dashed line denotes the expected percentage of outliers for a perfectly reliable 51-member ensemble, which is just under 4%. In addition the typical u-shape of rank histograms, which signifies that the analysis falls disproportionately often into the most extreme bins, is significantly improved (Figure 18d). Not only are the most extreme bins less populated, but the population is much closer to the expected value in the middle bins as well. Note, that including analysis uncertainty will further reduce the percentage of outliers (Section 2.4.) and will be accounted for as soon as refined estimates of the analysis error standard deviation are available.

The impact of the scheme is similarly positive in the Southern Hemisphere and especially in the Tropical band, which is highly underdispersive for all variables, since there are no initial perturbations except around tropical cyclones and the stochastic diabatic tendency scheme alone is not able to produce sufficient spread.

Currently, different formulations for the calculation of the dissipation rate are being tested with regard to their robustness to model resolution and model version and first experiments with an ensemble at the operational resolution of T399L62 are under way.

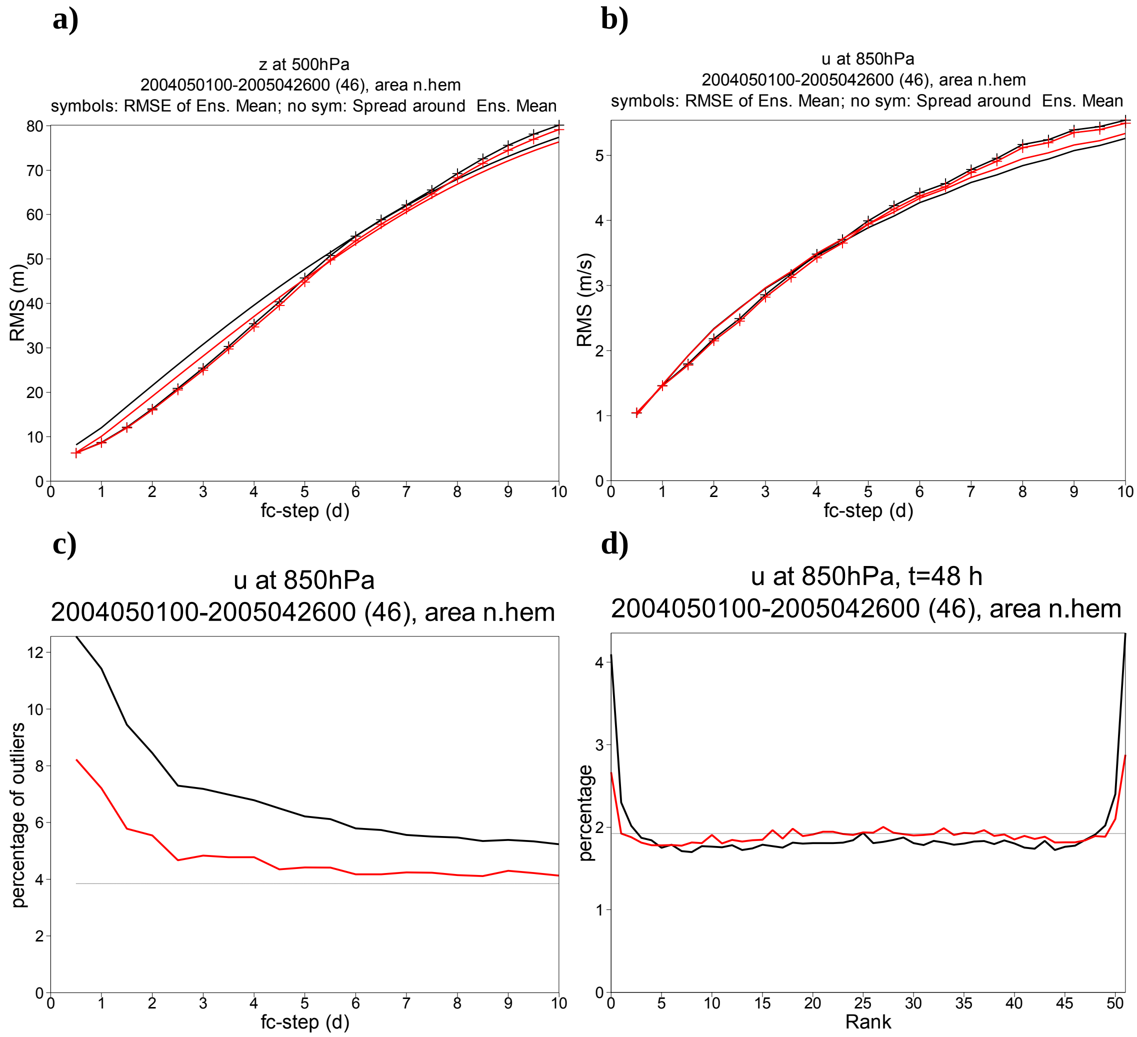


Figure 18: Ensemble mean error (solid with "+"symbols) and spread around the ensemble mean (solid, no symbols) for T255 ensemble with (red) and without spectral backscatter scheme (black) for a) z500 and b) u850. c) Percentage of outliers of u850 as function of forecast time. d) Rank histogram of u850 at a forecast time of 48h. Dashed lines in c) and d) denote the expected number of outliers for a perfectly reliable 51-member ensemble. All measures are computed for the Northern Hemisphere.

Table 2: Summary of the impact of the stochastic backscatter scheme on ensemble spread and skill for z500, u850 and T850 in a T255 50 member-ensemble with reduced initial perturbations in the Northern Hemisphere. The first and second columns list the skill score and, if applicable, the verification thresholds, e.g., the Brier skillscore is computed for +1 standard deviation (σ) and -1 standard deviation events. A positive impact for all forecast times is denoted by +; a strong positive impact for all forecast times (subjective measure) by ++. >/< denotes a positive/negative impact for short forecast times, but a negative/positive impact for longer forecast times.

Northern Hemisphere	Threshold	Z500	U850	T850
Spread		++	+	>
Brier skill score	> 1σ	++	+	++
	< 1σ	++	+	++
ROC area	> 1σ	++	+	++
	< 1σ	++	++	++
Rank probability skill score		+	++	++
Ignorance score	> 1σ	++	++	+
	< 1σ	++	++	+
Percentage of outliers		++	++	<
Rank histogram at 48h		++	+	+

5.2. TIGGE (the THORPEX Interactive Grand Global Ensemble project)

As of Feb 2005 about 13 meteorological operation and research centres over the world operate their own global ensemble prediction system (WMO 2005a). These global ensemble systems differ in the resolution of the analysis and forecast models, in the schemes used to parameterize physical processes, in the ensemble size and in the forecast length. Membership and resolution are dictated by computer resources availability, and communication bandwidth: up to now, all centres have opted for a membership between 20 and 50, since no strong evidence has been produced so far that a larger size would bring substantial impact on the accuracy of the ensemble probabilistic forecasts.

One of the main goals of the THORPEX Interactive Grand Global Ensemble (TIGGE), a key component of the WMO's THORPEX World Weather Research Programme, is to explore the benefits that could be achieved if all these ensemble forecasts were to be shared in real time, and used to generate multi-model ensemble products. Such an approach would possibly improve the simulation of model uncertainty, which so far has been proved to be one of the main weaknesses of the current systems (Buizza et al 2005). Other key objectives of TIGGE are: to enhance collaboration on the development of ensemble prediction, internationally and between operational centres and universities, and to develop new methods of combining ensembles from different sources and of correcting for systematic errors.

To achieve these objectives, it is necessary that the ensemble forecasts, observation data, and other existing datasets be shared between TIGGE users. At the first TIGGE workshop held on 1-3 March 2005 at ECMWF, three centres (CMA, ECMWF, and NCAR) indicated willingness to become TIGGE Data Centres, and 11 forecasting centres expressed interest in providing global EPS data to the TIGGE archives (WMO 2005b).

As of 20 June 2007, 5 centres are sending real time data to the 3 data centres. Table 3 is the summary of the EPSs of the confirmed 10 data providers.

At ECMWF, the first TIGGE data started been archived on the 1st of October 2006, with 3 providers sending their ensemble forecasts. Another 2 centres have joined since then, and 2 further centres are currently testing sending procedures as of 20 June 2007 (see the last column in Table 3).

Table 3 Key characteristics of the TIGGE global ensemble prediction systems.

Centre	Initial pert method (area)	Model error simul	Horizon res (tigge)	Vert res	Fcst length (days)	# pert mem	#runs per day (UTC)	# mem per day	Data From*

ECMWF	SVs(globe)	YES	TL399(N200**)
			TL255(128)
				62
				62
					0-10
					10-15
						50	2(00/12)	102	1 Oct 06

UKMO(UK)	ETKF(globe)	YES	(1.25x2/3 deg)		15	23	2(00/12)	48	1 Oct 06
JMA(Japan)	BVs (NH+TR)	NO	T106L40(1.25deg)	40	9	50	1(12)	51	1 Oct 06
NCEP(USA)	BVs(globe)	NO	T126(1.00 deg)		16	20***	4(00/06/12/18)	84	11 Nov 06
CMA(China)	BVs (globe)	NO	T213(0.5625deg)	31	10	14	2(00/12)	30	15 May 07
MSC(Canada)	Analyses cycl (globe)	YES	TL149		16	16	2(00/12)	34	On test
KMA(Korea)	BVs (NH)	NO	T213	40	10	16	2(00/12)	34	On test
CPTEC(Brazil)	EOF-based (40S:30N)	NO	T170	42	15	25	2(00/12)	52	
BMRC(Australia)	BVs(TR)	NO	TL119	19	10	32	2(00/12)	66	On test
Meteo France	BVs(local)	NO	TL358		2.5	10	1(18)	11	On test

* Based on the ECMWF TIGGE archive

** Reduced Gaussian grid

*** Start dates 18 UTC 27 Mar 2007. 15 members from 12 UTC 14 Dec 2006-12 UTC 27 Mar 2007. 11 members from 00 UTC 1 Nov 2006 - 06 UTC 14 Dec 2006.

Before testing combination methods, it is important to assess the relative skill of the available ensemble systems: preliminary results from the comparison of the skill of the four available ensemble forecasts (ECMWF, JMA, NCEP and UK MetOffice) based on 62 cases (28 March to 28 May 2007) are discussed here (this period includes forecasts for the first 62 dates since NCEP increased membership from 14 to 20, and for which the ensembles can be verified). These results are based on all the ensemble forecasts starting at 12 UTC of each day only. Geopotential height at 500 hPa (z500hPa) and temperature at 850 hPa level (t850hPa) forecasts from the ECMWF, UKMO, JMA, and NCEP ensembles have been verified, each against its own analysis (defined as the control forecast at the initial time) on a regular 1.25 degree grid (the finest common grid).

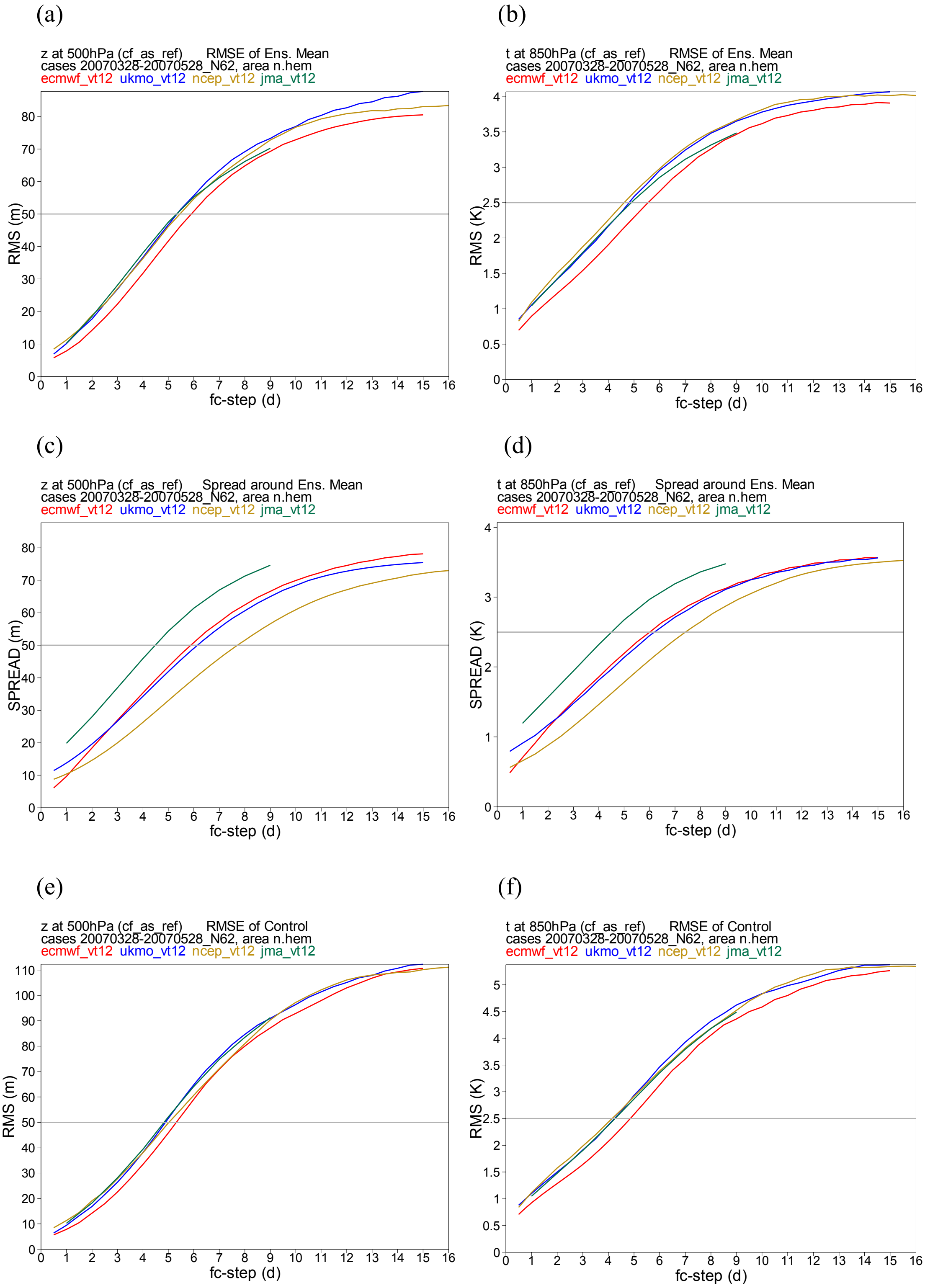


Figure 19(a) and(b) Root-mean-square-error of the ensemble-mean forecast; (c) and (d) spread around ensemble-mean; (e) and (f) root-mean-square-error of the control forecast for the 500 hPa geopotential (left panels) and 850 hPa temperature (right panels) averaged over the Northern Hemisphere for the period of 28 March - 28 May 2007. Red: ECMWF, blue: UKMO, orange: NCEP, green: JMA.

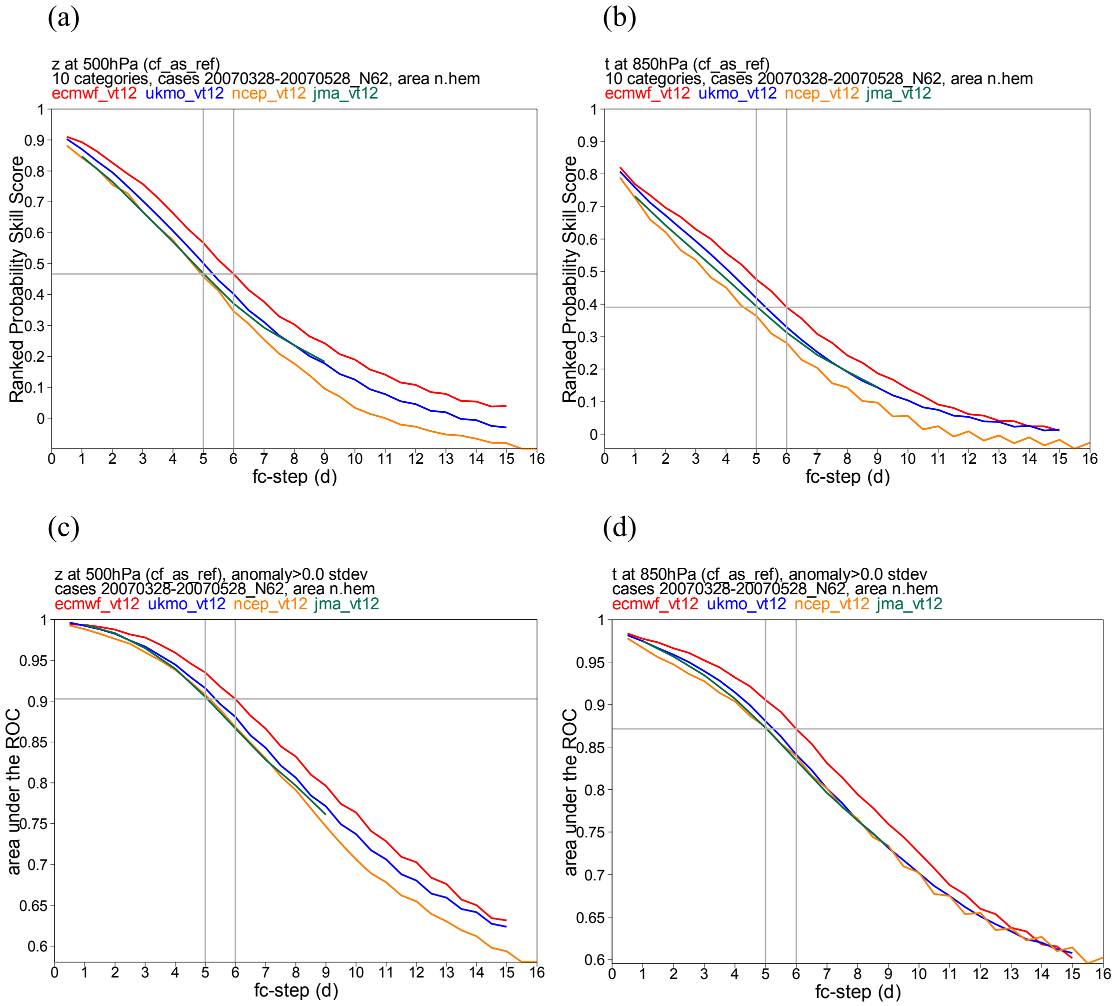


Figure 20a) Ranked probability skill score computed considering 10 categories and a climatological forecast as reference for the 500 hPa geopotential; (b) as (a), but for the 850 hPa temperature;(c) area under the ROC for positive anomalies of the 500 hPa geopotential; (d) as (c), but of the 850 hPa temperature over the Northern Hemisphere for the period of 28 March - 28 May 2007. Red: ECMWF, blue: UKMO, orange: NCEP, green: JMA.

Figure 19 shows the average error of the ensemble-mean (EM) and the ensemble standard deviation, and the error of the control forecast (CF) computed for z500hPa and t850hPa over the Northern Hemisphere (note that the scale of the ordinate axis in the ensemble-mean and the control forecasts are different). Results indicate that the ECMWF EM and CF have the lowest error for both parameters, and that the matching between the EM-error and the ensemble standard deviation is best for the ECMWF ensemble. Considering ensemble spread, JMA shows the largest and NCEP the smallest values, while ECMWF and UKMO show almost identical values after forecast day 2. Note that the relative performance of JMA-EM improves after 6 day, probably due to the smoothing linked to the large ensemble spread. Similar consideration can be drawn considering probabilistic forecasts (Figure 20), with the ECMWF ensemble showing the highest skill. At day 6, e.g., the difference in RPSS and ROCA between ECMWF and the second best system amounts to ~18 hours in predictability.

Since these results cover only 60 dates of a transition season, it is difficult to draw any strong statistically significant conclusion, and to identify clear reasons for the different relative performance of the four systems. Work is continuing to increase the number of cases to at least one season, and to extend the comparison to include other ensemble systems. Moreover, work will start on the development of a combination method that could generate combined products using all available TIGGE data (162 members at 12 UTC, with 315 members available per day). Key issues that will be investigated include: which is the best way to combine these forecasts to get more skillful prediction than a single good EPS? Is it better to use all the available members or only some best members? These and other related questions will be addressed to get optimally combined ensemble system with the ensemble data from 7 forecast centre (including Korea, China and Canada).

6. Calibration

6.1. Background

It is well known that post-processing or calibration methods can improve the skill of forecasts from direct model output (Wilks, 2006), and in fact such methods have been used already for a long time by individual Member States as can be seen e.g. in the reports on 'Verification of ECMWF products in Member States and Co-operating States' ("The green book"). Until now, the post-processing applied to ECMWF direct model output has been based on operationally available training datasets, which are either relatively short datasets or - if they cover longer times - are inconsistent datasets containing data from different model cycles or even different model resolutions. More recently it has been suggested that such calibration techniques can lead to even greater improvements if large datasets of consistent re-forecasts are available (Hamill et al., 2004, Hamill and Whitaker, 2007). Mainly two factors have prevented ECMWF from previously embarking on such a re-forecast programme: First of all, it had to be examined whether the level of improvements, which had been demonstrated only for forecasts with relatively low quality, could also be achieved for the higher quality ECMWF forecasts; and secondly, the significant computational resources associated with the production of large re-forecast datasets on an operational basis had to be found.

In order to address the first concern, some studies on the potential benefits of calibrating ECMWF EPS forecasts have been performed, and the results of these studies are summarized in this chapter. The second concern is being resolved with the planned implementation of the unified VarEPS-monthly forecast system (see Section 3). This new system offers the possibility of using the re-forecast dataset (originally produced only for the monthly forecast bias correction) additionally for calibrating the medium-range EPS forecasts. Hence, when the unified system will become operational, available re-forecasts can be used by Member States for their individual calibration techniques.

Here we describe the experiments performed and calibration methods used to assess the potential benefits of a larger re-forecast dataset, and present preliminary results of these studies offering some suggestions for the most efficient setup of a unified re-forecast suite at ECMWF.

6.2. Datasets and methods

6.2.1. Training datasets

The re-forecast dataset produced for this study covers the period 1 September to 1 December, with one re-forecast per week, i.e. 14 cases or startdates are available (01/09, 08/09,…,01/12). For each of these startdates, 20 re-forecast years covering the period 1982-2001 were run. The re-forecasts were produced with the model cycle and setup operational during Sep-Dec 2005 (CY29r2, T255), except that the initial conditions were taken from ERA-40 re-analysis. Furthermore, the re-forecast ensemble consists of only 15 members (14 perturbed + 1 control) instead of the full set of 51 members. Ideally, the re-forecast dataset should contain the same number of members, however, since the production of such a full set of re-forecasts is currently not affordable in an operational setting, this option was not considered in this study and only the maximum affordable number of members were produced for this test re-forecast dataset.

6.2.2. Calibration methods

In order to assess the level of improvements achievable by calibrating the direct model output, a number of calibration techniques, from simple bias correction to more advanced regression methods have been tested. Here we will present results achieved with two different methods:

a) Bias correction: In this simplest calibration scheme, the long-term systematic error *b* is estimated from the mean difference between the ensemble mean forecast $\overline{x}_{ens}$ and the observations *o* in training dataset

$$b = \frac{1}{N}\sum_{i=1}^{N} \overline{x}_{ens,i} - o_i \qquad (1)$$

with: *i*: training cases (i=1,…,N)

The correction factor *b* is applied to every ensemble member of the ensemble to be calibrated, i.e. the ensemble spread is unaffected by this procedure and only the ensemble mean is corrected. The systematic error is estimated individually at each location and for every lead time.

b) Non-homogeneous Gaussian Regression (NGR): In this calibration scheme the probability *P* for a future event (e.g. temperature below or equal a certain threshold *q*) is set according to the equation:

$$P(T \le q) = \Phi\left[\frac{q-(a+b\overline{x}_{ens})}{\sqrt{c+d{s_{ens}}^2}}\right] \qquad (2)$$

with: *Φ*: CDF of standard Gaussian distribution

s_{ens}: ensemble spread

a, *b*, *c*, *d*: regression coefficients

The parameters *a*, *b*, *c*, *d* are fit iteratively by minimizing the Continuous Ranked Probability Skill Score (CRPSS) for the training dataset. The NGR method takes into account existing spread-skill relationships by modelling the error term as a function of the ensemble spread. Though the shape of the final estimated forecast PDF is always Gaussian, its width is a function of the forecast ensemble variance and thus non-homogenous with respect to different forecasts to be calibrated. More details regarding the NGR method can be found e.g. in Gneiting et al. (2005).

6.3. Results

The calibration of near surface weather parameters offers the greatest scope for improving forecasts (Hamill and Whitaker, 2007). Therefore, here we present results for calibrating 2m temperature forecasts at 100 stations mainly located in central Europe. The global forecast fields are interpolated to these locations using the ECMWF 12-point interpolation scheme, and synoptic observations from the World Meteorological Organization (WMO) Global Telecommunication System (GTS) are used as verification.

The comparison of the bias-corrected and NGR-calibrated results with the direct model output demonstrates the great potential for improvements (Figure 21). Though the main improvement is achieved by the simple bias correction, the NGR method can add significantly to that (Figure 21a). During the early forecast range the bias correction improves the forecasts by up to three days, and the NGR method can add another one or two days to that. This extra improvement is reduced to about half a day for the later forecast range.

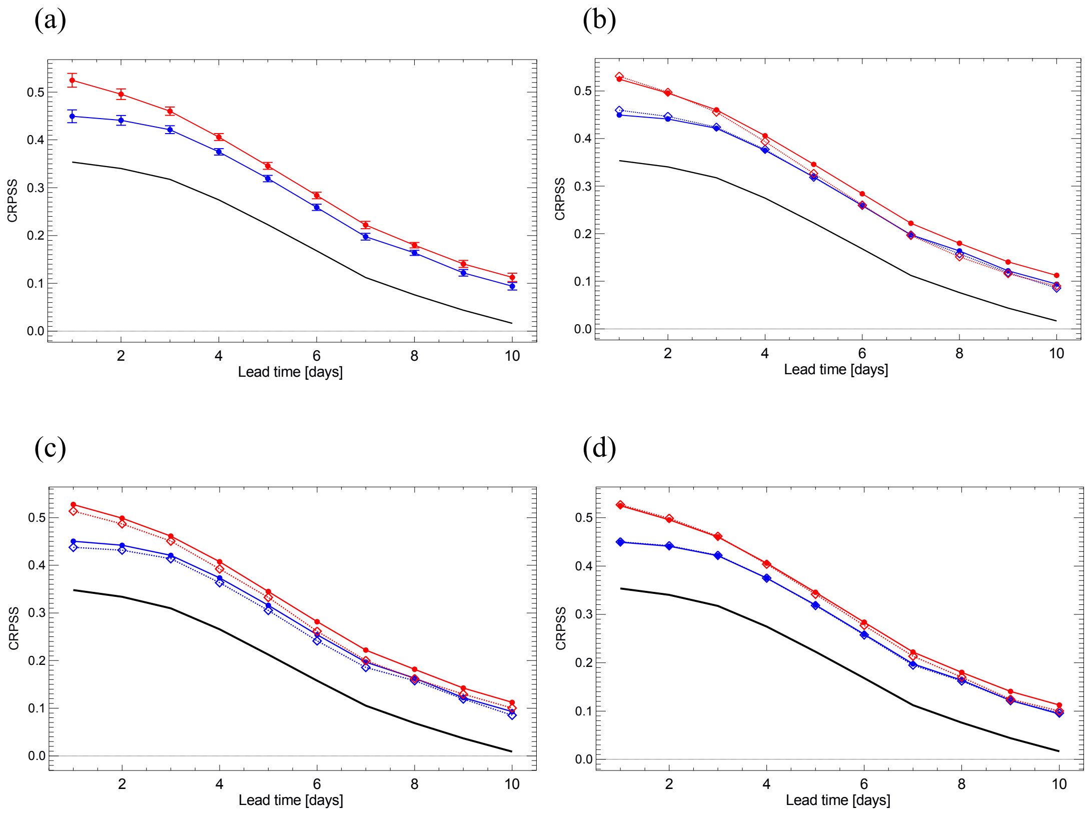


Figure 21: Continuous Ranked Probability Skill Scores (CRPSS) of 2m temperature predictions at 100 European stations and for 91 cases (1 Sep - 1 Dec 2005) versus lead time. (a) Black line: direct model output, blue line: calibrated predictions using Bias correction method, red: calibrated predictions using NGR method. Significance levels (0.05) are denoted by vertical/horizontal bars; (b) as (a), but showing in addition the results when using the operational forecasts of the previous 30 days as training dataset (dashed lines); (c) as (a), but showing in addition the results when using only 12 years instead of 20 years of the re-forecast dataset; (d) as (a), but showing in addition the results when using only 5 members instead of 15 members of the reforecast dataset.

Having established that the re-forecasts can be successfully used for calibrating EPS forecasts of weather parameters at local stations, it has been investigated how successful the calibration procedure is when replacing the re-forecast training dataset with operationally available previous forecasts (Figure 21b). It turns out that the bias correction results are hardly affect by using the operational EPS forecasts from the previous 30 days instead of the re-forecasts. However, the NGR method is more sensitive to the available training dataset, and the results deteriorate particularly for the longer forecast range.

In order to setup re-forecast suite optimized in terms of cost-benefits, the impact of available number of years and ensemble member in the re-forecast dataset has been tested. The results indicate that both bias correction and NGR method are more affected by a reduction in the number of years (Figure 21c) than by reducing the number of available ensemble members (Fig. 21d), with the NGR method generally being more sensitive to changes in the training dataset.

6.4. Summary and Outlook

This study on the potential of using re-forecasts for calibrating direct model output from the EPS has demonstrated the general scope for improvements, and that the re-forecast dataset can add some benefit when using advanced calibration methods. The results also suggest that a larger number of years in the re-forecast suite can improve the calibration more than an increased number of ensemble members. That is, when deciding on the configuration of the re-forecast suite, priority should be given to the number of years rather than the number of ensemble members.

Since the calibration procedure is most successful for EPS forecasts of near surface weather parameters at local stations, and traditionally such post-processing is done by the NMS's of the Member States, it is not planned that ECMWF will provide global model fields of calibrated EPS forecasts. Instead, it is expected that Member States will use the reforecast dataset themselves for their local calibration procedures. However, ECMWF will offer their expertise and tools for further collaborations in research on optimized calibration methods. Such collaborations and feedback on the utility of the re-forecast dataset is also important for future refinements and optimizations in the design of the re-forecast suite.

7. Combined Prediction Systems

Over the first few days of a forecast, when the flow is more predictable, it is possible that the higher resolution deterministic forecast could add skill to the lower resolution ensemble forecast. This may be particularly pertinent for a fine-scale weather parameter such as precipitation (Rodwell, 2006). Figure 22a shows schematically how one could incorporate the additional skill of a single high-resolution forecast into an ensemble prediction system. The figure shows a frequency plot of the forecast rainfall amounts from a ten-member EPS (orange squares) and the single high-resolution deterministic forecast (yellow rectangle). The high-resolution forecast has been drawn with an area (i.e. weight) equal to three ensemble members to reflect the possibility that it may be more skilful. The orange squares are all the same size because each ensemble member is equally likely. Based on the schematic, the combined forecast probability for the "event" that precipitation is greater than 1 mm would be 9/13. In practice, the weight to apply to the deterministic forecast is calculated so as to maximise the Brier Skill Score (BSS) of the combined prediction system (CPS). To avoid any artificial enhancement of skill, the weight determined for year *n* is used in the calculation of the BSS for a year *n*+1, etc. The weights determined here are a function of lead-time but independent of SYNOP station and, initially, independent of the time-of-year. This station independence ensures that the estimated skill does not over estimate the general level of skill for points where training data are not available. Although the BSS is used here, the same principles could be applied to a user's own definition of "value". An important question is how sensitive the weights are to particular measures of value or definitions of "event".

Figure 22b show the mean BSS for years 2001–2005 based on the operational 50-member EPS alone (black) and on the CPS (orange). The orange circles signify where the CPS is statistically significantly superior at the 5% level to the ensemble system alone (based on daily contributions to the BSS). It is clear that the incorporation of the single high-resolution DET forecast improves the skill at all lead-times, particularly at short lead-times. Further cross-validated tests reveal that the increased skill of the CPS occurs in all seasons with perhaps the biggest increases occurring in autumn and winter.

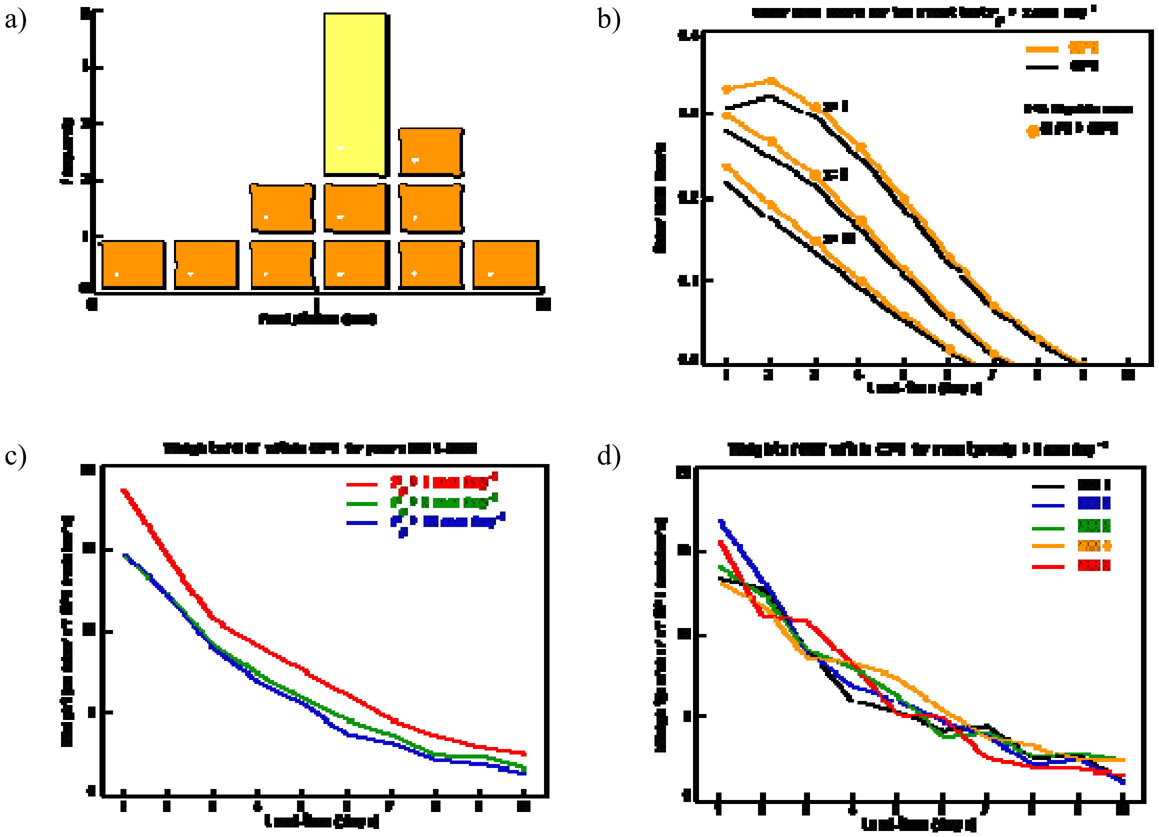


Figure 22: The combined prediction system (CPS). (a) Schematic of the methodology used to combine a single high resolution deterministic forecast (yellow rectangle) with a 10-member lower resolution ensemble forecast (orange squares). See text for more details. (b) The 2001-2005 mean of the daily station-mean Brier Skill Score (BSS) for the EPS (black) and the CPS (orange) for three 24-hour European SYNOP station precipitation thresholds (1mm, 5mm and 10mm). Orange circles indicate statistically significant (at the 5% level) improvements in skill. (c) The optimal weights applied to the deterministic forecast averaged over all years for the three 24-hour precipitation thresholds. (d) The optimal weights for the 5mm threshold for each year individually.

The optimal weight for the deterministic forecast in units of EPS members is relatively independent of event threshold (c) and stable over the years considered (Figure 22d). Generally the deterministic forecast has a weight of around 15 EPS members at day 1. Interestingly, this is roughly the number of T255 EPS members that could be made with the same computing power as a single T511 deterministic forecast. After day 1, the weight declines towards 1 EPS member by day 10.

Since September 2006, the new VAREPS has been operational. The operational configuration is 50 members, with T399 resolution to day 10 and then T255 to day 15. The dotted red curve in Figure 23a shows the BSS based on the operational VAREPS for the period 2006/10/01 to 2006/12/07 and for the event that 24-hour precipitation is greater than 1mm. It should be noted that the scores are now plotted to day 15. Generally, however, skill is lost by day 9. Motivated by the results of the CPS, a "high resolution" (HR) VAREPS configuration has been investigated. This HR configuration has a resolution of T799 until day 5 and then T399 to day 15. To use the same computing resources as the operational configuration, the HR configuration has just 10 members. The dotted blue curve shows the skill of the HR configuration. The increased resolution in the partially deterministic early period of the forecast leads to increased skill which is statistically significant (at the 5% level) at days 2 and 3. However, the smaller ensemble size leads to poorer scores beyond day 5 (statistically significantly worse at days 6 and 7). For a medium-range forecasting centre, it would appear, at first sight, that the operational configuration is preferential.

It has long been suspected that there could be some benefit from combining a previous forecast (for example yesterday's forecast) with the most recent forecast. Hoffmann and Kalnay (1983) investigated this possibility in an idealised context with "Lagged Average Forecasts". They were interested in minimising the mean squared (deterministic) error of a weighted average of time-lagged forecasts. A lagged average forecast based on a deterministic forecast did reduce this error. Less deterministic error improvement was found for a lagged average of an EPS ensemble-mean. It is unclear whether this latter result is valid for today's operational ensemble systems but, in any case, the true power of an ensemble is in its ability to span the space of possible future states in order to produce probabilistic forecasts. A "lagged combined" ensemble will have more members and may be better able to span this space. If an increased ensemble size is useful then the 10-member HR VAREPS configuration may particularly benefit from such a lagged approach. This possibility is investigated below.

Using the CPS methodology three VAREPS forecasts, starting on consecutive days, have been combined. In this case, optimal weights are found for each of the three ensembles. To ensure no artificial skill enhancement, the weights based on the first half of the period are applied in the second half and visa-versa. The solid curves in Figure 23a demonstrate that both VAREPS configurations do indeed benefit from this lagged approach. (The lead-time refers to that of the most recent forecast). In particular, the increased ensemble size improves the medium-range precipitation skill of the HR configuration (blue solid) so that it is equivalent to that of the lagged operational configuration (red solid). (This is the case even for 20mm rainfall events). Another advantage of the lagged approach is that it increases consistency in forecasts from one day to the next. This is because only a third of ensemble members are replaced each day. Inconsistency or "jumpiness" is a particular concern for market traders and can add volatility to prices. This is discussed further in section 8. For the 1mm event, the weight of the youngest of the three forecasts is around 0.4 at D+1 and this rises to around 0.7 in the medium range before dropping back towards ⅓ by the end of the forecast. The maximum in weight at the medium range may reflect the fact that the older forecasts contribute useful spread to the CPS prior to this.

The above results focus on Europe but a new global SYNOP climatology has been produced based on all global SYNOP reports from 1979. This has allowed us to estimate, for the first time, the skill of our Meteogram rainfall product throughout the world. Although present tropical rainfall skill is poor, the fact that we can now estimate this skill will provide further impetuous to make improvements to our representation of tropical physics.

Figure 23b shows the corresponding results for European 850hPa temperature (T850). Slightly more data was available when these scores were produced. In the short range, the non-lagged HR VAREPS (blue dotted) is marginally better than the operational VAREPS (red dotted) (statistically significantly better at days 1, 2 and 3). In the medium-range the HR VAREPS is worse (significantly worse at days 6, 7, 9, 10 and 15). The lagged approach does not greatly affect the short-range scores for either configuration and the HR VAREPS remains significantly better at days 1, 2 and 3. Lagging does, however, improve the medium-range scores for the HR VAREPS so that it is no longer significantly worse than the lagged operational VAREPS. The weight for the youngest ensemble is 0.87 for the operational VAREPS and 0.82 for the HR VAREPS at D+1 and rises to a maximum around D+5.

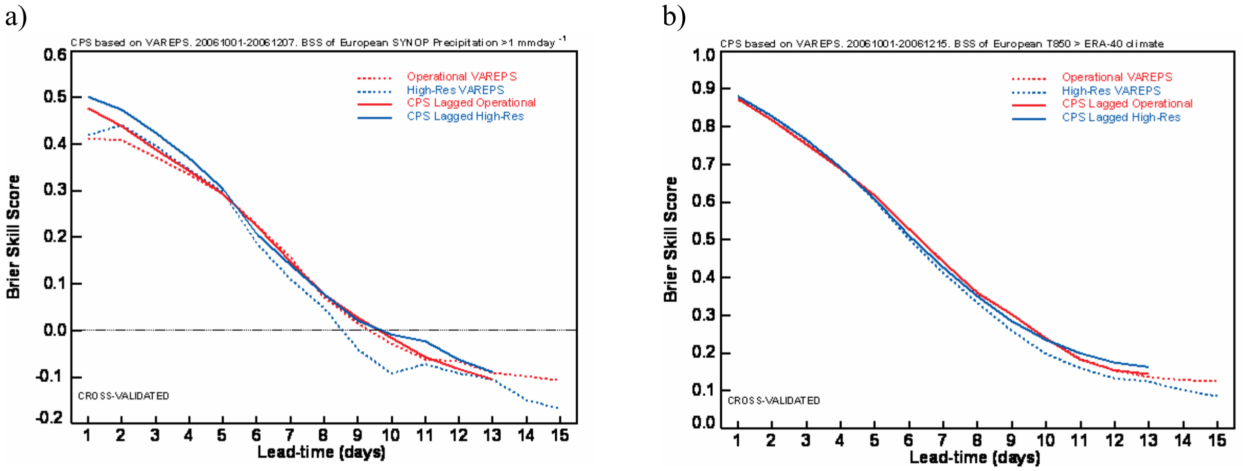


Figure 23: Mean of daily European Brier Skill Scores. (a) For the event that 24-hour accumulated precipitation at European SYNOP stations is greater than 1mm. (b) For the event that T850 is greater than ERA-40 climatology. Red dotted: the operational 40-member VAREPS (T399 to D+10, T255 thereafter). Blue dotted: the high-resolution 10-member VAREPS (T799 to D+5, T399 thereafter). Red solid: the CPS of three lagged operational VAREPS forecasts, with start dates separated by 1 day. Blue solid: the corresponding lagged CPS of the high-resolution VAREPS configuration. The period used in (a) is 2006/10/01 to 2006/12/07 and in (b) it is 2006/10/01 to 2006/12/15.

In conclusion, there is evidence that the HR configuration is preferable for probabilistic forecasts of precipitation (less so for T850 and Z500) if a forecaster is able to use lagged techniques. This result holds especially in the first few days of the forecast period. Further research is required, particularly into the prediction of severe weather such as flooding. Whether (and how) the VAREPS should be configured to optimise lagged combined forecasts requires further research. The CPS technique is, of course, also useful as a diagnostic tool. The output weights, for example, allow a comparison of forecast systems and even allow one to diagnose how well the VAREPS handles resolution truncation.

8. Use of EPS forecasts in decision-making processes

8.1. Demonstrating the value of EPS forecasts: Weather Roulette

The ultimate value of a weather forecast system lies in its ability to improve weather related decision-making processes. That is, decisions made with the knowledge of a forecast should be better than decisions made without this information. It is also known that the added uncertainty information given by probabilistic forecast systems can improve the value of the predictions compared to deterministic forecasts (Richardson, 2000). However, conventional scientific scores like the potential economic value or Brier Skill Score as described e.g. in Joliffe and Stephenson (2003) are not always the best tools to convince the general public of the benefits one can achieve when using probabilistic forecasts in their individual decision-making processes. Hence, a simple yet convincing tool has been developed to illustrate the value of probabilistic forecasts to a wider audience, the so-called Weather Roulette (Hagedorn et al., 2007).

The basic idea behind Weather Roulette is to 'literally' assess the outcome of the situation in which someone is saying: "I bet I can do better than using your forecast!" Being confident in the superiority of our forecast, we accept the challenge and agree on some fair betting rules. The rules for a fair game of Weather Roulette between two players, each using a different forecast system, are as follows.

The event to be predicted is one of five equally likely categories (normal, below normal, above normal, well-below normal, well-above normal) for the 2m temperature (T2m) at London-Heathrow. Each player opens a "weather casino" setting the odds for these five events according to their own forecast, i.e. low probabilities correspond to high odds (or payouts) and vice versa. In order to evaluate which forecast system leads to the higher return, both players, let's call them Adam and Eve, gamble the same amount of money in the casino of the opponent by distributing their capital according to their own forecast. Let's assume e.g. that Adam's forecast predicts a low probability of 0.2 for T2m being in the above normal category, i.e. Adam's casino offers a return of 5:1 for above normal T2m. Assuming further that Eve's forecast predicts above normal T2m with a probability of 0.5, and thus Eve invests 50% of her capital on this category. If the above normal category verifies, Adam has to pay out 5 times the money Eve set on the above normal category, i.e. Eve gets back 2.5 times of the total money she spent in Adam's casino. On the other hand, since the odds in Eve's casino for the verifying category were only 2:1, Adam will get back only twice his stake in the above normal category, which was only 20% of his total capital, i.e. he will get back only 0.4 times of the total money he spent in Eve's casino. Comparing Adam's and Eve's return on their investment (0.4 vs. 2.5) it is clear that in this case Eve's forecast system wins.

Playing Weather Roulette for a longer period and at more than one location can give useful insight into the relative performance of two forecast systems. An example of such a diagnostic is given in Figure 24, which shows the average daily return on investment when playing the dressed EPS forecast against the dressed high-resolution deterministic forecast. The results demonstrate that, apart from the very early forecast range, the EPS forecasts clearly lead to a higher return on investment.

Though this simple demonstration of the value of EPS forecasts can be a helpful tool in raising interest in the potential of EPS forecasts, the real value of a forecast indisputably depends strongly on the specific (and real!) application it is used for. Thus, we give below some examples of current applications of the EPS in the real world.

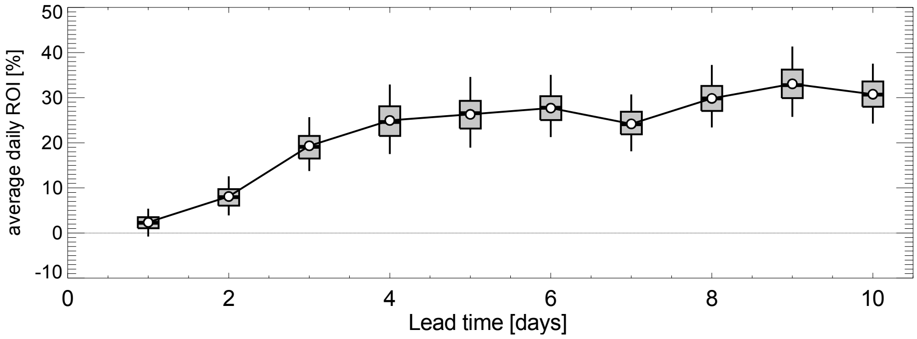


Figure 24: Average return on investment when playing the weather roulette game with the dressed EPS against the dressed high-resolution deterministic for predicting in which of the five categories (normal, below normal, above normal, well-below normal, well-above normal) the 2m temperature will fall at 100 locations during the period March-May 2007. Results are shown for lead times from 1-day to 10-day forecasts. Box-and-whisker symbols indicate 5, 25, 50, 75, and 95 percentiles of time-bootstrapped results.

8.2. Application of the EPS by end-users

8.2.1. Energy Trading

Weather is a dominant driver of energy prices traded in liberalised markets. Temperature, precipitation and wind speed are some of the weather variables which impact the supply/demand energy balance and it is the combined effect of a given weather pattern which must be considered, rather than each variable in isolation. Energy price predictions are obtained by modelling the expected supply/demand balance, and weather forecast information is one component of the input driving such models. Significant shifts in the predicted weather patterns between successive model runs often lead to increased volatility in market pricing as trader expectations change. Hence, while accurate forecasts are valuable, a good understanding of the uncertainty in forecasts provided is key. Consequently, probabilistic weather forecast information has become an essential component of a successful trading strategy.

Precipitation is a major driver in a number of power markets; daily precipitation forecasts are translated into potential daily power output according to where the precipitation falls with respect to the installed hydropower. Such daily precipitation energy forecasts are provided by a commercial organisation named ‘Point Carbon’, derived from the ECMWF precipitation predictions, and used by energy trading companies as input into their daily decision-making process on optimizing their energy trading strategy. Figure 25 shows the cumulative precipitation power for a sample region predicted by the high-resolution deterministic forecast started on 9 May 12z (EC12) in blue and the EPS members in grey. The deterministic forecast initialised on 8 May 12z is also shown (red). Imagine we are viewing this forecast on the morning of 10 May. The May 9 12z deterministic forecast and corresponding ensembles are the most recent forecasts available as the European markets start opening at 6-7am GMT, and are quite a shift from yesterday’s view, which was wet, but not very wet. Having seen today’s viewpoint, what forecast should we give to the energy analysts? How much rain will fall, and how much confidence do we have in our prediction? What are the alternative outcomes?

If we only use the deterministic runs, then the choice is between “expect one-and-a-half times normal” as suggested by the 12z deterministic forecast of 8 May, or “expect twice normal” as suggested by the 12z deterministic forecast of 9 May. Such “jumpiness” in the forecasts is particularly bad for energy trading applications, since a daily change in your buying and selling strategy can become very costly. Thus, information on the uncertainty in the forecasts can be extremely valuable for energy applications. Dressing the deterministic forecasts would give some idea of likely uncertainty based on historical data, but even from the two deterministic forecasts one can see that the variation is likely due to how many major precipitation events[1] will take place. It is here that the individual ensemble member forecasts come into their own: from first glance there are a number of ensemble members with no major precipitation events which seem to have lower cumulative amounts, some with one event and possibly a few with two. Distinguishing the ensembles by number of major precipitation events, it appears that many more members with one (or no) major precipitation event exist, and the probability for two events is low, but not negligible. When comparing the forecasts with the observed cumulative precipitation energy it becomes clear that mainly one major precipitation event around 13-15 May is the determining factor for the final accumulated value, which is - in this case - predicted very well by the deterministic forecasts. However, this deterministic forecast only becomes valuable through the additional confirming information from the EPS, which gives traders a far clearer picture of the drivers behind the possible outcomes and the timing (and uncertainty therein) of the precipitation events. Consequently forecasters can provide a better brief to the energy analysts resulting in a less volatile trading strategy, ultimately leading to real monetary benefits. One of the aims behind the research work on combined prediction systems (Section 7) is to provide reliable information on the relative

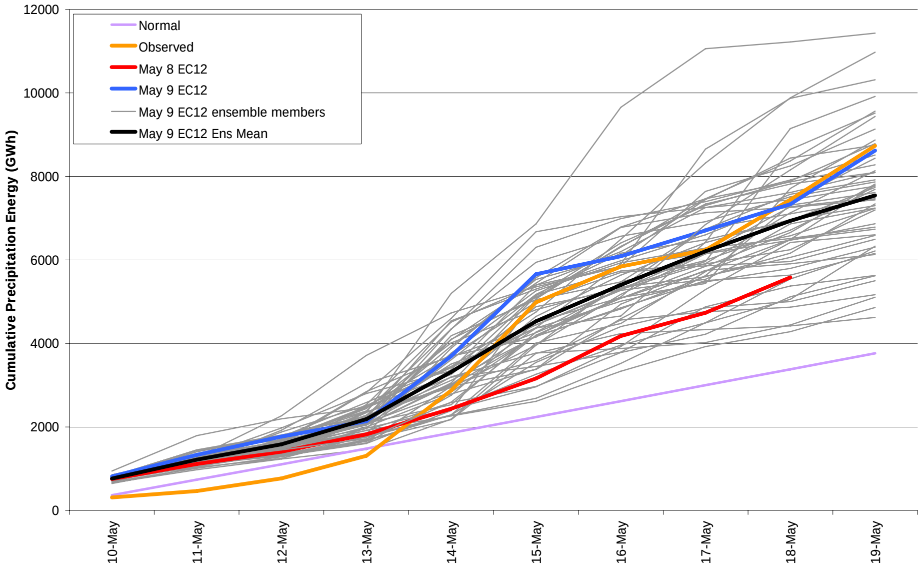


Figure 25: Cumulative precipitation energy forecast for a sample region, displaying the deterministic forecast initialized 9 May 12z (blue), 8 May 12z (red), all ensemble members from the forecast initialized 9 May 12z (grey). In addition, the ensemble mean (black), the climatological values (purple), and the observed values (orange) are shown. Daily precipitation energy forecast data source: Point Carbon.

[1] Defining a major precipitation event as at least one day of ~1500GWh, the 12z deterministic run from 9 May includes one event and the 12z deterministic from 8 May none.

value of the individual components of the suite of forecasts, and to objectively optimize their combined use in the energy markets.

8.2.2. Flood Forecasting

Following the disastrous floods in the Elbe and Danube in August 2002, the European Commission launched an activity on the development of a European Flood Alert system (EFAS). One feature of this forecasting system is the use of meteorological Ensemble Prediction Systems as input into the flood simulation model allowing one to estimate the uncertainty in combined meteorological and hydrological forecasts.

Figure 26 shows examples of such daily EFAS forecasts from 20th March to 8th April 2006, predicting threshold exceedances at the Zahorska Ves (Slovakia) station, where a flood event occurred on the 29th of March 2006. The forecasts are displayed as a matrix, with the forecast start dates changing from the top to bottom row and the lead times varying from left to right columns. For the simulations based on the deterministic weather forecasts (Figure 26&b), the highest EFAS threshold exceeded by the forecasts is displayed as colour-coded box. In the EPS-based flood forecasts (Figure 26c) the boxes are colour-coded according to the number of simulations above EFAS high threshold (EPS>HAL). For the case of the deterministic forecasts systems a flood warning is issued when the high alert level (red box) is given by two successive forecasts. Based on the DWD forecasts a flood warning would be issued with only 1-day lead (28th March) and the ECMWF deterministic system would issue a warning with 4-day lead (25th March). In contrast to that, there was already a signal of a probability of reaching high alert levels in the EPS-based simulations, with e.g. 21 out of 51 ensemble members (41%) reaching the high threshold in the forecast issued on 22nd March. The signal in EPS-based forecasts was persistent, with increasing probabilities as the forecast dates got closer to the event: more than 50% of simulations exceeding high thresholds with lead times of 6-7 days and more than 70% of simulations for lead times 4-5 days. The correct end of the flood event is predicted comparable by all forecasts.

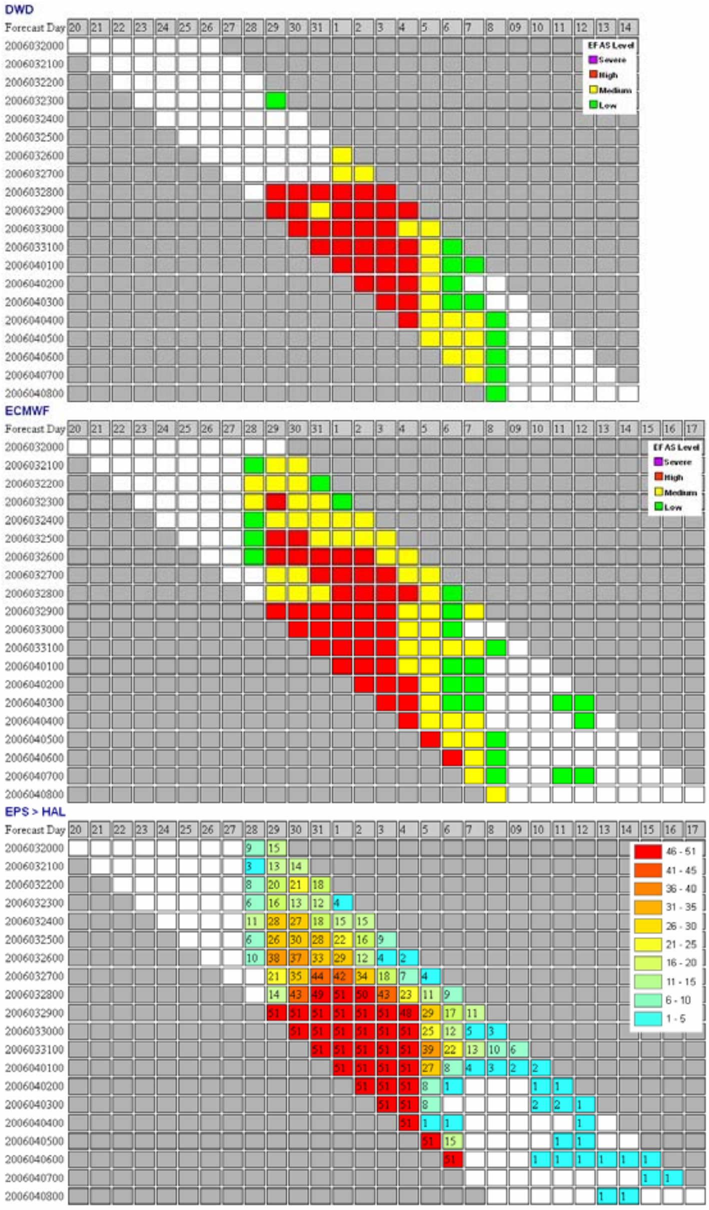


Figure 26: History of EFAS forecasted levels at the Zahorska Ves station in the Morava River for forecast dates from 20th March to 8th April (rows from top to bottom, lead time increasing from left to right columns).(a)Highest EFAS level exceeded by DWD 7-day forecasts; (b) same as (a) but for 10-day ECMWF deterministic forecasts; (c) number of ensemble members of the ECMWF EPS exceeding EFAS high threshold.

The skill of the EPS in early flood warning has been investigated using various skill scores to assess the performance of EFAS. One of the important results is that by using the EPS, events can be detected much earlier. Figure 27 displays the distribution of the gain in lead time achieved when using the EPS instead of deterministic forecasts. Positive numbers indicate that the lead time (advance warning) given by the EPS is longer than the lead time given by the deterministic forecasts. For example, using the EPS rather than a deterministic forecast leads in 8000 cases to a flood warning issued one day earlier when considering upstream areas between 4,000 and 10,000 km2 (Figure 27a). In general, the advance warning time given by the EPS is significantly longer compared to the deterministic forecasts (significant up to day 4). The average increase of lead time leaves more time to organize for instance:

- Change of working schedule
- Increased attention to affected areas
- Discussions in forecast teams (additional information to minimize risk of false alarms, orientation information)
- Included into warning message
- Release of reservoirs

Organisations which took actions based on the advanced warning have given an overwhelming positive feedback.

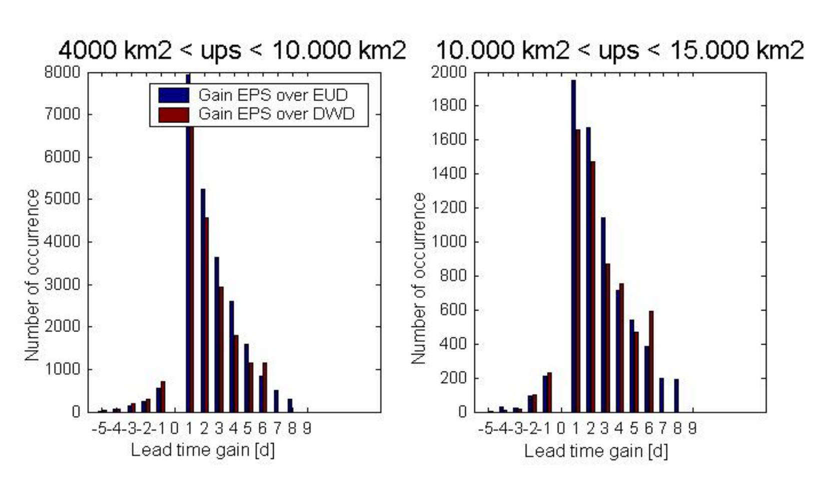


Figure 27: Histograms of the gain in lead time for flood warnings when using the probabilistic forecasts from the EPS instead of the deterministic forecasts from ECMWF (blue bars) and DWD (red bars).Positive (negative) numbers correspond to an increase (decrease) in warning time when using the EPS forecasts. The histograms consist of all flood warnings issued for European rivers in the period between 01.01.2001 - 31.12.2006; (a) forecasts issued when considering catchments with an upstream area between 4,000 and 10,000 km2; (b) forecasts issued when considering catchments with an upstream area between 10,000 and 15,000 km2.

8.3. Outlook

The above given examples on the use of EPS forecast in the commercial and public services sectors are only two out of a growing number of EPS applications. After the early years of the EPS, in which the idea and potential of probabilistic forecasting had to be explained and spread in the user community, EPS forecasts and products have more recently become essential components in many application areas. For example, the EPS is not only used by all major consulting companies on the European and global energy trading market (e.g. Point Carbon[2] , montelpowernews[3], WSI[4] , etc.), but also serves as an important source of information

[2] http://www.pointcarbon.com/

[3] http://www.montelpowernews.com/

[4] http://www.wsi.com/

in all decision-making processes related to actual energy production (e.g. scheduling of powering up/down stations, maintenance, work rota, etc.).

In addition to the well established use of the EPS in traditional markets, the emerging market of renewable energy production, in particular wind energy, is starting to develop interest in probabilistic weather forecasts. Contacts have been established with research groups in the field of wind energy forecasting in order to promote and establish the use of EPS forecasts and products amongst this community. In particular when trying to establish wind energy as a reliable component of the energy mix, improved forecasting capabilities including reliable uncertainty information is of crucial importance. The value of additional information on the range of uncertainty is demonstrated in Figure 28. In this example, the deterministic forecast (red) predicted a power production of 40% of the maximum capacity over Germany but only slightly more than 20% could be produced (green). In contrast, the probabilistic power prediction, using the EPS forecasts as forcing, reveals the range of uncertainties and shows a relatively high probability for lower power production (black lines). This information could have been used to cover for the shortfall in advance.

In the future, this research will be extended in order to establish how best to use probabilistic forecast information for such applications, and possibly to improve individual aspects of the EPS system specified by relevant end-user demands.

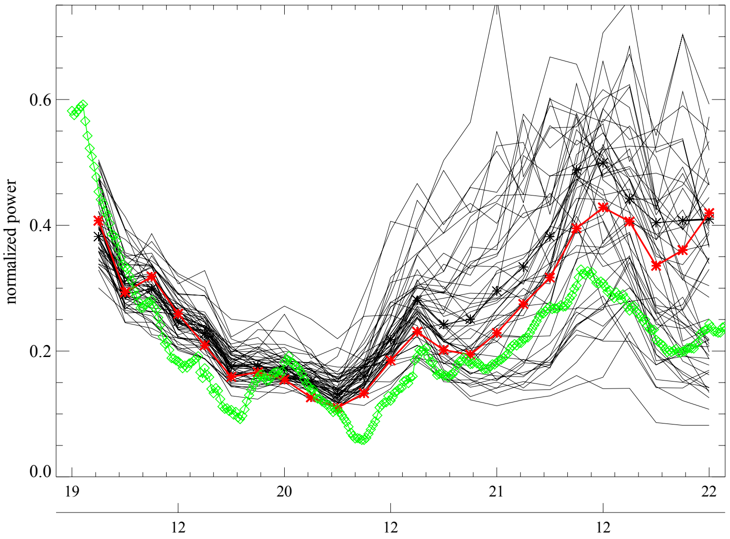
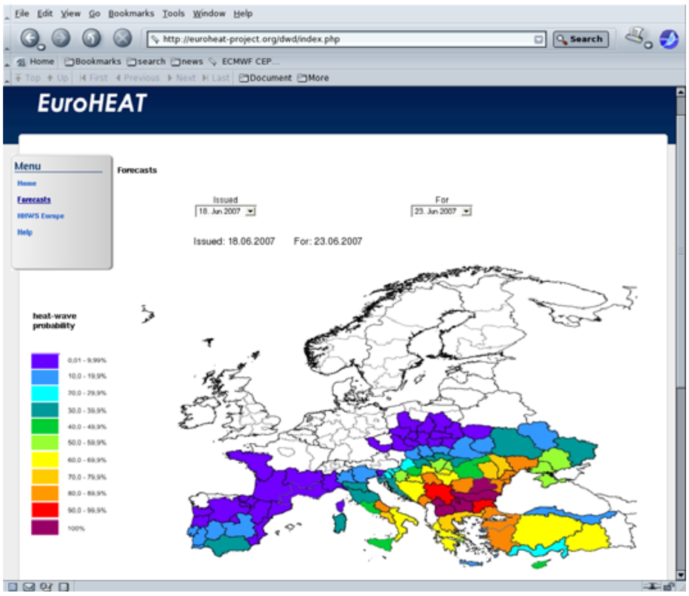


Figure 28: Prediction of wind power production over Germany for the period 19 to 21 March 2007, normalized by the maximum capacity. The observed power production is marked with green diamonds, the deterministic power prediction using the ECMWF high-resolution forecast as input is denoted with red stars, and the probabilistic prediction using the EPS ensemble members as input is displayed in black.

In addition to the various users from the energy sector, users from other application areas are starting to get interested in incorporating uncertainty information into their decision-making processes. Most recent activities include using the EPS for applications in the health, transport, and agriculture sectors. For example, the German Weather Service has developed a European Heat-Health-Warning-System (HHWS) based on ECMWF EPS information as part of the EU project EUROHEAT (Koppe, 2007). First results indicate that the EPS forecasts, in combination with a health related approach to determine warning levels, can give useful information with lead times up to approximately 8 days in comparison to the 3-5 days warning time given by the deterministic system. This HHWS is now operational and e.g. did predict very well the recent heat wave occurring in South-east Europe around 23 June 2007, with probabilities reaching even 100% for 5-day lead forecasts in the worst affected areas of the central Balkan (Figure 29). Such early warnings can be used to increase the preparedness of the health sector (planning of special care for vulnerable population, hospital services plan for rise in admissions, etc.) and the general public (planning of activities, increasing water intake, etc.).



http://euroheat-project.org/dwd/index.php

Forecasts

Issued: 18. Jun 2007

For: 23. Jun 2007

Issued: 18.06.2007 For: 23.06.2007

Figure 29: The Web-based decision support tool for Heat Health Warnings developed by the German Weather Service (DWD) based on the ECMWF EPS forecasts. The website displays the probabilities for exceeding the regional thresholds for heat. Forecasts are issued with lead times up to nine days, updated every day. The example displayed here shows the forecast for 23 June 2007, issued on 18 June 2007.

Although the above mentioned developments demonstrate a growing interest in using the EPS, there is always scope for improving the level of awareness and use of probabilistic forecasts. Apart from intensifying the use of EPS forecasts in the above mentioned sectors of energy, transport, etc., it seems that in particular the area of weather forecasts for the general public is a potential target area for increased incorporation of probabilistic information. Though traditionally ECMWF itself is less engaged with the end users, we believe that it is important to develop an education and feedback strategy incorporating the whole chain from forecast producers, via forecast providers to finally the forecast users. Thus, ECMWF will increase its efforts to promote and educate on the best use of EPS forecasts and products via the National Met Services of the Member States. Encouraging more exchange of information between all involved parties and engaging forecast providers and end-users in the process of product development will lead to an optimized utilization of the full potential of the EPS.

9. Future development

The development of the EPS can be grouped around the following areas: representation of forecast uncertainties, ensemble size, the resolution and complexity of the underlying forecast model, verification and application.

Whilst the singular vector methodology has proven useful over the last 15 years (as providing the means whereby EPS error and spread statistics are reasonably-well matched), as the underlying sources of forecast error are better understood, and hence better quantified, the dependence of the EPS on singular-vector perturbations may decrease. As discussed in the body of this paper, some significant work is now underway to develop an ensemble data assimilation (EDA) system. Whilst initial studies have shown EDA, by itself, to lead to very underdispersive ensemble forecasts, EDA nevertheless provides the basis for future developments. For example, through EDA it will be possible to incorporate more fully than has been hitherto possible, initial perturbations in land-surface variables and in the tropics in general. Hence, it is envisaged that EDA will replace the evolved singular-vector component of the current initial perturbations.

The incorporation of stochastic parametrisation into EDA, an ongoing activity, provides a first step towards incorporating model uncertainty into the specification of the initial conditions. Further developments within EDA will include correlated observation error statistics. Although such errors are pertinent to satellite data, correlated observation-error covariances can be considered to represent implicitly some of the observation representativity effects; if a set of regional observations is influenced by circulations whose scales are too small for the model to represent well dynamically, the assimilation of these observations into the model will give rise to correlated errors in associated grid-point fields.

It is clear that the development of techniques to represent model error will continue to be crucial for improving the reliability of EPS forecasts. At present three techniques are being developed to represent model uncertainty: the multi-model ensemble, perturbed parameters/parametrisations, and stochastic physics. Of these, stochastic physics has the potential to reduce systematic model error through noise-induced drift effects. However, stochastic physics schemes are still in their infancy. The initial types of scheme which represented uncertainty in the diabatic tendencies through stochastic modification of these tendencies, have been supplemented with backscatter schemes. However, a crucial aspect of these backscatter schemes is that the imposed near-grid forcing can project onto modes which have the desired upscale transfer properties. Work to develop the effectiveness of backscatter schemes must ensure that the energy of the imposed stochastic forcing is not immediately dissipated. It is possible that singular vector techniques could prove useful here - singular vector growth exhibits upscale effects.

The development of the THORPEX/TIGGE archive will provide the means to compare objectively, stochastic physics and perturbed parameter/parametrisation schemes with multi-model schemes. Similar types of study are already underway on seasonal/decadal timescales in the EU ENSEMBLES project - showing that the multi-model approach is still more than competitive with the others - and techniques developed here can be readily applied to the medium range problem.

One of the key questions in designing an EPS system is the balance between model resolution and ensemble size. In recent years, the EPS resolution has been chosen to be half the resolution of the (high-resolution) deterministic forecast and the ensemble size has been 50. An ensemble size of 50 allows probabilities with resolution of about 10% to be meaningfully given. In this sense, the EPS would be useful to users with cost/loss ratios of greater than about 0.1. Is this balance between resolution and ensemble size optimal?

It is proposed to study this question using the VAREPS technique which was implemented earlier this year to provide a unified medium range and monthly forecast system. It is possible to extend the implementation of VAREPS with resolution approaching that of the high-resolution deterministic forecast system in the early medium range. In this way, the potential benefits of a high-resolution EPS in the short and early medium range can be realised. On the other hand, whilst the computational cost of an ensemble scales linearly with both ensemble size and forecast length, the computational cost of a forecast model scales approximately cubically with horizontal resolution. Hence doubling resolution to take advantage of the benefits of high resolution in the short and early medium range will mean a substantial reduction in ensemble size. One question is whether one can compensate for such a reduction in ensemble size by lagging together earlier EPSs. This is not entirely clear as perturbations associated with earlier ensembles may not be orthogonal to perturbations associated with today's ensemble. Also, it may be possible to reject some lagged perturbations as unrealistic when propagated forward to the current analysis time. Answers to such questions will require much further research.

As well as increased forecast resolution, it is important to include in the underlying forecast model as much Earth-System complexity as is relevant. For example, in the current VAREPS system, ocean-atmosphere coupling is switched on after day 10. As well as influencing large-scale modes in the tropics on intraseasonal timescale, ocean-atmosphere coupling is believed to influence the development of synoptic-scale disturbances, especially in cases of severe weather development. As such it is hoped that within the planning period of the current 4-year plan, the EPS system will use a coupled ocean-atmosphere model from day 0. Other developments in Earth-System complexity will be incorporated into the EPS as and when these become available.

Verification continues to be an important issue. As has been shown in this paper, some simple user-relevant scores (cf the Weather Roulette example) have been developed to extend the range of possible verification techniques available at ECMWF. On the other hand, it is not ECMWF's remit to develop verification software for specific applications. However, EPS staff at ECMWF will continue to work with those in the various applications sectors for which a reliable EPS is a valuable tool, in order to maximise application of the EPS.

One specific application of the EPS that is already becoming important for Member States activities is the use of the EPS to provide both boundary and initial perturbations for limited area model ensembles (LAMEPS). In this way, detailed probabilistic forecasts for the short range can be made.

Over the past 15 years, the EPS has grown from an experimental tool, to an established part of numerical weather prediction. However, there are still fundamental problems to be solved before the EPS can be described as fully reliable. Most importantly, the problem of defining uncertainties in terms of known probability distributions is not fully solved. It is hoped that in the next 15 years will see a gradual transition of the many "unknown unknowns" with fully quantified "known unknowns".

10. References

Barkmeijer, J., M.V. Gijzen and F. Bouttier ,1998: Singular vectors and estimates of the analysis-error covariance matrix. Quart. J. Roy. Meteor. Soc., 124, 1695–1713.

Barkmeijer, J., Buizza, R., & Palmer, T. N., 1999: 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. Q. J. R. Meteorol. Soc., 125, 2333–2351.

Buizza, R., 1994: Sensitivity of optimal state structures, Quart. J. Roy. Meteor. Soc., 120, 429–451.

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

Buizza, R., M.J. Miller and T.N. Palmer, 1999: Stochastic simulation of model uncertainties in the ECMWF Ensemble Prediction System. Q.J.R.Meteorol. Soc., 125, 2887-2908.

Buizza, R., Houtekamer, P. L., Toth, Z., Pellerin, G., Wei, M., & Zhu, Y., 2005: A comparison of the ECMWF, MSC and NCEP Global Ensemble Prediction Systems. Mon. Wea. Rev., 133, 1076-1097.

Buizza, R. M. Miller and T.Palmer: Stochastic representation of model uncertainties in the ECMWF Ensemble Prediction System, Quarterly Journal of the Royal Meteorological Society, 125, 2887-2908

Buizza, R., J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G.Holt and F. Vitart, 2007: The new ECMWF VAREPS (Variable Resolution Ensemble Prediction System). Q.J.Roy. Meteorol. Soc., 133, 681-695.

Candille, G and O. Talagrand, 2005: Evaluation of probabilistic prediction systems for a scalar variable. Quart. J. Roy. Meteor. Soc., 131, 2131–2150.

Coutinho, M. M., B. J. Hoskins and R. Buizza, 2004: The Influence of Physical Processes on Extratropical Singular Vectors. J. Atmos. Sci., 61, 195–209.

Ehrendorfer, M., R. M. Errico and K. D. Raeder, 1999: Singular vector perturbation growth in a primitive equation model with moist physics. J. Atmos. Sci., 56, 1627–1648.

Ehrendorfer, M. and A. Beck, 2003: Singular vector-based multivariate normal sampling in ensemble prediction. ECMWF Tech. Memo. 416.

Gneiting, T., A.E. Raftery, A.H. Westveld III, and T. Goldman, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. Mon. Wea. Rev., 133, 1098-1118.

Hagedorn, R. et al., 2007: Playing weather roulette to evaluate probabilistic forecasts. Meteor. Appl., submitted.

Hamill, T.M., J.S. Whitaker, and X. Wei, 2004: Ensemble re-forecasting: Improving medium-range forecast skill using retrospective forecasts. Mon. Wea. Rev., 132, 1434-1447.

Hamill, T. M. and J. Juras, 2006: Measuring forecast skill: Is it real skill or is it the varying climatology? Quart. J. Roy. Meteor. Soc., 132, 2905–2923.

Hamill, T.M. and J.S. Whitaker, 2007: Ensemble calibration of 500 hPa Geopotential Height and 850 hPa and 2-Meter Temperatures Using Reforecasts. Mon. Wea. Rev., accepted

Hoffman, R. N., and E. Kalnay, 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.

Houtekamer, P. L., & Mitchell, H., 1998: Data assimilation using an ensemble Kalman filter. Mon. Wea. Rev., 126, 796–811.

Jolliffe, I.T. and D.B. Stephenson, 2003: Forecast verification: A practitioner's guide in atmospheric science. Wiley, 240 pp.

Jung, T. and M. Leutbecher, 2007: Performance of the ECMWF forecasting system in the Arctic during winter. Quart. J. Roy. Meteor. Soc. (in press).

Koppe, C., 2007: Climate Information Decision Support Tool for Heat in Europe, online publication: http://euroheat-project.org/dwd/flyer_background.pdf

Leutbecher, M. and T. N. Palmer, 2007: Ensemble forecasting. J. Comp. Phys. (in press).

Mahfouf, J.-F., 1999: Influence of physical processes on the tangent-linear approximation. Tellus, 51 A, 147–166.

Molteni, F. and T.N. Palmer, 1993: Predictability and finite-time instability of the northern winter circulation. Q.J.R.Met.Soc., 119, 269-298.

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

Murphy, J. and T.N. Palmer, 1986: A real time extended range forecast by an ensemble of numerical integrations. Met.Mag. 337 348.

Palmer, T. N., R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics and adaptive observations. J. Atmos. Sci., 55, 633–653.

Palmer, T.N., 2001: A nonlinear dynamical perspective on model error: a proposal for nonlocal stochastic-dynamic parametrisation in weather and climate prediction models. Q.J.R.Meteorol.Soc., 127, 279-304.

Palmer, T.N., F.Molteni., R. Mureau R., R.Buizza, P. Chapelet and J.Tribbia, 1993: Ensemble Prediction. In Processings 1992 ECMWF Seminar: Validation of Models over Europe, pp21-66, Reading UK.

Palmer, T.N.. A. Alessandri, U. Anderson and co-authors. 2004. The Development of a European Multi-model Ensemble System for Seasonal to Interannual Prediction (DEMETER). Bulletin of the American Meteorological Society, 85, 853-872.

Richardson, D.S., 2000: Skill and relative economic value of the ECMWF Ensemble Prediction System. Quart. J. Roy. Meteor. Soc., 126, 649-668.

Rodwell, M. J., 2006: Comparing and combining deterministic and ensemble forecasts: How to predict rainfall occurrence better. ECMWF Newsletter, 106, 17-23.

Roulston, M.S. and Smith, L.A., 2002: Evaluating probabilistic forecasts using information theory, Monthly Weather Review, 130, 1653-1660

Toth, Z. and E. Kalnay, 1993: Ensemble forcasing at NMC: the generation of perturbations. Bull Am. Meteorol. Soc.,

Saetra, Ø. & J.-R Bidlot, 2004: Potential Benefits of Using Probabilistic Forecasts for Waves and Marine Winds Based on the ECMWF Ensemble Prediction System. Wea. Forecasting, 19, 673-689.

Saetra, O., H. Hersbach, J.-R. Bidlot, D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. Mon. Wea. Rev, 132, 1487–1501.

Shutts, G.J, 2004: A stochastic kinetic energy backscatter algorithm for use in ensemble prediction systems, ECMWF Technical Memorandum, 449

Shutts, G., 2005: A kinetic energy backscatter algorithm for use in ensemble prediction systems. Q.J.R.Meteorol.Soc., 131, 3079-3102.

Tompkins, A. M. and Janisková, M., 2004: A cloud scheme for data assimilation: Description and initial tests. Quart. J. Roy. Meteor. Soc., 130, 2495–2517.

Uppala, S.M. and co-authors, 2005: The ERA-40 re-analysis. Quart. J. Roy. Meteor. Soc.,131, 2961–3012.

Vitart, F., S. Woolnough, M.A. Balmaseda and A. Tompkins, 2007: Monthly forecast of the Madden-Julian Oscillation using a coupled GCM. Mon. Wea. Rev., in press.

WMO, 2005a: Report of the 20th session of the CAS/JSC WG on numerical experimentation. Met Office, Exter, UK, 11-15 October 2004. WMO/TD-No. 1297, 59pp.

WMO, 2005b: Report of the 1st workshop on the THORPEX Interactive Grand Global Ensemble (TIGGE). Reading, UK, 1-3 March 2005. WMO/TD-No. 1273, WWRP/THORPEX No. 5, 39pp.

Walser, A., M. Arpagaus, C. Appenzeller and M. Leutbecher, 2006: The Impact of Moist Singular Vectors and Horizontal Resolution on Short-Range Limited-Area Ensemble Forecasts for Two European Winter Storms. Mon. Wea. Rev., 134, 2877–2887.

Wilks, D.S., 2006: Statistical Methods in the Atmospheric Sciences, 2nd Ed., Academic Press, 627pp.

Woolnough, S. J., F. Vitart and M. A, Balmaseda, 2007: The role of the ocean in the Madden-Julian Oscillation: Sensitivity of an MJO forecast to ocean coupling. Q. J. Roy. Meteor. Soc., 133, 117-128.