432

Verification statistics and evaluations of ECMWF forecasts in 2002-2003

F. Lalaurette, L. Ferranti, A. Ghelli, and G. van der Grijn

Operations Department

December 2003

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading, Berks RG2 9AX

library@ecmwf.int

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications.html

© Copyright 2003

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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.

1. Introduction

This document summarises the recent changes to the data assimilation/forecasting system (section 2). Verification results of the medium range free atmosphere forecast produced at ECMWF are then presented, including results from the EPS (section 3). A large part of this section is devoted to a comprehensive comparison with other centres providing global numerical weather forecasts. Section 4 deals with the verification of ECMWF weather parameters and oceanic wave forecasts, while section 5 provides insights on the performance of the seasonal forecast systems. A short technical note describing the scores used in this report is in 0.

The set of verification scores shown here is mostly consistent with previous years (ECMWF Tech. Memo. 346 and 414), in order to help compare the performance year after year. Aspects related to experimental products such as those for monthly forecasts are treated in separate documents.

Verification pages have been created on the ECMWF web server and are regularly updated. Currently they are accessible at the following addresses:
http://www.ecmwf.int/products/forecasts/d/charts/verification/ *(medium-range)*
http://www.ecmwf.int/products/forecasts/seasonal/verification/index.html *(seasonal range)*

2. Changes to the data assimilation/forecasting system

The list of changes to the system since the preparation of documents for the previous meeting of the Committee is as follows:

- 29 October 2002: NOAA17 data activated in assimilation;
- 14 January 2003: Cycle 25r4, a major upgrade of the model that included:
    1. Revised multi-incremental (T95/T159) 4D-Var algorithm, including a non-linear balance in the Jb, statistics from 4D-var ensembles, and revised Jc (more selective to gravity waves)
    2. GOES WV radiances, MODIS winds, more HIRS channels and SAR data are activated; SSM/I radiances directly assimilated
    3. Improved cloud-scheme numerics, revised cloud physics and convection scheme with new type and cloud top/base algorithms; checks all levels up to 700hPa for initiation of deep convection
    4. Rescaling of EPS initial perturbations to reflect new data assimilation error statistics.
- 19 February 2003: AIRS radiances are passively monitored;
- 4 March 2003: All operational forecast suites are switched to run on the new IBM clusters;
- 25 March 2003: ECMWF medium range operational forecasts (T511L60 deterministic and T255L40 EPS) are run twice per day (00 and 12UTC). The data assimilation cycle has remained unchanged.
- 31 March 2003: the Fujitsu computers are decommissioned. As a result and according to plan, the old version of the seasonal forecast suite (System 1) has been discontinued
- 29 April 2003: Cycle 26r1- a technical change introducing in operations the use of the Observation Data Base (ODB) for analysis input and feedback files (it has been used internally by the IFS since Cy22r3)

Note: All model changes since 1985 are described and updated in real time at:
http://www.ecmwf.int/products/data/operational_system/index.html

3. Verification for free atmosphere medium range forecasts

3.1 ECMWF scores

3.1.1. Extratropics

Figure 1 gives the evolution of the average forecast skill valid for consecutive 12-months periods since 1980. The forecast parameter is the 500 hPa height over Northern Hemisphere (extratropics only) and Europe, while the scoring method is a root mean square error normalised with reference to a forecast that would persist initial conditions into the future. The last month included in the statistics is July 2003. Clearly the levels of performance have been again unprecedented this year over the Northern Hemisphere, while average performance over Europe kept the very high levels reached last year. Figure 2 shows that the forecasts of 500 hPa anomalies from the climate kept a good correlation with the observed (analysed) anomalies consistently throughout last year, with a much reduced difference between the "bad" (summer) months and the "good" (winter) ones, most notably over Europe. In particular, the performance during summer 2003 seems to be well ahead of anything found previously. Although this remains to be confirmed over the years, the hope is that the extensive work initiated by the very poor performance in 1999 has paid off in bringing summer performance in line with the winter one. It was stressed last year that for the first time, performance over the Southern Hemisphere had reached the same level as in its Northern counterpart, this year the Boreal hemisphere appears to have taken the lead again. Not too much significance should be given to such fluctuations that may be related to transient changes in the general circulation. It is worth mentioning though that a lot of efforts have been devoted to improving the use of satellite data over land recently, which could result in paying more dividends over the Northern than the Southern hemisphere, more dominated by the oceans.

Figure 3 pictures the distribution of the anomaly correlation of the day 5 forecasts of 850hPa temperature with verifying analyses over Europe since 1985 for the winter and summer seasons. 2002-03 has not seen the record-breaking performance of the previous winter reached again, which is confirmed for Europe by other scores and other parameters as well. The situation though does not seem to be related to any major failure of the forecasting system, but rather to a very different flow pattern. This will be confirmed later in this report when comparison with other forecast centres performance will be discussed (section 3.2). On the other hand, summer 2003 has shown an unprecedented high level of correlation for most forecasts at Day 5 over Europe, which is possibly again related to the unusually stable, hot and dry weather. Model improvements have been introduced however over the last few years that were aimed specifically at improving performance in the summer, and therefore some of the improved performance is likely to be related to them. Another feature that can be seen from Figure 3 is that the EPS ensemble mean is an effective filter of random errors associated with the limited predictability by Day 5 of some of the weather system described by the full resolution (T511) model. As a result, the distribution of errors in winter is shifted towards higher anomaly correlations - the only year when this was not the case was 2000-01, but this has since been found to be the result of mishandling of the scaling of EPS initial perturbations at the time (see ECMWF Tech Memo 346). The same effect can also be found in summer.

Winter 2002-2003 was not one characterised by a large storm activity, but rather a large number of blocking episodes. The average performance scores however have shown that the EPS Control does not keep the same high level of skill as the T511 model does (Figure 4). This in a way is a positive result showing that the current model formulation (numerics and physics) takes a clear benefit from the high resolution, something that was not as clear some years ago. However, the loss of skill of the T255 model in the early ranges of the forecast takes its toll on the performance of the EPS itself, as can be seen in the same figure from the performance of the Ensemble mean that only crosses the T511 curve by Day 5. One of the few storms to have affected western Europe during the cold season happened on 27 October. Careful examination of the early stages of development of the storm have revealed that the coarse resolution, long time step version of the model used for the EPS may miss some of the most important aspect of the development of such storms (Figure 5). In this case, this for example has resulted in a systematic shift in the timing of the storm landing over the British Isles, most of EPS members seeing it later than the T511 model. The improvement in performance of the T255 Control is an area of investigation that is currently looked at in research mode. Some promising results have already been obtained that hopefully should result in better performances of the relatively coarse resolution of the model used in the EPS in the near future.

Proper estimates of the EPS skill however should explicitly address the probabilistic nature of ensemble forecasts. This is done using for example Brier Scores time series and ROC curves areas (Figure 6). The latter clearly shows that the EPS has followed the general trend of improvement of the forecast over the last few years, both for moderate (+/-4K) and larger (+/-8K) anomalies. The fact that it is more so on the ROCA curves than the Brier Skill score time series reflects the fact that the Relative Operating Characteristics curves are indicative of the signal detection properties of the forecast without being sensitive to systematic model biases (model anomalies are not bias corrected in the medium range forecast as they are for the seasonal forecast for example). That the ROCA shows a better skill for increasingly large anomalies, although surprising at first, has been reported by other centres using ensemble systems, and is likely to be one of the strong incentive for using dynamical rather than statistical methods for probability forecasts.

One of the remarkable results shown over the past few years is that the improvement of the deterministic forecast quality has translated into an improved consistency of the forecasts valid for the same date from one day to the next. This level of consistency has been further improved this year, as can be seen in Figure 7 from the time series of the average RMS difference between consecutive forecasts over Europe and the Northern Extratropics.

Finally, the performance of the deterministic forecast at predicting the flow patterns in the stratosphere has been illustrated recently by the remarkable performance in forecasting the very unusual breakdown of the winter polar vortex in the southern hemisphere last September (see ECMWF Newsletter 96). The time series of scores computed as part of the routine evaluation of the system for 50hPa height in the Northern extratropics is shown in Figure 8: June and July 2003 have reached record breaking low RMSE values, adding to the series of steps towards improvement seen over the last few years.

3.1.2. Tropics

The skill over the Tropics, as measured by root mean square vector errors of the wind forecast with respect to the model analysis, is shown in Figure 9. This year has further amplified the reduction of errors observed over the last few years both at 850 and 200 hPa. The change in trend seems to be well correlated in time with the introduction of Cy25r4 mid January this year, confirming the soundness of the modifications introduced both in the data assimilation and the physics on that occasion.

3.2 ECMWF vs other NWP centres

3.2.1. Deterministic (T511) model

The basic common ground for such an intercomparison is the regular exchange of scores between GDPS centres under WMO/CBS auspices following agreed standards of verification. Figure 10 shows time series of such scores over Northern Extratropics for both 500hPa height and Mean Sea Level Pressure. These curves confirm the very good performance achieved again this winter compared to other centres, although for winter months the gap with other models has been slightly reduced. This lead extends clearly more into the spring and summer seasons this year, not only at 500hPa but also mean sea level. The difference is even larger in the Southern Extratropics (Figure 11).

Scores with reference to each centre's own analysis over Europe are not exchanged under WMO/CBS auspices, but using forecast products exchanged on the GTS, verification statistics can be computed. Figure 12 shows distributions of 500hPa height anomaly correlations of day 5 forecasts with verifying analyses over Europe during the cold season (15 Oct.-15 Apr.) for ECMWF, the UK MetOffice, Deutscher Wetterdienst (DWD), US National Centre for Environmental Prediction (NCEP) and the Canadian Meteorological Centre. These distributions confirm the positive evaluation based on monthly hemispheric scores in the previous paragraph. The good quality of ECMWF forecasts not only shows there by an increased number of very good forecasts, but also in the fact that by Day 5, not a single 500 hPa height forecast did score worse than 50% correlation with the verifying analysis, something that none of the other forecasting systems shown here has achieved this winter.

The situation in the Tropics is summarised in Figure 13. ECMWF scores have for the first time this year reached levels that are smaller than other centres consistently since the introduction of Cy25r4 in January. This is of particular importance as the low level winds are an important forcing mechanism for both oceanic waves and sub surface ocean dynamics in the coupled model used for seasonal forecast (the version currently used for operational seasonal forecasts is however still Cy23r4)

Finally WMO exchanged scores are presented that use radiosondes as the verifying dataset over Europe (Figure 14). They confirm the conclusions drawn previously from the field verification against each model's own analysis with a model independent reference.

3.2.2. T255 (EPS Control) multi-analysis system

ECMWF has been running a multi-analysis, T255 forecasting system since March 2001 as a research diagnostic. Initial conditions for this system are kindly provided to ECMWF by the US NCEP, Deutscher Wetterdienst, Meteo-France and UK MetOffice from their own global forecasting system. After interpolation on the T255 model grid of the differences between these daily (12UTC) analyses and the one from ECMWF, 4 daily forecasts are run up to 10 days. An additional one is run using a consensus averaged from all 5 centres analyses. Verification has been collected since October 2001 for all these streams. An error in the set-up of the multi-analysis suite has been found however since then, which meant that the forecasts could not be compared with the EPS Control (differences in time step and initialisation). The situation has been corrected on 24 September 2002, but was in effect at the time when first results from the multi-analysis suite have been provided in the "Verification statistics and evaluations of ECMWF forecasts" report was produced in 2002 (ECMWF Tech. Memo. 414).

Figure 15, Figure 16 and Figure 17 show an average of scores (RMSE and anomaly correlation) over Europe, Northern and Southern Extratropics for the cold season (15 Oct. - 15 Apr.). The most noticeable feature is that in most cases, the closest to ECMWF control forecast is the one run from a consensus of all analyses. Clearly the benefit gained from using ECMWF own analysis is the biggest in the Southern Extratropics (Figure 17), where up to 6 days the control (ECMWF T255) forecasts outperforms all other runs, sometimes by a considerable amount. Results shown last year for Europe when running the model from either the consensus or NCEP analyses outperformed the EPS Control was clearly resulting from the inconsistent set-up of the experiment. None of this can be found this year.

4. Weather parameters and oceanic waves

4.1 Extratropics

Figure 18 shows the monthly mean and standard deviation of the 2m temperature and specific humidity errors over Europe up to July 2003, verified against synoptic observations (a correction for the difference between model and true orography was applied to the temperature forecast error). The springtime warm bias that did characterise 2000 and 2001 has been reduced for the second year in a row. The night time cold bias in February and March has been larger than has been the case over recent years. Although it has by no means reached the amplitude of the 1995-96 bias, careful examination has suggested that the model misrepresentation of heat fluxes near the ground when early spells of warm weather reaches soils with growing but still low vegetation cover might be to blame. This will remain an area for further research and will be carefully monitored in the coming months. The average performance of temperature forecasts however remains very good, with both daytime and nighttime error standard deviations from observations reaching record-breaking low values in July – it is worth noting that these standard deviations have been reduced by more than 1°C during daytime since the late eighties. The trend for specific humidity is also showing a reduction of errors – it should be noted that a small peak in daytime humidity biases in March coincides with the cold biases, supported the idea of an incorrect balance between latent and sensible heat fluxes. Measures of forecast skill with reference to persistence of 2m temperature forecasts are shown in Figure 19 separately for daytime and nightime conditions up to 5 days range. They show that there is still some skill in the deterministic forecast up to day 5, something that was not the case in the late nineties. Although there is a large level of season-to-season variability, it seems that there has been a step towards increased skill taken around 1999/2000, when both horizontal and vertical resolution have been increased.

Figure 20 shows monthly bias and standard deviations from observations for total cloud cover and 10m wind speed forecasts. A small trend towards increasingly negative daytime total cloud cover biases seem to appear that will deserve more investigations. On the positive side, the standard deviation from observed wind speed has reached in July unprecedented low values.

2002-2003 has been characterised in Europe by rather extreme temperatures, first with a cold winter associated with a large number of blocking events, and finally with a heat wave in July/ August that has been responsible for a large number of forest fires, and later to a humanitarian disaster when thousands of elderly people died from excessively hot weather conditions, mainly in France. Although we have shown that the deterministic forecast had skill in forecasting 2m temperature up to 5 days ahead, and we have addressed the skill of 850hPa temperature forecast maps over Europe in section 3.1.1, EPSgrams have been since 2000 another way to convey the information on ECMWF web service about both deterministic and ensemble forecasts for any given location over land. Some discussion relevant for the verification of the forecast parameters (deterministic values, EPS median and spread) offered on these graphics follows.

Time series of observed 2m temperatures and day 6 forecasts at London/Heathrow during the last winter (DJF) and summer (15 May-15 August) can be found in Figure 21 as an example. They confirm that the model has now not only skill in forecasting the large-scale weather patterns at this range, but also local parameters such as temperature. The deterministic, T511 model generates extreme values that compare well with observations, both on the cold and hot side, even though the precise timing can be shifted by one or two days at this forecast range. Table 1 summarises some basic verification statistics for both London Heathrow and Vienna, separately in winter and summer this year. Results there confirm the skill of the model compared to the forecast provided by a linear trend over the period, whether it is measured by the reduction of standard deviation of errors or by the correlation with observed anomalies. This skill is reduced for the low resolution T255 EPS Control. However the median of the ensemble, although a parameter that is less likely to generate extreme values, has more skill than the high resolution deterministic forecast both in correlation and standard deviation. It can also be seen in Figure 21 that in most cases, the EPS did generate a range of forecasts that did include verification (small number of "outliers" outside the range of the EPS min/max)

Table 1: Mean skill of Day 6 forecasts of 2m temperature at Heathrow and Vienna in 2002-2003. Correlations are taken between Observations and Forecasts from which the observed trend has been removed.

Location	(Obs – Trend) STD	(T511-Obs)		(T255-Obs)		(Med. – Obs)	
		Std	Corr	Std	Corr	Std	Corr
Heathrow (winter)	3.40K	2.71K	70.0%	2.90K	63.9%	2.30K	74.0%
Heathrow (summer)	3.59K	3.23K	60.5%	3.03K	60.5%	2.73K	65.5%
Vienna (winter)	4.31K	3.77K	57.3%	3.89K	54.5%	3.44K	62.3%
Vienna (summer)	4.11K	3.94K	49.0%	4.19K	48.7%	3.41K	52.0%

On Figure 21 the EPS extremes (ensemble min/max) and quartiles have also been reported which show that there is without doubt a large day to day variability in the spread of ensemble forecasts for a given forecast range and location. Whether or not these variations of spread can be related to similar variations in the distribution of errors is discussed hereafter.

EPSgrams highlight as spread the difference in value between the first and last quartile of the ensemble forecast ("blue box", see Figure 22). A perfect probabilistic forecast should be one when verifying observations lie outside the "blue box" with a frequency of occurrence of 50% exactly. If the error is measured by the absolute difference between the observed value and the median of the ensemble forecast, then (provided positive and negative errors can be assumed to have similar behaviour) exactly one half of the errors should exceed the standard deviation defined as (Q75-Q25)/2 (observation outside the blue box), while the other half should be smaller (observation in the blue box). In other words, the median of the distribution of errors should be exactly (Q75-Q25)/2 (Figure 22).

The distribution of both the EPS spread and absolute errors are pictured in Figure 24 for Day 6 forecasts of 2m temperature at Heathrow in 2002-2003. Clearly there is no correlation between the spread of the day and the error observed later. This however could have been anticipated, as the EPS is not expected to provide a deterministic estimation of errors but, at best, a reliable estimate of their probability distribution. An empirical estimate of the match between the forecast and verifying distributions can only be found if criteria are set that define similar properties that can be verified. A classical example is the reliability diagram, when 1) an event (e.g. T<0° C) is defined, 2) forecasts reaching a similar level of probability are gathered, and finally 3) the frequency of occurrence the event in observations conditional to the forecast reaching a certain probability value is estimated. The same methodology can be applied for the verification of the spread of the ensemble: if we define different categories of forecast spread, we should see the median of the forecast error distribution conditional to the spread falling in that category matching the forecasted spread. The different values of the median of the error distribution for different sizes of the conditional samples are shown in Figure 24. Although they seem to support the idea that the EPS spread values verify reasonably well when faced with conditional error distributions, the fit only converges slowly when the size of the samples increases. Over a year, defining 12 or even 6 categories of spread clearly leads to a very noisy estimate of the conditional error distribution – only when 3 categories (each populated by 121 events) are defined does the estimate of the conditional medians converge reasonably. The fit to the diagonal is then surprisingly good, with only a slight tendency to overestimate the very small values of spread.

In order to further discriminate between winter and summer time conditions, more than one verifying station is needed due to the very slow convergence of the empirical estimates of the medians. To keep the sample reasonably homogeneous while sampling different weather events, 8 stations from the Northern European plain have been selected (WMO Ids 03257, 03772, 06179, 06240, 06451, 07038, 07168, 10147, see Figure 23). The values of the EPS spread are verified for these stations separately in winter and summer in Figure 25. In winter, the values of spread correlate very well with the conditional distribution of errors, but there is clearly a small underestimation of the spread for all categories with the exception of the largest one. The verification for summer shows consistent results. The same verification for daily amounts of rain is shown in Figure 26. Again the fit of verification with the forecast is surprisingly good in winter – indeed this may reflect a high correlation between the spread and the amount of rain, in which case a reasonably good estimate of the spread good be made by using MOS techniques based on purely deterministic forecasts – something that should be investigated further. The verification for summer though is not as good as it is in winter. This might have been anticipated, as verification here is made using local observations: the subgrid scale variability of precipitation fields is likely to be higher during the convective season, something that cannot be accounted for explicitly by the ensemble. The fact that the spread is *overestimated* in summer is indeed surprising, and might be related to some overactive perturbations in the stochastic physics to be further investigated.

Back to the validation of deterministic forecasts of precipitation, the monthly mean error of the precipitation forecasts at day 3 over Europe is shown in Figure 27, for 00, 06, 12 and 18 UTC - again this year the biases seem to have been reduced, confirming the trend found in the previous years. This trend is confirmed when looking at the time series of Equitable Threat Scores (Figure 28). A WGNE initiative aiming at providing consistent intercomparison of precipitation forecasts has now reach a stage when several Meteorological services are providing a validation of different models (including ECMWF) using their observation network. Meteo-France has recently generated a consistent set of verification data from which Figure 29 has been extracted. It provides for a range of precipitation thresholds (from 0.1 mm/day up to 16mm/day) the probability of detection (Number of good positive forecasts over total number of occurrences of the event) and the false alarm ratio (number of wrong positive forecasts over total number of positive forecasts).

Observations have been upscaled to the model grid size. Good forecasts are close to the upper left corner of the diagram, while the distance from the diagonal indicates the Frequency Bias. ECMWF compares extremely well with other centres on this graph.

Figure 30 provides a probabilistic evaluation of EPS forecasts of precipitation in the usual form using Brier skill scores and in the form of ROC area time series similar to the ones previously shown for 850hPa temperature anomalies (Figure 6). Here again the peak performance reached during winter 2001-02 over Europe is not reproduced, but there seems to be some positive trend in the scores for moderately large amounts of rain (more than 10 and 20 mm/day).

4.2 Tropics

Verification of Tropical cyclones forecasts has again this year been conducted on a routine basis, both in deterministic and probabilistic mode (strike probability maps). The T511 deterministic forecast remains the most accurate both in terms of position and intensity (Figure 31). Strike probabilities that are about to be made available on ECMWF web service have also been verified with observation reports (Figure 32). The low probabilities are indeed quite reliable, but higher values are overconfident, which is probably related to the earlier stages of the forecast when the ensemble only has very small spread.

As an independent evaluation of ECMWF forecasts, Figure 33 and Figure 34 show results provided to ECMWF by Hong-Kong Observatory for North Western Pacific (0-45N, 105-125E). They show a nice reduction of errors for 2002. Comments included by Hong-Kong Observatory in their report follow:

1. The overall mean position errors in 2002 were significantly smaller than those for previous years. It is noted that the error for T+72h was smaller than that for T+48h. This is probably related to the small sample size of the verification data set. With a larger sample size in the expanded verification area, the mean errors increased with forecast range;
2. From the trend of annual mean errors as shown in [Figure 33], the performance of subjective forecasts follows closely the trend of ECMWF model performance in the past few years. Once again, this is a reflection of the contribution of ECMWF model outputs in HKO’s forecasting operations. The model skill in 2002 improved significantly compared with that in the past couple of years as shown in [Figure 34]
3. There was an apparent reduction of nearly 20% in the mean position errors in 2002 where the coverage of Quikscat data was good for TC initialisation.

4.3 Oceanic waves

Verification scores from the global oceanic wave products are shown in Figure 35 and Figure 36. They show that the steady improvement of the forecasts has continued this year, with a clear improvement of the Day 5 forecasts in both hemisphere.

5. Seasonal forecasts

5.1 The 2002-2003 El Nino forecasts

During late spring and summer 2002 an El Nino event with moderate amplitude developed and peaked around December 2002. The forecasts for Nino-3.4 have generally been good for this El Nino. Figure 37 shows Nino-3.4 predictions throughout the year with subsequent verification (heavy blue dashed line). In October and later November 2002 strong westerly winds had propagated from the Indian Ocean to the

Pacific. In response to that, El Nino could have intensified. However the ECMWF forecasts showed no intensification of El Nino, but rather a relatively rapid decline from November/December, which has subsequently verified well.

Although the Nino-3.4 forecasts were generally very good, some other regions showed errors: Nino-4 forecasts underestimated the persistence of warm conditions in the late months of 2002 and the early months of 2003, for example. This error was perhaps consistent with an earlier finding that this version of the seasonal forecast system does not perform particularly well in this region. The latest El Nino forecasts present a relatively large spread and indicate that a more likely scenario is for SSTs to be near normal.

5.2 Seasonal Forecast performance during 2002-2003

In summer 2002, when the El Nino was developing in the tropical Pacific, severe flooding in central Europe and a very dry monsoon were observed. It is quite possible that these events were connected. For example it is well known that drought years over India are associated with warm SST anomalies in the equatorial central and East Pacific. Although this relationship is far from perfect it is clear that the monsoon and ENSO are related in some fundamental manner. Rodwell and Hoskins (2000) suggested that the establishment of the typical summer weather pattern over Europe is linked with the monsoon circulation. Normally, as the monsoon season arrives, huge volumes of air begin to rise over the Indian subcontinent. These push north and west, allowing other columns of air to descend over the Mediterranean region, causing stable, high-pressure systems to develop, which repel unstable, rain-bearing weather. When the ascent associated with the monsoon is weak the descent over Southern Mediterranean is also weak so that low pressure systems can develop in the area. Precipitation anomalies from GPCP data averaged for June to August 2002 (Figure 38, upper panel) shows clearly the lack of rain over the Asian monsoon region, wet conditions over central and southern Europe and over central and Eastern tropical Pacific. Predictions started in May 2002 (Figure 38, lower panel) shows probabilities of about 70% of a dry summer over the Arabian sea over the most South tip of India, Madagascar and Indonesia. However forecasts underestimated the extension of dry anomalies over the Indian continent and over the Bay of Bengal. Probabilities of about 50% that dry anomalies would occur over Southern Europe were predicted. This might indicate the effect of the monsoon anomalies over the Mediterranean region in this case was not well represented by the seasonal forecast system. Further investigation to establish this is in progress.

During the past winter cold anomalies with large amplitudes persisted over Central Europe. None of the seasonal predictions ensembles showed high probability of cold anomalies over that area. In contrast, over the same area, more than one ensemble prediction gave probabilities of warm anomalies.

Comparison between seasonal predictions and atmospheric integrations forced by observed SST indicated that such anomalies did not respond directly to anomalies in the oceanic conditions, making them far less predictable than the ones observed over the North America sector.

For this particular cold event the seasonal forecast ensembles did not generate enough spread to include the observed anomaly. The observed temperature anomalies were even lying outside the model climate probability distribution in December (Figure 39). A larger ensemble spread and/or the use of a calibrated forecast might have improved those predictions - if not by adding predictability, at least by correctly indicating the lack of it when appropriate. In fact, for a correct interpretation of seasonal predictions the user needs to complement the forecast products with knowledge of the forecast skill. The site at http://www.ecmwf.int/products/forecasts/d/charts/seasonal/verification provides a comprehensive documentation of skill levels, using methods that have been agreed at the international (WMO) level for the evaluation of long-range forecast systems. Since last year the verification has been extended. A larger number of time-series is verified and the ranked probability skill scores have been introduced as an additional skill measure. Verification of the UKMO seasonal forecasts have been implemented and is available internally.

6. Summary

The forecasting system has reached again this year unprecedented levels of skill in many areas, both on the large scale (500hPa scores over Northern and Southern Extratropics, tropical wind errors at 850hPa and 200hPa) and in terms of weather parameters (smallest standard deviation in terms of 10m wind speed errors and smallest precipitation biases over Europe). The consistency of the 500hPa height forecasts valid for the same date has also been further increased. Measures of skill for 2m temperature and daily precipitation forecasts show that there is now some skill in the deterministic forecast of these parameters up to 5 days in Europe. Areas of concern however remain like the small biases developing in daytime cloud cover and early spring 2m temperatures, although the problem is far from reaching the scale that was observed in 1995-96. The difference in performance between the low and high resolution (both in time and space) versions of the model used in the medium range has also been found to have increased over the last few years, and to occasionally affect the EPS in failing to predict correctly the initial stages of development of mid latitude systems. Verification of the EPSgrams (median and inter-quartile spread) has however confirmed the average good level of performance of the system, even when compared to higher resolution forecasts. Comparison to other centres performance has confirmed ECMWF lead not only for the forecast of the large scale weather patterns, but also for daily amounts of rain (WGNE/Meteo-France study).

Finally the performance of the seasonal forecast system has been quite good in the tropical Pacific. Although Europe has been affected by large scale anomalies of the weather both during the summer and the winter, these seem not to have been strongly influenced by the oceanic circulation and therefore were not predicted accurately by the coupled model.

Annex A: A short note on scores used in this report

A.1 Deterministic upper-air forecasts

The verifications used follow WMO/CBS recommendations as closely as possible. Scores are computed from forecasts on a standard 2.5 x 2.5 grid limited to standard domains (bounding co-ordinates are reproduced in the figure inner captions) as this is the resolution used for most products exchanged on the GTS; when other centres' scores are produced, they have been provided as part of the WMO/CBS exchange of scores among GDPS centres, unless stated otherwise (Figure 12); when verification scores are computed using radiosonde data (Figure 14), the sondes have been selected following an agreement reached by data monitoring centres and published in WMO/WWW Operational Newsletter.

Root Mean Square Errors (RMSE) are the geographical average of the squared differences between the forecast and the analysis valid for the same time; when models are compared, each model uses its own analysis for verification; RMSE for winds (Figure 13, Figure 14) root the sum of the mean squared errors for the two components of the wind independently; Skill scores (Figure 1) are computed as the reduction of the RMSE which the model achieves with respect to persistence (forecast obtained by persisting the initial analysis over the forecast range); in mathematical terms:

$$SS = 100 * (1 - \frac{RMSE_f^2}{RMSE_p^2})$$

Figure 2, Figure 3, Figure 4, Figure 12, Figure 15, Figure 16 and Figure 17 are correlations in space between the forecast anomaly and the verifying analysis anomaly; anomalies with respect to NMC climate are available at ECMWF from the start of its operational activities in the late 1970s. Only for oceanic waves (Figure 35 and Figure 36) has the climate been derived from ECMWF analysis

A.2 Probabilistic forecasts

Events usually defined for the verification of medium-range probabilistic forecasts are anomalies with reference to a 10-year model climatology (1984-1993). This climatology is often referred to as the long-term climatology, as opposed to the sample climatology, which is simply the collation of the events occurring during the period considered for verification. Probabilistic skill is illustrated and measured in this report in the form of Brier Skill Scores and the area under Relative Operating Characteristics (ROC) curves.

The Brier Score (BS) is a measure of the distance between forecast probabilities and the verifying observations (which, as any deterministic system, takes only 0 or 1 as values). For a single event, it can be written as:

$$BS = (p - o)^2$$

As any probabilistic score, however, the BS only becomes significant when results are averaged over a large sample of independent events. Then its values range from zero (perfect, deterministic forecast) to 1 (consistently wrong, deterministic forecast).

the Reliability Skill Score is defined as:

$$BSS\_REL = (1 - \frac{BS\_REL}{BS_{cl}})$$

Time series of the Brier Skill Scores can be found in Figure 6 and Figure 30.

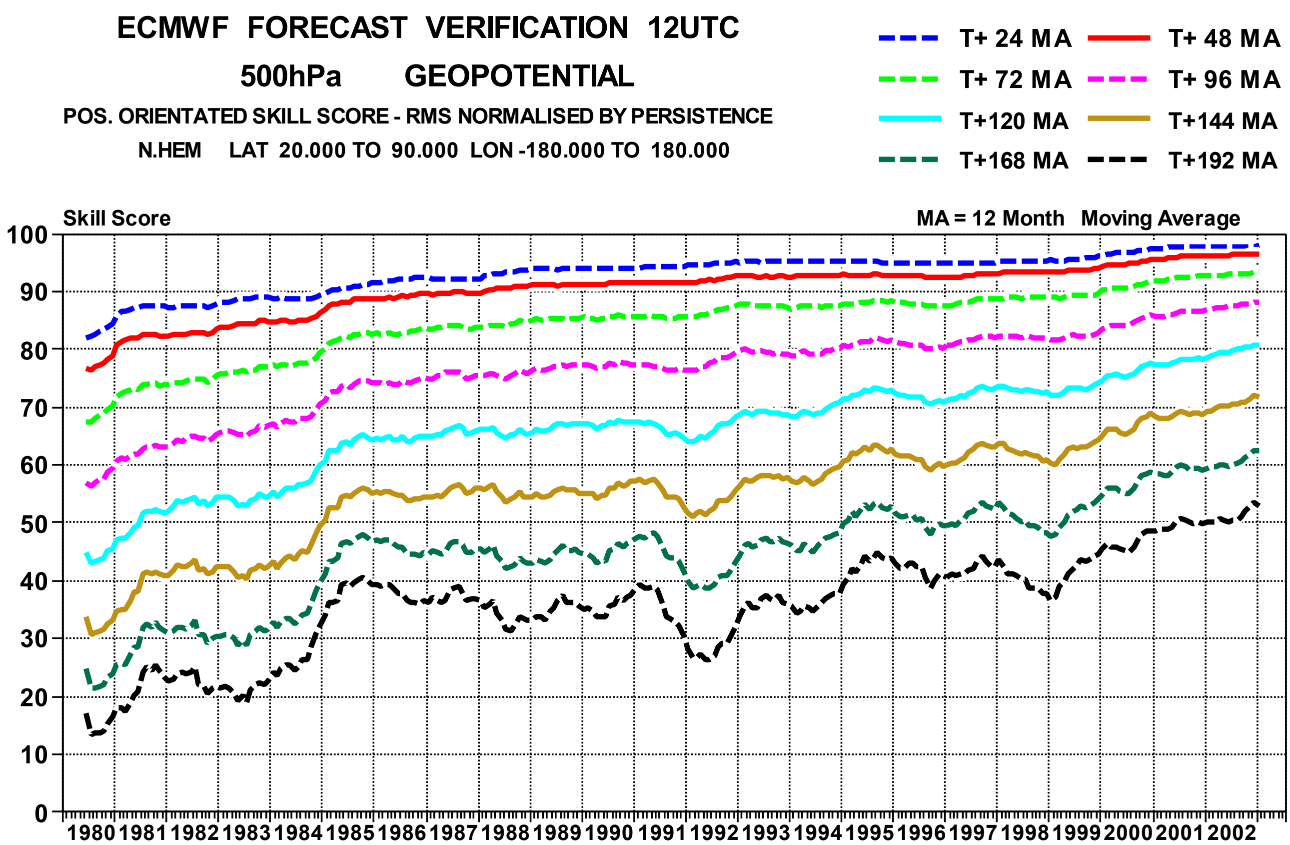
Relative Operating Characteristics curves show how much signal can be picked from the ensemble forecast: although a single valued forecast can be characterised by a unique false alarm (x-axis) and hit rate (y-axis), ensemble forecasts can be used to detect the signal in a different way depending whether one is more sensitive to the number of hits (the forecast will then be issued event if a relatively small number of members go for the event) of false alarms (one will then wait for a large proportion of members to forecast the event); the ROC curve simply shows the false alarm and hit rates associated to different thresholds (proportion of members, or probabilities) used before the forecast will be issued. Because the closer to the upper left corner (0 false alarm, 100% hits) the better, the area under the ROC curve (ROCA) is a good indication of the forecast skill (0.5 is no skill, 1 is a perfect detection). Time series of the ROCA are shown in Figure 6 and Figure 30.

A.3 Weather parameters (section 4)

Verification data are European 6-hourly SYNOP data (limiting area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the 4 closest grid points, provided the difference between the model and true orography is less than 500m. A crude quality control is applied to SYNOP data (maximum departure from the model forecast has to be less than 100mm, 25K, 20g.kg-1 or 15m.s-1 for precipitation, temperature, specific humidity and wind speed respectively). 2m temperatures are corrected for model/true orography differences using a crude constant lapse rate assumption, provided the correction is less than 4K amplitude (data are otherwise rejected).

When verification against analyses is mentioned for EPS forecasts of rainfall amounts (Figure 30), the 0-24h-model forecast is used as a proxy for a model-scale analysis.

List of Figures



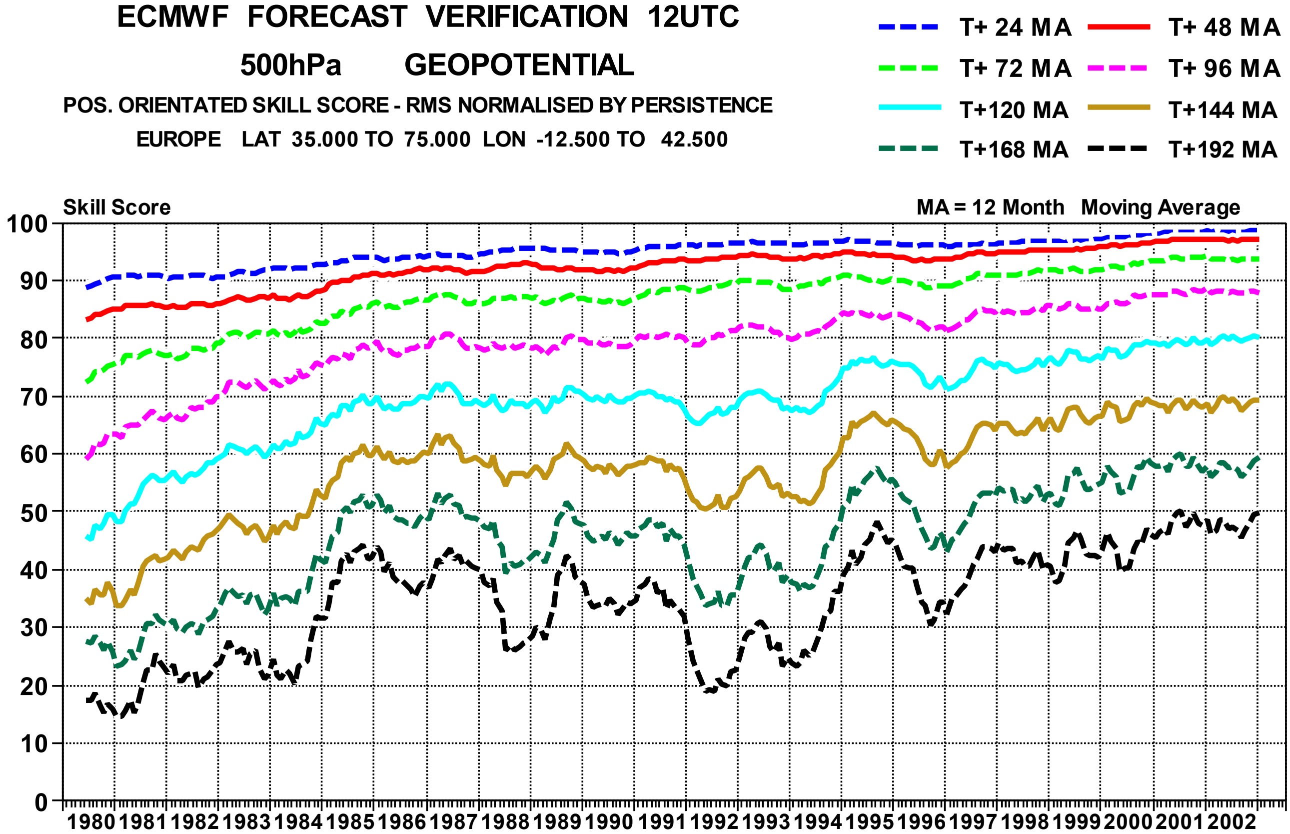


Figure 1: 500hPa height skill score (N. Hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)

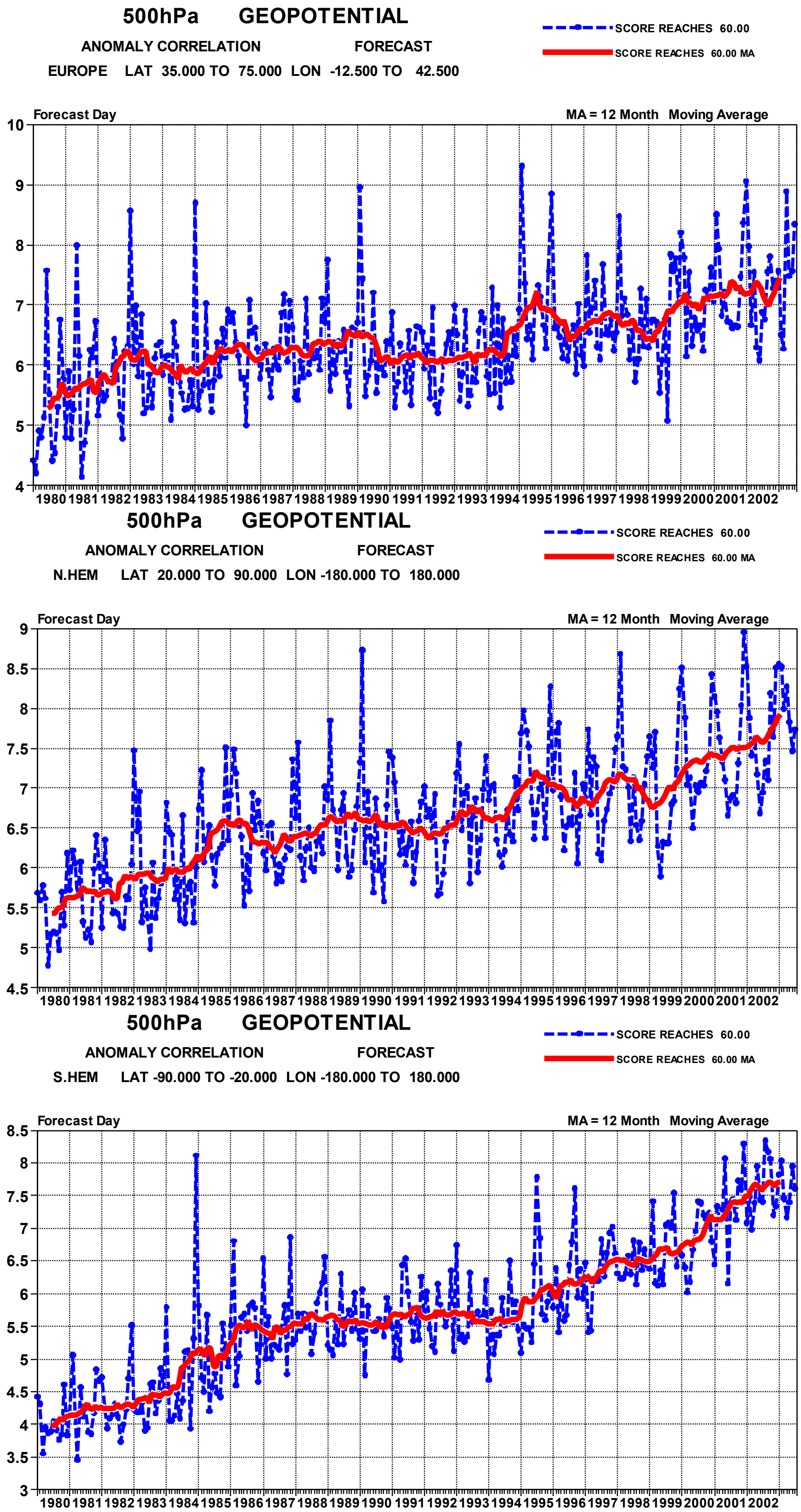
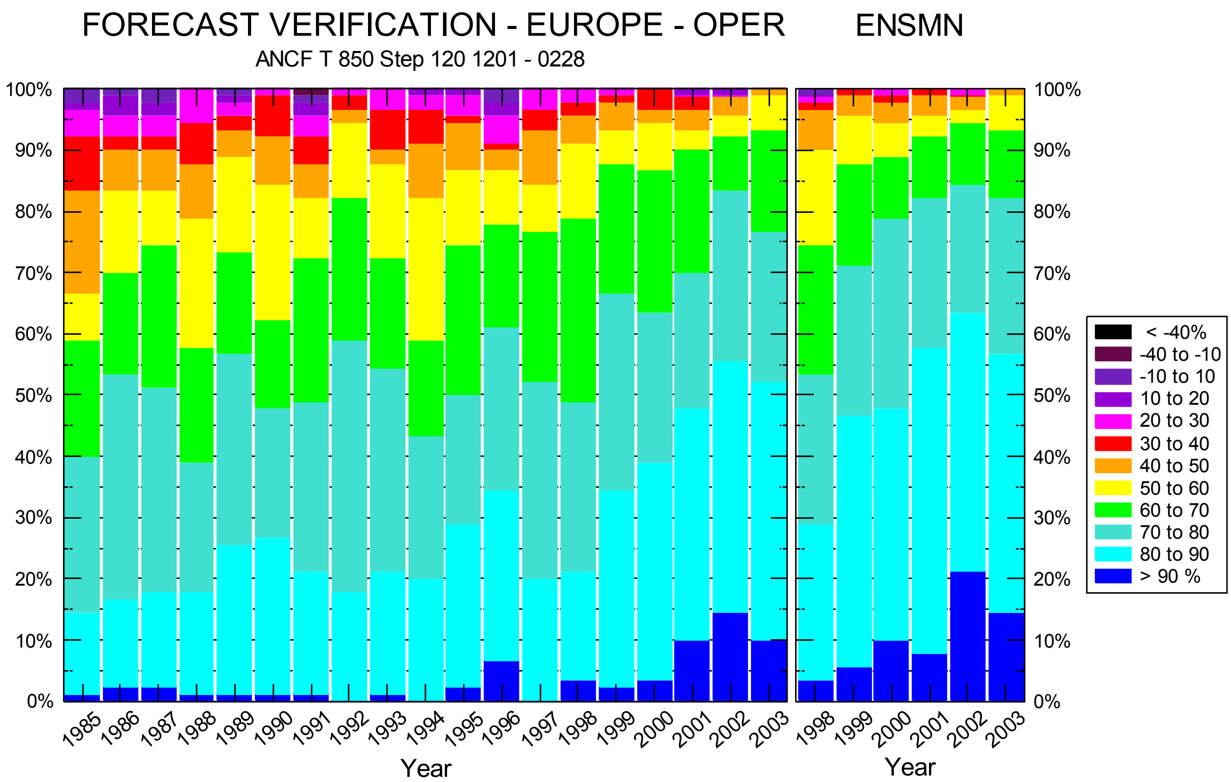


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the curves is the forecast range when the monthly average of the daily forecast anomaly correlation with observation (analysis) is falling below 60% for Europe, Northern and Southern Extratropics (the red curve is the 12-month moving average)



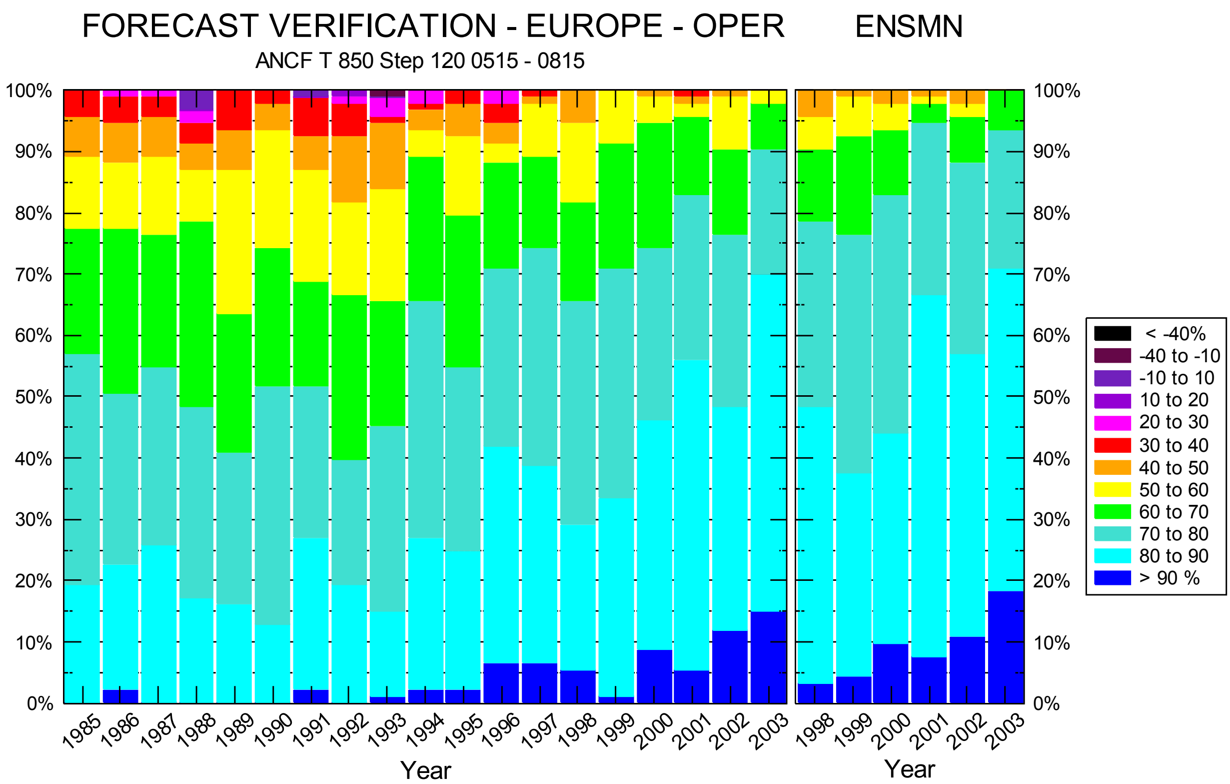
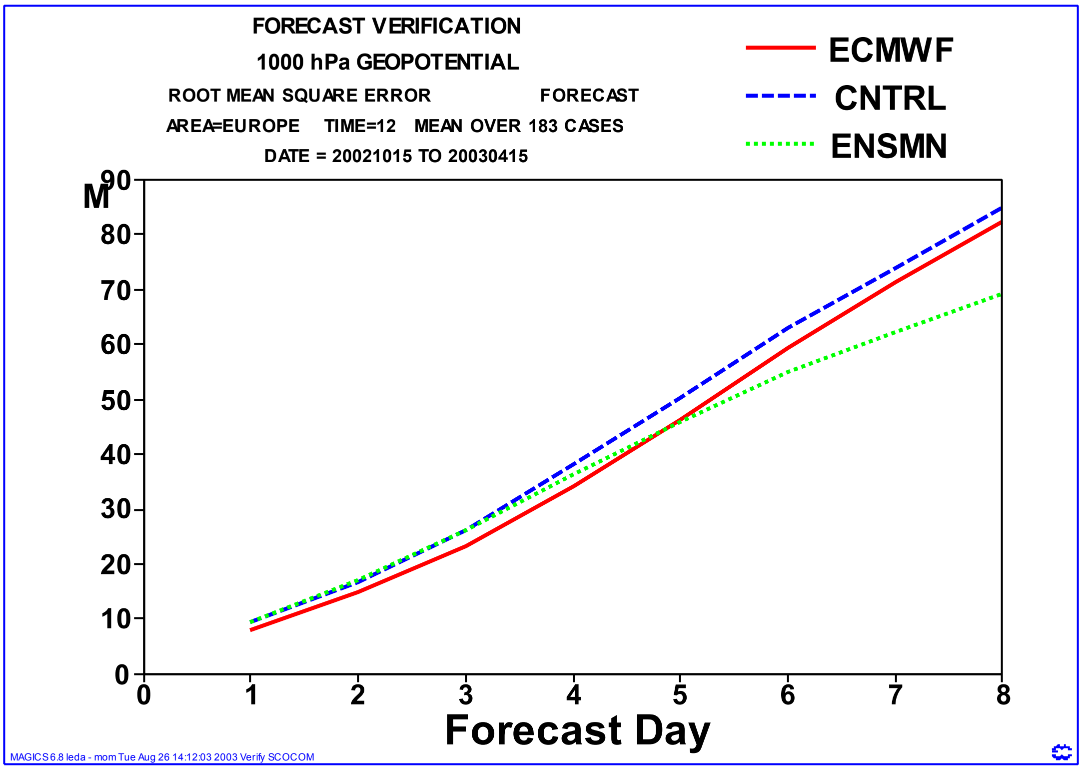


Figure 3: Cumulative distribution of Anomaly Correlation of 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (15-May-15 Aug., bottom) since 1984-85 for the deterministic, high resolution forecasts (left panels) and since 1996-97 for the EPS Ensemble mean (right panels).



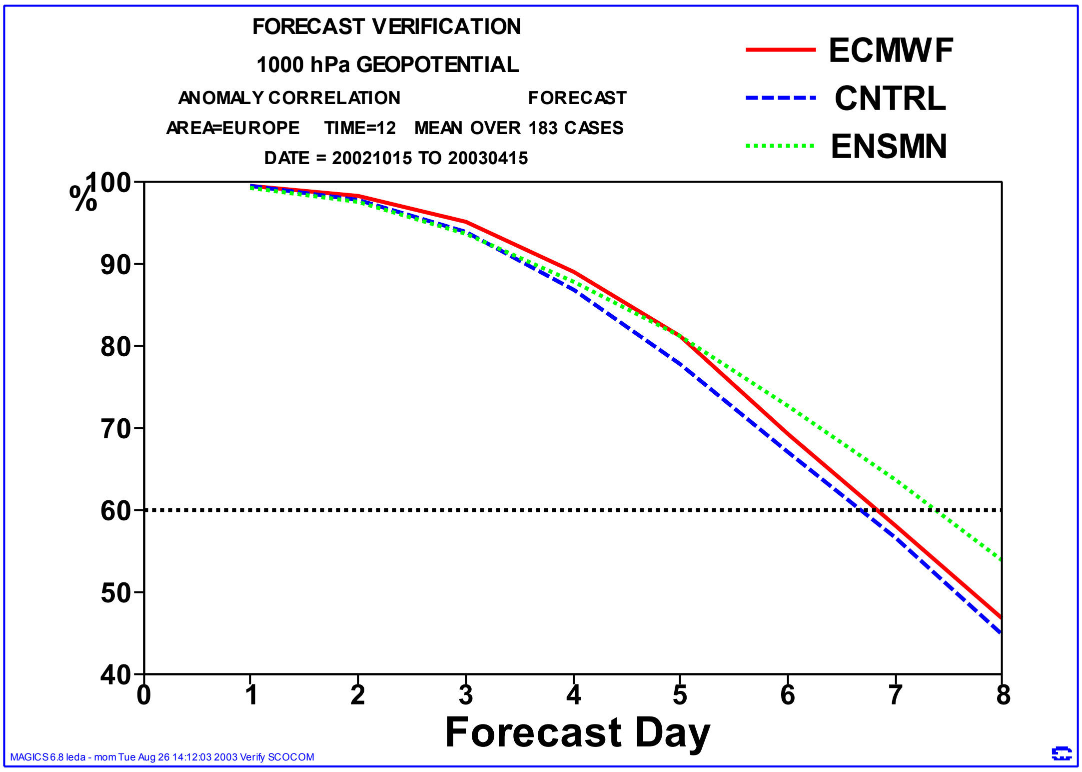


Figure 4: Comparison between scores from T511 (OPER, red), T255 (CNTRL, blue) and EPS ensemble mean (green) during the last cold season (15 Oct.-15 Apr.) over Europe for 1000hPa height forecasts.

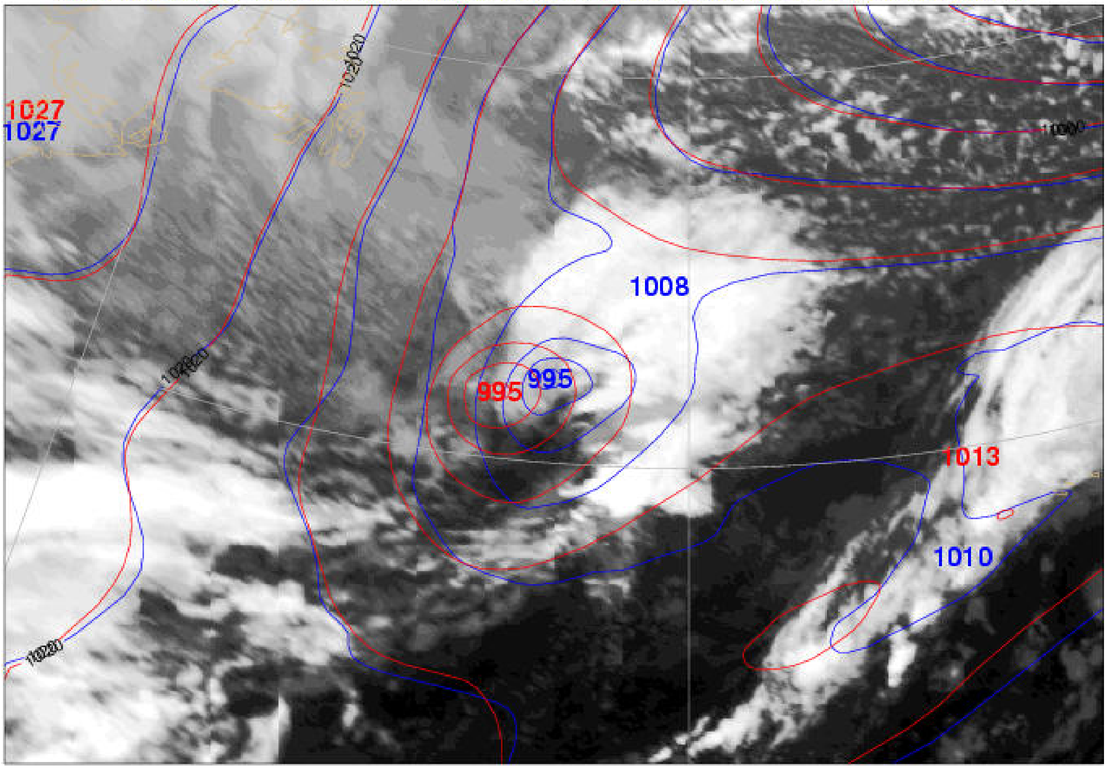


Figure 5: Early stage of development of the 27 October storm. 18h mean sea level pressure forecasts of the storm valid for 06UTC on 25 October are overlaid as blue (T511) and red (T255 EPS control) contours on the Meteosat 7 IR picture.

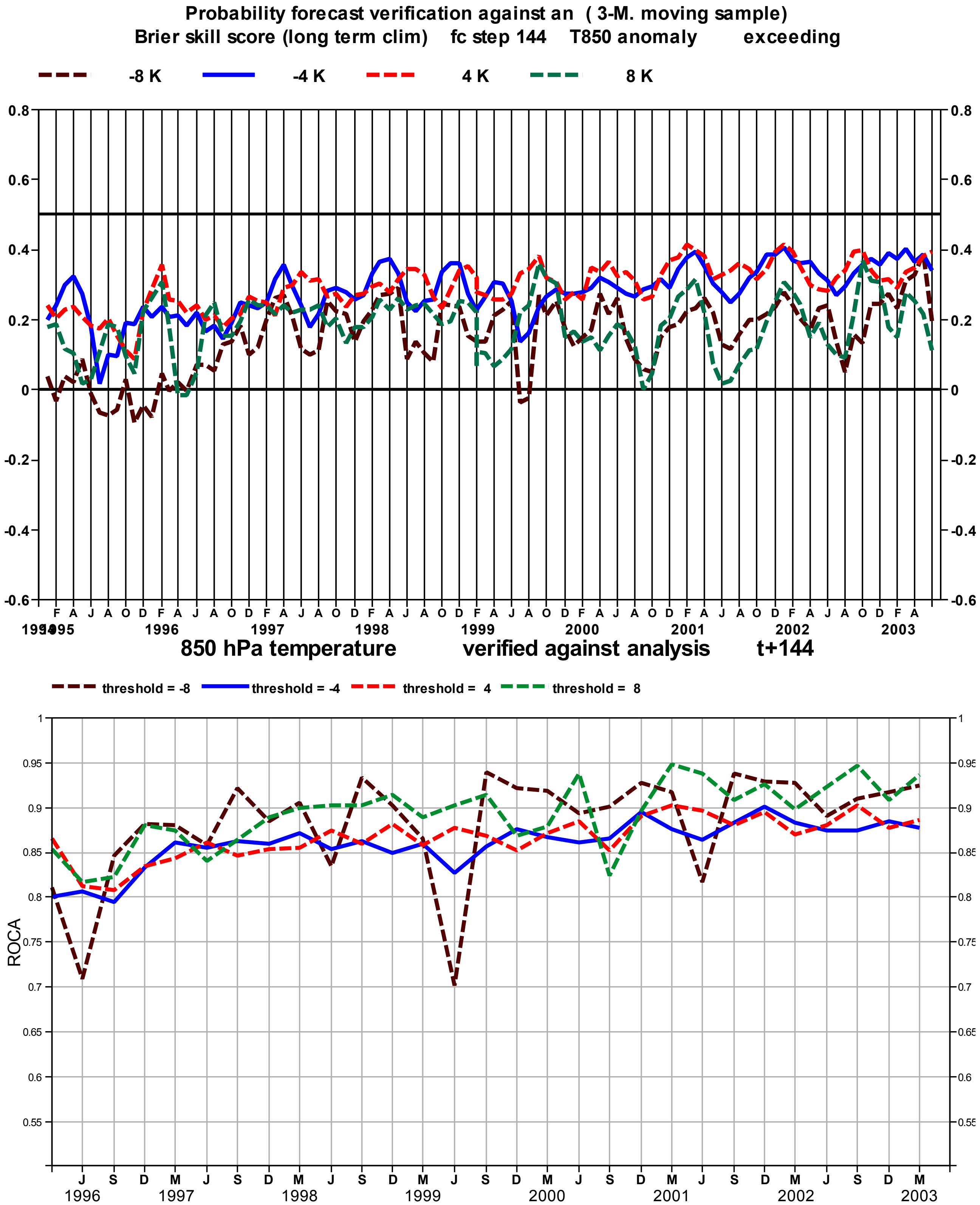
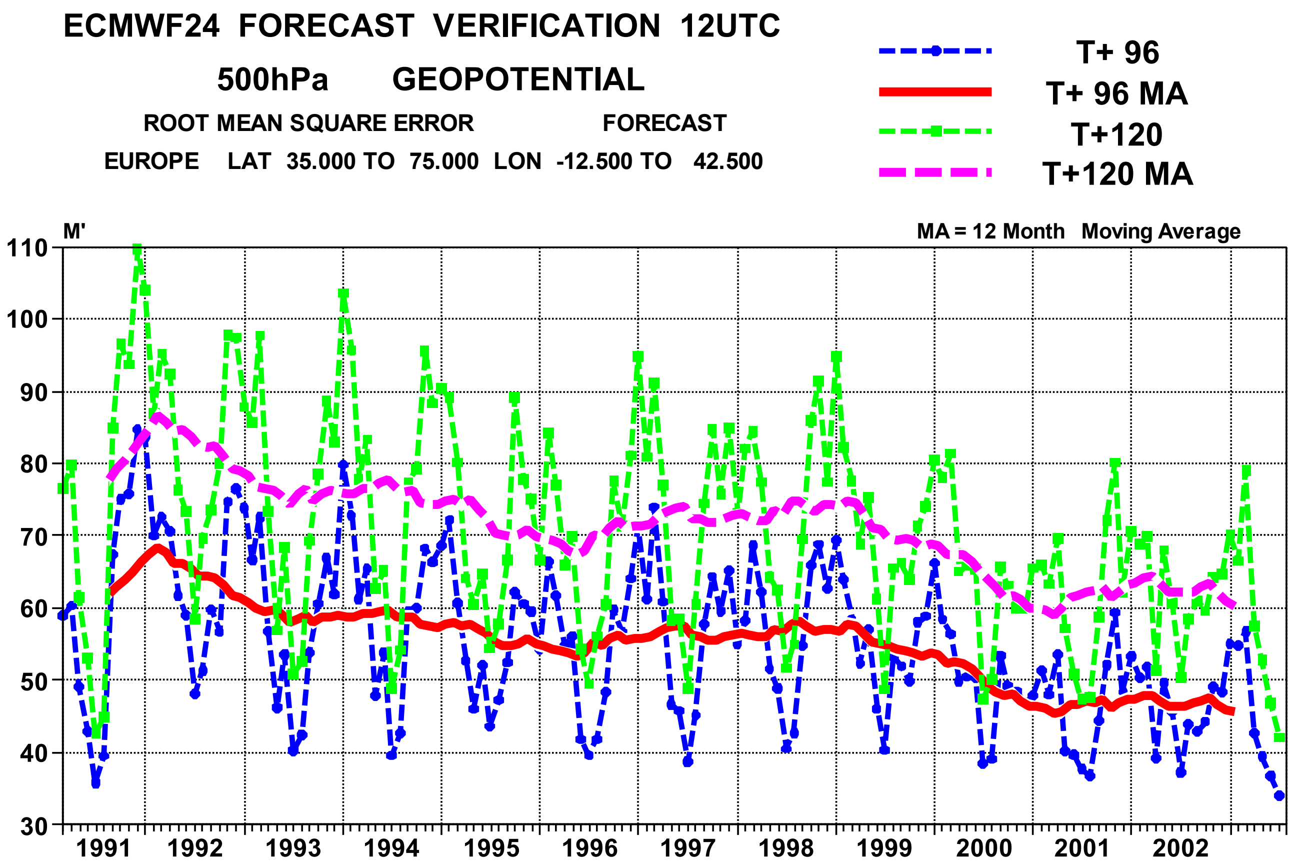


Figure 6: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics curve Area (ROCA, lower panel), the later showing the skill shown by the EPS at detecting a signal verified by the analysis out of the Day 6 probability forecast of 850hPa anomaly temperature (0.5 is no skill, 1 is a perfect detection).



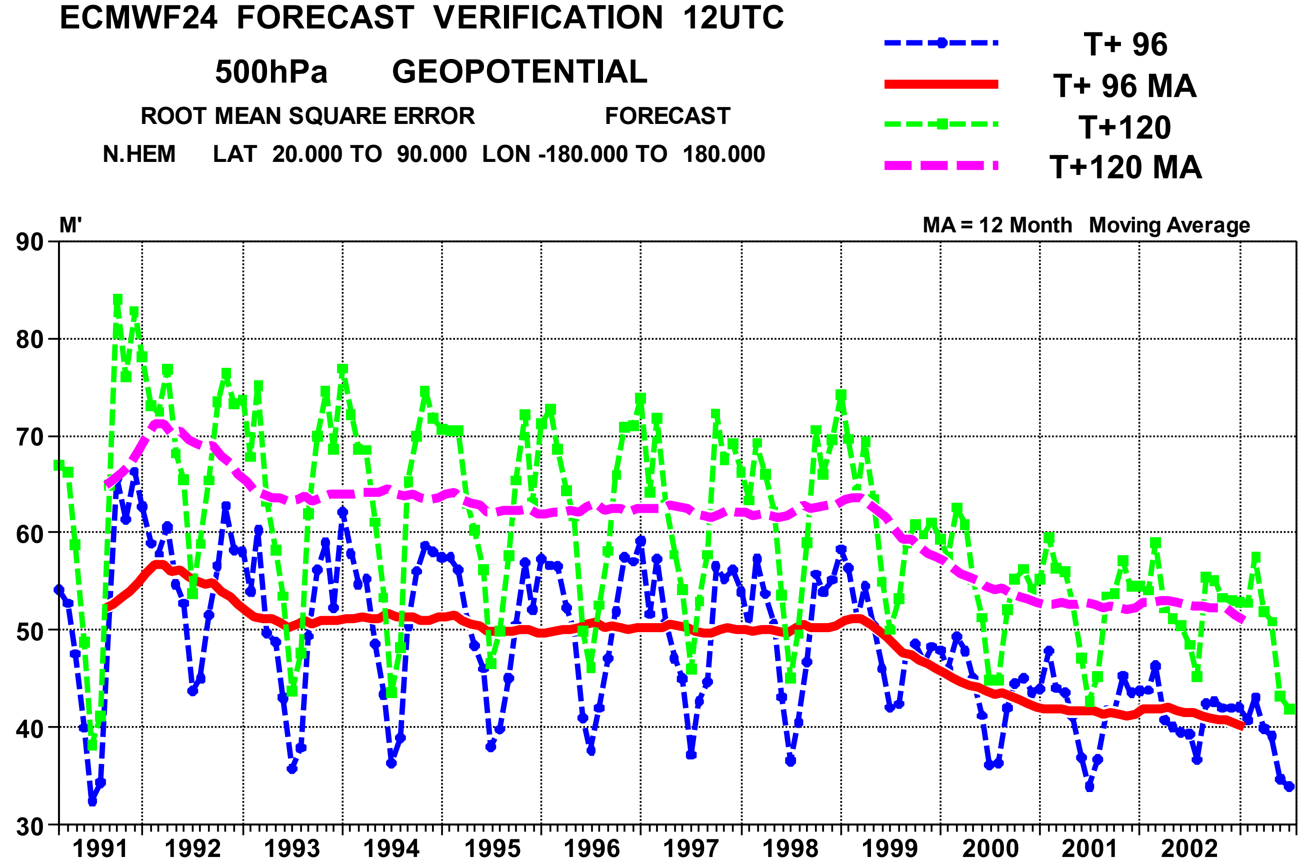


Figure 7: RMS of the difference between 24h-consecutive 500hPa height forecasts verifying the same day over Europe (upper panel) and Northern Extratropics (lower panel).

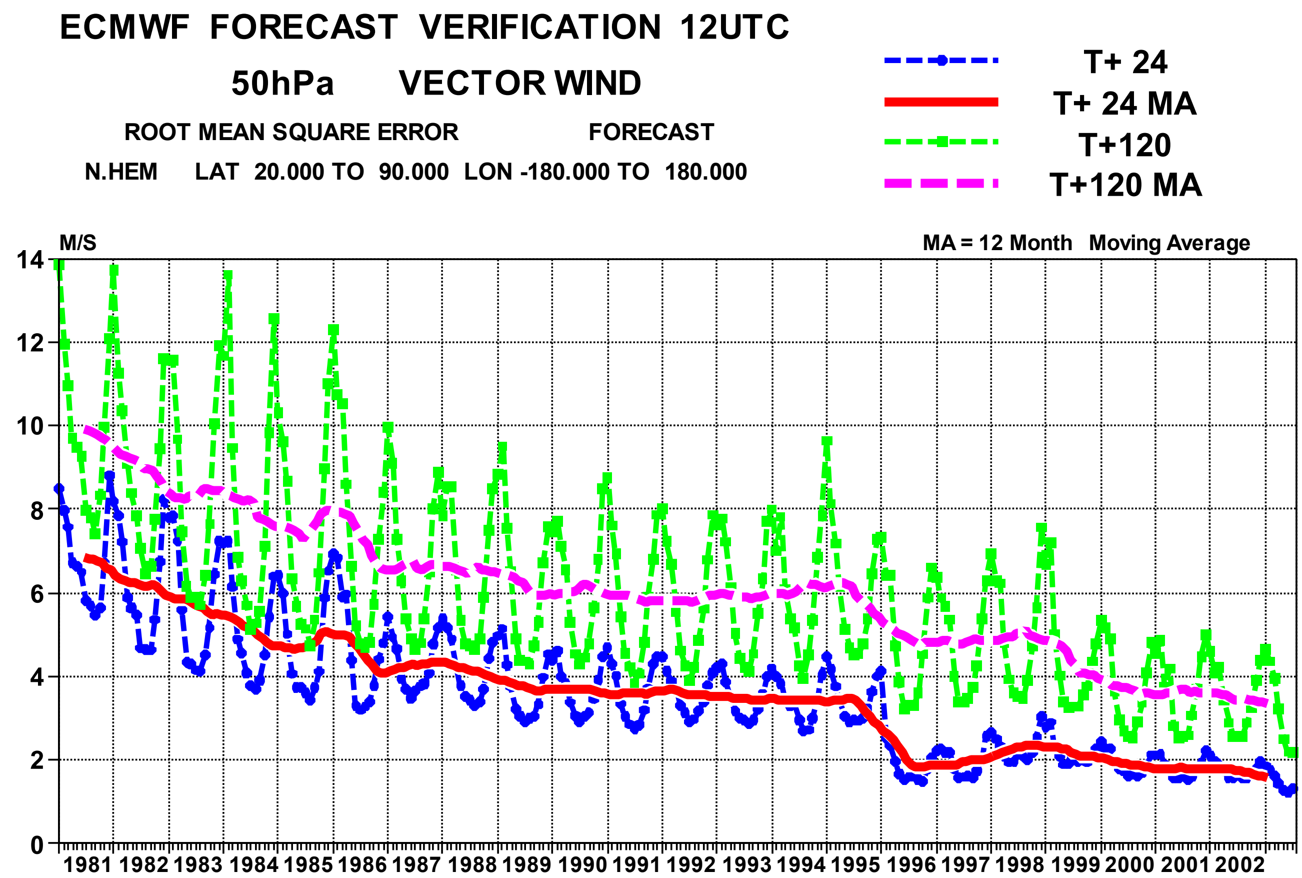
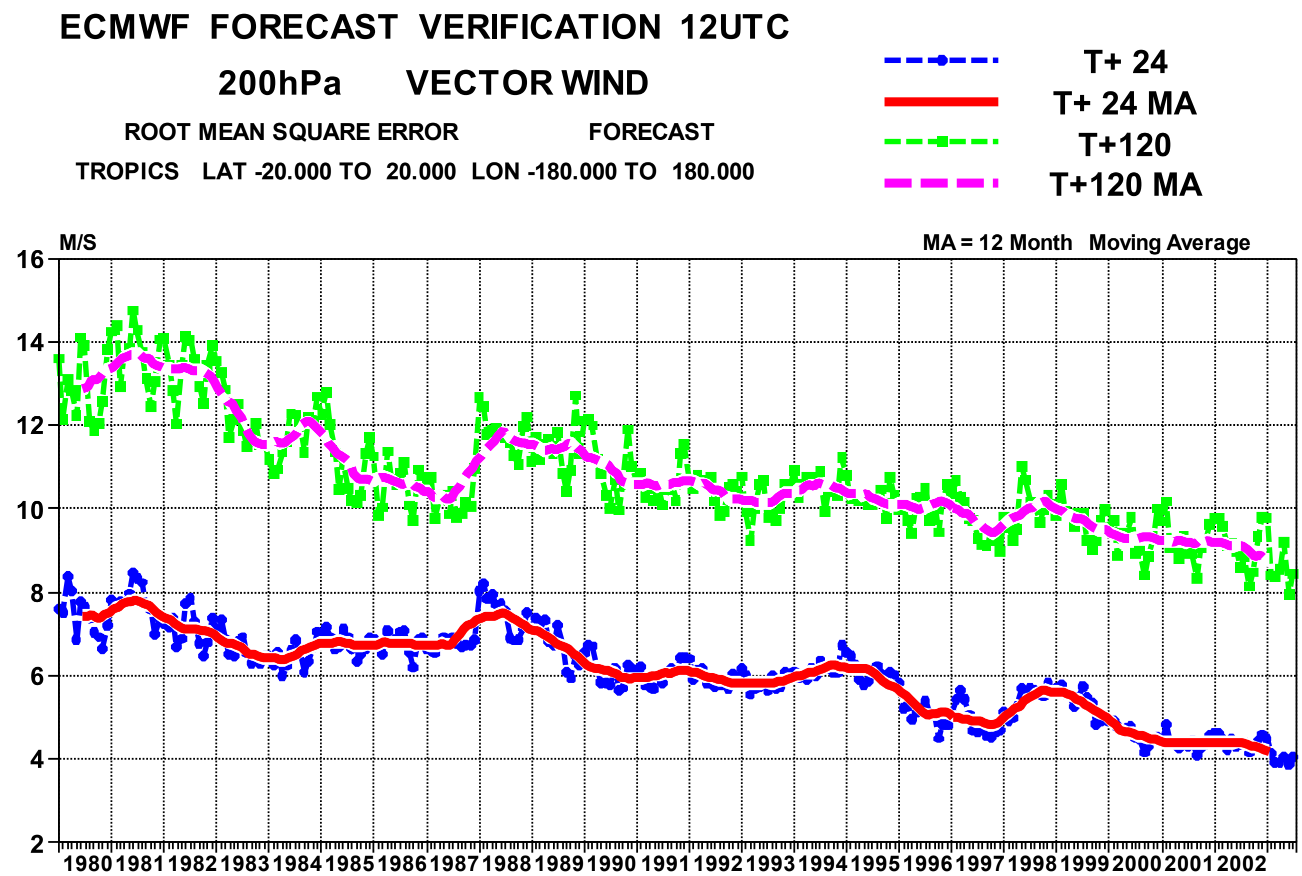


Figure 8: Model scores in the extratropical Northern Hemisphere stratosphere (50hPa height Day 1 and Day 5 forecasts RMSE)



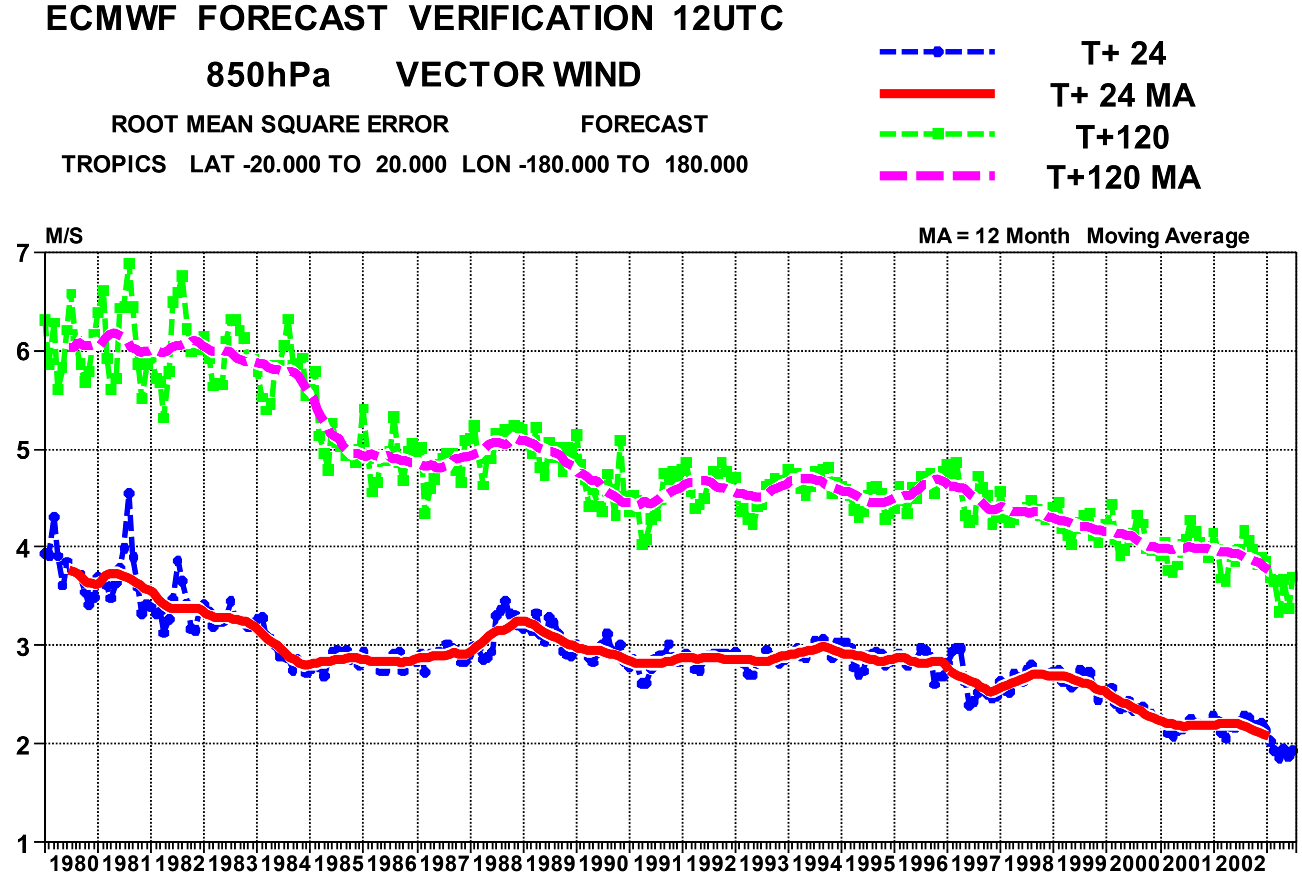
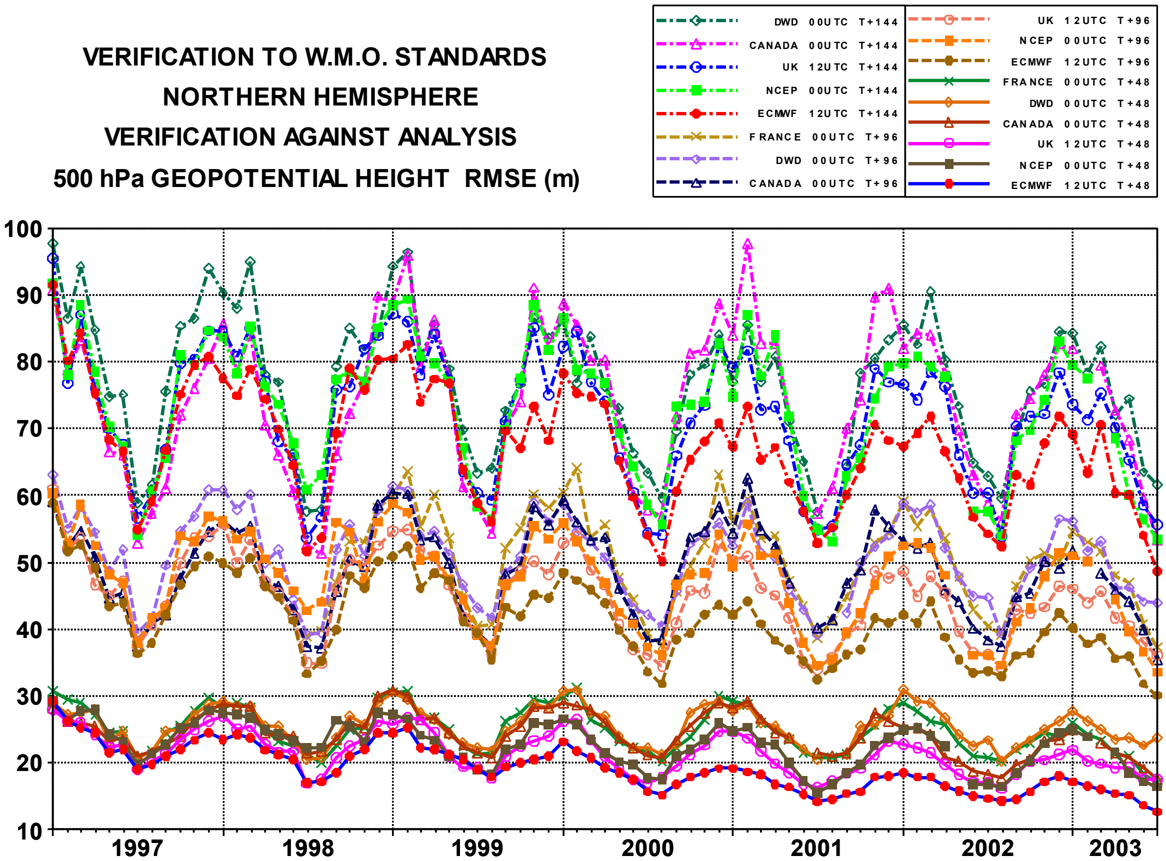


Figure 9: Model scores in the Tropics (root mean square errors for 200hPa and 850hPa wind)



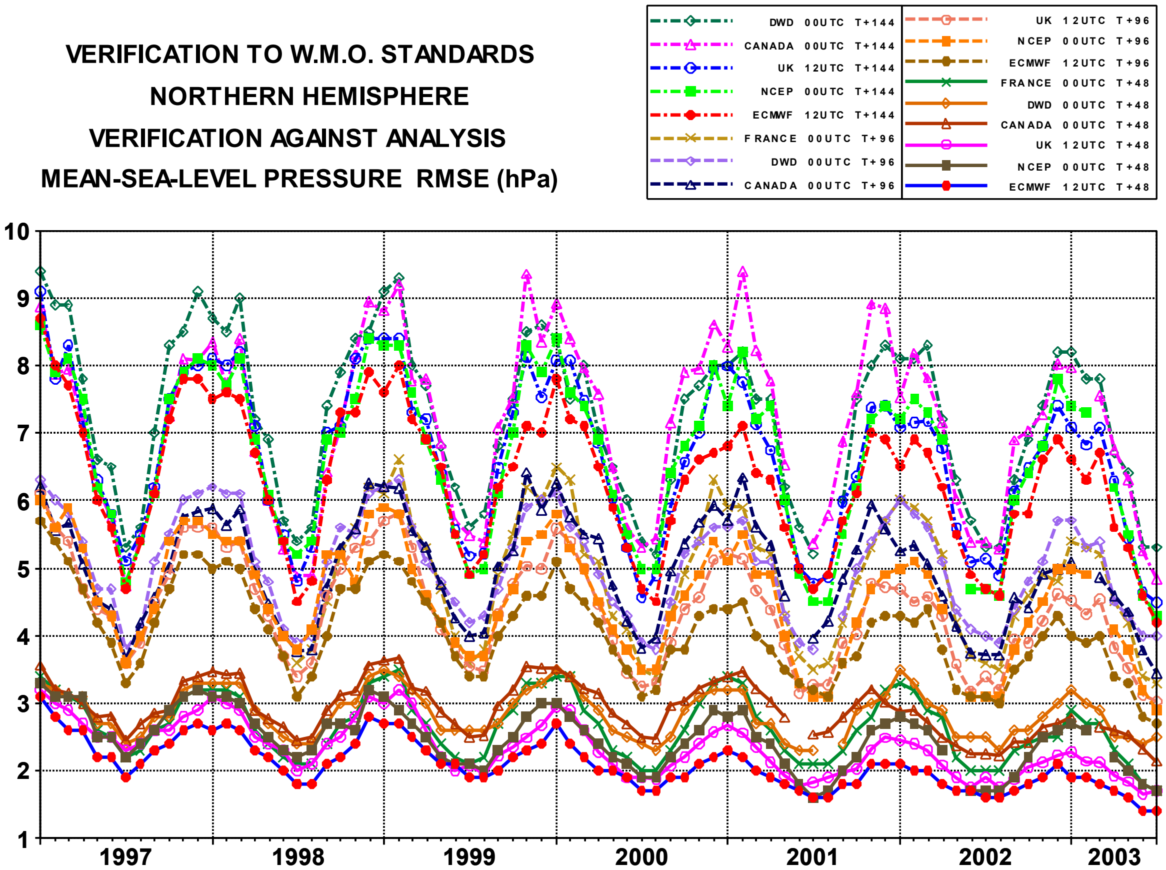
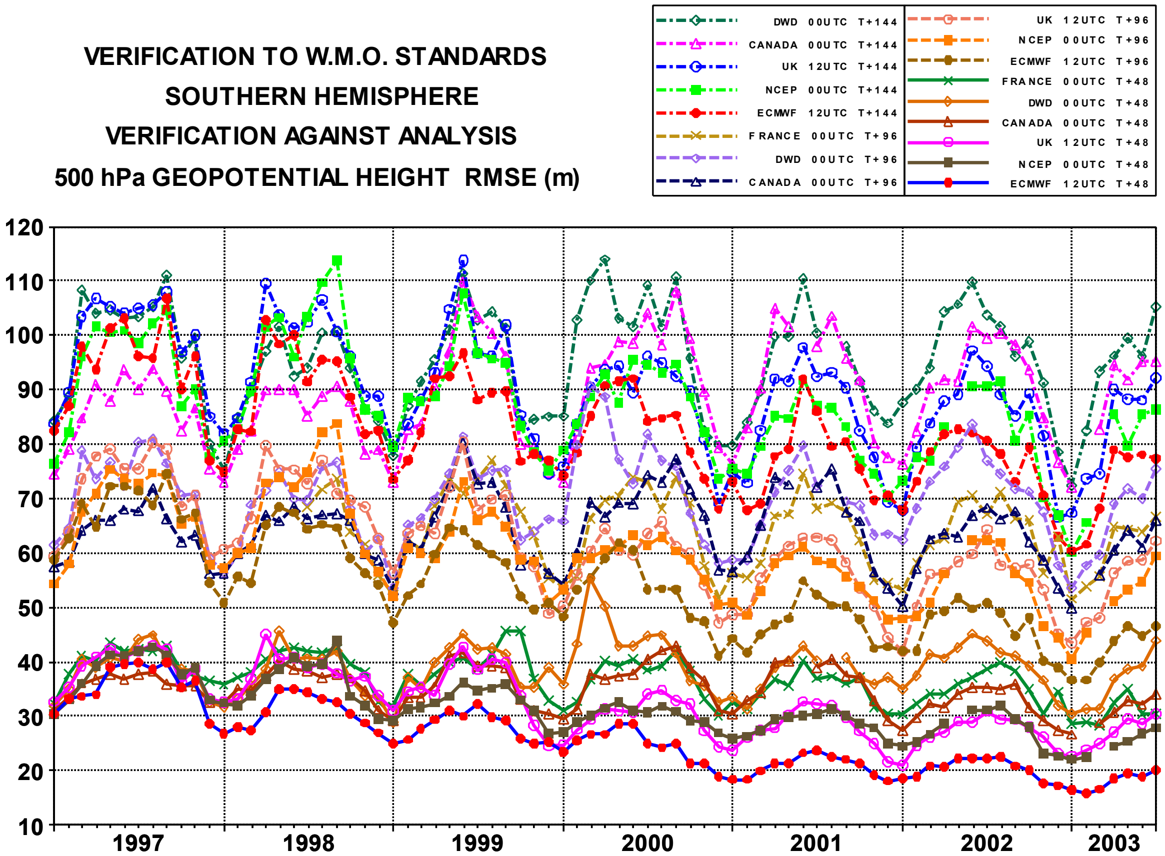


Figure 10: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)



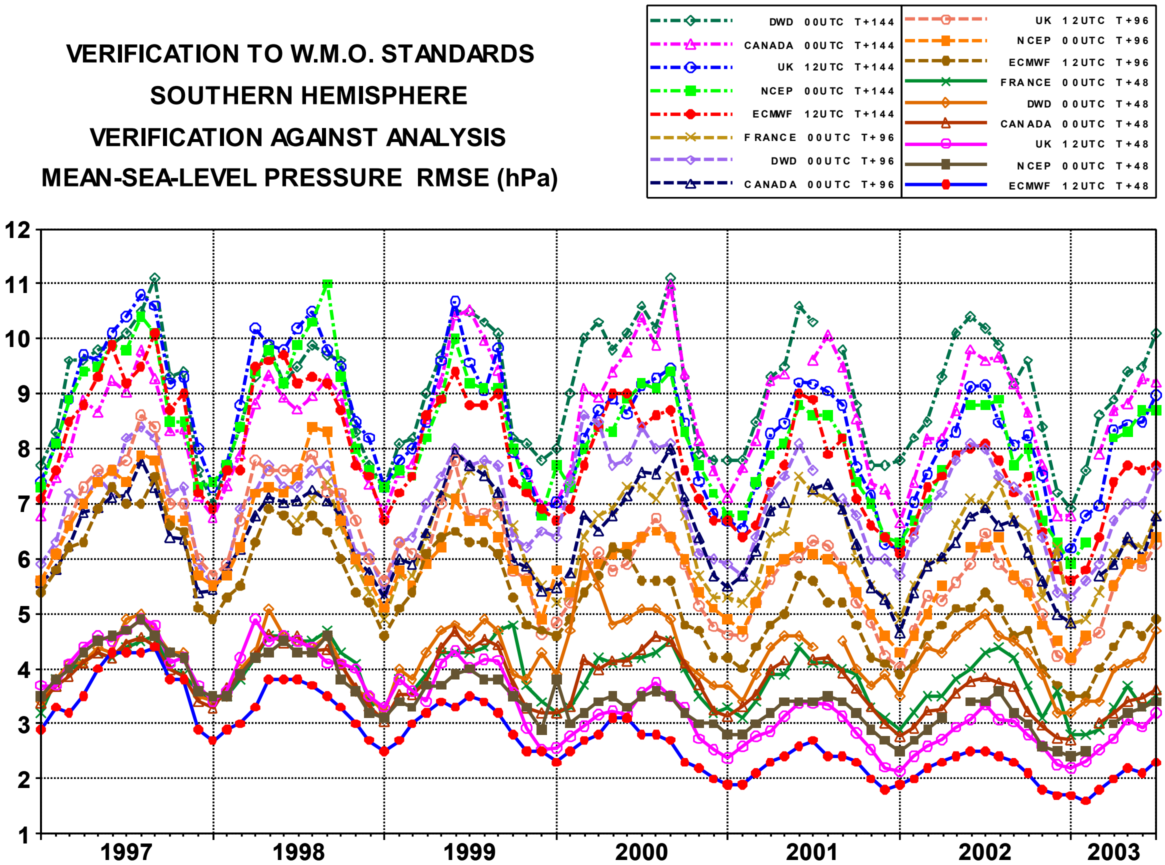


Figure 11: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)

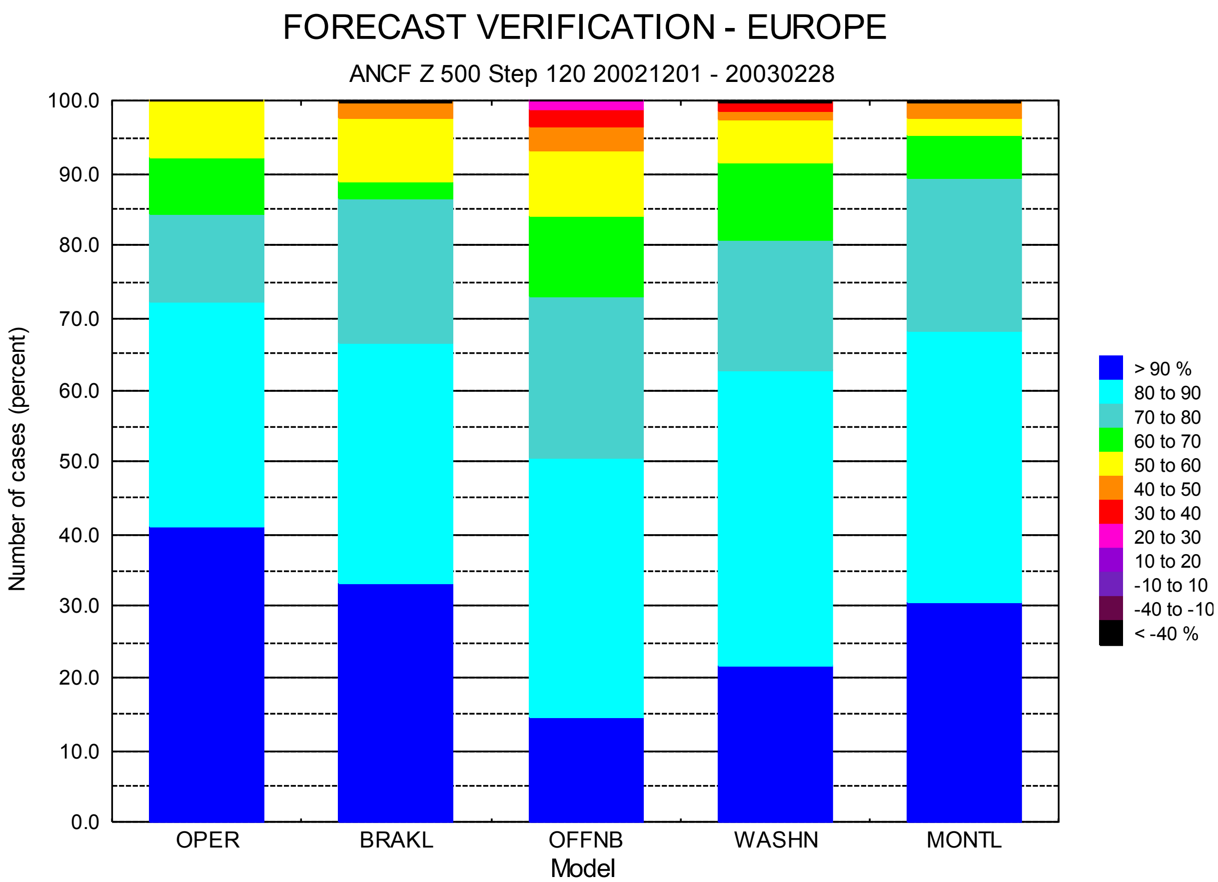
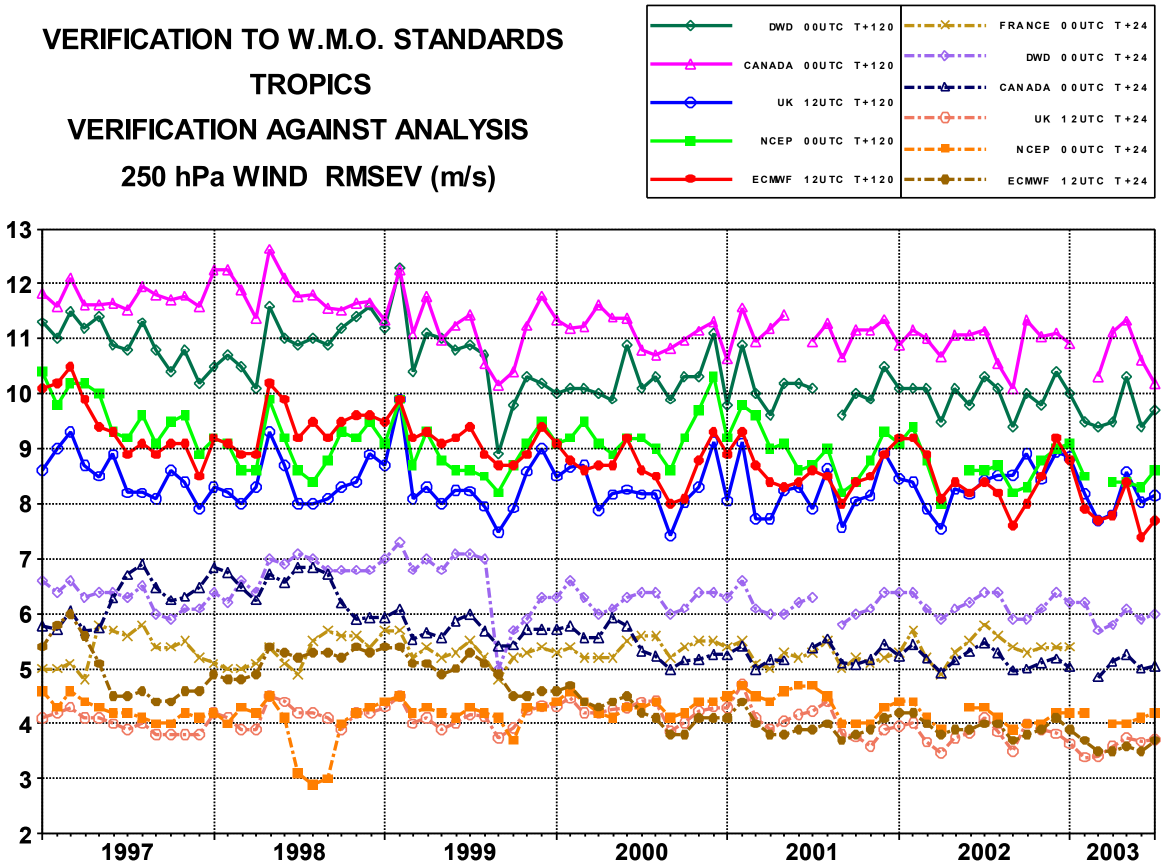


Figure 12: Distribution of anomaly correlation scores (500hPa height, day 5, Europe) for ECMWF (OPER) and four other NWP centres during the last winter season (Dec.-Feb.). Scores have been computed at ECMWF using GTS products.



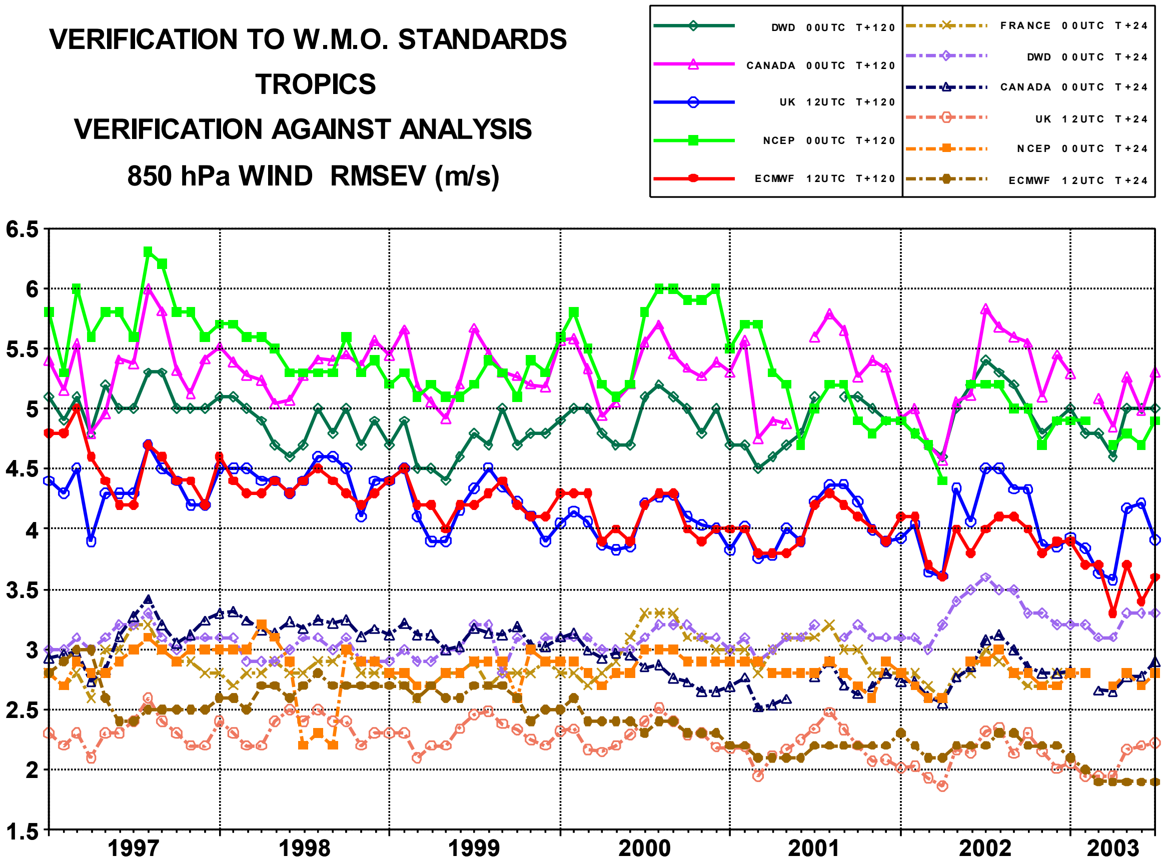
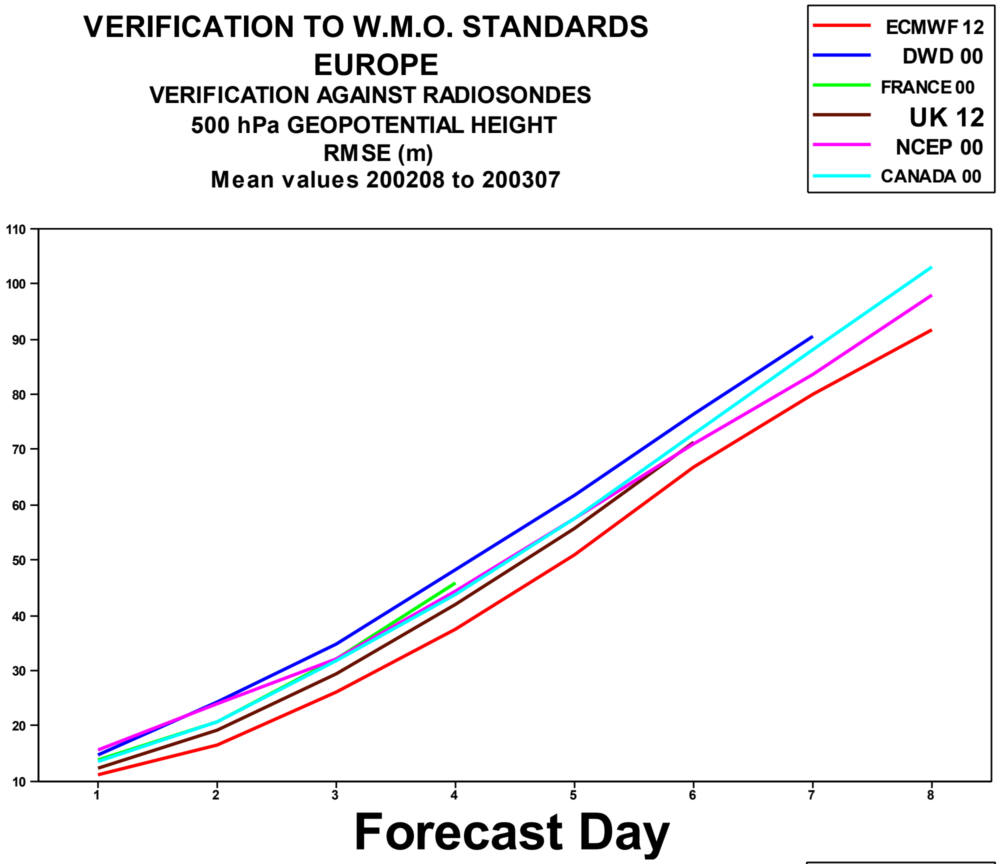


Figure 13: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250hPa and 850hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis



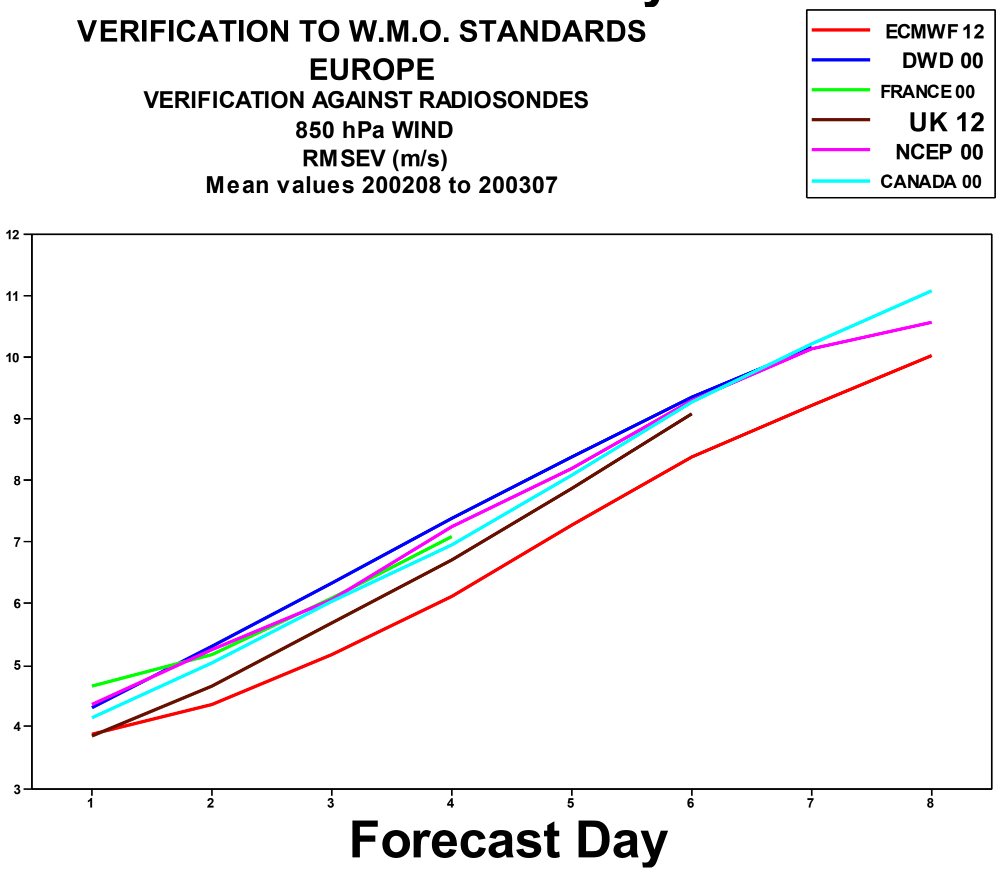
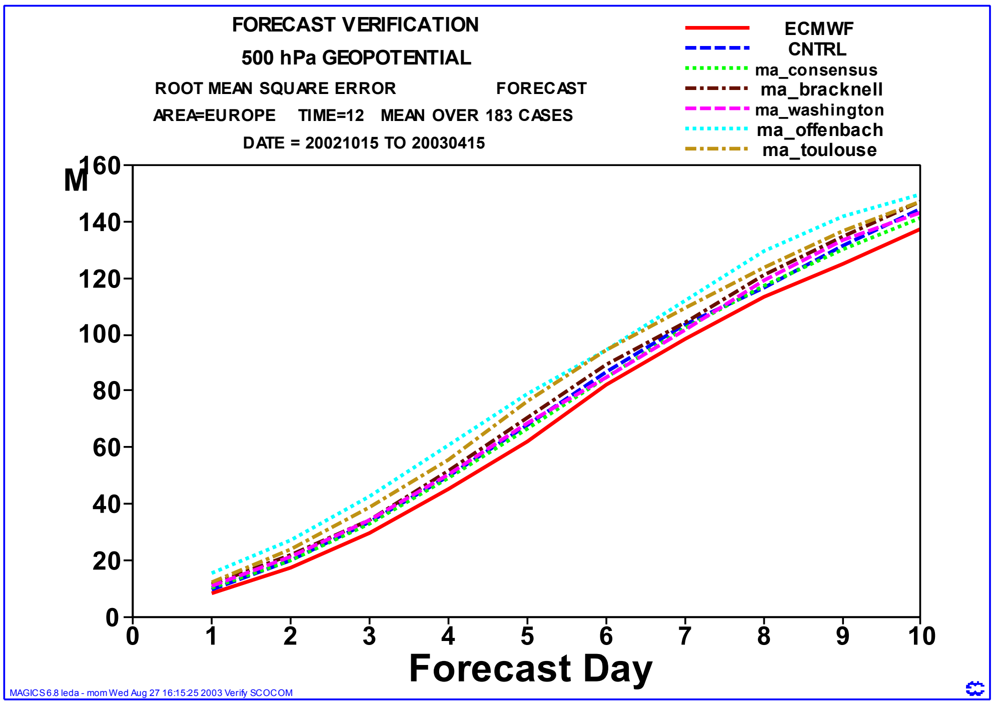


Figure 14: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)



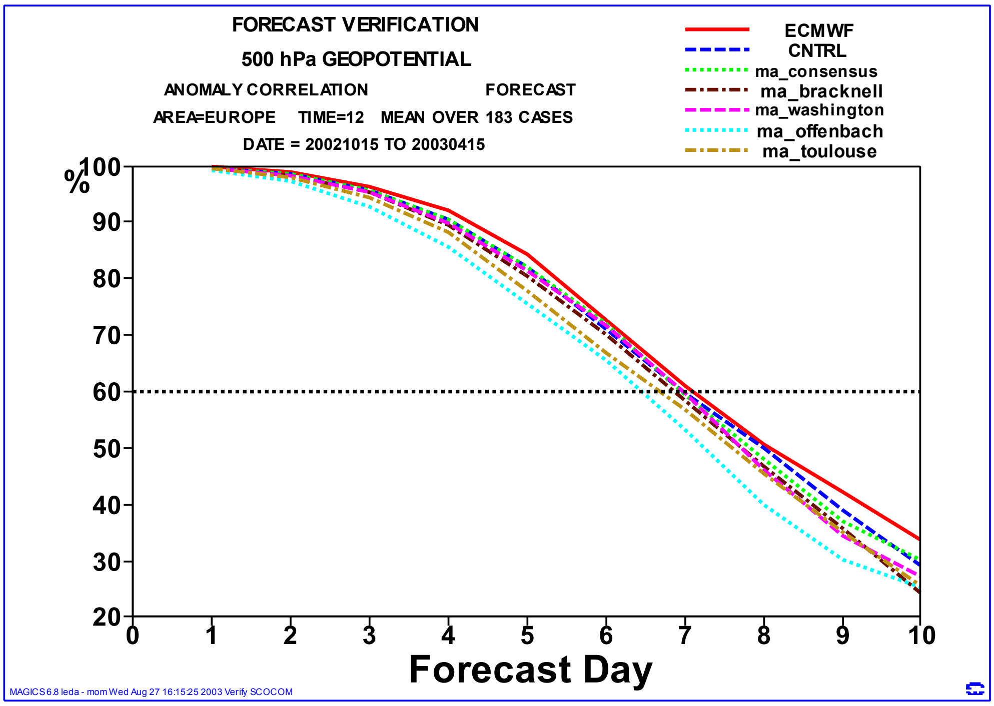
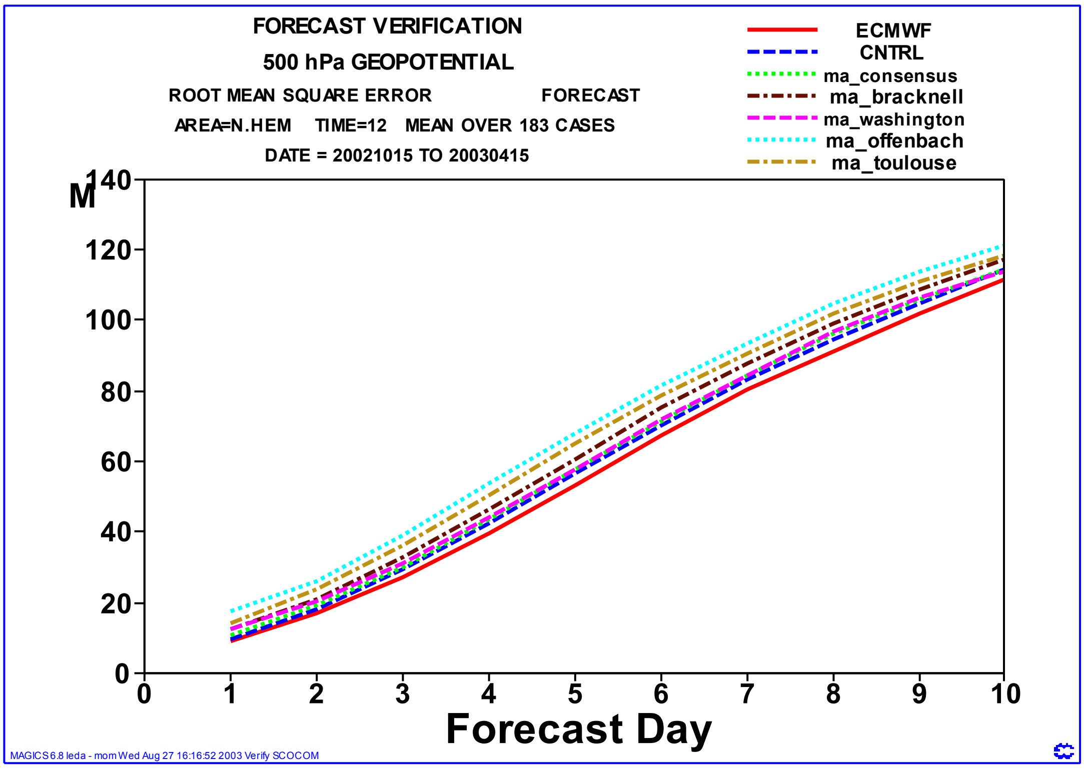


Figure 15: 500hPa height scores from the multi-analysis system (Europe, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution



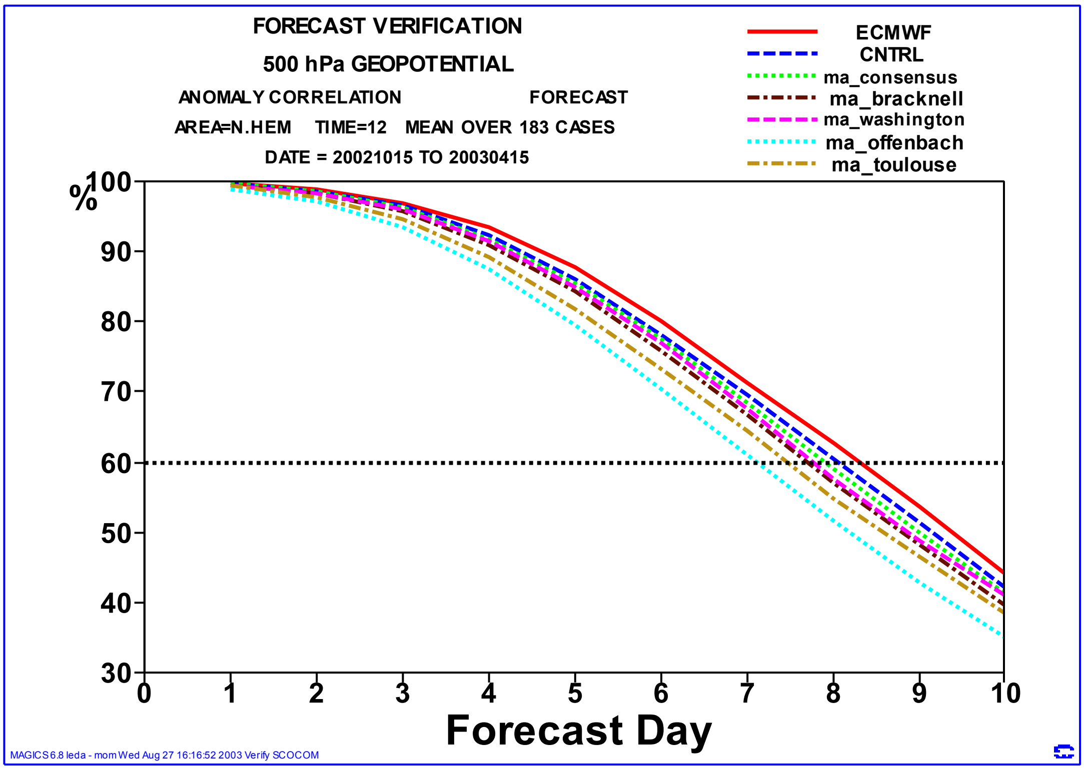
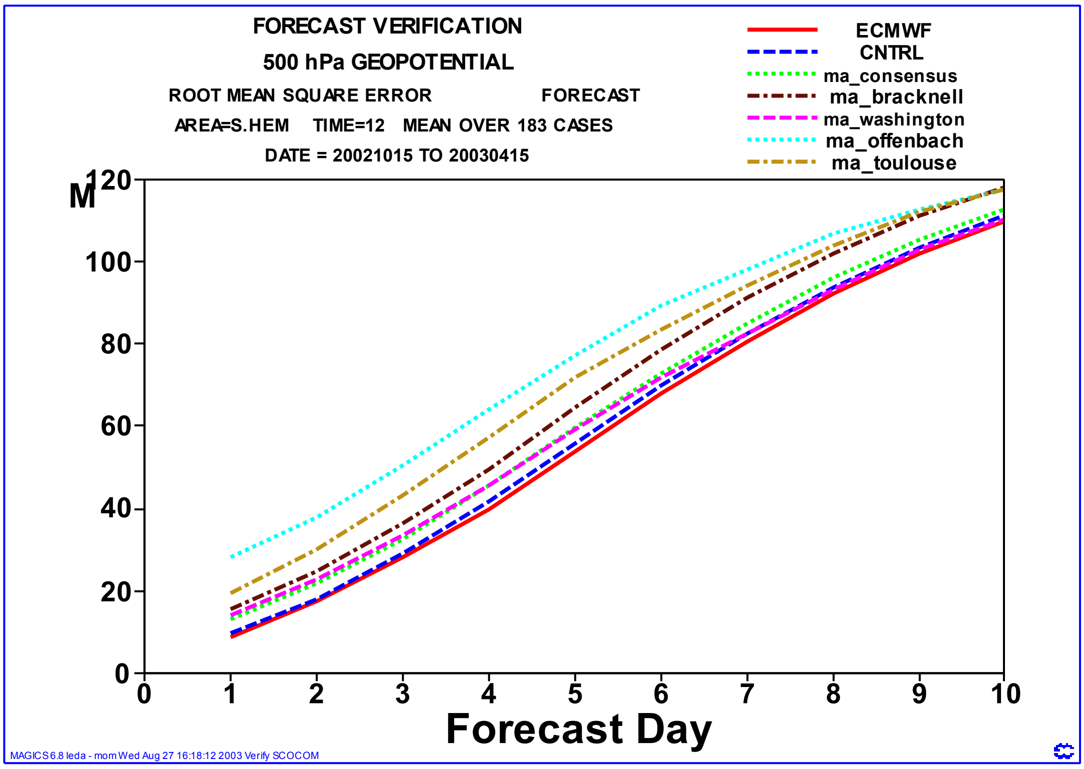


Figure 16: 500hPa height scores from the multi-analysis system (N. Extratropics, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution



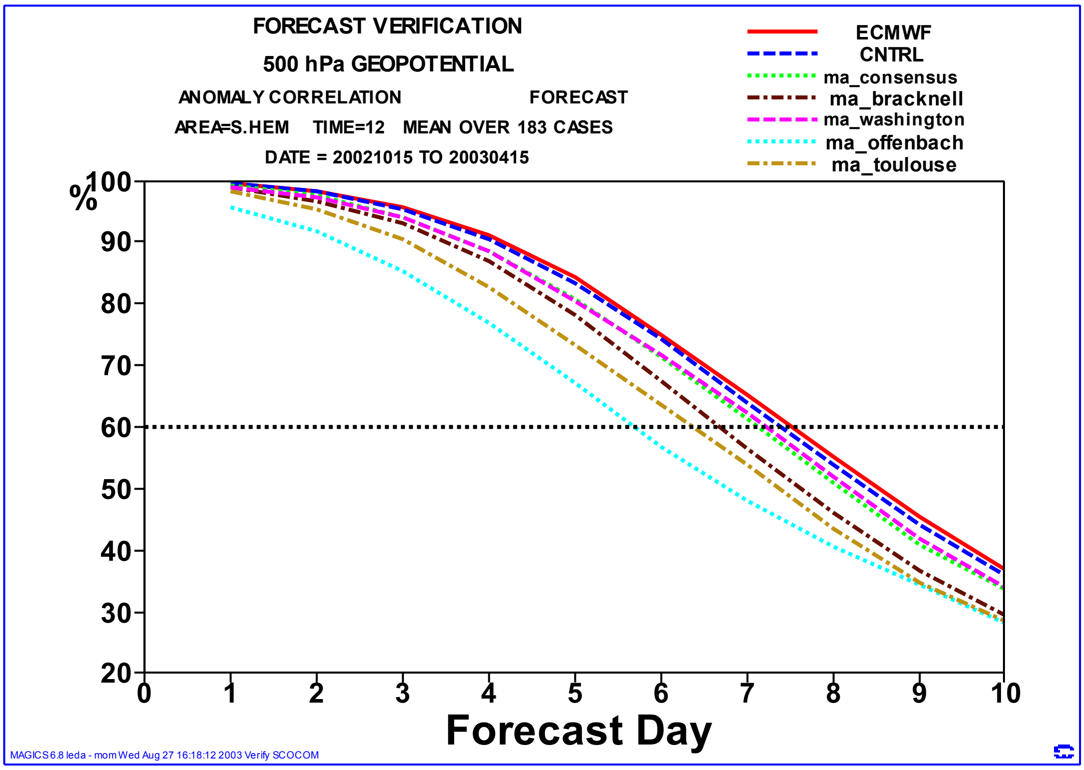
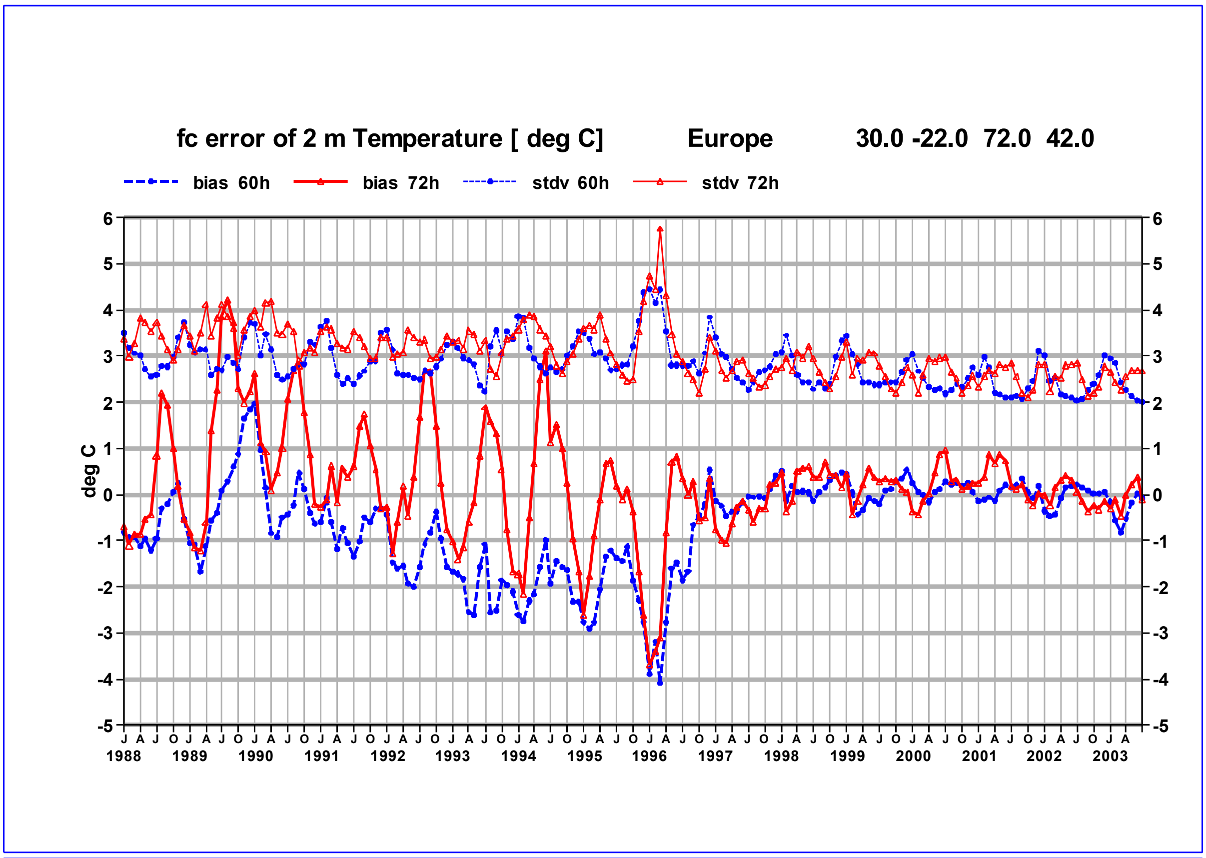


Figure 17: 500hPa height scores from the multi-analysis system (Southern Extratropics, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation ; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution



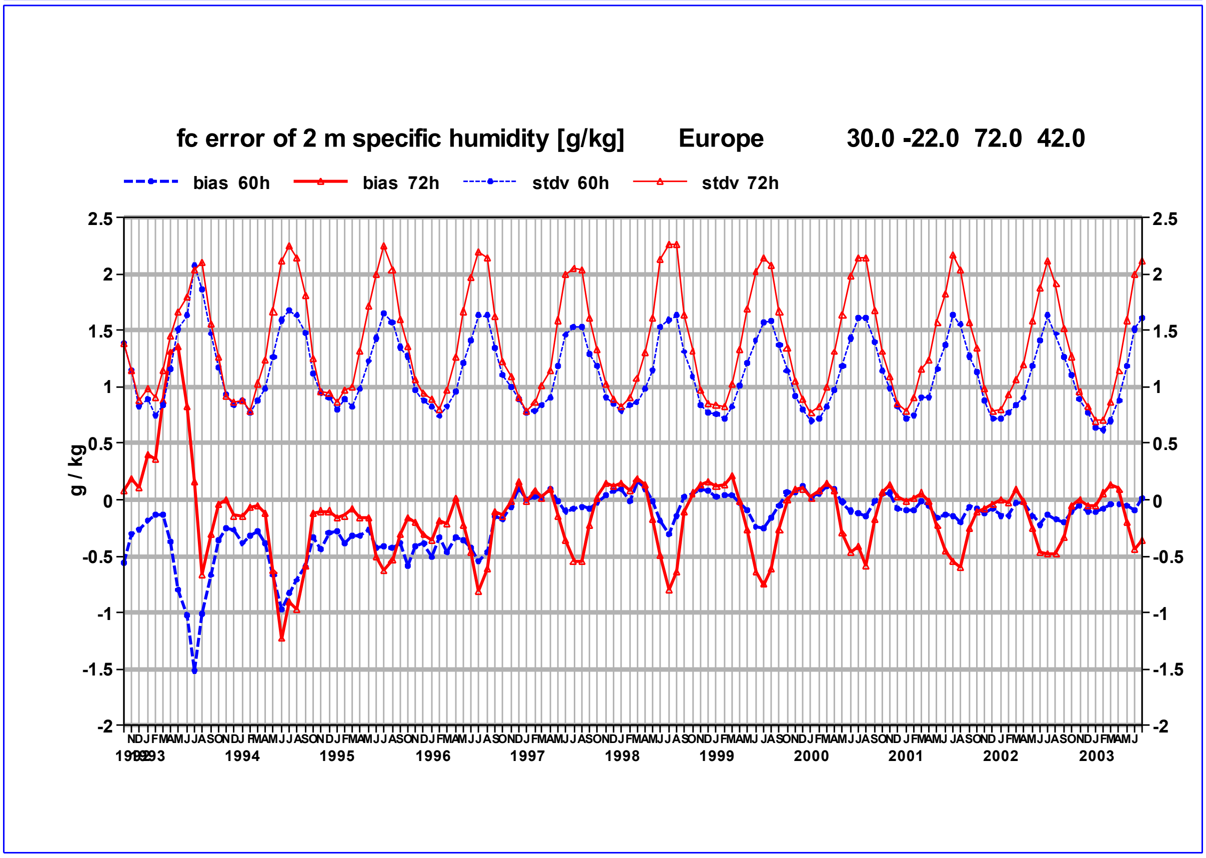
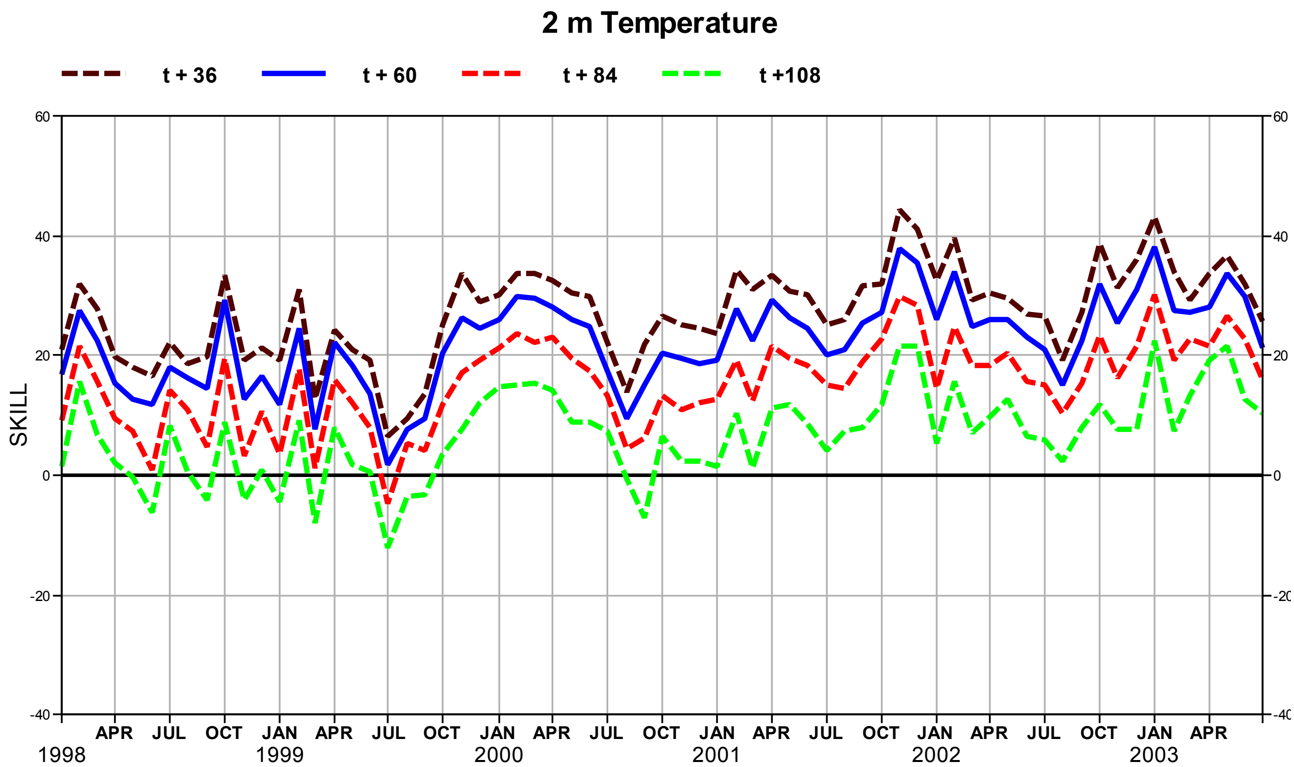


Figure 18: Verification against European SYNOP observations of 2m Temperature and specific humidity (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)



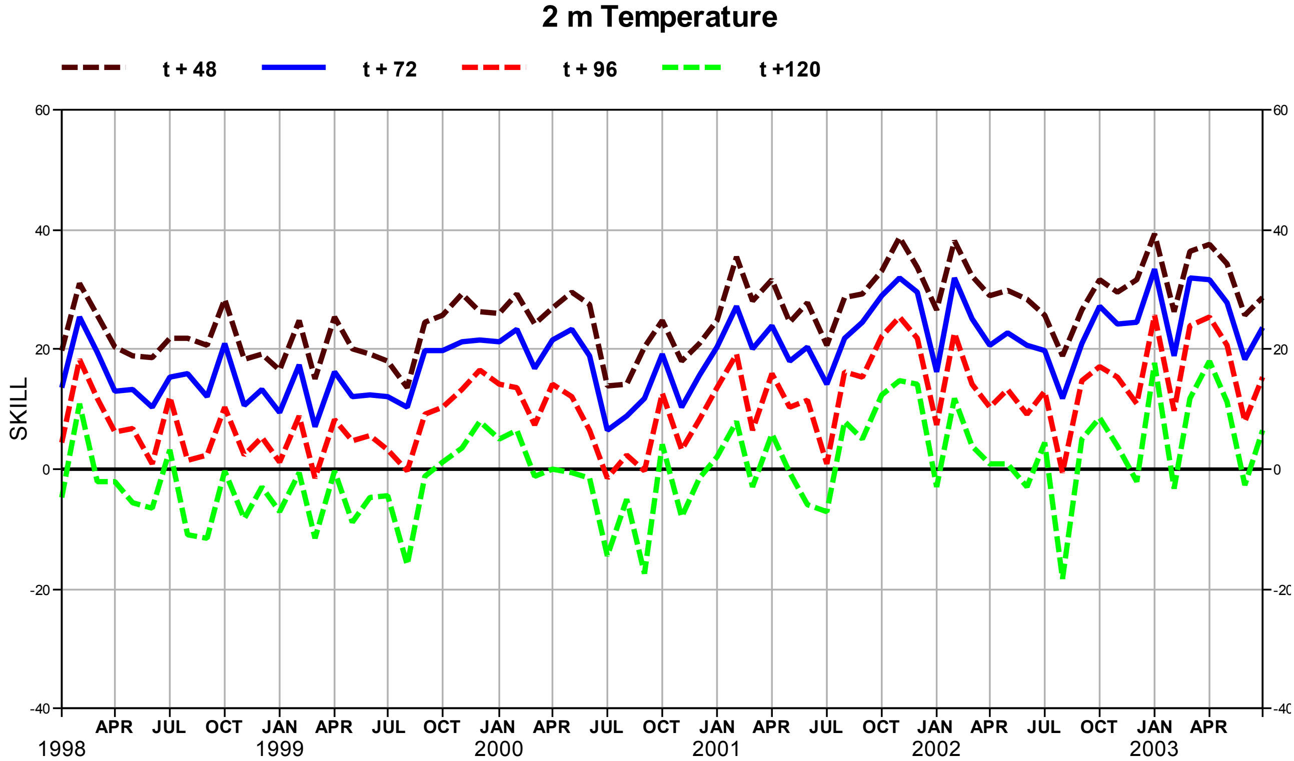
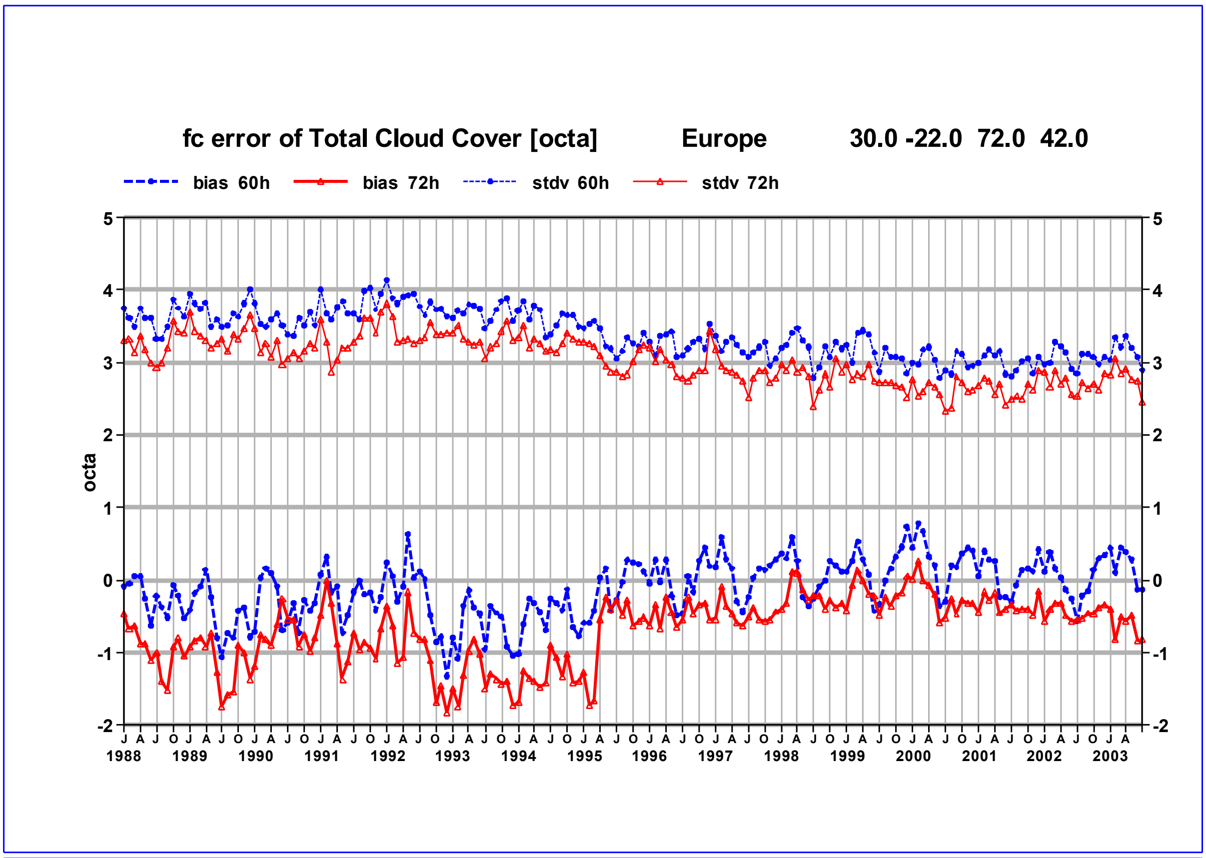


Figure 19: Error variance reduction with respect to persistence for 2m temperature forecasts over Europe during night time (00UTC, top) and daytime (12UTC, bottom) for different forecast ranges.



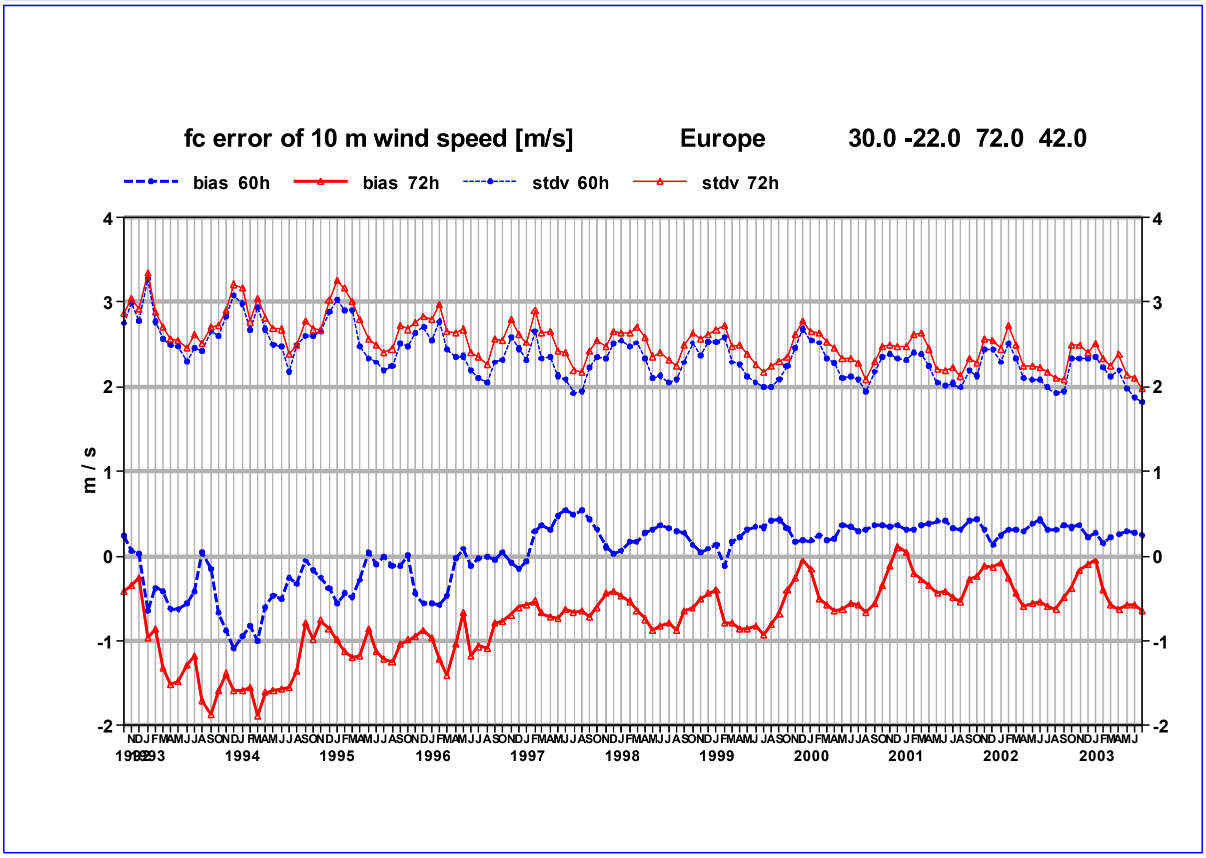
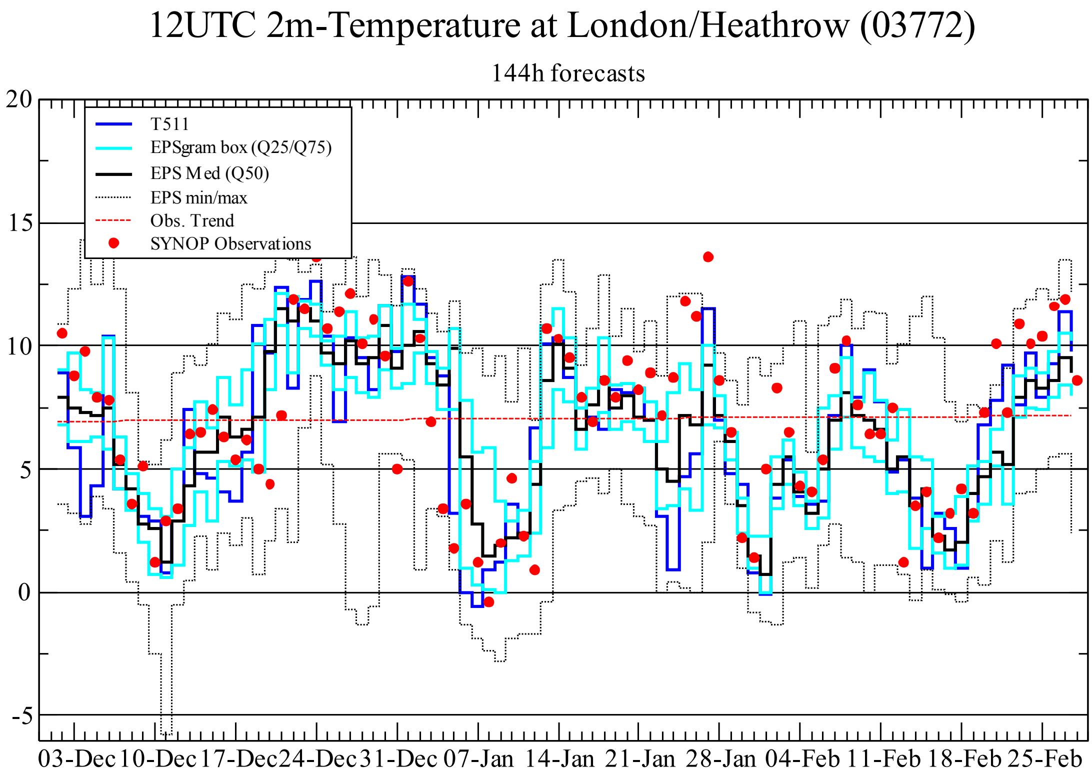


Figure 20: Scores against European SYNOPs of total cloud cover and 10m wind speed forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC).



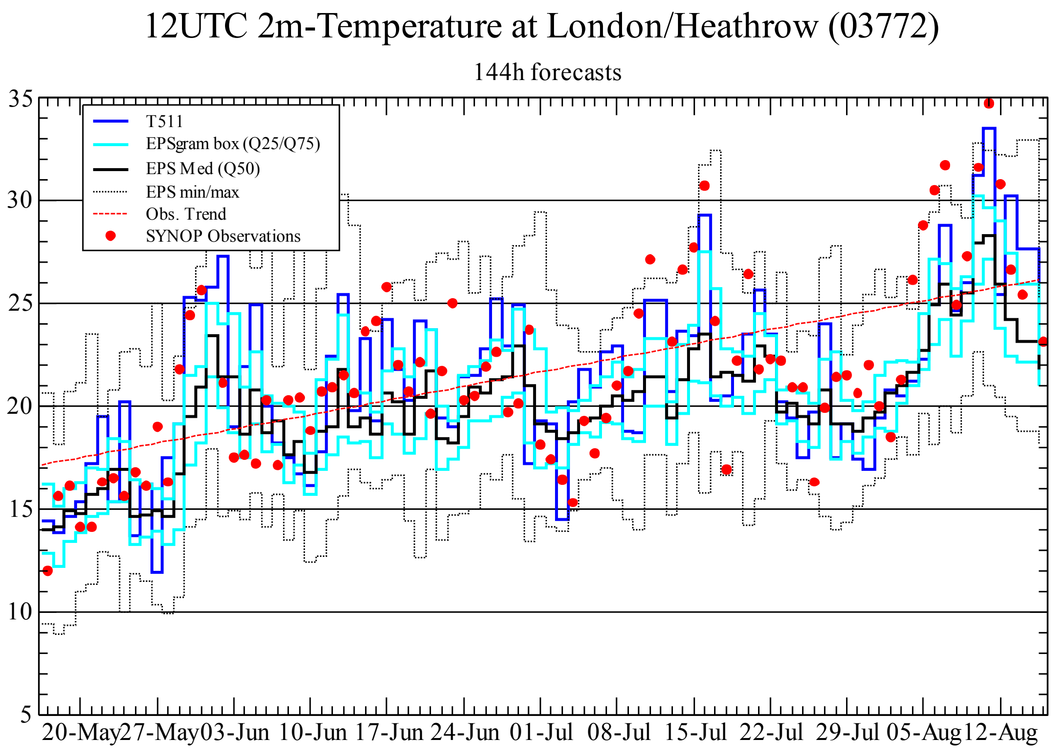


Figure 21: Time series of Day 6 forecasts of 2m temperatures at London (03772) during last winter (upper panel) and summer (lower panel). Verifying observations are also shown

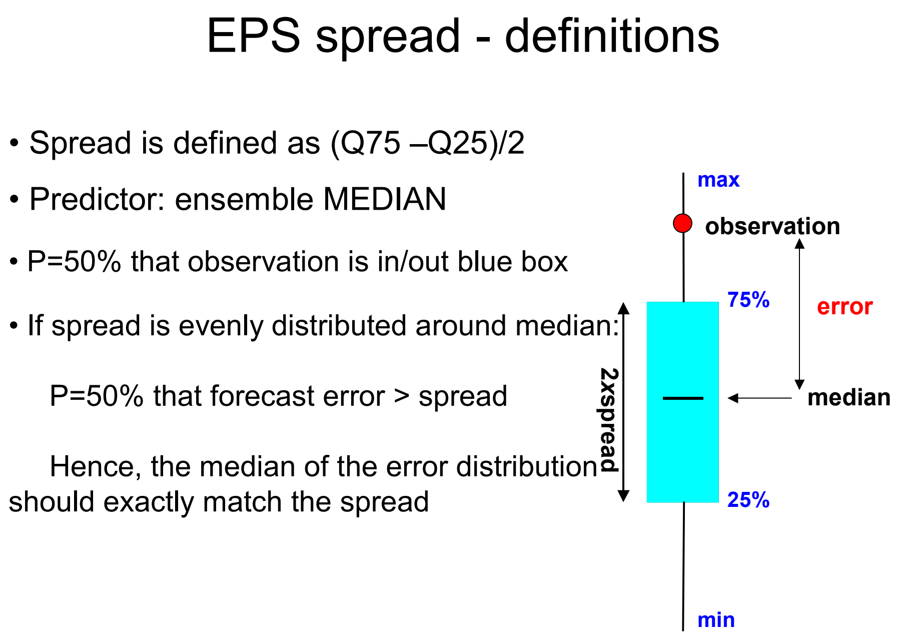


Figure 22: Schematic description of the spread skill relation that should be found in a perfect probabilistic forecast.

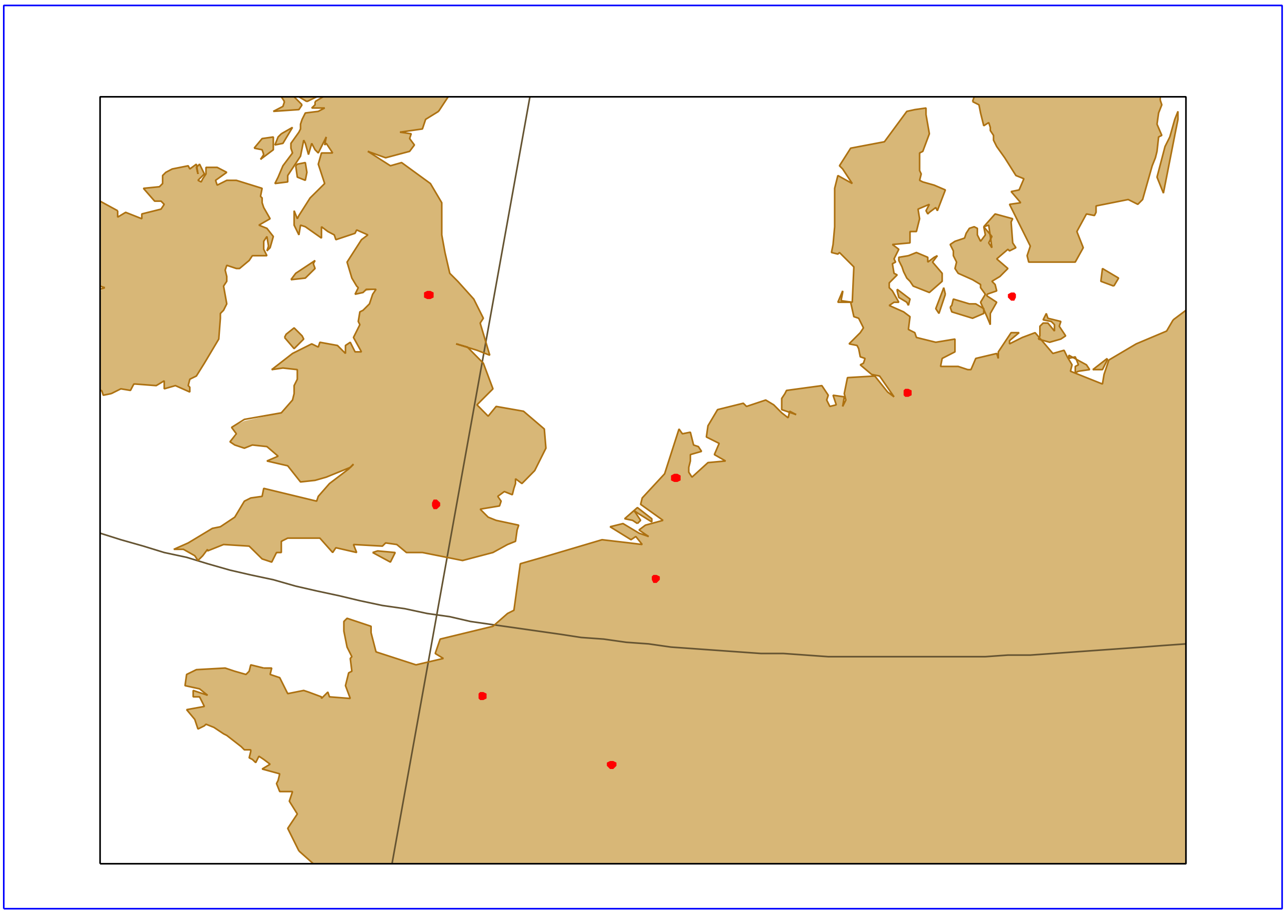


Figure 23: Stations used for the verification of the EPS spread (Figure 25 and Figure 26)

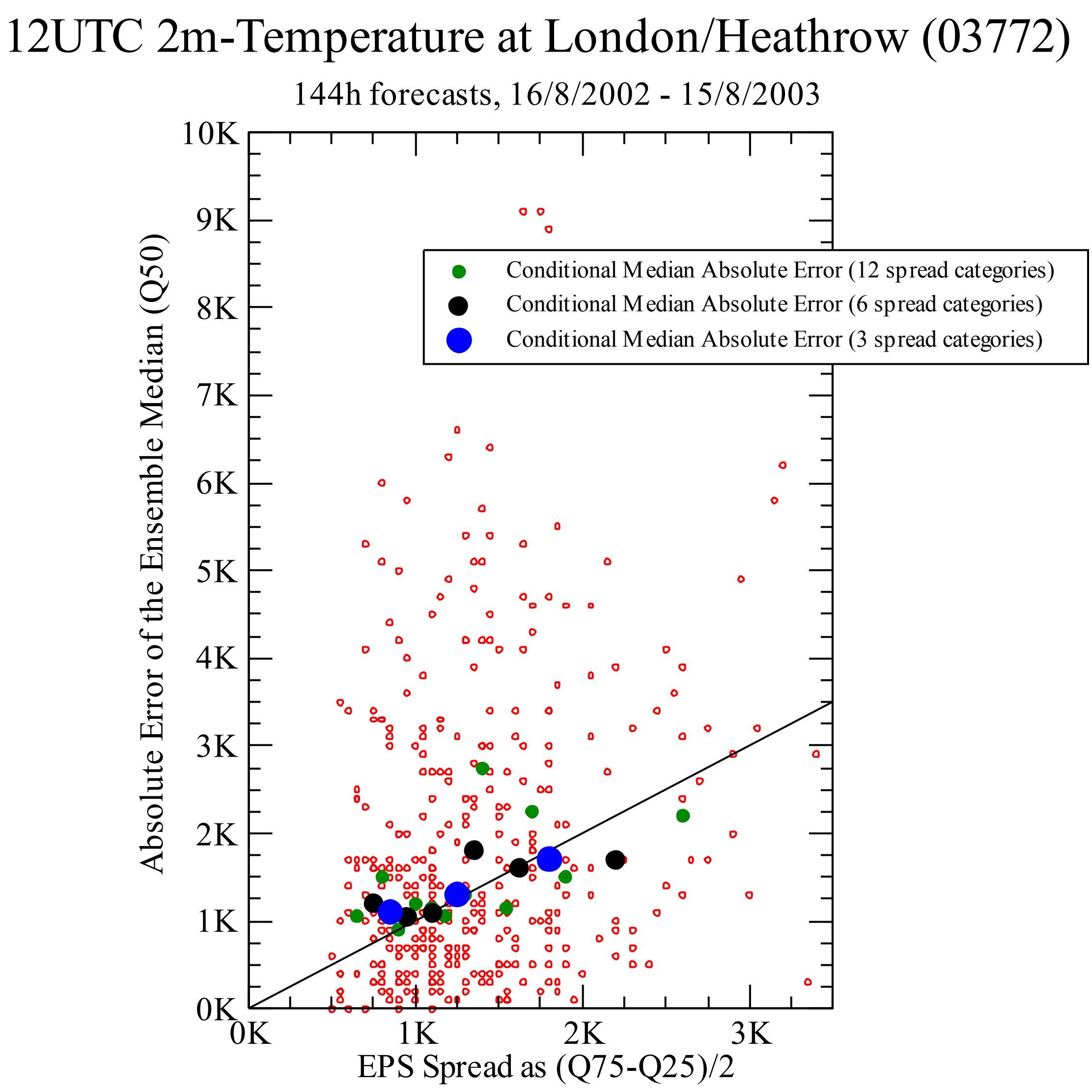


Figure 24: Scatter plot of absolute errors of the median of the D6 ensemble forecast of 2m temperature as a function of the EPS spread (interquartile half value) at London/Heathrow. The median of the distribution of these errors for different categories of spread are also show in large circles (see caption)

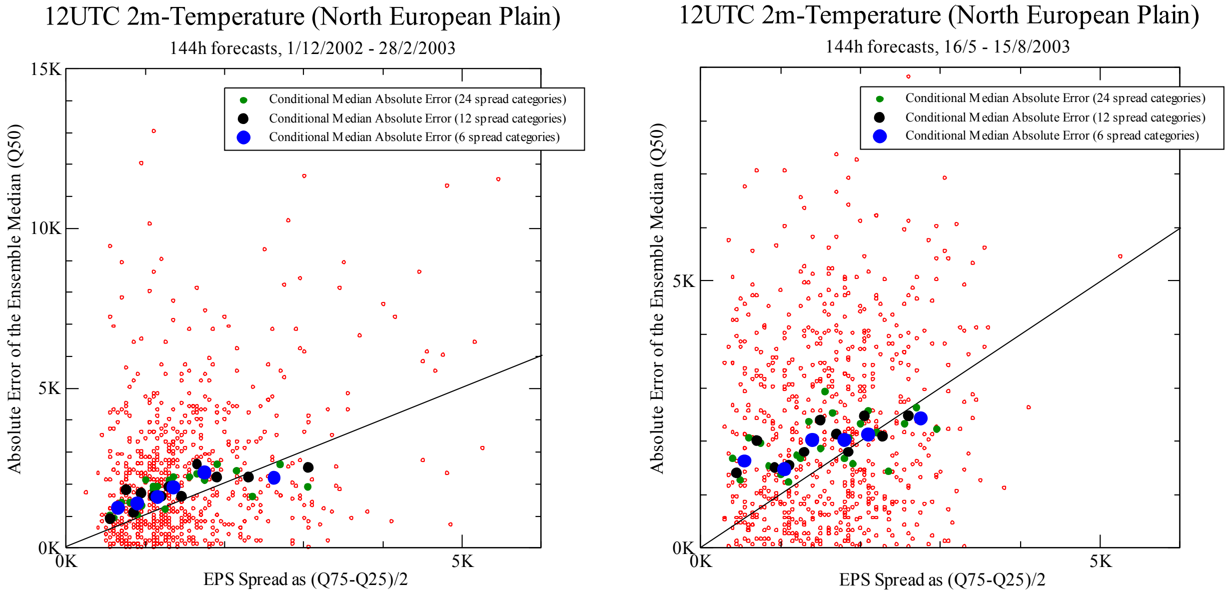


Figure 25: Scatter plot of absolute errors of the median as a function of the EPS spread (interquartile half value) for Day 6 2m temperature forecasts at a selection of stations in the North European plain (left: winter 2002-03; right: summer 2003).

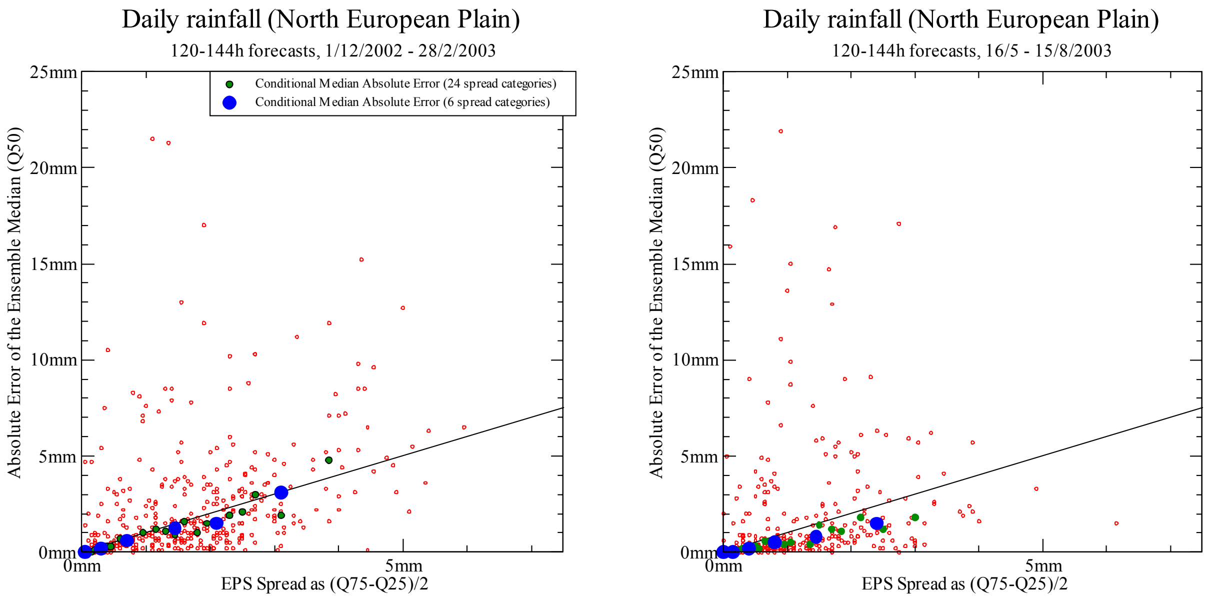


Figure 26: Scatter plot of absolute errors of the median as a function of the EPS spread (interquartile half value) for 120-144h forecasts of rain accumulation at a selection of stations in the North European plain (left: winter 2002-03; right: summer 2003).

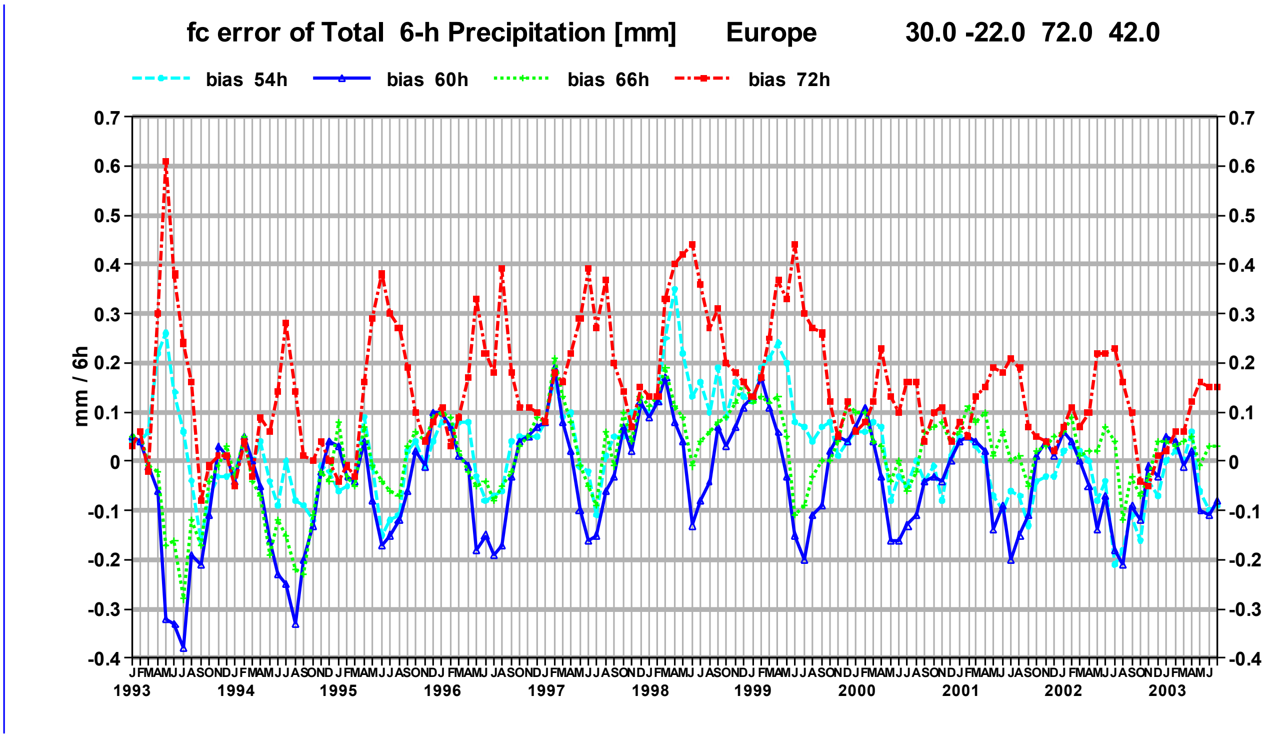


Figure 27: 6h-accumulated precipitation forecasts biases (T+54/60/66/72h) with respect to SYNOP

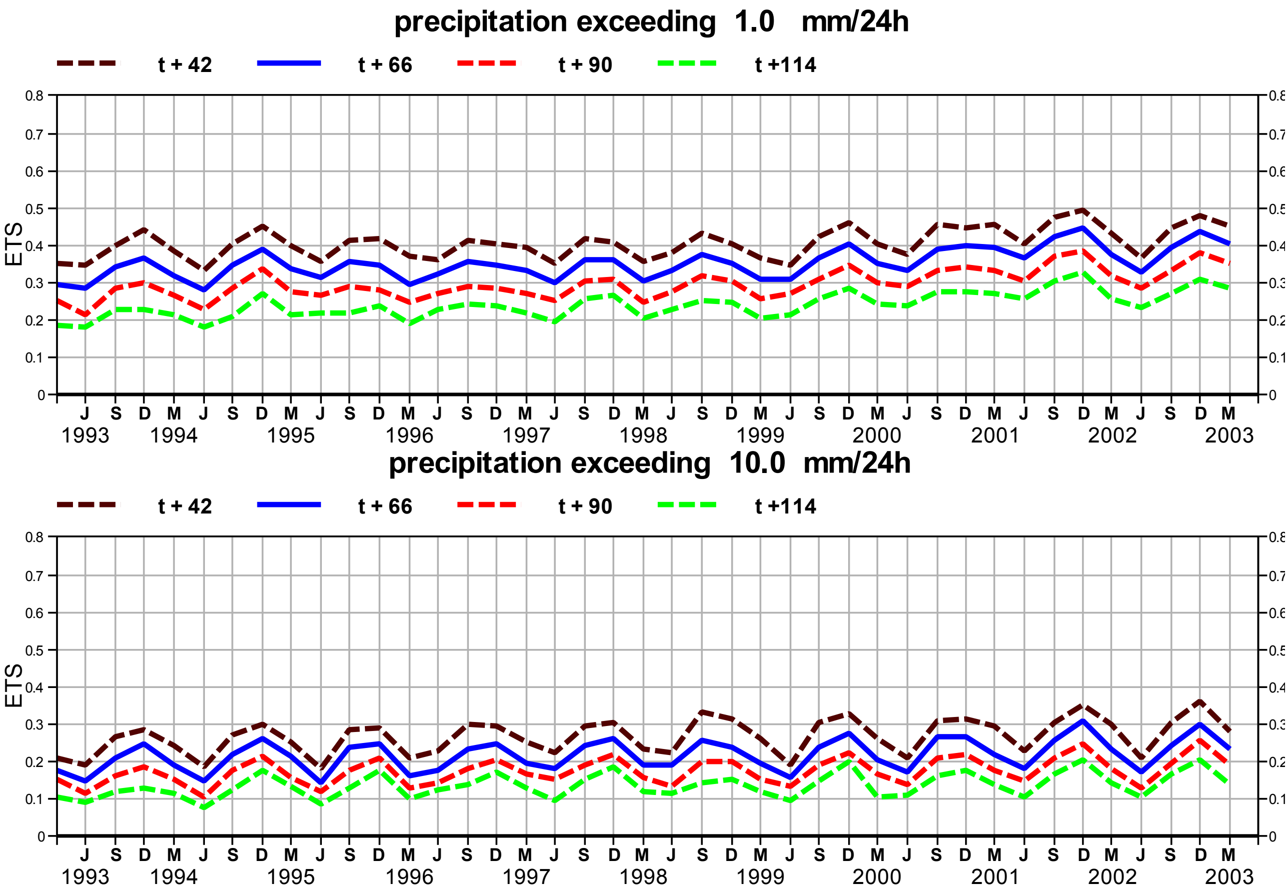


Figure 28: Time series of Equitable Threat Scores for the forecast of daily precipitation verified using SYNOP reports over Europe; Top: threshold 1mm, bottom: 10mm.

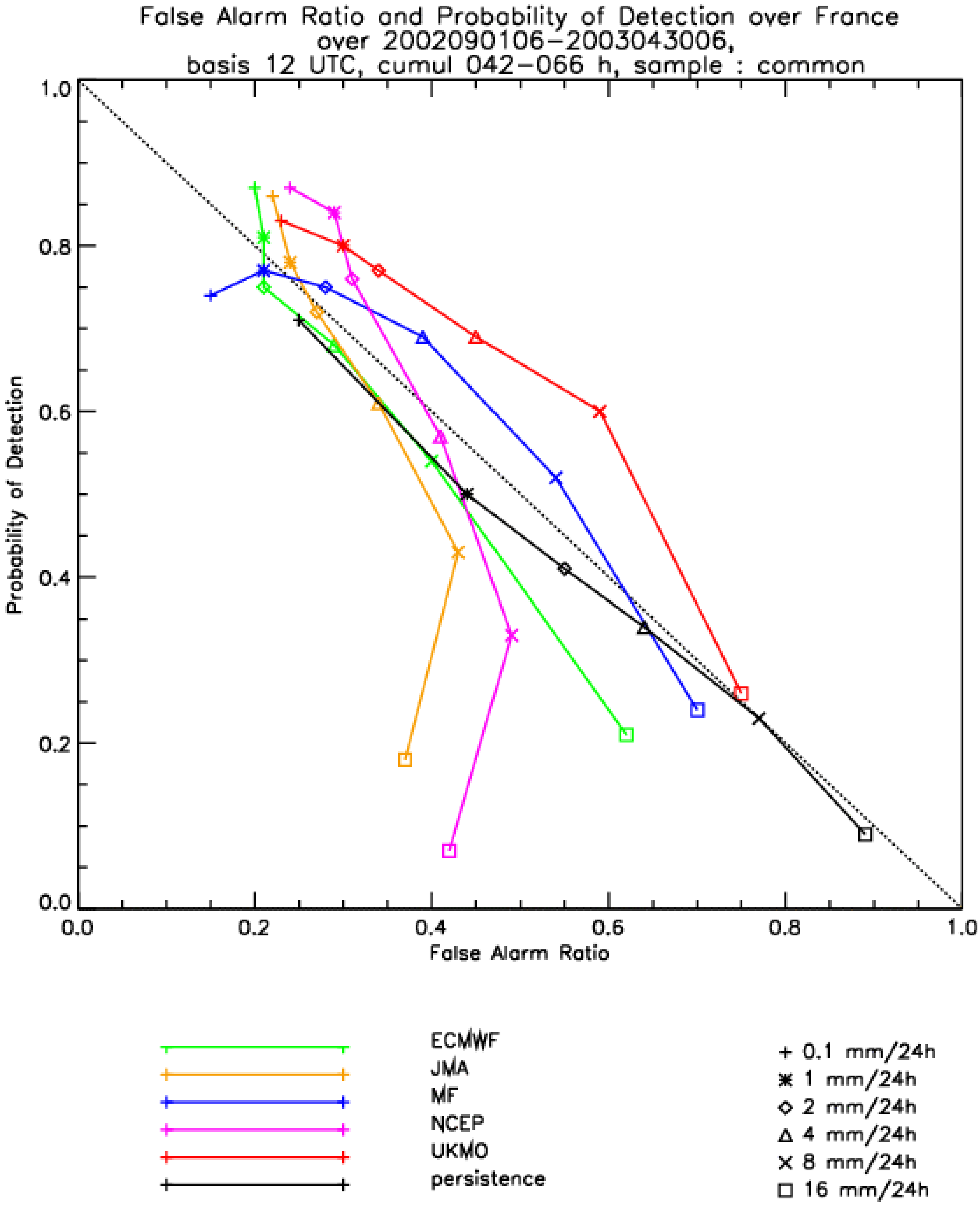


Figure 29: Intercomparison of precipitation forecasts against upscaled observations over France (courtesy from Meteo-France)

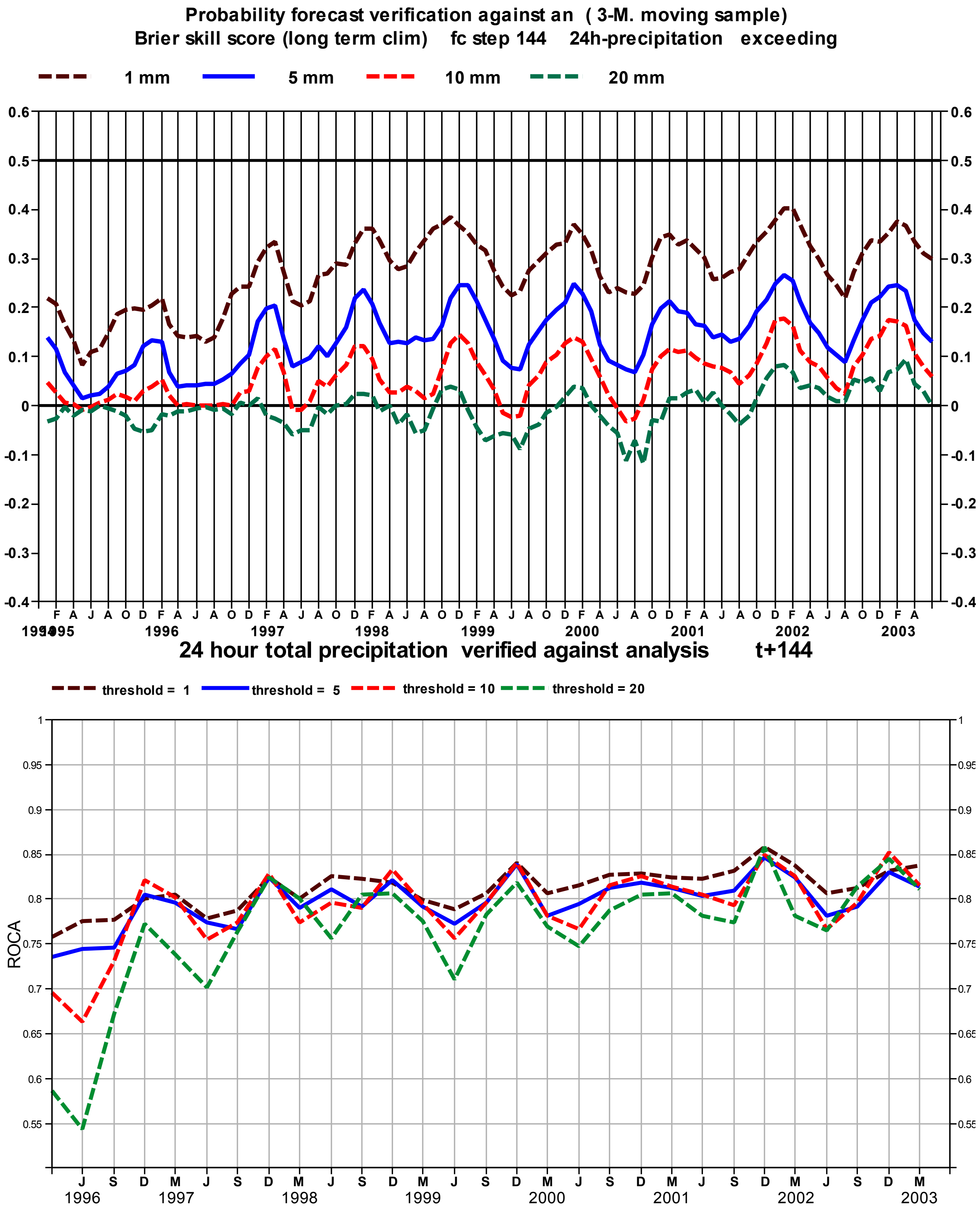


Figure 30: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics curve Area (ROCA, lower panel), the later showing the skill shown by the EPS at detecting a signal out of the 120-144h probability forecast for rain (0.5 is no skill, 1 is a perfect detection).

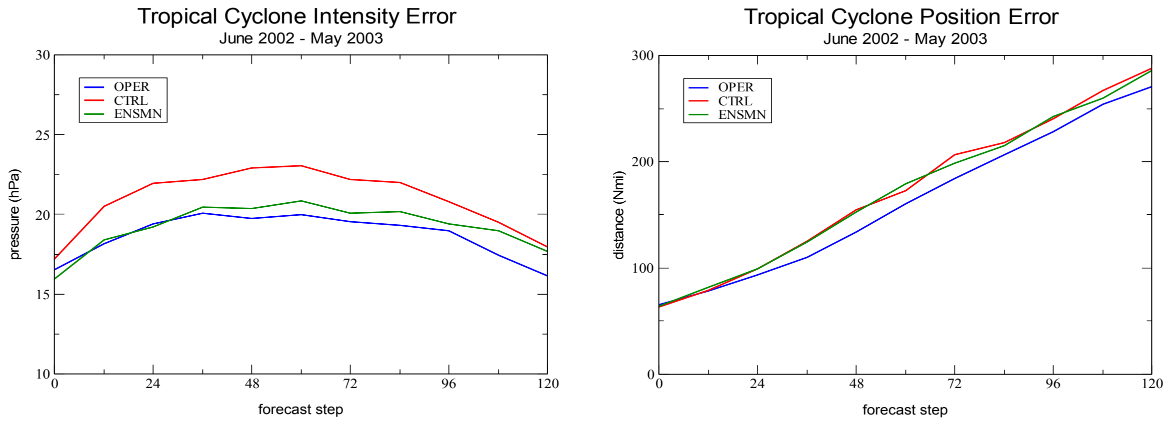


Figure 31: Verification of Tropical Cyclone forecasts from the deterministic, T511 forecast (blue), EPS T255 Control (red) and mean position/ intensity averaged among all cyclones tracked in each member of the ensemble forecast

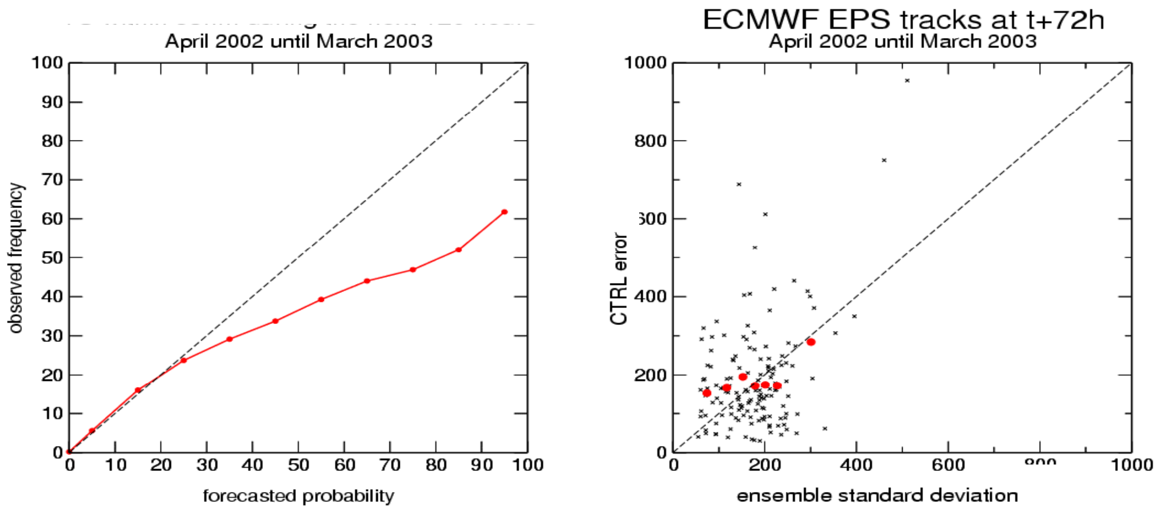


Figure 32: Left: Verification of EPS forecasts of Tropical Cyclones (probability of getting closer than 120km from a tropical cyclone within the next 120h, or "strike probability"); right: scatter plot (axis labelled in 1/60th of degrees)

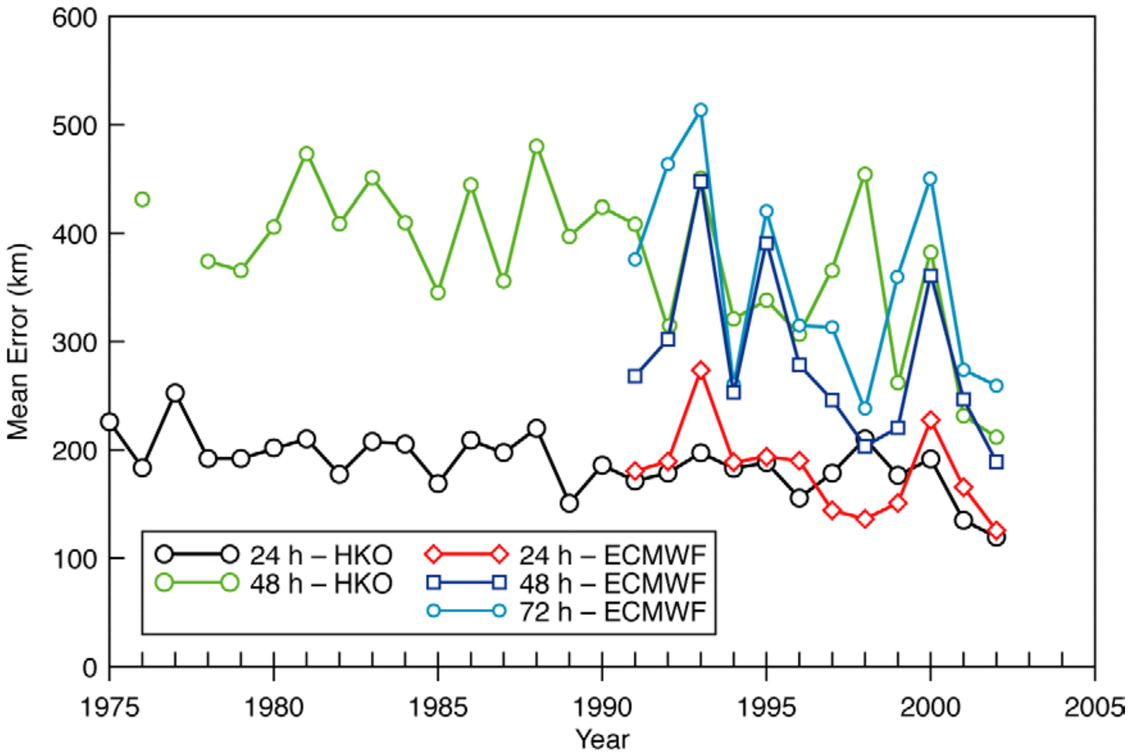


Figure 33: Mean errors for ECMWF forecasts and the HKO subjective forecasts for TCs over the verification area 10-30N, 105-125E (Courtesy from Hong-Kong Observatory)

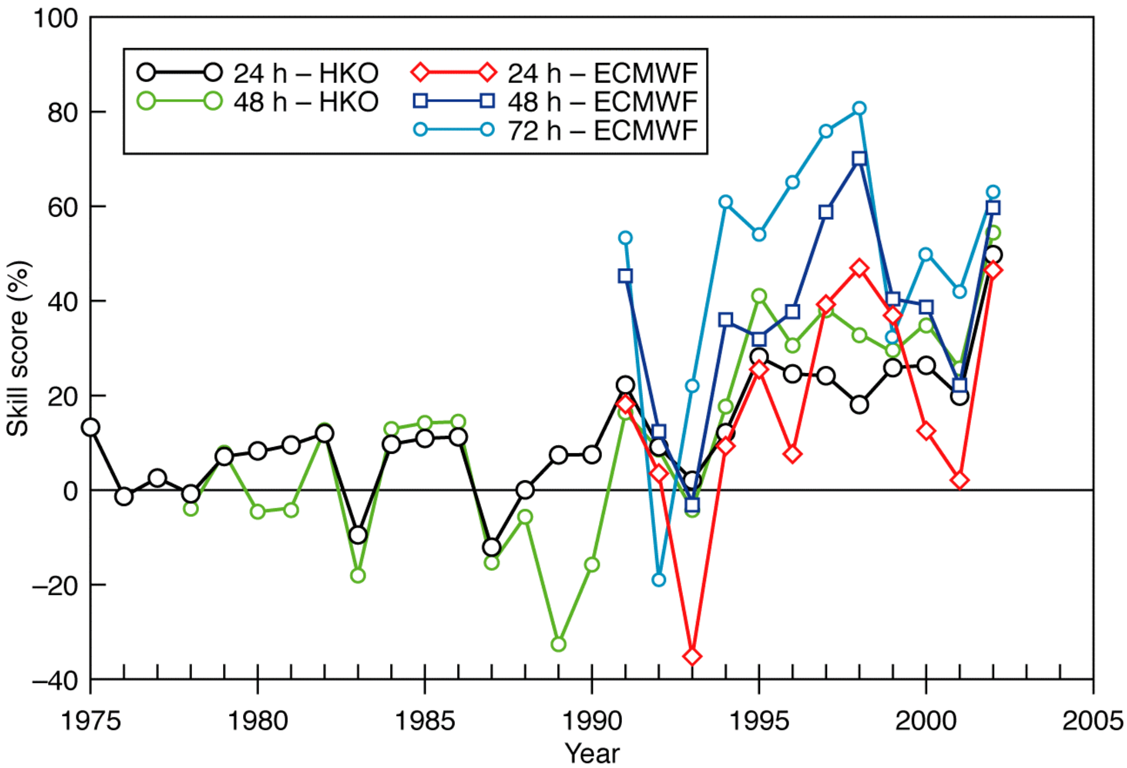
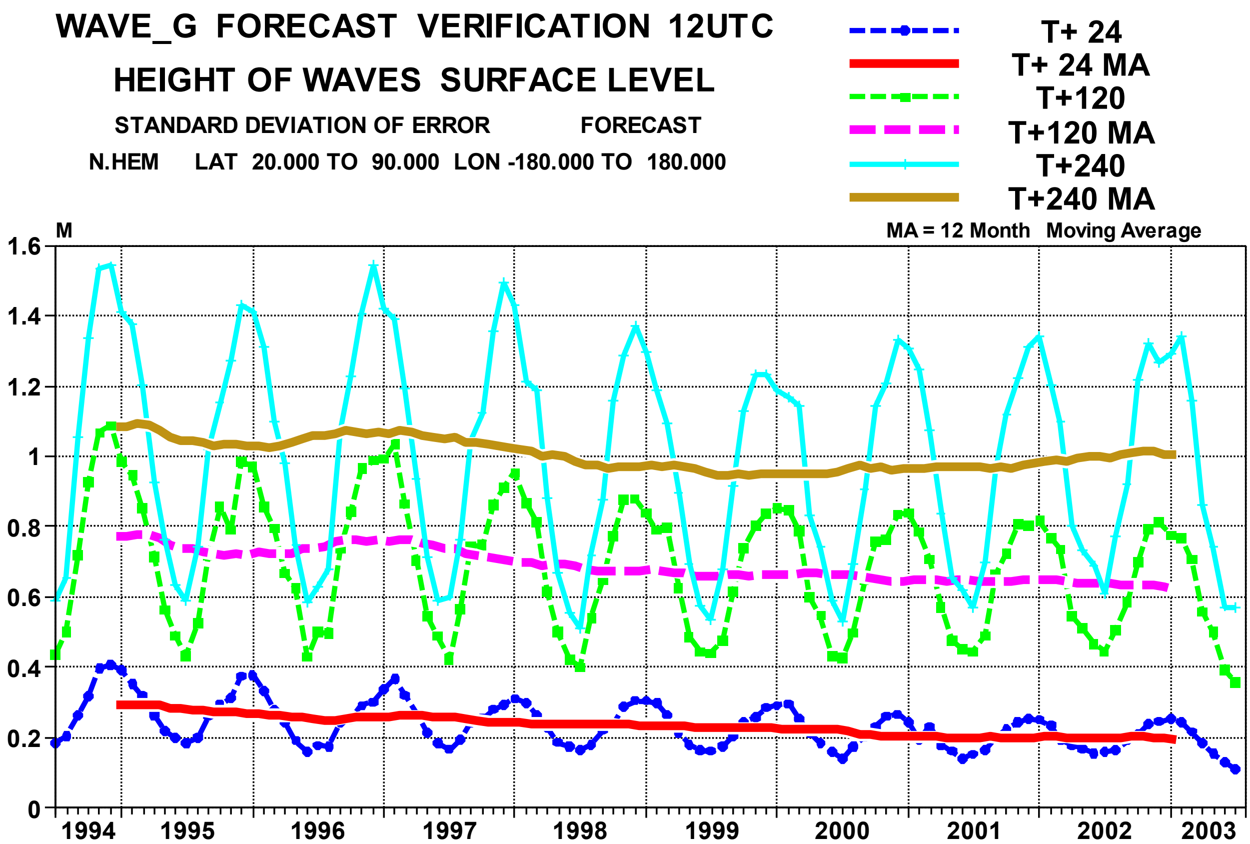


Figure 34: Skill relative to climatology-persistence for ECMWF model forecasts and the HKO subjective forecasts for TCs over the verification area 10-30N, 105-125E (Courtesy from Hong-Kong Observatory)



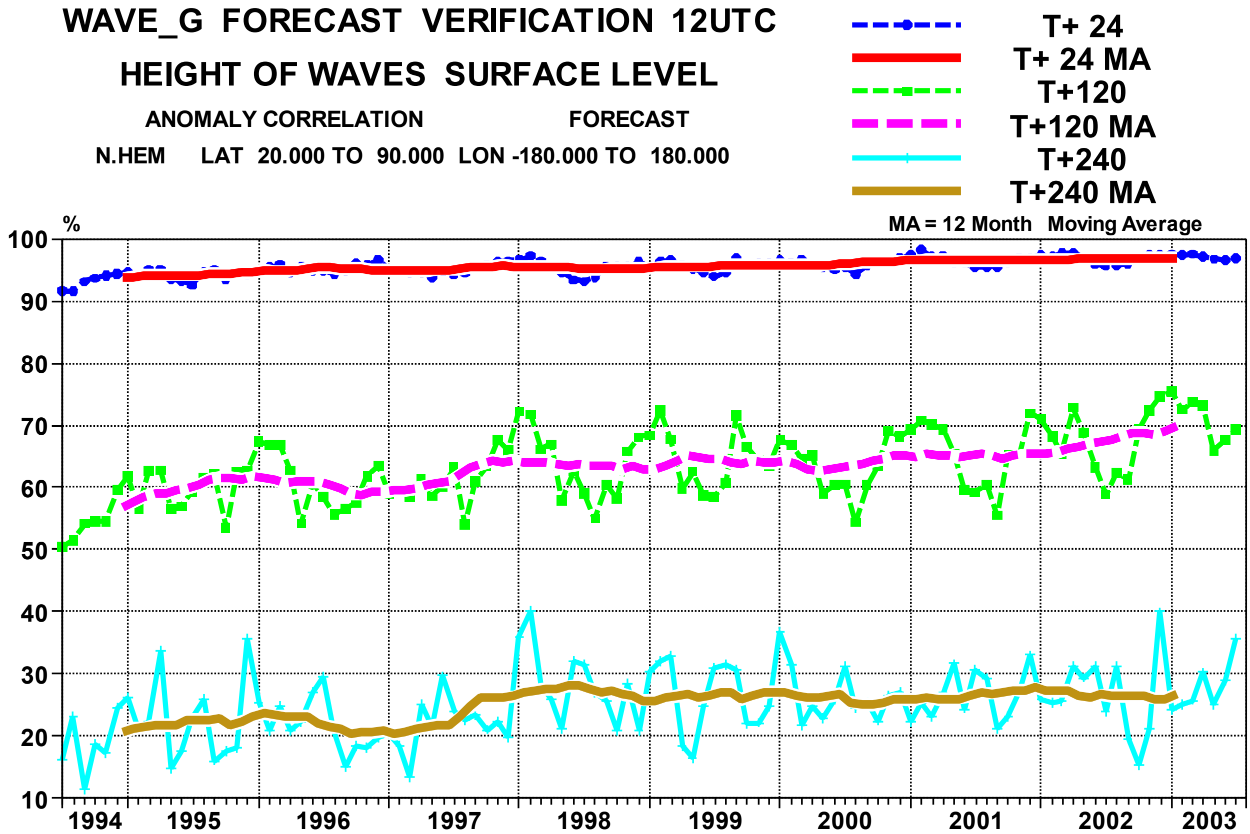
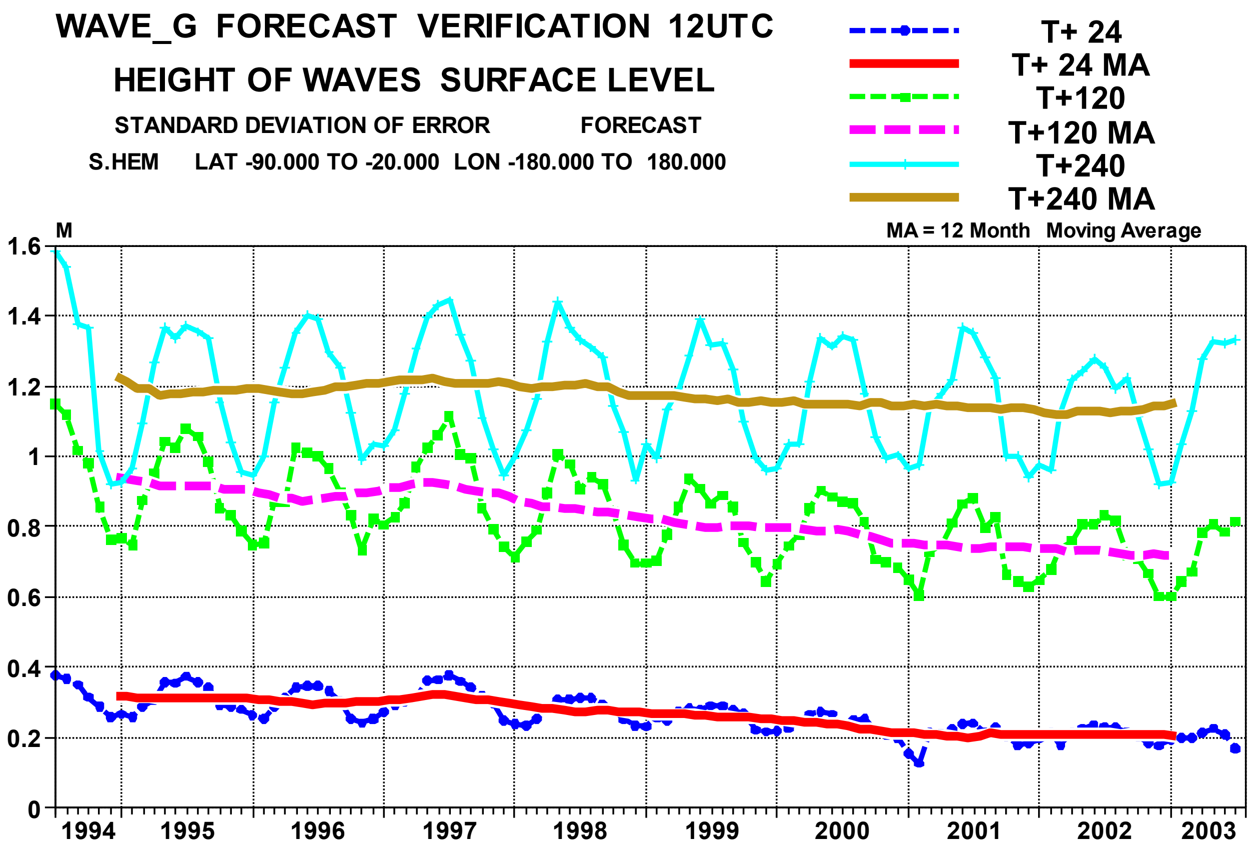


Figure 35: Scores (std and anomaly correlation) of oceanic wave heights verified against the analysis (Northern Extratropics)



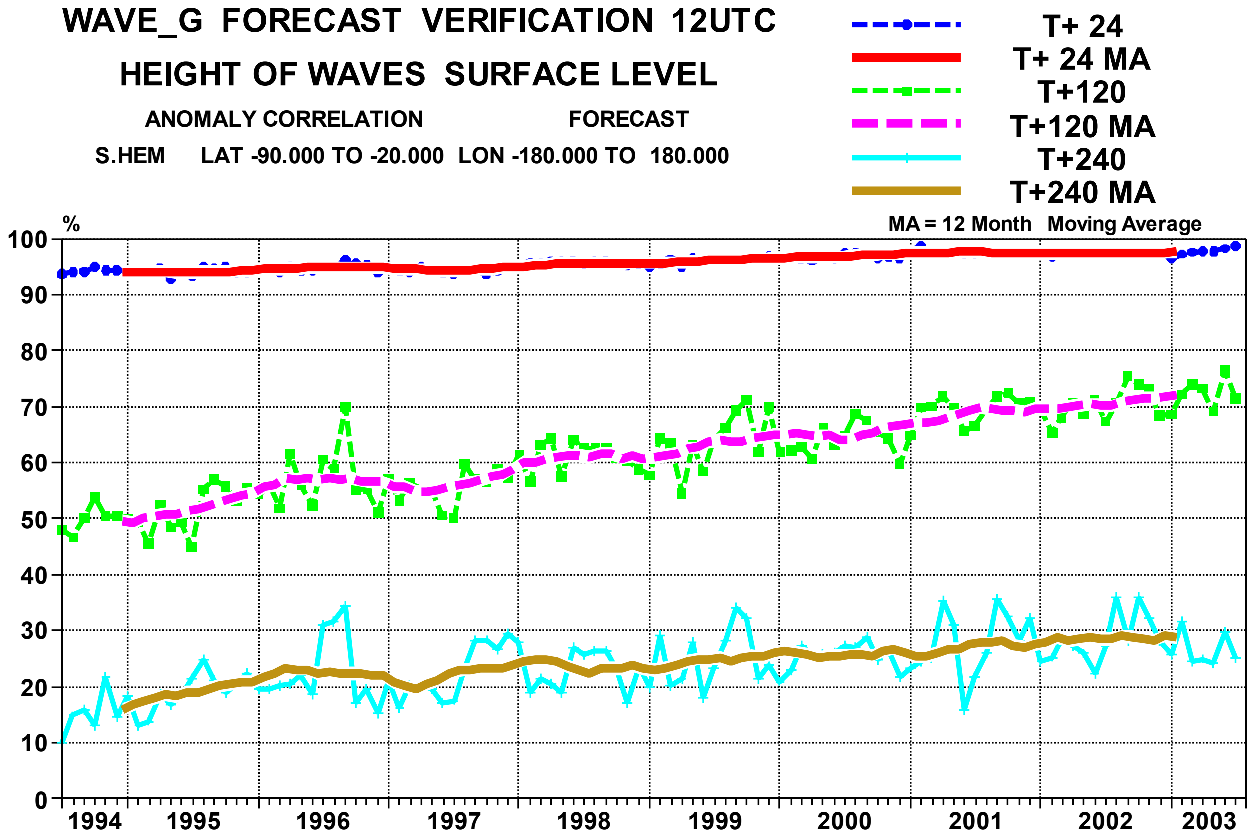


Figure 36: Scores (std and anomaly correlation) of oceanic wave heights verified against the analysis (Southern Extratropics)

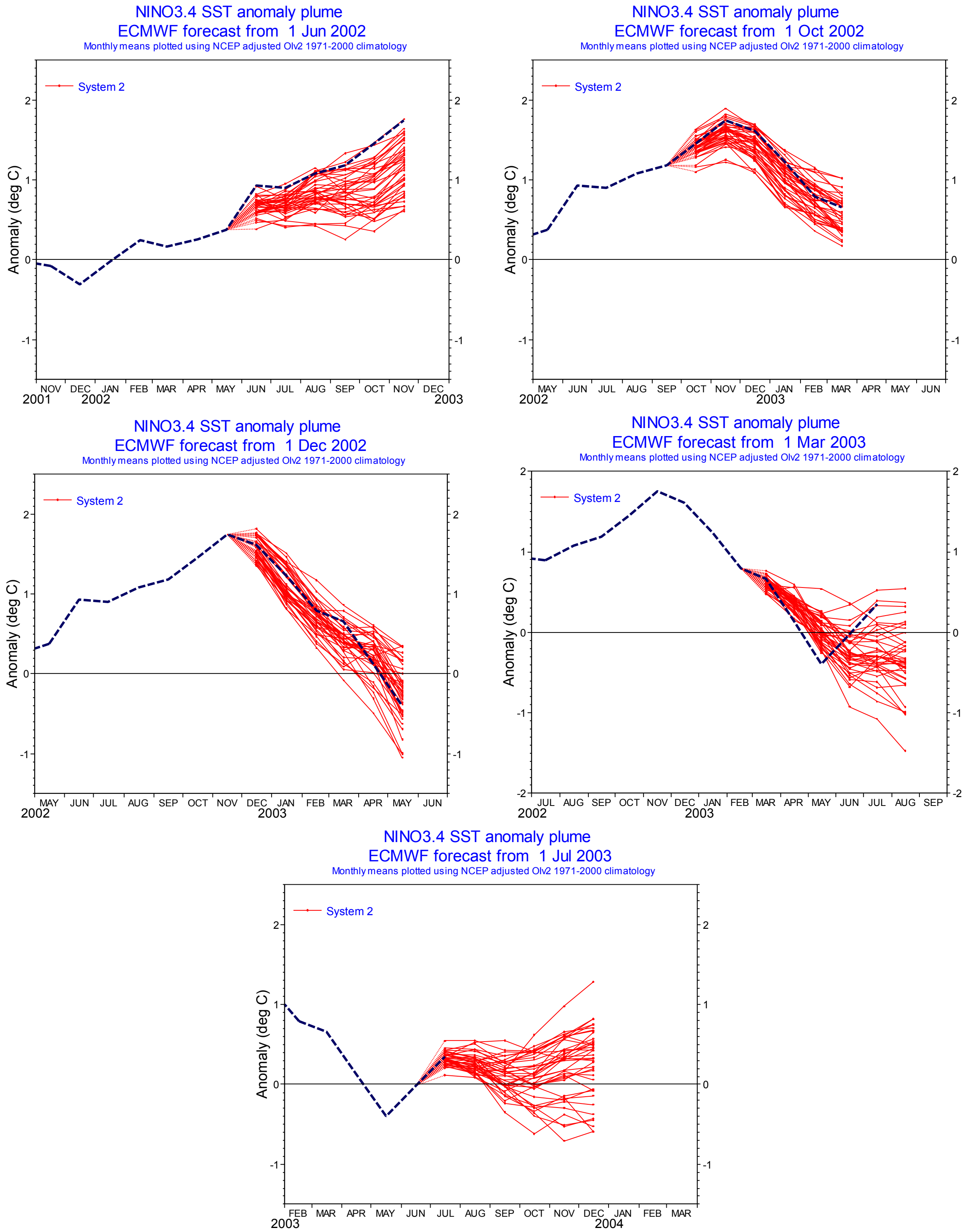


Figure 37: Plot of forecasts of Nino-3.4 at four start dates June, October, December 2002 March 2003 and July 2003. The red lines represent the 40 ensemble members. The heavy dashed line represents subsequent verification.

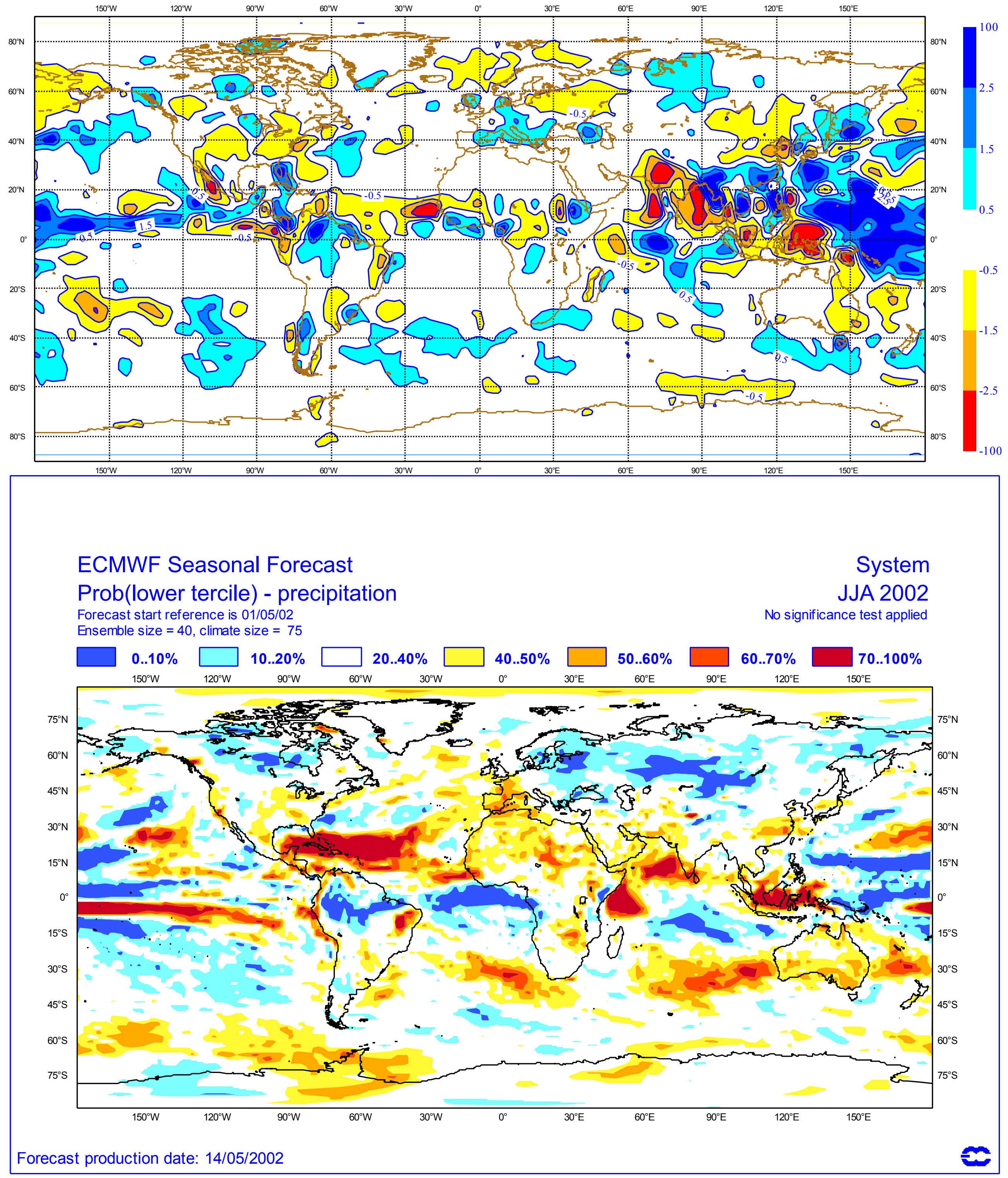


Figure 38: JJA 2002 precipitation anomalies from GPCP data set. Anomalies are computed with respect to mean climate 1987-2001 (upper panel) and) Probability of precipitation for JJA 2002 associated with the lower tercile from seasonal predictions started in May 2002.

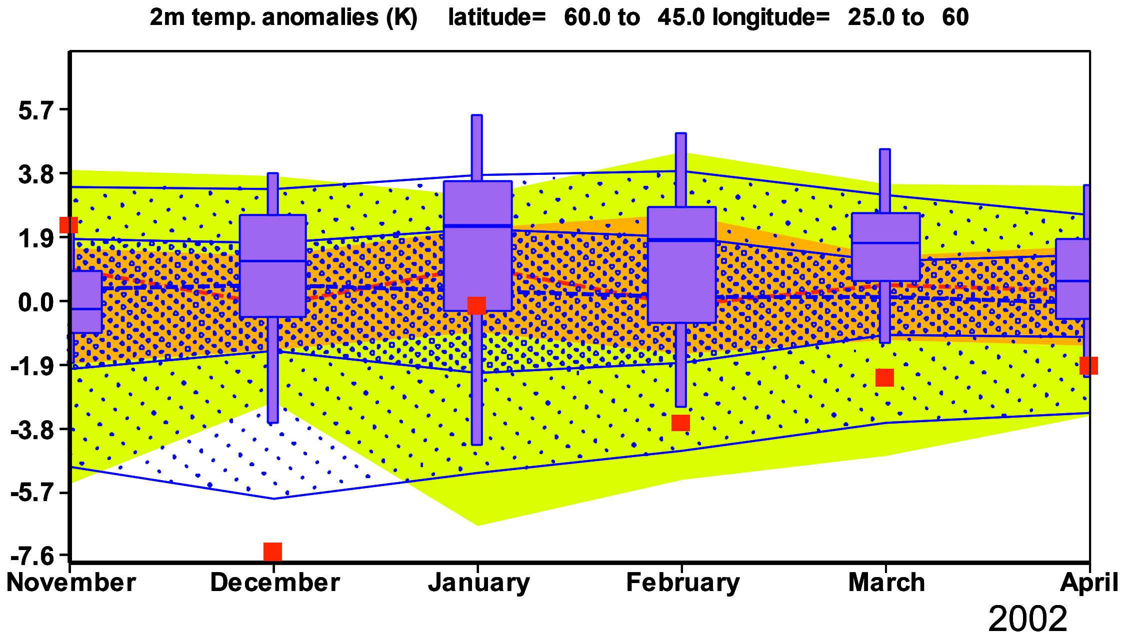


Figure 39: A climagram, showing 2 meter temperature predictions from November 2002. Predicted monthly-mean values are represented in blue and model climatological values are dotted blue areas within Q25/Q75. The climate extremes (95% and 5%) and the median are also shown. The same values from the forecast distribution are shown in the usual "box and whiskers" representation, while the ERA40 climate is in orange/yellow-green shading. Observed anomalies are red squares, showing the very cold anomaly observed last December in Eastern Europe