414

Verification statistics and evaluations of ECMWF forecasts in 2001-2002

F. Lalaurette, L. Ferranti, A. Ghelli,
Ø. Saetra and H. Böttger

Operations Department

July 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

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. A large part of this section 3 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 Annex A. A note describing pre-operational testing procedures is provided in Annex B.

The set of verification scores shown here is mainly consistent with previous years (Lalaurette and Ferranti, 2002), in order to help compare the performance year after year. Aspects related to experimental products such as those for severe weather or monthly forecasts are treated in separate documents.

Please note that in September 2000, verification pages were created on the ECMWF web server and 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:

- October 2001: the new seasonal forecast "System 2" runs in operational mode; its products have been used as the "operational" products on the web server in replacement of system 1 since June 2002 (products from system 1 are still generated and posted on the web for ECMWF Member States and Co-operating States);
- 22 January 2002: Introduction of Cycle 24r3. This version includes the following important changes that together will affect all components of the system (Data assimilation, atmospheric and oceanic waves forecasts, EPS):
    1. More data are activated (SeaWinds data from QUIKSCAT, less thinning of aircraft observations, improved selection and scan correction of ATOVS radiances);
    2. Pre-processing (bias correction) of SSMI data and redundancy checks for SYNOP SHIP, DRIBU, AIREP, TEMP and PILOT observations are refined;
    3. 4D-var analysis algorithms are upgraded: pre-conditioning is added to the minimisation, resulting in a 40% reduction of the number of adiabatic iterations; correlation functions slightly revised (compacted); the radiative transfer model used to assimilate satellite radiances has been completely re-written; the observation timeslot has been reduced from 1 hour to 30 minutes;
    4. Model changes include a new finite-element vertical discretization, small changes in the convective precipitation scheme and supersaturation checks and an improved temporal scheme for oceanic wave generation;

5. Initial EPS perturbations in the tropics are included. The perturbations are generated for a maximum of four target areas by Gaussian sampling of the 5 leading diabatic; singular vectors for each area. The Caribbean (0°-25°N and 100°-60°W) is always a target area, as is the area around every tropical storm of a category larger than 1 between 25°N and 25°S. (In the event that these criteria produce more than four target areas, the closest areas are merged.)

- 4 March 2002: New blacklisting procedure is introduced to avoid using METEOSAT AMV (ex-SATOB) winds during time-slots likely to be affected by solar-eclipse problems
- 5 March 2002: EPS wind gust bugfix (maximum was over 6 instead of 12h)
- 9 April 2002: Introduction of Cycle 25r1. This version includes the following changes:
  1. A revised shortwave radiation scheme with variable effective radius of liquid cloud water;
  2. Retuning of the land surface (TESSEL) parametrization to reduce winter/spring warm biases in low-level temperatures;
  3. Improved physics for the oceanic wave model;
  4. Improved wind-gust post-processing;
  5. The activation of new data streams in the assimilation: water-vapour radiances from Meteosat-7, SBUV and GOME ozone data, and European wind profilers;
  6. A bug in the convective momentum transfer was also fixed;

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 the free atmosphere medium-range forecasts

3.1 ECMWF scores

3.1.1 Deterministic (T511) model

Figure 1 gives the evolution of forecast skill as measured by the 12-month moving average from 1980 to 2001 of the 500 hPa height error normalised by persistence, for the Northern Hemisphere and the European area. The last month included in the statistics is July 2002. These curves show that the performance is still steadily improving for the fourth consecutive year up to Day 4 over Europe, up to Day 5 over the Northern Extratropics. Beyond these ranges, 12-month averages of deterministic scores still show quite a large amount of skill in the forecasts - the large variability of the errors makes it more difficult, however, to extract from a 12-month average a signal than can without doubt be attributed to model changes. Forecast ranges during which the monthly and annual mean anomaly correlation of the forecast with the verifying analysis stays beyond 60% can be found in Figure 2. One of the remarkable results here is that after several years when the performance over the Southern Hemisphere has improved dramatically (which is most likely due to the improved data assimilation of satellite radiances), both hemispheres have now reached annual mean forecast ranges of 7.5 days - a range that was not reached even during winter months before 1983. Another remarkable feature when comparing Northern and Southern Extratropics is that the variations in skill between summer and winter are much stronger in our Boreal hemisphere, reflecting the challenge that the correct treatment of physical processes over the large proportion of land surfaces and mountain ranges are posing for medium range forecasts in this hemisphere. From this point of view, it seems that the large dip in

ECMWF skill curves during spring/summer 1999 was not repeated, probably thanks to the numerous improvements, among other developments, of the land-surface parameterisation and radiation scheme introduced since then. Figure 3 shows the distribution of errors for 1000-hPa surfaces (which can be considered as very similar to pressure fields at sea level) this winter compared to the previous two. This shows that two 96h-forecasts out of three correlated better than 90% with the verifying analysis this winter, a result that improves on previous winters' already high level of performance. It is also worth mentioning that the number of relatively bad 96h-forecasts (less than 70% anomaly correlation) has been reduced for the second consecutive year: there were still 5% of them in 1999-2000, while there was none this winter.

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 4. This year has confirmed the reduction of errors both at 850 and 200 hPa that was observed last year. In the same Figure 4, information is provided regarding the quality of the stratospheric forecast (50-hPa level). Here again, the performance this year has kept the high level reached since the increase of vertical resolution introduced in March 1999 (cy19r2).

One of the noteworthy results shown over the past few years is that the improvement of the forecast quality means that the consistency of the forecast from one day to the next (when comparing for the same valid date) has improved a lot. This level of consistency has been kept this year, as can be seen in Figure 5 with the RMS difference between consecutive forecasts.

3.1.2 Ensemble Prediction System (EPS)

As shown in the previous paragraphs, the level of skill of the deterministic T511 model is now very high at predicting the general circulation up to day 4 to 5. Beyond this range, although run at lower (T255) resolution, the ensemble can be used to refine single-value estimates in order to filter out small scale, unpredictable features in the late medium range. This can be seen from Figure 6, where distributions of scores for 850hPa temperature anomalies are compared for the T511 and the Ensemble Mean forecast over Europe this winter. As expected, the T511 forecast estimate is better up to day 4, but from day 5 the distribution is clearly shifted in favour of the ensemble mean. This filtering effect could be achieved on the basis of statistical estimates of error-scale relationships: that the EPS is achieving more than this, can be seen in Figure 7, which shows that the variations in scattering (spread) of the ensemble members are a good predictor of the forecast skill - something that is due to the dynamical properties of the EPS system. On this Figure 7 have also been reported in blue values from the previous winter (2000-2001) which was affected by an underestimation of the ensemble spread until the rescaling of perturbations on 5 February 2001: this problem clearly affected the performance of the system, as reported at the time. The scatter plot lay mainly above the diagonal (too small spread in the forecast) and had no resolution (large dots aligned more along the horizontal than the diagonal): all these features have disappeared from the system this winter (red dots).

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, reliability diagrams (Figure 9) or ROC curves (Figure 10). Time series of Brier scores (Figure 8) show that the EPS has followed the general trend of improvement of the forecast this year: for moderate anomalies (+/-4K), the values reached at Day 6 today are around 0.35, values which were found for Day 4 in 1995 - this is a 2-day gain in predictability in 6.5 years. Although the trend is not as steady for strong anomalies (+/- 8K) the levels of skill achieved this year have also improved recently. Figure 9 and Figure 10 compare the performance of the forecasts in winter 2001-02 and 1999-2000 (winter 2000-01 is omitted due to the EPS problems that year; using 1999-2000 also has the advantage of showing the impact of the resolution upgrade in November 2000,). For all thresholds,

both the reliability diagrams and the ROC curves show a very clear improvement in performance over two years ago, which confirms the findings of studies conducted in pre-operational mode before the upgrade of the EPS from T159 to T255. The most significant development of the EPS this year is however found in the tropics, with the operational use of targeted, diabatic singular vectors around tropical cyclones and in the Caribbean basin (see section 2). Although some beneficial effects are expected from this procedure over Europe, when some tropical cyclones experience an extratropical transition (this should happen for the first time during the autumn 2002), most of the benefits have been found in the development of probabilistic products for tropical cyclone forecasts and are documented in a separate document dealing with Severe Weather developments.

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 11 shows time series of such scores over Northern Extratropics for both 500-hPa 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 from other models has been slightly reduced. This lead extends a bit more into the spring season this year, and the level of errors is again much smaller than in 1999. The difference is even larger in the Southern Extratropics, where it is can be seen throughout the year, including the warm season (Figure 12).

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 13 shows distributions of 500-hPa height anomaly correlations of day 5 forecasts with verifying analyses over Europe during the cold season (15 Oct.-15 Apr.) for ECMWF, the UK Met Office, Deutscher Wetterdienst, 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 situation in the Tropics is summarised in Figure 14. ECMWF scores are this year again similar to those from the UK Met Office, and both centres are now joined by US NCEP at 250hPa. A remarkable reduction in errors occurred in 1997 at 850hPa (first in a series of refinements of background error statistics) and since then, further developments of the model physics and data assimilation seem to have maintained a steady trend of error reductions.

Finally scores are presented using radiosondes as the verifying dataset, both in Europe and in the Tropics (Figure 15). They confirm the conclusions drawn previously from the field verification against each model's own analysis, although it can be seen that ECMWF low level forecasts in the Tropics compare better to the radiosondes than those of other centres.

3.2.2 T255 (EPS Control) multi-analysis system

In order to provide a diagnostic tool for a better understanding of the sources of forecast errors and, particularly, the relative parts of model and analysis errors, ECMWF has been running a multi-analysis, T255 forecasting system since March 2001. Initial conditions for this system are kindly provided to ECMWF by the US NCEP, Deutscher Wetterdienst, Météo-France and UK Met Office from their own global forecasting systems. After interpolation on the T255 model grid of the differences between these systems'

daily (12UTC) analyses and the one from ECMWF, 4 daily forecasts are run up to 10 days. An additional forecast is run using a consensus averaged from all 5 centres' analyses. Verification has been collected since October 2001 for all these streams. Figure 16, Figure 17 and Figure 18 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 notable feature is that in most cases, the closest to the ECMWF control forecast is the one run from a consensus of all the analyses. Clearly the benefit gained from using ECMWF's own analysis is the greatest in the Southern Extratropics (Figure 18), where up to 6 days, the control (ECMWF T255) forecast outperforms all other runs, sometimes by a considerable amount. The case of Europe (Figure 17) is, however, one in which the different combinations come much closer - indeed both the consensus and the forecast run from NCEP analysis outperformed the control run by day 6-7. More investigations will be needed to gain a better understanding of these results, and to determine whether they are due to the interpolation procedure (the 4D-var analysis aims at providing an optimum T511 trajectory, not a T255 one) or to sampling unusual meteorological situations.

4. Weather parameters and oceanic waves

4.1 Deterministic (T511) model

Figure 19 shows the monthly mean and standard deviation of the 2m temperature and specific humidity errors over Europe up to July 2002, 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 characterised the last two years has vanished, an expected feature after the re-tuning of land surface parameterisations on 9 April 2002. For the second consecutive year, daytime temperature random errors (STD) have been kept well below 3K, while summer time nighttime errors are only 2K on average. This winter has, however, seen slightly larger negative night time biases than in recent years, but still much smaller than those found in 1995/96 - indeed they are likely to be more related to different meteorological conditions than a drift in the model performance. The trend for humidity is mainly neutral, still showing a systematic dry bias at daytime compared with SYNOP reports. Part of the problem might be related to a systematic dry bias from radiosondes and a significant research effort has been devoted this year towards improving the diagnosis and ultimately the quality of humidity analysis. Figure 20 shows the same comparison for total cloud cover and 10m wind speed forecasts. Neither shows any noteworthy change in performance compared to previous years.

The monthly mean error of the precipitation forecasts at day 3 over Europe is shown in Figure 21, for 00, 06, 12 and 18 UTC - again no major change compared to last year, and the reduction in the diurnal cycle of convective activity since the changes introduced in 1999 are confirmed. In order to get some more specific insight into the realism of the model for high precipitation events, both model forecasts interpolated at SYNOP station location and observation distributions have been compared (Figure 22). The dataset involved has more than 120,000 daily observations from 1531 stations reporting on the GTS North of 20N. It can be seen that although there is a large representativity difference between model grid forecasts and SYNOP observations, model distributions have been coming gradually closer to observations over the last two years than was the case two years ago with the T319 model. Frequency Bias Indices (ratio of the number of cases when the model forecasts more than a given threshold of rain divided by the number of observations) are also shown for the three seasons in Figure 22. Although over the three winters all curves show a sharp downward trend towards zero for very large amounts (the FBI should be one for a perfect, infinite resolution

model) the model distribution stays within 20% of the observed frequencies from 10 to 40mm, which was far from the case with the T319 model in 1999-2000.

This year the verification procedures for 2m-temperature and precipitation have been enhanced to provide not only biases and standard deviation of errors, but also more elaborate measures of skill. Figure 23 shows the reduction of 2m-temperature errors with reference to persistence, while Figure 24 give the time series of Equitable Threat Scores of precipitation forecasts over recent years. The latter confirms that the maximum skill for precipitation forecasts is found during the winter season , while the use of persistence as a reference removes most of the seasonal cycle from 2m-temperature scores. There is a modest, positive trend for improvement of 2m-temperature forecasts, mainly during night time since October 1999 when higher vertical resolution was introduced in the planetary boundary layer; a modest improvement both in day time and at night time is also found since the horizontal resolution increase in November 2000. The trend for the improvement of precipitation forecasts (Figure 24) is more remarkable, notably for moderate to high rainfall events (>5mm/day). Work is currently in progress to extend these measures of skill to super-observations gathered using European high resolution datasets provided by our Member States and Co-operating States.

Verification scores from the global oceanic wave products are shown in Figure 25 and Figure 26. They show that the steady improvement of the forecasts has continued this year, most notably in the Southern Extratropics.

4.2 Ensemble Prediction System (EPS)

Time series of Brier Skill Scores for daily precipitation forecasts in excess of 1, 5, 10 and 20mm at day 4 and 6 are plotted in

Figure 27. Verification is provided in this figure by amounts of rain reported in SYNOP messages originating from Europe. Although such a procedure has its flaws (irregular coverage and limited representativity of isolated stations when compared to model precipitation fluxes), it is used here rather than using short-range model estimates of the precipitation fluxes, as these see their characteristics changing with the model resolution, physics and even data assimilation techniques along the 7 years covered by this graph. The improvement in skill of the ensemble forecast exhibited by these curves is quite remarkable. A closer inspection reveals that forecasts for small amounts (1mm) have gained a lot of skill since October 19991 when (among other changes) the vertical resolution of the PBL was increased and the large scale precipitation scheme was revised to cope with partial cloudiness - before this date, the 1mm threshold had less skill than larger thresholds due to an excessive number of light rain events. The other step upwards that seems to be detectable on the time series is in April 2000, corresponding to the increase in horizontal resolution introduced in November 2000.

Reliability diagrams for the verification of 24-h accumulated precipitation over Europe are shown in Figure 28. The reliability curves for the same period in 1999-2000 (before the horizontal resolution increase, and before the unintentional reduction of amplitude of EPS initial perturbations occurred) are also reported for reference in blue. These diagrams show a tendency for the EPS forecast to be more confident that amounts in excess of 5 and 10mm will happen than found in local observations. This feature has been shown to be related to the difference in scale between the observation and the model resolution. Statistical downscaling of

[1] Because 12 month samples are used, there is a 6 month negative phase lag between when a change is introduced and when it is to be detected in the time series: changes introduced in October 1999 are already affecting the Brier score plotted in April 1999, as this one gathers events from November 1998 up to October 1999.

EPS probabilities or upscaling rainfall observations using high-resolution networks lead to a much better agreement between forecasts and observations. While the latter procedure is being developed at ECMWF using additional data provided by our Member States and Co-operating states, downscaling is usually applied in Member States - a pilot study has been conducted at ECMWF in collaboration with the Hungarian Met Service in 2001, the method and results of which can be provided on request.

ROC curves for the same events are available in Figure 29. They show a remarkable improvement in the forecast of light rain events compared to two years ago. Smaller improvements are to be found for events in excess of 5 and 10mm per day. The 20mm event seems to show a degradation of the signal detection ability of the system since 1999-2000. Although this cannot be overstated due to the limited sample, the different meteorological context and the representativity problems highlighted above, there have been cases when excessive triggering of large amounts of rain were observed in the forecasts in recent months that might cause such increased false alarm rates. This potential problem and its relation to some of the physical process parametrizations is currently investigated.

The verification of wave ensemble products that have been generated in operations since June 1998 has made good progress this year (Saetra and Bidlot, 2002). In Figure 30, reliability diagrams for the day 5 forecast probabilities of wave heights in excess of 2, 4, 6 and 8 m are plotted. Generally, the results indicate good reliability, particularly for the 4 m threshold. For threshold values of 2 and 6 m, the reliability is also quite good, but there is a small tendency for the points to lie below the diagonal line, which indicates that high probabilities are forecasted slightly too often. The Brier score is smaller for the two largest threshold values. This is due to the fact that for threshold values of 6 and 8 metres, the vast majority of both forecasts and observations are in the two lowest forecasting classes, while for 2 and 4 m, the EPS forecasts are more evenly spread over the range of probabilities. For the 8 m threshold, the reliability curve shows the behaviour typical for situations with insufficient sample size (rare event).

In Figure 31, the 90-percentile (Q90) of the absolute error for significant wave height is given as a function of the ensemble spread for the day 5 forecast range. The observations in this case are the global altimeter data. The ensemble spread is defined as the difference between the upper and lower quartiles of the ensemble. The absolute error is defined as the distance between the observed value and the control forecast. The black solid line is the result when all data available globally are taken into account. The triangles mark the centre points of each bin for the spread. The number of cases that have been used for each bin is indicated by the histogram in the upper left corner. Q90 shows a clear dependency on the ensemble spread. The data have also been divided in different areas with different characteristics, and hence different variability. Of course, the choice of the percentile for the observed errors is more or less arbitrary; any other percentile gives qualitatively similar results. As an example, the Q75 fits roughly with the diagonal line. A very approximate rule of thumb may therefore be that the error in the wave forecasts is expected, with 75% probability, to be less than or equal to the inter-quartile range of the wave ensemble.

5. Seasonal forecasts

5.1 General comments

During the past year we have put a lot of effort into developing a comprehensive verification scheme using facilities (post-processing, archiving) provided by the new “system 2” forecasting system. In order to help evaluate the system, verification for both the “old” system 1 (to be phased out later this year) and the operational system 2 are computed and published whenever possible. Our aim is to provide the users with 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. Extra care has to be taken when dealing with these forecast ranges as both the temporal and spatial scales involved are so different from those involved in medium range, as is the signal over noise ratio. Significance testing methods are therefore given increasing importance in our system, something we hope will be increasingly reflected in the verification statistics provided to our users. More on the validation of System 2 can be found in Anderson *et al.*., 2003.

A WMO expert team in which ECMWF actively participated has been discussing these issues this year and has provided guidelines for a Standard Verification System for Long Range Forecast (SVS-LRF).

5.2 Results

Since the beginning of July 2001 verification is available at:

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/verification/

At this site, estimates of model bias for a wide range of variables, including zonal averages, time series of a set of indices of SST and large-scale patterns of variability such as: the Southern Oscillation Index (SOI), the Pacific North American Pattern (PNA) and the North Atlantic Oscillation (NAO) are available. A suite of verification scores for deterministic (e.g. spatial anomaly correlation and Mean Square Skill Score Error (MSSE)) and probabilistic forecasts can be viewed. This site contains verification for both seasonal forecast systems: System 2, the operational system, and System 1 the "old" system still running in parallel with the operational version. It should be stressed that the development of the new operational seasonal forecast verification system has benefited a great deal from developments in the framework of the DEMETER project.

The robustness of verification statistics is always a function of the sample size. In the case of seasonal forecast verification, the sample size of 15 years is considered just sufficient. For this reason verification is performed in cross-validation (Michaelson 1987) mode using the whole set of forecast data available: hindcasts and real time forecasts with no distinction.

For verification purposes, forecasts started in February, May, August and November are used. Results are shown as 90-day means with 1 and 3 month forecast lead. GPCP data are used to verify precipitation, while the other atmospheric parameters are verified against ERA 40. For the period when the re-analysis is not yet available, the operational analysis is used.

The verification period is 1987 to 2001 for System 2 and 1991 to 2001 for System 1. Every year the site will be updated by adding another year to the verification period. The maximum number of forecasts available for the entire verification period is used. For example for System 2 only 5 members are considered, since in the period 1987-2000 5 members are available. However, for the forecast started in November (and soon for the forecast started in May) an ensemble of 40 members is available, so for these particular forecasts all the 40 members are included in the verification.

The ensemble for System 1 consists of 11 members. It is worth mentioning that the System 1 ensemble is less homogeneous then the ensemble from System 2. In fact, the System 1 hindcast (1991-1996) consists of forecasts started on the first of the month with perturbed Sea Surface Temperature, while the real-time forecasts, up to Jan 1988 were run 3 times every week; later they were run daily starting from the current date.

5.2.1 Bias maps

In order to document the model drift, bias maps are produced although it should be understood that such biases are removed from the forecast products as part of the post-processing. The bias is the difference between the seasonal forecast mean climate and the corresponding verifying mean climate. Figure 32 compares biases from System 2 and System 1. Clearly the strong negative biases that were known as one of the characteristic signatures of System 1 have now disappeared to a large extent in the new, System 2 forecast system. Zonal mean biases are also computed for the whole latitudinal band and for the Atlantic and Pacific regions.

5.2.2 Forecast indices

Time-series anomalies of several indices averaged over a 3 -month period have been computed. For example the NINO3 verifications for both systems are generated and shown in Figure 33. Summer 2001 conditions were near normal, a feature well captured by both System 1 and System 2 February forecasts.

5.2.3 Single-value forecast scores

The most usual way to summarise the information from an ensemble of forecasts into one value is to use the grand mean of all members. It should be stressed that although such values are often referred to as deterministic forecasts, the grand mean involves a stochastic process removing random, unpredictable errors from the forecast. As a consequence, ensemble mean charts are unrealistic, if compared to analyses (many small scale features do not show up), but they are usually are the best single-value estimates, if evaluated using RMSE measures. For such forecasts the following verification data are provided both as global maps and averaged over pre-defined areas:

- Anomaly correlation computed for grid point values; correlation is based on the all years available, as indicated in the plot header.
- Grid point values of the Mean Square Skill Score (MSSS); the MSSS is positive (negative) when the accuracy of the forecasts is greater (less) than the accuracy of the reference forecasts. The reference forecast used for the skill scores is simply the sample climate computed by averaging the analysis. They are usually called "internal" climatological forecasts (Murphy 1988).

In addition, the ratio between the forecast and the analyses variances is provided, as it can be shown that the MSSS is the sum of an anomaly correlation and a variance ratio component. An example of such maps is shown in

Figure 34 for 850-hPa temperature 4-6 month forecasts, originated in February 1987-2001. It can be seen there that, although the forecasts correlate positively with the verifying analysis most of the time (upper panel), the anomaly amplitude in the ensemble mean is much smaller than analysed (bottom panel) resulting in MSSS that are negative over large areas, most notably in Europe (middle panel). Similar maps for System 1 (only 1990-2001 in this case) are in Figure 35. They show that System 2 in general produced an improvement in the summer forecasts - see Southern Africa and Australia and also some parts of Europe.

5.2.4 Probabilistic scores

The full content of the information provided by the seasonal ensemble forecasts is only accessible in multi-valued, probabilistic mode. Basic methods for verifying probabilistic forecasts have been in use for several years at ECMWF for medium-range EPS products and the methodology is now being naturally extended to seasonal forecasts. The Relative Operating Characteristics curve shows - for a range of different probability thresholds - hit rates versus false alarm rates of forecasts of a particular event in different regions. The event

thresholds are defined with respect to terciles from model and observations climatologies. The blue dotted lines indicate the 95% confidence interval for the hit rate values. Figure 36 shows the ROC diagrams for 850-hPa temperature summer forecasts over Europe, while Figure 38 is for rainfall anomalies in the Tropics. These curves indicate that most of the skill of the system lies in the forecast of rainfall anomalies in the tropics (the system is less good at forecasting "near normal" conditions there). By contrast, the signal detection for 850-hPa temperature anomalies over Europe is much weaker.

6. Summary

This year the very good medium-range forecast scores have continued to maintain the high levels reached in 2000-2001. The reduction of summer time forecast errors over Europe is evident but this issue is still kept under close scrutiny. The reduction of warm biases following the re-tuning of the physics is also apparent and the distribution of heavy precipitation events is coming closer to the observations, partly as a result of increased horizontal resolution in 2000. A lot has been done both to extend the verification system (seasonal forecasts and also multi-analysis ensembles and skill scores introduced for 2m-temperature and precipitation) and to make the results available on the web server in real-time to ECMWF Member States and Co-operating states. Following the introduction of geographical information on the quality of seasonal forecasts (see

Figure 34 for example), there are plans to extend this type of verification to medium-range ensemble forecasts.

References

Anderson, D., T. Stockdale, M. Balmaseda, L. Ferranti, F. Vitart, P. Doblas-Reyes, R. Hagedorn, T. Jung, A. Vidard, A. Troccoli and T. Palmer, 2003: Comparison of the ECMWF seasonal forecast Systems 1 and 2, including the relative performance for the 1997/8 El Nino. *ECMWF Tech. Memorandum*, **404**

Lalaurette, F. and L. Ferranti, 2001: Verification statistics and evaluations of ECMWF forecasts in 2000-2001. *ECMWF Tech. Memorandum*, **346**

Michaelson J. 1987: Cross-validation in statistical climate forecast models. J. Clim.Appl. Meteorol. 26 1589-1600.

Murphy 1988: Skill scores based on the mean square error and their relationship to the correlation coefficient. Mon Weather Review 116 2417-2424

Saetra, Ø and J-R Bidlot, 2002: Assessment of the ECMWF ensemble prediction system for waves and marine winds. *ECMWF Tech. Memorandum*, **388**

Simmons, A.J. and A. Hollingsworth, 2001: Some aspect of the improvement of skill in numerical weather prediction. *ECMWF Tech. Memorandum*, **342**

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

A.1 Deterministic upper-air forecasts (sections 3.1.1 and 3.2.1)

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 (boundary co-ordinates are reproduced in the figures' inner captions); when other centres' scores are produced, they have been provided as part of the WMO/CBS exchange of scores between GDPS centres; when verification scores are computed using radiosonde data (Figure 15), 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 4, Figure 14, Figure 15) 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:

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

Anomaly correlation scores are spatial correlations 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; they show for each month the average of the ranges at which the daily forecast dropped below 60% of anomaly correlation.

A.2 Probabilistic forecasts (section 3.1.2, 4.2 and 5.2.4)

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 reliability diagrams, Brier Skill Scores and the Relative Operating Characteristics (ROC) curves. The latter are also used for seasonal forecasts (section 5.2.4); events are also defined with respect to a model-based climatology in this case.

Reliability diagrams (Figure 9, Figure 28, Figure 30) simply show red curves that are frequencies of occurrence (y-axis) of a predefined event for forecasts that gave similar probabilities (x-axis) that this event should occur. A probabilistic forecast is perfectly reliable, if the red curve in these diagrams lies along the diagonal.

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

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

As with 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 BS can be split into the sum of the sample climate *uncertainty* and the forecast *reliability* (BS_REL), minus the forecast *resolution* (BS_RSL):

- *uncertainty* varies from 0 to 0.25 and indicates how close to 50% the occurrence of the event was during the sample period (uncertainty is 0.25 when the event is split equally into occurrence and non-occurrence);
- *reliability* tells how close the frequencies of observed occurrences are from the forecasted probabilities (on average, when an event is forecast with probability p, it should occur with the same frequency p); reliability is zero only if the reliability curve lies along the diagonal;
- *resolution* tells how informative the probabilistic forecast is; it varies from zero, for a system for which all forecasted probabilities verify with the same frequency of occurrence, to the sample uncertainty, for a system for which the frequency of verifying occurrences takes only values 0 or 100% (such a system resolves the forecast between occurring and non-occurring events perfectly);

From these components, skill scores can be derived:

- the Brier Skill Score (BSS,) is computed by reference to the BS of a probabilistic forecast that would consistently forecast the climate distribution:

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

- the Resolution Skill Score (BSS_RSL) is the ratio of resolution by uncertainty;
- the Reliability Skill Score is defined as:

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

these scores are shown in labels of reliability diagrams (Figure 9, Figure 28) and time series of the Brier Skill Scores can be found in Figure 8 and Figure 28.

Relative Operating Characteristics (ROC) curves show how much signal can be picked up 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. If one is more sensitive to the number of hits, the forecast will be issued even if a relatively small number of members forecast the event. If one is more sensitive too false alarms, one will wait for a large proportion of members to forecast the event. The ROC curve shows the false alarm and hit rates associated to different thresholds (proportion of members, or probabilities) used before the forecast will be issued. Such curves are shown in Figure 10, Figure 29, Figure 36 and Figure 37. A good system should have most points clustering near the upper left portion of the diagram - in each diagram, there is an indication of the area under the curve that can be used as a skill score.

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).

In Figure 22, the precipitation data have been collected from the full Northern Extratropics (north of 30°N) both from the model (18-42h forecast accumulation) and SYNOP observations (06-06UTC). The empirical distributions have then been obtained simply by ordering the data from the smallest to the largest amount. The stepwise signature for observations is simply a feature from observations that usually only report full integer values (e.g. 10, 11 or 12mm). The Frequency Bias Index also shown on this Figure is the ratio of the total number of forecasts to observations exceeding a given threshold.

Annex B: The development and testing of upgrades to the operational forecast system

It has been the working practice at ECMWF for many years to implement upgrades to the forecast system one or more times each year. This has enabled a consistent improvement in forecast skill and forecast products to be maintained over two decades. A summary of the main content, impact and pre-operational testing of the upgrades made since 1991 are presented each year in the SAC/TAC paper entitled Performance of the Operational Forecasting System, Part C: The impact of upgrades.

B.1 Research testing

In order to achieve this, it has been necessary to combine an often substantial number of changes in a new IFS cycle. Sometimes it is possible to package together changes involving only the forecast model, for example, but usually an IFS upgrade will contain a mixture of assimilation, model, data usage and technical revisions which will vary significantly in their impact on the analysis and forcast quality. The following discussion summarizes the methodology used to handle this rather complex and sometimes difficult pre-operational process. To virtually guarantee the quality of a set of meteorologically influential changes, it would be necessary to test each component of these changes over several years of assimilation and forecasts and to then repeat such extensive testing with combinations of changes. Such a factorial testing environment is clearly not practicable for any forecast centre and hence the following process has evolved as a judicious compromise.

All individual changes are tested by their initiator before being provided for combination and, unless they have negligible meteorological impact, will have been tested in full assimilation and forecast mode at operational resolution. Prior to this, individual changes will also have been tested in a variety of lower resolution assimilations and/or forecasts as part of their development. Sometimes individual changes will be grouped together earlier, based on scientific grounds, examples might include several changes all involving the incremental 4D-Var algorithm, or a group of interactive parametrization changes such as to radiation, cloud and convection schemes. Scientific discretion is also used to anticipate when individual changes from different components of the system have the potential for marked interactions. In such cases, efforts are made to combine these as early as possible, for example, a parametrization change that affects the surface winds and an ocean wave model revision.

Once the full combination of changes has been created, a rigorous series of tests is performed prior to the commencement of the formal pre-operational trials.

Simmons and Hollingsworth (2001) discuss, inter alia, the efficacy of recent changes and their impact on analyses and short-range forecasts. They note that examination of 1-day RMS. errors is a powerful measure of improvements and that this confirms that the current process for implementing changes produces consistent year-on-year improvements in the ECMWF forecast skill.

B.2 Pre-operational testing and the implementation process

Before preparing and implementing changes to the forecasting system, consideration is given to which users, i.e. Member States and Co-operating States, ECMWF internal users or NMHSs of the WMO and also the public, will be affected by the change and to what extent. The web-based information system has become a major communication system for the provision of real-time products from the Centre and has broadened the

global use of the products. This community, which includes users in the Member States, Co-operating States and beyond, also needs to be considered when planning the implementation of changes.

The Centre's 4-year-planning procedure outlines the planned major upgrades of the forecasting system. Member States and Co-operating States are kept informed of any impact that the changes may have on their own use of the products, i.e.

- expected changes in the meteorological characteristics of model results
- technical implications, such as GRIB changes, new parameters, change in the definition of parameters
- changes in the run time of the forecasting system.

Major upgrades in model resolution, for example, are notified with a substantial lead-time of several weeks to months.

When the Research Department has completed the testing of the upgrades of the forecasting system, the changes are handed over to Operations for e-suite testing. This consists of running the new cycle in parallel with the established operational cycle for several weeks or months. On commencing the e-suite, a schedule for the implementation of the changes is drawn up, together with a work plan depending on the nature of the changes. The work plan includes

(i) E-suite testing requirements

All components of the forecasting system are tested, particular attention is given to the components which are affected by the change, e.g. data assimilation, model, EPS, wave model, etc.

(ii) Product generation tests and archiving

This is also used for providing test data to Member States users.

(iii) Evaluation

The routine verification package is run for the e-suite. Depending on the nature of the change, the evaluation is focussed on the overall performance, the weather parameters, EPS performance, etc.

(iv) Visual inspection of all products

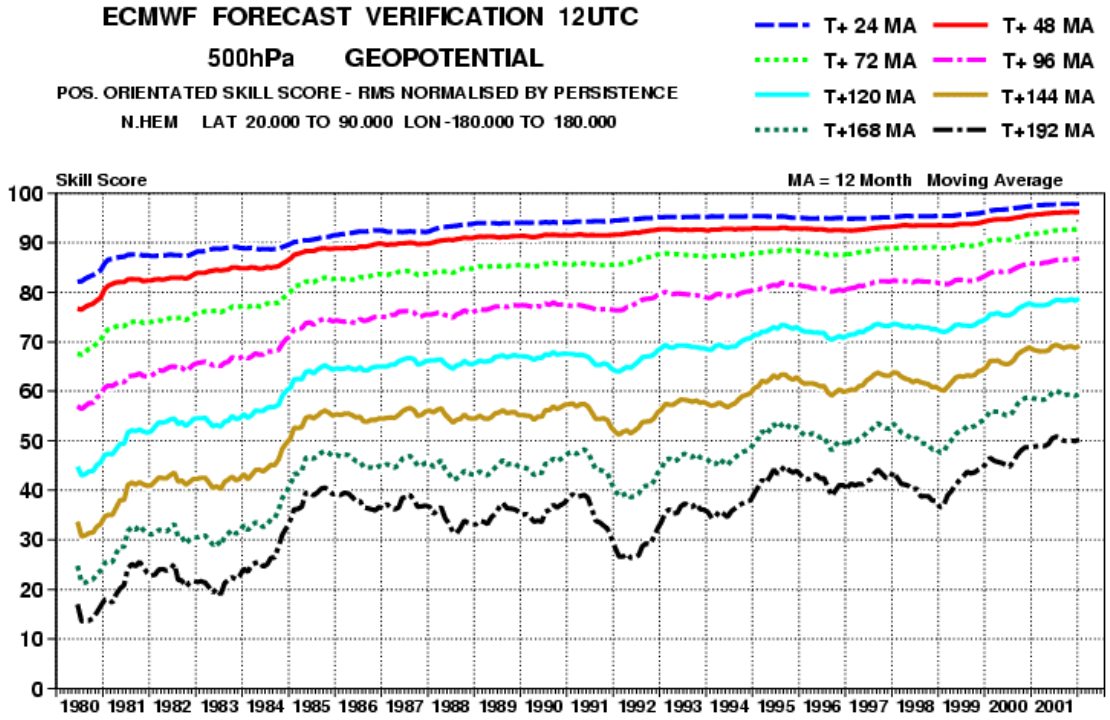
The product generation tests are complemented by visual inspections of the products generated for the Met. Ops. Room, the web and the dissemination.

(v) Synoptic studies

The objective verification is complemented by synoptic studies of particular events, e.g. heavy precipitation, cyclone development, tropical cyclone tracks, etc.

Member States and Co-operating States are provided with early access to e-suite results in MARS and additional test products through the dissemination on demand.

Operations and Research staff jointly monitor the progress towards the implementation of upgrades to the forecasting system and ensure that the appropriate actions are taken, according to the implementation schedule. Since the e-suite is restricted to a given season i.e. that preceding or containing the implementation date, the E-suite results are supplemented by previous research testing. This combined information gives as reliable a measure as possible of the meteorological impact of a new cycle and is used by the two Heads of Department in making the final joint decision on the implementation.



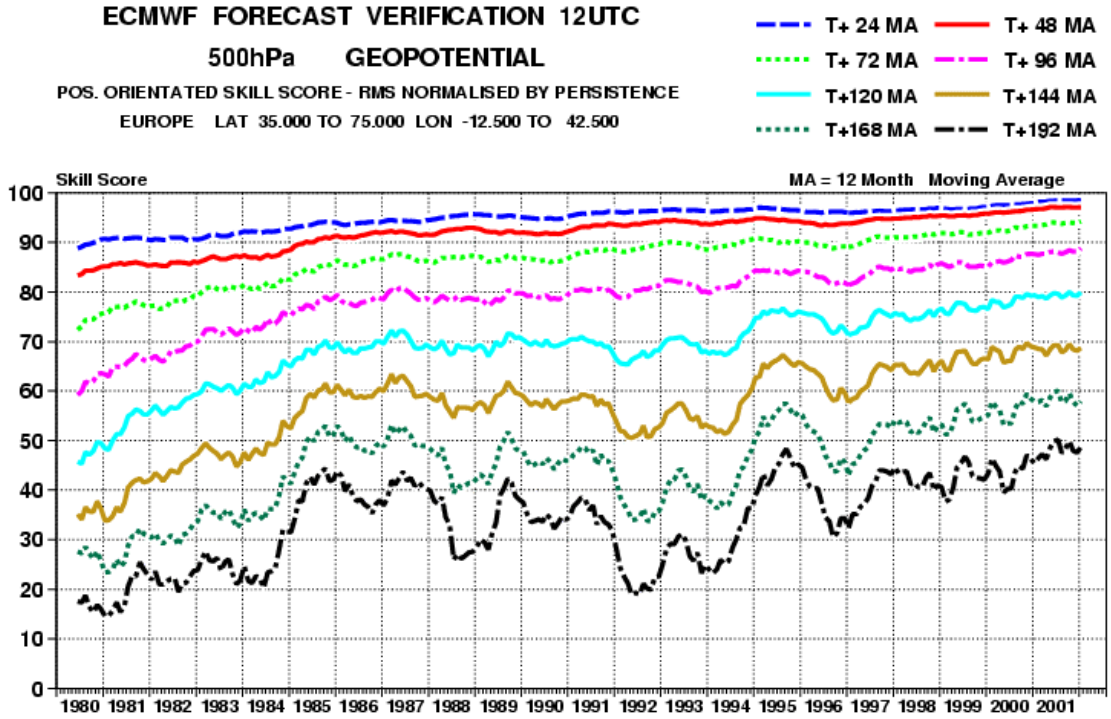
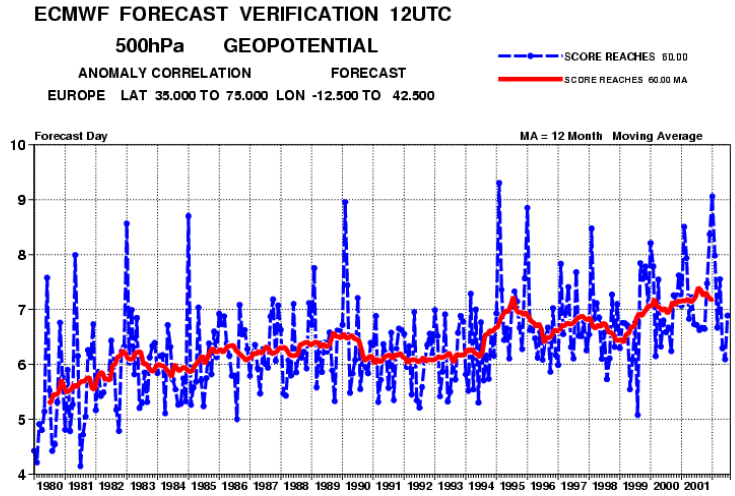
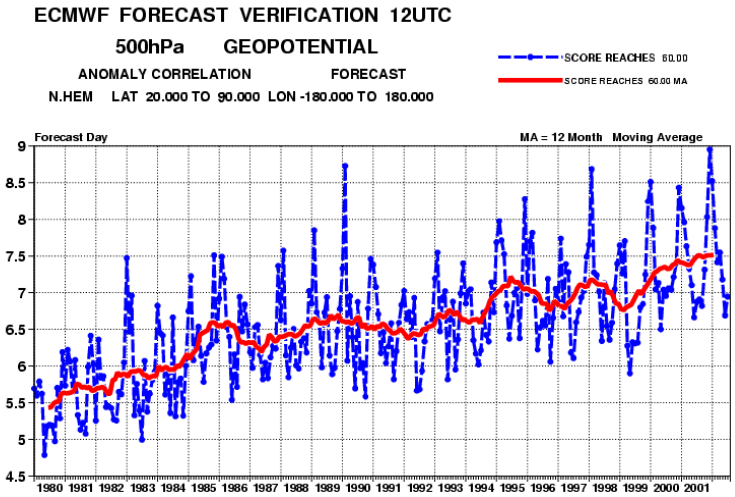


Figure 1: 500-hPa height skill score (N. Hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)





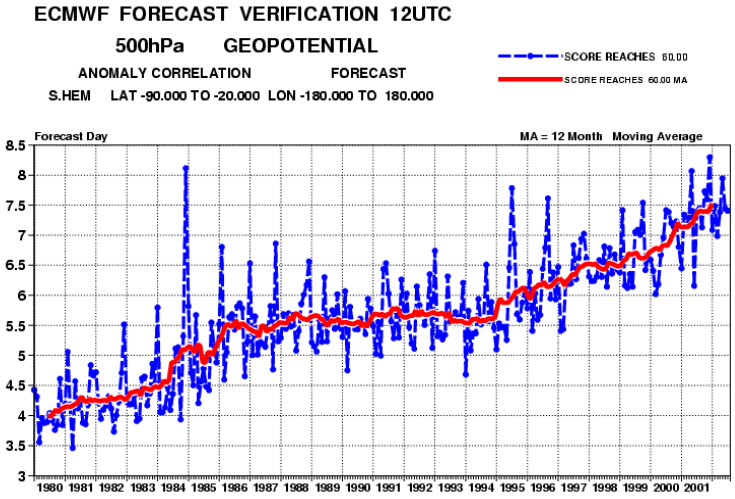


Figure 2: 500-hPa height monthly and annual means of the forecast range when the anomaly correlation is falling below 60% for Europe, Northern and Southern Extratropics

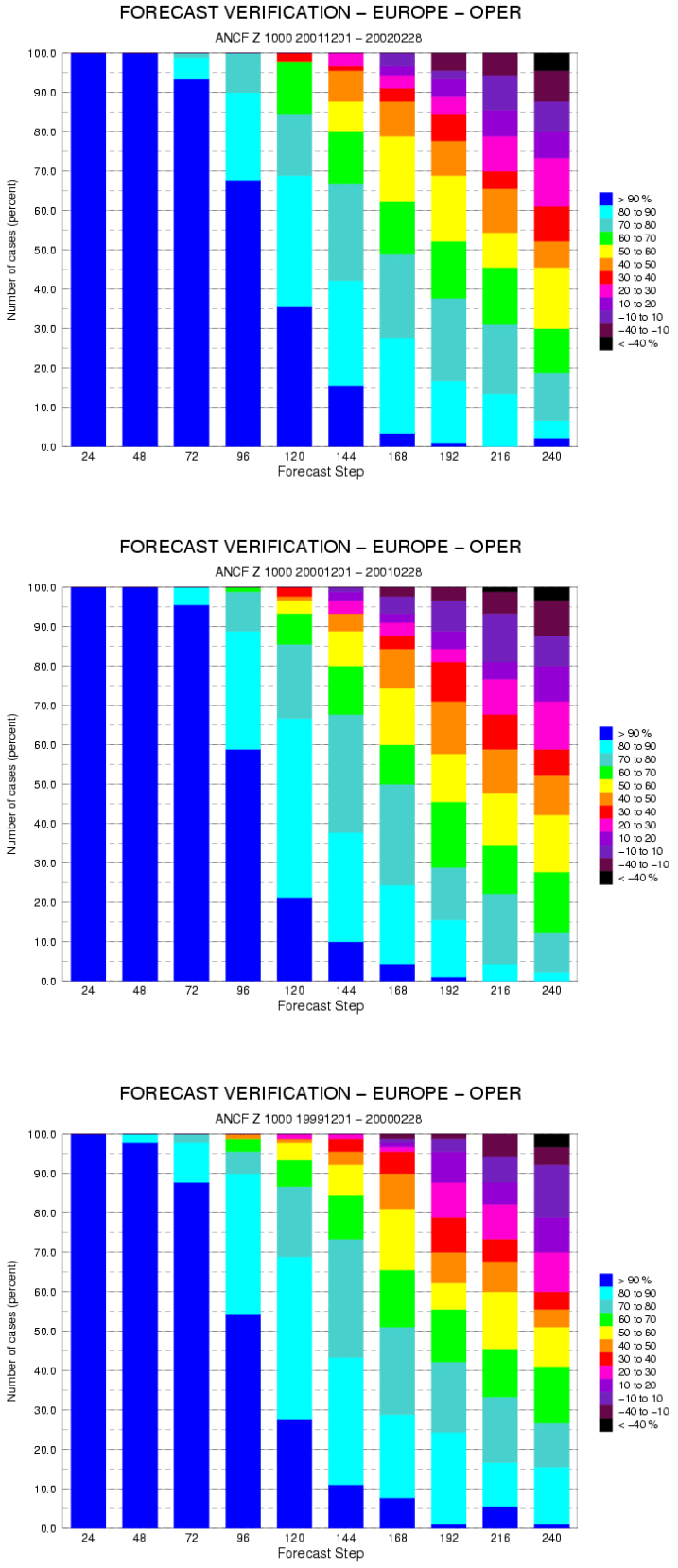
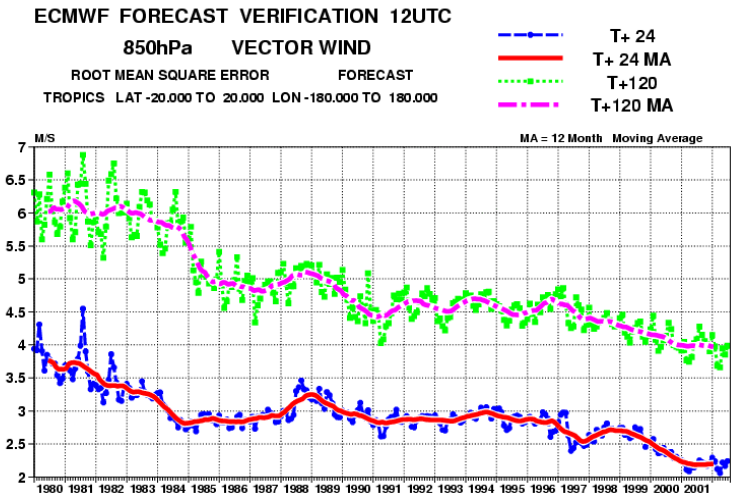
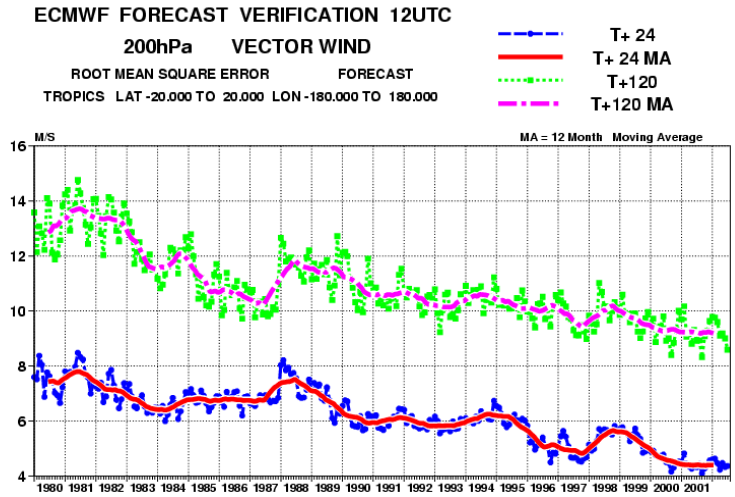


Figure 3: Cumulative frequency distribution - 1000-hPa Anomaly Correlation over Europe, last three winters: from top to bottom, DJF 2001-02, 2000-2001 and 1999-2000





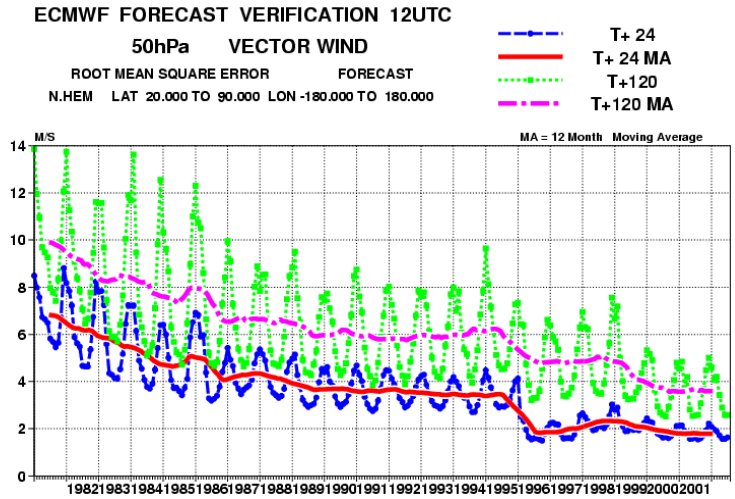


Figure 4: Model scores in the Tropics (root mean square errors against 850hPa and 200hPa wind analysis) and Northern Hemisphere stratosphere (same score, level 50hPa);

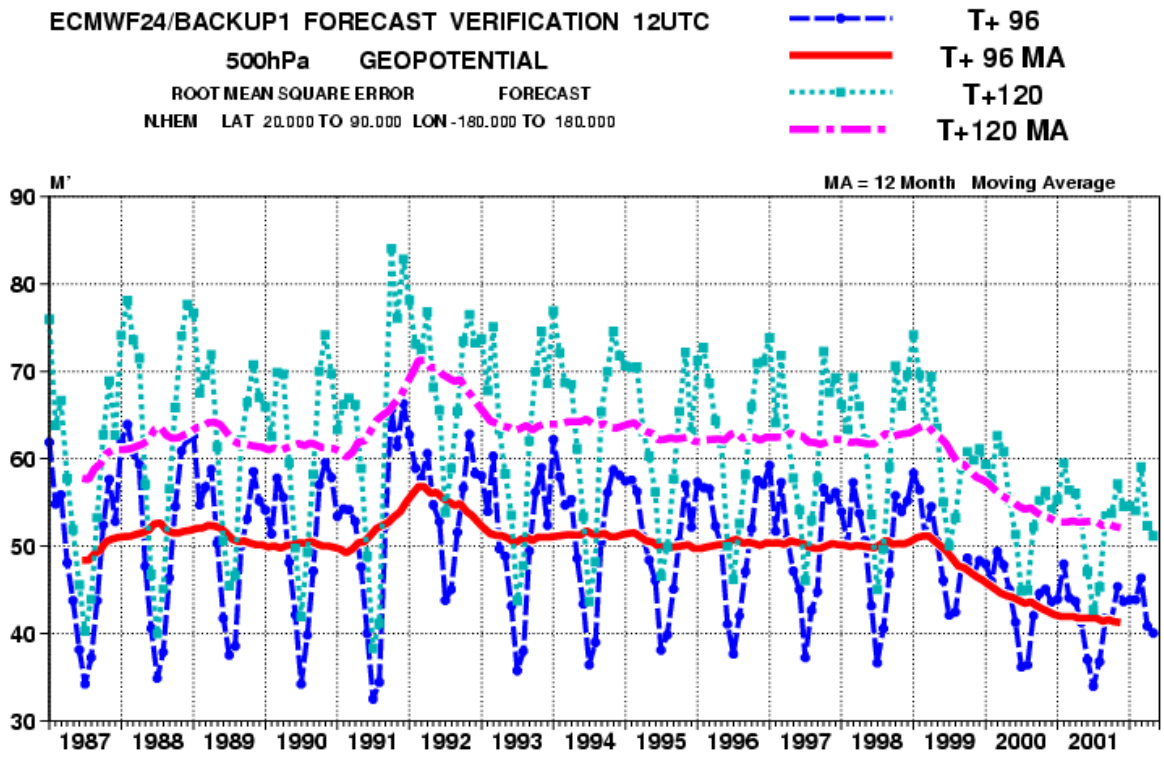


Figure 5: RMS of the difference between 24h-consecutive 500-hPa height forecasts verifying the same day over Northern Extratropics.

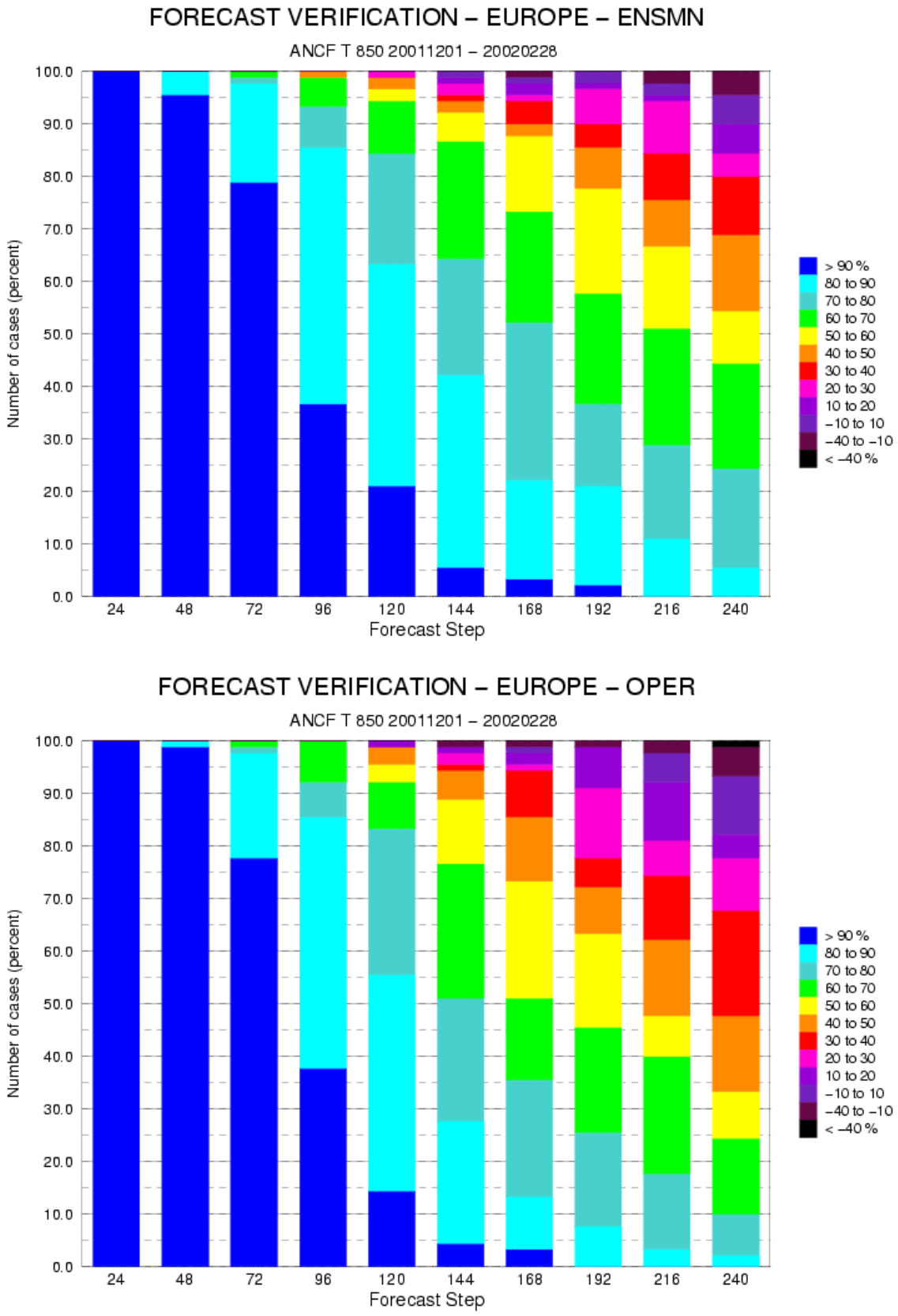


Figure 6: Distribution of T850 anomaly correlation scores over Europe this winter (DJF) for the EPS Ensemble Mean (top) and T511 forecast (bottom)

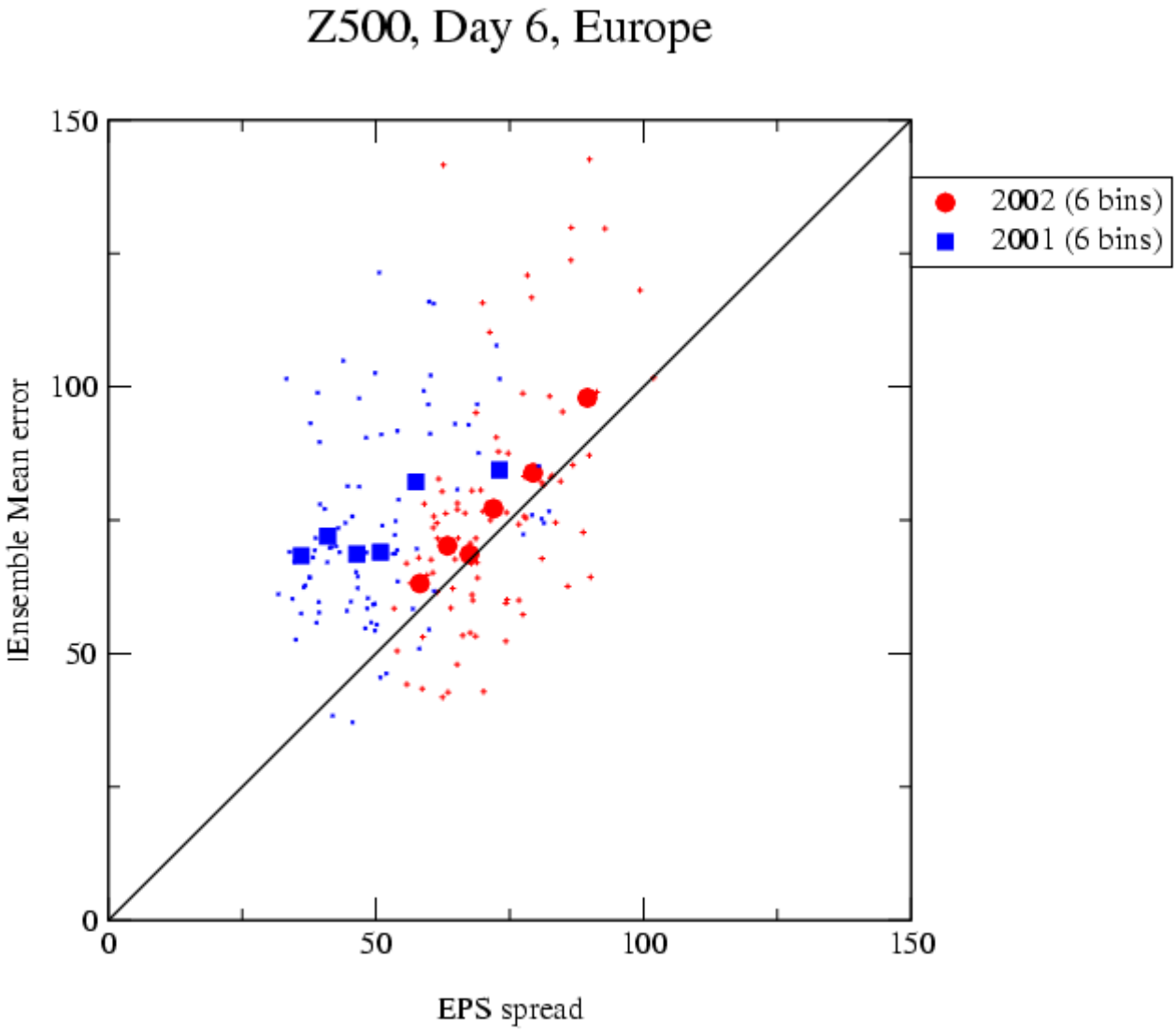
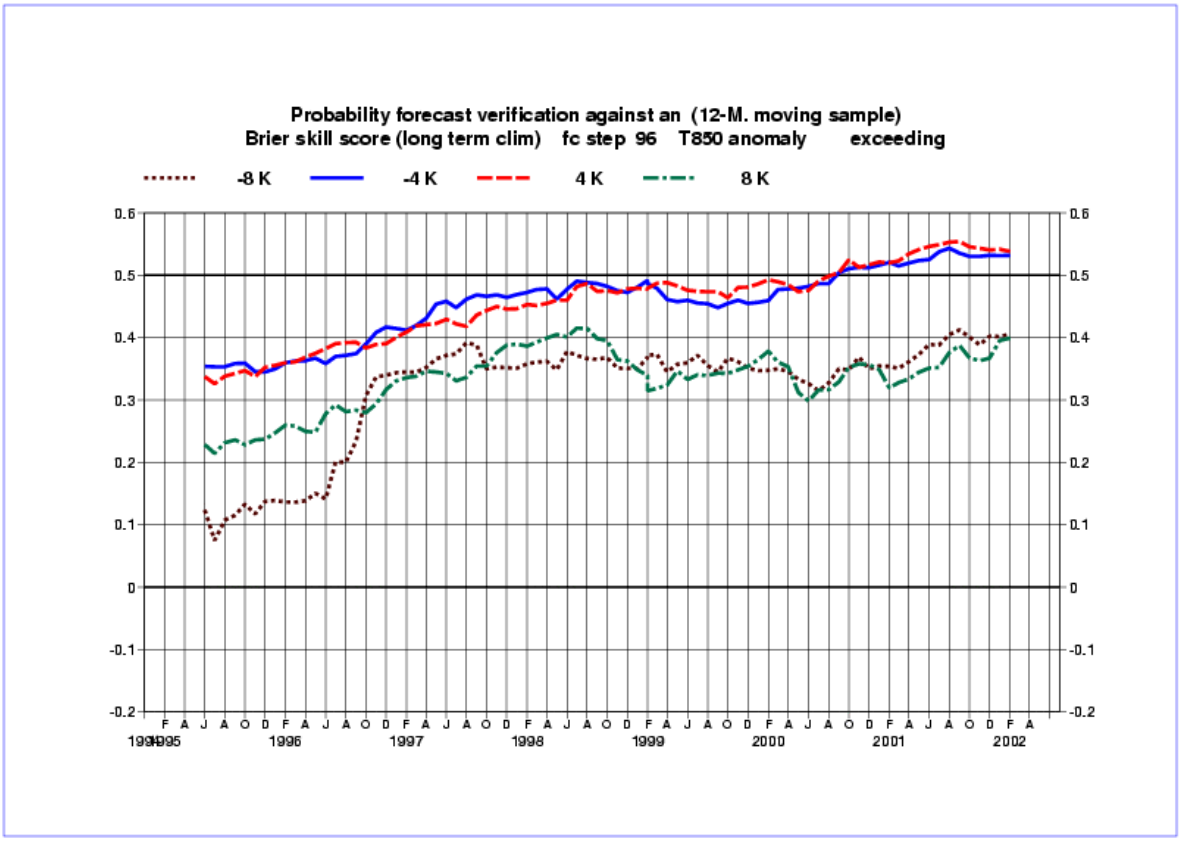


Figure 7: EPS Spread-skill scatter diagram: small dots are daily forecasts, large dots are averaging equally populated bins stratified by increasing EPS spread: they should ideally lie on the diagonal; all forecasts are Day 6 500-hPa height over Europe; blue is winter 2000/01, red in winter 2001/02



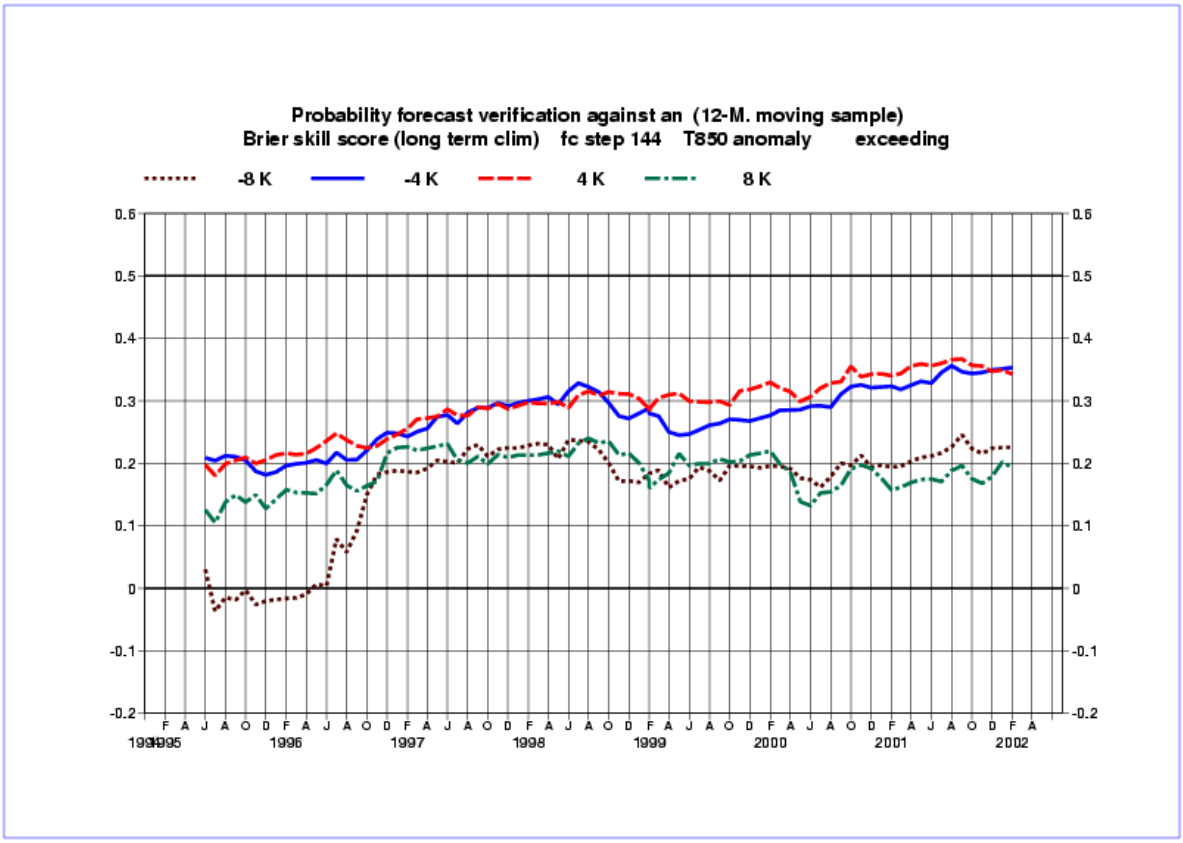


Figure 8: Time series of Brier Skill Scores for EPS 850-hPa temperature anomalies over Europe (12-months moving data samples)

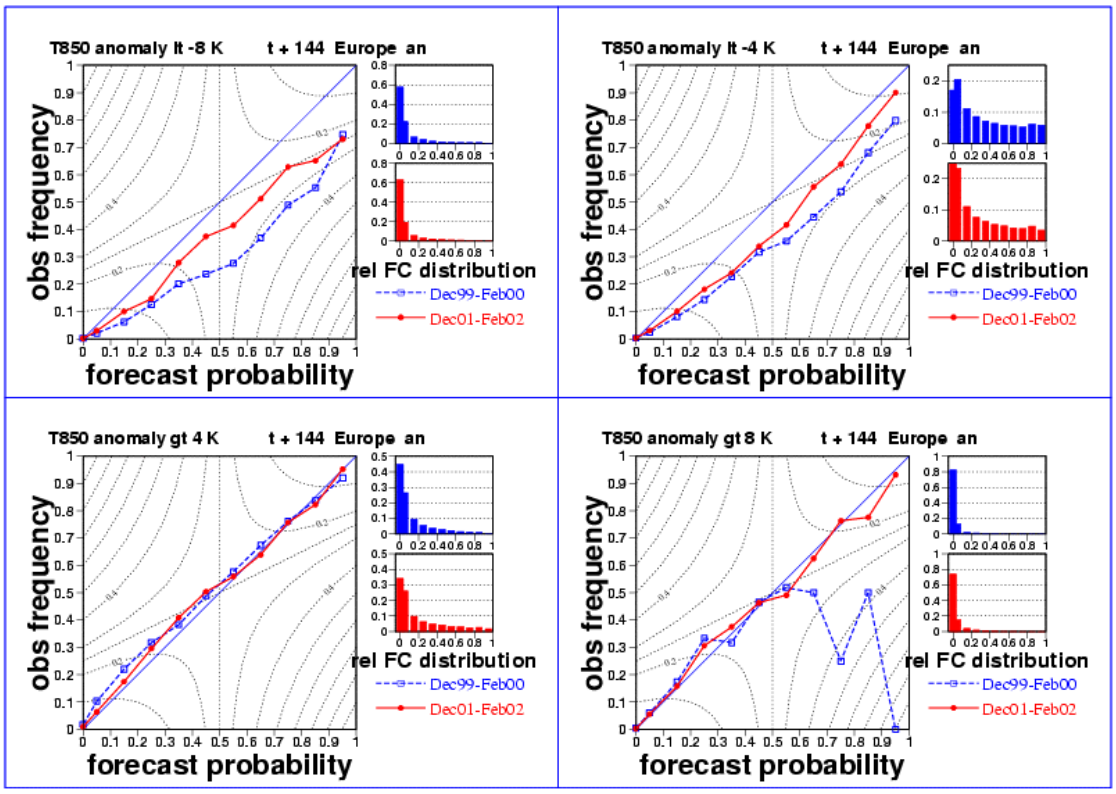


Figure 9: Reliability diagrams for Day 6 EPS forecasts of 850-hPa temperature anomalies over Europe in winter 2001-02 (red) and 1999-2000 (blue); thresholds(clockwise from top left) -8, -4. +8 and +4K

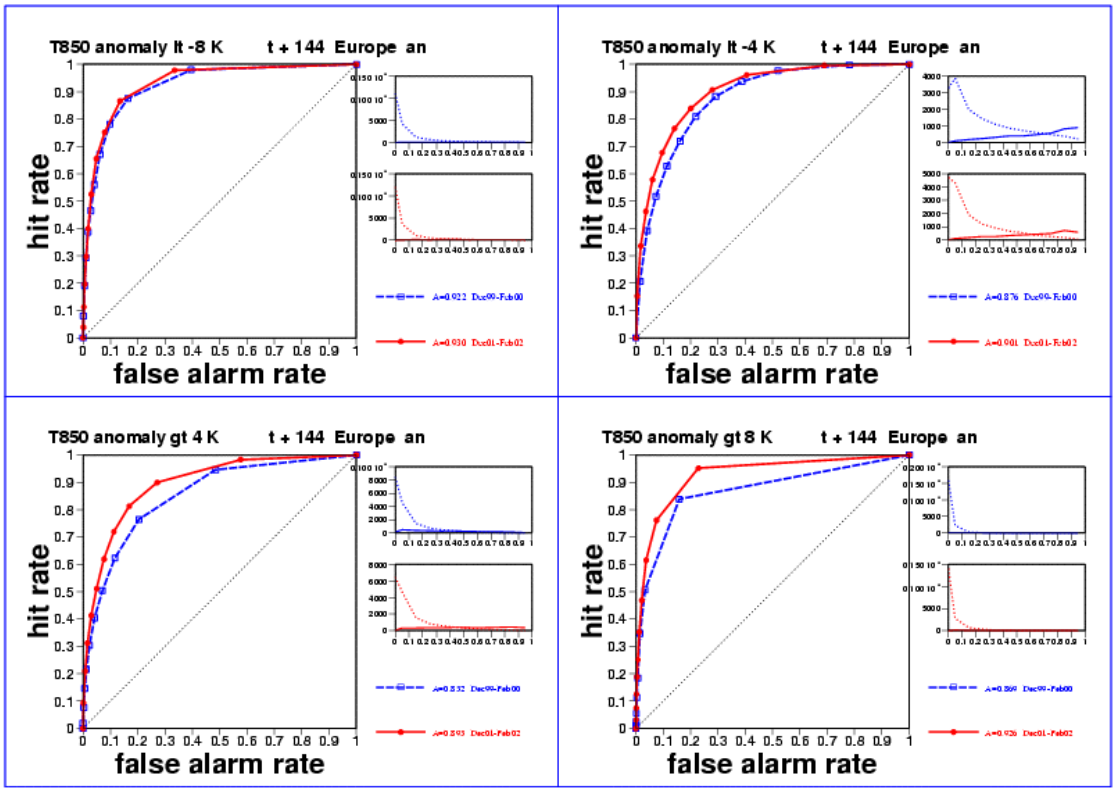
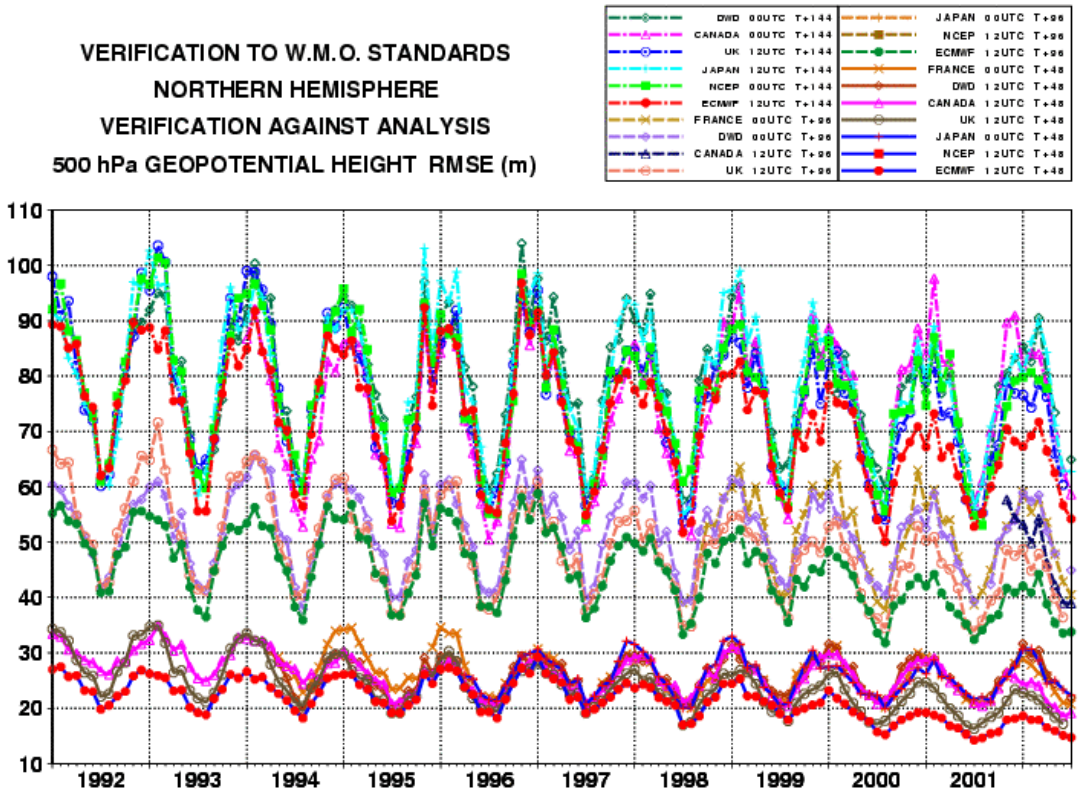


Figure 10: ROC curves for Day 6 EPS forecasts of 850-hPa temperature anomalies over Europe in winter 2001-02 (red) and 1999-2000 (blue); thresholds (clockwise from top lef)t: -8, -4. +8 and +4K



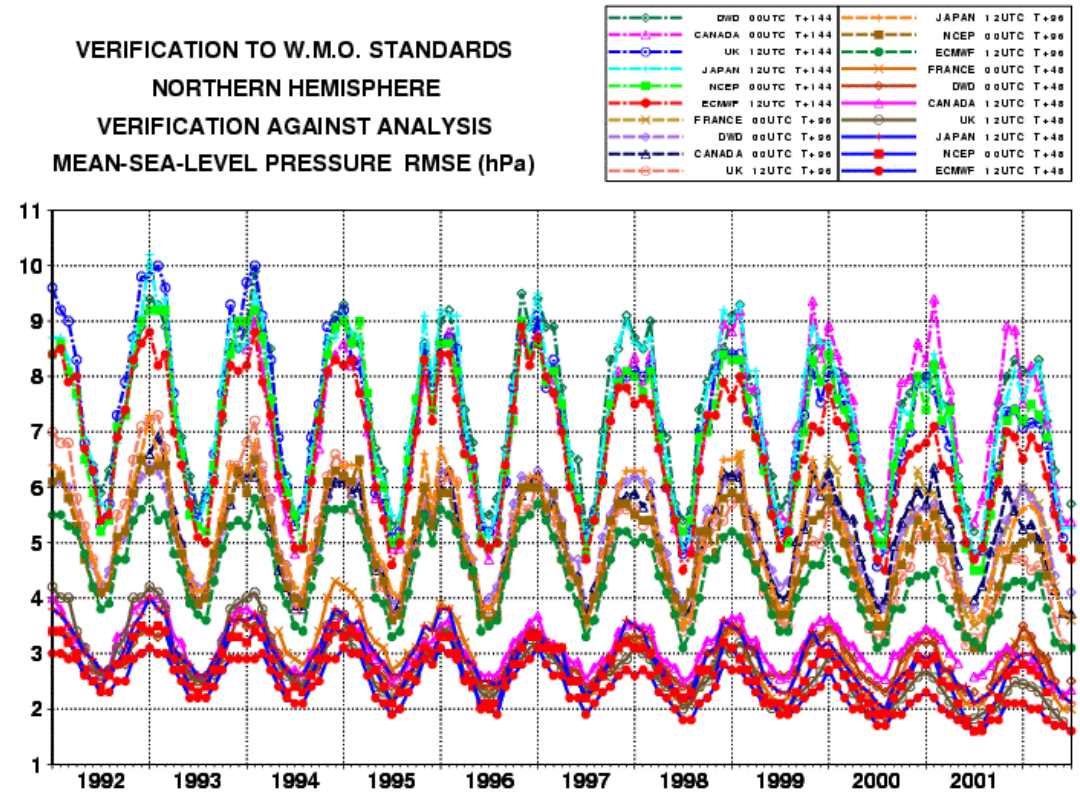
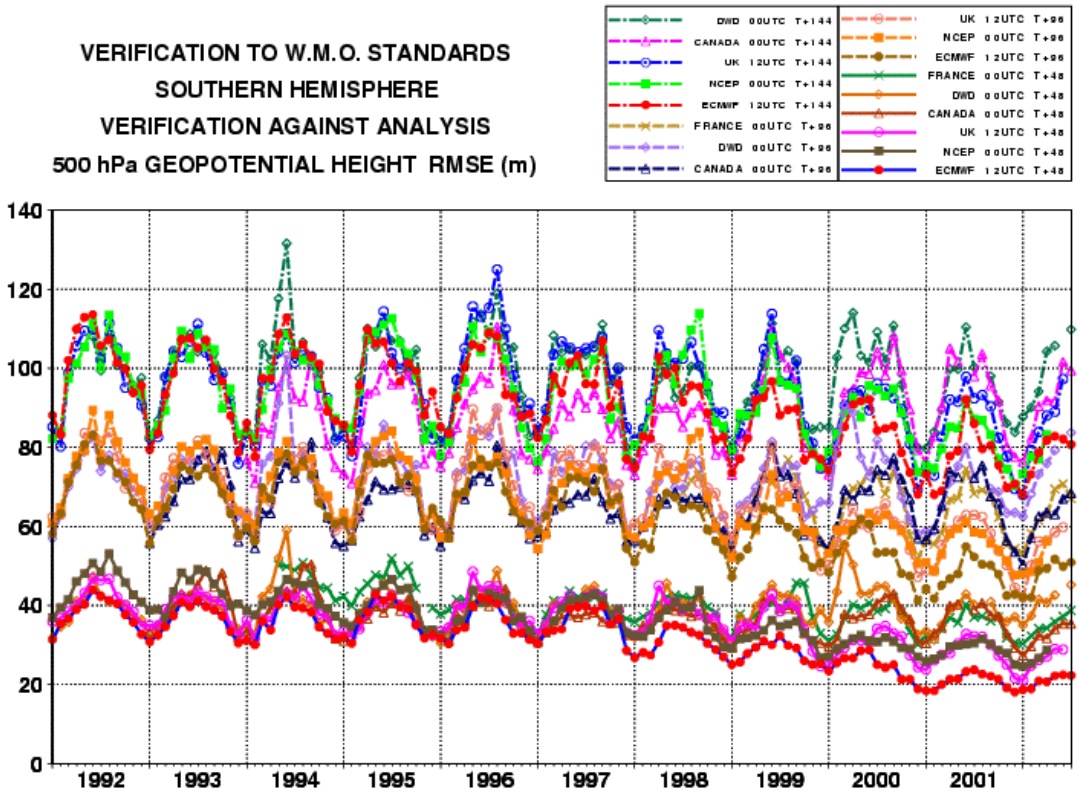


Figure 11: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)



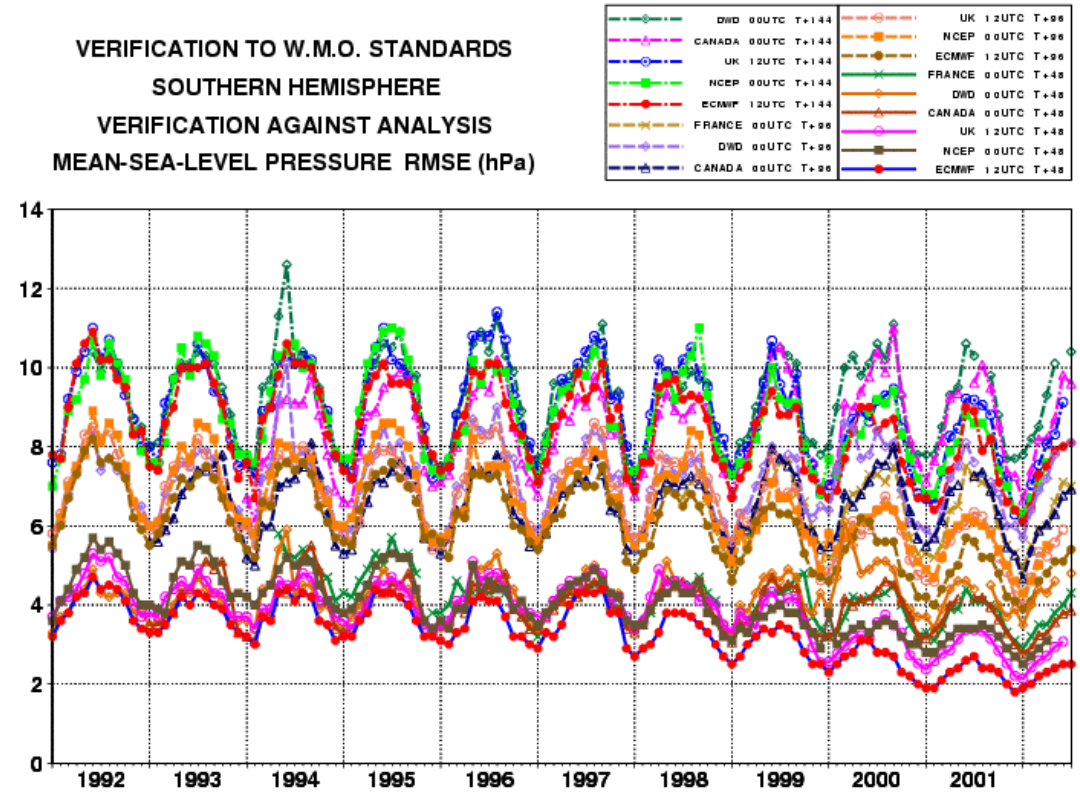


Figure 12: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)

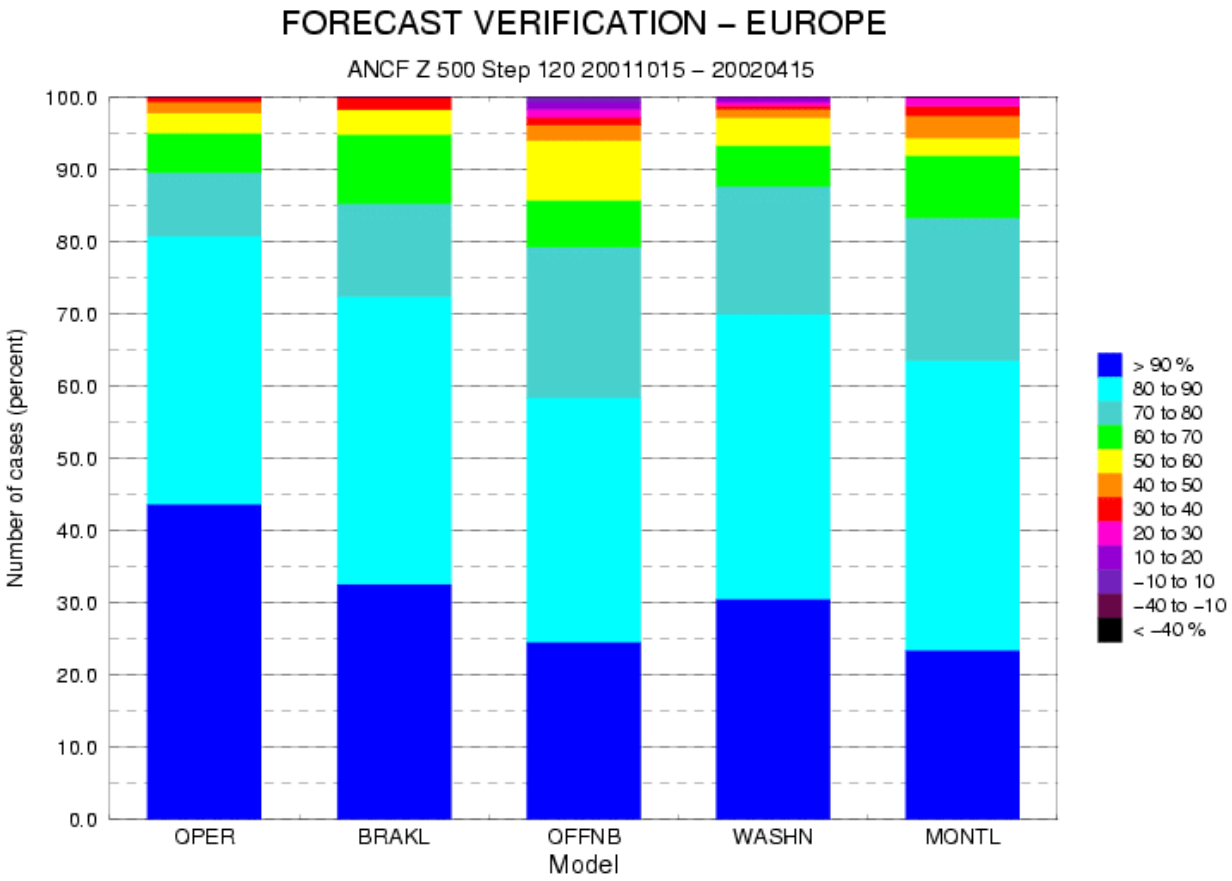
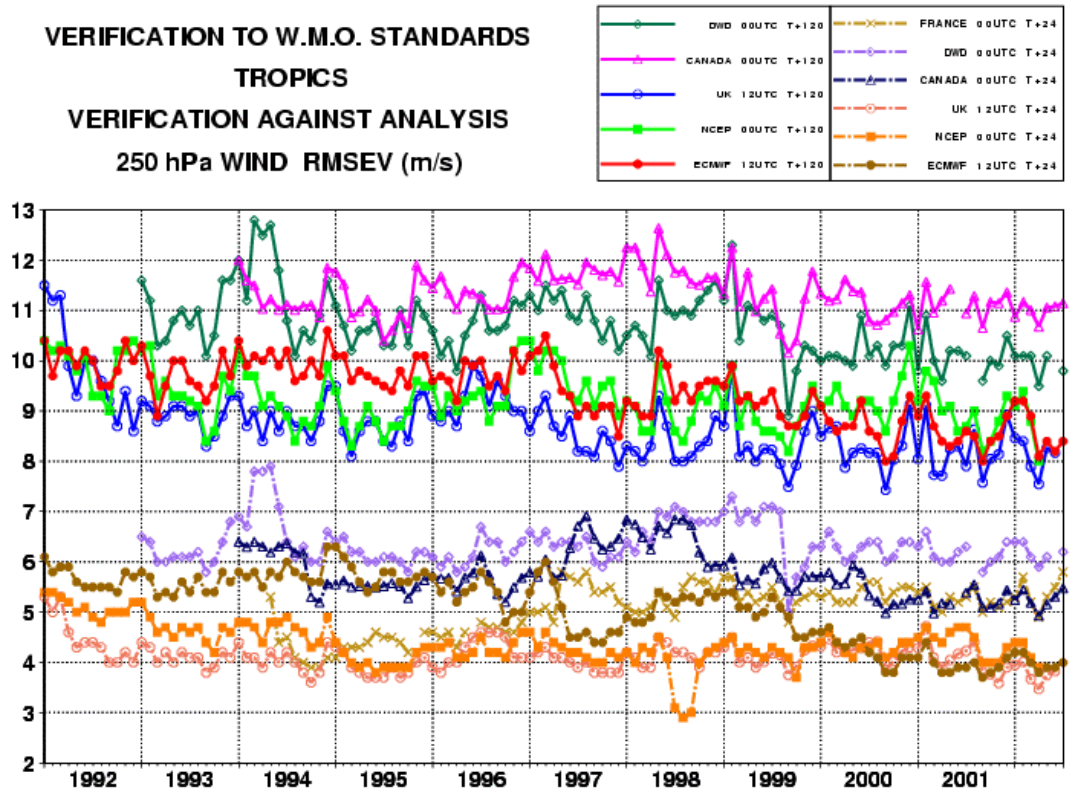


Figure 13: Distribution of anomaly correlation scores (500-hPa height, day 5, Europe) for ECMWF (OPER) and four other NWP centres during the cold season (15 Oct- 15 Apr.)



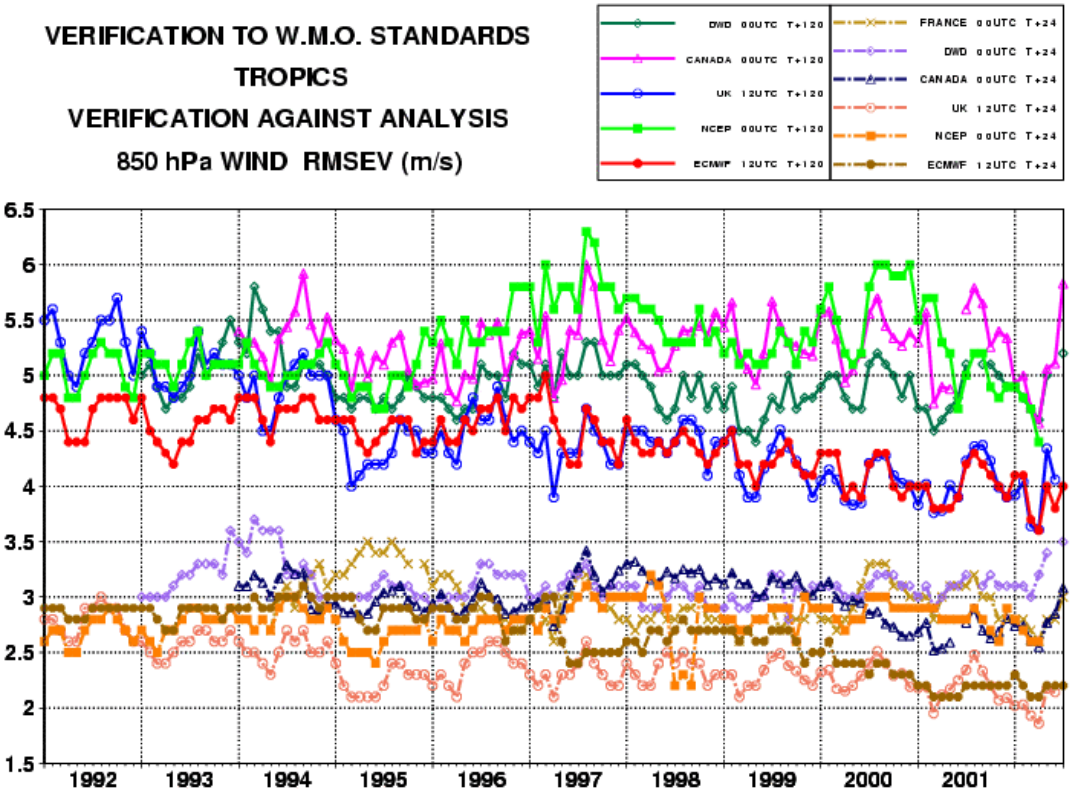
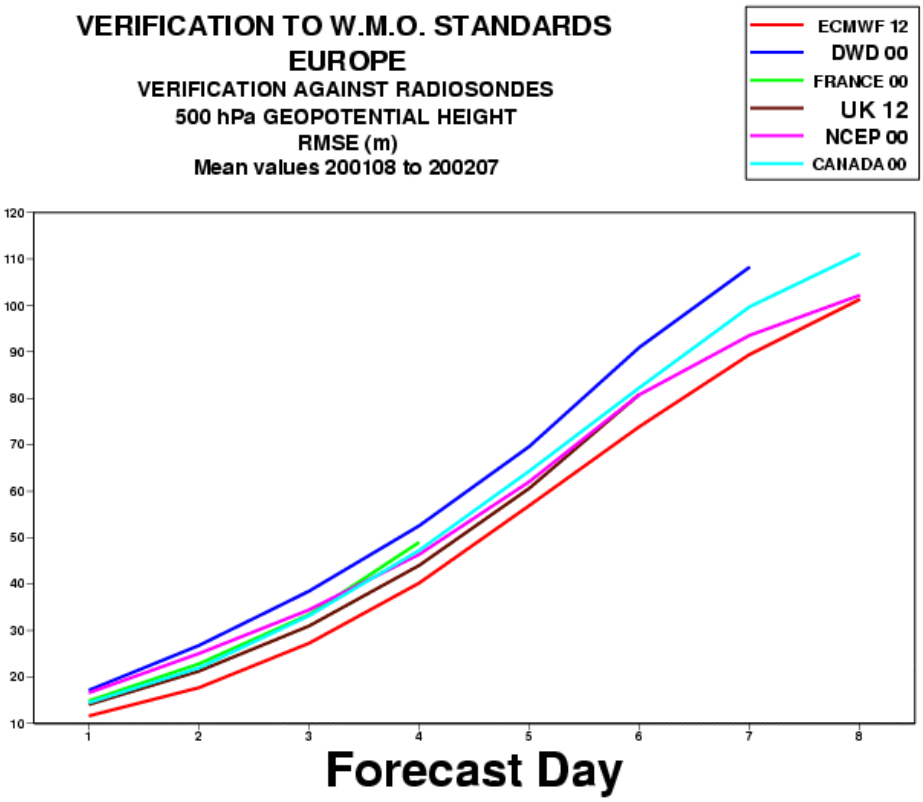


Figure 14: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250-hPa and 850-hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis



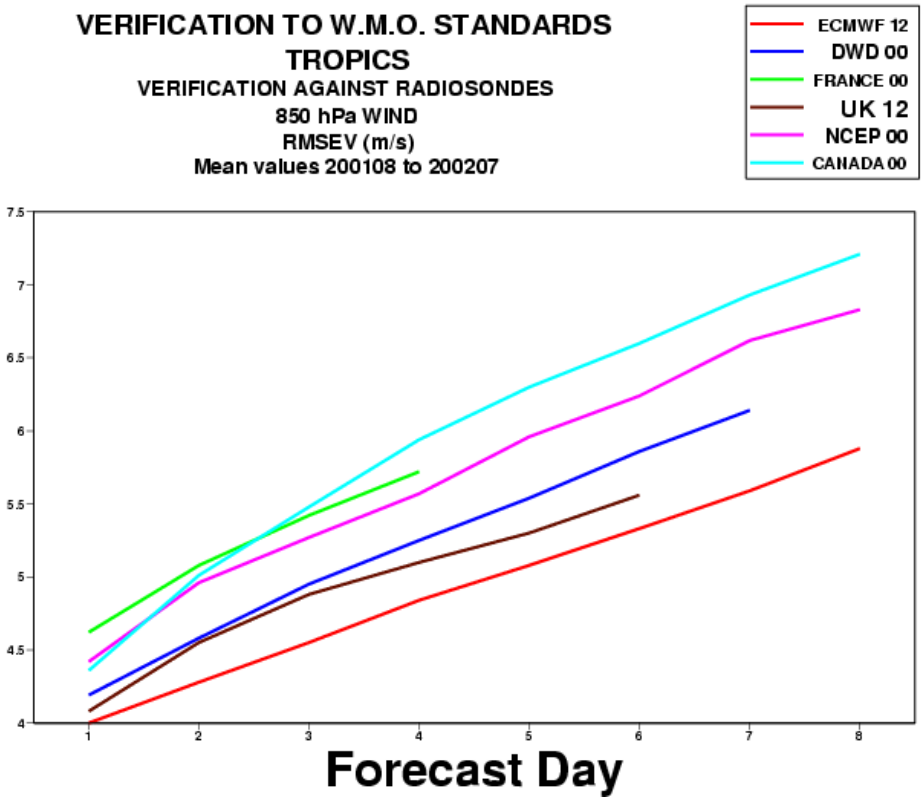


Figure 15: WMO/CBS exchanged scores using radiosondes: 500-hPa RMS error over Europe and 850-hPa wind errors in the Tropics (annual mean)

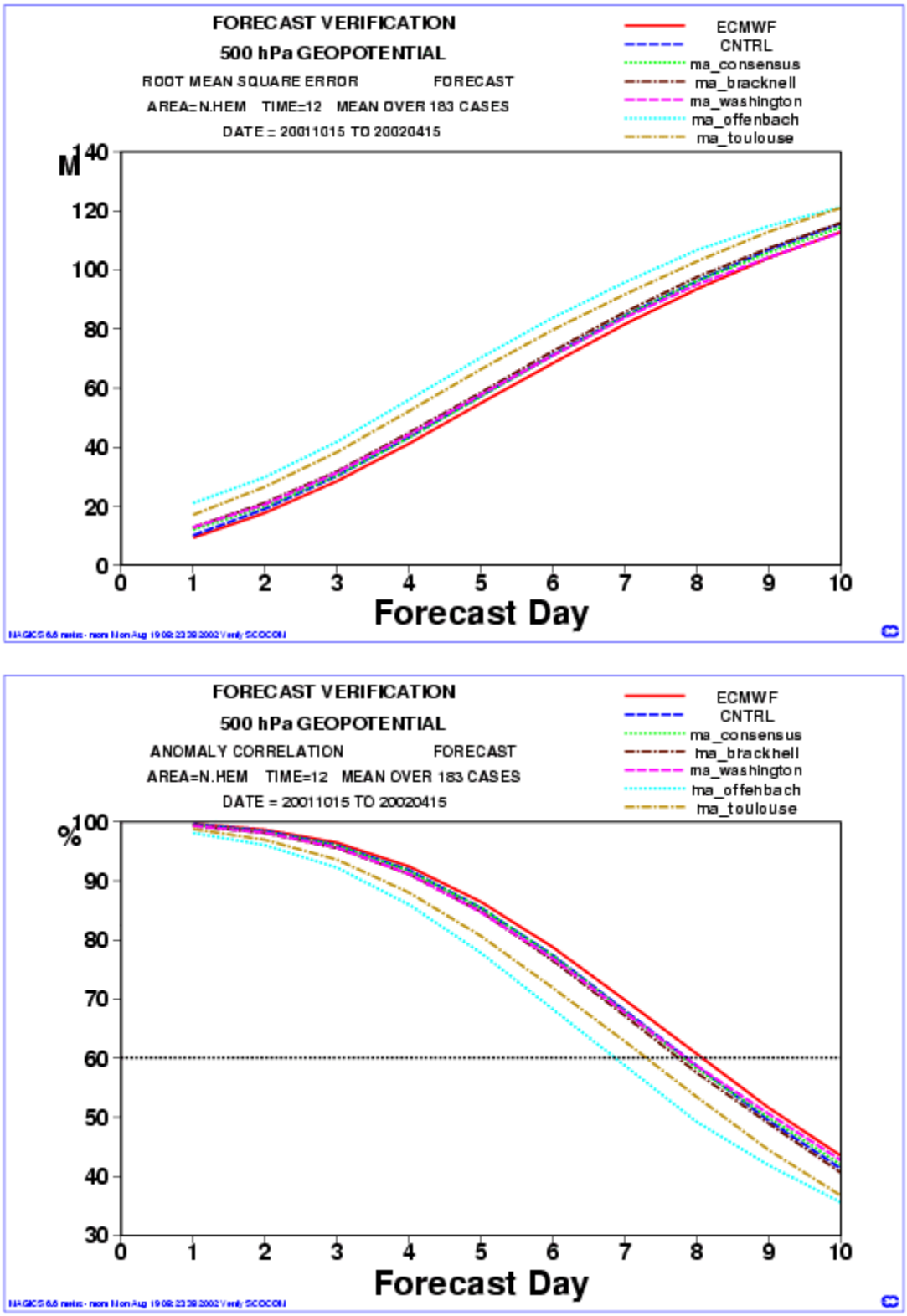


Figure 16: 500-hPa height scores from the multi-analysis system (Northern Extratropics, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution

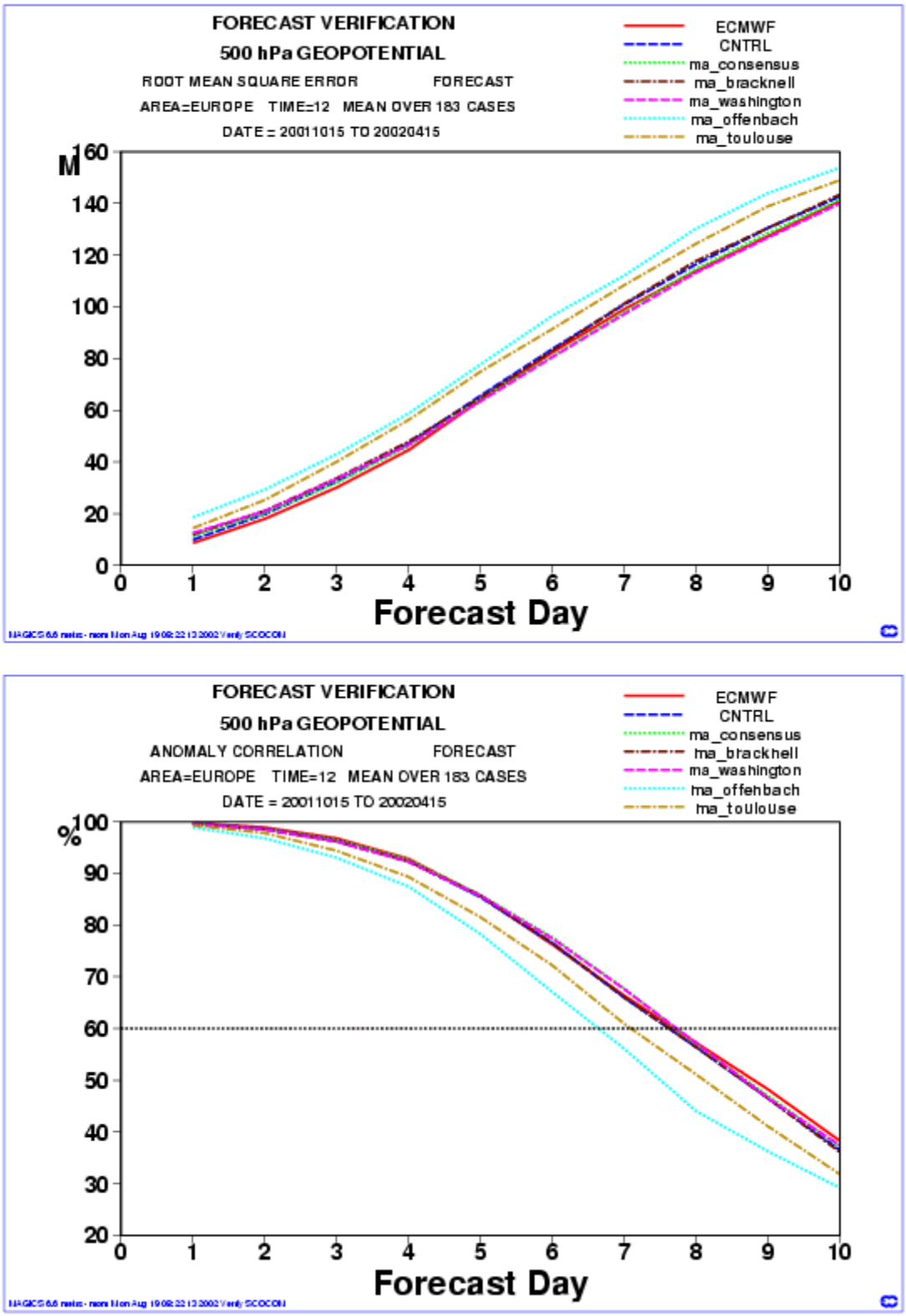


Figure 17: 500-hPa height scores from the multi-analysis system (Europe, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution

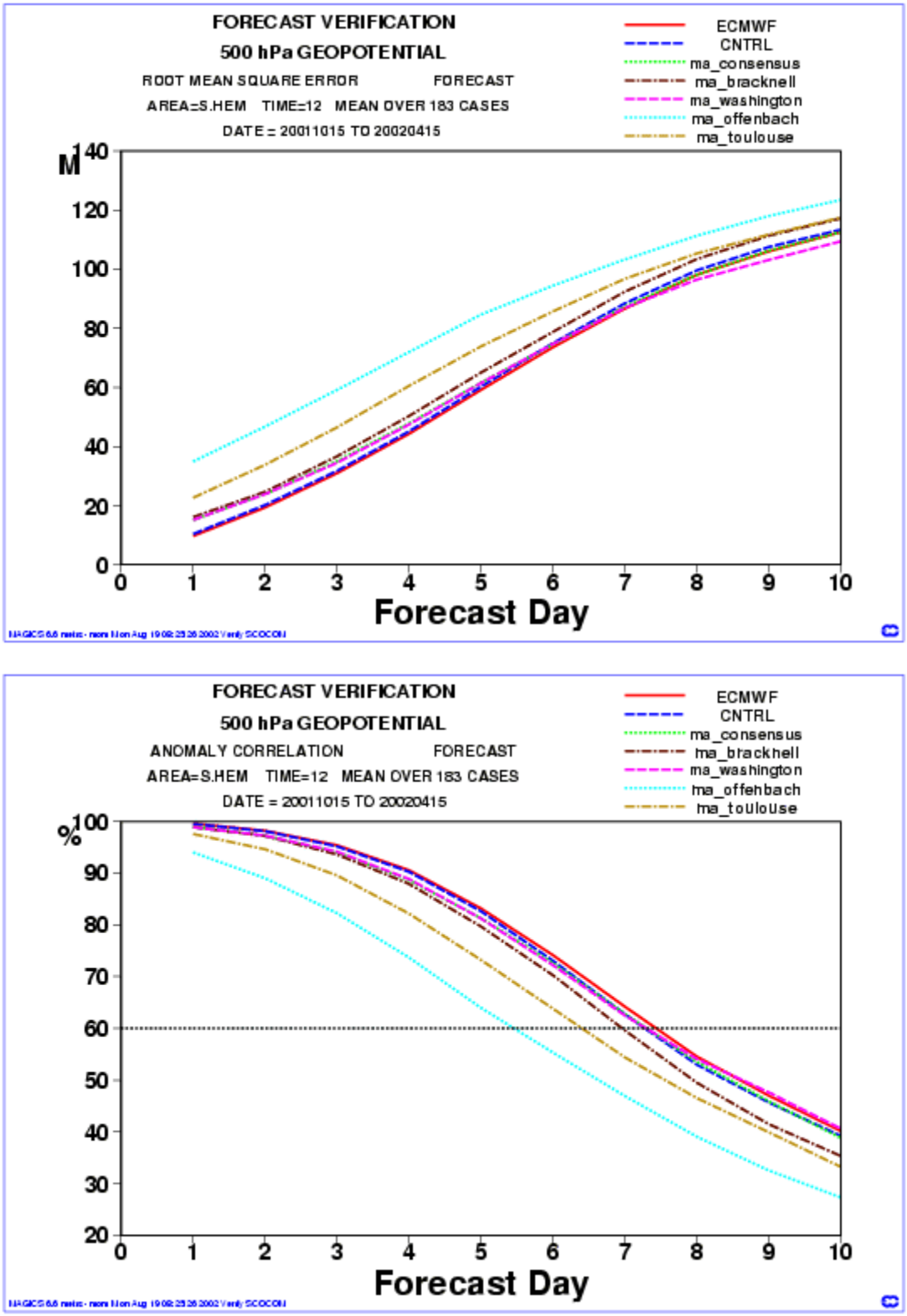


Figure 18: 500-hPa height scores from the multi-analysis system (Southern Extratropics, 15 Oct.-15 Apr.), Upper: RMSE, lower: Anomaly correlation ; both ECMWF T511 and CNTRL T255 forecasts from ECMWF analysis are scored – all other forecasts are run at T255 resolution

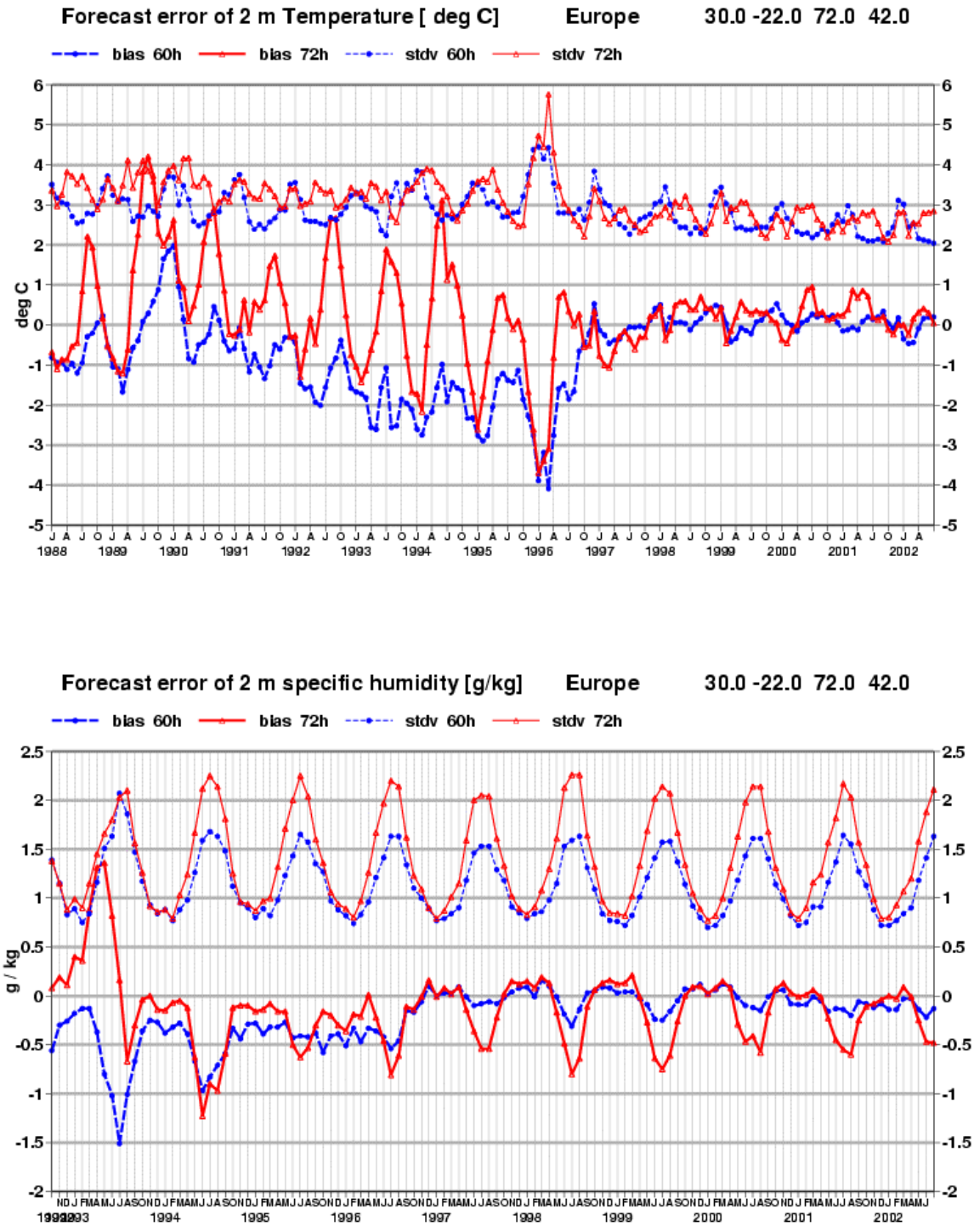


Figure 19: Scores against European SYNOPs of 2m-Temperature and specific humidity forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)

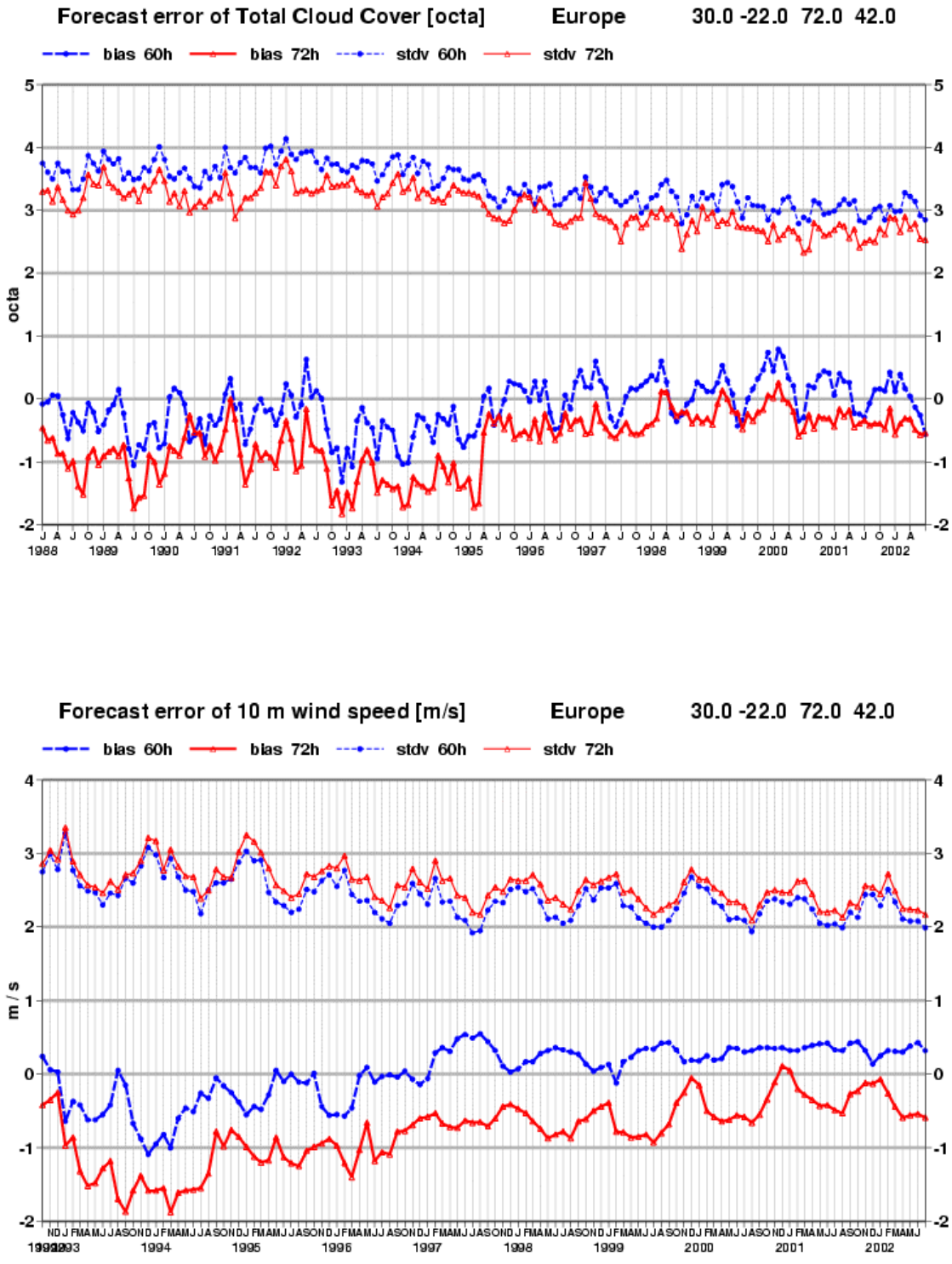


Figure 20: Scores against European SYNOPs of total cloud cover and 10m wind speed forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC).

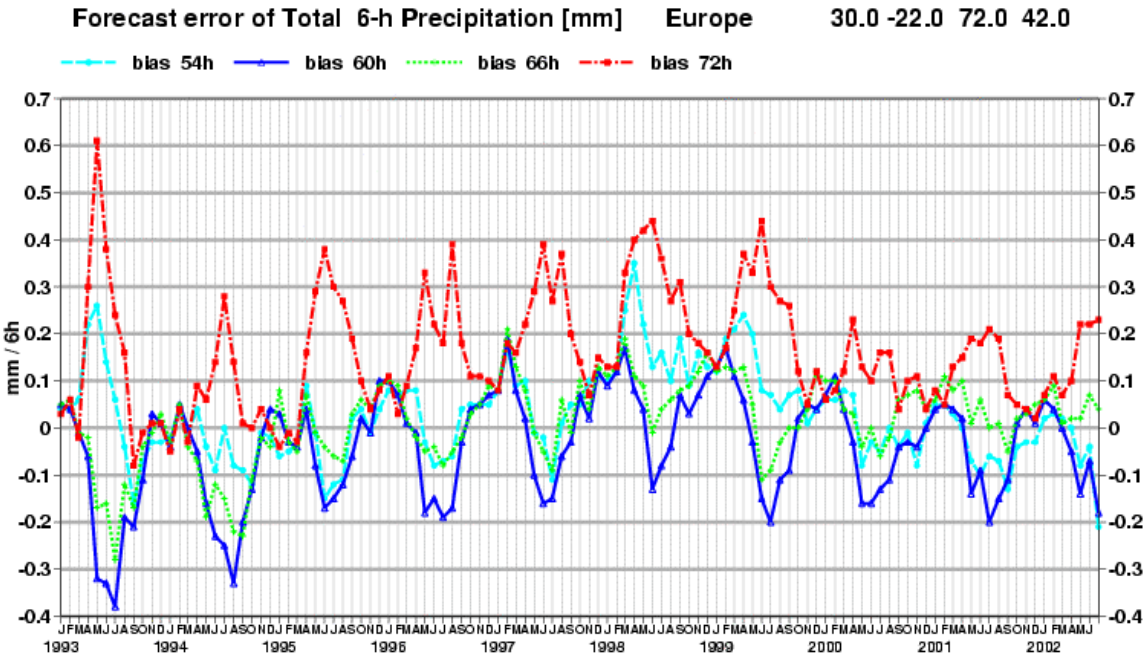
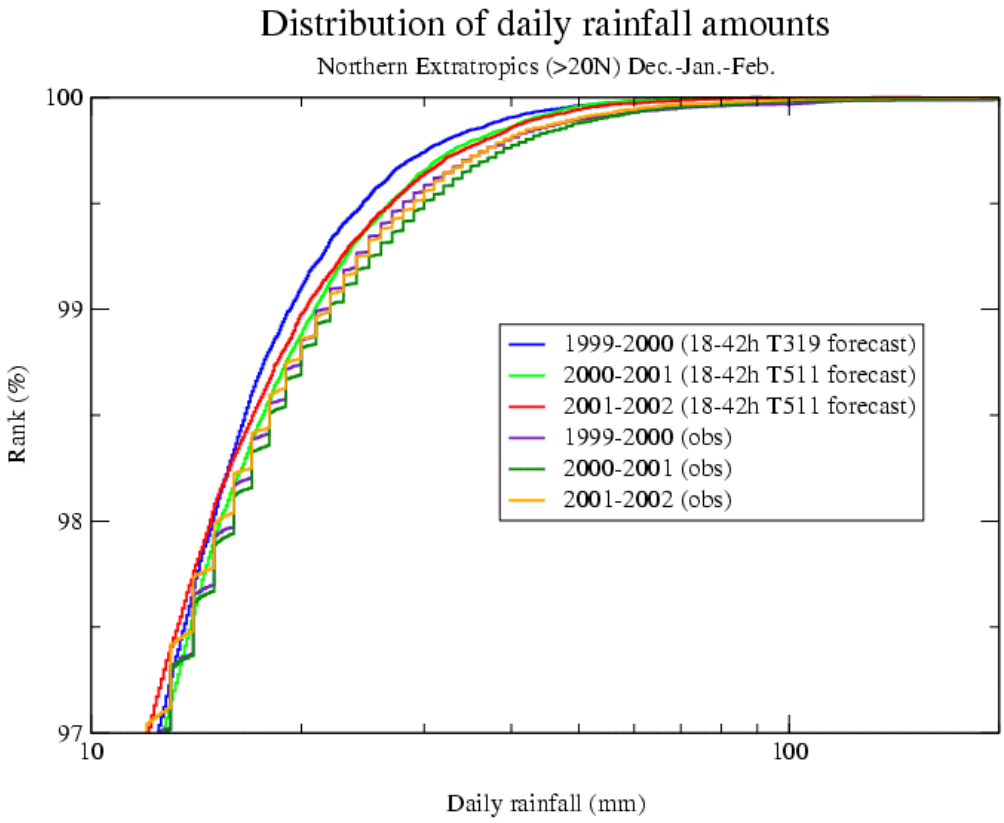


Figure 21: 6h-accumulated precipitation forecasts biases (T+54/60/66/72h) with respect to SYNOP



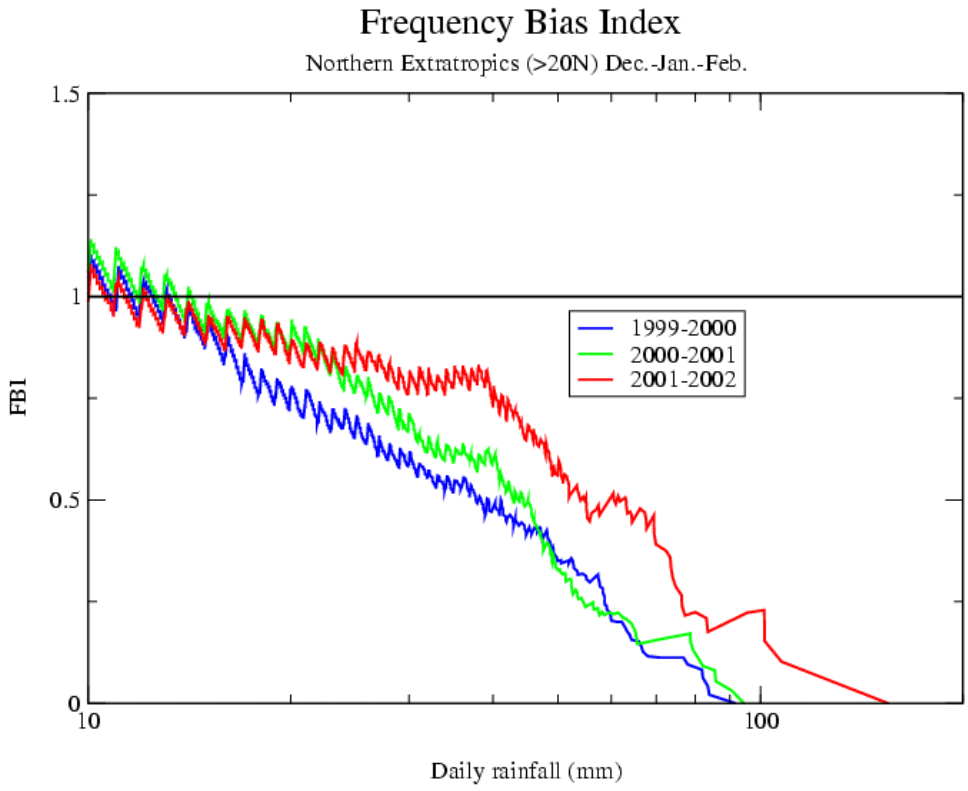


Figure 22: Comparison between model 18-42h forecast and observed (SYNOP) daily rainfall distributions; stepwise curves for observations reflect the discretisation (precipitation is often reported every mm); x-axis is in logarithmic scale beyond 10mm.Top: ranked distributions; bottom: bias index

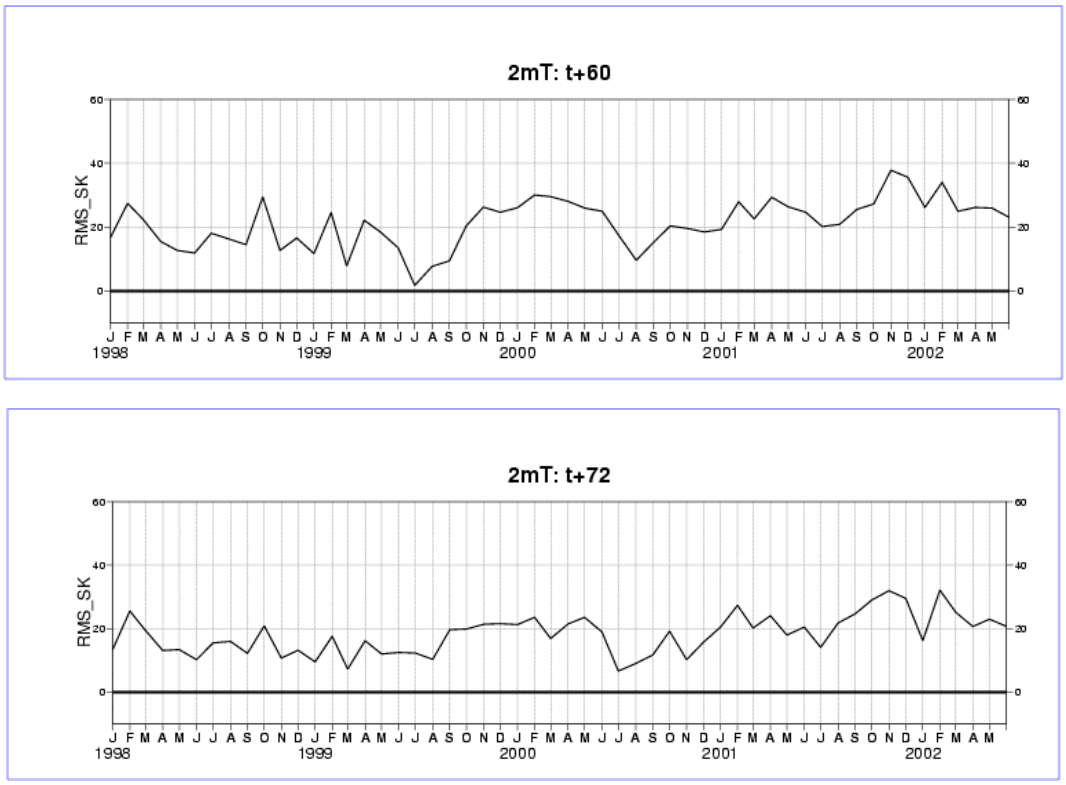


Figure 23: RMS error reduction with respect to persistence for 2m temperature forecasts over Europe during night time (60h forecast, top) and daytime (72h forecast, bottom

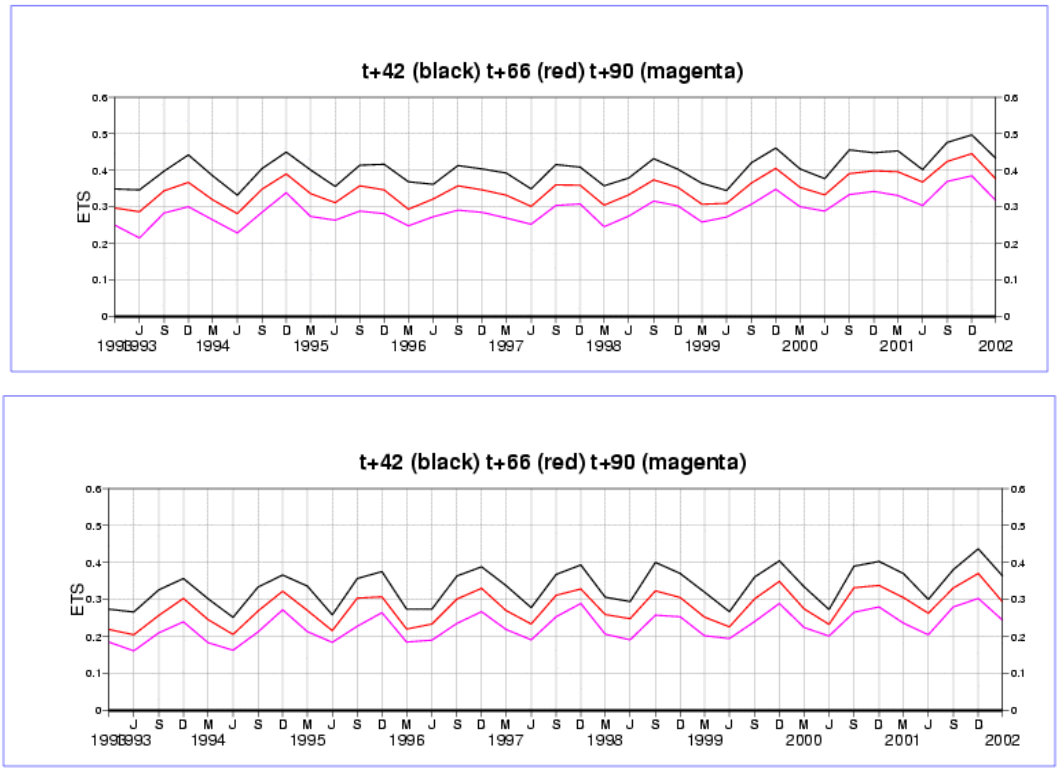
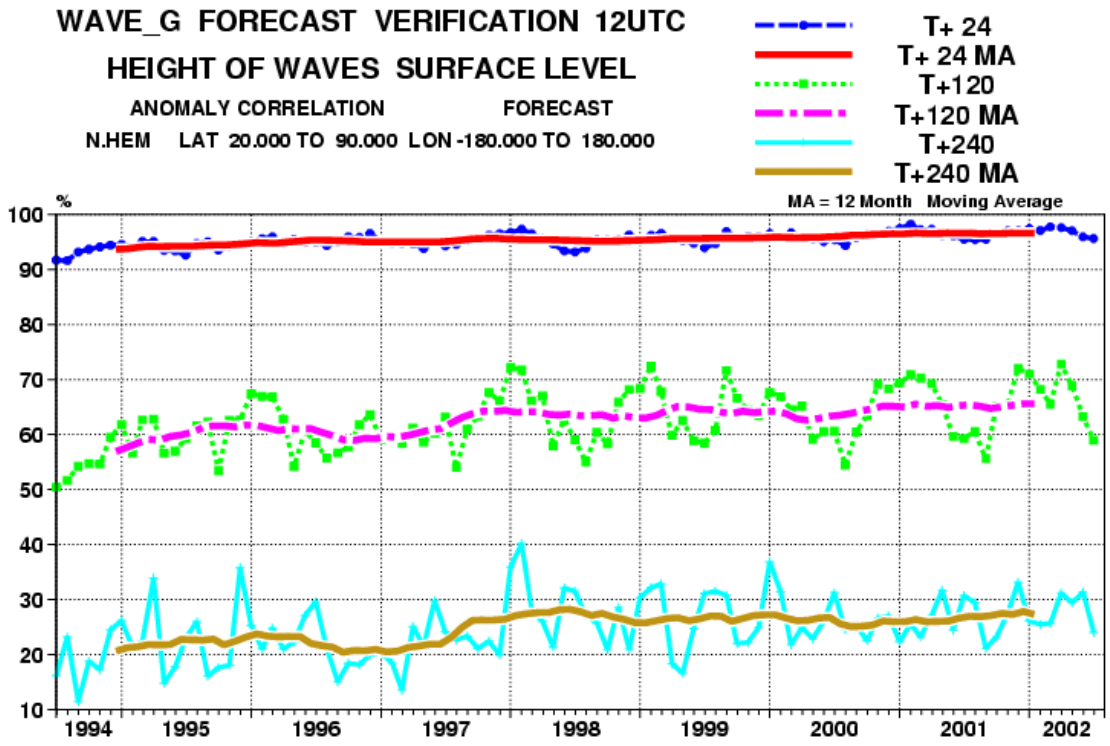


Figure 24: Equitable Threat Score (ETS) for daily precipitation forecasts over Europe in excess of 1mm (top) and 5mm (bottom)



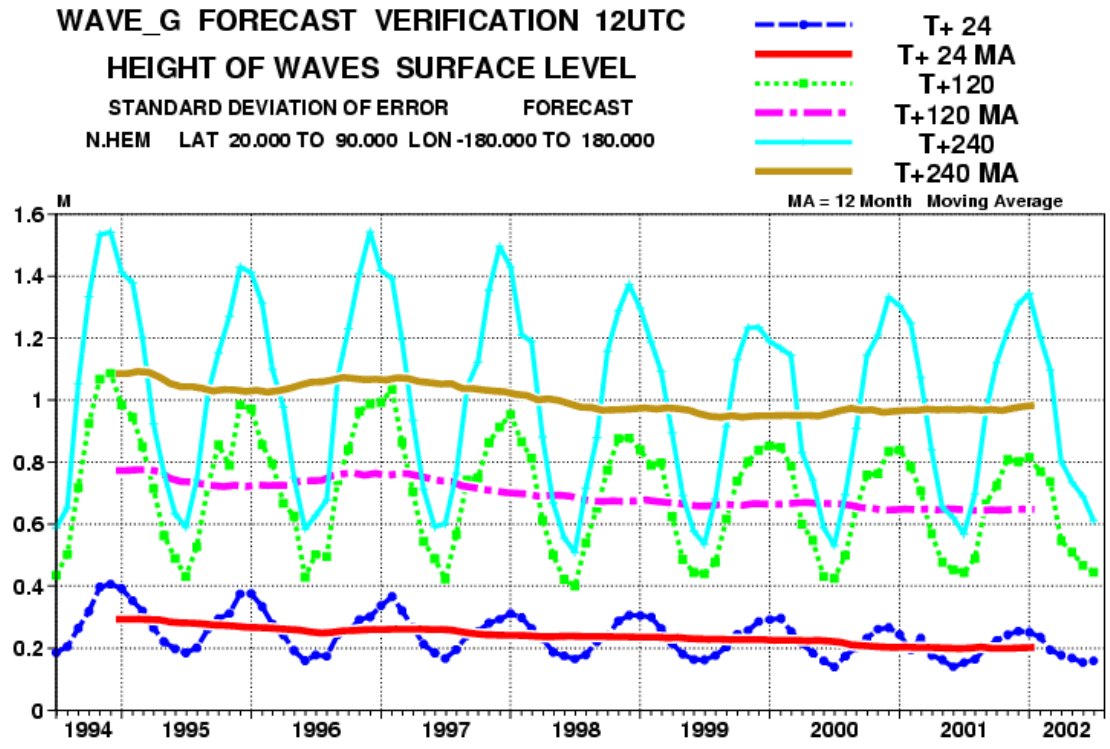
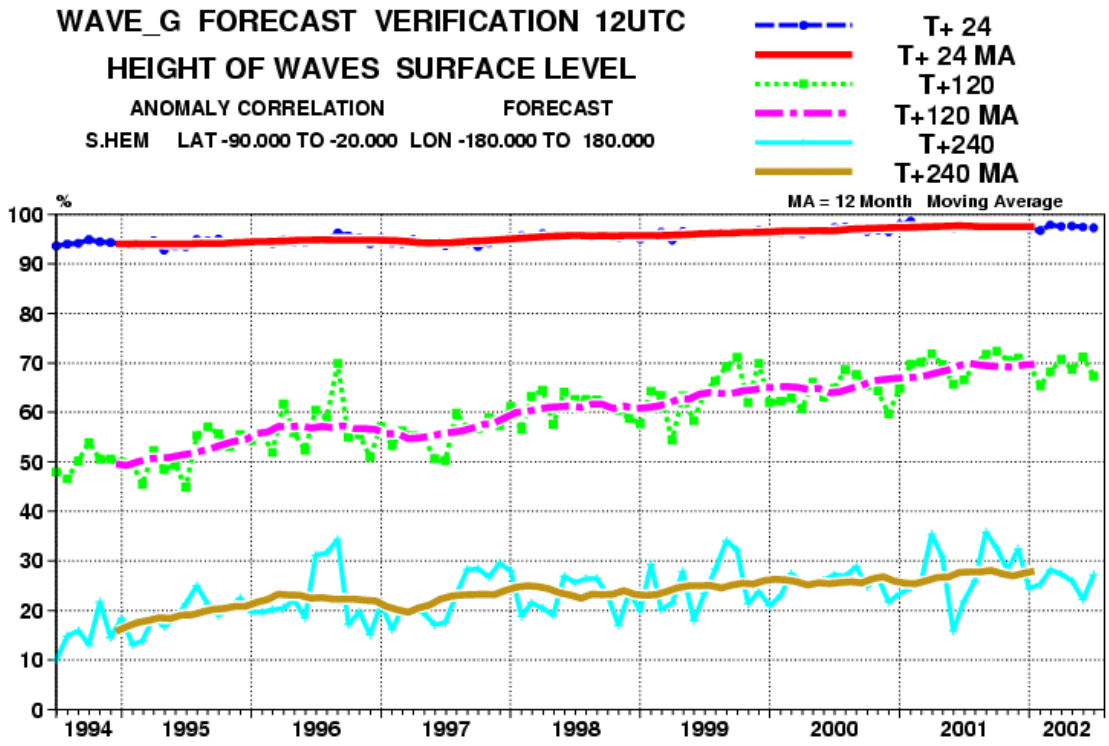


Figure 25: Scores (std and anomaly correlation) of oceanic wave heights verified against the analysis (Northern Extratropics)



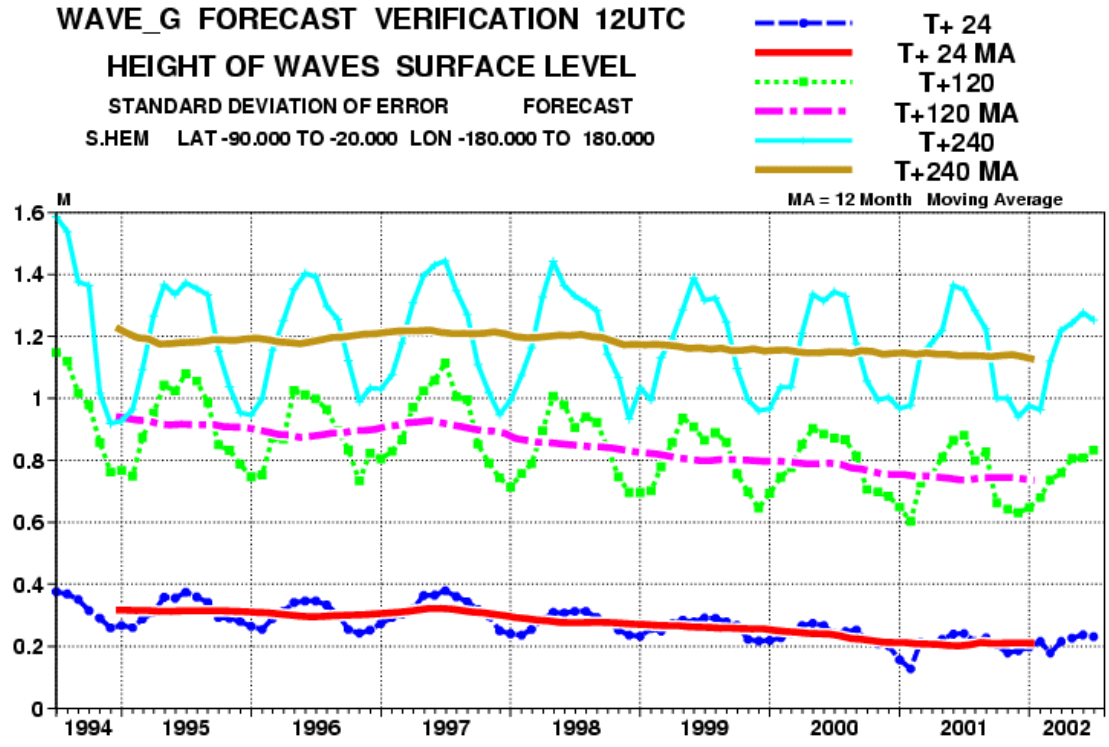


Figure 26: Scores (std and anomaly correlation) of oceanic wave heights verified against the analysis (Southern Extratropics)

Figure 27: Time series of Brier Skill Scores for EPS daily rainfall day 4 (top) and 6 (bottom) forecasts over Europe (12-months moving data samples, verifying observations gathered from SYNOP reports);

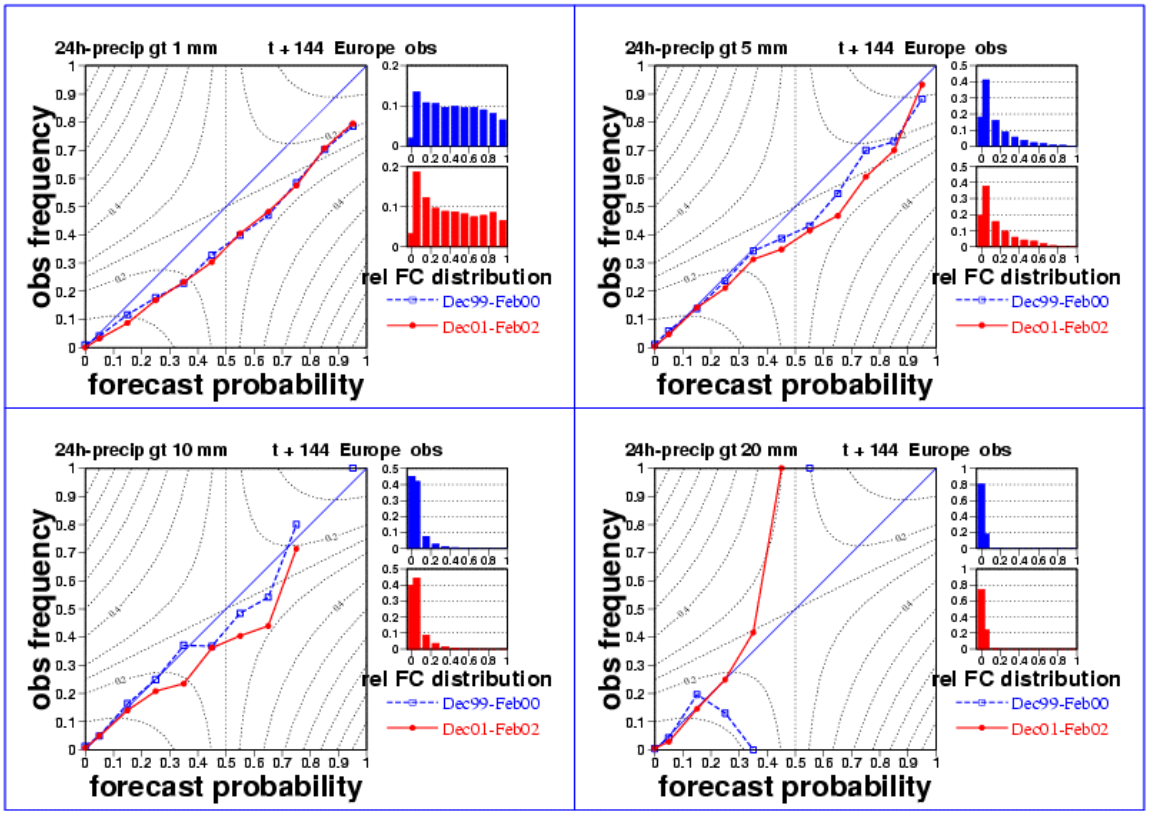


Figure 28: Reliability diagrams for Day 6 EPS forecasts of daily precipitation over Europe in winter 2001-02 (red) and 1999-2000 (blue); thresholds(clockwise from top left): 1, 5, 20 and 10mm.

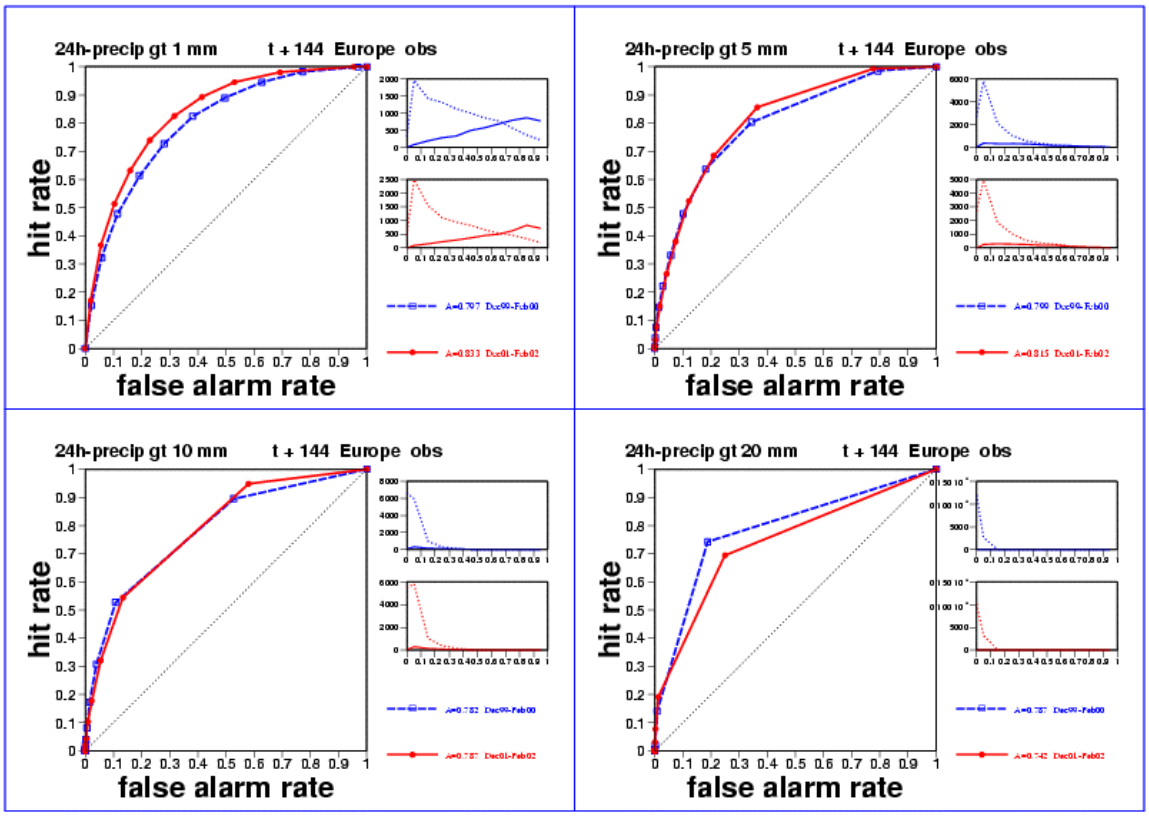


Figure 29: ROC curves for Day 6 EPS forecasts of daily precipitation over Europe in winter 2001-02 (red) and 1999-2000 (blue); thresholds(clockwise from top left): 1, 5, 20 and 10mm.

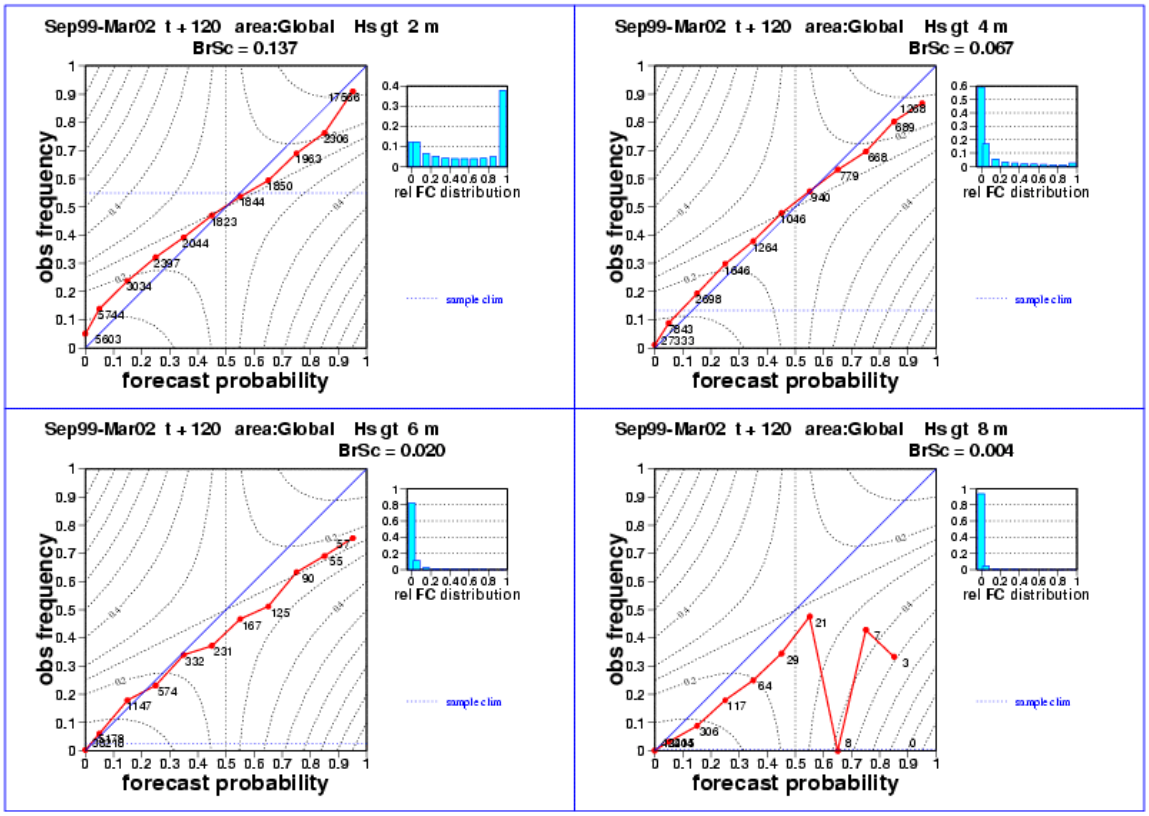


Figure 30: Reliability diagrams for Day 5 EPS forecasts of wave height (all areas); thresholds (clockwise from top left): 2,4,8 and 6m. Verification data are observations from buoys.

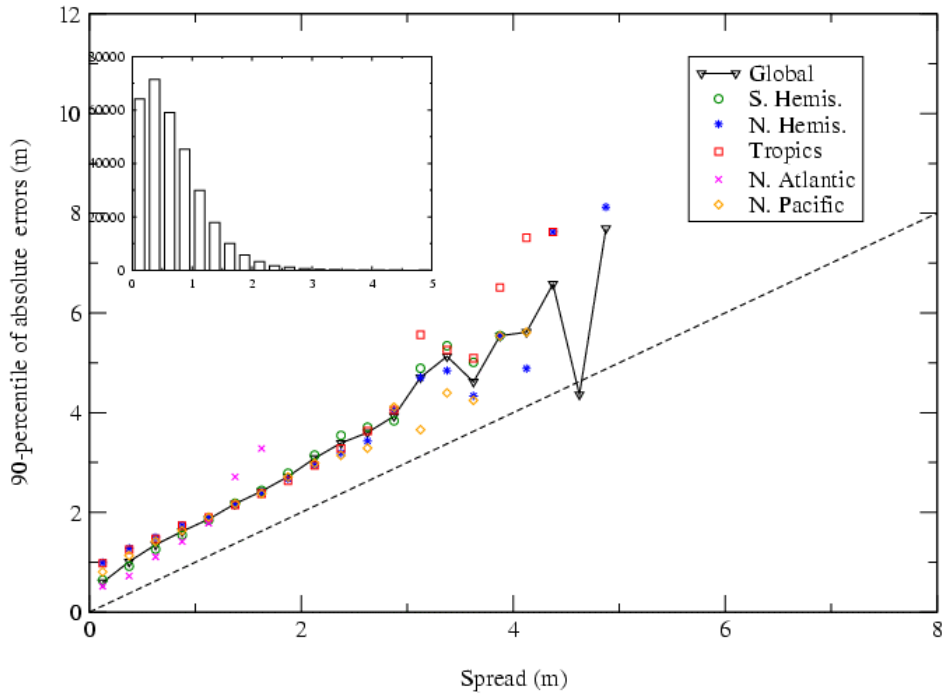


Figure 31: Day 5 EPS wave forecast spread-skill diagram (verification data are from ERS altimeter). The frequency in the various bins for ensemble spread are depicted in the bar diagram.

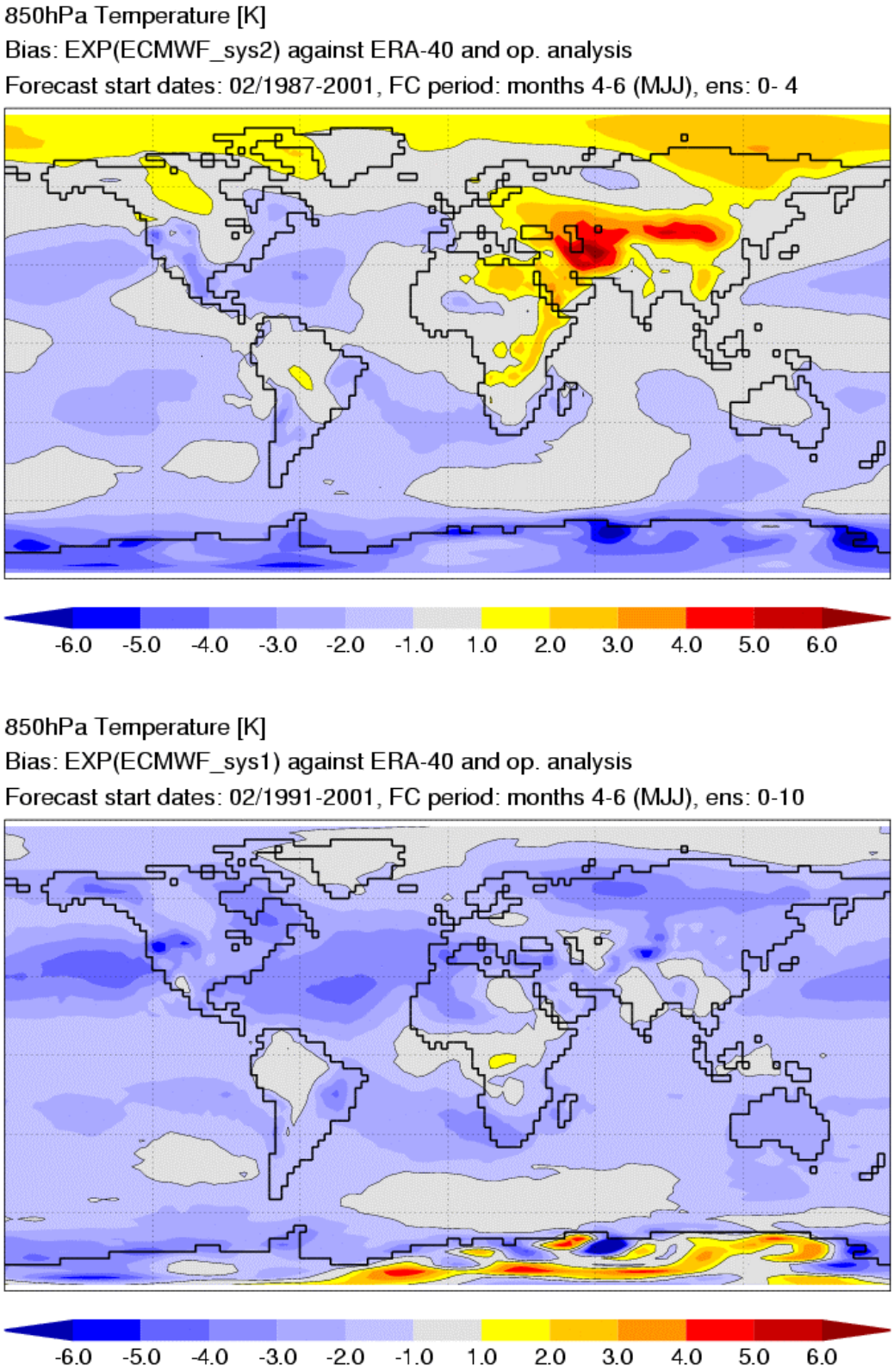


Figure 32: 850-temperature bias in the 4-6 month forecasts originating from February and verified against ERA-40 analyses. Top: System 2; Bottom: System 1

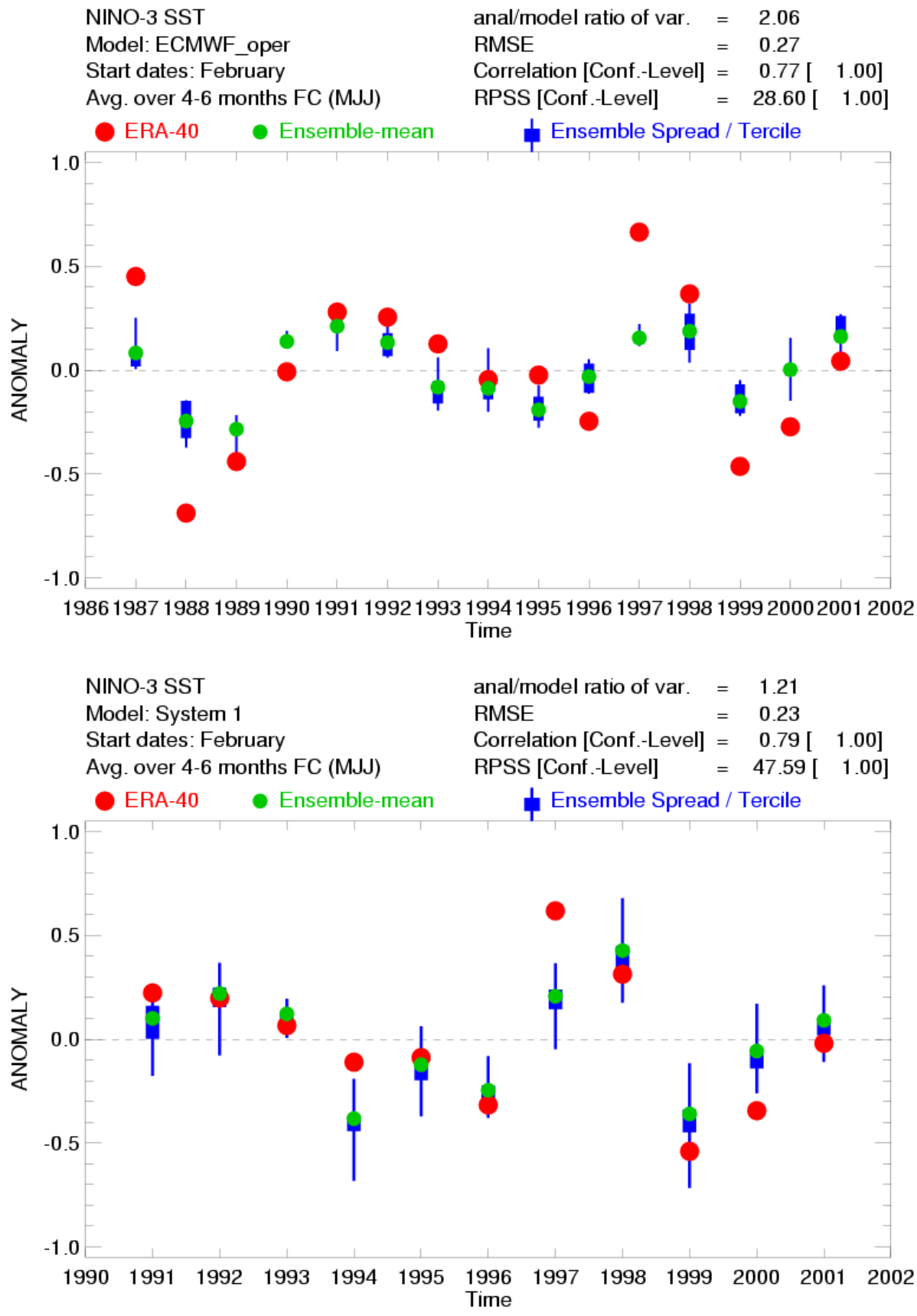


Figure 33: Verification (red dots) of Nino-3 SST forecasts; top: system 2 (operations) ; bottom: system 1

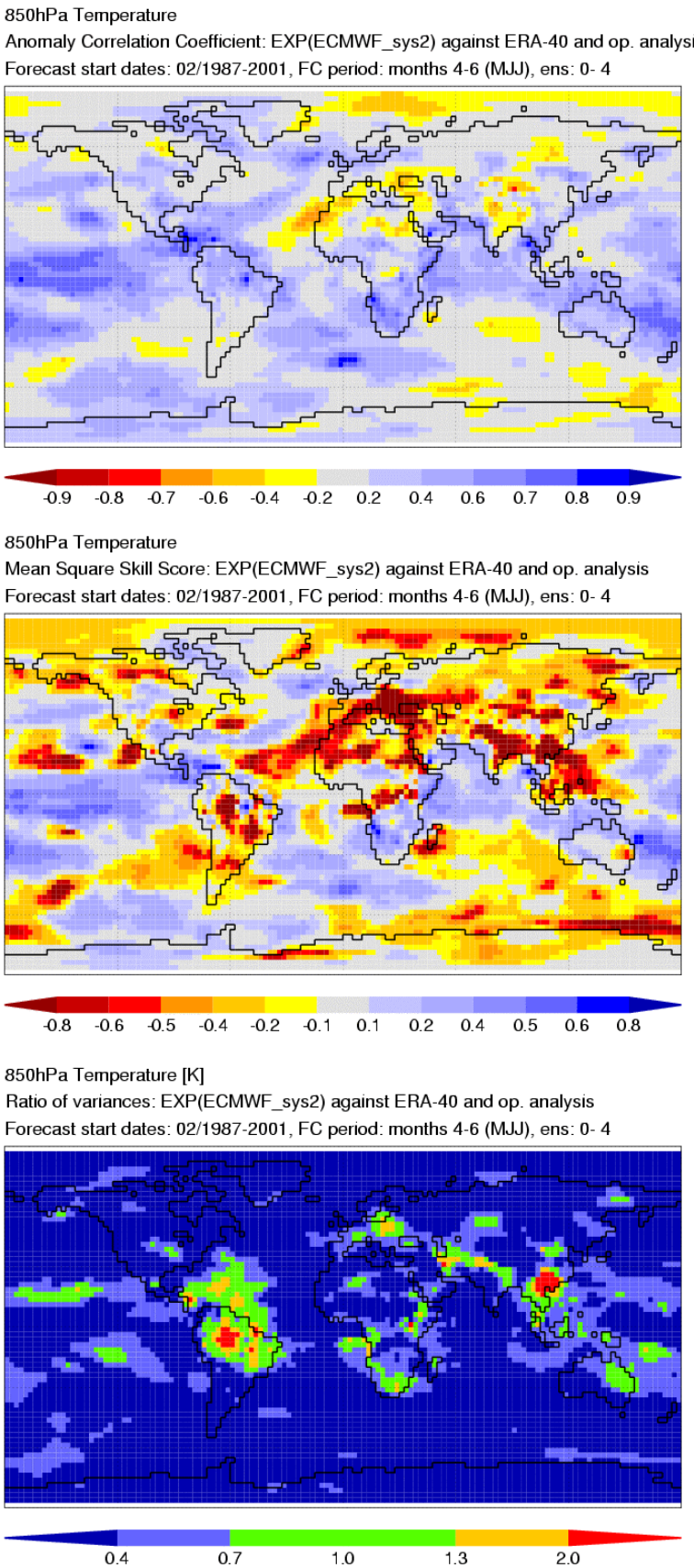


Figure 34: Top: Anomaly Correlation; Middle: Mean Square Skill Score; Bottom: Variance ratio of 850-hPa temperature 4-6 months forecasts originated in February 1987-2001

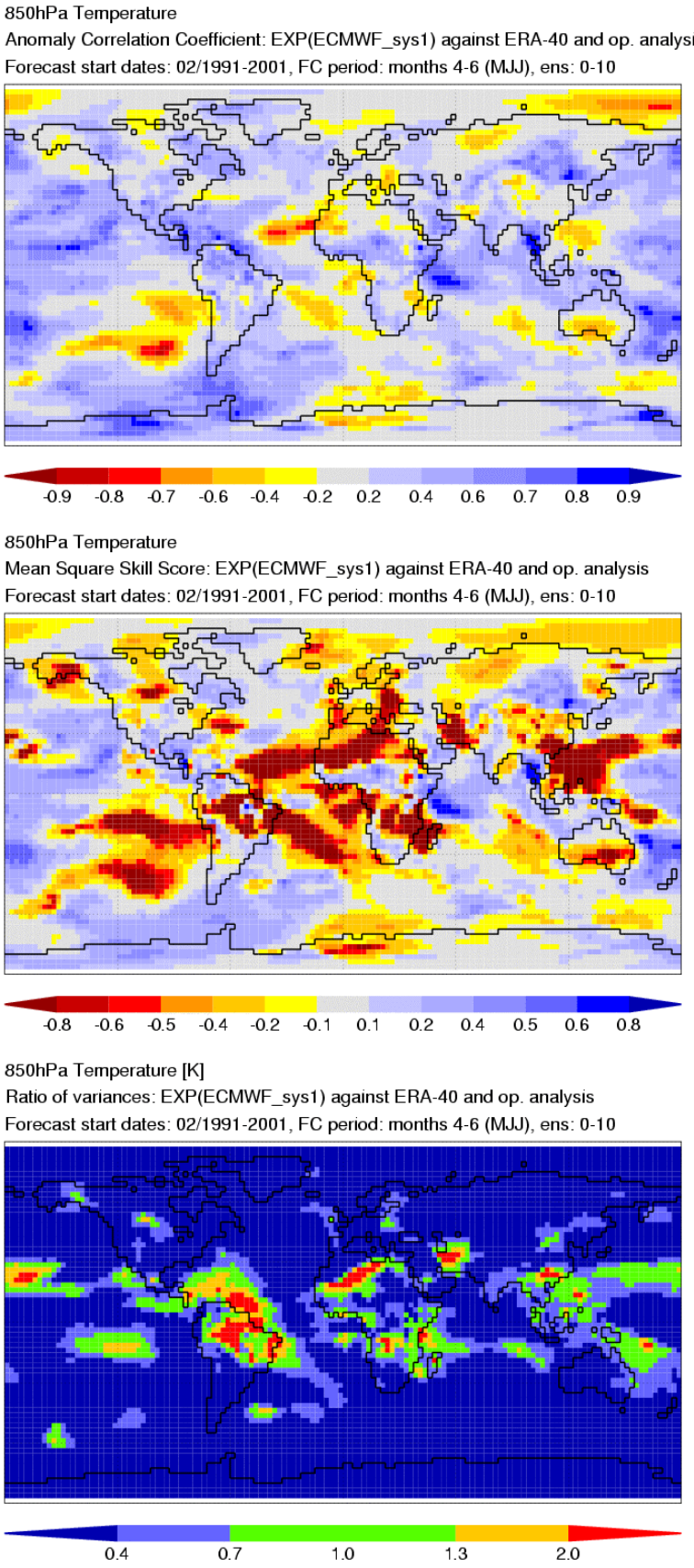


Figure 35: same as Figure 34, but for system 1 instead of system 2 (also, reference period is 1991-2001 instead of 1987-2001)

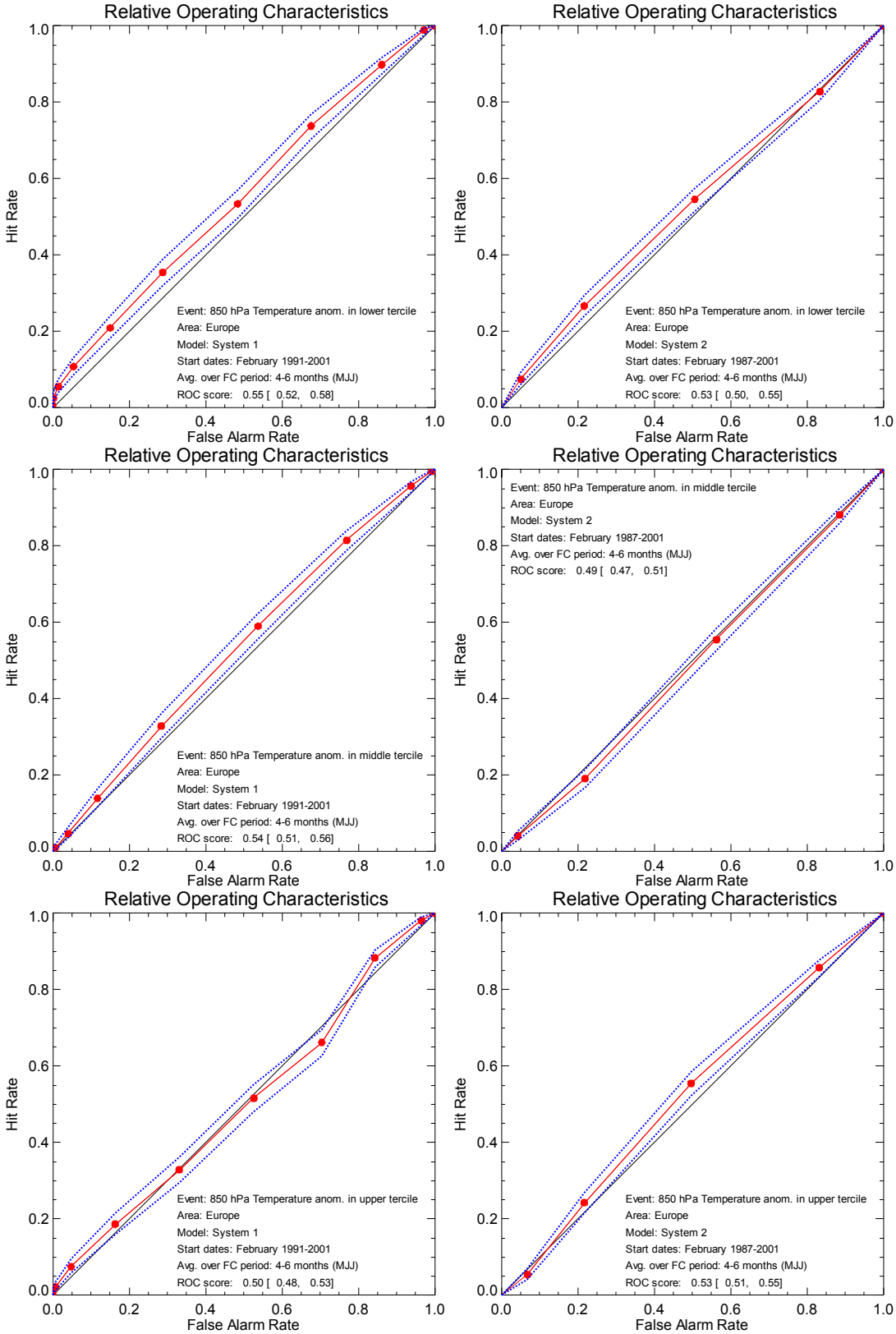


Figure 36: ROC diagrams for 850-hPa temperature anomaly 4-6 months forecasts over Europe originating in February 1991-2001 (left, system 1) and 1987-2001 (right, system 2); top row: below normal (lower tercile); middle: near normal (middle tercile); bottom row: above normal (upper tercile)

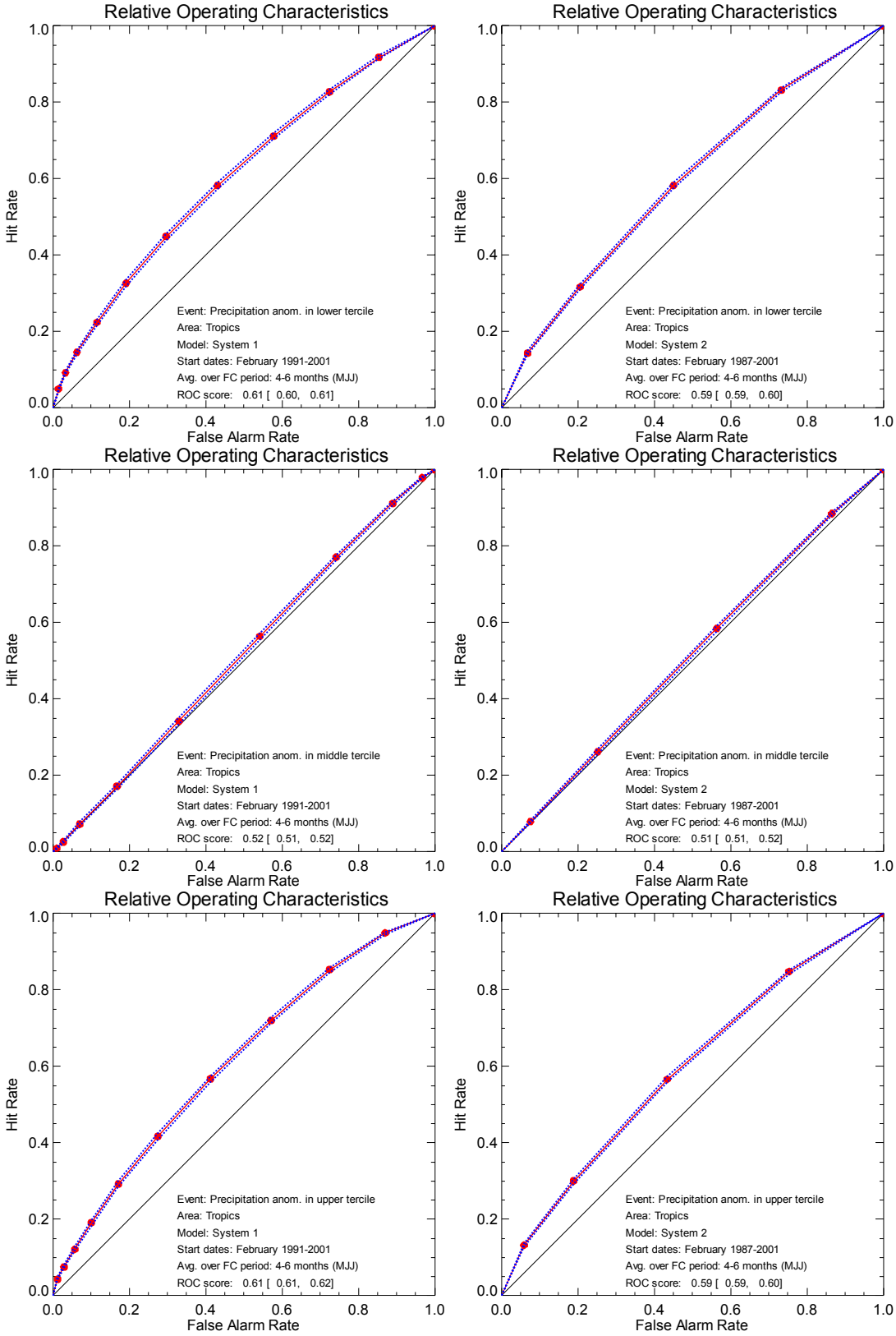


Figure 37: ROC diagrams for rainfall anomaly 4-6 months forecasts in the Tropics originating in February 1991-2001 (left, system 1) and 1987-2001 (right, system 2); top row: below normal (lower tercile); middle: near normal (middle tercile); bottom row: above normal (upper tercile)