TECHNICAL MEMORANDUM

504

Verification statistics and evaluations of ECMWF forecasts in 2005-2006

D. Richardson, J. Bidlot, R. Buizza, L. Ferranti, A Ghelli, G van der Grijn, F. Vitart and E. Zsoter

Operations/Research Department

October 2006

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

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

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

1. Introduction

This document presents recent verification statistics and evaluations of ECMWF forecasts. Recent changes to the data assimilation/forecasting and post-processing system are summarised in Section 2. Verification results of the medium-range free atmosphere ECMWF forecasts are presented in Section 3, including, when available, a comparison of ECMWF forecast performance with that of other global forecasting centres. Section 4 deals with the verification of ECMWF forecasts of weather parameters and ocean waves, while severe weather events are addressed in Section 5. Finally, Section 6 provides insights into the performance of monthly and seasonal forecast systems. A short technical note describing the scores used in this report is given in Annex A.

The set of verification scores shown here is mainly consistent with that of previous years, in order to aid comparison from year to year (ECMWF Tech. Memos. 346, 414, 432, 463, 501).

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/medium/verification/ (medium-range)

http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/ (monthly range)

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

2. Changes to the data assimilation/forecasting/post-processing system

The changes to the system since the preparation of documents for the last meeting of the Committee are summarised below.

Cycle 30r1 including a major resolution upgrade was introduced on 1 February 2006. The changes are:

- Increase in horizontal resolution to T799 for the deterministic forecast and the outer loops of 4D-Var
- Increase in vertical resolution to 91 levels; model top raised to 0.01 hPa
- Increase in horizontal resolution to T255 for the second inner loop of 4D-Var
- Increase in horizontal resolution to T399 and in vertical resolution to 62 levels (model top ~5 hPa) for the EPS
- Increase in vertical resolution to 62 levels for the monthly forecast system
- Increase in horizontal resolution to 0.36 degrees for the global ocean wave model
- Increase in wave spectral resolution to 24 directions and 30 frequencies (from 12 and 25 respectively) for the EPS ocean wave model
- Use of grid-point humidity and ozone in 4D-Var
- Revised coefficients (version 2.3) for the linearised ozone chemistry scheme, supplied by Daniel Cariolle, CERFACS
- Use of Jason altimeter wave height data and ENVISAT ASAR spectra in the wave model data assimilation; ERS-2 SAR spectra are no longer assimilated

Other changes to the operational forecasting system included

- 8 September 2005: AMSU-A radiances from the new NOAA-18 satellite were included in the operational assimilation
- 9 November 2005: Warmest (rather than central) field of view AIRS data included in the operational assimilation
- 30 November 2005: MHS humidity data from NOAA-18 and humidity data from dropsondes were included in the operational assimilation
- 7 February 2006: Revised satellite radiance bias files, reducing temperature biases of the analysis compared against aircraft and radiosondes. The new radiance bias correction is derived from a trial variational bias correction scheme which is being tested for full operational implementation in the next model cycle.

Note: All forecasting-system cycle 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-month periods since 1980. The forecast parameter is the 500 hPa height over the northern hemisphere (extratropics only) and Europe, while the scoring method is root mean square error, normalised with reference to a forecast that persists initial conditions into the future. The last month included in the statistics is July 2006. The trend for a marked improvement in the quality of the forecasts observed over the last few years reached a plateau in 2004, with some degradation in the later stages of the forecast, especially over Europe. Since mid-2005, the trend is again upwards, although the scores have yet to regain the levels of 2003-04. If reference to climatology rather than persistence is used and errors are measured by anomaly correlation rather than root mean square differences (Figure 2), this positive trend is confirmed, with consistently high skill in the first half of 2006, compared with the same period in previous years over Europe and both the northern and southern extratropics. In contrast to recent years, these two sets of figures give consistent signal. The dip in skill since 2003 in Figure 1 was attributed to more persistent circulation in winter 2004-05, making persistence more difficult to beat (Figure 3). Persistence for winter 2005-06 was similar to 2004-05, while June and July 2006 have very good persistence scores in Figure 3. It is encouraging that Figure 1 demonstrates a positive trend in recent months, despite this high persistence, indicating model improvements rather than a change in synoptic activity.

Figure 4 shows the distribution of anomaly correlation scores for day 7 forecasts of 850 hPa temperature over Europe in winter and summer. The percentage of good forecasts was slightly lower in winter 2005-06 than in the previous winter, although the proportion of excellent forecasts (correlation greater than 80%) was not reduced. There was a noticeable dip in the number of good forecasts in summer 2005 compared to the previous two exceptionally good years. This reduction was shared by other NWP centres and can be associated with an active westerly circulation pattern that dominated the summer circulation over central and

northern Europe. Initial results for summer 2006 (only June and July 2006 are included in Figure 4) indicate that performance is back to the high level of the previous years. While the trend for the ensemble mean scores matches that of the operational high resolution forecast, the ensemble mean is consistently better (at this forecast range day 7), whatever the conditions.

One of the noteworthy results over the past few years has been that the improvement of the deterministic forecast quality has translated into improved consistency in the forecasts valid for the same date from one day to the next. This level of consistency has stayed very high again this year, as can be seen from Figure 5, showing the time series of the average RMS difference between consecutive forecasts over Europe and the northern extratropics. This high consistency has not been adversely affected by the increase in resolution from T511 to T799 on 1 February.

The quality of ECMWF analyses and forecasts for the upper atmosphere has been recognised by several institutions. Among them is WMO/AREP, which uses them when preparing the bi-weekly WMO Antarctic Ozone Bulletins. The time series of wind scores at level 50hPa in the extratropics is shown in Figure 6. The very good performance reached over the past couple of years was maintained in 2005, although errors increased a little during 2005 and early 2006. This will be kept under review.

The provision of EPS verification scores to JMA was initiated in 2004. This is ultimately intended to provide WMO/CBS members with a comparison of EPS scores. There are, so far, only a few participants (JMA, KMA, ECMWF, and recently CMC), so the exchange is still considered to be pre-operational and no comparisons can be made. Data are exchanged in the form of contingency tables, from which a variety of scores can be derived. As an example, Brier skill scores and ROC areas for Europe are shown in Figure 7. The curves illustrate the long-term positive trend in EPS performance, however the level of noise attached to probabilistic verification scores, when applied to a small sample, makes it difficult to comment on year-to-year changes.

As an overall indication of probabilistic forecast performance, time series of the ranked probability skill score (RPSS) for 500 hPa height over the northern hemisphere and Europe are shown in Figure 8. Again, the long-term upward trend in skill is clear. Although also subject to sampling noise, it is noted that skill levels for Europe reached a new peak in early 2006.

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. The reduction in medium-range errors that followed the major model changes (both in the physics and data assimilation) early in 2003 (Cy25r4) has been maintained this year. Although the day-1 error has increased slightly when verified against the model analysis, verification against radiosonde observations shows a continued reduction in error during 2005.

3.2 ECMWF vs other NWP centres

3.2.1 Deterministic (T511/T799) model

The common ground for such a comparison is the regular exchange of scores between GDPFS centres under WMO/CBS auspices, following agreed standards of verification. Figure 10 shows time series of such scores over the northern extratropics for both 500hPa height and Mean Sea Level Pressure. Our lead over other centres was reduced last winter because of substantial improvements in NCEP forecasts. However, differences so far this summer are comparable to previous years. The gap is, in general, bigger in the

southern extratropics (Figure 11). The improved cold season performance from the Met Office noted in 2005 is continued in 2006, but improved ECMWF scores help maintain the substantial lead at longer range. The substantial improvement of NCEP forecasts is less noticeable in the southern hemisphere.

WMO exchanged scores also include verification against radiosondes over smaller areas such as Europe. Figure 12, showing both 500 hPa height and 850 hPa wind errors, confirms the good performance of our forecasts using this alternative reference.

The situation in the Tropics is summarised in Figure 13. Since mid-2005, the Met Office has had the lowest short-range errors, while performance at day 5 is similar for ECMWF and the Met Office. Although this verification against analyses shows a small increase in short-range error for ECMWF, the corresponding scores for radiosonde observations show a continuing trend of reducing errors.

4. Weather parameters and ocean waves

4.1 Weather parameters - deterministic and EPS

Long-term trends in mean error and standard deviation of error for 2m temperature, specific humidity, total cloud cover and 10 metre wind speed forecasts over Europe are shown in Figure 14 to Figure 17. Verification is against synoptic observations available on the GTS. A correction for the difference between model orography and station height was applied to the temperature forecasts, but no other post-processing has been applied to the model output. Temperature errors are generally similar to previous years, although there have been periods of large daytime errors in winter that appear to be at least partly associated with the representation of low cloud in periods of anticyclonic conditions. The long-term trend in winter wind-speed error has continued, with 2005-06 recording lower errors than previous winters (Figure 17).

The trend in precipitation skill for Europe is shown in Figure 18 using the True Skill Score (or Pierce's Skill Score) for thresholds of 1mm and 10mm per day. For both thresholds, 2005 had the highest ever summer skill, with particularly high scores for the higher threshold. This may be partially associated with the more active synoptic flow over northern Europe in summer 2005. Scores for winter 2005-06 also maintained the good levels achieved in recent winters.

Similar improvements in precipitation skill are found for the EPS. Figure 19 shows the long-term positive trend for EPS probability forecasts of precipitation over Europe for different thresholds, measured by Brier Skill Score (BSS) and ROC area. As for the deterministic forecasts, 2005 was the most skilful summer for all thresholds. Winter 2005-06 was also particularly skilful for the EPS, notably for the high 20mm/day threshold.

4.2 Ocean waves

The quality of the ocean wave model analysis continues to improve. The analysed wave height compared well to independent ocean buoy observations throughout last year and there is a continued improvement in the first half of 2006 (Figure 20), demonstrating the benefit of additional observational data from the Jason altimeter introduced in February 2006. Figure 20 also shows a time series of the analysis error for the 10 m wind over maritime regions. The analysed wind is compared to the 10 metre wind observed by buoys in the oceans. The error has steadily decreased since 1998, providing better quality winds for the forcing of the ocean wave model.

The good performance of the wave model forecasts is confirmed again this year, as shown in Figure 21 and Figure 22. Comparisons with other models using buoy observations support this positive assessment. In Figure 23, the forecasts from different models are verified against a set of northern hemisphere wave buoys. The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of buoys. The improvement in the ECMWF analysis and forecasts this year can clearly be seen in this comparison.

5. Severe weather

5.1 Extreme Forecast Index (EFI)

The Extreme Forecast Index (EFI) was developed at ECMWF as a generic tool to provide some guidance on potential extreme events. By comparing the EPS distribution of a chosen weather parameter to the model's climatological distribution, the EFI indicates occasions when there is an increased chance of an extreme event occurring. Initial verification results were presented in last year's report. These demonstrated that the EFI increases the detection ability for extreme precipitation, compared to using the EPS probability directly. A revised climatology for the EFI was introduced at the time of the model resolution upgrade on 1 February 2006. The new configuration ensures that the EFI climatology uses the same model version as the operational forecast and will allow straightforward development of an EFI for additional parameters (Zsoter, 2006).

Verification results for daily precipitation and for maximum temperature exceeding the 99.5% threshold of the climatology at stations included in the EUMETNET/ECSN Climate Atlas for Europe are shown in Figure 24. The sample is for the period July 2005 to May 2006; for dates before 1 February 2006, the EFI climatology was recalculated using the new method, but with the model version that was in operation at the time. The solid curves show the discriminating ability, using different EFI thresholds as decision criteria, while the dashed curves show the equivalent results, obtained by directly using the EPS probability of the extreme event. Using the EFI increases the ability to detect these events, compared to simply taking EPS probabilities at face value. There is less degradation with forecast range for temperature than for precipitation.

5.2 Tropical cyclones

Verification of Tropical Cyclone (TC) forecasts is a routine activity that has recently been given a higher profile, due to the release of real-time TC forecast products on the GTS in BUFR format. Several hundreds have been tracked since 2002 (Figure 25), making it a suitable sample for both deterministic and probabilistic verification. The 2005 North Atlantic hurricane season was exceptional, breaking records for numbers, intensity and destruction. Of interest for the European area, Vince was the first TC to strike the Iberian peninsula and the most north-easterly forming tropical cyclone on record; tropical storm Delta caused significant damage in the Canaries. In contrast, the 2006 North Atlantic season has had a quiet start.

Position and core pressure errors for the latest 12 month period (August 2005 to July 2006) are shown in Figure 26a. The high resolution deterministic (T511/T799) and EPS control are verified against the best track reported by WMO RSMCs. The "consensus" EPS forecast is also shown; this is the average position of all EPS ensemble members that have successfully tracked a TC (Ensemble Mean). Clearly the higher resolution forecast provides the best results for both position and core pressure. Compared to the same period one year before (Figure 26b), position errors are reduced at all forecast steps and core pressure errors are substantially

reduced. The resolution upgrade to T799 and T399 was introduced into operations in February 2006, so most of the forecasts in this sample are from the previous resolution system. Results for the period since the system upgrade (February to July 2006), compared with the same period in 2005, confirm a large reduction in core pressure error and consistent improvement in position. There is some suggestion that TCs in the T799 forecasts may be too deep at days 4 and 5. However the sample size is relatively small, and caution is always needed when comparing samples from different years. Some notable examples of tropical cyclone genesis during the forecast have also been seen. It is too soon to have any objective verification, but this aspect will be kept under review.

A probabilistic verification of the "Strike probability" product offered on the web (the probability at any geographical point that a reported TC will pass within 120km in the next 120h) is shown in Figure 27. The performance for August 2005 to July 2006 is comparable to that for the previous 12 months. Although the reliability appears to be slightly lower this year, the resolution (as indicated by the ROC) is slightly better; however both differences are small and probably not significant.

6. Monthly and Seasonal forecasts

6.1 The 2005-2006 El Niño forecasts

During the summer months of 2005 over the tropical Pacific the overall patterns of convection, SST, low-level winds and upper-level winds were near average, consistent with ENSO-neutral conditions. In June cold SSTs were observed just off the coast of Ecuador and Peru but by August these cold anomalies had disappeared.

From November 2005 to April 2006, SST anomalies in the Niño 3.4 region were negative with minimum value -0.9°C. During this period, enhanced convection (above average rainfall) was observed over India and across Indonesia, as well as over the north Atlantic and Central America; the latter was consistent with the very active Atlantic hurricane season. Later in March, the suppressed convection over the eastern tropical Pacific indicated that the atmospheric component was in the state associated with cold (La Niña) conditions, as well as its oceanic counterpart. The cold SST anomalies decreased quite quickly and by May 2006 they returned to near-normal values. Consistent with this, convection returned to near-average along the equator and over the Maritime Continent. Figure 28 shows Niño-3.4 predictions throughout the year, together with subsequent verification (heavy blue dashed line). In general, forecasts over the Nino-areas verified well for this year. However, predictions of la Niña onset appeared to be quite challenging, so that in the September and October forecasts (not shown) just one or two ensemble members were indicating the development of cold conditions. In the November forecast, the ensemble spread in the first 3 months is still quite large. In contrast, la Niña's demise proved to be more predictable, with the January forecast showing a much smaller spread.

6.2 Seasonal Forecast performance for the tropical Atlantic

6.2.1 Prediction of Atlantic hurricane frequency

From March to November 2005 the tropical North Atlantic SSTs were quite warm and an extremely active Atlantic hurricane season was observed. Forecasts of the frequency of tropical storms in different ocean basins are generated from the ECMWF seasonal forecast system and published on the web:

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/Tropical_storm/Tropical_storm_forecast

The tropical storm forecast published in June 2005 indicated higher than normal tropical storm frequency with 95% level of confidence. This was not just an isolated event: Figure 29 shows the skill of tropical storm predictions, started in June, for the period 1993-2005. The skill estimate is the correlation between observed and predicted numbers of tropical storms in each season. The figure shows both the extent of the anomalous hurricane season in 2005 and the performance of the seasonal forecast system in capturing the inter-annual variations (correlation is about 0.70).

The European multi-model Seasonal to Inter-annual Prediction (EUROSIP) system is being developed to provide combined products from different dynamical seasonal forecast systems (currently ECMWF, Met Office and Météo-France). The potential benefit of such a system is illustrated in Figure 30, which shows the inter-annual variations of tropical storms for the period 1993-2005, as simulated by the multi-model system (blue line), an empirical scheme (red line) and the observed values (black line). The overall skill of the EUROSIP prediction, measured in terms of correlation, is about 0.75 and is significantly larger than that of each individual EUROSIP component (not shown).

6.2.2 West African Monsoon

In April 2005 cold SST anomalies developed over the Guinea basin and persisted until July 2005. These anomalies have the local effect of reducing the monsoon rainfall along the Guinea coast. The signal of reduced rainfall over that region was well captured by the forecasts issued in May 2005. Forecasts can be found on the web:

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/charts/group/seasonal_charts_2tm/

6.3 Seasonal Forecast performance for the extratropics during 2005-2006

Although several components can contribute to the predictability of extreme anomalies, most of the seasonal forecast skill is associated with the predictions of El Niño cycles and its remote impacts. Given the neutral ENSO condition of summer 2005 and that the La Niña conditions in winter 2005-06 had a relatively small amplitude and short lifetime, predictions for the extratropics were expected to have limited skill.

Since autumn 2005, North America has experienced warm conditions, which intensified during winter and persisted into spring 2006. Although the warm conditions were not predicted for the autumn, a good indication of warm anomalies was given by the EUROSIP forecast for the winter. In addition ECMWF and EUROSIP forecasts initiated in February 2006 predicted the warmth well for the subsequent 3 months. It is possible that the anomalous snow cover and surface conditions might have played a role in augmenting/persisting such anomalies.

Europe and Asia experienced below-average temperatures over wide areas through the winter 2005-06. Neither ECMWF nor EUROSIP dynamical systems predicted those cold conditions. In contrast, for the first time the Met Office issued a public forecast for the winter. Their predictions, based to a considerable extent on a statistical scheme that forecasts the North Atlantic Oscillation (NAO), indicated a negative phase of the NAO, implying colder than average conditions over much of Europe and the UK and drier than average conditions over much of the UK. Although observed anomalies during winter 2005-06 differed substantially from a typical NAO anomaly, the Met Office forecast proved to be more useful in this instance than any of the available dynamical predictions. Van Oldenborgh et al. (2005) compared the skill of simple statistical ENSO models with that of the ECMWF operational forecast system over the 1987-2001 period. They showed that, although the statistical approach could be advantageous for some seasons and regions, there is a

real benefit in using a dynamical system. Nevertheless, winter 2005-06 over Europe is an interesting case that will be further explored, in order to isolate the sources of predictability and the model deficiencies in representing them.

6.4 Monthly Forecasts

6.4.1 Verification

A comprehensive verification site has been developed. This displays predicted and observed anomalies of each individual case since December 2005, together with deterministic and probabilistic scores.

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

As an example of the verification available on the web, Figure 31 shows the ROC score computed over each grid point for the 2m temperature monthly forecast anomalies at two different forecast ranges: day 12-18 and day 19-25. All the real-time monthly forecasts since 7 October 2004 have been used in this calculation. The red shades correspond to ROC scores larger than 0.5 (the monthly forecast has more skill than climatology) and the blue shades correspond to ROC scores below 0.5 (the monthly forecast has less skill than climatology). The anomalies are relative to the past 12-year model climatology. The monthly forecasts are verified against ERA40 reanalysis or the operational analysis, when ERA40 is not available.

Although these scores are strongly affected by sampling, they can provide the user with a first estimate of the spatial distribution of forecast skill.

6.4.2 Monthly forecast performance 2005-2006

ROC scores computed over each individual season since May 2003 suggest that the monthly forecasting system has performed rather well during the past year (red curve in Figure 32) for the time ranges day 12-18 and day 19-32. The model has consistently performed better than persistence of the previous weekly or two-weekly period (blue curve in Figure 32). ROC scores of the probability that the 2-metre temperature is in the upper tercile over the northern extratropics were particularly high during the last winter and spring. In particular, the ROC scores for spring were the highest for this season, since the start of the monthly forecasting system, for both day 12-18 and day 19-32. It is not clear whether these good performances are due to improvements in the model physics or to a more predictable large-scale circulation.

The skill of the monthly forecasting system has also been monitored in the tropics. The late onset of the 2005 Indian monsoon was well forecast by the monthly system. In 2006, the model produced surprisingly good forecasts of the Indian monsoon up to three weeks in advance. It also successfully predicted a dry period a few weeks later, four weeks in advance. Verification over the last four summers suggests that the model has some skill in predicting precipitation over India up to 3 weeks in advance. However, the skill of the monthly forecast system in predicting the African monsoon is very poor. This seems to be linked to the fact that the model does not propagate the ITCZ as far north in Africa as is observed.

The stratosphere-troposphere interaction is an important source of predictability for the monthly forecasting system (Baldwin et al, 2000). In fact, with a delay of a couple of weeks the sudden stratospheric warming can have a significant impact on the tropospheric circulation. Jung et al (2005) have shown that the ECMWF forecast system is able to simulate such propagation. Following a strong, sudden warming event that took place in the second half of January 2006, the skill of the monthly forecasting system to predict a stratospheric sudden warming event has been assessed. The model showed some skill in predicting the January 2006 sudden warming more than 10 days in advance, and a study using composites of all the observed sudden

warming events and their forecasts since 1990 suggests that the model has some skill in predicting stratospheric sudden warming events more than 2 weeks in advance.

References

Baldwin, M. P., D.B. Stephenson, D.W.J. Thompson, T.J. Dunkerton, A.J. Charlton, and A. O'Neill, 2003: Stratospheric memory and skill of extended-range weather forecasts. Science, 301, 636-640.

Jung, T., J. Barkmeijer and M.J. Rodwell, 2005: Sensitivity of the tropospheric circulation to changes in the strength of the stratospheric polar vortex. Submitted to Monthly Weather Review.

Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. *ECMWF Tech. Memo* **430.**

van Oldenborgh G, M.A. Balmaseda, L. Ferranti, T. N. Stockdale and D. L. T. Anderson, 2005: Did the ECMWF Seasonal Forecast Model Outperform Statistical ENSO Forecast Models over the Last 15 Years?. Journal of Climate Vol. 18 N.16 3240-3249.

Zsoter, E., 2006: Recent developments in extreme weather forecasting. *ECMWF Newsletter* **107,** *Spring 2006***.**

List of Figures

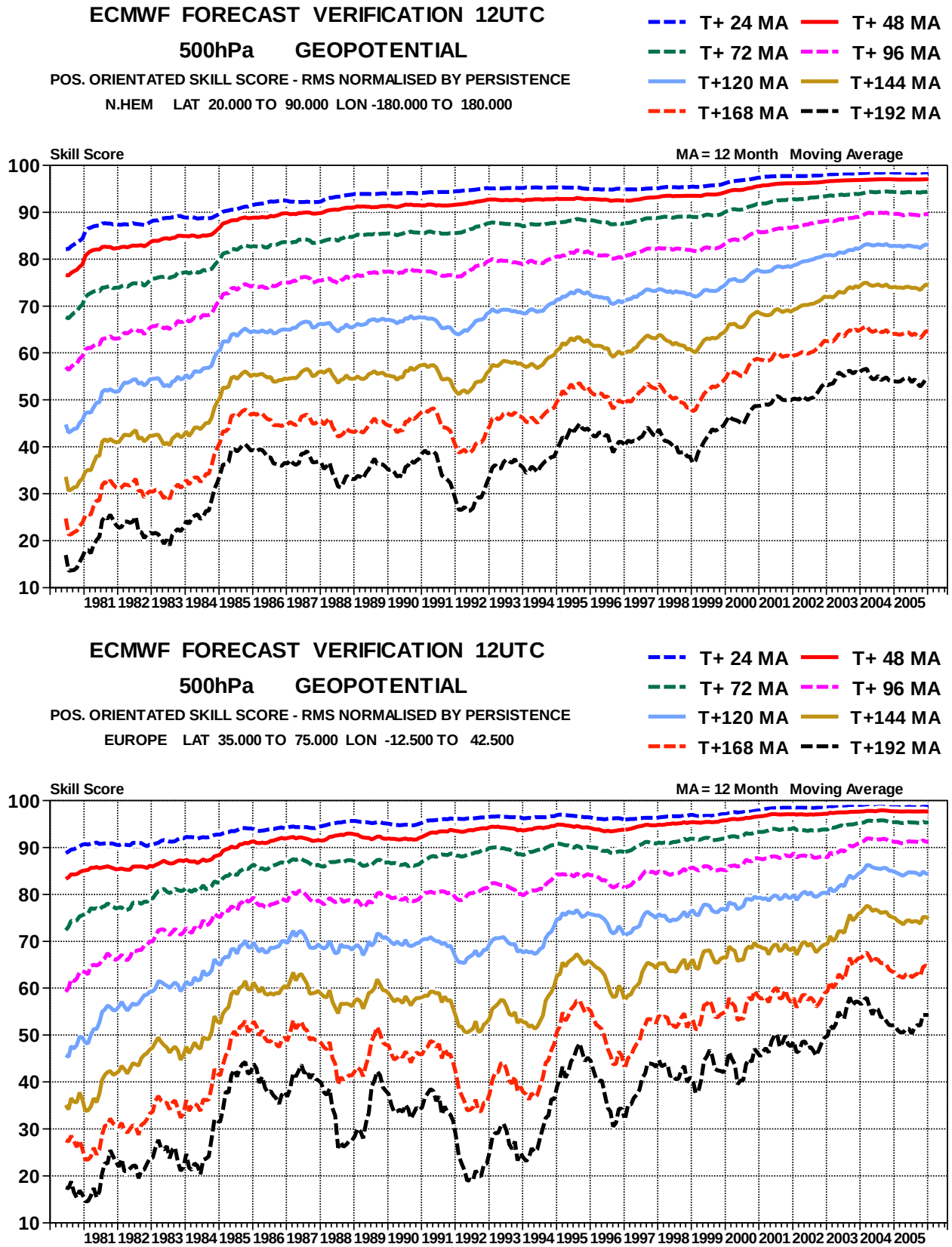


Figure 1: 500hPa height skill score (northern hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)

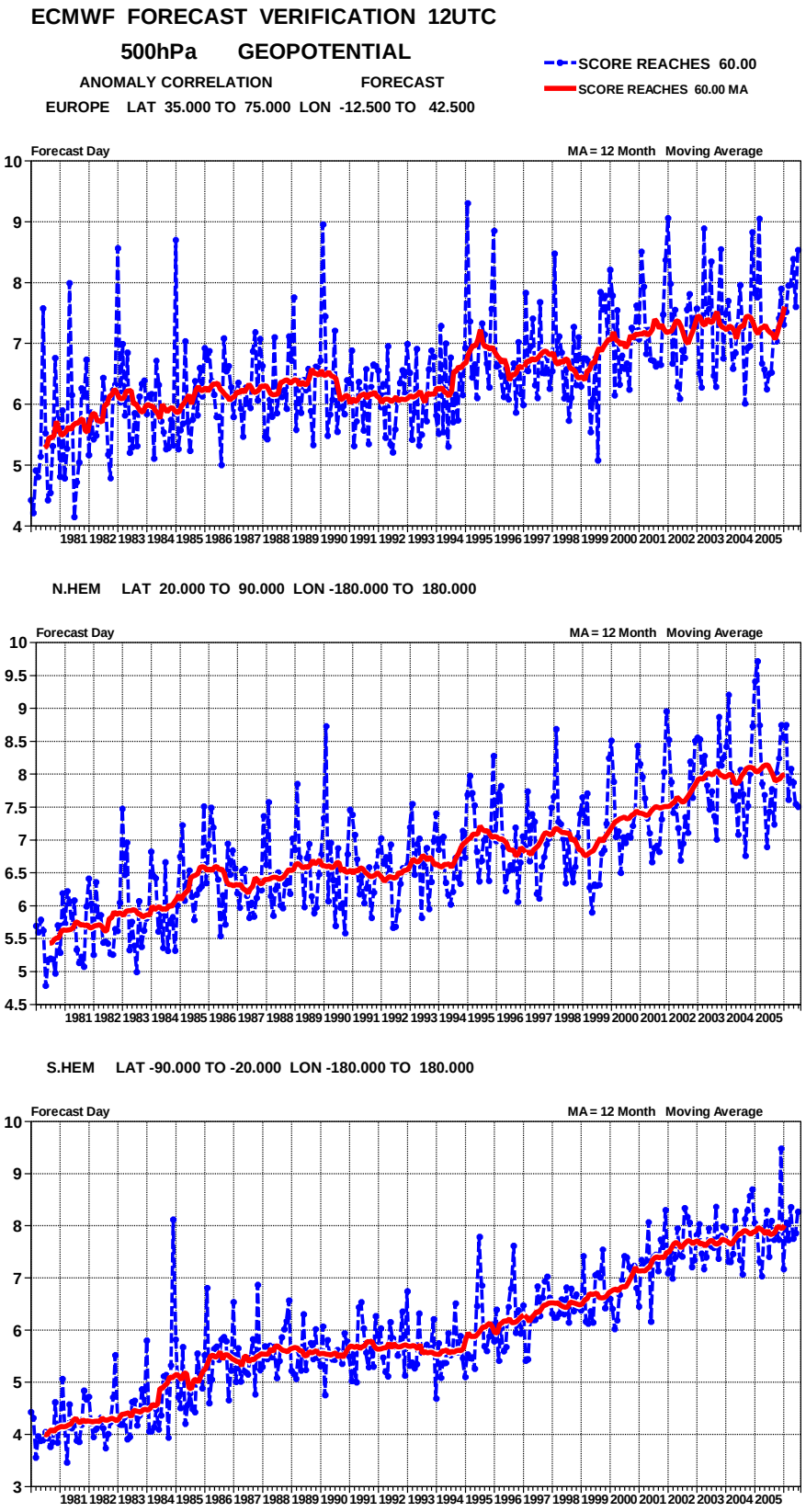


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the blue curves is the forecast range at which the monthly average of the forecast anomaly correlation with the verifying analysis falls below 60% for Europe, northern and southern extratropics (the red curve is the 12-month moving average)

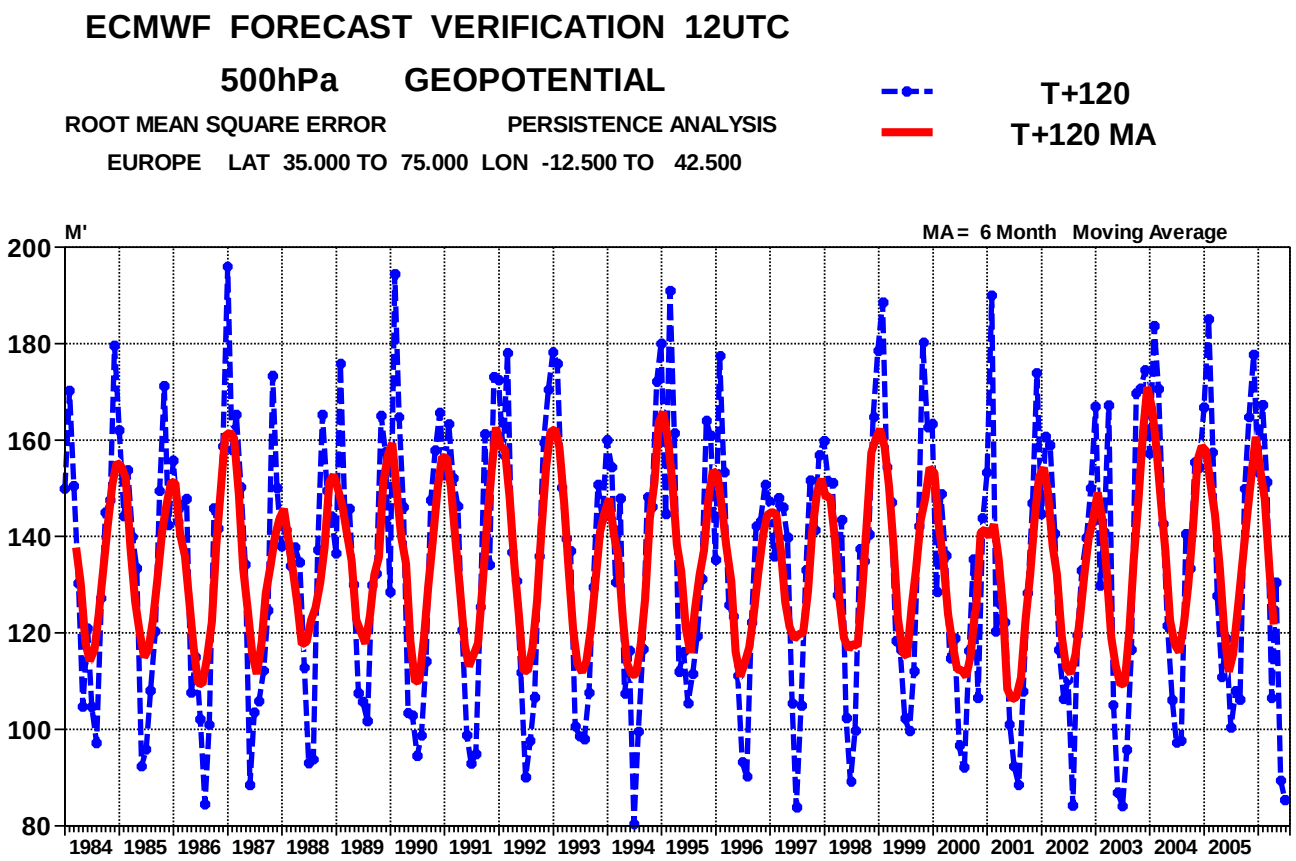


Figure 3: Root Mean Square Error made by persisting the analysis over 120h and verifying it as a forecast, monthly averages in blue, six-monthly moving averages in red. 500 hPa geopotential height over Europe.

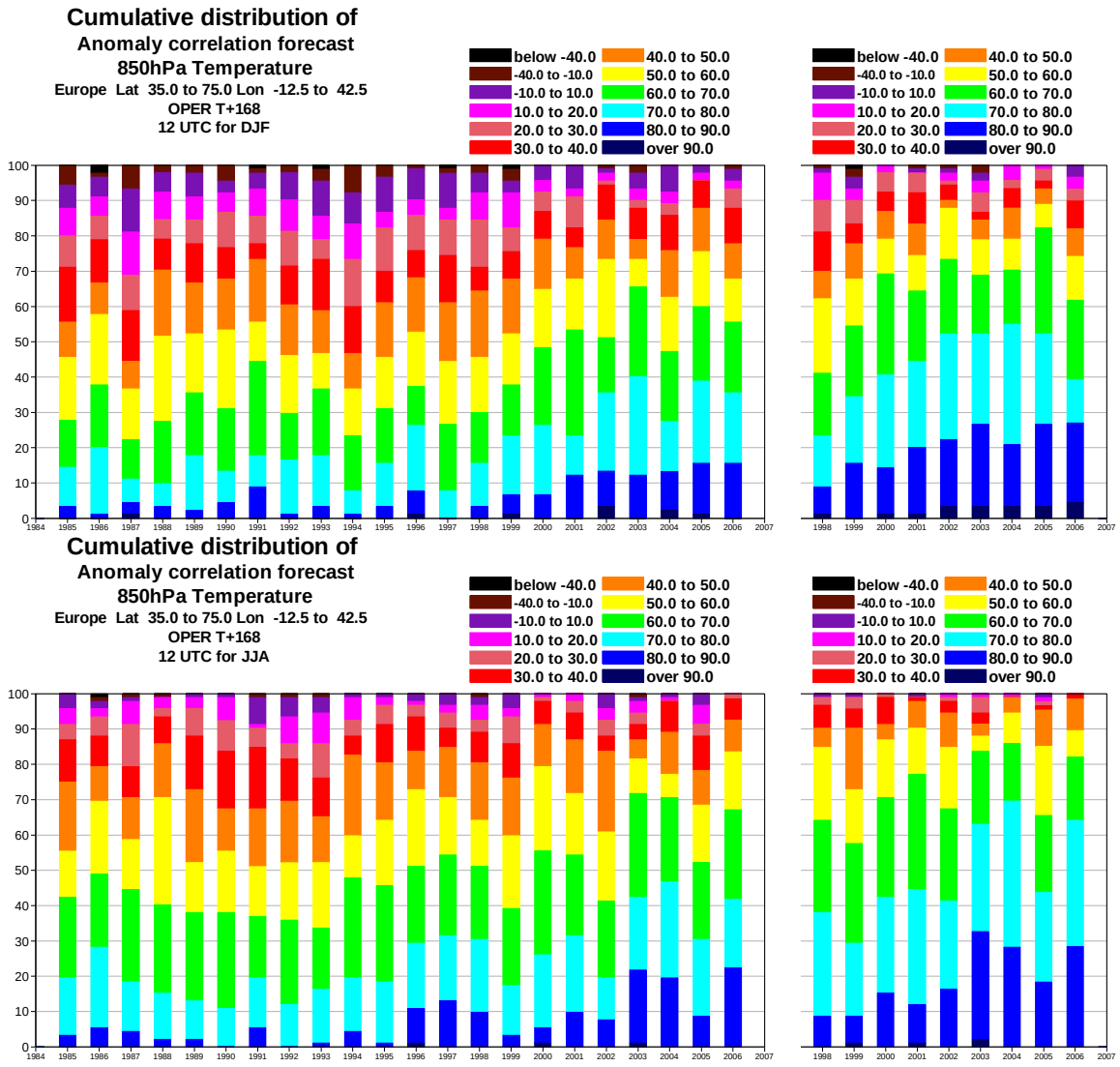


Figure 4: Cumulative distribution of Anomaly Correlation of the Day 7 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1984-85 for the deterministic, high resolution forecasts (left panels) and since 1997-98 for the EPS ensemble mean (right panels).

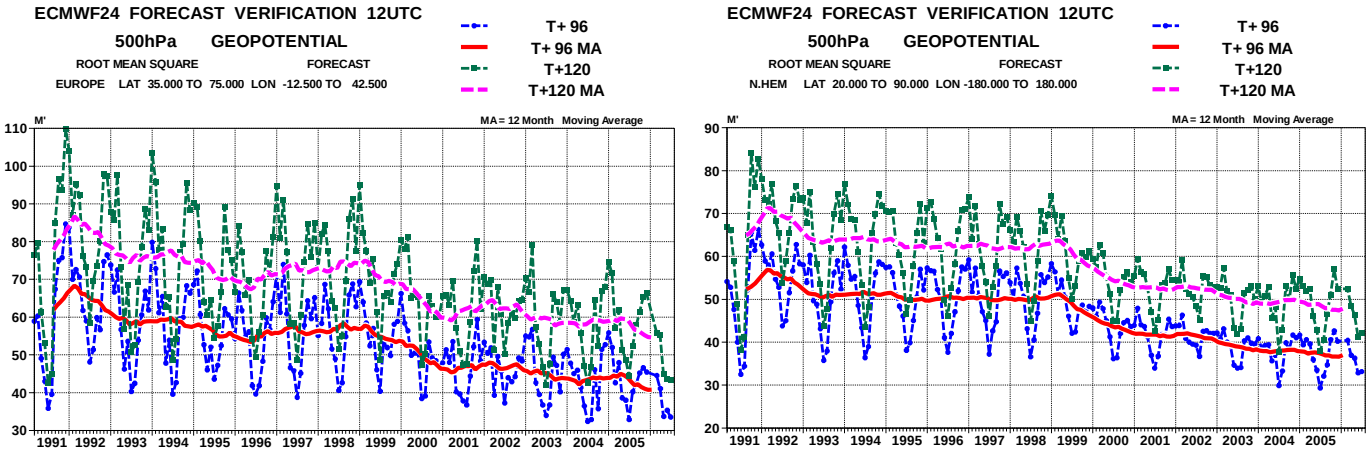


Figure 5: Consistency of the 500hPa height forecasts over Europe (left panel) and northern extratropics (right panel). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24h apart, for 96-120h (blue) and 120-144h (green). 12-month moving average scores are also shown. Last month is July 2006.

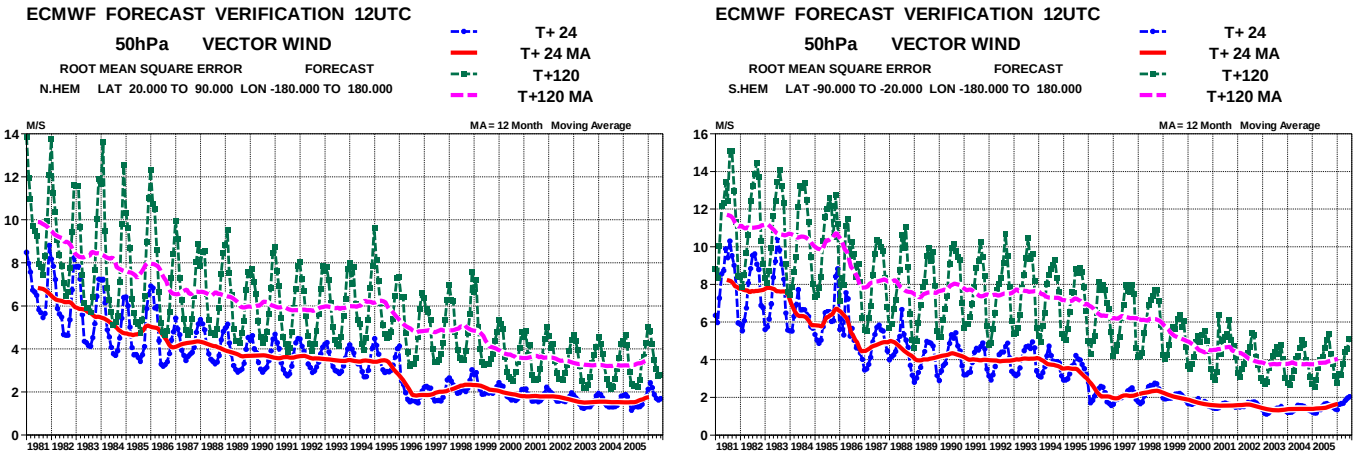
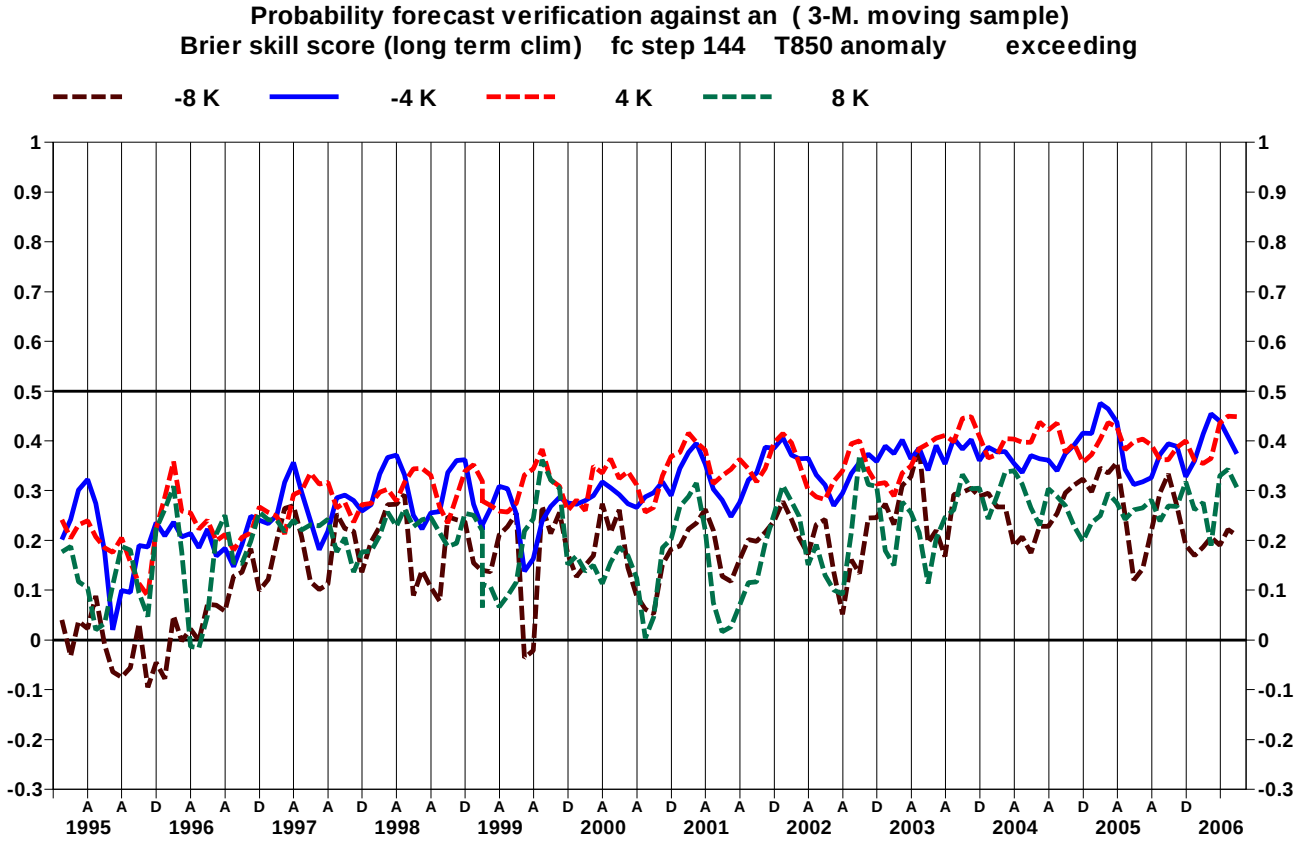


Figure 6: Model scores in the extratropical northern (left) and southern (right) hemisphere stratosphere (RMS vector wind error at 50hPa for 1-day and 5-day forecasts)



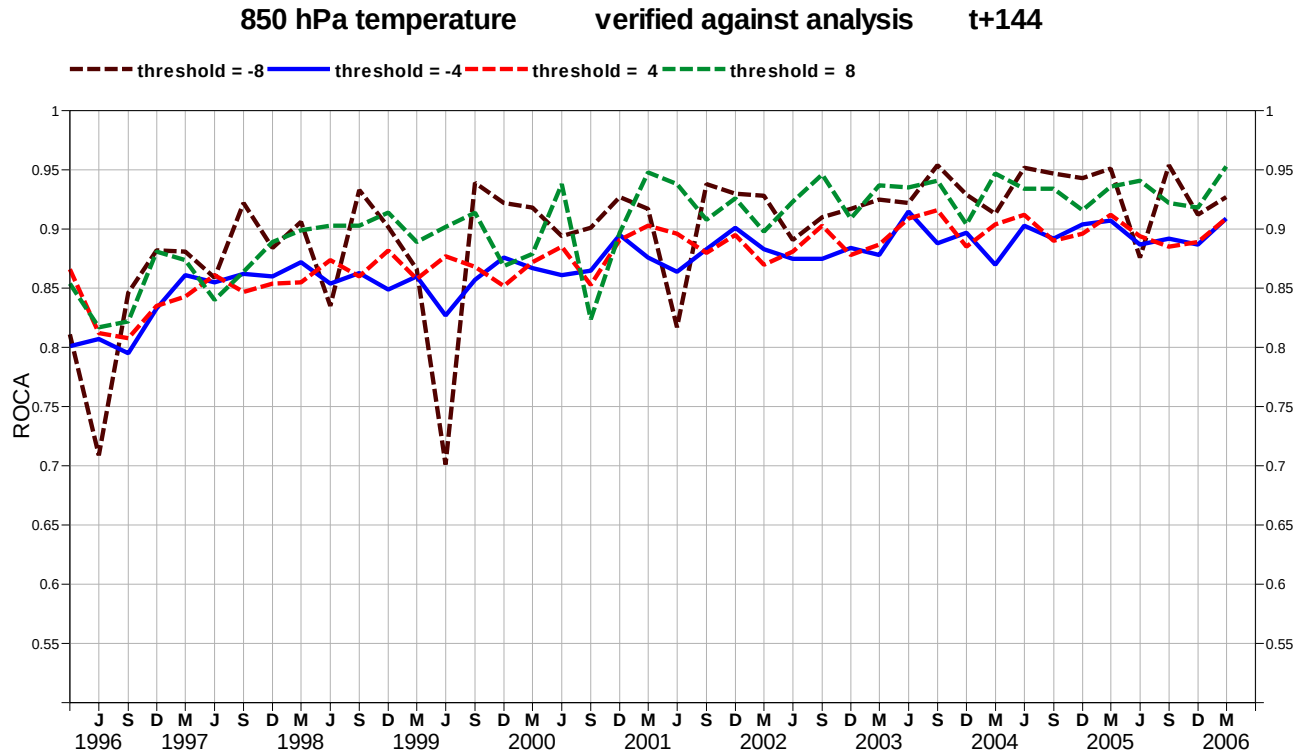


Figure 7: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA, bottom) for Europe for 144h EPS forecasts of 850hPa temperature anomalies exceeding a range of thresholds.

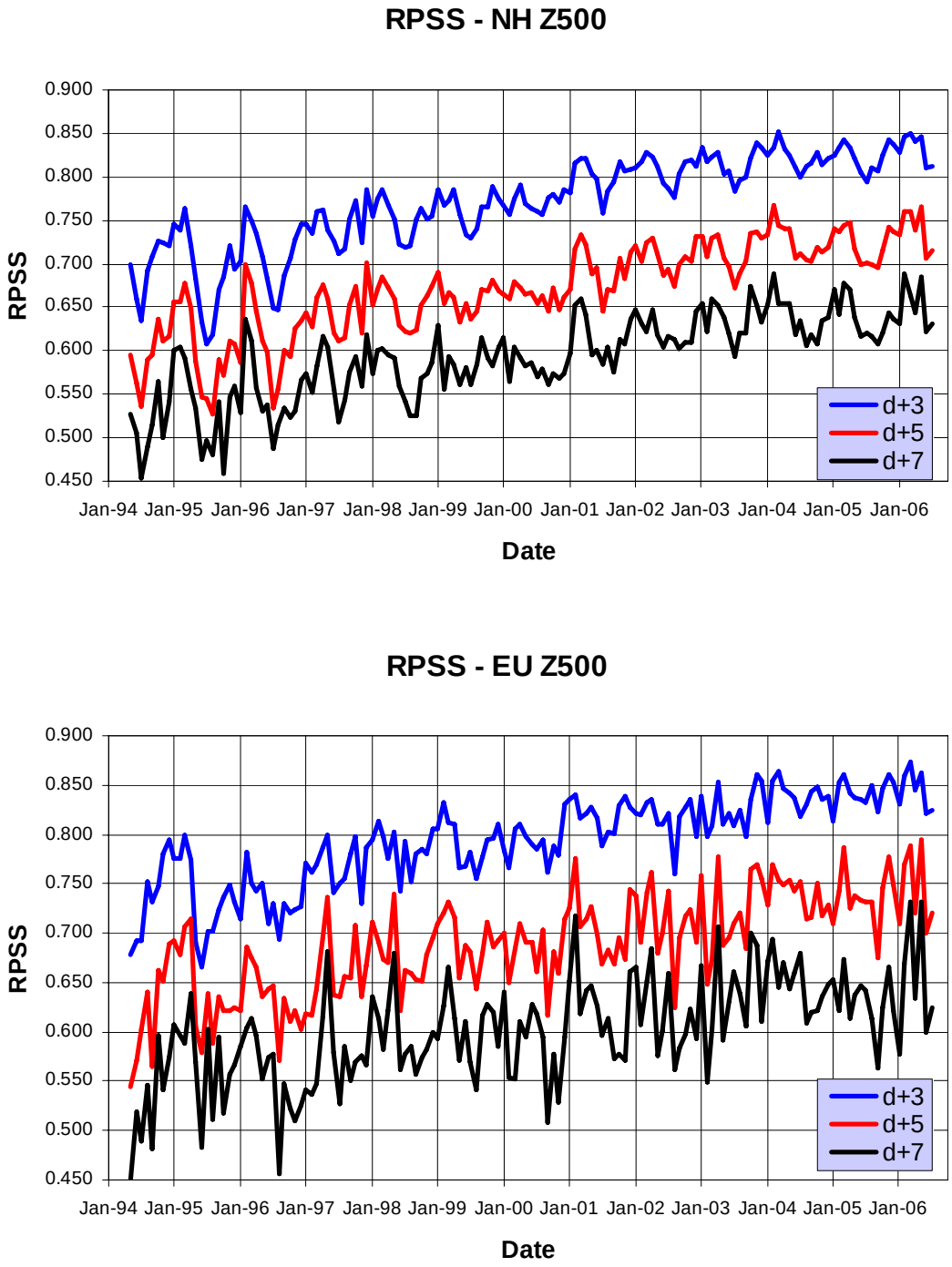


Figure 8: Time series of Ranked Probability Skill Score for EPS forecasts of 500 hPa geopotential height at day 3, 5 and 7 for the northern hemisphere extratropics (top) and Europe (bottom).

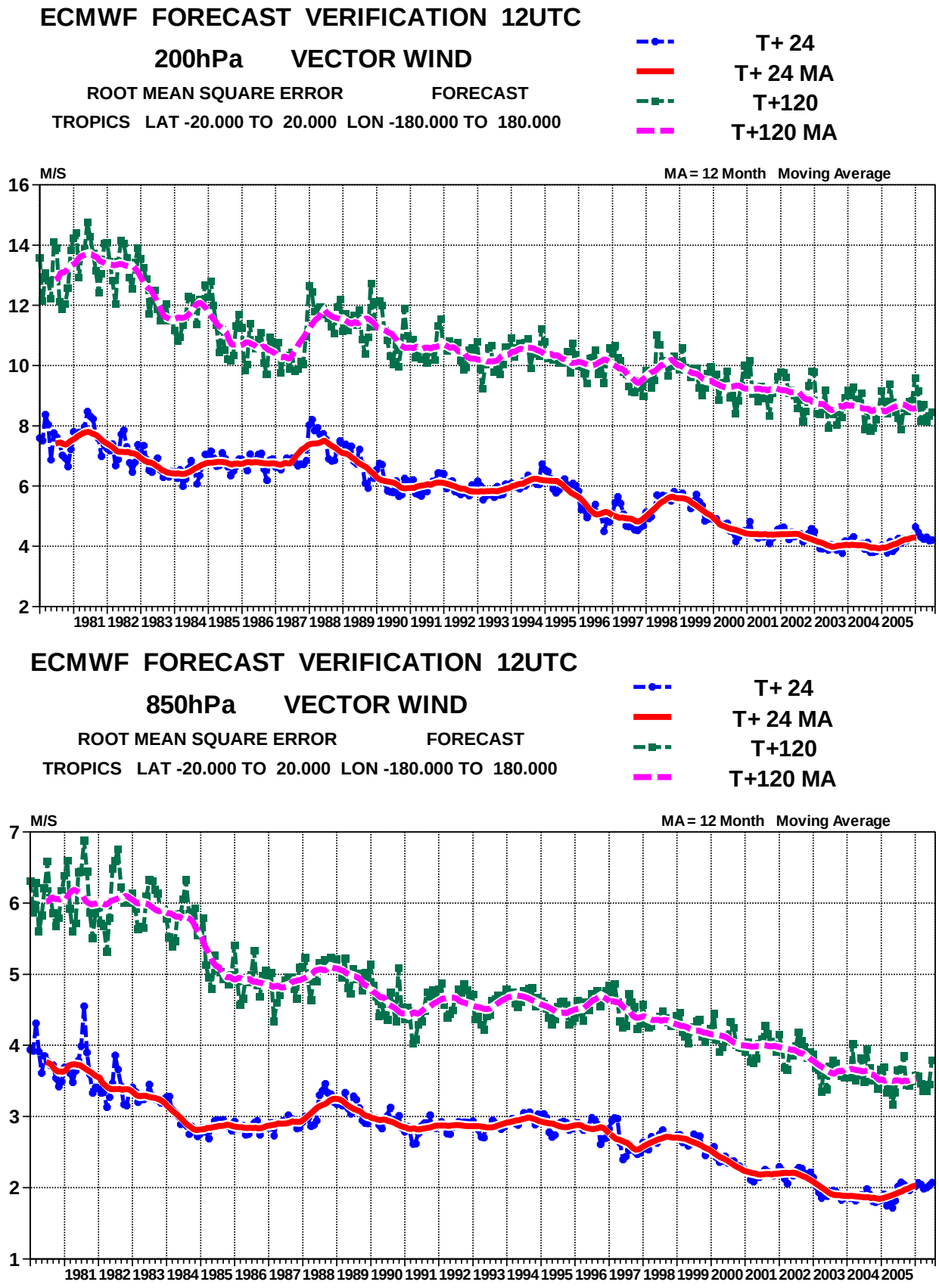


Figure 9: Model scores in the tropics (root mean square vector wind errors at 200hPa and 850hPa for 1-day and 5-day forecasts). Monthly mean and 12-month running mean.

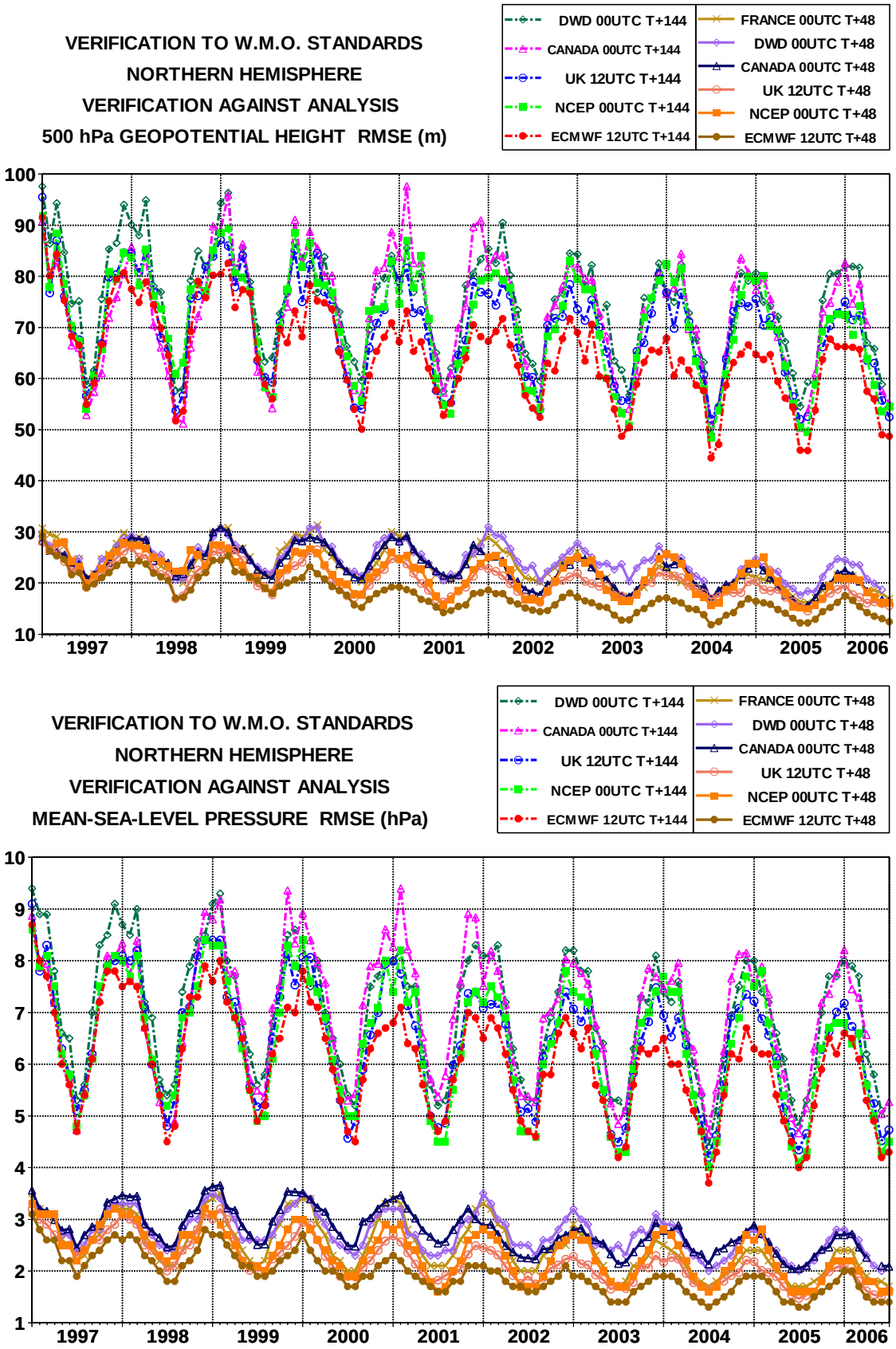
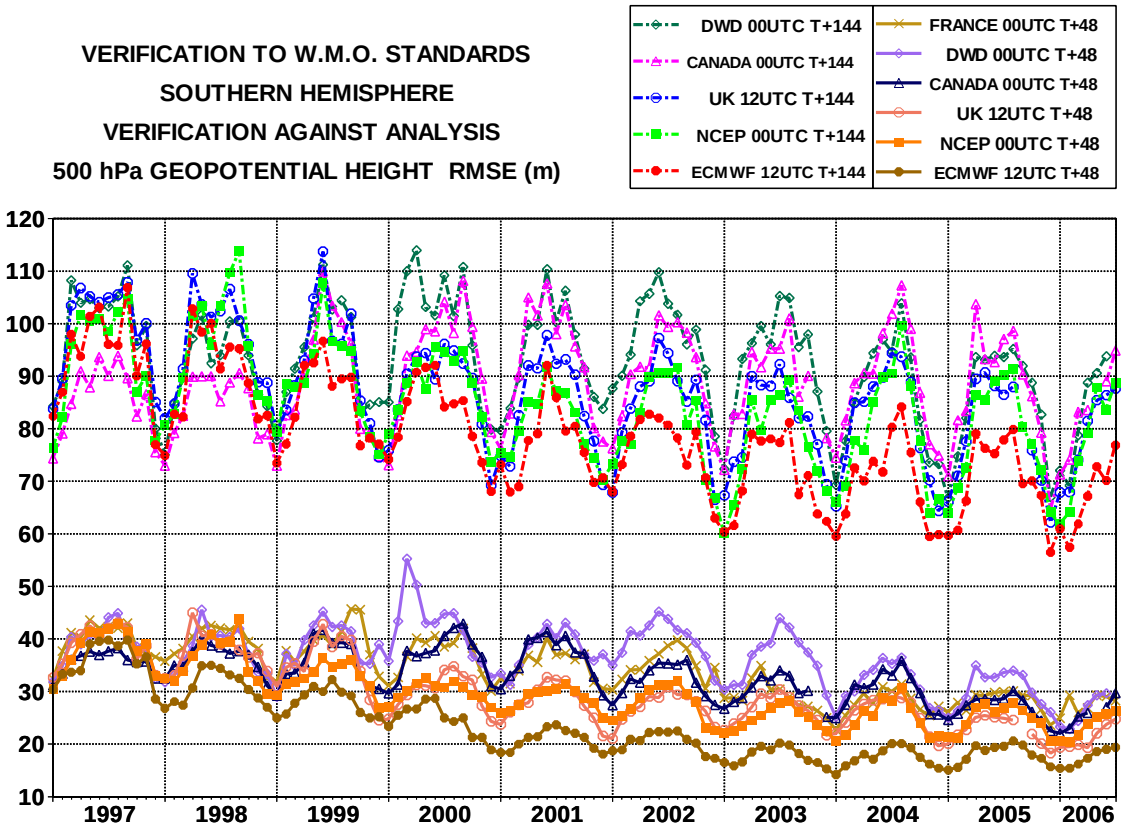


Figure 10: WMO/CBS exchanged scores (RMS error over northern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).



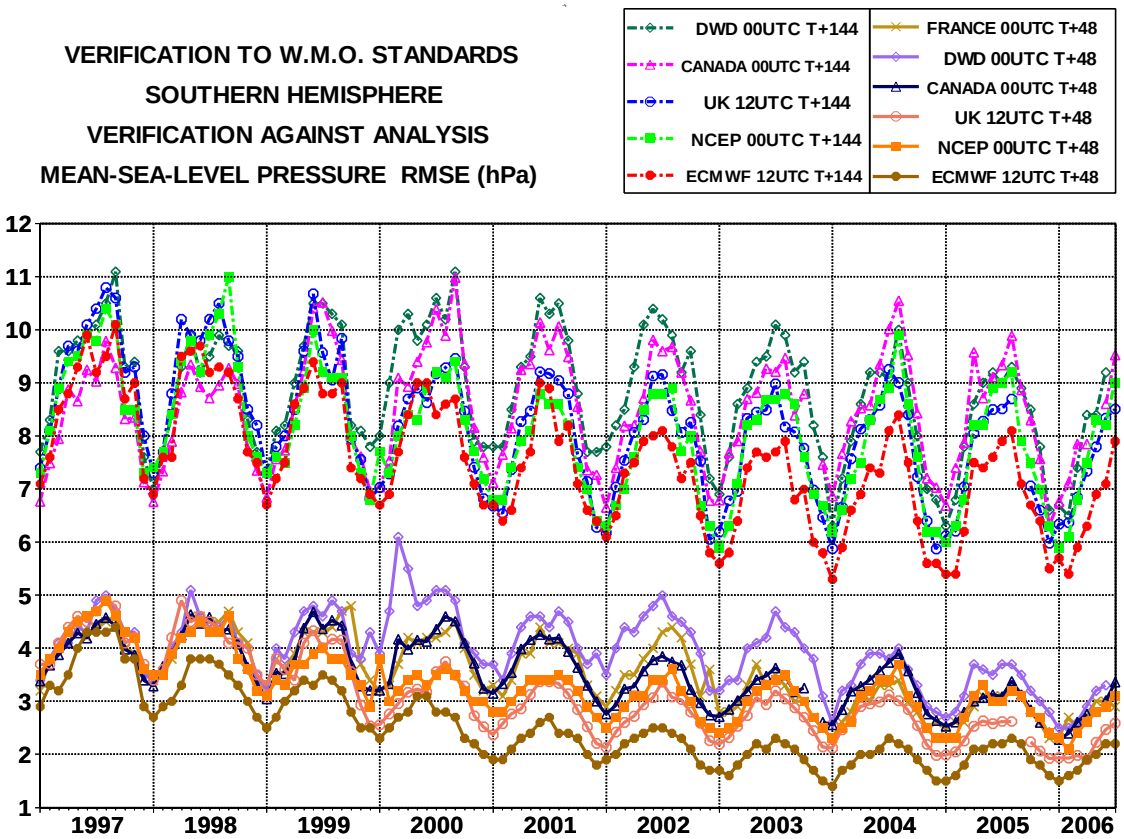
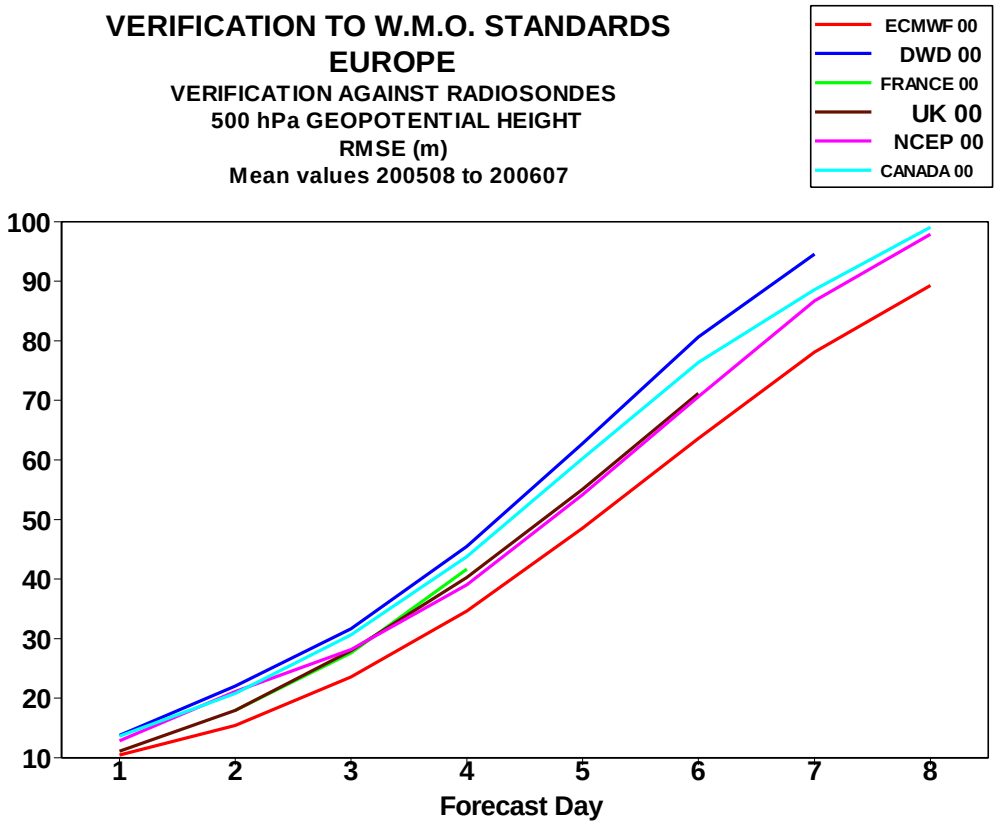


Figure 11: WMO/CBS exchanged scores (RMS error over southern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).



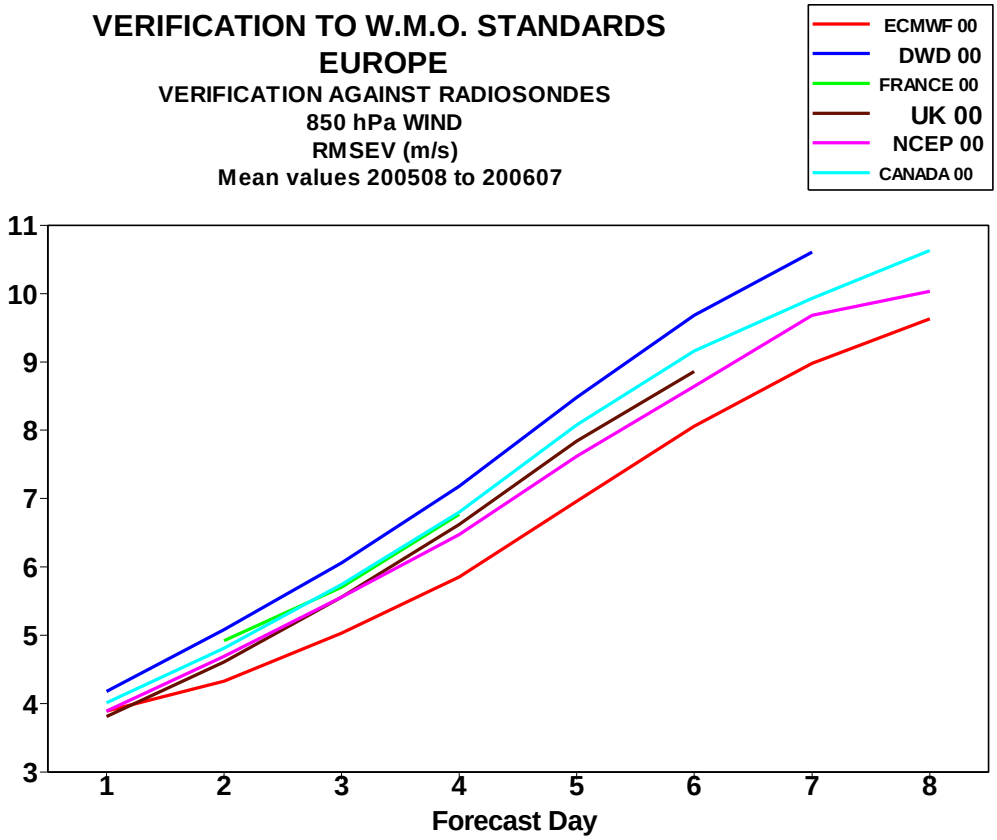
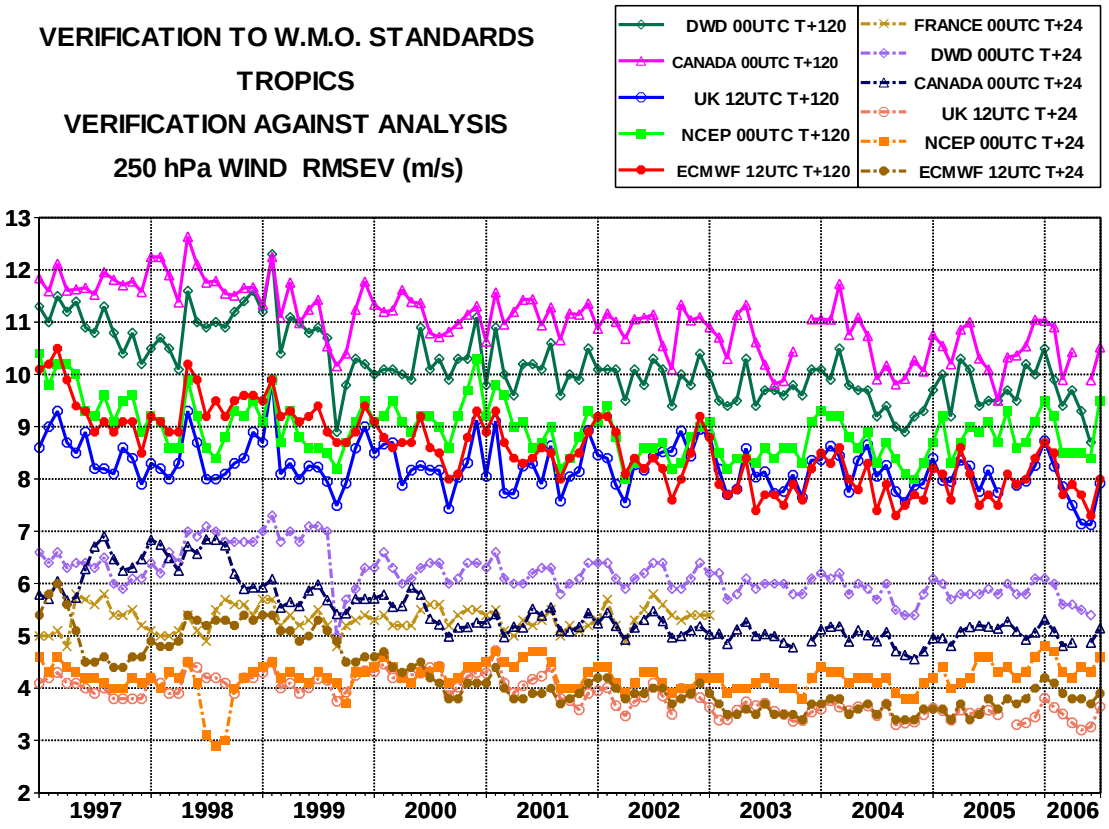


Figure 12: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)



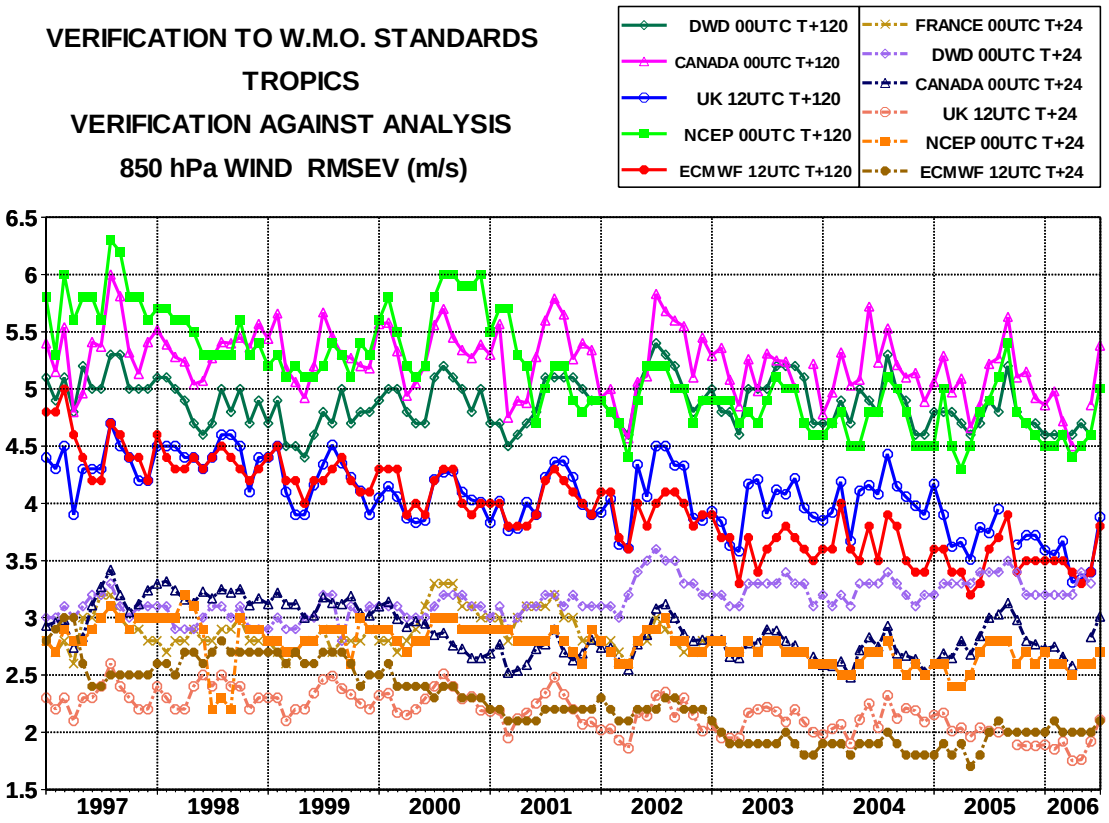


Figure 13: WMO/CBS exchanged scores (RMS vector error over the tropics, 250hPa and 850hPa wind forecast for day 1 and day 5).

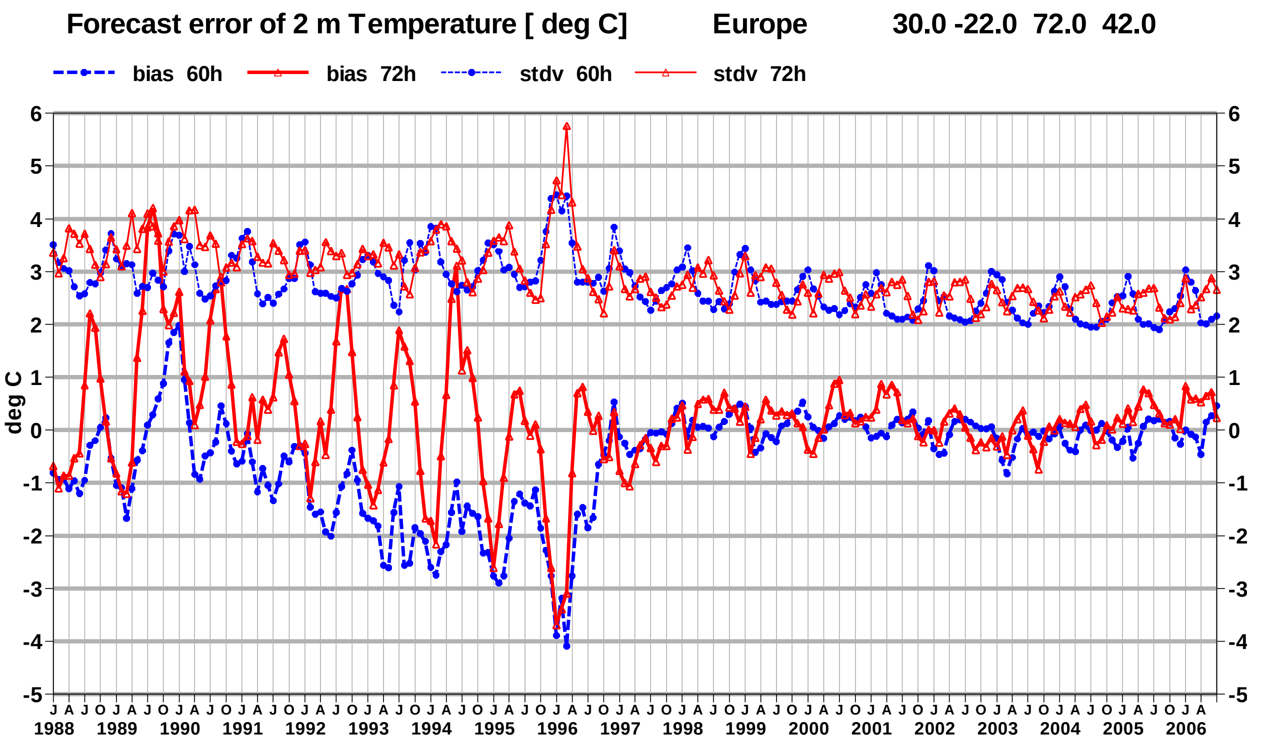


Figure 14: Verification of 2 metre temperature forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

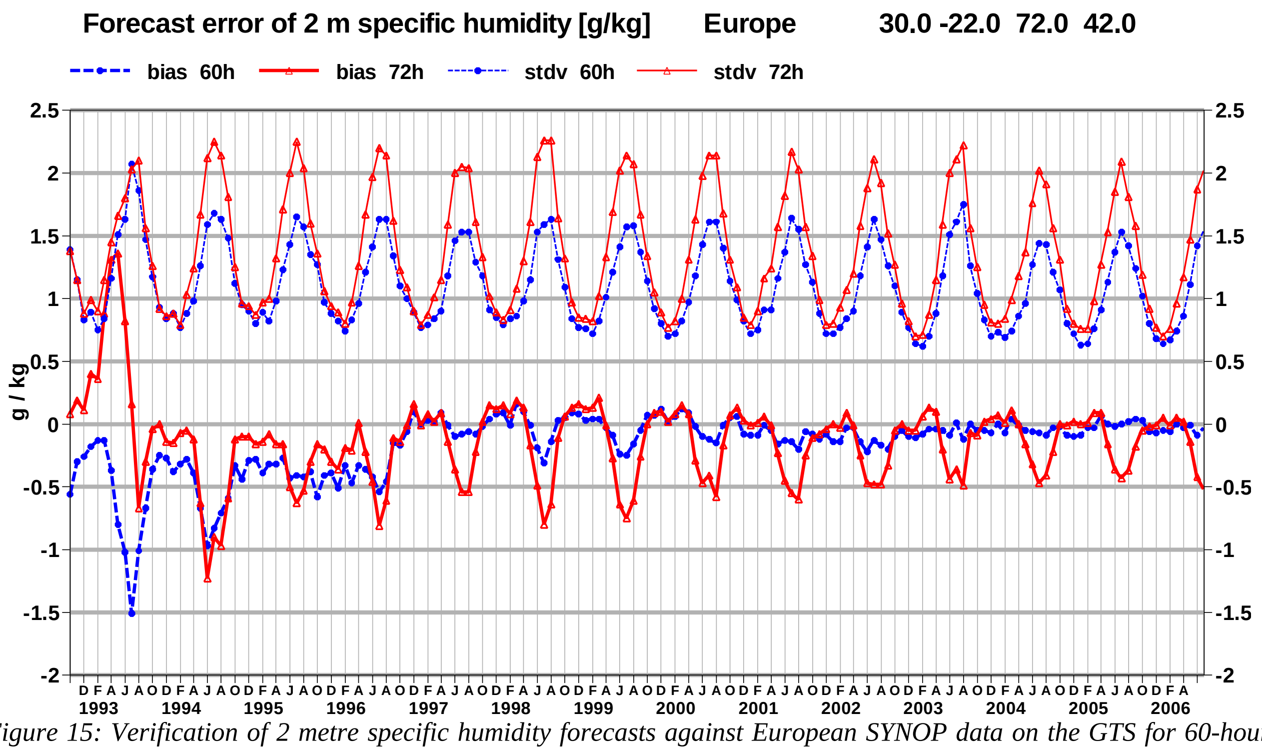


Figure 15: Verification of 2 metre specific humidity forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

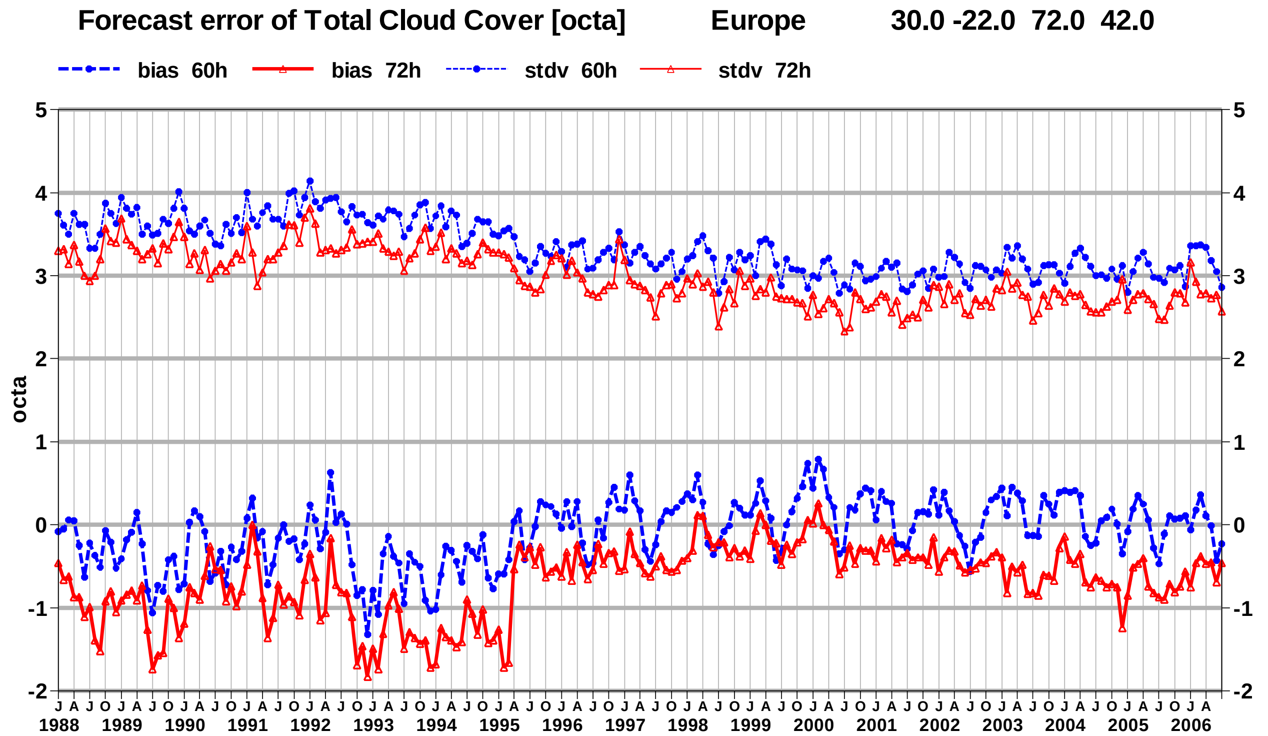


Figure 16: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

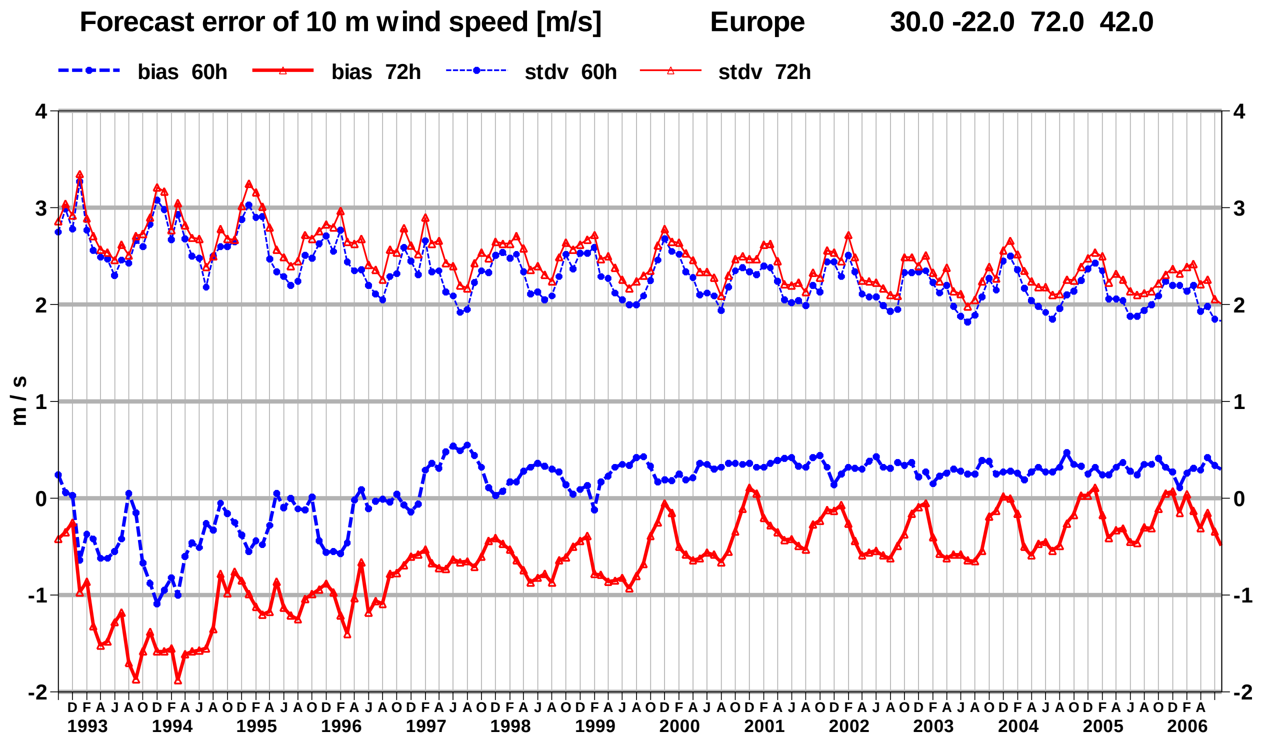


Figure 17: Verification of 10-metre wind speed forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

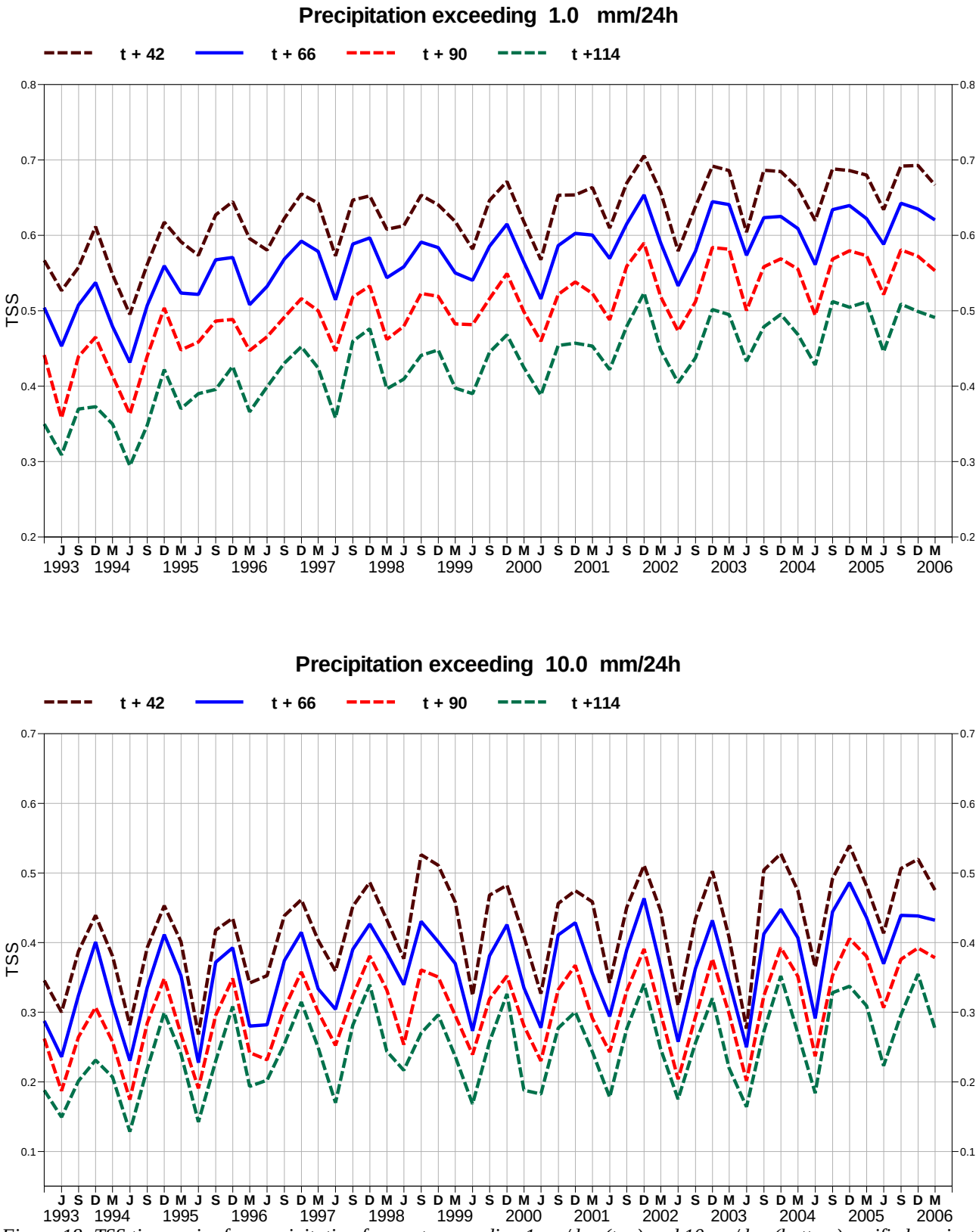
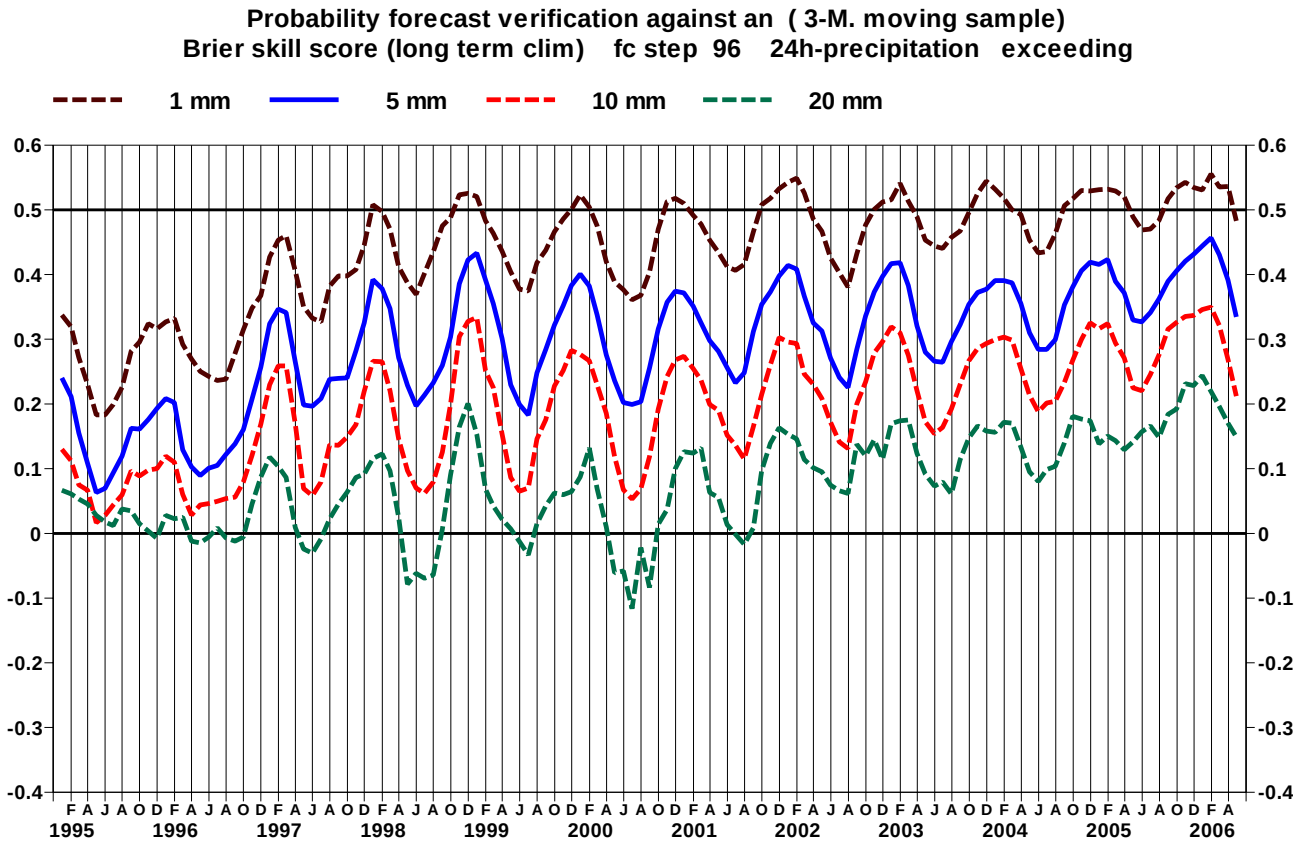


Figure 18: TSS time series for precipitation forecasts exceeding 1mm/day (top) and 10mm/day (bottom) verified against SYNOP data on the GTS for Europe. Curves are shown for the 24-hour accumulations up to 42, 66, 90, and 114 hours (from the forecasts starting at 12 UTC). 3-month mean scores (last point is March-May 2006).



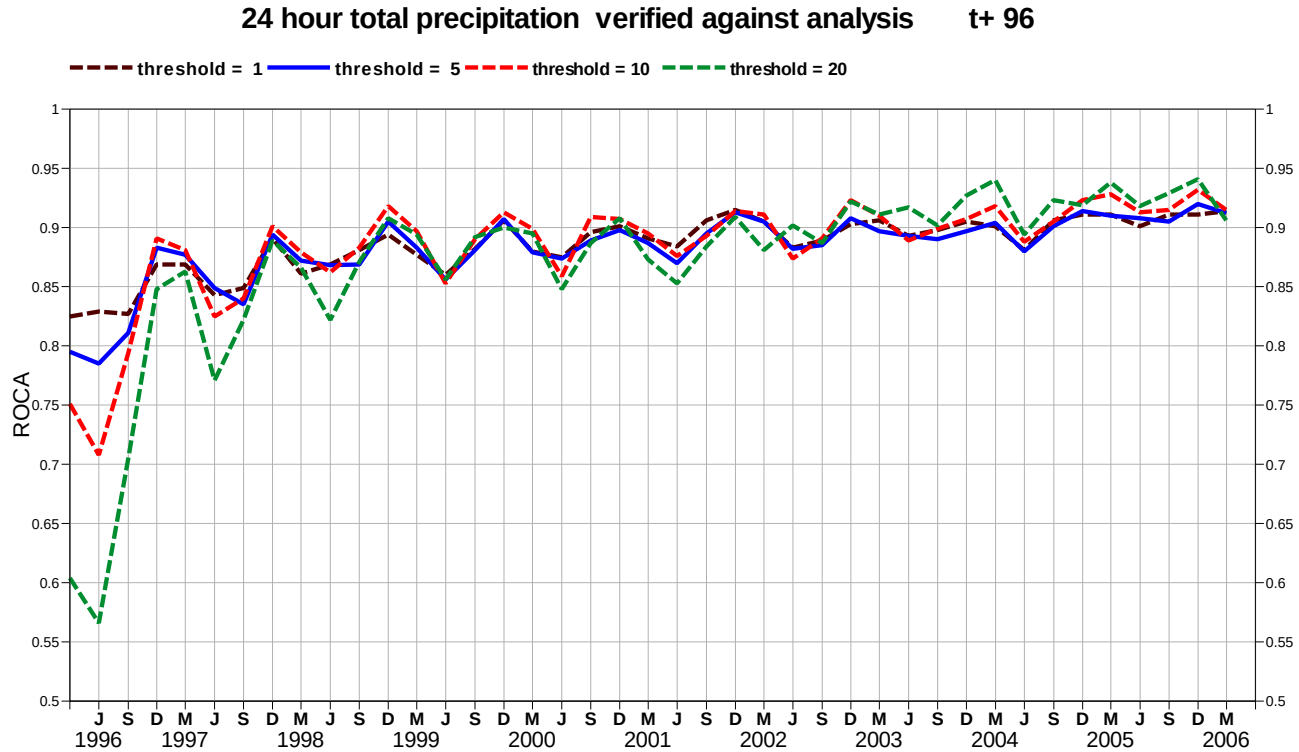


Figure 19: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA) for EPS probability forecasts of precipitation over Europe exceeding thresholds of 1, 5, 10 and 20 mm/day at day 4. The skill score is calculated for three-month running periods. The reference is the sample climate.

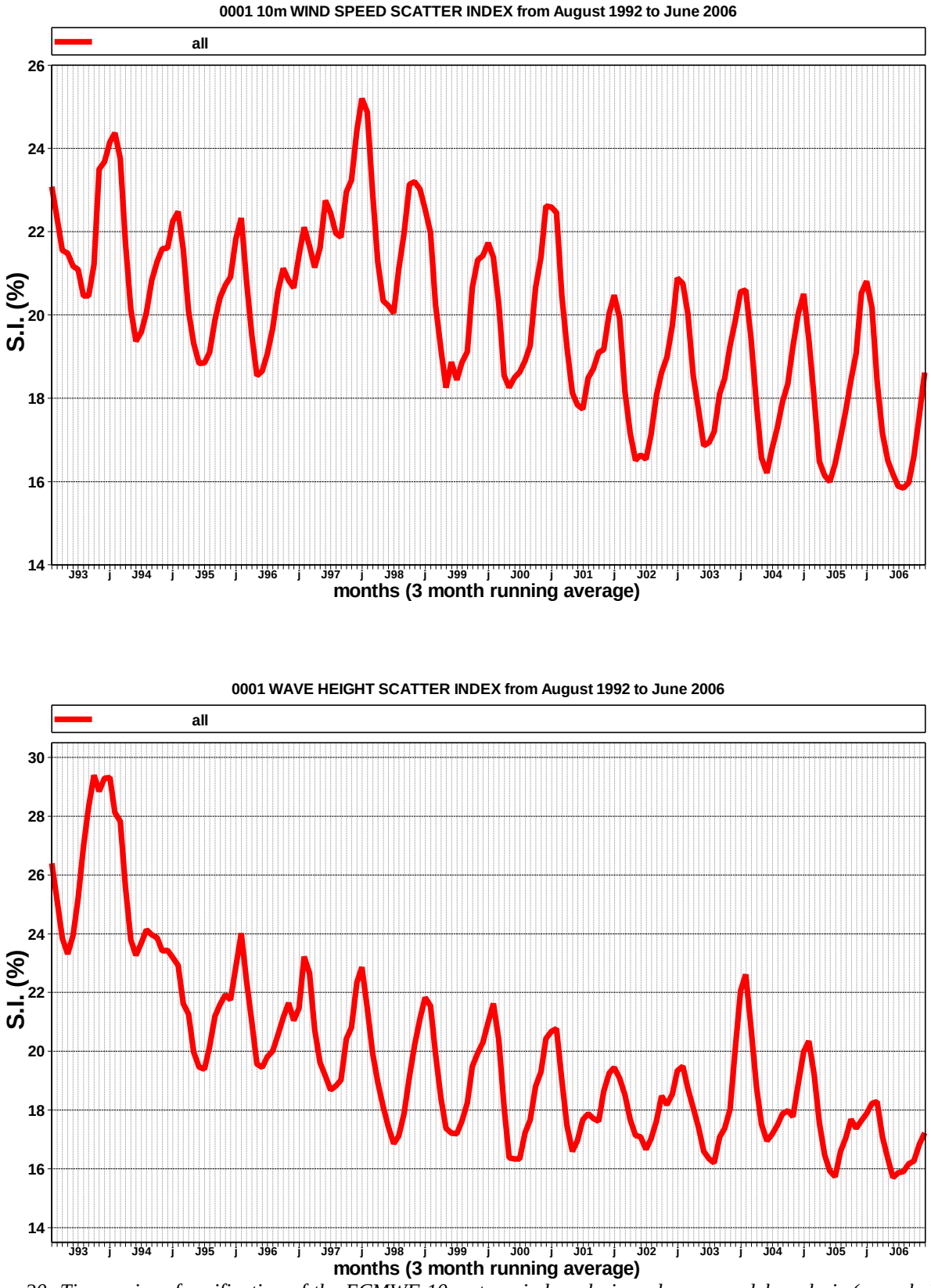
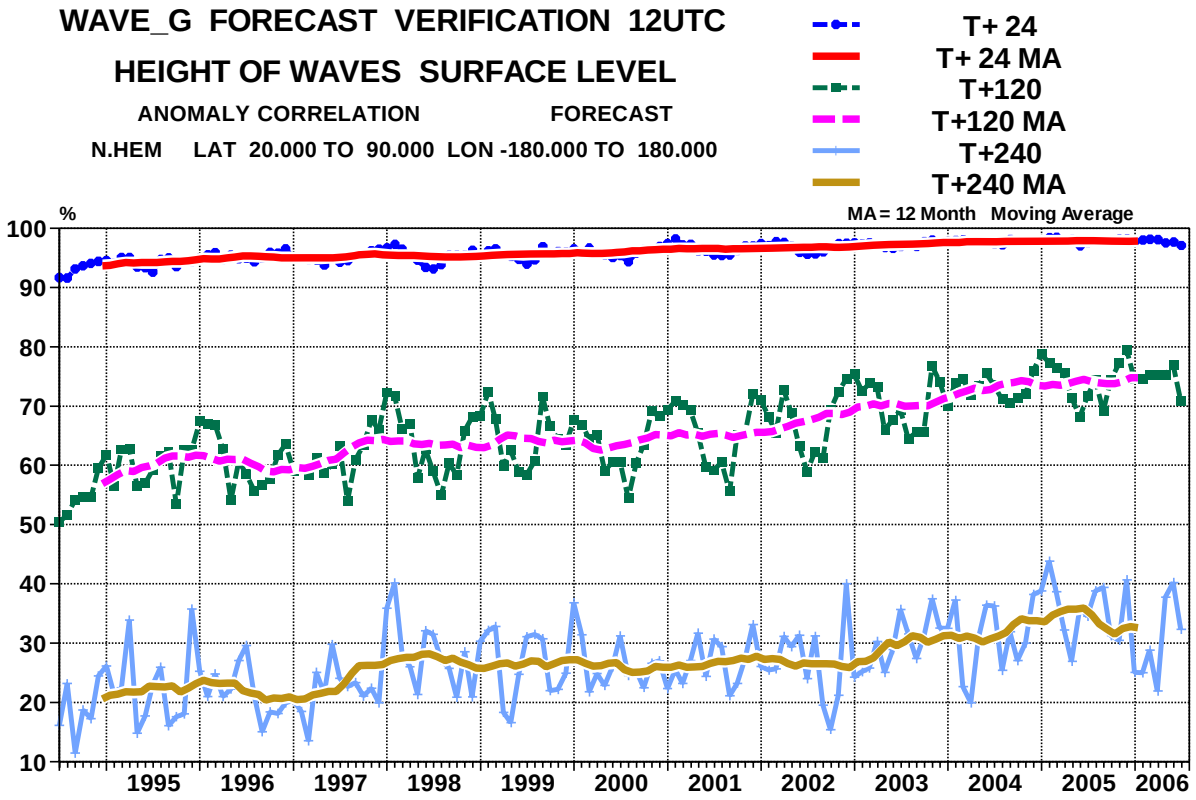


Figure 20: Time series of verification of the ECMWF 10 metre wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.



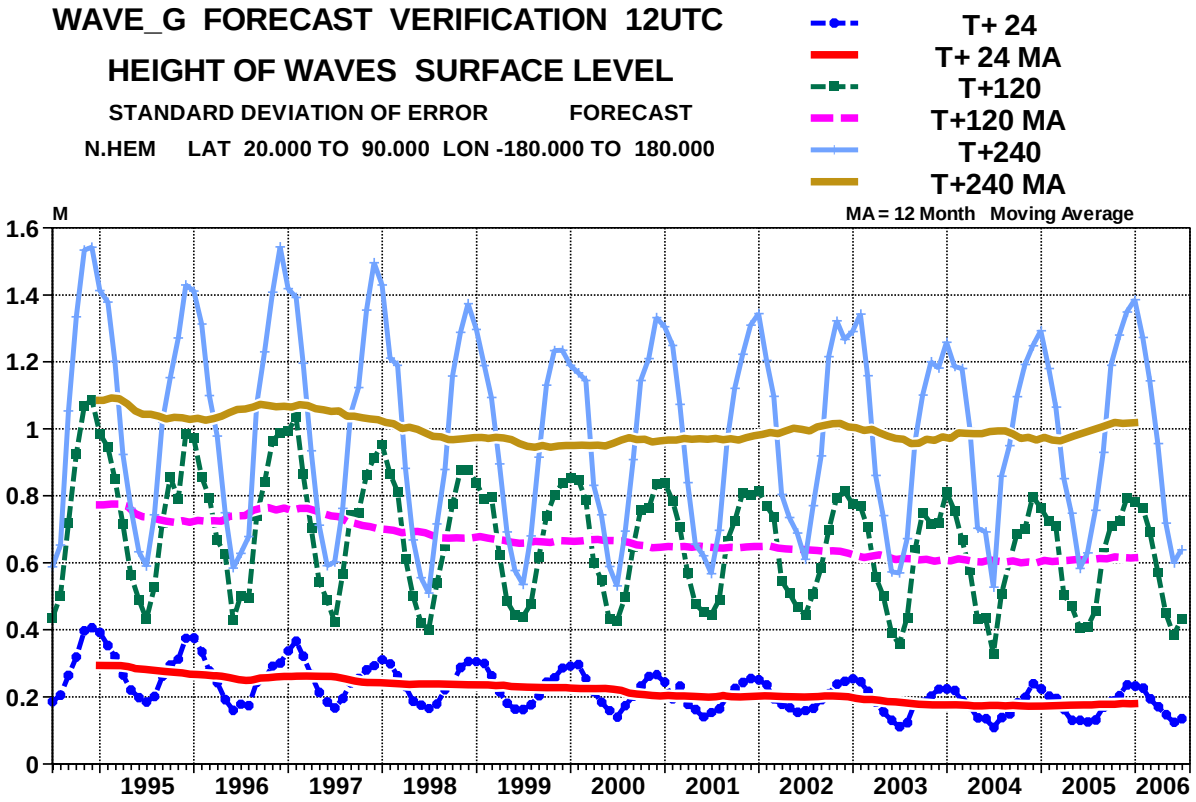
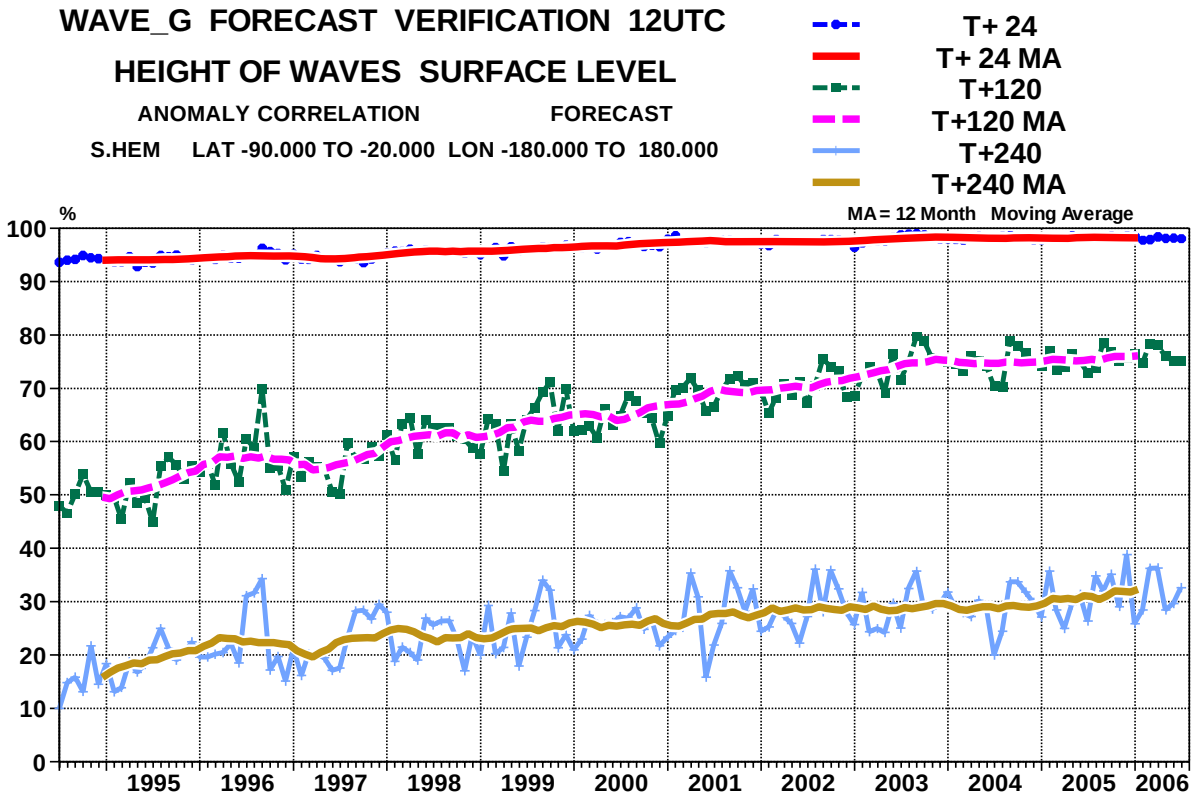


Figure 21: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (northern extratropics)



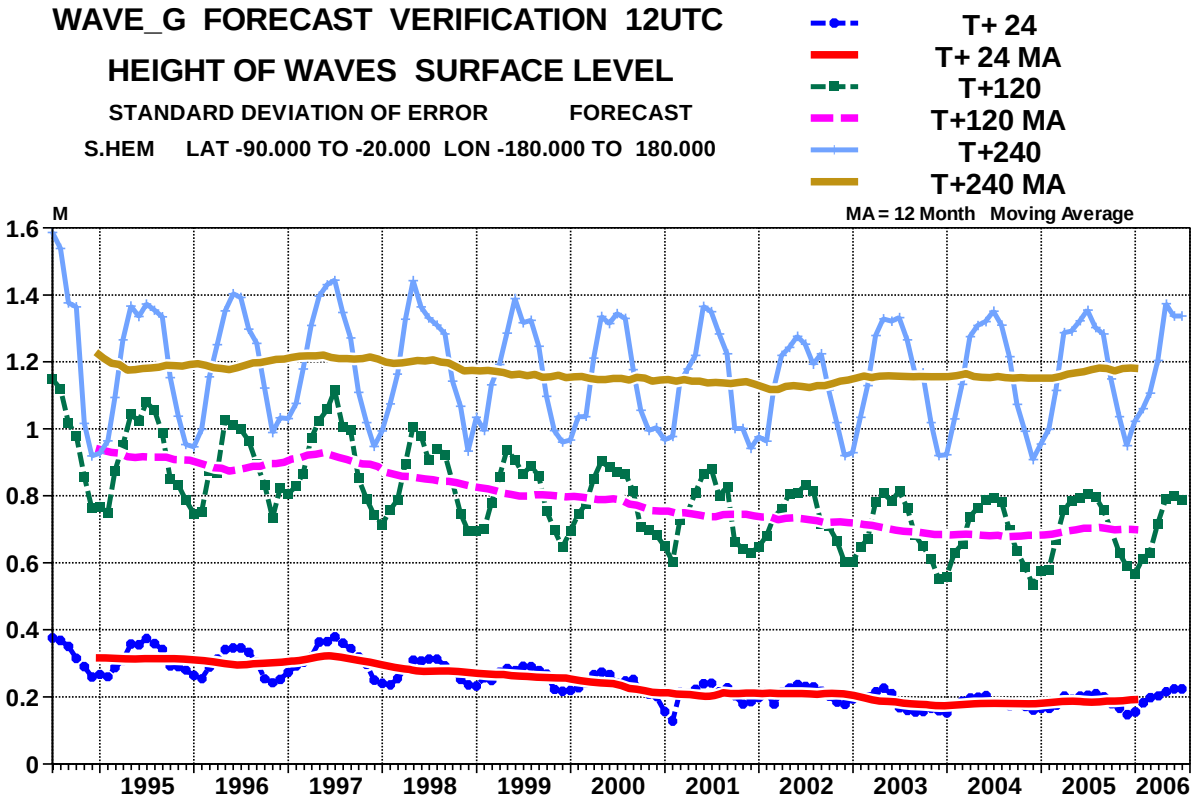


Figure 22: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (southern extratropics)

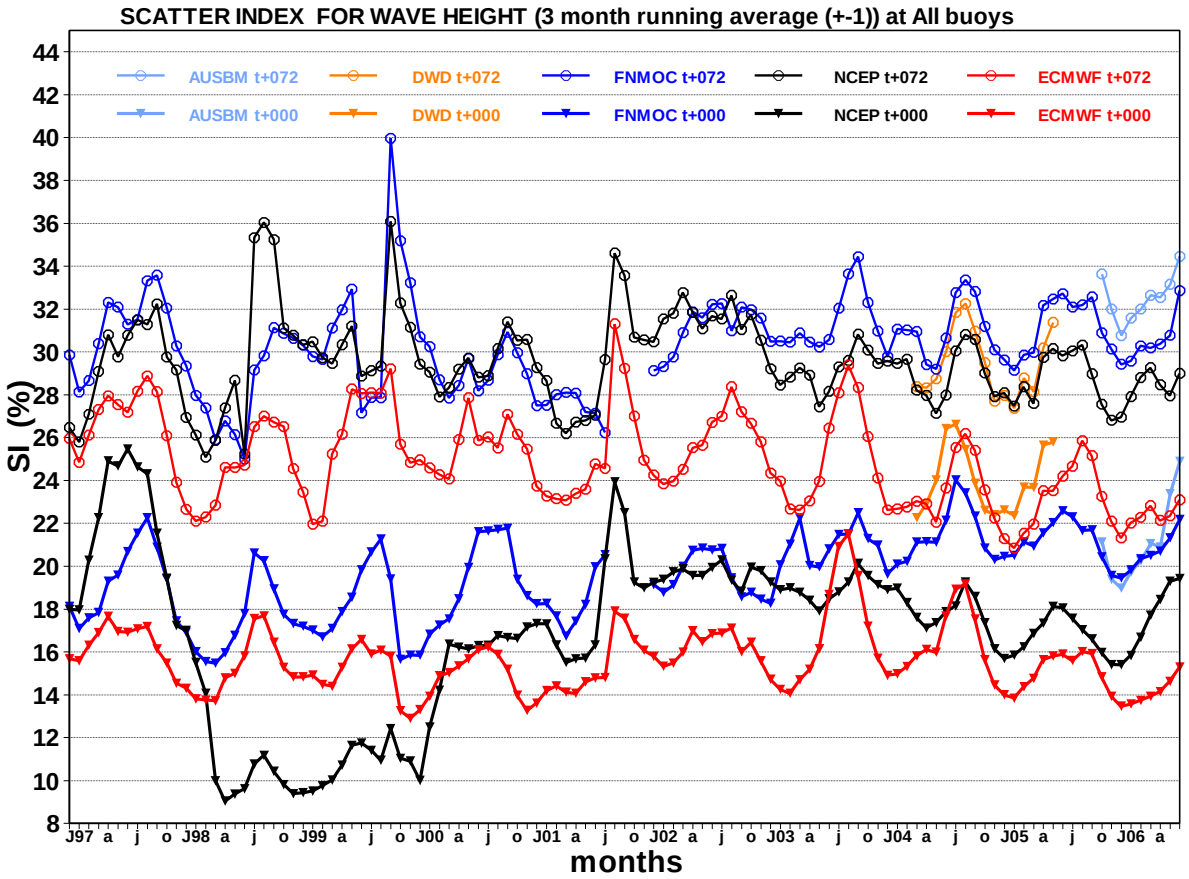


Figure 23: Verification of different model wave height forecasts using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value.

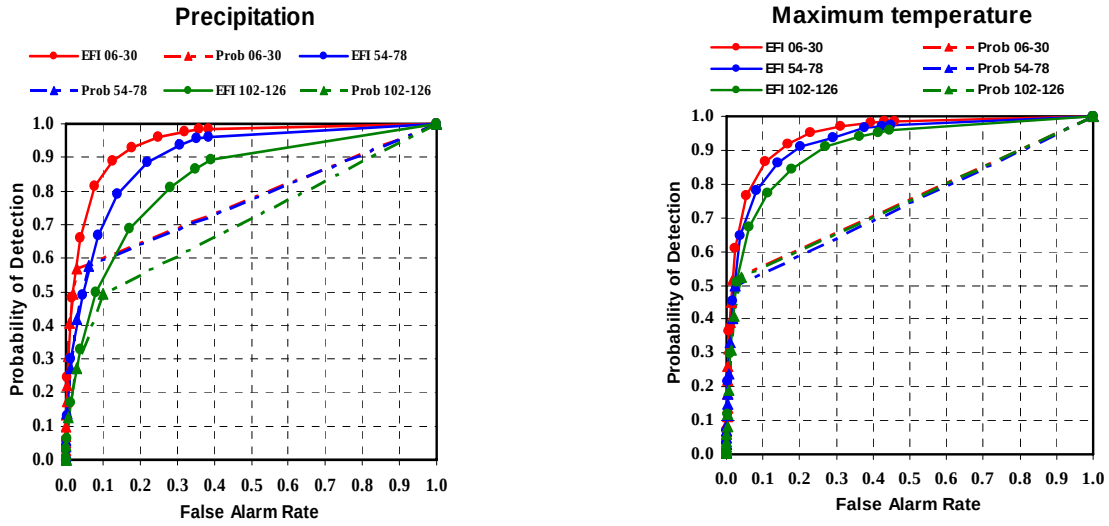


Figure 24: ROC verification of daily rainfall (left) and maximum temperature (right) exceeding the 99.5% threshold of the EUMETNET/ECSN station climatology. The performance using different thresholds of the EFI (solid lines) is compared to that using EPS probabilities of the event (dashed lines). Each colour is for a different forecast range: day 1 (red), day 3 (blue), day 5 (green). Sample contains all events over the period July 2005 to May 2006.

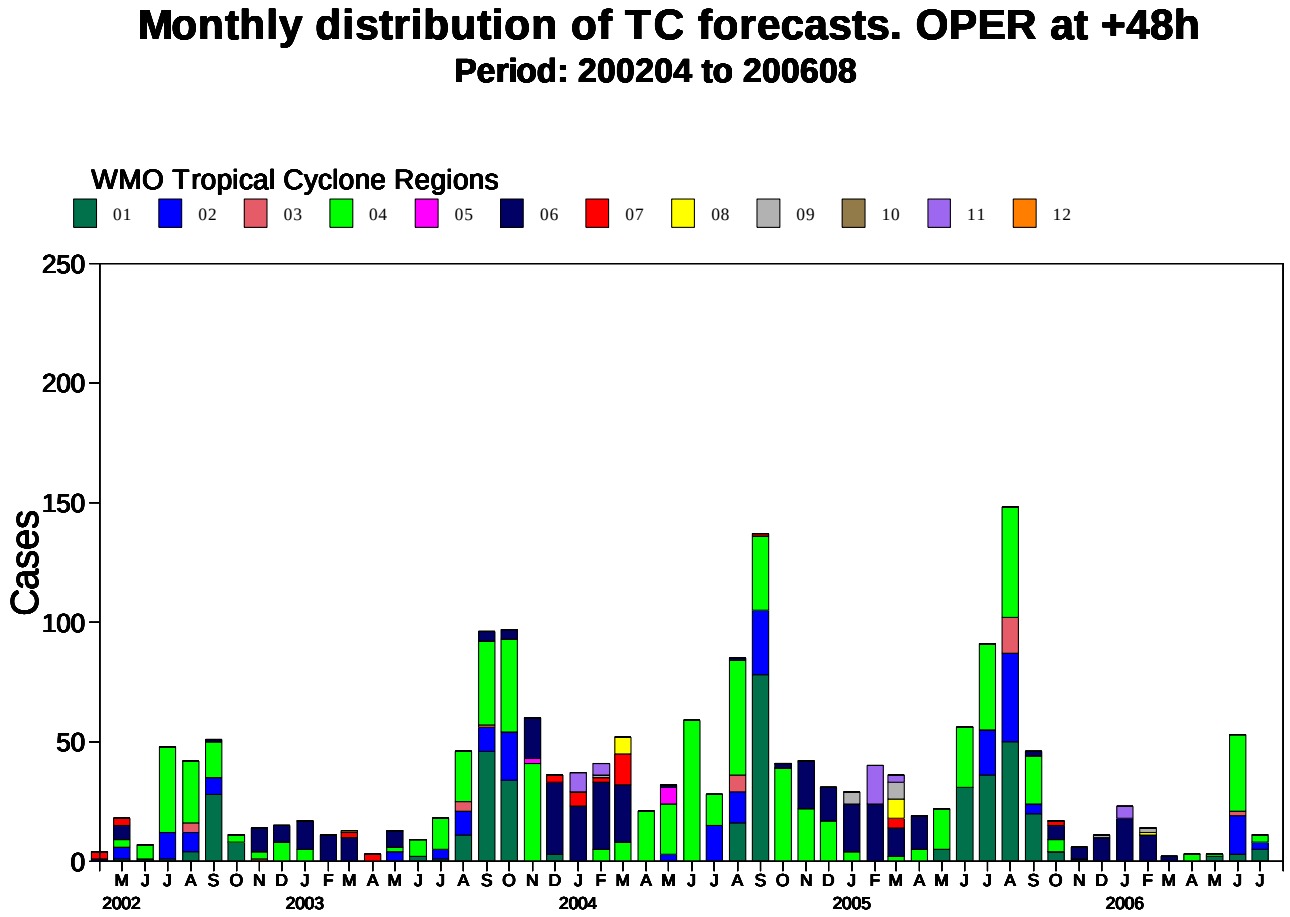
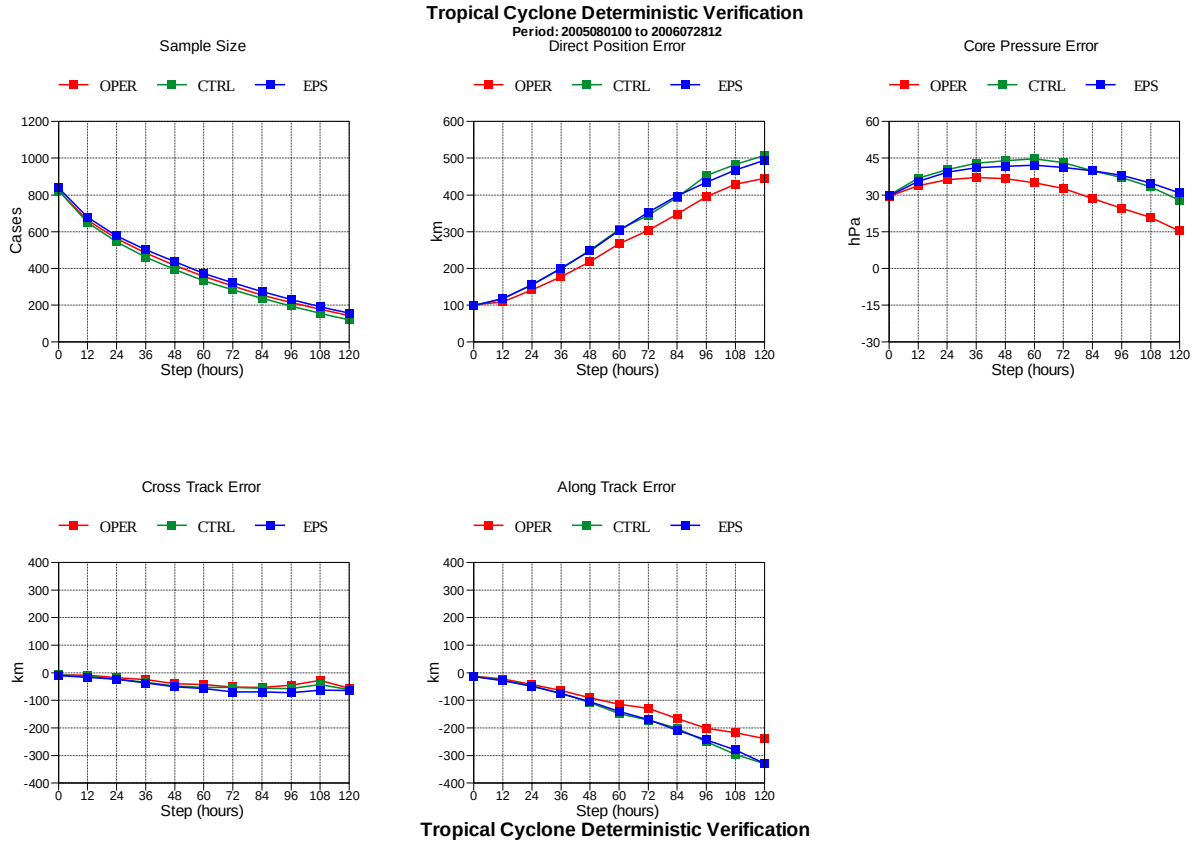


Figure 25: Number of tropical cyclones tracked by the operational deterministic T511/T799 day 2 forecast from January 2002 to July 2006. For each month, the number is split per WMO Tropical Cyclone region (1=NW Atlantic; 2=NE Pacific; 3=N Pacific; 4=NW Pacific; 5=N. Indian; 6= SW Indian; 7=SE Indian; 8/9/10=SW Pacific; 11/12=S. Pacific). Both 00 and 12UTC forecasts are tracked.

a)



b)

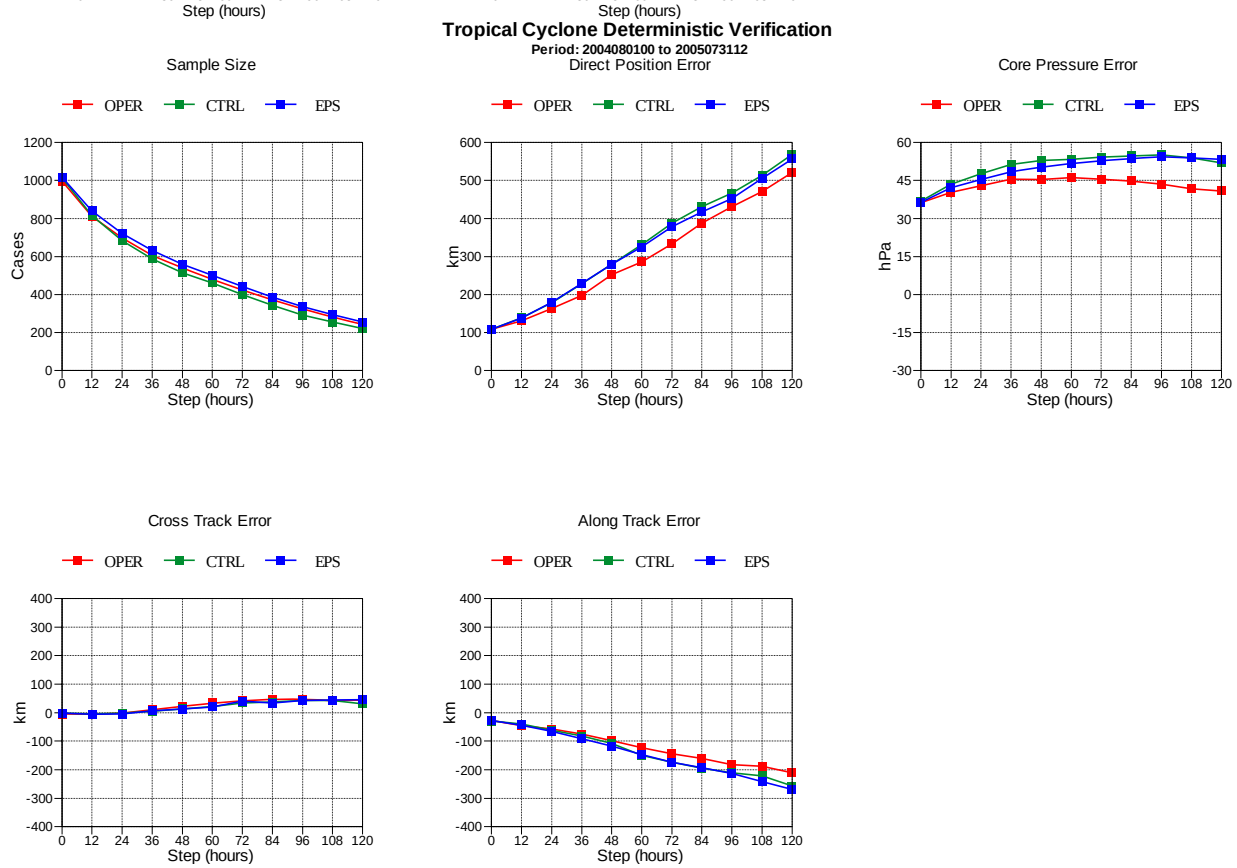


Figure 26 a): Verification of Tropical Cyclone forecasts from the operational deterministic T511/T799 forecast (red), EPS Control (green) and mean position/ intensity averaged among all cyclones tracked in each member of the ensemble forecast (blue) for the period August 2005 to July 2006. b): As a) but for the same period in 2004-2005.

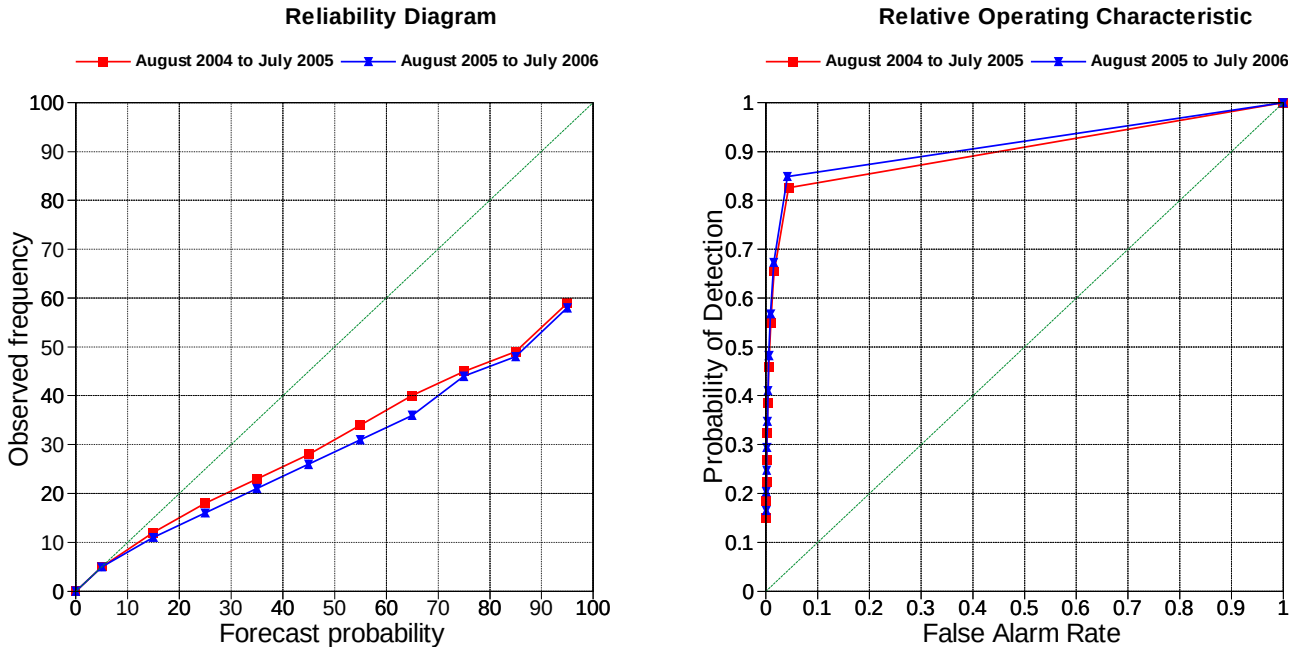


Figure 27: Probabilistic verification of TC strike probability forecasts for August 2005-July 2006 (blue) and August 2004-July 2005 (red). Left: reliability diagram (the closer to the diagonal the better); right: ROC diagram (the closer to the upper left corner the better).

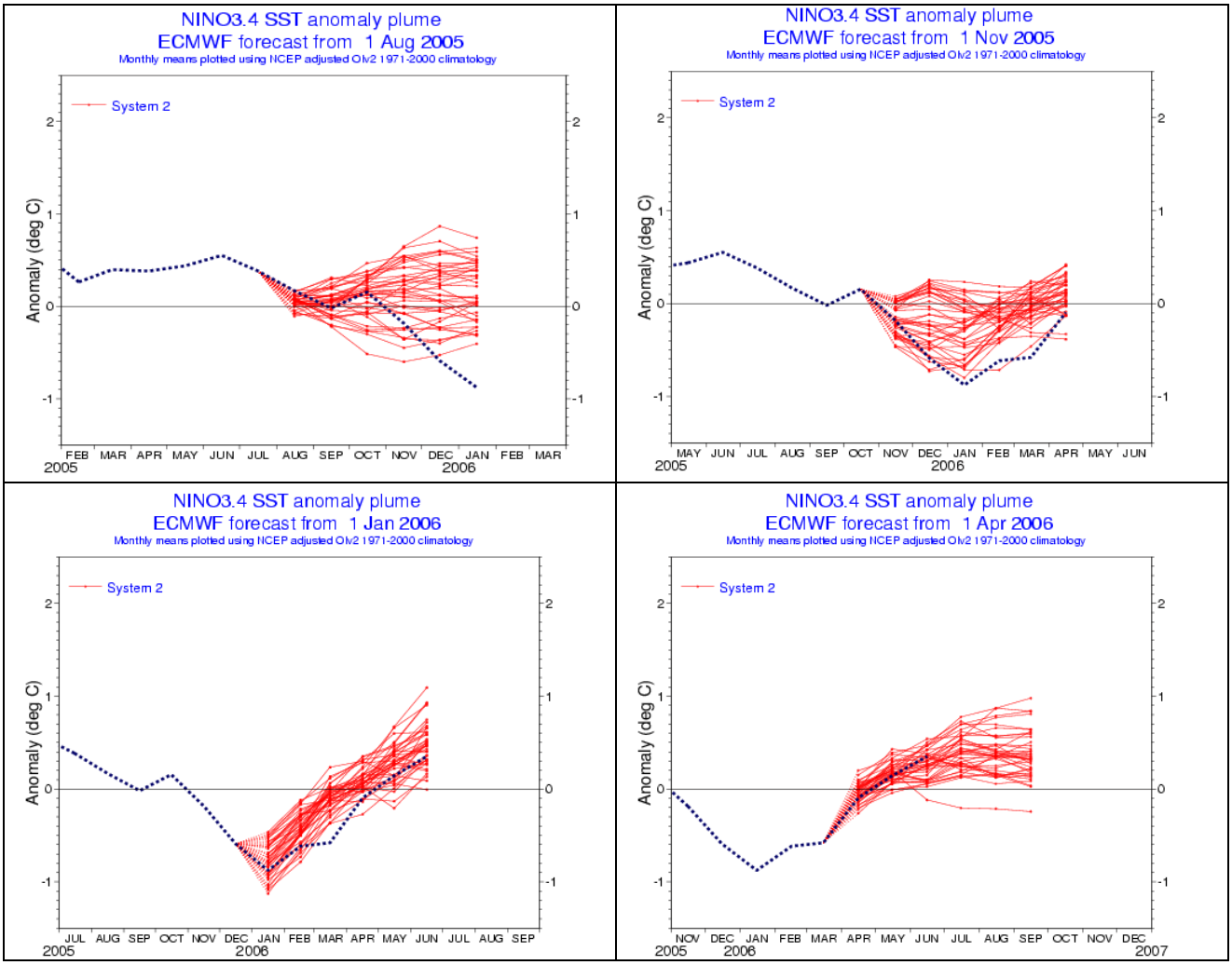


Figure 28: Plot of forecasts of Nino-3.4 SST anomalies from four start dates: August 2005, November 2005, January 2006 and April 2006. The red lines represent the 40 ensemble members. The dashed blue line represents subsequent verification.

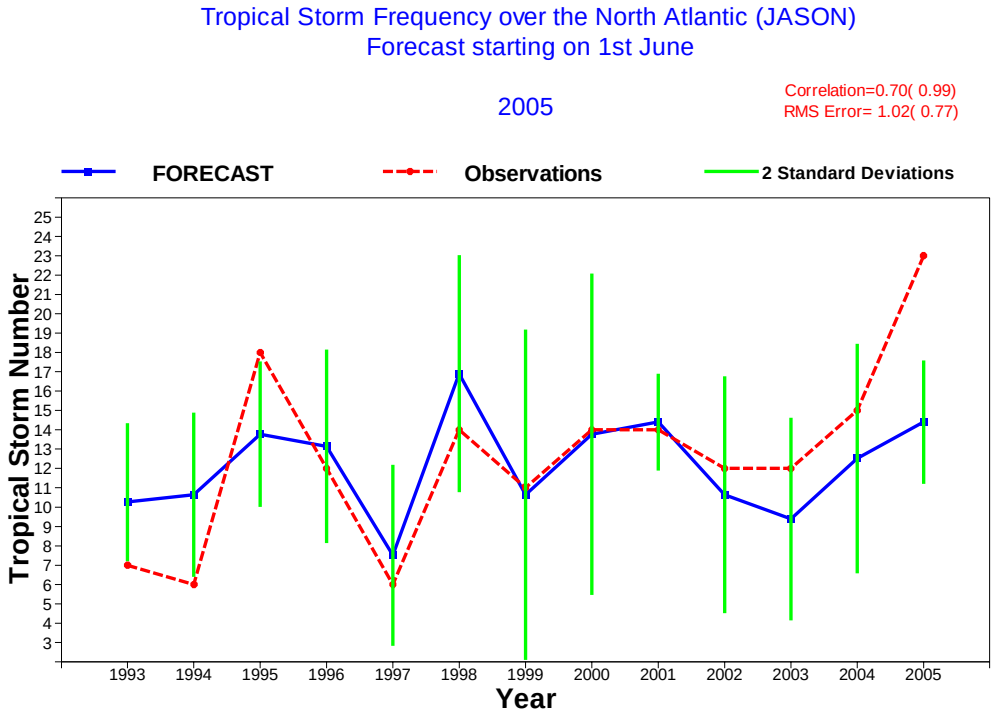


Figure 29: Time series of tropical storm frequency over the North Atlantic for the July to November season from 1993 to 2005. The red dashed line represents the observed values and the blue line the ensemble mean of the seasonal forecast started in June.

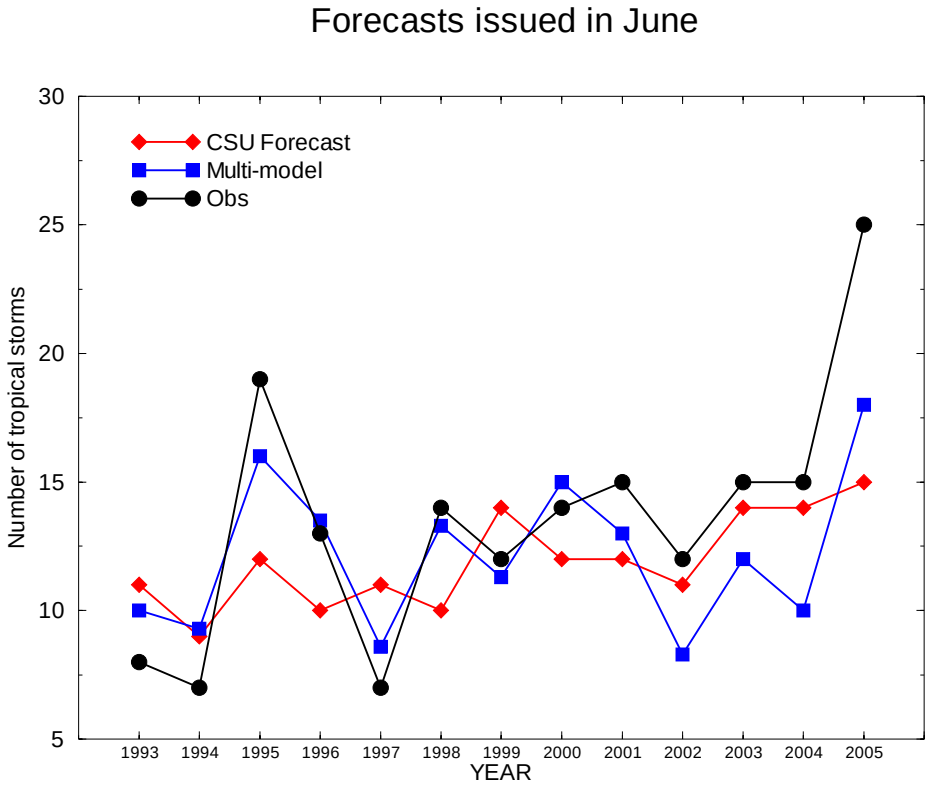


Figure 30: Time series of tropical storm frequency over the North Atlantic for the June to November season from 1993 to 2005. The black line represents the observed values, the blue line the EUROSIP multi-model ensemble mean forecast started in June, and the red line the empirical CSU forecast.

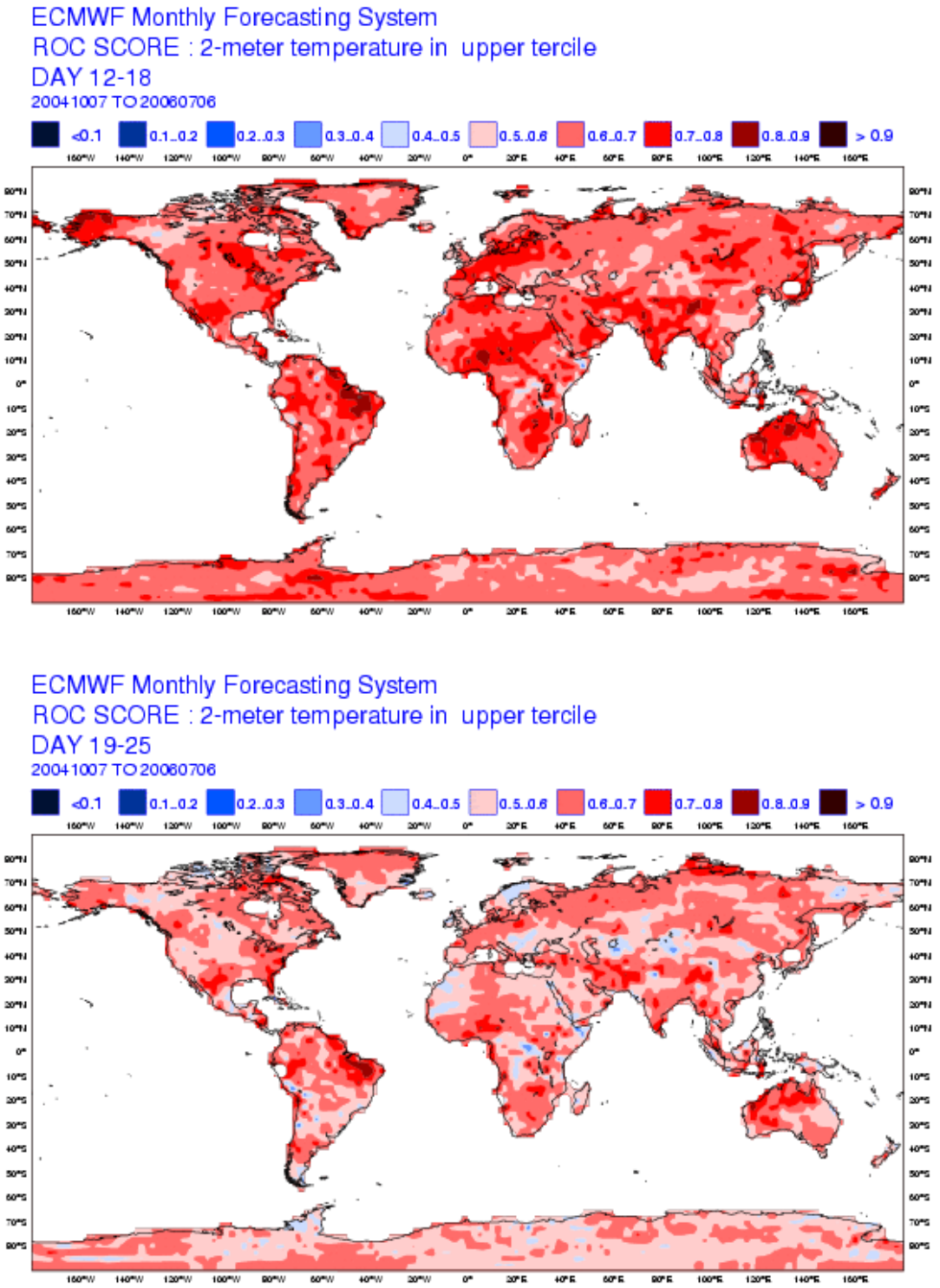


Figure 31: Spatial distribution of ROC area scores for the probability of 2m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 6 July 2006 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicated positive skill compared to climate.

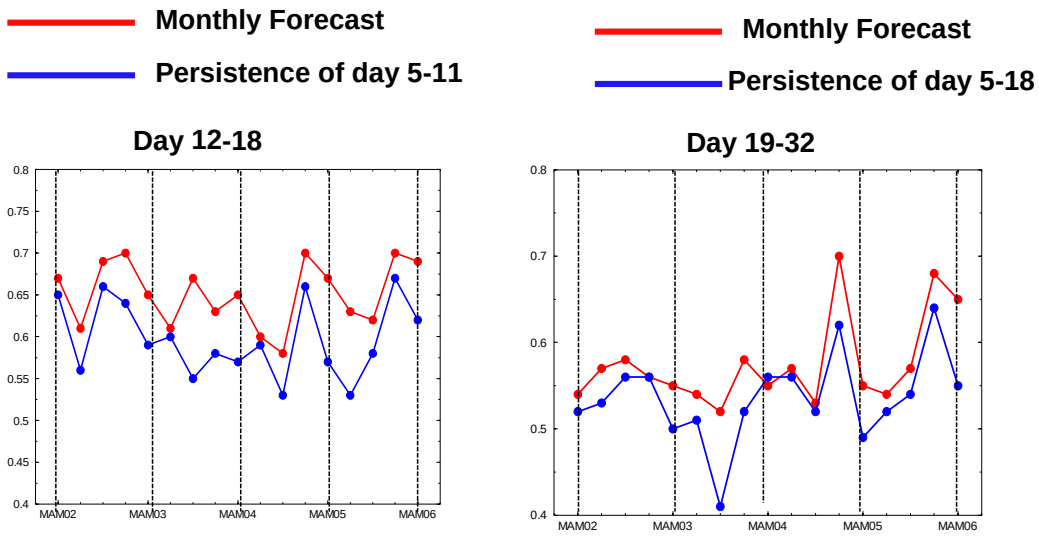


Figure 32: ROC area score of the probability that 2-metre temperature is in the upper third of the climate distribution, for each season since May 2002 over the northern extratropics. Only land points have been included. The red line represents the score of the operational monthly forecasting system. The blue line represents the score using persistence of the earlier part of the forecast.

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 - e.g. when verification scores are computed using radiosonde data (Figure 12), 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 square root of the geographical average of the squared differences between the forecast field and the analysis valid for the same time. When models are compared, each model uses its own analysis for verification; RMSE for winds (Figure 12, Figure 13) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

Skill scores (Figure 1) are computed as the reduction in Mean Square Error achieved by the model with respect to persistence (forecast obtained by persisting the initial analysis over the forecast range); in mathematical terms:

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

Figure 2 and Figure 4 show correlations in space between the forecast anomaly and the verifying analysis anomaly. Anomalies with respect to NMC Washington climate are available at ECMWF from the start of its operational activities in the late 1970s. Only for ocean waves (Figure 21, Figure 22) has the climate been derived from the ECMWF analysis.

A.2 Probabilistic forecasts

Events for the verification of medium-range probabilistic forecasts are usually defined as 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 Characteristic (ROC) curves.

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

$$BS = \left(p - o \right)^2$$

As for any probabilistic score, however, the BS only becomes significant when results are averaged over a large sample of independent events. Its value ranges from zero (perfect deterministic forecast) to 1 (consistently wrong deterministic forecast). The Brier Skill Score is defined as:

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

Time series of the Brier Skill Scores can be found in Figure 7.

There are four possible outcomes for a deterministic forecast of a dichotomous (yes/no) event: the event is forecast correctly (hit, H); the event is forecast and does not occur (False alarm, F); the event is correctly forecast not to occur (correct rejection, CR); or the event occurs but is not forecast (miss, M). The following measures are defined over a large sample:

Hit rate or Probability of Detection (POD) = H/(H+M)

False alarm rate = F/(F+CR)

False alarm ratio = F/(H+F)

Relative Operating Characteristic curves show how much signal can be gained 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 different ways, depending on whether 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) or 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 with the different thresholds (proportion of members or probabilities), used before the forecast will be issued (Figure 24).

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 perfect detection). Time series of the ROCA are shown in Figure 7 and Figure 32.

A.3 Weather parameters (Section 4)

Verification data are European 6-hourly SYNOP data (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 or 15m/s 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 for EPS forecasts of rainfall amounts is mentioned, the 0-24h-model forecast is used as a proxy for a model-scale analysis. A better alternative is to use an analysis derived from high-resolution networks upscaled to the model resolution. Although such data are not available in real time, ECMWF gets access to most networks in Europe and uses such analyses for internal purposes.