TECHNICAL MEMORANDUM

547

Verification statistics and evaluations of ECMWF forecasts in 2006-2007

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

Operations Department

January 2007

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

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

1. Introduction

This document 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, 504).

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.

12 September 2006: Cycle 31r1, including the following main changes:

- Revisions to the cloud scheme including treatment of ice supersaturation and new numerics
- Implicit computation of all convective transports
- New turbulent orographic form drag scheme and revisions to orographic wave drag
- Improved formulation of gusts for stochastic physics over orography
- Reduction of ocean surface relative humidity from 100% to 98% (due to salinity effects)
- Revised formulation of the assimilation of rain-affected radiances
- Variational bias correction of satellite radiances
- Thinning of low level AMDAR data (mainly affects Japanese AMDAR network)

This cycle included the necessary technical changes to enable the extension of the EPS to day 15 at reduced resolution (VarEPS). Cycle 31r1 has also been adopted for the seasonal forecasting System 3 and for ERA-Interim re-analysis.

2 November 2006: Passive monitoring of AMSU-A data from the MetOp-A satellite (launched on 19 October 2006).

28 November 2006: Variable resolution ensemble forecasting system (VarEPS). The horizontal resolution of the EPS is truncated from T399 to T255 at day 10 and the EPS forecasts are continued at this reduced resolution to day 15.

30 November 2006: Passive monitoring of HIRS and MHS data from MetOp-A.

12 December 2006: Cycle 31r2, introducing the assimilation of new satellite data: winds from MTSAT (Japanese GEO satellite), and GPS radio occultation data from CHAMP, GRACE and COSMIC.

11 January 2007: Operational assimilation of AMSU-A and MHS from MetOp-A.

March 2007: New seasonal forecasting system, System 3.

5 June 2007: Cycle 32r1, including the following main changes:

- Three-minimization version of 4D-Var assimilation scheme (T95/T159/T255) with improved moist linear physics (cloud and convection)
- Improved parametrization of the heterogeneous ozone chemistry
- New short-wave radiation scheme (RRTM-SW), plus McICA cloud-radiation interaction and MODIS albedo
- Retuned ice particle size
- Revised subgrid-orography scheme
- Explicit numerical treatment of convection in the moist tangent linear model used in the calculation of tropical singular vectors

12 June 2007: Operational assimilation of IASI radiances and ASCAT surface winds from MetOp-A.

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 shows the evolution of the skill of the deterministic forecast of 500 hPa height over the extra-tropical northern hemisphere and Europe since 1980. Each curve is a 12-month moving average of 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 June 2007. Figure 2 shows the equivalent evolution of performance using the anomaly correlation, where reference is to climatology instead of persistence. Both measures give a consistent signal of increasing skill over the past year. The positive trend is particularly striking for the anomaly correlation over Europe. The trends in Figure 1 are partly influenced by inter-annual variations in synoptic activity. Figure 3 shows that persistence errors were particularly low over the past year, making it more difficult to attain high skill using persistence as a reference. The improvements seen in Figure 1 and Figure 2 are thus a result of model improvements rather than changes in synoptic activity.

The consistently good performance over the last year can be seen in the scores for the individual months for all three regions in Figure 2. One notable reason for the overall high scores is a continuing reduction in the number of poor individual forecasts and corresponding increase in occurrence of skilful forecasts. This is illustrated in Figure 4, which shows the distribution of anomaly correlation scores for day 7 forecasts of 850 hPa temperature over Europe in winter and summer.

Figure 5 shows the time series of the average RMS difference between consecutive 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the downward trend indicates there is less "jumpiness" in the forecast from day to day.

The quality of ECMWF forecasts for the upper atmosphere is shown through the time series of wind scores at level 50hPa in the extratropics in Figure 6. In both hemispheres, warm season errors increased slightly in the last two years while cold season scores improved, reducing the overall difference in performance between winter and summer.

The trend in EPS performance is illustrated in Figure 7, which shows the evolution of the ranked probability skill score (RPSS) for 850 hPa temperature over the northern hemisphere and Europe. As for the deterministic forecast, the EPS skill was consistently good over the last year. The EPS performance benefited substantially from the increase in resolution in February 2006 (T255 to T399). This is apparent especially in the day 5 and day 7 scores over Europe in Figure 7 and is also clear in the results shown in Figure 8 and Figure 9. VarEPS was introduced at the end of November 2006, extending the EPS to 15 days with reduced resolution beyond day 10. The ensemble spread and ensemble-mean error over winter 2006-07 and 2005-06 are shown in Figure 8 for the extra-tropical northern hemisphere. The close match between spread and error in 2006-07 indicates the EPS is well tuned; this correspondence is maintained throughout the 15-day forecast range at both 500 hPa and 850 hPa. The error of the ensemble mean is lower in 2006-07 than in the previous winter and more closely matched by the spread. Figure 9 shows the RPSS for days 1 to 15 for winter 2006-07 over the extra-tropical northern hemisphere. For both 500 hPa height and 850 hPa temperature, there is a substantial improvement over previous years and positive skill is maintained out to the 15-day range in the new system.

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 10. Recent model changes have led to continued improvements, especially in the upper tropospheric winds. Although the day-1 error has increased slightly, when verified against the model analysis, verification against radiosonde observations shows a continued reduction in error.

3.2 ECMWF vs other NWP centres

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 11 shows time series of such scores over the northern extratropics for both 500hPa height and Mean Sea Level Pressure. All centres performed well over the last year, with lowest ever errors for both winter and summer periods. The ECMWF lead over other centres is comparable to the previous year. The gap is, in general, bigger in the southern extratropics (Figure 12). Most centres show smaller errors in the 2007 cold season than in previous years. The ECMWF lead is reduced in this period but is still consistent and substantial at longer range, especially for 500 hPa height.

WMO exchanged scores also include verification against radiosondes over smaller areas, such as Europe. Figure 13, showing both 500 hPa height and 850 hPa wind errors, confirms the good performance of the ECMWF forecasts using this alternative reference.

The situation in the tropics is summarised in Figure 14. 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 the short-range error for ECMWF remaining fairly constant for the last two years, the corresponding scores for radiosonde observations show a continuing trend of reducing errors. The most noticeable changes over the past year are the improvement in 850 hPa wind errors for the NCEP forecasts at short-range and the contrasting increase in errors for the Canadian forecasts.

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 15 to Figure 18. 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. The relatively large daytime bias in winter 2005-06 was associated with the representation of low cloud in periods of anticyclonic conditions. Biases have been consistently low in 2006-2007. The negative daytime bias in cloud cover has reduced substantially since 2005 (Figure 17) and error standard deviation is also decreasing. Physics changes introduced in model cycle 31r1 (September 2006) increased 10 m wind speeds globally, generally improving negative biases in many regions. Over Europe this resulted in a change from negative to positive overall bias for daytime forecasts (Figure 18), but did not adversely affect the error standard deviation.

The trend in precipitation skill for Europe is shown in Figure 19, using the True Skill Score (or Pierce’s Skill Score) for thresholds of 1mm and 10mm per day. The improvement in skill following the introduction of cycle 31r1 is particularly striking for low and moderate precipitation thresholds (1mm and also 5mm, not shown). The same signal can be seen in the scores for the EPS probability forecasts shown in Figure 20.

4.2 Ocean waves

The quality of the ocean wave model analysis continues to improve, as can be seen in the comparison with independent ocean buoy observations in Figure 21. The improvement in the analysis since the introduction of JASON altimeter data in February 2006 is clear.

Figure 21 also shows a time series of the analysis error for the 10 metre wind over maritime regions using the wind observations from the same set of buoys. The error has steadily decreased since 1998, providing better quality winds for the forcing of the ocean wave model. As for the wave height, there was a substantial improvement in the past year.

The good performance of the wave model forecasts is confirmed again this year, as shown in Figure 22 and Figure 23. This is particularly noticeable in the verification against observations and comparison with other wave models, as shown in Figure 24. The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of northern hemisphere buoys. Two additional centres (the Japanese Meteorological Agency, JMA, and the French naval Service Hydrographique et

Océanographique de la Marine, SHOM,) were added to this comparison in 2006. The French SHOM forecasts are the closest in performance to ECMWF; their wave model is driven by the ECMWF winds.

5. Severe weather

5.1 Verification of precipitation

The precipitation fields that are directly output from a forecast model should be treated as averages over the area of a model grid box. They do not directly correspond to the point observations that are often used for verification. There is likely to be a large discrepancy between the model value and the observations, especially in severe weather events when the focus is on the maximum precipitation that has occurred at points in an area. In real time, only the relatively sparsely distributed synoptic observations are available for verification over the European area. At ECMWF and elsewhere these are therefore used for verification of the model precipitation fields, ignoring the effect that neglecting the sub-gridscale variability will have on the scores.

This issue was addressed in a collaboration between ECMWF and Deutscher Wetterdienst on the verification of severe weather. In order to estimate the impact on verification of comparing area-mean forecasts with point observations, we have used the European high-resolution network of daily precipitation accumulations to construct a “perfect forecast” by computing the area-mean of observations within each model grid box. To assess the effect in severe weather situations, where the interest is in the extreme values, this “perfect forecast” is then verified against the maximum observation in the grid box.

The skill of the perfect model was assessed for maximum precipitation exceeding different thresholds of up to 100 mm per day, using the equitable threat score (ETS). By construction, the perfect model forecast (grid box average) will always be less than the verifying observation. In forecasting a severe event it is sensible to issue a warning, when the model is predicting a lower amount for the grid box as a whole. For each event, the model threshold that produced the highest ETS was selected.

Figure 25 shows the skill of the perfect model forecast for different precipitation amounts. Even for light precipitation, the skill is substantially below the maximum value of 1.0. For heavier precipitation the skill is lower, dropping below 0.5 for events greater than 40 mm per day. As a comparison, the skill of the ECMWF deterministic forecast is also shown; the data is for winter (December to February) 2004-2005, the most recent winter for which the high resolution data was available during this study. As might be expected, the forecast skill is considerably less than that of the perfect model. Nevertheless, the difference between the ECMWF and perfect model scores is much smaller than between either of these and the theoretical maximum ETS value of 1.0. Using point observations to directly verify gridded model precipitation forecasts inevitably generates an impression of poor skill, especially in predicting severe events.

5.2 Tropical cyclones

After the exceptional 2005 North Atlantic hurricane season, the 2006 season was close to normal. Nevertheless there were some notable events. Ex-hurricane Gordon struck north-east Iberia in November and subsequently, as an intense extratropical system, brought severe weather to parts of the UK and Ireland. This system was very well and consistently forecast by the high-resolution T799 system.

Average position and intensity errors for all tropical cyclones forecast over the three latest 12-month periods are shown in Figure 26. The resolution increase in February 2006 resulted in substantial improvements in intensity, with tropical cyclones appearing deeper throughout the forecast range. The core pressure is also much better in the analysis, as well as in the forecast. While initial position errors have not changed significantly, the past year has consistently seen the lowest errors at all forecast steps. The improvement is mainly due to a reduction in the along-track error. In general, tropical cyclones have moved too slowly in the forecast (negative along-track error); this slow bias is reduced, compared with the previous years.

The EPS tropical cyclone forecast is presented on the ECMWF web site as a strike probability: the probability at any location that a reported tropical cyclone will pass within 120 km of that location within the next 120 hours. Verification of these probabilistic forecasts for the three latest 12-month periods is shown in Figure 27. Reliability is generally similar in the three years. The signal detection capability (as indicated by the ROC) is higher in the most recent year. This is particularly evident in the modified ROC which uses the false alarm ratio instead of the false alarm rate on the horizontal axis (this removes the reference to the non-events in the sample and shows more clearly the reduction in false alarms in those cases where the event is forecast).

Some notable examples of tropical cyclone genesis during the forecast have also been seen, in both the deterministic forecast and the EPS. The routine identification and tracking of tropical cyclones that appear during the forecast (but are not present at analysis time) is under investigation.

6. Monthly and Seasonal forecasts

6.1 The 2006-2007 El Niño forecasts

By September 2006, sea-surface temperatures across the central and eastern Pacific were warmer than average. Figure 28 shows predictions for the Niño-3.4 region from September 2006 and January 2007, produced using the operational seasonal forecast System 2 and the new System 3, together with subsequent verification. Both systems were run in parallel from September 2006 and System 3 replaced System 2 as the operational system in March 2007. For further information about seasonal forecast System 3 and its products see ECMWF Newsletters 110 and 111.

Predictions initiated in September 2006 from System 2 gave a rather good indication of further warming up to 4 months into the forecast, although the later part of the forecast underestimated the transition to cooling. System 3 may have been less accurate in giving the magnitude of the warming observed during December 2006 but gave a more realistic indication of the temperature anomalies at longer range. Predictions initiated in January 2007 from both systems successfully forecast the El Niño's quick decline in the early part of the year. This transition was unusually rapid - transitions from warm to cold phases typically occur on a time scale of one year.

6.2 Seasonal Forecast performance for the tropics

ECMWF has issued global seasonal predictions every month since 1997. During this time the seasonal forecast system has been upgraded twice. Figure 29 shows the root mean square (RMS) errors for the predicted sea-surface temperature anomalies over the Equatorial Pacific, computed over a common set of hindcast cases for System 1 (used in 1997), for System 2 (operational from 2002) and for the present operational System 3 (implemented in March 2007). Figure 29 illustrates the sustained improvement of seasonal forecast performance.

Since the implementation of System 3, an additional “ENSO outlook” is made four times a year by extending the forecast up to 13 months (with a reduced ensemble size of 11 members, compared to the normal 41 members). Figure 30 shows an estimate of the forecast skill for temperature anomalies over the Nino 3.4

area, based on the period 1981-2005. These annual-range forecasts are available on the web, to Member State users, at http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/annual_range_forecast/

While the yearly-averaged skill estimates are useful overall indications, it is important for users to know how the seasonal forecast skill changes according to the time of year and to the forecast lead time. Figure 31 shows the anomaly correlation for the Southern Oscillation Index (SOI) as a function of forecast lead time in months (vertical axis) versus the "target" or verification month (horizontal axis). Correlations exceed 90% for forecasts verifying in September to January, with leads of up to 4 months. However, forecasts for the northern hemisphere summer months, most notably for June, are more difficult, with correlations as low as 50% at 6 month lead time. The sudden drop in skill near April-May is known as the spring barrier.

6.3 Seasonal Forecast performance for the extra-tropics

In conjunction with the extraordinary, rapid transition from a warm ENSO phase to a cold one, considerable intra-seasonal fluctuations have been observed in the extra-tropical circulation in the northern hemisphere. Some of these month-to-month variations were well represented by the forecast. The seasonal forecast also gave good indications of the mild winter-spring temperatures observed over Europe. The verification can be seen on the web site, for example from the predictions started in January and February 2007 for the southern European area at

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/seasonal_range_forecast/groupp/Climagrams_2mt.

In mid-latitudes, where predictability based on the ensemble mean is often low, skill is better assessed using probabilistic measures such as the Relative Operating Characteristic (ROC).

Figure 32 shows the area under the ROC for forecasts that the index of the Pacific North American pattern (PNA) will be in the lower third of the climate distribution. The negative phase of PNA pattern is typically related to cold conditions over the tropical Pacific (La Niña) and it generates warmer temperatures and weaker precipitation over the western USA. The graph displays the skill as a function of forecast lead in months versus the verification months (as in Figure 31). Although the skill is plotted for all the months, it is important to note that the PNA pattern is particularly relevant in the northern hemisphere winter, when it is one of the leading variability patterns. As expected, the seasonal forecast shows some skill between November and April, with forecasts for March exceeding a ROC value of 60% at leads of up to 3 months.

A complete set of verification statistics based on the hindcast integrations (1981-2005) from the System 3 is under development and will be made available on the ECMWF seasonal forecast web site.

6.4 Monthly Forecast verification statistics and performance

Except for the changes in common with the operational EPS forecast, no modification has been made to the monthly forecasting system in the past year. The present operational version of the monthly forecast is cycle 32r2, implemented on 7 June 2007.

Comprehensive verification for the monthly forecasts is available on the ECMWF web site at:
http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/.

Figure 33 shows the ROC score computed over each grid point for the 2-metre temperature monthly forecast anomalies at two forecast ranges: days 12-18 and days 19-25. All the real-time monthly forecasts since 7 October 2004 have been used in this calculation. The red colours correspond to ROC scores higher than 0.5 (the monthly forecast has more skill than climatology) and the blue colours 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 the ERA40 reanalysis or the operational analysis, when ERA40 is not available.

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

6.4.1 Monthly Forecasts performance 2006-2007

Figure 34 shows the probabilistic performance of the monthly forecast over each individual season since May 2002 for the time ranges days 12-18 and days 19-32. The figure shows the ROC scores for the probability that the 2-metre temperature is in the upper third of the climate distribution over the extra-tropical northern hemisphere. The model has consistently performed better than persistence of the previous 7-day or 14-day period. At days 12-18, scores reached their highest ever values for both winter 2006-07 and spring 2007. In contrast, persistence skill has been low throughout the past year, so the gap between persistence and model has grown substantially. For the forecast range 19-32 days, the scores for winter and spring are slightly lower than for the previous year. However, persistence scores are also lower and (as for the earlier forecast range) the gap between persistence and model skill is relatively large.

Over the tropics the monthly forecast performance has been generally good. The late onset of the West African monsoon, observed in conjunction with a suppressed convection phase associated with a Madden-Julian Oscillation (MJO) event, were both well predicted by the monthly forecast at day 12-18. The weekly variability of the Indian monsoon precipitation was also realistically reproduced at forecast range 12-18 days.

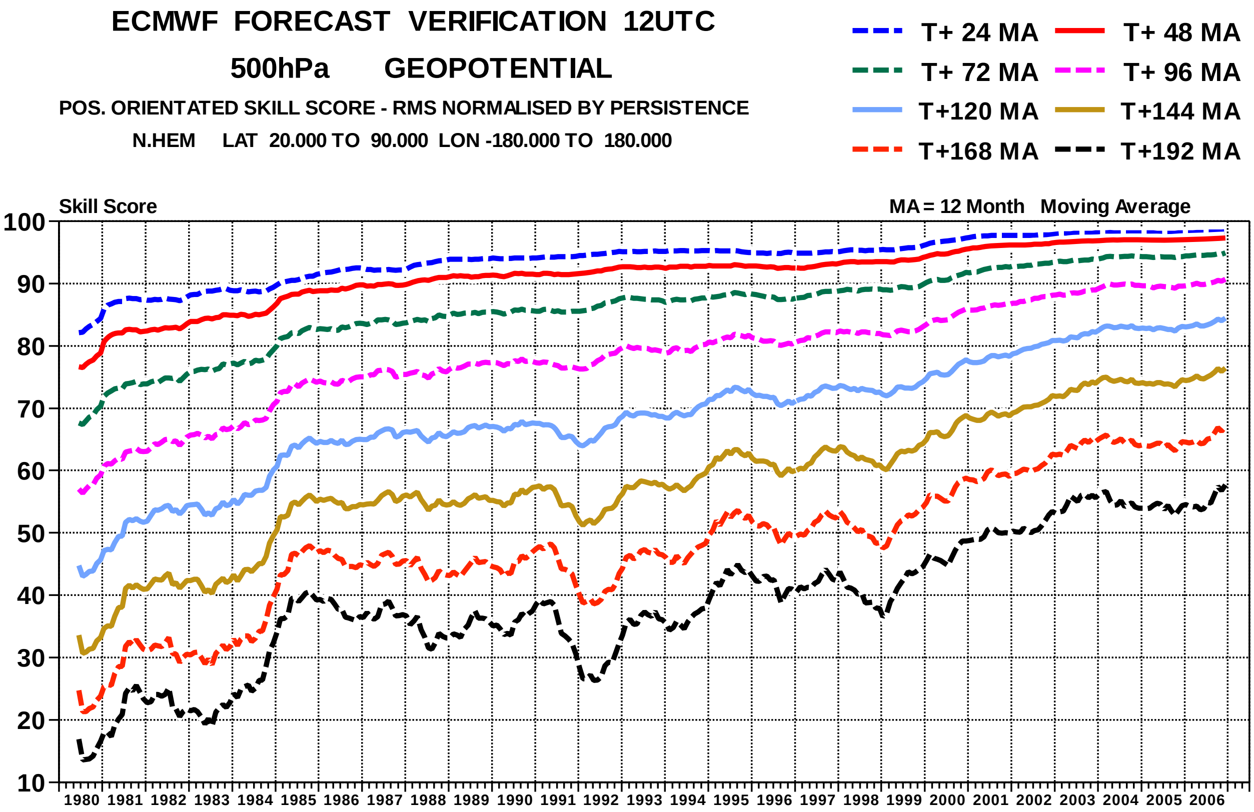
During December 2006 a strong episode of enhanced convection (above-average rainfall amounts) was observed across the Indian Ocean, as well as over the central equatorial Pacific, while suppressed convection (below-average rainfall amounts) was seen across Indonesia. These convection anomalies were associated with an intensification of Madden-Julian Oscillation (MJO) activity. A recent study from Vitart et al. (2007) has shown that the Monthly Forecasting system has some skill in predicting the evolution of the MJO up to 12-14 days in advance. However, the skill is less than that of some statistical models (up to 20 days). The results from Vitart et al. indicate a great potential for improving the skill of the current forecast system in predicting the MJO events.

7. References

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

Vitart, F., S.J. Woolnough, M.A. Balmaseda and A. Tompkins, 2007: Monthly forecast of the Madden-Julian Oscillation using a coupled GCM. *Monthly Weather Review*, **135**, 2700-2715

List of Figures



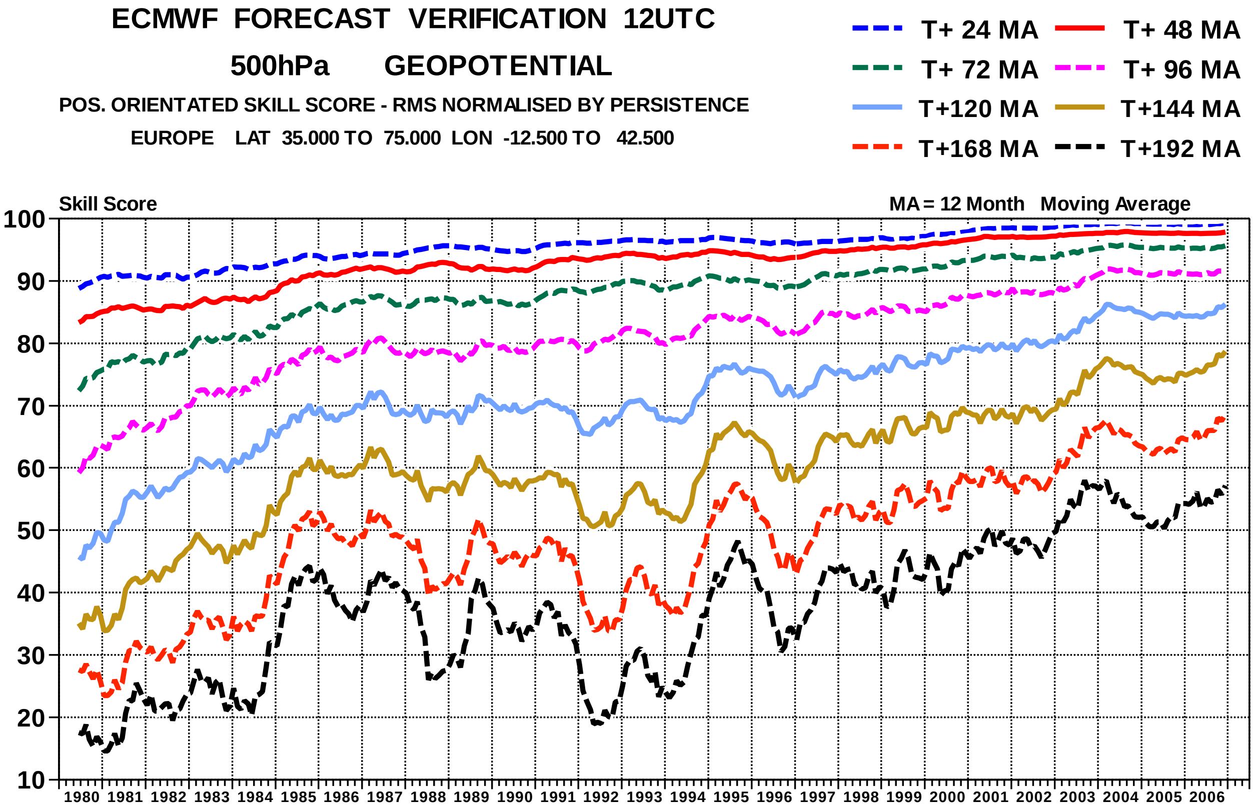


Figure 1: 500hPa height skill score for northern hemisphere (top) and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours

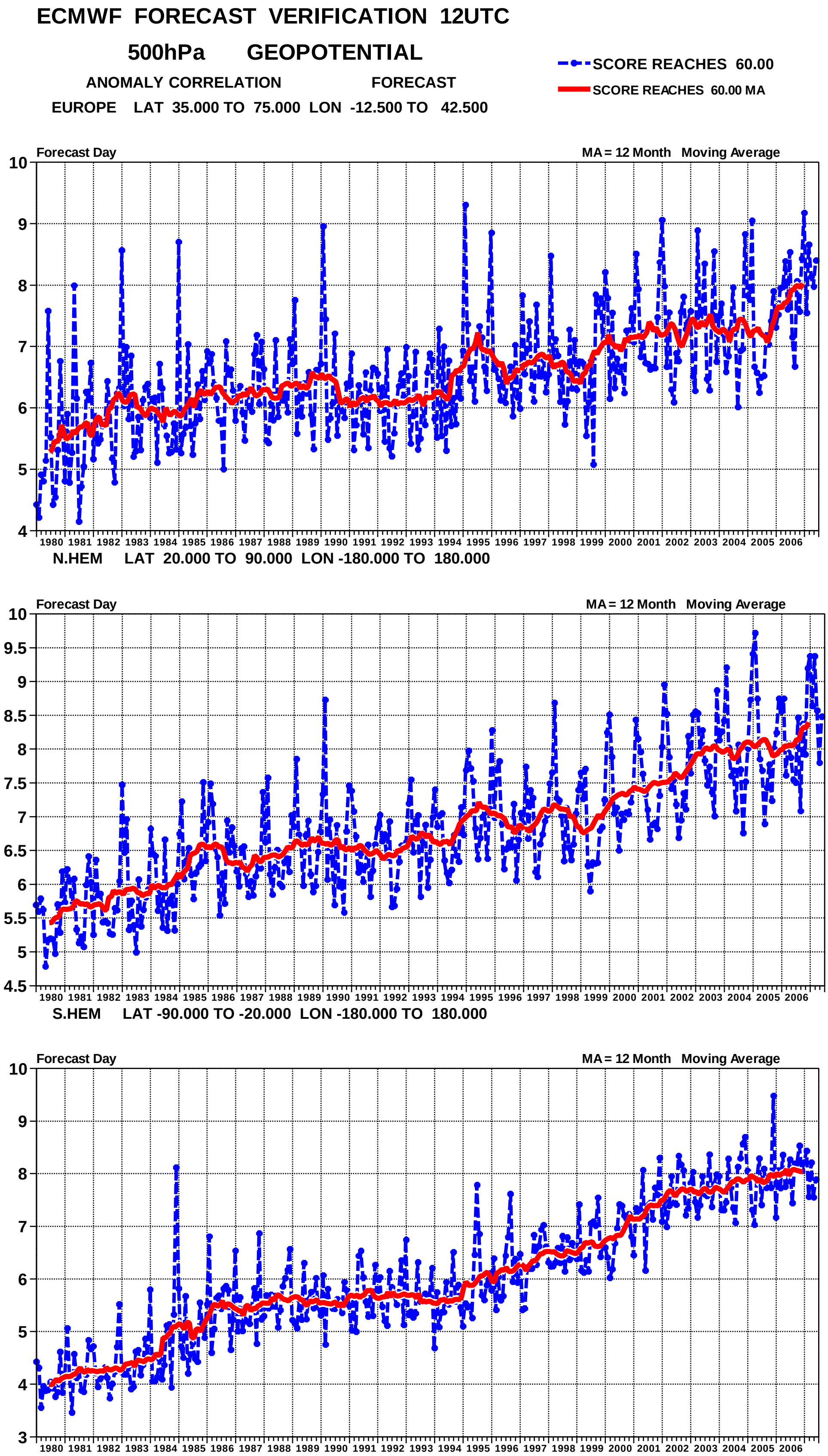


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the blue curves is the forecast range at which the monthly average of the forecast anomaly correlation with the verifying

analysis falls below 60% for Europe, northern and southern extratropics (the red curve is the 12-month moving average)

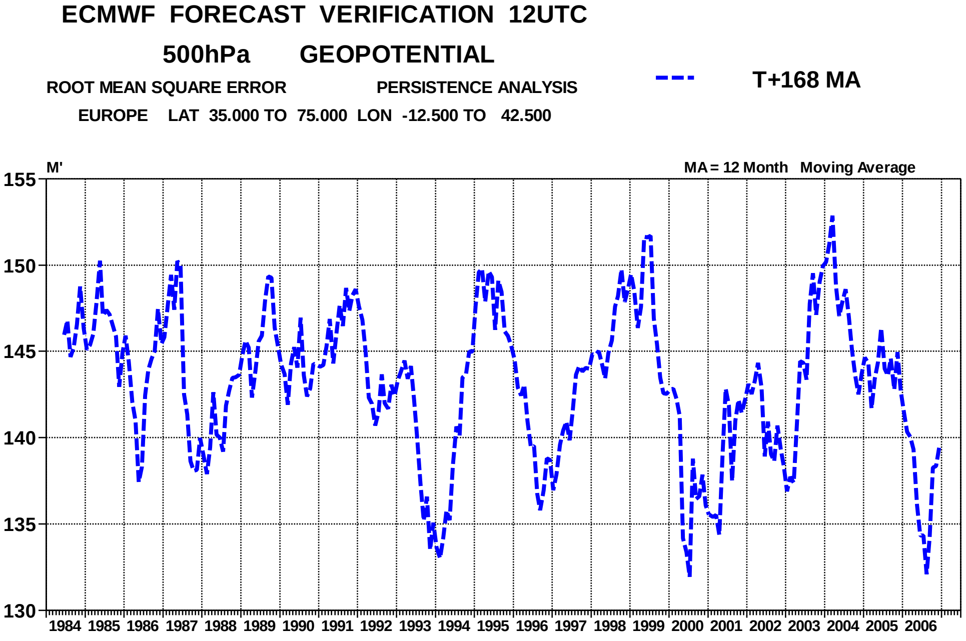


Figure 3: Root Mean Square Error of forecast made by persisting the analysis over168h and verifying it as a forecast for 500 hPa geopotential height over Europe. 12-month moving average.

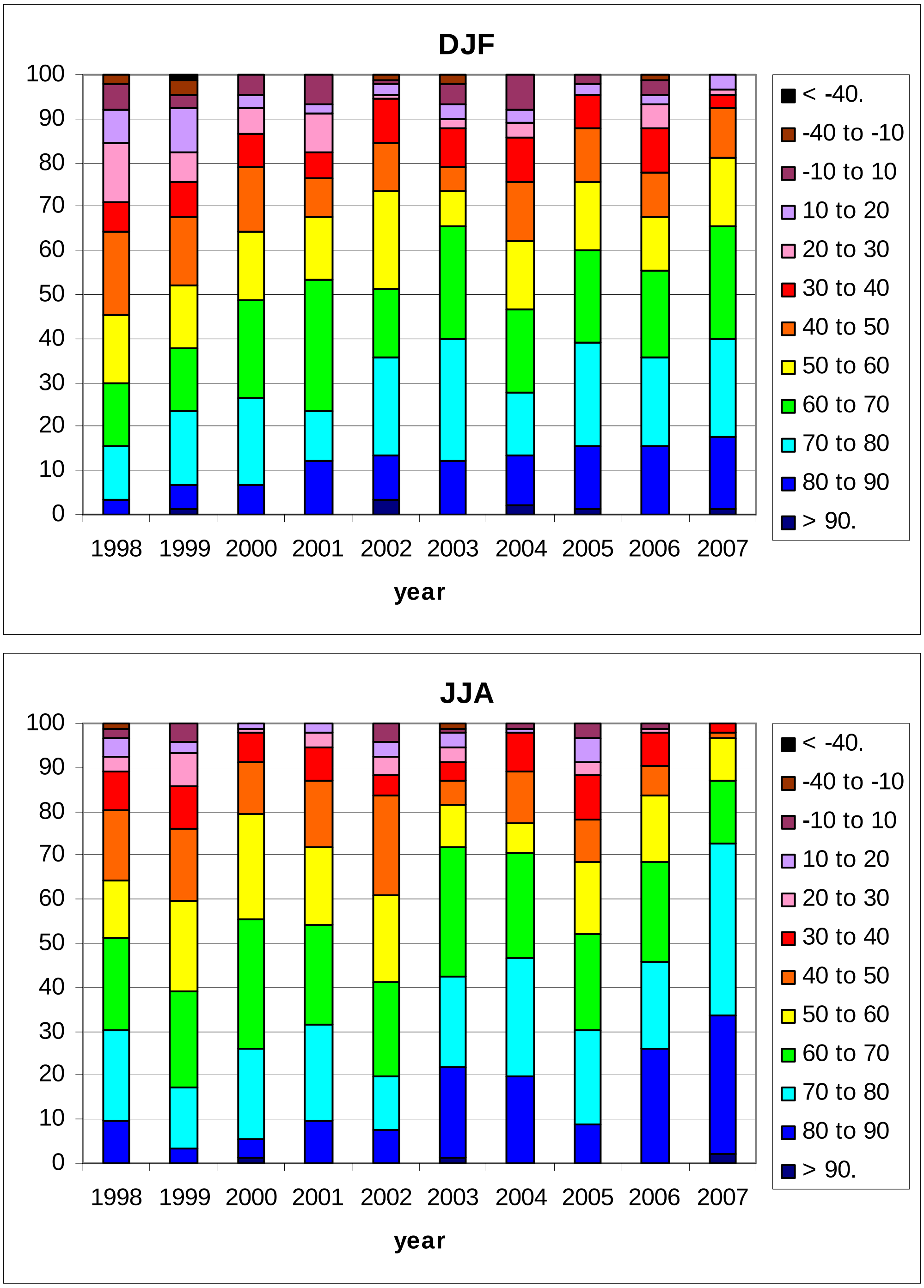


Figure 4: Cumulative distribution of Anomaly Correlation of the Day 7 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997-1998.

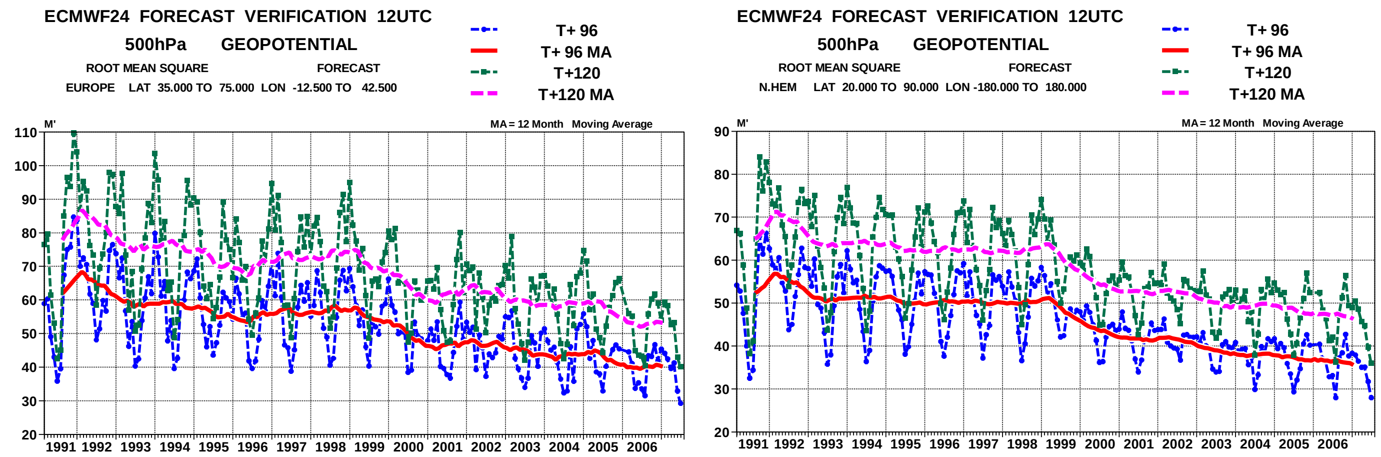


Figure 5: Consistency of the 500hPa height forecasts over Europe (left panel) and northern extratropics (right panel). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24h apart, for 96-120h (blue) and 120-144h (green). 12-month moving average scores are also shown. Last month is July 2007.

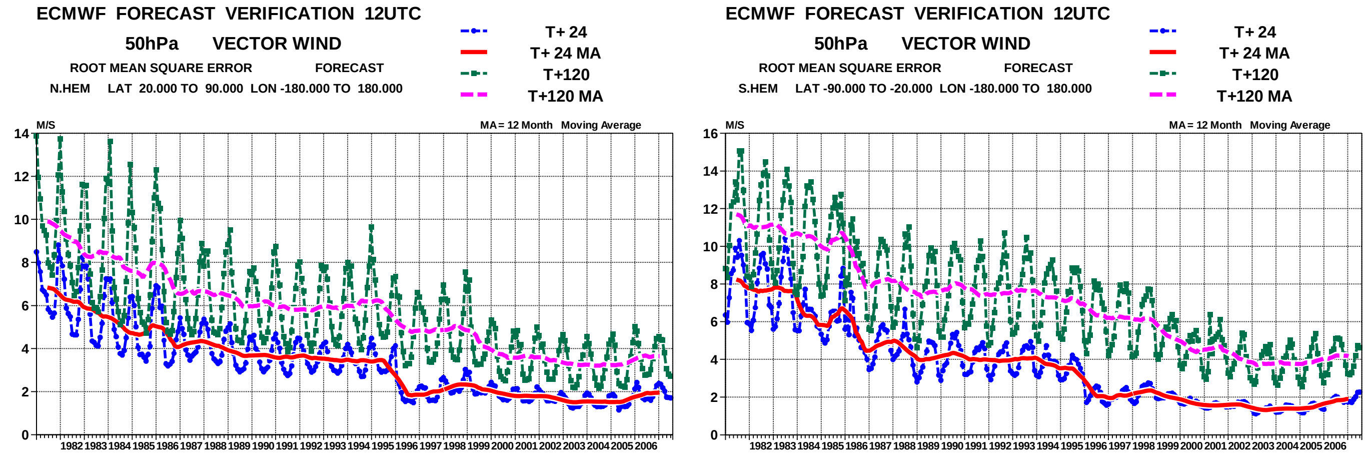
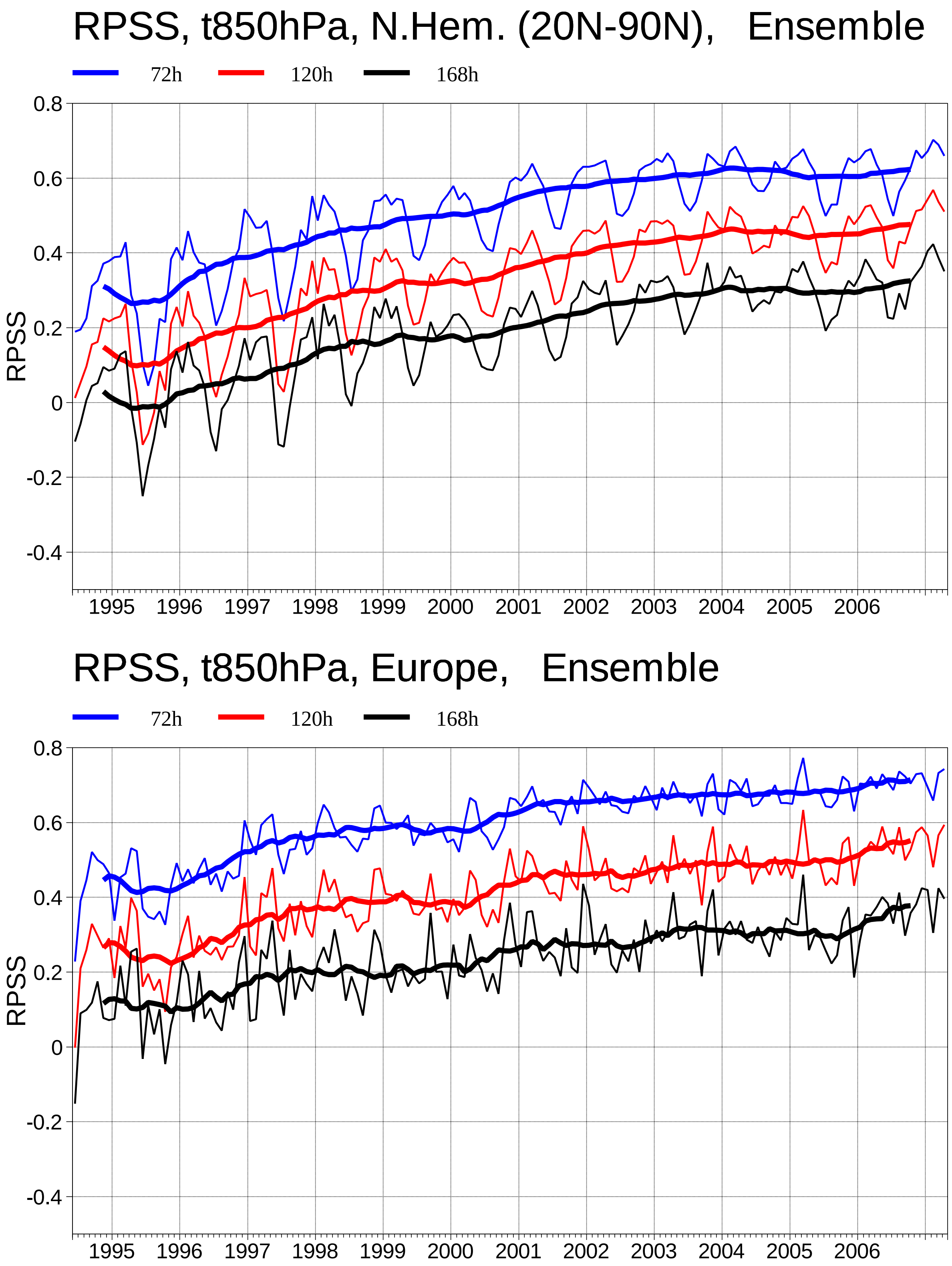


Figure 6: Model scores in the extratropical northern (left) and southern (right) hemisphere stratosphere (RMS vector wind error at 50hPa for 1-day and 5-day forecasts)



RPSS, t850hPa, N.Hem. (20N-90N), Ensemble

RPSS, t850hPa, Europe, Ensemble

Figure 7: Monthly score and 12-month running mean (bold) of Ranked Probability Skill Score for EPS forecasts of 850 hPa temperature at day 3 (blue), 5 (red) and 7 (black) for the northern hemisphere extratropics (top) and Europe (bottom).

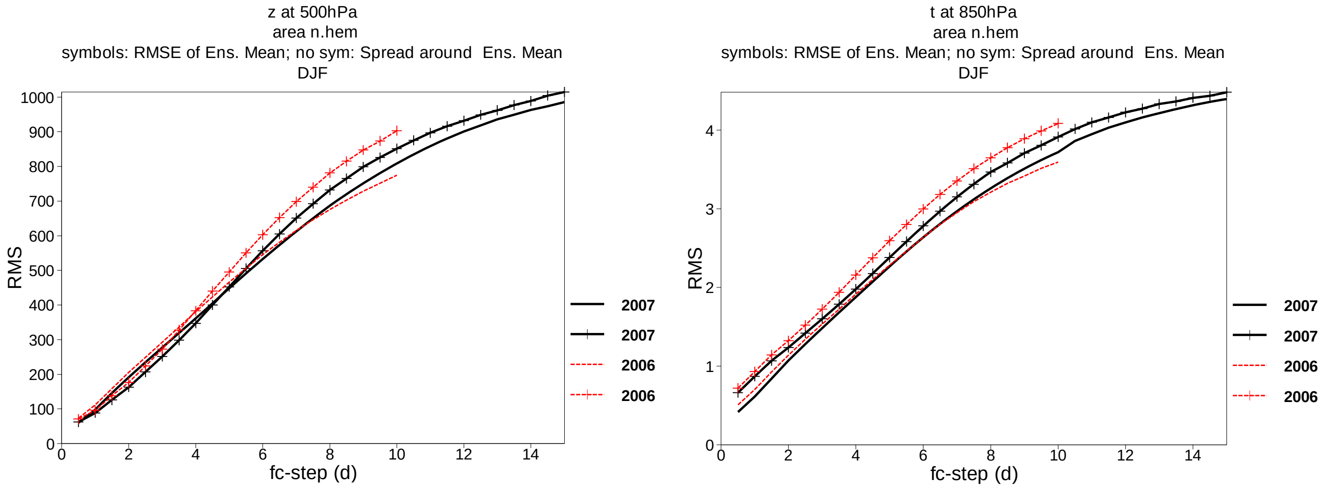


Figure 8: Ensemble spread (standard deviation) and root mean square error of ensemble-mean (lines with crosses) for 500 hPa height (left) and 850 hPa temperature (right) for winter 2006-07 (black) and 2005-06 (red) over the extra-tropical northern hemisphere.

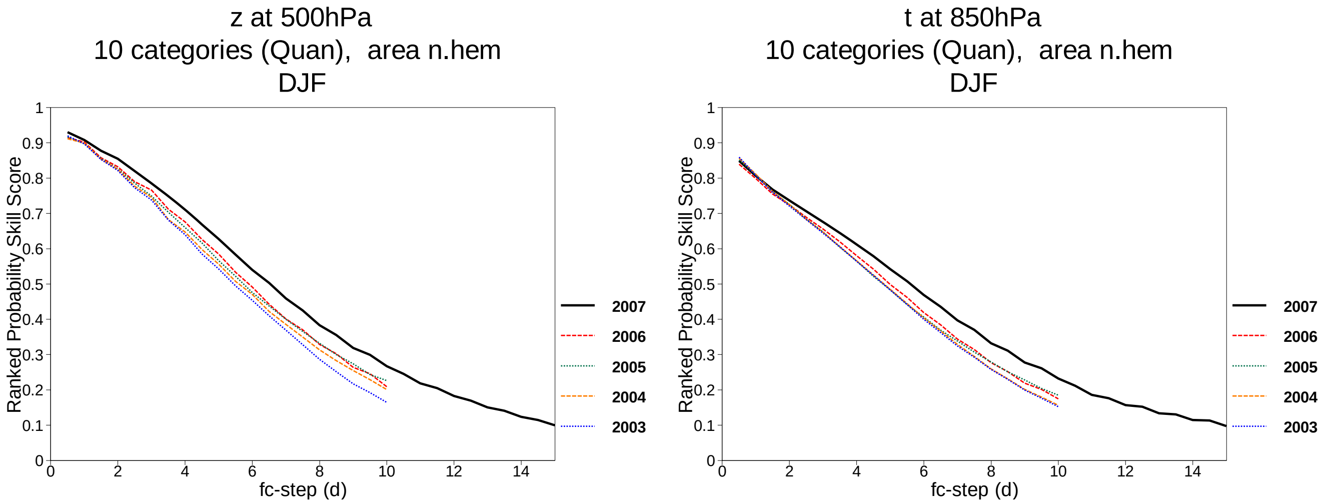


Figure 9: Ranked probability skill score for 500 hPa height (left) and 850 hPa temperature (right) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. The solid black line shows the skill from the VarEPS days 1-15 forecasts for winter 2006-07; there is a clear improvement over previous winters (the EPS only ran to 10 days in previous years).

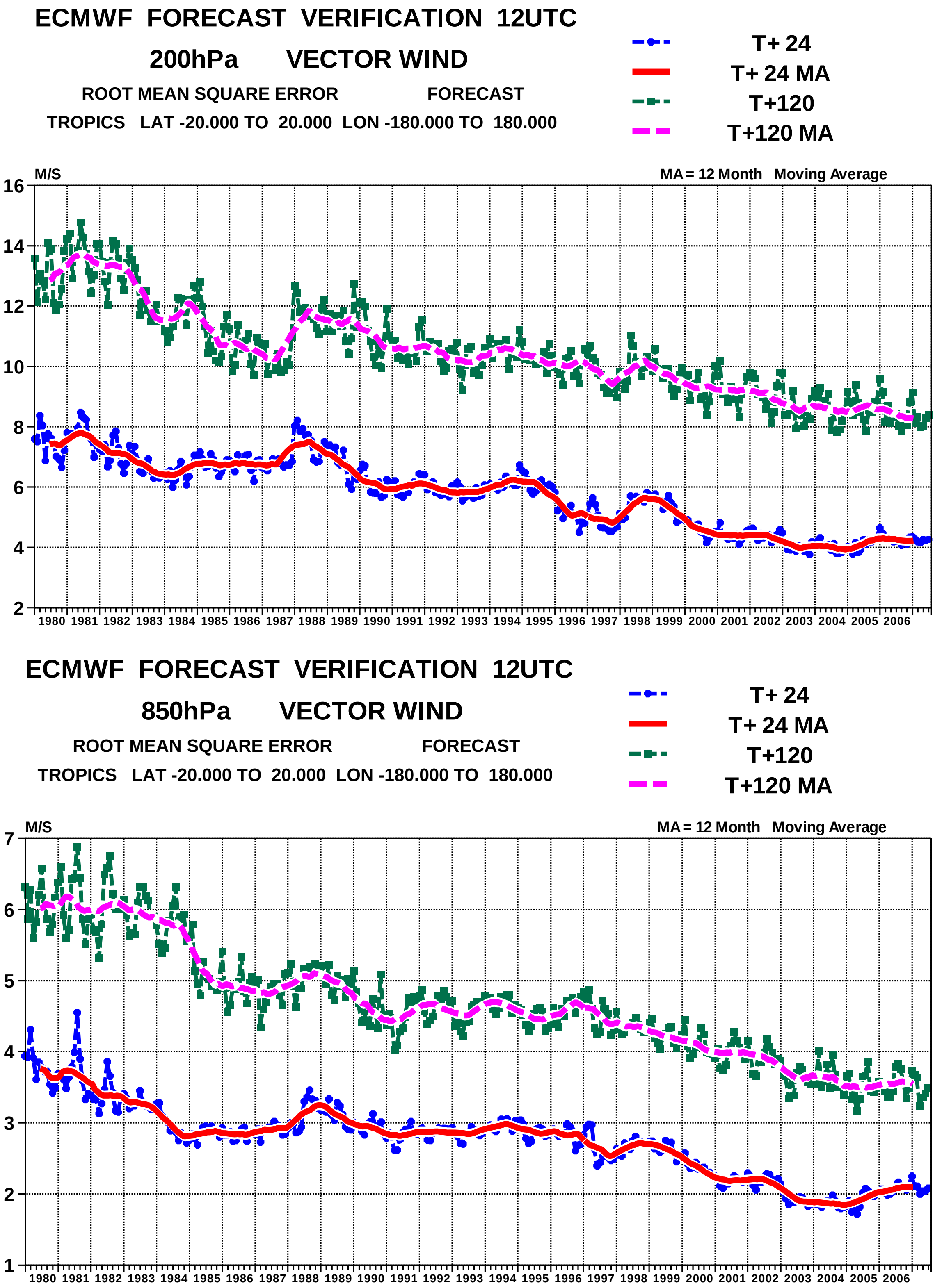


Figure 10: Model scores in the tropics (root mean square vector wind errors at 200hPa and 850hPa for 1-day and 5-day forecasts). Monthly mean and 12-month running mean.

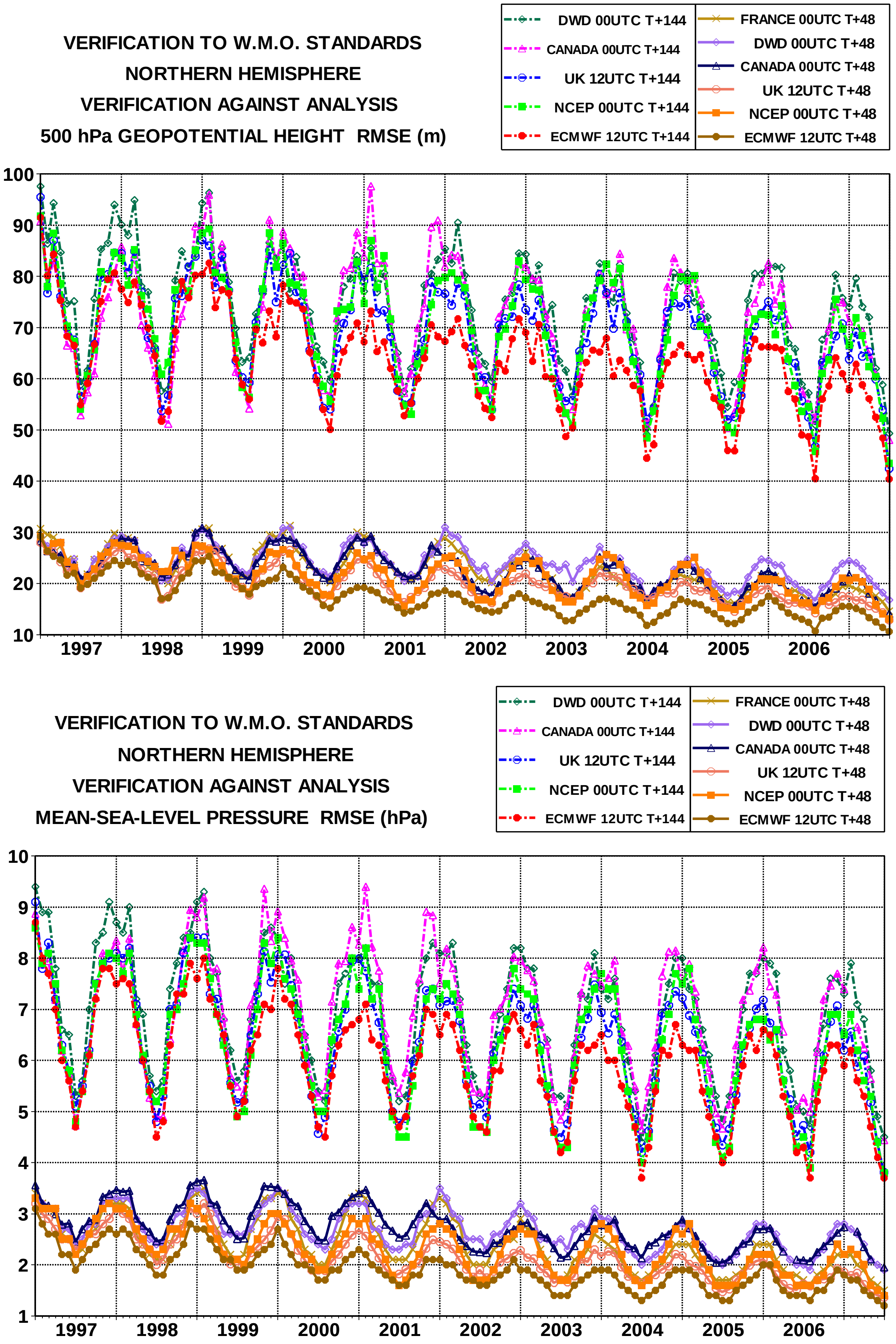
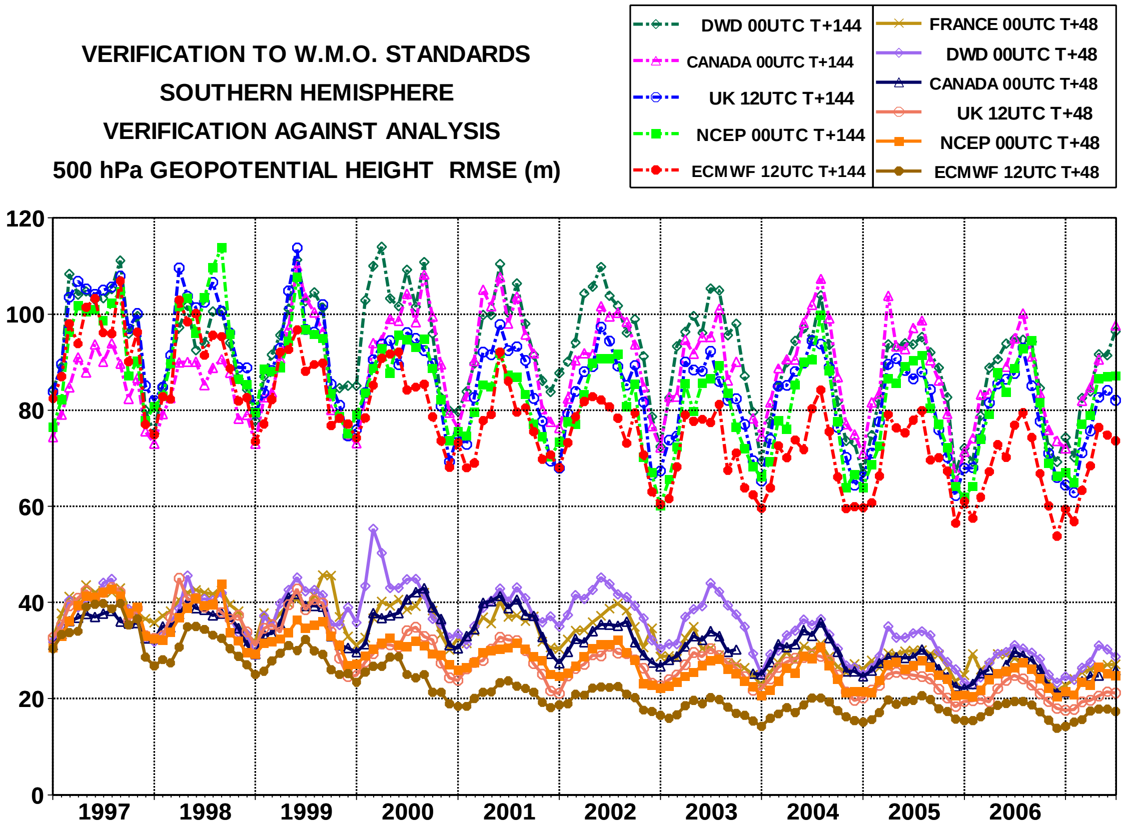


Figure 11: WMO/CBS exchanged scores (RMS error over northern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).



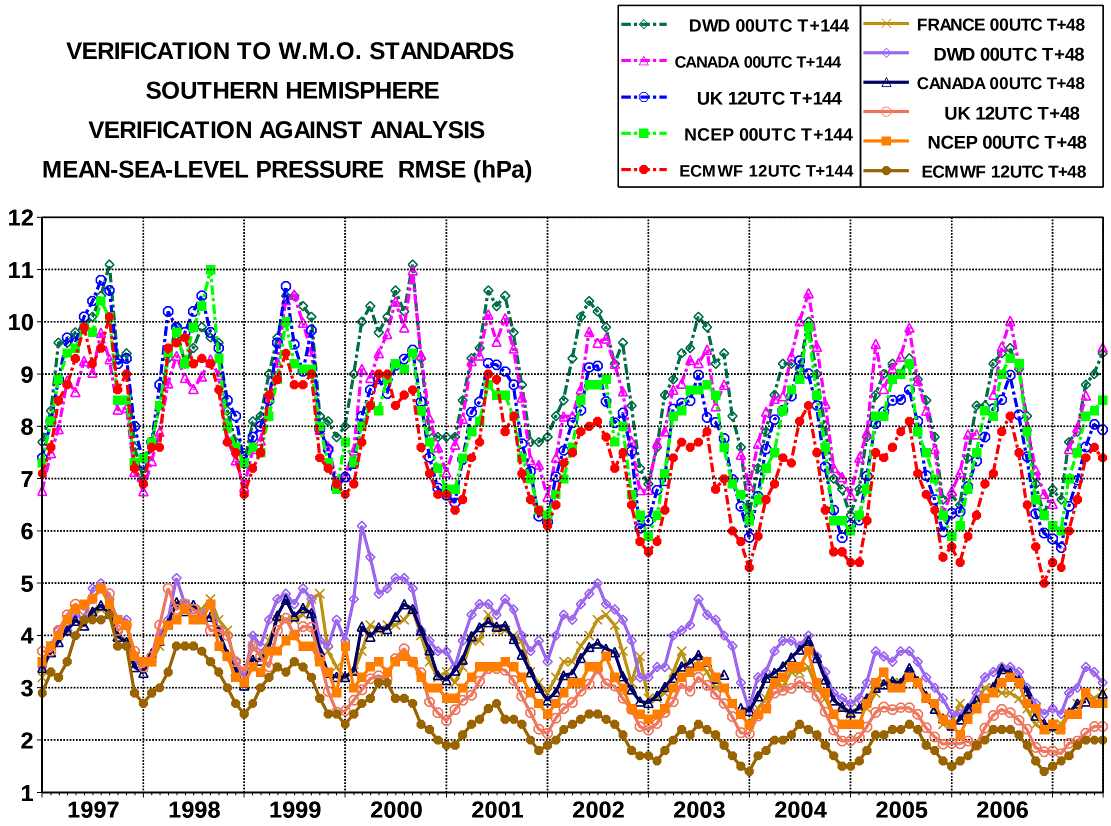
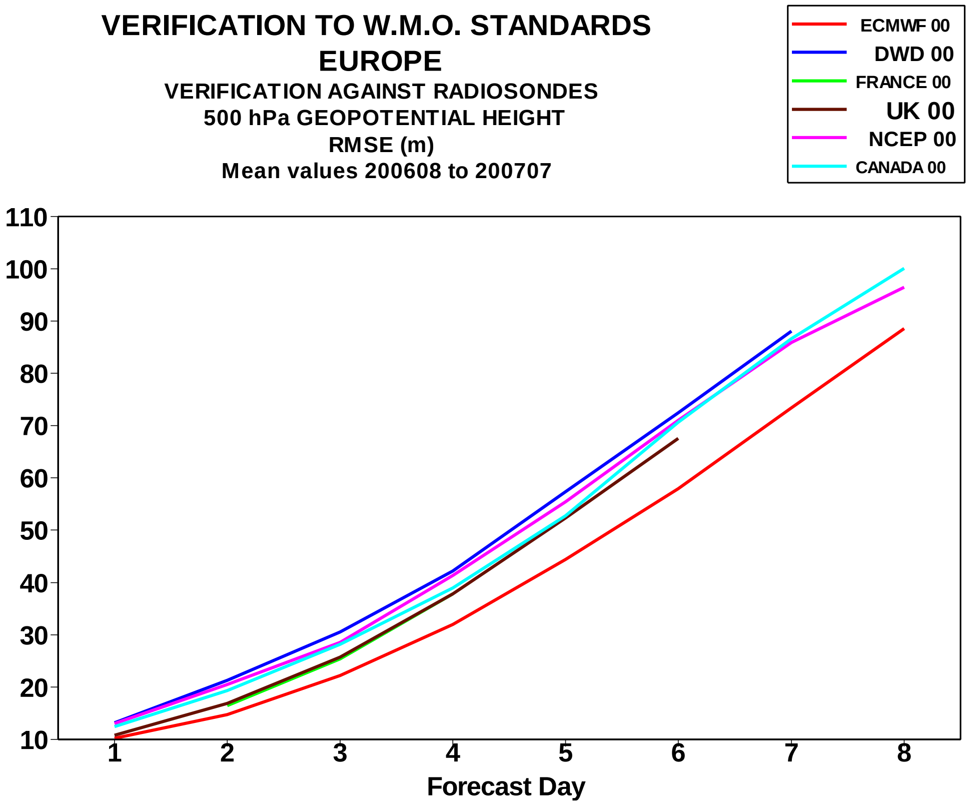


Figure 12: WMO/CBS exchanged scores (RMS error over southern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).



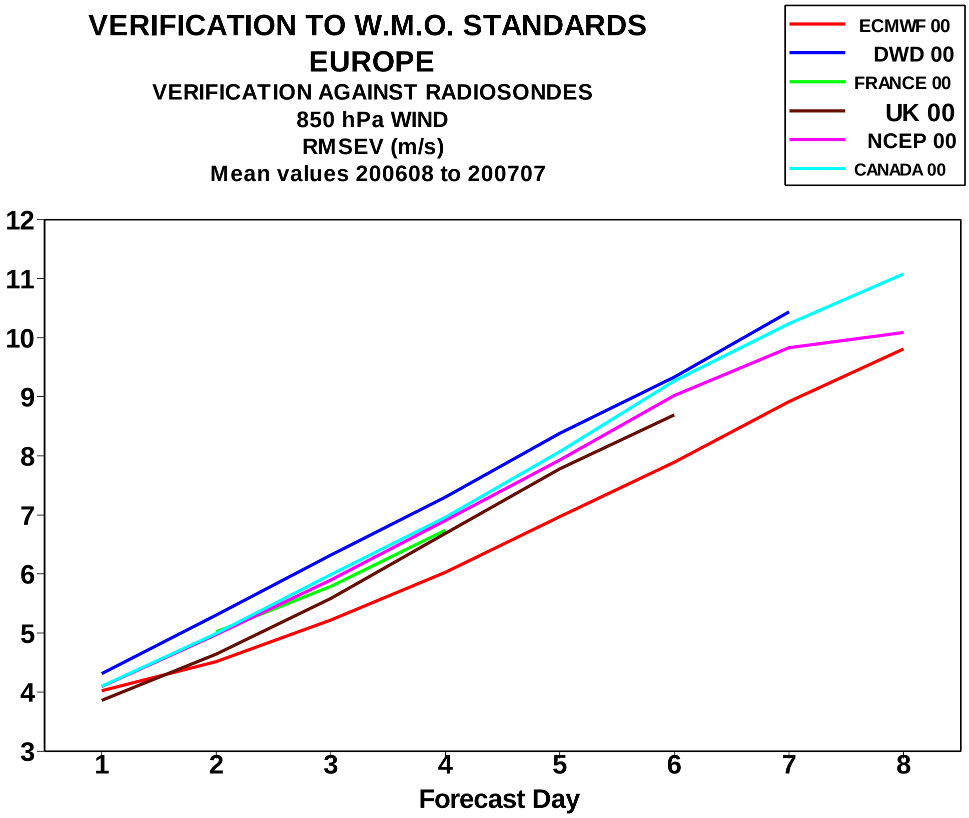
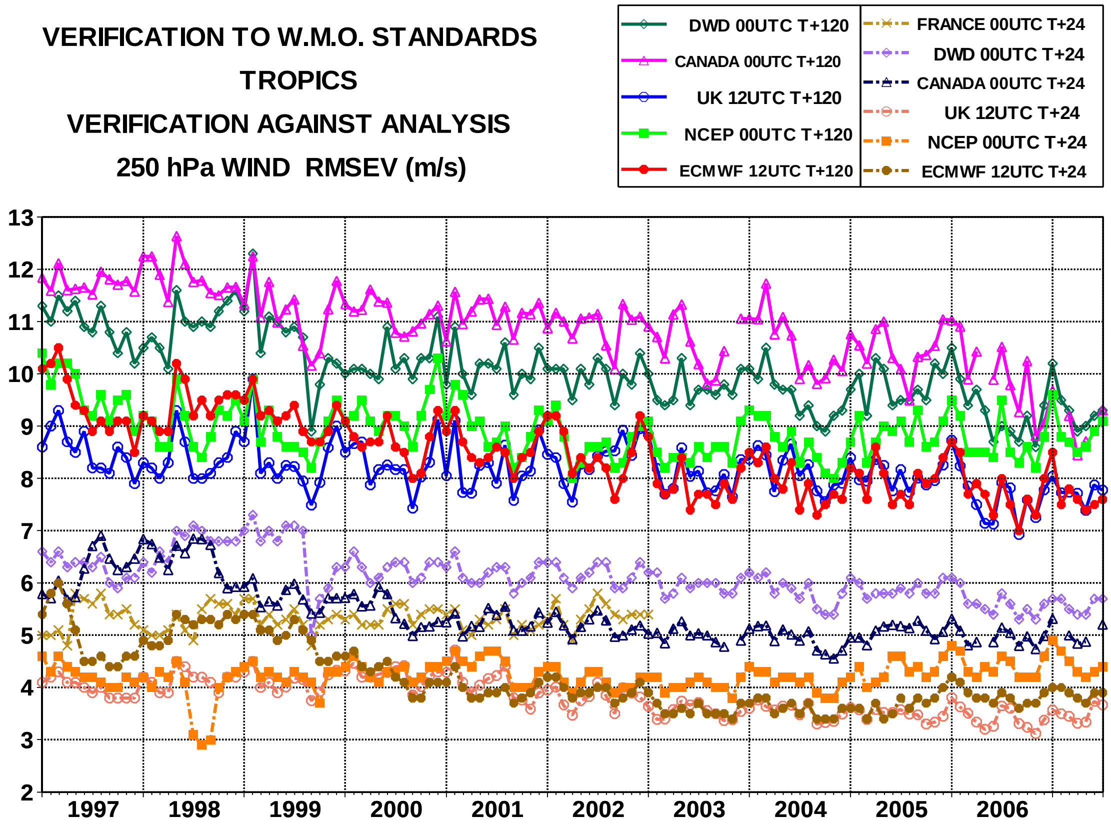


Figure 13: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)



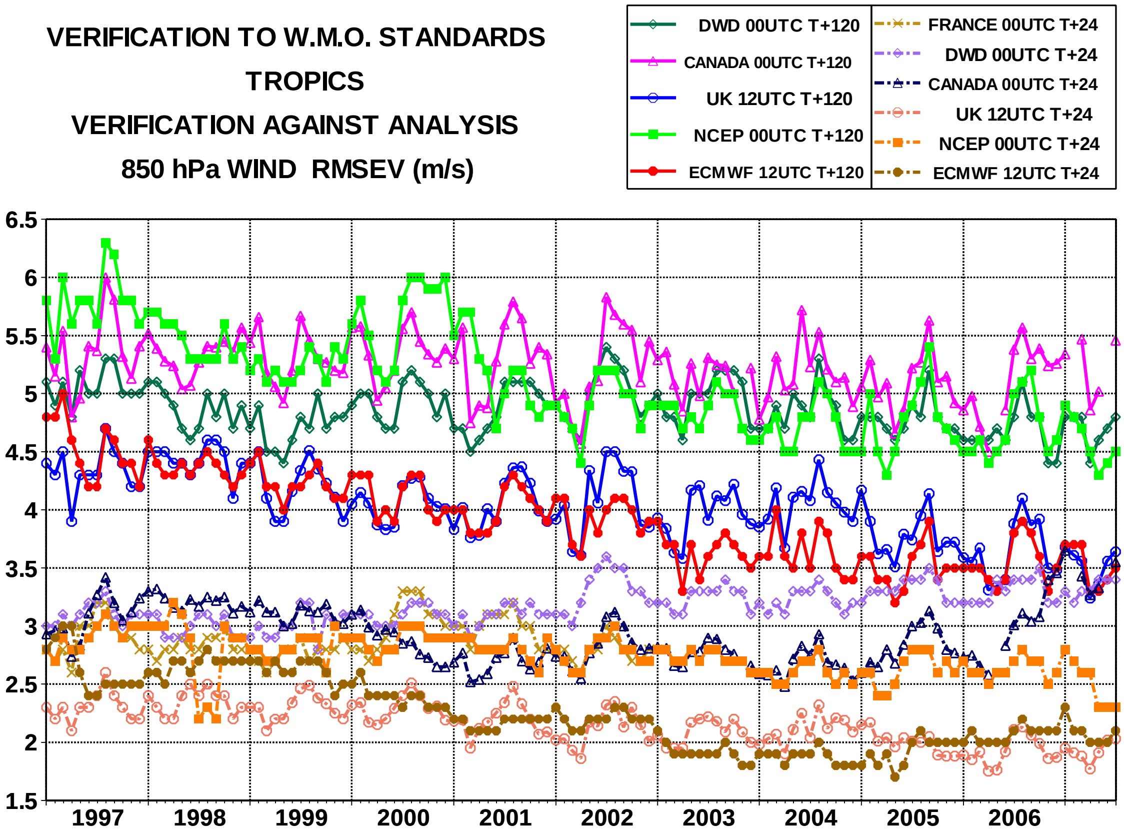


Figure 14: WMO/CBS exchanged scores (RMS vector error over the tropics, 250hPa and 850hPa wind forecast for day 1 and day 5).

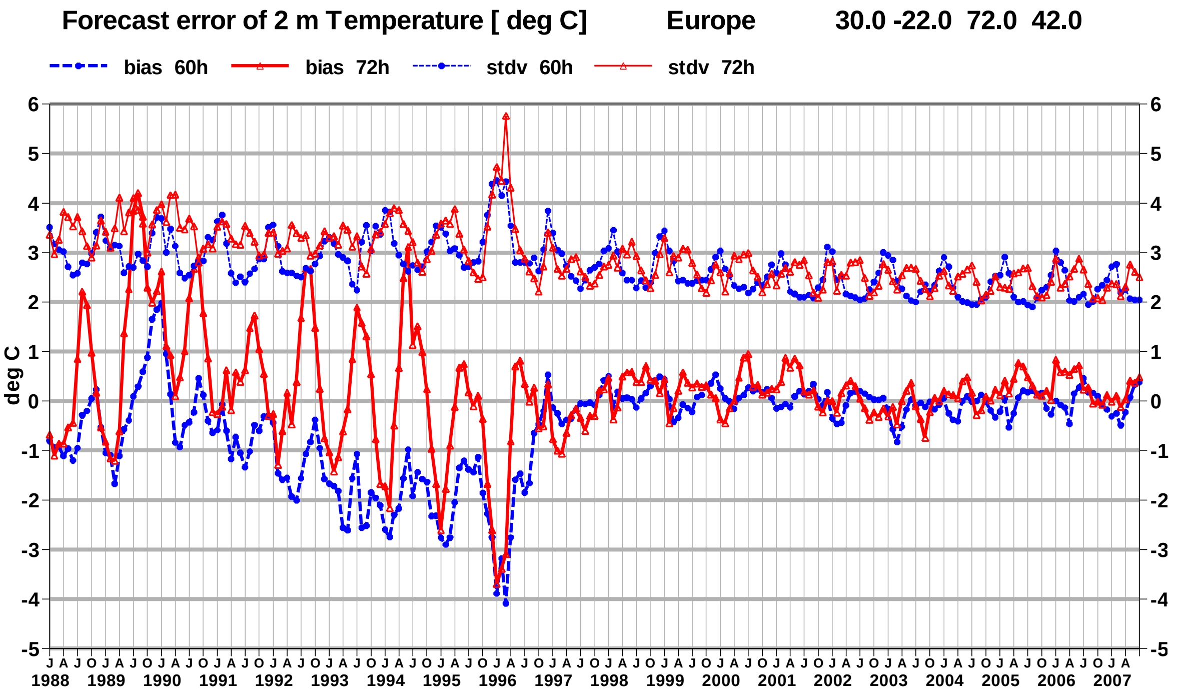


Figure 15: Verification of 2 metre temperature forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

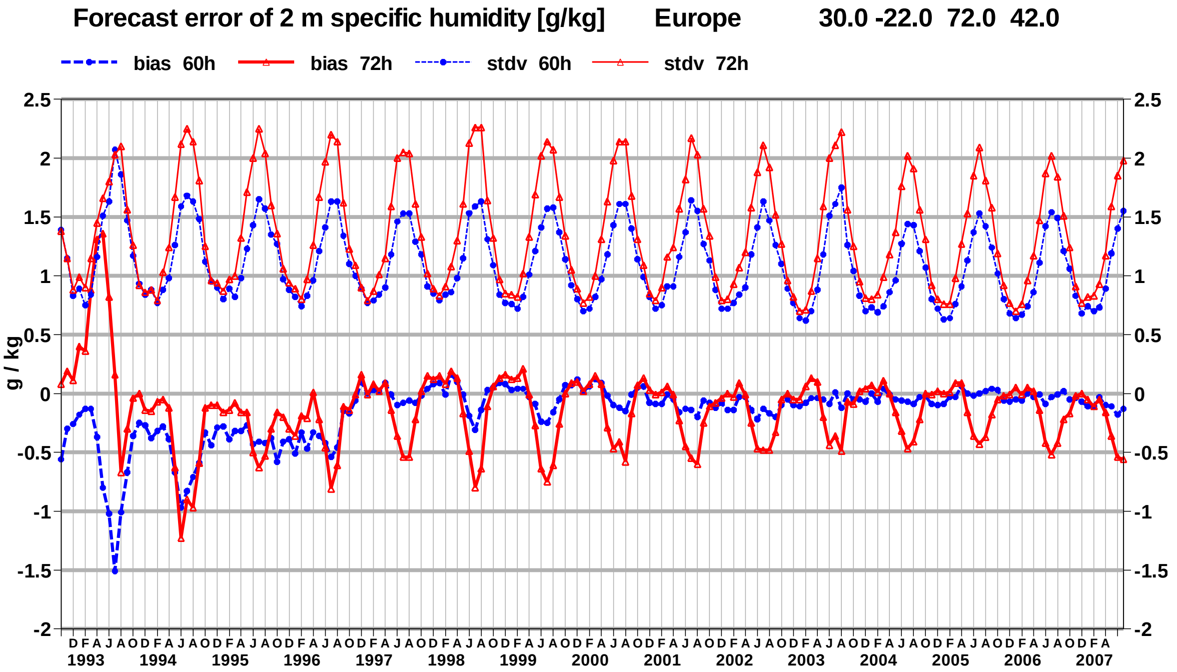


Figure 16: Verification of 2 metre specific humidity forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

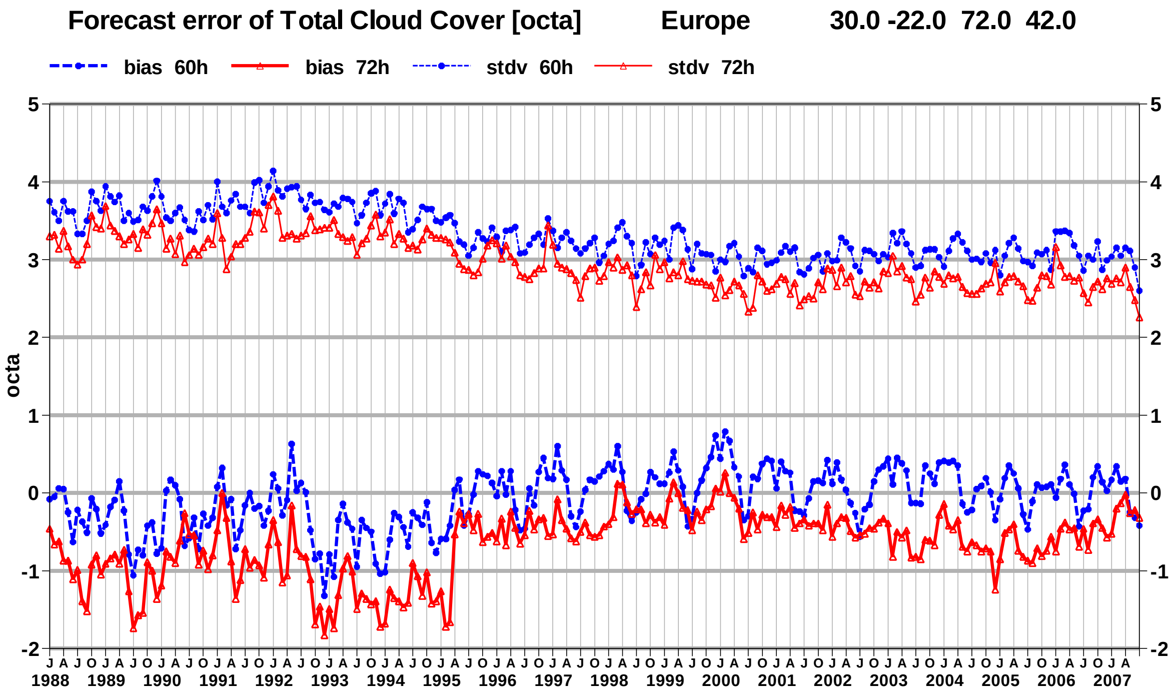


Figure 17: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

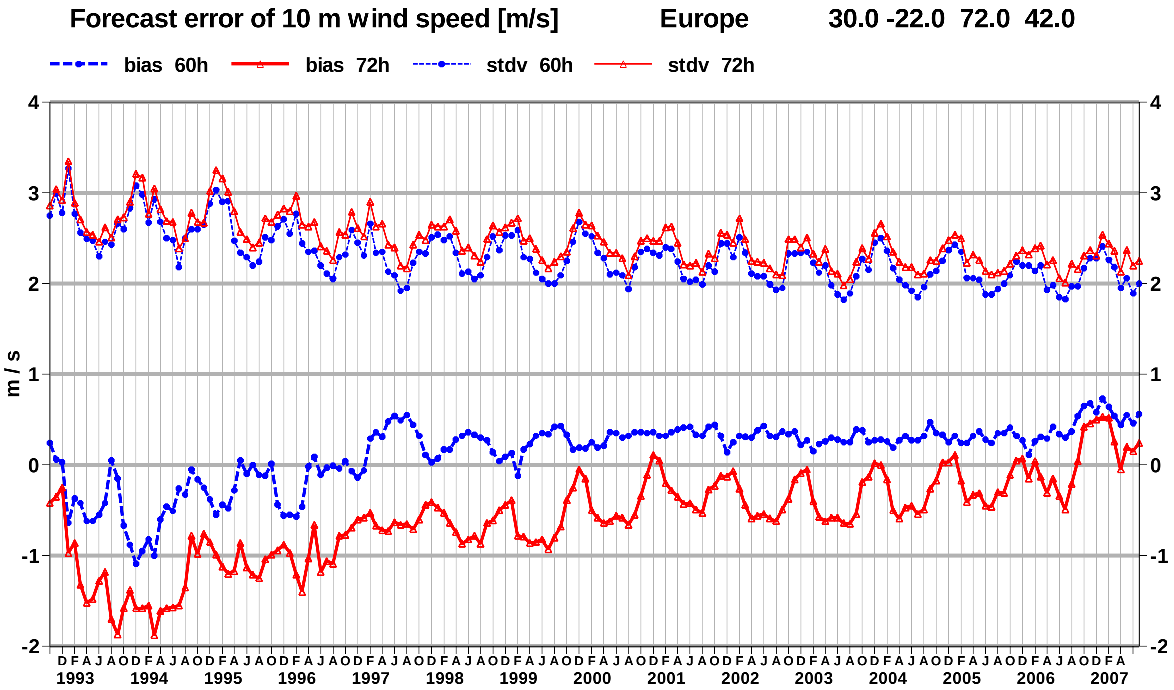
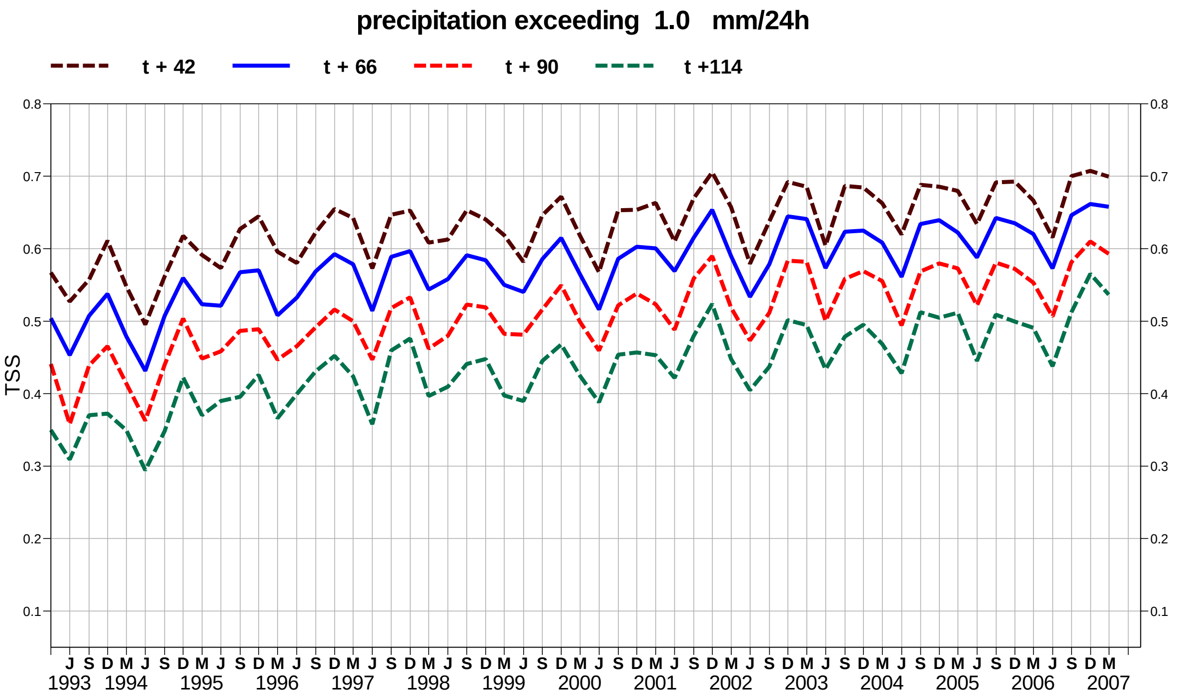


Figure 18: Verification of 10-metre wind speed forecasts against European SYNOP data on the GTS for 60-hour (nighttime) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.



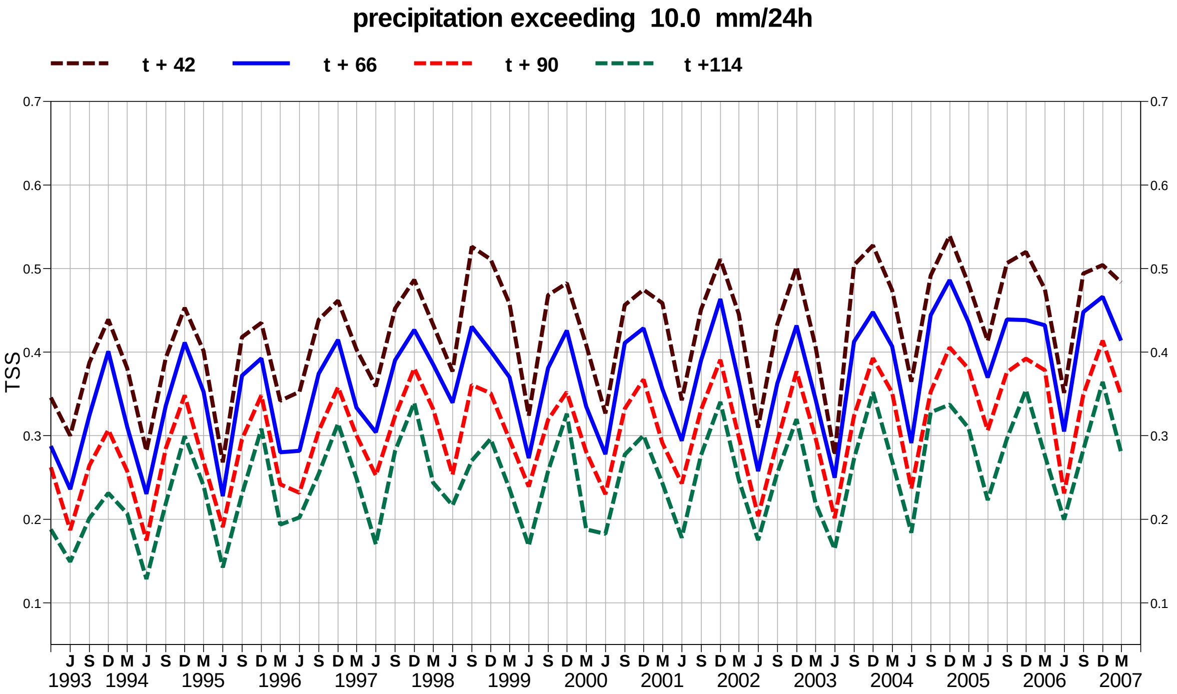
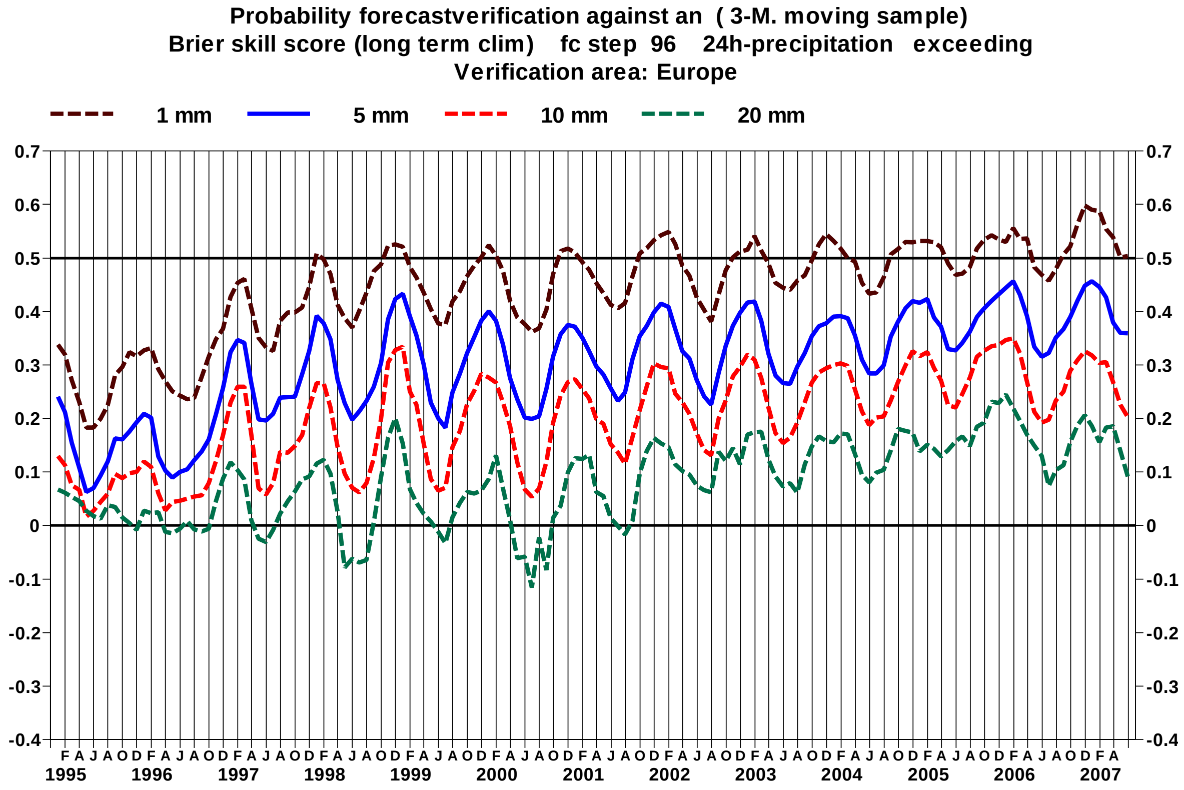


Figure 19: TSS time series for precipitation forecasts exceeding 1mm/day (top) and 10mm/day (bottom) verified against SYNOP data on the GTS for Europe. Curves are shown for the 24-hour accumulations up to 42, 66, 90, and 114 hours (from the forecasts starting at 12 UTC). 3-month mean scores (last point is March-May 2007).



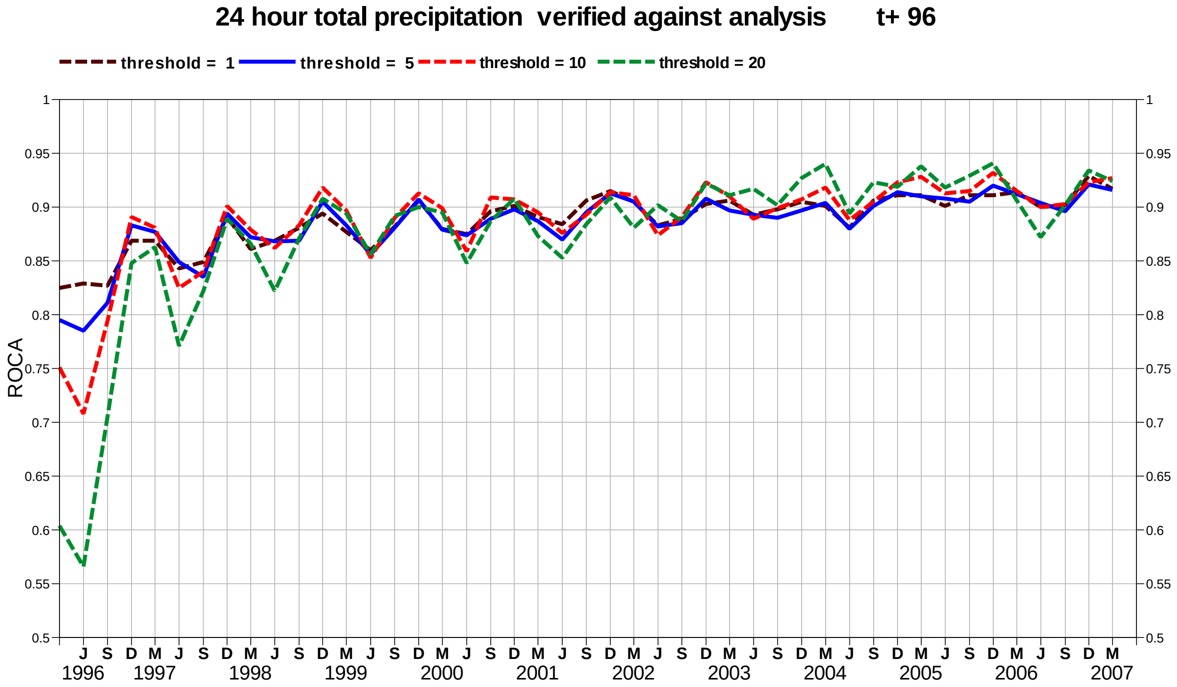
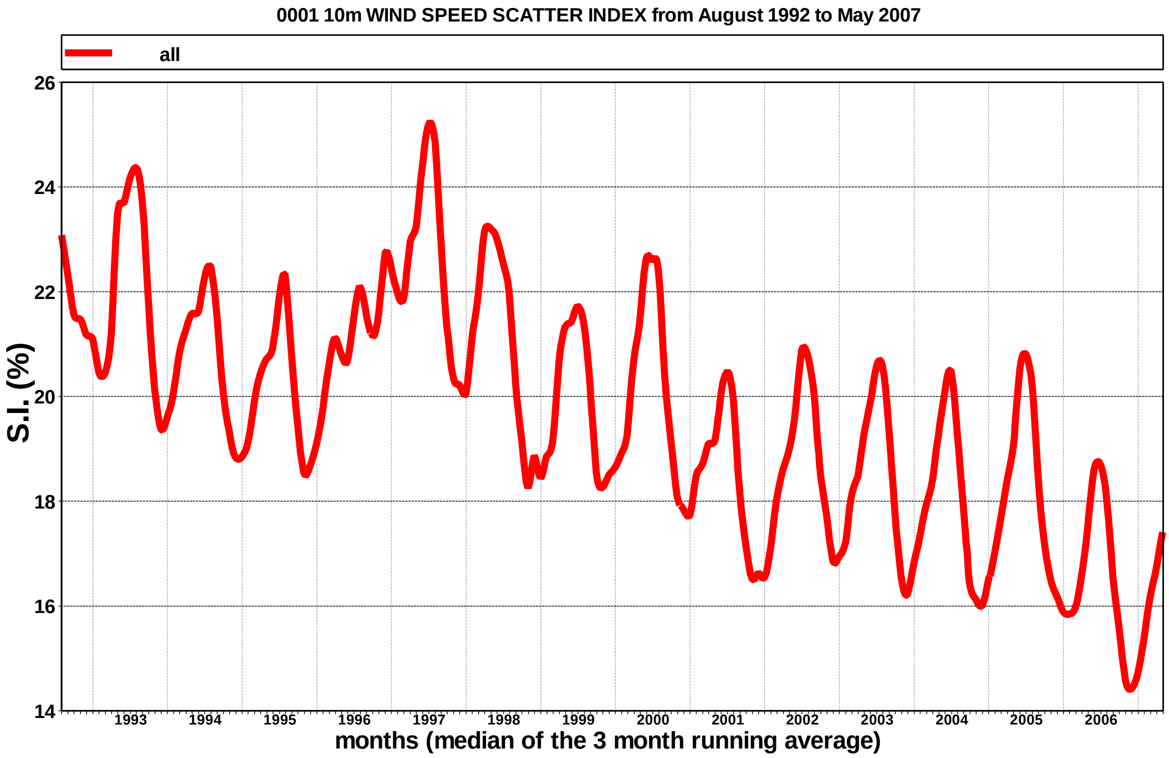


Figure 20: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA) for EPS probability forecasts of precipitation over Europe exceeding thresholds of 1, 5, 10 and 20 mm/day at day 4. The skill score is calculated for three-month running periods.



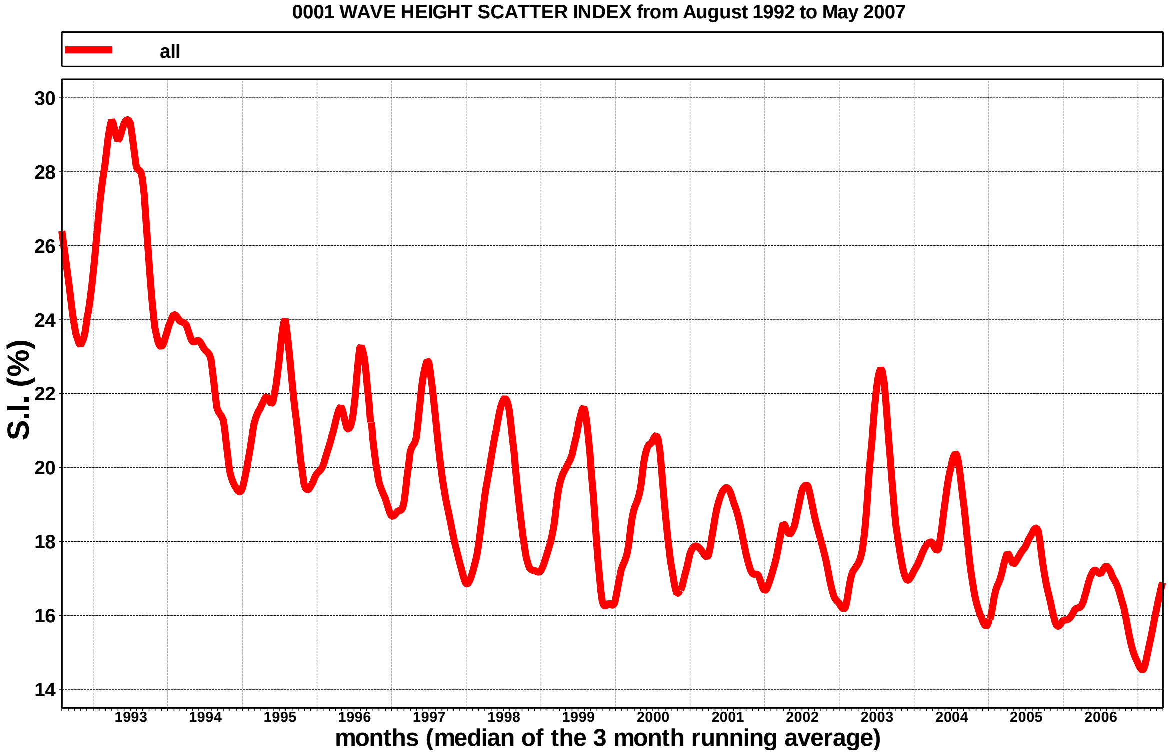
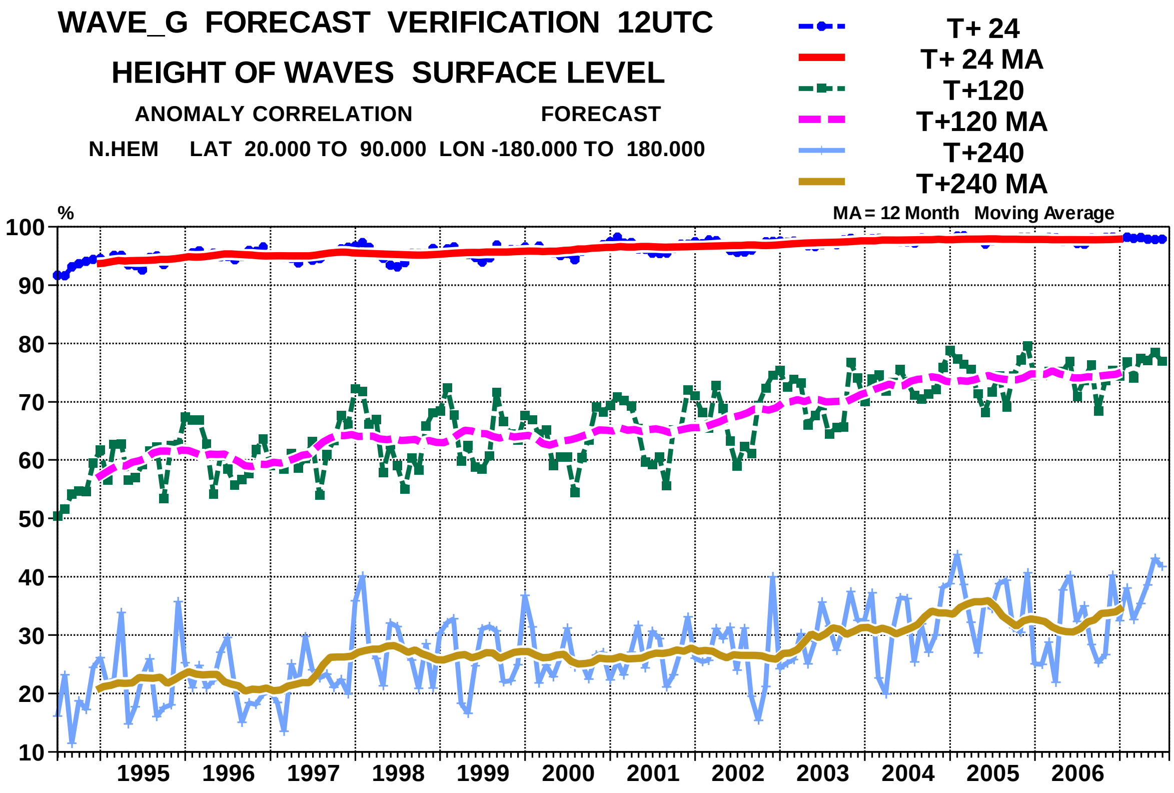


Figure 21: Time series of verification of the ECMWF 10 metre wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.



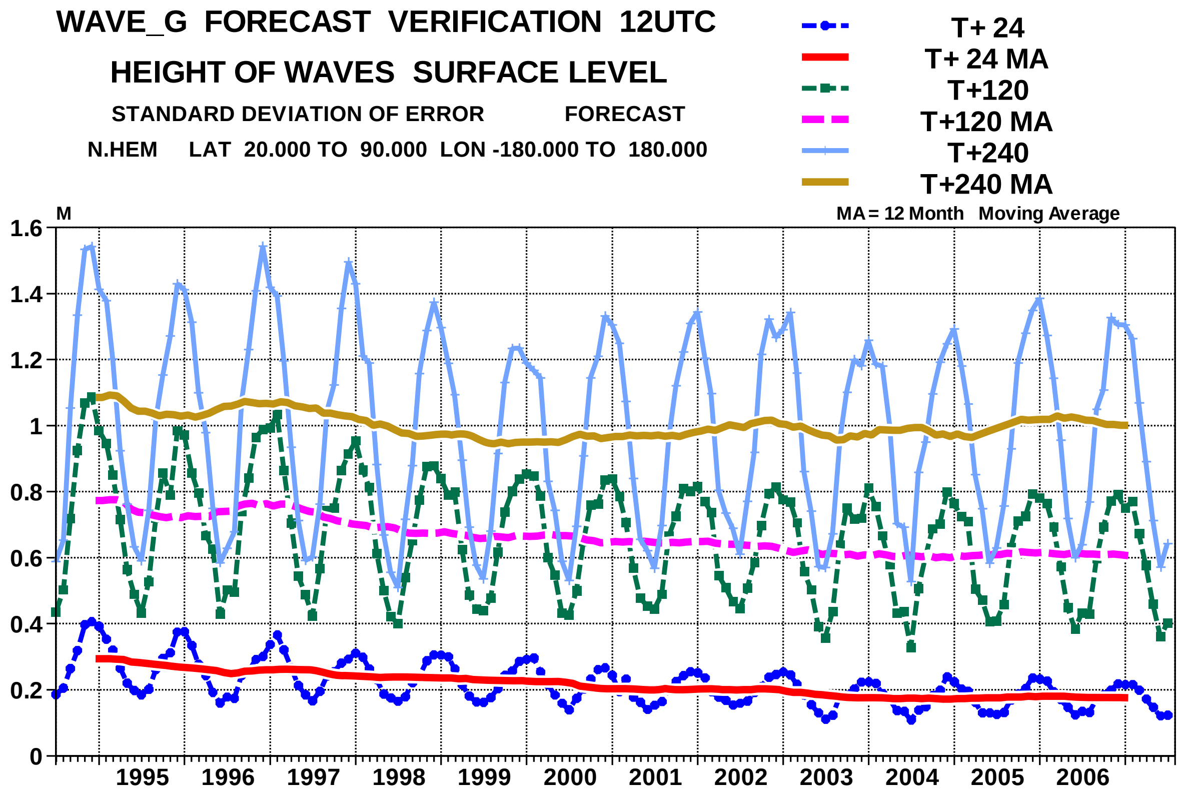
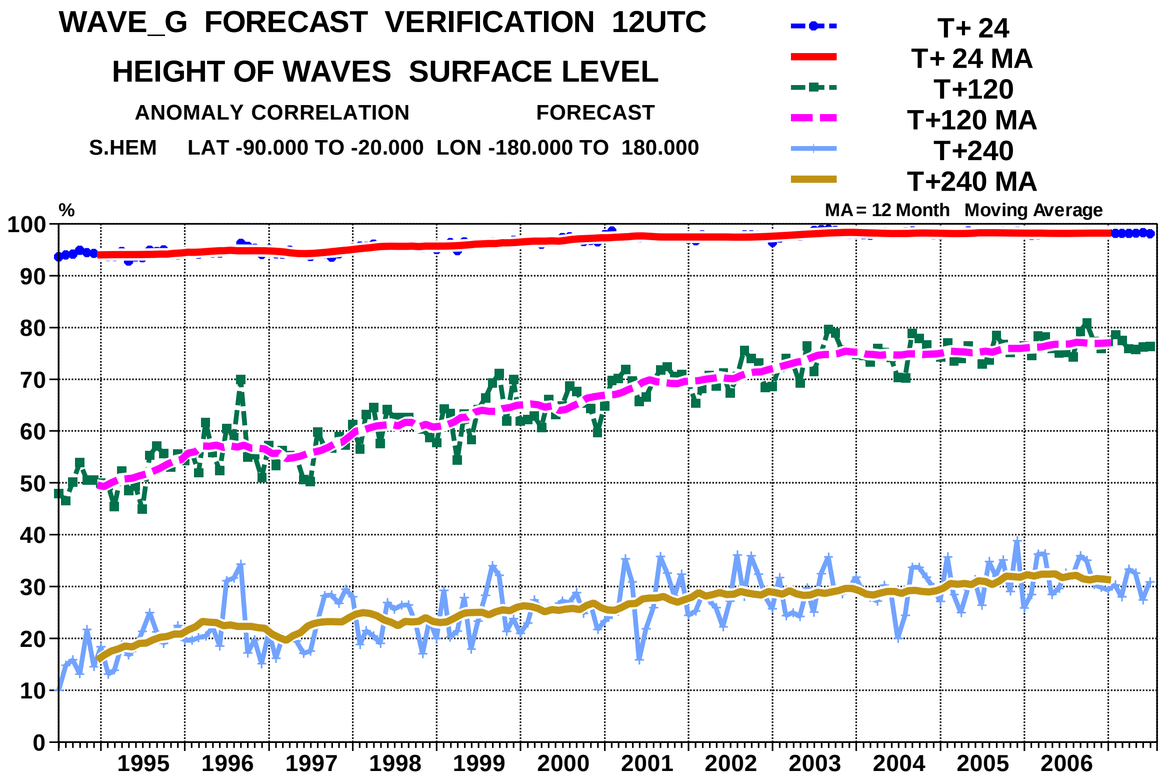


Figure 22: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (northern extratropics)



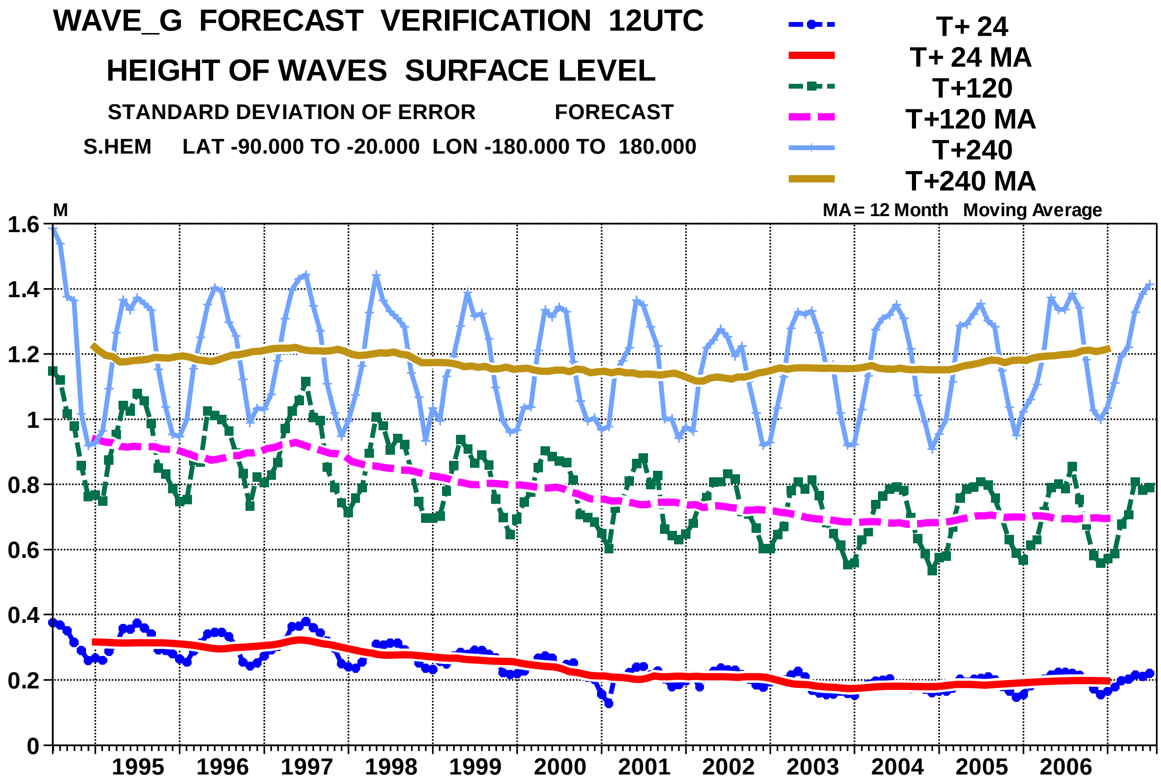


Figure 23: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (southern extratropics)

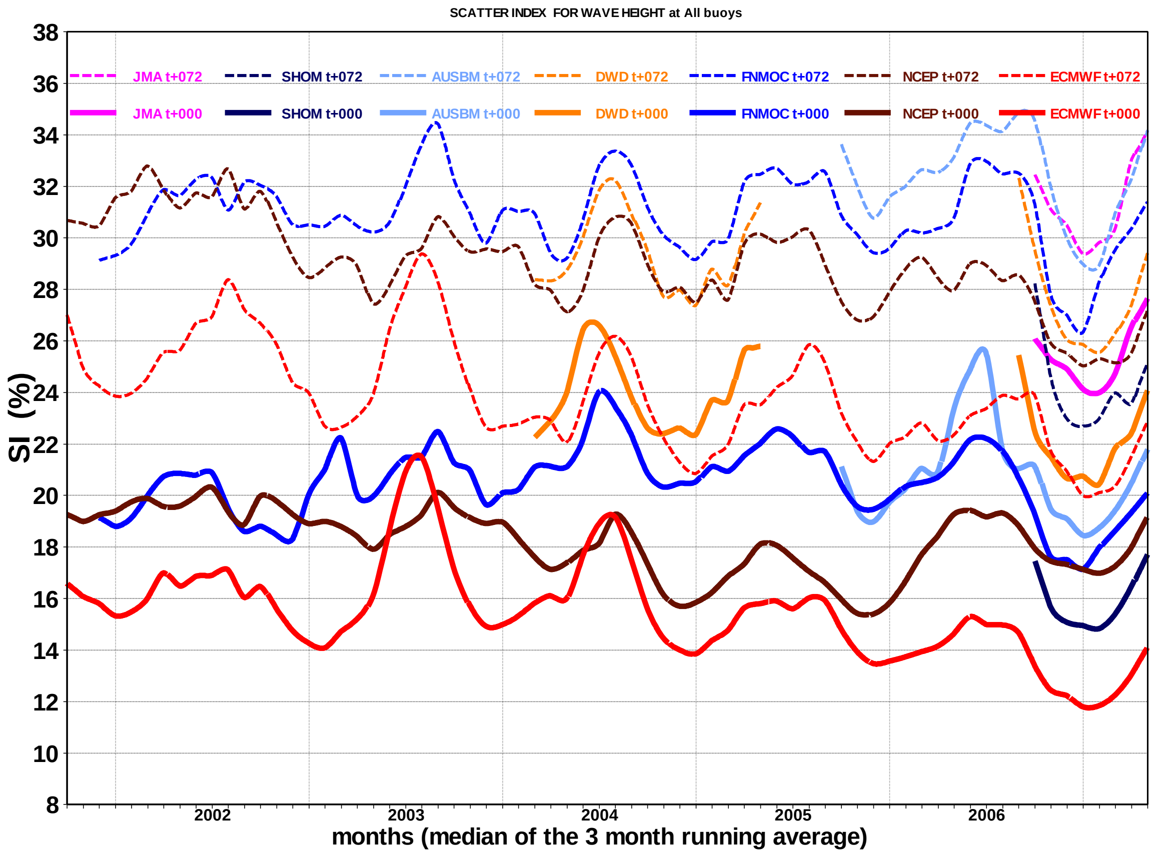


Figure 24: Verification of different model wave height forecasts using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; a three-month running mean is used.

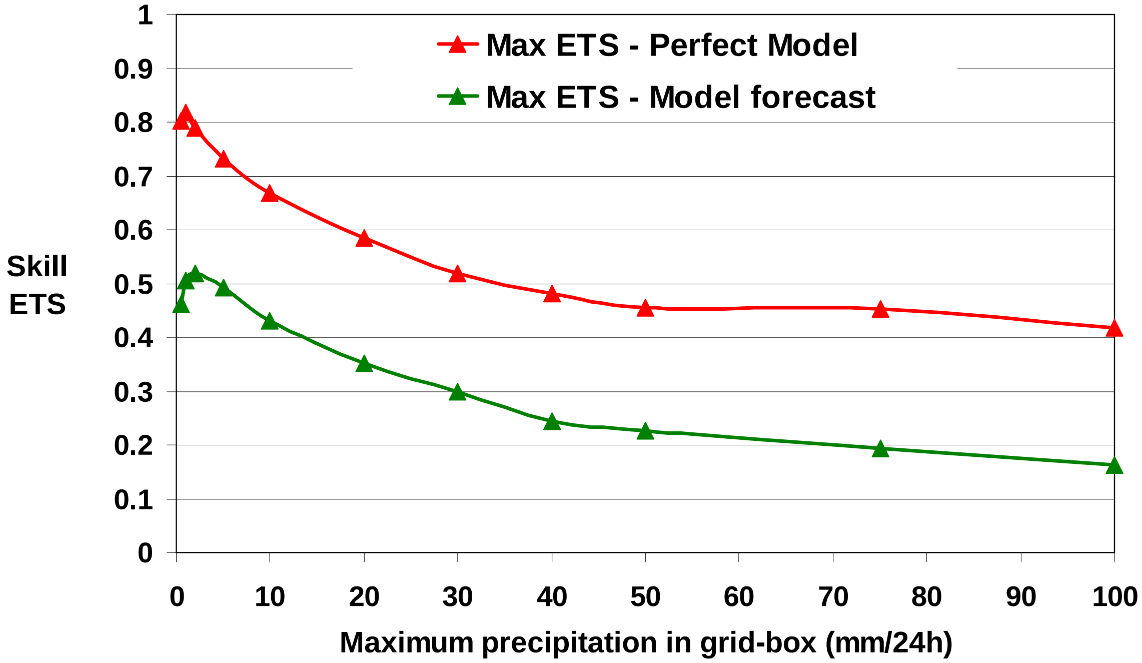


Figure 25: Verification of perfect model precipitation forecasts: equitable threat score (ETS) for forecasts of precipitation exceeding different thresholds for a "perfect model" (red) and for the 42-hour ECMWF operational forecast (green) for winter 2004-05.

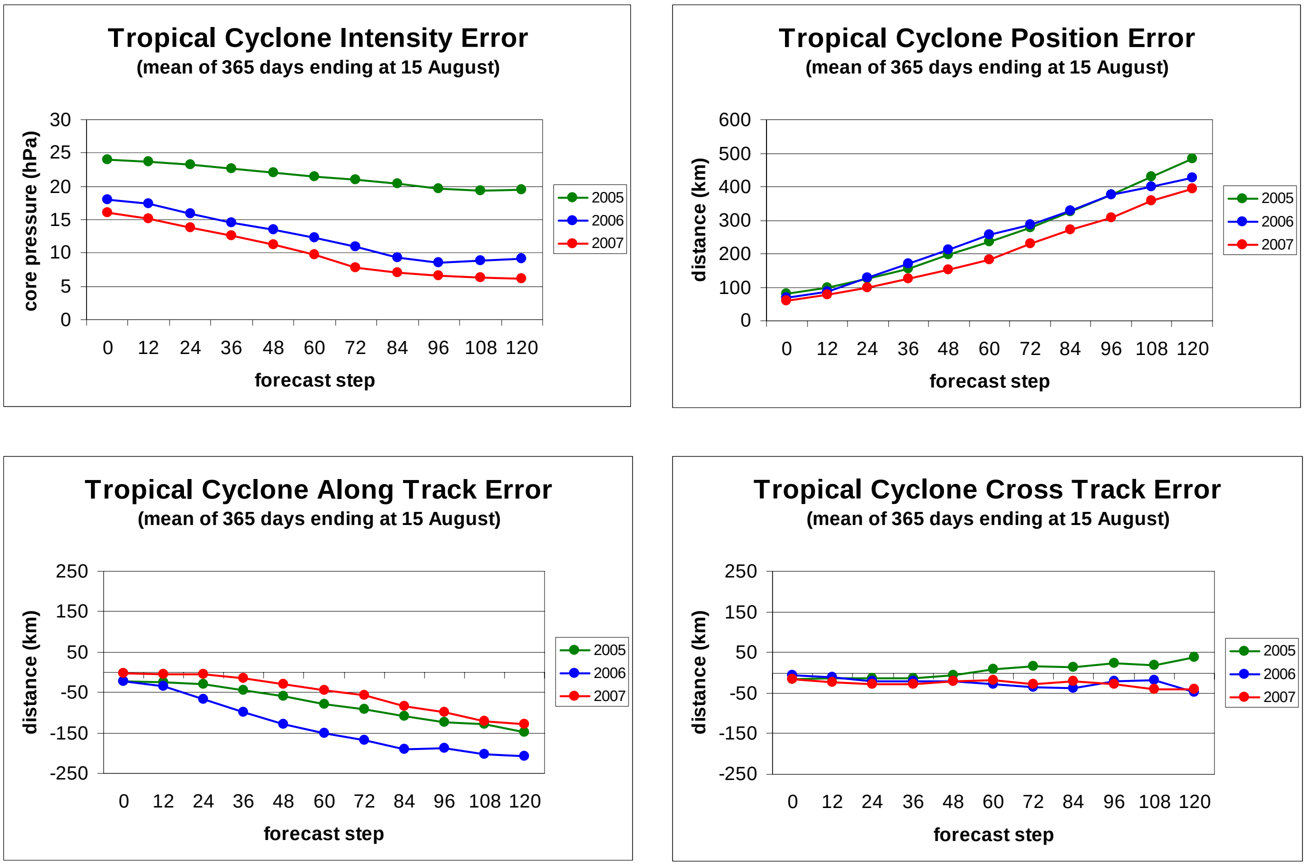


Figure 26: Verification of tropical cyclone predictions from the operational deterministic forecast for three 12-month periods: August 2004 - August 2005 (green), August 2005 - August 2006 (blue) and August 2006 - August 2007 (red). The upper panel shows the mean error in core pressure (left) and position (right). The lower panel shows the mean error in the direction of travel of the cyclone (along track error; negative values indicate slow bias) on the left and at right-angles to the direction of travel (cross track error) on the right.

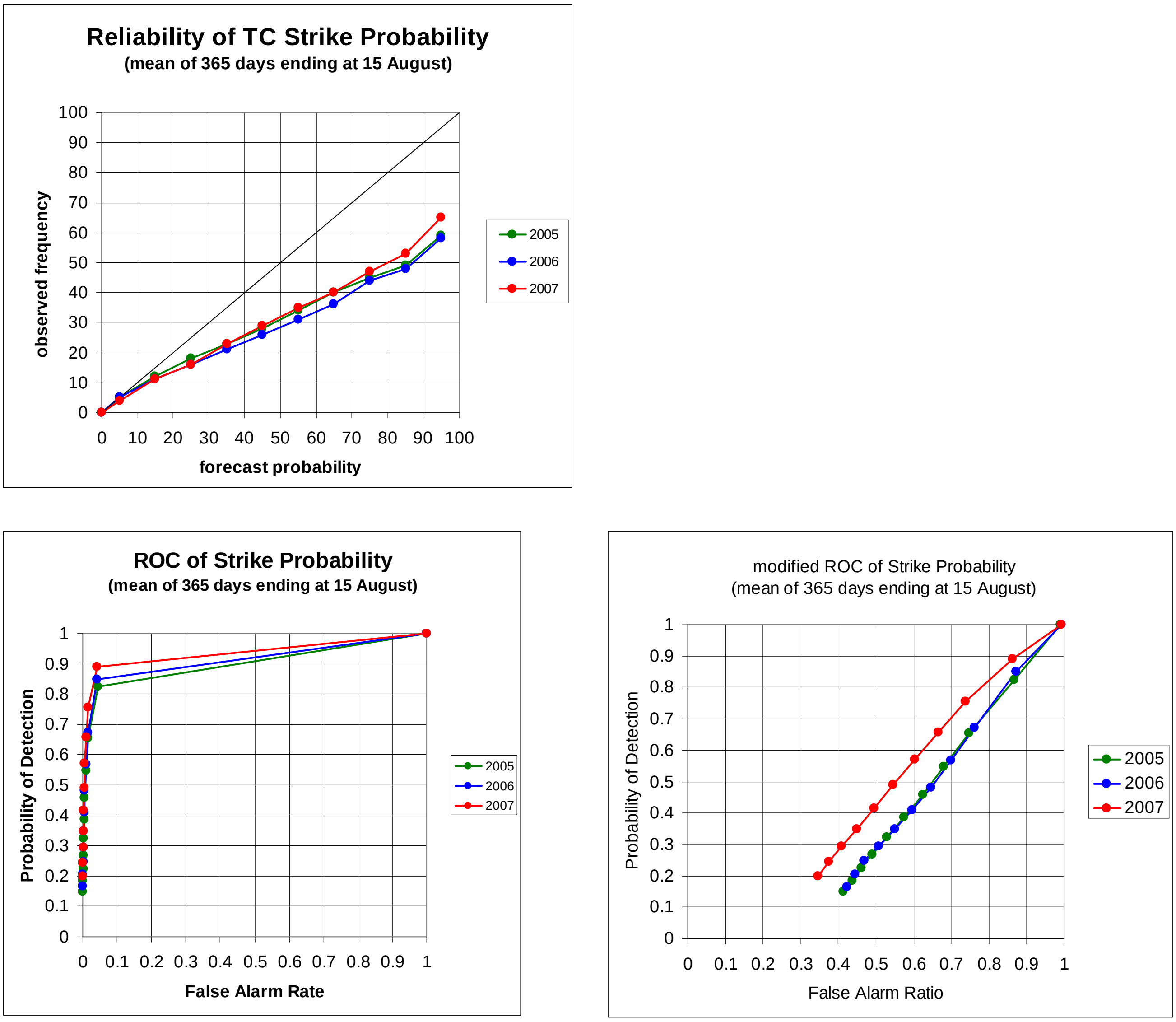


Figure 27: Probabilistic verification of EPS tropical cyclone forecasts for three 12-month periods: August 2004 - August 2005 (green), August 2005 - August 2006 (blue) and August 2006 - August 2007 (red). Upper panel shows reliability diagram (the closer to the diagonal the better). The lower panel shows (left) the ROC diagram (the closer to the upper left corner the better) and the modified ROC, where the false alarm ratio is used instead of the false alarm rate in the standard ROC.

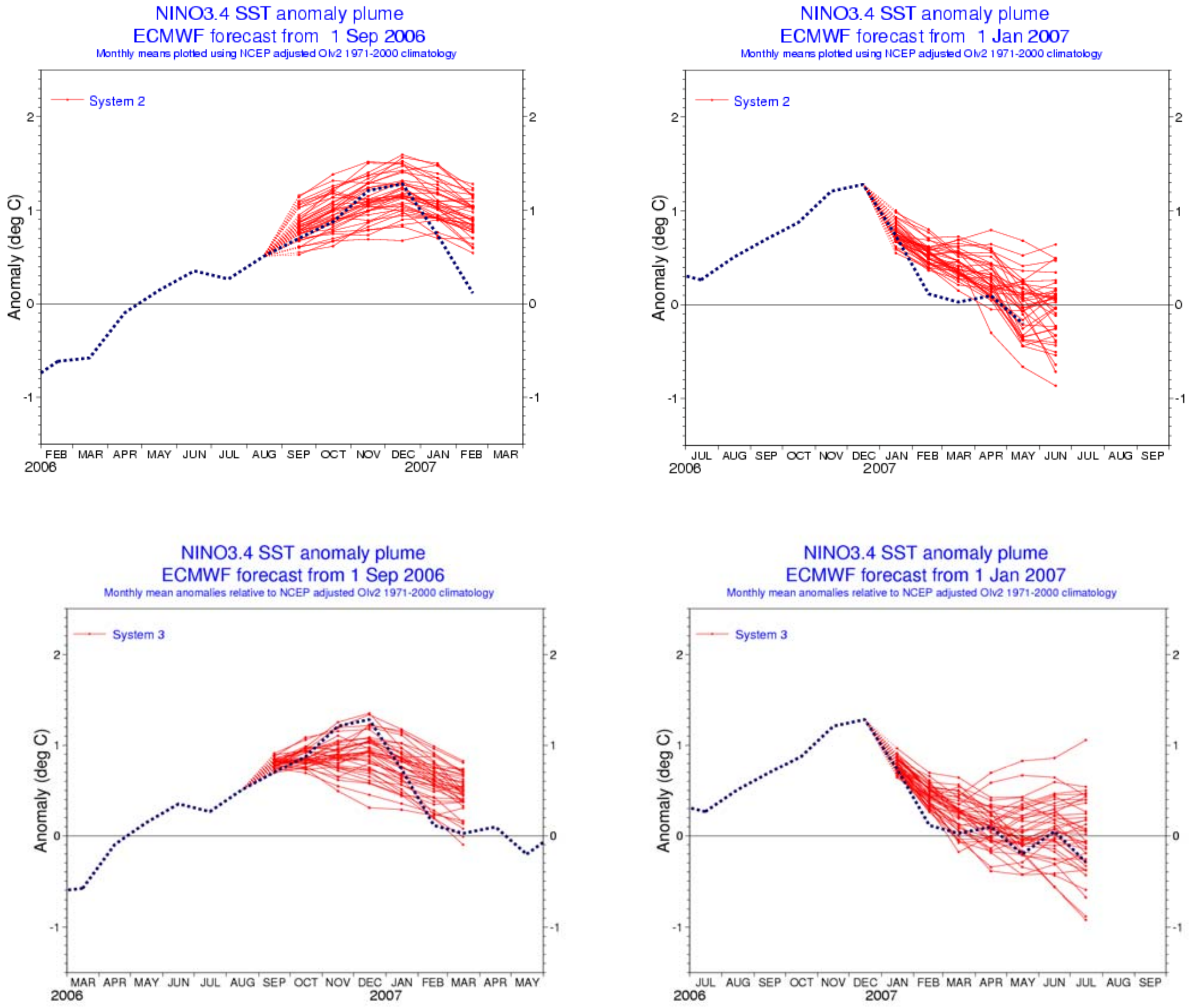


Figure 28: Plot of forecasts of SST anomalies over the Nino 3.4 region of the tropical Pacific from start dates September 2006 (left) and January 2007 (right) from seasonal forecast System 2 (top) and System 3 (bottom). The red lines represent the 40 ensemble members; dashed blue lines show the subsequent verification.

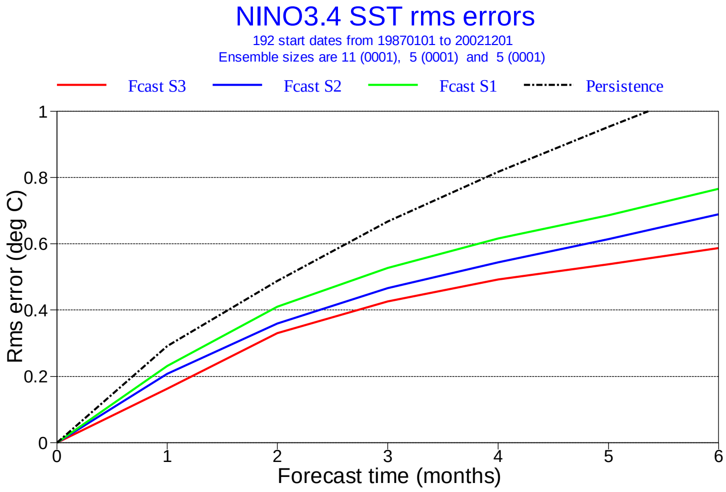


Figure 29: Root mean square errors for forecasts of SST in the Nino 3.4 region of the tropical Pacific from seasonal forecast System 1 (green), System 2 (blue) and System 3(red). Errors are averaged over a common set of 192 cases from start dates between 1987 and 2002. The error of a persistence forecast is also shown (black).

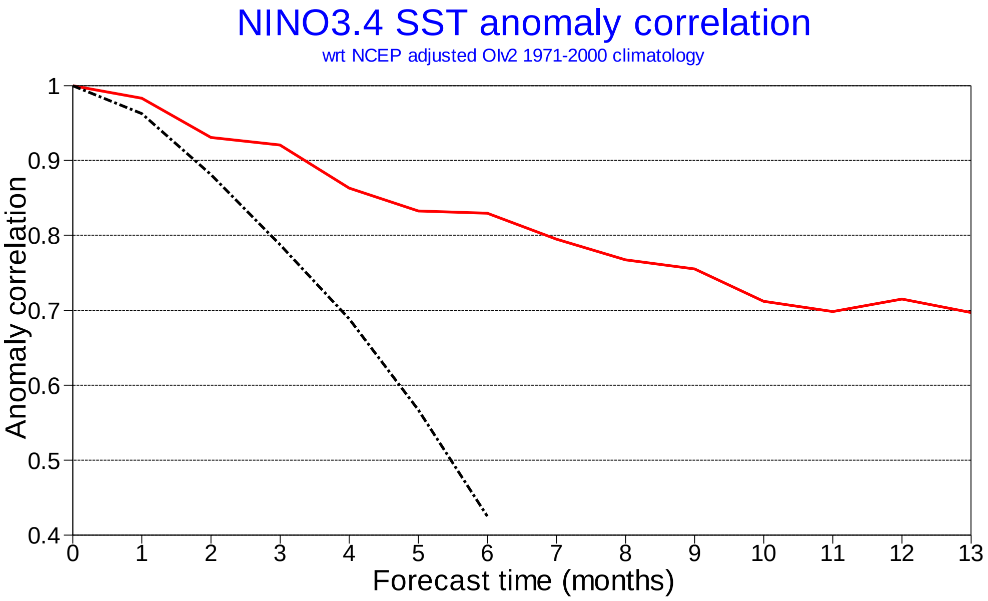


Figure 30: Anomaly correlation for forecasts of SST in the Nino 3.4 region of the tropical Pacific from the System 3 annual-range forecast system (red), compared to a persistence forecast (black), averaged over a set of cases with start dates between 1981 and 2005.

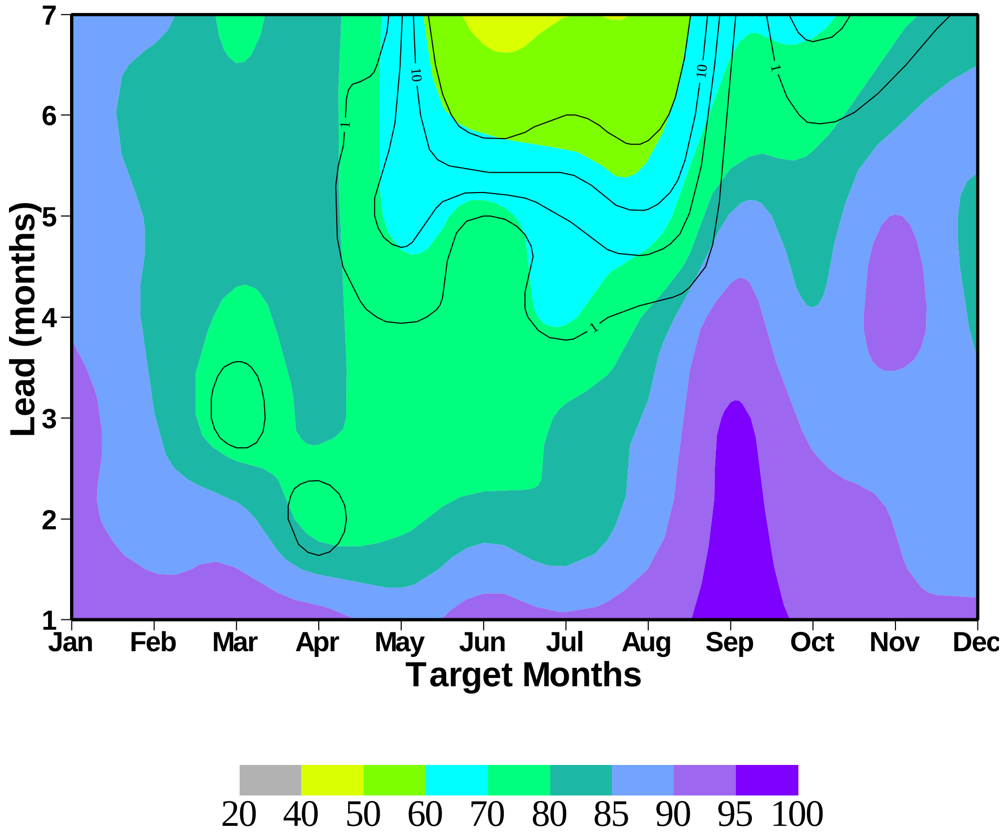


Figure 31: Anomaly correlation (%) of the ensemble-mean forecasts of the monthly mean Southern Oscillation Index (SOI), displayed as a function of forecast lead in months (vertical axis) versus the "target" or verification month (horizontal axis). The correlations are computed using the hind-cast integrations covering the period 1981-2005. Black solid lines indicates the probability of 1%, 5%, 10% and 20% that the forecast has no skill (estimated from a randomized sample of 10,000 cases).

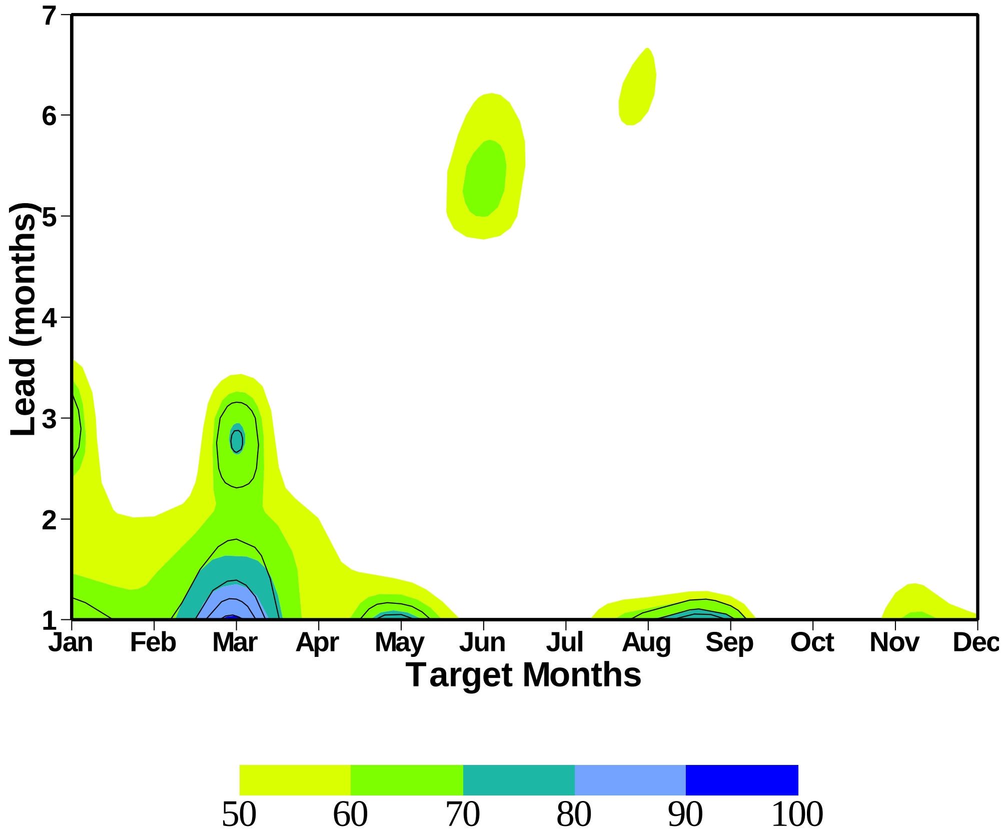


Figure 32: Area under ROC (%) for the lower tercile category of monthly mean PNA predictions, displayed as a function of forecast lead in months (vertical axis) versus the "target" or verification month (horizontal axis). The correlations are computed using the hind-cast integrations covering the period 1981-2005. Black solid lines indicates the probability of 1%, 5% 10% and 20% that the forecast has no skill (estimated from a randomized sample of 10,000 cases).

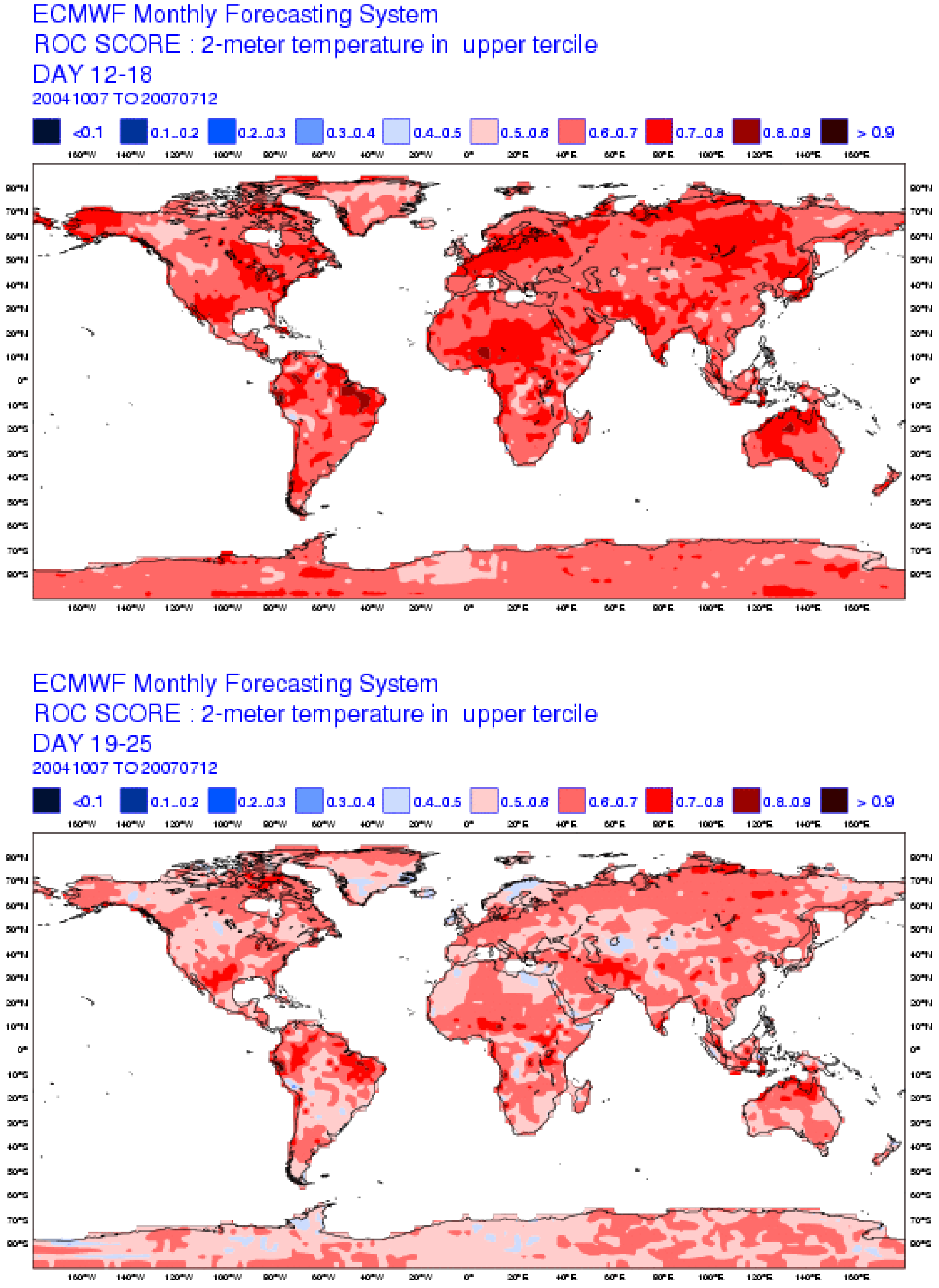

ECMWF Monthly Forecasting System
ROC SCORE : 2-meter temperature in upper tercile
DAY 12-18
20041007 TO 20070712
<0.1
0.1..0.2
0.2..0.3
0.3..0.4
0.4..0.5
0.5..0.6
0.6..0.7
0.7..0.8
0.8..0.9
> 0.9
160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E
80°N 70°N 60°N 50°N 40°N 30°N 20°N 10°N 0° 10°S 20°S 30°S 40°S 50°S 60°S 70°S 80°S
ECMWF Monthly Forecasting System
ROC SCORE : 2-meter temperature in upper tercile
DAY 19-25
20041007 TO 20070712
<0.1
0.1..0.2
0.2..0.3
0.3..0.4
0.4..0.5
0.5..0.6
0.6..0.7
0.7..0.8
0.8..0.9
> 0.9
160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0° 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E
80°N 70°N 60°N 50°N 40°N 30°N 20°N 10°N 0° 10°S 20°S 30°S 40°S 50°S 60°S 70°S 80°S

Figure 33: 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 12 July 2007 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicates positive skill compared to climate.

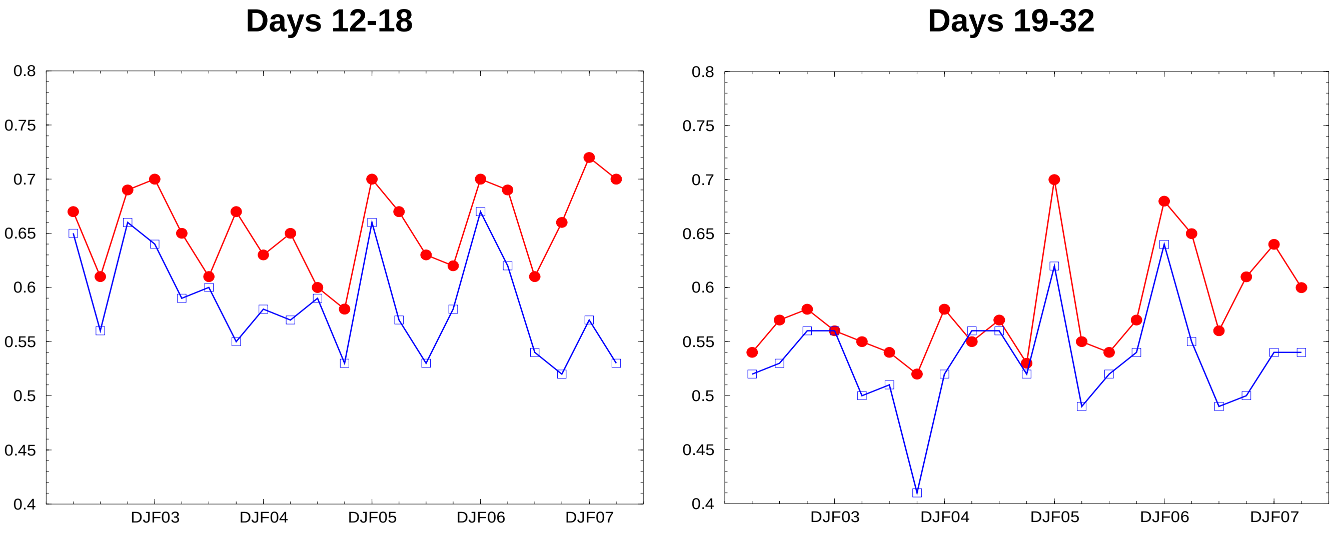


Figure 34: Area under ROC for the probability that 2-metre temperature is in the upper third of the climate distribution. Scores are calculated for each 3-month season since spring (March-May) 2002 for all land points in the extra-tropical northern hempishere. The red line shows the score of the operational monthly forecasting system for forecast days 12-18 (7-day mean) and 19-32 (14-day mean). As a comparison, the blue line shows the score using persistence of the preceding 7-day or 14-day period of the forecast.

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 13), 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 13, Figure 14) 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. For ocean waves (

Figure 22,

Figure 23) the climate has 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 (BSS), Ranked Probability Skill Scores (RPSS), and the area under Relative Operating Characteristic (ROC) curves.

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

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

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:

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

Where BS_{cl} is the Brier Score for a climate forecast (forecast probability is constant and equal to the climatological probability of the event). Time series of the Brier Skill Scores can be found in

Figure 20.

For multiple-category events, the Ranked Probability Score (RPS) is used. The RPS measures the distance between cumulative probabilities over the set of (k) events.

$$RPS = \frac{1}{k-1}\sum_{k}\left(\sum_{j\le k} p_j - \sum_{j\le k} o_j\right)^2$$

The RPS is equivalent to the average of the Brier Scores for exceeding the thresholds that separate the categories. The Ranked Probability Skill Score (RPSS) is defined similarly to the BSS, with the reference score being the RPS for a constant forecast of the climatological probability for each category. For the EPS, upper-air verification, the climatology is based on ERA-40 analyses for 1979-2001. The RPS uses 10 climatologically equally-likely categories, so is equal to the average of BS for exceeding 10, 20, 30, …, 90 % of the climate distribution. The RPSS thus gives an overall measure of the probabilistic skill of the EPS at predicting a range of events.

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 27). Figure 27 also shows a “modified ROC” plot of hit rate against false alarm ratio.

Since 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 20 and Figure 34.

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.