TECHNICAL MEMORANDUM

578

Verification statistics and evaluations of ECMWF forecasts in 2007-2008

D.S. Richardson, J. Bidlot, L. Ferranti, A. Ghelli, M. Janousek, M. Leutbecher, F. Prates, F. Vitart and E. Zsoter

Operations Department

November 2008

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 ECMWF medium-range free atmosphere 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 the annex to this document.

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, 547).

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 the last report are summarised below.

6 November 2007: Cycle 32r3, including the following main changes:

- New formulation of convective entrainment and relaxation timescale
- Reduction in free atmosphere vertical diffusion
- New soil hydrology scheme
- New radiosonde temperature and humidity bias correction
- Increase in number of radio occultation data from COSMIC
- Assimilation of AMSR-E, TMI and SSMIS window channels (clear sky)
- Assimilation of SBUV (NOAA-17, NOAA-18) and monitoring of OMI ozone data
- Initial EPS perturbation amplitude reduced by 30%
- EPS singular vectors targeted on tropical cyclones computed with the new moist physics package in the tangent-linear and adjoint models

29 January 2008: revised production schedule; product availability times improved by 10 to 25 minutes, depending on product type.

11 March 2008: integration of Monthly Forecasting System with the medium-range Ensemble Prediction System (EPS). The main changes are:

- Daily ocean-coupling for day 10 to 15 of EPS forecasts from 00 UTC analyses

- Use of persisted SST anomalies in all uncoupled atmospheric forecasts
- Modified EFI products using the new unified re-forecasts
- New GRIB description for all monthly forecast products, analogous to existing medium-range EPS data

20 May 2008: Operational assimilation of MetOp GRAS radio occultation data.

3 June 2008: Cycle 33r1, including the following main changes:

- Improved moist physics in tangent linear/adjoint model used in 4D-Var assimilation
- Re-tuned entrainment in convection scheme
- Bug fix to scaling of freezing term in convection scheme
- Additional shear term in diffusion coefficient of vertical diffusion
- Increased turbulent orographic form drag
- Fix for soil temperature analysis in areas with 100% snow cover
- Change in surface roughness for momentum and change in post-processing of two-metre temperature and specific humidity
- Assimilation of AMSR-E and TMI radiances in 1D+4D-Var; assimilation OMI ozone data
- Usage of all four wind solutions for QuikSCAT in assimilation, rather than only two previously
- Extended coverage and increased resolution for the limited area wave model
- Improved shallow water physics and modified advection scheme for ocean wave models
- Introduction of two new wave model parameters: maximum wave height and corresponding wave period

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 July 2008. 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 continued high skill, in particular consolidation of the increased skill over Europe. Figure 3 shows that synoptic activity was relatively high over Europe in the past year, following a period of relatively persistent flow. Since the skill in Figure 1 is relative to the persistence error, such variations in

synoptic activity can affect the scores, and may contribute to the peak of skill in 2007. It is reassuring that Figure 2 also shows continued high skill for Europe.

The consistently good performance over the last year can be seen in the scores for the individual months in Figure 2. 2007 was the first year where the anomaly correlation remained above 60% for at least 7.5 days every month over the northern hemisphere (7 days for Europe). This consistently good performance has been maintained through the first half of 2008.

One notable reason for the overall high scores is a continuing reduction in the number of poor individual forecasts and corresponding increase in the 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 4 and 3 day (and 6 and 5 day) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the general downward trend indicates that there is less 'jumpiness' in the forecast from day to day. There was a small increase in this measure following the introduction of model cycle 32r3 in November 2007, consistent with the increase in model activity in that cycle. Previous cycles underestimated activity slightly in mid-latitudes, and more significantly in the tropics. Changes to the physical parametrizations in 32r3 addressed these deficiencies.

The quality of ECMWF forecasts for the upper atmosphere in the extratropics is shown through the time series of wind scores at 50hPa in Figure 6. In both hemispheres, scores for the last year are similar to those for the previous year.

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: the higher skill level has been maintained throughout the past year.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extra-tropical northern hemisphere for the last three winters are shown in Figure 8. The increase in model activity in Cycle 32r3 (introduced in November 2007) resulted in a significant increase in ensemble spread. The amplitude of the initial perturbations was therefore reduced (by 30%) to maintain the agreement between spread and error. This change resulted in an improved match between spread and error for 500 hPa height; in particular the substantial over-dispersion of the EPS in the early forecast range in previous cycles is no longer apparent. There is now more under-dispersion of the EPS for temperature at 850 hPa to around day seven. However, analysis uncertainty should be taken into account to obtain a reliable relationship between spread and error in the first few days.

Figure 9 shows the skill of the EPS using the Ranked Probability Skill Score (RPSS) for days 1 to 15 for winter over the extra-tropical northern hemisphere. In November 2006 the EPS was extended to 15 days, at reduced horizontal resolution beyond day 10. Skill in the extended range for winter 2007-08 matches that for 2006-07, confirming the positive skill at this forecast range.

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. The increase in error at 850 hPa at the end of 2007 is associated with the introduction of cycle 32r3. Changes to the physical parametrizations in this cycle increased model activity to higher but more realistic levels, especially in the tropics.

3.2 ECMWF vs other NWP centres

The common ground for such a comparison is the regular exchange of scores between WMO designated Global Data-processing and Forecasting Systems (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. ECMWF continues to maintain its lead over the other centres; however both the Met Office and NCEP show noticeable improvements in winter 2007-08, resulting in a smaller gap than most recent winter periods. The gap is, in general, bigger in the southern extratropics (Figure 12). Most centres show smaller errors in the southern hemisphere 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 regions such as Europe. Figure 13, showing both 500 hPa height and 850 hPa wind errors averaged over the past 12 months, confirms the good performance of the ECMWF forecasts using this alternative reference relative to the other centres.

The situation in the tropics is summarised in Figure 14. Since mid-2005, the Met Office has had the lowest short-range errors, while at day 5 ECMWF and the Met Office performance is similar. 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. In general the performance over the past year follows the trend of previous years and there is no adverse effect on the scores from the increase in atmospheric activity introduced in model cycle 32r3 in November 2007. The negative daytime bias in cloud cover has reduced substantially since 2005 (Figure 17) and is now close to zero for both daytime and night-time; 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. Wind bias has been consistent over the past year, while night-time error standard deviations in particular have been reduced somewhat.

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 1 mm and 10 mm per day. Results for the past year confirm the striking improvement that occurred following the introduction of cycle 31r1 in September 2006. For the higher threshold of 10 mm/day the performance over the past year has been exceptional. The same overall trend 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.

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. The Met Office was added in 2007.

5 Severe weather

5.1 Extreme Forecast Index (EFI)

The Extreme Forecast Index (EFI) was developed at ECMWF as a tool to provide some general 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 risk of an extreme event occurring. The model climatology used for the EFI was changed with the integration of the EPS and monthly forecast systems in March 2008. Previously, the climate distribution was generated from a set of 48-hour EPS control forecasts run from initial conditions of the 30-year period 1971-2000. With the introduction of the unified system, the EFI uses the same set of model re-forecasts that is produced to calibrate the monthly forecast products. This climatology comprises 5-member ensembles of 32-day forecasts starting on the same day and month as the real-time monthly forecast for each of the past 18 years. 5 weeks of the re-forecasts are combined for the EFI climate. Although the sample size is substantially less that for the previous EFI climate, the new system provides model climate distribution for all forecast steps and gives the potential to extend the EFI to a longer forecast range than at present.

Verification of the EFI has been made using synoptic observations over Europe available on the GTS. An extreme event is taken as occurring if the observation exceeds the 95th percentile of the observed climate for that station (calculated from a 15-year sample, 1993-2007). The ability of the EFI to detect extreme events is assessed using the Relative Operating Characteristic (ROC). Results are presented in Figure 25 for the 2007-08 cold season (October-March) for precipitation and 10m wind. During this period, both old and new EFI

climatologies were available, allowing a comparison of the impact of the change. As can be seen from Figure 25, the change in EFI climate does not significantly affect the performance. For both precipitation and wind the EFI demonstrates substantial ability to detect extreme events. The performance decreases relatively slowly with forecast lead time.

5.2 Tropical cyclones

The 2007 North Atlantic hurricane season was close to normal despite an early start. Two major hurricanes, Dean and Felix, caused loss of life and significant damage in many parts of Central America.

Average position and intensity errors for all tropical cyclone forecasts over the three latest 12-month periods are shown in Figure 26. A significant reduction in intensity errors for 2007/2008 can be seen when compared with the two previous periods. The core pressure is notably better in the analysis and throughout the five day forecast. This fact might be related to important changes implemented in the model physics over the past year. Position errors have gradually decreased over recent years, although the initial errors (the analysis errors) have not changed significantly. There is a clear tendency in the forecast for tropical cyclones to move too slowly (negative along-track error), despite some improvements over the last few years. From the plots of along-track error we see that the performance in 2008 is slightly worse than the performance in the previous year.

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 during the next 120 hours. Verification of these probabilistic forecasts for the three latest 12-month periods is shown in Figure 27. Results show an over-confidence for the three periods, with small variations from year to year. The signal detection capability (as indicated by ROC) has improved this 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).

6 Monthly and seasonal forecasts

6.1 Monthly forecast verification statistics and performance

The monthly forecasting system has been integrated with the medium-range Ensemble Prediction System (EPS). The new combined system enables users to be provided with EPS output uniformly up to 32 days ahead, once a week. It also introduces a coupled ocean-atmosphere model for the forecast range day 10 to 15 for the forecast started from the 00 UTC analysis, on a daily basis. The integrated EPS was implemented on 11 March 2008 and the first monthly run with the new system was on 13 March.

The main changes introduced include: the use of persisted SST anomalies in all atmospheric forecasts, the daily ocean-coupling for days 10 to 15 of EPS forecasts initiated from 00 UTC analyses. For further information see:

- www.ecmwf.int/publications/newsletters/pdf/115.pdf contains an article showing a comparison between the performance of the new system and the previous monthly forecasting system.
- www.ecmwf.int/products/changes/vareps-monthly provides a technical description
- Comprehensive verification for the monthly forecasts is available on the ECMWF website at: http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/.

Figure 28 shows the ROC score computed over each grid point for the 2m 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). Currently the anomalies are relative to the past 18-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 subject to sampling, they can provide the user with a first estimate of the forecast skill's spatial distribution.

6.1.1 Monthly forecasts' performance 2007-2008

Figure 29 shows the probabilistic performance of the monthly forecast over each individual season since September 2005 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 performed consistently better than persistence of the previous 7-day or 14-day period. At days 12-18, scores reached their highest ever values for winter 2007-08. MAM 2008 scores at day 12-18 are slightly lower than the ones for MAM 2007 but this could just be due to variability from year to year. For the forecast range 19-32 days, the scores for MAM 2008 are low and closer to persistence than in previous years.

Figure 30 shows the skill in terms of ROC for the probability of 2m temperature being in the upper third of the distribution, computed for the period during which the monthly forecast has been operational. The skill has been calculated for 2 areas: northern extra-tropics (solid line) and Europe (dashed line) and for 3 forecast ranges: day 5-11 (black), day 12-18 (red) and day 19-32 (blue). A positive trend with higher skill in the latest years is evident at all forecast ranges and for both areas. For the most recent years, skill over Europe is higher than that over northern extratropics at forecast ranges 5-11 and 12-18 days. In contrast, at forecast range 19-32 the skill over Europe is lower than the skill over the northern extratropics for all the years considered. It is plausible that, at this extended forecast range, the main source of predictability is associated with the SST variability over the tropical Pacific, consequently the Pacific sector could exhibit a higher level of skill, when compared with the skill over the European sector.

6.2 The 2007-2008 El Niño forecasts

During the period 2007-08 the sea-surface temperatures (SST) across the central and eastern Pacific went through a transition phase with the development of La Niña conditions. Figure 31 shows predictions for the NINO 3.4 region (170W-120W, 5N-5S) from February 2007, June 2007, October 2007 and February 2008 with subsequent verification.

As shown in the previous report, predictions initiated in December 2006 and January 2007 successfully forecasted the El Niño's quick decline in the early part of 2007. This transition was unusually rapid. Typically transitions from warm to cold phases occur on a time scale of one year. In this case, by February 2007 the SST over the Tropical Pacific was close to its climatological value. Forecasts issued in February 2007 showed some probability of development into La Niña later in the year. Subsequent forecasts (not shown) successfully predicted the development of the cooling. Although predictions initiated in June 2007 overestimated the cooling rate in the first part of the forecast, overall they gave a realistic indication of the

intensity of the anomalies for the late part of 2007. The amplitude of the anomalies observed during La Niña 2008 are comparable with those observed during strong cold events such as 1999-2000 and 1989. After a further deepening of the cold anomalies with the minimum values observed in February 2008, the SST anomalies started to decline. Forecasts issued in October 2007, although underestimating the further cooling, successfully predicted the timing of the La Niña's decline. February 2008 predictions showed a rather realistic decline of the cold anomalies, although the return towards neutral conditions was underestimated in the late part of the forecast.

6.3 Seasonal Forecast performance for the tropical SST

Recently, area averages of sea surface temperature anomalies computed over a set of 10 'key regions' have been added to the range of seasonal forecast products at:

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

Verification of these new products is available at the same website; Figure 32 shows a few examples. The seasonal forecast skill for SST anomalies averaged over: the NINO 3.4 area (top panel), the Eastern Tropical Indian Ocean (middle panel) and the Southern Tropical Atlantic (bottom panel) is estimated by the anomaly correlation and it is displayed as a function of forecast lead time in months (vertical axis) versus the 'target' or verification month (horizontal axis). In the seasonal forecast the skill varies more strongly with the target season than with the lead time. For the NINO 3.4 area correlations exceed 80% for forecasts verifying in September to April, with leads of up to 5 months. However, forecasts for the northern hemisphere summer months, most notably for May-June, are more difficult. The sudden drop in skill near April-May, known as the spring barrier, is related to the ENSO seasonal cycle and is a feature common to many seasonal forecast systems. Although the main source of predictability on the seasonal time scale is related to the variability of SST anomalies over the tropical Pacific, tropical SST over other basins play a significant role in modulating the climate of several regions. For example, the cooling over the Eastern Tropical Indian Ocean area (90-110E, Eq.-10S) is typically associated with heavy rainfall over East Africa and severe droughts over the Indonesian region. It is interesting to see that for the Eastern Tropical Indian Ocean the anomalies exceed 80% for forecasts verifying in August to November with a lead time of up to 3 months. For the southern tropical Atlantic area (30W-15E, Eq.-20S) correlations with a lead time of 2 months exceed 80% for the March-July target months.

6.4 Seasonal forecast performance for the global domain

A set of verification statistics based on the hindcast integrations (1981-2005) from the operational System 3 has been produced and is also available on the ECMWF website at:

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

Spatial skill distributions, as well as skill estimates accumulated over pre-defined regions, are available for a large set of atmospheric parameters. The skill estimates are based on deterministic and probabilistic measures and are computed for: 1 month, 2 months, up to 4 months lead times. Figure 33 shows the spatial distribution of the ROC skill scores for the probability of 2m temperature anomalies being in the upper third of the climatological distribution valid for the June-August period at 1 month (top panel) and 2 months (bottom panel) lead time. The skill scores are computed by using the climate as the reference forecast. For the verification data, a combination of the ERA40 reanalysis and the ECMWF operational analysis for the most recent years, has been used. Orange to red areas indicate regions where the seasonal forecast has higher

skill than a forecast based on climatological values. From 1 to 2 months lead time the skill decreases only marginally.

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-Oscillation using a coupled GCM. *Monthly Weather Review*, **135**, 2700-2715.

List of Figures

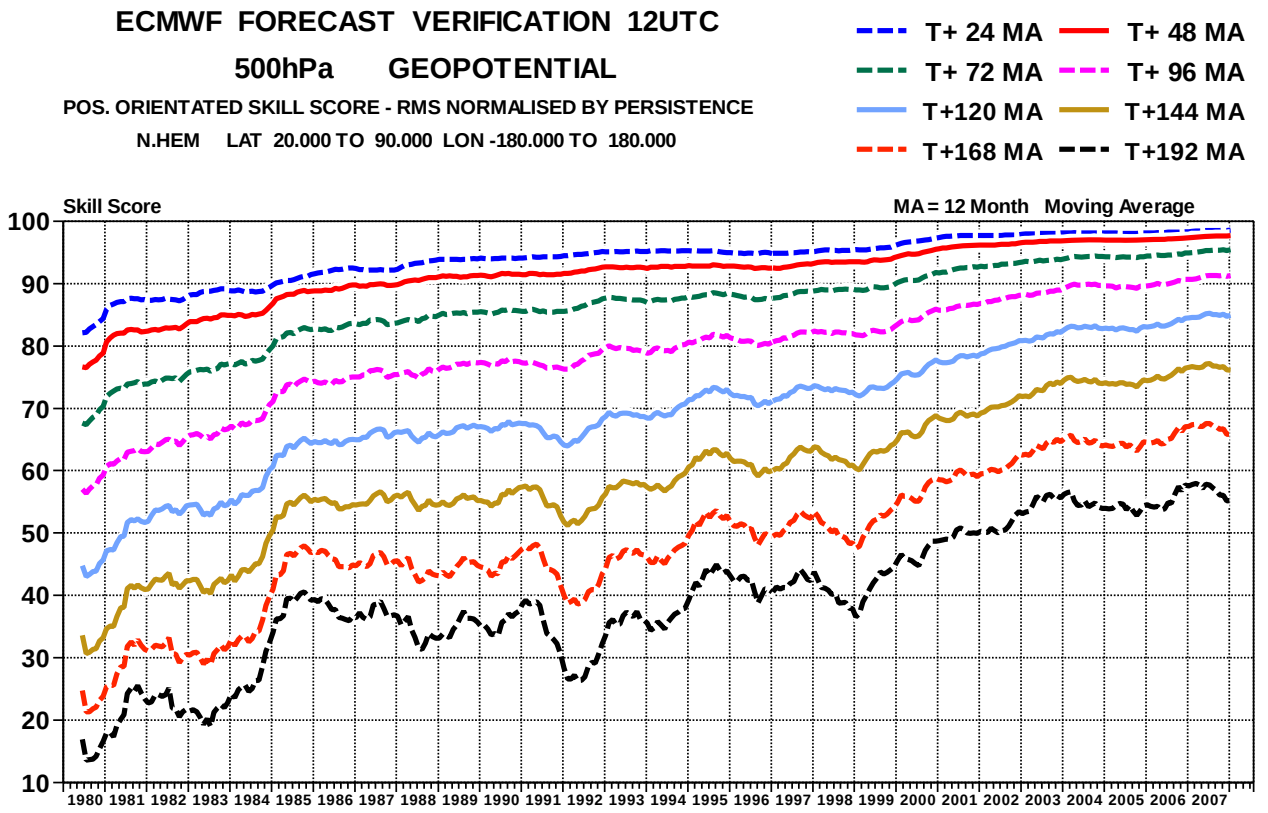


Figure 1: 500hPa height skill score for northern hemisphere (top) and Europe (bottom), 12-month moving averages, forecast ranges from 24 to 192 hours.

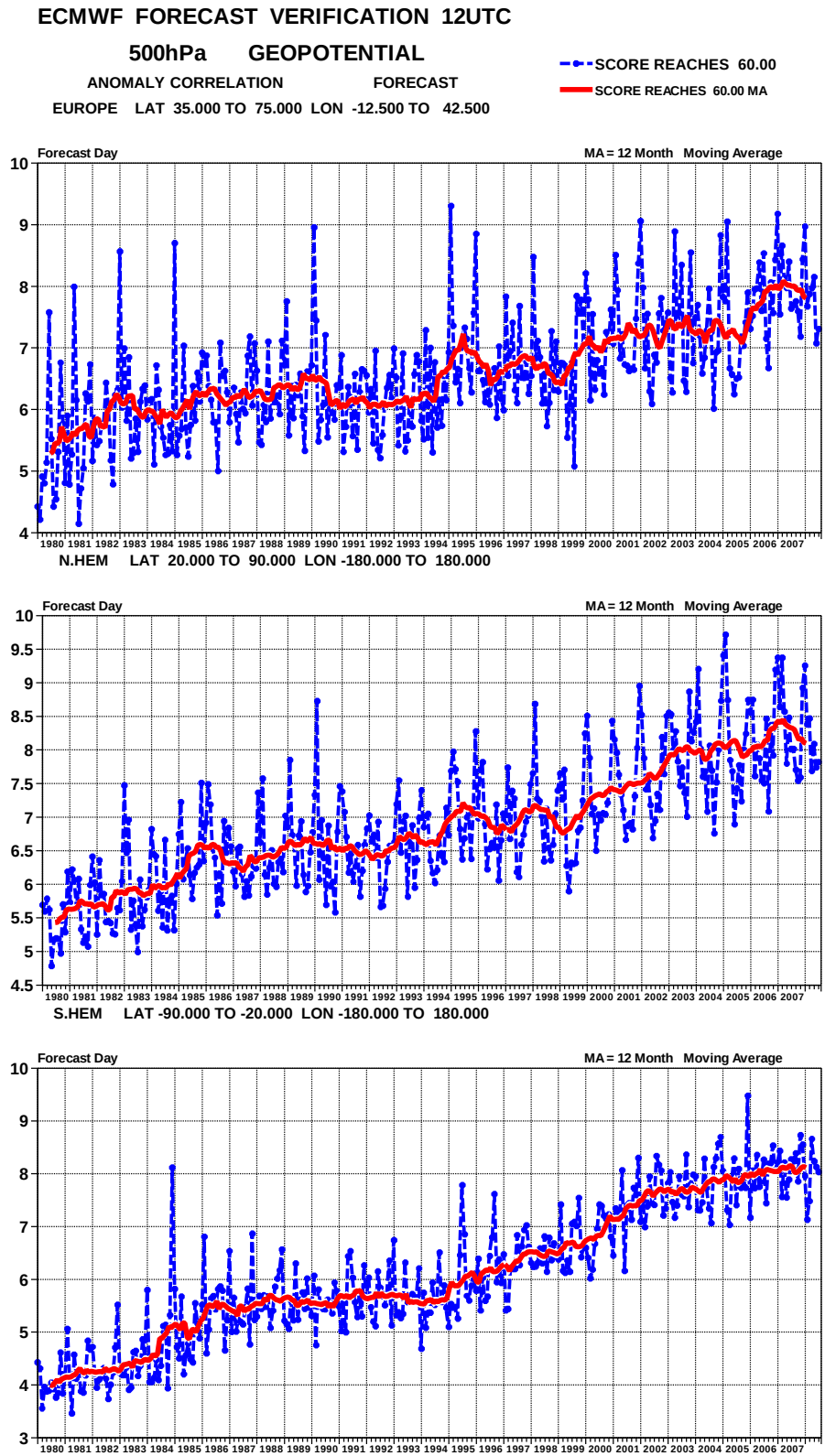


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the blue curves is the forecast range at which the monthly average of the forecast anomaly correlation with the verifying analysis falls below 60% for Europe, northern and southern extratropics (the red curve is the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

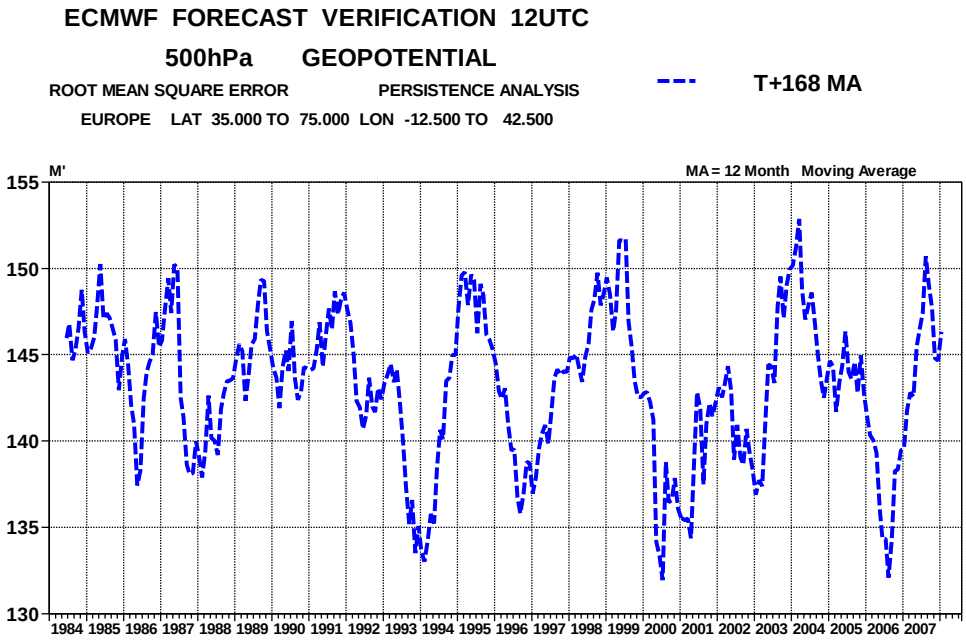


Figure 3: Root Mean Square Error of forecast made by persisting the analysis over 168h and verifying it as a forecast for 500 hPa geopotential height over Europe. 12-month moving average.

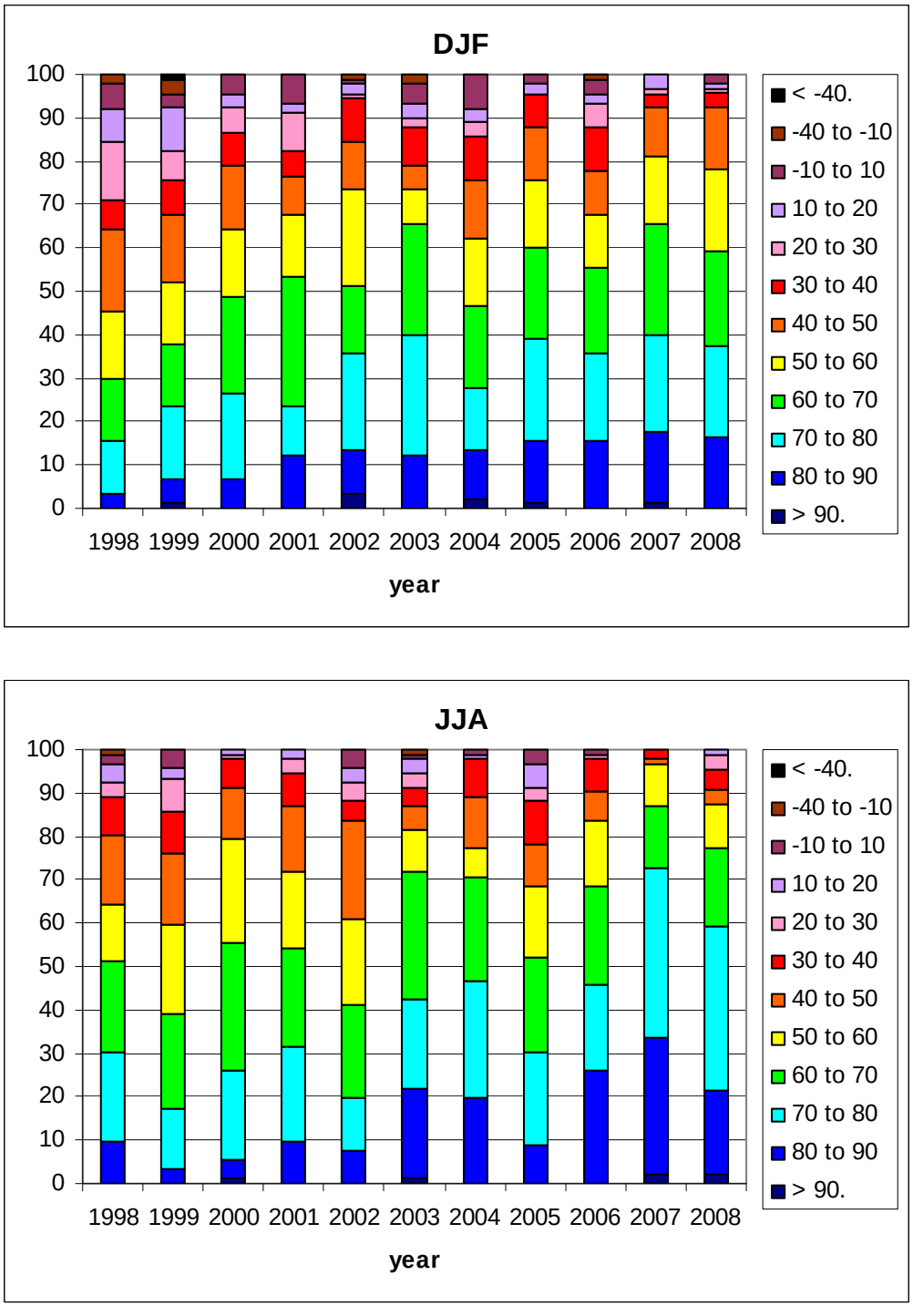


Figure 4: Distribution of Anomaly Correlation of the Day 7 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997-1998.

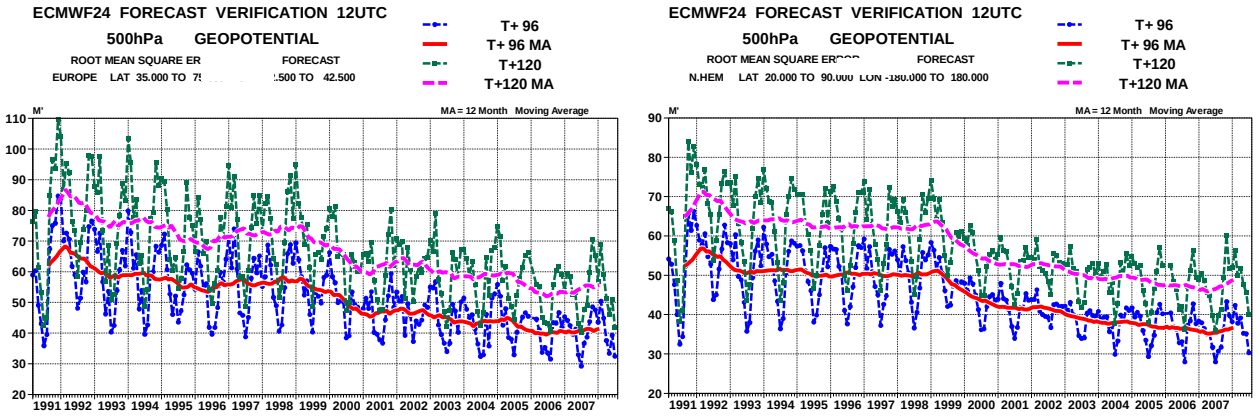


Figure 5: Consistency of the 500hPa height forecasts over Europe (left panel) and northern extratropics (right panel). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24h apart, for 96-120h (blue) and 120-144h (green). 12-month moving average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

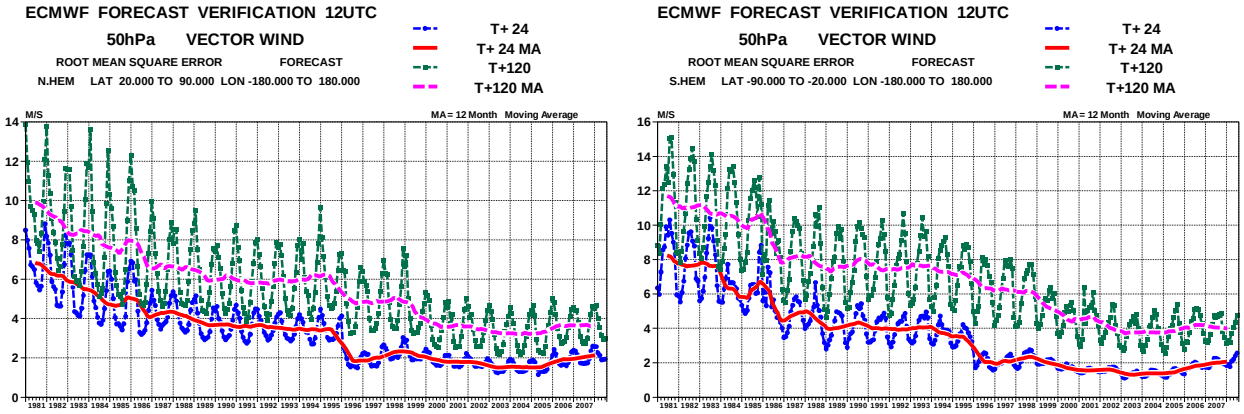


Figure 6: Model scores in the extratropical northern (left) and southern (right) hemisphere stratosphere (RMS vector wind error at 50hPa for 1-day and 5-day forecasts).

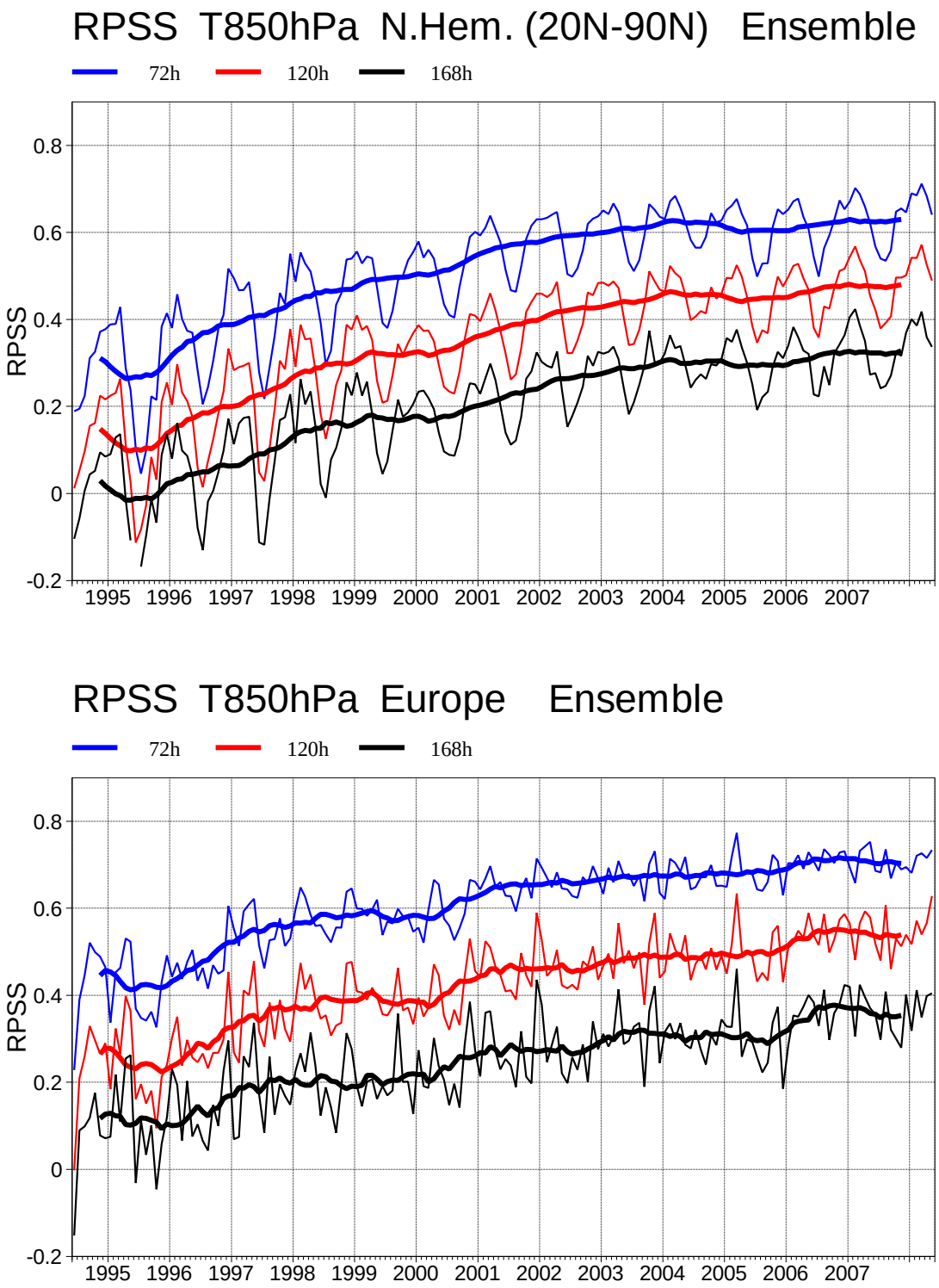


Figure 7: Monthly score and 12-month running mean (bold) of Ranked Probability Skill Score for EPS forecasts of 850 hPa temperature at day 3 (blue), 5 (red) and 7 (black) for the northern hemisphere extratropics (top) and Europe (bottom).

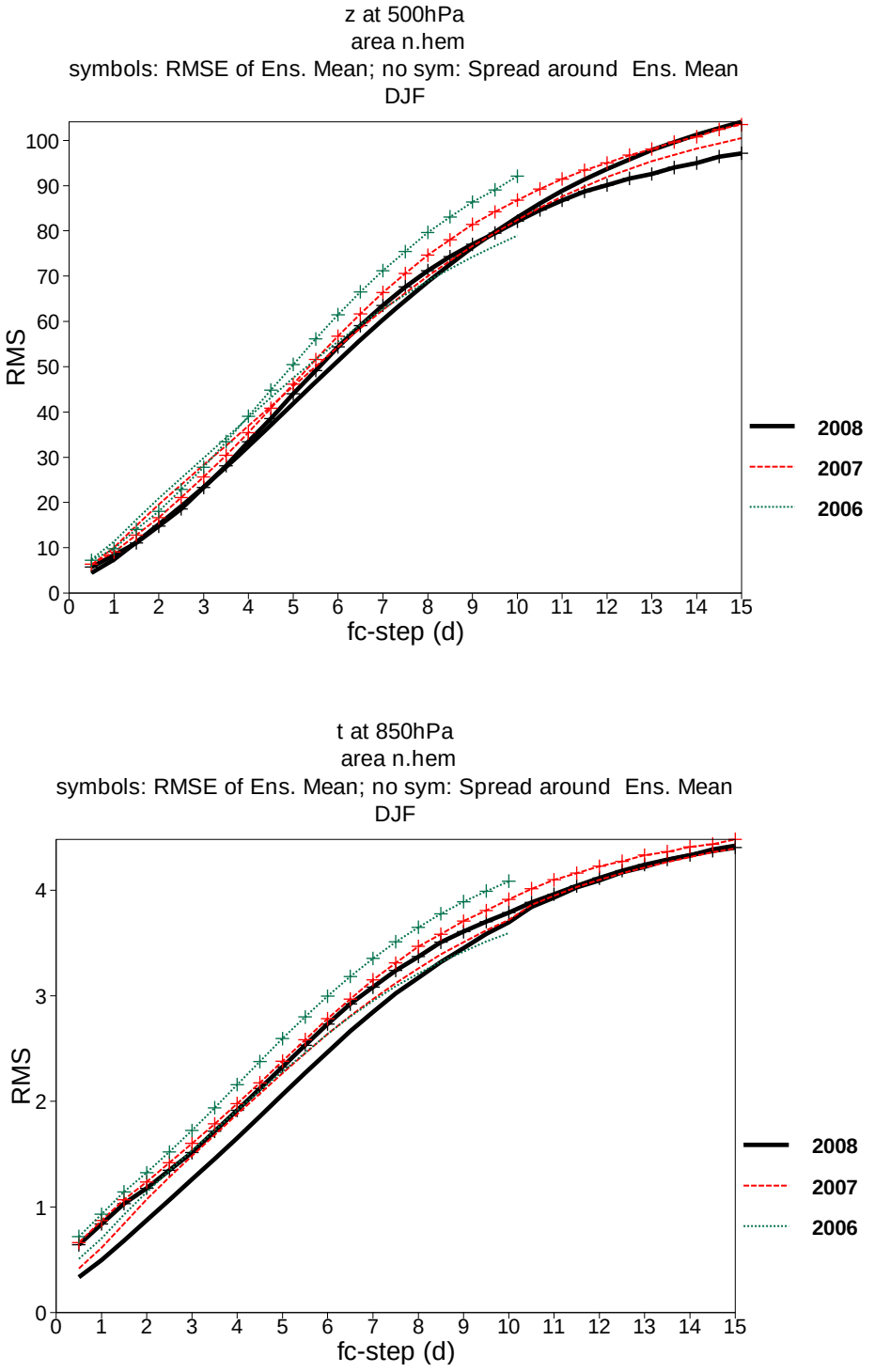


Figure 8: Ensemble spread (standard deviation) and root mean square error of ensemble-mean (lines with crosses) for 500 hPa height (top) and 850 hPa temperature (bottom) for winter 2007-08 (black), 2006-07 (red) and 2005-06 (green) over the extra-tropical northern hemisphere.

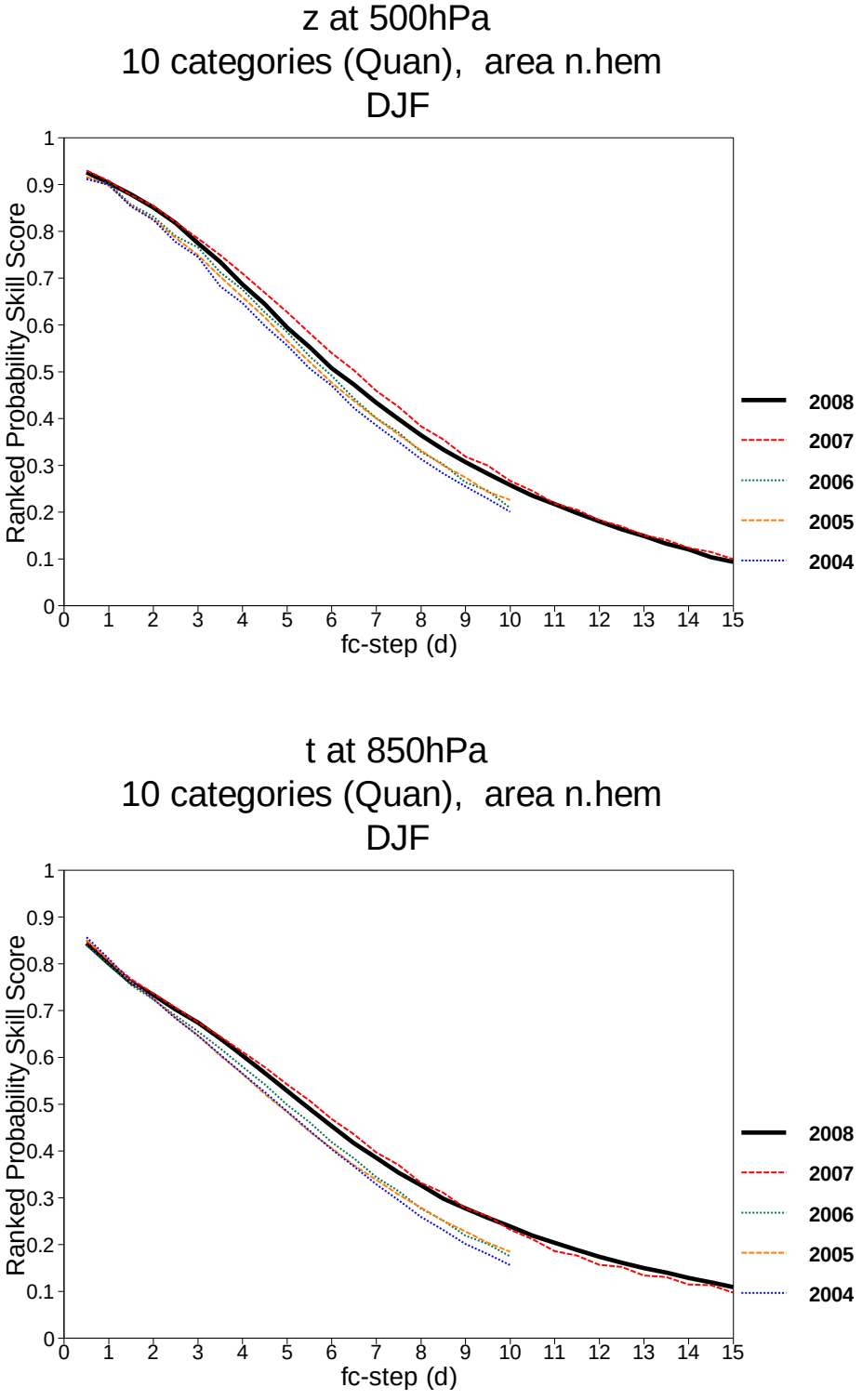
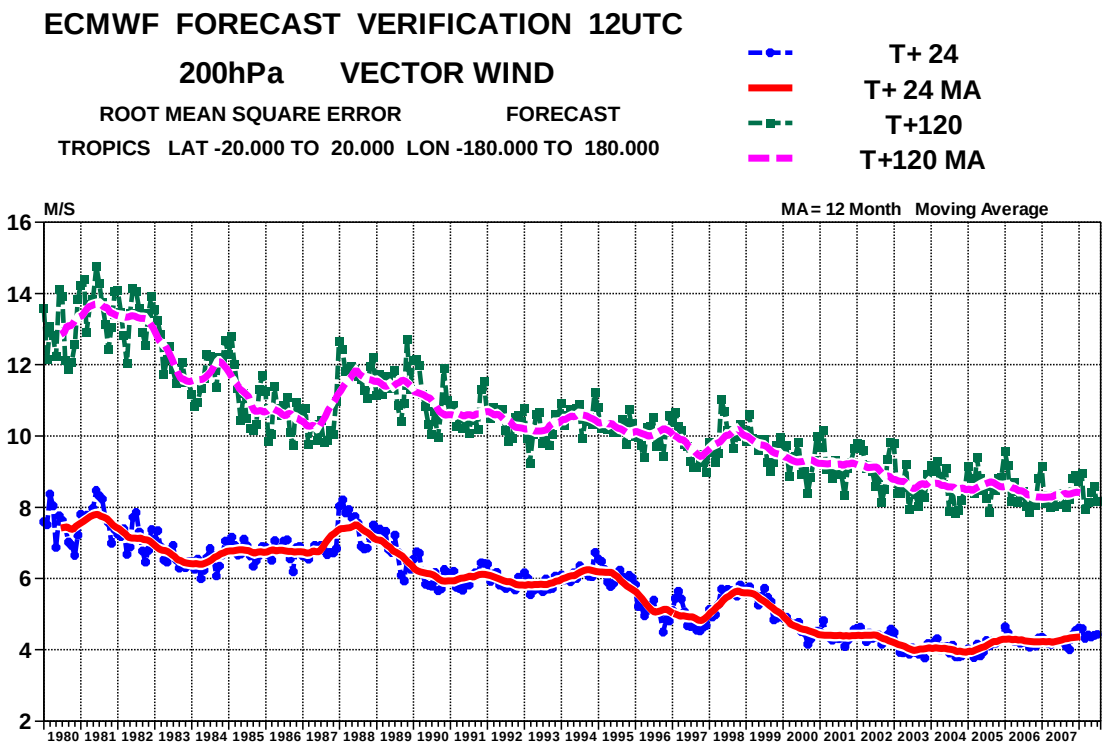


Figure 9: Ranked probability skill score for 500 hPa height (top) and 850 hPa temperature (bottom) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. Skill from the VarEPS days 1-15 forecasts is shown for winter 2007-08 (black) and 2006-07 (red); the EPS only ran to 10 days in previous years.



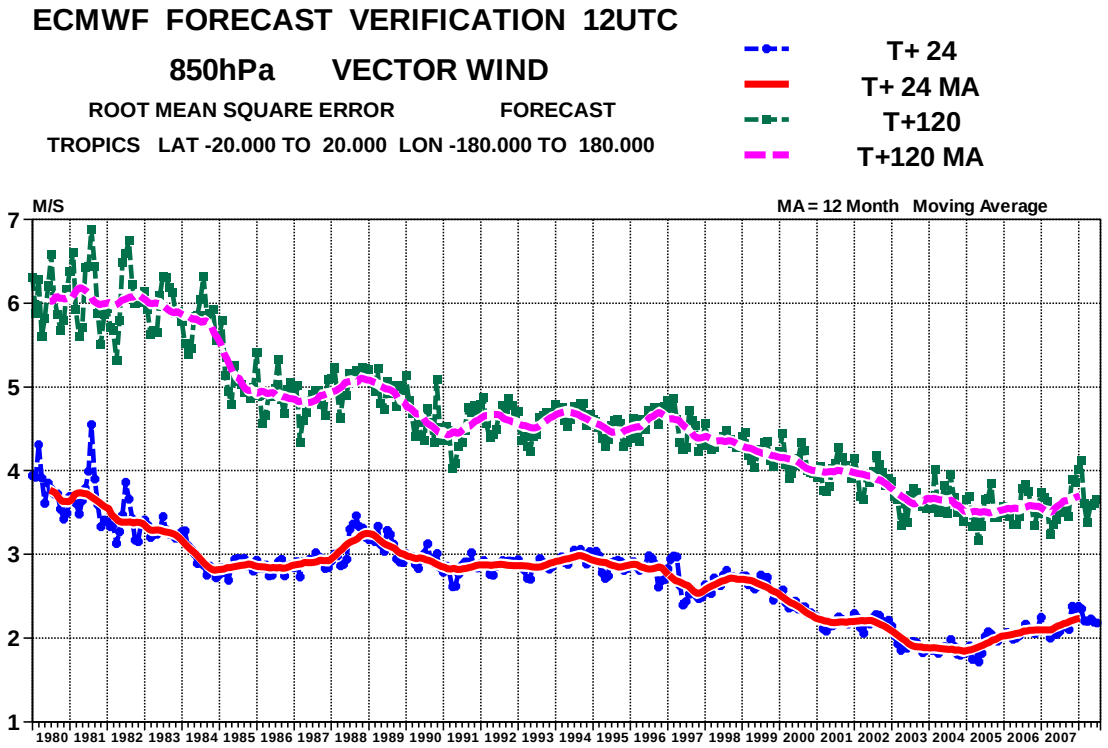


Figure 10: Model scores in the tropics (root mean square vector wind errors at 200hPa and 850hPa for 1-day and 5-day forecasts). Monthly mean and 12-month running mean.

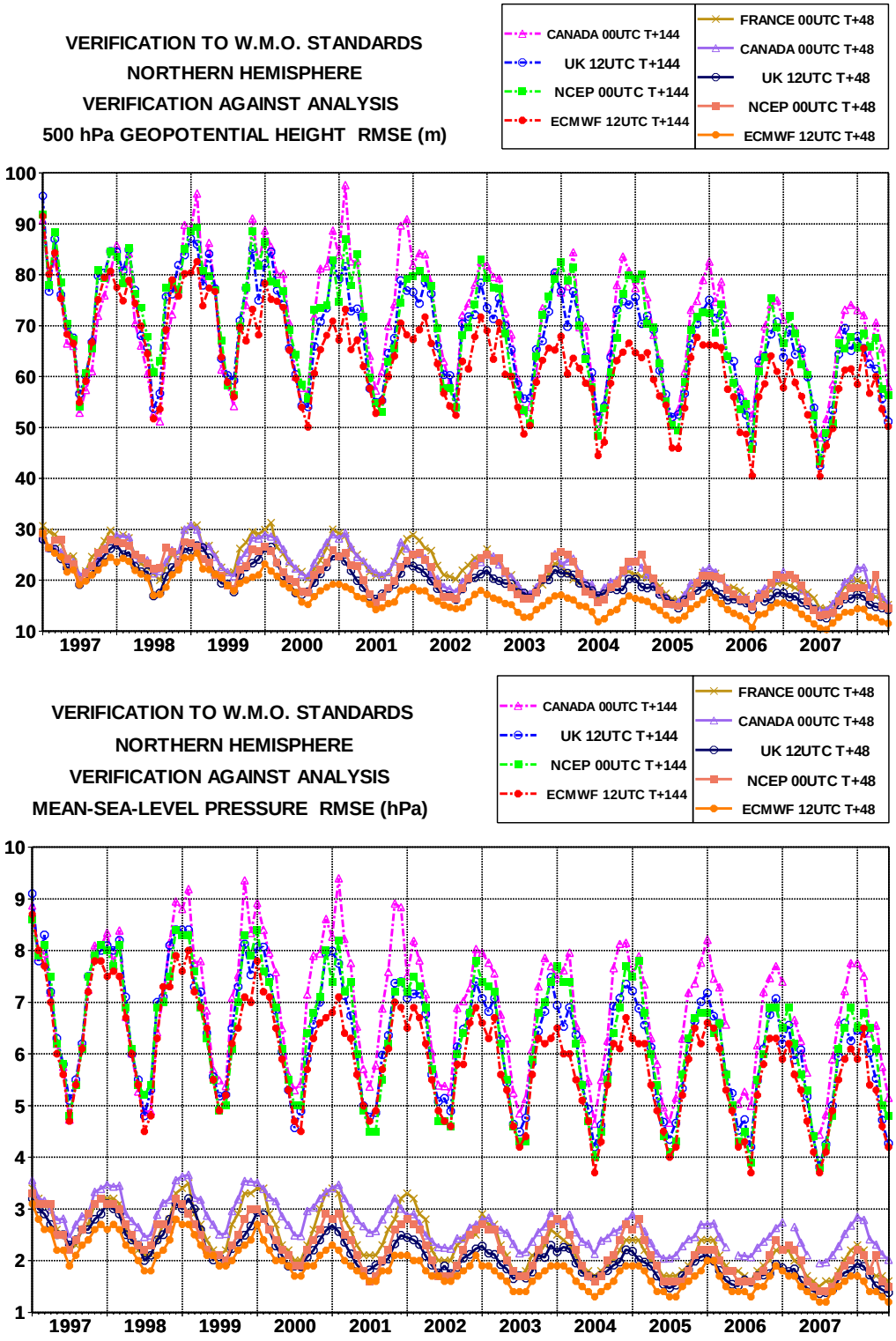


Figure 11: WMO/CBS exchanged scores (RMS error over northern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).

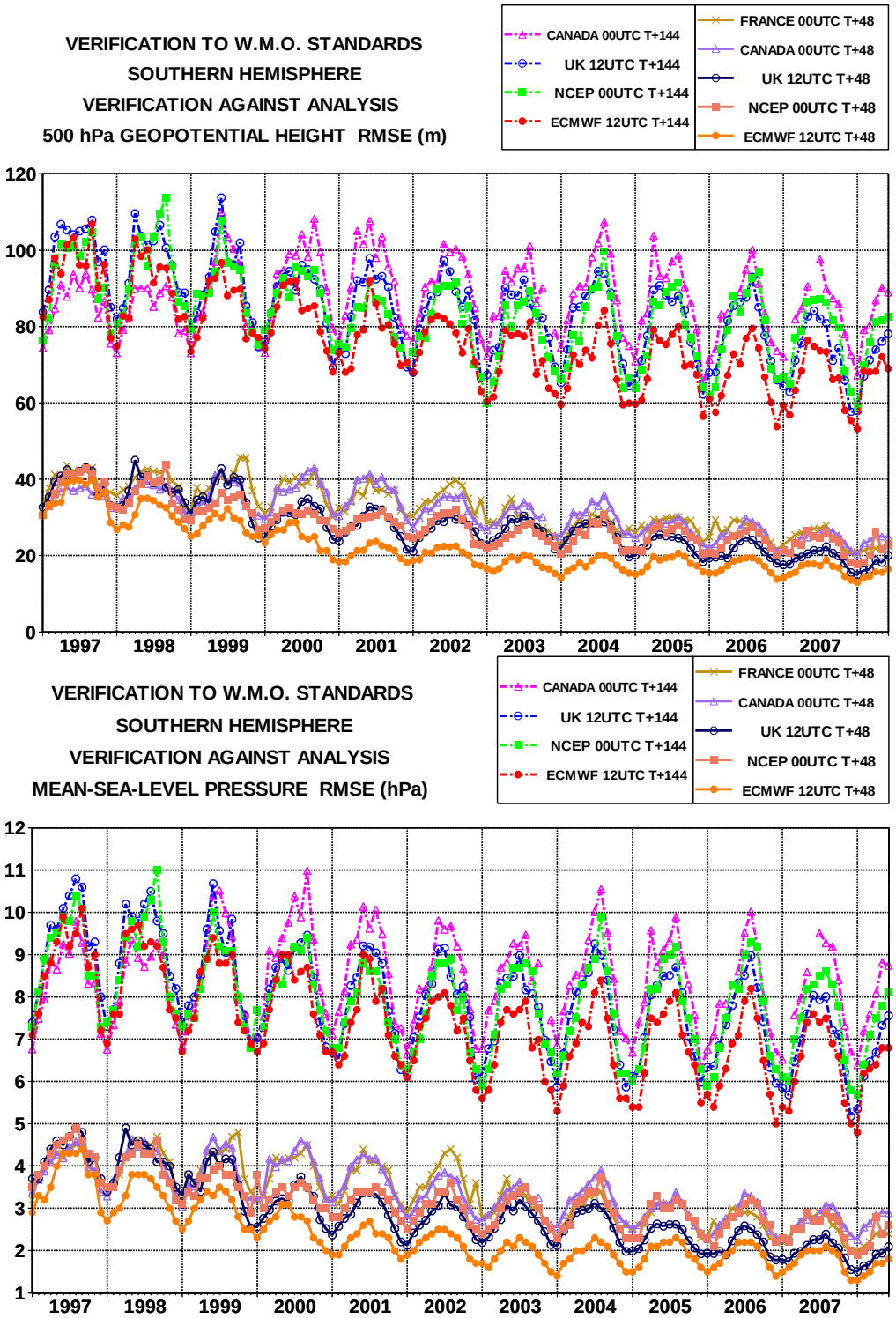


Figure 12: WMO/CBS exchanged scores (RMS error over southern extratropics, 500hPa geopotential height and MSLP for 2-day and 6-day forecasts).

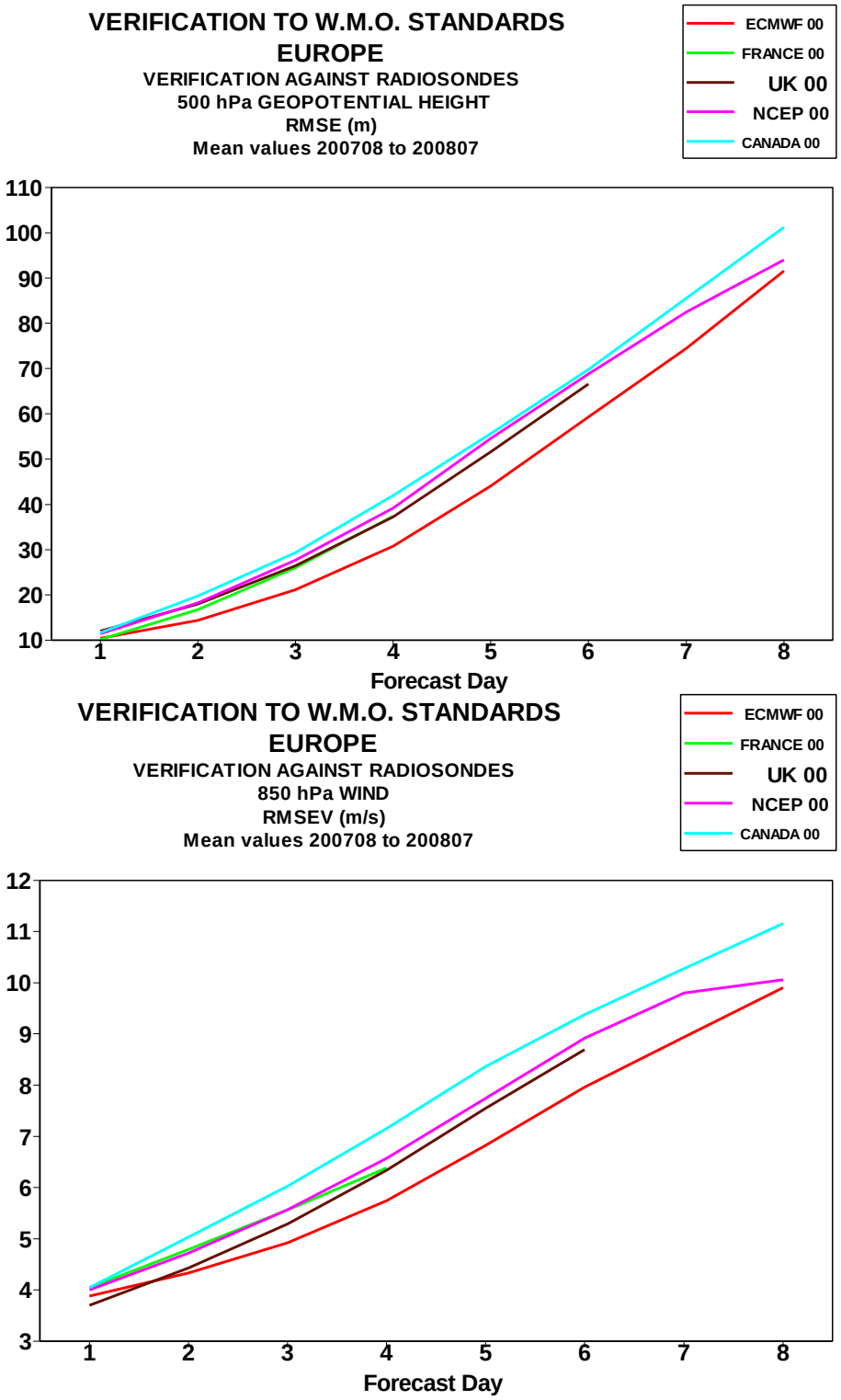


Figure 13: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean).

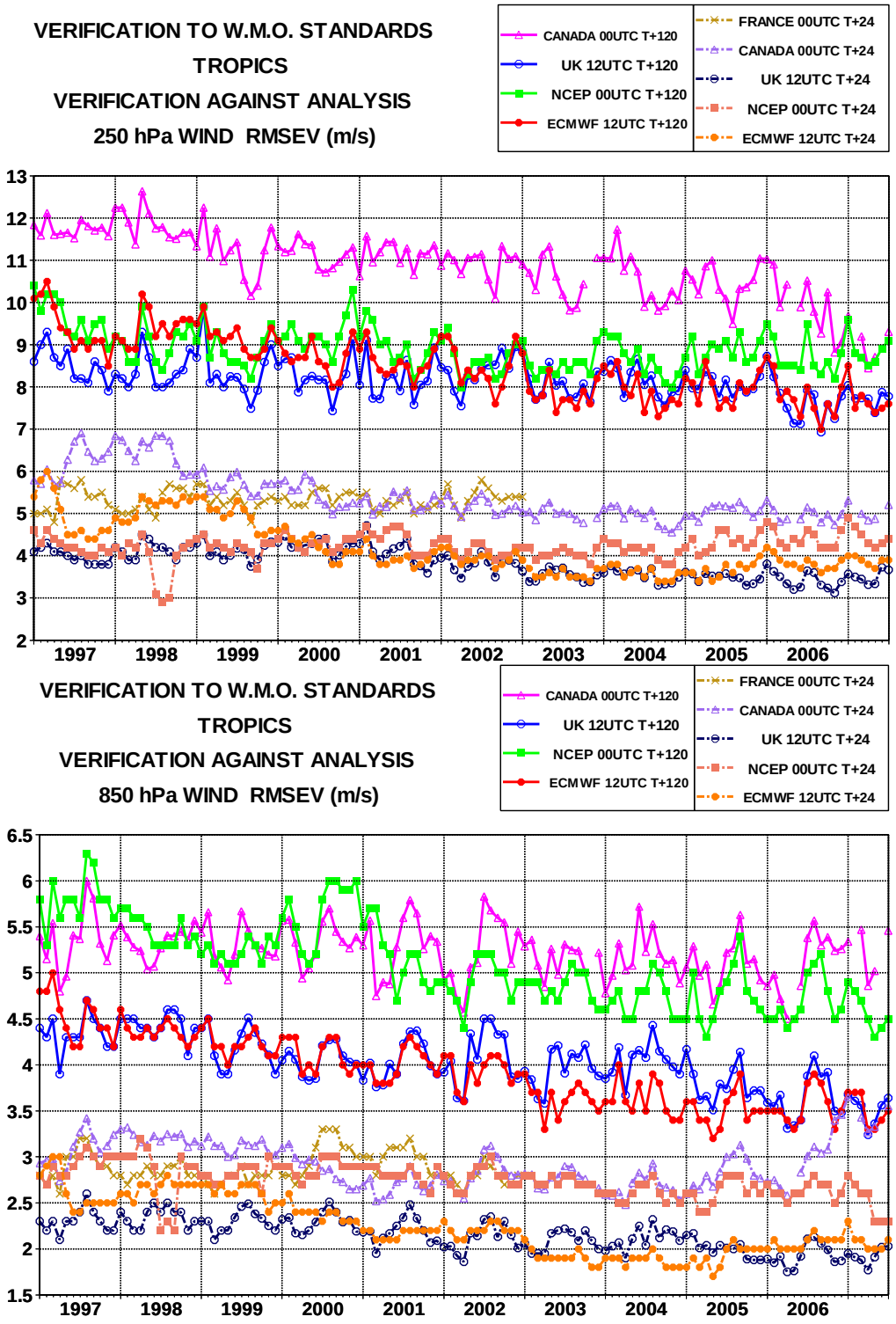


Figure 14: WMO/CBS exchanged scores (RMS vector error over the tropics, 250hPa and 850hPa wind forecast for day 1 and day 5).

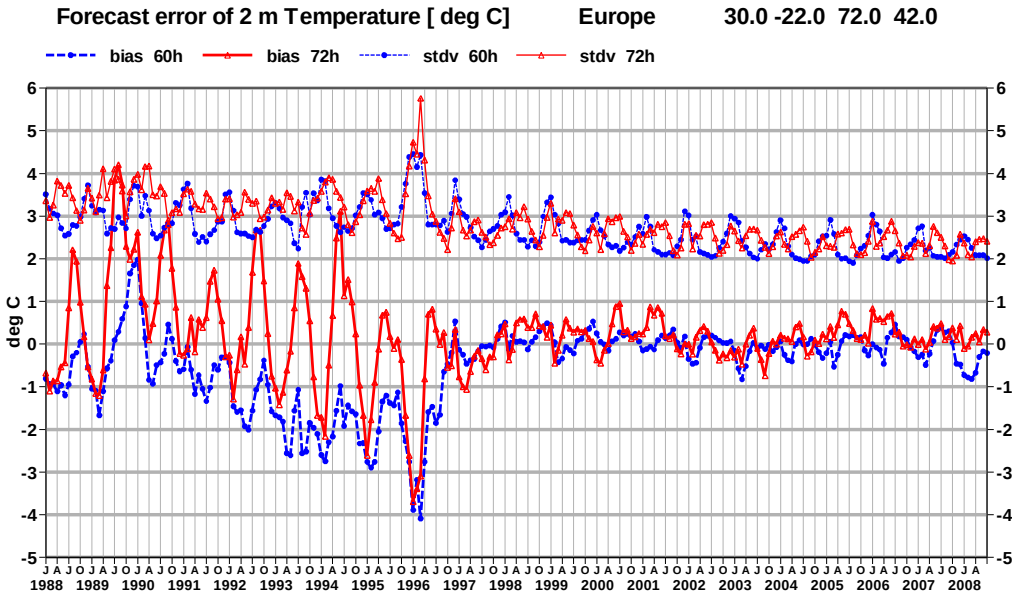


Figure 15: Verification of 2 metre temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

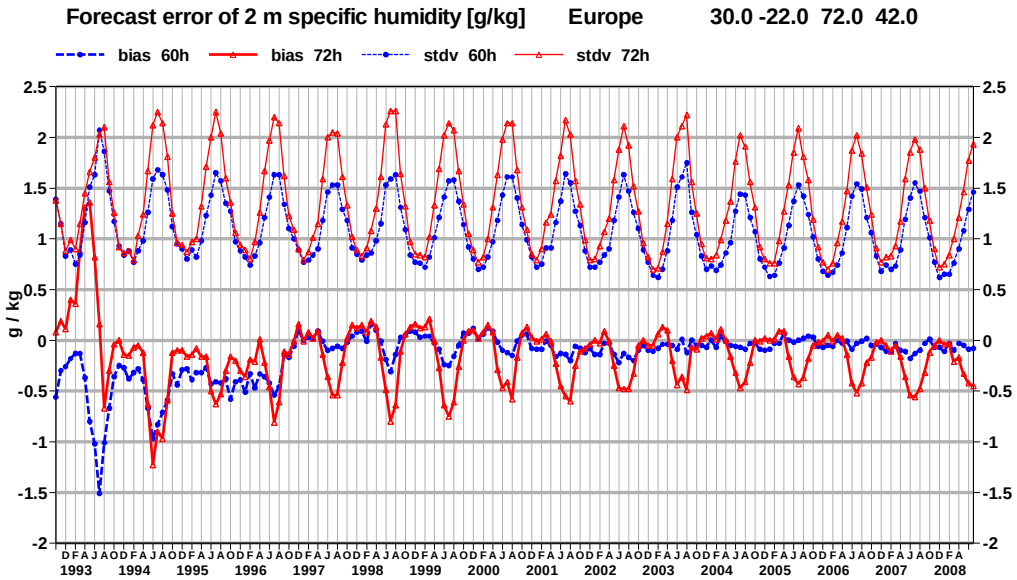


Figure 16: Verification of 2 metre specific humidity forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

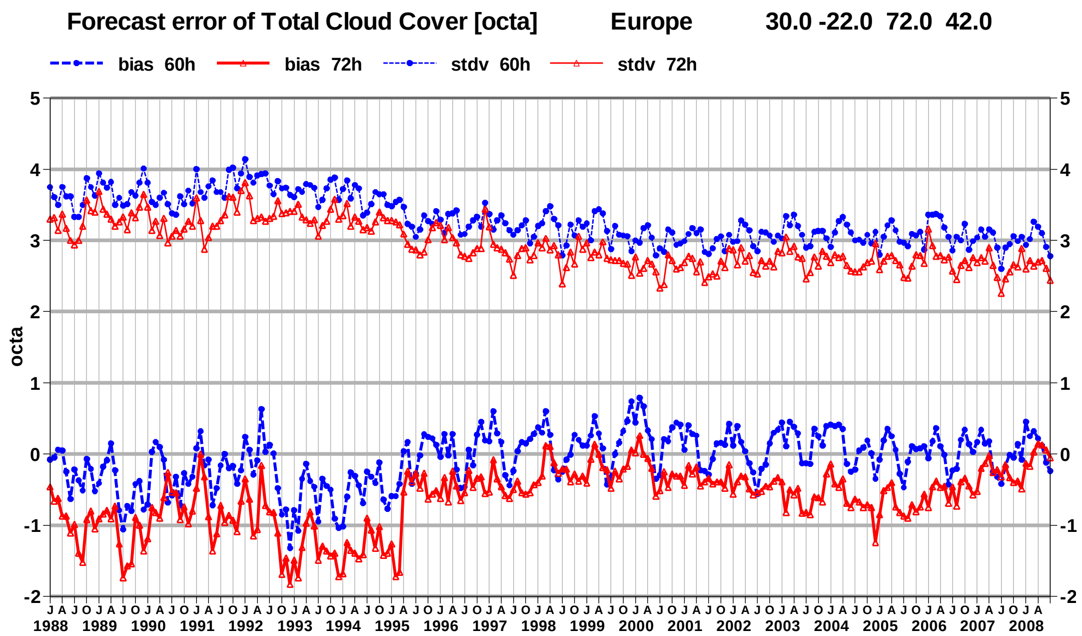


Figure 17: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

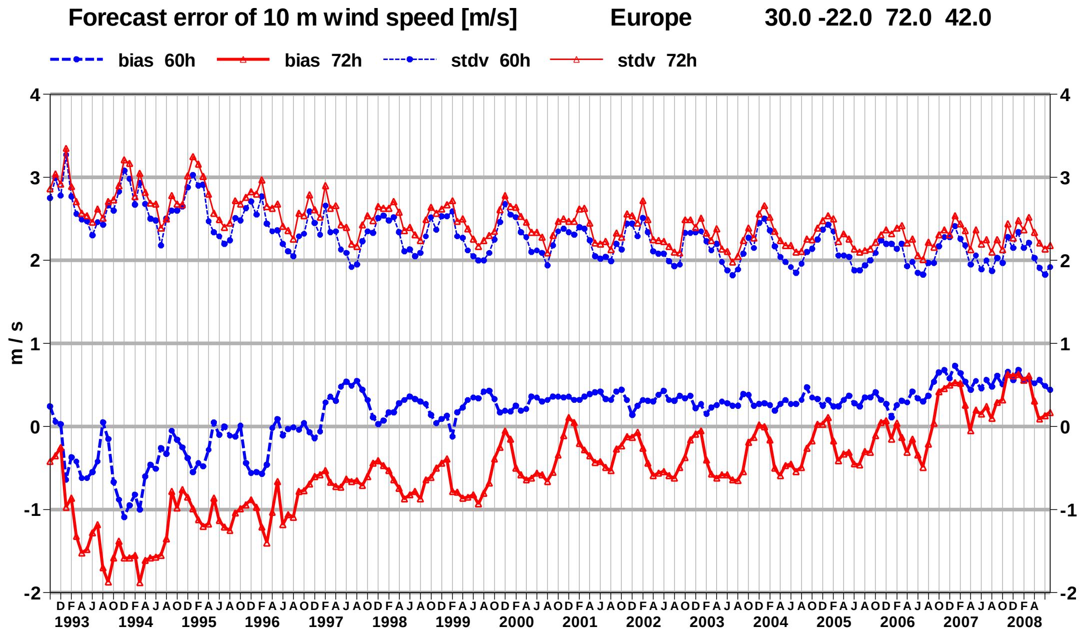
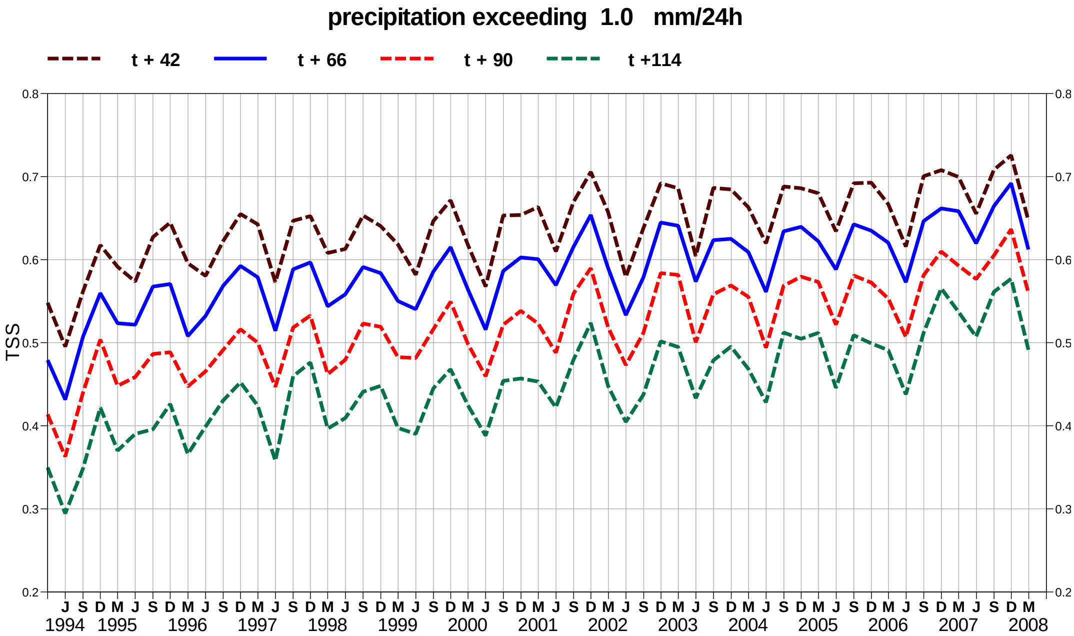


Figure 18: Verification of 10-metre wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.



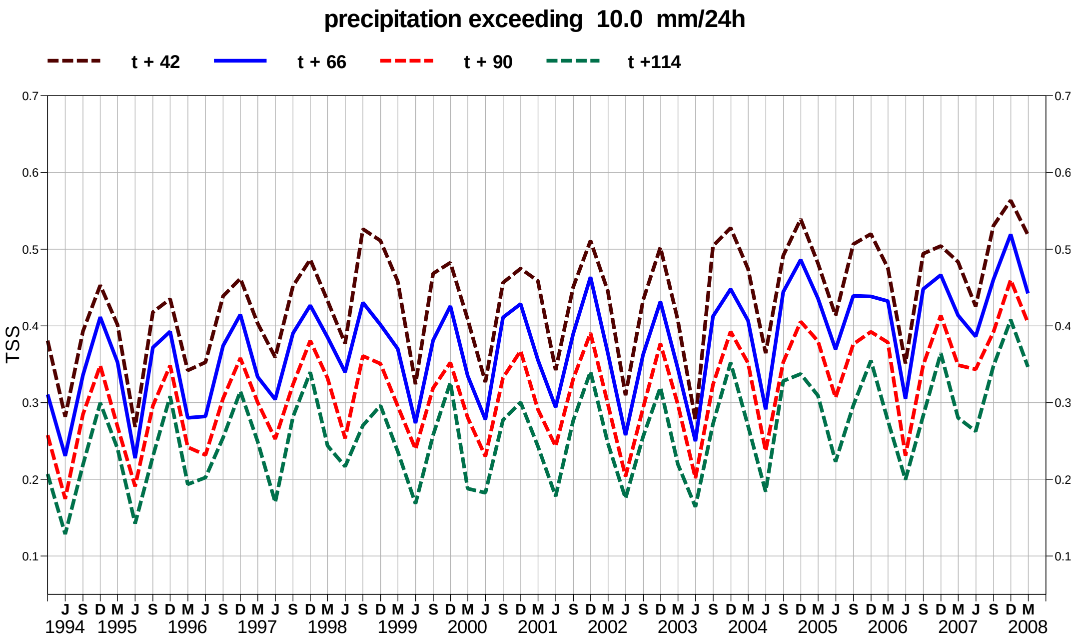


Figure 19: TSS time series for precipitation forecasts exceeding 1mm/day (top) and 10mm/day (bottom) verified against SYNOP data on the GTS for Europe. Curves are shown for the 24-hour accumulations up to 42, 66, 90, and 114 hours (from the forecasts starting at 12 UTC). 3-month mean scores (last point is March-May 2008).

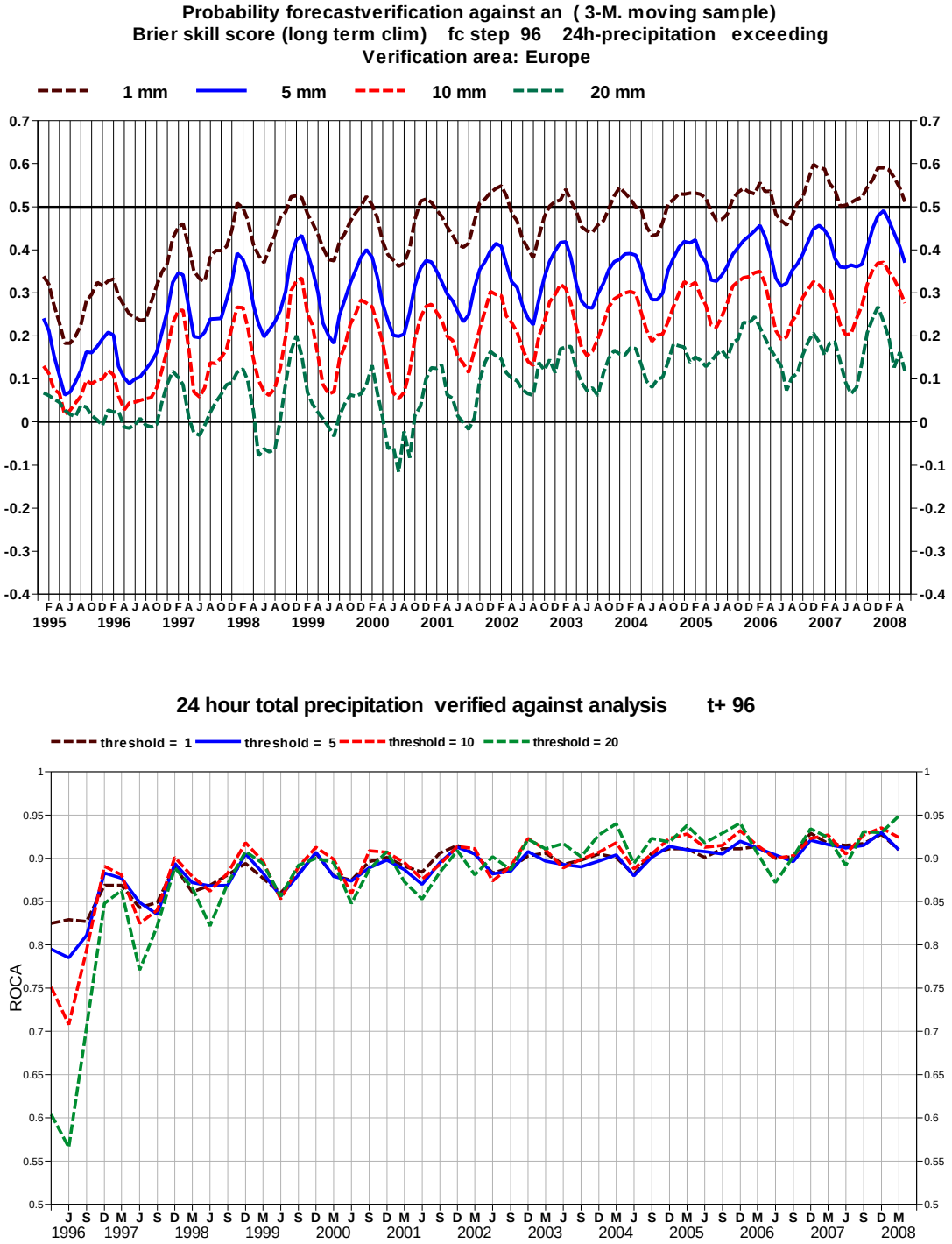


Figure 20: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA) for EPS probability forecasts of precipitation over Europe exceeding thresholds of 1, 5, 10 and 20 mm/day at day 4. The skill score is calculated for three-month running periods.

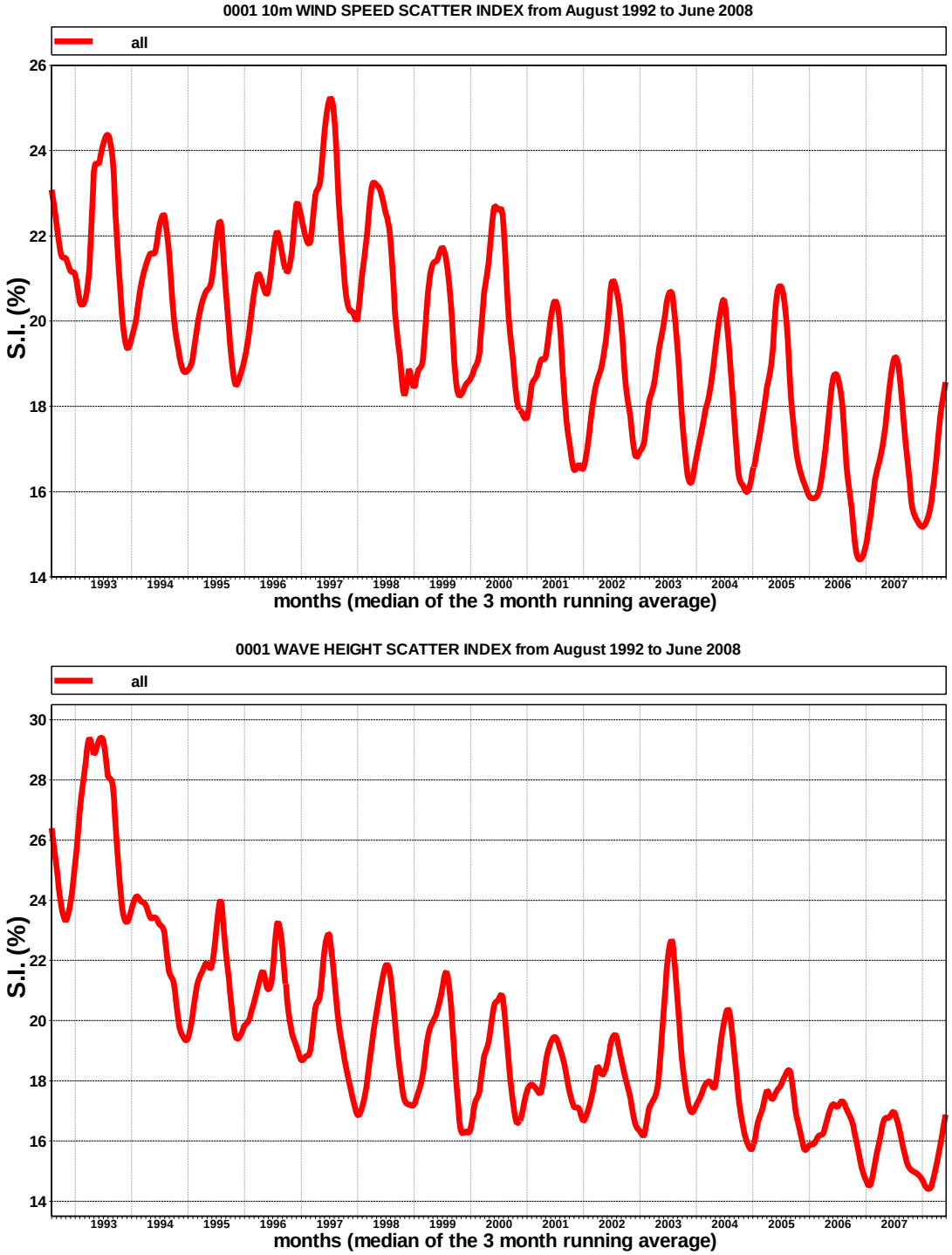


Figure 21: Time series of verification of the ECMWF 10 metre wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

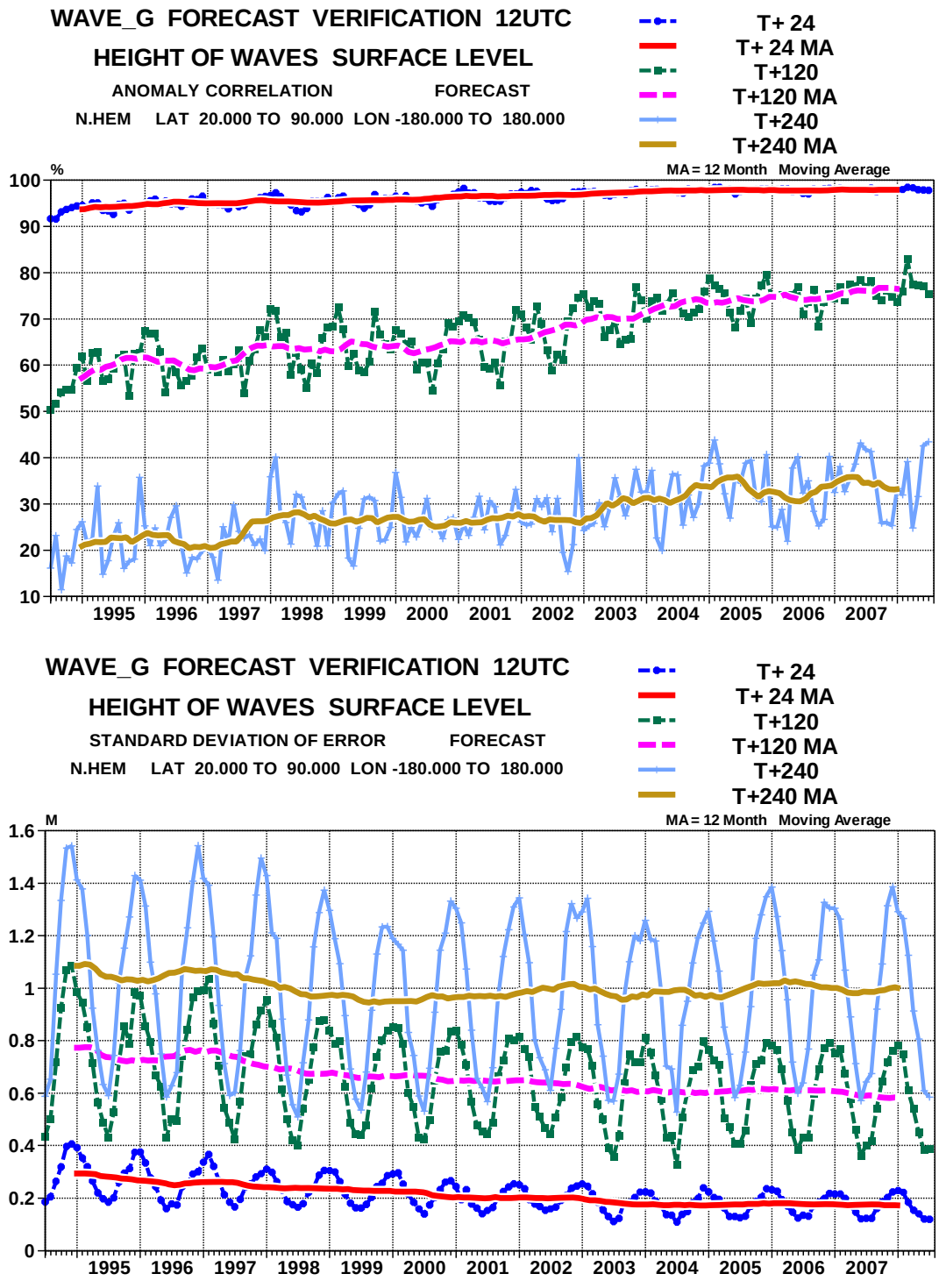
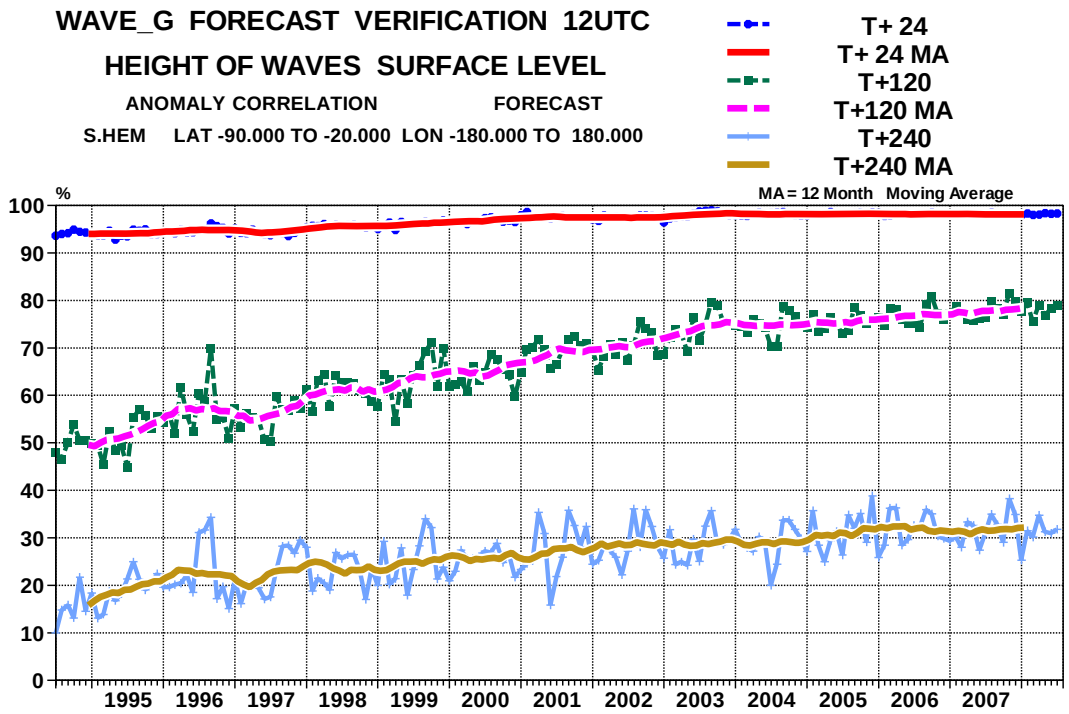


Figure 22: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (northern extratropics).



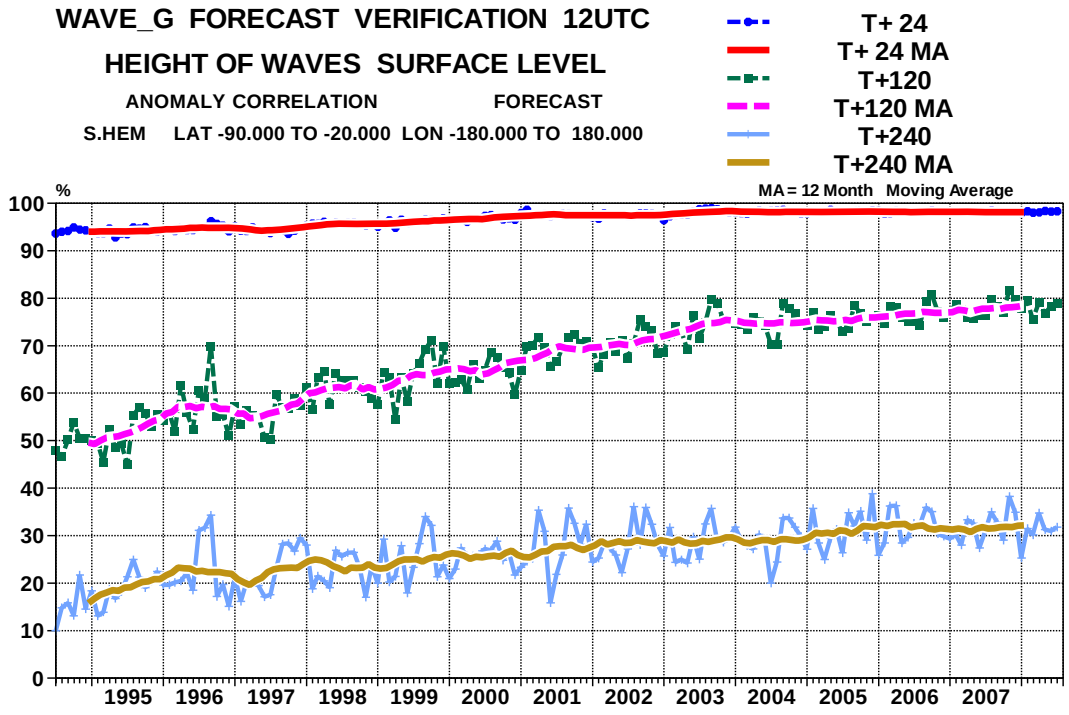


Figure 23: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (southern extratropics).

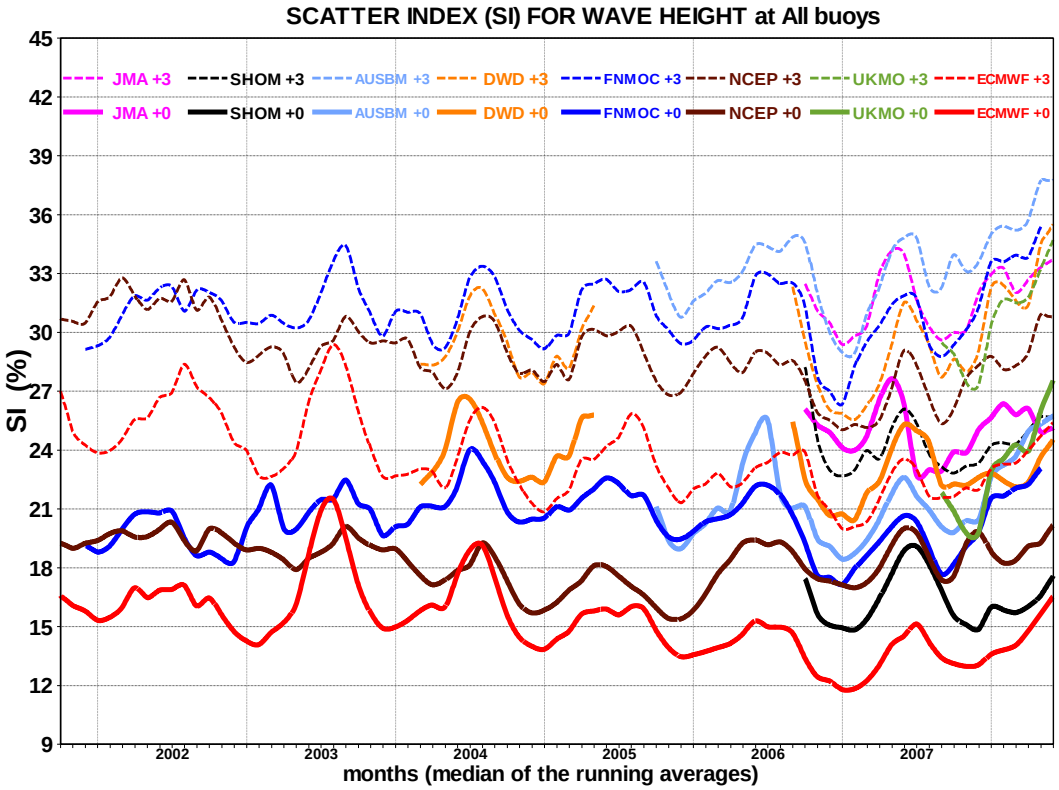


Figure 24: Verification of different model wave height forecasts using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; a three-month running mean is used.

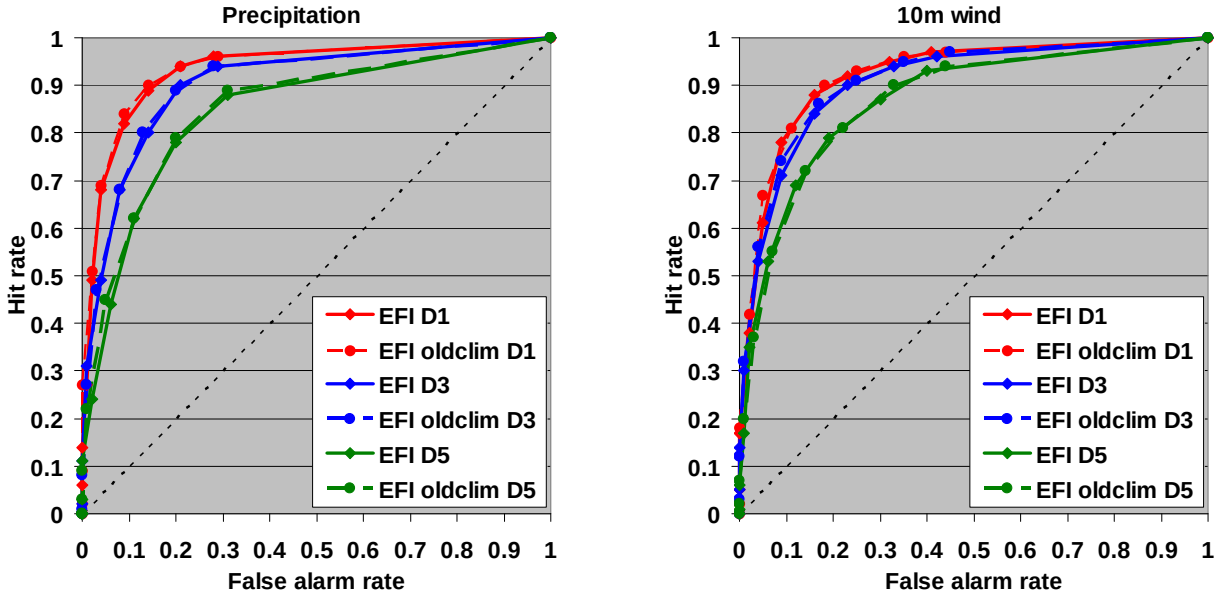


Figure 25: Verification of Extreme Forecast Index (EFI) for precipitation (left) and 10m wind (right) over Europe for October 2007 - March 2008. Extreme event is taken as an observation exceeding 95th percentile of station climate. Hit rates and false alarm rates are calculated for EFI exceeding different thresholds. Results are shown for forecast days 1 (red), 3 (blue) and 5 (green) using both old (dashed lines) and new (solid lines) EFI climates (the new climatology was introduced into operations in March 2008).

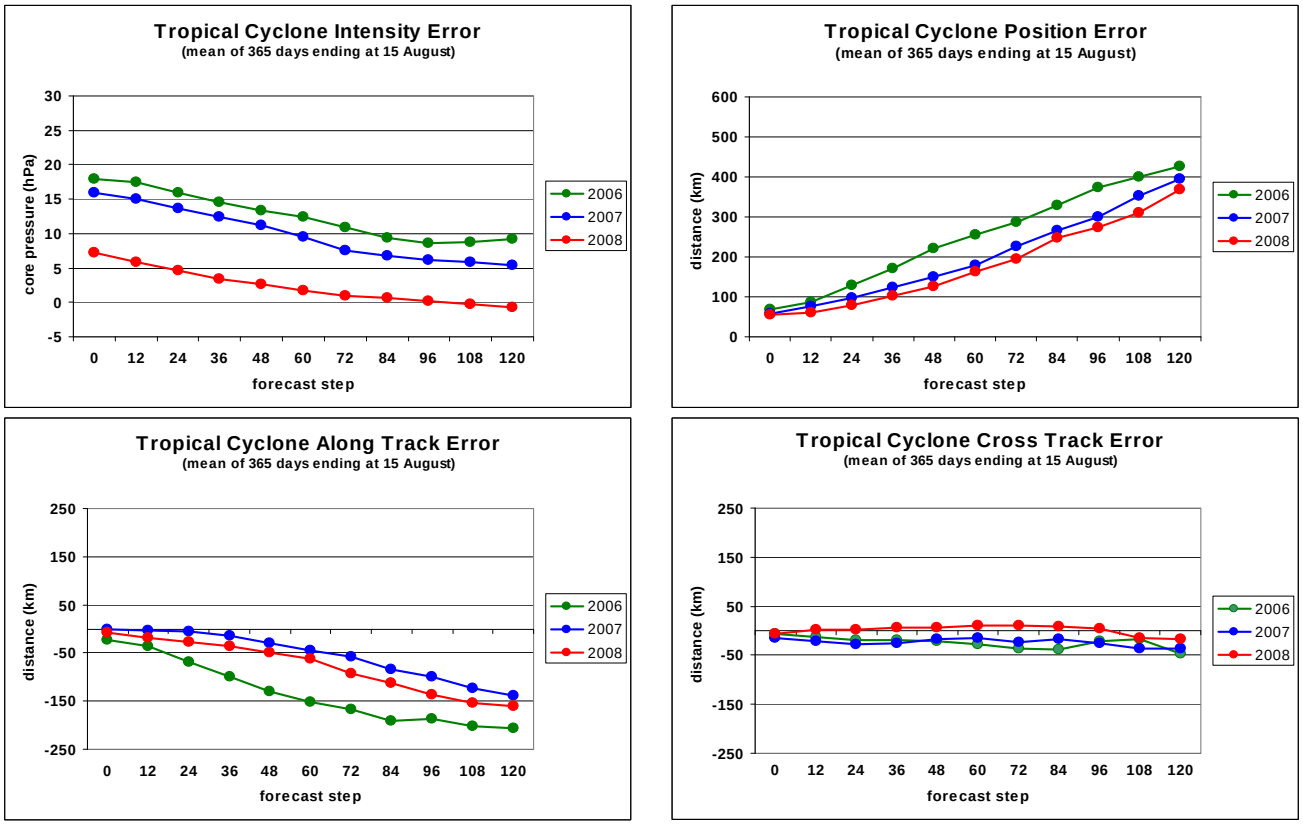


Figure 26: Verification of tropical cyclone predictions from the operational deterministic forecast for three 12-month periods: August 2005 - August 2006 (green), August 2006 - August 2007 (blue) and August 2007 - August 2008 (red). The upper panel shows the mean error in core pressure (left) and position (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. Within each year, the sample size is the same at each forecast step (but the number of cyclones varies from year to year). Verifications made against tropical cyclone observations reported on the GTS.

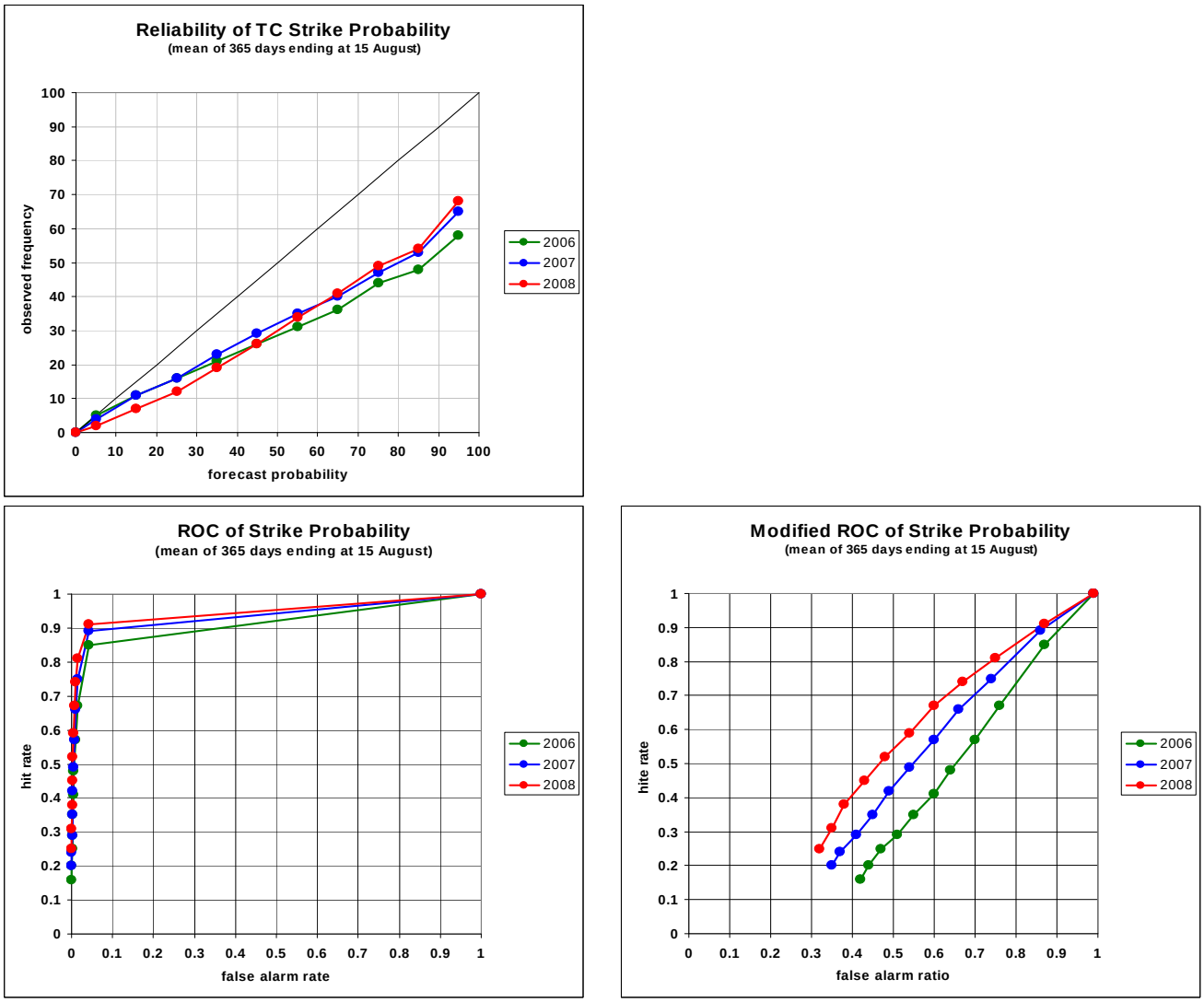


Figure 27: Probabilistic verification of EPS tropical cyclone forecasts for three 12-month periods: August 2005 - August 2006 (green), August 2006 - August 2007 (blue) and August 2007 - August 2008 (red). Upper panel shows reliability diagram (the closer to the diagonal the better). The lower panel shows (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.

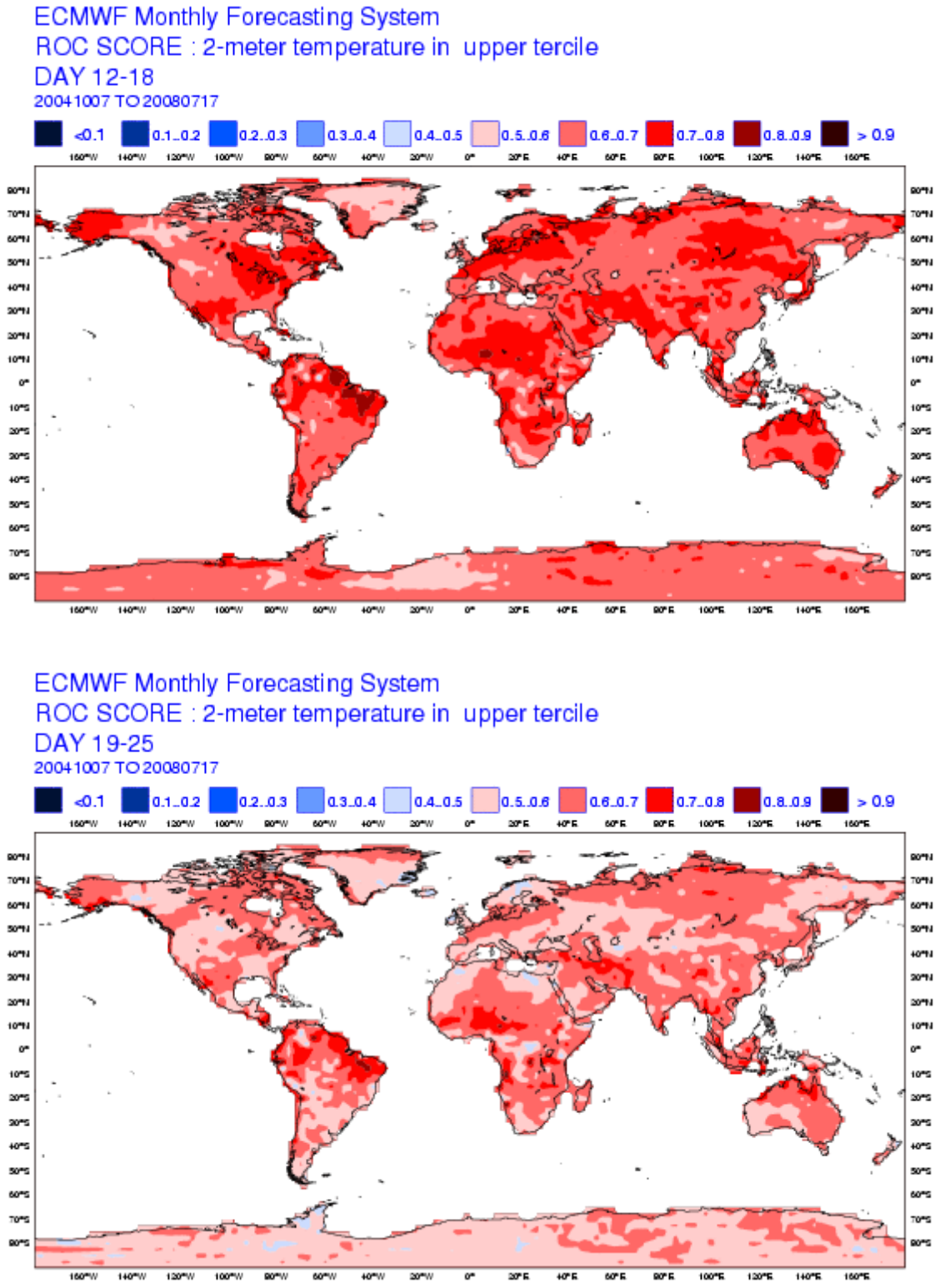


Figure 28: Spatial distribution of ROC area scores for the probability of 2m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 17 July 2008 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicates positive skill compared to climate.

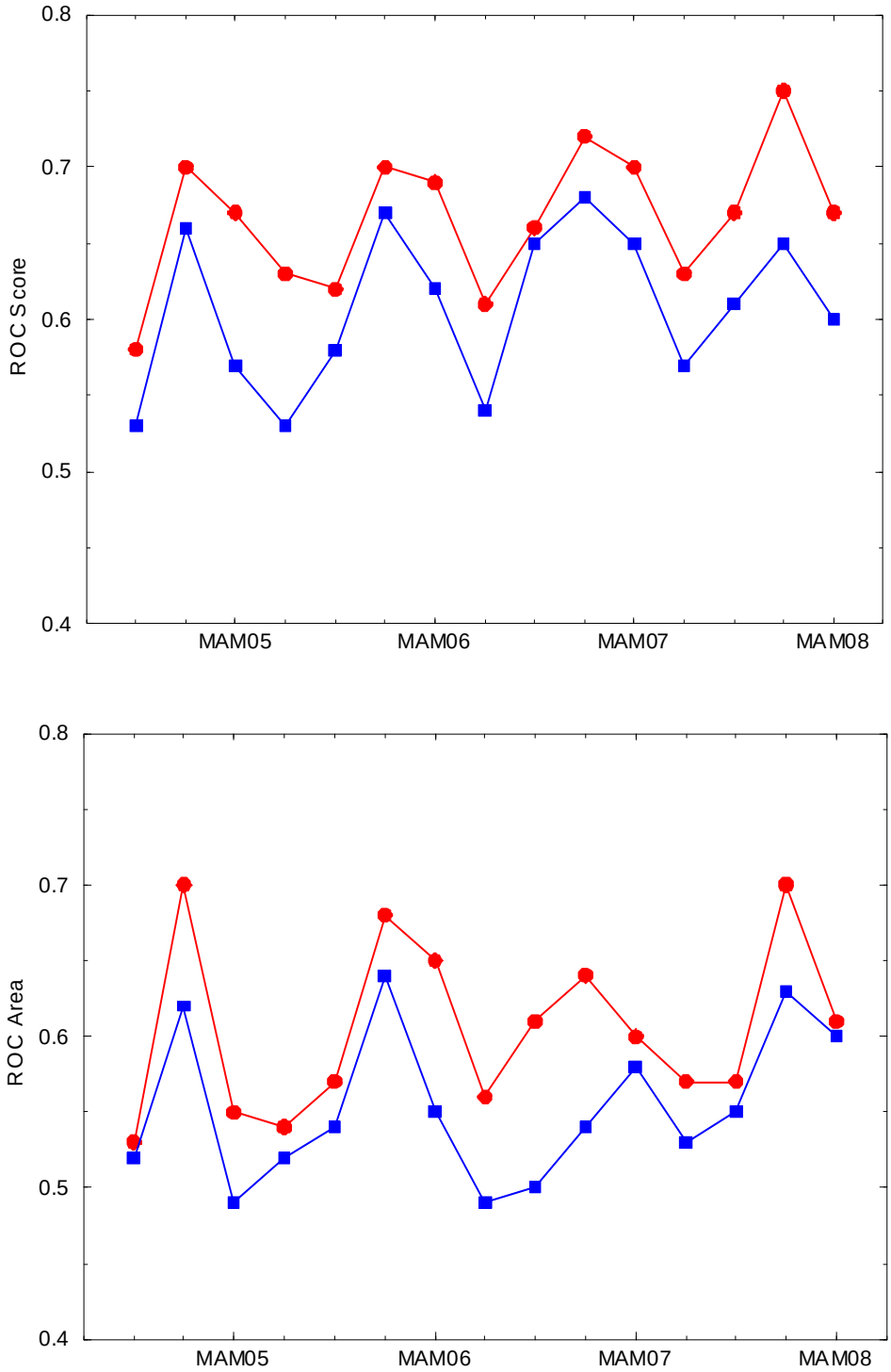


Figure 29: Area under ROC for the probability that 2-metre temperature is in the upper third of the climate distribution. Scores are calculated for each 3-month season since autumn (September-November) 2005 for all land points in the extra-tropical northern hemisphere. The red line shows the score of the operational monthly forecasting system for forecast days 12-18 (7-day mean) (top panel) and 19-32 (14-day mean) (bottom panel). As a comparison, the blue line shows the score using persistence of the preceding 7-day or 14-day period of the forecast.

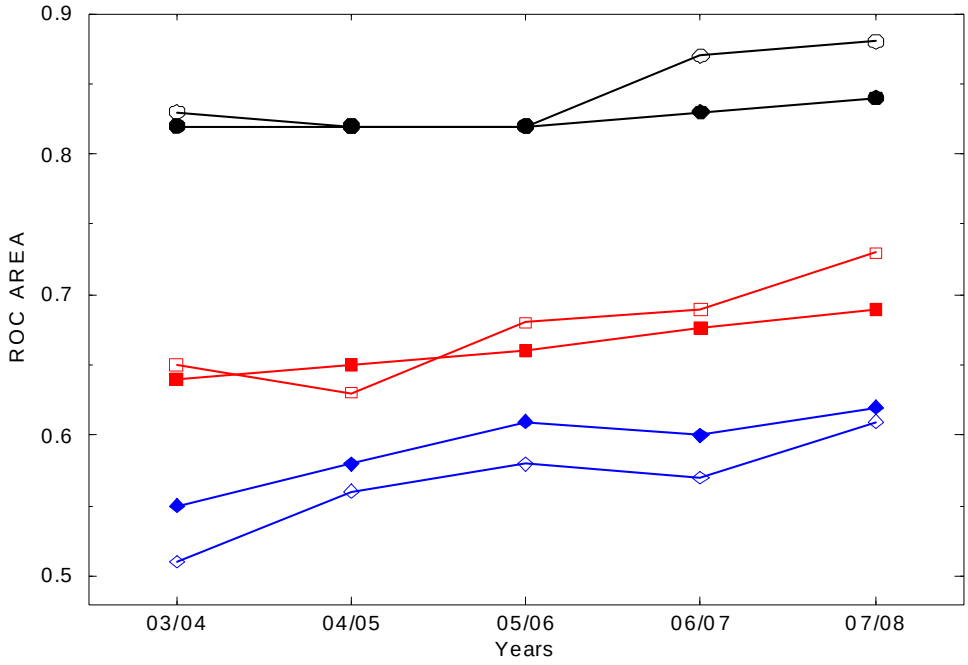


Figure 30: Area under ROC for the probability that 2-metre temperature is in the upper third of the climate distribution. Scores are calculated for each year for all land points over: the extra-tropical northern hemisphere (solid lines) and Europe (dashed lines). Black lines show the scores for the forecast range 5-11 days, red line shows for forecast days 12-18 and 19-32.

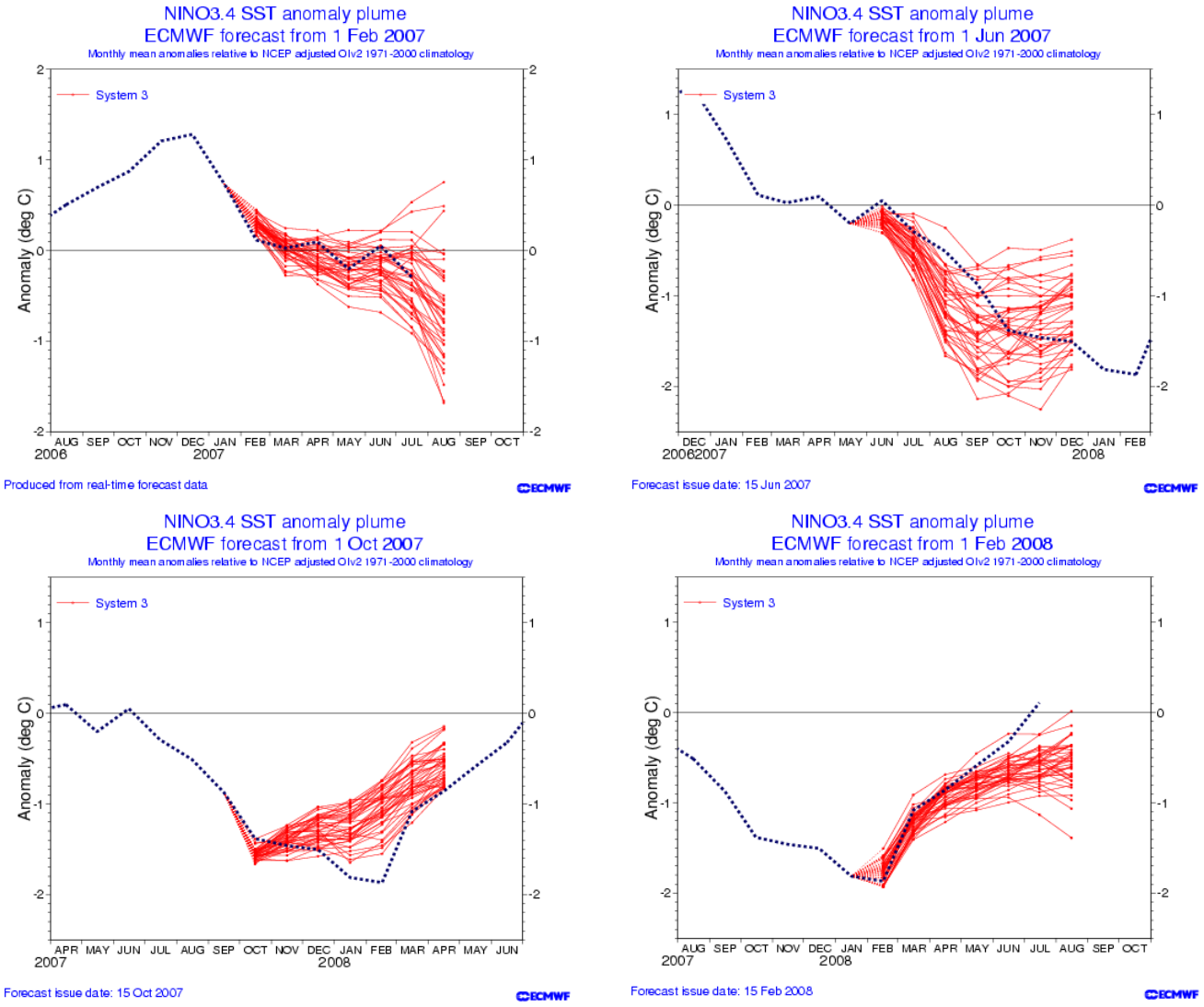


Figure 31: Plot of forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from start dates February (top left), June (top right), October (bottom left) 2007 and February 2008 (bottom right). The red lines represent the 40 ensemble members; dashed blue lines show the subsequent verification.

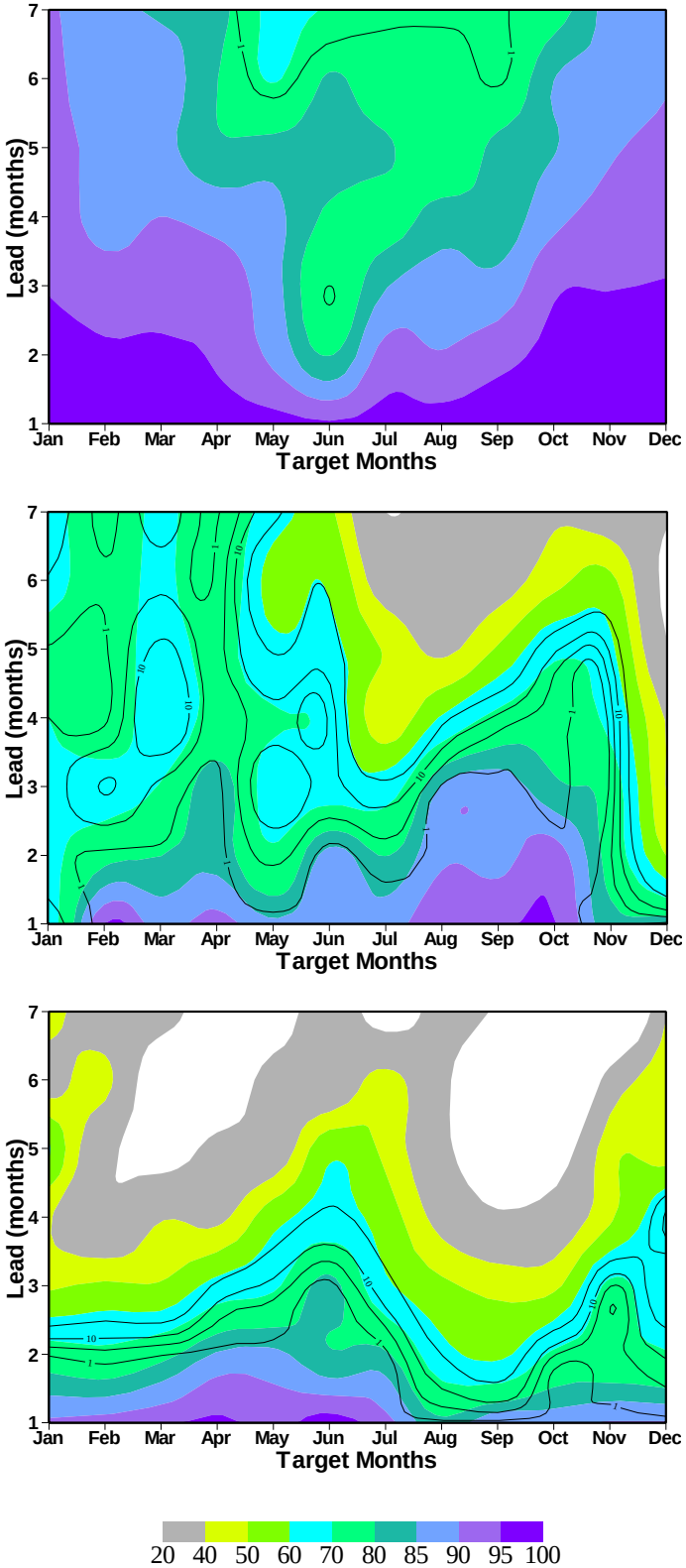


Figure 32: Anomaly correlation (%) of the ensemble-mean forecasts for SST anomalies over: NINO 3.4 area (top), Eastern tropical Indian Ocean (middle) and Southern tropical Atlantic (bottom). The monthly mean values, displayed as a function of forecast lead in months (vertical axis) versus the 'target' or verification month (horizontal axis). The correlations are computed using the hindcast 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).

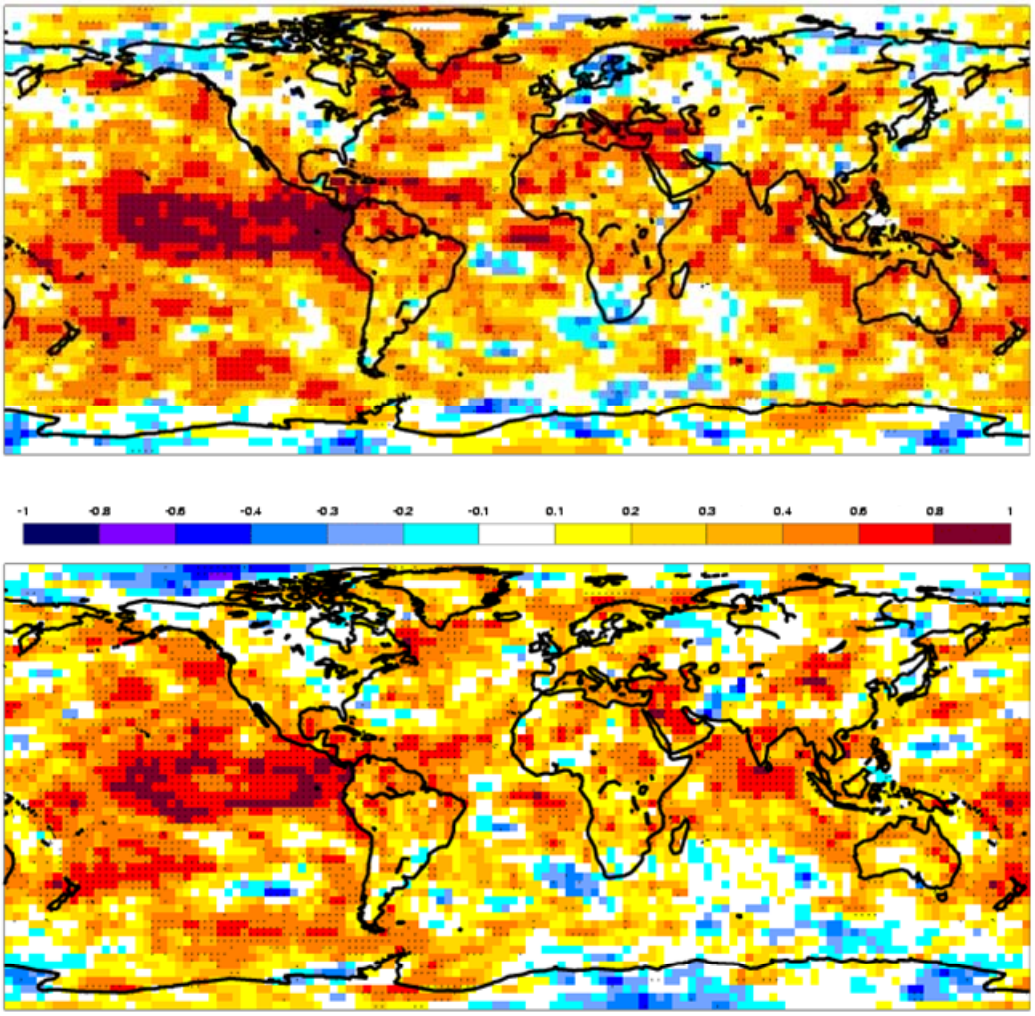


Figure 33: Spatial distribution of ROC skill score for the probability of 2m temperature anomalies being in the upper third of the climatological distribution. The scores are based on 25 years of past forecast (1981-2005) and are valid for the three-month period June to August. Scores for the forecast initiated in May are shown in the top panel and scores for the forecast initiated in April are shown in the bottom panel.

Annex - 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:

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

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:

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

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 29.

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).

For verification of EPS precipitation forecasts against analysis, 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.