TECHNICAL MEMORANDUM

654

Verification statistics and evaluations of ECMWF forecasts in 2010-2011

D.S. Richardson, J. Bidlot, L. Ferranti, A. Ghelli, T. Haiden, T. Hewson, M. Janousek, F. Prates and F. Vitart

Operations Department

December 2011

Presented to the SAC, 40th Session, 3–5 October 2011

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/

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.

At its 42nd Session (October 2010), the Technical Advisory Committee endorsed a set of two primary and four supplementary headline scores to monitor trends in overall performance of the operational forecasting systems. These new headline scores are included in the current report, and the organisation of the report has been revised accordingly. As in previous reports a wide range of complementary verification results are included and, to aid comparison from year to year, the set of additional verification scores shown here is mainly consistent with that of previous years (ECMWF Tech. Memos. 346, 414, 432, 463, 501, 504, 547, 578, 606, 635). A short technical note describing the scores used in this report is given in the annex to this document.

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)

http://www.ecmwf.int/products/forecasts/wavecharts/index.html#verification (ocean waves)

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 Technical Advisory Committee are summarised below.

9 November 2010 Cycle 36r4, including the following main changes:

- Five-species prognostic microphysics scheme, introducing cloud rain water content and cloud ice water content as new model variables
- Retuning and simplification of convective entrainment/detrainment and land/sea dependent threshold for precipitation
- Retuning of sub-grid-scale orographic gravity wave drag
- Adjustment to diffusion in stable boundary layers near the surface
- All-sky improvements of microwave radiance assimilation

- Adaptation to neutral wind of the observation operator for scatterometer data
- New soil-moisture analysis scheme (SEKF, simplified ensemble Kalman filter)
- New snow analysis based on OI (Optimum Interpolation) and upgrade of NESDIS snow cover data to 4 km resolution
- Monthly varying climatology of leaf area index (LAI), based on MODIS data
- 2 rather than 3 updates of the outer loop performed by the 4D-Var of the early delivery suite (EDS)

The main changes to the EPS included in this cycle are:

- Revision of stochastically perturbed physical tendencies
- Introduction of spectral stochastic backscatter scheme
- Retuned initial perturbation amplitudes

18 May 2011 Cycle 37r2, including the following main meteorological changes:

- Use of background error variances from the ensemble of data assimilations (EDA) by the deterministic 4D-Var assimilation
- Improvements to the cloud scheme formulation
- Changes to the all-sky assimilation of microwave data
- Reduction of AMSU-A observation errors and adjustments to MODIS AMVs
- Accounting for tangent point drift in GPS radio occultation

In addition to meteorological changes, this cycle included a major technical change: from this cycle onwards the WMO FM-92 GRIB Edition 2 (GRIB-2) format is used for the encoding of model-level data.

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 Europe and the extra-tropical northern and southern hemispheres since 1981. Each point on the curves shows the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the anomaly correlation (ACC) between forecast and verifying analysis falls below 80%. The hemispheric scores have been consistently very good over the past year, remaining above 80% to around forecast day 6 and up to day 7 in early winter in both hemispheres. As is typical for a smaller region, there is more variability in the scores for Europe, but the 80% threshold is also reached typically at around forecast day 6. Previous editions of this report have shown the forecast range at which the ACC falls below 60%. For reference, the curves for this 60% threshold are shown in Figure 2. The use of the 80% threshold refocuses the Centre's headline score on the 6 to 7 day range; it produces scores that vary more smoothly with time and trends that are overall very similar to results based on the 60% threshold.

As noted in last year's report, the record high scores achieved for Europe and the northern hemisphere in winter 2009-10 were partially a result of large anomalies associated with the strong negative phase of the North Atlantic Oscillation circulation pattern that dominated that winter. Similar conditions occurred in December 2010 (associated with the very cold early winter over northern Europe) and again this partly accounts for the record skill shown in Figure 1 (top) for December over Europe.

Figure 3 shows the evolution of performance using a skill measure based on root mean square error and using persistence as a reference instead of climatology (as used for the ACC). 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 2011. These confirm the good overall performance shown in the ACC scores. Figure 4 shows that overall synoptic activity over Europe has increased somewhat over the last year, but, in general, has been at fairly average levels compared to previous years.

Figure 5 illustrates the forecast performance for 850 hPa temperature over Europe. The distribution of daily anomaly correlation scores for day 7 forecasts is shown for each winter (December to February, top) and summer (June to August, lower panel) season since winter 1997-98. The exceptional winter 2009-10 performance is apparent for the 850 temperature scores, with a greater fraction of the individual forecasts achieving very high ACC scores than in previous years. The proportion of very good forecasts for winter 2010-11 is, as expected, lower than 2009-10 but compares well with previous years. Summer 2011 scores are comparable with those for 2010, with relatively few occasions of moderate or poor skill at the 7-day range.

Figure 6 shows the time series of the average RMS difference between 4 and 3 day (blue) and 6 and 5 day (red) 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 level of consistency between consecutive forecasts has been maintained since this model change.

The quality of ECMWF forecasts for the upper atmosphere in the extratropics is shown through the time series of wind scores at 50 hPa in Figure 7. 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 8, which shows the evolution of the Continuous Ranked Probability Skill Score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. As for the deterministic forecast, the EPS skill reached record levels in winter 2009-10. There has been some reduction from these record levels, especially over Europe, as might be expected and as was seen also for the deterministic forecast. However, there is still a strong signal that the EPS performance in 2010 and 2011 has been consistently high when compared to previous years. A number of changes have been made to the EPS in this period, including improvements to both the initial perturbations (36r2, June 2010) and representation of model uncertainties (cycle 35r3, September 2009; cycle 36r4, November 2010, see section 2) and the increase in resolution (36r1, January 2010). The sustained high skill is consistent with the improvements from these model changes.

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 9. The match between the spread and error improved in 2010-11 compared to the previous two winter seasons. The over-dispersion of the EPS for 500 hPa height in the early forecast range, noted in previous years, is not apparent for 2010-11 and the under-dispersion at longer ranges is reduced. In general, the EPS is still under-dispersive for temperature at 850 hPa, although uncertainty in the verifying analysis should be taken into account when considering the relationship between spread and error in the first few days. The introduction of the ensemble of data assimilations (EDA) in the initial perturbations (June 2010) improved the short-range over-dispersion, while this, together with the changes to the representation of model uncertainty, improved the overall dispersion of the EPS for both parameters.

Figure 10 shows the skill of the EPS using CRPSS 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 has been consistent for the first three winter seasons since this extension, confirming the positive skill at this forecast range. The performance in winter 2009-10 was clearly exceptional compared to earlier years. In part, as for the deterministic forecast, the anomalous flow made some contribution to the high scores. Skill for 2010-11 was equal to that for 2009-10 for the early part of the forecast. Although not maintaining the same record levels at longer ranges, the performance for 2010-11 was substantially higher than for earlier years.

3.1.2 Tropics

The forecast performance over the tropics, as measured by root mean square vector errors of the wind forecast with respect to the analysis, is shown in Figure 11. 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. The performance in the tropics has been consistent over the last 3 years.

3.2 ECMWF vs other NWP centres

The common ground for comparison is the regular exchange of scores between WMO designated global data-processing and forecasting system (GDPFS) centres under WMO/CBS auspices, following agreed standards of verification. The new scoring procedures for upper-air fields used in the rest of this report were approved for use in this score exchange by the 16th WMO Congress in 2011 and will be implemented at the other participating centres in the coming months. However, the scores from all centres for the period of this report have been computed using the existing procedures. Therefore the ECMWF scores shown in this section are also computed using these existing procedures (as presented in previous reports).

Figure 12 (northern hemisphere extratropics) and Figure 13 (southern hemisphere extratropics) show time series of such scores for both 500 hPa geopotential height and mean sea level pressure (MSLP). ECMWF continues to maintain a lead over the other centres; as in previous years, this is larger for the southern hemisphere. Overall, however, the difference in performance between centres is decreasing. In particular the UK Met Office forecasts have reduced the overall gap for the northern hemisphere compared to previous years.

WMO exchanged scores also include verification against radiosondes over regions such as Europe Figure 14, showing both 500 hPa geopotential height and 850 hPa wind forecast 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 comparison for the tropics is summarised in Figure 15 (verification against analyses) and Figure 16 (verification against observations). When verified against the centres' own analyses, the UK Met Office has had the lowest short-range errors since mid-2005, while at day 5 ECMWF and the UK Met Office performance is similar. The errors of the JMA (Japan Meteorological Agency) forecast system have steadily decreased over several years and are now comparable with those of the UK Met Office model at both short and medium ranges. In the tropics, verification against analyses (Figure 15) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations, the ECMWF, UK Met Office and JMA models have very similar short-range errors.

4 Weather parameters and ocean waves

4.1 Weather parameters - deterministic and EPS

The new supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 17. The left-hand panel shows the lead time for which the SEEPS skill for the deterministic forecast for 24-hour precipitation over the extra-tropics remains above 45%. The right-

hand panel shows the lead time for which the CRPSS for the EPS probability forecast of 24-hour precipitation over the extra-tropics remains above 10%. Both scores are verified against station observations. The increase in skill of the deterministic forecast in 2010 is associated with the 5-species prognostic microphysics scheme introduced on 9 Nov 2010 (Cy36r4); the increased skill of the EPS forecast between mid-2009 and mid-2010 is associated with the EPS resolution increase on 26 January 2010 (Cy36r1). The temporal averaging of the scores leads to step-wise changes in model skill appearing as gradual changes over 12 months in the plots.

ECMWF has begun a routine comparison of the precipitation forecast skill of ECMWF and other centres for both the deterministic forecast and the EPS (using the TIGGE ensembles archived in MARS). Results using these same headline scores for the last 12 months show a consistent clear lead for ECMWF with respect to the other centres (Figure 18). The reversed ranking of the JMA and UK Met Office at short lead times is due to a greater drop in skill in the JMA model during the northern hemisphere convective season (JJA) compared to the other models.

The summer 2011 edition of the ECMWF Newsletter (No. 128) contains an article about the new SEEPS score used for the verification of the deterministic precipitation forecasts.

Long-term trends in mean error and standard deviation of error for 2 m temperature, specific humidity, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 19 to Figure 23. 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.

In recent winters and early spring there have been significant negative night-time temperature biases over Europe (Figure 19). A particular factor in winter 2010-11 was an increase in the number of large negative errors (forecast was too cold) at night, especially in cold conditions (Figure 23, right panel). These large errors occurred particularly in Scandinavia and Eastern Europe. There was also a reduction in the number of large positive errors in these areas. Overall, these events resulted in the night-time temperature bias (Figure 23, left panel) being more negative in these regions than in previous years. Investigations identified a problem related to the model's handling of low cloud in very cold conditions and model improvements to address the issue have been developed. A summary of this specific issue is presented in document ECMWF/TAC/43(11)6.

4.2 Ocean waves

The quality of the ocean wave model analysis is shown in the comparison with independent ocean buoy observations in Figure 24. The top panel of Figure 24 shows a time series of the analysis error for the 10 m wind over maritime regions using the wind observations from these buoys. The error has steadily decreased since 1997, providing better quality winds for the forcing of the ocean wave model; in general the wind errors have been lower in 2010-11 than in the previous year. The errors in the wave analysis have been consistently good over the past year, particularly in the most recent months. The long-term trend in the performance of the wave model forecasts is shown in Figure 25 and Figure 26; overall, the performance in 2010-11 is similar to that for the previous 12-months.

ECMWF maintains a regular inter-comparison of performance between wave models from different centres on behalf of the Expert Team on Waves and Storm Surges of the WMO-IOC Joint Technical

Commission for Oceanography and Marine Meteorology (JCOMM). The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of ocean buoys (mainly located in the northern hemisphere). An example of this comparison is shown in Figure 27 for the most recent 3-month period (May-July 2011). ECMWF forecast winds are used to drive the wave model of Météo France; the wave models of the two centres are also similar, hence the closeness of their errors in Figure 27. Of the centres not using ECMWF winds, the UK Met Office has the lowest errors for both wind speed and wave height.

A comprehensive set of wave verification charts is now available on the ECMWF website, including the figures shown in this report: http://www.ecmwf.int/products/forecasts/wavecharts/

5 Severe weather

Supplementary headline scores for severe weather are:

- the skill of the extreme forecast index (EFI) for 10 m wind verified using the relative operating characteristic area (Section 5.1)
- the tropical cyclone position error for the deterministic forecast (Section 5.2)

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. Verification of the EFI has been performed using synoptic observations over Europe from the GTS. An extreme event is judged to have occurred, 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). The headline measure, skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead), is shown in Figure 28 (top), together with the corresponding results for 24-hour total precipitation (bottom left) and 2 m temperature (bottom right). Each curve shows a four-season running mean of ROC area skill scores from 2004 to 2011; the final point on each curve includes the spring (March-May) season 2011. For all parameters, there is a clear improvement during this period.

5.2 Tropical cyclones

The 2010 North Atlantic hurricane season was, after a slow start, a very active season, consistent with the La Niña conditions. From April 2010 onwards, the seasonal tropical storm predictions consistently indicated enhanced activity over the Atlantic for the 2010 season (see section 6.3, Figure 35).

The tropical cyclone position error for the 3-day deterministic forecast is one of the supplementary headline scores for severe weather. The average position errors for the deterministic medium-range forecasts of all tropical cyclones (all ocean basins) over the last nine 12-month periods are shown in Figure 29. Errors in the forecast intensity of the tropical cyclones (represented by the reported sea-level pressure at the centre of the system) are also shown in Figure 29. The position errors (top left panel, Figure 29) are similar for the last three years, at around 200 km on average for the 3-day

forecast. The bottom right panel of Figure 29 shows the average speed error for the tropical cyclones for the last three years. Typically the cyclones move too slowly in the forecast (by around 1 km/hour) compared to the observed speed; there has been some reduction of this slow bias over the period, but because of the substantial year to year variations in the number and intensity of cyclones, there is some uncertainty in these figures. The mean error (bias) in the tropical cyclone intensity has reduced in the last year, indicating that, on average, forecast cyclones are just slightly too weak (top right panel Figure 29). The mean absolute error of the tropical cyclone intensity is slightly lower than in previous years (bottom right panel Figure 29). As for the speed errors, there is a relatively large uncertainty in these scores because of the year to year variations in the number of storms.

The EPS tropical cyclone forecast is presented on the ECMWF website 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 30. Results show an overconfidence for the three periods, with small variations from year to year. The skill is shown by the ROC and the modified ROC which uses the false alarm ratio instead of the false alarm rate on the horizontal axis (this removes the reference to non-events in the sample and shows more clearly the reduction in false alarms in those cases where the event is forecast). Both measures show similar performance to the previous year.

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) since March 2008. The combined system enables users to be provided with EPS output uniformly up to 32 days ahead, once a week. It also introduced a coupled ocean-atmosphere model for the forecast range day 10-15 for the forecast started from the 00 UTC analysis, on a daily basis.

Figure 31 shows the ROC area score computed over each grid point for the 2 m 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). This is now achieved in all regions; stronger shades indicate the regions of higher skill. Currently, the anomalies are relative to the past 18-year model climatology. The monthly forecasts are verified against the ERA-Interim reanalysis or the operational analysis, when ERA-Interim is not available. Although these scores are strongly subject to sampling, they provide users with a first estimate of the forecast skill's spatial distribution, showing that the monthly forecasts are more skilful than climatology over all areas.

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

Figure 32 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 m temperature is in the upper third of the climate distribution over the extra-tropical northern hemisphere. For the 12–18 day period, the monthly forecast has shown a clear substantial advantage over persistence of the medium-range (days 5-11) forecast throughout the last year. For the 19–32 day range, the system had only a small lead for the 2010-11 winter, but

performed well compared to persistence for other seasons. The exceptionally high scores reached in winter 2009-10 for forecast ranges 12-18 and 19-32 days were associated with the very persistent negative North Atlantic Oscillation (NAO) conditions of that winter.

A strong negative NAO pattern in winter is typically associated with anomalously cold weather over northern Europe, as occurred in late 2011 and in much of the 2009-10 winter. A preliminary study of the flow-dependent behaviour of the monthly forecast has investigated its ability to reproduce these 2 m temperature patterns in its weekly mean predictions; data for the October to April period over Europe from the last seven years was used. Figure 33 shows the composite map of temperature anomalies for those weeks with the strongest projection onto this NAO circulation pattern during this period. The analysis fields (left) show substantial cold temperatures extending over a large area of northern Europe, with coherent areas of positive temperature anomalies to the south-east and west over the Atlantic. The 12-18 day forecast (right) reproduces this pattern well. The magnitude of the temperature anomaly is not as strong as in the analysis, but this is at least partly a statistical consequence of selecting cases based on the magnitude of the observed pattern (conditional sampling). The NAO pattern can persist for several weeks, as happened during winter 2009-10. To evaluate the performance of the monthly forecast in predicting the onset of this regime, the evaluation was repeated for the subset of cases in which the strong negative NAO pattern was first established (i.e. the previous week did not have a strong observed signal; this happened in 24 of the 35 cases; the remaining 11 cases were persistent situations, where the strong signal was already established). This confirmed the ability of the monthly forecast to predict the establishment of a strong event at the 12-18 day range.

However, the monthly forecast is not so successful in predicting the opposite phase of this circulation regime (associated with mild winters over Europe, not shown). Verification of the probabilistic forecasts confirms that the monthly forecast is more skilful in forecasting cold events than in forecasting the milder winter conditions.

This flow-dependent verification will be further developed, to provide a more comprehensive evaluation of the monthly forecasting system's ability to predict regimes and regime transitions.

6.2 The 2010-2011 El Niño forecasts

The 2009-10 El Niño event peaked in December 2009 and dissipated quickly by May 2010. This was followed by a transition to cooler conditions and a La Niña event persisted through the second half of 2010. The La Niña declined during the first half of 2011 and sea surface temperatures in the tropical Pacific were close to normal in June and July. The development and decline of the La Niña was well forecast by the seasonal forecast system (Figure 34). This was also well forecast in general by the EUROSIP partners, although there was some uncertainty between the different systems on the magnitude and duration of the event in the earlier forecasts (start dates in May and August 2010).

6.3 Tropical storm predictions from the seasonal forecasts

From April 2010 onwards, the seasonal tropical storm predictions consistently indicated enhanced activity over the Atlantic for the 2010 season, consistent with the forecasts of La Niña conditions. The June forecast predicted between 18 and 26 named tropical storms in the Atlantic (Figure 35); 19 were observed, the highest number since 2005. Over the western Pacific, seasonal predictions indicated reduced activity compared to normal, again consistent with La Niña, and this also verified. The

forecast for the eastern Pacific also gave a significant signal for reduced tropical storm frequency; this was also correct, although it should be remembered that evaluation using the seasonal reforecasts shows that the skill of the tropical storm forecast in the eastern Pacific is generally low. However, verification shows that the skill in predicting ACE (accumulated cyclone energy) over the Atlantic basin, calculated using the most recent 20 years, is substantial, with a correlation of 0.65 between ensemble mean forecast and observation (Figure 36). There is also moderate skill overall for the western Pacific: the correlation for that region is 0.53.

6.4 Seasonal forecast performance for the global domain

A set of verification statistics based on hindcast integrations (1981-2005) from the operational System 3 has been produced and is presented alongside the forecast products on the ECMWF website, for example:

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

A set of verification statistics based on the hindcast integrations (1987-2005) from the operational EUROSIP multi-model is under development. The skill measures are the same as those used to evaluate the ECMWF seasonal forecast system.

As discussed above, the seasonal forecasting system performed well in predicting the decline of La Niña during the first half of 2011 (Figure 34). It provided a good prediction of the atmospheric anomalies for 500 hPa height over the Pacific region during December - February and temperature and precipitation anomalies related to La Niña, such as the increased rainfall over north-east Australia. However, the cold temperatures over northern and eastern Europe last winter were not well represented by the seasonal forecast.

References

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

List of Figures

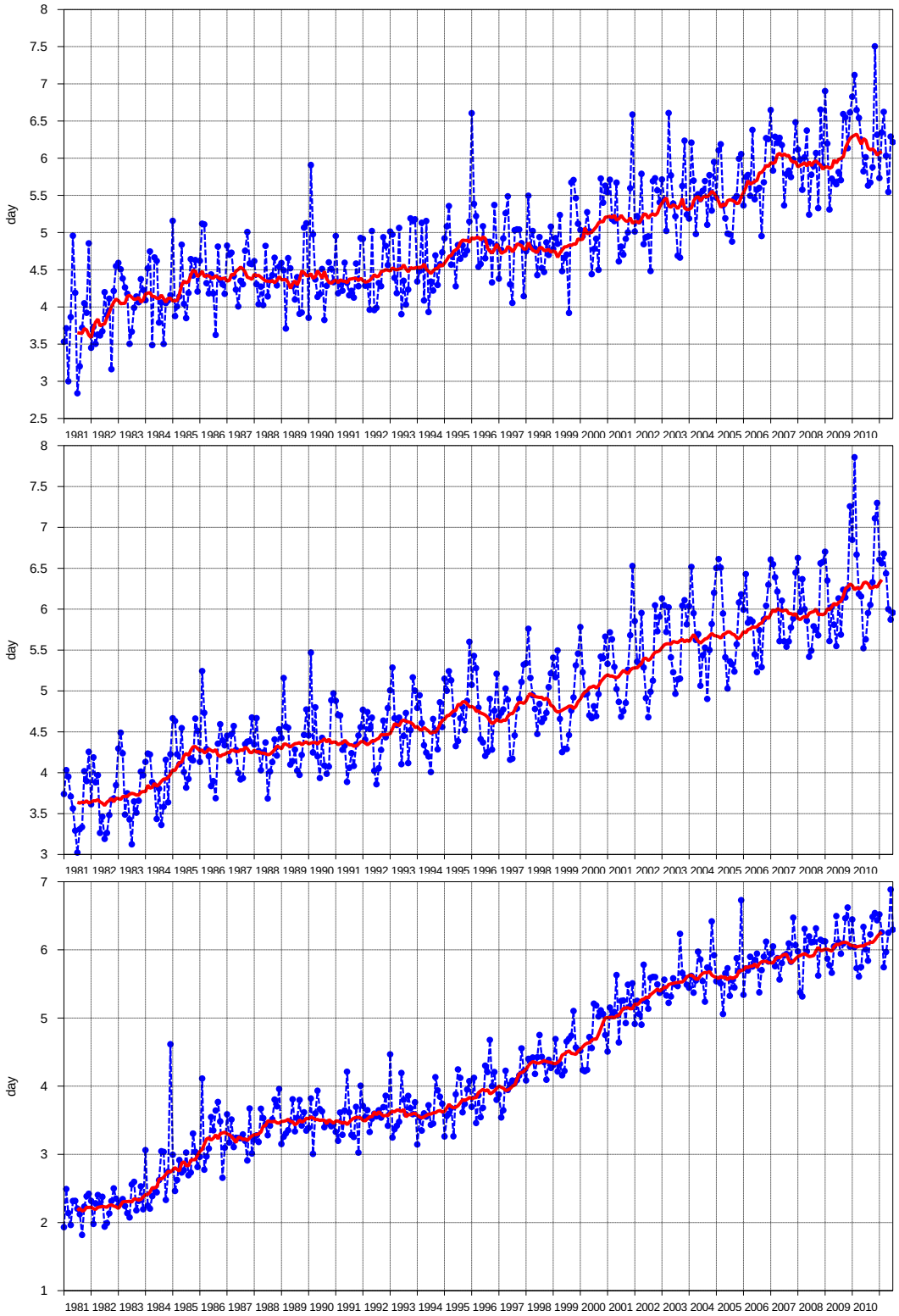


Figure 1: Primary headline score for the deterministic forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).

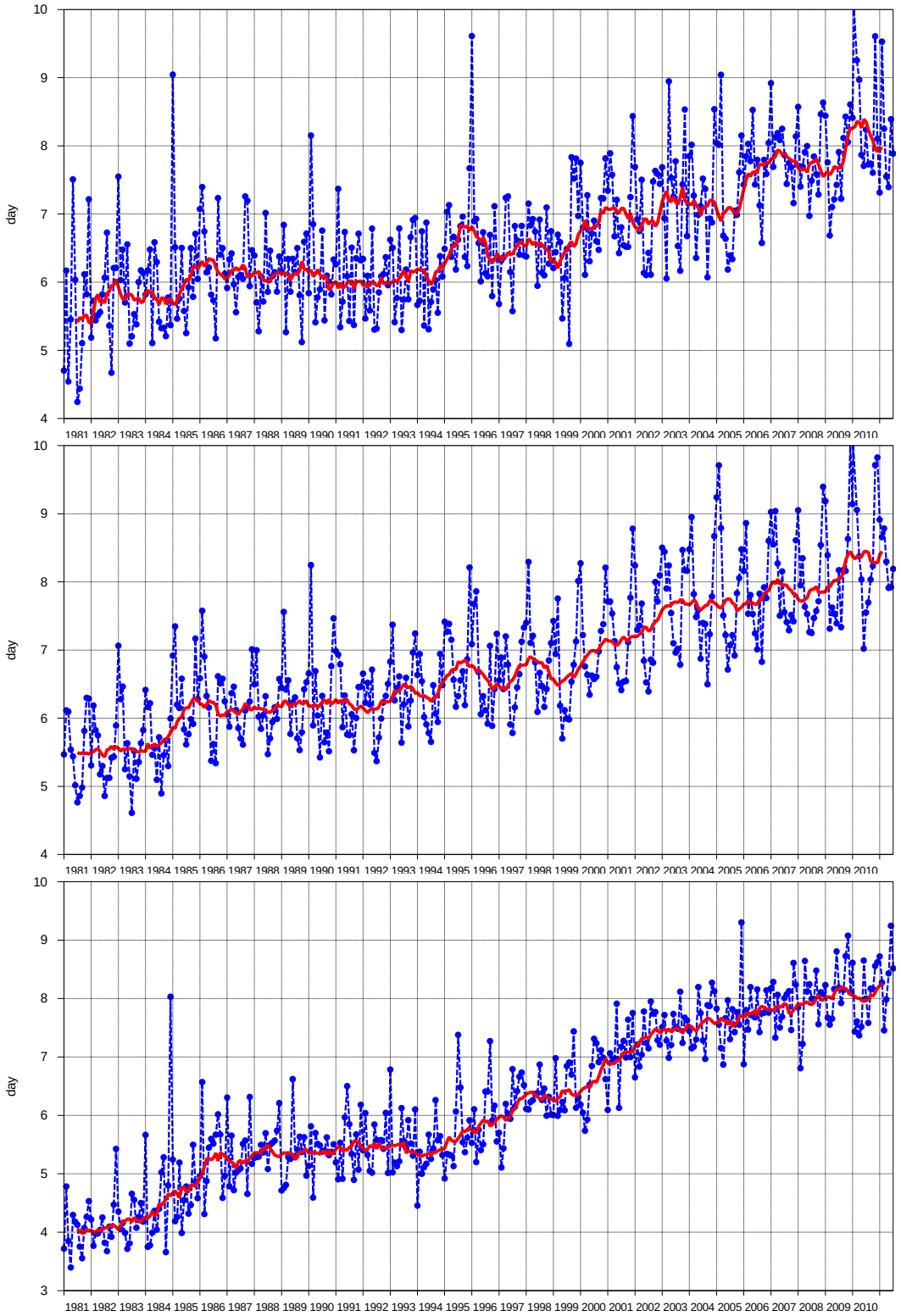
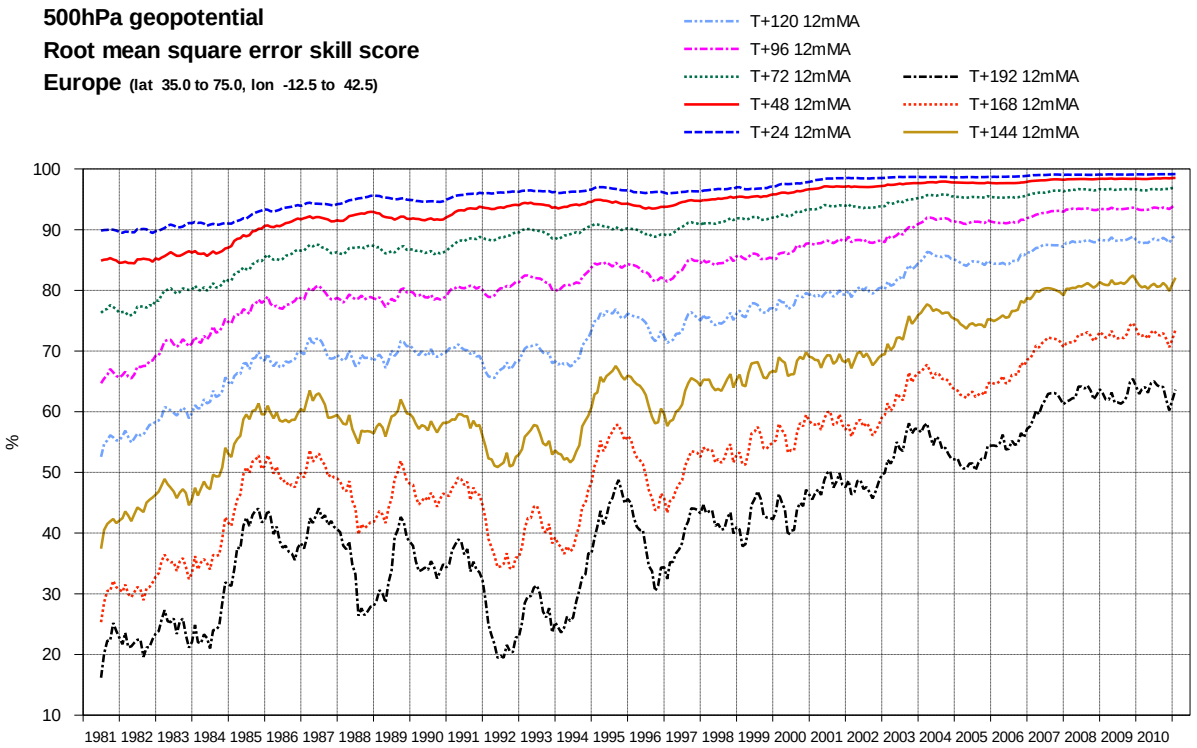


Figure 2: As Figure 1 but showing the old headline score based on the anomaly correlation reaching the threshold value 60% (cf. TAC reports on verification scores from previous years). If the monthly mean correlation remains above 60% throughout the 10-day forecast range, this is indicated by the absence of a blue symbol for that month (e.g. northern hemisphere and Europe for February 2010).



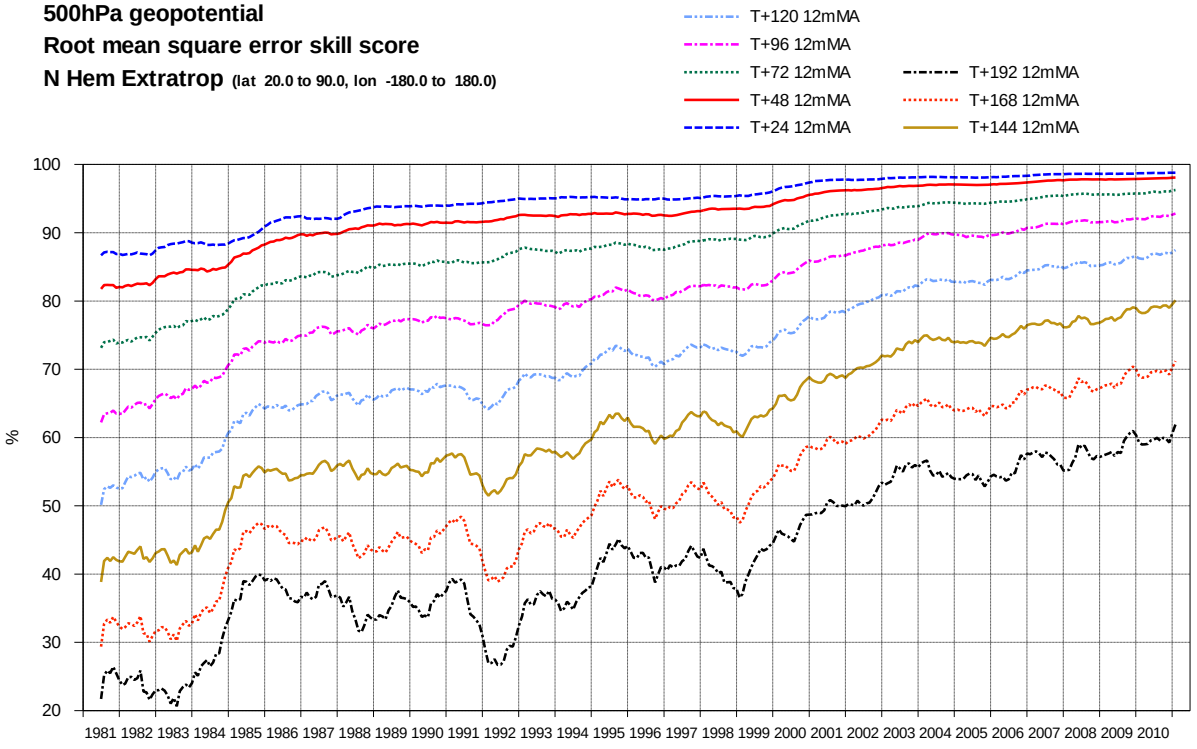


Figure 3: 500 hPa geopotential height skill score for Europe (top) and the northern hemisphere extra-tropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2010 - July 2011.

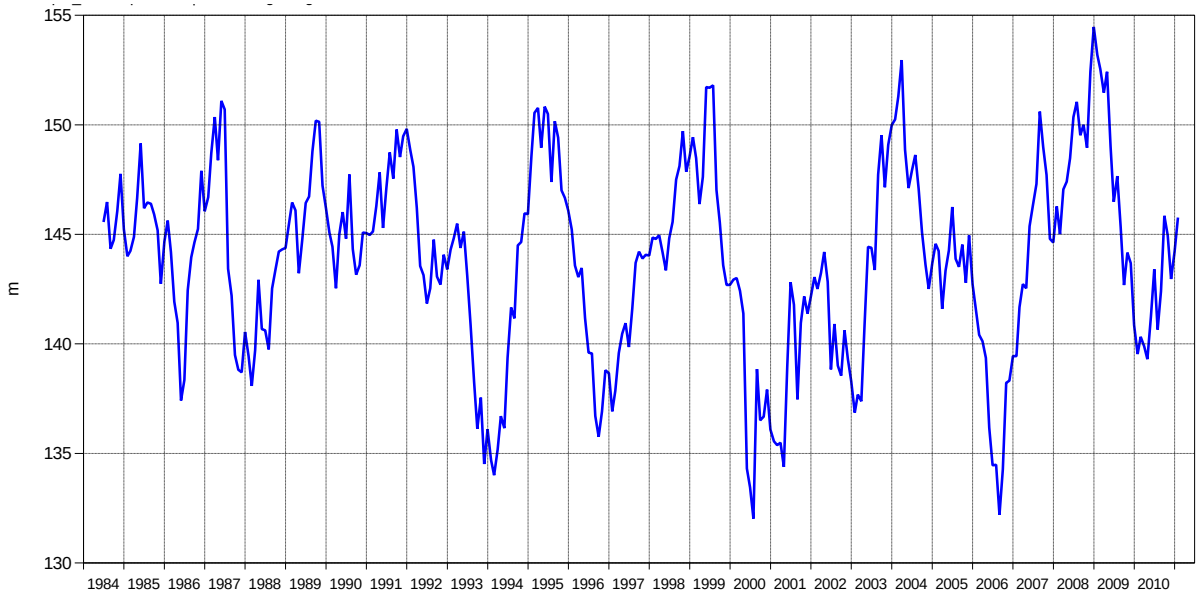


Figure 4: Root mean square error of forecasts made by persisting the analysis over 7 days (168 hours) and verifying it as a forecast for 500 hPa geopotential height over Europe. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2010 - July 2011.

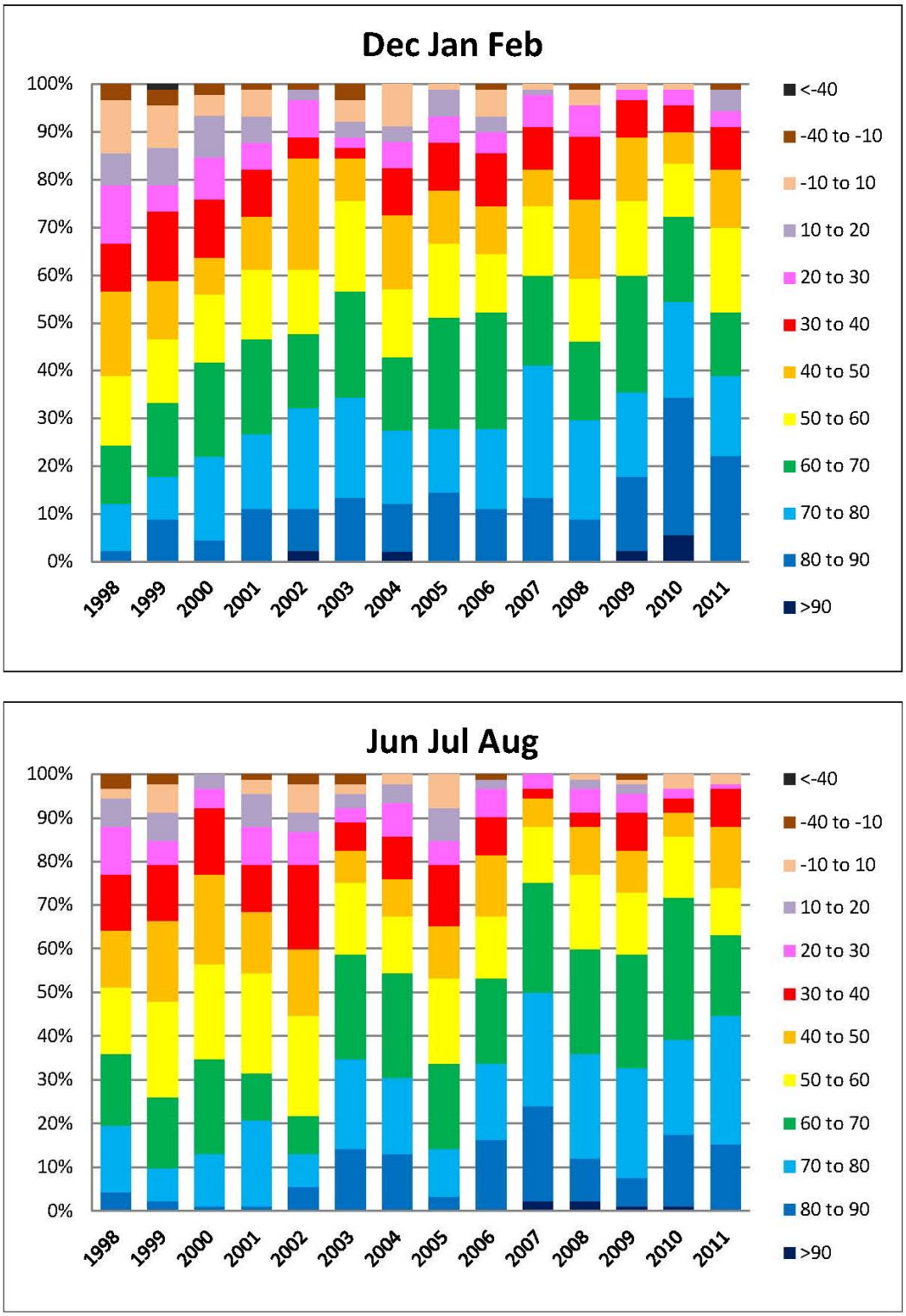


Figure 5: Distribution of Anomaly Correlation of the Day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997-1998.

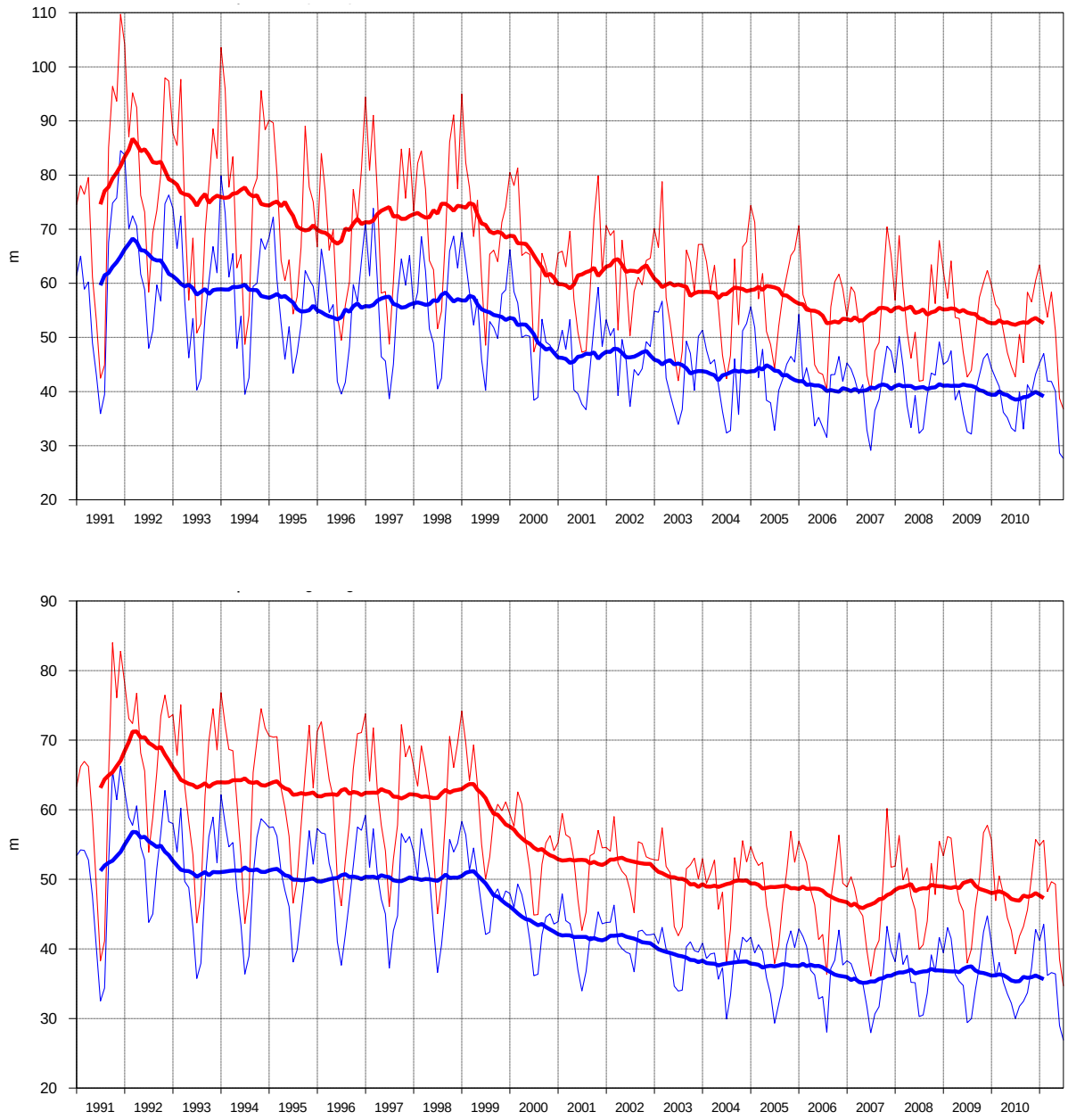


Figure 6: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96-120 h (blue) and 120-144 h (red). 12-month moving average scores are also shown (in bold).

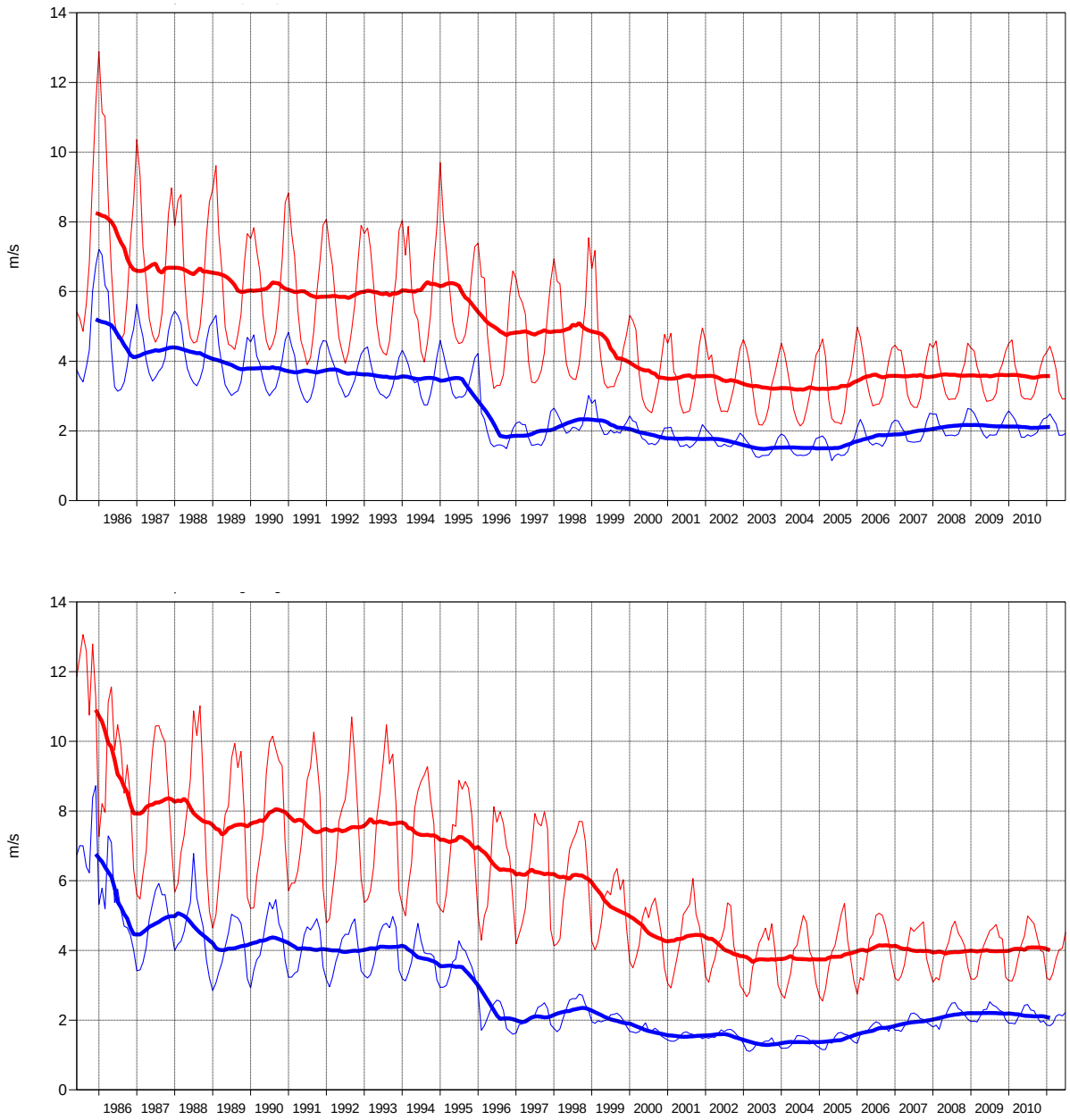


Figure 7: Model scores in the northern (top) and southern (bottom) extra-tropical stratosphere. Curves show the monthly average RMS vector wind error at 50 hPa for 1-day (blue) and 5-day (red) forecasts. 12-month moving average scores are also shown (in bold).

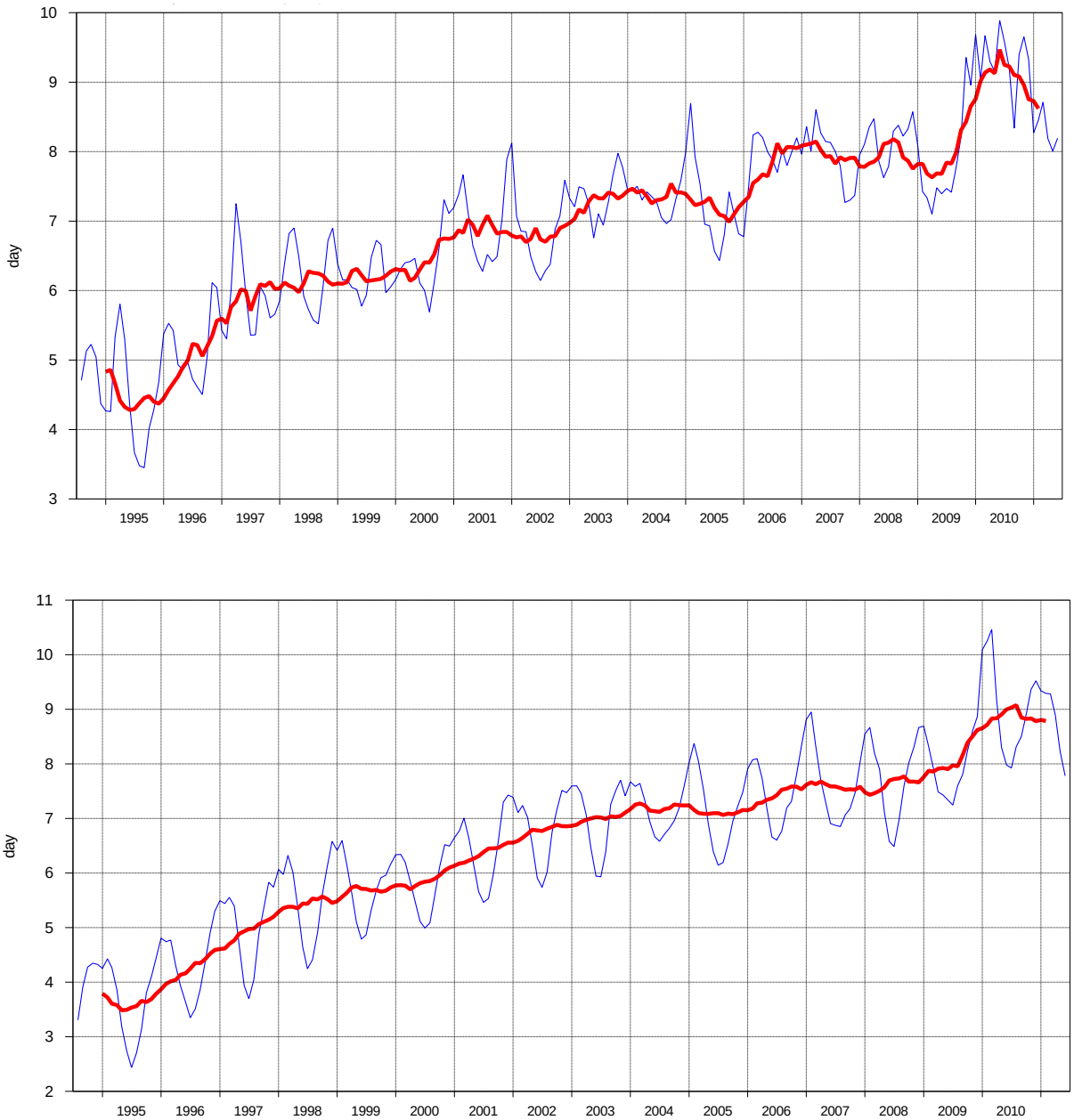


Figure 8: Primary headline score for the EPS probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance – each point on curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous ranked probability skill score falls below 25% for Europe (top), northern hemisphere extratropics (bottom).

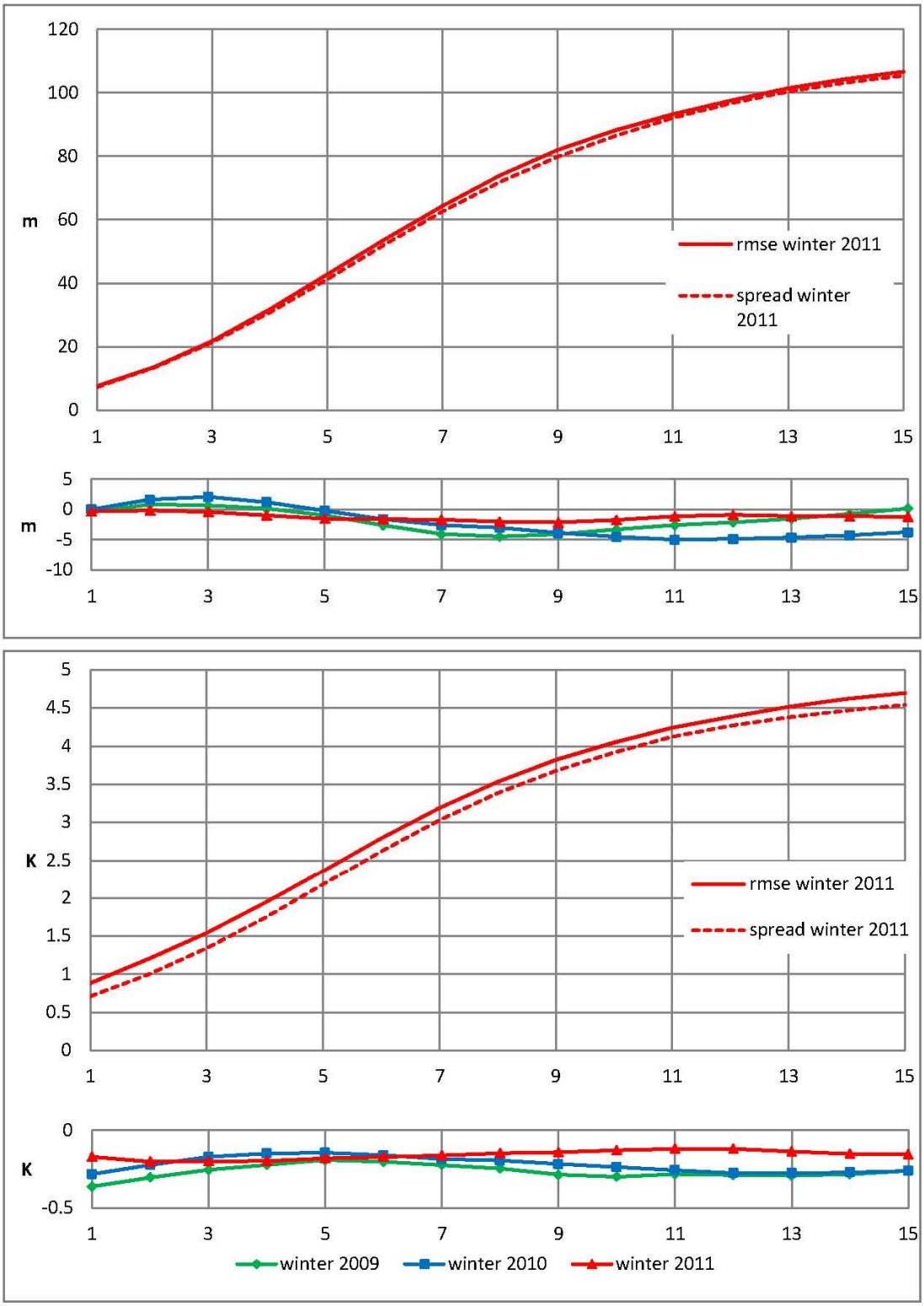


Figure 9: Ensemble spread (standard deviation, dashed lines) and root mean square error of ensemble-mean (solid lines) for winter 2010-2011(upper figure in each panel), complemented with differences of ensemble spread and root mean square error of ensemble-mean for last 3 winter seasons (lower figure in each panel, negative values indicate spread is too small); plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extra-tropical northern hemisphere for forecast days 1 to 15.

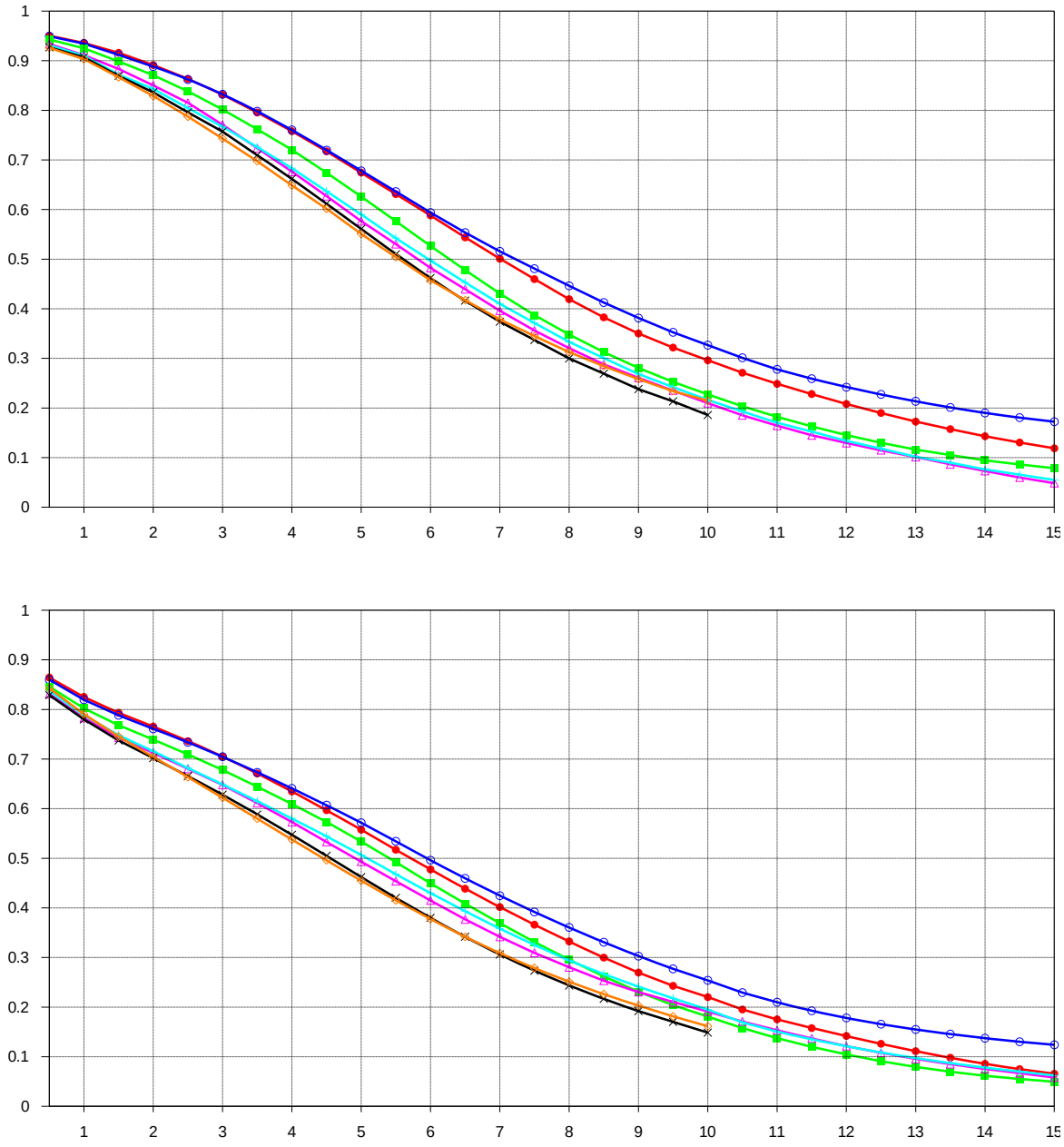


Figure 10: Continuous ranked probability skill score for 500 hPa height (top) and 850 hPa temperature (bottom) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. Skill from the EPS day 1-15 forecasts is shown for winters 2010-11 (red), 2009-10 (blue), 2008-09 (green), 2007-08 (magenta) and 2006-07 (cyan). The EPS only ran to 10 days in previous years: 2005-06 (black), 2004-05 (orange).

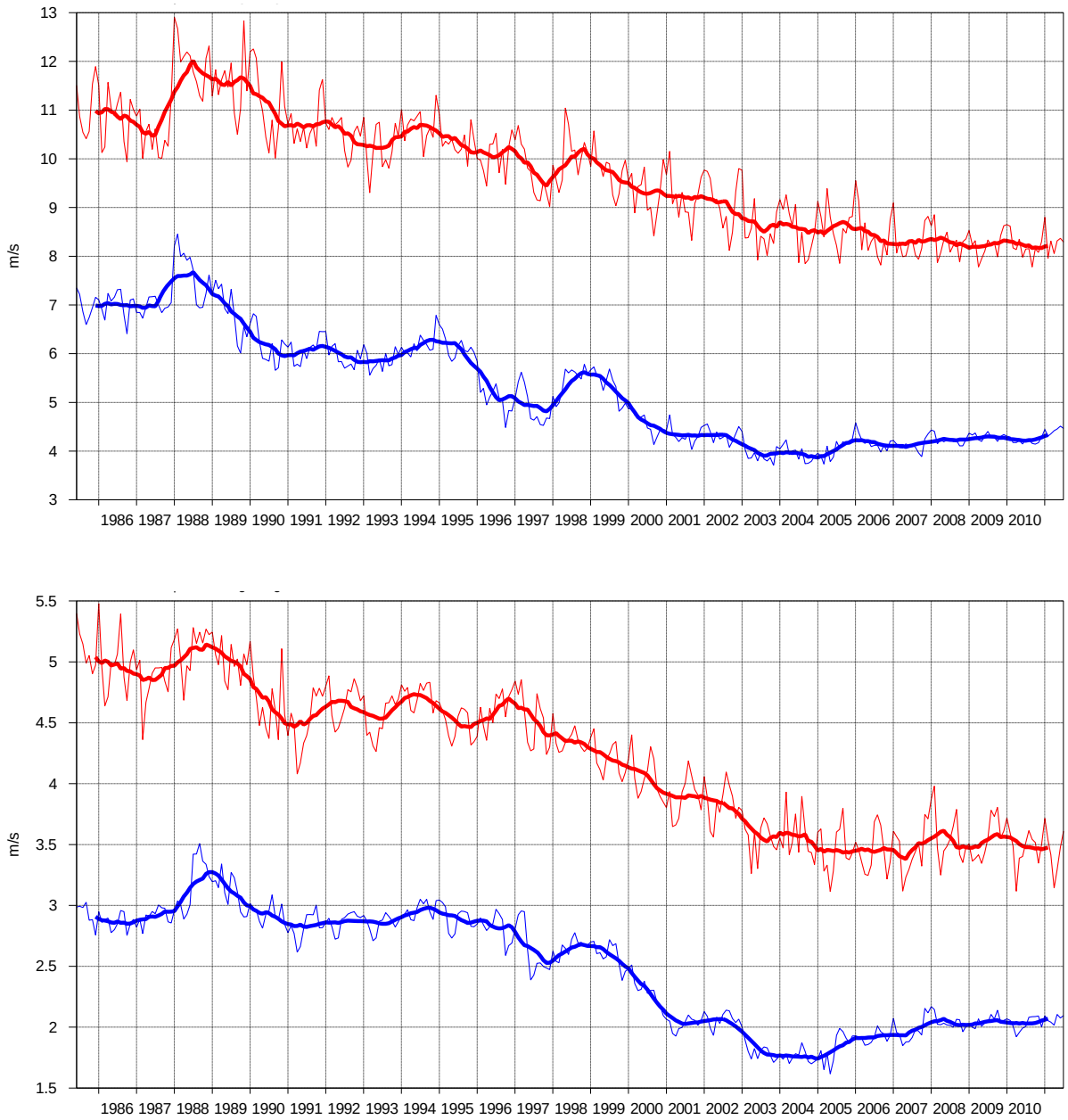


Figure 11: Model scores in the tropics. Curves show the monthly average root mean square vector wind errors at 200 hPa (top) and 850 hPa (bottom) for 1-day (blue) and 5-day (red) forecasts. 12-month moving average scores are also shown (in bold).

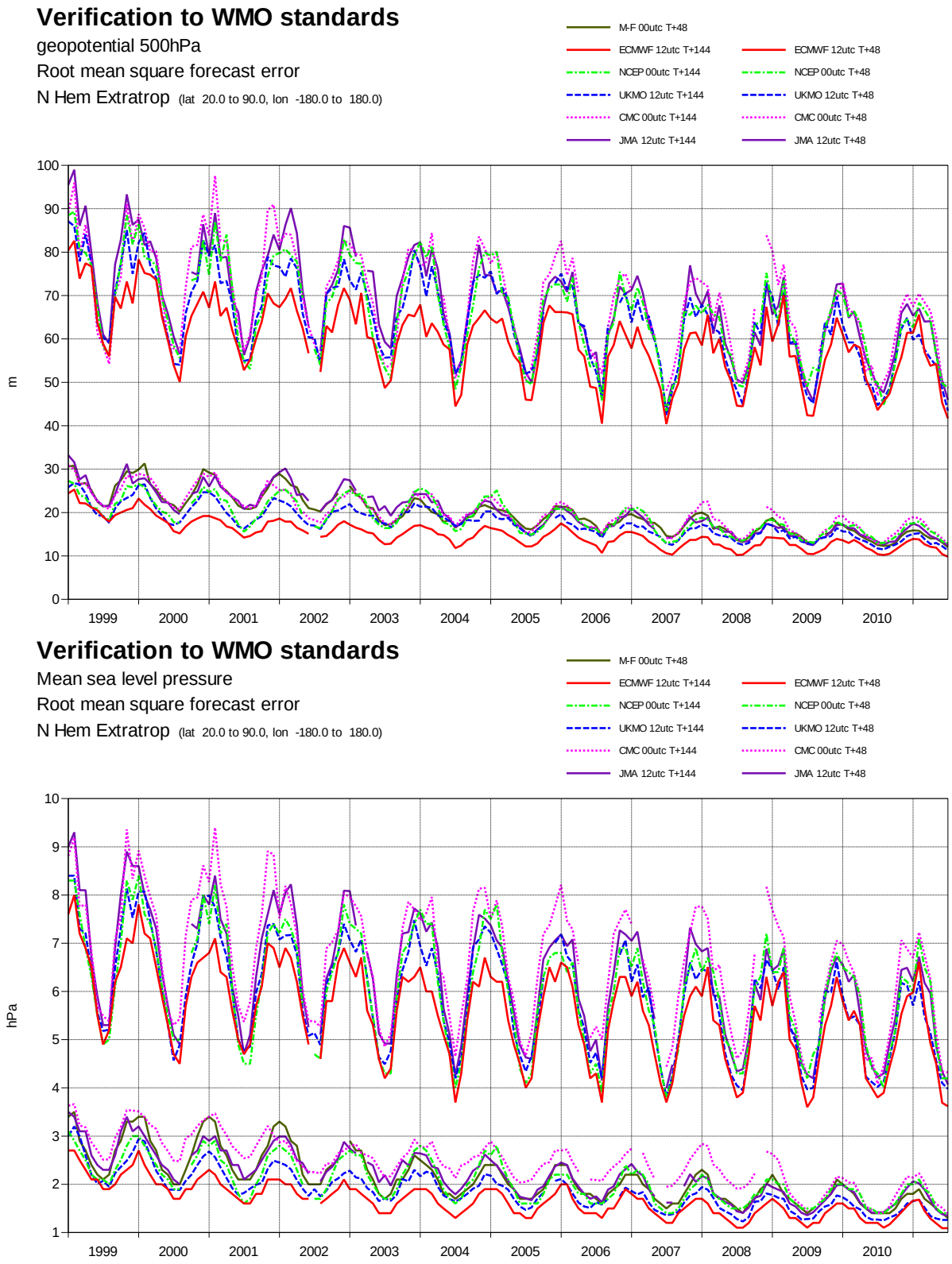
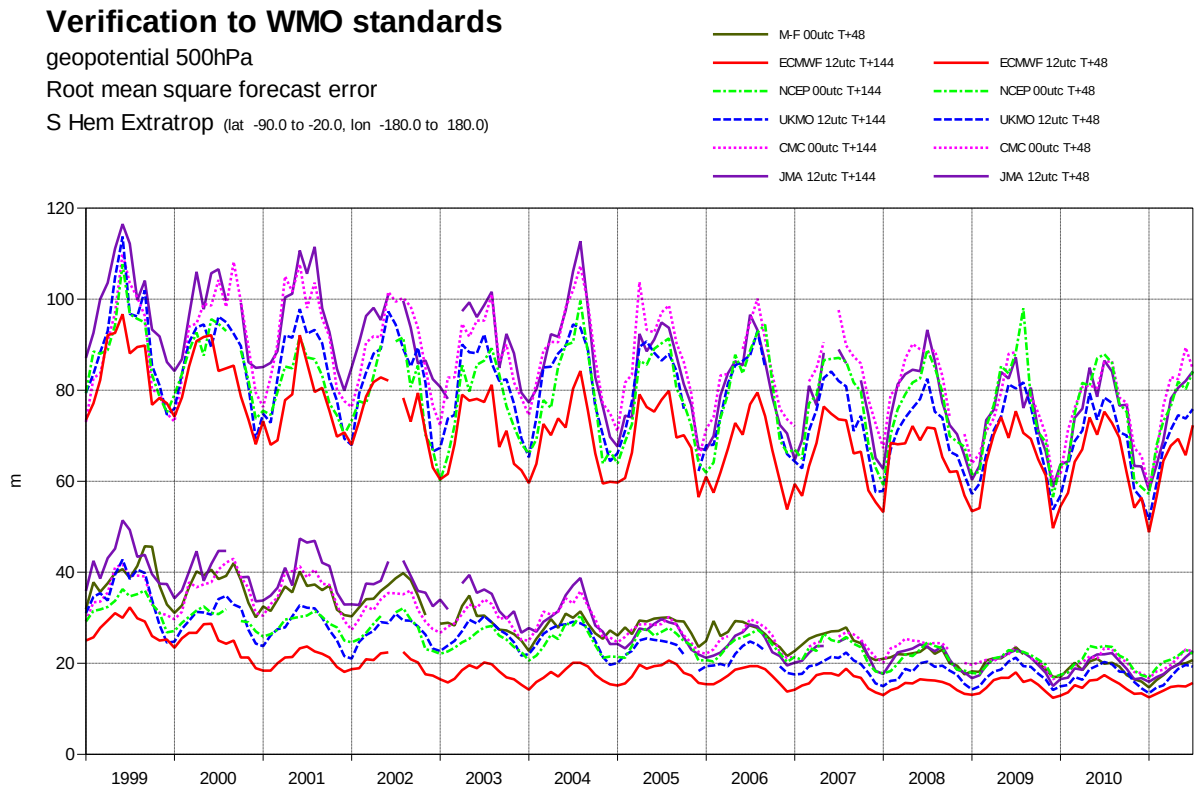


Figure 12: WMO/CBS exchanged scores from global forecast centres. RMS error over northern extratropics for 500 hPa geopotential height (top) and MSLP (bottom). In each panel the upper curves show the 6-day forecast error and the lower curves show the 2-day forecast error. Each model is verified against its own analysis. JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Meteorological Office, NCEP = U.S. National Centers for Environmental Prediction, M-F = Météo France.



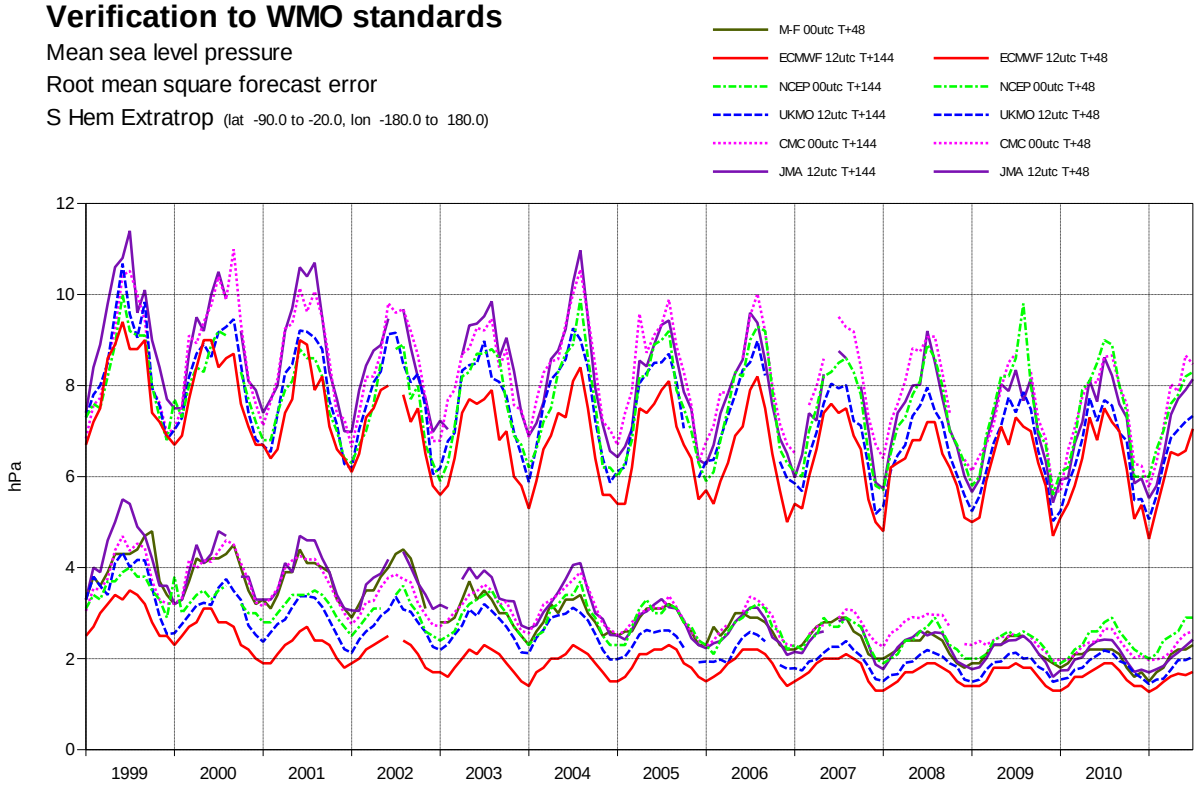


Figure 13: As Figure 12 for the southern hemisphere.

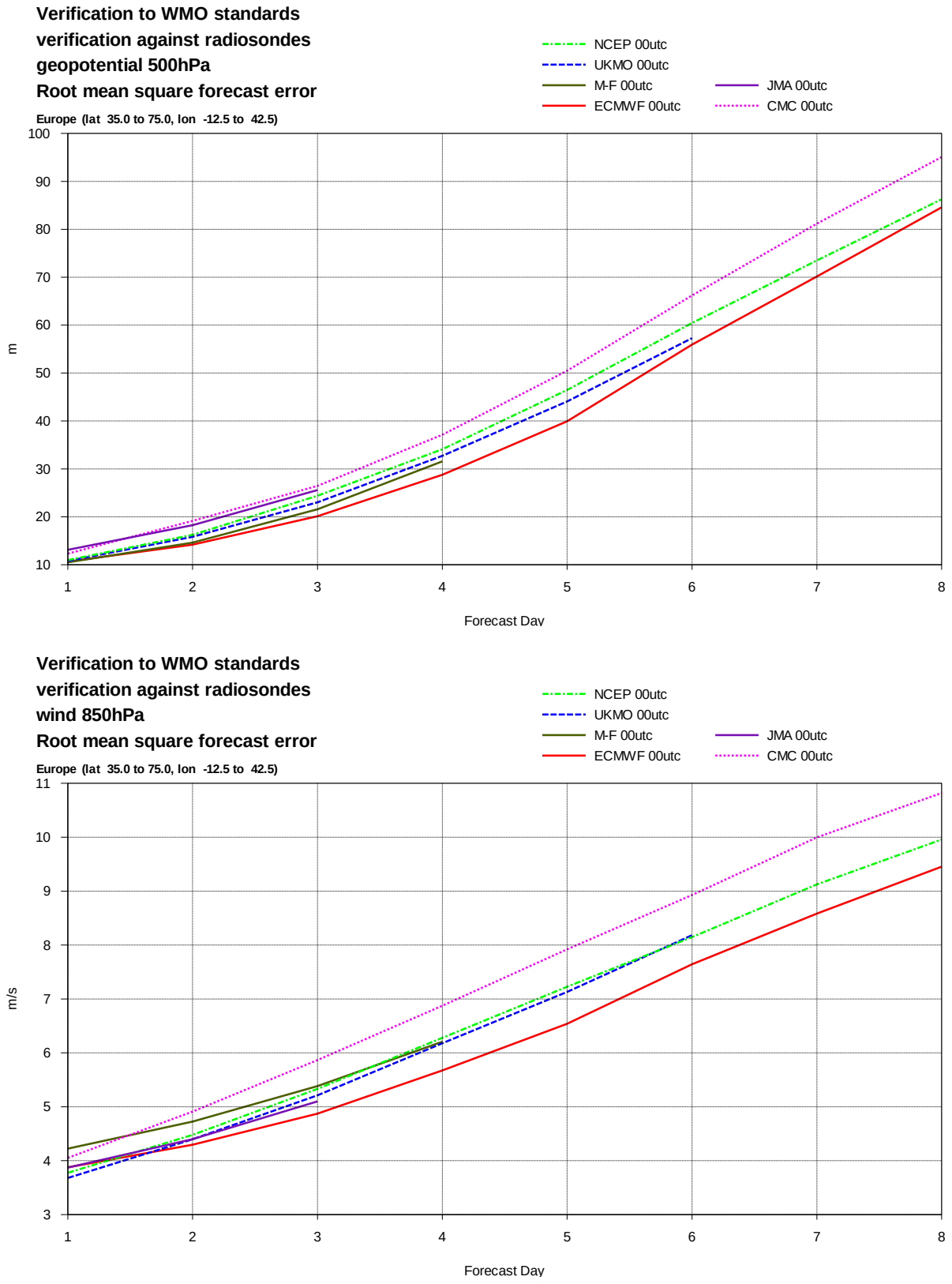


Figure 14: WMO/CBS exchanged scores using radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2010 – July 2011).

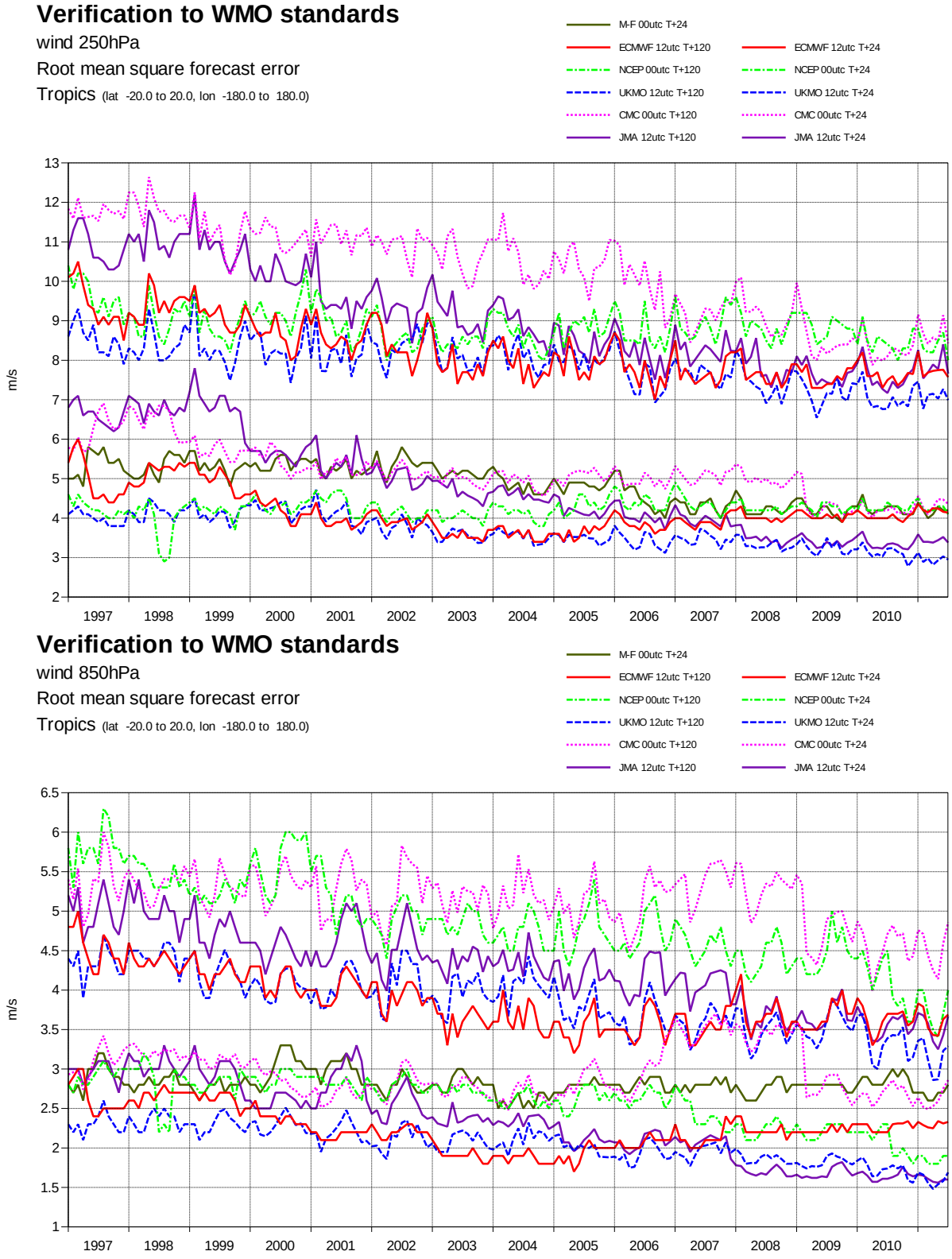


Figure 15: WMO/CBS exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the 5-day forecast error and the lower curves show the 1-day forecast error. Each model is verified against its own analysis.

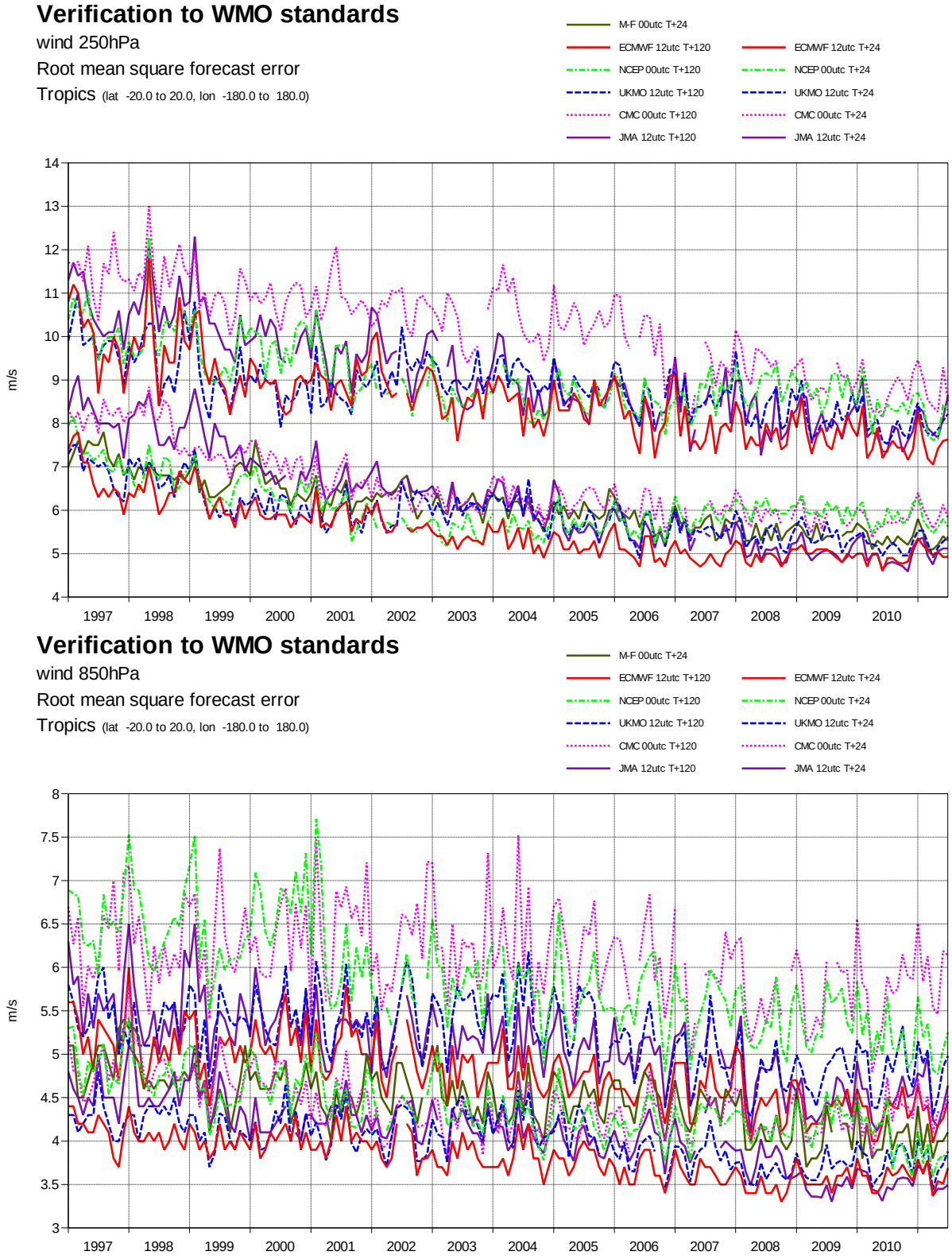


Figure 16: As Figure 15 for scores computed against radiosonde observations.

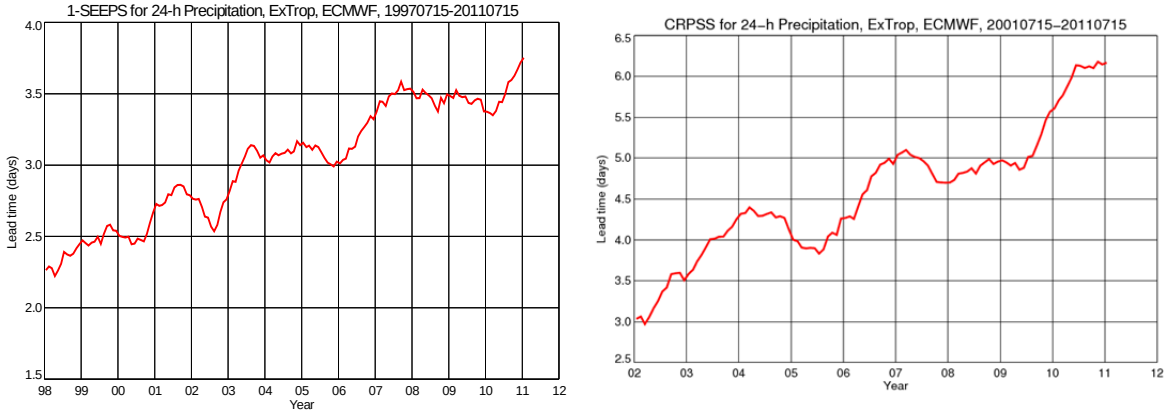


Figure 17: Supplementary headline scores for deterministic (left) and probabilistic (right) precipitation forecasts. Each curve shows the number of days for which the centred 12-month mean skill remains above a specified threshold for precipitation forecasts over the extra-tropics. In both cases the verification is for 24-hour total precipitation verifying against available synoptic observations in the extra-tropics; each point is calculated over a 12-month period, plotted at the centre of the period. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated.

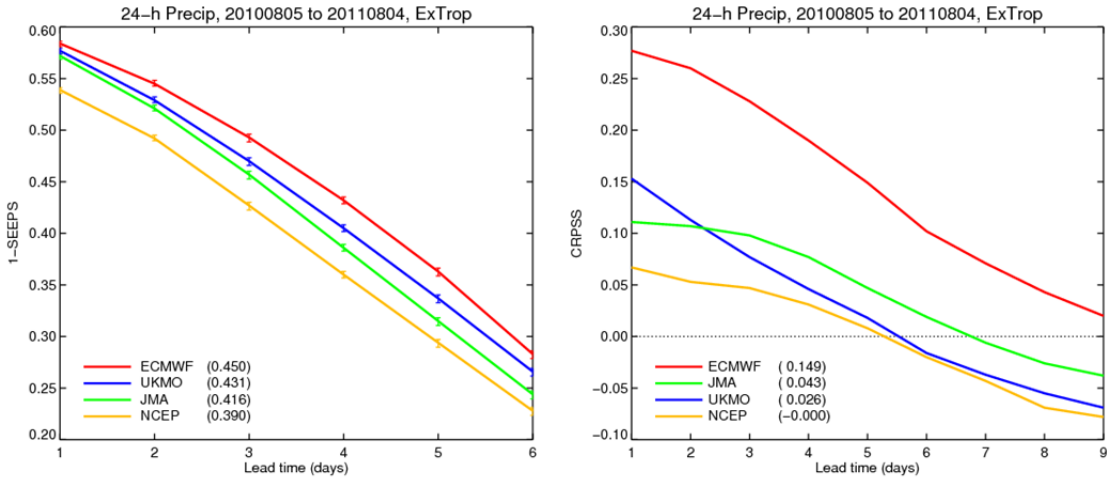


Figure 18: Comparison of precipitation forecast skill for ECMWF, the Met Office (UKMO), Japan Meteorological Agency (JMA) and NCEP using the new supplementary headline scores for precipitation. Left: deterministic; right: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2010-August 2011. Bars indicate 95% confidence intervals for the deterministic skill.

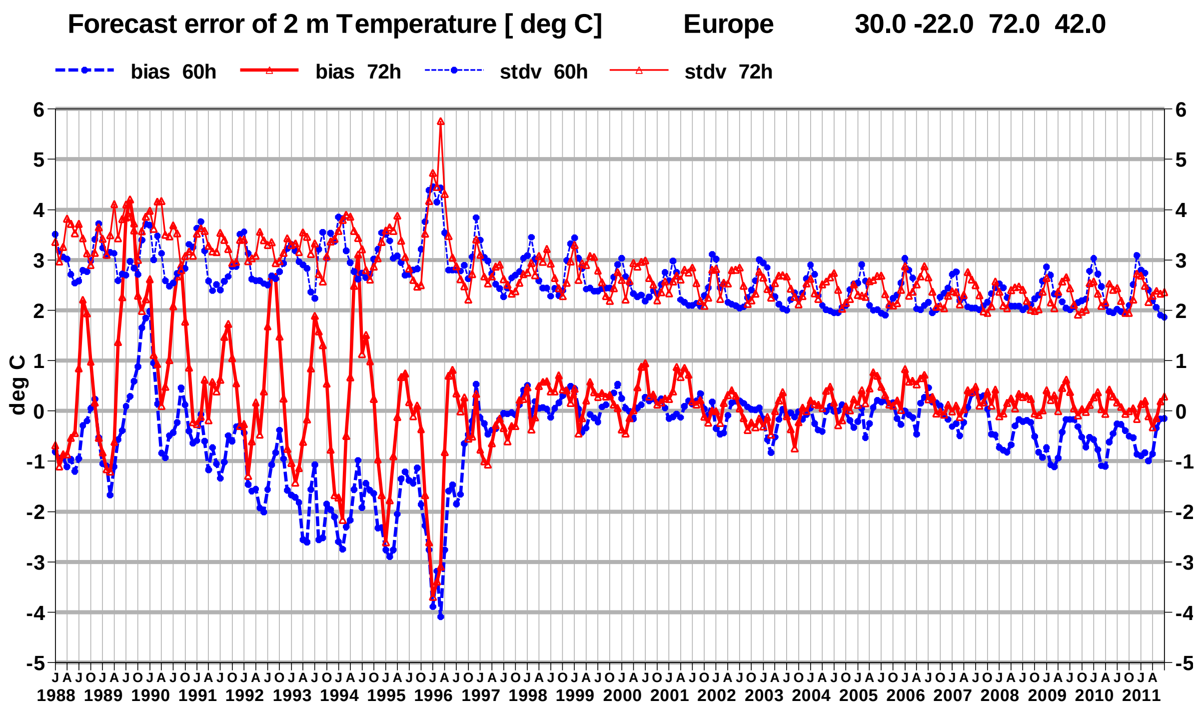


Figure 19: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

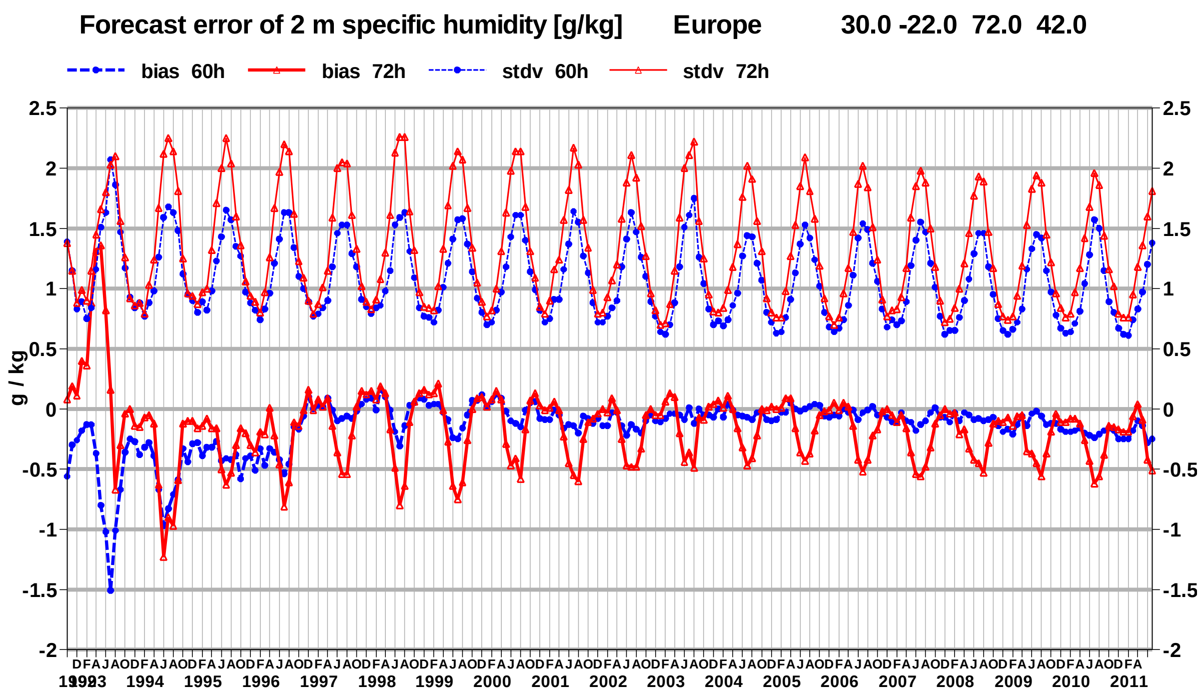


Figure 20: Verification of 2 m specific humidity forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

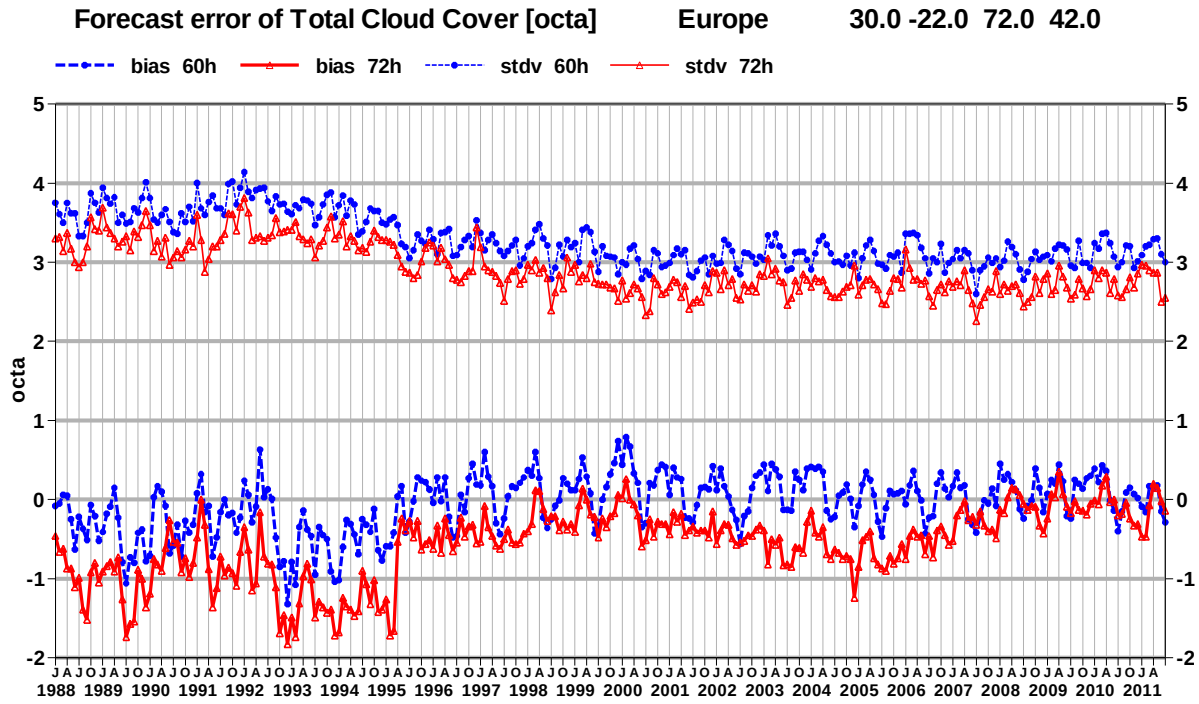


Figure 21: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

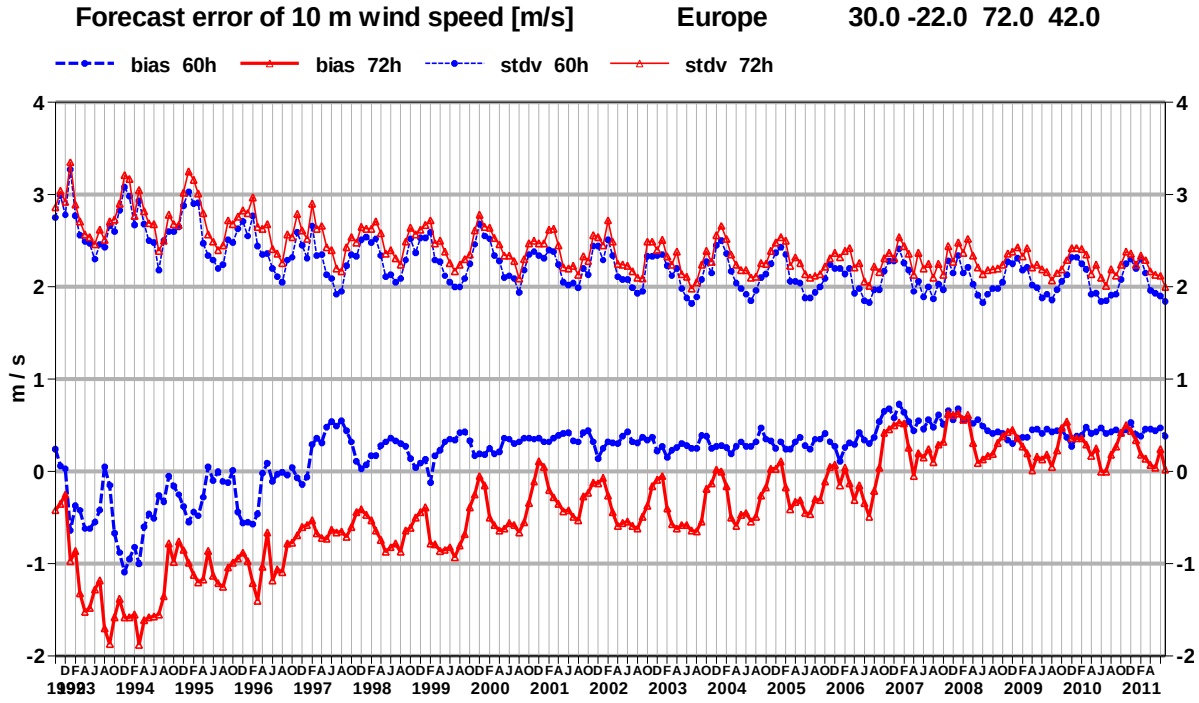


Figure 22: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

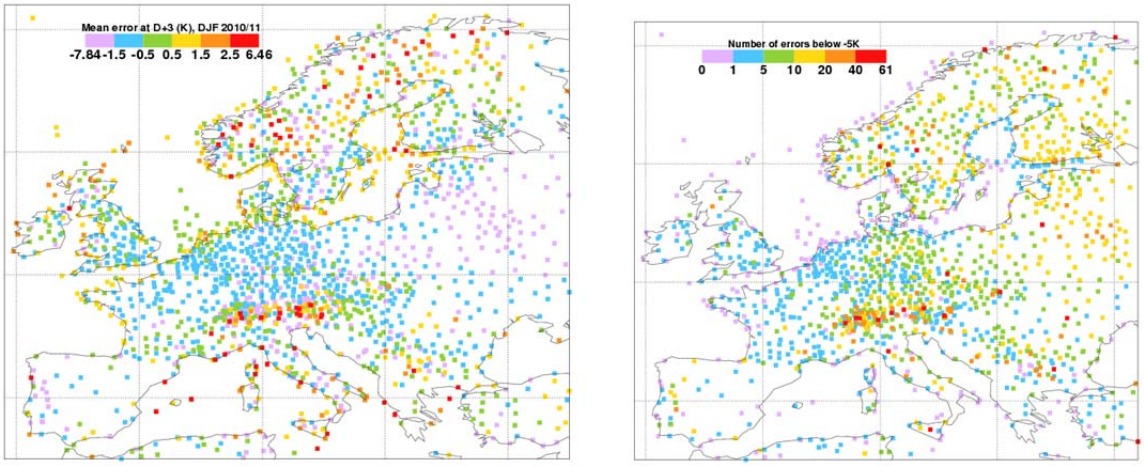


Figure 23: Night-time 2 m temperature errors during winter 2010-11 (December – February). Left: mean 2 m temperature errors at 00 UTC for 3-day forecast. Right: number of occasions with large negative temperature errors (< -5C).

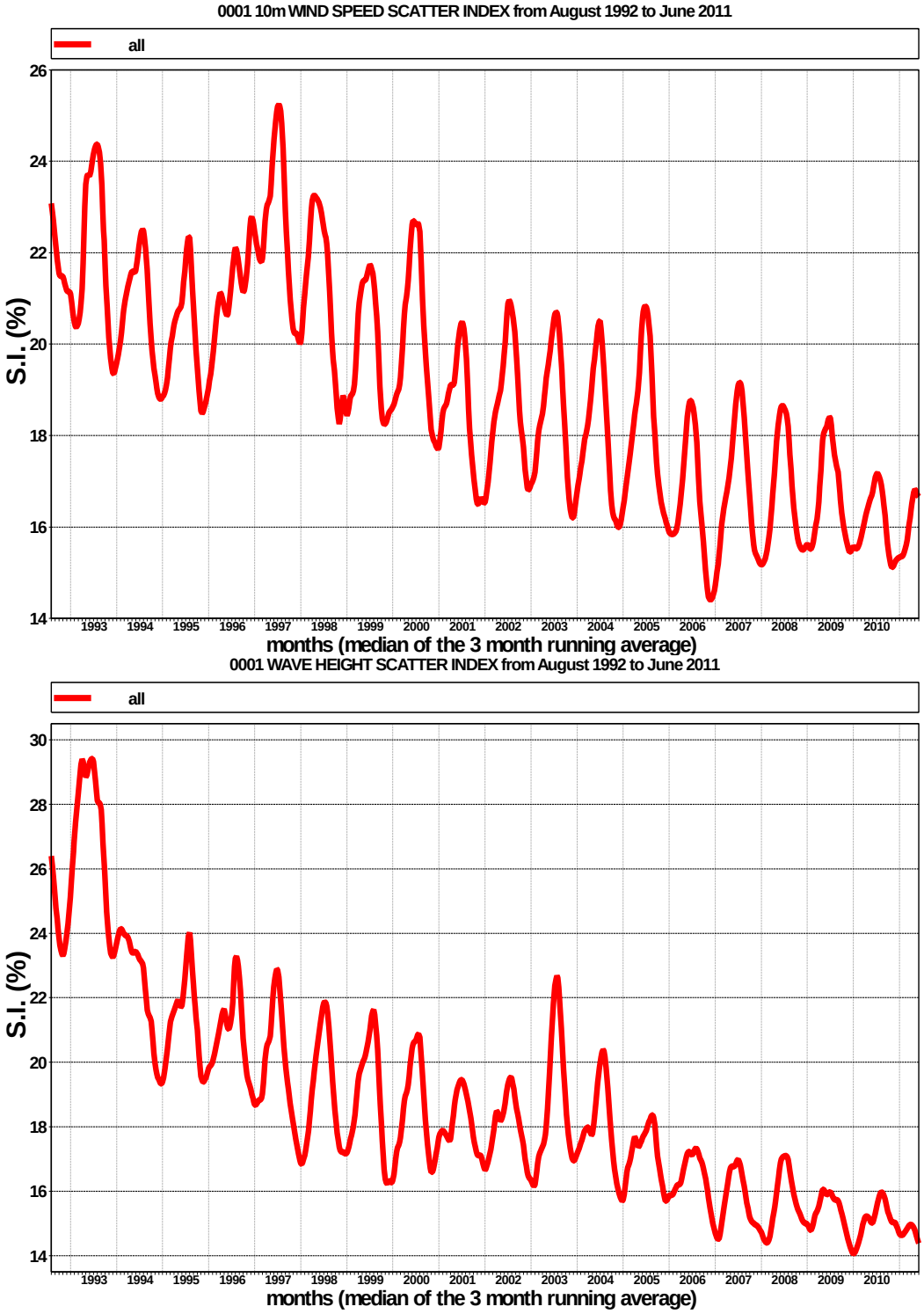


Figure 24: Time series of verification of the ECMWF 10 m wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

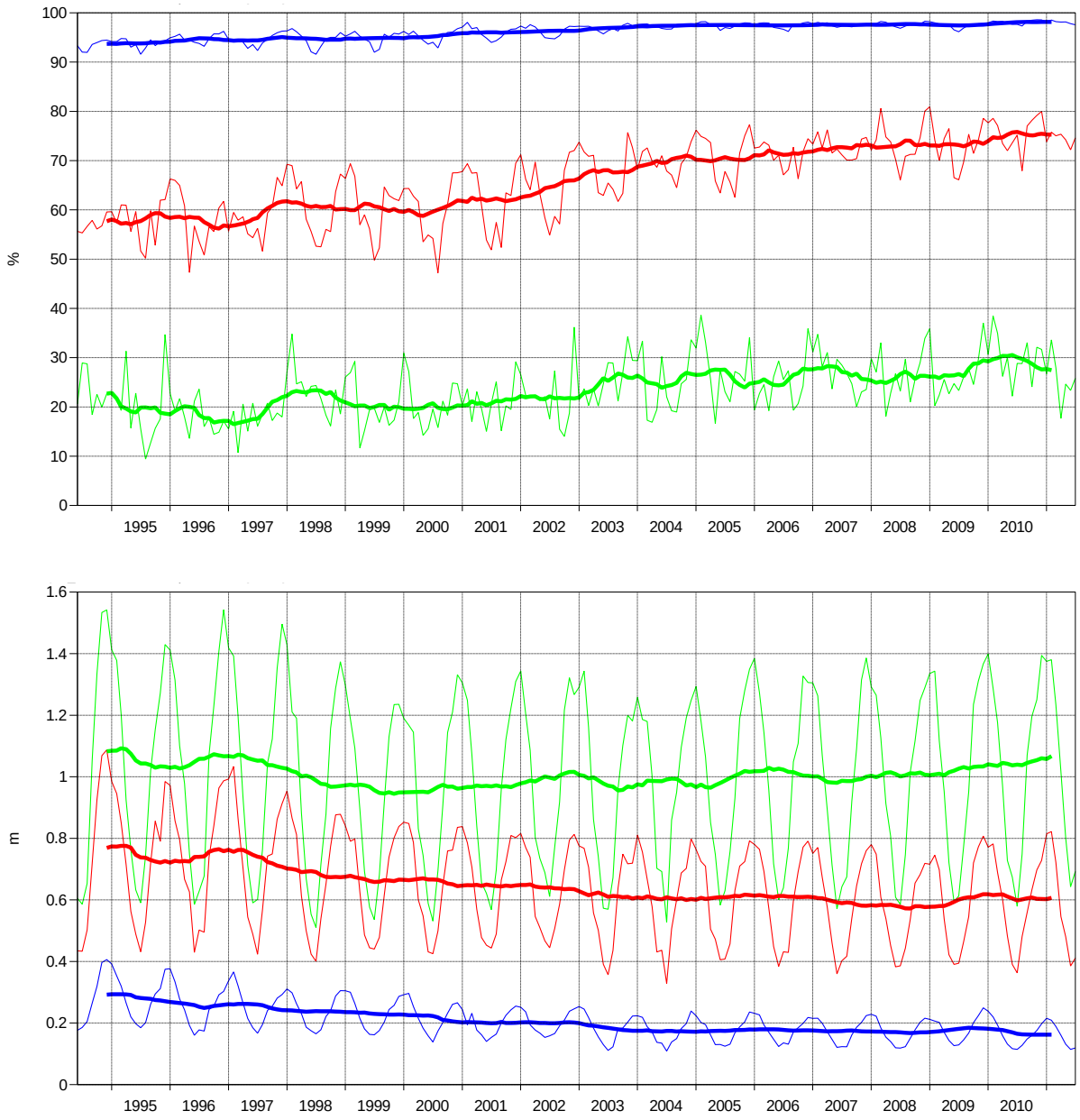


Figure 25: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of anomaly correlation (top) and error standard deviation (bottom) for ocean wave heights verified against analysis for the northern extratropics at day 1 (blue), 5 red) and 10 (green).

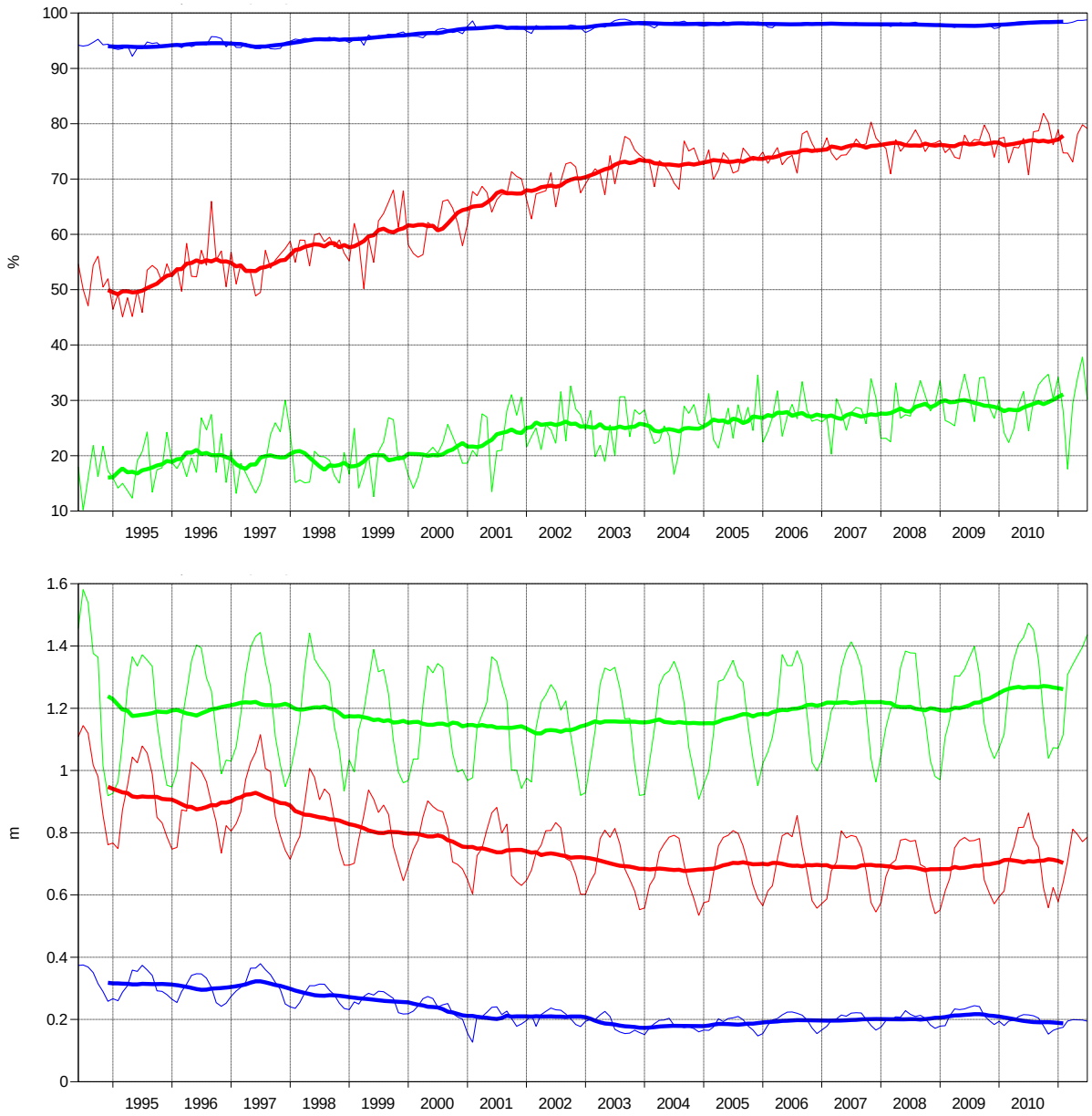


Figure 26: As Figure 25 for the southern hemisphere.

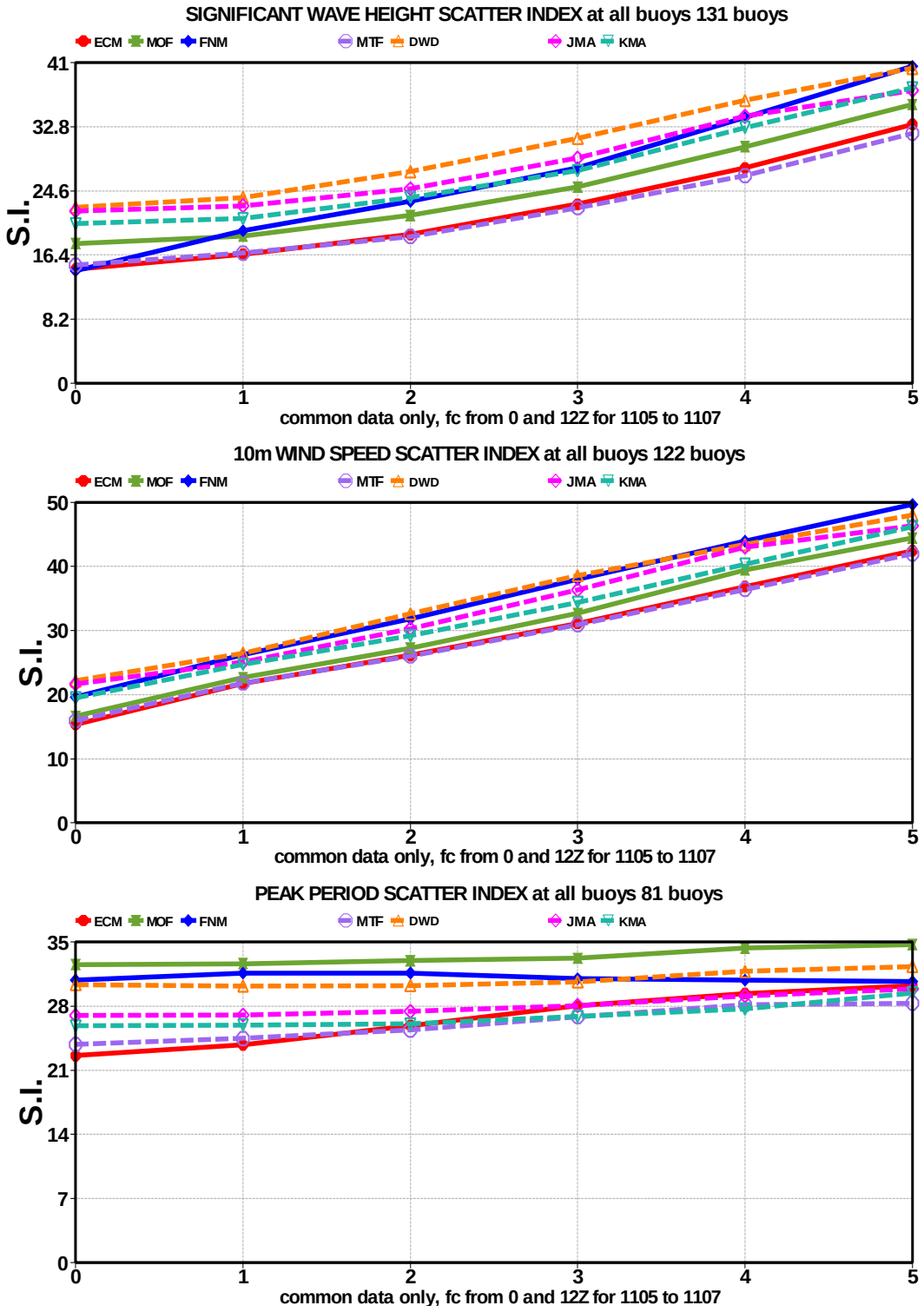


Figure 27: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 3-month period May-July 2011. The x-axis shows the forecast range in days from analysis (step 0) to day 5. MOF: the Met Office, UK; FNM: Fleet Numerical Meteorology and Oceanography Centre, USA; NCP: National Centers for Environmental Prediction, USA; MTF: Météo France; DWD: Deutscher Wetterdienst, BoM: Bureau of Meteorology, Australia; JMA: Japan Meteorological Agency.

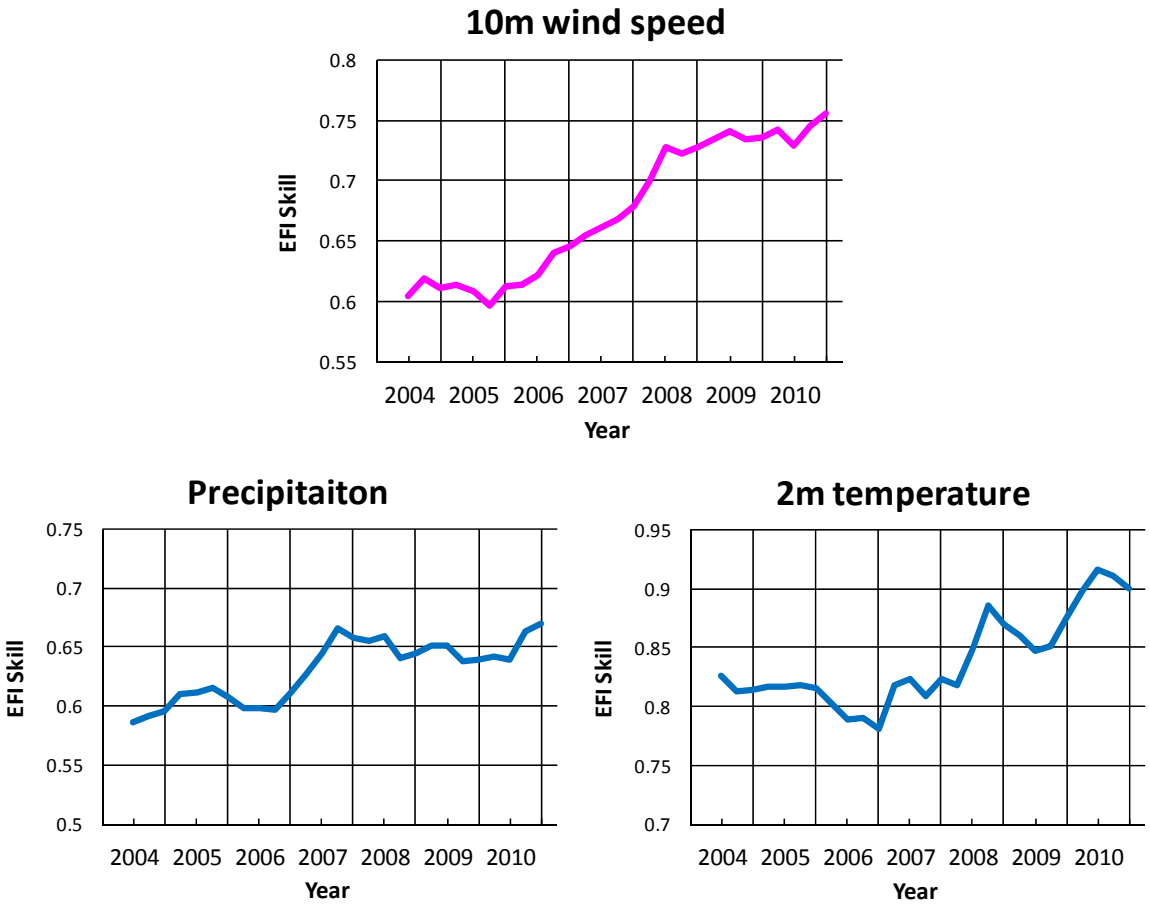


Figure 28: Verification of Extreme Forecast Index (EFI). Top panel: supplementary headline score - skill of the extreme forecast index (EFI) for 10m wind speed at forecast day 4 (24-hour period 72 - 96 hours ahead);an extreme event is taken as an observation exceeding 95th percentile of station climate, curves show a four-season running mean of ROC area skill scores (final point includes spring (March-May) 2011).Bottom panels show the equivalent ROC area skill scores for the precipitation (left) and 2m temperature (right) EFI forecasts.

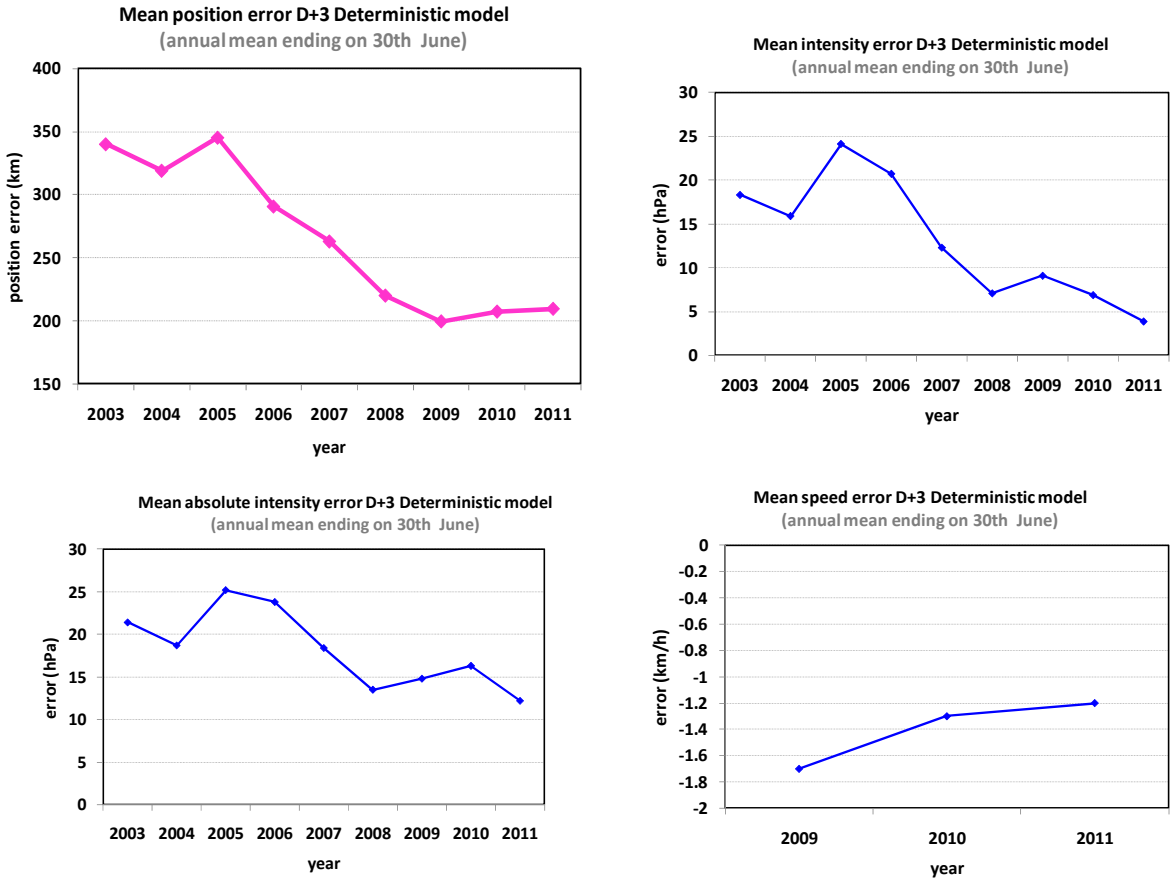


Figure 29: Verification of tropical cyclone predictions from the operational deterministic forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported in real-time via the GTS. Top left: supplementary headline score - the mean position error (km) of the 3-day deterministic forecast. Top right: mean error (bias) in the cyclone intensity (difference between forecast and reported central pressure; positive error indicates the forecast pressure is less deep than observed. Bottom left: mean absolute error of the intensity. Bottom right: mean speed error (negative values indicate the forecast is too slow compared to the observed).

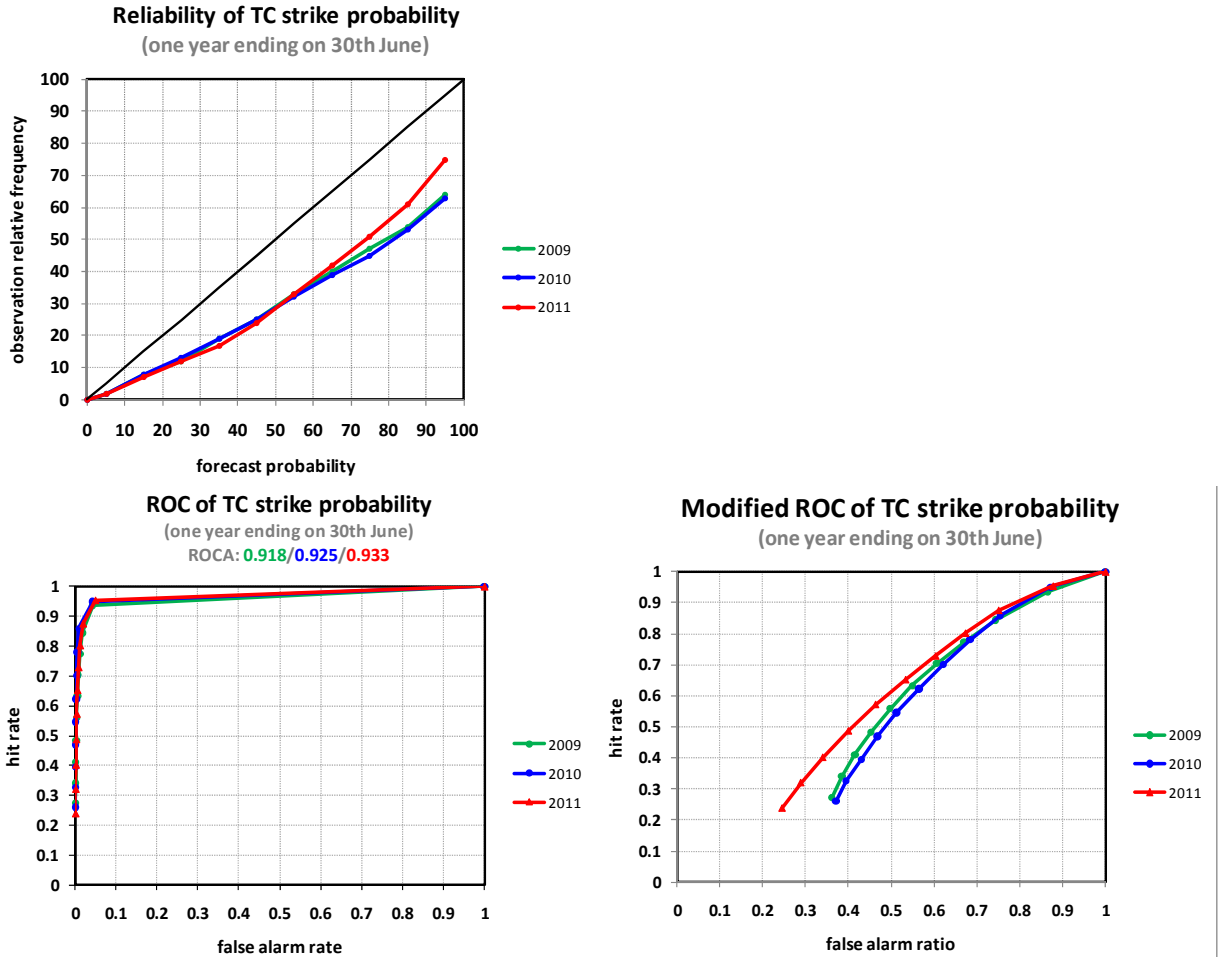


Figure 30: Probabilistic verification of EPS tropical cyclone forecasts for three 12-month periods: July 2008 - June 2009 (green), July 2009-June 2010 (blue) and July 2010-June 2011 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) the ROC diagram and the modified ROC, where the false alarm ratio is used instead of the false alarm rate in the standard ROC. For both ROC and modified ROC, the closer the curve is to the upper left corner, the better (indicating a greater proportion of hits and fewer false alarms).

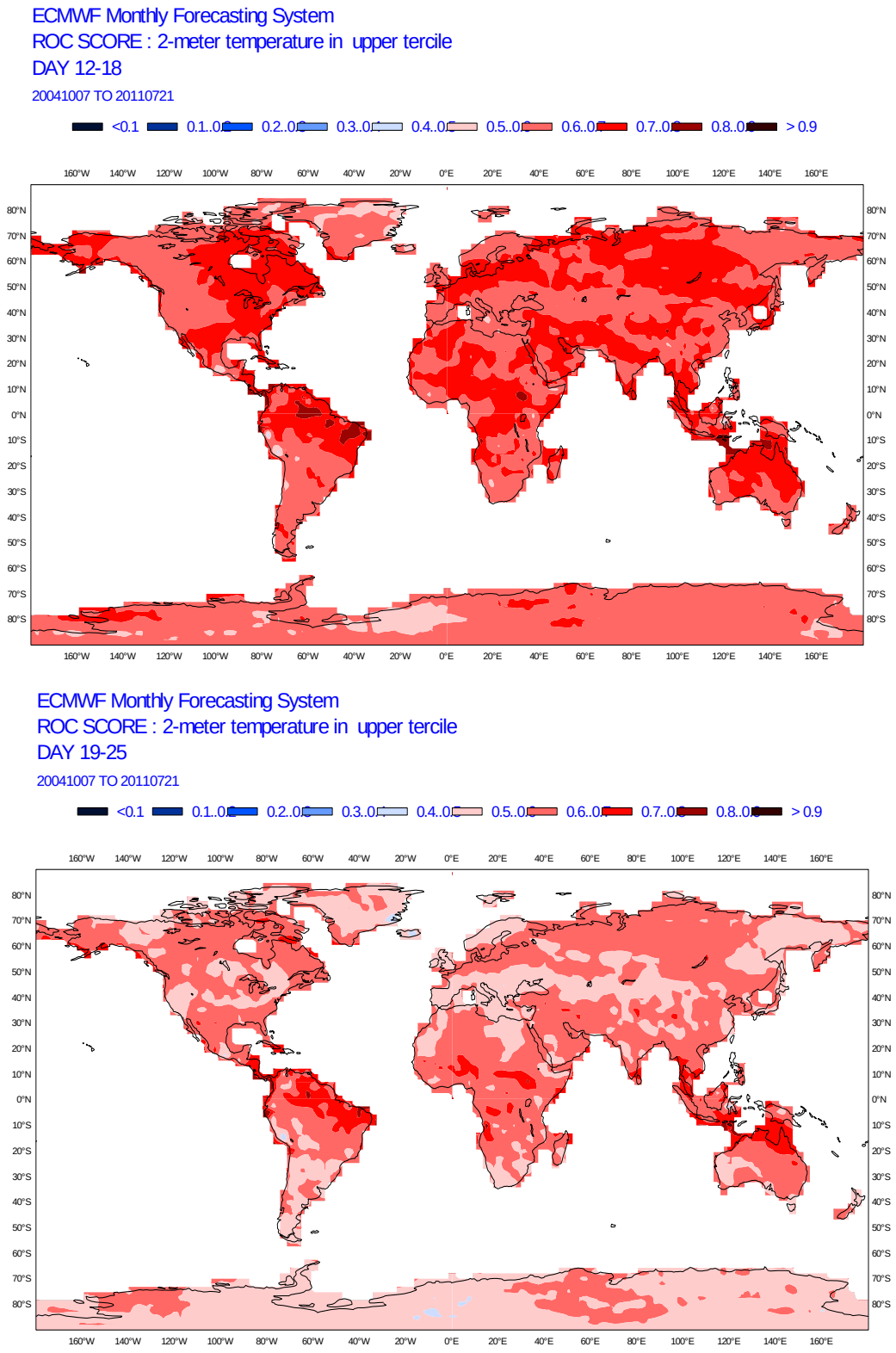


Figure 31: Monthly forecast verification. Spatial distribution of ROC area scores for the probability of 2 m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 21 July 2011 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Stronger red shading indicates higher skill compared to climate.

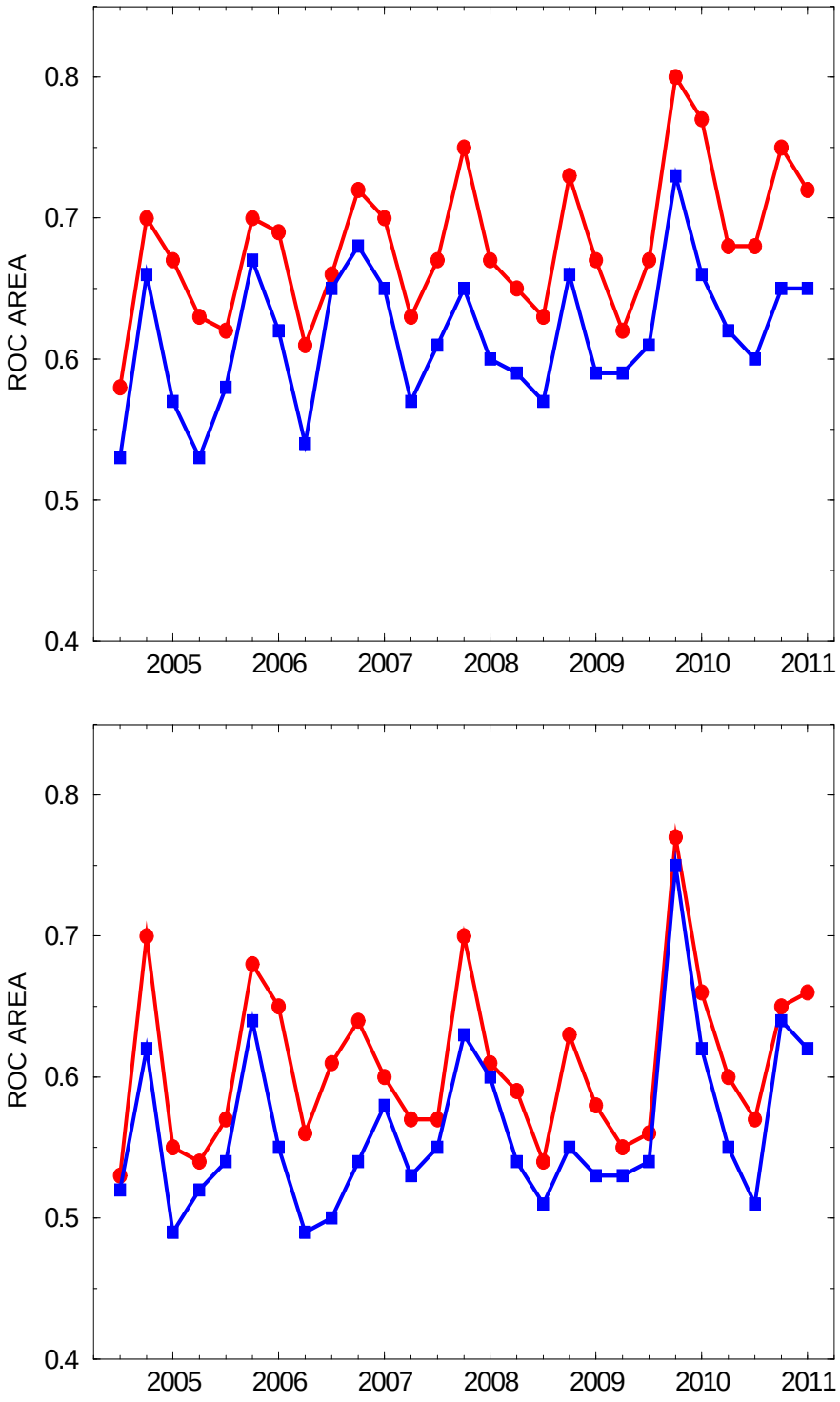


Figure 32: Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution. Scores are calculated for each 3 month season since autumn (September-November) 2004 for all land points in the extra-tropical northern hemisphere. The red line shows the score 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. The last point on each curve is for the spring (March-May) season 2011.

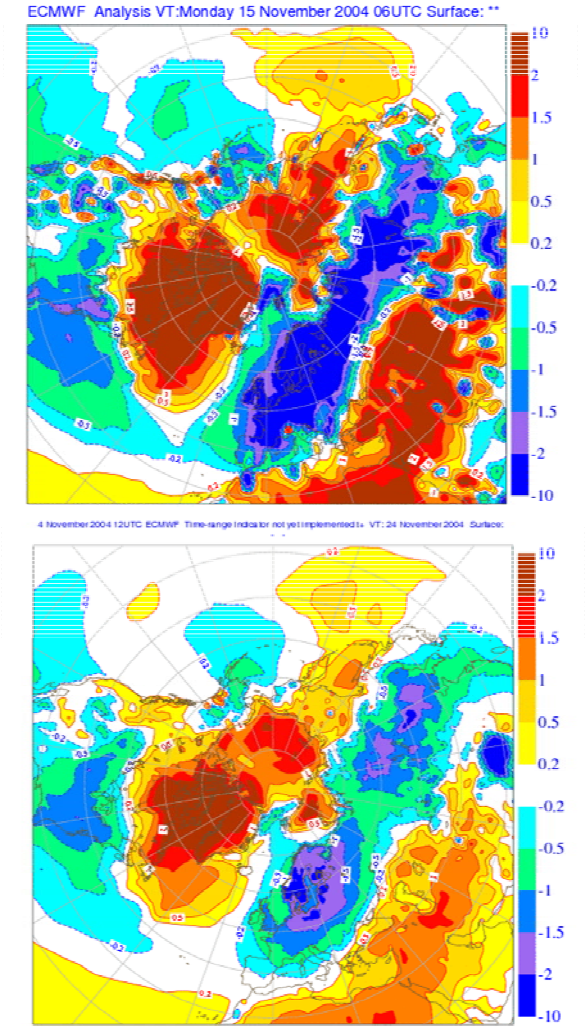


Figure 33: Composite maps of the 2 m temperature anomaly for "cold" cases for the analysis (left) and the 12-18 day ensemble-mean forecast (right). Maps show averages of the weekly-mean temperature anomalies for 35 cases with the strongest signal (in the analysis) for this pattern, associated with the negative NAO circulation regime.

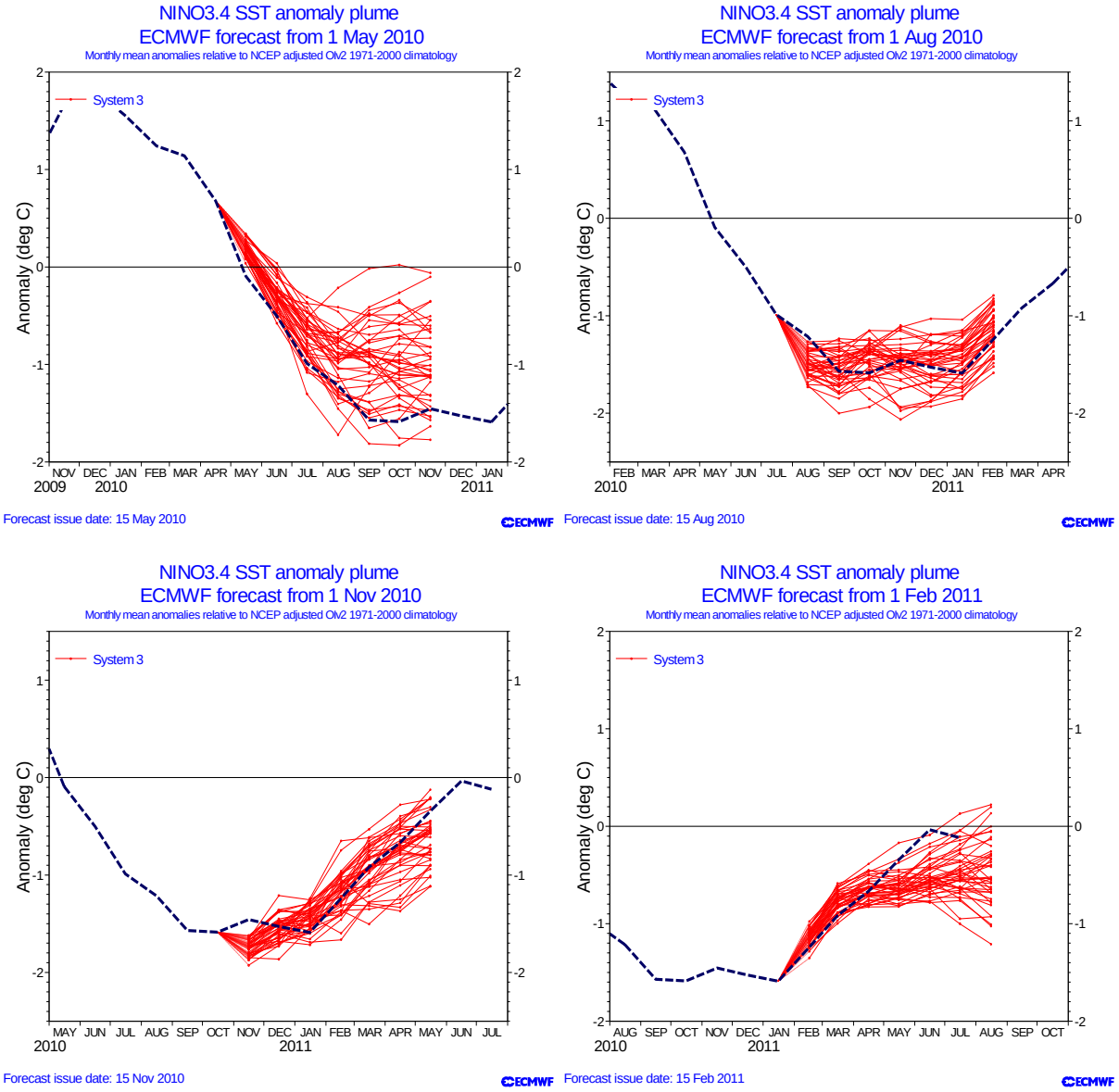


Figure 34: Plot of ECMWF seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from May 2010 (top left), August 2010 (top right), November 2010 (bottom left) and February 2011(bottom right). The red lines represent the ensemble members; dashed blue lines show the subsequent verification.

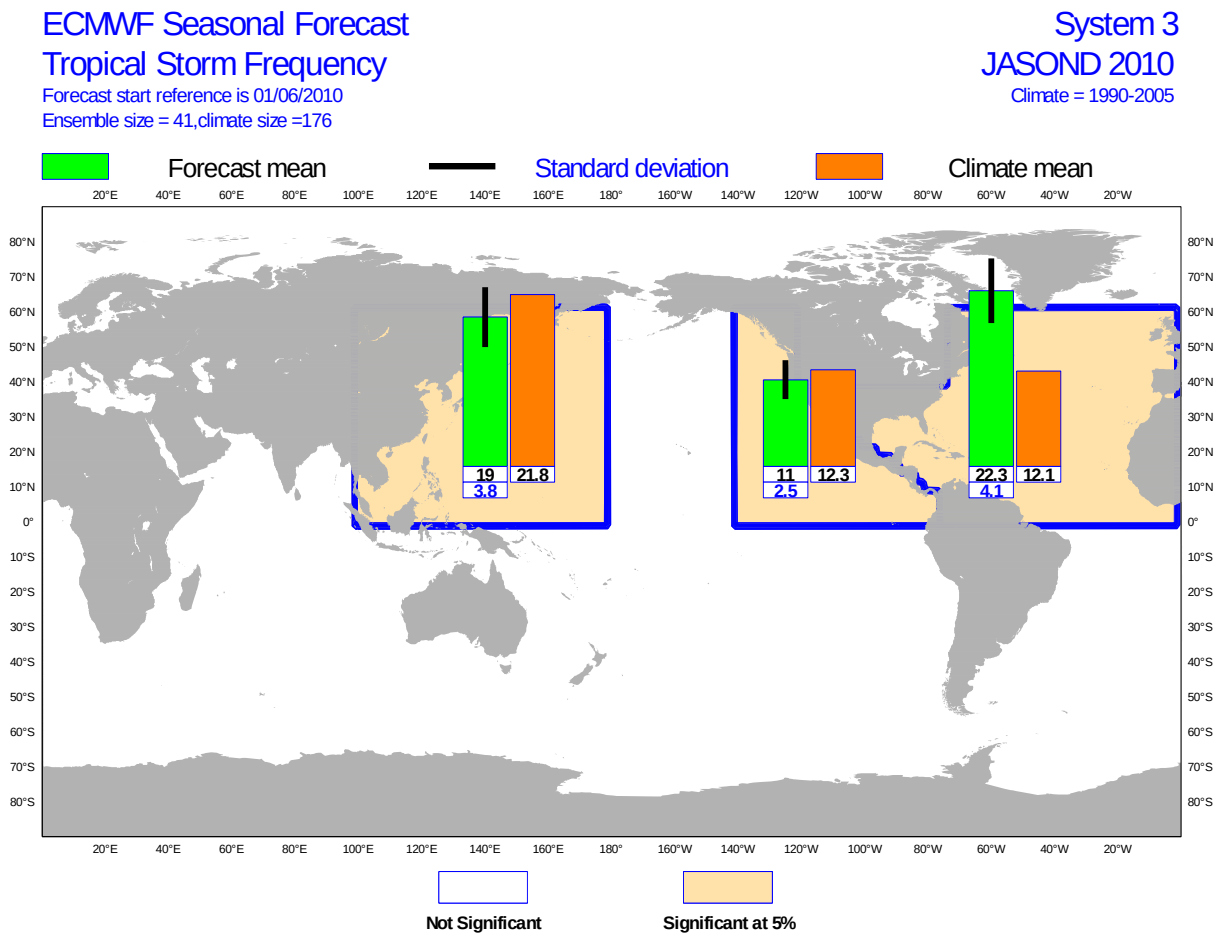


Figure 35: Tropical storm frequency forecast issued in June 2010 for the 6-month period July-December 2010. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The black bars represent ±1 standard deviation within the ensemble distribution; these values are indicated by the blue number. The 41-member ensemble forecast is compared with the climatology. A Wilcoxon-Mann-Whitney (WMW) test is then applied to evaluate if the predicted tropical storm frequencies are significantly different from the climatology. The ocean basins where the WMW test detects significance larger than 90% have a shaded background.

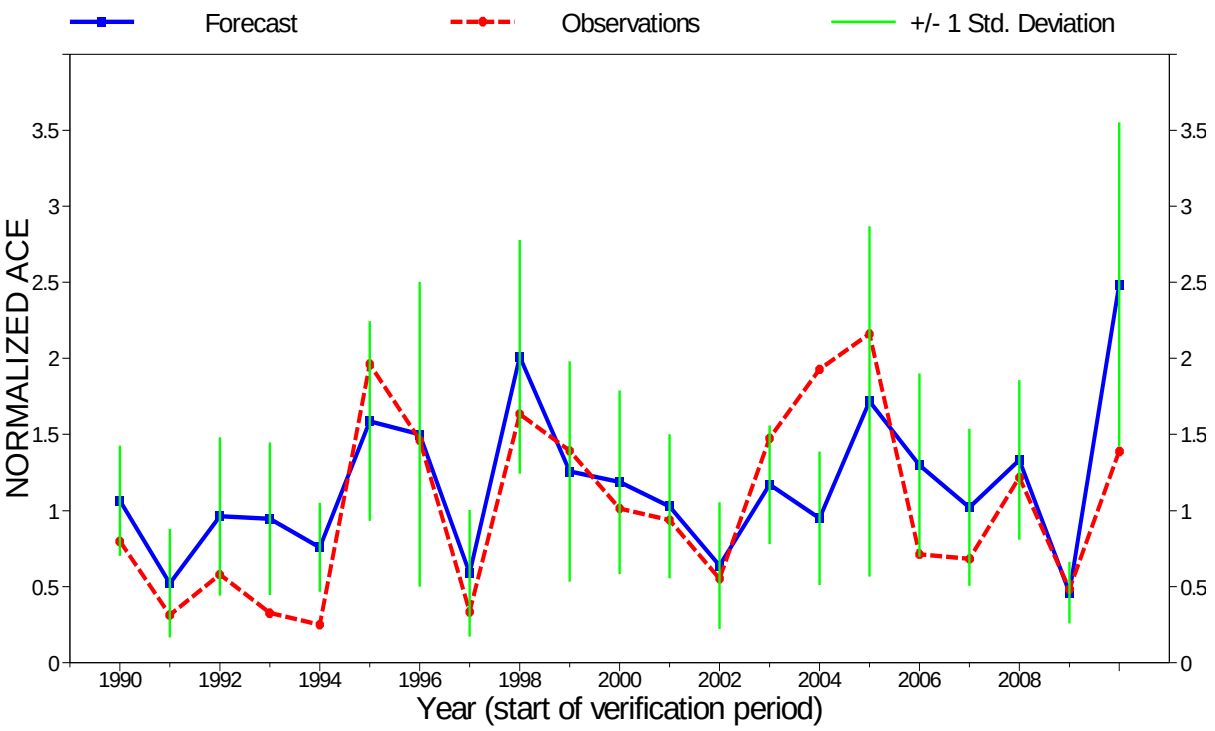


Figure 36: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July-December 1990 to July-December 2010. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation), the red dotted line shows the observation. Forecasts are from ECMWF seasonal forecast system 3: for 1990-2005 these are based on the 11-member re-forecasts; from 2006 onwards they are from the operational 40-member seasonal forecast ensemble. Start date of the forecast is 1 June.

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 1.5 x 1.5 grid (computed from spectral fields with T120 truncation) limited to standard domains (bounding co-ordinates are reproduced in the figure inner captions), as this is the resolution agreed in the updated WMO/CBS recommendations approved by the 16th WMO Congress in 2011. 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 14), the sondes have been selected following an agreement reached by data monitoring centres and published in WMO/WWW Operational Newsletter.

Root Mean Square Errors (RMSE) are the 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 14 and Figure 15) 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 3) 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 1 and Figure 5 show correlations in space between the forecast anomaly and the verifying analysis anomaly. Anomalies with respect to ERA-Interim analysis climate are available at ECMWF from early 1980s. For ocean waves (Figure 25, Figure 26) the climate has been also derived from the ERA-Interim analyses.

A. 2 Probabilistic forecasts

Events for the verification of medium-range probabilistic forecasts are usually defined as anomalies with reference to a suitable climatology. For upper-air parameters, the climate is derived from ERA-Interim analyses for the 20-years period 1989-2008. Probabilistic skill is evaluated in this report using the Continuous Ranked Probability Skill Score (CRPSS) and the area under Relative Operating Characteristic (ROC) curve.

The Continuous Ranked Probability Score (CRPS) is an integral measure that the quality of the forecast probability distribution

$$CRPS = \int_{-\infty}^{\infty} \left[P_f(x) - P_a(x) \right]^2 dx$$

where P_f is forecast probability cumulative distribution function (CDF) and P_a is analysed value expressed as a CDF. CRPS is computed discretely following Hersbach, 2000. CRPSS is then computed as

$$CRPSS = 1 - \frac{CRPS}{CRPS_{clim}}$$

where CRPSclim is the CRPS of a climate forecast (either ERA-Interim analysis-based or observation-based climatology). CRPSS is used to measure the long-term evolution of skill of ECMWF Ensemble Forecasting System (Figure 8) and its inter-annual variability (Figure 10). 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 30). Figure 30 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 32.

A.3 Weather parameters (Section 4)

Verification of the deterministic precipitation forecasts is made using the newly developed 'SEEPS' score (Rodwell et al., 2010). SEEPS (Stable Equitable Error in Probability Space) uses three categories: 'dry', 'light precipitation' and 'heavy precipitation'. Here 'dry' is defined, with reference to WMO guidelines for observation reporting, to be any accumulation (rounded to the nearest 0.1 mm) that is less than or equal to 0.2 mm. To ensure that the score is applicable for any climatic region, the 'light' and 'heavy' categories are defined by the local climatology so that 'light' precipitation occurs twice as often as 'heavy' precipitation. A global 30-year climatology of SYNOP station observations is used (the resulting threshold between the 'light' and 'heavy' categories is generally between 3 and 15 mm for Europe, depending on location and month). SEEPS is used to compare 24-hour accumulations derived from global SYNOP observations (exchanged over the Global Telecommunication System; GTS) with values at the nearest model grid-point. 1-SEEPS is used for presentational purposes (Figure 17, Figure 18) as this provides a positively-oriented skill score.

The EPS precipitation forecasts are evaluated with the CRPSS (Figure 17, Figure 18). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 19-Figure 22), 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 four closest grid points, provided the difference between the model and true orography is less than 500 m. A crude quality control is applied to SYNOP data (maximum departure from the model forecast has to be less than 25 K, 20 g/kg or 15 m/s for temperature, specific humidity and wind speed respectively). 2 m temperatures are corrected for model/true orography differences, using a crude constant lapse rate assumption, provided the correction is less than 4 K amplitude (data are otherwise rejected).

References

Hersbach, H., 2000*:* Decomposition of the Continous Ranked Probability Score for Ensemble Prediction System. *Wea. Forecasting,* **15,** 559-570.

Rodwell, M.J., D.S. Richardson, T.D. Hewson & T. Haiden, 2010: A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. R. Meteorol. Soc*., **136**, 1344–1363.