TECHNICAL MEMORANDUM

635

Verification statistics and evaluations of ECMWF forecasts in 2009-2010

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

Operations Department

October 2010

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. A short technical note describing the scores used in this report is given in the annex to this document.

In order to aid comparison from year to year, the set of verification scores shown here is mainly consistent with that of previous years (ECMWF Tech. Memos. 346, 414, 432, 463, 501, 504, 547, 578, 606).

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 Committee's last session are summarised below.

8 September 2009: Cycle 35r3, including the following main changes:

- Assimilation of cloud-affected radiances for infra-red instruments
- Improved assimilation of satellite channels that are sensitive to the land surface
- Assimilation of total column water vapour data from the MERIS instrument (over land)
- Variational bias correction for ozone satellite data
- Improved quality control (using Huber norm) of conventional observations
- Improved background-error statistics for humidity, new humidity formulation in 4D-Var
- Weak-constraint 4D-Var taking into account systematic model errors in the stratosphere
- Non-orographic gravity wave scheme
- New trace gas climatology
- Further revision of the snow scheme

- Wave damping in wind input source term for ocean waves
- Revised stochastic physics (model perturbations) for EPS

26 January 2010: Cycle 36r1, including major increase in horizontal resolution for the deterministic and EPS forecast systems:

- Deterministic forecast and analysis horizontal resolution increased from T799 (25 km) to T1279 (16 km).
- EPS horizontal resolution increased from T399 (50 km) to T639 (32 km) to day 10 and from T255 (80 km) to T3l9 (63 km) beyond day 10.
- Global wave model resolution increased from 0.36 to 0.25 degrees in the deterministic model and from 1.0 to 0.5 degrees for the EPS.
- Correction of short-wave radiation interaction with clouds.

The following corrections to the handling of land surface parameters were also implemented with cycle 36r1:

- Re-activation of use of NESDIS satellite snow cover product in snow analysis
- Correction of snow density update in the presence of fresh snow
- Correction to daily update of MODIS-based monthly albedo.

22 June 2010: Cycle 36r2, revised initial perturbations for the EPS: differences between members of an ensemble of data assimilations (EDA) are used instead of the evolved singular vectors to create initial spread between EPS forecast members. Initial-time singular vectors continue to be used in conjunction with the EDA perturbations.

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 hemisphere 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 2010. For both regions skill has been consistently high, reaching new record levels relative to persistence. Figure 2 shows the evolution of performance using the anomaly correlation (ACC), where reference is to climatology instead of persistence. In February 2010, a new landmark was achieved when the average score remained above 60% throughout the 10-day forecast range for both the European region and the extra-tropical northern hemisphere as a whole. The ACC was 67% at day 10 for the northern hemisphere and 61% for Europe.

These are the highest scores ever reached by the forecasting system. The 2009-10 winter season was unusual over the northern hemisphere with a strong negative phase of the North Atlantic Oscillation and Arctic Oscillation circulation patterns. These are typically associated with cold weather in northern Europe and more active weather systems and heavy rain fall affecting south west Europe. The exceptional scores are partially a result of the large anomalies this winter. However, the high scores for both the ACC and the skill relative to persistence confirm that the ECMWF forecast system performed consistently well in predicting these anomalous weather conditions. Figure 3 shows that overall synoptic activity has reduced from the very high levels of the previous 12 months over Europe. As noted above, the somewhat more stable flow has partially contributed to the higher scores compared to last year.

Figure 4 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 also apparent for the 850 temperature scores, with a greater fraction of the individual forecasts achieving very high ACC scores than in previous years. Summer 2010 scores also show a good performance with relatively few occasions of moderate or poor skill at the 7-day range.

Figure 5 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 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 Europe and the northern hemisphere. As for the deterministic forecast, the EPS skill reached record levels in winter 2009-10. Over Europe in particular, these very high scores compared to previous years have been maintained throughout 2010. A number of changes have been made to the EPS over the past year, including improvements to both the initial perturbations and representation of model uncertainties and the increase in resolution (see Section 2). Although the high skill is partly influenced by the strongly anomalous circulation pattern, it is also consistent with the improvements from these model changes. However, it is too early to confirm unambiguously the long-term impact of these changes on the overall skill levels of the EPS.

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 match between the spread and error is similar for the three winter seasons. There is a small over-dispersion of the EPS for 500 hPa height in the early forecast range and some under-dispersion at longer ranges. In

general the EPS is 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 recent introduction of the ensemble of data assimilations (EDA) in the initial perturbations and the planned changes to the representation of model uncertainty are expected to improve the overall dispersion of the EPS for both parameters.

Figure 9 shows the skill of the EPS using 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 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 the earlier years. In part, as for the deterministic forecast, the anomalous flow made some contribution to the high scores.

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 10. 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 two 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. Figure 11 shows time series of such scores over the northern extratropics for both 500 hPa geopotential height and mean sea level pressure (MSLP). For both parameters, medium-range forecast errors for all models were lower in winter 2009-10 than in winter 2008-09. ECMWF continues to maintain its lead over the other centres. Overall, however, the difference in performance between centres is decreasing; in particular the Canadian forecast errors have recently reduced substantially and are now more in line with those of the other global centres. In general, the ECMWF lead has been greater in the southern hemisphere extratropics (Figure 12). However, improvements in the UK Met Office forecasts have reduced the overall gap compared to previous years.

WMO exchanged scores also include verification against radiosondes over regions such as Europe. Figure 13, showing both 500 hPa geopotential 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 comparison for the tropics is summarised in Figure 14 (verification against analyses) and Figure 15 (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 Met Office performances are similar. The errors of the JMA forecast system have steadily decreased over several years and are now comparable with those of the Met Office model at both short and medium ranges. In the tropics, verification against analyses (Figure 14) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations, the ECMWF, Met Office and JMA models have very similar short-range errors.

The large increase in 850 hPa wind error against analysis in the Canadian forecasts from 2006 and sudden drop in early 2009 (Figure 14) are related to the verification procedure and do not reflect differences in model performance. This does, however, demonstrate the importance of consistent verification methodology, when comparing forecasts from different centres. This matter is being addressed by a new WMO CBS Co-ordination Group on forecast verification that is reviewing the current procedures used for these WMO standard scores; the group will report its findings and proposals to the CBS in autumn 2010.

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 2 m temperature, specific humidity, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 16 to Figure 19. 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.

Winter 2009-10 had significant negative night-time temperature bias over Europe, similar to that of the past two winters (Figure 16). Large temperature errors were particularly associated with discrepancies in snow cover in the model. On some occasions, after significant snowfall events, snow remained too long on the ground in the model analysis. This led to substantial temperature errors in the subsequent ECMWF forecasts. The problem of excessive snow in the model analysis became particularly apparent to users in March, as the spring snow-melt began. The main impact on users was that 2 m temperature forecasts were consistently too cold in areas where snow remained in the model after they had, in reality, become snow-free. A number of modifications have been made to the snow analysis to address these problems; a new snow analysis scheme is being tested as part of the next model cycle (the issues and modifications are explained in more detail in the document on the snow analysis, ECMWF/TAC/42(10)12).

Time series of precipitation skill for Europe is shown in Figure 20, using the True Skill Score (or Pierce's Skill Score) for thresholds of 10 mm and 20 mm per day. As noted in previous reports, there has been a consistent improvement since the introduction of cycle 31r1 in September 2006. For both 10 mm/day and the higher threshold of 20 mm/day, the skill in 2009-10 was similar to that for 2008-09: below the exceptional performance in 2007-08 but consistently higher than in previous years. The same overall trend can be seen in the scores for the EPS probability forecasts shown in Figure 21 for a range of precipitation thresholds at day 4.

4.2. Ocean waves

The quality of the ocean wave model analysis is shown in the comparison with independent ocean buoy observations in Figure 22. In general the errors in the analysis have decreased in 2009-10, compared to previous years. The top panel of Figure 22 shows a time series of the analysis error for the 10 m wind over maritime regions using the wind observations from the same set of buoys. The error has steadily decreased since 1997, providing better quality winds for the forcing of the ocean wave model and this year has been similar to last year.

The good performance of the wave model forecasts is confirmed this year, as shown in Figure 23 and Figure 24. This is particularly noticeable in the verification against observations and comparison with other wave models, as shown in Figure 25. 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). The ECMWF forecast winds are used to drive the wave models of Météo-France and SHOM (Service Hydrographique et Océanographique de la Marine, France); the wave models of these centres are also similar, hence the closeness of the errors of the three centres. 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 web site, including the figures shown in this report: http://www.ecmwf.int/products/forecasts/wavecharts/

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. 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). Results for precipitation, 10 m wind speed and 2 m temperature are presented in Figure 26 for each season from winter 2004-05 to spring 2010. For all parameters, there is a clear improvement during this period, especially for the 5-day forecast.

5.2. Tropical cyclones

After a very active North Atlantic hurricane season in 2008, the 2009 season was below normal, with nine tropical storms, including three hurricanes. This was well predicted by the seasonal forecast system (see Section 6.3, Figure 33).

Average position and intensity errors for the deterministic medium-range forecasts of all tropical cyclones (all ocean basins) over the last seven 12-month periods are shown in Figure 27. A significant reduction in both position and intensity errors was reported for 2007-08, compared with the previous periods. This improved performance is confirmed for 2008-09 and the most recent period. Both the position errors (top right panel of Figure 27) and the mean intensity errors (bottom left panel) are very similar for the last three years. The mean absolute error of the TC intensity has increased somewhat for 2009-10 (bottom right panel of Figure 27), although there is a relatively large uncertainty in these scores as indicated by the bars in the figure.

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 28. Results show an over-confidence 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 the 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 new, 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 to 15 for the forecast started from the 00 UTC analysis, on a daily basis.

Figure 29 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). 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 provide the user 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/.

6.1.1. Monthly forecast performance 2009-2010

Figure 30 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.

The monthly forecast system has continued to perform well. The exceptionally high scores reached last winter for forecast ranges 12-18 and 19-32 days are very evident in Figure 30, consistent with the exceptional scores noted for the medium-range deterministic (Figure 2) and EPS (Figure 7) forecasts. The scores for the persistence forecast (blue curves) were also the very high, showing the influence of the atmospheric flow (very persistent negative North Atlantic Oscillation (NAO) conditions) on the performance measures. Despite such high values for persistence in winter 2009-10, comparing the two curves shows that the monthly forecast system outperformed this reference by a similar margin to that achieved in other seasons. The monthly forecast score for spring 2010 was also exceptionally high compared to previous spring seasons.

In early January 2010 temperatures were substantially colder than normal across the whole of northern and western Europe. This was part of a wider pattern of very cold temperatures across much of the northern hemisphere. The onset of the large-scale pattern was predicted by the monthly system between two and three weeks in advance.

6.2. The 2009-2010 El Niño forecasts

The 2009/2010 El Niño event was of moderate intensity. Warm conditions developed during 2009; the El Niño peaked in December 2009 and dissipated quickly by May 2010. This was followed by a transition to cooler conditions and has currently reached a weak La Niña situation. The annual-range outlook from November 2008 gave an early indication of the El Niño development (Figure 31). The development and decline of the El Niño was consistently well forecast by the seasonal forecast system (Figure 32), including the forecasts from EUROSIP partners. Due to the seasonality of the predictive skill, the uncertainty in the amplitude of the El Niño was particularly large in the March and April forecasts; however the observed values remained within the predicted range. The large spread of EUROSIP in the first month is because the upgraded UK Met-Office GLOsea4 model is based on a 'lagged start' approach, combined with stochastic perturbations generated during the model integration (both stochastic backscatter and perturbed physics) to generate ensemble members; the initial conditions span an average period of a month.

6.3. Tropical storm predictions from the seasonal forecasts

The seasonal forecast predictions for the 2009 Atlantic tropical storm season verified well: a range of storms between 6 to 10 was forecast and 9 were observed (Figure 33). Over the Western Pacific, seasonal predictions indicated a slightly enhanced activity while a slightly below-normal activity was observed; however the forecast signal did not reach the 90% significance level (basin not shaded on map, Figure 33). The forecast for the Eastern Pacific did give a significant signal for reduced tropical storm frequency; this was not correct, as the season was in fact more active than normal. Evaluation using the seasonal re-forecasts confirmed that the skill of the tropical storm forecast in the Eastern Pacific was very low. This contrasts with the performance for the Atlantic basin: Figure 34 shows the skill in predicting ACE (Accumulated Cyclone Energy) over the Atlantic basin calculated using the most recent 20 years is substantial, with a correlation between ensemble mean forecast and observation of 0.73.

Since April 2010 the seasonal tropical storm predictions have indicated enhanced activity over the Atlantic. This is consistent with the forecast of the transition from El Niño to La Niña conditions: La Niña typically contributes to increased Atlantic hurricane activity by decreasing the vertical wind shear over the Caribbean Sea and tropical Atlantic Ocean. However the early part of the 2010 Atlantic season has had slightly below average activity.

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 presented alongside the forecast products on the ECMWF web site, 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.

The seasonal forecast system provided a good prediction of the atmospheric anomalies for 500 hPa height and 2 m temperature over the Pacific region during the 2009-10 winter, consistent with the good forecasts for the El Niño. During the same period the Atlantic region was dominated by a strong negative phase of the North Atlantic Oscillation (NAO), with associated cold temperatures over northern and western Europe. This was not well predicted beyond the first month of the seasonal forecast.

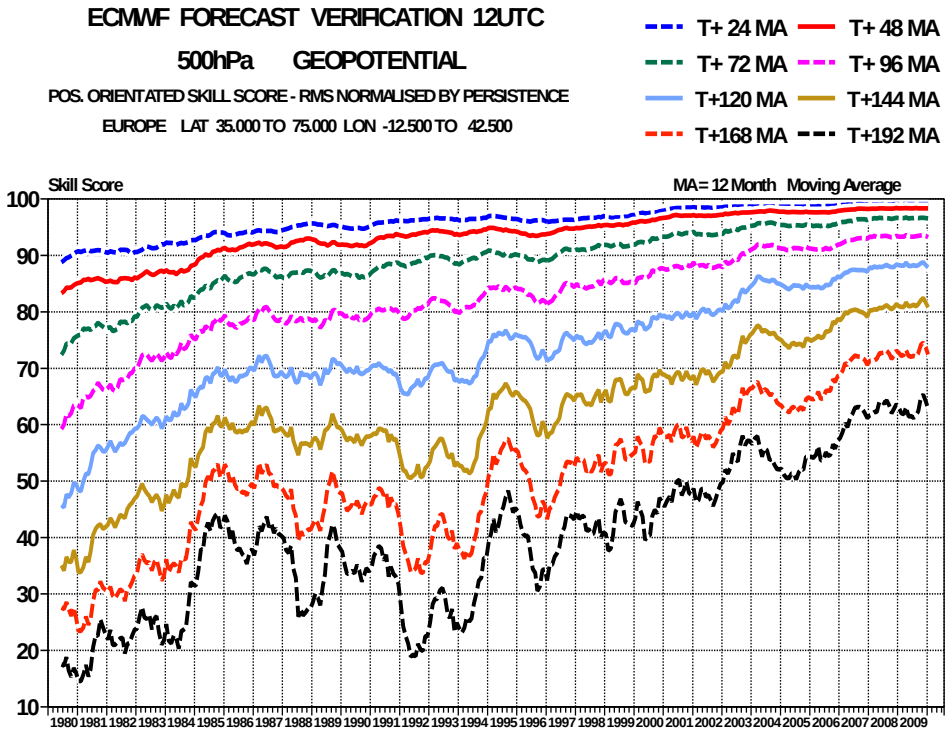
Results from a number of sensitivity experiments indicate slightly better forecasts when the ECMWF model is forced by observed SSTs, or when the Tropics are relaxed towards observations. However the amplitude of the 2009-10 winter anomalies was not well captured in either case. Model results also showed that there is no significant impact when the stratosphere is relaxed towards observations. Given that a dominant source of predictability for this very unusual winter has not been found, it may be that the unusual circulation during the winter of 2009-10 was a result of the internal (extratropical) atmospheric dynamics, making it difficult for seasonal forecasting systems to predict the onset of the negative phase of the NAO.

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



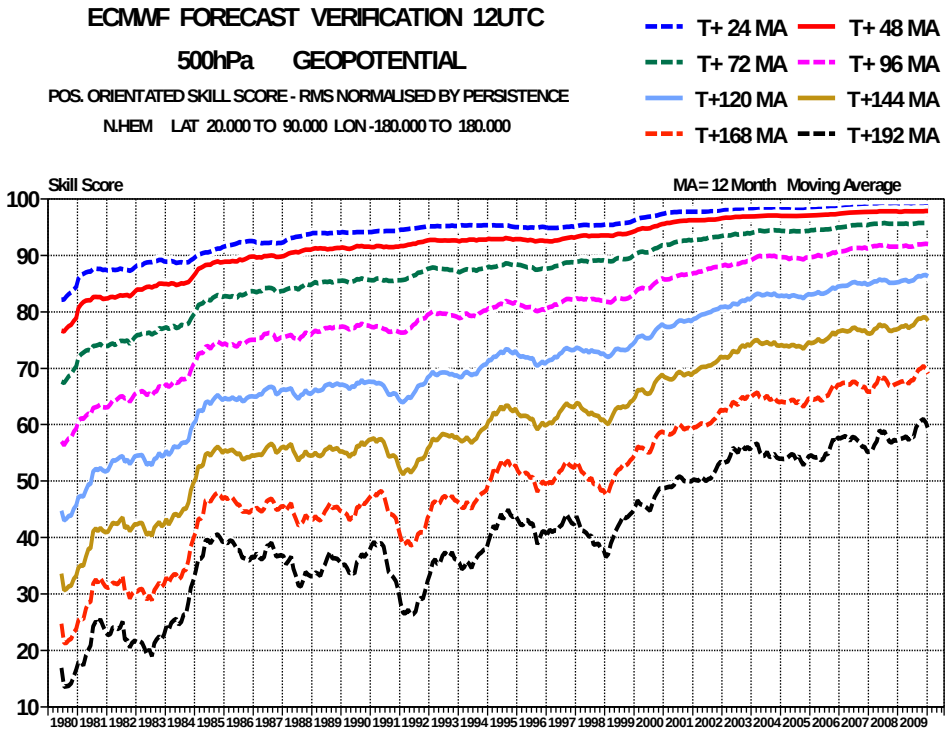


Figure 1: 500 hPa geopotential height skill score for Europe (top) and the northern hemisphere extra-tropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2009 - July 2010.

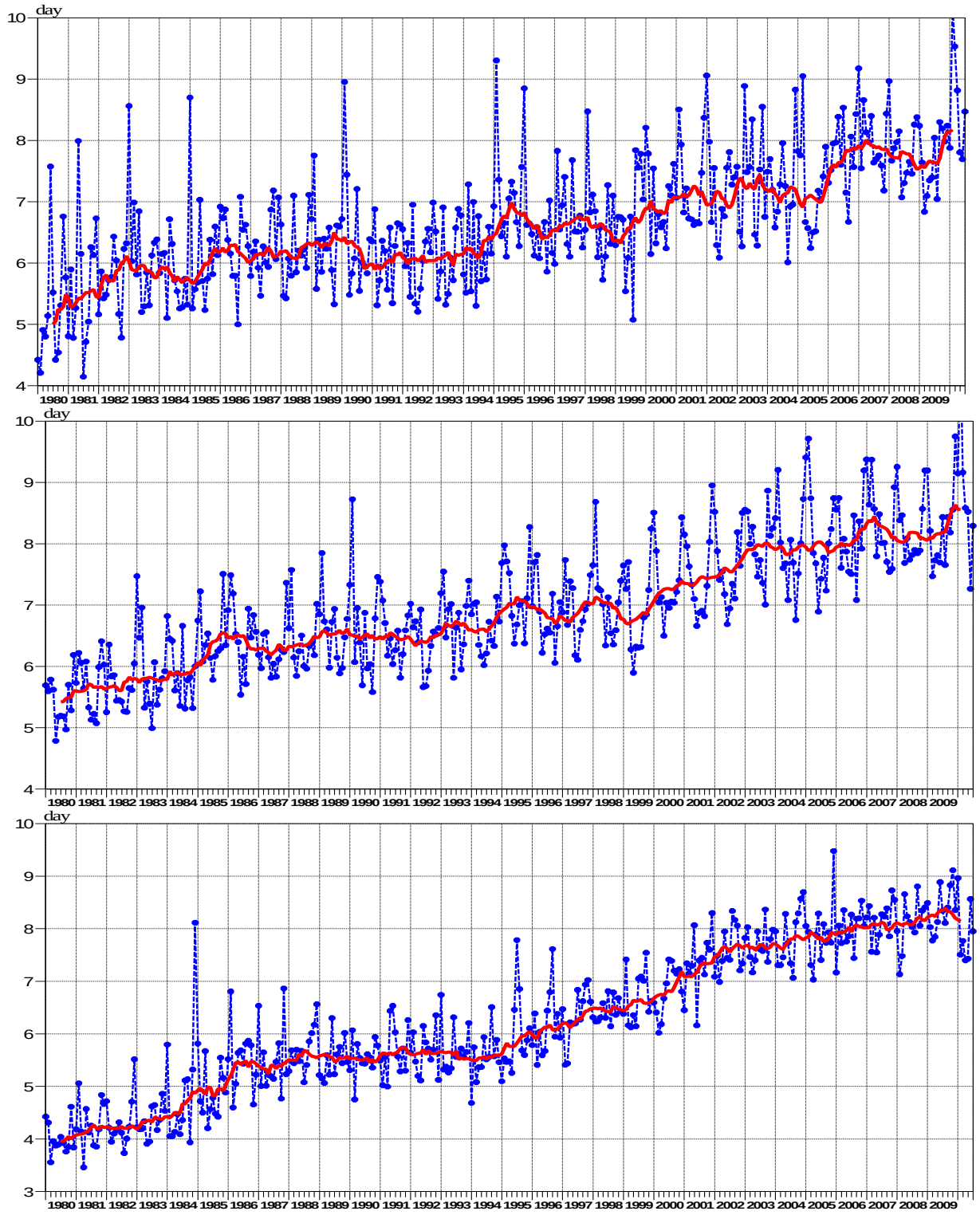


Figure 2: Evolution with time of the 500 hPa geopotential height forecast performance – each point on curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation with the verifying analysis falls below 60% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom). 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).

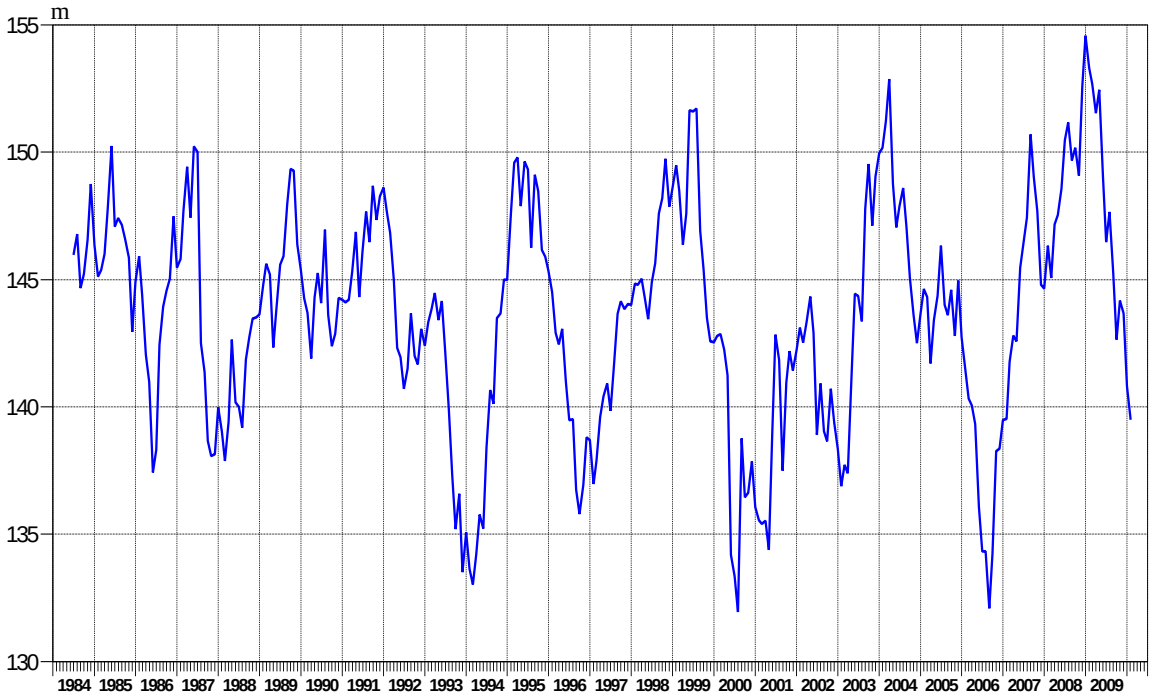


Figure 3: Root mean square error of forecasts made by persisting the analysis over 7 days (168 hours) and verifying it as a forecast for 500 hPa geopotential height over Europe. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2009 - July 2010.

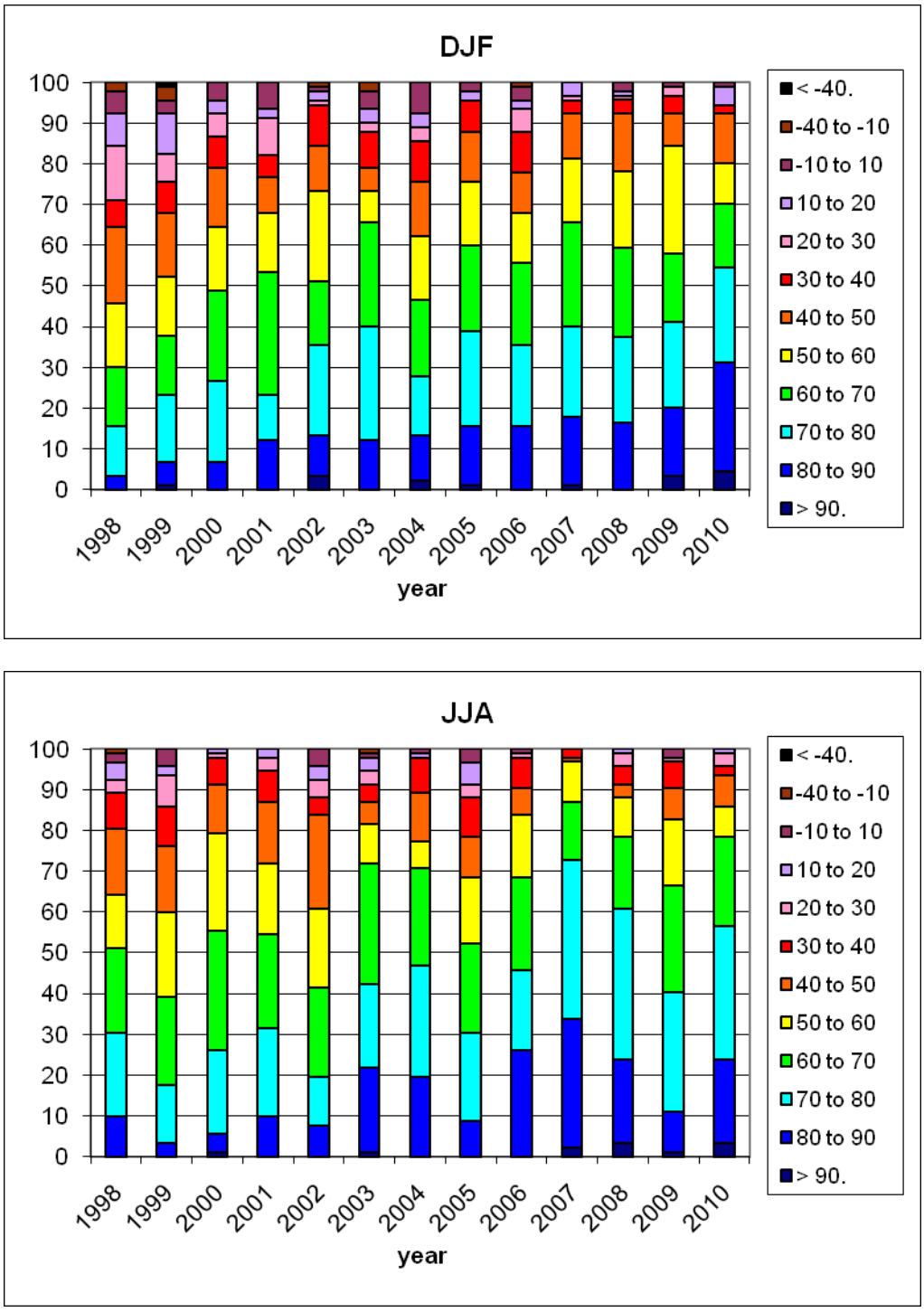


Figure 4: Distribution of Anomaly Correlation of the Day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997-1998.

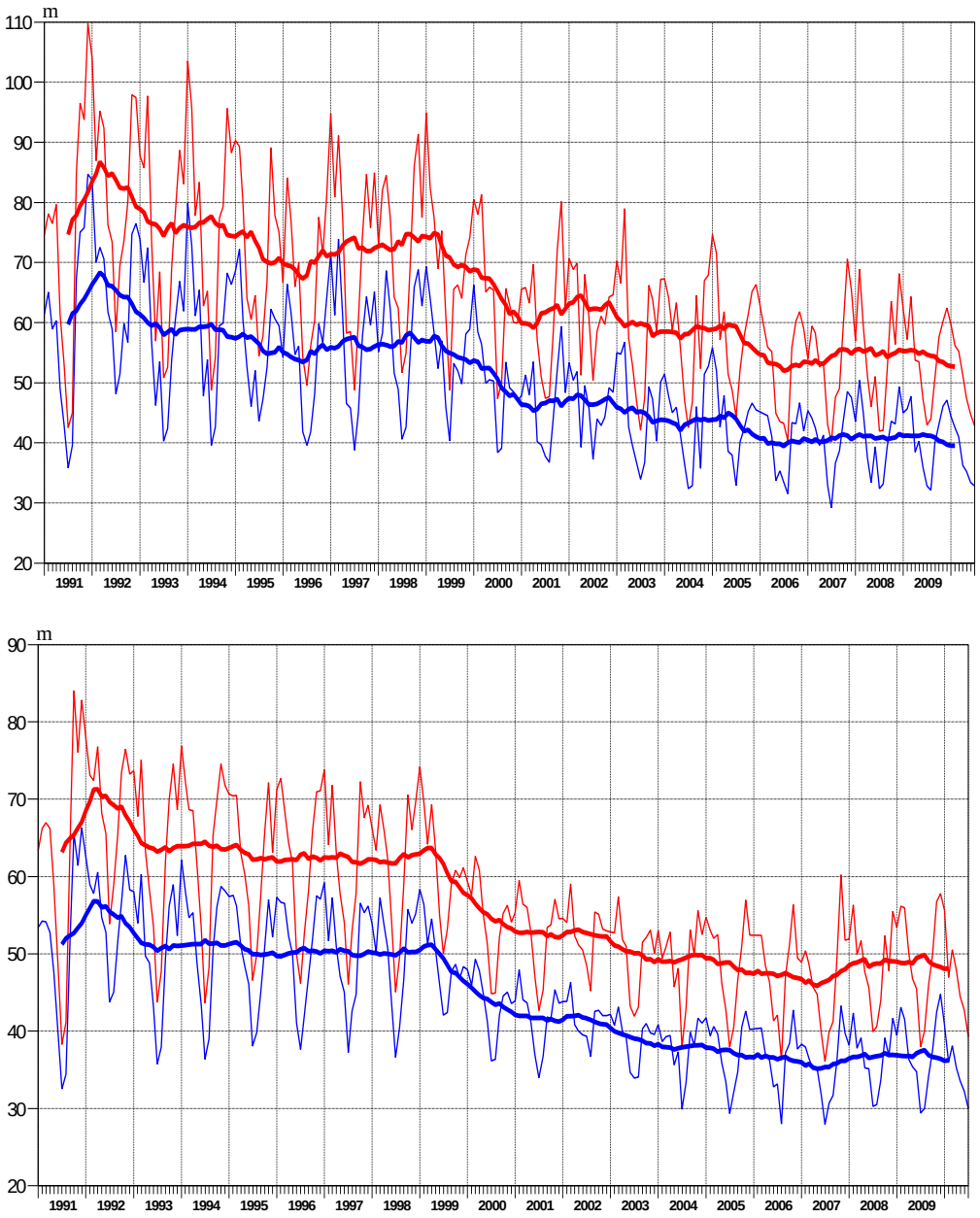


Figure 5: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96-120 h (blue) and 120-144 h (red). 12-month moving average scores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

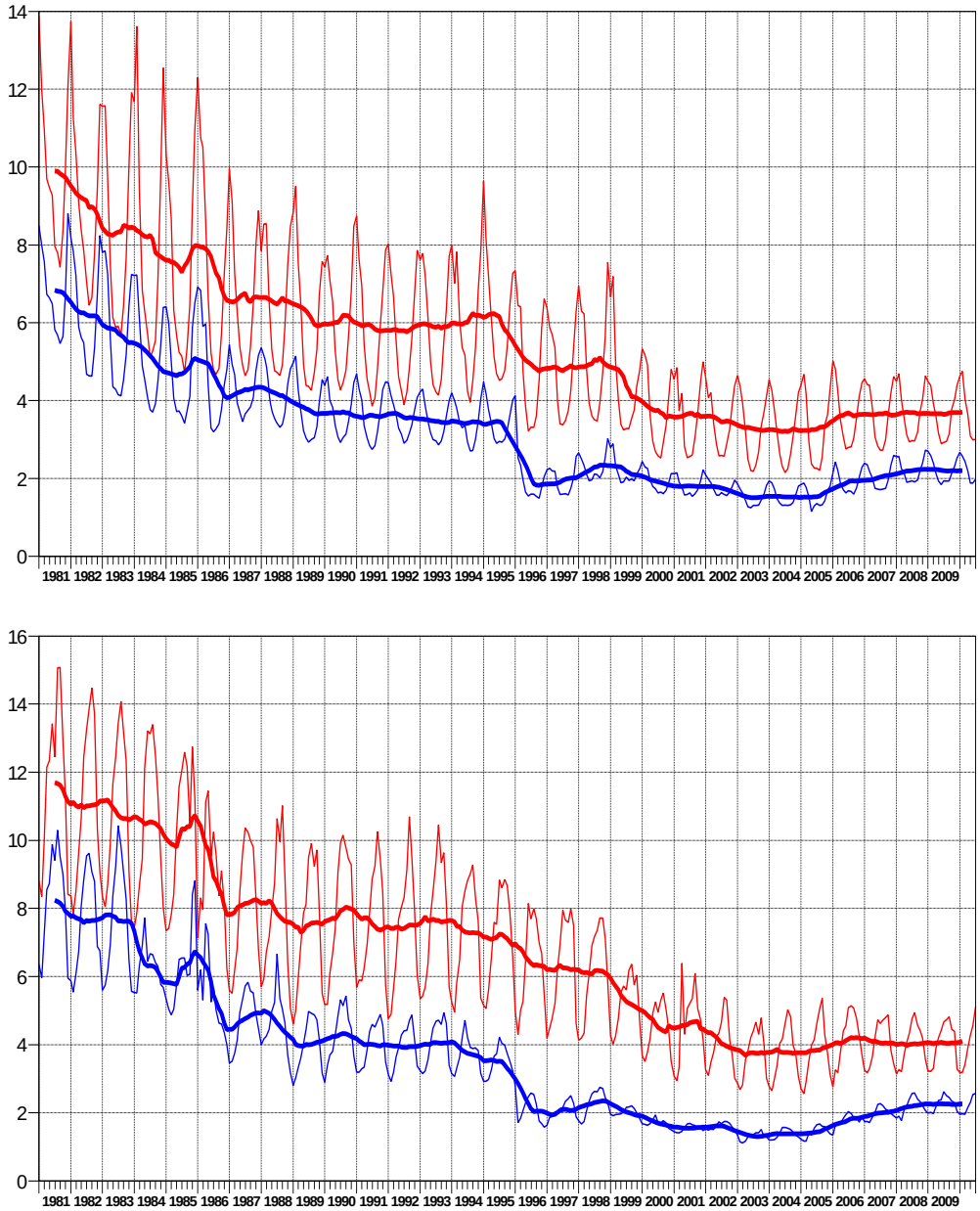


Figure 6: Model scores in the northern (top) and southern (bottom) extra-tropical stratosphere. Curves show the monthly average RMS vector wind error at 50 hPa for 1-day (blue) and 5-day (red) forecasts. 12-month moving average scores are also shown (in bold).

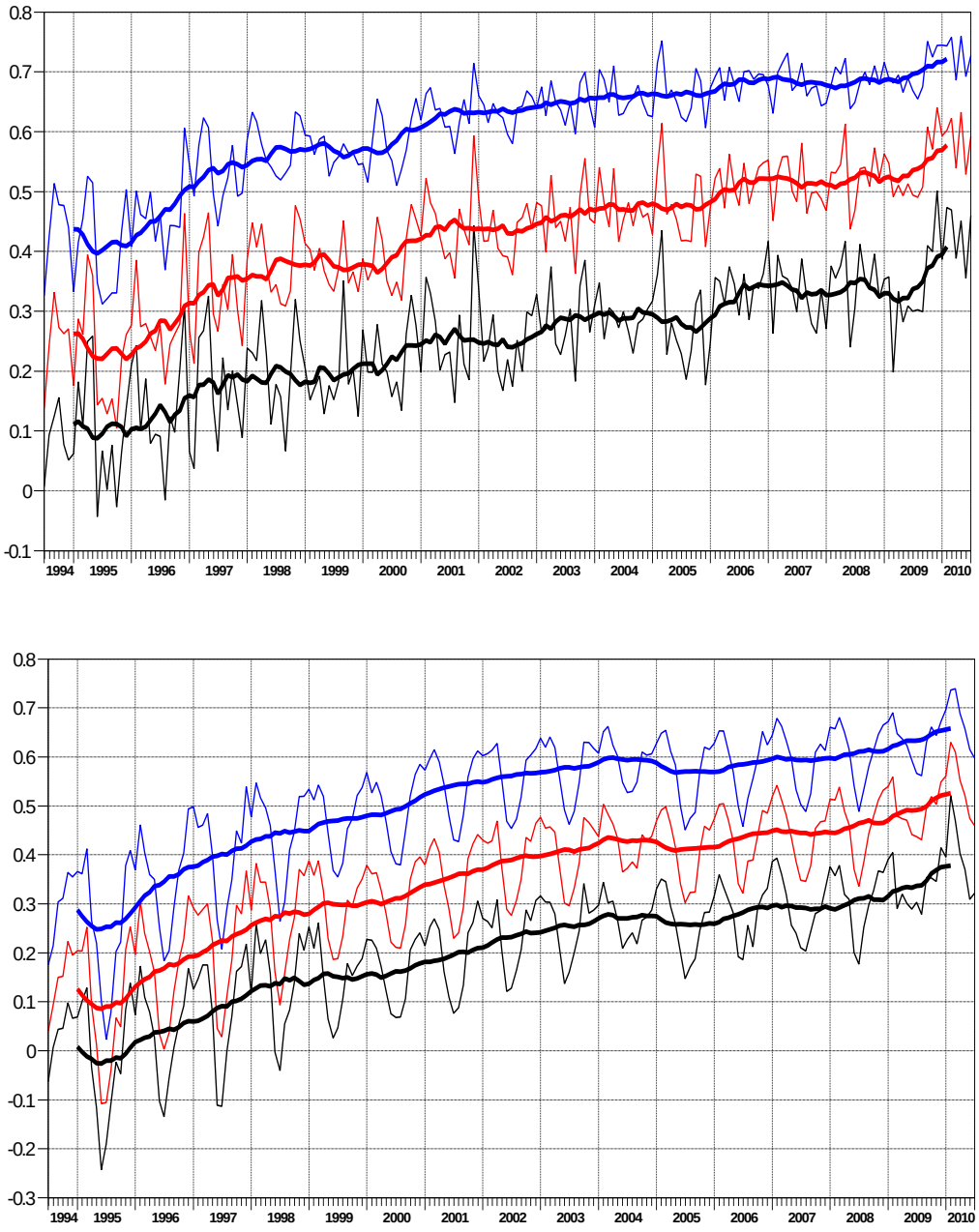


Figure 7: Monthly score and 12-month running mean (bold) of Ranked Probability Skill Score for EPS forecasts of 850 hPa temperature at day 3 (blue), 5 (red) and 7 (black) for Europe (top) and the northern hemisphere extratropics (bottom).

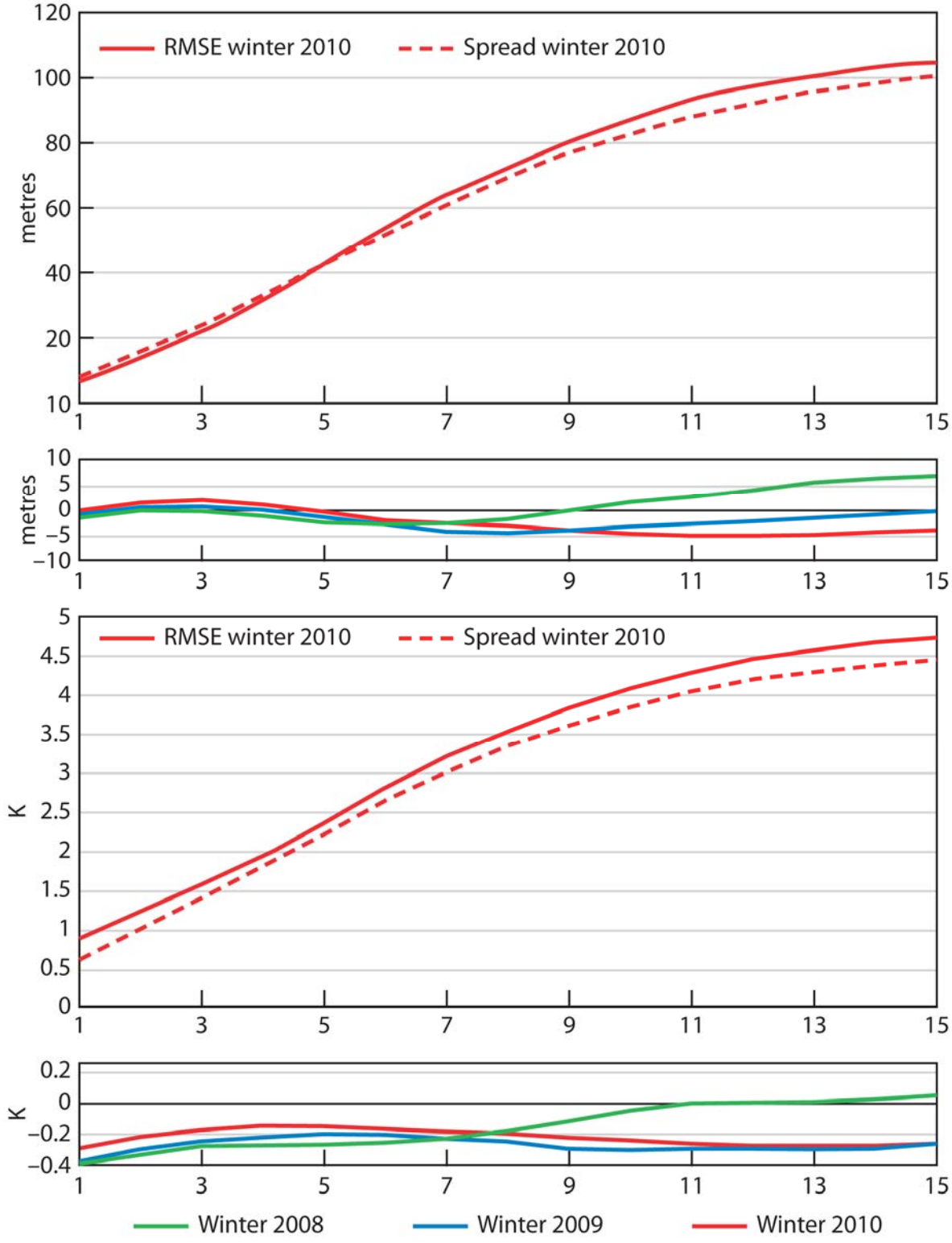


Figure 8: Ensemble spread (standard deviation, dashed lines) and root mean square error of ensemble-mean (solid lines) for winter 2009-2010(upper figure in each panel), complemented with differences of ensemble spread and root mean square error of ensemble-mean for last 3 winter seasons (lower figure 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.

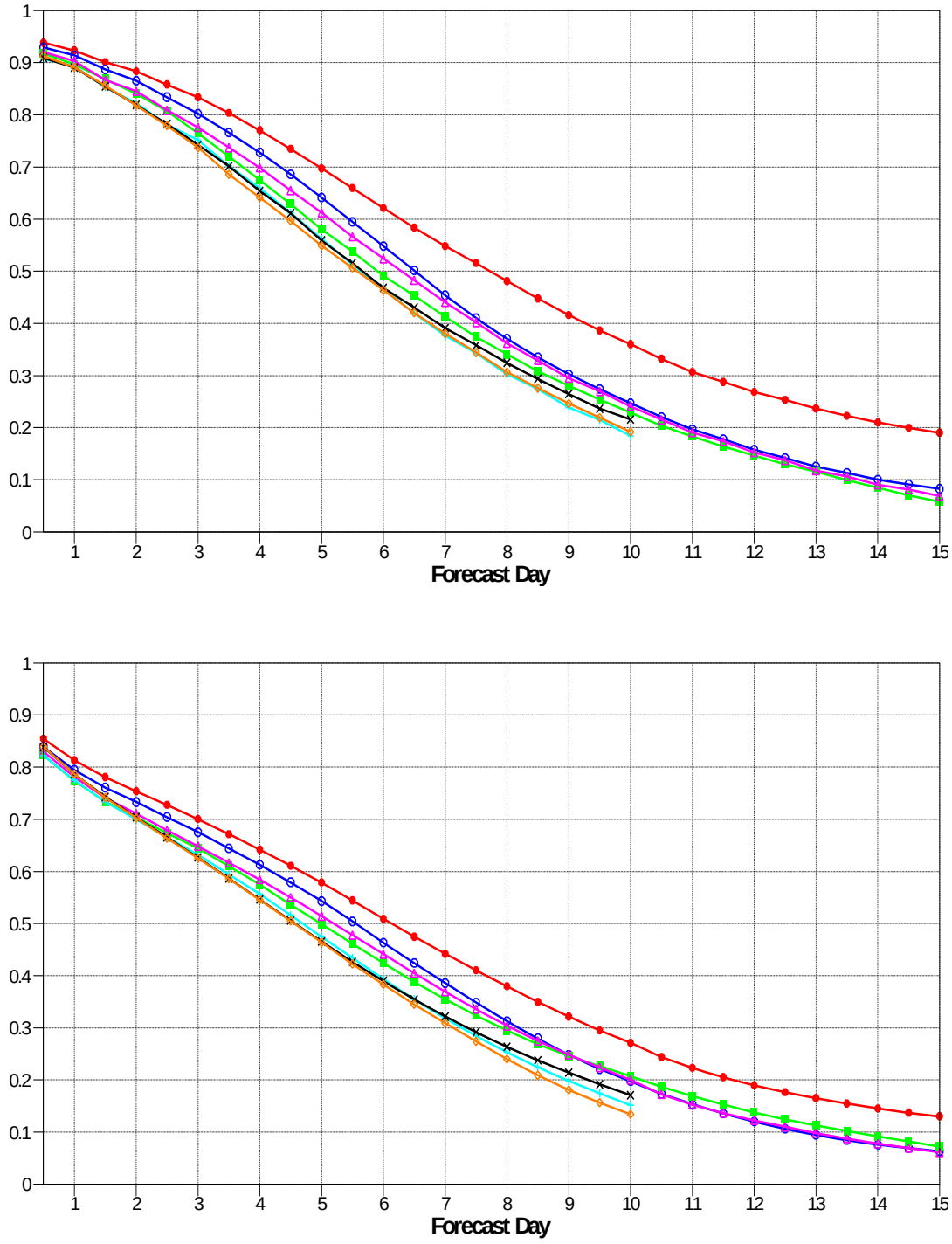


Figure 9: Ranked probability skill score for 500 hPa height (top) and 850 hPa temperature (bottom) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. Skill from the EPS day 1-15 forecasts is shown for winter 2009-10 (red), 2008-09 (blue), 2007-08 (green) and 2006-07 (magenta). The EPS only ran to 10 days in previous years: 2005-06 (cyan), 2004-05 (black), 2003-04 (orange).

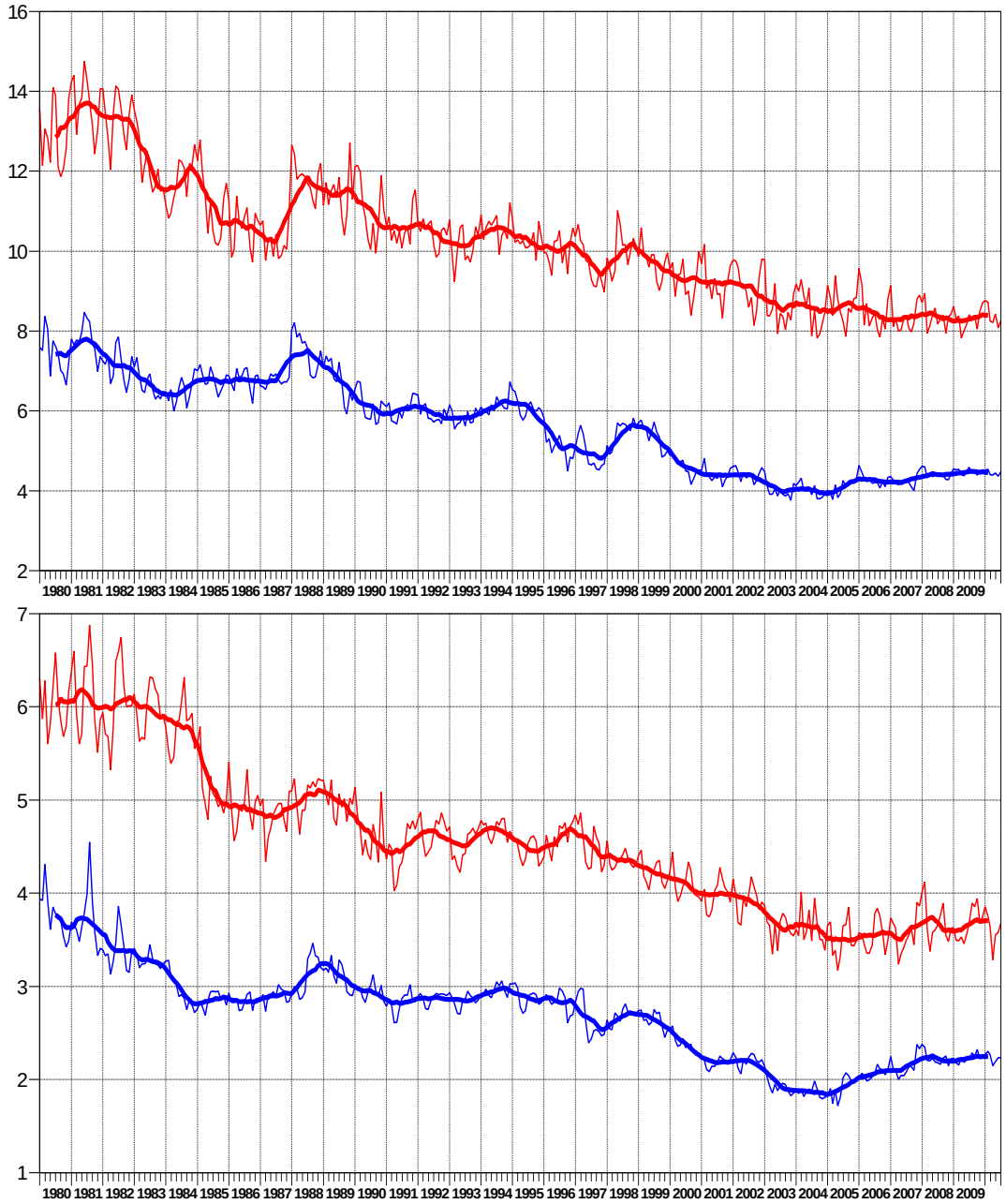


Figure 10: Model scores in the tropics. Curves show the monthly average root mean square vector wind errors at 200 hPa (top) and 850 hPa (bottom) for 1-day (blue) and 5-day (red) forecasts. 12-month moving average scores are also shown (in bold).

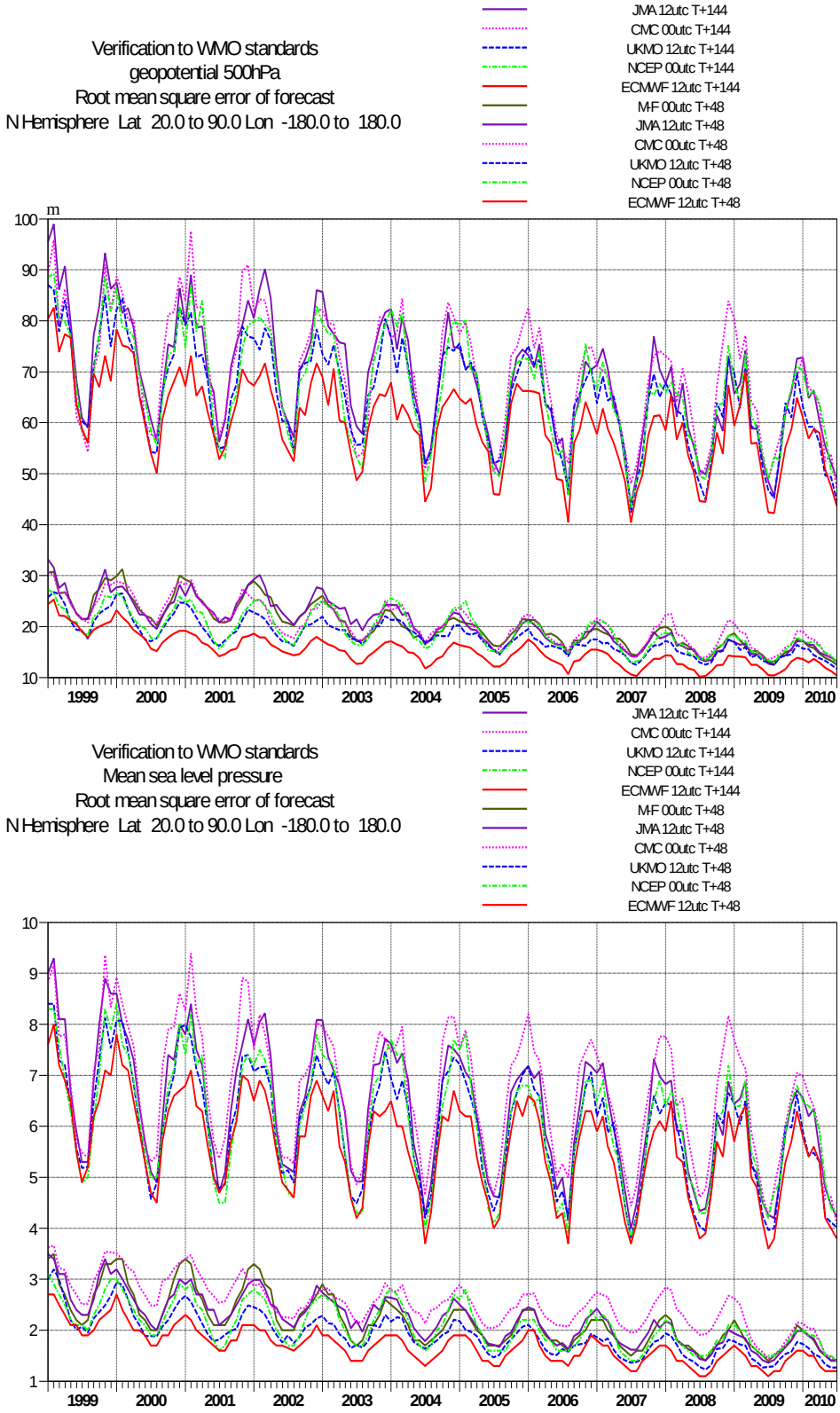


Figure 11: WMO/CBS exchanged scores from global forecast centres. RMS error over northern extratropics for 500 hPa geopotential height (top) and MSLP (bottom). In each panel the upper curves show the 6-day forecast error and the lower curves show the 2-day forecast error. Each model is verified against 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.

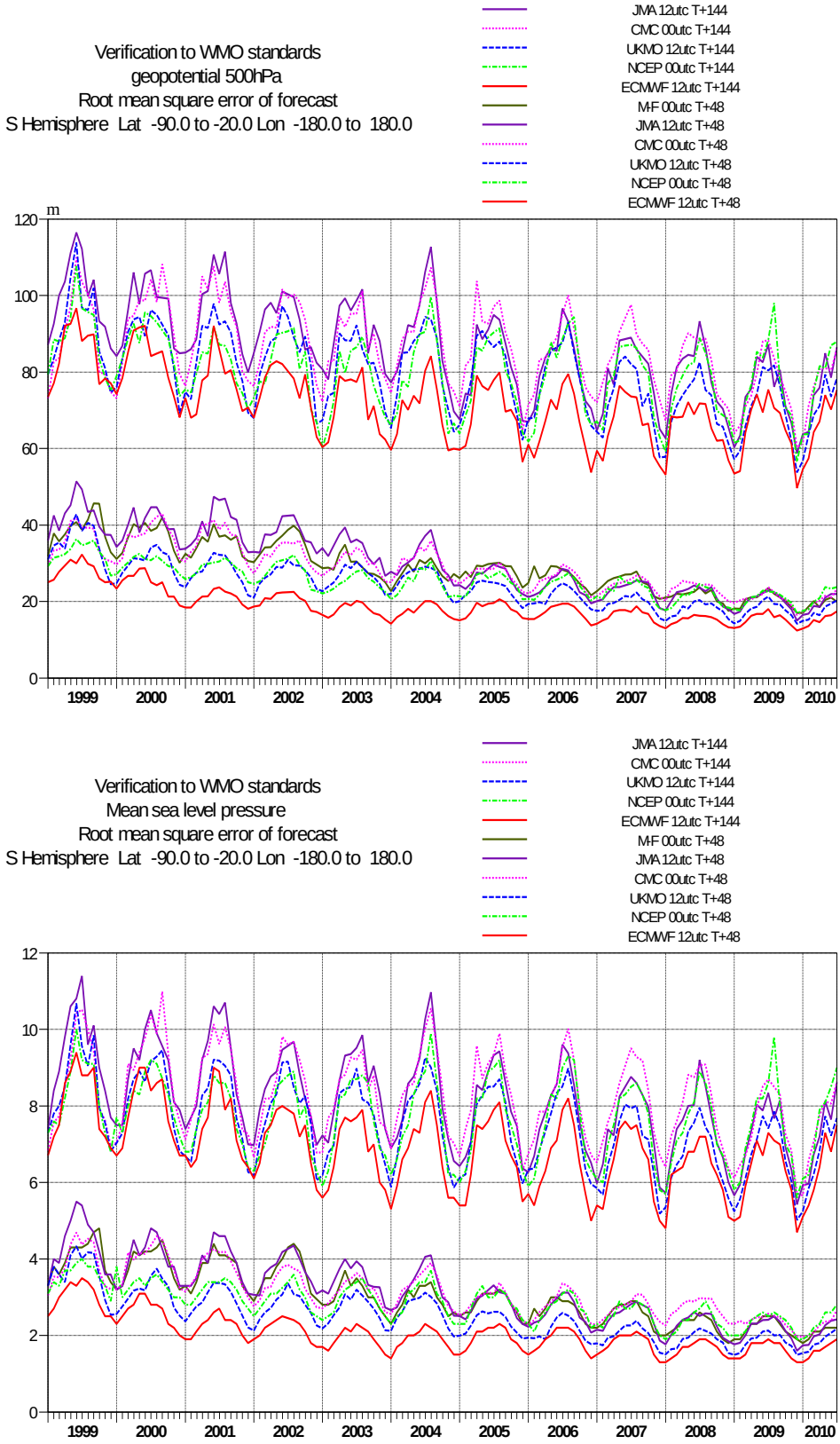
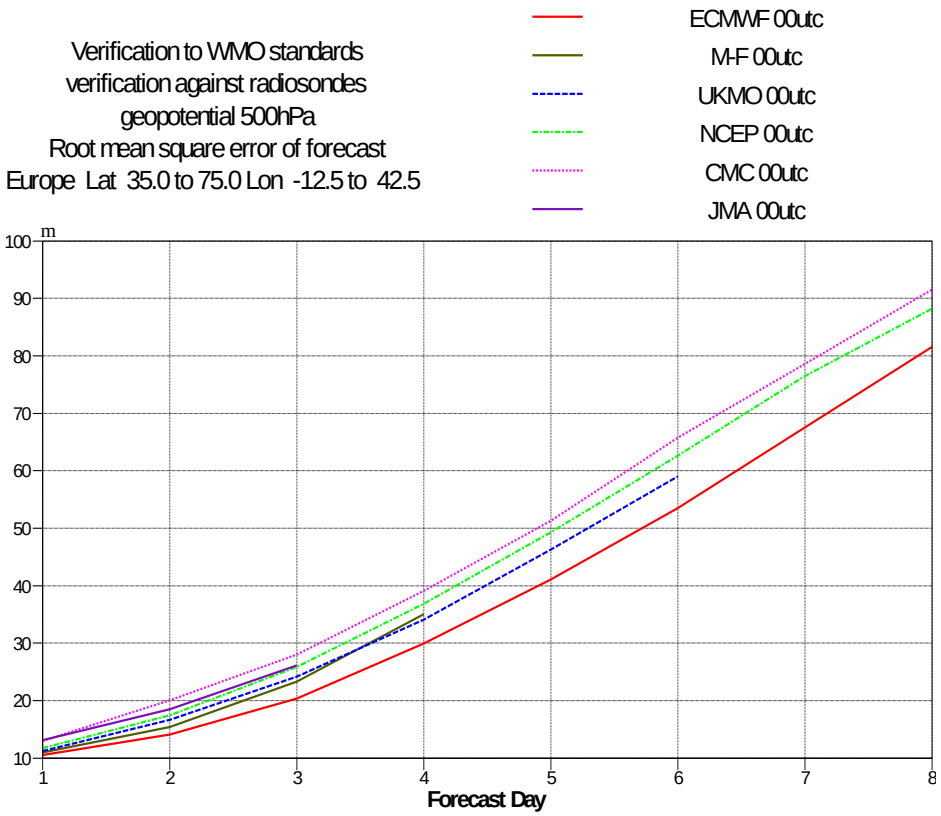


Figure 12: As Figure 11 for the southern hemisphere.



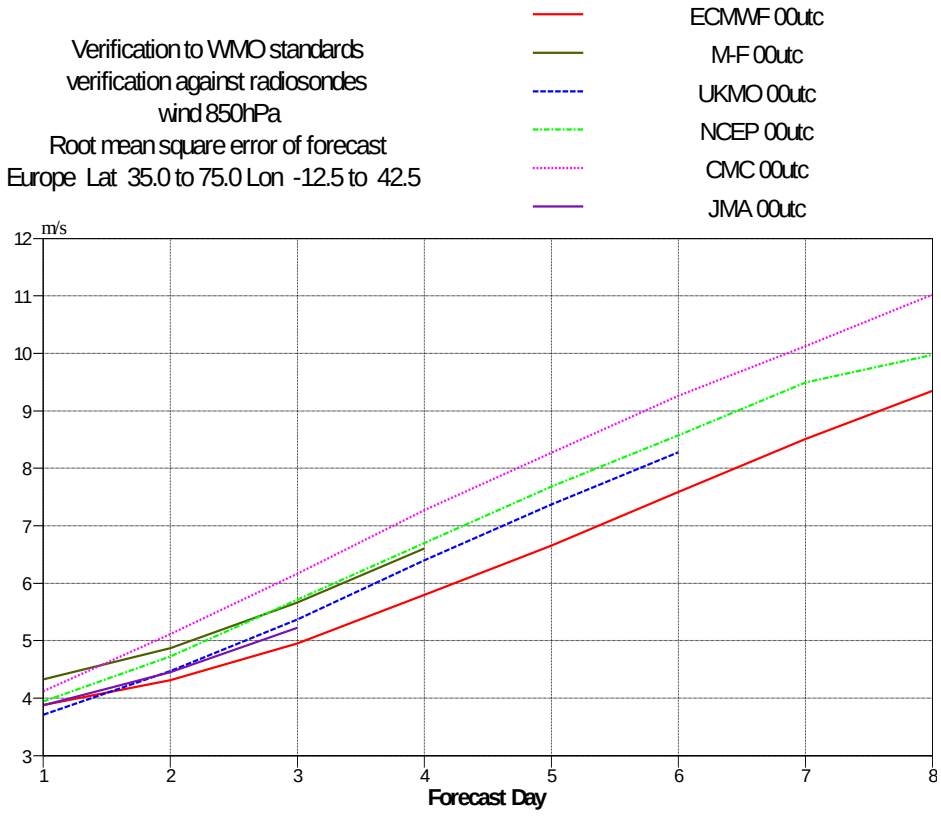


Figure 13: WMO/CBS exchanged scores using radiosondes: 500 hPa height (top) and 850 hPa wind(bottom) RMS error over Europe (annual mean August 2009 – July 2010).

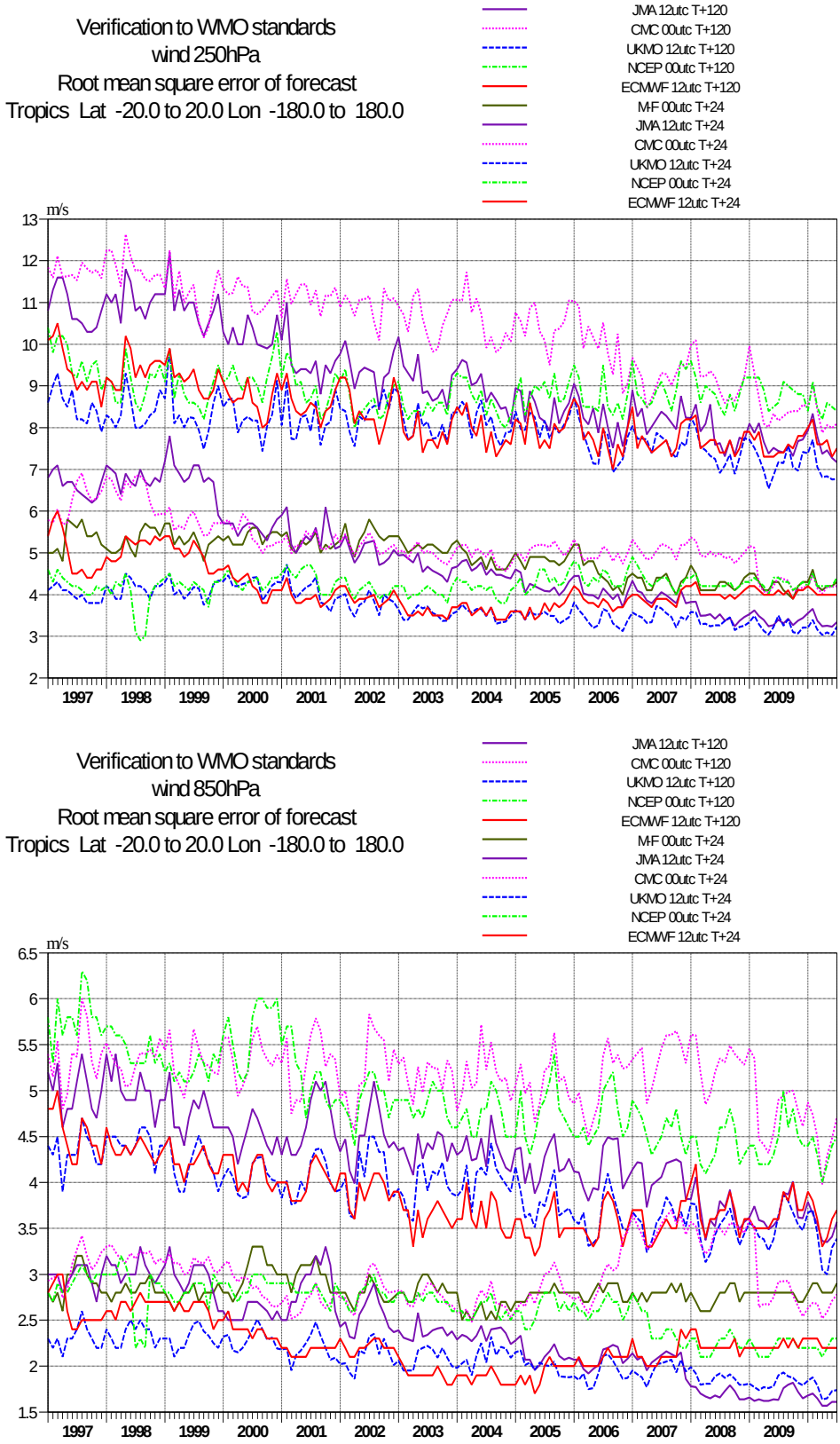
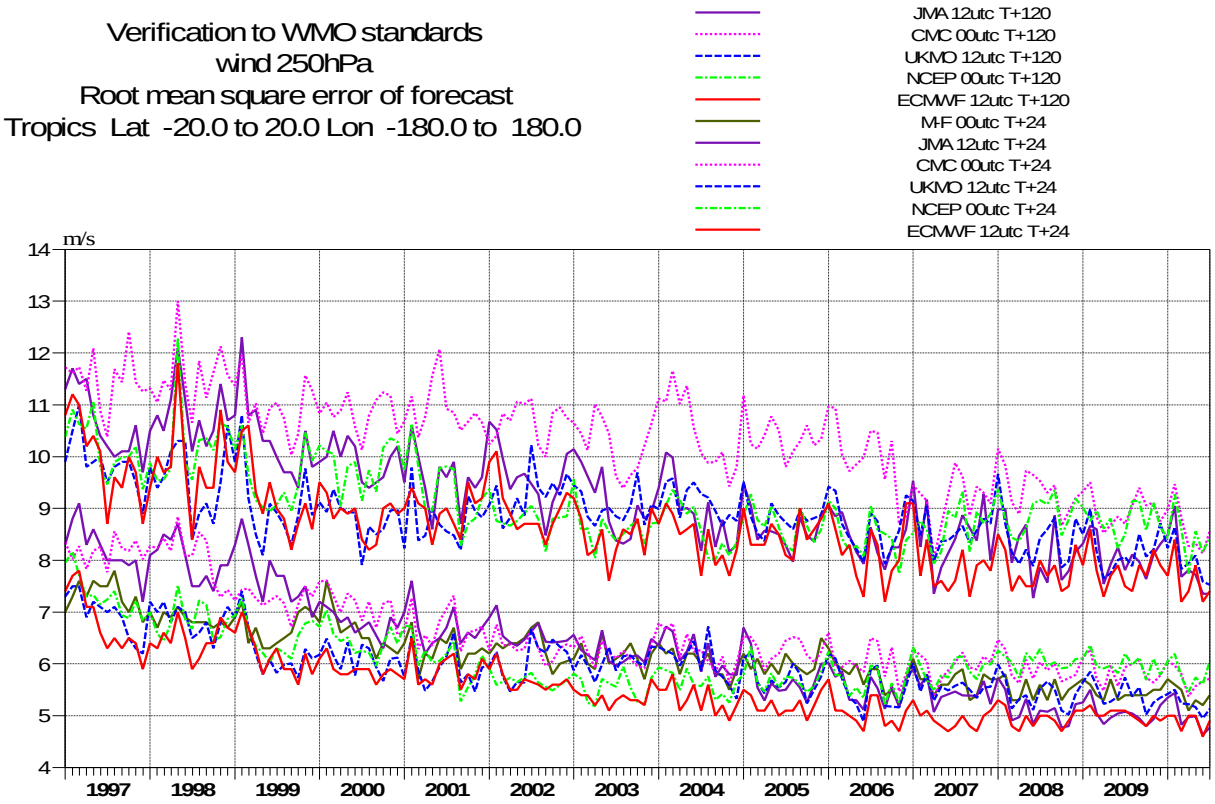


Figure 14: WMO/CBS exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the 5-day forecast error and the lower curves show the 1-day forecast error. Each model is verified against its own analysis.



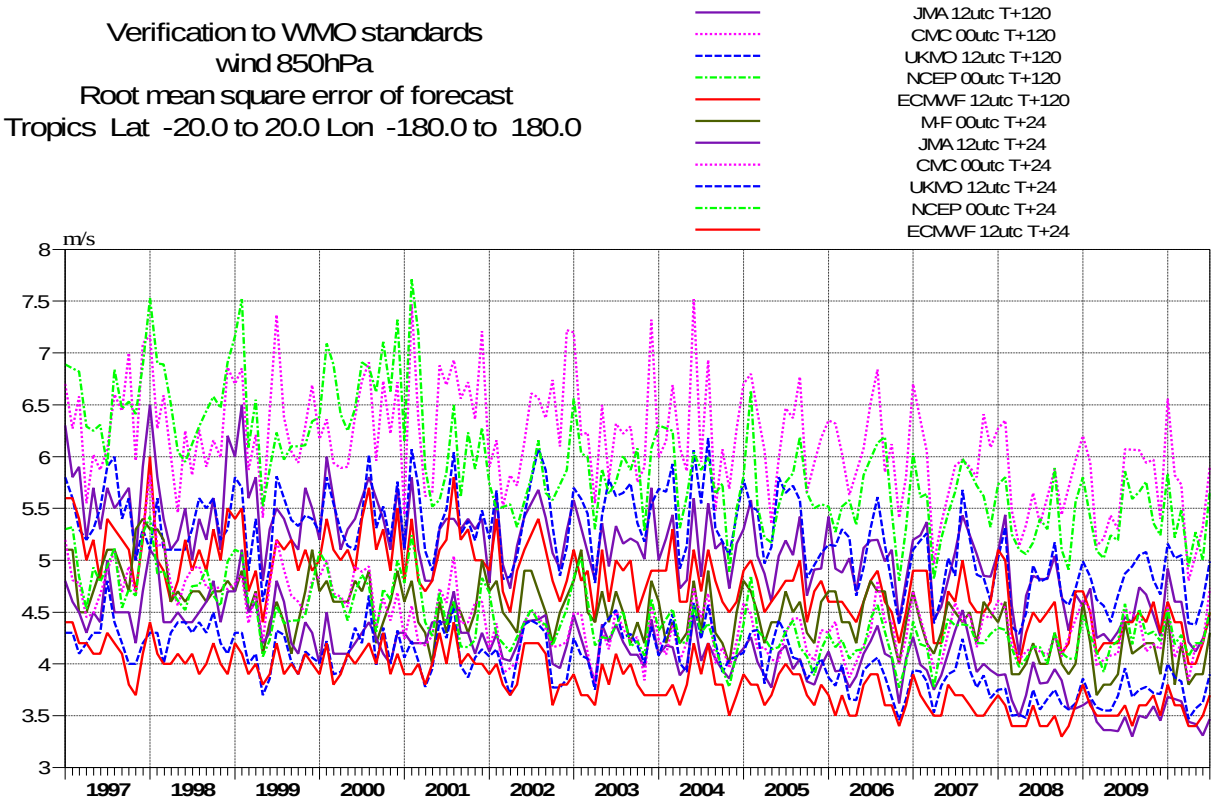


Figure 15: As Figure 14 for scores computed against radiosondes observations.

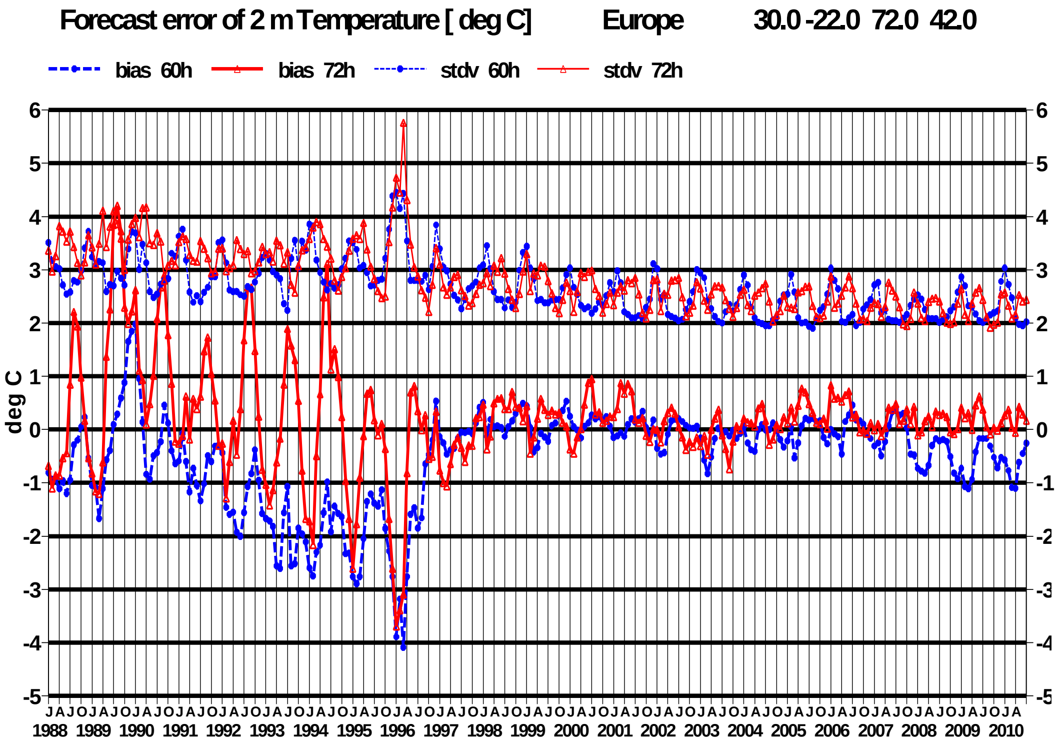


Figure 16: Verification of 2 metre temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

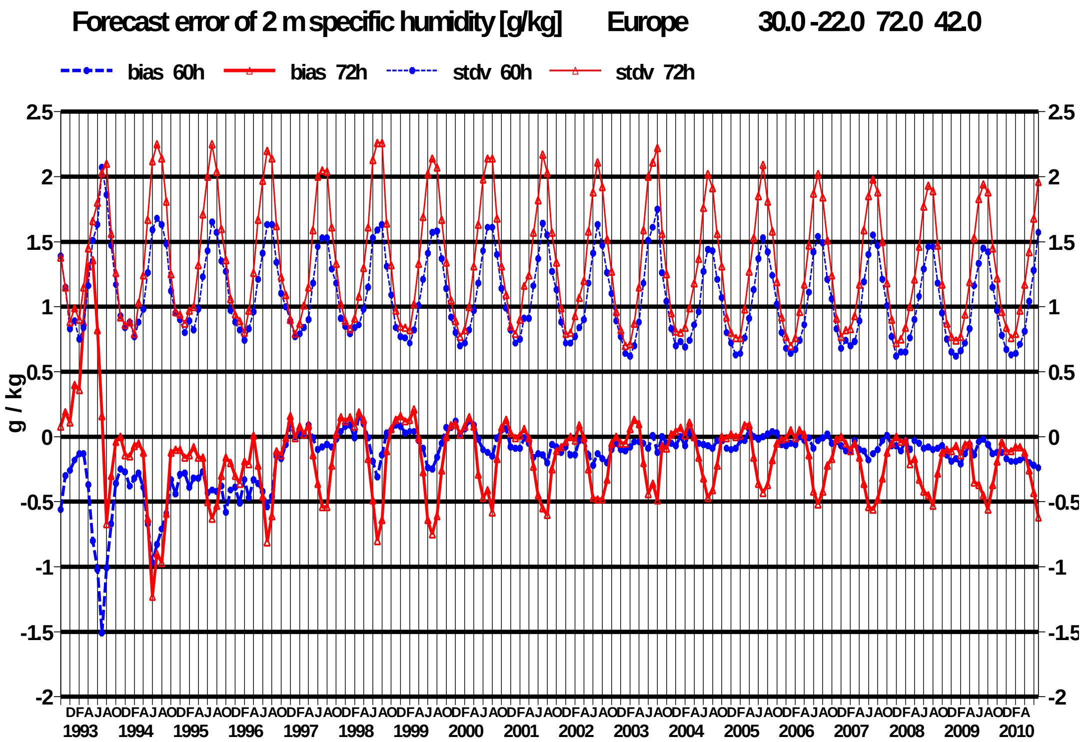


Figure 17: Verification of 2 metre specific humidity forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

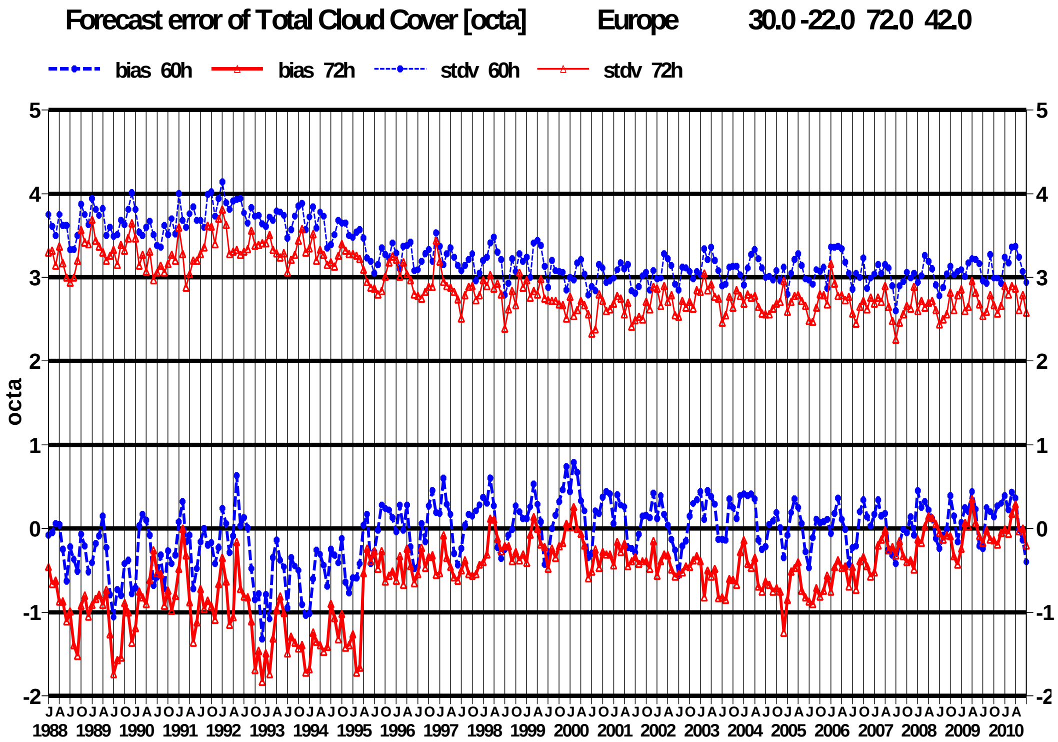


Figure 18: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

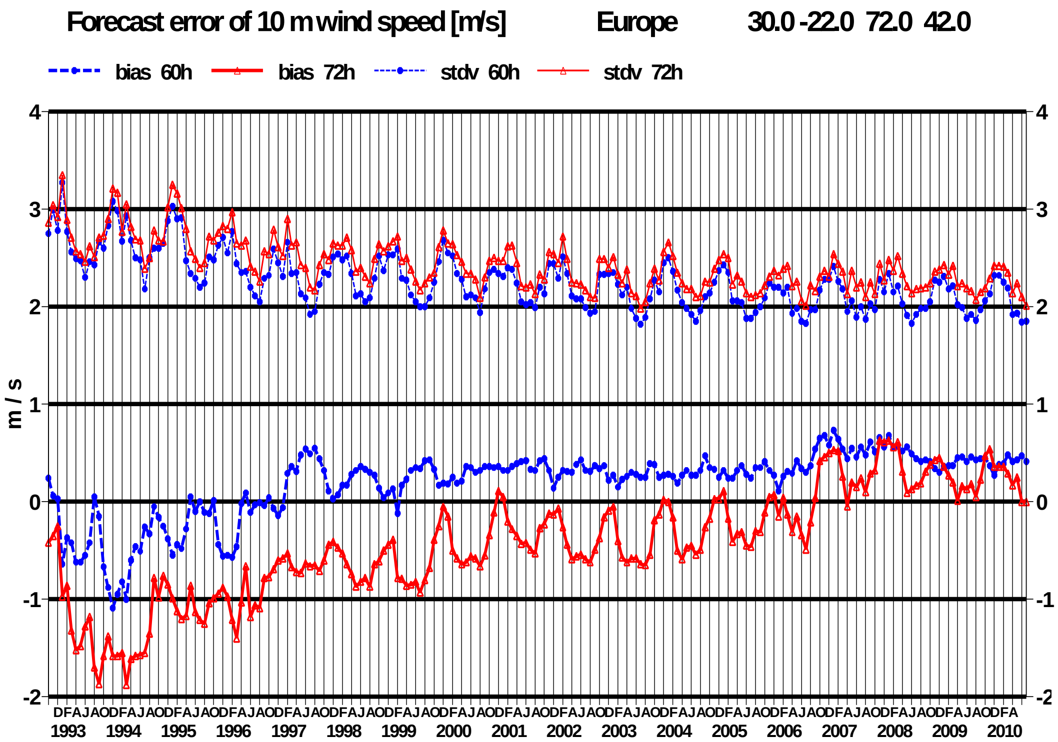
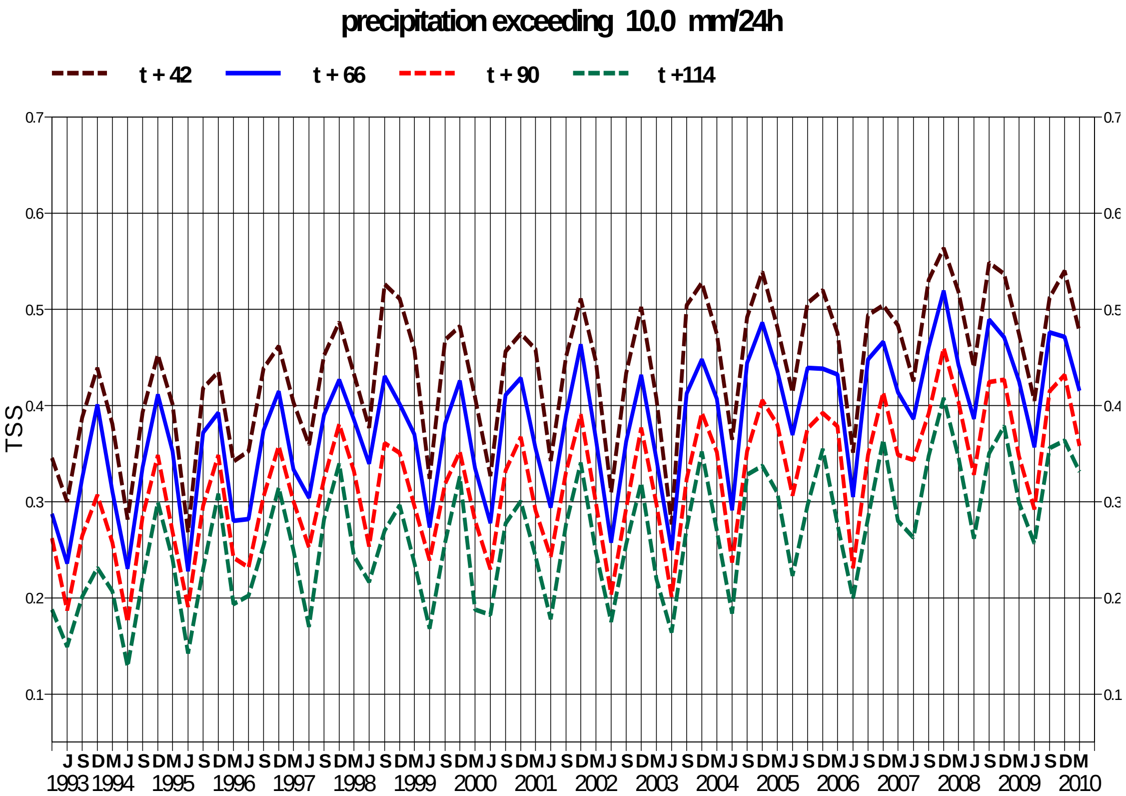


Figure 19: Verification of 10 metre wind speed forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.



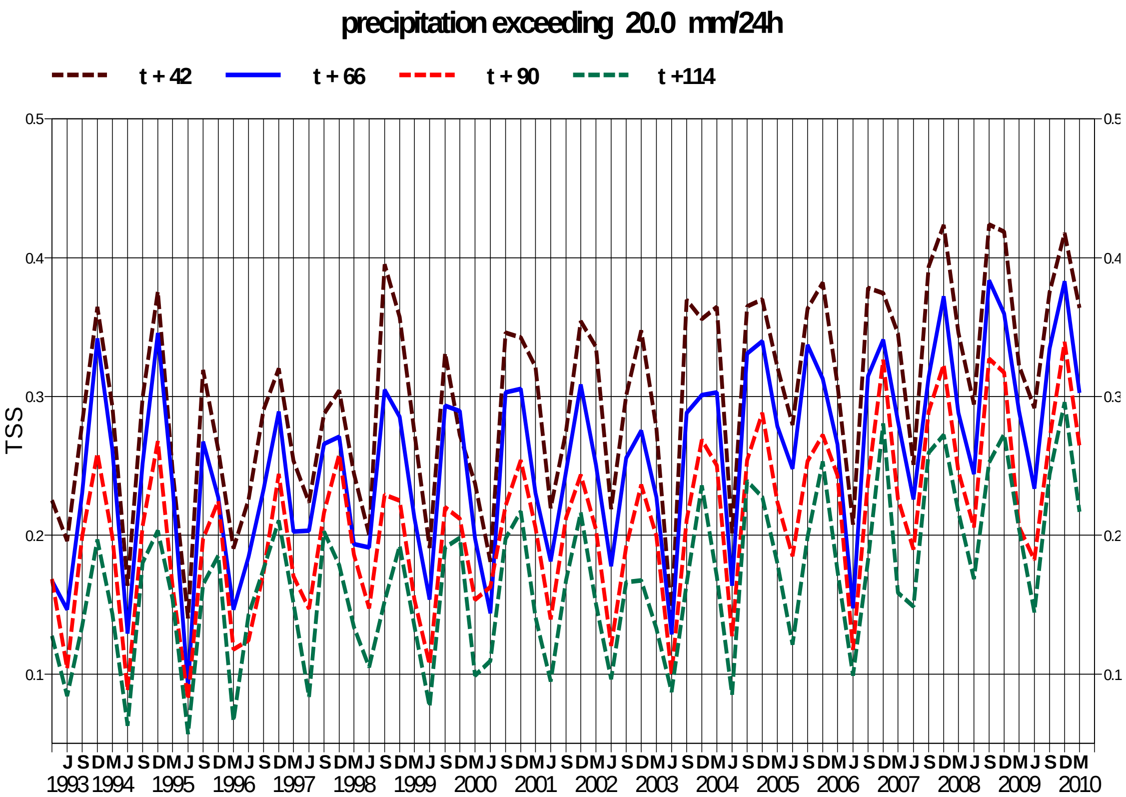


Figure 20: True Skill Score (TSS) time series for precipitation forecasts exceeding 10 mm/day (top) and 20 mm/day (bottom) verified against SYNOP data on the GTS for Europe. Curves are shown for the 24 hour accumulations up to 42, 66, 90, and 114 hours (from the forecasts starting at 12 UTC). 3 month mean scores (last point is March-May 2010).

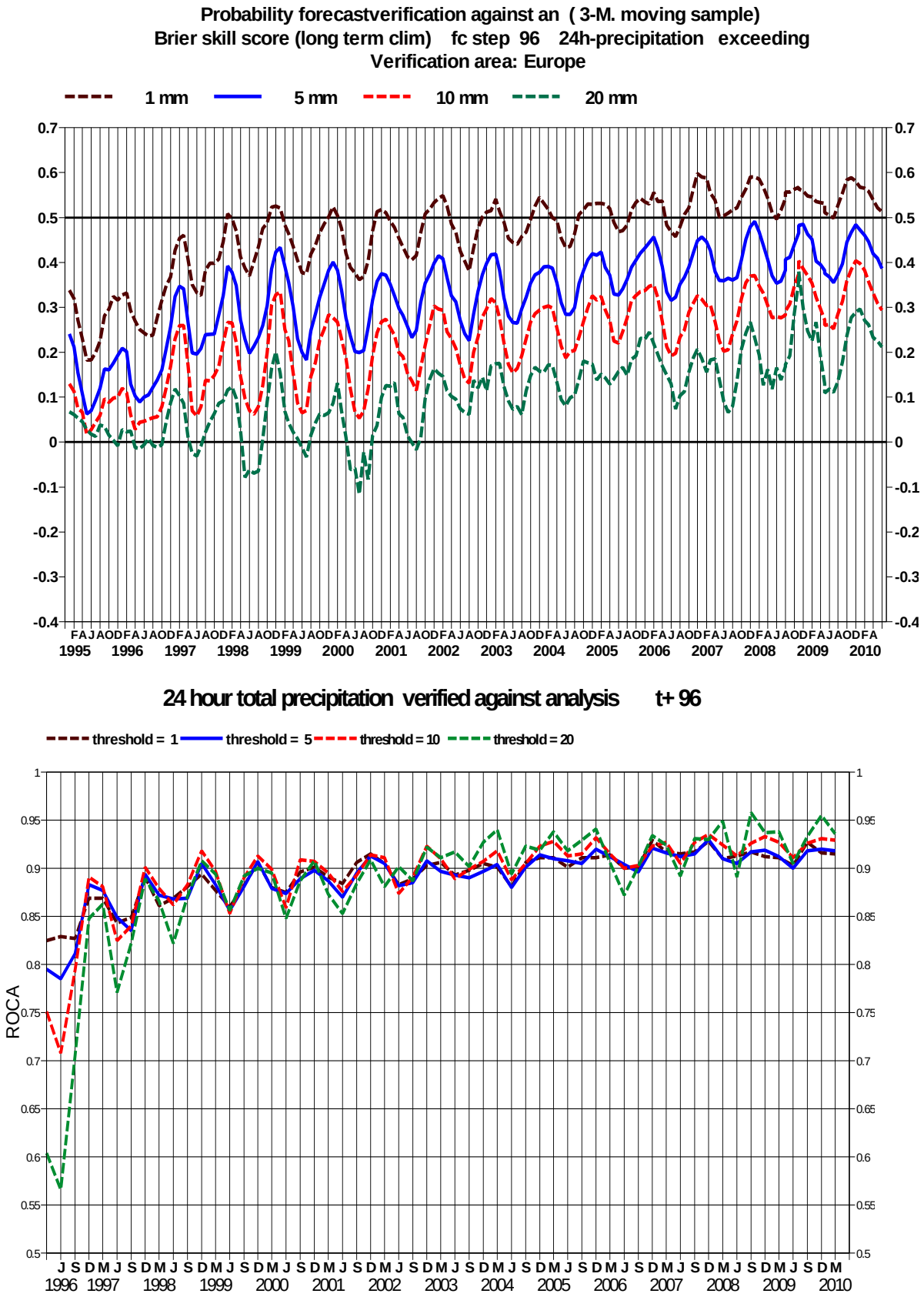


Figure 21: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA) for EPS probability forecasts of precipitation over Europe exceeding thresholds of 1, 5, 10 and 20 mm/day at day 4. The skill score is calculated for three-month running periods.

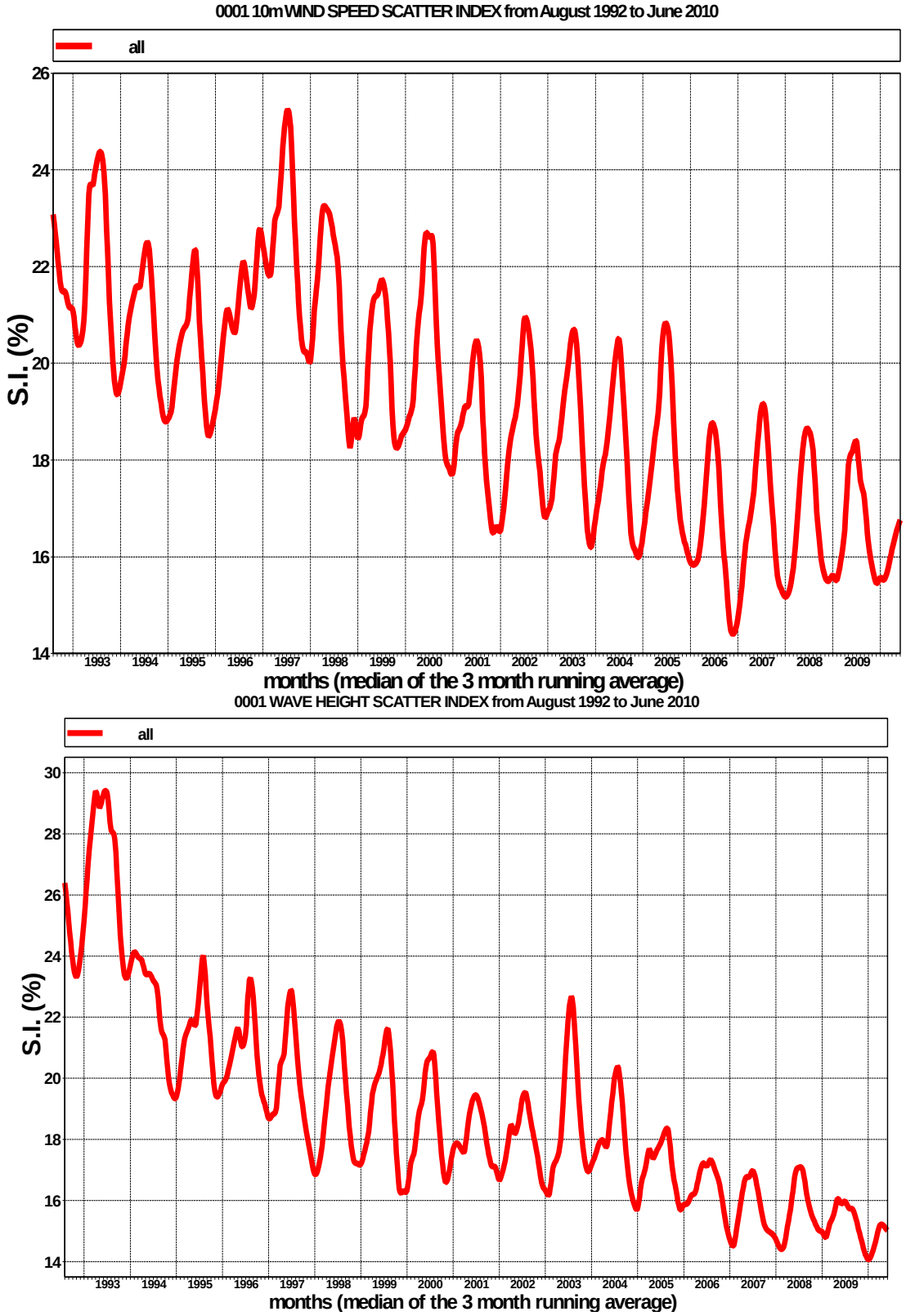


Figure 22: Time series of verification of the ECMWF 10 metre wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

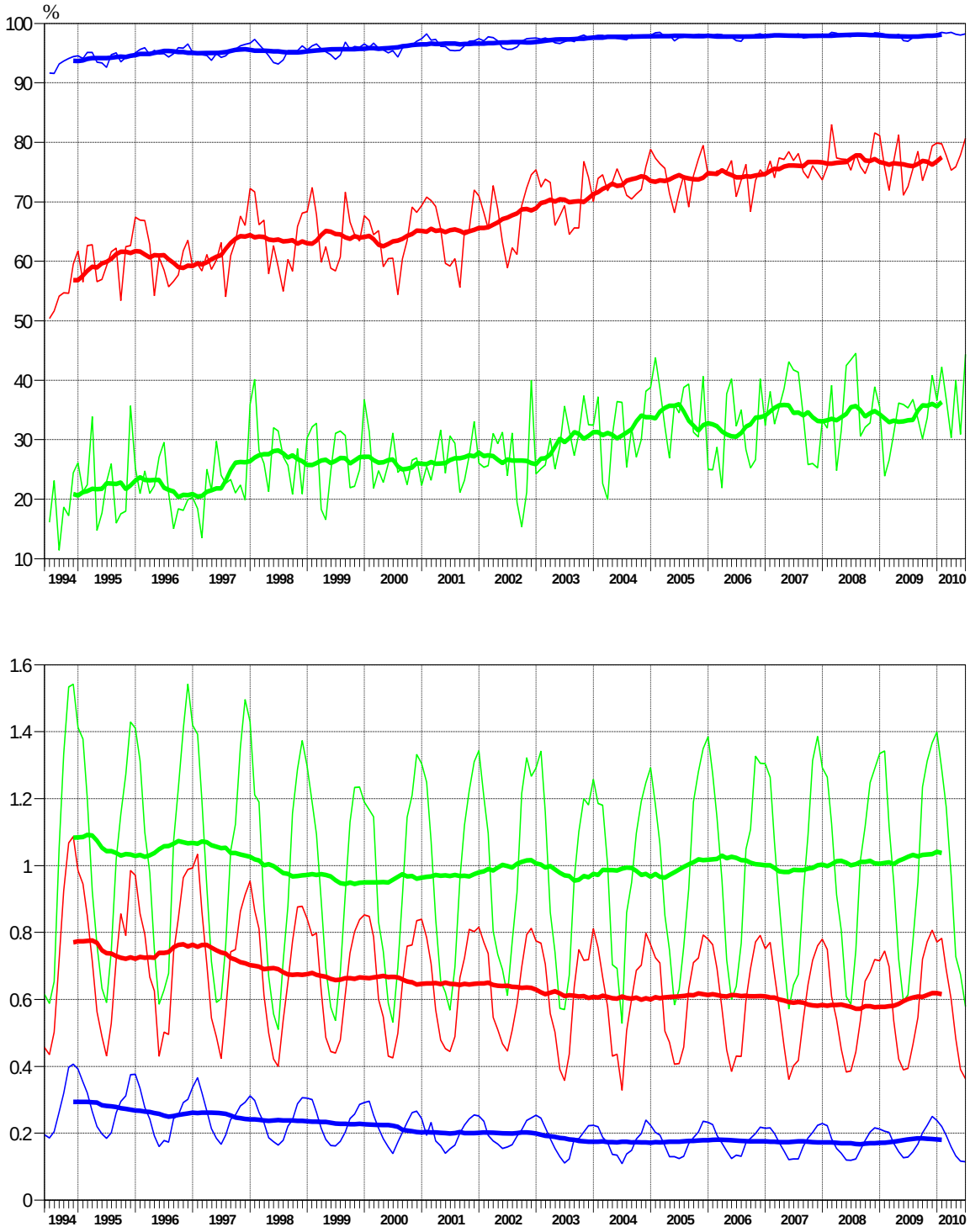


Figure 23: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of anomaly correlation (top) and error standard deviation (bottom) for ocean wave heights verified against analysis for the northern extratropics at day 1 (blue), 5 red) and 10 (green).

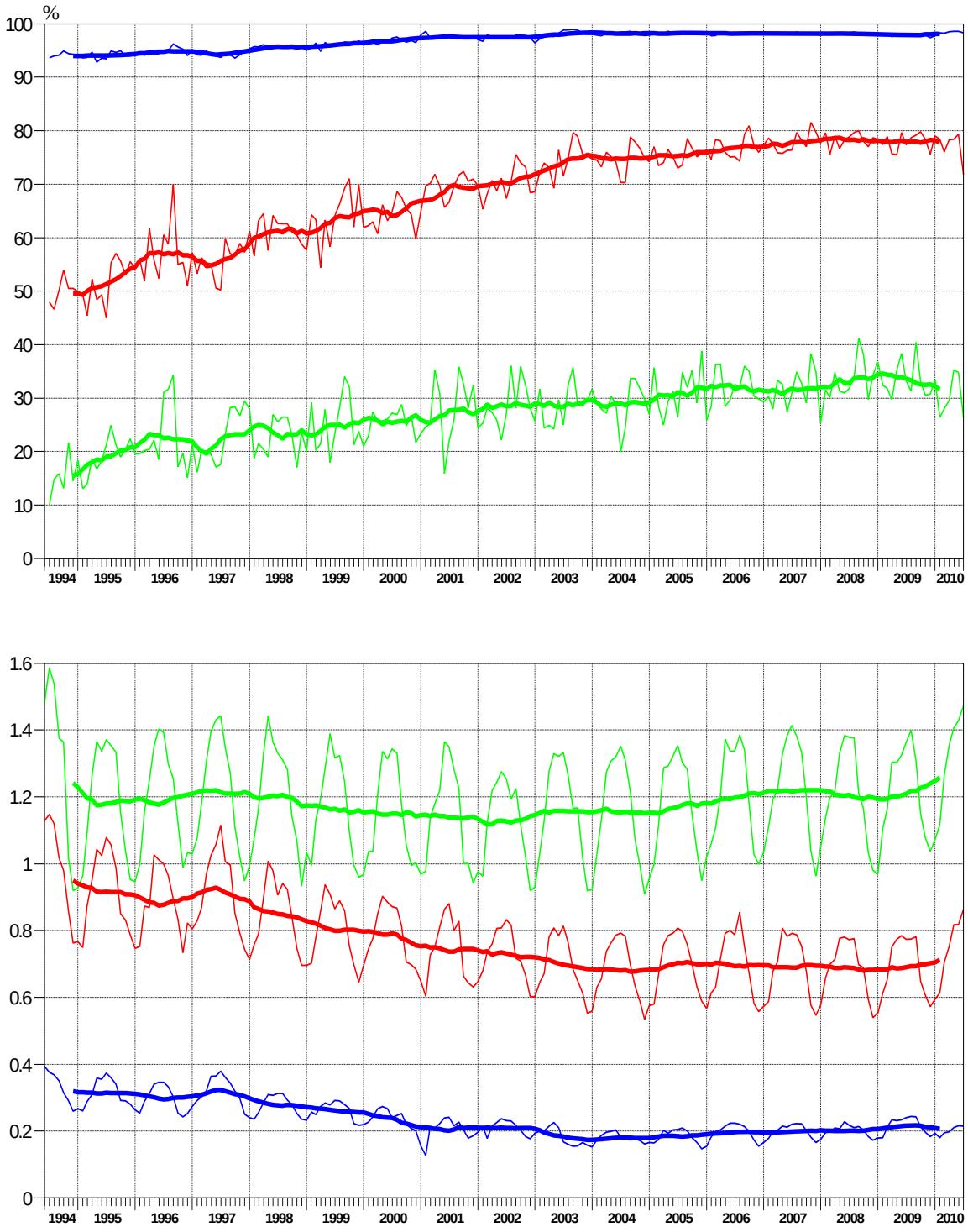


Figure 24: As Figure 23 for the southern hemisphere.

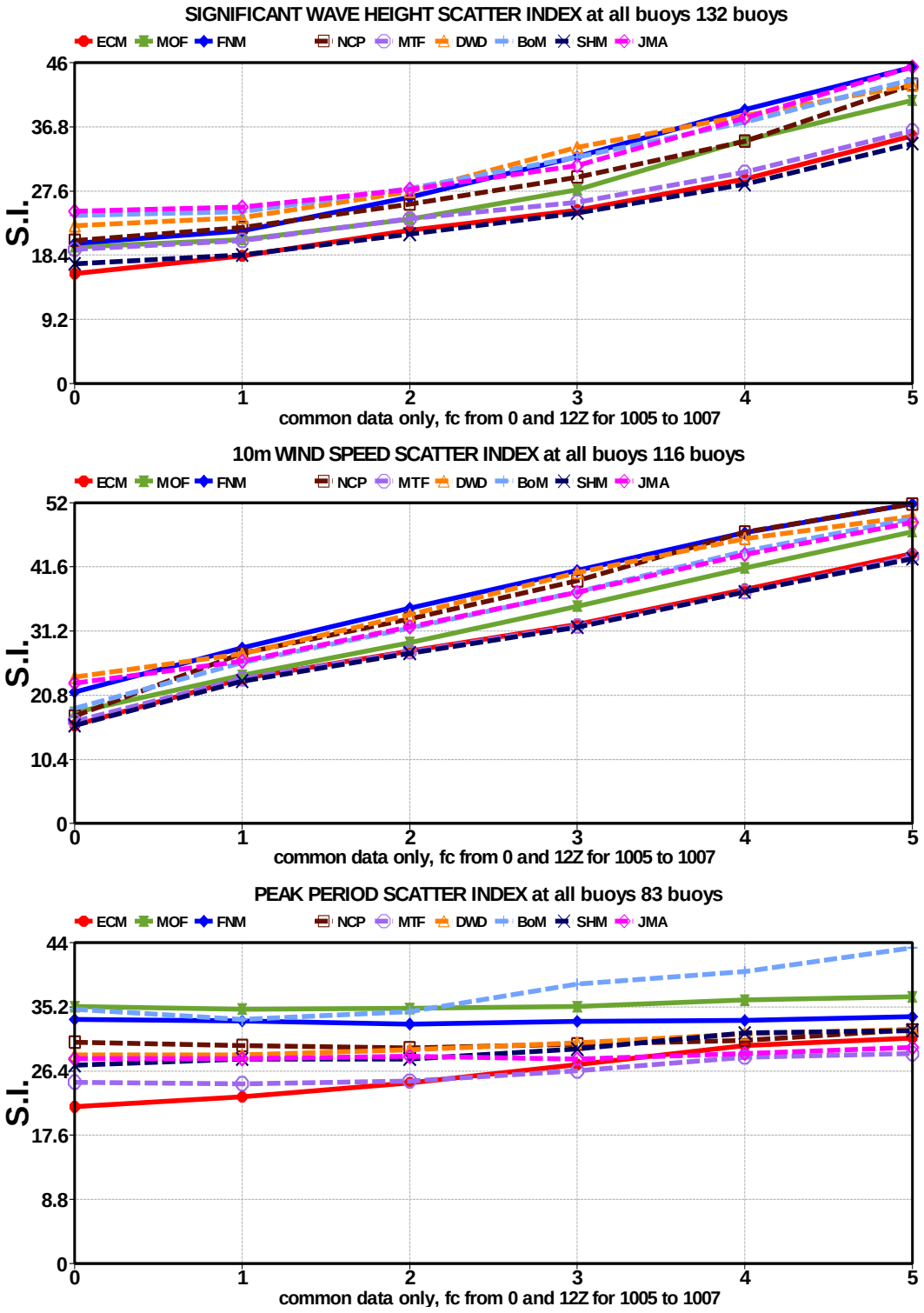


Figure 25: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 3-month period May-July 2010. 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; SHM: Service Hydrographique et Océanographique de la Marine, France; JMA: Japan Meteorological Agency.

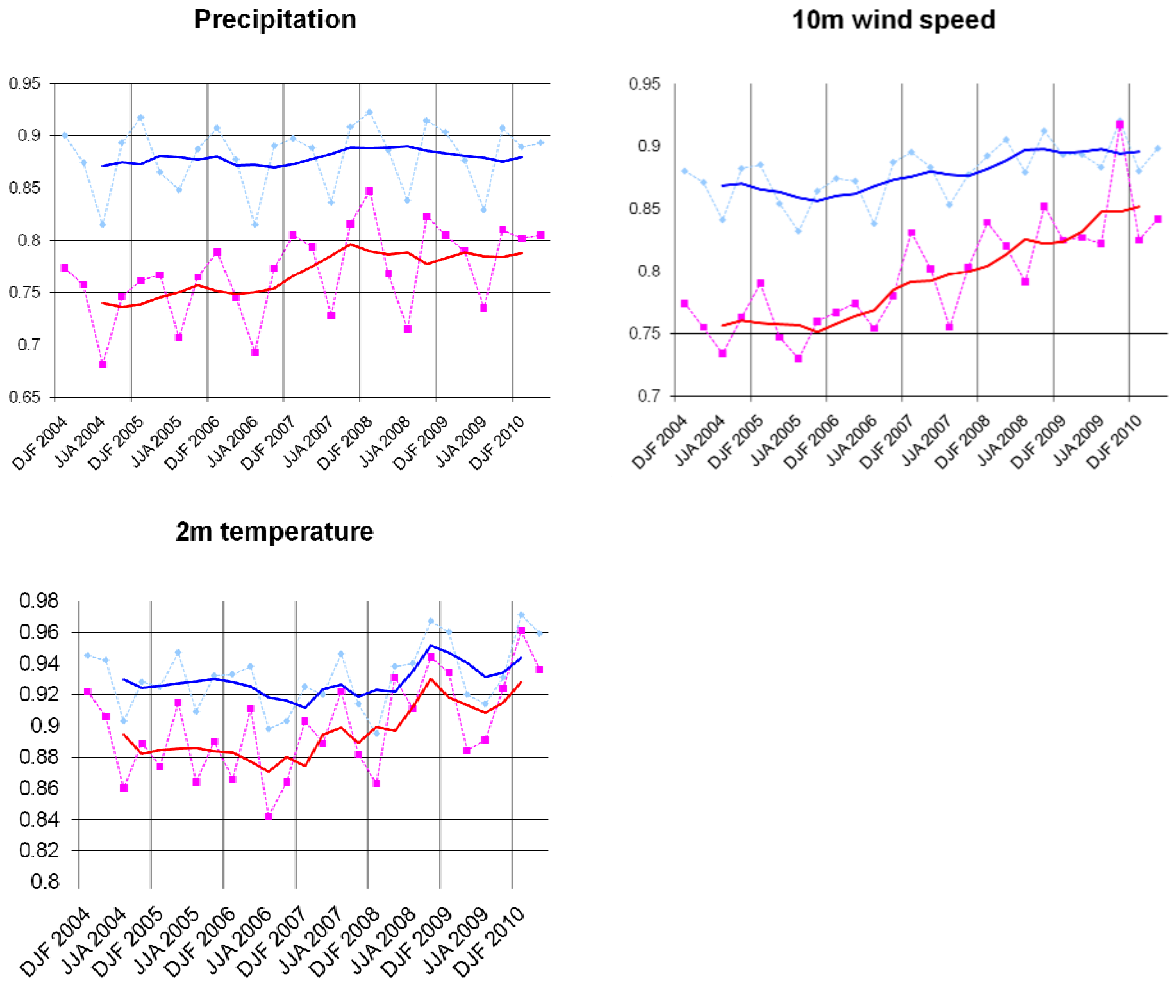


Figure 26: Verification of Extreme Forecast Index (EFI) for precipitation, 10 m wind speed and 2 m temperature over Europe. Extreme event is taken as an observation exceeding 95th percentile of station climate. Hit rates and false alarm rates are calculated for EFI exceeding different thresholds. Curves show the ROC area calculated for each 3-month season from winter (December-February, DJF) 2004 - 2005 to spring (March-May, MAM) 2010 for day 2 (light blue dashed) and day 5 (magenta dashed). Solid lines show running mean of seasonal scores averaged over 4 seasons for: day 2 (blue) and day 5 (red); last point is for average from summer (JJA) 2009 to spring 2010.

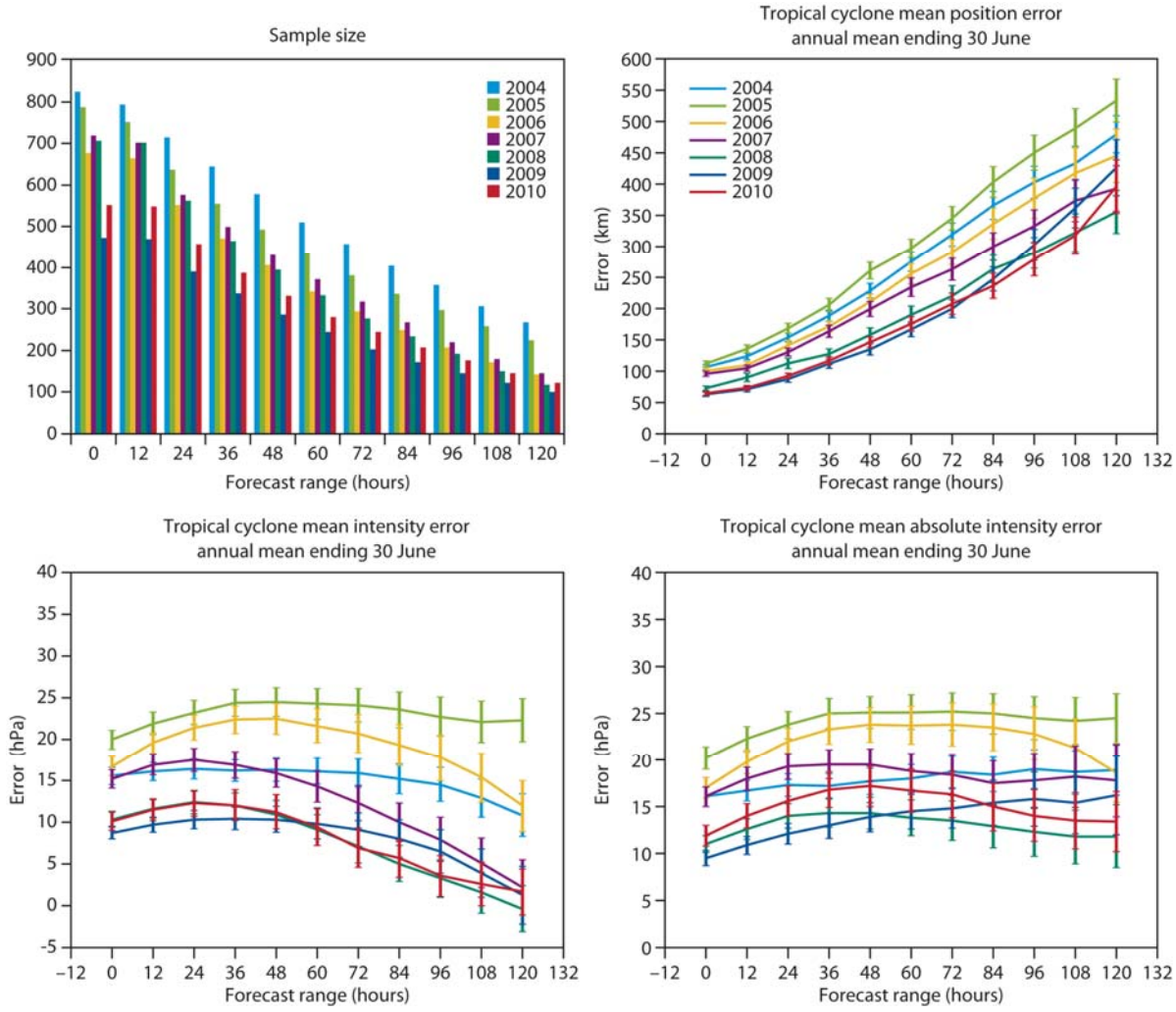


Figure 27: Verification of tropical cyclone predictions from the operational deterministic forecast. Results are shown for 12-month periods ending on 30 June. The latest period, 1 July 2009 to 30 June 2010, is shown in red; other years are coloured as indicated in the legend (same for all panels). Verification is against the observed position reported in real-time via the GTS. The top right panel shows the mean position error (average over all cases of the distance between forecast and observed position; always positive). The bottom left panel shows the 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). The bottom right panel shows the mean absolute error of the intensity. The sample size at each forecast step for each year is shown in the top left panel: there are substantially fewer events at later forecast steps than earlier in the forecast and hence there will be greater uncertainty in the scores at the later ranges; the uncertainty in the scores in indicated by the 90% confidence interval (based on T-test).

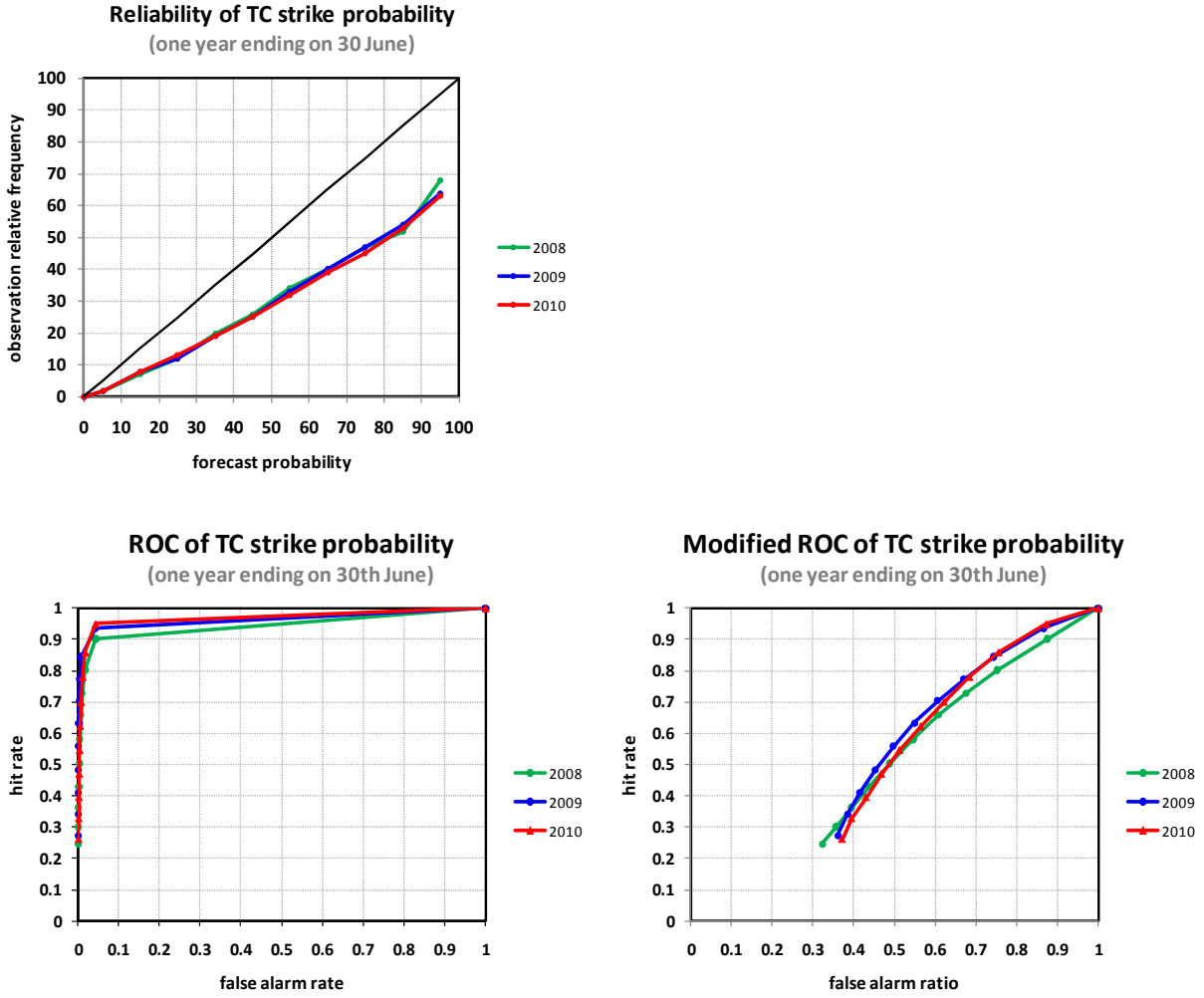


Figure 28: Probabilistic verification of EPS tropical cyclone forecasts for three 12-month periods: July 2007 - June 2008 (green), July 2008 - June 2009 (blue) and July 2009 - June 2010 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) 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).

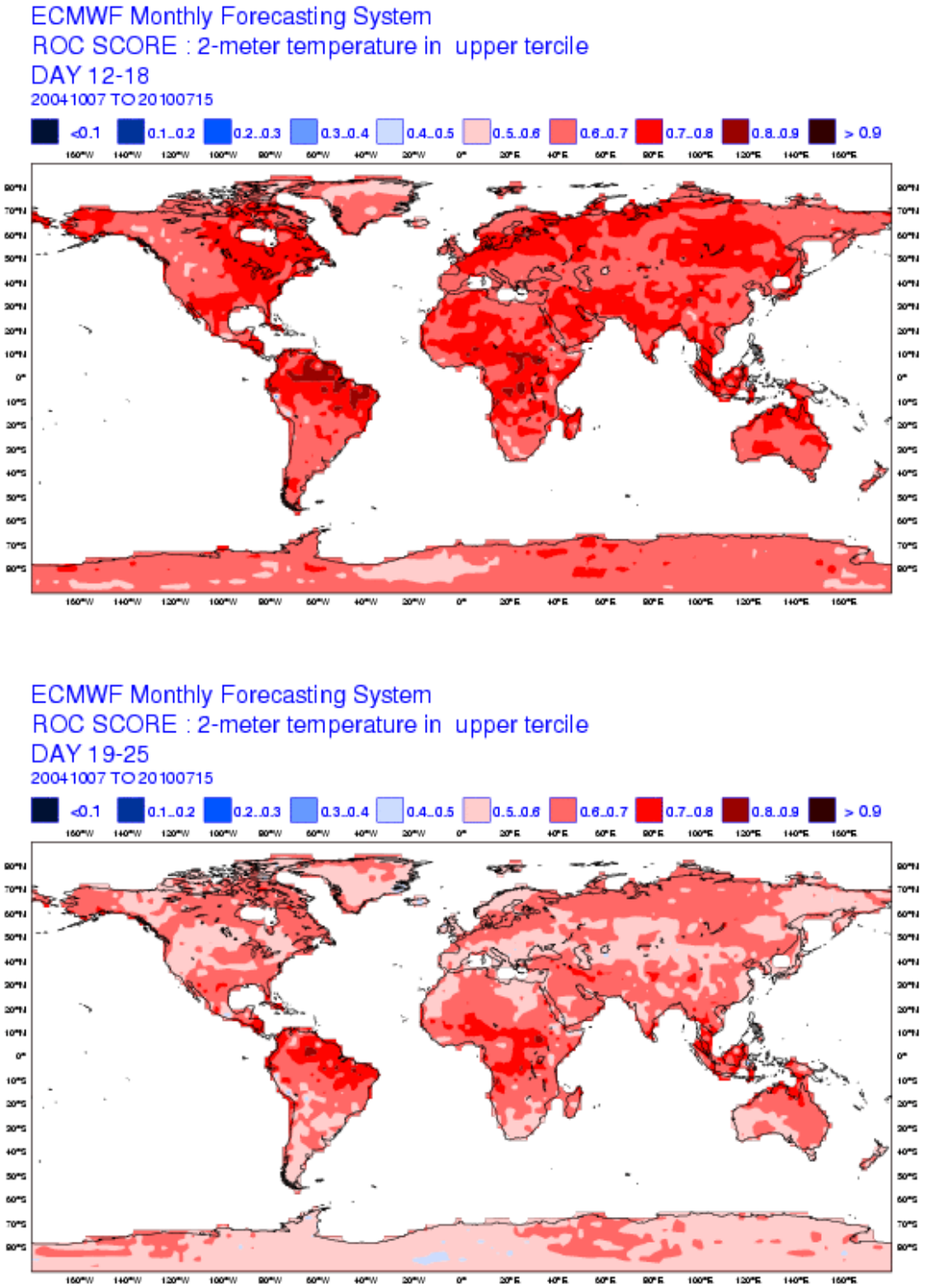


Figure 29: Monthly forecast verification. Spatial distribution of ROC area scores for the probability of 2 m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 15 July 2010 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicates positive skill compared to climate.

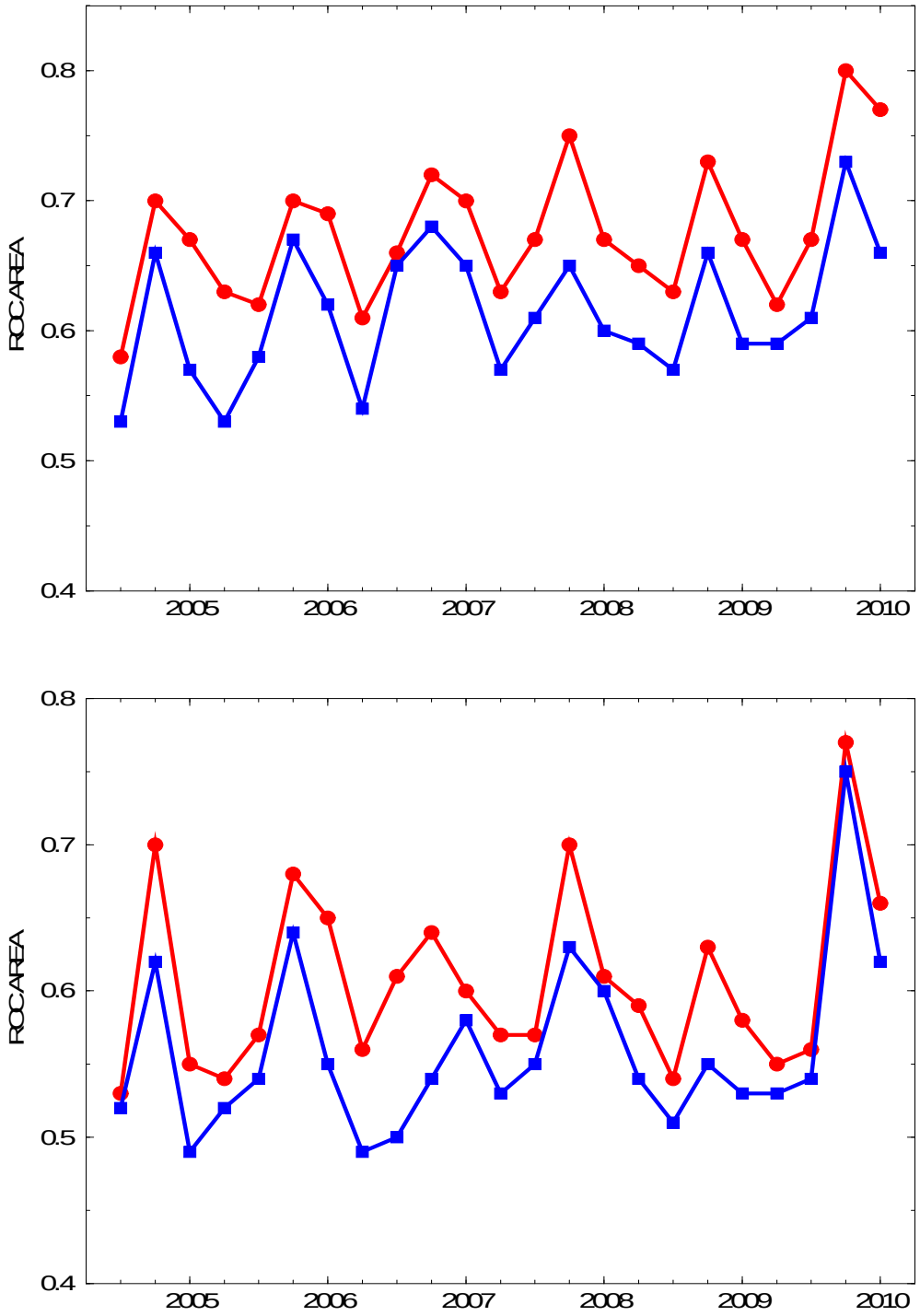


Figure 30: Area under the ROC curve for the probability that 2 metre temperature is in the upper third of the climate distribution. Scores are calculated for each 3 month season since autumn (September-November) 2004 for all land points in the extra-tropical northern hemisphere. The red line shows the 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 2010.

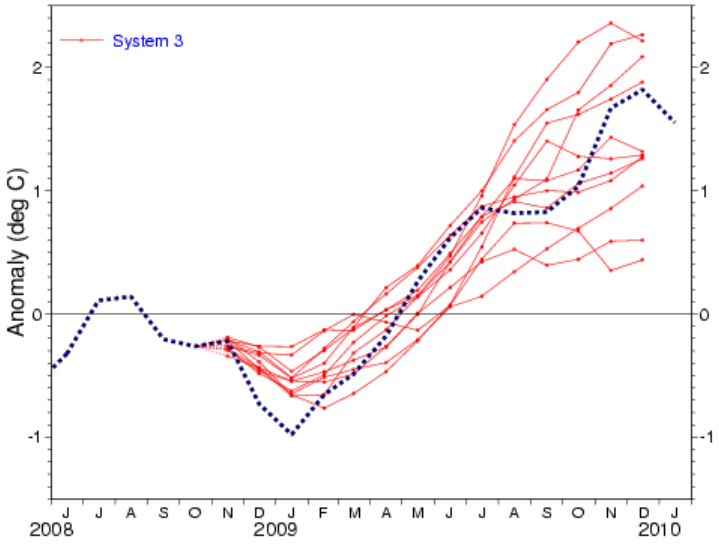


Figure 31: Plot of ECMWF13-month forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from November 2008. The red lines represent the 11 ensemble members; dashed blue lines show the subsequent verification.

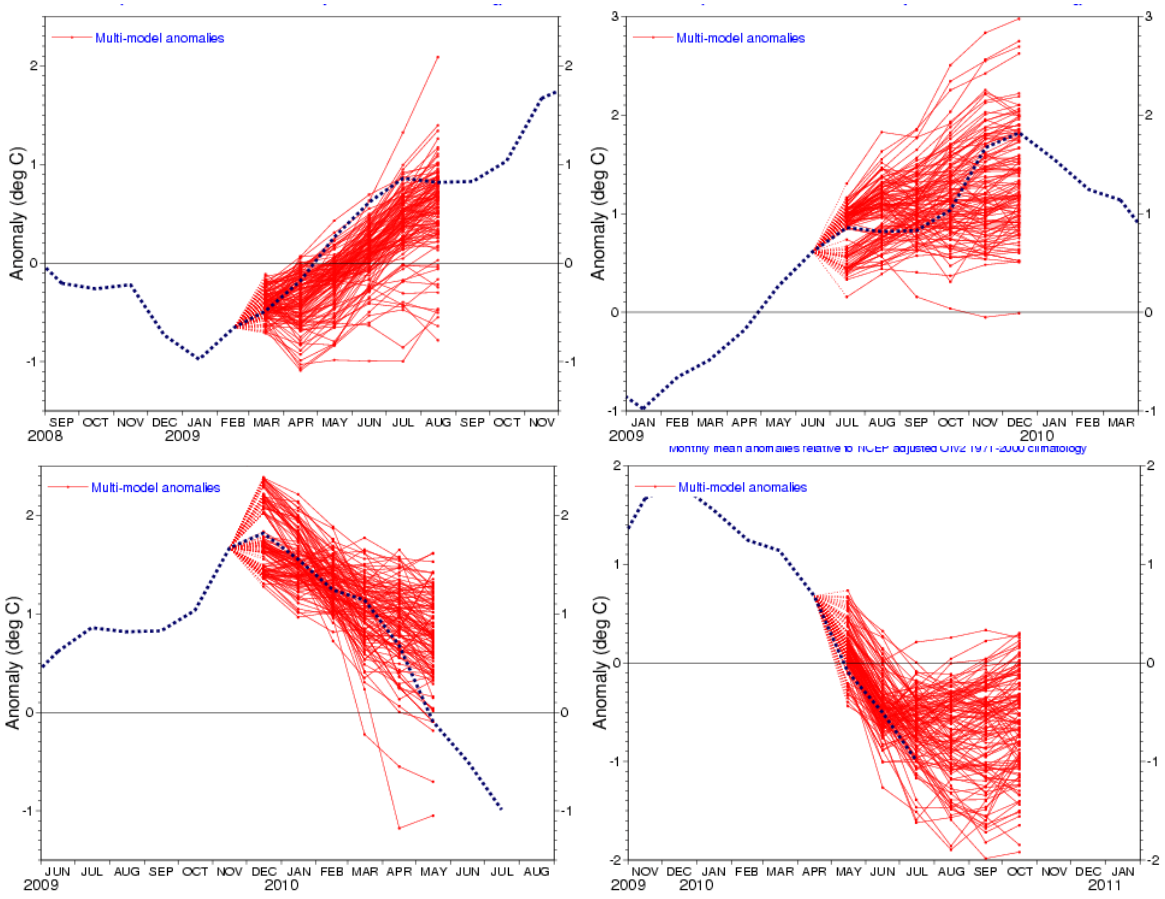


Figure 32: Plot of EUROSIP forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from March 2009 (top left), July 2009 (top right), December 2009 (bottom left) and May 2010 (bottom right). The red lines represent the ensemble members; dashed blue lines show the subsequent verification. EUROSIP comprises seasonal forecast ensembles from ECMWF, Metéo-France and the Met Office.

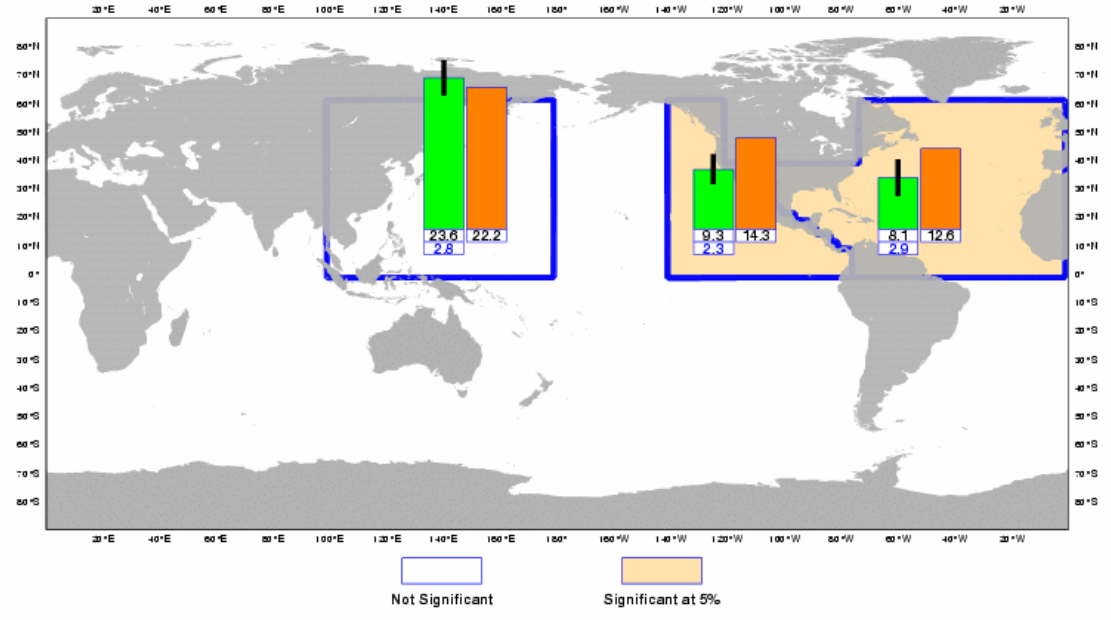


Figure 33: Tropical storm frequency forecast issued in May 2009 for the 6-month period June-November 2009. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The black 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 a significance larger than 90% have a shaded background.

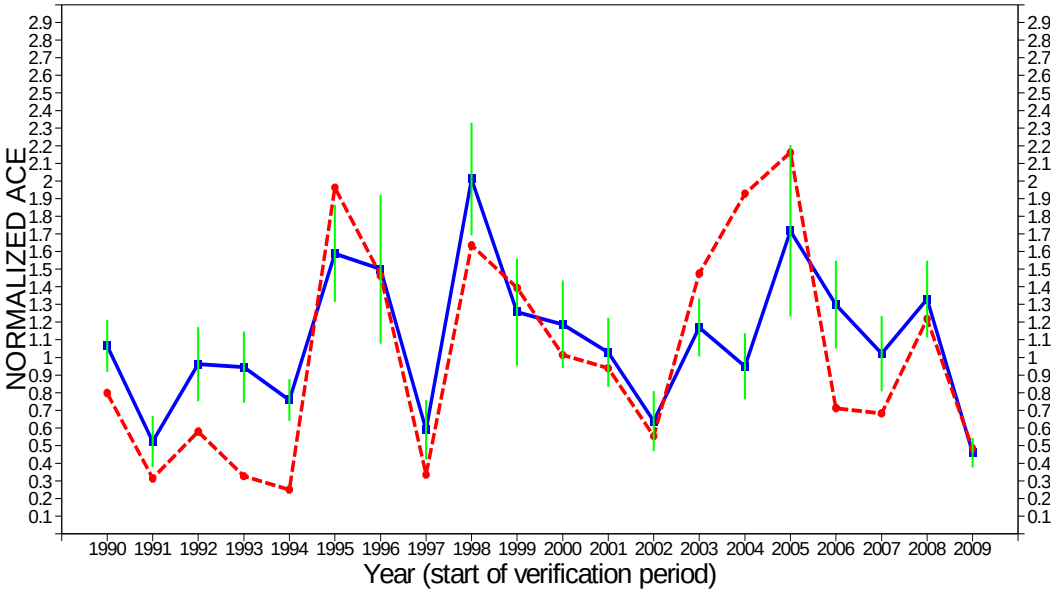


Figure 34: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July – December 1990 to July- December 2009. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation), the red dotted line shows the observation. Forecasts are from ECMWF seasonal forecast system 3: for 1990 to 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.

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 23, Figure 24) 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 21.

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 28). Figure 28 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 21 and Figure 30.

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 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 100 mm, 25 K, 20 g/kg or 15 m/s for precipitation, 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).

For verification of EPS precipitation forecasts against analysis, the 0-24 h 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.