TECHNICAL MEMORANDUM

710

Evaluation of ECMWF forecasts, including 2012–2013 upgrades

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

Forecast Department

November 2013

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

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

© Copyright 2013

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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

1 Introduction

Recent changes to the ECMWF forecasting system are summarised in section 2. Verification results of the ECMWF medium-range free atmosphere forecasts are presented in section 3, including, where available, a comparison of ECMWF's 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 products.

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 system. These new headline scores are included in the current report. As in previous reports a wide range of complementary verification results is 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, 654, 688). 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:

- www.ecmwf.int/products/forecasts/d/charts/medium/verification/ (medium-range)
- www.ecmwf.int/products/forecasts/d/charts/mofc/verification/ (monthly)
- www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/seasonal_range_forecast/ (seasonal)
- www.ecmwf.int/products/forecasts/wavecharts/index.html#verification (ocean waves)

2 Changes to the ECMWF forecasting system

On 25 June 2013, a new model cycle of the Integrated Forecasting System (IFS) was implemented. The cycle contains the long anticipated upgrade of the vertical resolution in the high-resolution forecast model (HRES), the main assimilation (4DVAR), the ensemble of data assimilations (EDA) and the Boundary-Conditions (BC) optional programme from 91 to 137 levels. The model top remains at 0.01 hPa.

A number of other changes were introduced with this cycle to enhance the cycle's performance and to prepare for future upgrades, namely the revision of background error variances at 137 levels based on IFS Cy38r1, the adaptation of EDA calibration and filtering for 137 levels, the deactivation of the model error cycling in the stratosphere, the modification of surface drag parametrization and test parcel entrainment in boundary layer and shallow convection, an adjustment of non-orographic gravity wave drag to be consistent with the seasonal forecast System-4, an oxygen absorption correction in the radiation scheme, the revision of Sea-ice/SST (sea surface temperature) quality control over the Caspian Sea and the correction of the glacier mask over Iceland.

IFS Cy38r2 increases the vertical resolution of the model throughout the troposphere and stratosphere. It enables a better representation of physical processes: clouds, inversions and vertically propagating gravity waves, for example. The forecast impact in terms of objective verification scores against observations and analyses are summarised in the score card (Figure 1).

Tropospheric upper-air scores are overall slightly positive in the northern hemisphere and mainly neutral for Europe and the southern hemisphere. The performance in the tropics is mixed, with some negative results compared to observations but neutral against analyses. In the extratropics the main negative impacts are for the upper-tropospheric relative humidity (300 hPa, Figure 1). The main positive impacts are for geopotential in the lower stratosphere, and to a lesser extent in the troposphere (see also Figure 2).

For precipitation and temperature the overall conclusion is a slight improvement in the extratropics and a slight degradation in the tropics. The scores for 10 m wind show neutral to slightly positive impact in both extratropics and tropics. There is an overall slight reduction of wind speed, most notable in Europe at 12 UTC. No significant differences have been found between the synoptic performance of the pre-operational e-suite and the operational forecast.

Tropical cyclone tracks and intensity have been compared for all tropical cyclones available in the research and pre-operational e-suites. There is a slight improvement for the position errors from day-3 onwards, although this is not statistically significant. The impact is neutral for tropical cyclone intensity.

Later this year (cycle 39r1), the ENS vertical resolution will also be enhanced. It is planned that the ENS will then use the 91 level configuration that was operational for HRES, BC and EDA before 38r2 and that is used also for SEAS (seasonal forecasting System-4). This will reduce the diversity of vertical resolution configurations at ECMWF and thus enhance consistency between the components of the forecasting system (IFS). The ENS model top will thereby be raised to 0.01 hPa (from the current 5 hPa), making it possible to explore more of the medium to long range predictability of large-scale weather changes exerted by the stratosphere.

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 3 shows the evolution of the skill of the high-resolution forecast of 500 hPa height over Europe and the extratropical 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%. In the northern hemisphere and in Europe scores have generally been at a high level, similar to last year. In January 2013 a record score of more than eight forecast days was achieved for Europe. Such winter-time high values are partially the result of large anomalies associated with the negative phase of the North Atlantic Oscillation (NAO). The effects on scores of the year-to-year variations in the predictability of the atmosphere can be accounted for by comparing the operational model performance with that of the ERA-Interim forecasts, which use a fixed version of the ECMWF model and assimilation system. This comparison shows that the high January score in Europe and the northern extratropics was due to higher-than-normal predictability associated with large-scale variability. In the southern hemisphere scores have stabilized on a level slightly below the maximum

that has been reached so far (in the year 2011). Comparison with ERA-Interim shows that the slight decrease after 2011 resulted from atmospheric variability.

Figure 4 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 (RMS) error, normalised with reference to a forecast that persists initial conditions into the future. The last month included in the statistics is July 2013. Figure 5 shows the RMS errors for Europe of the six-day forecast and the persistence forecast (the reference system for Figure 4). The level of activity has further decreased over the last year, and persistence forecast errors have reduced accordingly. This relative improvement in the performance of the reference forecast (persistence) suggests that the drop in skill in the later forecast steps in Europe in Figure 4 is partly due to atmospheric variability.

Figure 6 illustrates the forecast performance for 850 hPa temperature over Europe. The distribution of daily ACC scores for day-7 forecasts is shown for each winter (December–February, top panel) and summer (June–August, lower panel) season since winter 1997–98. For winter 2012–13 the proportion of good forecasts with ACC>70% is ranking second only to the exceptional winter 2009–10. After the exceptionally good scores in summer 2012, scores in summer 2013 went down as expected but still keeping the high level seen in recent years.

Figure 7 shows the time series of the average RMS difference between four- and three-day (blue) and six- and five-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. The level of consistency between consecutive forecasts has further increased in the last year. This is most readily apparent in the northern extratropics for 96–120 h, where both summer and winter values of RMS differences have reached their lowest values so far.

The quality of ECMWF forecasts for the upper atmosphere in the northern hemisphere extratropics is shown through time series of temperature and wind scores at 50 hPa in Figure 8. For both temperature and vector wind, scores have improved compared to last year. This is also true for the southern hemisphere (not shown). The improvement of temperature scores at 50 hPa is mainly driven by a reduction of negative bias due to changes to the cloud ice parameterization in cycle 38r1 (June 2012).

The trend in ensemble performance is illustrated in Figure 9, 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 high-resolution forecast, the ensemble 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 high-resolution forecast. However, the ensemble performance has been consistently high, and for Europe the skill of individual months at the beginning of 2013 has been comparable to the record levels of 2010. A number of changes have been made to the ensemble configuration since 2010, including improvements to both the initial perturbations (cycle 36r2, June 2010) and representation of model uncertainties (cycle 35r3, September 2009; cycle 36r4, November 2010) and the increase in resolution (cycle 36r1, January 2010) and further redefinition of perturbations using the ensemble of data assimilations (cycle 38r1, June 2012). 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 extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 10. The match between the spread and error improved in 2012–13 compared to the previous winter season. The slight under-dispersion of the ensemble for 500 hPa height has been further reduced, although beyond day 11 some over-dispersion has developed. The under-dispersion for temperature at 850 hPa has been further reduced, although uncertainty in the verifying analysis should be taken into account when considering the relationship between spread and error in the first few days.

Figure 11 shows the skill of the ensemble using CRPSS for days 1 to 15 for winter over the extratropical northern hemisphere. The performance in winters 2009–10 and 2010–11 was clearly exceptional. In part, as for the high-resolution forecast, the anomalous flow made some contribution to the high scores. Although not maintaining the same record levels as these two winters, the performance for 2012–13 compares well with previous years and exceeds that of 2011–12.

3.1.2 Tropics

The forecast performance over the tropics, as measured by RMS vector errors of the wind forecast with respect to the analysis, is shown in Figure 12. At 200 hPa (upper panel) the 5-day forecast has continued to improve, and the 1-day forecast has improved relative to last year, although it is still slightly higher than the minimum which was reached in 2003–2004. At 850 hPa (lower panel) error reductions are somewhat more pronounced. Both upper and lower level improvements are in part associated with changes introduced with cycle 38r1 in June 2012.

3.2 ECMWF versus other numerical weather prediction 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 commission for basic systems (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 are now being implemented at participating centres. ECMWF ceased computation of scores using previous procedures in December 2011. Therefore the ECMWF scores shown in this section are a combination of scores using the old (December 2011 and before) and new procedures (for 2012). The scores from other centres for the period of this report have been computed still using the previous procedures. For the scores presented here the impact of the changes is relatively small for the ECMWF forecasts and does not affect the interpretation of the results.

Figure 13 (northern hemisphere extratropics) and Figure 14 (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.

WMO-exchanged scores also include verification against radiosondes over regions such as Europe. Figure 15, 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 16 (verification against analyses) and Figure 17 (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 more similar. At the beginning of 2012 the errors of the ECMWF forecast at 850 hPa have shifted to a slightly lower level due to a change in the computation of the score. Instead of sampling the full fields on a 2.5° grid, fields are now spectrally truncated equivalent to 1.5° resolution, in accordance with WMO guidelines. The errors of the Japan Meteorological Agency (JMA) 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 16) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 17), the ECMWF, UK Met Office and JMA models have very similar short-range errors.

4 Weather parameters and ocean waves

4.1 Weather parameters – high-resolution and ensemble

The supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 18. The upper panel shows the lead time for which the stable equitable error in probability space (SEEPS) skill for the high-resolution forecast for 24-hour precipitation over the extratropics remains above 45%. This threshold has been chosen such that the score measures the skill at a lead time of 3–4 days. The lower panel shows the lead time for which the CRPSS for the probability forecast of 24-hour precipitation over the extratropics remains above 10%. This threshold has been chosen such that the score measures the skill at a lead time of approximately 6 days. Both scores are verified against station observations. The increase in skill of the high-resolution forecast in 2010 was associated with the 5-species prognostic microphysics scheme introduced on 9 November 2010 (cycle 36r4); the increased skill of the ensemble forecast between mid-2009 and mid-2010 was associated with the resolution increase on 26 January 2010 (cycle 36r1). The temporal averaging of the scores leads to step-wise changes in model skill appearing as gradual changes over 12 months in the plots.

The decrease of the score for the high-resolution forecast in 2012 is due to atmospheric variability, as shown by comparison with the ERA-Interim reference forecast (dashed line in Figure 18, upper panel). By taking the difference between the operational and ERA-Interim scores much of this variability is removed, and the effect of model upgrades is seen more clearly (Figure 19). The slight improvement in the probabilistic score (lower panel in Figure 18) is also due to atmospheric variability. The CRPS of the climatology forecast, which is used as a reference for the CRPSS (see Appendix A.2), decreased (i.e. improved) over the period 2010–2011, which has masked improvements due to model upgrades during that time. In 2012, however, this trend has reversed, so that some of the improvement has become visible in the CRPSS.

ECMWF performs a routine comparison of the precipitation forecast skill of ECMWF and other centres for both the high-resolution forecast and the ensemble forecasts using the TIGGE data archived in the Meteorological Archival and Retrieval System (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 20). The reversed ranking of the JMA and National Centers for Environmental

Prediction (NCEP) ensemble forecasts relative to the Met Office at short lead times is due to a greater drop in skill in these models during the northern hemisphere convective season (JJA).

Compared to other global models, the ECMWF precipitation forecast previously showed a relative weakness in the first day of the forecast. It was most visible in the scores for Europe but could also be seen in the extratropics in general. While ECMWF had the best forecast from day 2 onwards, it dropped behind the UK Met Office model at day 1 during the non-convective season. This is no longer the case, primarily due to changes in cloud microphysics introduced with cycle 38r1 in June 2012.

Trends in mean error and standard deviation over the last 10 years of error for 2 m temperature, 2 m dew point, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 21 to Figure 24. Verification is against synoptic observations available on the Global Telecommunication System (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. An exception is the marked change in the 10 m wind speed bias (Figure 24) associated with the introduction of cycle 37r3 in November 2011: the change in surface roughness in this cycle generally reduced 10 m wind speeds over land, resulting in improved bias against observations. This also led to a decrease of the non-systematic error of the 10 m wind speed forecast. Slight improvements in both bias and error standard deviation over the last two years are also apparent for total cloud cover (Figure 23).

A recurring feature of the 2 m temperature forecast is a negative night-time temperature bias over Europe in winter and early spring (Figure 21). Model improvements to address the issue were implemented in November 2011 (cycle 37r3), which led to an overall reduction of the negative night-time bias in Europe in winter by 0.2–0.3 K. However, the problem was not solved entirely, and in winter 2012–13 the negative bias increased again compared to 2011–12 (Figure 25). Comparison of error distributions in these two winters for the operational forecast and ERA-Interim (Figure 26) shows a decrease of skill (reduction of the peak, increases in the tails of the distribution) which can be attributed to atmospheric variability. However, the overall shift towards the cold side of the distribution is seen in the operational forecast but not in ERA-Interim.

A problem of the 2 m temperature forecast, which has recently been investigated, is a too-rapid afternoon cooling during spring, leading to substantial negative biases at 12, 15, and 18 UTC in a belt centred around 60 N on the Eurasian continent. Figure 27 shows how the negative errors generally progress westwards, in accordance with local time. The negative afternoon bias is most pronounced in forested, snow-covered areas, and preliminary investigations have shown that it results from the way 2 m temperature is computed for open areas (low vegetation tiles) within forested grid boxes.

To complement the evaluation of surface weather forecast skill, routine verification of radiation and cloudiness using satellite data has been established (see also ECMWF Newsletter No. 135). Here we show results obtained for verification against the top of the atmosphere (TOA) reflected solar radiation products (daily totals) from the Climate Monitoring Satellite Application Facility (CM-SAF) based on Meteosat data. Fluxes have been made non-dimensional and normalized by scaling with a latitudinally and seasonally varying clear-sky flux at the surface. Figure 28 shows the mean error and standard deviation of the error at forecast day 3 of the TOA reflected radiation for the year 2012. As

in previous years the largest negative bias (underestimation of cloud cover and/or cloud reflectance) is found in the Southern Ocean and in the subtropical stratocumulus (Sc) regions. A positive bias is present in areas dominated by trade cumulus (Cu). The existence of biases of different sign in Sc and Cu cloudiness has been a longstanding problem that is being addressed through changes made to the formulation of boundary-layer cloud triggering (introduced in cycle 38r2) and cloud microphysics (Ahlgrimm and Forbes, 2013).

Verification against CM-SAF TOA reflected solar radiation shows a reduction of the non-systematic error relative to ERA-Interim in recent years, both in the extratropics and tropics (Figure 29), that can be attributed to the combined effect of a series of model changes beginning with the introduction of the five-species prognostic microphysics scheme in November 2010 (cycle 36r4). A decrease in error standard deviation for cloud cover during daytime is also noticeable in the verification results against SYNOP observations (Figure 23).

4.2 Ocean waves

The quality of the ocean wave model analysis is shown in the comparison with independent ocean buoy observations in Figure 30. The top panel of Figure 30 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 wind fields for the forcing of the ocean wave model. Similar to the decrease in 10 m wind speed forecast error over land (Figure 24), wind errors are consistently low from 2010 onwards. Errors in the wave analysis have also been consistently low over the last year although they did not reach the record low values of winter 2011–12. The long-term trend in the performance of the wave model forecasts is shown in Figure 31 and Figure 32. The general trend of increasing performance in both hemispheres continues and has become slightly stronger over the last two years.

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 33 for the most recent 12-month period (September 2012 – August 2013). ECMWF forecast winds are used to drive the wave model of Météo France; the wave models of the two centres are similar, hence the closeness of their errors in Figure 33. ECMWF outperforms the other centres with regard to wind speed and wave height, while Météo France has the highest skill in forecasting the peak period, followed by ECMWF and JMA. Of the centres not using ECMWF wind fields, the UK Met Office and the National Centers for Environmental Prediction (NCEP) have the lowest errors for both wind speed and wave height; the NCEP forecasts have improved significantly because of recent improvements to the atmospheric model and the introduction of a new ocean wave model in May 2012.

A comprehensive set of wave verification charts is now available on the ECMWF website:

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 high-resolution forecast (Section 5.2)

5.1 Extreme Forecast Index (EFI)

The Extreme Forecast Index (EFI) was developed at ECMWF as a tool to provide early warnings for potential extreme events. By comparing the ensemble 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 34 (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 2012; the final point on each curve includes the spring (March–May) season 2013. For all three quantities EFI skill is maintained at a similar level to the previous year.

5.2 Tropical cyclones

The 2012 North Atlantic hurricane season had an above average number of tropical storms (15 compared to 12 in the climate mean, see also Figure 42). The tropical cyclone position error for the three-day high-resolution forecast is one of the supplementary headline scores for severe weather. The average position errors for the high-resolution medium-range forecasts of all tropical cyclones (all ocean basins) over the last ten 12-month periods are shown in Figure 35. Errors in the forecast intensity of tropical cyclones, represented by the reported sea-level pressure at the centre of the system, are also shown.

The position errors (top panel, Figure 35) have reached their smallest value so far (180 km) for the three-day forecast. This represents about 10% improvement compared to the ~200 km error which was typically seen in recent years. The bottom left panel of Figure 35 shows the average speed error for tropical cyclones for the last three years. Typically tropical cyclones move too slowly in the forecast (by around 1 km/h) compared to the observed speed; while the previous 12-month period showed no overall bias in speed of movement, it has increased again to about -1.5 km/h. However, because of the substantial year-to-year variations in the number and intensity of cyclones, there is some uncertainty in these figures. In spite of the increased negative bias the corresponding mean absolute speed error (bottom right panel) has decreased. Both the mean error (bias) and mean absolute error in tropical cyclone intensity (centre panels in Figure 35) have increased slightly compared to last year but they remain at a generally low level. As with the speed errors, there is a relatively large uncertainty in these scores because of the year-to-year variations in the number and character of storms.

The ensemble 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 36. Results show over-confidence for the three periods, with small variations from year to year. Compared to the previous two years reliability has decreased. The skill is shown by the ROC and the modified ROC, which uses the false alarm ratio (fraction of yes forecasts that turn out to be wrong) instead of the false alarm rate (ratio of false alarms to the total number of non-events) 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. With regard to ROC, the performance was in between those of the previous two years.

5.3 Additional severe-weather diagnostics

In order to extend severe-weather diagnostics at ECMWF, additional scores are being tested. While most standard scores tend to degenerate to trivial values for rare events, some have been specifically designed to address this issue. The latest among these is the symmetric extremal dependence index, SEDI (Ferro and Stephenson, 2011; see Annex A.4), which is used here to compare the skill of the operational high-resolution and the ERA-Interim forecast. Both forecasts are verified against station observations. Figure 37shows the skill in forecasting events above the 98th climate percentile in Europe. For 24-h precipitation (left panel), the gain in skill of the operational forecast relative to ERA-Interim is about one forecast day and it is mainly due to a higher hit rate. For 10 m wind speed the gain is close to two forecast days and it is mainly due to a lower false alarm rate. Forecast skill as measured by SEDI for high percentile events is generally higher for 24-h precipitation than for 10 m wind speed.

Whereas SEDI is computed based on calibrated forecasts and therefore measures potential skill, the potential economic value (PEV) gives the actual skill in terms of the gain (relative to climatology) obtained by performing action and non-action following the forecast guidance (see Annex A.4). As in the case of SEDI, it is applied here to the 98th percentile of 24-h precipitation and 10 m wind speed in Europe. Figure 38 shows that for the HRES forecast the maximum PEV on forecast day 1 is about 0.3 for both precipitation, and wind. At longer lead times positive PEV values exist only for a narrow range of cost–loss ratios. However, the PEV of the ENS forecast is considerably higher, especially for precipitation. This indicates the potential benefit of probabilistic forecast information in decision making. As with SEDI, forecasts of 10 m wind speed are less skilful overall than 24-h precipitation.

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 since March 2008. The combined system made it possible to provide users with ensemble output uniformly up to 32 days ahead, once a week. A second weekly run of the monthly forecast was introduced in October 2011, running every Monday (00 UTC) to provide an update to the main Thursday run.

Figure 39 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 20-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 40 shows the probabilistic performance of the monthly forecast over each individual season since September 2004 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 extratropical northern hemisphere. Both for the 12–18 day and 19–32 day periods, the advantage over persistence of the medium-range (days 5–11) forecast has increased compared to the previous year. Even for the 19–32 day range, the system shows a substantial lead compared to persistence of the 5–18 day forecast for all 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 NAO conditions of that winter.

6.2 Seasonal forecast performance

6.2.1 Seasonal forecast performance for the global domain

A new version (System 4) of the seasonal component of the IFS was implemented in November 2011. System 4 uses a new ocean model (NEMO instead of HOPE) and a more recent version of the ECMWF atmospheric model (cycle 36r4) run at higher resolution. The forecasts contain more ensemble members (51 instead of 41) and the re-forecasts have more members (15) and cover a longer period (30 years instead of 25).

A set of verification statistics based on re-forecast integrations (1981–2010) from System 4 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 comprehensive description and assessment of System 4 is provided in ECMWF Technical Memorandum 656, available from the ECMWF website:

http://www.ecmwf.int/publications/library/do/references/show?id=90277

6.2.2 The 2012–2013 El Niño forecasts

The 2011 La Niña declined during the second half of 2011 and sea surface temperatures in the tropical Pacific had changed to warm (El Niño) conditions by June 2012 (Figure 41). Probably due to the reduced amplitude of these anomalies, the atmospheric circulation never evolved into a state of El Niño conditions. Forecasts in 2011 captured well the change from La Niña to El Niño but only very few members indicated the short-lived (~6 months) nature of the warm episode. From December 2012 to the present the anomalies evolved into slightly cold conditions, while the majority of ensemble members suggested a return to warm anomalies. However, observations were generally within the range of uncertainty provided by the ensemble forecast. The multi-model EUROSIP forecasts performed slightly better in the sense that the ensemble was better centred on the observations but otherwise exhibited the same weaknesses.

6.2.3 Tropical storm predictions from the seasonal forecasts

Seasonal tropical storm predictions indicated slightly reduced activity (compared to climatology) over the Atlantic for the 2012 season, consistent with the prediction of a moderately strong El Niño episode. The June forecast predicted between 7 and 15 named tropical storms in the Atlantic, which is a statistically significant signal at the 90% level (Figure 42). During the 2012 season 15 named tropical storms occurred, exceeding the climatological mean (12). The failure of the forecast to predict above-normal activity is related to the overestimation of the strength of the El Niño episode (Figure 41).

The seasonal forecasts indicated slightly weaker than normal tropical storm activity also for both the western and eastern Pacific. For the western Pacific this was correct, while for the eastern Pacific slightly above normal activity was observed, similar to in the Atlantic basin. With System 4 the skill in predicting accumulated cyclone energy (ACE) over the Atlantic basin, calculated using the most recent 20 years, has increased, with a correlation of 0.72 between ensemble mean forecast and observation (Figure 43). Skill in the western Pacific has also increased with correlations approaching a value of around 0.7.

6.2.4 Extratropical seasonal forecasts

Europe, Central Asia, and North America experienced extreme cold conditions during March 2013. Seasonal (March–May) temperature anomalies, which exceeded 1.5 standard deviations from the 30-year climate mean in several regions (Figure 44, lower panel), were associated with the persistence of the negative phase of the NAO. While the extreme cold over Europe in March was well predicted 12–18 days in advance by the monthly forecasting system, for the seasonal mean both ECMWF's system 4 and the Eurosip multi-model systems failed to give a strong signal. A number of potential drivers such as tropical SST anomalies, sudden stratospheric warmings and Madden-Julian Oscillation activities, and the state of the sea-ice in the Arctic were analysed. Results obtained so far indicate that all the above drivers contributed to the evolution of the cold anomalies.

7 References

Ahlgrimm, M., and R. Forbes, 2013: Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. Mon. Wea. Rev., 141, (accepted).

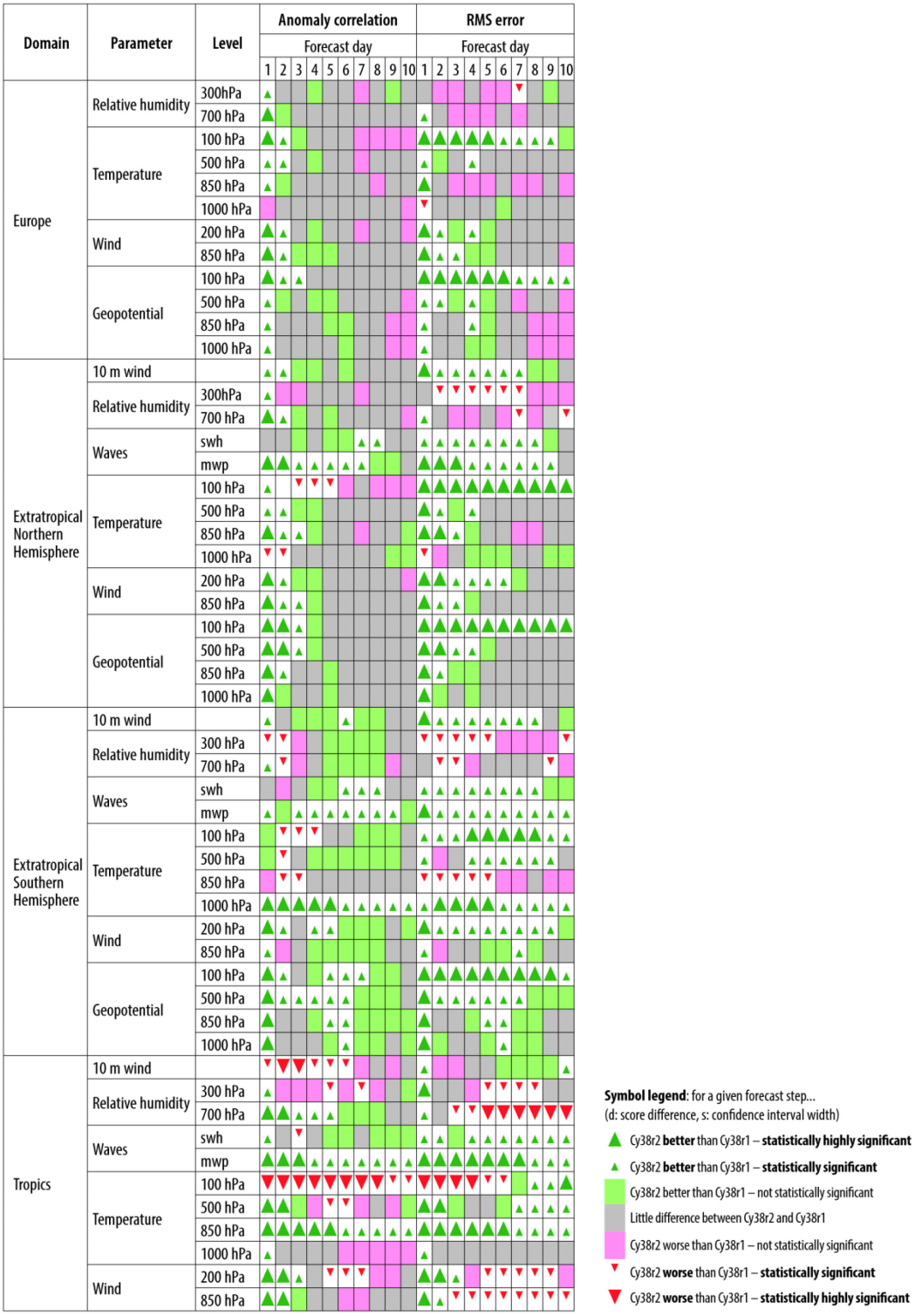


Figure 1: Scorecard for cycle 38r2 versus 38r1 high-resolution forecasts verified by the respective analysis at 00 and 12 UTC.

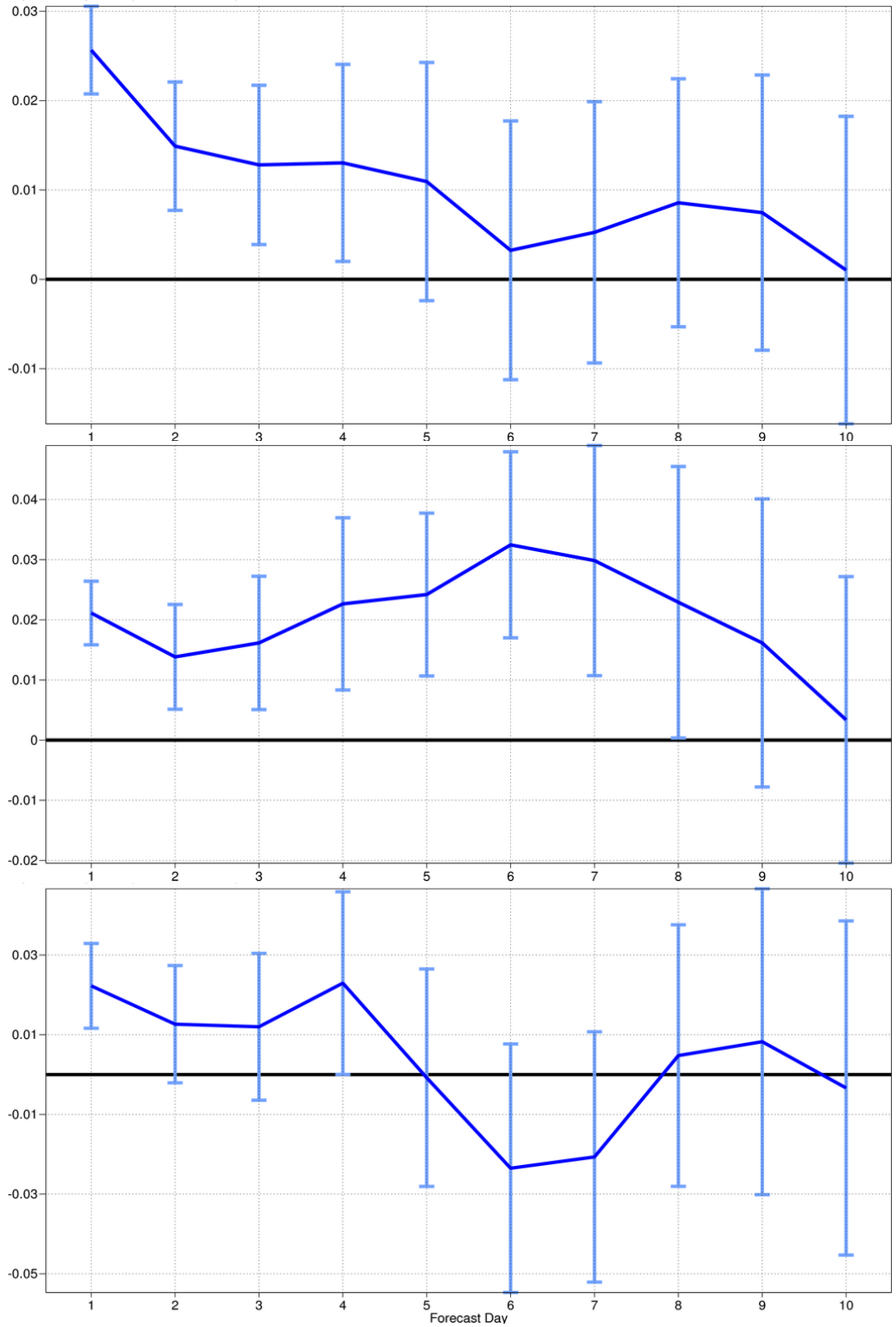


Figure 2: Normalised difference in root mean square error between cycle 38r2 and 38r1 for 500 hPa height high-resolution forecast over northern hemisphere extratropics (top), southern hemisphere extratropics (middle) and Europe (bottom) from combined testing in RD and OD (440 cases). Positive values indicate improvements for 38r2.

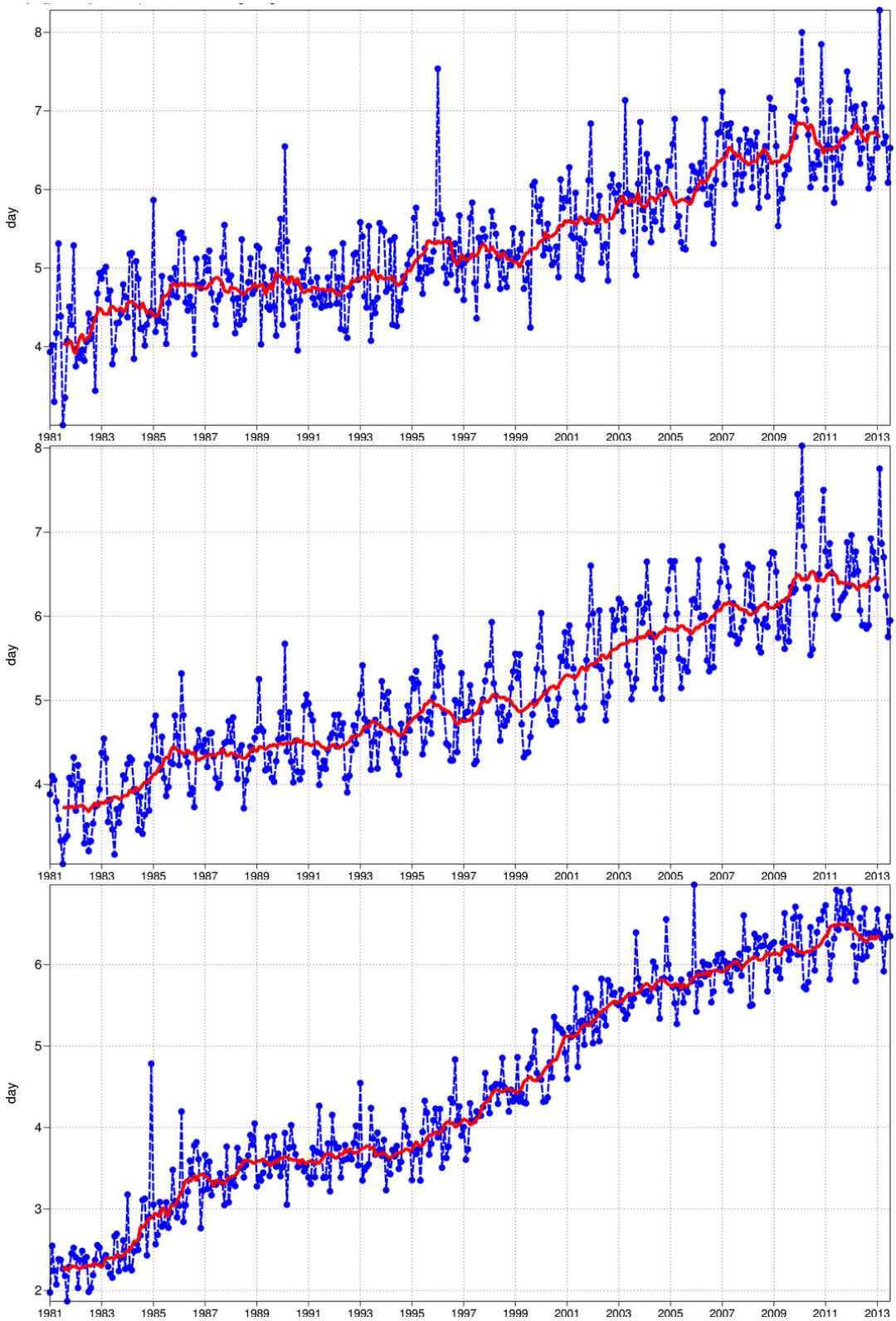


Figure 3: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).

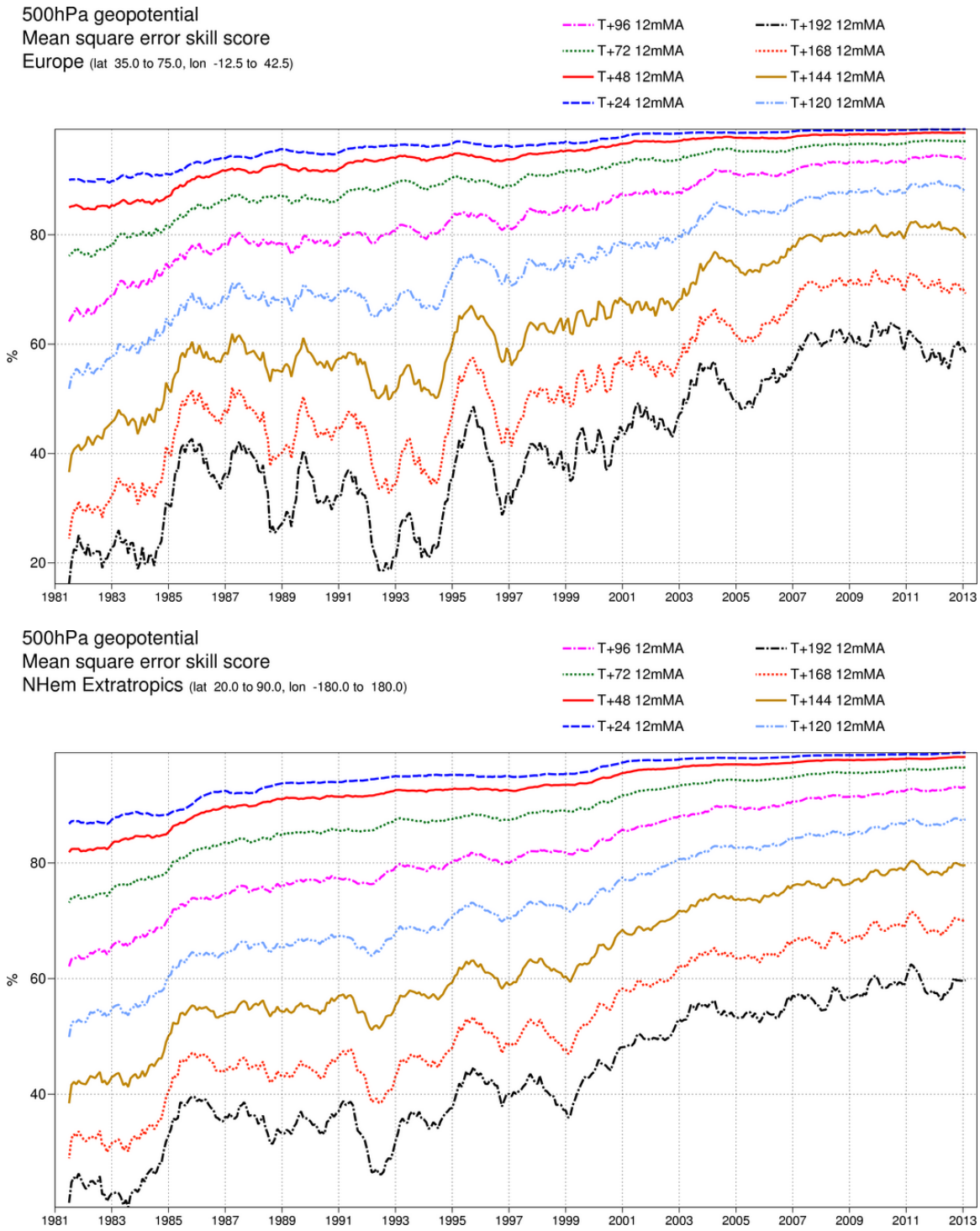


Figure 4: 500 hPa geopotential height skill score for Europe (top) and the northern hemisphere extratropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2012–July 2013.

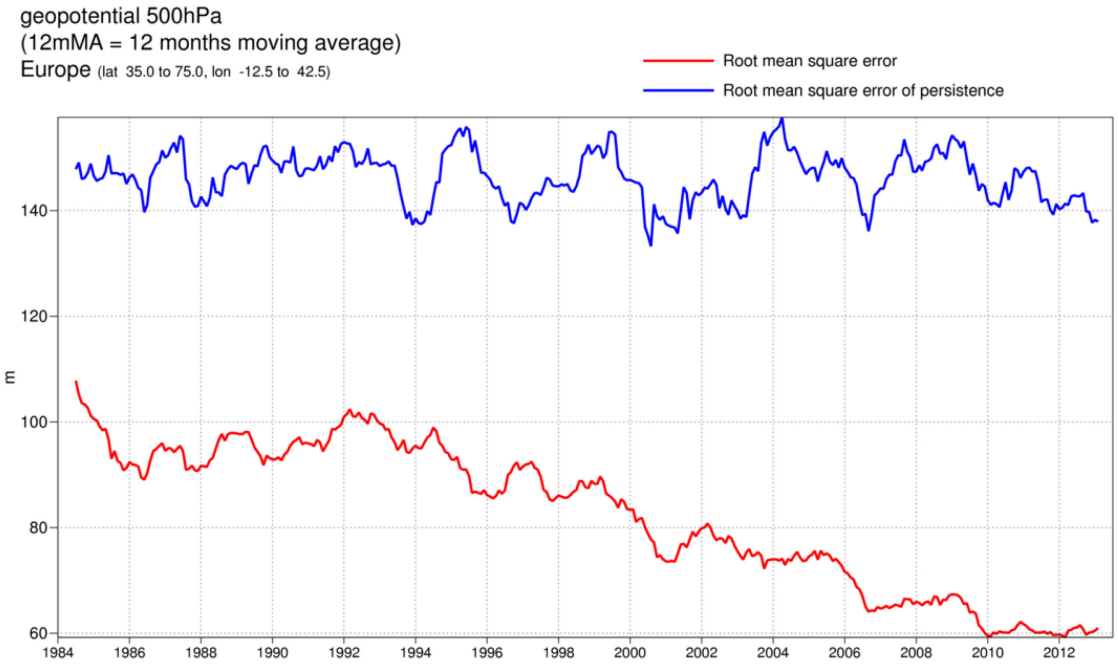


Figure 5: Root mean square (RMS) error of forecasts made by persisting the analysis over 6 days (144 hours) and verifying it as a forecast for 500 hPa geopotential height over Europe (blue). The RMS error of the forecast at day 6 is shown in red. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2012–July 2013.

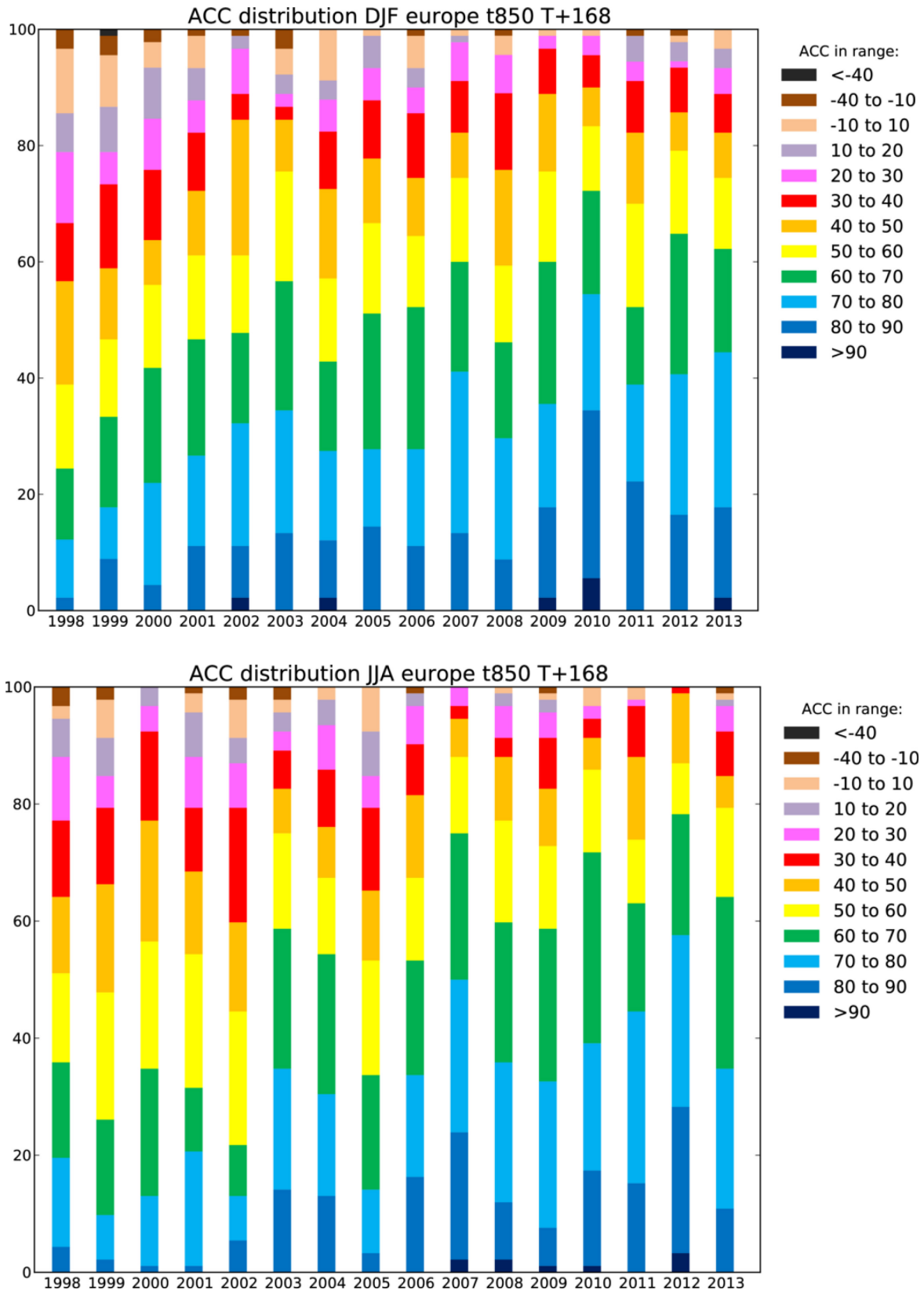


Figure 6: Distribution of ACC of the day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997–1998.

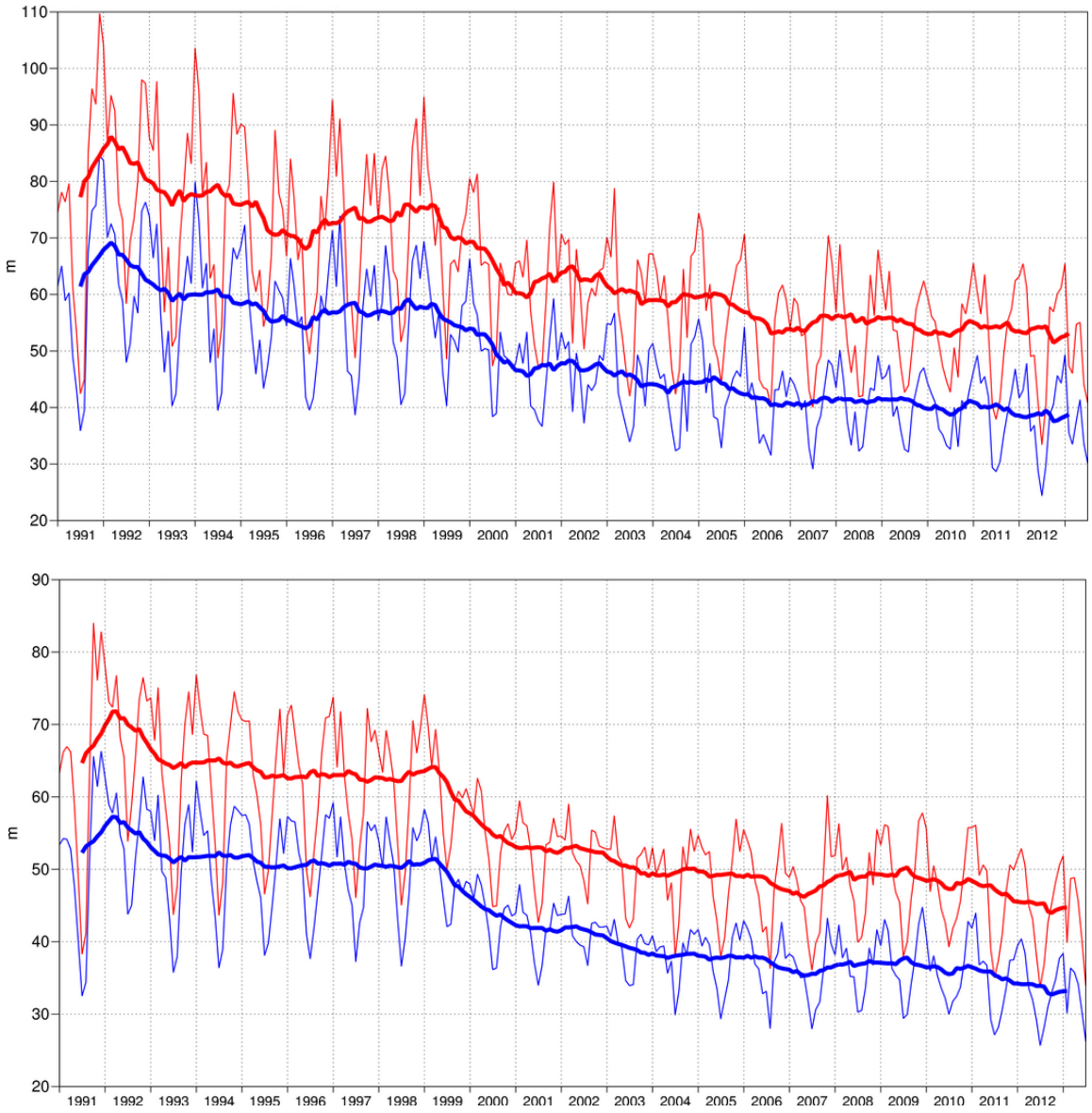


Figure 7: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores are also shown (in bold).

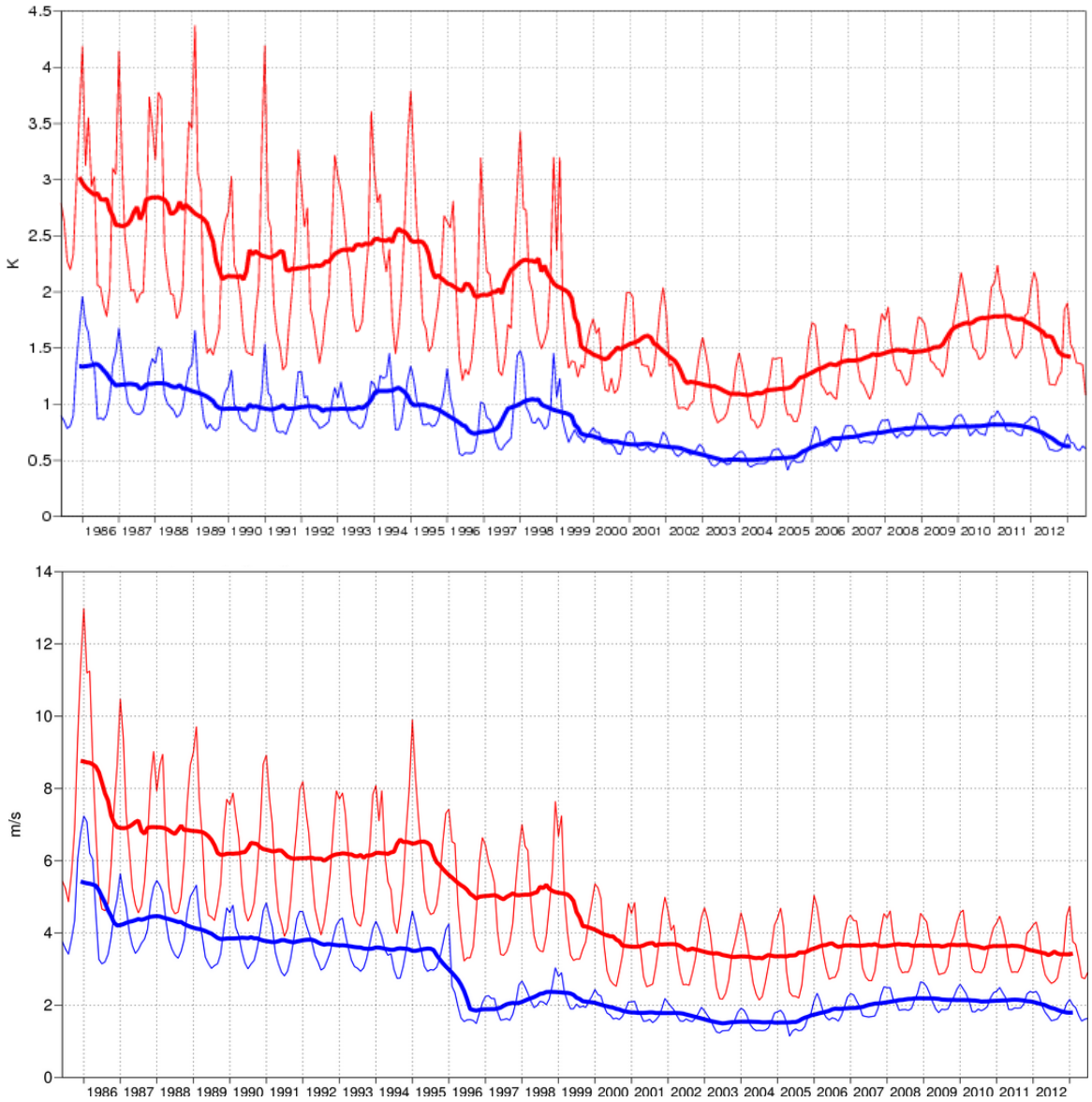


Figure 8: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).

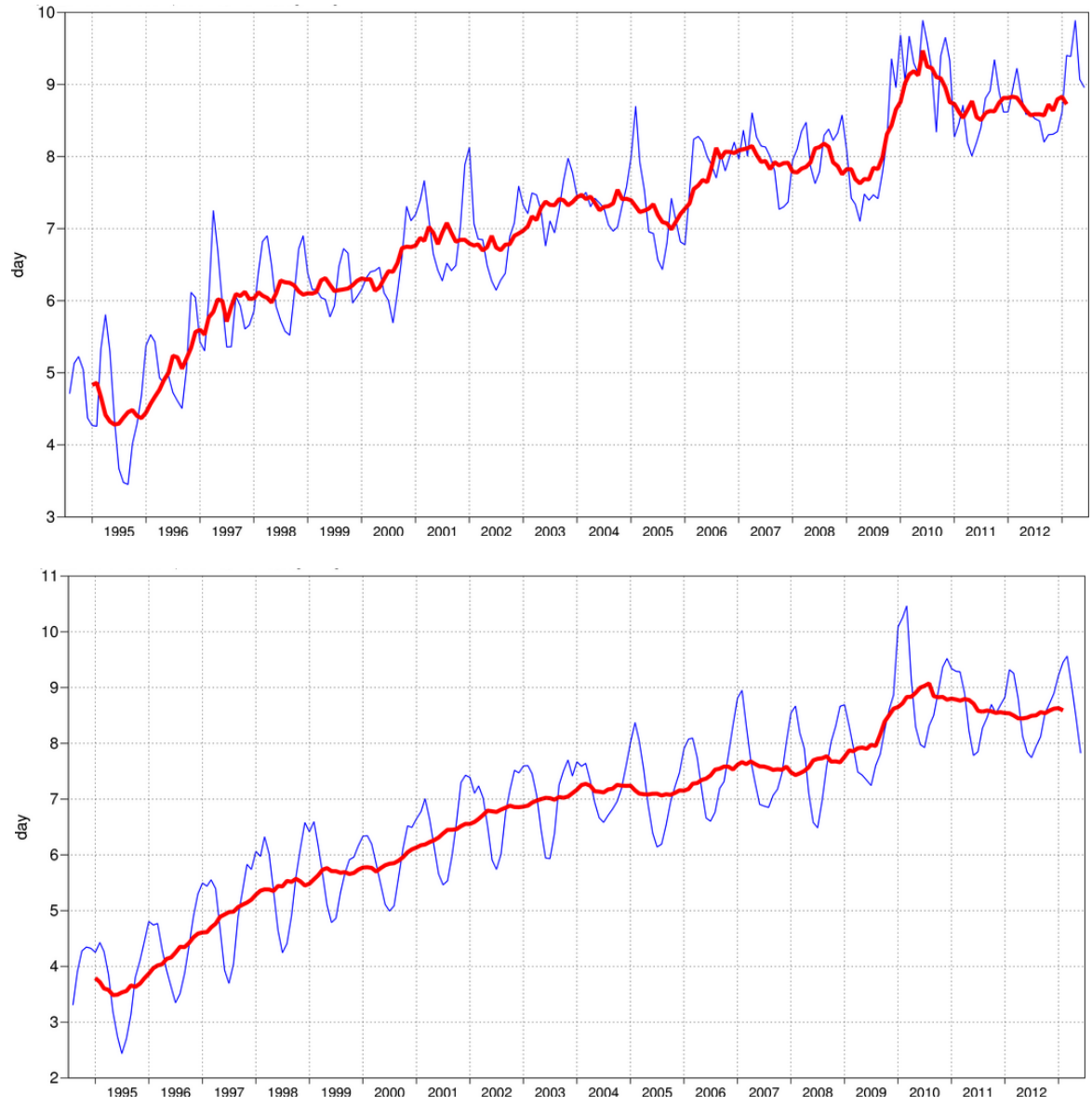


Figure 9: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance – each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous ranked probability skill score (CPRSS) falls below 25% for Europe (top), northern hemisphere extratropics (bottom).

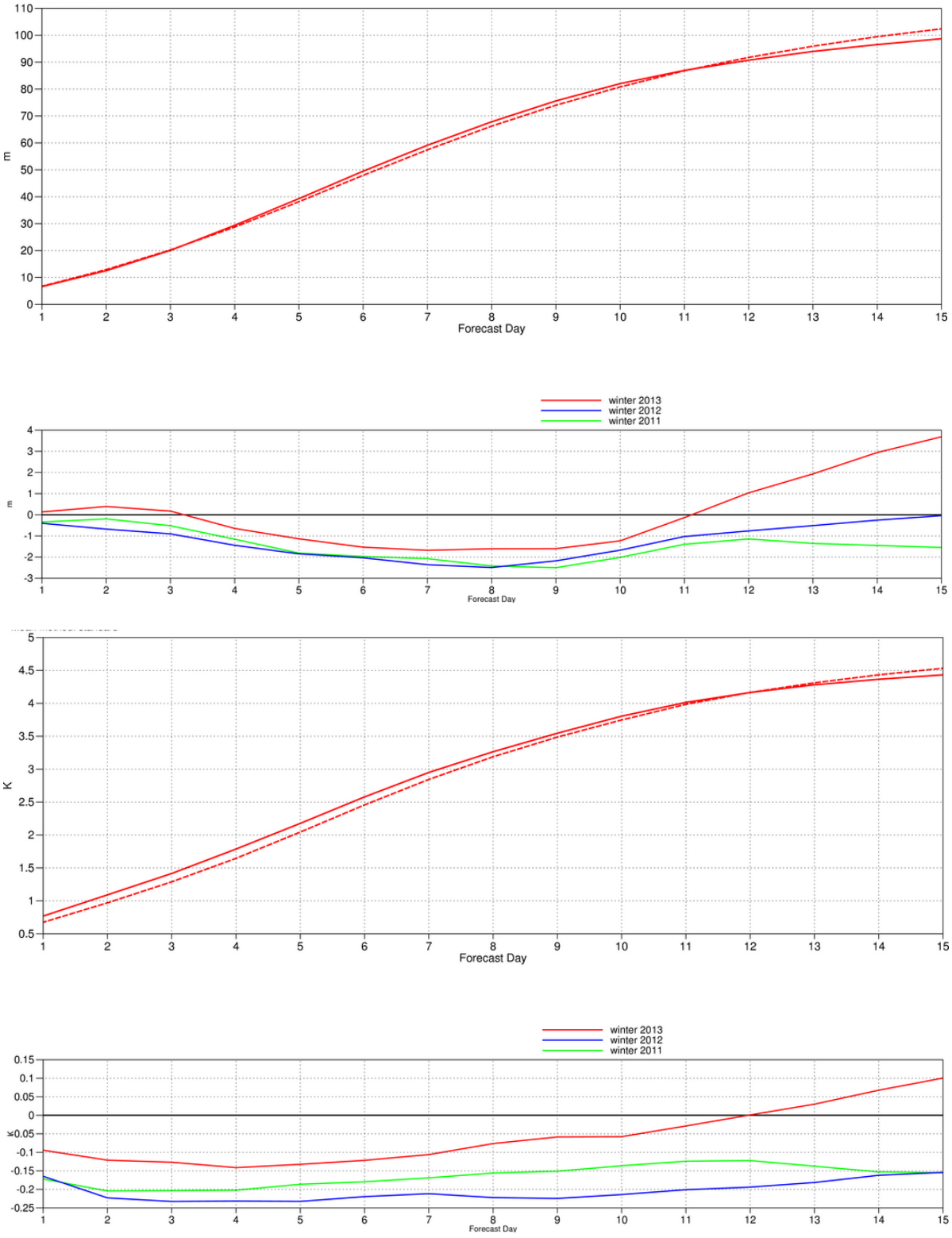


Figure 10: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2012–2013 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate spread is too small); plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extratropical northern hemisphere for forecast days 1 to 15.

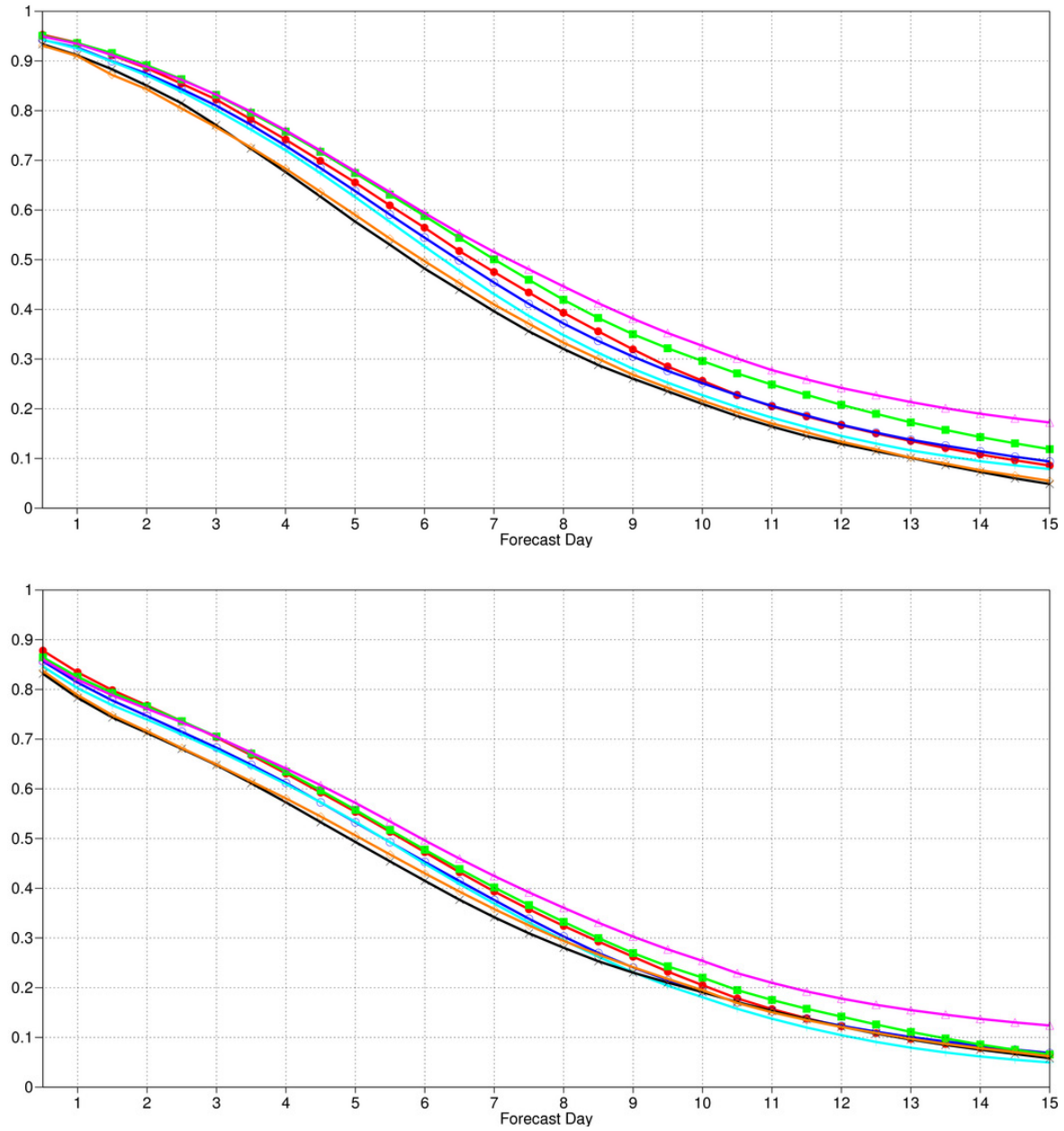


Figure 11: CPRSS for 500 hPa height (top) and 850 hPa temperature (bottom) ensemble forecasts for winter (December–February) over the extratropical northern hemisphere. Skill from the ensemble day 1–15 forecasts is shown for winters 2012–13 (red), 2011–12 (blue), 2010–11 (green), 2009–10 (magenta), 2008–09 (cyan), 2007–08 (black) and 2006–07 (orange).

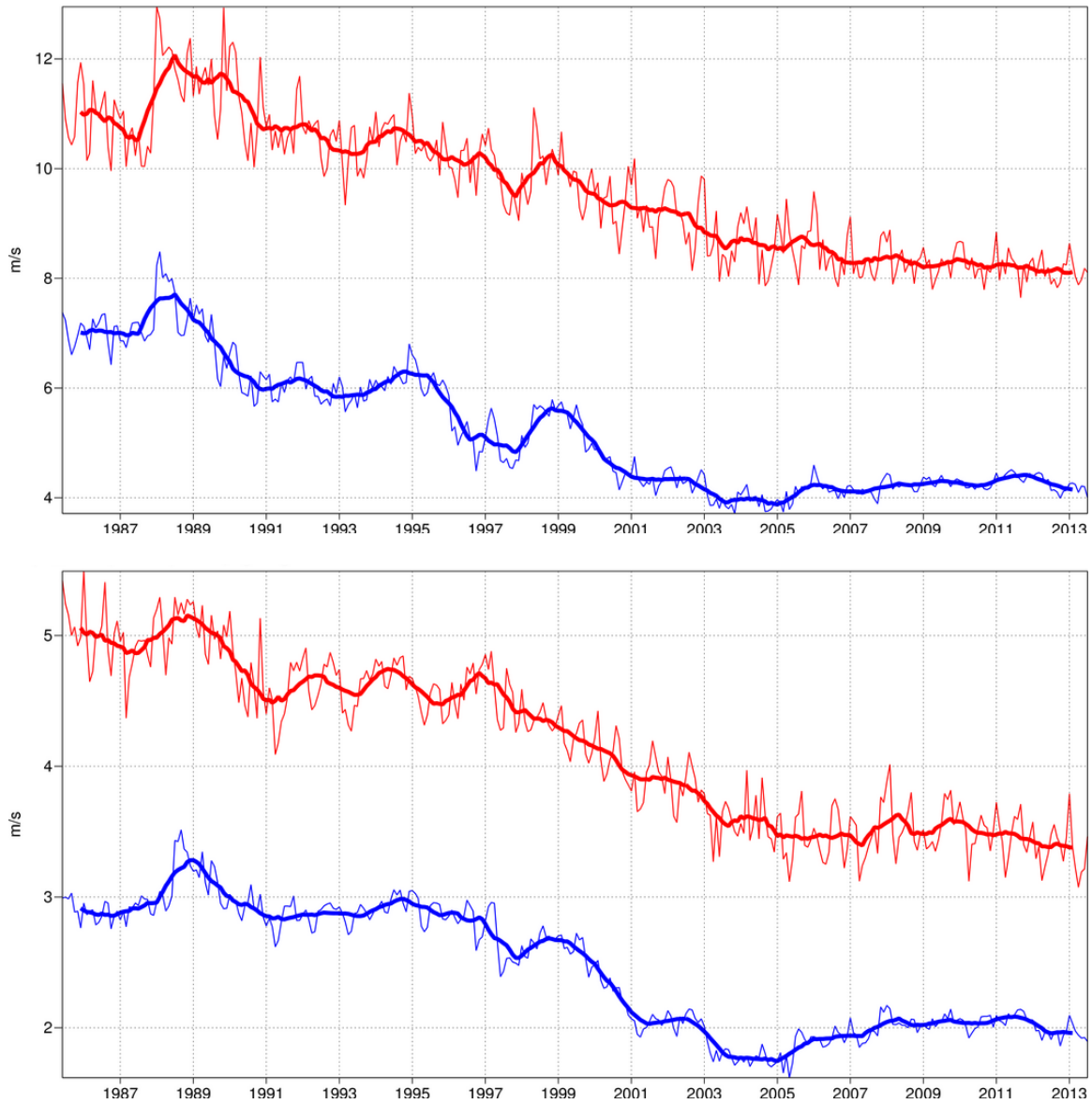
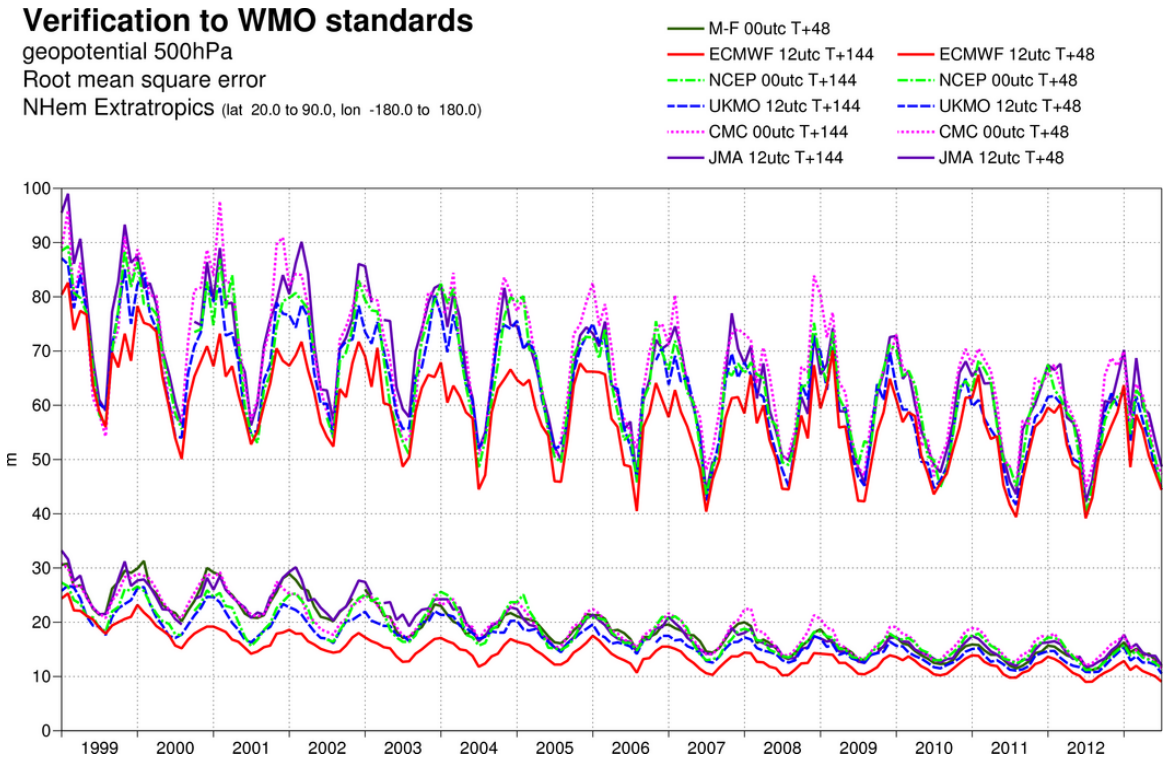


Figure 12: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).



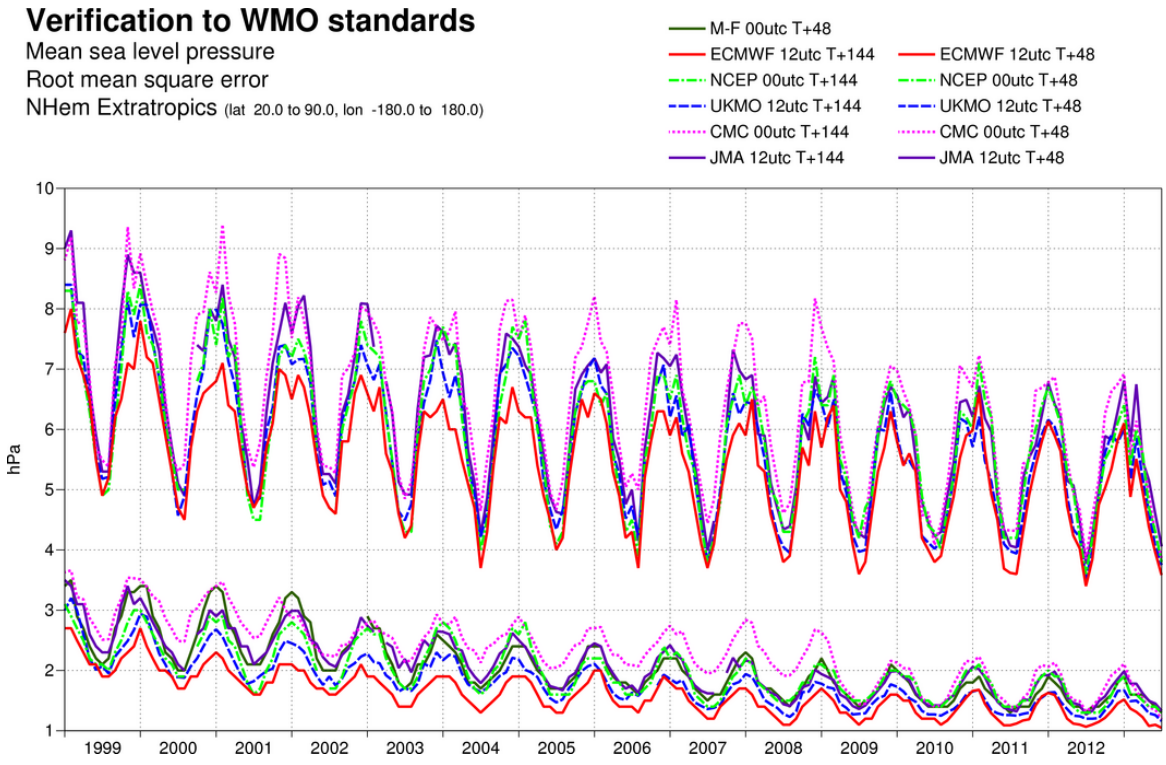
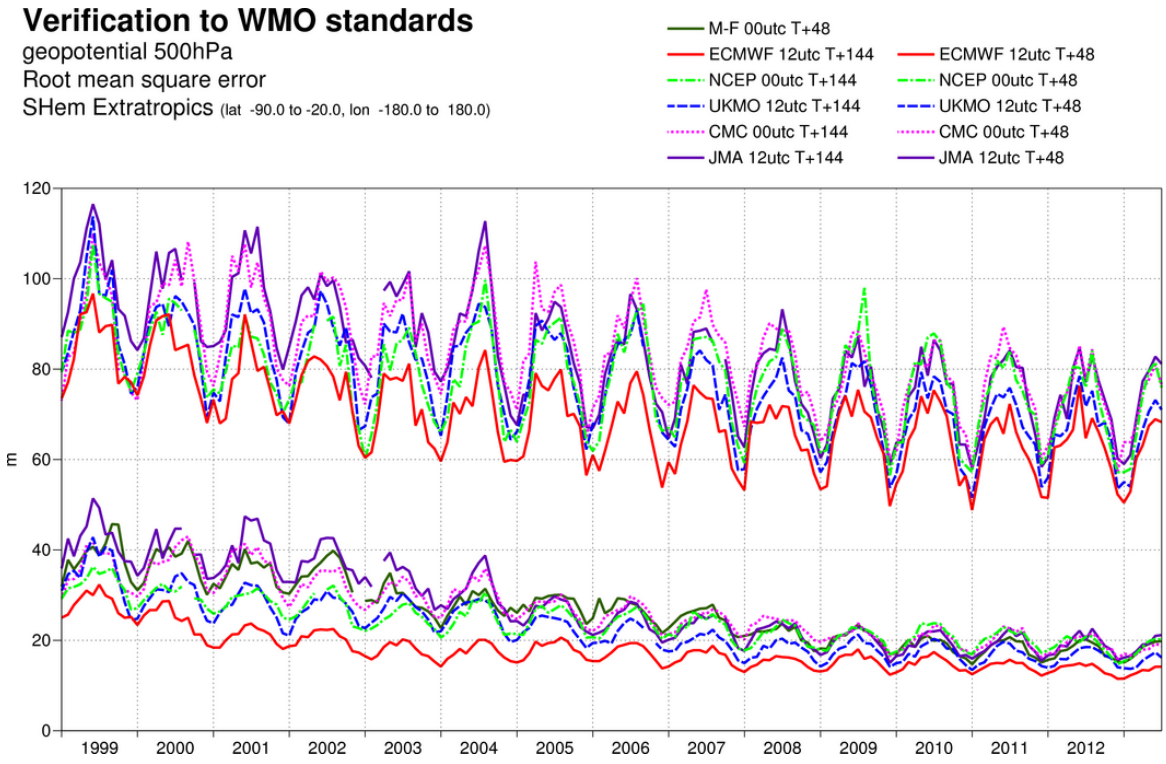


Figure 13: WMO-exchanged scores from global forecast centres. RMS error over northern extratropics for 500 hPa geopotential height (top) and mean sea level pressure (bottom). In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model 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.



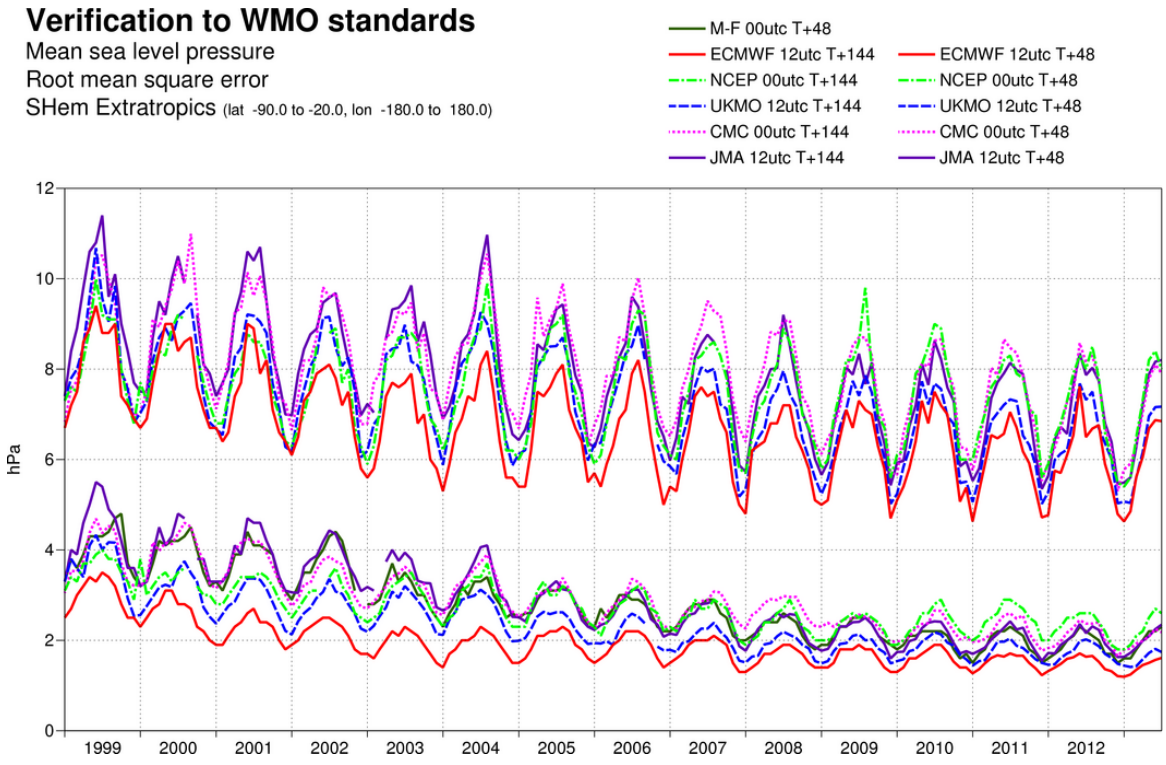
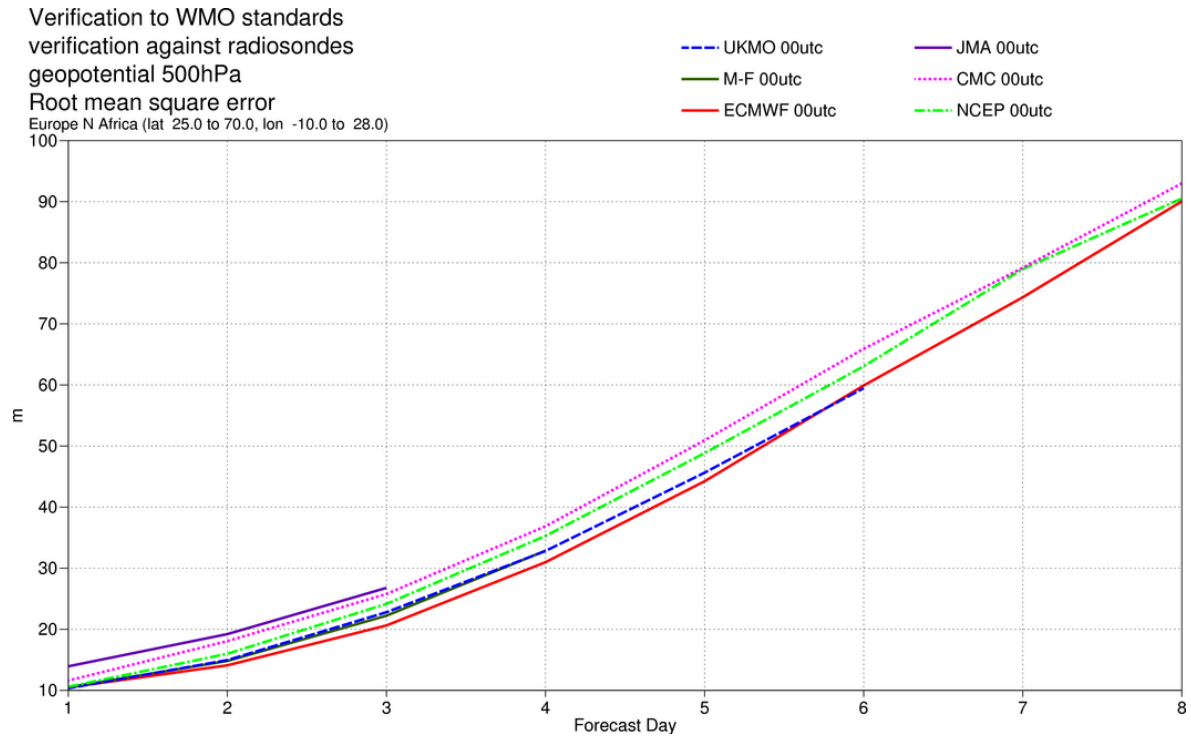


Figure 14: As Figure 13 for the southern hemisphere.



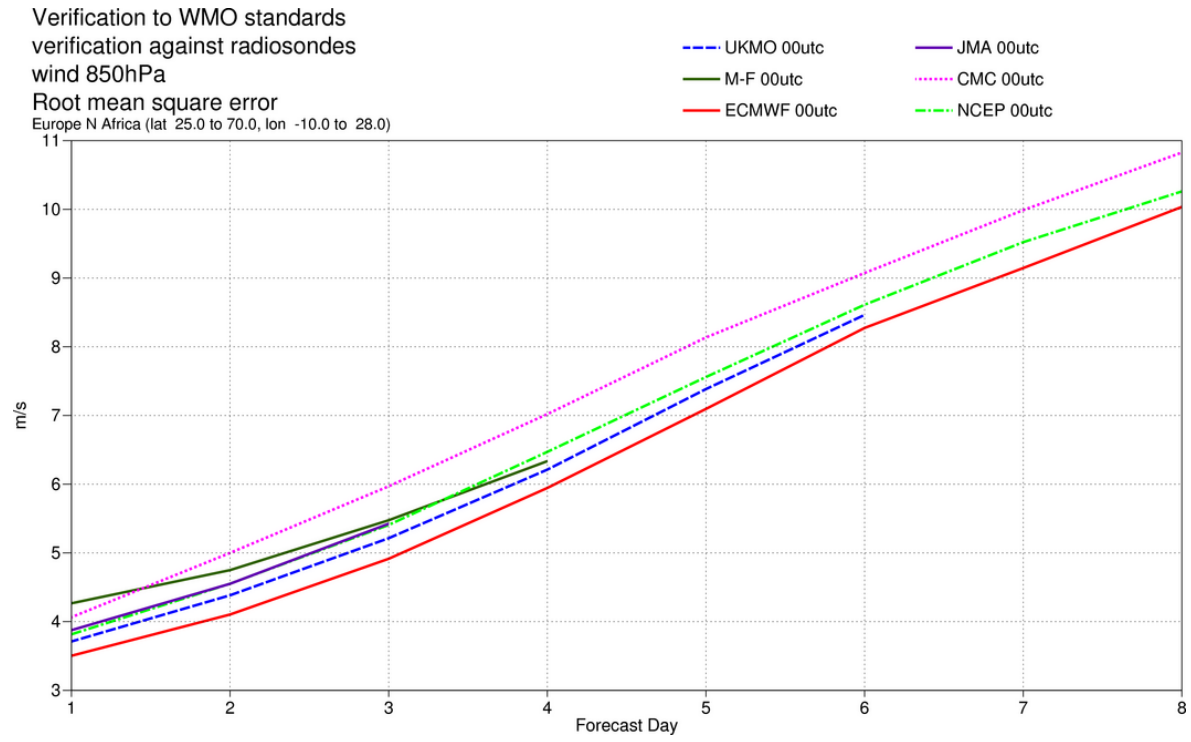


Figure 15: WMO-exchanged scores using radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2012–July 2013).

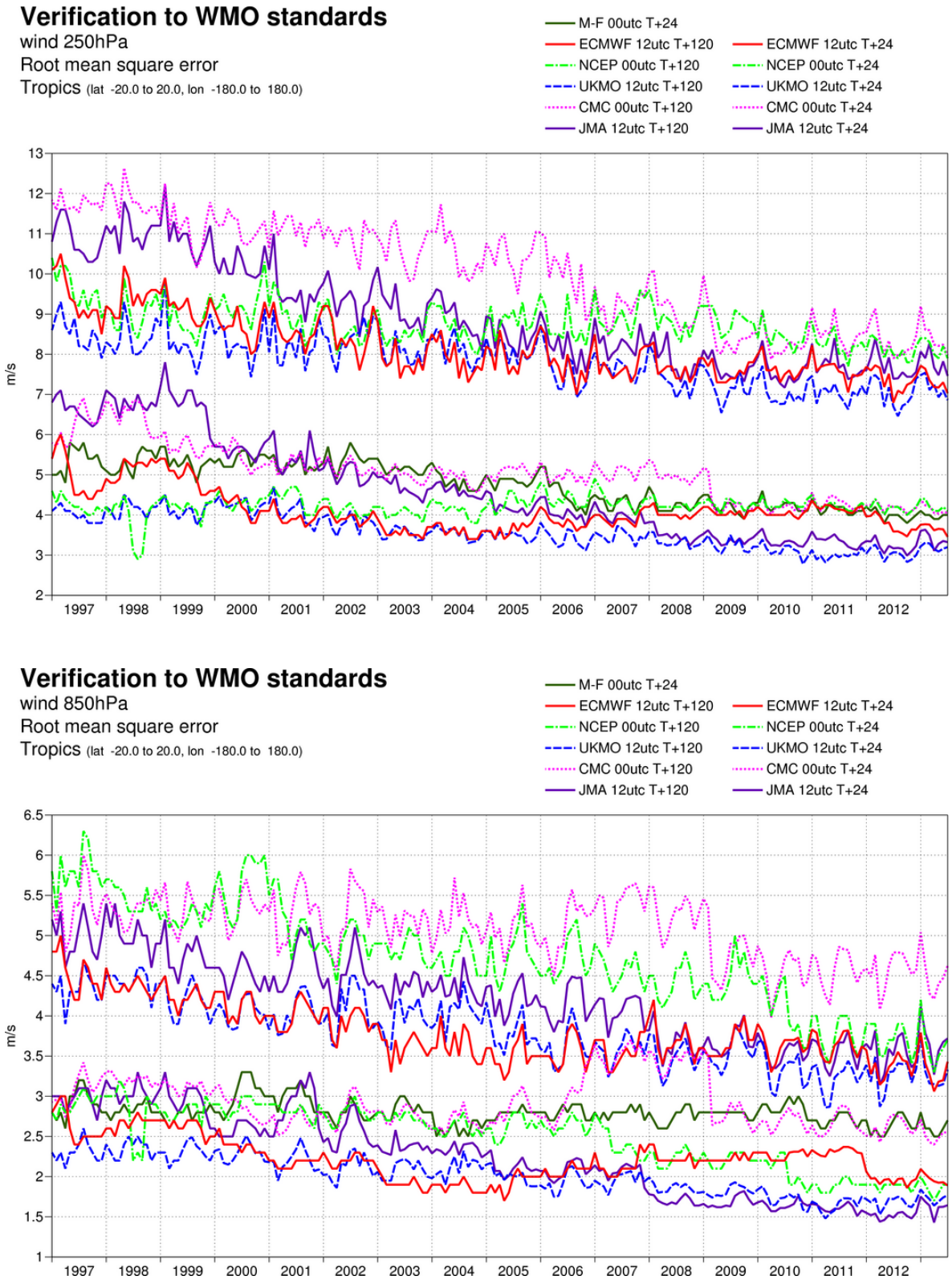
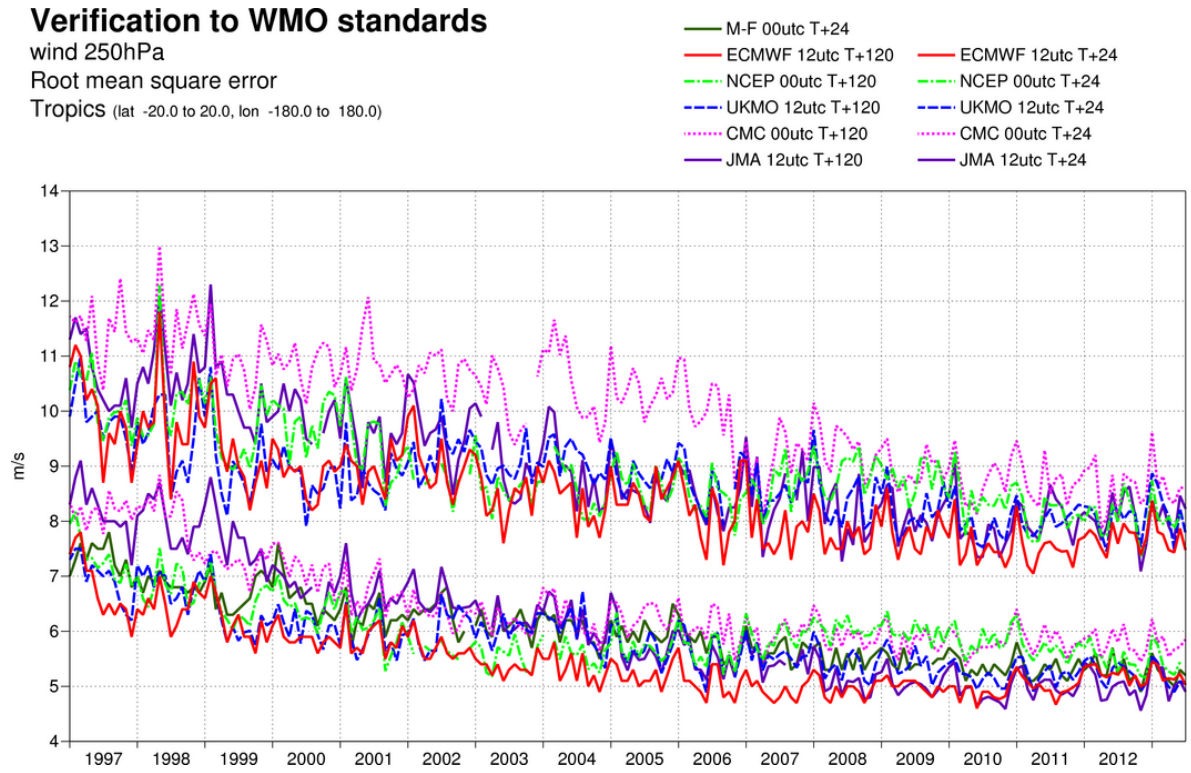


Figure 16: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



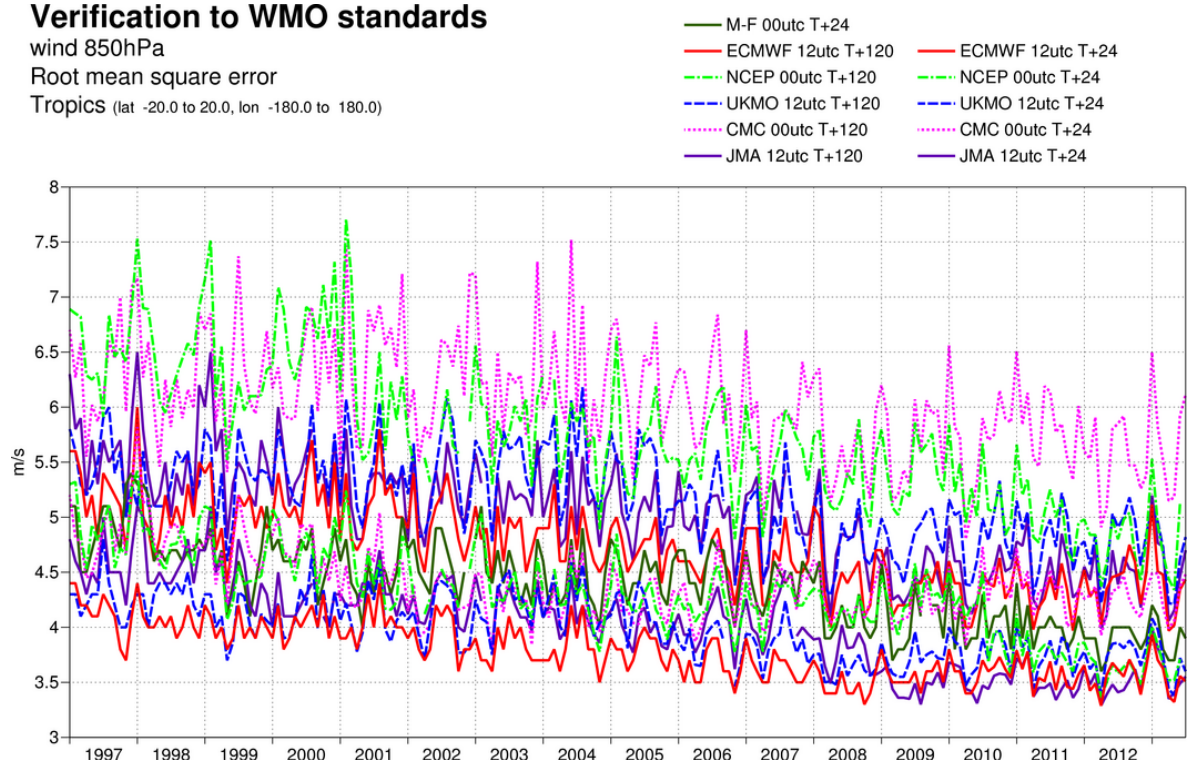


Figure 17: As Figure 16 for scores computed against radiosonde observations.

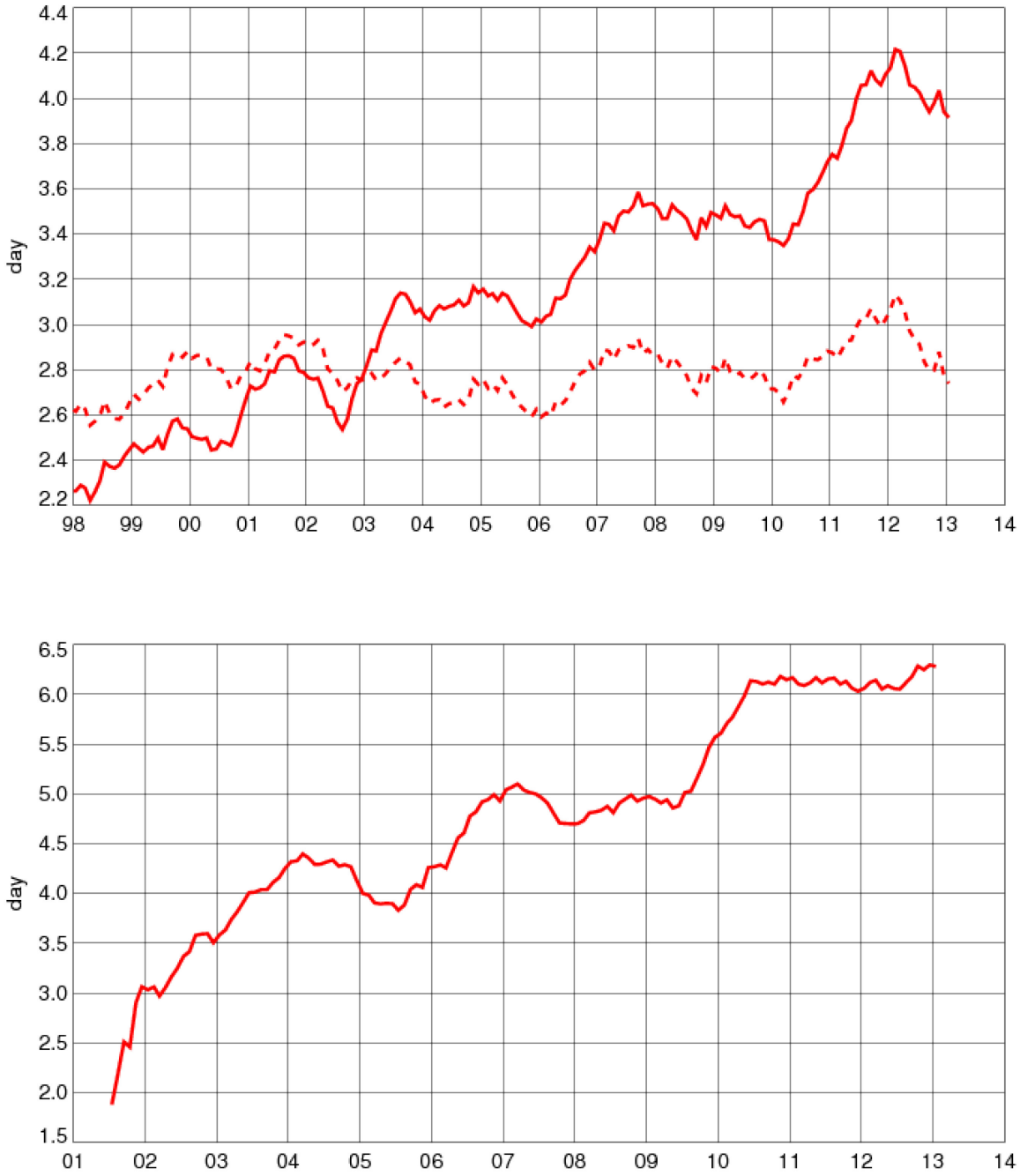


Figure 18: Supplementary headline scores for deterministic (top) and probabilistic (bottom) precipitation forecasts (continuous curves). The dashed curve shows the deterministic headline score for ERA-Interim as a reference. Each curve shows the number of days for which the centred 12-month mean skill remains above a specified threshold for precipitation forecasts over the extratropics. In both cases the verification is for 24-hour total precipitation verified against available synoptic observations in the extratropics; 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.

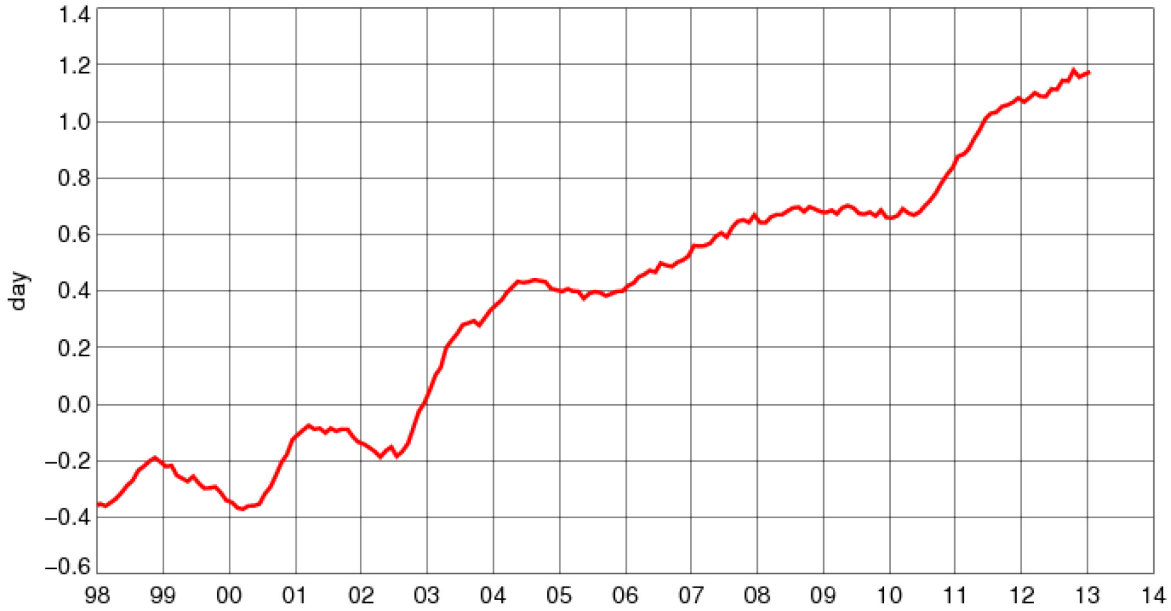


Figure 19: Difference between the operational forecast and ERA-Interim of the supplementary headline score for deterministic precipitation forecasts. The curve is the difference between the two curves in the upper panel of Figure 18.

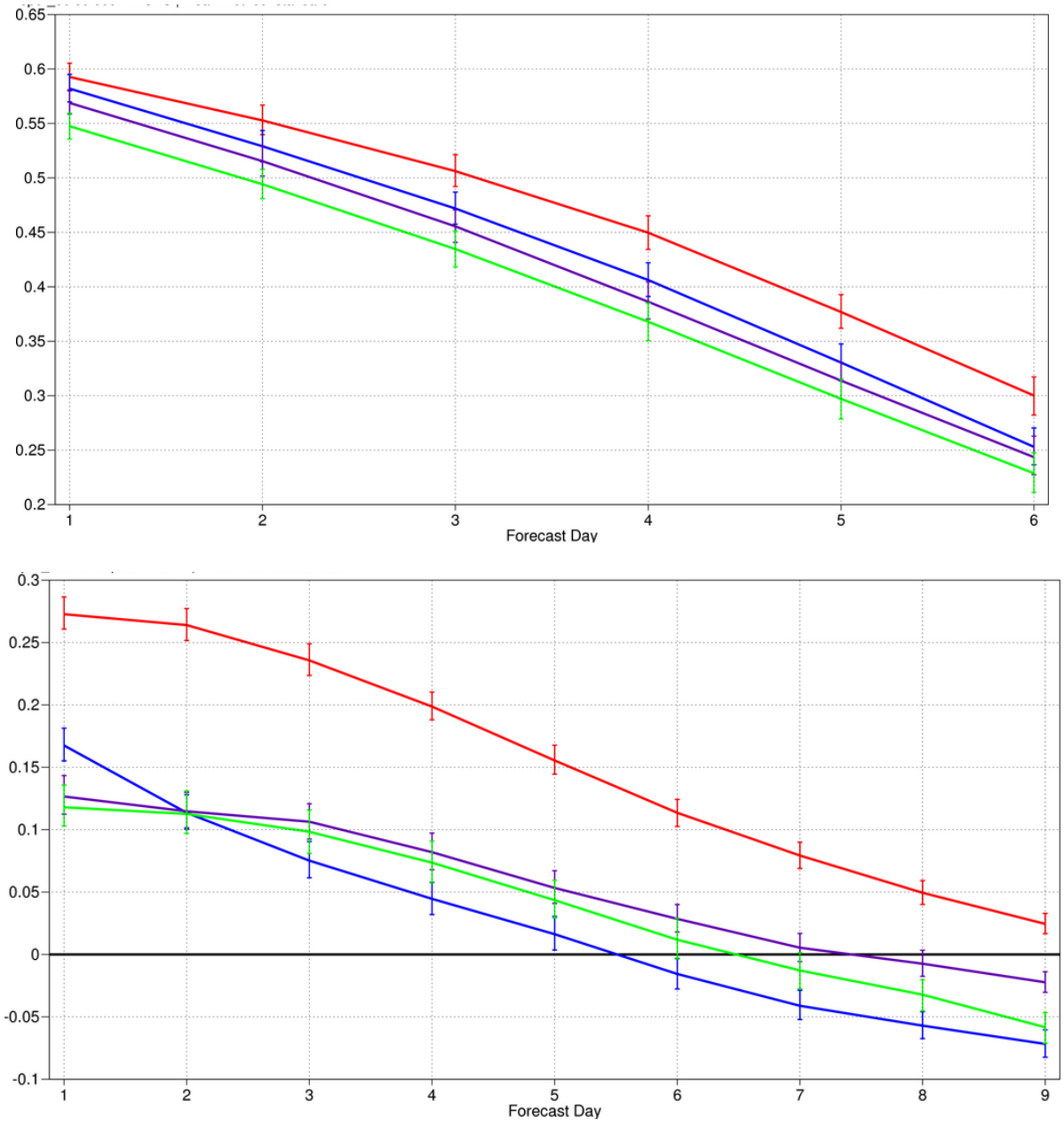


Figure 20: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2012–July 2013. Bars indicate 95% confidence intervals.

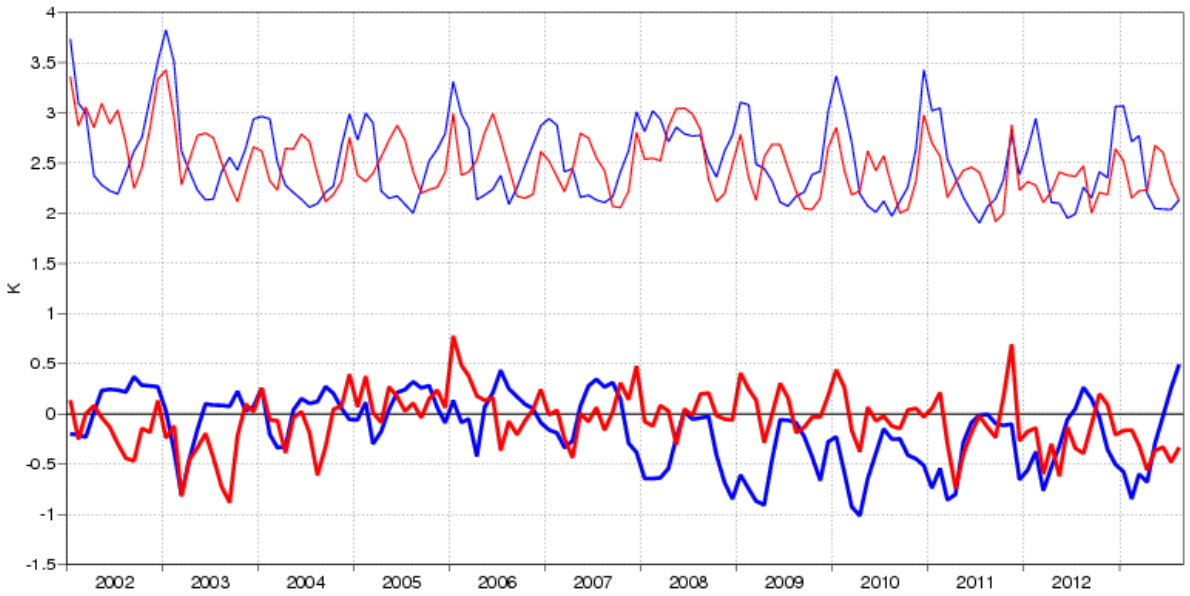


Figure 21: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

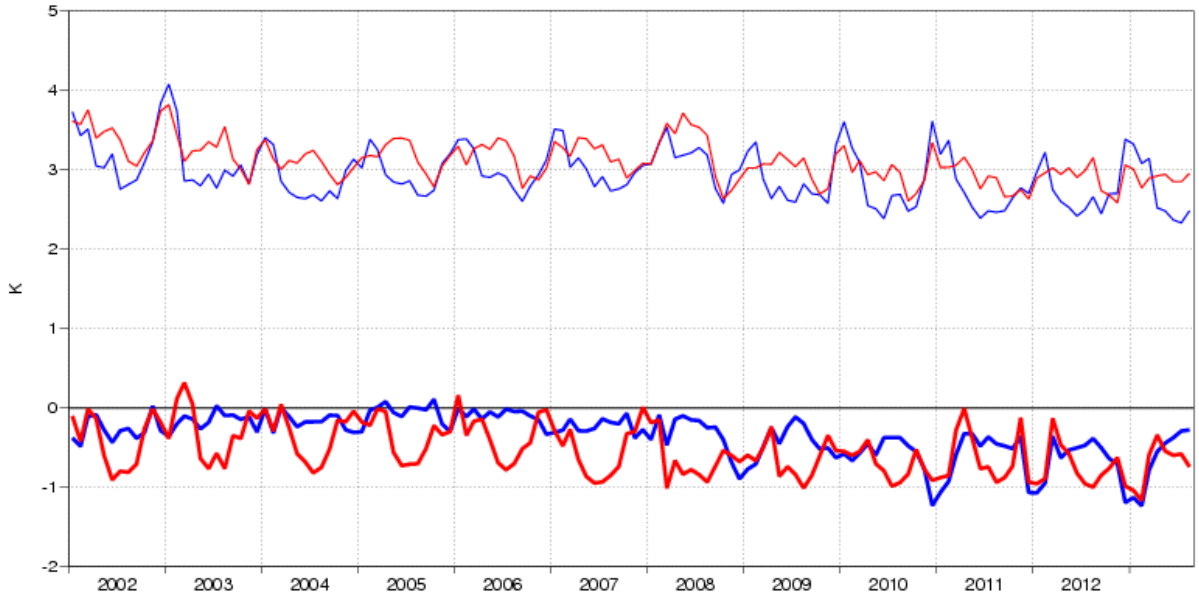


Figure 22: Verification of 2 m dewpoint forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

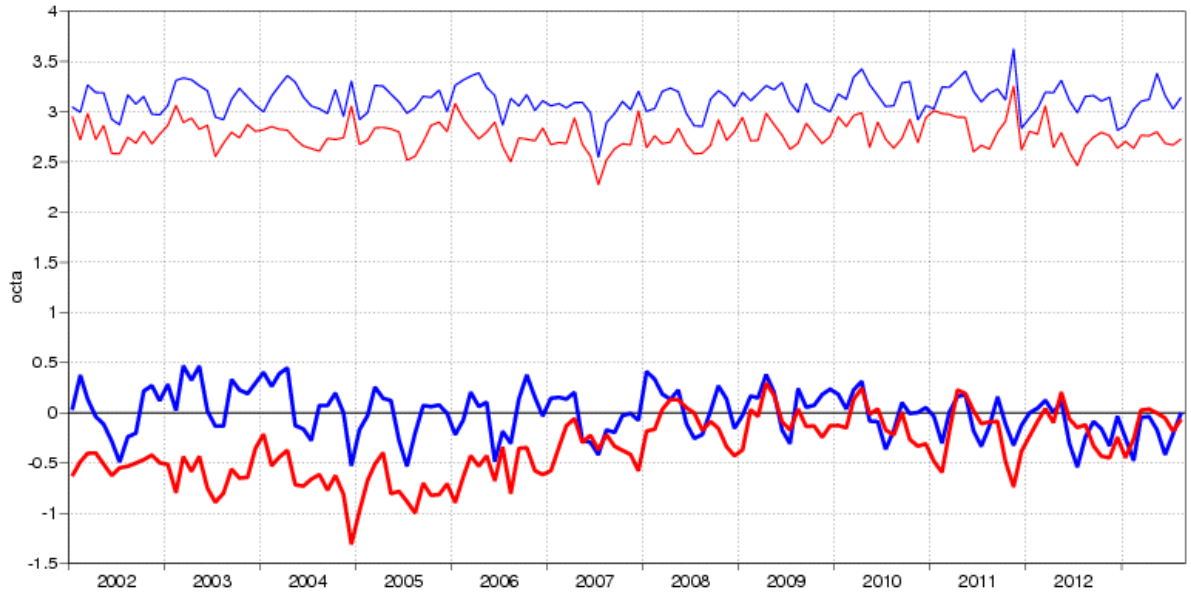


Figure 23: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

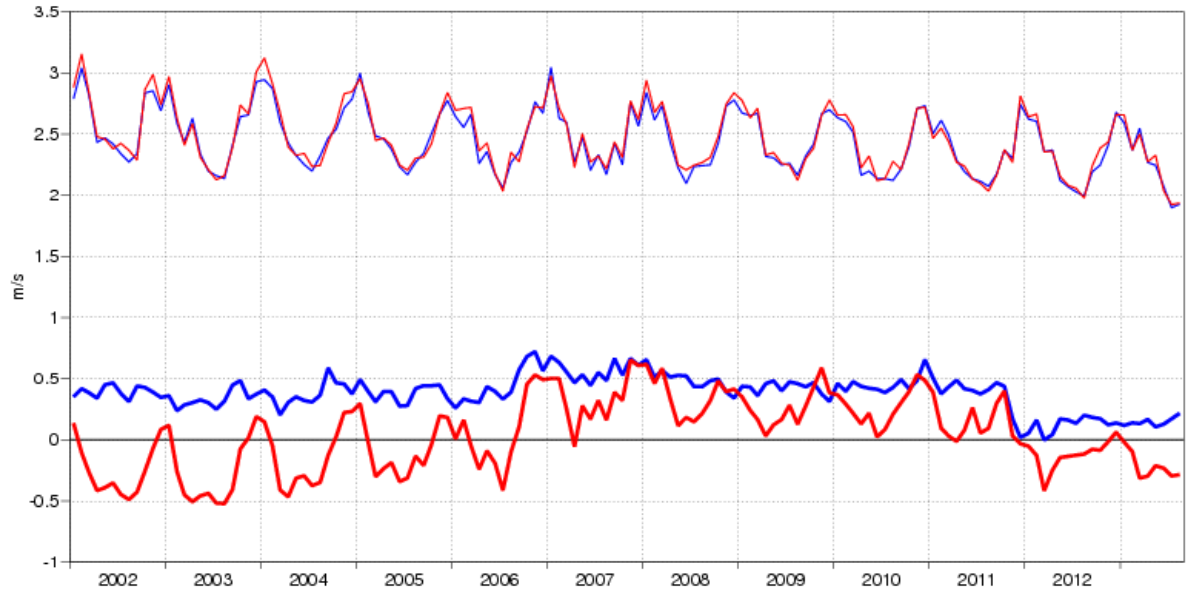


Figure 24: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

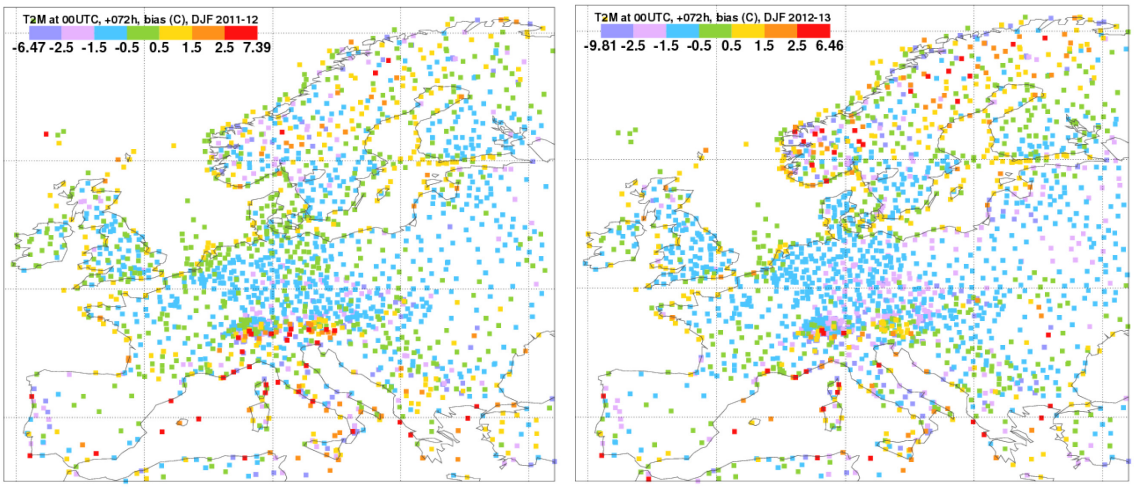


Figure 25: Night-time 2 m temperature mean errors during winters (Dec–Feb) 2011–12 and 2012–13.

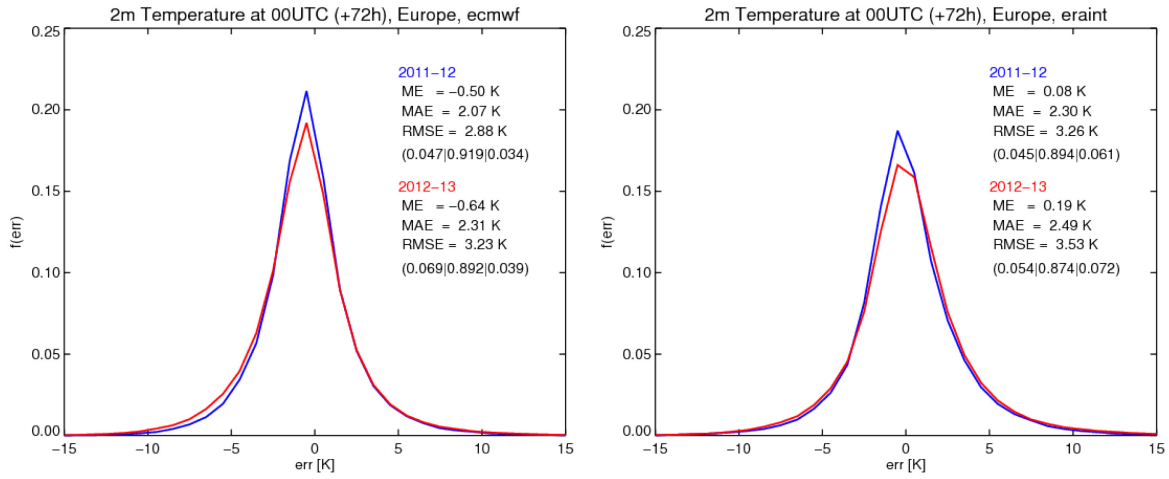


Figure 26: Error distributions for Europe, comparison of winters 2011–12 and 2012–13 for the operational run (left) and ERA-Interim (right). Also shown are mean error (ME), mean absolute error (MAE) and root mean squared error (RMSE) for the two winters. Fractions of cases with errors <-5 K, between -5 and +5 K, and >+5 K are given in parentheses.

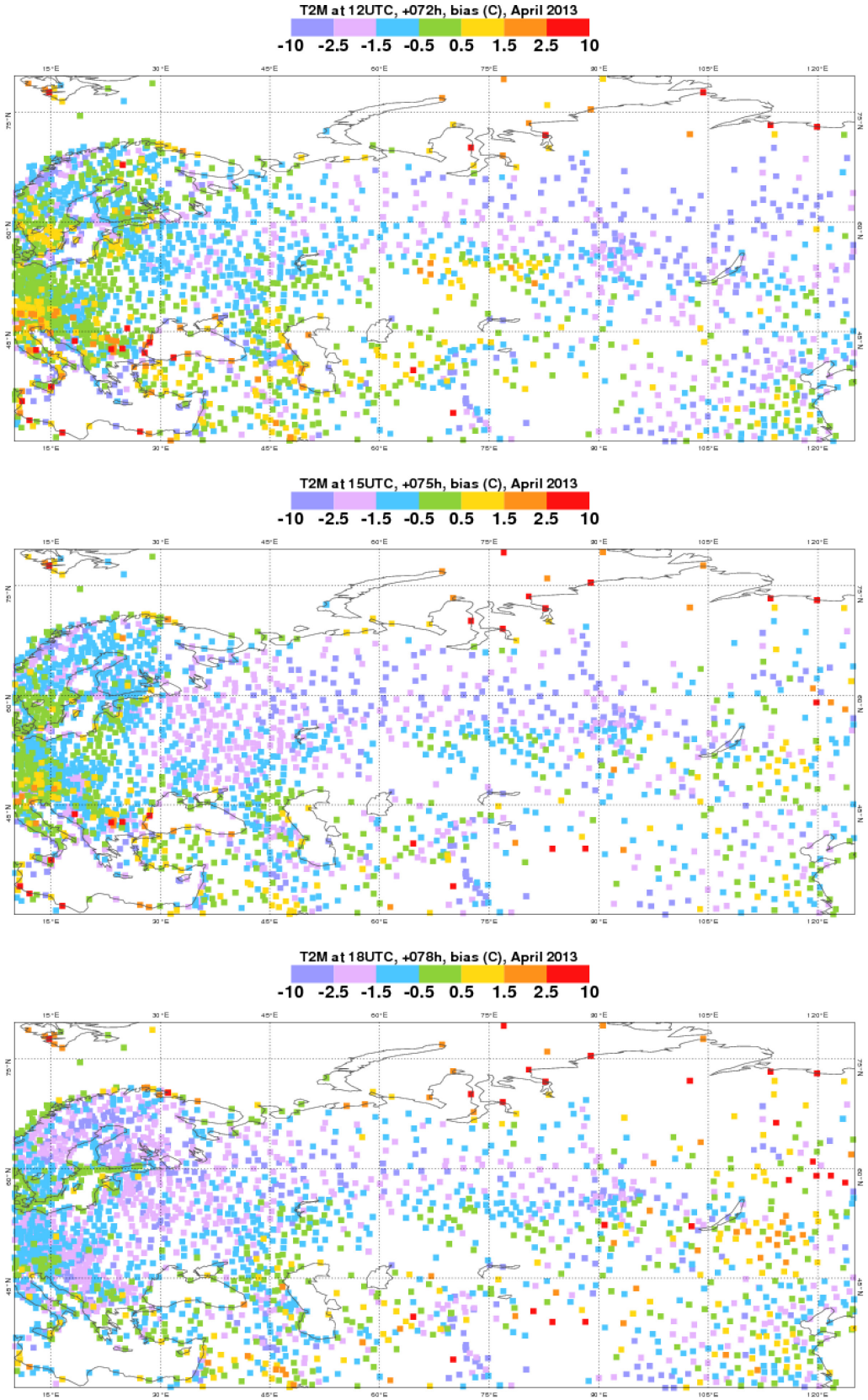


Figure 27: Mean errors of the 2 m temperature forecast at forecast day 3 for 12, 15, and 18 UTC during April 2013.

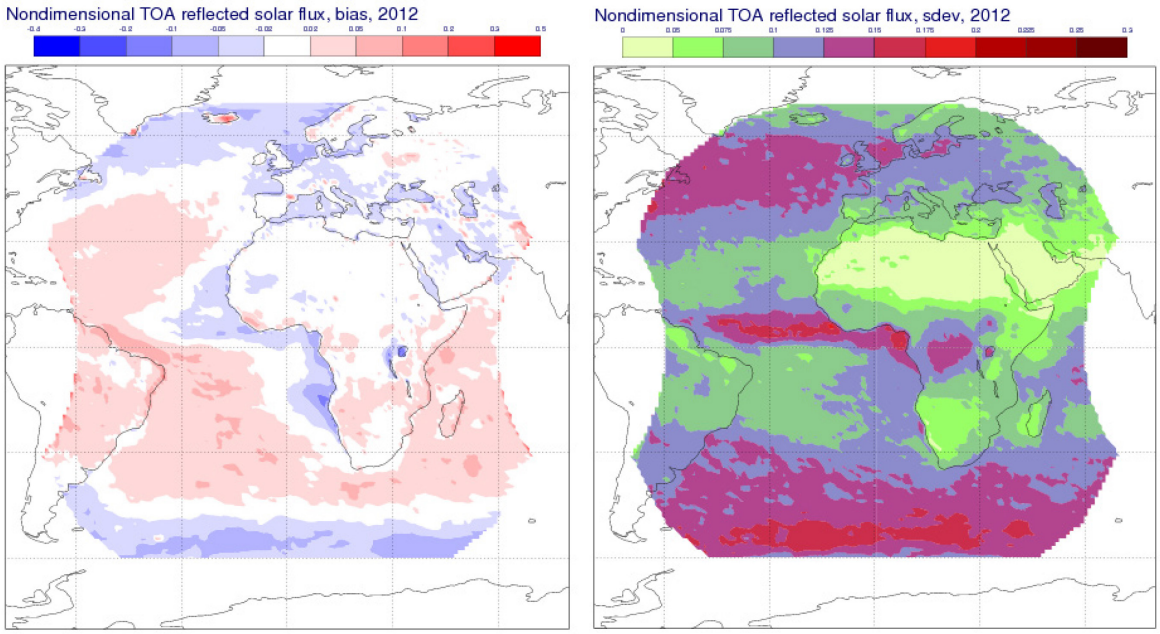


Figure 28: Mean error (left) and standard deviation of the error (right) for the high-resolution operational forecasts of daily means of non-dimensional top of the atmosphere reflected solar radiation at forecast day 3 in the year 2012.

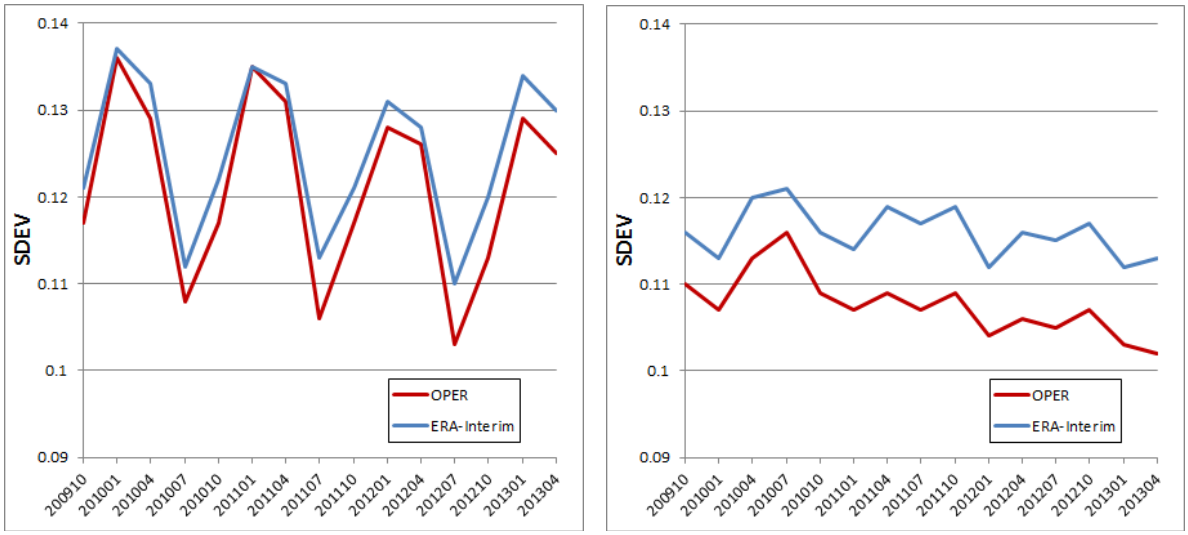
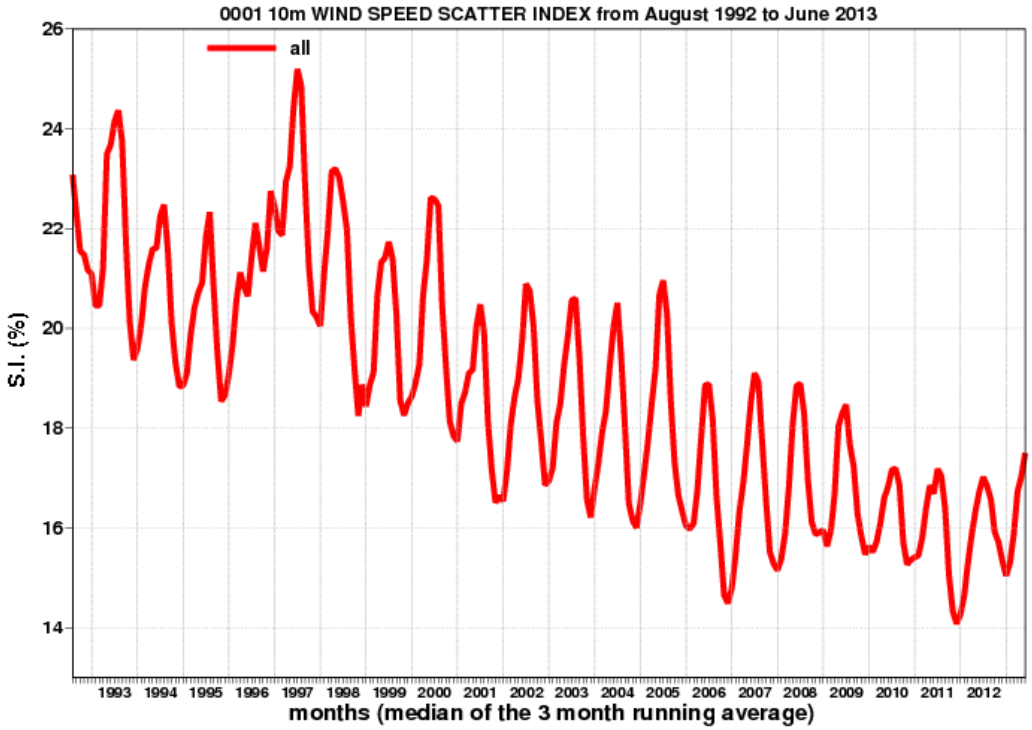


Figure 29: Seasonal averages of the day 3 forecast error standard deviation of normalized TOA reflected solar flux (daily totals) in the parts of the northern hemisphere extratropics (left panel) and tropics (right panel) which are covered by the CM-SAF product in Figure 28.



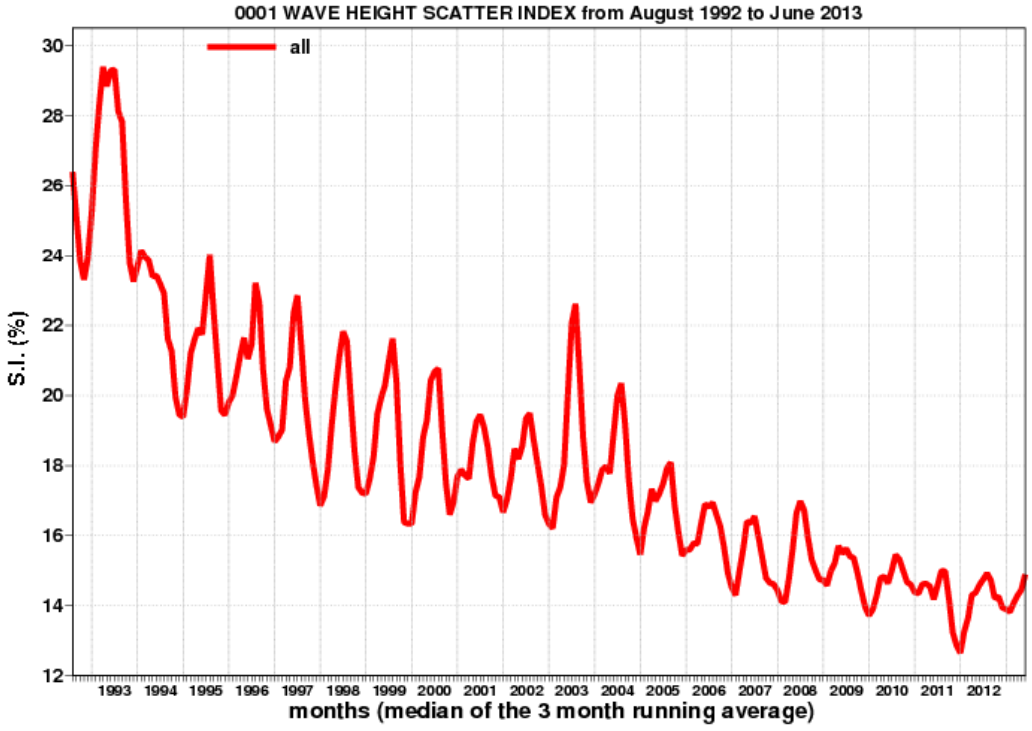


Figure 30: Time series of verification of the ECMWF 10 m wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

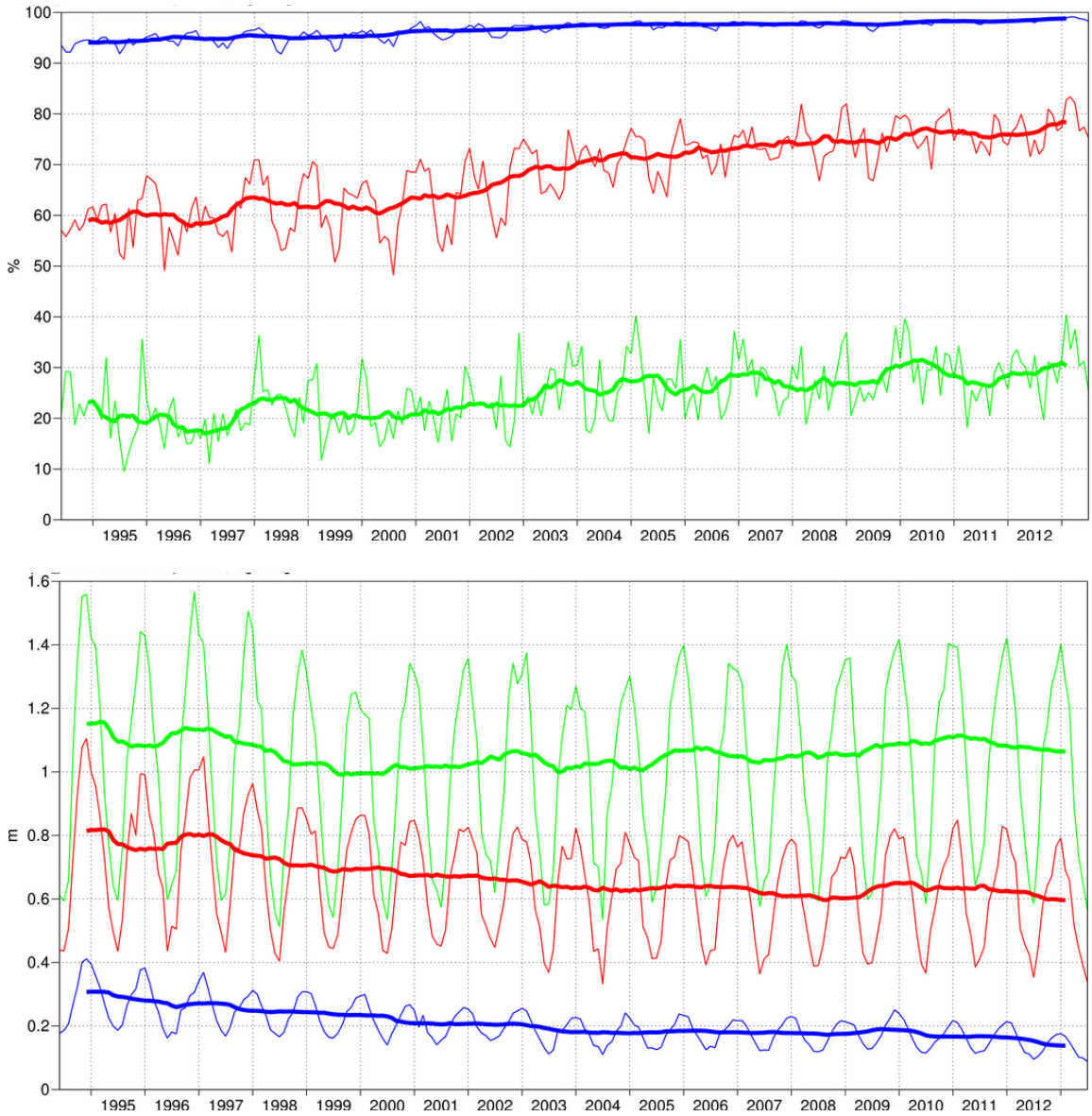


Figure 31: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC (top) and error standard deviation (bottom) for ocean wave heights verified against analysis for the northern extratropics at day 1 (blue), 5 (red) and 10 (green).

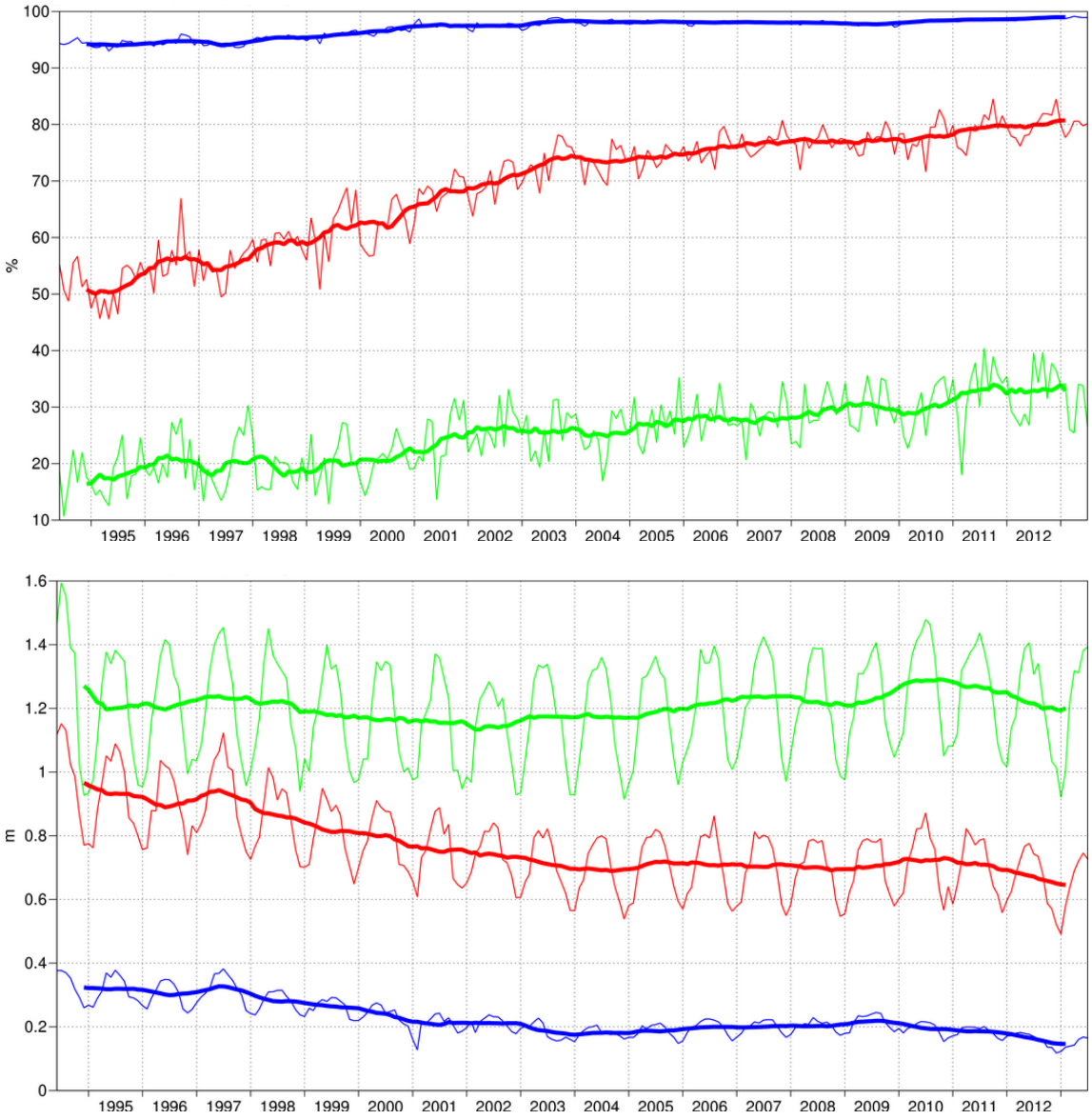


Figure 32: As Figure 31 for the southern hemisphere.

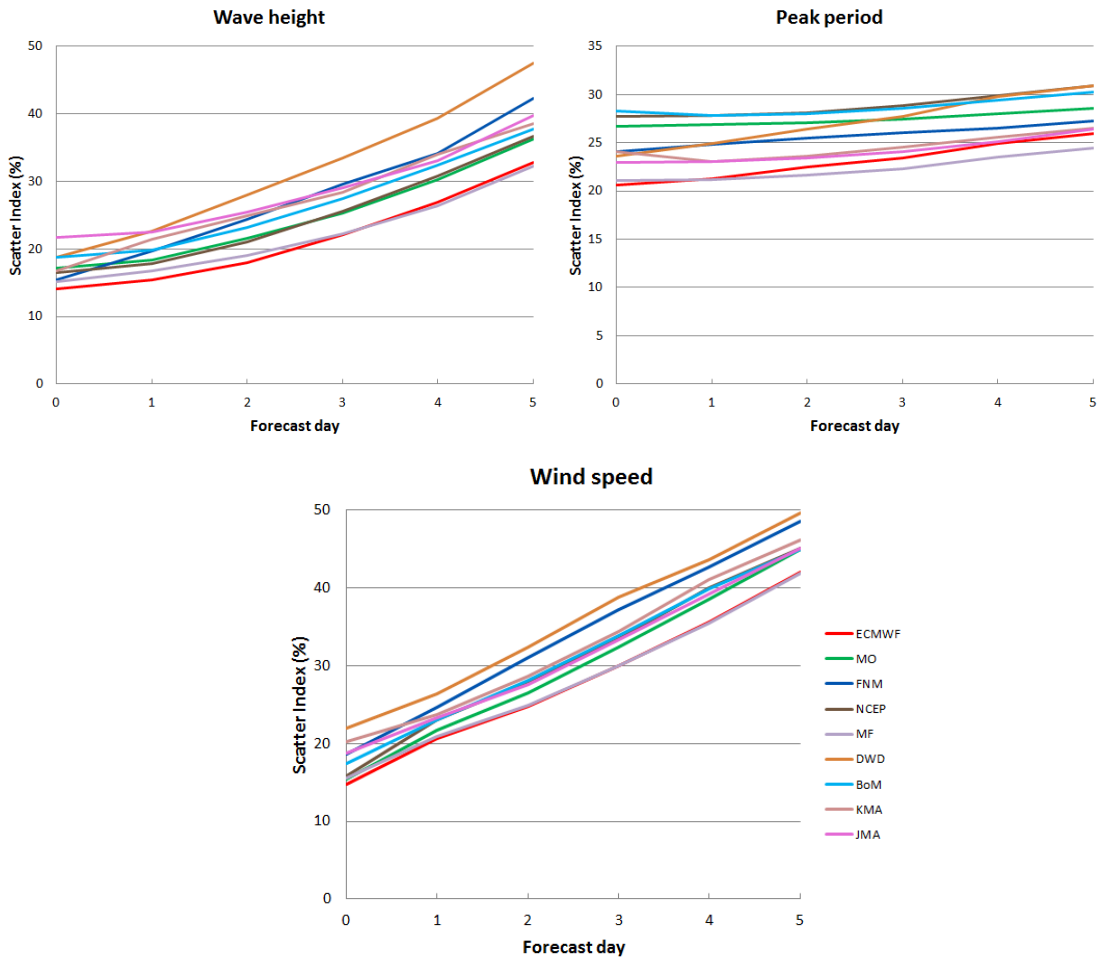


Figure 33: Verification of different model forecasts of wave height, peak wave period and 10 m wind speed using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period September 2012 –August 2013. The x-axis shows the forecast range in days from analysis (step 0) to day 5. MO: the Met Office, UK; FNM: Fleet Numerical Meteorology and Oceanography Centre, USA; NCEP: National Centers for Environmental Prediction, USA; MF: Météo France; DWD: Deutscher Wetterdienst, BoM: Bureau of Meteorology, Australia; KMA: Korea Meteorological Administration; JMA: Japan Meteorological Agency

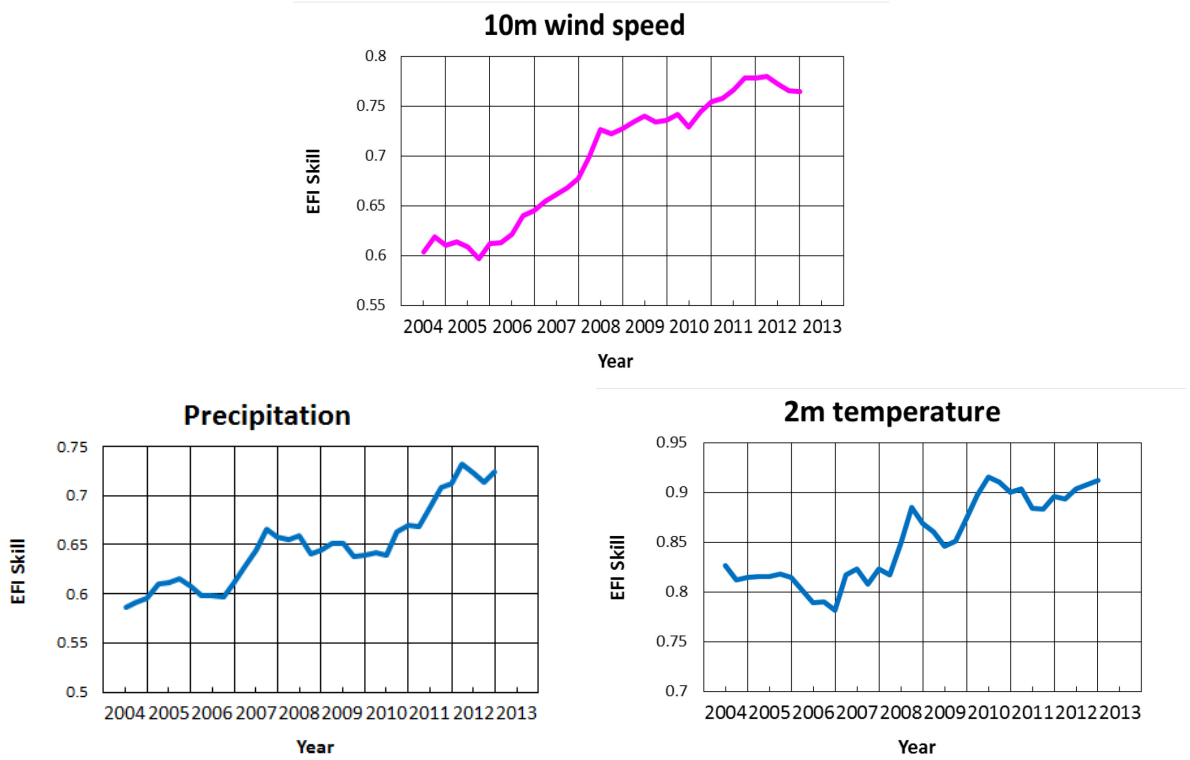


Figure 34: Verification of Extreme Forecast Index (EFI). Top panel: supplementary headline score – skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate, curves show a four-season running mean of relative operating characteristic (ROC) area skill scores (final point includes spring (March–May) 2013). Bottom panels show the equivalent ROC area skill scores for the precipitation (left) and 2 m temperature (right) EFI forecasts.

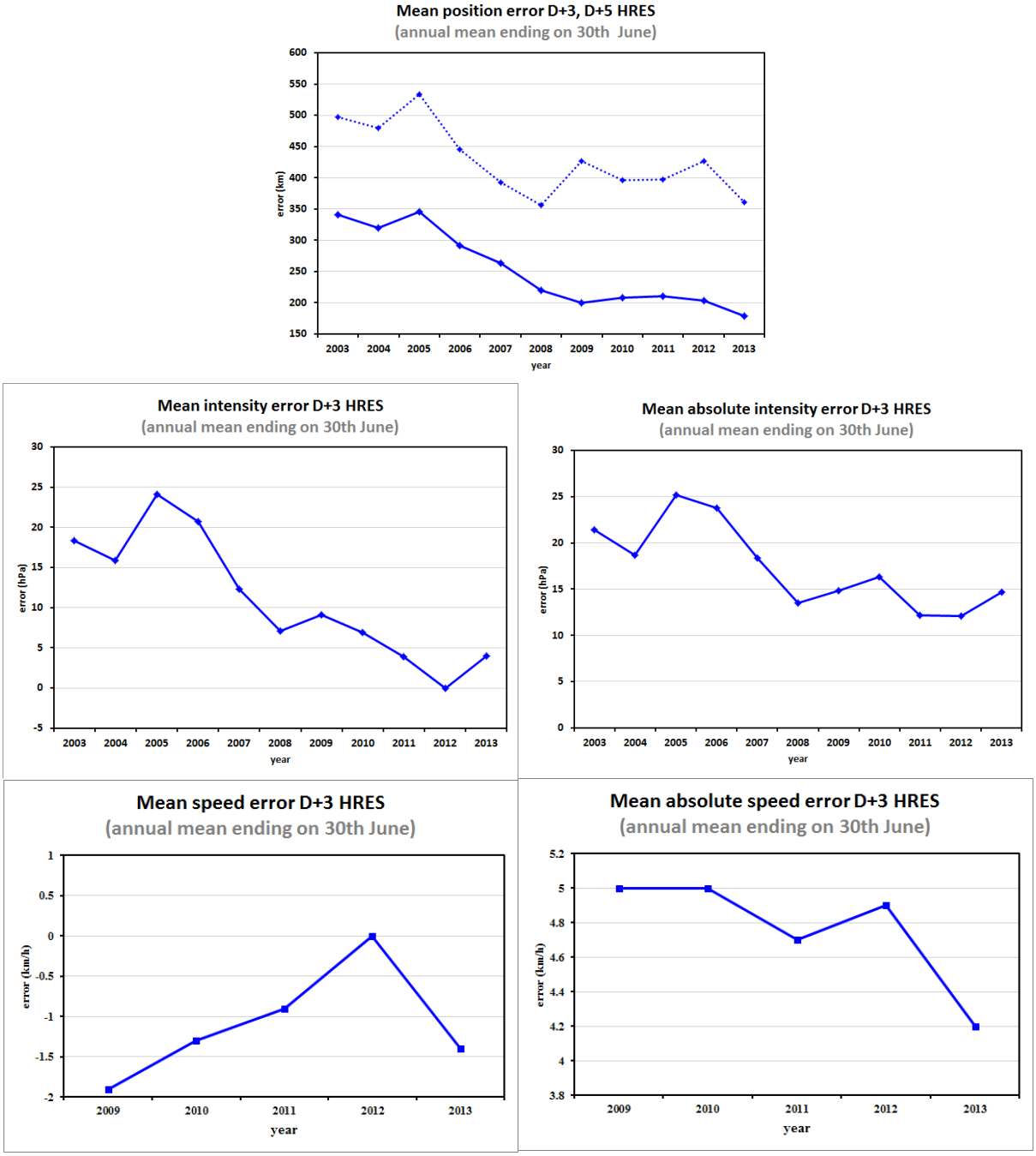


Figure 35: Verification of tropical cyclone predictions from the operational high-resolution forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported in real time via the GTS. Top left: supplementary headline score – the mean position error (km) of the three-day high-resolution forecast. The error for day 5 is included for comparison. Centre left: 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. Centre right: mean absolute error of the intensity. Bottom left: mean speed error; negative values indicate the forecast is too slow compared to the observed cyclones. Bottom right: mean absolute error of the speed.

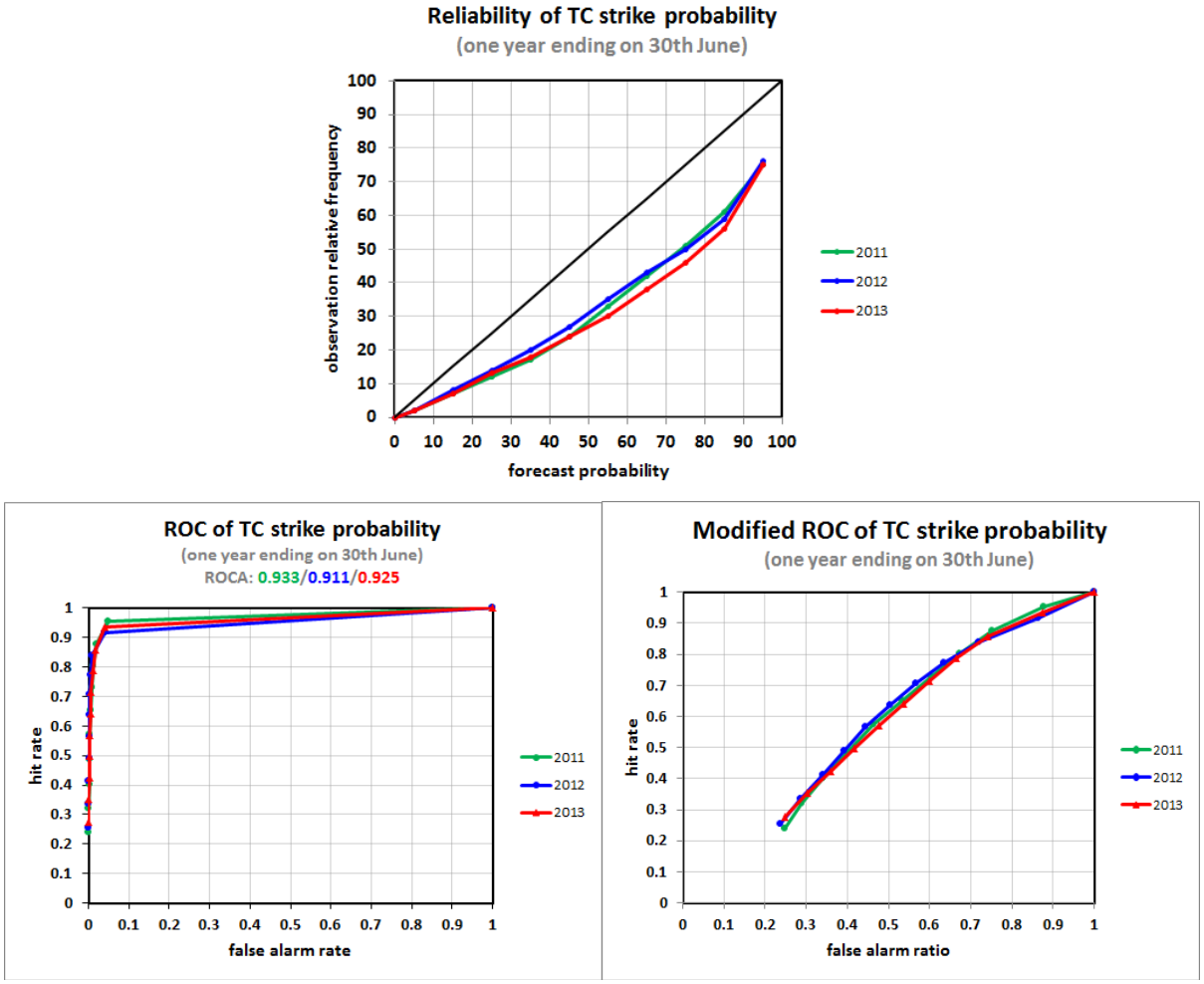


Figure 36: Probabilistic verification of ensemble tropical cyclone forecasts for three 12-month periods: July 2010–June 2011 (green), July 2011–June 2012 (blue) and July 2012–June 2013 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) 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).

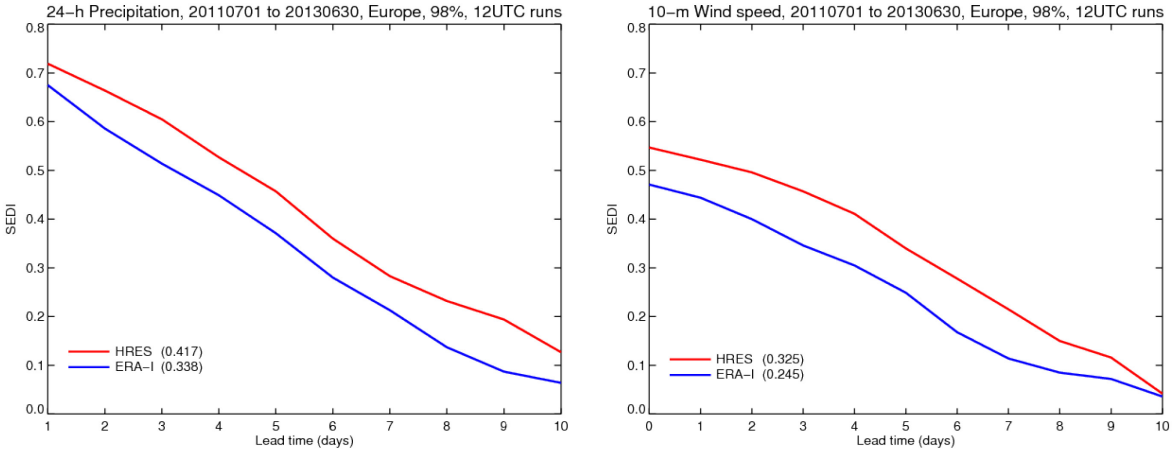


Figure 37: Skill of the HRES forecast (red) and ERA-Interim (blue) in predicting 24-h precipitation amount (left panel) and 10 m wind speed (right panel) above the 98th climate percentile in Europe as measured by the SEDI score. Verification period is 2 years (July 2011–June 2013). Mean values over all lead times given in parentheses.

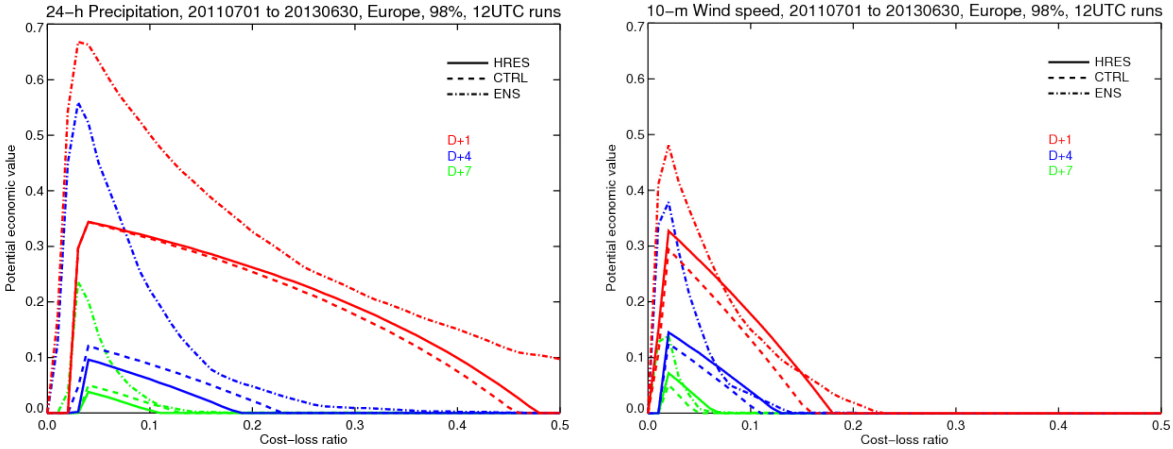


Figure 38: Potential economic value of the HRES forecast (continuous), the ENS control forecast (dashed), and the ENS forecast (dash-dotted) in predicting 24-h precipitation amount (left panel) and 10 m wind speed (right panel) above the 98th climate percentile in Europe. Colours indicate forecast days 1 (red), 4 (blue), and 7 (green). Users' cost–loss ratios are typically quoted to be in the range 0.01–0.2.

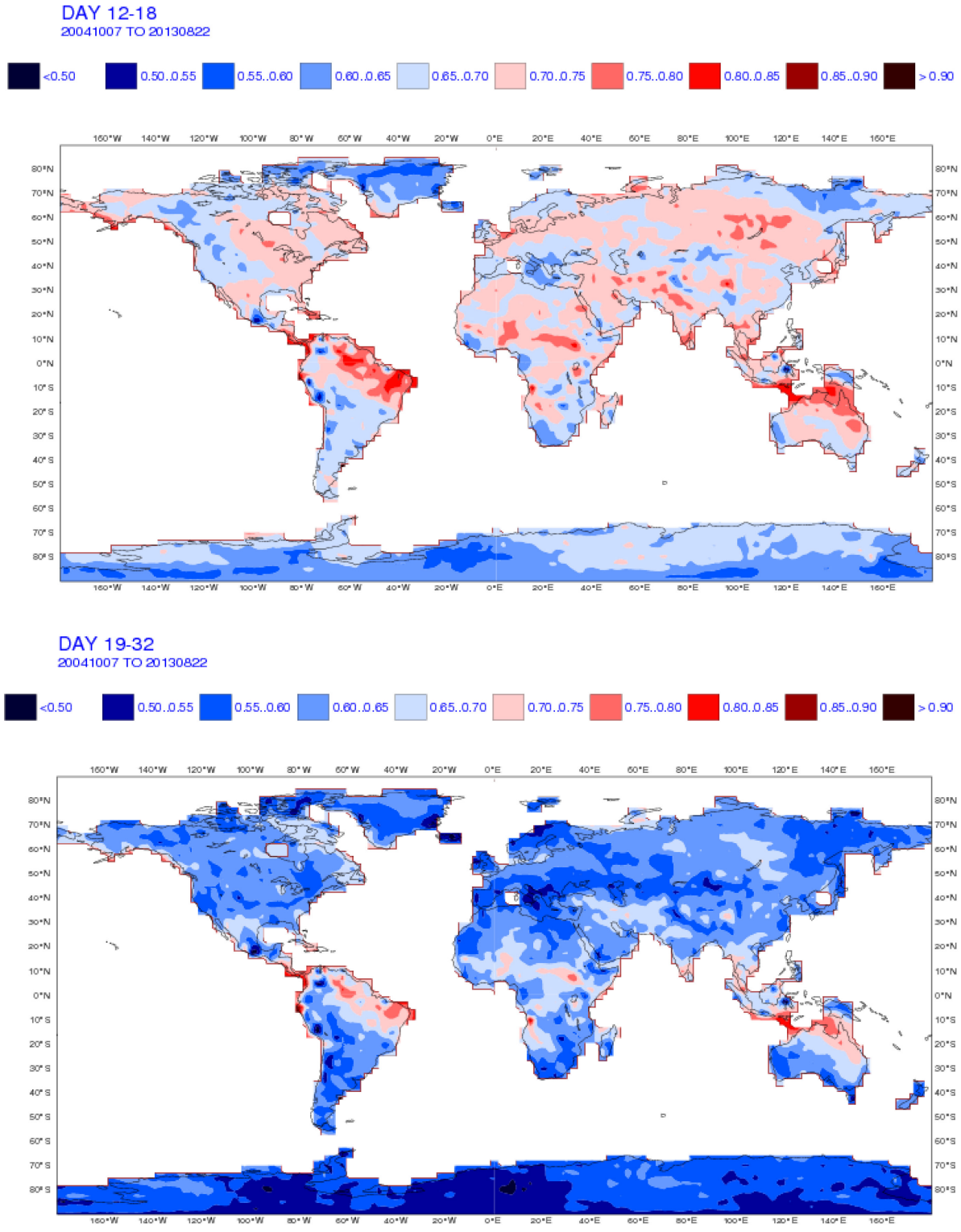


Figure 39: Monthly forecast verification. Spatial distribution of ROC area scores for the probability of 2 m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 22 Aug 2013 for two seven-day forecast ranges: days 12–18 (top) and days 19–25 (bottom). Warmer colour indicates higher skill compared to climate.

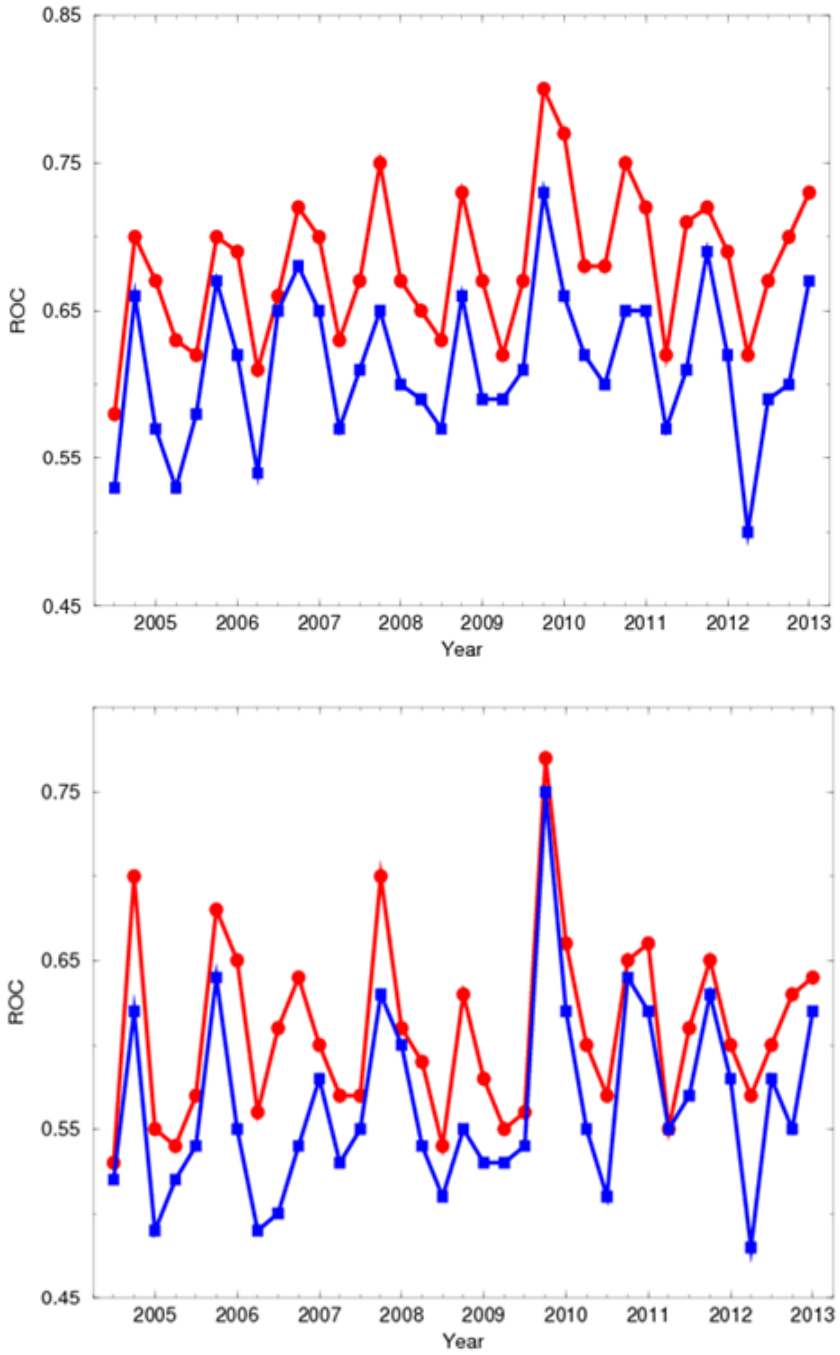


Figure 40: Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution. Scores are calculated for each three-month season since autumn (September–November) 2004 for all land points in the extratropical northern hemisphere. The red line shows the 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 2013.

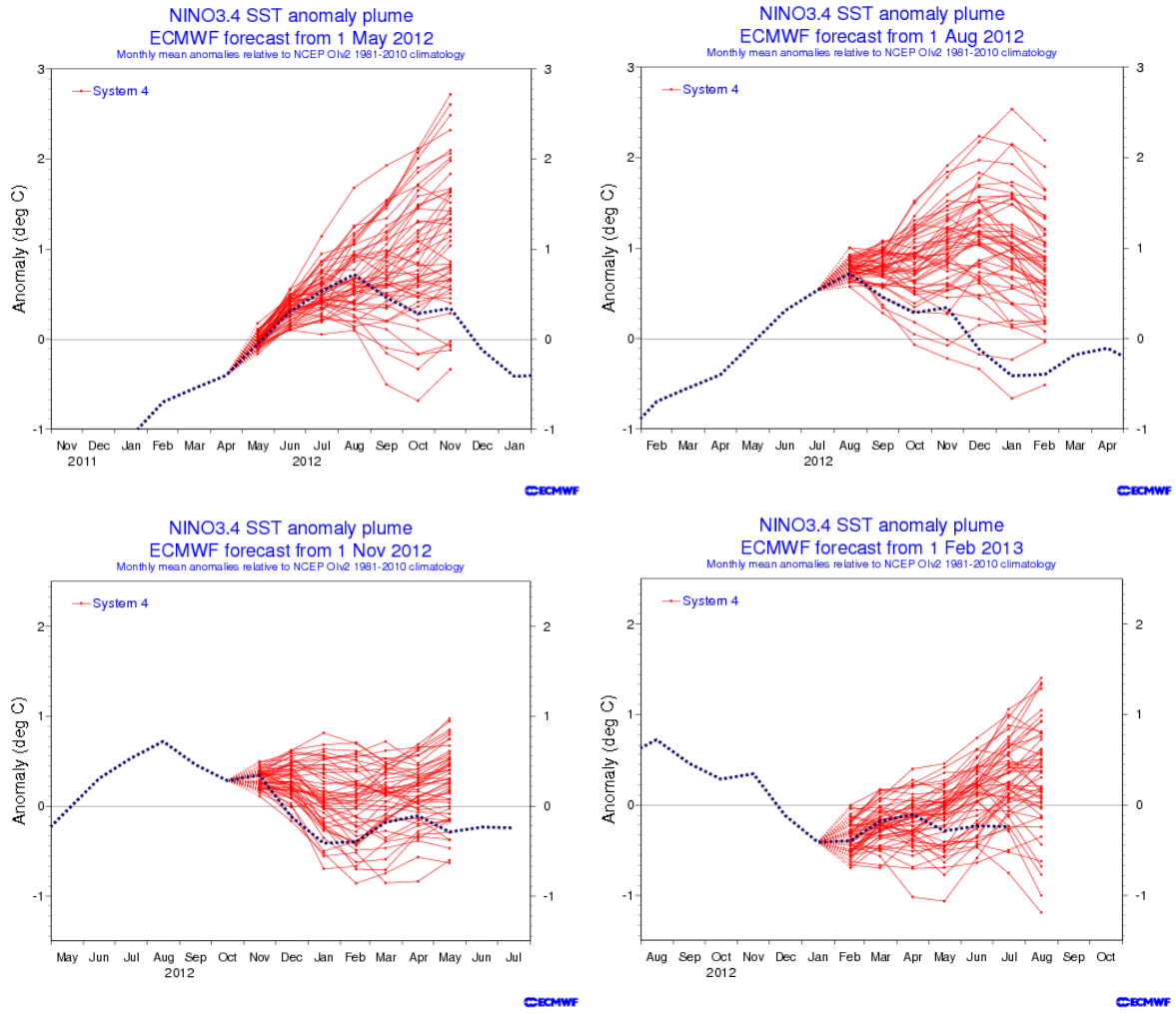


Figure 41: ECMWF seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from May 2012 (top left), August 2012 (top right), November 2012 (bottom left) and February 2013 (bottom right). The red lines represent the ensemble members; dashed blue lines show the subsequent verification.

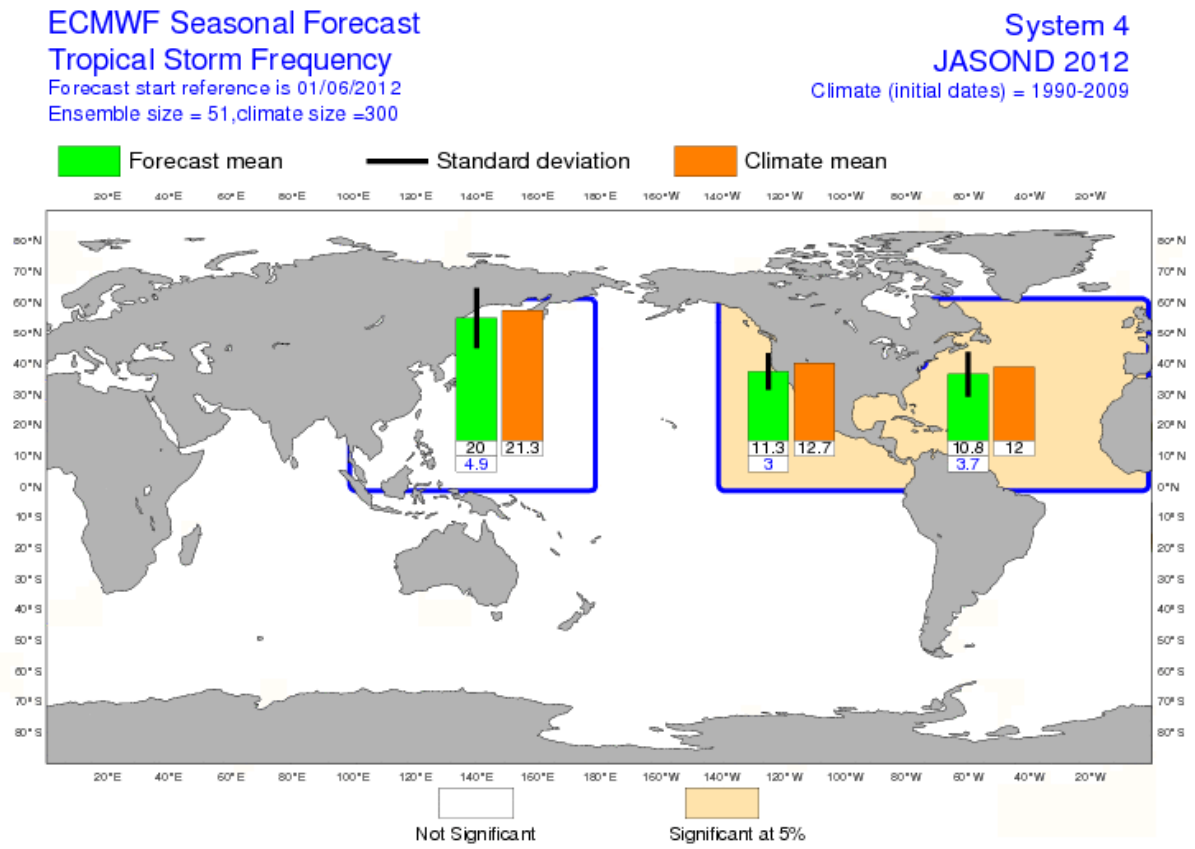


Figure 42: Tropical storm frequency forecast issued in June 2012 for the six-month period July–December 2012. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The black bars represent ±1 standard deviation within the ensemble distribution; these values are indicated by the blue number. The 51-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.

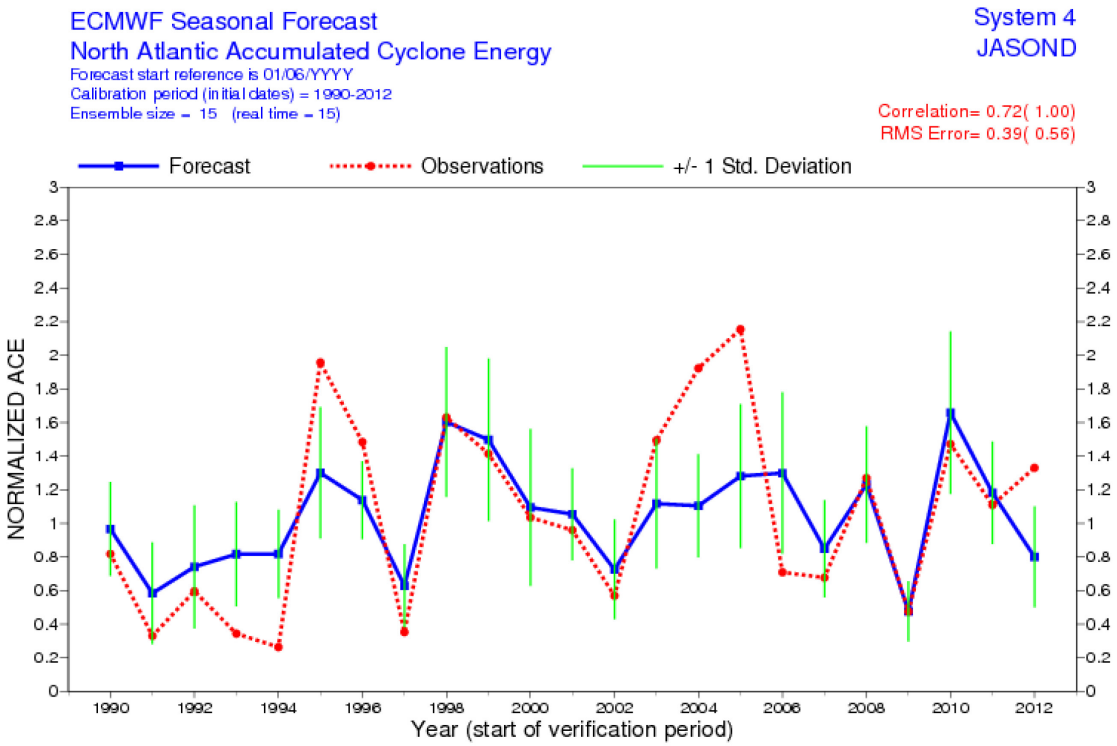


Figure 43: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2011. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows the observation. Forecasts are from System 3 of the seasonal component of the IFS: 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.

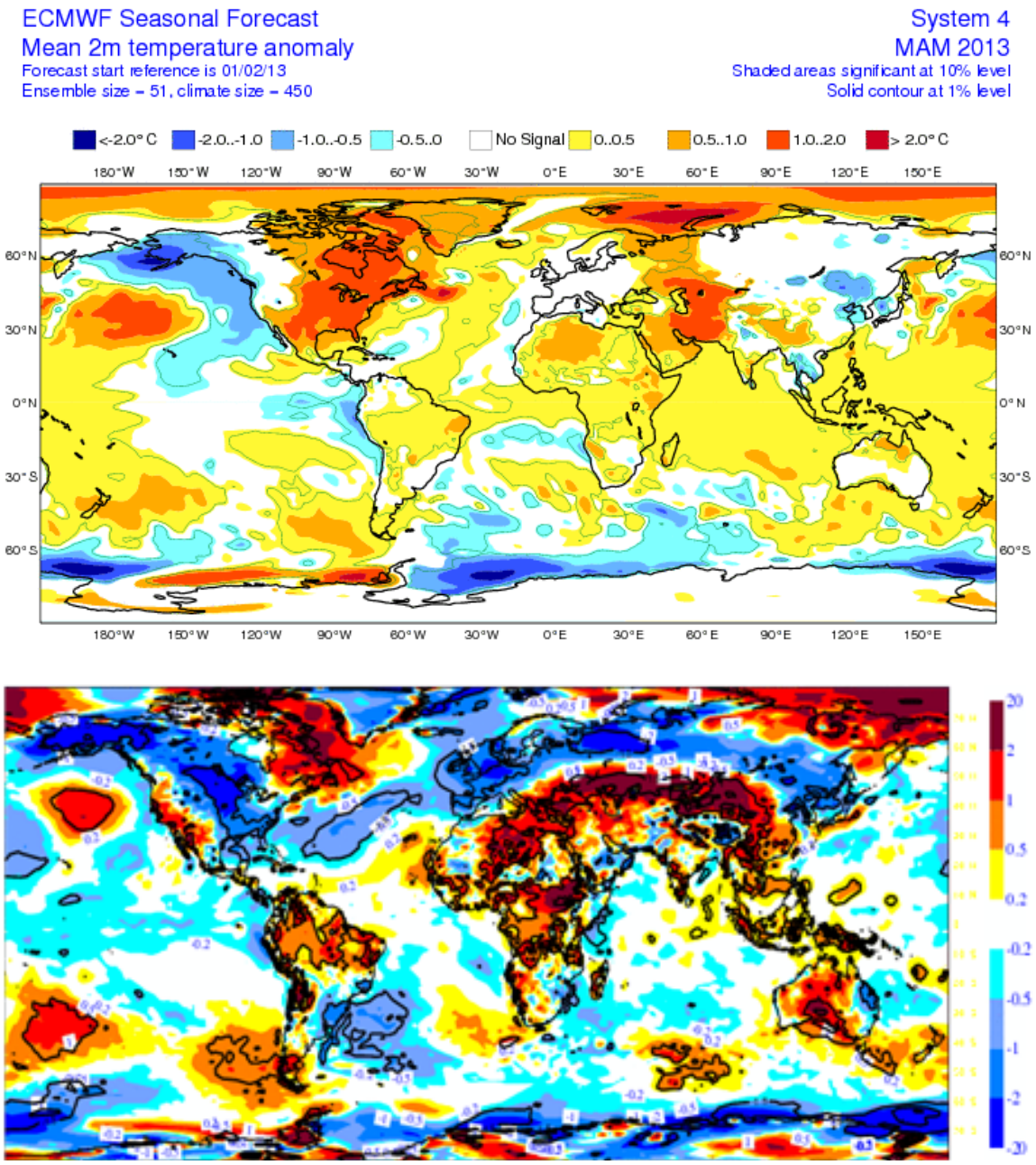


Figure 44: Anomaly of 2 m temperature as predicted by the seasonal forecast system for March–May 2013 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

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 × 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 15), the sondes have been selected following an agreement reached by data monitoring centres and published in the 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 15,

Figure 16) 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 4) are computed as the reduction in RMSE 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 3 and Figure 6 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 31, Figure 32) 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-year 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), an integral measure of the quality of the forecast probability distribution, is computed as

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

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

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

where CRPSclim is the CRPS of a climate forecast (based either on the ERA-Interim analysis or observed climatology). CRPSS is used to measure the long-term evolution of skill of the IFS ensemble (Figure 9) and its inter-annual variability (Figure 11).

ROC 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 the forecast user 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 is issued (Figure 36). Figure 36 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 40.

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 18, Figure 20) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 18, Figure 20). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 21 to Figure 24), 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 differences between model and true orography, using a crude constant lapse rate assumption provided the correction is less than 4 K amplitude (data are otherwise rejected).

A.4 Verification of rare events (Section 5.3)

Experimental verification of deterministic forecasts of rare events is performed using the symmetric extremal dependence index SEDI, which is computed as

$$SEDI = \frac{\log F - \log H - \log(1-F) + \log(1-H)}{\log F + \log H + \log(1-F) + \log(1-H)}$$

where F is the false alarm rate and H is the hit rate. In order to obtain a fair comparison between two forecasting systems using SEDI, the forecasts need to be calibrated (Ferro and Stephenson, 2011). Therefore SEDI is a measure of the potential skill of a forecast system. In order to get a fuller picture of the actual skill, the frequency bias of the uncalibrated forecast can be analysed. Another score which measures actual skill is the potential economic value (Richardson, 2000). It is computed as

$$PEV(\alpha) = \frac{\min(\alpha, B) - F\alpha(1-B) + HB(1-\alpha) - B}{\min(\alpha, B) - \alpha B}$$

where B is the base rate (observed frequency of occurrence) of the event, and α is the cost–loss ratio, which forms the x-axis of the PEV plot. The PEV can be interpreted as the economic gain (relative to climatology) obtained by performing action or non-action depending on the forecast. The relative value of a particular forecasting system depends on parameters α and B which are external to the system, and H and F which are model dependent (Richardson, 2000).

References

Ferro, C. A. T., and D. B. Stephenson, 2011: Extremal dependence indices: improved verification measures for deterministic forecasts of rare binary events. Wea. Forecasting, 26, 699–713.

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

Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. Q. J. R. Meteorol. Soc., 126, 649–667.

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