TECHNICAL MEMORANDUM

688

Verification statistics and evaluations of ECMWF forecasts in 2011-2012

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

Operations Department

November 2012

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

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

Verification pages have been created on the ECMWF web server and are regularly updated. Currently they are accessible at the following addresses:

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

2. Changes to the ECMWF forecasting system

The changes to the integrated forecasting system (IFS) since the preparation of documents for the last meeting of the Technical Advisory Committee are summarised below.

15 November 2011 Cycle 37r3, including the following main meteorological changes:

- Modification of the entrainment and detrainment of convection
- Modification of the supersaturation and deposition rate for clouds
- Modification of the surface roughness
- Assimilation of accumulated rainfall from NEXRAD radar data from the USA
- Assimilation of ozone observations from infrared radiances
- Bias correction of aircraft temperature observations
- Cycling of stratospheric model error (for the weak-constraint 4D-Var)

- Use of the latest version of the NWP-SAF radiative transfer model (RTTOV-10 including FASTEM-4)
- Retuning of cloud detection for the advanced infrared sounder data

Changes specific to the ensemble are:

- Use of the NEMO ocean model (instead of HOPE) in ensemble and use of the NEMOVAR ocean data assimilation system
- Coupling of the ensemble to the ocean model from day 10 onwards for the forecast from 12 UTC (as already done for the 00 UTC forecast)

This cycle also includes hourly post-processing of model data to 90 hours in support of the Boundary Conditions (BC) Optional Programme, including hourly wave data for all four forecast cycles. Hourly post-processed products were also introduced for the European limited area wave model.

19 June 2012 Cycle 38r1, including the following main meteorological changes:

- Updated background-error covariance statistics for data assimilation based on the cycle 37r2 EDA
- Improved statistical filtering of EDA sampled background errors
- Assimilation of MHS channel 5 over land and the all-sky radiance product from SEVIRI on board Meteosat-9
- Modified convective downdraught entrainment
- Improved de-aliasing of the pressure gradient term, reducing numerical noise, allowing a reduction of the horizontal diffusion in the forecast
- Improved description of swell in the ocean wave model

Changes specific to the ensemble are:

- Use of a new surface reanalysis to initialise the surface fields in the ensemble re-forecasts
- Extension of the ensemble re-forecasts from 18 to 20 years, all based on ERA-Interim initial data
- Redefinition of the EDA perturbations using the EDA ensemble mean instead of the EDA control as the reference

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 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%. The hemispheric scores have been consistently very good over the past year, remaining above 80% to around forecast day 6 and up to day 7 in early winter in both hemispheres. As is typical for a smaller region, there is more variability in the scores for Europe, but the 80% threshold is also reached typically at around forecast day 6.

As noted in last year's report, the record high scores achieved for Europe and the northern hemisphere in winters 2009–10 and 2010–2011 were partially a result of large anomalies associated with the strong negative phase of the North Atlantic oscillation circulation pattern that dominated those winters. The effects of the year-to-year variations in 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 although the scores were lower in winter 2011–12, the lead of the operational forecast over ERA-Interim was actually larger than in the previous winters, consistent with the benefits expected from the upgrades to the model made during the year.

Figure 2 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 2012. Figure 3 shows the RMS error of the seven-day persistence forecast (the reference system for Figure 2) for Europe. After a number of years of relatively high synoptic activity (and correspondingly poor persistence forecasts), the level of activity has been reduced in the last year and persistence forecast errors have reduced accordingly. This relative improvement in the performance of the reference forecast (persistence) is the reason for the drop in skill in the later forecast steps seen in Figure 2.

Figure 4 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) and summer (June–August, lower panel) season since winter 1997–98. The exceptional winter 2009–10 and 2010–11 performance is apparent for the 850 temperature scores, with a greater fraction of the individual forecasts achieving very high ACC scores than in previous years. The proportion of very good forecasts for winter 2011–12 is, as expected, lower than previous two winters but compares well with previous years. Summer 2012 scores are exceptionally good with a large proportion of very good and a significant reduction of poor forecasts at the seven-day range.

Figure 5 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. 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 ensemble performance is illustrated in Figure 7, 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, there is still a strong signal that the ensemble performance has been consistently high since 2010, when compared to previous years. A number of changes have been made to the ensemble configuration in this period, 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, see section 2). 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 the last three winters are shown in Figure 8. The match between the spread and error did not change significantly in 2011–12 compared to the previous winter season. The over-dispersion of the ensemble for 500 hPa height in the early forecast range, noted in previous years, is no longer apparent and the under-dispersion at longer ranges is further reduced. In general, the ensemble is still under-dispersive for temperature at 850 hPa, although uncertainty in the verifying analysis should be taken into account when considering the relationship between spread and error in the first few days. The introduction of the ensemble of data assimilations (EDA) in the initial perturbations (June 2010) improved the short-range over-dispersion, while this, together with the changes to the representation of model uncertainty, improved the overall dispersion of the ensemble for both parameters. Figure 9 shows the skill of the ensemble using CRPSS for days 1 to 15 for winter over the extratropical northern hemisphere. In November 2006 the ensemble was extended to 15 days, at reduced horizontal resolution beyond day 10. The ensemble has demonstrated consistent skill out to 15 days since this extension, confirming the positive skill at this forecast range. 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 the previous two winters, the performance for 2012–12 compares well with previous years.

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 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 four years.

3.2. ECMWF versus other NWP centres

The common ground for comparison is the regular exchange of scores between WMO designated global data-processing and forecasting system (GDPFS) centres under WMO CBS auspices, following agreed standards of verification. The new scoring procedures for upper-air fields used in the rest of this report were approved for use in this score exchange by the 16th WMO Congress in 2011 and 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 11 (northern hemisphere extratropics) and Figure 12 (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 13, 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 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 UK Met Office performance is similar. The errors of the JMA (Japan Meteorological Agency) forecast system have steadily decreased over several years and are now comparable with those of the UK Met Office model at both short and medium ranges. In the tropics, verification against analyses (Figure 14) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations, the ECMWF, UK Met Office and JMA models have very similar short-range errors.

4. Weather parameters and ocean waves

4.1. Weather parameters - high-resolution and ensemble

The new supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 16. 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%. 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%. Both scores are verified against station observations. The increase in skill of the high-resolution forecast in 2010 is 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 is 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.

While both supplementary headline scores for precipitation show an increase of skill over the last 10–15 years (Figure 16) they differ with respect to inter-annual variations. The new microphysics scheme introduced in November 2010, for example, which resulted in a substantial improvement in SEEPS for the high-resolution forecast, did not increase the probabilistic score (CRPSS Figure 16, lower panel). Investigation has shown that this is due to different sensitivities of the two scores to precipitation intensity and to the annual variations in observed precipitation. Figure 17 shows that the basic error measures for the ensemble control forecast (SEEPS and mean absolute error) and the ensemble probabilistic forecast (continuous ranked probability score, CRPS) all decreased (i.e. improved) during the 2010–2012 period. Figure 17 also shows (dotted lines) the corresponding scores for reference systems: the deterministic forecasts from the ERA-Interim system (which uses a fixed version of the model) and the CRPS for a fixed climatology forecast; these scores have also decreased substantially over the same period, indicating that the year-to-year variations in the observed precipitation has influenced all the performance measures. In the case of SEEPS, which gives large weight to light precipitation, a net improvement remains. In the case of CRPSS the increase in skill is masked by the (coincidental) increase in skill of the climatology forecast over the same period, and by the fact that light precipitation cases have less weight in CRPS than in SEEPS. The increase in climatological CRPS skill during that period can be traced back to a decrease in the number of observed heavy precipitation cases, from which the climatology forecast benefits more than the model forecast. In SEEPS these cases have relatively less weight, and the evolution of the score is more strongly determined by changes in light precipitation forecast skill.

ECMWF has begun 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 18). The reversed ranking of the JMA and Met Office ensembles at short lead times is due to a greater drop in skill in the JMA model during the northern hemisphere convective season (JJA) compared to the other models.

Compared to other global models, the ECMWF precipitation forecast shows a relative weakness in the first day of the forecast. It is most visible in the scores for Europe but can also be seen in the extratropics in general (Figure 18, top panel). While ECMWF has the best forecast from day 2 onwards, it drops behind the Met Office model at day 1 during the non-convective season. This does not occur in the tropics, where the lead of ECMWF relative to the other models is consistent throughout the six-day forecast range.

The relative weakness of extratropical ECMWF SEEPS scores at day 1 is related to an over-forecasting of light precipitation events when no precipitation was observed (see also Haiden et al. 2012). The frequency distribution of ECMWF forecasts at day 2 is closer to the observed distribution than it is at day 1. Both the convective and the large-scale part of the precipitation forecast contribute to the problem. This behaviour (too often forecasting light precipitation at the short range) is not so apparent for the Met Office or JMA models.

The summer 2011 edition of the ECMWF Newsletter (No. 128) contains an article about the SEEPS score used for the verification of the deterministic precipitation forecasts. A detailed comparison of the performance of the different global forecast models using SEEPS has been published in Haiden et al. 2012.

Long-term trends in mean error and standard deviation of error for 2 m temperature, specific humidity, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 19 to Figure 22. 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. There is a clear change in the 10 m wind speed bias (Figure 22) 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.

In recent winters and early spring there have been significant negative night-time temperature biases over Europe (Figure 19). A particular factor in winter 2010–11 was an increase in the number of large negative errors (forecast was too cold) at night, especially in cold conditions. These large errors occurred particularly in Scandinavia and Eastern Europe. Investigations identified a problem related to the model's handling of low cloud in very cold conditions. Model improvements to address the issue were implemented in November 2011 (cycle 37r3); pre-operational testing confirmed this significantly reduced the number of large errors and overall reduced the negative night-time bias in Europe in winter by 0.2–0.3 K. A comparison of the geographical distribution of the operational forecast biases in the winters 2010–11 and 2011–12 shows large reduction in bias over north-eastern Europe (Figure 23). However, to a large extent these differences are due to inter-annual variability, which is evident from a comparison with ERA-Interim (Figure 24). The distribution of forecast errors shows nearly the same improvement from 2010–11 to 2011–12 (narrowing of the distribution, reduction of errors) in ERA-Interim as in the operational forecast. Only the mean error (bias) behaves differently. While it has become slightly less positive in ERA-Interim, it is slightly less negative in the operational forecast. The difference (0.27 K) is consistent with what was expected from the model cycle changes.

Figure 25 shows the geographical distribution of mean night-time (top) and day-time (bottom) temperature errors for June to August 2012 (right) compared to 2011 (left). In general the variation and magnitude of the bias is similar for both years, although there is some increase in night-time bias over eastern Central Europe and part of the Balkans. Preliminary investigations into possible causes have not yet provided conclusive results and this may also be more a reflection of inter-annual variability.

To complement the evaluation of surface weather forecast skill, a new initiative towards operational verification of cloudiness and radiation using satellite data has been started. First results have been obtained for verification against the downward surface solar radiation product (daily totals) from the Climate Monitoring Satellite Application Facility (CM-SAF) based on Meteosat data. Fluxes have been made non-dimensional by scaling with a latitudinally and seasonally varying clear-sky flux. Figure 26 (top panels) shows the mean error and standard deviation of the error at forecast day 1 for the year 2011. The largest positive biases (overestimation of solar radiation at the surface) are found in the Southern Ocean, off the coast of south-western Africa, and in the Sahara region. Comparison with biases in the top of the atmosphere (TOA) reflected solar radiation (not shown) confirms an underestimation of cloudiness in the two ocean regions, whereas the systematic error in the Sahara region is at least partly due to a bias in the CM-SAF product. The area of strongest negative bias near the Horn of Africa is mainly due to the aerosol climatology used in the model. In this region it overestimates the aerosol content during the summer months.

The non-systematic part of the short-range forecast error (Figure 26, top right panel) is largest in the tropics and in the Southern Ocean, with a somewhat less pronounced belt of higher values in northern mid-latitudes. The smallest non-systematic errors are found over subtropical land areas. In order to take the vastly different predictability of cloudiness in these different regions into account, a skill score based on the comparison with the climatological forecast has been developed (Figure 26, bottom panels). Skill is generally highest in the extratropics (with the exception of the Southern Ocean south of 50°S). In Europe, Mediterranean land areas show the highest skill. A large area of low skill can be seen in the southern sub-tropical Atlantic. More localised low skill in Brazil was found to be co-located with strong gradients in leaf-area index and soil moisture. Time-series (Figure 27) of daily radiation values suggest an overestimation of cloudiness during the dry season (June–September) in this area.

4.2. Ocean waves

The quality of the ocean wave model analysis is shown in the comparison with independent ocean buoy observations in Figure 28. The top panel of Figure 28 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; in general the wind errors have been lower in 2011–12 than in the previous year. The errors in the wave analysis have been consistently small over the past year, particularly in the most recent months. The long-term trend in the performance of the wave model forecasts is shown in Figure 29 and Figure 30; overall, the performance in 2010–11 is similar to that for the previous 12 months.

ECMWF maintains a regular inter-comparison of performance between wave models from different centres on behalf of the Expert Team on Waves and Storm Surges of the WMO-IOC Joint Technical Commission for Oceanography and Marine Meteorology (JCOMM). The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of ocean buoys (mainly located in the northern hemisphere). An example of this comparison is shown in

Figure 31 for the most recent three-month period (May–July 2012). 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 their errors in Figure 31. 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, including the figures shown in this report: http://www.ecmwf.int/products/forecasts/wavecharts/

5. Severe weather

Supplementary headline scores for severe weather are:

- The skill of the Extreme Forecast Index (EFI) for 10 m wind verified using the relative operating characteristic area (Section 5.1)
- The tropical cyclone position error for the 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 32 (top), together with the corresponding results for 24-hour total precipitation (bottom left) and 2 m temperature (bottom right). Each curve shows a four-season running mean of ROC area skill scores from 2004 to 2011; the final point on each curve includes the spring (March–May) season 2012. For 2 m temperature the EFI skill is maintained at a similar level to the previous year, while for precipitation and wind speed, there has been a clear improvement during the last year.

5.2. Tropical cyclones

The 2011 North Atlantic hurricane season had an above average number of tropical storms, consistent with the La Niña conditions. From May 2011 onwards, the seasonal tropical storm predictions consistently indicated enhanced activity over the Atlantic for the 2011 season (see section 6.3, Figure 38).

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 33. Errors in the forecast intensity of tropical cyclones (represented by the reported sea-level pressure at the centre of the system) are also shown in Figure 33.

The position errors (top left panel, Figure 33) are similar for the last four years, at around 200 km on average for the three-day forecast. The bottom right panel of Figure 33 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/hour) compared to the observed speed; there has been substantial reduction of this slow bias over last three years, with the latest 12-month period showing no overall bias in speed of movement. However, because of the substantial year-to-year variations in the number and intensity of cyclones, there is some uncertainty in these figures. The mean error (bias) in the tropical cyclone intensity has also improved in the last three years, showing no overall bias in intensity for the last year (top right panel, Figure 33). The mean absolute error of the tropical cyclone intensity is similar to last year (bottom right panel, Figure 33). As for the speed errors, there is a relatively large uncertainty in these scores because of the year-to-year variations in the number of storms.

The 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 34. Results show 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 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 since March 2008. The combined system enabled users to be provided 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 35 shows the ROC area score computed over each grid point for the 2 m temperature monthly forecast anomalies at two forecast ranges: days 12–18 and days 19–25. All the real-time monthly forecasts since 7 October 2004 have been used in this calculation. The red colours correspond to ROC scores higher than 0.5 (the monthly forecast has more skill than climatology). This is now achieved in all regions; stronger shades indicate the regions of higher skill. Currently, the anomalies are relative to the past 18-year model climatology. The monthly forecasts are verified against the ERA-Interim reanalysis or the operational analysis when ERA-Interim is not available. Although these scores are strongly subject to sampling, they provide users with a first estimate of the forecast skill's spatial distribution, showing that the monthly forecasts are more skilful than climatology over all areas.

Comprehensive verification for the monthly forecasts is available on the ECMWF website at:

http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/.

Figure 36 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. For the 12–18 day period, the monthly forecast has shown a clear substantial advantage over persistence of the medium-range (days 5–11) forecast throughout the last year. For the 19–32 day range, the system had a small but consistent 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. The 2011–2012 El Niño forecasts

The 2010 La Niña declined during the first half of 2011 and sea surface temperatures in the tropical Pacific were close to normal in June and July. This was followed by a return to La Niña conditions during the second half of 2011 and subsequent decline and transition to warm (El Niño) conditions by July 2012. Forecasts in 2011 captured well the decline of La Niña but initially showed considerable uncertainty over the subsequent development (Figure 37, top). However, once the cold conditions became established, forecasts gave a good indication of the transition that occurred during 2012 (Figure 37, top). The uncertainty in the earlier forecasts was equally apparent in the multi-model EUROSIP forecasts.

6.3. Tropical storm predictions from the seasonal forecasts

From May 2011 onwards, the seasonal tropical storm predictions indicated the probability of enhanced activity over the Atlantic for the 2011 season. The June forecast predicted between 9 and 17 named tropical storms in the Atlantic (Figure 38), although this was not a statistically significant signal. All later forecasts did give significant signals for above normal activity, consistent with the increasing confidence in the La Niña conditions discussed in the previous section. During the 2011 season 19 named tropical storms occurred, the same number as in 2010, although fewer of these developed into hurricanes in 2011 than in 2010.

The seasonal forecasts indicated near-normal tropical storm activity for both the western and eastern Pacific. These were also correct, although it should be remembered that evaluation using the seasonal re-forecasts shows that the skill of the tropical storm forecast in the eastern Pacific is generally low. However, verification shows that the skill in predicting accumulated cyclone energy (ACE) over the Atlantic basin, calculated using the most recent 20 years, is substantial, with a correlation of 0.65 between ensemble mean forecast and observation (Figure 39). There is also moderate skill overall for the western Pacific: the correlation for that region is 0.53.

6.4. Seasonal forecast performance for the global domain

A 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

As discussed above, the seasonal forecasting system performed well in predicting the evolution of La Niña from late 2011 (Figure 37). The System 4 forecast from November also provided a good prediction of the dipole over the Eurasian region during December to February. However the forecast for North America was less successful: in common with the other EUROSIP models, the forecast was for colder than normal in the north-west and warmer in the south-east of the continent, consistent with a typical teleconnection response to La Niña. In fact most of the North American continent, apart from the south-west, was warmer than normal in winter 2011–12.

References

Haiden, T., M. J. Rodwell, D. S. Richardson, A. Okagaki, T. Robinson, and T. Hewson, 2012: Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Wea. Rev.*, **140,** 2720-2733.

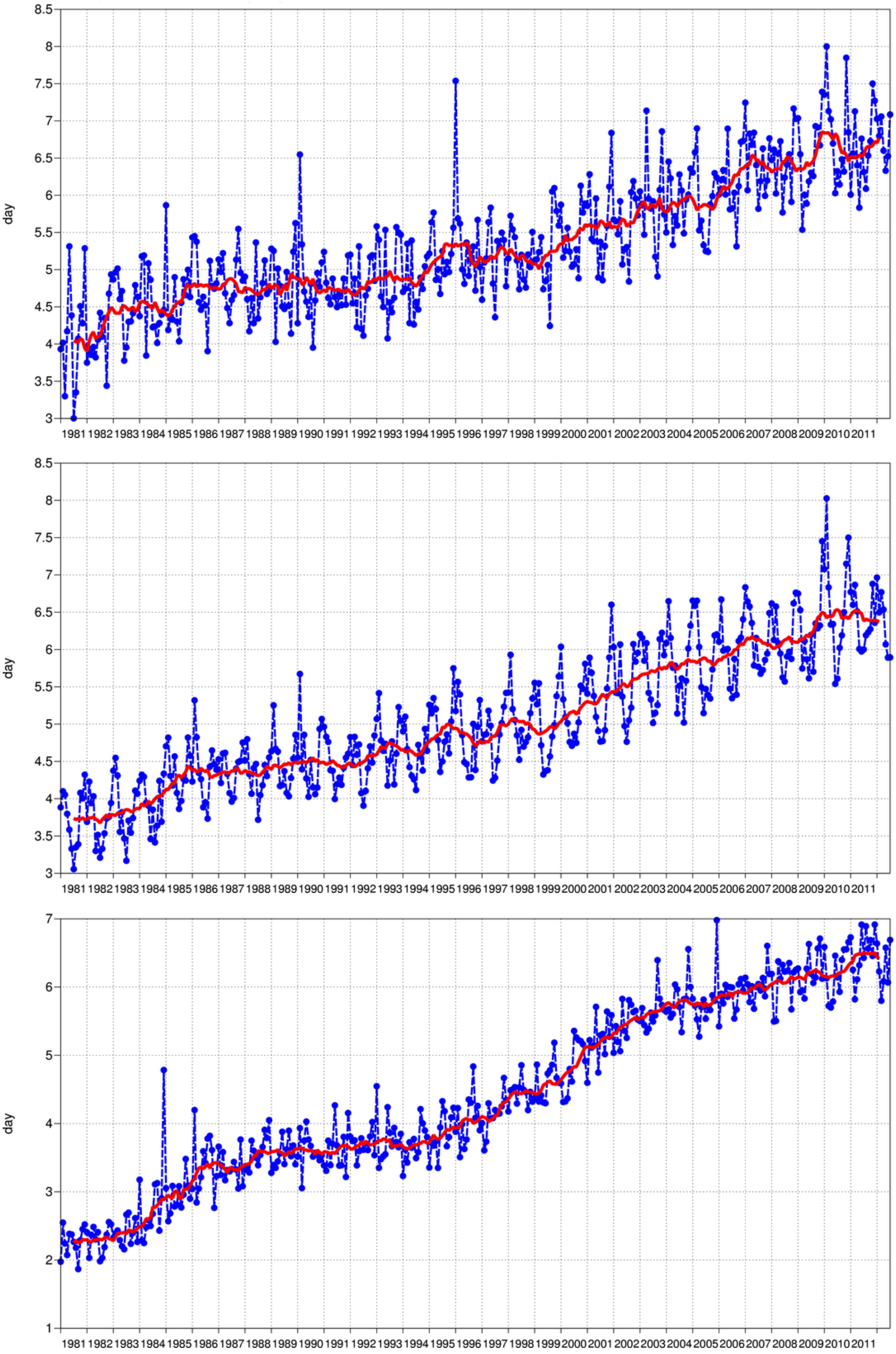
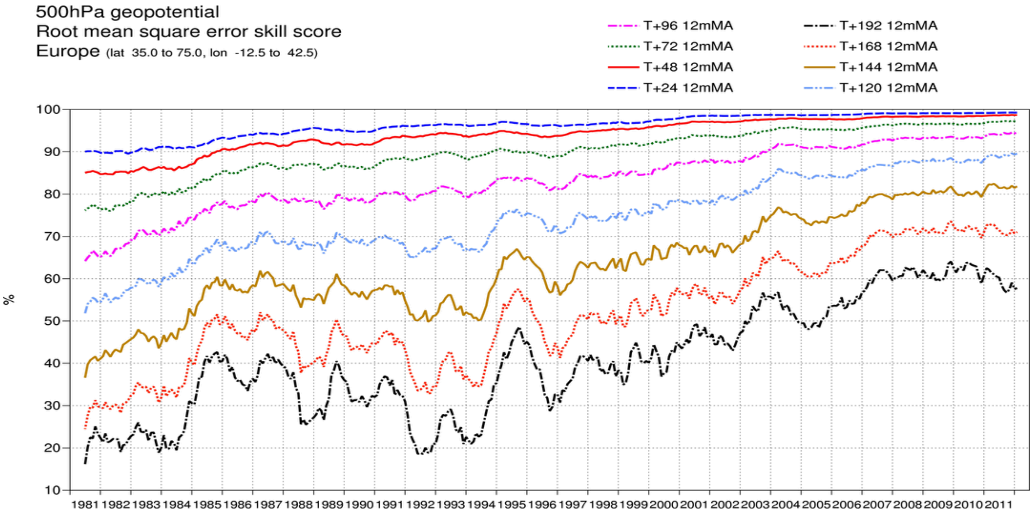


Figure 1: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly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



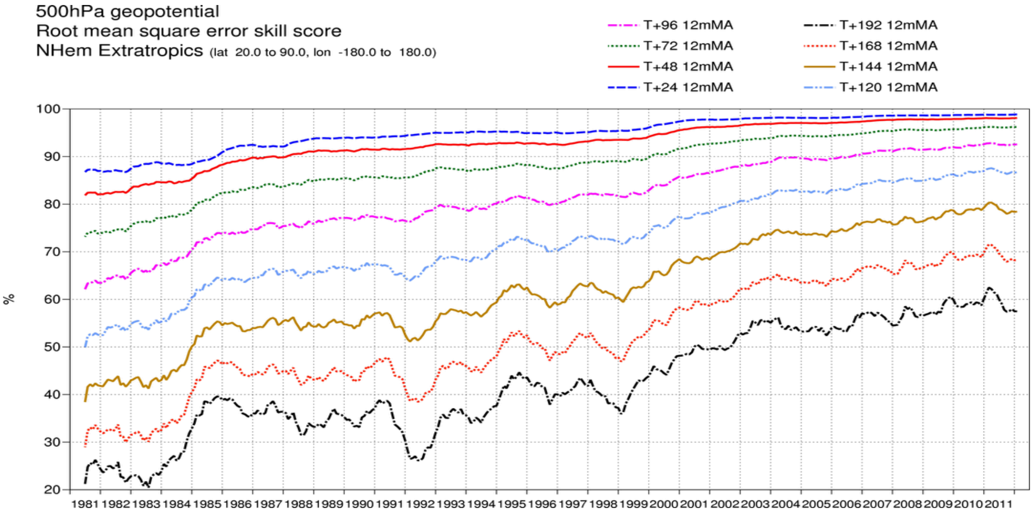


Figure 2: 500 hPa geopotential height skill score for Europe (top) and the northern hemisphere extratropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2011–July 2012.

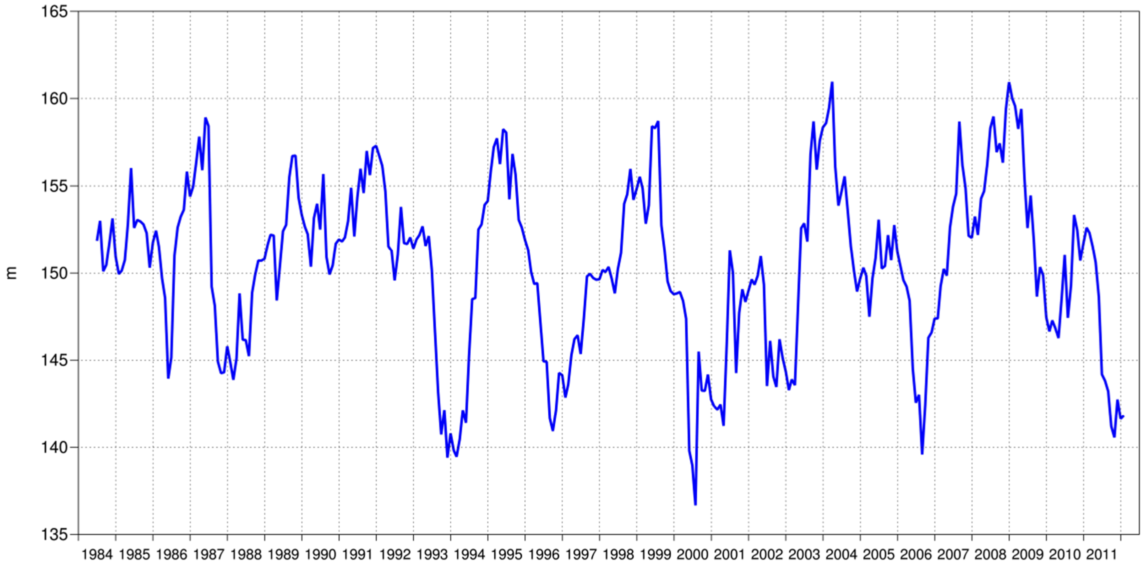
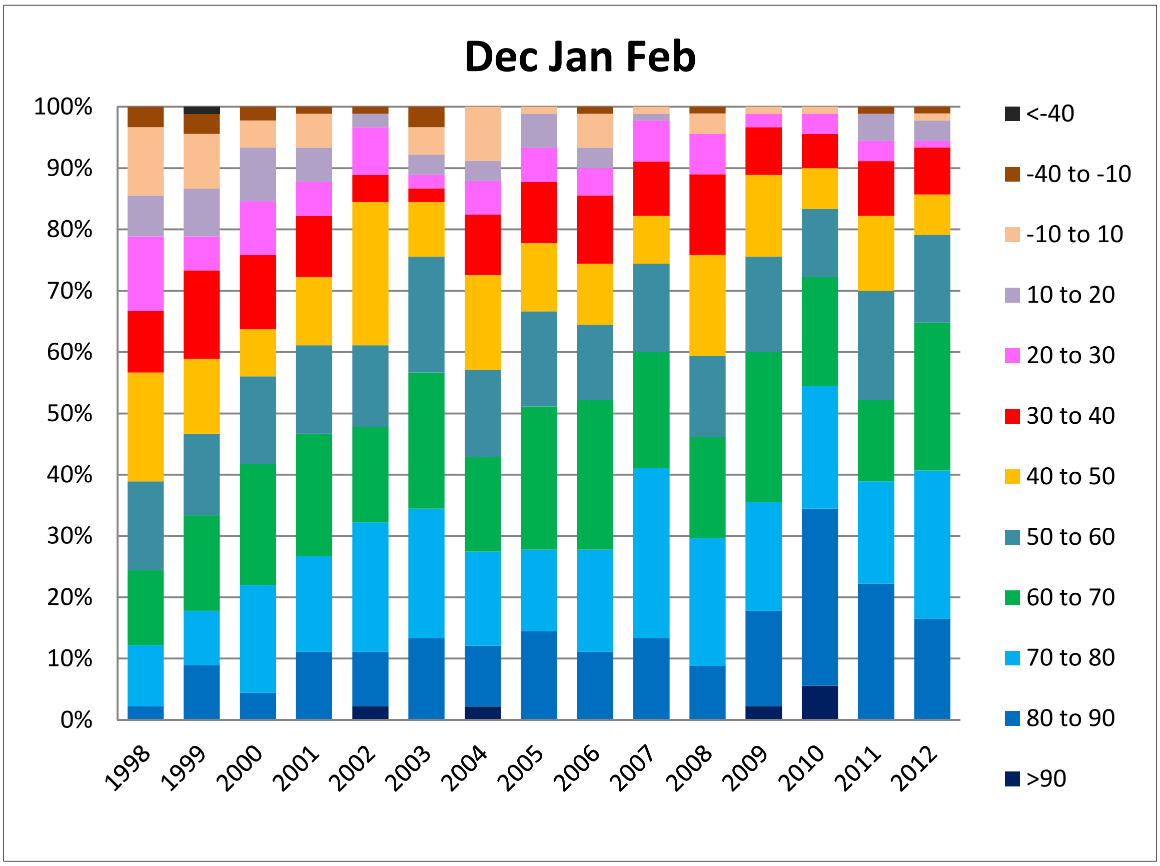


Figure 3: Root mean square (RMS) error of forecasts made by persisting the analysis over 7 days (168 hours) and verifying it as a forecast for 500 hPa geopotential height over Europe. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2011–July 2012.



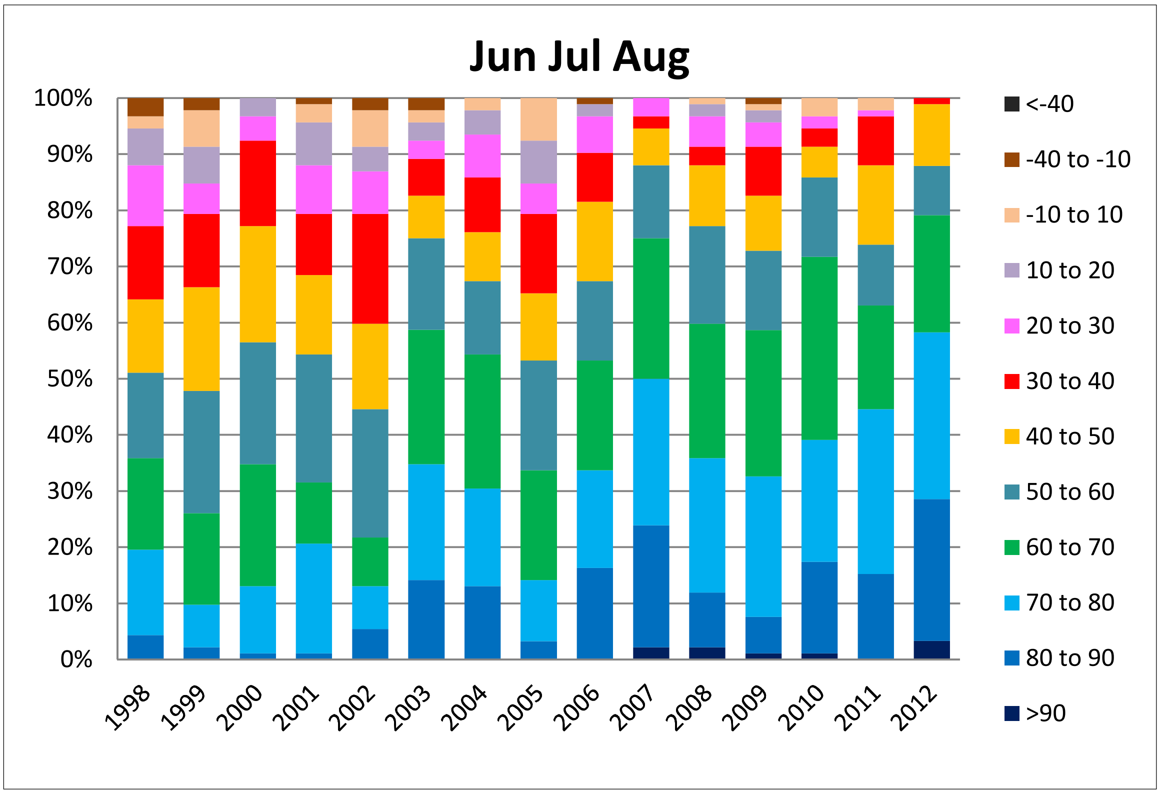
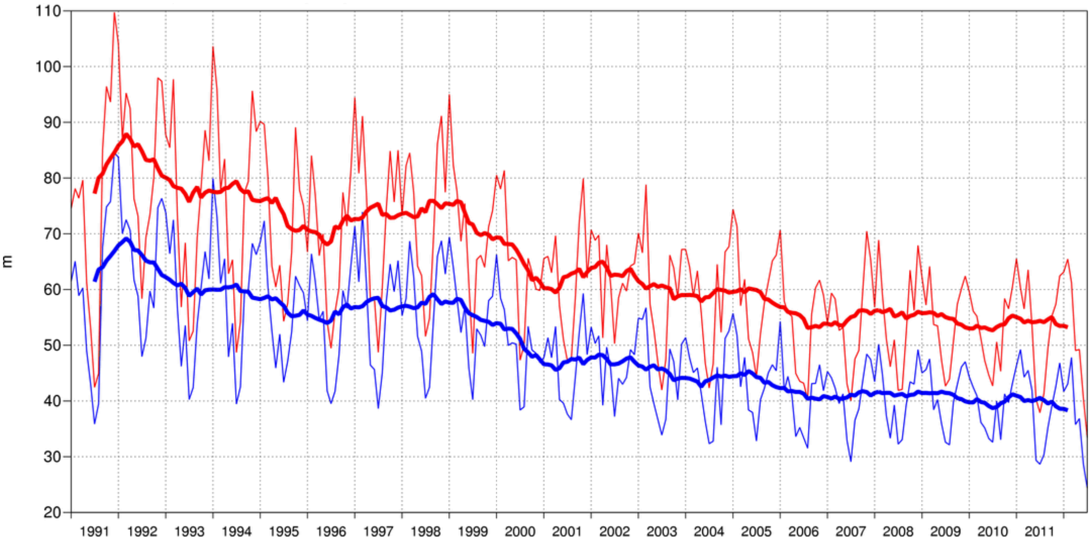


Figure 4: Distribution of ACC of the day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997–1998.



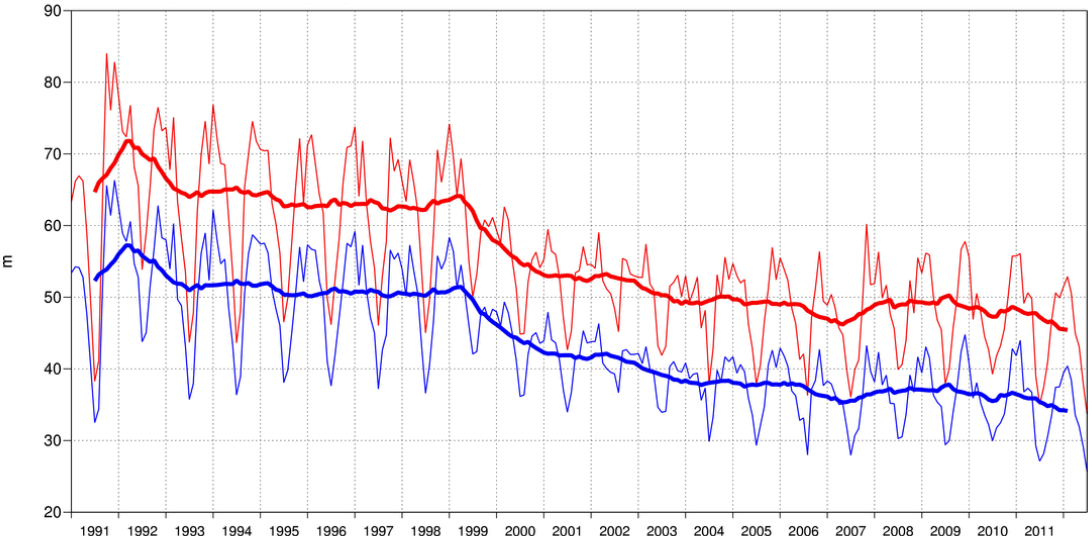


Figure 5: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores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

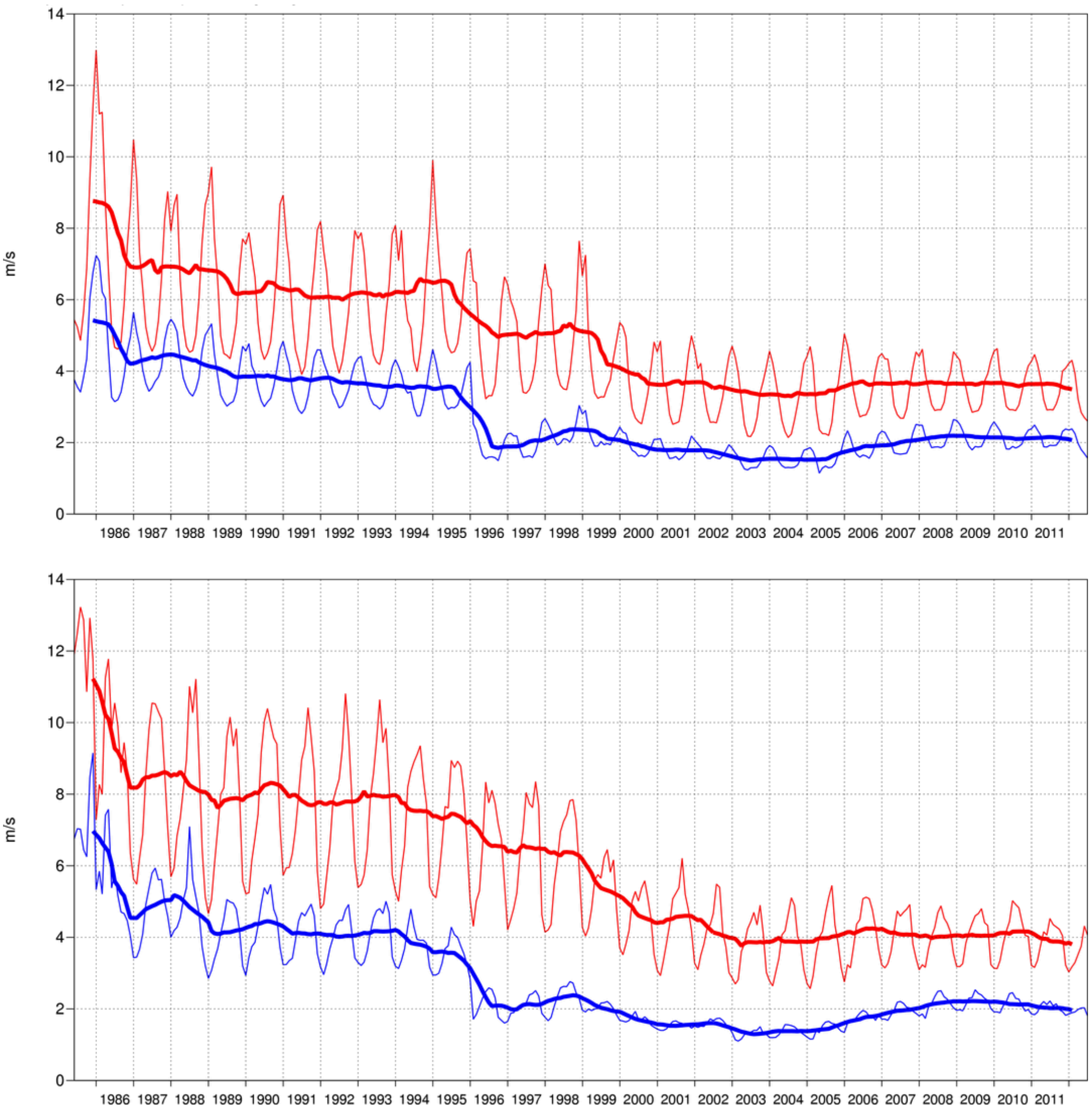


Figure 6: Model scores in the northern (top) and southern (bottom) extratropical stratosphere. Curves show the monthly average RMS vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).

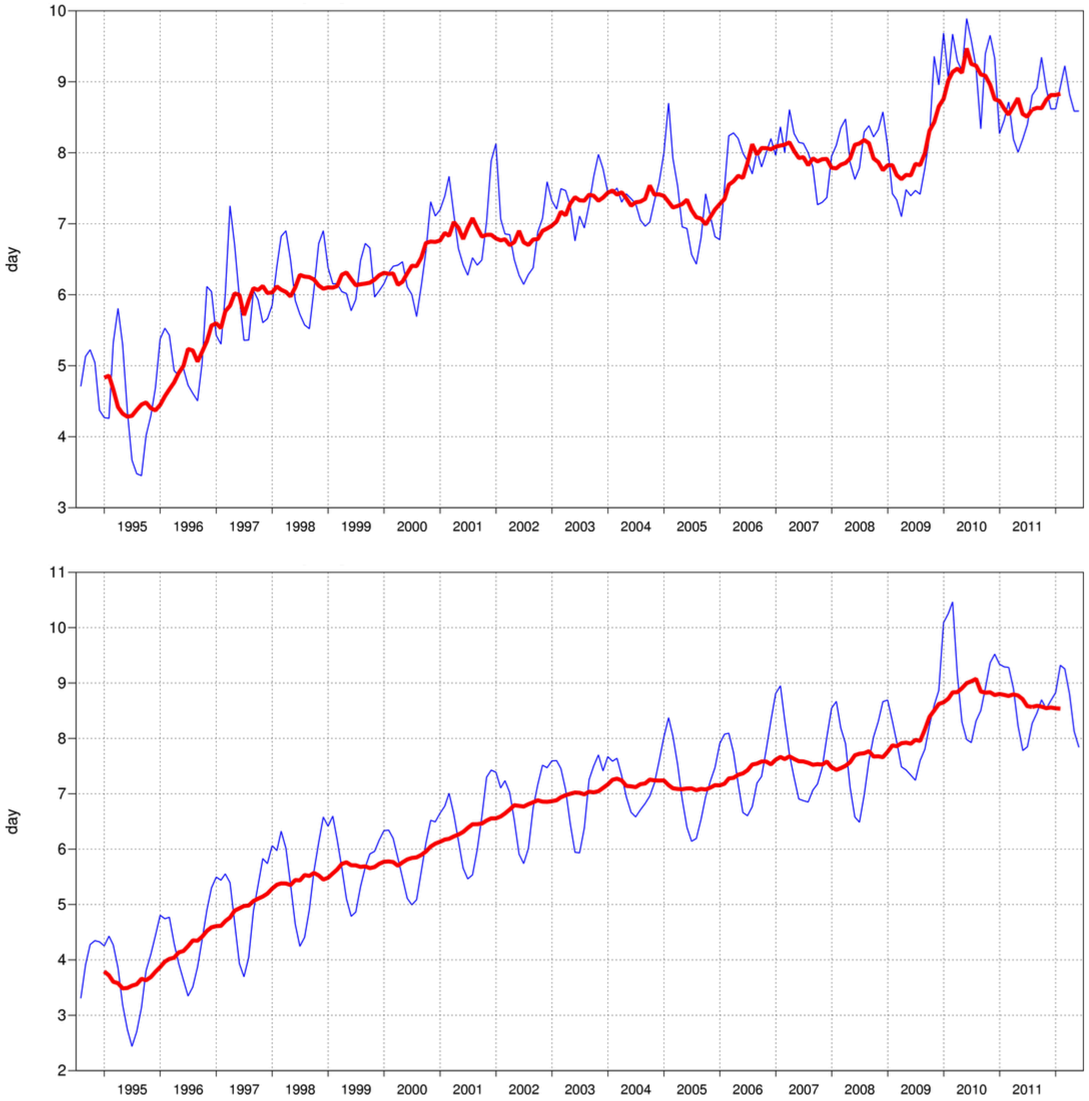


Figure 7: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance – each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous ranked probability skill score (CPRSS) falls below 25% for Europe (top), northern hemisphere extratropics (bottom).

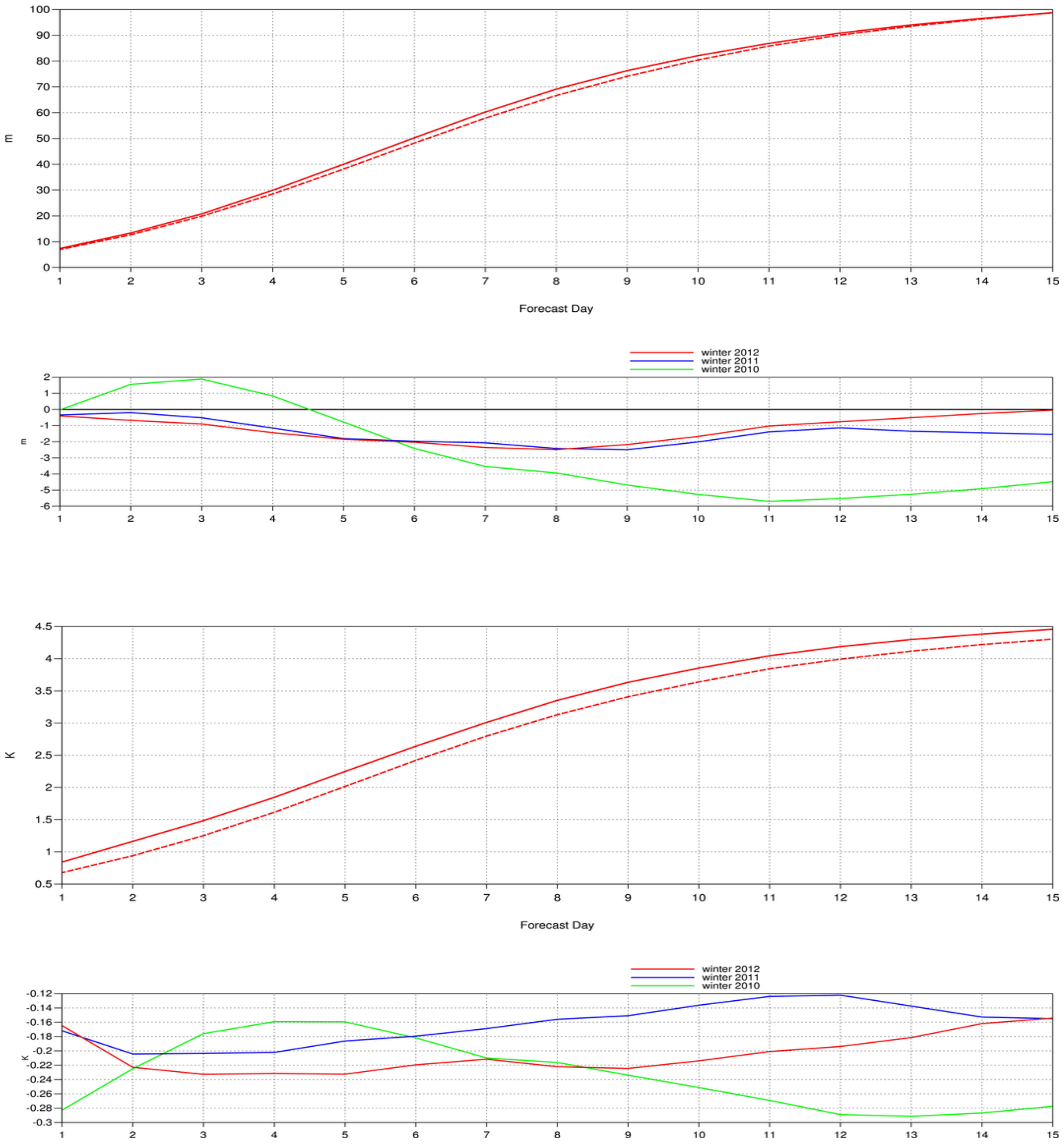
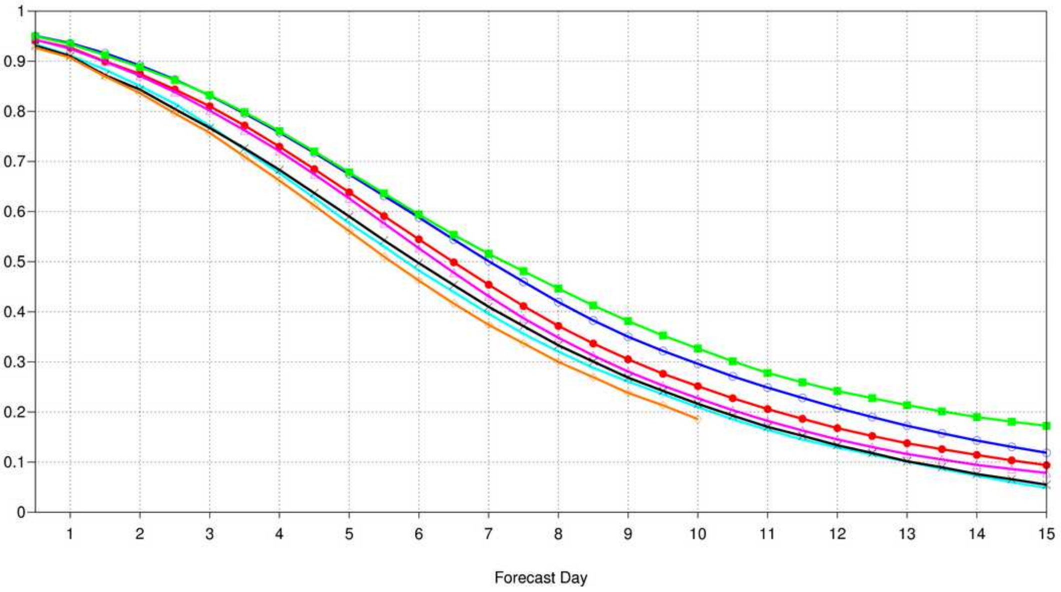


Figure 8: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2011–2012 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.



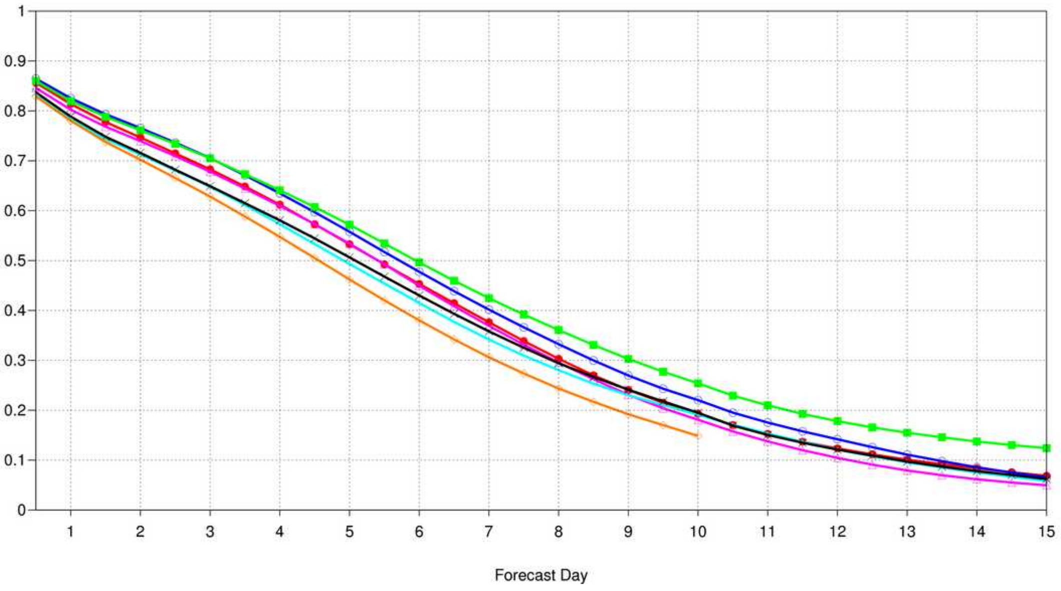


Figure 9: CPRSS for 500 hPa height (top) and 850 hPa temperature (bottom) ensemble forecasts for winter (December–February) over the extratropical northern hemisphere. Skill from the ensemble day 1–15 forecasts is shown for winters 2011–12 (red), 2010–11 (blue), 2009–10 (green), 2008–09 (magenta) and 2007–08 (cyan). The ensemble only ran to ten days in 2005–06 (orange).

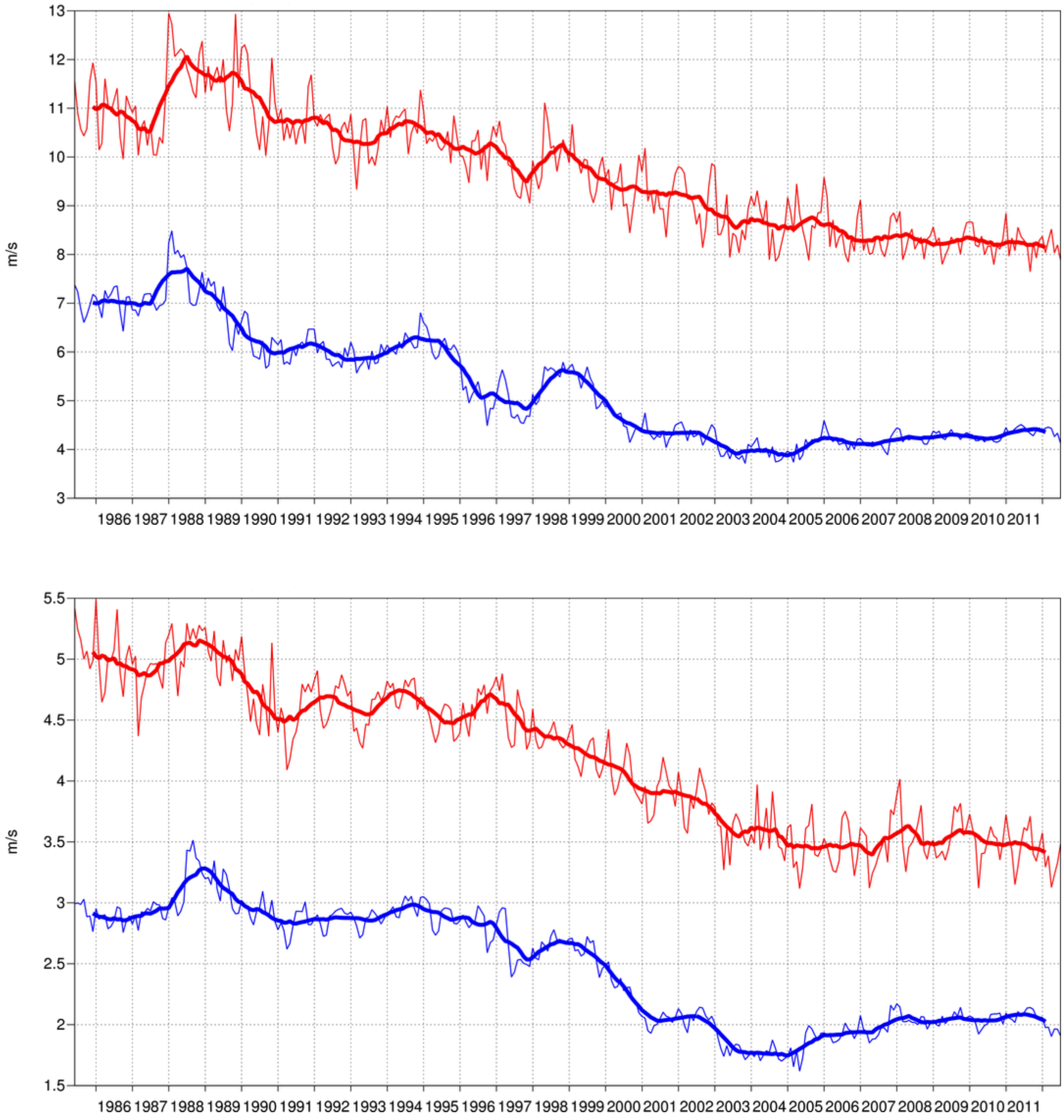
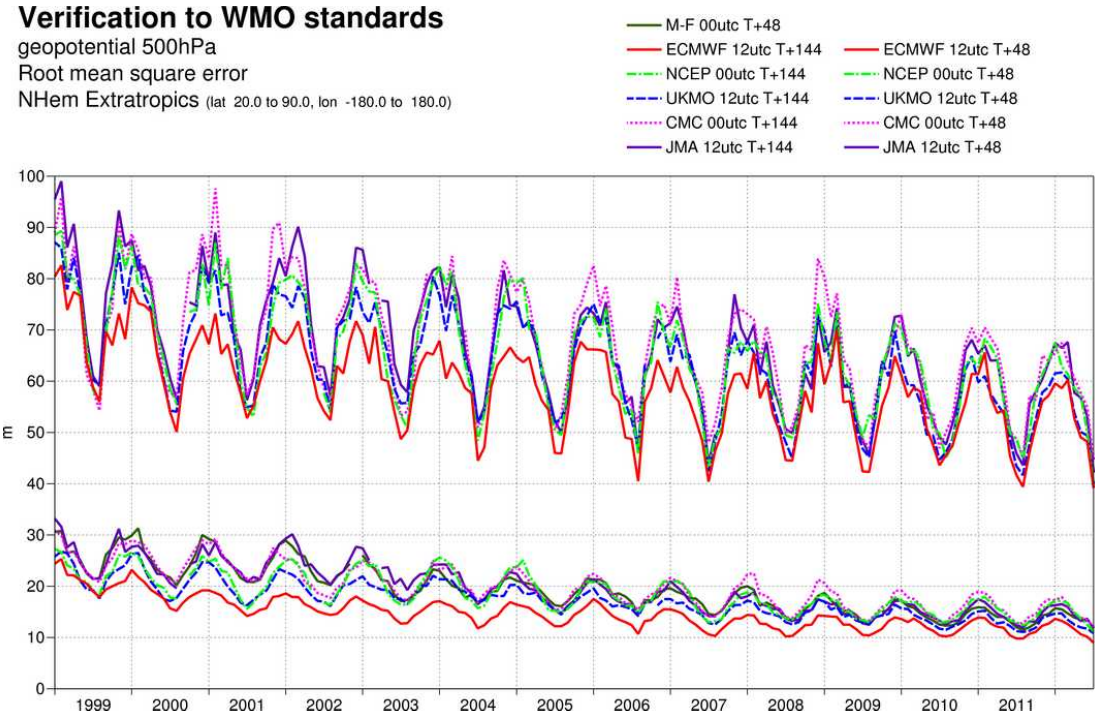


Figure 10: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).



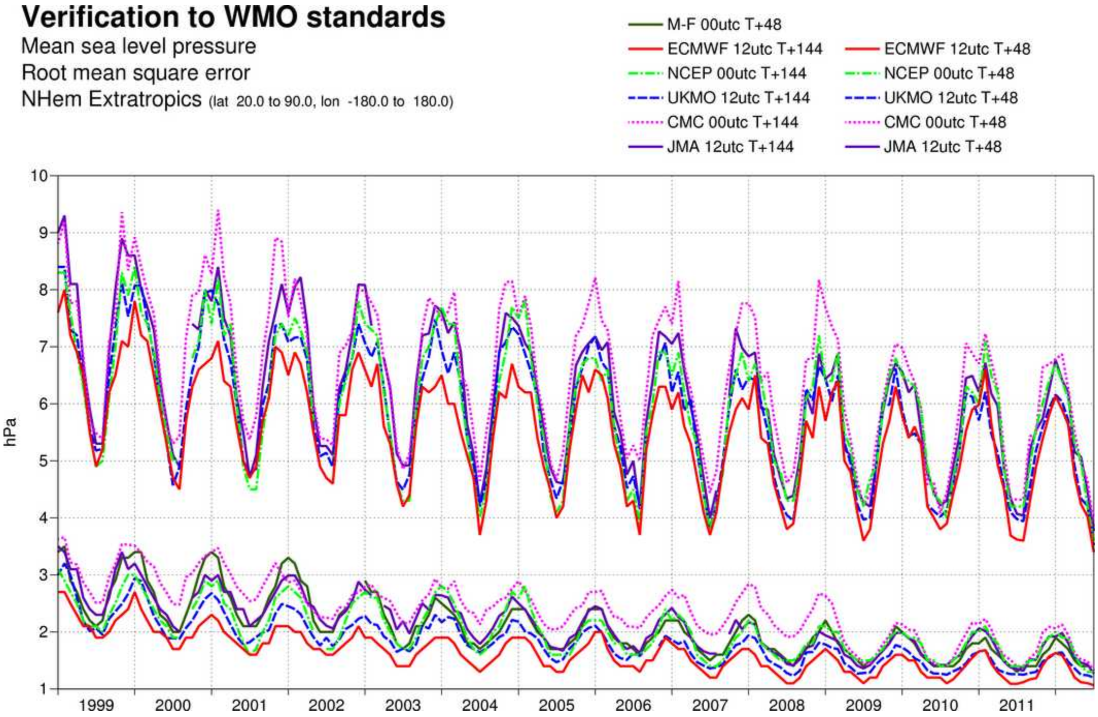
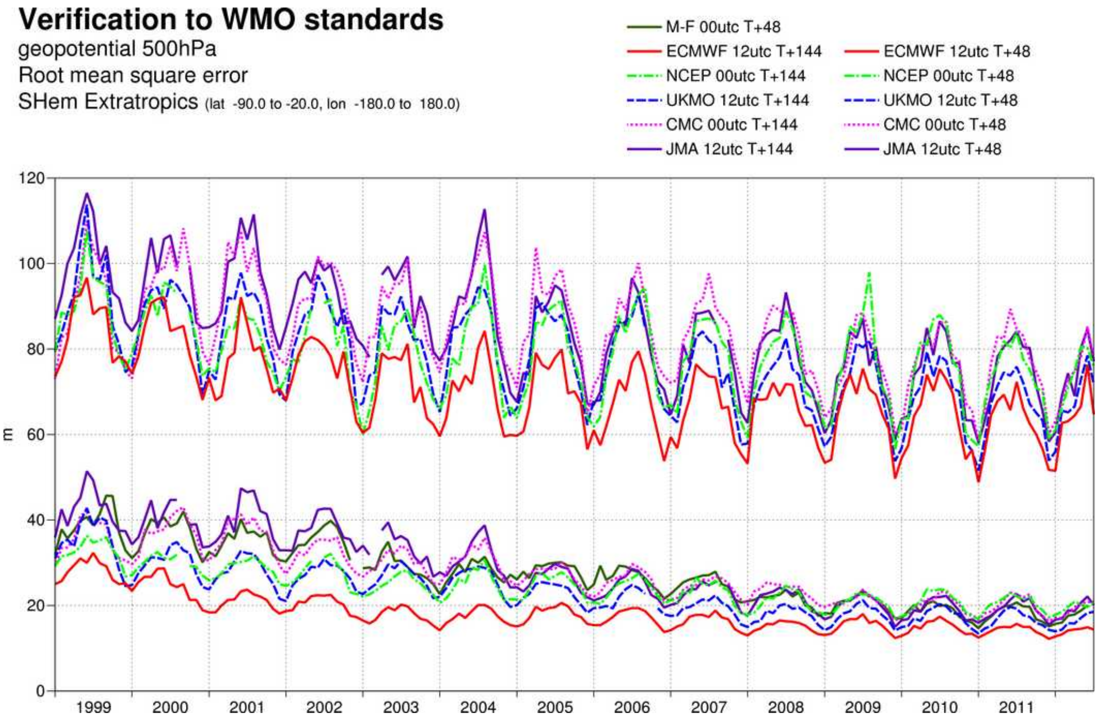


Figure 11: WMO-exchanged scores from global forecast centres. RMS error over northern extratropics for 500 hPa geopotential height (top) and mean sea level pressure (bottom). In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model 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.



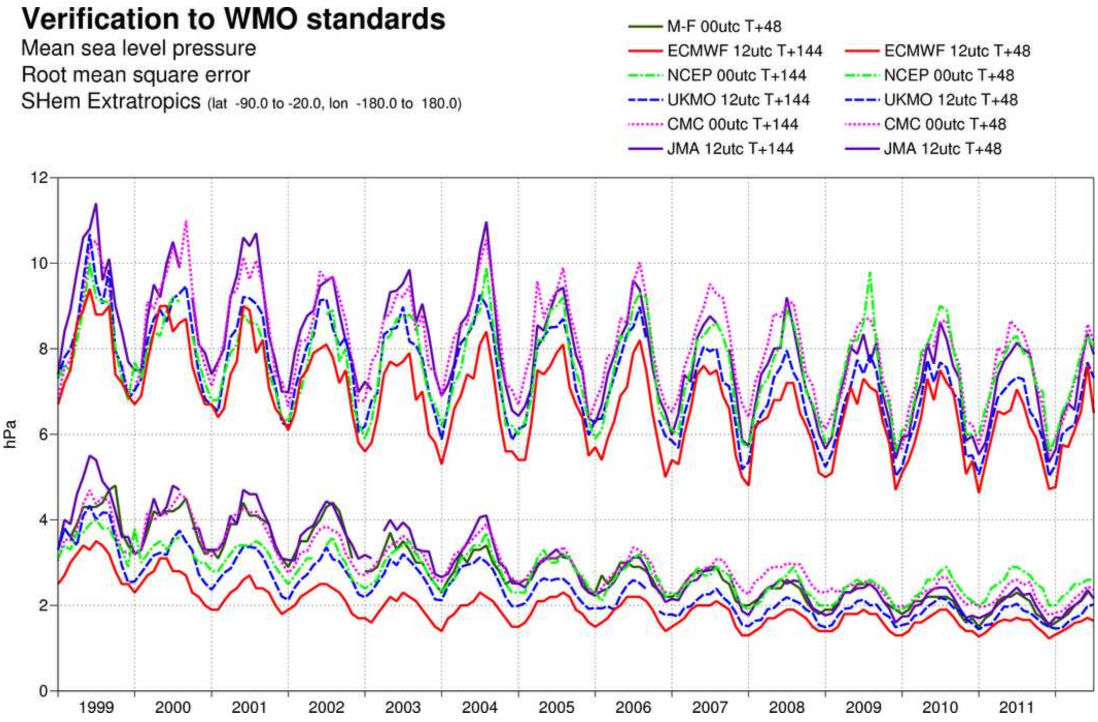
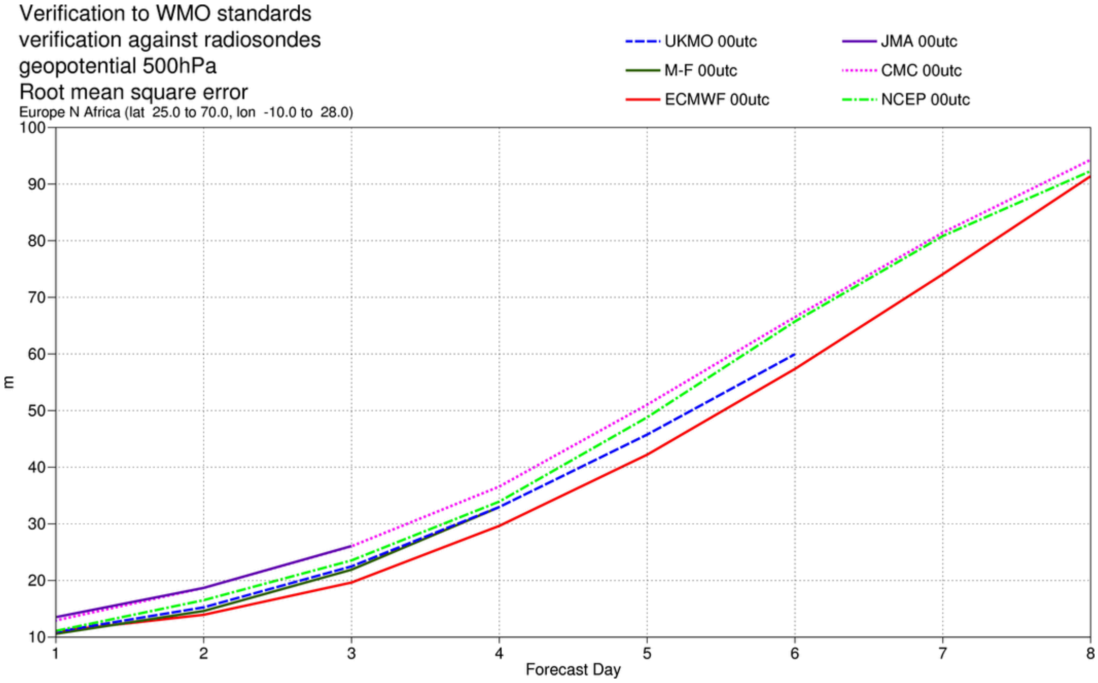


Figure 12: As Figure 11 for the southern hemisphere.



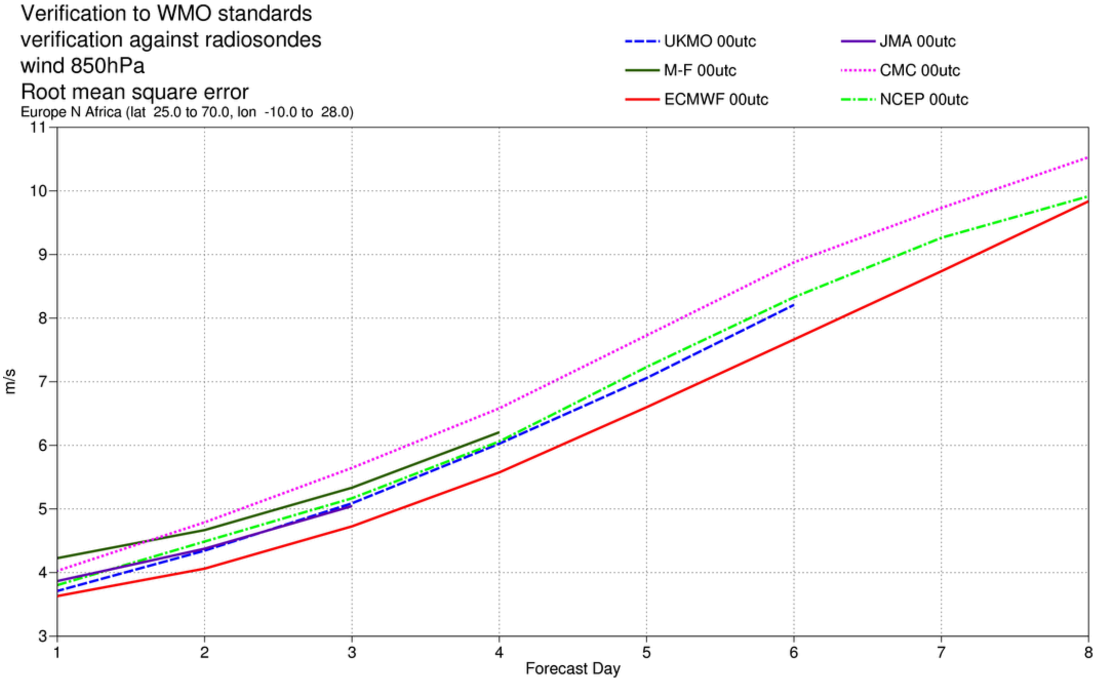
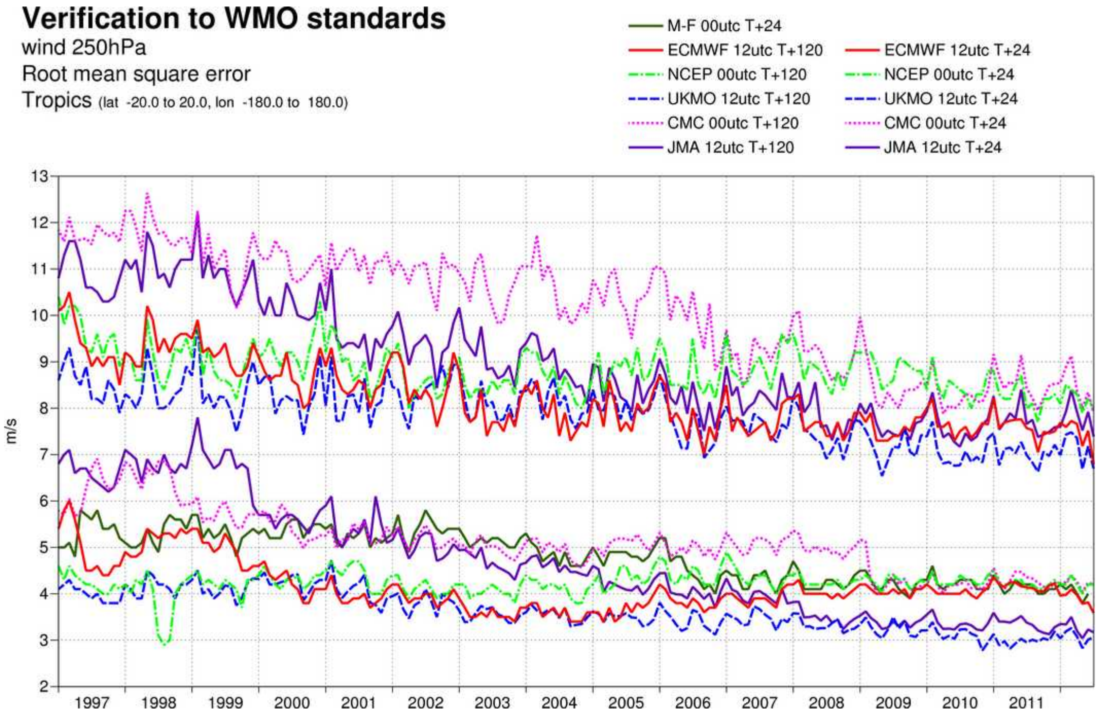


Figure 13: WMO-exchanged scores using radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2011–July 2012).



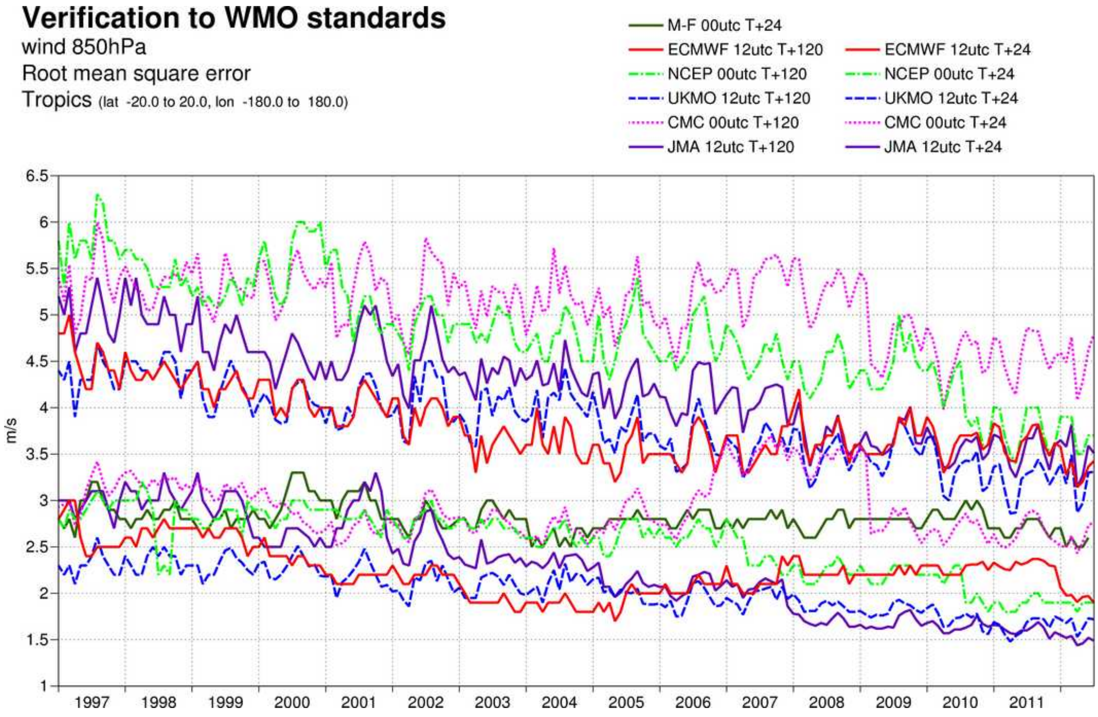
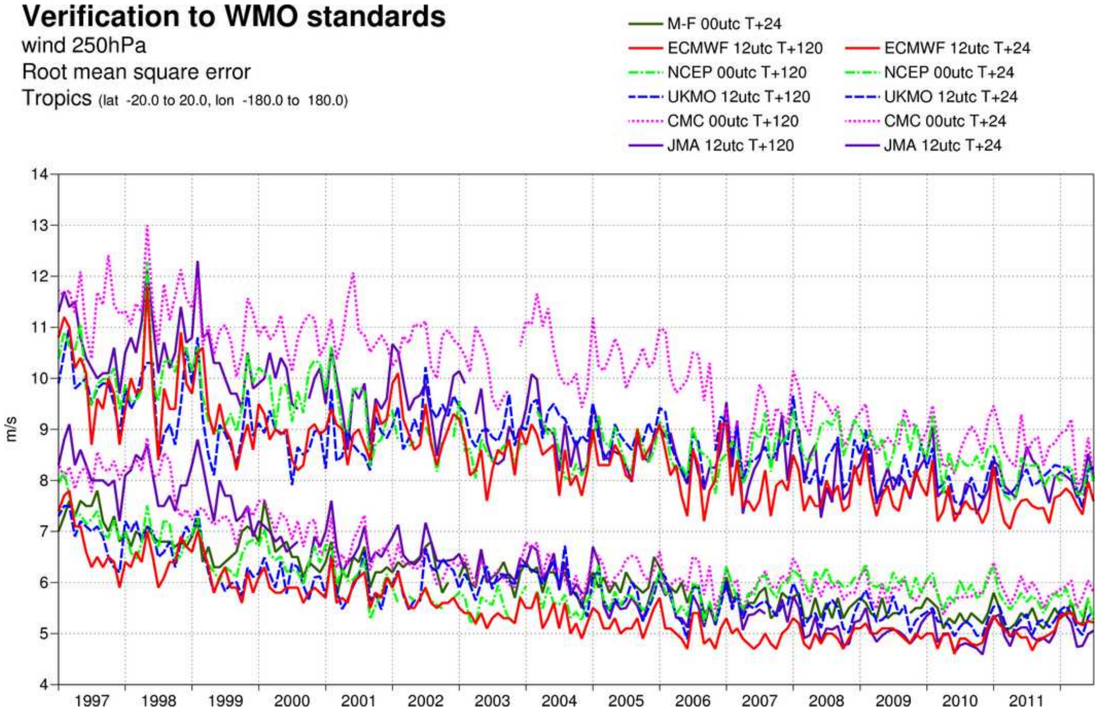


Figure 14: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



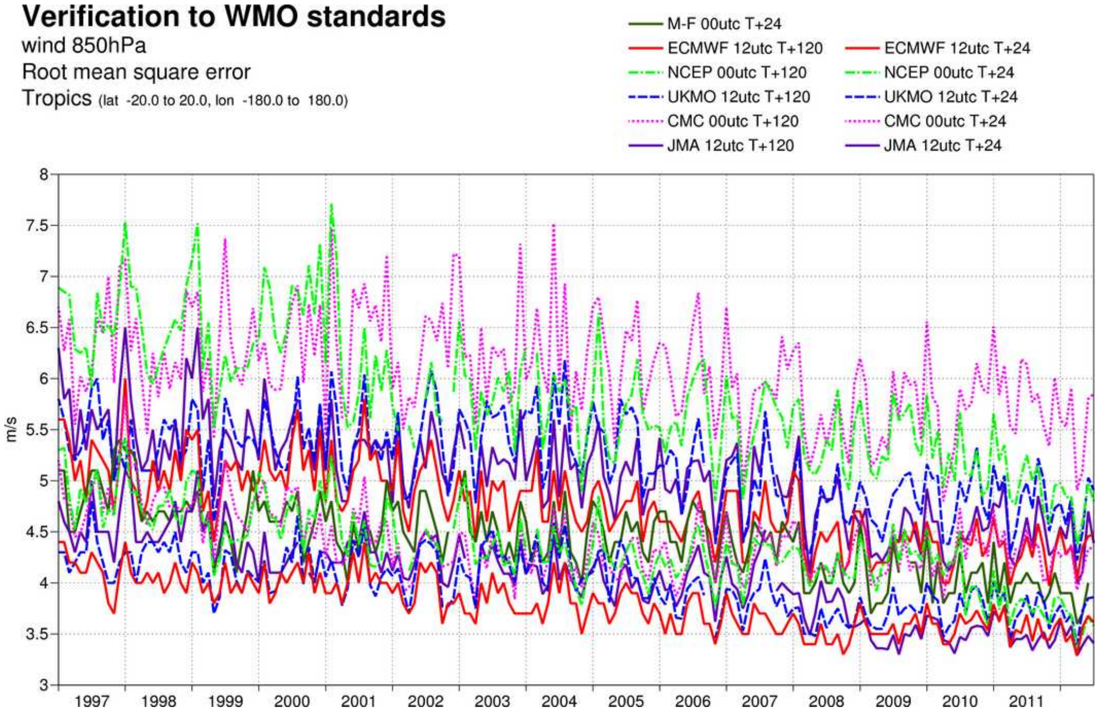
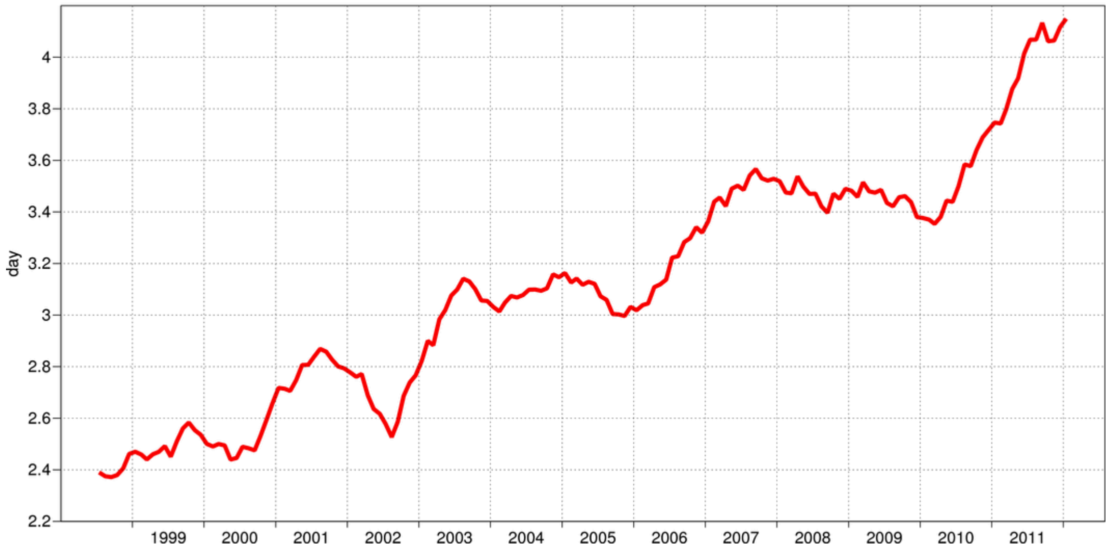


Figure 15: As Figure 14 for scores computed against radiosonde observations.



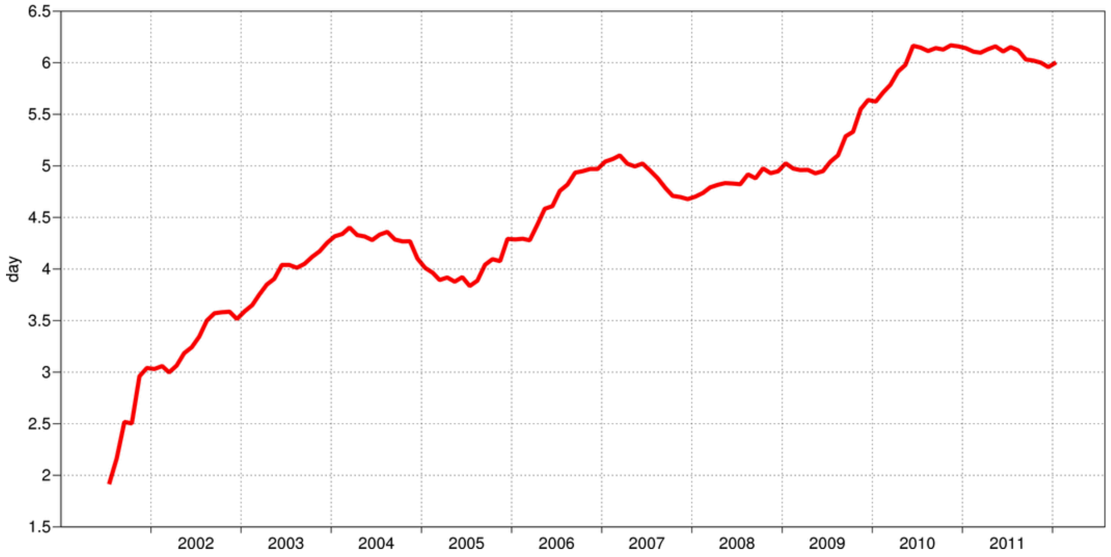


Figure 16: Supplementary headline scores for deterministic (top) and probabilistic (bottom) precipitation forecasts. Each curve shows the number of days for which the centred 12-month mean skill remains above a specified threshold for precipitation forecasts over the extratropics. In both cases the verification 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.

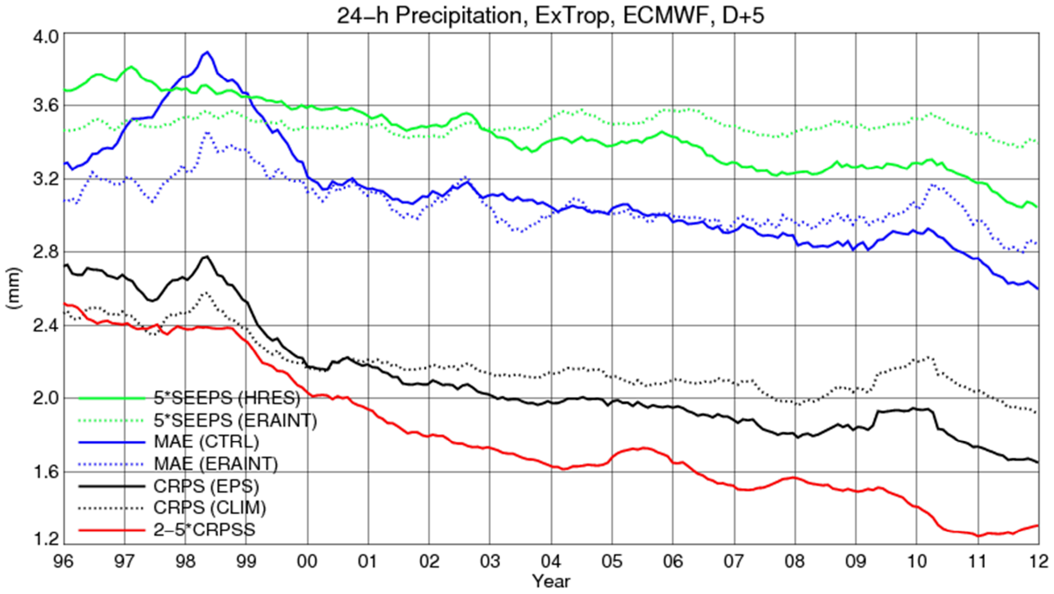


Figure 17: Comparison of the stable equitable error in probability space (SEEPS) score (green) of the high-resolution and the ERA-Interim forecast, the mean absolute error (blue) of the ensemble control and ERA-Interim, the continuous ranked probability score (CRPS, black) of the ensemble and climate, and the linearly scaled CRPSS, red. All model forecasts are for the extratropics, D+5. Values are centred running averages over one year.

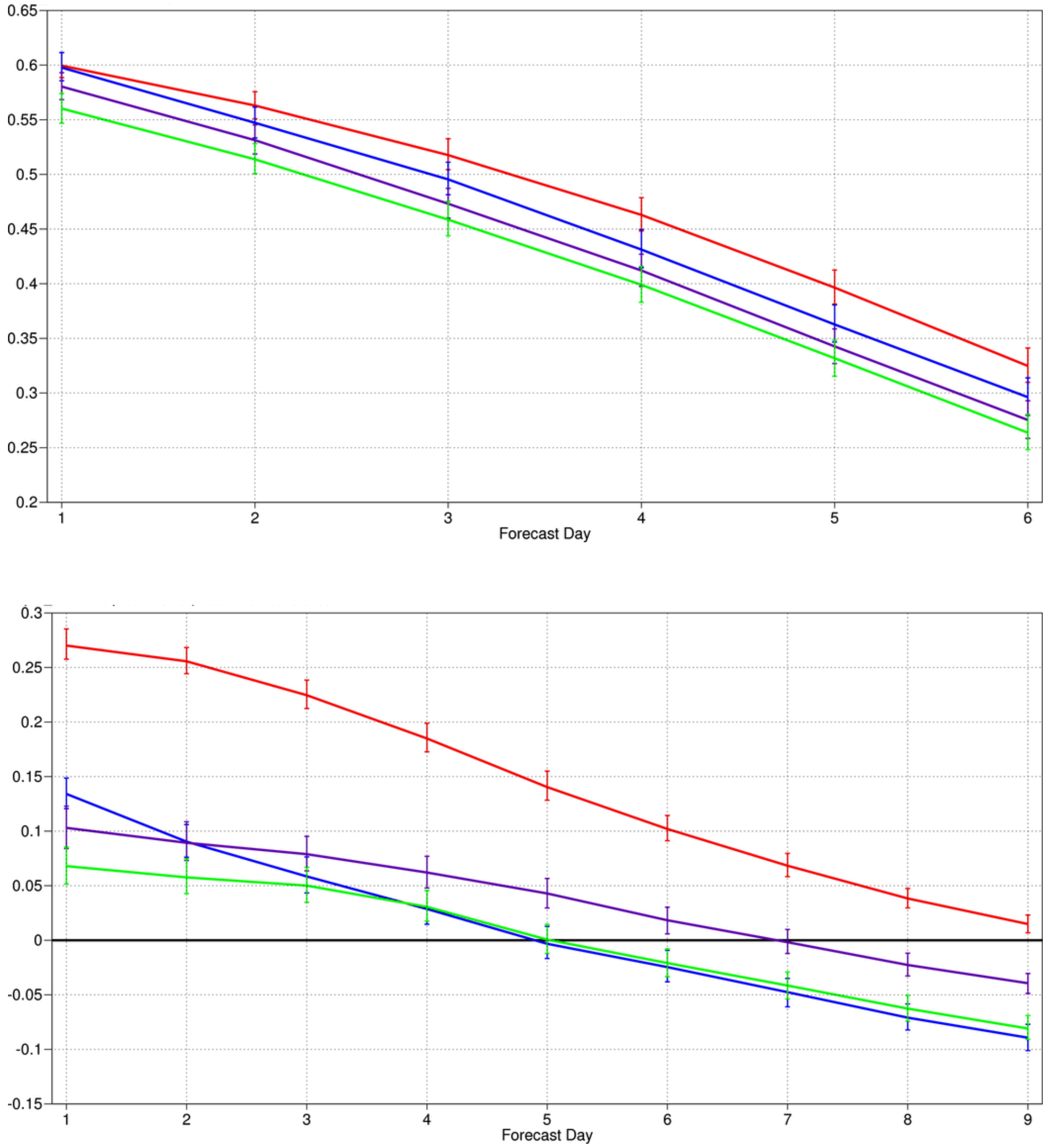


Figure 18: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2011 to July 2012. Bars indicate 95% confidence intervals.

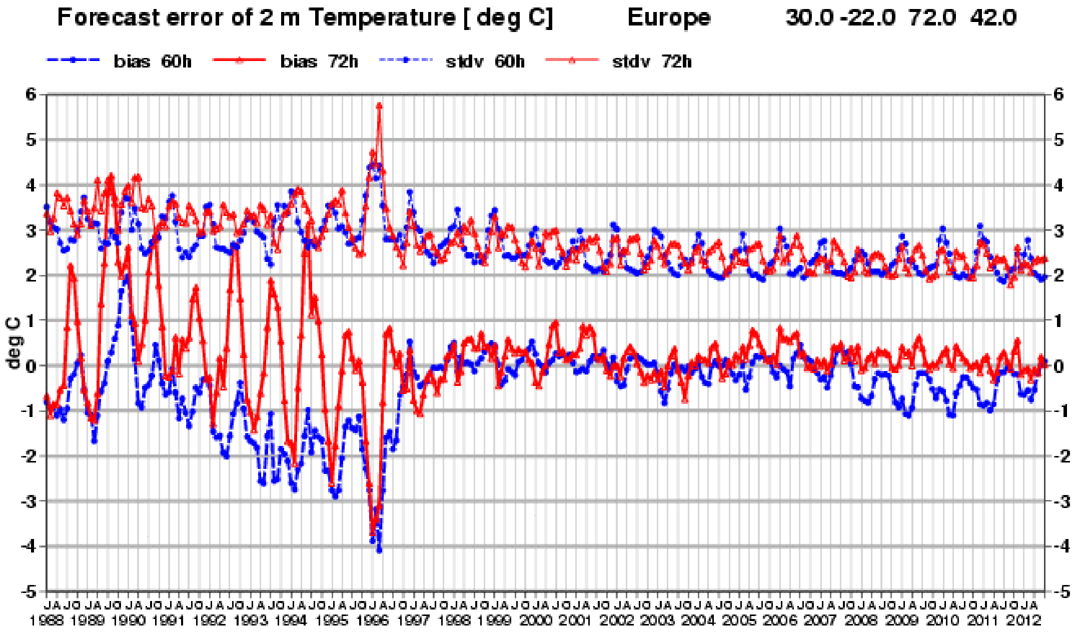


Figure 19: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

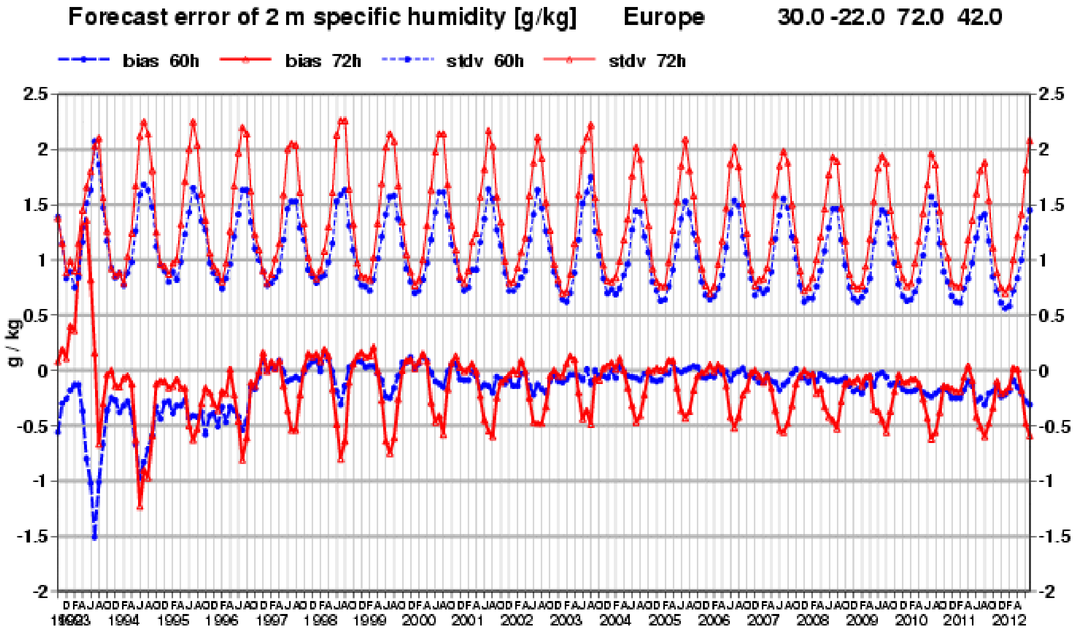


Figure 20: Verification of 2 m specific humidity forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

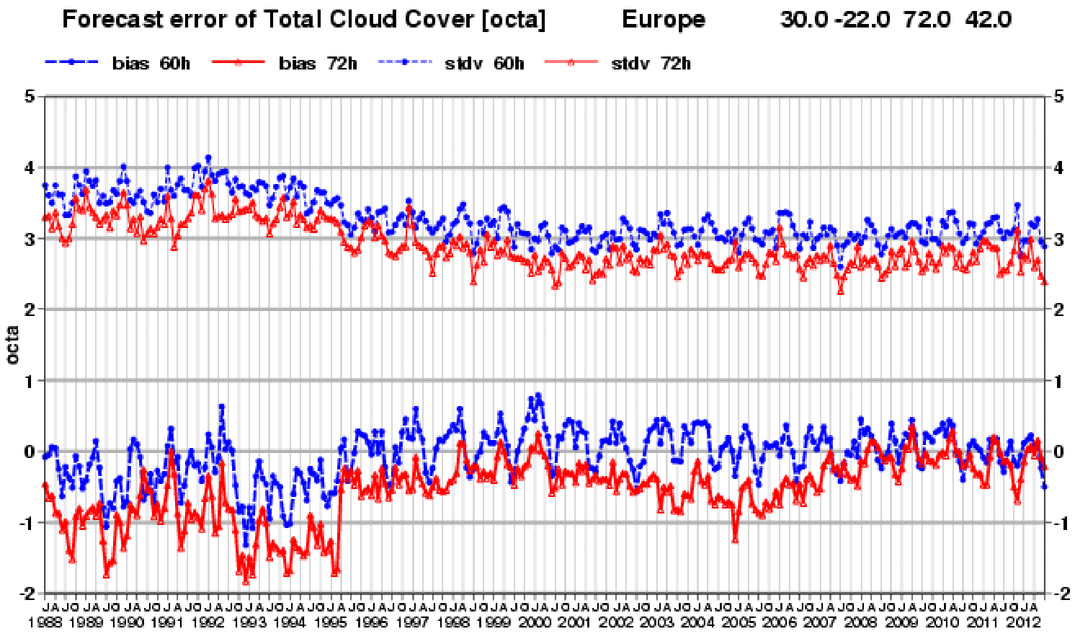


Figure 21: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

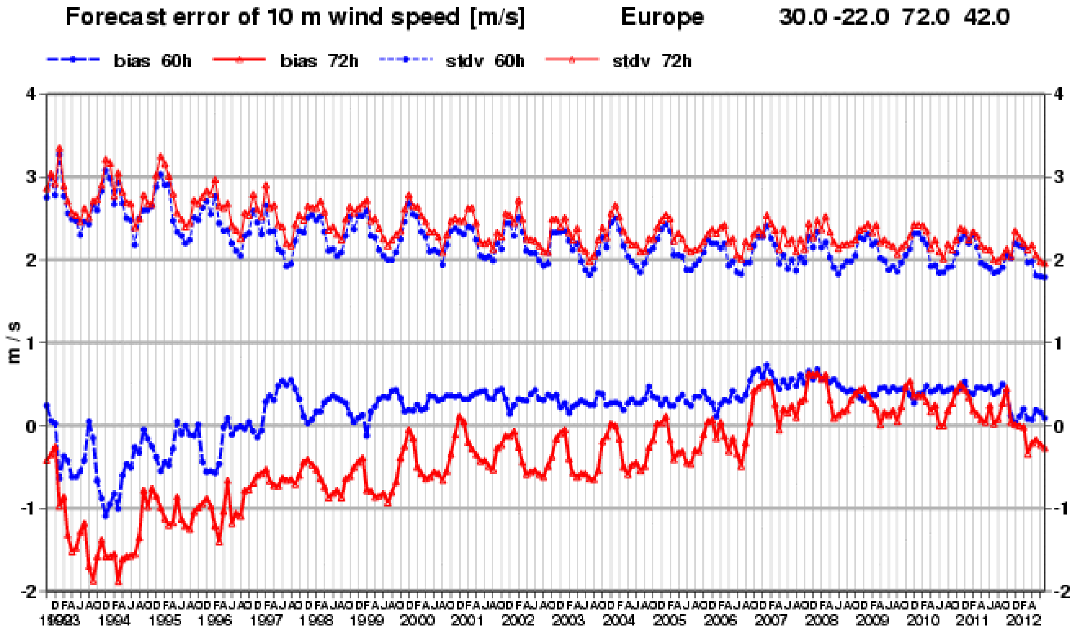


Figure 22: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

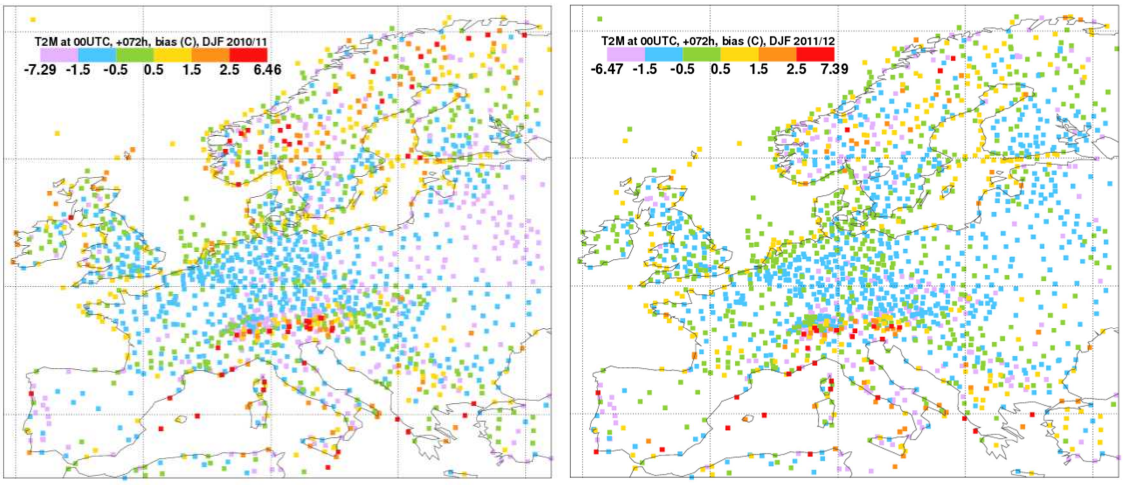

T2M at 00UTC, +072h, bias (C), DJF 2010/11
-7.29 -1.5 -0.5 0.5 1.5 2.5 6.46
T2M at 00UTC, +072h, bias (C), DJF 2011/12
-6.47 -1.5 -0.5 0.5 1.5 2.5 7.39

Figure 23: Night-time 2 m temperature mean errors during winters (December–February) 2010–11 and 2011–12.

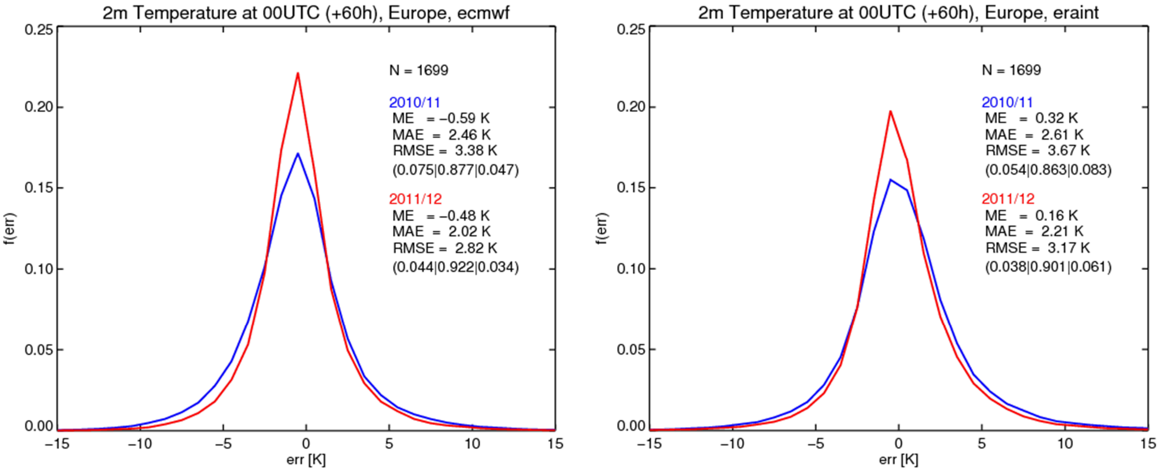


Figure 24: Error distributions for Europe, comparison of winters 2010–11 and 2011–12 for the operational run (left) and ERA-Interim (right). N gives the average number of stations available during the period. Also shown are mean error (ME), mean absolute error (MAE) and root mean squared error (RMSE) for the two winters. Fractions of cases with errors <-5C, between -5 and +5C, and >+5C given in parentheses.

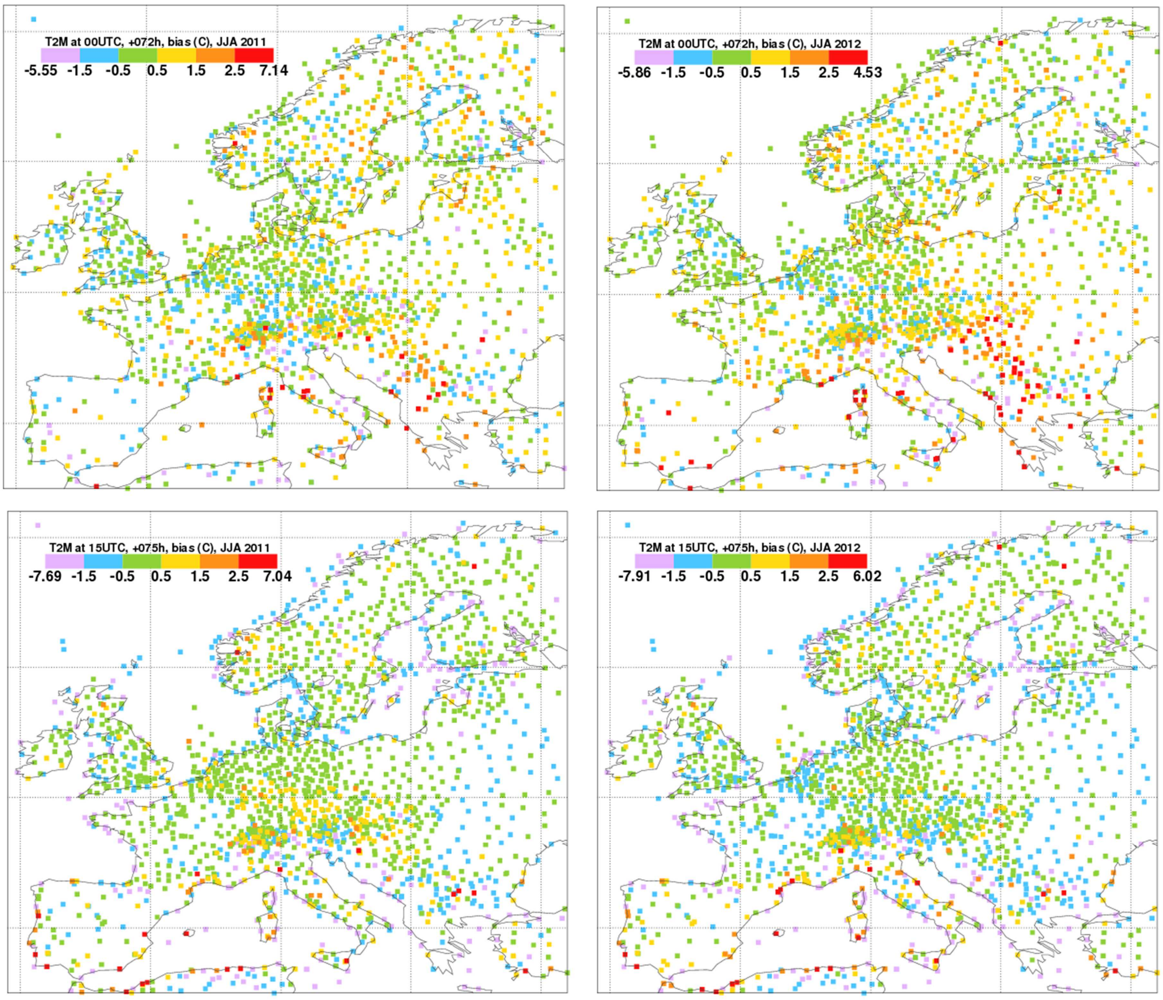

T2M at 00UTC, +072h, bias (C), JJA 2011
-5.55 -1.5 -0.5 0.5 1.5 2.5 7.14
T2M at 00UTC, +072h, bias (C), JJA 2012
-5.86 -1.5 -0.5 0.5 1.5 2.5 4.53
T2M at 15UTC, +075h, bias (C), JJA 2011
-7.69 -1.5 -0.5 0.5 1.5 2.5 7.04
T2M at 15UTC, +075h, bias (C), JJA 2012
-7.91 -1.5 -0.5 0.5 1.5 2.5 6.02

Figure 25: Night-time (top) and day-time (bottom) 2 m temperature mean errors during summers (June–August) 2011 and 2012.

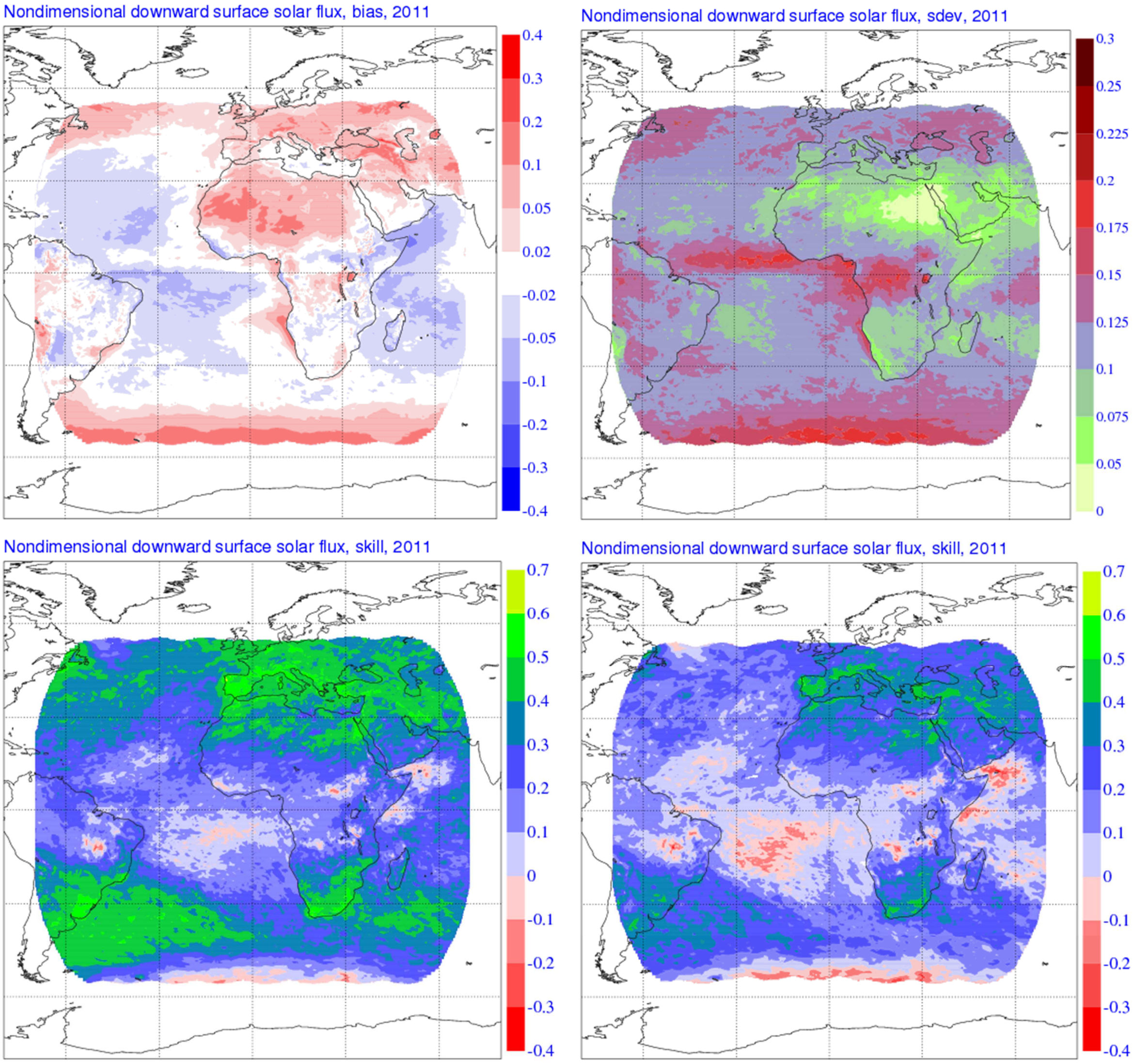

Nondimensional downward surface solar flux, bias, 2011
0.4
0.3
0.2
0.1
0.05
0.02
-0.02
-0.05
-0.1
-0.2
-0.3
-0.4
Nondimensional downward surface solar flux, sdev, 2011
0.3
0.25
0.225
0.2
0.175
0.15
0.125
0.1
0.075
0.05
0
Nondimensional downward surface solar flux, skill, 2011
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
Nondimensional downward surface solar flux, skill, 2011
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4

Figure 26: Top panels: mean error (left) and standard deviation (right) of the error for the high-resolution operational forecasts of daily means of non-dimensional downward solar radiation at the surface at forecast day 1. Bottom panels: forecast skill (compared to climatology) of non-dimensional downward solar radiation at the surface at forecast day 1 (left) and 3 (right).

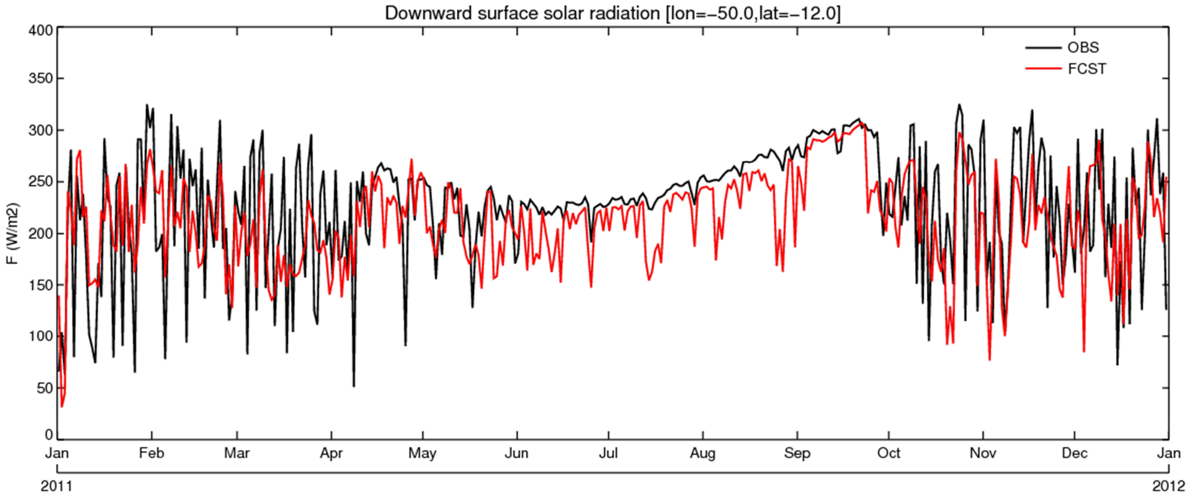
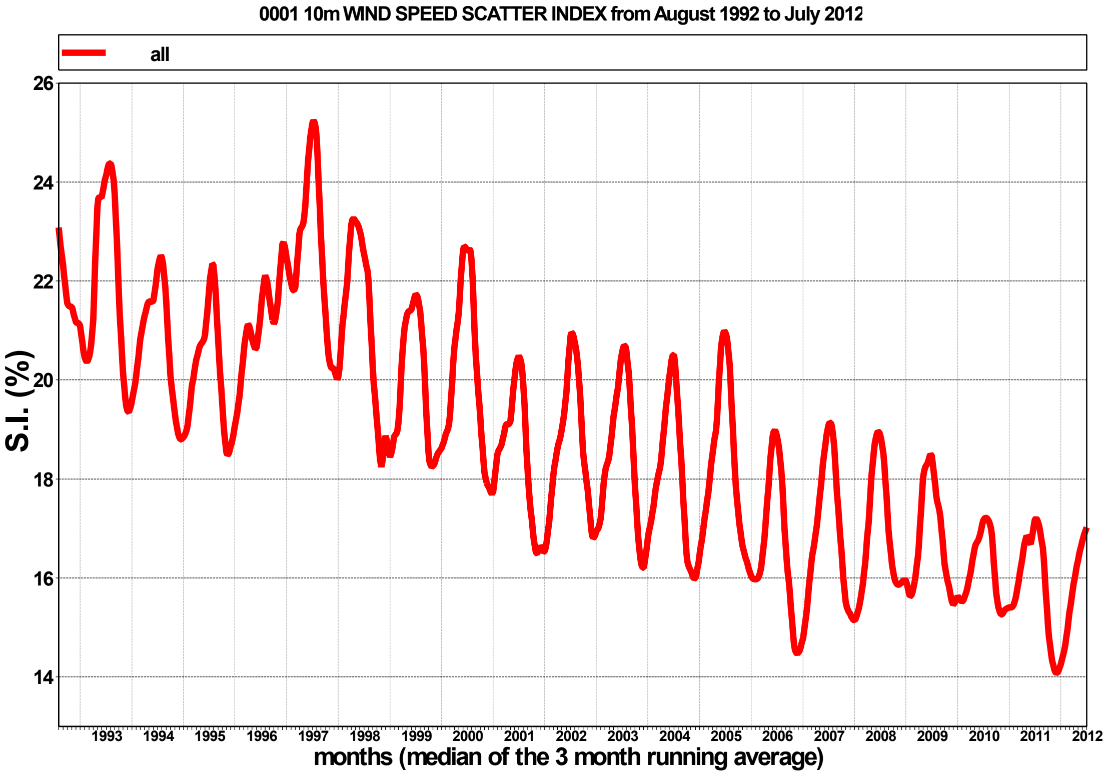


Figure 27: Observed and forecast diurnal totals of downward surface solar radiation values in the year 2011 at a location in Brazil.



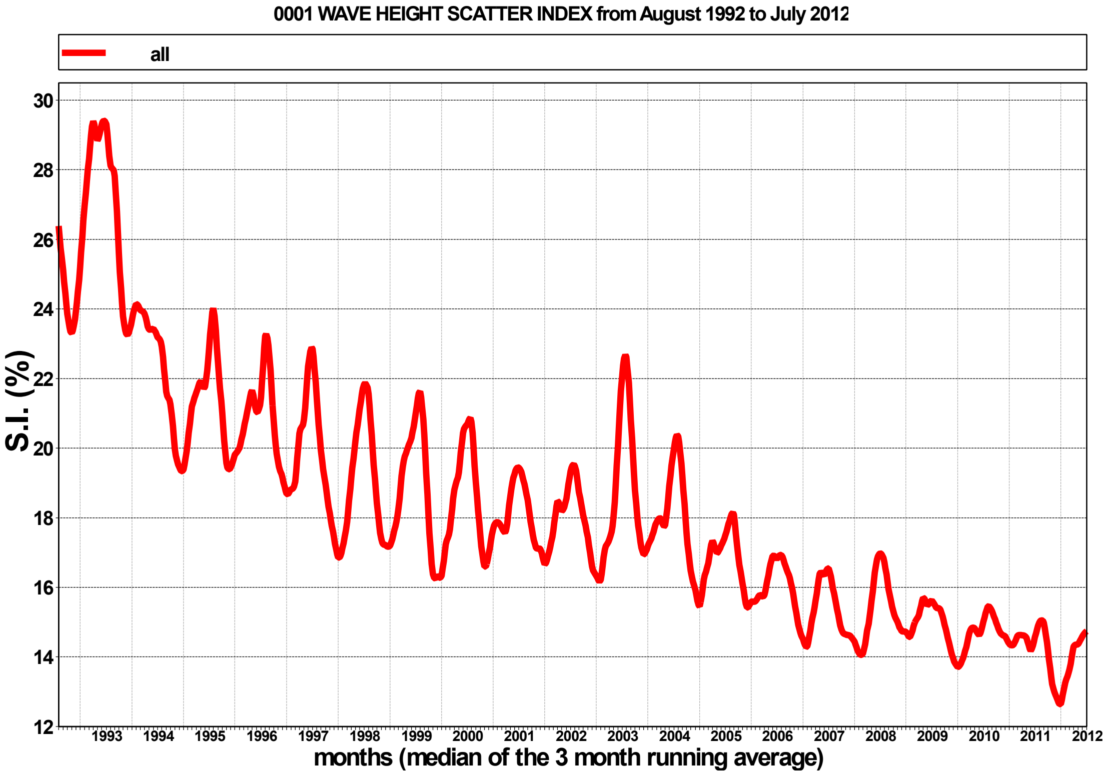


Figure 28: Time series of verification of the ECMWF 10 m wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

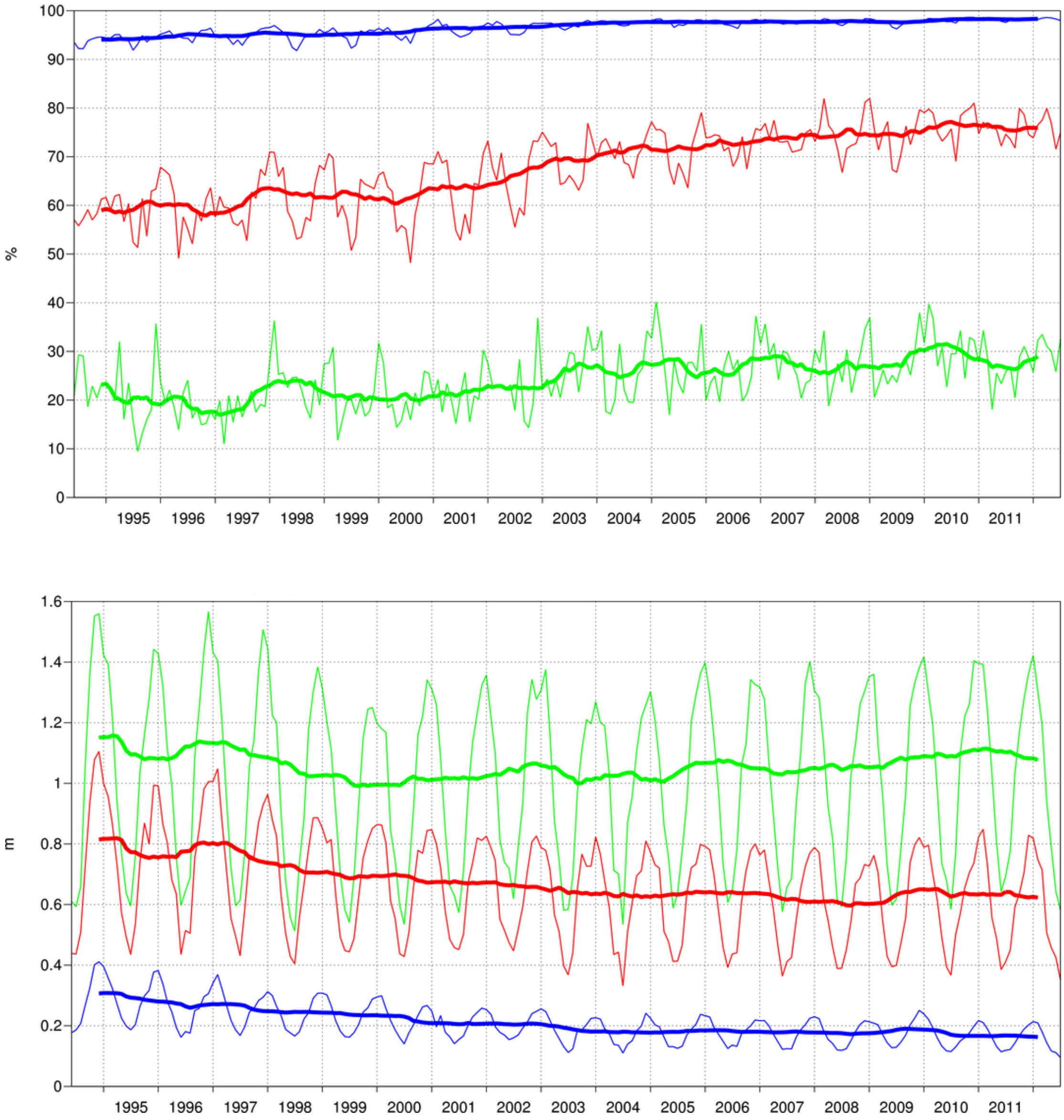


Figure 29: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC (top) and error standard deviation (bottom) for ocean wave heights verified against analysis for the northern extratropics at day 1 (blue), 5 (red) and 10 (green).

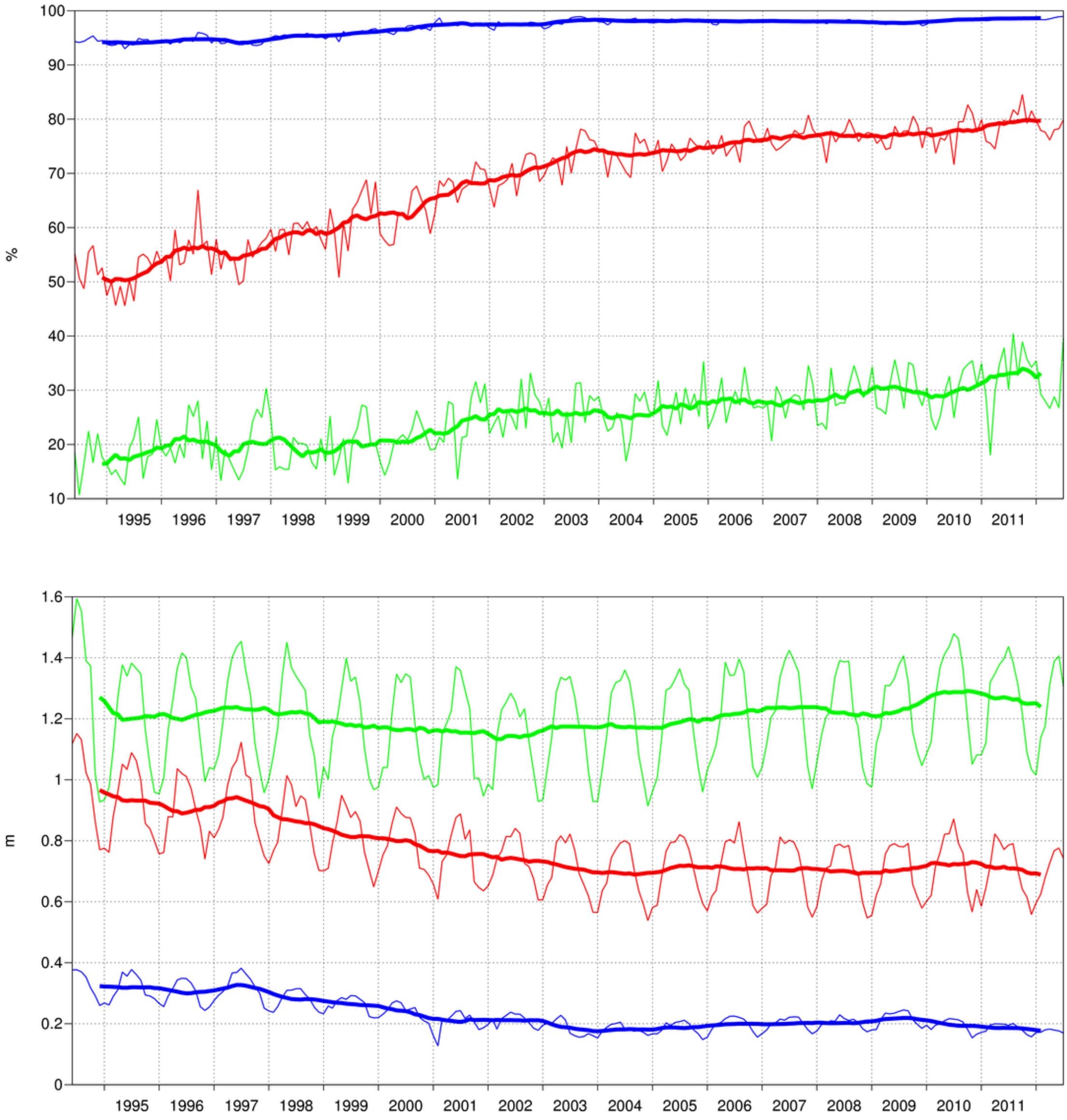


Figure 30: As Figure 29 for the southern hemisphere.

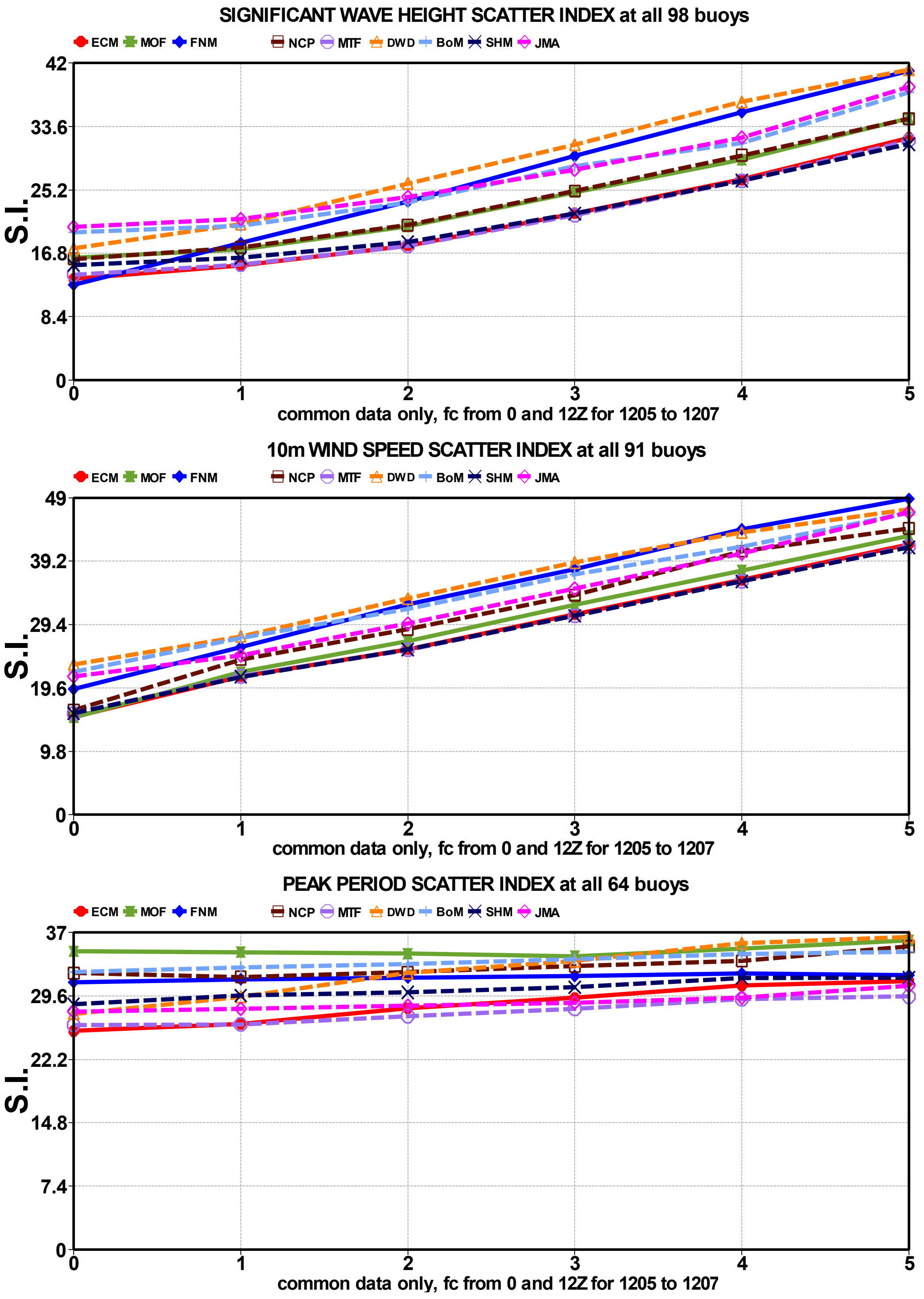


Figure 31: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the three-month period May–July 2011. The x-axis shows the forecast range in days from analysis (step 0) to day 5. MOF: the Met Office, UK; FNM: Fleet Numerical Meteorology and Oceanography Centre, USA; NCP: National Centers for Environmental Prediction, USA; MTF: Météo France; DWD: Deutscher Wetterdienst, BoM: Bureau of Meteorology, Australia; SHM: Service Hydrographique et Océanographique de la Marine, France JMA: Japan Meteorological Agency.

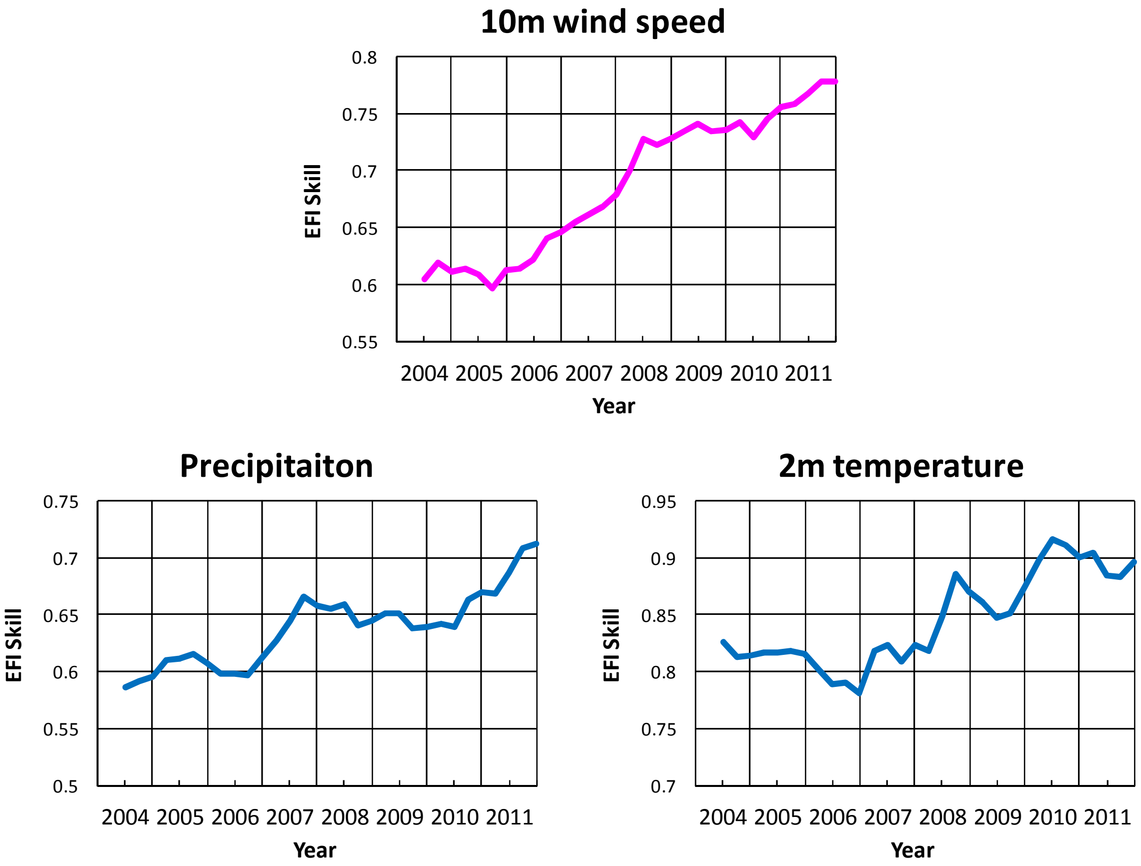


Figure 32: Verification of Extreme Forecast Index (EFI). Top panel: supplementary headline score - skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate, curves show a four-season running mean of relative operating characteristic (ROC) area skill scores (final point includes spring (March–May) 2012).Bottom panels show the equivalent ROC area skill scores for the precipitation (left) and 2 m temperature (right) EFI forecasts.

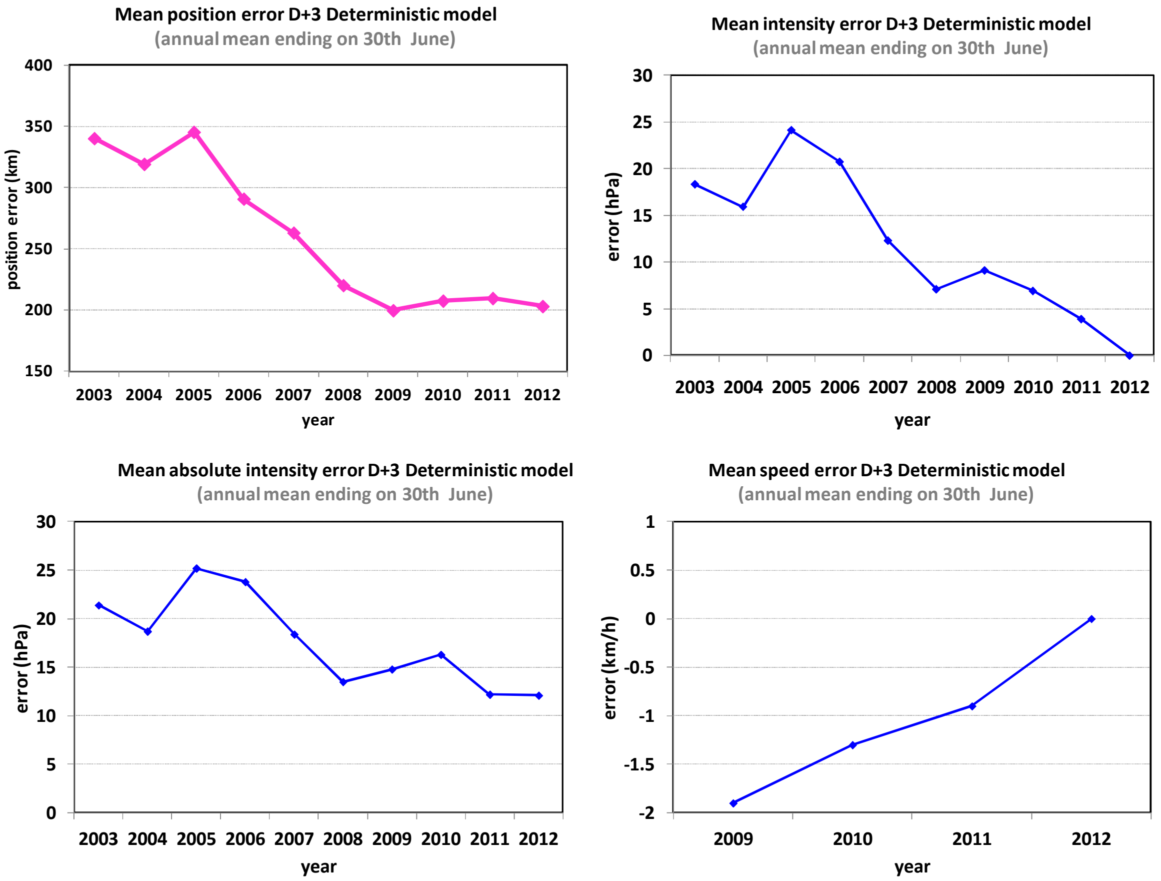


Figure 33: Verification of tropical cyclone predictions from the operational high-resolution forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported in real time via the GTS. Top left: supplementary headline score - the mean position error (km) of the three-day high-resolution forecast. Top right: mean error (bias) in the cyclone intensity (difference between forecast and reported central pressure; positive error indicates the forecast pressure is less deep than observed. Bottom left: mean absolute error of the intensity. Bottom right: mean speed error; negative values indicate the forecast is too slow compared to the observed cyclones.

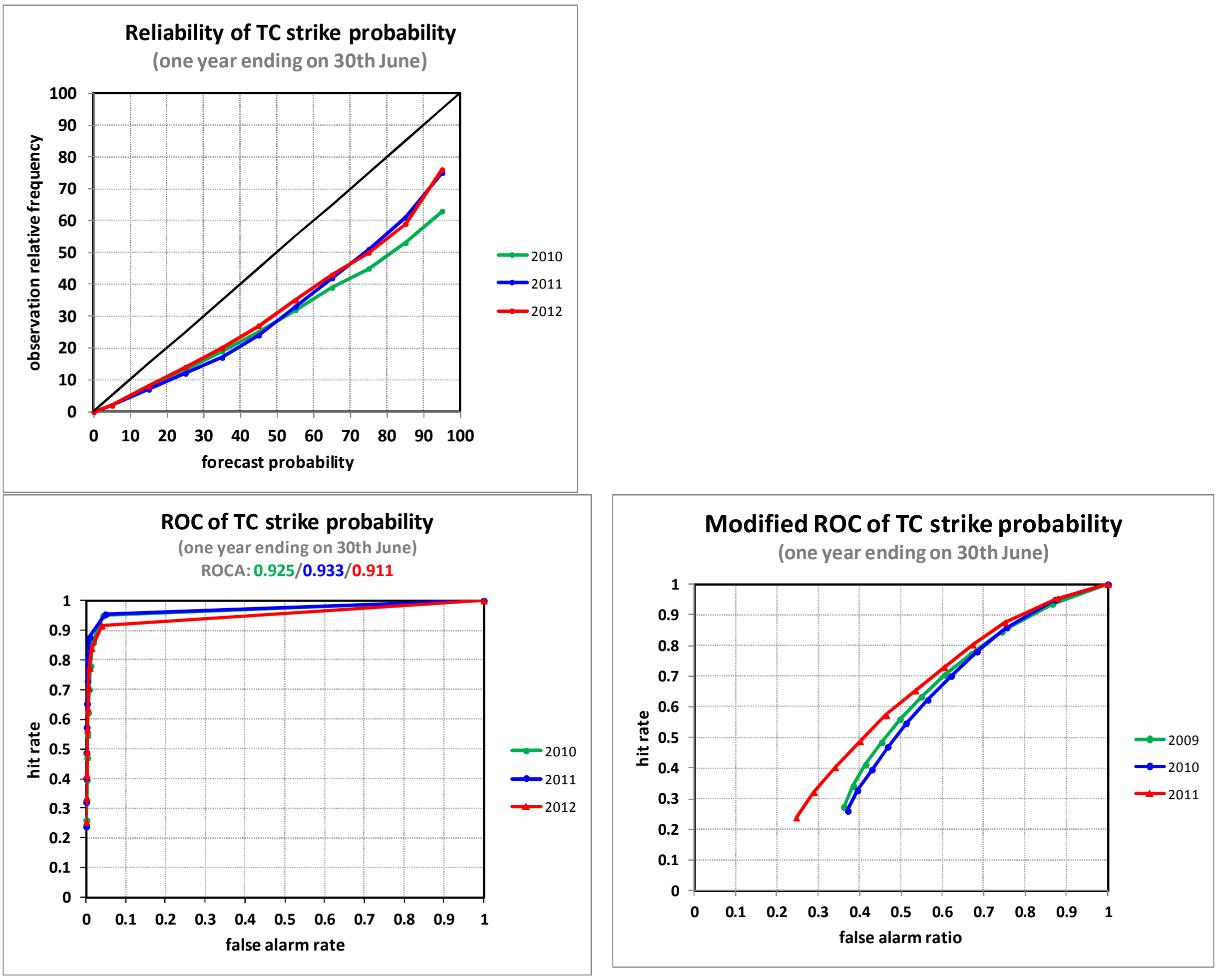
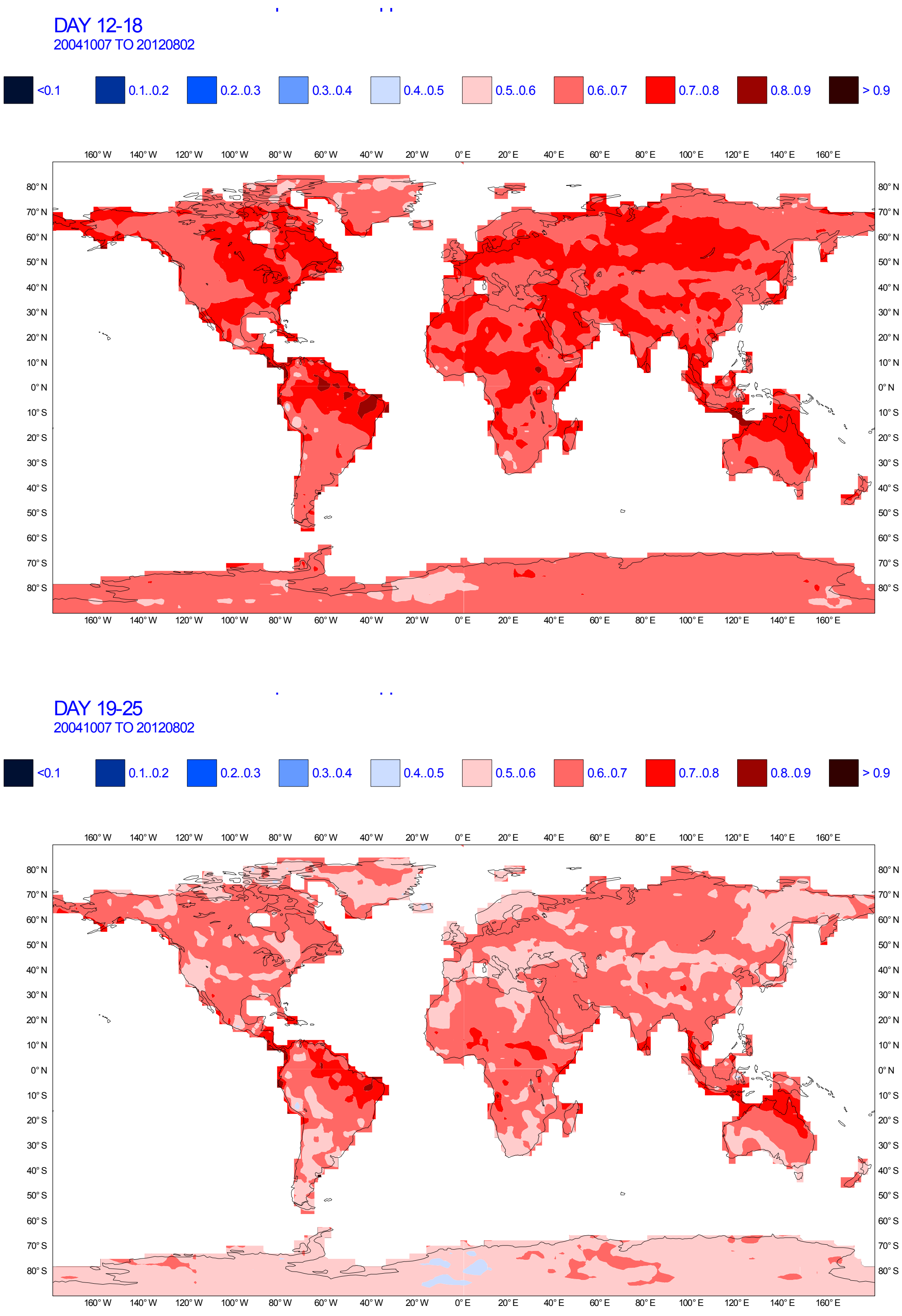


Figure 34: Probabilistic verification of ensemble tropical cyclone forecasts for three 12-month periods: July 2009–June 2010 (green), July 2010–June 2011 (blue) and July 2011–June 2012 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) 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).



DAY 12-18

20041007 TO 20120802

DAY 19-25

20041007 TO 20120802

Figure 35: 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 2 August 2012 for two seven-day forecast ranges: days 12–18 (top) and days 19–25 (bottom). Stronger red shading indicates higher skill compared to climate.

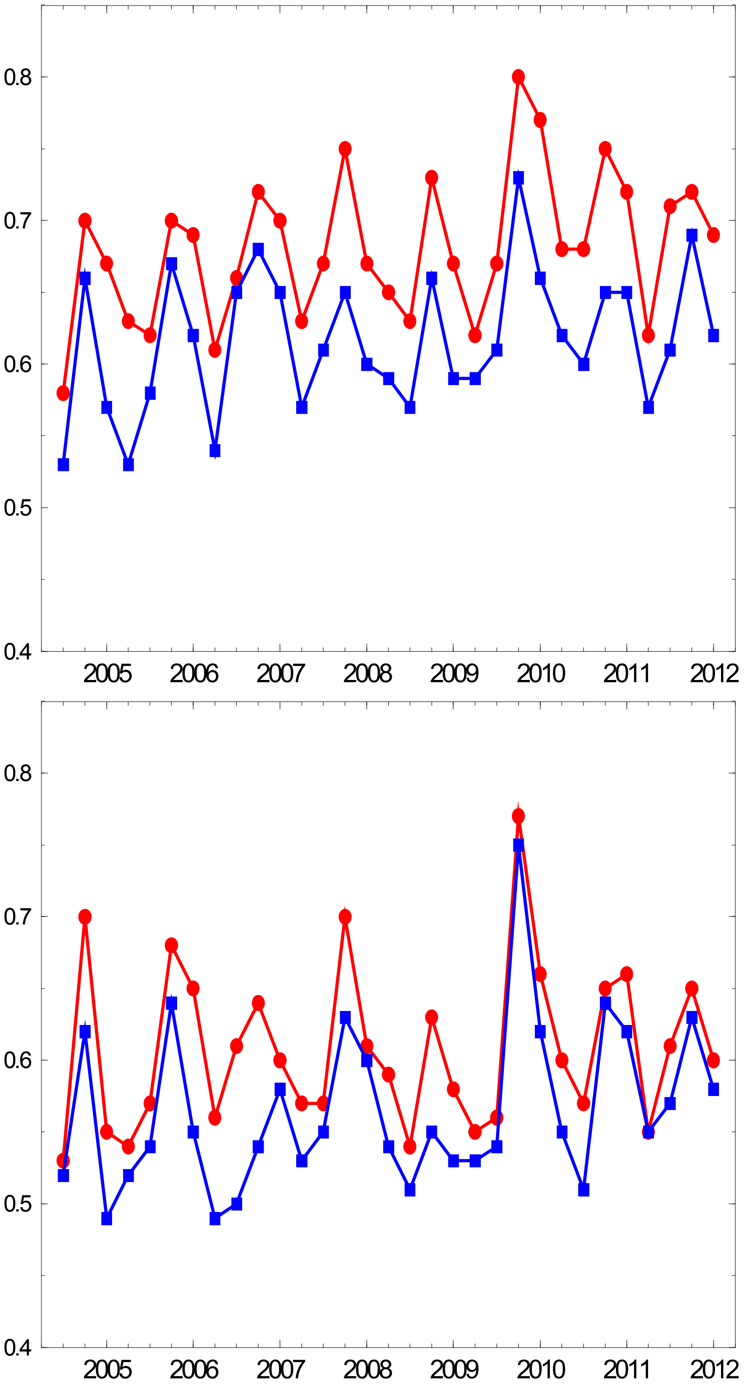
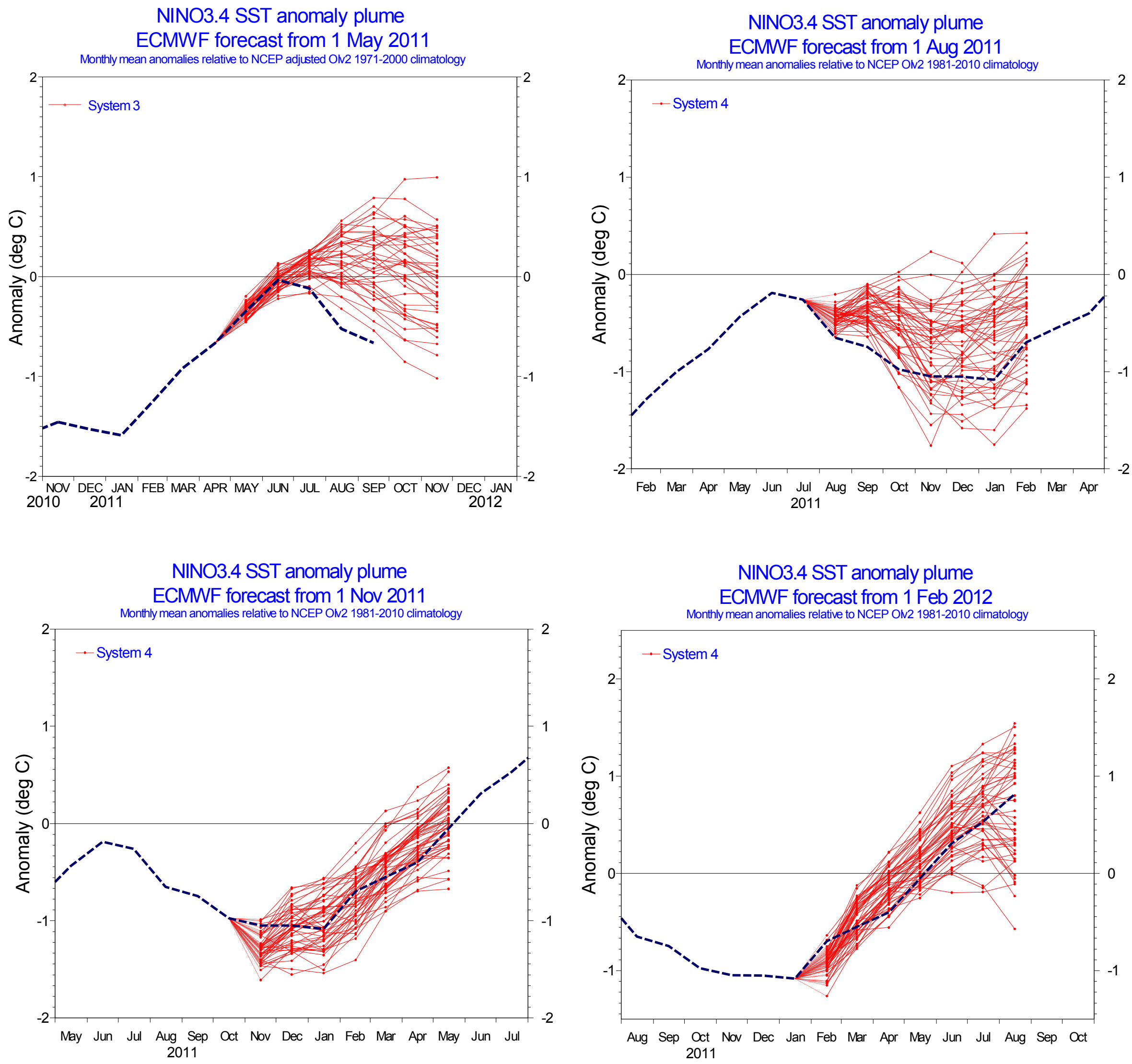


Figure 36: Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution. Scores are calculated for each three-month season since autumn (September–November) 2004 for all land points in the extratropical northern hemisphere. The red line shows the 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 2012.



NINO3.4 SST anomaly plume
ECMWF forecast from 1 May 2011
Monthly mean anomalies relative to NCEP adjusted OIv2 1971-2000 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Aug 2011
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Nov 2011
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Feb 2012
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

Figure 37: ECMWF seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from May 2011 (top left), August 2011 (top right), November 2011 (bottom left) and February 2012 (bottom right). The red lines represent the ensemble members; dashed blue lines show the subsequent verification. Note that the forecast from May 2011 was made using the System 3 of the seasonal component of the IFS, while the other three forecasts are from System 4.

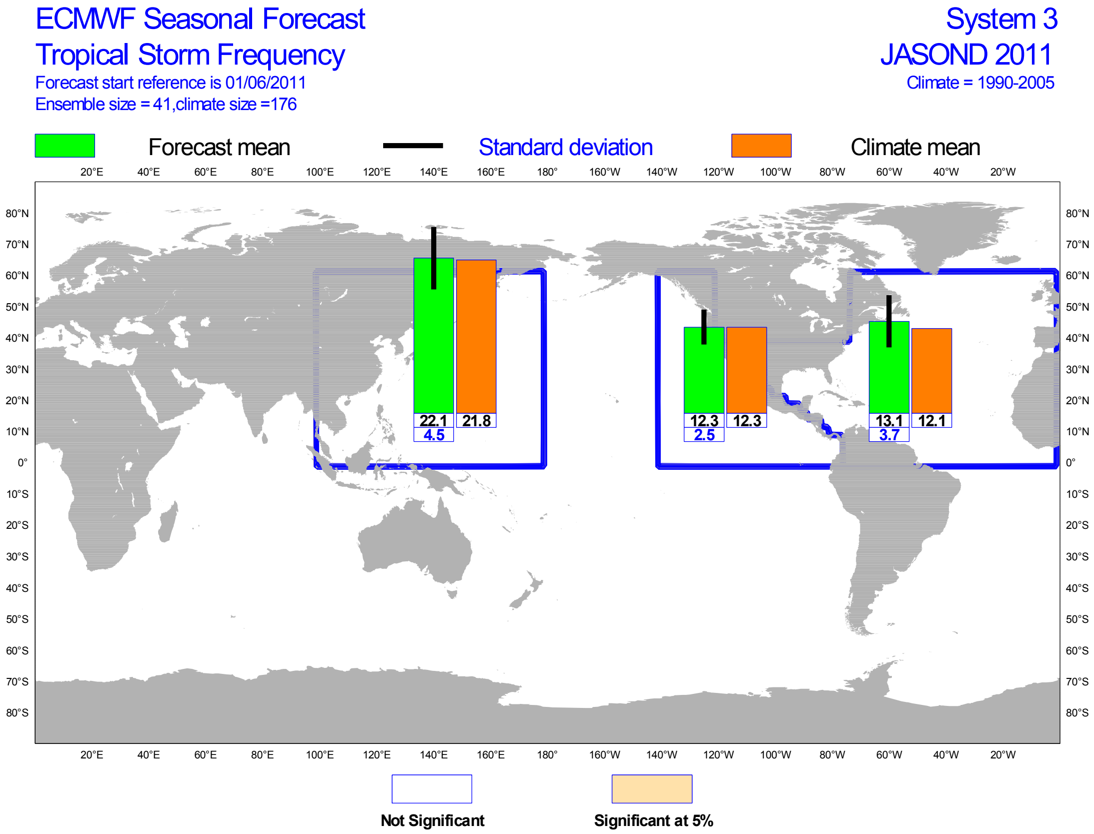


Figure 38: Tropical storm frequency forecast issued in June 2011 for the six-month period July–December 2011. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The black bars represent ±1 standard deviation within the ensemble distribution; these values are indicated by the blue number. The 41-member ensemble forecast is compared with the climatology. A Wilcoxon-Mann-Whitney (WMW) test is then applied to evaluate if the predicted tropical storm frequencies are significantly different from the climatology. The ocean basins where the WMW test detects significance larger than 90% have a shaded background

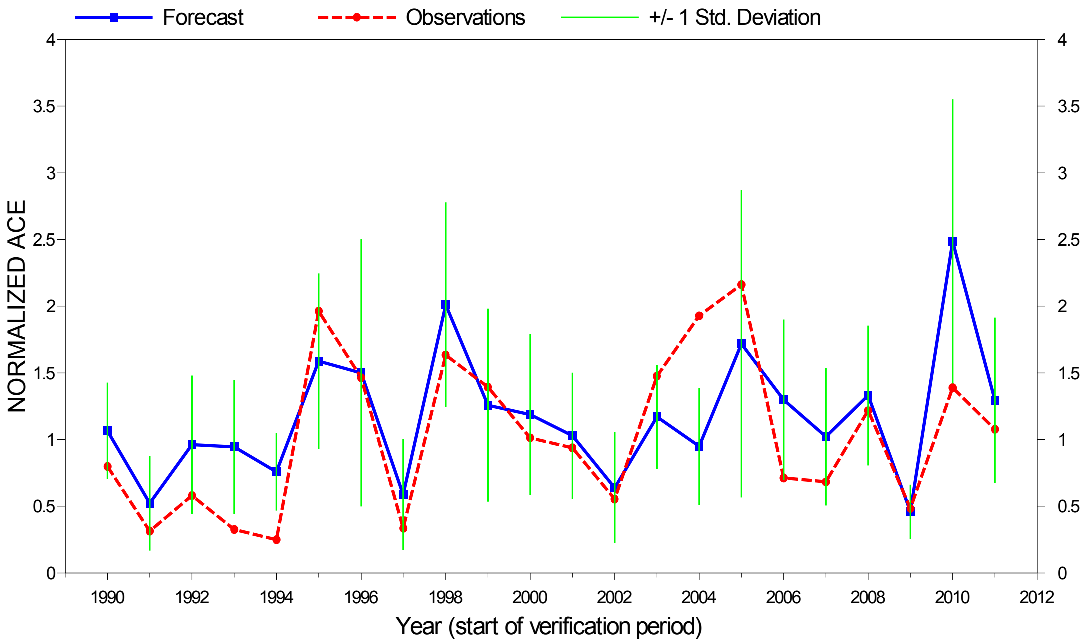


Figure 39: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2011. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows the observation. Forecasts 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.

Annex: A short note on scores used in this report

A. 1 Deterministic upper-air forecasts

The verifications used follow WMO CBS recommendations as closely as possible. Scores are computed from forecasts on a standard 1.5 x 1.5 grid (computed from spectral fields with T120 truncation) limited to standard domains (bounding co-ordinates are reproduced in the figure inner captions), as this is the resolution agreed in the updated WMO CBS recommendations approved by the 16th WMO Congress in 2011. When other centres' scores are produced, they have been provided as part of the WMO CBS exchange of scores among GDPS centres, unless stated otherwise - e.g. when verification scores are computed using radiosonde data (Figure 13), 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 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 2) 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 1 and Figure 4 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 29, Figure 30) 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

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

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

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

where $CRPS_{clim}$ 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 7) and its inter-annual variability (Figure 9).

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 34). Figure 34 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 36.

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

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 16, Figure 18). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 19 to Figure 22), verification data are European 6-hourly SYNOP data (area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the four closest grid points, provided the difference between the model and true orography is less than 500 m. A crude quality control is applied to SYNOP data (maximum departure from the model forecast has to be less than 25 K, 20 g/kg or 15 m/s for temperature, specific humidity and wind speed respectively). 2 m temperatures are corrected for 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).

References

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

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