TECHNICAL MEMORANDUM

742

Evaluation of ECMWF forecasts, including 2013-2014 upgrades

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

Research and Forecast Department

December 2014

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/en/research/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's forecast performance with that of other global forecasting centres. Section 4 presents 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, 710). A short technical note describing the scores used in this report is given in the annex to this document.

Verification pages have mostly been moved to the new ECMWF web and are regularly updated. They are accessible at the following address: www.ecmwf.int/en/forecasts/charts

by choosing ‘Verification’ under the header ‘Medium Range’

(medium-range and ocean waves)

by choosing ‘Verification’ under the header ‘Extended Range’

(monthly)

by choosing ‘Verification’ and ‘Seasonal forecasts‘ under the header ‘Long Range’

(seasonal)

2. Changes to the ECMWF forecasting system

With model cycle 38r2, the vertical resolution upgrade for the high-resolution forecast model (HRES), the ensemble of data assimilations (EDA), the main assimilation (4DVAR) and the boundary conditions (BC) optional programme had been implemented on 25 June 2013. The overall impact is already slightly positive and it is expected that the full potential of a fundamentally better resolution of physical processes will be fully exploited in the near future.

The subsequent cycle, 40r1, has been implemented on 19 November 2013 and comprised the combination of changes for the ensemble forecasts (ENS), namely the tendency coupling of ocean with atmosphere from initial time using a new NEMO version and the inclusion of wave effects on ocean circulation, an increased vertical resolution, and the addition of land surface parameter perturbations to the ENS initial conditions. The cycle also includes major model changes: The convection scheme has been revised to address the longstanding problem of updraft initiation too early in the day thus improving the diurnal cycle of precipitation. Secondly, the modification of vertical diffusion and orographic drag alleviates systematic wind shear and wind turning errors in the boundary layer. These model changes also prove to be very beneficial over the entire troposphere in terms of large-scale predictive skill, most pronounced over the northern hemisphere.

Now, ENS uses the previous vertical discretisation of the HRES system with 91 levels (L91). The ENS had used 62 levels since February 2006 (cycle 30r1). The change from L62 to L91 raises the model top from 5 hPa to 0.01hPa and more than doubles the number of levels between 100 hPa and 5 hPa.

The decision for implementing L91 was based on extensive ENS experimentation comparing vertical discretisation L62 with L91 and other configurations over 61 cases in two periods. Results showed a considerable positive impact on ENS skill measures in the stratosphere from raising the model top (more than 1-day gain in skill at 50 hPa zonal wind in the northern extra-tropics at day-7). Furthermore, raising the model top also led to a statistically significant positive impact in the troposphere (2-hour gain in skill at day-7 for 500 hPa geopotential in the northern extra-tropics).

The EDA has been further enhanced by increasing the number of ensemble members from 10 to 25, thus improving statistical representativeness of derived background error statistics. In addition, full covariance statistics are provided to 4DVAR. Instead of using a climatological background error covariance model, an online estimate of the background error covariances is updated at each assimilation cycle from the available EDA ensemble perturbations over the most recent 12 days. This introduces another degree of flow-dependency in the horizontal structure functions and the vertical correlations which results in an average error reduction in geopotential of 1-2%.

In the past, several components of the snow analysis have been improved, for example, the addition of regional station observational network data and the introduction of the optimum interpolation technique that produces the analysis. With cycle 40r1, the weight given to observations has been reduced to avoid cases where snow cover data obtained from satellite observations was found to be out of date and could eliminate freshly fallen snow in the analysis. A related change is the revision of the glacier mask over Iceland accounting for the observed reduction of glacier extent due to global change.

Cycle 40r1 also includes a number of changes to satellite data usage. More temperature and humidity observations over land and sea-ice from AMSU-A/B and MHS as well as cloud affected radiances over ocean will be used, the satellite radiance quality control will include model background error estimates derived from the EDA, and the estimation of observation errors and quality control for Atmospheric Motion Vectors (AMV) has been fundamentally improved.

Situation-dependent observation errors also facilitate the harmonisation and simplification of the quality control applied to AMVs, where observations are compared to the model. This has previously been very strict in the ECMWF system. The situation dependent observation errors lead to a more appropriate weighting of the observations in the analysis, allowing a less stringent data selection and thus an increased use of AMVs.

Cycle 40r1 produces a substantial improvement of temperature, geopotential and wind predictive skill for HRES over the northern hemisphere in the troposphere and low-mid stratosphere, and, to a lesser degree also over Europe. ENS verification shows a slight reduction of both spread and error, except for wind. Continuous Ranked Probability Skill Scores (CRPSS) improve in general, and these improvements are statistically significant for the first days.

In the southern hemisphere, a small degradation of low-level temperatures is found for HRES that is only apparent when forecasts are compared to analyses rather than observations. In the tropics, the elimination of a term defining unresolved wind shear leads to increased lower level wind errors in both HRES and ENS. Again, this impact is not seen when verified with observations. Lastly, the convection change produces a small increase of upper tropospheric temperature bias. The two latter features will be addressed by the forthcoming model cycle.

In long coupled model integration the negative SST bias in the tropical Pacific (maximum September-November) has been improved, mostly due to the model changes themselves. Improvements of SSTs in the northern Pacific can be associated with changes applied to the ocean model.

The overall performance of this cycle is shown in Figure 1 in terms of anomaly correlations and root-mean-square errors when verified with analysis and relative to the previous cycle 38r2. Cycle 40r1 has been migrated to the new Cray XC-30.

3. Verification for free atmosphere medium-range forecasts

3.1. ECMWF scores

3.1.1. Extratropics

Figure 2 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 both hemispheres and over Europe scores have been consistently high. Resulting 12-month means are close to or, as in the northern hemisphere, slightly exceeding the highest previous values. The effect 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 recent increase of anomaly correlation in the northern extratropics was not due to atmospheric variability but represents an actual increase in skill. In Europe, as well as in the southern hemisphere, the skill relative to ERA-Interim has stabilized on the high level reached in 2012.

Atmospheric variability affects different skill measures in different ways, as discussed in ‘An evaluation of recent performance of ECMWF’s forecasts’ in ECMWF Newsletter No. 137 (Autumn 2013). As a complementary measure of performance, Figure 3 shows the evolution of skill 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. In the northern hemisphere a continuing upward trend can be observed, especially at shorter lead times. The apparent decrease in skill in the later forecast steps over Europe is partly due to atmospheric variability, as can be seen from Figure 4 which shows the RMS errors for Europe of the six-day forecast and the persistence forecast (the reference system for Figure 3). The error of the six-day forecast has been consistently low over the last four years, however the level of activity as measured by the error of the persistence forecast, used here as a reference, has been at its lowest in 30 years. .

Figure 5 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. In terms of the error distribution, 2014 was very good. In winter, the number of large errors (ACC<50%) was smaller than in any previous winter. The results for summer show that forecasts in 2014 were quite skilful as well but did not reach the very high values seen in 2007 and 2012. However, the number of very large errors (ACC<20%) was smaller than in all previous summers since 1998.

Figure 6 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. Overall, the level of consistency between consecutive forecasts has slightly increased further in the last year. Values for summer 2014 are very similar to the exceptionally low values in summer 2012.

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 7. Five-day temperature scores have improved compared to last year. Scores for one-day forecasts of temperature as well as forecasts of vector wind have been stable. For temperatures (upper panel), a recent improvement from changes to ice cloud model physics is observed, reflecting the sensitivity of the upper troposphere - lower stratosphere (UTLS) performance to clouds and their impact on heating. For winds (lower panel) scores have been fairly constant over time but recent improvements are observed here as well.

While keeping an emphasis on tropospheric performance, model biases across the UTLS range have now started to be addressed. This has been partly accomplished by the vertical resolution upgrade and a modification of non-orographic gravity wave drag and vertical diffusion, a revised ensemble of data assimilations (EDA) sampling and filtering, and improved numerics in the semi-Lagrangian scheme.

The trend in ensemble performance is illustrated in Figure 8, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. As for the 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 the skill in winter 2013-14 in the northern extratropics has been very similar 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 and representation of model uncertainties, the increase in resolution inJanuary

2010, and further redefinition of perturbations using the ensemble of data assimilations. 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 and summer, as well as the difference between ensemble spread and ensemble-mean error for the last three winters and summers, are shown in Figure 9 and Figure 10. The match between the spread and error in 2014 is very similar to previous years. For 500 hPa height the ensemble spread is now very close to the error up to day 6 while at larger lead times there is some over-dispersion in winter and under-dispersion in summer. The under-dispersion for temperature at 850 hPa in both seasons is still present, 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 2013-14 has been the highest so far for 500 hPa height up to forecast day 7, and for 850 hPa temperature up to day 12. It is only exceeded at the longer ranges by the exceptional winter 2009–10, when anomalous flow made some contribution to the high scores.

In order to have a benchmark for the ENS, the CRPS has been computed for a ‘dressed’ HRES. This also helps to distinguish the effects of ensemble configuration developments from pure model developments. The dressing uses the mean error and standard deviation of the previous 30 days to generate a Gaussian distribution around the HRES. Figure 12 shows the evolution of the CRPS for the ENS and for the dressed HRES over the last 10 years for temperature at 850 hPa at forecast day 5. In the northern hemisphere the skill of the ENS relative to the reference forecast was about 6 % in 2005 and has reached a value of 14% after the change of resolution associated with the introduction of model cycle 36r1 (Jan 2010). For the southern hemisphere the corresponding values are 11% and 16%. Both the dressed HRES and ENS have further improved afterwards, however their relative difference has remained nearly constant. Figure 13 shows the skill of the ENS relative to the dressed HRES for different lead times. It can be seen that the relative benefit of the ENS strongly increases with lead time. For forecast day 1, the ENS has reached a level of skill similar to the dressed HRES only after the 2010 resolution upgrade. It is worth noting, however, that using the forecast error for dressing of the HRES is equivalent to generating a perfectly calibrated ensemble. Thus this sort of reference forecast represents a rather challenging benchmark. The recent drop in relative skill which is most visible for day 1 appears for the most part to be due to improvements in the HRES associated with changes introduced in June and November 2013 (model cycles 38r2 and 40r1).

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 14. At 200 hPa (upper panel) the 1-day forecast has continued to improve (although it is still slightly higher than the minimum which was reached in 2003–2004), and the 5-day forecast has been similar to the previous two years. At 850 hPa (lower panel) the error at day 1 has been slightly reduced while at day 5 it has slightly increased. Relative to ERA-Interim the forecast skill has further increased.

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 onwards). 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 15 (northern hemisphere extratropics) and Figure 16 (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.

WMO-exchanged scores also include verification against radiosondes over regions such as Europe. Figure 17, (Europe) and Figure 18 (northern hemisphere extratropics) 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 19 (verification against analyses) and Figure 20 (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 19) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 20), 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 21. The upper panel shows the lead time at which the stable equitable error in probability space (SEEPS) skill for the high-resolution forecast for precipitation accumulated over 24 hours over the extratropics drops below 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 at which the CRPSS for the probability forecast of precipitation accumulated over 24 hours over the extratropics drops below 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.

Much of the recent variation of the score for the high-resolution forecast is due to atmospheric variability, as shown by comparison with the ERA-Interim reference forecast (dashed line in Figure 21, upper panel). By taking the difference between the operational and ERA-Interim scores most of this variability is removed, and the effect of model upgrades is seen more clearly (upper panel in Figure 22). While the largest improvement is associated with the introduction of the five-species microphysics in November 2010 (cycle 36r4), microphysics changes in subsequent cycles led to a further increase in skill. The probabilistic score (lower panel in Figure 21) shows some recent improvement after the stagnant period 2010-2012 which was partly 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 (lower panel in Figure 22), which has masked improvements due to model upgrades during that time. In 2012, however, this trend has reversed, so that the increase in skill has become more visible again 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 23). The comparatively low skill 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).

There is an overestimation of the frequency of light precipitation in most global and regional models. In the ECMWF model this leads to frequency biases on the order of 1.2-1.4 for precipitation amounts >1 mm/24h in the extra-tropics, only a limited part of which can be explained by the representativeness mismatch due to finite model resolution. The issue will be partially addressed by cloud physics changes introduced with cycle 40r3. However there is some indication that the changes to the deep convection scheme introduced with model cycle 40r1 (November 2013) may have had an adverse impact on the frequency of occurrence of light precipitation in some areas such as the European Alps. This will be further investigated.

Trends in mean error and standard deviation over the last 10 years of error for 2 m temperature, 2 m dewpoint, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 24 to Figure 27. 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 was a marked change in the 10 m wind speed bias (Figure 27) 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. Also, the non-systematic error of the 10 m wind speed forecast has decreased in recent years. Improvements in both bias and error standard deviation over the last two years are also apparent for total cloud cover (Figure 26).

A recurring feature of the 2 m temperature forecast is a negative night-time temperature bias over Europe in winter and early spring (Figure 24). Model changes implemented in November 2011 (cycle 37r3) led to a slight reduction of the negative night-time bias in Europe in winter of the order of 0.2–0.3 K. However, much of the problem persists, as is apparent for the winter 2013–14 which shows a negative bias very similar to the previous winter, although the geographical distribution of the bias is different (Figure 28). Comparison of error distributions in these two winters for the operational forecast and ERA-Interim (Figure 29) shows a similar increase of skill (heightening of the peak, decreases in the tails of the distribution) in both forecasts and can therefore largely be attributed to atmospheric variability.

A problem of the 2 m temperature forecast which has recently been addressed is too-rapid afternoon cooling during spring in some areas, contributing to substantial negative biases especially at higher latitudes. This is most pronounced in forested, snow-covered areas such as Scandinavia, and is related to the way 2 m temperature is computed for open areas (low vegetation tiles) within forested grid boxes. A solution to the problem has been found and will be implemented in model cycle 40r3.

The issue of the underestimation of the strength of surface-based inversions over snow under clear-sky, calm conditions was investigated. Experiments with a revised computation of 2 m temperature under such conditions did so far not give positive results.

The forecast of 2 m humidity in Europe has exhibited a dry bias during daytime in summer (Figure 25), which was related to the too strong mixing of humidity out of the surface layer. This has improved markedly with the changes to the deep convection scheme introduced with model cycle 40r1 in November 2013 (Bechtold et al., 2014).

For total cloudiness (Figure 26) the standard deviation of the forecast error is stable at the low level reached in 2012, and the negative bias was further reduced. A similar result can be seen for 10 m wind speed (Figure 27) where the standard deviation is consistently low compared to earlier years, and the night-time bias is very small. However the daytime bias, which changed sign with the retuning of roughness lengths in November 2011 (model cycle 37r3), has become slightly more negative.

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 normalized by scaling with a latitudinally and seasonally varying clear-sky flux at the surface. Figure 30 shows the mean error and standard deviation of the error at forecast day 3 of the TOA reflected radiation for the year 2013. Compared to 2012 the negative bias (underestimation of cloud cover and/or cloud reflectance) in the Southern Ocean has decreased, but it has slightly increased in the subtropical stratocumulus (Sc) regions off the western coast of Africa. The positive bias in areas dominated by trade cumulus (Cu) has decreased compared to 2012. The biases of different sign in Sc and Cu areas are expected to decrease further in the near future due to changes in cloud microphysics to be introduced with model cycle 40r3 (Ahlgrimm and Forbes, 2014). The error standard deviation has decreased compared to 2012, most notably in the Southern Ocean.

To reduce the effect of atmospheric variability on scores, the verification against CM-SAF data is also performed for ERA-Interim. There is a substantial increase in skill of the operational high-resolution forecast relative to ERA-Interim in recent years, both in the extratropics and tropics (Figure 31), 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 total cloud cover during daytime is also noticeable in the verification results against SYNOP observations (Figure 32).

ERA-Interim is useful as a reference forecast for the HRES as it allows to filter out much of the effect of atmospheric variations on scores. Figure 32 shows the evolution of skill relative to ERA-Interim for various upper-air and surface parameters. The top panel shows the relative skill at day 5, while the bottom panel shows improvements in lead time. From the top panel it can be seen that the largest relative improvements (15-20% since 2002) have been achieved for upper-air and dynamic fields, followed by 2 m temperature and 10 m wind speed. The skill of total cloud cover has stagnated for an extended period and started to increase only with more recent cycle changes. This type of plot does not take into account the fact that a given relative improvement may inherently be more difficult to achieve for some parameters than for others. To highlight this aspect, the bottom panel of Figure 32 shows improvements in terms of lead time. Using this measure, the improvement of the upper air parameters and MSLP become very similar (~0.8 forecast days since 2002), and forecast skill for total cloud cover does not lag behind as much. However, the improvement for 10 m wind speed appears disproportionally large. This is because parameters for which the forecast already has a relatively large error at initial time, such as 10 m wind speed evaluated against SYNOP, tend to exhibit a rather weak lead-time dependence of skill. Thus the 4% increase in skill for wind speed in 2011-12 translates into 1.5 forecast days. Nevertheless it highlights the fact that the changes to the roughness length introduced in Nov 2011 (model cycle 37r3) not only reduced the overall bias but led to a substantial reduction of the non-systematic error as well.

4.2. Ocean waves

The quality of the ocean wave model analysis and forecast is shown in the comparison with independent ocean buoy observations in Figure 33. The top panel of Figure 33 shows time series of the forecast error for 10 m wind speed using the wind observations from these buoys. The forecast error has steadily decreased since 1997 and it has reached its lowest value so far in the winter season 2013-14. Errors in the wave height forecast have been the lowest so far in the 1-5 day range. The long-term trend in the performance of the wave model forecasts is shown in Figure 34 and Figure 35. The general trend of increasing performance in both hemispheres has continued.

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 36 for the 12-month period June 2013 – May 2014. 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 36. ECMWF outperforms the other centres with regard to wind speed and wave height (although SHM is very close in wind speed), while Météo France has the highest skill in forecasting the peak period. 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.

A comprehensive set of wave verification charts is available on the ECMWF website at:
old.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 potentially 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 37 (top), together with the corresponding results for 24-hour total precipitation (centre) and 2 m temperature (bottom). Each curve shows a four-season running mean of ROC area skill scores from 2004 to 2013; the final point on each curve includes the spring (March–May) season 2014. For 10 m wind speed and precipitation, EFI skill has reached its highest values so far. For 2 m temperature, EFI skill is generally higher than for the other two parameters and appears to remain close to a value of 0.9.

5.2. Tropical cyclones

The 2013 North Atlantic hurricane season had a close to average number of tropical storms (13 compared to 12 in the climate mean, see also Figure 45). The tropical cyclone position error for the three-day high-resolution forecast is one of the two 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 38. 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 comparison of HRES and ENS control demonstrates the benefit of higher resolution for tropical cyclone forecasts.

The HRES position errors (top panel, Figure 38) have reached their second smallest value so far (slightly worse than in the previous year) for the three-day forecast. The same is true for the mean absolute speed errors at D+3. Typically tropical cyclones move too slowly in the forecast, however this negative bias has been relatively small in recent years. Because of the substantial year-to-year variations in the number and intensity of cyclones, there is some uncertainty in these figures. Both the mean error (bias) and mean absolute error in tropical cyclone intensity (upper central panels in Figure 38) have increased. 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 bottom panel of Figure 38 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. Whereas the forecast is underdispersive before the resolution upgrade in 2010, the spread-error relationship is very good since then. The figure also shows that the HRES position error has been generally smaller than the ensemble mean error at forecast day 3 (although similar recently), and vice versa at forecast day 5.

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 39. Results show over-confidence for the three periods, which appears to have increased from year to year. The reason for this behaviour (especially in combination with the increasing position forecast skill shown in Figure 38) is not clear and requires further study. 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. Differences between the last two consecutive years of these two measures are of different sign and not considered significant.

5.3. Additional severe-weather diagnostics

In order to extend severe-weather diagnostics at ECMWF, additional scores are being tested (see also ECMWF Newsletter No. 139). While many scores tend to degenerate to trivial values for rare events, some have been specifically designed to address this issue. Here we use the symmetric extremal dependence index, SEDI (Annex A.4), to compare the skill of the operational high-resolution and the ERA-Interim forecast. Both forecasts are verified against SYNOP observations. Figure 40 shows the time-evolution of skill in forecasting events above the 98th climate percentile in Europe, corresponding to a once in fifty days event. For 24-h precipitation (upper panel), the gain in skill of the operational forecast relative to ERA-Interim amounts to about one forecast day and is mainly due to a higher hit rate. For 10 m wind speed the gain is between one and two forecast days and 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 (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. The verification period is July 2013 – June 2014. Figure 41 shows that the maximum PEV on forecast day 4 is about 0.4 for precipitation, and 0.2 for wind speed. For precipitation, the HRES has higher skill than ERA-Interim for all potential users, while for wind speed ERA-Interim has higher skill for users in a certain range of cost-loss ratios. At larger lead times positive PEV values exist only for a narrow range of cost–loss ratios. As with SEDI, forecasts of 10 m wind speed are less skilful than those of 24-h precipitation. The PEV of the ENS, shown in Figure 41 for comparison, demonstrates the benefit for a wide range of potential users of a probabilistic forecast of extremes.

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 42 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. Anomalies are relative to the past 20-year model climatology. ROC scores are everywhere higher than 0.5 (the monthly forecast has more skill than 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 limitations, 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/en/forecasts/charts

Figure 43 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 improvement over persistence of the medium-range (days 5–11) forecast is similar to the previous year. 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://old.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://old.ecmwf.int/publications/library/do/references/show?id=90277

6.2.2. The 2013–2014 El Niño forecasts

The year 2013 was characterized by slightly cold conditions in the eastern tropical Pacific. The majority of ensemble members of the forecasts made in spring and summer of 2013 (upper two panels in the left column of in Figure 44) predicted a return to warm conditions, which did not materialize. However, the spread within the ensemble was wide, and it included the possibility of temperatures remaining slightly on the cold side. The autumn forecast captured the basic characteristics of the next months' evolution somewhat better, suggesting first a slight drop, then a change to warm anomalies. This transition, which finally happened in spring 2014 was very well captured in the forecast made in winter 2013-14 (lower left panel in Figure 44). The multi-model EUROSIP forecasts (right column) performed slightly better in the sense that the ensemble was better centred on the observations.

6.2.3. Tropical storm predictions from the seasonal forecasts

The 2013 North Atlantic hurricane season was exceptionally quiet with an accumulated cyclone energy index (ACE) of just 35% of the 1950-2012 climate average (see Figure 46), although the number of tropical storms which formed in 2013 (13 named storms) was slightly above average (12). Seasonal tropical storm predictions from System 4 indicated slightly below average activity compared to climatogy over the Atlantic. The June forecast predicted 11 (with a range from 7 to 14) tropical storms in the Atlantic (Figure 45) and an ACE of 80% of the observed climatology (+/-20%). Most other seasonal forecast models predicted an active to very active 2013 Atlantic tropical storm season.

System 4 predicted below average activity over the eastern North Pacific (ACE 20% below normal) and normal tropical storm activity over the western North Pacific (Figure 45). The 2013 eastern Pacific hurricane season was tied for the most active since 1992 for the number of tropical storms (16 tropical stormed formed over the eastern North Pacific between July and December), although most of the storms remained weak. The ACE was 10% below average. 24 tropical storms formed over the western North Pacific in 2013, which is slightly above average (21.3). However, the ACE over the western North Pacific was also about 10% below average.

In summary, the drop of ACE in the Atlantic sector was captured by the forecast, although the magnitude of the decrease was underestimated. The tendency of the model to underforecast the number of storms is consistent with its too strong tendency towards a return of El Nino conditions in this period (Figure 44).

6.2.4. Extratropical seasonal forecasts

The recent winter was characterized by an exceptional zonal flow across the Atlantic with maximum westerly anomalies located at about 40N. It is a seasonal record over the ERA-Interim period. The extent of these anomalies could only be partially captured by the projections onto the positive phase of the NAO because the gradient between the Icelandic low and the high over the Azores, on which it based, is located further north. Although the seasonal forecasts (System 4 and Eurosip) predicted enhanced zonal flow over the Atlantic, the extent of the low anomalies centred over the North Atlantic and the UK was underestimated (Figure 47).

The severe cold conditions experienced in the U.S. Midwest were not anticipated by the seasonal forecasts (Figure 48). Predicted SSTs were in a broad sense realistic showing a warm southern Indian Ocean and southern west Pacific and large warm anomalies over the Gulf of Alaska.

Further diagnostics suggest that there was a dynamical linkage between the Atlantic circulation pattern and the flow over North America and the North Pacific, indicating that part of the potential predictability of the anomalies in winter 2013-14 originated in the tropics.

References

Ahlgrimm, M., and R. Forbes, 2014: Improving the representation of low clouds and drizzle in the ECMWF model based on ARM observations from the Azores. *Mon. Wea. Rev.,* **142,** 668-685.

Bechtold, P., N. Semane, P. Lopez, J.-P. Chaboreau, A. Beljaars, and N. Bormann, 2014: Representing equilibrium and nonequilibrium convection in large-scale models. *J. Atmos. Sci.,* **71,** 734-753.

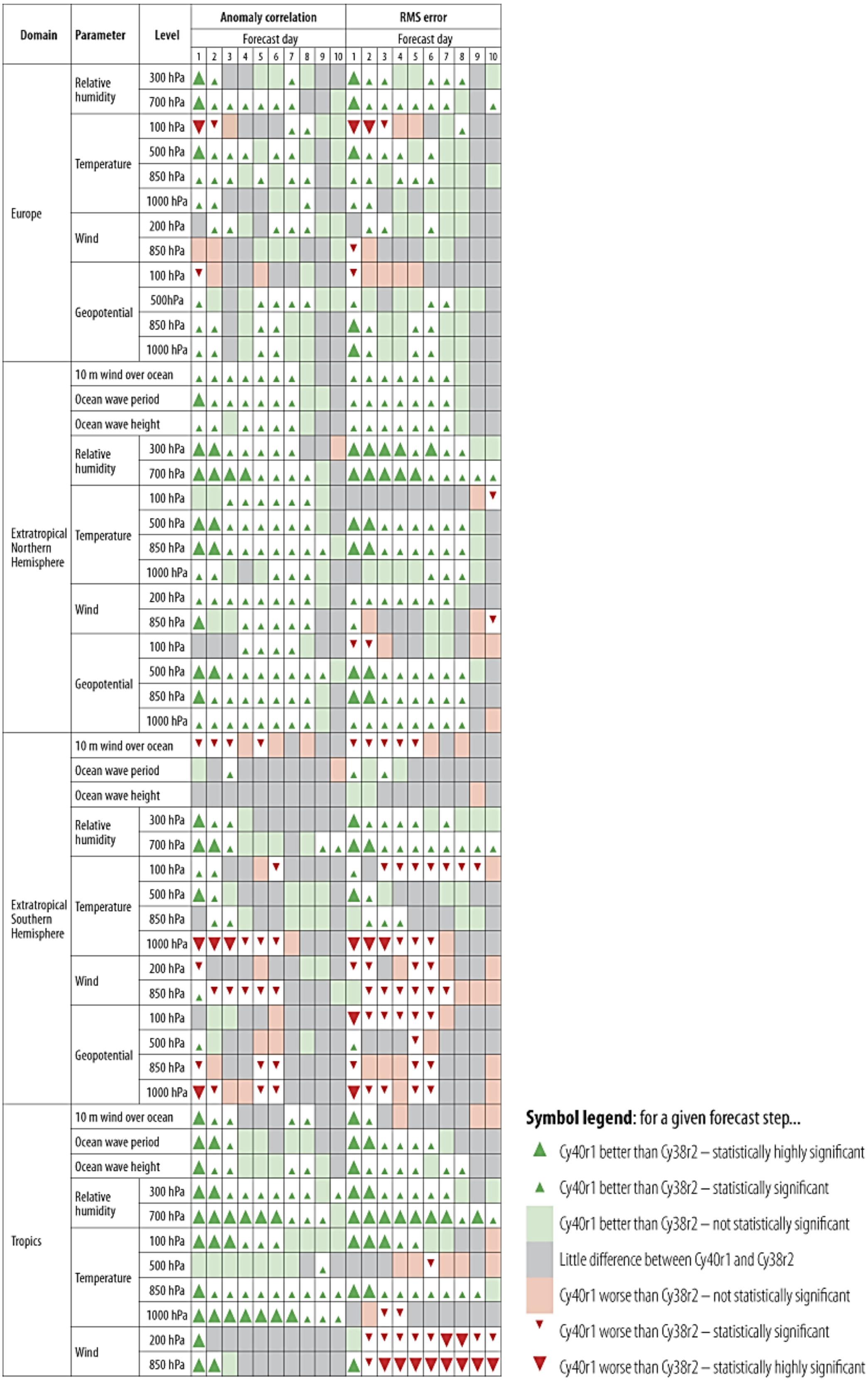


Figure 1: Summary score card for Cy40r1. Score card for cycle 40r1 versus cycle 38r2 verified by the respective analyses at 00 and 12 UTC for 244 days in the period 6 February 2012 to 3 November 2013. Verification is also carried out against observations, but this is not shown.

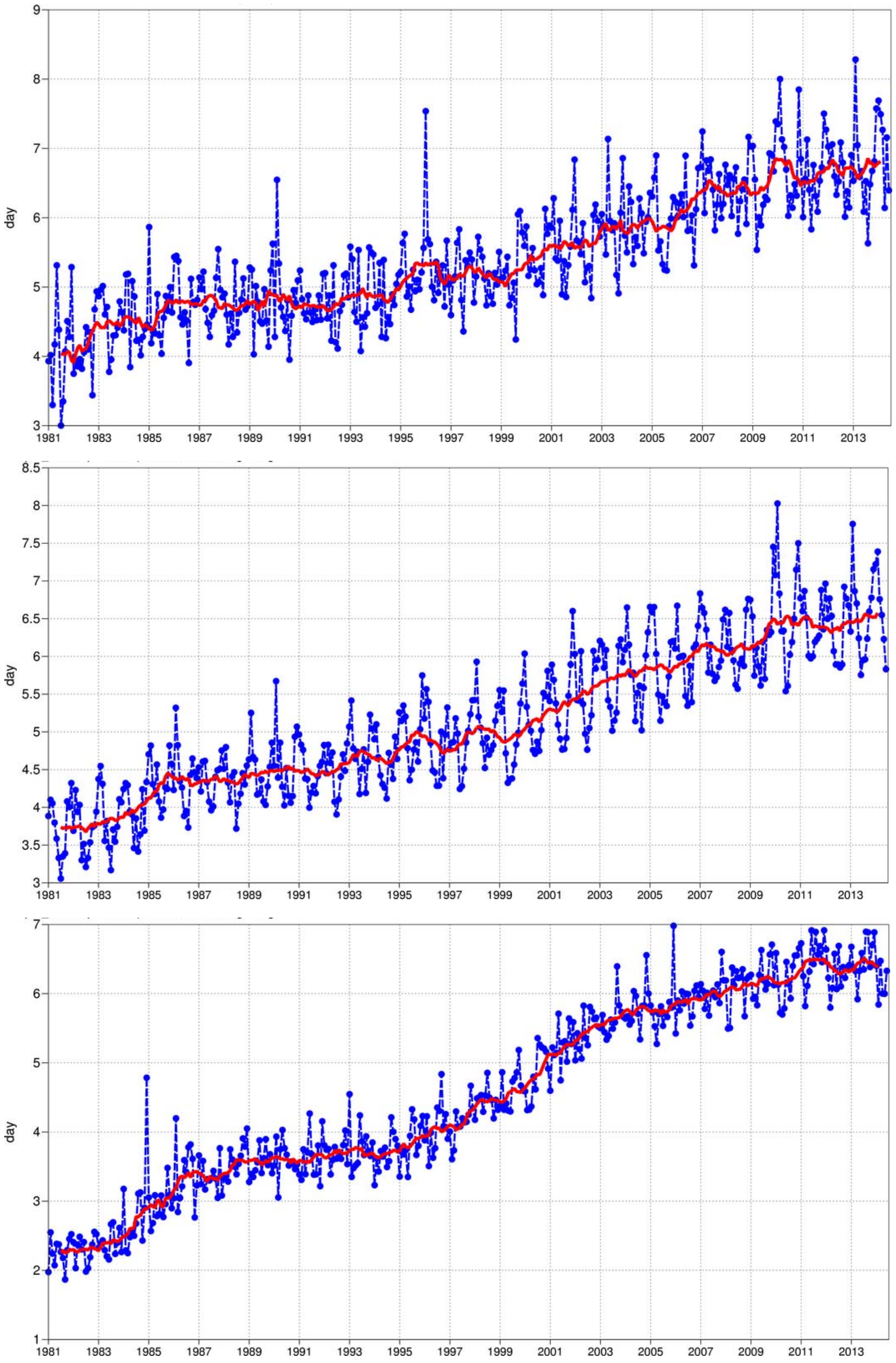
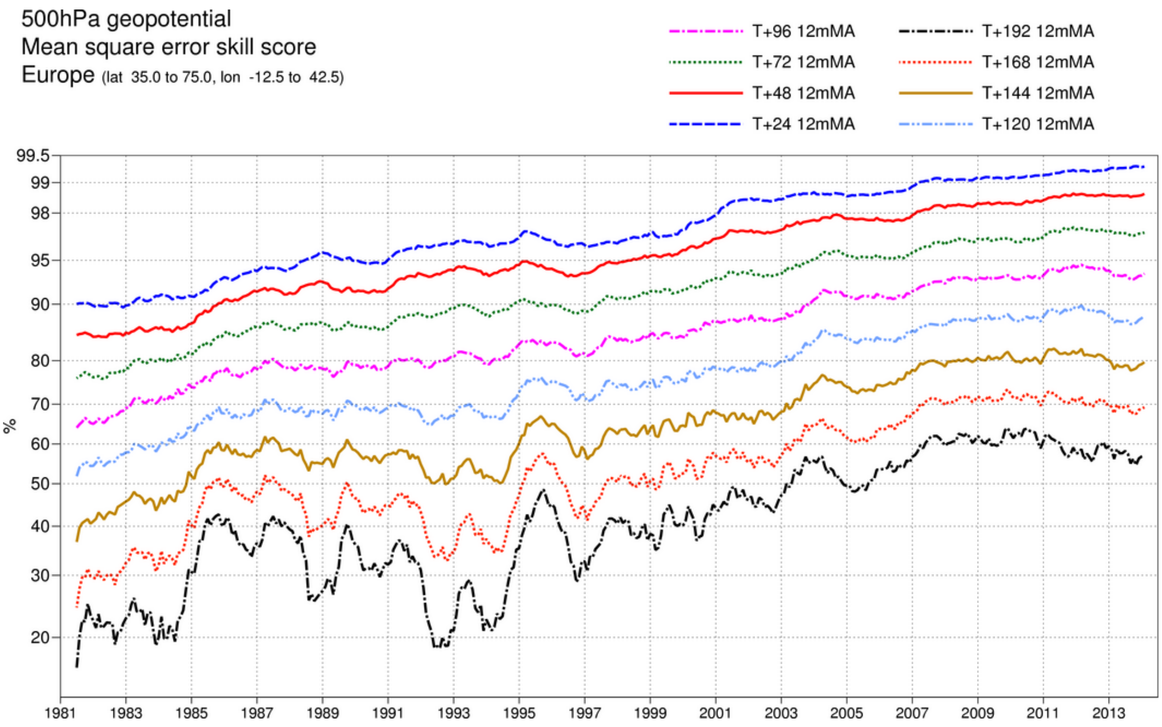


Figure 2: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).



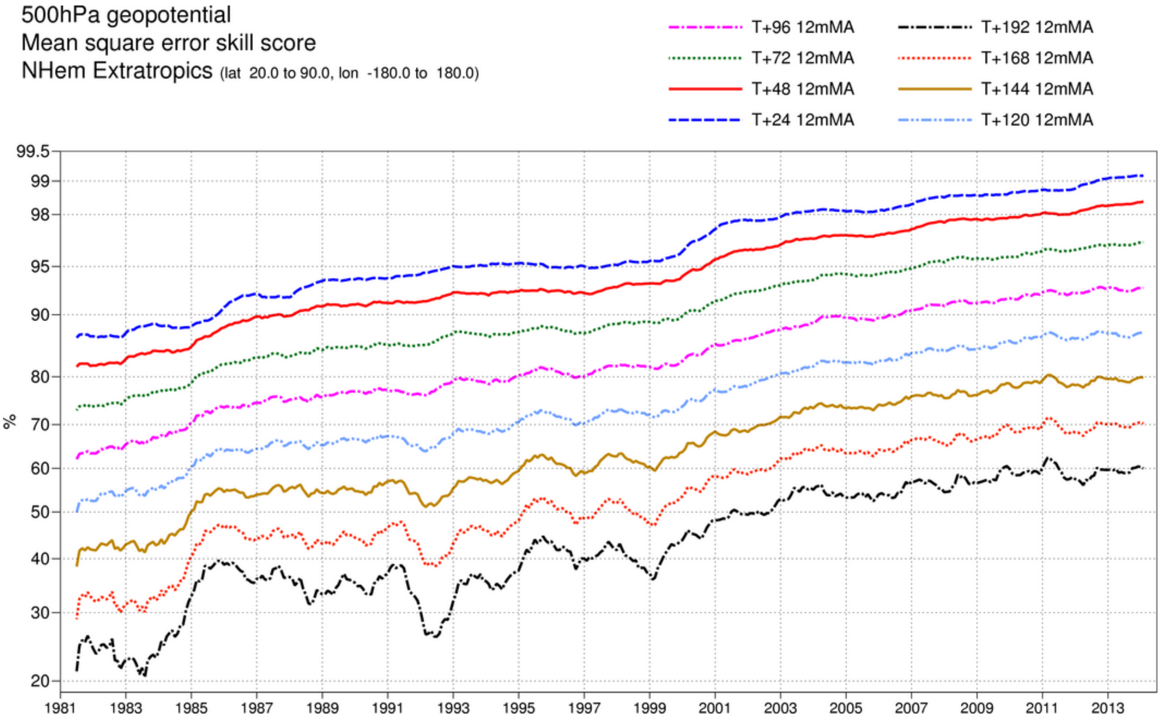


Figure 3: 500 hPa geopotential height mean square error skill score for Europe (top) and the northern hemisphere extratropics (bottom), showing 12-month moving averages for forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2013–July 2014.

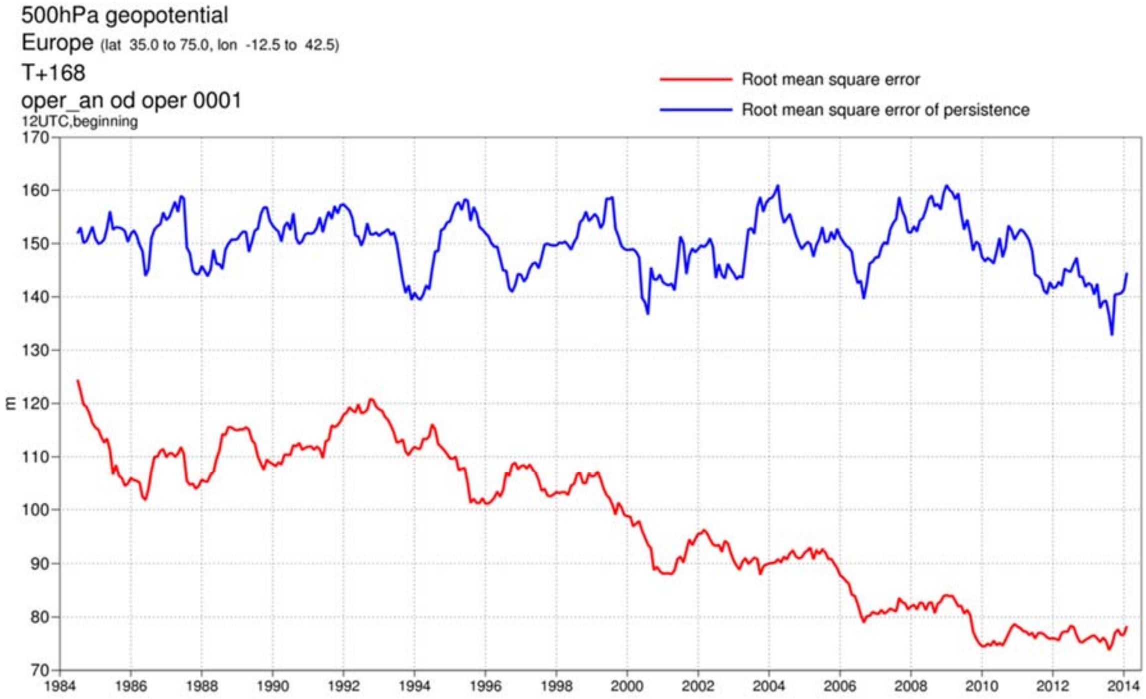
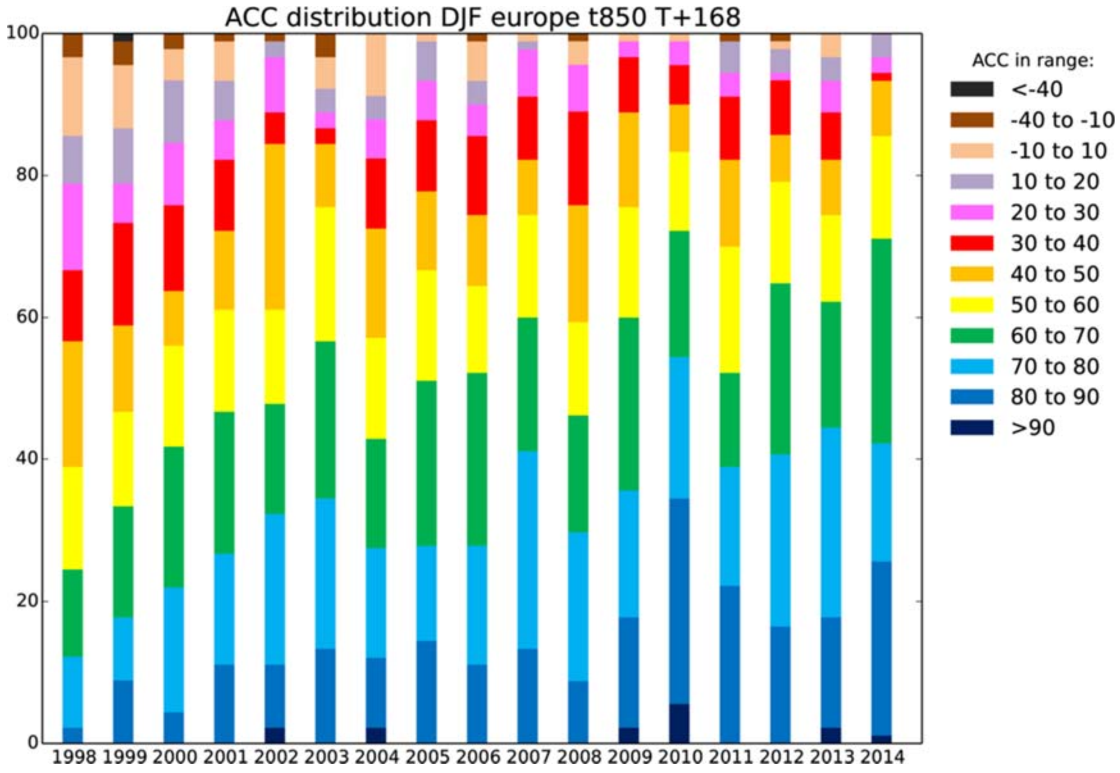


Figure 4: Root mean square (RMS) error of forecasts made by persisting the analysis over 6 days (144 hours) and verifying it as a forecast for 500 hPa geopotential height over Europe (blue). The RMS error of the forecast at day 6 is shown in red. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2013–July 2014.



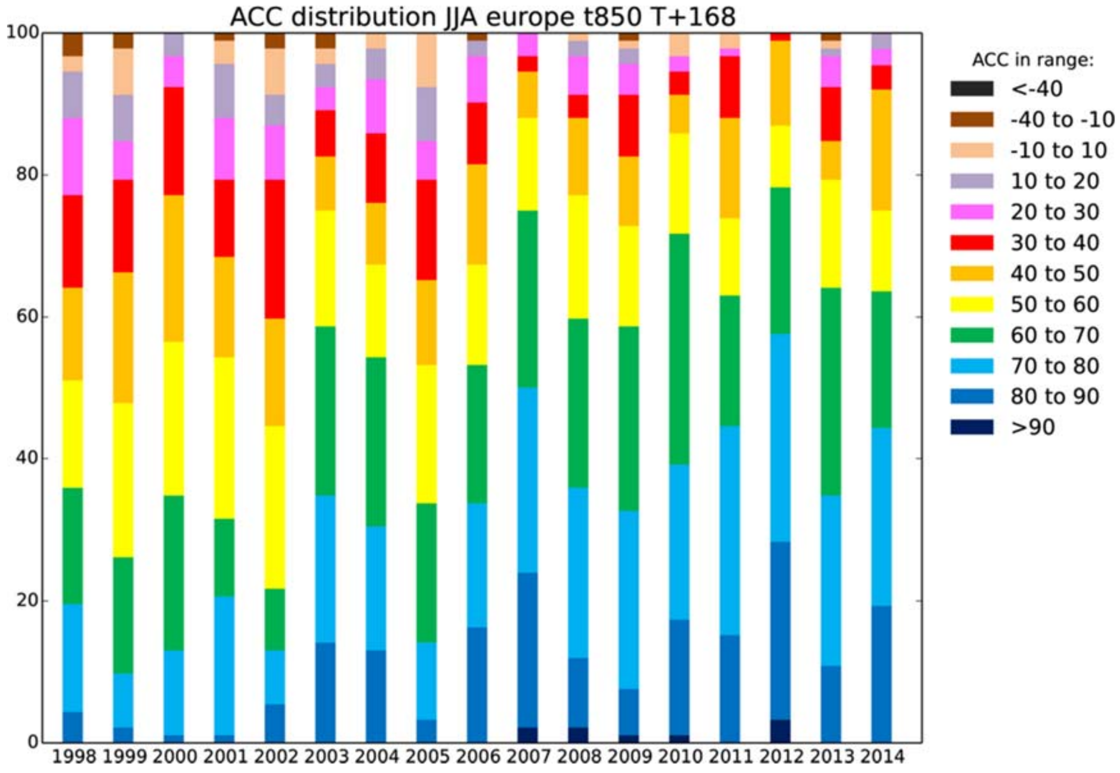
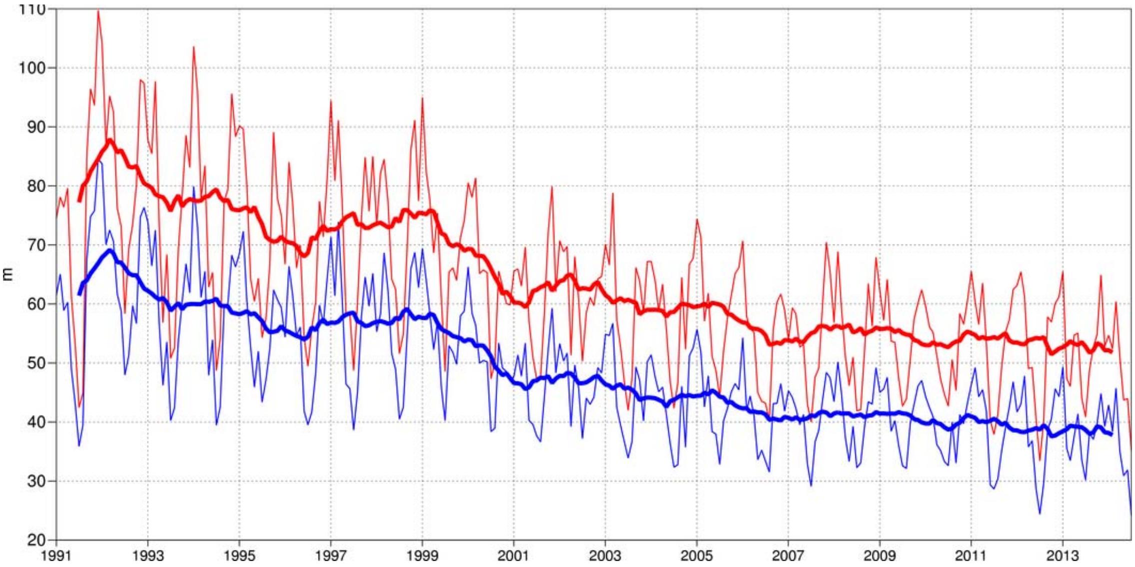


Figure 5: Distribution of ACC of the day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997–1998.



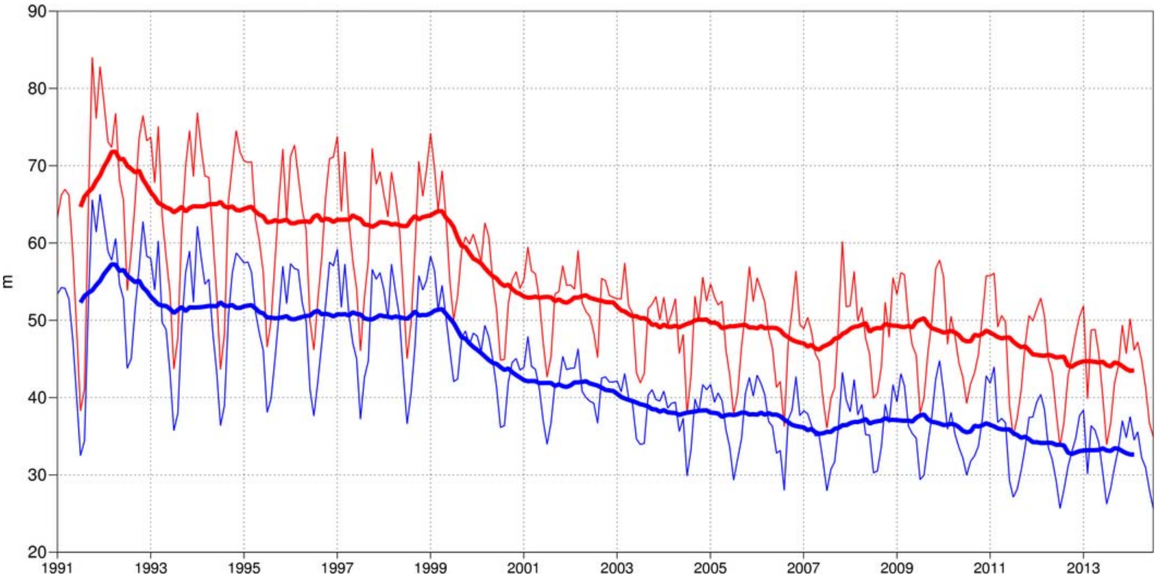
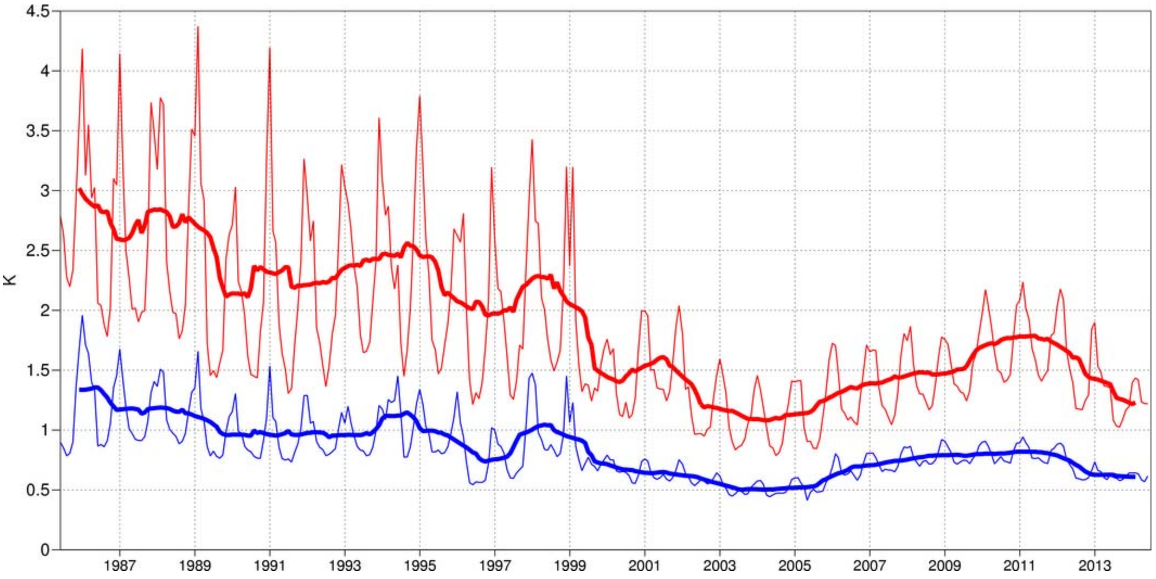


Figure 6: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores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



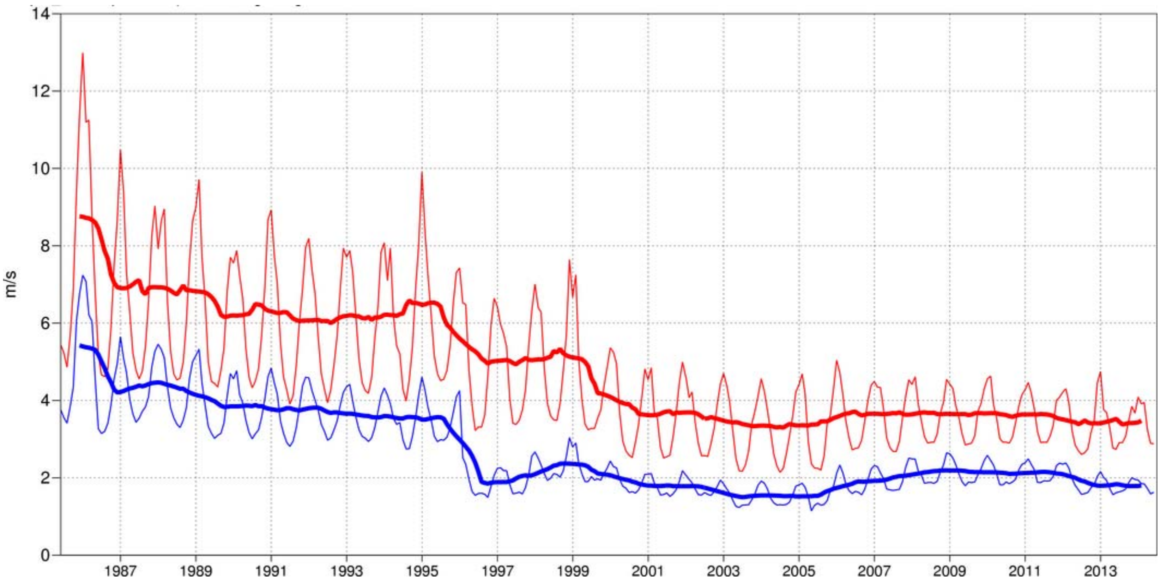
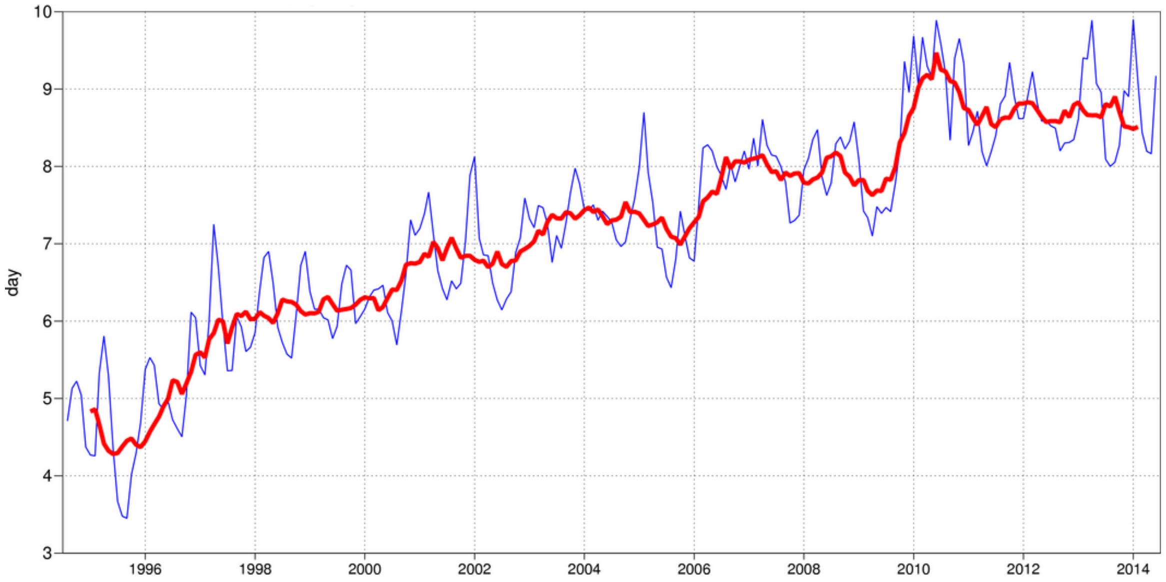


Figure 7: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).



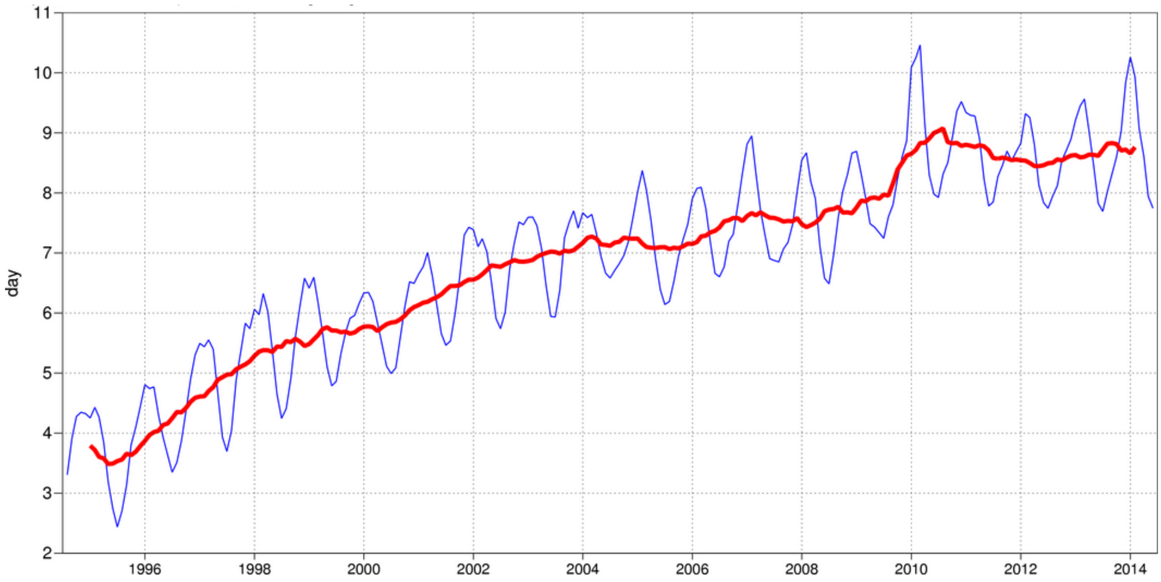


Figure 8: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance – each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous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

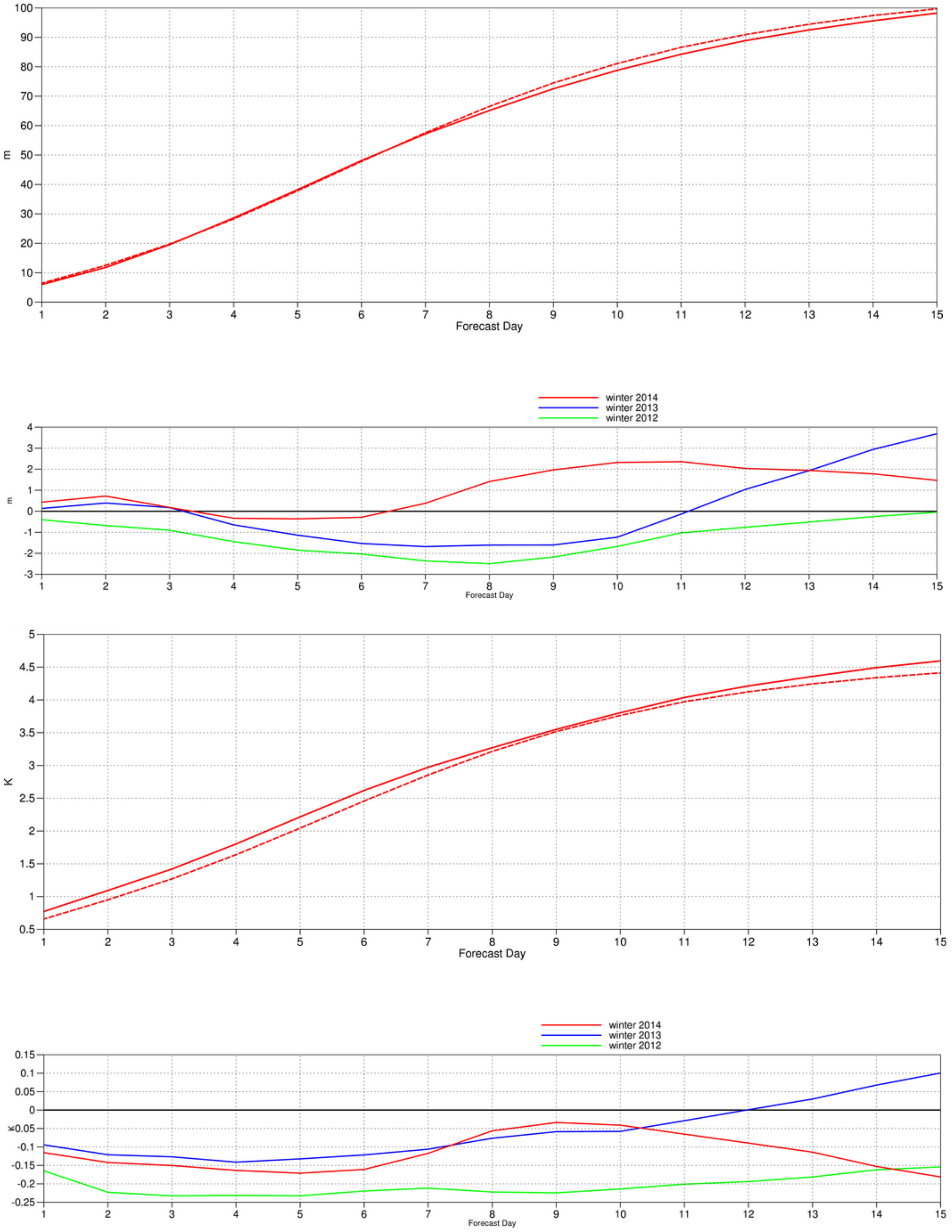


Figure 9: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2013–2014 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.

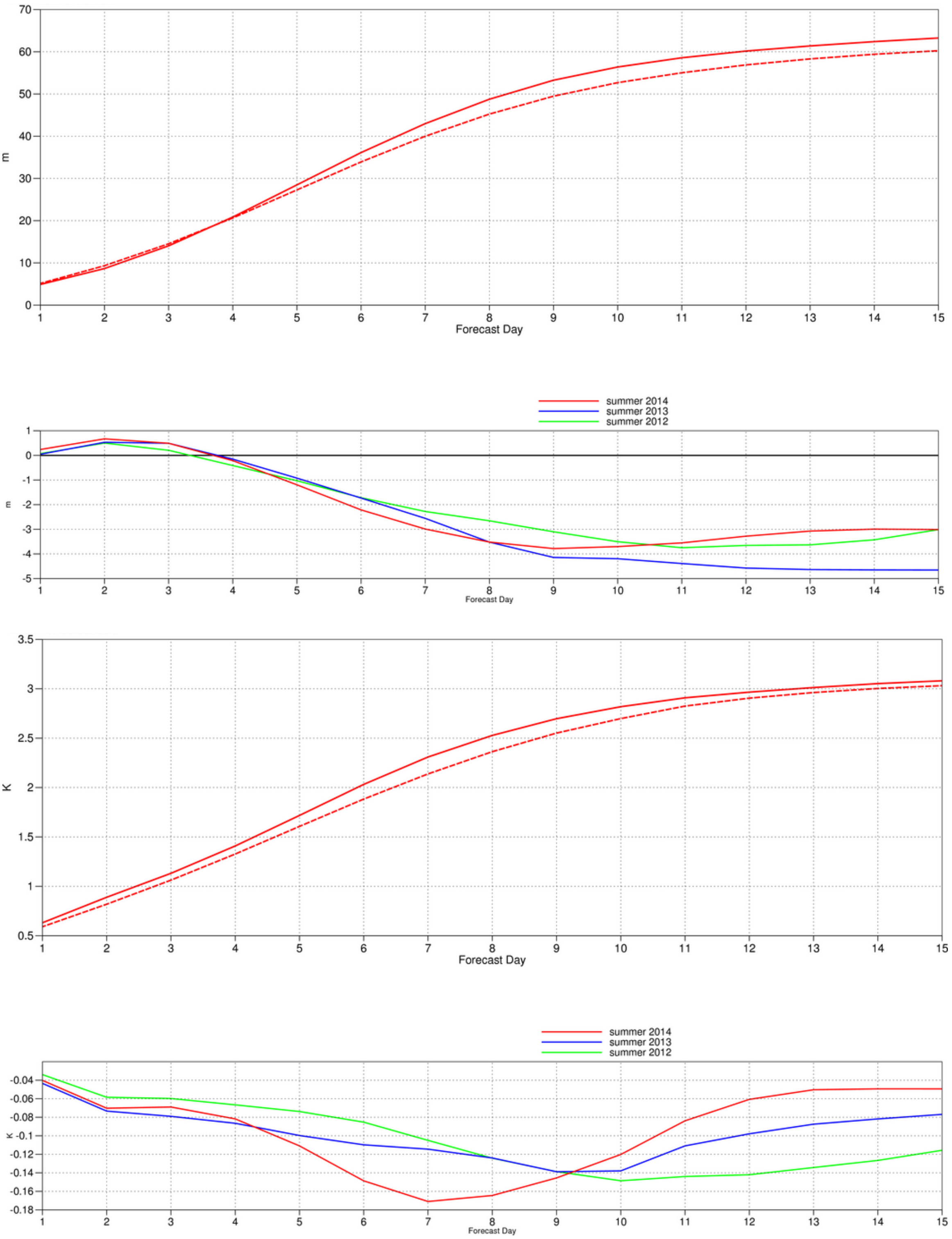
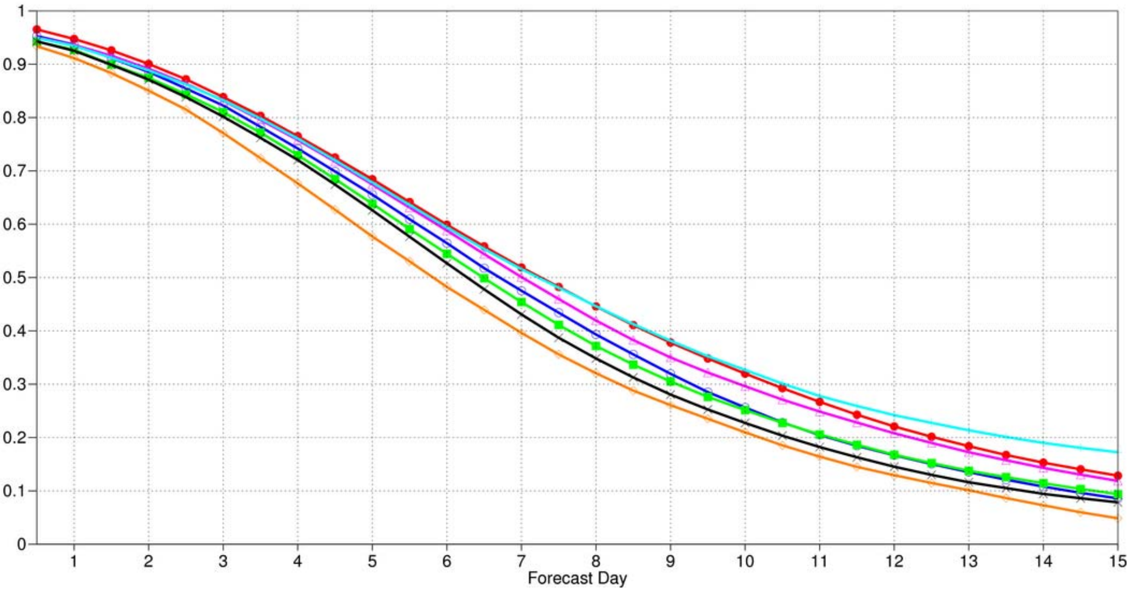


Figure 10: As Figure 9 for summer seasons.



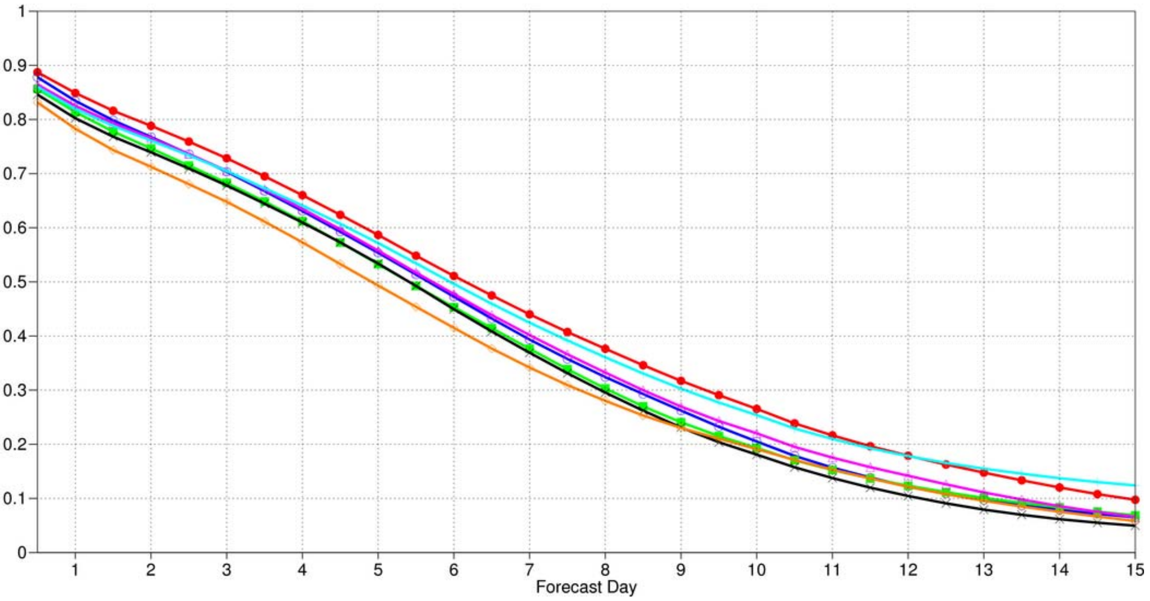
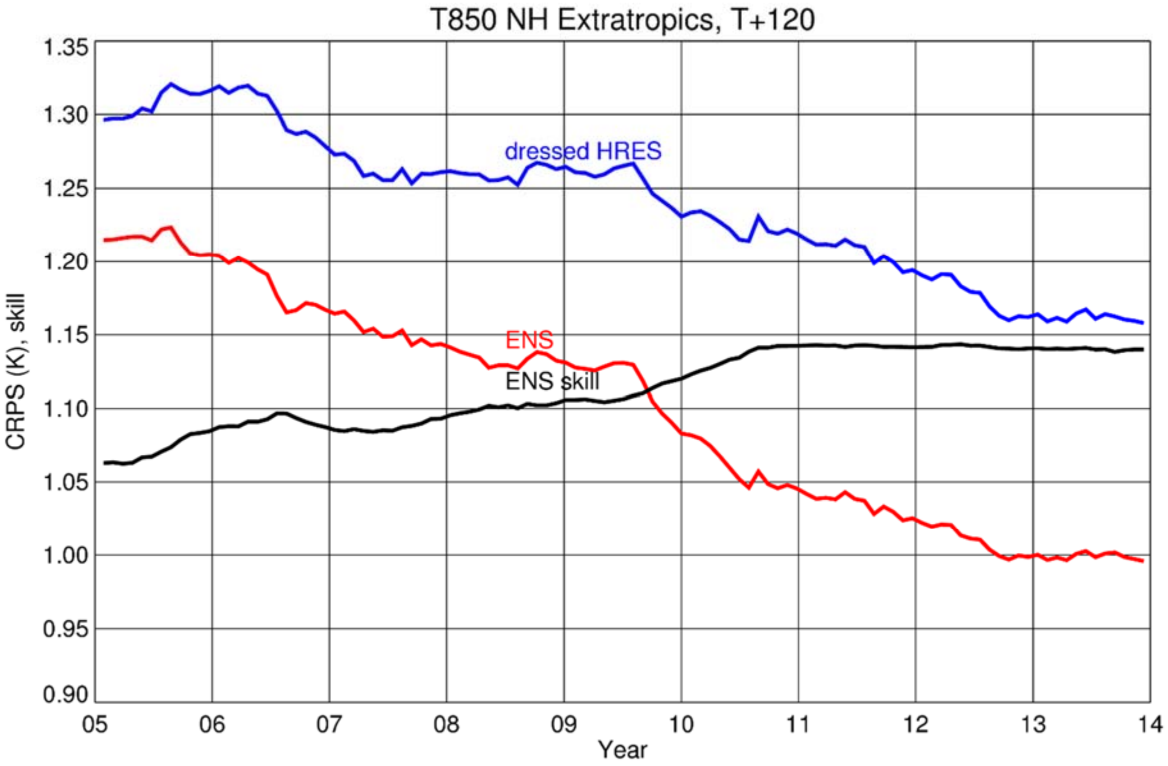


Figure 11: CRPSS for 500 hPa height (top) and 850 hPa temperature (bottom) ensemble forecasts for winter (December–February) over the extratropical northern hemisphere. Skill from the ensemble day 1–15 forecasts is shown for winters 2013–14 (red), 2012–13 (blue), 2011–12 (green), 2010–11 (magenta), 2009–10 (cyan), 2008–09 (black) and 2007–08 (orange).



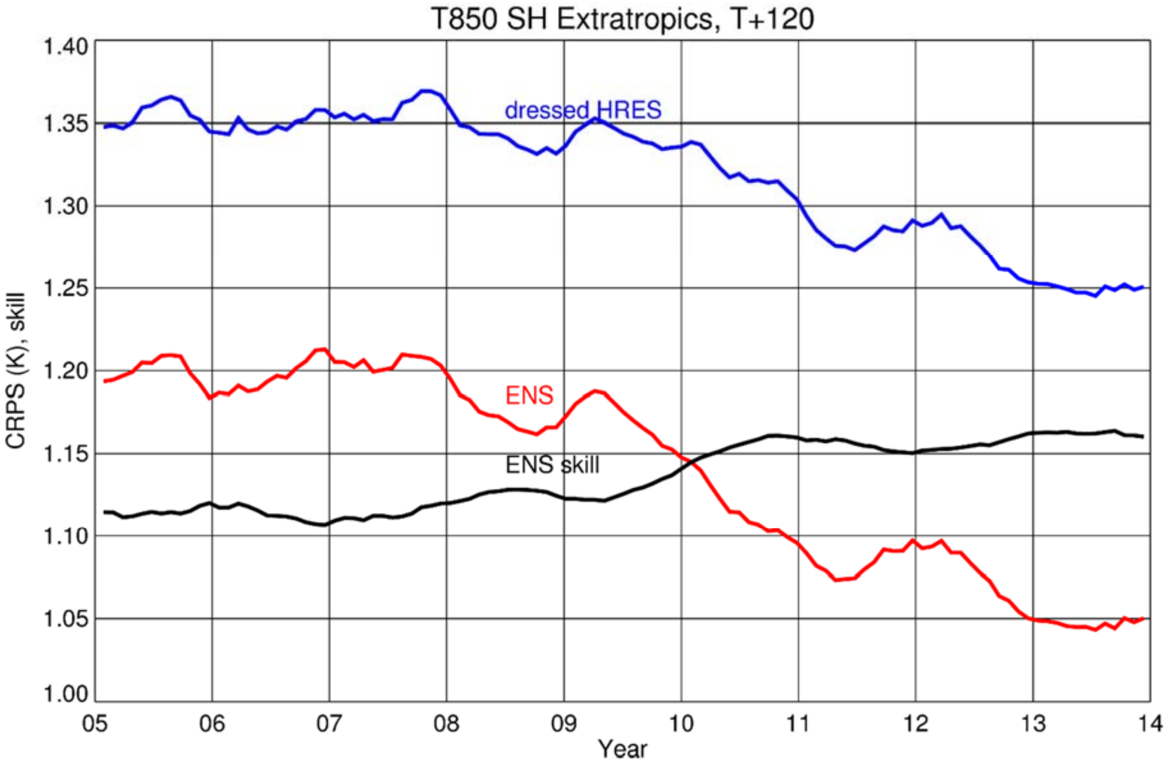
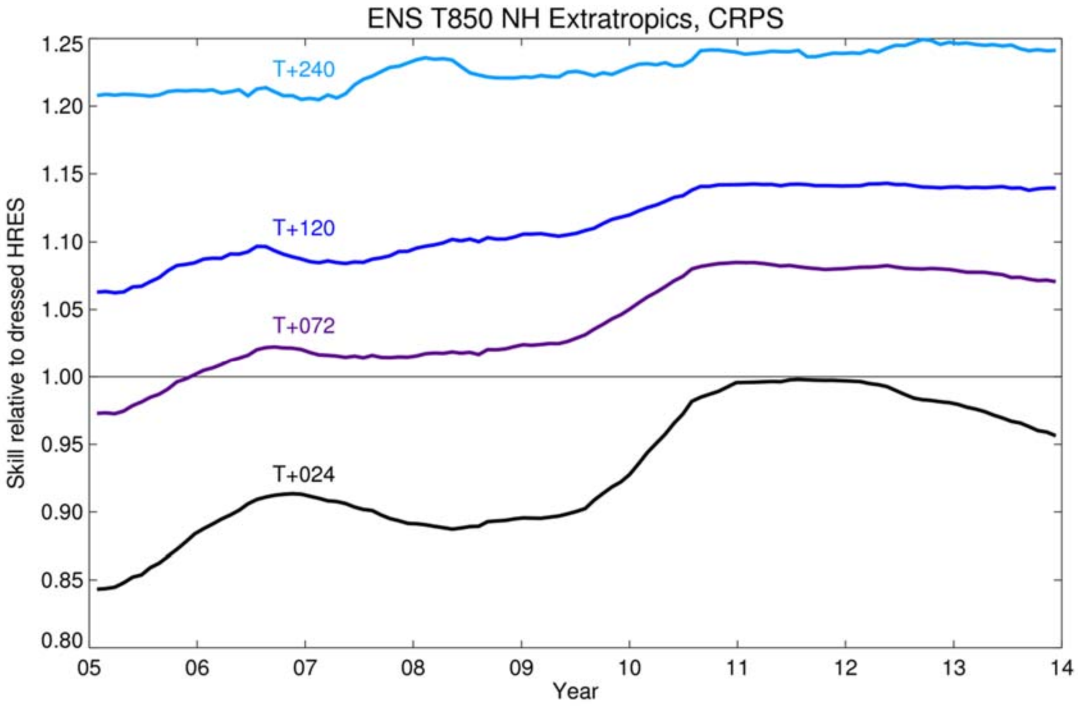


Figure 12: CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5. Scores are shown for the ensemble forecast (red) and the dressed HRES (blue). Black curves show the skill of the ENS relative to the dressed HRES. Values are running 12-month averages. Note that for CRPS (red and blue curves) lower values are better, while for CRPS skill (black curve) higher values are better.



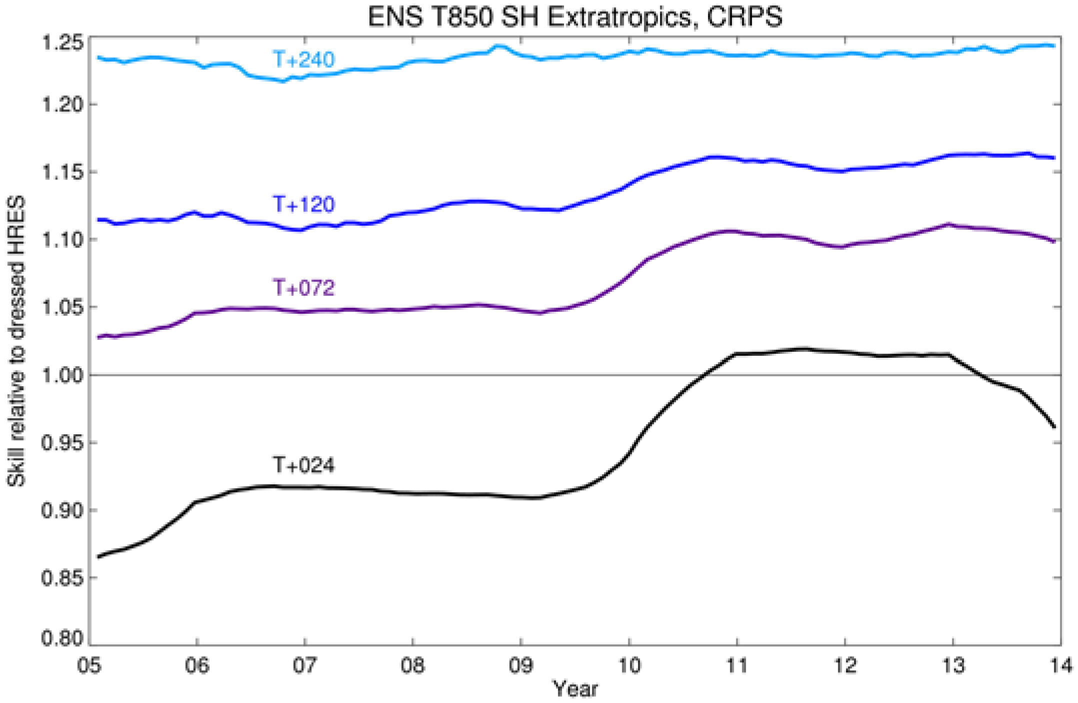


Figure 13: CRPS skill of the ENS relative to the dressed HRES for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics. Values are running 12-month averages.

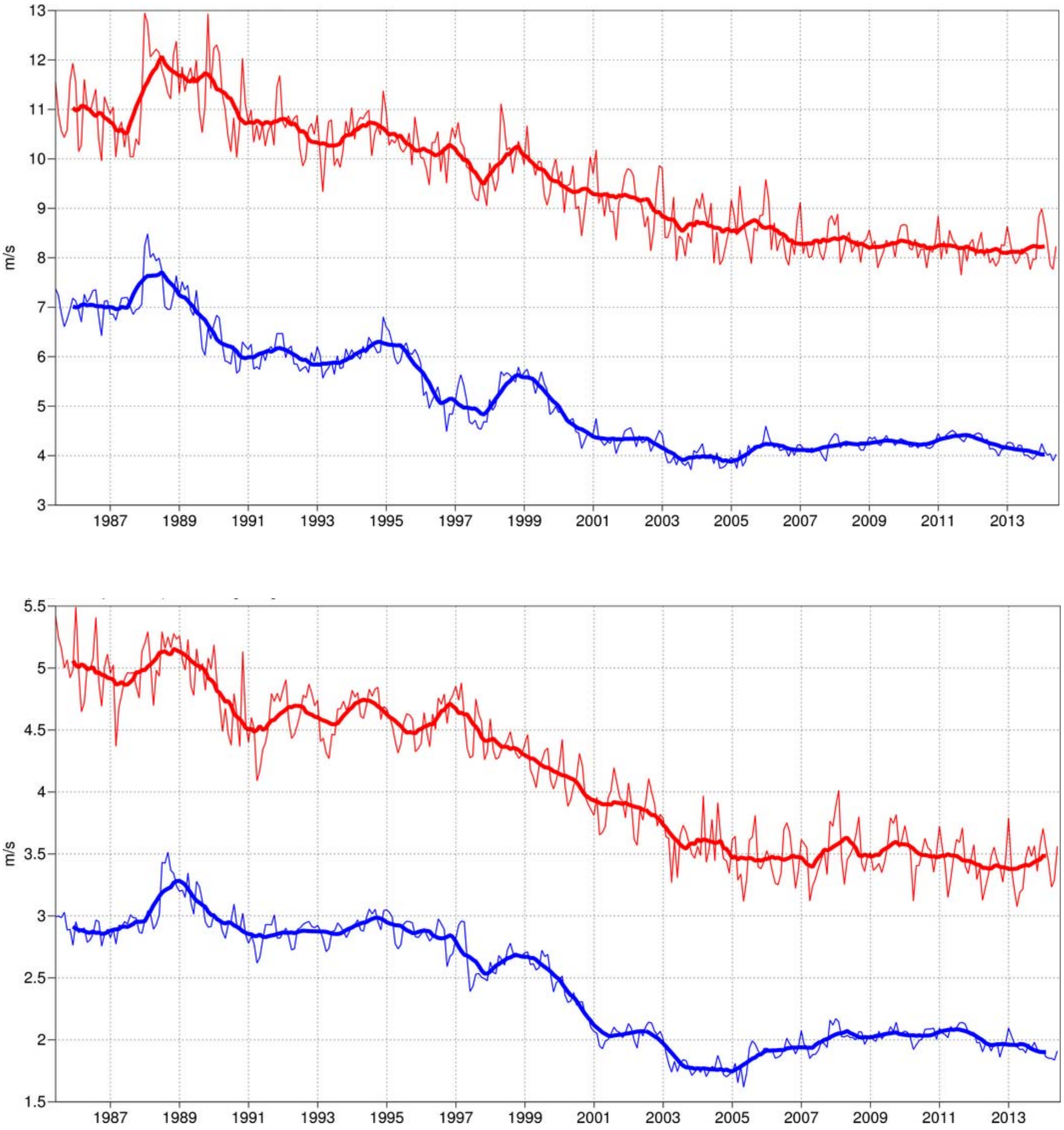


Figure 14: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts. 12-month moving average scores are also shown (in bold).

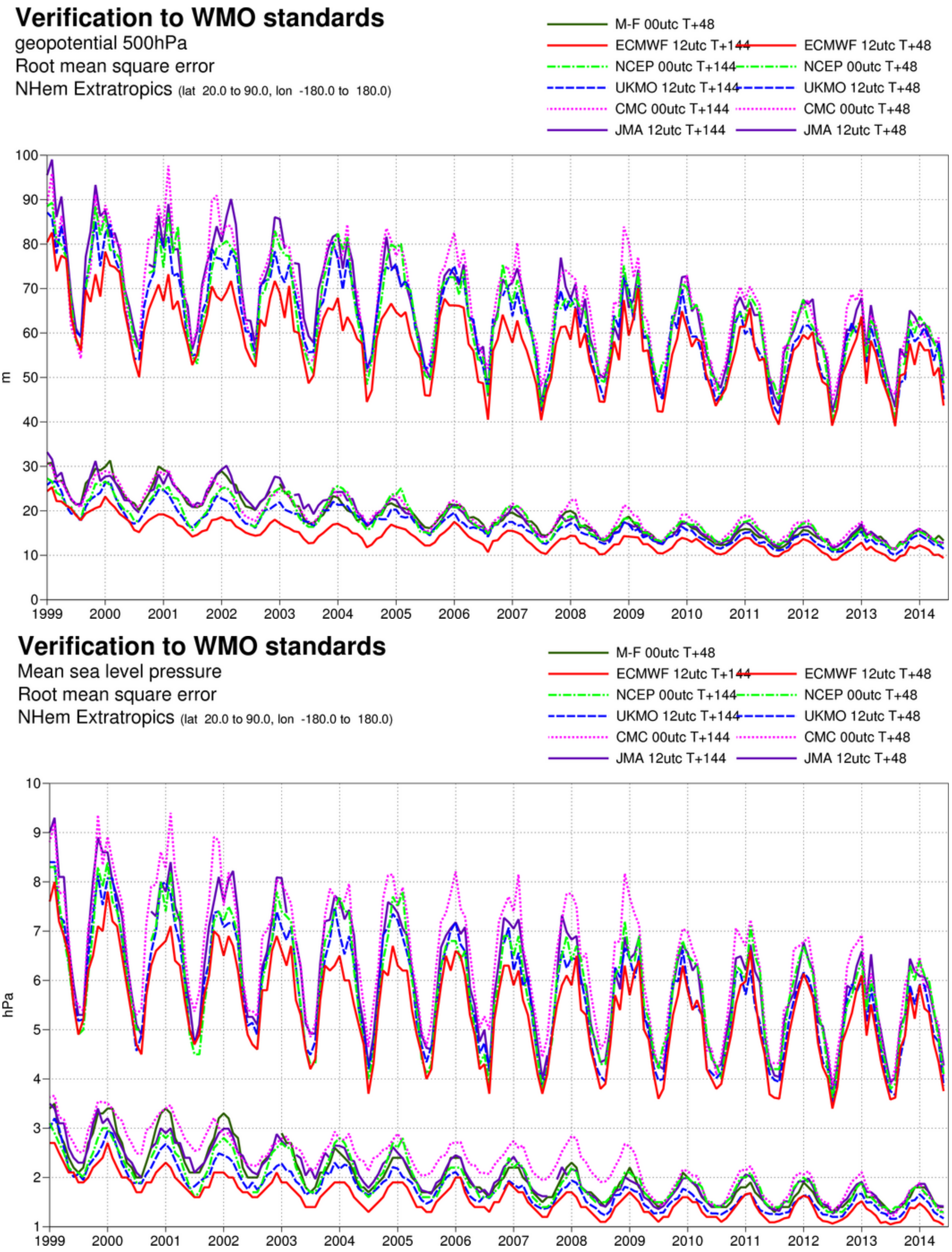
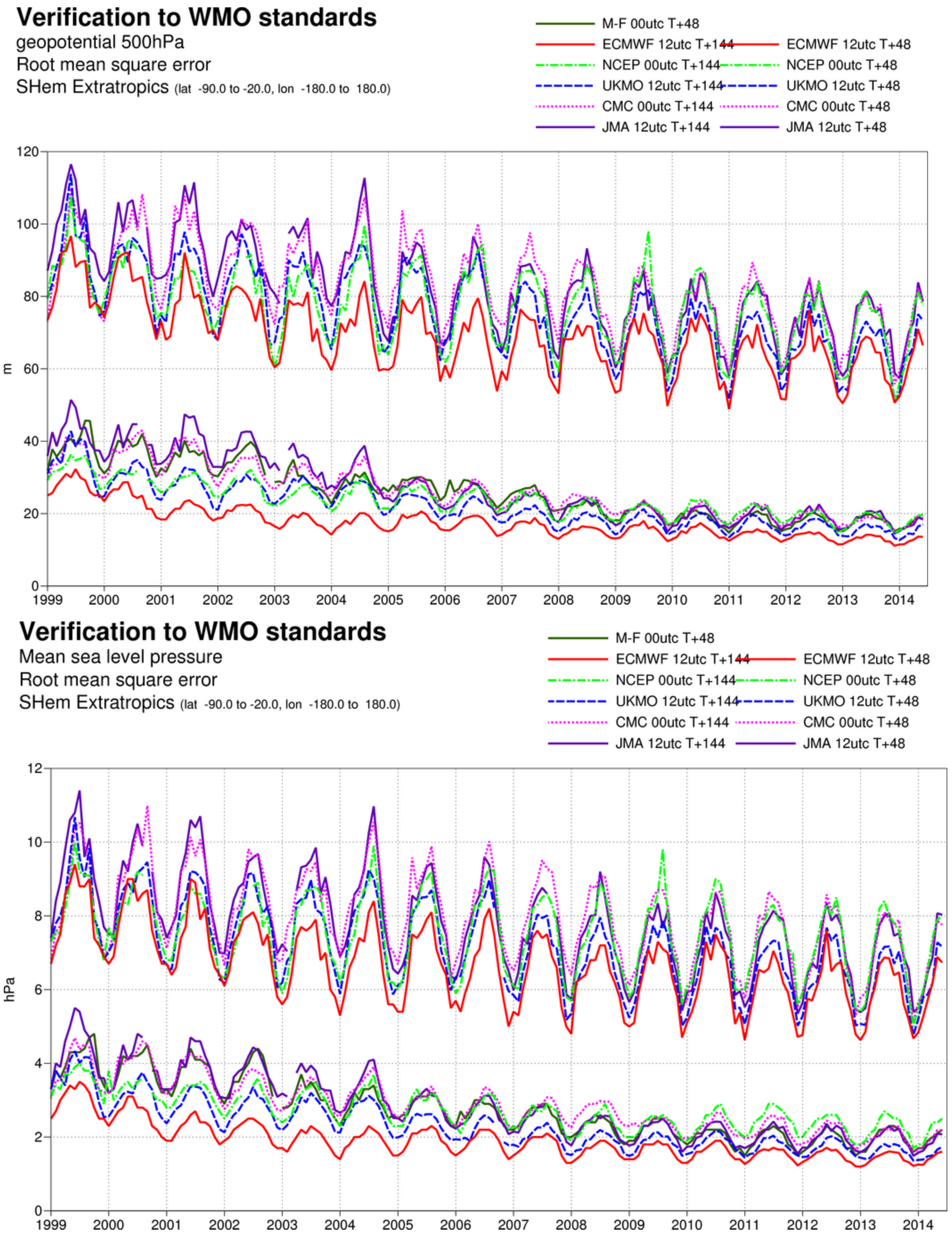
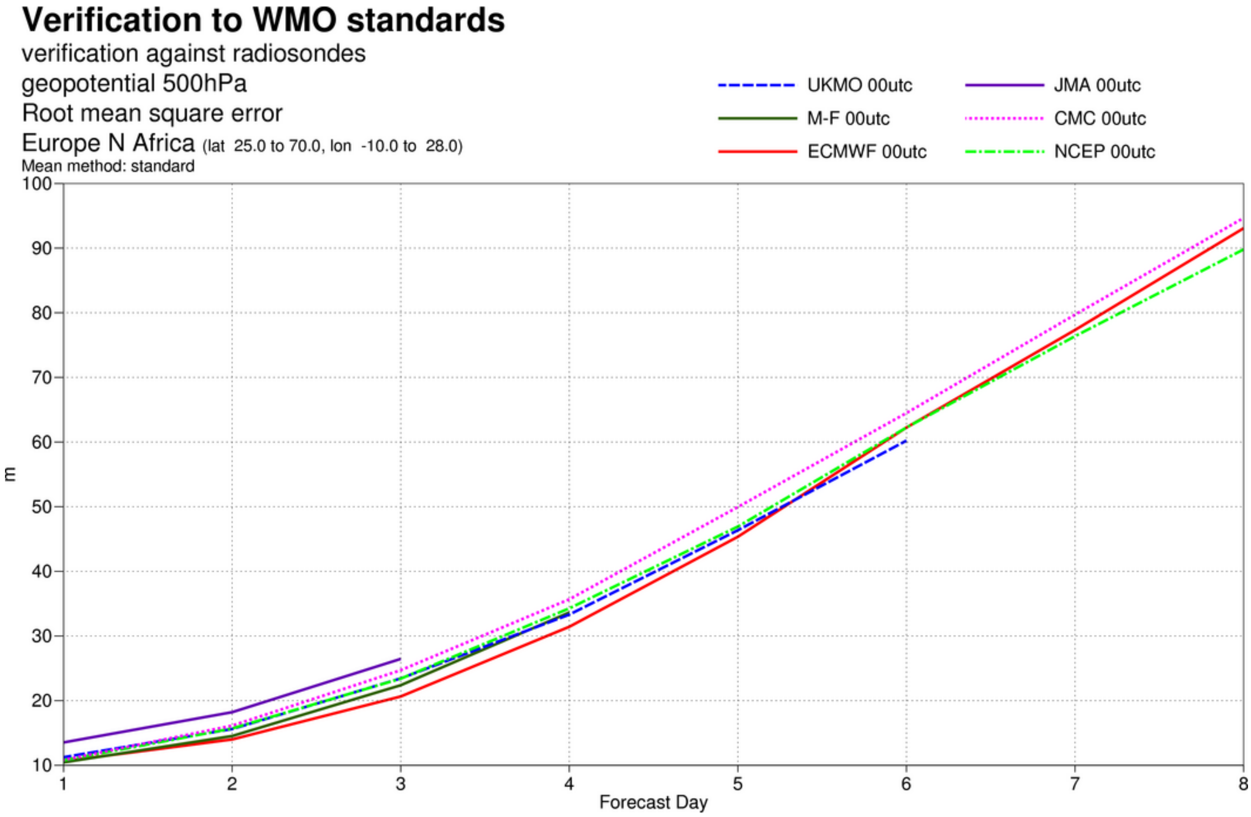


Figure 15: WMO-exchanged scores from global forecast centres. RMS error over northern extratropics for 500 hPa geopotential height (top) and mean sea level pressure (bottom). In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model 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.


Verification to WMO standards
geopotential 500hPa
Root mean square error
SHem Extratropics (lat -90.0 to -20.0, lon -180.0 to 180.0)

Verification to WMO standards
Mean sea level pressure
Root mean square error
SHem Extratropics (lat -90.0 to -20.0, lon -180.0 to 180.0)

Figure 16: As Figure 15 for the southern hemisphere.



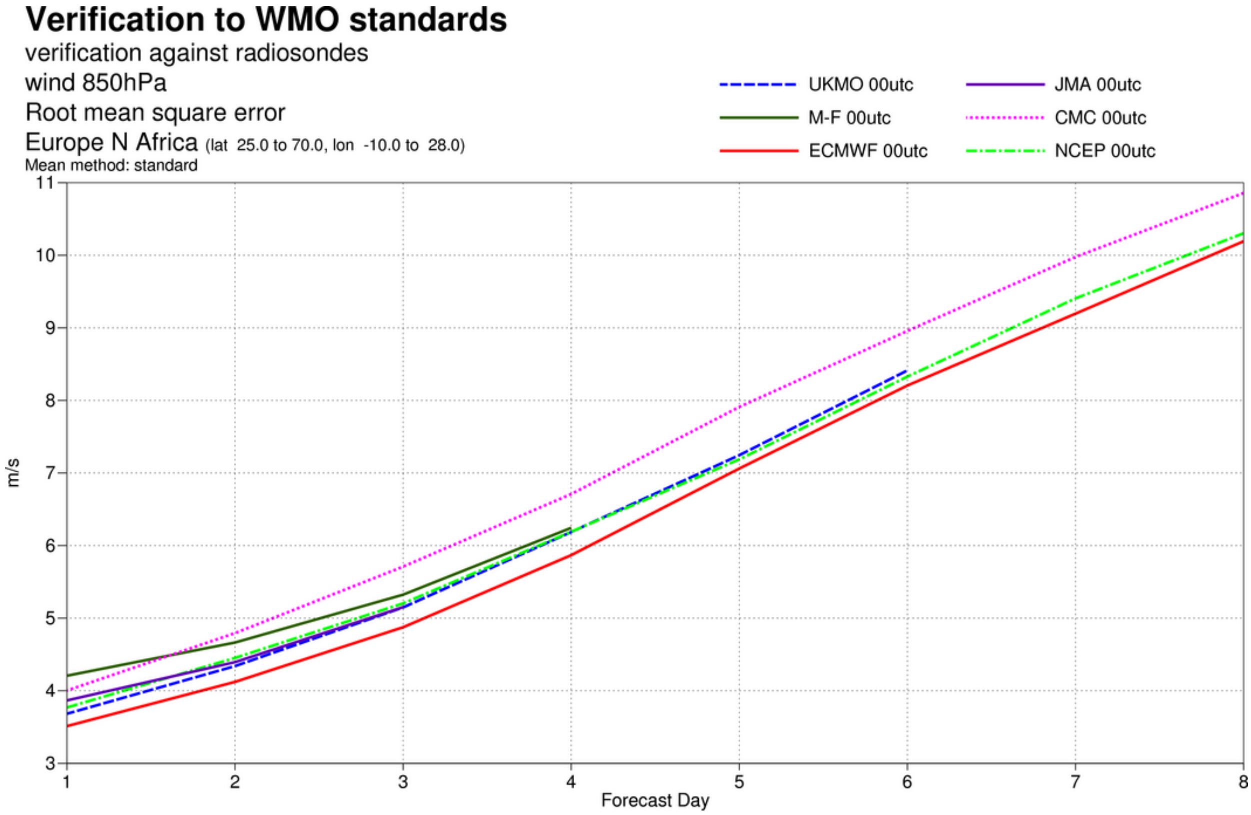
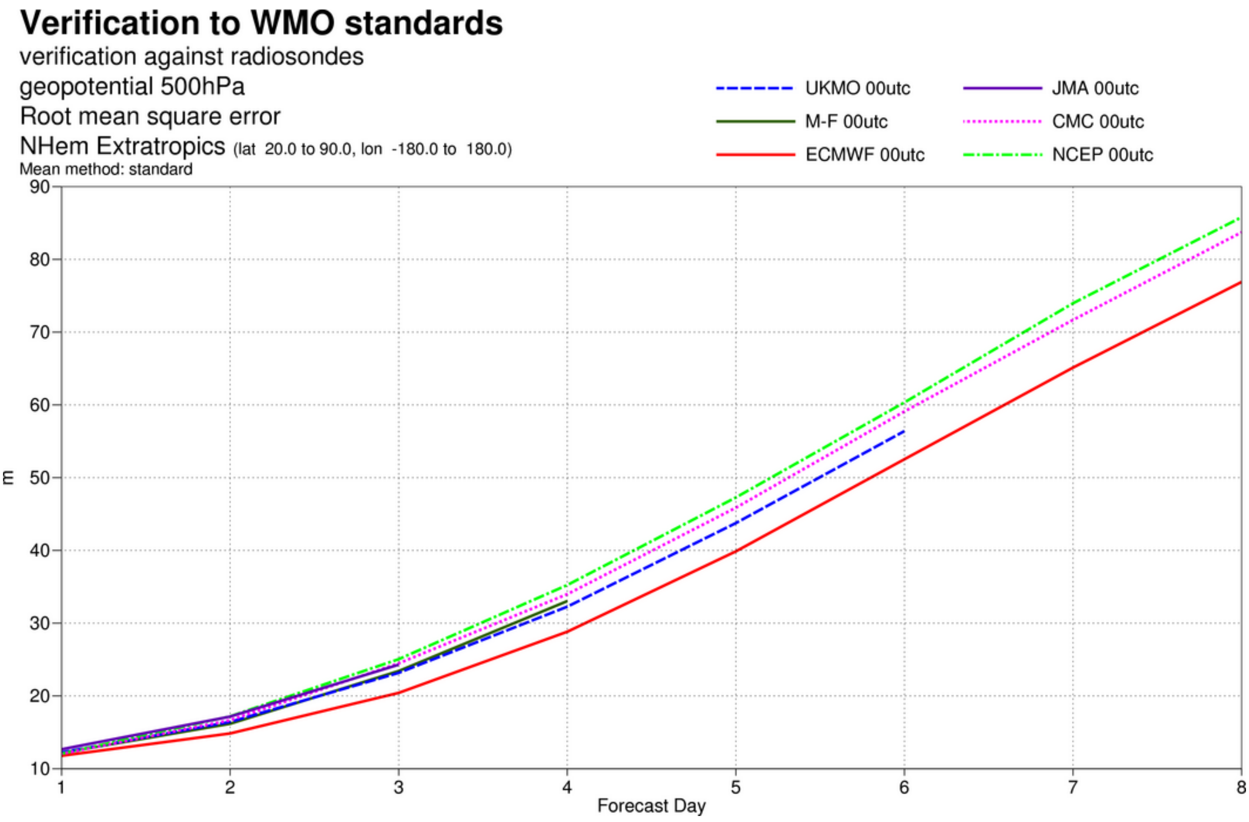


Figure 17: WMO-exchanged scores using radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2013–July 2014).



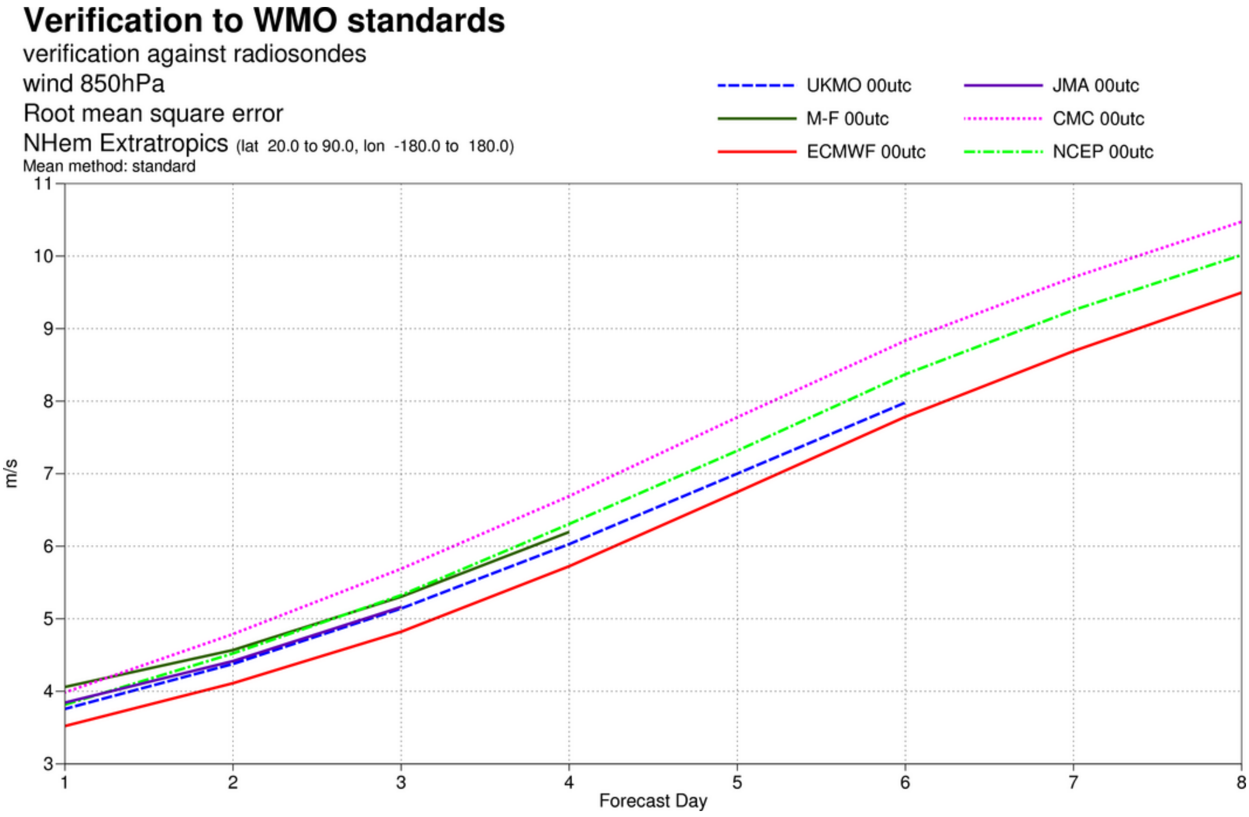


Figure 18: As Figure 17 for the northern hemisphere extratropics.

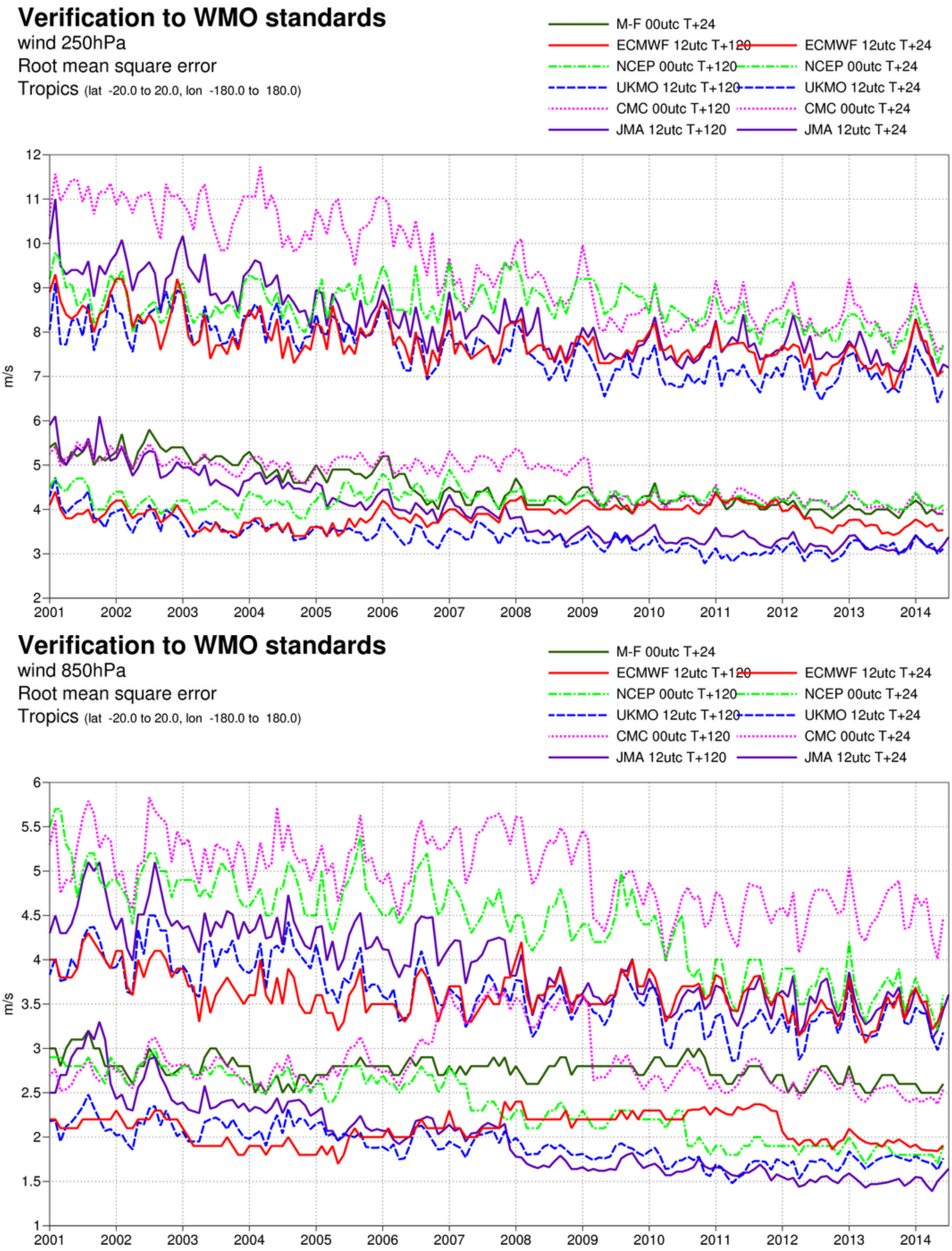
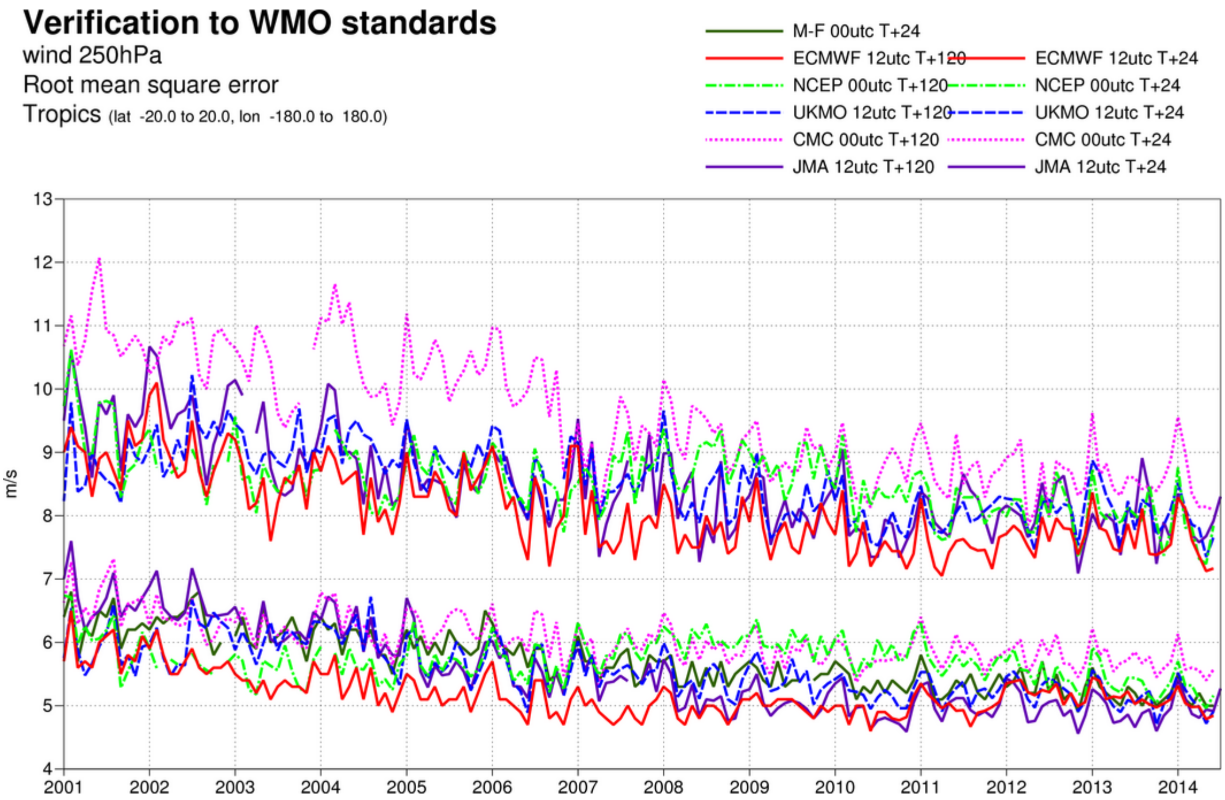


Figure 19: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



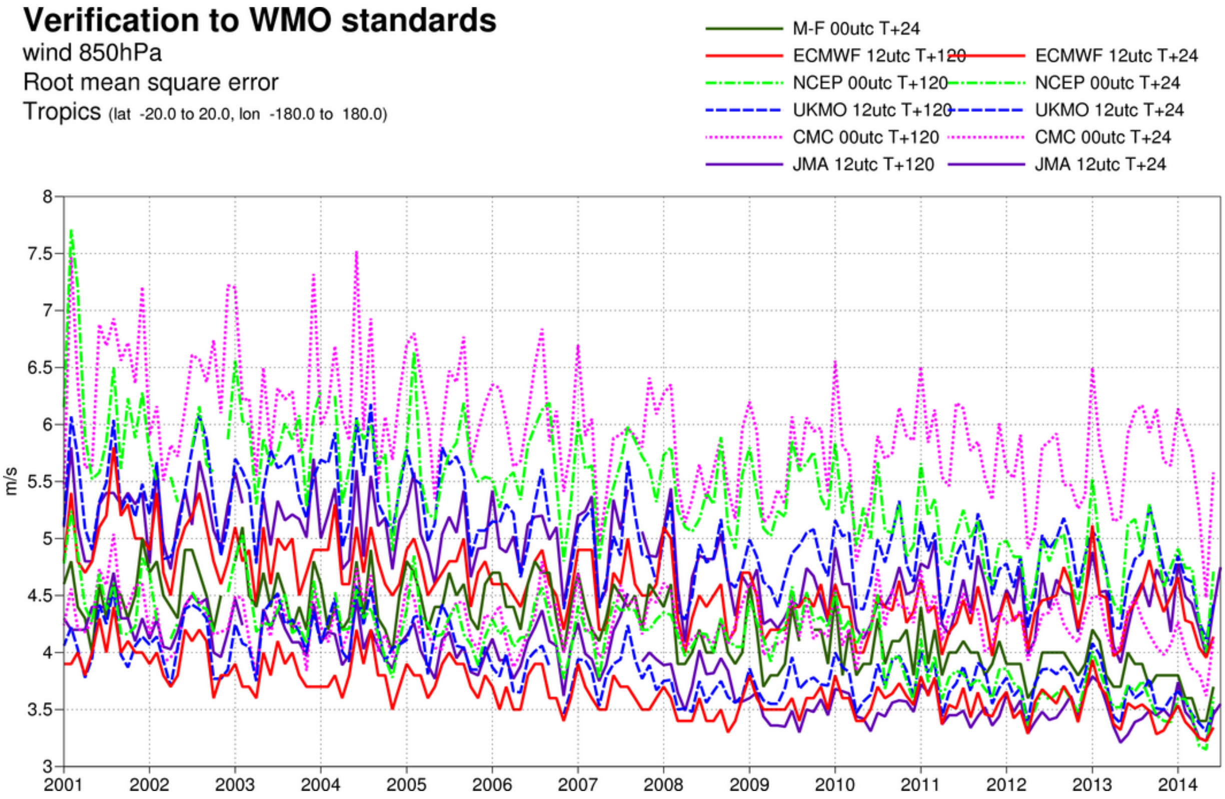
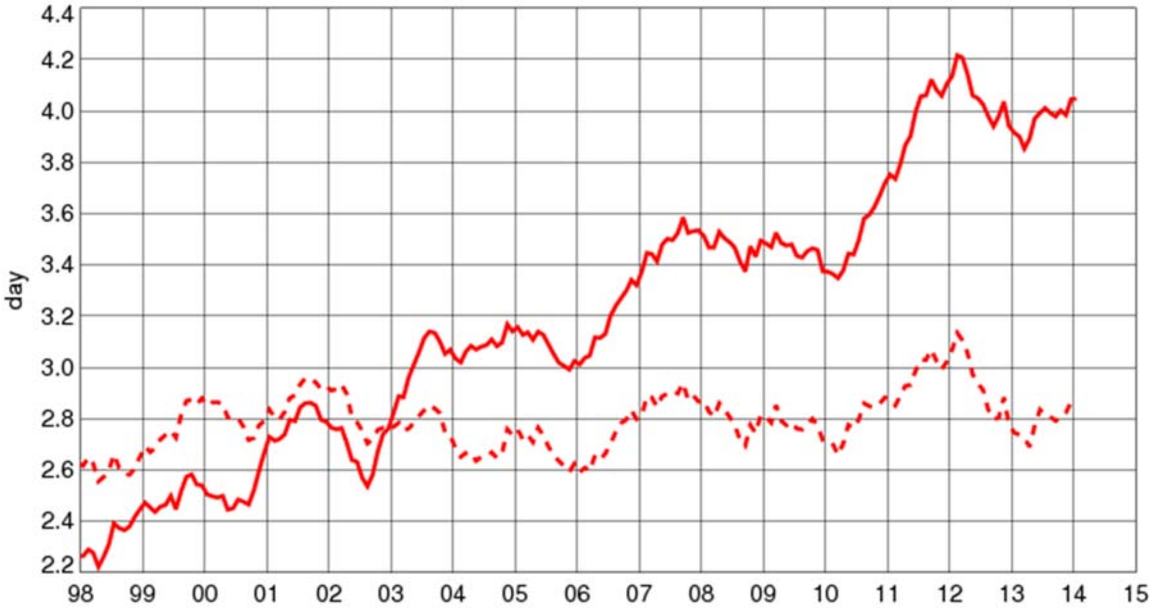


Figure 20: As Figure 19 for scores computed against radiosonde observations.



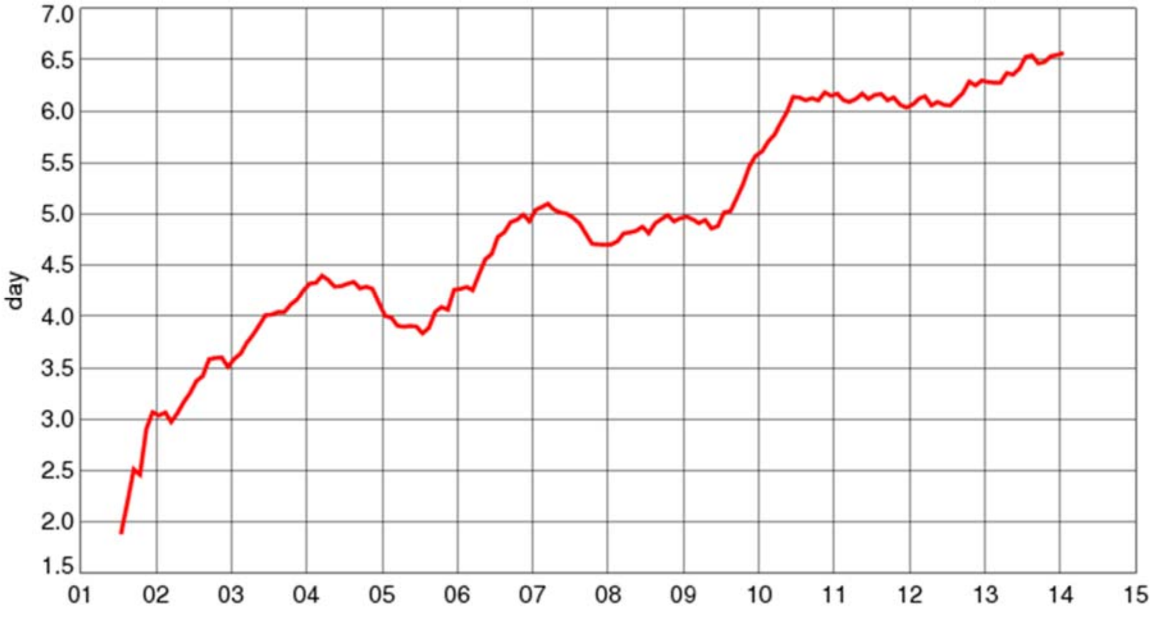
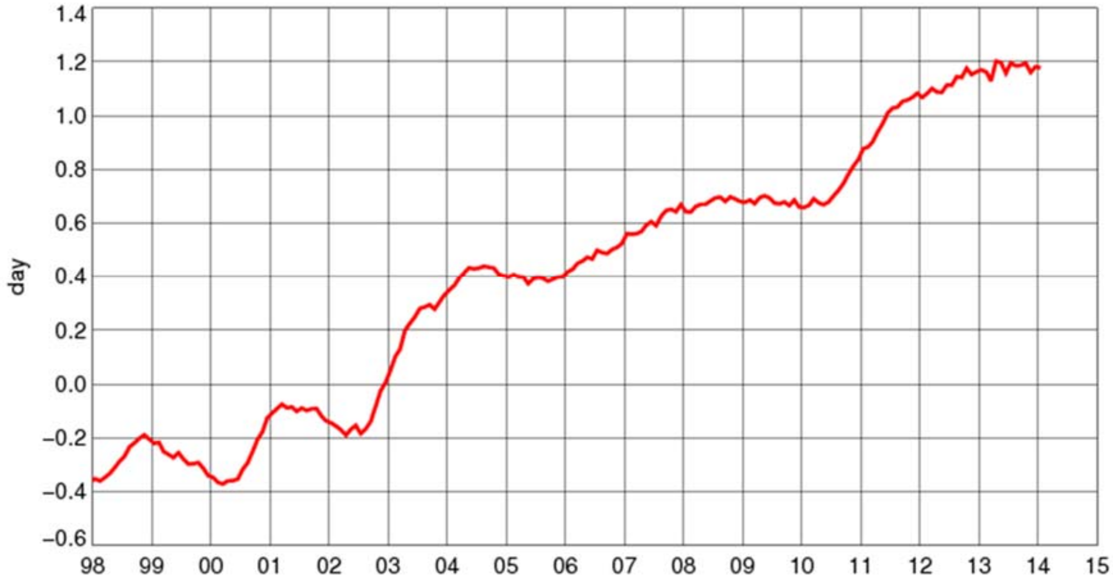


Figure 21: Supplementary headline scores for deterministic (top) and probabilistic (bottom) precipitation forecasts (continuous curves). The dashed curve shows the deterministic headline score for ERA-Interim as a reference. Each curve shows the number of days for which the centred 12-month mean skill 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.



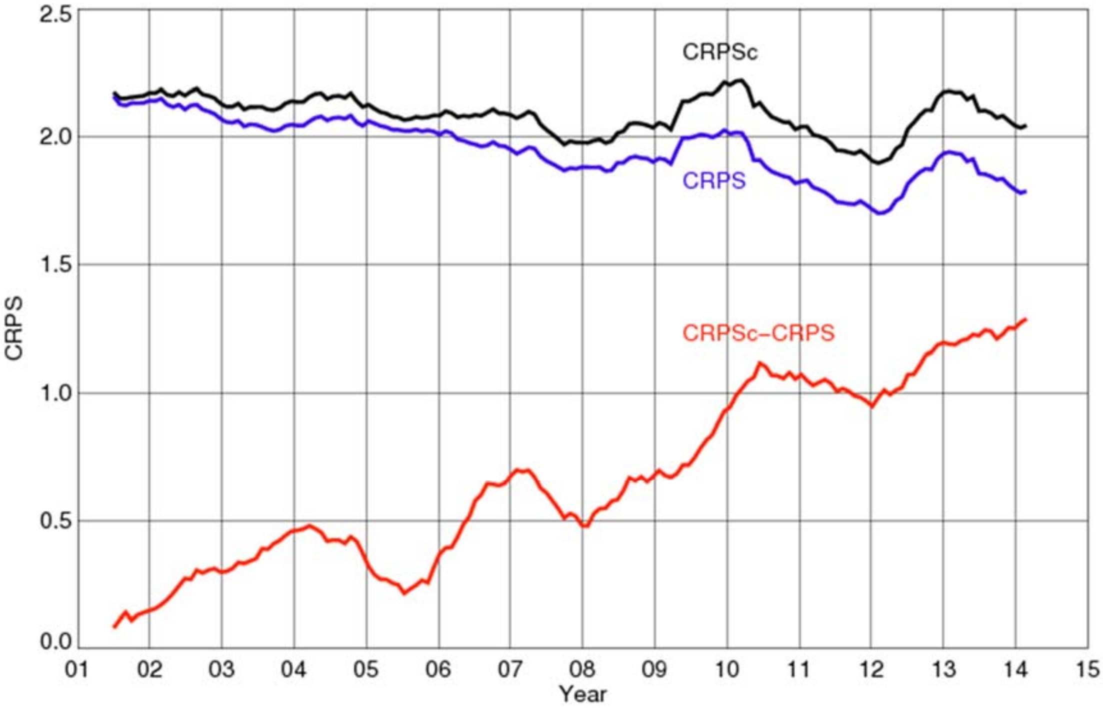
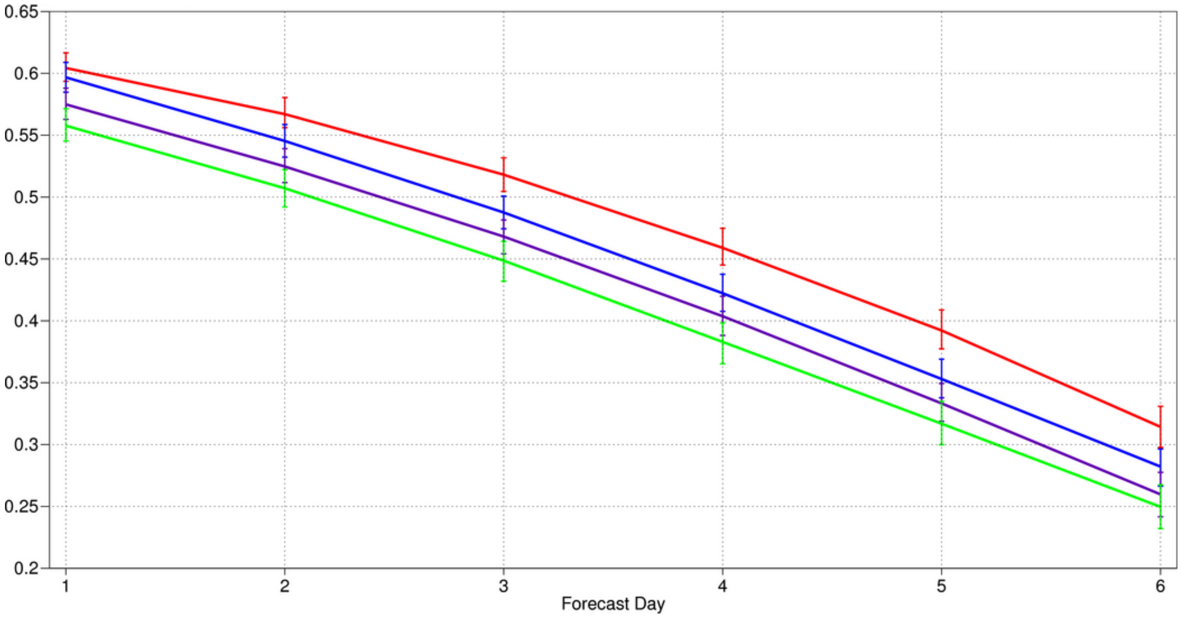


Figure 22: The top panel shows the difference between the operational forecast and ERA-Interim of the supplementary headline score for deterministic precipitation forecasts. The curve is the difference between the two curves in the upper panel of Figure 21. The lower panel shows the CRPS for probabilistic precipitation forecasts at day 6 in the extratropics in blue, the corresponding CRPS of the climate (which is used as a reference in the headline score shown in the lower panel of Figure 21) in black, and the difference in red.



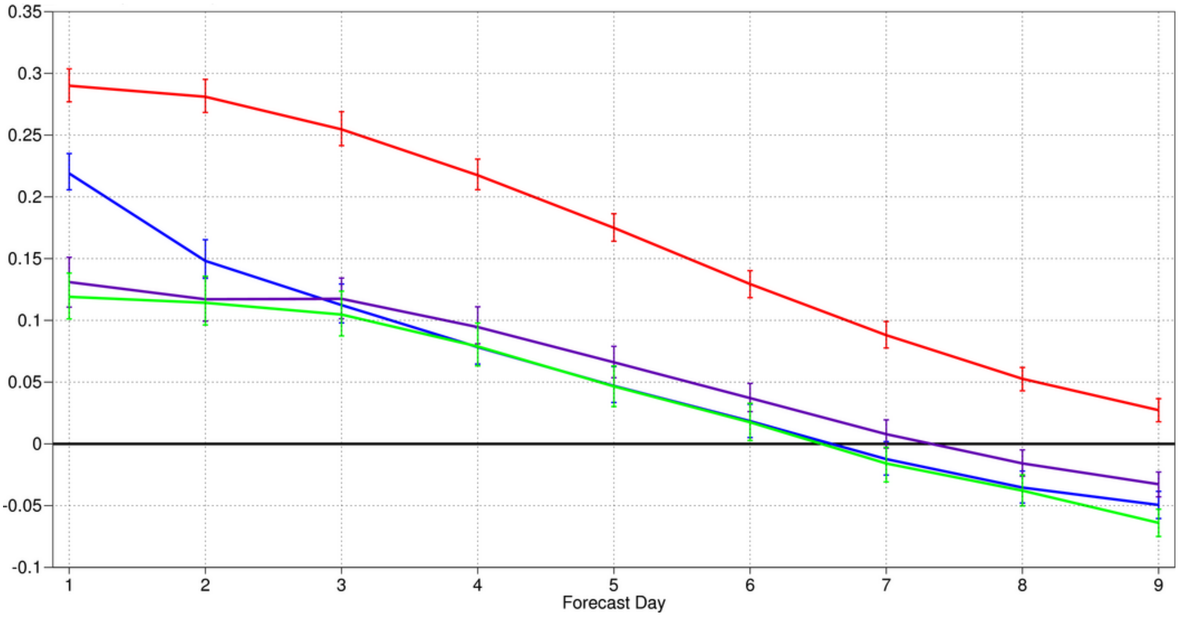


Figure 23: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2013–July 2014. Bars indicate 95% confidence intervals.

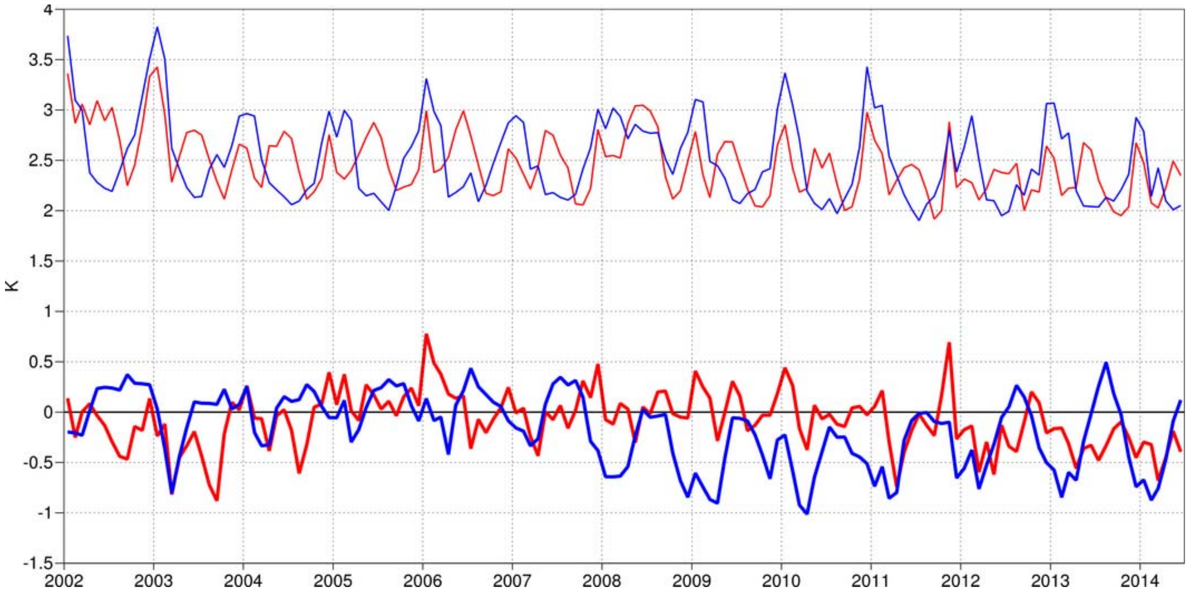


Figure 24: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

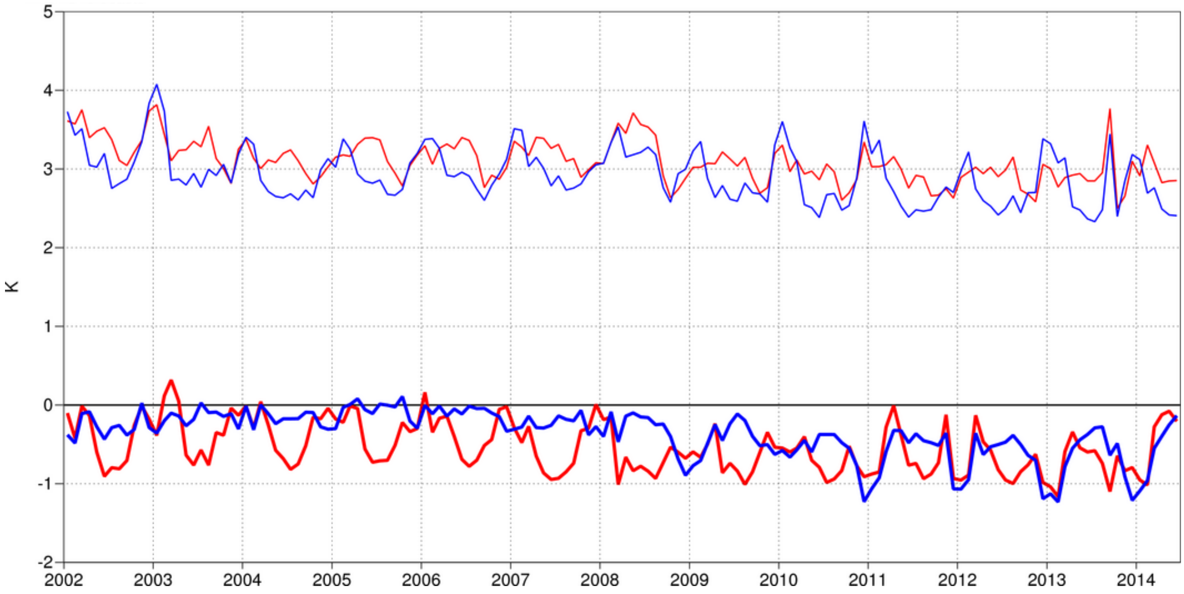


Figure 25: Verification of 2 m dewpoint forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

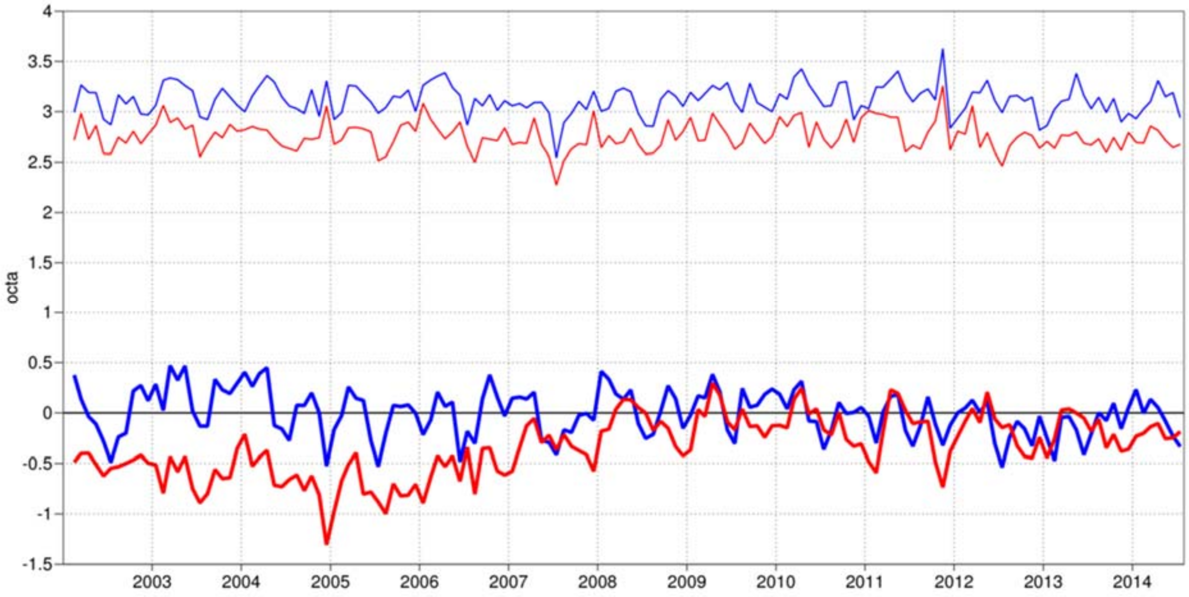


Figure 26: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error. [Note that the vertical position of the curves has changed compared to last year's report, which contained a scaling error.]

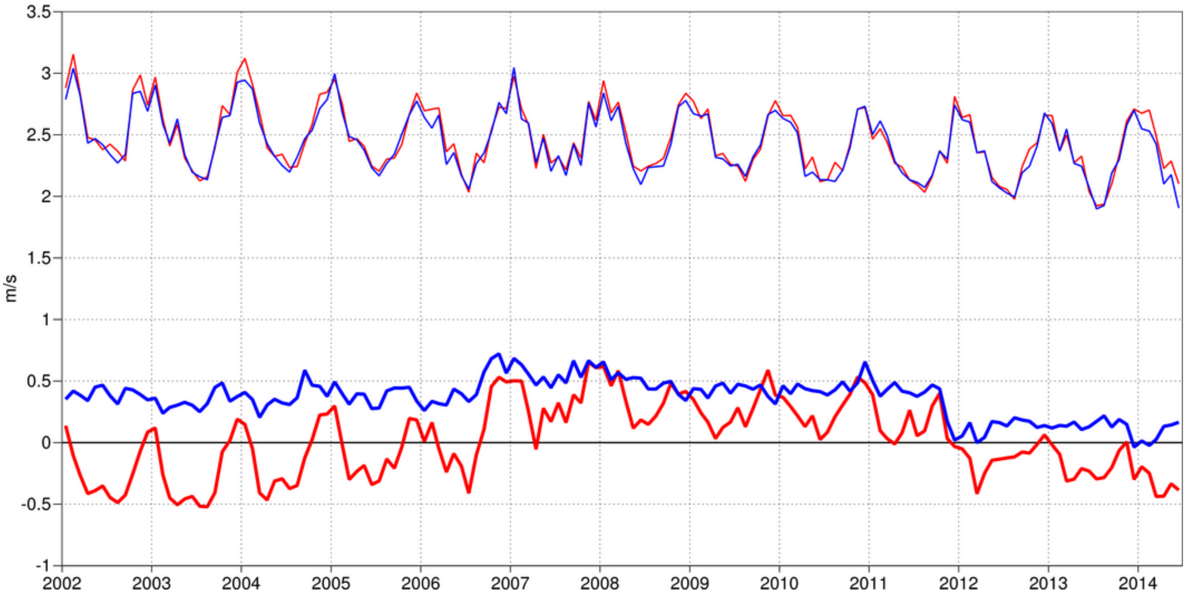


Figure 27: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

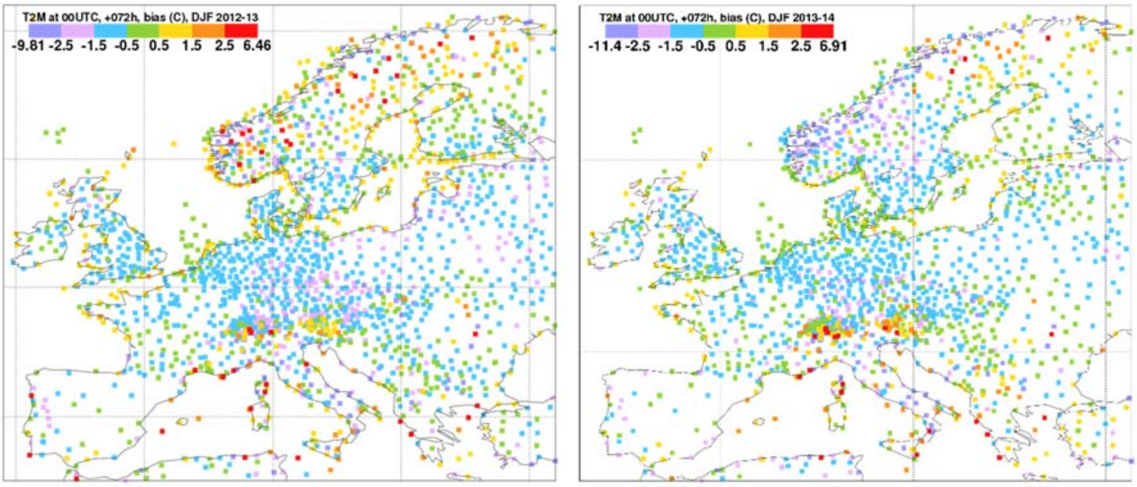

T2M at 00UTC, +072h, bias (C), DJF 2012-13
-9.81 -2.5 -1.5 -0.5 0.5 1.5 2.5 6.46
T2M at 00UTC, +072h, bias (C), DJF 2013-14
-11.4 -2.5 -1.5 -0.5 0.5 1.5 2.5 6.91

Figure 28: Night-time 2 m temperature mean errors during winters (Dec–Feb) 2012–13 and 2013–14.

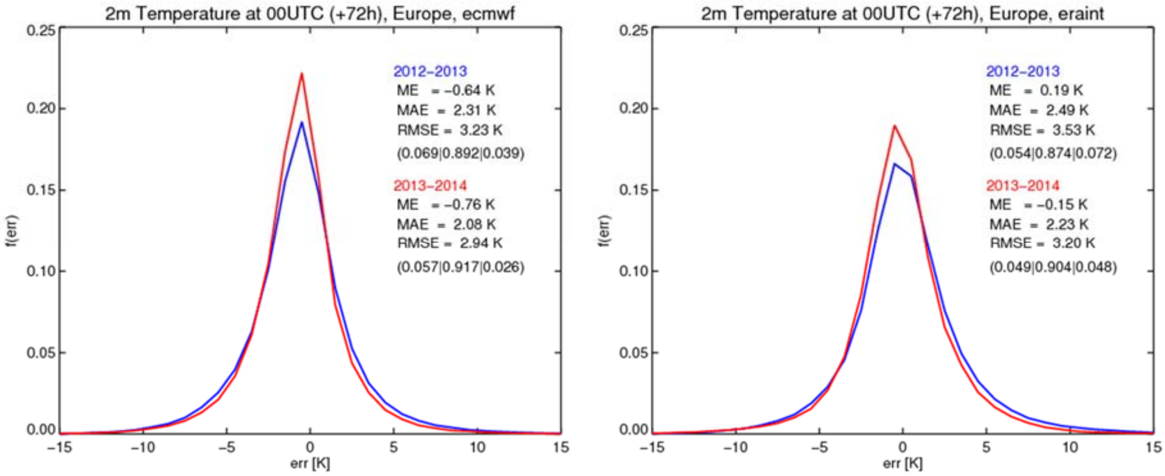


Figure 29: Error distributions for Europe, comparison of winters 2011–12 and 2012–13 for the operational run (left) and ERA-Interim (right). Also shown are mean error (ME), mean absolute error (MAE) and root mean squared error (RMSE) for the two winters. Fractions of cases with errors <-5 K, between -5 and +5 K, and >+5 K are given in parentheses.

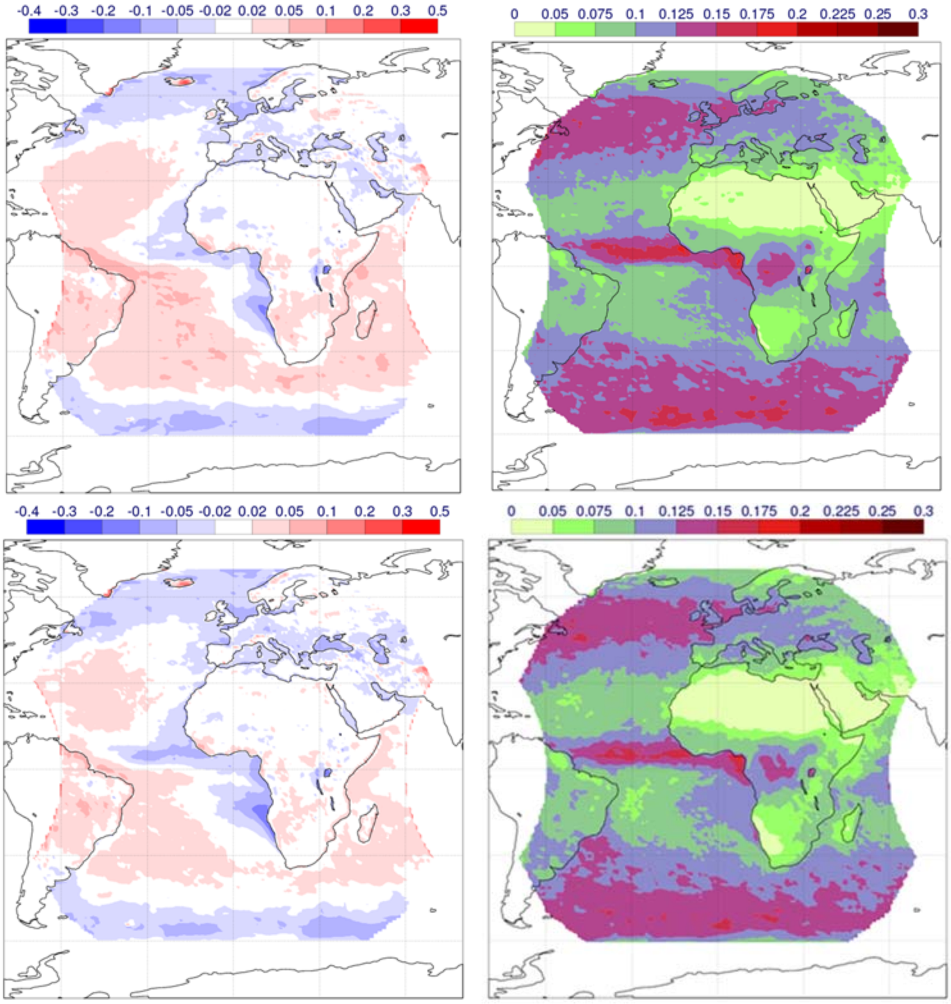


Figure 30: Mean error (left) and standard deviation of the error (right) for the high-resolution operational forecasts of daily means of normalized top of the atmosphere reflected solar radiation at forecast day 3 in the years 2012 (top) and 2013 (bottom).

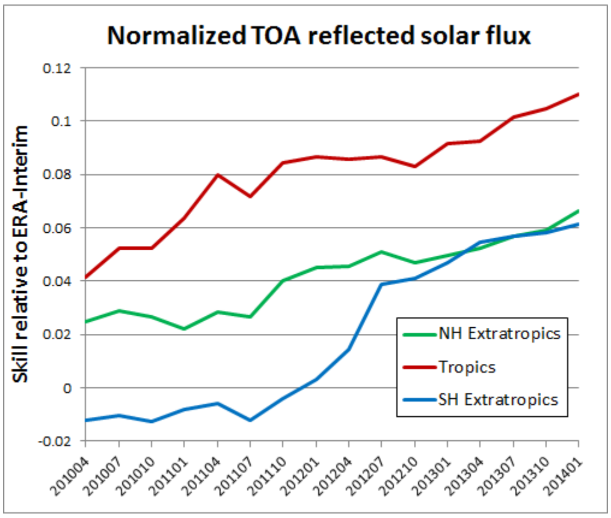


Figure 31: 12-month running average of the day 3 forecast skill relative to ERA-Interim of normalized TOA reflected solar flux (daily totals) in the parts of the northern hemisphere extratropics (green), tropics (red), and southern hemisphere extratropics (blue) which are covered by the CM-SAF product in Figure 30.

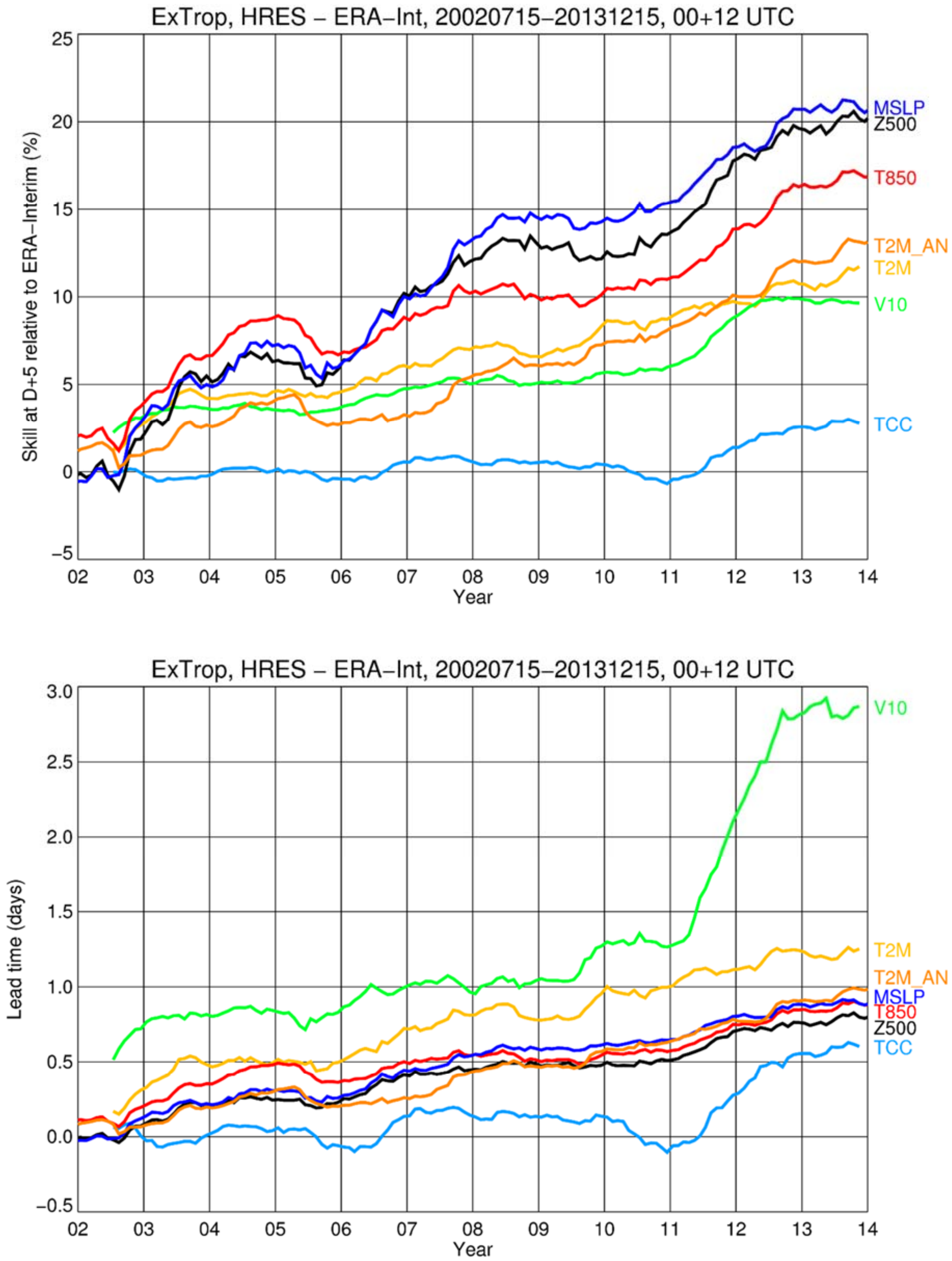
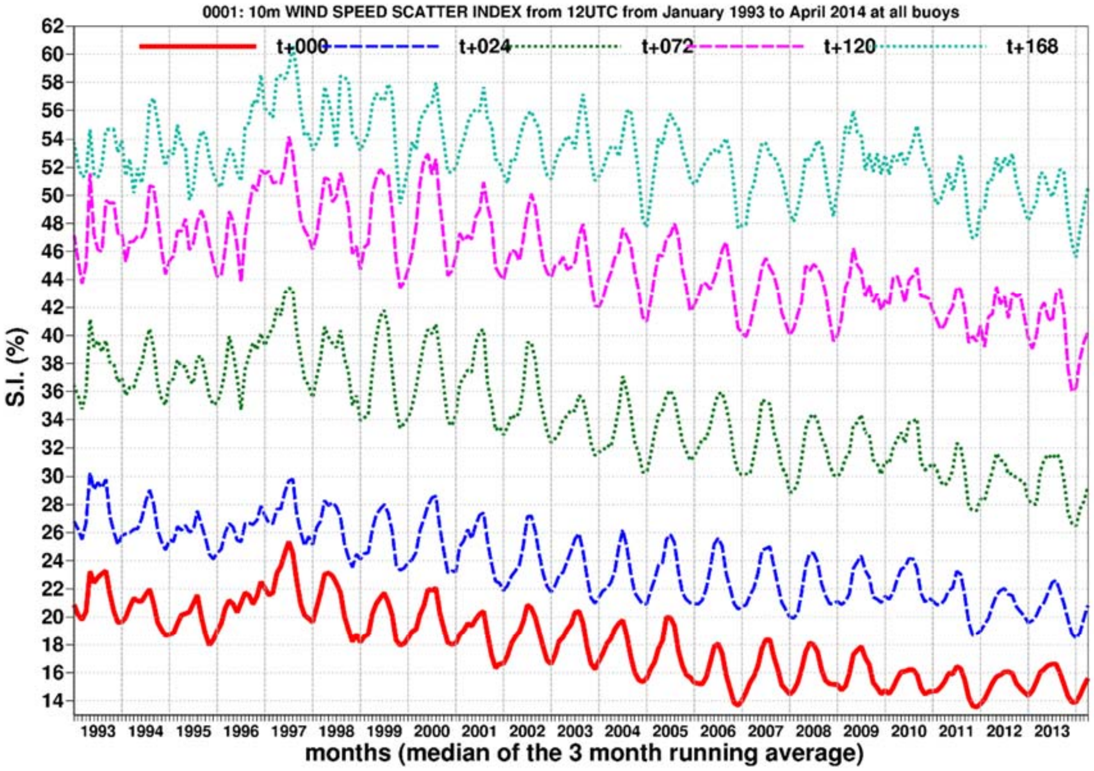


Figure 32: Evolution of skill of the HRES forecast relative to ERA-Interim expressed as relative skill at forecast day 5 (top), and in terms of increase of lead time for the forecast skill which was reached at day 5 in 2002 (bottom). Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), mean sea level pressure (MSLP) and 2 m temperature (T2M_AN), using RMSE as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC), using error standard deviation as a metric.



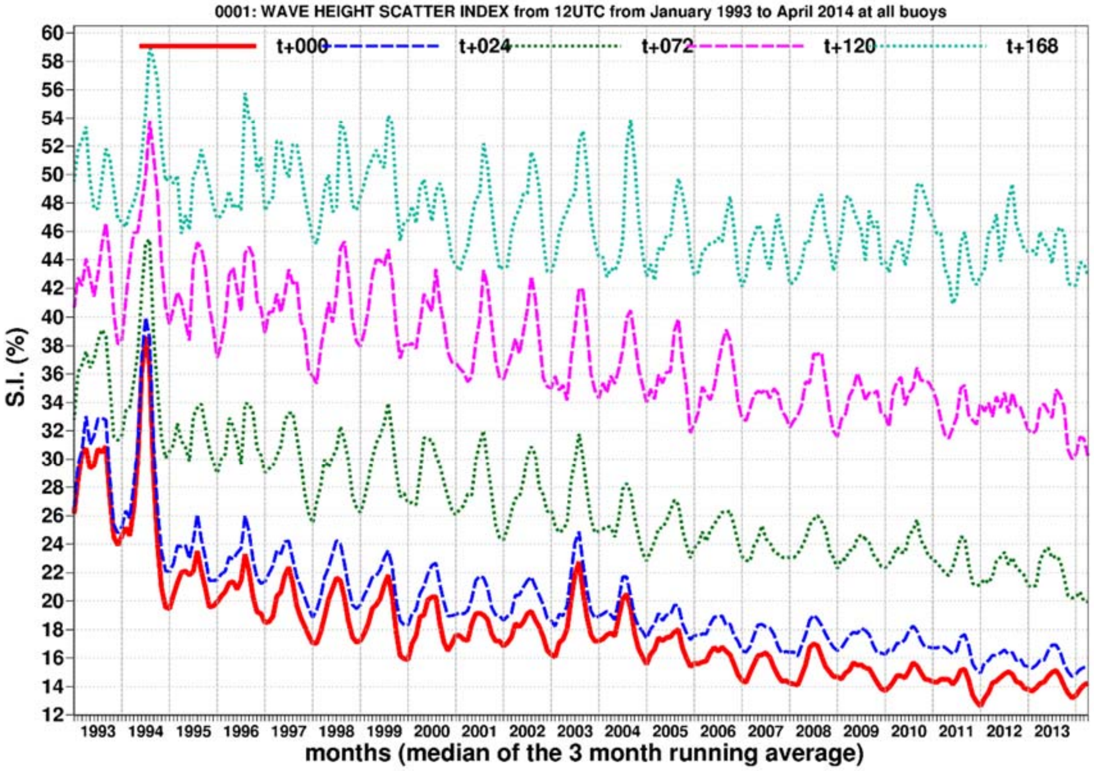
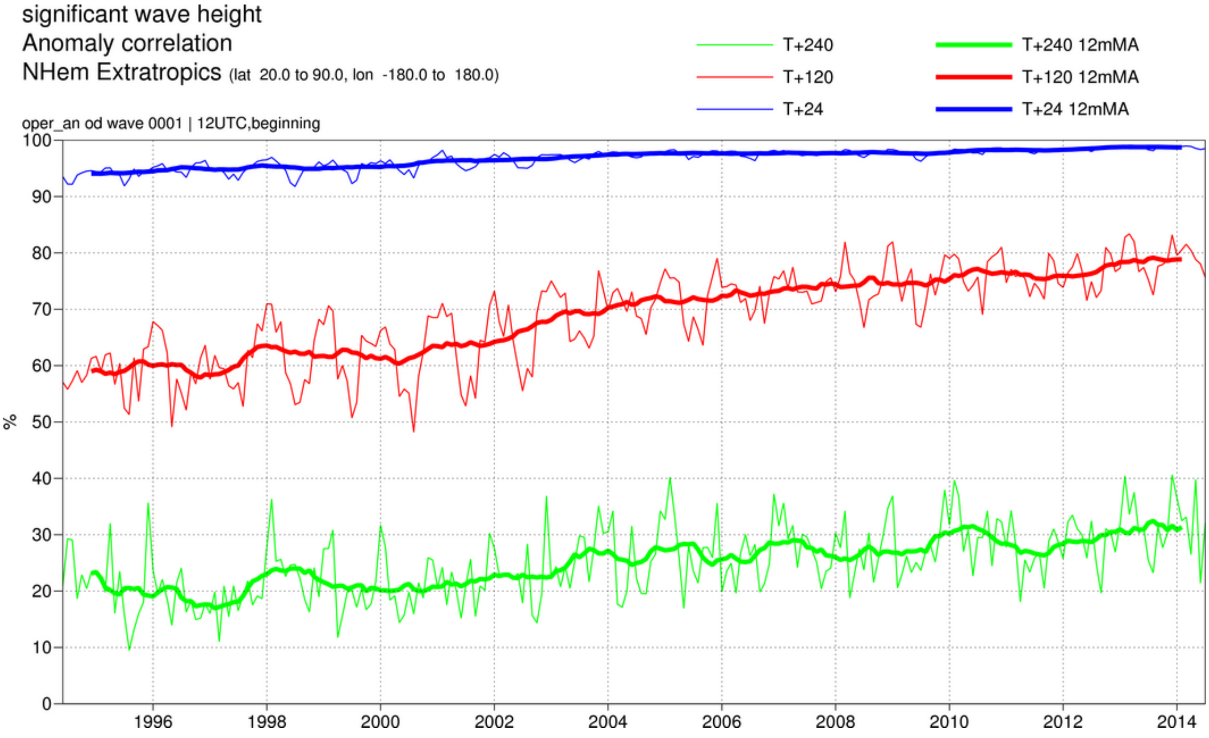


Figure 33: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.



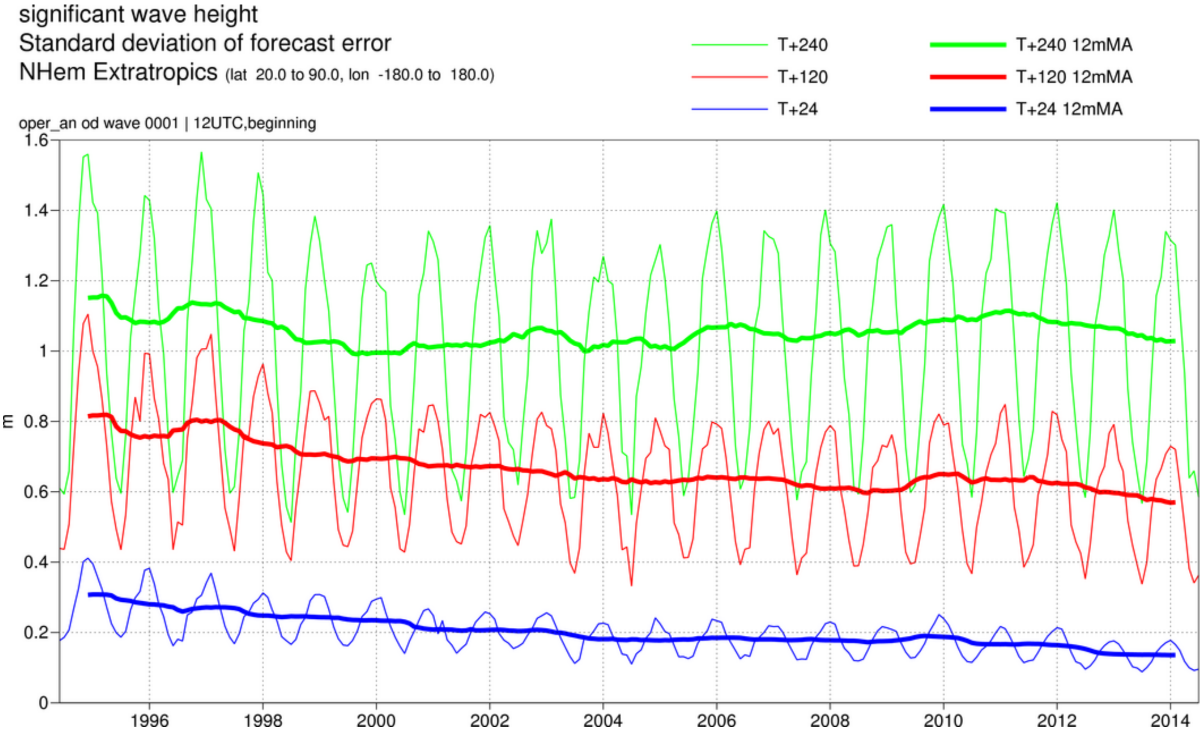
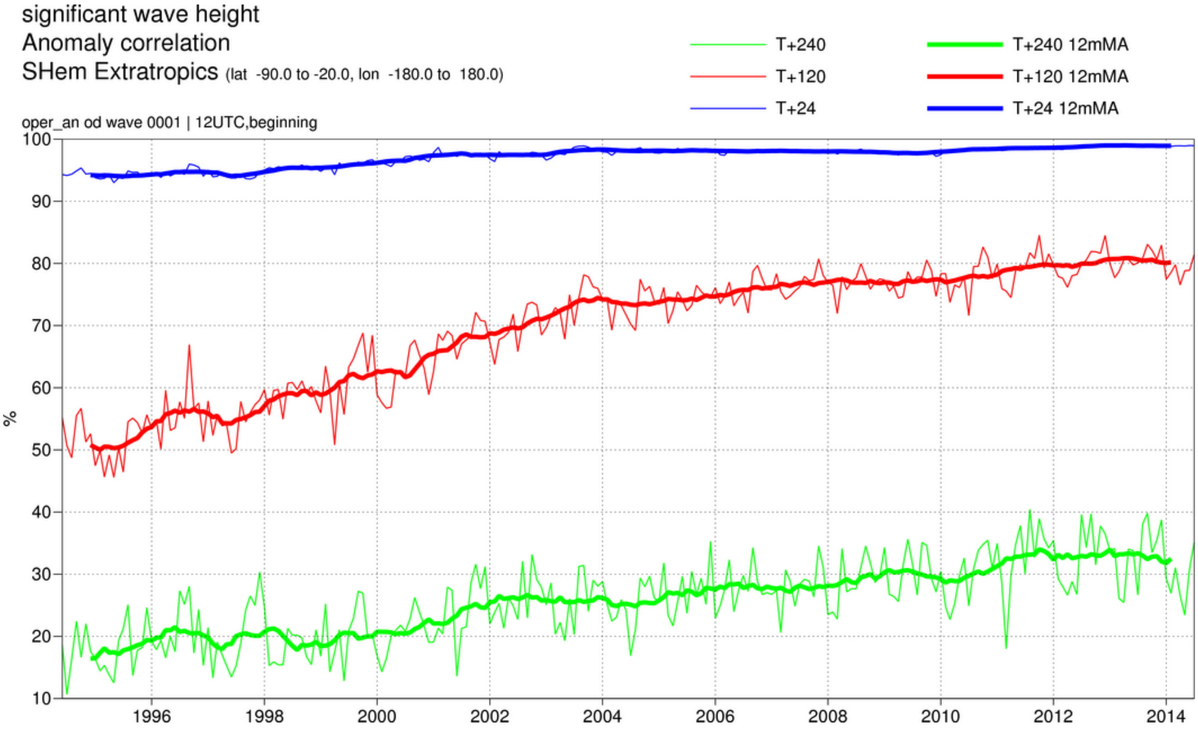


Figure 34: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC (top) and error standard deviation (bottom) for ocean wave heights verified against analysis for the northern extratropics at day 1 (blue), 5 (red) and 10 (green).



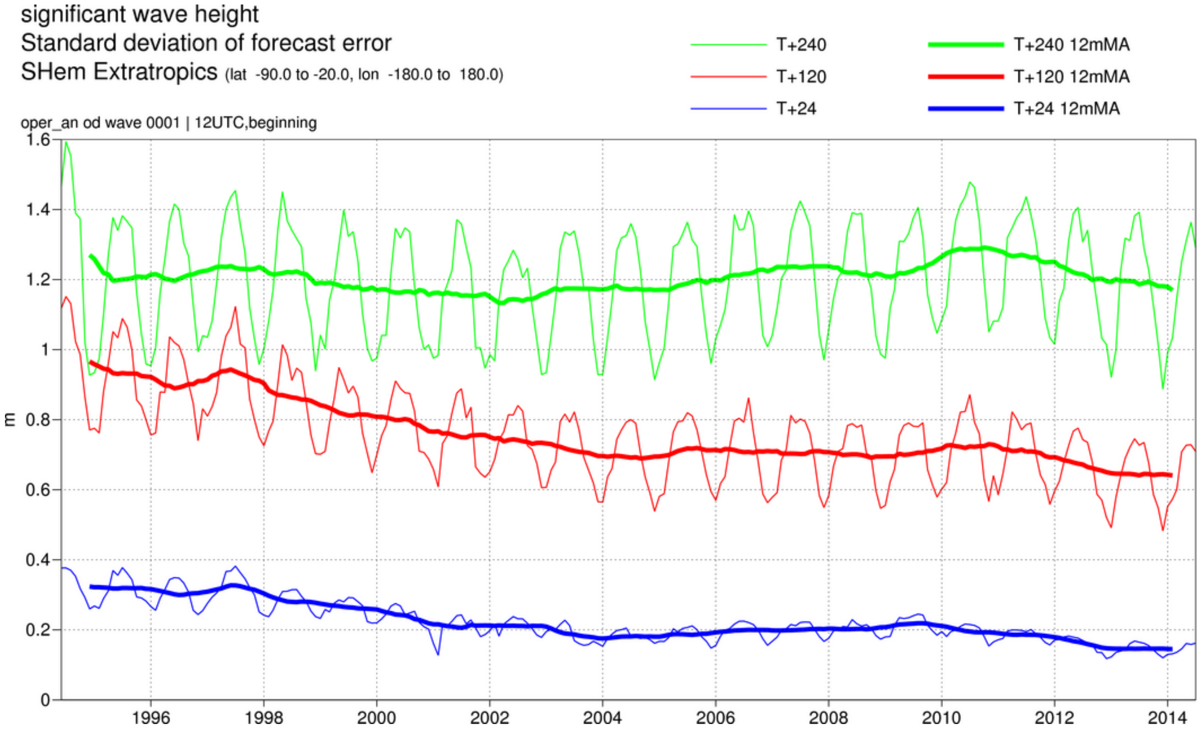


Figure 35: As Figure 34 for the southern hemisphere.

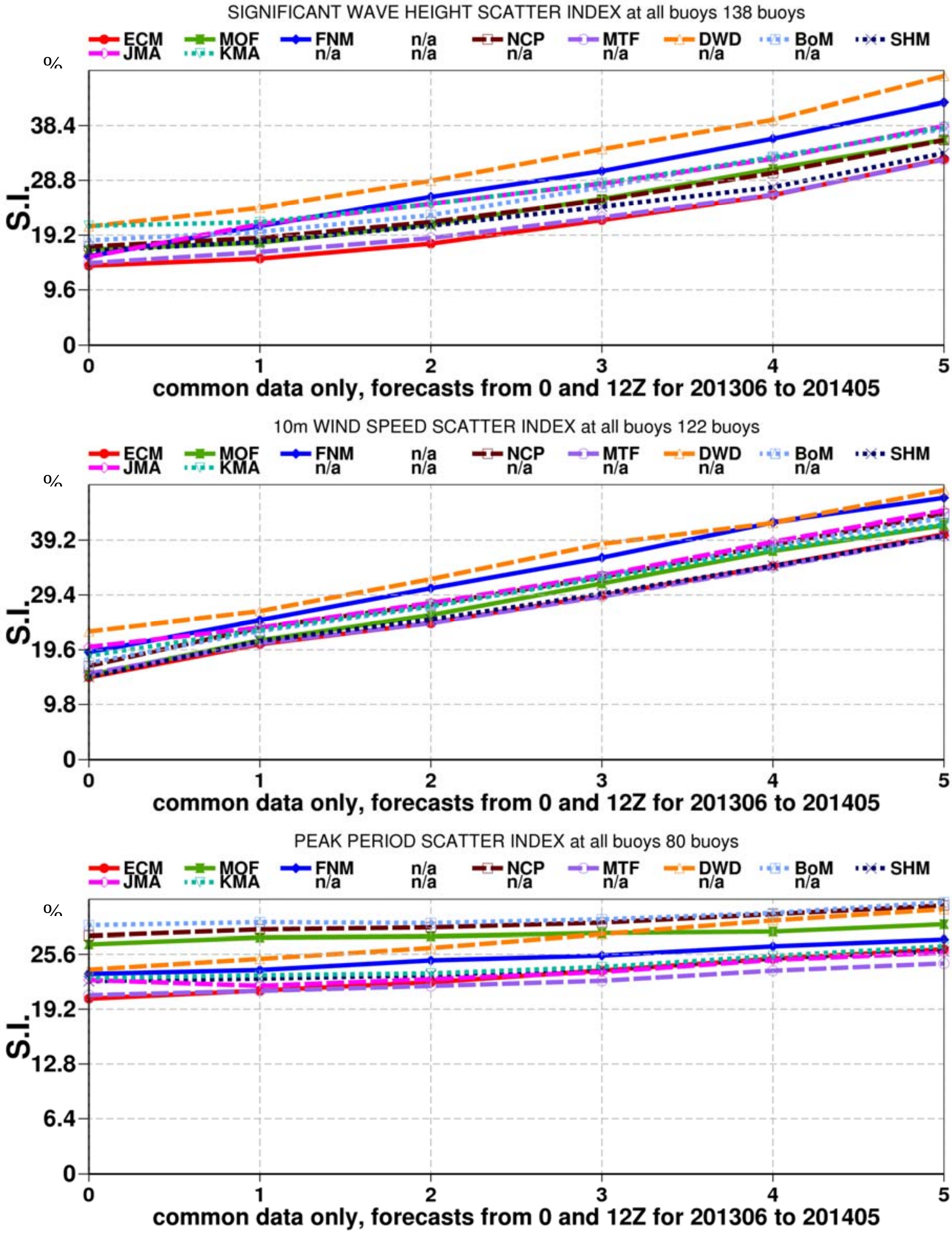


Figure 36: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period June 2013 – May 2014. The x-axis shows the forecast range in days from analysis (step 0) to day 5. MOF: 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; KMA: Korea Meteorological Administration.

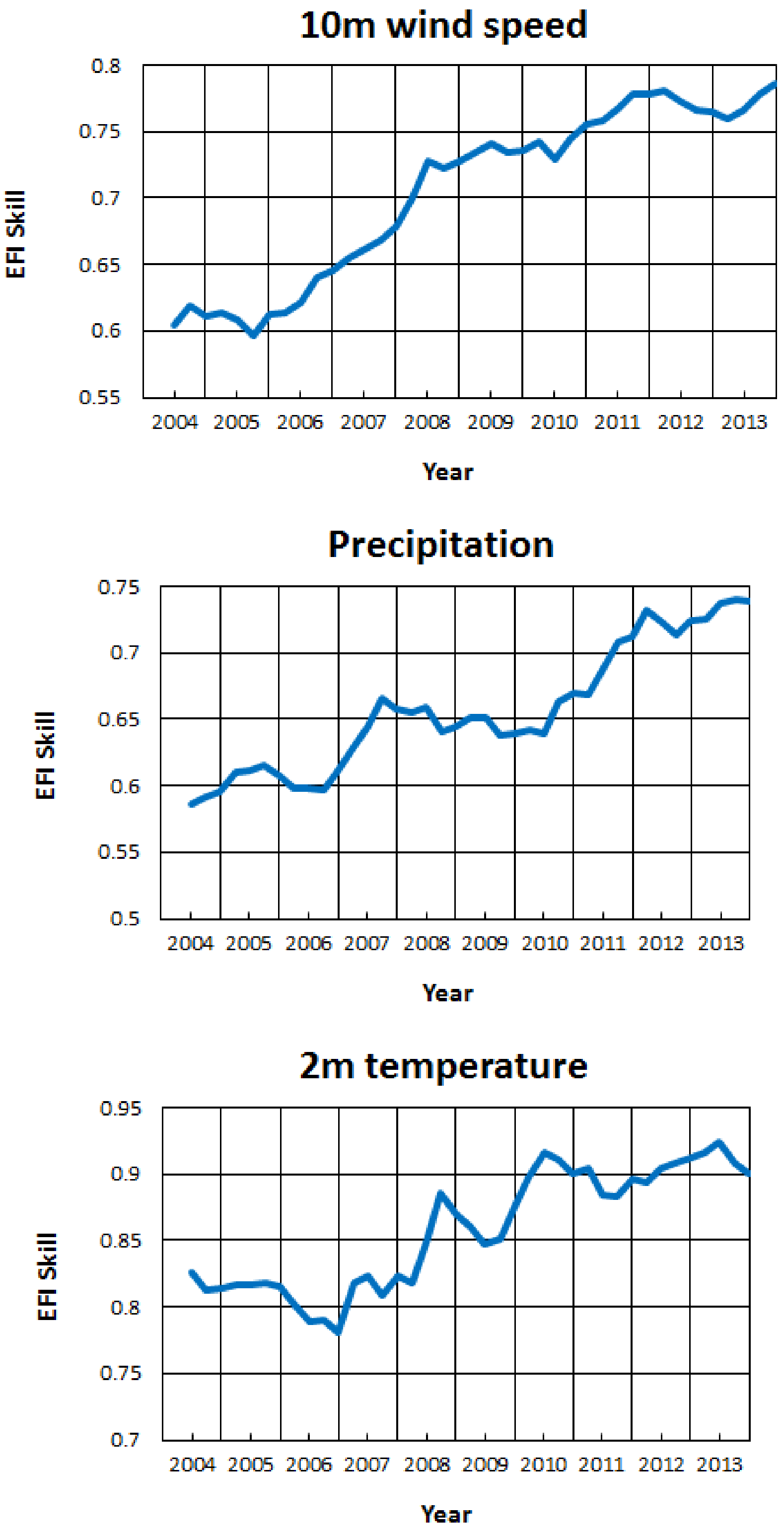


Figure 37: Verification of Extreme Forecast Index (EFI). Top panel: supplementary headline score – skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate, curves show a four-season running mean of relative operating characteristic (ROC) area skill scores (final point includes spring (March–May) 2014). Centre and bottom panels show the equivalent ROC area skill scores for precipitation EFI forecasts and for 2 m temperature EFI forecasts.

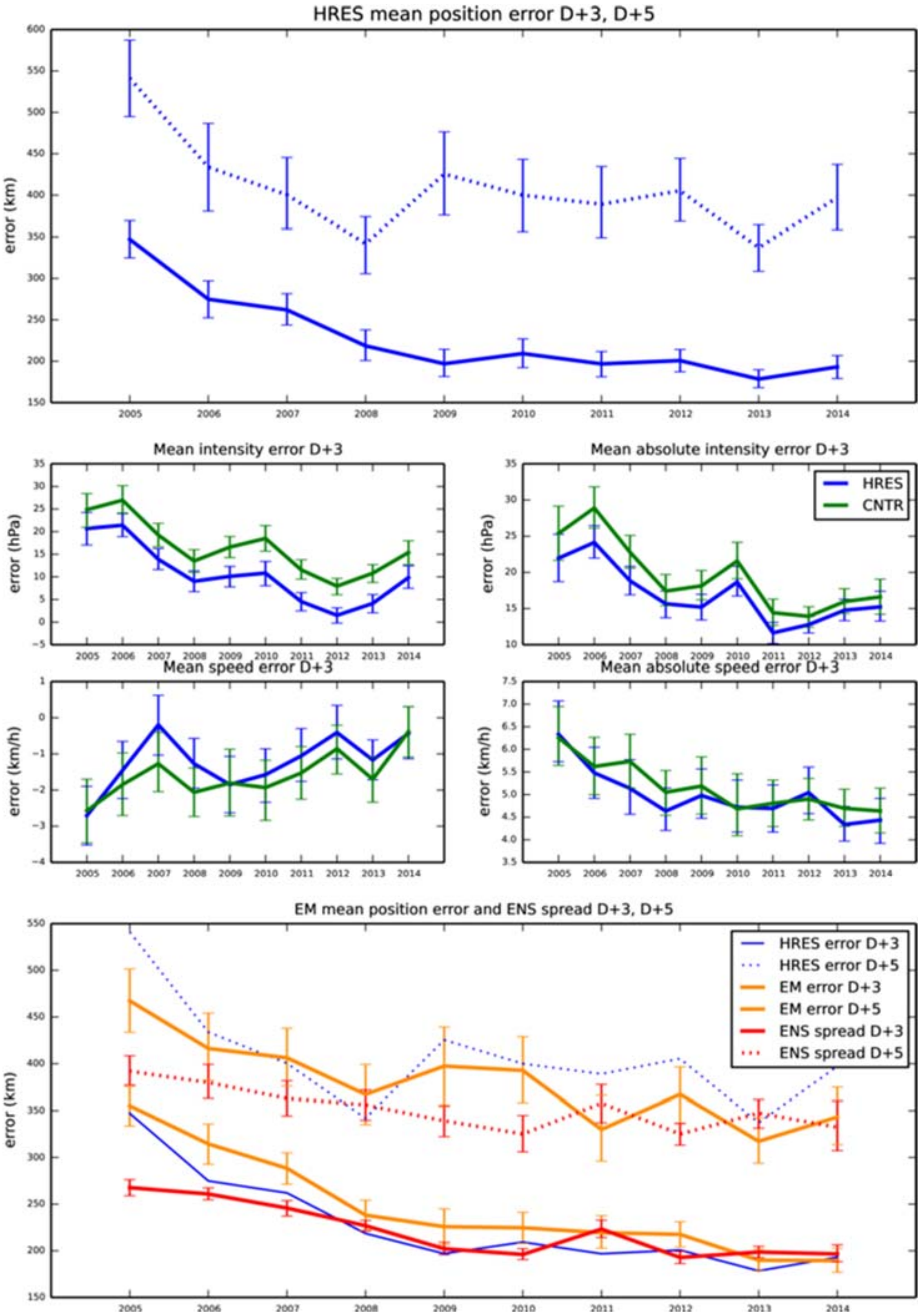


Figure 38: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported via the GTS. Top panel supplementary headline score – the mean position error (km) of the three-day high-resolution forecast. The error for day 5 is included for comparison. Centre four panels show 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), mean absolute error of the intensity and mean and absolute error of cyclone motion speed for cyclone forecast both by HRES and ENS control. Bottom panel shows mean position error of ensemble mean (mean of cyclones forecast by ensemble members) with respect to the observed cyclone (cyan curve) and ensemble spread (mean of distances of ensemble cyclones from the ensemble mean; red curve); for comparison the HRES position error (from the top panel) is plotted as well (blue curve).

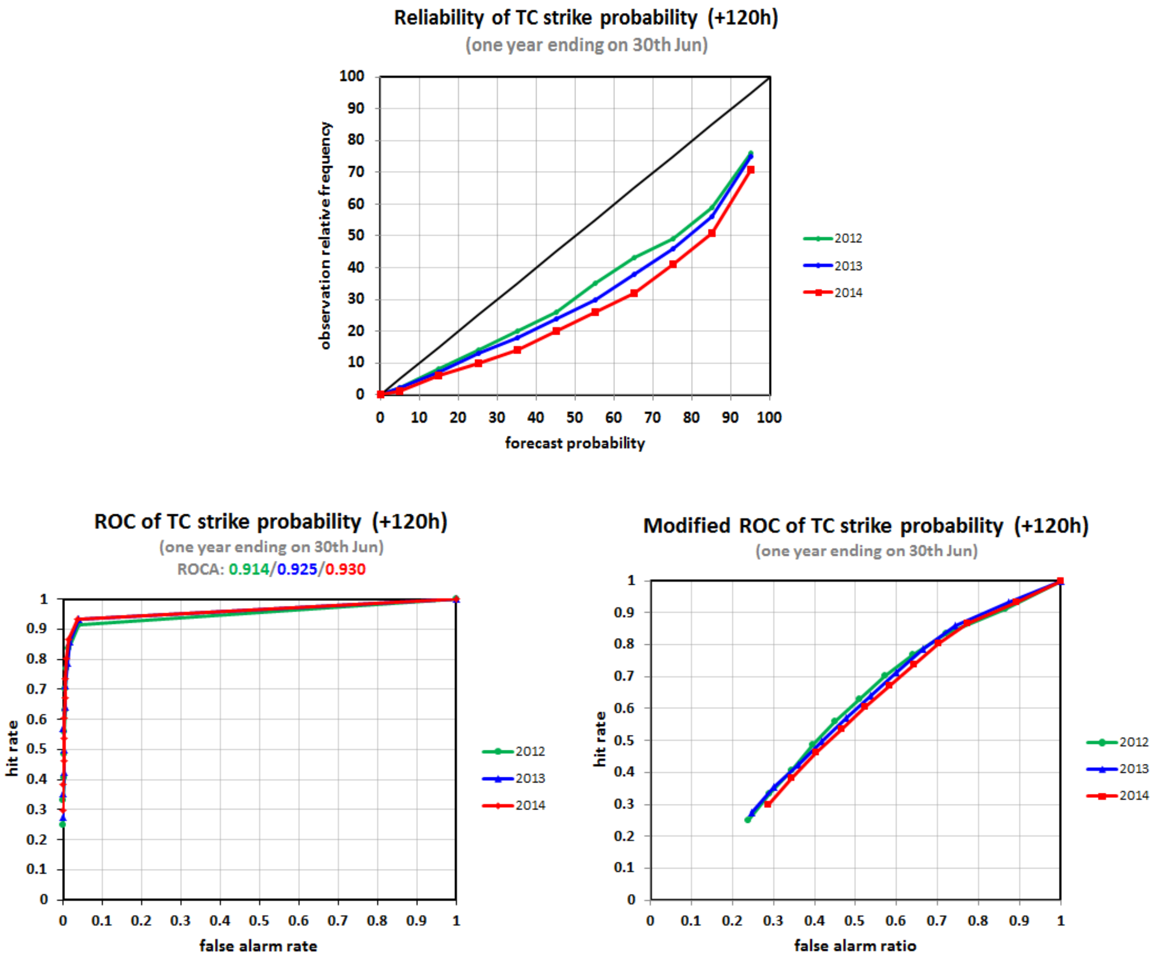
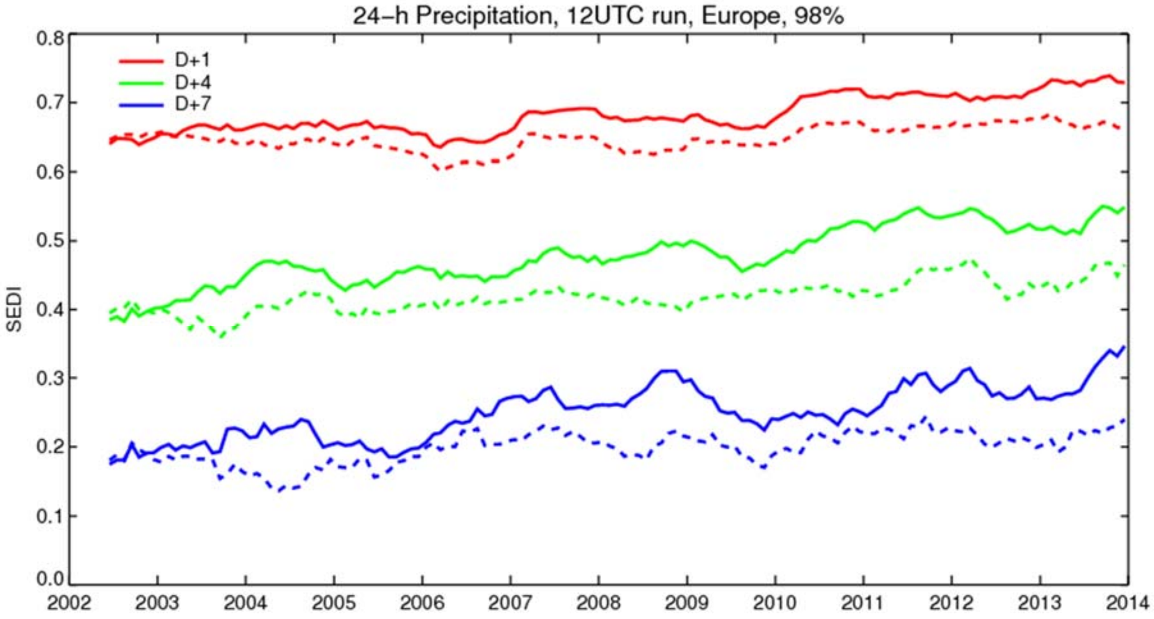


Figure 39: Probabilistic verification of ensemble tropical cyclone forecasts for three 12-month periods: July 2011–June 2012 (green), July 2012–June 2013 (blue) and July 2013–June 2014 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) 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).



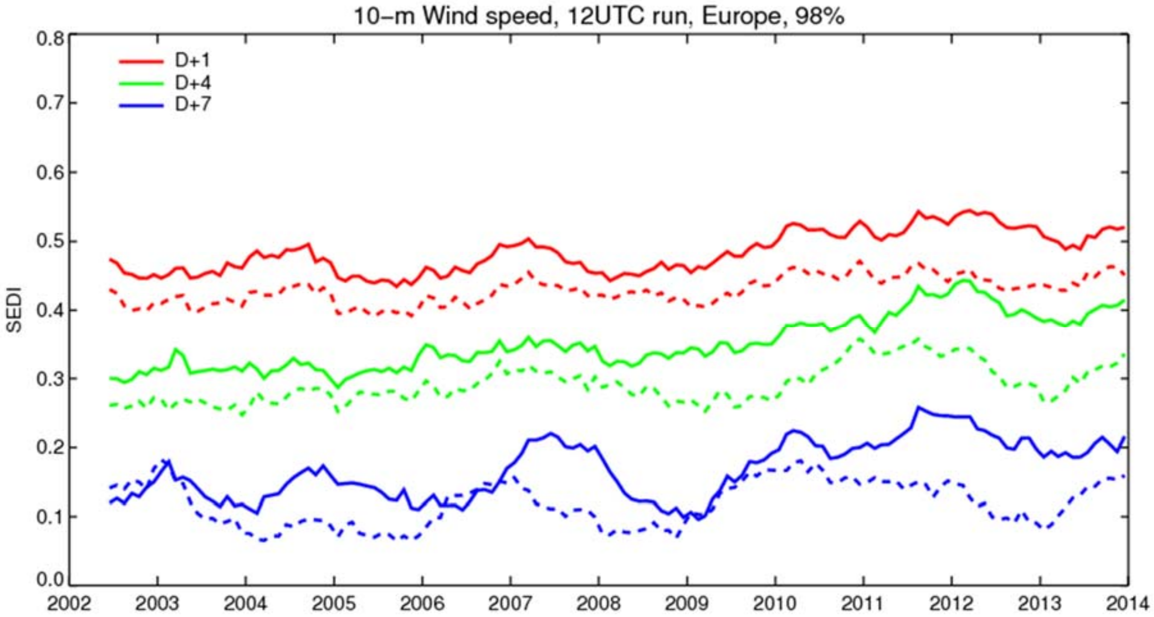


Figure 40: Skill of the HRES forecast (continuous) and ERA-Interim (dashed) in predicting 24-h precipitation amounts (top) and 10 m wind speeds (bottom) above the 98th climate percentile in Europe as measured by the SEDI score for forecast days 1, 4, and 7. Curves show 12-month running averages.

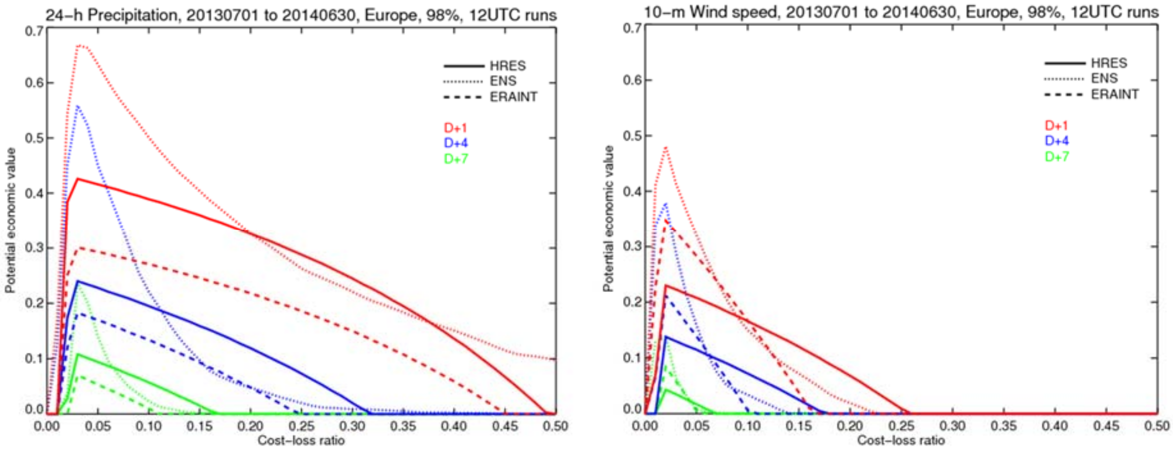


Figure 41: Potential economic value of the HRES forecast (continuous), the ensemble forecast (dotted), and ERA-Interim (dashed) in predicting 24-h precipitation amounts (left panel) and 10 m wind speeds (right panel) above the 98th climate percentile in Europe in the period July 2013 – June 2014. Colours indicate forecast days 1 (red), 3 (green), 5 (blue), and 7 (yellow). Cost–loss ratios are typically in the range 0.01–0.2.

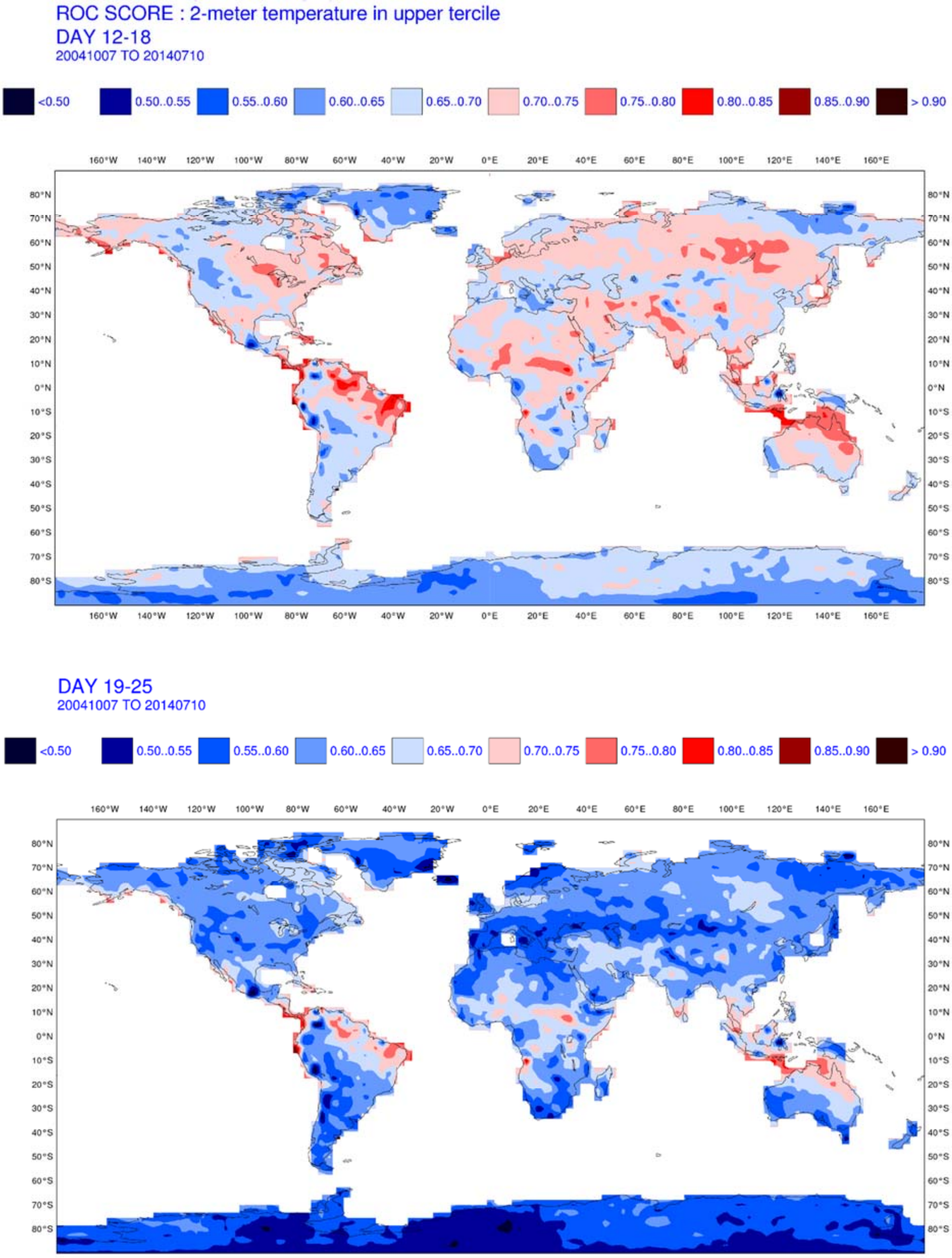

ROC SCORE : 2-meter temperature in upper tercile
DAY 12-18
20041007 TO 20140710
<0.50
0.50..0.55
0.55..0.60
0.60..0.65
0.65..0.70
0.70..0.75
0.75..0.80
0.80..0.85
0.85..0.90
> 0.90
160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°E 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E
80°N 70°N 60°N 50°N 40°N 30°N 20°N 10°N 0°N 10°S 20°S 30°S 40°S 50°S 60°S 70°S 80°S
DAY 19-25
20041007 TO 20140710
<0.50
0.50..0.55
0.55..0.60
0.60..0.65
0.65..0.70
0.70..0.75
0.75..0.80
0.80..0.85
0.85..0.90
> 0.90
160°W 140°W 120°W 100°W 80°W 60°W 40°W 20°W 0°E 20°E 40°E 60°E 80°E 100°E 120°E 140°E 160°E
80°N 70°N 60°N 50°N 40°N 30°N 20°N 10°N 0°N 10°S 20°S 30°S 40°S 50°S 60°S 70°S 80°S

Figure 42: 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 18 July 2013 for two seven-day forecast ranges: days 12–18 (top) and days 19–25 (bottom). Stronger red shading indicates higher skill compared to climate.

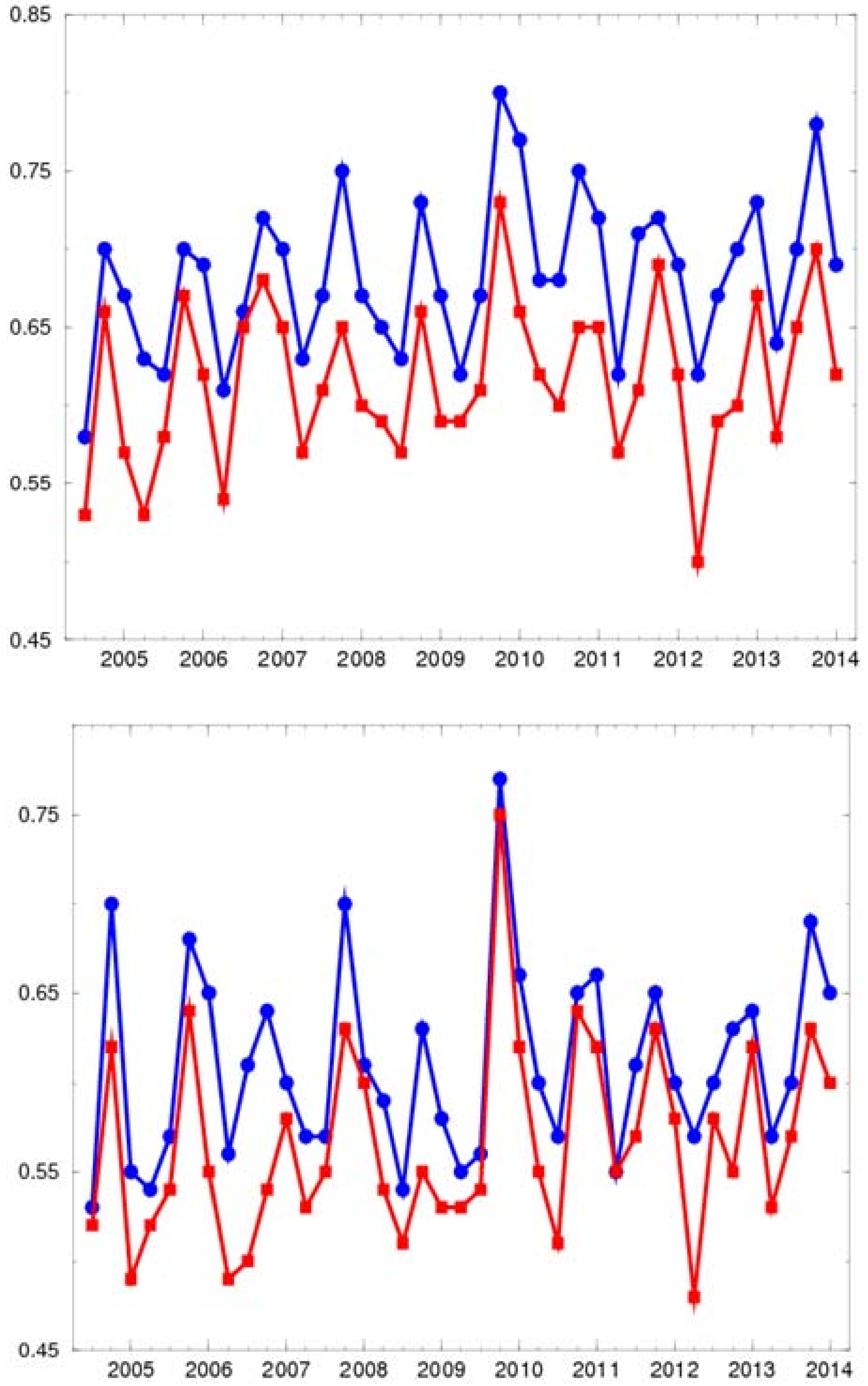


Figure 43: Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution. Scores are calculated for each three-month season since autumn (September–November) 2004 for all land points in the extratropical northern hemisphere. The blue line shows the 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 red 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 2014.

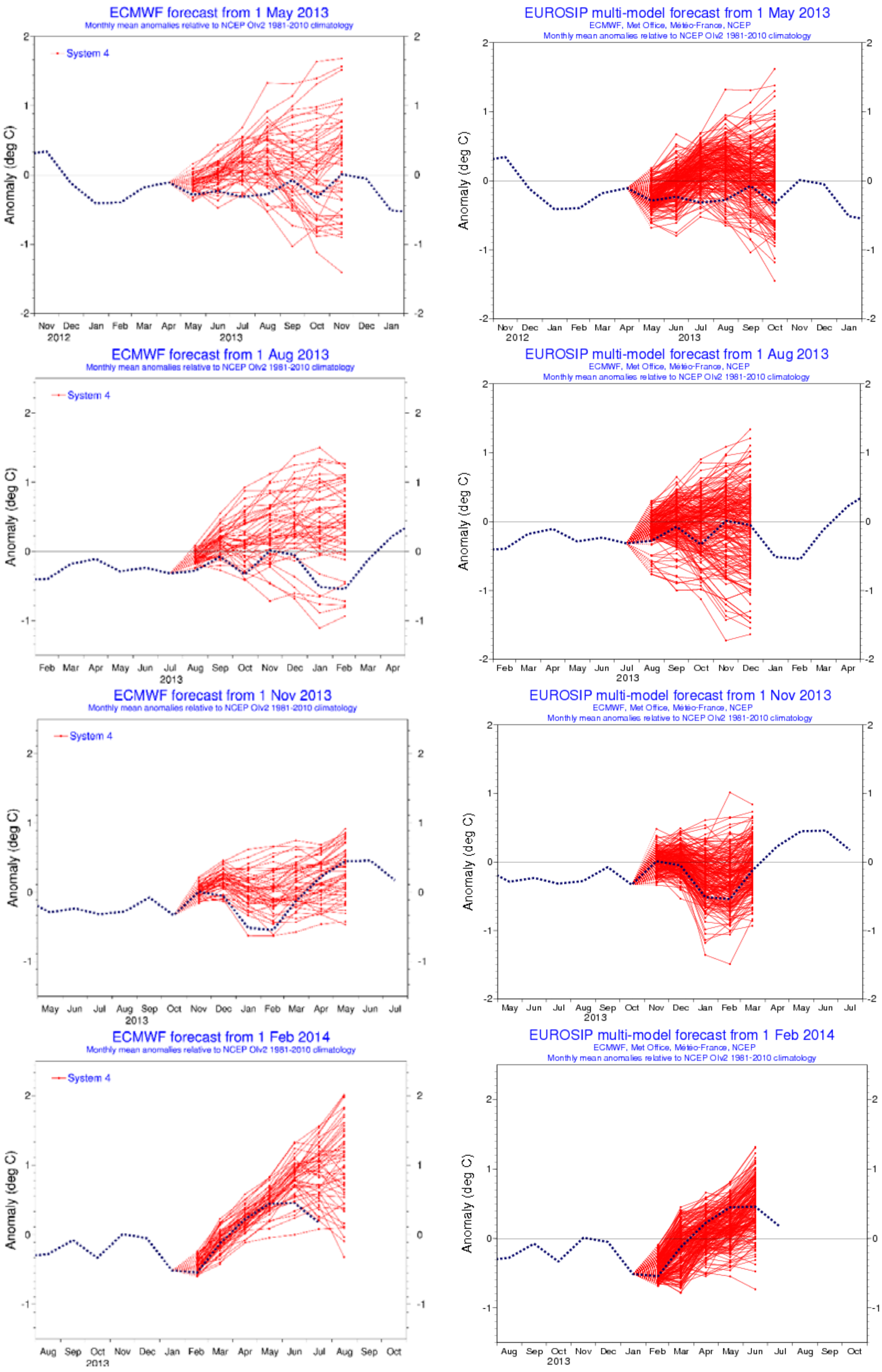


Figure 44: ECMWF (left column) and EUROSIP multi-model forecast (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2013, August 2013, November 2013 and February 2014. The red lines represent the ensemble members; dashed blue lines show the subsequent verification.

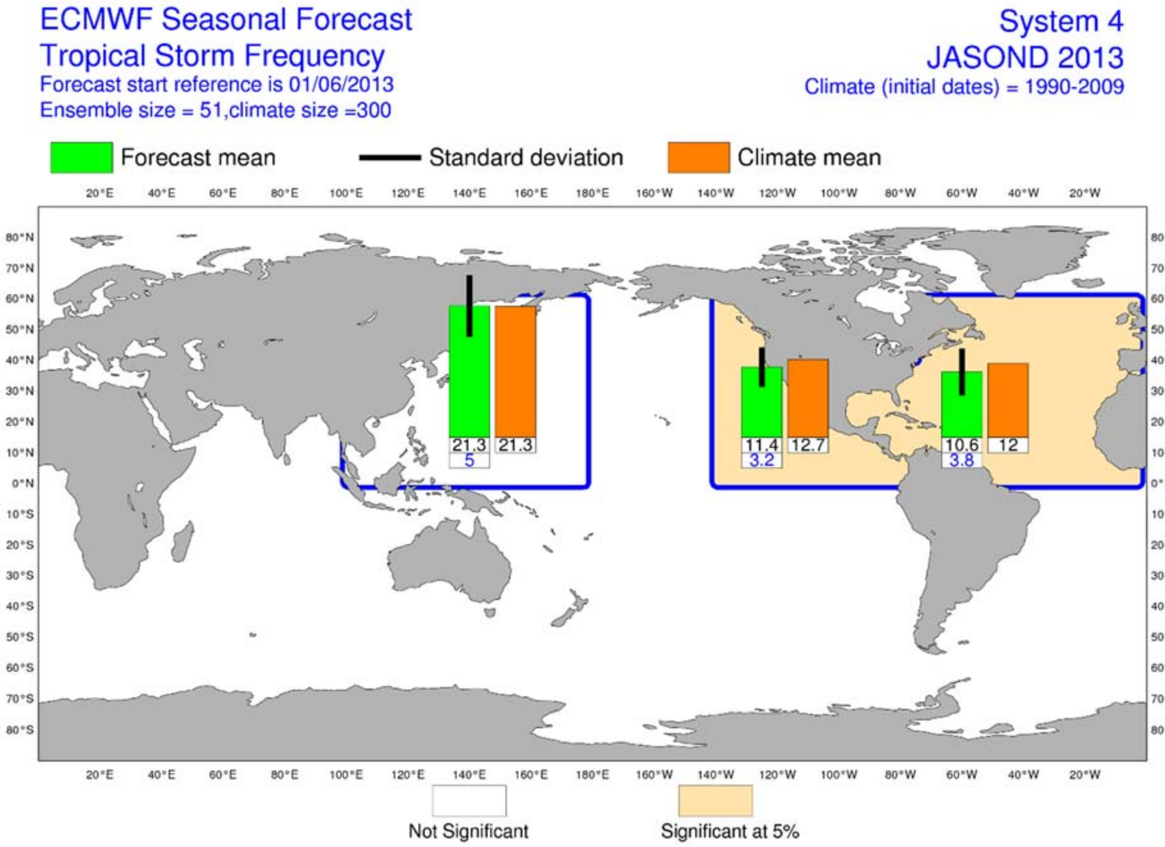


Figure 45: Tropical storm frequency forecast issued in June 2013 for the six-month period July–December 2013. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The 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.

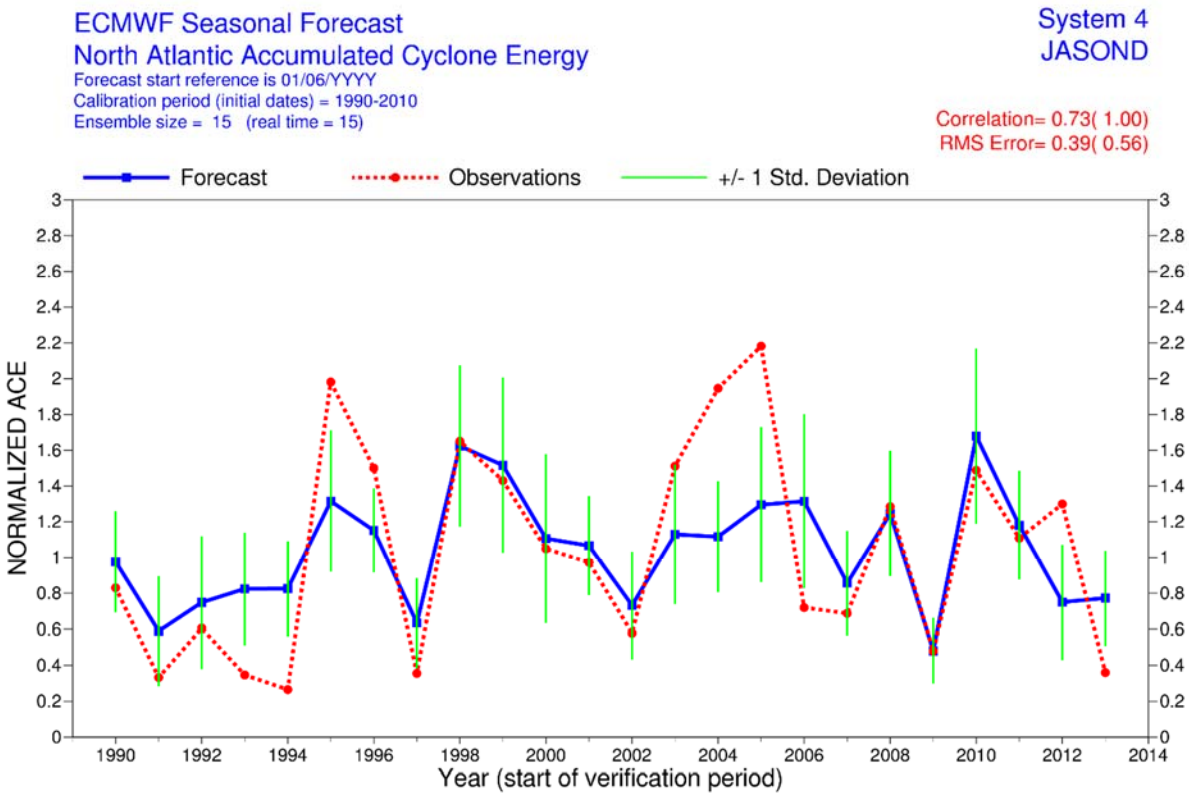


Figure 46: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2013. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts are from System 4 of the seasonal component of the IFS: these are based on the 15-member re-forecasts; from 2011 onwards they are from the operational 51-member seasonal forecast ensemble. Start date of the forecast is 1 June.

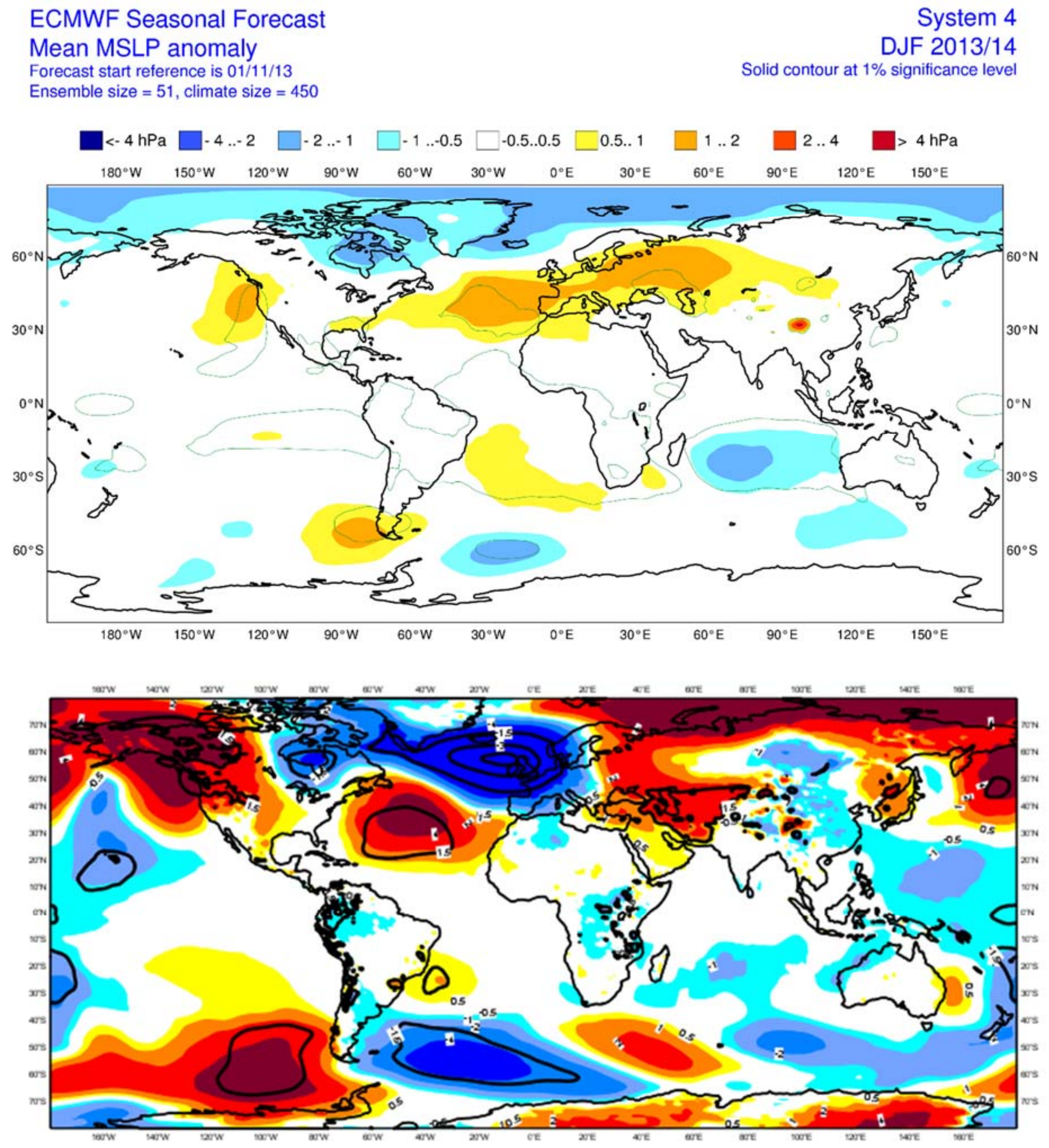


Figure 47: Anomaly of mean sea level pressure as predicted by the seasonal forecast from Nov 2013 for DJF 2013-14 (upper panel), and verifying analysis (lower panel).

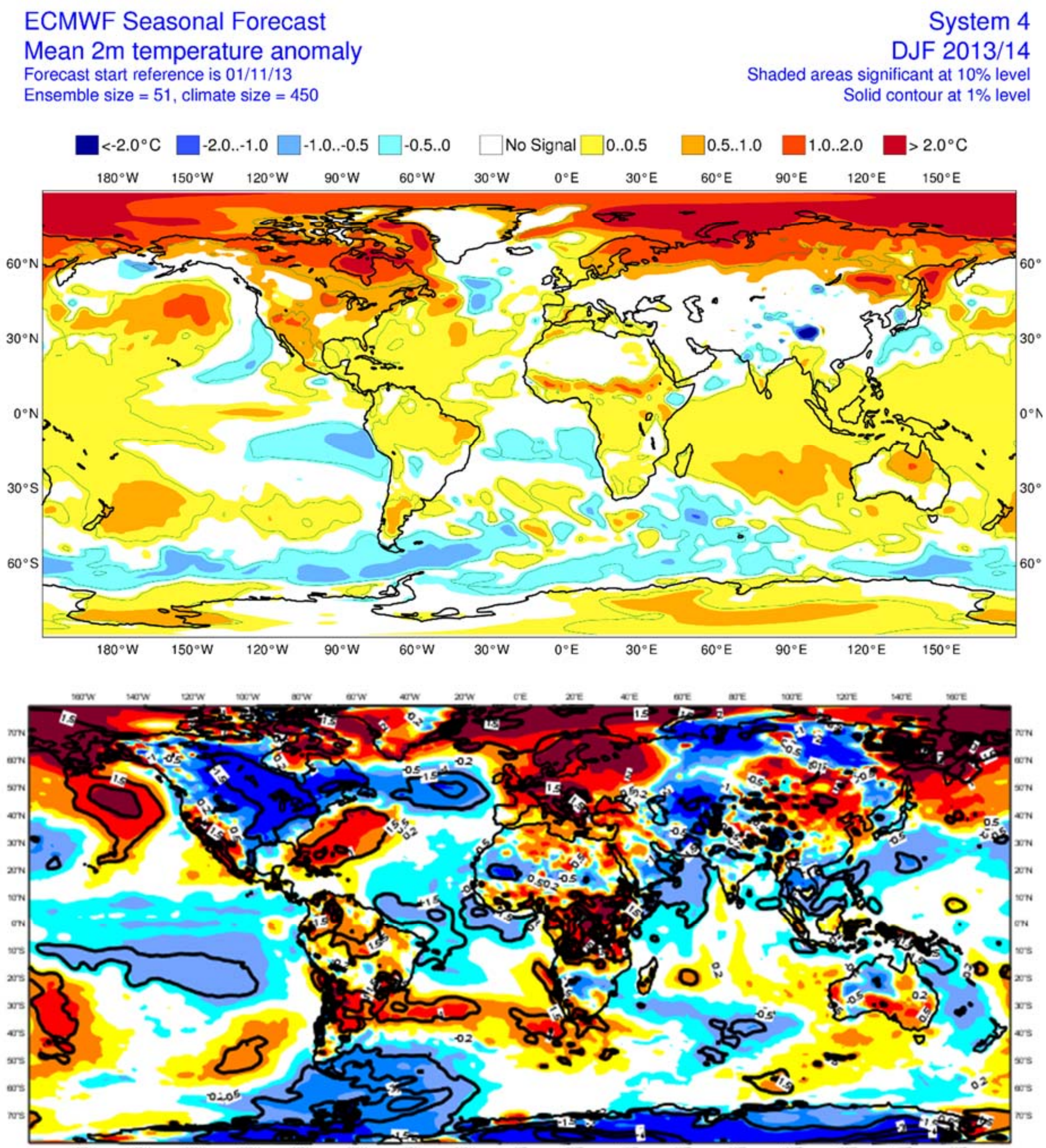


Figure 48: Anomaly of 2 m temperature as predicted by the seasonal forecast from Nov 2013 for DJF 2013-14 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

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 17), 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 17, Figure 19) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

Skill scores (Figure 3) are computed as the reduction in 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 2 and Figure 5 show correlations in space between the forecast anomaly and the verifying analysis anomaly. Anomalies with respect to ERA-Interim analysis climate are available at ECMWF from early 1980s. For ocean waves (Figure 34, Figure 35) 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 8) 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 39). Figure 39 also shows a modified ROC plot of hit rate against 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).

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 43.

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 21, Figure 23) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 21, Figure 23). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 24 to Figure 27), 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