TECHNICAL MEMORANDUM

765

Evaluation of ECMWF forecasts, including 2014-2015 upgrades

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

Research and Forecast Department

November 2015

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 upper-air 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 is addressed in section 5. Finally, section 6 discusses 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 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, 742). 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 website 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

In November 2014, ECMWF implemented an intermediate cycle (40r1.1) of its Integrated Forecasting System (IFS) that added the capability to actively assimilate all conventional observational data in BUFR format (binary code). This modification was needed since WMO allowed providers to stop disseminating data in Traditional Alphanumeric Codes (TAC) format in November 2014. WMO decided to move to a representation of observations in BUFR because of the lack of flexibility of the TAC format used for the exchange of surface and upper air observations for the last 50 years. The new BUFR format allows substantially more data to be provided, for example much higher vertical resolution in radiosonde reports, than was possible with the TAC format. However, a significant and continuing effort has been necessary to monitor the transition to the new format as numerous errors and quality issues have been identified as data providers introduce the change. ECMWF has been playing an active role in this effort in close collaboration with the Member States, EUMETNET's Observations Programme and the WMO.

A new model cycle (41r1) was implemented on 12 May 2015.

Cycle 41r1 introduces a lake parametrization, based on the FLake model, which is applied to all resolved and sub-grid scale lakes. The work on lakes resulting in this implementation has benefitted from a multi-year collaboration with the lake-NWP community in Europe and in particular recognizes the scientific co-ordination of the Deutsche Wetterdienst. This implementation improves 2-metre temperature forecasts in the vicinity of small lakes and near coastlines not represented in the previous model.

Ocean wave forecasts benefit from the extension of the high-resolution wave model from the European and North Atlantic region to the whole of the globe. These stand-alone forecasts are driven by the high resolution forecast HRES, are performed at a higher resolution than the coupled wave model, and include a forcing by ocean currents.

New land-sea mask, orography and climate fields (glacier information, surface albedo) have been introduced, as well as new data for lake depth and other lake parameters. The new model also uses new CO_2, O_3 and CH_4 climatologies from the latest MACC-II reanalysis.

A revised vertical interpolation in the semi-Lagrangian advection scheme reduces gravity wave noise during sudden stratospheric warming events.

The inner-loop resolutions of the 4DVAR data assimilation system have been upgraded to T255 (80 km) for each of the three iterations of the outer loops to produce finer scale increments. The background error covariances are made more flow-dependent by reducing the sampling window and averaging the statistics over shorter past periods, these dynamical statistics being used jointly with a climatology. A range of additional satellite observations improves the representation of land surface, sea ice and ocean wave parameters.

Monthly ensemble forecasts and re-forecasts have been extended from 32 to 46 days. The extended forecasts should be used with care but results have shown that there is positive skill in some aspects of forecasts in the 30–46 day range. The ensemble forecast (ENS) re-forecast dataset is significantly enhanced, with re-forecasts running twice a week, for Mondays and Thursdays (previously just Thursdays), and with the size of each re-forecast ensemble increased from 5 to 11 members. This provides a substantial increase in the sample size for the model climates for the medium-range Extreme Forecast Index (EFI)/Shift of Tails (SOT) and the extended-range (monthly) forecast anomaly products.

A summary of forecast performance is provided as a scorecard in Figure 1.

The new model cycle improves both high-resolution forecasts (HRES) and ensemble forecasts (ENS) throughout the troposphere and in the lower stratosphere. Improvements are seen both in verification against the model analysis and verification against observations.

Cycle 41r1 brings consistent gains in forecast performance at the surface for total cloud cover and precipitation. Improvements in the modelling of cloud and precipitation reduce the predicted occurrence of drizzle in situations where large-scale precipitation dominates, and they increase the amount of rainfall in forecasts of intense events, leading to a better match with observations. Improvements are also seen for 2-metre temperature and 2-metre humidity in parts of the northern hemisphere and the tropics. Cycle 41r1 also introduces a number of new output parameters, such as precipitation type, including freezing rain.

The average position error for tropical cyclones is slightly reduced, and tropical cyclones are generally forecast to be more intense. For example, IFS Cycle 41r1 performed better than Cycle 40r1 in predicting the track of tropical cyclone Pam, which devastated Vanuatu in the South

Pacific in March 2015. In HRES, the sea level pressure minimum at the centre of tropical cyclones is on average slightly lower at all lead times. Up to and including day 3 this makes the forecast better, by reducing the slight positive bias. From day 5 onwards, however, the pre-existing bias towards over-deepening has increased slightly.

The new model cycle is described in greater detail at

http://www.ecmwf.int/en/forecasts/documentation-and-support/changesecmwf-model/cycle-41r1.

3. Verification for upper-air medium-range forecasts

3.1. ECMWF scores

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 equal to or slightly exceeding the highest previous values.

A complementary measure of performance is the root mean square (RMS) error of the forecast. Figure 3 shows RMS errors for both extratropical hemispheres of the six-day forecast and the persistence forecast. The error of the six-day forecast has further decreased in the hemispheric averages.

Figure 4 shows the time series of the average RMS difference between four- and three-day (blue) and six- and five-day (red) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the general downward trend indicates that there is less “jumpiness” in the forecast from day to day. The level of consistency between consecutive forecasts has increased further in the last year. In 2014 the 12-month moving averages of RMS differences reached their lowest values so far.

The quality of ECMWF forecasts for the upper atmosphere in the northern hemisphere extratropics is shown through time series of temperature and wind scores at 50 hPa in Figure 5. Scores for one-day forecasts of temperature as well as forecasts of vector wind have been stable since last year.

The verification of model forecasts in the stratosphere is currently performed against analyses and radiosonde observations. Both datasets have limitations that reduce our ability to properly assess model developments in the stratosphere. GPS radio occultation (RO) observations represent an alternative way of verification in the stratosphere. They have a good vertical resolution, global and homogeneous distribution (around 3000 profiles per day), and they do not require bias correction.

Since RO measurements are assimilated directly in the form of bending angles, one option is to perform the verification using this quantity, which is primarily sensitive to variations in temperature in the stratosphere. RO bending angles can provide a robust measure of forecast error changes between model cycles. Figure 6 (top left panel) shows the reduction in error standard deviation of bending angles due to the most recent model upgrade. The bottom panel shows the corresponding time-series of bending angle departures for both model versions, indicating that the largest reductions in the random part of the forecast error occur during stratospheric warming events.

Since verification results for bending angles can be difficult to interpret, temperature retrievals from GPS-RO represent an alternative way of assessing the impact of model changes on systematic temperature forecast errors in the lower and middle stratosphere. However, GPS-RO temperature retrievals require a priori information about the upper atmosphere. In order to have robust results when comparing IFS cycles it is important to use the same prior information for RO temperature retrievals, which in this case is provided by the operational 6-hour forecast. The upper right panel in Figure 6 shows the reduction in the standard deviation of the temperature departures corresponding to the improvement in bending angle departures shown in the left panel.

The trend in ENS 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 HRES, the ENS 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 HRES. However, the ENS 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 in January 2010, and further redefinition of perturbations using the ensemble of data assimilations. The slightly decreasing trend in 2014 is due to atmospheric variability.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 8. The match between spread and error in 2014 is similar to previous years, although slightly stronger under-dispersion can be seen in the medium range. 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.

A good match between spatially and temporally averaged spread and error is a necessary but not a sufficient requirement for a well-calibrated ensemble. It should also be able to capture day-to-day changes, as well as geographical variations, in predictability. This can be assessed using spread-reliability diagrams. Forecast values of spread over a given region and time period are binned into equally populated spread categories, and for each bin the average error is determined. In a well-calibrated ensemble the resulting line should be close to the diagonal. Figure 9 and Figure 10 show spread-reliability plots for 500 hPa geopotential and 850 hPa temperature in the northern extratropics (top), Europe (centre), and the tropics (bottom, in Figure 10 only) for different global models. Spread reliability generally improves with lead time. At day 1 (left panels), forecasts tend to be more strongly under-dispersive at low spread values than at day 6 (right panels). ECMWF performs very well, with its spread reliability usually closest to the diagonal. The stars in the plots mark the average values, corresponding to Figure

8, and ideally should lie on the diagonal, as closely as possible to the lower left corner. Also in this respect, ECMWF performs best overall.

In order to have a benchmark for the ENS, the CRPS has been computed for a 'dressed'ERA-I. This also helps to distinguish the effects of IFS developments from pure atmospheric variability. The dressing uses the mean error and standard deviation of the previous 30 days to generate a Gaussian distribution around the ERA-I. Figure 11 shows the evolution of the CRPS for the ENS and for the dressed ERA-I 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 15% in 2005 and is approaching 30% in 2015. It is worth noting that using the forecast error for dressing of the ERA-I is equivalent to generating a nearly perfectly calibrated ensemble. Thus this sort of reference forecast represents a challenging benchmark.

The forecast performance over the tropics, as measured by RMS vector errors of the wind forecast with respect to the analysis, is shown in Figure 12. At 200 hPa (upper panel) the 1-day forecast has continued to improve slightly (although it is still slightly higher than the minimum which was reached in 2003–2004), while the 5-day forecast error has increased. Similarly, at 850 hPa (lower panel) the error at day 1 has been slightly reduced while at day 5 it has increased. The increase at 850 hPa is also seen in ERA-Interim (not shown) and in forecasts of other centres. It occurs for verification against analysis and does not appear when the forecast is verified against observations (cf. Section 3.2, Figure 16 and Figure 17). Note that scores for wind speed in the tropics are generally sensitive to inter-annual variations of tropical circulation systems such as the Madden-Julian oscillation, or the number of tropical cyclones.

3.2. WMO scores - comparison with other 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 (until December 2011) and new procedures (from 2012 onward). The scores from other centres for the period of this report have been computed still using the previous procedures. For the scores presented here the impact of the changes is relatively small for the ECMWF forecasts and does not affect the interpretation of the results.

Figure 13 shows time series of such scores for 500 hPa geopotential height in the northern and southern hemisphere extratropics. Over the last 10 years errors have decreased for all models, especially during the winter season. ECMWF continues to maintain a lead over the other centres.

WMO-exchanged scores also include verification against radiosondes over regions such as Europe. Figure 14 (Europe), and Figure 15 (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 16 (verification against analyses) and Figure 17 (verification against observations). When verified against the centres' own analyses, the Japan Meteorological Agency (JMA) forecast has the lowest error in the short range (day 1) while in the medium range, ECMWF and JMA are the leading models in the tropics. 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. In the tropics, verification against analyses (Figure 16) is very sensitive to the analysis, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 17), the ECMWF forecast has now the smallest overall errors both in the short and medium ranges.

4. Weather parameters and ocean waves

4.1. Weather parameters – high-resolution and ensemble

The supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 18. The top 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 bottom 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 HRES is due to atmospheric variability, as shown by comparison with the ERA-Interim reference forecast (dashed line in Figure 18, top 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 (centre panel in Figure 18). While the largest improvement is associated with the introduction of the five-species microphysics in November 2011 (cycle 36r4), microphysics changes in subsequent cycles led to a further increase in skill. The probabilistic score (lower panel in Figure 18) shows some recent improvement after the stagnant period 2010–2012 which was again 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, which has masked improvements due to model upgrades during that time. In 2012, however, this trend has reversed, so that model improvements have 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 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 the HRES leading with respect to the other centres from day 3 onwards while the Met-Office model is leading at day 1(Figure 19, upper panel), and for the ENS a consistent clear lead for ECMWF over the whole lead time range (Figure 19, bottom panel).

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 20 to Figure 23. 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. For 2 m temperature and dewpoint, the error standard deviation (upper curves in each plot) has been comparatively small. However, negative biases with marked annual cycles persist, especially at night-time. For total cloud cover (Figure 22) the bias has been small in recent years, and the error standard deviation has shown little change. For wind speed (Figure 23) the reduced level of night-time bias associated with the change in surface roughness in November 2011 has been maintained. However the daytime bias has become slightly more negative.

To complement the evaluation of surface weather forecast skill, 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 are shown. There is an increase in the skill of the operational high-resolution forecast relative to ERA-Interim in recent years, both in the extratropics and tropics (Figure 24), 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).

ERA-Interim is useful as a reference forecast for the HRES as it allows filtering out much of the effect of atmospheric variations on scores. Figure 25 shows the evolution of skill at day 5 relative to ERA-Interim in the northern hemisphere extratropics for various upper-air and surface parameters. The metric used is the RMSE for upper-air fields and the error standard deviation for the surface fields. Curves show 12-month running mean values. 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, which had been stable prior to 2011, started to increase as a result of more recent cycle changes. For mean sea levelpressure, the highest 12-month skill so far was reached at the end of 2014. Both for 500 hpa geopotential and 850 hPa temperature, maxima so far occurred early in 2014, and further increases are expected from the model upgrade in May 2015.

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 26. The top panel of Figure 26 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 2001 and it has reached its lowest value so far in the winter season 2013–14. Errors in the wave height forecast in 2014–15 have been the lowest so far in the 1–3 day range, and among the lowest at 5 days. The long-term trend in the performance of the wave model forecasts is also seen in the verification against analysis. Anomaly correlation for significant wave height has reached some of its highest values in 2014 (Figure 27).

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 28 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 28. ECMWF outperforms the other centres with regard to wind speed and peak period. For wave height, Météo France and ECMWF forecasts have highest skill.

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

http://www.ecmwf.int/en/forecasts/charts

under 'Ocean waves'.

5. Severe weather

Supplementary headline scores for severe weather are:

- The skill of the Extreme Forecast Index (EFI) for 10 m wind speed 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 29 (top), together with the corresponding results for 24-hour total precipitation (centre) and 2 m temperature (bottom). Each curve shows seasonal values, as well as the four-season running mean, of ROC area skill scores from 2004 to 2014; the final point on each curve includes the spring (March–May) season 2015. For all three parameters, ROC skill has stabilized on a high level, with some inter-annual variations due to atmospheric variability.

5.2. Tropical cyclones

The tropical cyclone position error for the 3-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 30. 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 and ENS position errors (top and bottom panels, Figure 30) have reached their lowest values so far. The same is true for the mean absolute speed errors of the HRES and the CTRL 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 30) 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 30 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. The forecast was under-dispersive before the resolution upgrade in 2010, but the spread-error relationship has improved 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 very 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 31. Results show a certain amount of over-confidence, with little change from year to year. Skill is shown by the ROC and the modified ROC, the latter using 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 three consecutive years of these two measures are not considered significant.

5.3. Additional severe-weather diagnostics

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 evaluate heavy precipitation forecast skill of the HRES. Forecasts are verified against synoptic observations. Figure 32 shows the time-evolution of skill expressed in terms of forecast days for 24-hour precipitation exceeding 20 mm in Europe. The gain in skill amounts to about two forecast days over the last 15 years and is primarily due to a higher hit rate. A more detailed evaluation of heavy precipitation forecast skill can be found in ECMWF Newsletter No. 144.

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. In IFS cycle 41r1 (May 2015) the monthly ensemble forecasts and re-forecasts have been extended from 32 to 46 days.

Figure 33 shows the probabilistic performance of the monthly forecast over the extratropical northern hemisphere for summer (JJA, top panels) and winter (DJF, bottom panels) seasons since September 2004 for week 2 (days 12–18, left panels) and week 3+4 (days 19–32 right panels). Curves show the ROC score for the probability that the 2 m temperature is in the upper third of the climate distribution in summer, and in the lower third of the climate distribution in winter. Thus it is a measure of the ability of the model to predict warm anomalies in summer and cold anomalies in winter. For reference, the ROC score of the persistence forecast is also shown in each plot. Forecast skill for week 2 exceeds that of persistence by about 10%, for weeks 3 to 4 (combined) by about 5%. In the weeks 3 to 4 (14-day period), summer warm anomalies appear to have slightly higher predictability than winter cold anomalies, although the latter has increased in recent winters (with the exception of 2012). Overall, the skill of the forecast is more stable from year to year than the skill of persistence, both in summer and winter.

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

http://www.ecmwf.int/en/forecasts/charts

6.2. Seasonal forecast performance

6.2.1. Seasonal forecast performance for the global domain

The current version (System 4) of the seasonal component of the IFS was implemented in November 2011. It uses the ocean model NEMO and ECMWF atmospheric model cycle 36r4. The forecasts contain 51 ensemble members and the re-forecasts 15 ensemble members, covering a period of 30 years.

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.

A comprehensive description and assessment of System 4 is provided in ECMWF Technical Memorandum 656, available from the ECMWF website.

6.2.2. The 2014–2015 El Niño forecasts

The year 2014 was characterized by a change from slightly cold to slightly warm conditions in the eastern tropical Pacific. The majority of ensemble members of the forecasts made in spring and summer of 2014 (upper two panels in the left column of Figure 34) predicted a more substantial strengthening of warm anomalies than what was observed. The autumn forecast captured the basic characteristics of the next months' evolution better, suggesting little change. The winter forecast (bottom left panel) again predicted a strong evolution towards El Nino conditions, which this time agreed very well with observations. The multi-model EUROSIP forecasts (right column of Figure 34) performed slightly better in the first half of the period, in the sense that the ensemble was better centred on the observations. The Feb 2015 forecast of the strong El Nino development in 2015 was predicted more clearly (with greater sharpness) in the ECMWF model.

The ECMWF forecast system predicts the continued strengthening of the El Nino in the coming months, with the peak expected around December 2015. Previous experience shows that for very large El Nino events (specifically 1997) the model tends to exaggerate the amplitude of SST anomalies due to non-linearities in the model bias characteristics. The 2015 El Nino is likely to be very strong, but more likely than not still weaker than the one in 1997. The longer range forecast (13 months lead) suggests that El Nino is likely to end sometime in April/May/June 2016, with temperatures either normal or below normal by July 2016.

6.2.3. Tropical storm predictions from the seasonal forecasts

The 2014 North Atlantic hurricane season was quiet with an accumulated cyclone energy index (ACE) of just 67% of the 1950–2012 climate average (see Figure 35). The number of tropical storms which formed in 2014 (8 named storms) was also below average (12). Seasonal tropical storm predictions from System 4 indicated below average activity compared to climatology over the Atlantic. The June forecast predicted 9 (with a range from 6 to 12) tropical storms in the Atlantic (Figure 36) and an ACE of 60% of the observed climatology (+/- 20%). Most other seasonal forecast models predicted a below average 2014 Atlantic tropical storm season due to the presence of a moderate El-Nino event.

Figure 36 shows that System 4 predicted above average activity over the eastern North Pacific (although with the ACE 10% below normal) and slightly enhanced activity over the western North Pacific (ACE 20% above normal). The 2014 eastern Pacific hurricane season was well above average (19 tropical storms formed over the eastern North Pacific between July and December) and the ACE was 43% above the 1981–2010 average, which makes it the seventh-highest since 1971. Only 17 tropical storms formed over the western North Pacific in 2014 between July and December, which is below average (21.3). There was no clear signal in the forecast over this basin. The drop of the ACE in the Atlantic sector was well captured by the forecast, with almost the same amplitude as observed, although most ensemble members predicted at the time a stronger El-Nino (conducive to a reduction in tropical cyclone activity in the Atlantic) than observed (Figure 34).

6.2.4. Extratropical seasonal forecasts

2 m temperatures in the northern-hemisphere winter (DJF 2014–15) were characterized by strong warm anomalies over northern Eurasia, which were captured quite well by the seasonal forecast (Figure 37). The western parts of Europe, in contrast, were influenced by a persistent cold anomaly over the North Atlantic. This anomaly was also captured by the model but it did not extend quite as far into Europe as was observed. The seasonal forecast was not able to predict the large-scale cold anomaly over much of the eastern United States and Canada.

During the northern-hemisphere summer (JJA 2015), central and southern Europe experienced extremely persistent heat, leading to a magnitude of seasonal 2 m temperature anomaly of more than 2 standard deviations above normal (Figure 38). In some European countries all-time record temperatures were surpassed. Temperatures in northern Europe were at the same time below normal, with a steep gradient of the anomaly at about 50-55 degrees north. The seasonal forecast predicted positive anomalies in southern Europe, and non-significant anomalies in northern Europe. The extension of the anomalous warmth into Asia was well captured.

				anomaly correlation	RMS error	SEEPS
Europe	against analysis	Geopotential	100hPa	▴▲▴▴▴▴▨▨▨🟩	▲▲▲▲▲▴▴🟩▨🟩	
			500hPa	▴▴▴🟩🟩🟩▨▨▨▨	▲▴▴▴▴🟩▨▨▨▨	
			850hPa	▴▴▴🟩🟩🟩🟩🟩▨▨	▴▲▴▴▴🟩▨▨▨▨	
			1000hPa	▴▴▴🟩▴▴🟩🟩▨▨	▲▲▴▴▴🟩▨▨▨▨	
		MSL pressure		🟩▴▴▨🟩▨▨▨▨🟥	▴▴▴🟩▨▨▨▨▨🟥	
		Temperature	100hPa	▲▲▲▴▴▴▨▨▨▨	▲▲▲▲▲▲▴▴🟩▴	
			500hPa	▴▴▴▴🟩🟩▨▨▨🟩	▴▴▴▴▴▴▨▨▨▨	
			850hPa	▲▴▴🟩▴▴▨▨▨▨	▲▴▴🟩▴🟩▨▨▨▨	
			1000hPa	▴▴▴▴▴▴🟩🟩🟩▨	▴▴▴▴▴▴▴🟩▨▨	
		Wind	200hPa	▲▴▴▴▴▴▨▨🟩▨	▲▲▴▴▴🟩▨▨▨▨	
			850hPa	▴▴▴▴▴🟩▨▨▨▨	▴▴▴▴▴🟩▨▨▨🟥	
		Relative humidity	300hPa	▴🟩▴🟩▨▨🟥▨▨▨	▴▴▴🟩🟩▨🟥▨▨▨	
			700hPa	▲▴▴▴▴▴▨▨▨▨	▲▲▲▴▴▴▨🟩▴🟩	
	against observations	Temperature	100hPa	🟩▨🟩🟩▨▨▨▨▨▨	▲▲▲▲▴▴▴▨▨🟩	
			200hPa	🟩▴▴▨▨🟥🟥▨▨▨	🟩🟩▴▨▨🟥🟥▨▨▨	
			850hPa	▴▨▨▨▨🟩▨▨▨▨	▨▨▨▨▨🟩▨▨▨▨	
		2m temperature			▴▴▲▴▴▴▴🟩▴▨	
		Wind	100hPa	▴▴▴🟩▨▨▨▨▨▨	▴▴🟩▨▨▨🟥🟥▨▨	
			200hPa	▨▴🟩🟩▨▨▨▨▨▨	🟩🟩🟩🟩▨🟥🟥▨▨▨	
			850hPa	🟩🟩🟩▨▨▨▨▨▨▨	🟩▴🟩▨▨▨▨▨▨▨	
		10m wind			🟥▨▨▨▨▨▨▨▨▨	
		2m dew-point			▼🟥🟩▴▴🟩▴▴▴🟩	
		Total cloud cover			▲▲▲▴🟩▴▴🟩▨▨	
		24h precipitation				▴▴▴🟩▨▨▨🟥▨▨
Extratropical Northern Hemisphere	against analysis	Geopotential	100hPa	▲▲▲▲▲▴▴▴▴▴	▲▲▲▲▲▲▲▲▲▴	
			500hPa	▲▲▲▴▴▴▴🟩🟩🟩	▲▲▲▴▴▴▴▴▴🟩	
			850hPa	▲▲▲▴▴▴▴▴🟩🟩	▲▲▲▴▴▴▴▴▴▴	
			1000hPa	▲▲▴▴▴▴▴▴🟩▨	▲▲▴▴▴▴▴▴🟩🟩	
		MSL pressure		▴▴▴🟩🟩▴▴🟩▨▨	▴▴▴🟩🟩▴▴🟩▨▨	
		Temperature	100hPa	▲▲▲▲▴▴▴▴▴▴	▲▲▲▲▲▲▲▲▲▲	
			500hPa	▲▲▲▴▴▴▴▴🟩🟩	▲▲▲▴▴▴▴▴▴▴	
			850hPa	▲▴▴▴▴▴▴🟩▨▨	▴▨▨▨▨🟩🟩▨▨▨	
			1000hPa	▴▴▴▲▴▴▴▴▴🟩	▲▲▲▲▴▴▴▴▴▴	
		Wind	200hPa	▲▲▲▴▴▴▴▴▴▴	▲▲▲▲▴▴▴▴▴▴	
			850hPa	▲▴▴▴▴▴▴▴▴▴	▲▴▴🟩▴▴▴▴▴▴	
		10m wind over ocean		▴▴🟩🟩🟩🟩▴🟩▴▴	🟩▨▨▨▨▨▨▨▴▴	
		Ocean wave height		▼▾▾▨🟩▴▴🟩🟩▨	▼▾🟥▨▨🟩▴🟩▴🟩	
		Ocean wave period		▼▼▼▼▾🟥▨▨▨▨	▼▼▼▼▾▾▨▨▨▨	
		Relative humidity	300hPa	▲▴▴▴▴▴🟩▨▨▨	▲▴▴▴▴▴▴▴▴▴	
			700hPa	▲▴▴▴▴▴🟩🟩▴▴	▲▲▲▲▲▲▲▲▲▲	
	against observations	Temperature	100hPa	▴▴▴▴▨▨🟩▨🟩▨	▲▲▲▲▲▲▲▴▴▴	
			200hPa	▴▴▴▴▨▨🟥▨▨▨	▴▴▴🟩▨▨▨▨🟩🟩	
			850hPa	▾▾▨🟥▨▨▨▨▨▨	▾▾🟥🟥▨▨▨▨▨▨	
		2m temperature			▲▲▲▲▴▴▴▴▴🟩	
		Wind	100hPa	▲▲▴▴▴🟩▨▨▨🟩	▲▲▴▴▴▴▨▨▨▨	
			200hPa	▲▴▴▴🟩▨▨▨▨🟩	▲▴▴▴🟩▨▨▨▨🟩	
			850hPa	▾▨▨▨🟩🟩▨▨▨▨	▾▨▨▨🟩🟩▨▨🟩🟩	
		10m wind			▲▲▲▲▴▴▨🟥▨▨	
		2m dew-point			▴▲▲▲▲▲▴▴▴▴	
		Total cloud cover			▴▴▴▴▴▴▴▴▴▴	
		24h precipitation				🟩▨▨🟩▨▨▨▨▨▨
Extratropical Southern Hemisphere	against analysis	Geopotential	100hPa	▴▴▴▴▴▴▴▴🟩🟩	▲▲▲▲▲▲▲▲▲▲	
			500hPa	▲▲▲▴▴▴▴▴▴▴	▲▲▲▴▴▴▴▴▴▴	
			850hPa	▲▲▲▴▴▴▴▴▴▴	▲▲▲▴▴▴▴▴▴▴	
			1000hPa	▲▲▲▴▴▴▴▴▴▴	▲▲▲▴▴▴▴▴▴▴	
		MSL pressure		▲▲▴▴▴▴▴▴▴▴	▲▲▴▴▴▴▴▴▴▴	
		Temperature	100hPa	▲▲▲▲▲▲▴▴▴▴	▲▲▲▲▲▲▲▲▲▲	
			500hPa	▲▲▲▴▴▴▴🟩▴▴	▲▲▲▲▴▴▴▴▴▴	
			850hPa	▲▴▴🟩🟩🟩🟩🟩▨▨	▨▾▾▾▾▾🟥▾▾▾	
			1000hPa	▲▲▲▲▴▴▴▴▴▨	▲▲▴▴▴🟩🟩🟩🟩▨	

				anomaly correlation	RMS error	SEEPS
		Wind	200hPa	▴▴▴▴▴▴▴▴🟩🟩	▲▴▴▴▴▴▴▴▴▴	
			850hPa	▲▲▴▴▴▴▴▴▴▴	▲▲▲▴▴▴▴▴▴▴	
		10m wind over ocean		▲▴▴▴▴▴▴▴▴▴	▴▴▴▴▴🟩▨🟩▨▴	
		Ocean wave height		▼▼🟥▨▨▨🟩▴▴▴	▼▼🟥▨▨▨🟩▴🟩▴	
		Ocean wave period		▼▼▼▼▼▾🟥🟥▨🟩	▼▼▼▼▼▾▾▾▨▨	
		Relative humidity	300hPa	▴▴▴▴▴▴▴▴▴▴	▨🟩▴▴▴▴▴▴▴▴	
			700hPa	▲▴▴▴▴🟩▨▨▨▨	▲▲▲▲▲▲▲▲▲▲	
	against observations	Temperature	100hPa	▴▴▴▴▴▴▴▴▴▴	▲▲▲▲▲▲▲▲▲▲	
			200hPa	▴▴▴▴▴▴▴▴▴🟩	▴▴▴▴▴🟩▴▴▴🟩	
			850hPa	🟩🟩▨🟩▨▨▴▨🟩🟥	▨▨▨🟩🟩▨▴▨▨🟥	
		2m temperature			▴▴▴▴▴▴▴▴🟩▨	
		Wind	100hPa	▲▲▲▴▴▴▴🟩🟩🟩	▲▲▲▲▴▲▴▴▴▴	
			200hPa	▨▴▴▴▴▴▴▴🟩🟩	▨▴▴▴▴▴▴▴🟩▴	
			850hPa	▴▴🟩▴🟩▴🟩▨▨▨	▴▴▴▴▴▴▴🟩▨🟩	
		10m wind			▾▼▾▾▾🟥▨🟥▨🟥	
		2m dew-point			▨🟥▾▨▨🟩▴🟩▨▨	
		Total cloud cover			▲▴▴▴▴▴▴🟩▴🟩	
		24h precipitation				▨▨▨▨🟩▨▴🟩▨▨
Tropics	against analysis	Temperature	100hPa	▲▲▲▲▲▲▲▲▲▲	▲▲▲▲▲▲▲▲▲▲	
			500hPa	▲▲▲▲▲▲▲▲▲▲	▲▲▲▲▲▲▲▲▲▲	
			850hPa	▲▲▴▴▴▴▴🟩▨🟩	▲▴▨▨▨▨🟩▨▨▨	
			1000hPa	▨▾▾🟥🟥▨🟥🟥▾▾	▲▲▲▲▲▲▲▲▲▴	
		Wind	200hPa	▲▲▲▲▲▲▲▲▲▲	▲▲▲▲▲▲▲▲▲▲	
			850hPa	▲▲▲▲▲▲▲▴▴▴	▲▲▲▴▴▴▴▴🟩🟩	
		10m wind over ocean		▲▲▲▲▲▲▲▴▴▴	▲▲▲▲▴▴🟩▨▾🟥	
		Ocean wave height		▼▼▼▼▼▾▾▾▾▾	▼▼▼▼▼▼▼▼▼▾	
		Ocean wave period		▼▼▼▼▼▾▾▾▾🟥	▼▼▼▼▼▼▼▼▾▾	
		Relative humidity	300hPa	🟩▴▴▴▴▴▴▴▴▴	▼▼▾▾▾▾▾▾▾▾	
			700hPa	▲▴▴▴▴▴▴▨▨▨	▲▲▲▲▲▲▲▲▲▲	
	against observations	Temperature	100hPa	▲▲▲▲▲▲▲▲▲▴	▲▲▲▲▲▲▲▲▲▲	
			200hPa	▲▲▲▲▲▲▲▴▲▴	▲▴🟩▨🟥🟥🟥🟥▨🟥	
			850hPa	▨▨🟩▨▨▨▨🟩▨▨	▴▴▴▴▴▴▴▴🟩🟩	
		2m temperature			▲▲▲▲▲▲▲▲▲▲	
		Wind	100hPa	▲▲▲▲▲▲▲▴▴▴	▲▲▲▲▲▲▲▲▲▲	
			200hPa	▴🟩▴▴▴▴▴🟩🟩🟩	▴▨▴▴▴▴▴▴▴▴	
			850hPa	▲▲▲▴▴▴▴▴▴▴	▴▴▴▴▴▴🟩🟩🟩▨	
		10m wind			▼▼▼▼▼▼▼▼▼▾	
		2m dew-point			▲▲▲▲▲▲▲▲▲▲	
		Total cloud cover			▲▲▲▲▲▲▲▲▲▲	
		24h precipitation				▴▴▴▴🟩🟩🟩▨▨▨

Symbol legend: for a given forecast step... (d: score difference, s: confidence interval width)

▲ CY41r1 **better** than CY40r1 **statistically highly significant** (the confidence bar above zero by more than its height) (d/s>3)

▴ CY41r1 **better** than CY40r1 **statistically significant** (d/s≥1)

🟩 CY41r1 better than CY40r1, yet not statistically significant (d/s≥0.5)

▨ not really any difference between CY40r1 and CY41r1

🟥 CY41r1 worse than CY40r1, yet not statistically significant (d/s≤-0.5)

▾ CY41r1 **worse** than CY40r1 **statistically significant** (d/s≤-1)

▼ CY41r1 **worse** than CY40r1 **statistically highly significant** (the confidence bar below zero by more than its height) (d/s<-3)

Figure 1: Summary score card for Cy41r1. Score card for cycle 41r1 versus cycle 40r1 verified by the respective analyses and observations at 00 and 12 UTC for 704 forecast runs in the period 2 January 2014 to 10 May 2015.

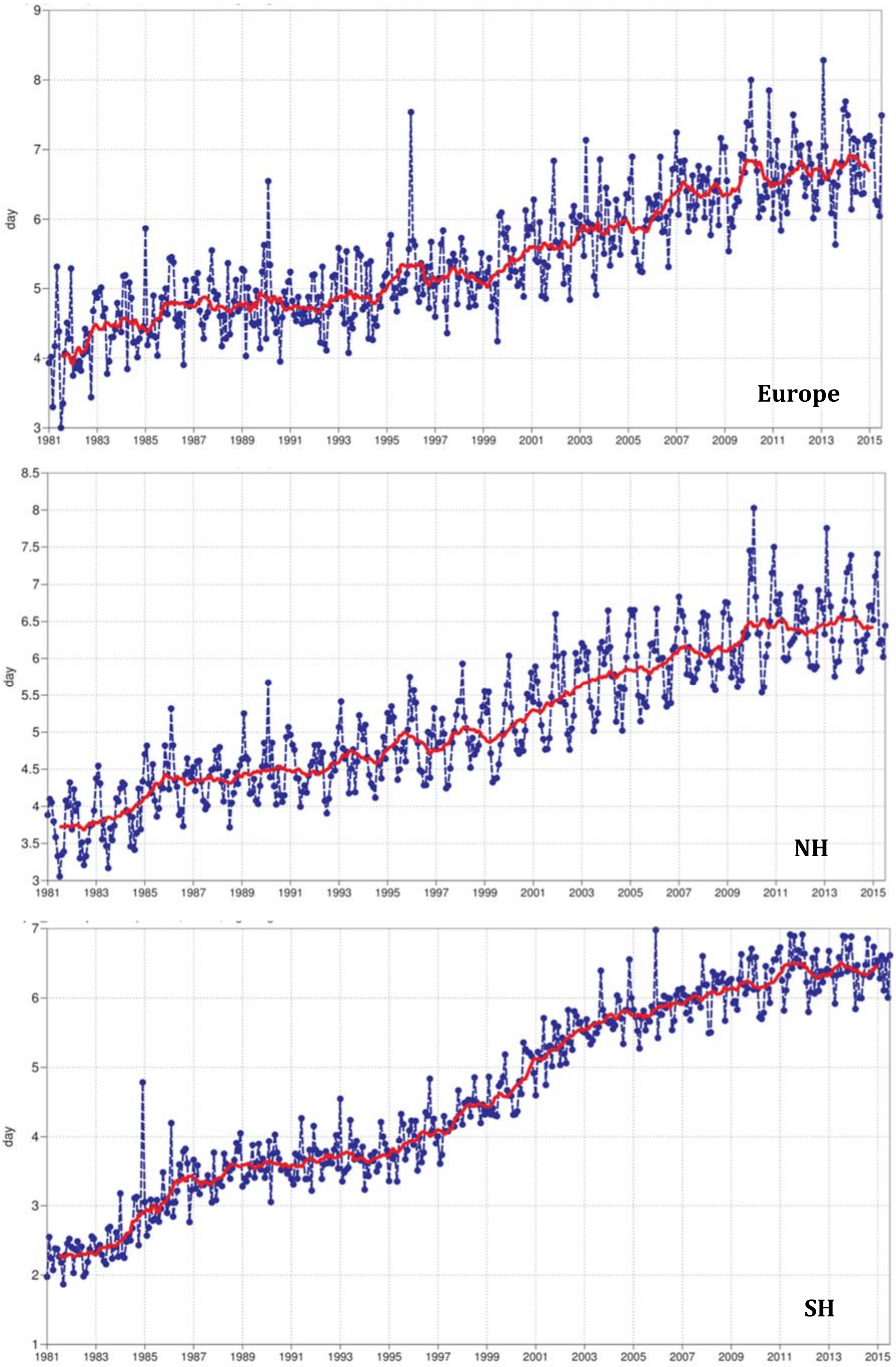


Figure 2: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly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

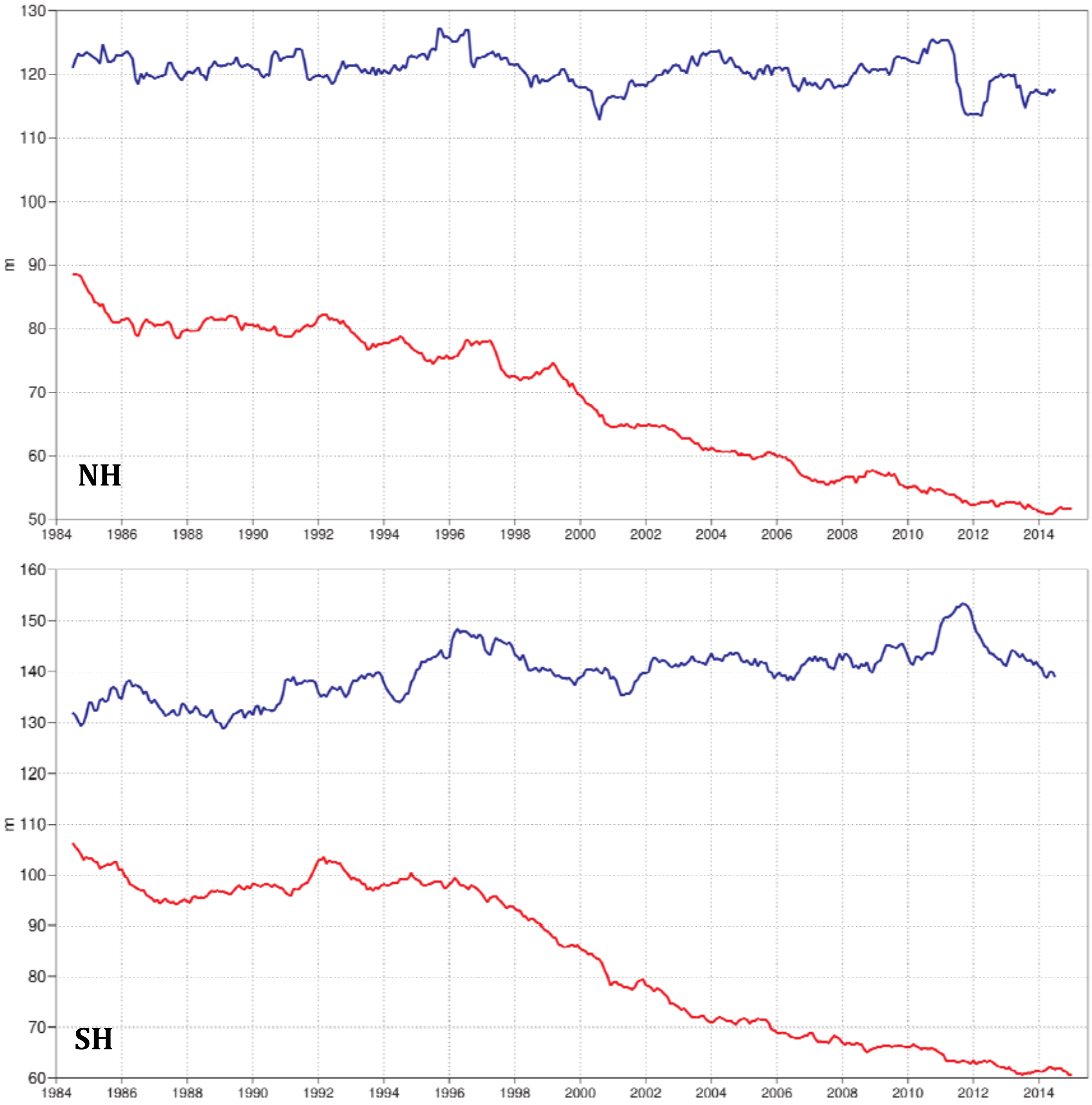


Figure 3: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2014–July 2015. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

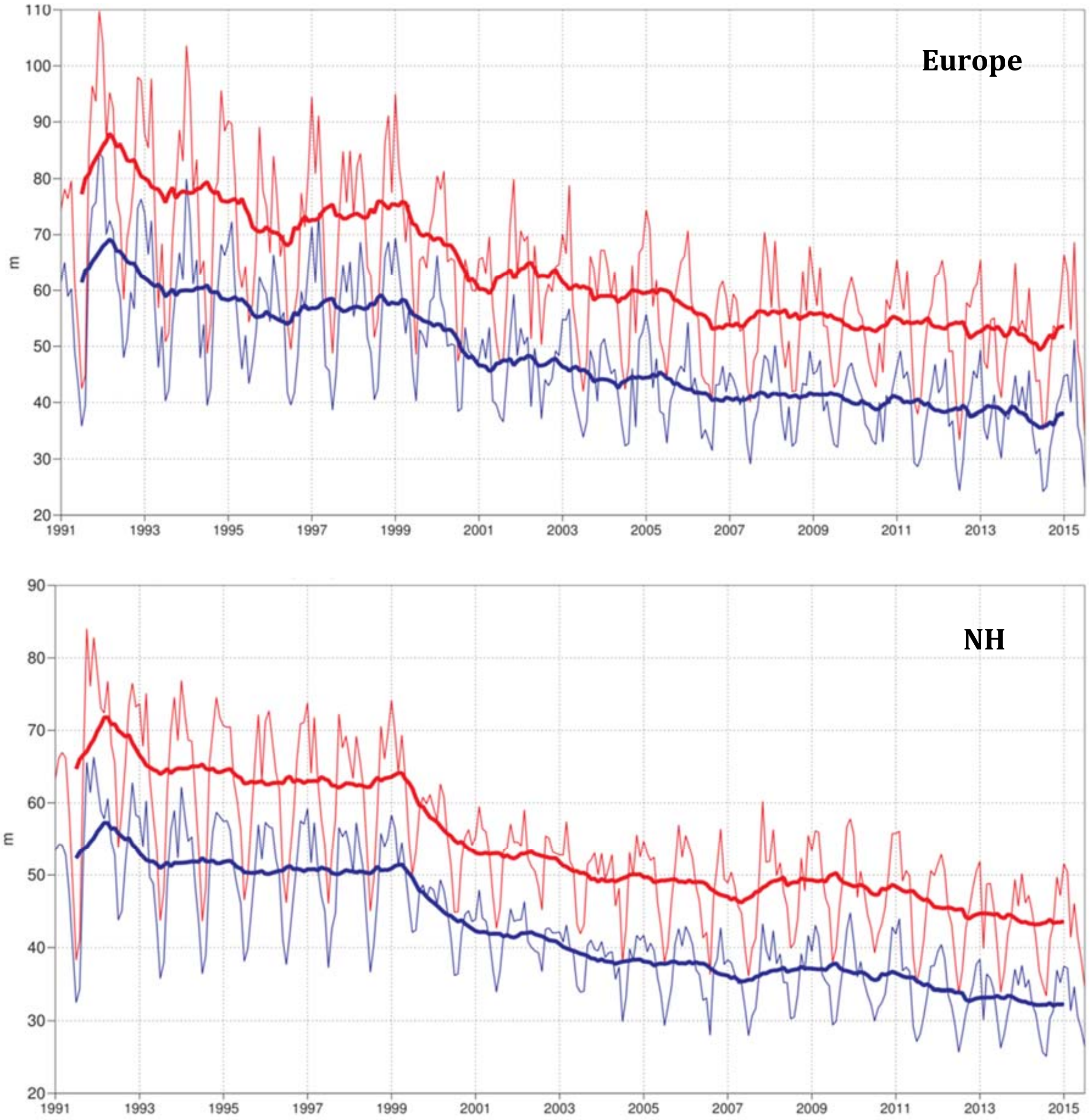


Figure 4: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores are also shown (in bold).

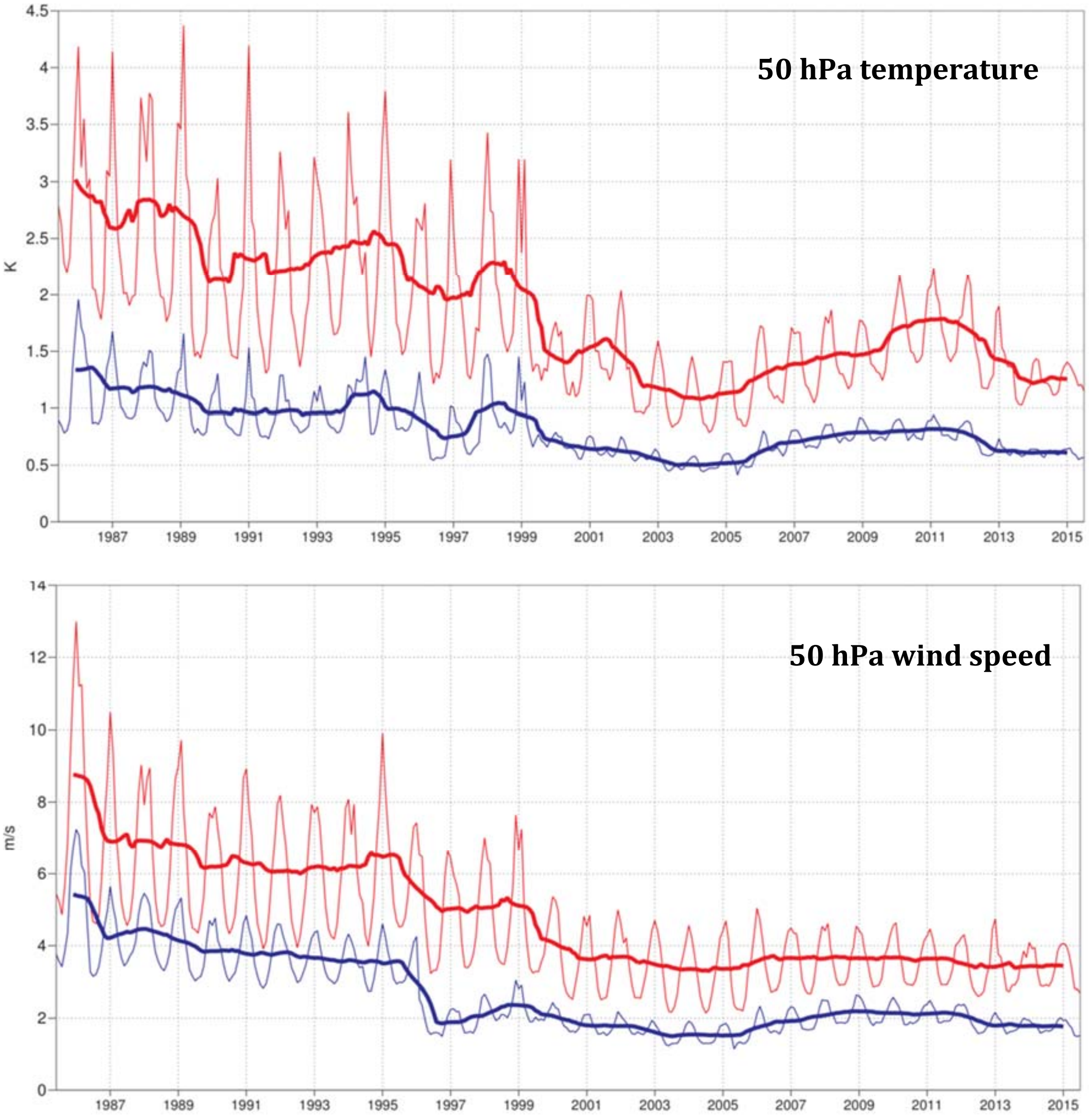


Figure 5: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

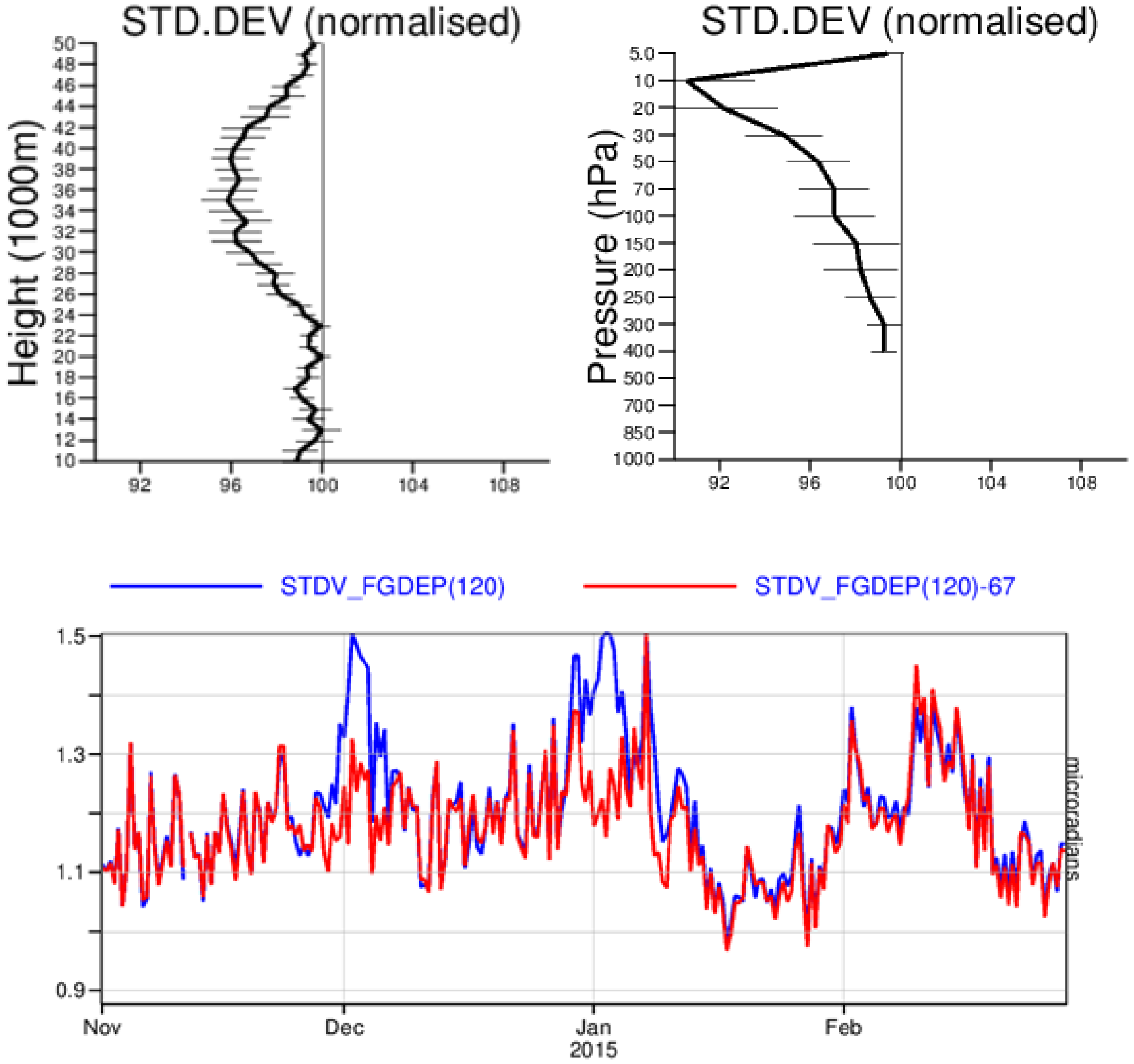
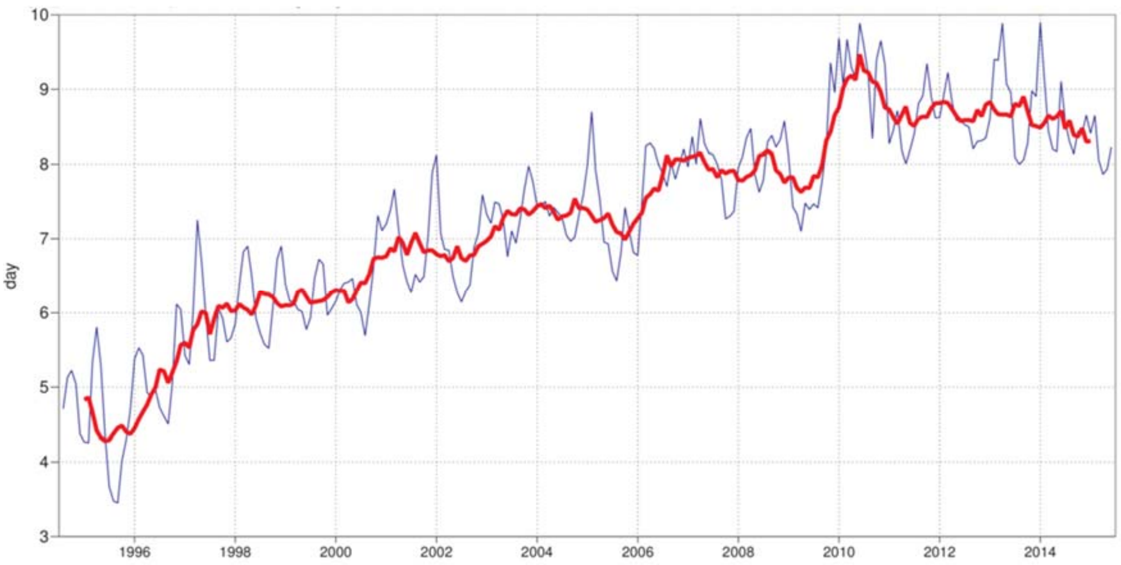


Figure 6: Normalised differences between model cycles 41r1 and 40r1 of the standard deviation of day 5 forecast departures (normalised by the observation error) of GPSRO bending angles (upper left panel) and GPSRO temperature retrievals (upper right panel) over the northern hemisphere extra-tropics. Values below 100 indicate a reduction in the standard deviation compared to the reference. The bottom panel shows corresponding time series of the standard deviation of 5-day forecast departures of GPSRO bending angles for the layer between 40 and 49 km (blue: 40r1, red: 41r1).



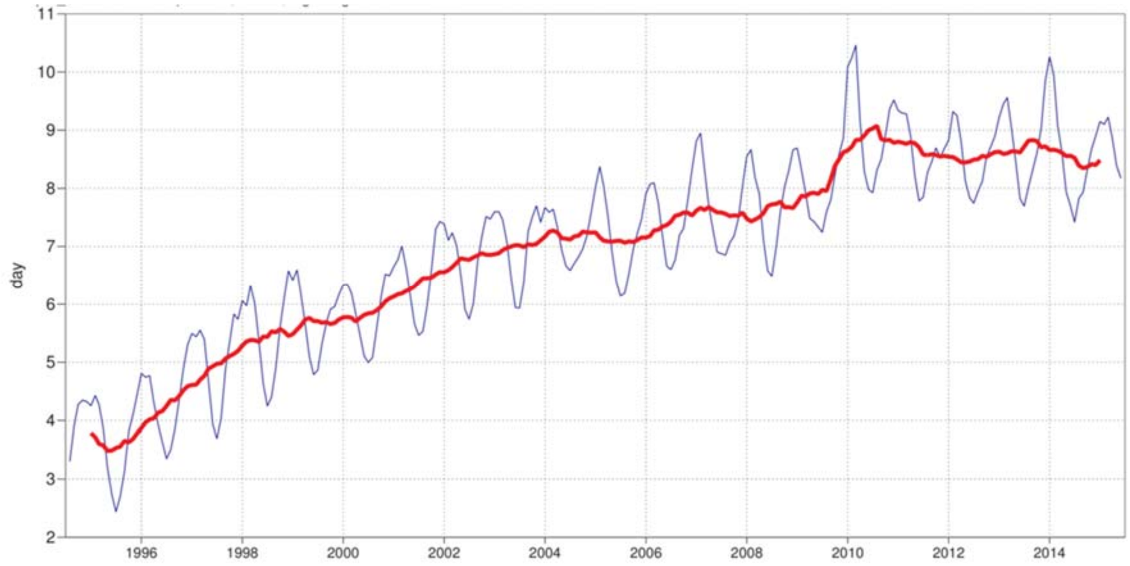


Figure 7: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

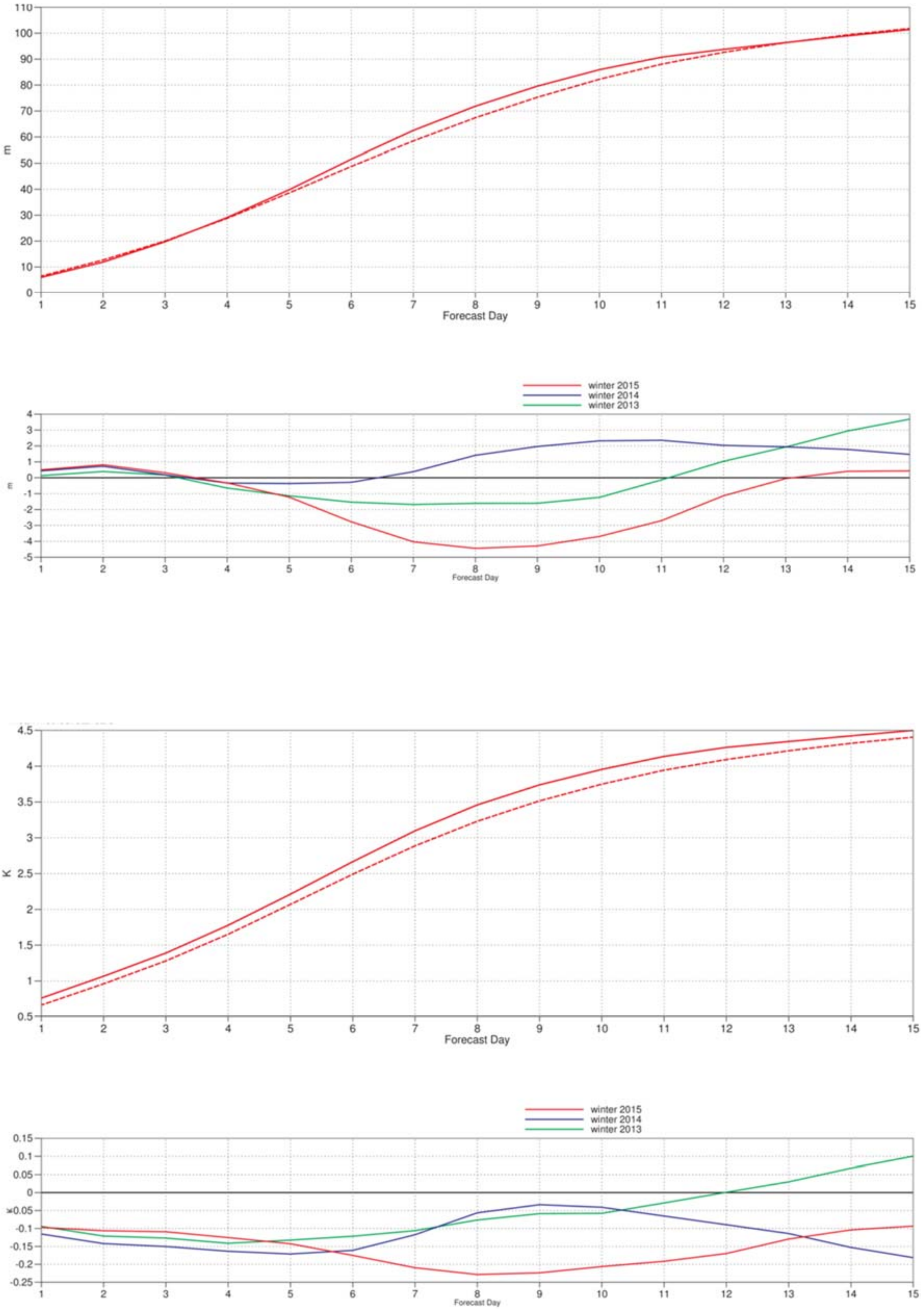


Figure 8: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2014–2015 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate spread is too small); verification is against analysis, plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extratropical northern hemisphere for forecast days 1 to 15.

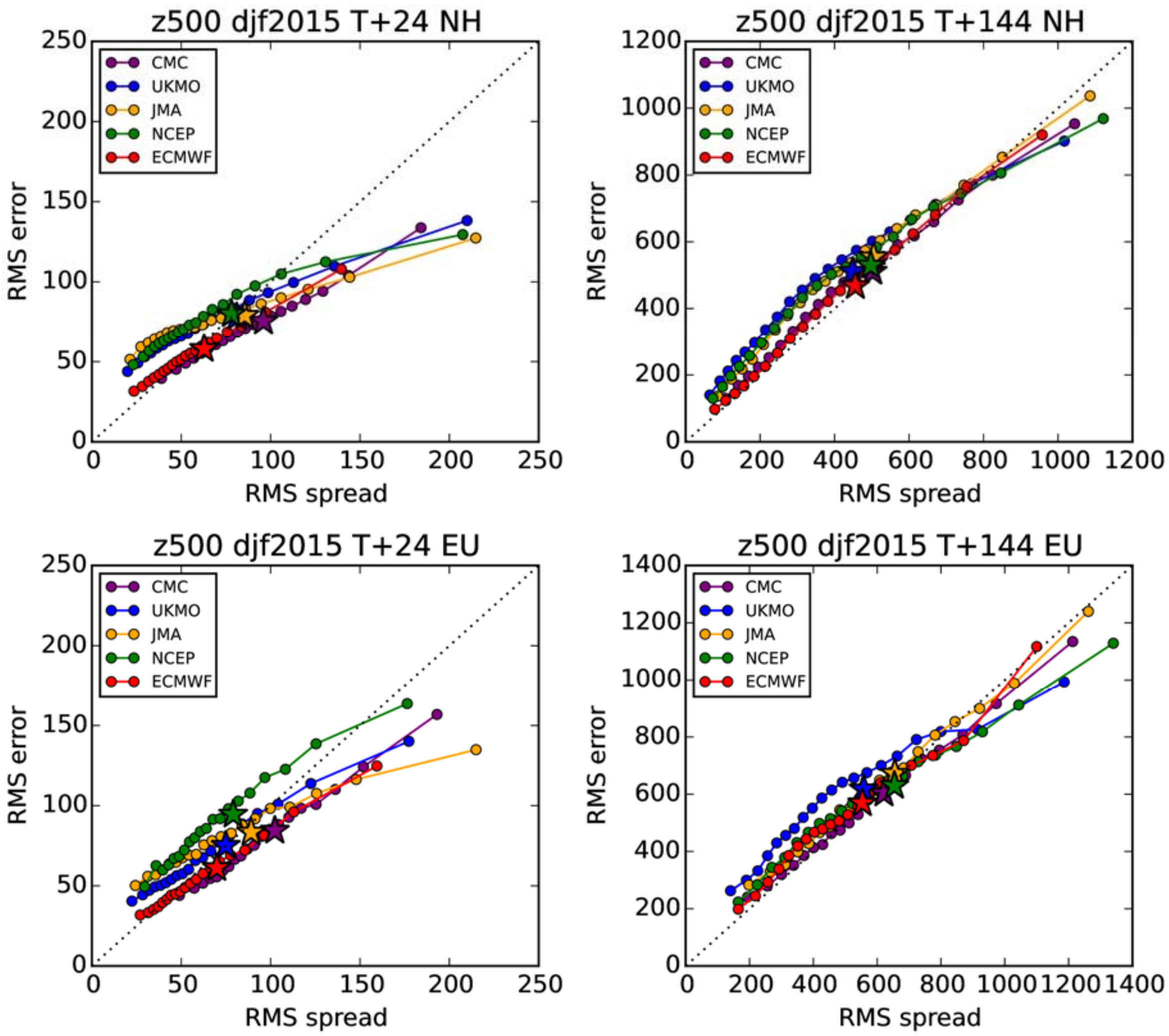


Figure 9: Ensemble spread reliability of different global models for 500 hPa geopotential in DJF 2014–15 in the northern hemisphere extra-tropics (top) and in Europe (bottom) for day 1 (left) and day 6 (right) , verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

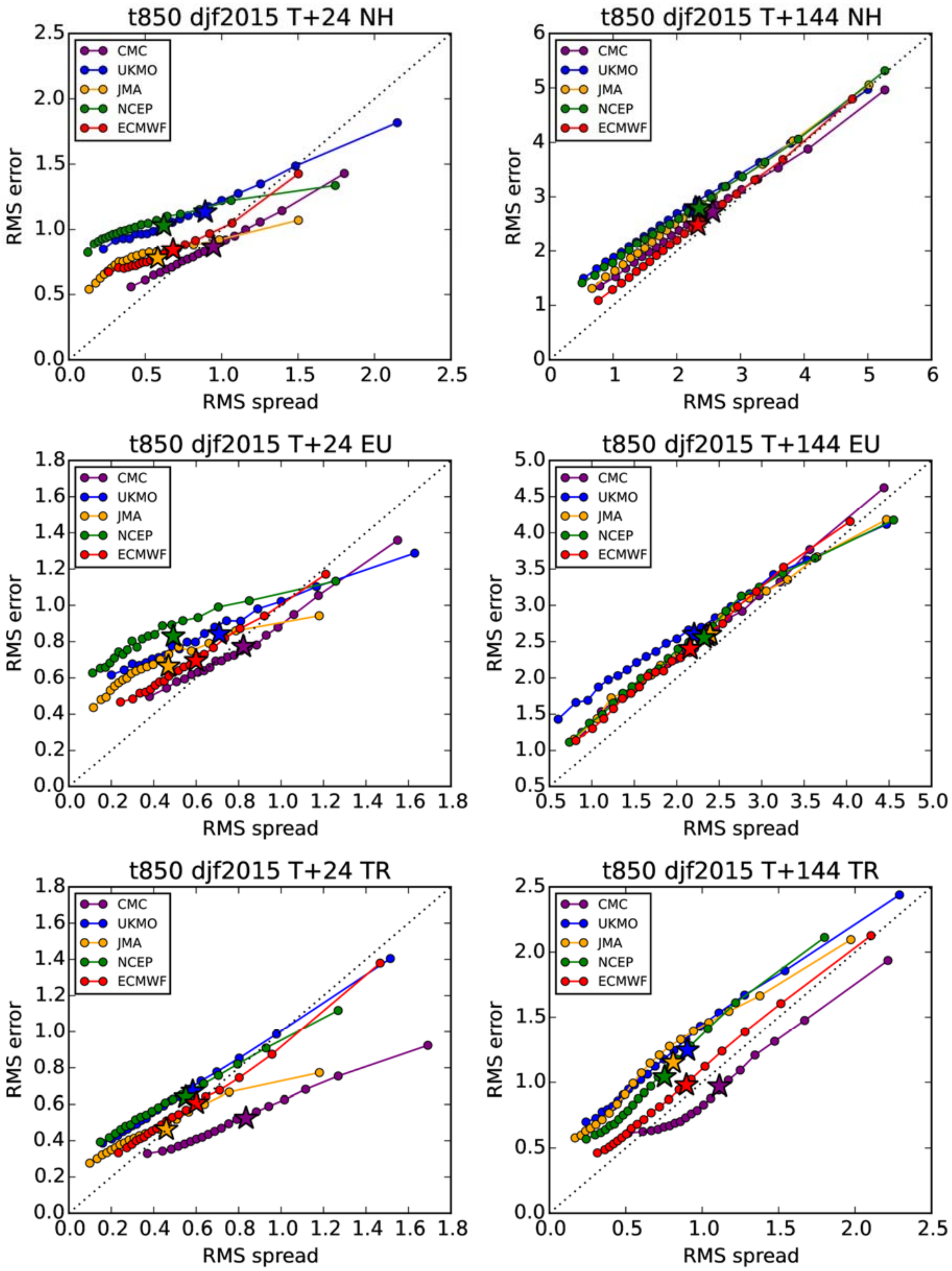
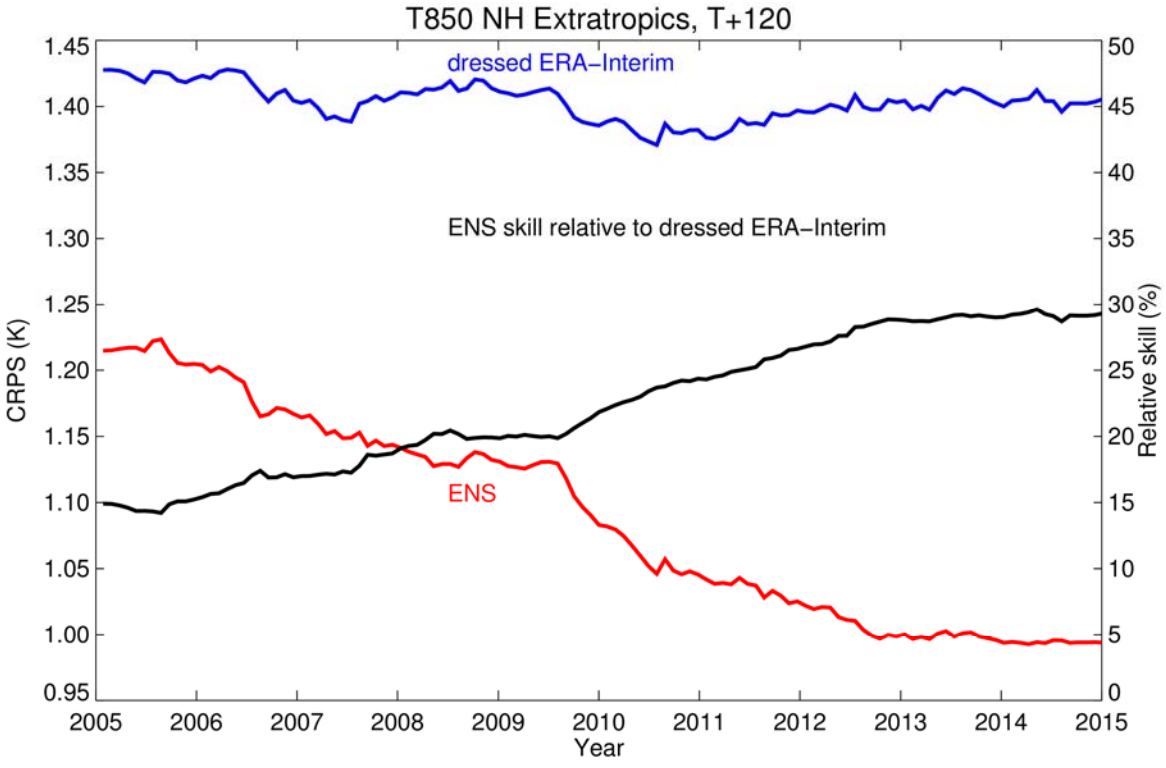


Figure 10: Ensemble spread reliability of different global models for 850 hPa temperature in DJF 2014–15 in the northern hemisphere extra-tropics (top), Europe (centre), and the tropics (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.



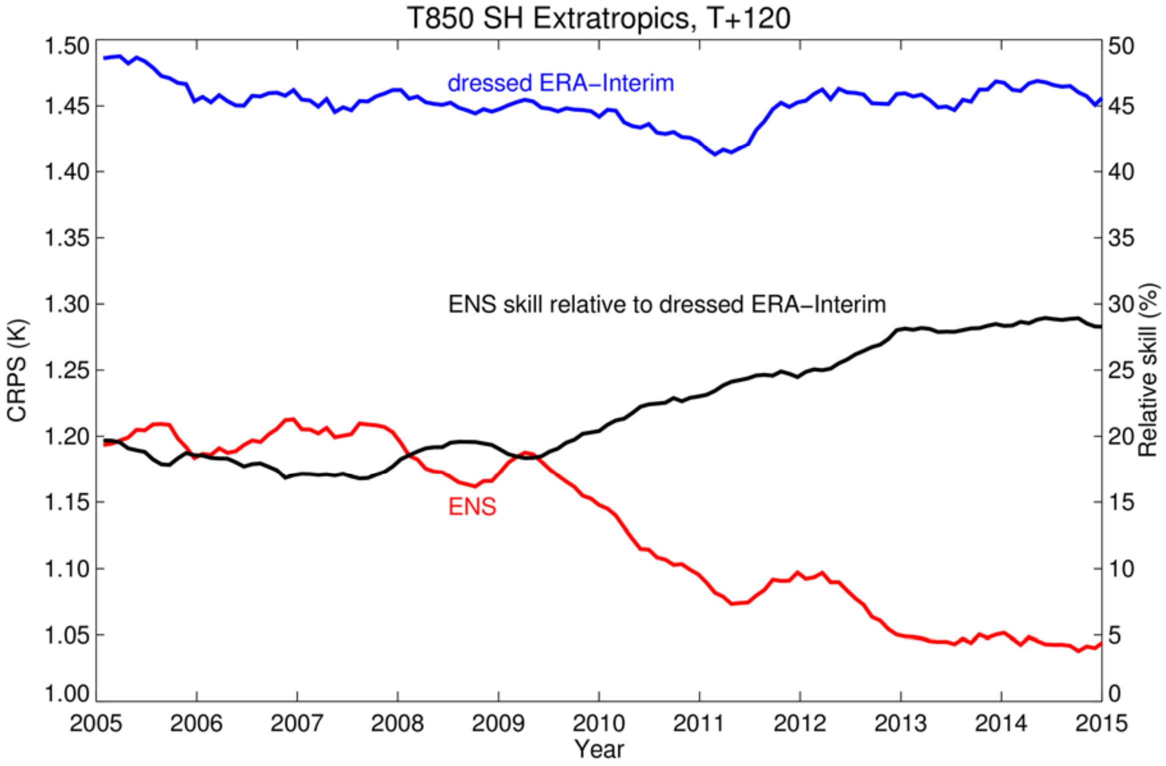
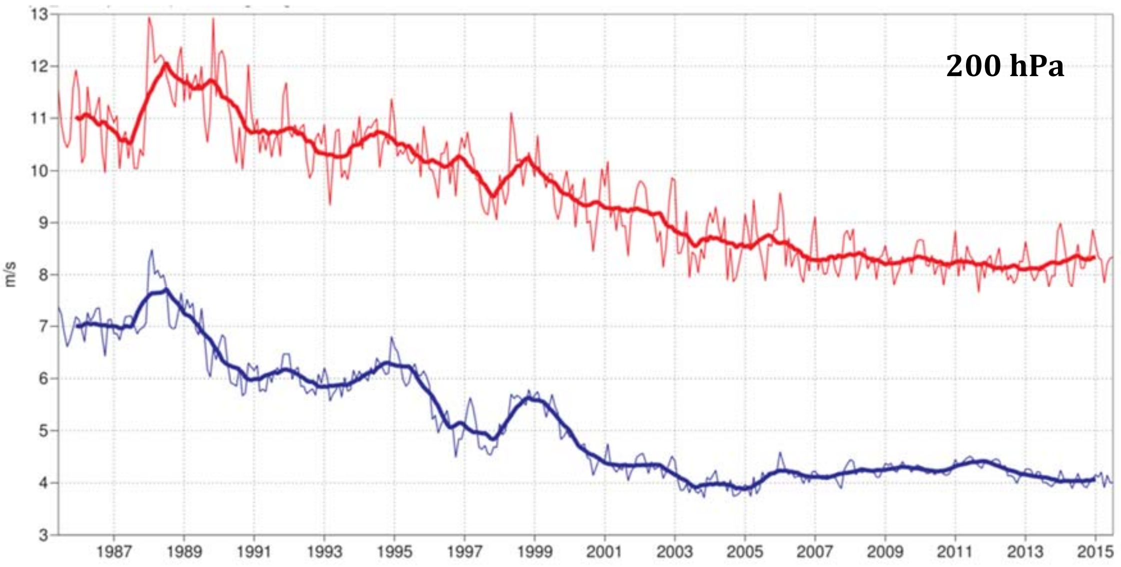


Figure 11: CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5, verified against analysis. Scores are shown for the ensemble forecast (red) and the dressed ERA-Interim forecast (blue). Black curves show the skill of the ENS relative to the dressed ERA-Interim forecast. 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**.**



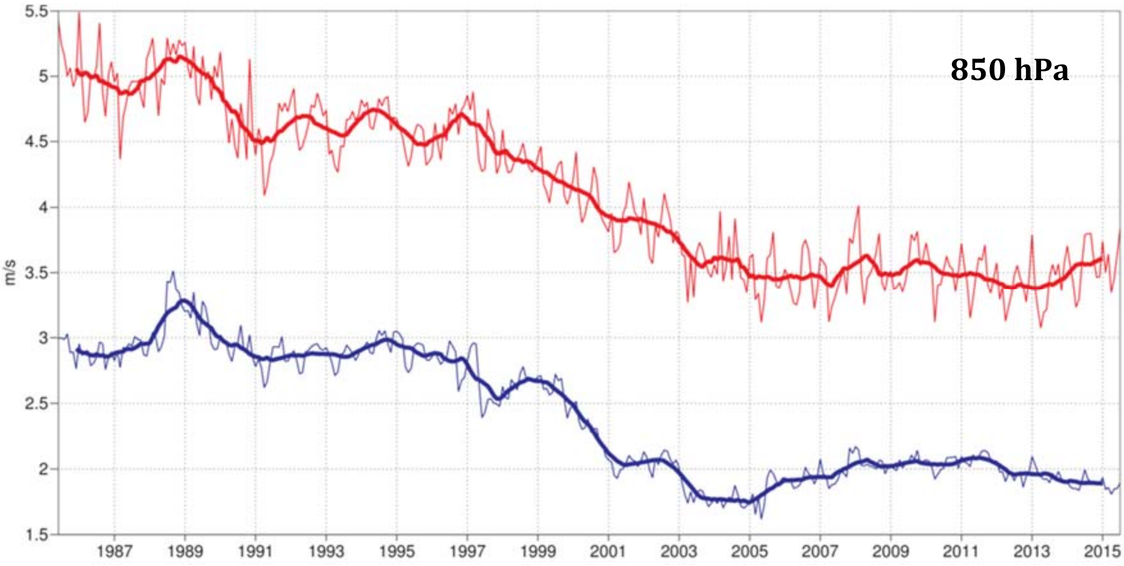
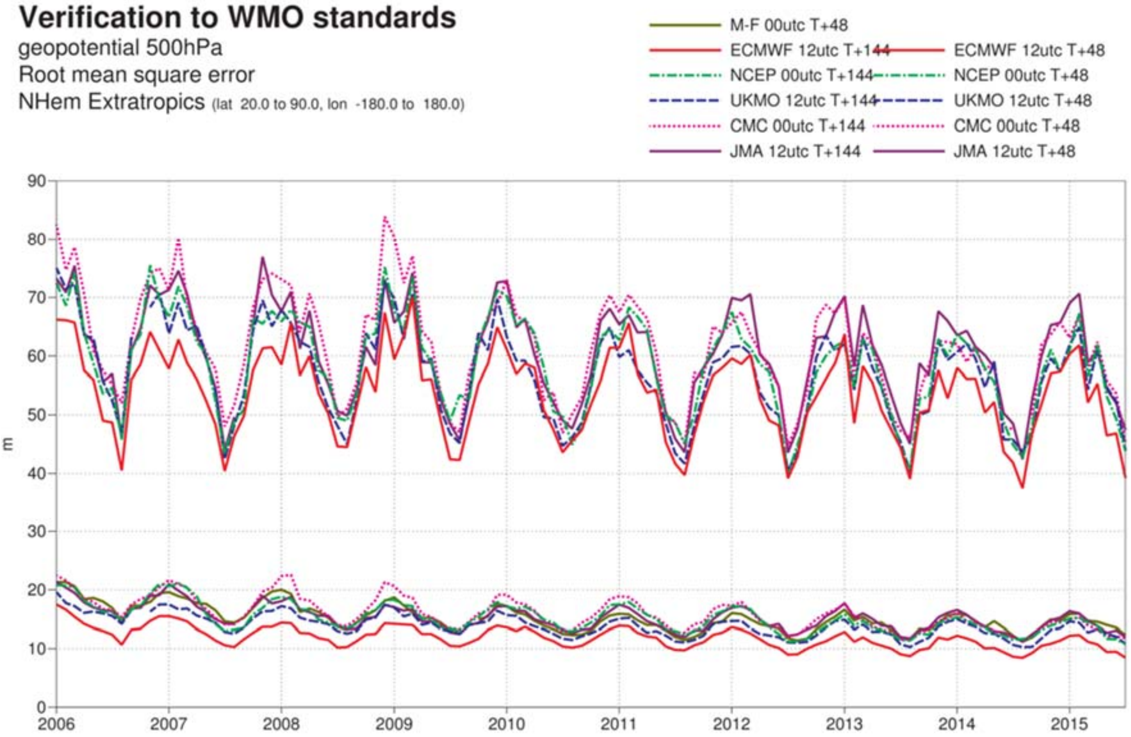


Figure 12: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).



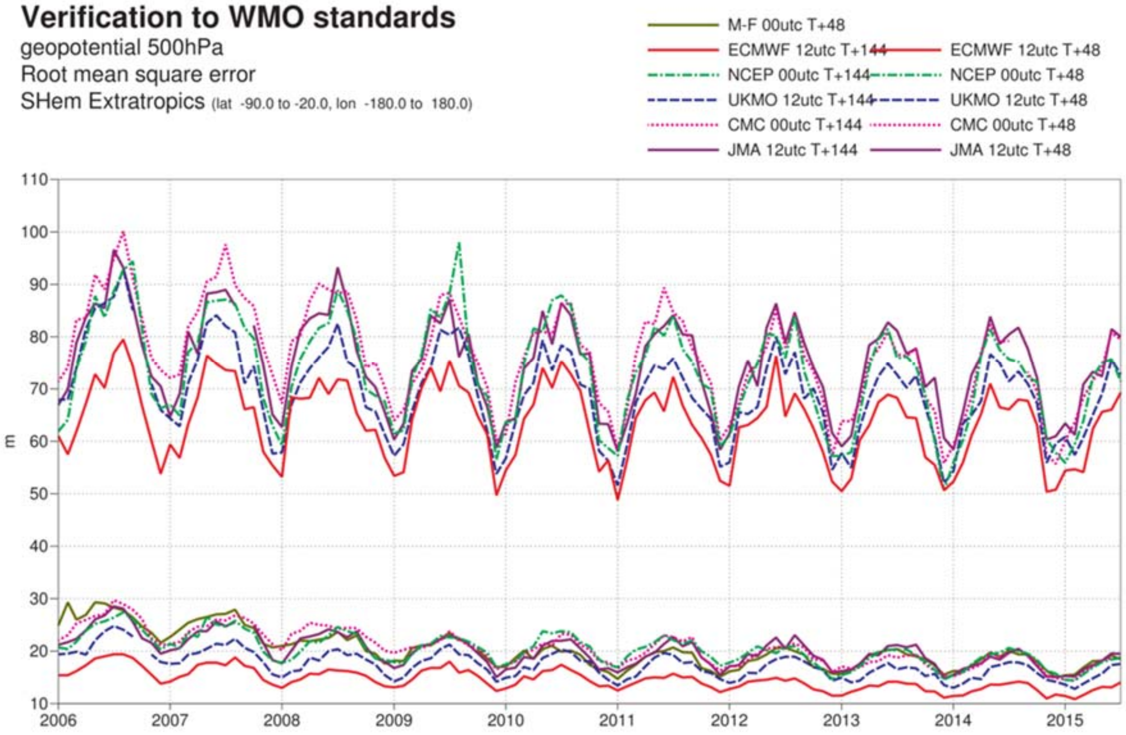
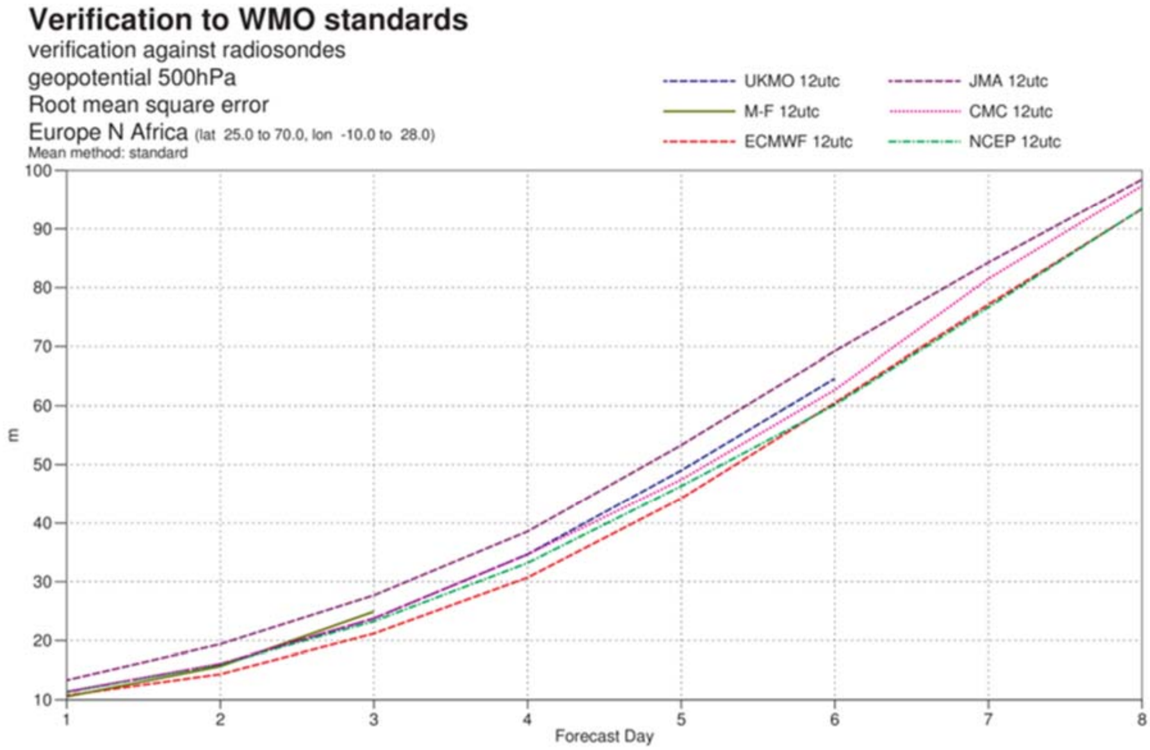


Figure 13: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top) and southern (bottom) extratropics. In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model is verified against its own analysis. JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, NCEP = U.S. National Centers for Environmental Prediction, M-F = Météo France.



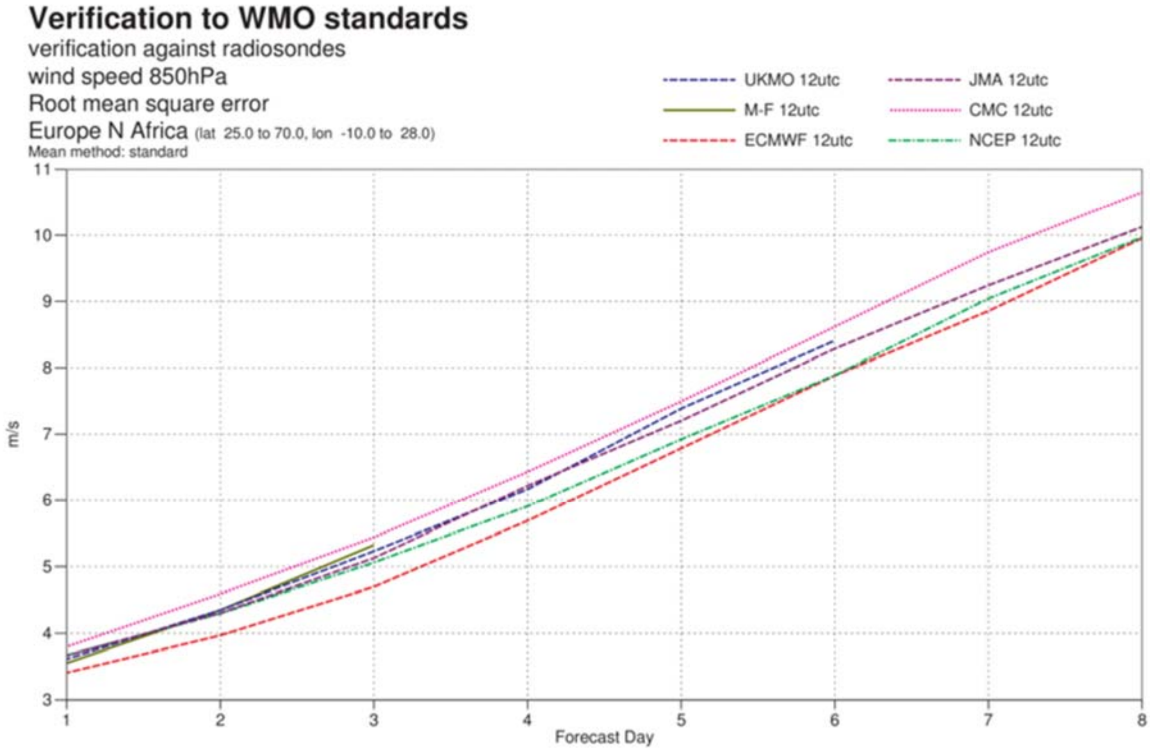
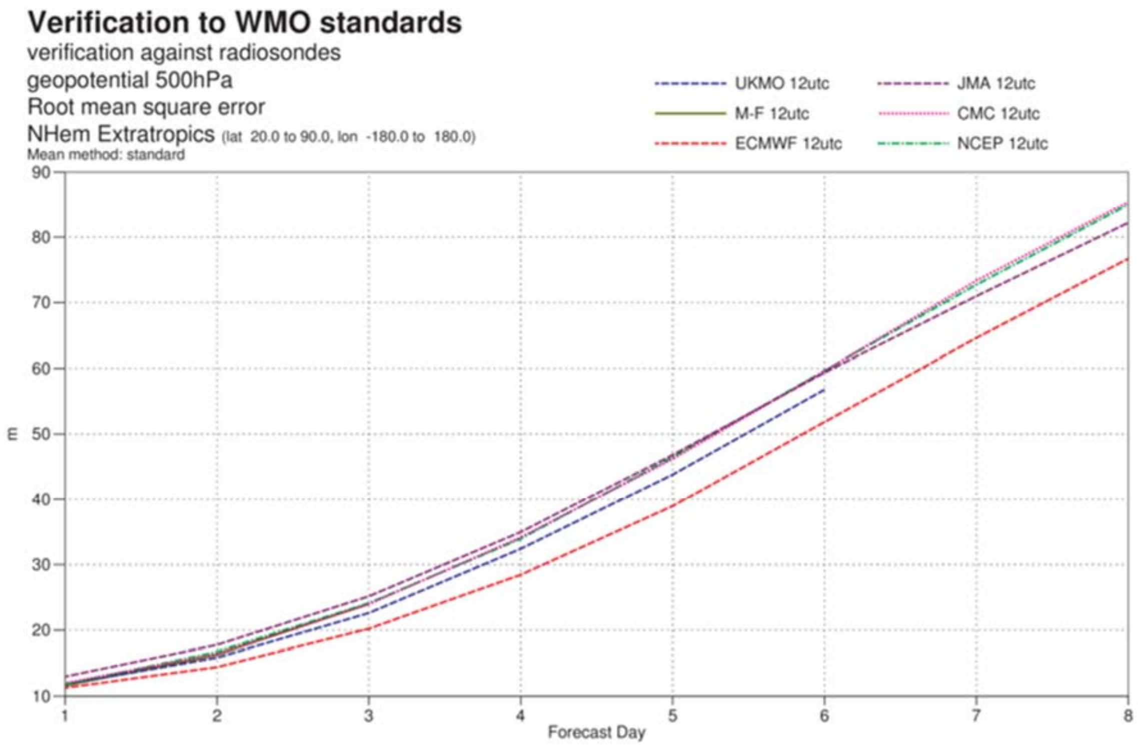


Figure 14: WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2014–July 2015).



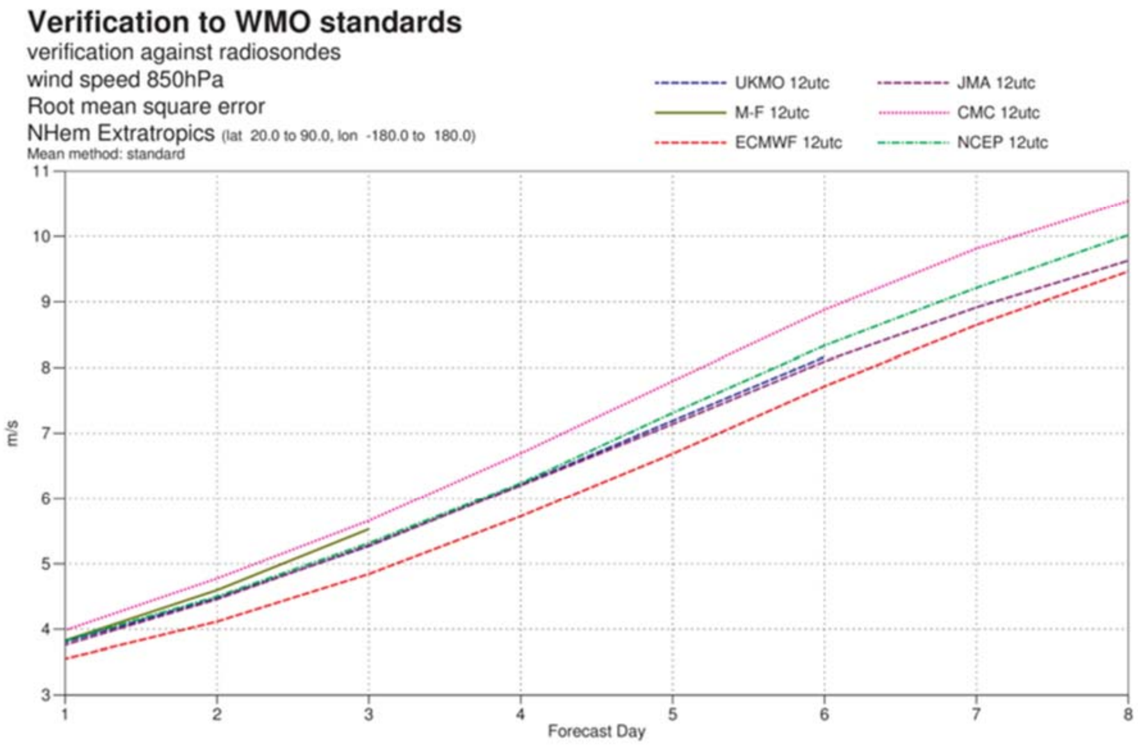
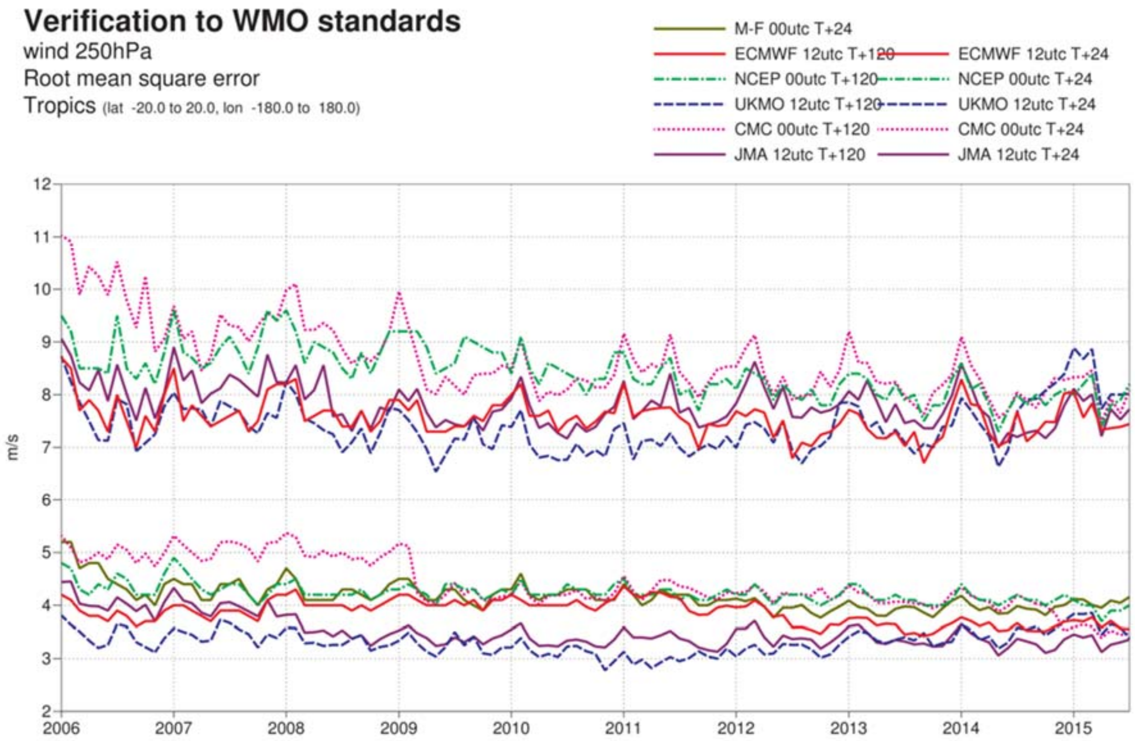


Figure 15: As Figure 14 for the northern hemisphere extratropics.



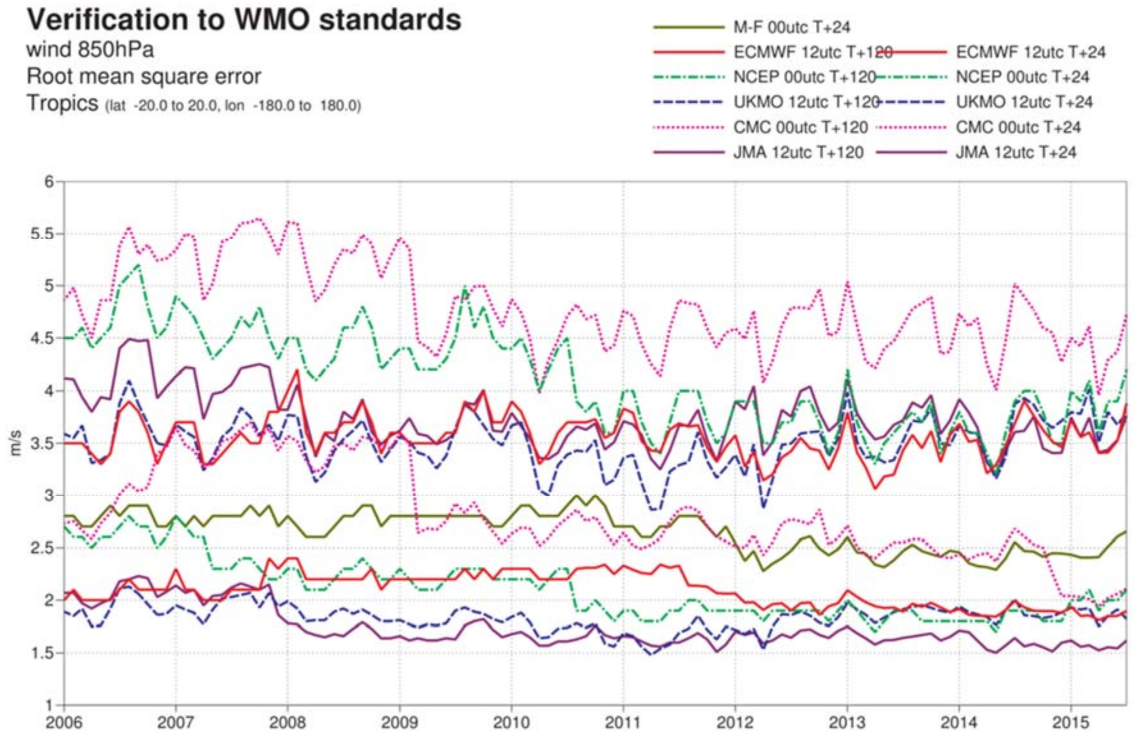
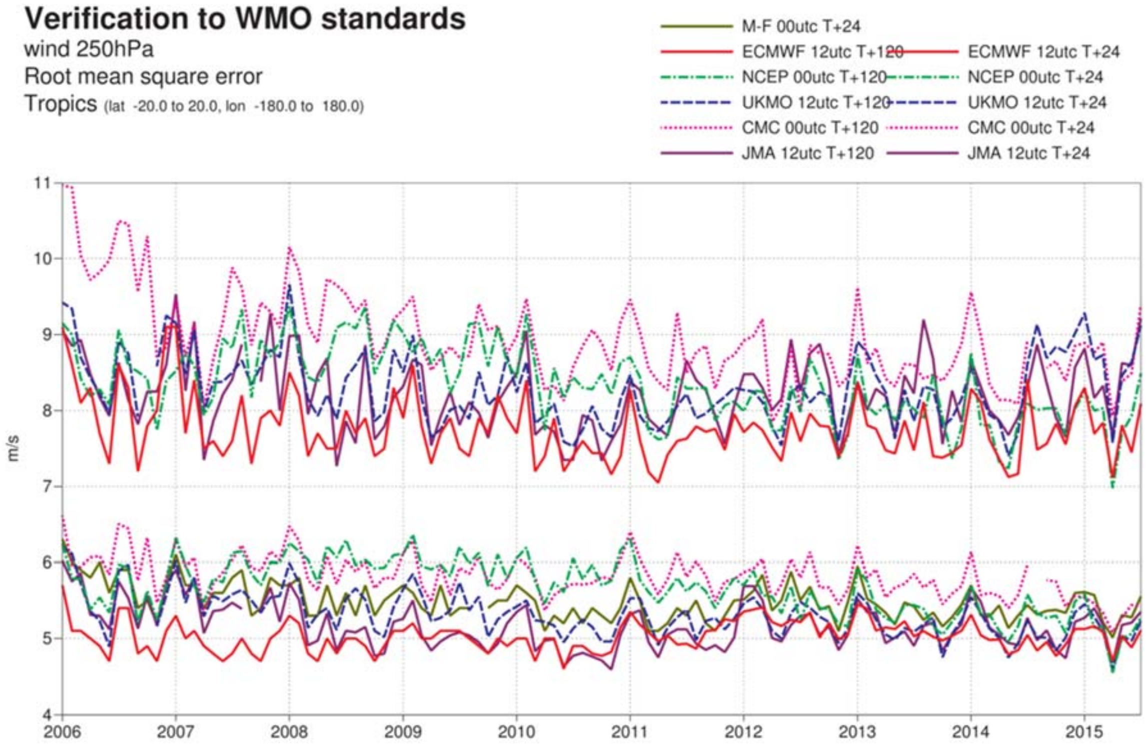


Figure 16: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



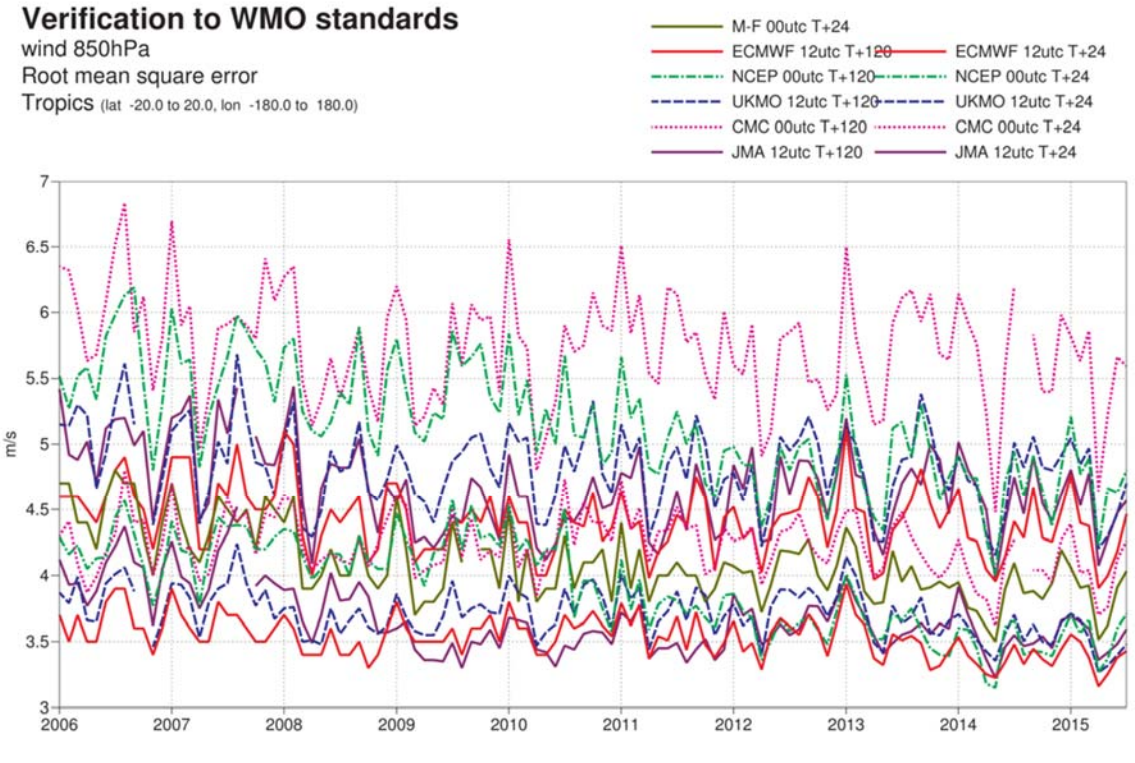


Figure 17: As Figure 16 for verification against radiosonde observations.

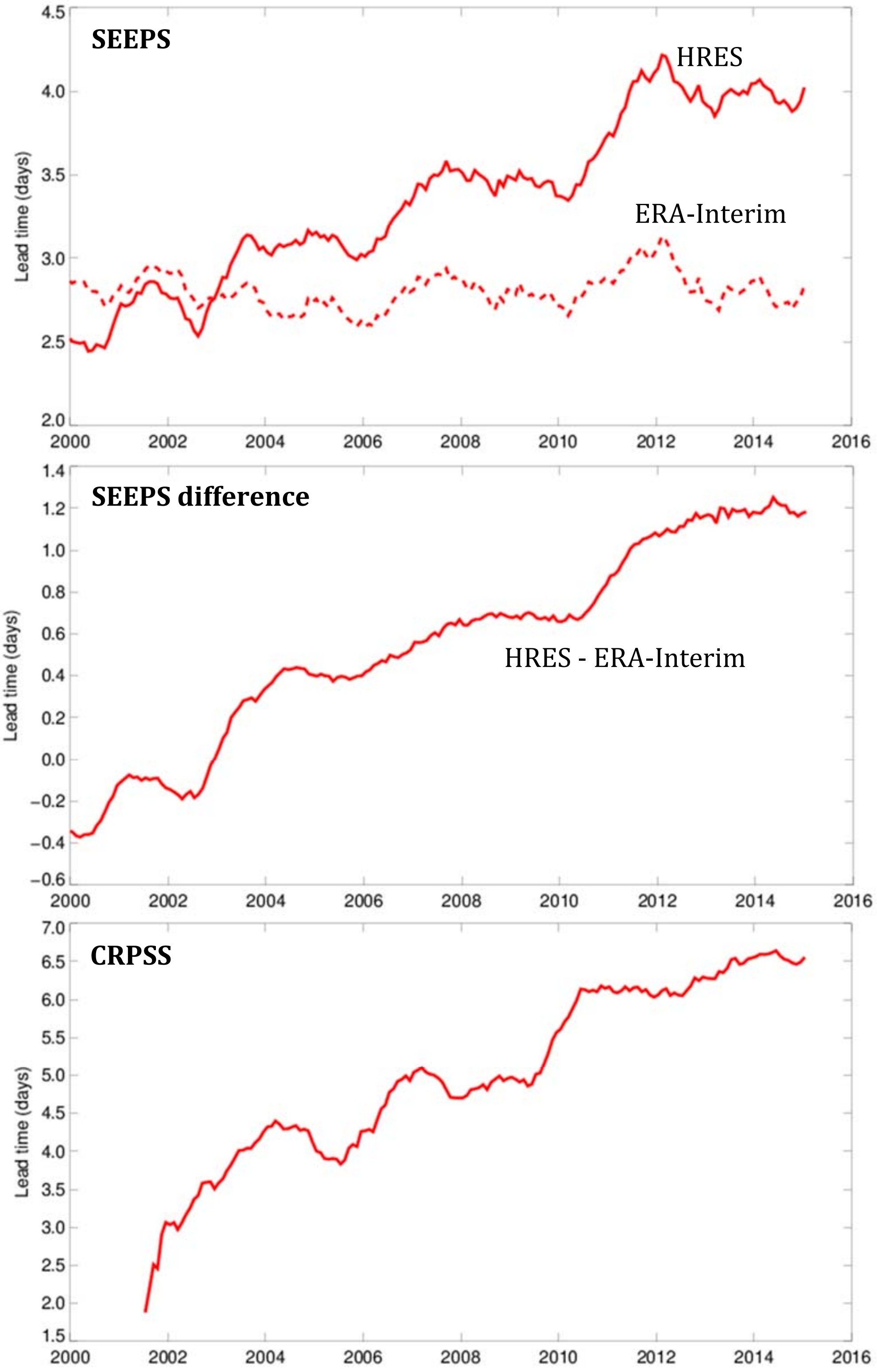


Figure 18: Supplementary headline scores for deterministic (top, centre) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics; each point is calculated over a 12-month period, plotted at the centre of the period. The dashed curve shows the deterministic headline score for ERA-Interim as a reference. The centre panel shows the difference between the operational forecast and ERA-Interim. Curves show the number of days for which the centred 12-month mean skill remains above a specified threshold. The forecast day on the y-axis is the end of the 24-hour period over which the precipitation is accumulated.

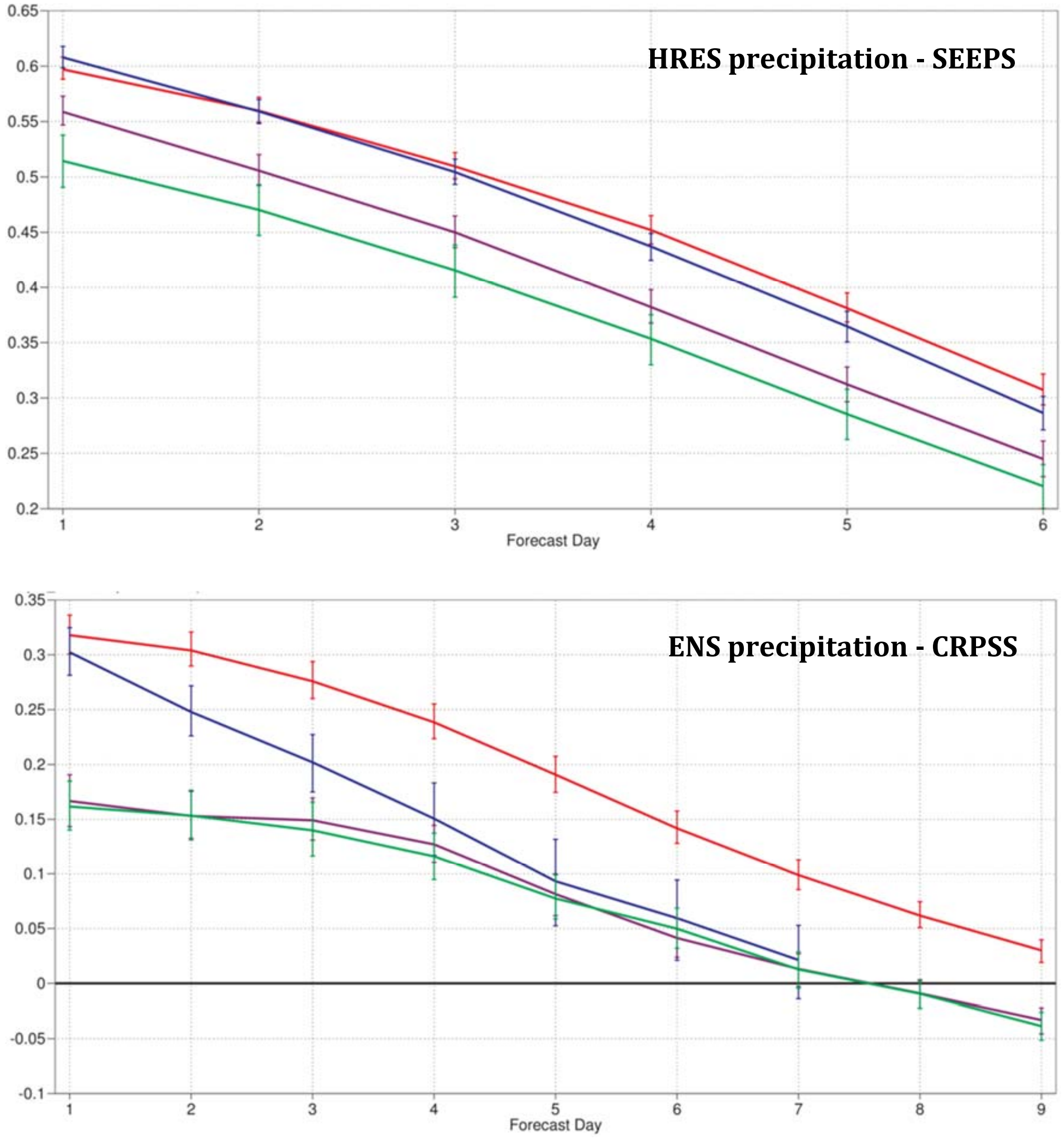


Figure 19: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 18. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2014–July 2015. Bars indicate 95% confidence intervals.

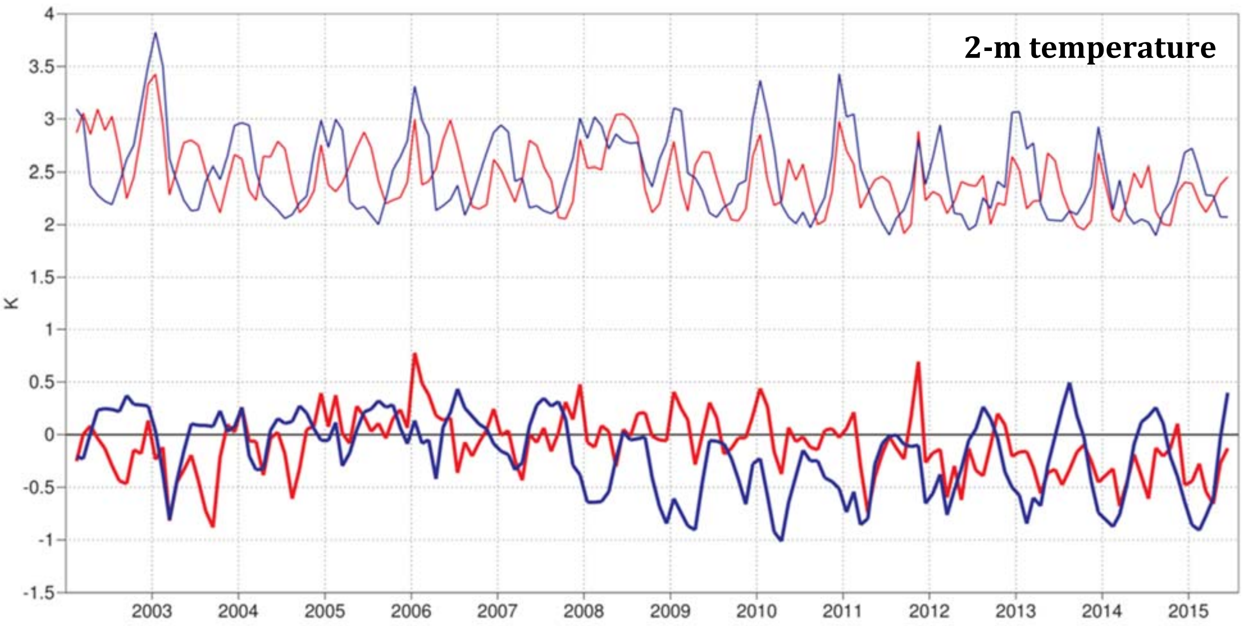


Figure 20: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

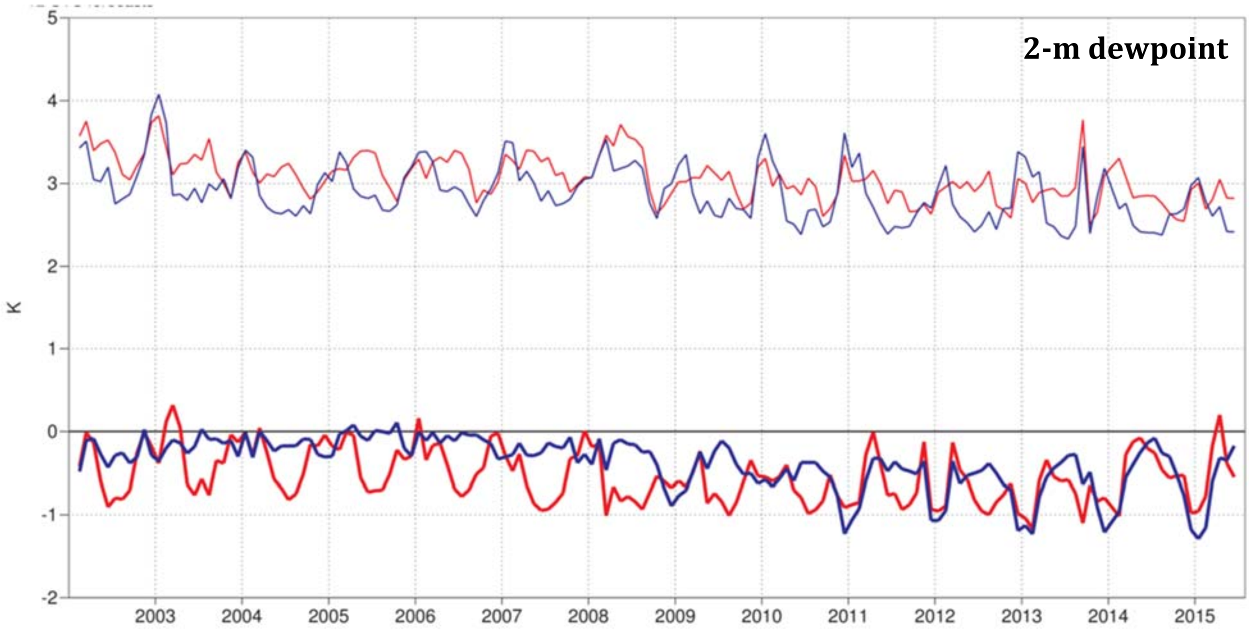


Figure 21: Verification of 2 m dewpoint forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

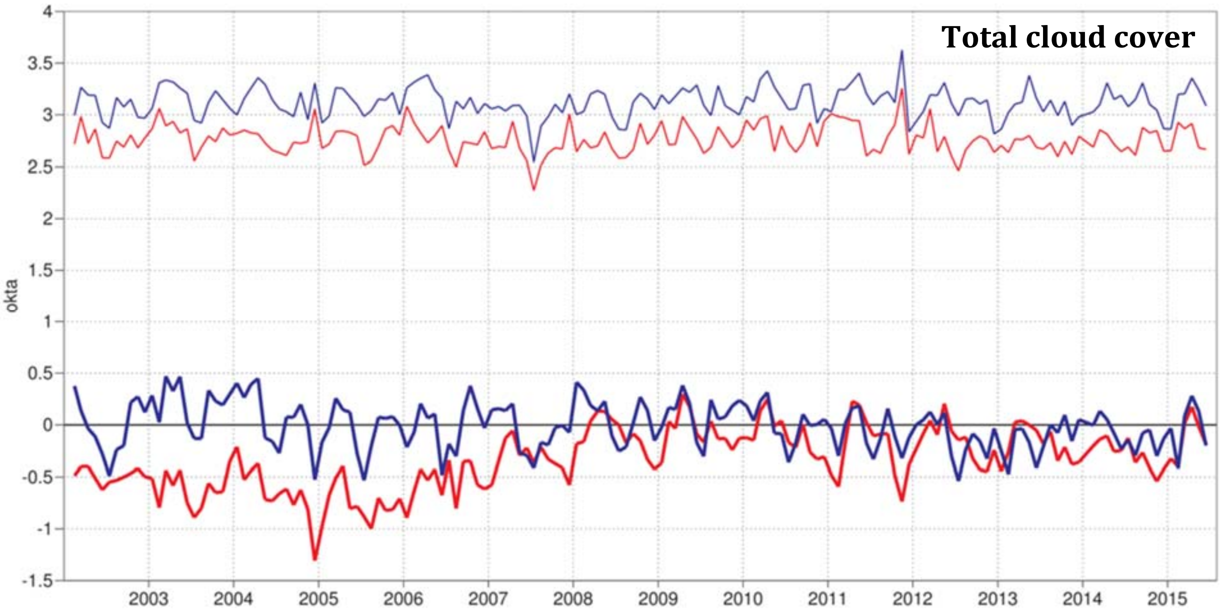


Figure 22: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

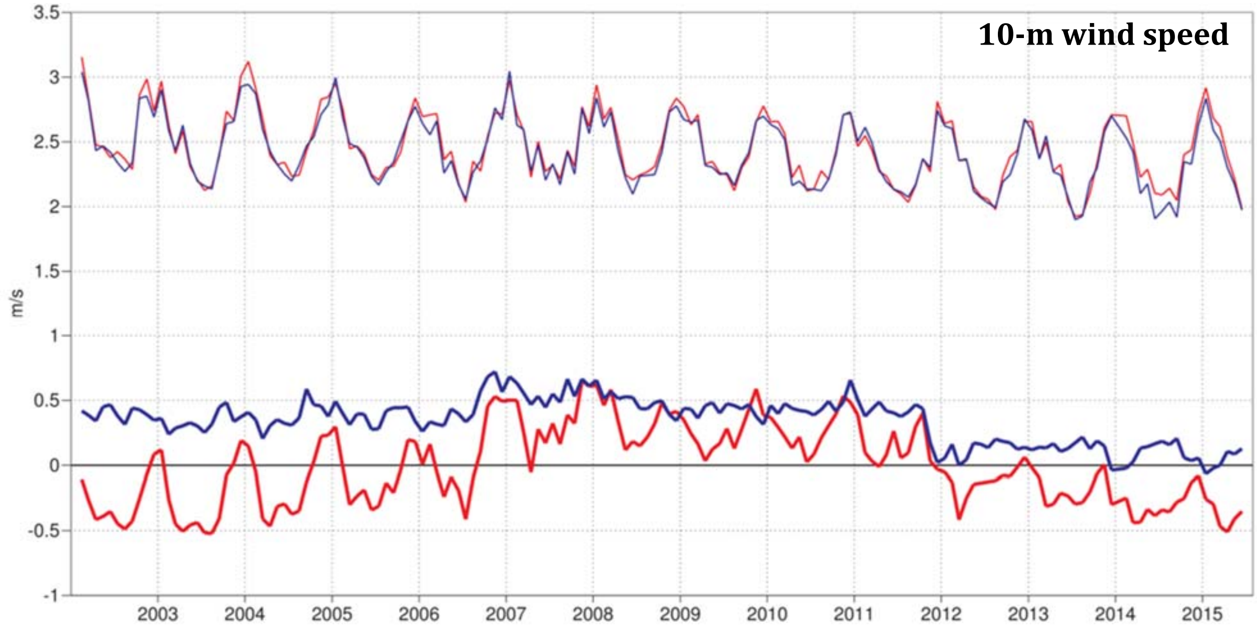


Figure 23: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

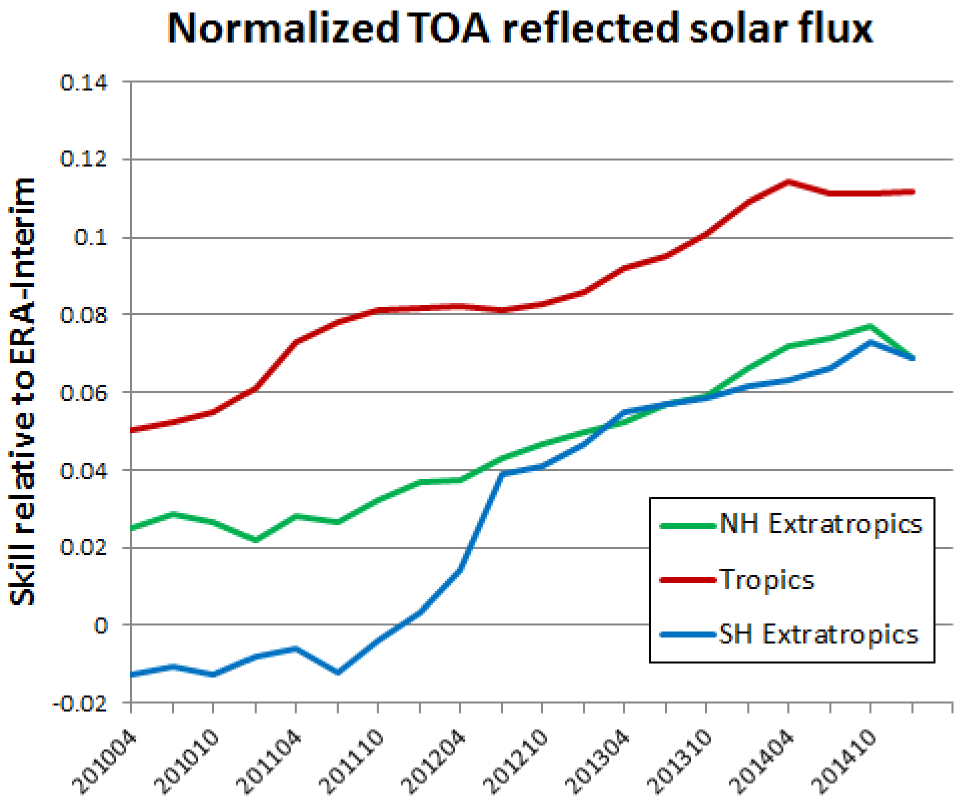


Figure 24: 12-month running average of the day 3 forecast skill relative to ERA-Interim of normalized TOA reflected solar flux (daily totals), verified against satellite data. The verification has been carried out for those parts of the northern hemisphere extratropics (green), tropics (red), and southern hemisphere extratropics (blue) which are covered by the CM-SAF product (approximately 70 S to 70 N, and 70 W to 70 E).

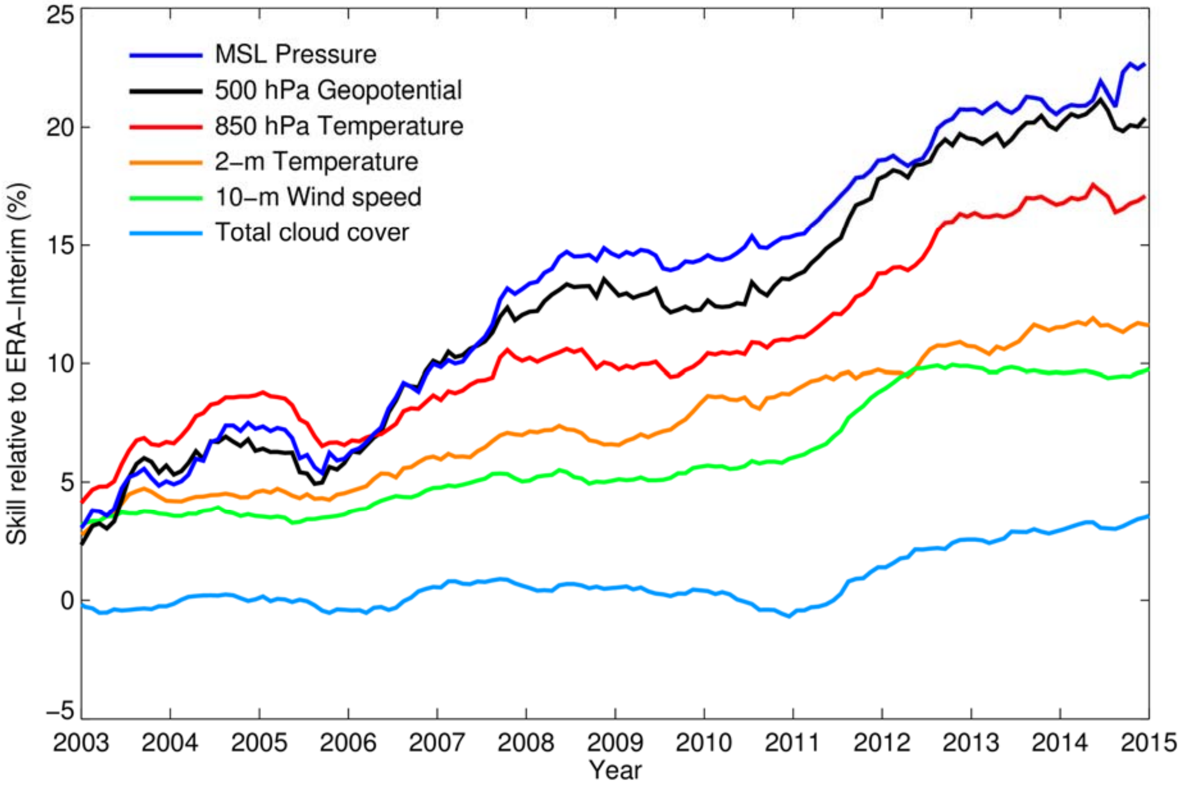


Figure 25: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA-Interim. Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), and mean sea level pressure (MSLP), using RMSE as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC), using error standard deviation as a metric.

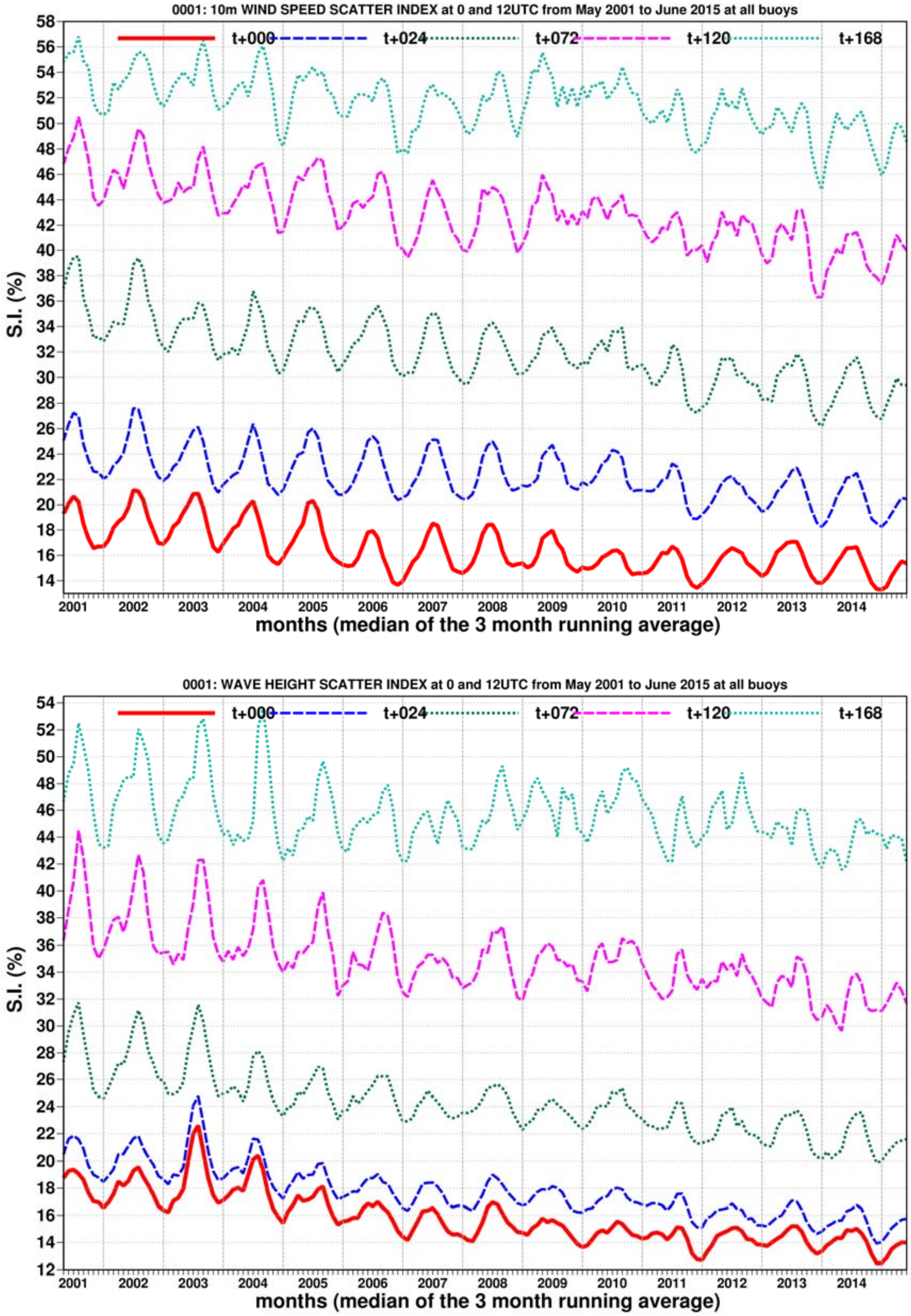
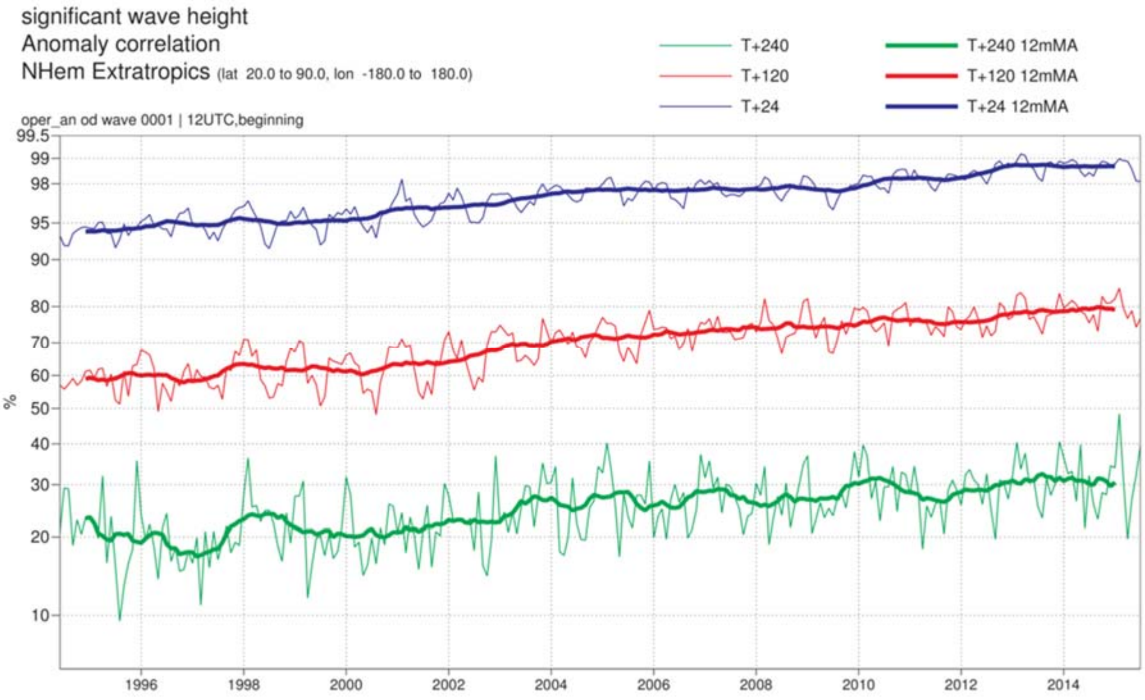


Figure 26: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.



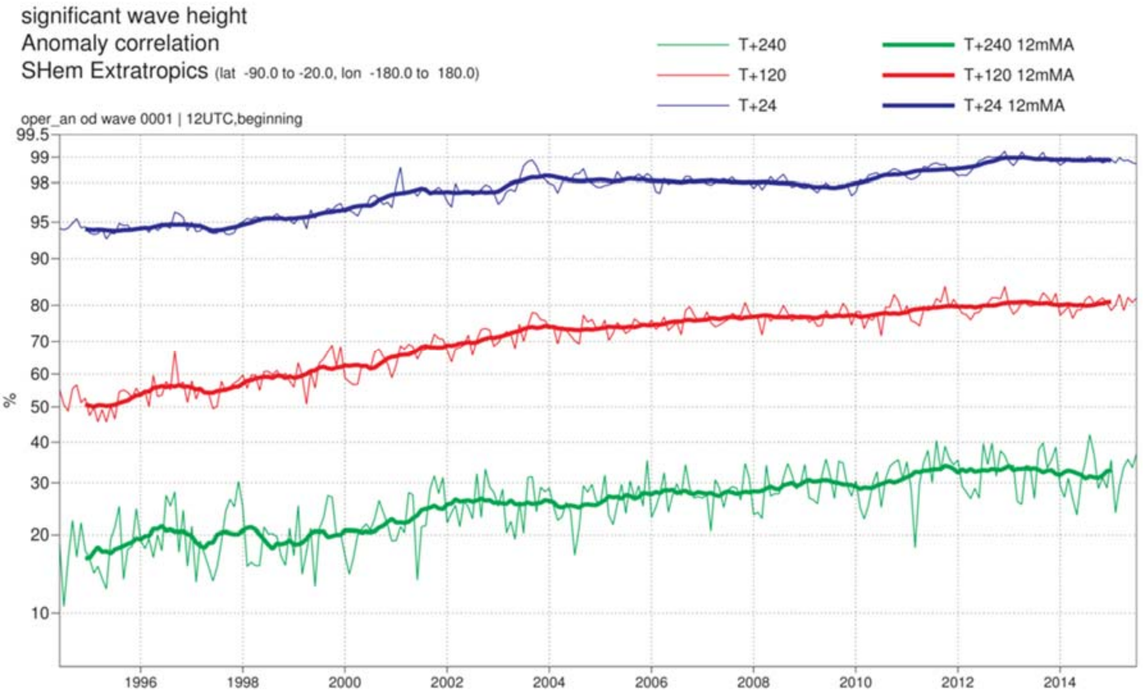


Figure 27: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

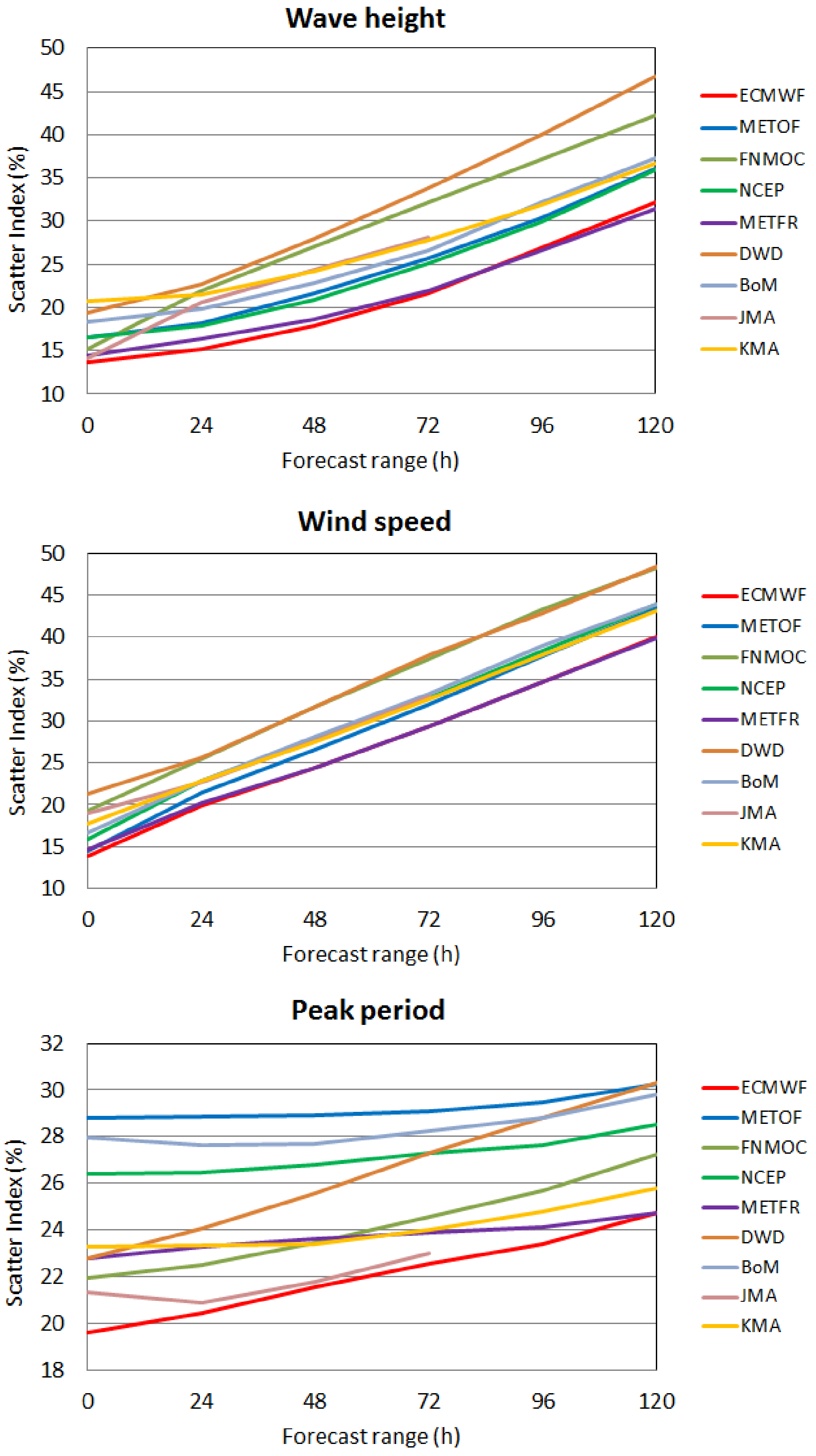


Figure 28: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period July 2014–June 2015. 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; JMA: Japan Meteorological Agency; KMA: Korea Meteorological Administration.

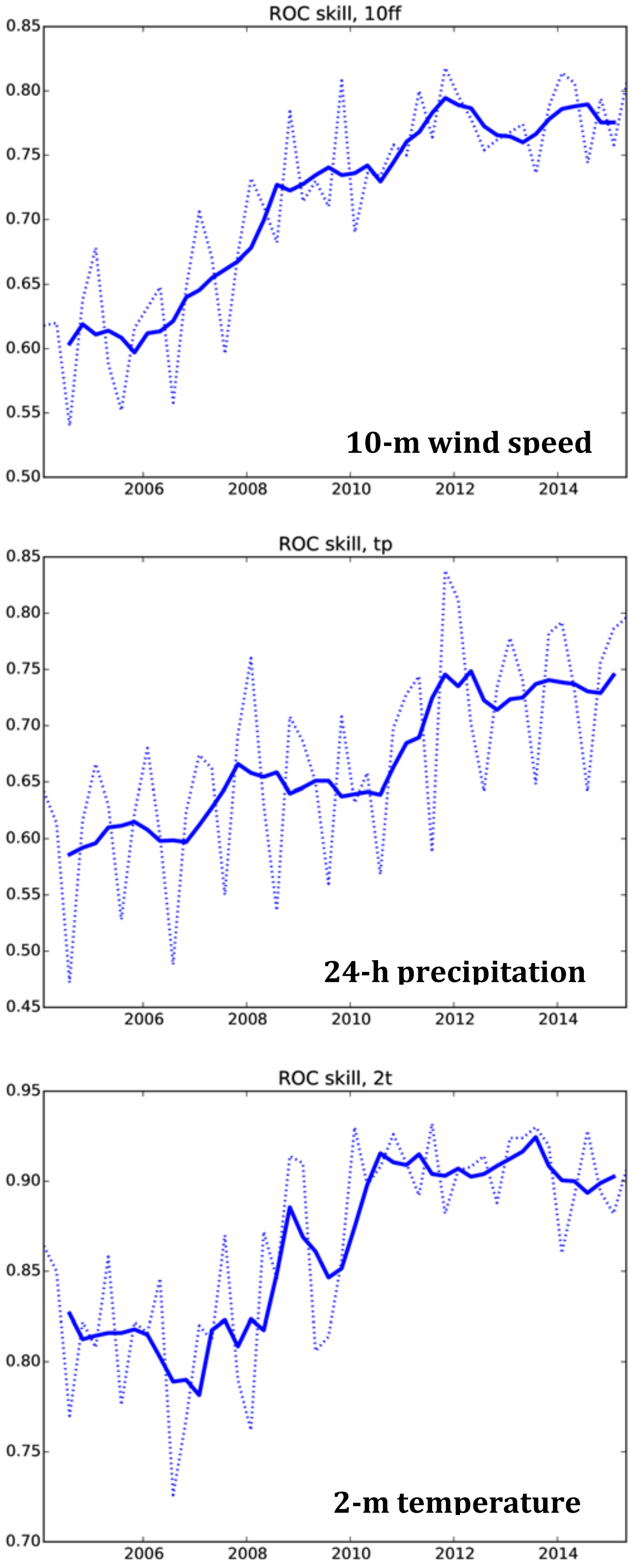


Figure 29: Verification of Extreme Forecast Index (EFI) against analysis. Top panel: supplementary headline score – skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate. Curves show seasonal values (dotted) and four-season running mean (continuous) of relative operating characteristic (ROC) area skill scores. Centre and bottom panels show the equivalent ROC area skill scores for precipitation EFI forecasts and for 2 m temperature EFI forecasts.

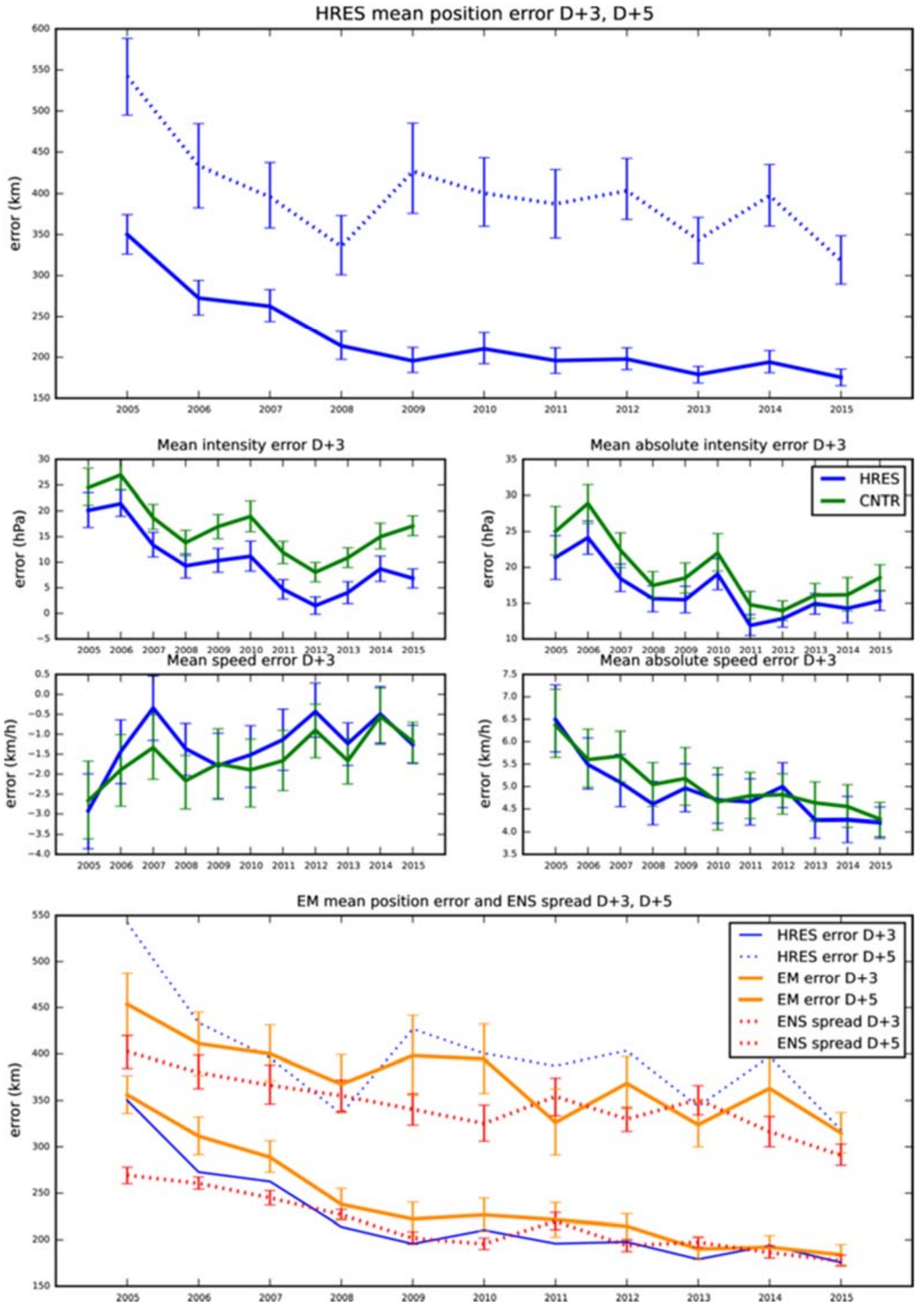


Figure 30: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported via the GTS. Top panel supplementary 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).

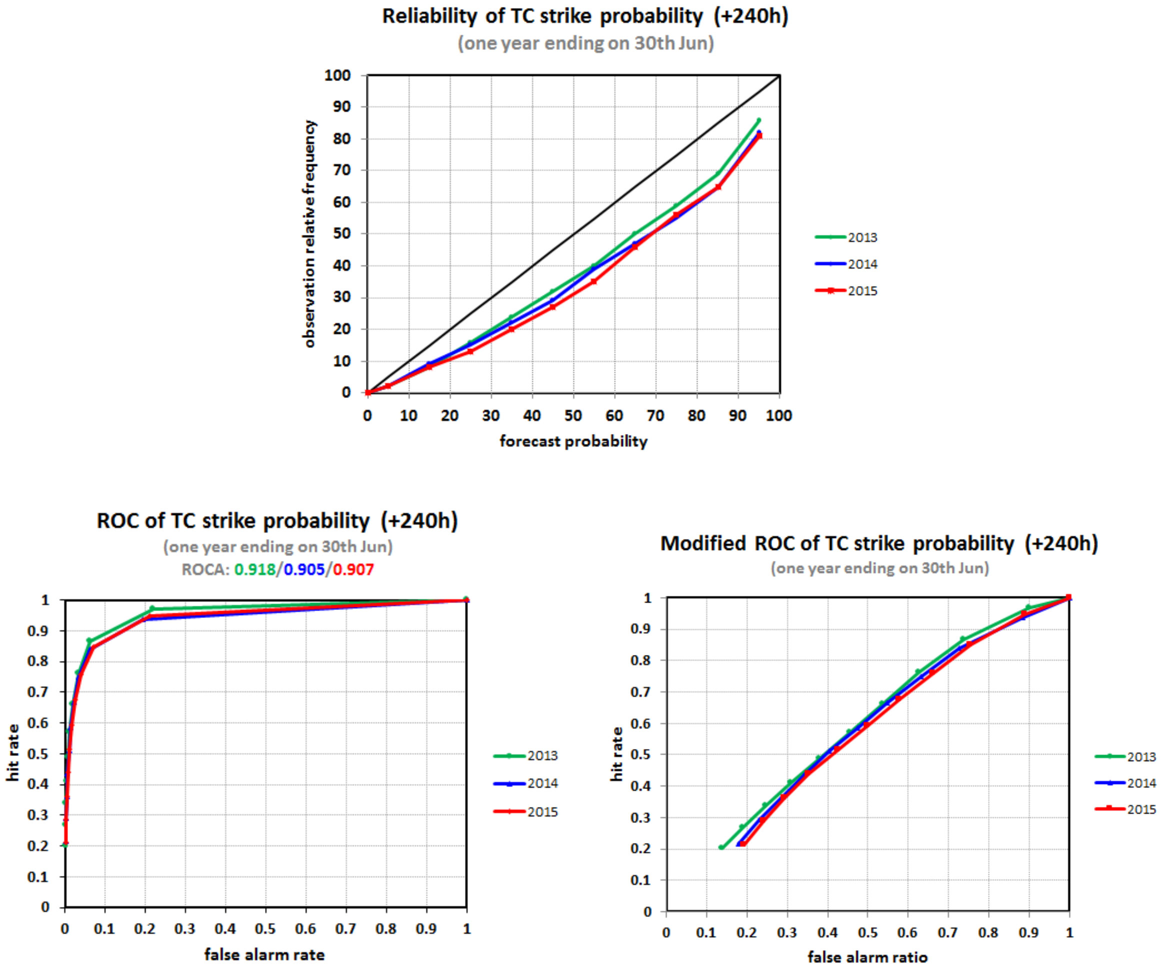


Figure 31: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2012–June 2013 (green), July 2013–June 2014 (blue) and July 2014–June 2015 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) the standard ROC diagram and (right) a modified ROC diagram, where the false alarm ratio is used instead of the false alarm rate. 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.

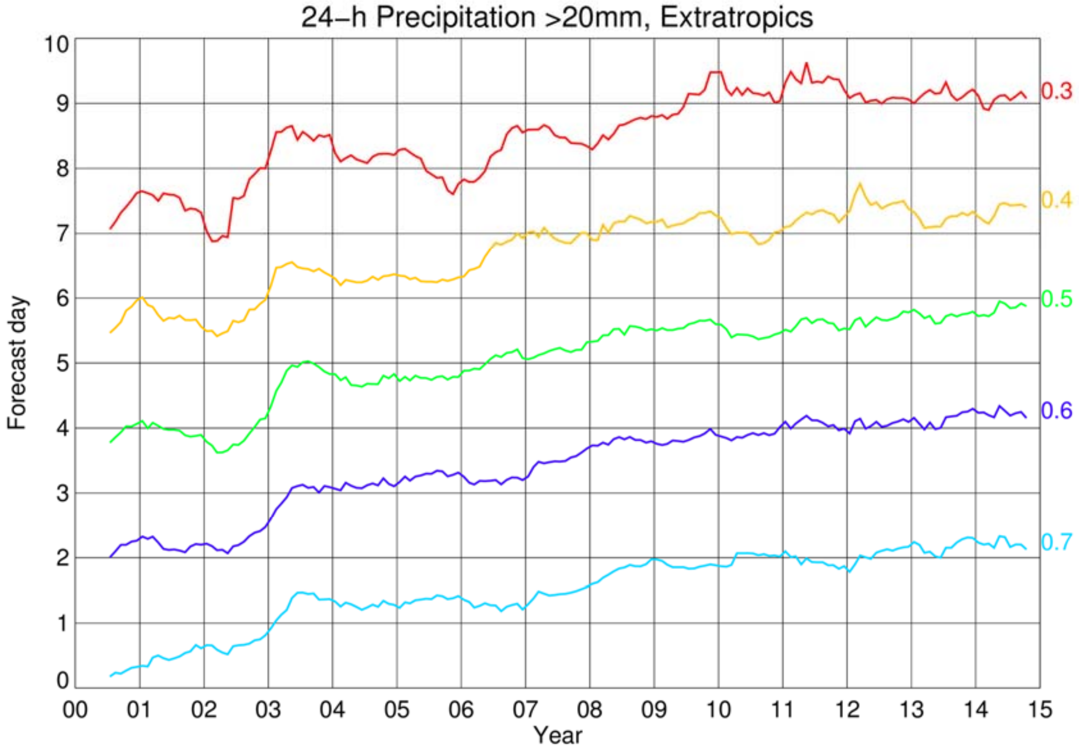


Figure 32: Evolution of skill of the HRES forecast in predicting 24-h precipitation amounts >20 mm in the extra-tropics as measured by the SEDI score, expressed in terms of forecast days. Verification is against SYNOP observations. Numbers on the right indicate different SEDI thresholds used. Curves show 12-month running averages.

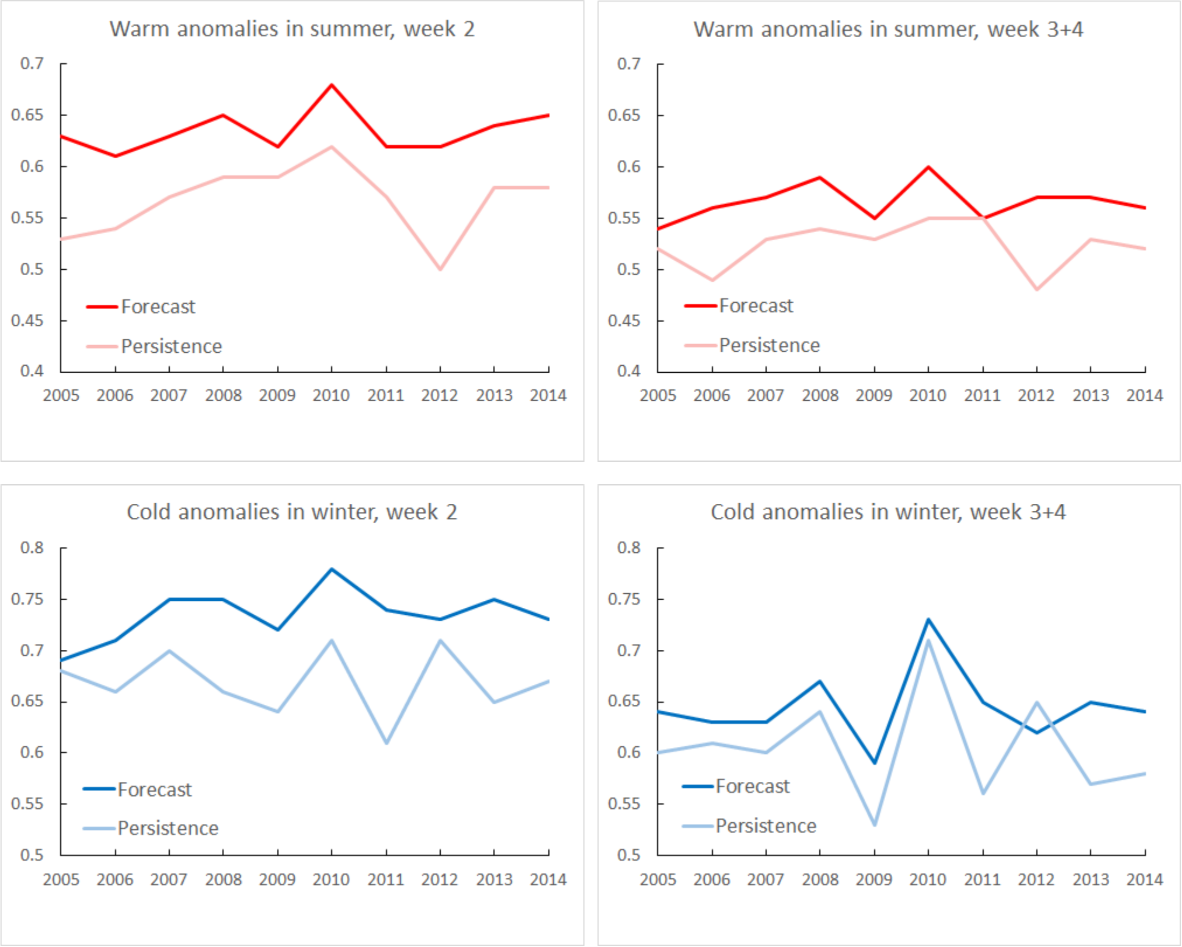


Figure 33: Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all land points in the extra-tropical northern hemisphere. Left panels show the score of the operational monthly forecasting system for forecast days 12–18 (7-day mean), and right panels for forecast days 19–32 (14-day mean). As a reference, lighter coloured lines shows the score using persistence of the preceding 7-day or 14-day period of the forecast.

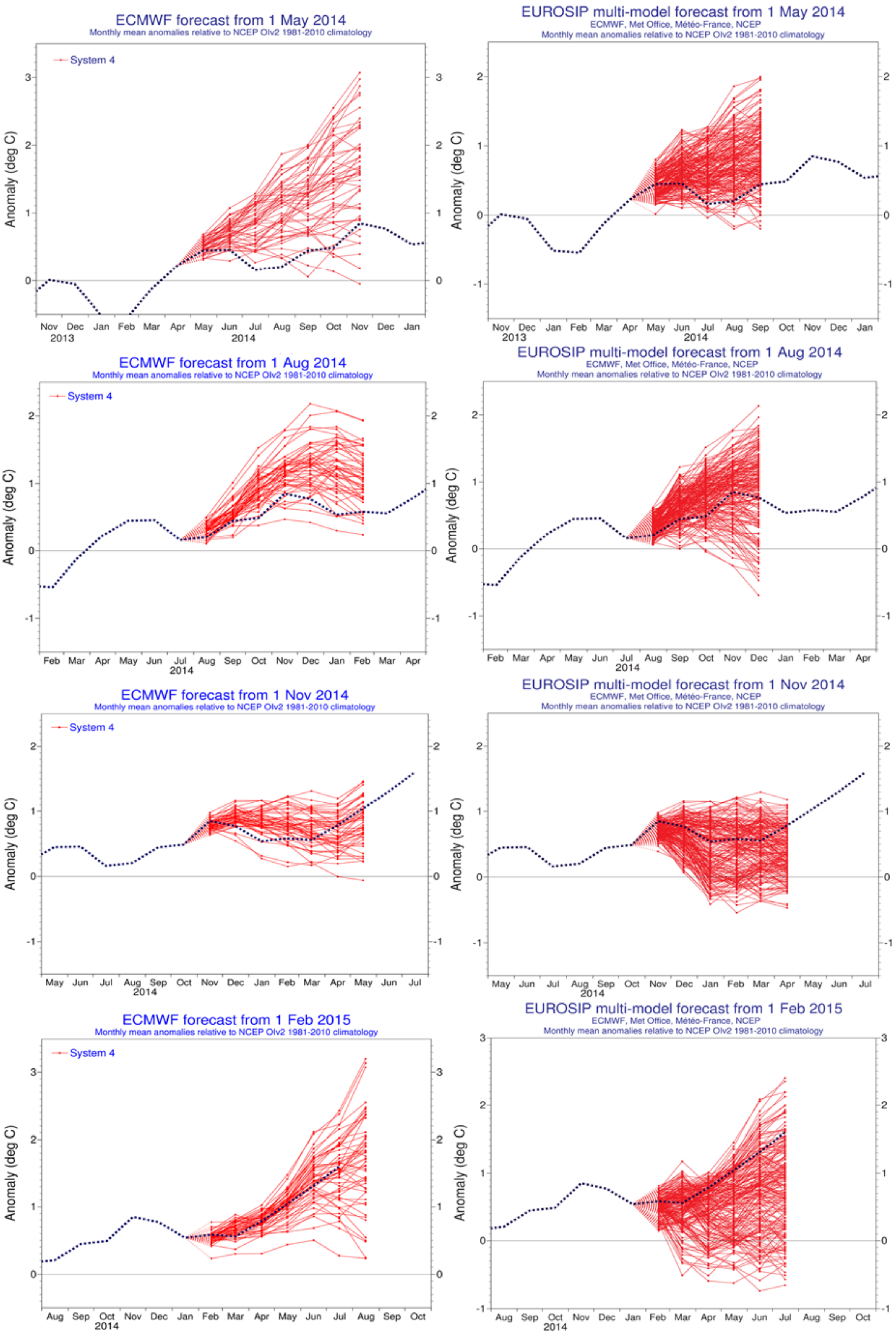


Figure 34: ECMWF (left column) and EUROSIP multi-model forecast (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2014, August 2014, November 2014 and February 2015. The red lines represent the ensemble members; dotted blue line shows the subsequent verification.

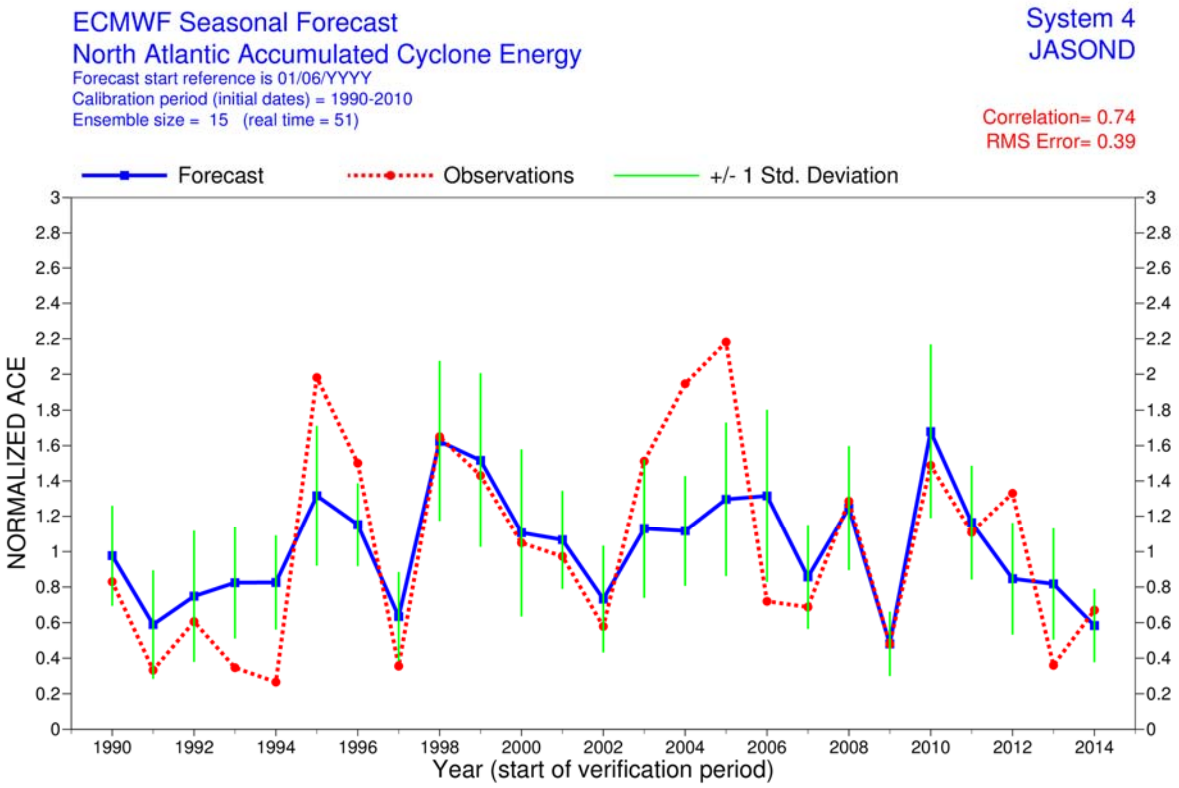


Figure 35: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2014. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts 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.

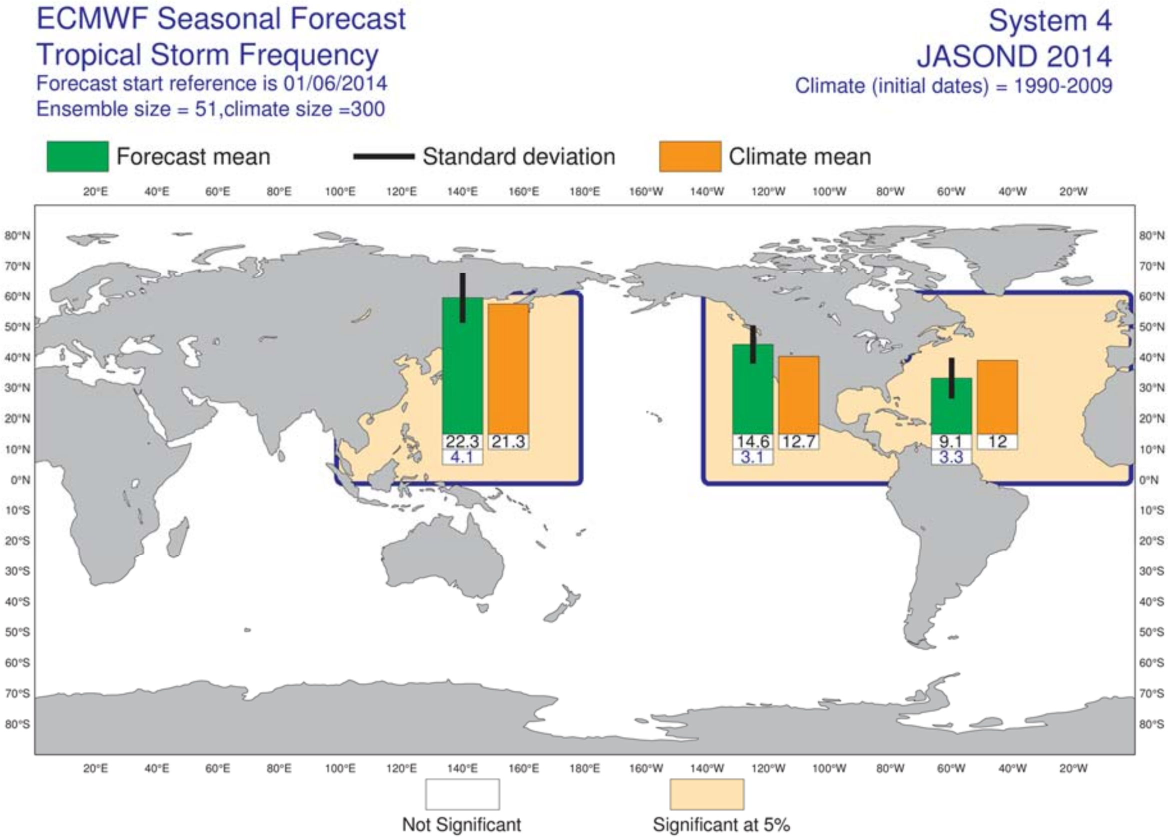


Figure 36: Tropical storm frequency forecast issued in June 2014 for the six-month period July–December 2014. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The 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.

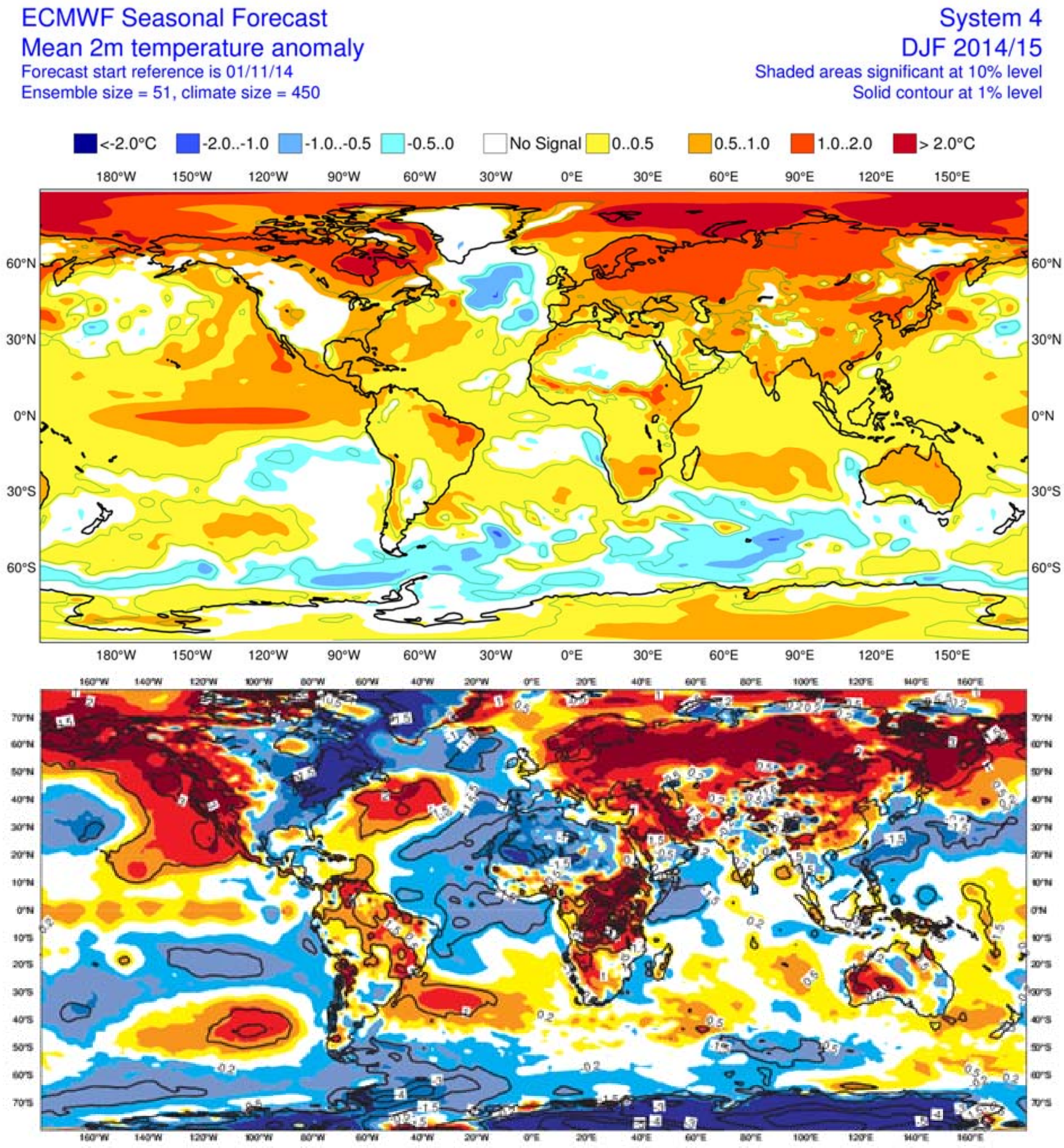


Figure 37: Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2014 for DJF 2014/15 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

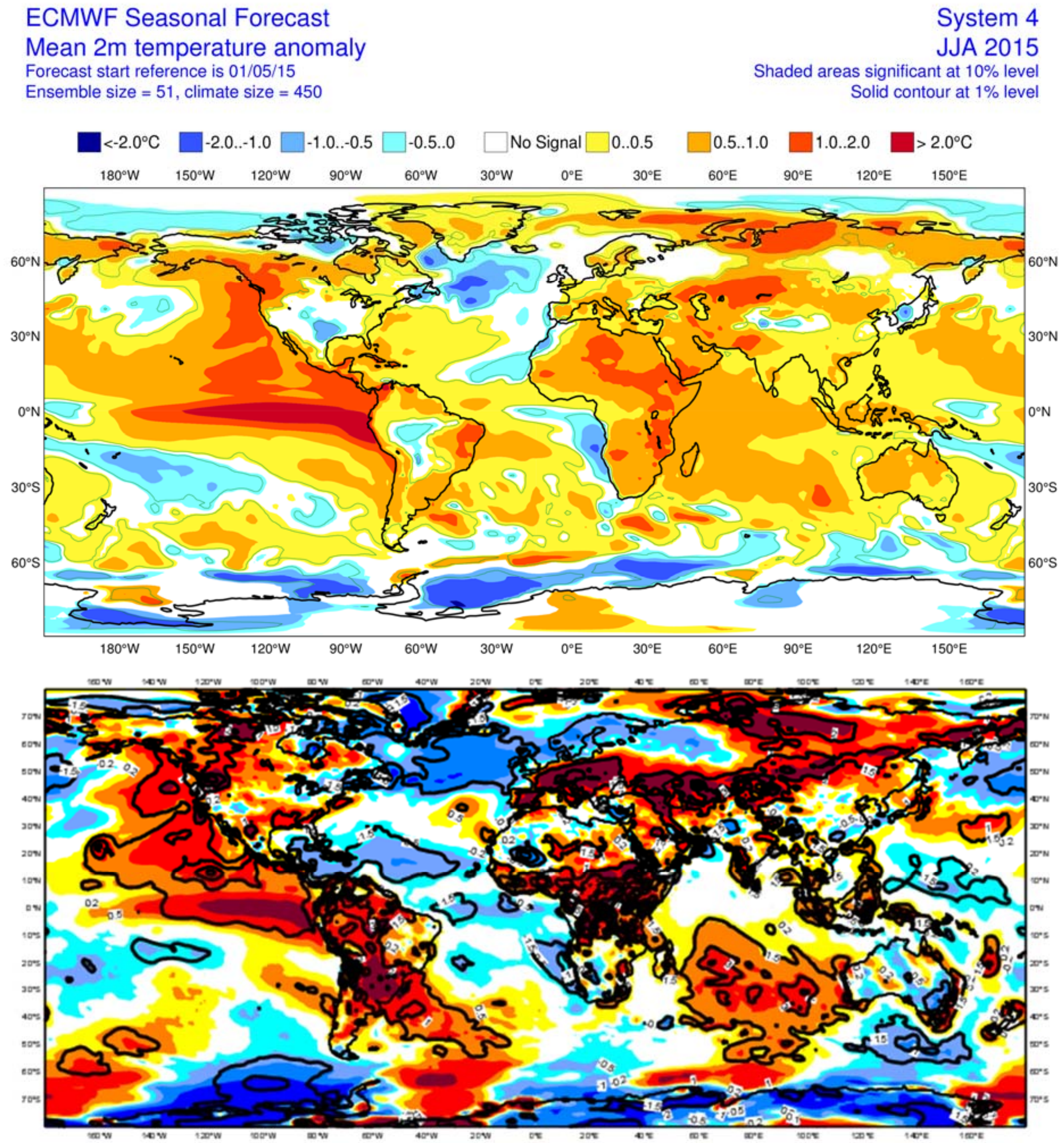


Figure 38: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2015 for JJA 2015 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

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 14), 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 14, Figure 16) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

Skill scores 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 shows 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 27) 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 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 31). Figure 31 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 33.

A. 3 Weather parameters (Section 4)

Verification of the deterministic precipitation forecasts is made using the newly developed SEEPS score (Rodwell et al., 2010). SEEPS (stable equitable error in probability space) uses three categories: dry, light precipitation, and heavy precipitation. Here “dry” is defined, with reference to WMO guidelines for observation reporting, to be any accumulation (rounded to the nearest 0.1 mm) that is less than or equal to 0.2 mm. To ensure that the score is applicable for any climatic region, the “light” and “heavy” categories are defined by the local climatology so that light precipitation occurs twice as often as heavy precipitation. A global 30-year climatology of SYNOP station observations is used (the resulting threshold between the light and heavy categories is generally between 3 and 15 mm for Europe, depending on location and month). SEEPS is used to compare 24-hour accumulations derived from global SYNOP observations (exchanged over the Global Telecommunication System; GTS) with values at the nearest model grid-point. 1-SEEPS is used for presentational purposes (Figure 18, Figure 19) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 18, Figure 19). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 20 to Figure 23), 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

Experimental verification of deterministic forecasts of rare events is performed using the symmetric extremal dependence index SEDI (Figure 32), 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.

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.