TECHNICAL MEMORANDUM

792

Evaluation of ECMWF forecasts, including the 2016 resolution upgrade

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

December 2016

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

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 consistent with that of previous years (ECMWF Tech. Memos. 346, 414, 432, 463, 501, 504, 547, 578, 606, 635, 654, 688, 710, 765). A short technical note describing the scores used in this report is given in the annex to this document.

Verification pages are regularly updated, and 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

On 8 March 2016, ECMWF upgraded the horizontal resolution of its analyses and forecasts. The upgrade has a horizontal resolution that translates to about 9 km for HRES and the data assimilation (the outer loop of the 4D-Var) and to about 18 km for the ENS up to day 15. The resolution of the ENS extended (day 16 up to day 46) is about 36 km. The new cycle is labelled 41r2 and includes a number of enhancements to the model and data assimilation listed below.

2.1 Resolution upgrade

The detailed specification of the resolution upgrades included in IFS cycle 41r2 are:

- Introduction of a new form of the reduced Gaussian grid, the octahedral grid, for HRES, ENS and ENS Extended;
- Horizontal resolution of the HRES increased from TL1279 / N640 to TCO1279 / O1280, where subscript C stands for cubic and O for octahedral, with a model time step of 450s;

- Horizontal resolution of the ENS increased from TL639 / N320 to TCO639 / O640 for ENS (Days 0 - 15) with a model time step of 720s and from TL319 / N160 to TCO319 / O320 for ENS Extended (Days 16 - 46) with a model time step of 1200s;
- For the medium-range ENS there will no longer be a decrease of resolution at day 10: the ENS Days 11 - 15 will be run at the same TCO639 / O640 resolution as ENS Days 0 - 10;
- Increase of the HRES-WAM resolution from 0.25 to 0.125 degrees and the ENS-WAM Days 0 - 15 from 0.5 to 0.25 degrees;
- Horizontal resolution of the EDA outer loop is increased from TL399 to TCO639 with its two inner loops increased from TL159 / TL159 to TL191 / TL191, respectively;
- Horizontal resolution of the three 4DVar inner loops is increased from TL255 / TL255 / TL255 to TL255 / TL319 / TL399, respectively.

2.2 Meteorological content of IFS cycle 41r2

2.2.1 Data assimilation

- Compute scale-dependent hybrid B (background error covariance) by adding samples from latest EDA forecast to static climatological B with increasing weight of today's EDA for smaller wavelengths (30% up to T63, growing to a maximum 93% at T399).
- The EDA now cycles its own background error and covariance estimates, rather than using climatological estimates.
- Change to use the Sonntag equation for saturation vapour pressure in humidity observation operators to improve saturation calculation for very cold temperatures (colder than -40C).
- Assimilation of aircraft humidity data implemented.

2.2.2 Satellite

- GPSRO (radio occultation) observation errors based on a physical error propagation model are increased by 25% to account for missing sources of error (e.g. observation error correlations, forecast model error). Improves lower stratosphere/tropopause winds and temperatures.
- Activated SSMIS F-18 humidity sounding channels over ocean and extended all-sky assimilation to snow covered land surfaces.
- Improved specification of AMSU-A observation errors based on satellite (due to instrument noise characteristics and ageing) and situation (cloud, orography) thereby increasing the number of observations assimilated.
- Improved aerosol detection and screening for IASI infrared satellite data.
- Increased use of Atmospheric Motion Vectors (AMVs), including extension in latitudinal coverage from geostationary platforms from 60 to 64 degrees zenith angle and addition of Meteosat mid-height AMVs derived from infrared imagery.

- Revised data selection (screening) of cold-air outbreaks in low-peaking all-sky microwave channels to allow more data to be assimilated.
- Updated microwave observation operator coefficient files (54-level RTTOV files with latest spectroscopy)

2.2.3 Numerics

- Changed from a linear reduced Gaussian to a cubic octahedral reduced Gaussian grid for HRES, ENS and 4DVAR outer loops. The spectral truncation is unchanged.
- For the EDA, the spectral truncation of the outer loop is increased from TL399 to TCO639 with the corresponding cubic octahedral reduced Gaussian grid.
- Increased semi-lagrangian departure point iterations from 3 to 5 to remove numerical instabilities near strong wind gradients, particularly improving East Asia (downstream of the Himalayas) and improved representation of tropical cyclones.
- Changed formulation of the horizontal spectral diffusion to a spectral viscosity with significantly reduced damping at the small scales.
- Removed dealiasing filter on rotational part of the wind as no longer needed for cubic grid (no aliasing).
- Reduced diffusion in the sponge layer near the top of the model (above level 30) scaled by grid resolution rather spectral resolution, due to new cubic grid.

2.2.4 Physics

- Improved representation of radiation-surface interactions with approximate updates every timestep on the full resolution grid leads to a reduction in 2m temperature errors near coastlines.
- Included surface-tiling for long-wave radiation interactions to reduce occasional too cold 2m temperature errors over snow.
- Improved freezing rain physics and an additional diagnostic for freezing rain accumulation during the forecast.
- Introduced resolution dependence in the parametrization of non-orographic gravity wave drag, reducing with resolution and improving upper stratospheric wind and temperature for HRES and ENS.
- Changed the parcel perturbation for deep convection to be proportional to the surface fluxes, reducing overdeepening in tropical cyclones.
- Increased cloud erosion rate when convection is active, to reduce cloud cover slightly and improve radiation, particularly over the ocean.
- Improvements of linear physics used in the data assimilation for gravity wave drag, surface exchange and vertical diffusion, improving near-surface winds over ocean in the short-range.

- Correction to solar zenith angle for the sunshine duration diagnostic. For clear sky days the sunshine duration increases by 2 hours, now in good agreement with observations. For cloudy days, sunshine duration may now be overestimated due to an existing underestimation of cloud optical thickness.
- Improved solar zenith angle calculation removes stratospheric temperature dependence on radiation timestep and reduces anomalous small amplitude fluctuations in incoming solar radiation around the equator.

2.2.5 Ensemble

- Modified SKEB (Stochastic Kinetic Energy Backscatter) stochastic physics necessary for the new cubic grid, removing the numerical dissipation estimate from the dissipation rate. Reduces ensemble spread slightly, but this is then consistent with reduced error (RMSE) in the new cycle.
- Modified singular vector calibration to compensate for increased variance from the higher resolution EDA.

2.3 Meteorological impact of the new cycle

Comparison of scores between IFS cycle 41r2 and IFS cycle 41r1 for HRES can be found in the IFS cycle 41r2 scorecard (Figure 1).

2.3.1 Upper air

The new model cycle (41r2) provides improved HRES and ENS performance throughout the troposphere. In the HRES there is a significant reduction of forecast errors in upper-air fields in the extra-tropics. Error reductions on the order of 2-3% are found for most upper-air parameters and levels. The improvement in the primary headline score for the HRES (lead time at which the 500 hPa geopotential anomaly correlation drops below 80%) is about 2 hours (0.08 days). Improvements are seen both in verification against the model analysis and verification against observations. In the tropics, evaluation against model analysis shows an apparent degradation in the short and near-medium range, mostly due to a more active analysis resulting from the increase in resolution of the EDA. Verification against observations, however, gives neutral to positive results in the tropics, except for temperature at 500 hPa, which shows a slight degradation.

The root mean square error (RMSE) and anomaly correlation for temperature are both improved in the extra-tropics, but there is a small (0.2 K) mean cooling in the upper troposphere. As the mean geopotential in the lower stratosphere is sensitive to changes in the vertically integrated tropospheric temperature, this shows up as an increased RMSE for geopotential at 100 hPa. The upper air scores over East Asia are significantly better associated with an improved representation of the flow downstream of the Himalayas due to the new cubic grid and more stable numerics. The overall kinetic energy spectra of the model is significantly improved with an increase in the energy towards the smaller scales.

Changes in skill of the ENS are generally similar to the HRES, with improvements in the extra-tropics on the order of 2-3% (CRPS reduction), and degradations in the tropics when verified against analysis. The improvement in the primary headline score for the ENS (lead time at which the CRPSS of the 850

hPa temperature drops below 25%) is ~0.2 days. The overall kinetic energy spectra of the model is significantly improved with an increase in the energy towards the smaller scales.

2.3.2 Weather parameters

The increased resolution leads to a better representation of coastlines and orography with potential for improved local prediction. The new model cycle yields consistent gains in forecast performance in the tropics and extra-tropics for 2m temperature, 2m humidity, and 10m wind speed. Precipitation forecasts are slightly improved in the extra-tropics and slightly deteriorated in the tropics. Mostly neutral results are found for forecasts of total cloud cover.

The increase in forecast skill for 2m temperature as measured by the reduction of RMSE is about 3% in the northern hemisphere extra-tropics and 1% in the tropics. There is a mean cooling in the northern hemisphere of about 0.05 K. Changes to the calculation of radiative fluxes lead to particular improvements in near coastal 2m temperatures at places where surface conditions vary abruptly.

For 2m dewpoint, RMS error reductions of 2% are observed in the NH, and neutral results are obtained for the tropics. There is an overall reduction of 2m dewpoint on the order of 0.05 to 0.1 K.

The RMSE for 10m wind speed improves by about 2% overall. There is no significant change in the mean in the northern hemisphere, and a reduction on the order of 0.05 m s-1 in the tropics.

Forecast skill for 24 hour precipitation totals shows an overall slight improvement in the northern hemisphere, and a small (1%) degradation in the tropics. This degradation in the tropics is seen in the SEEPS score but not in the RMSE.

Total cloud cover shows improvements in RMSE on the order of 0.5% in the tropics and neutral results in the extra-tropics.

There is a substantial reduction in localized (unrealistic) precipitation extremes over orography. The improvement is due to the cubic grid representation and modifications in the semi-Lagrangian advection scheme.

2.3.3 Tropical cyclones

The structural representation of tropical cyclones is improved with a more clearly defined eye and better resolved rainbands. Evidence from case studies shows that the increase in resolution leads to improved forecasts of tropical cyclone intensity in the ENS. Initial ensemble spread is also improved for tropical cyclones by the increased resolution in the EDA. For HRES, the tropical cyclone impact of the resolution change is smaller. Case studies show a better representation of the precipitation pattern around the core of tropical cyclones in the new cycle. This improvement is due to changes in model numerics (move to cubic grid and changes in the semi-Lagrangian scheme).

2.3.4 Wave forecast

Results for ocean wave height are positive, except for a deterioration in the very short range (day 1) in the tropics when verified against the analysis. A similar short-range degradation is seen for 10m wind speed over ocean areas. This is due to an increase in activity of the low-level wind and wave analyses

associated with the move from TL1279 to TCO1279. When verified against observations (buoys), no degradation is seen. Wave period has a mixed signal and may require some retuning in the next cycle.

2.3.5 Monthly forecast

Results suggest a generally neutral effect on upper-air and near surface skill scores in the tropics and for the MJO. For extra-tropical skill there is a slight improvement coming from the change to a cubic grid. Tropical cyclone sub-seasonal prediction is also improved.

2.3.6 Data assimilation / analysis

The kinetic energy spectrum has changed in the analysis even more than for the HRES. Whereas the analysis used to have less energy in the smaller scales compared to the forecast, both now have the same improved energy spectra. The background error variances derived from the higher-resolution EDA are larger in many areas, particularly in the tropics, leading to closer analysis fit to observations. Observation-minus-background departure statistics have improved for wind profile data.

The new model cycle is also described at

https://software.ecmwf.int/wiki/display/FCST/Detailed+information+of+implementation+of+IFS+cycle+41r2

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 the 12-month mean scores have reached their highest values so far; in Europe they are approximately equal to the highest previous values reached in 2014. This latest increase in skill is due to the last two model upgrades (cycle 41r1 in May 2015 and cycle 41r2 in March 2016).

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 both 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 in the northern extratropics has increased further in the last year. Curves for Europe are subject to larger inter-annual variability and do not show a clear signal. Here, jumpiness is close to the lowest values seen in 2014.

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 relatively stable since last year.

The trend in ENS performance is illustrated in Figure 6, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. In both areas the 12-month mean ENS skill has reached levels close to those in winter 2009–10, where predictability was exceptionally high. In Europe, the 3-month mean has for the first time exceeded a value of 10 days during the 2015-16 winter.

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 7. For 500 hPa geopotential height, the match between spread and error in the winter 2015-16 has been exceptionally good over the first 10 days of the forecast. It has noticeably improved compared to previous years. For 850 hPa temperature, the winter 2015-16 is comparable to the average of the two previous winters. A general under-dispersion for temperature at 850 hPa (also in summer, not shown) 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 is close to the diagonal. Figure 8 and Figure 9 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 9 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 well, with its spread reliability usually closest to the diagonal. The stars in the plots mark the average values, corresponding to Figure 7, and ideally should lie on the diagonal, as closely as possible to the lower left corner. Also in this respect ECMWF usually performs best, with the exception of 850 hPa temperature in the tropics in the short range, where JMA has the lowest error.

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 ERA-I. Figure 10 shows the evolution of the CRPS for the ENS and for the dressed ERA-I over the last 11 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 has started to exceed 30% during 2015. In the southern hemisphere, the corresponding value is about 28%. 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 11. At 200 hPa (upper panel) both the 1-day and 5-day forecast errors have changed very little. At 850 hPa (lower panel) the error at day 1 has been stable as well, while at day 5 it has increased over the last year, and most recently decreased again. 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 15 and Figure 16). 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. New scoring procedures for upper-air fields were approved for use in this score exchange by the 16th WMO Congress in 2011 and are 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 12 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. A noticeable reduction of RMSE at day 6 in the southern extratropics can be seen in 2015-16, while in the northern extratropics this signal is less pronounced. At day 1, the overall lead of ECMWF has not changed much over the last year.

WMO-exchanged scores also include verification against radiosondes over regions such as Europe. Figure 13 (Europe), and Figure 14 (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 15 (verification against analyses) and Figure 16 (verification against observations) which show vector wind errors for 250 hPa and 850 hPa. 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. In the tropics, verification against analyses (Figure 15) is sensitive to details of the analysis method, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 16), the ECMWF forecast has now the smallest overall errors both in the short and medium range. However, ECMWF's lead in the tropics is smaller than it is in the extratropics.

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 17. 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 17, 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 17). For example, the increase in skill at the beginning of 2015 is also seen in ERA-Interim, and the difference has been relatively stable at the time. Towards the end of 2015, however, there is an increase in the skill difference due to the last model upgrade (cycle 41r2) which would have been masked by atmospheric variability if only the HRES scores were considered. The probabilistic precipitation score (lower panel in Figure 17) also shows a pronounced improvement in 2015 and now exceeds 7 days.

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 days 1 and 2 (Figure 18, upper panel). For the ENS there is a consistent clear lead of ECMWF over the whole lead time range (Figure 18, bottom panel).

Trends in mean error (bias) and standard deviation over the last 10 years for 2 m temperature, 2 m dewpoint, total cloud cover, and 10 m wind speed forecasts over Europe are shown in Figure 19 to Figure 22. Verification is against synoptic observations received via the Global Telecommunication System (GTS). A correction for the difference between model orography and station height was applied to the temperature forecasts. No other post-processing has been applied to the model output.

In general, the performance over the past year is similar to previous years. For 2 m temperature (Figure 19) the error standard deviation (upper curves) has been somewhat higher than in the previous year, while the bias has remained slightly negative during daytime and become more strongly positive at night in summer. This problem has been partially addressed in cycle 41r2, and night-time bias for summer 2016 is expected to be reduced compared to 2015. Dewpoint shows little change over the last year. For total cloud cover (Figure 21) the error standard deviation has shown little change as well, however the bias has been further reduced both during the day and at night. For wind speed (Figure 22) the bias is very similar to last year, while the standard deviation in winter 2015-16 was the lowest so far.

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 has been a sustained increase in the skill of the operational high-resolution forecast relative to ERA-Interim since 2010 in the extratropics (Figure 23), which 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). Note the steepening of the curve for the northern extratropics due to the last model upgrade (cycle 41r2). In the tropics, the skill relative to ERA-Interim has increased in a similar fashion up to 2014 but shown some slight downward trend over the last two years.

ERA-Interim is useful as a reference forecast for the HRES as it allows filtering out some of the effect of atmospheric variations on scores. Figure 24 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 error standard deviation. Curves show 12-month running mean values. It can be seen that the largest relative improvements (around 20%) since 2002 have been achieved for upper-air and dynamic fields, followed by 2 m temperature, 10 m wind speed, and total cloud cover. All parameters show the beneficial effect of the most recent model upgrades in 2015 and 2016. It is worth noting that even the skill for total cloud cover, which has been stagnant prior to 2012, shows substantial improvement in recent years. Note also that the full effect of the resolution upgrade in March 2016 is not yet visible in the 12-month running average, as it only includes 5 months of cycle 41r2.

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 25. The top panel of Figure 25 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. Errors in the wave height forecast have reached their lowest values by far in 2016, especially in the medium range. 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 in the medium-range has increased further in 2015 (Figure 26).

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 27 for the 12-month period June 2015 – May 2016. ECMWF forecast winds are used to drive the wave model of Météo France and the French marine (SHOM); the wave models of these are similar, hence the closeness of their errors in Figure 27. For both wave height and peak period, ECMWF generally outperforms centres which are not using ECMWF's wind forecasts.

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 moving 15-year sample). 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 28 (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 2015; the final point on each curve includes the spring (March–May) season 2016. For all three parameters, ROC skill has stabilized on a high level, with some inter-annual variations due to atmospheric variability. The seasonal skill of 2m temperature has reached its highest value so far in spring 2016.

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 29. Errors in the forecast central pressure of tropical cyclones 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 29) are comparable to last year, when they reached their lowest values so far. Mean absolute speed errors of the HRES and the CTRL at D+3 have further decreased. Typically tropical cyclones move too slowly in the forecast, however this negative bias has been relatively small in recent years. The mean error (bias) in tropical cyclone central pressure (upper left of the central panels in Figure 29) has decreased as well. In the HRES it has been close to zero in 2016. However, there is substantial compensation between over- and underforecast storms, and the mean absolute errors in central pressure have increased in recent years both in the HRES and the CTRL (upper right of the central panels in Figure 29).

The bottom panel of Figure 29 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 30. Results show a certain amount of over-confidence, however, reliability has increased compared to previous years (top panel). 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 31 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. As for other surface fields, a positive signal from the last two model upgrades can be seen across the lead-time range.

6 Monthly and seasonal forecasts

6.1 Monthly forecast verification statistics and performance

With the introduction of IFS cycle 41r1 (May 2015) the monthly ensemble forecasts and re-forecasts, which are run twice a week, were extended from 32 to 46 days. Since the resolution upgrade in March 2016 (IFS cycle 41r2) the ENS-extended benefits from being run at the highest resolution (18 km) over the first 15 days. From days 16 to 46 the resolution is 36 km.

Figure 32 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, both the absolute and the relative skill have not shown a systematic improvement in recent years.

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, at

www.ecmwf.int/en/forecasts/charts

by choosing 'Verification' and 'Seasonal forecasts' under the header 'Long Range'. A comprehensive description and assessment of System 4 is provided in ECMWF Technical Memorandum 656, available from the ECMWF website:

http://www.ecmwf.int/en/research/publications

6.2.2 The 2014–2015 El Niño forecasts

The year 2015 was characterized by a change from slightly warm to very warm conditions in the eastern tropical Pacific associated with a strong El Nino. Both the strengthening and weakening phases were very well captured in the forecast, with the ECMWF model providing useful guidance on the 6-month timescale for this event (left column of Figure 34). Previous experience showed 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. This was also the case for this event, however, the overestimation was not substantial, and the observed peak was well within the ensemble spread. The multi-model EUROSIP forecasts (right column of Figure 34) got the peak slightly better, however the overall spread was larger than in the ECMWF forecast, and the weakening phase was not as well predicted in the forecast from November 2015.

6.2.3 Tropical storm predictions from the seasonal forecasts

Due to El Nino, the 2015 North Atlantic hurricane season was relatively quiet with an accumulated cyclone energy index (ACE) of about 60% of the 1990–2010 climate average (see Figure 34). The number of tropical storms which formed in 2015 (11 named storms) was slightly below average (12). Seasonal tropical storm predictions from System 4 indicated below average activity compared to climatology over the Atlantic. The June forecast predicted 8 (with a range from 5 to 11) tropical storms in the Atlantic (Figure 35) and an ACE of 50% of the observed climatology. Most other seasonal forecast models predicted a below average 2015 Atlantic tropical storm season due to the presence of El-Nino.

Figure 35 shows that System 4 predicted average activity over the eastern North Pacific, and slightly enhanced activity over the western North Pacific. In the western North Pacific the forecast was very good, with 22 storms predicted, and 23 observed. In the eastern North Pacific, storm activity was

underestimated, with 16 observed and 13 predicted. In contrast to the North Atlantic, El Nino has counteracting effects on the North Pacific (higher SSTs but also enhanced shear), so that the predictability of tropical cyclone number is lower there. The 2015 tropical storm season also saw the formation of the most intense hurricane in the western hemisphere so far: Patricia (October 2015), off the Mexican coast, with a minimum core pressure of 872 hPa and a maximum 1-min sustained wind speed of 96 m s-1.

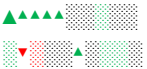
6.2.4 Extratropical seasonal forecasts

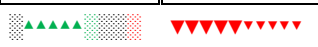
The presence of the strong El Nino substantially enhanced predictability on the seasonal time-scale. 2 m temperatures in the northern-hemisphere winter (DJF 2015–16) were characterized by strong warm anomalies over North America and northern Eurasia, which were captured quite well by the seasonal forecast (Figure 36). The westernmost parts of Europe were influenced by a persistent cold anomaly over the North Atlantic, which was also captured very well by the model. Also in the southern hemisphere, the general pattern of warm and cold anomalies was predicted to a large degree.

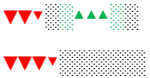
During the northern-hemisphere summer (JJA 2016), sea surface temperatures in the eastern tropical Pacific had reverted back from the previous El Nino to slightly negative (La Nina type) anomalies (Figure 37). In terms of large-scale anomaly patterns the seasonal 2 m temperature forecast captured most of the major features, however compared to DJF 2015-16, there was less skill on the smaller, sub-continental scales.

7 Action required

The Committee is invited to take note of the information presented.

				ACC	RMSe/STDe	SEEPS
Europe	against analysis	Geopotential	100hPa			
			500hPa			
			850hPa			
			1000hPa			
		MSL pressure				
		Temperature	100hPa			
			500hPa			
			850hPa			
			1000hPa			
		Wind	100hPa			
			200hPa			
			850hPa			
		Relative humidity	300hPa			
			700hPa			
	against observations	Geopotential	100hPa			
			500hPa			
		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	100hPa			
			200hPa			
			850hPa			
		10m wind over ocean				
		Ocean wave height				
		Ocean wave period				
		Relative humidity	300hPa			
			700hPa			

				ACC	RMSe/STDe	SEEPS
	against observations	Geopotential	100hPa			
			500hPa			
		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			
		Wind	100hPa			
			200hPa			
			850hPa			
		10m wind over ocean				
		Ocean wave height				
		Ocean wave period				
		Relative humidity	300hPa			
			700hPa			
	against observations	Geopotential	100hPa			
			500hPa			
		Temperature	100hPa			
			200hPa			
			850hPa			
		2m temperature				
		Wind	100hPa			
			200hPa			
			850hPa			
		10m wind				
		2m dew-point				
		Total cloud cover				
		24h precipitation				
Tropics		Temperature	100hPa			

				ACC	RMSe/STDe	SEEPS
	against analysis		500hPa			
			850hPa			
			1000hPa			
		Wind	100hPa			
			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)
▲ CY41r2 **better** than CY41r1 **statistically highly significant** (the confidence bar above zero by more than its height) (d/s>3)
▲ CY41r2 **better** than CY41r1 **statistically significant** (d/s≥1)
CY41r2 better than CY41r1, yet not statistically significant (d/s≥0.5)
not really any difference between CY41r1 and CY41r2
CY41r2 worse than CY41r1, yet not statistically significant (d/s≤-0.5)
▼ CY41r2 **worse** than CY41r1 **statistically significant** (d/s≤-1)
▼ CY41r2 **worse** than CY41r1 **statistically highly significant** (the confidence bar below zero by more than its height) (d/s<-3)

Figure 1: Summary score card for Cy41r2. Score card for cycle 41r2 versus cycle 41r1 verified by the respective analyses and observations at 00 and 12 UTC for 670 forecast runs in the period 1 January 2015 to 8 March 2016.

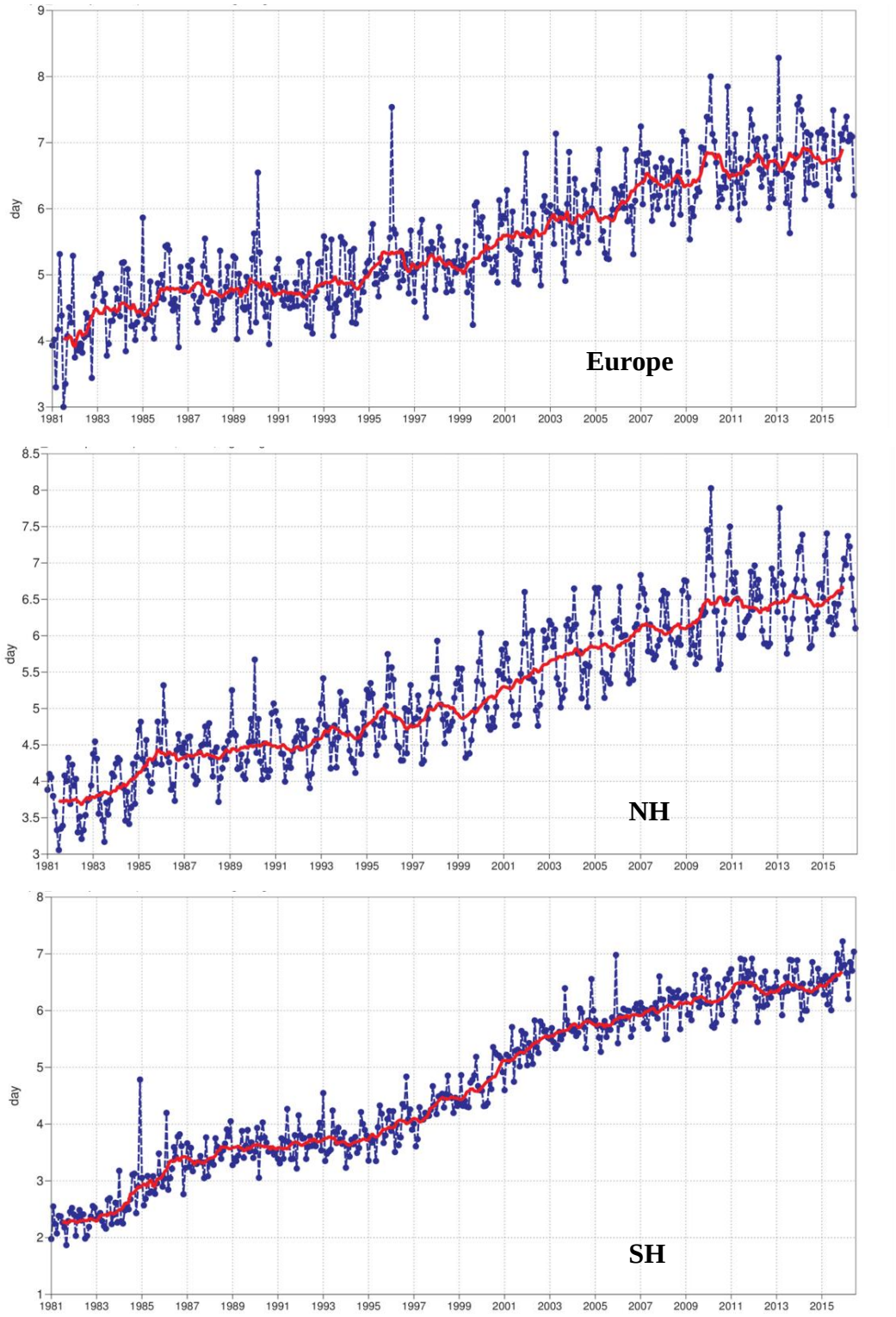


Figure 2: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly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

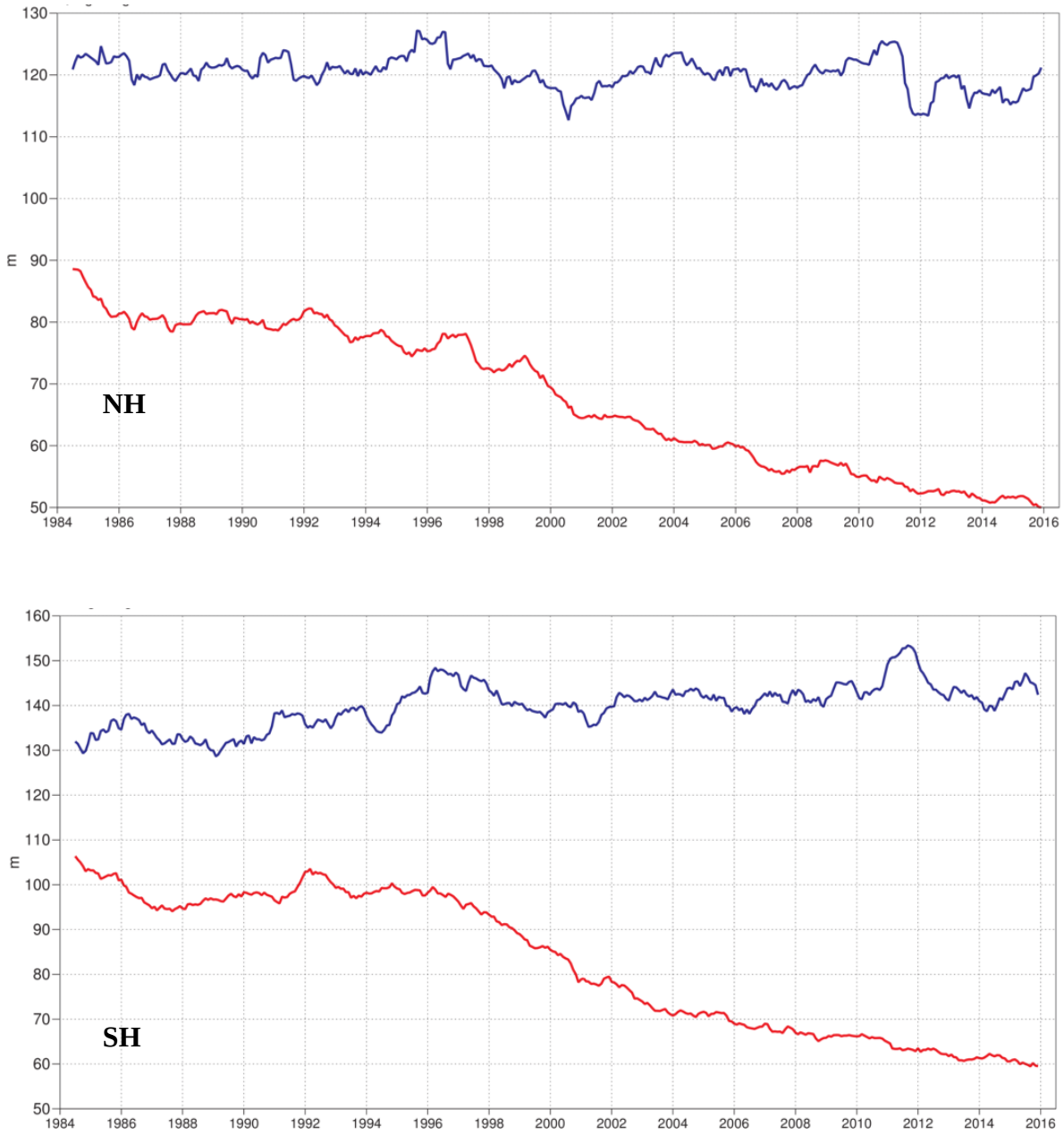


Figure 3: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2015–July 2016. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

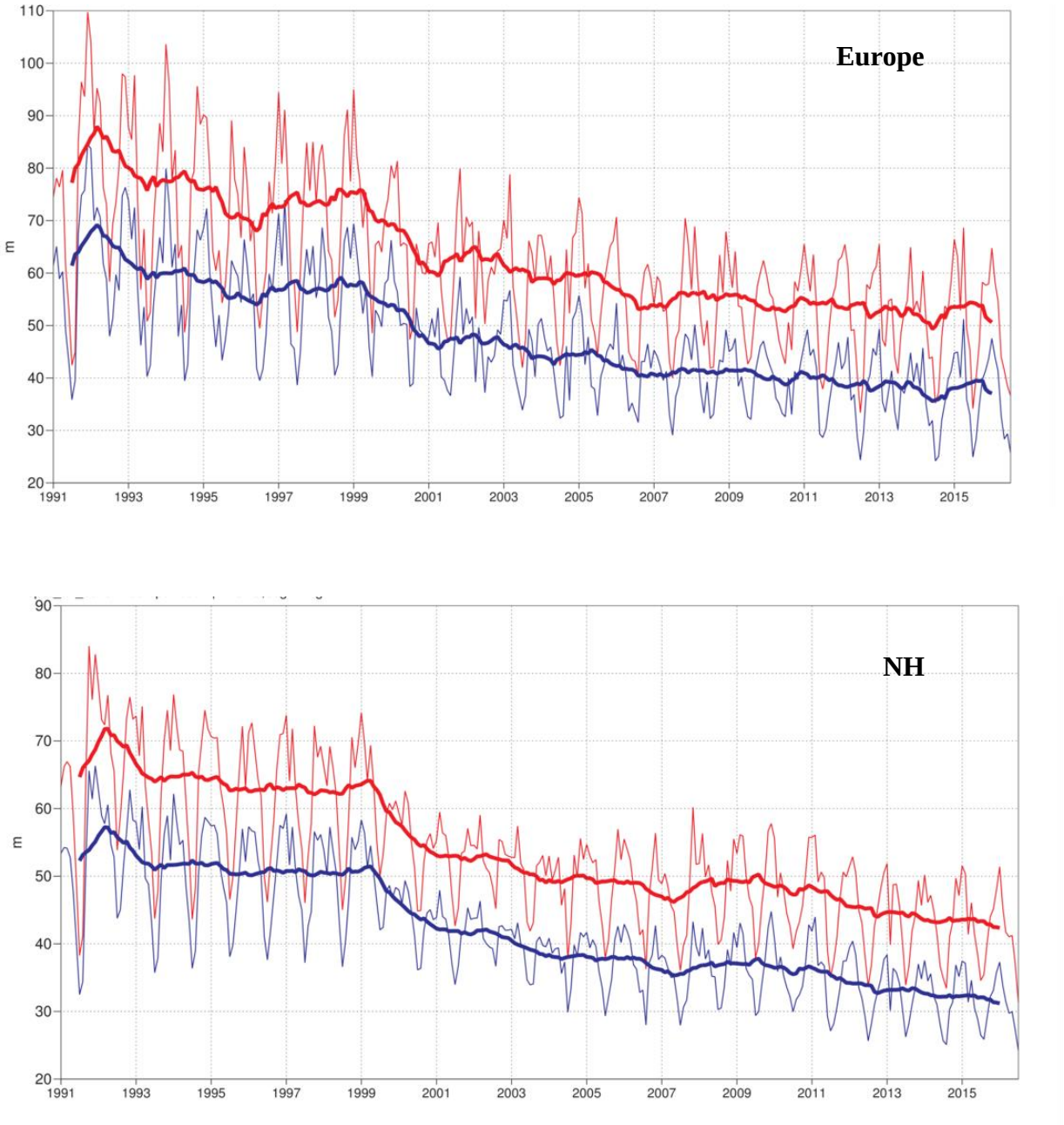


Figure 4: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores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

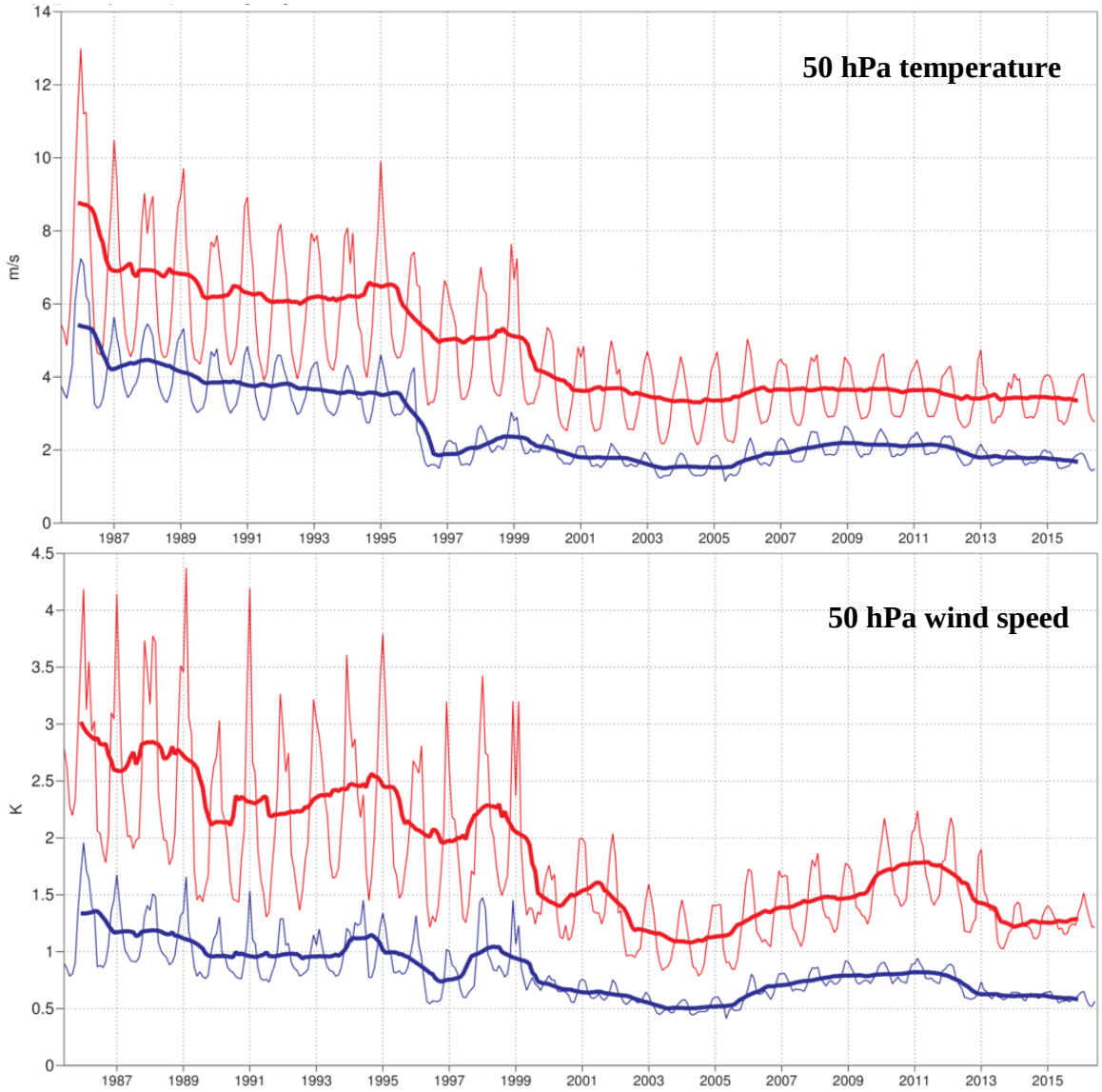


Figure 5: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also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

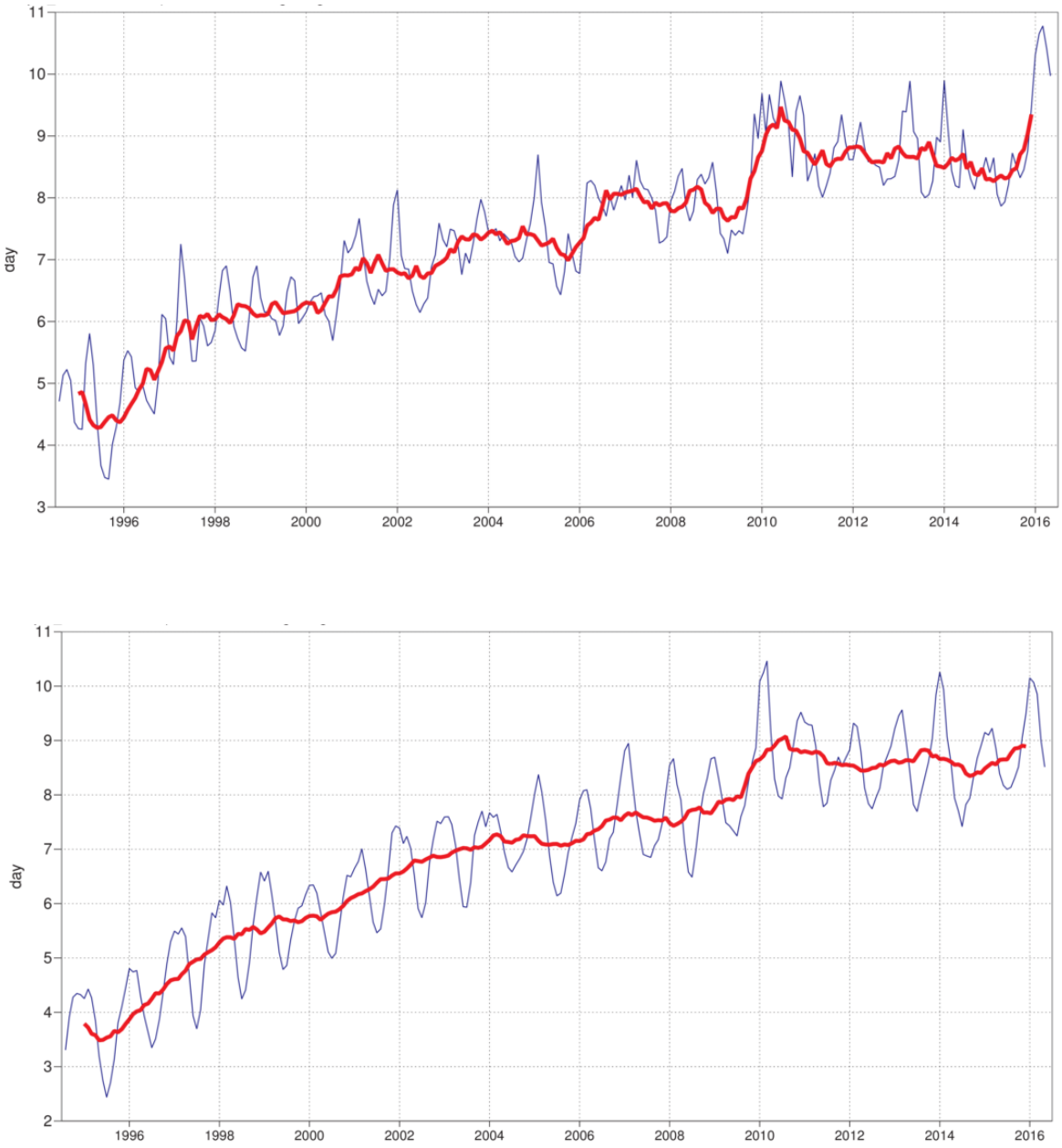


Figure 6: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous ranked probability skill score (CPRSS) falls below 25% for Europe (top), northern hemisphere extratropics (bottom).

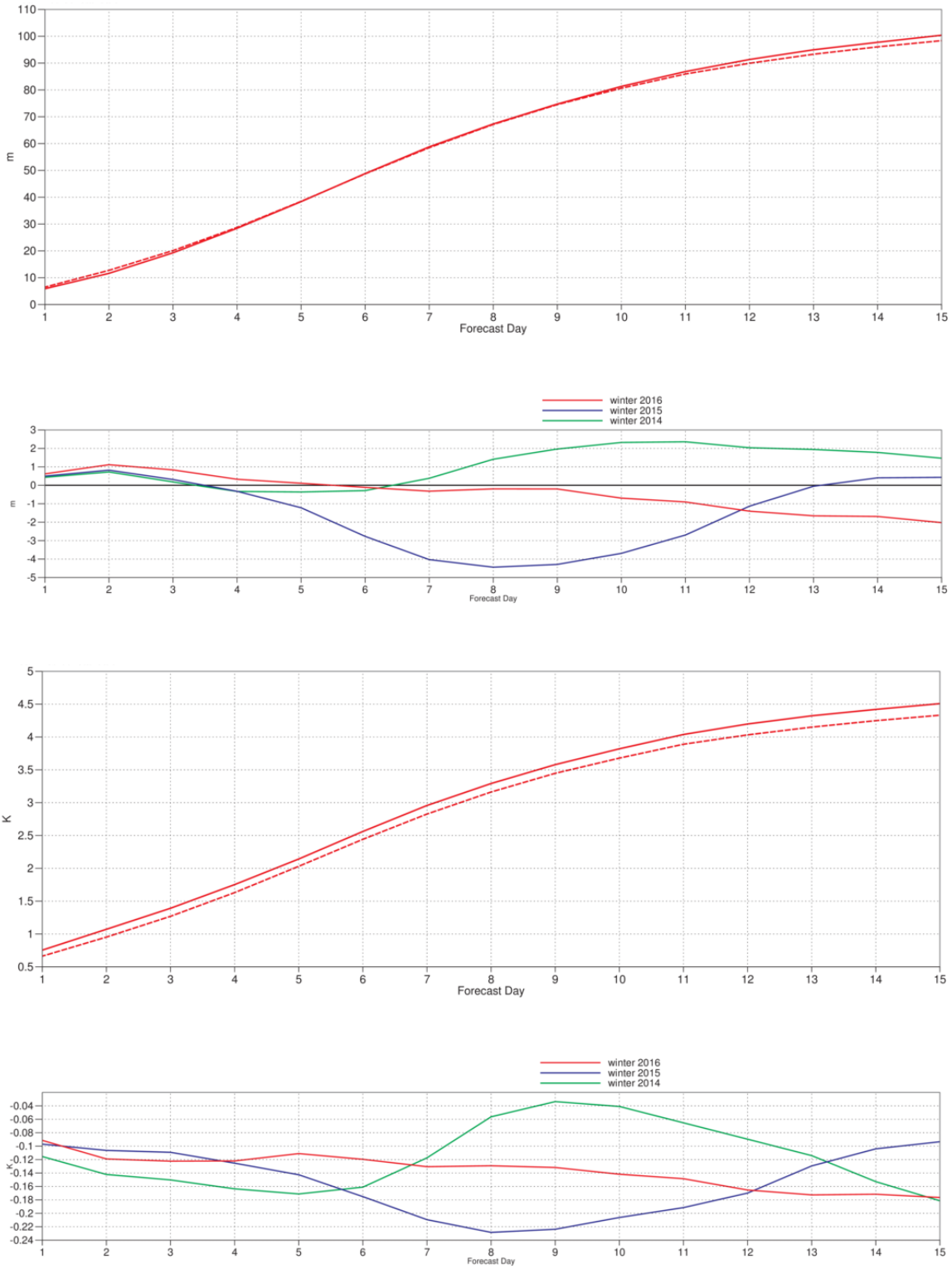


Figure 7: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2015–2016 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.

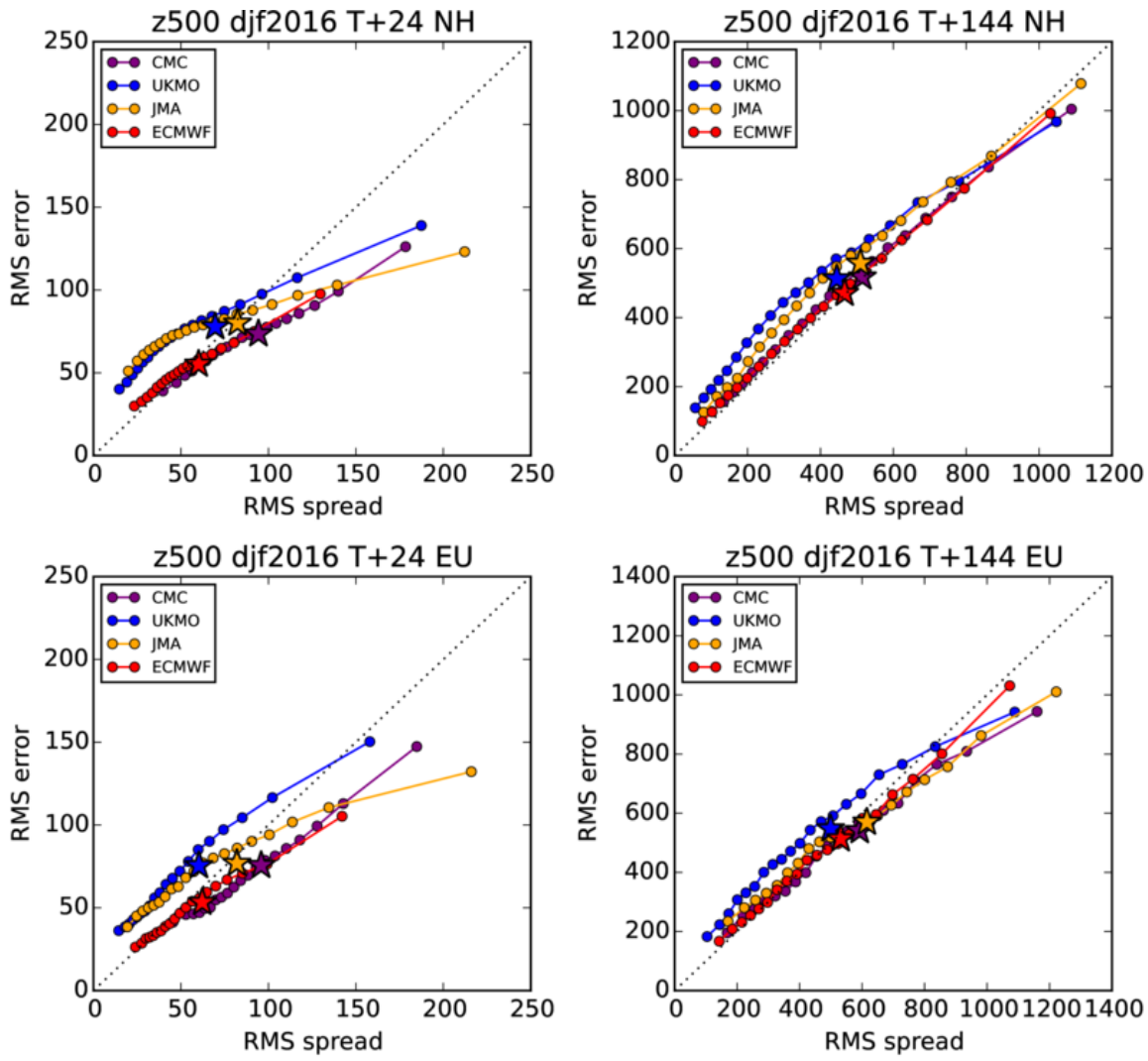


Figure 8: Ensemble spread reliability of different global models for 500 hPa geopotential in DJF 2015–16 in the northern hemisphere extra-tropics (top) and in Europe (bottom) for day 1 (left) and day 6 (right) , verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

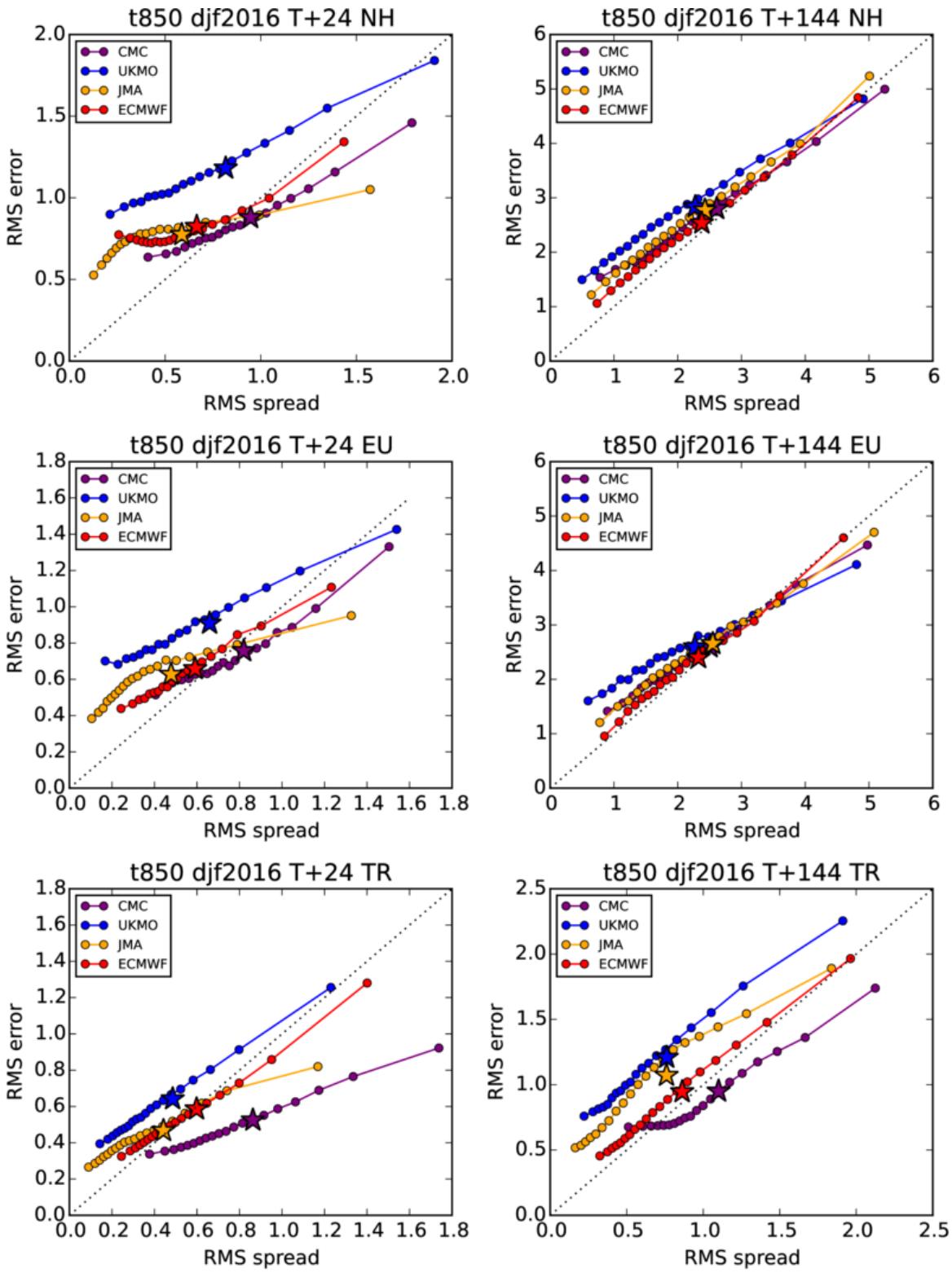


Figure 9: Ensemble spread reliability of different global models for 850 hPa temperature in DJF 2015–16 in the northern hemisphere extra-tropics (top), Europe (centre), and the tropics (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

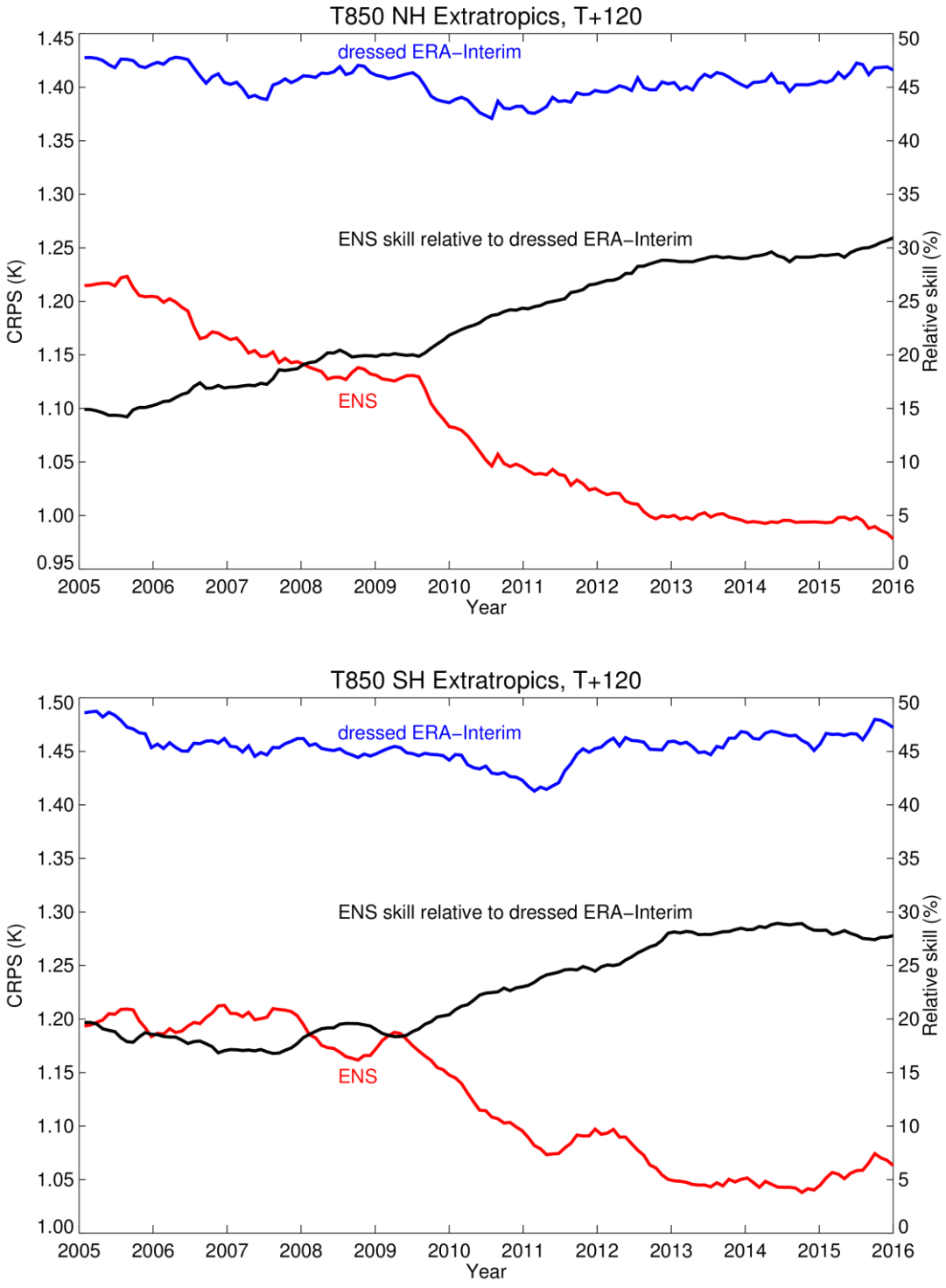


Figure 10: CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5, verified against analysis. Scores are shown for the ensemble forecast (red) and the dressed ERA-Interim forecast (blue). Black curves show the skill of the ENS relative to the dressed ERA-Interim 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.

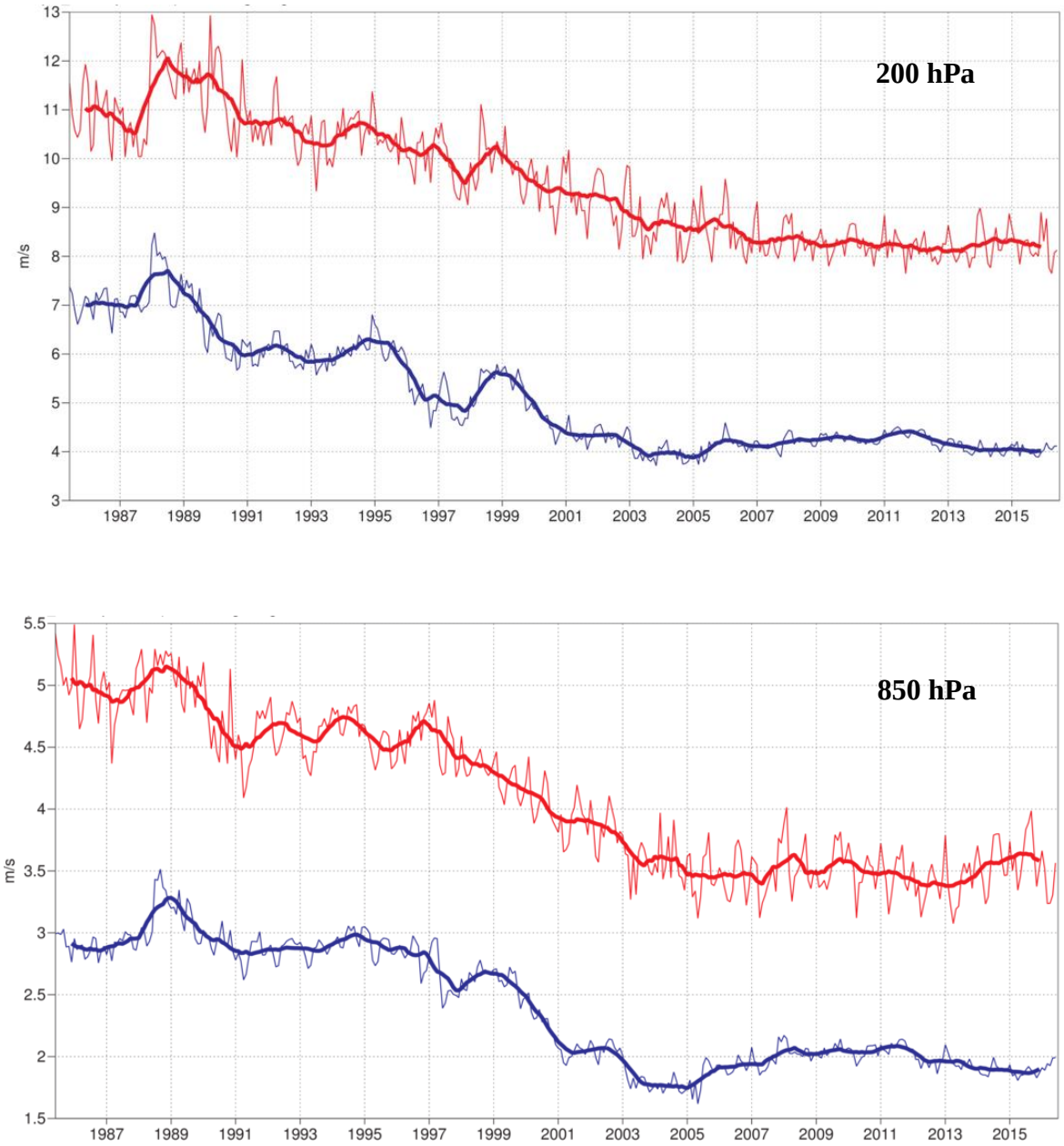
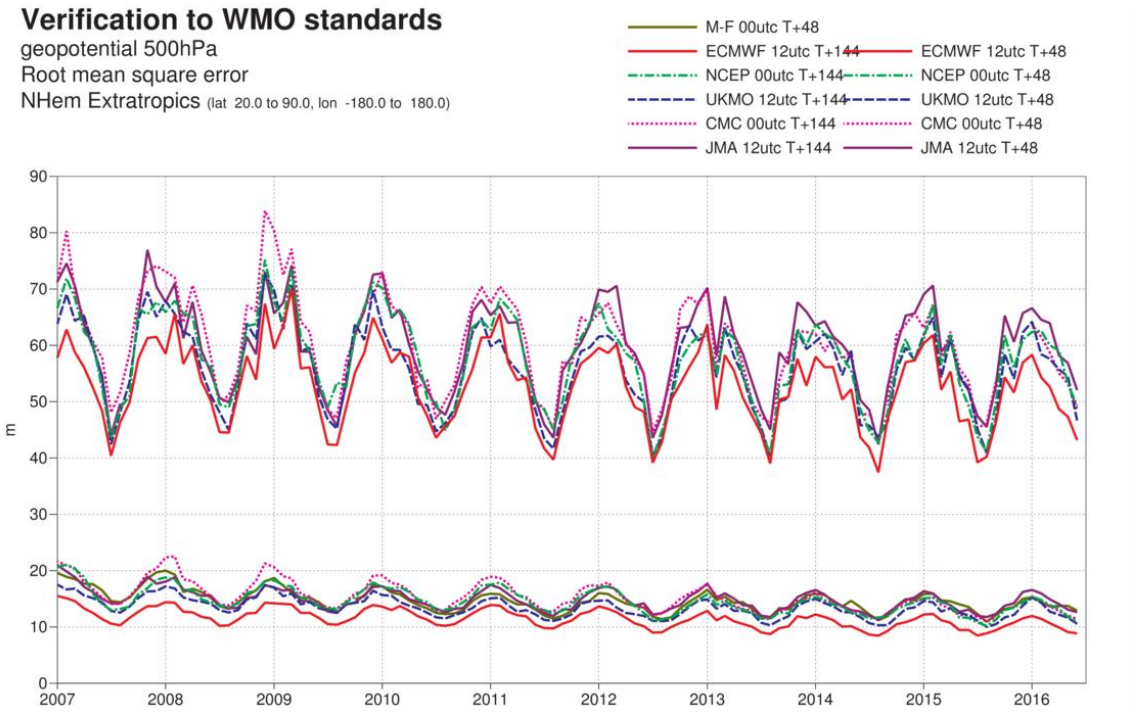


Figure 11: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).



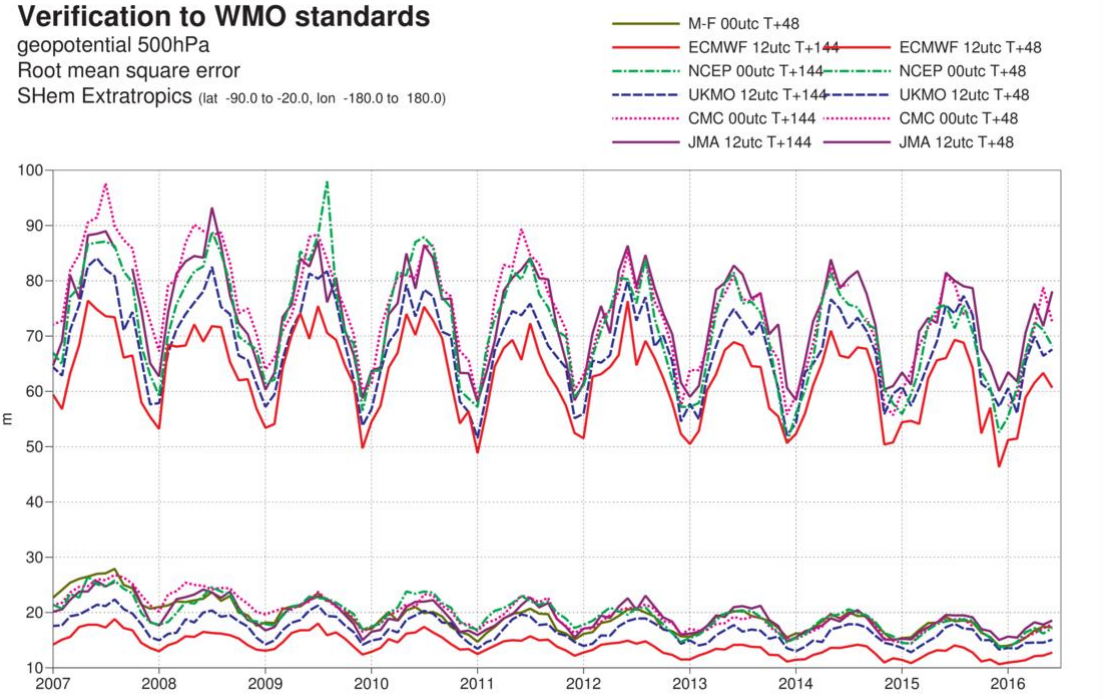
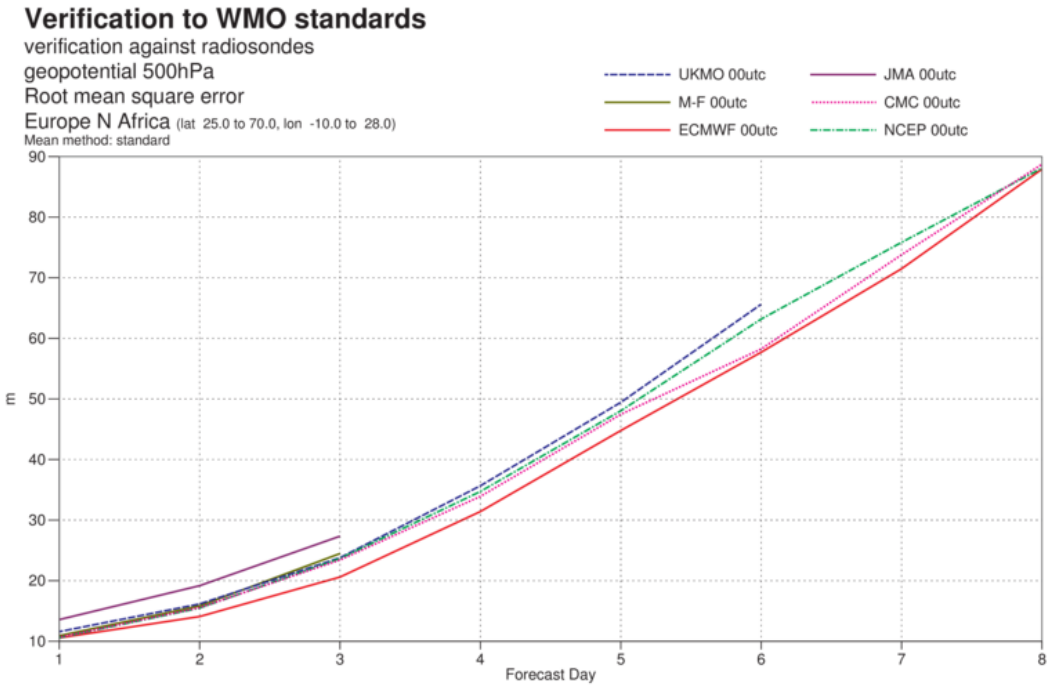


Figure 12: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top) and southern (bottom) extratropics. In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model is verified against 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.



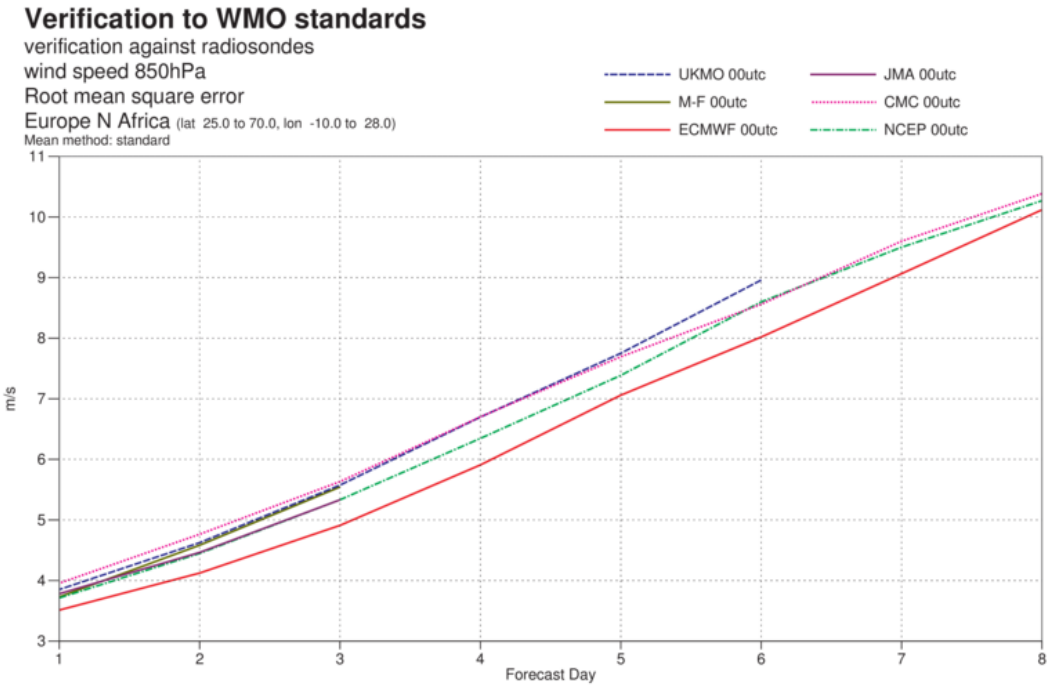
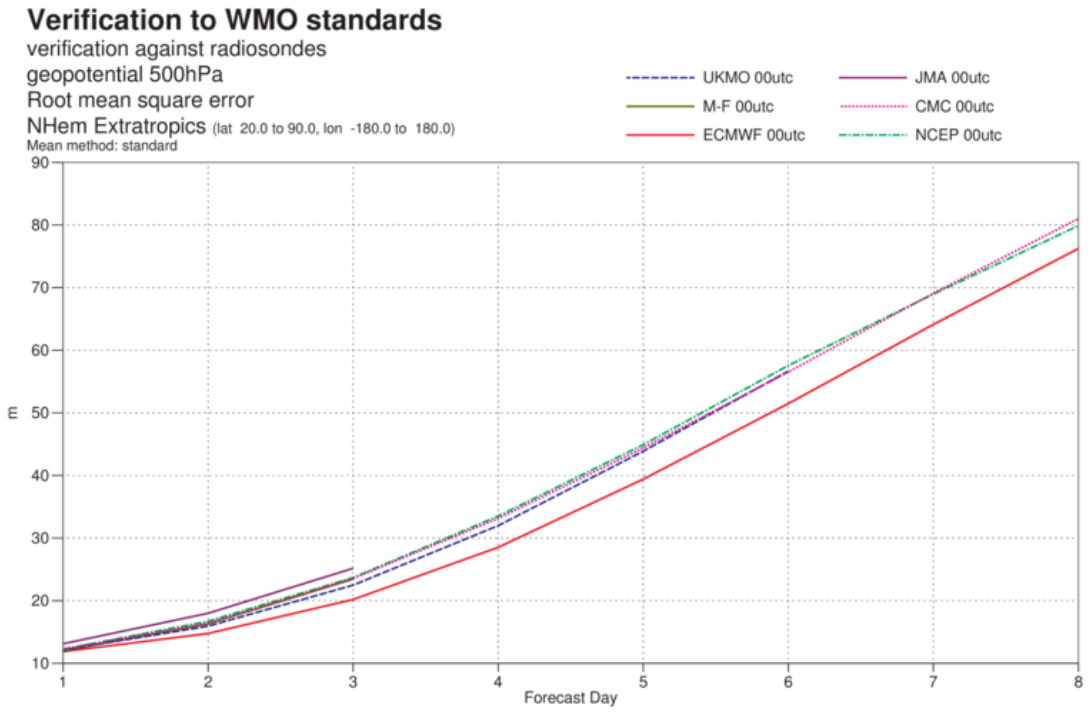


Figure 13: WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2015–July 2016).



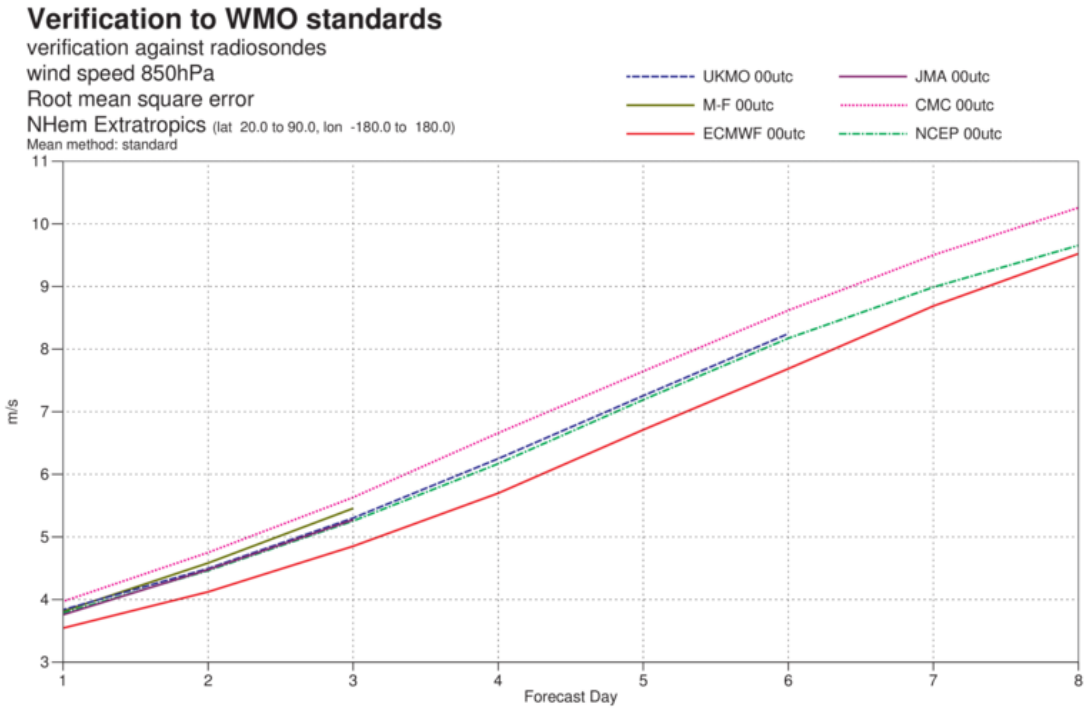
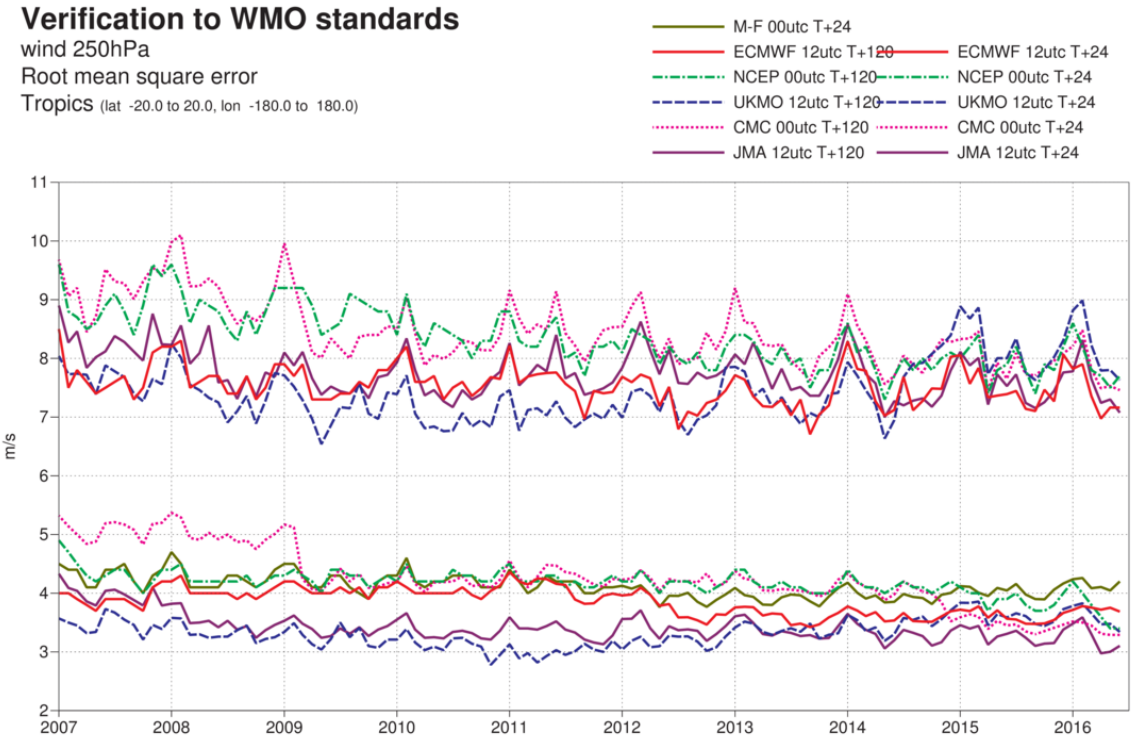


Figure 14: As Figure 13 for the northern hemisphere extratropics.



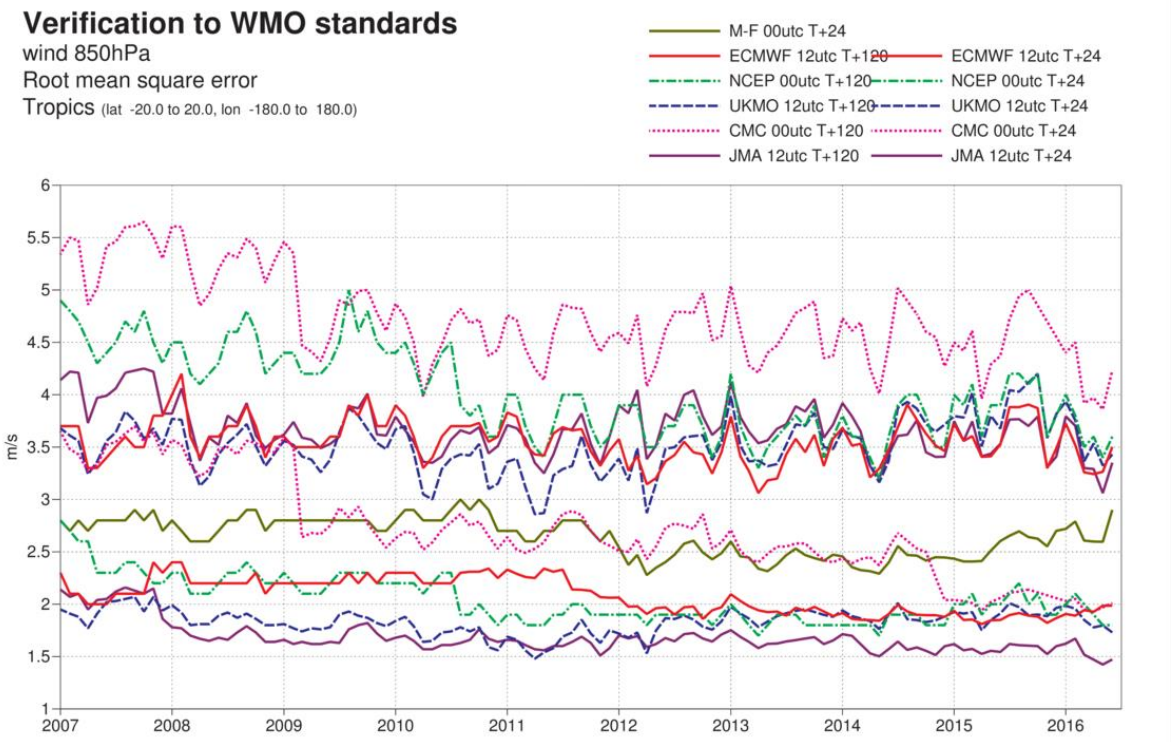
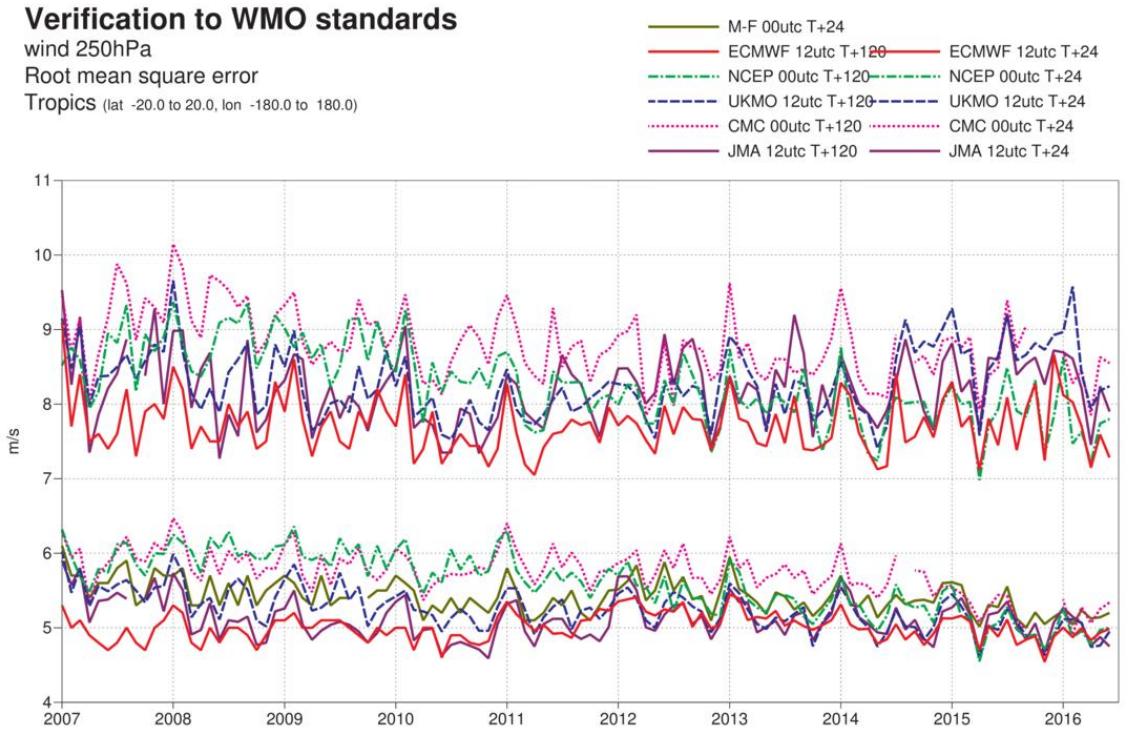


Figure 15: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



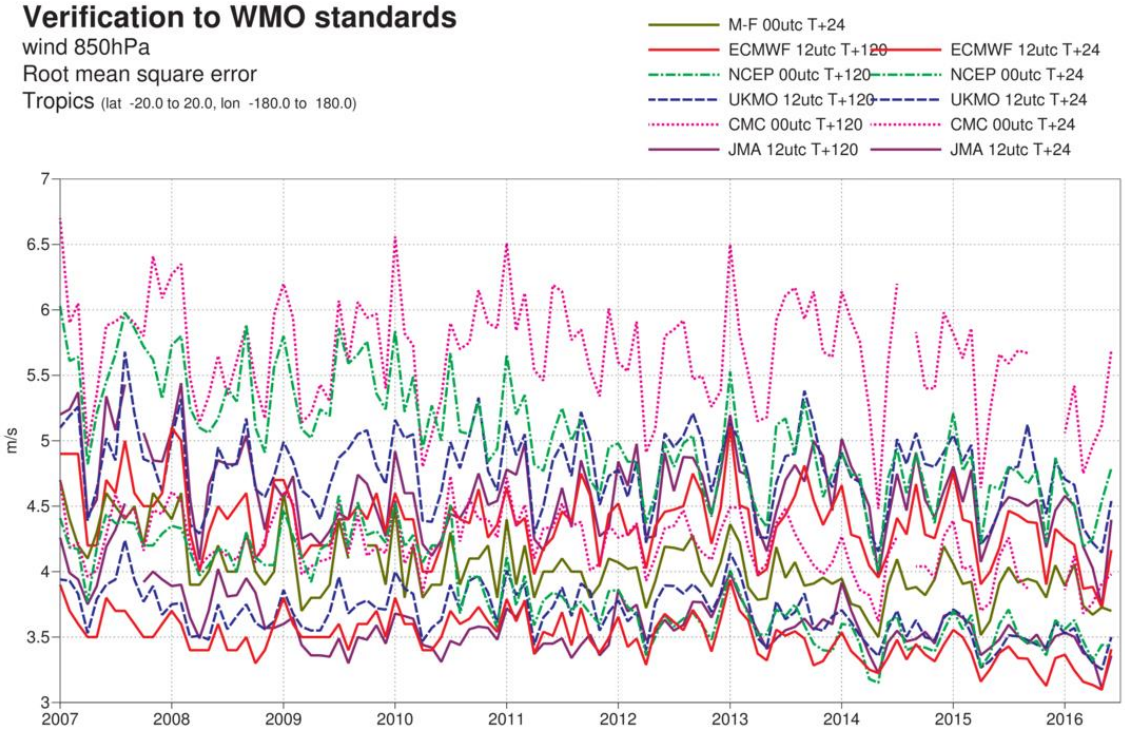


Figure 16: As Figure 15 for verification against radiosonde observations.

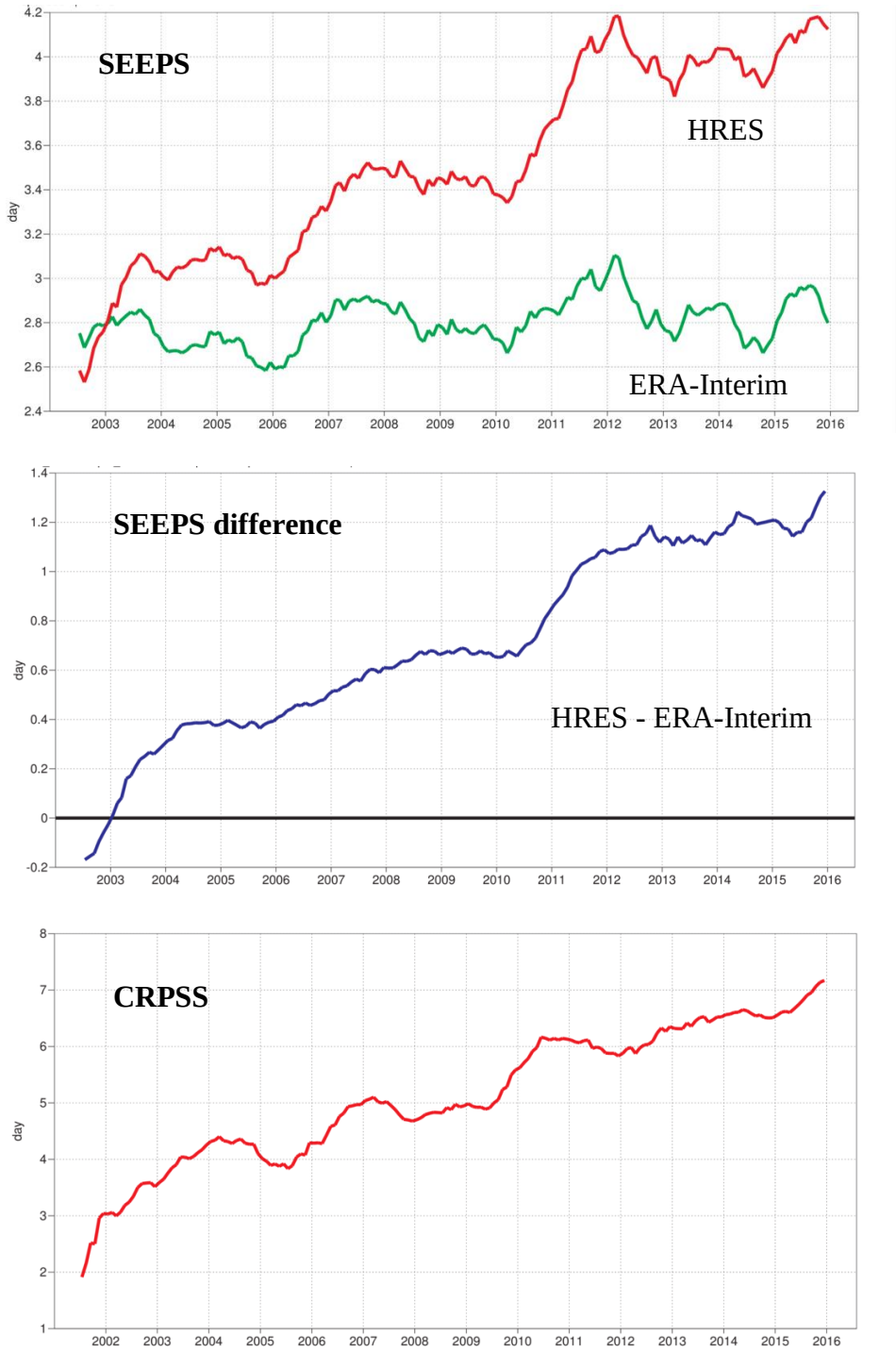


Figure 17: Supplementary headline scores for deterministic (top, centre) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics; each point is calculated over a 12-month period, plotted at the centre 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.

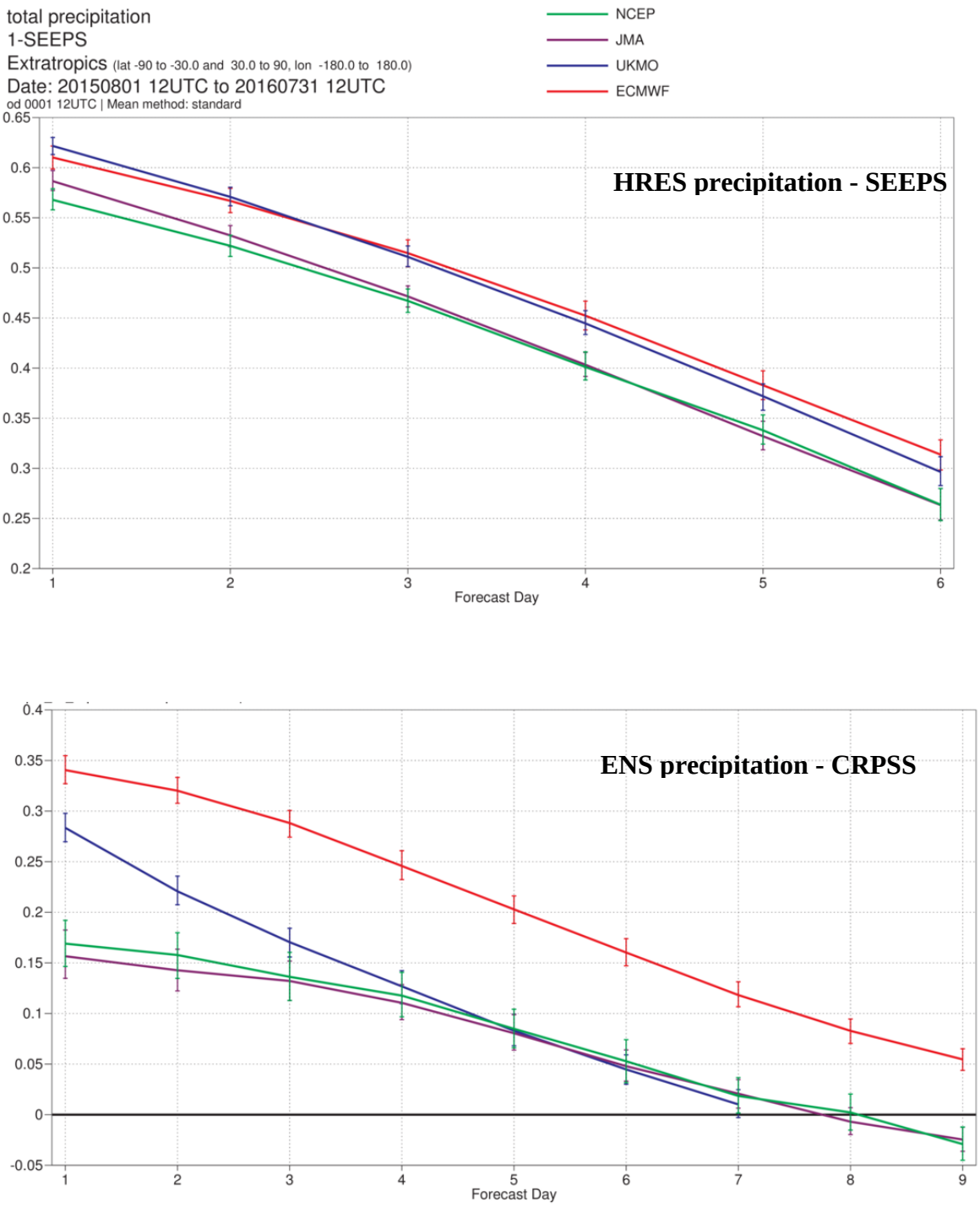


Figure 18: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 17. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2015–July 2016. Bars indicate 95% confidence intervals.

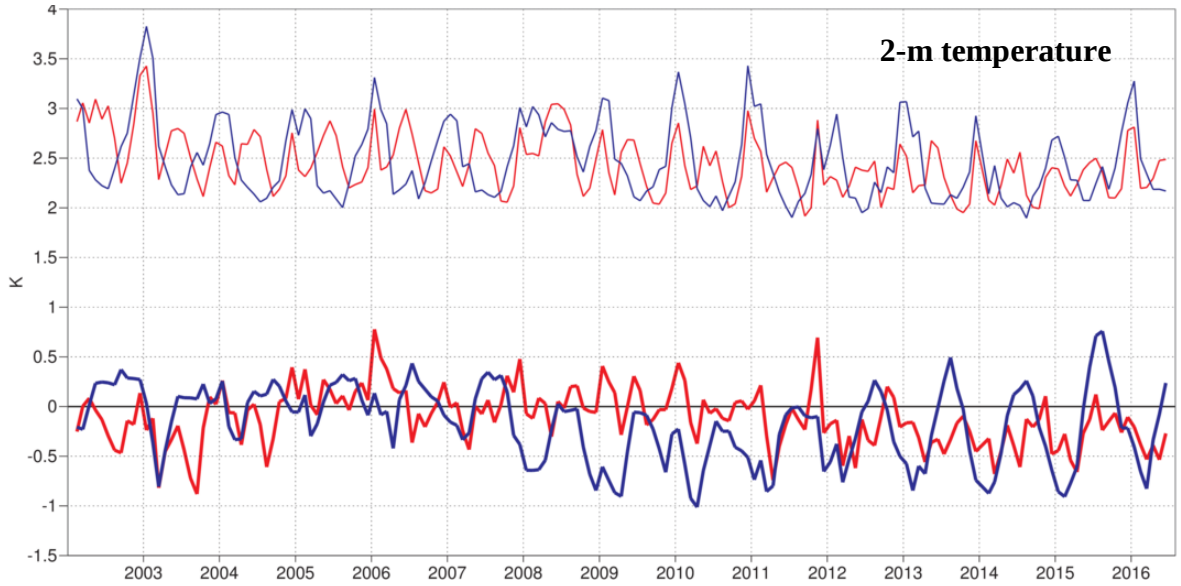


Figure 19: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

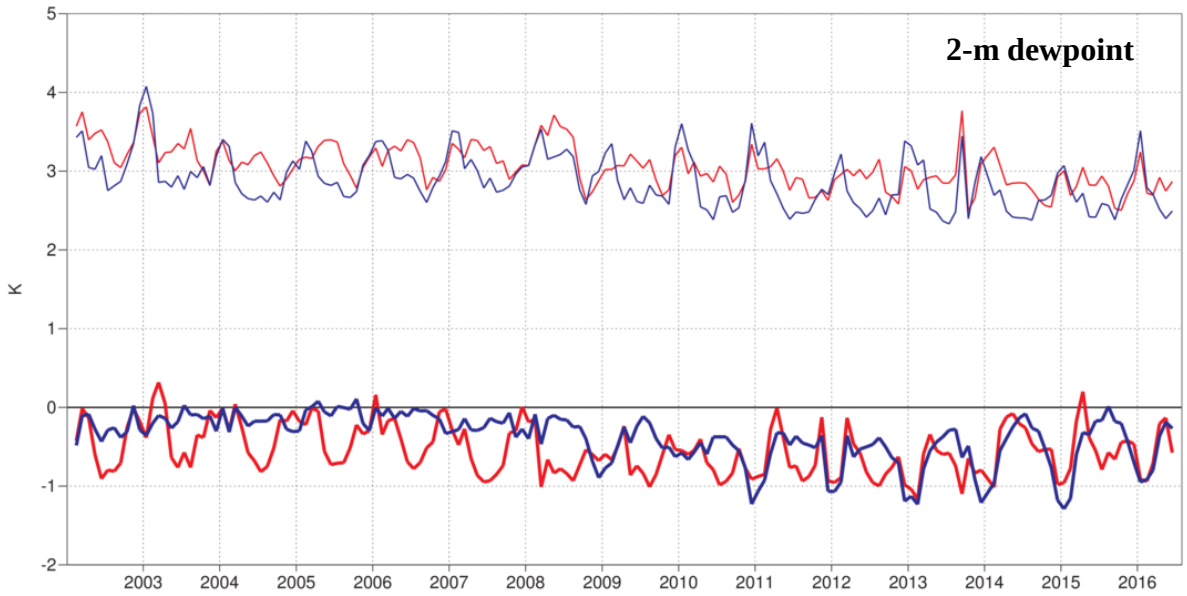


Figure 20: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

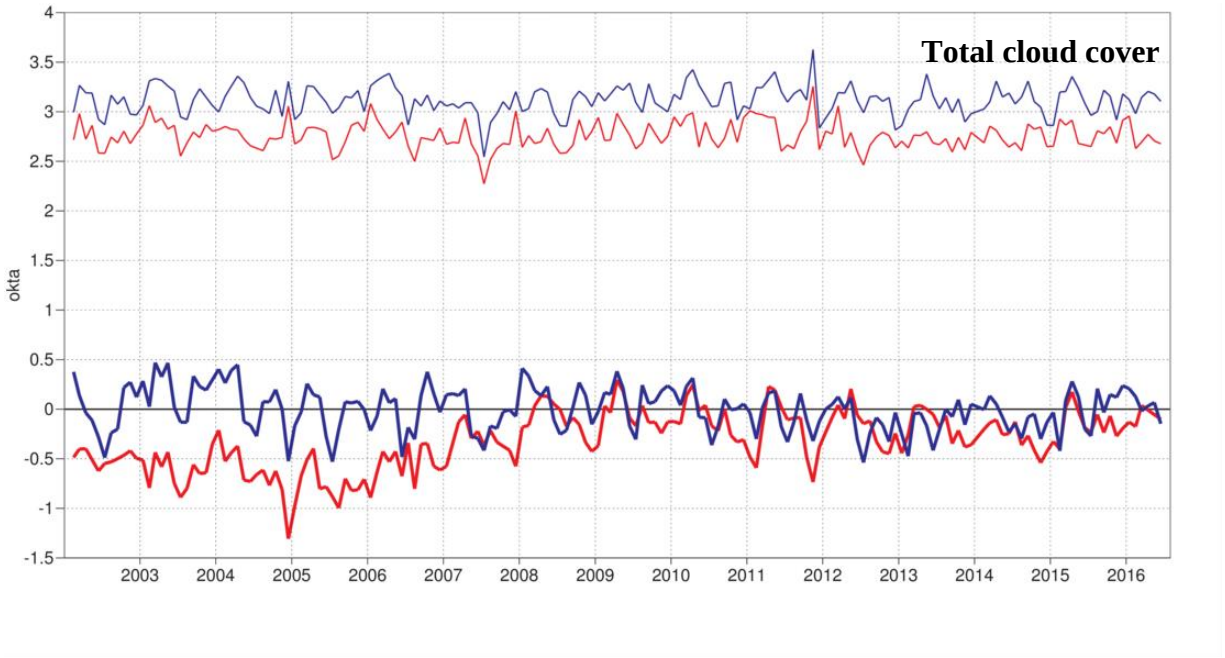


Figure 21: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

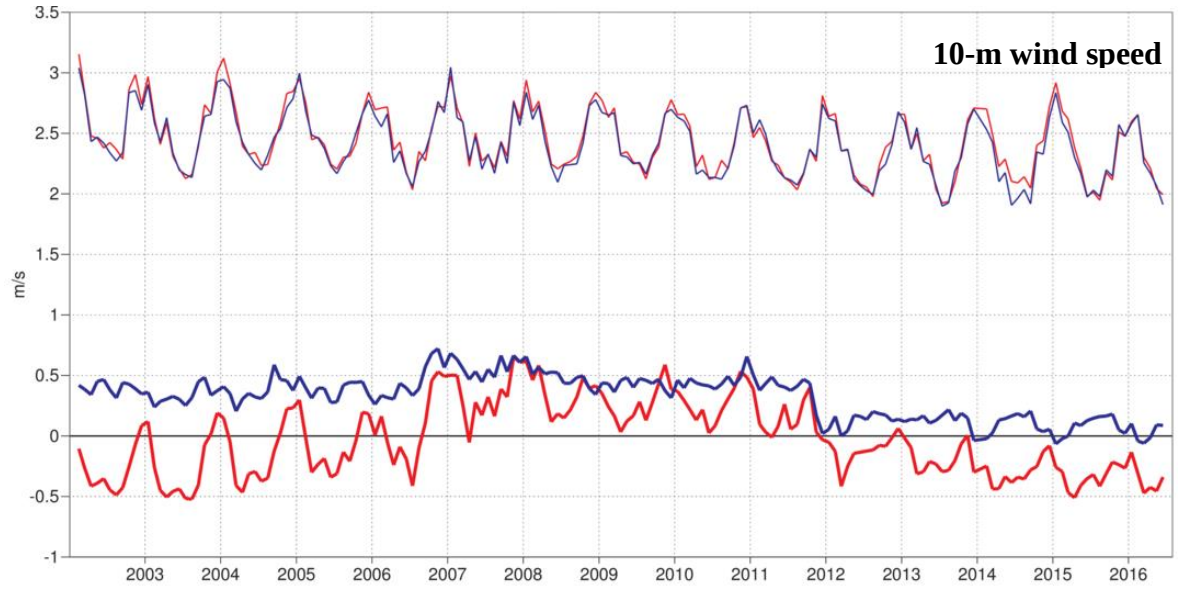


Figure 22: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

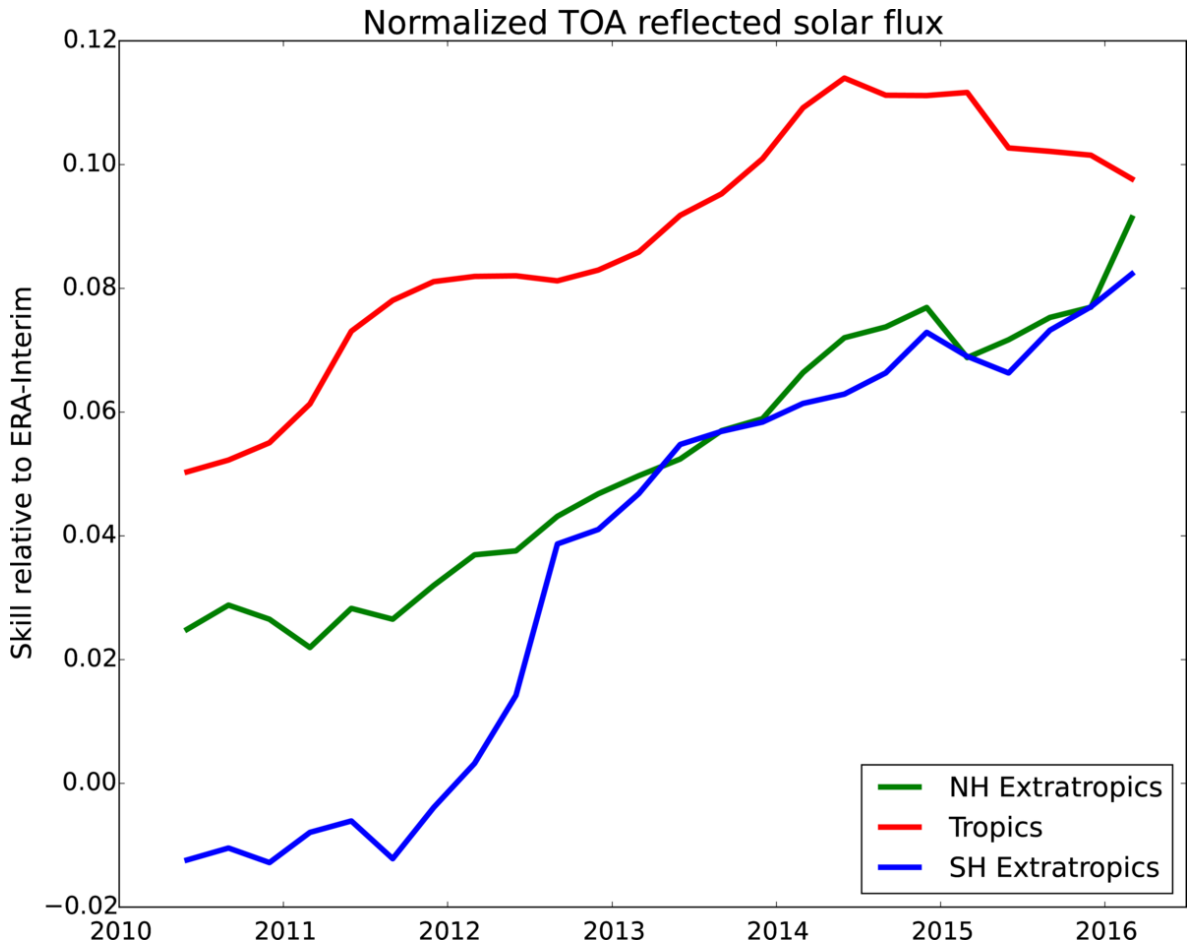


Figure 23: 12-month running average of the day 3 forecast skill relative to ERA-Interim of normalized TOA reflected solar flux (daily totals), verified against satellite data. The verification has been carried out for those parts of the northern hemisphere extratropics (green), tropics (red), and southern hemisphere extratropics (blue) which are covered by the CM-SAF product (approximately 70 S to 70 N, and 70 W to 70 E).

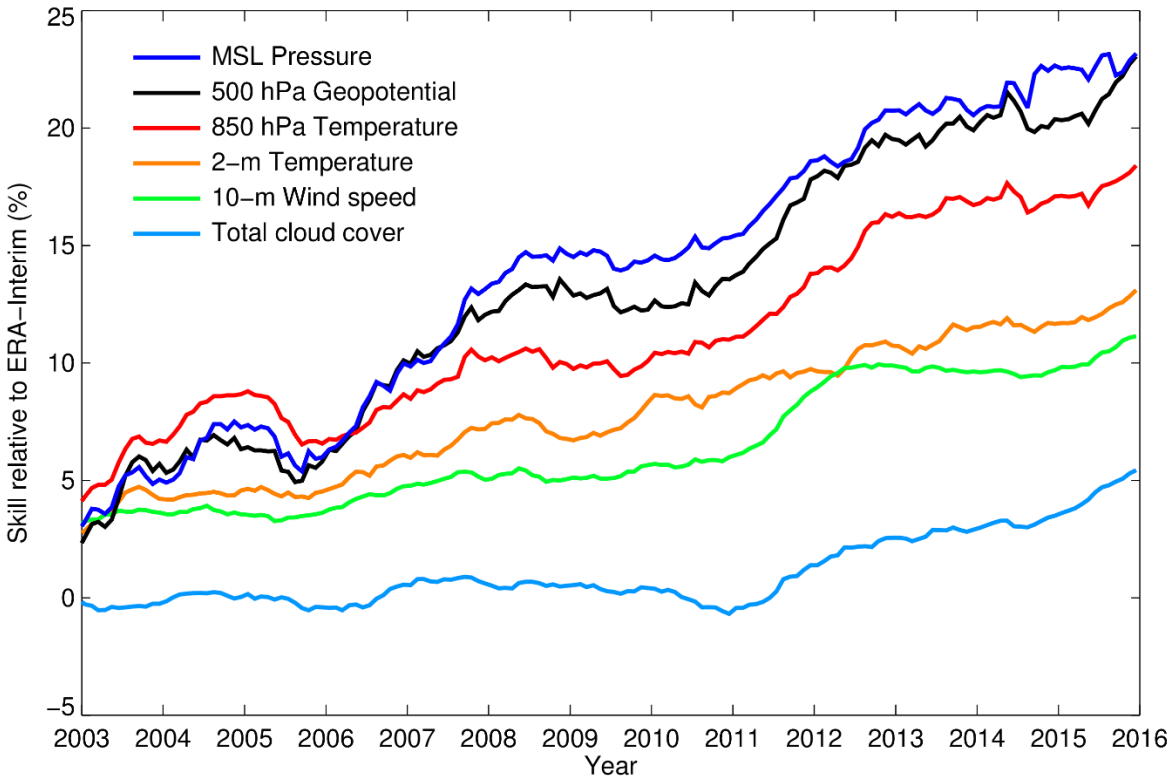


Figure 24: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA-Interim. Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), and mean sea level pressure (MSLP), using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC).

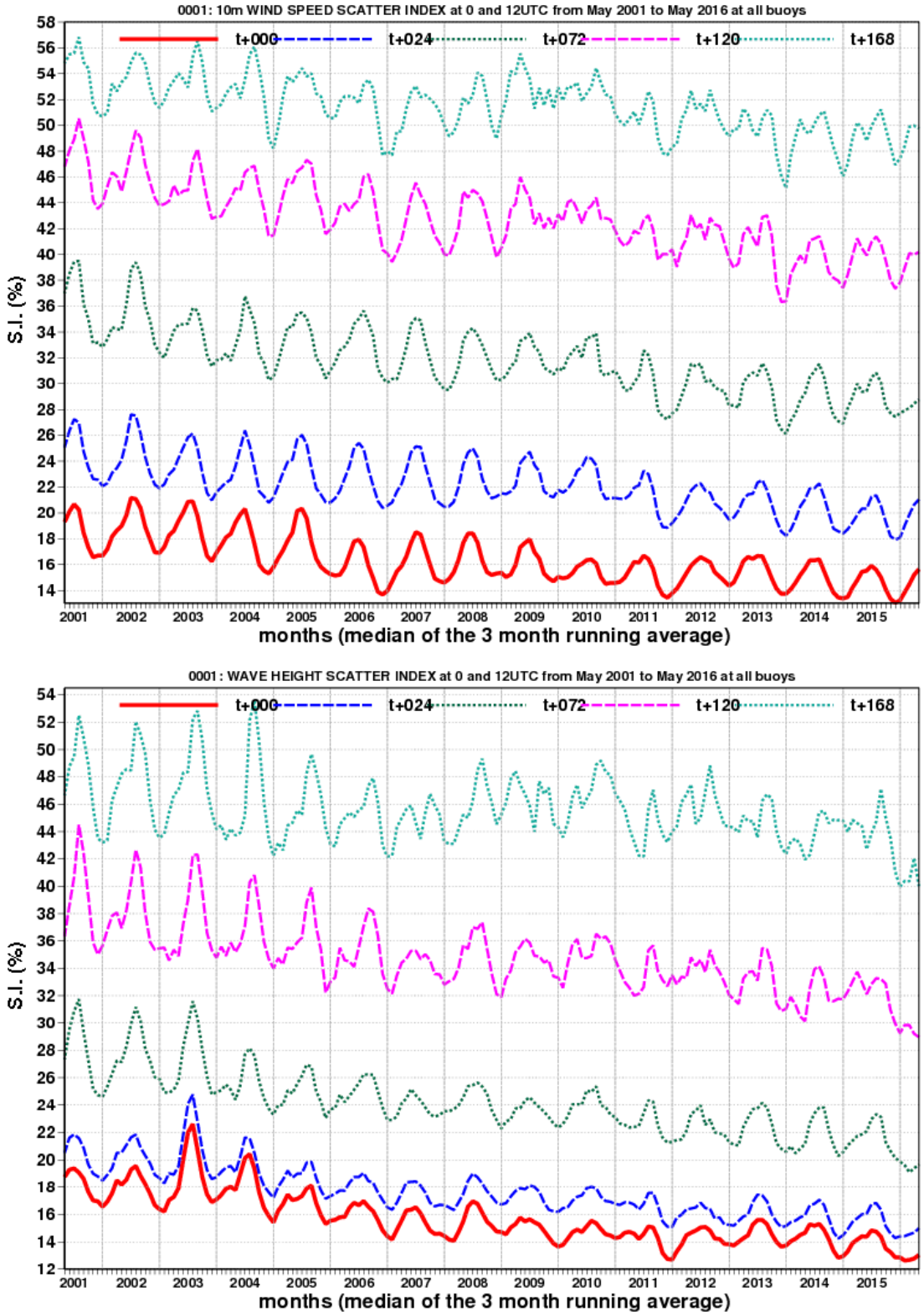


Figure 25: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

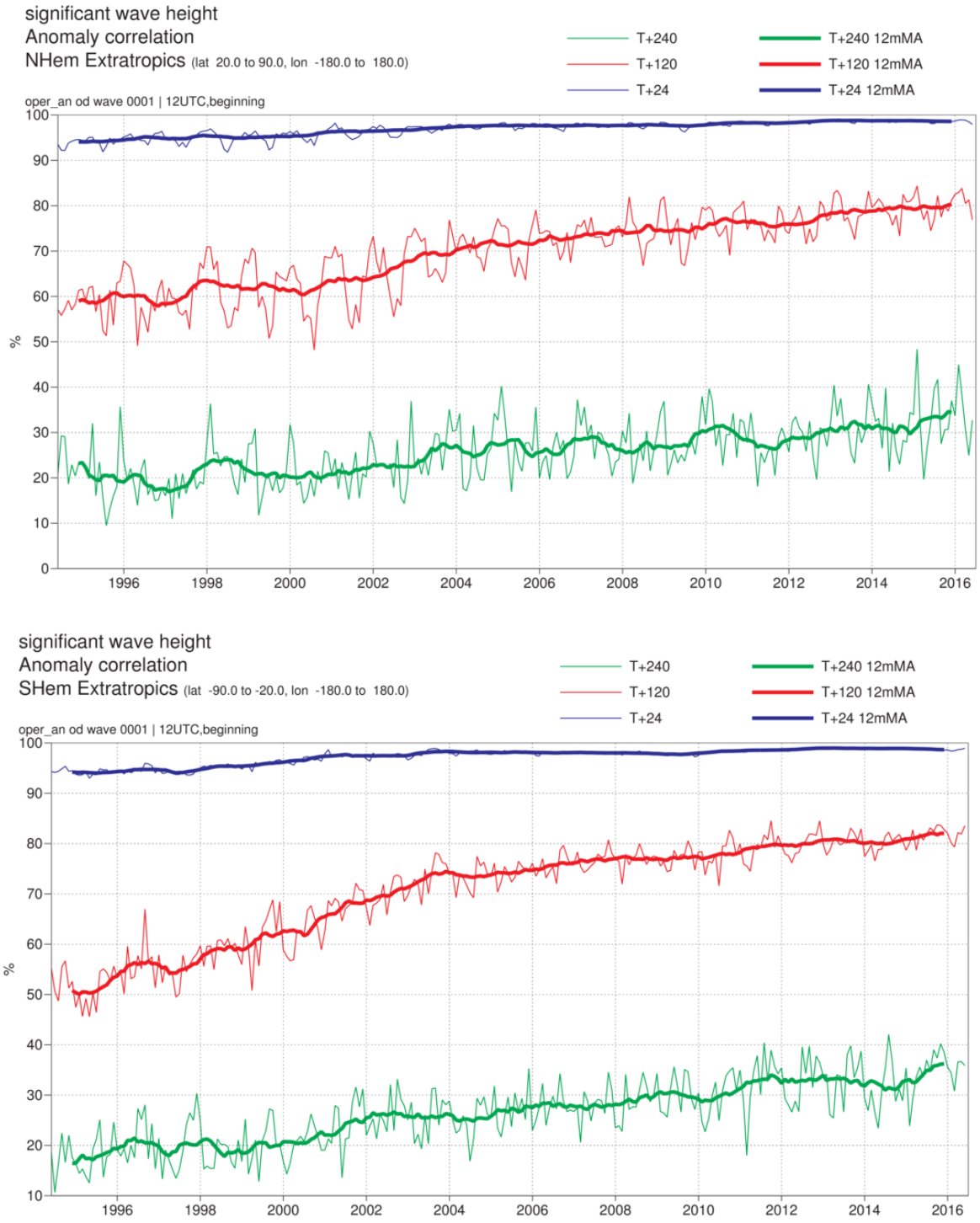


Figure 26: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

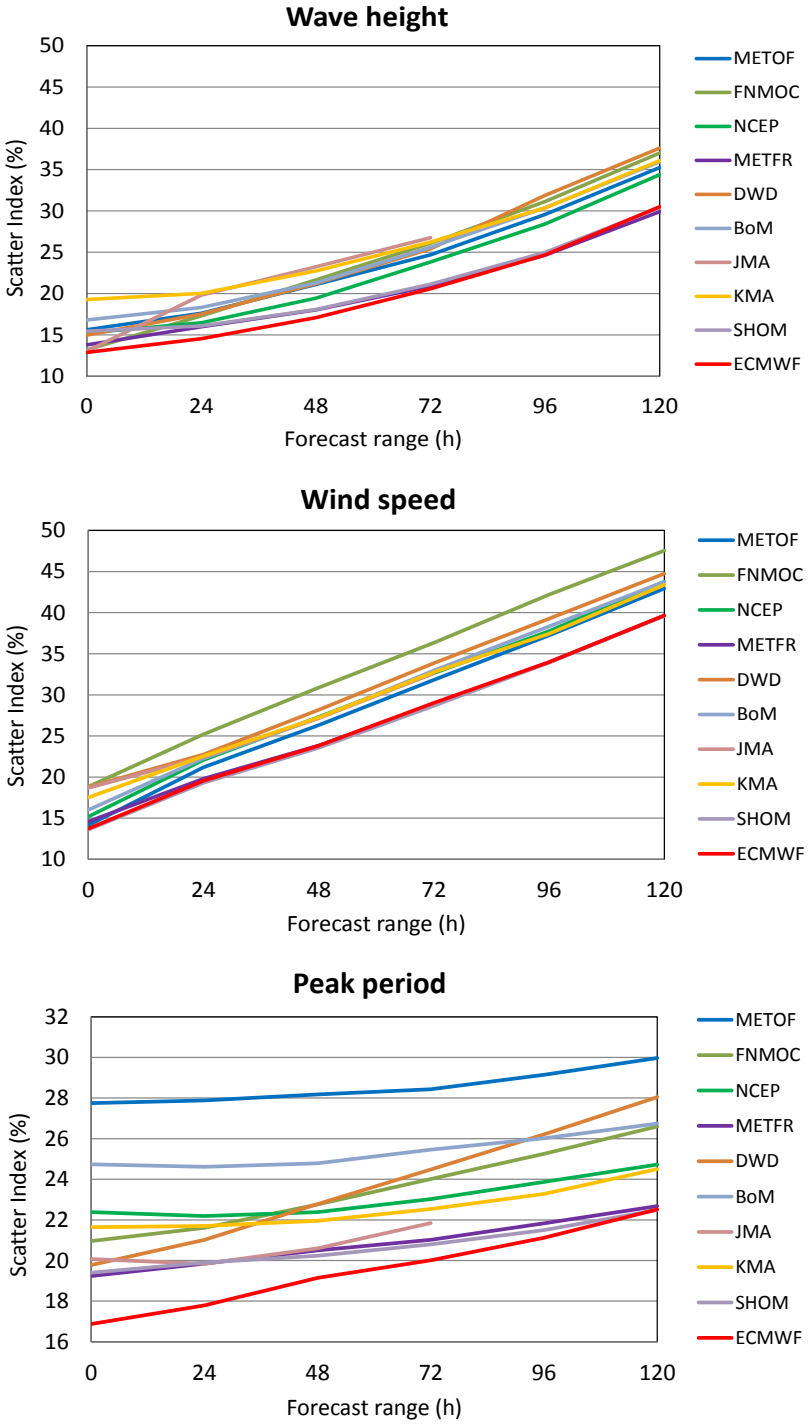


Figure 27: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period June 2015–May 2016. 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; SHOM: Hydrographic and Oceanographic Service of the French Marine.

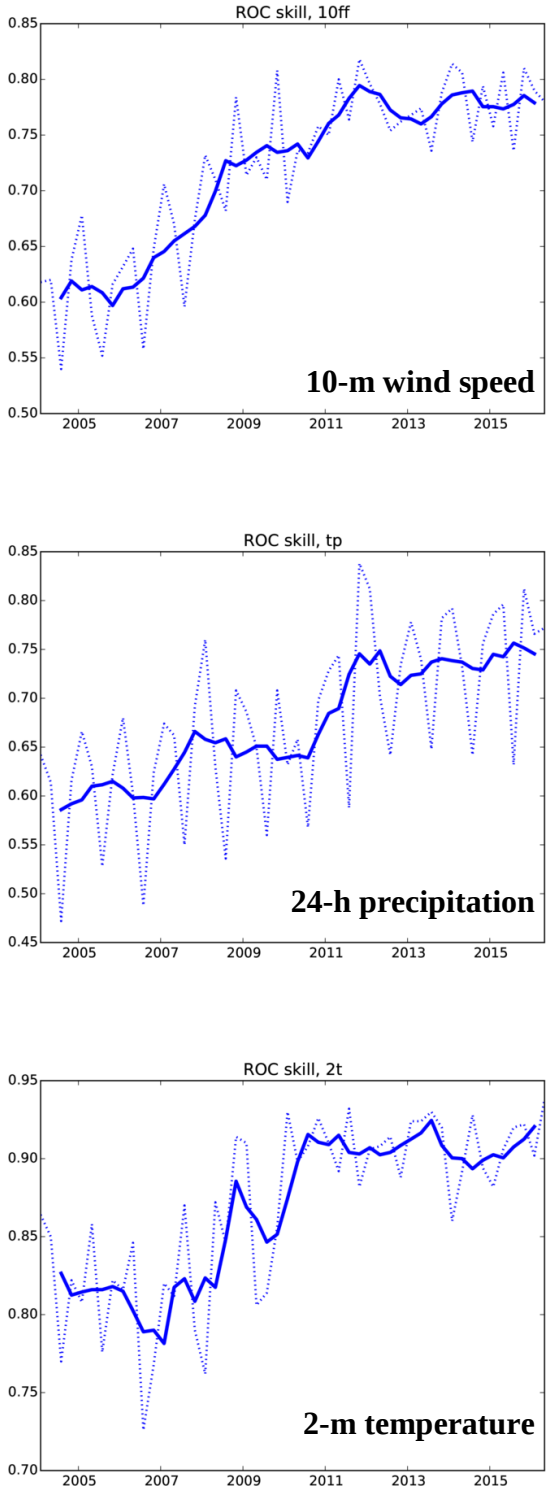


Figure 28: Verification of Extreme Forecast Index (EFI) against analysis. Top panel: supplementary headline score – skill of the EFI for 10 m wind speed at forecast day 4 (24-hour period 72–96 hours ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate. Curves 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.

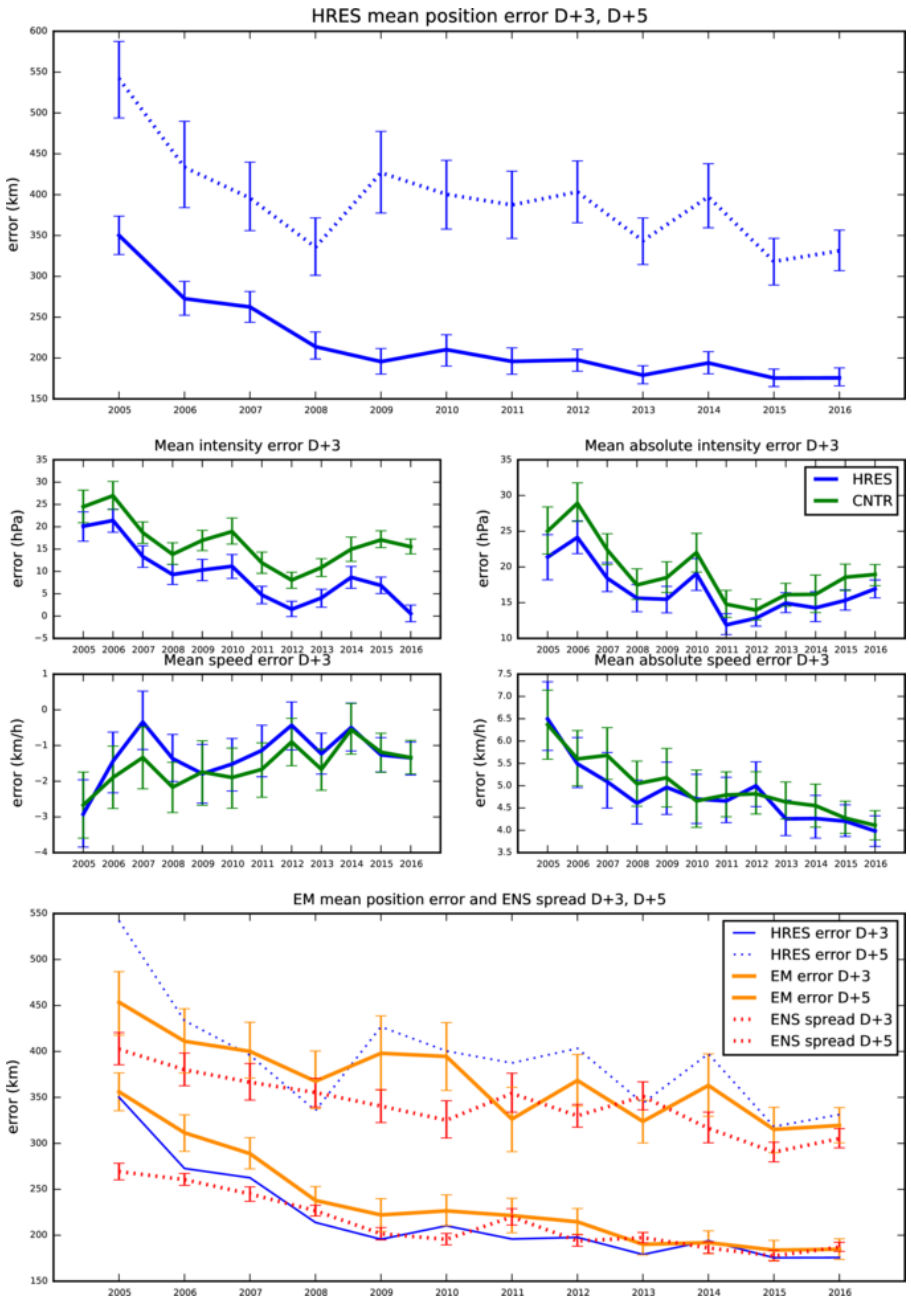


Figure 29: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported via the GTS. Top panel supplementary 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).

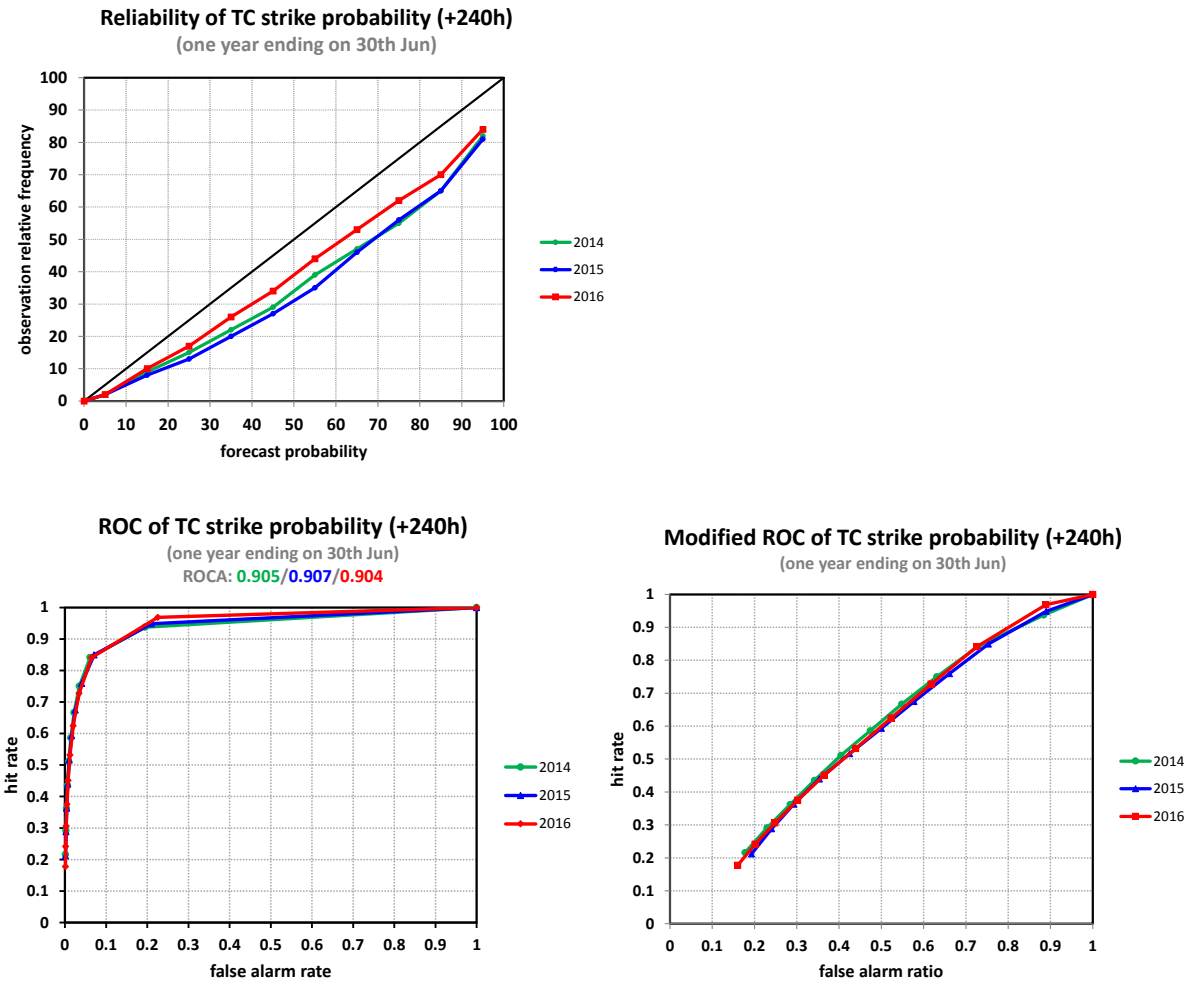


Figure 30: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2013–June 2014 (green), July 2014–June 2015 (blue) and July 2015–June 2016 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

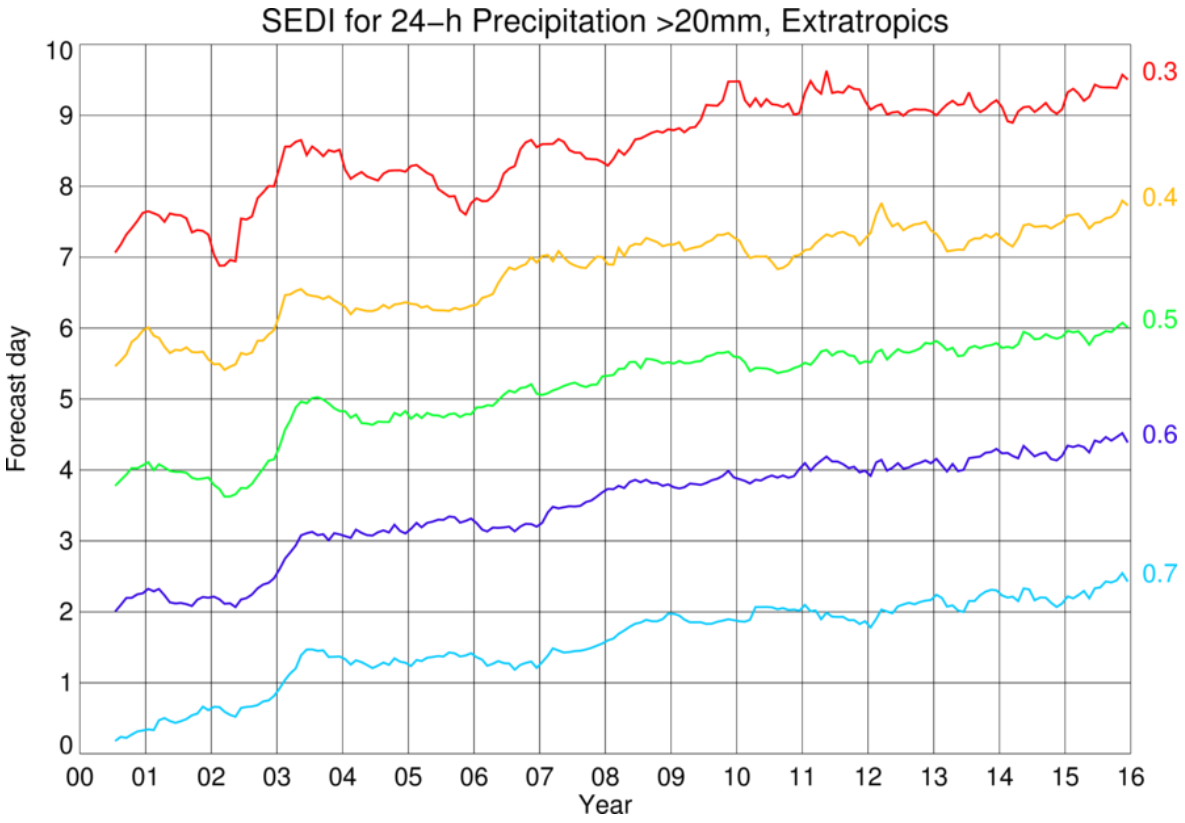


Figure 31: Evolution of skill of the HRES forecast in predicting 24-h precipitation amounts >20 mm in the extra-tropics as measured by the SEDI score, expressed in terms of forecast days. Verification is against SYNOP observations. Numbers on the right indicate different SEDI thresholds used. Curves show 12-month running averages.

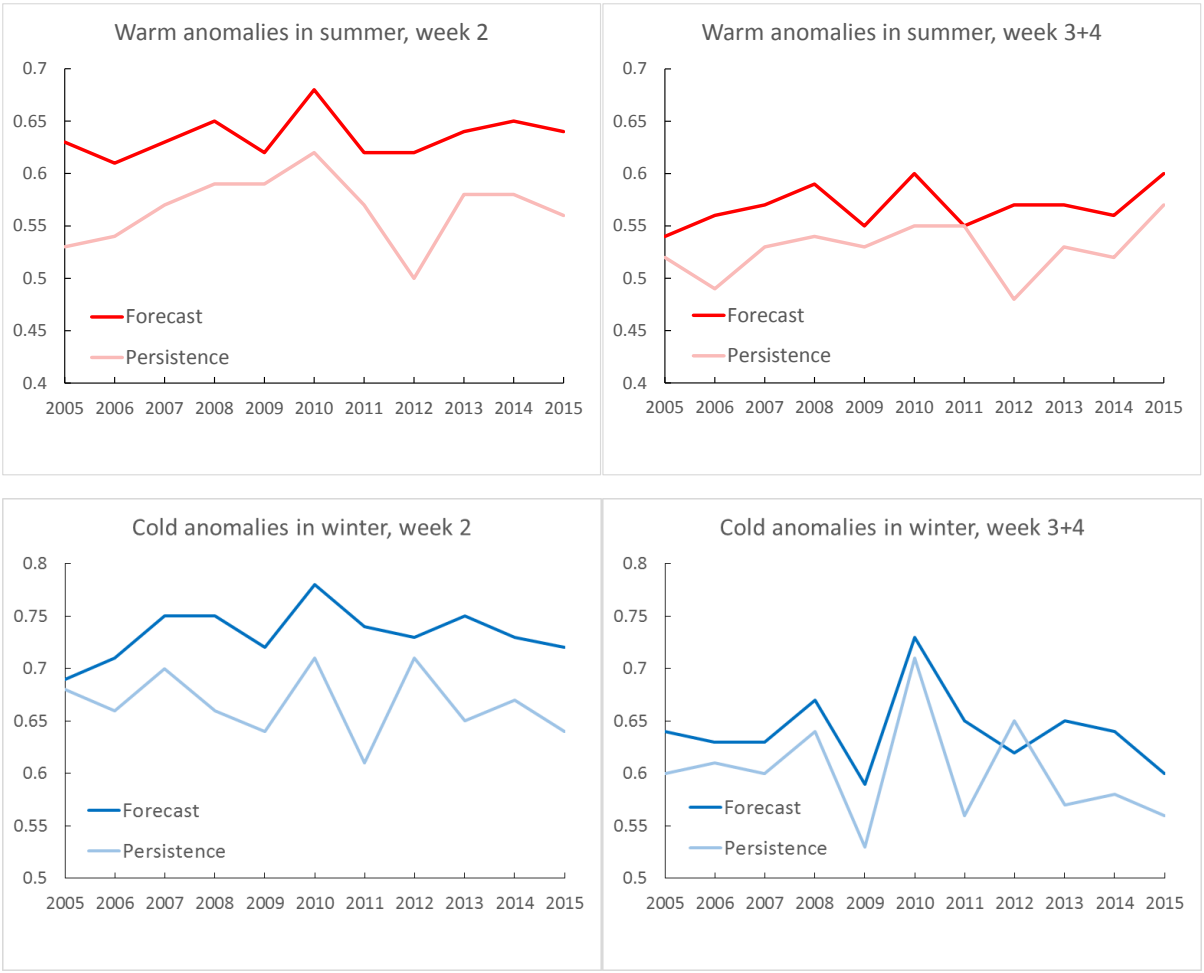


Figure 32: Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all land 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.

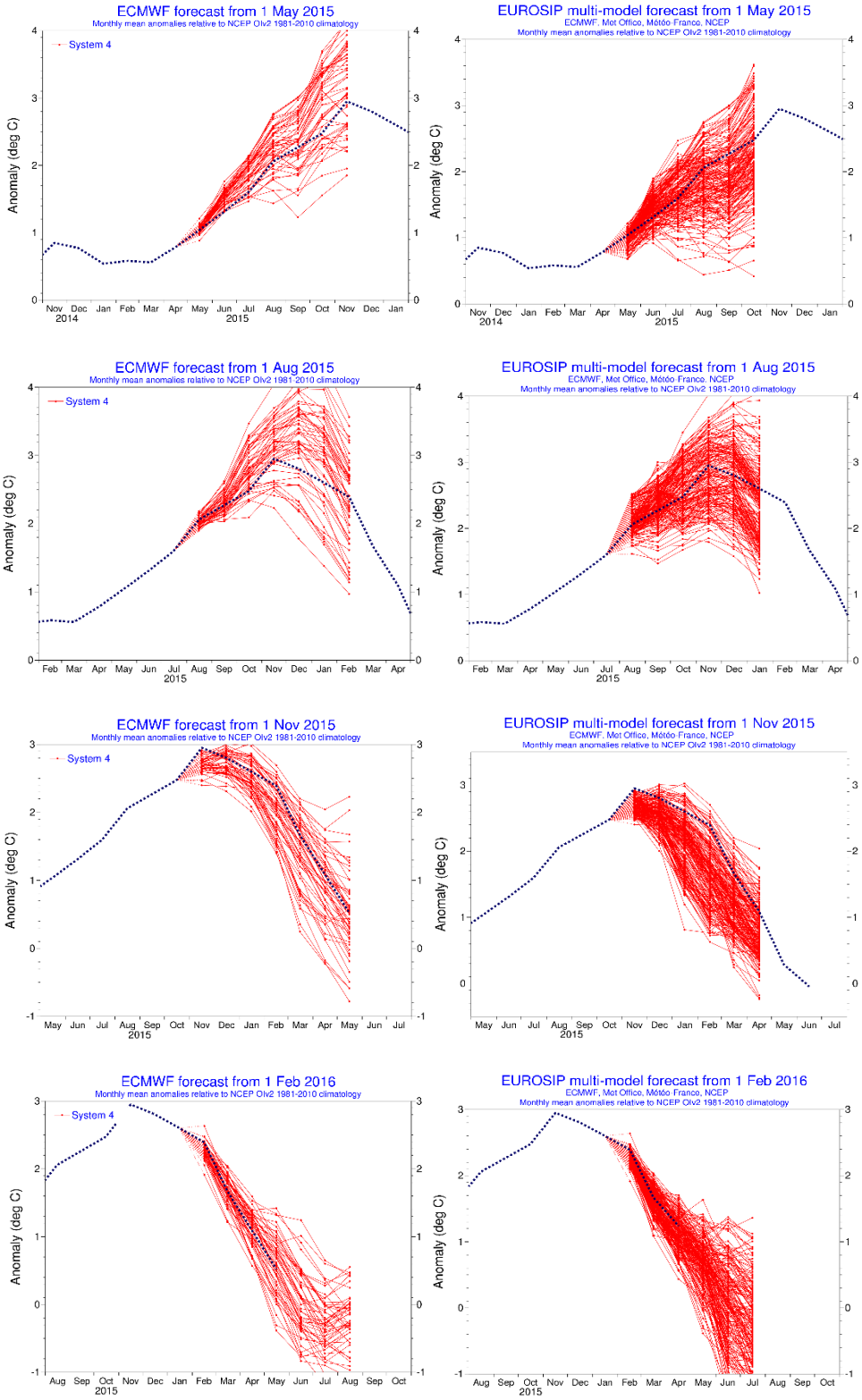


Figure 33: ECMWF (left column) and EUROSIP multi-model forecast (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2015, August 2015, November 2015 and February 2016. The red lines represent the ensemble members; dotted blue line shows the subsequent verification.

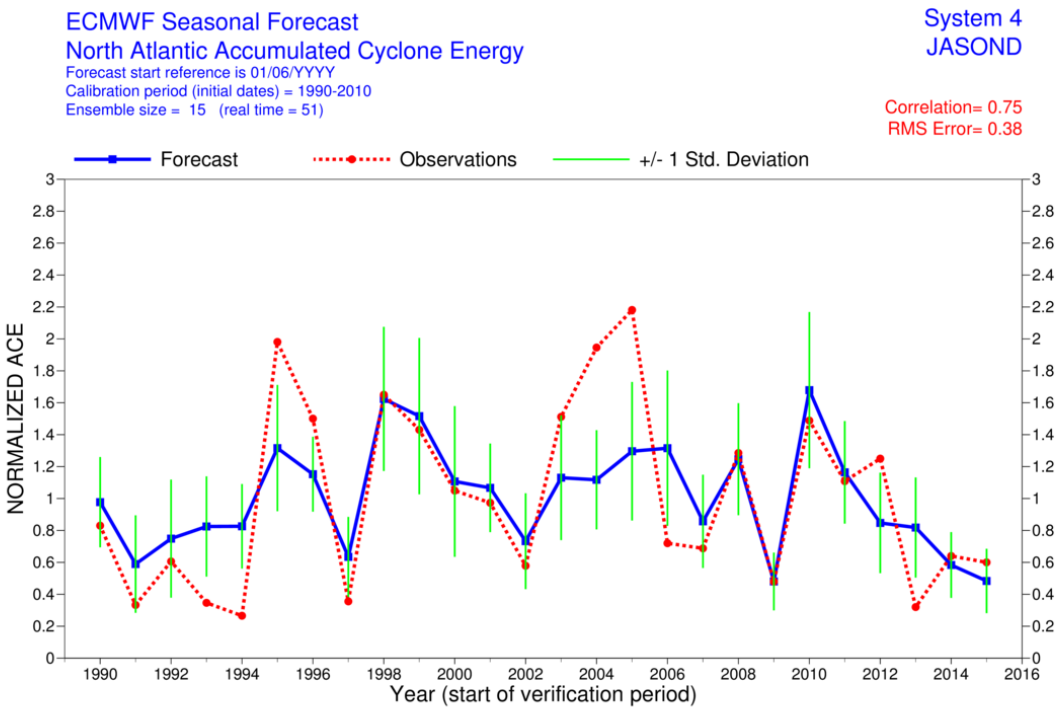


Figure 34: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2015. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts 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.

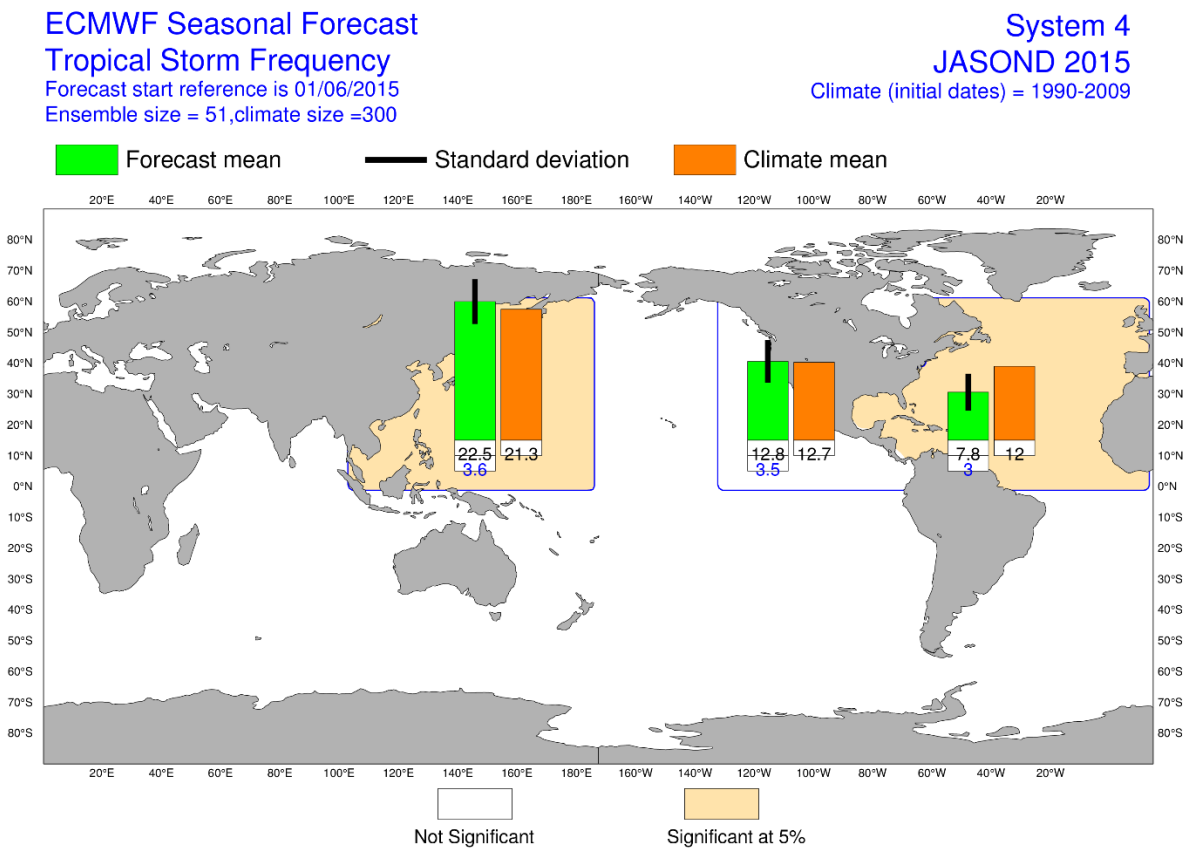


Figure 35: Tropical storm frequency forecast issued in June 2015 for the six-month period July–December 2015. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The 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.

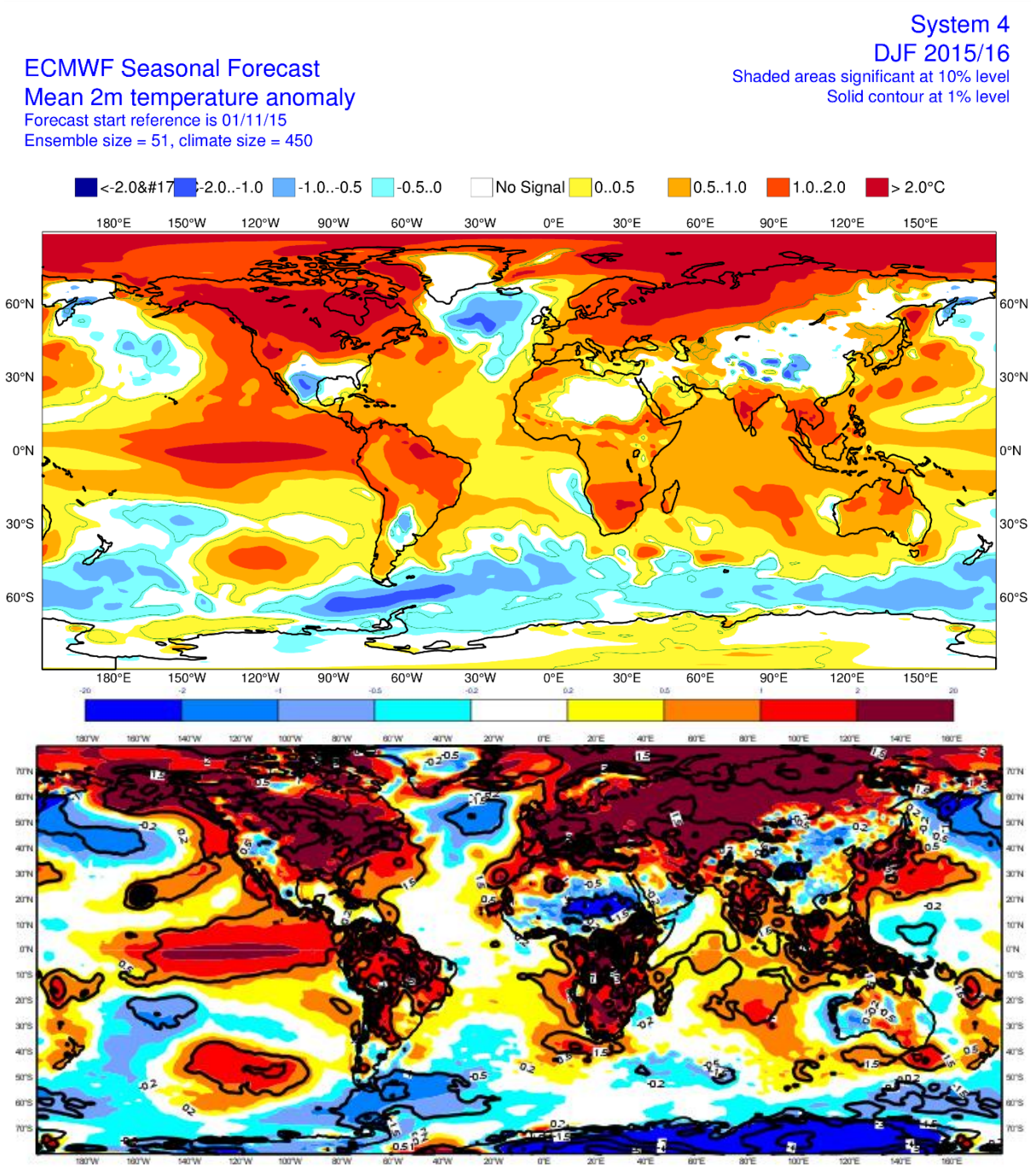


Figure 36: Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2015 for DJF 2015/16 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

A short note on scores used in this report

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 13), the sondes have been selected following an agreement reached by data monitoring centres and published in the WMO WWW Operational Newsletter.

Root mean square errors (RMSE) are the square root of the geographical average of the squared differences between the forecast field and the analysis valid for the same time. When models are compared, each model uses its own analysis for verification; RMSE for winds (Figure 13, Figure 15) 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 26) the climate has been also derived from the ERA-Interim analyses.

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_f 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 6) and its inter-annual variability (Figure 10).

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 30). Figure 30 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 32.

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

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 17, Figure 18). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 19 to Figure 22), verification data are European 6-hourly SYNOP data (area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the four closest grid points, provided the difference between the model and true orography is less than 500 m. A crude quality control is applied to SYNOP data

(maximum departure from the model forecast has to be less than 25 K, 20 g/kg or 15 m/s for temperature, specific humidity and wind speed respectively). 2 m temperatures are corrected for differences between model and true orography, using a crude constant lapse rate assumption provided the correction is less than 4 K amplitude (data are otherwise rejected).

4 Verification of rare events

Experimental verification of deterministic forecasts of rare events is performed using the symmetric extremal dependence index SEDI (Figure 31), 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.