TECHNICAL MEMORANDUM

817

Evaluation of ECMWF forecasts, including 2016-2017 upgrades

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

Forecast Department

December 2017

European Centre for Medium-Range Weather Forecasts
Europäisches Zentrum für mittelfristige Wettervorhersage
Centre européen pour les prévisions météorologiques à moyen

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:
http://www.ecmwf.int/publications/

Contact: library@ecmwf.int

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

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

1. Introduction

Recent changes to the ECMWF forecasting system are summarised in section 2. Verification results of the ECMWF medium-range 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. Two additional scores, proposed by the TAC Subgroup on Verification 2016 to become additional headline scores, have been included for the first time. 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, 792). A short technical note describing the scores used in this report is given at the end of 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

2.1. IFS Cycle 43r1 (22 November 2016)

IFS Cycle 43r1 was an upgrade with many scientific contributions, including changes in data assimilation (both in the EDA and the 4D-Var); in the use of observations; and in modelling. Moreover, ENS hourly fields were made available up to T+90 for the Boundary Conditions optional programme.

2.1.1. Data assimilation methodology

The sea-surface temperature (SST) perturbations used in the EDA have been upgraded to a recently developed climatology based on the HadISST.2 dataset. This makes the perturbations statistically consistent with the error characteristics of the analysis cycles.

The EDA-derived background error estimates used in the high-resolution 4D-Var are now computed at spectral resolution TL399 (previously TL159) and a new wavelet-based filtering algorithm is used to control sampling noise. The background error variance has been increased by ~16%.

The weak constraint option of 4D-Var has been reactivated using a model error forcing term active in the stratosphere above 40 hPa and a new estimate of the model error covariance matrix.

The land surface assimilation of SYNOP screen level observations now accounts for the vertical distance between the observations and model grid points. A new vertical structure function has been introduced that follows the approach used at Environment Canada and at Météo-France in MESAN-SAFRAN. The vertical correlation is expressed as a Gaussian function, consistent with that used for snow depth analysis. This gives more weight to observations from stations that are vertically closer to the model grid point (and less to observations less representative of the model altitude).

A new ocean analysis/re-analysis (ORAS5), based on NEMOVAR with a higher-resolution version of the ocean model NEMO (0.25 degrees with 75 vertical layers: ORCA025Z75) has been implemented. This uses the same ocean model version (NEMO v3.4.1) as ENS. ORAS5 uses a new perturbation strategy for the surface fluxes and to simulate observation errors. It also includes an improved quality-control scheme for ocean observations. Sea ice is assimilated within NEMOVAR, with a weakly coupled assimilation to the ocean dynamics. The analyses have been run from 1975 and continue in real-time to provide initial conditions for the ENS forecasts and re-forecasts.

2.1.2. Satellite observations

Radiance assimilation will now take the viewing geometry more fully into account, by evaluating the radiative transfer along slantwise paths (instead of vertically). This is done for all clear-sky sounder radiances when interpolating model fields to observation locations.

A better treatment of observation uncertainty for IASI and CrIS has led to updated observation error covariance matrices and a change of ozone anchor channels in bias correction.

The channel selection for the hyperspectral infrared instrument CrIS has been revised and now uses 117 rather than 77 channels.

The aerosol detection scheme for IASI has been revised making it independent of the bias correction. The scheme is also applied to both CrIS and AIRS.

2.1.3. Model changes

A new CAMS ozone climatology is now used, consisting of monthly means of a re-analysis of atmospheric constituents (CAMSiRA) for the period 2003 to 2014.

Changes to boundary layer cloud for marine stratocumulus and at high latitudes.

Modifications to surface coupling for 2 metre temperature.

Assimilation of snowfall from the NEXRAD RADAR network over the USA.

New model output fields include four cloud and freezing diagnostics (for aviation), a new direct-beam solar radiation diagnostic and improvements to the sunshine duration diagnostic.

2.1.4. Medium-range/monthly ensemble

An interactive sea-ice model (the Louvain-la-Neuve Sea Ice Model - LIM2) has been introduced so that sea-ice cover now evolves dynamically. Previously it was persisted for 15 days; over the next 30 days of the forecast, it was relaxed towards the climatology of the previous 5 years.

Ocean initial conditions are taken from ORAS5 instead of ORAS4.

A global fix for tendency perturbations in the stochastic model error scheme SPPT to improve global momentum, energy and moisture conservation properties.

The land initial conditions of the ENS re-forecasts are taken from a new land surface simulation at the native ENS resolution (TCO639, ~16km), replacing the previous configuration that used ERA-Interim/Land (at TL255 resolution, ~80 km)

2.1.5. Meteorological impact of the new cycle

A comparison of scores between IFS cycle 43r1 and IFS cycle 41r2 for HRES and ENS is provided in the form of scorecards in Figure 1 and in Figure 2.

Upper air

The new model cycle provides improved high-resolution forecasts (HRES) and ensemble forecasts (ENS) throughout the troposphere and lower stratosphere. In the extra-tropics, error reductions of the order of 0.5-1% 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 1 h.

Improvements are most consistently seen in verification against the model analysis. In the tropics, there is a small degradation (both against analysis and observations) of temperature near the tropopause in terms of root mean square error (RMSE) but not in terms of anomaly correlation. This is due to a slight cooling caused by a modification in the treatment of cloud effects in the vertical diffusion scheme, which overall leads to improved cloud cover. While there is a consistent gain for upper-air parameters on the hemispheric scale, some continental-scale areas, such as North America and East Asia, show statistically significant improvements only at some levels and for some parameters.

Increases in upper-air skill of the ENS are generally similar to the HRES, with a substantial gain for mean sea level pressure. 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 small (of the order of 0.5 h). The spread-error relationship is generally improved, partly due to reduced error and partly due to increased spread. For some parameters this improvement is quite significant, such as the 850 hPa wind speed in the tropics, where the under-dispersion is reduced by about 20% in the medium range.

Weather parameters and waves

The new model cycle yields consistent gains in forecast performance in the tropics and extra-tropics for total cloud cover, mostly due to a reduction of the negative bias in low cloud cover.

Changes in precipitation over land areas are small and overall neutral.

The increase in forecast skill for 2 m temperature is most pronounced in the short- and medium-range, where it amounts to ~1% reduction of the RMSE in the northern hemisphere extra-tropics, and up to 2% over some land areas such as Europe and North America. In the tropics there is an increase of 0.5-1% in the RMSE for 2 m temperature, connected to a slight increase of the overall cold bias at low latitudes. In the ENS there is a significant improvement in 2 m temperature amounting to a 3% reduction in the continuous ranked probability score (CRPS) in Europe.

There is an increase of the RMSE of 2 m humidity by about 1% in winter associated with the introduction of limited evapotranspiration when the uppermost soil layer is frozen. This change contributes to the improvements in 2 m temperature.

10 m wind speed shows error reductions of 0.5-1% over the ocean, leading to improvements in significant wave height and mean wave period, especially in the tropics and southern hemisphere. Over land areas, changes in 10 m wind speed forecast skill are generally neutral to slightly positive.

Monthly forecast

Verification results show a modest positive effect on skill scores although the differences are not statistically significant. There is a substantial improvement in the skill scores for the Madden-Julian Oscillation (MJO), corresponding to a gain in lead time of 0.5-1 day at a forecast range of 4 weeks. Also, MJO spread is increased, bringing it closer to the RMSE. Verification of precipitation against analysis shows some degradation in the tropics which is not statistically significant, and a reduction of precipitation biases in the northwest Pacific.

Sea ice

The new cycle introduces a prognostic sea-ice model, leading to a significant reduction of the RMSE of sea ice fraction in the later medium range.

2.2. IFS Cycle 43r3 (11 July 2017)

This upgrade of ECMWF’s Integrated Forecasting System implemented on 11 July improves forecast skill in medium-range and monthly forecasts. IFS Cycle 43r3 includes changes in the model and in the assimilation of observations. Model changes include a new radiation scheme, improvements in the modelling of convection, and a new aerosol climatology. Changes in data assimilation and in the way dropsonde observations are handled have improved the accuracy of the initial conditions on which forecasts are based, especially for tropical cyclones.

2.2.1. Data assimilation

Improved humidity background error variances directly from the EDA like for all other variables.

Revised wavelet filtering of background error variances and revised quality control of drop-sonde wind observations in 4D-Var to improve tropical cyclone structures.

2.2.2. Satellite observations

Increased use of microwave humidity sounding data by adding new sensors (SAPHIR, GMI 183 GHz channels).

Activation of 118 GHz channels over land from MWHS-2 instrument on-board FY-3C. Harmonised data usage over land and sea-ice for microwave sounders (adding MHS channel 4 over snow, adding some ATMS channels, lower observation errors for MHS data over land).

Improved screening of infrared observations for anomalously high atmospheric concentrations of hydrogen cyanide (HCN) from wildfires.

Improved quality control for radio occultation observations and radiosonde data.

2.2.3. Model changes

New, more efficient radiation scheme with reduced noise and more accurate longwave radiation transfer calculation.

New aerosol climatology based on 'tuned' CAMS aerosol re-analysis including dependence on relative humidity.

Increased super-cooled liquid water at colder temperatures (down to -38C) from the convection scheme.

Visibility calculation changed to use 'tuned' CAMS aerosol climatology.

2.2.4. Meteorological impact of the new cycle

A comparison of scores between IFS cycle 43r3 and IFS cycle 43r1 for HRES is provided in the form of a scorecard in Figure 3.

Results for the HRES are positive, with many of the scores over NH, SH and Europe indicating statistically significant improvements at the 95% level up to about forecast day-5 when forecasts are verified against own analysis. When forecasts are verified against observations, the positive impact of 43r3 is also evident. Improvements are larger in summer than in winter and are to a large extent due to improvements in the humidity background error, as well as changes to the deep convection scheme and the aerosol climatology, which improved the temperature gradient between extra-tropics and tropics. Improvements are significant for temperature and vector wind throughout the extra-tropical troposphere. In the tropics there is some deterioration in temperature and humidity at certain vertical levels associated with the changes to the deep convection scheme. Surface parameters show partially statistically significant improvements both in the tropics and extra-tropics (2 m humidity, 10 m wind speed, total cloud cover, precipitation), except for 2 m temperature which shows neutral results.

Over the ocean, statistically significant improvements are seen for verification against own analysis for 10 m wind speed, significant wave height, and mean wave period.

Results for the ENS (Figure 4) are mainly positive and similar to the HRES both for upper-air and surface variables for the NH, SH, and Europe when verified against analysis. In the tropics there is some deterioration in upper tropospheric wind speed and lower tropospheric temperature associated with reduced spread. There are also some slight deteriorations in tropical 2 m temperature and precipitation scores.

Changes in the tropical cyclone analysis are notable, with the cyclone structure defined in a better way. At forecast day-1 there is a marginally significant improvement in position error; the improvement is undetectable thereafter. Tropical cyclone intensity (as measured by central pressure) is slightly reduced from day-2 onwards: for lead times beyond four days this has a beneficial effect since it reduces the existing negative bias in tropical cyclone central pressure in such forecasts.

3. Verification of upper-air medium-range forecasts

3.1. ECMWF scores

Figure 5 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 the signal-to-noise ratio is smaller due to the smaller area, and the positive trend there is less pronounced than in the hemispheric scores.

A complementary measure of performance is the root mean square (RMS) error of the forecast. Figure 6 shows RMS errors for both extratropical hemispheres of the six-day forecast and the persistence forecast. In both hemispheres the error of the six-day forecast continues to decrease at the rate seen in recent years.

Figure 7 shows the time series of the average RMS difference between four- and three-day (blue) and six- and five-day (red) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the general downward trend indicates that there is less “jumpiness” in the forecast from day to day. The level of consistency between consecutive forecasts in the northern extratropics has increased further in the last year. Curves for Europe are subject to larger inter-annual variability so that the downward trend does not show up as clearly as in the extratropics.

The quality of ECMWF forecasts for the upper atmosphere in the northern hemisphere extratropics is shown through time series of temperature and vector wind scores at 50 hPa in Figure 8. There are downward trends in RMSE for wind speed and for day-1 temperature. The recent increasing trend for day-5 temperature (as well as the maximum around 2011) is mainly due to an increase in bias, as confirmed by the trend in error standard deviation (not shown) for day-5 temperature, which is downward as well.

The trend in ENS performance is illustrated in Figure 9, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. In Europe the recent 12-month mean ENS skill is comparable to 2010, when predictability was exceptionally high. In the extratropics, the 12-month mean has for the first time exceeded 9 days over an extended period.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 10. For 500 hPa geopotential height, the forecast shows a stronger under-dispersion than in previous years from day-10 onwards. For 850 hPa temperature, the under-dispersion has been reduced up to day-7 but increased for larger lead times.

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 11 and Figure 12 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 12 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 10, and ideally should lie on the diagonal, as closely as possible to the lower left corner. Also in this respect ECMWF usually performs best among the global models, with the exception of 850 hPa temperature in the tropics in the short range, where JMA has the lowest error (although ECMWF has the better match between error and spread).

In order to have a benchmark for the ENS, the CRPS has been computed for a ‘dressed’ ERA-Interim. 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-Interim. Figure 13 shows the evolution of the CRPS for the ENS and for the dressed ERA-Interim over the last 12 years for temperature at 850 hPa at forecast day-5. The improvement due to recent model upgrades shows up clearly in a reduction of CRPS and increase in relative skill. In the northern hemisphere the skill of the ENS relative to the reference forecast has now reached 33%. In the southern hemisphere, the corresponding value is 30%.

The forecast performance in the tropics, as measured by RMS vector errors of the wind forecast with respect to the analysis, is shown in Figure 14. At 200 hPa (upper panel) the 1-day forecast errors have changed very little, while the 5-day forecast errors have reached their lowest level so far. A somewhat similar behaviour is seen at 850 hPa (lower panel) where the recent error reduction is more visible at day-5 than at day-1. 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 model inter-comparison plots shown in this section are based on 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.

Figure 15 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, and ECMWF continues to maintain its lead over other centres.

WMO-exchanged scores also include verification against radiosondes over regions such as Europe. Figure 16 (Europe), and Figure 17 (northern hemisphere extratropics) showing both 500 hPa geopotential height and 850 hPa wind forecast errors averaged over the past 12 months, confirm the leading performance of the ECMWF forecasts relative to the other centres in direct verification against observations.

The comparison for the tropics is summarised in Figure 18 (verification against analyses) and Figure 19 (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 18) is sensitive to details of the analysis method, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 19), the ECMWF forecast has the smallest overall errors in the medium range. However, ECMWF’s lead in the tropics is smaller than 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 20. 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 SYNOP observations.

The deterministic precipitation forecast has reached its highest level of skill so far (Figure 20). The lead time at which the given threshold is reached has increased by almost one forecast day since 2009 when the SEEPS score was developed. There is considerable variation in the score due to atmospheric variability, as shown by comparison with the ERA-Interim reference forecast (green line in Figure 20, 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 20).

Since ERA-Interim is beginning to lose skill due to its inability to adapt to some of the changes in the observation system, also the difference between HRES and ERA5 is shown in Figure 20. The improvement of the HRES relative to ERA5 is about 0.1 forecast days smaller than relative to ERA-Interim. It is also apparent that ERA5 removes variations due to atmospheric variability even more efficiently than ERA-Interim (see also Figure 27 which shows this for other weather parameters).

The probabilistic precipitation score (lower panel in Figure 20) shows a long-term improvement as well, however the peak at the end of 2015 is partly due to atmospheric variability, hence the values seen at the end of 2016 are more representative of the actual current level of skill.

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 both the HRES and ENS leading with respect to the other centres (Figure 21). While other centres have positive skill out to day-6, ECMWF’s probabilistic precipitation forecasts retain positive skill up to day-9.

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 22 to Figure 25. Verification is performed against synoptic observations received via the Global Telecommunication System (GTS). The matching of forecast and observed value uses the nearest grid-point method. A correction for the difference between model orography and station height is applied to the temperature forecasts.

For 2 m temperature (Figure 22) there is a visible reduction in the error standard deviation (upper curves) in the last two years. With regard to bias (lower curves), the springtime cold bias still persists, however the night-time warm bias in summer is getting smaller. Also the 2 m dewpoint (Figure 23) shows a reduction of the error standard deviation, however the daytime negative bias has become slightly bigger in recent years. For total cloud cover (Figure 24) the error standard deviation is showing little change, while the bias is now quite small both during the day and at night. For wind speed (Figure 25) the error standard deviation continues to decrease, with the winter 2016-17 showing about 10% reduction in error compared to previous winter seasons. In terms of bias, the weak trend to more negative values during the day appears to continue.

To complement the evaluation of surface weather forecast skill, verification is also performed against top of the atmosphere (TOA) reflected solar radiation (daily totals) from the Climate Monitoring Satellite Application Facility (CM-SAF), based on Meteosat data. Shown is the relative improvement compared to ERA-Interim (Figure 26). In the northern extratropics we see a continuation of the upward trend of recent years. In the tropics, the slight downward trend since 2014 has been halted.

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 27 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. All parameters show the beneficial effect of recent model upgrades. It is worth noting that even the skill for total cloud cover, which has been stagnant prior to 2012, shows some moderate improvement in recent years.

Now that ERA5 forecasts have become available for certain periods, the presumed loss of skill of ERA-Interim (mainly due to the fact that it cannot adapt to some of the changes to the observing system) can be quantified. The thick curves in Figure 27 show the skill of HRES v ERA-Interim, but corrected for the difference between ERA-Interim and ERA5. It can be seen that the corrected gain in skill especially for upper-air parameters (including msl pressure) is lower by 3–4%, and that there is less impact on surface parameters. It is also apparent that ERA5 removes variations due to atmospheric variability more efficiently than ERA-Interim, i.e. the resulting curves are smoother (see also Figure 20 which shows this for precipitation). Once the whole verification period will be covered by ERA5, it will replace ERA-Interim as a reference system for evaluation.

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 28. The top panel of Figure 28 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 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. In both hemispheres, anomaly correlation for significant wave height in the medium-range has reached its highest value so far (Figure 29).

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 30 for the 12-month period June 2016–May 2017. ECMWF forecast winds are used to drive the wave model of Météo France, hence the almost identical wind errors of Météo France and ECMWF in Figure 30. For both wave height and peak period, ECMWF generally manages to outperform the other centres.

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 by the blue lines in Figure 31 (top), together with results for days 1–3 and day-5. Corresponding results for 24-hour total precipitation (centre) and 2 m temperature (bottom) are shown as well. Each plot contains seasonal values, as well as the four-season running mean, of ROC area skill scores from 2004 to 2016; the final point on each curve includes the spring (March–May) season 2017. For all three parameters, ROC skill has stabilized on a high level, with some inter-annual variations due to atmospheric variability. The clearest indication for a positive trend in recent years is seen for 24-h precipitation.

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 32. Errors in the forecast central pressure of tropical cyclones are also shown. The comparison of HRES and ENS control (central four panels) demonstrates the benefit of higher resolution for tropical cyclone forecasts.

Both HRES and ENS position errors (top and bottom panels, Figure 32) have reached their lowest values so far. Mean absolute intensity errors of the HRES and the CTRL at D+3 have decreased compared to the previous year but are still larger than in 2011–12. Mean absolute speed errors have slightly increased compared to the previous year.

The bottom panel of Figure 32 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. The forecast was generally under-dispersive before the resolution upgrade in 2010, but the spread-error relationship has improved since then. However, in the latest season, there has been a change to over-dispersive spread. This is a result of the resolution increase in the EDA which will be corrected in model cycle 45r1 by removing the enhanced inflation of singular vectors in the tropics compared to the extra-tropics. The figure also shows that the HRES position and ENS position errors have become very similar recently.

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 33. Results show a certain amount of over-confidence, however, reliability has been higher in 2016 and 2017 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. Highest values of these two measures are found in the 2017 season.

5.3. Additional severe-weather diagnostics

While many scores tend to degenerate to trivial values for rare events, some have been specifically designed to avoid this problem. 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 34 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 recent 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 35 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 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 (SEAS4) 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 SEAS4 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 SEAS4 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 2016 was characterized by a change to slightly negative anomalies from very warm conditions in the eastern tropical Pacific associated with a strong El Nino. This was well captured in the forecast (top row), although the duration of the subsequent period of negative anomalies (7 months) was somewhat underestimated. Here EUROSIP gave a slightly better guidance, albeit with relatively large spread (Figure 36, 2nd row). The return to positive anomalies was very well predicted by ECMWF (Figure 36, 3rd row), however the return to a full-fledged El Nino, suggested by many ensemble members, did not verify (Figure 36, 4th row).

6.2.3. Tropical storm predictions from the seasonal forecasts

The 2016 Atlantic hurricane season had a total of 15 named storms including 7 hurricanes and 4 major hurricanes. It was an active season (the most active since 2012) with an accumulated cyclone energy index (ACE) of about 140% of the 1990–2010 climate average (Figure 37). The seasonal forecast predicted a noticeable increase in ACE compared to 2015, however the observed increase was even larger. Seasonal tropical storm predictions from SEAS4 indicated average activity compared to climatology over the Atlantic. Similarly, the number of tropical storms which formed in 2016 (15) was above average (12) whereas the forecast predicted 11.5 (with a range from 8.1 to 14.9) tropical storms in the Atlantic (Figure 38).

Figure 38 also shows that SEAS4 predicted above average activity over the eastern North Pacific, and average activity over the western North Pacific. In the western North Pacific, SEAS4 predicted 21 storms, and 24 were observed in the period July to December. In the eastern North Pacific, although SEAS4 did predict an above average season, the tropical storm activity was underestimated, with 19 observed and 14 predicted.

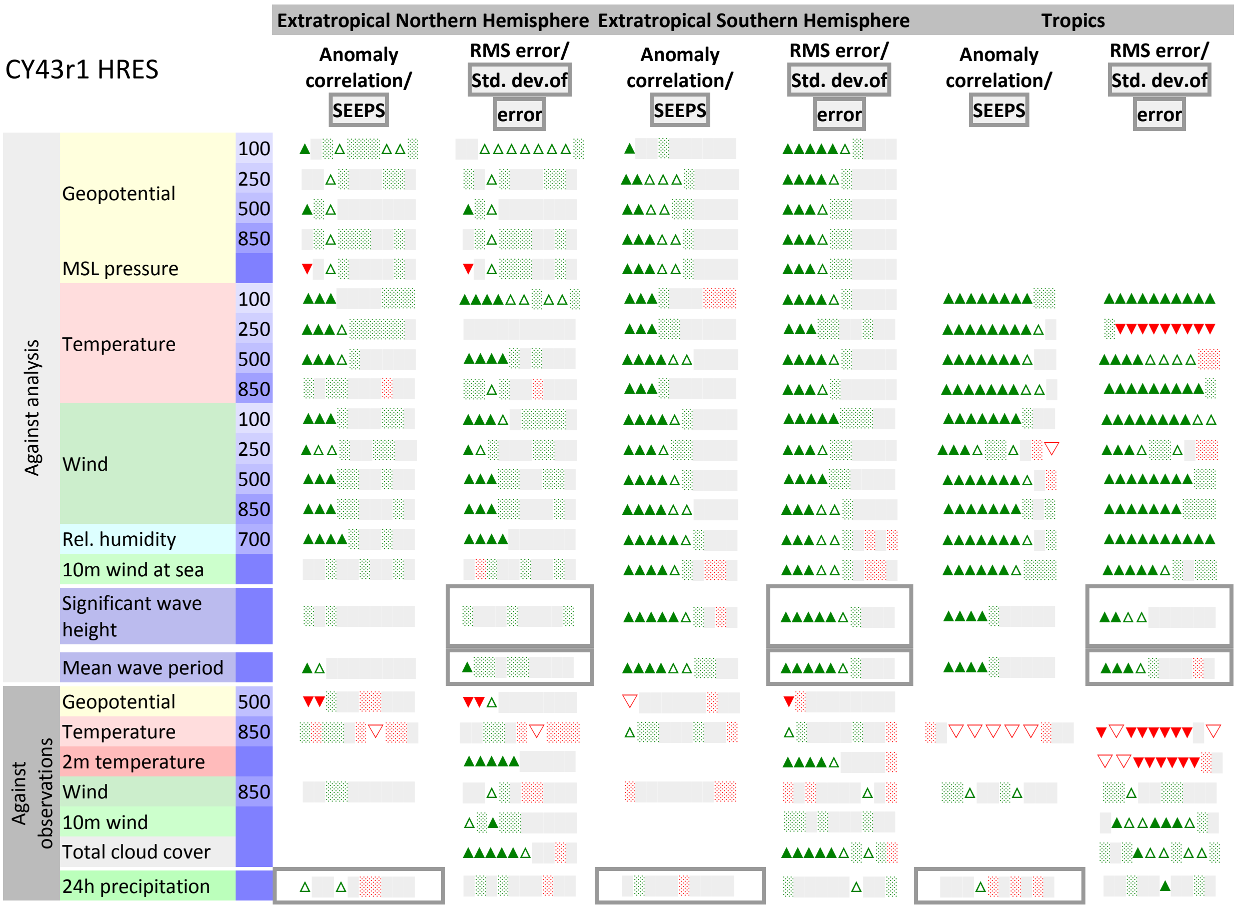
6.2.4. Extratropical seasonal forecasts

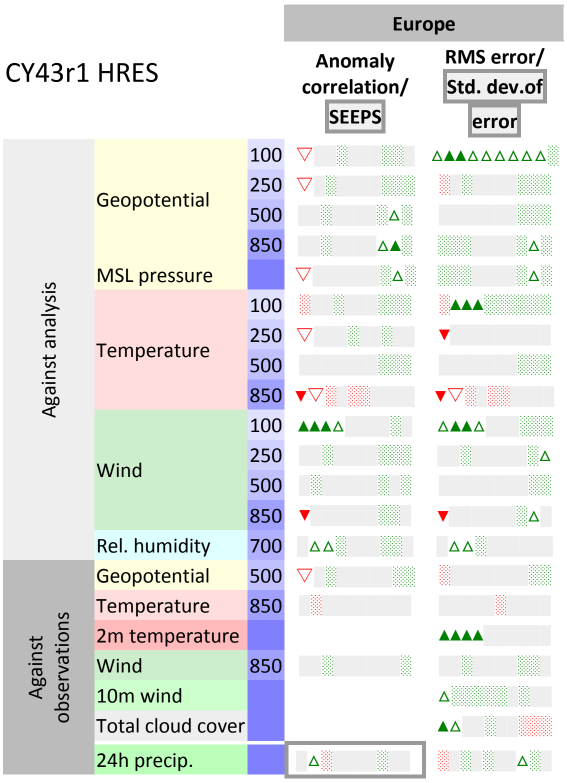
Since the El Nino of 2015–16 had ended in early summer 2016, a general drop in seasonal predictive skill had to be expected for 2016–17. Nevertheless, there were some large-scale temperature anomalies which the forecast captured to some extent.

The pattern of 2 m temperature in the northern-hemisphere winter (DJF 2016–17) was characterized by strong warm anomalies in the Arctic, and over North America and Eurasia. These warm anomalies, which are a combination of the effect of global warming and inter-annual variability, were captured reasonably well by the seasonal forecast in North America, but not in Eurasia, where over a large area no statistically significant anomalies were predicted (Figure 39). In the southern hemisphere, the general large-scale pattern of warm and cold anomalies was predicted quite well.

Parts of Europe experienced a hot summer season in 2017. For the northern-hemisphere summer (JJA 2017) the forecast predicted positive anomalies over Southern Europe and the Mediterranean, as well as parts of Siberia (Figure 40). Again, part of this pattern is due to global warming, as confirmed by comparison with anomalies based on a more recent climatological reference period (not shown). The verifying analysis confirms the basic pattern, however the observed cold anomaly stretching from Scandinavia eastward was only indicated as a minimum in warming.

Climagrams for Northern and Southern Europe for summer and autumn 2017 are shown in Figure 41. Red squares indicate observed monthly anomalies. In Northern Europe (top panel), the anomalies of individual months were not captured. In Southern Europe, the observed strongly positive anomaly in June was close to the 95th percentile of the forecast distribution. The magnitudes of the positive anomalies in May and July were predicted quite well.





Symbol legend: for a given forecast step...

▲ CY43r1 better than CY41r2 statistically significant with 99.7% confidence

△ CY43r1 better than CY41r2 statistically significant with 95% confidence

▒ CY43r1 better than CY41r2 statistically significant with 68% confidence

□ not really any difference between CY41r2 and CY43r1

▒ CY43r1 worse than CY41r2 statistically significant with 68% confidence

▽ CY43r1 worse than CY41r2 statistically significant with 95% confidence

▼ CY43r1 worse than CY41r2 statistically significant with 99.7% confidence

Figure 1: Summary score card for Cy43r1. Score card for HRES cycle 43r1 versus cycle 41r2 verified by the respective analyses and observations at 00 and 12 UTC for 743 forecast runs in the period 2 November 2015 to 22 November 2016. Boxes indicate that symbols refer to the second score indicated at the top of the column.

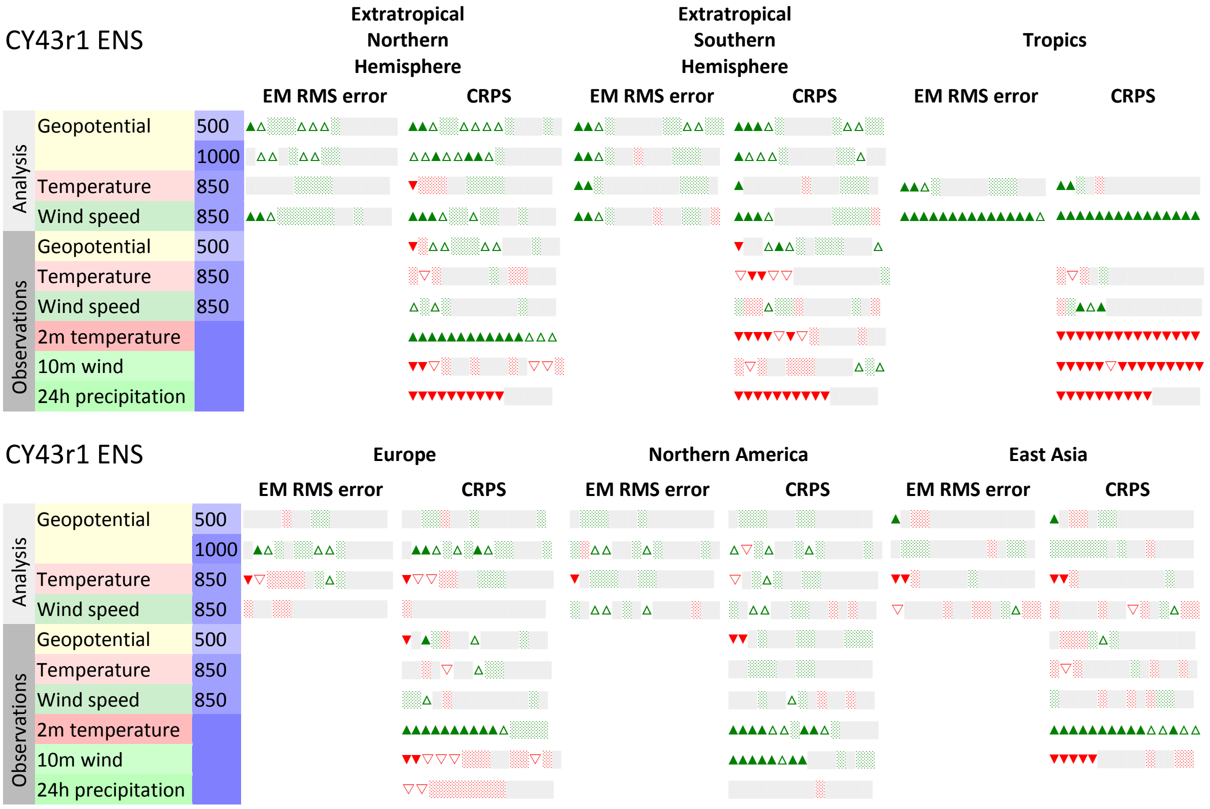
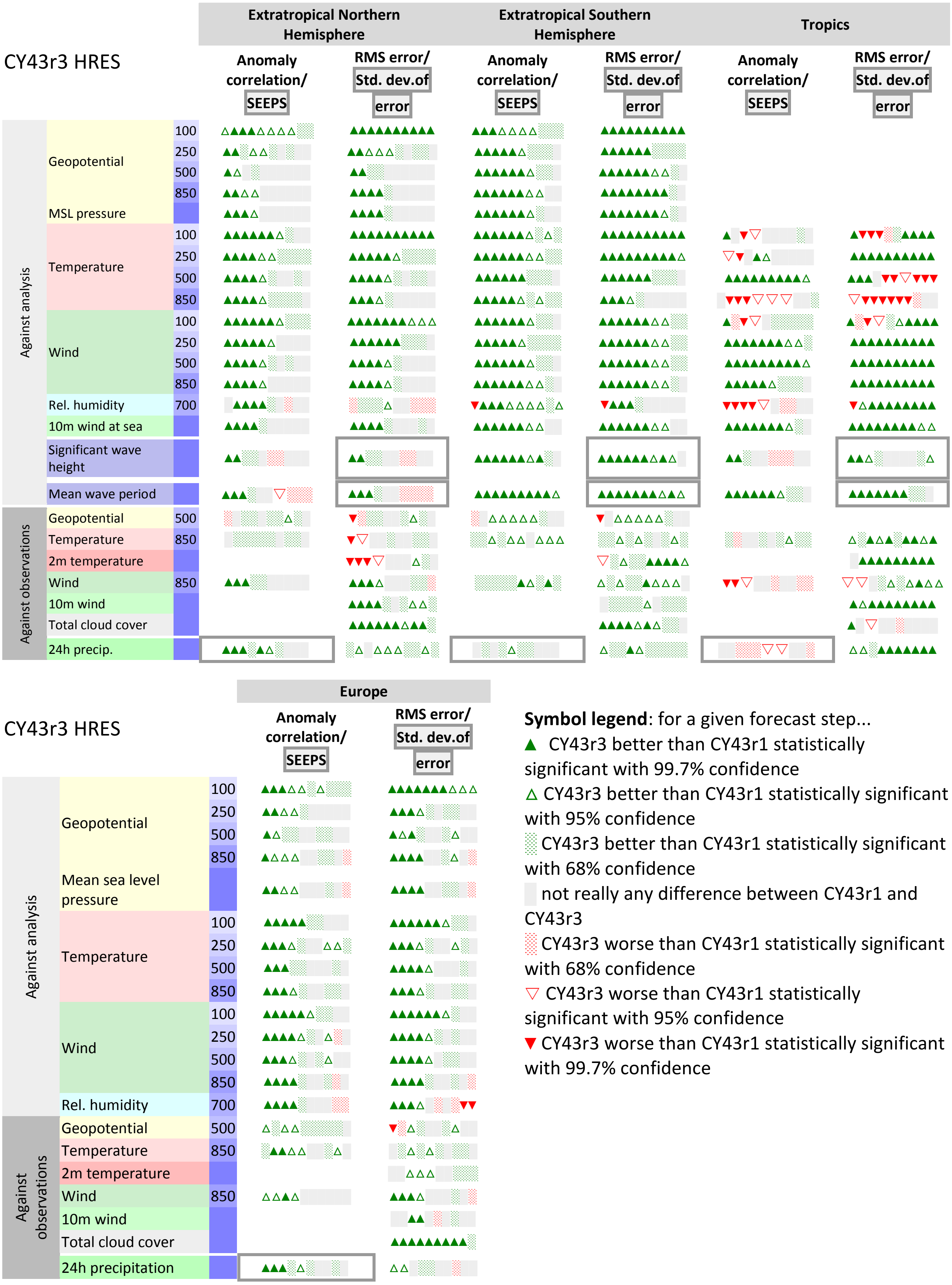


Figure 2: Summary ENS score card for Cy43r1. Score card for ENS cycle 43r1 versus cycle 41r2 verified by the respective analyses and observations at 00 Z for 237 forecast runs in the period 1 June 2016 to 22 November 2016.



Symbol legend: for a given forecast step...

▲ CY43r3 better than CY43r1 statistically significant with 99.7% confidence

△ CY43r3 better than CY43r1 statistically significant with 95% confidence

🟩 CY43r3 better than CY43r1 statistically significant with 68% confidence

⬜ not really any difference between CY43r1 and CY43r3

🟥 CY43r3 worse than CY43r1 statistically significant with 68% confidence

▽ CY43r3 worse than CY43r1 statistically significant with 95% confidence

▼ CY43r3 worse than CY43r1 statistically significant with 99.7% confidence

Figure 3: Summary score card for HRES Cy43r3. Score card for HRES cycle 43r3 versus cycle 43r1 verified by the respective analyses and observations at 00 and 12 UTC for 809 forecast runs in the period 1 June 2016 to 11 July 2017. Boxes indicate that symbols refer to the second score indicated at the top of the column.

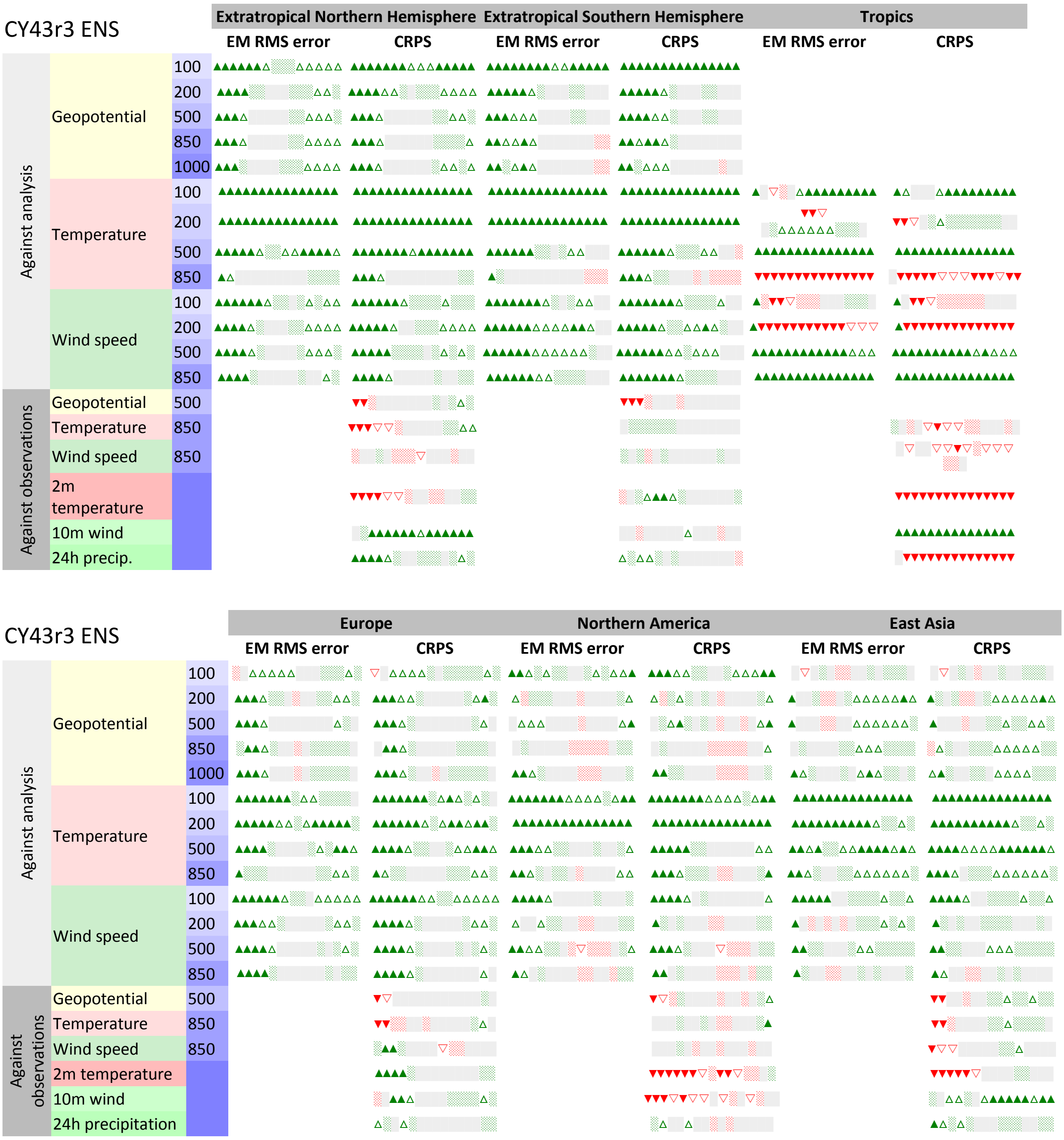


Figure 4: Summary score card for ENS Cy43r3. Score card for ENS cycle 43r3 versus cycle 43r1 verified by the respective analyses and observations at 00 UTC for 234 forecast runs in the period 1 June 2016 to 11 July 2017.

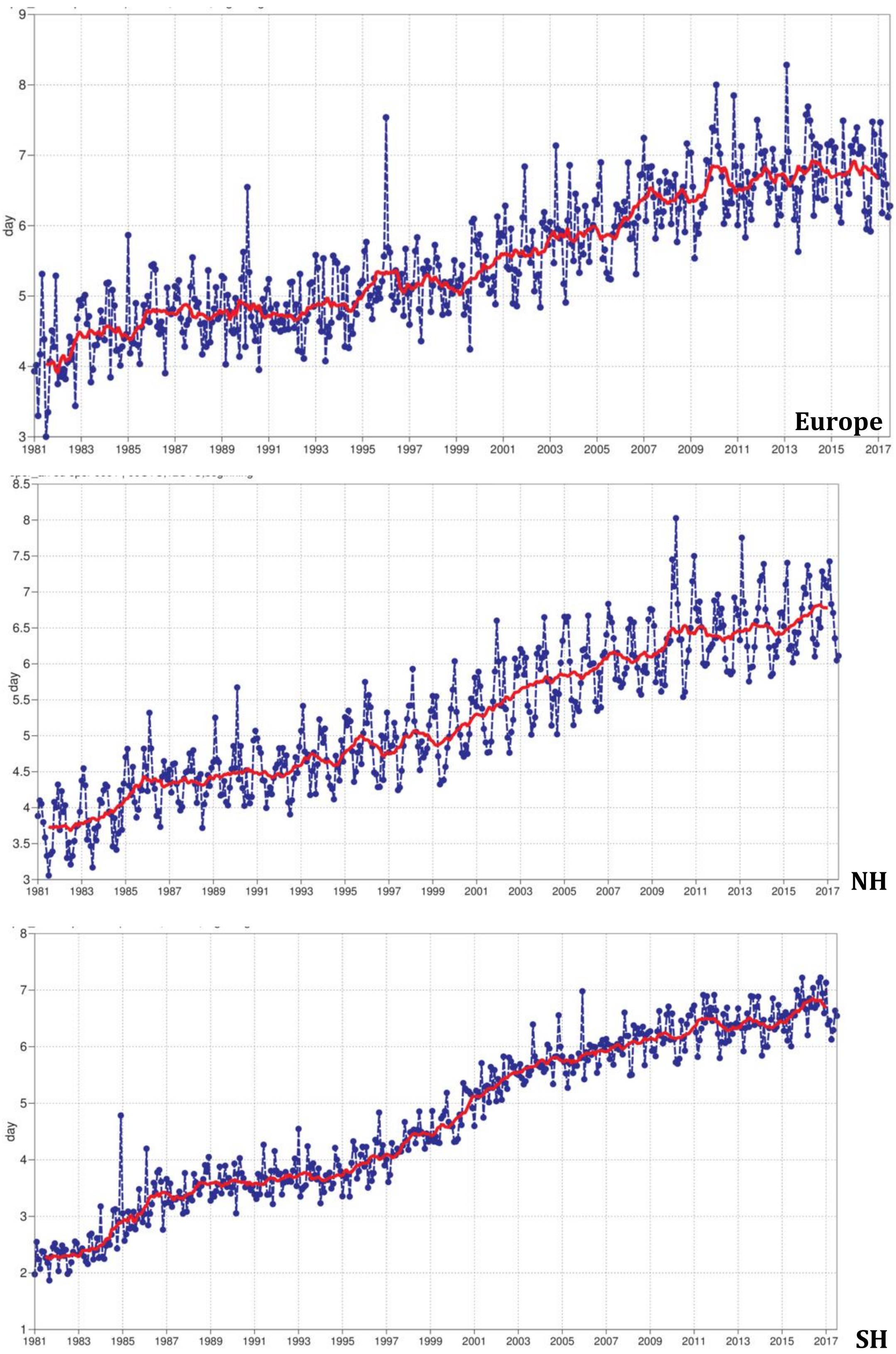
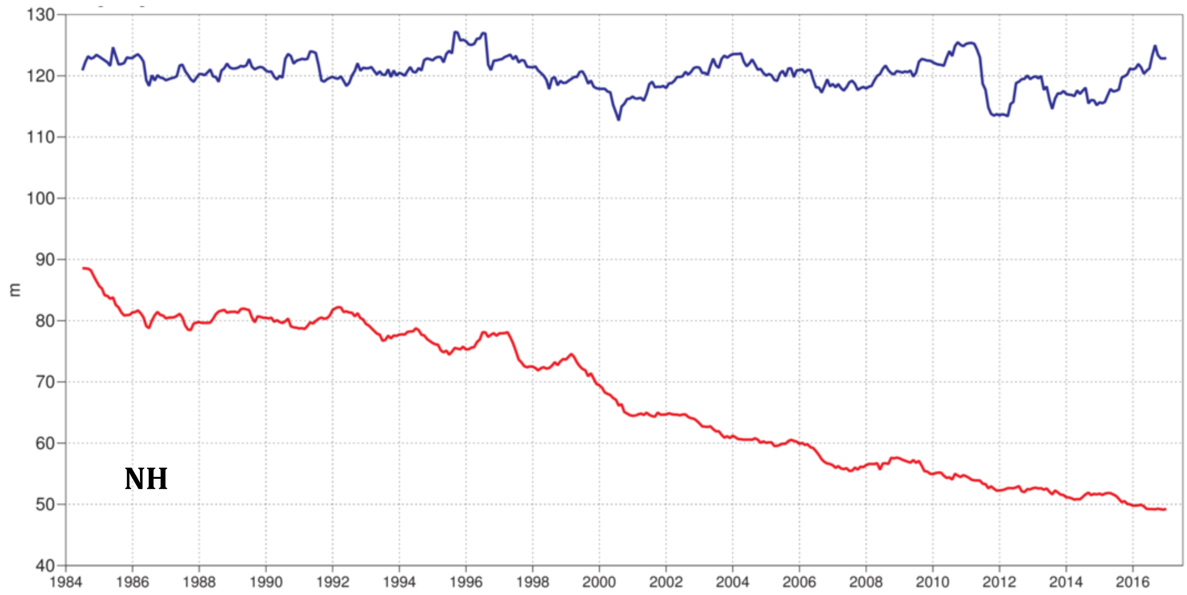


Figure 5: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).



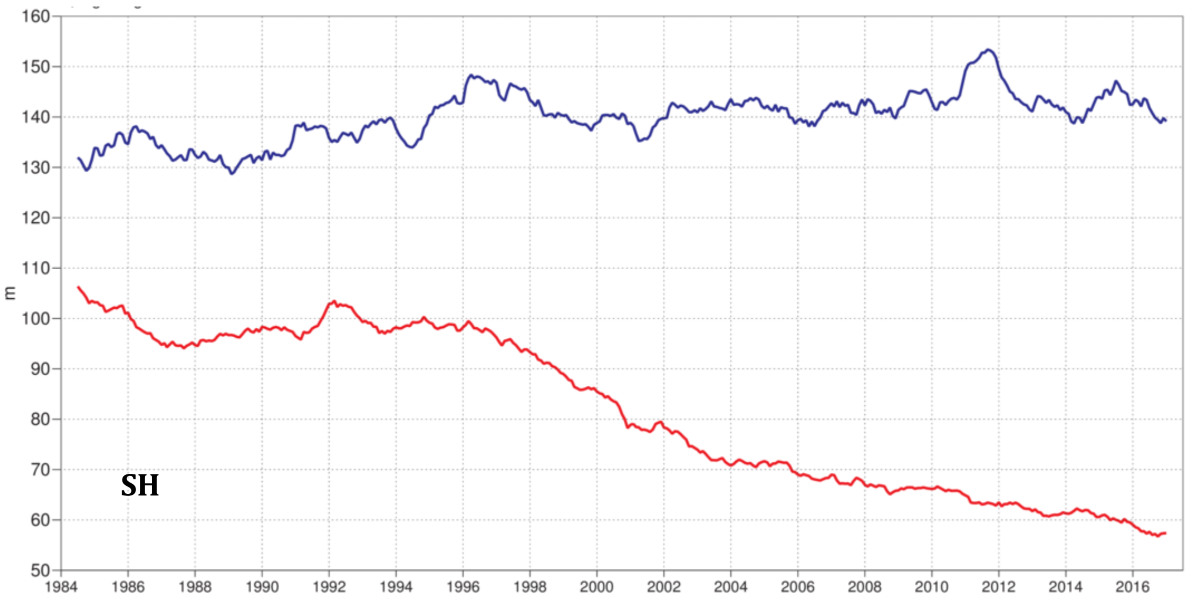


Figure 6: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2016–July 2017. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

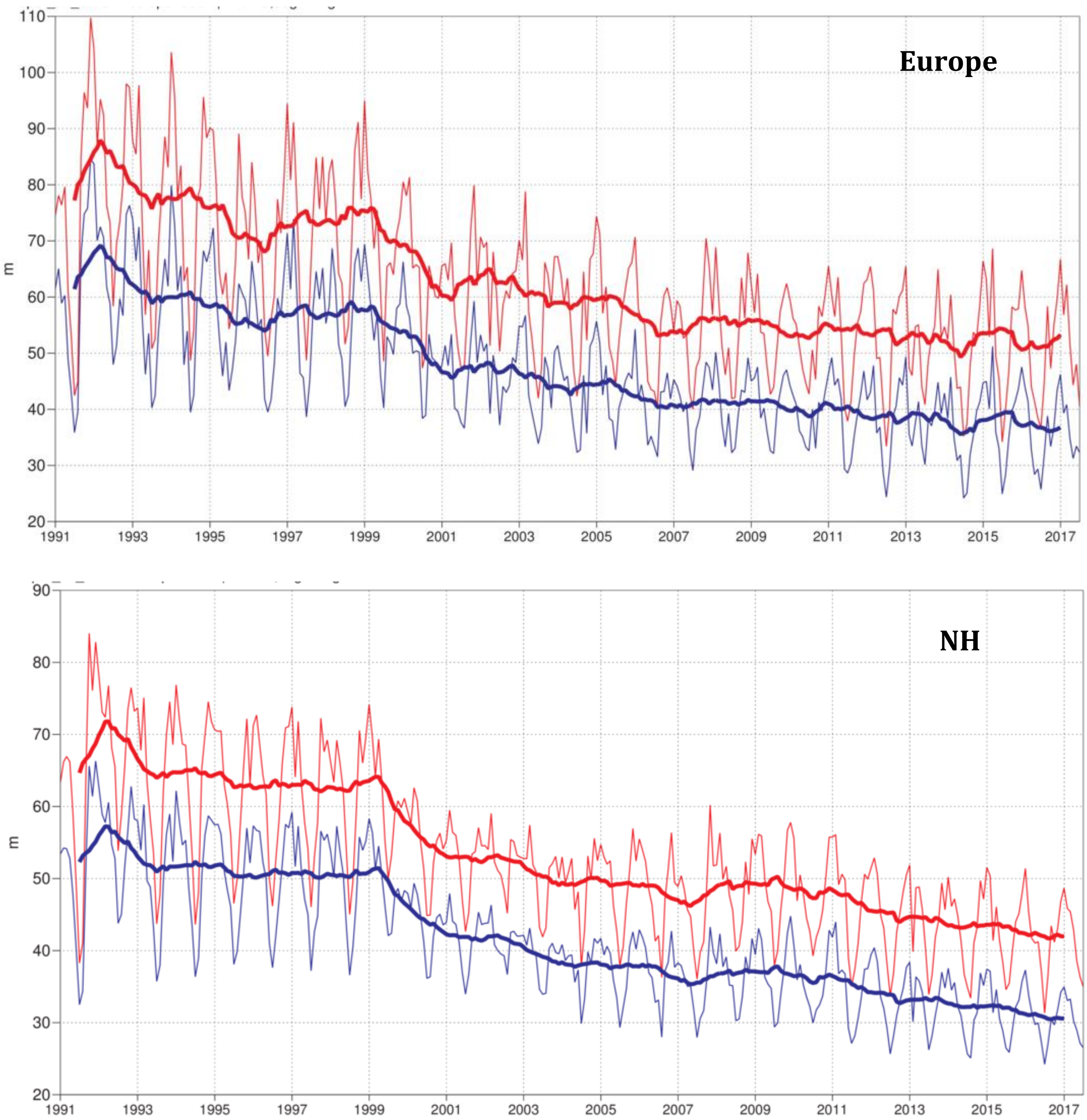


Figure 7**:** Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average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

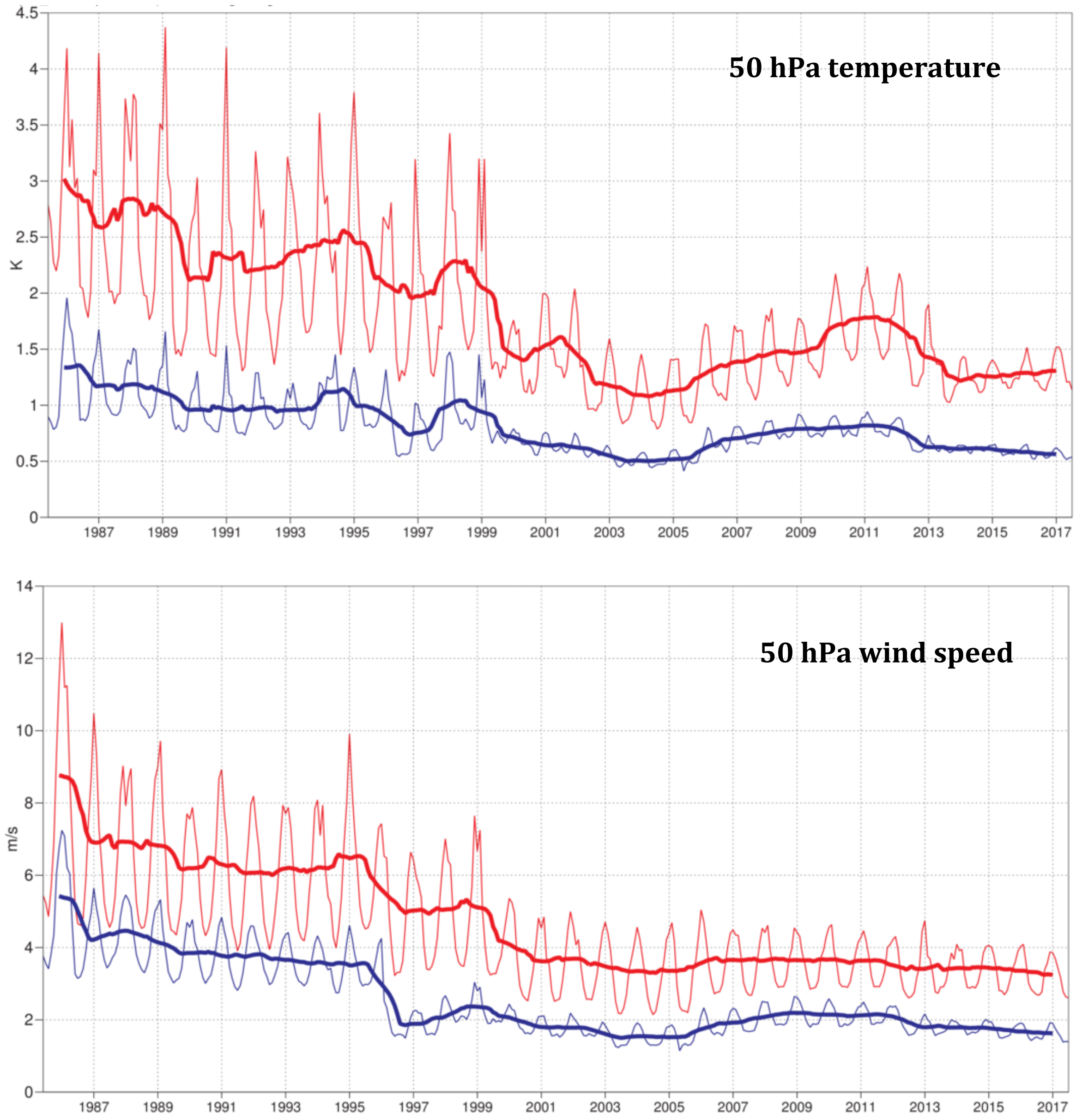
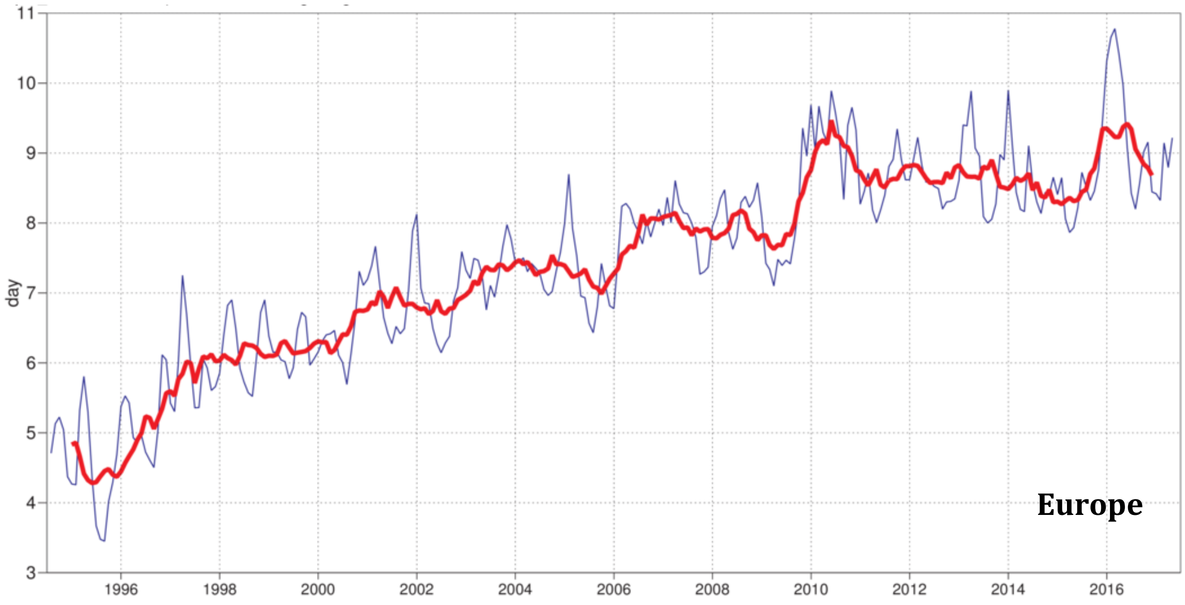


Figure 8**:** Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are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



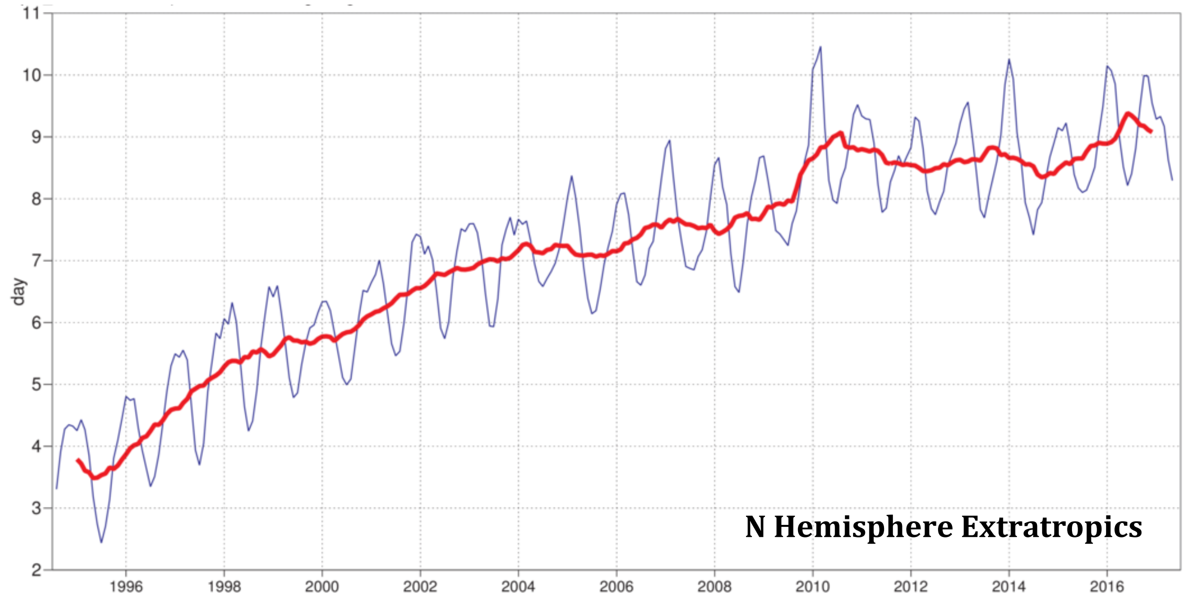


Figure 9**:** Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

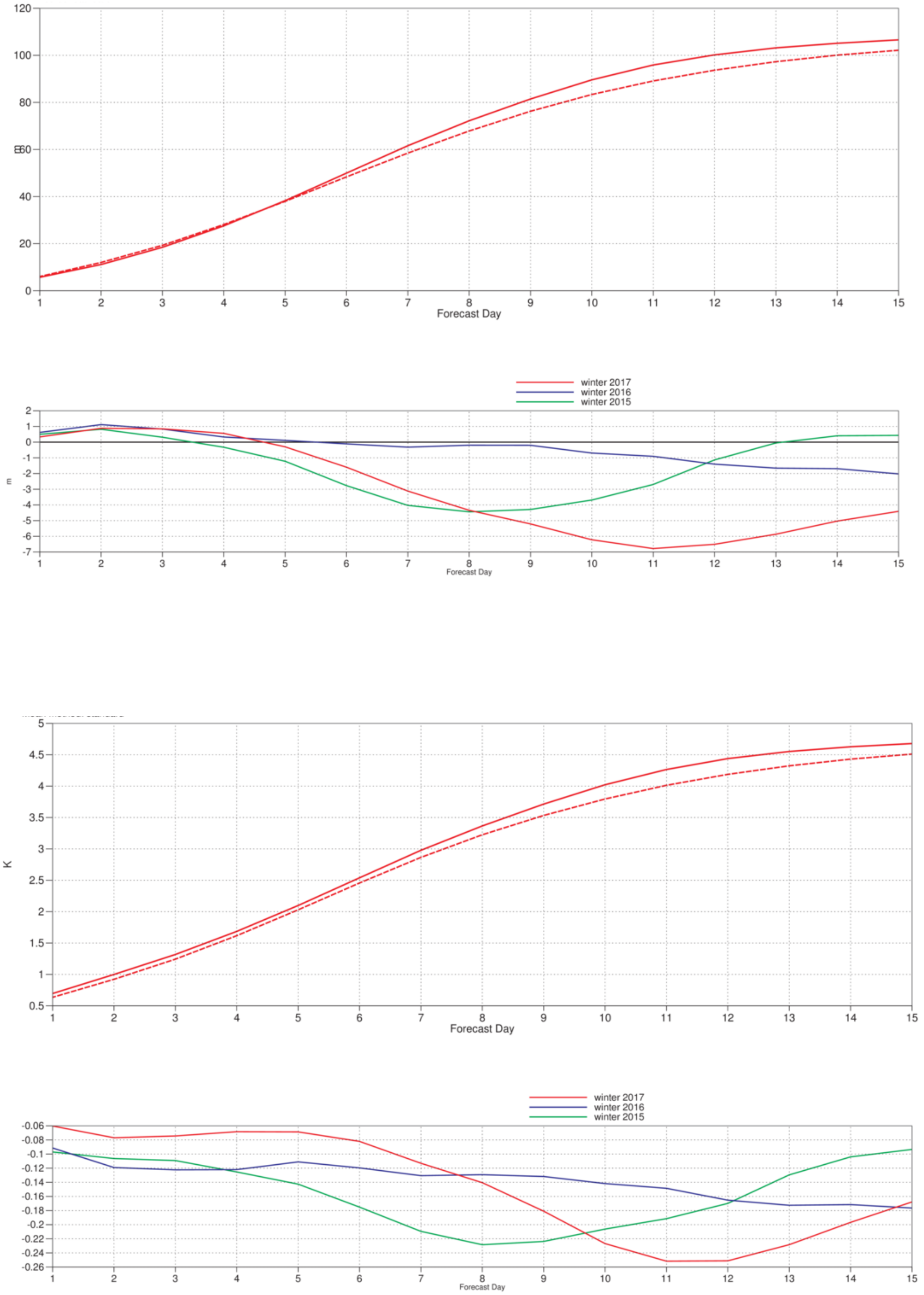


Figure 10: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2016–2017 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.

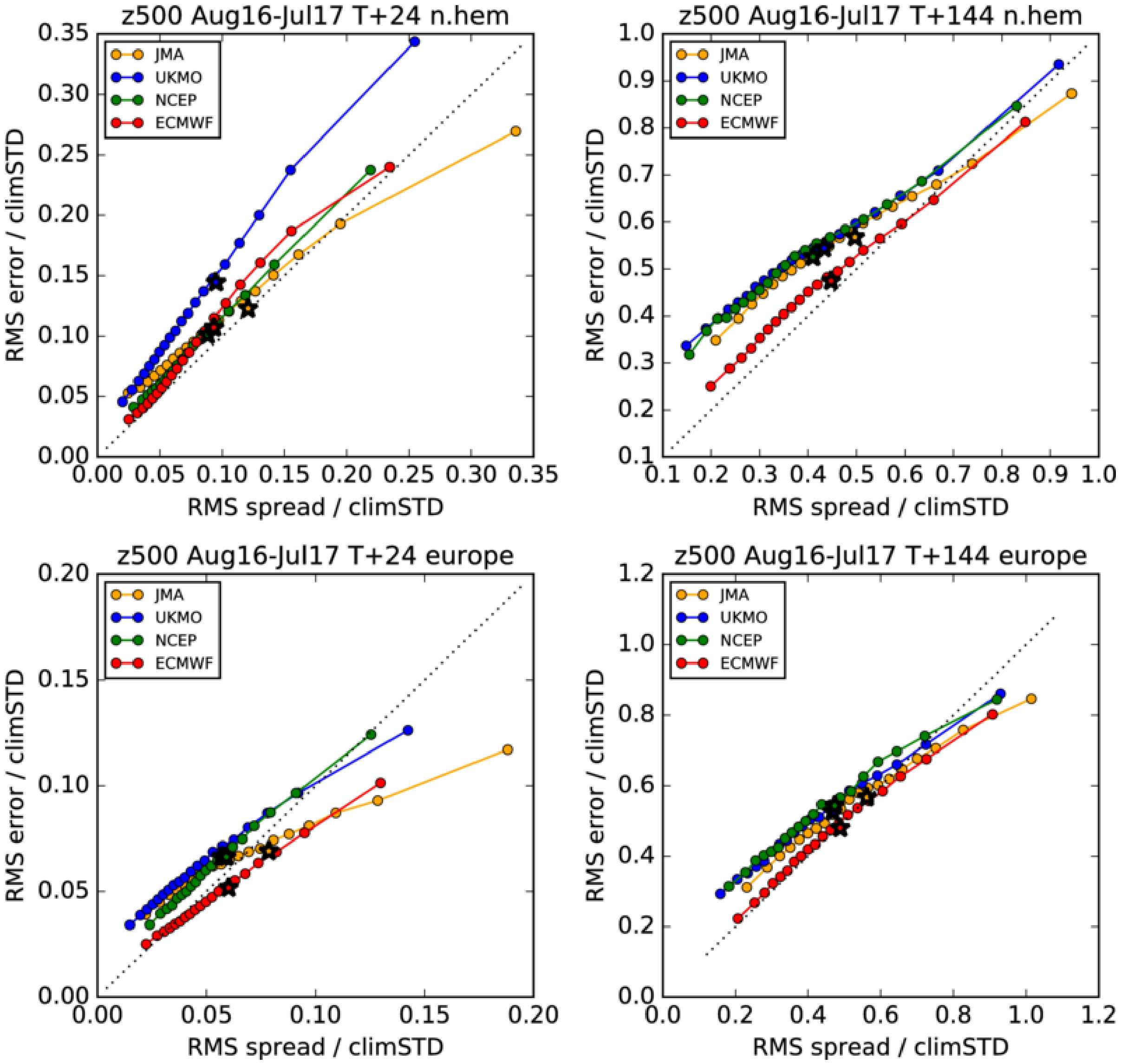


Figure 11: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2016 – July 2017 in the northern hemisphere extra-tropics (top) and in Europe (bottom) for day 1 (left) and day 6 (right) , verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship. Both spread and error have been normalised by the climate standard deviation before clustering.

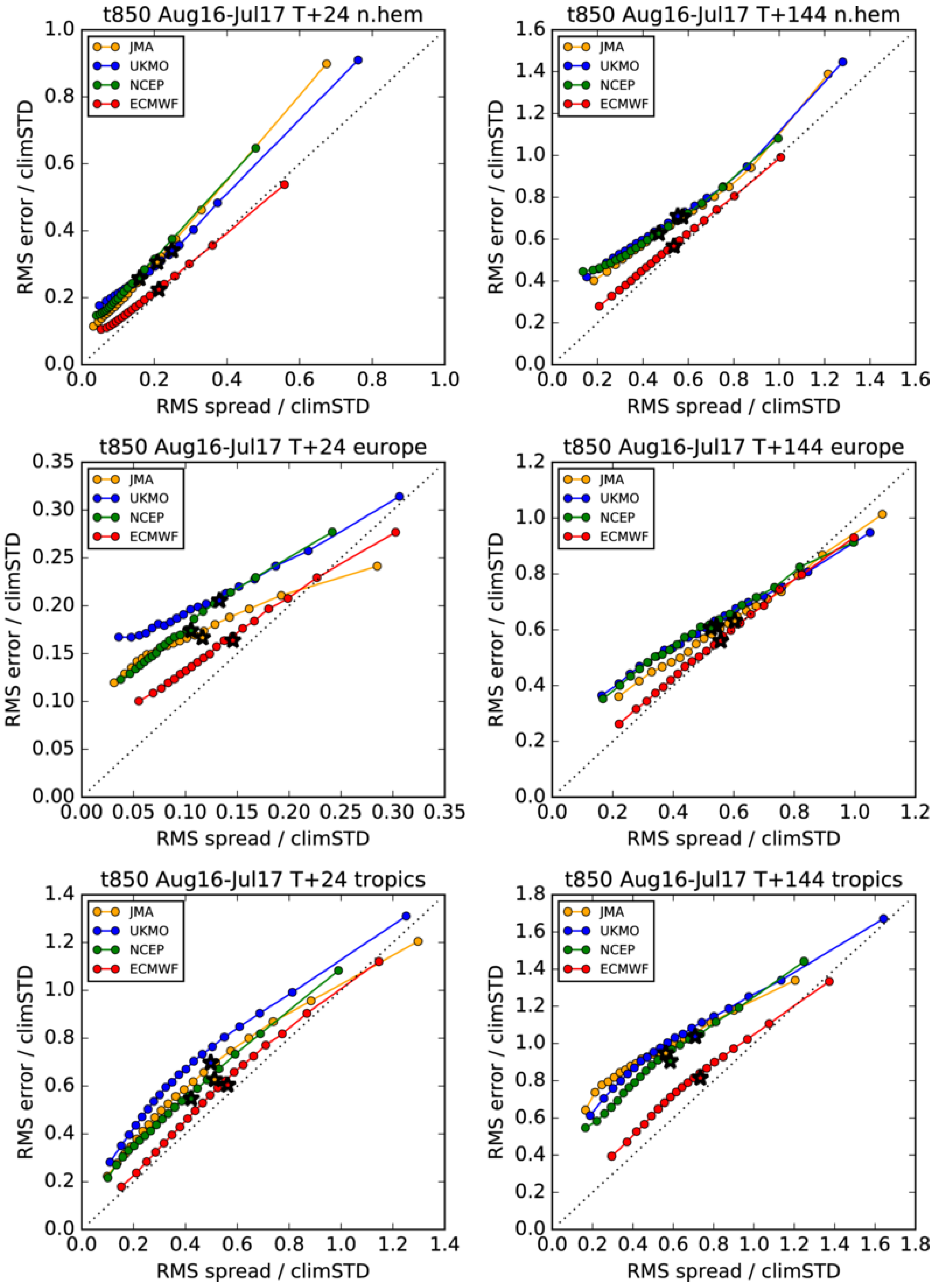
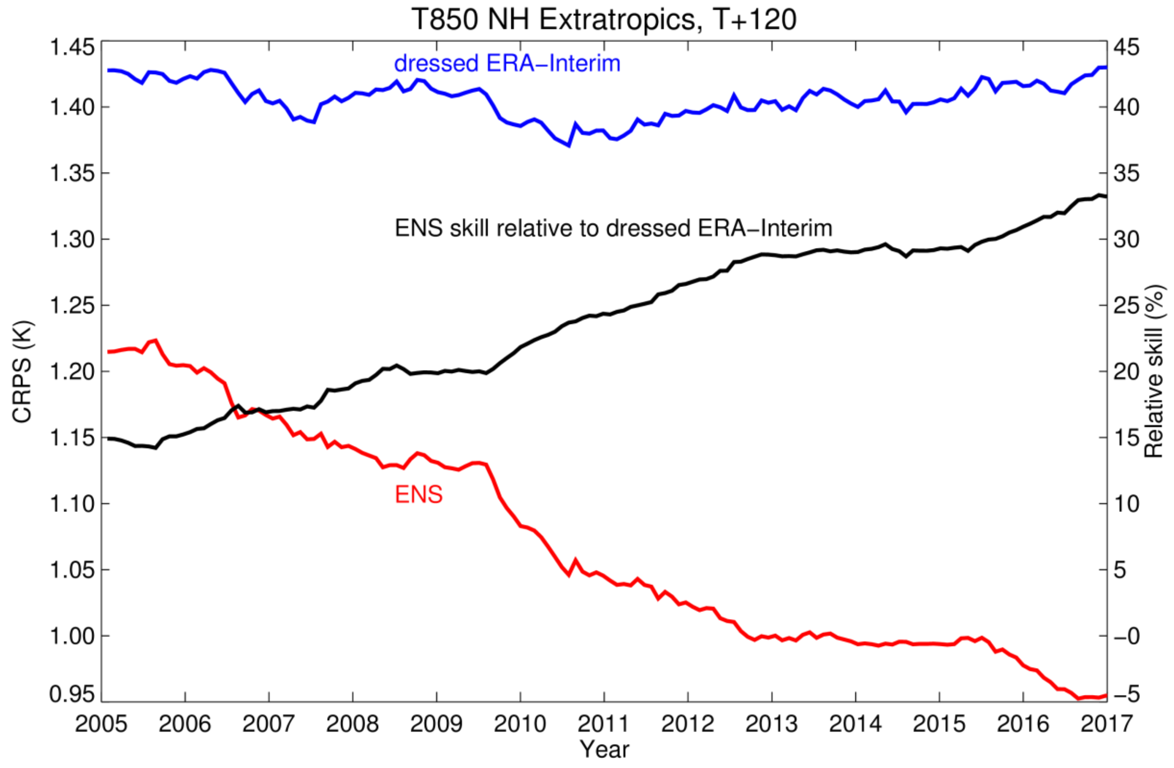


Figure 12**:** Ensemble spread reliability of different global models for 850 hPa temperature for the period August 2016 – July 2017 in the northern hemisphere extra-tropics (top), Europe (centre), and the tropics (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship. Both spread and error have been normalised by the climate standard deviation before clustering.



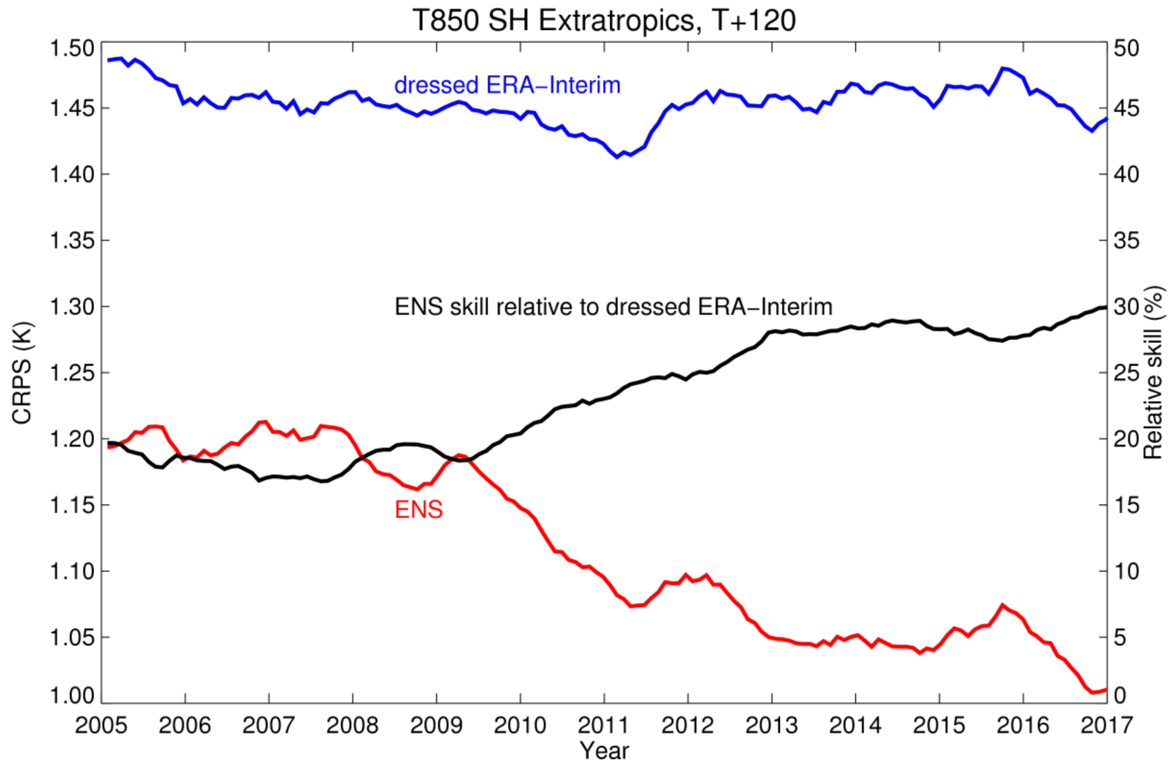
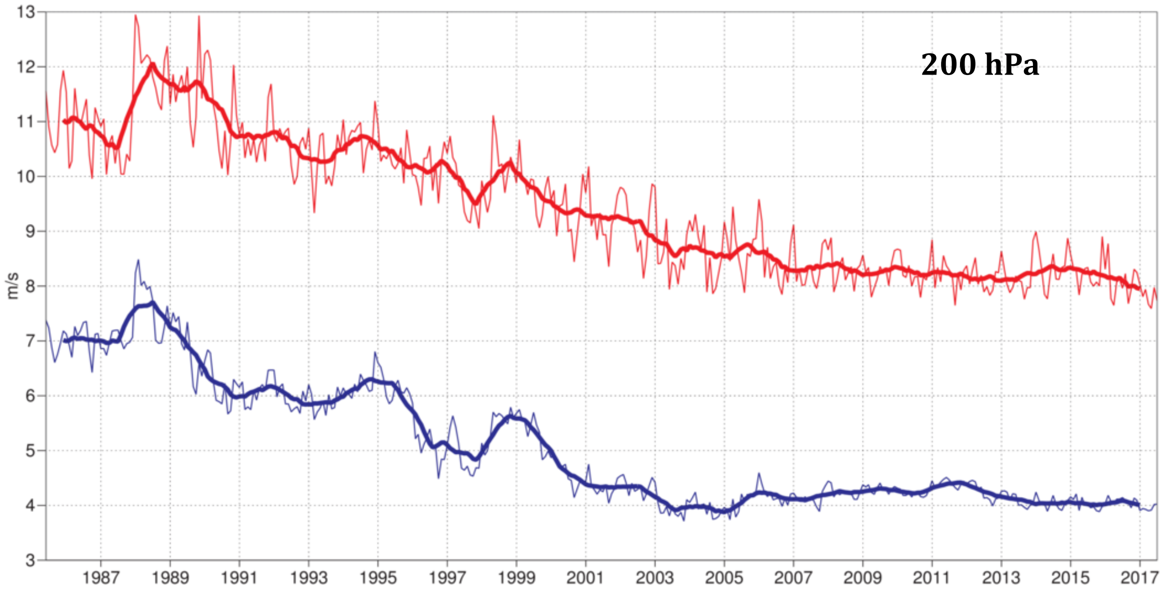


Figure 13**:** CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5, verified against analysis. Scores are shown for the ensemble forecast (red) and the dressed ERA-Interim forecast (blue). Black curves show the skill of the ENS relative to the dressed ERA-Interim 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**.**



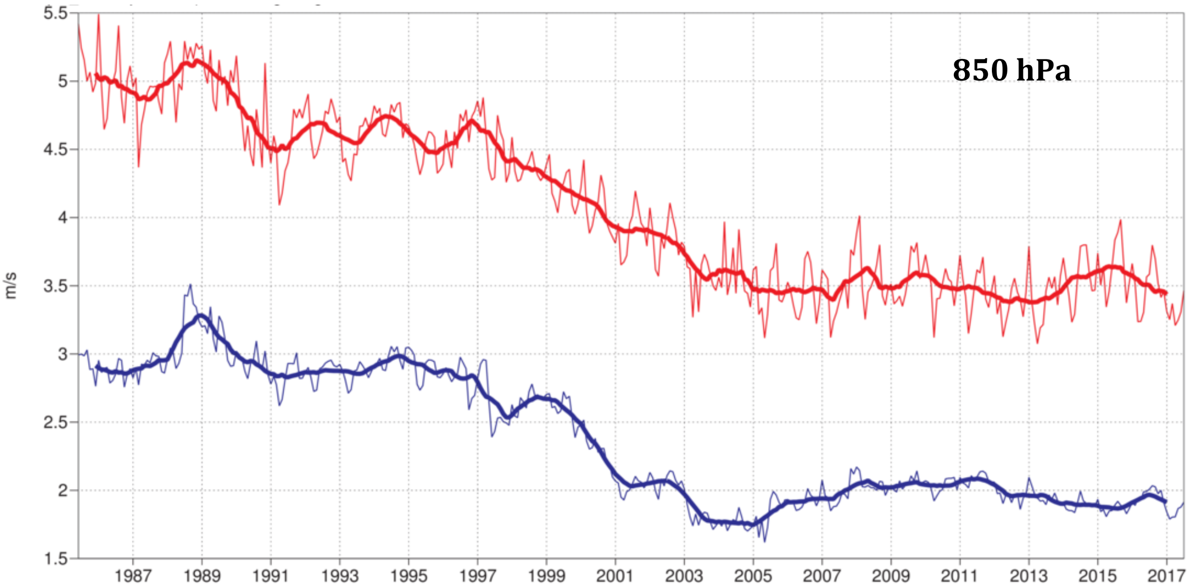
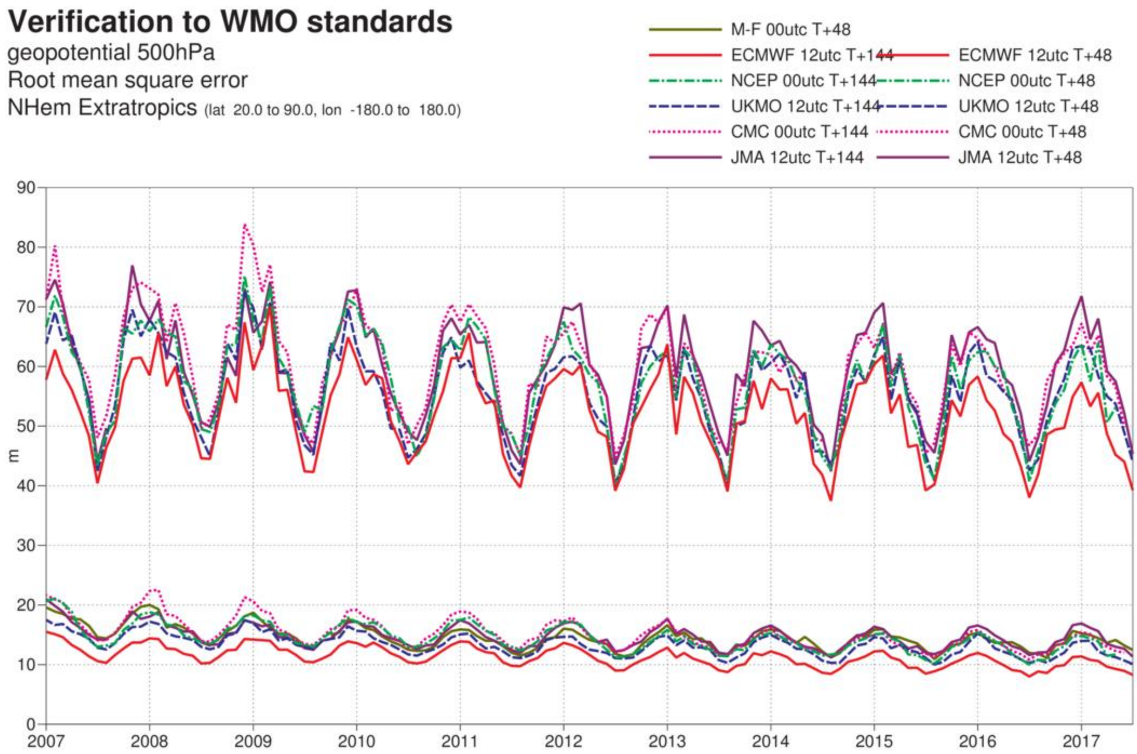


Figure 14: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).



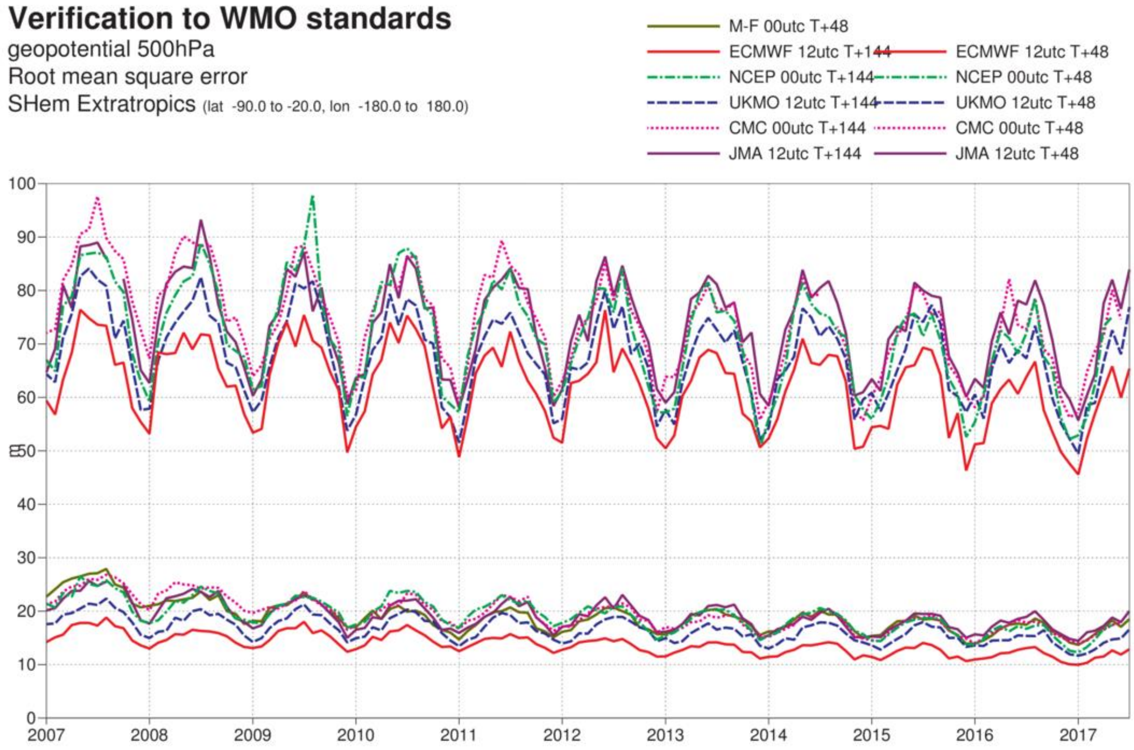
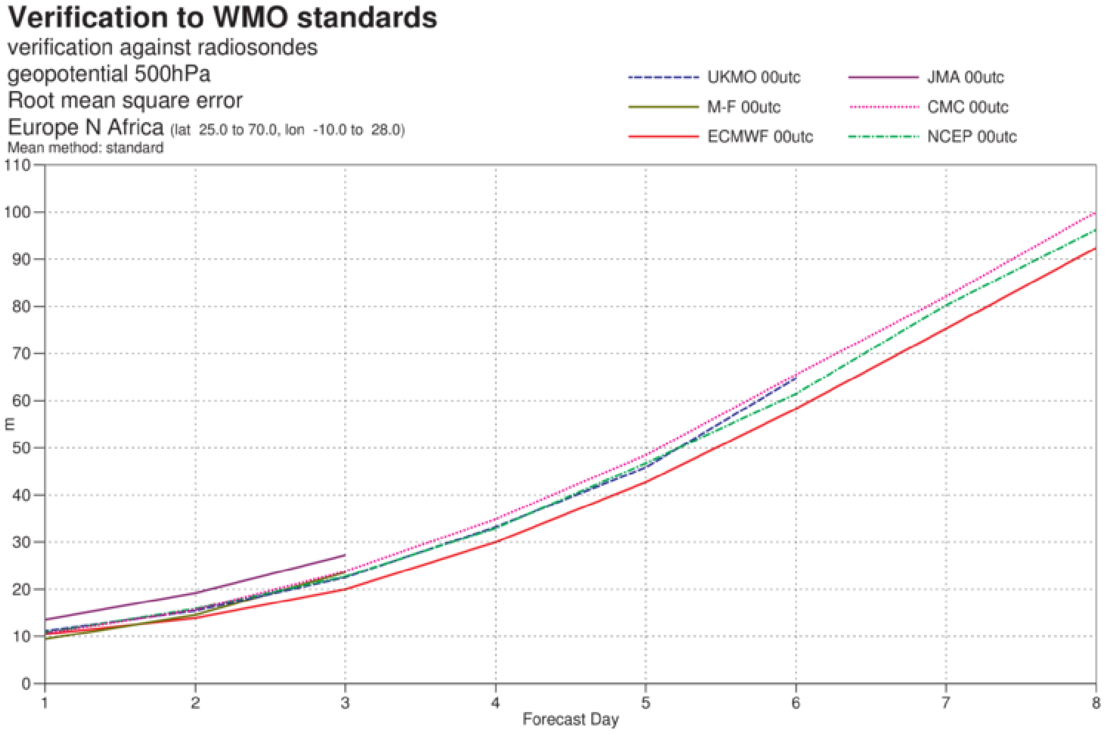


Figure 15: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top) and southern (bottom) extratropics. In each panel the upper curves show the six-day forecast error and the lower curves show the two-day forecast error. Each model is verified against 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.



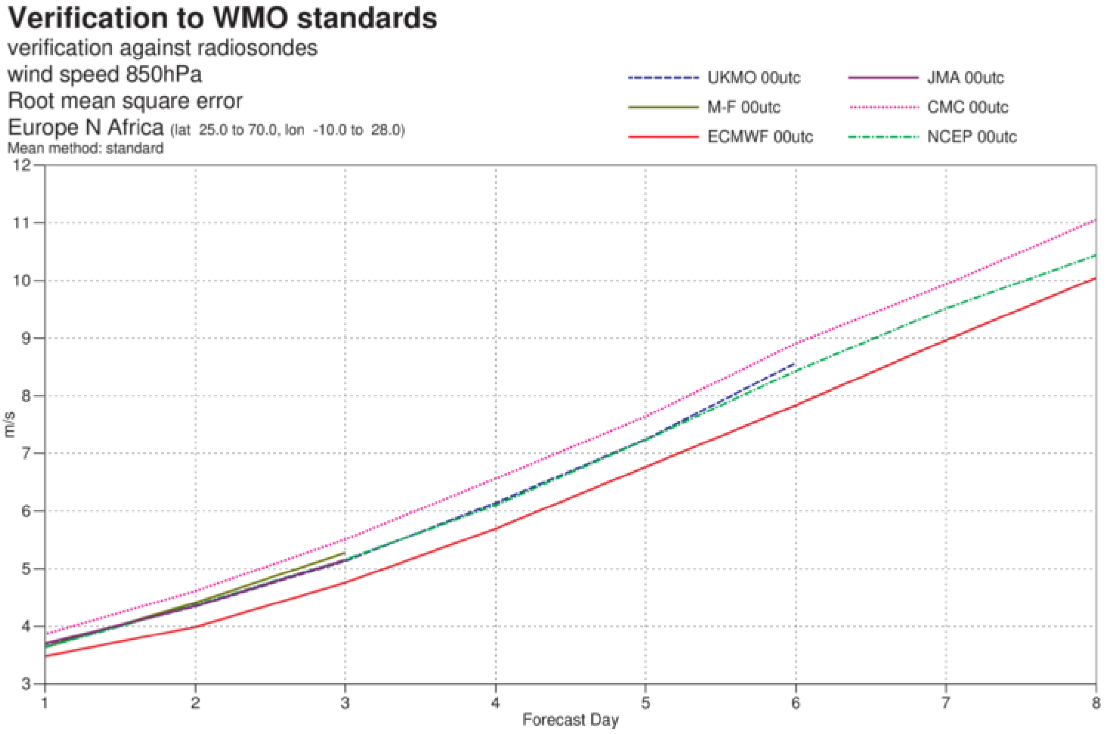
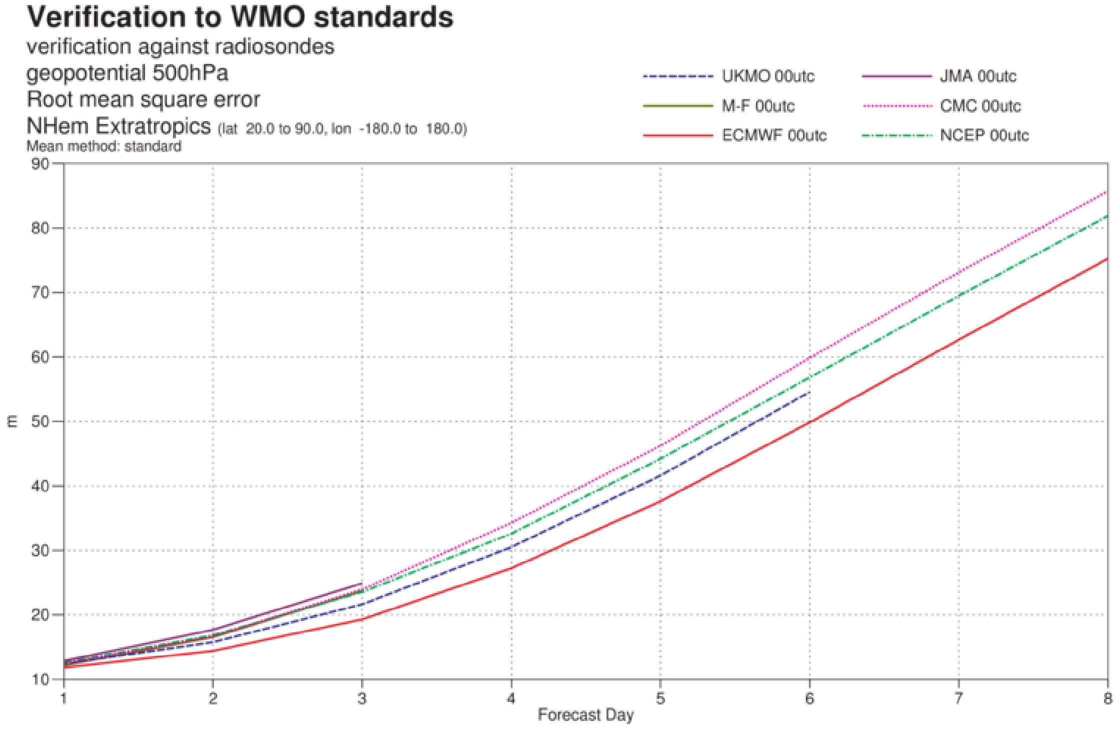


Figure 16: WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2016–July 2017).



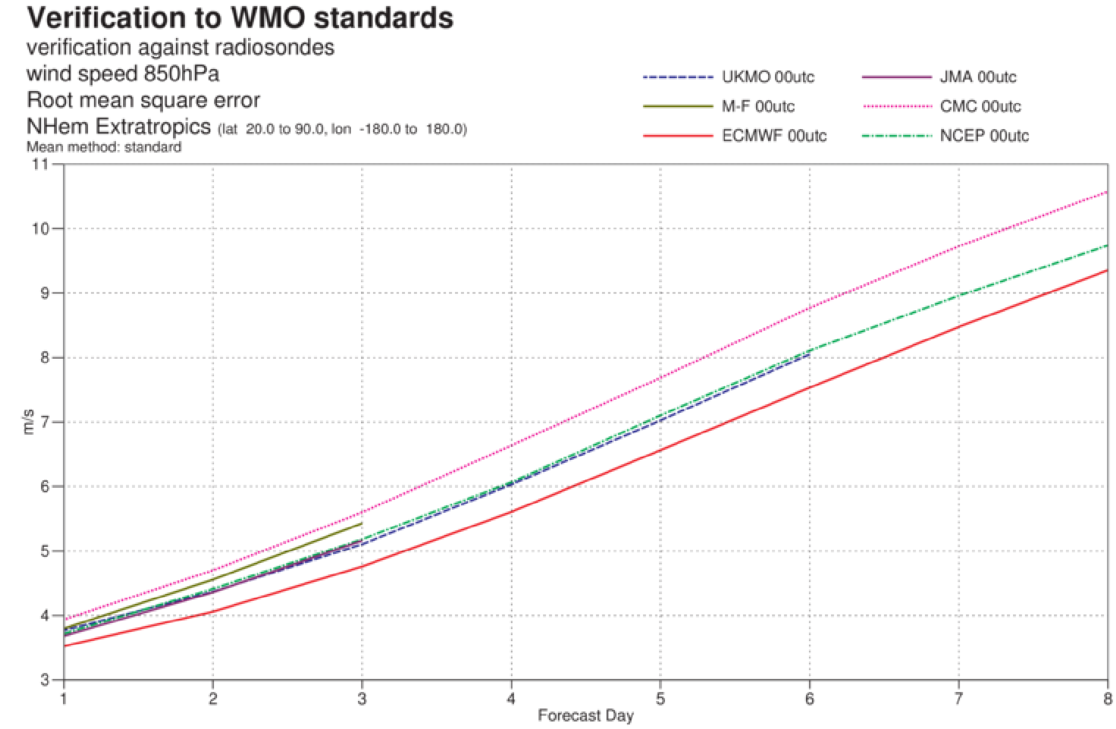
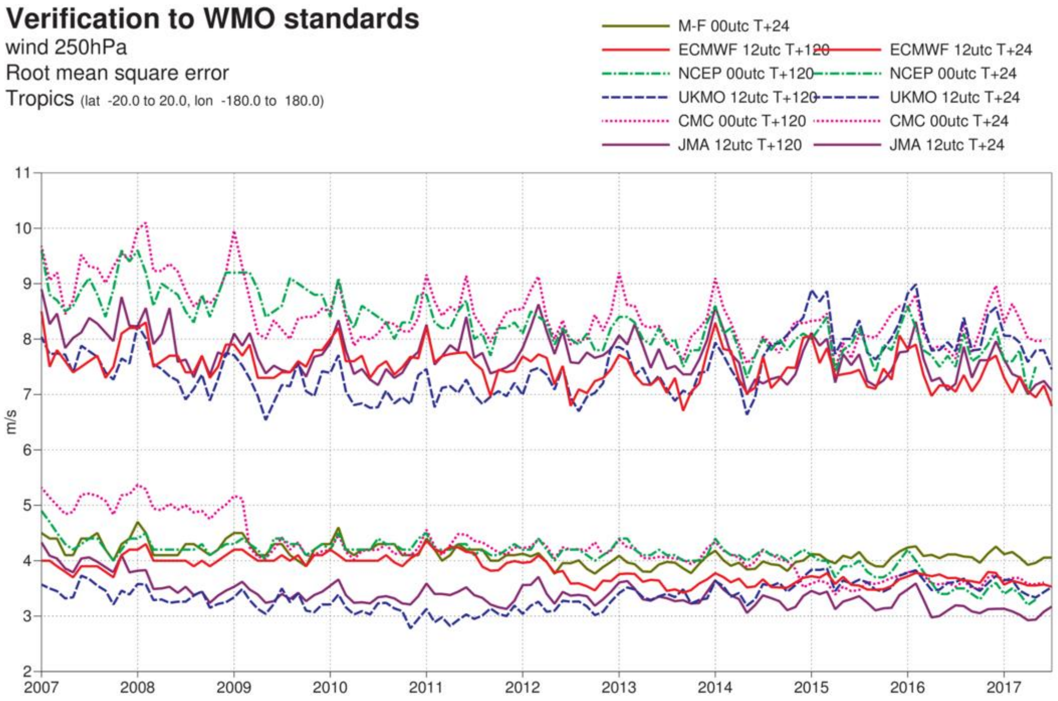


Figure 17**:** As Figure 16 for the northern hemisphere extratropics.



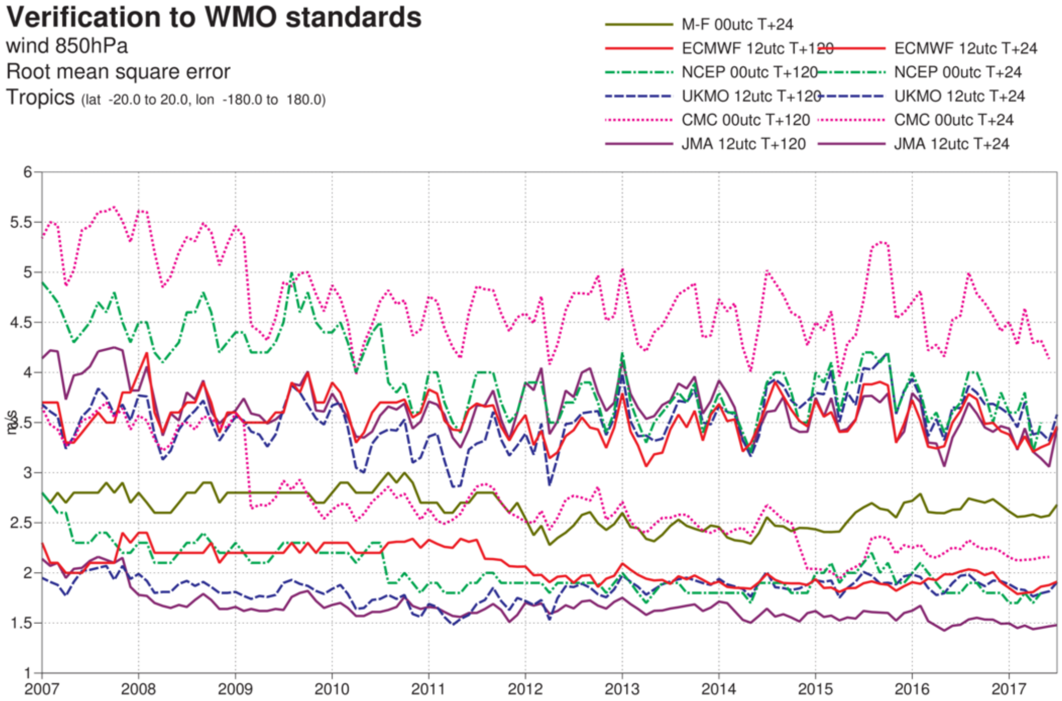
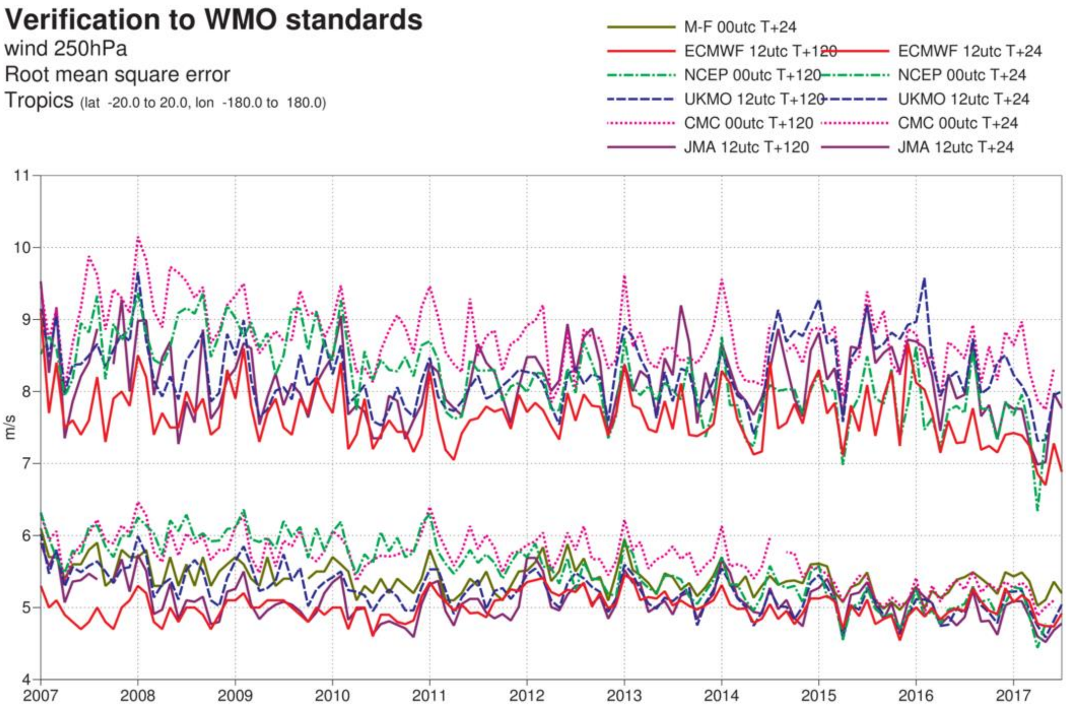


Figure 18**:** WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel the upper curves show the five-day forecast error and the lower curves show the one-day forecast error. Each model is verified against its own analysis.



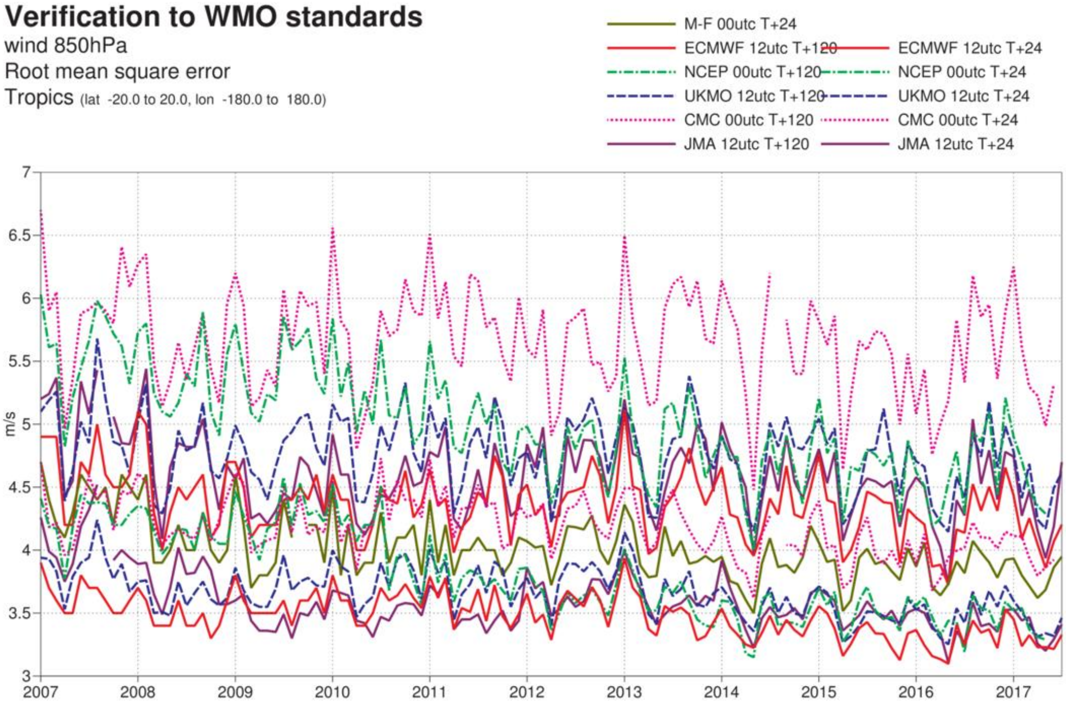


Figure 19**:** As Figure 18 for verification against radiosonde observations.

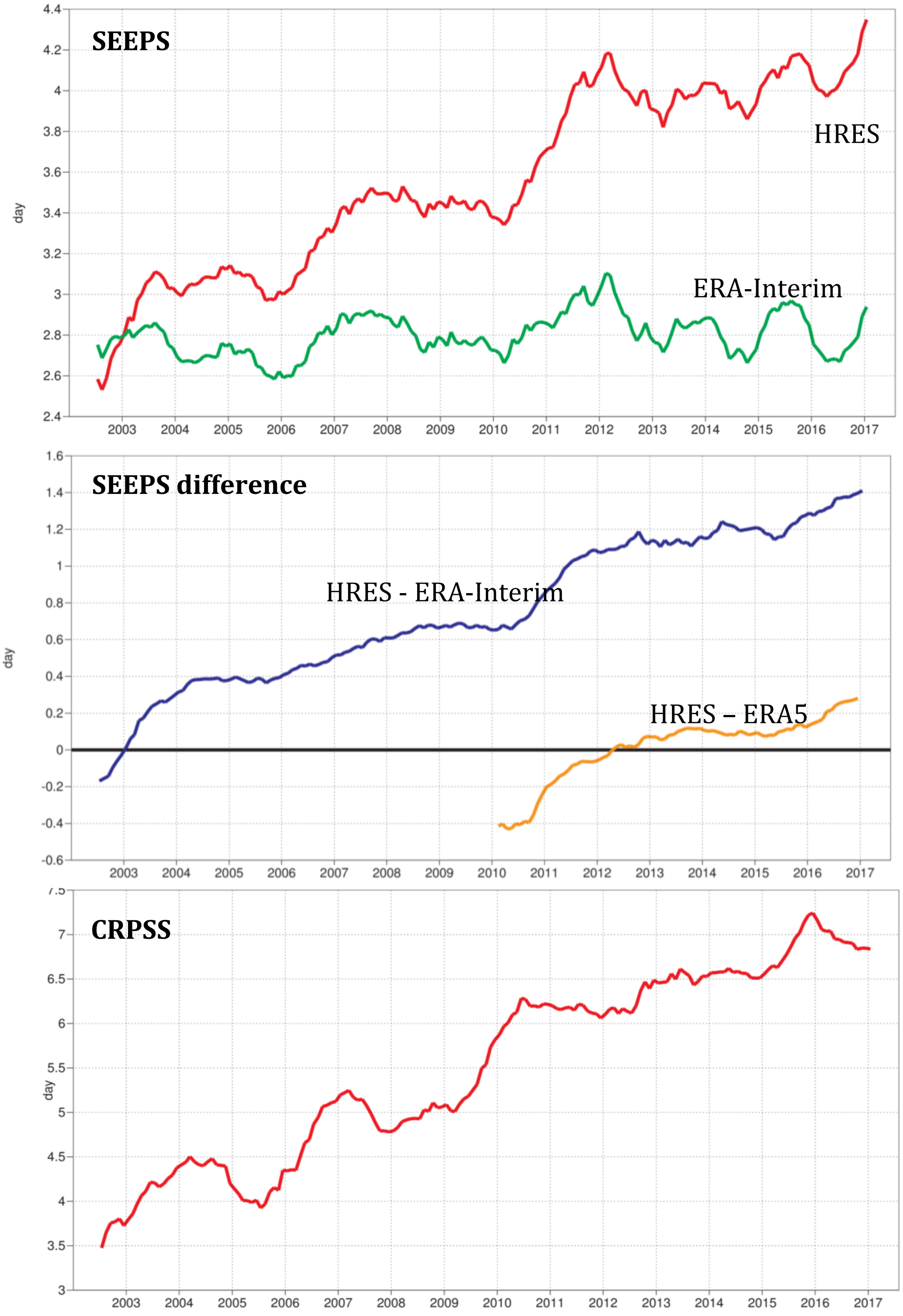


Figure 20**:** Supplementary headline scores for deterministic (top, centre) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics; each point is calculated over a 12-month period, plotted at the 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.

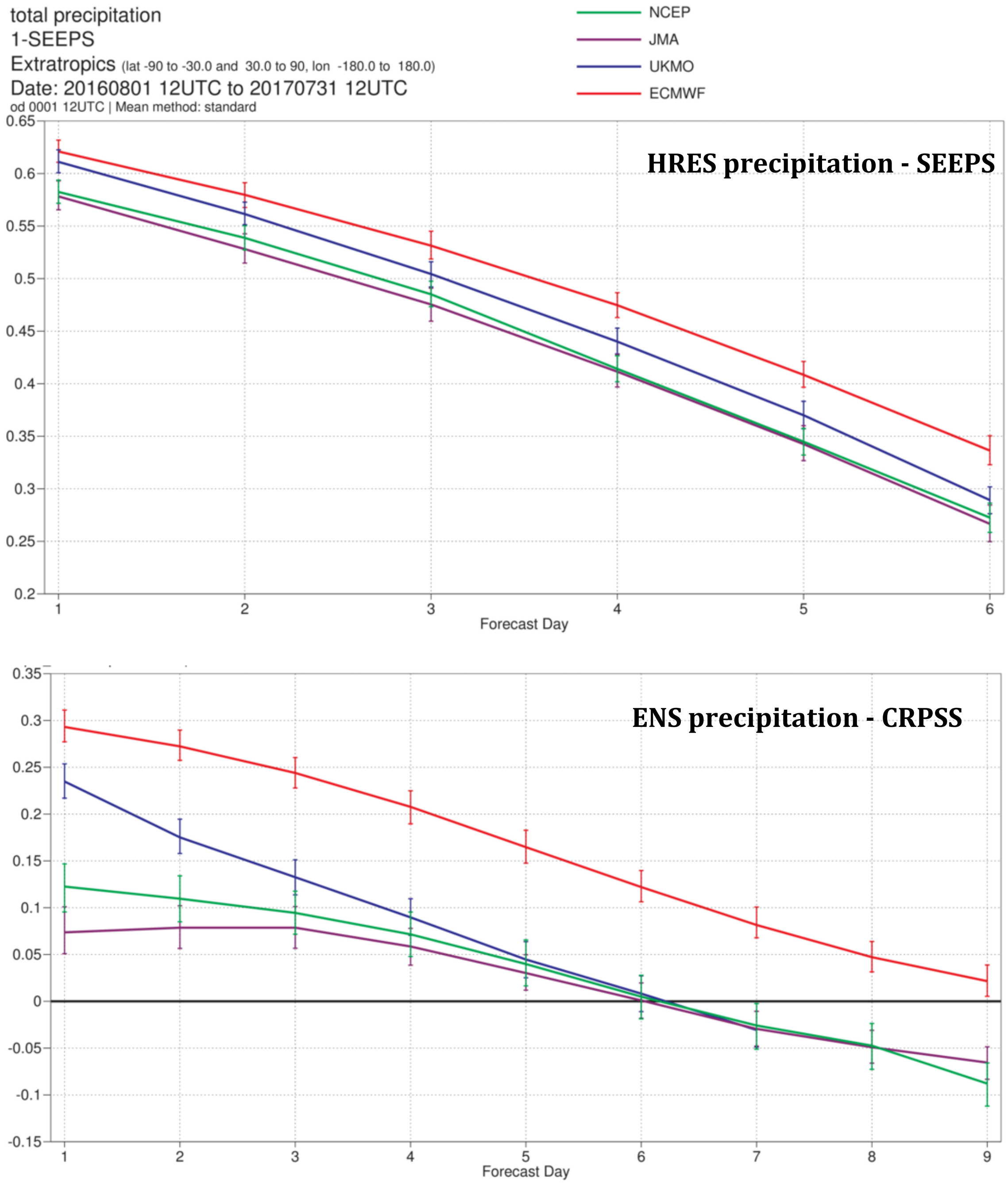


Figure 21: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure **20**. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2016–July 2017. Bars indicate 95% confidence intervals.

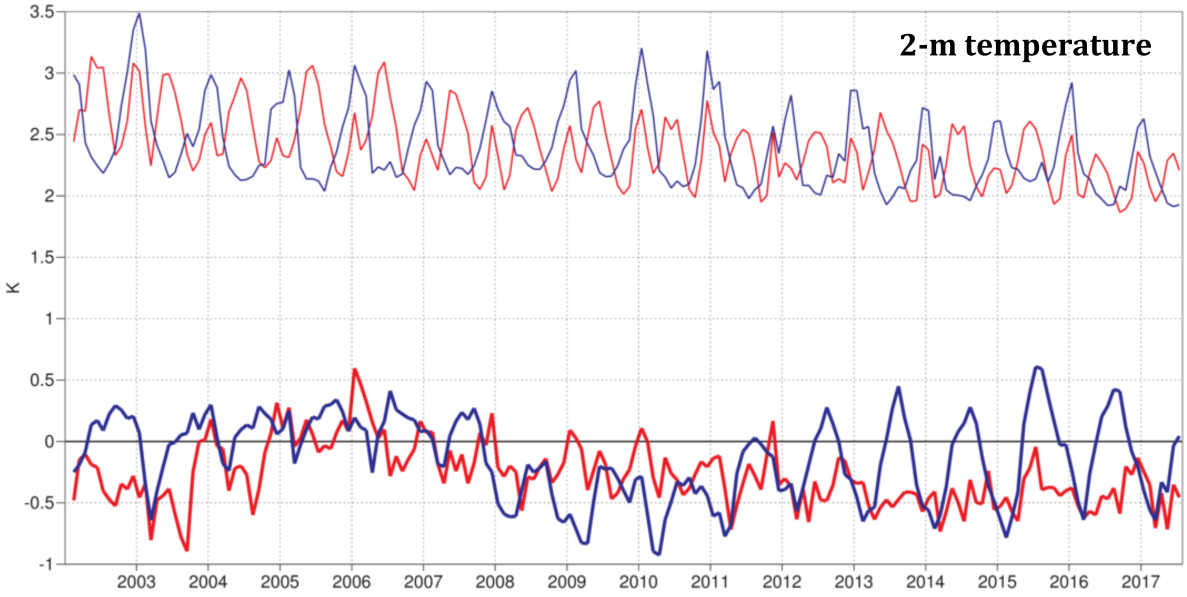


Figure 22: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

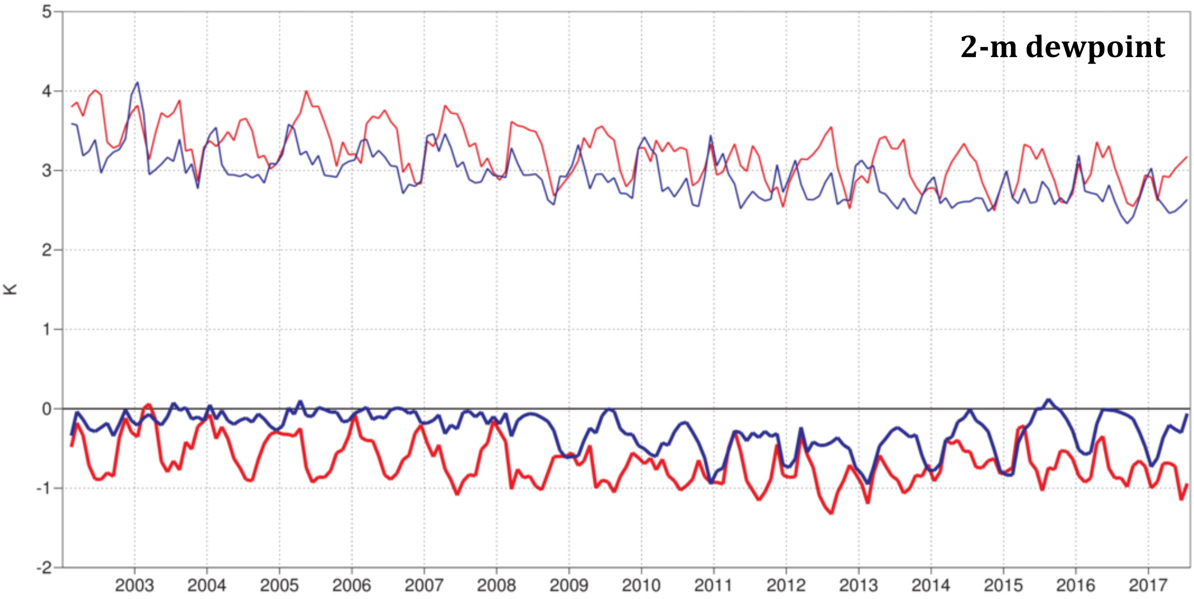


Figure 23: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

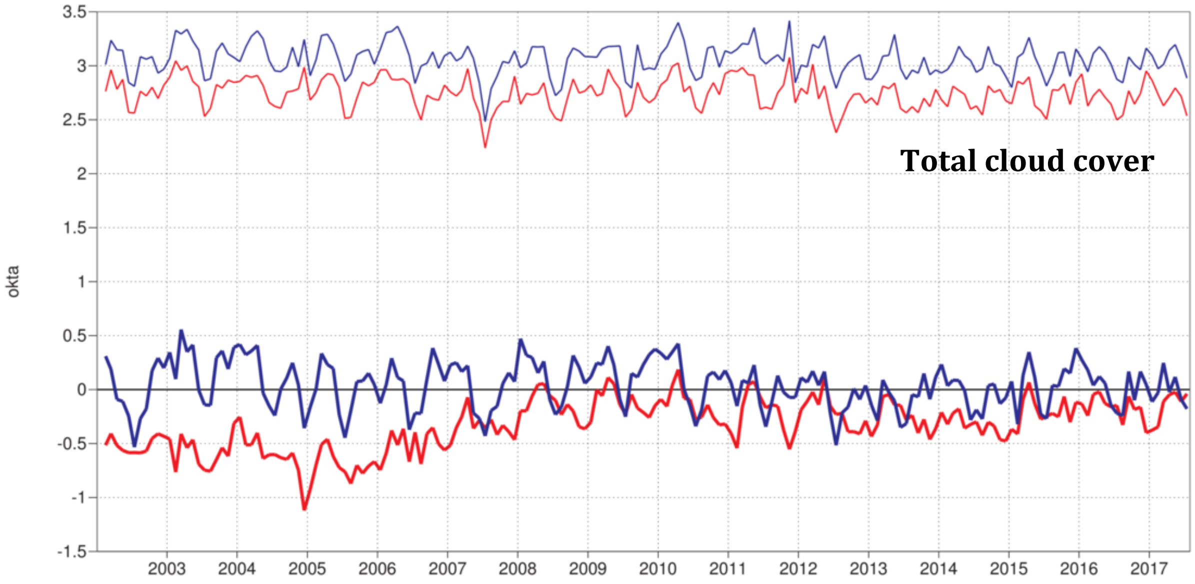


Figure 24: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

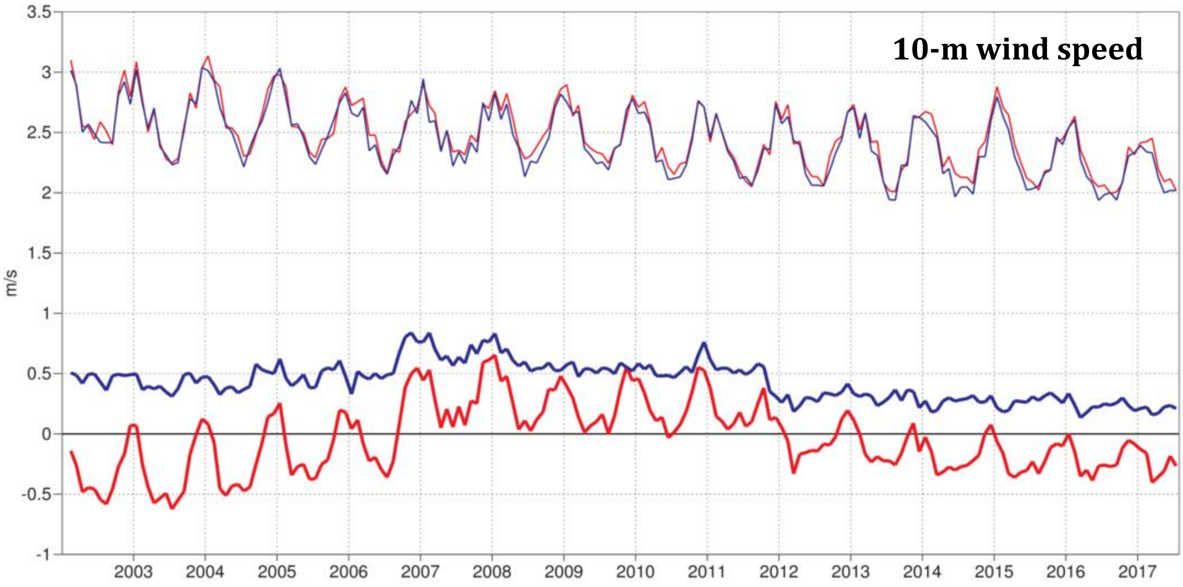


Figure 25: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

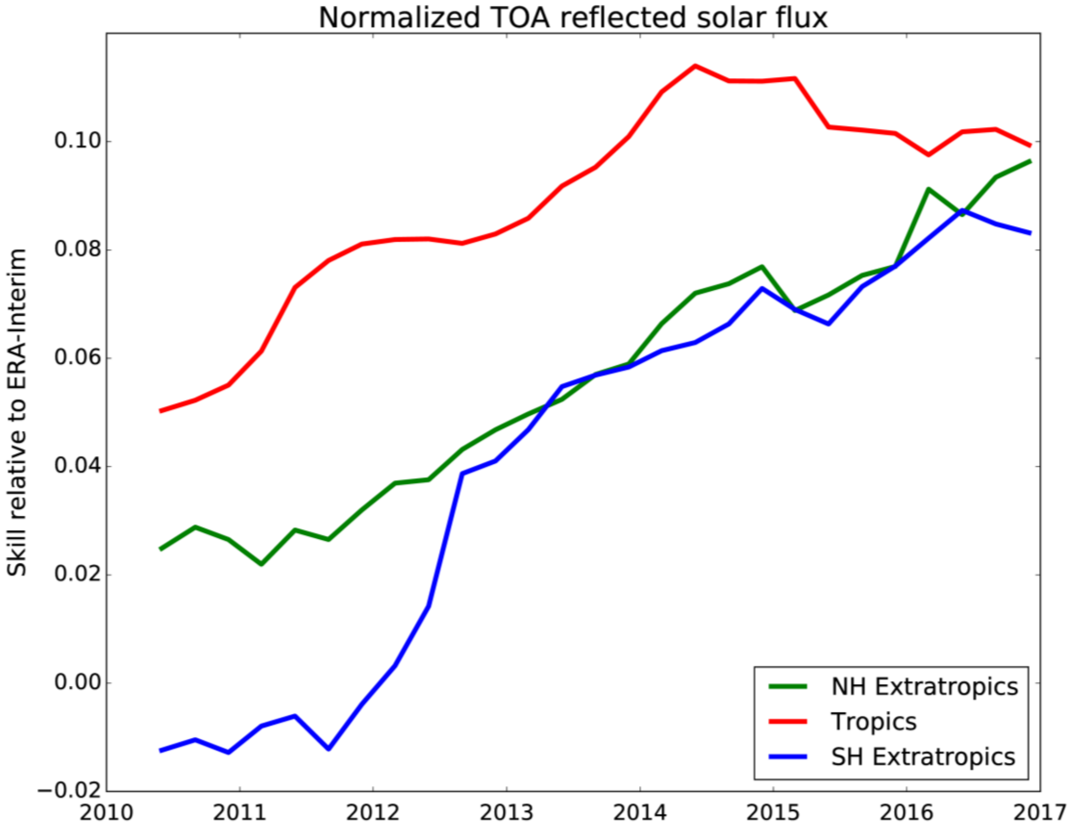


Figure 26**:** 12-month running average of the day 3 forecast skill relative to ERA-Interim of normalized TOA reflected solar flux (daily totals), verified against satellite data. The verification has been carried out for those parts of the northern hemisphere extratropics (green), tropics (red), and southern hemisphere extratropics (blue) which are covered by the CM-SAF product (approximately 70 S to 70 N, and 70 W to 70 E).

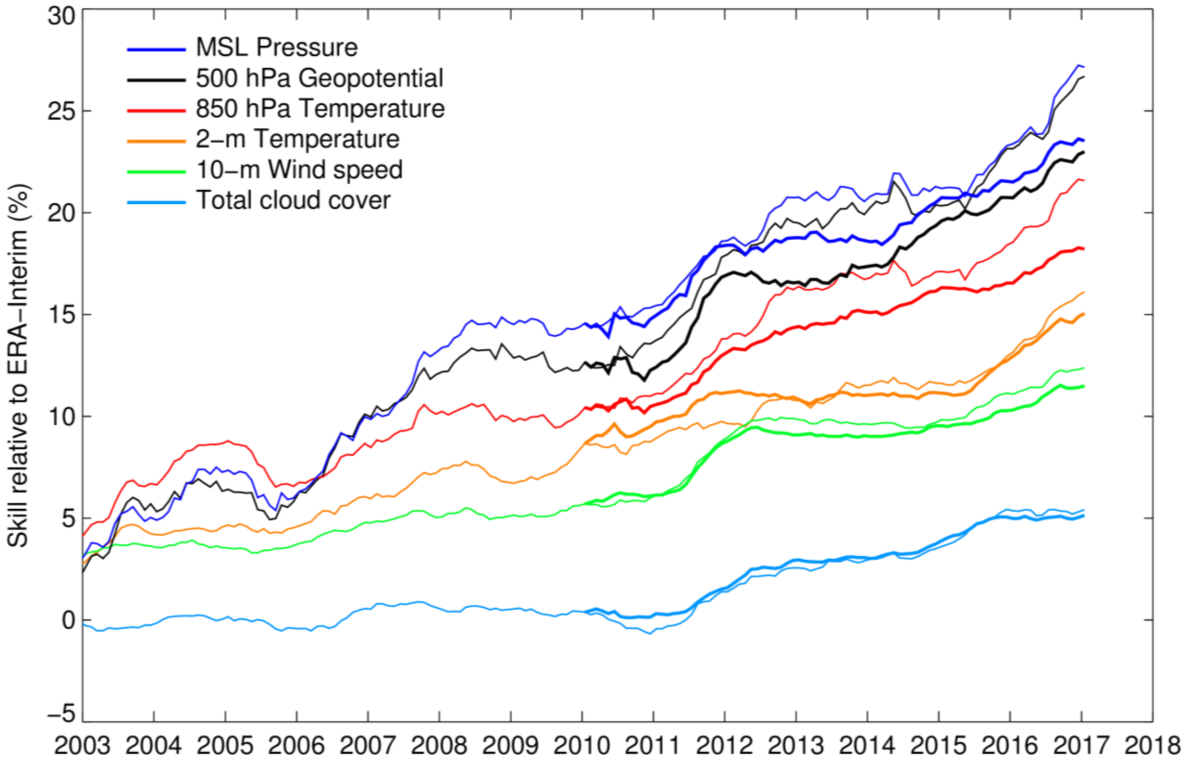


Figure 27: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA-Interim (thin lines). Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), and mean sea level pressure (MSLP), using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC). Thick lines from 2010 onwards show relative skill HRES v ERA-Interim, but with correction for loss of skill of ERA-Interim v ERA5.

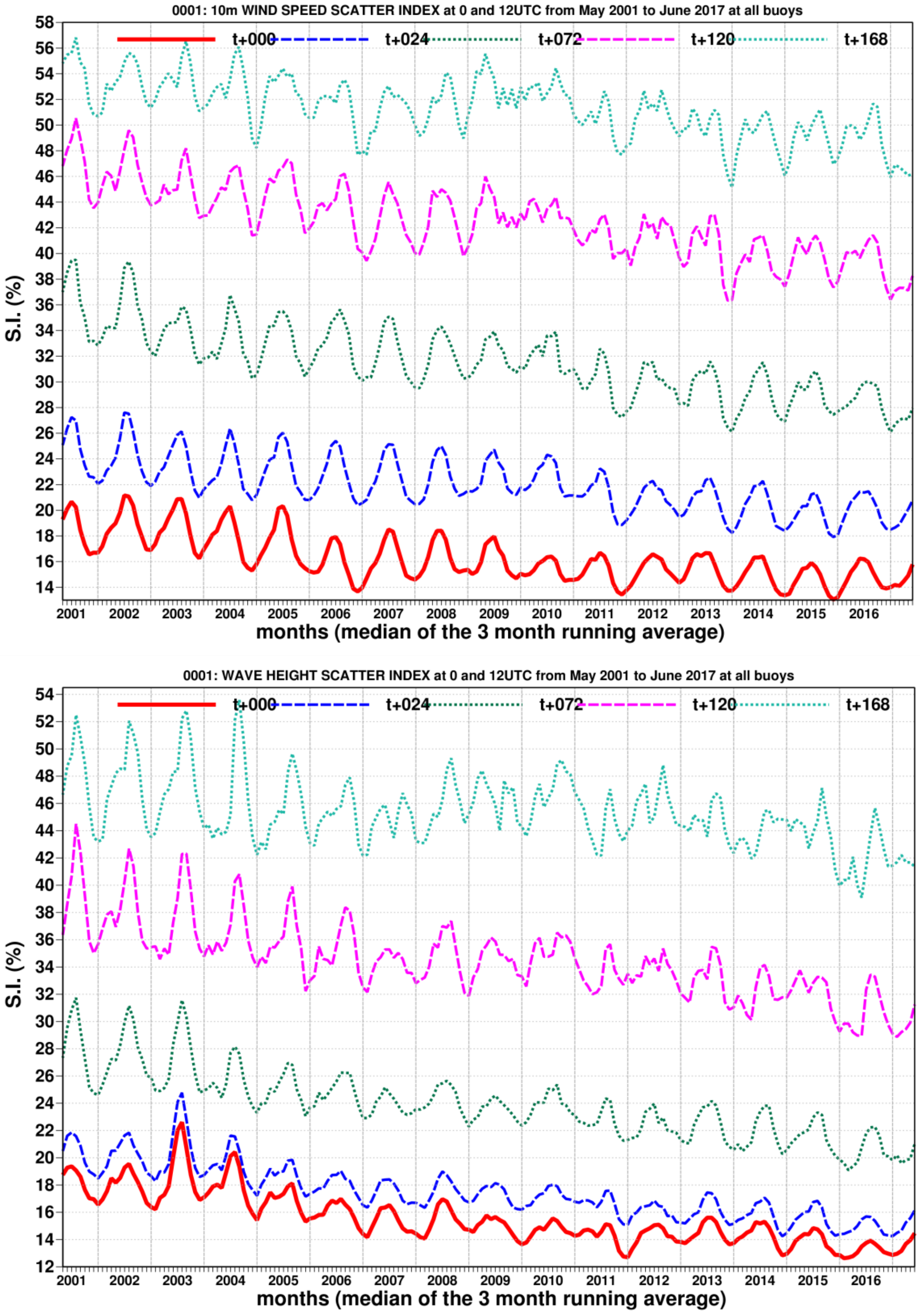
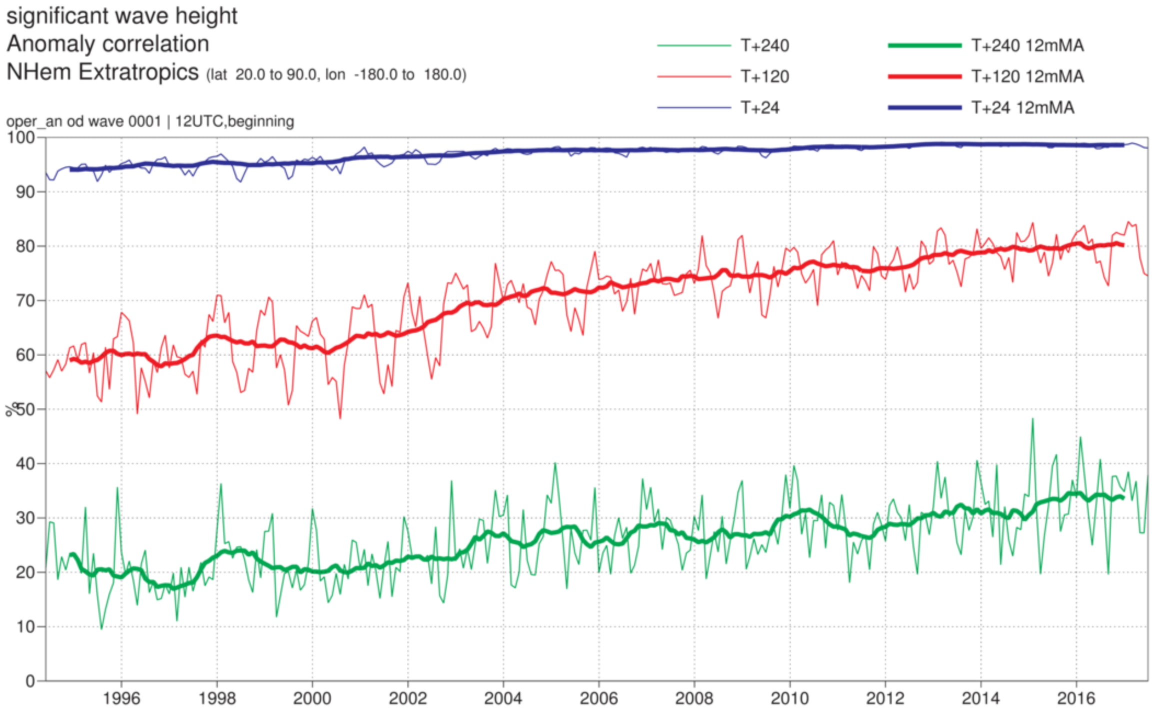


Figure 28: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running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



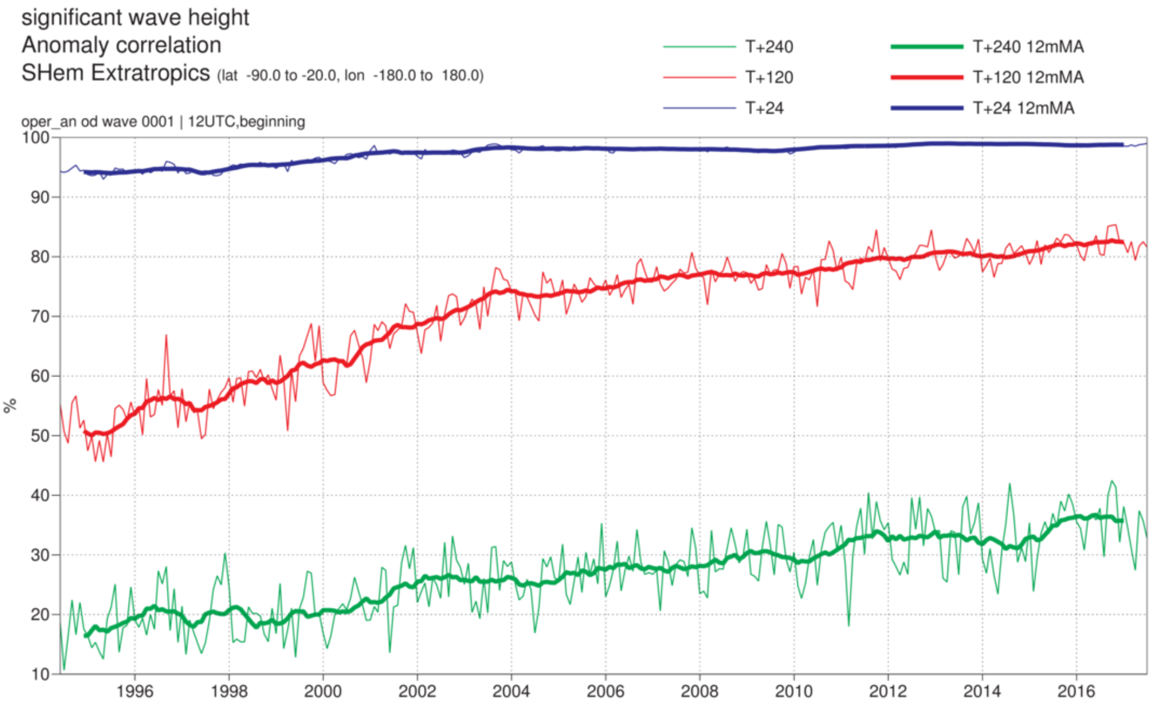


Figure 29: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

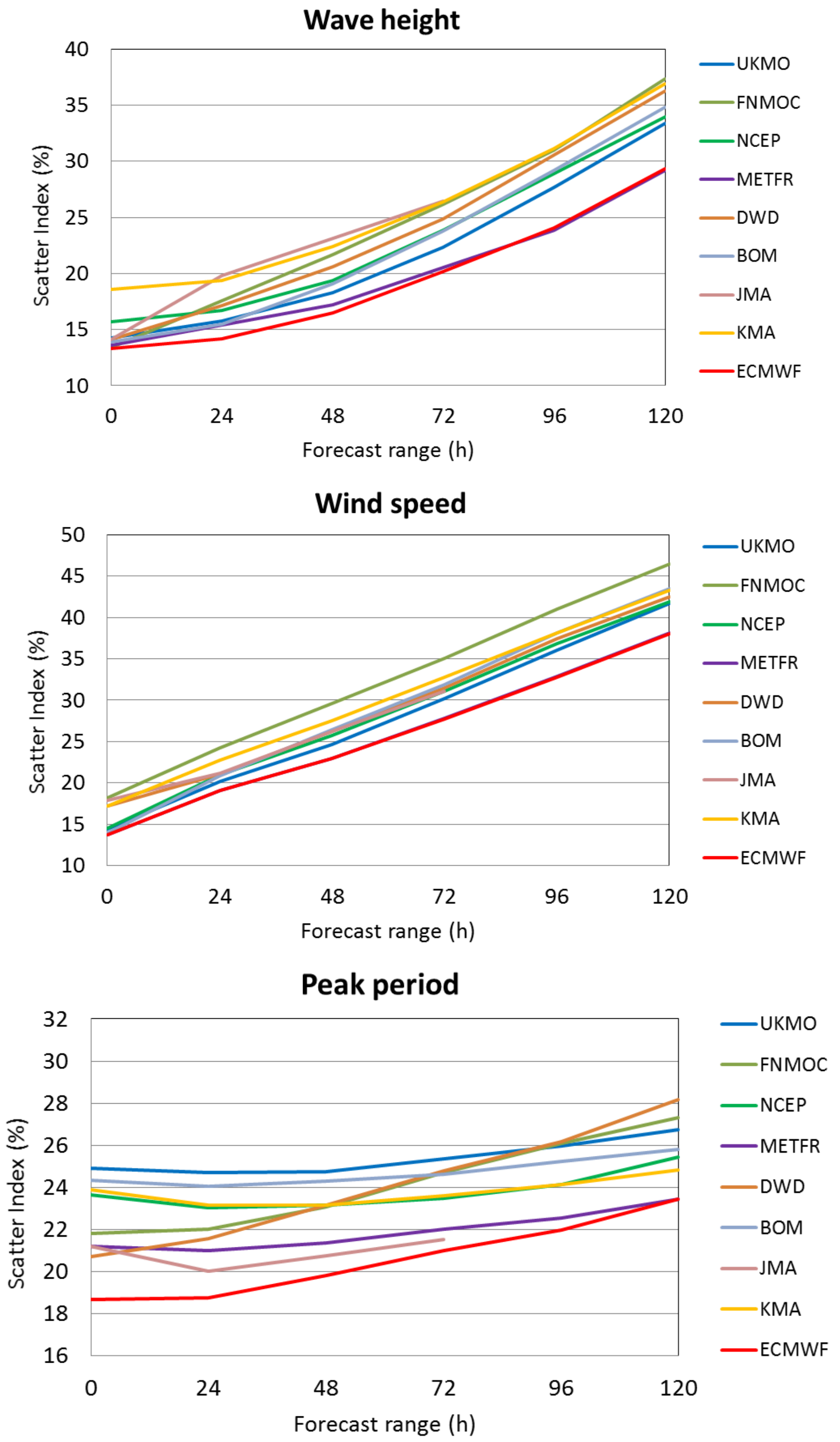


Figure 30: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period 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.

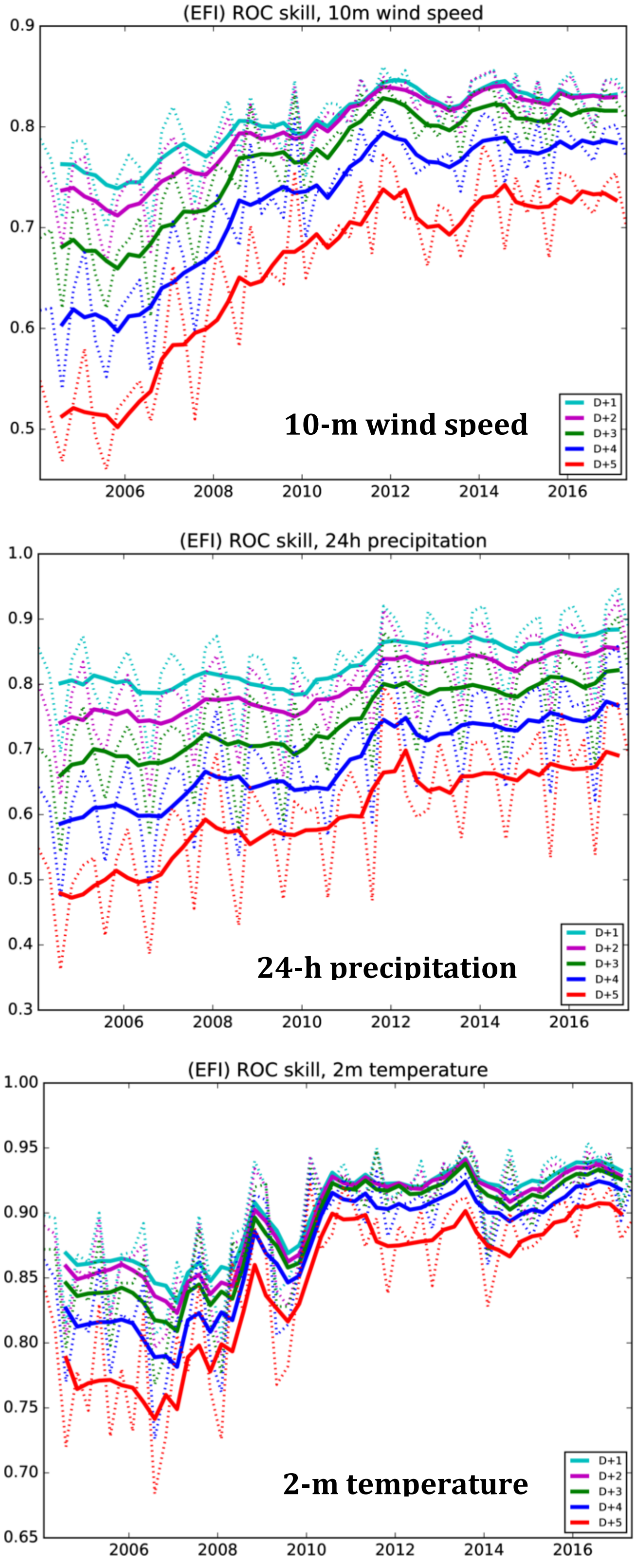


Figure 31: Verification of Extreme Forecast Index (EFI) against analysis. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 5 (24-hour period 96-120 hours ahead); skill at day 4 (blue line) is the supplementary headline score; an extreme event is taken as an observation 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.

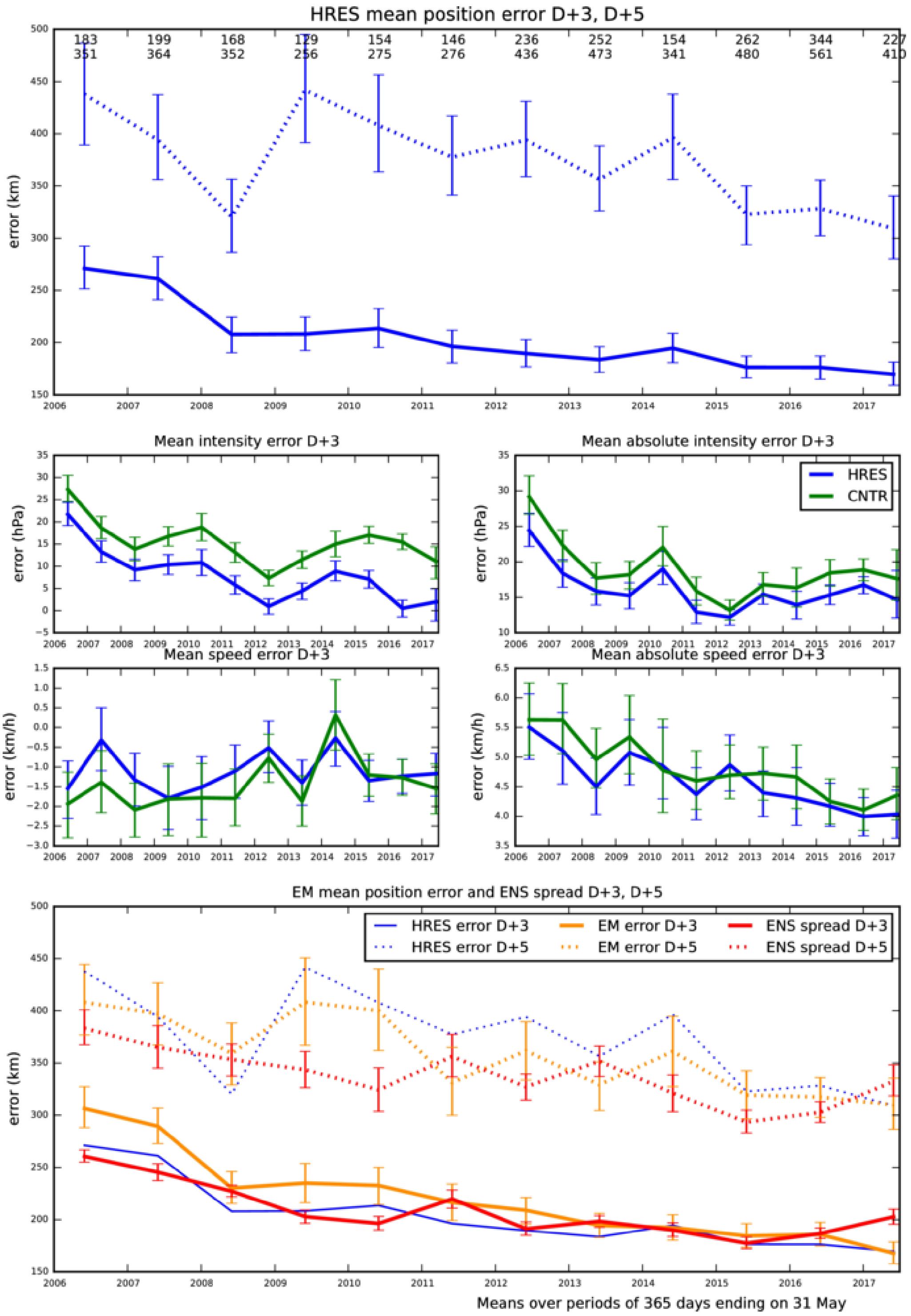


Figure 32: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 30 June. Verification is against the observed position reported via the GTS. Top panel supplementary 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 (orange 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).

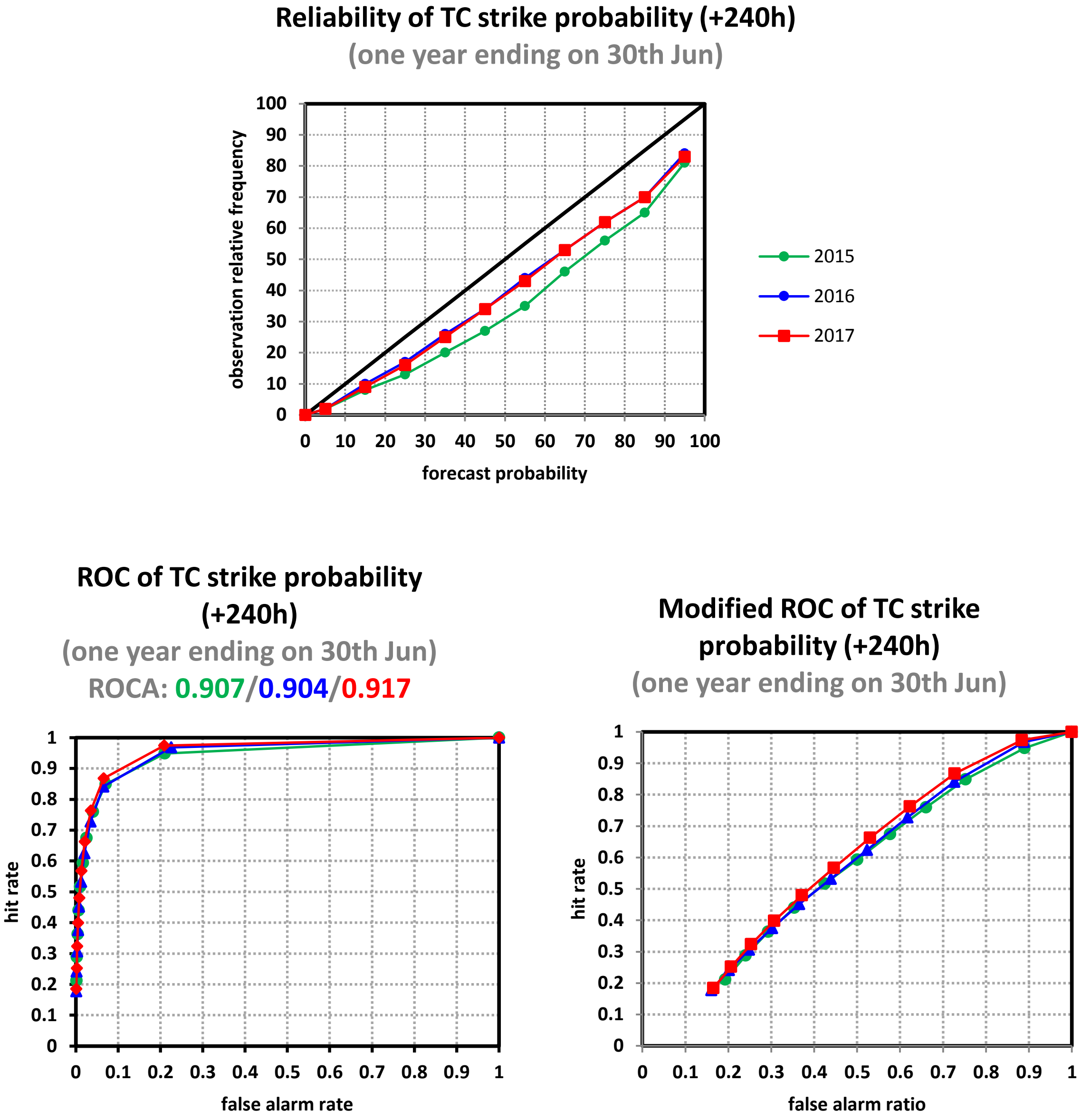


Figure 33: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2014–June 2015 (green), July 2015–June 2016 (blue) and July 2016–June 2017 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

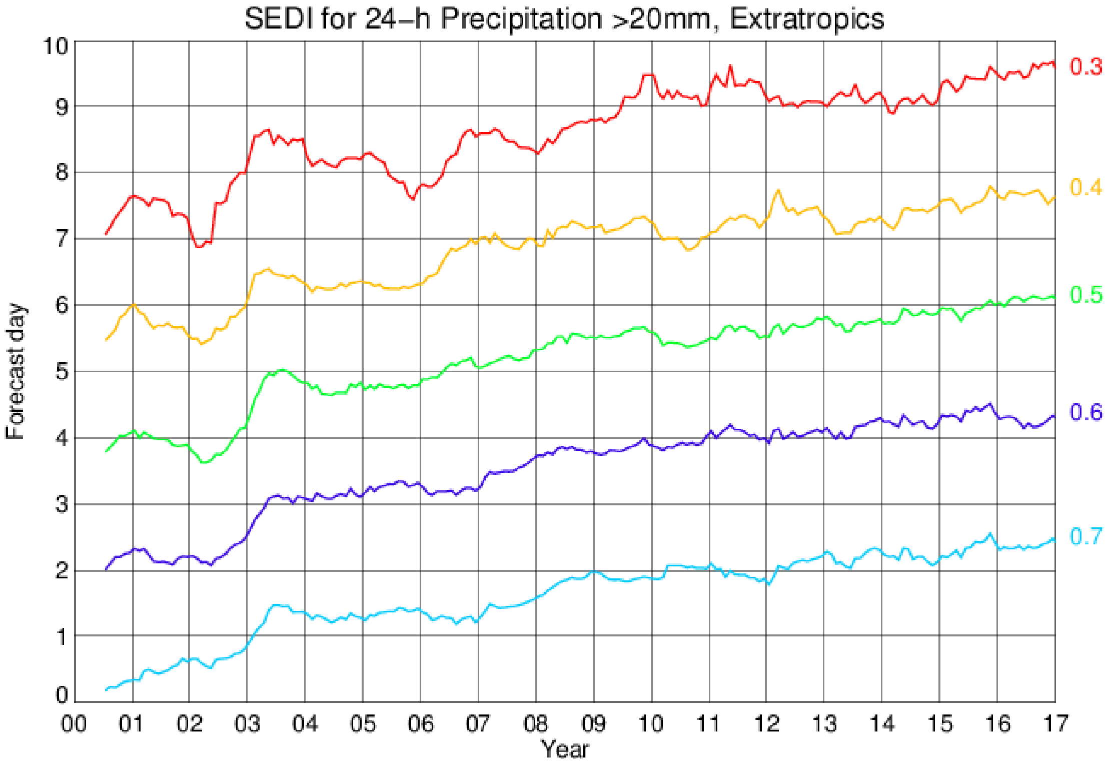


Figure 34: Evolution of skill of the HRES forecast in predicting 24-h precipitation amounts >20 mm in the extra-tropics as measured by the SEDI score, expressed in terms of forecast days. Verification is against SYNOP observations. Numbers on the right indicate different SEDI thresholds used. Curves show 12-month running averages.

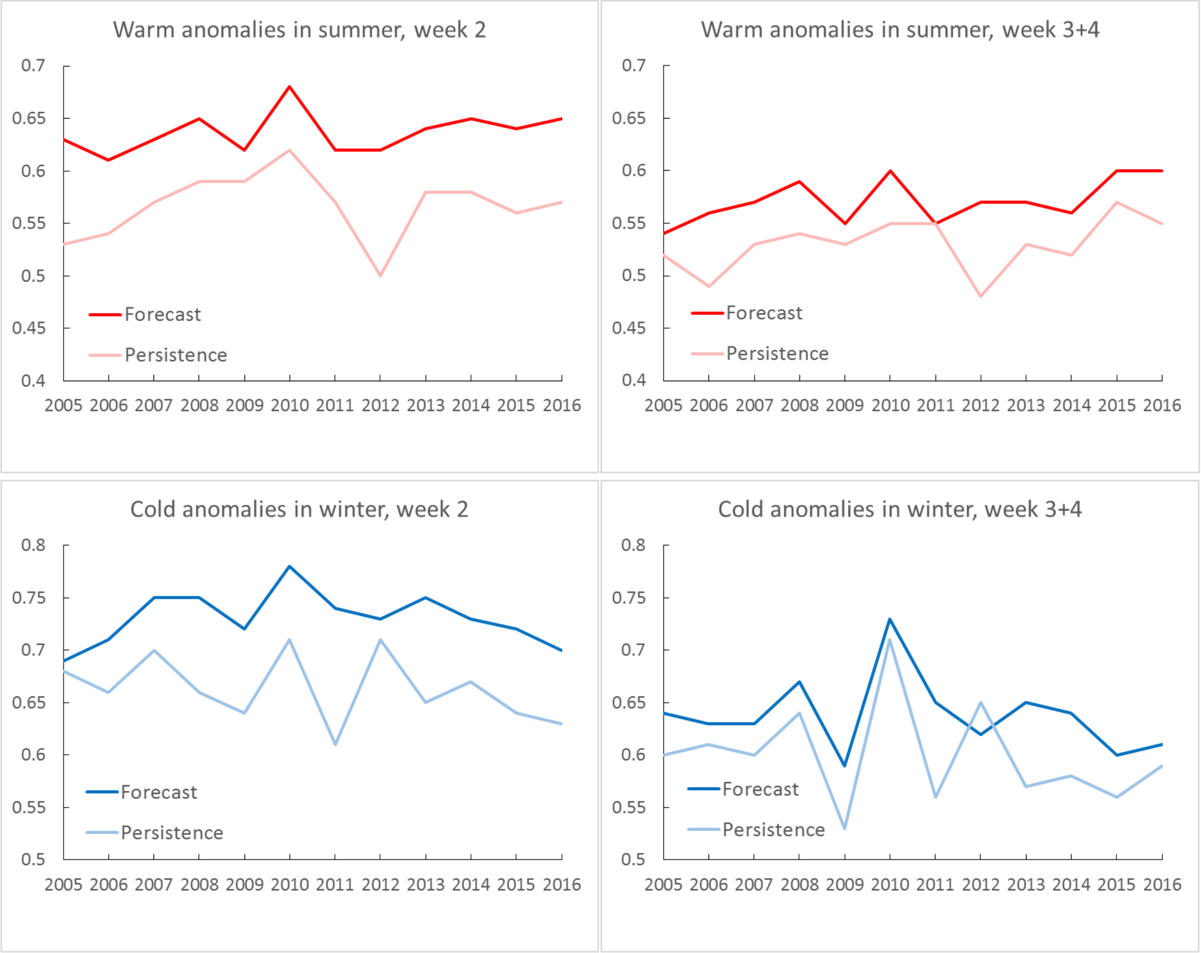


Figure 35: Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all land 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.

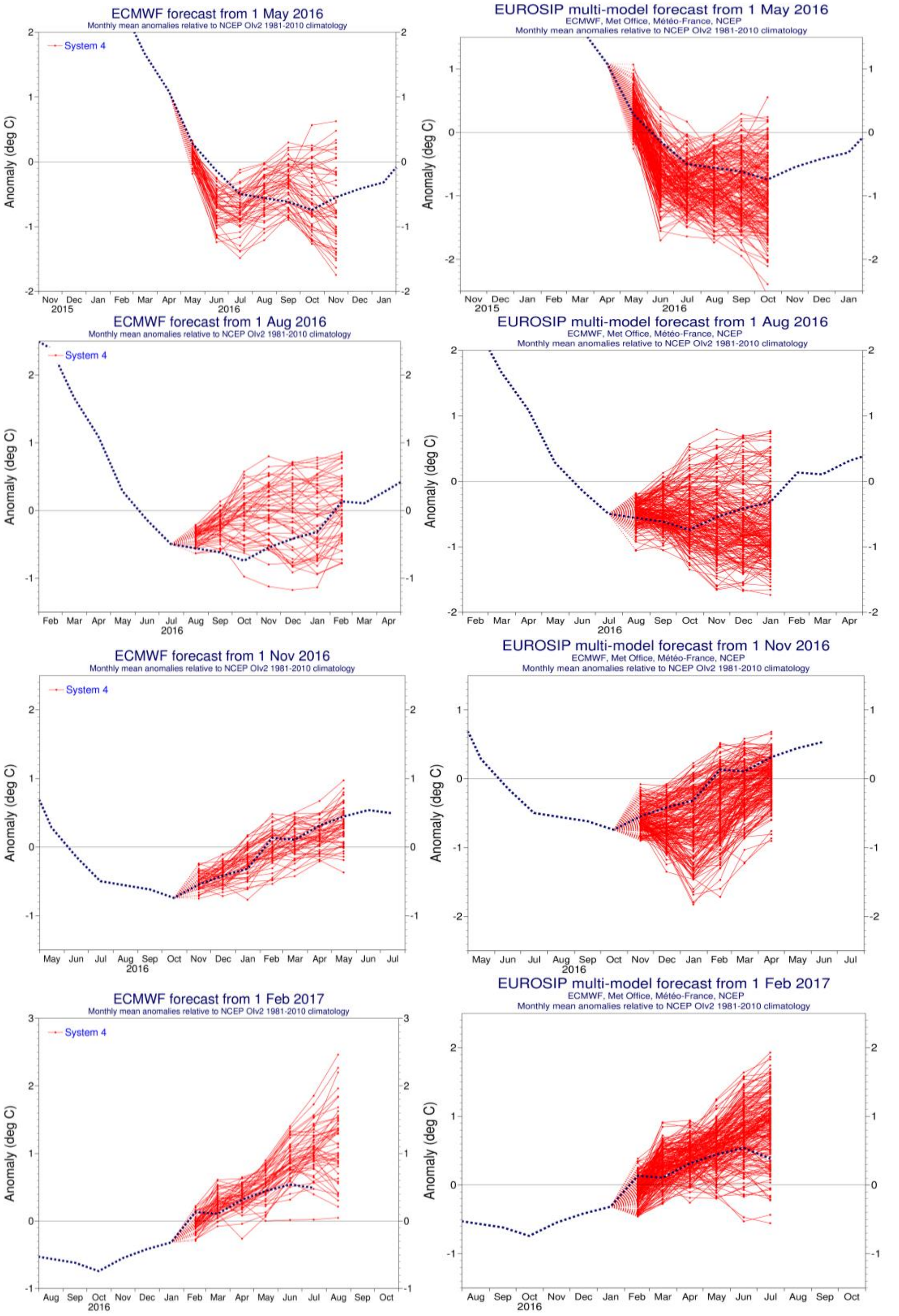


Figure 36: ECMWF (left column) and EUROSIP multi-model forecast (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2016, August 2016, November 2016 and February 2017. The red lines represent the ensemble members; dotted blue line shows the subsequent verification.

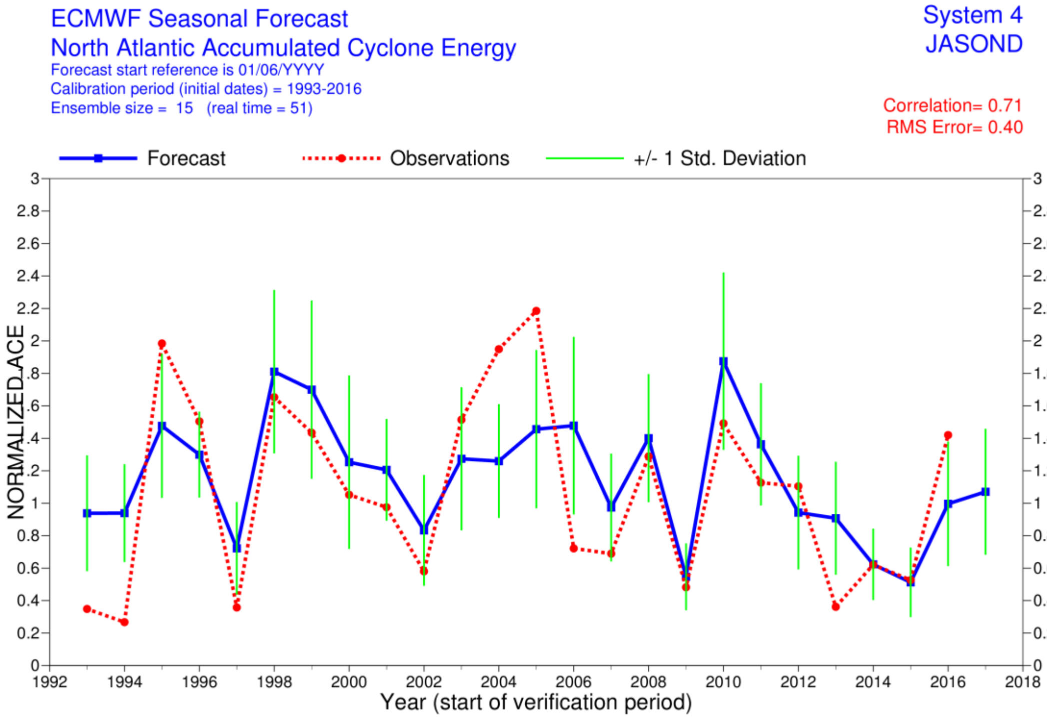


Figure 37: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2016. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts are from SEAS4 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.

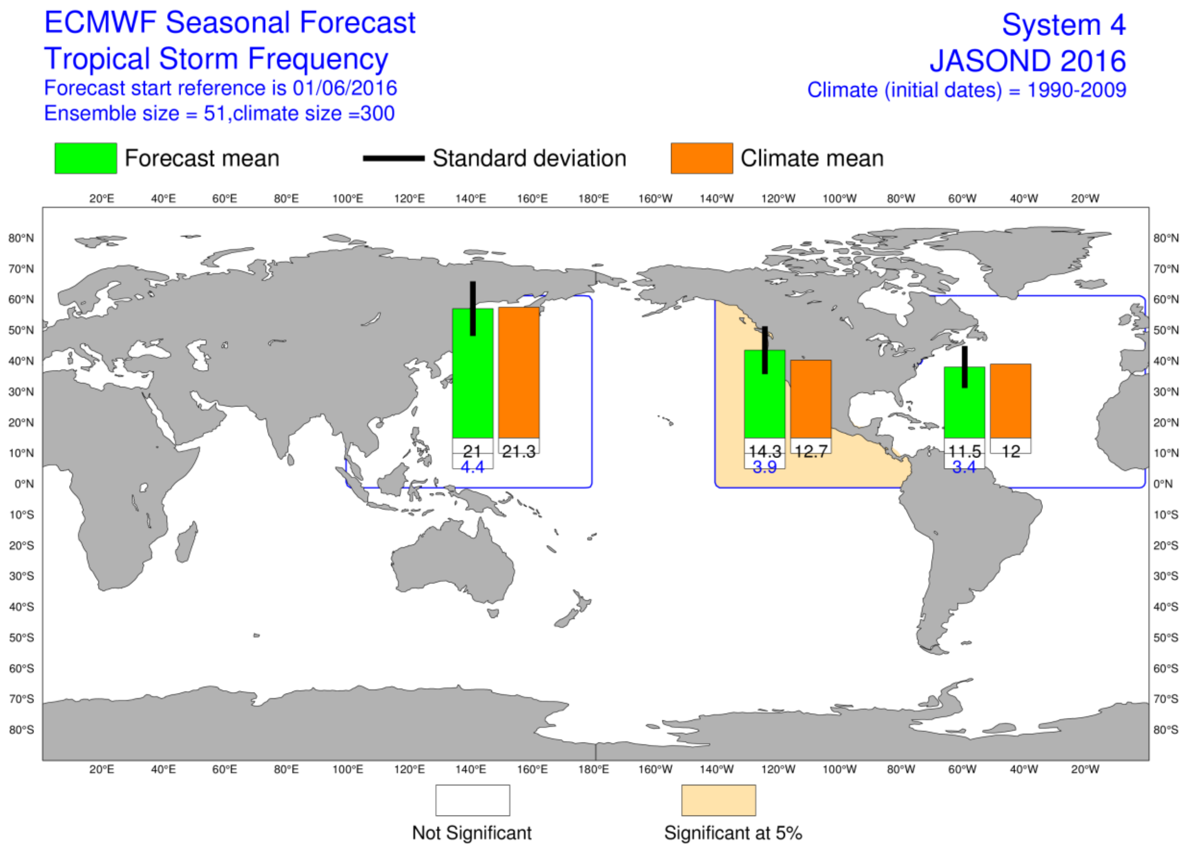


Figure 38**:** Tropical storm frequency forecast issued in June 2016 for the six-month period July–December 2016. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. 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.

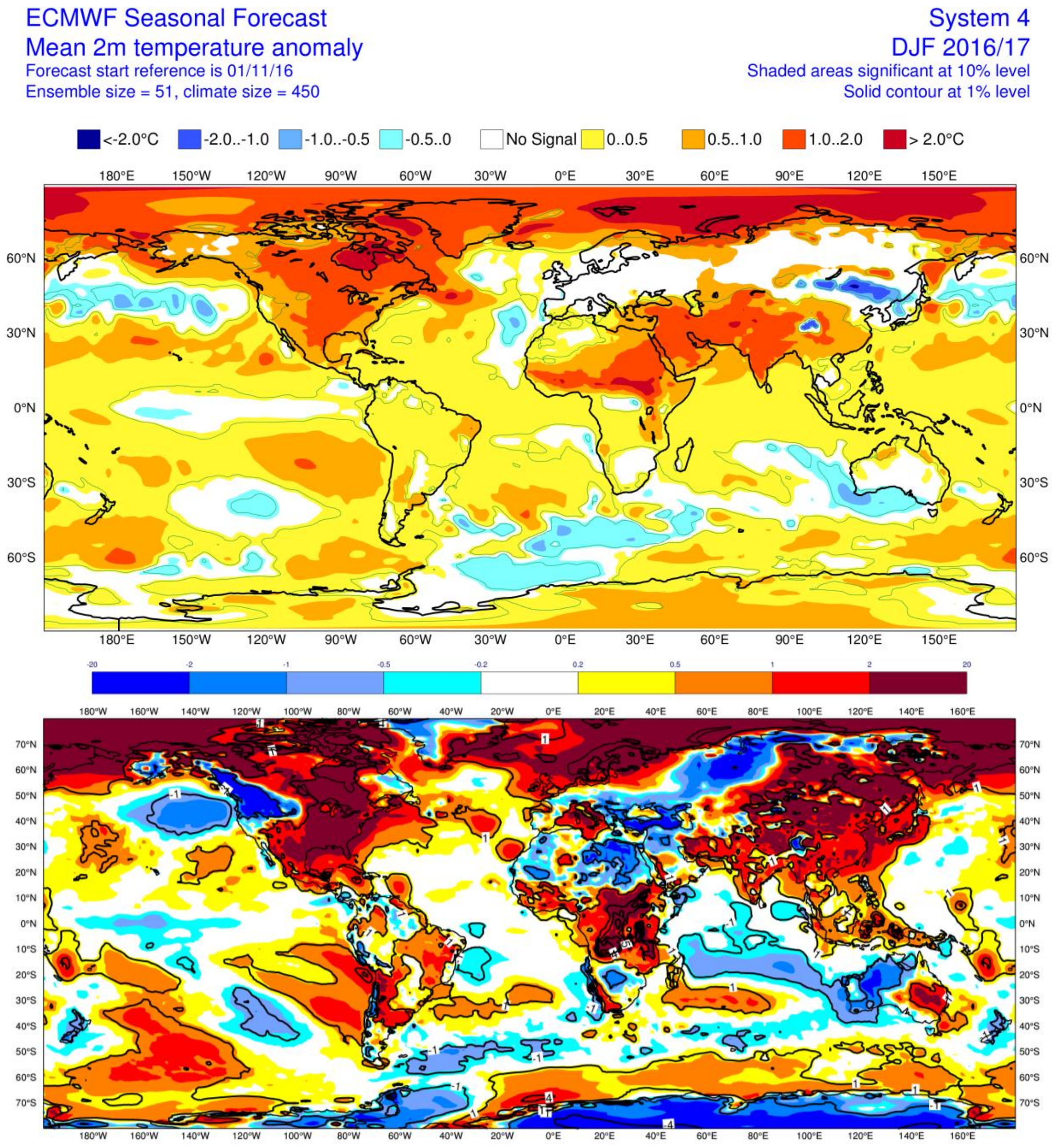


Figure 39**:** Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2016 for DJF 2016/17 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

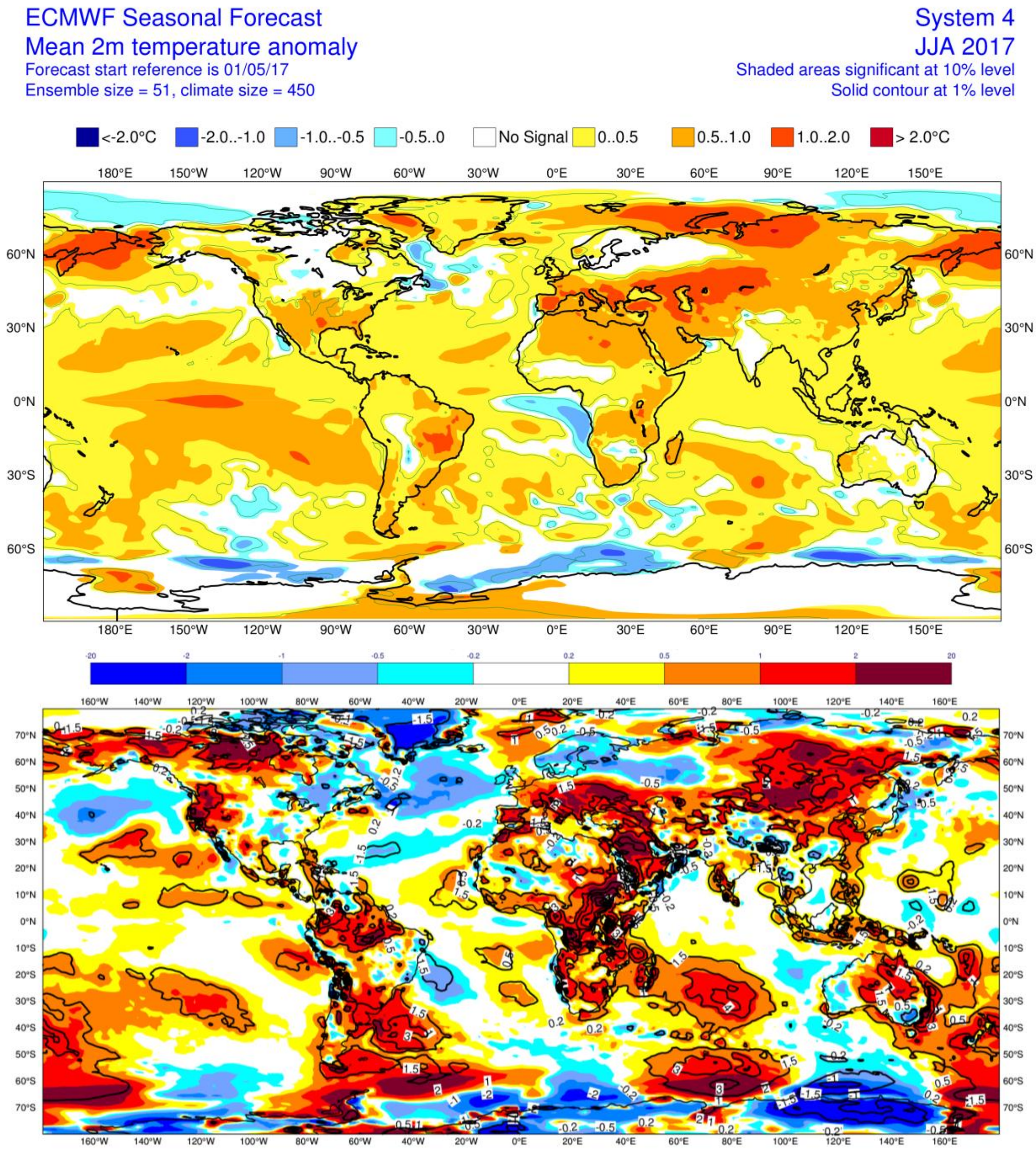

ECMWF Seasonal Forecast
Mean 2m temperature anomaly
Forecast start reference is 01/05/17
Ensemble size = 51, climate size = 450
System 4
JJA 2017
Shaded areas significant at 10% level
Solid contour at 1% level
<-2.0°C
-2.0..-1.0
-1.0..-0.5
-0.5..0
No Signal
0..0.5
0.5..1.0
1.0..2.0
> 2.0°C
-20
-2
-1
-0.5
-0.2
0.2
0.5
1
2
20

Figure 40: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2017 for JJA 2017 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

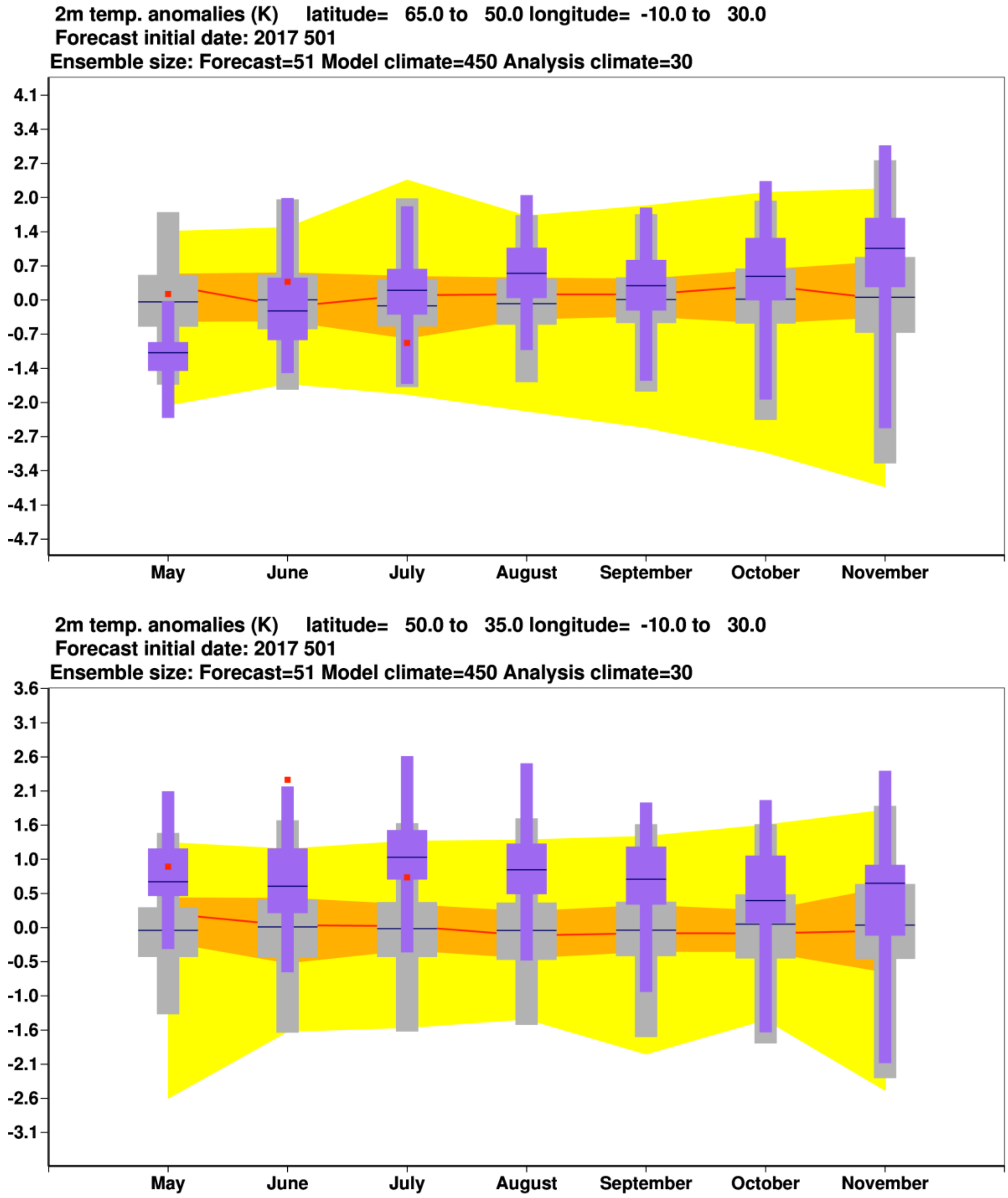


Figure 41: Long-range forecast of 2 m temperature anomalies from May 2017 for JJA 2017 for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-4 hindcasts is shown in grey, and the analysis in the 30-year hindcast period is shown in yellow and orange. The limits of the purple/grey whiskers and yellow band correspond to the 5th and 95th percentiles, those of the purple/grey box and orange band to the lower and upper tercile, and medians are represented by lines. The verification from operational analyses is shown as a red square. Areal averages have been computed using land fraction as a weight, in order to isolate temperature variations over land.

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 16), 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 16, Figure 18) 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 5 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 29) 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 9) and its inter-annual variability (Figure 13).

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 33). Figure 33 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 35.

A. 3 Weather parameters

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

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 20, Figure 21). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 22 to Figure 25), 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 34), 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

Buizza, R., J.-R. Bidlot, Janousek, M., Keeley, S., Mogensen, K., and Richardson, D., 2017: New IFS cycle brings sea-ice coupling and higher ocean resolution. *ECMWF Newsletter* n. **150**, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pg 14–17.

Buizza, R., Bechtold, P., Bonavita, M., Bormann, N., Bozzo, A., Haiden, T., Hogan, R., Holm, E., Radnoti, G., Richardson, D., and Sleigh, M., 2017: IFS Cycle 43r3 brings model and assimilation updates. *ECMWF Newsletter* n. **152**, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pg. 18–22.

Buizza, R., Anderson, E., Forbes, R., and Sleigh, M., 2017: The ECMWF Research to Operations (R2O) process. *ECMWF Research Department Technical Memorandum* n. **806** ECMWF, Shinfield Park, Reading RG2 9AX, UK, pp. 16.

Sleigh, M. et al., 2017: *ECMWF Newsletter* n. **152**, ECMWF, Shinfield Park, Reading RG2-9AX, UK, pg. 17.

Lavers, D. A., E. Zsoter, D. S. Richardson, and F. Pappenberger, 2017: An assessment of the ECMWF extreme forecast index for water vapour transport during boreal winter. *Weather and Forecasting* (in press).

Magnusson, L., 2017: Diagnostic methods for understanding the origin of forecast errors. *Q.J.R. Meteorol. Soc*, **143**: 2129–2142. doi:10.1002/qj.3072.

Rodwell, M.J., D.S. Richardson, D.B. Parsons, and H. Wernli, 2017: Flow-dependent reliability: A path to more skillful ensemble forecasts. *Bull. Amer. Meteor. Soc*. (submitted).

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.