Technical
Memo

ECMWF
European Centre for Medium-Range
Weather Forecasts

853

Evaluation of ECMWF forecasts, including the 2019 upgrade

T. Haiden, M. Janousek, F. Vitart, L. Ferranti
and F. Prates (Forecast Department)

November 2019

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, RG2 9AX, UK

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

1 Introduction

The most recent change to the ECMWF forecasting system (IFS Cycle 46r1, on 11 June 2019) is 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.

As in previous reports a wide range of 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, 817, 831). A few new plots have been included in response to specific requests by ECMWF's Technical Advisory Committee. 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 Meteorological content of IFS Cycle 46r1

On 11 June 2019, ECMWF implemented a substantial upgrade of its Integrated Forecasting System (IFS). IFS Cycle 46r1 includes changes in the model and in the data assimilation procedure used to generate the initial conditions for forecasts. The upgrade has had a very positive impact on the skill of medium-range and extended-range ensemble forecasts (ENS) and medium-range high-resolution deterministic forecasts (HRES). It follows the implementation of IFS Cycle 45r1 in June 2018, which brought coupling to all ECMWF forecasts, from one day to one year ahead, by including ocean and sea-ice models in the HRES configuration. Cycle 46r1 brings major changes in many areas, including:

In data assimilation: continuous data assimilation (an extra 4D-Var outer loop, an increase from 6 to 8 hours in the early-delivery assimilation window length, and an extension in the observation cut-off time); twice the number of members in the Ensemble of Data Assimilations (EDA); weakly-coupled data assimilation for sea-surface temperature in the tropics; consistent spatial interpolation of the model to observation locations in trajectories and minimisations; use of the EDA to calculate Jacobians in the soil-moisture analysis.

In the use of observations: assimilation of the SMOS neural-network soil-moisture product; assimilation of SSMIS-F17 satellite data at 150h GHz and GMI satellite data at 166 GHz; improved use of land/sea mask in the field of view for microwave imagers; introduction of inter-channel observation error correlations for ATMS and geostationary water-vapour channels; slant path calculations for geostationary radiances; usage of geostationary radiances at higher zenith angles; consistent infrared aerosol detection.

In the model: improvements in the convection scheme (entrainment, CAPE closure, shallow convection); activation of long-wave scattering in the radiation scheme; 3D rather than 2D aerosol climatology; correct scaling of dry mass flux in the diffusion scheme; improvement of the tangent linear and adjoint of the semi-Lagrangian departure point scheme in the polar-cap area; new parametrization for wind input and open ocean dissipation of the wave model; increase in the frequency of the ensemble radiation time step from 3 hours to 1 hour.

2.1.1 Data assimilation and observations

The continuous data assimilation scheme enables the use of later-arriving observations and, crucially, decouples the starting time of the assimilation calculations from the observational cut-off time. This permits the beneficial introduction of an additional outer loop without affecting delivery time. In addition, the early-delivery assimilation window length has been increased from 6 hours to 8 hours, thus ensuring that all observations that have arrived can be assimilated. For more details, see Lean et al. (2019).

The number of EDA members has increased from 25 to 50. The computational resources required are roughly the same as before as a result of efficiency improvements. The increase in the number of EDA members improves the HRES analysis by providing better background error variance and covariance estimates. Furthermore, it is now possible to assign a unique EDA perturbation to each ensemble forecast member, which makes the ensemble forecast members exchangeable. For more details, see Lang et al. (2019).

In the newly developed ocean–atmosphere weakly-coupled data assimilation, the atmospheric analysis sea-surface temperature in the tropics is taken from the ECMWF OCEAN5 near-real-time analysis, rather than from the OSTIA product directly. This results in improved forecast scores for near-surface temperature and humidity in the tropics compared to the analysis. For more details, see the article on weakly coupled data assimilation by Browne et al. (2019).

For the surface analysis of soil moisture, the Simplified Extended Kalman Filter (SEKF) described by de Rosnay et al. (2013) has been significantly upgraded to improve computational efficiency, by computing its Jacobians directly from the EDA rather than with perturbed nonlinear trajectories. This reduces the SEKF computing cost, compared to previous IFS cycles, by more than a factor of three in the operational HRES configuration. The EDA-Jacobian approach in the SEKF also enhances the coupling between the land and atmospheric assimilation systems by ensuring more dynamic Jacobian estimates than in the previous finite-difference approach.

Cycle 46r1 has introduced a package of changes to microwave all-sky assimilation. This includes the assimilation of SSMIS-F17 satellite data at 150h GHz and GMI satellite data (vertical and horizontal polarisation radiances) at 166 GHz, which bring new information on humidity and wind over tropical and subtropical oceans, as well as improving the use of the land–sea mask in the field of view for microwave imagers. Each microwave observation has a footprint depending on its frequency. We use the 10 GHz footprint for AMSR2 and GMI and the 19 GHz footprint for SSMIS-FOV to compute how the land–sea mask is affected by this footprint. This land–sea mask is more accurate than that used in Cycle 45r1, which depends on the resolution of each loop.

Inter-channel observation error correlations have been introduced for ATMS satellite data, which results in ATMS observations being assimilated, on average, with more weight. This has resulted in significant and consistent improvements in the fit of the short-range forecasts used in the data assimilation system (first-guess fit) to independent observations sensitive to temperature, humidity and wind, indicating improved forecasts of these variables.

Similarly, inter-channel observation error correlations have been introduced for geostationary satellite water vapour channels, affecting SEVIRI (Meteosat Second Generation) and AHI (Himawari) instruments, to provide the best first-guess fit to water vapour channels on other instruments, as well as impact at longer lead times.

A further upgrade to the use of geostationary radiances is to account for slanted paths within the radiative transfer calculation. This change enables us to use data up to zenith angles of 74°, thus improving coverage at the edges of the geostationary disks. This is particularly significant in the North Atlantic, where previously a significant amount of Meteosat-10 data was not used.

In addition, the SMOS (Soil Moisture and Ocean Salinity) neural-network soil moisture satellite product is now assimilated along with the ASCAT level-2 surface soil moisture satellite product. The impact of using SMOS neural network data and the EDA Jacobians on medium-range weather forecasts is near neutral. However, there is a small but significant improvement in 2-metre temperature forecasts in the short range in the northern hemisphere.

2.1.2 Model changes

In Cycle 46r1, the ENS radiation time step has been reduced from 3 hours to 1 hour, as is already the case for the HRES. Forecast skill is improved almost everywhere as a result, including a substantial error reduction for 2-metre temperature forecasts. Much of the improvement can be attributed to the faster coupling of radiation, clouds and the surface. Over tropical land areas, the root-mean-square error in low clouds has been reduced by as much as 15%. More frequent radiation updates incur an overall cost increase in the operational ENS of only about 3%. This was made possible in part because the new radiation scheme introduced in IFS Cycle 43r3 (ecRad) is significantly cheaper than its predecessor.

In addition, long-wave radiation scattering has been turned on in the radiation scheme, which leads to a slight warming of the surface and a reduction in the root-mean-square error in tropospheric temperature forecasts of around 0.5%. A key innovation in the implementation is to represent longwave scattering by clouds but to neglect it for aerosols (Hogan & Bozzo, 2018). This brings virtually all the benefits whilst enabling several optimisations to be performed, such that the overall cost of the radiation scheme when longwave scattering is included is very slightly reduced.

The 2D aerosol climatology used in the radiation scheme has been replaced by a new 3D aerosol climatology. This change has some positive impacts on lower tropospheric temperature and winds, especially along coastlines affected by seasonal biomass burning interacting with boundary layer clouds. Bigger positive impacts can be seen in the stratosphere, where the root-mean-square error of the temperature field in the 50–100 hPa layer near the summer pole decreases by 10% due to a similar reduction in the temperature bias.

Changes in the convection scheme include an increase in test-parcel entrainment; a correction for the denominator in the convective available potential energy (CAPE) closure (improving the tangent-linear approximation); and, for shallow convection, a relative-humidity-dependent area fraction for evaporation (previously a constant value).

A modification in the semi-Lagrangian advection scheme in tangent linear and adjoint coding results in improving the departure-point calculation near the polar cap area. This was a long-standing problem, which has in the past occasionally given rise to instabilities.

The changes introduced in the land-surface scheme aim to minimise the occurrence of spikes in the maximum 2-metre temperature. This was done by adjusting the wet-tile skin conductivity. This modification partially solves the spike problem, lowering the frequency of its occurrence by almost half, with a slightly positive net overall impact. In Cycle 46r1, the amount of rain that can refreeze when intercepted by the snowpack has been corrected, leading to improved handling of episodic snow events. Previously, unphysical accumulations of snow in rainy conditions were locally observed during wintertime.

A new wave physics parametrization for wind input and open ocean dissipation has been implemented in Cycle 46r1. It is based on the work of Ardhuin et al. (2010) and on an initial implementation in the Météo-France version of the wave model code. Because the wave model is coupled to the atmosphere, the new configuration was set up to yield a similar level of feedback in the form of a sea-state-dependent Charnock coefficient. This yields slightly larger ocean surface roughness under typical tropical wind conditions than before. The main benefit of the changes is on the wave parameters, partly addressing the issue of overprediction of long swell energy and the small underestimation in the storm tracks. Based on new parametrizations developed by Peter Janssen (2017) and Augustus Janssen, the freak wave parameter calculation has been updated. The main impact is an enhanced probability of larger waves in shallow water compared to the old version.

2.2 Meteorological impact of the new cycle

IFS Cycle 46r1 brings substantial improvements in forecast skill for both HRES (Figure 1) and ENS (Figure 2). Medium-range forecast errors in the extratropics are reduced by 1–5% for upper-air parameters and by 0.5–2% for surface parameters. Improvements of this magnitude are seen in verification against both the analysis and observations. In terms of lead time, upper-air improvements amount to a gain of around 2–3 hours. In the tropics, HRES results are predominantly positive, but there are some increases in temperature and humidity errors, mainly seen in verification against the analysis. For temperature, these are due to changes in the analysis and the introduction of the 3D aerosol climatology. ENS results in the tropics are also mixed. In addition to the changes mentioned already, they are affected by a minor reduction in spread (around 1%) due to changes in the deep convection scheme. Wave parameters (significant wave height and mean wave period) in the HRES are improved substantially by 5–10% due to the upgrade in the ocean wave model. Increased wave activity leads to some degradation in wave height at longer lead times in the ENS.

Precipitation forecast skill increases in the extratropics by about 0.5% in the ENS and 1% in the HRES. Other weather parameters, such as 2-metre temperature and 2-metre dewpoint, 10-metre wind speed and total cloud cover improve by about 1% in the ENS, and by 0.5–1% in the HRES when verified against observations. In the tropics, slightly reduced spread and increased bias lead to a very small (0.1–0.2%) degradation in ENS precipitation. Scores in the tropics show strong improvements for 2-metre temperature (4–8% against the analysis both in ENS and HRES, 1–2% against observations in the ENS). Tropical cyclone forecast skill is neutral overall, with a slight reduction in track error, consistent with improved winds in the tropics.

The extended-range impact of model changes associated with 46r1 is neutral, except for a small degradation of 2-metre temperature and precipitation skill scores in the tropics, linked to a small, but statistically significant, reduction of the ensemble spread. For the reforecasts, use of ERA5 instead of ERA-Interim as initial condition (from 46r1 onwards) gives significant improvements in weeks 1-2 in the extratropics, and up to week 4 in the tropics (Figure 3).

The new IFS cycle 46r1 will use the ERA5 data to initialize the re-forecasts and use ERA5 EDA to perturb the re-forecasts initial conditions.

3 Verification of upper-air medium-range forecasts

3.1 ECMWF scores

Figure 4 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 Europe, the 12-month running mean has reached 7 days in 2018, which is the highest value so far. This is partly due to an increase in predictability in 2018, as seen in ERA5 scores (not shown). Similarly, in the northern hemisphere, the 12-month running mean in 2018 slightly exceeded the previous highest value from 2016. Comparison with ERA5 indicates that northern hemispheric predictability was similar in the two years, and lower in 2017. In the southern hemisphere, the 12-month running mean in 2018 has been lower than in 2016 due to a reduction in predictability. However, relative to ERA5, skill has further increased since 2016.

A complementary measure of performance is the root mean square (RMS) error of the forecast. Figure 5 shows RMS errors for both extratropical hemispheres of the six-day forecast and the persistence forecast. In both hemispheres, the 12-month running mean RMS error of the six-day forecast has reached the lowest values so far.

Figure 6 shows the time series of the average RMS difference between four- and three-day (blue) and six- and five-day (red) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time. Consistent with the decrease in RMS error (Figure 5), the 12-month running mean of this metric has reached its lowest value so far in both hemispheres.

The quality of ECMWF forecasts in the upper atmosphere in the northern hemisphere extratropics is shown through time series of temperature and vector wind scores at 50 hPa in Figure 7. Downward trends seen in recent years are continuing, and for wind speed (lower panel), lowest 12-month running average values so far have been reached. For temperature (upper panel) at day 5, values are still higher than the minimum in 2003 due to a slightly larger bias in 2018 than in 2003.

The trend in ENS performance is illustrated in Figure 8, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. At these relatively large lead times (around day 9), year-to-year variations in atmospheric predictability affect the score evolution more strongly than in the early medium-range. Values since 2016 have been generally higher than before 2016 (with the exception of the high-predictability period in 2010), however on a year-to-year basis, no clear trend can be identified due to relatively large inter-annual variations. It should be noted that cycle 46r1 (implemented in June 2019) brings substantial improvements for the ENS (Figure 2). However, the full effect on 12-month running average scores will become visible only in 2020.

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 9. Both for 500 hPa geopotential height and 850 hPa temperature, forecasts show a good overall match between spread and error. For 850 hPa temperature, the under-dispersion at longer lead times has been further reduced, however it has increased in the early medium-range.

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 10 and Figure 11 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 11 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 9, and ideally should lie on the diagonal, and 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 the Japan Meteorological Agency (JMA) has the lowest error (although ECMWF has the better match between error and spread).

To create a benchmark for the ENS, the CRPS is also computed for a ‘dressed’ ERA5 forecast (replacing ERA-Interim, which was used for this purpose in previous years). This allows one to better distinguish the effects of IFS developments from those of atmospheric variability and produces a more robust measure of ENS skill. The dressing uses the mean error and standard deviation of the previous 30 days to generate a Gaussian distribution around ERA5. Figure 12 shows the evolution of the CRPS for the ENS and for the dressed ERA5 over the last 12 years for temperature at 850 hPa at forecast day-5. In both hemispheres, the skill of the ENS relative to the reference forecast is now slightly above 15%. As noted above, cycle 46r1 (implemented in June 2019) brings further improvements for the ENS which are, however, not yet visible in 12-month running averages.

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 13. Both at 200 hPa and 850 hPa, errors have decreased recently and reached (or are close to) their lowest values ever. 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 14 shows time series of such scores for 500 hPa geopotential height in the northern and southern hemisphere extratropics. Different from previous years, 12-month running averages, as well as a shorter period, are shown to better identify recent inter-annual trends. Over the last 8 years errors have decreased for all models, while ECMWF continues to maintain the lead.

WMO-exchanged scores also include verification against radiosondes. Figure 15 (Europe), and Figure 16 (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 ECMWF medium-range forecasts relative to the other centres when verified against observations. In the short range, ECMWF's lead is generally small (day 2), or essentially zero (day 1).

The WMO model intercomparison for the tropics is summarised in Figure 17 (verification against analyses) and Figure 18 (verification against observations), which show vector wind errors for 250 hPa and 850 hPa. When verified against the centres' own analyses, the JMA forecast has the lowest error in the short range (day-2) while in the medium-range, both ECMWF and JMA are the leading models in the tropics. In the tropics, verification against analyses (Figure 17) is sensitive to details of the analysis method, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 18), the ECMWF forecast has the smallest overall errors in the medium range.

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 19. 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%. The threshold has been chosen in such a way that the score measures the skill at a lead time of 3–4 days. The centre panel shows the differences of this score between HRES and ERA-Interim and between HRES and ERA5. 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 in such a way that the score measures the skill at a lead time of 6-7 days. Both scores are based on verification against SYNOP observations.

The deterministic precipitation forecast has reached its highest level of skill so far (red line in Figure 19). 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 19, top panel) or with the ERA5 reference forecast (light blue line in Figure 19, top panel). By taking the difference between the operational and ERA-Interim or ERA5 scores, most of this variability is removed, and the effect of model upgrades is seen more clearly (centre panel in Figure 19). The nearly linear increase indicates that each of the recent model upgrades contributed to improvements in the extratropical precipitation forecast.

The probabilistic precipitation score (lower panel in Figure 19) shows a long-term improvement as well. The peak at the end of 2015 was partly due to increased atmospheric variability, while the same values seen in 2018 are now 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 20). ECMWF’s probabilistic precipitation forecasts retain positive skill beyond day 9.

Trends in mean error (bias) and standard deviation for 2 m temperature, 2 m dewpoint, total cloud cover, and 10 m wind speed forecasts over Europe are shown in Figure 21 to Figure 24. 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 standard correction of 0.0065 K m^{-1} for the difference between model orography and station height is applied to the temperature forecasts.

For 2 m temperature (Figure 21), the reduction in error standard deviation (upper curves) which started around 2016, is continuing. The biases in 2 m temperature (lower curves) have however not been substantially reduced, except that the large annual variation of the night-time bias (blue curve) has become slightly smaller in recent years. The increasingly pronounced negative bias in spring is under investigation. Similar to 2 m temperature, 2 m dewpoint (Figure 22) shows a reduction of the error standard deviation. Here, negative biases have been slightly reduced as well. Systematic errors in near-surface parameters have been investigated in the USURF project (‘Understanding uncertainties in surface-atmosphere exchange’), which has helped to identify the causes of some of these biases and informed possible future model changes (Haiden et al., 2018; Schmederer et al., 2019).

For total cloud cover (Figure 23) both the error standard deviation and the bias show little change. For wind speed (Figure 24) the error standard deviation has reached its lowest values ever in summer 2018. There is no significant trend in the bias.

It is worth noting that the mean errors documented in Figure 21 to Figure 24 do not show the full range of biases on the regional scale, due to compensation effects. For example, in winter there is a positive night-time bias in 2 m temperature of several K in northern Scandinavia, while in the rest of Europe there is a negative bias of 0.5-1 K. As a result of USURF, these issues are now better understood, which helps to address them in future model cycles.

ERA5 (in the past, ERA-Interim) is useful as a reference forecast for the HRES, as it allows filtering out some of the effects of atmospheric variations on scores. Figure 25 shows the evolution of skill at day 5 relative to ERA5 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. Improvements in near-surface variables are generally smaller than those for upper-air parameters, partly because they are verified against SYNOP, which implies a certain representativeness mismatch. Over the last year, values have been largely stagnant, however further improvements are expected to result from cycle 46r1 (not yet apparent in 12-month running averages).

The fraction of large 2 m temperature errors in the ENS has been adopted as an additional ECMWF headline score. An ENS error is considered ‘large’ in this context whenever the CRPS exceeds 5 K. Figure 26 shows that in the annual mean (red curve) this fraction has decreased from about 7% to 5% over the last 15 years, and that there are large seasonal variations, with values in winter more than twice as high as in summer. Recent model upgrades, such as the resolution increase in 2016, have improved this score both in summer and winter. In winter, there has been a continuous decrease of the fraction of large errors from about 10% 10 years ago to about 7.5% in 2019, which amounts to a relative decrease of 25% over the last 10 years.

A similar measure of the skill in predicting large 10 m wind speed errors in the ENS is shown in Figure 27. Here, a threshold of 4 m/s for the CRPS is used, to obtain similar fractions as for temperature. As for temperature, the 2016 resolution upgrade has resulted in a substantial decrease of the large error fraction. The longer-term improvement shows a reduction of large wind speed errors from 5-6% to about 4%.

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. Recent errors in both 10 m wind speed and in the wave height forecast are comparable to those of the last two years. The long-term trend of improving performance of the wave model forecasts is also seen in the verification against analysis (Figure 29). Cycle 46r1 brings substantial improvements to the wave forecast (see Figure 1 and Figure 2) which are not yet visible in the 12-month running averages.

ECMWF has become the WMO Lead Centre for Wave Forecast Verification and in this role it collects forecasts from various forecast centres to verify them against buoy observations. An example of this comparison is shown in Figure 30 for the 3-month period March-May 2019. For wave height, ECMWF and Meteo-France (which uses ECMWF winds) generally lead other centres, while for peak period, the MetOffice has a clear lead, possibly in part due to grid refinement near coasts, where most of the measurements are made.

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

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

under ‘Ocean waves’. Verification results from the WMO Lead Centre for Wave Forecast Verification can be found at

https://confluence.ecmwf.int/display/WLW/WMO+Lead+Centre+for+Wave+Forecast+Verification+LC-WFV

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 2019. For wind speed, the highest skill so far has been reached both in the seasonal means and in the 12-month running average. For precipitation, the recovery from a drop in skill in 2018, which was due to a decrease in predictability (as concluded from comparison with ERA5), is ongoing. For temperature, ROC skill has reached a plateau in recent years, with some inter-annual variations.

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 at day 5 (top and bottom panels, Figure 32) have reached their lowest values so far. Mean absolute errors of intensity and speed of the HRES at D+3 have further decreased.

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. Relative to the error, the spread has however decreased slightly more in this year compared to the previous year. The figure also shows that the HRES position and ENS position errors are very similar in recent years.

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 240 hours. Verification of these probabilistic forecasts for the three latest 12-month periods is shown in Figure 33. Results show a small decrease in reliability compared to previous years (top panel), consistent with the slightly smaller spread. 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. For both measures, skill has decreased in the 2019 season.

5.3 Additional severe-weather diagnostics

The symmetric extremal dependence index, SEDI (Annex A.4), is used 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. There has been a continuous improvement in recent years across the whole lead time range up to 10 days. As for precipitation forecast skill in general, the positive effect of recent model upgrades can be seen.

6 Monthly and seasonal forecasts

6.1 Monthly forecast verification statistics and performance

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). In 2018, week 2 forecast skill for summer warm anomalies was unusually high, but persistence was also at its highest level within the period shown. The corresponding week 3+4 forecast skill and persistence were also near the upper end of values seen so far. Skill for winter cold anomalies in 2018 was close to average, however in week 2 there is an increasingly consistent margin relative to persistence in recent years.

Because of the low signal-to-noise ratio of real-time forecast verification in the extended range, re-forecasts are a useful additional resource for documenting trends in skill. Figure 36 shows the skill of the ENS in predicting 2 m temperature anomalies in week 3 in the northern extratropics. This is an additional headline score of ECMWF which was recommended by the TAC Subgroup on Verification. Verification against both SYNOP and ERA-Interim analyses shows that there has been a substantial increase in skill from 2005-2012, and little change (against analysis), and a slight decrease (against observations) thereafter. Note that the verification is based on a sliding 20-yr period, and is therefore less sensitive to changes from year to year than the real-time forecast evaluation but some sensitivity remains, e.g. due to major El Nino events falling within, or dropping out of, the sliding period.

An evaluation of forecast skill from the medium to the extended range in terms of large-scale Euro-Atlantic regimes and their effect on severe cold anomalies in Europe has been given by Ferranti et al. (2018).

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 SEAS5 of the seasonal component of the IFS (implemented in November 2017) includes updated versions of the atmospheric (IFS) and interactive ocean (NEMO) models and adds the interactive sea ice model LIM2. While re-forecasts span 36 years (from 1981 to 2016), the re-forecast period used to calibrate the forecasts when creating products uses the more recent period 1993 to 2016. Compared to the previous version, SEAS5 shows an improvement in SST drift, especially in the tropical Pacific, and improvements in the prediction skill of Arctic sea ice.

A set of verification statistics based on re-forecast integrations from SEAS5 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 user guide for SEAS5 is provided at:

https://www.ecmwf.int/sites/default/files/medialibrary/2017-10/System5_guide.pdf

6.2.2 The 2017–2018 El Niño forecasts

The year 2018 was characterized by a change from weakly negative to positive SST anomalies in the eastern tropical Pacific. This transition was captured in the forecast (Figure 37, top row), with EUROSIP and C3S giving a slightly better indication than ECMWF of the small magnitude of positive anomalies in summer 2018, with larger anomalies in the following months. However, the ECMWF forecast issued in August predicted the further temporal evolution well, providing a significantly narrower plume than EUROSIP or C3S. Forecasts in consecutive seasons were quite good as well, with observations within the ensemble spread for all three systems, although there was a slightly stronger tendency in the ECMWF forecast compared to EUROSIP and C3S to overestimate the magnitude of the warm anomalies at longer ranges.

6.2.3 Tropical storm predictions from the seasonal forecasts

The 2018 Atlantic hurricane season had a total of 15 named storms including 8 hurricanes and 2 major hurricanes. It was the third consecutive above-average and damaging season, with an accumulated cyclone energy index (ACE) of about 125% of the 1993–2015 climate average (Figure 38). Seasonal tropical storm predictions from SEAS5 indicated a below average level of activity over the Atlantic (ACE of about 50% (+/- 20%) of the 1993–2015 climate average). Similarly, the number of tropical storms (15) which formed in 2018 was above average (13) whereas the forecast predicted 8 (with a range from 6 to 10) tropical storms in the Atlantic (Figure 39). The following forecasts, issued in July and August, also predicted a below average season. This poor seasonal forecast can partly be related to the tendency for an overestimation of the amplitude of the 2018-19 El-Nino (conducive to reduced tropical cyclone activity over the North Atlantic) by SEAS5.

The figure also shows that SEAS5 predicted average activity over the eastern North Pacific, and above average activity over the western North Pacific (ACE of about 120% of the 1993–2015 climate average).

The 2018 Pacific typhoon season was an above-average season producing 29 storms, 13 typhoons, and 7 super typhoons, with an ACE about 20% above average, as predicted by SEAS5. The eastern North Pacific hurricane season was an above-average season with 19 named storms from July to December, while SEAS5 predicted only 12.

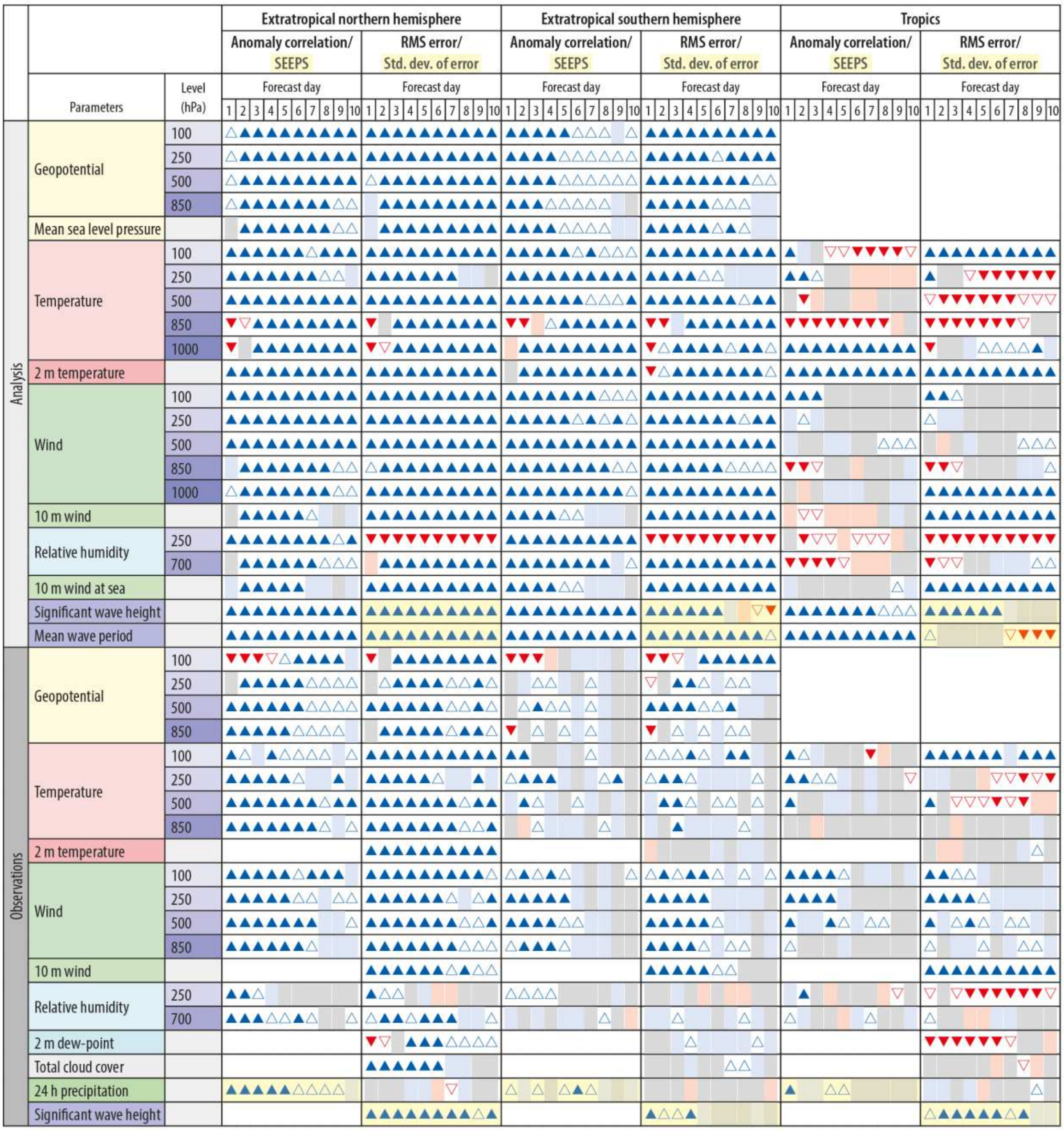
6.2.4 Extratropical seasonal forecasts

Because of the lack of a strong El Nino or La Nina signal, low seasonal predictive skill was likely in 2018. The pattern of 2 m temperature in the northern-hemisphere winter (DJF 2018–19) was characterized by strong warm anomalies in Europe, large parts of Siberia, and Alaska. A pronounced cold anomaly was present over Canada and the western United States. The high-latitude warm anomalies are a combination of the effect of global warming and inter-annual variability, and were captured reasonably well by the seasonal forecast (Figure 40). However, anomaly patterns in Europe, as well as the cold anomaly in Canada, were not correctly predicted.

Large parts of Europe experienced a very hot summer season in 2019. For the northern-hemisphere summer (JJA 2019) the forecast predicted positive anomalies over Central and Southern Europe, which agreed with observations (Figure 41). The forecast also qualitatively captured the small-scale negative anomaly near Portugal and the fact that the positive anomalies would not extend into Finland. Anomalies of both signs over the North Atlantic were also predicted quite well. Major discrepancies between forecast and analysis in the Northern Hemisphere occurred in western Canada, where the model failed to predict strong cold anomalies, and in northeast Siberia, where large warm anomalies were not predicted.

In the tropics and the Southern Hemisphere, remarkable similarities between predicted and analysed anomalies can be seen in the Indian Ocean / Australian sector, while over the Atlantic and Pacific sectors there is less agreement.

Climagrams for Northern and Southern Europe for winter 2018-19 and summer 2019 are shown in Figure 42. Red squares indicate observed monthly anomalies. The sign of the forecast anomaly is often predicted correctly, and the observations usually fall within the ensemble distribution. A notable exception was the cold May in southern Europe (lower panels), where the observed anomaly was outside the ensemble even at very short range (1 month).


Extratropical northern hemisphere
Extratropical southern hemisphere
Tropics
Anomaly correlation/ SEEPS
RMS error/ Std. dev. of error
Forecast day
1 2 3 4 5 6 7 8 9 10
Parameters
Level (hPa)
Analysis
Observations
Geopotential
100
250
500
850
Mean sea level pressure
Temperature
1000
2 m temperature
Wind
10 m wind
Relative humidity
700
10 m wind at sea
Significant wave height
Mean wave period
2 m dew-point
Total cloud cover
24 h precipitation

Symbol legend: for a given forecast step...

▲ 46r1 better than 45r1 statistically significant with 99.7% confidence

△ 46r1 better than 45r1 statistically significant with 95% confidence

🟦 46r1 better than 45r1 statistically significant with 68% confidence

⬜ no significant difference between 45r1 and 46r1

🟧 46r1 worse than 45r1 statistically significant with 68% confidence

▽ 46r1 worse than 45r1 statistically significant with 95% confidence

▼ 46r1 worse than 45r1 statistically significant with 99.7% confidence

Figure 1: Summary score card for IFS Cycle 46r1. Score card for HRES cycle 46r1 versus cycle 45r1 verified by the respective analyses and observations at 00 and 12 UTC for 690 forecast runs in the period June 2017 to June 2019. Yellow colouring indicates that symbols refer to the second score indicated at the top of the column.

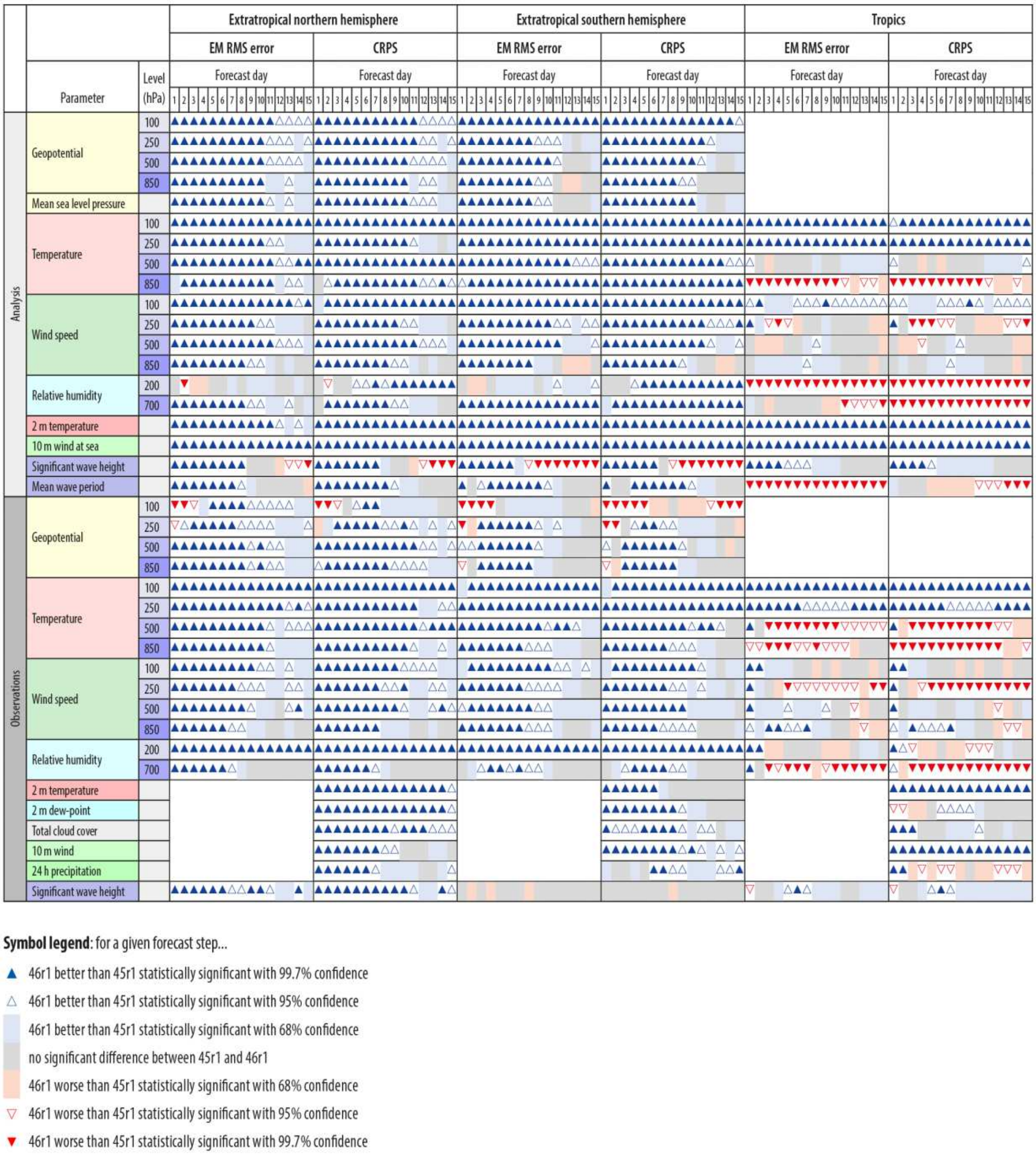


Figure 2: Summary ENS score card for IFS Cycle 46r1. Score card for ENS cycle 46r1 versus cycle 45r1 verified by the respective analyses and observations at 00 UTC for 282 ENS forecast runs in the period June 2017 to June 2019.

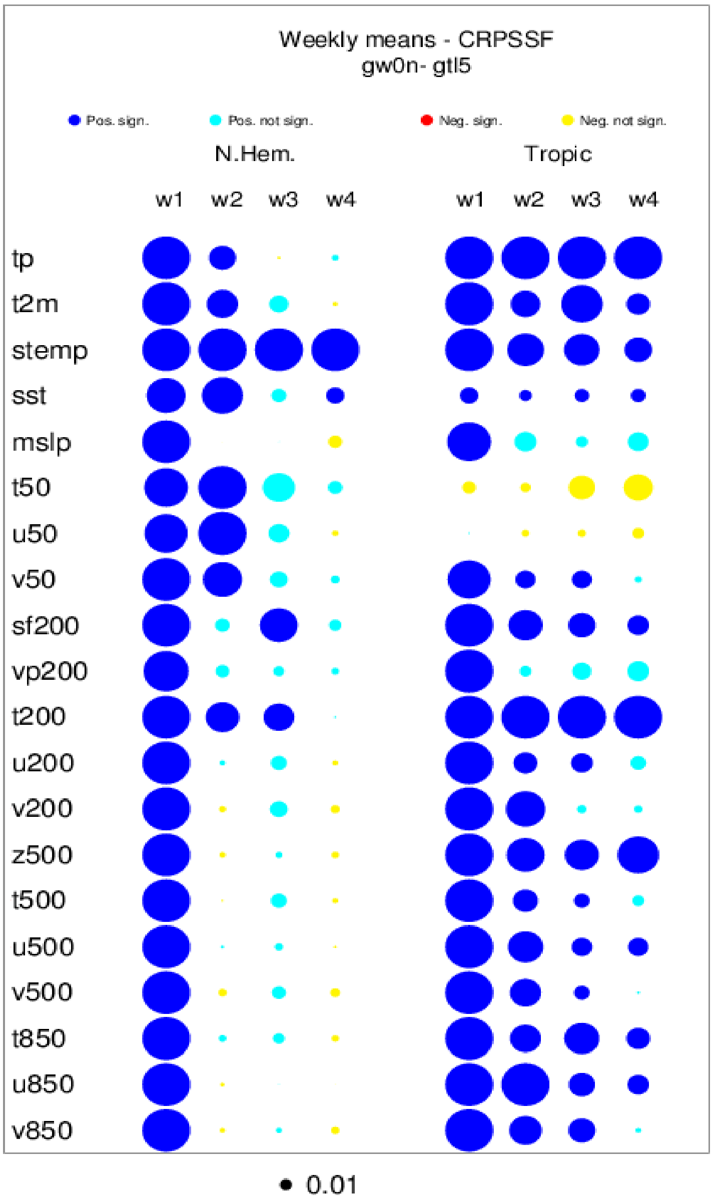


Figure 3: Improvements in the skill of reforecasts in the extended range from the use of ERA5 as initial condition (operational from 46r1 onwards). Columns show score differences for weekly means in the Northern Extratropics and the Tropics. Size of circles shows magnitude of difference, colour indicates statistical significance.

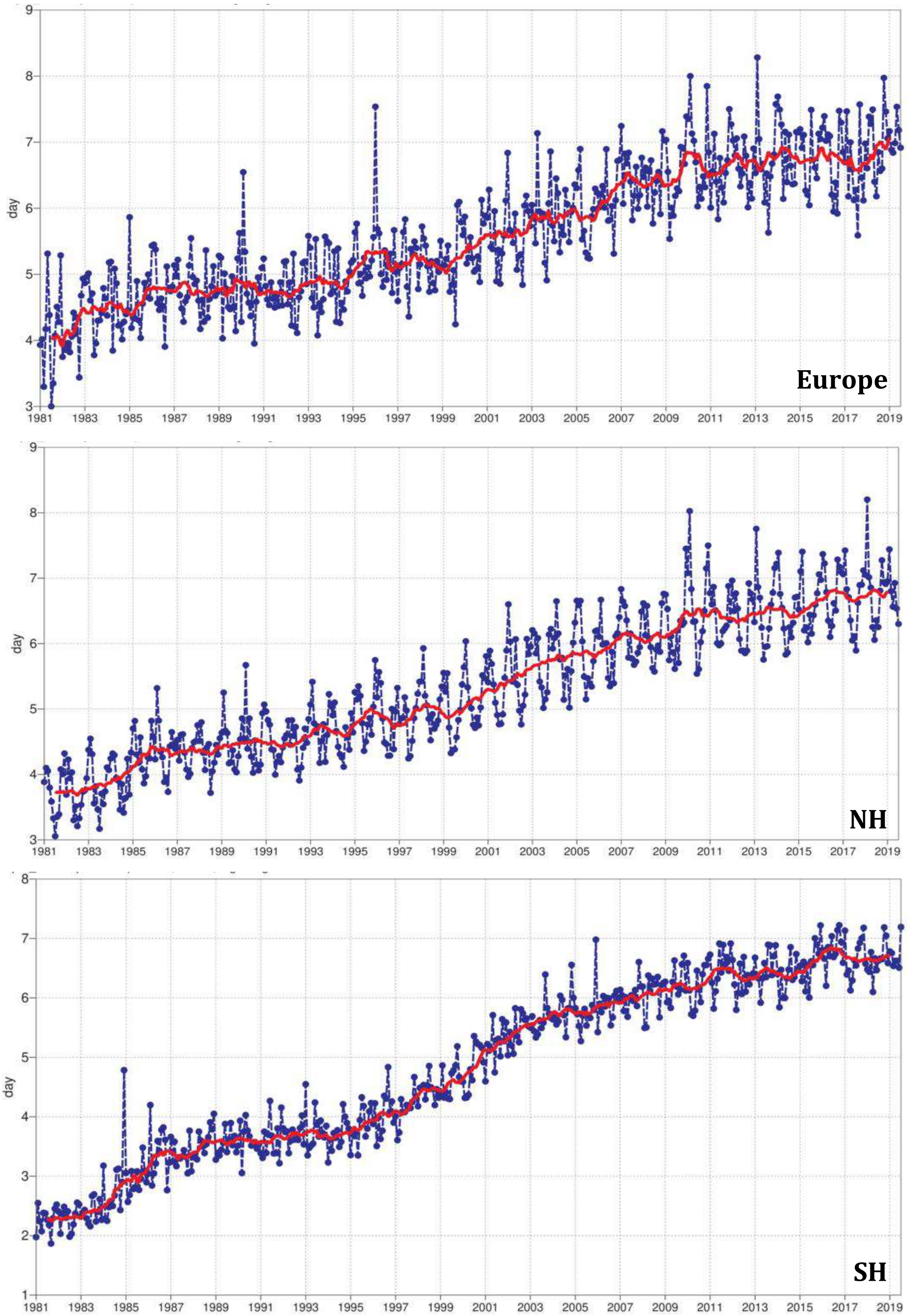
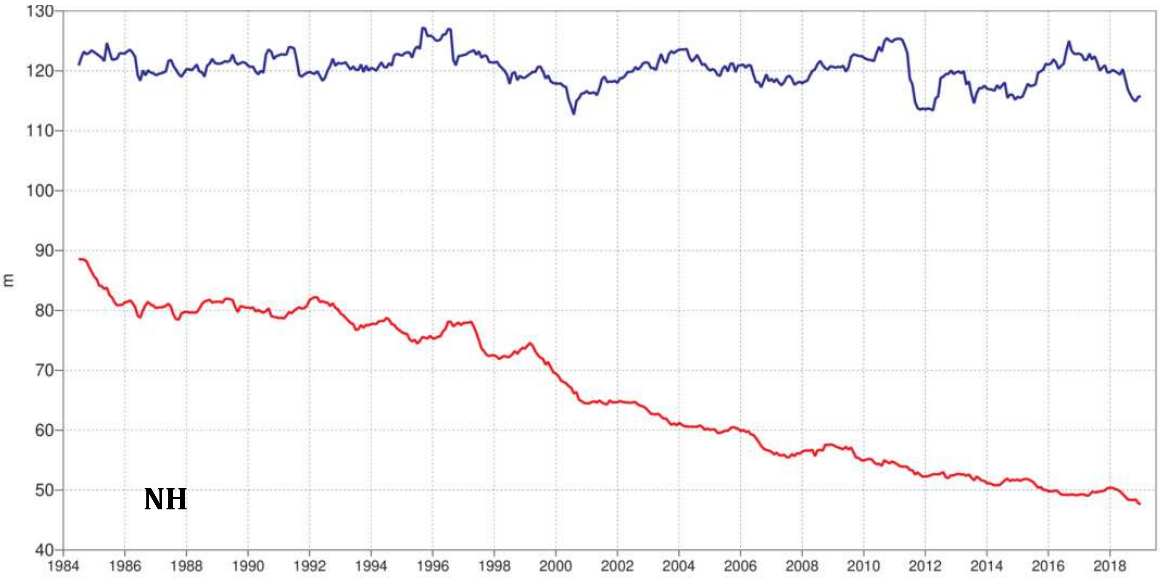


Figure 4: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).



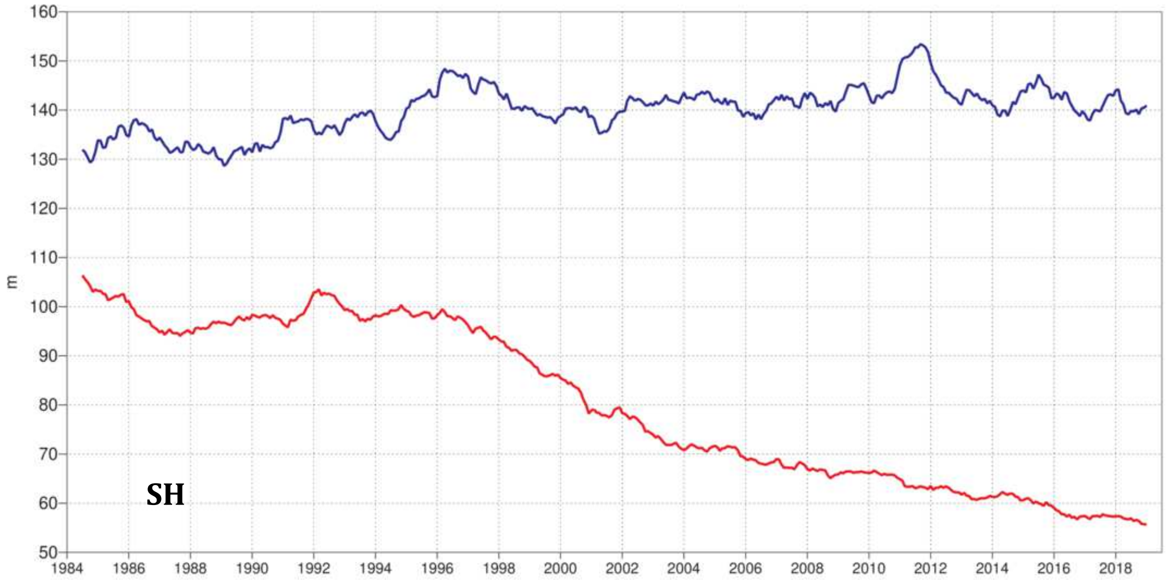
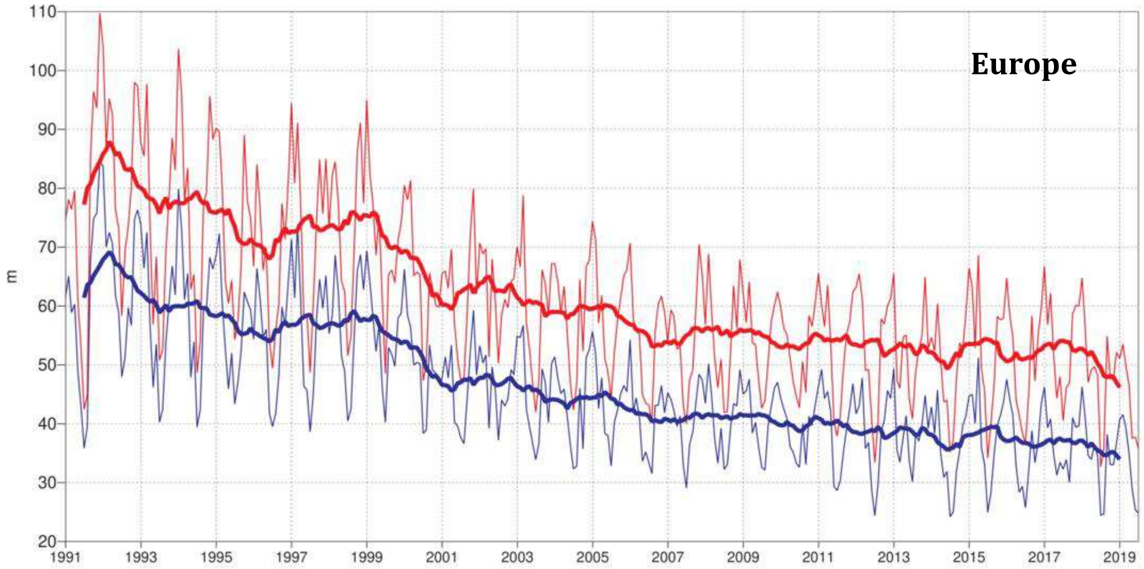


Figure 5: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2018–July 2019. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).



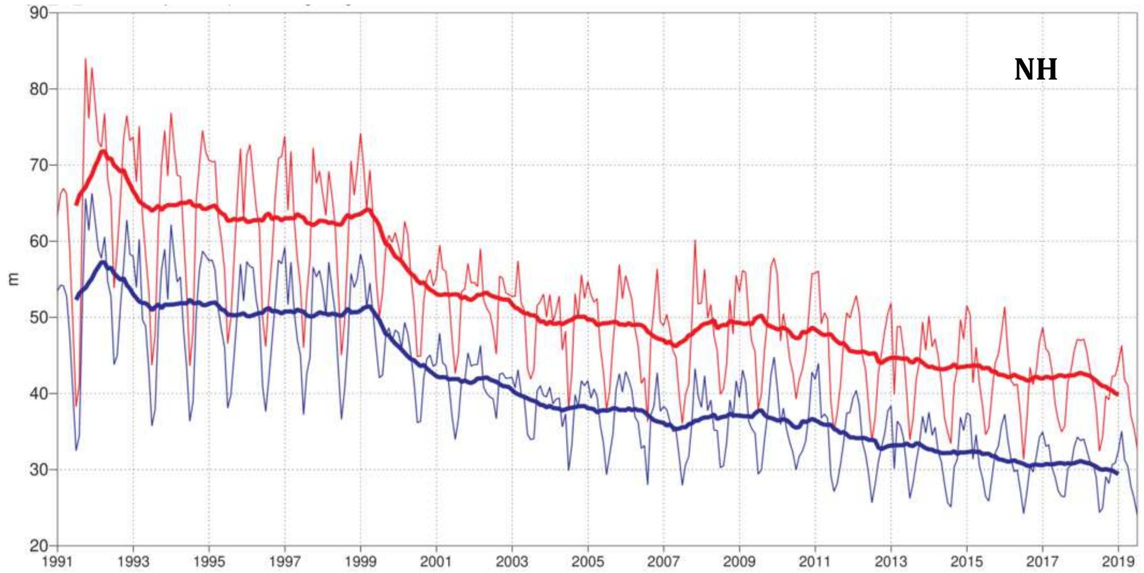
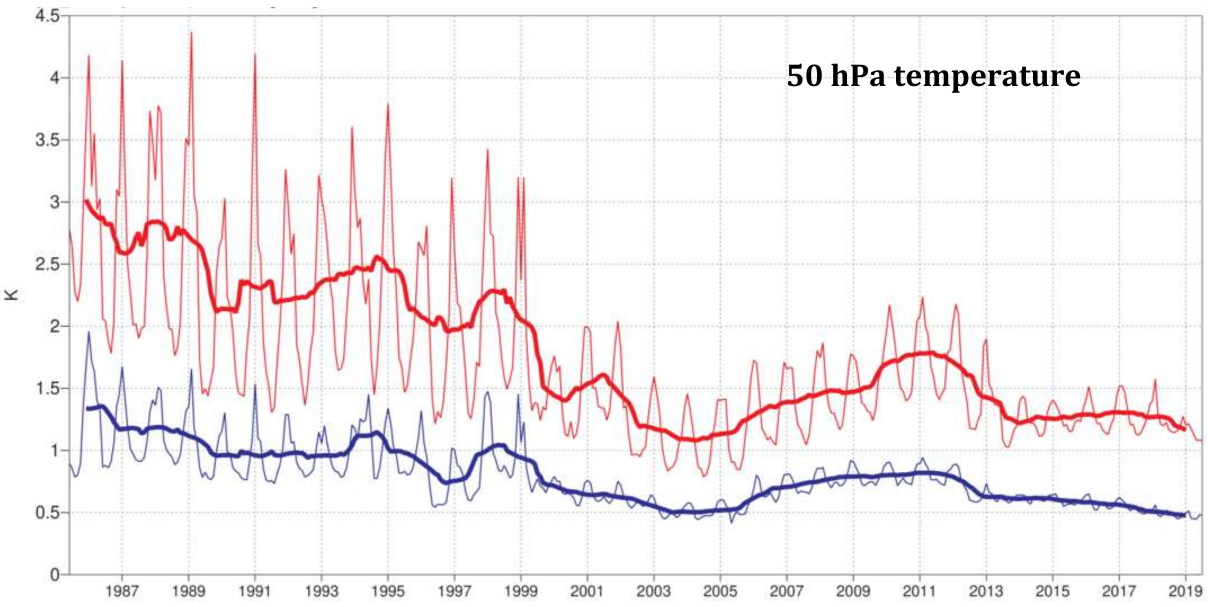


*Figure 6**:** Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average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



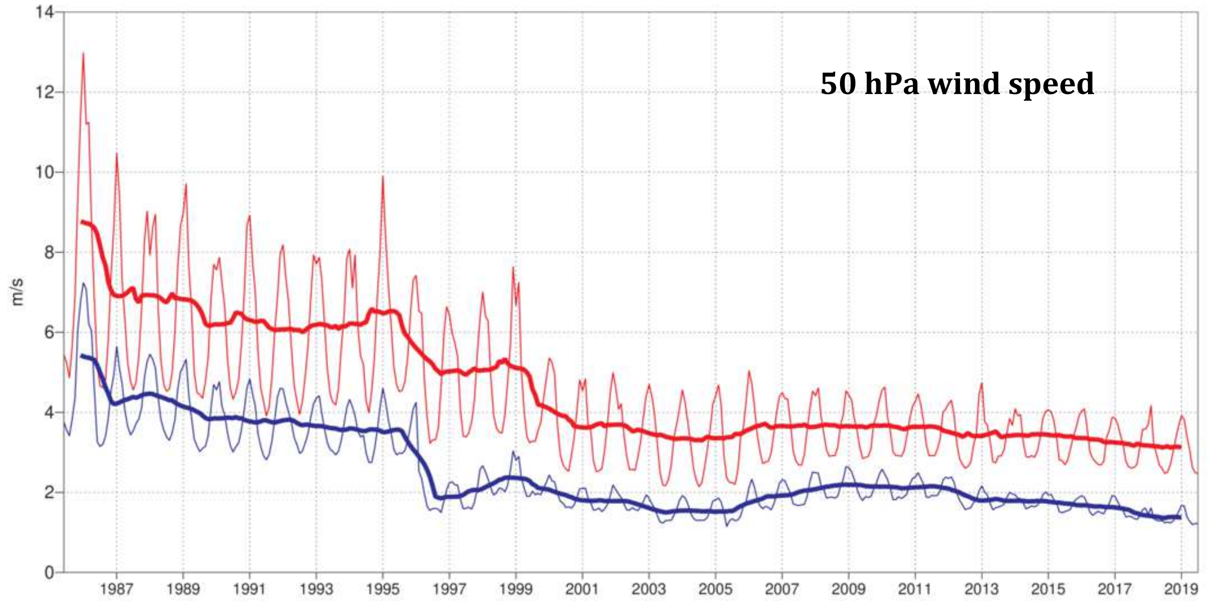
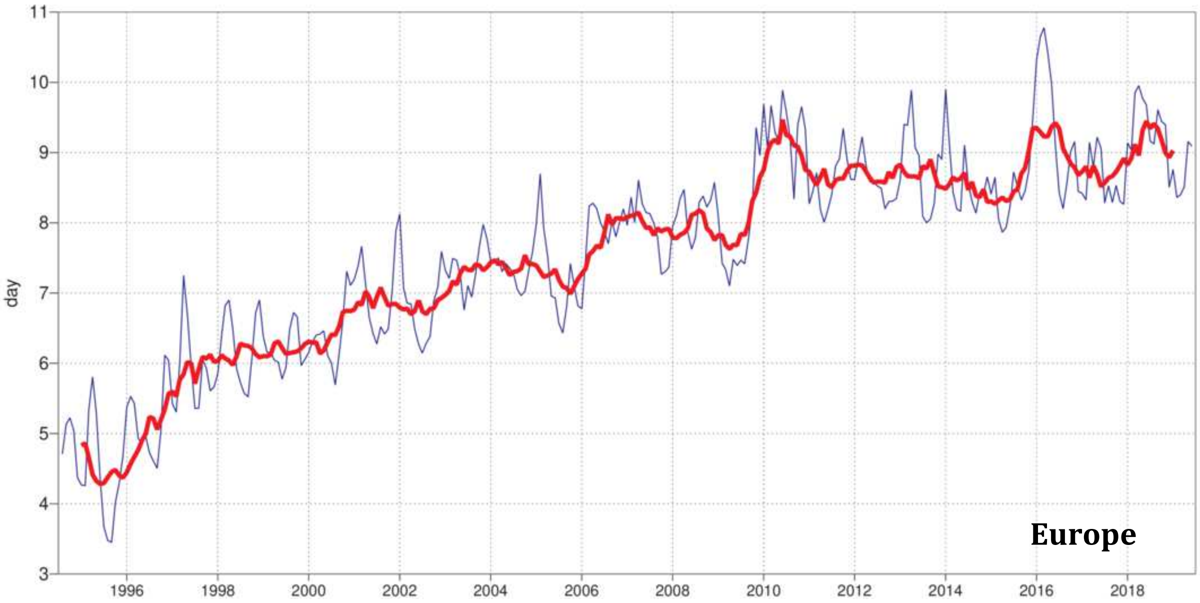


Figure 7: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also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



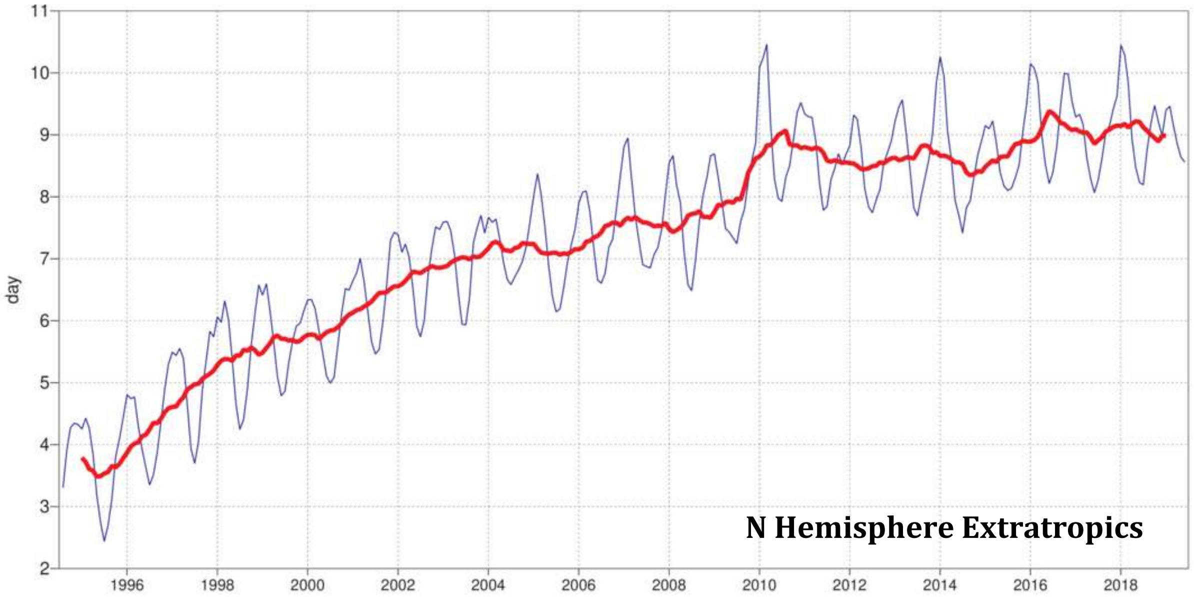


*Figure 8**:** Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

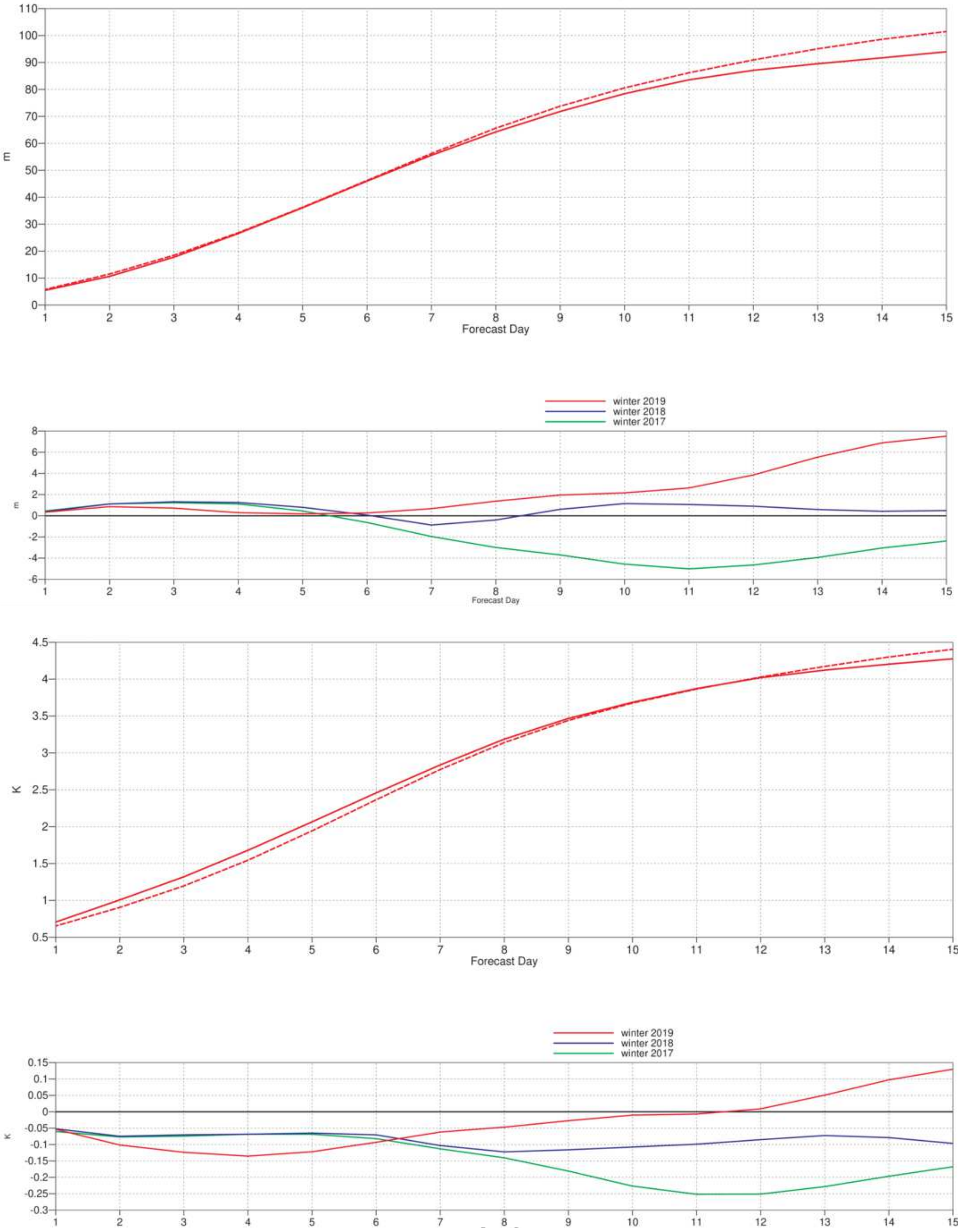


*Figure 9**:** Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2018–2019 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values 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.*

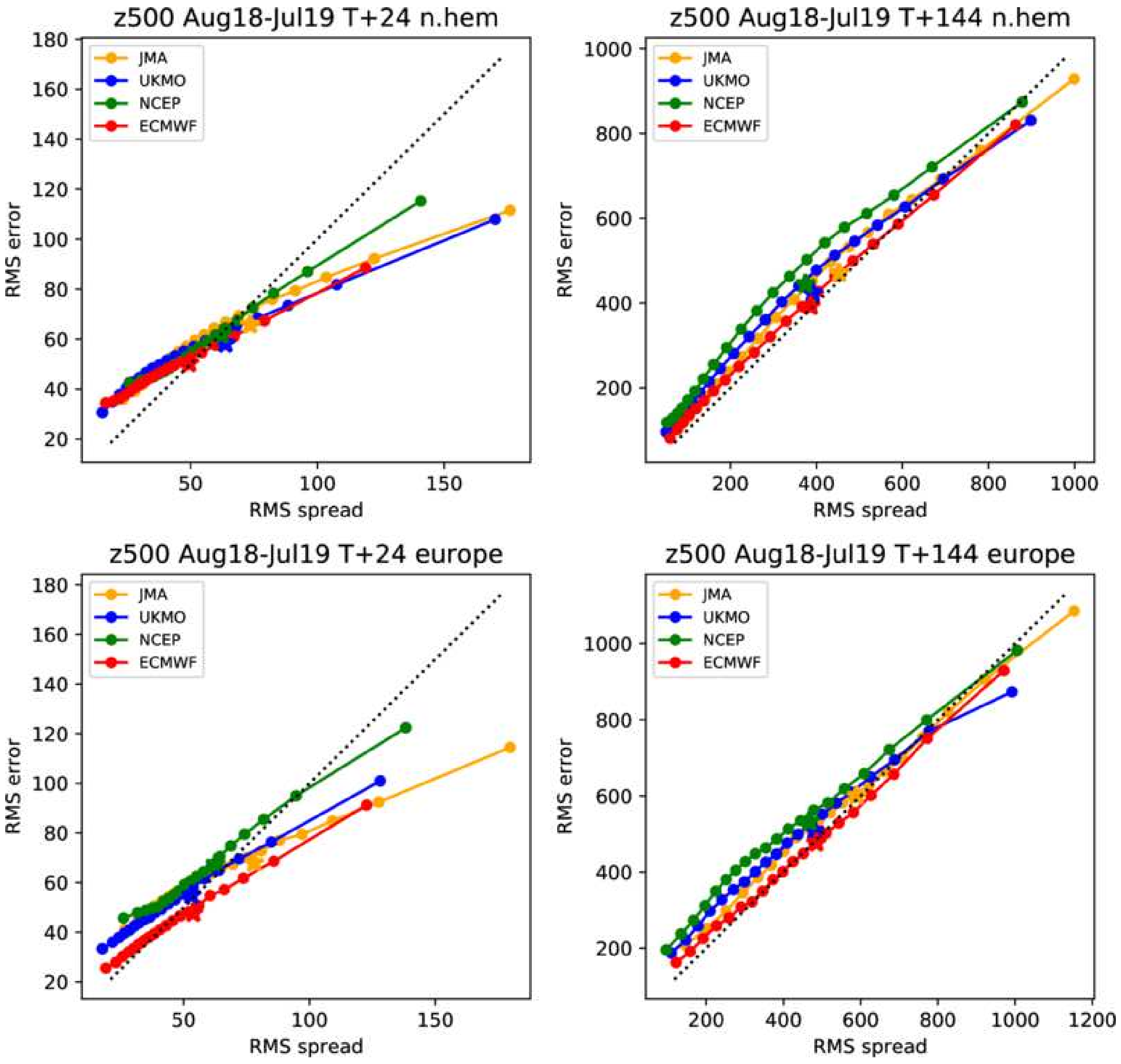


Figure 10: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2018–July 2019 in the northern hemisphere extra-tropics (top) and in Europe (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship. Due to random outages in the data supply, NCEP curves are based on a significantly reduced data set (50%).

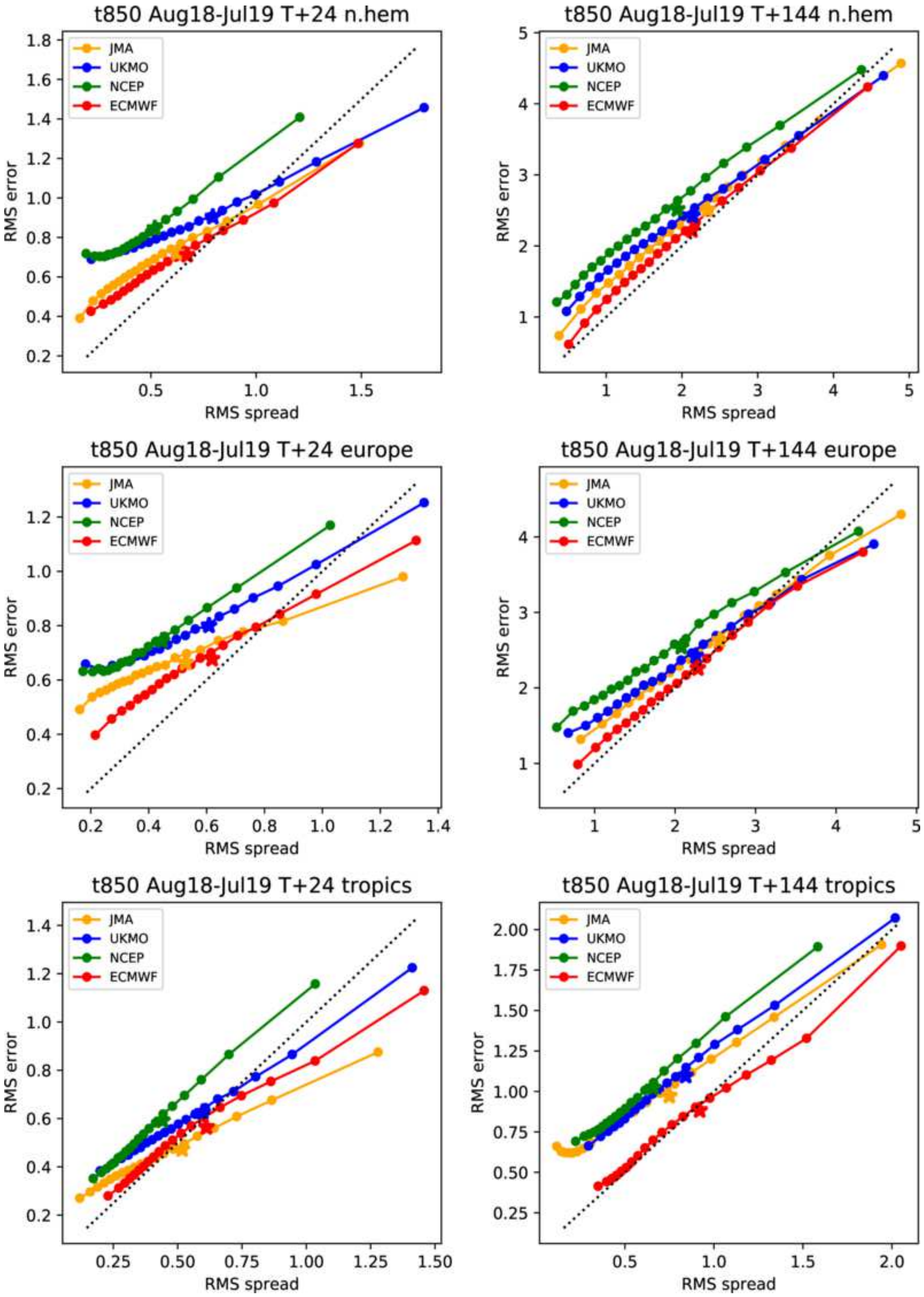
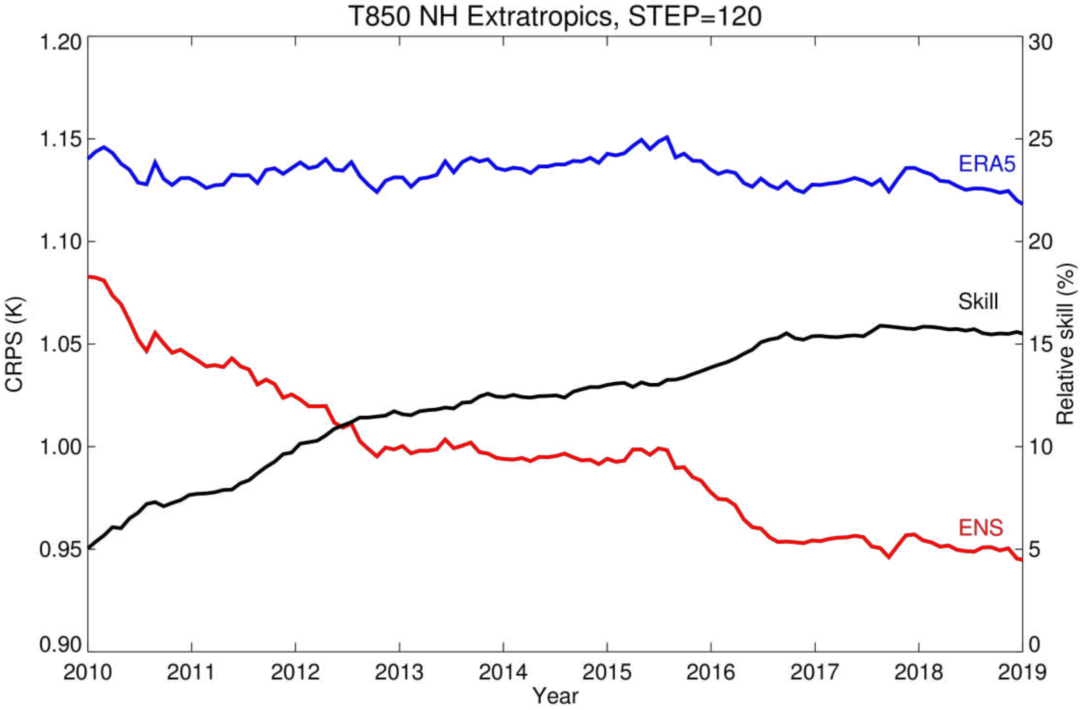


Figure 11: Ensemble spread reliability of different global models for 850 hPa temperature for the period August 2018–July 2019 in the northern hemisphere extra-tropics (top), Europe (centre), and the tropics (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship. Due to random outages in the data supply, NCEP curves are based on a reduced data set (60%).



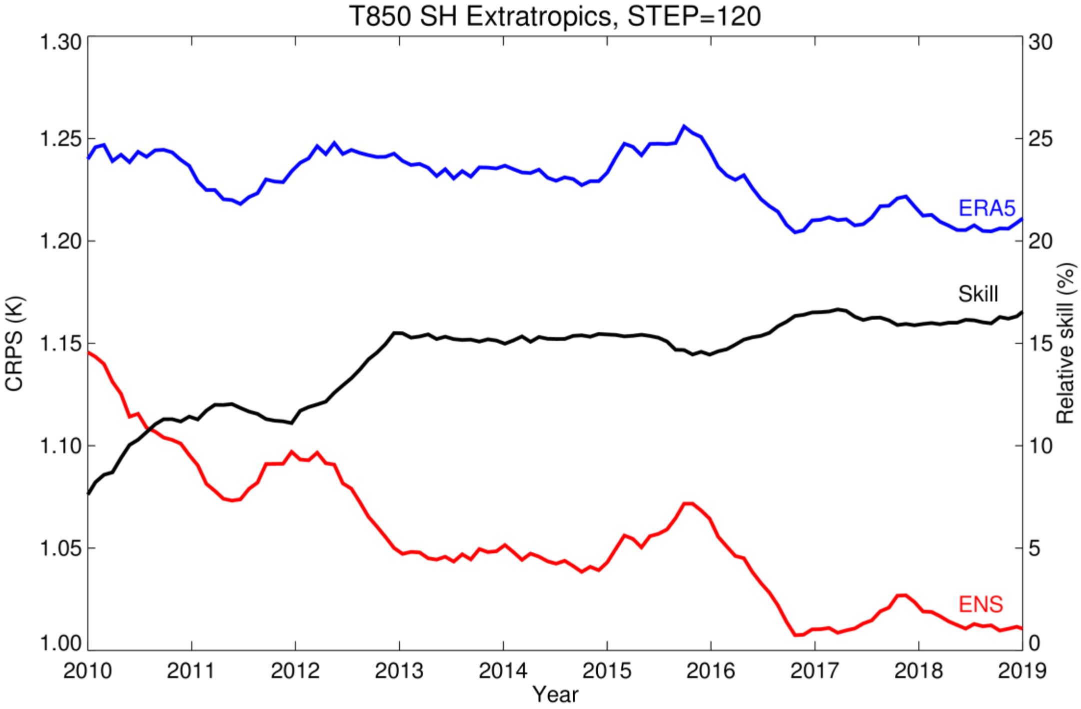


Figure 12: CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5, verified against analysis. Scores are shown for the ensemble forecast (red) and the dressed ERA5 forecast (blue). Black curves show the skill of the ENS relative to the dressed ERA5 forecast. 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.

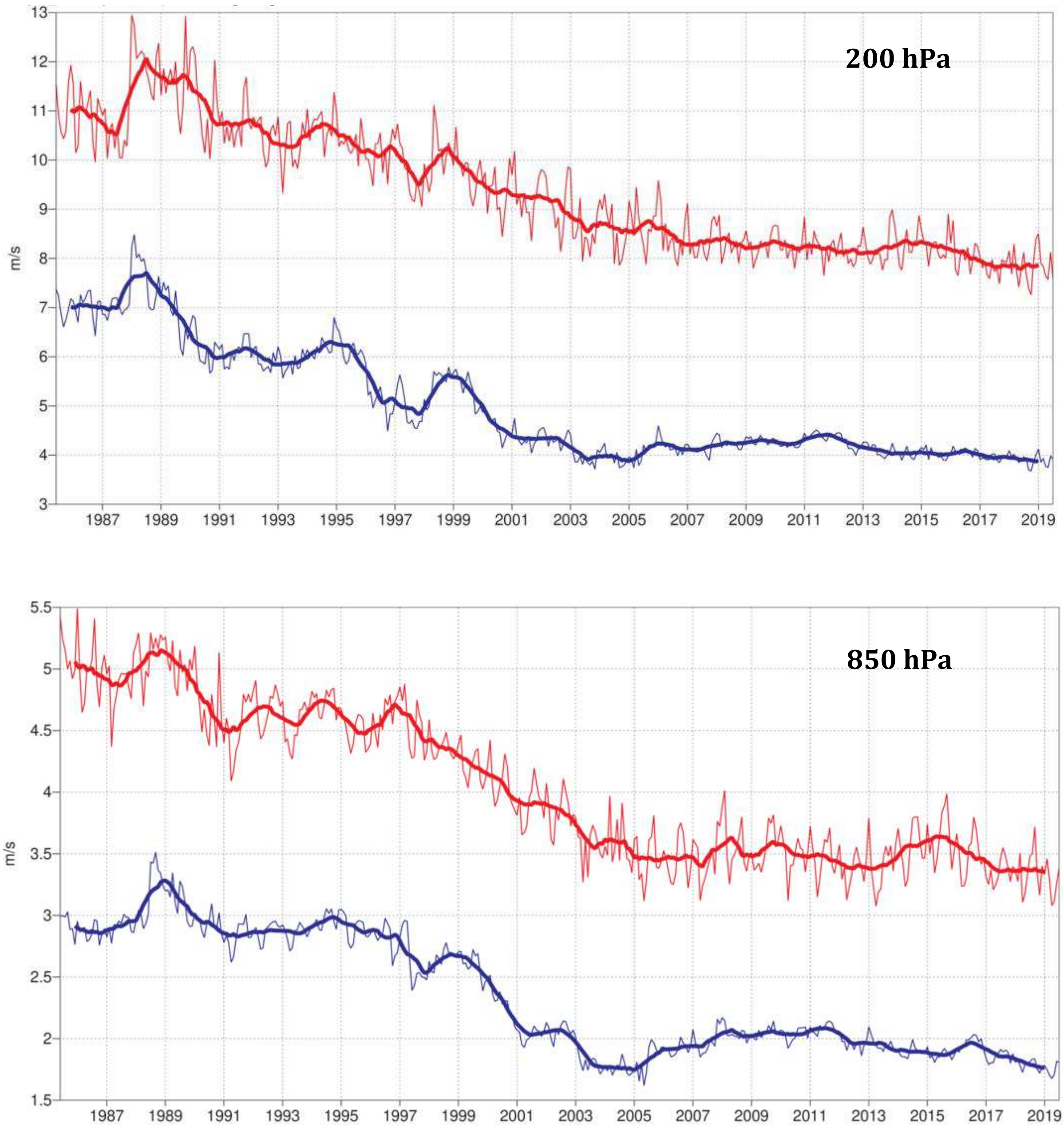


*Figure 13***:** *Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).*

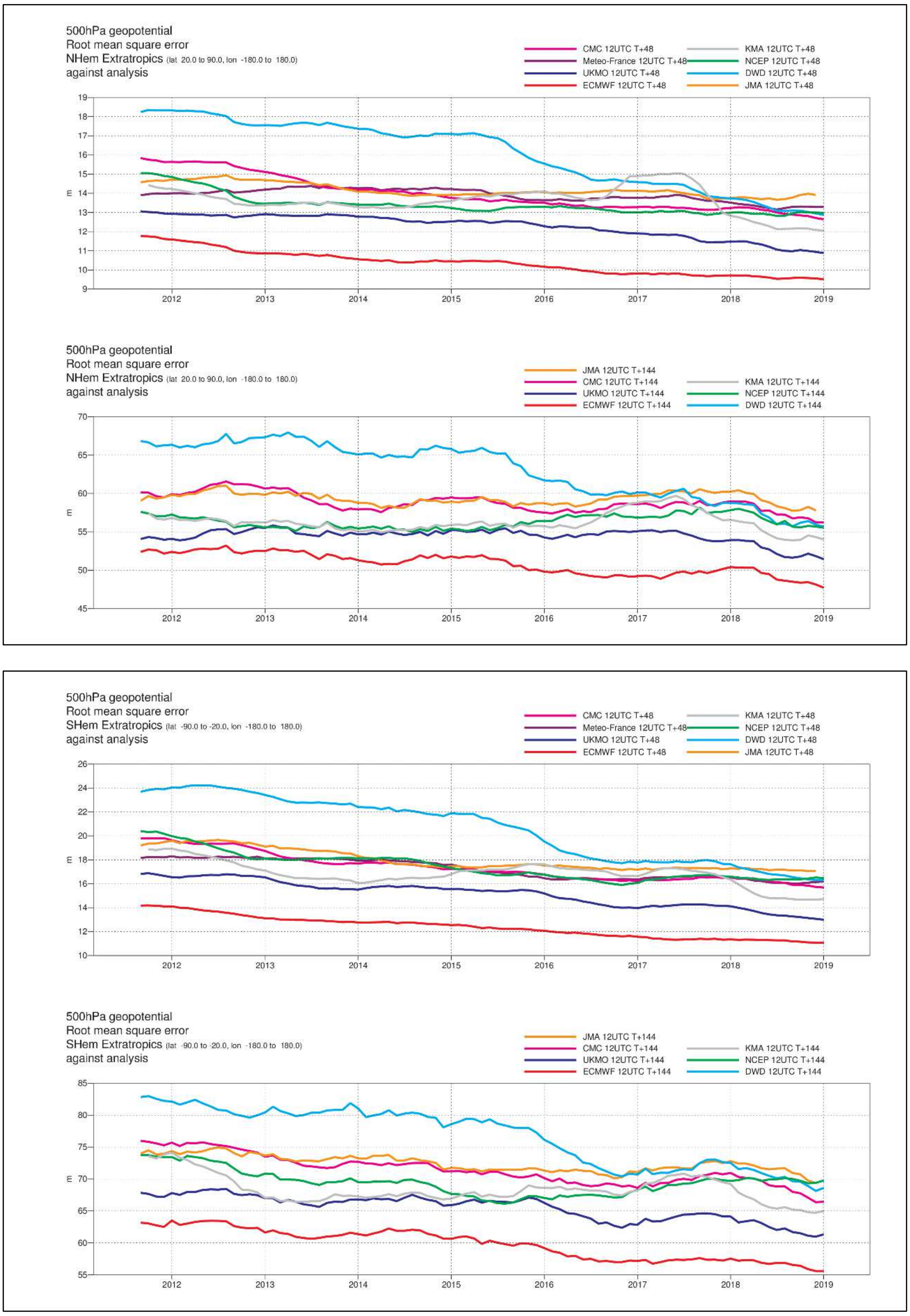
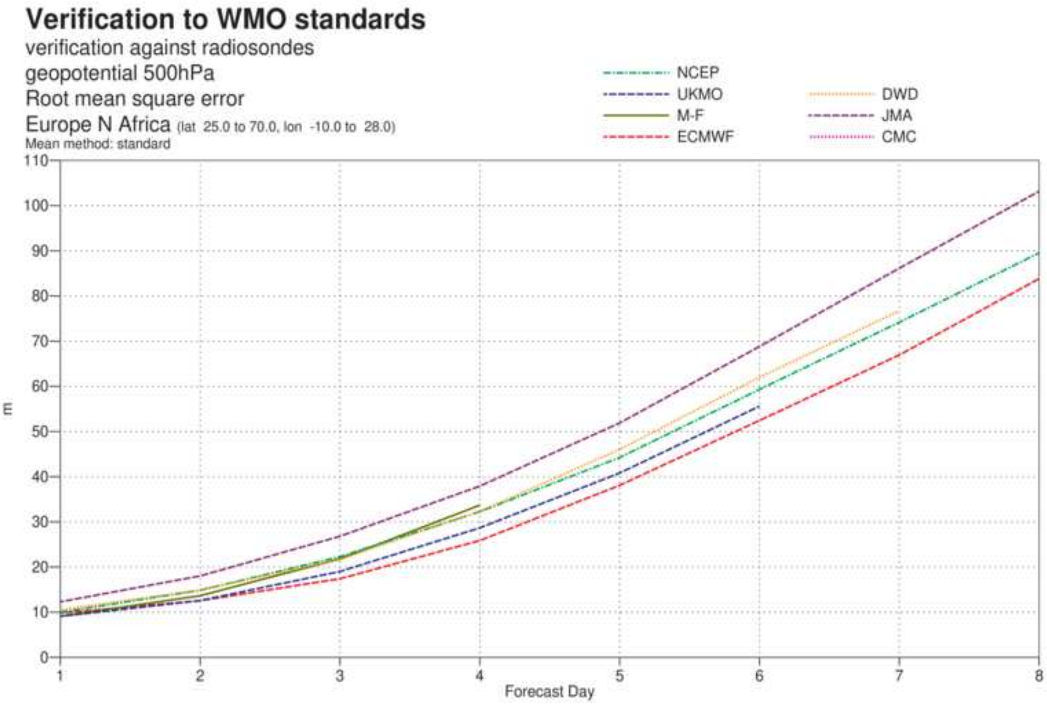


*Figure 14**:** WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top box) and southern (bottom box) extratropics. In each box the upper plot shows the two-day forecast error and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. 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, Météo France, DWD = Deutscher Wetterdienst.*



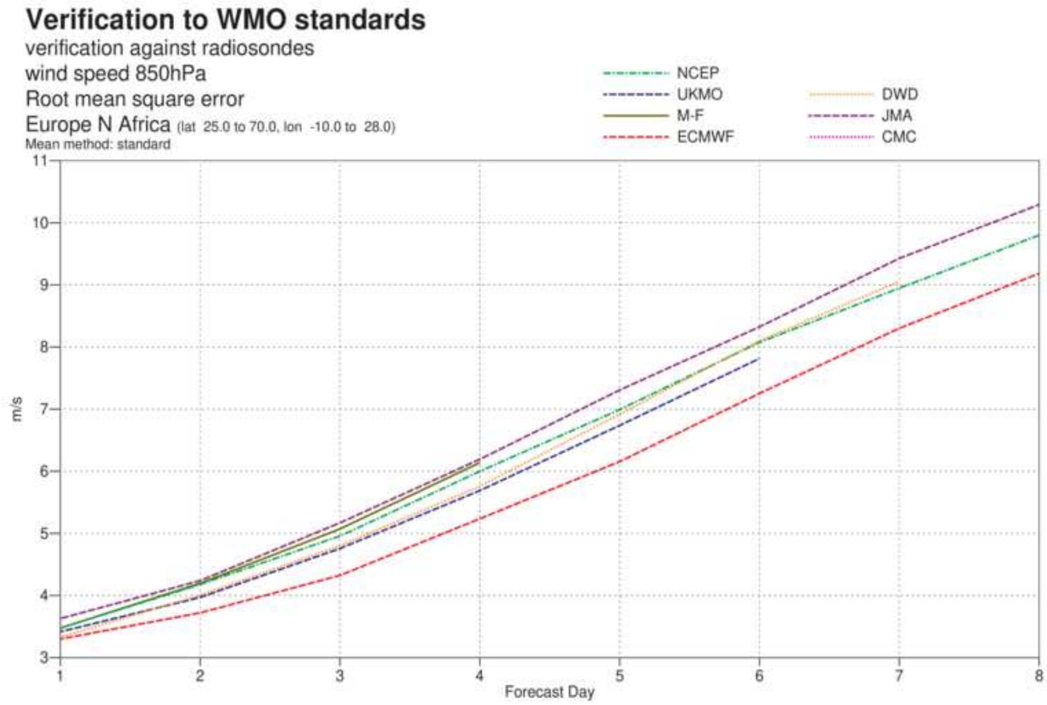
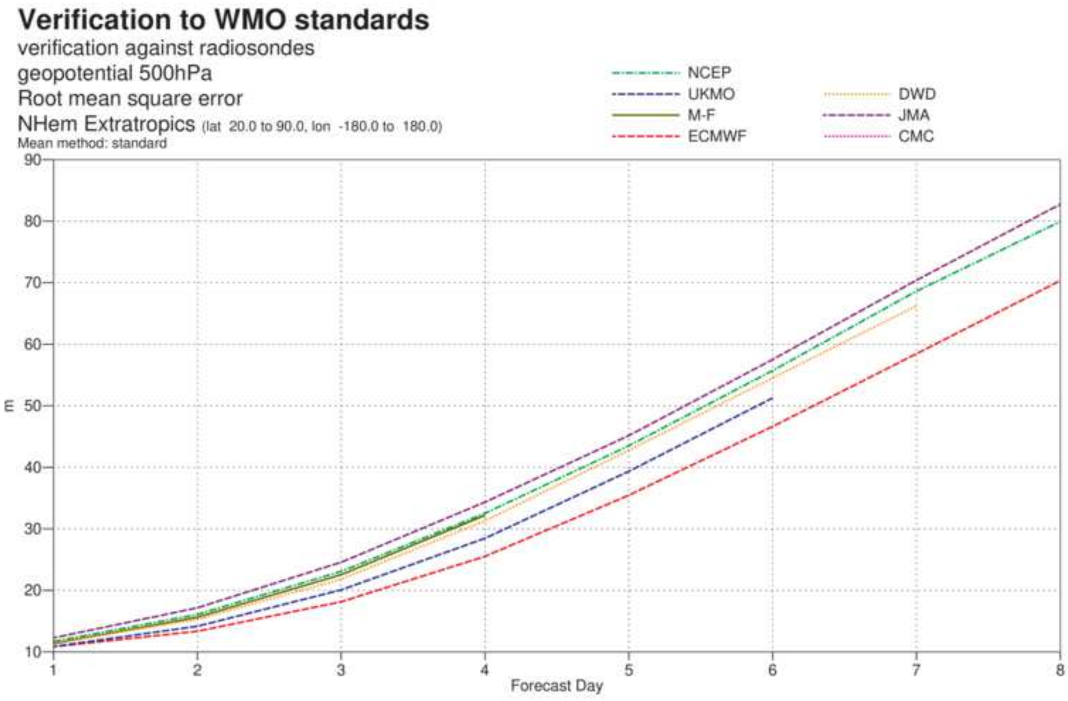


*Figure 15***:** *WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2018–July 2019) of forecast runs initiated at 12 UTC.*



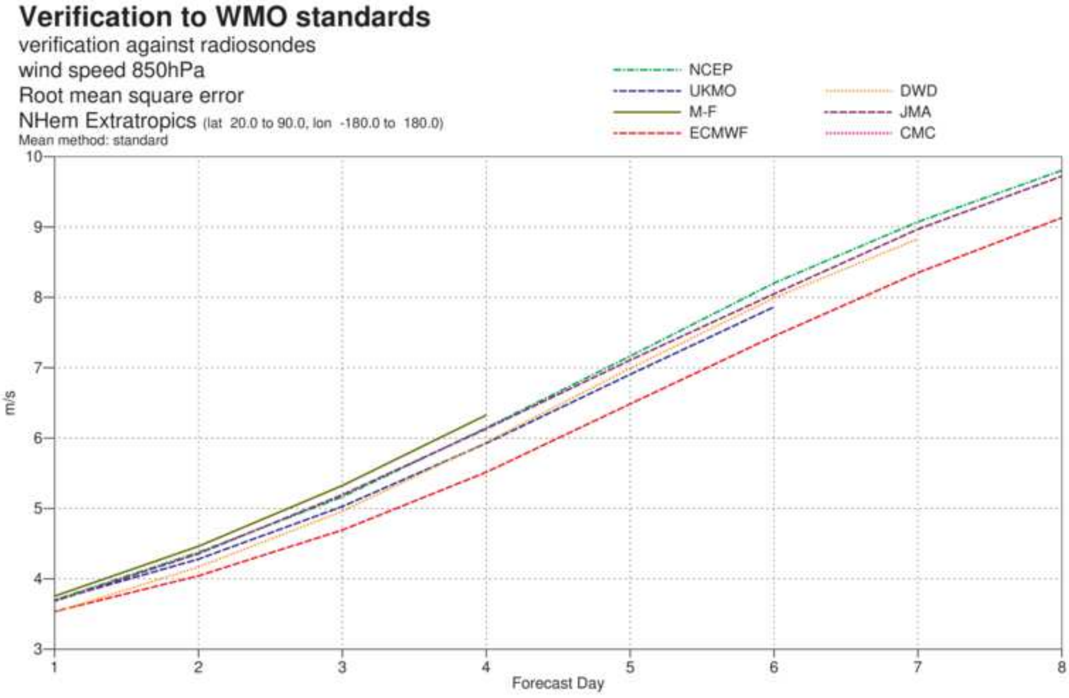


Figure 16: As Figure 15 for the northern hemisphere extratropics.

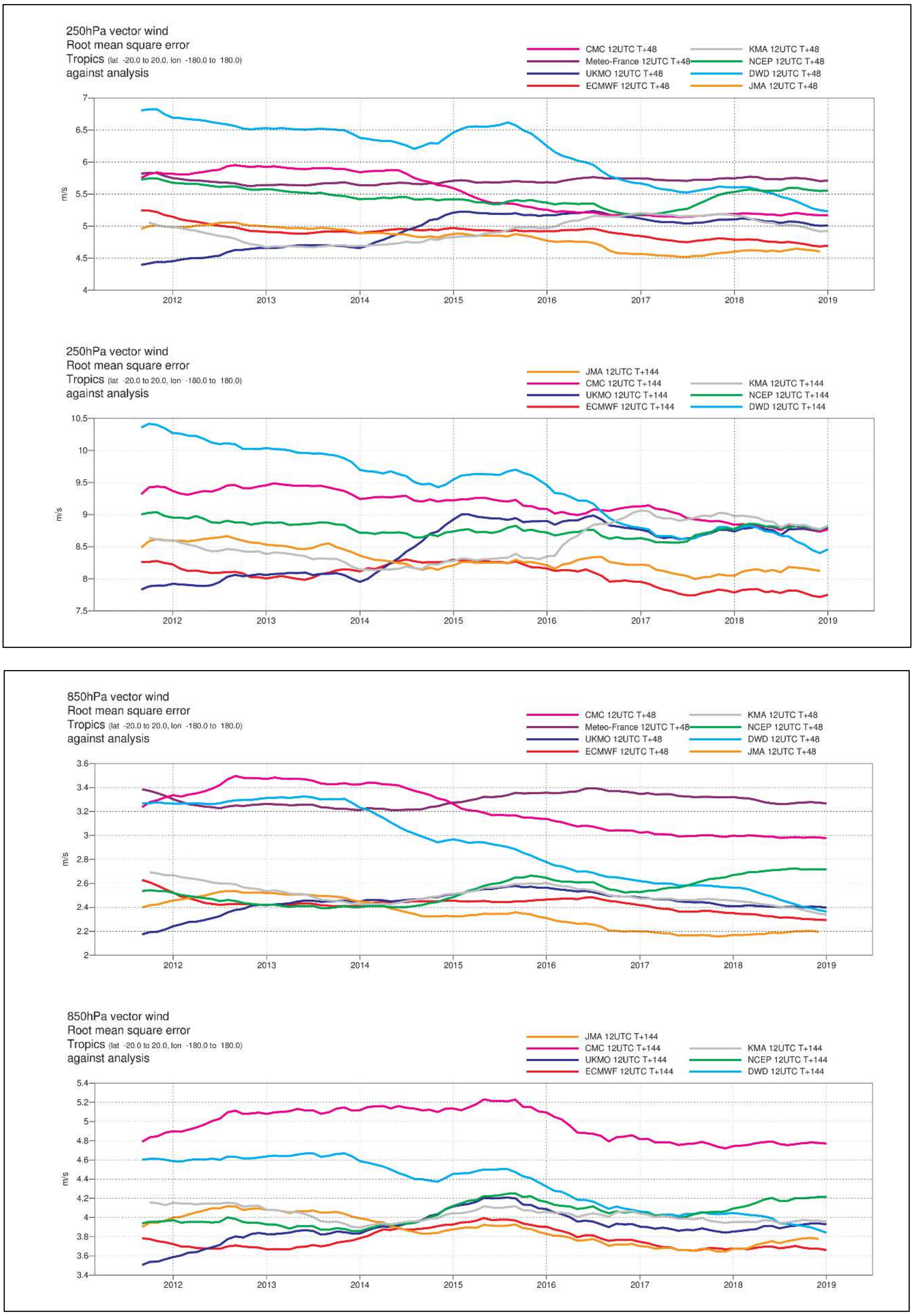


Figure 17: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top box) and 850 hPa (bottom box). In each box the upper plot shows the two-day forecast error and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis.

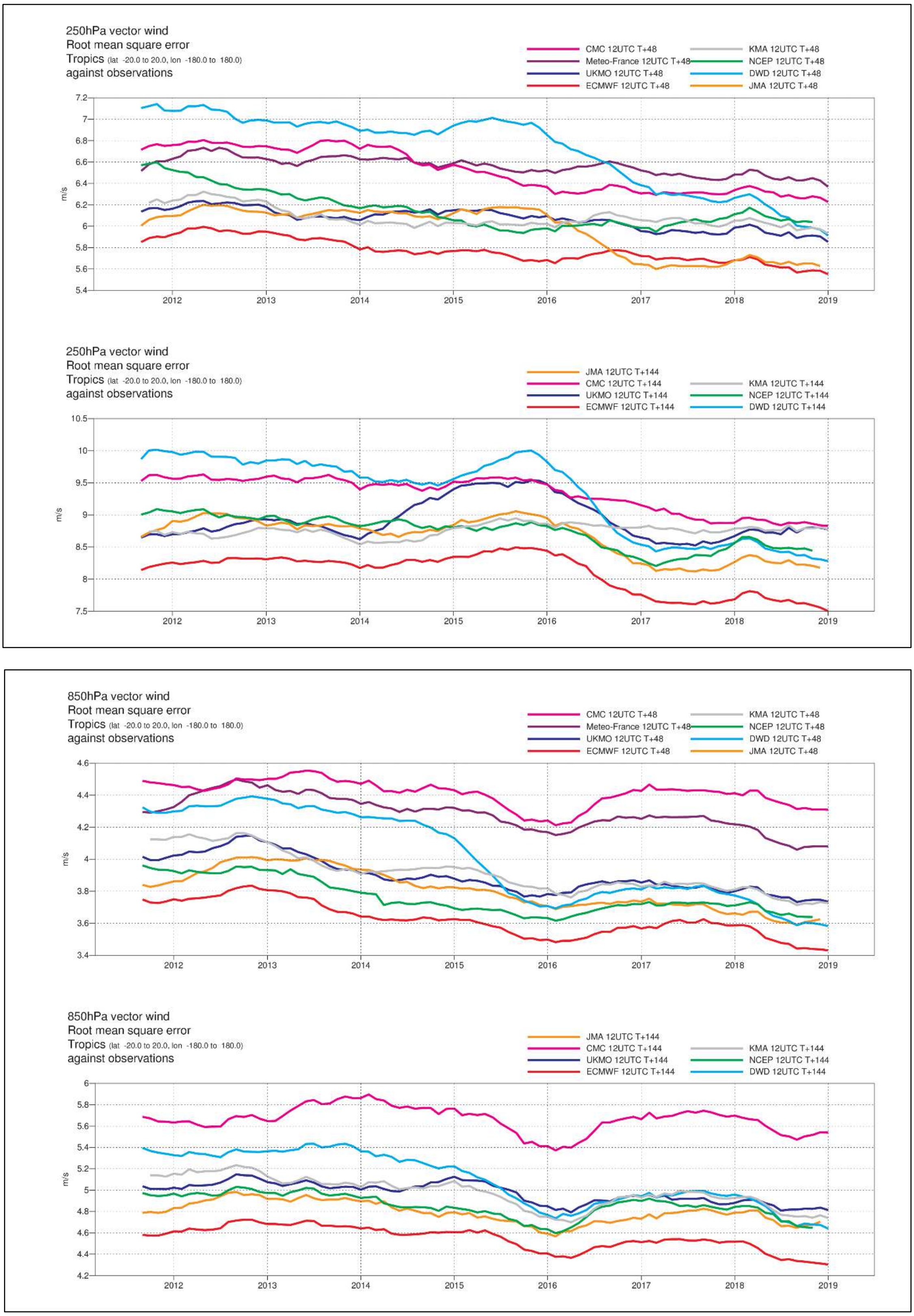


Figure 18: As Figure 17 for verification against radiosonde observations.

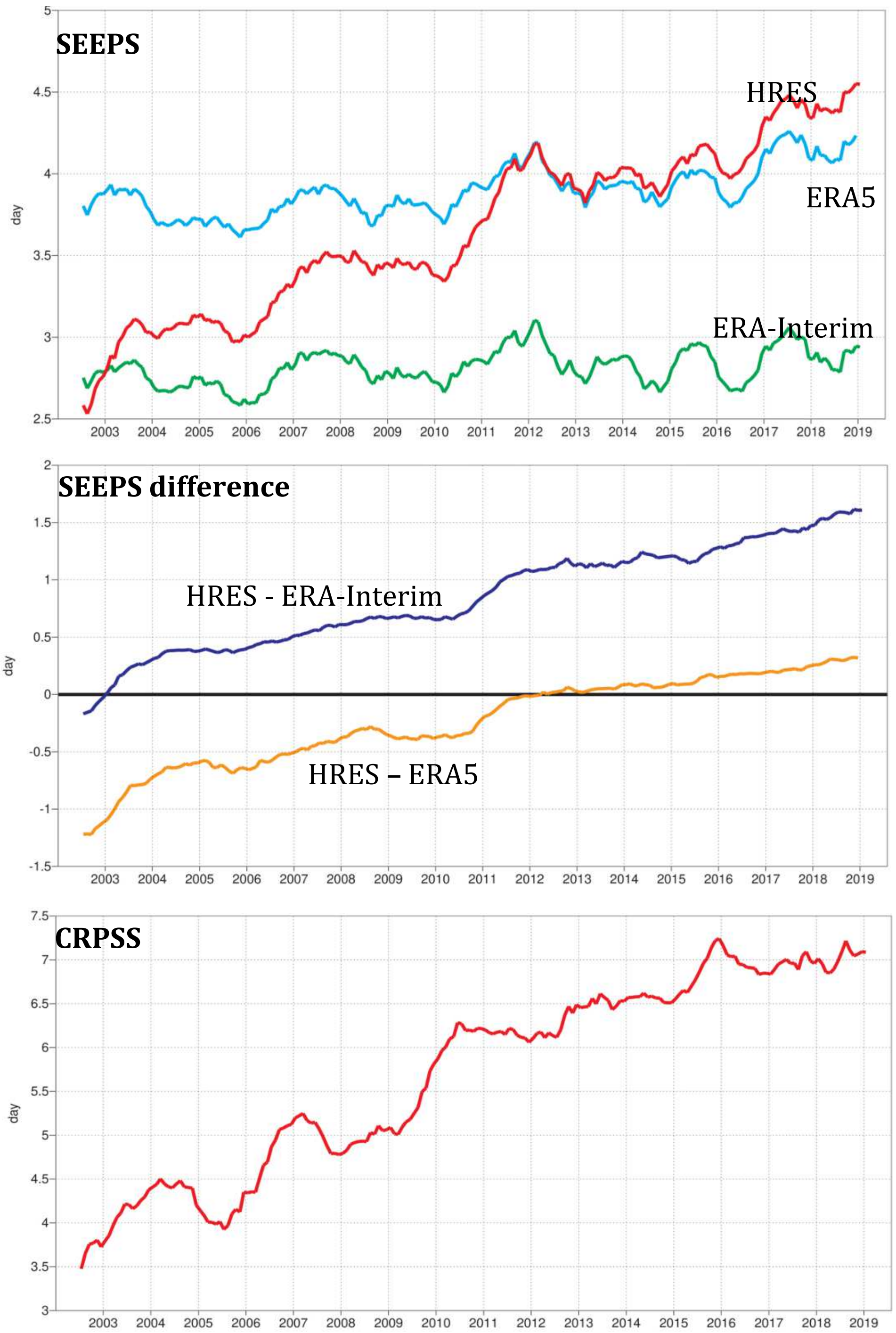
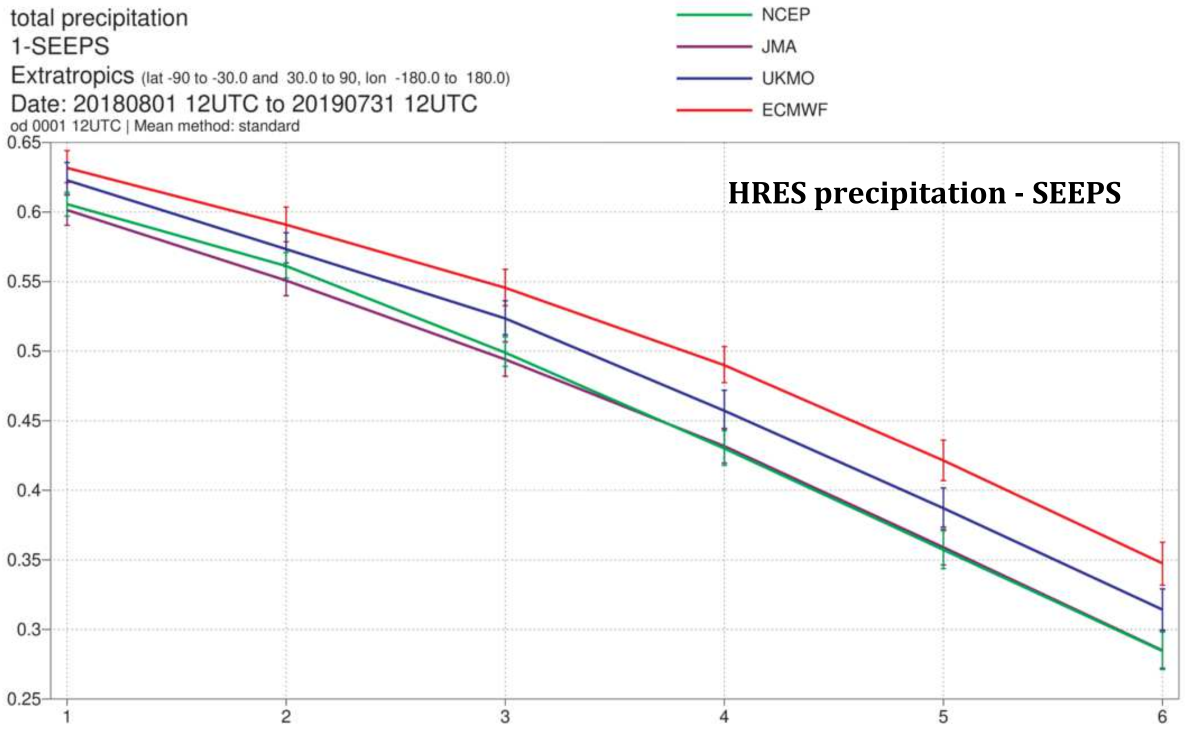


Figure 19: Supplementary headline scores (red) for deterministic (top, centre) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number of days for which the centred 12-month 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. The green and light blue curves in the top panel show the deterministic headline score for ERA-Interim and ERA5, respectively. The centre panel shows the difference between the operational forecast and ERA-Interim (blue), and between the operational forecast and ERA5 (yellow).



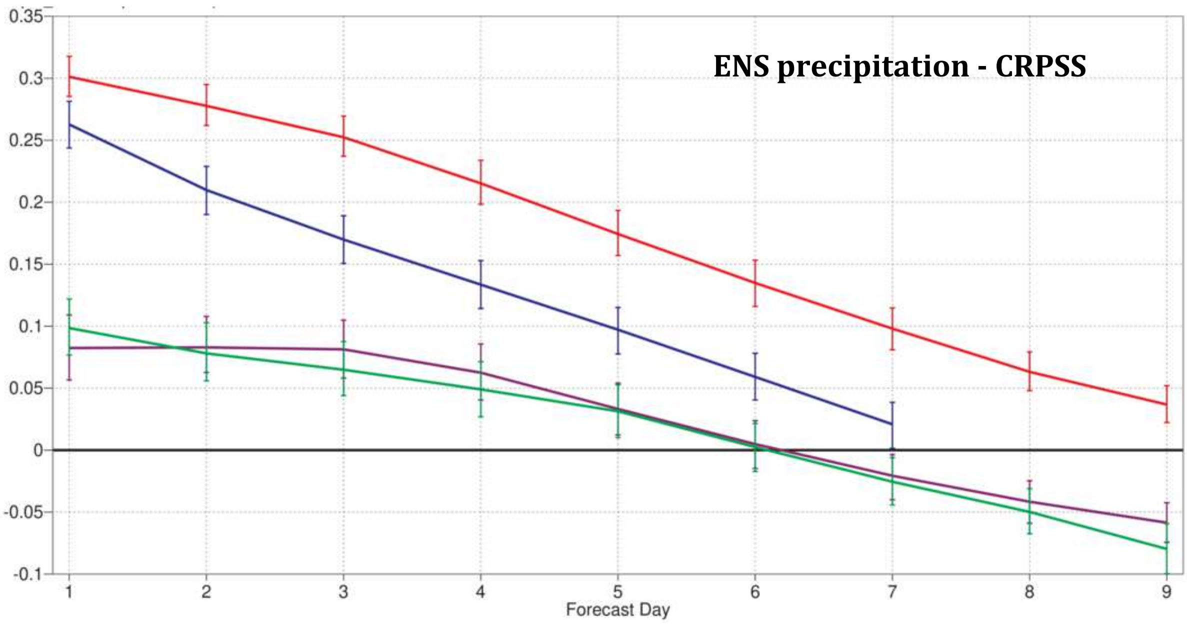


*Figure 20**:** Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure **19**. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2018–July 2019. Bars indicate 95% confidence intervals.*

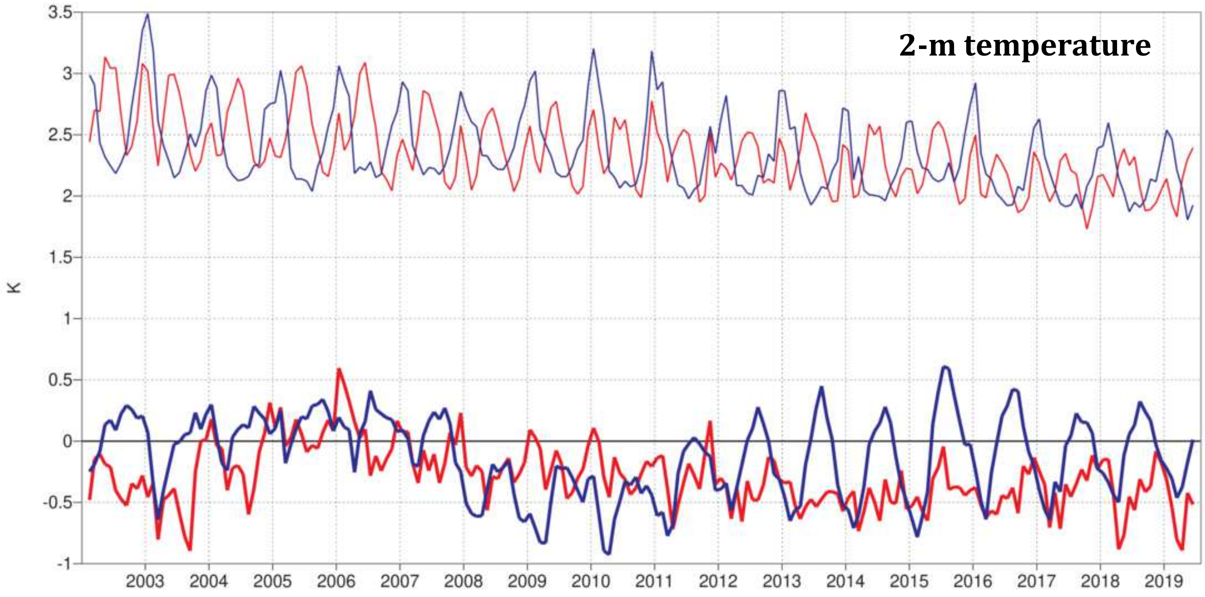


*Figure 21**:** Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.*

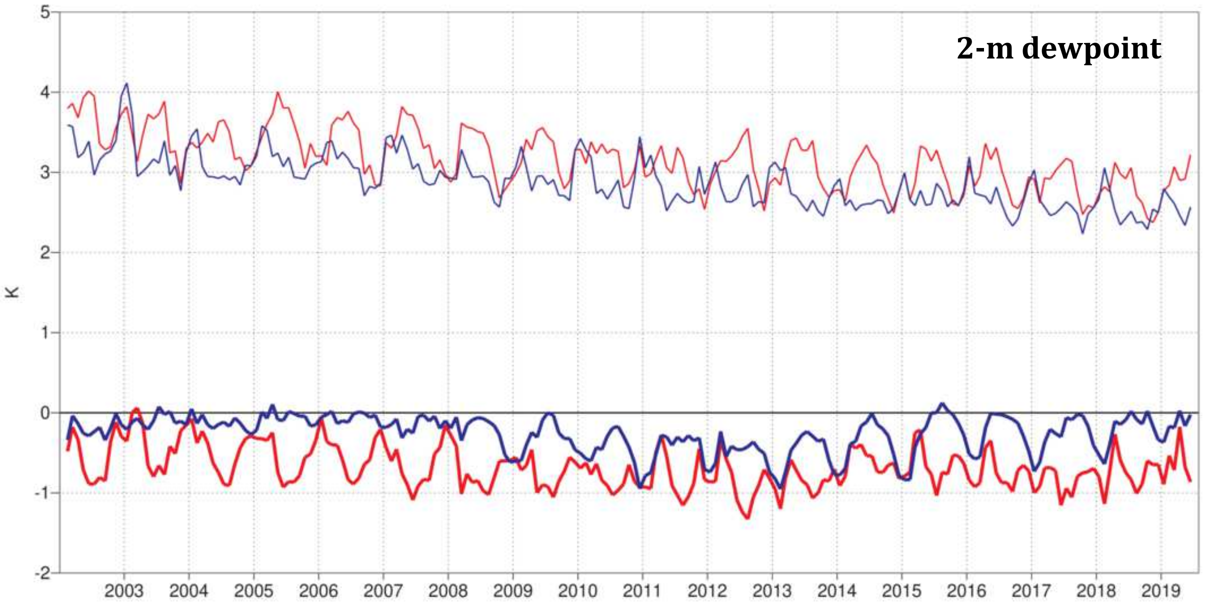


*Figure 22**:** Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.*

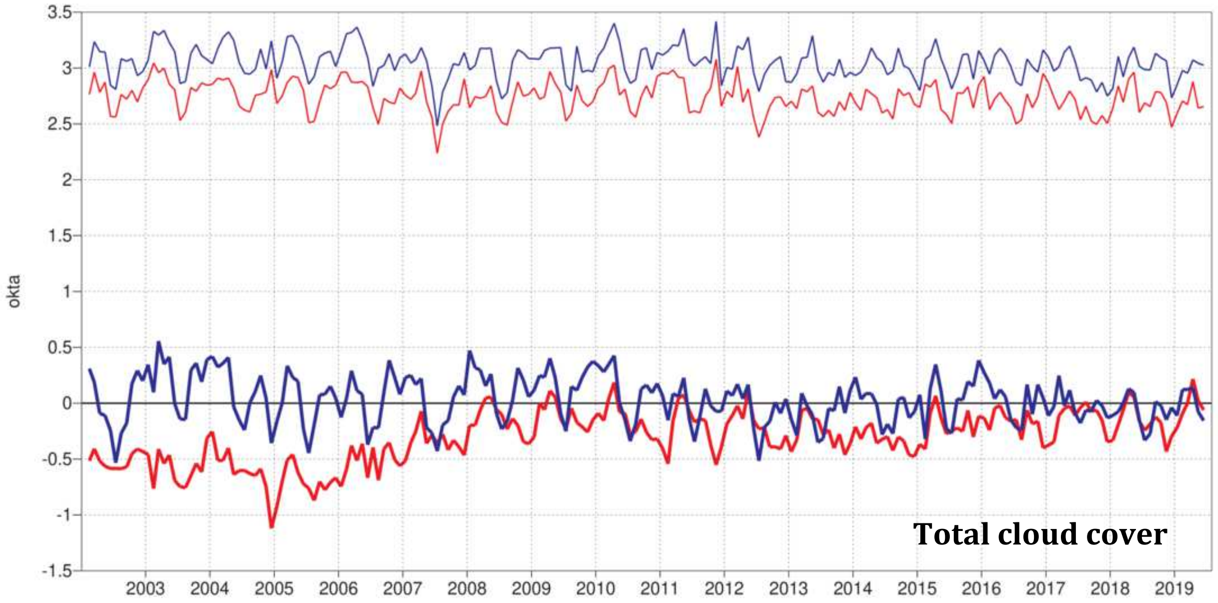


*Figure 23**:** Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.*

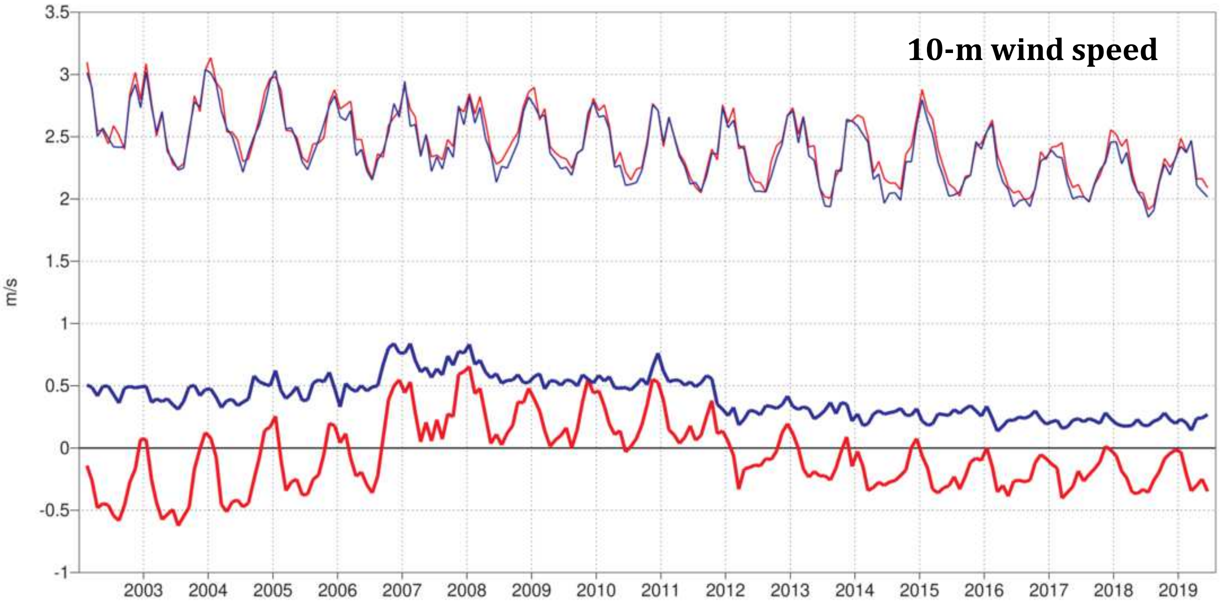


*Figure 24**:** Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.*

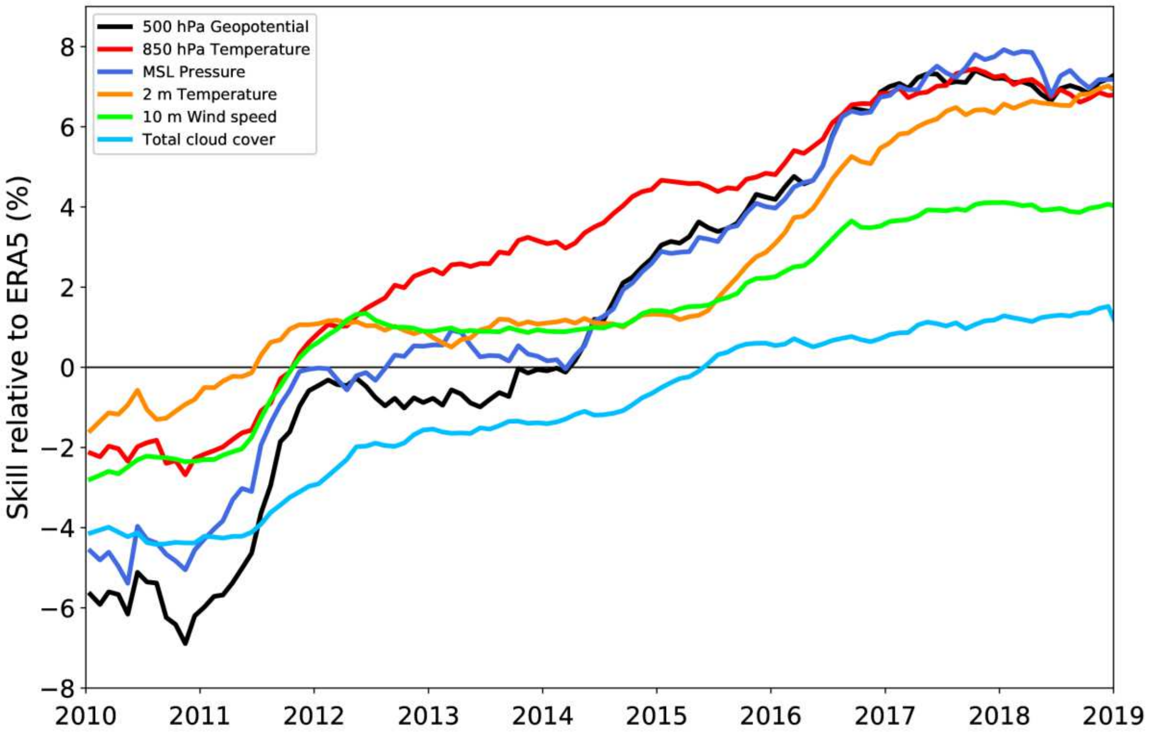


Figure 25: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA5. Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), and mean sea level pressure (MSLP), using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC).

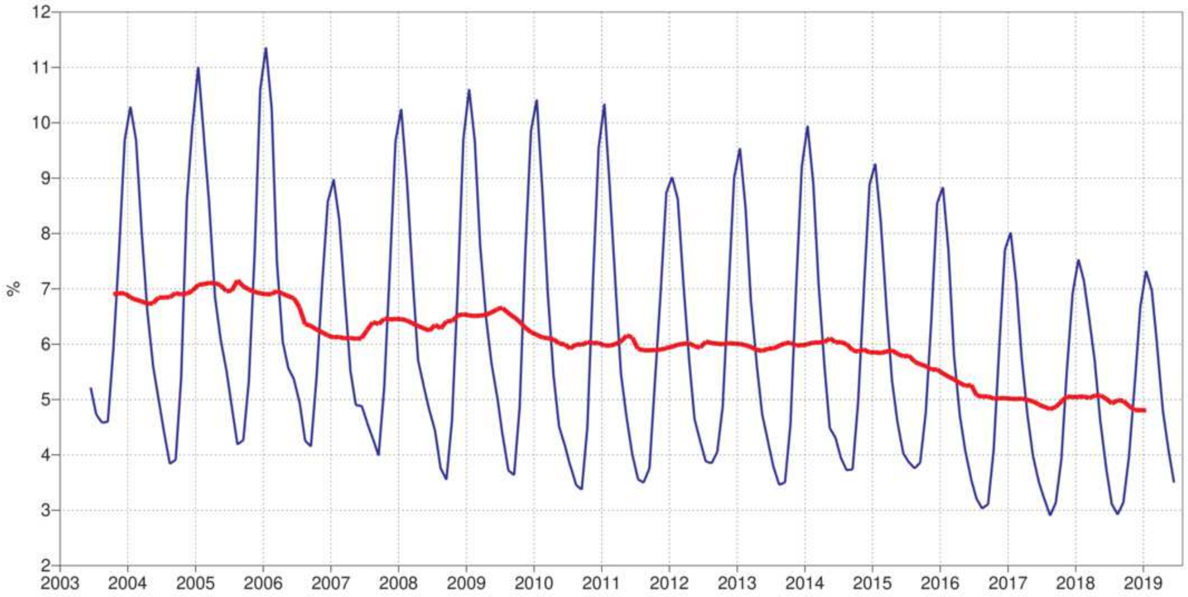


Figure 26: Evolution of the fraction of large 2m temperature errors (CRPS>5K) in the ENS at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

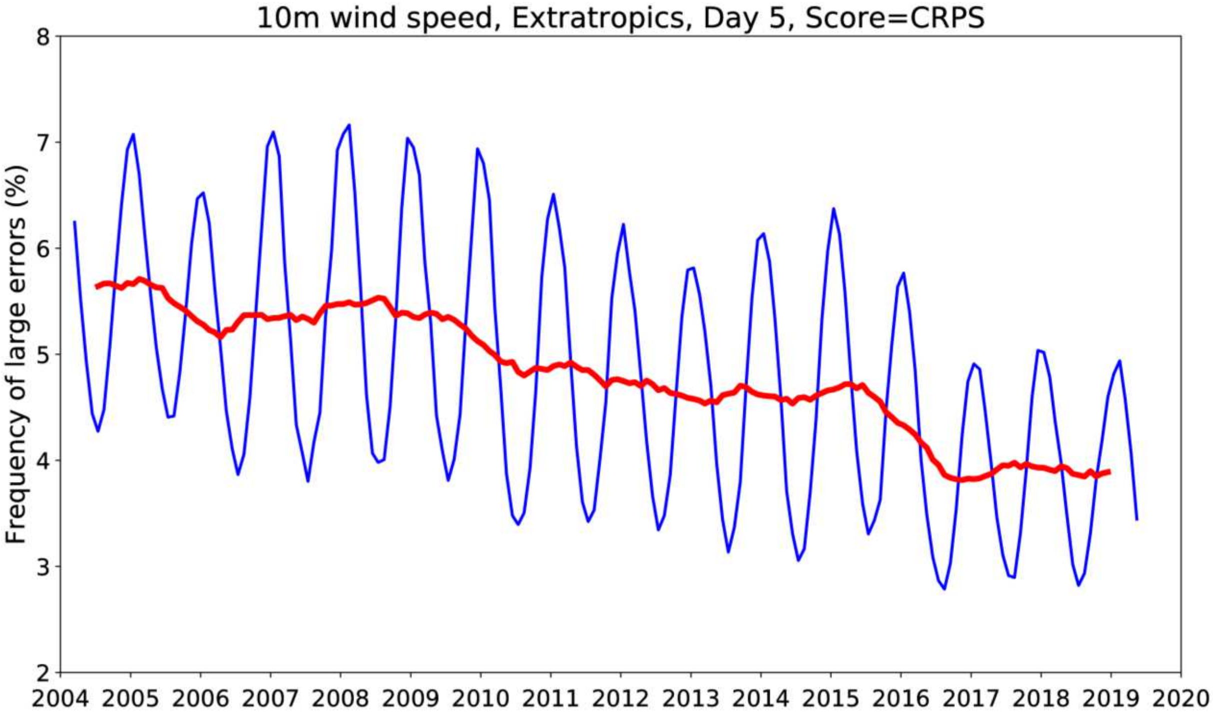


Figure 27: Evolution of the fraction of large 10m wind speed errors (CRPS>4m/s) in the ENS at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

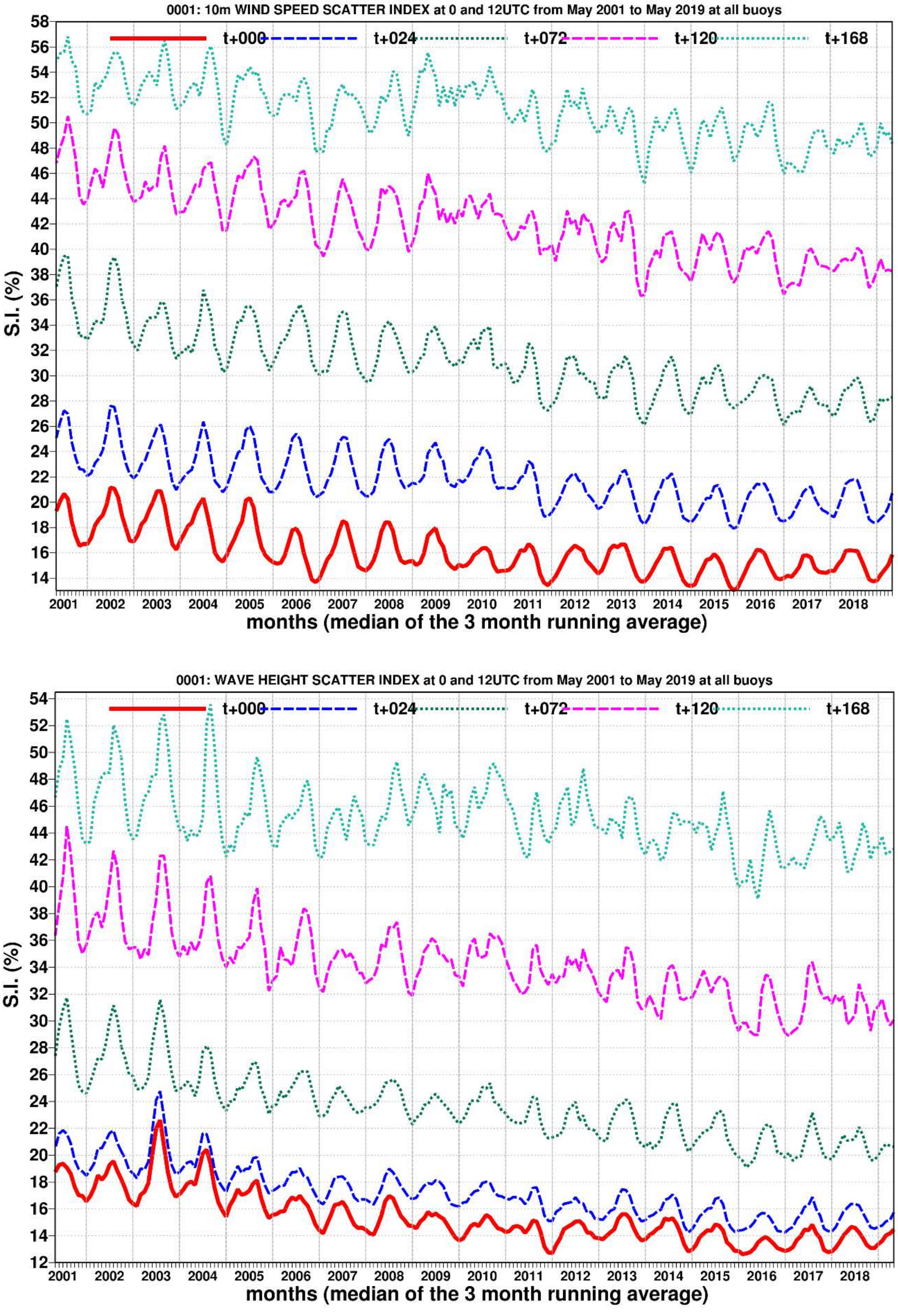
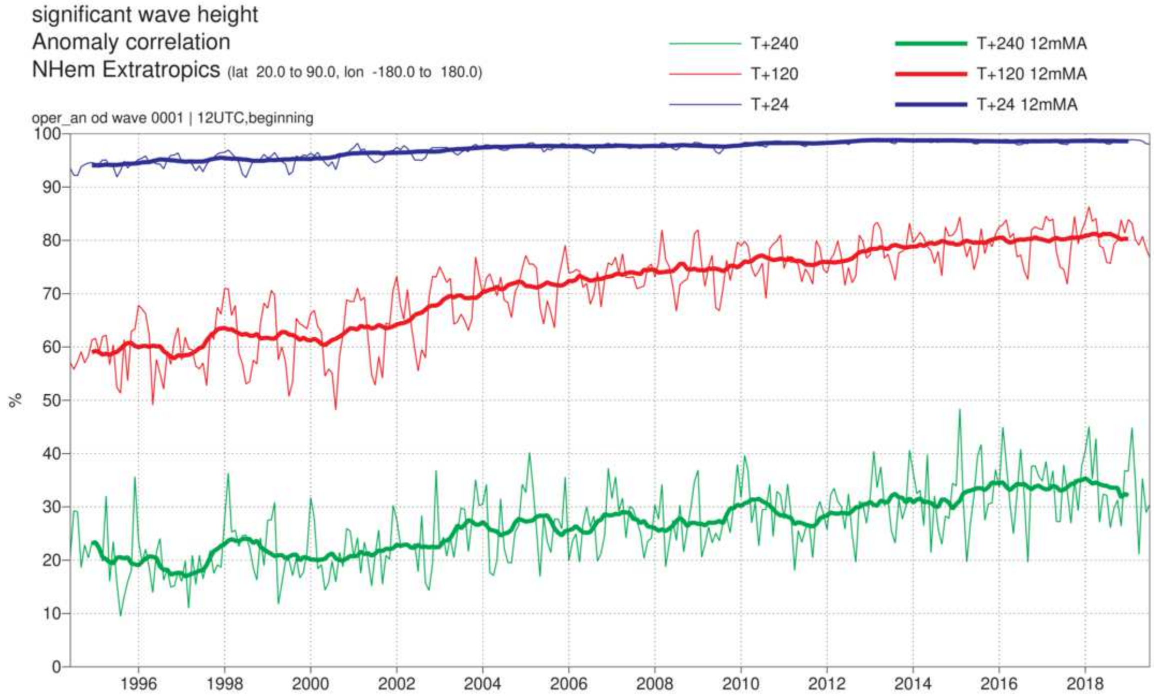


Figure 28: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running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



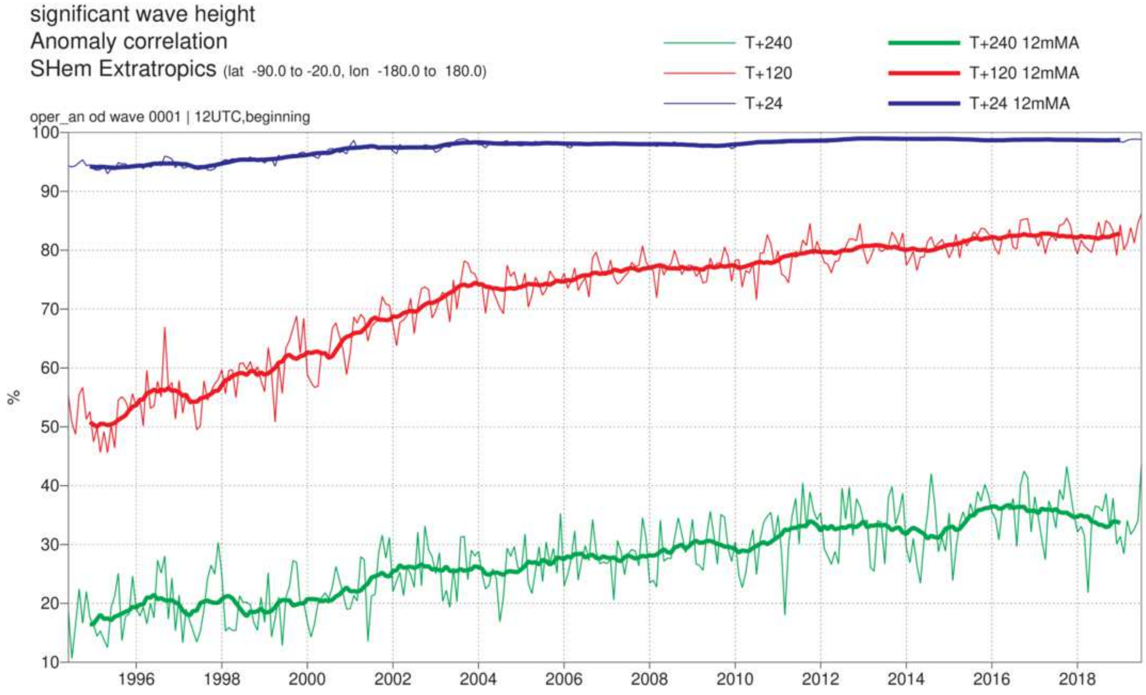


Figure 29: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

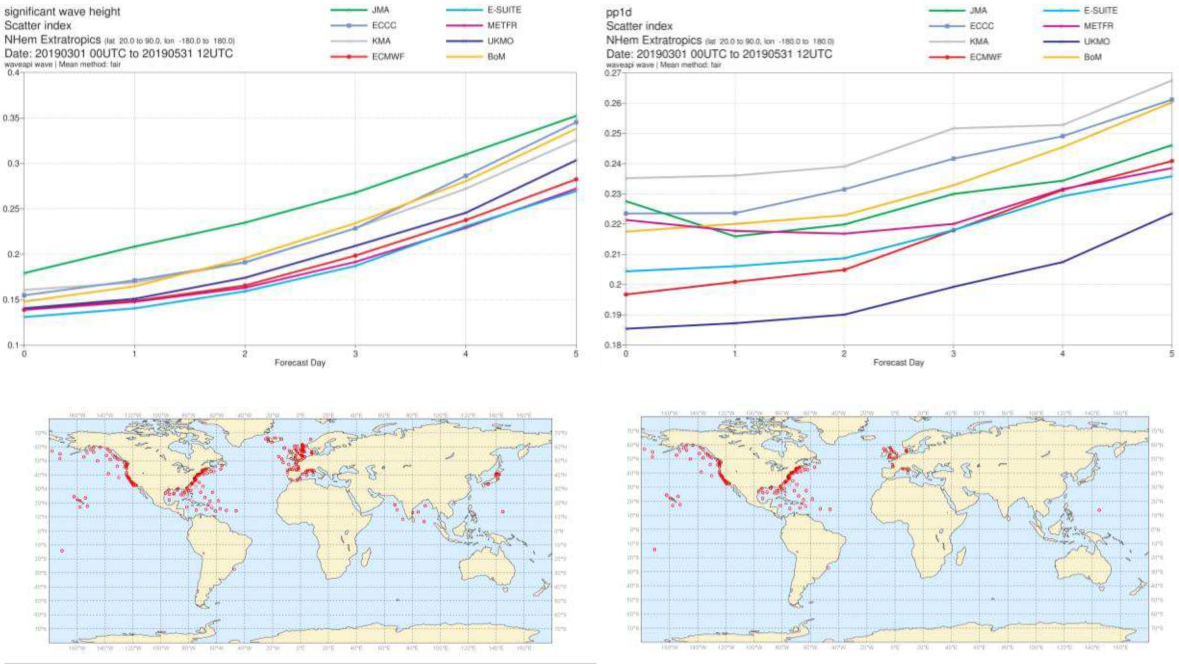


Figure 30: Verification of forecasts of wave height and peak wave period (upper panels) using observations from wave buoys (lower panels). The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 3-month period March-May 2019. UKMO: Met Office, UK; METFR: Météo-France; BoM: Bureau of Meteorology, Australia; JMA: Japan Meteorological Agency; KMA: Korea Meteorological Administration; ECCC: Environment and Climate Change Canada; E-SUITE: ECMWF cycle 46r1. Providing model intercomparison plots is part of ECMWF's verification activities as the WMO Lead Centre for Wave Forecast Verification.

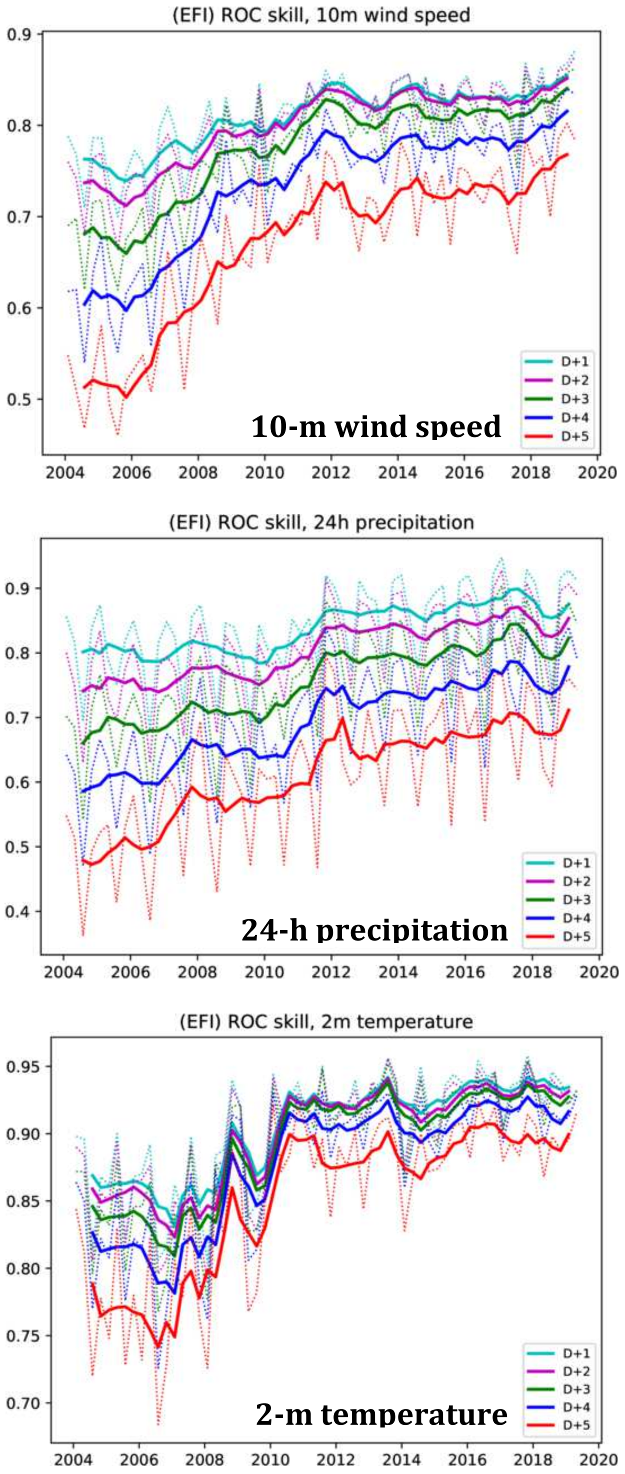


Figure 31: Verification of Extreme Forecast Index (EFI) against analysis. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 5 (24-hour period 96–120 hours ahead); skill at day 4 (blue line) is the supplementary headline score; an extreme event is taken as an observation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

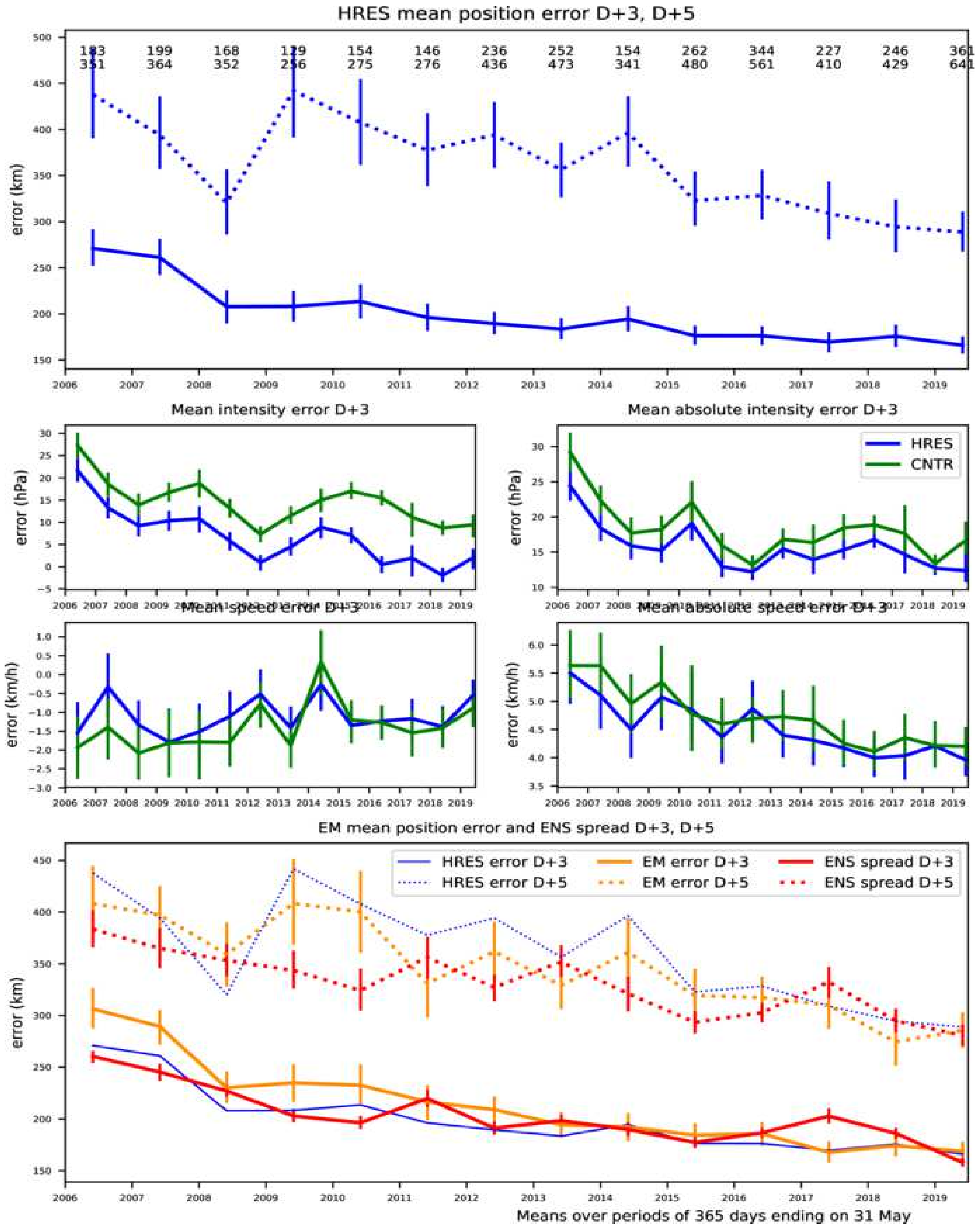


Figure 32: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary 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).

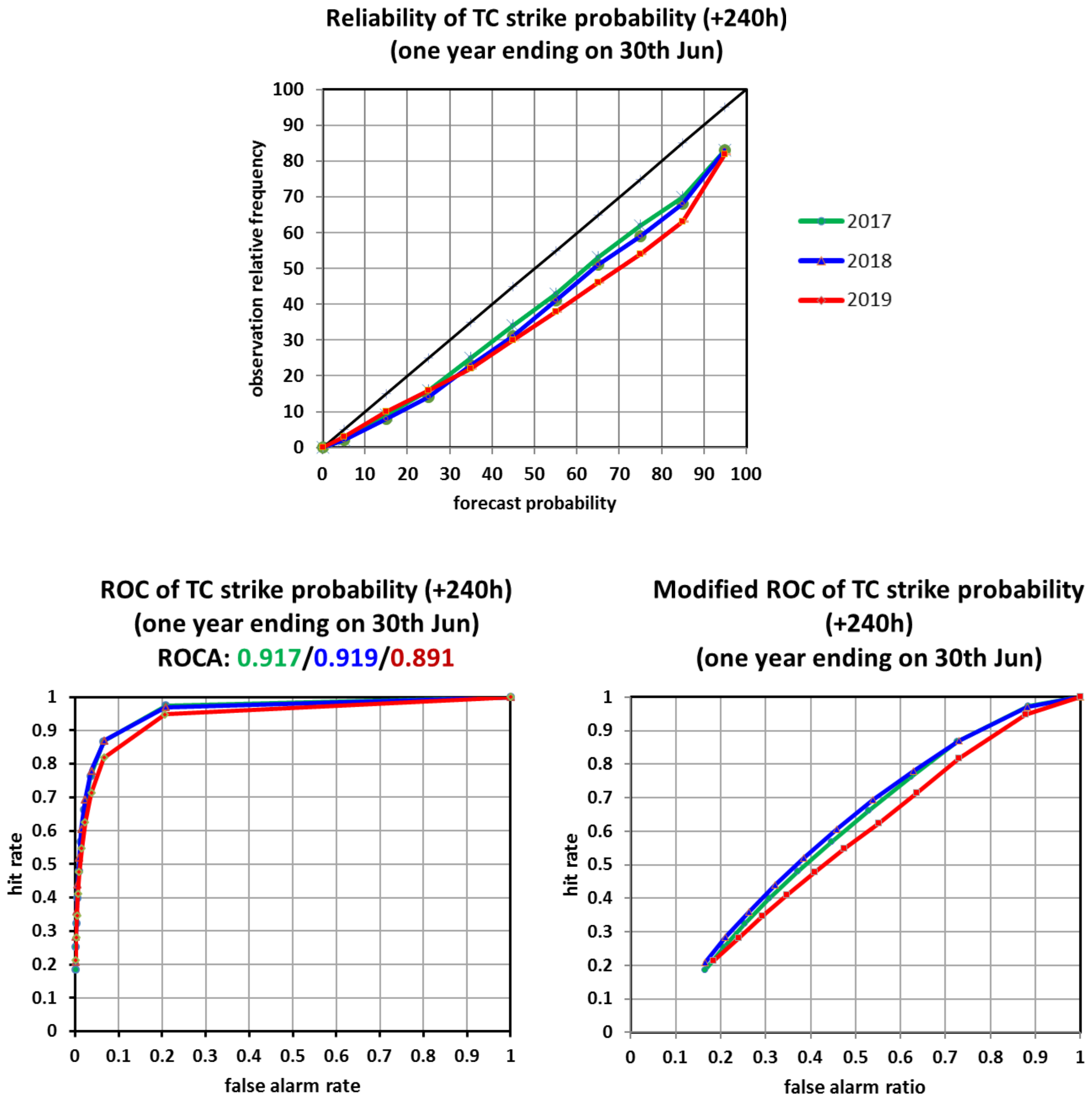


Figure 33: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2016–June 2017 (green), July 2017–June 2018 (blue) and July 2018–June 2019 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

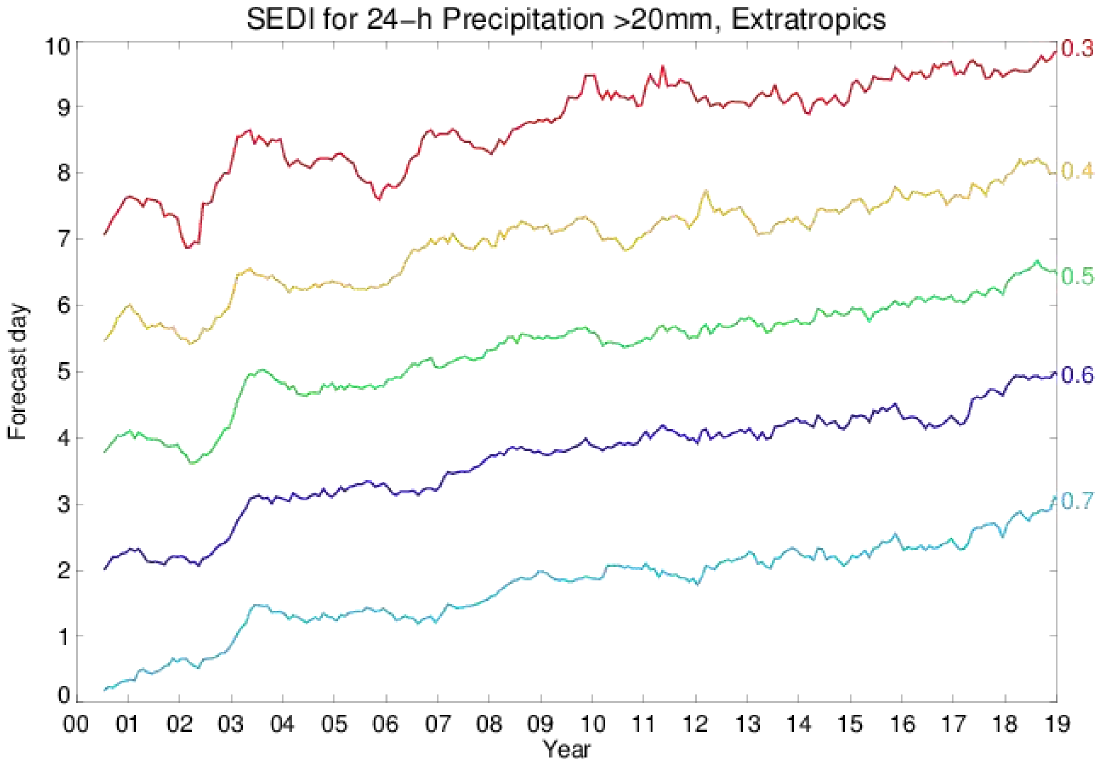


Figure 34: Evolution of skill of the HRES forecast in predicting 24-h precipitation amounts >20 mm in the extra-tropics as measured by the SEDI score, expressed in terms of forecast days. Verification is against SYNOP observations. Numbers on the right indicate different SEDI thresholds used. Curves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

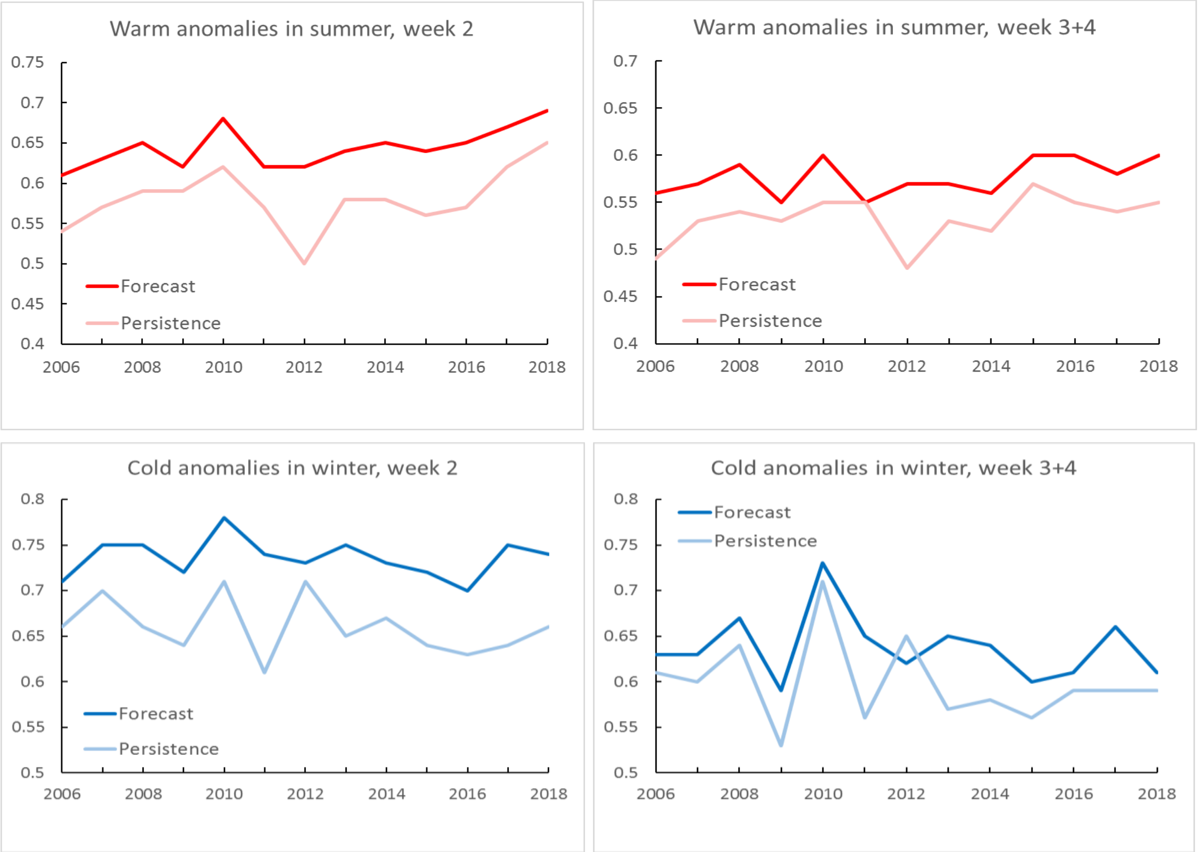


*Figure 35**:** Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all 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 show the score using persistence of the preceding 7-day or 14-day period of the forecast.*

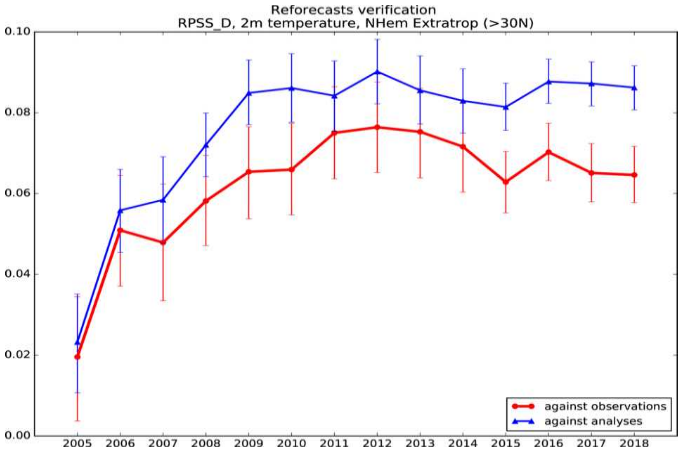


Figure 36: Skill of the ENS in predicting weekly mean 2m temperature anomalies (terciles) in week 3 in the northern extratropics. Verification against own analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.

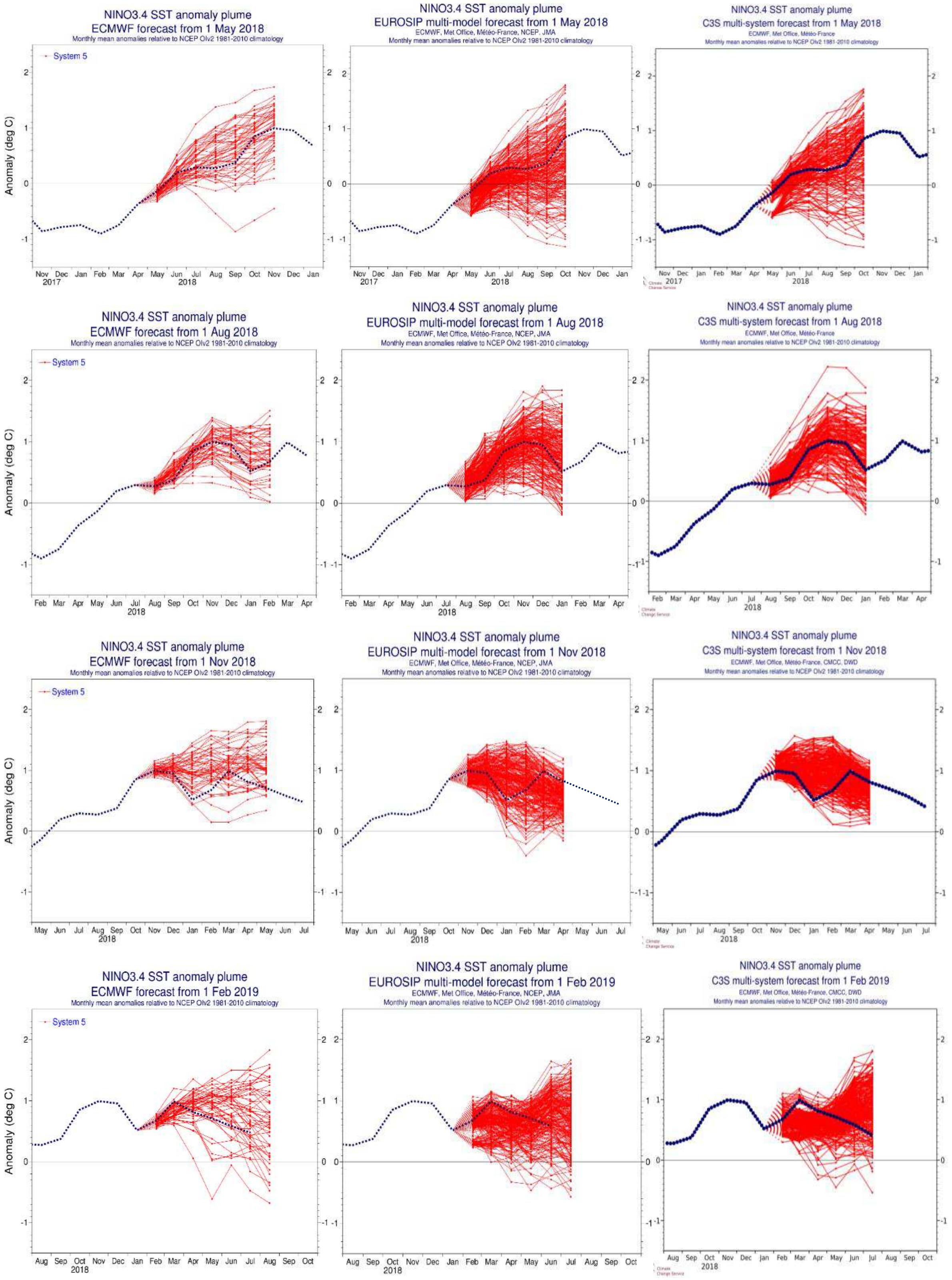


Figure 37: ECMWF (left column), EUROSIP multi-model (centre column) and Copernicus Climate Change Service (C3S, right column) multi-model seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2018, August 2018, November 2018 and February 2019. The red lines represent the ensemble members; dotted blue line shows the subsequent verification.

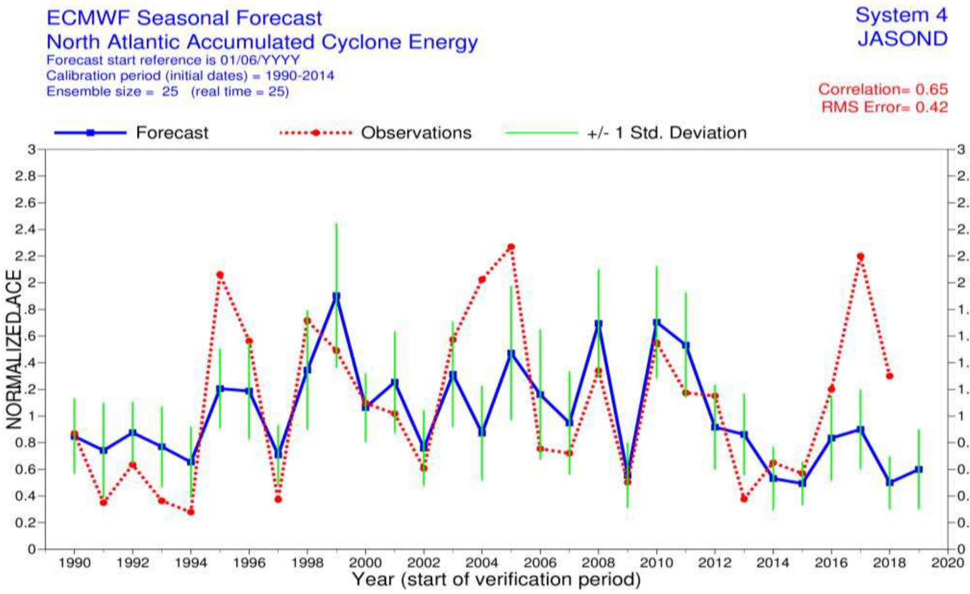


Figure 38: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2018. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts are from SEAS5 of the seasonal component of the IFS: these are based on the 25-member re-forecasts; from 2017 onward they are from the operational 51-member seasonal forecast ensemble. Start date of the forecast is 1 June.

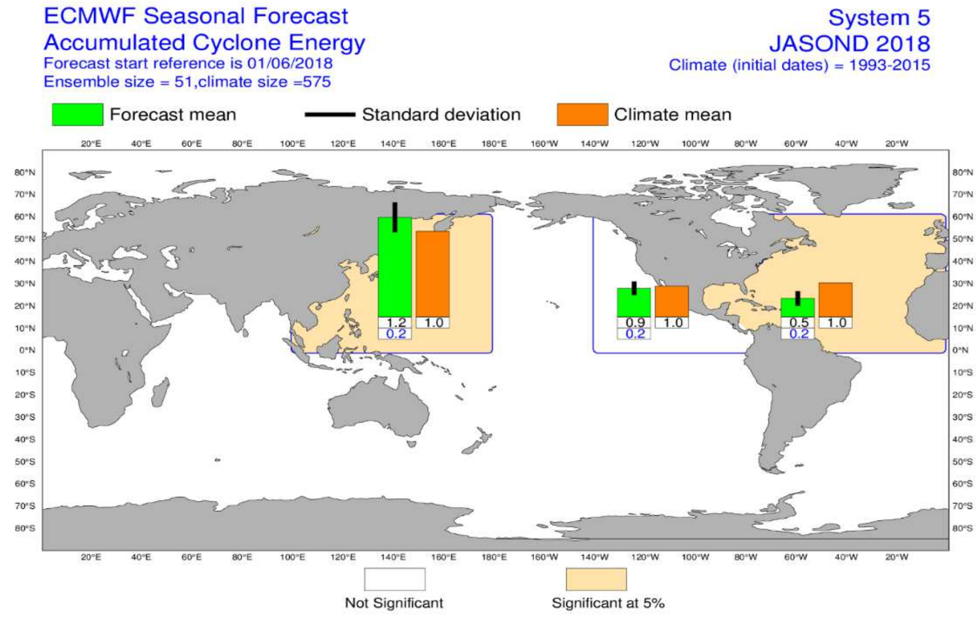


*Figure 39**:** Tropical storm frequency forecast issued in June 2018 for the six-month period July–December 2018. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. 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.*

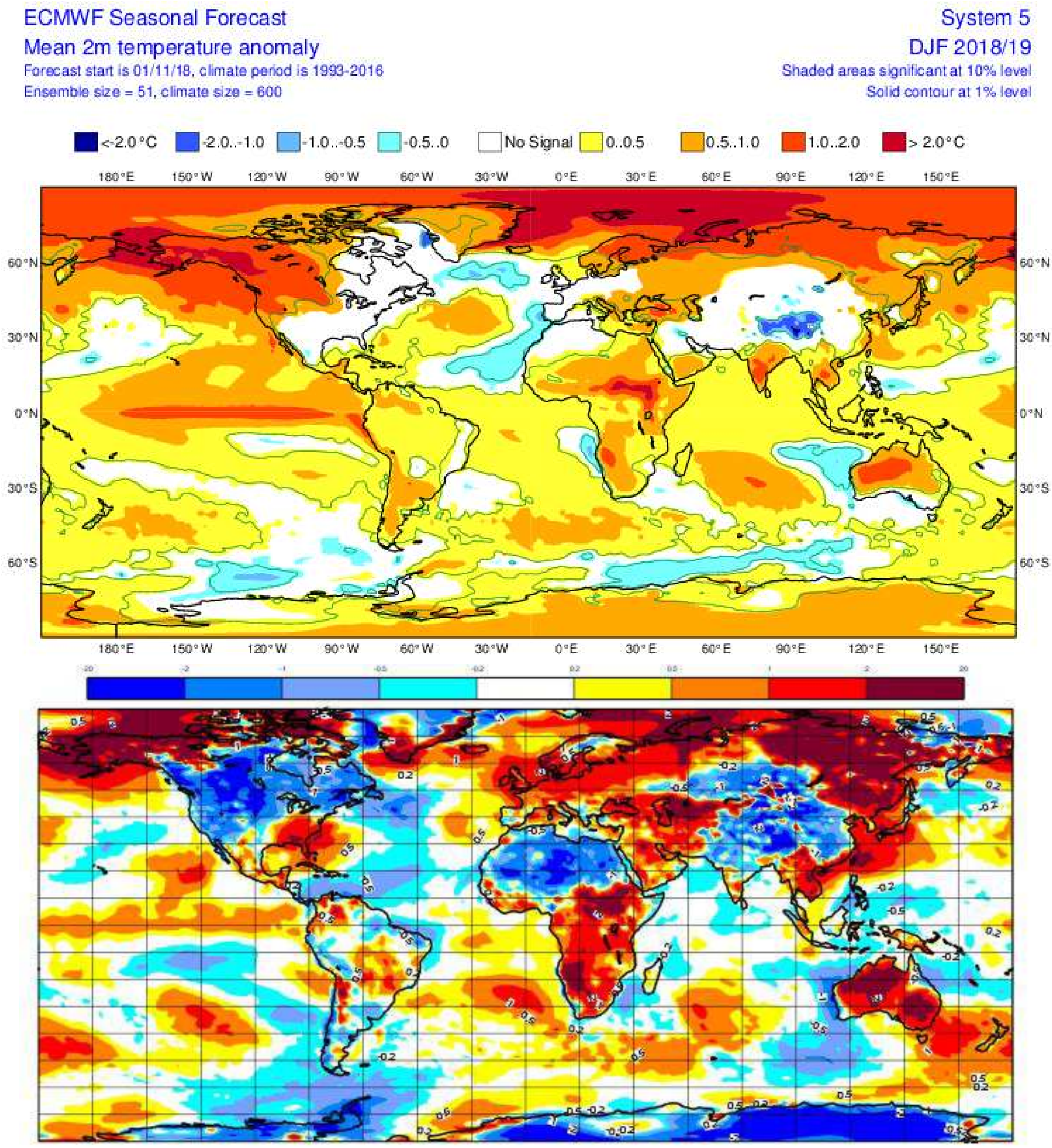


Figure 40: Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2018 for DJF 2018/19 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

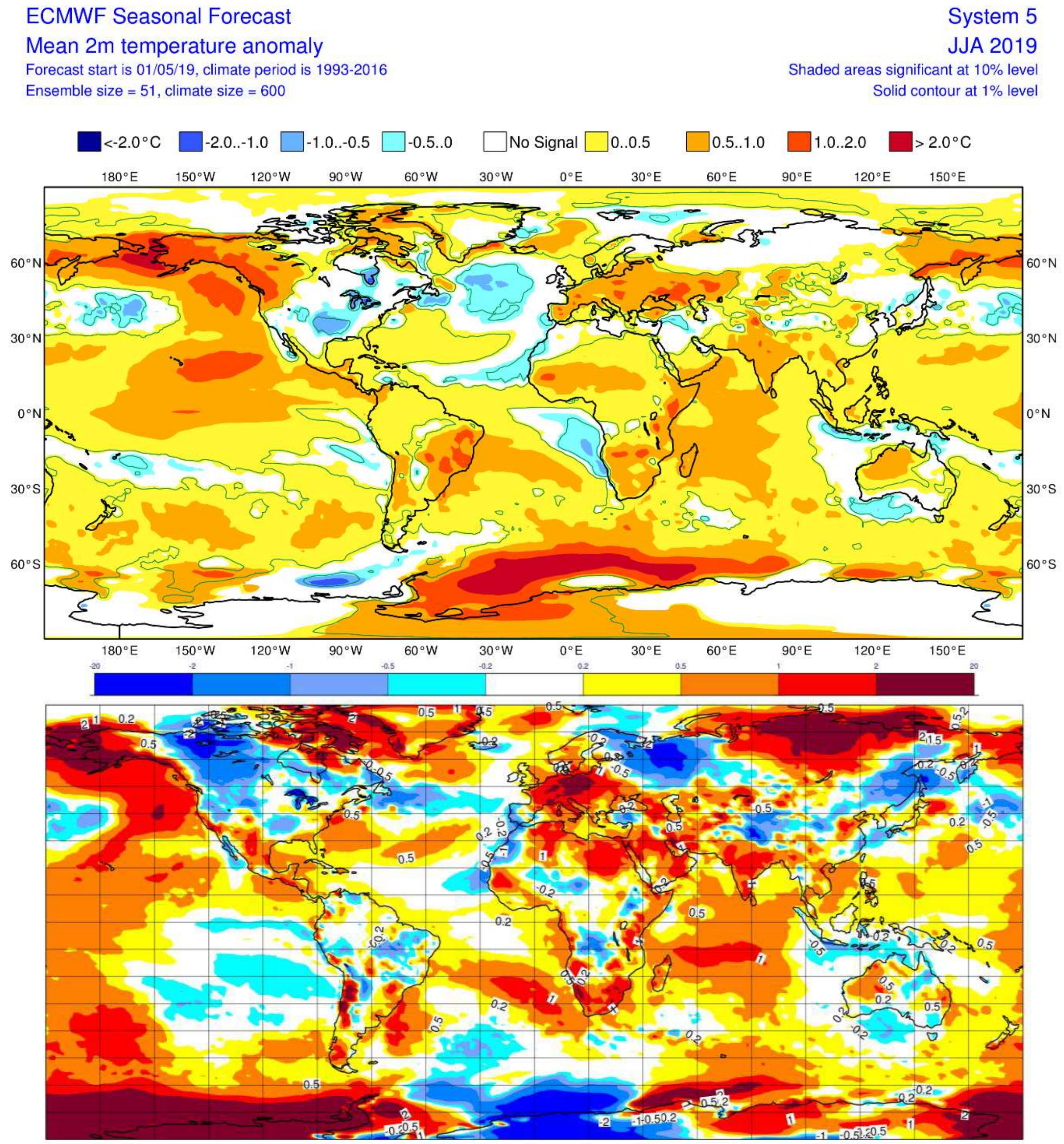


Figure 41: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2019 for JJA 2019 (upper panel) and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

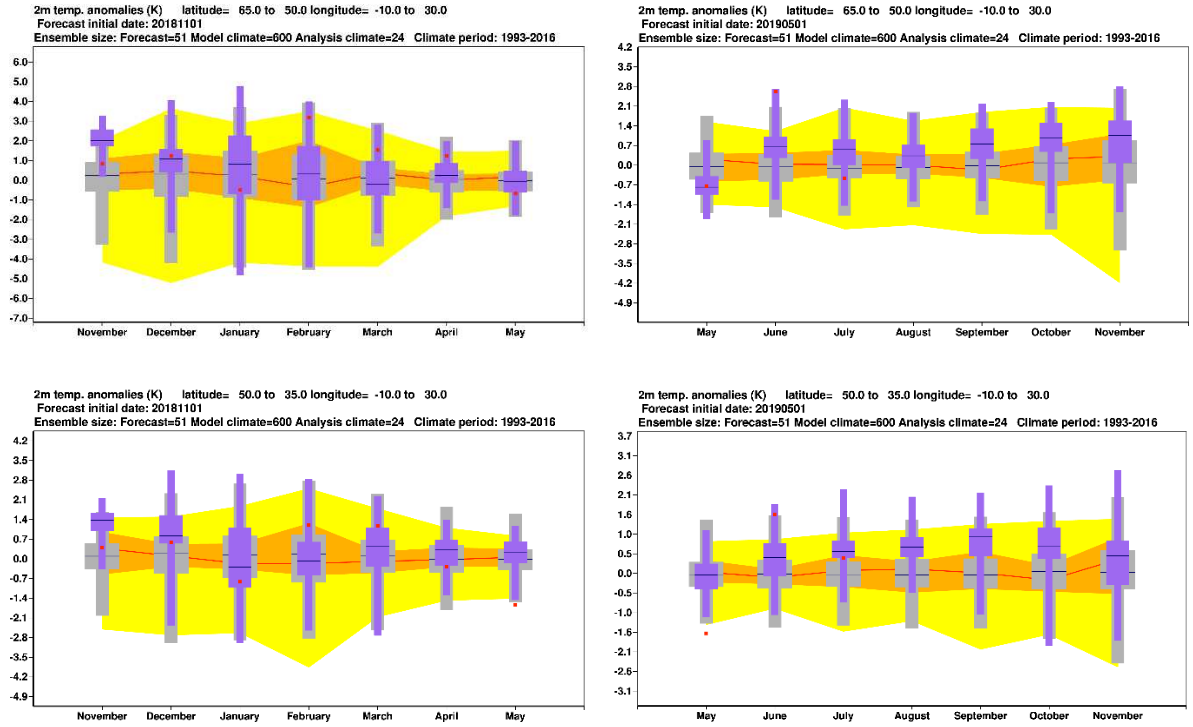


Figure 42: Long-range forecast of 2 m temperature anomalies from November 2018 for DJF 2018–19 (left panels) and from May 2019 for JJA 2019 (right panels) for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 hindcasts is shown in grey, and the analysis in the 24-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.

Appendix A: 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 15), 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 15, Figure 17) 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:

$$\mathrm{SS} = 100 * \left(1 - \frac{\mathrm{RMSE}_f^2}{\mathrm{RMSE}_p^2}\right)$$

Figure 4 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 8) and its inter-annual variability (Figure **12**).

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.

The comparison of spread and skill (Figure 9 to Figure **11**) takes into account the effect of finite ensemble size N by multiplying spread by the factor (N+1)/(N-1).

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

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure **19**, Figure 20). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 21 to Figure 24), 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

$$SEDI = \frac{logF - logH - log(1-F) + log(1-H)}{logF + logH + log(1-F) + log(1-H)}$$

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

Ardhuin, F., E. Rogers, A. Babanin, J.-F. Filipot, R. Magne, A. Roland, A. van der Westhuysen, P. Queffeulou, J.-M. Lefevre, L. Aouf & F. Collard, 2010: Semi-empirical dissipation source functions for windwave models: part I, definition, calibration and validation. J. Phys. Oceanogr., 40 (9), 1917–1941.

de Rosnay, P., M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel & L. Isaksen, 2013: A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF. Q.J.R. Meteorol. Soc., 139, 1199–1213, doi:10.1002/qj.2023.

Ferranti, L., L. Magnusson, F. Vitart and D.S. Richardson, 2018: How far in advance can we predict changes in large-scale flow leading to severe cold conditions over Europe? Q.J.R. Meteorol. Soc, 144**,** doi:10.1002/qj.3341.

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.

Haiden, T., I. Sandu, G. Balsamo, G. Arduini and A. Beljaars, 2018: Addressing biases in near-surface forecasts. ECMWF Newsletter No. 157, 20-25.

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

Hogan, R.J. & A. Bozzo, 2018: A flexible and efficient radiation scheme for the ECMWF model. J. Adv. Modeling Earth Sys., 10 (8), 1990–2008, doi:10.1029/2018MS001364.

Janssen, P. 2017: Shallow water version of the freak wave warning system. ECMWF Technical Memorandum No. 813.

Lang, S., E. Hólm, M. Bonavita & Y. Tremolet, 2019: A 50-member Ensemble of Data Assimilations. ECMWF Newsletter No. 158, 27–29, doi:10.21957/nb251xc4sl.

Lean, P., M. Bonavita, E. Hólm & T. McNally, 2019: Continuous data assimilation for the IFS. ECMWF Newsletter No. 158, 21–26, doi:10.21957/9pl5fc37it.

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.

Schmederer, P., I. Sandu, T. Haiden, A. Beljaars, M. Leutbecher, and C. Becker, 2019: Use of super-site observations to evaluate near-surface temperature forecasts. ECMWF Newsletter No. 161, 32-38.