TECHNICAL MEMORANDUM

831

Evaluation of ECMWF forecasts, including the 2018 upgrade

T. Haiden, M. Janousek, J. Bidlot,
R. Buizza, L. Ferranti,
F. Prates, and F. Vitart

Forecast Department

October 2018

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

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

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

1 Introduction

The most recent change to the ECMWF forecasting system (IFS Cycle 45r1, on 5 June 2018) 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 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, 817). 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 45r1

On 5 June 2018, ECMWF implemented a substantial upgrade of its Integrated Forecasting System (IFS). IFS Cycle 45r1 brings coupling to all ECMWF forecasts, from forecast day 1 to one year, by including the three-dimensional ocean and sea-ice model in the single high-resolution forecast (HRES). In line with increased ocean–atmosphere coupling, the sea-ice product used by the atmospheric analysis in IFS Cycle 45r1 is provided by ECMWF's ocean analysis (OCEAN5).

2.1.1 Data assimilation

For the first time, the atmospheric assimilation makes use of the OCEAN5 sea-ice analysis in the surface analysis of the high-resolution and EDA analyses. This enhances the coupling between the ocean and atmosphere. OCEAN5 makes use of LIM2 and assimilates the OSTIA-OSI-SAF product instead of OSTIA. Relative humidity increments are calculated using temperature instead of virtual temperature. The weak constraint model error forcing is applied at every time step instead of every hour to avoid shocks in the model integration. Changes to the tangent-linear (TL) and adjoint (AD) physics have led to a dramatic reduction in the number of spuriously large analysis increments in both the EDA and the high-resolution 4D-Var analysis.

With IFS Cycle 45r1, an increased number of observations are assimilated, such as non-surface-sensitive infrared (IR) channels over land and the assimilation of all- sky microwave (MW) sounding channels over coasts. Changes also include the introduction of RTTOV-12 and new microwave instrument coefficients, a retuning of the radiosonde observation error, and introduction of a scheme to account for radiosonde drift (Ingleby et al., 2018).

Wave height data from JASON-3 and Sentinel-3A altimeters is assimilated. BUFR SYNOP observations are used in the surface analysis, including more than 200 additional snow depth observations in China.

2.1.2 Model changes

Changes to the forecast model include: introduction of coupling of HRES to the 3-dimensional ocean model NEMO, with a 0.25 degree resolution and 75 layers, and LIM2 sea ice model (as in ENS, see Keeley and Mogensen, 2018); improved numerics for warm-rain cloud microphysics and the vertical extrapolation for semi-Lagrangian trajectories; improved representation of supercooled liquid water in convective clouds (Forbes et al., 2016); improved representation of mid- to upper-stratospheric water vapour; new output parameters including maximum CAPE and CAPE-shear in the last six hours of the forecast and lightning flash density; new bathymetry (water depth) in the wave model, mainly affecting wave fields in coastal areas.

The coupled model configuration with the community ocean model NEMO (the Nucleus for European Modelling of the Ocean, http://www.nemo-ocean.eu/) and LIM2 (the Louvain-la-Neuve sea-ice model developed at the Belgian Université catholique de Louvain) has been used in the medium-range/monthly ensemble (ENS) since 2016 (Buizza et al., 2017). Its introduction in the HRES enables rapidly interacting processes (e.g. during tropical cyclones) to be better described. In line with increased ocean–atmosphere coupling, the sea-ice product used by the atmospheric analysis in IFS Cycle 45r1 is provided by ECMWF's ocean analysis (OCEAN5). The upgrade introduces full ocean coupling in the tropics for both HRES and ENS, but it retains partial coupling in the extratropics. Partial coupling, as implemented in ENS in 2016, couples the sea-surface temperature tendencies rather than the actual sea-surface temperature field from the ocean model during the first four days of the forecast.

2.1.3 Model uncertainties (EDA, ENS)

There is an improved flow-dependent error representation in the SPPT scheme via reduced spread in clear-sky regions (due to unperturbed radiative tendency in clear sky), the activation of tendency perturbations in the stratosphere, and weaker tapering of perturbations in the boundary layer. Further changes are a reduction in the amplitude of the SPPT perturbation patterns (by 20%); introduction of the cycling of stochastic physics random fields in the EDA, and adoption of the same SPPT configuration in EDA as in ENS; deactivation of the stochastic backscatter (SKEB) scheme due to improved model error representation by the SPPT scheme (see above), leading to a 2.5% cost saving in ENS.

2.2 Meteorological impact of the new cycle

A comparison of parallel runs of the operational cycle (43r3) and the new cycle (45r1) indicates an overall positive impact in the tropics for both HRES and ENS Figure 1 and Figure 2. For the extratropics, results are mixed, with an overall slightly positive impact on the HRES scores, while for the ENS the sign of the impact depends on the geographical region and the variable.

2.2.1 Upper air

The new cycle leads to improvements in HRES upper-air fields. When these fields are verified against the model analysis, a positive signal is seen throughout the troposphere for most parameters, except temperature in the lower troposphere at shorter ranges. The latter is mainly a result of changes to the analysis, linked to changes in the stochastic scheme used in the EDA.

In cycle 45r1, changes to SPPT have induced an increase in the EDA spread, especially in the troposphere, with the result that the analysis is drawn closer to the observations and further away from the first guess. This, on average, improves EDA reliability, but can have a negative impact on the short-range forecast error evaluated against the analysis. Thus, care must be taken in interpreting the scorecard values computed against analyses (the top half of the scorecards) versus the values computed against observations (the lower half of the scorecards).

Forecast verification against observations shows a neutral impact. Upper-air improvements are more pronounced in the tropics, especially for wind and temperature. When verified against observations, upper-air changes are overall positive in the tropics except for relative humidity, and neutral to slightly positive in the extratropics. Upper-air results for ENS verified against the analysis are mostly positive in the tropics but more neutral in the extratropics. The negative signal for temperature in the lower troposphere at shorter lead times is again mainly due to changes in the analysis.

Against observations, results are mostly negative in the extratropics at short lead times and significantly positive in the tropics, with the exception of relative humidity at 700 hPa. The negative impact in the extratropics is partly due to a slight reduction in ensemble spread associated with the transition to a physically more realistic SPPT scheme. Whether or not this reduced spread is genuinely detrimental depends on how significant the impact of observation errors is in the verification; this has not been routinely taken into account so far. Experimental verification against radiosonde data that takes observation error into account indicates that a large fraction of the negative ENS results become statistically non-significant.

2.2.2 Weather parameters and waves

There is an overall improvement in 2-metre temperature both in the HRES and ENS, particularly for Europe. The impact on 2-metre humidity is largely neutral for HRES and positive for ENS, particularly in the tropics, while for 10-metre wind speed the impact is largely neutral in the HRES but slightly negative in the ENS. Precipitation in the HRES is improved in terms of categorical verification (e.g. the SEEPS score), and near-coastal precipitation in warm-rain dominated situations is significantly improved due to changes in the cloud physics (Forbes et al., 2018). However, the model changes lead to more activity at higher precipitation rates in active regions such as the East Asian monsoon, and as a result error measures such as RMSE or CRPS (for the ENS) are increased. The negative signal for significant wave height against analysis is a result of changes to the analysis resulting from a large increase in observation usage.

Cycle 45r1 includes lightning density as a new forecast parameter. For a more detailed description of this product, its parameterization, and evaluation results against satellite and ground-based observations, see Lopez (2018).

The coupling of the HRES from day zero and the full coupling in the tropics for HRES and ENS have an impact on the verification against the analysis of near-surface fields. In the tropics, before the upgrade the high-resolution atmospheric analysis saw the sea-surface temperature (SST) from OSTIA, while the coupled forecast model is initialised from the OCEAN5 SST.

2.2.3 Monthly forecast

Changes in scores for the monthly system are generally positive in the tropics with significant improvements in weeks 1 and 2 across the range of parameters, and weeks 3 and 4 for some parameters (Figure 3). In the extratropics, there are significant improvements in week 1 and largely neutral results in weeks 2 to 4. There is an indication of a positive effect on skill across all parameters in Europe. Before the upgrade, there was too little spread in the MJO Index. Changes in 45r1 to the SPPT scheme have now brought the spread and error into close agreement throughout the 30-day forecast range. The underestimation of the MJO Index amplitude has been significantly reduced throughout the forecast

2.2.4 Tropical cyclones

The implementation of the ocean-atmosphere coupling in the HRES removes the overall negative bias in tropical cyclone central pressure and thereby reduces the mean absolute intensity error by about 10% in the short range and by about 20% from day 5 onwards (Buizza et al., 2018). Evaluations indicate statistically neutral results for the position error. For further details on the influence of ocean–atmosphere coupling on tropical cyclone intensity forecasts, see Mogensen et al. (2018).

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 February 2018, the highest score of any individual month so far was reached in the northern hemisphere. As in the winter 2010, this resulted from the combined effect of large anomalies in the large-scale flow pattern and the long-term increase of skill of the IFS. The 12-month running mean in 2018 is at about the same level as in the previous year. Comparison with ERA5 (not shown) confirms that the slight downward trend in 2017 in both hemispheres is due to atmospheric variability. In Europe, the signal-to-noise ratio is smaller due to the smaller area, and the positive trend is less pronounced there in recent years compared to the hemispheric scores.

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 RMS error of the six-day forecast has remained at about the same level over the last 2 years.

Figure 6 shows the time series of the average RMS difference between four- and three-day (blue) and six- and five-day (red) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the general downward trend indicates that there is less “jumpiness” in the forecast from day to day. The 12-month running mean of the level of consistency between consecutive forecasts has levelled off in the last year. However, some further reduction is anticipated due to the very low monthly values reached in summer 2018 (thin curves) both in Europe and in the extratropics.

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 in RMSE for wind speed and for day-1 temperature, which have been seen in recent years, are continuing. The 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 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. On a 5-year timescale, there is weak positive trend visible in the extratropics. In Europe, no clear trend can be identified on top of the large inter-annual variations.

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 very good overall match between spread and error. For 850 hPa temperature in particular, the under-dispersion has been reduced significantly compared to previous years.

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). Note that both Figure 10 and Figure 11 show the unnormalized spread and error, in contrast to the previous year (ECMWF Tech Memo No. 817), where normalized values were shown. It was found that the normalization using the ERA Interim climatology worked fine for ECMWF but created some spurious effects for 850 hPa temperature for the other centres.

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 ERA-Interim. Figure 12 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. In both hemispheres, the skill of the ENS relative to the reference forecast is now around 16%.

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. For a lead time of 5 days (red curves), 12-month running mean errors have reached 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. 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. 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 the ECMWF medium-range forecasts relative to the other centres when verified against observations. At the shortest range (day 1), this lead is less pronounced, especially in Europe.

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-1) 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. However, ECMWF's lead in the tropics is smaller than it is in the extratropics.

4 Weather parameters and ocean waves

4.1 Weather parameters – high-resolution and ensemble

The supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 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 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.

In 2017, the deterministic precipitation forecast has reached its highest level of skill so far (red line in Figure 19). The lead time at which the given threshold is reached has increased by 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 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 probabilistic precipitation score (lower panel in Figure 19) 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 in 2017 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 20). 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 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) there is a visible reduction in the error standard deviation (upper curves) in the last 2–3 years. The biases in 2 m temperature (lower curves) have not changed substantially, except that the large annual variation of the night-time bias (blue curve) has been reduced in recent years. Similar to 2 m temperature, the 2 m dewpoint (Figure 22) shows a reduction of the error standard deviation, while biases have not changed. During daytime, there is a dry bias of 0.5-1 K. Systematic errors in near-surface parameters are currently investigated in a comprehensive study, which has already identified the causes of some of these biases (Haiden et al., 2018).

For total cloud cover (Figure 23) the error standard deviation is showing little change, as well as the bias, which is relatively small (less than 0.5 okta) both during the day and at night. 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. The causes of these differences are currently under investigation. Systematic errors in cloudiness, as well as the treatment of snow cover in the model appear to play a role (Haiden et al., 2018).

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 and ERA5 (Figure 25), with the standard deviation of the error used as a metric. The most clear-cut continuation of the upward trend of recent years is seen in the SH extratropics, due to an improved treatment of cloud ice in the convective parametrization that was introduced in model cycle 43r3. This had a positive effect especially on the cloudiness over the Southern Ocean. In the tropics, and relative to ERA5 also in the NH extratropics, no significant trend appears in recent years.

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 26 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. All parameters show the beneficial effect of recent model upgrades. 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). However, 2 m temperature has improved by an amount comparable to 850 hPa temperature.

Following a recommendation by the TAC Subgroup on Verification, 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, when the CRPS exceeds 5 K. Figure 27 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. It can be seen that recent model upgrades, such as the resolution increase in 2016, have improved this score both in summer and winter.

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. In the northern hemisphere, anomaly correlation for significant wave height 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 2017–May 2018. 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.

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 intensity errors of the HRES and the CTRL at D+3 have further decreased and have again reached the lowest levels seen so far (2011–12). Mean absolute speed errors are comparable to previous years.

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 2017 there has been a change to over-dispersive spread. This has been addressed 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 slightly increased over-confidence 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 2018 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.

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.

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.

A comprehensive 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 was implemented in November 2017, replacing System 4, which was implemented in 2011. SEAS5 includes updated versions of the atmospheric (IFS) and interactive ocean (NEMO) models and adds the interactive sea ice model LIM2. Ocean horizontal and vertical resolution have been increased. Ocean and land initial conditions have been updated, and the re-forecast ensemble size has been increased from 15 to 25. 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. SEAS5 highlights include a marked 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 2017 was characterized by a change from weakly positive to weakly negative SST anomalies in the eastern tropical Pacific. At first this was poorly captured in the forecast (Figure 37, top row), with EUROSIP providing somewhat better guidance in this case than the ECMWF forecast alone, albeit at the cost of rather large spread. The return to weakly positive anomalies was very well predicted by ECMWF (Figure 37, 3rd and 4th rows).

6.2.3 Tropical storm predictions from the seasonal forecasts

The 2017 Atlantic hurricane season had a total of 17 named storms including 10 hurricanes and 6 major hurricanes. It was a catastrophic season and the most active since 2005 with an accumulated cyclone energy index (ACE) of about 225% of the 1993–2015 climate average (Figure 38). Seasonal tropical storm predictions from SEAS5 indicated an average level of activity over the Atlantic (ACE of about 90% of the 1993–2015 climate average). Similarly, the number of tropical storms (17) which formed in 2017 was above average (13) whereas the forecast predicted 13 (with a range from 8.4 to 15) tropical storms in the Atlantic (Figure 39).

Also at shorter (monthly) ranges, forecasts failed to capture the heightened Atlantic TC activity. The poor seasonal and extended range TC prediction this year was mainly due to the absence of a strong El Nino or La Nina, and the fact that the Atlantic TC in the model is more strongly correlated with El Nino / La Nina than it is in the real atmosphere, where local SSTs (which were high in 2017) have a stronger influence.

The figure also shows that SEAS5 predicted average activity over the eastern North Pacific, and slightly below average activity over the western North Pacific (ACE of about 90% of the 1993–2015 climate average). In the western North Pacific, although SEAS5 predicted a slightly below average season, the tropical storm activity was underestimated, with 25 observed and 19 predicted tropical storms in the period July to December. In the eastern North Pacific, SEAS5 did predict an average season with 13 tropical storms while 14 were observed from July to December.

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 2017. 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 2017–18) was characterized by strong warm anomalies in the Arctic and in northern parts of Eurasia and in Alaska. A pronounced cold anomaly was present over Canada. 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). For Canada, near-neutral conditions were predicted. Anomaly patterns within Europe, in particular the cold conditions in parts of western Europe and Scandinavia, were poorly predicted.

Large parts of Europe experienced another hot summer season in 2018. For the northern-hemisphere summer (JJA 2018) the forecast predicted positive anomalies over Southern Europe and the Mediterranean. However, as for other models contributing to C3S, the large and persistent positive anomalies observed in large parts of Europe were not seen in the forecast (Figure 41). In contrast, temperature anomalies over the North Atlantic were well represented.

It should be noted that the analysis anomalies used to verify the seasonal and sub-seasonal forecasts are based on operational analysis departures from the ERA-Interim climate. This practice results from the immediate

availability of the operational analysis. However, the method is not ideal since the model version of ERA-Interim increasingly differs from the latest operational model. In the case of JJA 2018, for example, we have noticed that the signal of cold anomalies over Greenland (Figure 41, lower panel), which is not seen in the forecast, is amplified in the verifying analysis due to a different treatment of albedo between ERA-Interim and the operational model.

Climagrams for Northern and Southern Europe for winter 2017-18 and summer 2018 are shown in Figure 42. Red squares indicate observed monthly anomalies. Within the first two to three months of the forecast, the sign of the forecast anomaly is generally predicted correctly, and the observations fall well within the ensemble distribution. A notable exception is the summer forecast for northern Europe (upper right panel), where the observed anomalies exceed the 95th percentile of the ENS both in May and July 2018.

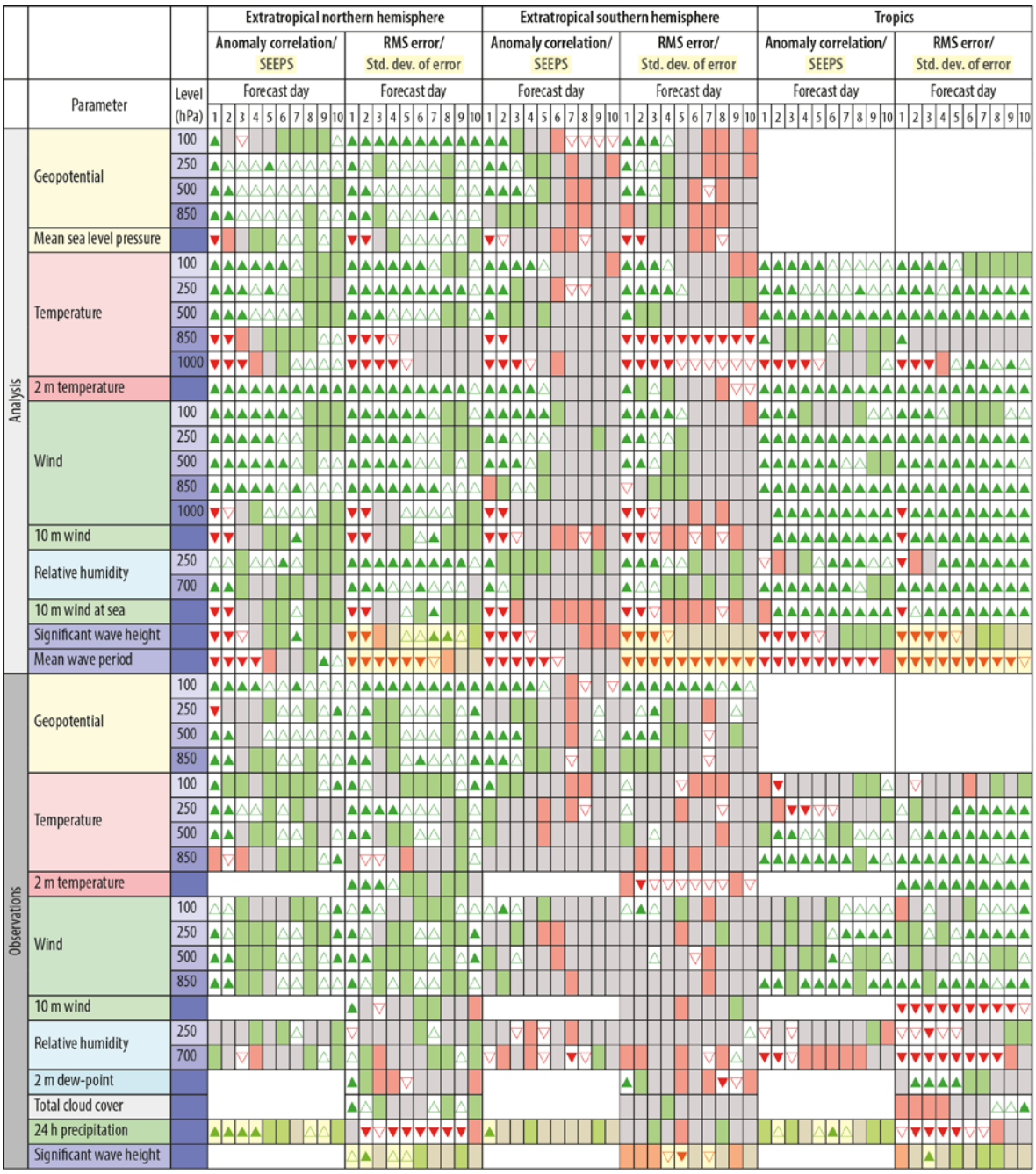
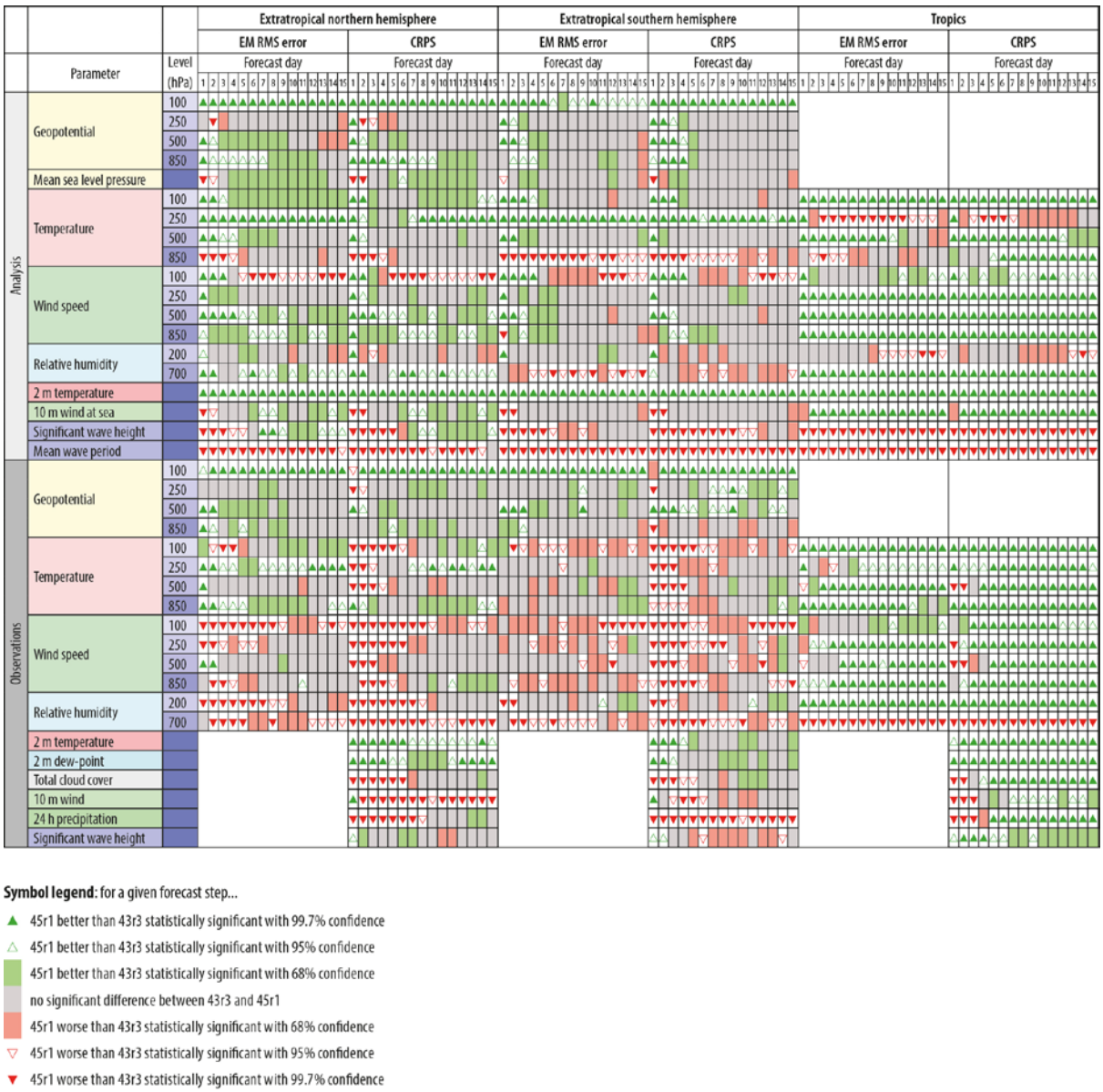


Figure 1: Summary score card for IFS Cycle 45r1. Score card for HRES cycle 45r1 versus cycle 43r3 verified by the respective analyses and observations at 00 and 12 UTC for 855 forecast runs in the period December 2016 to June 2018. Yellow colouring indicates that symbols refer to the second score indicated at the top of the column.



Symbol legend: for a given forecast step...

- ▲ 45r1 better than 43r3 statistically significant with 99.7% confidence
- △ 45r1 better than 43r3 statistically significant with 95% confidence
- 45r1 better than 43r3 statistically significant with 68% confidence
- no significant difference between 43r3 and 45r1
- 45r1 worse than 43r3 statistically significant with 68% confidence
- ▽ 45r1 worse than 43r3 statistically significant with 95% confidence
- ▼ 45r1 worse than 43r3 statistically significant with 99.7% confidence

Figure 2: Summary ENS score card for IFS Cycle 45r1. Score card for ENS cycle 45r1 versus cycle 43r3 verified by the respective analyses and observations at 00 and 12 UTC for 408 ENS forecast runs in the period December 2016 to June 2018.

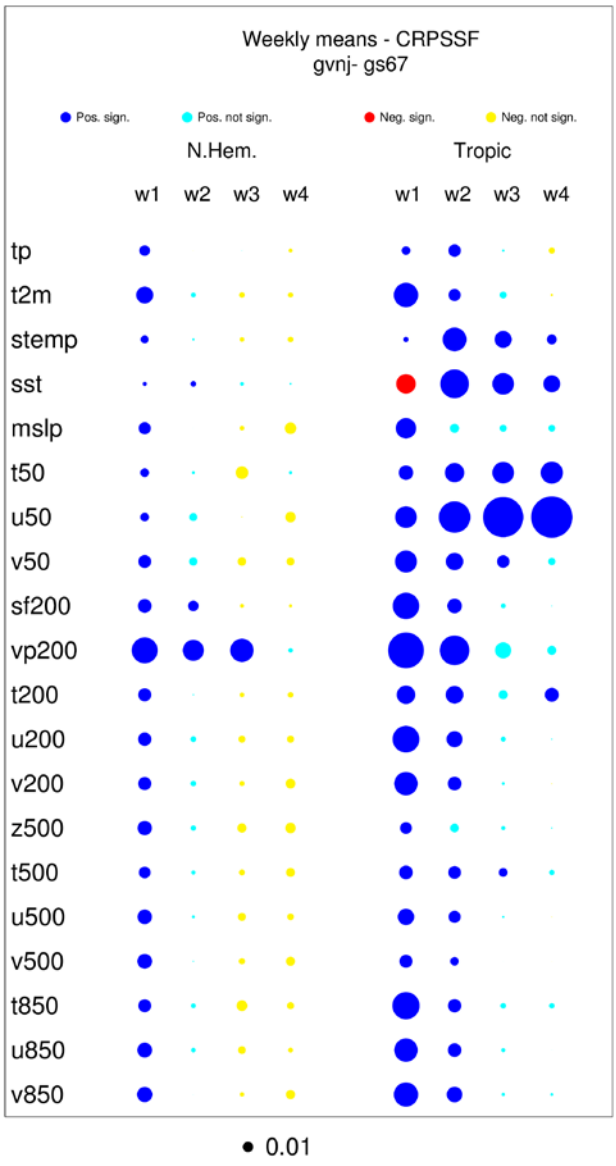


Figure 3: Scorecard for Cycle 45r1 for the extended range. Columns show score differences between 45r1 and 43r3 for weekly means in the Northern Extratropics and the Tropics. Size of circles shows matgnitude of difference, colour indicates statistical significance. Verification based on 15 members 12 times a year (1st of each month).

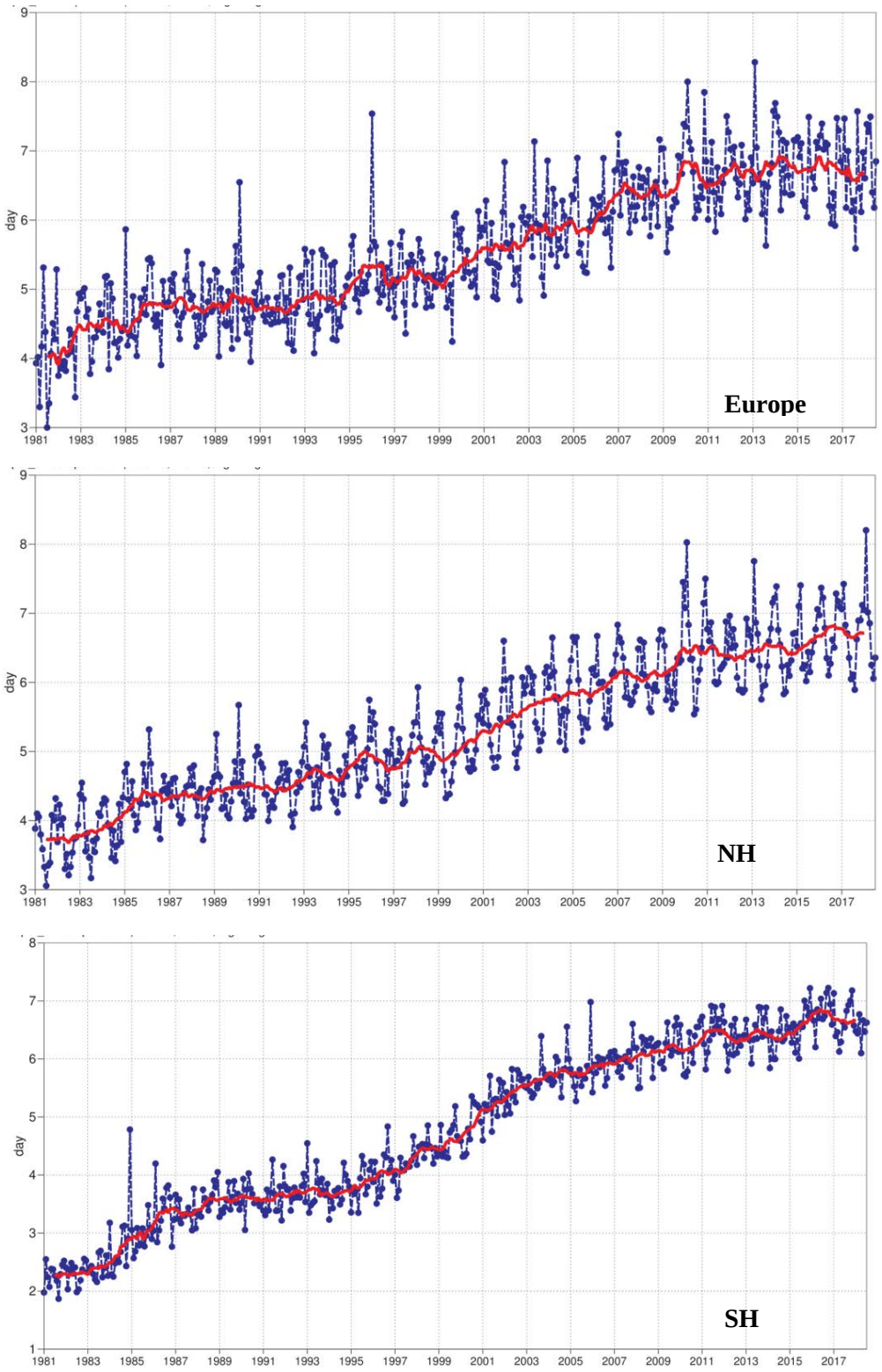


Figure 4: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly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

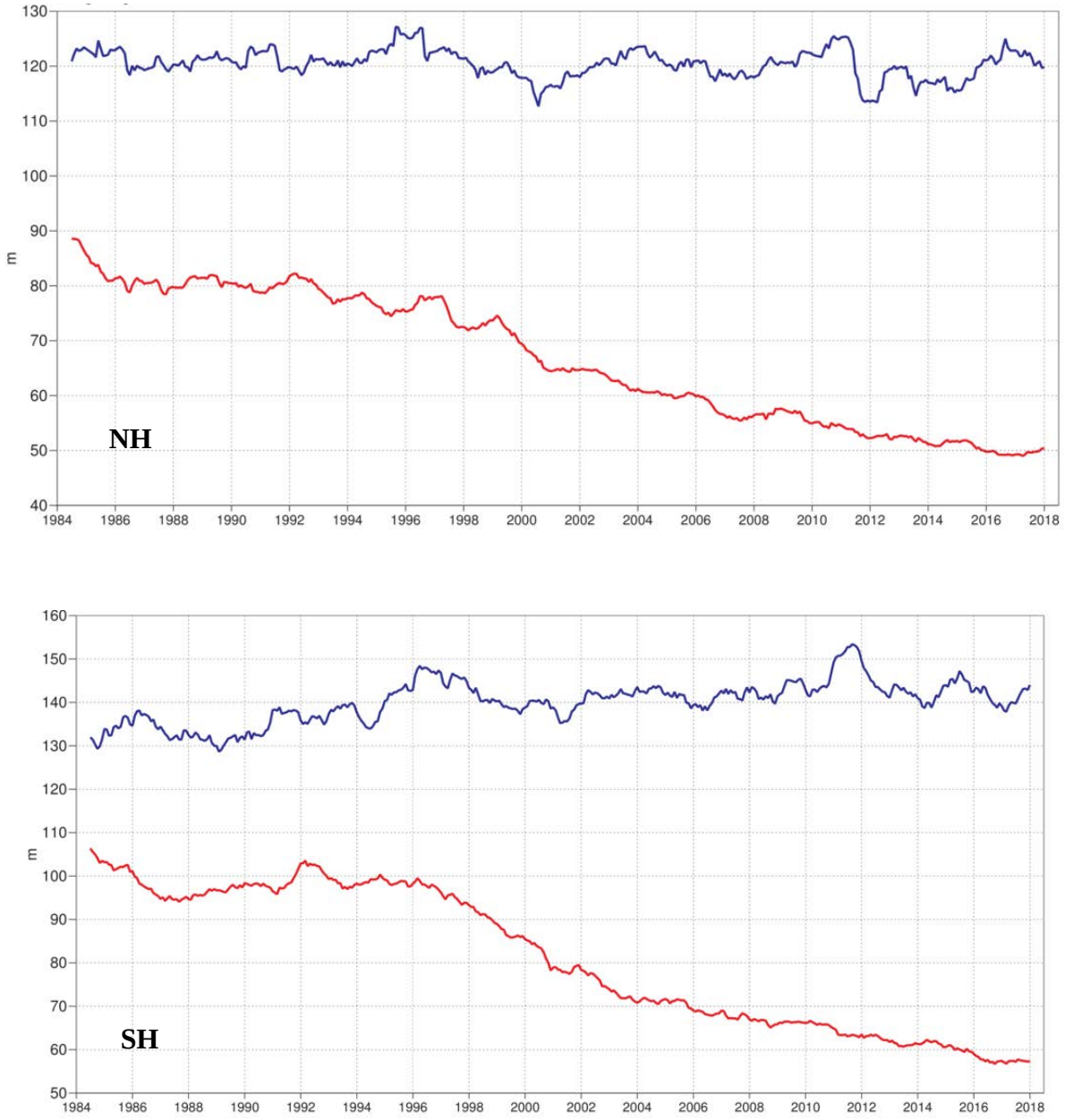


Figure 5: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2017–July 2018. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

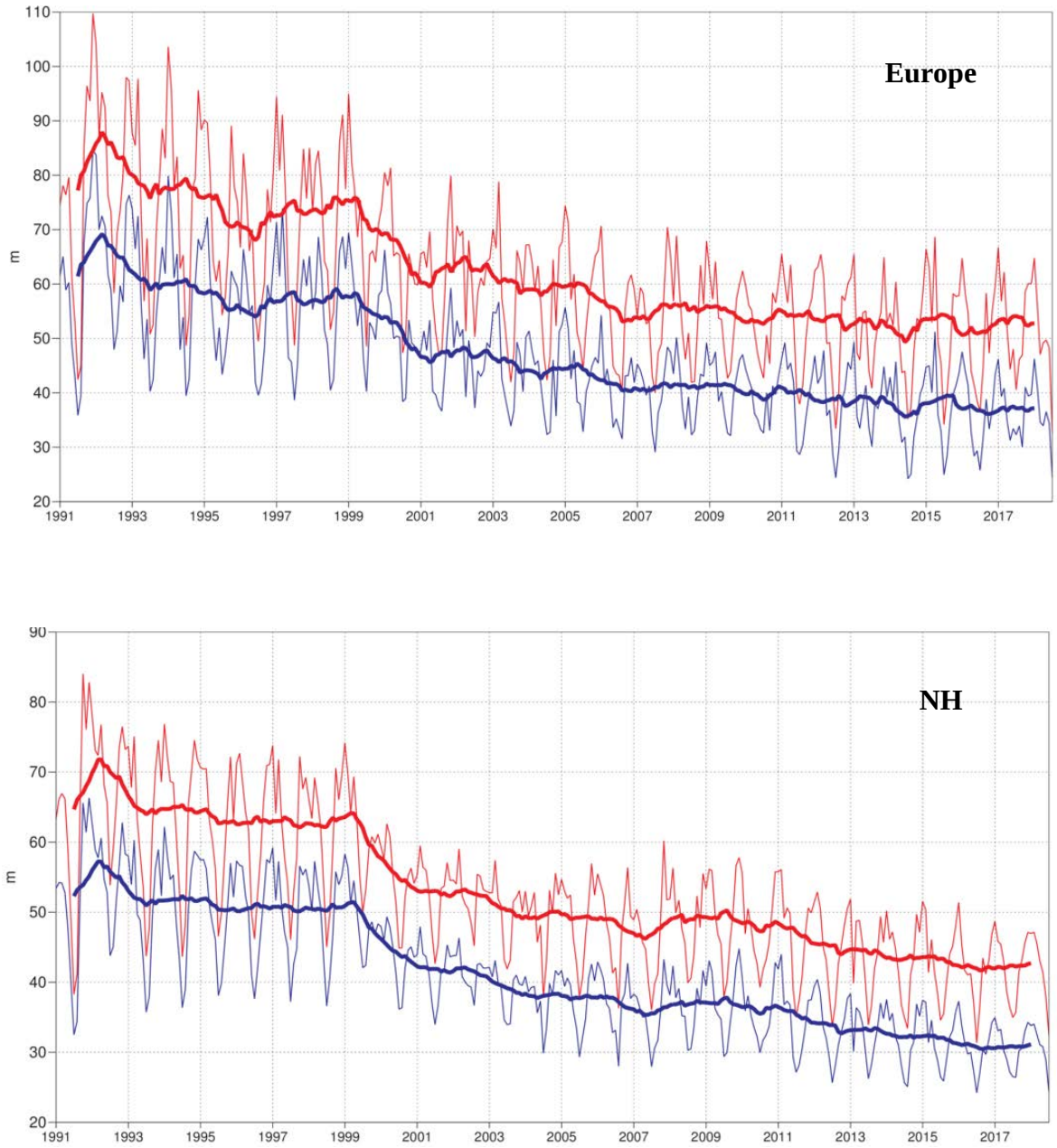


Figure 6: Consistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores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

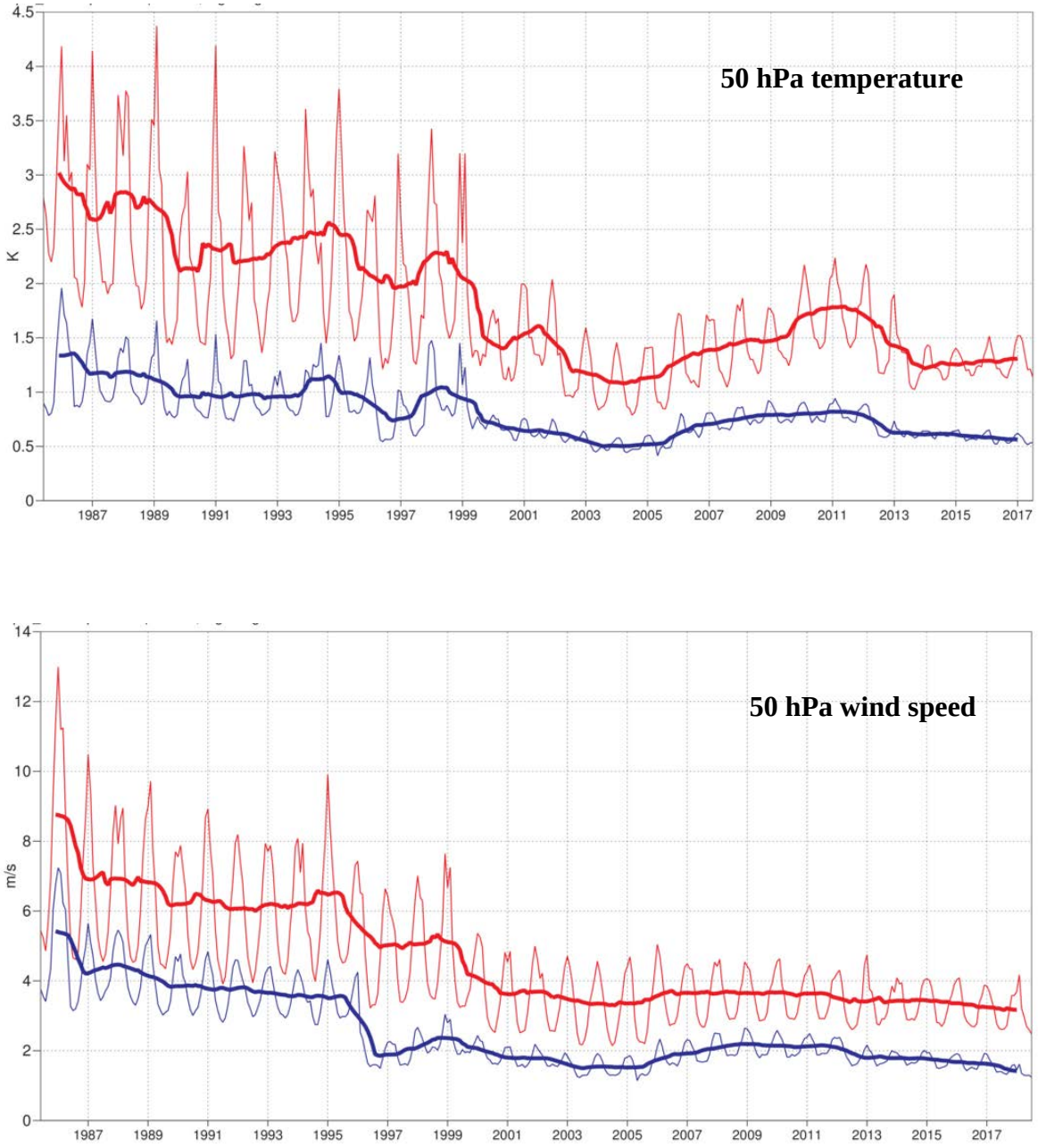


Figure 7: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also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

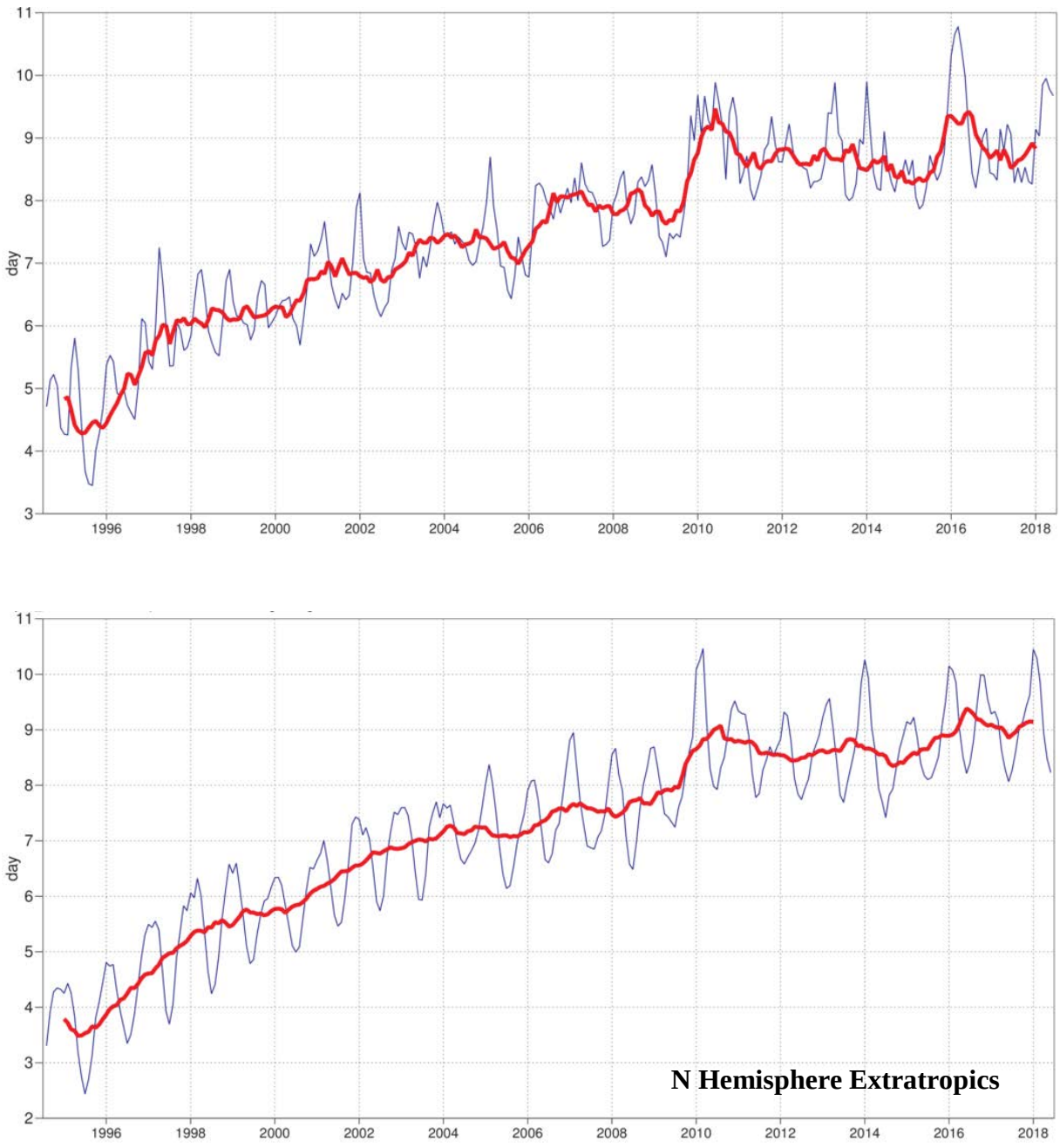


Figure 8: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

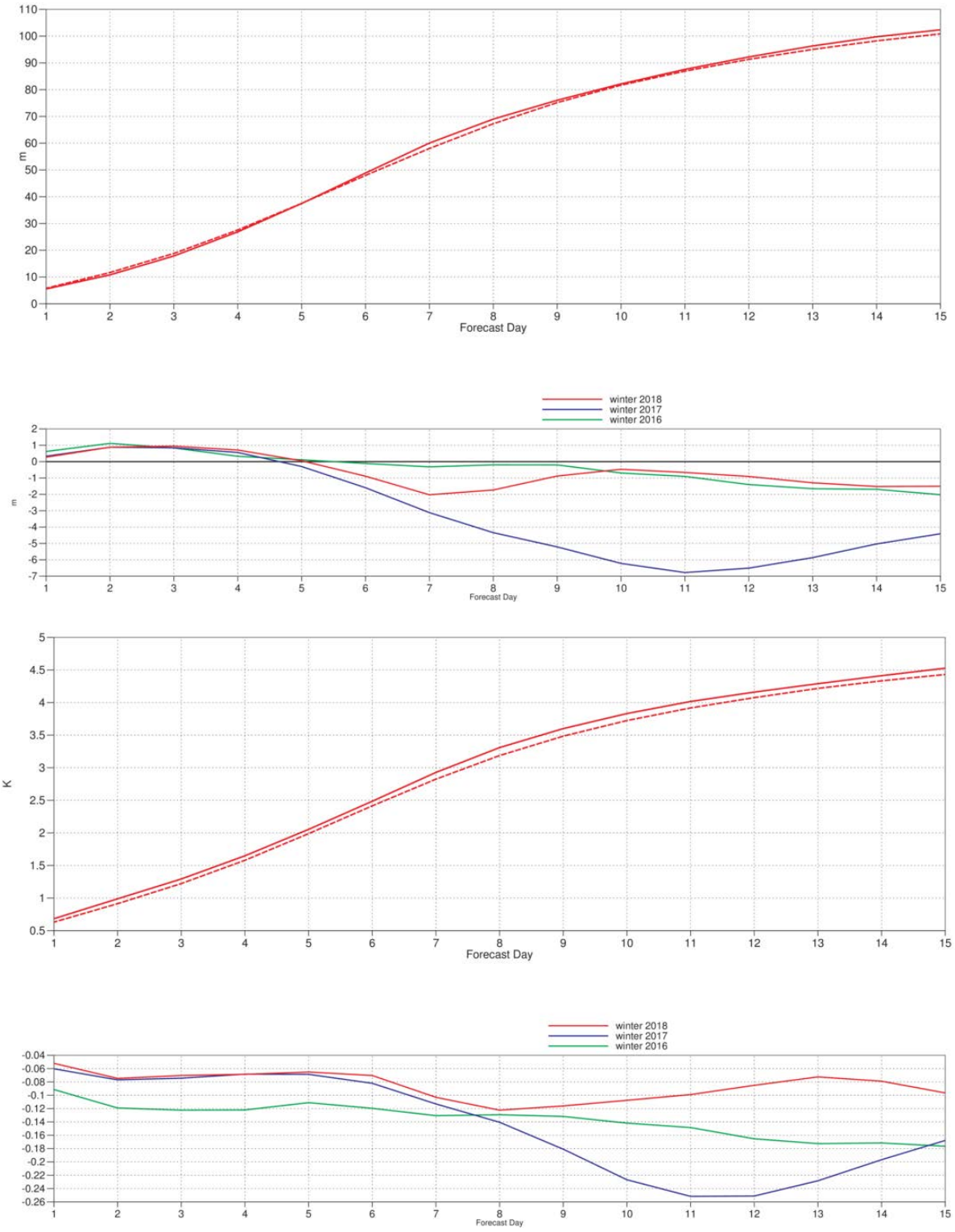


Figure 9: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2017–2018 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate 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.

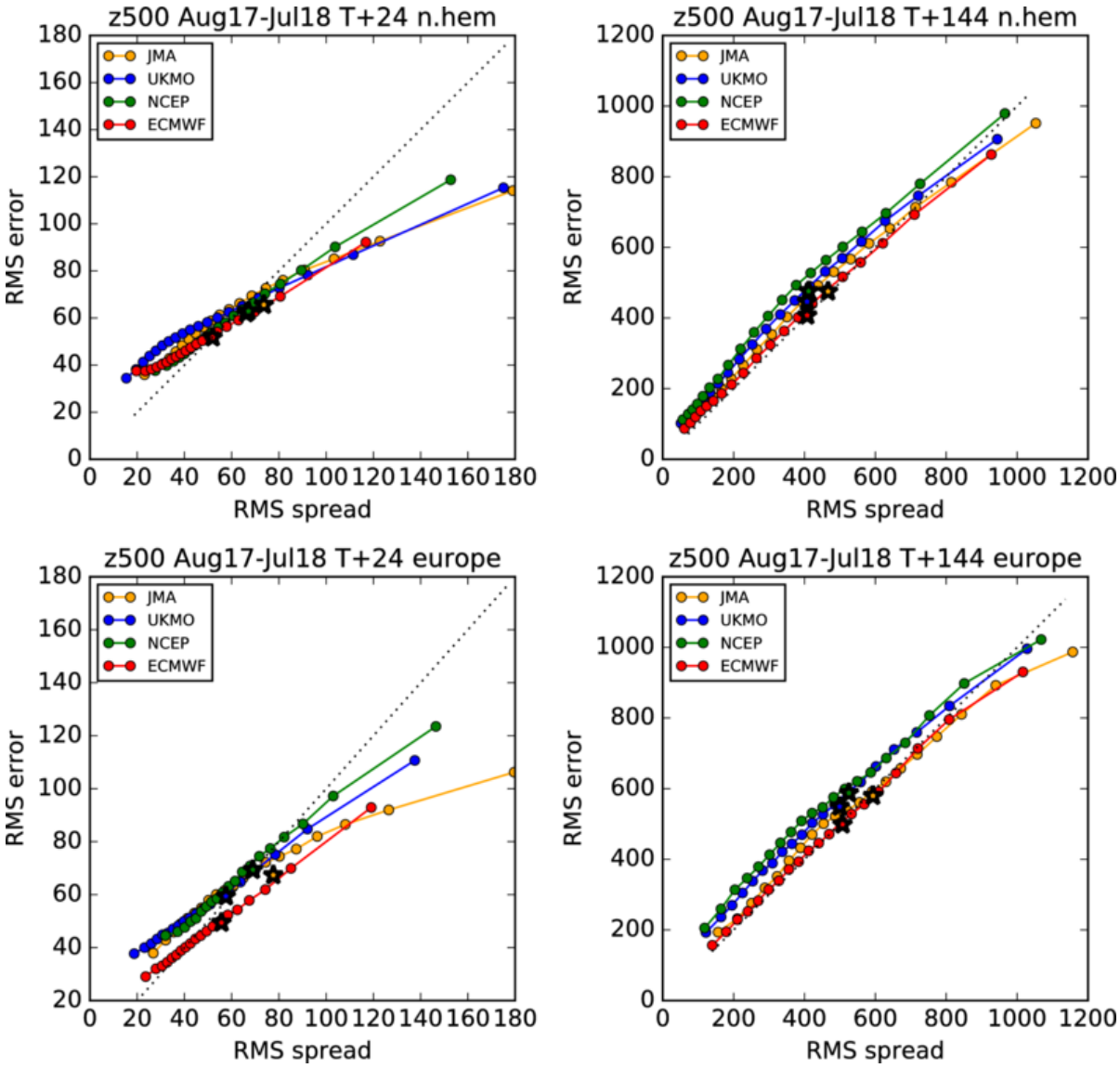


Figure 10: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2017–July 2018 in the northern hemisphere extra-tropics (top) and in Europe (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship. Due to random outages in the data supply NCEP curves are based on a reduced data set (60%).

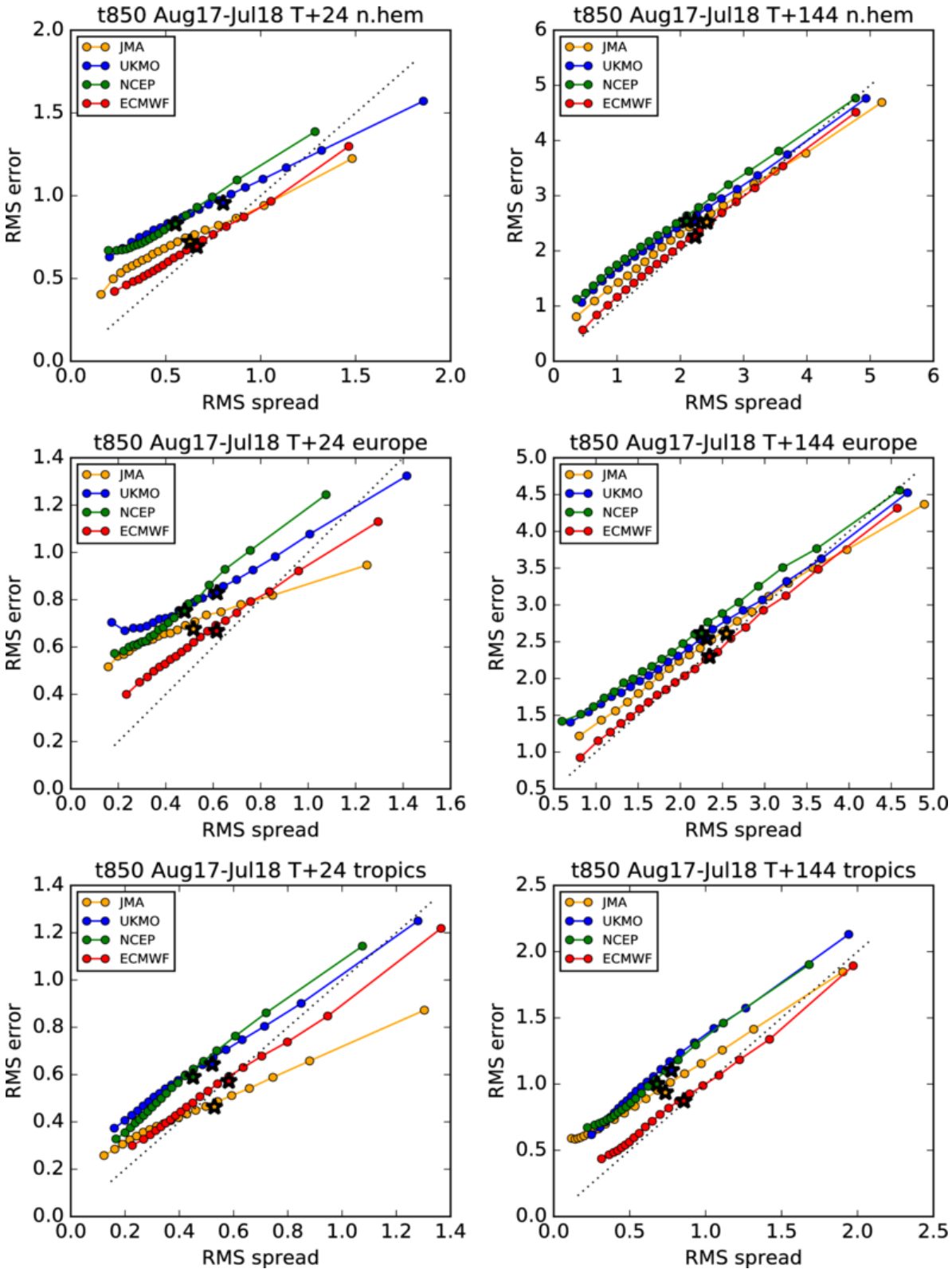


Figure 11: Ensemble spread reliability of different global models for 850 hPa temperature for the period August 2017–July 2018 in the northern hemisphere extra-tropics (top), Europe (centre), and the tropics (bottom) for day 1 (left) and day 6 (right), verified against analysis. Circles show error for 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%).

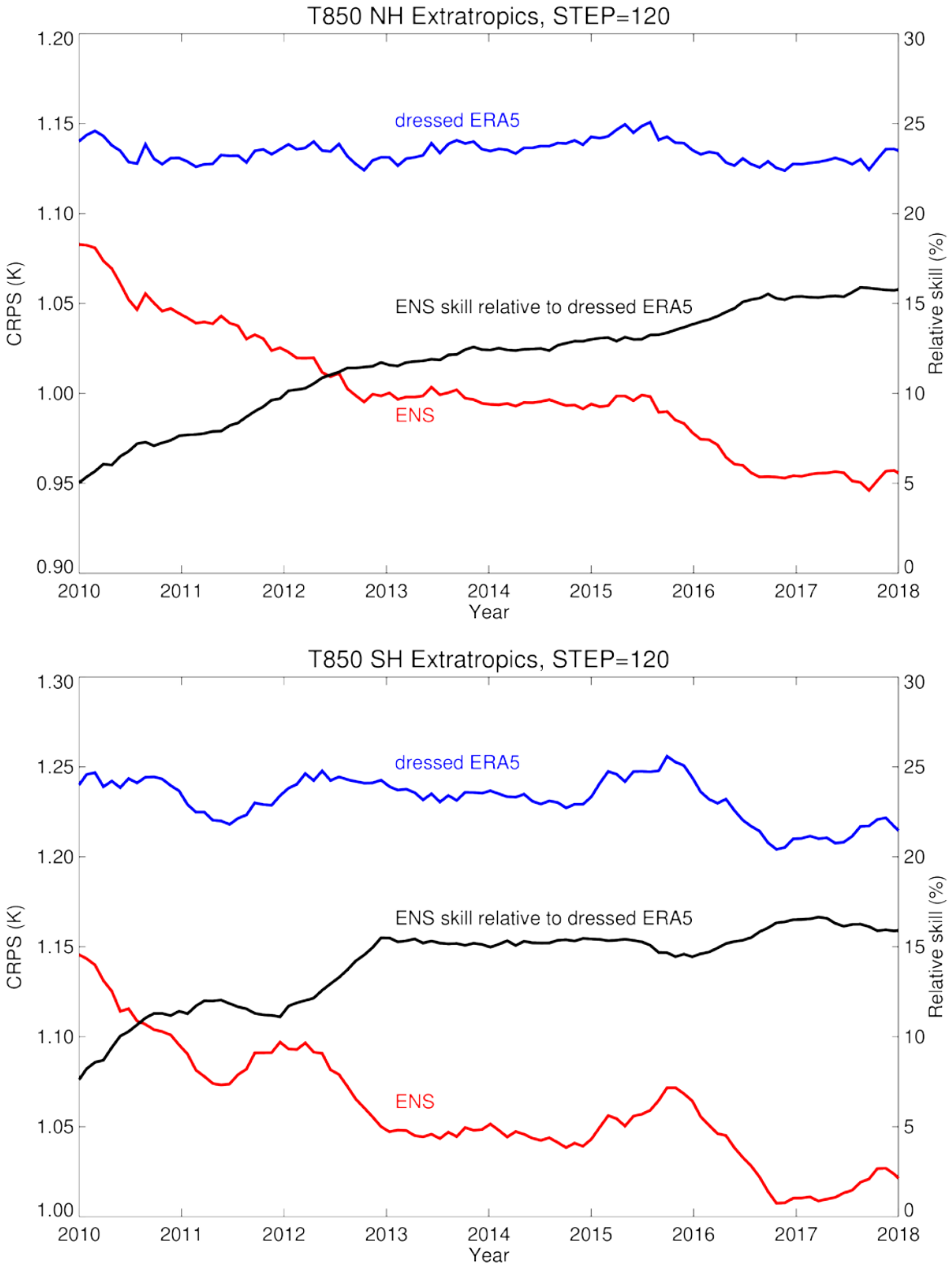


Figure 12: CRPS for temperature at 850 hPa in the northern (top) and southern (bottom) extratropics at day 5, verified against analysis. Scores are shown for the ensemble forecast (red) and the dressed ERA5 forecast (blue). Black curves show the skill of the ENS relative to the dressed ERA5 forecast.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

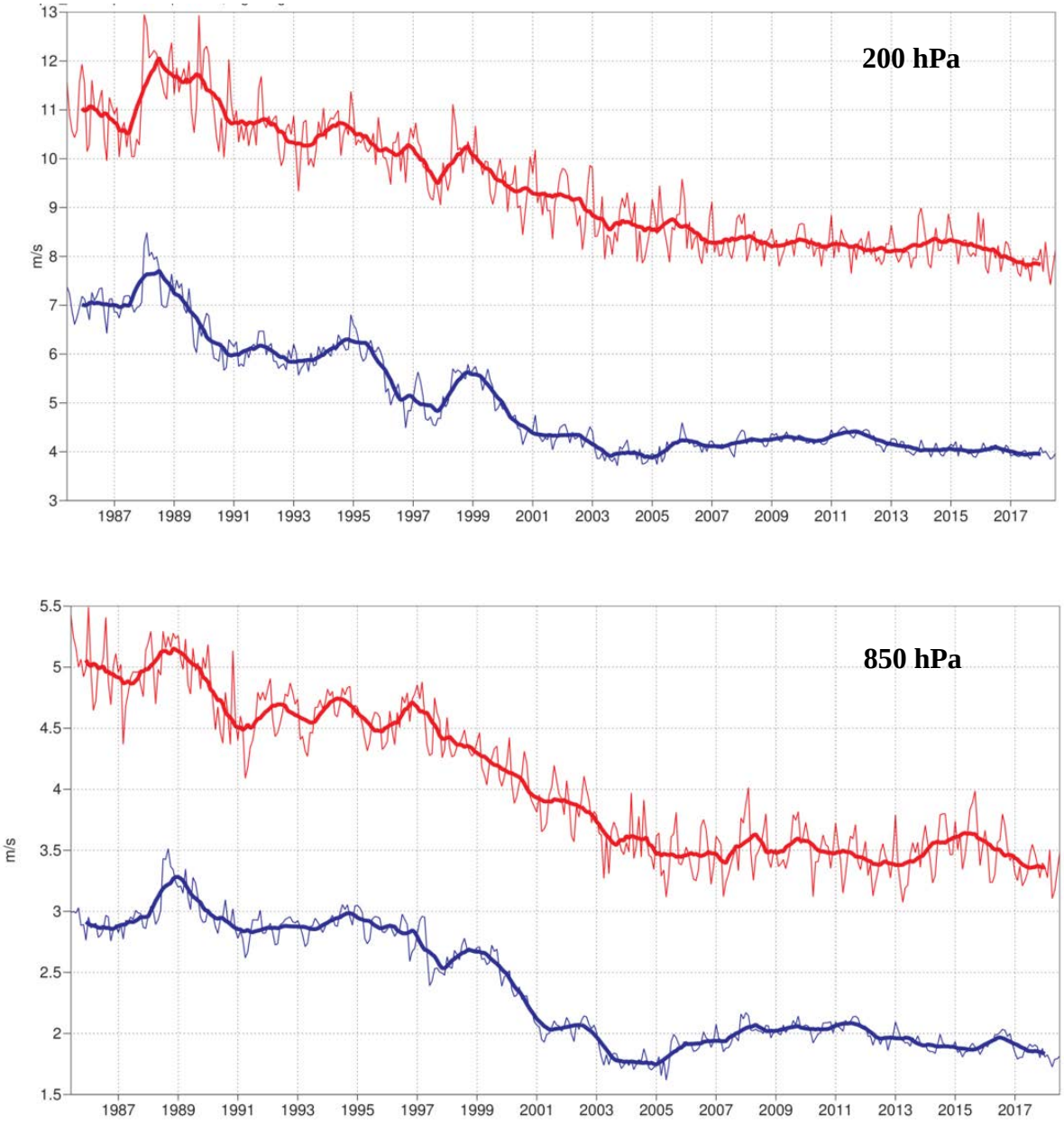


Figure 13: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

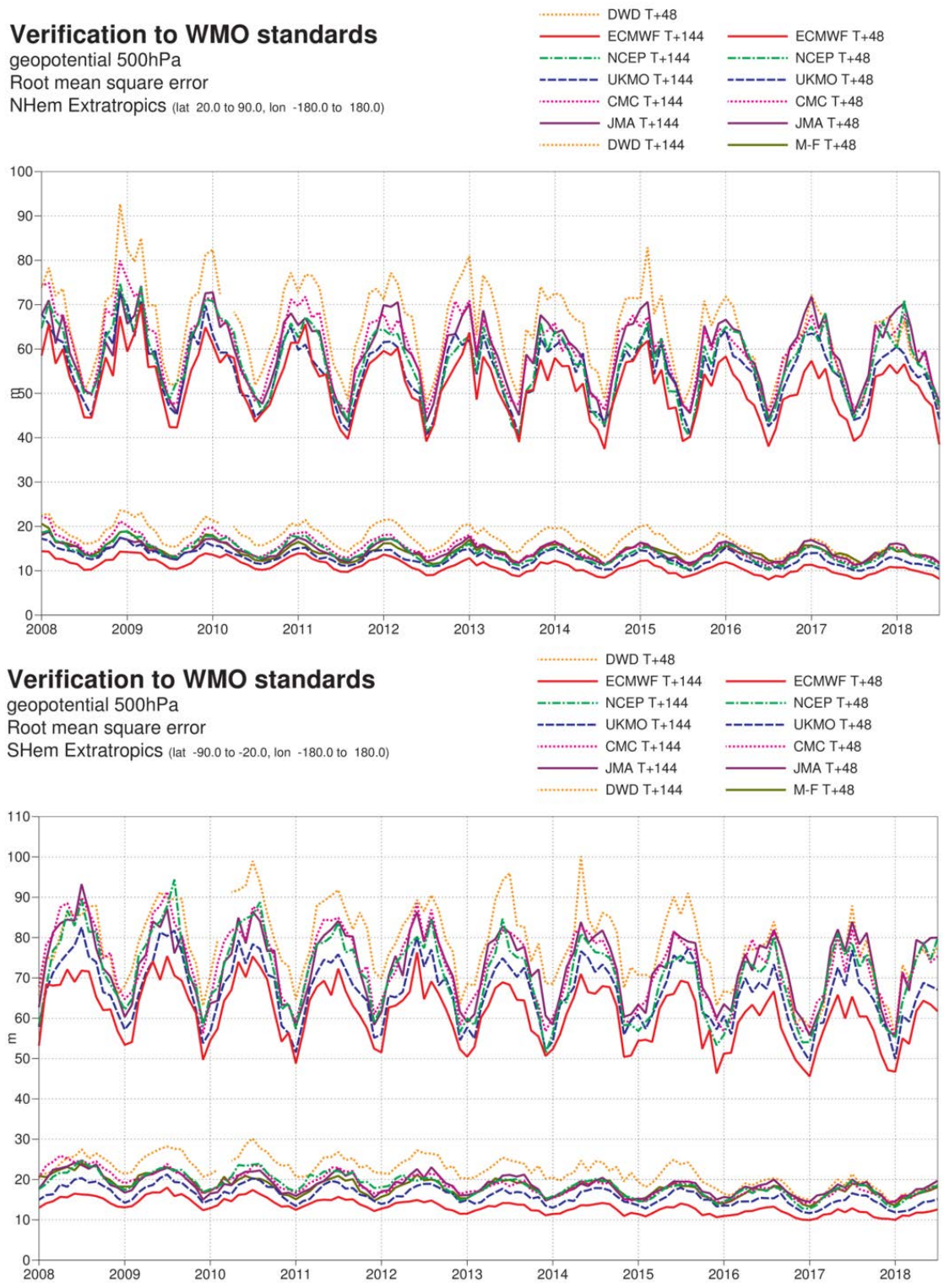


Figure 14: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top) and southern (bottom) extratropics. In each panel, the upper curves show the six-day forecast error and the lower curves show the two-day forecast error of model runs initiated at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

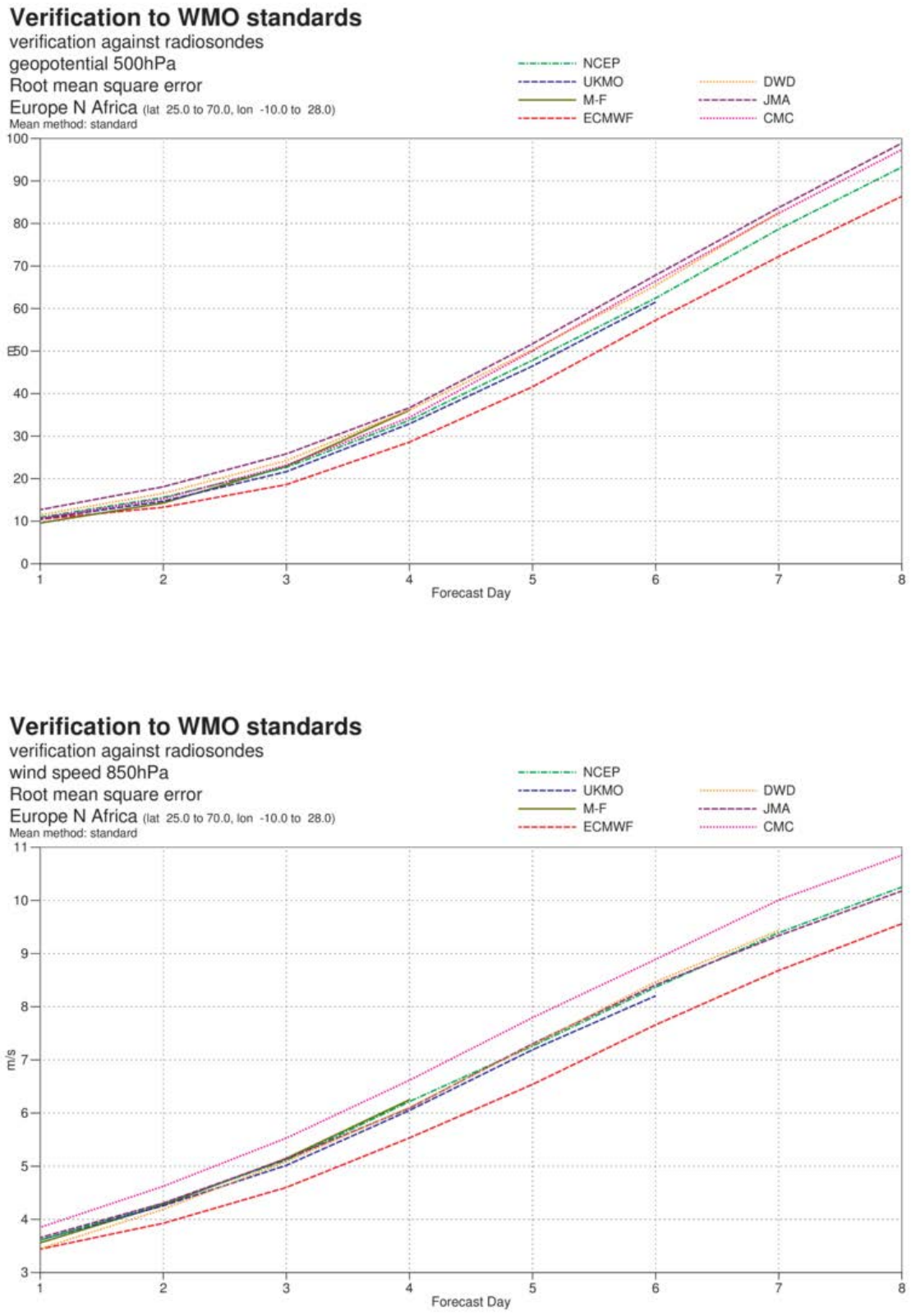
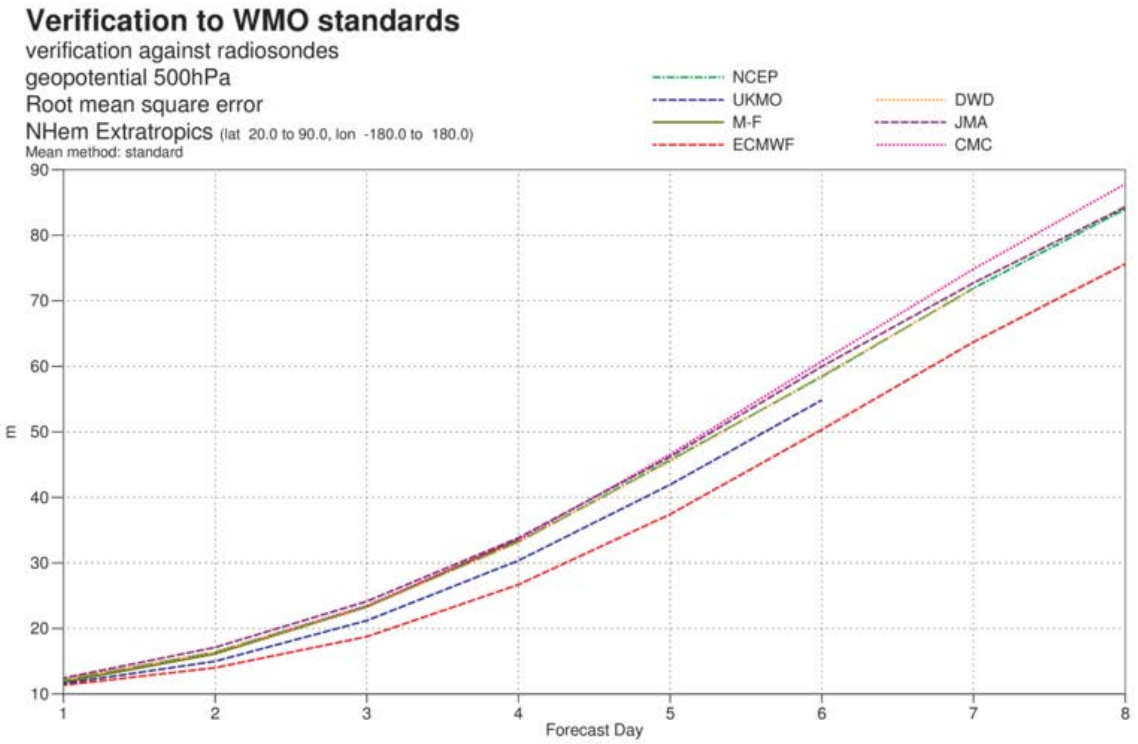


Figure 15: WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe (annual mean August 2017–July 2018) of forecast runs initiated at 12 UTC.



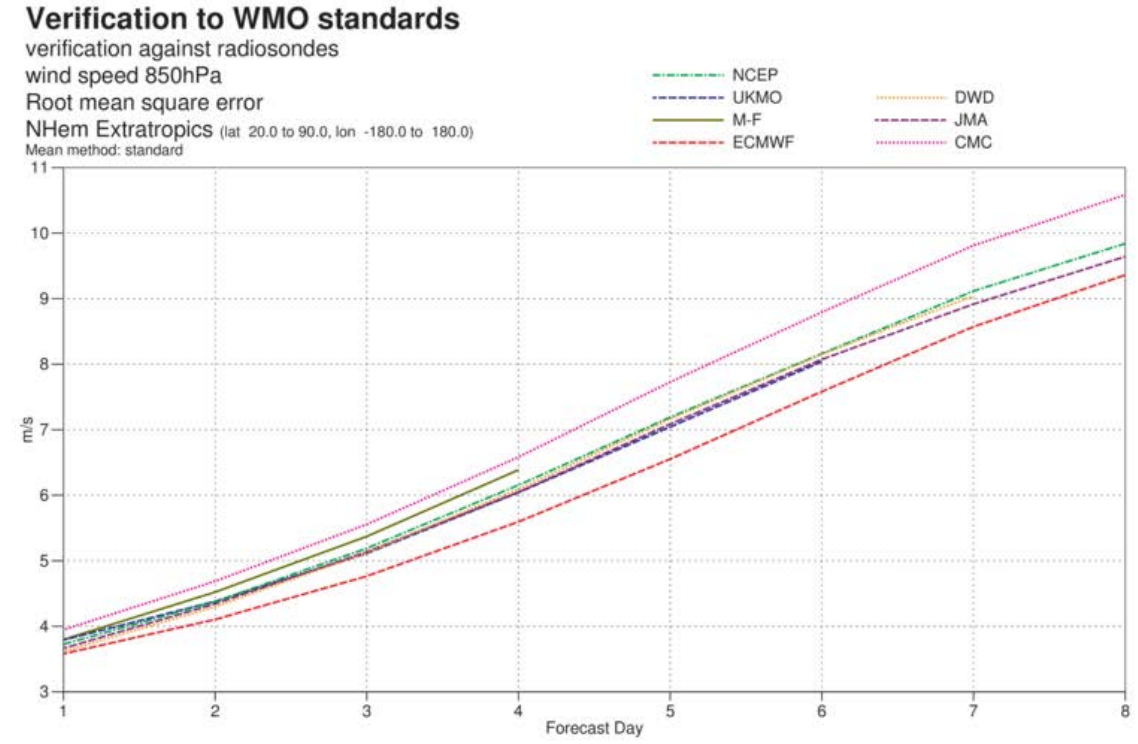
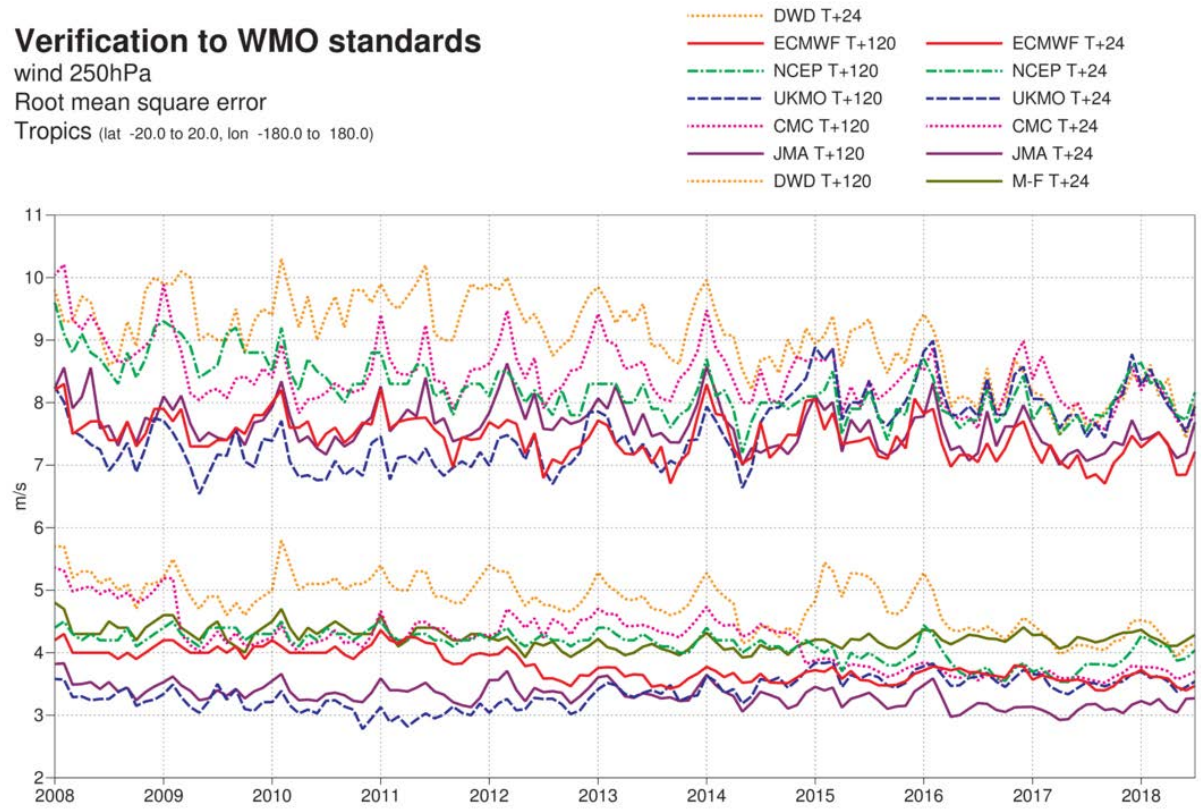


Figure 16: As Figure 15 for the northern hemisphere extratropics.



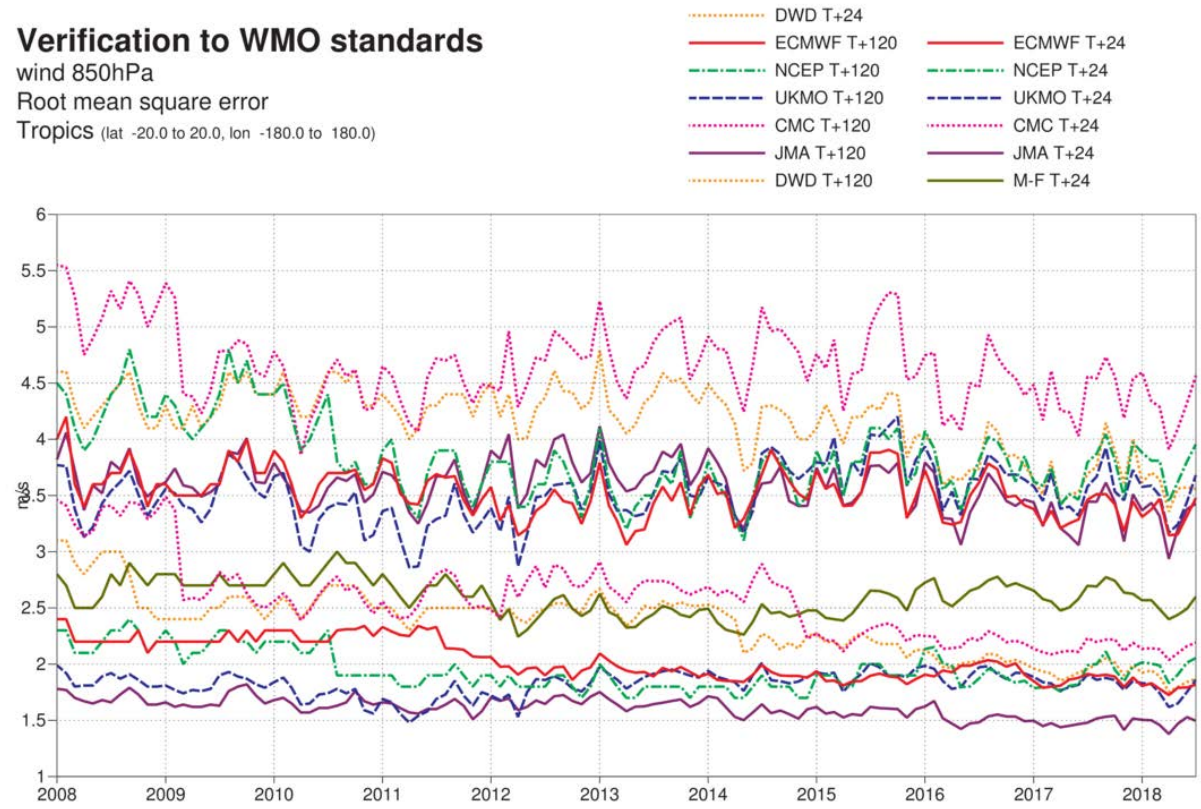
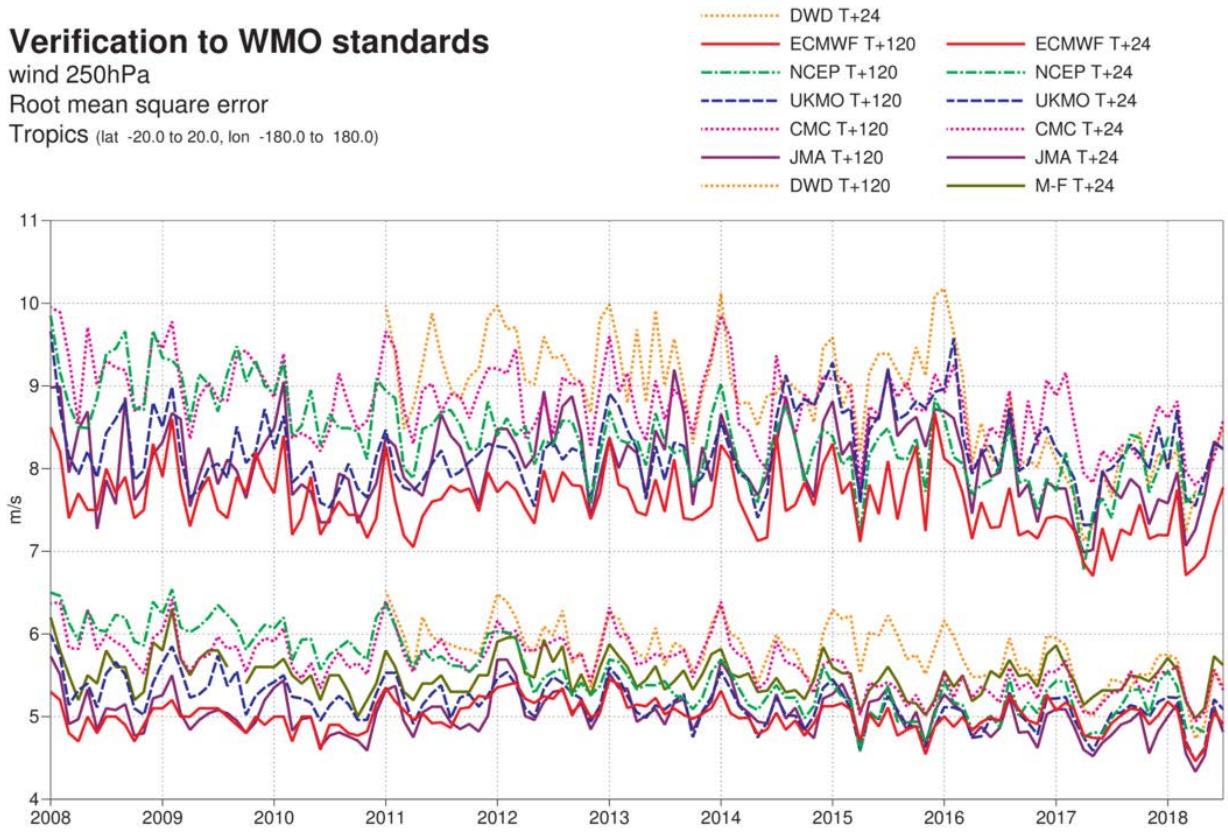


Figure 17: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top) and 850 hPa (bottom). In each panel, the upper curves show the five-day forecast error and the lower curves show the one-day forecast error of forecast runs initiated at 12 UTC. Each model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



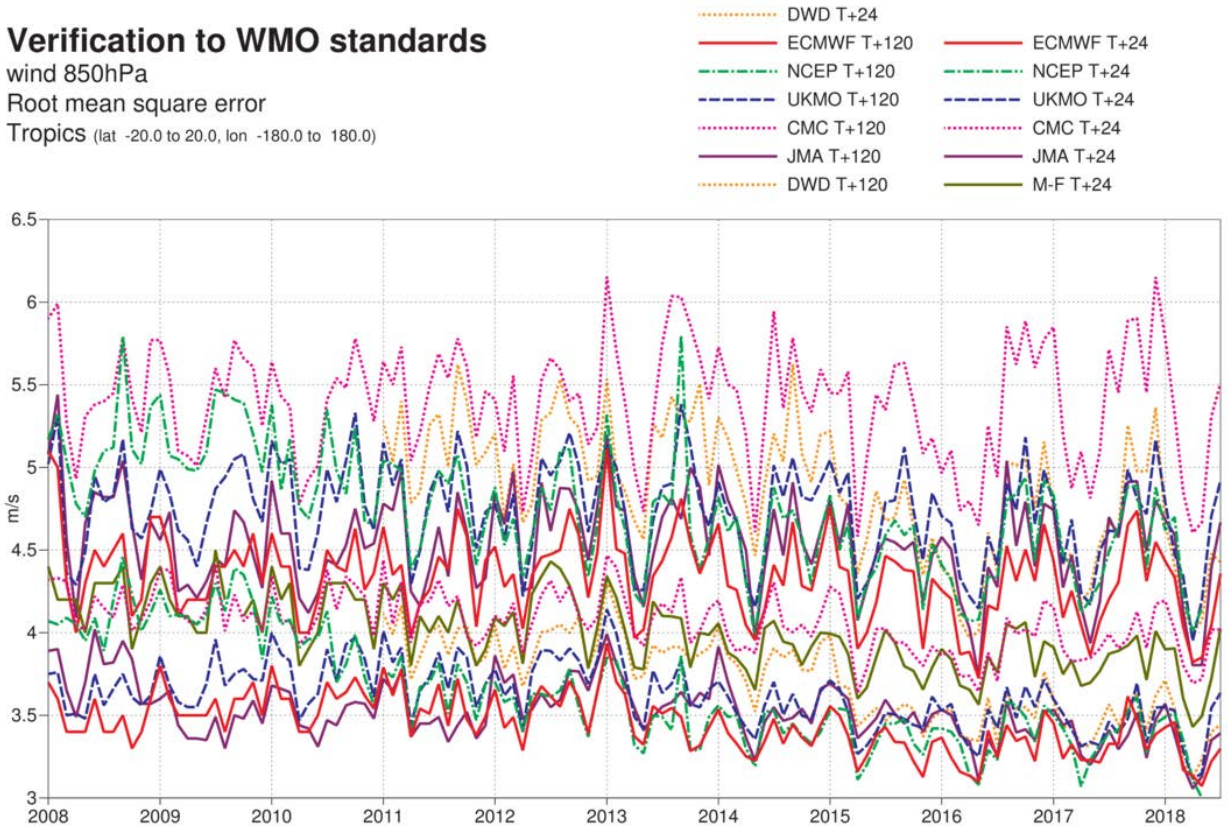


Figure 18: As Figure 17 for verification against radiosonde observations.

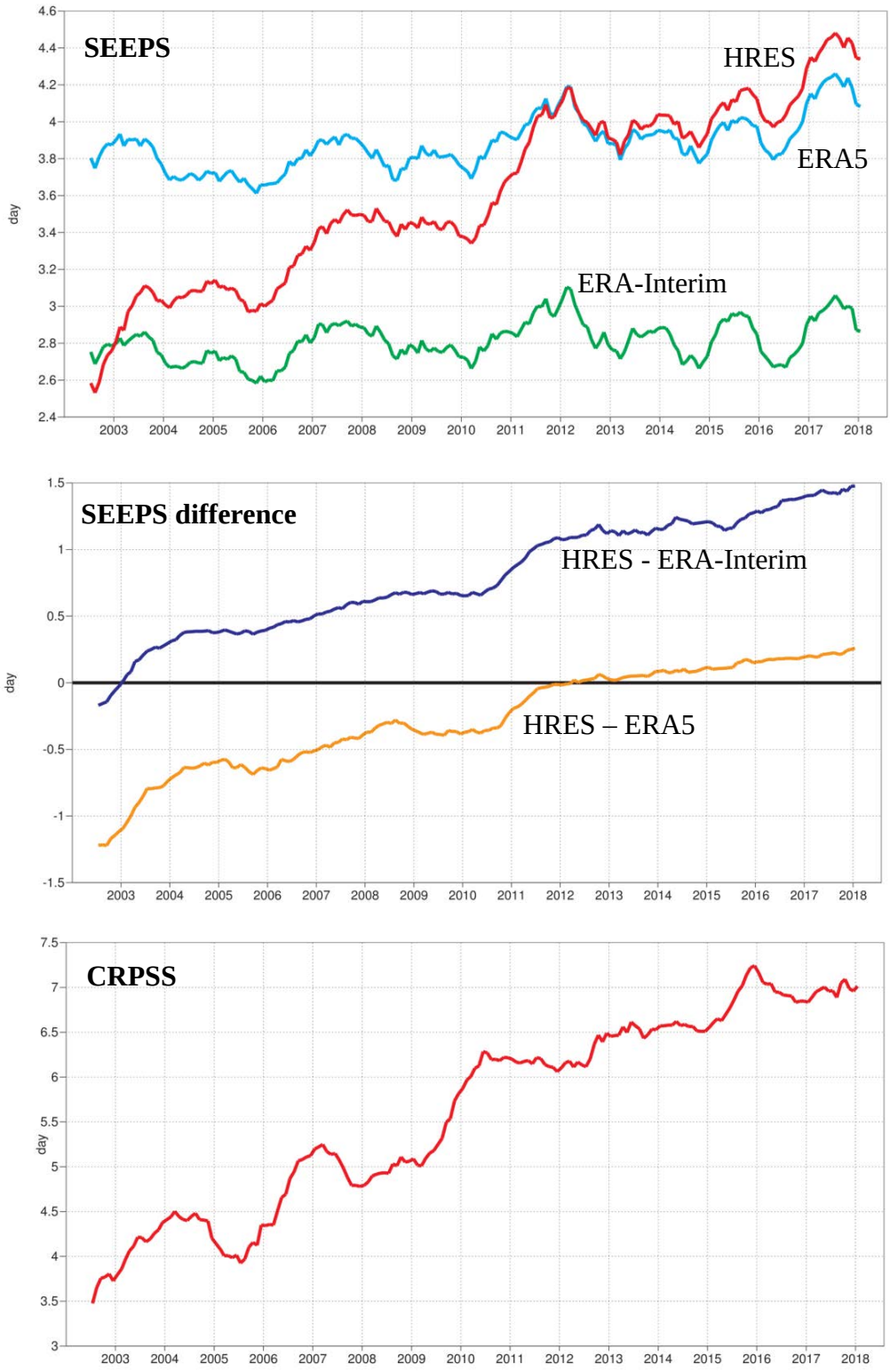


Figure 19: Supplementary headline scores (red) for deterministic (top, centre) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number of days for which the centred 12-month 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).

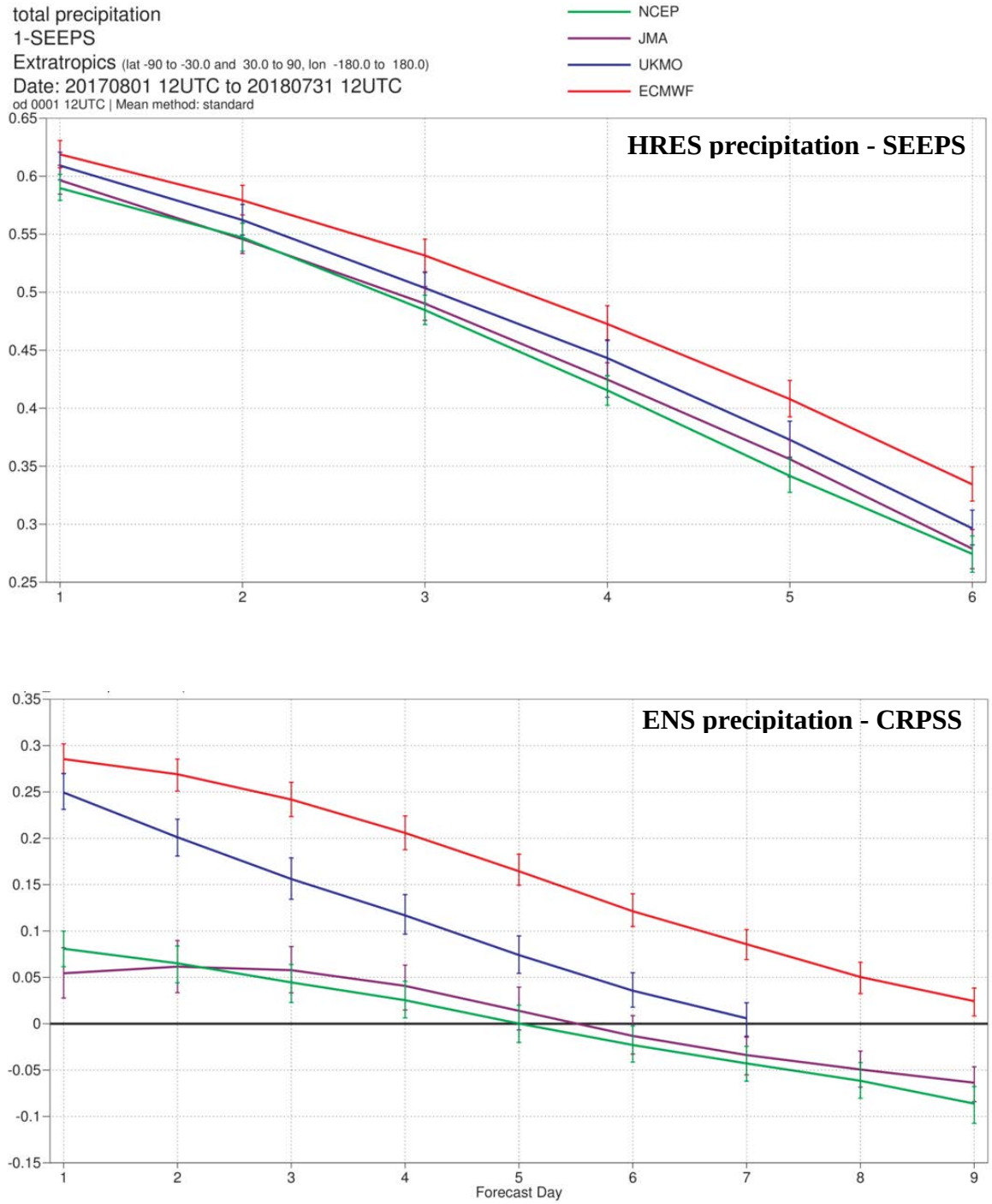


Figure 20: Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 19. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2017–July 2018. Bars indicate 95% confidence intervals.

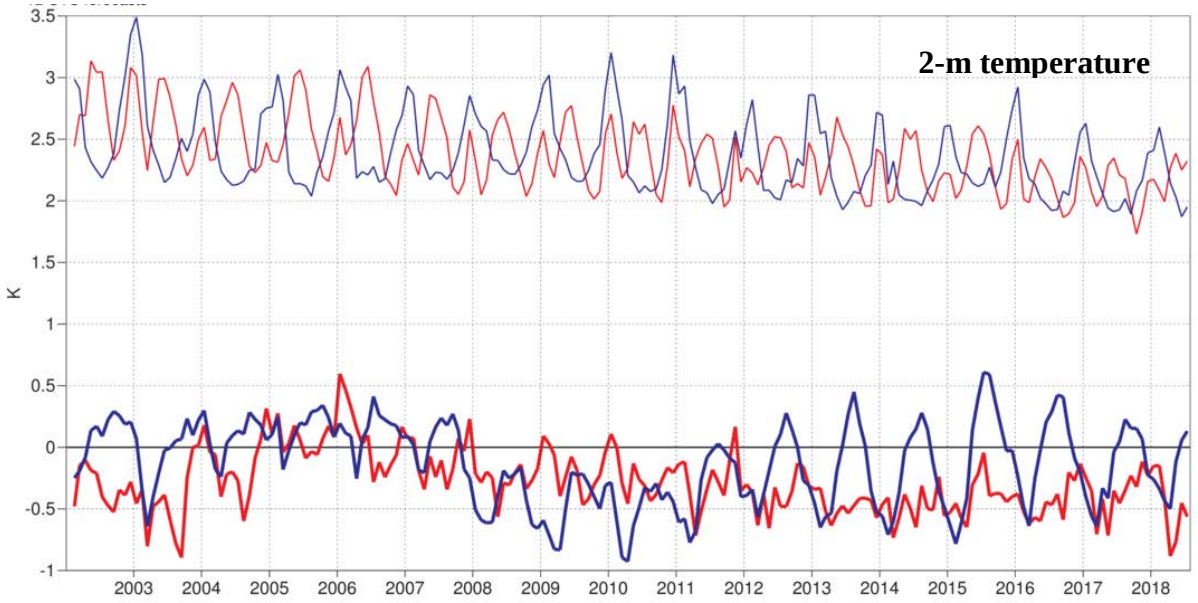


Figure 21: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

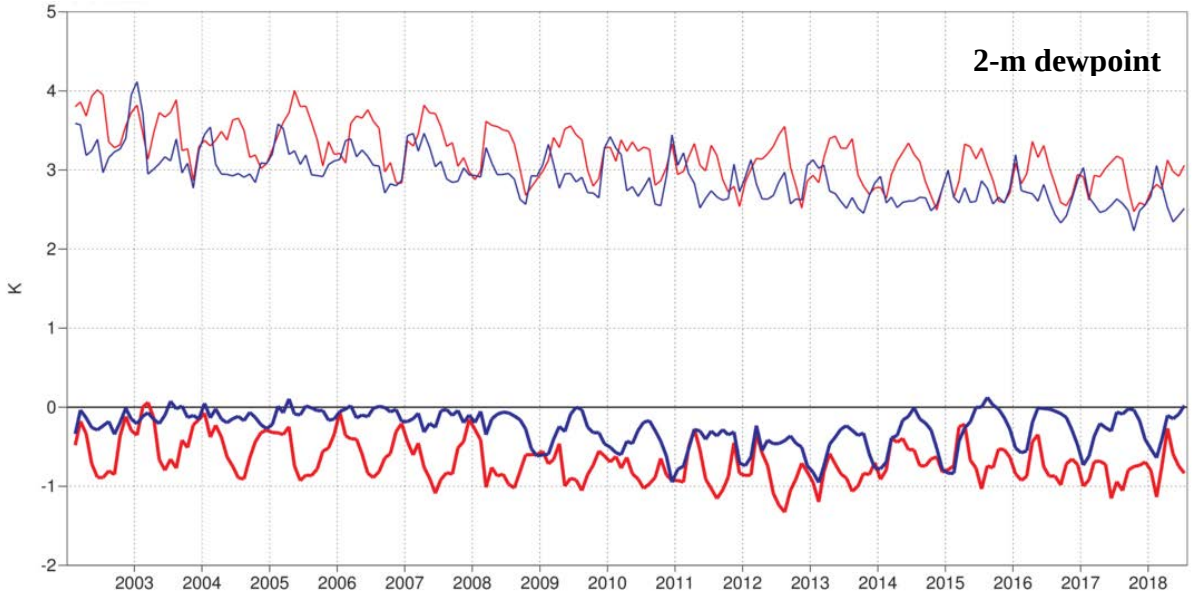


Figure 22: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

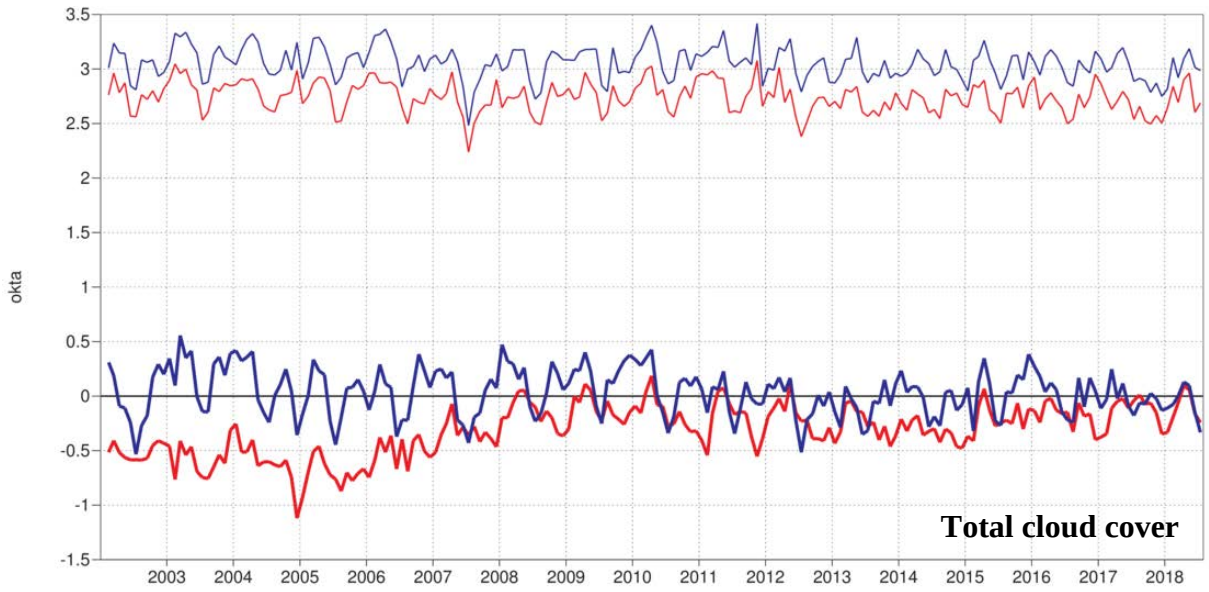


Figure 23: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

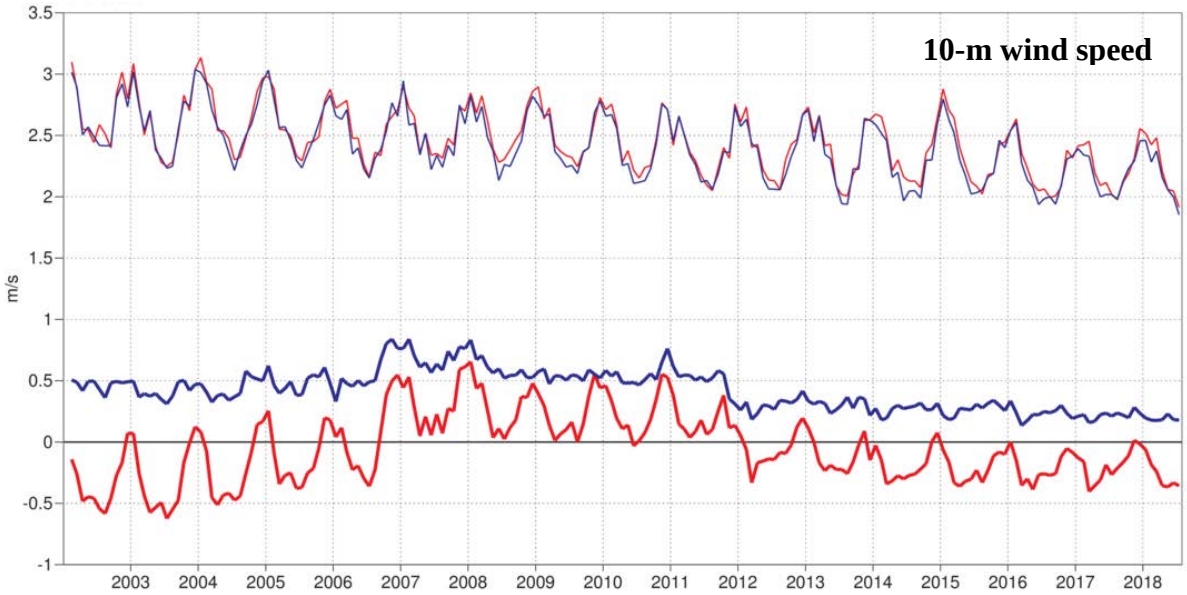


Figure 24: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

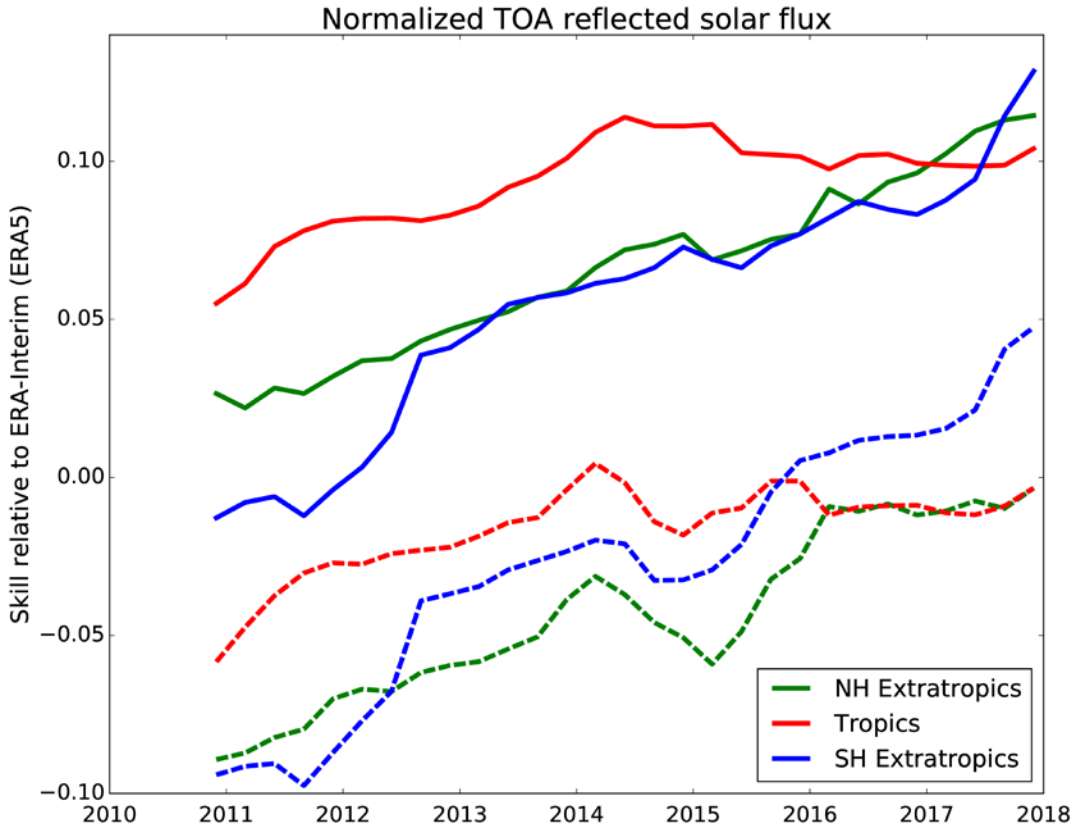


Figure 25: 12-month running average of the day 3 forecast skill relative to ERA-Interim (continuous curves) and relative to ERA5 (dashed) of normalized TOA reflected solar flux (daily totals), verified against satellite data. The verification has been carried out for those parts of the northern hemisphere 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).

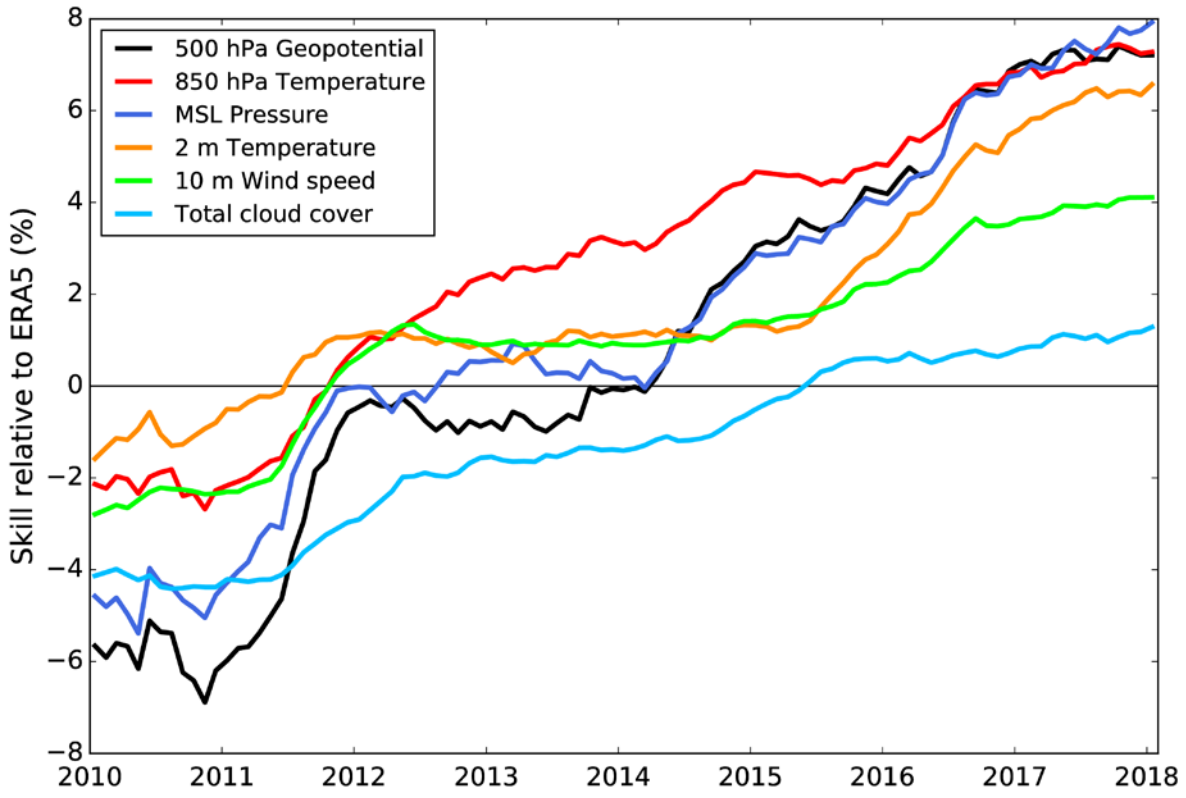


Figure 26: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA5. Verification is against analysis for 500 hPa geopotential (Z500), 850 hPa temperature (T850), and mean sea level pressure (MSLP), using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature (T2M), 10 m wind speed (V10), and total cloud cover (TCC).

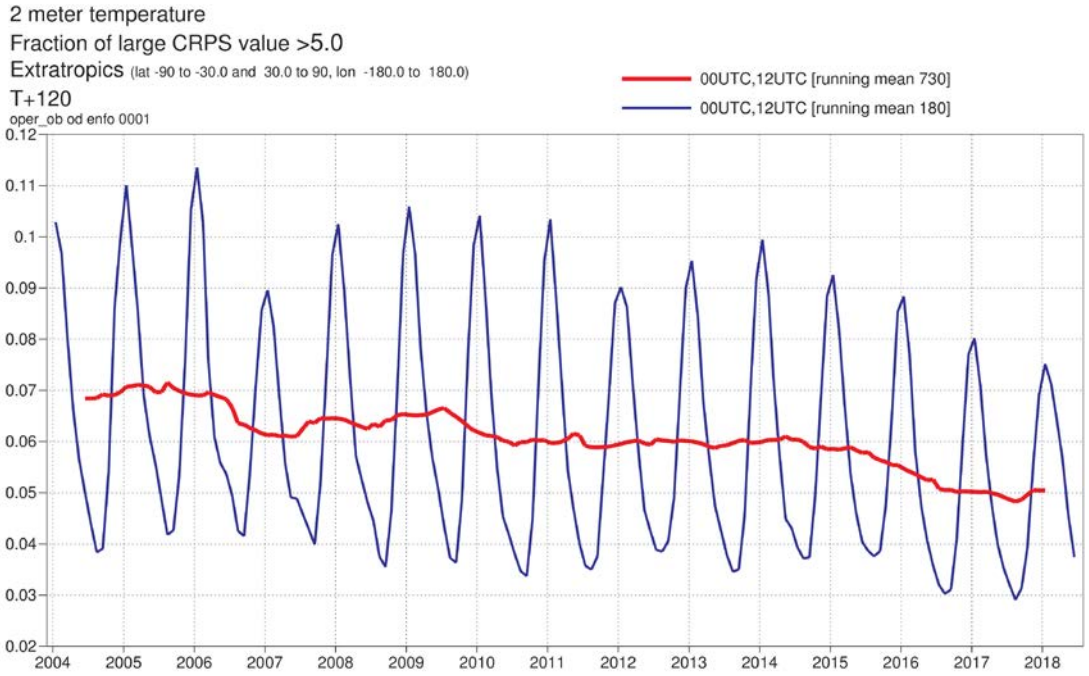


Figure 27: Evolution of the fraction of large 2m temperature errors (CRPS>5K) in the ENS at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

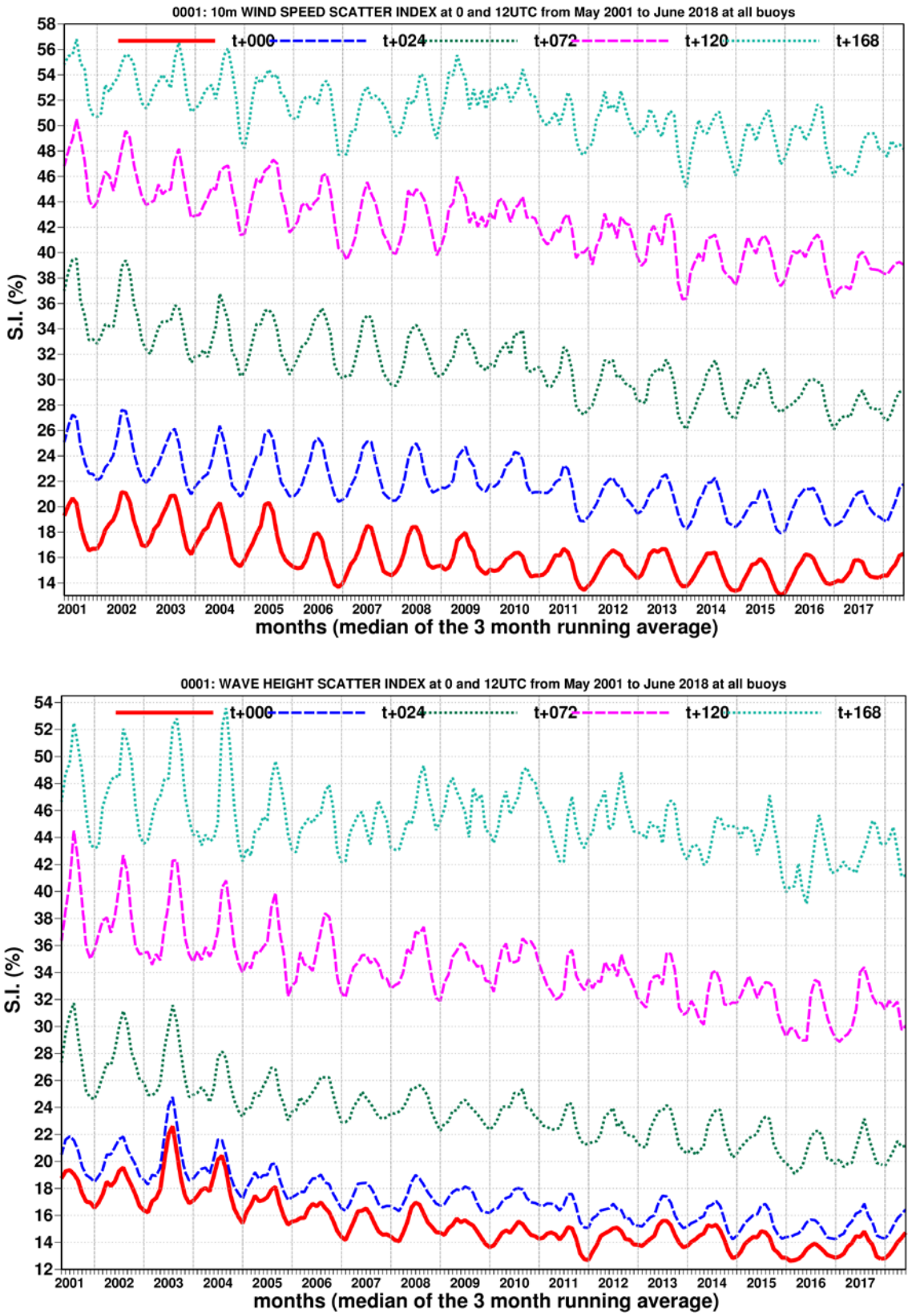


Figure 28: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running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

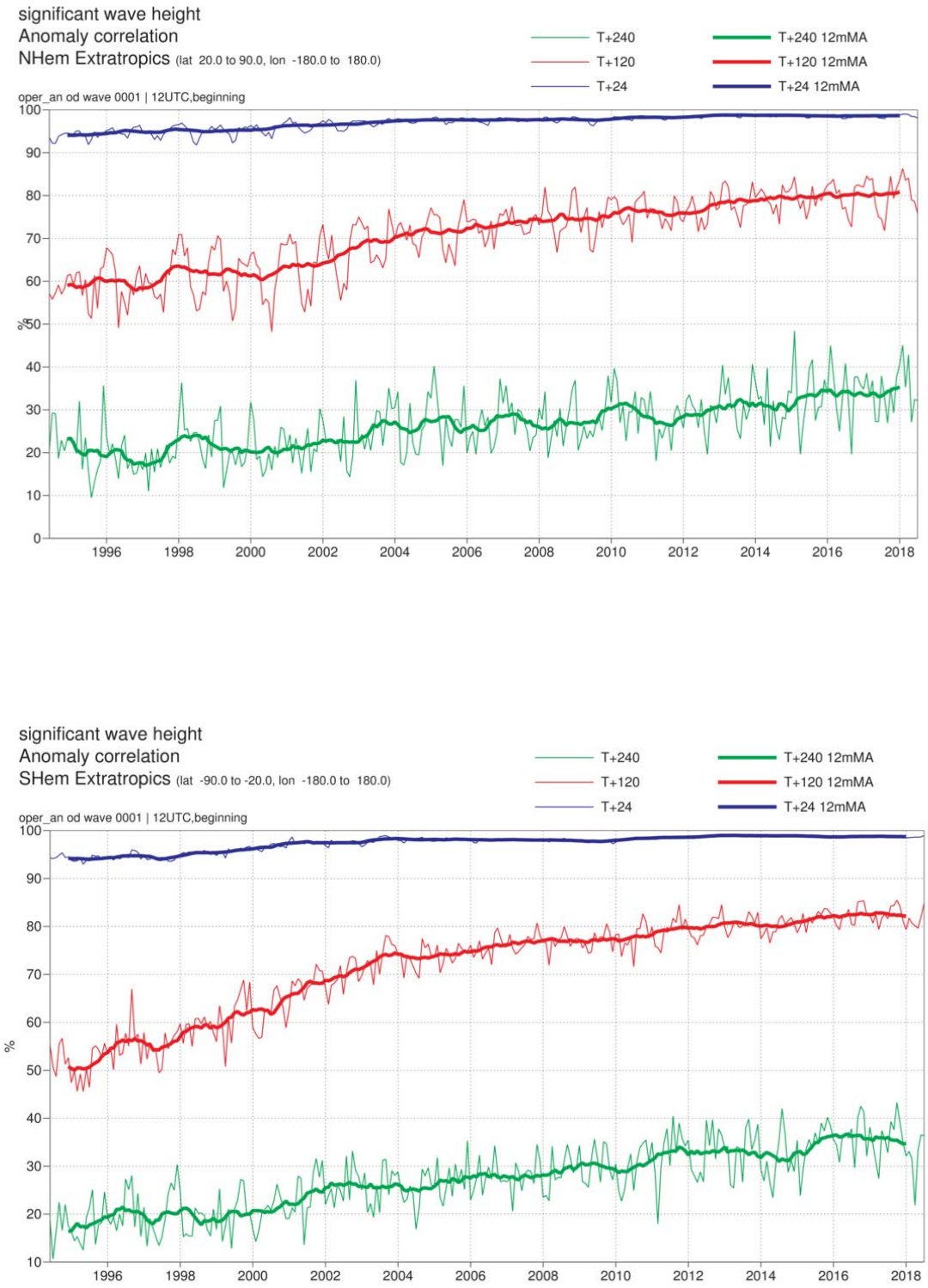


Figure 29: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

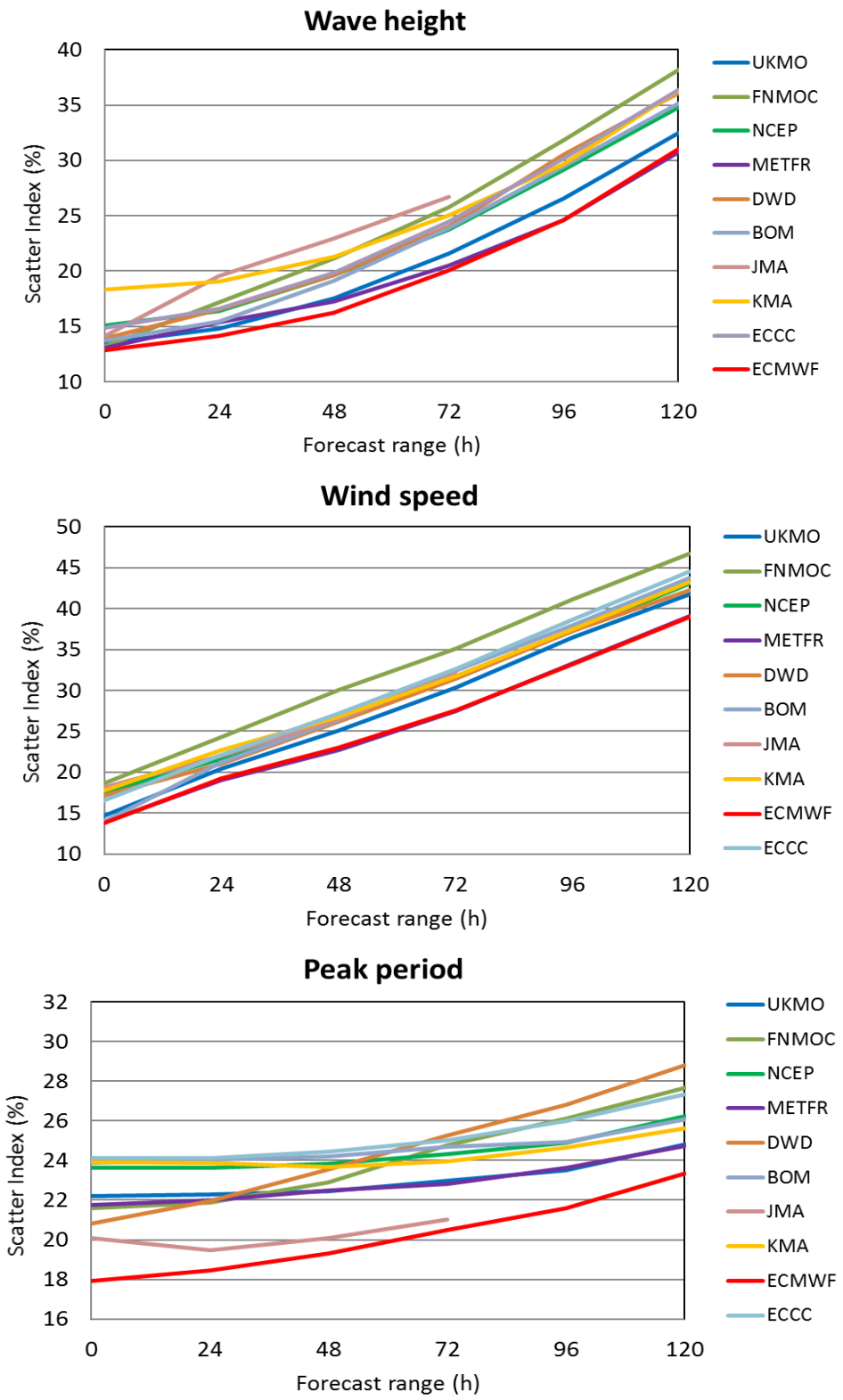


Figure 30: Verification of forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 12-month period June 2017–May 2018. 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; ECCC: Environment and Climate Change Canada.

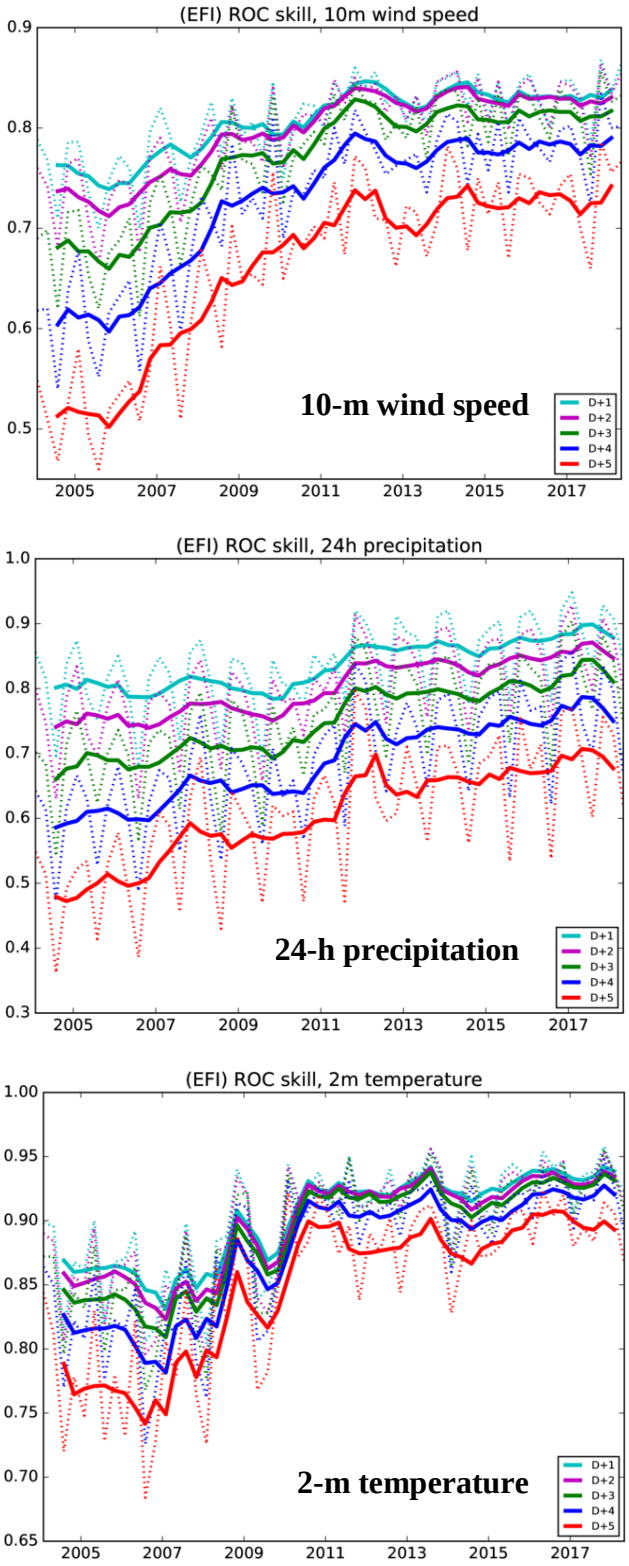


Figure 31: Verification of Extreme Forecast Index (EFI) against analysis. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 5 (24-hour period 96–120 hours ahead); skill at day 4 (blue line) is the supplementary headline score; an extreme event is taken as an observation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

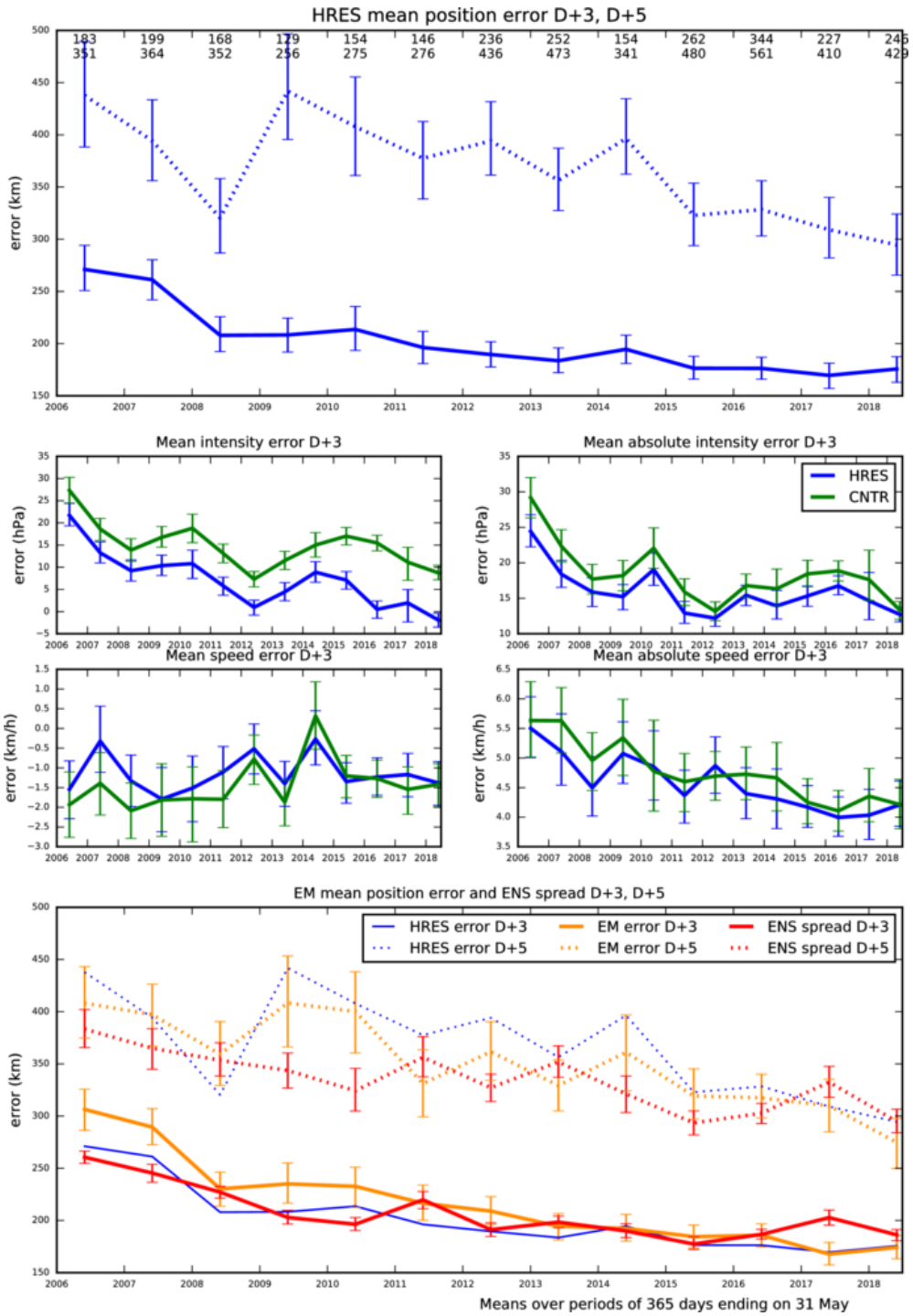


Figure 32: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary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

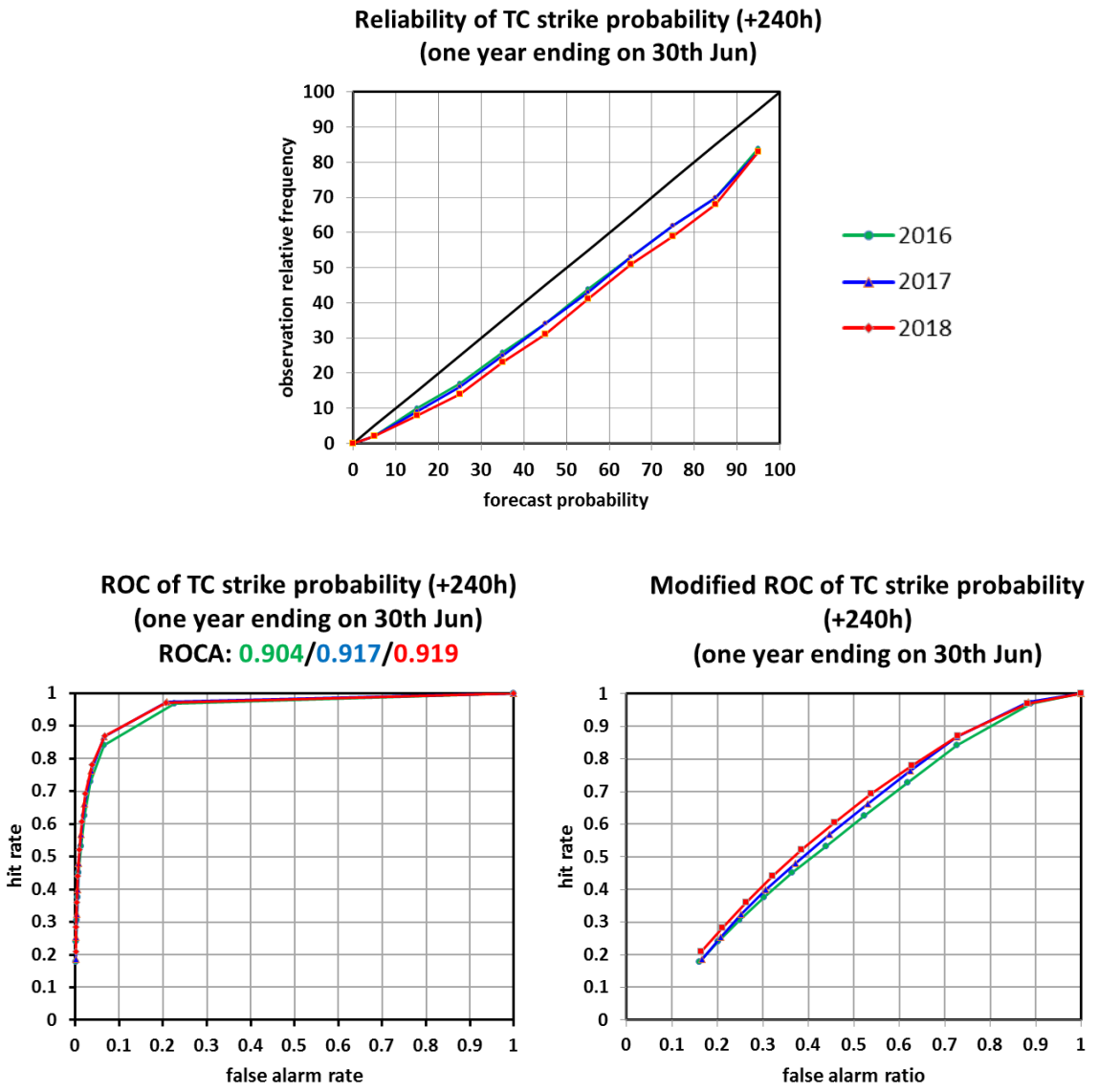


Figure 33: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2015–June 2016 (green), July 2016–June 2017 (blue) and July 2017–June 2018 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

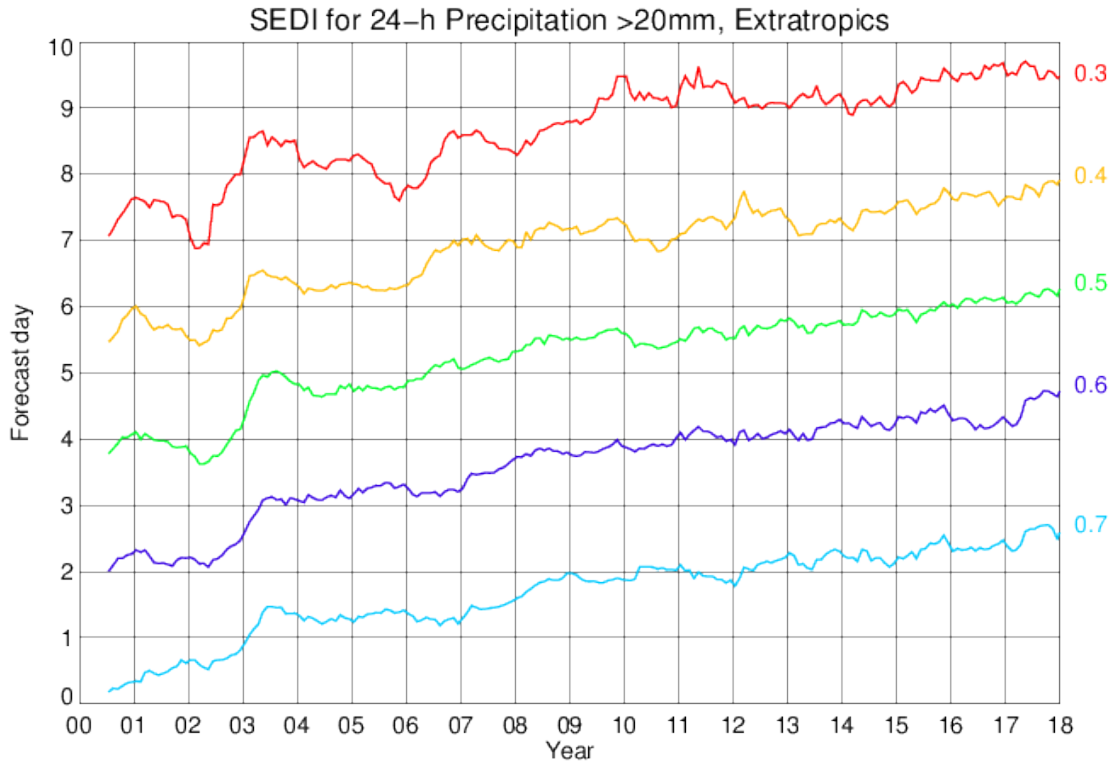


Figure 34: Evolution of skill of the HRES forecast in predicting 24-h precipitation amounts >20 mm in the extra-tropics as measured by the SEDI score, expressed in terms of forecast days. Verification is against SYNOP observations. Numbers on the right indicate different SEDI thresholds used. Curves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

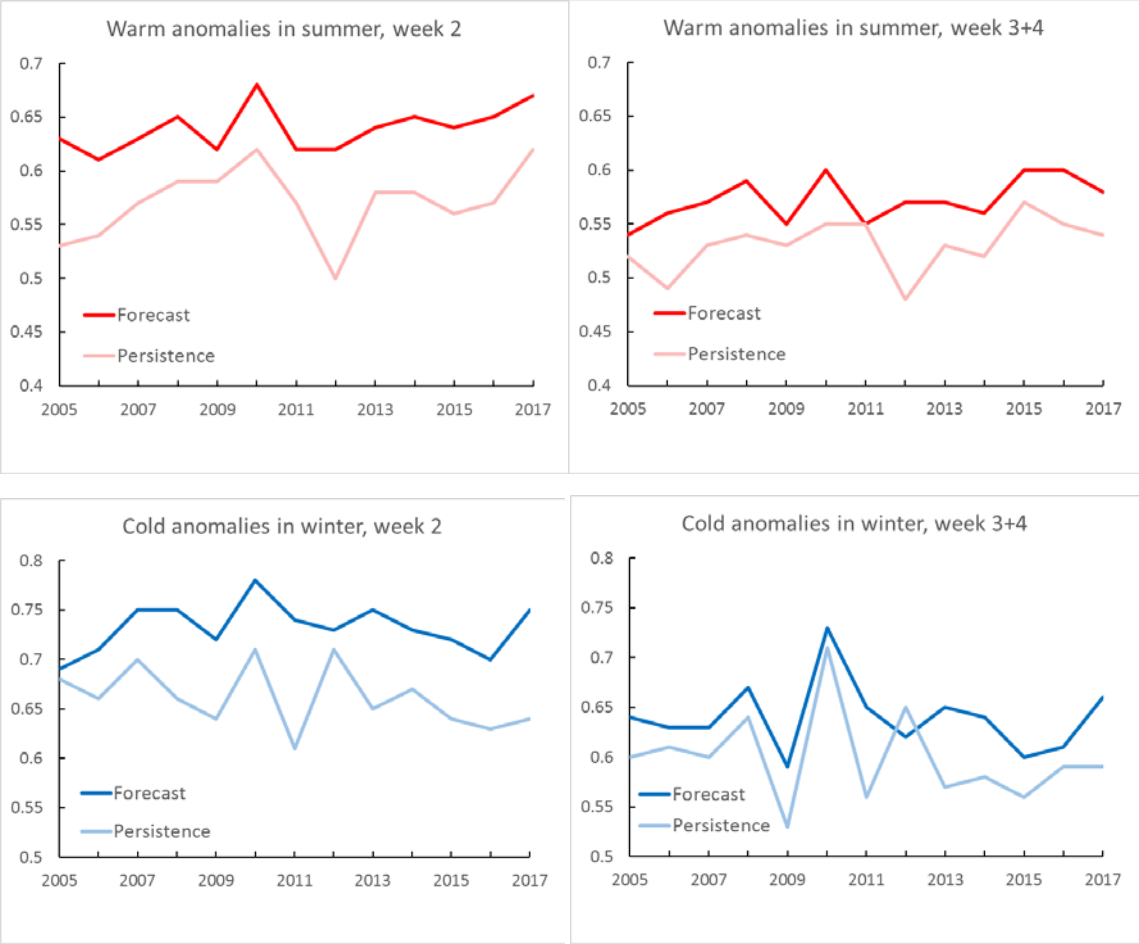


Figure 35: Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all land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

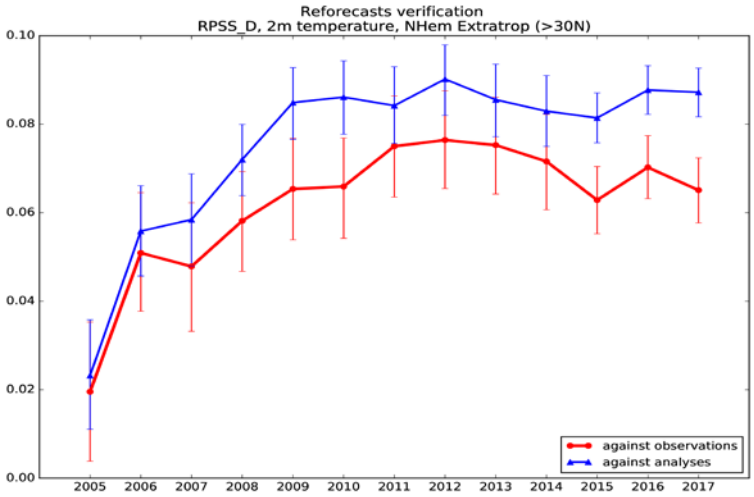


Figure 36: Skill of the ENS in predicting weekly mean 2m temperature anomalies (terciles) in week 3 in the northern extratropics. Verification against own analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.

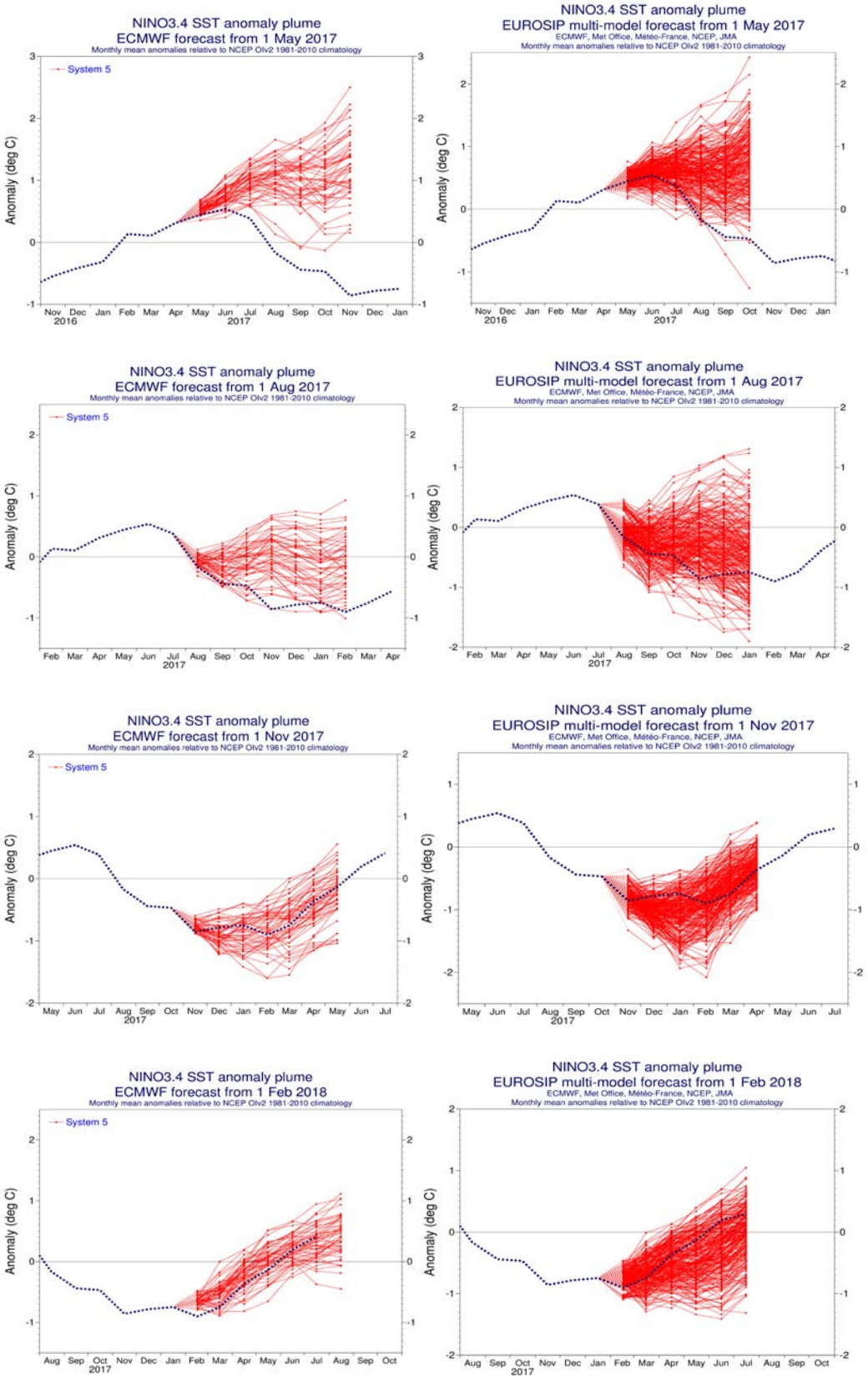


Figure 37: ECMWF (left column) and EUROSIP multi-model forecast (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2017, August 2017, November 2017 and February 2018. The red lines represent the ensemble members; dotted blue line shows the subsequent verification.

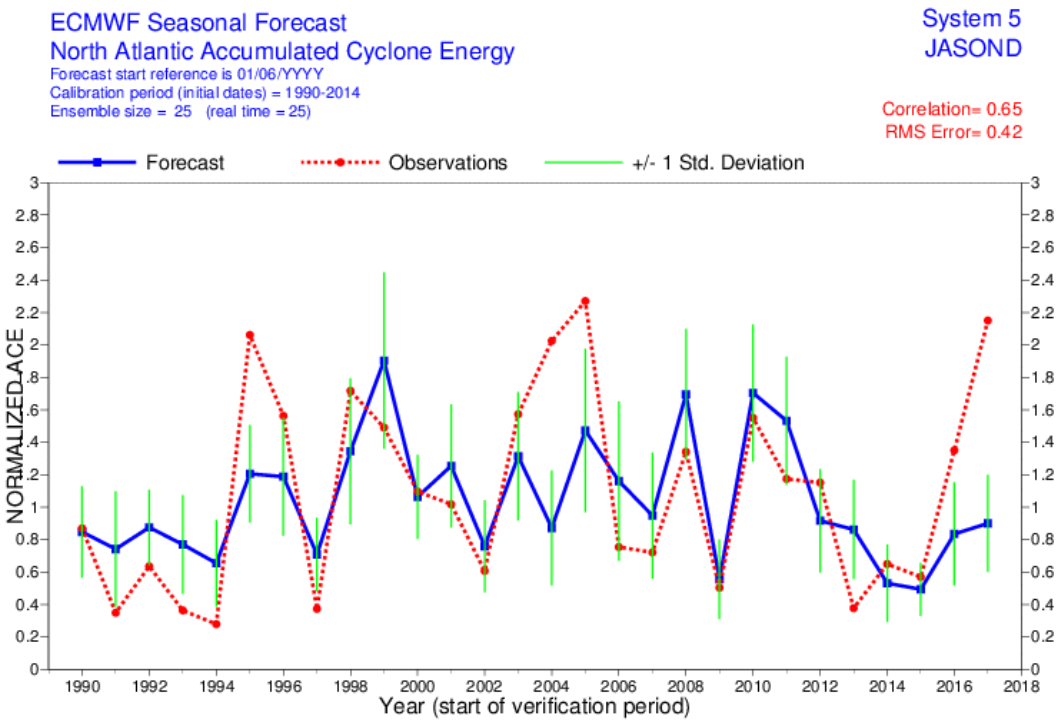


Figure 38: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1990 to July–December 2017. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts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

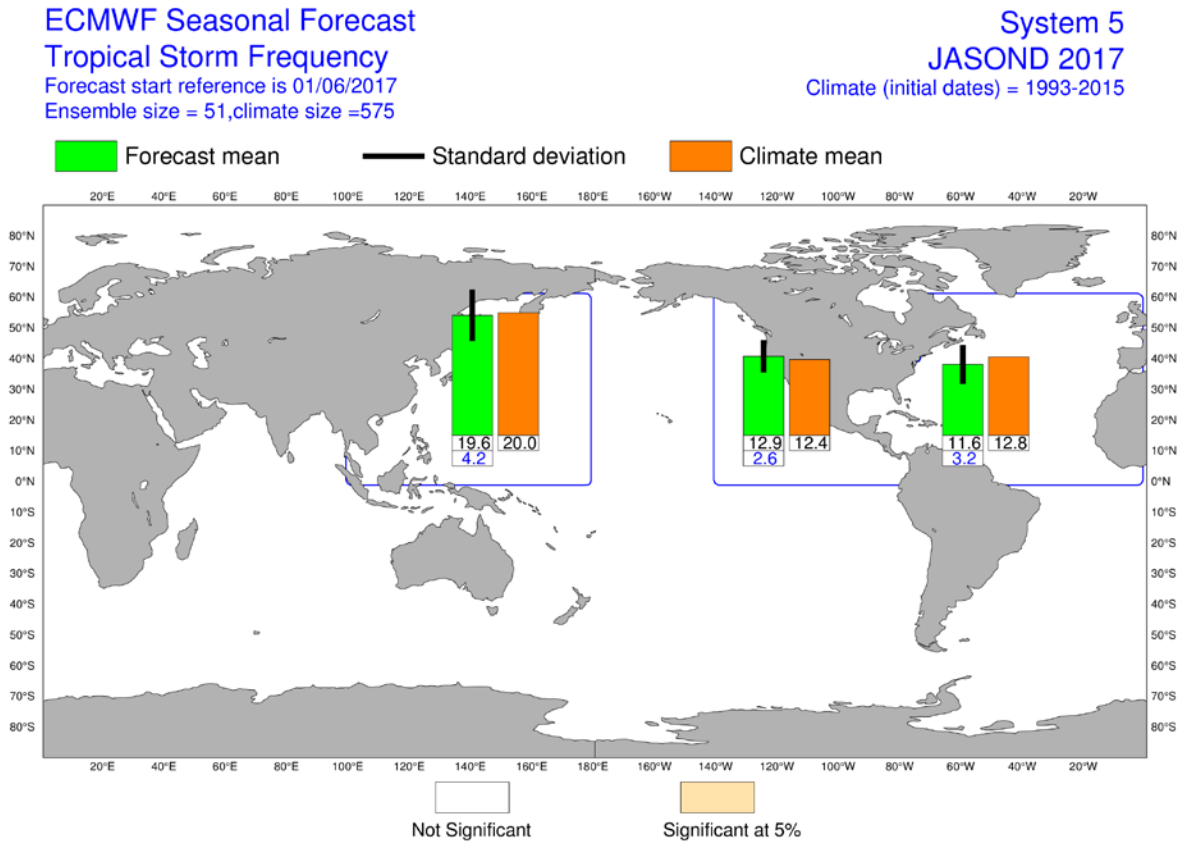


Figure 39: Tropical storm frequency forecast issued in June 2017 for the six-month period July–December 2017. Green bars represent the forecast number of tropical storms in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black underneath. The 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.

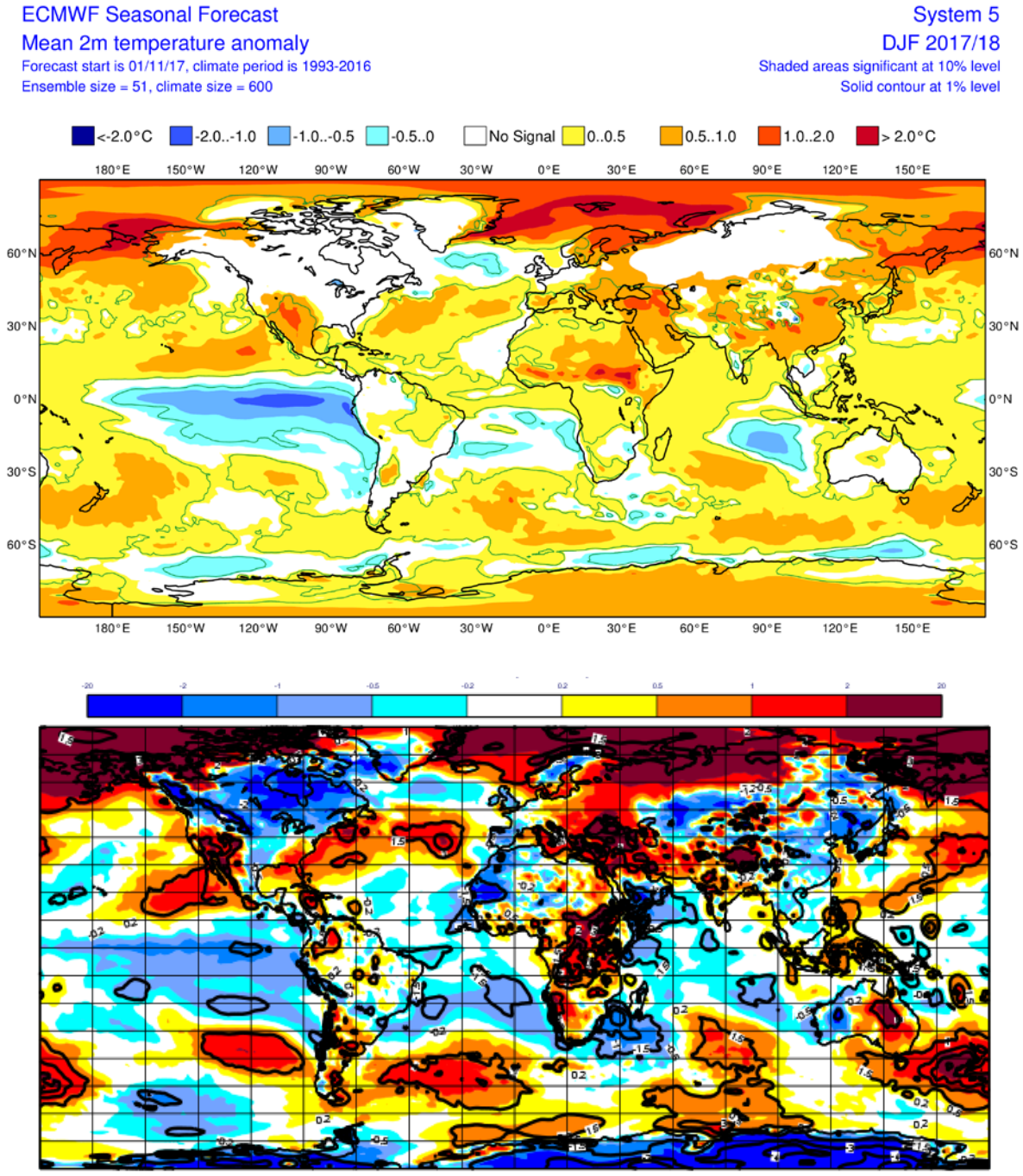


Figure 40: Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2017 for DJF 2017/18 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

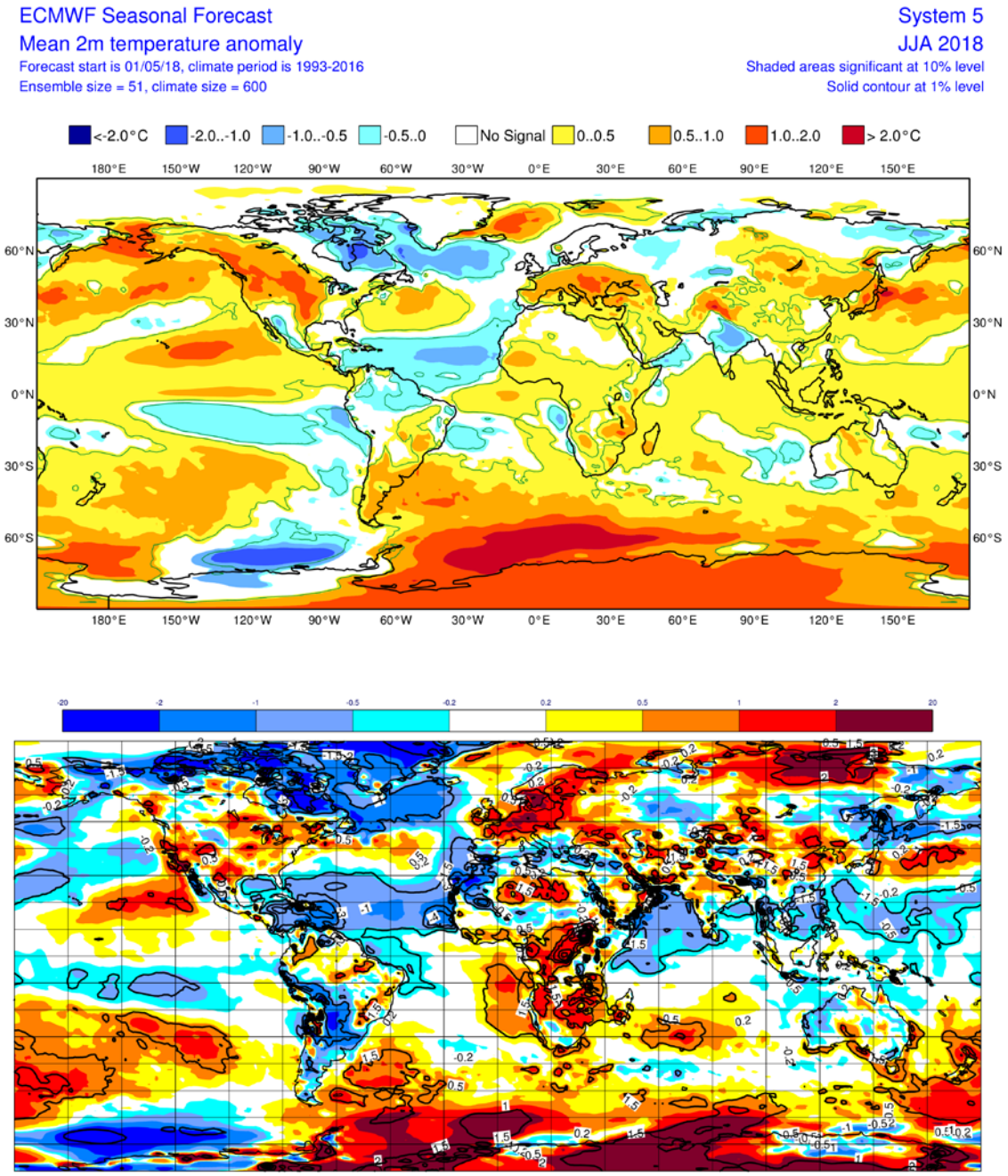


Figure 41: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2018 for JJA 2018 (upper panel), and verifying analysis (lower panel). Black contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

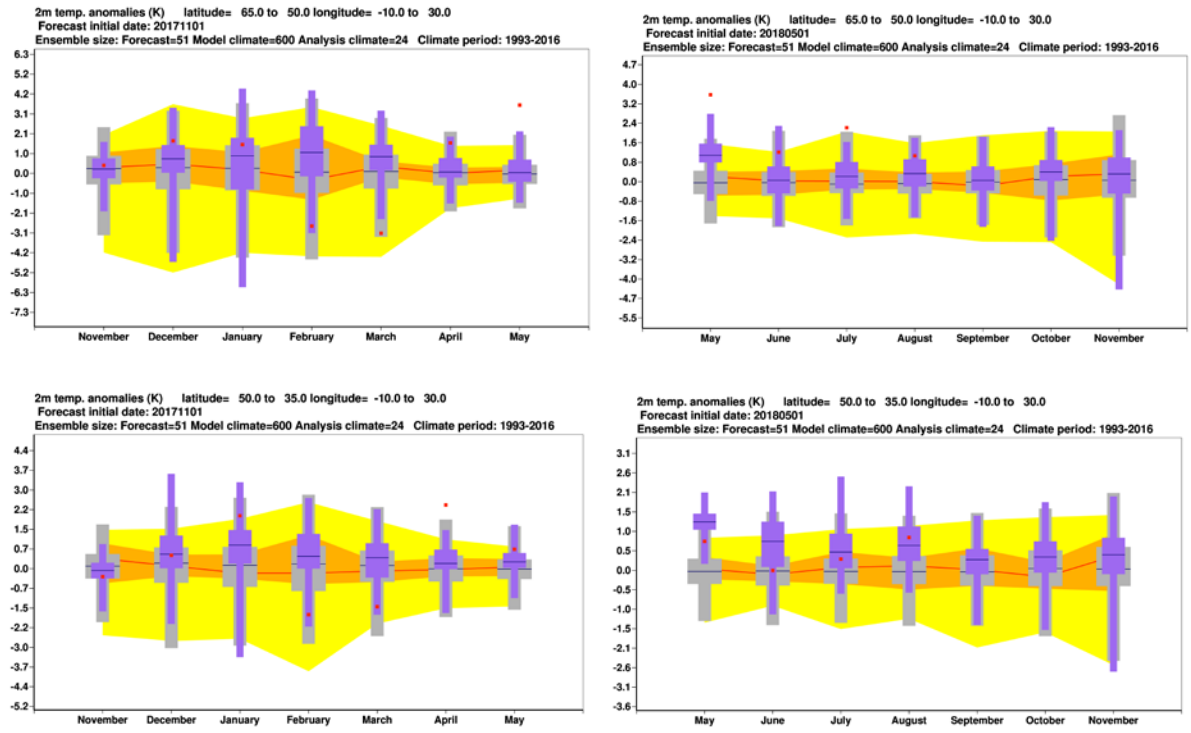


Figure 42: Long-range forecast of 2 m temperature anomalies from November 2017 for DJF 2017–18 (left panels) and from May 2018 for JJA 2018 (right panels) for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 hindcasts is shown 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.

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, 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 the effect of finite ensemble size N into account 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 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. For a fair comparison between two forecasting systems using SEDI, the forecasts need to be calibrated (Ferro and Stephenson, 2011). Thus, SEDI is a measure of the potential skill of a forecast system. 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, M. Janousek, S. Keeley, K. Mogensen and D. Richardson, 2017: New IFS cycle brings sea-ice coupling and higher ocean resolution. ECMWF Newsletter No. 150, 14–17.

Buizza, R., P. Bechtold, M. Bonavita, N. Bormann, A. Bozzo, T. Haiden, R. Hogan, E. Holm, G. Radnoti, D. Richardson and M. Sleigh, 2017: IFS Cycle 43r3 brings model and assimilation updates. ECMWF Newsletter No. 152, 18–22.

Buizza, R., E. Andersson, R. Forbes and M. Sleigh, 2017: The ECMWF Research to Operations (R2O) process. ECMWF Technical Memorandum No. 806, 16pp.

Buizza, R., G. Balsamo and T. Haiden, 2018: IFS upgrade brings more seamless coupled forecasts. ECMWF Newsletter No. 156, 18-22.

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.

Forbes, R., A. Geer, K. Lonitz and M. Ahlgrimm, 2016: Reducing systematic error in cold-air outbreaks. ECMWF Newsletter No. 146, 17–22.

Forbes, R., 2018: Improved precipitation forecasts in IFS Cycle 45r1. ECMWF Newsletter No. 156, 4.

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.

Ingleby, B., L. Isaksen, T. Kral, T. Haiden and M. Dahoui, 2018: Improved use of atmospheric in situ data. ECMWF Newsletter No. 155, 20–25.

Keeley, S. and K. Mogensen, 2018: Dynamic sea ice in the IFS. ECMWF Newsletter No. 156, 23–29.

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. Wea. Forecasting, 32, 1667-1674.

Lopez, P., 2018: Promising results for lightning predictions. ECMWF Newsletter No. 155, 14–19.

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.

Mogensen, K., L. Magnusson, J.-R. Bidlot and F. Prates, 2018: Ocean coupling in tropical cyclone forecasts. ECMWF Newsletter No. 154, 29–34.

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.

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., 99, 1015-1026.