Technical
Memo

ECMWF
European Centre for Medium-Range
Weather Forecasts

902

Evaluation of ECMWF forecasts, including the 2021 upgrade

T. Haiden, M. Janousek, F. Vitart,
Z. Ben Bouallegue, L. Ferranti, F. Prates
and D. Richardson

September 2022

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.

Abstract

This report provides a summary of ECMWF's forecast performance, covering medium, extended, and seasonal forecast ranges. It includes a short description of the changes implemented as part of the upgrade to model cycle 47r3 in October 2021 and the meteorological impact of the upgrade. There are several so-called headline scores that have been adopted by ECMWF in collaboration with its member states to monitor the evolution of various aspects of forecast skill. The report gives updates on these scores, as well as supplementary scores to help provide a more complete assessment of forecast skill. The primary focus of this summary is the medium range, and specifically the forecast performance for upper-air variables. It is shown that model cycle 47r3 led to an increase in upper-air forecast skill, such that ECMWF's lead compared to other global centres is being maintained. For surface parameters it is shown that the forecast skill of the Extreme Forecast Index (EFI) has reached a new high point. Over the ocean, where ECMWF has traditionally been leading in terms of significant wave height but not peak wave period, due to 47r3 it now also leads for the latter. On the seasonal timescale, the recent continuation and re-strengthening of La Nina was not predicted well, however the European anomalously warm summer 2022 was indicated in the forecasts with a consistent signal.

Plain Language Summary

This report summarizes ECMWF's forecast performance for the whole range of forecast lead times from a few days up to several months ahead. It also describes the changes that were made to the forecasting system in October 2021, and how they affected the skill of the forecasts. An important aspect of forecast performance is the skill of the model in predicting the larger-scale flow of the atmosphere. For this reason, a large part of the verification results deals with so-called 'upper-air' variables which define this flow. It is shown that the changes made to forecasting system in October 2021 did indeed increase the performance of the forecast in this respect. Model output more directly related to the weather experienced at the ground has also been verified, most importantly the Extreme Forecast Index (EFI) which is used to provide guidance in situations of potentially hazardous weather. ECMWF's skill in predicting the EFI has been the highest ever. Also, significant progress has been made in predicting the peak wave period over the oceans, such that ECMWF is now leading in this regard. At long ranges of several months ahead, the main source of predictability is the sea surface temperature in the tropical eastern Pacific. A persistent and recurring cold anomaly there ('La Nina') has not been forecast particularly well over the last year. Nevertheless, the extremely hot summer 2022 in Europe was indicated by the long-range forecasts initialized in late spring.

1 Introduction

The most recent change to the ECMWF forecasting system (IFS Cycle 47r3, on 12 October 2021) is summarised in section 2. Verification results of ECMWF medium-range upper-air forecasts are presented in section 3, including some comparisons of ECMWF's forecast performance with that of other global forecasting centres. Section 4 presents the evaluation 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 has been included and, to aid comparison from year to year, the set of plots shown 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, 853, 880, 884). One new plot has been added to highlight the shortwave radiation aspect of ECMWF's forecast performance. A short technical note describing some of 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:

https://apps.ecmwf.int/webapps/opencharts/

by choosing 'Verification' and

- 'Medium Range' (medium-range and ocean waves)
- 'Extended Range' (monthly)
- 'Long Range' (seasonal)

2 Changes to the ECMWF forecasting system

On 12 October 2021, ECMWF performed a major upgrade of the Integrated Forecasting System (IFS). IFS Cycle 47r3 includes changes to the forecast model and the data assimilation. It improves the assimilation and observations usage and provides a significantly improved physical basis for moist processes. Apart from immediate gains in forecast skill this facilitates further development of the IFS and its future application at convection-permitting resolutions.

2.1 Model and data assimilation changes

The new model cycle incorporates a major revision of the treatment of cloud by using a more consistent formulation of boundary layer turbulence, shallow convection, and sub-grid cloud. It employs a simplified and more consistent treatment of sub-grid cloud saturation adjustment, a consistent treatment of subgrid cloud from boundary layer turbulent mixing, a consistent computation of mixing height for the unstable turbulent boundary layer and convection scheme, and a change from double to single iteration of the turbulent mixing scheme. Other specific changes are:

- New method for computing inversion strength based on moist entropy for distinguishing stratocumulus and cumulus cloud
- Limit to convective overshoot based on tropopause stability

- New parametrized deep convection closure with an additional dependence on total advective moisture convergence
- Change from exponential-exponential cloud vertical overlap to random-exponential overlap in closer agreement with observations
- Inclusion of the vapour deposition process for growth of falling snow particles
- Change from linear to cubic interpolation for cloud liquid, ice, rain, and snow semi-Lagrangian departure point calculations, including 3D quasi-monotone limiter
- Interpolation of cloud and precipitation to radiation grid changed from in-cloud to grid-mean
- Inclusion of full supersaturation adjustment in the ensemble SPPT stochastic perturbations
- Mass-weighting and relaxation timescale introduced for ensemble SPPT stochastic perturbations
- Revised simplified moist physics and associated tangent-linear and adjoint
- Bug fix for vertical interpolation of 3D aerosol climatology
- Improved calculation of extinction coefficients for near-surface visibility in fog, rain, and snow
- Revised gustiness parametrization
- Improved calculation of the peak wave period for multi-peaked ocean wave spectra

Changes to the assimilation system are:

- New RTTOV coefficients for hyperspectral infrared (IR) sounders
- New height reassignment for low level AMVs
- Adding representativeness error to the total observation error for Aeolus
- Weak-constraint 4D-Var activation in the stratosphere for the EDA system
- Assimilation of all-sky AMSU-A

2.2 Meteorological impact of the new cycle

Figure 1 and Figure 2 show medium-range scorecards for Cycle 47r3 relative to own analysis and observations, for the HRES and ENS, respectively. There are many positive impacts of the Cycle, particularly on upper-air scores and tropical cyclone tracks. There are also some deteriorations, as discussed below. The description of meteorological impact for this upgrade has been taken from the ECMWF Newsletter article by Forbes et al. (2021) which provides additional background on the various model and assimilation changes.

Upper-air geopotential and wind in the first few days of the forecast are significantly improved by up to a few per cent for the northern hemisphere 500 hPa geopotential anomaly correlation, reducing with lead time. Upper-air winds are particularly improved in the tropics throughout the medium range, by up to 7%, reducing with lead time. Tropical upper-air temperatures are improved in HRES but degraded in the ENS from a small (~0.2 K) increase in bias due to a warming by the stochastic perturbations (partly mitigated by the SPPT changes mentioned above). Low level temperatures (including 850 hPa and 2 m temperature) are approximately neutral versus observations but degraded versus analyses in the subsidence regions over subtropical oceans, where the temperature at 850 hPa is very sensitive to small changes in boundary layer height.

For tropical cyclones, there is a 10% improvement in the track location errors in both HRES and in the ensemble mean of ENS (Figure 3) from forecast days 2 to 5, due to a combination of the additional observations assimilated in cloudy regions and model changes which improve the steering flow. With little change in spread, this results in an improvement in the statistical reliability of the tracks. The tropical cyclone central pressure is shallower on average by 2–3 hPa in Cycle 47r3. The difference can be greater than this in the rapidly deepening phase, with Cycle 47r2 closer to the ‘best-track’ central pressure data reported by the official global monitoring centres, but of opposite sign in the later stages, where Cycle 47r3 is closer to the ‘best-track’ data.

The impact on near-surface parameters is more mixed. There is a small improvement in 2 m temperature in the extratropics, but a small degradation in the tropics. Two-metre dew point and 10 m wind also show small deteriorations. A 3% increase in bias of total cloud cover as well as an increase in small scale variability and more binary (0/1) cloud cover leads to a degradation in the calculated scores. With such a major physics change, it is inevitable that there are some degradations, and these will be investigated further for later IFS cycles.

There are significant changes in the characteristics of precipitation, including enhanced fine-scale structures, reduced areal coverage and higher peak precipitation rates. The PDF (probability density function) of precipitation rate is improved, with reduced occurrence of light precipitation rates and increased occurrence of high precipitation rates in convective regimes, but similar precipitation accumulations overall. This is particularly evident over continental regions such as the USA and Africa, in better agreement with radar and satellite-based precipitation estimates. Along the intertropical convergence zone (ITCZ) there are reductions in the number of overactive quasi-stationary precipitation cells, which has been a longstanding problem in the IFS. The overall precipitation scores show some positive signals. For tropical precipitation, the HRES shows improvements of 1–2% in the deterministic SEEPS (stable equitable error in probability space) score, and the ENS shows improvements of about 0.6% in the fair CRPS (continuous ranked probability score). In the extratropics, the HRES impact is generally neutral while the ENS again shows improvements.

For the extended range, the impact of the physics changes on full-resolution ensemble re-forecasts includes a general increase in spread of a few per cent, particularly in the tropics. Although there are some increases in bias (for example 850 hPa temperature), consistent with the medium range, the impact on bias-corrected scores is approximately neutral (slightly positive in the tropics). The forecast skill of the Madden–Julian Oscillation (MJO) is slightly improved. The overall increase in spread of the MJO index leads to a slight over-dispersion, especially from the 850 hPa zonal wind component of the index. However, the MJO amplitude is increased for lead times greater than five days, reducing the amplitude bias (e.g. from –15% to –10% at day 10) and the eastward phase bias. There is no significant impact on the frequencies of Euro-Atlantic regimes.

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 the northern extratropics, the 12-month averaged score has reached its highest value so far. In Europe, where the score naturally exhibits larger interannual variations, as well as in the southern extratropics, no new high point has been reached but values have been consistently high over the last 3 years.

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. Similar to anomaly correlation, in the northern hemisphere the 12-month running mean RMS error of the six-day forecast has reached its lowest value. In the southern hemisphere values have been consistently low over the past 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 inconsistency between successive 12 UTC forecasts for the same verification time. Values have slightly increased in Europe after a minimum in 2019, however in the northern extratropics as a whole, the lowest values so far have been reached.

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. The RMSE is at its lowest values for both parameters, and there has been little change in the last couple of years. Comparison with other centres in terms of 100 hPa temperature scores (Figure 8, top panel) show that ECMWF is maintaing a substantial lead. Stratospheric scores for some parameters improved substantially due to changes made in cycles 46r1 and 47r1 by reducing long-standing biases in the stratosphere (Sleigh et al., 2020). The centre and bottom panels in Figure 8 show HRES stratospheric temperature scores for a range of stratospheric levels. They have improved markedly due to recent model upgrades.

The trend in ENS performance is illustrated in Figure 9, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern hemisphere. Both in Europe and the northern extratropics, the 12-month running mean of this score has been going down in 2021 as it returned from a high value driven by high predictability in the winter 2020-21. Since the interannual variability is primarily driven by the winter season, the summer minima give a more robust indication of the longer-term trend. Starting in 2019, the extratropical summer minima have been generally at a higher level than before.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 10. Both for 500 hPa geopotential height and 850 hPa temperature, forecasts show a good overall match between spread and error. For 500 hPa geopotential, there is some overdispersion especially in week two of the forecast which is larger than in previous years. This is partly related to the longer standing issue of having too much spread along mid-latitude storm tracks. In summer (not shown) an underdispersion of similar magnitude is seen.

A good match between spatially and temporally averaged spread and error is necessary but not sufficient for a well-calibrated ensemble. It should also be able to capture day-to-day changes in predictability, as well as their geographical variations. This can be assessed using spread-reliability diagrams. Forecast values of spread over a given region and time period are binned into equally populated spread categories, and for each bin the average error is determined. In a well-calibrated ensemble, the resulting line is close to the diagonal. Figure 11 and Figure 12 show spread-reliability plots for 500 hPa geopotential and 850 hPa temperature for different global models. Spread reliability generally improves with lead time. At day 1 (left panels), forecasts are only moderately skilful in ‘predicting’ the average error, resulting in curves that deviate significantly from the diagonal, while at day 6 (right panels) most models are capturing spatio-temporal variations in error rather well. Overall, ECMWF performs best, with its spread reliability closest to the diagonal. The stars in the plots mark the average values, corresponding to Figure 10, and ideally should lie on the diagonal, as close to the lower left corner as possible. In this regard ECMWF performs best among the global models, with the exception of 850 hPa temperature at day 1, where the Japan Meteorological Agency (JMA) forecast has the lowest error (although ECMWF has slightly better overall match between error and spread).

To create a benchmark for the ENS, the CRPS is also computed for a ‘dressed’ ERA5 forecast. This allows 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 13 shows the evolution of CRPS skill of the ENS relative to the ERA5 reference for some upper-air parameters. At forecast day 5 (upper panel) the positive effect of 47r3 in 2021 is clearly visible, leading to a forecast performance which for most parameters is at its highest level so far. At forecast day 10, however, the interannual variability is too large to identify a potential signal of improvement from 47r3 for the parameters shown.

The forecast performance in the tropics, as measured by RMS vector errors of the wind forecast with respect to the analysis, is shown in Figure 14. Both the 200 hPa and 850 hPa errors show a clear signal of improvement from model cycle 47r3 and have all reached their lowest values so far.

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 Observation, Infrastructure and Information Systems (Infrastructure Commission) auspices, following agreed standards of verification.

Figure 15 shows time series of such scores for 500 hPa geopotential height in the northern and southern hemisphere extratropics. In the northern hemisphere, errors have decreased for all models, while ECMWF continues to maintain its lead. In the southern hemisphere, the gap between ECMWF and other centres has decreased slightly.

WMO-exchanged scores also include verification against radiosondes. Figure 16 (Europe), and Figure 17 (northern hemisphere extratropics) show 500 hPa geopotential height and 850 hPa wind forecast errors averaged over the past 12 months. While ECMWF does not lead at all forecast ranges, it has the best performance in the medium range when verified against observations.

The WMO model intercomparison for the tropics is summarised in Figure 18 (verification against analyses) and Figure 19 (verification against observations), which show vector wind errors for 250 hPa and 850 hPa. When verified against the centres' own analyses, the JMA forecast generally has the lowest error in the short range (day-2) while in the medium-range, ECMWF is leading for wind at 250 hPa, and both are tied for wind at 850 hPa. In the tropics, verification against analyses (Figure 18) is sensitive to details of the analysis method, in particular its ability to extrapolate information away from observation locations. When verified against observations (Figure 19), the ECMWF forecast has the smallest overall error.

3.3 CAMS scores

The Copernicus Atmospheric Monitoring Service (CAMS) uses the same model cycle as HRES but has lower horizontal resolution (40 km grid spacing), does not use the EDA, has prognostic aerosols interacting with radiation, and only extends to day 5. Figure 20 shows that in terms of 500 hPa geopotential in the extratropics, the meteorological skill of CAMS forecasts has dropped with the new model cycle 47r3. The cause of the decrease is under investigation and will be addressed in the next model upgrade. In some continental areas in the tropics such as India, and parts of Africa and South America, CAMS slightly outperforms HRES for lower atmospheric (850 hPa) temperature, indicating the benefit of prognostic aerosol on the meteorological forecast. Routine verification of the CAMS atmospheric composition forecast is carried out by the CAMS Evaluation and Quality Assurance (EQA) with reports being published at https://atmosphere.copernicus.eu/eqa-reports-global-services .

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 21. The top left 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. For comparison the same score is shown for ERA5. The top right panel shows the score difference between HRES and ERA5. The bottom left 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%, the bottom right panel shows the lead time where the Diagonal Skill Score (DSS) drops below 20%. The ENS thresholds have been chosen in such a way that the scores measure the skill at a lead time of about 7 days. All plots are based on verification against SYNOP observations.

The deterministic precipitation forecast has reached its highest level of skill so far. However, a similar increase is seen in the ERA5 reference forecast (black line in Figure 21, top left panel) such that the difference between the operational and ERA5 scores and HRES (upper right panel in Figure 21) is more or less flat. Note that the scores for ERA5 have moved outside the previously seen range between 3.5 and 4 days. More detailed analyses have shown that this is neither due to a change over time in the number of stations used in the verification, nor the fact that a fixed climatology (1980-2009) is used to compute SEEPS. One possibility is that a change in precipitation climate (e.g. a decreasing fraction of precipitation falling as snow) is causing the apparent increase in skill, but this hypothesis still needs to be confirmed.

The probabilistic precipitation headline score CRPSS (lower left panel in Figure 21) has not shown much change in recent years but is now more consistently above forecast day 7. It should be noted that in addition to the difference HRES vs ENS also the scores used (SEEPS vs CRPSS) measure different aspects of the forecast. SEEPS, as a categorical score in probability space, does not penalize errors at high precipitation values as much as the CRPSS. The DSS (lower right panel) measures, like SEEPS, errors in probability space and puts more weight on the discrimination aspect of the forecast, while the CRPSS is more sensitive to the reliability/calibration of the forecast. The discrimination ability of the ENS has in fact reached its highest value so far (as seen in the DSS), but the reliability has decreased somewhat so that the CRPSS has not increased in the same manner.

ECMWF performs a routine comparison of precipitation forecast skill for ECMWF and other centres for both the high-resolution and the ensemble forecasts using the TIGGE data archived in the Meteorological Archival and Retrieval System (MARS). Results using these same headline scores for the last 12 months show the HRES leading with respect to the other centres (Figure 22). ECMWF’s probabilistic precipitation forecasts are comparable in skill at day 1 to some of the other centres but clearly leading in the medium range.

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 23 to Figure 26. Verification is performed against SYNOP observations. 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 23), the daytime negative bias in spring has become smaller in 2021. The daytime error standard deviation has not changed much overall, however in some of the recent winter months it has been smaller than seen so far. The nighttime 2m temperature errors have have been very similar to the previous year. For 2 m dewpoint (Figure 24), the error standard deviation has increased recently. Similarly, for total cloud cover (Figure 25) there has been an increase in error standard deviation, as well as change of sign in bias. These are negative side effects of the comprehensive moist physics upgrade in model cycle 47r3 and will be addressed in upcoming cycles (Forbes et al., 2021). The error standard deviation of 10 m wind speed is comparable to the previous year (Figure 26).

ERA5 is useful as a reference forecast for the HRES, as it allows filtering out some of the effects of atmospheric variations on scores. Figure 27 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 that is a function of model resolution. For upper-air variables (verification against analysis), the positive effect from 47r3 is clearly visible. Note that the drop in the second half of 2020 not tied to a particular model cycle and appears to be related to atmospheric variability, notably an unusually high Arctic Oscillation index in JFM 2020. With this flow pattern, the HRES outperformed ERA5 somewhat more than usual during this particular winter. A similar feature (although less pronounced) has been observed for some of the other centres. For the surface parameters (verification against SYNOP), a weak positive effect is visible for 10 m wind speed, as well as a negative effect on total cloud cover.

As the verification of total cloud cover against SYNOP observations is affected by a significant representativeness mismatch and a generally large observation uncertainty, we also look at the skill of predicting radiation fluxes. Figure 28 shows how the 5-day forecast of the TOA net shortwave radiation has improved over time. ERA5 is included for comparison, showing that the recent slight uptick in RMSE is due to natural variability.

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’ whenever the CRPS exceeds 5 K. Figure 29 shows that in the annual mean (red curve) this fraction has decreased from about 7% to 4.5% over the last 15 years, and that there are large seasonal variations, with values in winter more than twice as high as in summer.

An analogous measure of the skill in predicting large 10 m wind speed errors in the ENS is shown in Figure 30. Here, a threshold of 4 m/s for the CRPS is used, to obtain similar fractions as for temperature. While the 12-month average value of the score has not changed much in 2021, the summer minimum has been the lowest so far.

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 31. While errors in 10 m wind speed have not become smaller in the last 2-3 years, wave height forecasts show a continued improving trend in the early medium range (forecast days 3 and 5 in the lower panel). This is seen even more clearly in the verification against analysis for the northern hemisphere (Figure 32, upper panel).

ECMWF is the WMO Lead Centre for Wave Forecast Verification, and in this role, it collects forecasts from wave forecasting centres to verify them against buoy observations. In the extratropics (Figure 33), ECMWF generally leads other centres in significant wave height, while for peak period ECMWF was until recently within the bundle of models. Changes to the compuattion of peak period in model cycle 47r3 however brought a substantial improvement, such that ECMWF is now leading for this parameter as well. This can also be seen in the scores for the tropics (Figure 34).

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

https://apps.ecmwf.int/webapps/opencharts/

by choosing ‘Verification’ and ‘Ocean waves’ (under ‘Parameters’).

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 the left column of Figure 35 (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. For 10 m wind speed and 2 m temperature, the 12-month average skill at day 5 (red curves) has reached its highest values ever in 2021-22. For precipitation, there has been a decrease in 2021 in the medium range, which is likely due to interannual variability, as it is not shown in an alternative score (see next paragraph).

A complementary way of verifying extremes is to use the Diagonal Elementary Skill Score DESS (Bouallegue et al., 2018), as shown in the right column of Figure 35 for the same three variables. It is based on verification in probability space, and like the ROC area, it emphasizes the discrimination aspect of the forecast. As for the EFI, the 95th quantile is used, but for wind and temperature, instantaneous rather than daily averages are used. Another difference between the two methods is that in the computation of the DESS, observation uncertainty (representativeness) has been explicitly accounted for using the method described in Bouallegue et al. (2020).

In terms of the DESS metric, forecast skill at day 5 has reached its highest values in 2021-22 for both precipitation and 2 m temperature.

5.2 Tropical cyclones

The tropical cyclone position error at day 3 of the HRES 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) are shown in Figure 36. 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 some aspects of tropical cyclone forecasts.

The HRES position error at day 3 (top panels, Figure 36) has slightly decreased compared to the previous year but not yet reached the low level of 2019. Relative to ERA5, the HRES has improved both at day

3 and at day 5 and reached the largest lead on ERA5 so far. The ENS mean absolute error of intensity has reached its lowest value so far, while there has been little change in ENS mean speed errors. The mean absolute error of speed in the HRES has reached its lowest value so far.

The bottom panel of Figure 36 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. The forecast continues to be slightly underdispersive, more so at day 3 than at day 5.

The ensemble tropical cyclone forecast is presented on the ECMWF website as a strike probability: the probability at any location that a reported tropical cyclone will pass within 120 km during the next 240 hours. Verification of these probabilistic forecasts for the three latest 12-month periods is shown in Figure 37. Results show a decrease in reliability compared to the two 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. The reliability of strike probability has slightly increased, while its ROC skill has slightly decreased. In terms of modified ROC skill there has been a small improvement.

6 Monthly and seasonal forecasts

6.1 Monthly forecast verification statistics and performance

Figure 38 shows the probabilistic performance of the monthly forecast over the extratropical northern hemisphere for summer (JJA, top panels) and winter (DJF, bottom panels) seasons 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. 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. Note that persistence is definer here as the persistence of the week 1 forecast into week 2, and persistence of the week 2 forecast into weeks 3+4.

Forecast skill for week 2 exceeds that of persistence on average by 5-10%, for weeks 3 to 4 (combined) by 0-5%. In weeks 3 to 4 (14-day period), summer warm anomalies appear to have slightly higher predictability than winter cold anomalies. There is a statistically significant long-term increase in skill for week 2 both in terms of absolute skill and skill above persistence. For weeks 3 to 4, no statistically significant trend is seen.

Because of the low signal-to-noise ratio of real-time forecast verification in the extended range (Figure 38), re-forecasts are a useful additional resource for documenting trends in skill. Figure 39 shows the skill of the ENS in predicting 2 m temperature anomalies in week 3 in the northern extratropics. Verification against both SYNOP observations and ERA5 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. However, a marked increase is seen in 2020-21, which is mainly due to ERA5 replacing ERA-Interim as initial condition for the reforecasts. Due to this change, the reforecast skill has ‘caught up’ and become more representative of real-time forecast skill. Note also that the verification is based on a sliding 20-year 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 Niño 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:

https://apps.ecmwf.int/webapps/opencharts/

by choosing 'Verification' and 'Extended' under 'Range'.

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 includes an interactive ocean (NEMO) model and 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. 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

https://apps.ecmwf.int/webapps/opencharts/

by choosing 'Verification' and 'Long' (under '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 2021-22 El Niño forecasts

The year 2021 was characterized by La Nina conditions at the beginning and end of the year, with close to neutral conditions during the summer months. This is a continuation of a multi-year La Nina state, which has rarely happened in recent decades. During 2021, SEAS5 forecasts showed a persistent signal for a return to neutral, or even slightly positive, conditions (Figure 40, left column), The C3S multi-model ensemble (Figure 40, right column), due to its naturally larger spread, better covered the observed evolution but also had a certain bias towards earlier return to neutral/positive.

6.2.3 Tropical storm predictions from the seasonal forecasts

The 2021 Atlantic hurricane season had a total of 21 named storms, which is the third highest number recorded, including 7 hurricanes and 4 major hurricanes. The accumulated cyclone energy index (ACE) was about 145% of the past 10-year (2010-2019) climate average (Figure 41) which makes it a very active, but not record-breaking season. Seasonal tropical storm predictions from SEAS5 indicated correctly a higher level of activity over the Atlantic (ACE of about 120% (+/- 40%) of the past 10-year average). The number of tropical storms (21) was slightly underpredicted (17) by SEAS5. Subsequent forecasts, issued in July and August, also predicted an above average intensity of the tropical cyclone season.

Figure 42 shows that SEAS5 predicted average activity over the eastern North Pacific, and below average activity over the western North Pacific (ACE of about 70% of the 2011-2020 climate average). The 2021 western Pacific typhoon season was a below-average season producing 22 storms, 9 typhoons, with an ACE about 30% below average, which is consistent with the SEAS5 forecast. The eastern North Pacific hurricane season was a near-normal active season with an ACE close to climatology, which is consistent with the SEAS5 prediction. Overall, SEAS5 tropical cyclone activity forecasts issued on 1st May 2021 verified well over all three ocean basins.

6.2.4 Extratropical seasonal forecasts

The seasonal forecast for boreal winter 2021-22 was reasonably skillful in the northern hemisphere, especially over the North Pacific, where the spatial pattern of warm and cold anomalies was well captured (Figure 43). As usual, over continents the skill was not quite as high, for example the eastward extent of the North American cold anomaly was underestimated, and the magnitude of the positive anomaly over much of mid-latitude Asia was underestimated. Here we compare an ensemble mean forecast with the actual outcome, so we cannot expect the full magnitude of the observed anomalies to be captured. Like SEAS5, most other centres predicted a warm anomaly in Siberia where a cold anomaly was observed.

Summer 2m temperature anomalies (Figure 44) still showed a distinctive La Nina pattern over the Pacific, which was captured reasonably well. A strong signal for a warm anomaly in Europe was present in the forecast, although its extension to, and overall magnitude in Scandinavia and Siberia was underestimated. The strongly negative temperature anomaly in the region of Pakistan was predicted. It was linked to a strongly positive precipitation anomaly in the area due to a very strong monsoon, leading to widespread severe flooding in the area. The fact that the negative phase of Indian Ocean Dipole (IOD) had a substantial amplitude may have increased the predictability for this anomaly.

Since the ensemble mean carries only part of the information provided by an ensemble, we also look at the forecast distribution in the form of quantile (climagram) plots. Climagrams for Northern and Southern Europe for winter 2021-22 and summer 2022 are shown in Figure 45. Red squares indicate observed monthly anomalies. As in previous years, both in winter and summer, warm anomalies are generally better predicted than cold ones, partly due to the global warming signal present in the forecast. The exceptionally warm period from May-July in Southern Europe (lower right panel) was indicated by an unusually strong signal in the forecast.

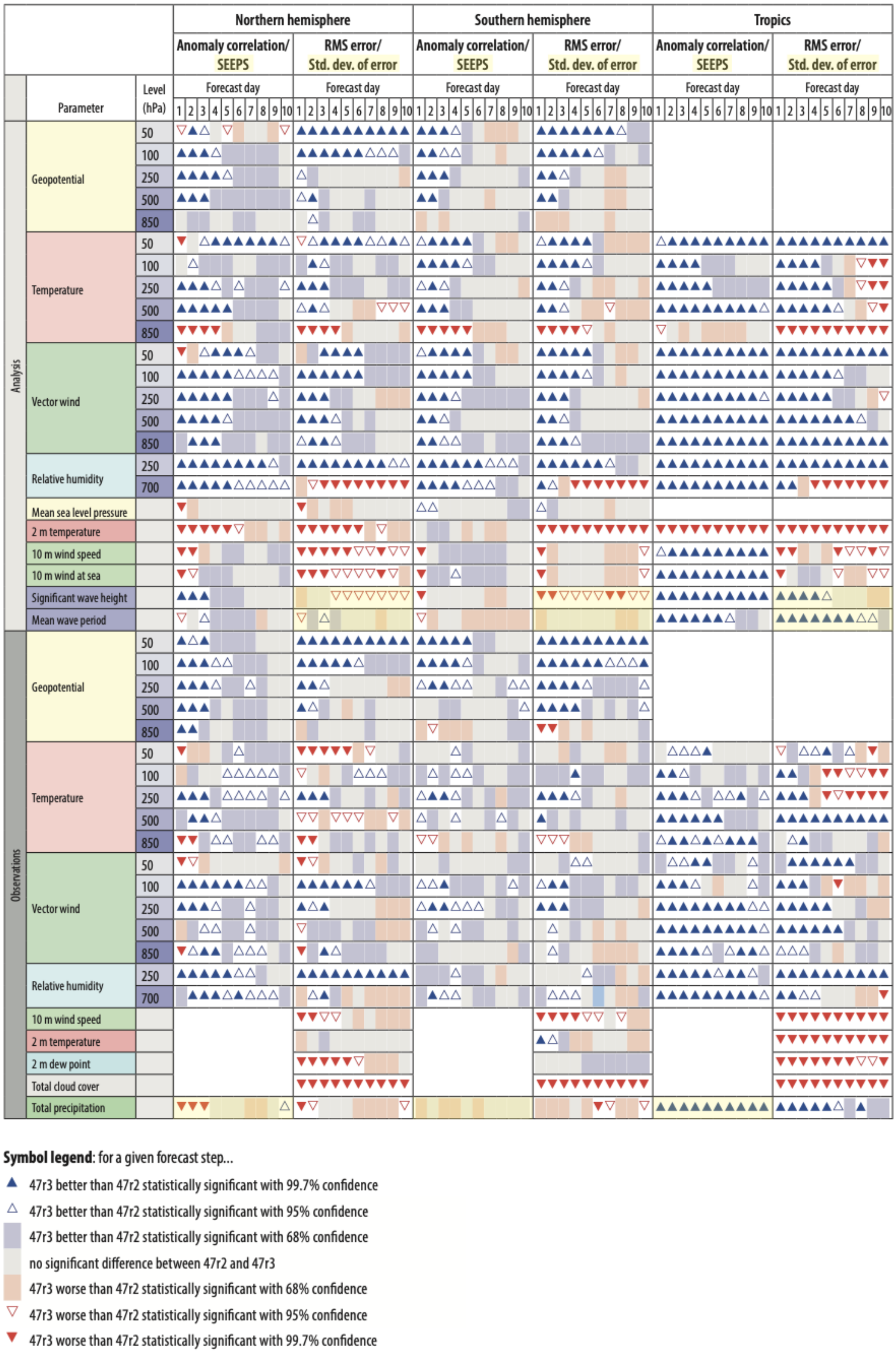


Figure 1: Summary HRES score card for IFS Cycle 47r3. Score card for HRES cycle 47r3 versus cycle 47r2 verified by the respective analyses and observations at 00 and 12 UTC for about 650 forecast runs in the period June-August 2020 and December 2020-August 2021. Yellow colouring indicates that symbols refer to the second score indicated at the top of the column.

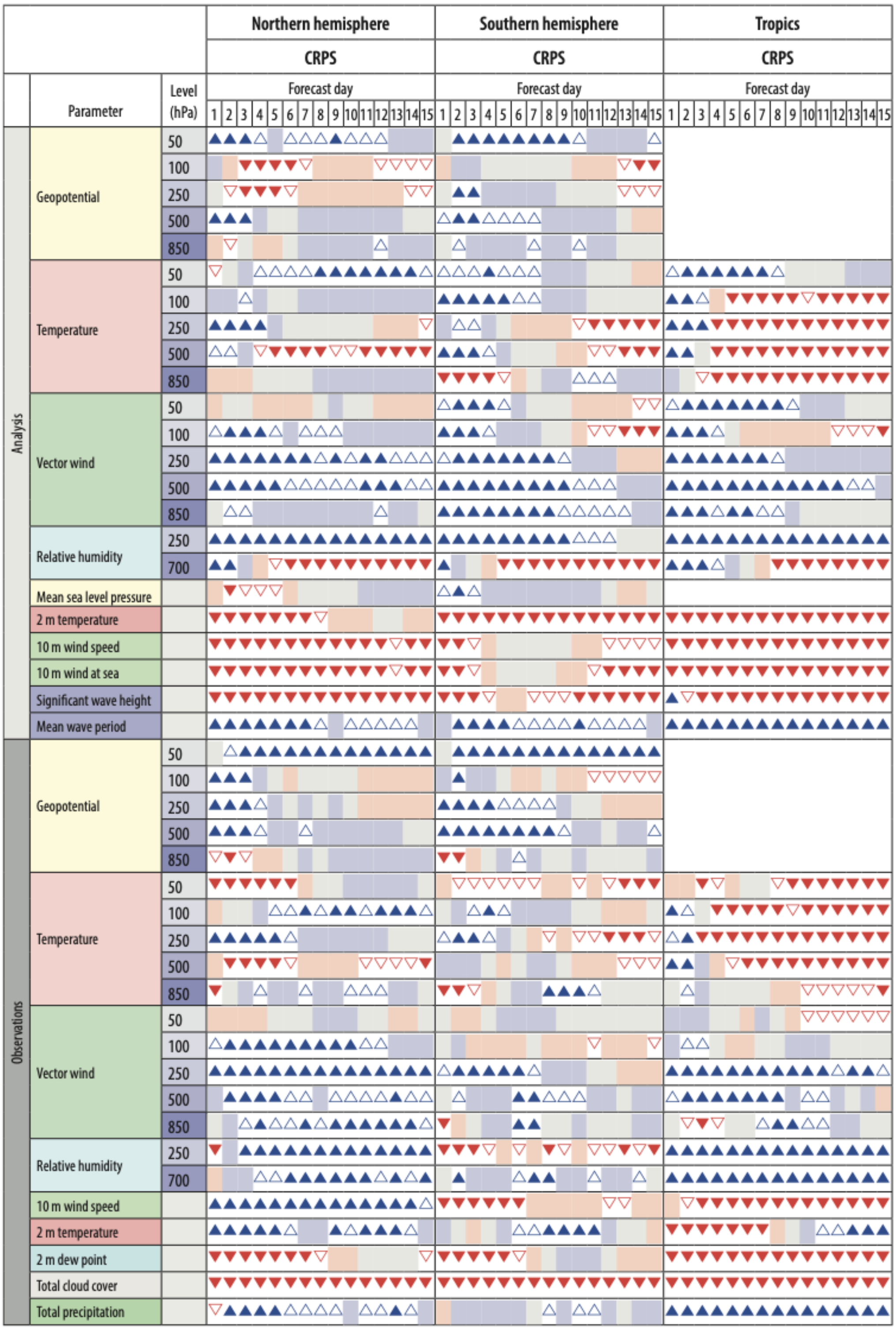


Figure 2: Summary ENS score card for IFS Cycle 47r3. Score card for ENS cycle 47r3 versus cycle 47r2 verified by the respective analyses and observations at 00 UTC for about 305 ENS forecast runs in the period June - August 2020 and December 2020 - August 2021.

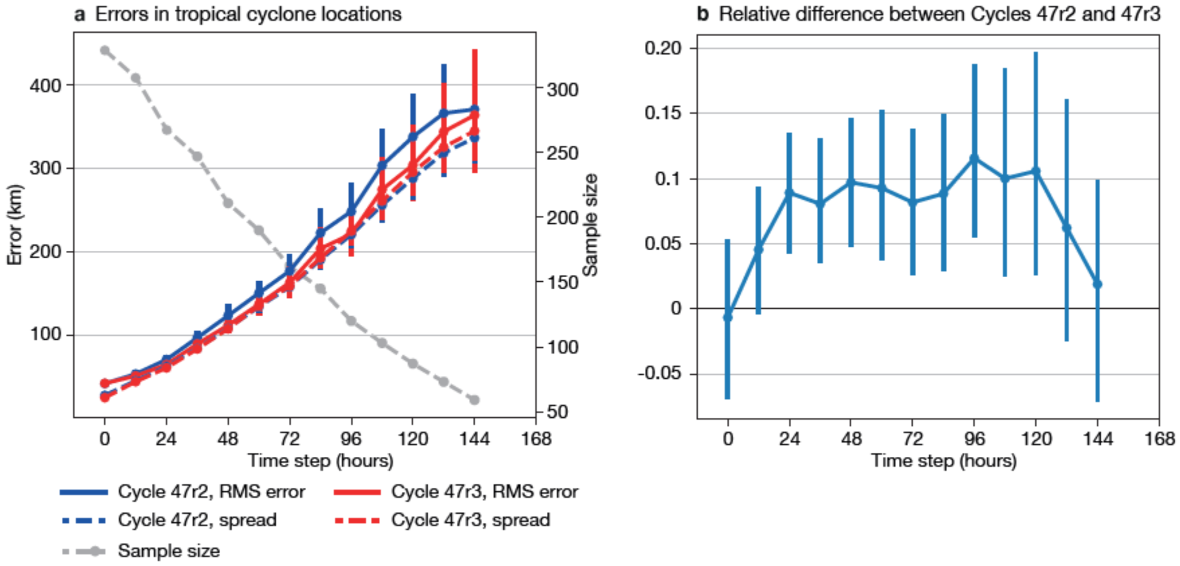


Figure 3: The charts show (a) root-mean-square (RMS) location errors (solid lines) in the ensemble mean of tropical cyclone positions in Cycle 47r2 (blue) and 47r3 (red), along with the standard deviation (‘spread’, dashed lines) among ensemble members, and (b) the normalised difference in ensemble mean location error between Cycles 47r2 and 47r3 (positive values indicate improved position in 47r3). Results are based on all TC basins for the period from 2 December 2020 to 30 August 2021. The dashed grey line in the left-hand panel and the right-hand side scale indicate the number of tropical cyclones which could be evaluated at each lead time. The bars indicate 95% confidence intervals. From Forbes et al. (2021).

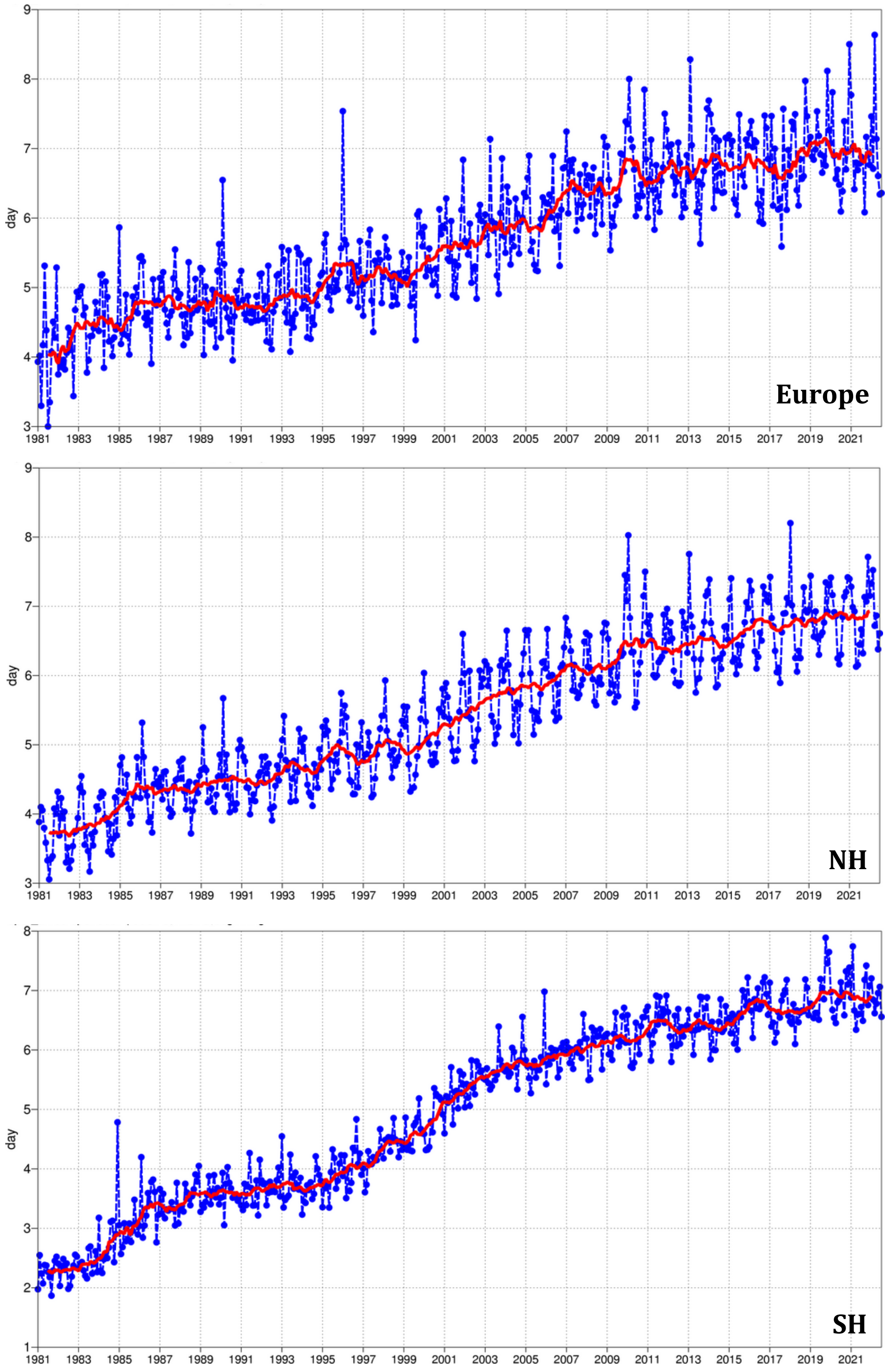
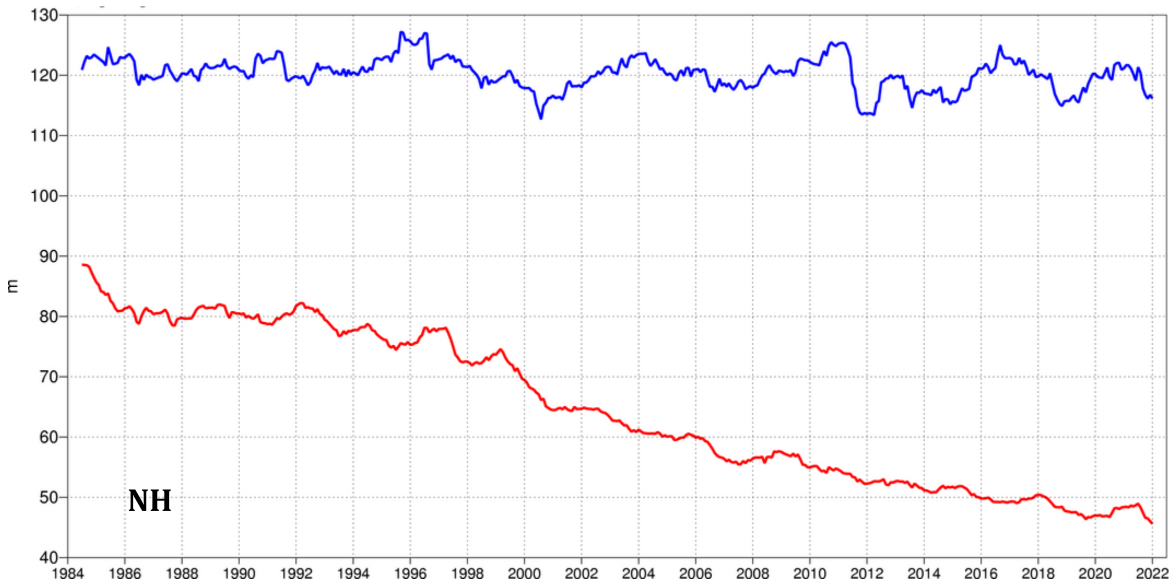


Figure 4: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly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



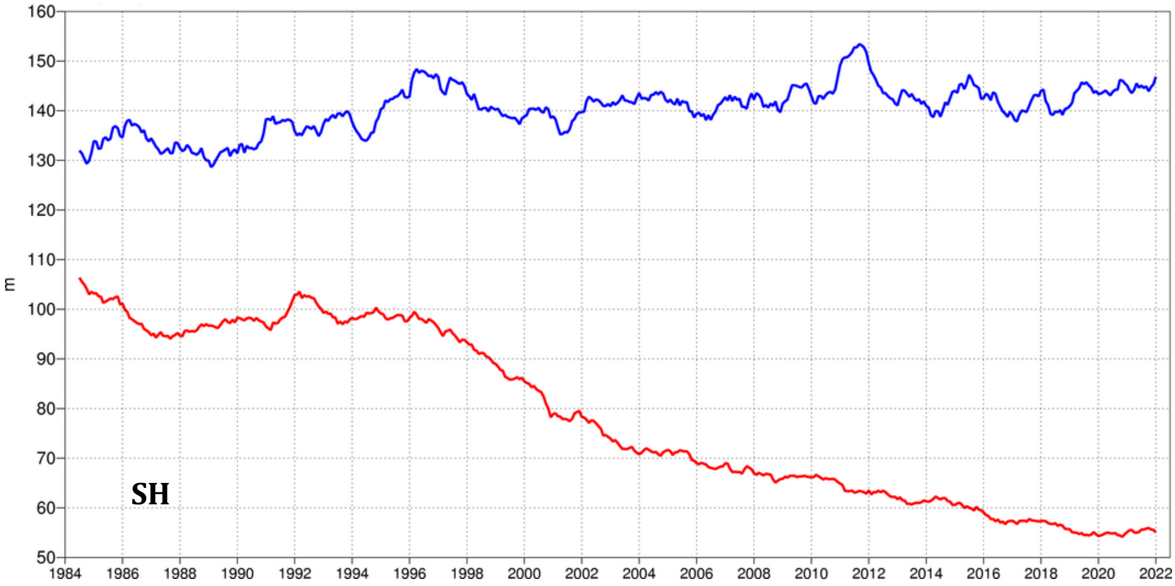
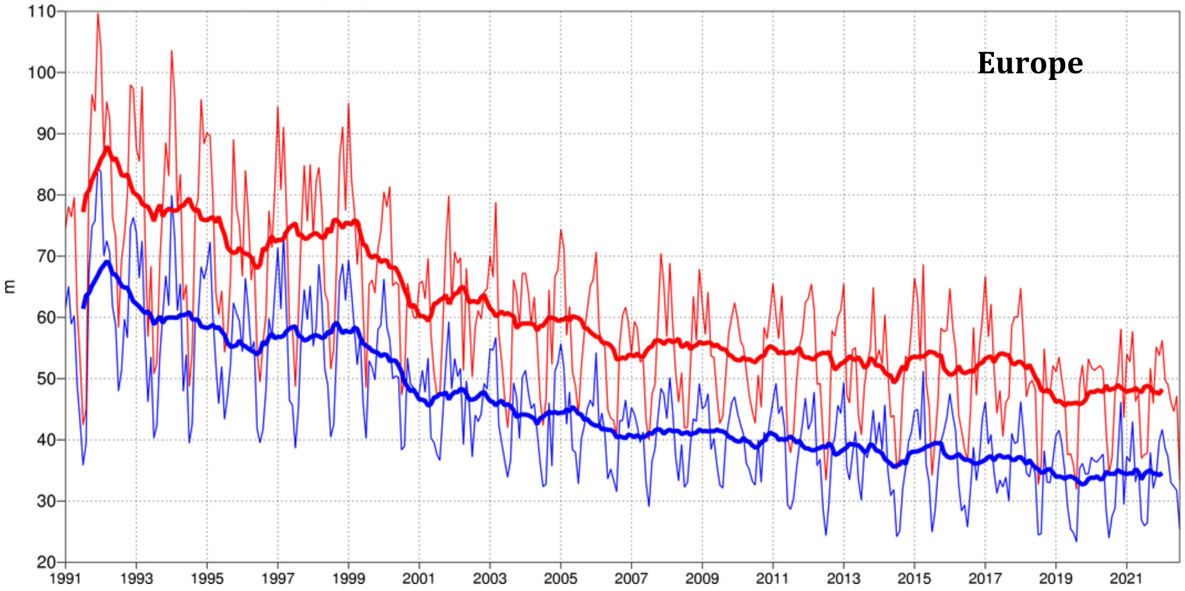


Figure 5: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2021–July 2022. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).



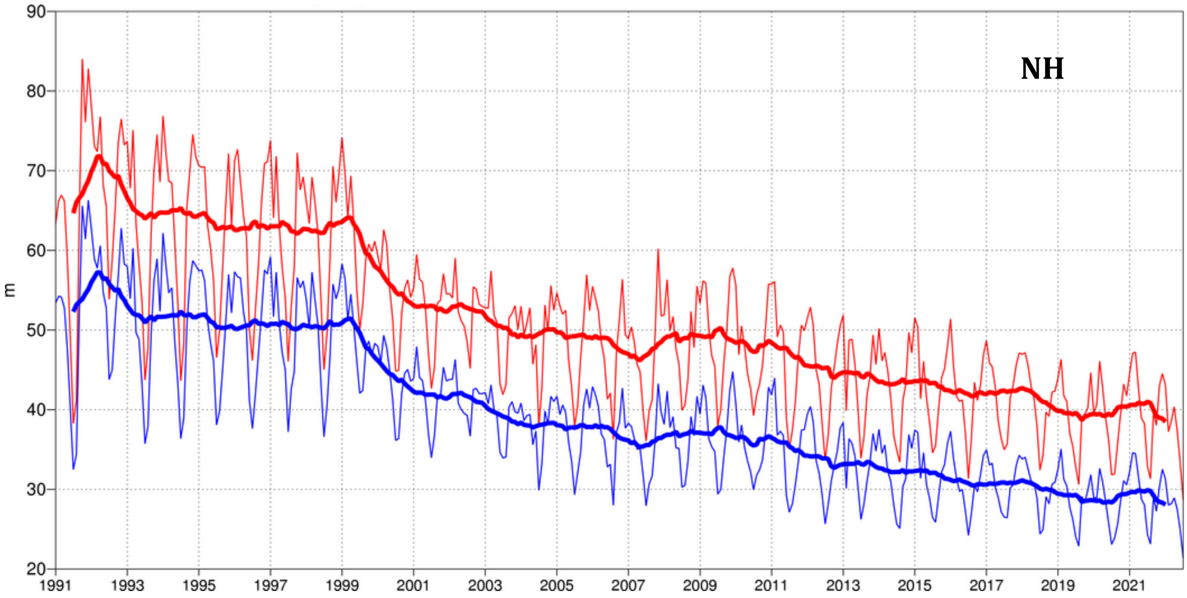


Figure 6**:** A measure of inconsistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month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

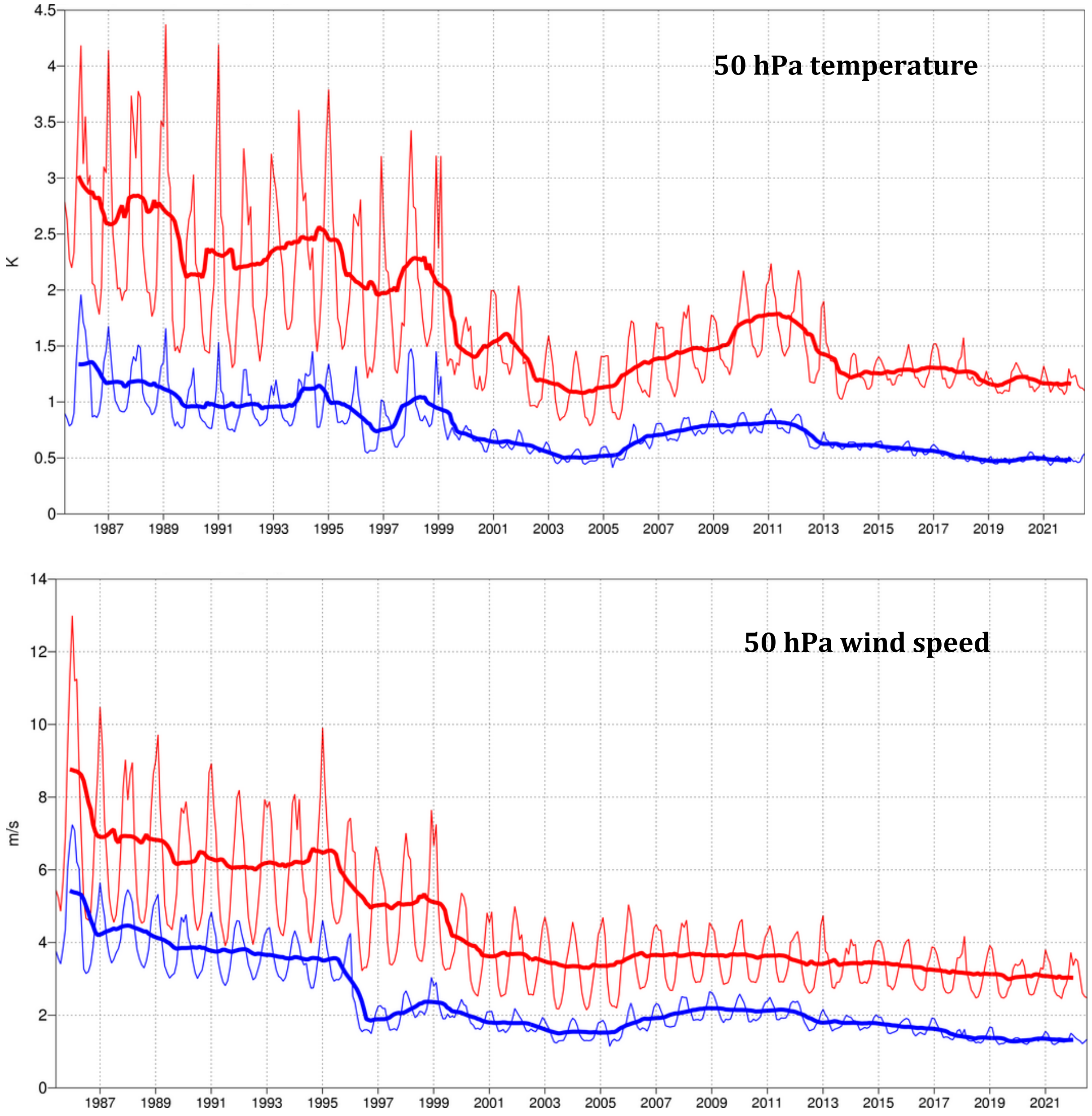
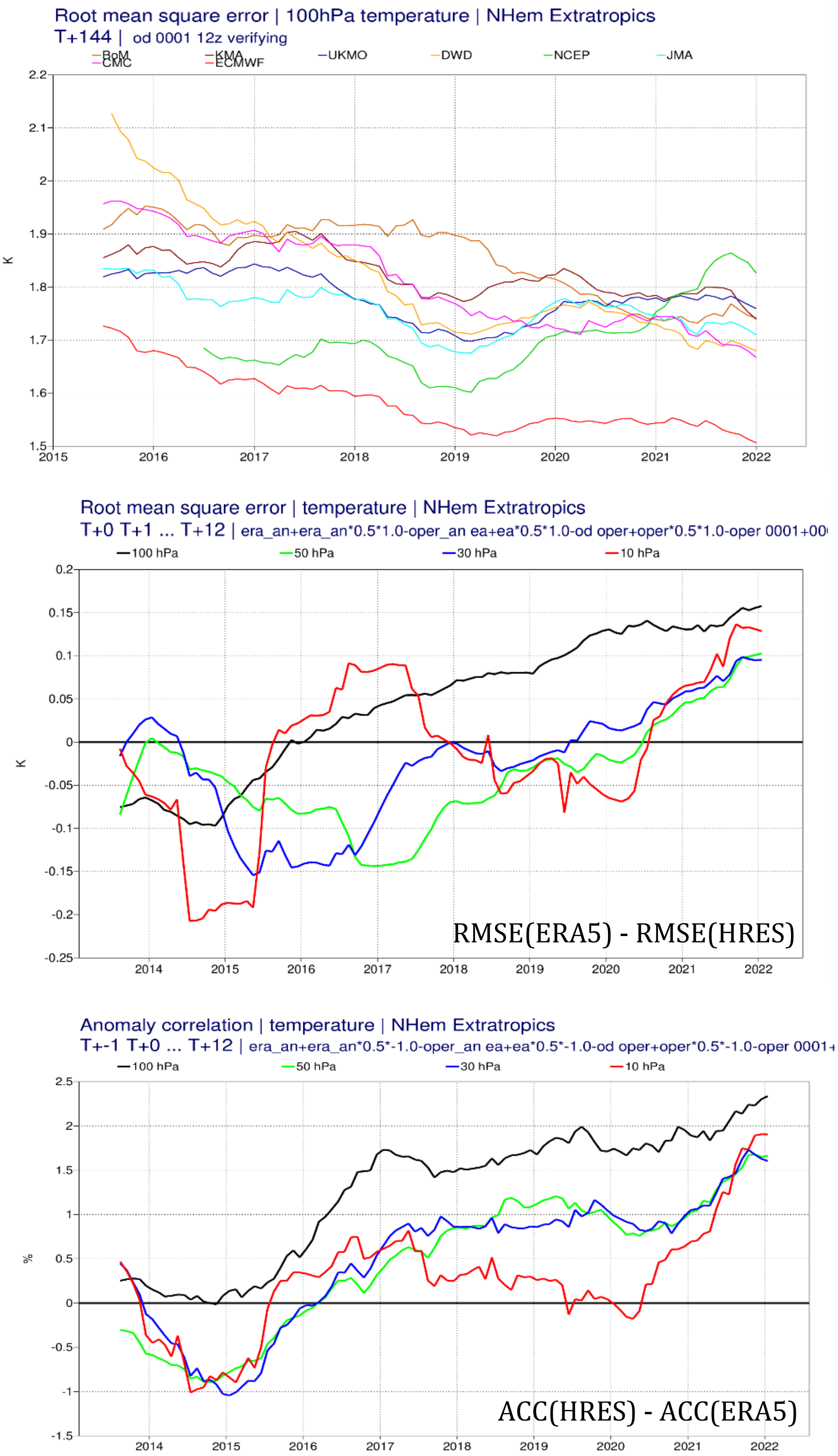


Figure 7**:** Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are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



Root mean square error | 100hPa temperature | NHem Extratropics
T+144 | od 0001 12z verifying

Root mean square error | temperature | NHem Extratropics
T+0 T+1 ... T+12 | era_an+era_an*0.5*1.0-oper_an ea+ea*0.5*1.0-od oper+oper*0.5*1.0-oper 0001+00

Anomaly correlation | temperature | NHem Extratropics
T+-1 T+0 ... T+12 | era_an+era_an*0.5*-1.0-oper_an ea+ea*0.5*-1.0-od oper+oper*0.5*-1.0-oper 0001+

Figure 8: Stratospheric scores at a lead time of +144 h. Top: global model intercomparison of the 100 hPa temperature RMSE in the northern extratropics based on the WMO exchange of scores. Centre: difference in RMSE of temperature between ERA5 and HRES at four different stratospheric levels. Bottom: difference in anomaly correlation of temperature between HRES and ERA5 at four different stratospheric levels. Curves in all three plots are 12-month running averages.

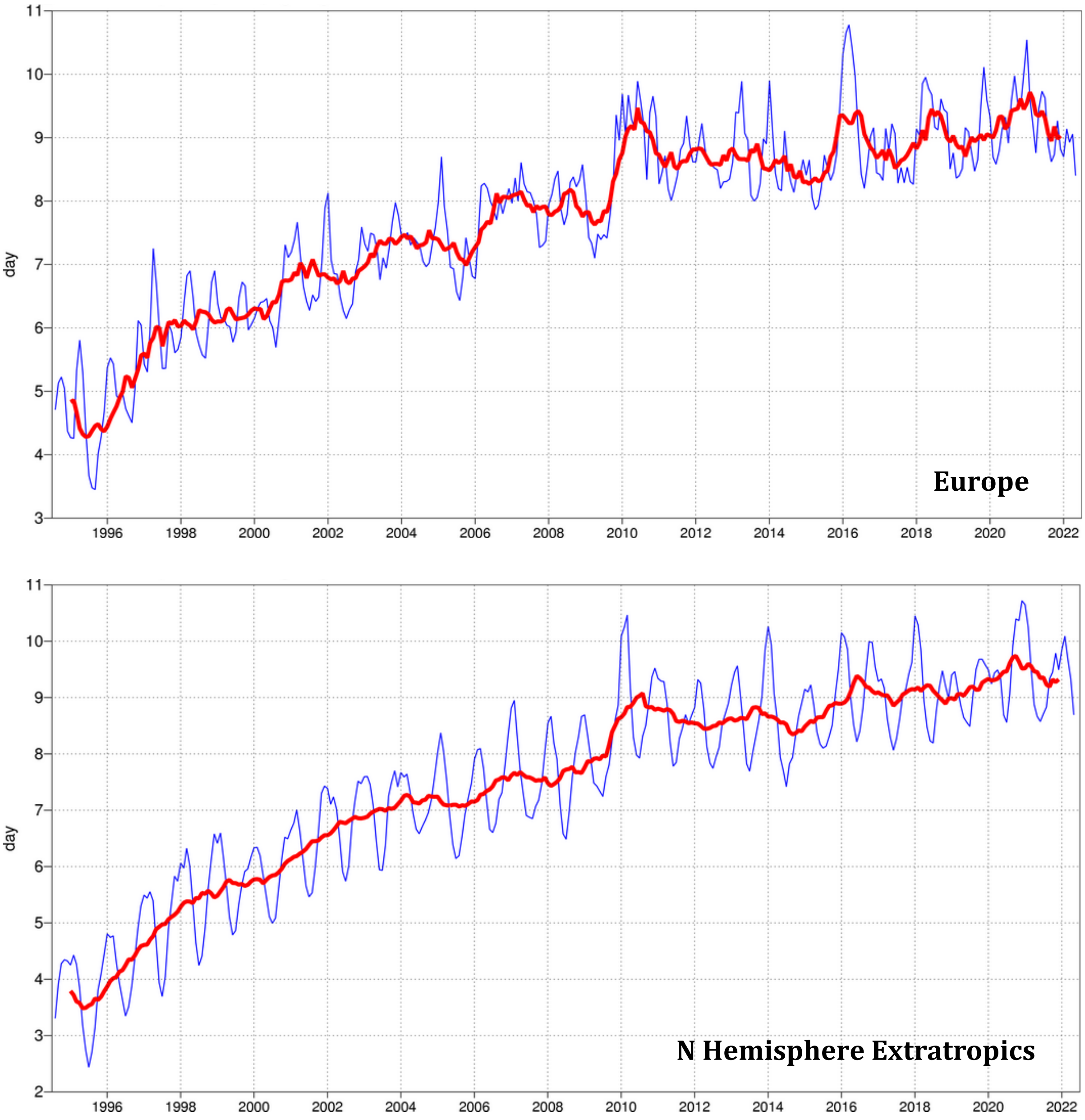


Figure 9**:** Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

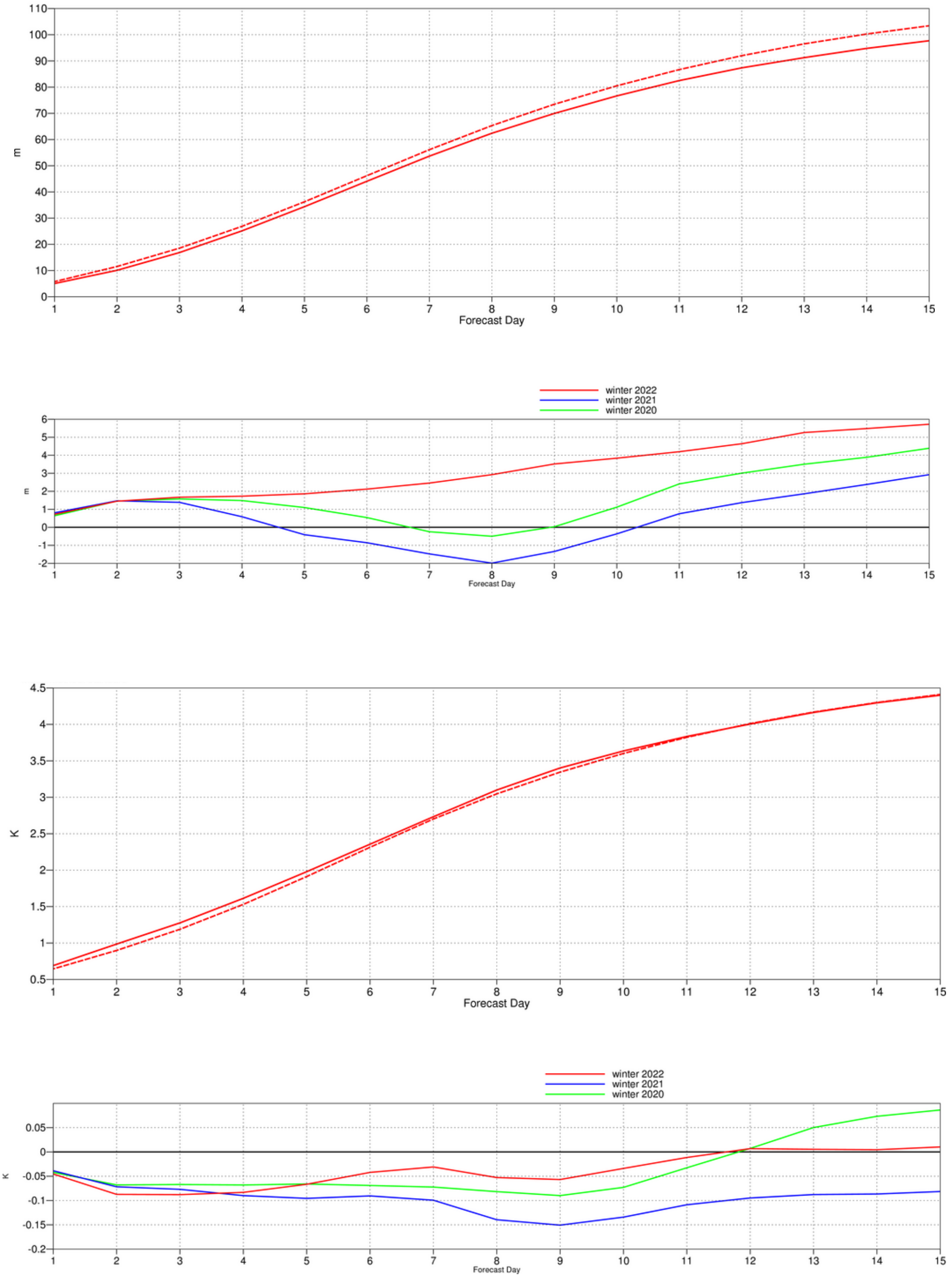


Figure 10**:** Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2021–2022 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate spread is too small); verification is against analysis, plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extratropical northern hemisphere for forecast days 1 to 15.

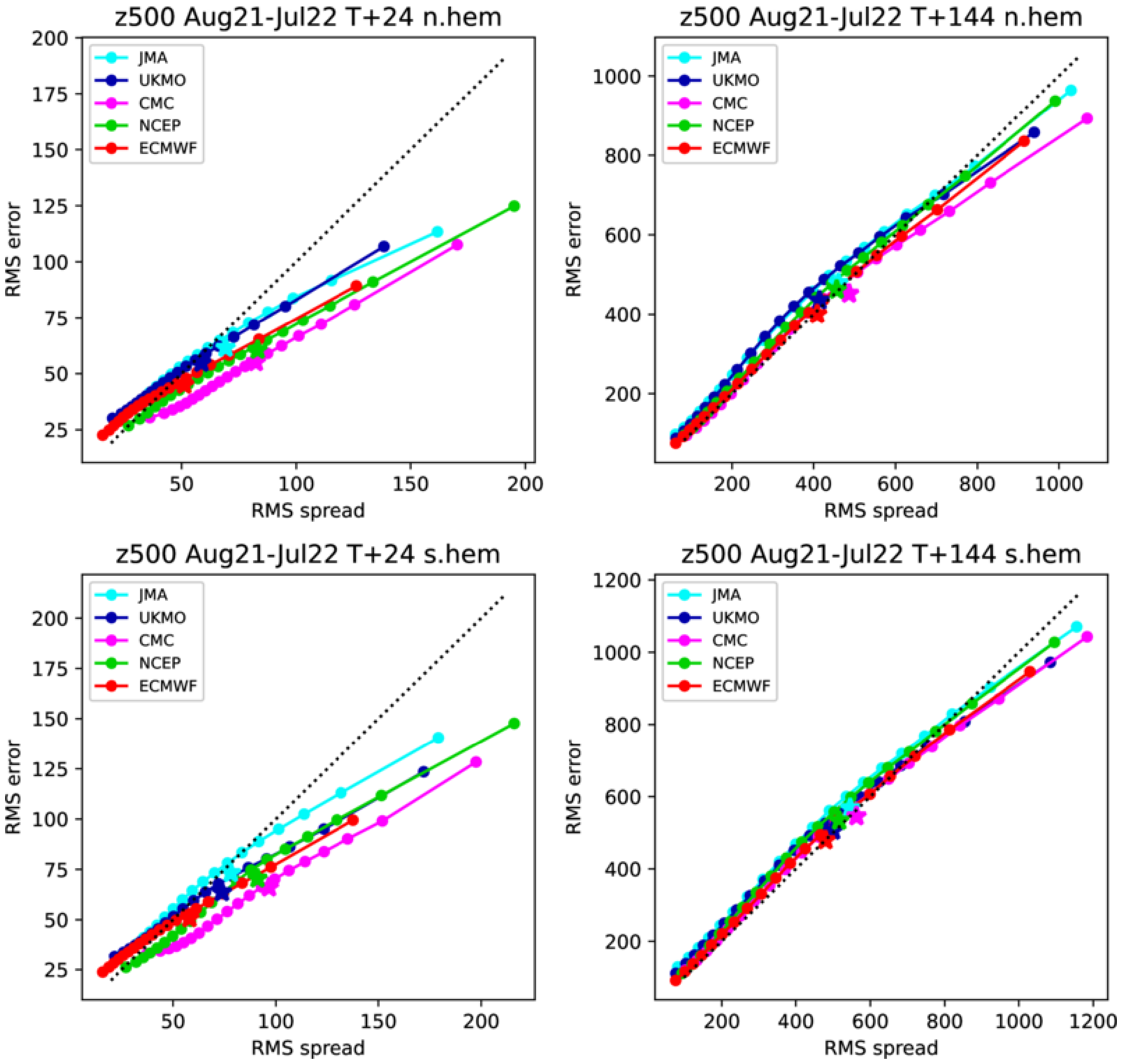


Figure 11**:** Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2021–July 2022 in the northern (top) and southern (bottom) hemisphere extra-tropics for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

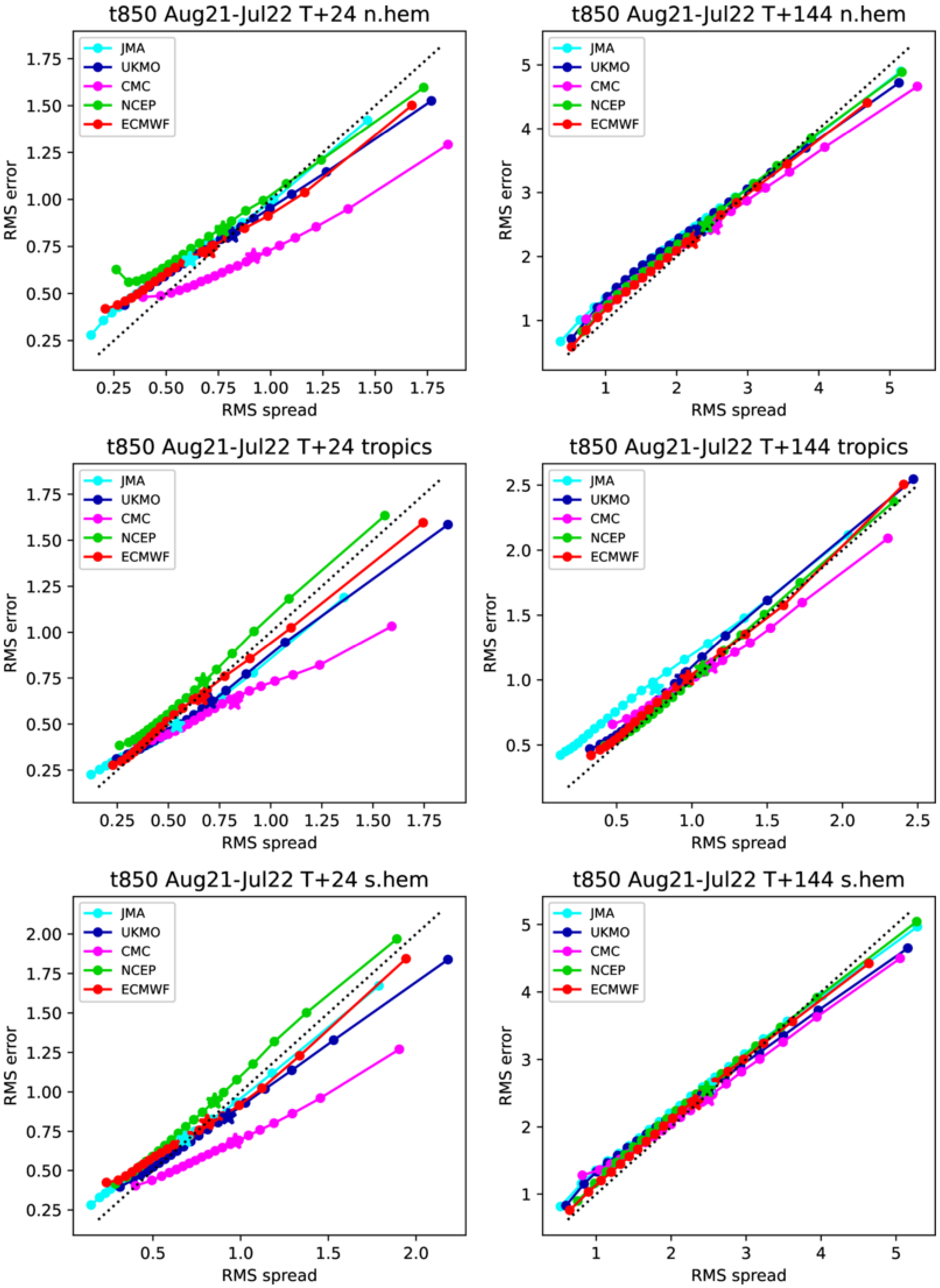


Figure 12**:** As Figure 11 for 850 hPa temperature, and including the tropics.

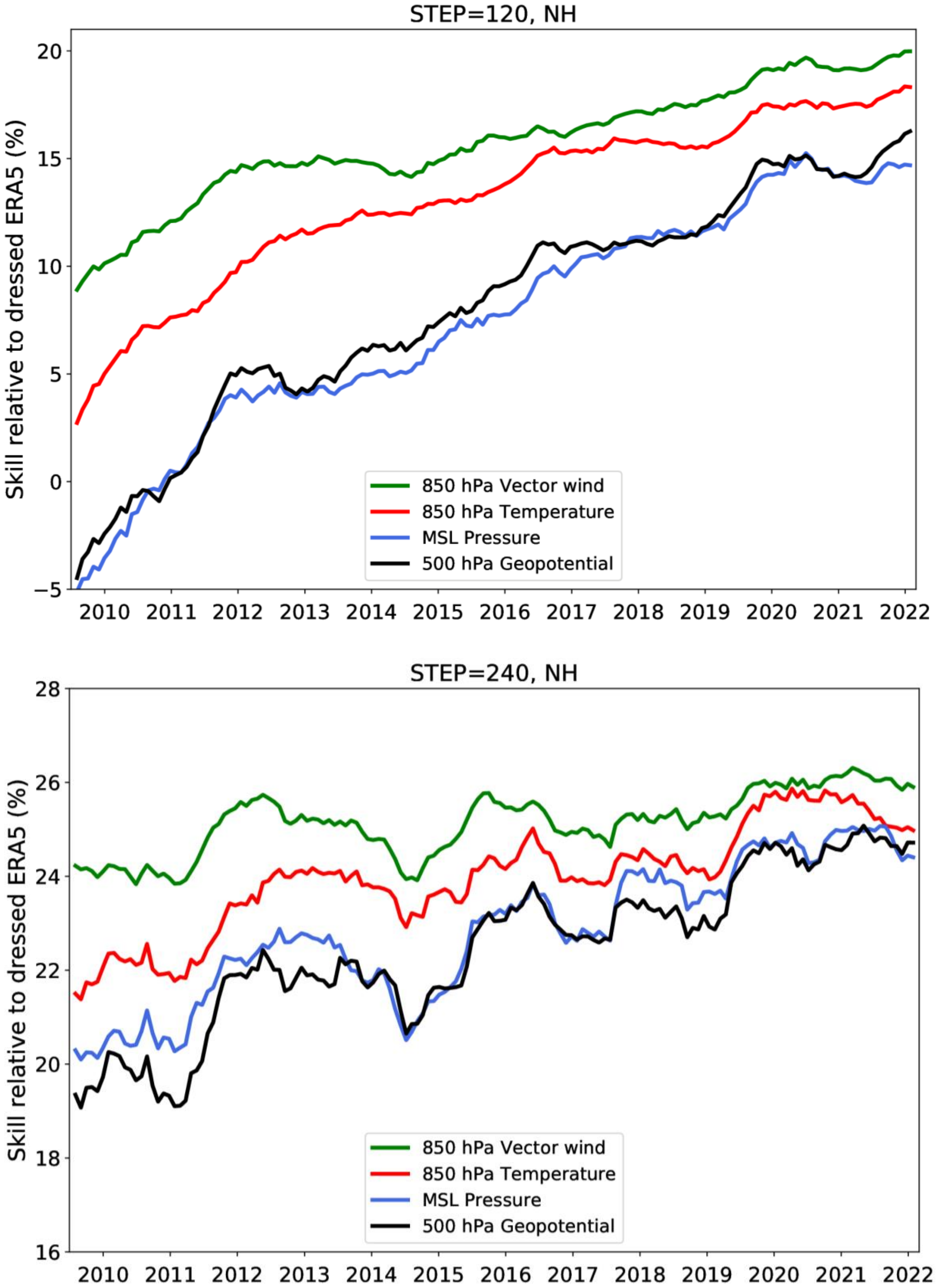
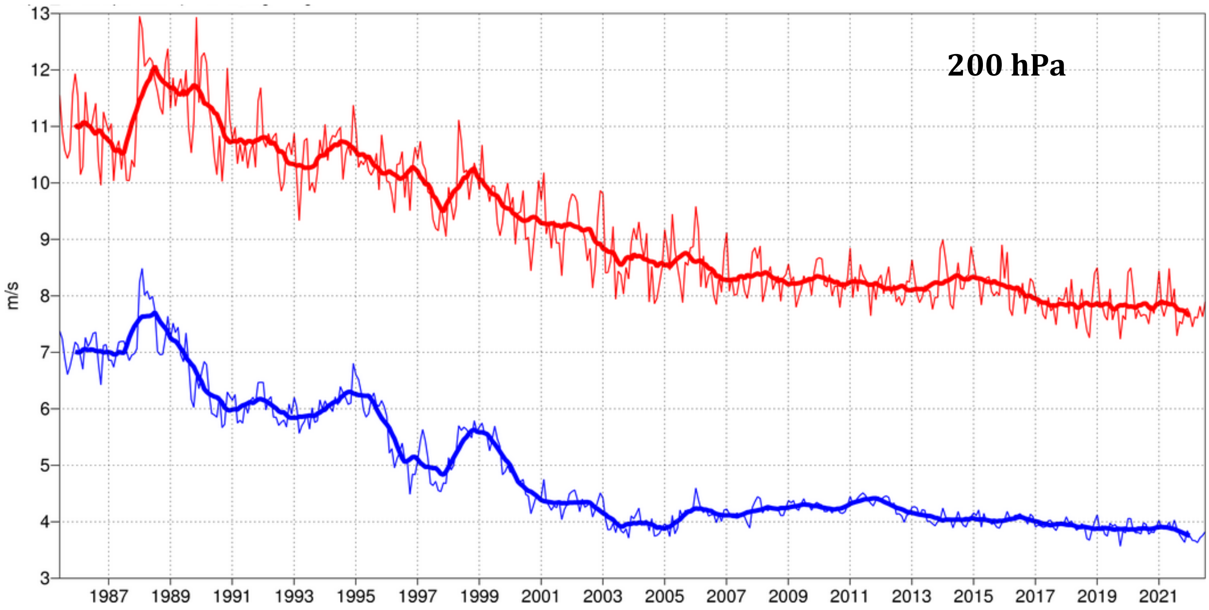


Figure 13**:** Skill of the ENS at day 5 (top) and day 10 (bottom) for upper-air parameters in the northern extra-tropics, relative to a Gaussian-dressed ERA5 forecast. Values are running 12-month averages, and verification is performed against own analysis.



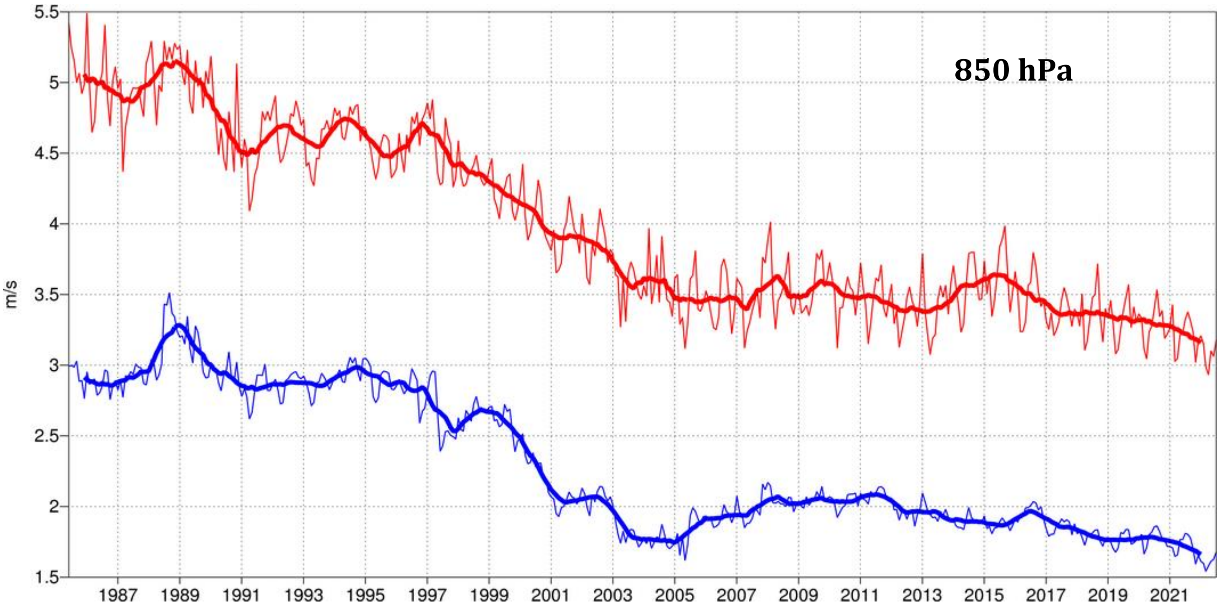


Figure 14**:** Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

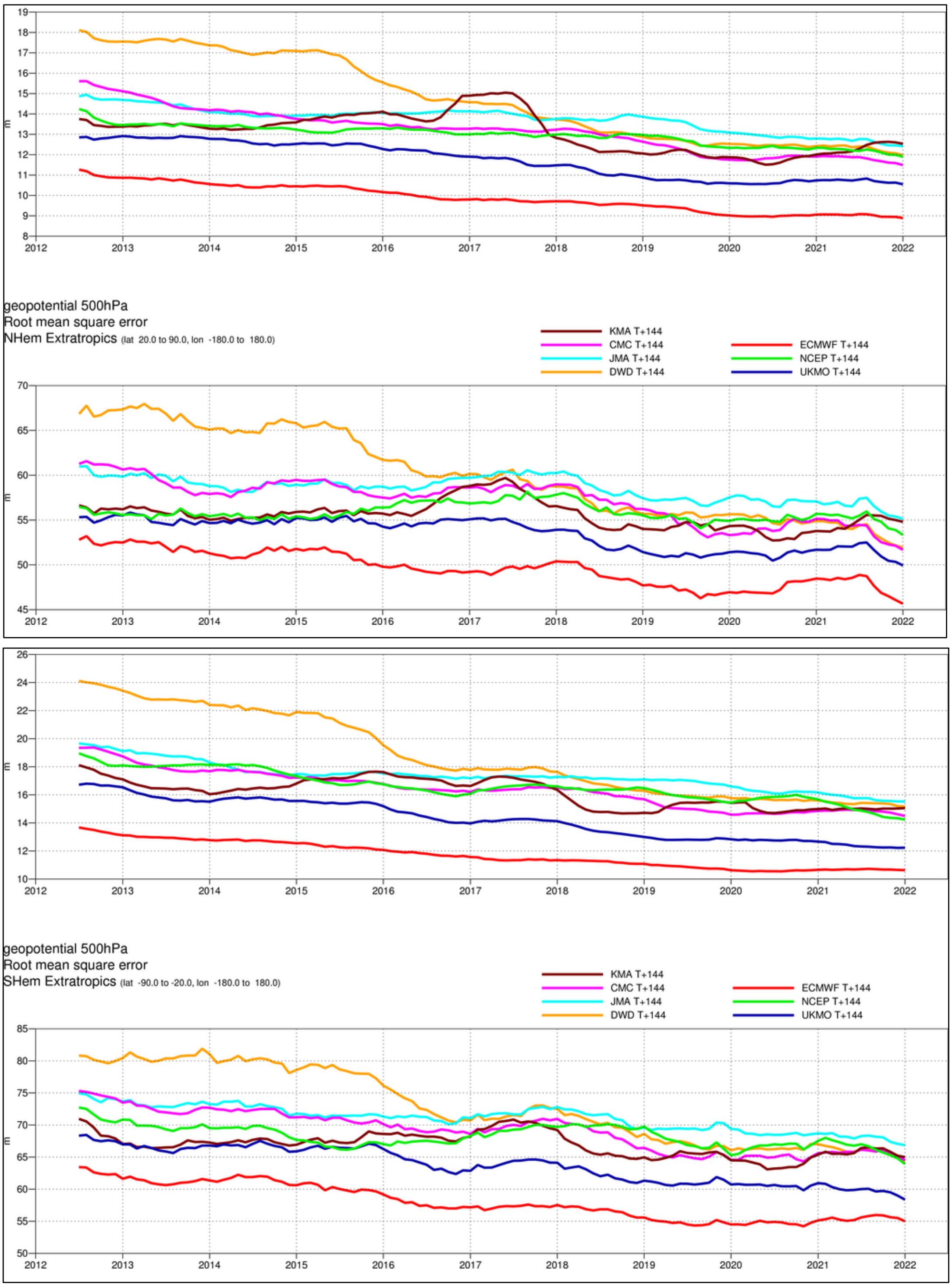
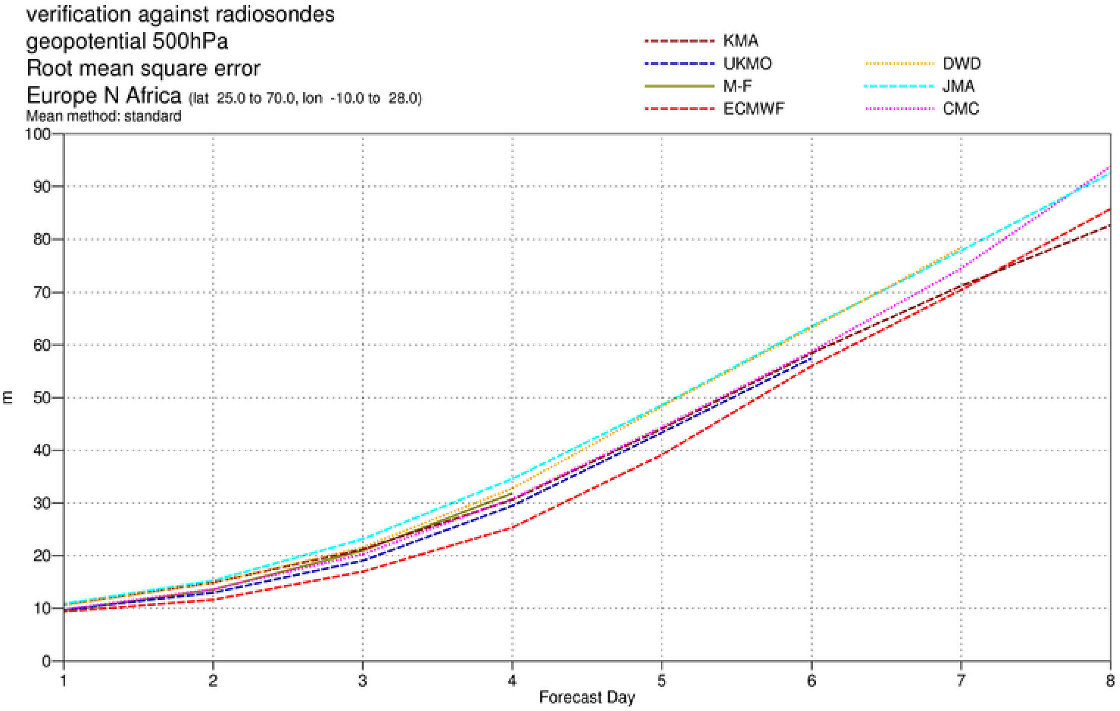


Figure 15: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top box) and southern (bottom box) extratropics. In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis. JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, KMA = Korea Meteorological Administration, NCEP = U.S. National Centers for Environmental Prediction, DWD = Deutscher Wetterdienst.



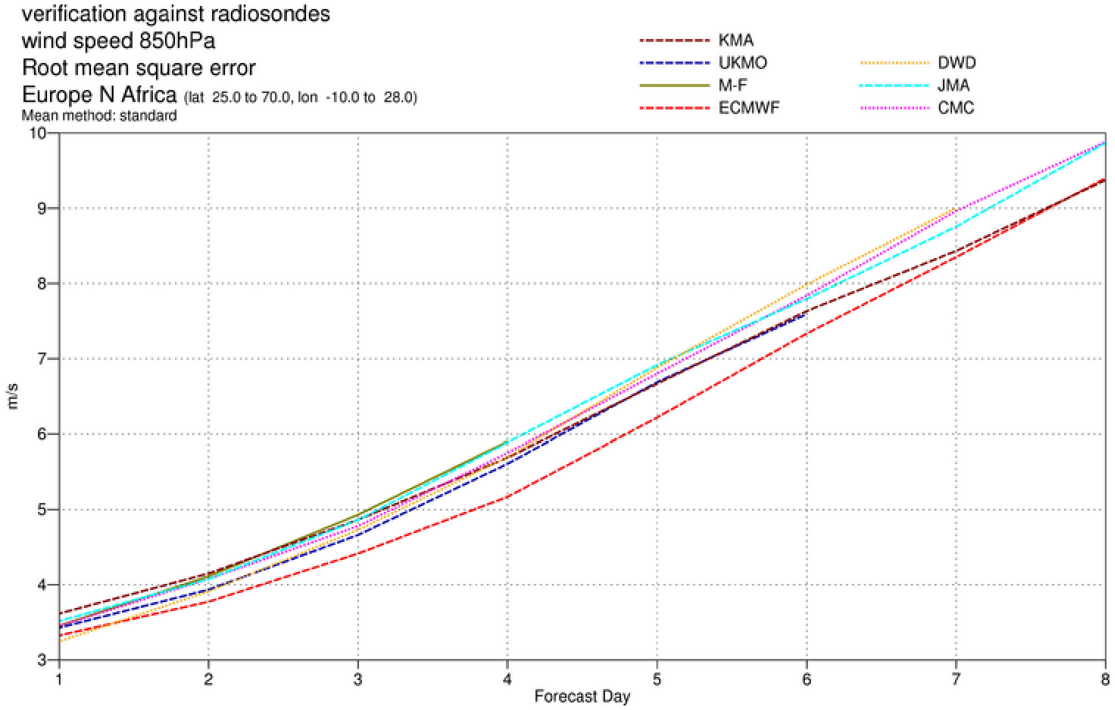
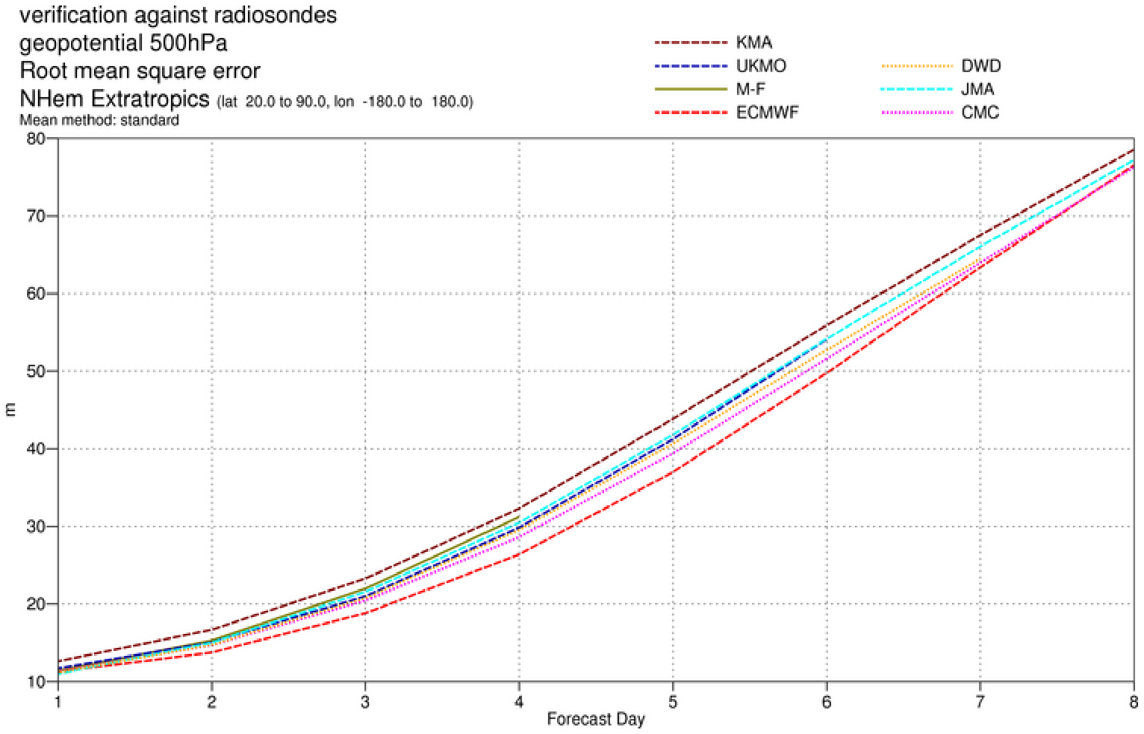


Figure 16**:** WMO-exchanged scores for verification against radiosondes: 500 hPa height (top) and 850 hPa wind (bottom) RMS error over Europe and North Africa (annual mean August 2021–July 2022) of forecast runs initialized at 12 UTC.



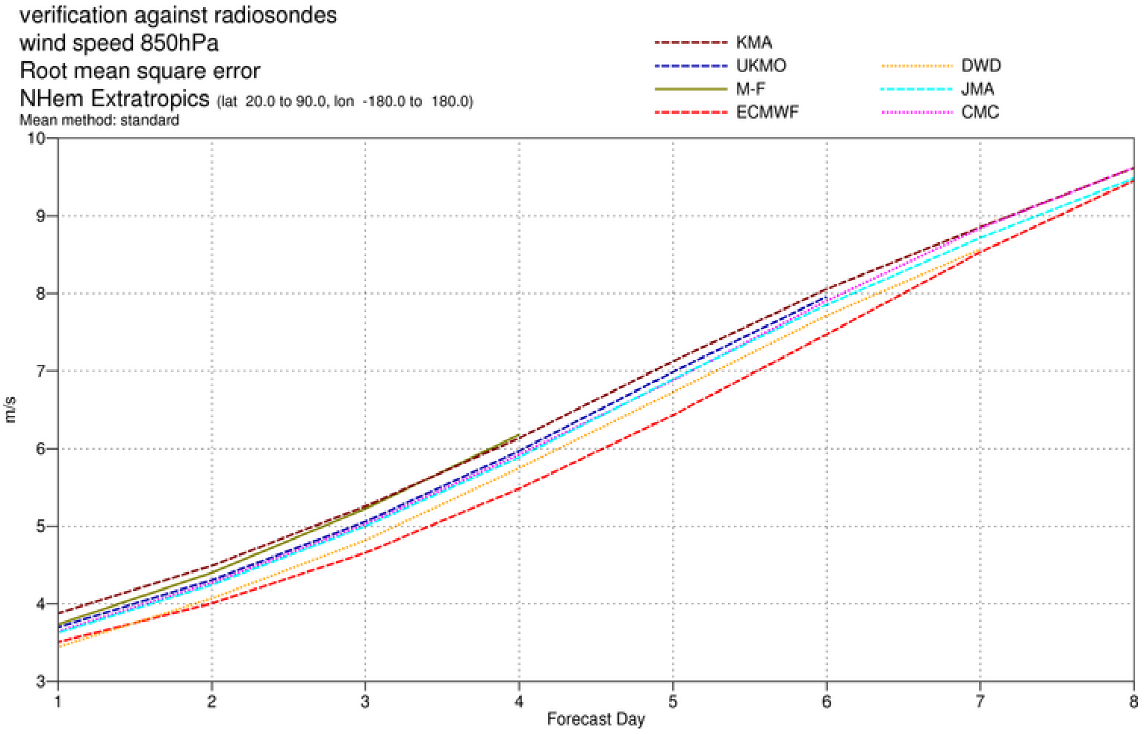


Figure 17**:** As Figure 16 for the northern hemisphere extratropics.

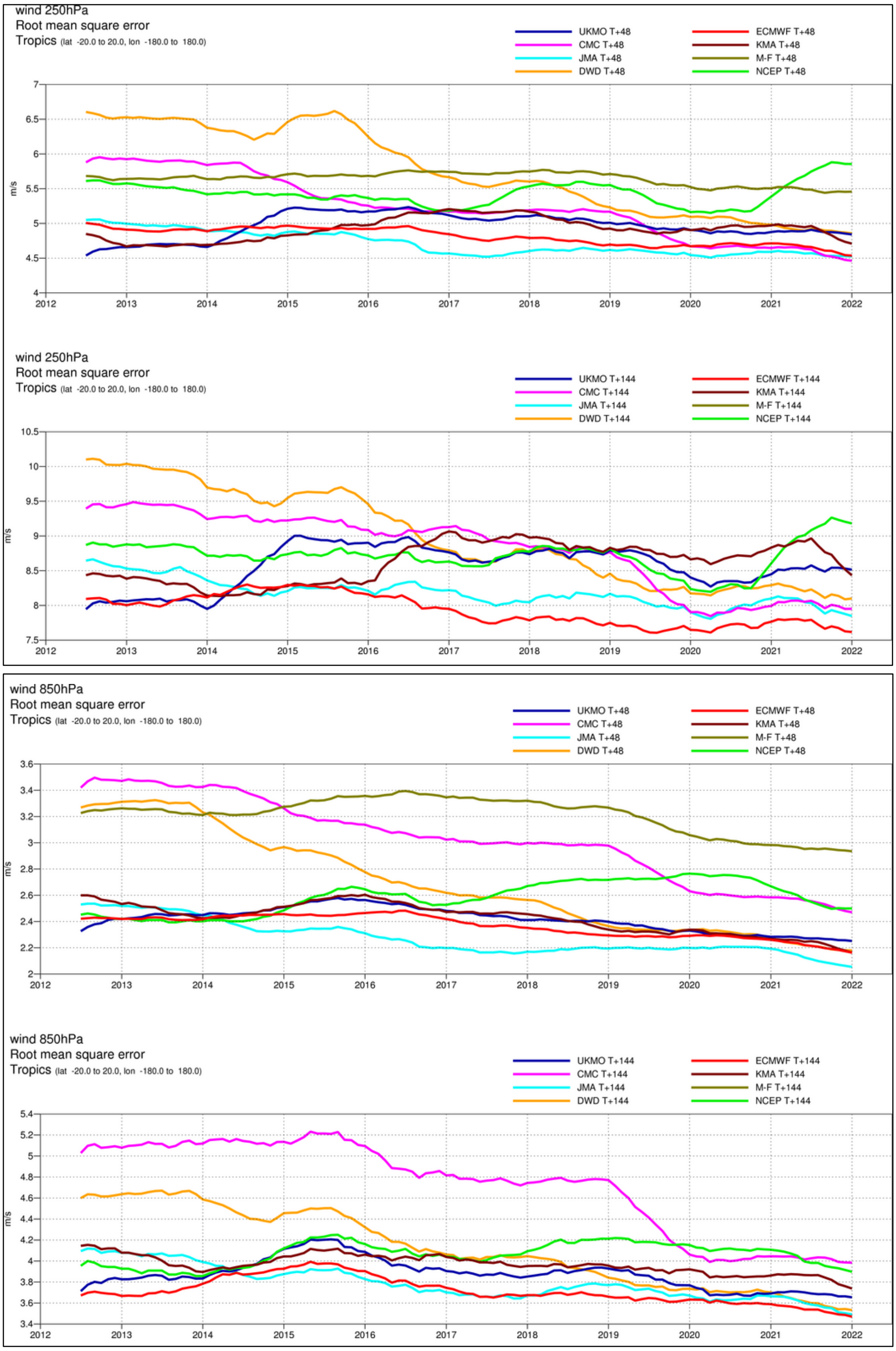


Figure 18**:** WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top box) and 850 hPa (bottom box). In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis.

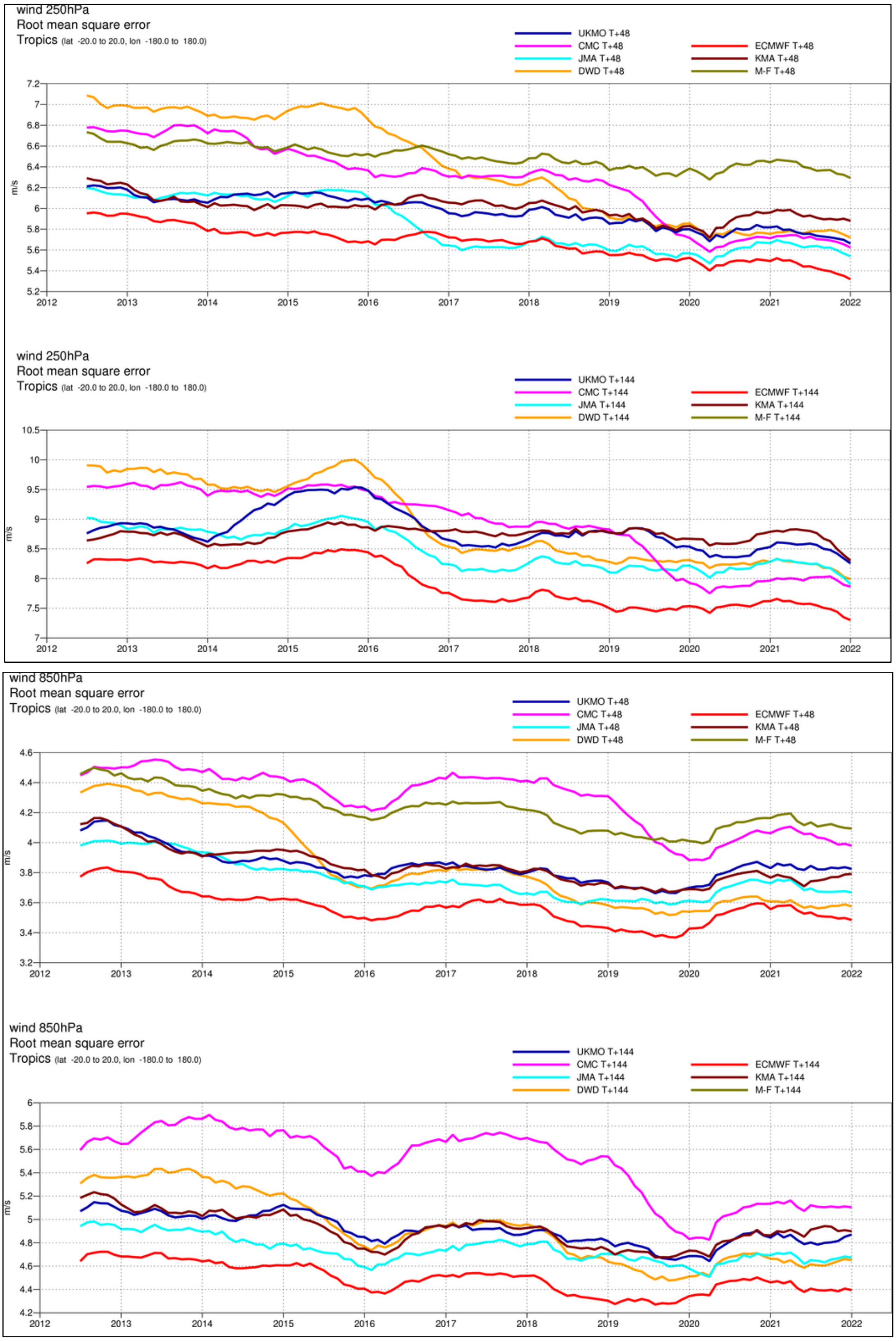


Figure 19**:** As Figure 18 but for verification against radiosonde observations.

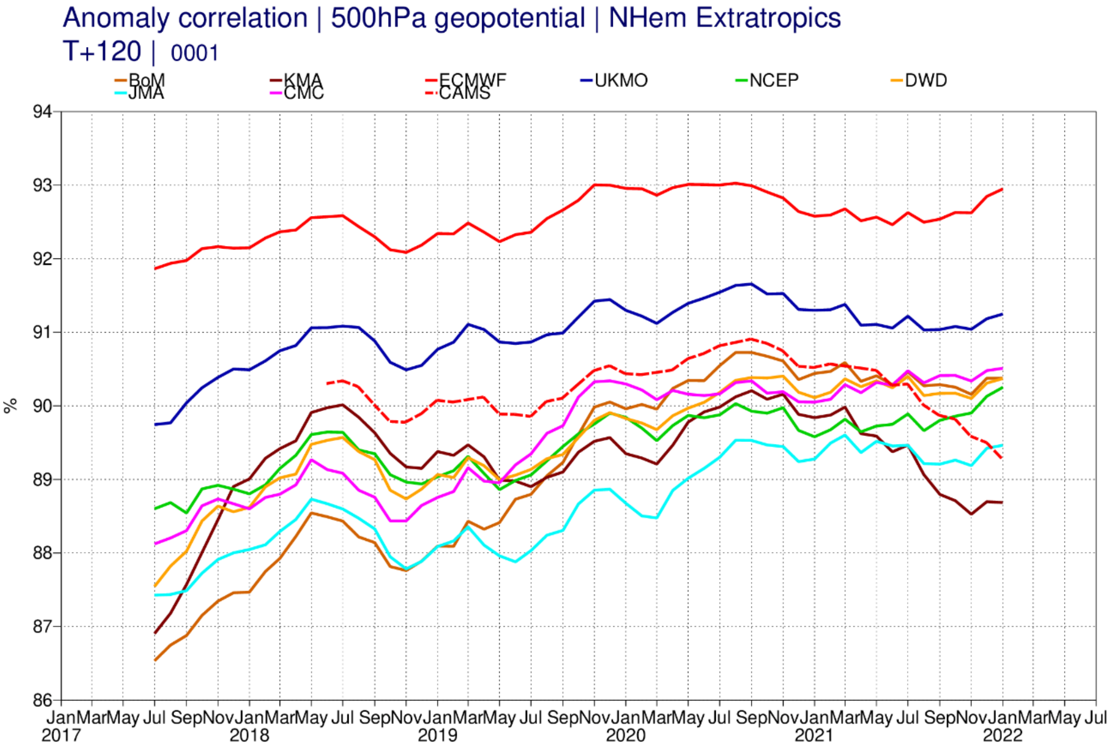


Figure 20: Anomaly correlation of 500 hPa geopotential in the northern hemisphere extratropics at day 5. CAMS forecast (red, dashed) shown in comparison to the HRES (red) and forecasts from other global centres.

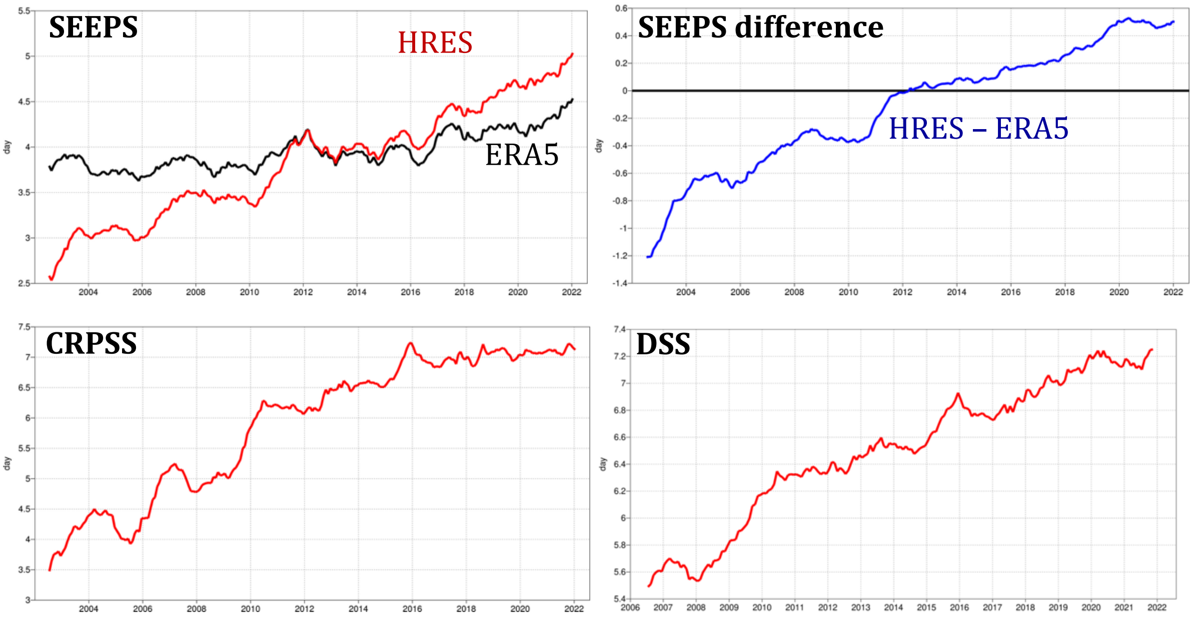
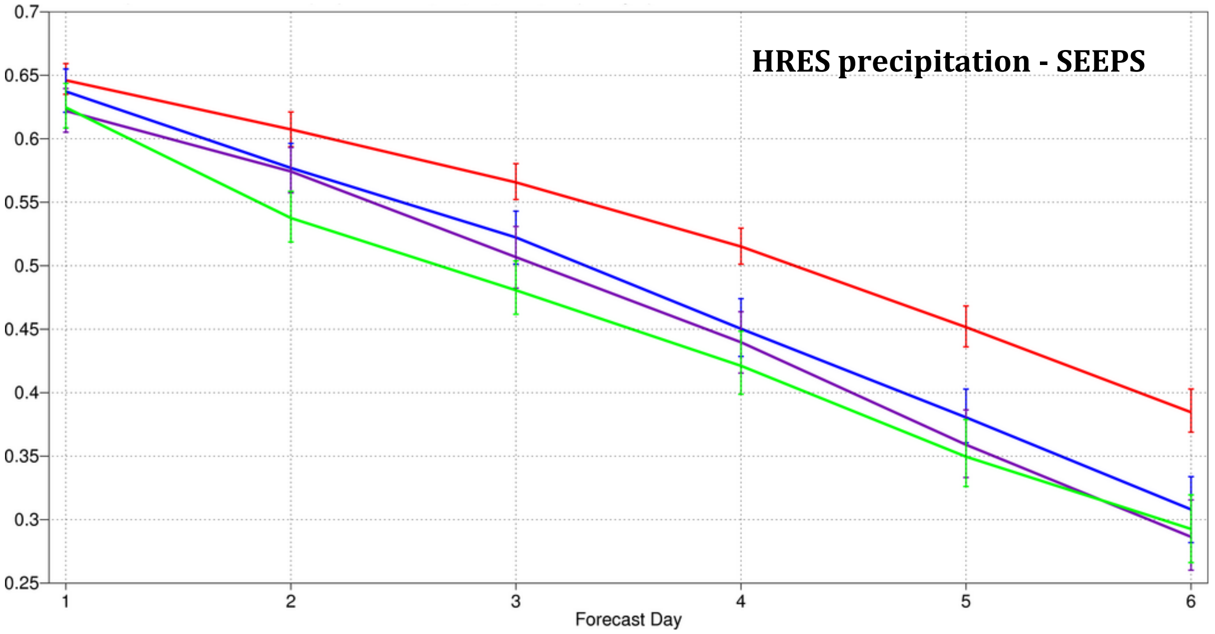


Figure 21**:** Supplementary headline scores (left column) and additional metrics (right column) for deterministic (top) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number 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 black curve in the top left panel shows the deterministic headline score for ERA5, and the top right panel shows the difference between the operational forecast and ERA5 (blue). Probabilistic scores in the bottom row are the Continuous Ranked Probability Skill Score (CRPSS) and the Diagonal Skill Score (DSS).



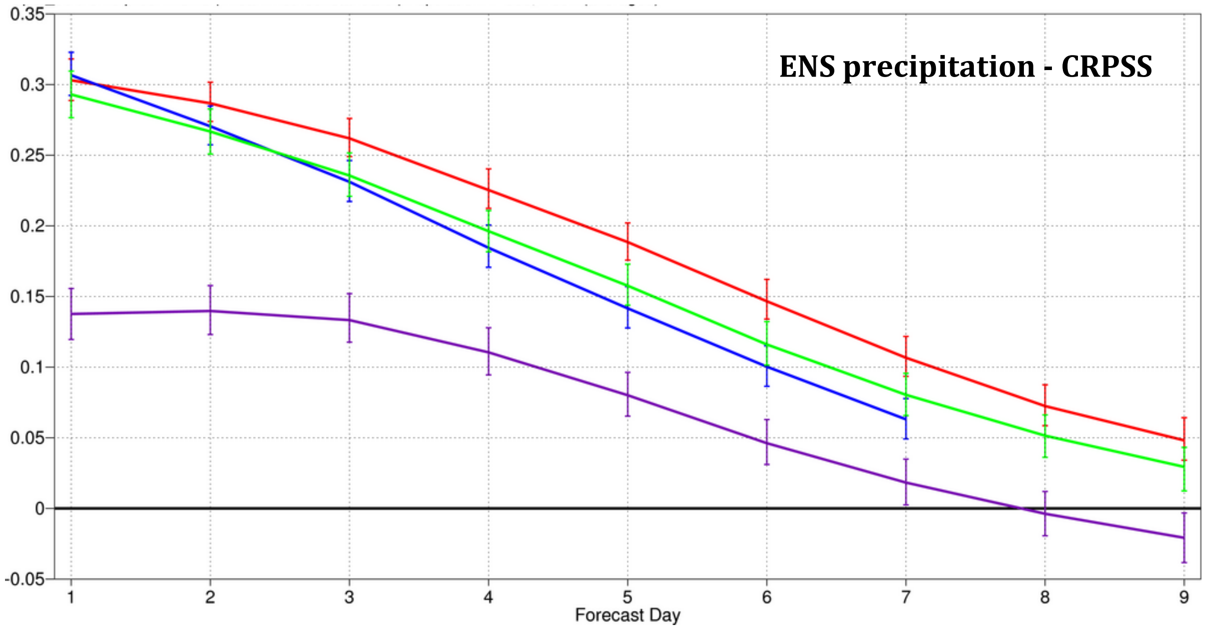


Figure 22**:** Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 21. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2021–July 2022. Bars indicate 95% confidence intervals.

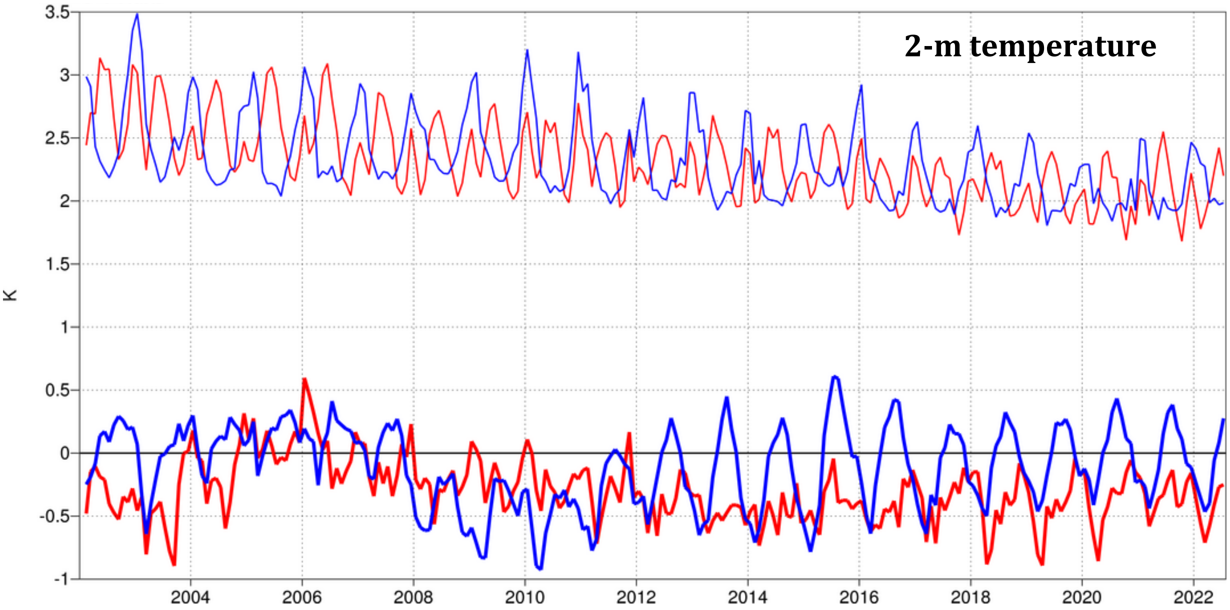


Figure 23**:** Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

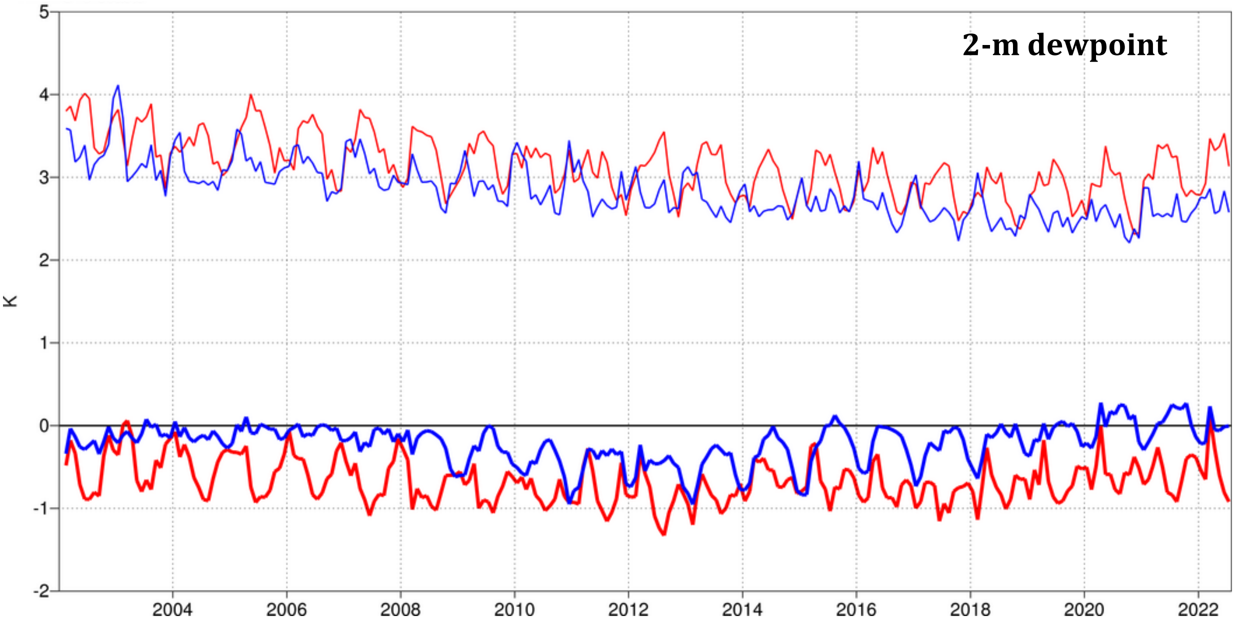


Figure 24**:** Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

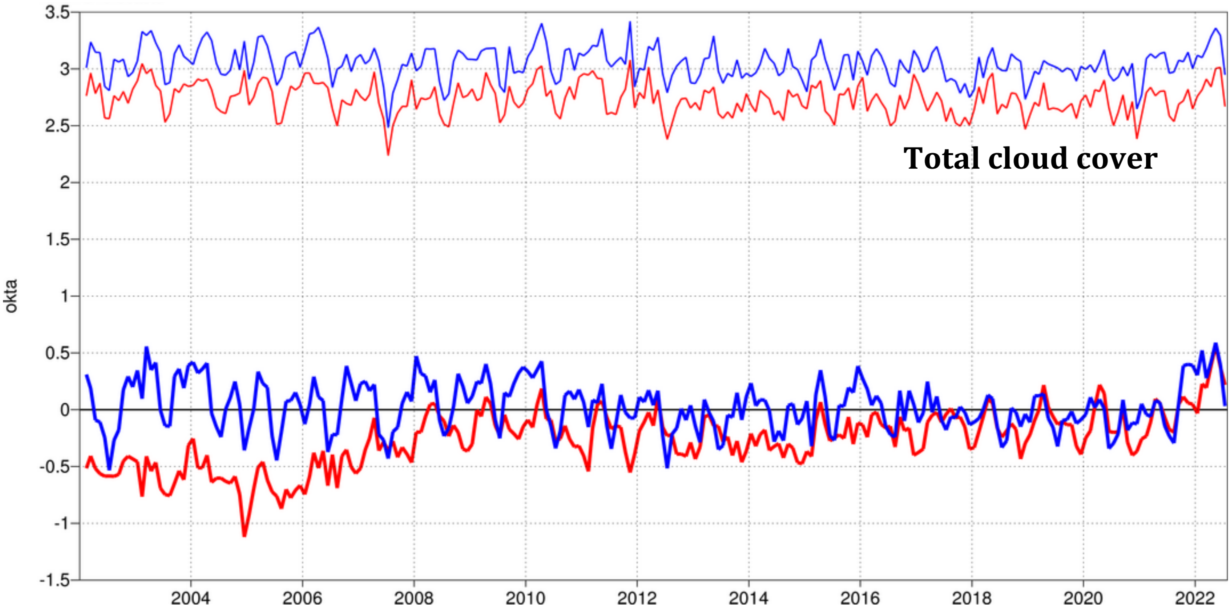


Figure 25**:** Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

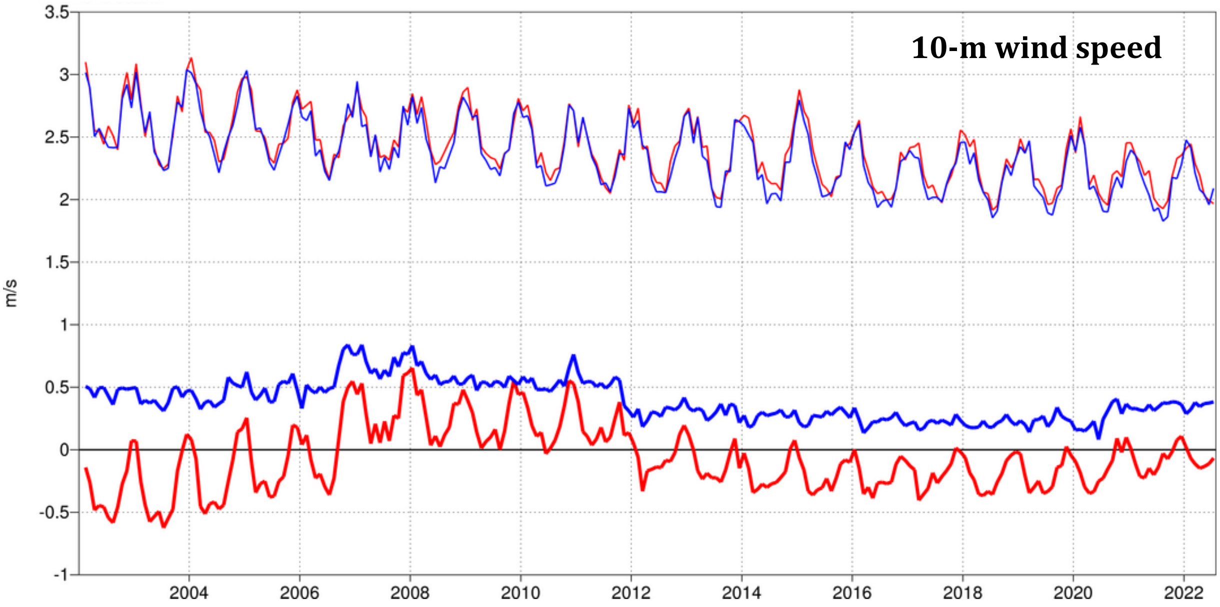


Figure 26**:** Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

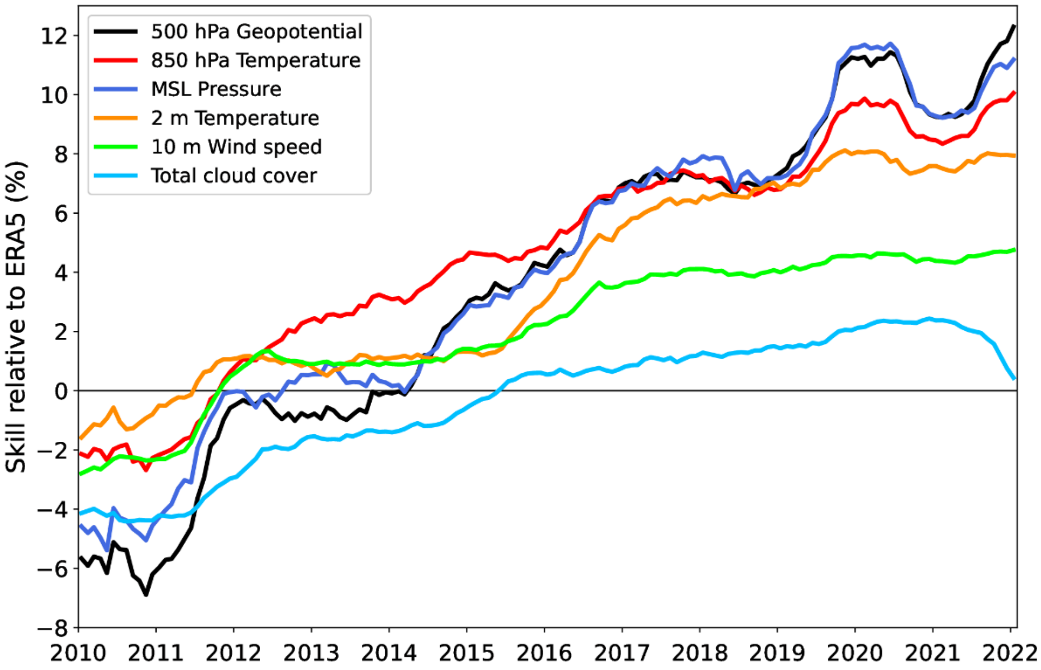


Figure 27: Evolution of skill of the HRES forecast at day 5, expressed as relative skill compared to ERA5. Verification is against analysis for 500 hPa geopotential, 850 hPa temperature, and mean sea level pressure, using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature, 10 m wind speed, and total cloud cover.

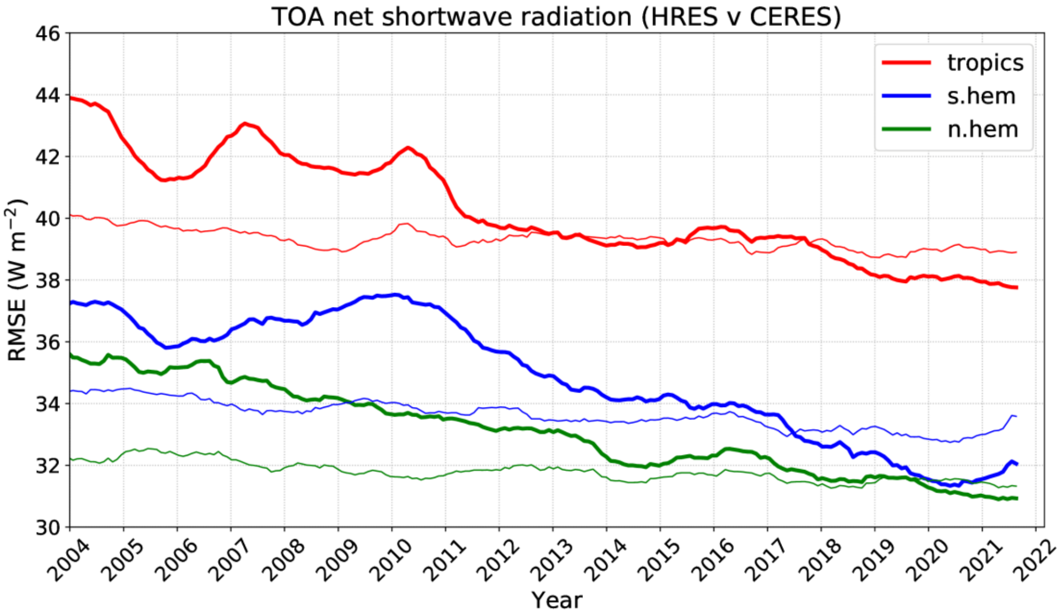


Figure 28: Evolution of the RMSE of the HRES forecast at day 5 (bold lines) of the top of the atmosphere (TOA) net shortwave radiation for the two extratropical hemisphere and the tropics. Thin lines show the RMSE of the ERA5 forecast for comparison. Verification is against CERES satellite data.

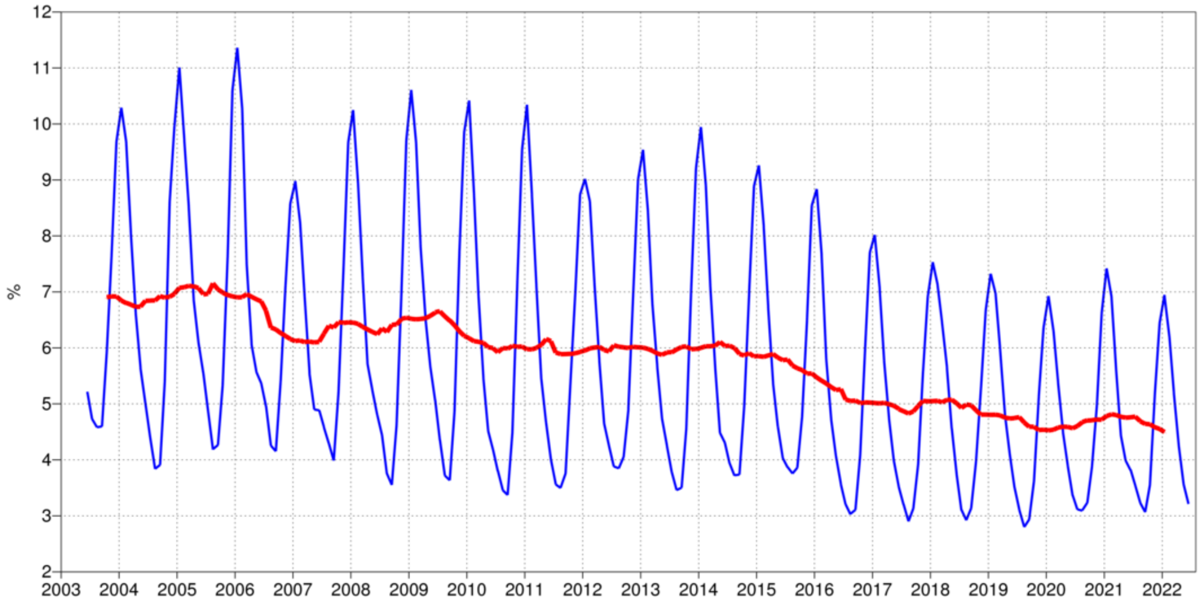


Figure 29: Evolution of the fraction of large 2m temperature errors (CRPS>5K) in the ENS at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

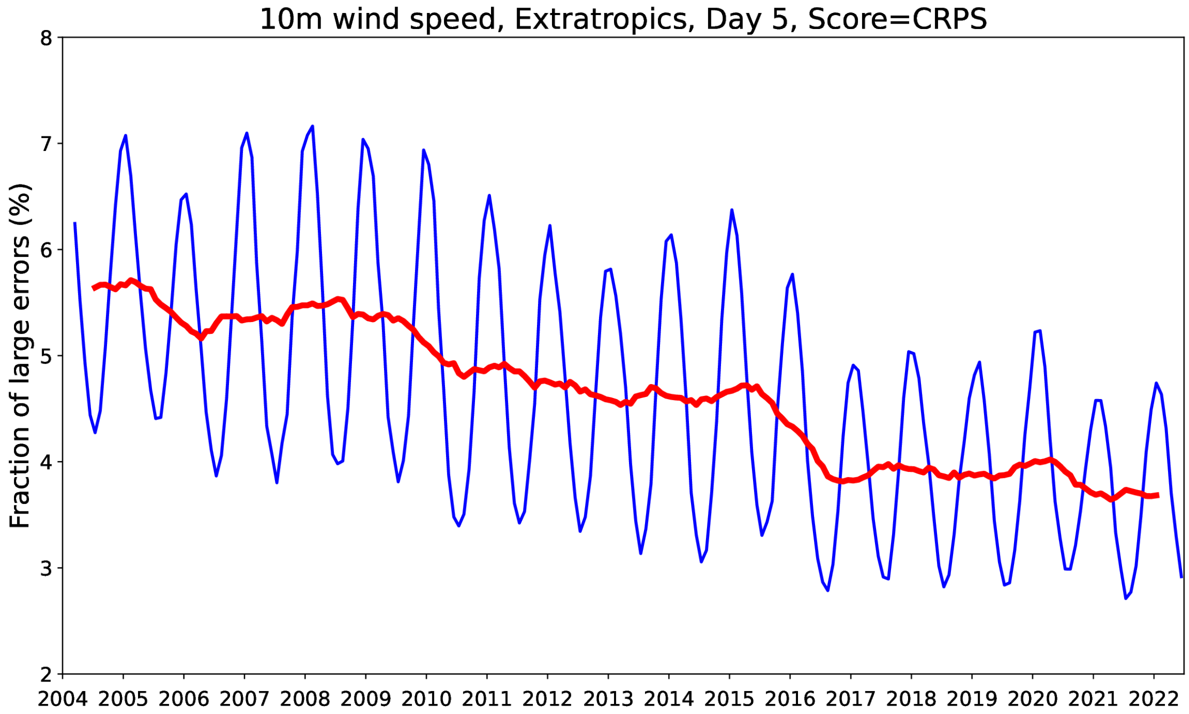
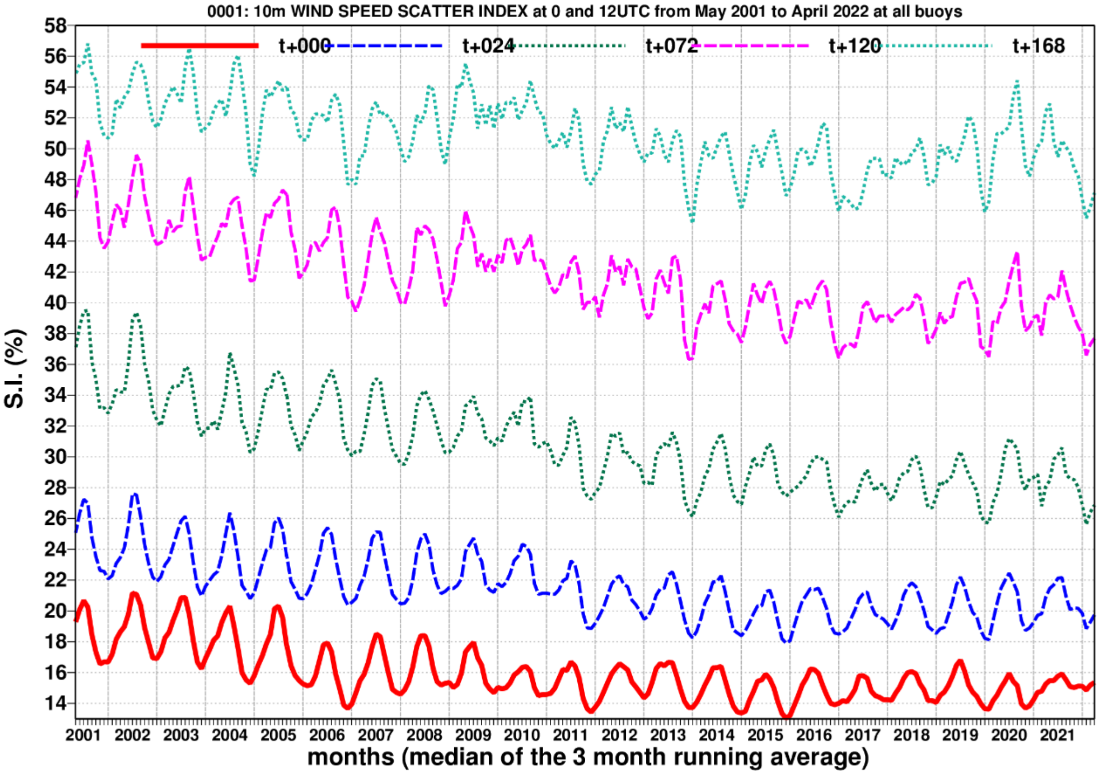


Figure 30: Evolution of the fraction of large 10m wind speed errors (CRPS>4m/s) in the ENS at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.



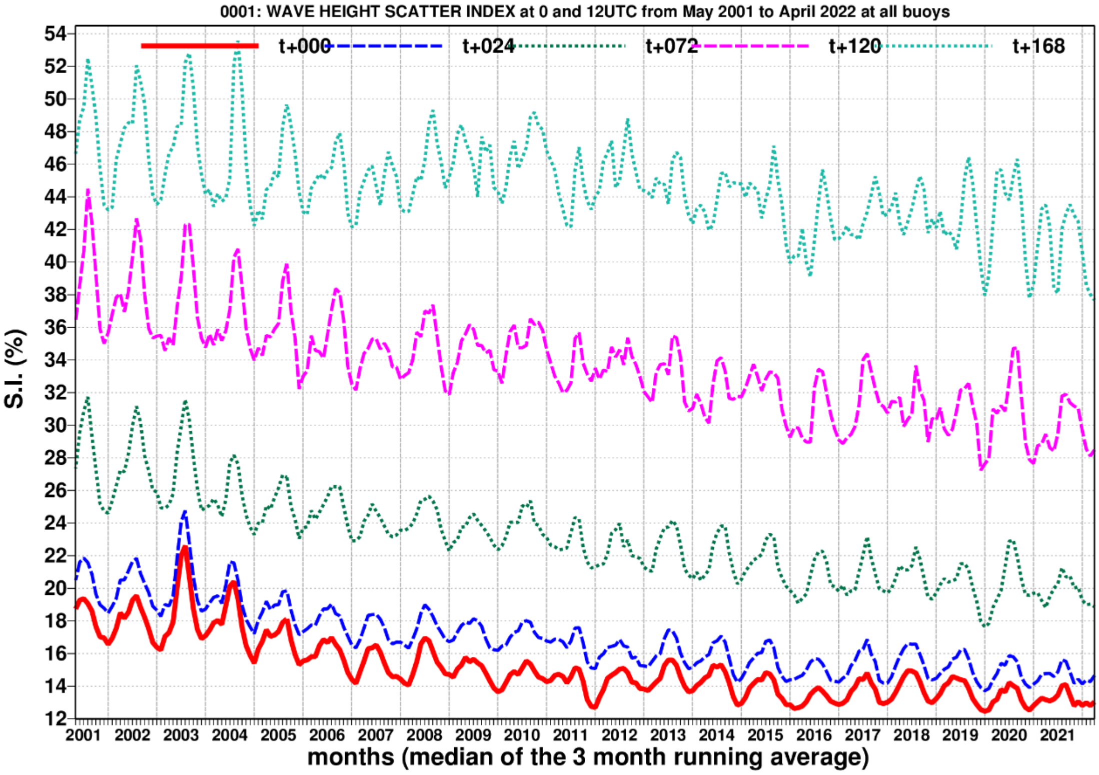


Figure 31: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave model forecast (wave height, bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.

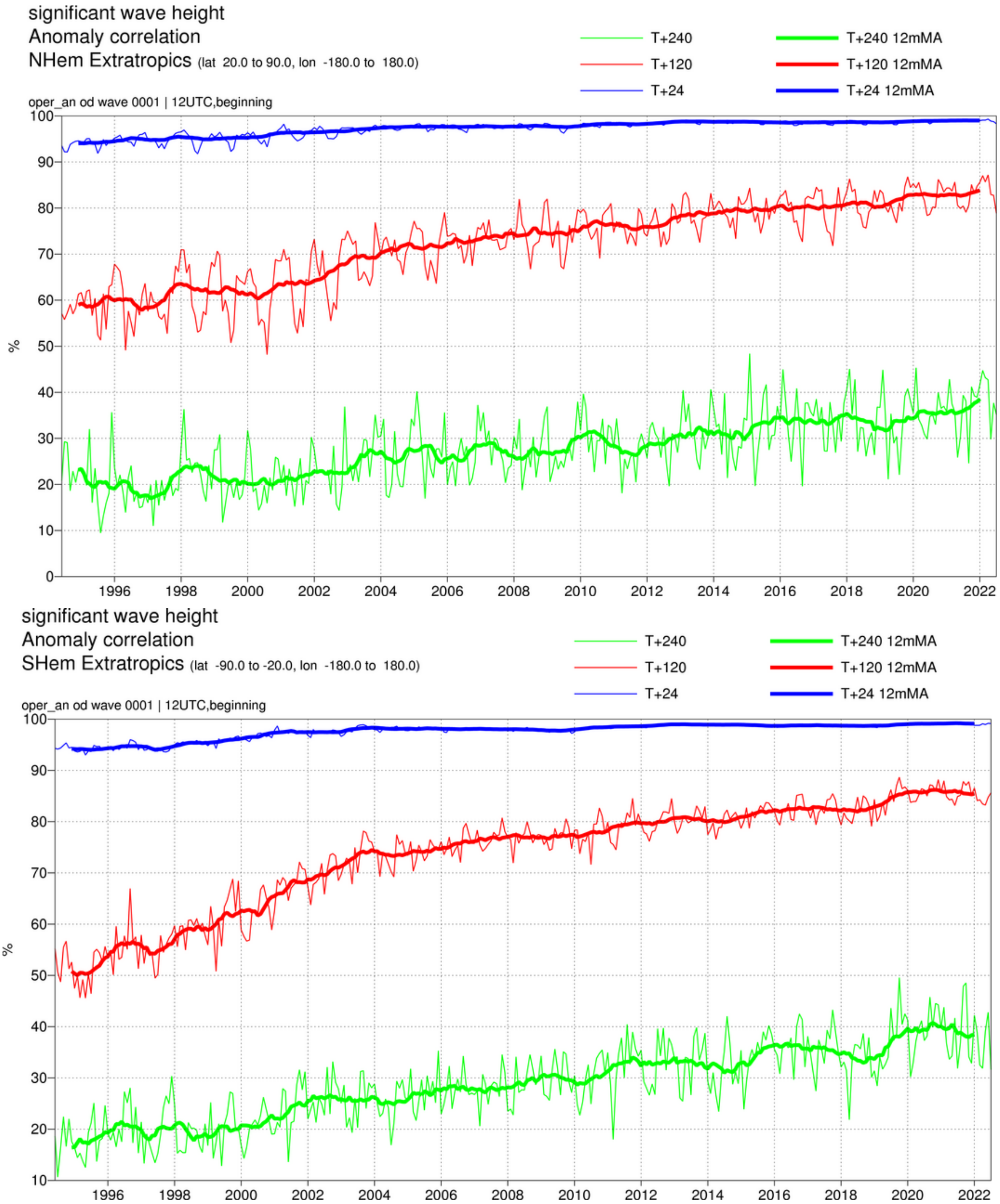


Figure 32: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

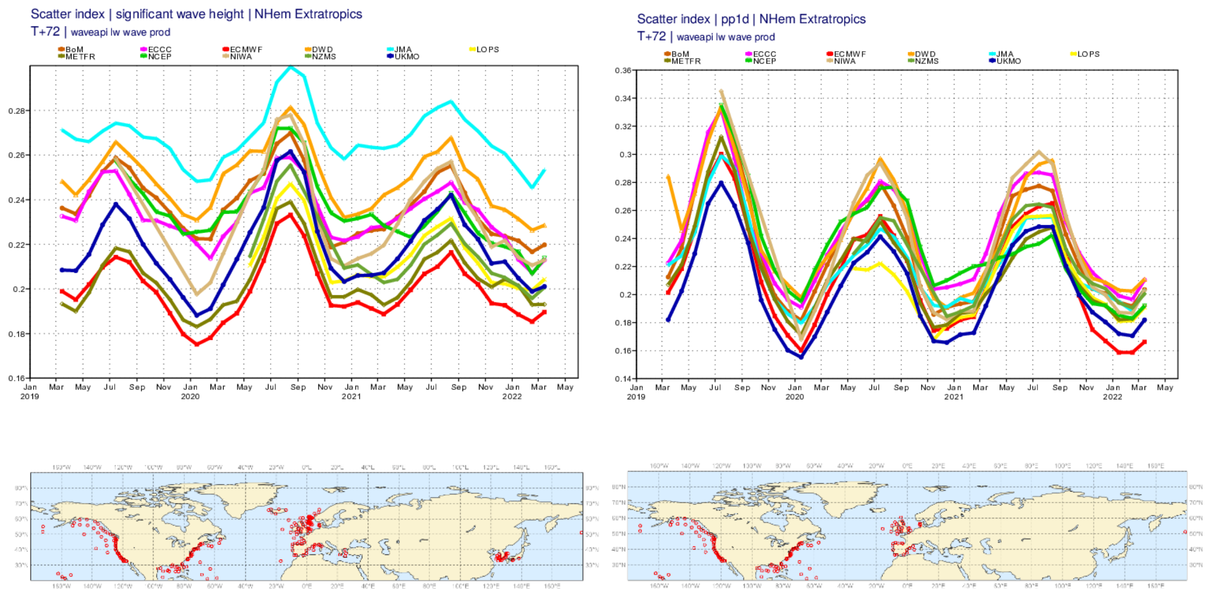


Figure 33: Verification of forecasts of wave height and peak wave period (upper panels) at +72 h using observations from wave buoys (lower panels). The scatter index (SI) is the standard deviation of error normalised by the mean observed value. METFR: Météo-France; JMA: Japan Meteorological Agency; ECCC: Environment and Climate Change Canada; BoM: Bureau of Meteorology, Australia; LOPS: Laboratory for Ocean Physics and Satellite remote sensing, France; NZMS: New Zealand Meteorological Service; DWD: Deutscher Wetterdienst, Germany; UKMO: Met Office, UK; NCEP: National Centers for Environmental Prediction, USA; NIWA: National Institute of Water and Atmosperic Research, New Zealand.

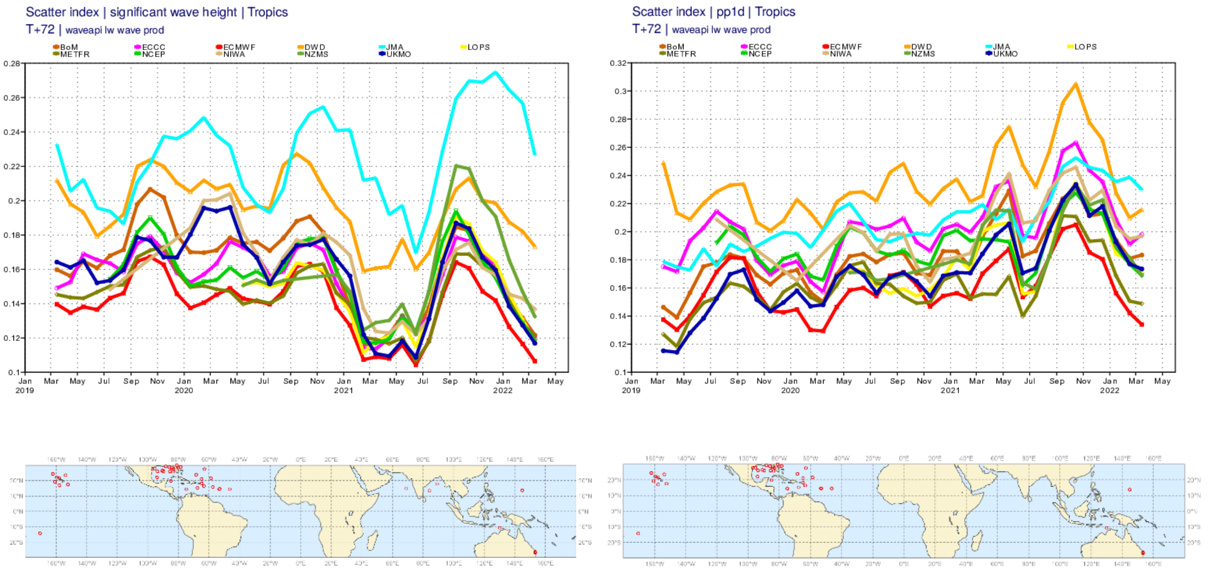


Figure 34: As Figure 33, but for the tropics.

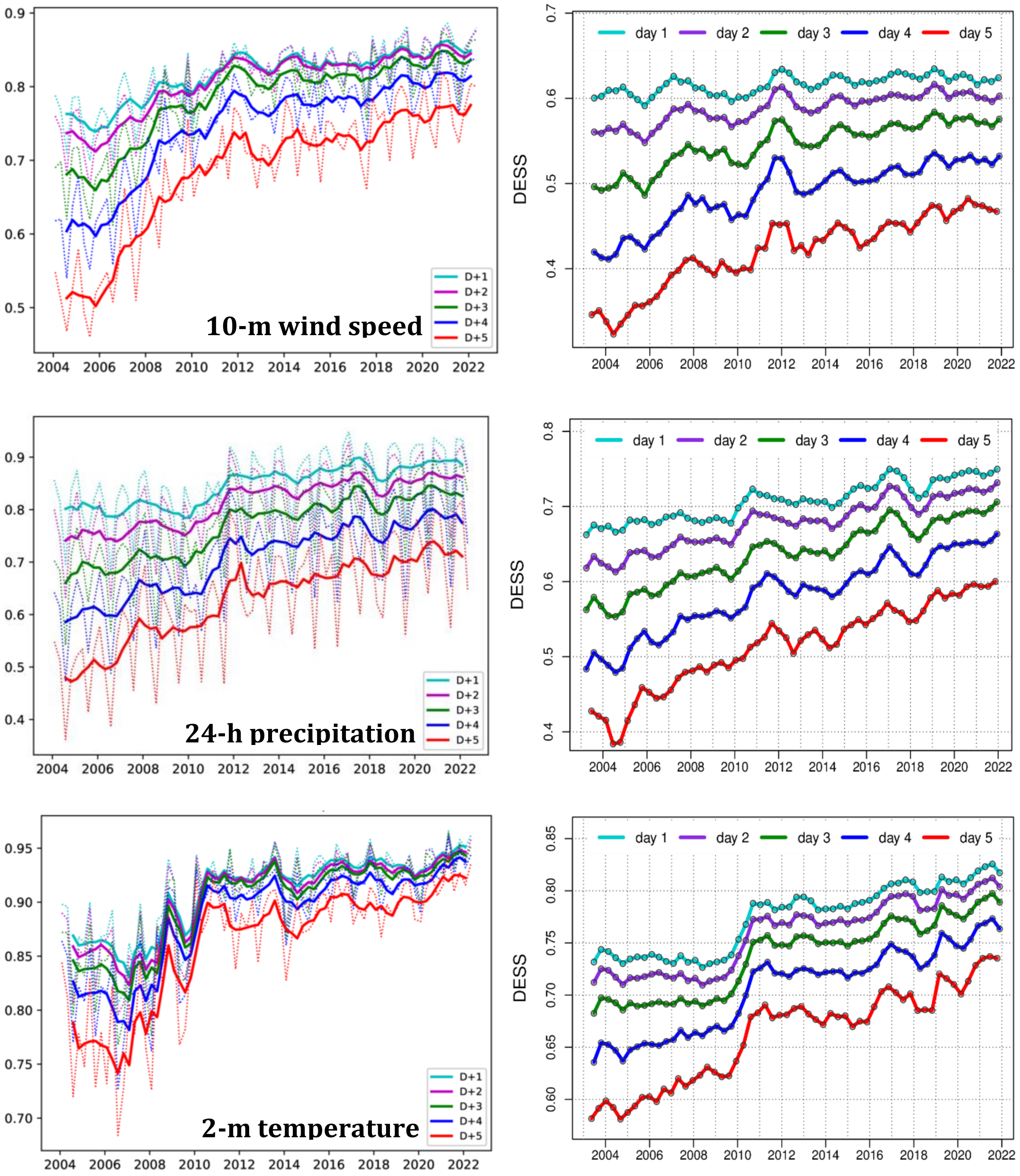


Figure 35: Verification of Extreme Forecast Index (EFI) against analysis (left column). Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 5 (24-hour period 96–120 hours ahead); skill at day 4 (blue line) is the supplementary headline score; an extreme event is taken 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 on the left show the equivalent ROC area skill scores for precipitation EFI forecasts and for 2 m temperature EFI forecasts. Diagonal elementary skill score (DESS) for the 95th percentile for the same three variables, taking observation uncertainty into account (right column).

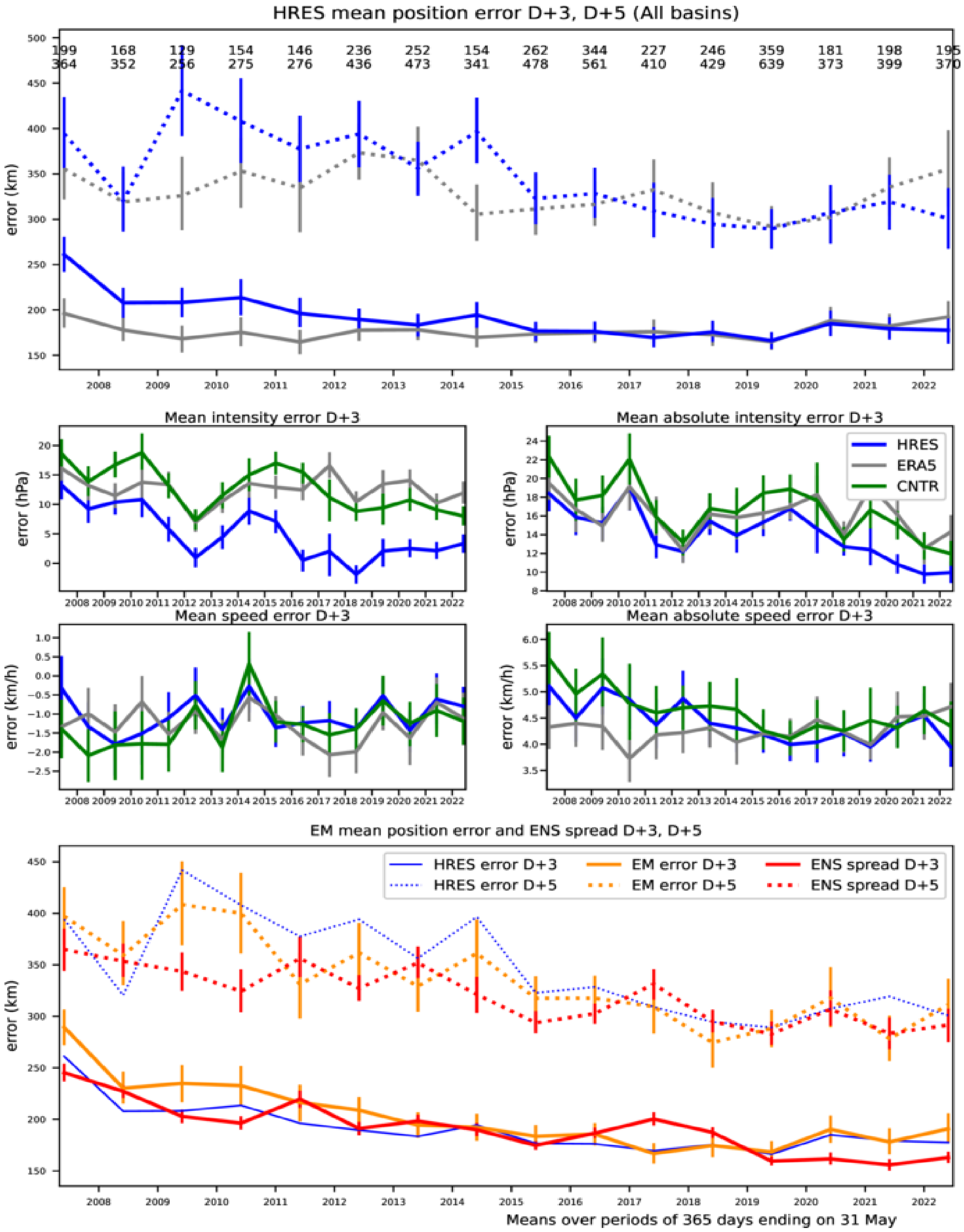


Figure 36: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary 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). For reference, errors of tropical cyclone forecasts by ERA5 are shown in grey.

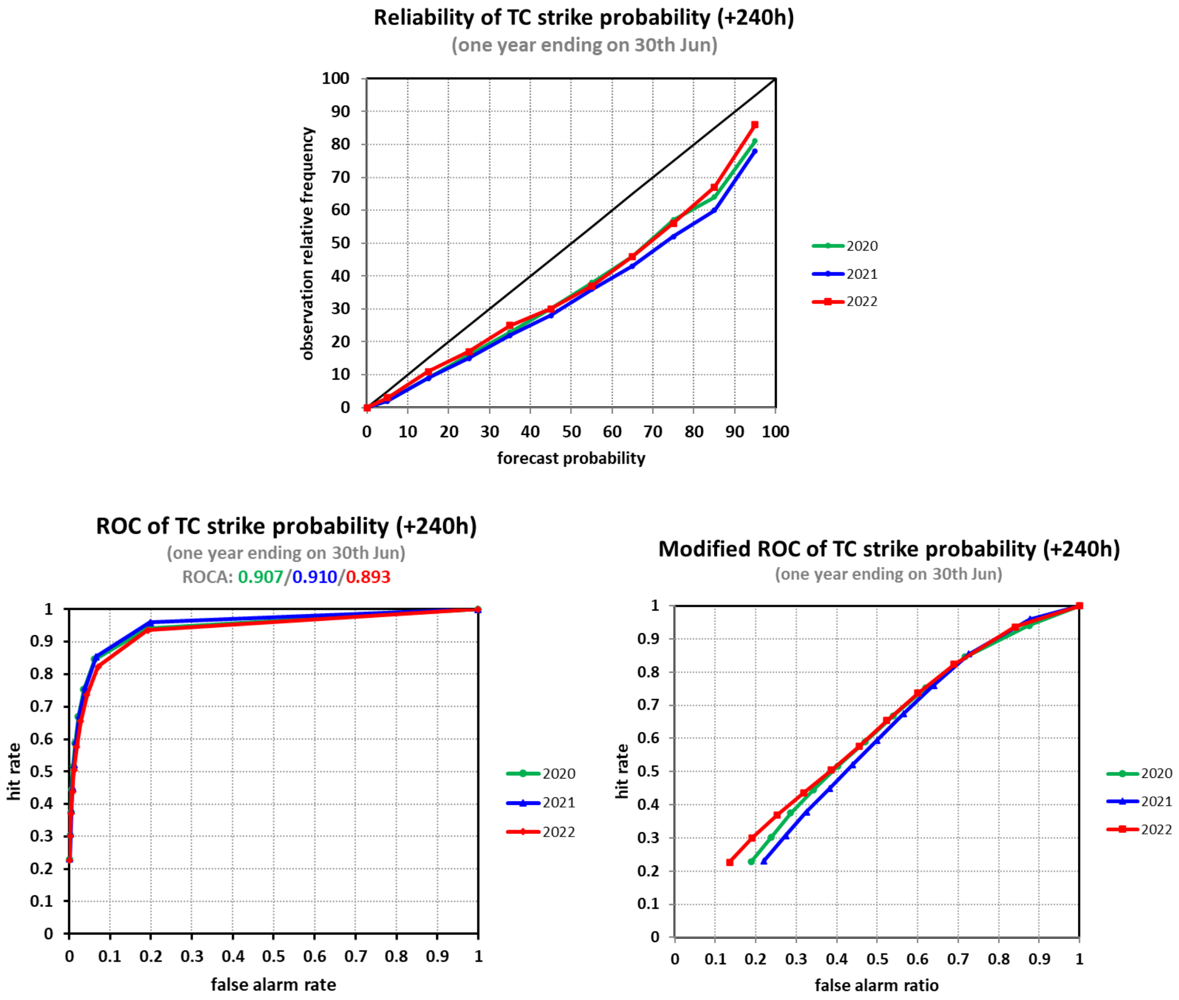


Figure 37: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2019–June 2020 (green), July 2020–June 2021 (blue) and July 2021–June 2022 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

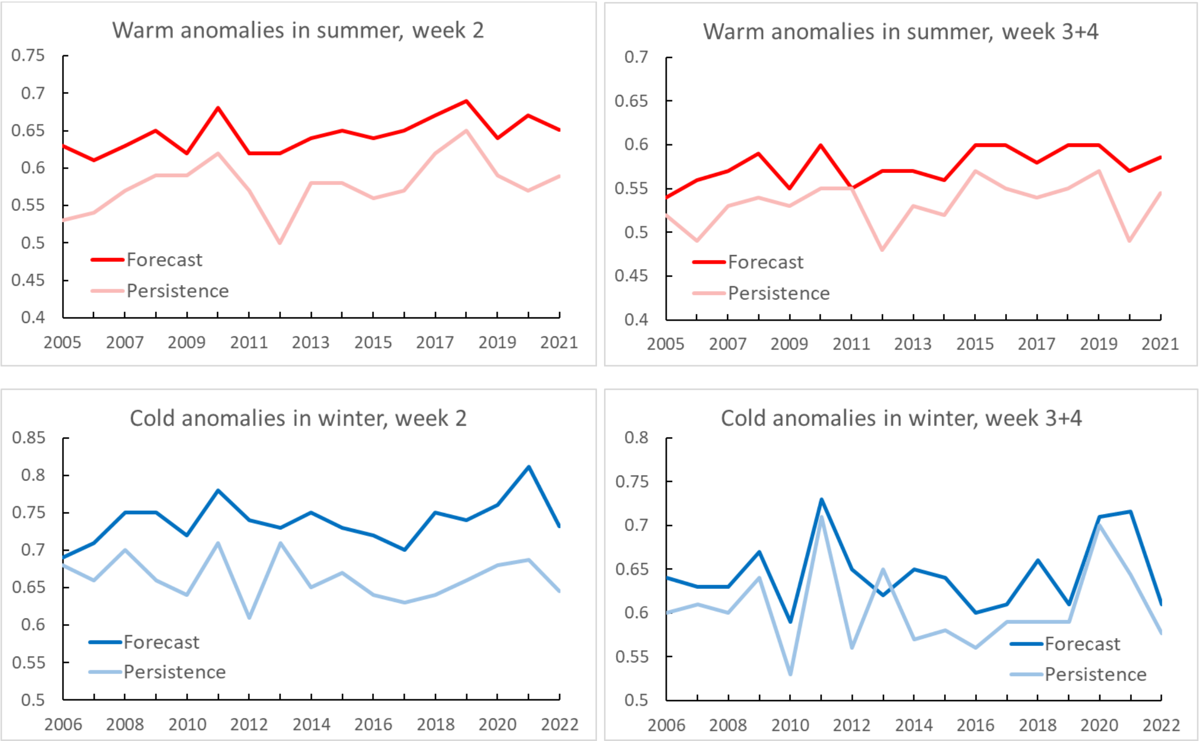


Figure 38**:** Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all 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.

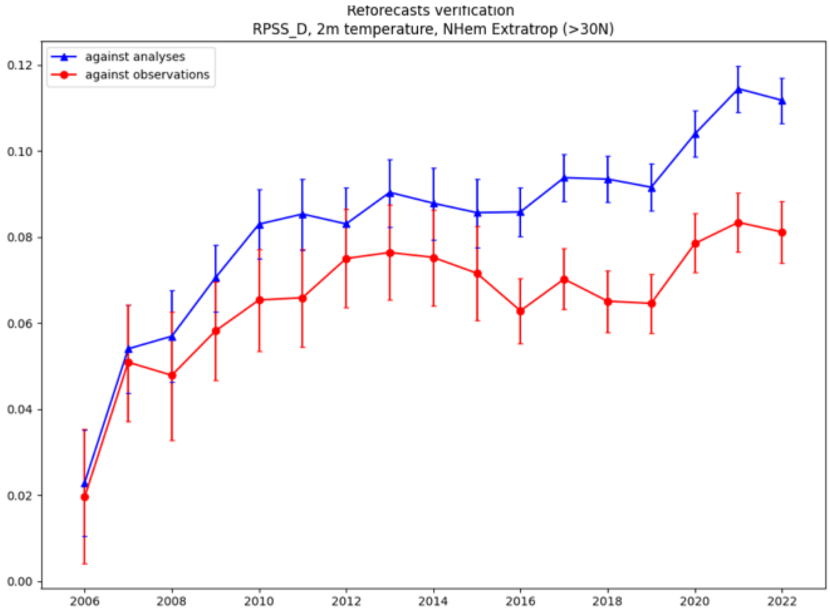
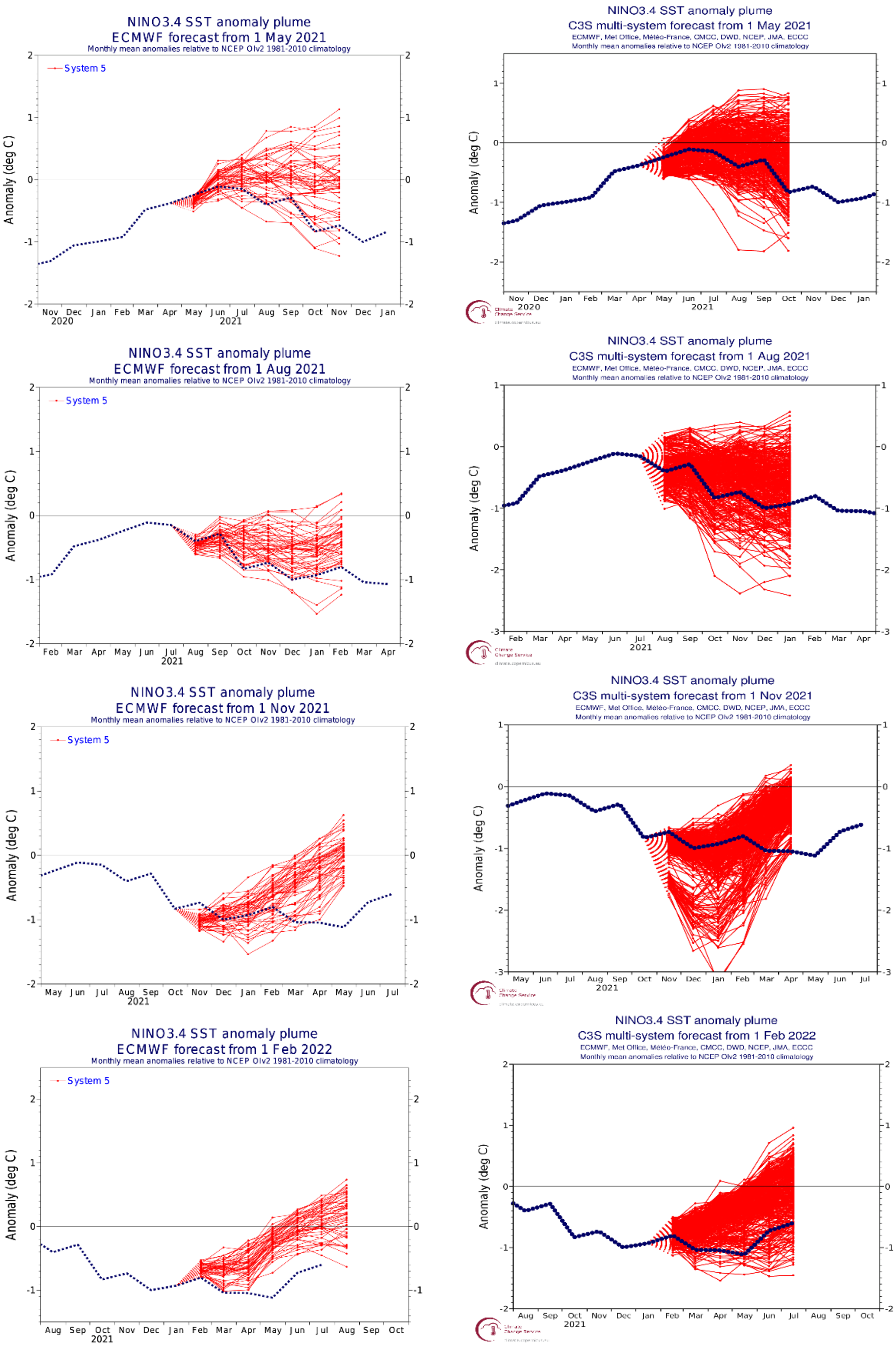


Figure 39: Skill of the ENS in predicting weekly mean 2m temperature anomalies (terciles) in week 3 in the northern extratropics. Verification against ERA5 analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.



NINO3.4 SST anomaly plume
ECMWF forecast from 1 May 2021
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 May 2021
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Aug 2021
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Aug 2021
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Nov 2021
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Nov 2021
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Feb 2022
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Feb 2022
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to NCEP OIv2 1981-2010 climatology

System 5
Anomaly (deg C)
Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan
2020
2021
Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr
May Jun Jul Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul
Aug Sep Oct Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct
Climate Change Service

Figure 40: ECMWF System 5 (left column), and Copernicus Climate Change Service multi-model (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2021, August 2021, November 2021, and February 2022. The red lines represent the ensemble members; dotted blue line shows the subsequent verification. The C3S multi-model forecast includes forecasts from ECMWF, MetOffice, Meteo-France, CMCC, DWD, NCEP, JMA, and ECCC.

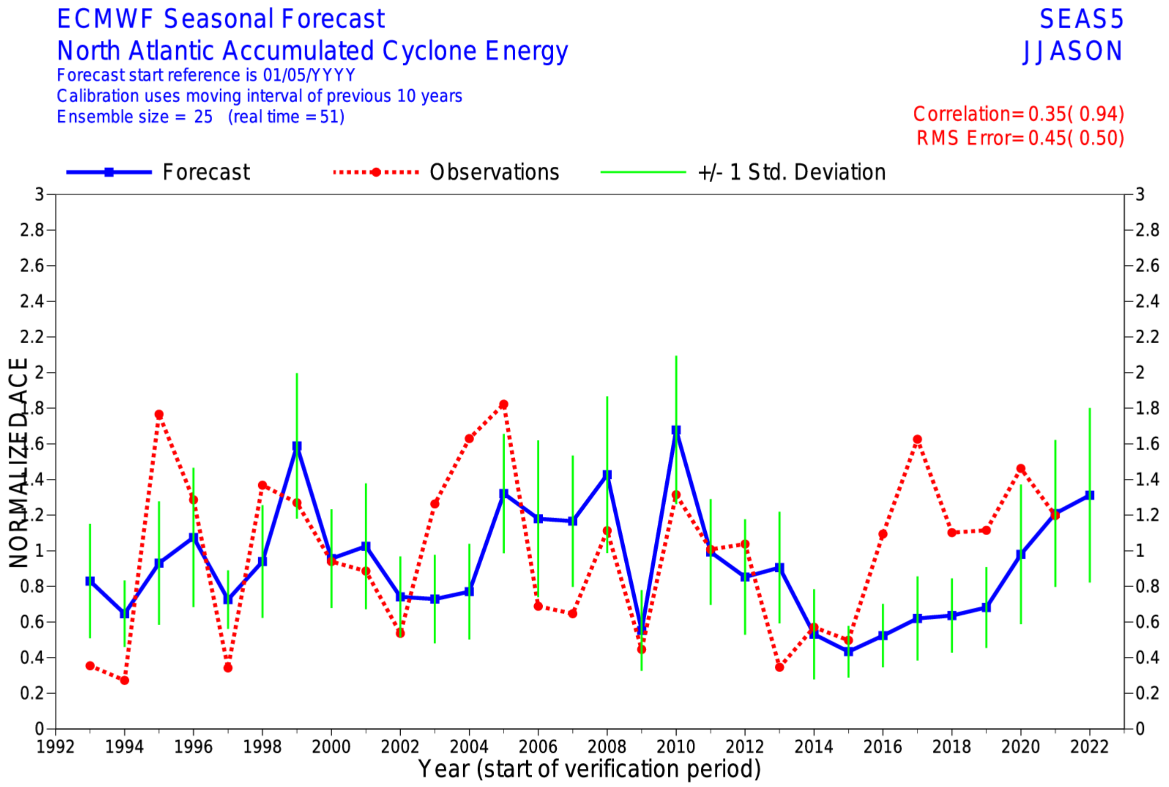


Figure 41: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1993 to July–December 2022. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts 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. Note that this plot is based on the new forecast calibration (based on the most recent 10 year running mean, rather than the fixed period 1993-2015 used before).

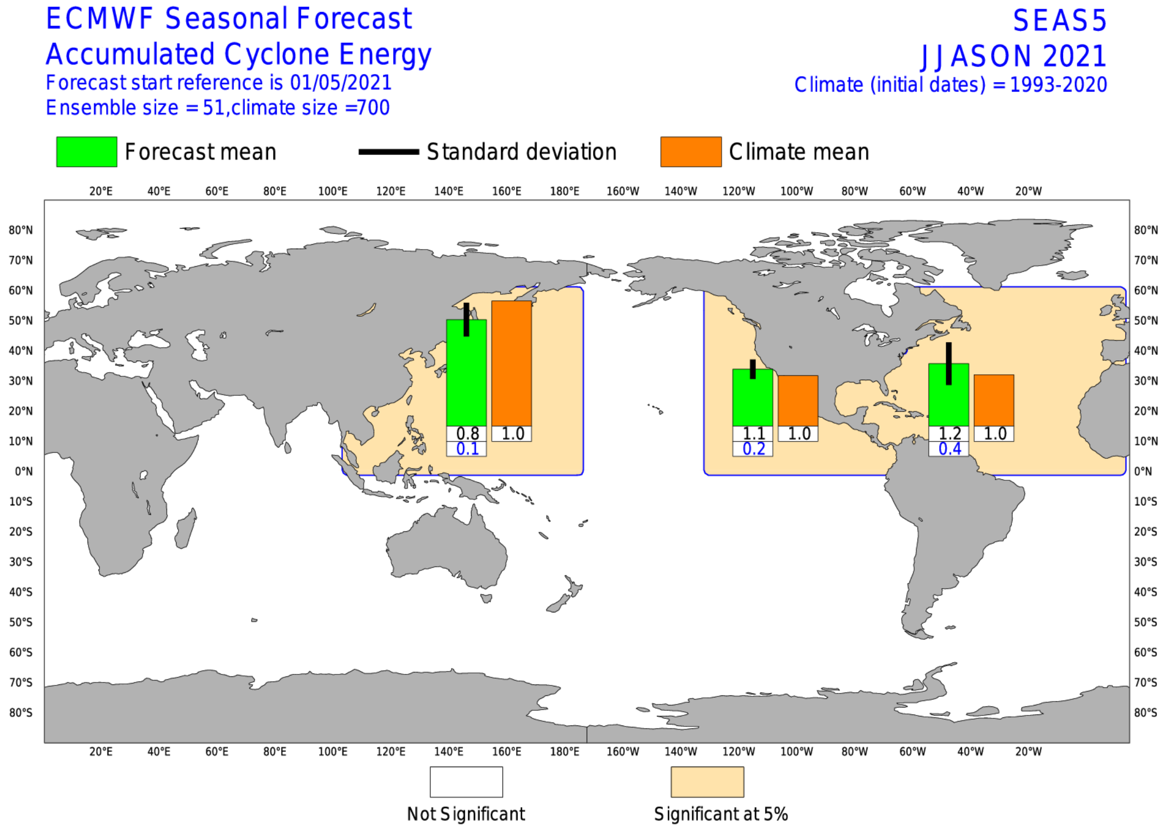


Figure 42**:** Forecast of tropical storm accumulated cyclone energy (ACE, normalized) issued in May 2021 for the six-month period June–November 2021. Green bars represent the forecast ACE in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black 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 ACE is significantly different from the climatology. The ocean basins where the WMW test detects significance larger than 90% have a shaded background.

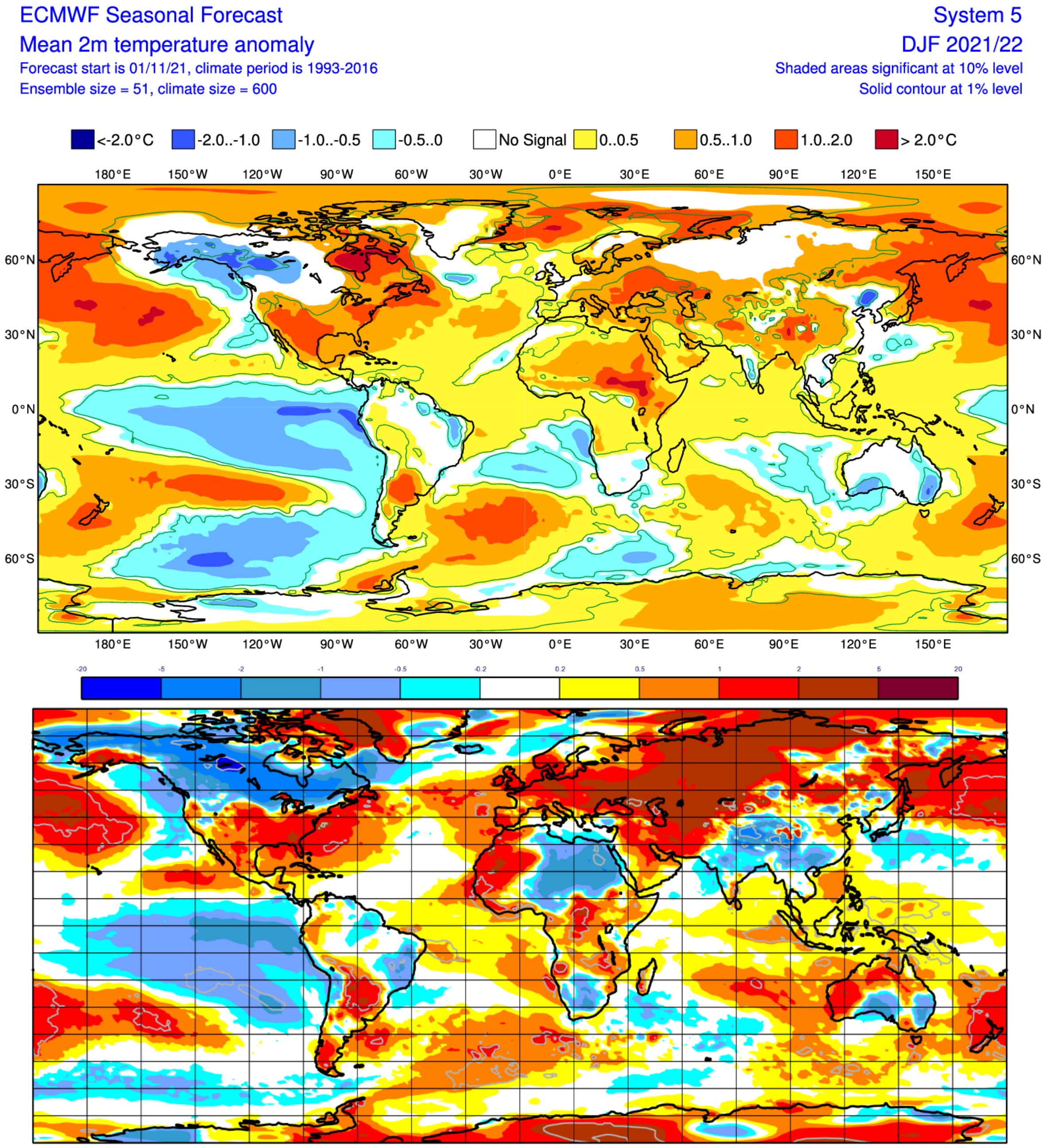


Figure 43**:** Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2021 for DJF 2021/22 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

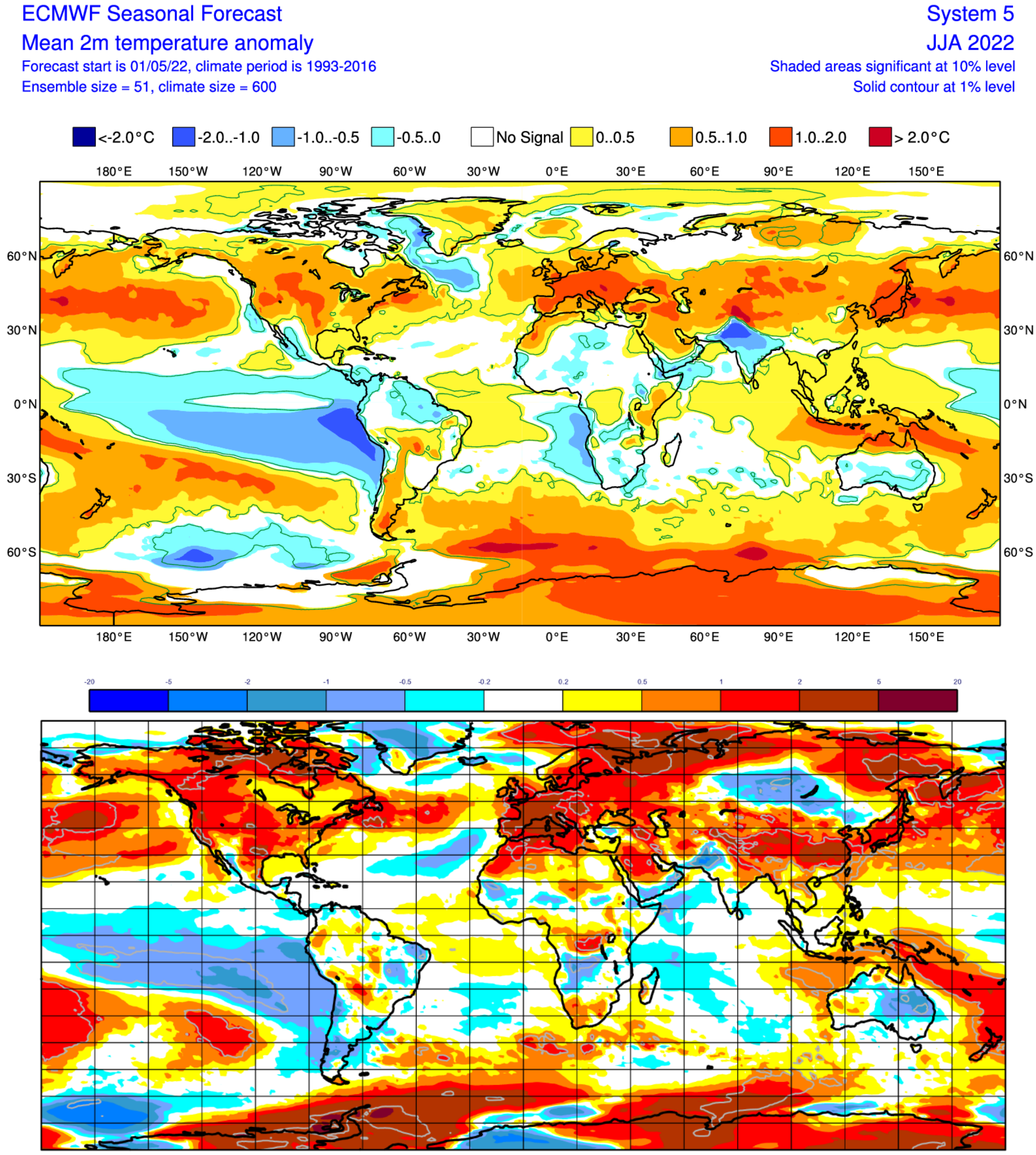


Figure 44: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2022 for JJA 2022 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

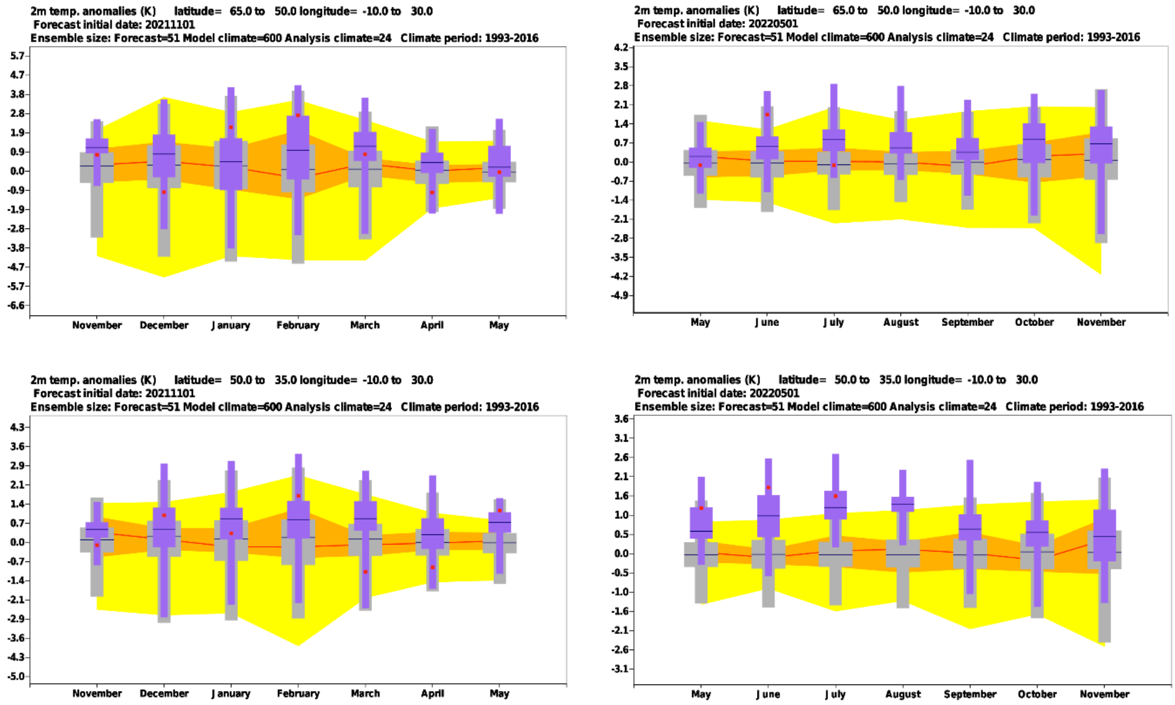


Figure 45: Long-range forecast of 2 m temperature anomalies from November 2021 for DJF 2021–22 (left panels) and from May 2022 for JJA 2022 (right panels) for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 hindcasts is shown 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 to isolate temperature variations over land.

A short note on scores used in this report

A. 1 Deterministic upper-air forecasts

The verifications used follow WMO CBS recommendations as closely as possible. Scores are computed from forecasts on a standard 1.5 × 1.5 grid (computed from spectral fields with T120 truncation) limited to standard domains (bounding co-ordinates are reproduced in the figure inner captions), as this is the resolution agreed in the updated WMO CBS recommendations approved by the 16th WMO Congress in 2011. When other centres' scores are produced, they have been provided as part of the WMO CBS exchange of scores among GDPS centres, unless stated otherwise – e.g. when verification scores are computed using radiosonde data (Figure 16), the sondes have been selected following an agreement reached by data monitoring centres and published in the WMO WWW Operational Newsletter.

Root mean square errors (RMSE) are the square root of the geographical average of the squared differences between the forecast field and the analysis valid for the same time. When models are compared, each model uses its own analysis for verification; RMSE for winds (Figure 16, Figure 18) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

Skill scores are computed as the reduction in RMSE achieved by the model with respect to persistence (forecast obtained by persisting the initial analysis over the forecast range); in mathematical terms:

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

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 32) the climate has been also derived from the ERA-Interim analyses.

A. 2 Probabilistic forecasts

Events for the verification of medium-range probabilistic forecasts are usually defined as anomalies with reference to a suitable climatology. For upper-air parameters, the climate is derived from ERA-Interim analyses for the 20-year period 1989–2008. Probabilistic skill is evaluated in this report using the continuous ranked probability skill score (CRPSS) and the area under relative operating characteristic (ROC) curve.

The continuous ranked probability score (CRPS), an integral measure of the quality of the forecast probability distribution, is computed as

$$CRPS = \int_{-\infty}^{\infty} \left[P_f(x) - P_a(x)\right]^2 dx$$

where P_f is forecast probability cumulative distribution function (CDF) and P_a is analysed value expressed as a CDF. CRPS is computed discretely following Hersbach, 2000. CRPSS is then computed as

$$CRPSS = 1 - \frac{CRPS}{CRPS_{clim}}$$

where $CRPS_{clim}$ is the CRPS of a climate forecast (based either on the ERA-Interim analysis or observed climatology). CRPSS is used to measure the long-term evolution of skill of the IFS ensemble (Figure 9) and its inter-annual variability (Figure 13).

ROC curves show how much signal can be gained from the ensemble forecast. Although a single valued forecast can be characterised by a unique false alarm (x-axis) and hit rate (y-axis), ensemble forecasts can be used to detect the signal in different ways, depending on whether the forecast user is more sensitive to the number of hits (the forecast will be issued, even if a relatively small number of members forecast the event) or of false alarms (one will then wait for a large proportion of members to forecast the event). The ROC curve simply shows the false alarm and hit rates associated with the different thresholds (proportion of members or probabilities) used, before the forecast is issued (Figure 37). Figure 37 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 38.

The comparison of spread and skill (Figure 10 to Figure 12) 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 21, Figure 22) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 21, Figure 22). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 23 to Figure 26), 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).

References

Ben Bouallegue, Z., T. Haiden, and D. S. Richardson, 2018: The diagonal score: definition, properties, and interpretations. Q. J. R. Met. Soc., 144, 1463-1473.

Ben Bouallegue, Z., T. Haiden, N. J. Weber, T. M. Hamill, and D. S. Richardson, 2020: Accounting for representativeness in the verification of ensemble precipitation forecasts. Mon. Wea. Rev., 148, 2049-2062.

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.

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

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.

Forbes, R., P. Laloyaux, and M. Rodwell, 2021: IFS upgrade improves moist physics and use of satellite observations. ECMWF Newsletter No. 169, 17-24.

Sleigh, M., P. Browne, C. Burrows, M. Leutbecher, T. Haiden, and D. Richardson, 2020: IFS upgrade greatly improves forecasts in the stratosphere. ECMWF Newsletter No. 164, 18-23.