Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

918

Evaluation of ECMWF forecasts

Thomas Haiden, Martin Janousek,
Frederic Vitart, Maliko Tanguy,
Fernando Prates, Matthieu Chevallier

September 2024

Series: ECMWF Technical Memoranda

A full list of ECMWF Publications can be found on our web site under:

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

Contact: library@ecmwf.int

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

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

Abstract

This report provides a summary of ECMWF's forecast performance, covering medium, sub-seasonal, and seasonal forecast ranges. Headline scores 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. An important recent change has been the increase in resolution of the ensemble forecast. With the implementation of model cycle 48r1 in June 2023 it now matches the high-resolution run, and this brings clear improvements in forecast skill. The primary focus of this summary is the medium range, specifically the forecast performance for upper-air variables. It is shown that in this respect ECMWF continues to have an overall lead among centres. For surface variables, other centres have partially taken the lead, especially in the short range, but significant improvements of ECMWF forecasts due to model cycle 48r1 can be seen there as well, such as a reduction of large 2-m temperature and 10-m wind speed errors in the ensemble forecast. In the sub-seasonal forecast range, the increase of the frequency of forecasts from bi-weekly to daily and the increase of the number of ensemble members from 50 to 100 has enhanced the effective value of the forecast. On the seasonal timescale, the change of Pacific SSTs from near-neutral to El Niño conditions in 2023 and return towards neutral conditions in 2024 was well predicted. For the first time this report also includes scores from machine-learning forecasts (AIFS), higher-resolution forecasts (DestinE continuous Extremes Digital Twin), and hydrological forecasts from the Copernicus Emergency Management Service (CEMS), in addition to atmospheric composition forecasts from the Copernicus Atmosphere Monitoring Service (CAMS).

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. ECMWF uses a set of headline scores to monitor the evolution of forecast skill over time, and this report gives the latest updates on these scores and on additional measures of forecast quality. An important recent change has been the increase in resolution of the ensemble forecast to match the high-resolution run, which brings clear improvements in forecast skill. This report mainly reports on the skill of the model in predicting the larger-scale flow of the atmosphere several days ahead. For this reason, a large part of the verification results deals with so-called 'upper-air' variables which define this flow. In this respect ECMWF has a clear lead among centres. For surface variables, some centres have overtaken ECMWF, especially at shorter ranges. However, improvements coming from the most recent model upgrade can be seen in a reduction of large 2-m temperature and 10-m wind speed errors in the ensemble forecast. For lead times of several weeks ahead, ECMWF provides now daily instead of bi-weekly forecasts, and 100 ensemble members instead of 50. This has enhanced the effective value of the forecast. In seasonal prediction, the change of Pacific SSTs from near-neutral to El Niño conditions in 2023 and return towards neutral conditions in 2024 was well predicted. For the first time this report also includes scores from machine-learning forecasts (AIFS), higher-resolution forecasts (DestinE continuous Extremes Digital Twin) and hydrological forecasts from the Copernicus Emergency Management Service (CEMS), in addition to atmospheric composition forecasts from the Copernicus Atmosphere Monitoring Service (CAMS).

1 Introduction

This report presents a summary of verification results from ECMWF's operational forecasting system including the medium-range, sub-seasonal and seasonal forecast, as well as the Copernicus Atmospheric Monitoring Service (CAMS) and, for the first time, the European Flood Awareness System (EFAS). To provide a reference for the medium-range scores, forecasts from ERA5 are evaluated as well. For the first time, scores from the higher resolution forecast (4.4 km) run with the global continuous Extremes Digital Twin (Extremes DT) developed in Destination Earth (DestinE) are included, as well as machine learning (ML) forecasts, particularly the AIFS, which has been developed in-house using the Anemoi framework which is being co-developed by ECMWF and Member States.

Verification results of ECMWF medium-range upper-air forecasts are presented in section 2, including some comparisons of ECMWF's forecast performance with that of other global forecasting centres. Section 3 presents the evaluation of ECMWF forecasts of weather parameters and ocean waves, while severe weather is addressed in section 4. Finally, section 5 discusses the performance of monthly and seasonal forecast products. In many of the plots, the beneficial effect of model cycle 48r1 is apparent. This cycle has been implemented in June 2023, shortly before the beginning of this year's reporting period. A comprehensive description of the changes introduced with 48r1 has been given last year in Tech Memo 911 (see also Lang et al., 2023).

Note that the abbreviation 'HRES' is still being used in this report for the deterministic run, because although ENS and HRES have been at the same resolution since the implementation of 48r1 in June 2023, the HRES and ENS control forecast (CTRL) are not yet bit-identical. From cycle 49r1 onwards, when HRES and CTRL are bit-identical, the name HRES will be replaced by CTRL. While the ENS scores are the most relevant, we present the deterministic ones first for historical reasons.

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 Technical Memoranda 346, 414, 432, 463, 501, 504, 547, 578, 606, 635, 654, 688, 710, 765, 792, 817, 831, 853, 880, 884, 902, 911). 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:

https://charts.ecmwf.int

by choosing 'Verification' and

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

2 Verification of upper-air medium-range forecasts

2.1 ECMWF scores

Figure 1 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%. For all three domains the 12-month averaged score has reached a new high point during the reporting period. The blue line shows that this has been the result of consistently high monthly scores rather than new record values in individual months. Note that in the northern extratropics, skill during JJA 2024 has been low, causing a drop in the 12-month average over the most recent period.

A complementary measure of performance is the root mean square (RMS) error of the forecast. Figure 2 shows RMS errors for both extratropical hemispheres of the six-day forecast, and the persistence forecast as a reference. In the northern hemisphere the 12-month running mean RMS error of the six-day forecast has reached its best (=lowest) value during the last year of the time series. In the southern hemisphere, values have been consistently low since 2020. A slight increase in 2022-23 appears to be driven by a decrease in persistence (seen as increased error of the persistence forecast).

Figure 3 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. Apart from inter-annual variability there has been no significant change in this metric since 2019.

The quality of ECMWF forecasts in the stratosphere in the northern hemisphere extratropics is shown through time series of temperature and vector wind scores at 50 hPa in Figure 4. There has been some further small reduction in error, especially for wind at day 5. Comparison with other centres in terms of 100 hPa temperature scores (Figure 5, top panel) shows that ECMWF is maintaining a substantial lead and managed to slightly reduce errors further in 2023. The centre and bottom panels in Figure 5 show HRES stratospheric temperature RMSE skill relative to ERA5 for a range of stratospheric levels. In the extratropics, results show consistently high values for 100 hPa but some drop at higher stratospheric levels, while in the tropics the highest values so far have been reached at levels of 30 and 50 hPa.

The trend in ENS performance is illustrated in Figure 6, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern extratropics. Both in Europe and the northern extratropics, the 12-month running mean of this score has reached its highest value so far, giving a clear signal of the beneficial effect of model cycle 48r1 which included an increase in ensemble resolution, among other changes.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 7. Forecasts show a very good overall match between spread and error, which has further improved for 500 hPa geopotential in 2024 compared to the two previous years.

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. Figures 8 and 9 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 well. In the medium-range ECMWF generally performs well, with its spread reliability closest to the diagonal. The stars in the plots mark the average values, corresponding to Figure 7, and ideally should lie on the diagonal and as close to the lower left corner as possible. In this regard ECMWF ranks among the best of the global models included here, with the exception of 850 hPa temperature at day 1, where the Japan Meteorological Agency (JMA) forecast exhibits the best spread reliability and lowest errors, and 500 hPa geopotential at day 1, where some of the other models have better spread reliability.

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 thus giving a clearer indication of ENS skill improvements. The dressing uses the mean error and standard deviation of the previous 30 days to generate a Gaussian distribution around ERA5. It represents a challenging benchmark since 50 ensemble members are compared against a continuous distribution. Figure 10 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 model cycle 48r1 in 2023 is clearly visible, leading to a forecast performance which is at its highest level so far. At forecast day 10, interannual variability is larger, making the signal of improvement from 48r1 less clear, however for 850 hPa vector wind and temperature the highest values so far have been reached.

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 11. Errors have been consistently low after the implementation of 47r3 in 2021.

2.2 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 Integrated Processing and Prediction System (WIPPS) designated centres, following agreed standards of verification.

Figure 12 shows time series of such scores for 500 hPa geopotential height verified against own analysis in the northern and southern hemisphere extratropics. In both hemispheres, ECMWF continues to maintain its lead. WMO-exchanged scores also include verification against radiosondes. Figure 13 (Europe), and Figure 14 (northern hemisphere extratropics) show 500 hPa geopotential height, 850 hPa temperature, 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 overall performance in the medium range when verified against observations. DWD tends to have the lowest errors at day 1 for both temperature and wind at 850 hPa.

The WMO model intercomparison for the tropics is summarised in Figure 15 (verification against analyses) and Figure 16 (verification against observations), which show vector wind errors for 250 hPa and 850 hPa. When forecasts are verified against each centre's own analysis, ECMWF does not generally have the lead, especially at shorter lead times (T+48). In the tropics, verification against analyses (Figure 15) is more sensitive to the details of the analysis method than in the extratropics. Smoother analyses (and forecasts) can help to reduce the RMSE. When verified against observations (Figure 16), different degrees of smoothness are less of an advantage, and the ECMWF forecast has the smallest overall error.

2.3 AIFS and other machine-learning forecasts

In 2023, ECMWF has started to run, in addition to the IFS, data-driven (machine-learning, ML) forecasts which use ERA5 analyses for training, and the HRES analysis as initial condition (Bouallegue et al. 2023, Lang et al. 2024). ECMWF's ML forecasting system AIFS is now routinely evaluated alongside other ML forecasts such as Google Deepmind's GraphCast and Huawei's Pangu as well as the IFS. Figure 17 shows that the upper-air skill of AIFS (red dashes) is very similar to that of GraphCast (olive), and that the IFS 'physics-based NWP' skill is comparable to Pangu's ML forecast skill (grey).

The skill of forecasts from AIFS increased substantially after two major upgrades of the system in winter 2024. The first increased the horizontal resolution to N320 (approximately 0.25°) and the second introduced total and convective precipitation. It is now similar in performance to GraphCast for some parameters and for others it is the leading Machine Learning (ML) model. In addition to smaller forecast errors, AIFS has been found to provide generally less jumpy forecasts. For example, in terms of the difference between consecutive forecasts validating at the same time, AIFS lies roughly halfway between the HRES and the ENS mean at day 5, and closer to the HRES at days 10-15. It also produces improved tropical cyclone track predictions when compared to the operational (physics-based) IFS forecast.

Figure 18 compares the drop of 500 hPa geopotential ACC with lead time between the IFS ensemble, IFS deterministic runs (HRES, CTRL), AIFS, and GraphCast. The two ML forecasts can be seen in between the smoother ENS mean, and the deterministic physics-based NWP forecasts. AIFS and GraphCast forecast skill are very similar. Further comparisons of ML vs physics-based NWP is given in the section about weather parameters below. Note that from model cycle 49r1 onwards (to be implemented November 2024), HRES and CTRL will become bit-identical.

2.4 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 Ensemble of Data Assimilation, has prognostic aerosols interacting with radiation, and only extends to day 5. Figure 17 shows that in terms of 500 hPa geopotential in the extratropics, the meteorological skill of CAMS forecasts is on par with those centres other than ECMWF that generally lead the ranking. In terms of atmospheric composition, routine evaluations confirm the consistently good performance of the CAMS forecast to predict Saharan Dust episodes, wild-fire plumes and air pollution events in Europe, China, and North-America as well as variations in Antarctic stratospheric ozone. For Europe specifically, AOD forecasts for several major Saharan dust episodes in spring 2024 gave a good indication of the timing and magnitude of enhanced dust concentrations. There is also a routine intercomparison of air quality forecast for North America of several global and regional models led by Environment Canada as part of WMO's Global Air Quality Forecasting and Information System (GAFIS) initiative. Their latest report (Presseau et al., 2024) shows that CAMS is competitive with other air quality forecasts in predicting NO_2, surface ozone, and PM2.5 (Figure 19). 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.

2.5 Hydrological forecasts

As part of its role as the Hydrological Forecast Computational centre of the Copernicus Emergency Management Service (CEMS), ECMWF has produced operationally river discharge simulations and forecasts over Europe, published in the European Flood Awareness System (EFAS) since 2012 (see EFAS wiki pages for detailed technical information regarding forcings and model setup).

Here we compare the skill of simulations forced with observations for EFAS versions 3, 4, and 5 which were implemented on 13 May 2019, 14 October 2020, and 20 September 2023, respectively. This is the first systematic comparison of the system's evolution and is based on a network of over 1,800 observational sites. Major changes include temporal resolution (from version 3 to 4), spatial resolution (from version 4 to 5) as well as an increasing number of stations used for calibration. The evaluation covers the period from 1993 to 2017, with some sites containing missing data. The modified Kling-Gupta Efficiency (KGE'; Kling et al., 2012) is used as the metric for this comparative analysis. KGE' combines correlation, bias ratio, and variability ratio, and is widely employed in hydrological assessments. The full KGE' formulation is provided in Annex A.4.

Figure 20 presents box plots comparing the range of KGE' and its components across the three model versions for all 1,807 stations with observed data. Overall, KGE' has improved in the newer versions, with the median increasing from 0.576 for EFAS3 to 0.663 for EFAS4, and to 0.688 for EFAS5. Additionally, the range of KGE' values has narrowed with fewer outliers in later versions. Upon examining the individual components of KGE', improvements are most evident in the correlation component. While the bias ratio and variability ratio show clear improvements from EFAS3 to EFAS4, the median slightly degrades from EFAS4 to EFAS5, though the range tightens with fewer outliers.

Figure 21 displays similar box plots for a subset of 778 stations with an upstream area smaller than 1500 km². For smaller catchments, the results are consistent with the general trends, but the improvements are more pronounced, with the median increasing from 0.460 for EFAS3 to 0.557 for EFAS4, and to 0.665 for EFAS5.

While this evaluation assesses the simulation ability of the hydrological model when forced with observed precipitation, we plan to include verification of EFAS operational forecasts in future issues of this report, to monitor the combined skill of both the meteorological and hydrological modelling.

3 Weather parameters and ocean waves

3.1 Weather parameters – high-resolution and ensemble

The supplementary headline scores for deterministic and probabilistic precipitation forecasts are shown in Figure 22. 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.

SEEPS deterministic precipitation forecast skill has been dropping since 2022. There was no model upgrade in 2022, and as shown by the ERA5 reference curve (black line in Figure 22, top left panel) this return to lower levels reflects interannual variability. The difference between HRES and ERA5 scores (upper right panel in Figure 22) shows little change in the last couple of years.

Probabilistic precipitation headline scores CRPSS and DSS are shown in the bottom panels in Figure 22. 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. In contrast to the deterministic skill, there is a clear positive signal in probabilistic scores due to the increase in spatial resolution of the ENS with model cycle 48r1.

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 at ECMWF. Results using these same headline scores for the last 12 months show the HRES leading with respect to the other centres from day 2 onwards (Figure 23). ECMWF's probabilistic precipitation forecasts are more skilful than those of other centres from day 3 onwards.

Trends in mean error (bias) and error standard deviation for 2 m temperature, 2 m dewpoint, total cloud cover, and 10 m wind speed forecasts over Europe are shown in Figures 24-27. 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.

Errors in 2 m temperature (Figure 24), have not changed significantly in recent years except for negative daytime biases (red bold curve), which have become somewhat smaller. For 2 m dewpoint (Figure 25), the negative bias during daytime in summer has persisted in recent years, and so has the level of error standard deviation. For total cloud cover (TCC, Figure 26) there has been an increase in error standard deviation, as well as bias. This is a consequence of the comprehensive moist physics upgrade in model cycle 47r3 which led to high cloud cover becoming much more binary (more often =100% whereas before it was often partial cloud cover) so for a deterministic forecast where cloud can be in the wrong place, this will typically lead to an increase in error (Forbes et al., 2021). While model cycle 48r1 did not address this specific aspect, there are plans to investigate whether TCC scores can be improved in upcoming model cycles. It is worth noting that the shortwave radiation has actually been improved by the change (see Figure 29, as discussed below) since there has been a compensation between changes in cloud cover and cloud optical depth. The error standard deviation of 10 m wind speed has been a bit smaller than in the previous year (Figure 27). The night-time positive wind speed bias is due to insufficient calming of the wind in the presence of surface inversions and is an issue that is being worked on.

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 28 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 (rather than own analysis), which implies a certain representativeness mismatch that adds to other sources of error. Overall, there is little improvement in the non-systematic errors of the deterministic forecast of surface parameters over the last few years, and a degradation specifically for TCC, consistent with results shown in Figures 24-27.

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 shortwave radiation fluxes. Figure 29 shows how the 5-day forecast of the TOA net shortwave radiation has improved over time. ERA5 is included for comparison, which shows that HRES errors have reached their lowest values so far relative to ERA5.

The fraction of large 2 m temperature errors in the ENS is one of ECMWF's headline scores. An ENS error is considered 'large' whenever the CRPS exceeds 5 K. Figure 30 shows that in the annual mean (red curve) this fraction has decreased from about 7% to 4% over the last 20 years. There are large seasonal variations, with values in winter about twice as high as in summer, however the amplitude of this annual variation is decreasing. A clear signal of improvement both in winter and summer, as well as the annual mean, can be seen in 2023 due to model cycle 48r1.

An analogous measure of the skill in predicting large 10 m wind speed errors in the ENS is shown in Figure 31. Here, a threshold of 4 m s^{-1} for the CRPS is used, to obtain similar fractions as for temperature. In recent years there has been a substantial reduction in the number of large 10 m wind speed errors, such that their fraction went down from a little less than 4% to 3.2%, which represents a relative improvement of about 20%. As for 2 m temperature, we see further improvements due to the model upgrade in 2023.

A comparison of the RMSE of 2 m temperature and 10 m wind speed at forecast day 5 from various ML forecasts and the HRES is shown in Figure 32. The strong improvement of the AIFS in winter 2023-24 has brought it into a clear lead for 2 m temperature compared to the other ML models, and on par with GraphCast for 10 m wind speed.

Since summer 2023, DestinE Extremes DT 4.4 km forecasts out to day 5 are being produced on a daily basis. Figure 33 shows time-series of RMSE of 2 m temperature and 10 m wind speed in Europe and the northern extratropics for DestinE Extremes DT and HRES. Overall, DestinE has smaller errors than HRES by a few percent, except for wintertime T2m in some areas, and total cloud cover, where both forecasts have very similar errors (Figure 34). Comparison of upper-air scores shows that errors in DestinE Extremes DT are very similar to those of HRES (not shown).

3.2 Ocean waves

The quality of the ocean wave model analysis and forecast is shown in the comparison with independent ocean buoy observations in Figure 35. After a slight increase of 10 m wind speed errors in the analysis and short-range forecast in recent years, they have gone down again and are close to the lowest levels seen in the timeseries. Wave height forecast scores have remained at the level seen in previous years. This is true also when verified against own analysis (Figure 36).

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. Both in the extratropics (Figure 37) and the tropics (Figure 38), ECMWF leads other centres in significant wave height and (after an upgrade in the computation of peak period in model cycle 47r3) also in peak period.

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

https://charts.ecmwf.int

by choosing 'Verification' and 'Ocean waves' (under 'Parameters').

Verification results from the WMO Lead Centre for Wave Forecast Verification, which are updated at 3-monthly intervals, can be found at

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

4 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 (ROC) area (Section 4.1)
- The tropical cyclone position error for the high-resolution forecast (Section 4.2)

4.1 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 is performed using synoptic observations 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 darker green lines in Figure 39 (top), together with results for days 1–3 and days 5-7. It has reached its highest value so far both in the northern extratropics and in Europe. Similarly, 24-hour total precipitation scores (centre) have reached some of their highest values in 2023. For 2 m temperature, the 12-month average skill has dropped in recent years compared to the highest values so far which occurred in 2021-22. Note however that the absolute level of skill for 2 m temperature is higher than for the other two parameters, in particular at longer lead times.

4.2 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 40. Errors in the forecast central pressure of tropical cyclones are also shown.

In terms of absolute skill, the HRES position error (top panel, Figure 40) has been larger than in the previous year, however the ERA5 curve shows that this is due to interannual variability and that at day 3 the HRES has had the smallest errors compared to those of ERA5. Intensity errors of the ENS control have improved significantly and are now similar to those of the HRES because of the ENS spatial resolution upgrade in cycle 48r1. The bottom panel of Figure 40 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. While errors have been larger, the match between error and spread has been particularly good in the most recent year.

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 41. 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. In terms of all three metrics, the most recent year shows the highest skill, demonstrating the beneficial effect of the ENS resolution upgrade, among other changes, in model cycle 48r1.

Figure 42 shows how AIFS and HRES tropical cyclone position and central pressure errors compare. The top panel shows a large (25%) reduction of the mean absolute position error in AIFS compared to HRES. For central pressure, both the mean error (bias) and mean absolute error are smaller in HRES, partly due to the lower resolution of the AIFS and its training data.

5 Sub-seasonal and seasonal forecasts

5.1 Sub-seasonal forecast verification statistics and performance

Figure 43 shows the probabilistic performance of the sub-seasonal 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 defined here as the persistence of the week 1 forecast into week 2, and persistence of the week 2 forecast into weeks 3+4.

After the unusually high skill in summer 2022, values for 2023 have gone down to the more typical high values seen in recent years. For the winter 2023-34, week 2 skill has been high, while the skill of persistence has been at the low end, thus the skill above persistence was the highest seen so far. This is even more pronounced at weeks 3+4, where persistence had very low skill, creating a large skill-above-persistence value for the forecast. The low persistence appears to be related to the absence of prolonged blocking periods during winter 2023-24 in the northern hemisphere.

Because of the low signal-to-noise ratio of real-time forecast verification in the sub-seasonal lead time range (Figure 43), re-forecasts are a useful additional resource for documenting trends in skill. Figure 44 shows the skill of the IFS 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. There has been no significant change since 2021. A slight drop in recent years is more visible for verification against observations than verification against analysis. Note that there have been important changes in the sub-seasonal forecast with model cycle 48r1 such as an increase in the number of ensemble members from 50 to 100, and an increase in the frequency of runs from bi-weekly to daily. These changes have increased the benefit of the forecast for users, however they are not shown by the headline score based on reforecast verification. The improvement due to the increased number of ensemble members can be seen in Figure 45 which was included in last year's report (ECMWF Tech Memo 911) but is reproduced here for convenience. It shows beneficial effects across parameters and forecast ranges (weeks 1-4).

More verification results for the sub-seasonal forecasts are available on the ECMWF website at https://charts.ecmwf.int by choosing 'Verification' and 'Extended' under 'Range'.

5.2 Seasonal forecast performance

5.2.1 Seasonal forecast performance for the global domain

The current version SEAS5 of the seasonal component of the forecasting system is based on IFS cycle 43r1. While the ocean model and initial conditions are the same as used in the sub-seasonal forecast including resolution (TCo319), SEAS5 has 91 levels, whereas the sub-seasonal forecast has 137. There are also some minor differences in the model physics, mainly in the treatment of aerosols. 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://charts.ecmwf.int

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

5.2.2 The 2022-23 El Niño forecasts

The year 2023 was characterized by a transition from near-neutral to El Niño conditions which peaked in December. SEAS5 forecasts gave a very good signal for this transition, for the timing of the maximum, as well as the return to more neutral conditions in 2024 (Figure 46, left column). It did however overestimate the magnitude of the warm anomalies. The C3S multi-model ensemble (Figure 46, right column), with its naturally larger spread, gave a better estimate of the magnitude. The large spread of the C3S ensemble is mostly due to different biases of contributing models. The tendency towards the cold side in some of these models balances the tendency of SEAS5 towards the warm side.

5.2.3 Tropical storm predictions from the seasonal forecasts

The 2023 Atlantic hurricane season had a total of 15 named storms, including 7 hurricanes and 3 major hurricanes. The accumulated cyclone energy index (ACE) was about 20% above the 10-year (2013-2022) climate average (Figure 47) which makes it an active season in terms of the number of tropical storms (climate average is about 13) and ACE. Seasonal tropical storm predictions from SEAS5 initialized on 1st May 2023 correctly indicated a higher number of tropical storms (17 +/- 4) over the Atlantic and larger ACE (about 40% above average) than climatology. Subsequent forecasts, issued in July and August, correctly predicted an above average intensity of the tropical cyclone season, although overall SEAS5 overestimated the intensity of the 2023 hurricane season.

Figure 48 shows that SEAS5 predicted average activity over the eastern North Pacific and slightly below average over the western North Pacific (ACE of about 90% of the 2013-2022 climate average). The 2023 western Pacific typhoon season was the third consecutive below-average typhoon season and the third most inactive season on record in terms of terms of named storms, with only 17 storms developing including 10 typhoons and an ACE close to climatology. This is consistent with the SEAS5 forecast, although SEAS5 predicted a larger number of tropical cyclones than observed. The 2023 eastern North Pacific hurricane season was an above average active season with an ACE 20% larger than climatology, while SEAS5 predicted average activity.

For 2024, SEAS5 initialized on 1st May predicted a largely above-normal season over the Atlantic (ACE twice as large as climatology), consistent with La-Nina and warmer tropical Atlantic SSTs. The subsequent forecasts issued in June, July and August continued predicting an active Atlantic season, but with increasingly lower magnitude. SEAS5 predicted a below average tropical cyclone activity over the western North Pacific and average activity over the eastern North Pacific.

5.2.4 Extratropical seasonal forecasts

The seasonal forecast of temperature anomalies for DJF 2023/24 was very good in terms of large-scale patterns in the southern hemisphere, especially over the ocean (Figure 49). This is not surprising given the strong forcing provided by El Niño. In the northern hemisphere, the extreme warm anomalies over the continental areas were correctly predicted, however the very distinct and rather persistent cold anomaly in Scandinavia was missed. A more detailed analysis of this feature has shown that it was not predicted even on the sub-seasonal timescale beyond week 2. A similar cold anomaly in the easternmost parts of Siberia was at least slightly indicated.

Predicted summer 2m temperatures (Figure 50) indicated widespread positive anomalies over the northern hemisphere, in accordance with observations. A large area of positive anomalies covering Africa and central/eastern Europe was captured. Also, some of the smaller areas with negative-to-neutral anomalies were indicated. The main difference between forecasts and analyses in the northern hemisphere was a negative anomaly in the eastern North Atlantic that was not present in the forecast, as well as the warm/cold patterns in the Arctic, e.g. along the Siberian Arctic coast.

Since the ensemble mean carries only part of the information provided by the ENS, we also look at the forecast distribution in the form of quantile (climagram) plots. Climagrams for Northern and Southern Europe for winter 2023-24 and summer 2024 are shown in Figure 51. Red squares indicate observed monthly anomalies. As in previous years, both in winter and summer, warm anomalies are generally better predicted than cold ones. The persistent cold anomaly in Northern Europe in winter was missed, and investigations showed that predictability of this anomaly was low even on the sub-seasonal timescale. In contrast, the forecast for summer in Northern Europe was rather good, including a qualitatively correct prediction of the intra-seasonal variation of warm anomalies. In Southern Europe, observed anomalies were positive throughout, which was predicted. Some of the verifying positive anomalies fell even outside the range spanned by the ENS.

ECMWF's Copernicus Climate Change Service (C3S) provides verification maps of the seasonal forecast for individual models and the multi-model ensemble at

https://confluence.ecmwf.int/display/CKB/C3S+seasonal+forecasts+verification+plots

Examples of these plots, which are based on reforecasts for the period 1993-2016 are given in Figures 52 and 53 for the ROC area and linear correlation, respectively. Note that these reforecasts are created using the same forecast model, initial conditions and ensemble generation as the real-time forecasts, thus representing as close as possible a parallel of having created forecasts with the same methodology, in the past.

The verification plots show generally high correlation and ROC area in the tropics and indicate Europe as an area of mostly low or non-significant skill. However, the multi-model forecast adds some skill over the Eurasian continent. Also, in the summer period, where ECMWF has very low skill in the Arctic, it provides some skill. In the tropics, the overall pattern of ROC area and correlation is very similar between ECMWF's forecast and the multi-model forecast.

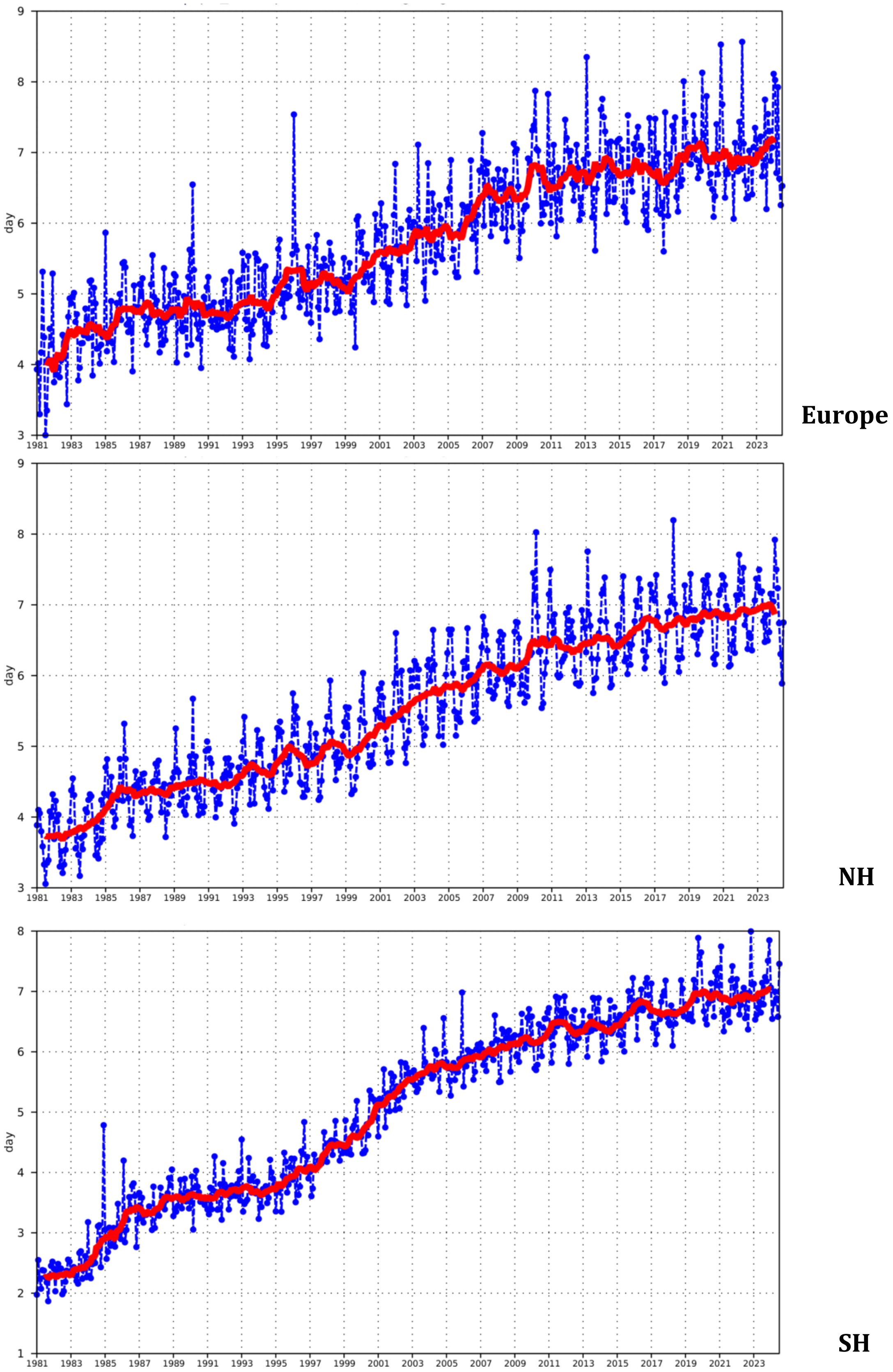


Figure 1: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).

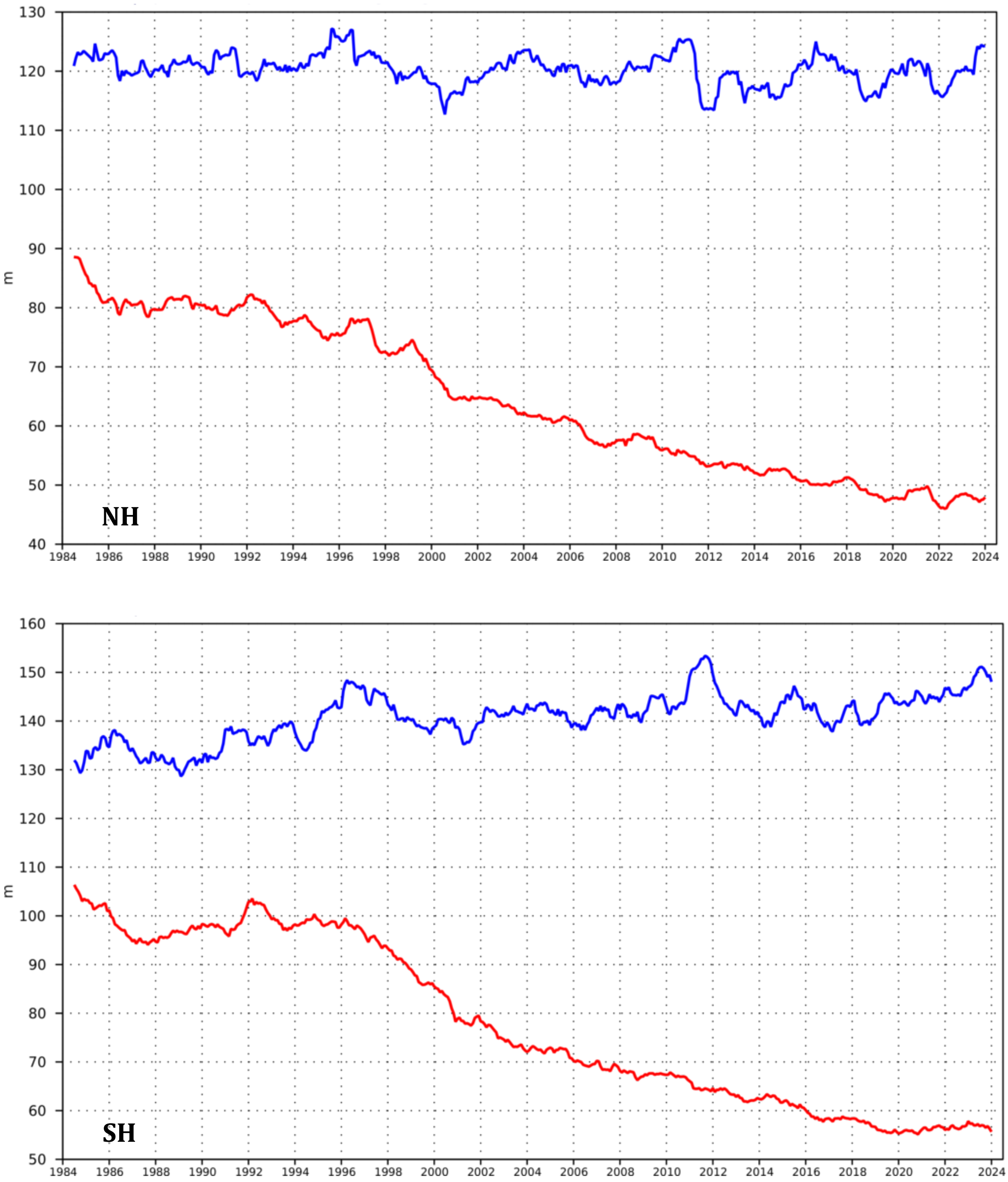


Figure 2: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2023–July 2024. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

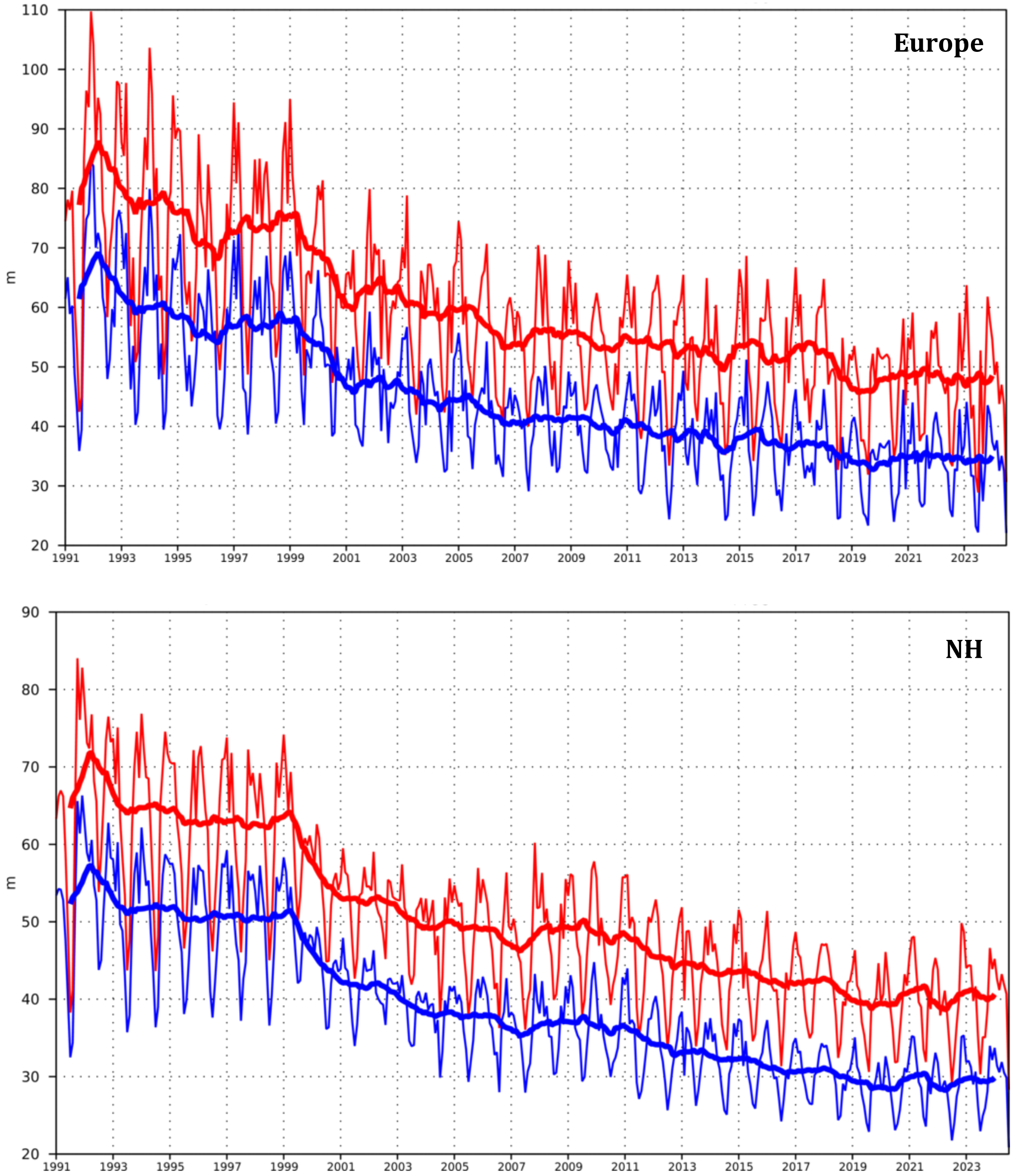


Figure 3**:** A measure of inconsistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores are also shown (in bold).

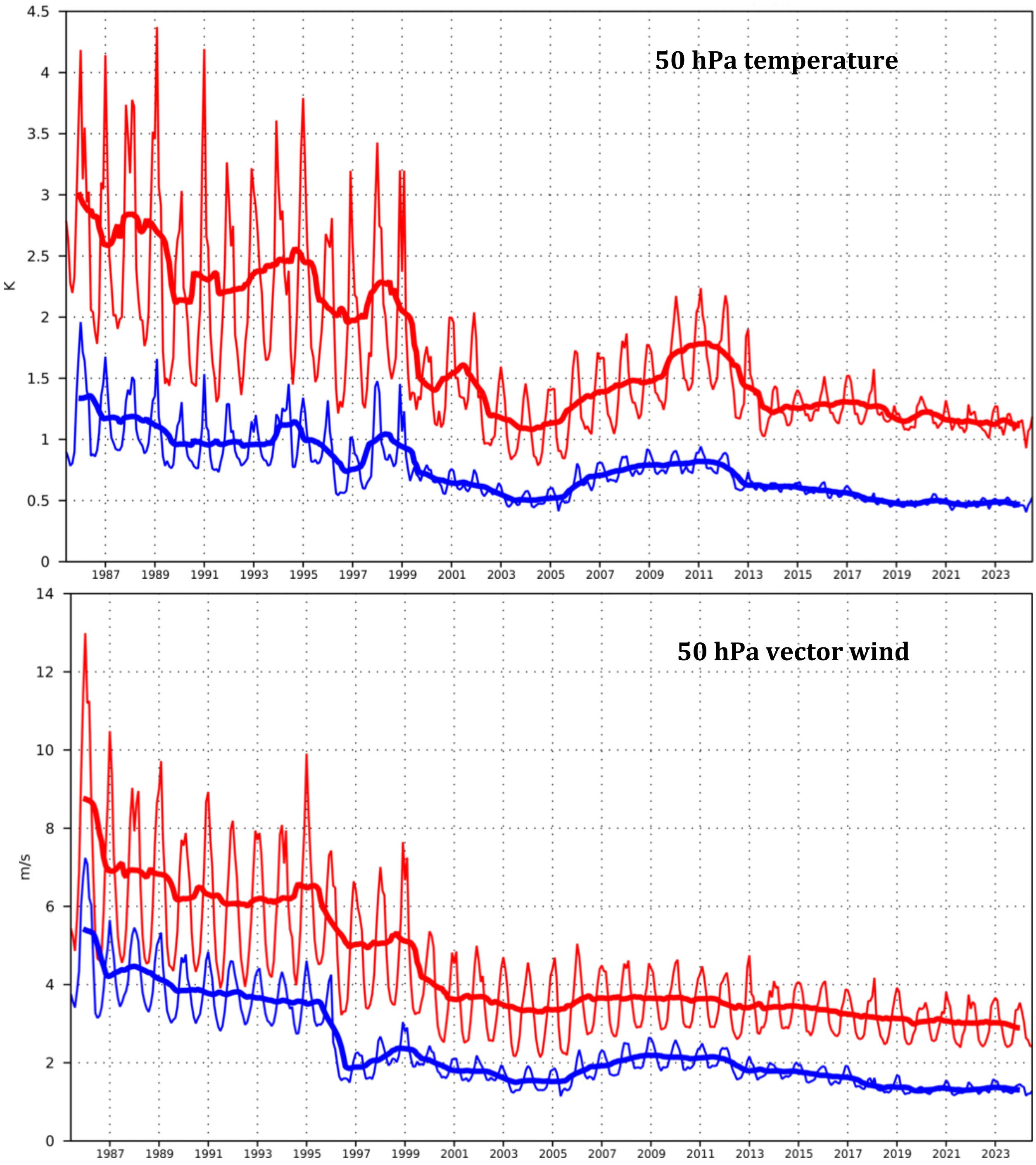


Figure 4**:** Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

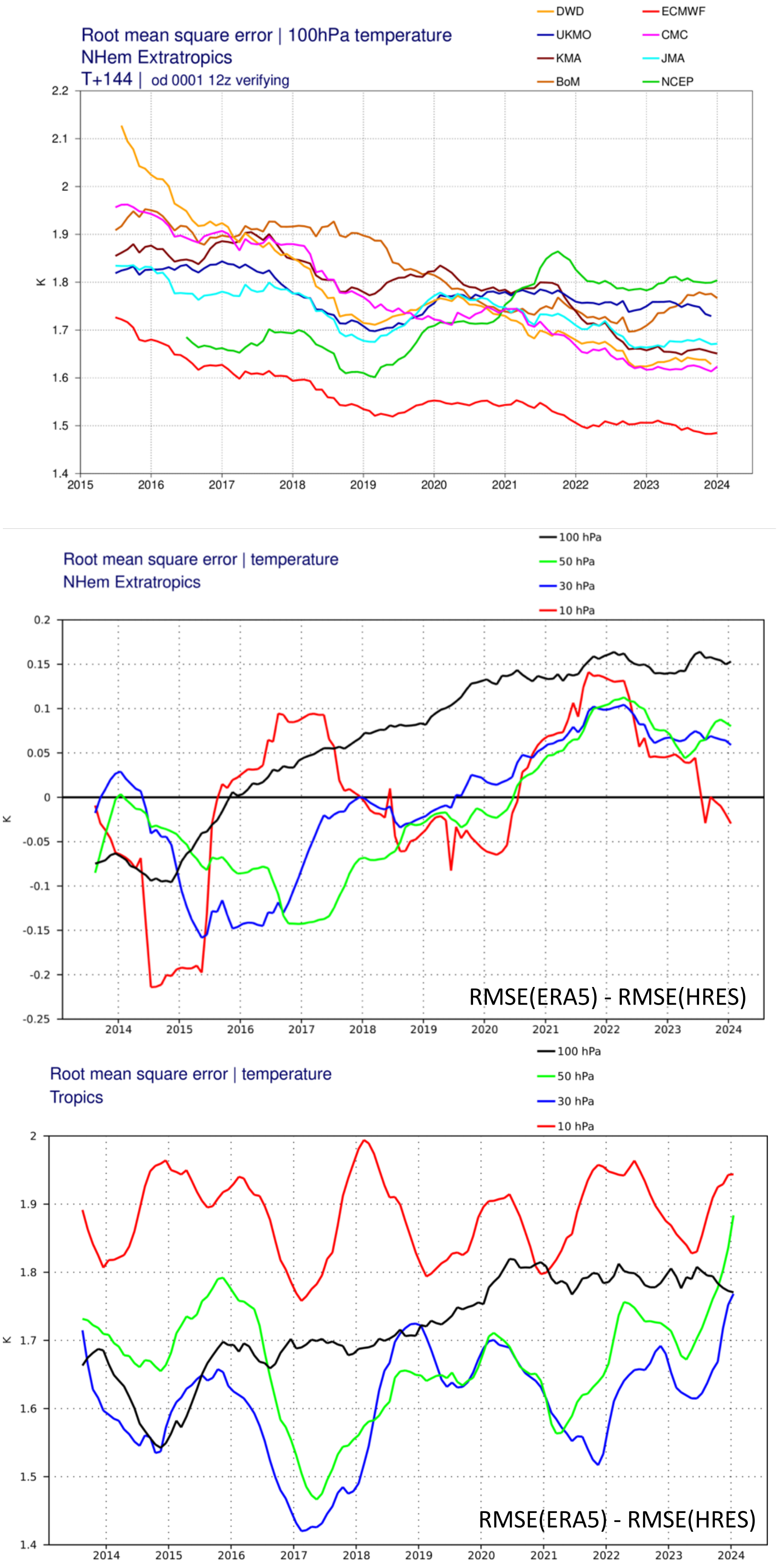


Figure 5: Stratospheric scores at a lead time of +144 h. Top: global model intercomparison of the 100 hPa temperature RMSE in the northern extratropics based on the WMO exchange of scores. Centre: difference in RMSE of temperature between ERA5 and HRES at four different stratospheric levels in the northern

extratropics. Bottom: same as centre, but for the tropics. Curves in all three plots are 12-month running averages.

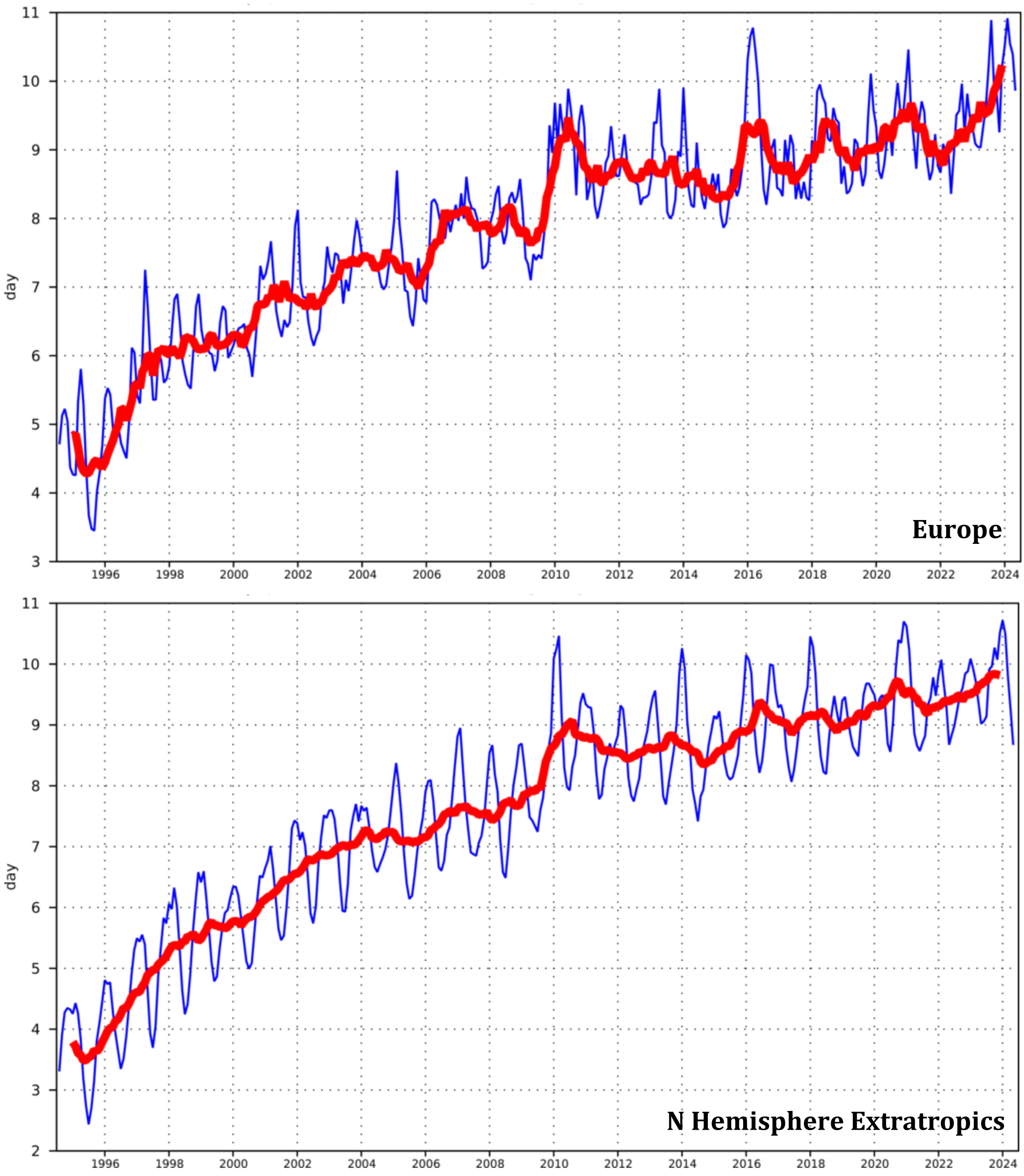


Figure 6**:** Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month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

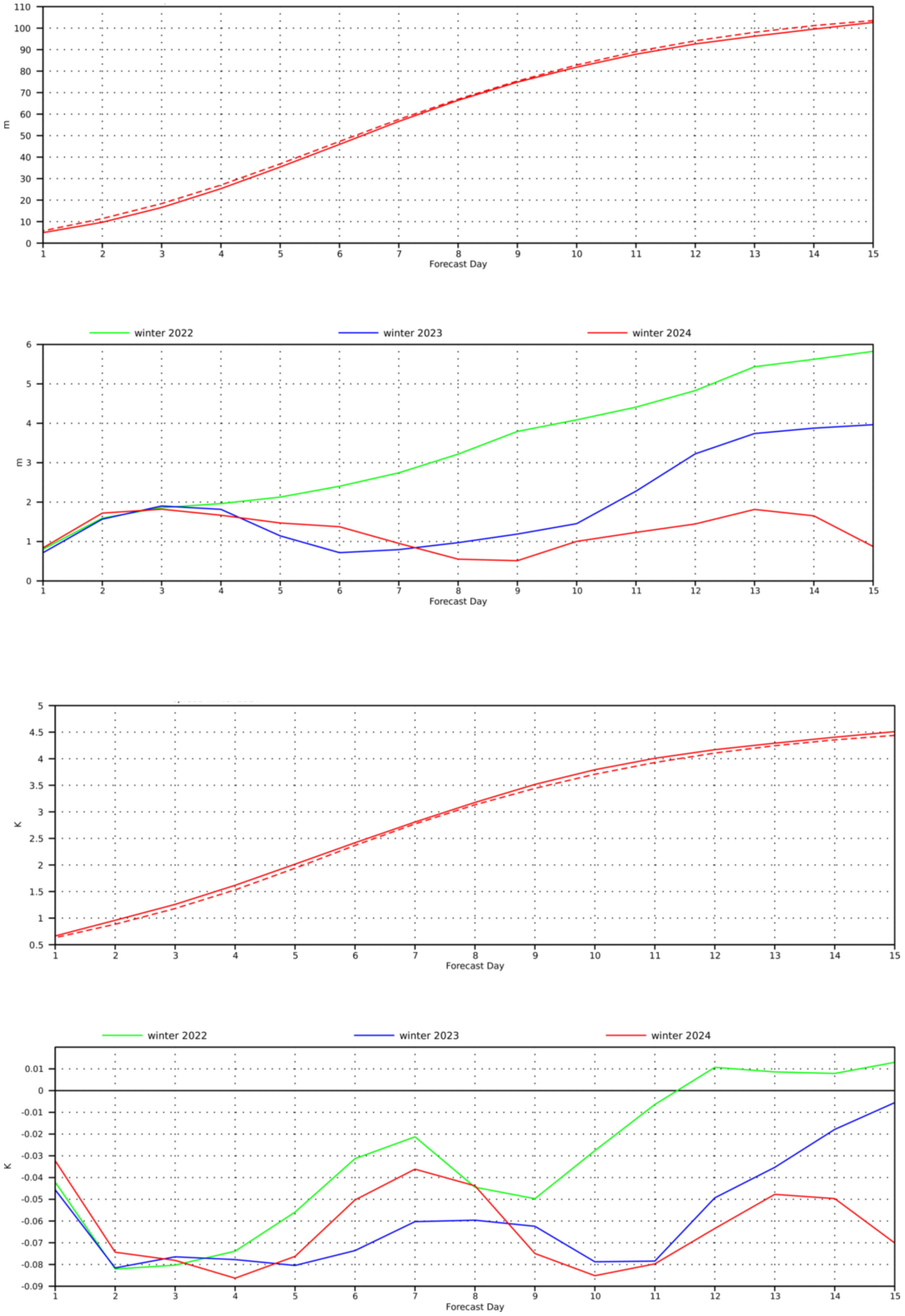


Figure 7**:** Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2023–2024 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values indicate spread is too small); verification is against analysis, plots are for 500 hPa geopotential (top) and 850 hPa temperature (bottom) over the extratropical northern hemisphere for forecast days 1 to 15.

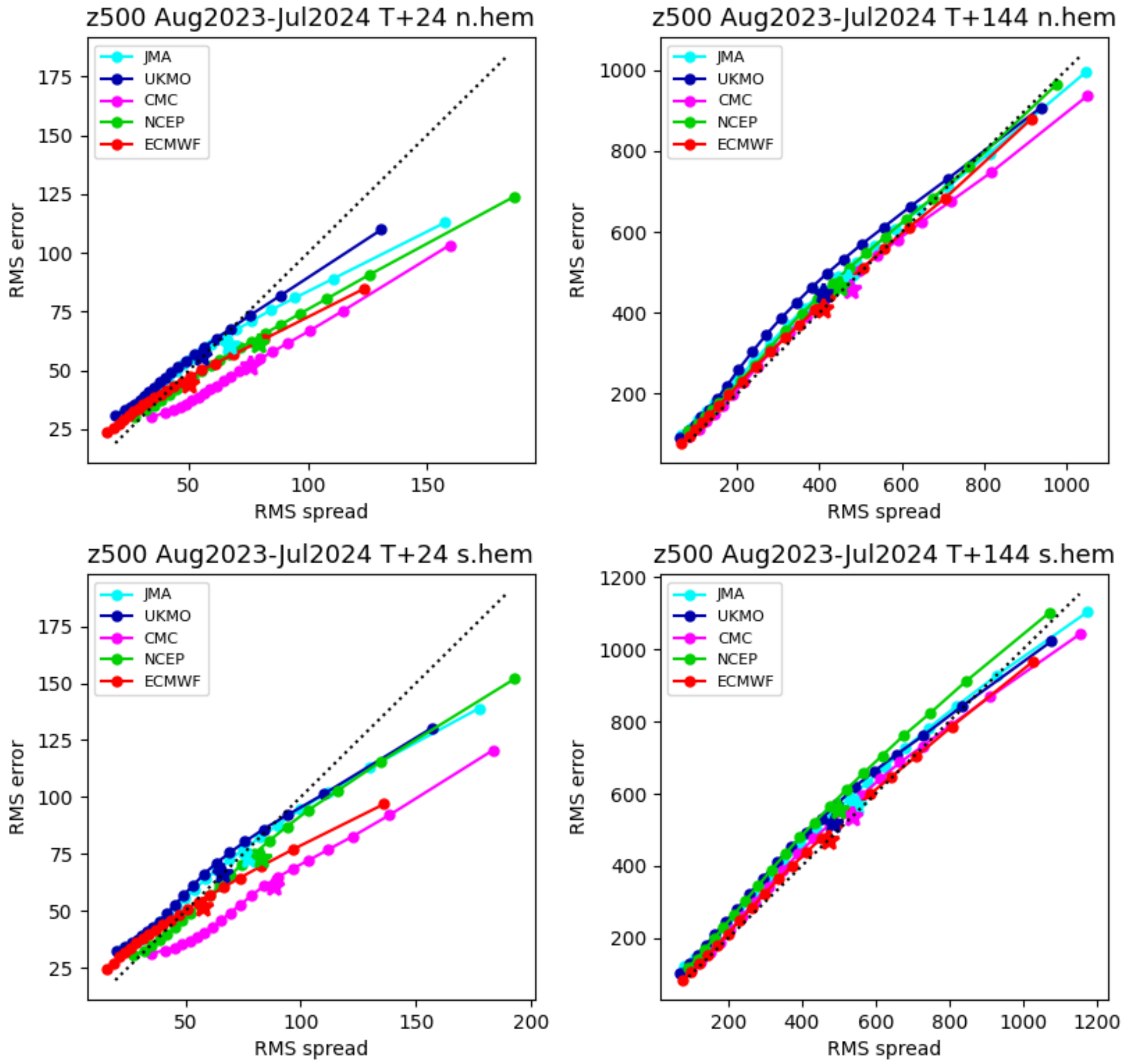


Figure 8: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2023–July 2024 in the northern (top) and southern (bottom) hemisphere extra-tropics for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

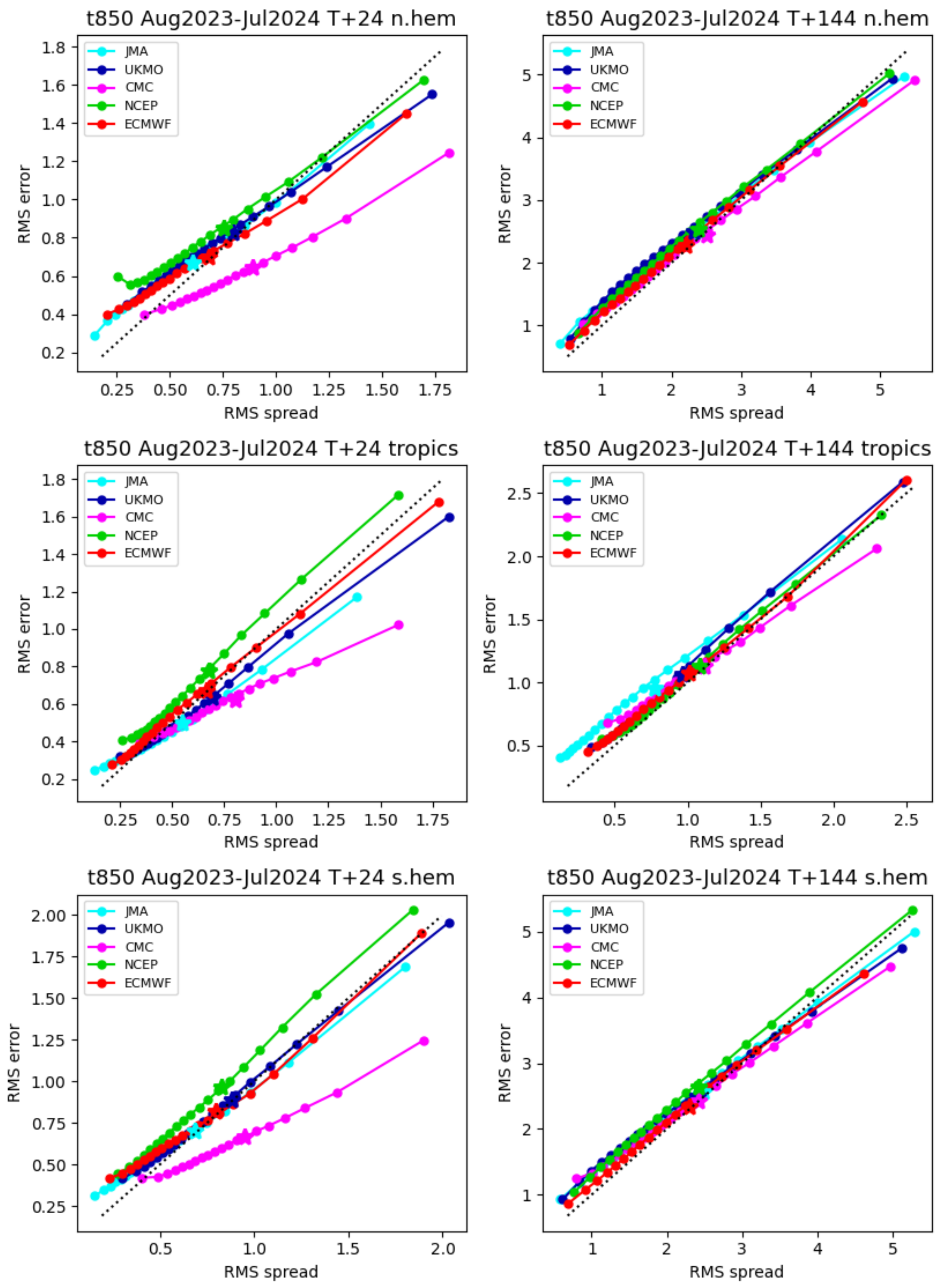


Figure 9**:** As Figure 8 but for 850 hPa temperature, and including the tropics.

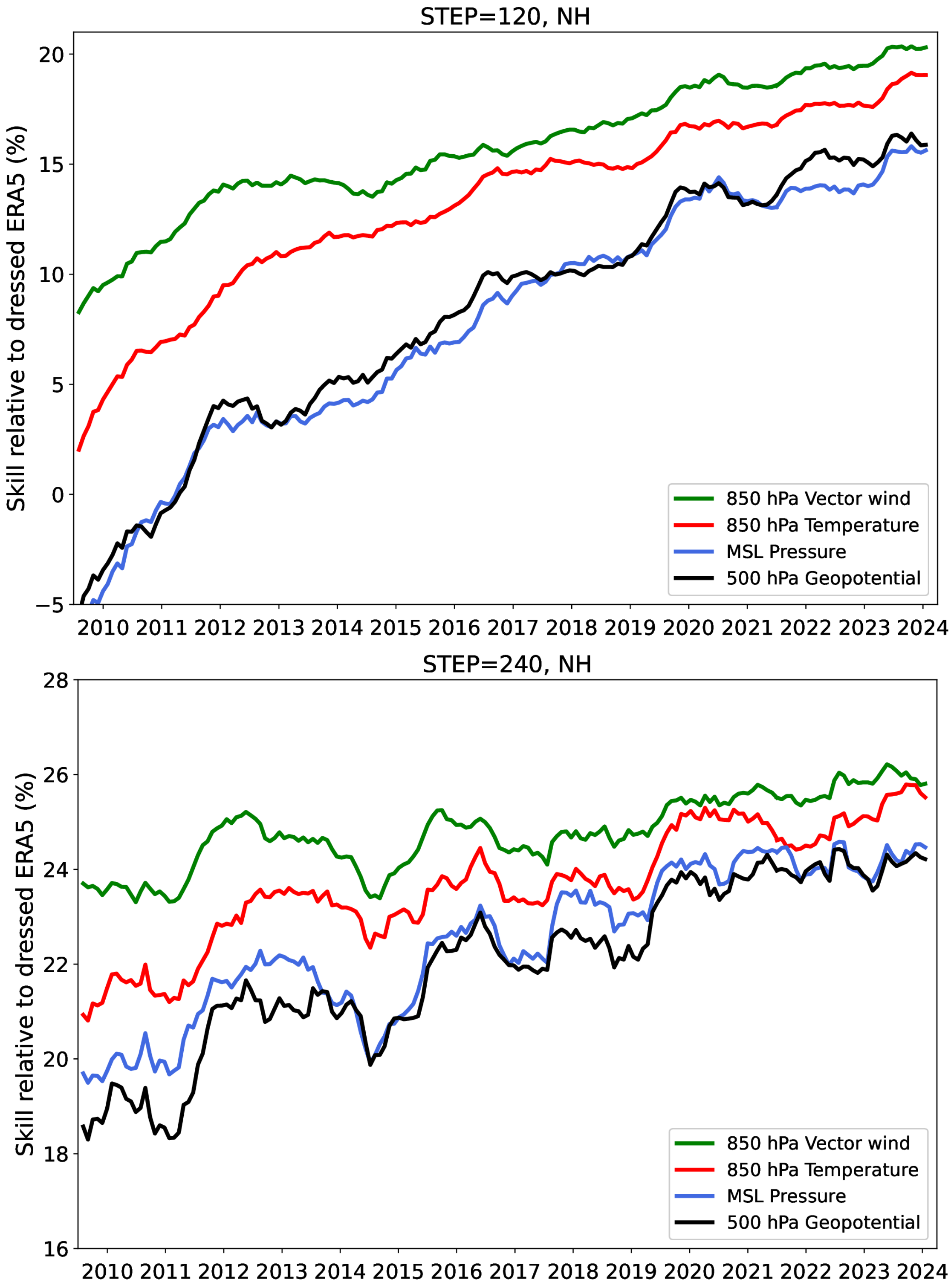


Figure 10**:** Skill of the ENS at day 5 (top) and day 10 (bottom) for upper-air parameters in the northern extratropics, relative to a Gaussian-dressed ERA5 forecast. Values are running 12-month averages, and verification is performed against own analysis.

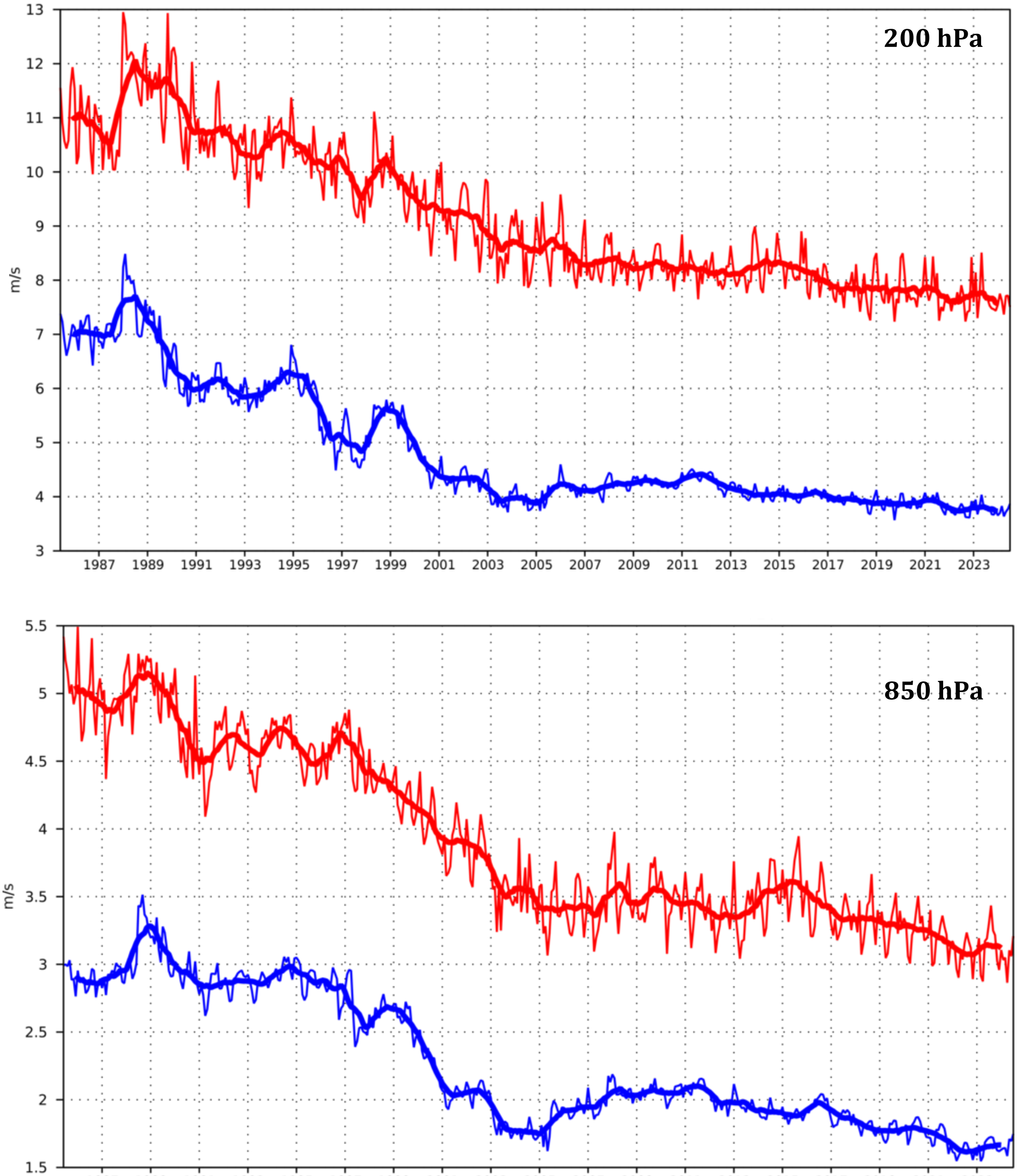


Figure 11**:** Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

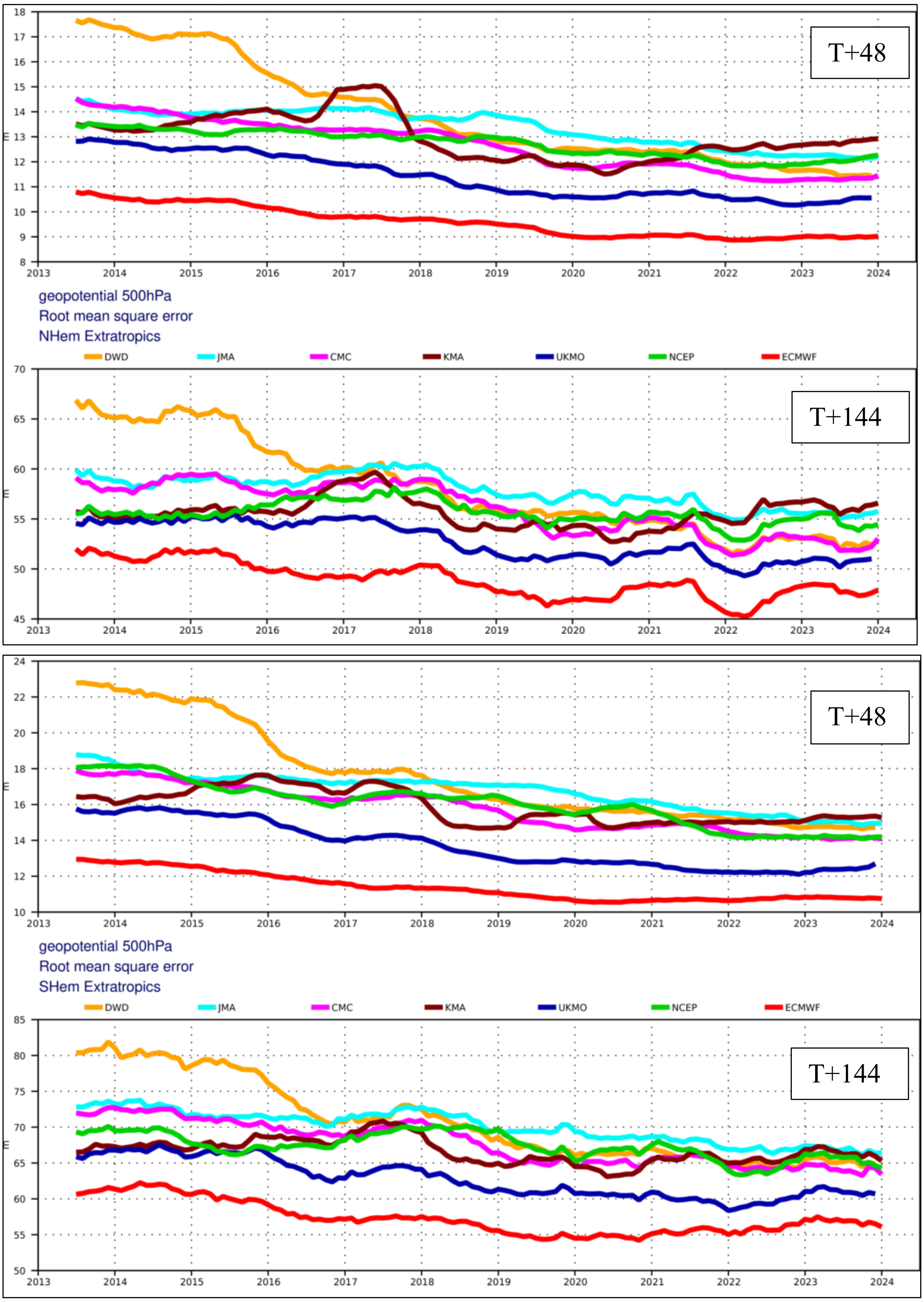


Figure 12**:** WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top box) and southern (bottom box) extratropics. In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated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

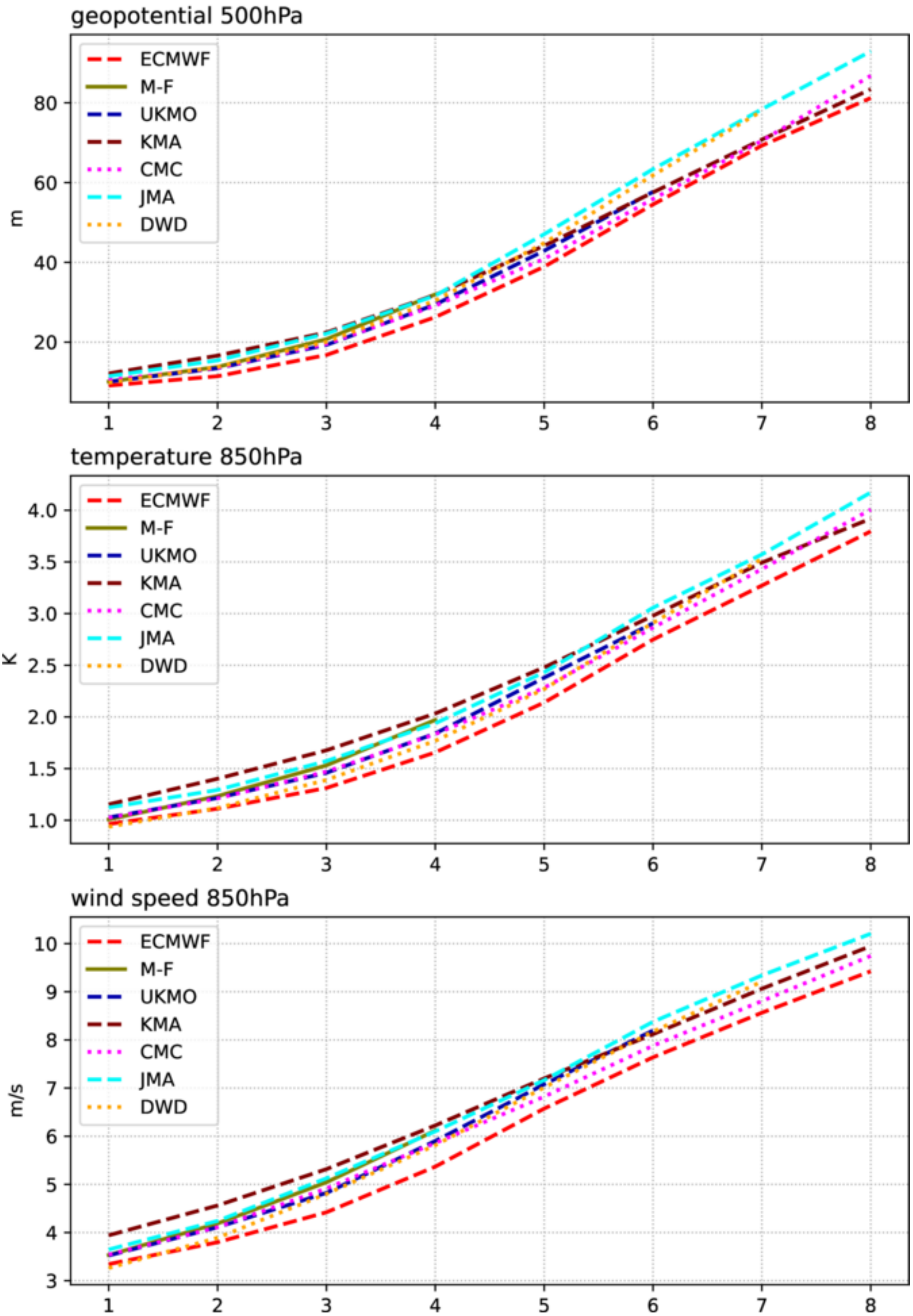


Figure 13**:** WMO-exchanged scores for verification against radiosondes: 500 hPa height (top), 850 hPa temperature (middle), and 850 hPa wind (bottom) RMS error over Europe and North Africa (annual mean August 2022–July 2023) of forecast runs initialized at 12 UTC. M-F = Météo-France, JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, KMA = Korea Meteorological Administration, DWD = Deutscher Wetterdienst.

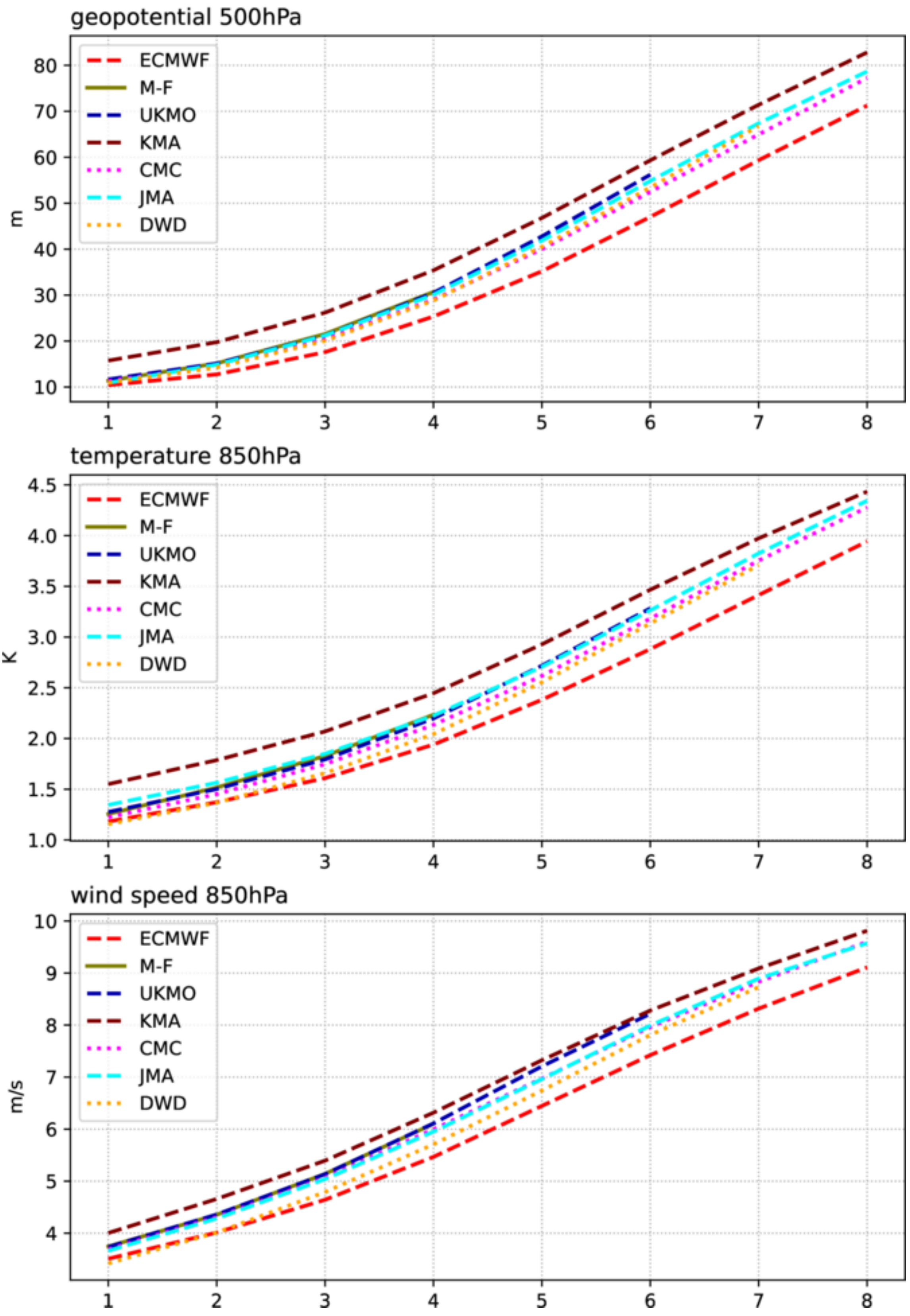


Figure 14: As Figure 13 for the northern hemisphere extratropics.

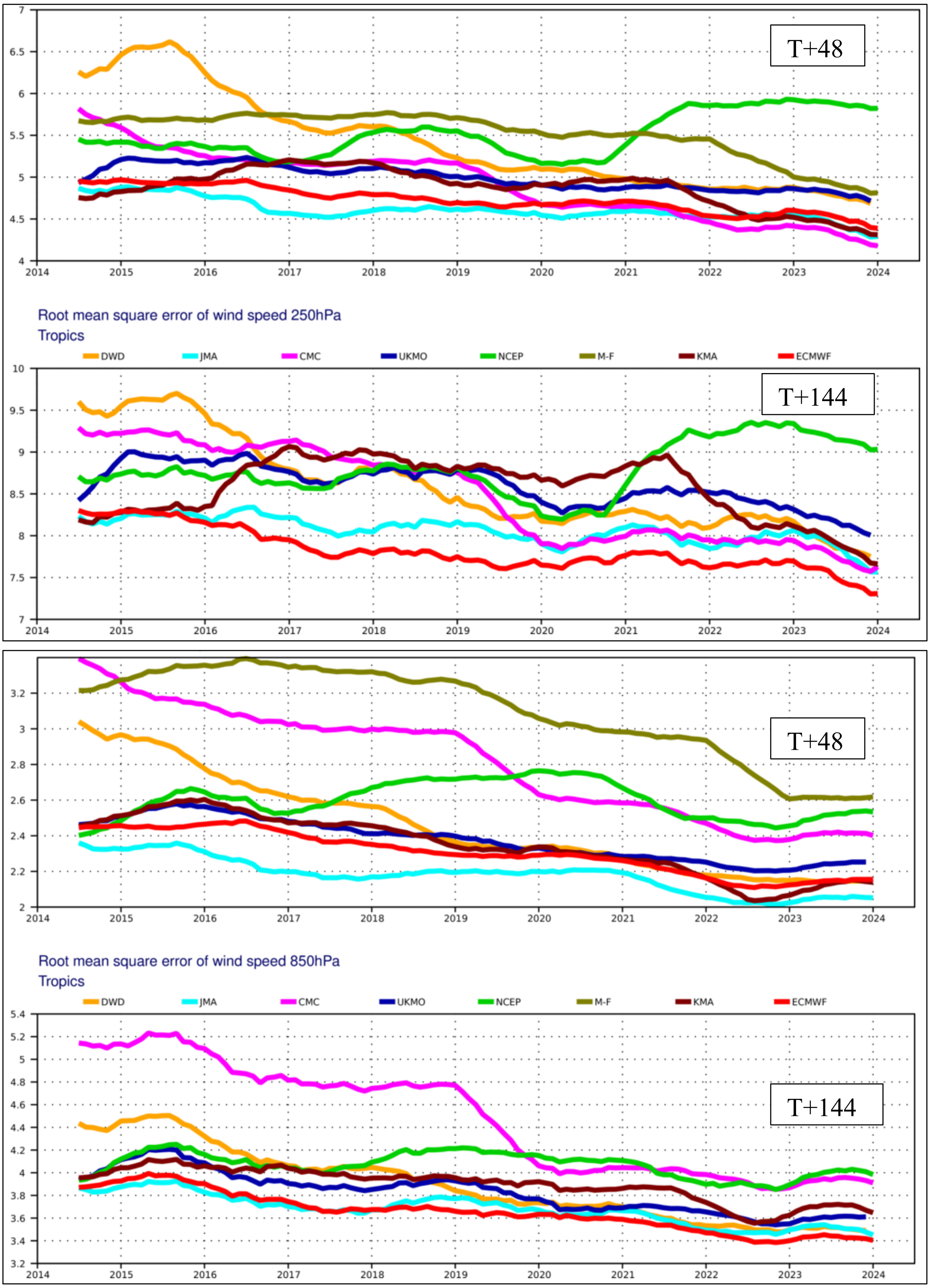


Figure 15**:** WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top box) and 850 hPa (bottom box). In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis.

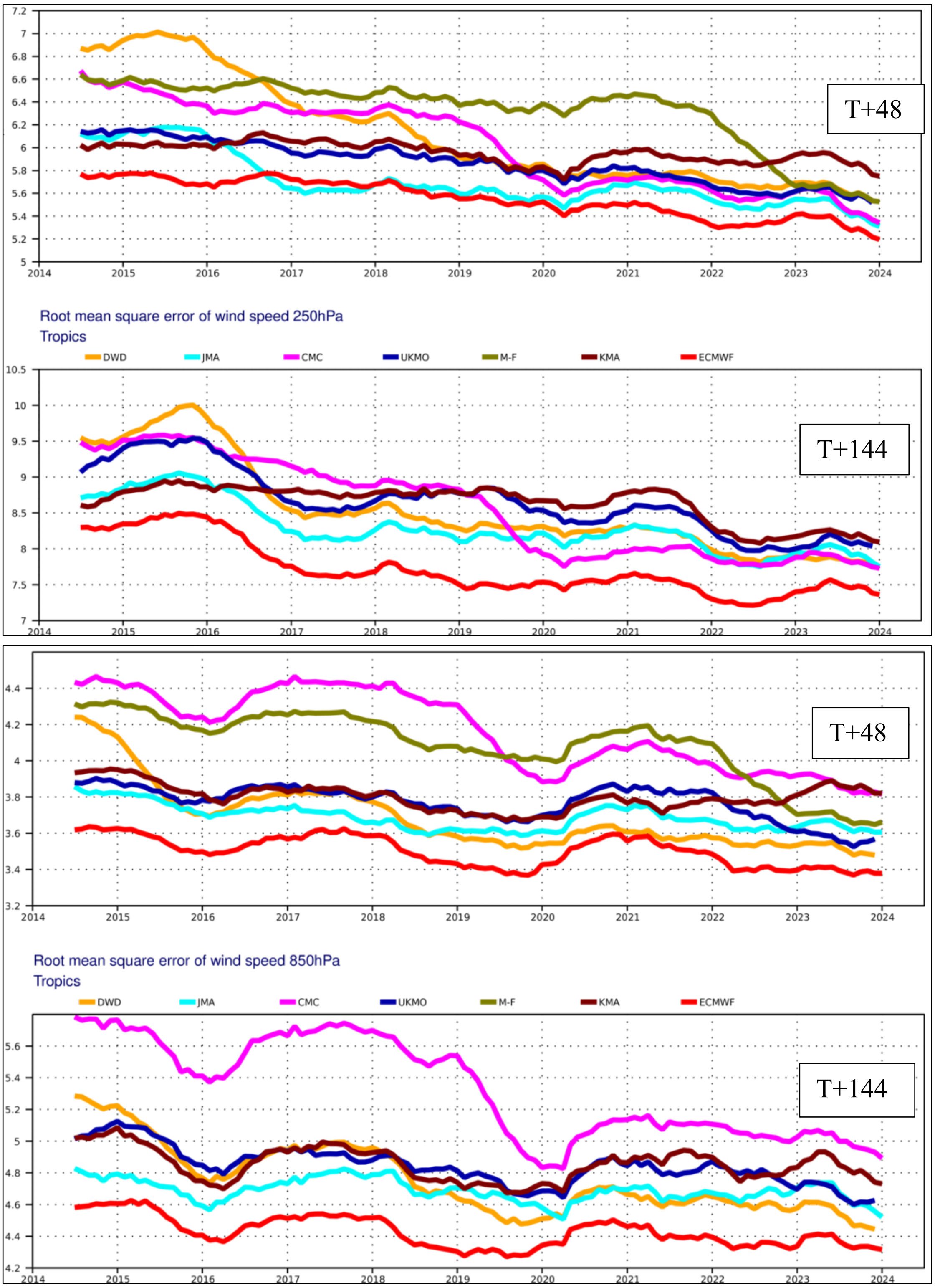


Figure 16**:** As Figure 15 but for verification against radiosonde observations.

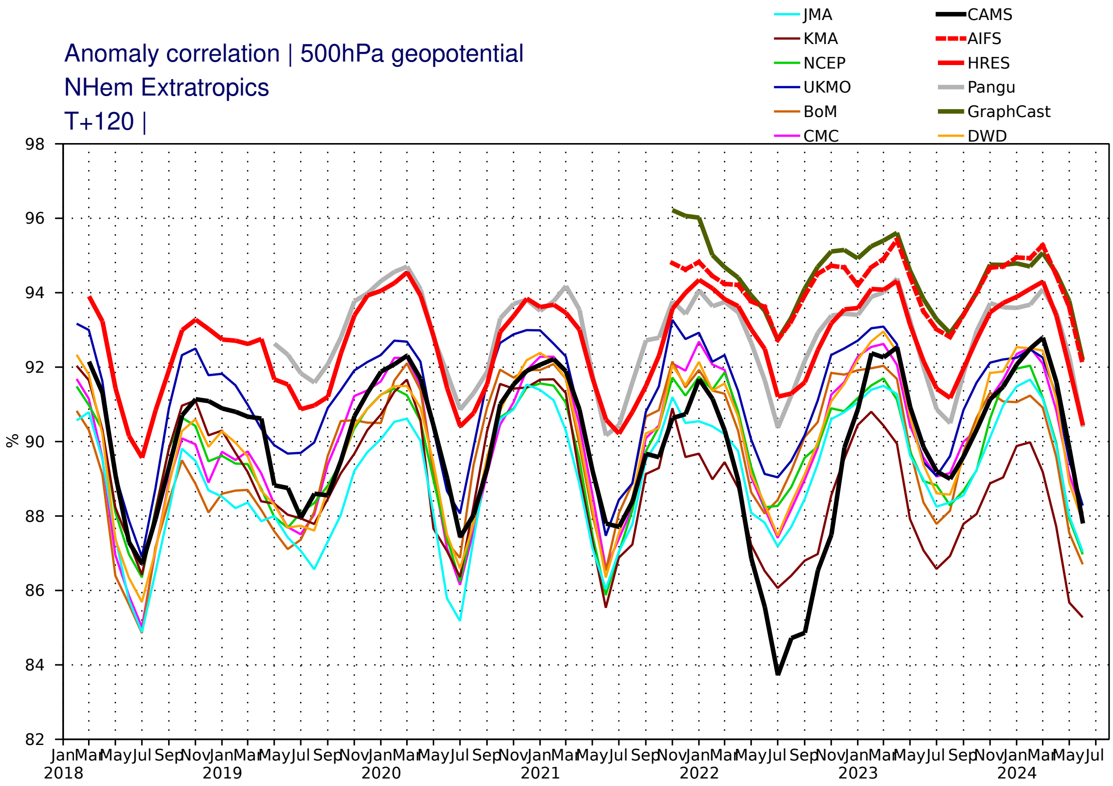


Figure 17: Anomaly correlation of 500 hPa geopotential in the northern hemisphere extratropics at day 5. CAMS forecast (black) shown in comparison to the HRES (red) and forecasts from other global centres (thin lines). Also shown are forecasts from machine learning (ML) models: GraphCast (olive), Pangu (grey), and AIFS (red dashes).

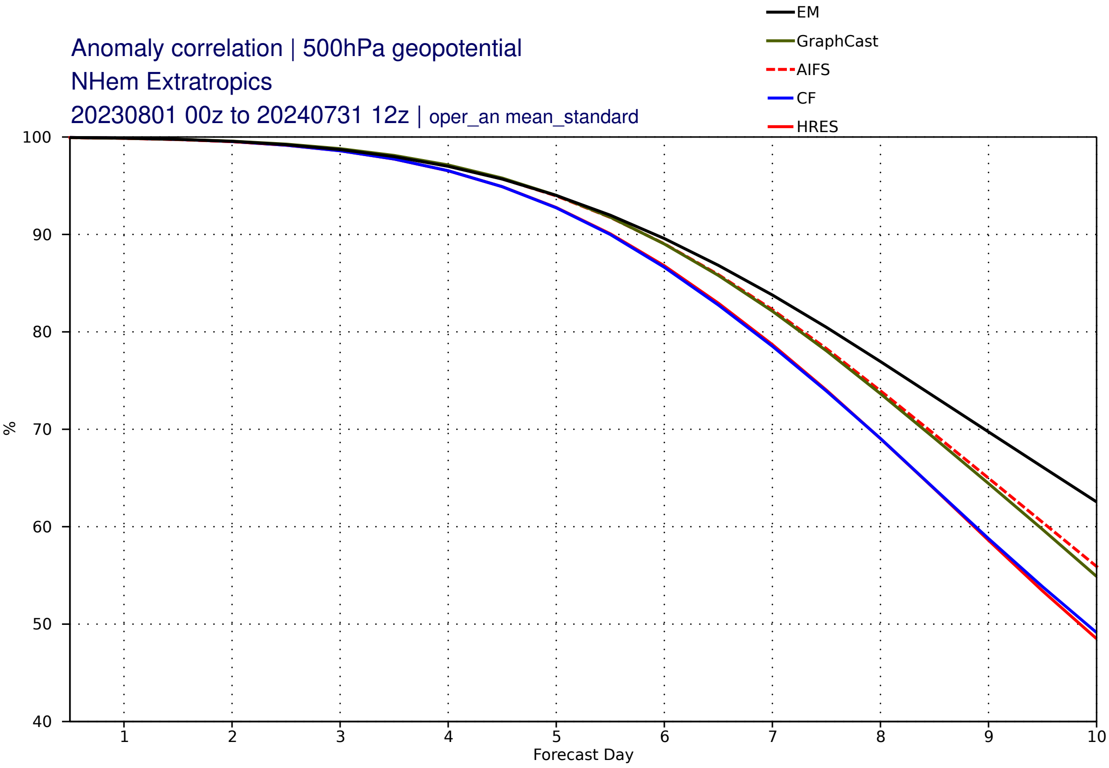


Figure 18: Anomaly correlation of 500 hPa geopotential in the northern extratropics for the 12-month period Aug 2023 to July 2024. Black: ENS mean, olive: GraphCast ML forecast, red dashes: AIFS, blue: ENS control, red: HRES.

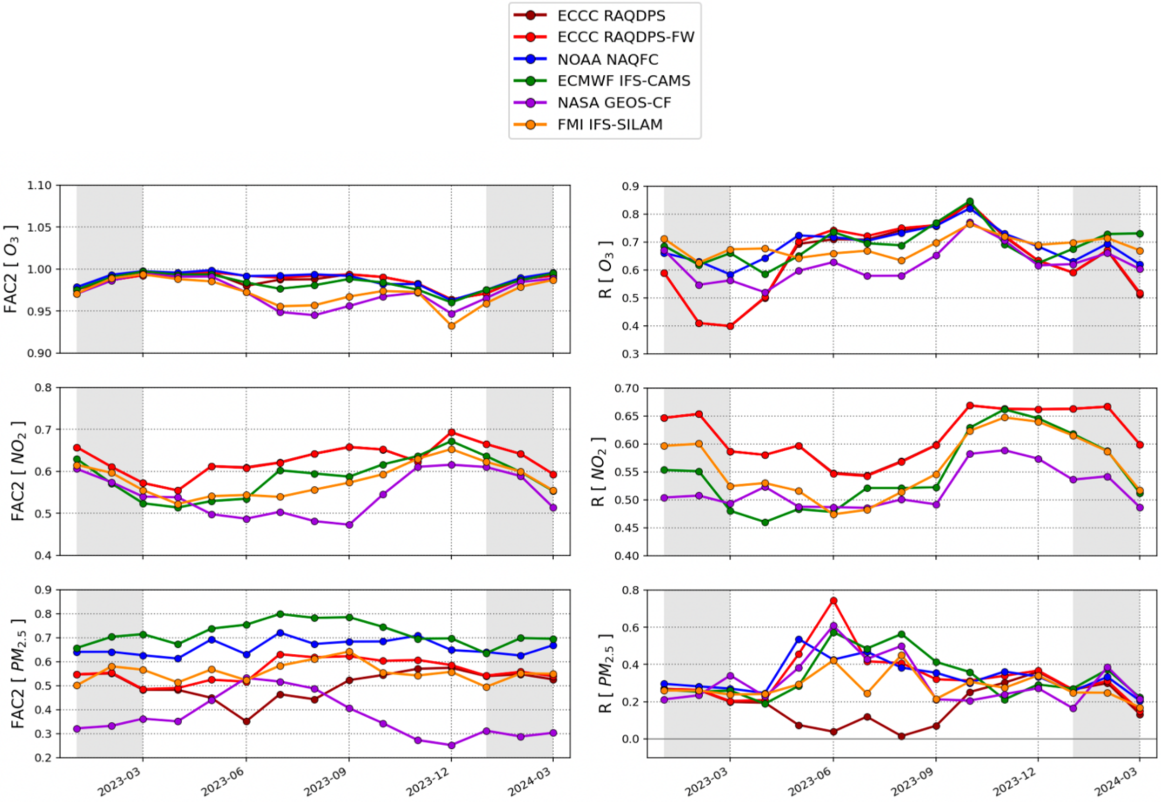


Figure 19: Verification of Ozone (top), NO_2 (centre), and $PM_{2.5}$ (bottom) over a common North American domain in terms of the factor-of-2-fraction, i.e. the fraction of cases where the forecast was within a factor of two of the observation (left), and linear correlation (right). Shown is a model intercomparison over a 15-month period in terms of monthly scores. ECMWF's CAMS forecast is shown in green. Taken from Presseau et al. (2024).

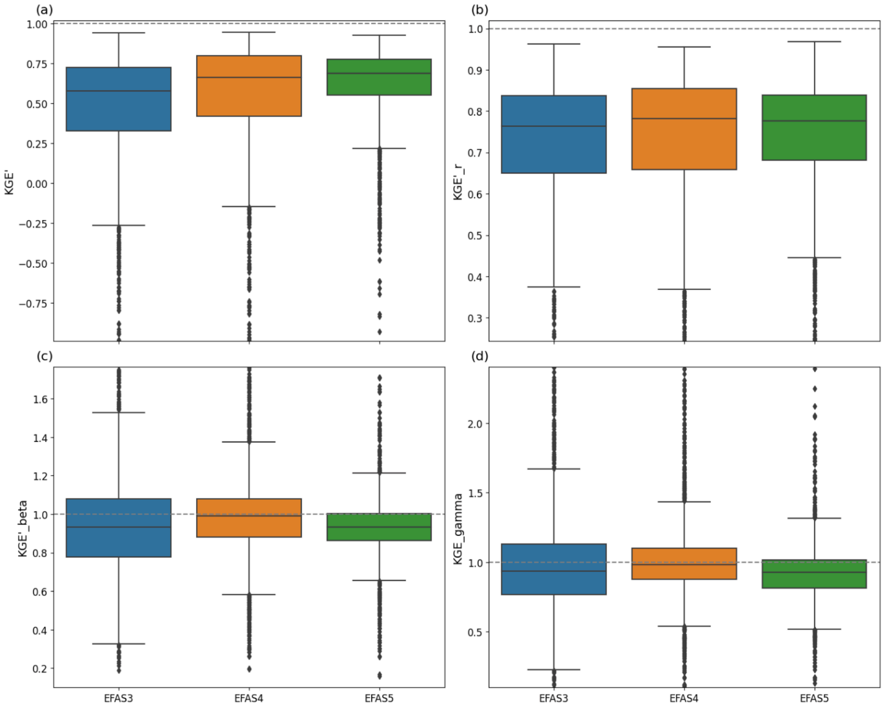


Figure 20: Range of (a) the Modified Kling-Gupta Efficiency (KGE'), and its three components (b) correlation (KGE'_r), (c) bias ratio (KGE'_beta) and (d) variability ratio (KGE'_gamma), for three successive versions of EFAS (EFAS3 in blue, EFAS4 in orange and EFAS5 in green). The statistics are based on data from 1,807 stations with observed river discharge across Europe, obtained from the CEMS database. Box plots show the data distribution, with the box representing the interquartile range (IQR), the line inside indicating the median, and whiskers extending to 1.5 times the IQR. Outliers are shown as individual points beyond the whiskers. The grey dashed line shows the optimum for KGE' and each of its components.

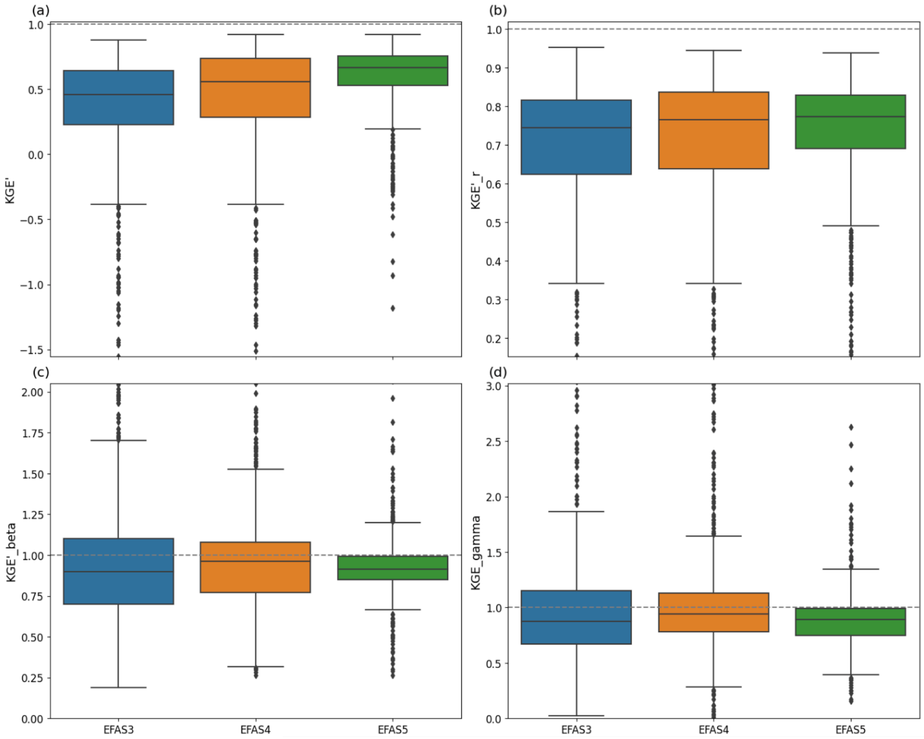


Figure 21**:** Same as Figure 20, but for a subset of 778 stations from the CEMS dataset with an upstream area smaller than 1500 km^2.

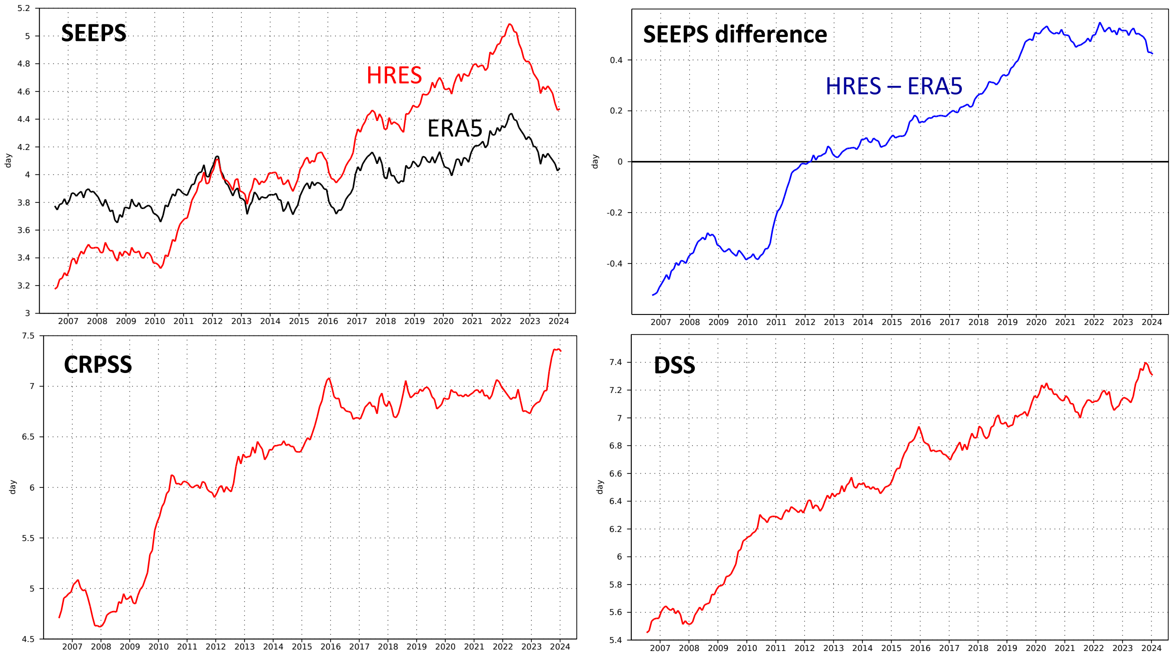


Figure 22: Supplementary headline scores (left column) and additional metrics (right column) for deterministic (top) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number 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).

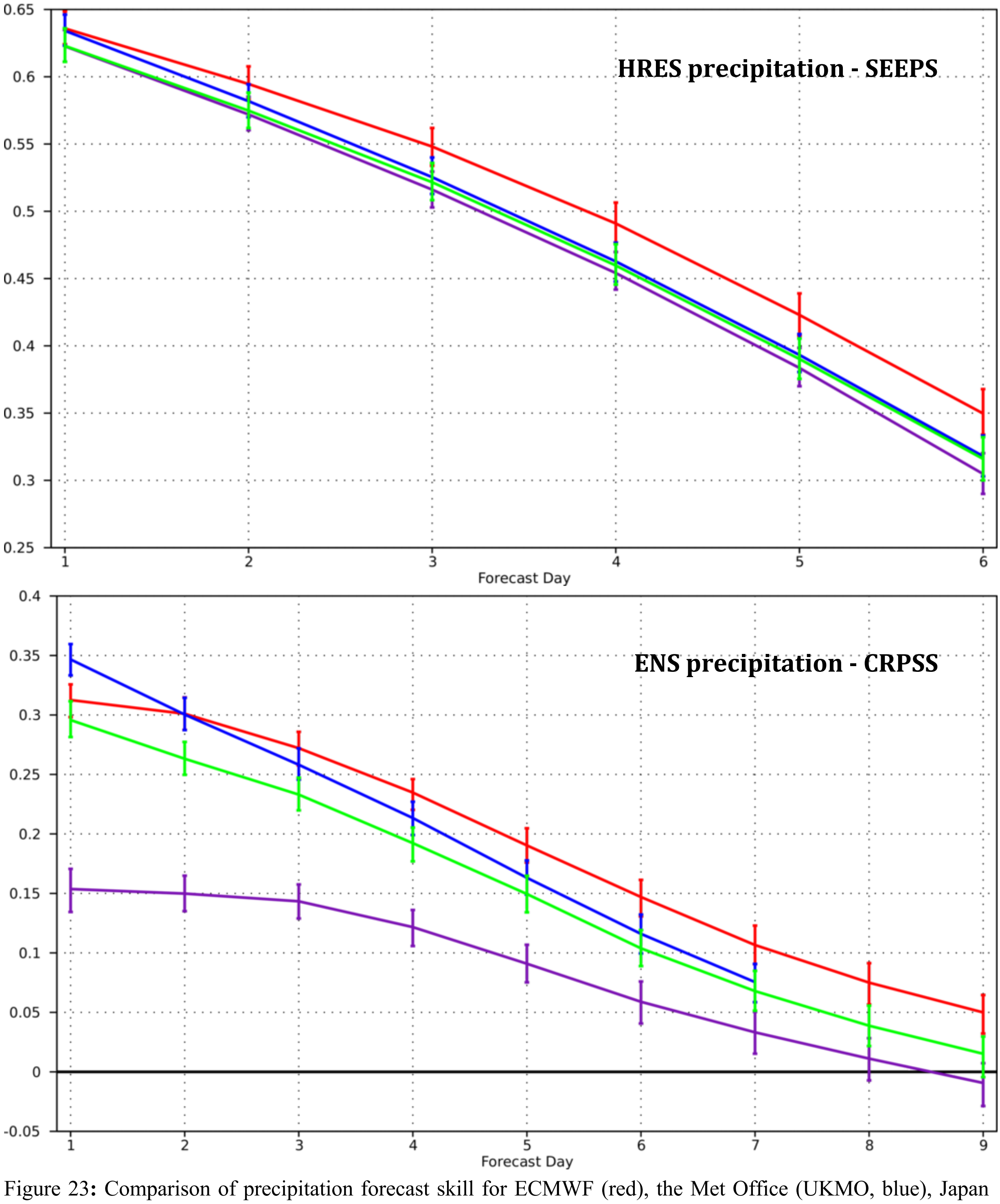


Figure 23**:** Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 22. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2023–July 2024. Bars indicate 95% confidence intervals. Note that for the CRPSS plot, scores for other centres have been computed based on data in the TIGGE archive, i.e. from forecasts at a lower resolution than their native grids.

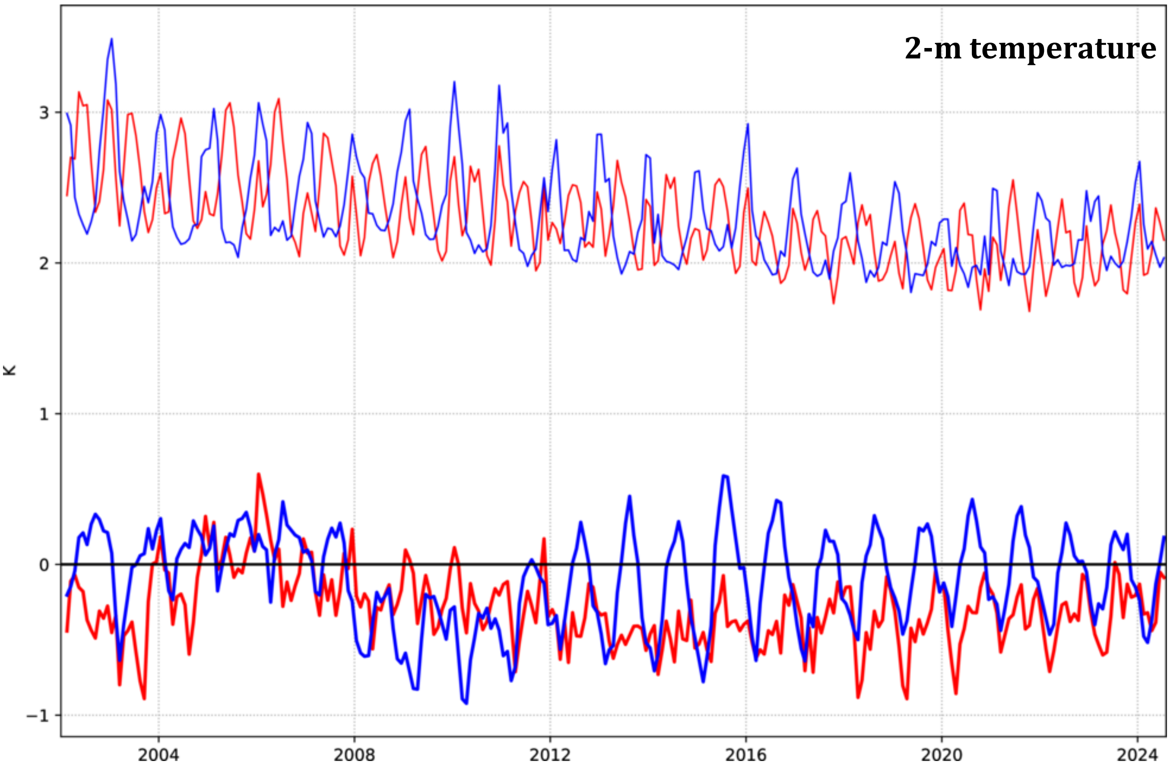


Figure 24**:** Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

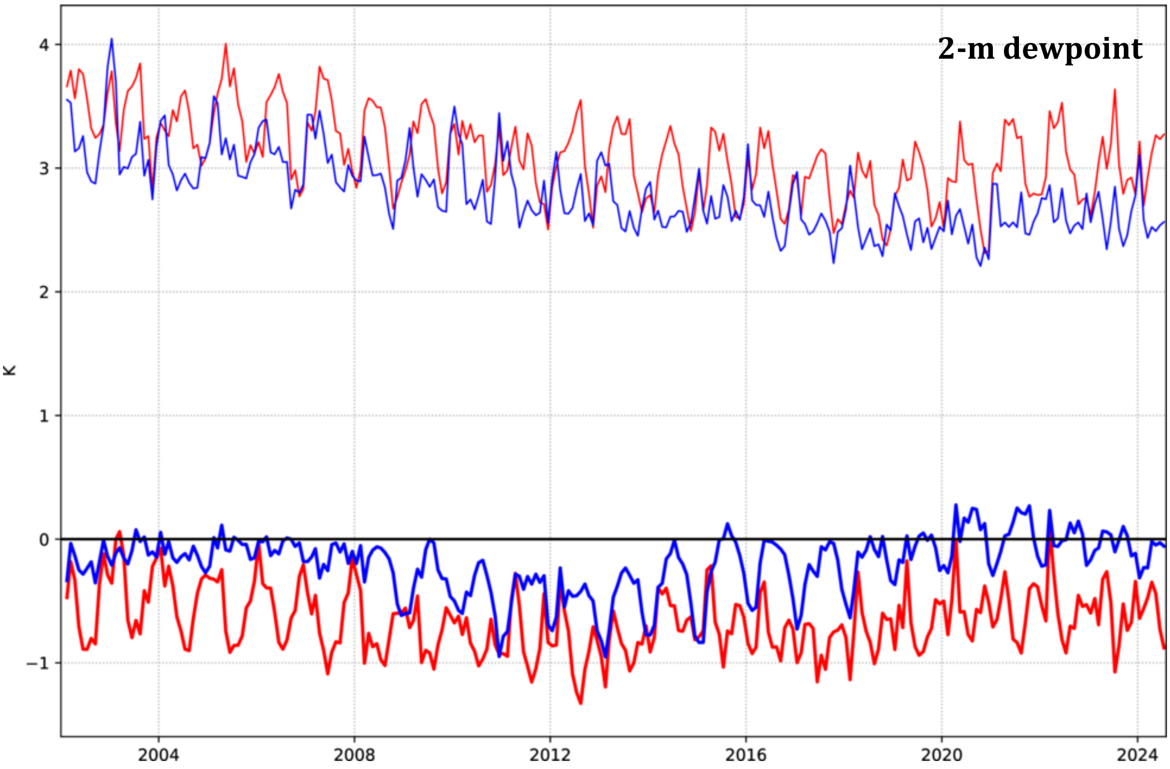


Figure 25**:** Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

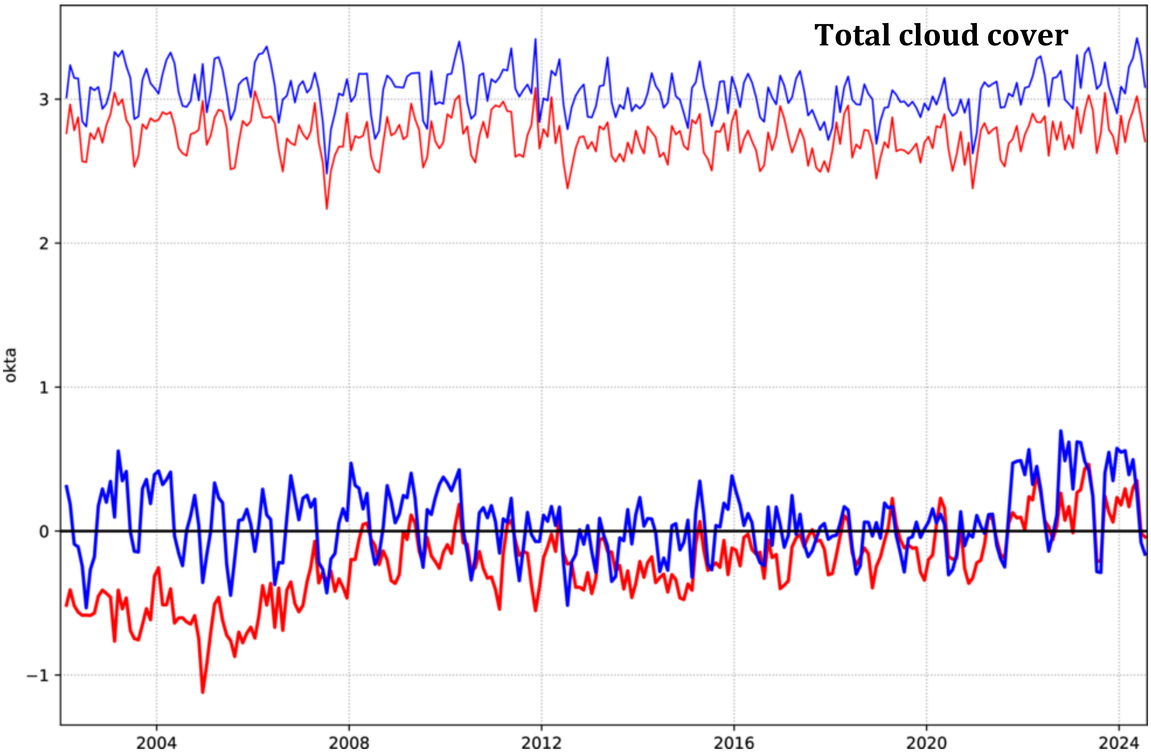


Figure 26: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

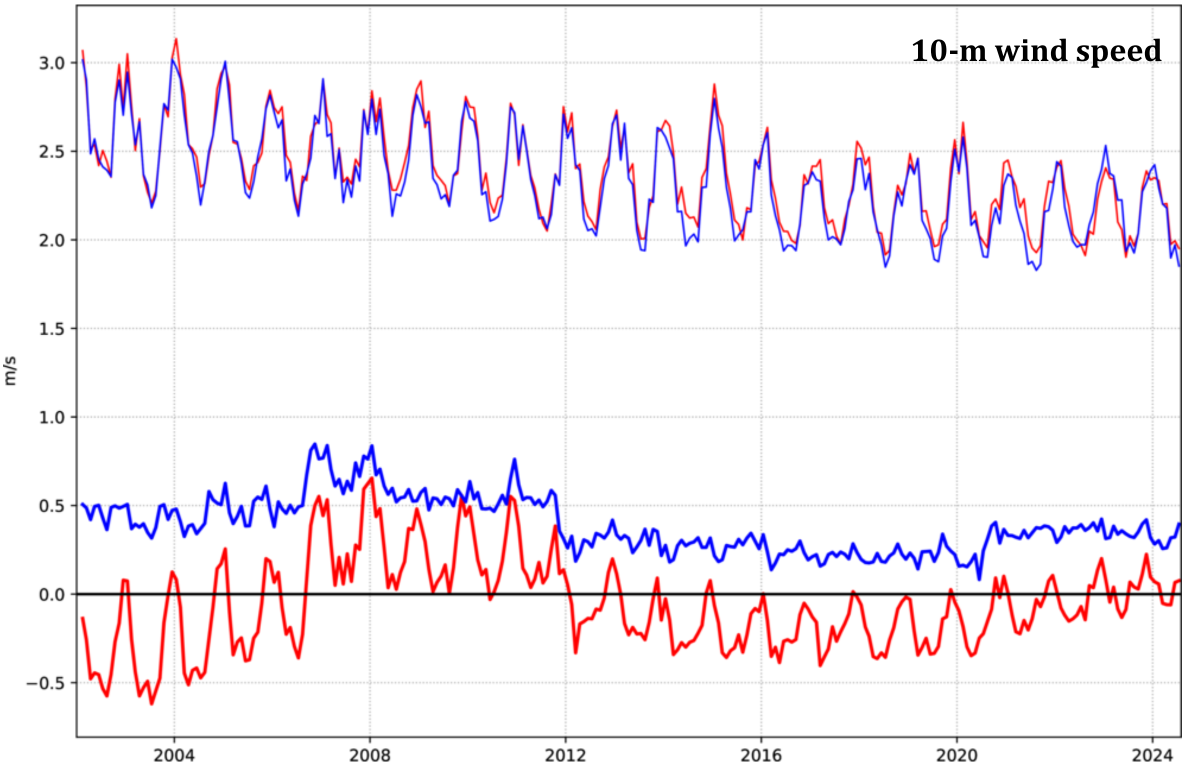


Figure 27: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

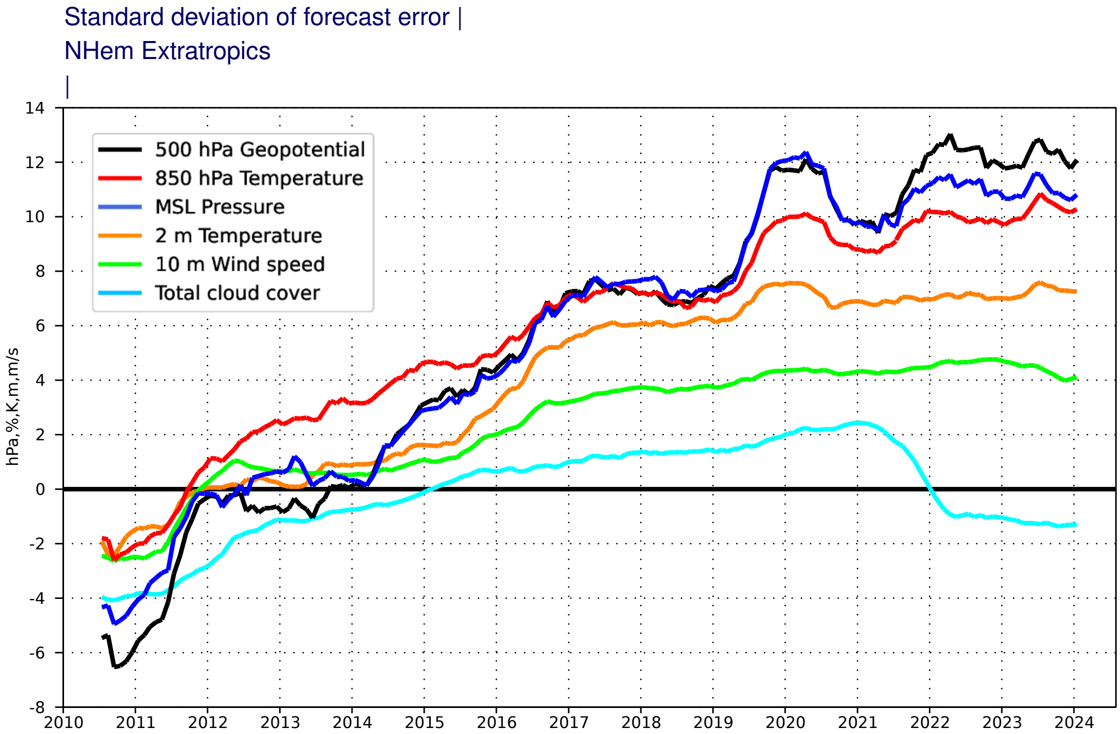


Figure 28: Evolution of skill of the HRES forecast at day 5 in the northern hemisphere extratropics, expressed as relative skill compared to ERA5. Verification is against analysis for 500 hPa geopotential, 850 hPa temperature, and mean sea level pressure, using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature, 10 m wind speed, and total cloud cover.

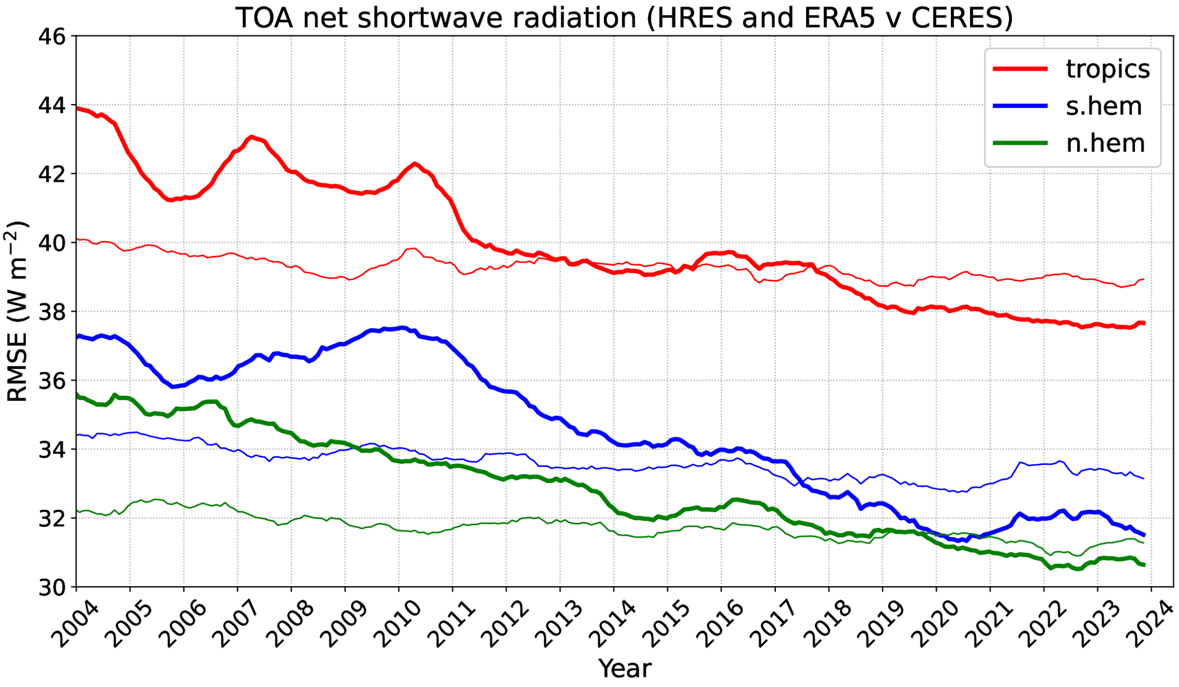


Figure 29: Evolution of the RMSE of the HRES forecast at day 5 (bold lines) of the top of the atmosphere (TOA) net shortwave radiation for the two extratropical hemisphere and the tropics. Thin lines show the RMSE of the ERA5 forecast for comparison. Verification is against CERES satellite data.

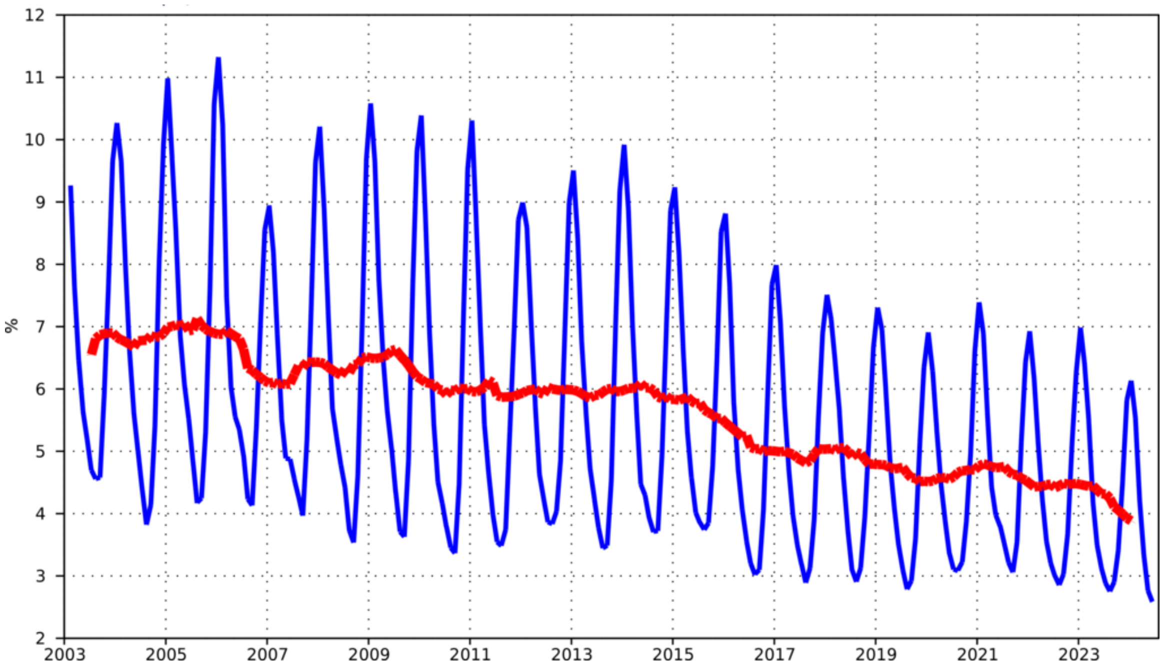


Figure 30: Evolution of the fraction of large ENS 2m temperature errors (CRPS>5K) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

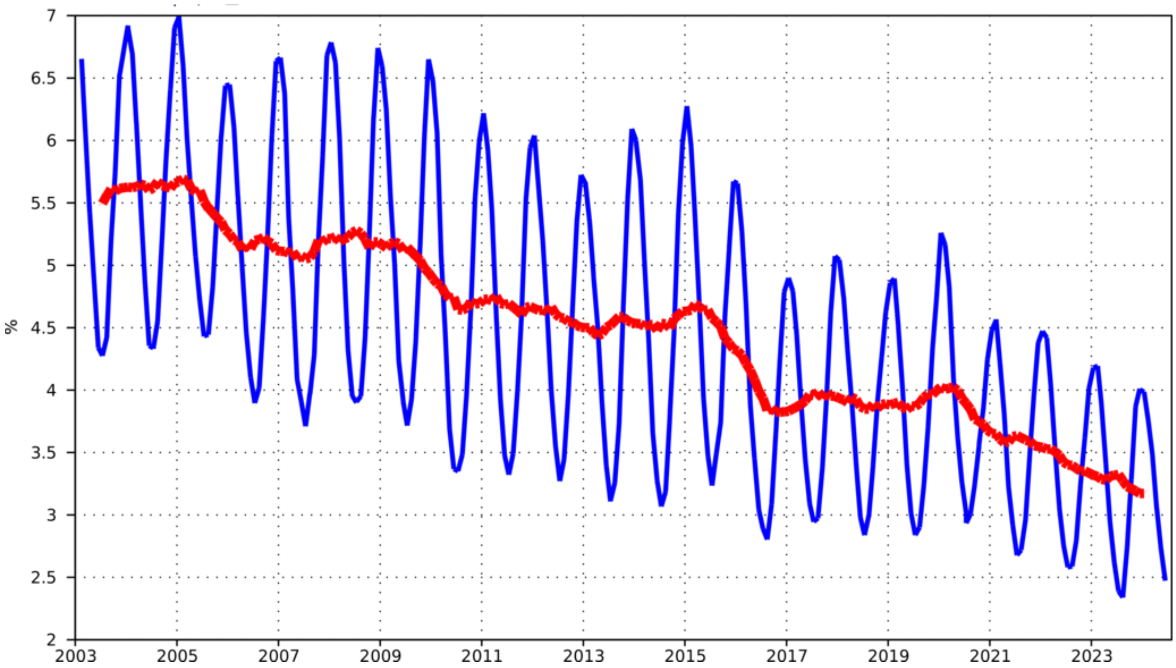


Figure 31: Evolution of the fraction of large ENS 10m wind speed errors (CRPS>4m/s) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

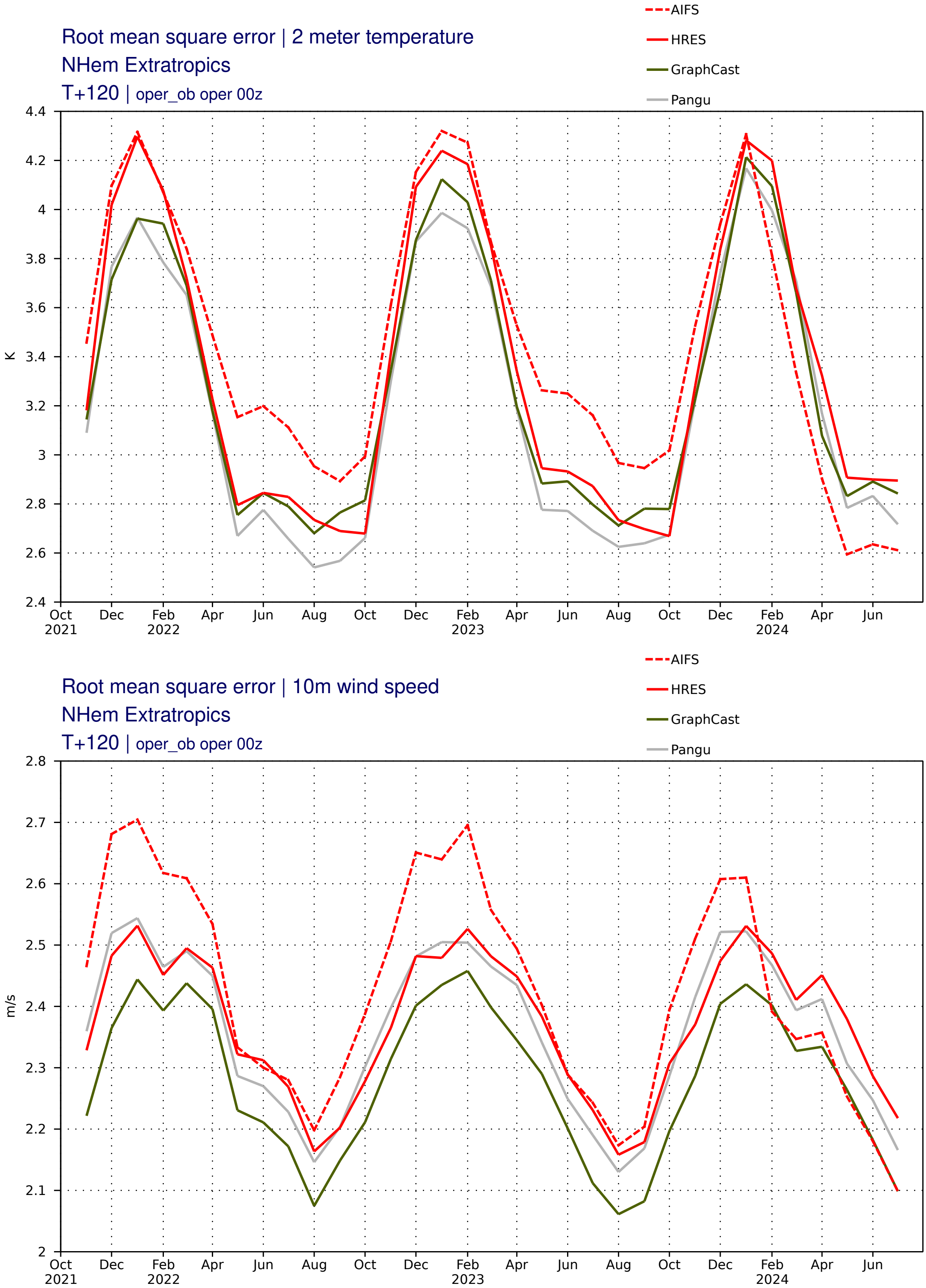


Figure 32: RMSE of 2 m temperature (top) and 10 m wind speed (bottom) at day 5 in the northern extratropics, comparing various ML forecasts with the HRES. Verification is against SYNOP observations. Shown are monthly averages.

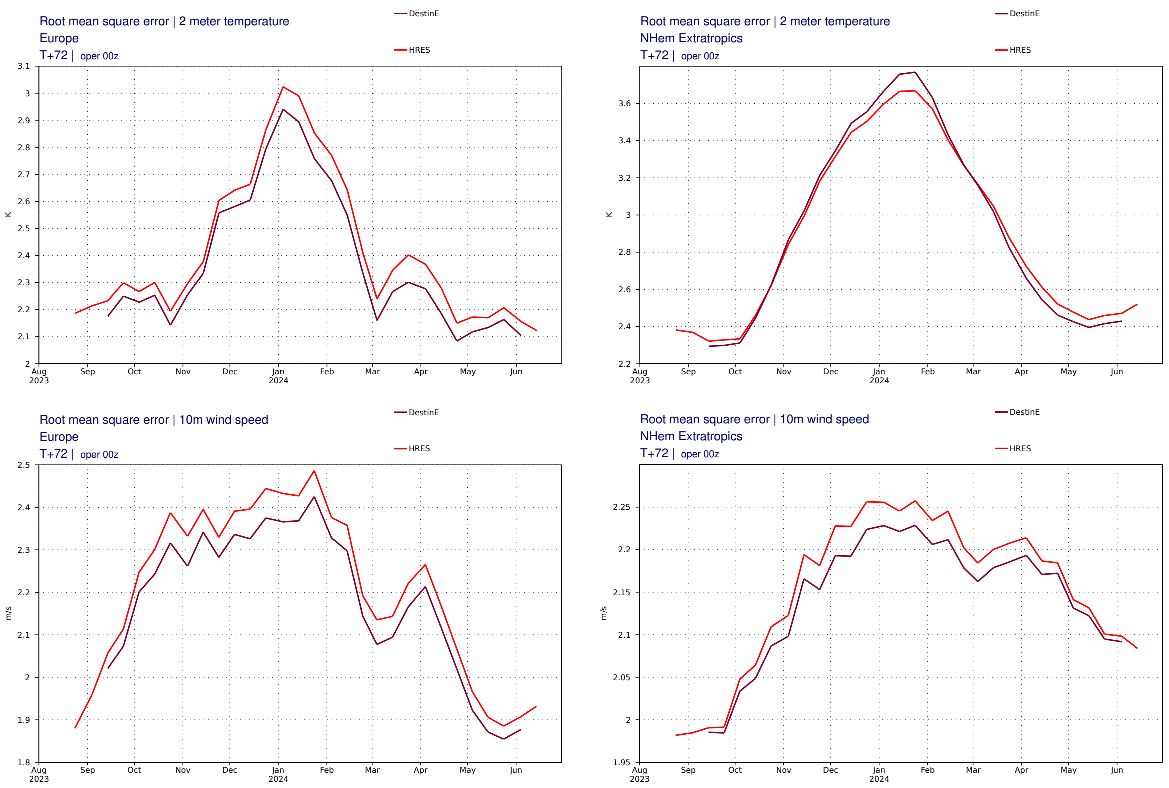


Figure 33: RMSE of HRES and DestinE Extremes DT (4.4 km) 00Z runs for 2 m temperature (top) and 10 m wind speed forecasts (bottom) at day 3 in Europe (left panels) and the northern extratropics (right panels). Verification is against SYNOP observations, shown are monthly averages.

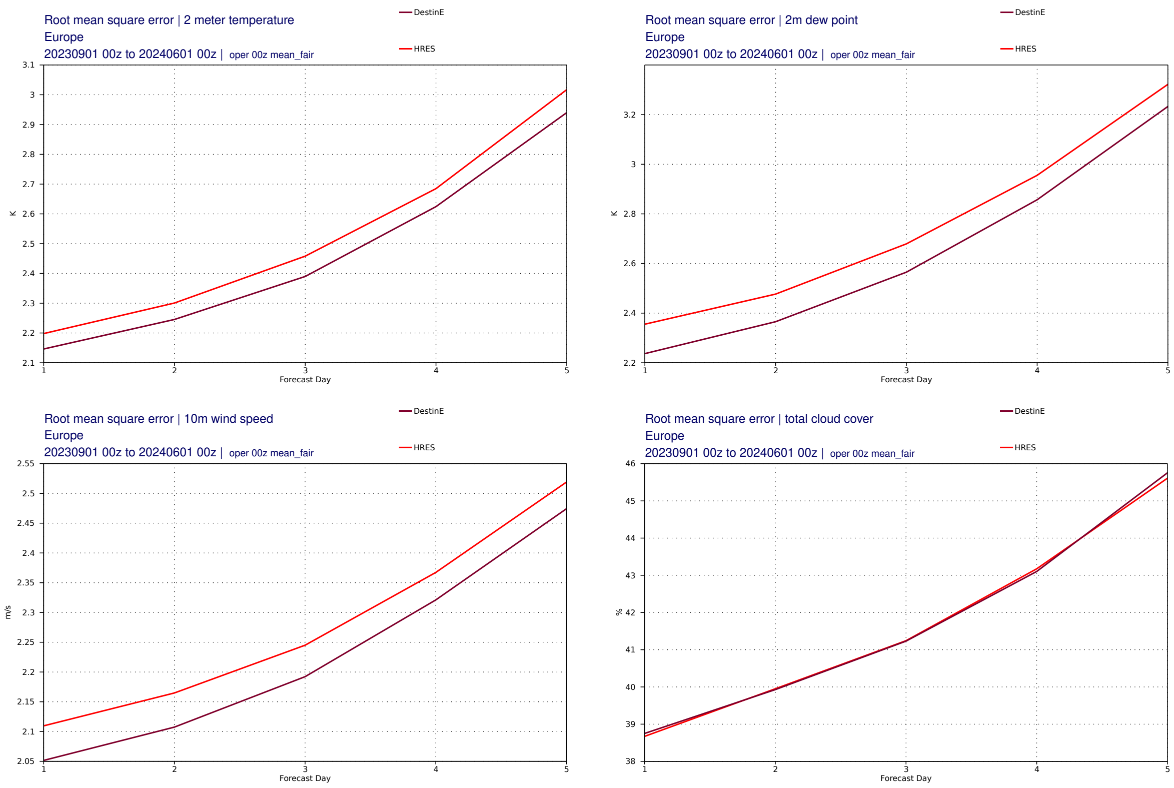
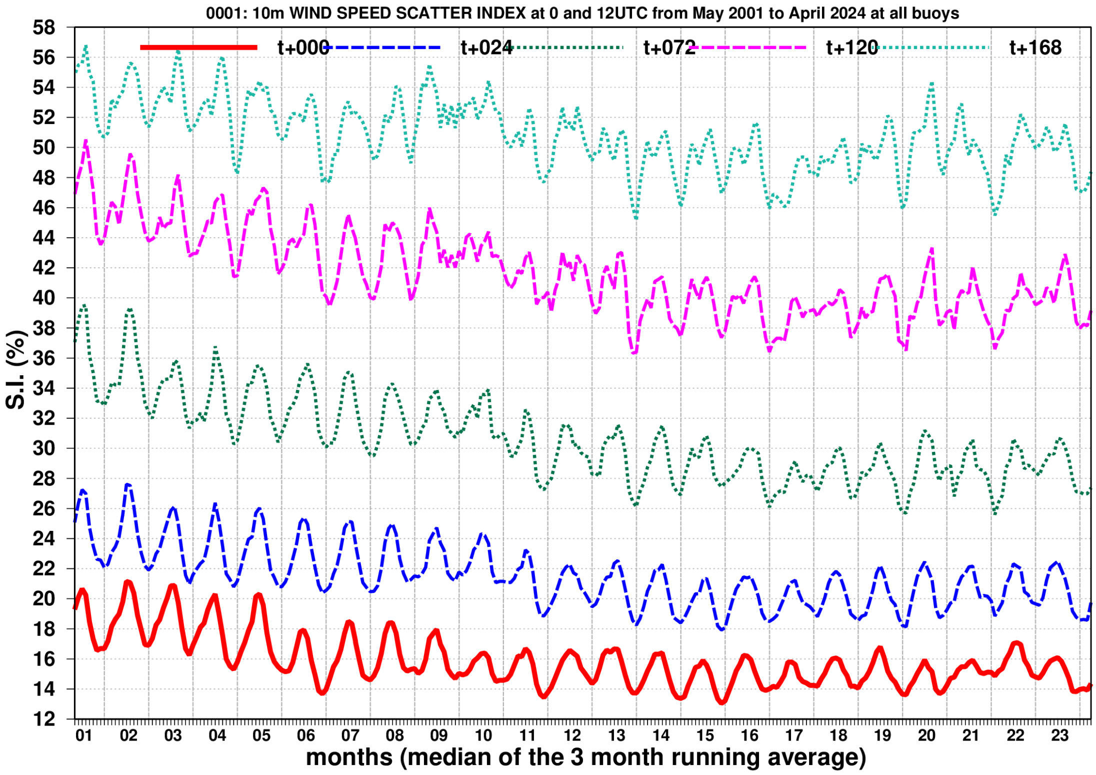


Figure 34: RMSE of HRES and DestinE Extremes DT (4.4. km) 00Z runs for various weather parameters as a function of lead time. Verification is against SYNOP observations. Verification period is Sep 2023 to June 2024.



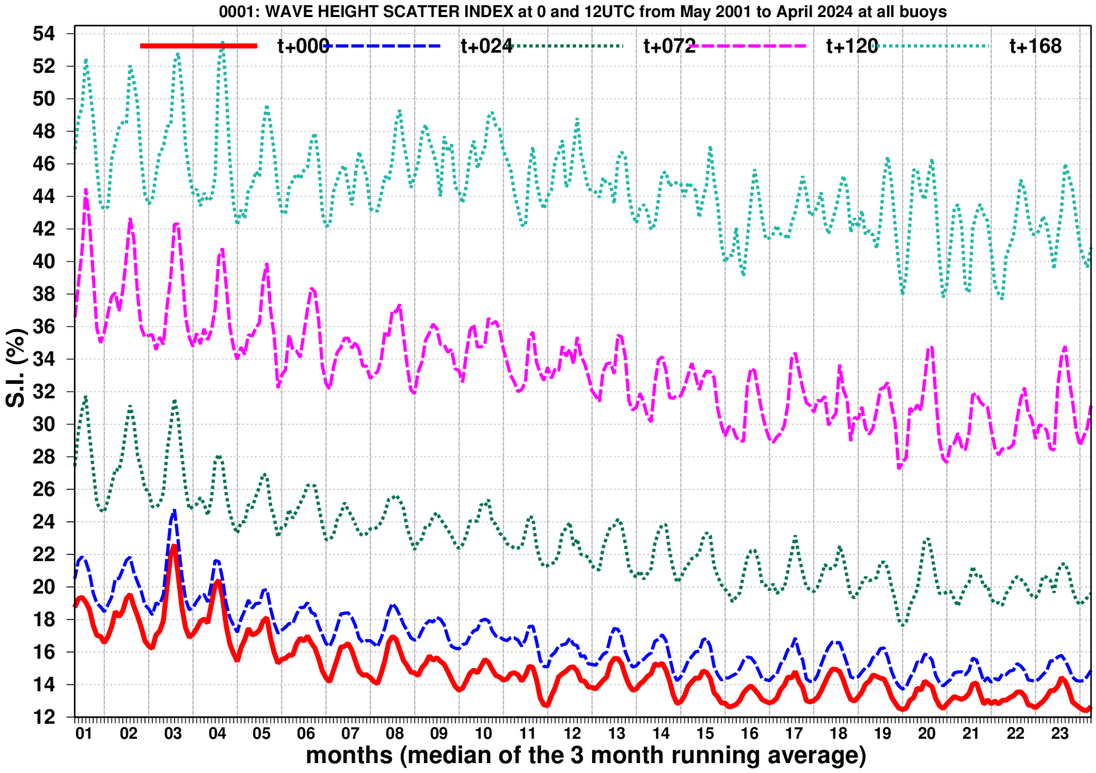
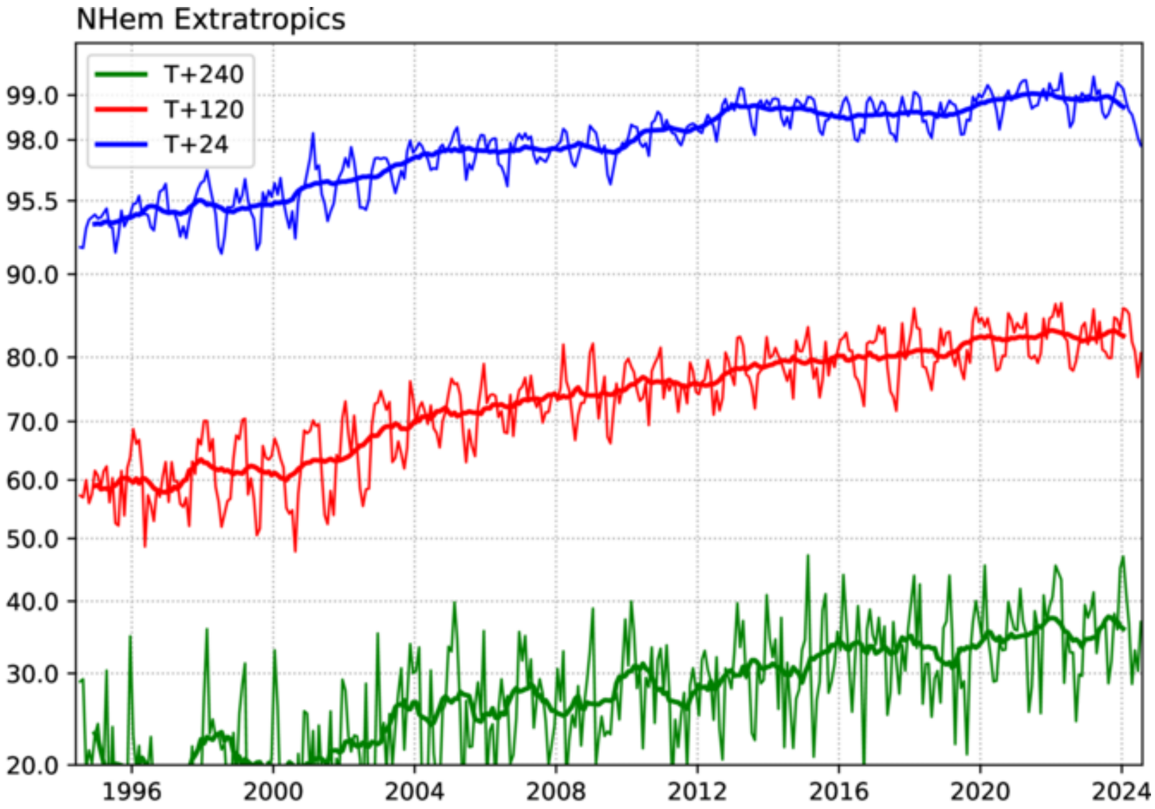


Figure 35: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave height forecast (bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is applied.



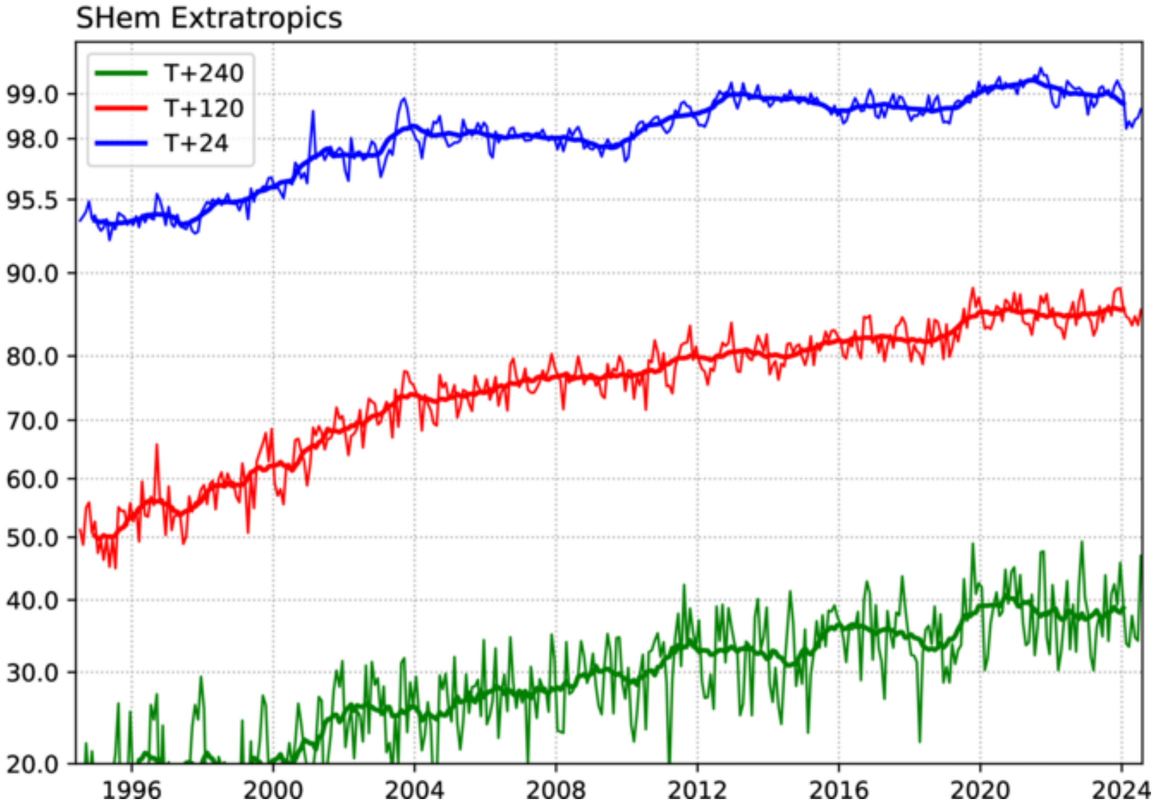


Figure 36: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

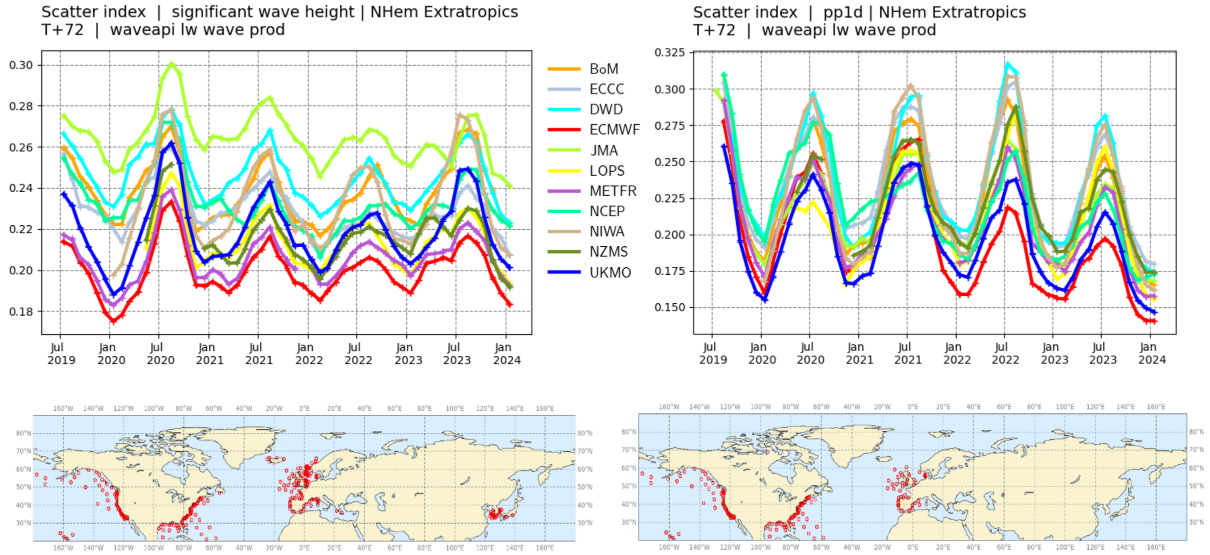


Figure 37: Verification of forecasts of wave height and peak wave period (upper panels) at +72 h using observations from wave buoys (lower panels). The scatter index (SI) is the standard deviation of error normalised by the mean observed value. NCEP: National Centers for Environmental Prediction, USA; 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; NIWA: National Institute of Water and Atmosperic Research, New Zealand.

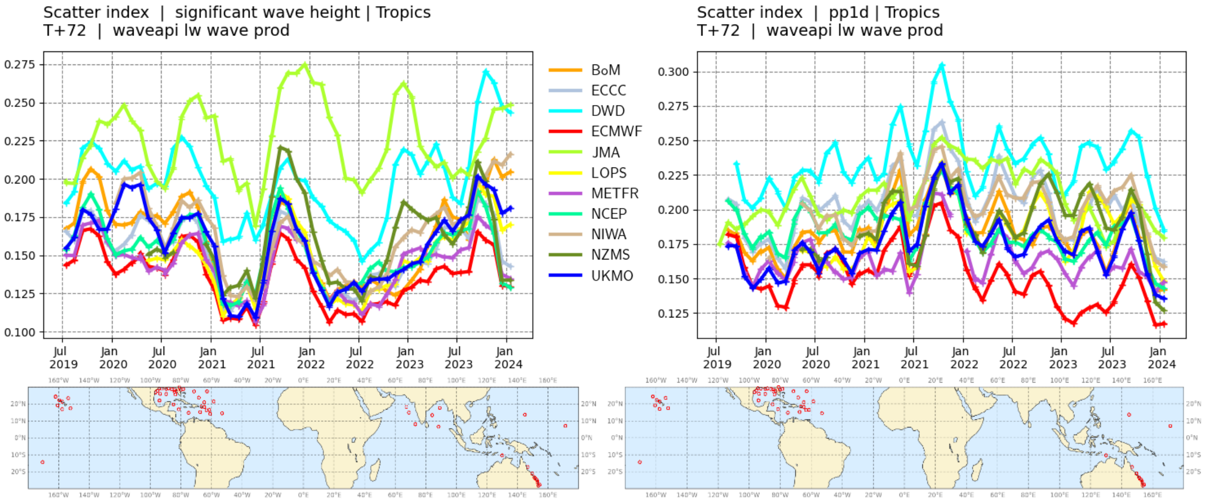


Figure 38: Same as Figure 37, but for the tropics.

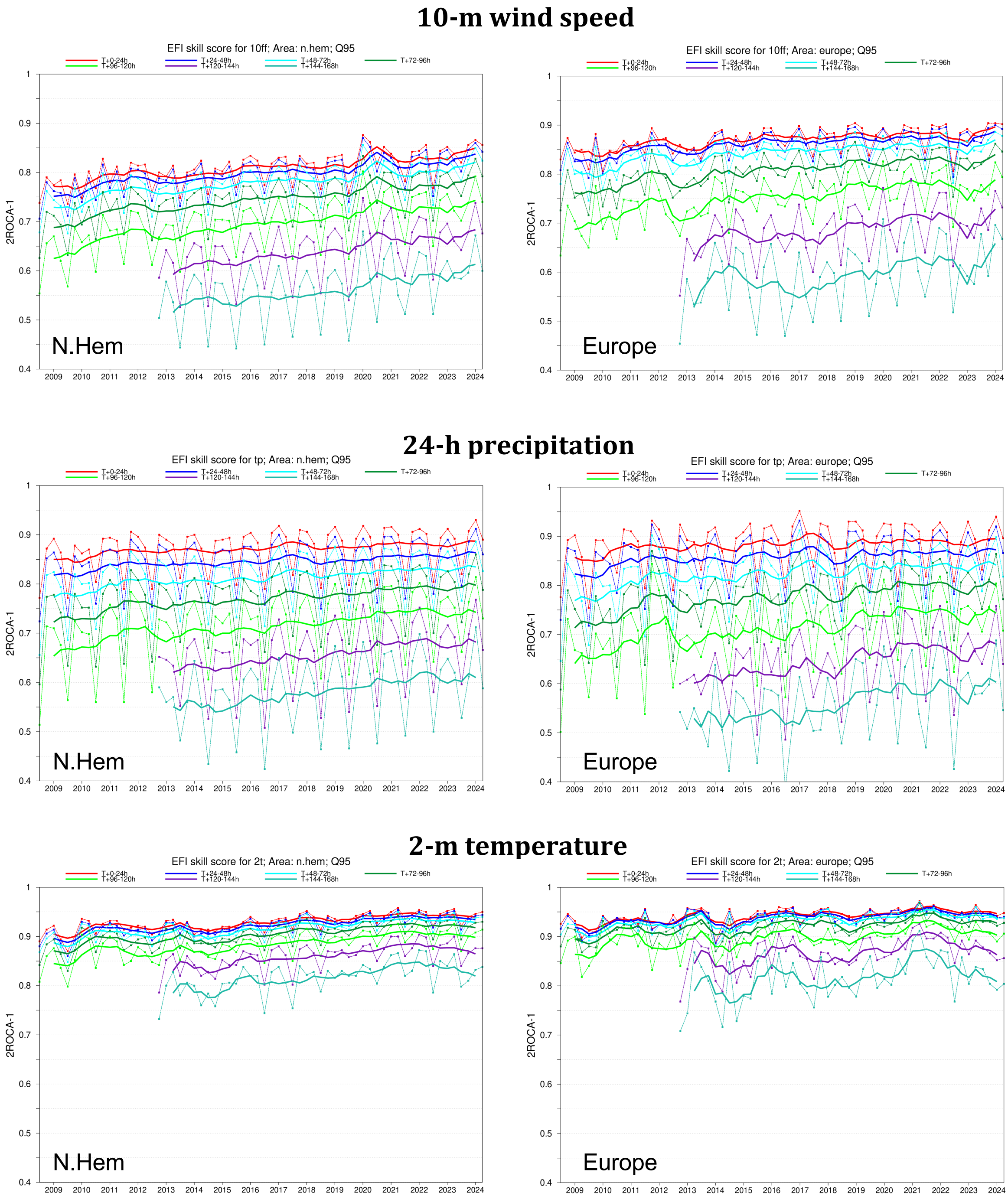


Figure 39: Verification of the Extreme Forecast Index (EFI) against SYNOP observations. Left column: Northern Extratropics, right column: Europe. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 7 (24-hour period 144-168 hours ahead); skill at day 4 (light green line) is the supplementary headline score; an extreme event is taken as an observation exceeding 95th percentile of station climate. Curves show seasonal values (thin) and four-season running mean (bold) of relative operating characteristic (ROC) area skill. Centre and bottom rows show the equivalent ROC area skill for precipitation EFI forecasts and for 2 m temperature EFI forecasts.

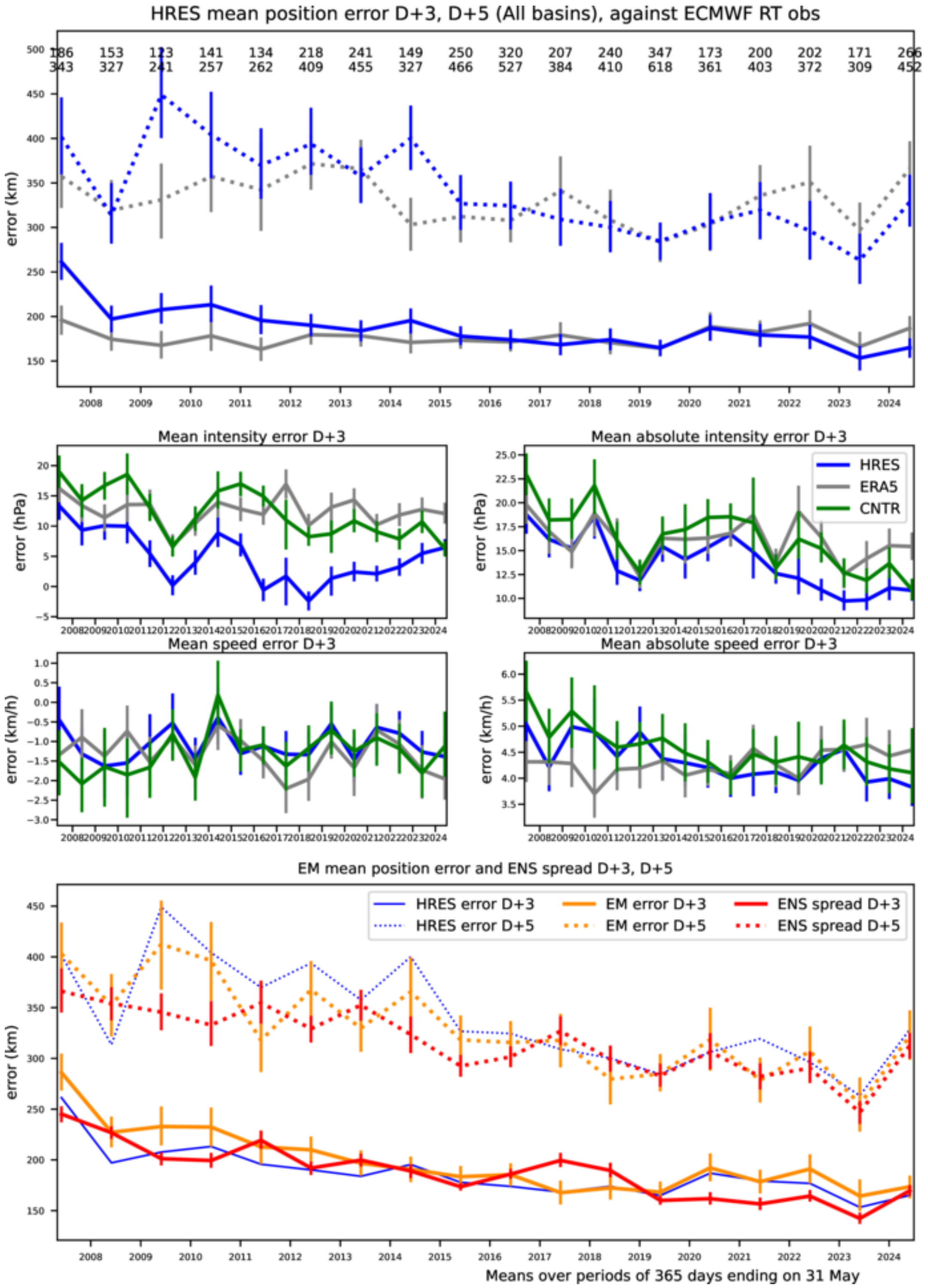


Figure 40: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary 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.

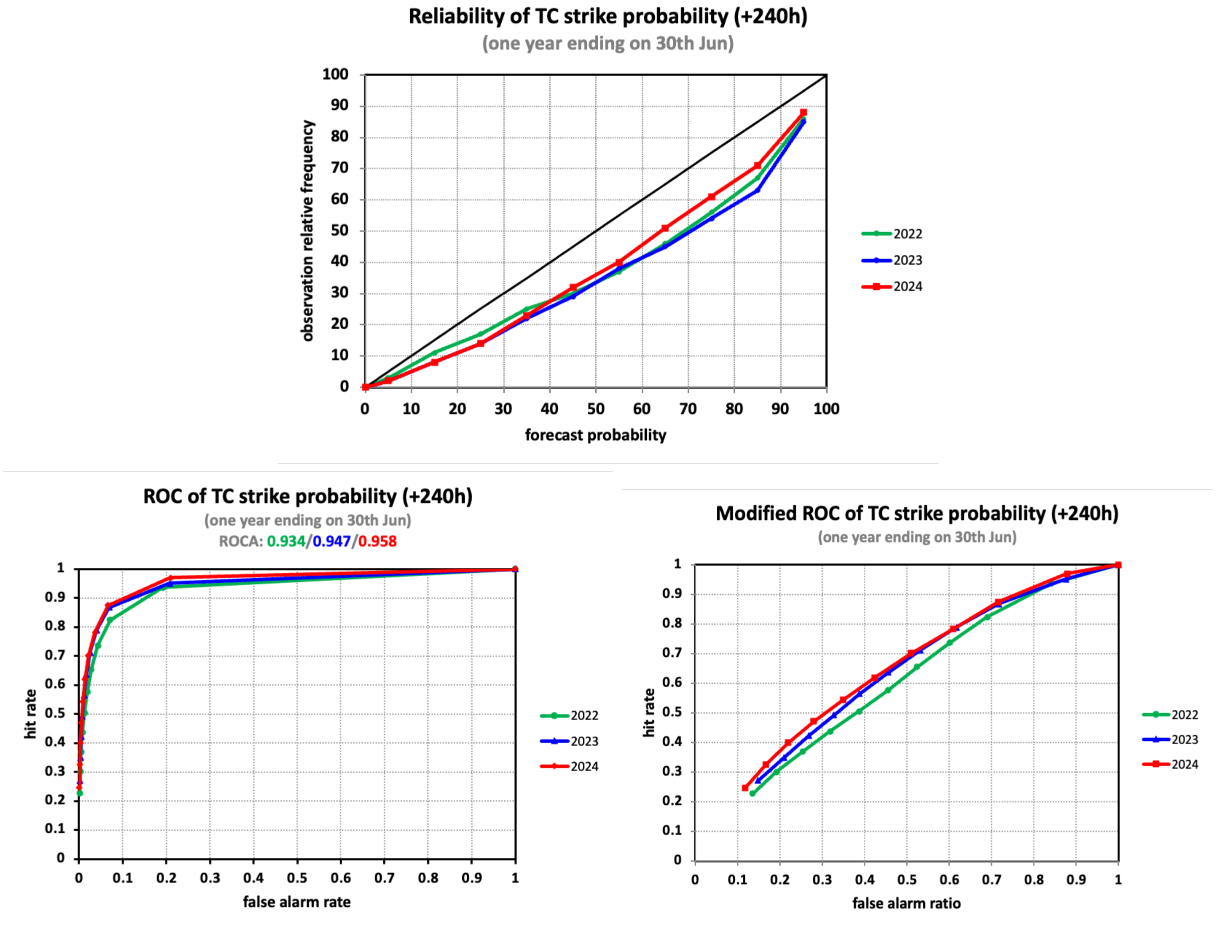


Figure 41: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2021–June 2022 (green), July 20222–June 2023 (blue) and July 2023–June 2024 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

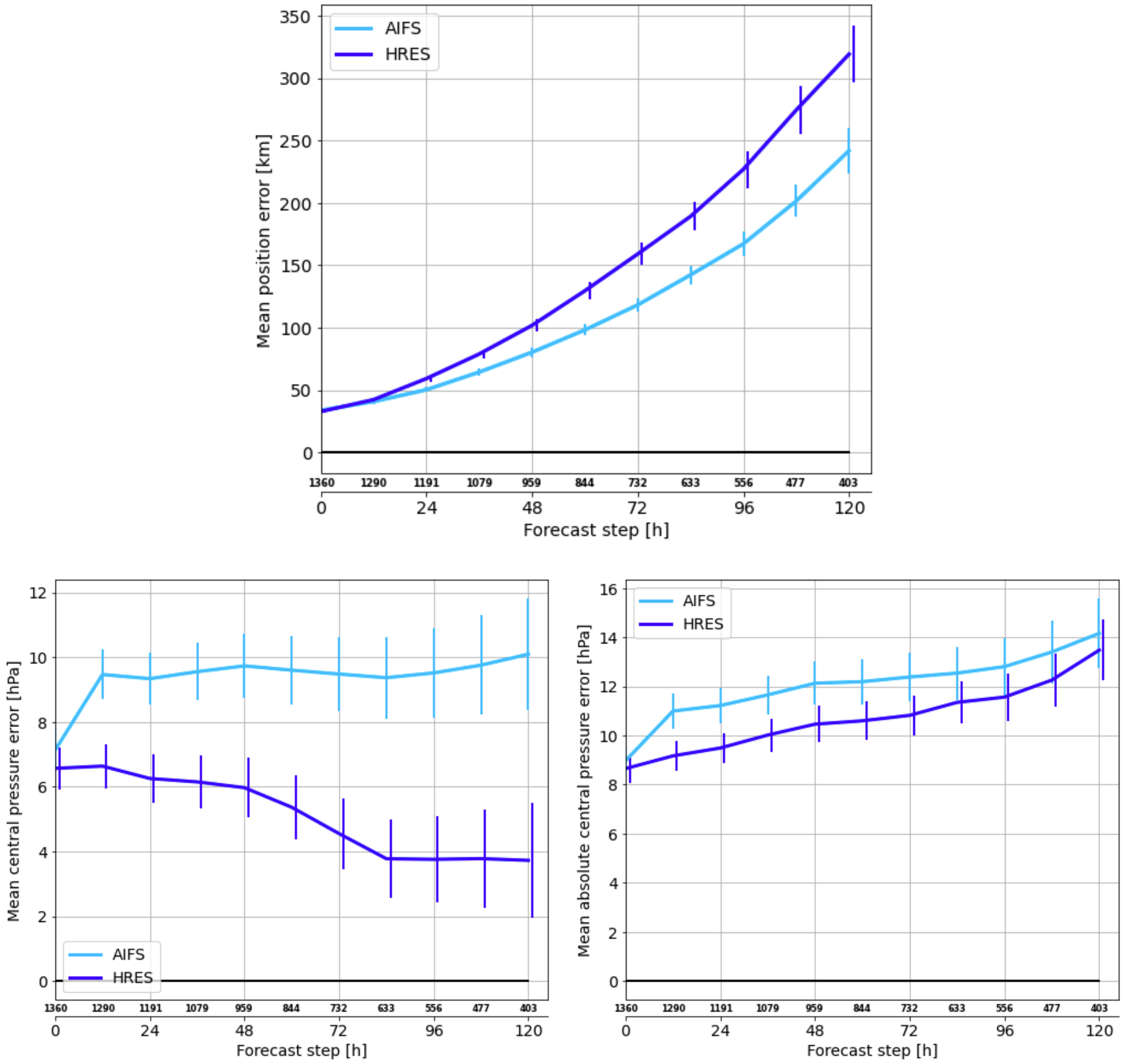


Figure 42: Comparison of HRES and AIFS errors in forecasting tropical cyclone position (MAE, top panel) and tropical cyclone central pressure (lower left: ME, lower richt: MAE). This is based on a global verification against the International Best Track Archive for Climate Stewardship (IBTrACS v4r01). Verification period is Jan 2022 – Dec 2023, including 00 and 12 UTC forecasts. Forecasts are homogenized to have a consistent number of cases between models. The verification is based on TCs that are present in the observation database at the forecast initial time. For each lead time, the number of cases is displayed directly below the graphs. Vertical bars indicate the 2.5%–97.5% confidence intervals.

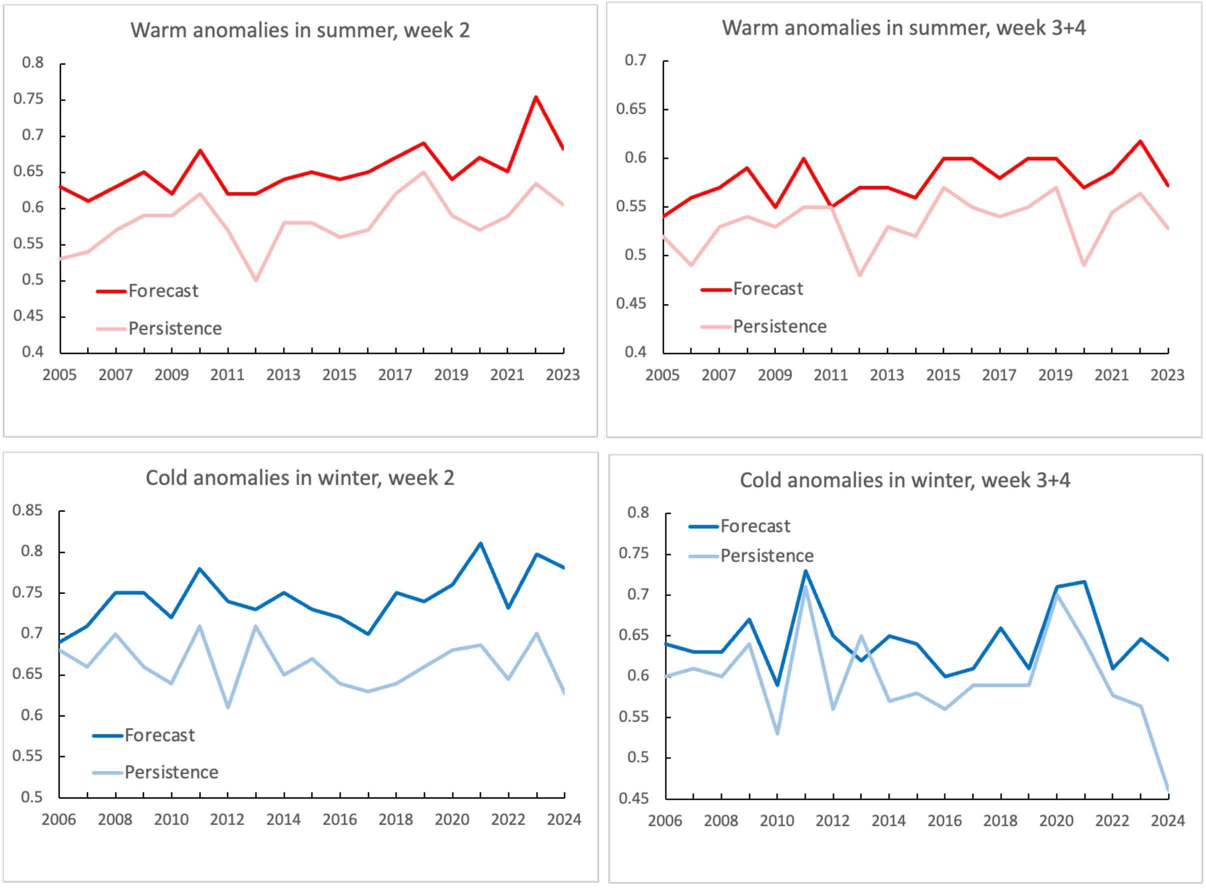


Figure 43**:** Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all 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.

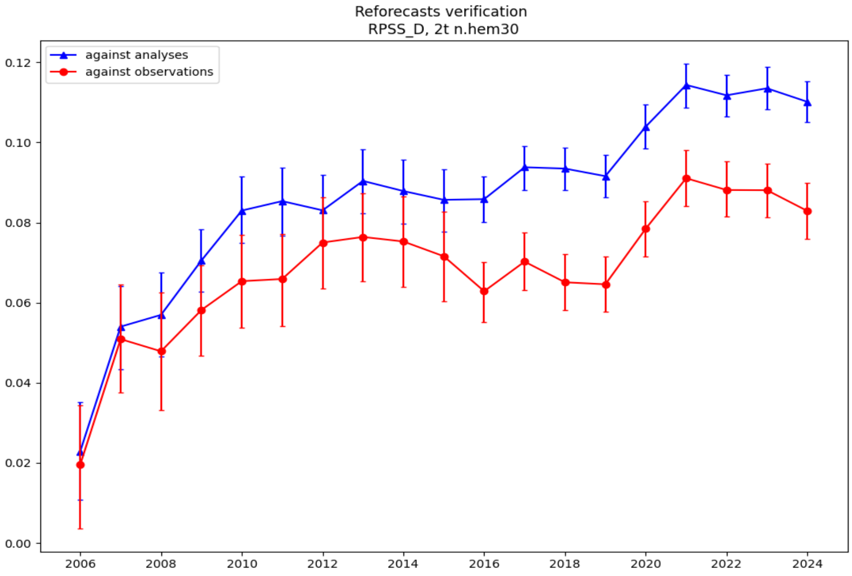


Figure 44: Skill of the sub-seasonal forecast in predicting weekly mean 2 m temperature anomalies (terciles) in week 3 in the northern extratropics (north of 30°N). Verification against ERA5 analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.

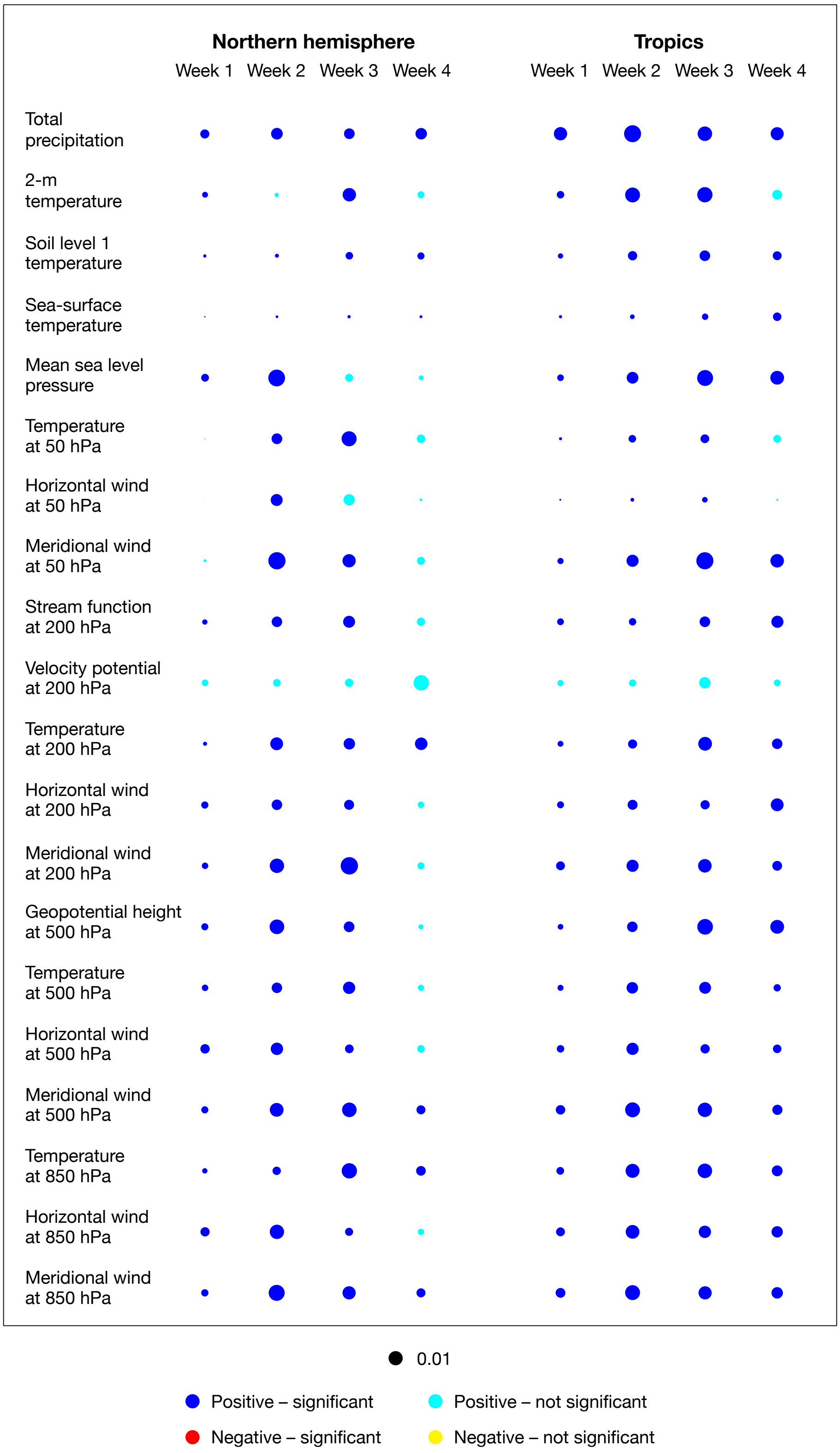
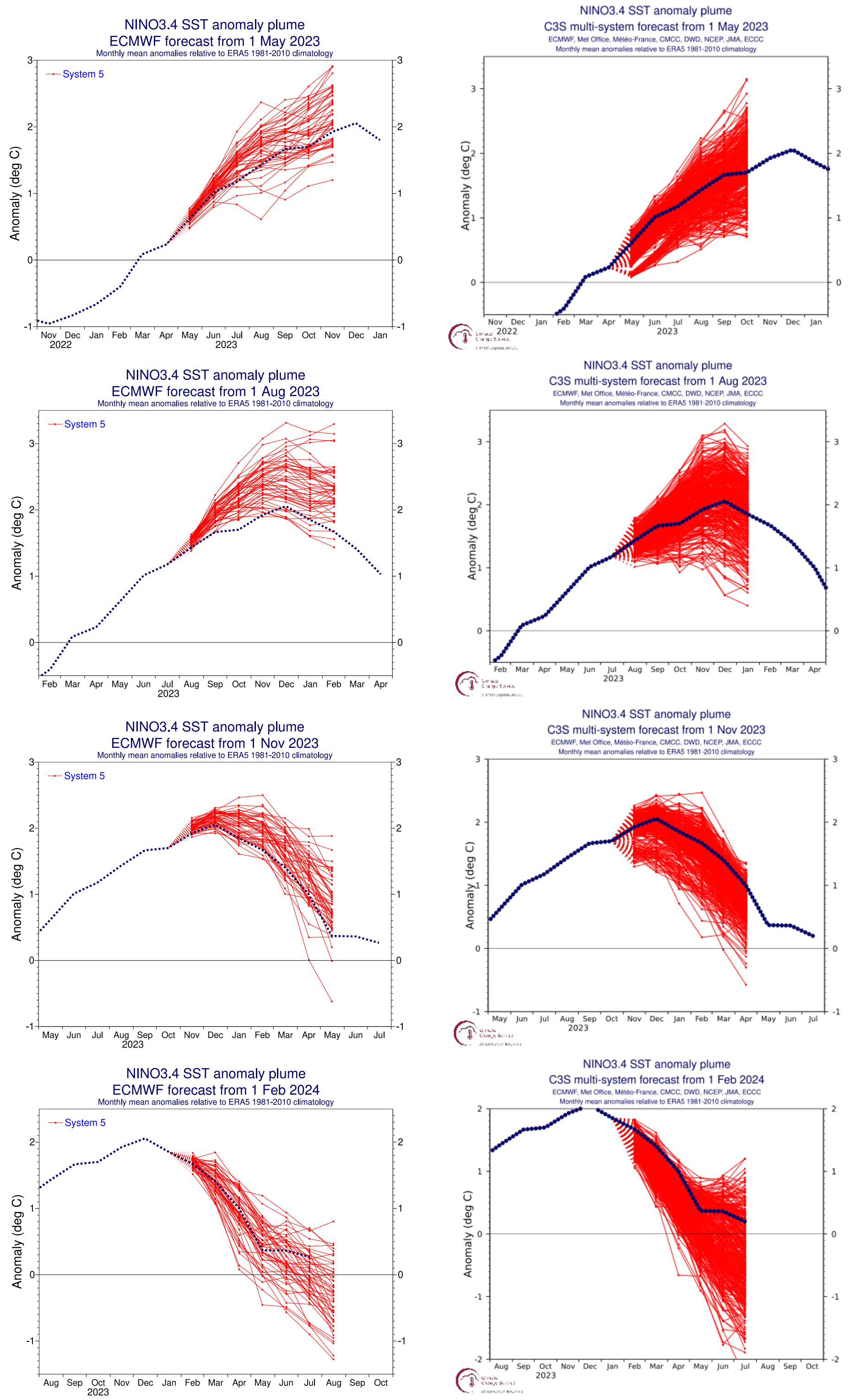


Figure 45: Scorecard showing the difference in the Continuous Ranked Probability Skill Score (CRPSS) between the 101- and 51-member ensembles for 20 variables, four lead times (weeks 1 to 4) and two regions (northern hemisphere on the left and tropics on the right). The blue (red) and cyan (yellow) colours indicate an improvement (a degradation), respectively. Blue or red indicate that the difference is statistically significant using a 10,000 bootstrap resampling technique. Re-forecasts were produced the first of each month over the period 1989–2016. From Vitart et al. (2022).



NINO3.4 SST anomaly plume
ECMWF forecast from 1 May 2023
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 May 2023
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Aug 2023
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Aug 2023
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Nov 2023
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Nov 2023
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
ECMWF forecast from 1 Feb 2024
Monthly mean anomalies relative to ERA5 1981-2010 climatology

NINO3.4 SST anomaly plume
C3S multi-system forecast from 1 Feb 2024
ECMWF, Met Office, Météo-France, CMCC, DWD, NCEP, JMA, ECCC
Monthly mean anomalies relative to ERA5 1981-2010 climatology

System 5
Anomaly (deg C)
Nov Dec Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec Jan
2022
2023
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
-2
-1
0
1
2
3

Figure 46: 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 2023, August 2023, November 2023, and February 2024. 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.

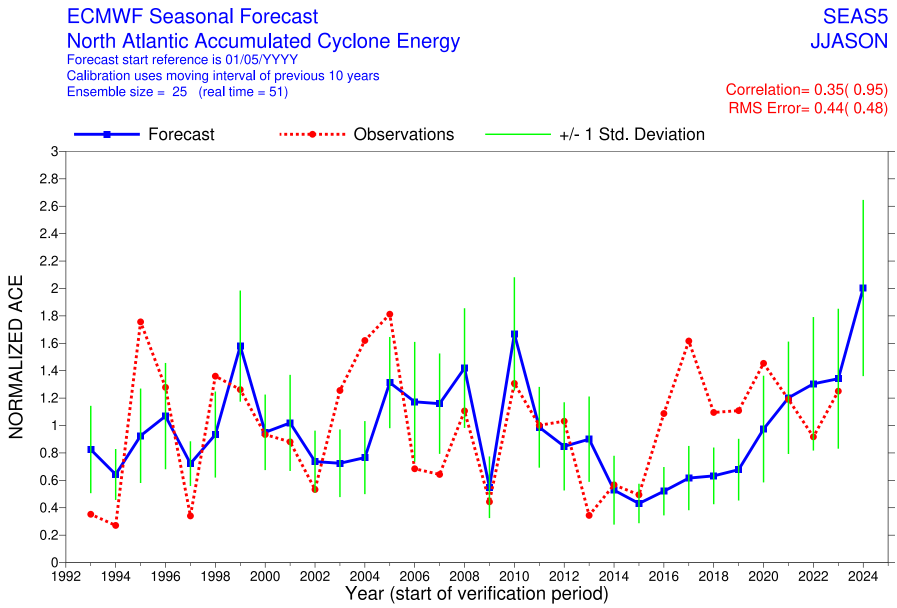


Figure 47: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1993 to July–December 2024. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts 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).

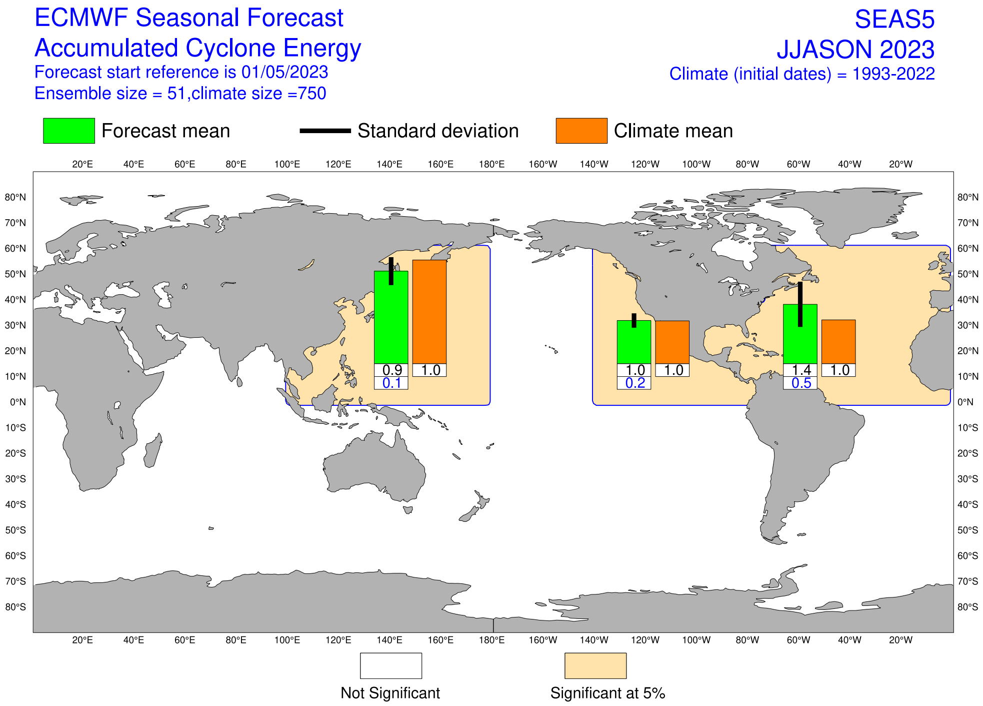


Figure 48**:** Forecast of tropical storm accumulated cyclone energy (ACE, normalized) issued in May 2023 for the six-month period June–November 2023. Green bars represent the forecast ACE in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black 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.

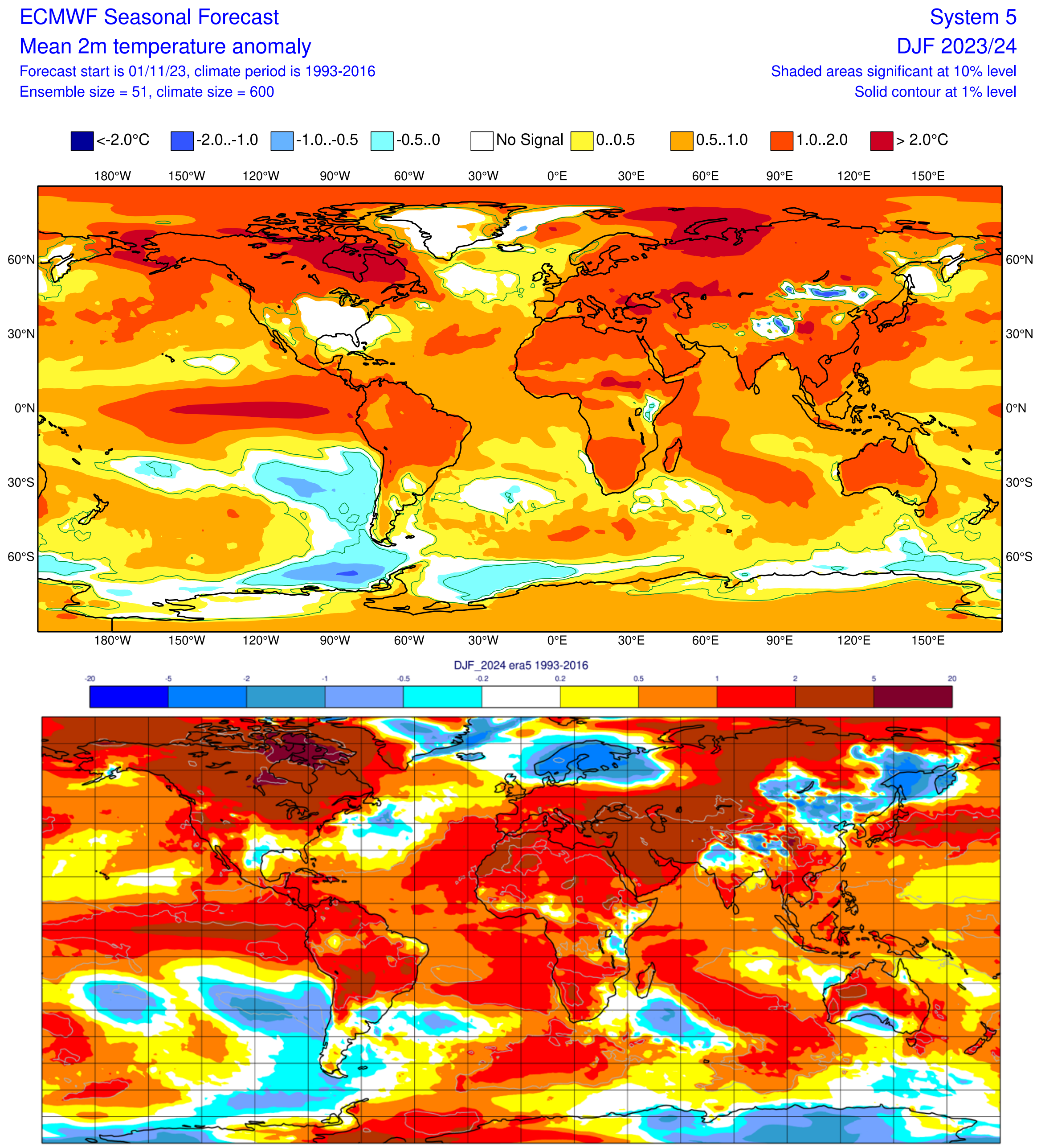


Figure 49**:** Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2023 for DJF 2023/24 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

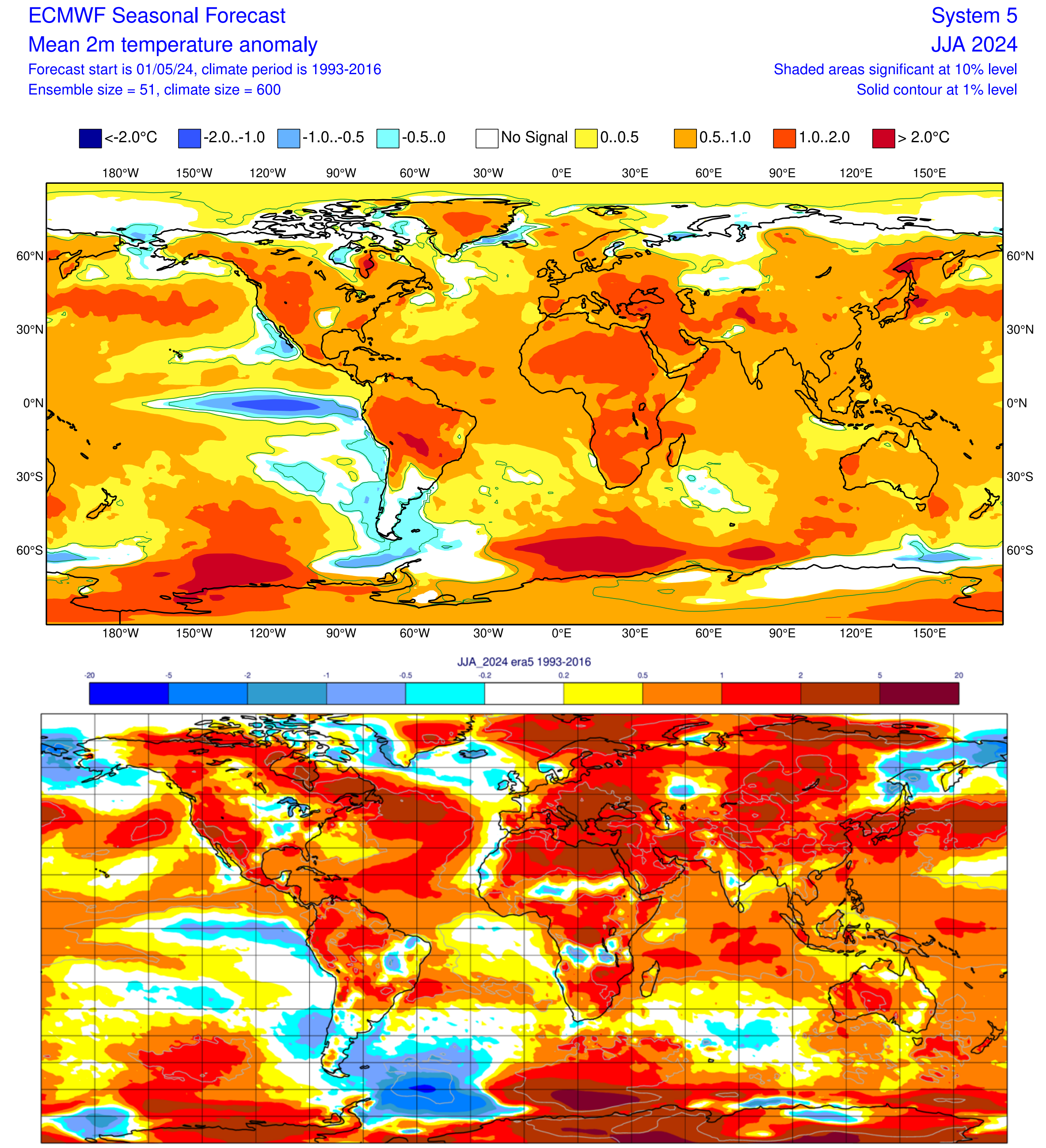

ECMWF Seasonal Forecast
Mean 2m temperature anomaly
Forecast start is 01/05/24, climate period is 1993-2016
Ensemble size = 51, climate size = 600
System 5
JJA 2024
Shaded areas significant at 10% level
Solid contour at 1% level
<-2.0°C
-2.0..-1.0
-1.0..-0.5
-0.5..0
No Signal
0..0.5
0.5..1.0
1.0..2.0
> 2.0°C
180°W 150°W 120°W 90°W 60°W 30°W 0°E 30°E 60°E 90°E 120°E 150°E
60°N 30°N 0°N 30°S 60°S
JJA_2024 era5 1993-2016
-20 -5 -2 -1 -0.5 -0.2 0.2 0.5 1 2 5 20

Figure 50: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2024 for JJA 2024 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

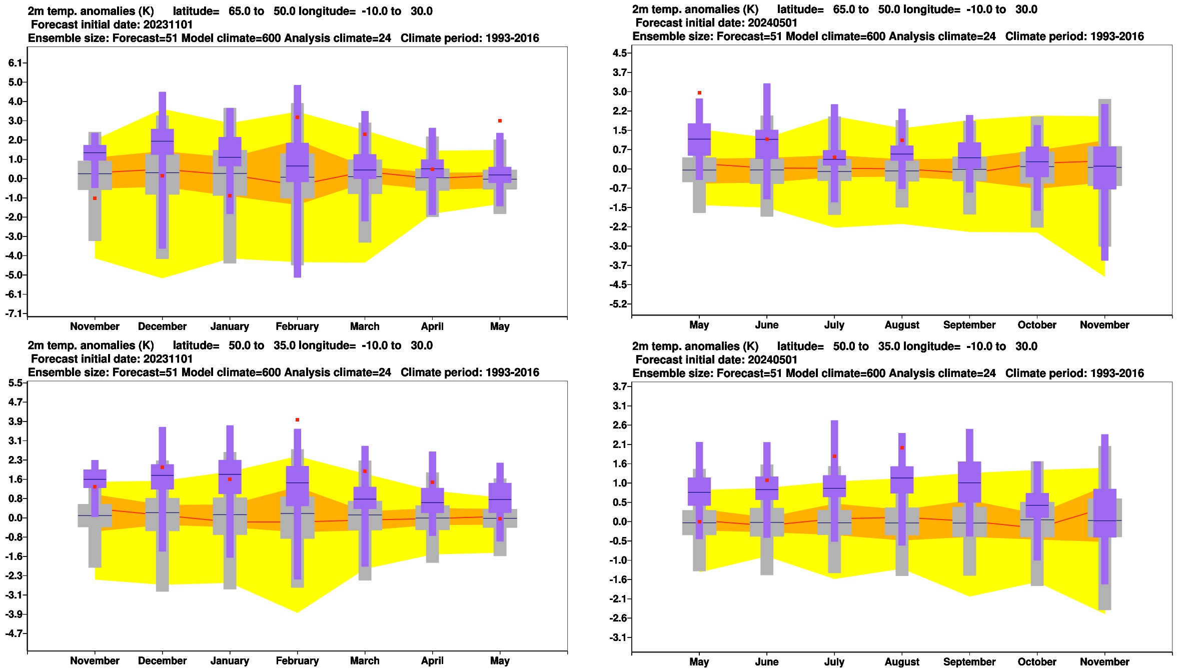
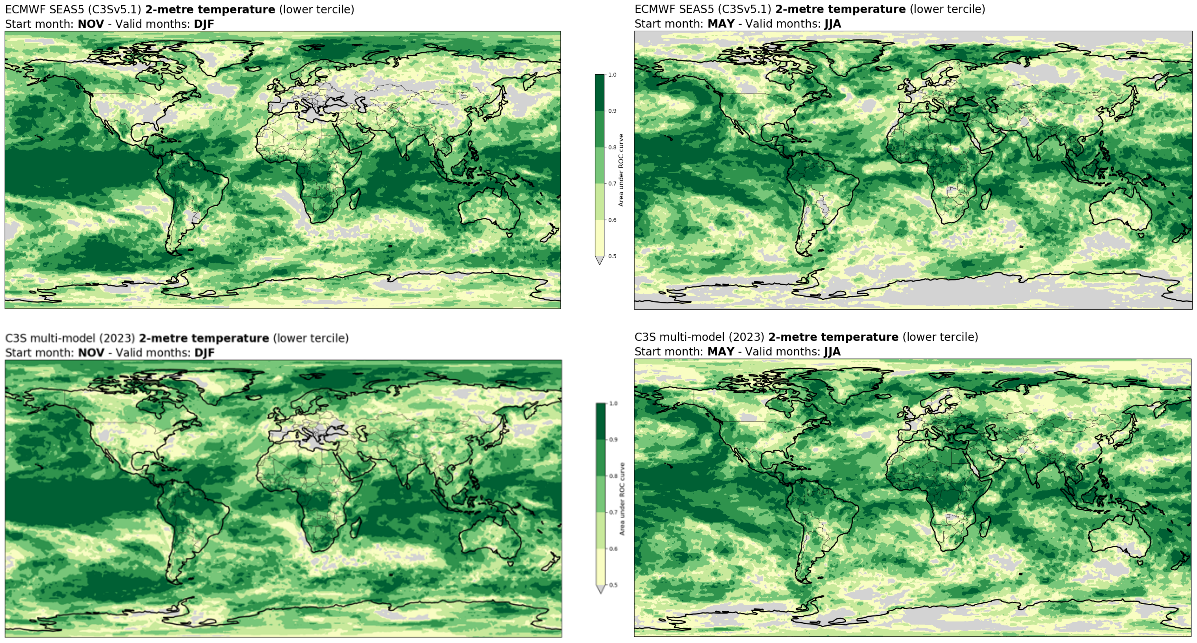


Figure 51: Verification of long-range forecasts of 2 m temperature anomalies from November 2023 for DJF 2023–24 (left panels) and from May 2024 for JJA 2024 (right panels) for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 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.



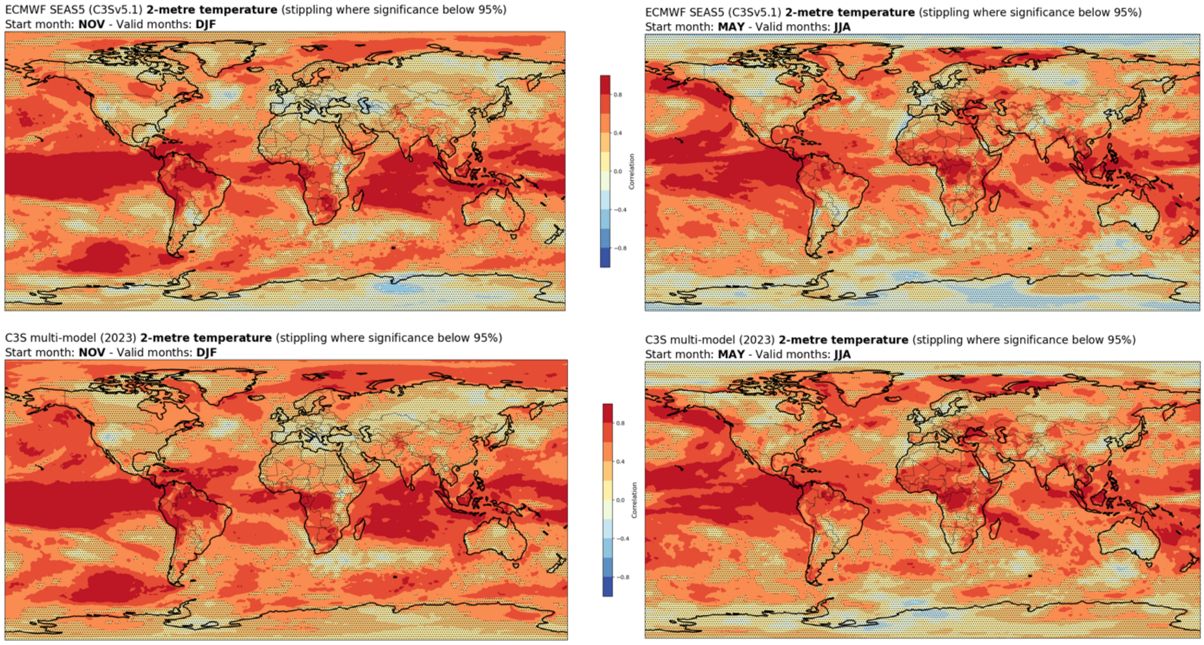
ECMWF SEAS5 (C3Sv5.1) **2-metre temperature** (lower tercile)
Start month: **NOV** - Valid months: **DJF**

ECMWF SEAS5 (C3Sv5.1) **2-metre temperature** (lower tercile)
Start month: **MAY** - Valid months: **JJA**

C3S multi-model (2023) **2-metre temperature** (lower tercile)
Start month: **NOV** - Valid months: **DJF**

C3S multi-model (2023) **2-metre temperature** (lower tercile)
Start month: **MAY** - Valid months: **JJA**

Figure 52: ROC area of seasonal forecasts of 2m temperature in DJF from forecasts initialized in November (left panels), and in JJA from forecasts initialized in May (right panels) over the period 1993-2016. Top panels: ECMWF forecasts, bottom panels: C3S multi-model forecasts. Verification is against ERA5 analyses.



ECMWF SEAS5 (C3Sv5.1) **2-metre temperature** (stippling where significance below 95%)
Start month: **NOV** - Valid months: **DJF**

ECMWF SEAS5 (C3Sv5.1) **2-metre temperature** (stippling where significance below 95%)
Start month: **MAY** - Valid months: **JJA**

C3S multi-model (2023) **2-metre temperature** (stippling where significance below 95%)
Start month: **NOV** - Valid months: **DJF**

C3S multi-model (2023) **2-metre temperature** (stippling where significance below 95%)
Start month: **MAY** - Valid months: **JJA**

Figure 53: As in Figure 52 but for linear temporal (Spearman) correlation. Stippling indicates significance below 95%.

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 (Figures 13, 14, 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 (Figures 13, 15) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

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

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

Figure 1 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 36) 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 for upper-air (Figure 6) and surface variables (Figure 22).

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 41). Figure 41 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 converted to skill via ROCAskill=2·ROCA-1 are for example shown in Figure 39.

The comparison of spread and skill (Figures 7-9) 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 (Figures 22 and 23) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figures 22 and 23). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figures 24-27), verification data are European 6-hourly SYNOP data (area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the four closest grid points, provided the difference between the model and true orography is less than 500 m. A crude quality control is applied to SYNOP data (maximum departure from the model forecast has to be less than 25 K, 20 g/kg or 15 m/s for temperature, specific humidity and wind speed respectively). 2 m temperatures are corrected for differences between model and true orography, using a crude constant lapse rate assumption provided the correction is less than 4 K amplitude (data are otherwise rejected).

A.4 Hydrological evaluation metric

The modified Kling-Gupta Efficiency (KGE') is calculated using the formulation by Kling et al. (2012), as detailed below:

$$KGE' = 1 - \sqrt{(r-1)^2 - (\beta-1)^2 - (\gamma-1)^2}$$

$$\beta = \frac{\mu_s}{\mu_o}$$

$$\gamma = \frac{CV_s}{CV_o} = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o}$$

Where:

- *KGE'* is the modified KGE-statistic (dimensionless),
- *r* is the correlation coefficient between simulated and observed runoff (dimensionless),
- β is the bias ratio (dimensionless),
- γ is the variability ratio (dimensionless),
- μ is the mean runoff in m^3/s,
- *CV* is the coefficient of variation (dimensionless),
- σ is the standard deviation of runoff in m^3/s,
- Indices *s* and *o* represent simulated and observed runoff values, respectively.

KGE', *r*, β and γ have their optimum at unity. For the variability ratio γ is calculated using CV_s/CV_o instead of σ_s/σ_o, as proposed in the original KGE formulation by Gupta et al. (2009). This adjustment prevents cross-correlation between the bias and variability ratios, which can occur, for example, when the precipitation inputs are biased.

References

Ben Bouallegue, Z., M. C. A. Clare, L. Magnusson, E. Gascon, M. Maier-Gerber, M. Janousek, M. Rodwell, F. Pinault, J. S. Dramsch, S. T. K. Lang, B. Raoult, F. Rabier, M. Chevallier, I. Sandu, P. Dueben, M. Chantry, F. Pappenberger, 2023: The rise of data-driven weather forecasting. Early online release https://arxiv.org/abs/2307.10128

Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, Journal of Hydrology, 377(1-2), pp.80-91, https://doi.org/10.1016/j.jhydrol.2009.08.003

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

Kling, H., Fuchs, M., Paulin, M. (2012). Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios, Journal of Hydrology, 424-425, pp.264-277, https://doi.org/10.1016/j.jhydrol.2012.01.011

Lang, S., D. Schepers, amd M. Rodwell, 2023: IFS upgrade brings many improvements and unifies medium-range resolutions. ECMWF Newsletter No. 176, 23-30.

Lang, S., and co-authors, 2024: AIFS – ECMWF's data-driven forecasting system. arXiv:2406.01465 [physics.ao-ph], https://doi.org/10.48550/arXiv.2406.01465

Presseau, C., D. Kornic, S. J. Peng, C. Stroud, V. Savic-Jovcic, and A. Lupu, 2024: Air-Quality Multi-Model Verification for North America 2024-01 to 2024-03. Environment and Climate Change Canada, 10p.

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.

Vitart, F., M. A. Balmaseda, L. Ferranti, and M. Fuentes, 2022: The next extended-range configuration for IFS Cycle 48r1. ECMWF Newsletter No. 173, 23-28.