Technical Memo

ECMWF
European Centre for Medium-Range Weather Forecasts

911

Evaluation of ECMWF forecasts, including the 2023 upgrade

Thomas Haiden, Martin Janousek, Frédéric Vitart, Zied Ben-Bouallegue, Fernando Prates

September 2023

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

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 48r1 in June 2023 and the meteorological impact of the upgrade. Scorecards show that the increase in ENS resolution to match that of the HRES, as well as other changes in cycle 48r1, bring a substantial increase in forecast skill. 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. The primary focus of this summary is the medium range, and specifically the forecast performance for upper-air variables. It is shown that in this respect ECMWF has a clear lead among centres. For surface variables, other centres have however partly taken the lead, especially at shorter ranges. Significant improvements of ECMWF due to recent model upgrades can be seen in a substantial reduction of large ENS 10-m wind speed errors. In the extended range, a distinct improvement in the forecasting of 2m temperature anomalies can be seen in week 2 but only marginal improvements in weeks 3 and 4. On the seasonal timescale, the change from La Nina to El Nino conditions was predicted well, and the anomalously warm northern hemispheric summer season 2023 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 June 2023, and how they affected the skill of the forecasts. An important part of these changes is the increase in resolution of the ensemble forecast to match the high-resolution run, which brings clear improvements in forecast skill. 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. In this respect ECMWF has a clear lead among centres. For surface variables on the other hand, some other centres have however overtaken ECMWF at shorter ranges. A large improvement coming from recent model upgrades can be seen in a reduction of large 10-m wind speed errors in the ensemble forecast. In the extended range, the forecast of 2m temperature has improved, but not much change is seen for weeks 3 and 4. In the seasonal forecast, the change from El Nino to La Nina conditions was predicted well, and the very warm northern hemispheric summer season 2023 was present in the forecast.

1 Introduction

This report presents a summary of verification results from ECMWF's operational forecasting system including the medium-range, extended-range and seasonal forecast, as well as CAMS and, to provide a reference for the medium-range scores, forecasts from ERA5.

The most recent change to the ECMWF forecasting system (IFS Cycle 48r1, on 27 June 2023) 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, 902). 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' (monthly)
- 'Long Range' (seasonal)

2 Changes to the ECMWF forecasting system

On 27 June 2023, ECMWF performed a major upgrade of the Integrated Forecasting System (IFS). IFS Cycle 48r1 includes an increase in ENS horizontal resolution so that it now matches that of the HRES (TCo1279, 9 km). The number of ensemble members in the extended range is increased from 51 to 101, and it is run daily instead of twice weekly. It runs from day 0 at a constant spatial resolution and is not an extension of the medium-range ensemble anymore. Changes to observation usage, data assimilation, and the model include, among other improvements, higher inner-loop resolution in the data assimilation system, the assimilation of surface-sensitive microwave imager channels over land and cold ocean surfaces, and a multi-layer snow scheme in the forecast model.

2.1 Model and data assimilation changes

Besides the resolution upgrade of the ENS, several changes in the data assimilation and model formulation have been included in cycle 48r1. They are listed below.

Changes in observation usage

- Improved observation pre-processing
- Assimilation of microwave imagers over land surfaces
    - 89 GHz, 150/166 GHz channels of GMI, SSMIS + GMI 183 GHz over land
    - Addition of 37 GHz channels, addition of AMSR2; improved bias correction, QC, and error models

Changes to the assimilation system

- Increase of HRES 4D-Var inner loop resolution to T_L511
- Reduced thinning of ASCAT L2 products
- Various optimisations for hyperspectral IR sounders
    - Unified VarBC setup for IR sounders
    - Allowing usage of all pixels from IASI
    - Aerosol type classification in IR data
    - Update on the IR trace gas detection
- Upgrade RTTOV to v13
    - Latest version of RTTOV: technical upgrade + additional capabilities to prepare for future changes
    - Microwave gas optical depth coefficient file upgrade, using new predictors (v13)
    - Major scientific upgrade of cloud and precipitation microphysics in RTTOV-SCATT
- ATMS over snow, Lambertian, slant-path
    - Activation of ATMS humidity channels over snow
    - ATMS Lambertian surface reflection over snow and sea-ice
    - Slant-path interpolation for selected MW sensors assimilated in the all-sky system
- Improved treatment of surface-sensitive channels in all-sky
    - Assimilation poleward of 60 degrees over land and ocean; relying on new sea-ice detection
    - Improved treatment of mixed land-water and water-sea-ice scenes

Changes to the model

- Improved water and energy conservation (dynamics and physics)
- Radiatively interactive prognostic ozone using new Hybrid Linear Ozone (HLO) scheme
- Multi-layer snow scheme
- New precipitation category - freezing drizzle
- Revised climate fields – improved orographic fields for atmospheric drag and water related representation (i.e. glacier mask, land-sea mask, lake cover, lake depth)
- Revised computation of Semi-Lagrangian advection departure points
- New model top sponge layer formulation and semi-Lagrangian vertical filter
- Revised SPPT, removed cloud saturation adjustment from tendency perturbations

2.2 Meteorological impact of the new cycle

The following summary is a slightly abridged version of the description in Lang et al. (2023). The scorecard summarising the ENS score changes is shown in Figure 1. Most ENS scores of surface variables, such as 2 m temperature, 10 m wind and total precipitation, are markedly improved, in the range of 2% to 6%. Most upper-air variables are improved as well, by around 1% to 3%. Stratospheric winds are improved, but some degradations of stratospheric temperatures due to increased biases can be observed. There are also improvements in scores over the Arctic and Antarctic, partly associated with increased spread generated by the multi-layer snow scheme. The impact of this scheme on 2 m temperatures in snow-prone regions includes reduced biases (Figure *3*), an improved daily cycle (Arduini et al., 2019), and reduced snow depth forecast errors.

The ensemble spread of upper-air variables (ensemble standard deviation, not shown) is reduced in the mid- latitudes (by around 1% to 2%) but mainly increased in the tropics (by around 2% to 3%), and the ensemble spread of surface variables is increased (by around 2% to 6%).

Cycle 48r1 improves ENS tropical cyclone track and intensity forecasts, with position errors reduced by up to 10% (Figure 5) and core pressure errors reduced by around 20%. The reduced track errors are mainly associated with a reduced slow-propagation bias of the forecast model. The intensity forecast is improved because the higher horizontal resolution allows for a better representation of the strong horizontal gradients associated with intense systems like tropical cyclones. This is also reflected in the strongly increased intensity spread of the ENS. The track spread, on the other hand, is very similar to that of the previous cycle (not shown).

The scores of the HRES forecast (now at the same spatial resolution as the ENS) show overall improvements as well but to a lesser extent than the ENS (Figure 2). Upper-air tropospheric scores are improved by around 1% to 3% in the northern hemisphere and in the tropics. In line with the ENS, stratospheric winds are improved, but some degradations of stratospheric geopotential can be observed due to increased biases. These impacts are mainly related to the new HLO scheme. Upper-level Arctic and Antarctic scores are improved, in line with the ENS scores. Surface winds in the tropics appear degraded when verified against ECMWF's analysis. This apparent degradation is caused by the large increase of scatterometer observations and the associated decorrelation of analysis and forecast errors. Verification against observations does not show the degradation. HRES northern and southern hemisphere 2 m temperature scores also show some degradation, which comes from increased (more realistic) analysis and forecast activity generated by the multi-layer snow scheme. Tropical cyclone track and intensity forecast scores in HRES are very similar to those for the previous cycle.

The forecast model changes result in small improvements of the weekly mean extended-range scores. Mainly, however, it is the increased ensemble size of the extended-range system, as well as the increased run frequency (daily instead of twice weekly), which lead to substantial skill improvements (Vitart et al., 2022), as seen in Figure 4.

3 Verification of upper-air medium-range forecasts

3.1 ECMWF scores

Figure 6 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 a new high point, mostly due to improved ACC values in the summer half-year, where this score is usually lowest. 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 couple of years.

A complementary measure of performance is the root mean square (RMS) error of the forecast. Figure 7 shows RMS errors for both extratropical hemispheres of the six-day forecast and the persistence forecast. Like anomaly correlation, in the northern hemisphere the 12-month running mean RMS error of the six-day forecast reached its best (=lowest) value during the last year of the time series. In the southern hemisphere, values have been consistently low since 2020 but not improved further.

Figure 8 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 after 2019.

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 9. 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 10, top panel) shows that ECMWF is maintaining a substantial lead which has, however, slightly decreased relative to some of the centres. The centre and bottom panels in Figure 10 show HRES stratospheric temperature RMSE skill relative to ERA5 for a range of stratospheric levels. There has been some drop in the extratropics, and in the tropics at lower stratospheric levels, and improvement at higher levels in the tropics.

The trend in ENS performance is illustrated in Figure 11, 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 still has not quite reached the high value that was driven by high predictability of 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. In recent years 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 12. 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 in week two of the forecast. 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. For 850 hPa temperature, there is a general slight underdispersion. As for 500 hPa geopotential, spread-error mismatch for 2023 lies in between those of 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. Figure 13 and Figure 14 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, in the medium range. The stars in the plots mark the average values, corresponding to Figure 12, and ideally should lie on the diagonal and as close to the lower left corner as possible. In this regard ECMWF overall performs best among 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 the MetOffice has a 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 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. Note that this represents a challenging benchmark since we compare 50 ensemble members against a continuous distribution. Figure 15 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, interannual variability is larger, making it less clear whether there is a 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 16. Especially for the 850 hPa errors there is a clear signal of improvement from model cycle 47r3.

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, following agreed standards of verification.

Figure 17 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, however the gap between ECMWF and other centres is decreasing somewhat.

WMO-exchanged scores also include verification against radiosondes. Figure 18 (Europe), and Figure 19 (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 20 (verification against analyses) and Figure 21 (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. In the tropics, verification against analyses (Figure 20) 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 21), smoothness is less of an advantage, and 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 22 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. There was a significant but transient drop in skill in 2022 as a result of the use of non-optimal background error covariances in the CAMS data assimilation. After the issue was identified and corrected, skill went back up to previous levels. 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.

3.4 Data-driven forecasts

In 2023, ECMWF has started to run, in addition to the IFS, data-driven (machine-learning, ML) forecasts using ERA5 for training and HRES initial conditions. As these forecasts are still experimental, no scores are included in this report which generally focuses on operational performance. Once they become part of the operational suite, they will be included. A comprehensive evaluation of one specific ML forecast run by ECMWF versus the IFS can be found in Bouallegue et al (2023).

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 23. 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 SEEPS deterministic precipitation forecast skill has increased considerably in recent years, although there has been a drop in 2022. Since there was no model cycle change in 2022, this drop reflects interannual changes in predictability. Also, the ERA5 reference forecast (black line in Figure 23, top left panel) shows both the multi-year increase, and the drop in 2022, such that the difference between operational and ERA5 scores and HRES (upper right panel in Figure 23) does not show a net increase in the last couple of years.

The probabilistic precipitation headline score CRPSS (lower left panel in Figure 23) shows a slight decrease in recent years. It should be noted that in addition to the difference of 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. More detailed analysis shows that the discrimination ability of the ENS precipitation has in fact increased, and this leads to flat (i.e. non-decreasing) DSS, but the reliability has decreased somewhat so that as a result the CRPSS has decreased as well.

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 24). ECMWF's probabilistic precipitation forecasts are more skilful than those of other centres from day 3 onwards. It should be noted that relative to previous years especially the UKMO precipitation forecasts have gained on those of ECMWF in the short 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 25 to Figure 28. 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 25), both daytime and night-time errors have not changed significantly compared to last year. For 2 m dewpoint (Figure 26), the negative bias during daytime in summer has increased in recent years, and so has the error standard deviation. For total cloud cover (Figure 27) 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 and is already being addressed in model cycle 48r1 (Forbes et al., 2021). It is worth noting that the shortwave radiation has actually been improved by the change (see Figure 30, 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 is comparable to the previous year (Figure 28. The night-time positive wind speed bias due to insufficient calming of the wind in the presence of surface inversions 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 29 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. Note that the drop in the second half of 2020 not tied to a particular model cycle but rather 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 winter. For the surface parameters (verification against SYNOP), 10 m wind speed forecast skill is slightly higher than before, and the negative effect of 47r3 on total cloud cover can be seen as well.

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 30 shows how the 5-day forecast of the TOA net shortwave radiation has improved over time. ERA5 is included for comparison, showing that the changes in the extratropics in the last couple of years are mainly due to

interannual variability, while in the tropics an improvement is seen both in absolute terms and relative to ERA5.

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 31 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. There has been no change in 2022, consistent with the absence of model upgrades in that year.

An analogous measure of the skill in predicting large 10 m wind speed errors in the ENS is shown in Figure 32. Here, a threshold of 4 m s^{-1} for the CRPS is used, to obtain similar fractions as for temperature. Over the last 3 years there has been a sizeable reduction in the number of large 10m wind speed errors, such that their fraction went down from a little less than 4% to 3.3%, which represents a relative improvement of about 15%.

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 33. While errors in 10 m wind speed have slightly increased in the analysis and short-range forecast in recent years, wave height forecast scores have remained stable (see also Figure 34). The most likely reason is that due to the increased use of satellite data in the initialization of the wave forecast, the analysis draws not as close to the in-situ buoy observations as it did in the past.

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 35), ECMWF generally 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. The same can be seen in the scores for the tropics (Figure 36).

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

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 (ROC) 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 37 (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 dropped somewhat compared to its highest values so far in 2021-22. For precipitation, there has been a stronger decrease which is mostly due to interannual variability as there was no model change in 2022.

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 37 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, recent forecast skill changes are largely similar to those of the EFI ROC area skill, confirming that they are not particular to the specific score used.

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 38. 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.

In terms of absolute skill, the HRES position error (top panel, Figure 38) has been smaller in 2023 than ever before, both for day 3 and day 5 of the forecast. Comparison with ERA5 shows that the drop relative to the previous year largely reflects increased predictability from 2022 to 2023. Similarly, changes in intensity and speed errors are mirrored by the ERA5. It is worth

noting that 48r1 brings a substantial improvement in intensity for the ENS (Figure 5) but as it has been implemented only in June 2023, the operational verification plots cannot show this yet.

The bottom panel of Figure 38 shows the spread and error of ensemble forecasts of tropical cyclone position. For reference, the HRES error is also shown. As for the HRES, ENS position errors are smaller than ever before. The reduction compared to the year before is especially clear for the ENS at day 5, where it changed from 280 to 250 km, i.e. it dropped by about 10%, keeping in mind that ERA5 shows it too and there was no upgrade in 2022, so due to interannual variability. Nevertheless, for forecast users it means that position forecasts were more accurate during that year. The slight underdispersion for ENS position is similar in magnitude to the year before.

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 39. 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. Both for the reliability of strike probability and ROC skill, 2023 lies in between the two previous years, while in terms of modified ROC skill there has been a small improvement.

6 Extended-range and seasonal forecasts

6.1 Extended-range forecast verification statistics and performance

Figure 40 shows the probabilistic performance of the extended-range 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.

In summer 2022, forecast skill for warm anomalies has reached new high points, especially in week 2, but also in weeks 3 to 4. In week 2, the skill above persistence was quite high as well. The skill of predicting cold anomalies in week 2 in the winter 2022-23 has been the 2nd highest so far, while the skill for weeks 3 to 4 has been close to average. However, the skill above persistence for cold anomalies in weeks 3 to 4 has been high in 2022-23. In the longer term, a clear trend of increasing skill over the last decade can be seen for warm anomalies in summer

in week 2, and to a lesser extent, weeks 3 to 4. For cold anomalies in winter, there is a weak trend in week 2 but no statistically significant trend in weeks 3 to 4.

Because of the low signal-to-noise ratio of real-time forecast verification in the extended range (Figure 40), re-forecasts are a useful additional resource for documenting trends in skill. Figure 41 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. There has been little change from 2022 to 2023, consistent with the fact that no new model cycle was implemented in 2022.

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 presented by Ferranti et al. (2018).

Comprehensive verification for the monthly forecasts is available on the ECMWF website at:

https://charts.ecmwf.int

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 forecasting system is based on IFS cycle 43r1. While the ocean model and initial conditions are the same as used in the extended range including resolution (TCo319), SEAS5 has 91 levels, whereas the extended range 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

6.2.2 The 2022-23 El Niño forecasts

The year 2022 was characterized by weak La Nina conditions for most of the year, and a transition to El Nino conditions starting in autumn. From early to mid-2022, SEAS5 forecasts showed a premature signal for this transition but from autumn 2022 onwards, the forecast was very good, as well as sharp in the sense of small spread (Figure 42, left column). The C3S multi-model ensemble (Figure 42, right column), due to its naturally larger spread, better covered the observed evolution in 2022 but showed quite large spread for the transition and therefore gave a less precise timing of the actual onset of El Nino. 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 that was beneficial for C3S earlier in 2022, had an adverse effect during the time of transition (late 2022 and early 2023).

6.2.3 Tropical storm predictions from the seasonal forecasts

The 2022 Atlantic hurricane season had a total of 14 named storms (compared to 21 in the previous year), including 8 hurricanes and 2 major hurricanes. The accumulated cyclone energy index (ACE) was about 90% of the past 10-year (2012-2021) climate average (Figure 43) which makes it an active season in terms of the number of tropical storms (climate average is about 12) but less active than average for ACE. Seasonal tropical storm predictions from SEAS5 indicated correctly a higher number of tropical storms (16 +/- 5) over the Atlantic but overestimated the ACE (1.3 +/- 0.5 instead of 0.90). Subsequent forecasts, issued in July and August, also predicted an above average intensity of the tropical cyclone season. Other seasonal forecasts overestimated the Atlantic tropical cyclone activity in 2022 as well, see https://en.wikipedia.org/wiki/2022_Atlantic_hurricane_season). The reason for this overestimation is not understood since the usual hurricane season predictors (ENSO, local SSTs) were all in favour of a more intense hurricane season.

Figure 44 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 2012-2021 climate average). The 2022 western Pacific typhoon season was a below-average season producing 25 storms, 10 typhoons, with an ACE about 30% below average, which is consistent with the SEAS5 forecast. The 2022 eastern North Pacific hurricane season was a near-normal active season with an ACE close to climatology, consistent with the SEAS5 prediction. Overall, SEAS5 tropical cyclone activity forecasts issued on 1st May 2022 verified well over the two Pacific ocean basins.

For 2023, SEAS5 predicts an above-normal season over the Atlantic, a slightly below-average season over the western North Pacific and normal season over the east Pacific.

6.2.4 Extratropical seasonal forecasts

The seasonal forecast of temperature anomalies for DJF 2022/23 was quite good in terms of large-scale patterns (Figure 45). The tropical Pacific was still in a weak La Nina state, and the canonical response of the Pacific North-American (PNA) pattern is seen both in the forecast and observations. Over the North American continent, the observed cold-warm distribution was

however somewhat different from the predicted, with the cold anomaly extending further south towards California, and a more extensive cold anomaly in Arctic Canada and parts of Greenland. In the Euro-Atlantic sector, the strong positive anomaly over northern Europe was indicated, although in the analysis it stretched further down (with high magnitude) towards the Mediterranean and norther Africa. A cold anomaly in north-eastern Siberia was missed. Patterns in tropical and subtropical areas were generally well predicted in terms of the sign of the anomaly.

Predicted summer 2m temperature anomalies (Figure 46) indicated mostly positive anomalies over the northern hemisphere, in accordance with observations. Very warm conditions in much of Europe, Asia, and North Americe were well 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 strongly positive signal in the north Atlantic region that was not forecast. In the southern hemisphere, the general longitudinal patterns were qualitatively indicated but the magnitude and spatial extent of the cold anomalies (like the one in the southern Pacific) was underestimated. Overall, the JJA 2023 seasonal forecast was however rather good.

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 2022-23 and summer 2023 are shown in Figure 47. 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 main cold anomalies that were missed were the cold December in Northern Europe at two months lead time, and the cold May in Southern Europe at one month lead time. An incorrect sign of the anomaly on that timescale suggests that it developed as a result of nonlinear flow interactions beyond the limit of predictability, rather than teleconnections between the tropics and extratropics, which would be represented in the model to some degree. All verifying anomalies fell however within the range spanned by the ENS.

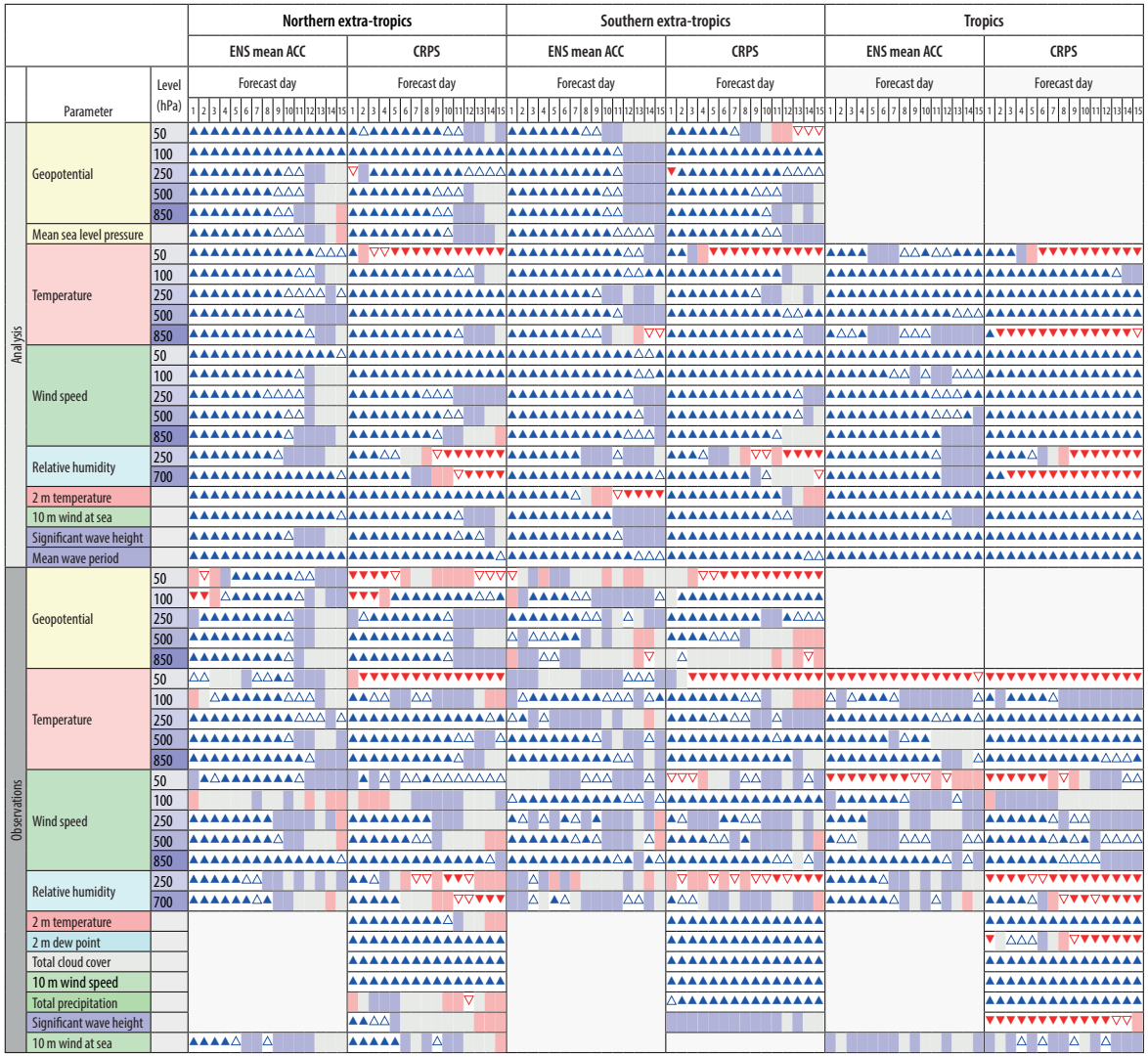
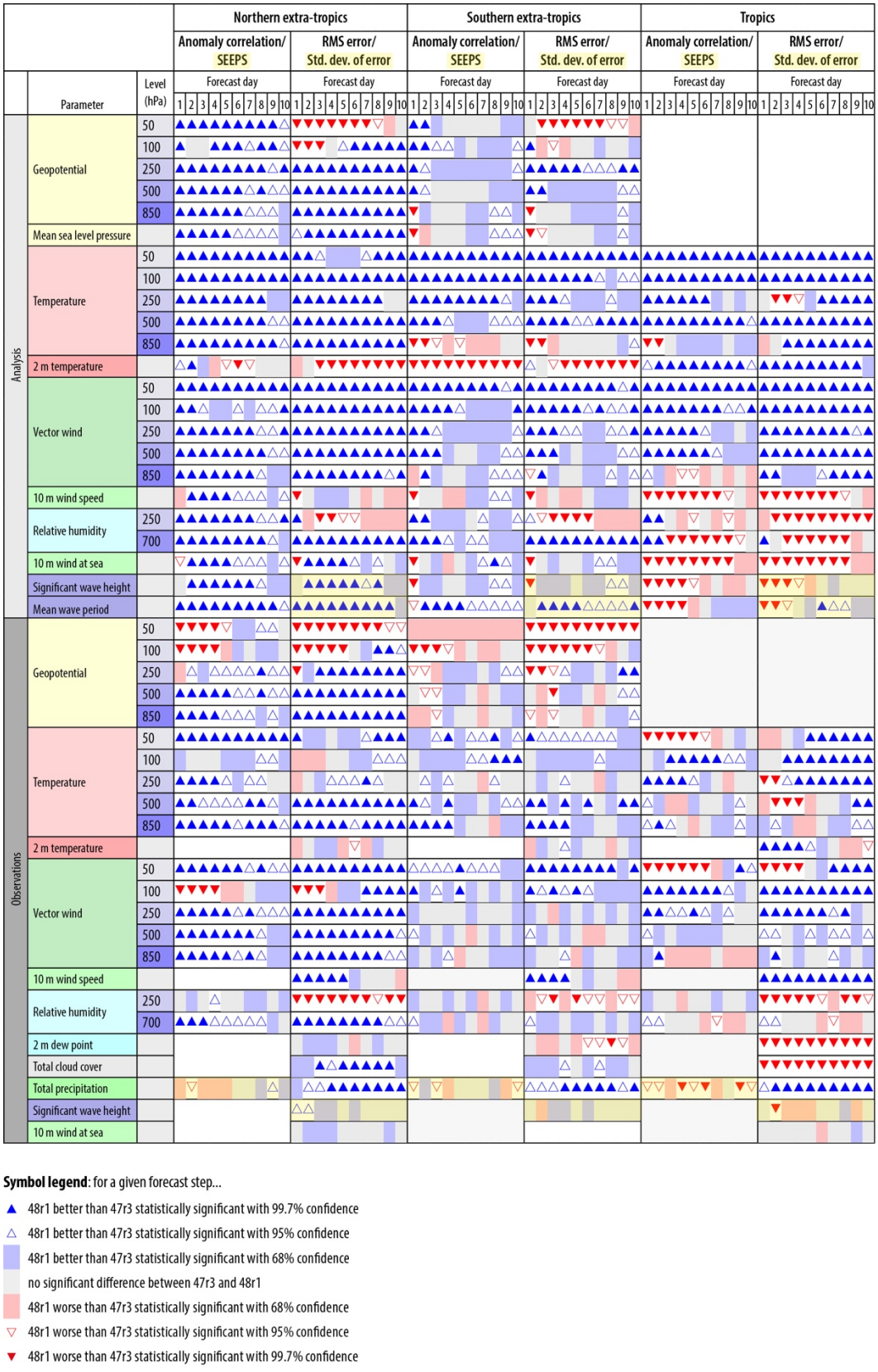


Figure 1: Summary ENS score card for IFS Cycle 48r1. Score card for ENS cycle 48r1 versus cycle 47r3 verified by the respective analyses and observations at 00 and 12 UTC for about 550 forecast runs in the period June 2020 to April 2023. From Lang et al. (2023).



Symbol legend: for a given forecast step...

- ▲ 48r1 better than 47r3 statistically significant with 99.7% confidence
- △ 48r1 better than 47r3 statistically significant with 95% confidence
- 48r1 better than 47r3 statistically significant with 68% confidence
- no significant difference between 47r3 and 48r1
- 48r1 worse than 47r3 statistically significant with 68% confidence
- ▽ 48r1 worse than 47r3 statistically significant with 95% confidence
- ▼ 48r1 worse than 47r3 statistically significant with 99.7% confidence

Figure 2: Summary HRES score card for IFS Cycle 48r1. Score card for ENS cycle 48r1 versus cycle 47r3 verified by the respective analyses and observations at 00 and 12 UTC for about 1200 forecast runs in the period June 2020 to April 2023. Yellow colouring indicates that symbols refer to the second score indicated at the top of the column. From Lang et al. (2023).

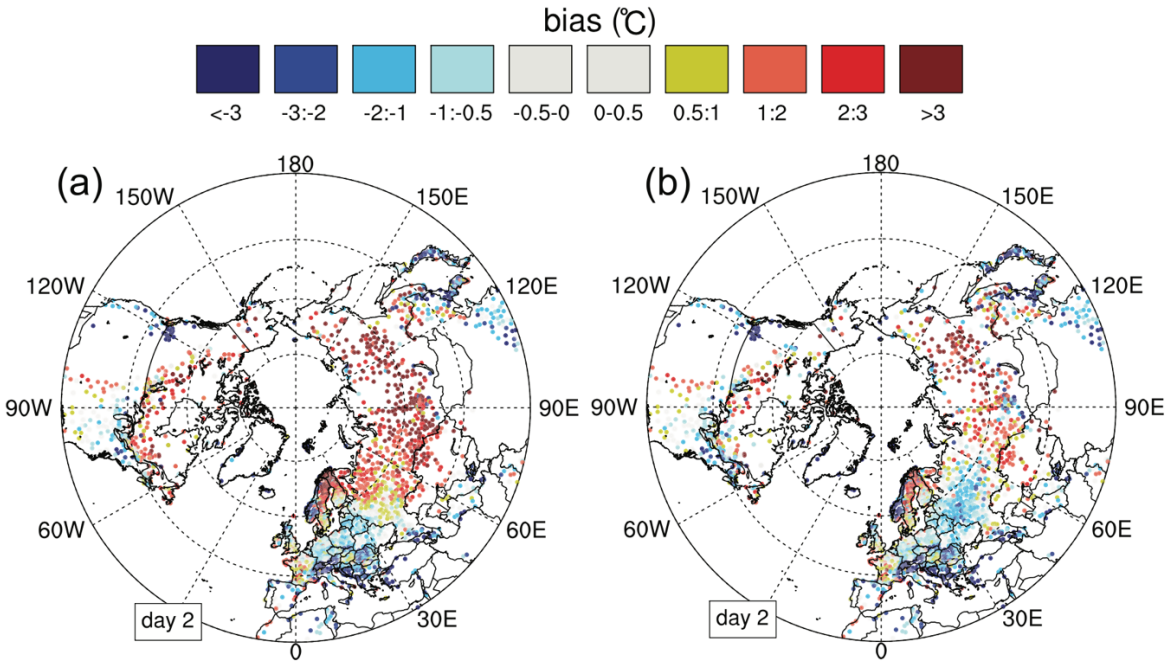


Figure 3: Bias of the daily minimum 2-m temperature of forecasts in DJF 2016/2017 using the (a) single-layer and (b) the multilayer snow scheme, at a lead time of 2 days (Arduini et al., 2019).

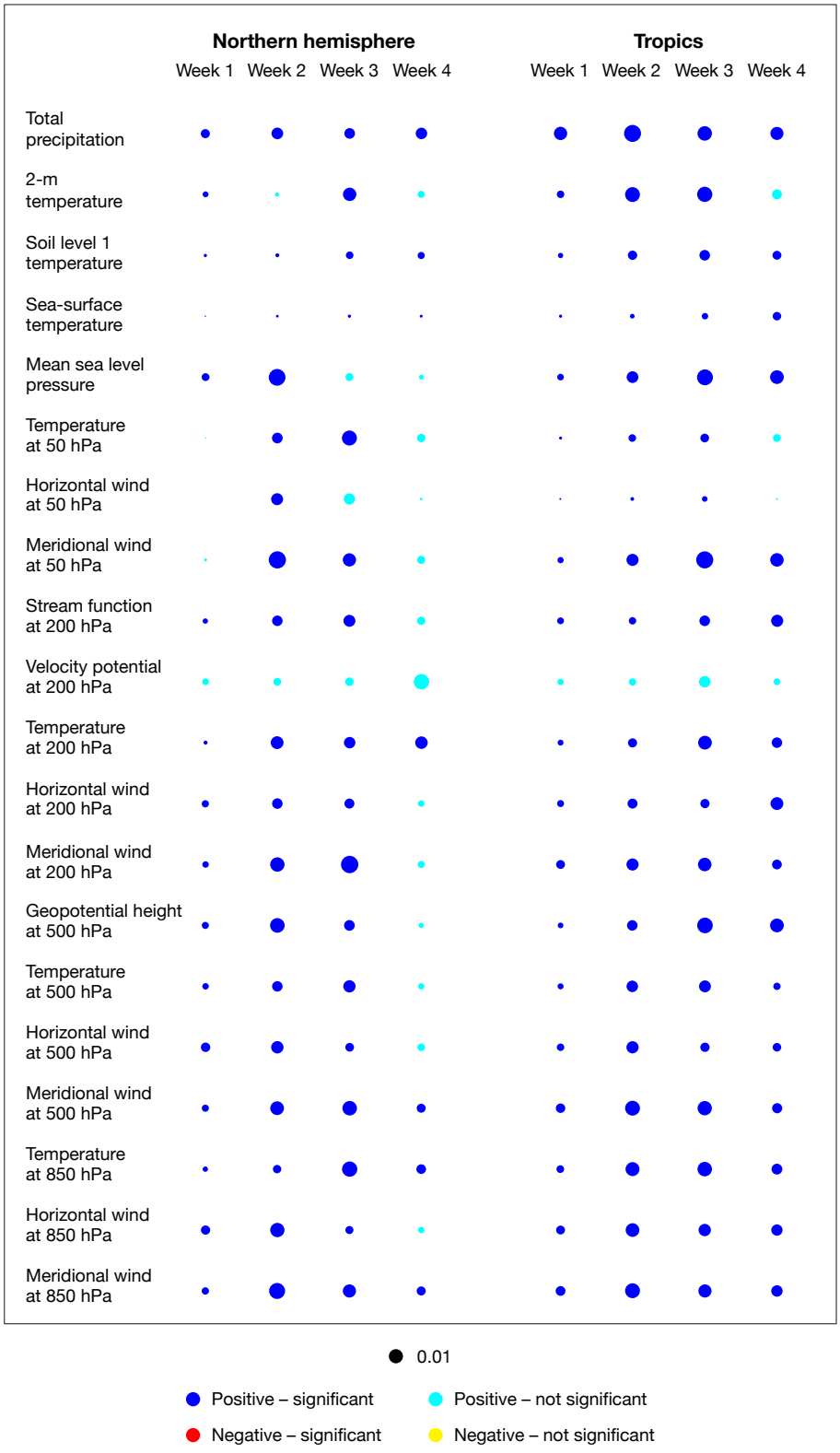


Figure 4: Scorecard showing the difference in the Continuous Ranked Probability Skill Score (CRPSS) between the 101- and 51-member ensembles for 20 variables, four lead times (weeks 1 to 4) and two regions (northern hemisphere on the left and tropics on the right). The blue (red) and cyan (yellow) colours 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).

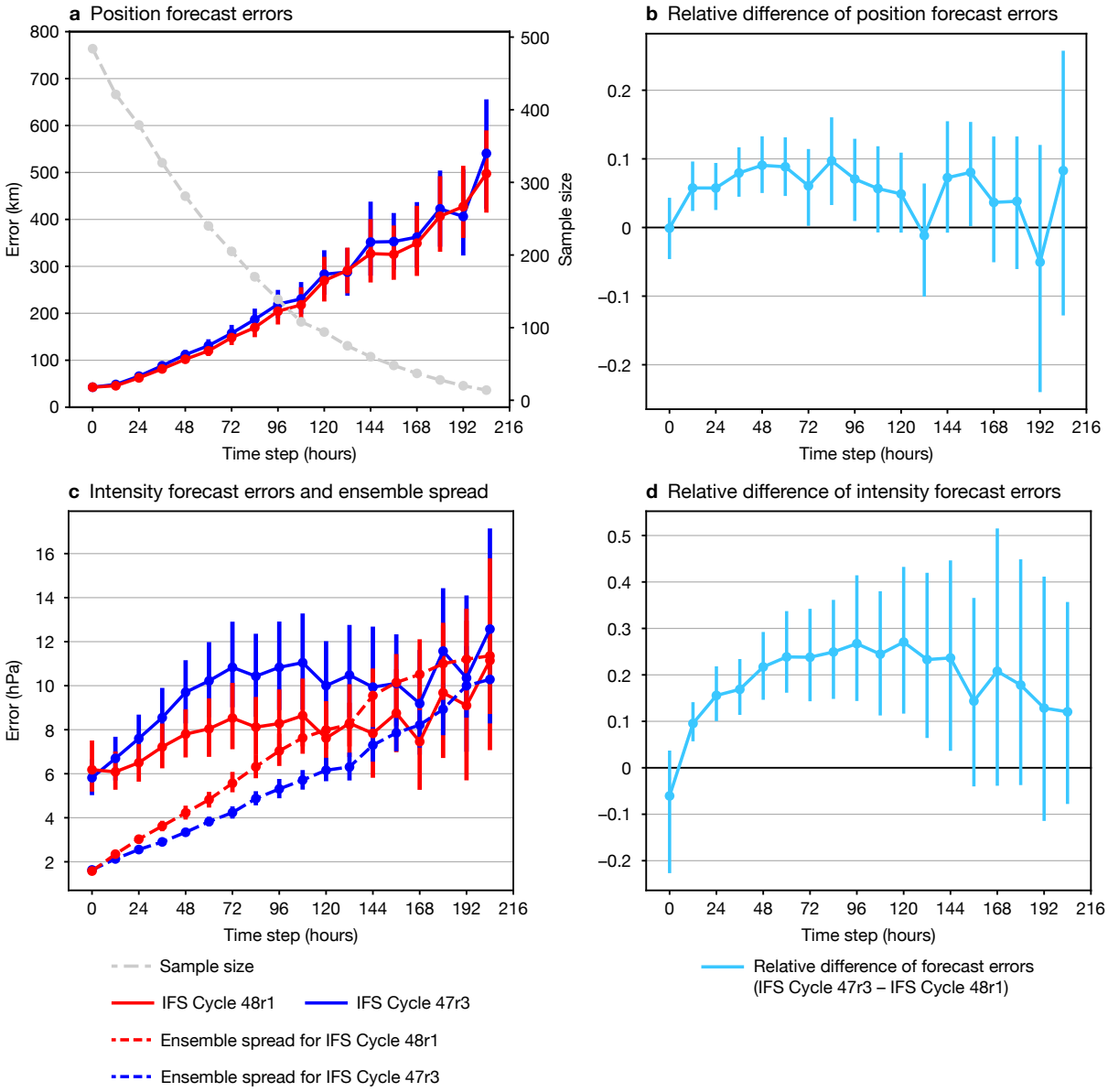


Figure 5: ENS tropical cyclone position and intensity forecast errors in IFS Cycle 48r1 compared to IFS Cycle 47r3, showing (a) position forecast errors of Cycle 48r1 and Cycle 47r3, (b) the relative difference of position forecast errors, with positive values showing improved forecast errors for Cycle 48r1, (c) intensity (central pressure) forecast errors (solid lines) and ensemble spread (dashed lines) of Cycle 48r1 and Cycle 47r3, and (d) the relative difference of intensity forecast errors, with positive values showing improved forecast errors for Cycle 48r1. The vertical error bars represent 95% confidence intervals. From Lang et al. (2023).

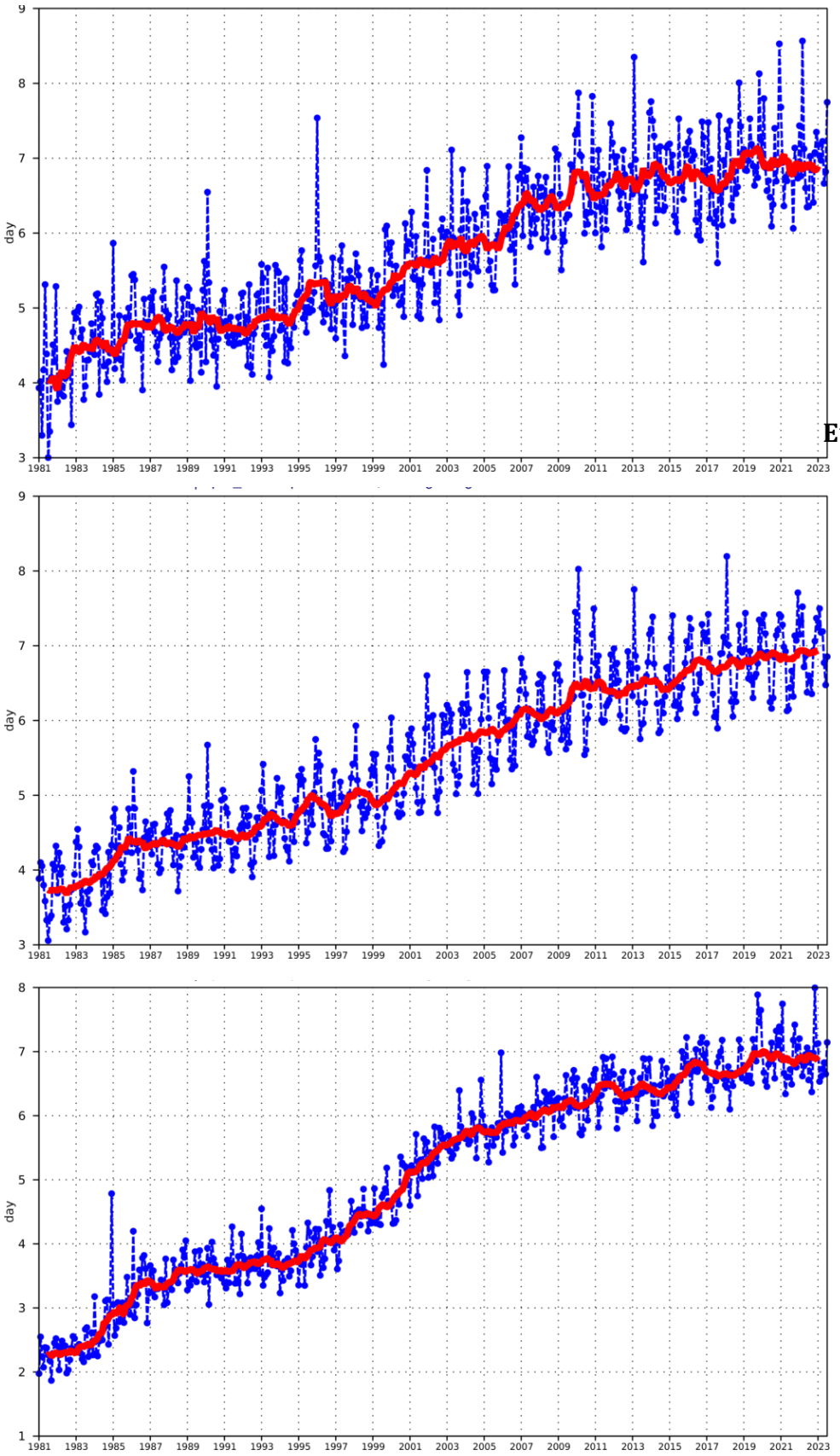


Figure 6: Primary headline score for the high-resolution forecasts. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom).

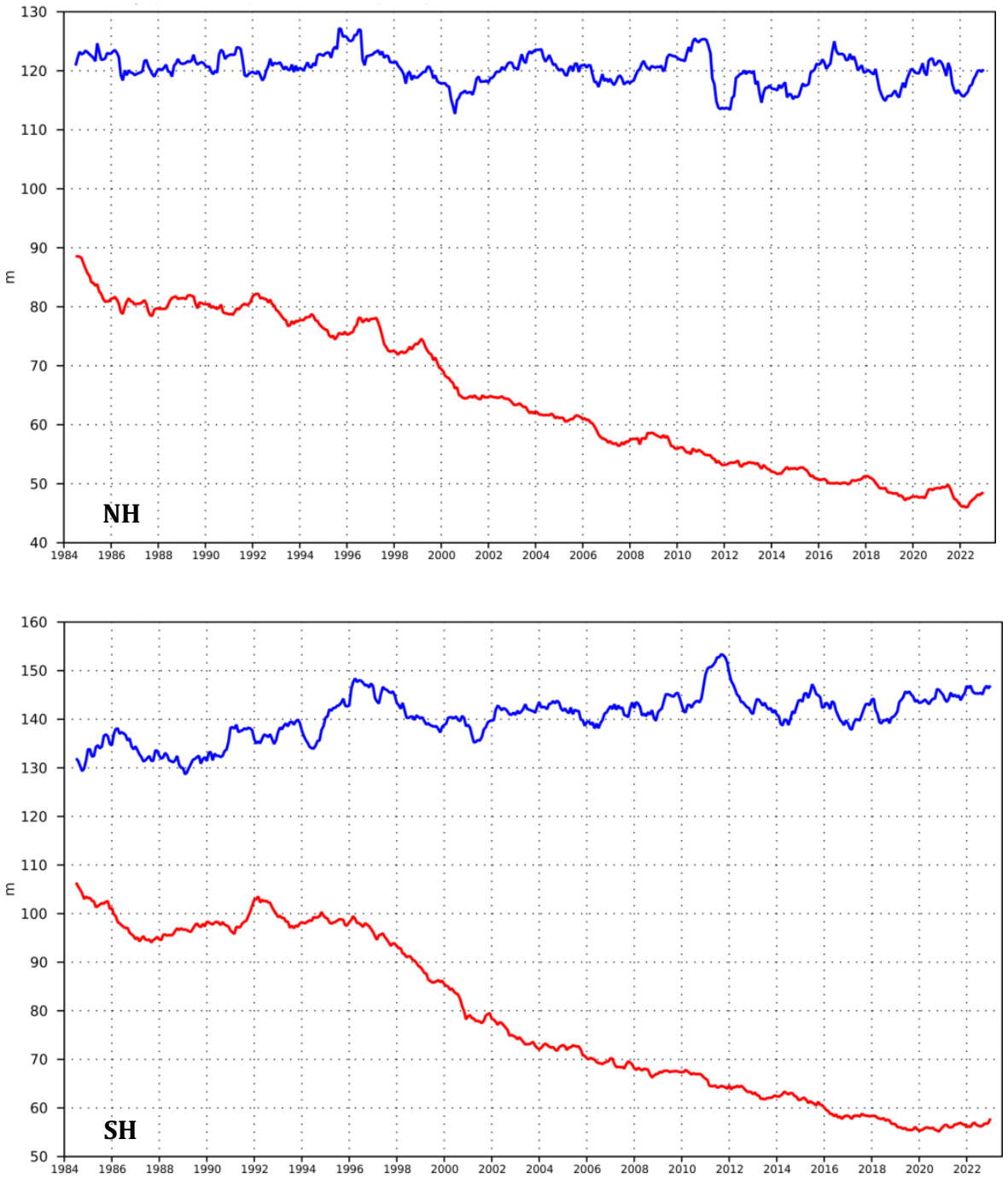


Figure 7: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2022–July 2023. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

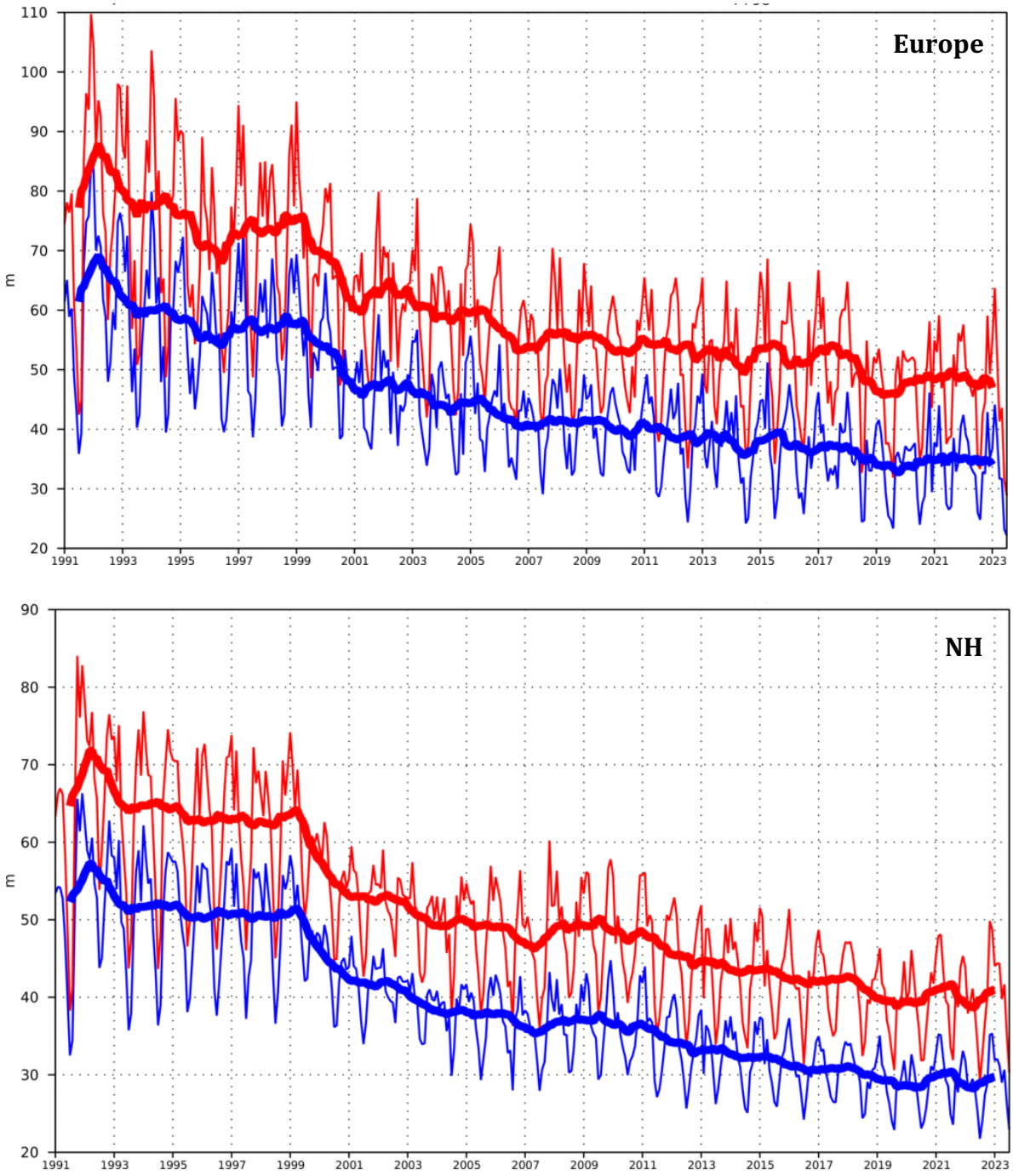


Figure 8**:** A measure of inconsistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores are also shown (in bold).

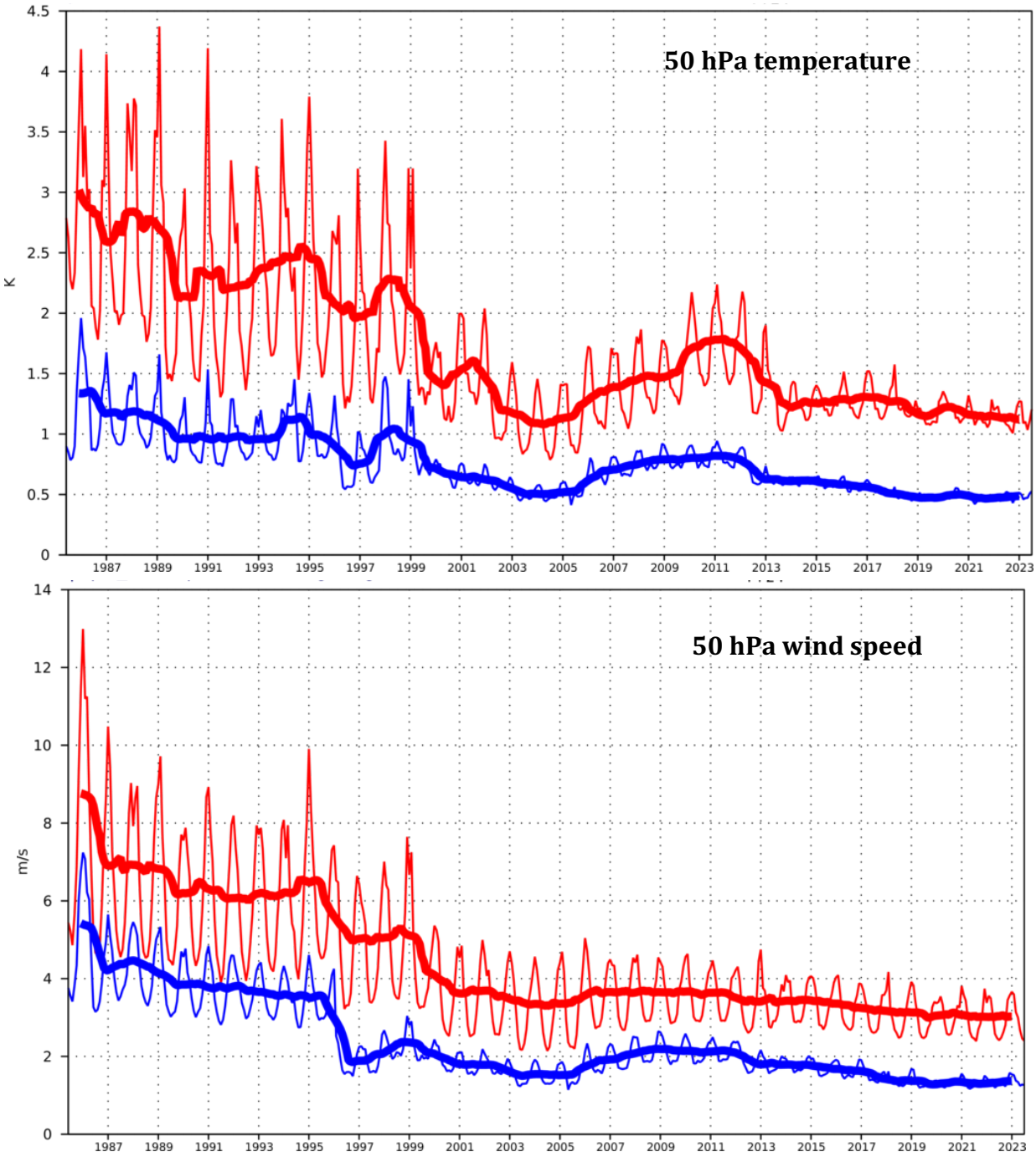


Figure 9**:** Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

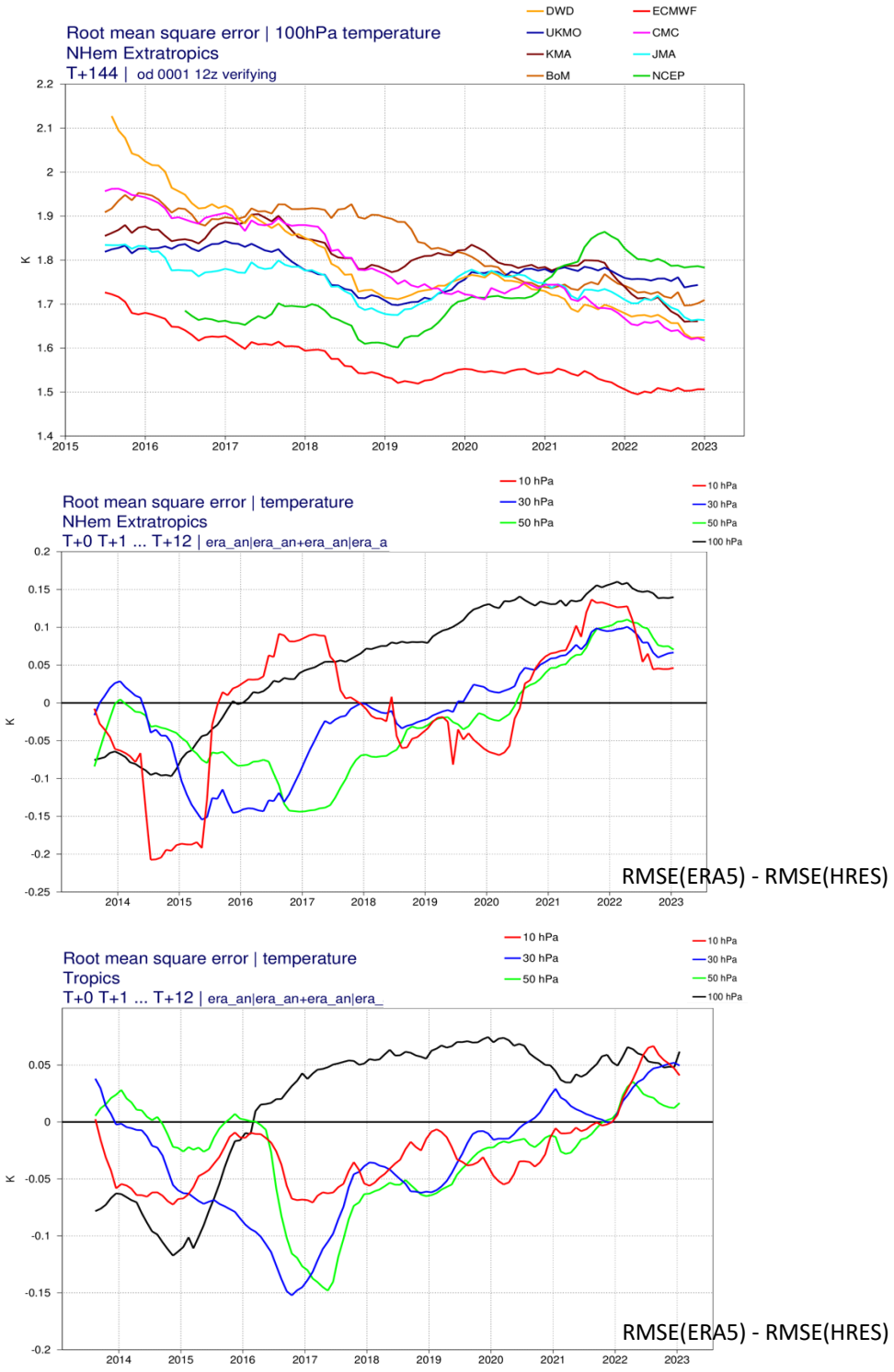


Figure 10: Stratospheric scores at a lead time of +144 h. Top: global model intercomparison of the 100 hPa temperature RMSE in the northern extratropics based on the WMO exchange of scores. Centre: difference in RMSE of temperature between ERA5 and HRES at four different stratospheric levels in the northern extratropics. Bottom: same as centre, but for the tropics. Curves in all three plots are 12-month running averages.

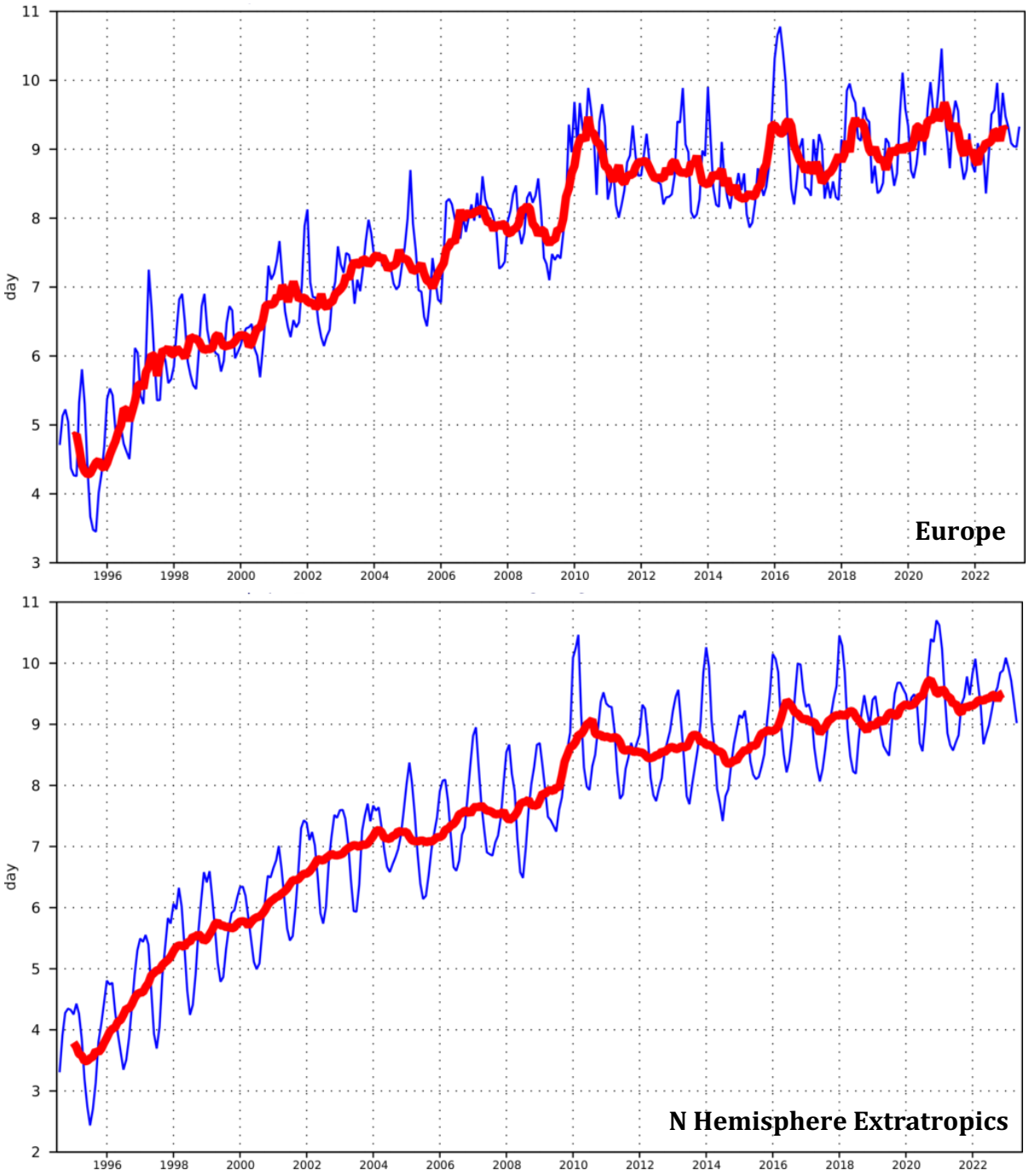


Figure 11: Primary headline score for the ensemble probabilistic forecasts. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous ranked probability skill score (CPRSS) falls below 25% for Europe (top), northern hemisphere extratropics (bottom).

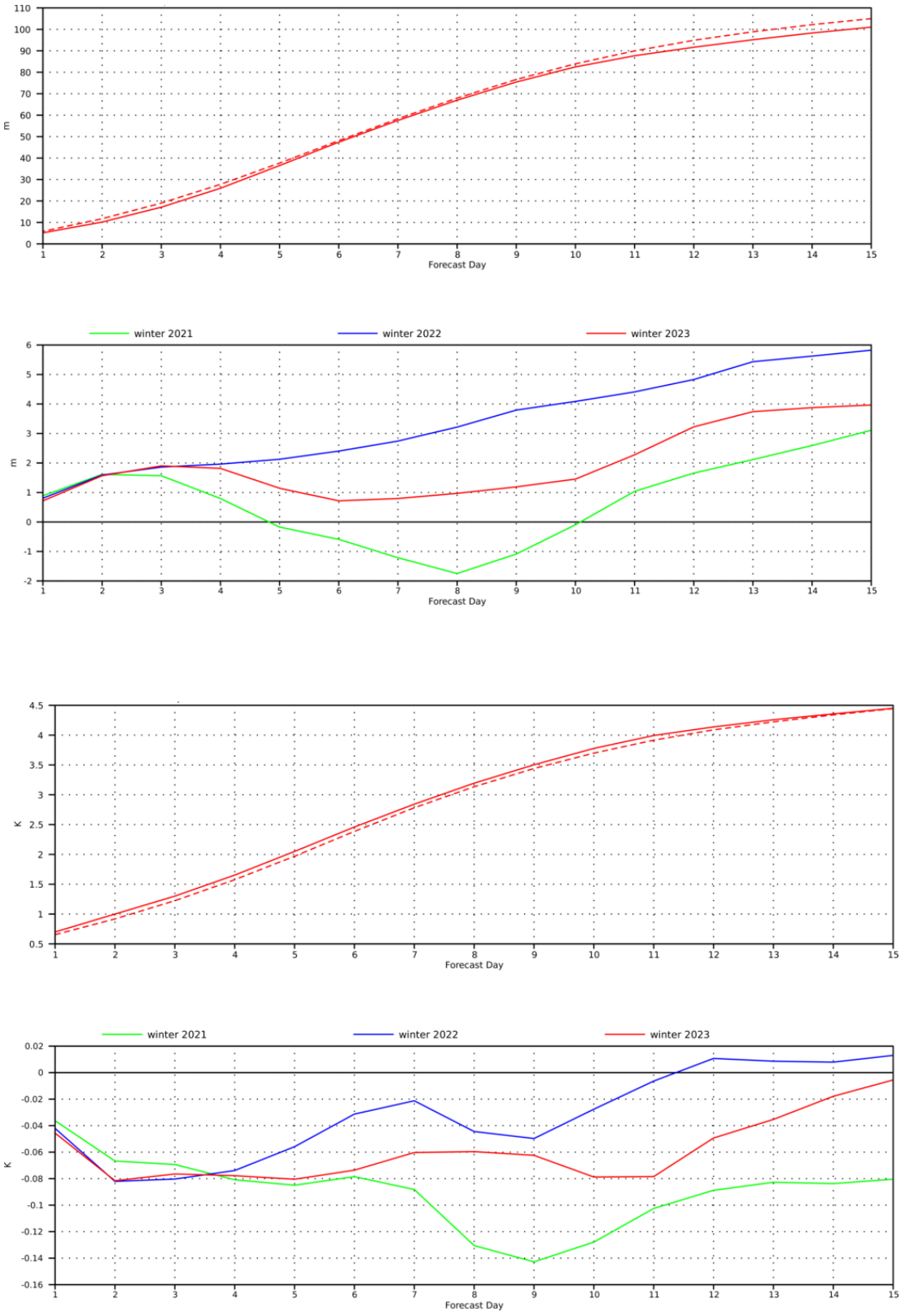


Figure 12**:** Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2022–2023 (upper figure in each panel), and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons (lower figure in each panel, negative values 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.

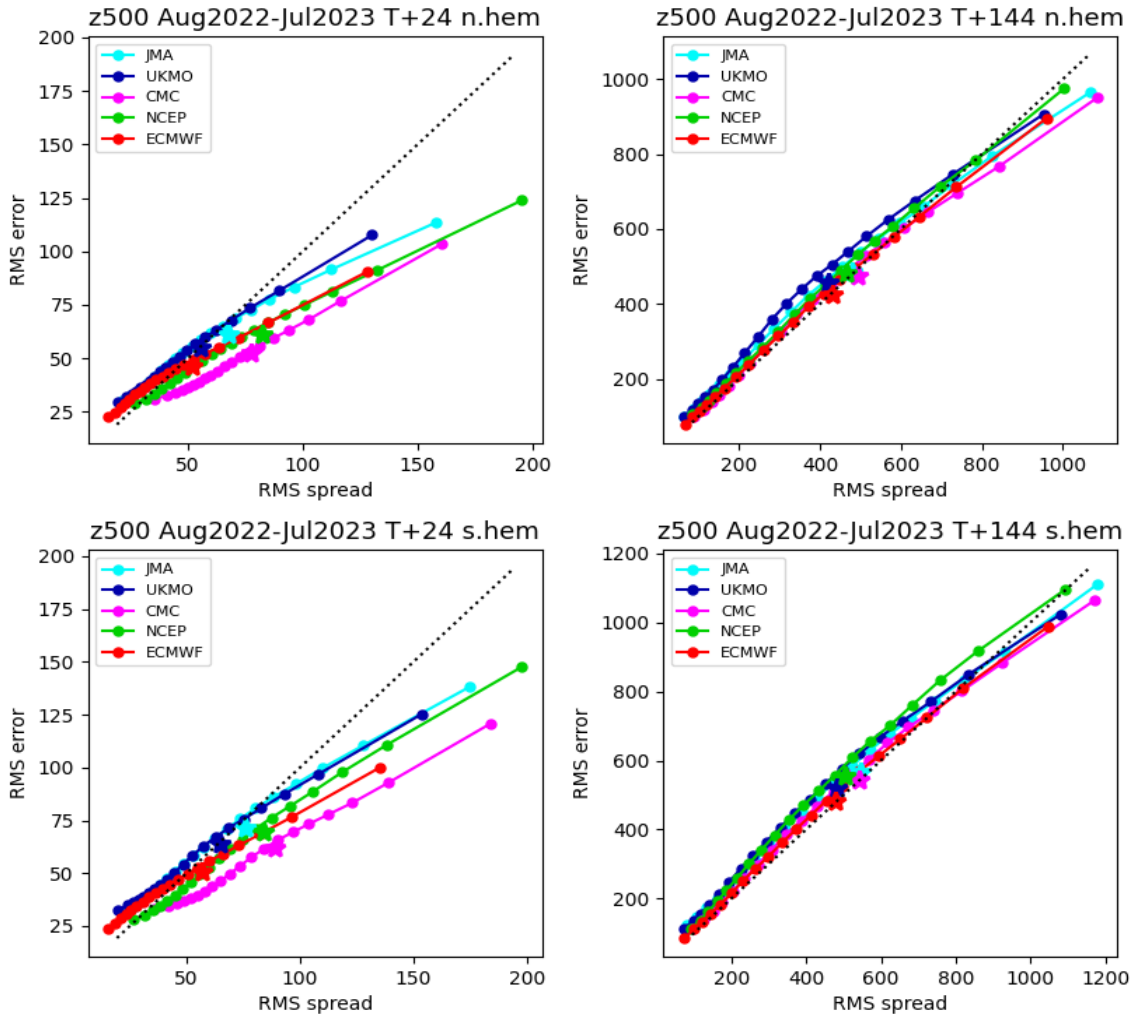


Figure 13**:** Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2022–July 2023 in the northern (top) and southern (bottom) hemisphere extra-tropics for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, stars show average error-spread relationship.

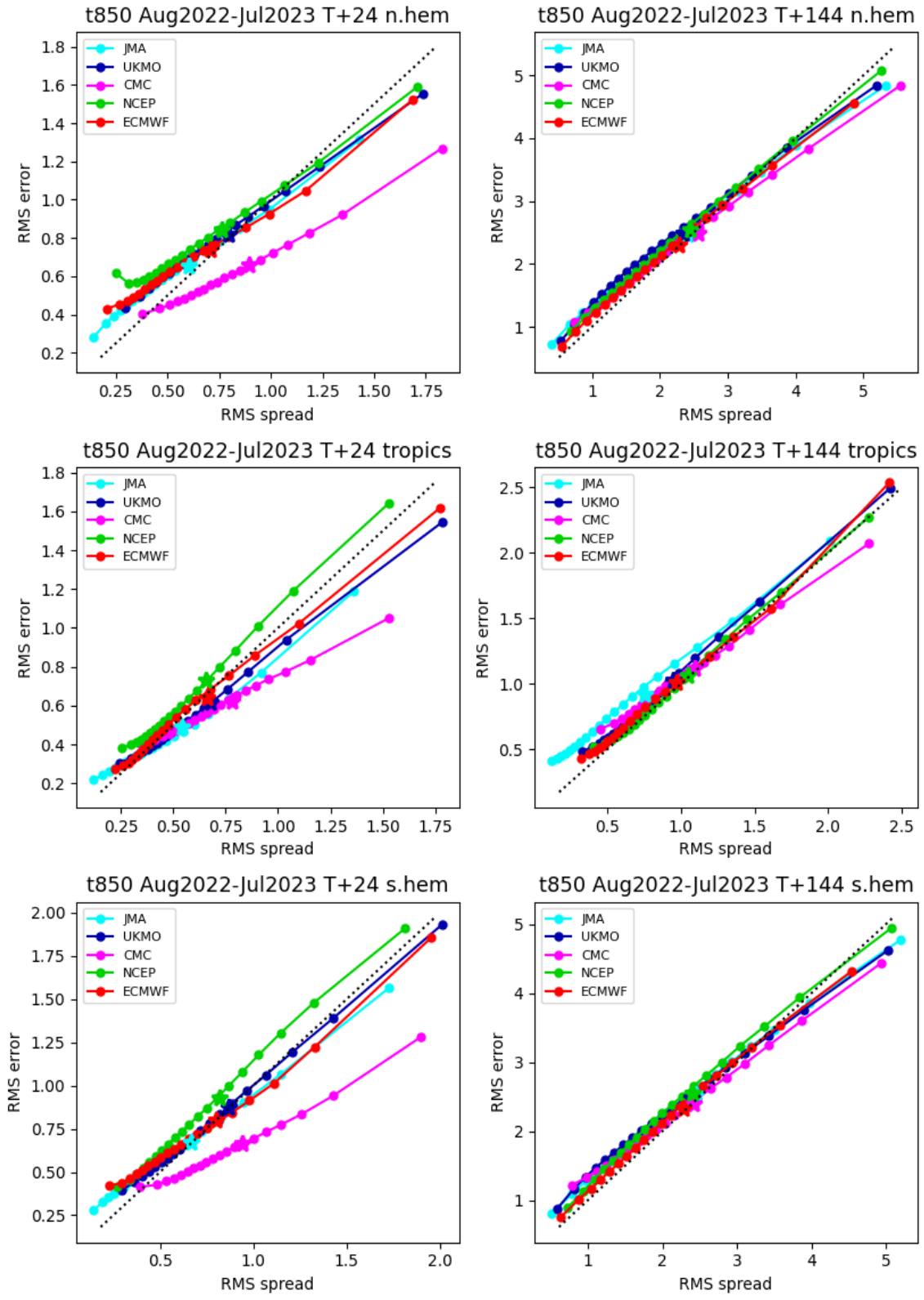


Figure 14**:** As Figure 13 for 850 hPa temperature, and including the tropics.

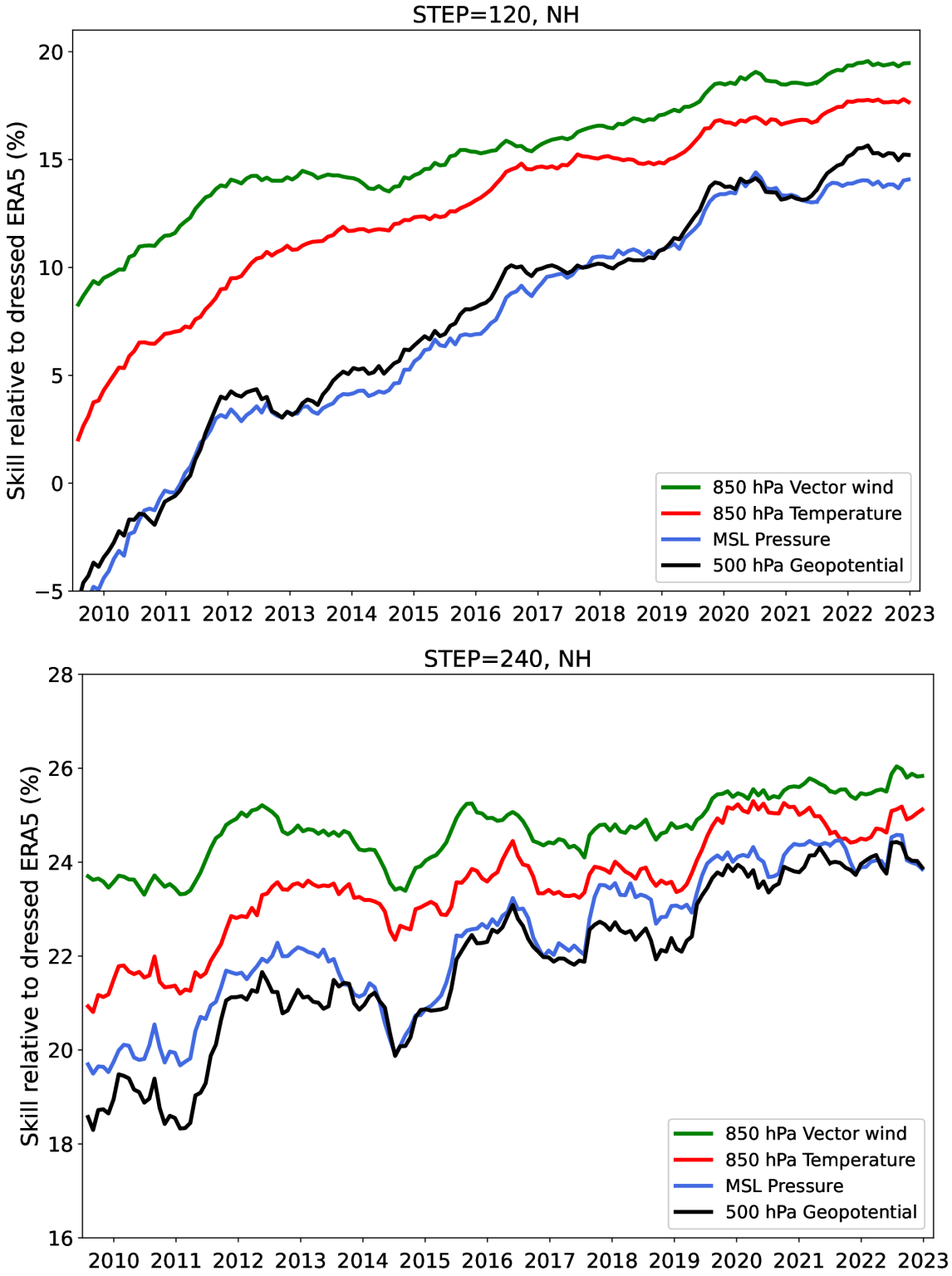


Figure 15: Skill of the ENS at day 5 (top) and day 10 (bottom) for upper-air parameters in the northern extra-tropics, relative to a Gaussian-dressed ERA5 forecast. Values are running 12-month averages, and verification is performed against own analysis.

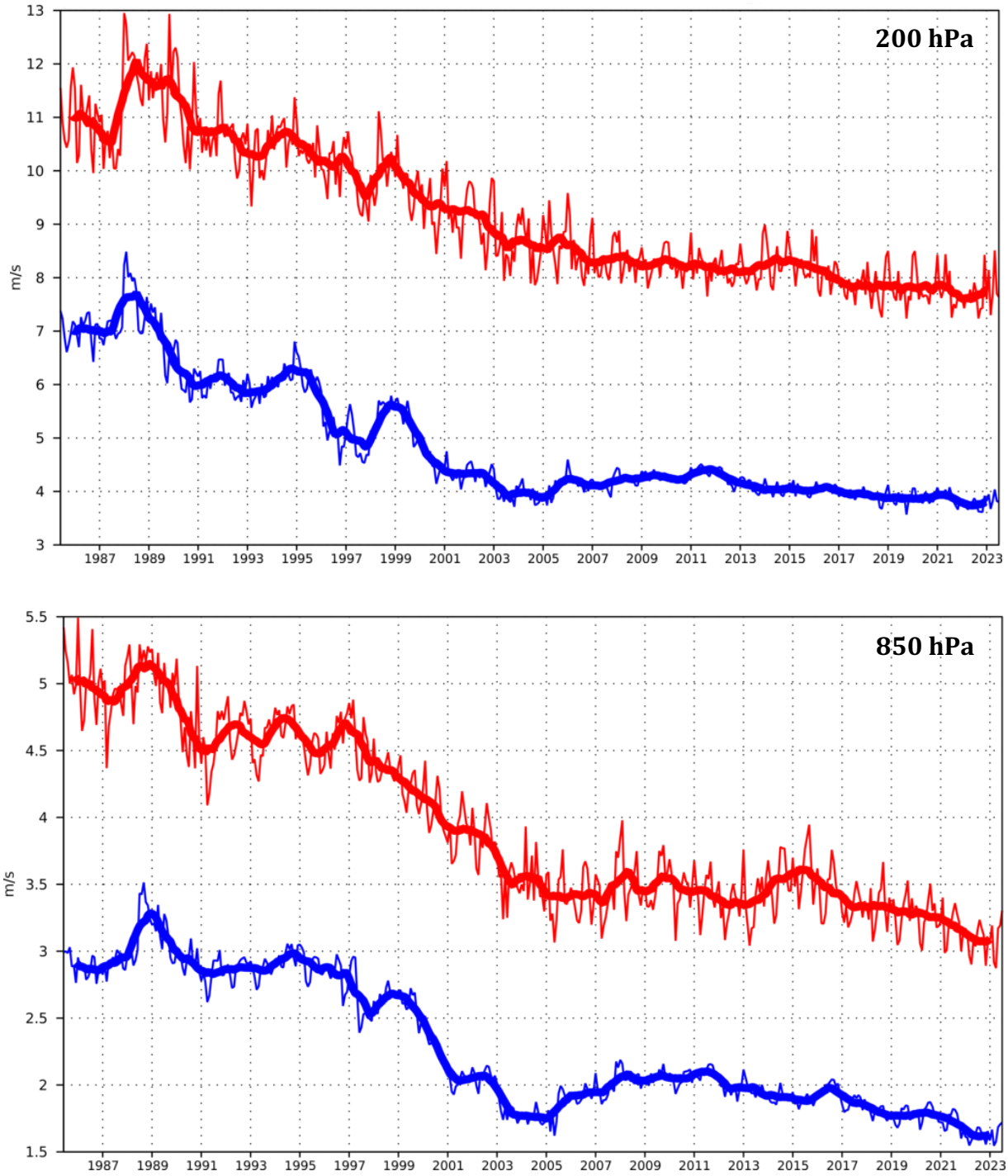


Figure 16**:** Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 200 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold).

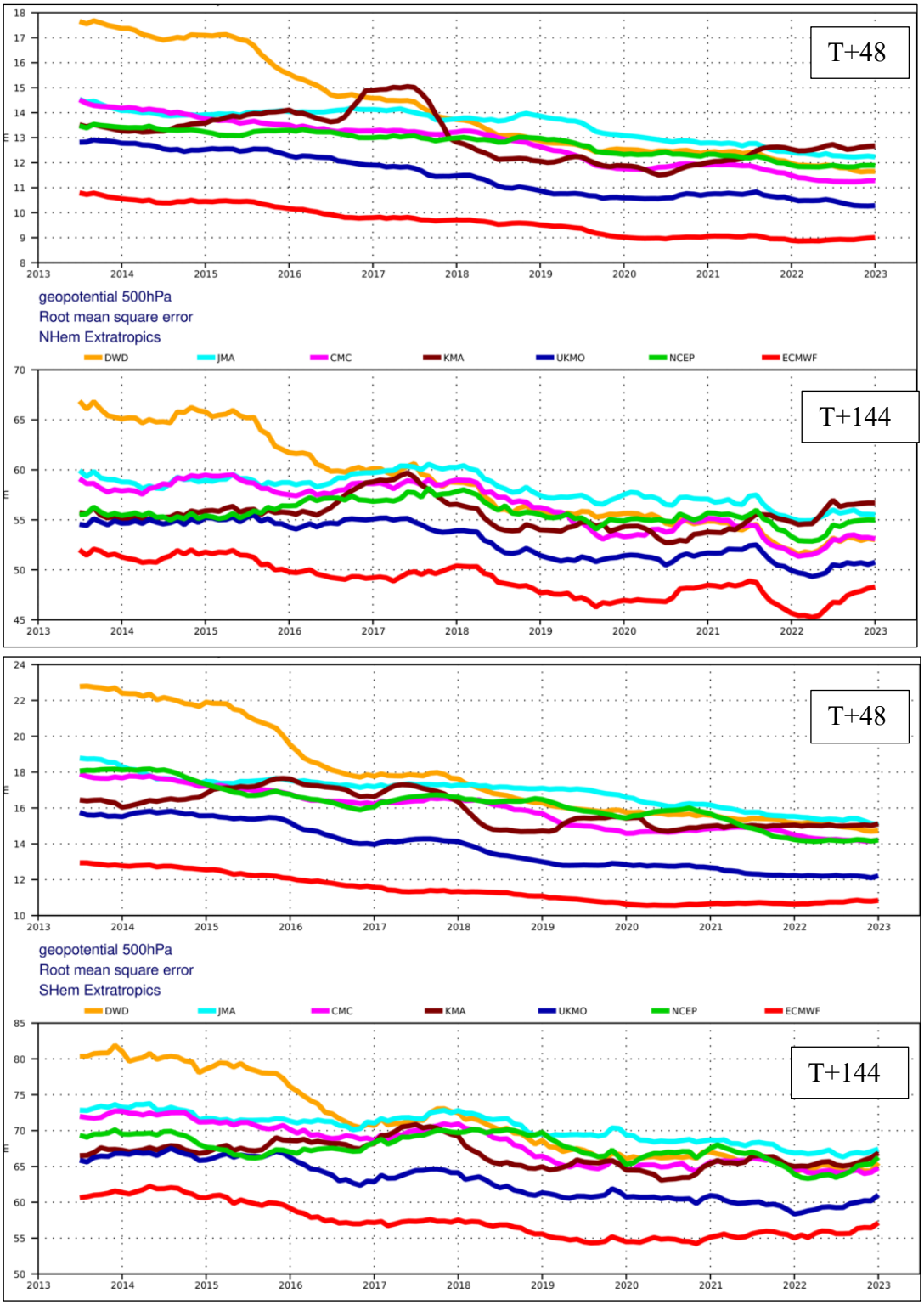


Figure 17: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top box) and southern (bottom box) extratropics. In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated 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.

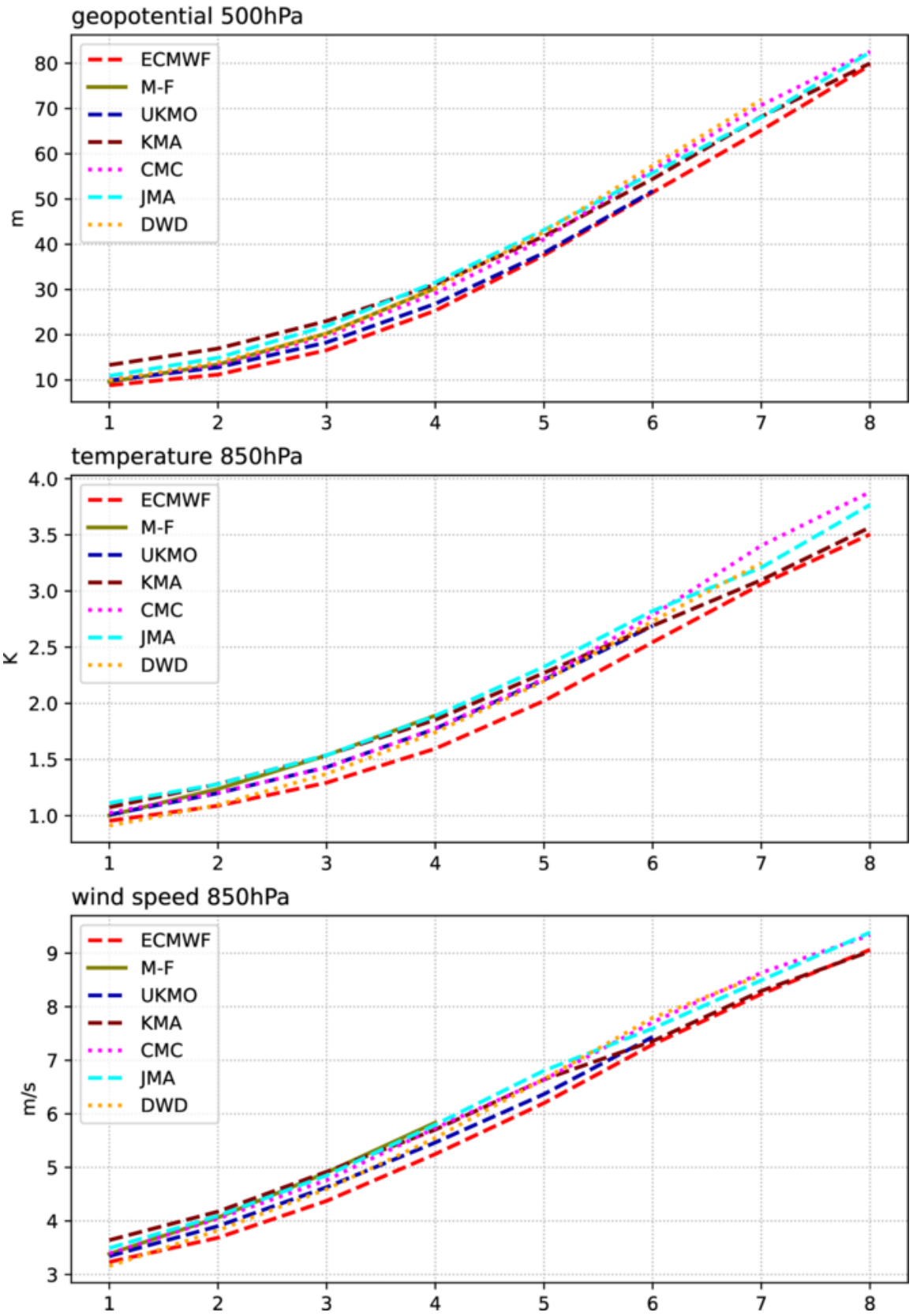


Figure 18**:** WMO-exchanged scores for verification against radiosondes: 500 hPa height (top), 850 hPa temperature (middle), and 850 hPa wind (bottom) RMS error over Europe and North Africa (annual mean August 2022–July 2023) of forecast runs initialized at 12 UTC. M-F = Météo-France, JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, KMA = Korea Meteorological Administration, DWD = Deutscher Wetterdienst.

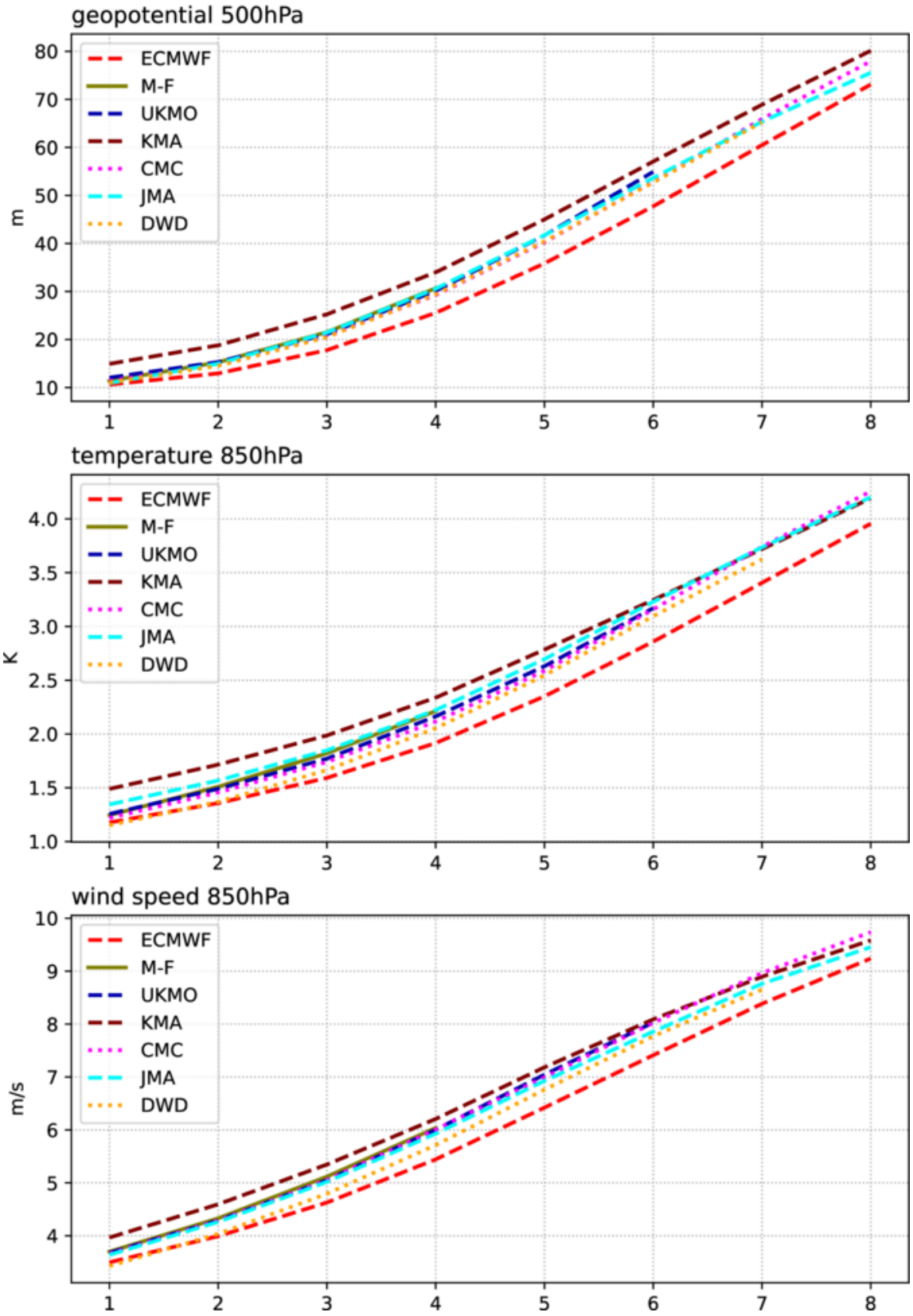


Figure 19**:** As Figure 18 for the northern hemisphere extratropics.

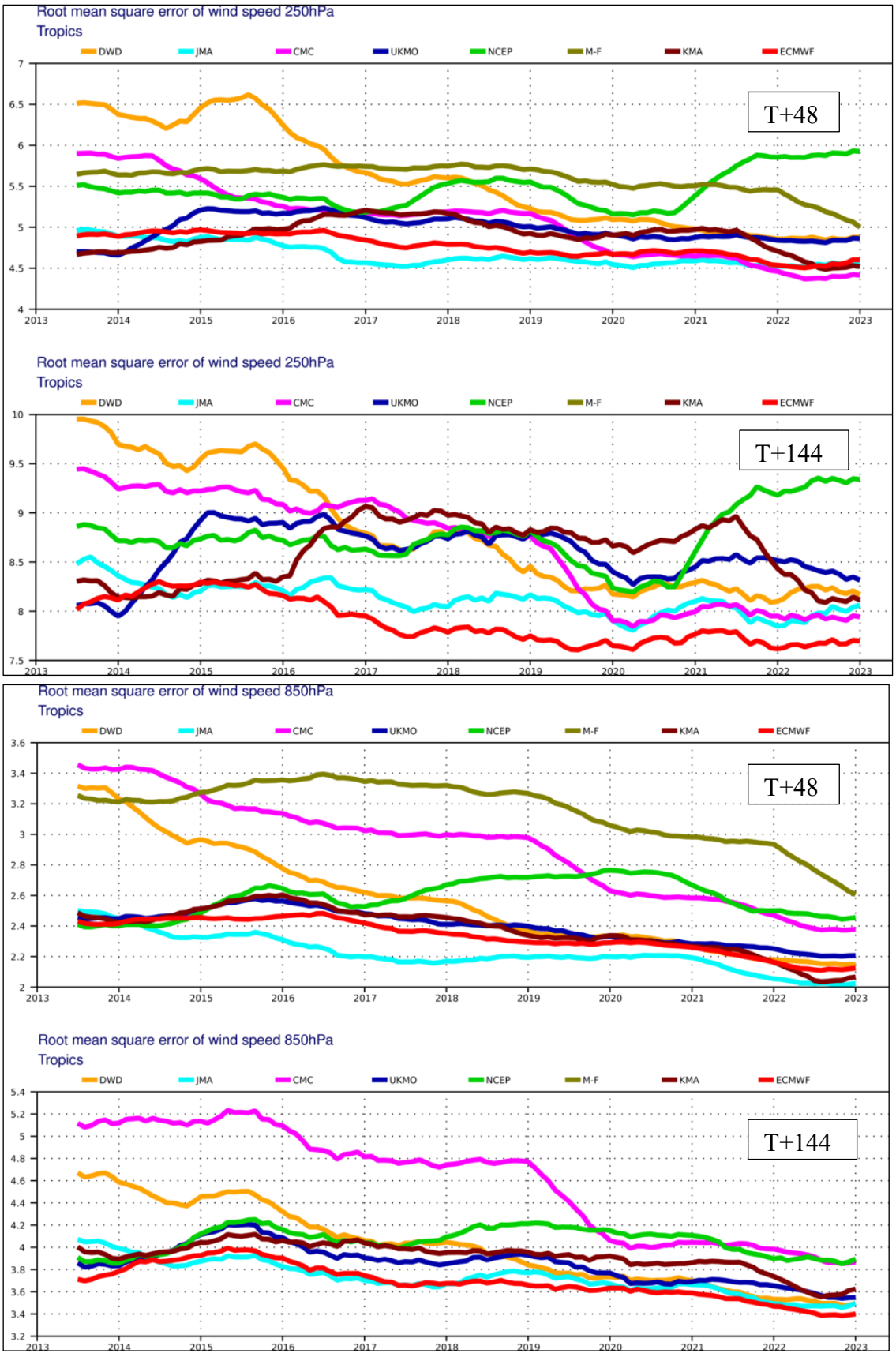


Figure 20**:** WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top box) and 850 hPa (bottom box). In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis.

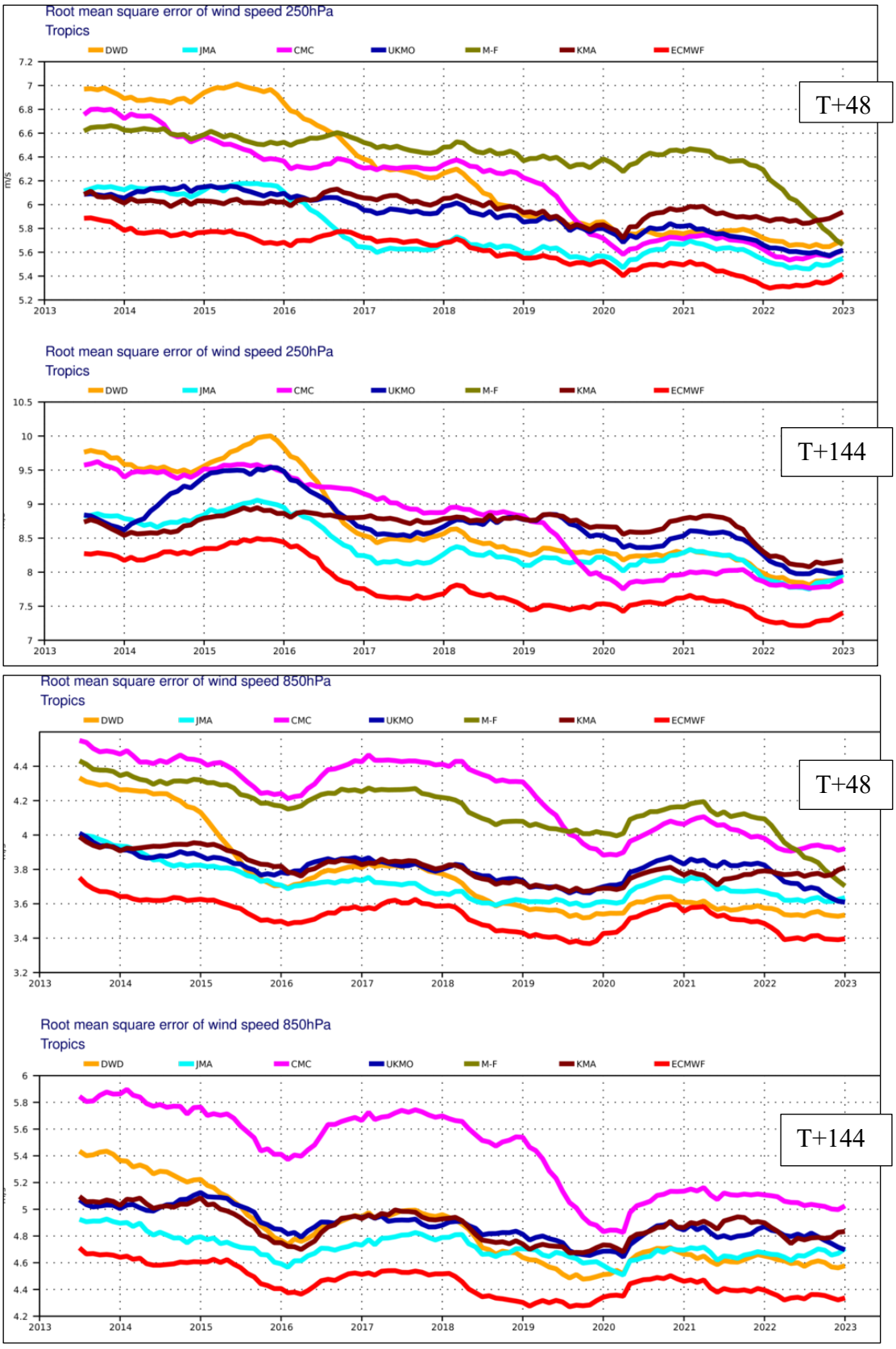


Figure 21: As Figure 20 but for verification against radiosonde observations.

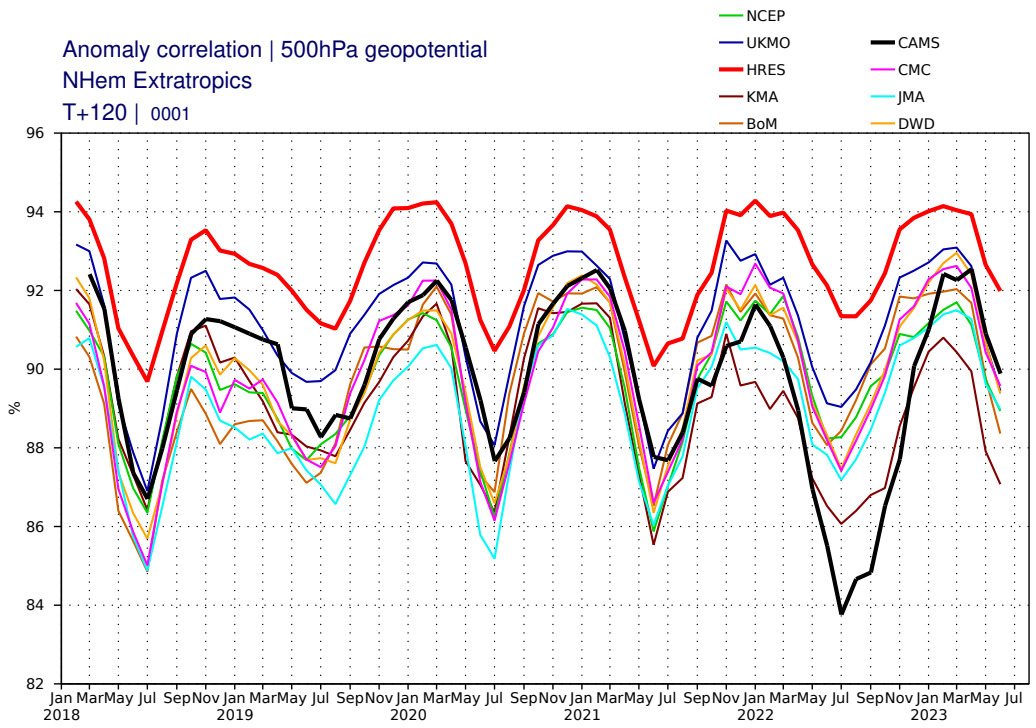


Figure 22: Anomaly correlation of 500 hPa geopotential in the northern hemisphere extratropics at day 5. CAMS forecast (black) shown in comparison to the HRES (red) and forecasts from other global centres (thin lines).

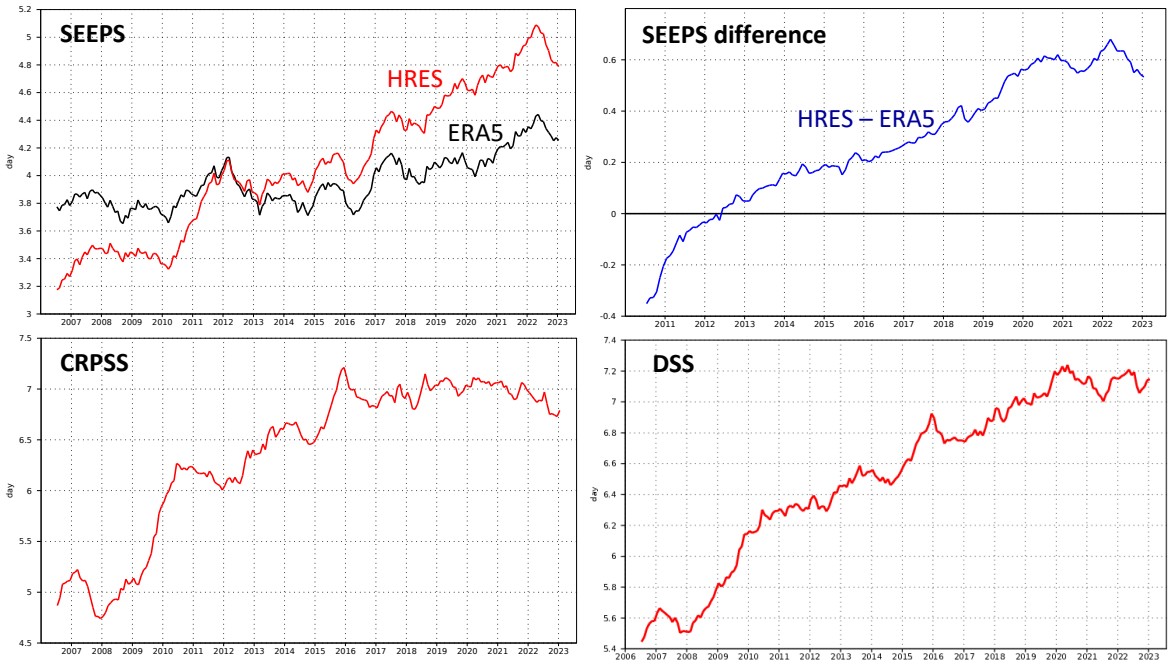


Figure 23**:** Supplementary headline scores (left column) and additional metrics (right column) for deterministic (top) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number 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).

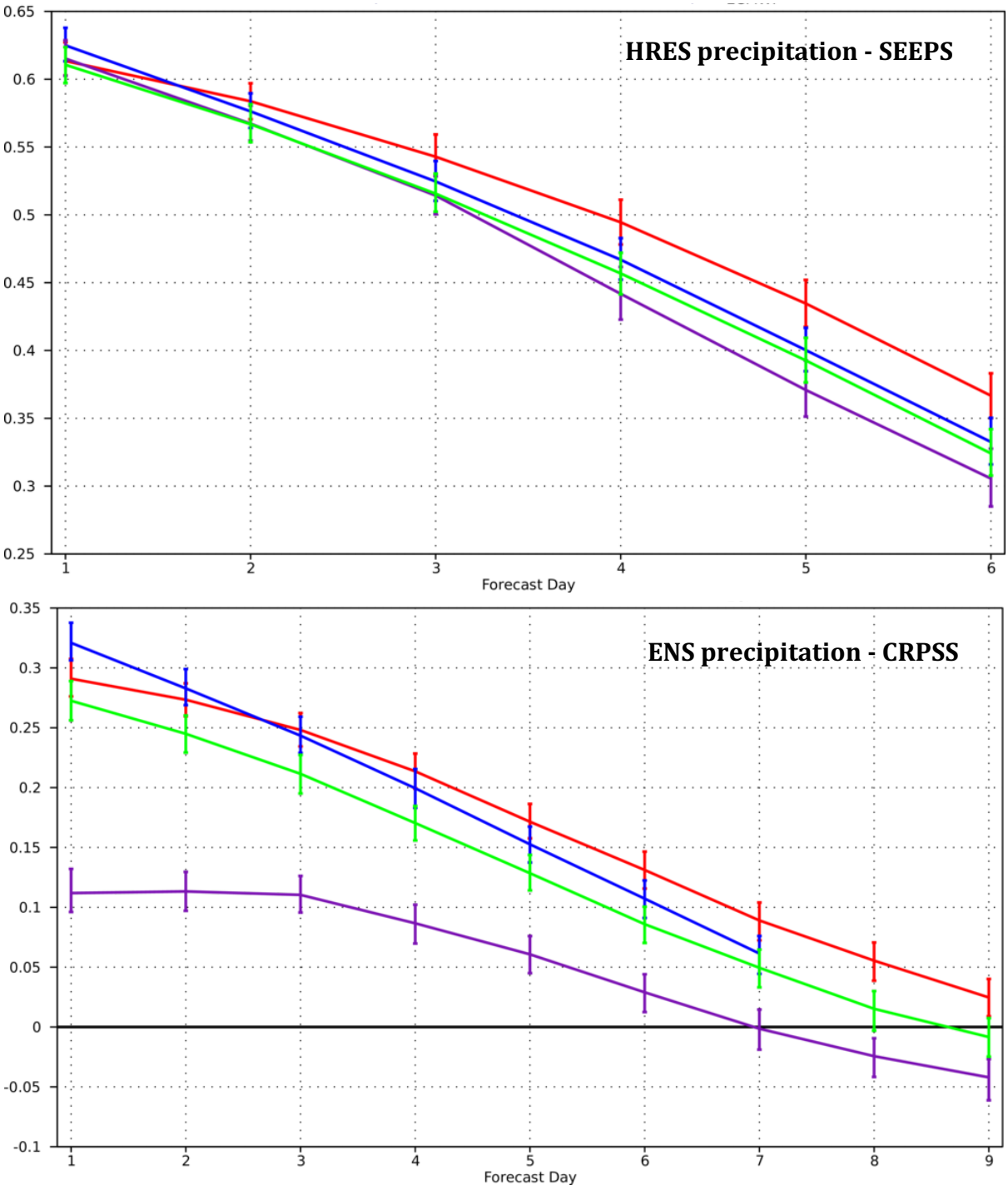


Figure 24**:** Comparison of precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, magenta) and NCEP (green) using the supplementary headline scores for precipitation shown in Figure 23. Top: deterministic; bottom: probabilistic skill. Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2022–July 2023. Bars indicate 95% confidence intervals.

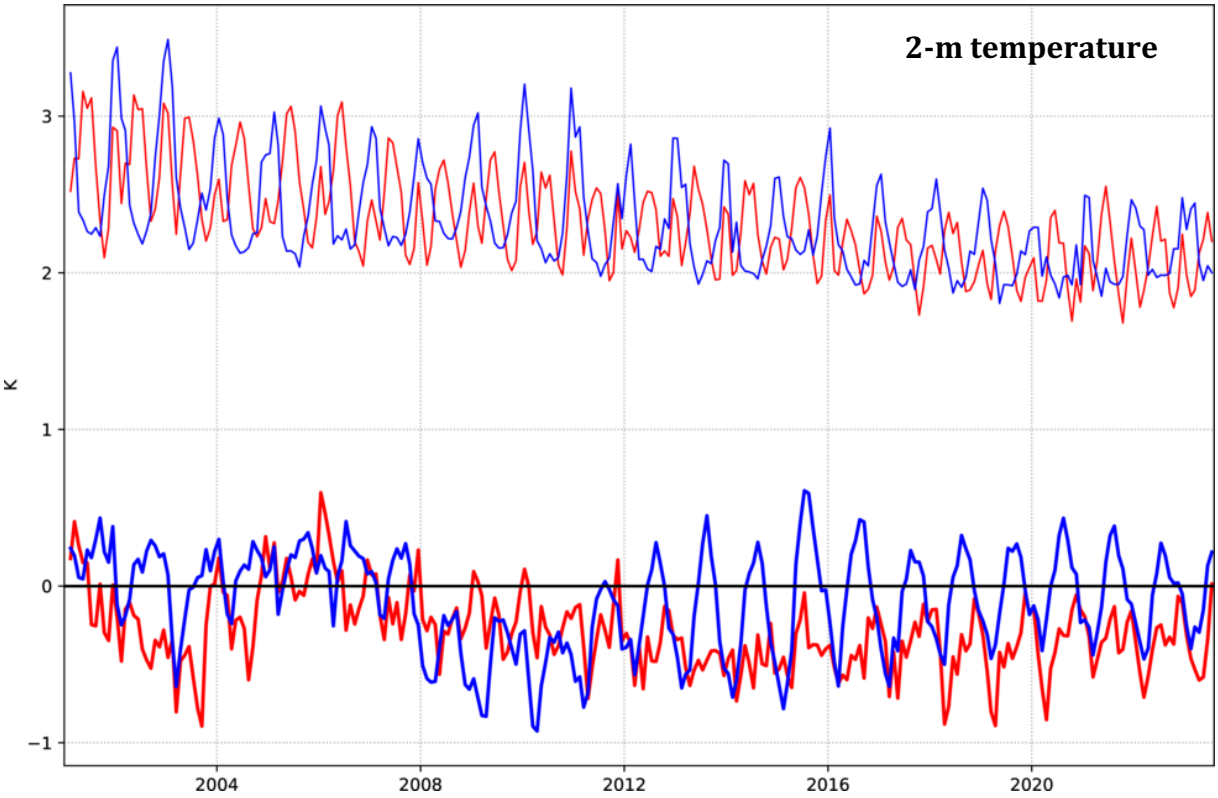


Figure 25: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves are standard deviation of error.

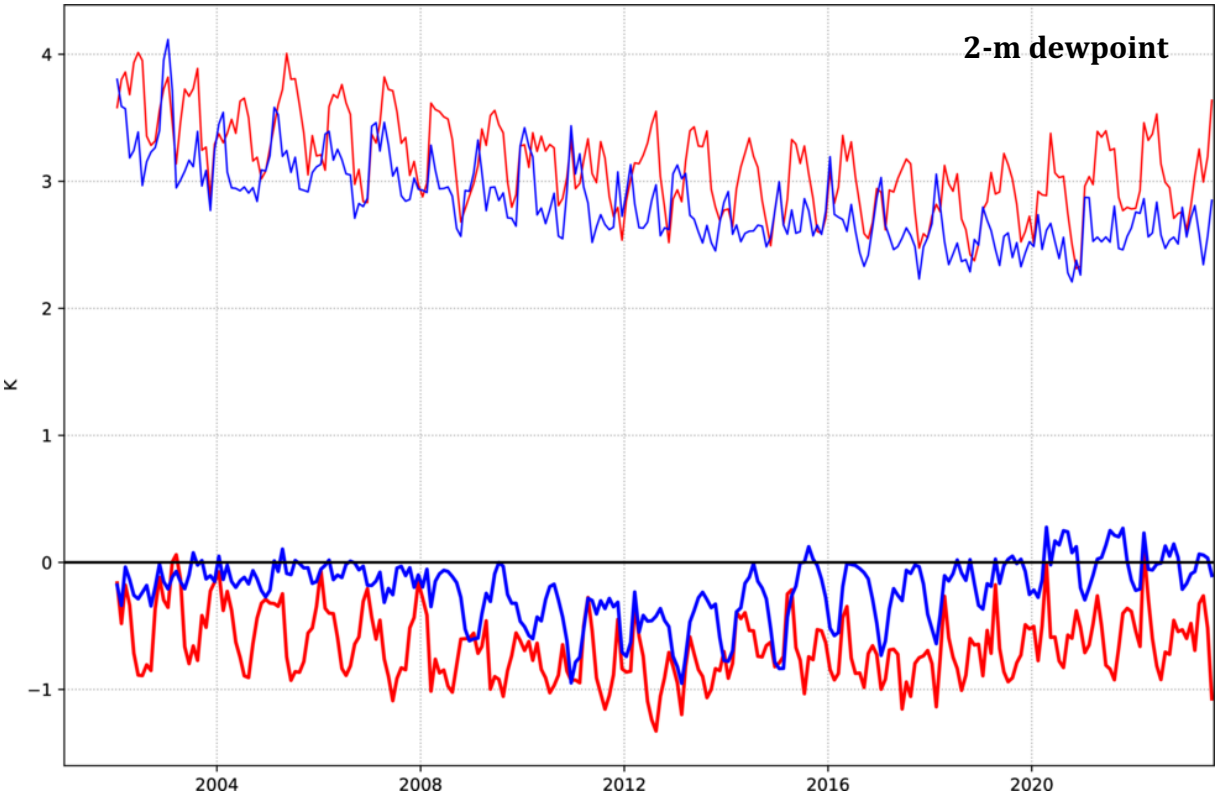


Figure 26: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

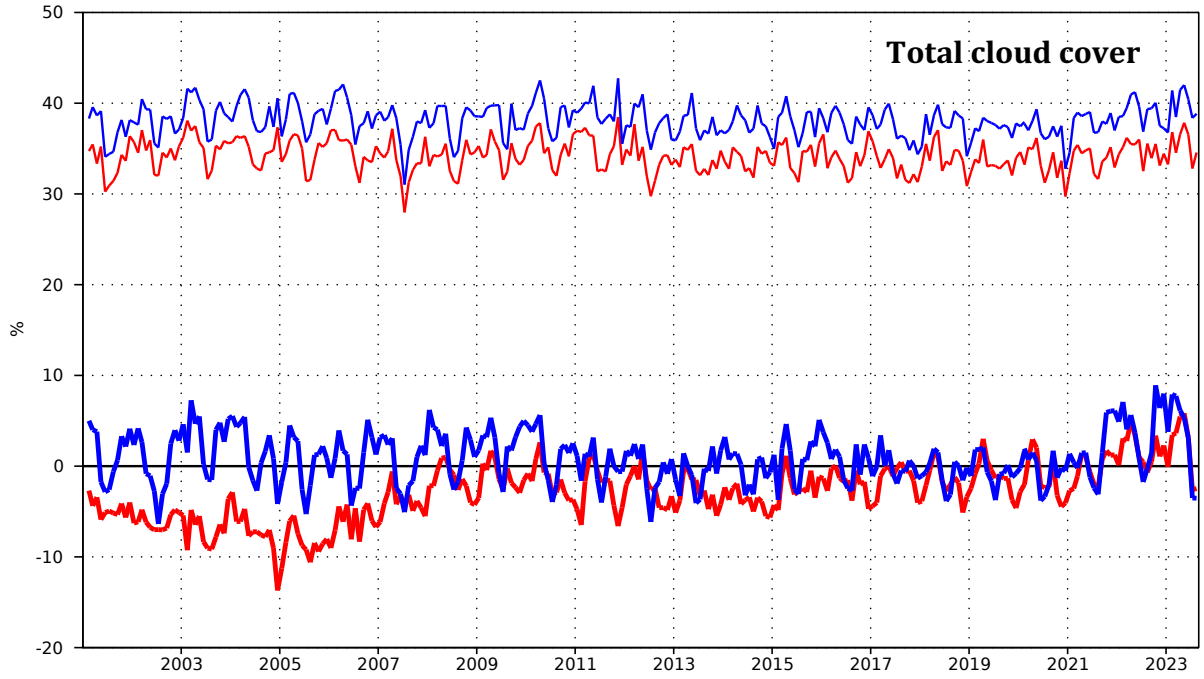


Figure 27: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

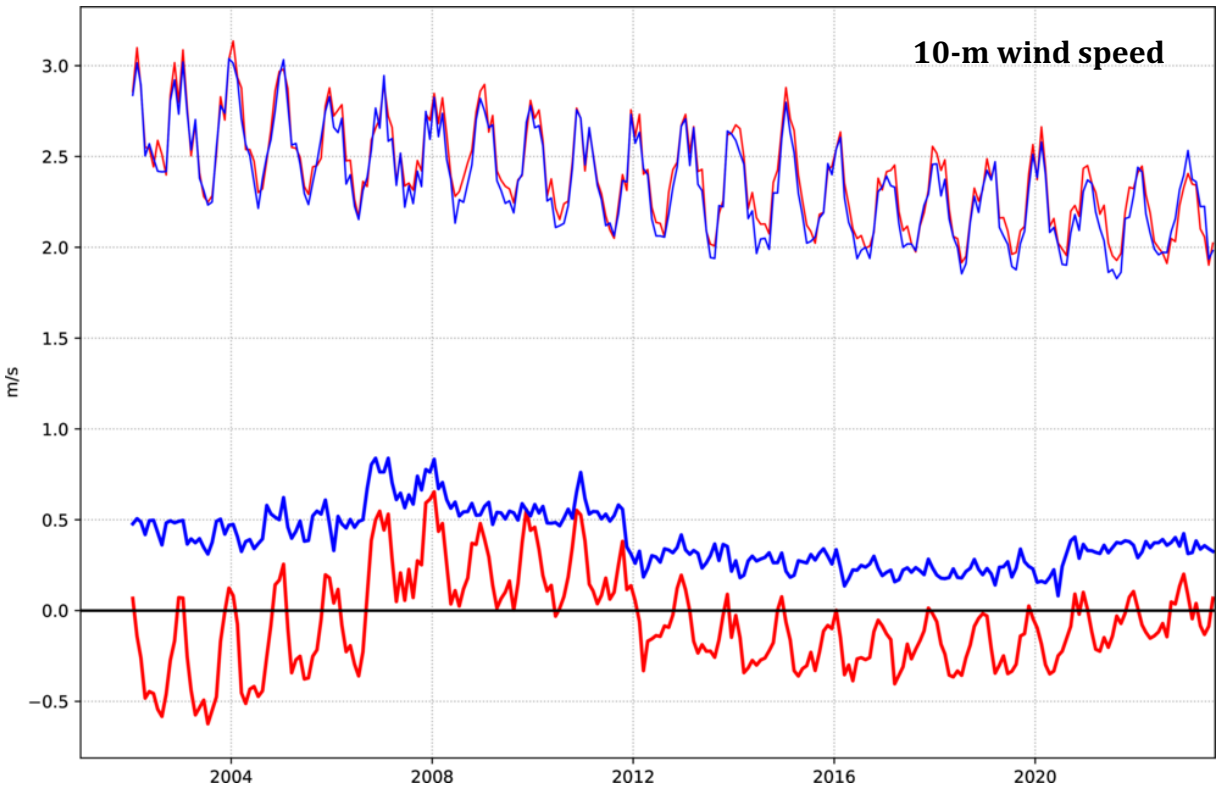


Figure 28: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves shows bias, upper curves show standard deviation of error.

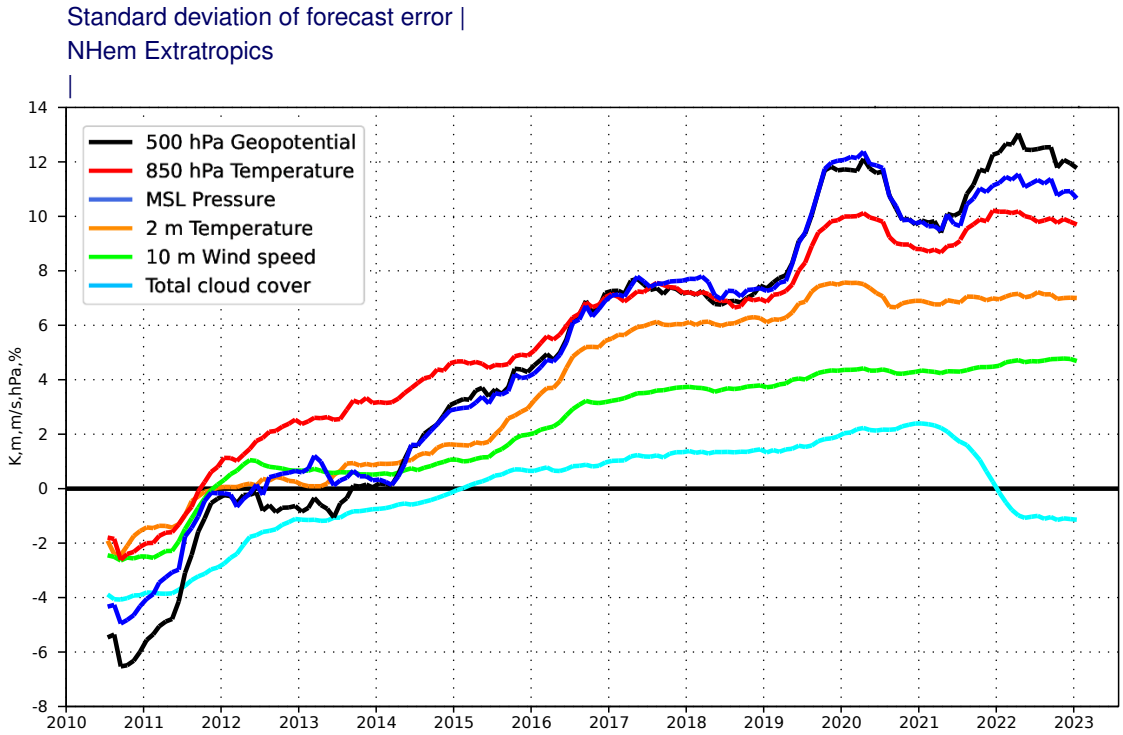


Figure 29: Evolution of skill of the HRES forecast at day 5 in the northern hemisphere extratropics, expressed as relative skill compared to ERA5. Verification is against analysis for 500 hPa geopotential, 850 hPa temperature, and mean sea level pressure, using error standard deviation as a metric. Verification is against SYNOP for 2 m temperature, 10 m wind speed, and total cloud cover.

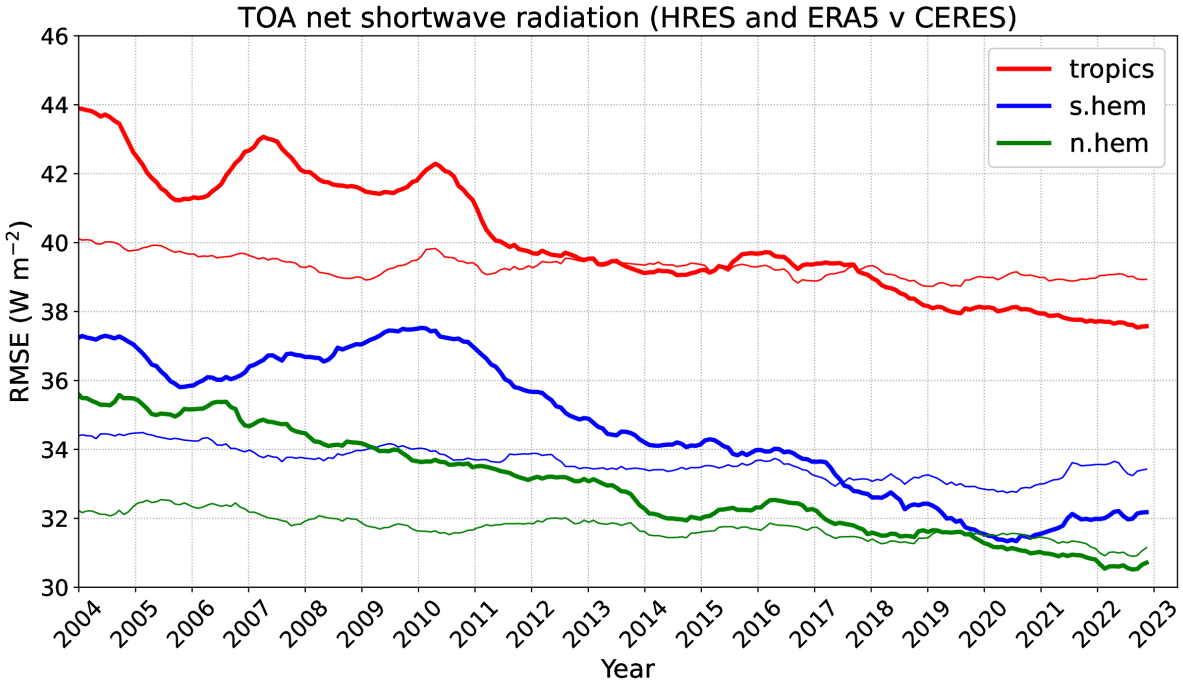


Figure 30: Evolution of the RMSE of the HRES forecast at day 5 (bold lines) of the top of the atmosphere (TOA) net shortwave radiation for the two extratropical hemisphere and the tropics. Thin lines show the RMSE of the ERA5 forecast for comparison. Verification is against CERES satellite data.

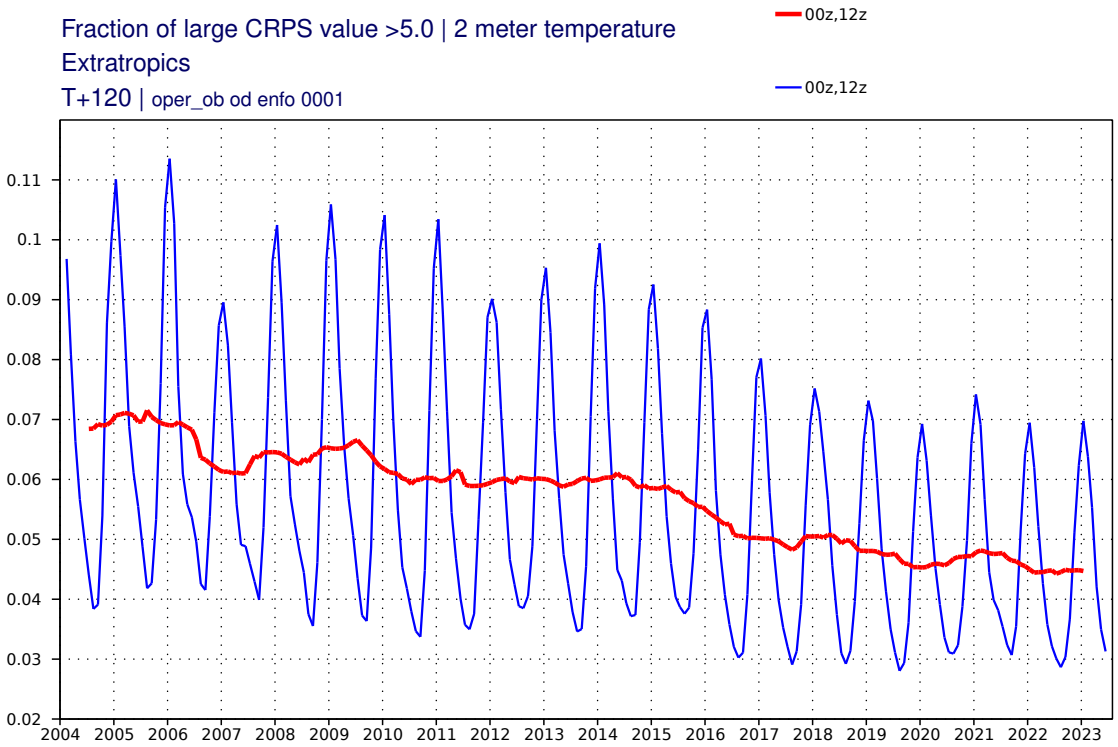


Figure 31: Evolution of the fraction of large ENS 2m temperature errors (CRPS>5K) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

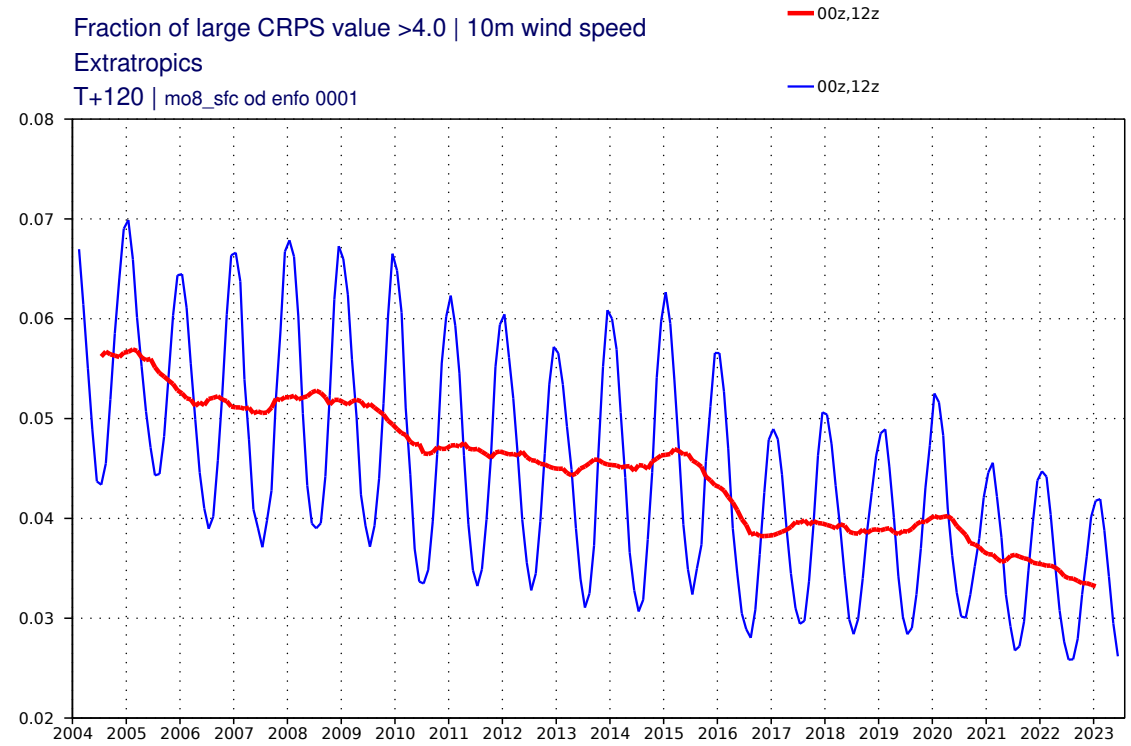


Figure 32: Evolution of the fraction of large ENS 10m wind speed errors (CRPS>4m/s) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

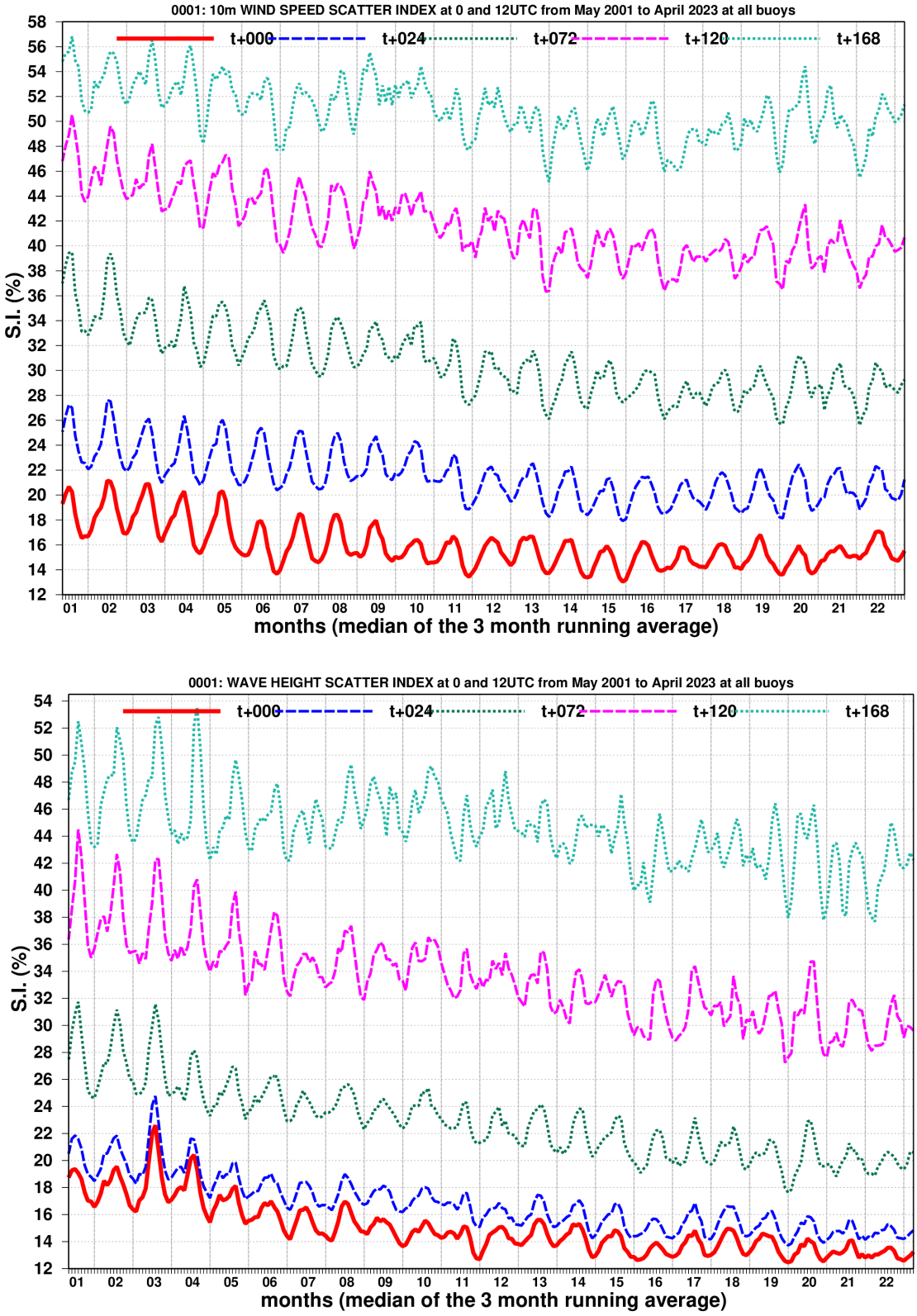


Figure 33: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave height forecast (bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is applied.

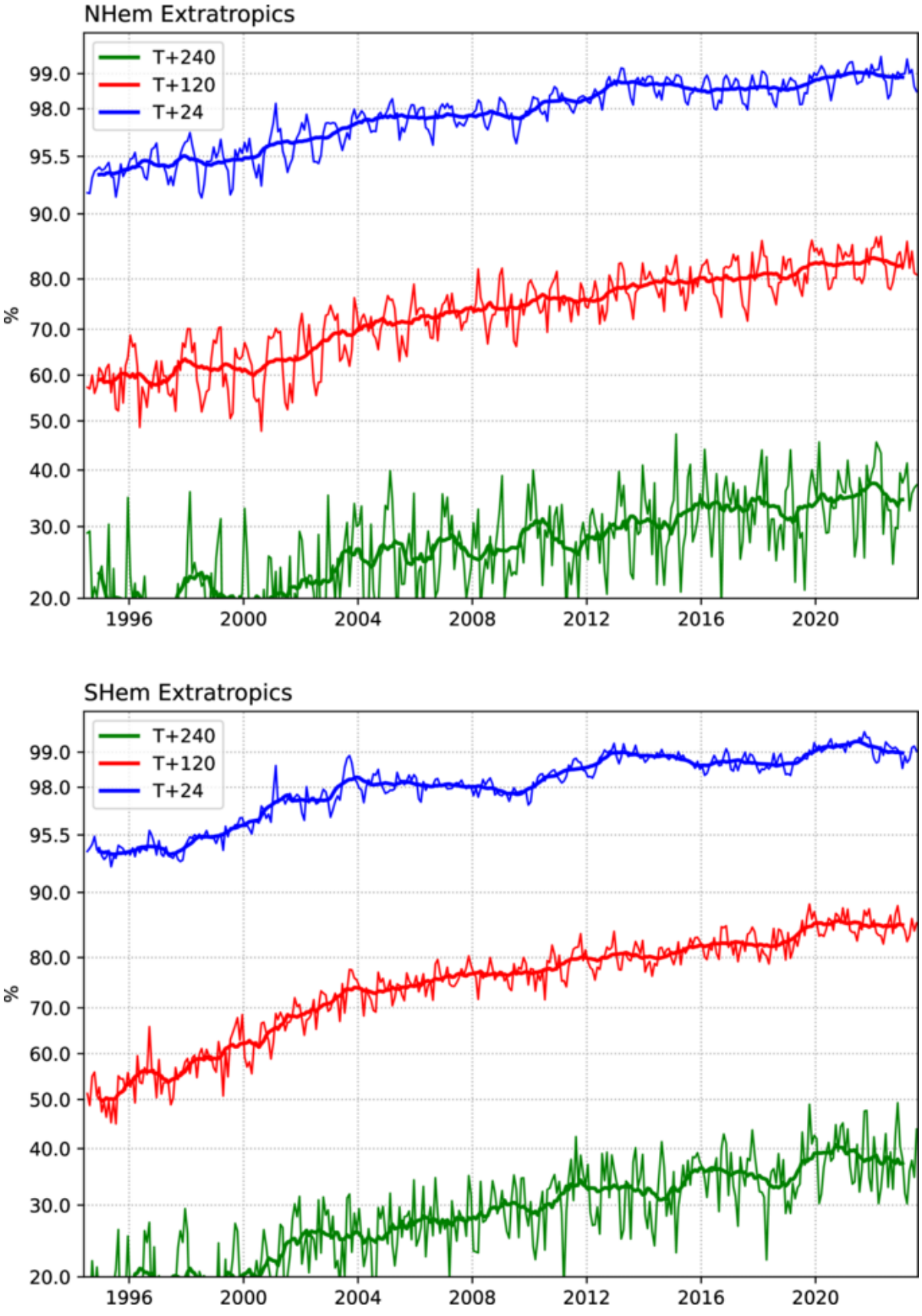


Figure 34: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

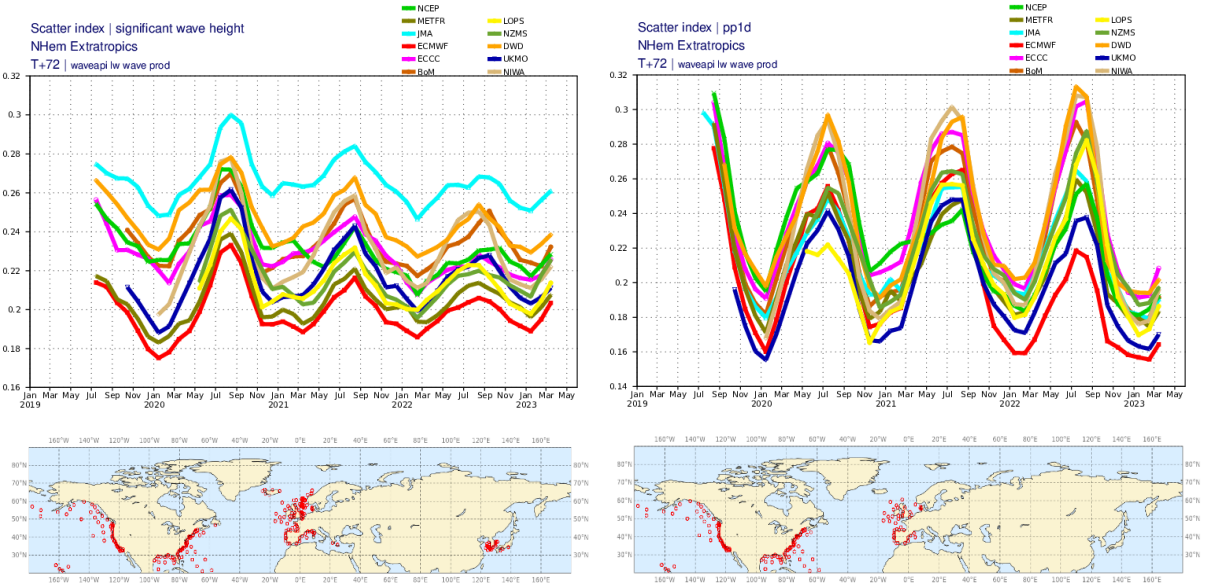


Figure 35: Verification of forecasts of wave height and peak wave period (upper panels) at +72 h using observations from wave buoys (lower panels). The scatter index (SI) is the standard deviation of error normalised by the mean observed value. NCEP: National Centers for Environmental Prediction, USA; 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.

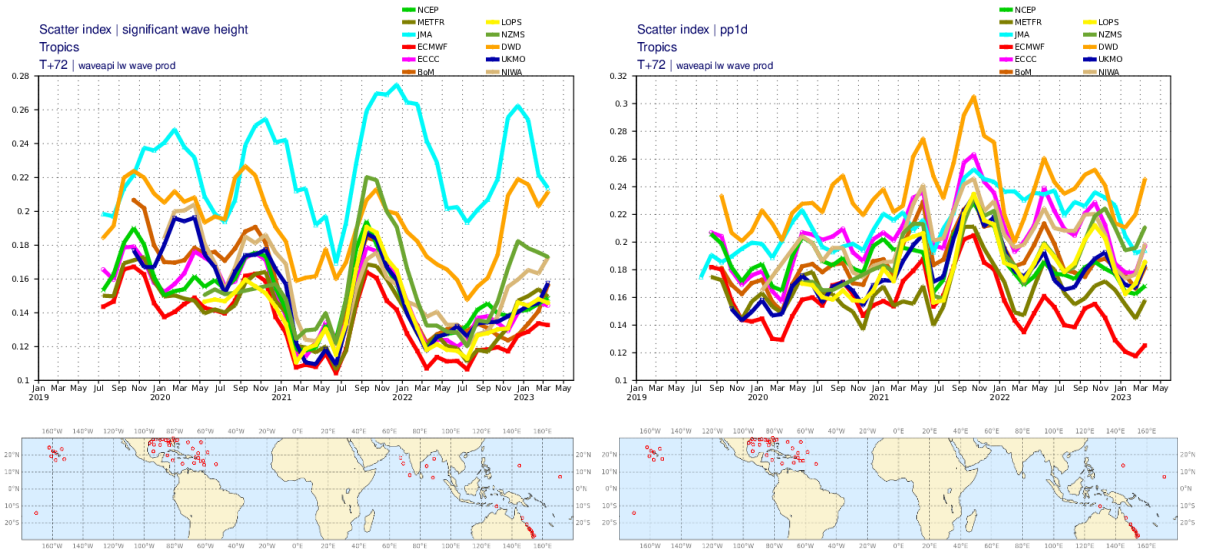


Figure 36: As Figure 35, but for the tropics.

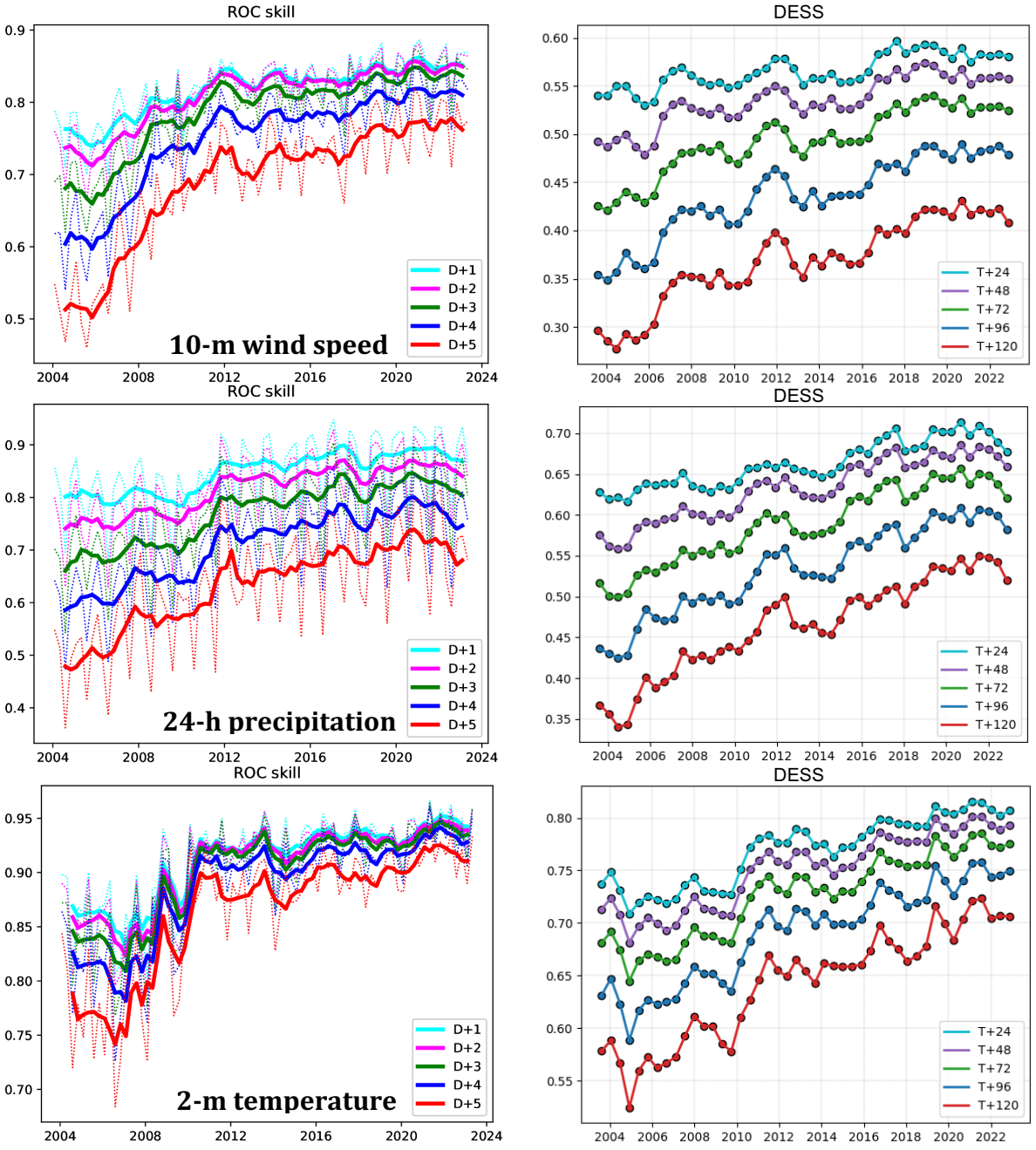


Figure 37: Left column shows verification of the Extreme Forecast Index (EFI) against SYNOP observations. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 5 (24-hour period 96–120 hours ahead); skill at day 4 (blue line) is the supplementary headline score; an extreme 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. Centre and bottom panels on the left show the equivalent ROC area skill for precipitation EFI forecasts and for 2 m temperature EFI forecasts. Right column shows the diagonal elementary skill score (DESS) for the 95th percentile for the same three variables, verified against SYNOP, taking observation uncertainty into account.

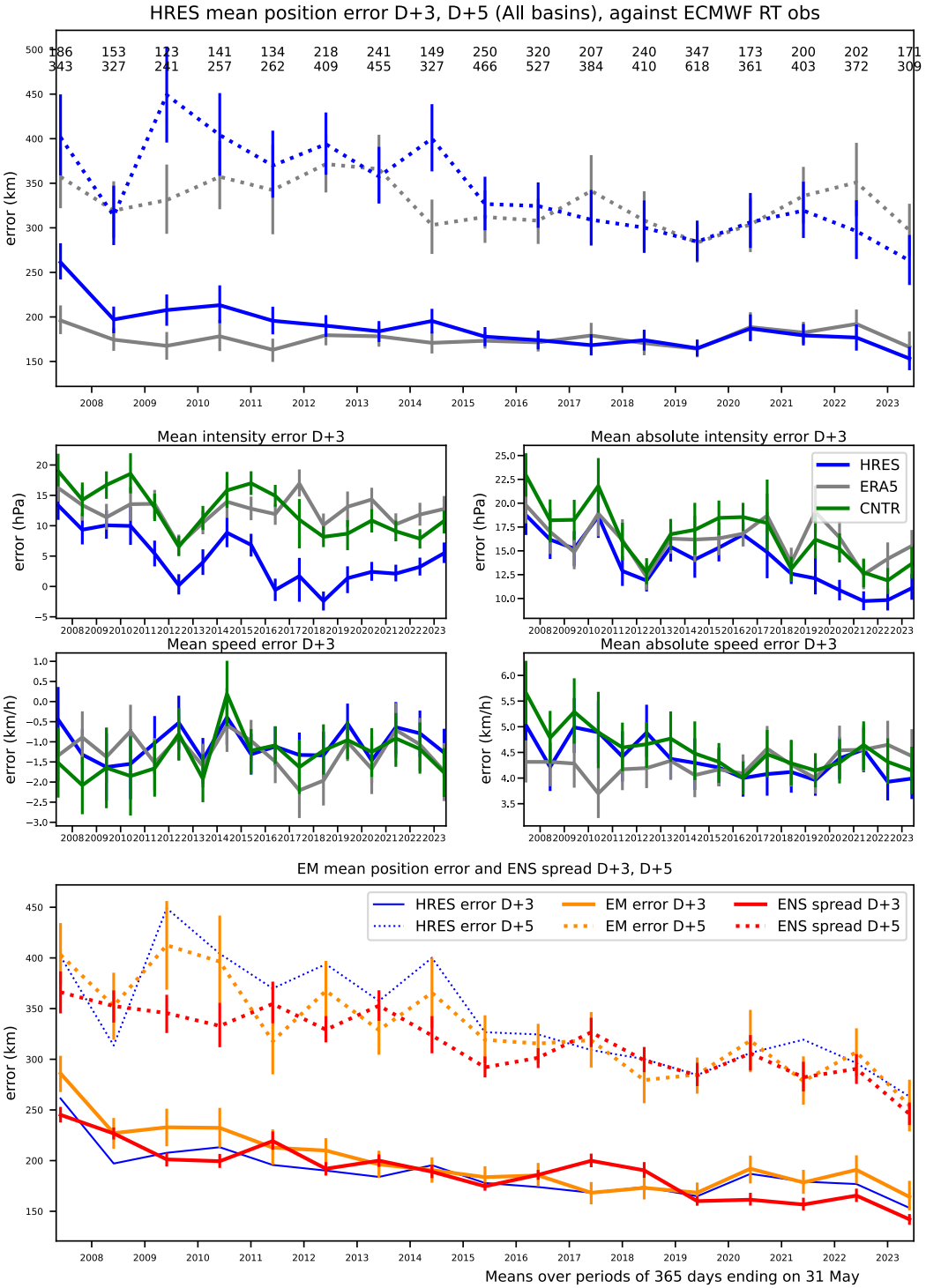


Figure 38: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary 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.

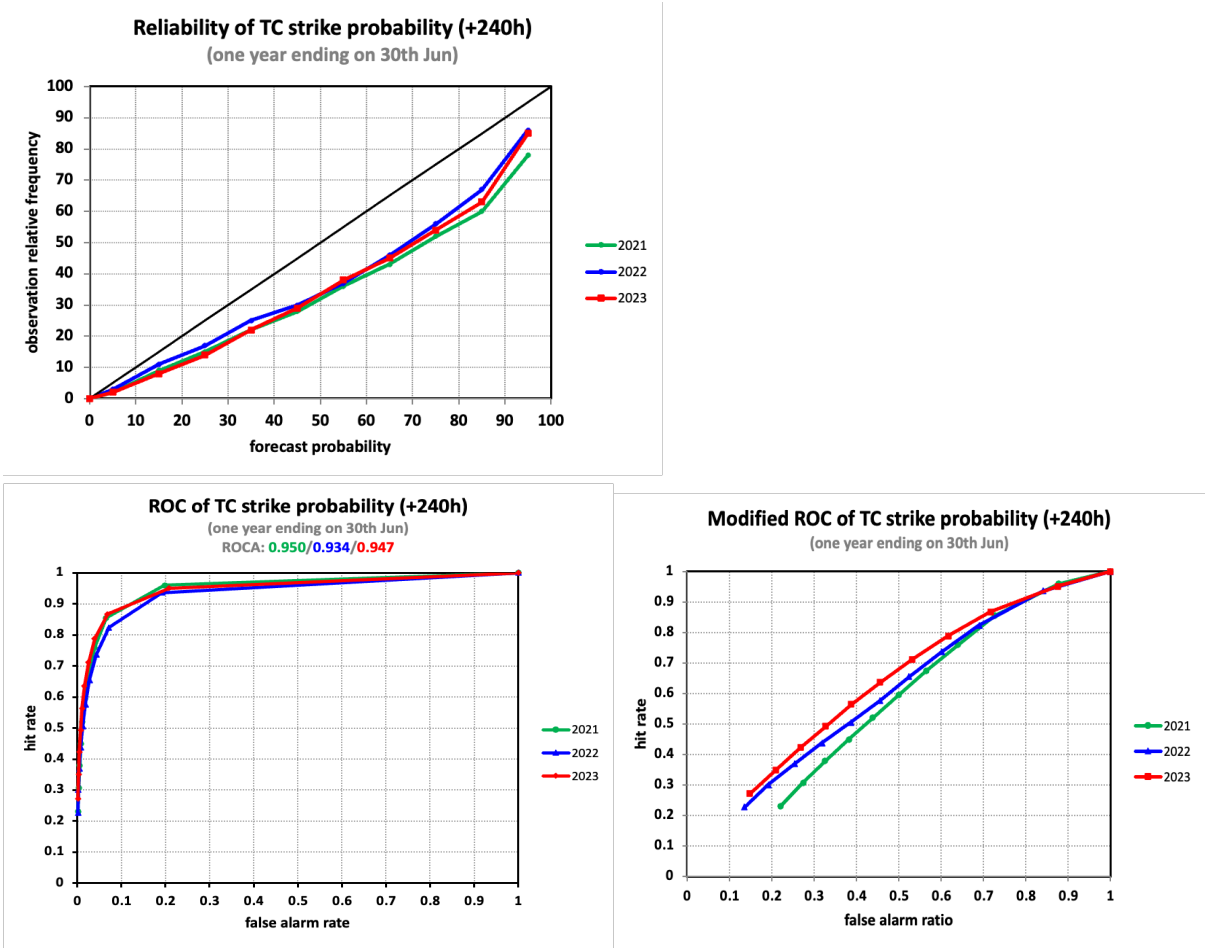


Figure 39: Probabilistic verification of ensemble tropical cyclone forecasts at day 10 for three 12-month periods: July 2019–June 2020 (green), July 2020–June 2021 (blue) and July 2021–June 2022 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (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.

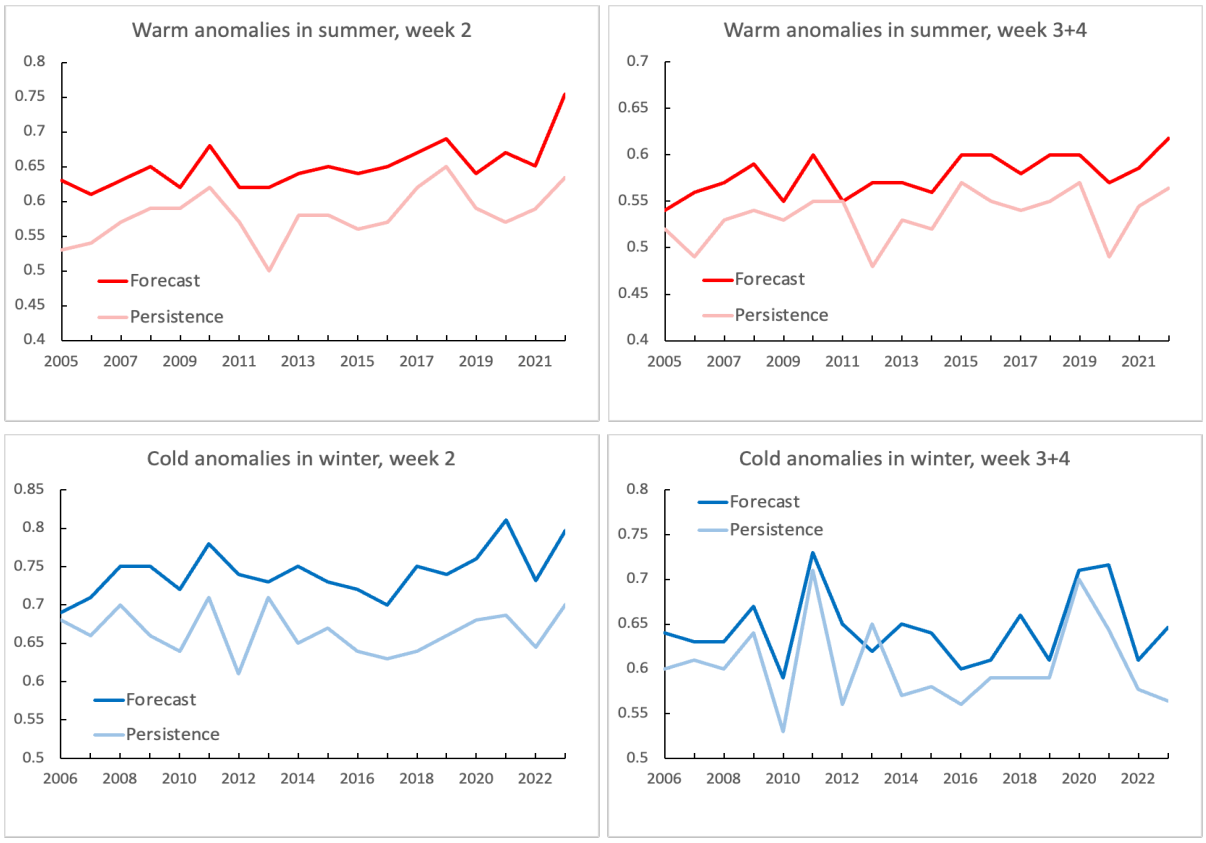


Figure 40**:** Verification of the monthly forecast against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all 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.

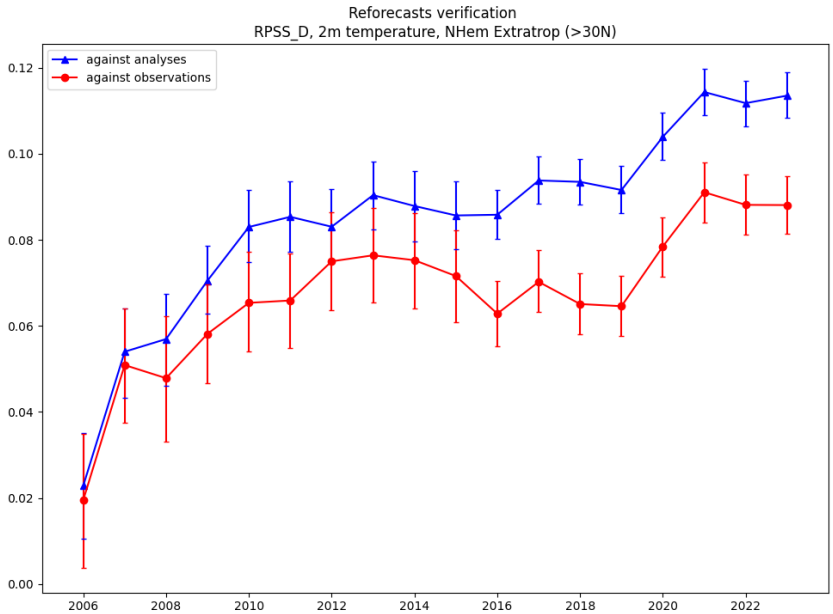


Figure 41: Skill of the ENS in predicting weekly mean 2m temperature anomalies (terciles) in week 3 in the northern extratropics. Verification against ERA5 analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.

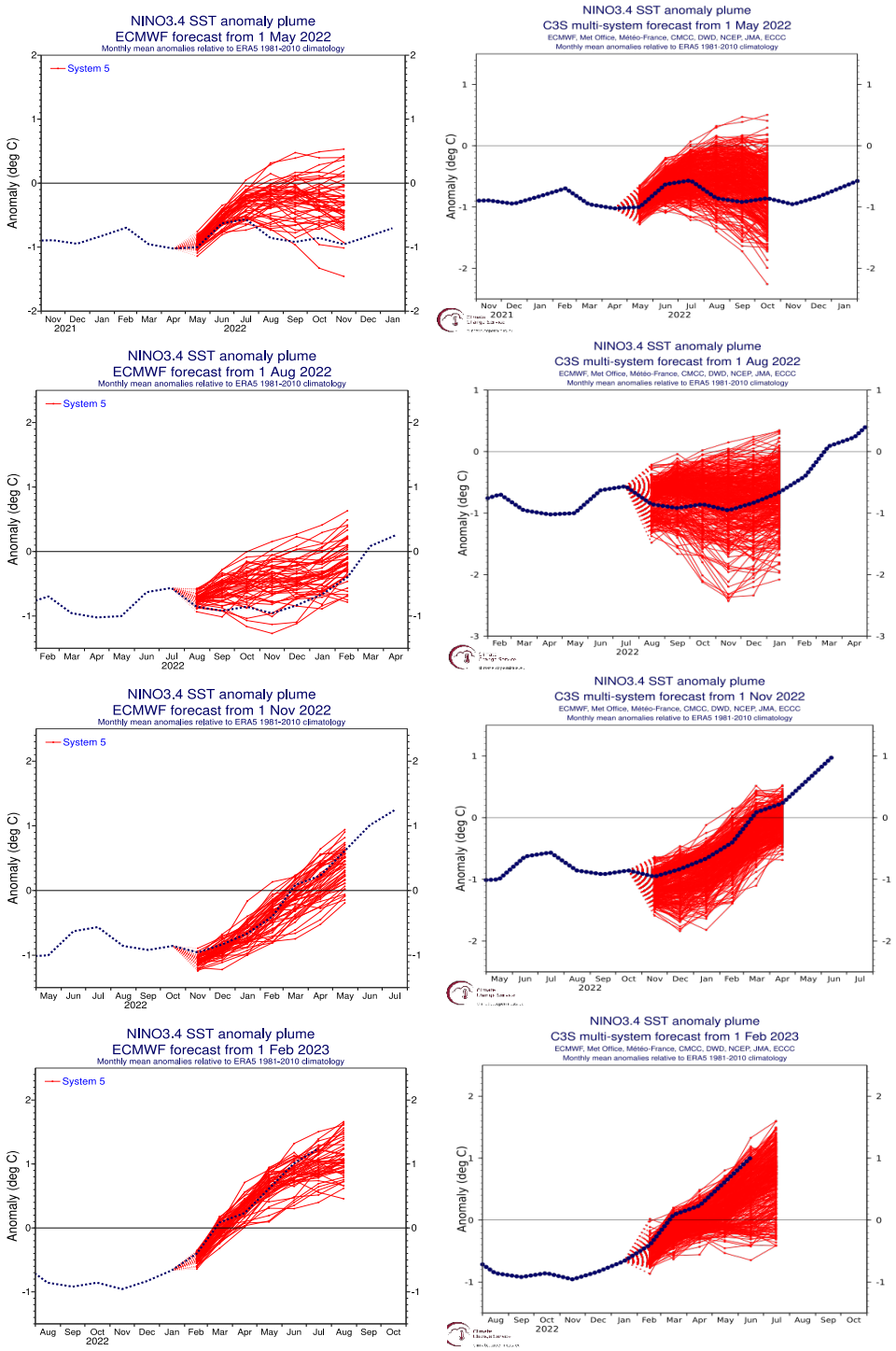


Figure 42: ECMWF System 5 (left column), and Copernicus Climate Change Service multi-model (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) May 2022, August 2022, November 2022, and February 2023. The red lines represent the 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.

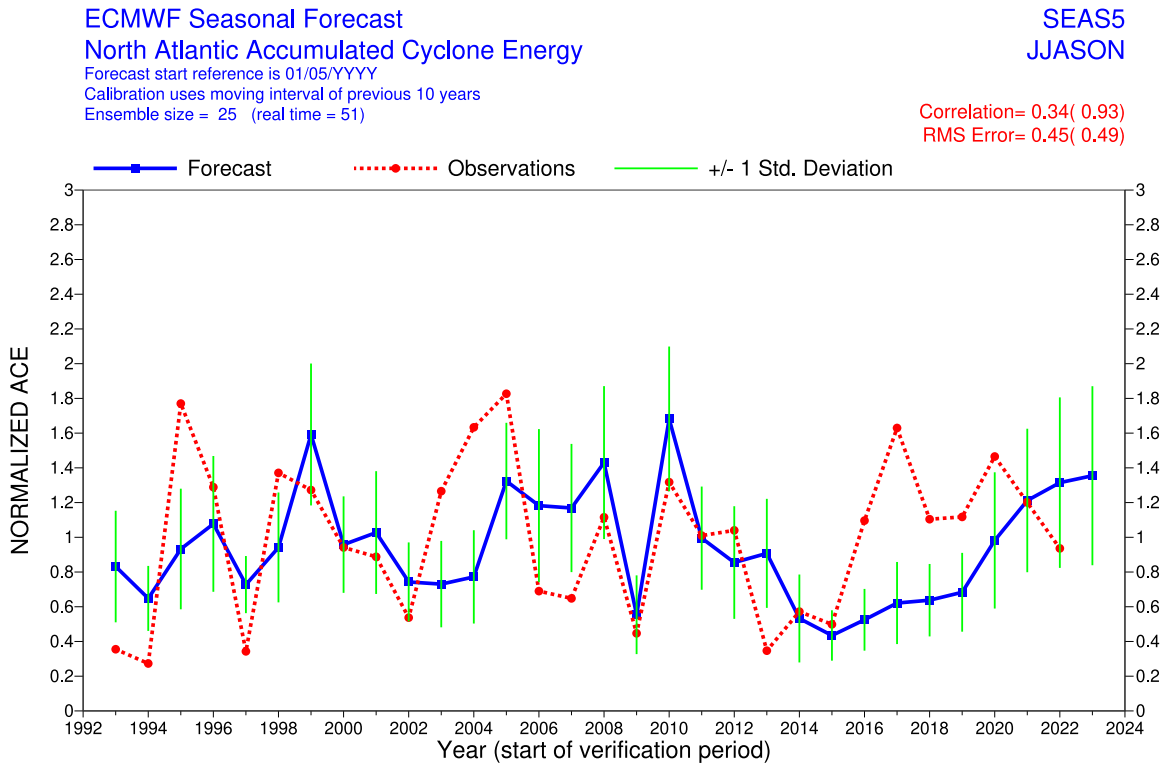


Figure 43: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1993 to July–December 2023. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (±1 standard deviation); red dotted line shows observations. Forecasts 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).

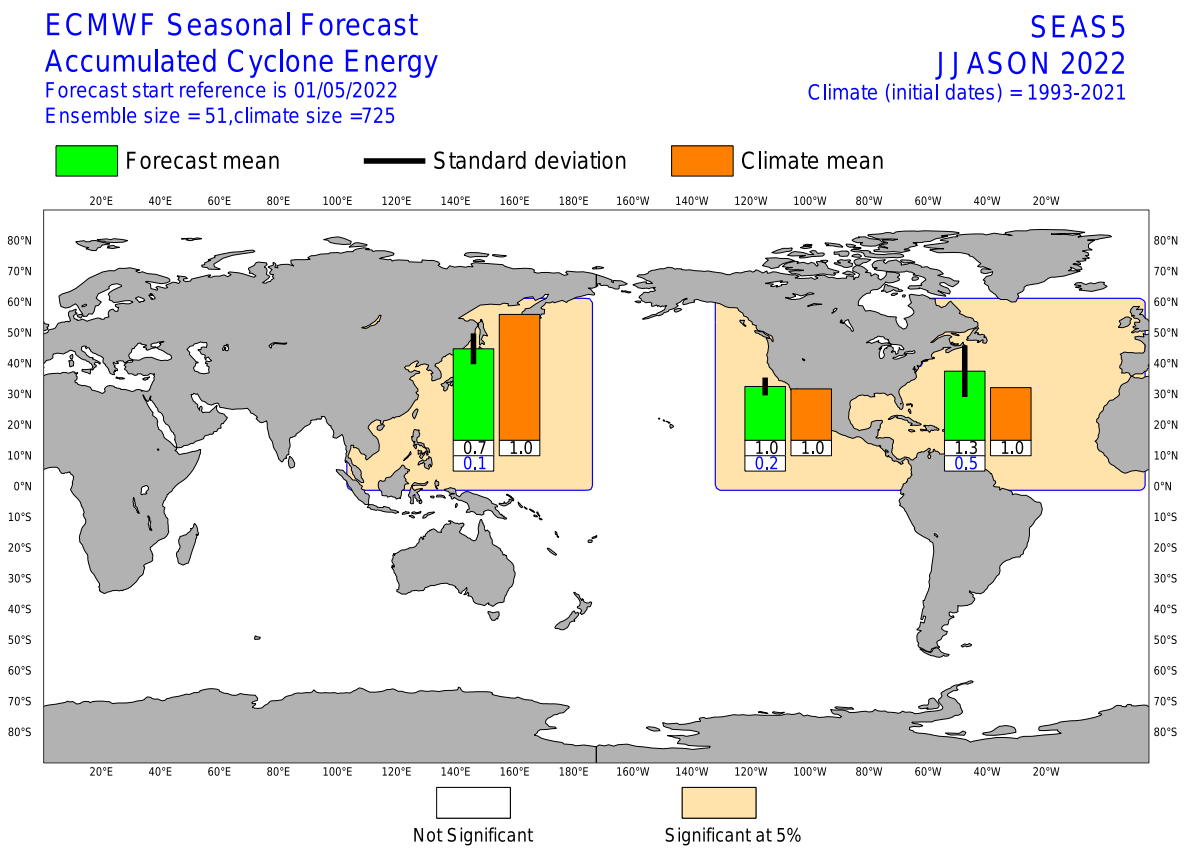


Figure 44**:** Forecast of tropical storm accumulated cyclone energy (ACE, normalized) issued in May 2022 for the six-month period June–November 2022. Green bars represent the forecast ACE in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black 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.

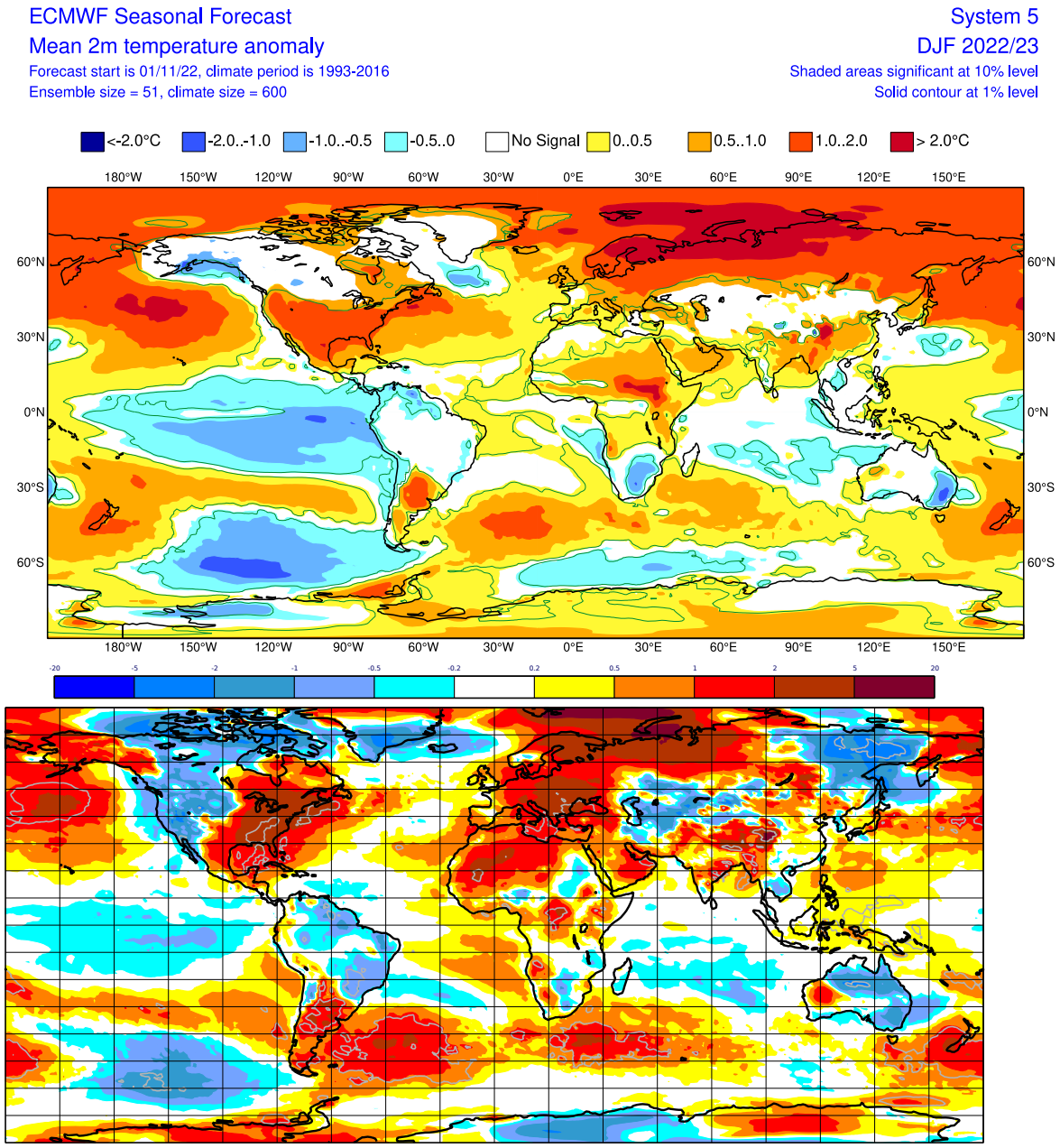


Figure 45**:** Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2022 for DJF 2022/23 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

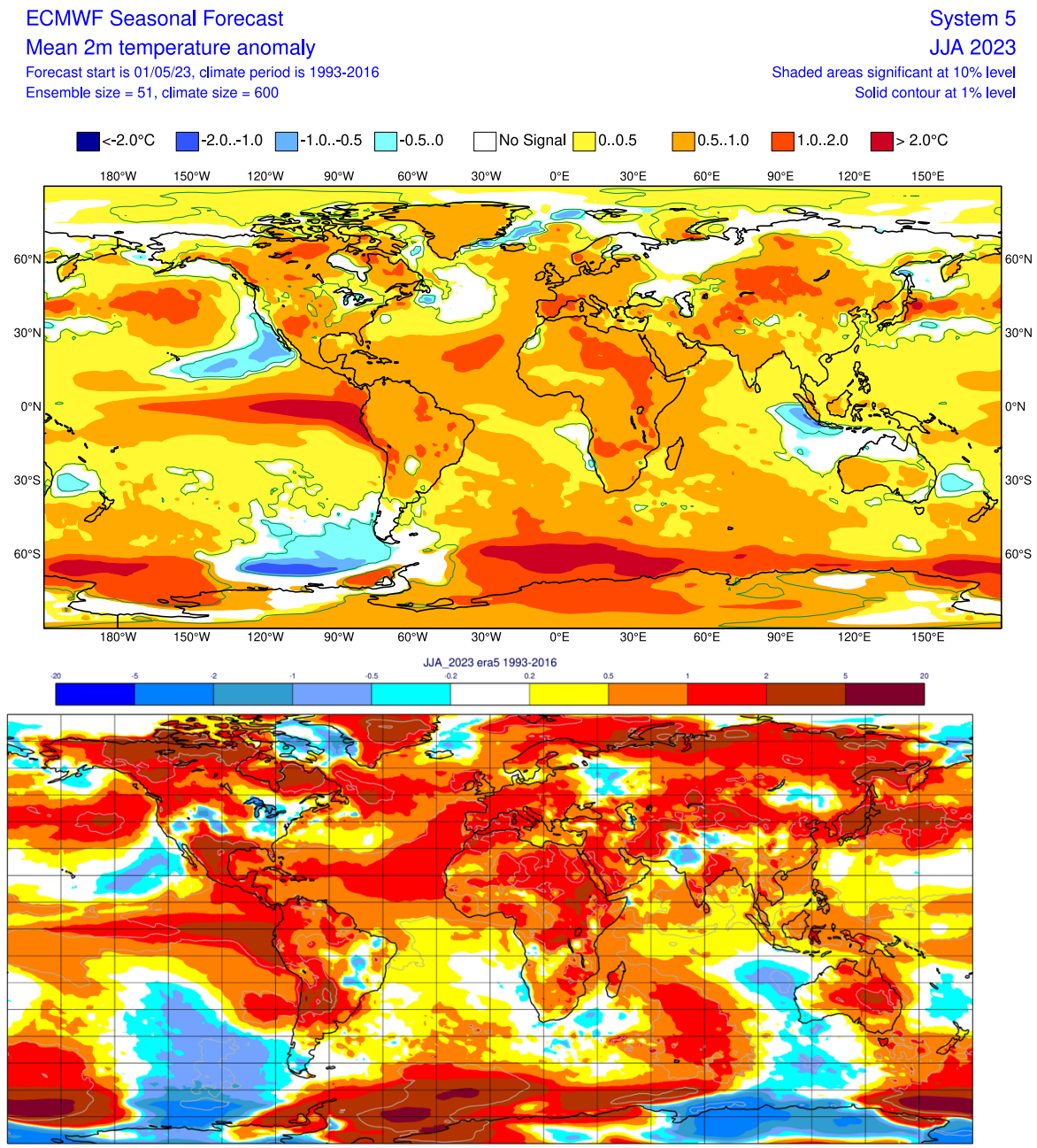


Figure 46: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2023 for JJA 2023 (upper panel) and verifying analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

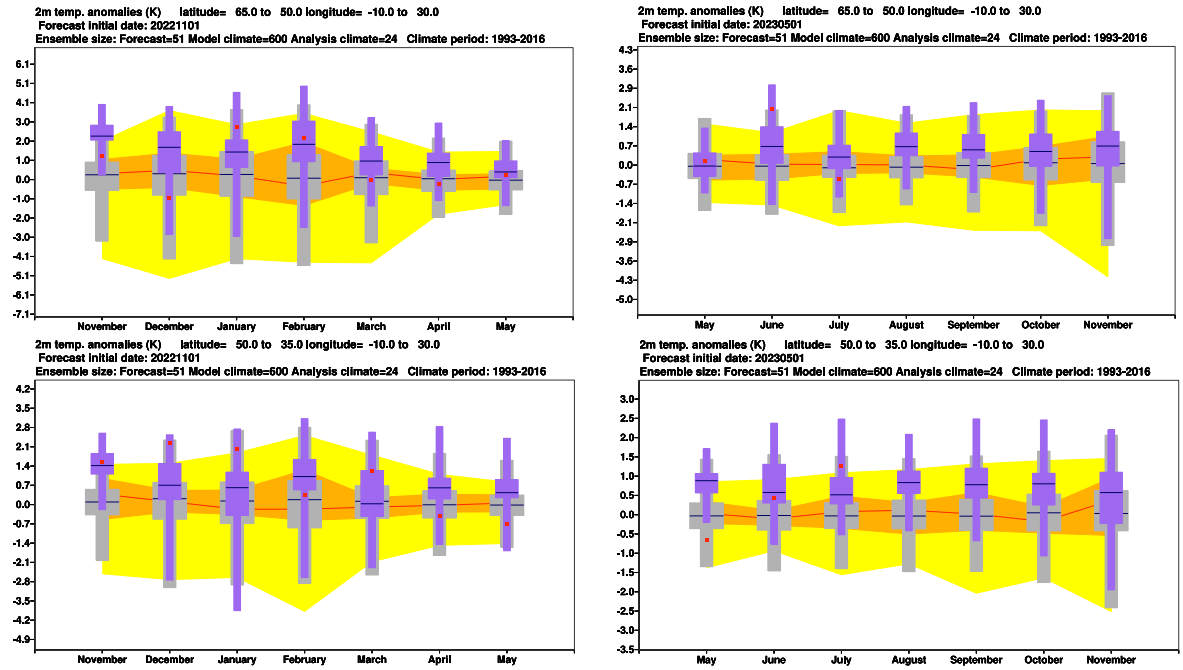


Figure 47: Verification of long-range forecasts of 2 m temperature anomalies from November 2022 for DJF 2022–23 (left panels) and from May 2023 for JJA 2023 (right panels) for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 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 18), 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 18, Figure 20) 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 6 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 34) 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 11) and its inter-annual variability (Figure 15).

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 39). Figure 39 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 40.

The comparison of spread and skill (Figure 12 to Figure 14) 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 23, Figure 24) as this provides a positively oriented skill score.

The ensemble precipitation forecasts are evaluated with the CRPSS (Figure 23, Figure 24). Verification is against the same set of SYNOP observations as used for the deterministic forecast.

For other weather parameters (Figure 25 to Figure 28), 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

Arduini, G., G. Balsamo, E. Dutra, J.J. Day, I. Sandu, S. Boussetta et al., 2019: Impact of a multi-layer snow scheme on near-surface weather forecasts. Journal of Advances in Modeling Earth Systems, 11, 4687–4710. https:// doi.org/10.1029/2019MS001725

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.

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

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.

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

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.