

938

Evaluation of ECMWF forecasts

Thomas Haiden, Martin Janousek,
Frederic Vitart, Michael Maier-Gerber,
Cathy Wing Yi Li, Fernando Prates,
Matthieu Chevallier

September 2026

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

© Copyright 2026

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. The content of this document is available for use under a Creative Commons Attribution 4.0 International Public License. See the terms at <https://creativecommons.org/licenses/by/4.0/>.

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 adopted by ECMWF in collaboration with its member states monitor the evolution of various aspects of forecast skill. The report gives updates on these scores, as well as supplementary scores to provide a more complete assessment of forecast skill. The implementation of IFS Cycle 50r1 in May 2026 has led to further 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. ECMWF is actively working on this aspect, but the gap has not yet been fully closed. The data-driven forecasting system AIFS is included in many of the plots, showing superior skill in most upper-air and surface parameters. In the sub-seasonal forecast range, increasing skill of the IFS above persistence is being observed. On the seasonal timescale, the change in 2025-26 of Pacific SSTs from La Niña to a strong El Niño was well predicted by SEAS5, but cold temperature anomalies over continental areas, especially Eurasia, remain difficult to predict. This report also includes scores from higher-resolution forecasts (DestinE Continuous Extremes Digital Twin), and 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. A major model upgrade in May 2026 has brought further improvements in forecast skill. This report mainly reports on the skill of the models 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, ECMWF is working on improving those as well. Scores of ECMWF's machine-learning forecasts (AIFS) are included in many of the plots and often show higher skill than the physics-based IFS. For lead times of several weeks ahead, the added value of forecasts relative to persistence appears to be growing. In seasonal prediction, the change of Pacific SSTs in 2025-26 from La Niña to strong El Niño conditions was well predicted. This report also includes scores from higher-resolution forecasts (DestinE Continuous Extremes Digital Twin) and atmospheric composition forecasts from the Copernicus Atmosphere Monitoring Service (CAMS).

1	INTRODUCTION	3
2	IFS AND AIFS UPGRADES	3
2.1	IFS CYCLE 50R1 OVERVIEW	3
2.2	IFS CYCLE 50R1 ASSIMILATION CHANGES	4
2.3	IFS CYCLE 50R1 USE OF OBSERVATIONS.....	5
2.4	IFS CYCLE 50R1 MODEL CHANGES	5
2.5	IFS CYCLE 50R1 METEOROLOGICAL IMPACT	6
2.6	CY50R1 CHANGES IN CAMS	7
2.7	AIFS-SINGLE v2.....	8
2.8	AIFS-ENS v2.....	9
3	VERIFICATION OF UPPER-AIR MEDIUM-RANGE FORECASTS.....	9
3.1	ECMWF SCORES	9
3.2	WMO SCORES - COMPARISON WITH OTHER CENTRES.....	11
3.3	AIFS.....	12
3.4	CAMS SCORES	13
4	WEATHER PARAMETERS AND OCEAN WAVES	13
4.1	WEATHER PARAMETERS – HIGH-RESOLUTION AND ENSEMBLE.....	13
4.2	OCEAN WAVES	15
5	SEVERE WEATHER.....	16
5.1	EXTREME FORECAST INDEX (EFI)	16
5.2	TROPICAL CYCLONES	17
6	SUB-SEASONAL AND SEASONAL FORECASTS.....	18
6.1	SUB-SEASONAL FORECAST VERIFICATION STATISTICS AND PERFORMANCE.....	18
6.2	SEASONAL FORECAST PERFORMANCE	18
6.2.1	<i>Seasonal forecast performance for the global domain.....</i>	<i>18</i>
6.2.2	<i>Tropical sea-surface temperature (SST) forecasts</i>	<i>19</i>
6.2.3	<i>Tropical storm predictions from the seasonal forecasts.....</i>	<i>19</i>
6.2.4	<i>Extratropical seasonal forecasts</i>	<i>20</i>
7	FIGURES	21
A.1	DETERMINISTIC UPPER-AIR FORECASTS	67
A.2	PROBABILISTIC FORECASTS	67
A.3	WEATHER PARAMETERS	68
	REFERENCES.....	69

1 Introduction

This report presents a summary of verification results from ECMWF's physics-based and machine-learning operational forecasting systems IFS and AIFS, as well as Copernicus Atmospheric Monitoring Service (CAMS) forecasts. For the IFS, evaluated forecast lead times range from short range to seasonal. Scores of higher resolution (4.4 km) forecast runs from the global continuous Extremes Digital Twin (Extremes DT) developed within Destination Earth (DestinE) are included as well.

On 12 May 2026, IFS Cycle 50r1 became operational, alongside an update of the ML forecasting system to AIFS v2. A brief description of the main changes to IFS and AIFS and their impact on the forecast is given in sections 2 and 3, respectively.

Note that for the IFS deterministic run, the abbreviation 'IFS-CF' is being used throughout this report since HRES and ensemble control have become bit-identical from Cycle 49r1 onwards. For multi-year curves of scores, this means that IFS-CF denotes the HRES prior to 49r1, and the ensemble control from 49r1 onwards.

Verification results of ECMWF medium-range upper-air forecasts are presented in section 4, including some comparisons of ECMWF's forecast performance with that of other global forecasting centres. Section 5 presents the evaluation of ECMWF forecasts of weather parameters and ocean waves, while severe weather is addressed in section 6. Finally, section 7 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 (since 2020, these can be found in ECMWF Technical Memoranda 880, 884, 902, 911, 918, 931). A short technical note describing some of the scores used in this report is given at the end of this document.

Regularly updated verification results can also be found at <https://charts.ecmwf.int> by choosing 'Verification' in the menu bar on the left and

- 'Medium (15 days)' (medium range, including ocean waves),
- 'Sub-seasonal' (sub-seasonal), or
- 'Seasonal' (seasonal).

2 IFS and AIFS upgrades

2.1 IFS Cycle 50r1 overview

The following description of IFS Cycle 50r1 and its meteorological impact is taken from the ECMWF implementation page at

<https://confluence.ecmwf.int/spaces/FCST/pages/567162191/Implementation+of+IFS+Cycle+50r1>

and from an ECMWF Newsletter article by Polichtchouk et al. (2025).

IFS Cycle 50r1, which was implemented on 12 May 2026, introduces a new ocean and sea ice configuration based on the NEMO4-SI³ model (a state-of-the-art modelling framework used for

research activities and forecasting services in ocean and climate sciences – Nucleus for European Modelling of the Ocean coupled with the Sea Ice Modelling Integrated Initiative which is developed by a European consortium), alongside a new ocean data assimilation system. The level of ocean-atmosphere coupling in the four-dimensional variational (4D-Var) data assimilation system is enhanced by using outer-loop coupling, a method that increases the consistency of the ocean and atmospheric initial conditions. An additional 12-hour window ocean analysis is now running in parallel to the atmospheric analysis. Together, these changes improve the representation of ocean-sea ice-atmosphere interactions. The cycle also revises the treatment of convection with the aim of improving aspects such as inland propagation of convective precipitation. In the stratosphere, reduced vertical diffusion (small-scale mixing of air between different heights) improves representation of the quasi-biennial oscillation (QBO) of tropical zonal winds between easterly and westerly as well as humidity. The data assimilation will also be able to extract stratospheric humidity information from observations and make use of more 2-metre temperature observations. Cycle upgrades to the ensemble system include scale-selective re-centering in the Ensemble of Data Assimilations (EDA) and a revised Stochastically Perturbed Parametrizations (SPP) scheme, which represents model uncertainty, to reduce excessive 10-metre wind spread. Additional improvements include a new glacier scheme and refined wave-sea ice coupling. These developments result in more realistic coupled forecasts, better use of observations, through both assimilating more data and extracting more information from them, and reduced computational cost.

A list of individual changes to the data assimilation, changes in the use of observations, and model changes, is given below.

2.2 IFS Cycle 50r1 assimilation changes

- New ocean and sea ice ensemble analysis system
- Outer-loop coupled Data Assimilation of ocean and sea-ice
- A weak constraint formulation is extended to the boundary layer which allows for assimilation of many more T2m observations than in previous cycles
- Coupled ocean-atmosphere assimilation of microwave imagers and geostationary infrared data giving increments to ocean and sea-ice as well as upper air
- Unified moisture and temperature analysis and flow dependant background error in the land data assimilation system
- EDA configuration change: perturbed members' resolution reduction (TL399 to TL255)
- Single precision trajectories in 4DVar
- Introduction of time-varying correction in Weak Constraint 4D-Var with extension to the troposphere
- Reintroduction of stratospheric humidity assimilation from radiosondes up to 60 hPa
- An upgrade of the Radiative Transfer for TOVS (RTTOV) model to version 14, improving the simulation of satellite radiances for assimilation
- Increased number of wave observations being assimilated

2.3 IFS Cycle 50r1 use of observations

- Upper air observations
 - Allowing specific humidity increments above the tropopause constrained by high quality radiosondes
 - The length of the observation timeslots has been reduced from 30 minutes to 15 minutes, allowing for a more accurate comparison between the model and the observations.
 - Introduction of wind tracing with-ozone sensitive data
- Ocean observations
 - Activation of in situ temperature/salinity profiles (e.g. ARGO, moored buoy, ship subsurface, mammal, gliders, AXBTs etc)
 - Activation of Sea Level Anomaly (SLA) along-track products from Altika/Cryosat-2/Jason-3/Sentinel-3a/Sentinel-3b

2.4 IFS Cycle 50r1 model changes

- New ocean and sea ice model based on NEMO4-SI³, with the updates including:
 - Improved numerical schemes and physical parametrisation
 - Introduction of multi-category sea ice model with prognostic salinity and melt pond dynamics
 - Replacement of the LIM2 sea ice model with the more advanced SI³ model
 - Climatological sea ice albedo replaced with albedo diagnosed by sea ice model SI³
 - Introduction of the Generic Length Scale (GLS) turbulence scheme for better vertical mixing processes
 - Use of the extended tripolar ORCA grid (eORCA025) (grid extends further towards the South Pole)
- New ensemble analysis system for ocean and sea ice based on ORAS6 reanalysis. This is used for initialising both the forecasts and the re-forecasts.
The updates include:
 - Hourly surface forcing from ERA5 atmospheric reanalysis. This allows much more accurate representation of short-term variability, such as diurnal cycle of sea surface temperature, which is not captured with daily forcing used in earlier systems.
 - Introduction of a flow-dependent background error covariance using ensemble of data assimilations
 - First-time use of variational assimilation of sea surface temperature (SST) with flow-dependent errors.
- Wave model:
 - New waves and sea ice interaction
Wave-current interaction (refraction, Doppler shift)
 - Revised wave model bathymetry

- New glacier parametrisation scheme for ecLand component replaces the previous binary glacier mask with one that accounts for how much of the grid area is covered by ice and uses a four-layer land-ice scheme
- Snow (single layer representation) on sea ice tile – in NWP, reduces the warm bias seen in winter over the ice surface especially in cloud free situations
- Modified Stochastically Perturbed Parametrisation (SPP) configuration to reduce 10-metre wind spread in the ensemble, and related misalignment of wind speed and wave height in perturbed members
- Scale-selective EDA re-centring for ENS perturbed initial conditions to address issues with tropical cyclone initialisation arising from high-resolution EDA outer loop
- The IFS has been modified to account for the effects of solar eclipses using accurate astronomical computations of the sun's and of the moon's position
- Model physics:
 - Convection and microphysics changes to improve the propagation of precipitation from ocean across land
 - Physics call reordering with vertical diffusion now called last
 - Reduced vertical diffusion in stable conditions in the stratosphere
- Coupling:
 - Coupling of snow depth and sea ice thickness from the sea ice model to the atmosphere, to allow snow over sea ice and variable ice thickness to be represented in the atmospheric forecast model
 - Partial coupling is switched off to enable fully coupled ocean-atmosphere forecasts, meaning the atmosphere now uses sea surface temperatures (SSTs) directly from the ocean model

2.5 IFS Cycle 50r1 meteorological impact

Scorecards summarising forecast performance for the IFS control forecast and the IFS-ENS are shown in Figure 1 and Figure 2, respectively. The scorecard for the IFS control (Figure 1) is predominantly blue, indicating that the new IFS Cycle 50r1 generally improves the forecast skill across a wide range of variables and lead times. The scorecard for the ensemble (Figure 2) is more mixed but, on balance, the upgrade brings a net improvement.

Improvements in both the control forecast and ENS include:

- Forecasts of QBO improved by up to 15%, due to reduced vertical diffusion and the revised non-orographic gravity wave drag formulation.
- Temperatures and winds at 850 and 250 hPa in the tropics improved by up to 7%, with contributions from the updated aerosol climatology, revised convection scheme and reordering of physics.
- Forecasts of total cloud cover, dewpoint temperature and 10-metre winds over sea improved by 1-2%, largely due to changes in the physics of the model.

Degradations in both the control forecast and ENS include:

- 250 hPa relative humidity – evaluated using root mean square error (RMSE) and continuous ranked probability (CRPS) scores – degraded by around 2%.
- 50 hPa geopotential height (using RMSE and CRPS) degraded by up to 15%, mainly due to a shift in mean values associated with the revised aerosol climatology.
- Forecasts of 2-metre temperature in the tropics degraded by 1-2% (using RMSE and CRPS), due to a bias shift linked to changes in the physics of the model (including aerosol climatology update), resulting in slightly colder-than-observed temperatures over land.

Forecast accuracy for total precipitation and 10-metre winds has improved by 1-2% in both the control forecast and the ensemble mean, based on RMSE. The Stable Equitable Error in Probability Space (SEEPS) score also shows improvement for total precipitation in the control forecast. However, when evaluated for the ENS using CRPS, these variables show degradations of up to 3%. These degradations in 10-metre winds can largely be explained by the revision of the SPP scheme, which significantly reduces ensemble spread (by 15% for 10-metre winds), addressing the spurious wind extremes and spread present in Cycle 49r1. However, when accounting for observation uncertainty in the form of representativeness errors (Bouallège, 2020), verification shows that 10-metre wind CRPS scores are nearly comparable between Cycles 49r1 and 50r1, and the apparent degradation in total precipitation CRPS is halved.

2.6 Cy50R1 changes in CAMS

The Cy50R1 brings a wide range of upgrades to IFS-COMPO and IFS-GHG across assimilation, emissions, and model changes in processes. On the assimilation side, background errors for NWP are updated and now taken from the operational EDA. Background errors for ozone have been updated to remove spurious long-range correlations. In IFS-GHG, the EDA-based B matrix used in 49R1 has been tapered in the stratosphere for CH₄ and near the surface for CO₂. In terms of observations, Sentinel-3 AOD over ocean and NOAA-21 OMPS limb profiler ozone data have been actively assimilated in IFS-COMPO, and TROPOMI CH₄ retrievals have been actively assimilated in IFS-GHG.

In IFS-COMPO, emissions have been updated through the introduction of CAMS-GLOB-ANT-M1, a merged inventory combining regional datasets for North America (USEPA), Europe (CAMS-REG), and China (MEIC) with EDGAR 6.2 elsewhere, alongside the activation of a weekly cycle in anthropogenic emissions, online BVOC emissions simulation, and the new GFAS v1.4.2 fire emissions based on MODIS and VIIRS. Finally, several model changes have been implemented. For the aerosol developments, an equilibration time parameterization for nitrate/ammonium formation has been implemented. The hydrophilic growth for organic matter, secondary organic aerosols and fine-mode nitrate has been revised. The hydrophilic growth factor of sea salt for sedimentation has been corrected, and the sedimentation velocity has been computed online for all species using the Stokes formula. For the reactive gases developments, the heterogeneous chemistry parameterization has been updated, and the SO₂ + H₂O₂ aqueous phase reaction and the ozone dry deposition via ground resistance have been revised.

2.7 AIFS-Single v2

On 12 May 2026, the day IFS Cycle 50r1 was implemented, also the AIFS underwent an upgrade. Both AIFS-Single and AIFS-ENS were upgraded to v2. Changes to AIFS-Single include:

- The introduction of a new wave component, including 11 wave variables, marking ECMWF's first operational data-driven wave forecasts.
- The addition of a new snow variable to the existing land component.
- Improved vertical velocities, by changing parameter W from a prognostic to a diagnostic field.
- An improved representation of the stratosphere, with the addition of pressure level fields at 10hPa.
- An improved training regime, using 2 years of additional training data compared to AIFS Single v1.1.

In terms of meteorological impact, a general improvement in forecast accuracy is observed due to including more recent training data in the fine-tuning of AIFS v2. Additionally, AIFS v2 forecasts are less smooth than AIFS v1.1 forecasts. This increased forecast activity leads to small skill degradations at longer lead times for some variables.

The skill for all variables at 50hPa and 100hPa in the northern and southern hemisphere has significantly improved. The presence of 10hPa also means sudden stratospheric warmings are present in the AIFS v2 forecast. The skill at 50hPa and 100hPa in the tropics remains largely the same but will be targeted in upcoming cycles.

The vertical velocities in AIFS v2 are much more physically consistent than in AIFS v1.1. The Hadley Circulation convection cells are now properly formed, and the strength of these cells agrees with physical reasoning. It is likely other vertical flows have also improved.

ECMWF has started to issue data-driven wave forecasts for the first time, leading to a substantial improvement in medium-range skill of wave forecasts. Significant wave height forecast performance improves by around 10% compared to operational IFS wave forecasts. However, some degradation and a negative bias is observed in the Antarctic during the summer in comparison to the IFS Cycle 50r1 wave model version, linked to AIFS v2 forecasts smoothing along the sea-ice edge. This will be addressed in future developments.

ECMWF is issuing data-driven snow cover forecasts for the first time. The AIFS forecasts show improved performance compared to IFS Cycle 50r1 forecasts, with the predicted snow cover fraction closer to observations. This indicates that AIFS v2 has a more accurate representation of snow distribution across the model domain.

For 2m temperature, degradations are observed in the Arctic, primarily linked to changes in sea-ice-atmosphere coupling introduced in Cycle 50r1, which impacts the initial conditions given to AIFS v2 and the limited amount of Cycle 50r1 data means the model has not yet fully learnt this new coupling. In particular, AIFS v2 has seen very limited examples of the new

wintertime sea-ice coupling in the Arctic, meaning these impacts are most pronounced during Arctic winter conditions. This degradation is expected to disappear when additional 50r1 data is included in future training cycles.

2.8 AIFS-ENS v2

The operational release of AIFS-ENS v2 introduced scientific and architectural changes to the AIFS ensemble model.

- The introduction of a new wave component, including 11 wave variables, marking ECMWF's first operational data-driven wave forecasts.
- The addition of a new snow variable to the existing land component.
- The addition of 2 variables already present in AIFS-Single (vsw and cp), to harmonise the two models.
- An improved representation of the stratosphere, with the addition of pressure level fields at 10hPa.

Architectural changes include:

- Replacing the afCRPS loss function used in AIFS ENS v1 with a multi-scale loss. See <https://arxiv.org/pdf/2506.10868>.
- Imposing the same variable bounding as is used by AIFS-Single, improving physical consistency.
- Revised graph features, utilising more edges in the decoder.
- An improved training regime, using 1 years of additional training data compared to AIFS ENS v1.

3 Verification of upper-air medium-range forecasts

3.1 ECMWF scores

Figure 3 shows the evolution of the skill of the IFS control forecast (IFS-CF) 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 (middle panel), the 12-month averaged score has been consistently high during the reporting period. In Europe, scores have been slightly lower than before, mostly due to the absence of individual months with high predictability which were more frequent in previous years. The positive effect of IFS Cycle 49r1 (implemented on 12 Nov 2024) is masked in these plots by a general slight decrease in upper-air predictability in the last two years but becomes apparent when compared to ERA5 performance (see Figures 12 and 32). Any impact from IFS Cycle 50r1 (implemented on 12 May 2026) is not yet expected to be clearly visible in these timeseries as the period covered is still too short.

Also included in Figure 3 are shorter curves showing the same scores for AIFS-Single. These are on the order of half a day of lead time above those of the IFS and otherwise exhibit similar intra- and interannual variations.

A complementary measure of performance is the root mean square error (RMSE) of the forecast. Figure 4 shows RMSE for both extratropical hemispheres of the six-day forecast, and the persistence forecast as a reference. In both hemispheres the 12-month running mean RMSE of the six-day forecast has reached its lowest value during the last year of the time series.

Figure 5 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. It measures the inconsistency between successive 12 UTC forecasts for the same verification time. For the northern hemisphere a drop is apparent in 2024, but it appears to be only partly due to Cycle 49r1, as values increase slightly again in 2025.

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 6. For IFS, there has been no significant change in the reporting period. AIFS has substantially larger errors than IFS at 50 hPa, however there is some improvement expected to come from AIFS v2. Comparison of IFS with other centres in terms of 100 hPa temperature scores (Figure 7, top panel) shows that ECMWF maintains a substantial lead. The centre and bottom panels in Figure 7 show IFS-CF stratospheric temperature RMSE skill for a range of stratospheric levels. At 30 and 50 hPa, errors have increased after they reached their lowest values in 2024, while at 100 hPa they show little change in the extratropics and a decrease in the tropics.

ENS performance is illustrated in Figure 8, which shows the evolution of the continuous ranked probability skill score (CRPSS) for 850 hPa temperature over Europe and the northern extratropics. The recent drop in Europe is related to decreasing atmospheric predictability as already mentioned in the deterministic verification (Figure 3). For the northern extratropics as a whole, the 12-month running average score has been consistently high, and the peak in winter 2025-26 has been the highest ever. For AIFS, as the period it has been operational is just slightly longer than 12 months, the resulting 12-month average curve is still very short, but it is included in Figure 8 in dark red, showing that it leads the IFS by about half a day around day 10.

In a well-tuned ensemble system, the RMS error of the ensemble mean forecast should, on average, match the ensemble standard deviation (spread). The ensemble spread and ensemble-mean error over the extratropical northern hemisphere for last winter, as well as the difference between ensemble spread and ensemble-mean error for the last three winters, are shown in Figure 9. While forecasts show a good match between IFS spread and error for 500 hPa geopotential in the first few days of the forecast, there has been stronger under-dispersiveness at longer lead times in 2026 than in previous years. For 850 hPa temperature, the under-dispersiveness has increased compared to previous years. AIFS has a slightly better spread-error match than IFS due to its spread being similar to IFS and its error smaller than IFS.

A good match between spatially and temporally averaged spread and error (as shown in Figure 9) is necessary but not sufficient for a well-calibrated ensemble. It also needs to 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 10 and Figure 11 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 on average only moderately skilful in ‘predicting’ the average error, resulting in curves that deviate significantly from the diagonal, while at day 6 (right panels) models are generally capturing spatio-temporal variations in error well. In the medium-range, ECMWF performs well overall, with its spread reliability closest to the diagonal. The stars in the plots mark the average values, corresponding to Figure 9, and ideally should lie on the diagonal and as 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 some of the 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 IFS-ENS, the CRPS is also computed for a ‘dressed’ ERA5 forecast. This allows for better distinguishing the effects of IFS developments from those of atmospheric variability, thus giving a clearer indication of 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 12 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 Cycle 49r1 in 2024 is clearly visible. The period covering Cycle 50r1 is still too short in these curves for a clear signal to emerge. At forecast day 10, interannual variability is larger, and the signal of improvement from 49r1 starts from a previous dip in values. However, for MSL pressure the highest values so far have been reached at this lead time.

Forecast performance in the tropics, as measured by RMS vector errors of the wind forecast with respect to the analysis, is shown in Figure 13. 12-month running average errors have slightly increased in the last couple of years both at 250 and 850 hPa. Figure 13 also shows that errors in the tropics are substantially smaller in AIFS than in IFS, especially in the medium range.

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

Figure 14 shows the 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 15 (Europe), and Figure 16 (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 highest overall performance in the short and medium range when verified against observations. DWD tends to have the lowest errors among the models at day 1 for some parameters, although only slightly lower than ECMWF.

The WMO model intercomparison for the tropics is summarised in Figure 17 (verification against analyses) and Figure 18 (verification against observations), which show vector wind errors for 250 hPa and 850 hPa. When forecasts are verified against each centre's own analysis, ECMWF does not lead at 850 hPa at shorter lead times (T+48). In the tropics, verification against analyses (Figure 17) is more sensitive to the details of the analysis method than in the extratropics. Smoother analyses (and forecasts) can help reduce the RMSE. When verified against observations (Figure 18), smoothness is less of an advantage, and the ECMWF forecast has the smallest overall error.

3.3 AIFS

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 IFS-CF analyses for both training and as initial condition (Lang et al. 2024). ECMWF's deterministic ML forecasting system AIFS-Single, which has become fully operational on 25 February 2025, is now routinely evaluated alongside the IFS and included in IFS plots throughout this report (see e.g. Figures 3, 6, 8, 9, 19, 20, 21, 46). Figure 19 shows that the upper-air skill of AIFS exceeds that of the IFS by an amount comparable to the IFS lead over other centres' physics-based models.

In addition to smaller forecast errors, AIFS has been found to provide generally less jumpy forecasts. It also produces improved tropical cyclone track predictions compared to the IFS (Figure 46).

Figure 20 compares the decrease of 500 hPa geopotential ACC with lead time between the IFS-CF and AIFS-Single, as well as the ensemble means of IFS and AIFS. As expected, the ensemble means drop of less sharply than the single runs, and the AIFS curves are above their IFS counterparts. The difference between the AIFS and IFS ensemble means is smaller than the difference between their single runs.

Figure 21 shows the long-term evolution of upper air deterministic skill in the northern extratropics for various global centres, for ECMWF's IFS-CF, AIFS-Single, and ERA5 as a reference. The lead of the AIFS on the IFS is about one-third of a forecast day, similar to the lead of the IFS over the other centres.

On 1 July 2025, ECMWF put the ensemble version of the AIFS into operations. The AIFS-ENS provides 51 different scenarios at a resolution of 31 km. Figure 22 shows a comparison of CRPSS as a function of lead time between IFS-ENS and AIFS-ENS for a 12 month period, both for upper-air (850 hPa temperature) and surface (2 m temperature), illustrating the substantial benefit of AIFS-ENS, except for T850 at Day 1.

3.4 CAMS scores

The Copernicus Atmospheric Monitoring Service (CAMS) uses the same model cycle as the IFS 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. It is referred to as IFS-COMPO. Figure 19 shows that in terms of 500 hPa geopotential in the extratropics, the meteorological skill of IFS-COMPO 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 IFS-COMPO in predicting Saharan Dust episodes, wild-fire and volcanic plumes and air pollution events in Europe, North America and Asia as well as variations in Antarctic stratospheric ozone. For instance, the upgrade in anthropogenic emissions to CAMS-GLOB-ANT-M1 in Cy50R1 generally improves forecast performance of primary species such as NO₂, SO₂ and CO. The upgrade in ozone deposition scheme lowers the forecasted surface ozone concentrations, leading to overall improvement in biases in North America and Europe. There is also a routine intercomparison of air quality forecast for North America of several global and regional models led by Environment and Climate Change Canada as part of WMO's Global Air Quality Forecasting and Information System (GAFIS) initiative. Their latest report (Presseau et al., 2026) shows that IFS-COMPO is competitive with other air quality forecasts in predicting surface NO₂, ozone, and PM_{2.5} (Figure 23). Routine verification of the IFS-COMPO atmospheric composition forecast is carried out by the CAMS Evaluation and Quality Assurance (EQA) with quarterly reports being published at <https://atmosphere.copernicus.eu/eqa-reports-global-services>. Scores for ozone columns and AOD shown in the latest EQA report (Benedictow et al., 2026) indicate a gradual improvement of the model forecasts from 48r1 (Figure 24).

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 25. 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 IFS-CF 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 CRPSS and DSS 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.

While SEEPS deterministic precipitation forecast skill undergoes strong interannual variations (mirrored by the ERA5 curve) due to changing atmospheric conditions (upper left panel in

Figure 25), the difference between IFS-CF and ERA5 scores (upper right panel in Figure 25) shows the highest level of skill so far.

Probabilistic precipitation headline scores CRPSS and DSS are shown in the bottom panels in Figure 25. 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. Both curves are subject to considerable noise, with no clear trend apparent over the last couple of years.

ECMWF performs a routine comparison of SEEPS deterministic precipitation forecast skill between global centres. Results show the IFS-CF leading relative to other centres from day 2 onwards (Figure 26).

While ECMWF has access to some centre's deterministic precipitation forecasts at their native resolution (to produce comparisons such as for SEEPS in Figure 26), the same is not available for their ensemble forecasts. Ensemble skill comparisons need to utilise data from the TIGGE archive, where other centres provide their forecasts at lower resolution, typically at 0.25 or 0.5 degrees. Figure 27 (upper panel) shows such a comparison between the IFS-ENS downgraded to a 0.25 deg grid and other centres. Note that the exact method of how other centres upscale their forecasts to 0.25 or 0.5 deg may differ from the method used here for the ECMWF data (grid-box mean). This may adversely affect spread, and therefore CRPSS, especially in the short range. For precipitation, while ECMWF is very close to DWD from day 5 onwards, there appears to be a weakness compared to other centres in terms of ensemble precipitation at shorter lead times. For 2 m temperature (Figure 27, lower panel), ECMWF is second to DWD in the short range and leading in the medium range.

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 28-31. 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, 2m dewpoint and total cloud cover (Figures 28, 29 and 30) there are no significant changes in either standard deviation or bias. Regarding total cloud cover, it is worth noting that the top of atmosphere shortwave radiation, which is determined largely by cloudiness, has further improved (see Figure 33 discussed below). The night-time positive wind speed bias, mostly due to insufficient calming of the wind in the presence of surface inversions, has remained low (Figure 31).

ERA5 is useful as a reference forecast for the IFS-CF, as it allows filtering out some of the effects of atmospheric variations on scores. Figure 32 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. There are substantial improvements in the non-systematic errors of both upper-air and surface parameters in 2024 from Cycle 49r. The effect of Cycle 50r1 in 2026 is not yet clearly visible, but the skill of T850 has reached its highest value so far.

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 33 shows how the 5-day forecast of the TOA net shortwave radiation has improved over time. ERA5 is included for comparison, which shows that IFS-CF errors have reached their lowest values relative to ERA5 and, for the northern extratropics, also in absolute terms.

The fraction of large 2 m temperature errors in the IFS-ENS is one of ECMWF's headline scores. In this context, an error is defined as large whenever the CRPS exceeds 5 K. Figure 34 shows that in the annual mean (red curve) this fraction has decreased from about 7% to 3.5% 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 strong signal of improvement due to Cycle 49r1 can be seen both in winter and summer, as well as the annual mean. This improvement is primarily due to a reduction of errors during winter.

An analogous measure of the skill in predicting large 10 m wind speed errors in the IFS-ENS is shown in Figure 35. 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%, which represents a relative improvement of about 25%. As for 2 m temperature, we see a further substantial improvement due to Cycle 49r1 and a slight increase thereafter due to higher wintertime errors.

A comparison of the RMSE of 2 m temperature and 10 m wind speed at forecast day 5 between AIFS-Single and IFS-CF is shown in Figure 36. The strong improvement of AIFS-Single in winter 2023-24 has put it into a clear lead for both 2 m temperature and 10 m wind speed.

Since summer 2023, DestinE Extremes Digital Twin (DT) 4.4 km forecasts out to day 5 are produced daily. Figure 37 shows time-series of RMSE of 2 m temperature and 10 m wind speed in Europe and the northern extratropics for the DestinE Extremes DT and IFS-CF. Overall, DestinE has smaller errors than IFS-CF by a few percent, and the implementation of Cycle 49r1 had a slightly larger beneficial effect on DestinE than on the IFS-CF, e.g. for T2m. Comparison of the RMSE of 850 hPa temperature (lower right panel in Figure 38) shows that also upper-air errors in DestinE Extremes DT are slightly smaller than those of the IFS-CF.

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 39. While 10 m wind speed errors have slightly increased in the analysis and short-range forecast in recent years, wave height forecast errors have reached a new minimum in the very short range and are relatively stable at other lead times. When verified against own analysis (Figure 40), scores of significant wave height have

slightly improved over the last year in the northern extratropics, and remained about constant in the southern extratropics.

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 41) ECMWF leads other centres in significant wave height and peak period. In the tropics (Figure 42), ECMWF is still performing well but does not have a clear lead.

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 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 dark green lines in the top panels of Figure 43, alongside results for other lead times up to day 7. In Europe (right panel) it has reached its highest value so far. 24-hour total precipitation scores (centre panels) have dropped slightly after reaching their highest values in 2025. For 2 m temperature, the 12-month average skill has been relatively flat, however the absolute level of skill for 2 m temperature is significantly higher than for the other two parameters, in particular at longer lead times.

5.2 Tropical cyclones

The tropical cyclone (TC) position error at day 3 of the IFS-CF is one of the two supplementary headline scores for severe weather. The average position errors for the high-resolution medium-range forecasts of all TCs (all ocean basins) are shown in Figure 44. Errors in the forecast of central pressure and propagation speed of TCs are also shown.

The IFS-CF position error (top panel, Figure 44) at day 3 has reached its 2nd-lowest value so far, slightly above 150 km. Relative to ERA5, errors have been roughly constant over the last few years. At day 5 errors are generally higher than 5-10 years ago. TC intensity errors have improved significantly relative to ERA5 due to the IFS-ENS spatial resolution upgrade in Cycle 48r1. The bottom panel of Figure 44 shows the spread and error of ensemble forecasts of TC position. After a few years of the forecast being slightly under-spread, it has been slightly over-dispersive in the last two years.

The IFS and AIFS ensemble TC forecasts are 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 45. 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 reliability and modified ROC, the most recent year shows lower skill than the previous two years, and for the standard ROC it is very similar to the previous year. The AIFS-ENS has higher reliability and modified ROC than the IFS-ENS, similar to the values IFS-ENS had in 2025.

Figure 46 shows how AIFS-Single and IFS-CF tropical cyclone position and central pressure errors compare. The top panel shows a substantial reduction of the mean position error in AIFS-Single compared to IFS-CF, equivalent to 0.5-1 forecast days. The IFS-CF on the other hand provides more accurate TC intensity forecasts than AIFS-Single, primarily due to a pronounced weak-intensity bias in AIFS-Single. The initial intensity bias in IFS gradually decreases and becomes negative, while that of AIFS-Single generally grows with lead time. This behaviour is mirrored in the wind speed curves. For the respective ensembles (IFS and AIFS, not shown), the behaviour is very similar. As in the deterministic case, the difference in position error between the ensemble means of AIFS and IFS is of the order of 0.5-1 forecast days. For track, AIFS ENS has a spread comparable to IFS ENS. Combined with the smaller mean errors, this results in overdispersion. For intensity, AIFS ENS is highly underdispersive, considerably more so than IFS ENS.

6 Sub-seasonal and seasonal forecasts

6.1 Sub-seasonal forecast verification statistics and performance

Figure 47 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 model's ability 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.

Skill has generally increased over the last 20 years both relative to persistence and in absolute terms, except for cold anomalies in weeks 3+4. The increase is more pronounced in week 2. The latest summer and winter seasons show a level of skill similar to recent years.

Because of the low signal-to-noise ratio of real-time forecast verification in the sub-seasonal lead time range (Figure 47), re-forecasts are a useful additional resource for documenting trends in skill. Figure 48 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 in this metric over the last few years.

Note that there have been important changes in the sub-seasonal forecast with 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 real-time forecast for users but cannot be shown by a headline score based on reforecast verification.

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’.

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 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

6.2.2 Tropical sea-surface temperature (SST) forecasts

The second half of 2025 and first half of 2026 saw a transition from La Niña to strong El Niño conditions. The seasonal forecast gave a good signal for this transition but initially underestimated the magnitude of El Niño (Figure 49, left column). Early in 2026 however the forecast started to indicate a very strong El Niño which, at the time of writing, appears likely to occur. The C3S multi-model ensemble (Figure 49, right column), with its naturally larger spread, gave a somewhat better estimate of the 2025 cold anomaly, but also underestimated the speed of the transition to El Niño and its 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.

6.2.3 Tropical storm predictions from the seasonal forecasts

The 2025 Atlantic hurricane season had a total of 13 named storms, including 5 hurricanes and 4 major hurricanes. Despite a below-average number of named storms and hurricanes, the season had a slightly above-normal accumulated cyclone energy (ACE) about 10% above the 10-year (2015-2024) climate average (Figure 50). This makes it a season close to climatology in terms of the number of tropical storms (climate average is about 14.5) and ACE, which is consistent with neutral ENSO conditions in the Pacific and small SST anomalies over the Atlantic. Seasonal tropical storm predictions from SEAS5 initialized on 1st May 2025 correctly indicated a number of tropical storms close to climatology (14.5 +/- 4.4) over the Atlantic and an ACE about 10% larger than climatology. Subsequent forecasts, issued in June, July and August, also predicted an Atlantic hurricane season close to normal.

Figure 51 shows that SEAS5 also predicted average activity over the eastern North Pacific and a significantly below average typhoon season over the western North Pacific (ACE of about 70% of the 2013-2024 climate average). The 2025 western Pacific typhoon season was a below average season in terms of ACE (ACE 20% below climatology) consistent with the SEAS5 seasonal prediction. The 2025 eastern Pacific hurricane season was a normal season with an ACE close to climatology, as predicted by SEAS5.

6.2.4 Extratropical seasonal forecasts

The seasonal forecast of 2 m temperature for DJF 2025-26 was not particularly skilful in terms of large-scale anomaly patterns, especially in the Eurasian sector (Figure 52). At that time there was no strong forcing from tropical Pacific SSTs, just a weak cold anomaly. For Europe a warm anomaly was predicted, which materialized in Europe's southwestern half only while in the northeastern parts a distinct cold anomaly was present. For North America the forecast was closer to the observed tri-pole pattern, with a belt of cold anomalies from Alaska through western Canada to the U.S. east coast.

The seasonal forecast of 2 m temperature in JJA 2026 (Figure 53) was overall more skilful, probably helped by the strong forcing from tropical Pacific SSTs due to the developing strong El Nino. The positive anomaly in Europe was predicted, however in the ensemble mean there was no indication that summer 2026 would become the most extreme observed so far in several European countries. There was also some indication of less heat in northeastern parts of Europe, as observed. Major differences between forecast and observations can be seen over the U.S.

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 2025-26 are shown in Figure 54 and for summer 2026 in Figure 55. Red squares indicate observed monthly anomalies. For northern Europe, the pronounced cold anomalies in Jan-Feb 2026 were not predicted. In Southern Europe there is a somewhat better match between forecast and observations, but this is mostly due to the absence of observed cold anomalies. Predictions for summer 2026 in Northern Europe show that the observed positive anomaly in June was indicated in the forecast by a long tail towards strong positive anomalies rather than a shift of the ensemble mean. For Southern Europe, the positive anomalies throughout JJA 2026 were indicated, however the magnitude of the observed August positive anomaly even exceeded the most extreme ensemble member.

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 56 and 57 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.

7 Figures



Figure 1: Summary IFS-CF score card for IFS Cycle 50r1 versus Cycle 49r1 verified by the respective analyses and observations for ~1100 forecasts initialised between May 2024 and May 2026. Grey frames indicate that results refer to the second score indicated at the top of the column.



Figure 2: Summary IFS-ENS score card for IFS Cycle 50r1 versus Cycle 49r1 verified by the respective analyses and observations for ~470 forecasts initialised between May 2024 and May 2026.

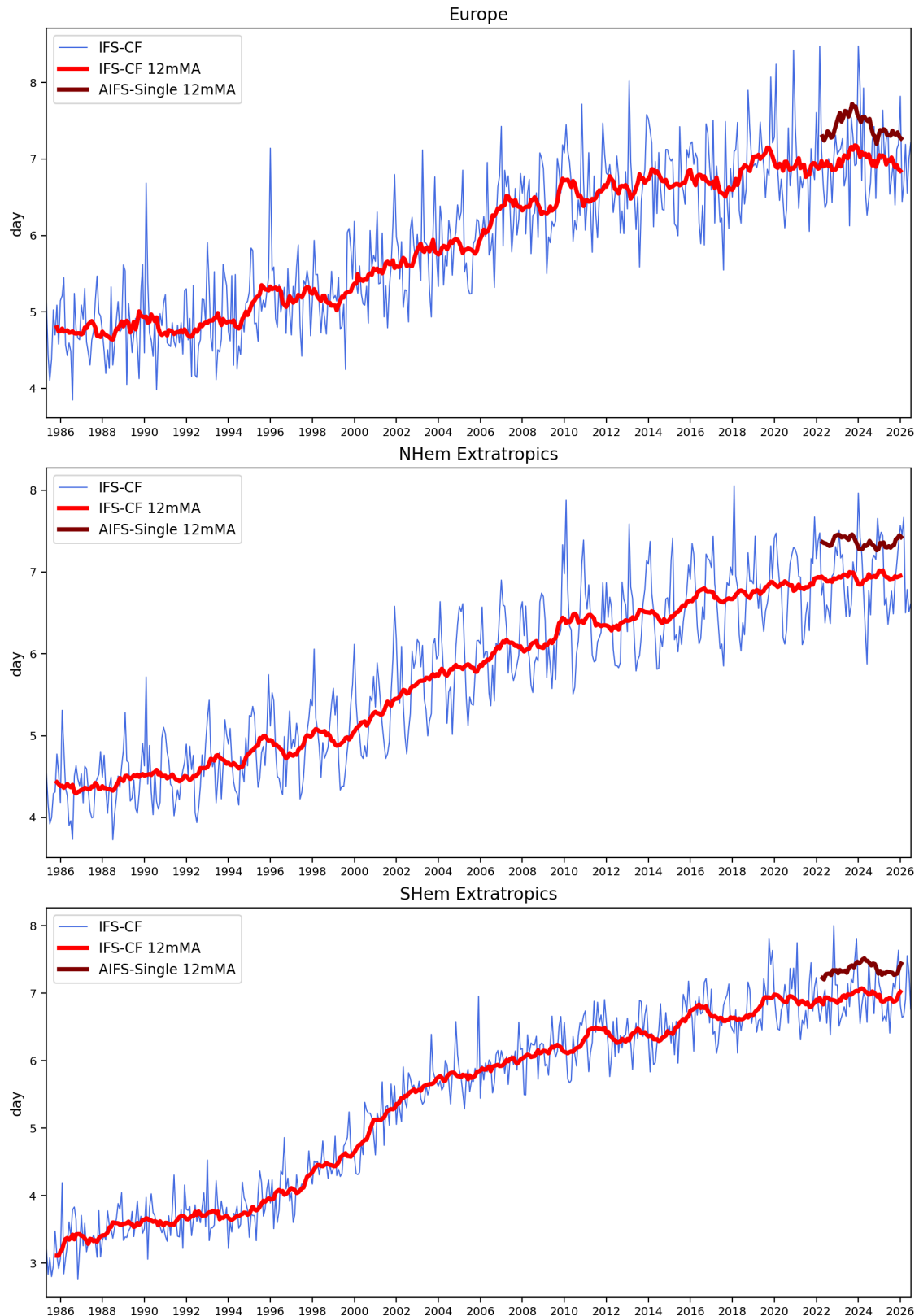


Figure 3: Primary headline score for the IFS-CF. Evolution with time of the 500 hPa geopotential height forecast performance – each point on the curves is the forecast range at which the monthly mean (blue lines) or 12-month mean centred on that month (red line) of the forecast anomaly correlation (ACC) with the verifying analysis falls below 80% for Europe (top), northern hemisphere extratropics (centre) and southern hemisphere extratropics (bottom). For reference, the corresponding 12-month running mean for AIFS-Single is shown in dark red.

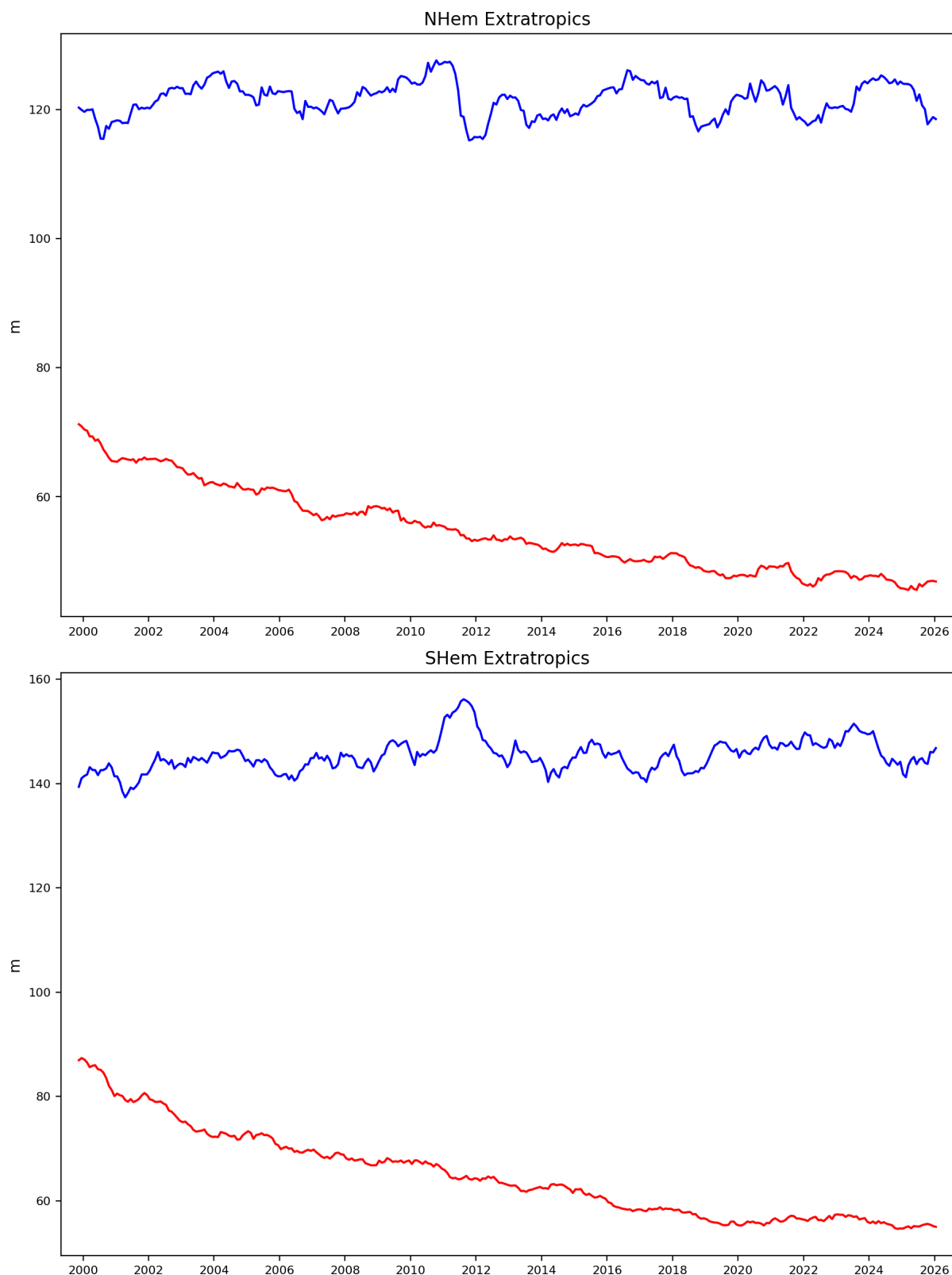


Figure 4: Root mean square (RMS) error of forecasts of 500 hPa geopotential height (m) at day 6 (red), verified against analysis. For comparison, a reference forecast made by persisting the analysis over 6 days is shown (blue). Plotted values are 12-month moving averages; the last point on the curves is for the 12-month period August 2025–July 2026. Results are shown for the northern extra-tropics (top), and the southern extra-tropics (bottom).

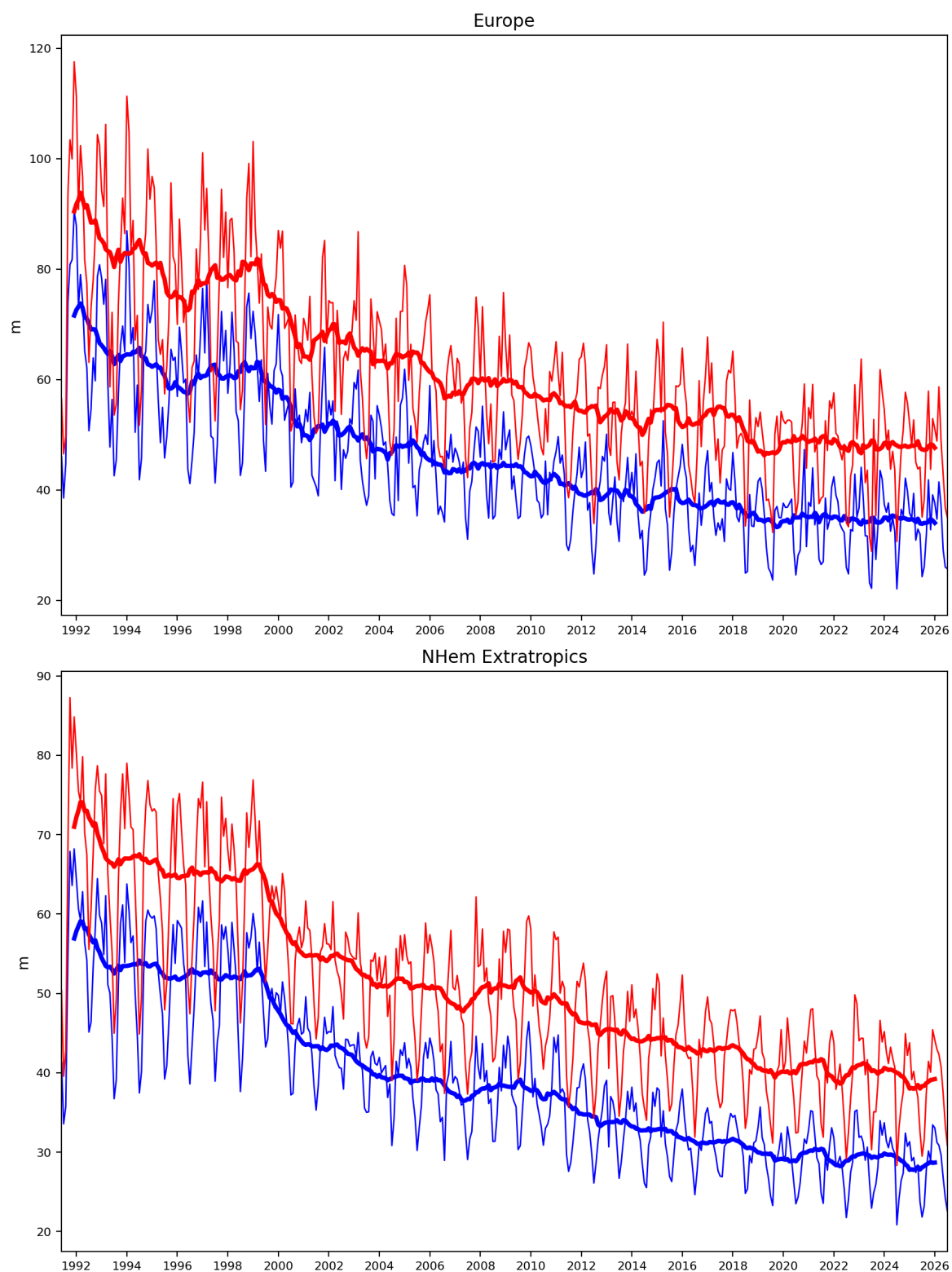


Figure 5: A measure of inconsistency of the 500 hPa height forecasts over Europe (top) and northern extratropics (bottom). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96–120 h (blue) and 120–144 h (red). 12-month moving average scores are also shown (in bold).

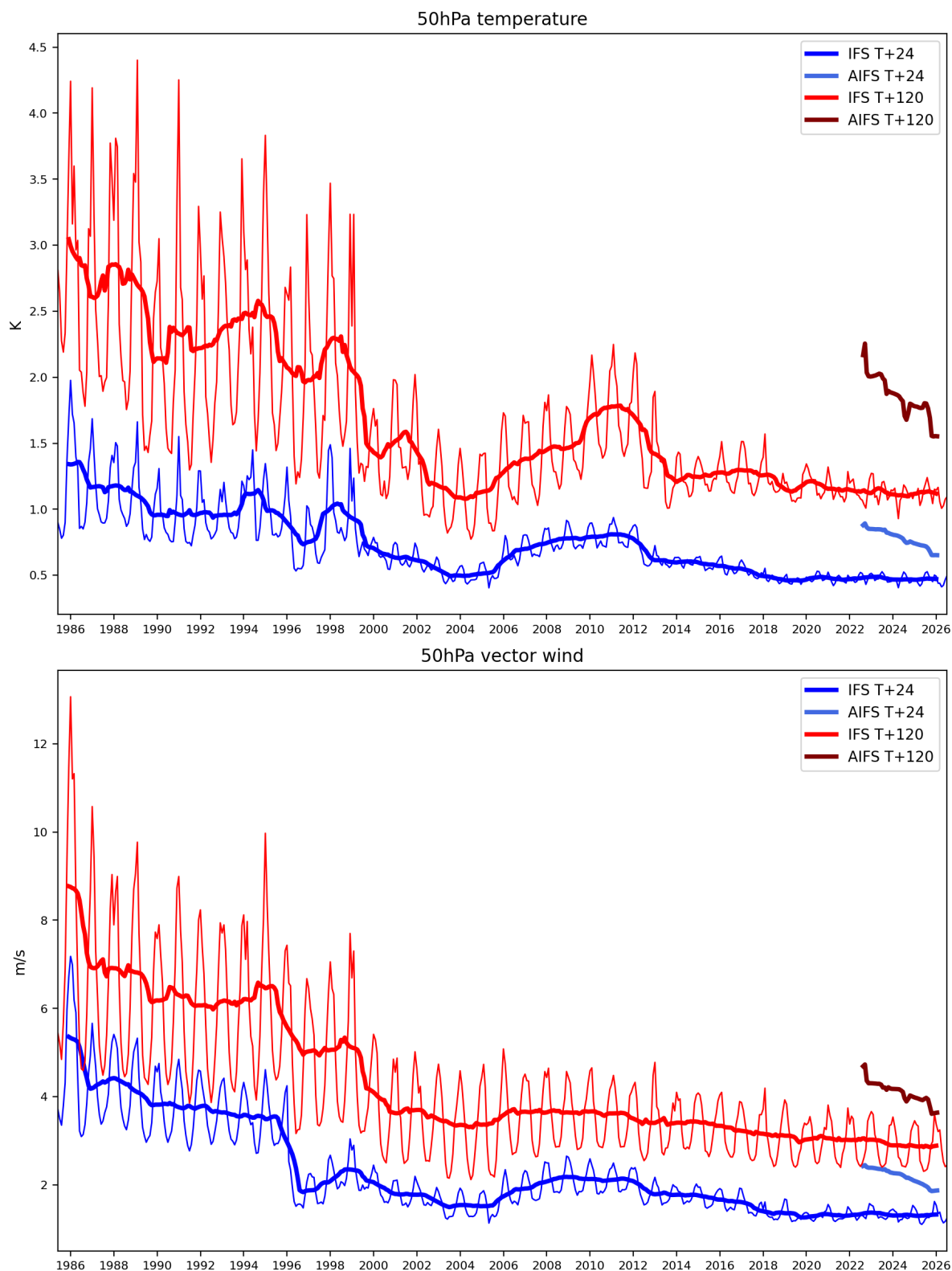


Figure 6: Model scores for temperature (top) and wind (bottom) in the northern extratropical stratosphere. Curves show the monthly average RMS temperature and vector wind error at 50 hPa for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold), including the AIFS-Single scores.

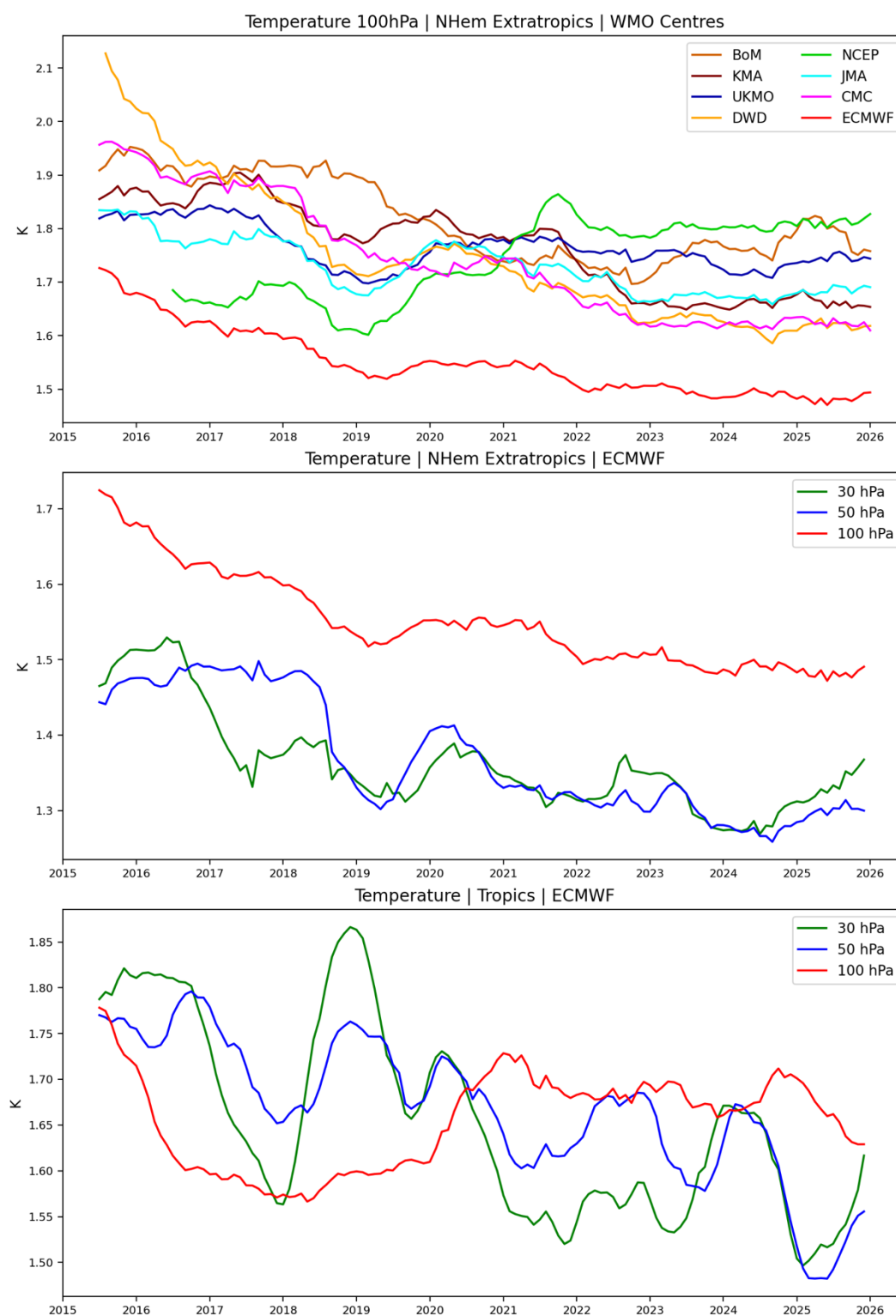


Figure 7: Stratospheric scores at a lead time of +144 h. Top: global model intercomparison of the 100 hPa temperature RMSE in the northern extratropics based on the WMO exchange of scores. Centre: RMSE of temperature at four different stratospheric levels in the northern extratropics. Bottom: same as centre, but for the tropics. Curves in all three plots are 12-month running averages.

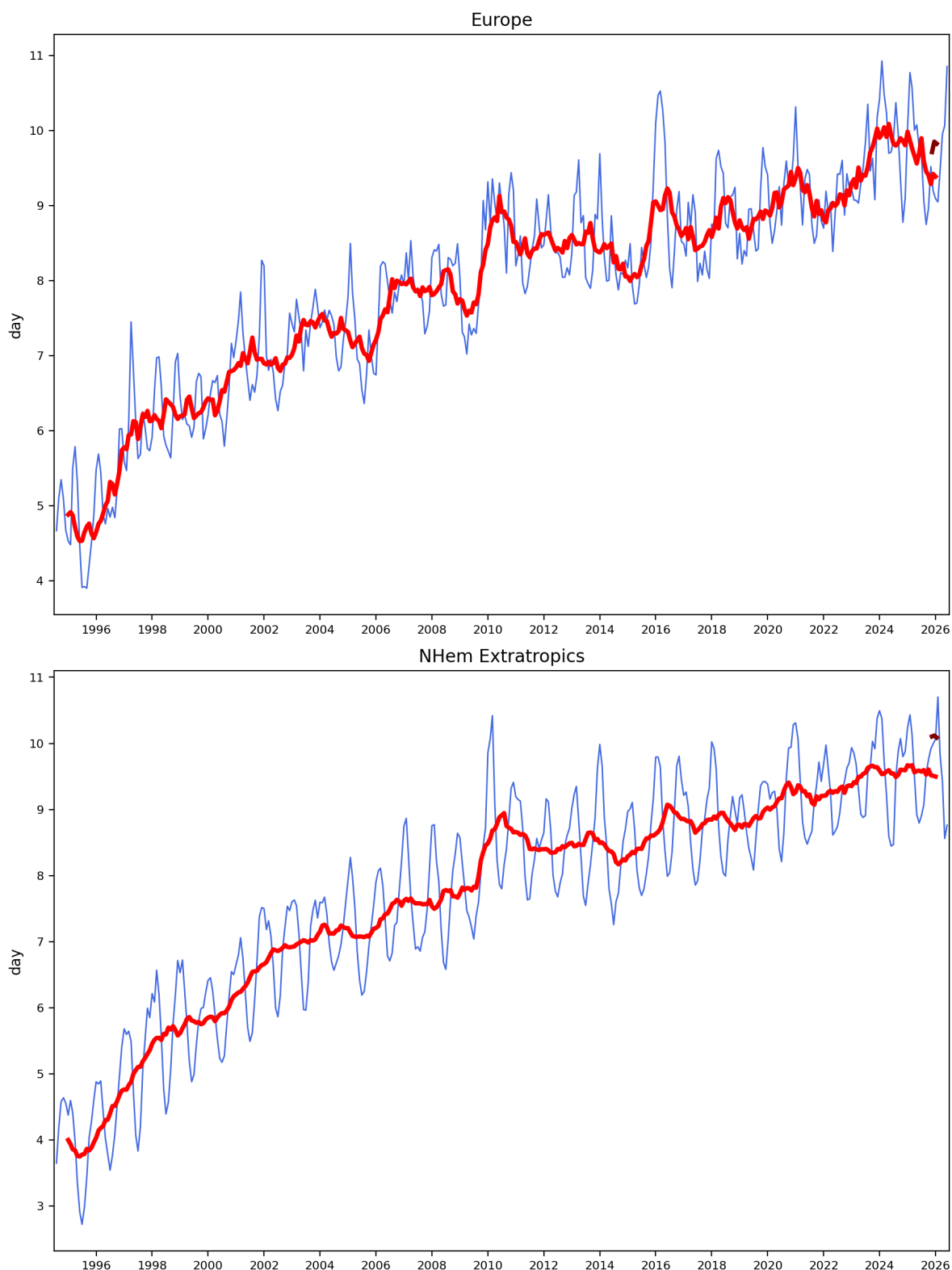


Figure 8: Primary headline score for the IFS-ENS. Evolution with time of 850 hPa temperature ensemble forecast performance, verified against analysis. Each point on the curves is the forecast range at which the 3-month mean (blue lines) or 12-month mean centred on that month (red line) of the continuous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

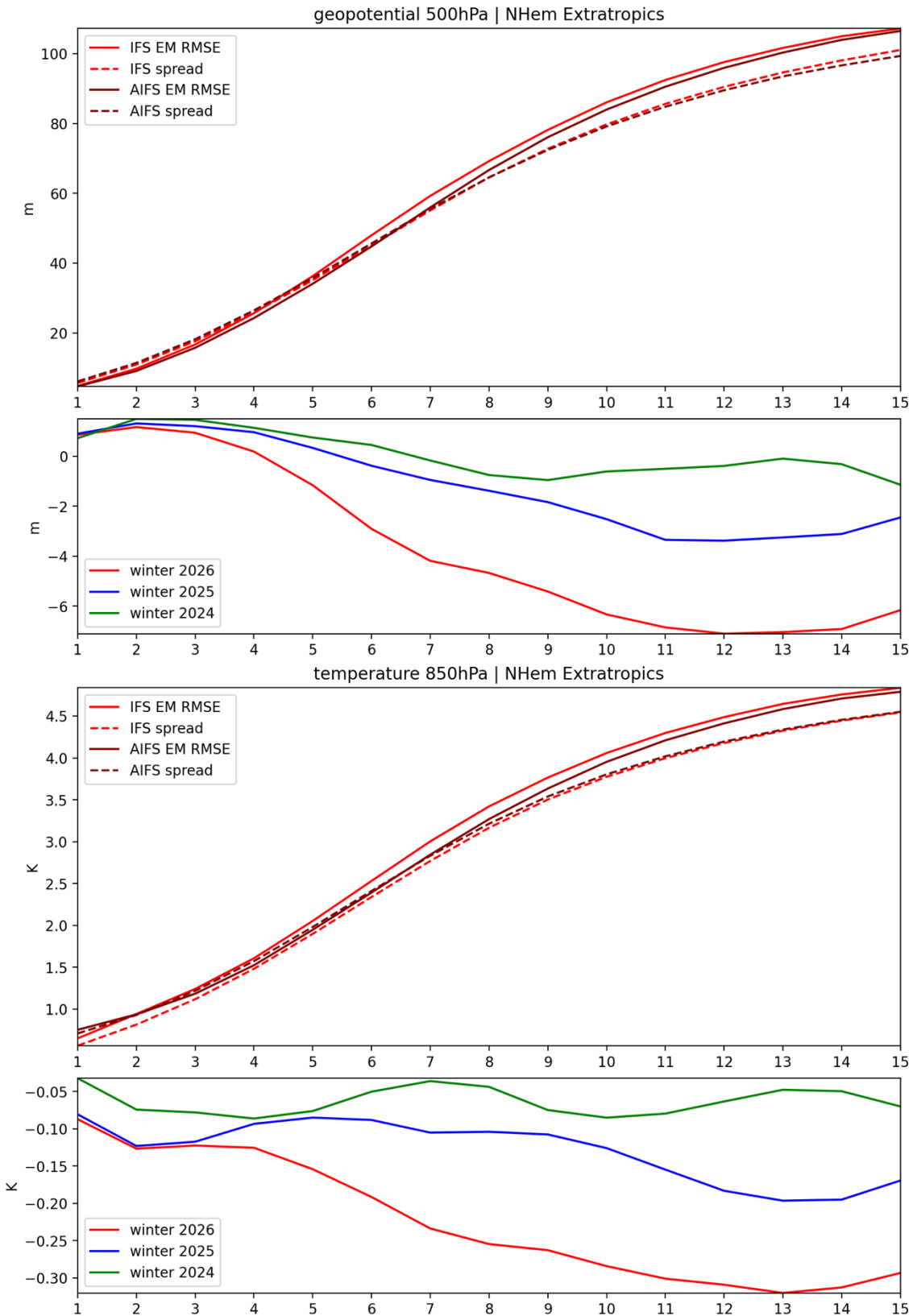


Figure 9: Ensemble spread (standard deviation, dashed lines) and RMS error of ensemble-mean (solid lines) for winter 2025–2026 (upper figure in each panel) for IFS and AIFS, and differences of ensemble spread and RMS error of ensemble mean for last three winter seasons for IFS (lower figure in each panel, 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. For reference, the corresponding scores for AIFS-ENS are shown in dark red.

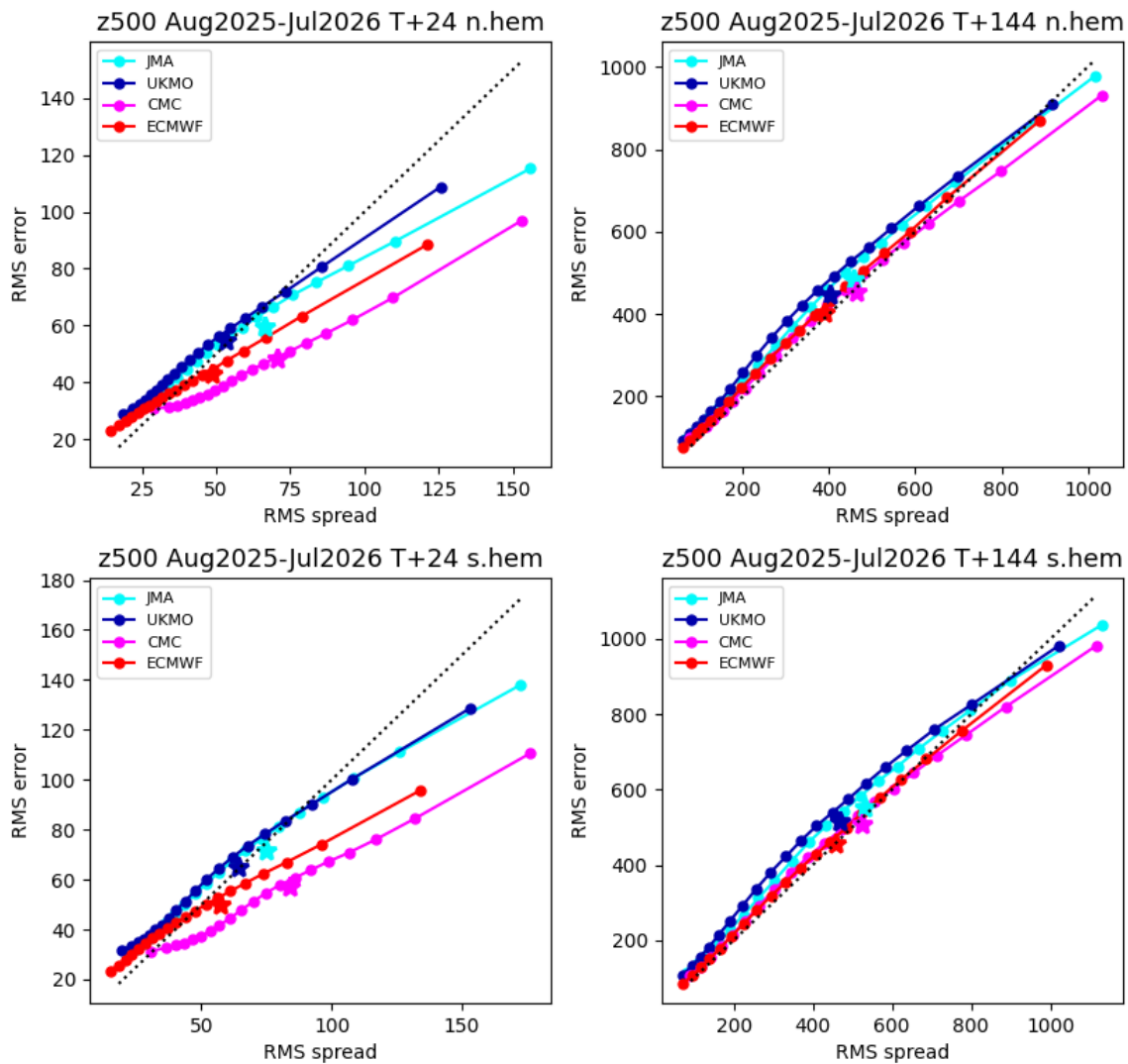


Figure 10: Ensemble spread reliability of different global models for 500 hPa geopotential for the period August 2025–July 2026 in the northern (top) and southern (bottom) hemisphere extra-tropics for day 1 (left) and day 6 (right), verified against analysis. Circles show error for different values of spread, and stars show average error-spread relationship

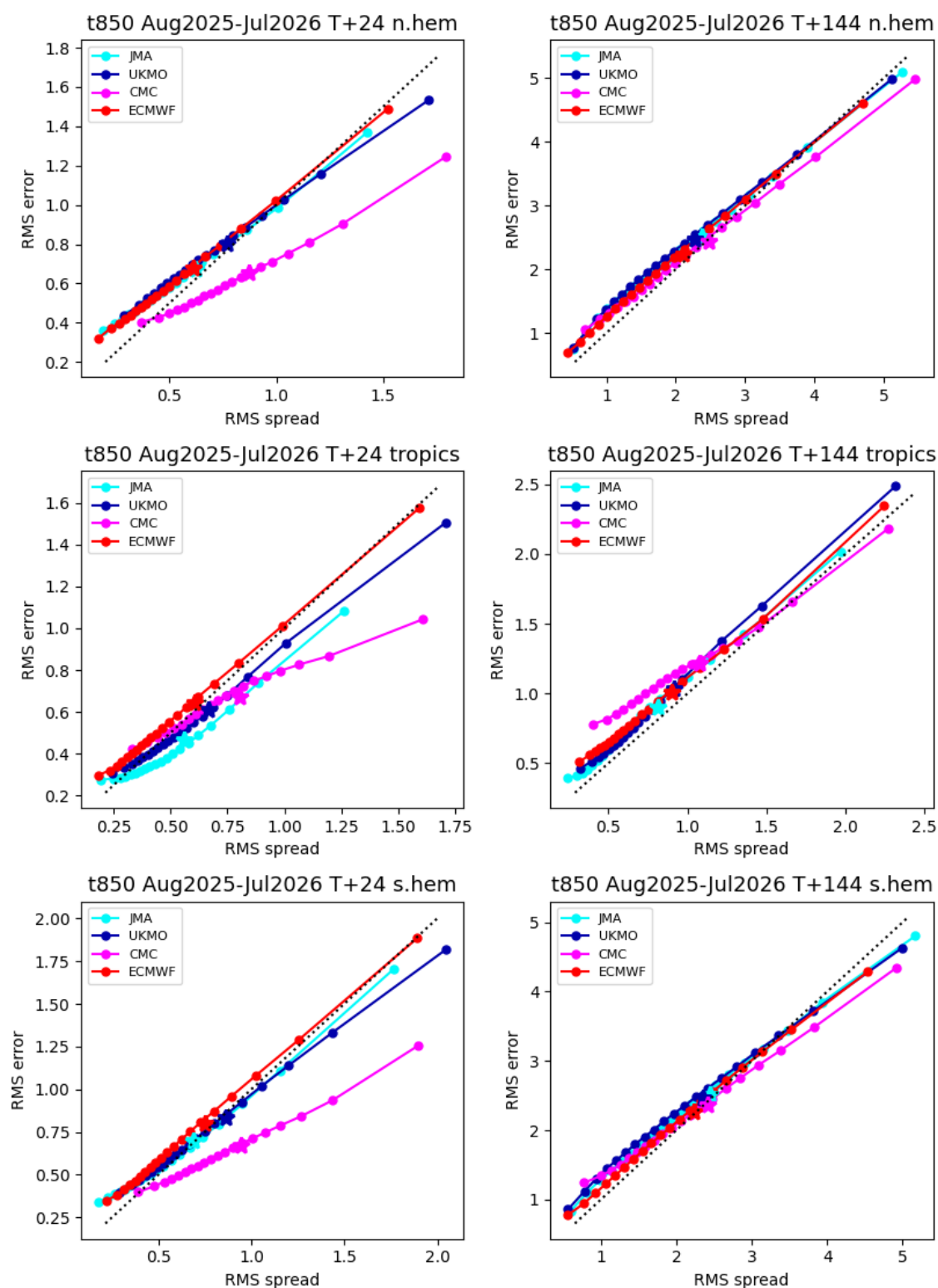


Figure 11: As Figure 8 but for 850 hPa temperature and including the tropics.

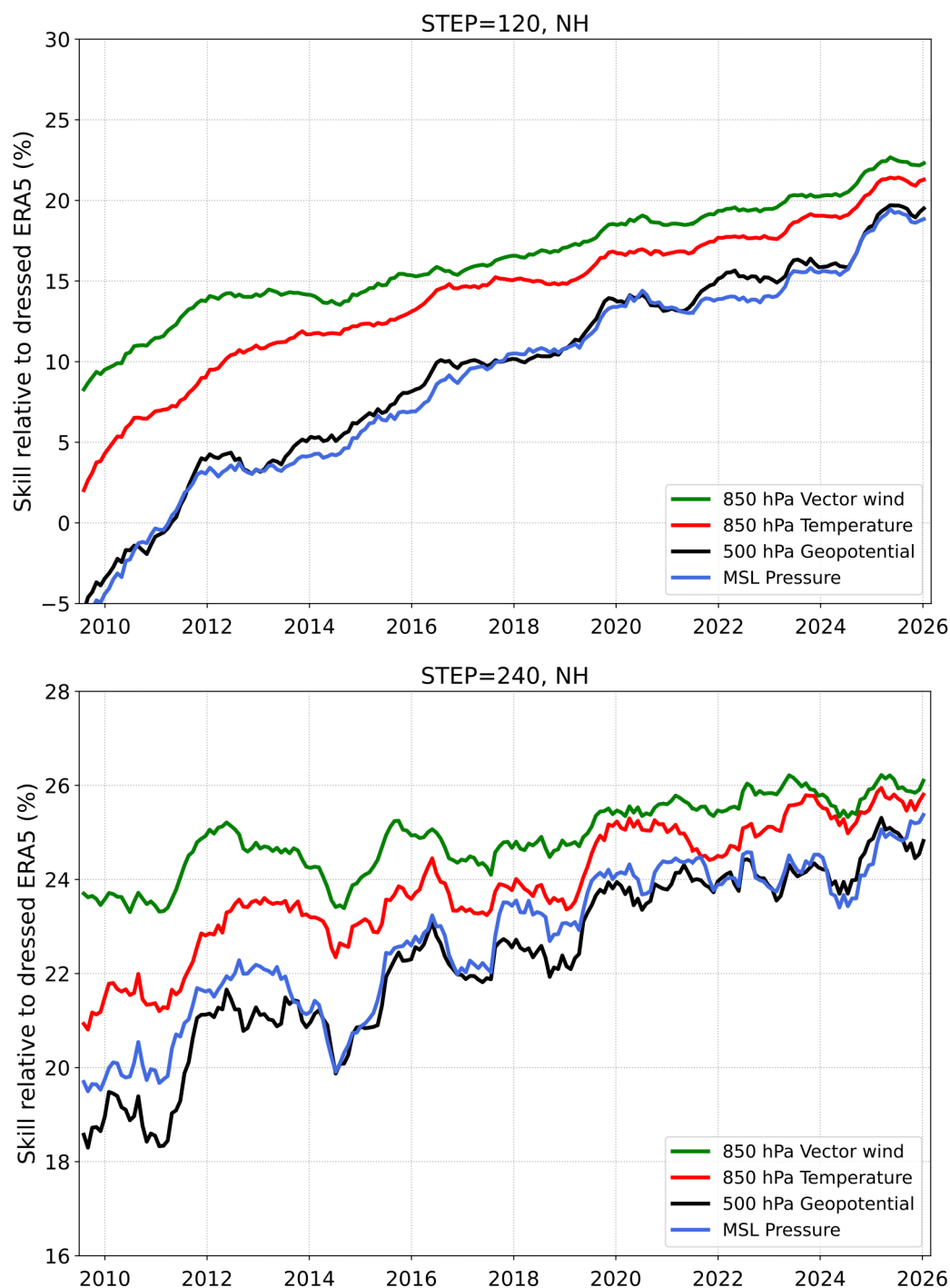


Figure 12: Skill of the IFS-ENS at day 5 (top) and day 10 (bottom) for upper-air parameters in the northern extra-tropics, relative to a Gaussian-dressed ERA5 forecast. Values are running 12-month averages, and verification is performed against own analysis.

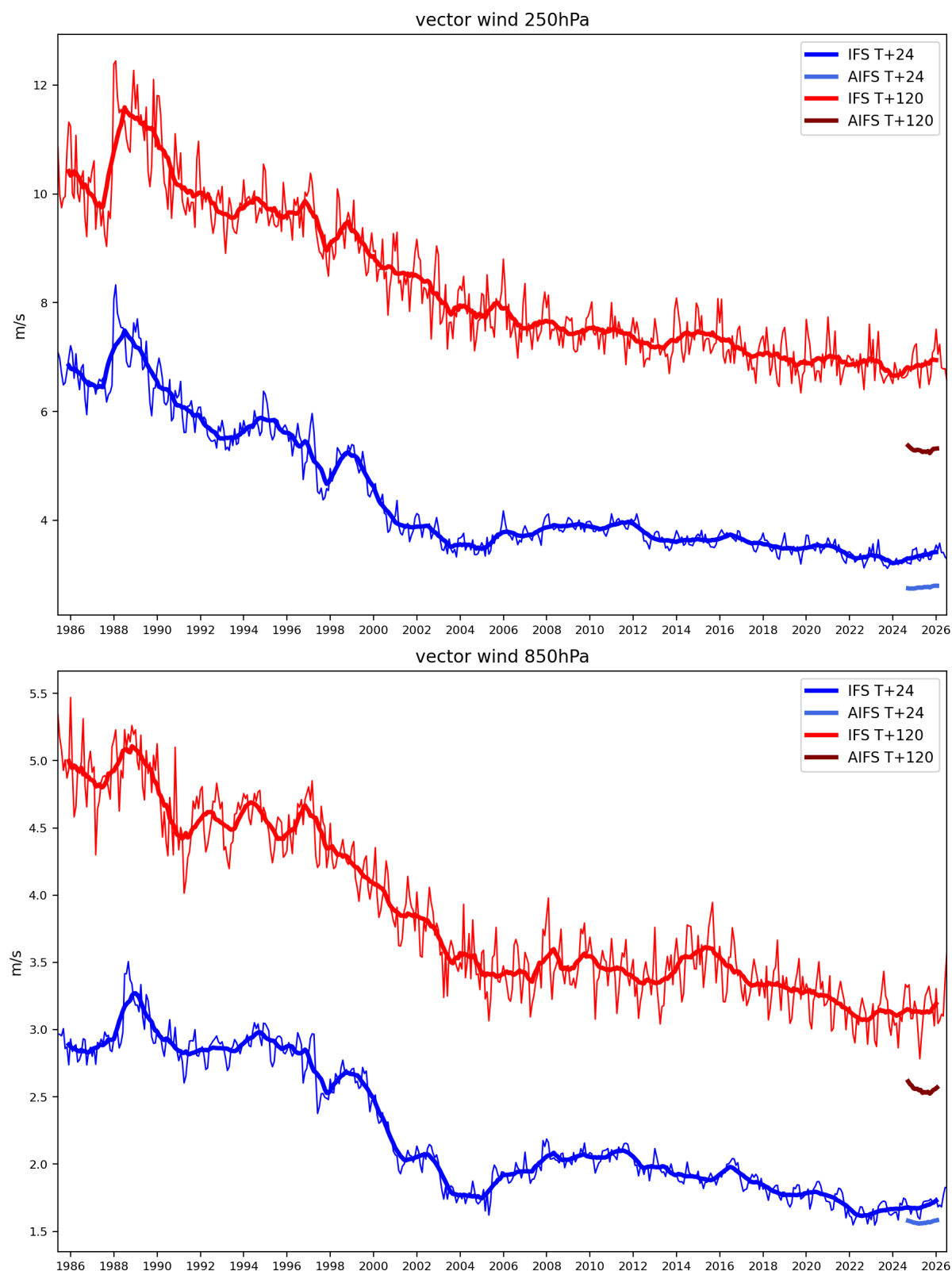


Figure 13: Forecast performance in the tropics. Curves show the monthly average RMS vector wind errors at 250 hPa (top) and 850 hPa (bottom) for one-day (blue) and five-day (red) forecasts, verified against analysis. 12-month moving average scores are also shown (in bold). For reference, the corresponding 12-month running mean for AIFS-Single is shown in dark red (T+120) and light blue (T+24).

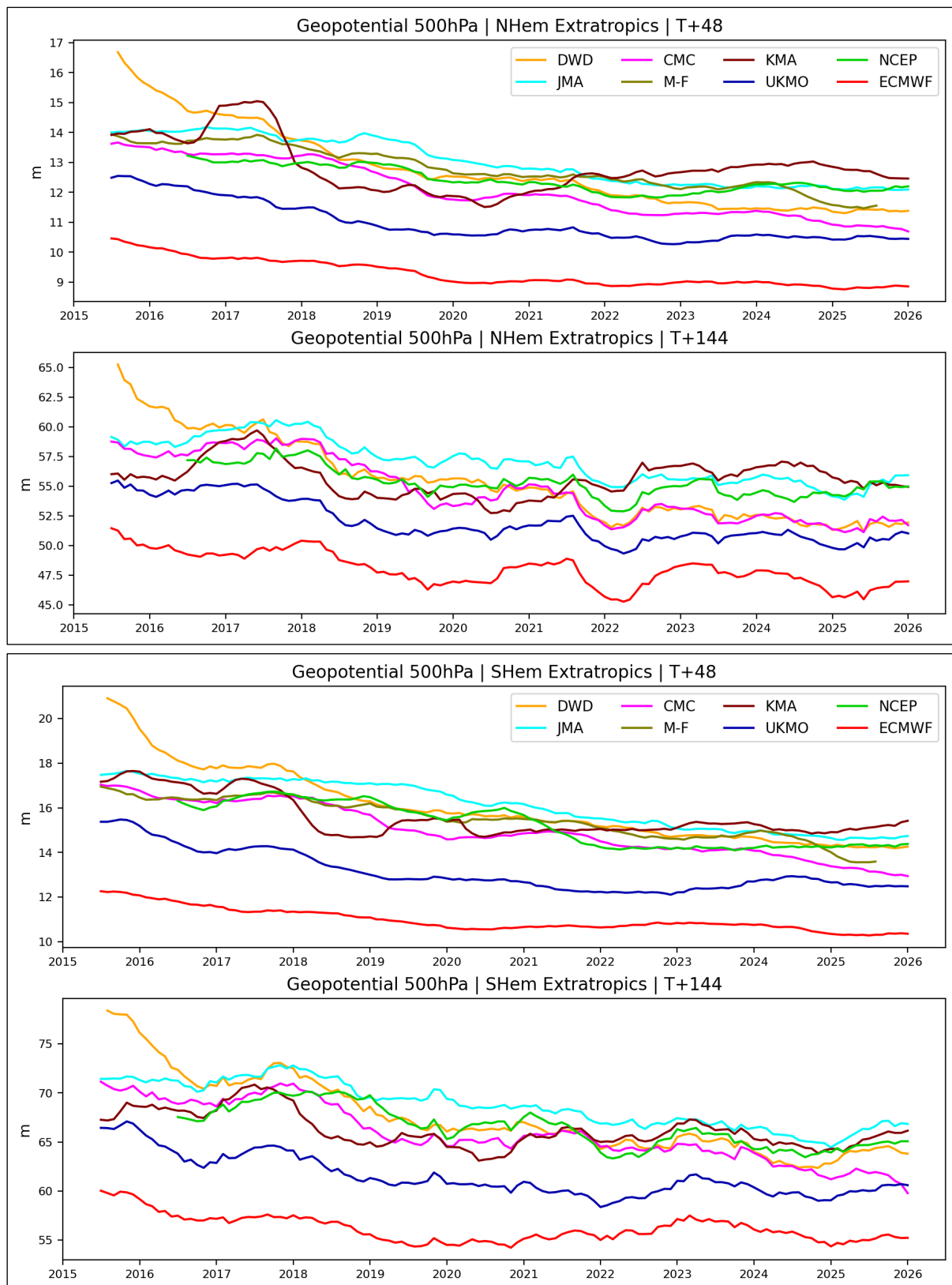


Figure 14: WMO-exchanged scores from global forecast centres. RMS error of 500 hPa geopotential height over northern (top box) and southern (bottom box) extratropics. In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model is verified against its own analysis. JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, KMA = Korea Meteorological Administration, NCEP = U.S. National Centers for Environmental Prediction, DWD = Deutscher Wetterdienst.

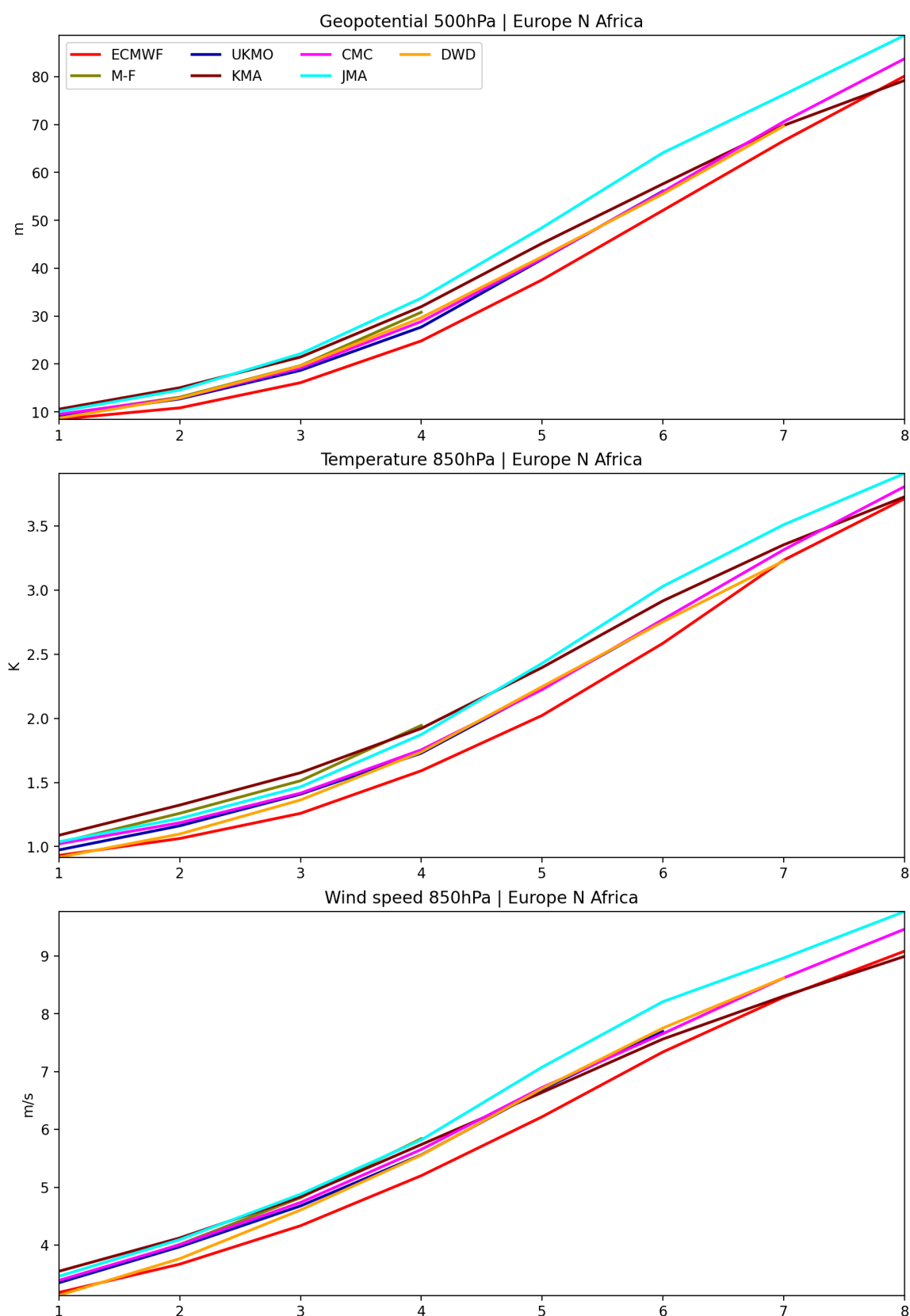


Figure 15: WMO-exchanged scores for verification against radiosondes: 500 hPa height (top), 850 hPa temperature (middle), and 850 hPa wind (bottom) RMS error over Europe and North Africa (annual mean August 2025–July 2026) of forecast runs initialized at 12 UTC. M-F = Météo-France, JMA = Japan Meteorological Agency, CMC = Canadian Meteorological Centre, UKMO = the UK Met Office, KMA = Korea Meteorological Administration, DWD = Deutscher Wetterdienst.

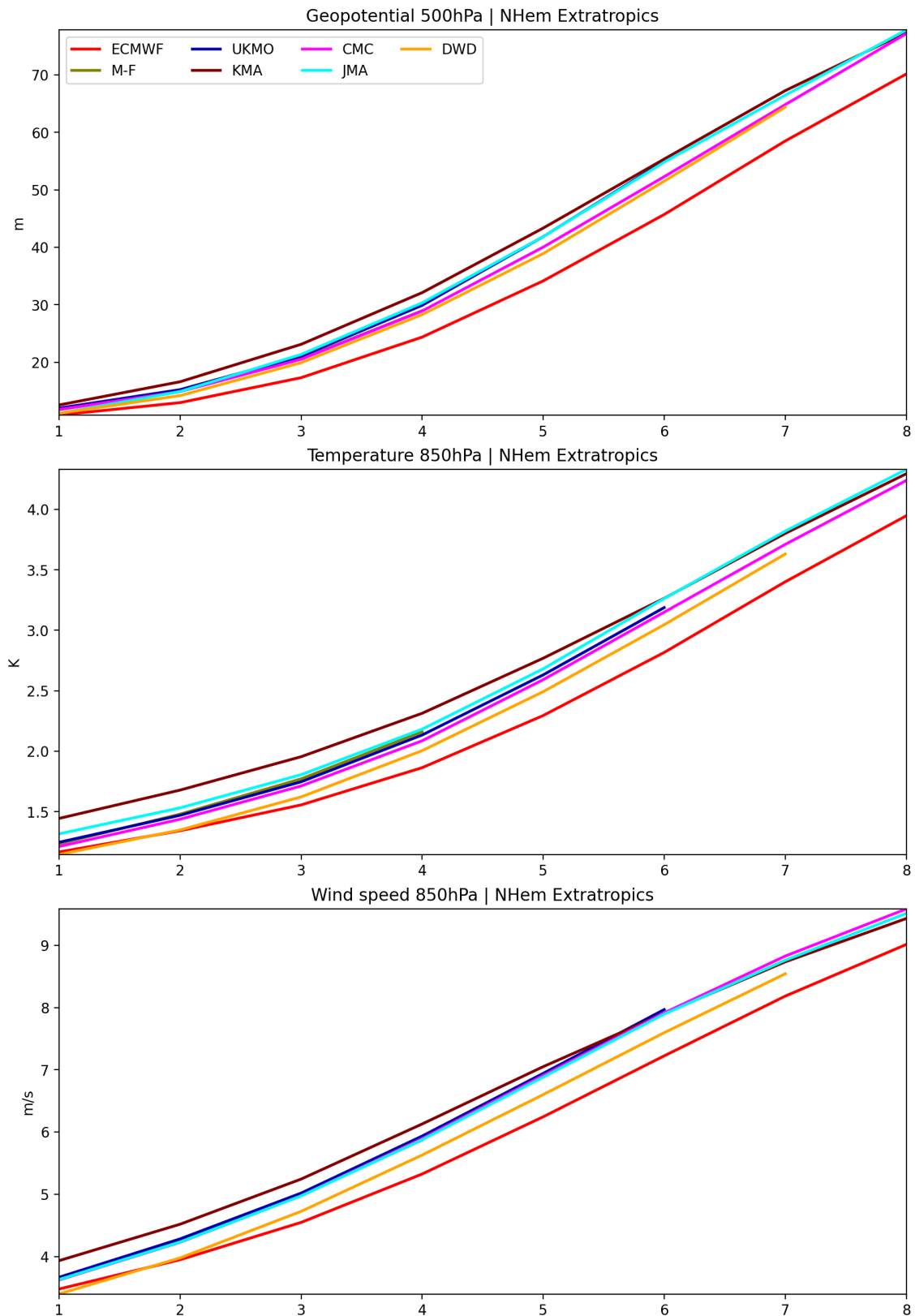


Figure 16: As Figure 15 but for the northern hemisphere extratropics.

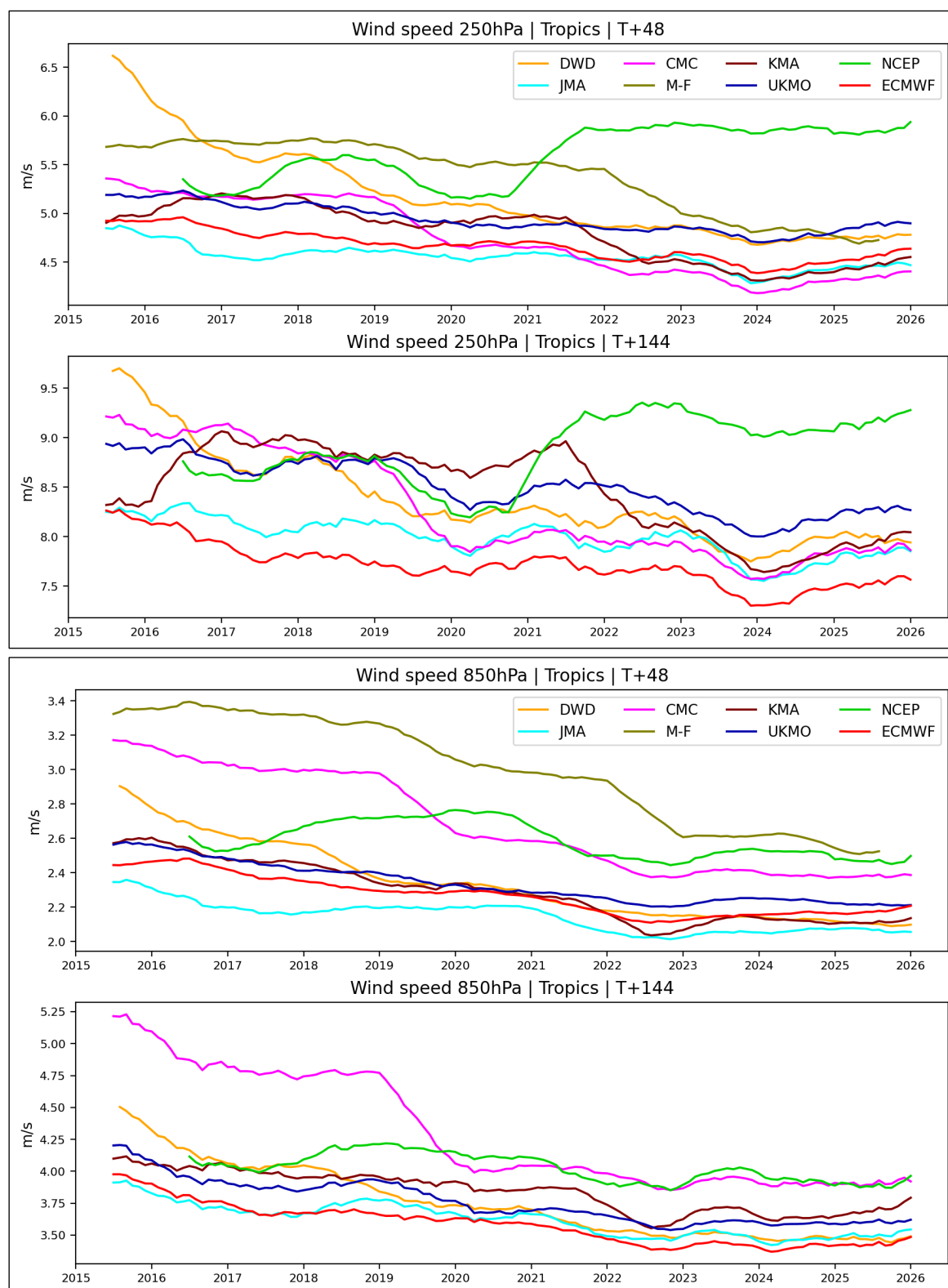


Figure 17: WMO-exchanged scores from global forecast centres. RMS vector wind error over tropics at 250 hPa (top box) and 850 hPa (bottom box). In each box the upper plot shows the two-day forecast error, and the lower plot shows the six-day forecast error of model runs initiated at 12 UTC. Each model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

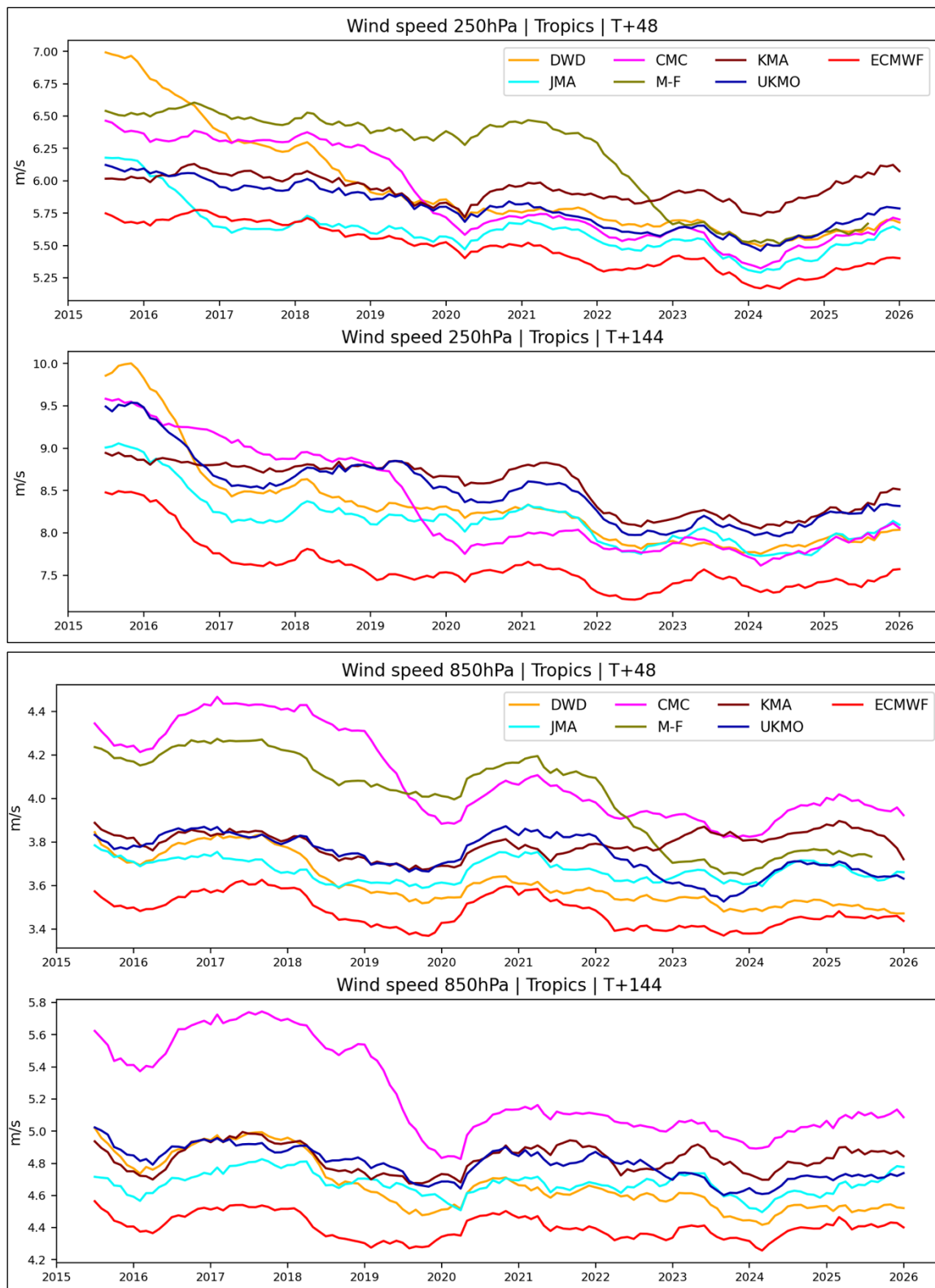


Figure 18: As Figure 17 but for verification against radiosonde observations.

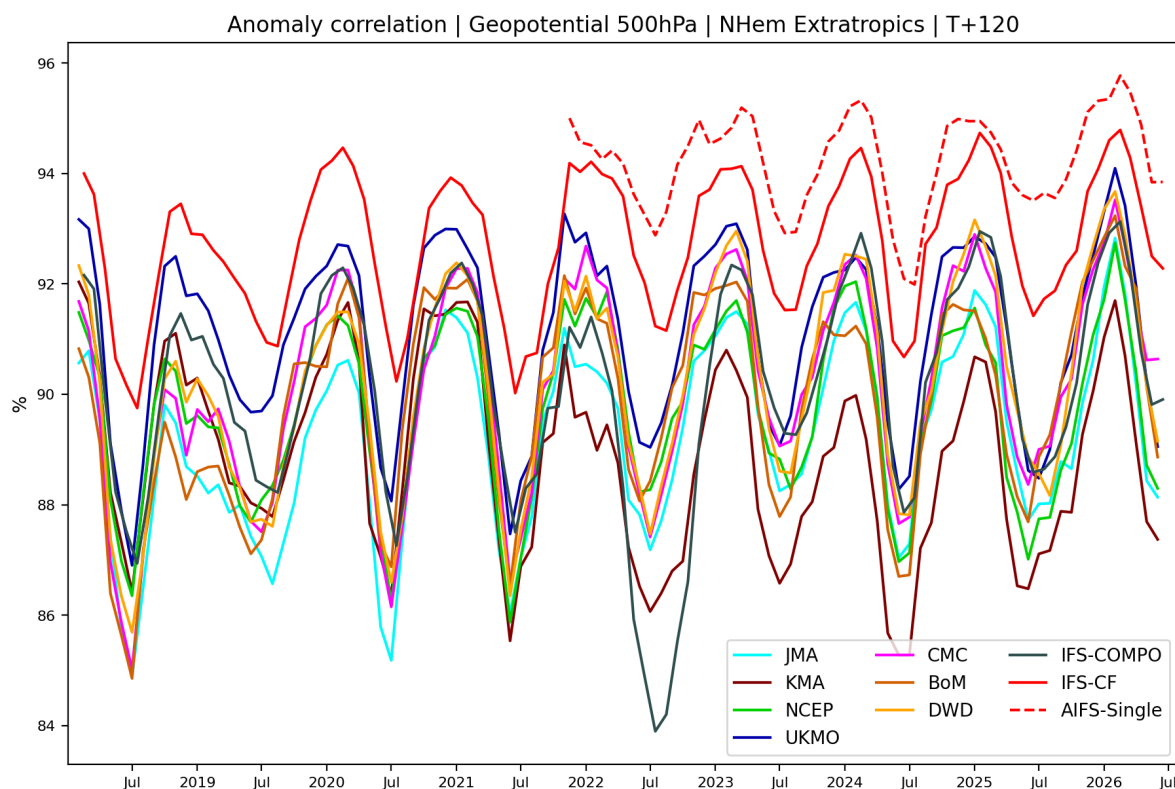


Figure 19: 3-monthly values of anomaly correlation of 500 hPa geopotential in the northern hemisphere extratropics for various centres at day 5. From ECMWF, the CAMS forecast IFS-COMPO (black) and AIFS-Single (red dashes) are included in addition to the IFS-CF (red solid curve).

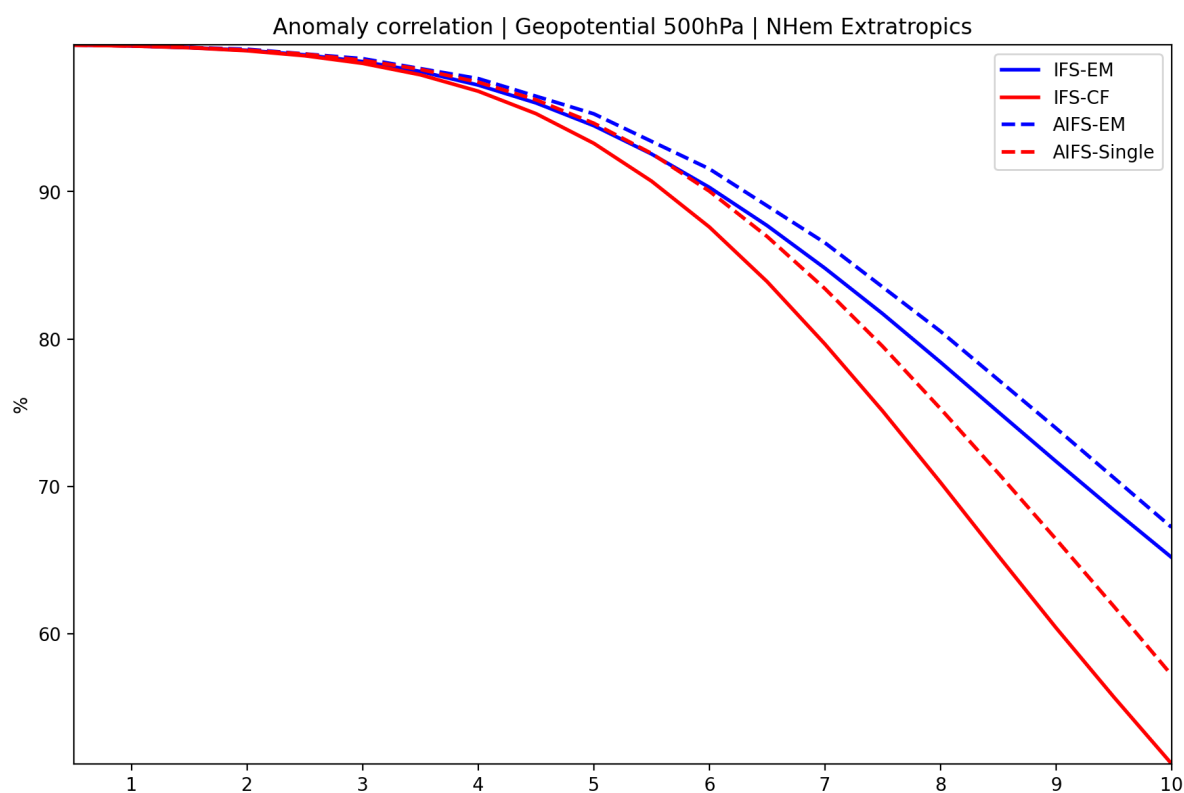


Figure 20: Anomaly correlation of 500 hPa geopotential in the northern extratropics for the 12-month period Aug 2025 to July 2026 for IFS (solid) and AIFS (dashed). Blue: ensemble mean, red: single runs.

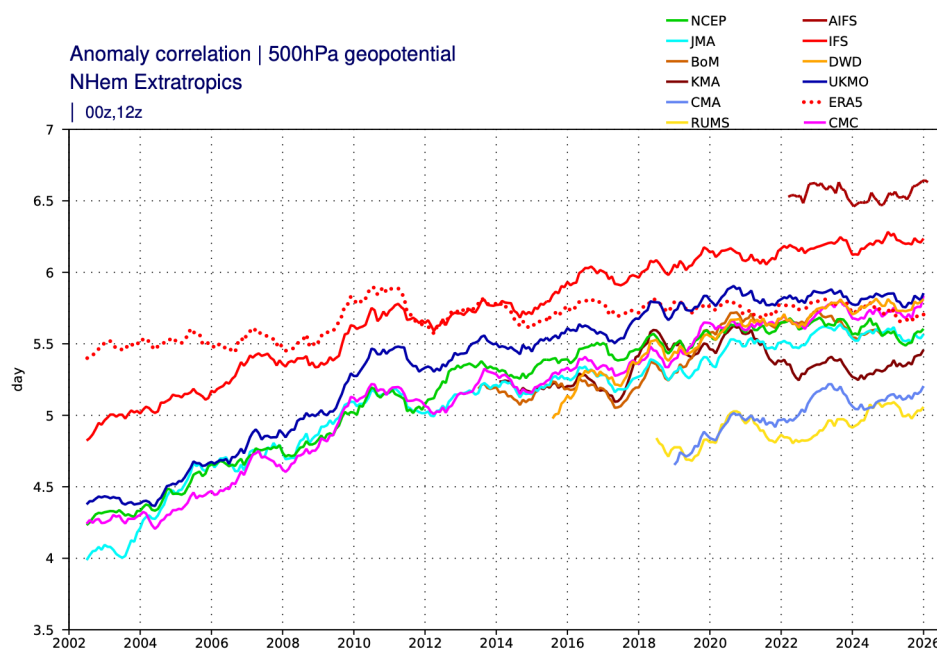


Figure 21: Lead time at which the 500 hPa anomaly correlation in the northern extratropics drops to 86% (specific threshold chosen such that lead time does not exceed forecast range of any centre). Comparison of physics-based forecasts from other global centres, IFS-CF (red), AIFS-Single (red dashes), ERA5 (red dots) for reference, and other ML models run at ECMWF.

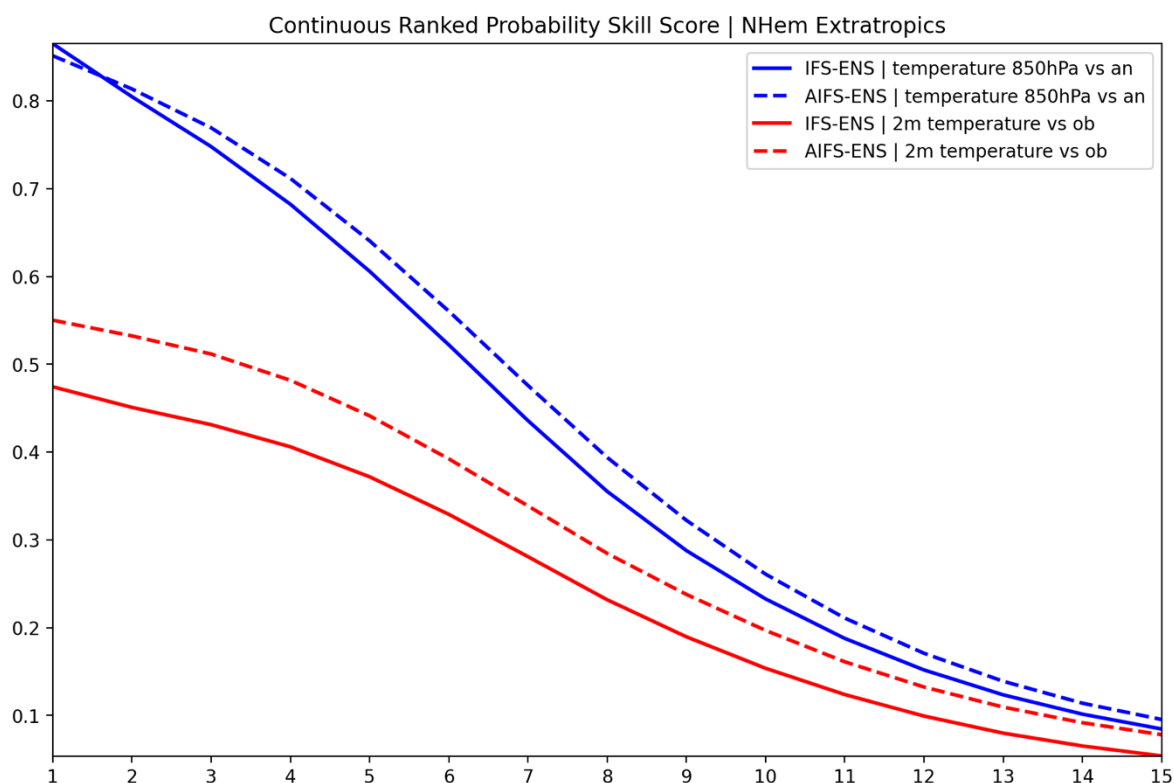


Figure 22: Comparison of IFS-ENS (solid) and AIFS-ENS (dashed) scores for 850 hPa temperature (blue) and 2 m temperature (red) in the northern extratropics. Shown is the CRPSS evaluated against analysis for 850 hPa temperature and evaluated against SYNOP for 2 m temperature. Verification period is August 2025 – July 2026 (12 months).

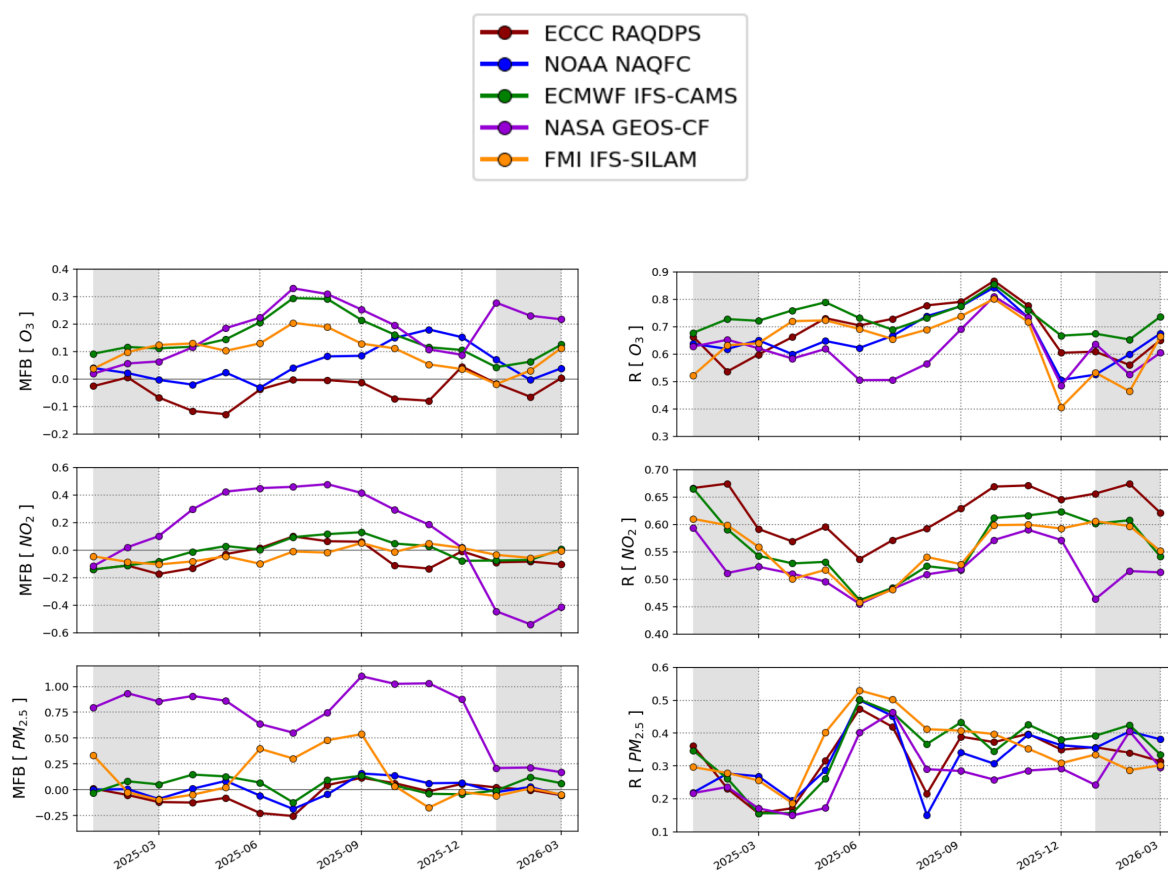


Figure 23: Verification of Ozone (top), NO_2 (centre), and $PM_{2.5}$ (bottom) over a common North American domain in terms of the mean fractional bias (left), and linear correlation (right). Shown is a model intercomparison over a 15-month period in terms of monthly scores. ECMWF's IFS-COMPO forecast is shown in green. Taken from Presseau et al. (2026).

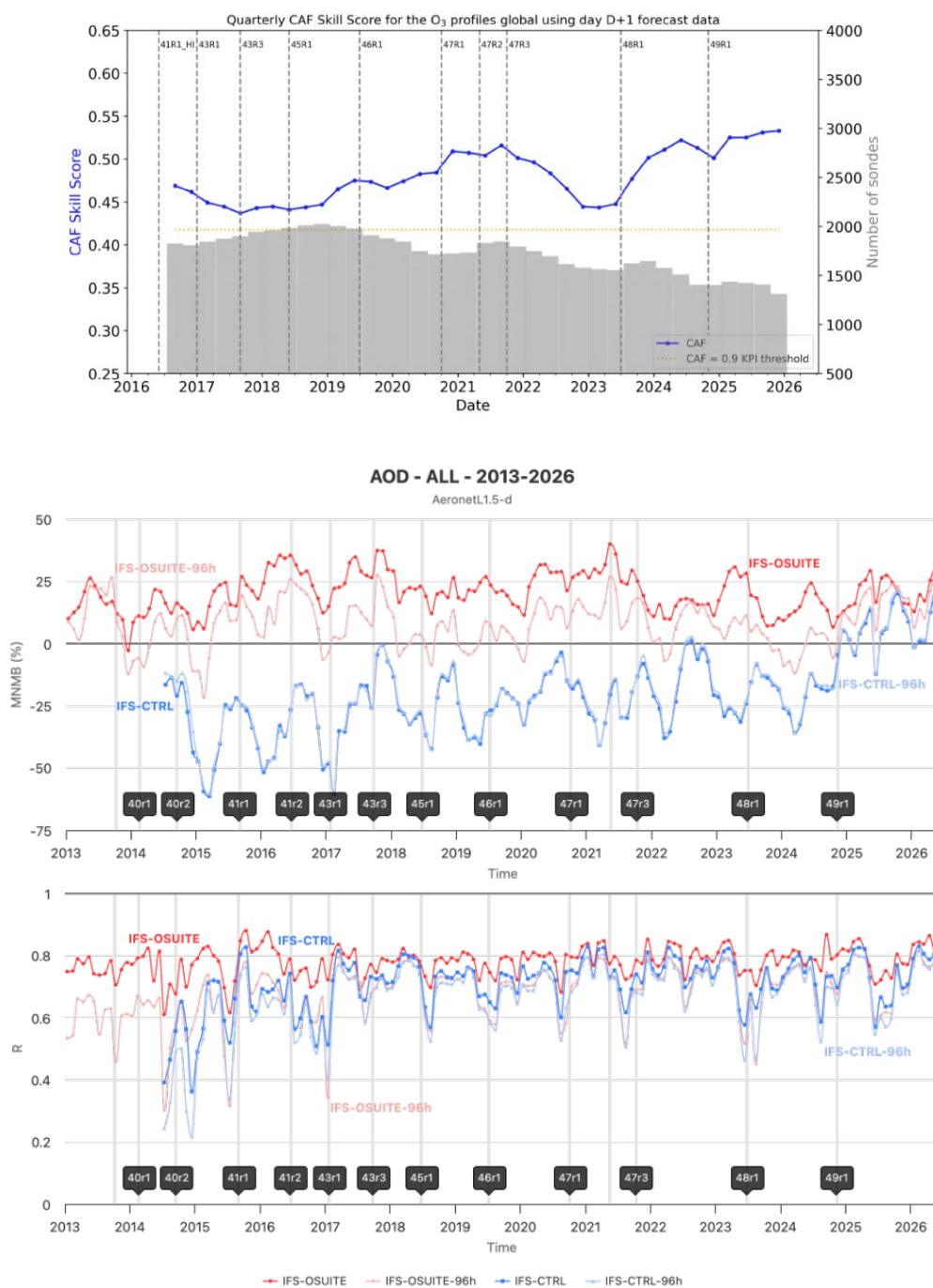


Figure 24: CAMS headline scores reported in the latest EQA report for the quarter DJF 2026 (Benedictow et al., 2026). *Top*: time series of the Common Area Fraction (CAF) values calculated between IFS-COMPO day 1 forecast data and balloon ozone sonde measurements from globally-distributed stations. The horizontal orange dashed line indicates the target KPI at CAF=0.9. The grey bars show the number of sondes used in each reported quarter. The drop of the CAF score in mid 2022 is associated with a slight degradation in Cycle 47r3, which stops with the implementation of Cycle 48r1. *Bottom*: Modified normalized mean bias (MNMB) and Correlation coefficient (R) of the aerosol optical depth at 550nm (AOD) from the IFS-COMPO o-suite at forecast day 1 (red) and at forecast day 4 (light red); control at forecast day 1 (blue) and at forecast day 4 (light blue) against daily matching AERONET Version 3 level 1.5 data. Model version changes are marked as vertical black dashed lines (top panel) and as grey vertical lines (bottom panel).

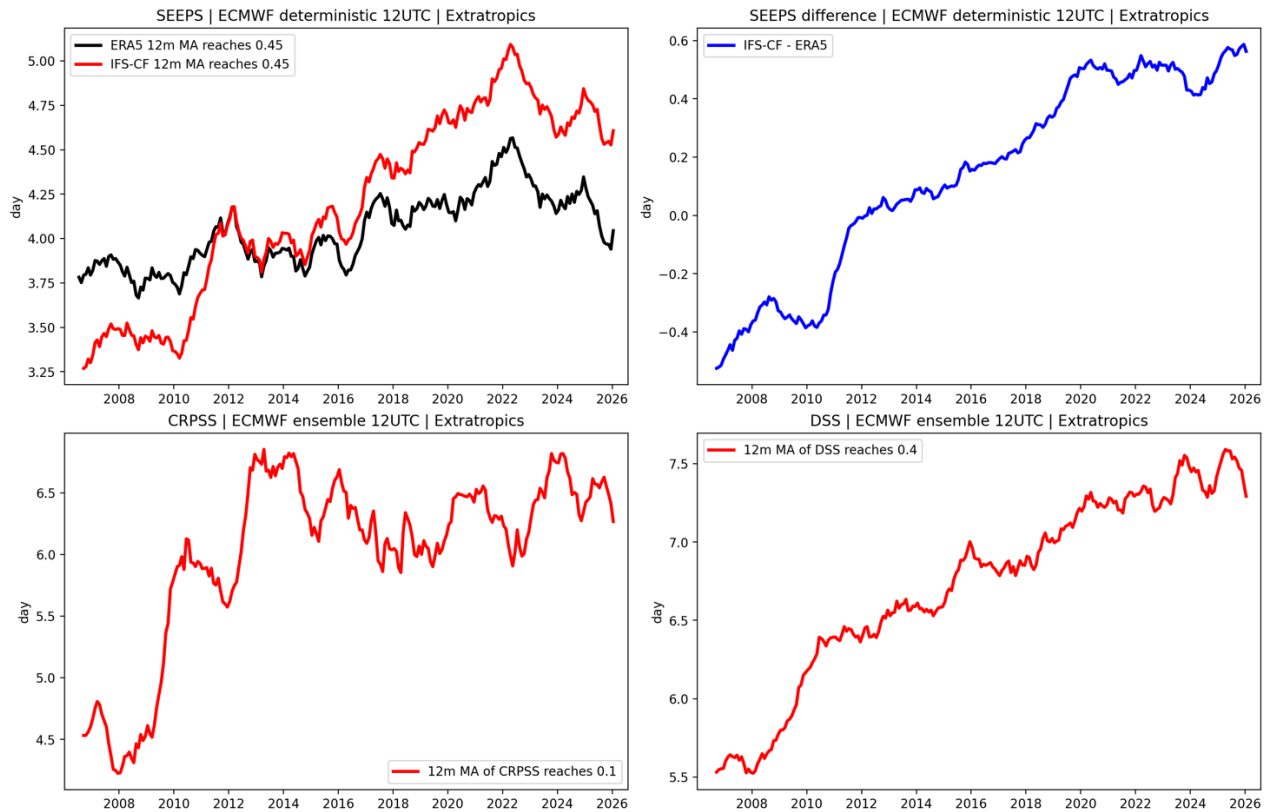


Figure 25: Supplementary headline scores (left column) and additional metrics (right column) for deterministic (top) and probabilistic (bottom) precipitation forecasts. The evaluation is for 24-hour total precipitation verified against synoptic observations in the extratropics. Curves show the number 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 IFS-CF and ERA5 (blue). Probabilistic scores in the bottom row are the Continuous Ranked Probability Skill Score (CRPSS) and the Diagonal Skill Score (DSS).

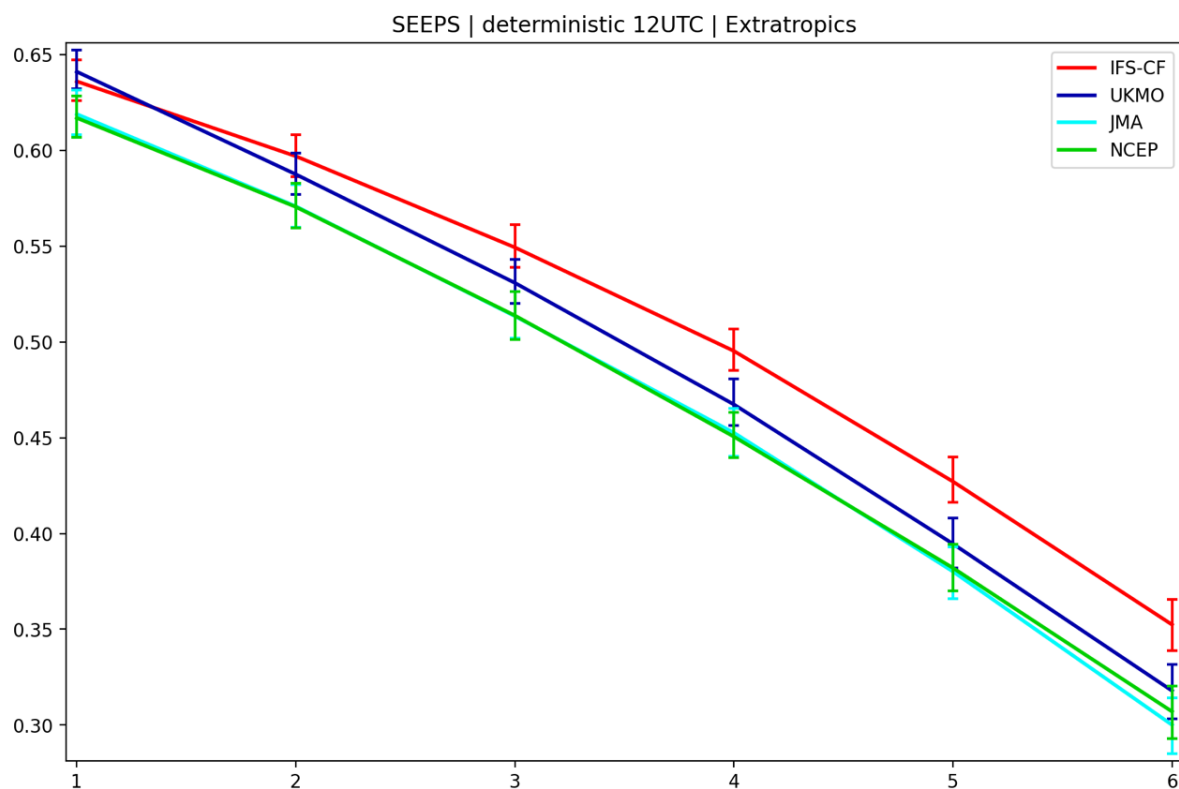


Figure 26: Comparison of SEEPS deterministic precipitation forecast skill for ECMWF (red), the Met Office (UKMO, blue), Japan Meteorological Agency (JMA, cyan) and NCEP (green). Curves show the skill computed over all available synoptic stations in the extratropics for forecasts from August 2025–July 2026. Bars indicate 95% confidence intervals.

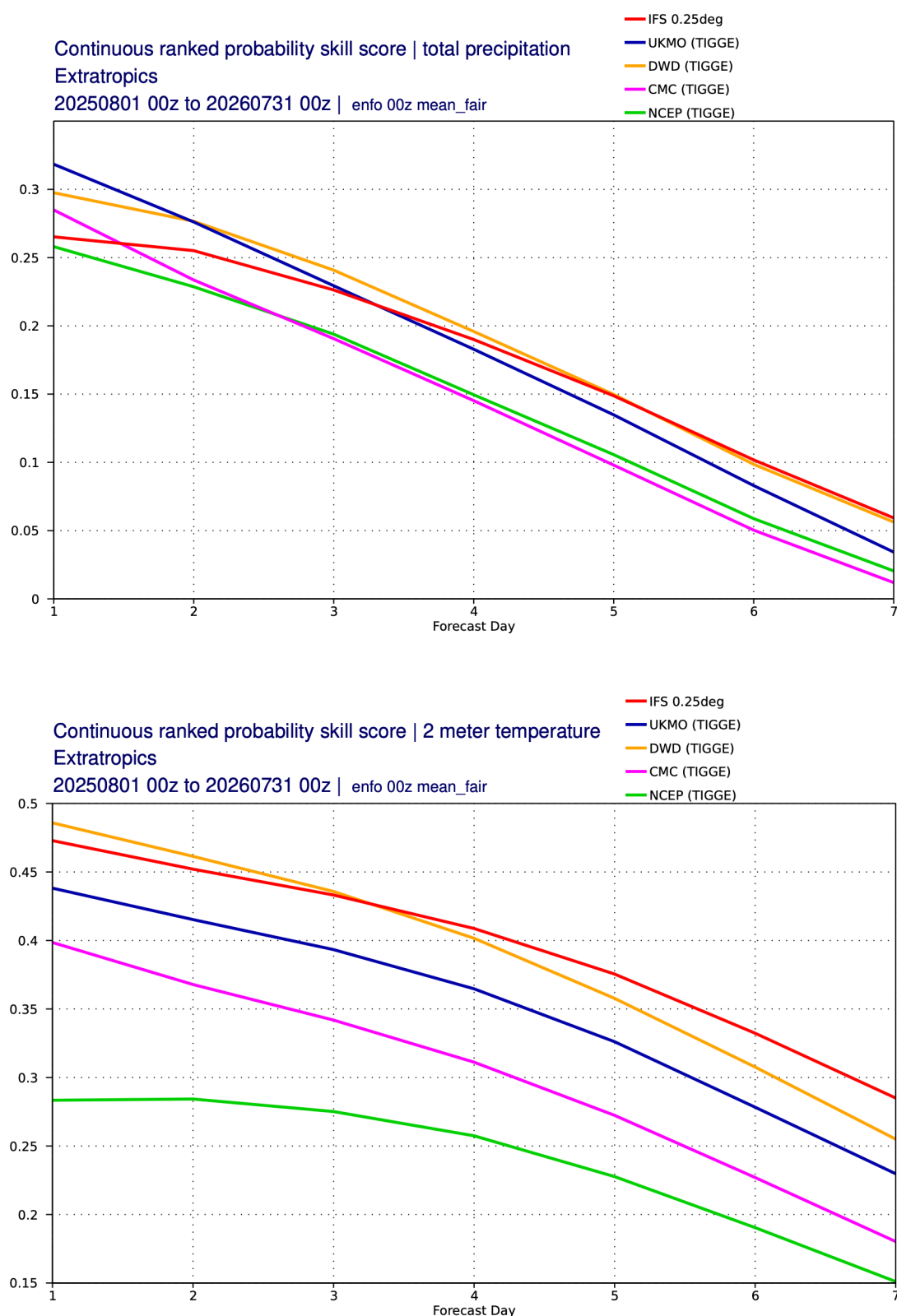


Figure 27: IFS-ENS forecast skill compared to other centres for precipitation (top) and 2 m temperature (bottom) using data from the TIGGE archive. Note that the resolution of the IFS-ENS forecast has been degraded to 0.25 deg to better match the resolution of TIGGE datasets from other centres shown (ranging from 0.25 to 0.5 deg). Verification is against SYNOP.

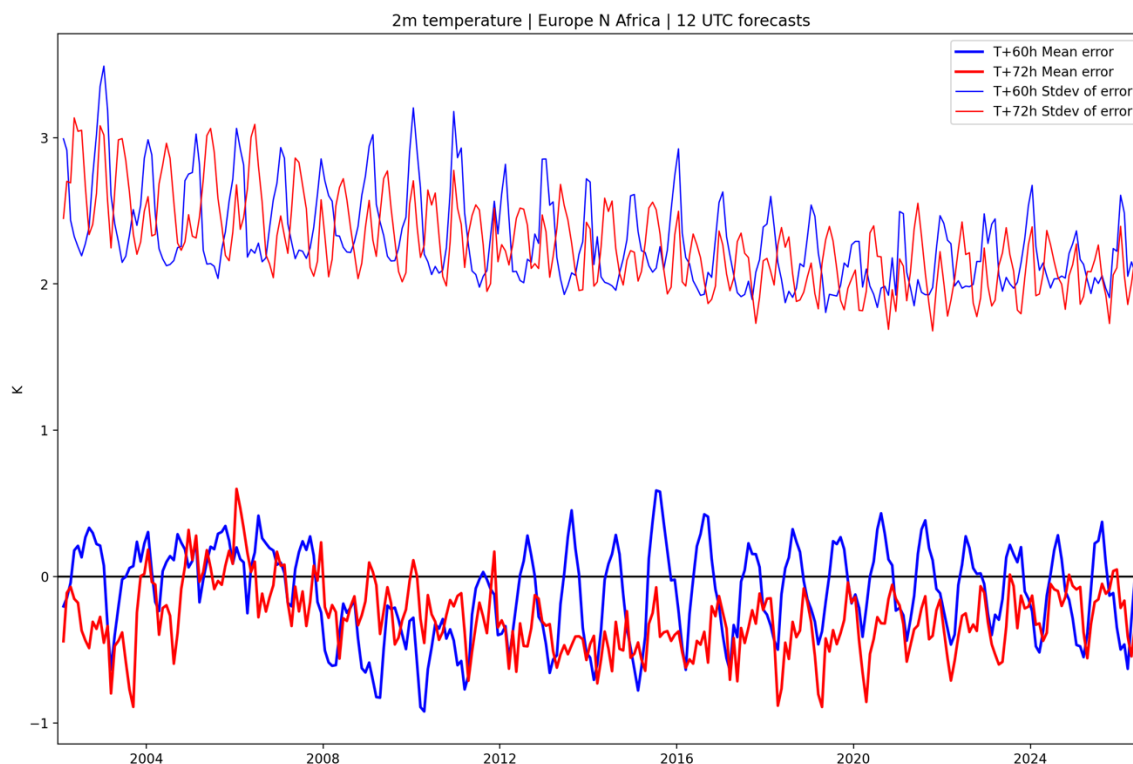


Figure 28: Verification of 2 m temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves show bias, upper curves are standard deviation of error.

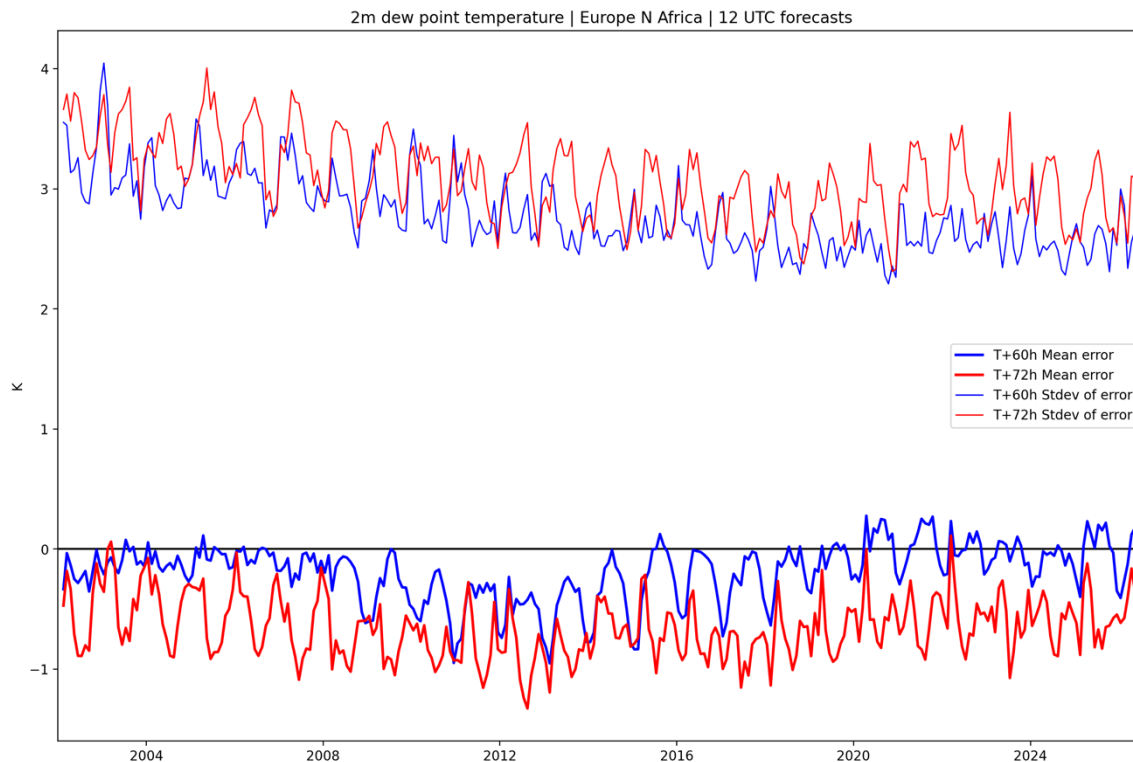


Figure 29: Verification of 2 m dew point forecasts against European SYNOP data on the Global Telecommunication System (GTS) for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves show bias, upper curves show standard deviation of error.

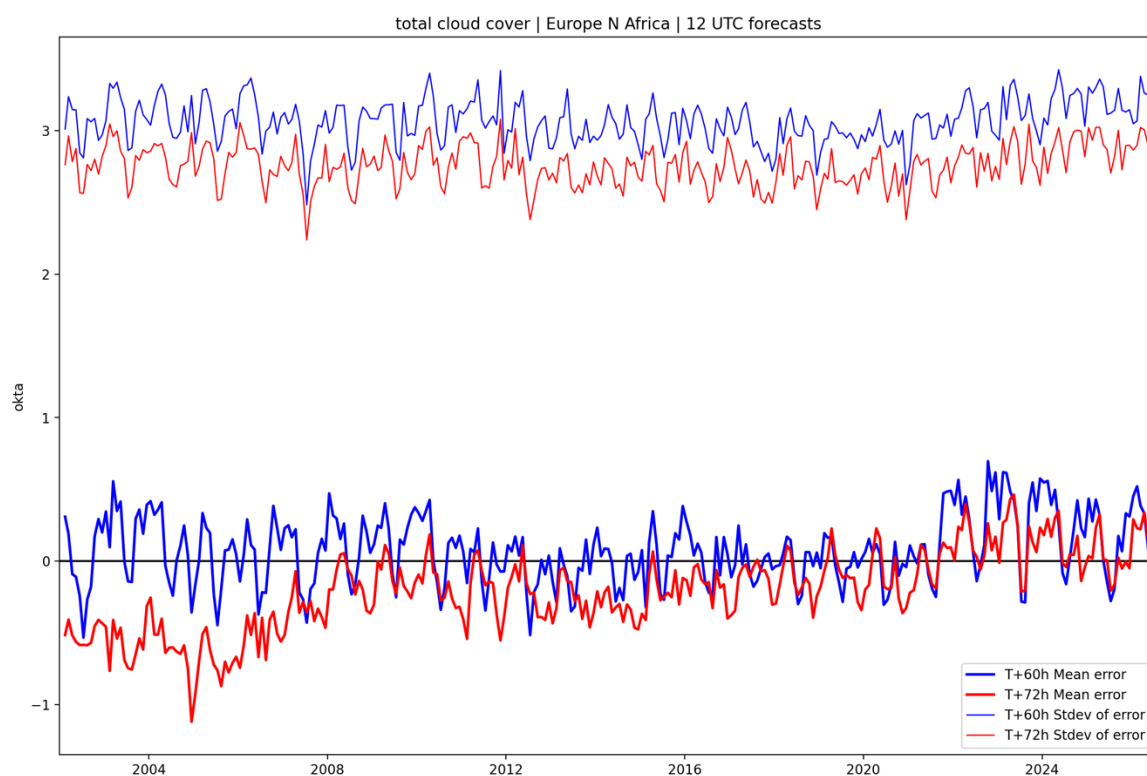


Figure 30: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves show bias, upper curves show standard deviation of error.

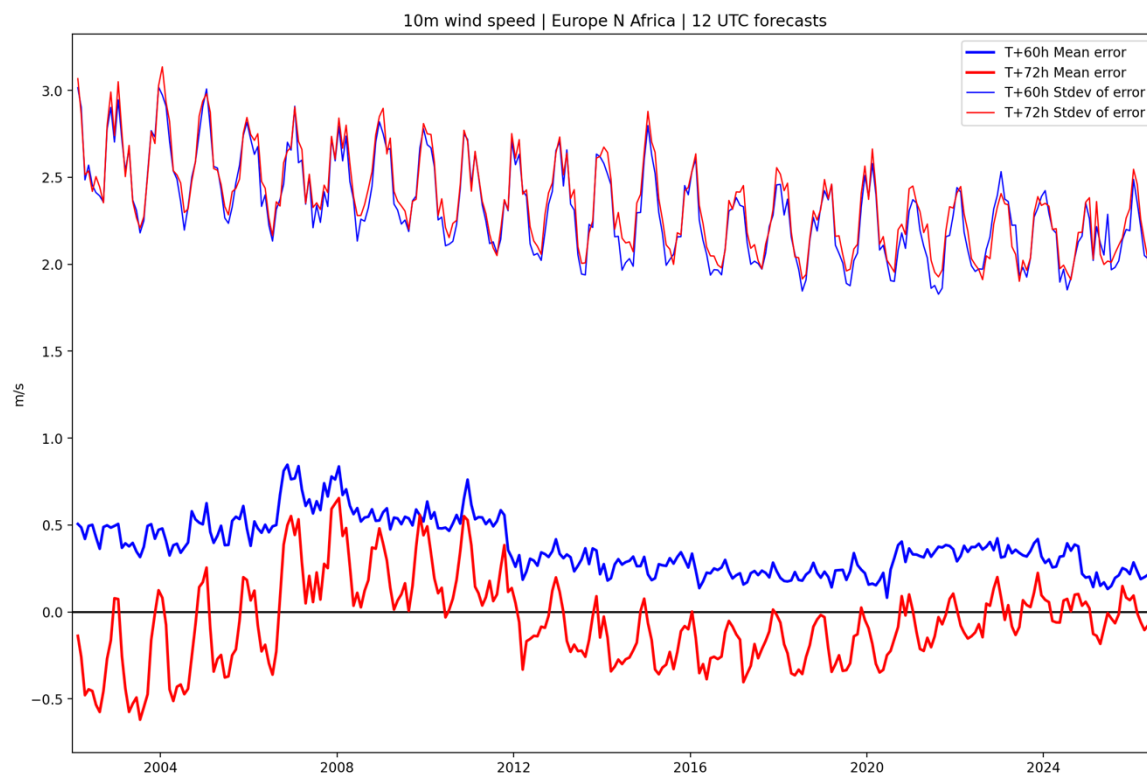


Figure 31: Verification of 10 m wind speed forecasts against European SYNOP data on the GTS for 60-hour (night-time, blue) and 72-hour (daytime, red) forecasts. Lower pair of curves show bias, upper curves show standard deviation of error.

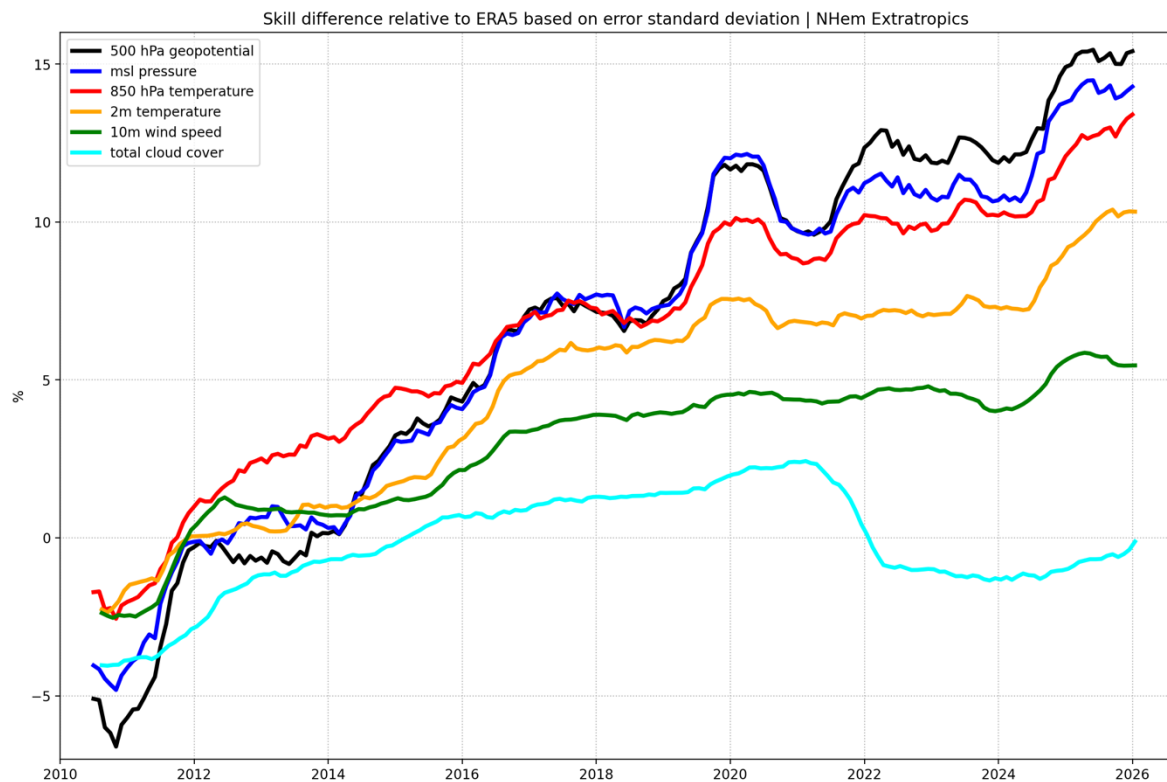


Figure 32: Evolution of skill of the IFS-CF at day 5 in the northern hemisphere extratropics, expressed as relative skill compared to ERA5 using error standard deviation as a metric. Verification is against analysis for 500 hPa geopotential, 850 hPa temperature, and mean sea level pressure, and against SYNOP for 2 m temperature, 10 m wind speed, and total cloud cover.

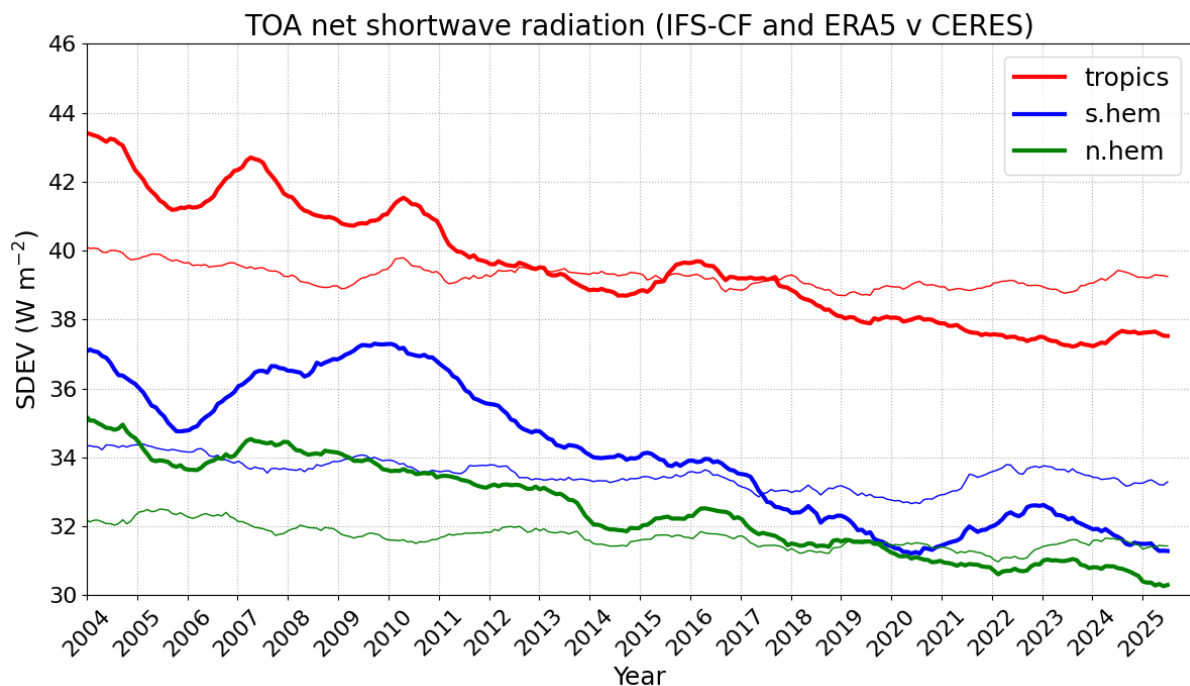


Figure 33: Evolution of the RMSE of the IFS-CF forecast at day 5 (bold lines) of the top of the atmosphere (TOA) net shortwave radiation for the two extratropical hemisphere and the tropics. Thin lines show the RMSE of the ERA5 forecast for comparison. Verification is against CERES satellite data.

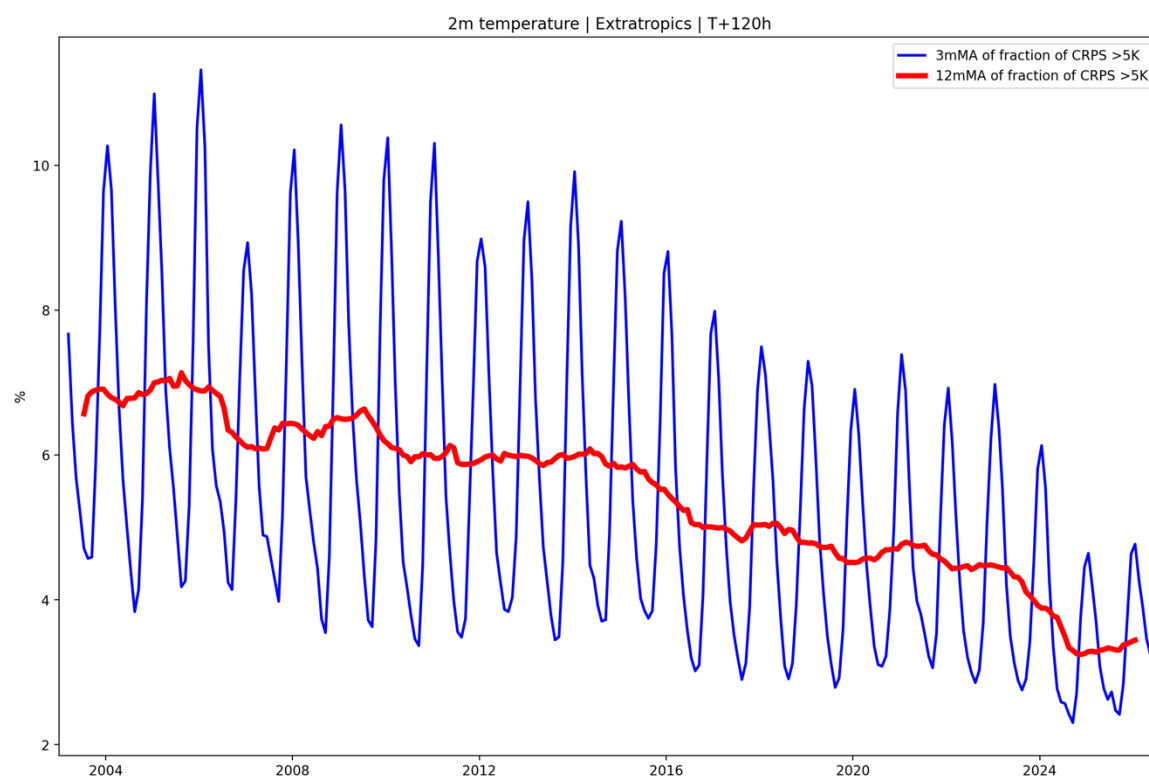


Figure 34: Evolution of the fraction of large IFS-ENS 2m temperature errors (CRPS>5K) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

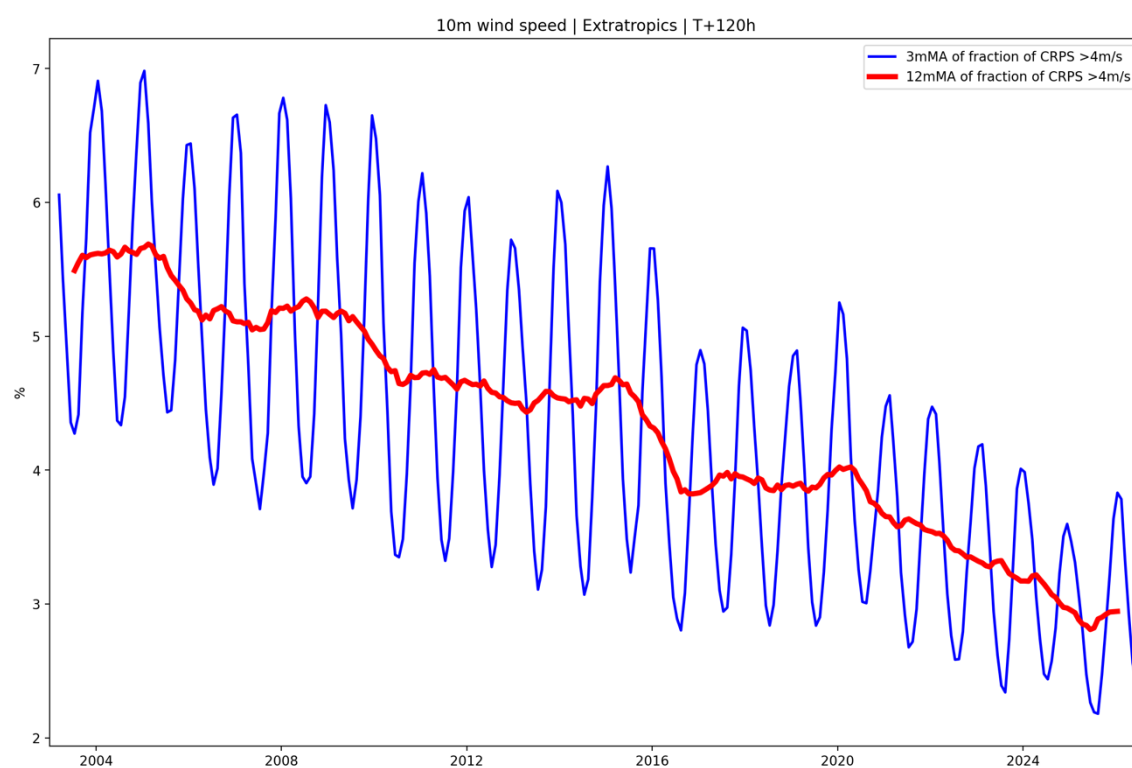


Figure 35: Evolution of the fraction of large IFS-ENS 10m wind speed errors (CRPS>4m/s) at forecast day 5 in the extratropics. Verification is against SYNOP observations. 12-month running mean shown in red, 3-month running mean in blue.

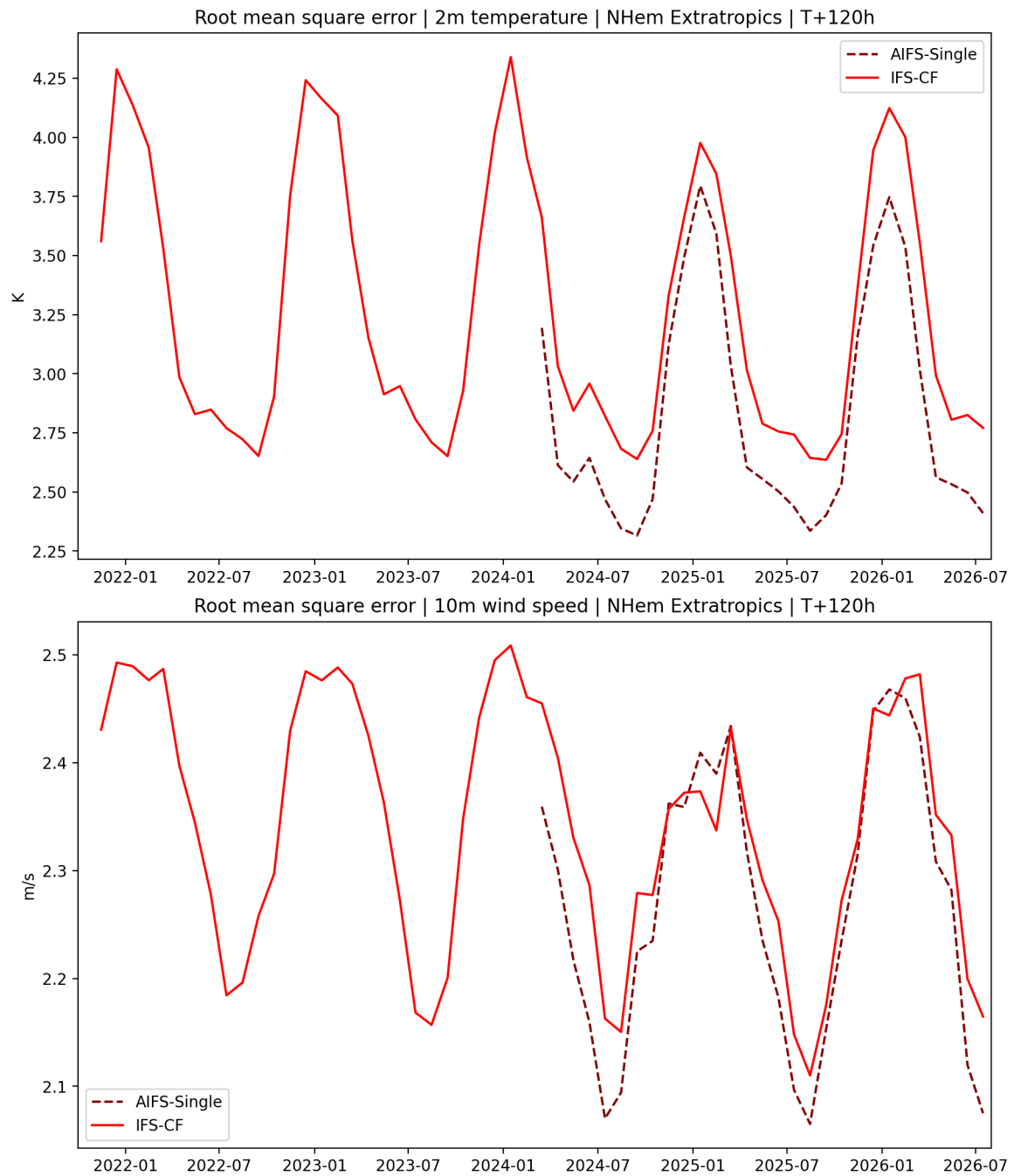


Figure 36: RMSE of 2 m temperature (top) and 10 m wind speed (bottom) at day 5 in the northern extratropics, comparing IFS-CF and AIFS-Single. Verification is against SYNOP observations. Shown are monthly averages.

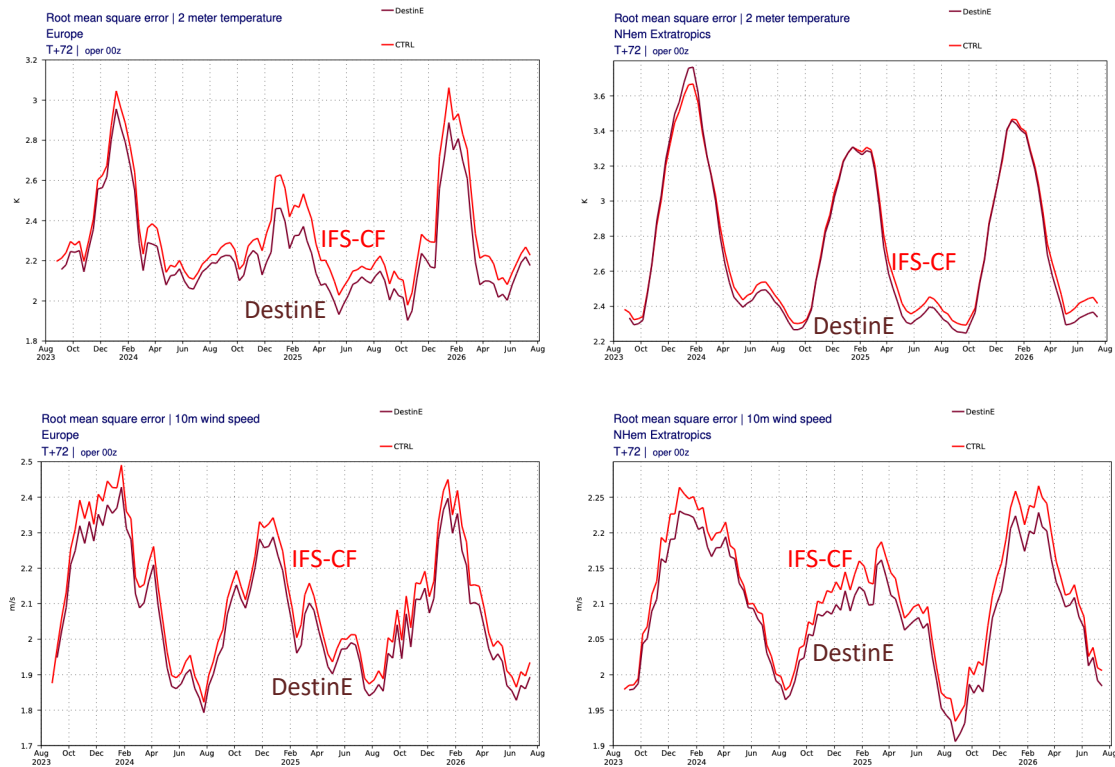


Figure 37: RMSE of IFS-CF and DestinE Extremes DT (4.4 km) 00Z runs for 2 m temperature (top) and 10 m wind speed forecasts (bottom) at day 3 in Europe (left panels) and the northern extratropics (right panels). Verification is against SYNOP observations, shown are monthly averages.

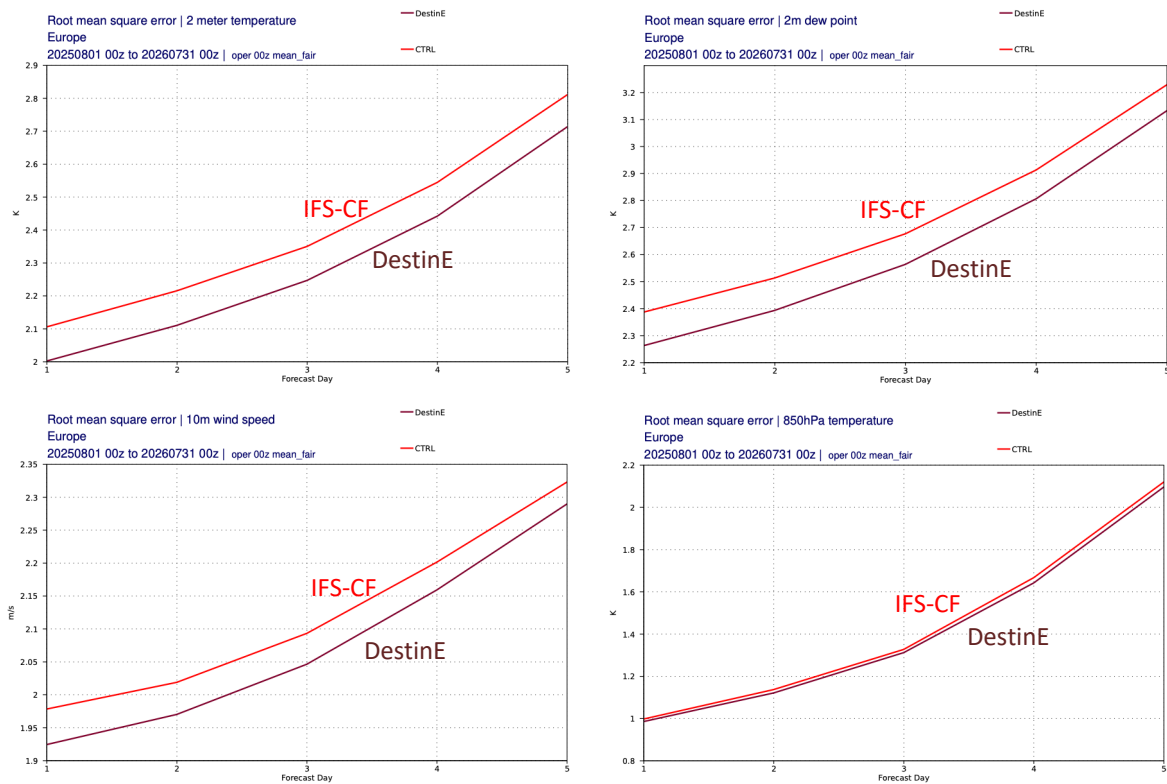


Figure 38: RMSE of IFS-CF and DestinE Extremes DT (4.4 km) 00Z runs for three weather parameters (2 m temperature, 2 m dewpoint, 10 m wind speed) verified against SYNOP, and one upper-air parameter (850 hPa temperature) verified against radiosondes, as a function of lead time. Verification period is 12 months, from August 2025 to July 2026.

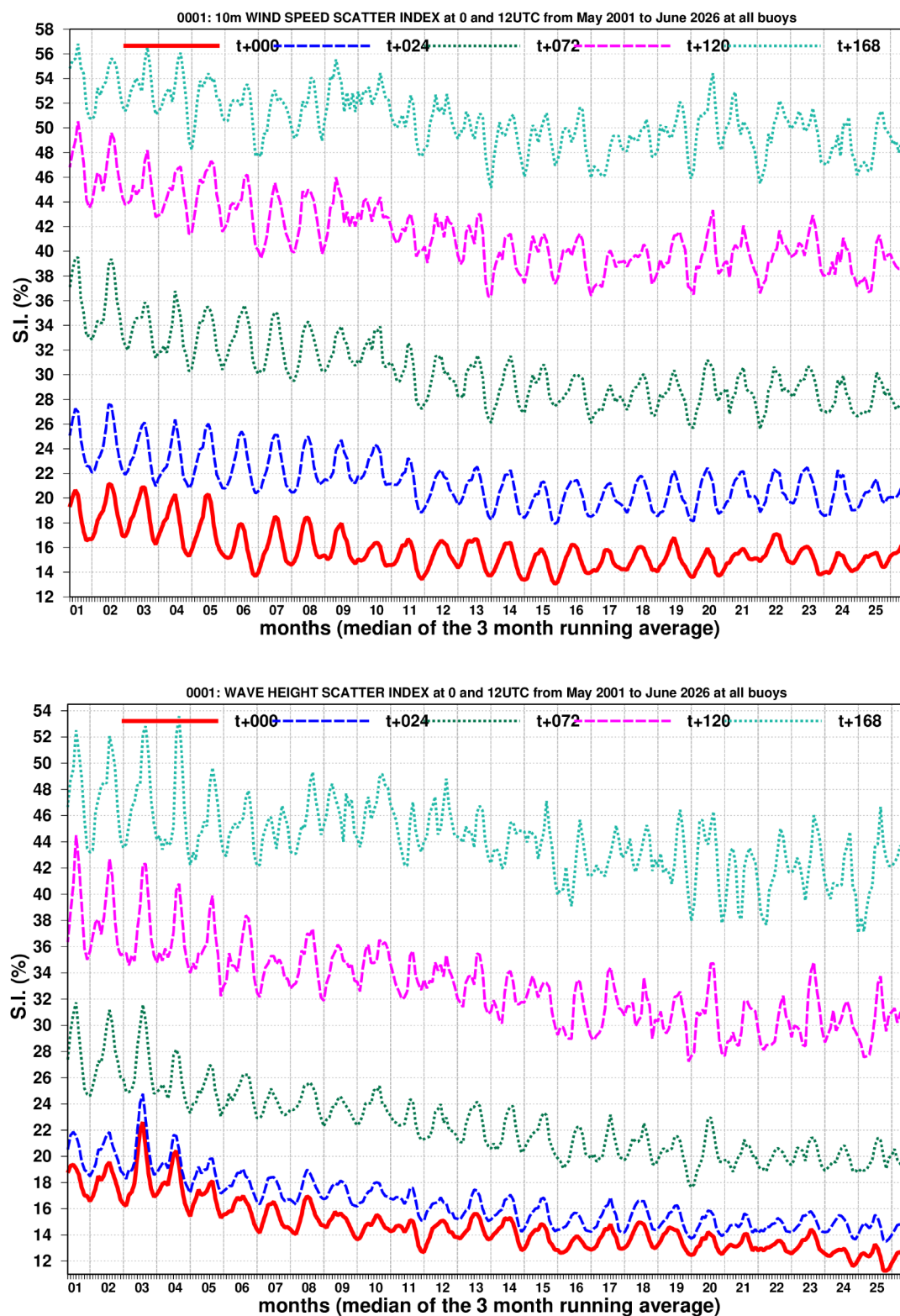


Figure 39: Time series of verification of the ECMWF 10 m wind forecast (top panel) and wave height forecast (bottom panel) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is applied.

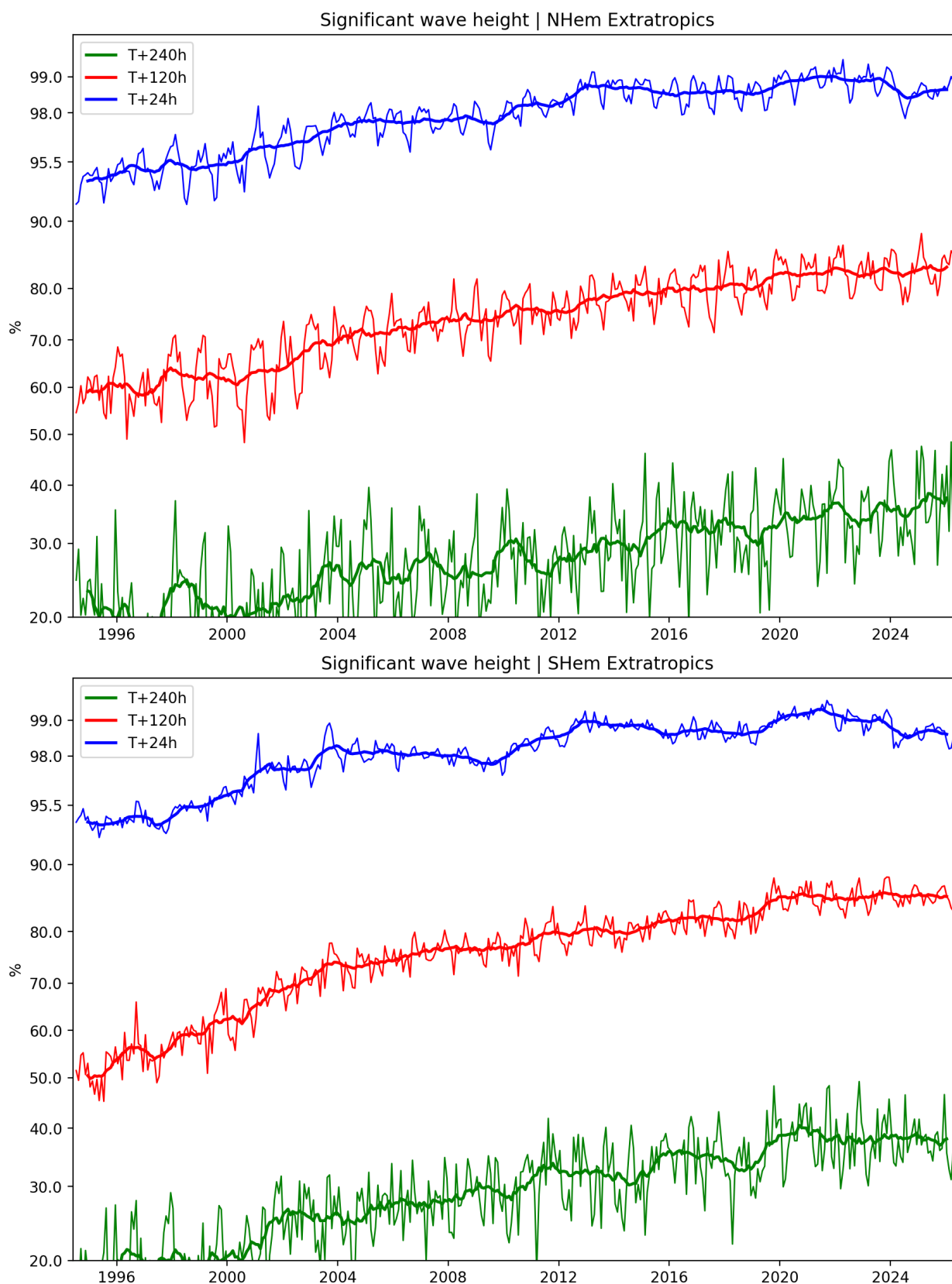


Figure 40: Ocean wave forecasts. Monthly score and 12-month running mean (bold) of ACC for ocean wave heights verified against analysis for the northern (top) and southern extratropics (bottom) at day 1 (blue), 5 (red) and 10 (green).

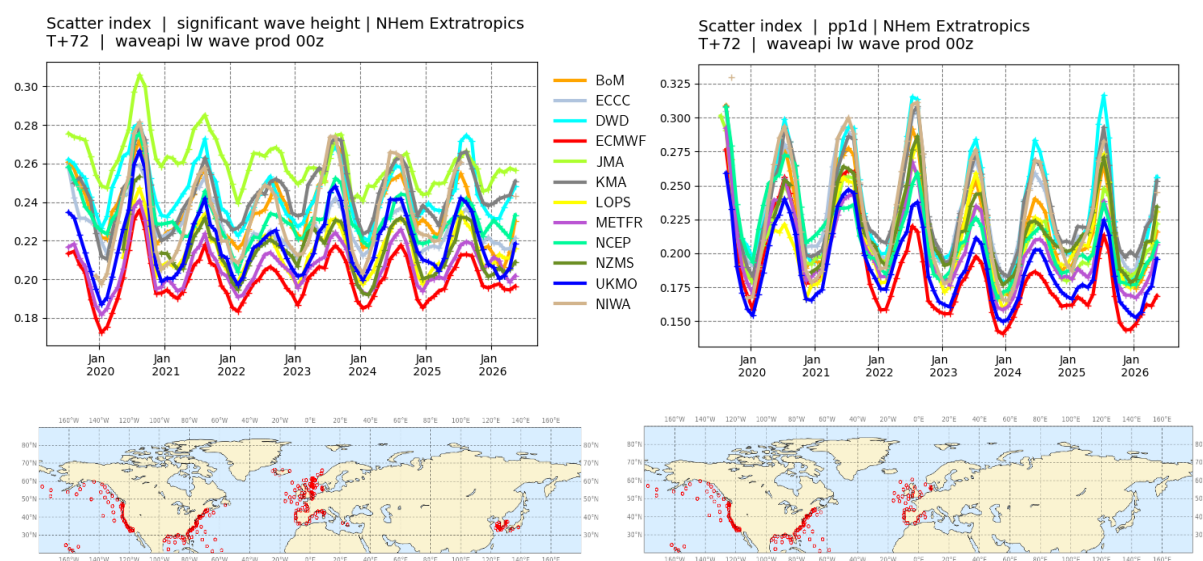


Figure 41: Verification of forecasts of wave height and peak wave period (upper panels) at +72 h using observations from wave buoys (lower panels). The scatter index (SI) is the standard deviation of error normalised by the mean observed value. NCEP: National Centers for Environmental Prediction, USA; 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 Atmospheric Research, New Zealand.

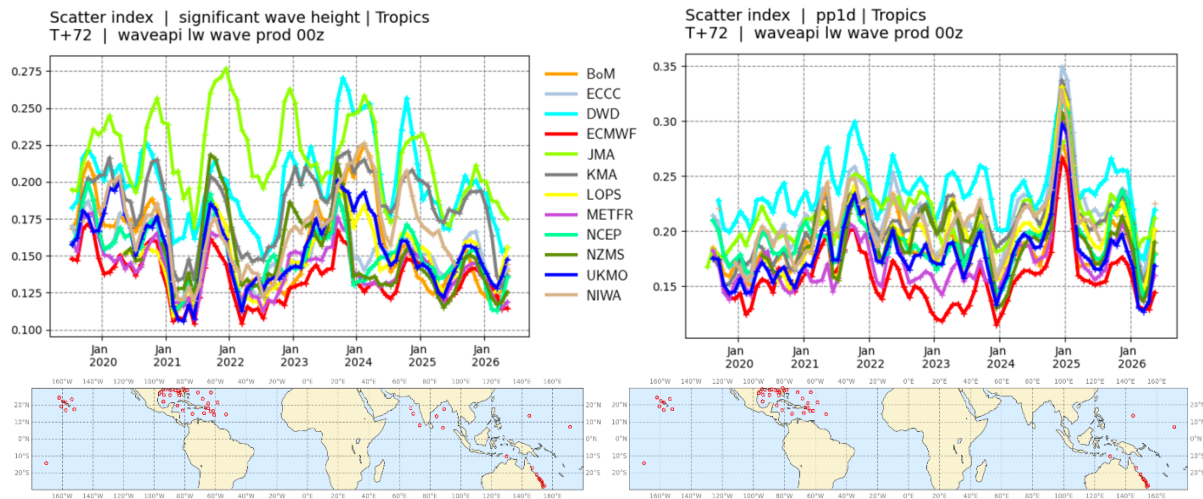
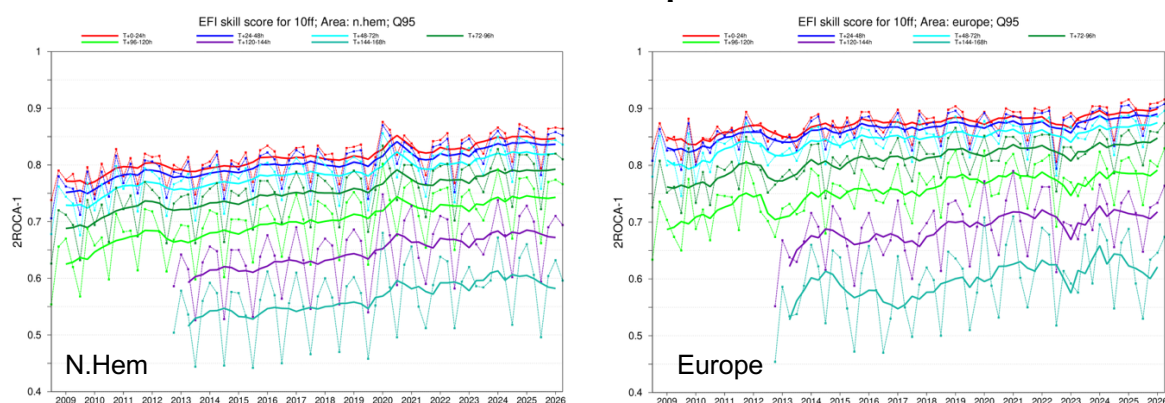
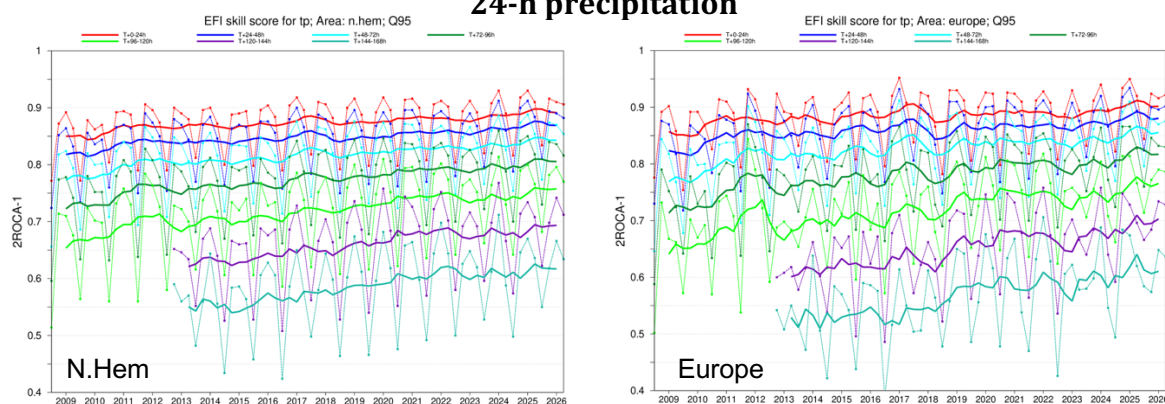


Figure 42: Same as Figure 41 but for the tropics.

10-m wind speed



24-h precipitation



2-m temperature

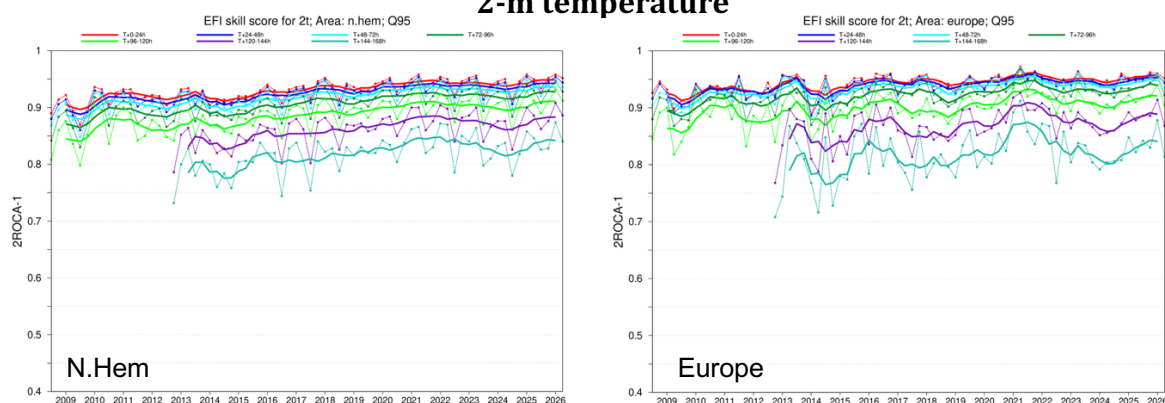


Figure 43: Verification of the Extreme Forecast Index (EFI) against SYNOP observations. Left column: Northern Extratropics, right column: Europe. Top panel: skill of the EFI for 10 m wind speed at forecast days 1 (first 24 hours) to 7 (24-hour period 144-168 hours ahead); skill at day 4 (light green 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.

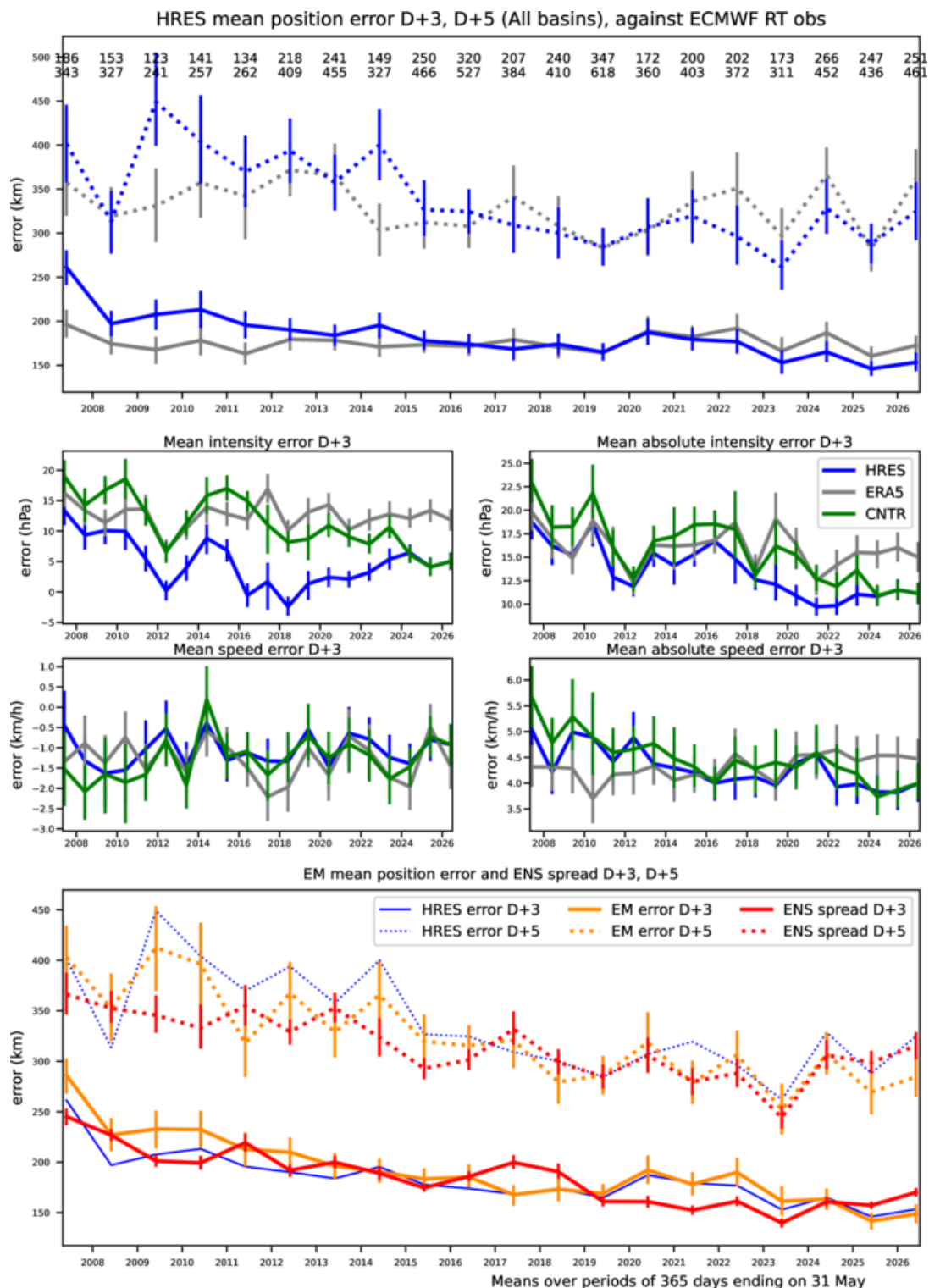


Figure 44: Verification of tropical cyclone predictions from the operational high-resolution and ensemble forecast. Results are shown for all tropical cyclones occurring globally in 12-month periods ending on 31 May. Verification is against the observed position reported via the GTS. Top panel supplementary 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 from HRES (up to 2024) and IFS-CF. 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 IFS-CF 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.

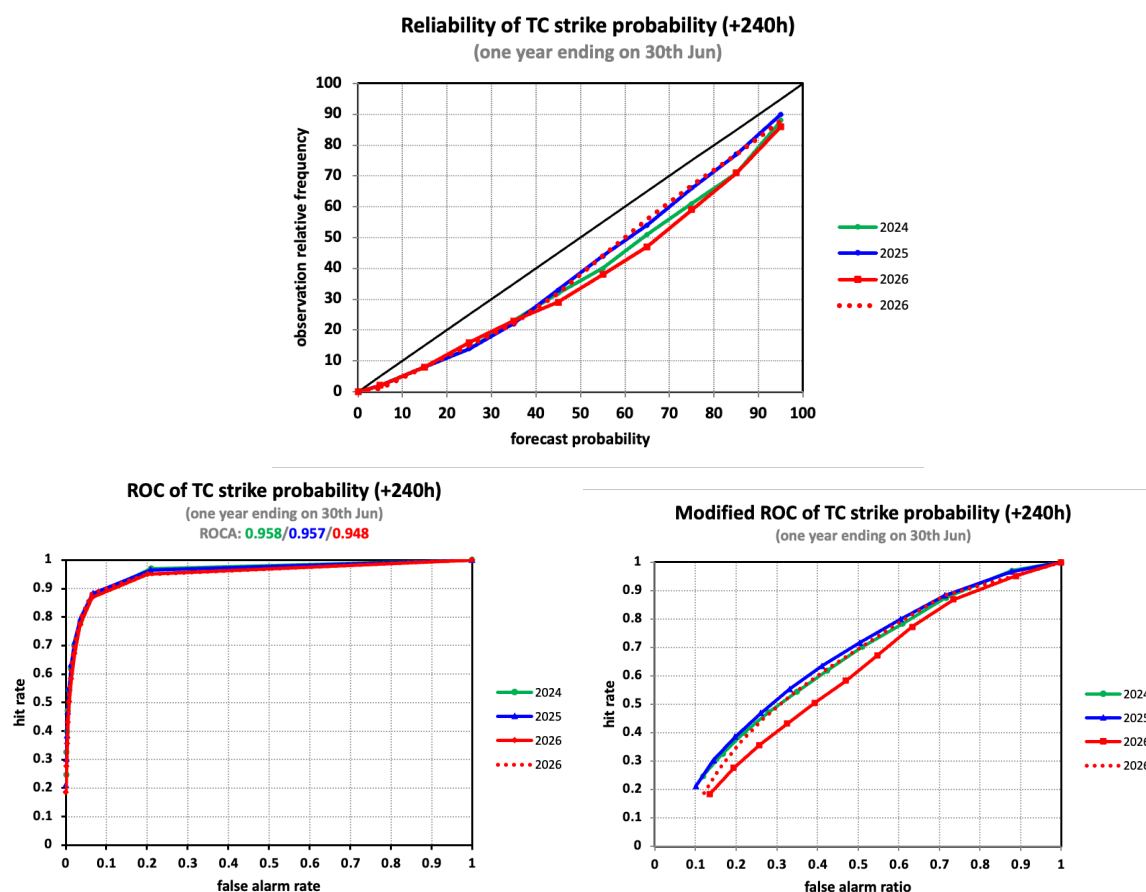


Figure 45: Probabilistic verification of IFS-ENS tropical cyclone forecasts at day 10 for three 12-month periods: July 2023–June 2024 (green), July 2024–June 2025 (blue) and July 2025–June 2026 (red). Red dotted curves show AIFS-ENS (for 2026 only). Upper panel shows reliability diagram (the closer to 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.

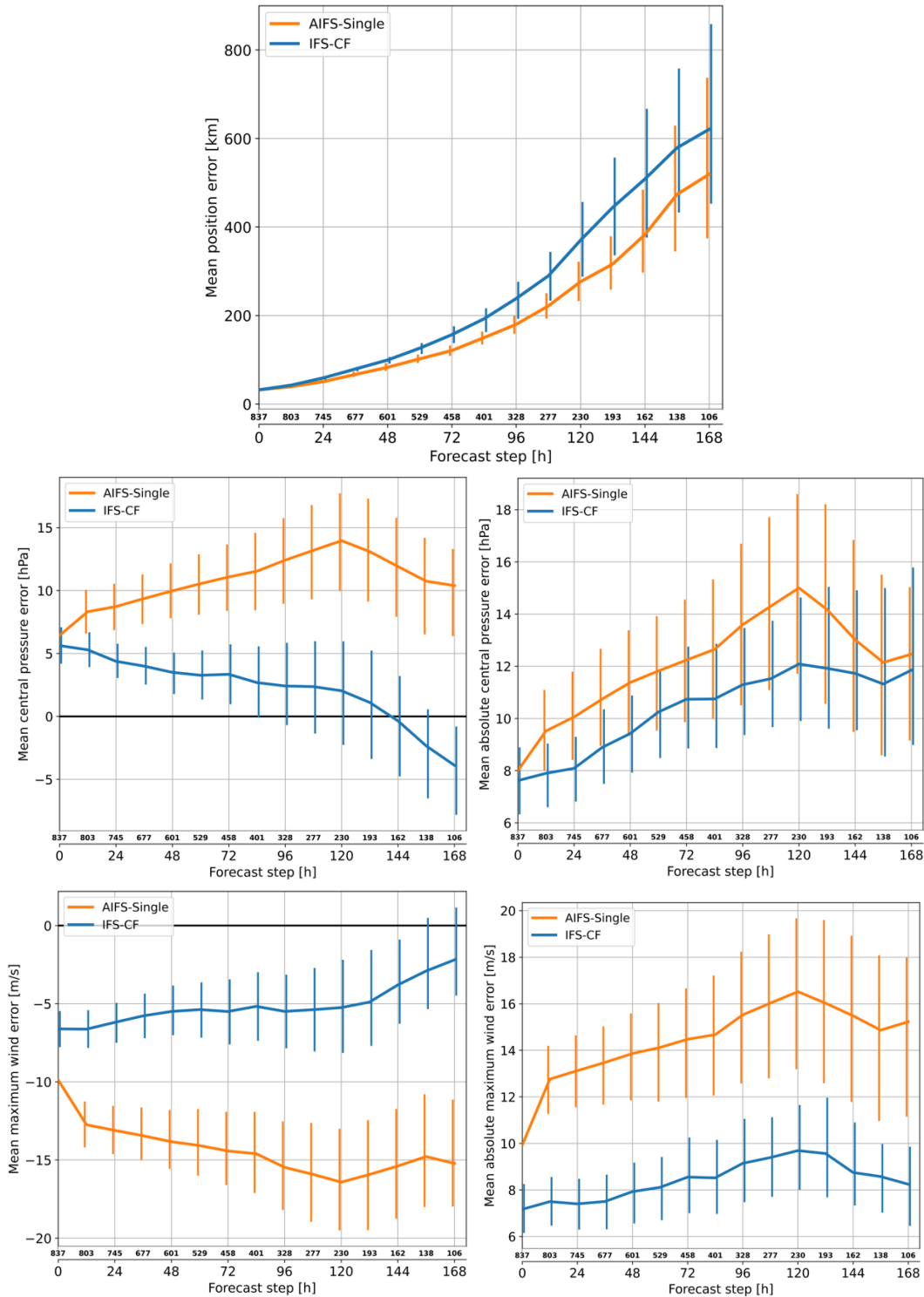


Figure 46: Comparison of IFS-CF (blue) and AIFS-Single (orange) errors in forecasting tropical cyclone position (MAE, top panel), tropical cyclone central pressure (left: ME, right: MAE), and wind speed (left: ME, right: MAE). This is based on a global verification against the International Best Track Archive for Climate Stewardship (IBTrACS v4r01). Verification period is July 2025 – June 2026, 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. An intensity threshold of 17 m/s was applied to filter the observational dataset for TCs that reached tropical storm strength. For each lead time, the number of cases is displayed directly below the graphs. Vertical bars indicate 2.5%–97.5% confidence intervals.

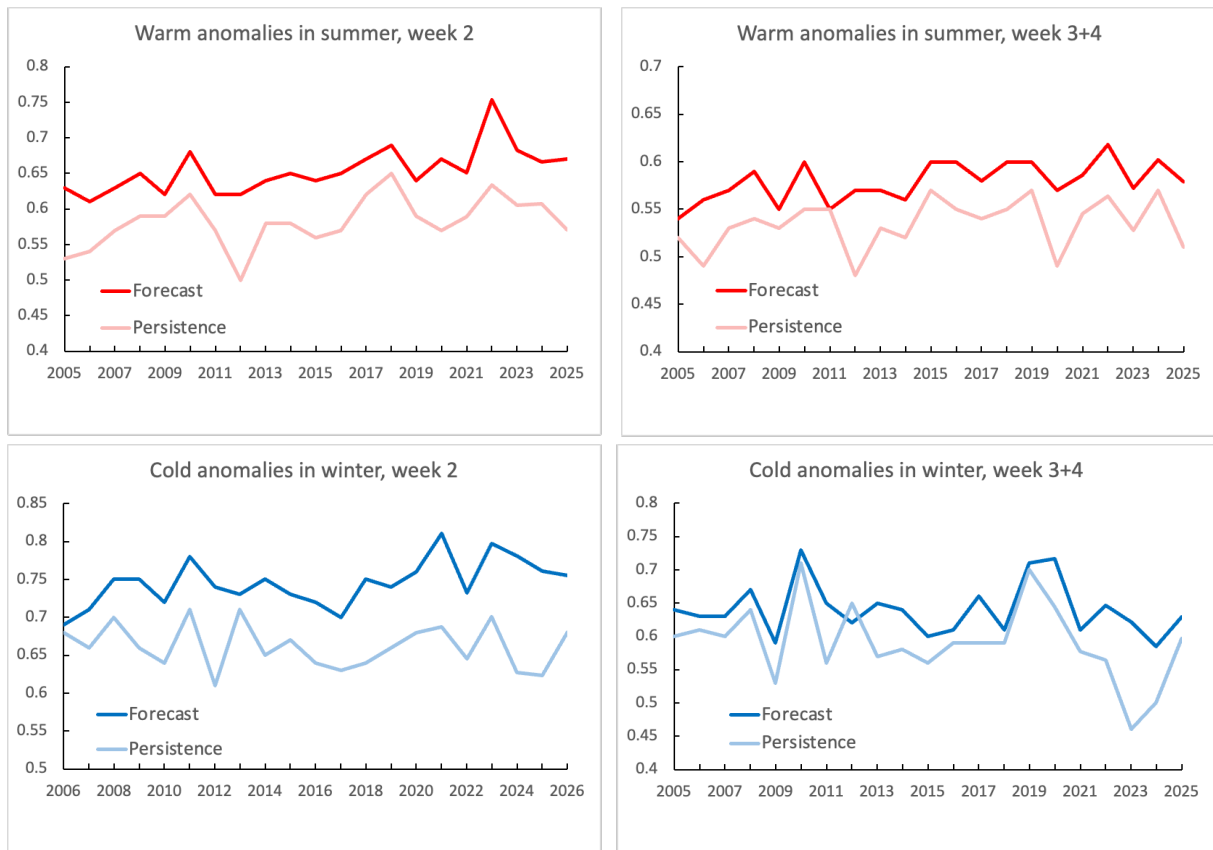


Figure 47: Verification of the sub-seasonal forecast IFS-SUBS against analysis. Area under the ROC curve for the probability that 2 m temperature is in the upper third of the climate distribution in summer (top) and in the lower third in winter (bottom). Scores are calculated for each three-month season for all land points in the extra-tropical northern hemisphere. Left panels show scores 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 scores of persistence of the preceding 7-day period (for week 2) or 14-day period (for weeks 3+4) of the forecast.

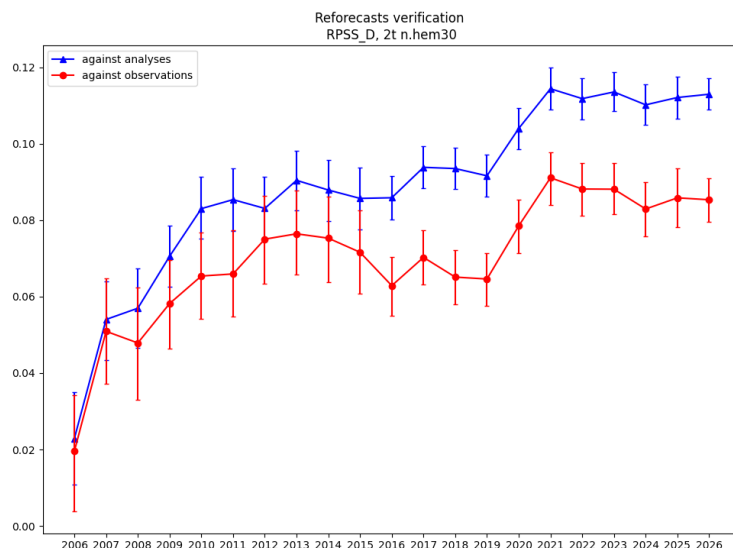


Figure 48: Skill of IFS-SUBS in predicting weekly mean 2 m temperature anomalies (terciles) in week 3 in the northern extratropics (north of 30°N). Verification against ERA5 analysis shown in blue, verification against SYNOP observations shown in red. Verification metric is the Ranked Probability Skill Score.

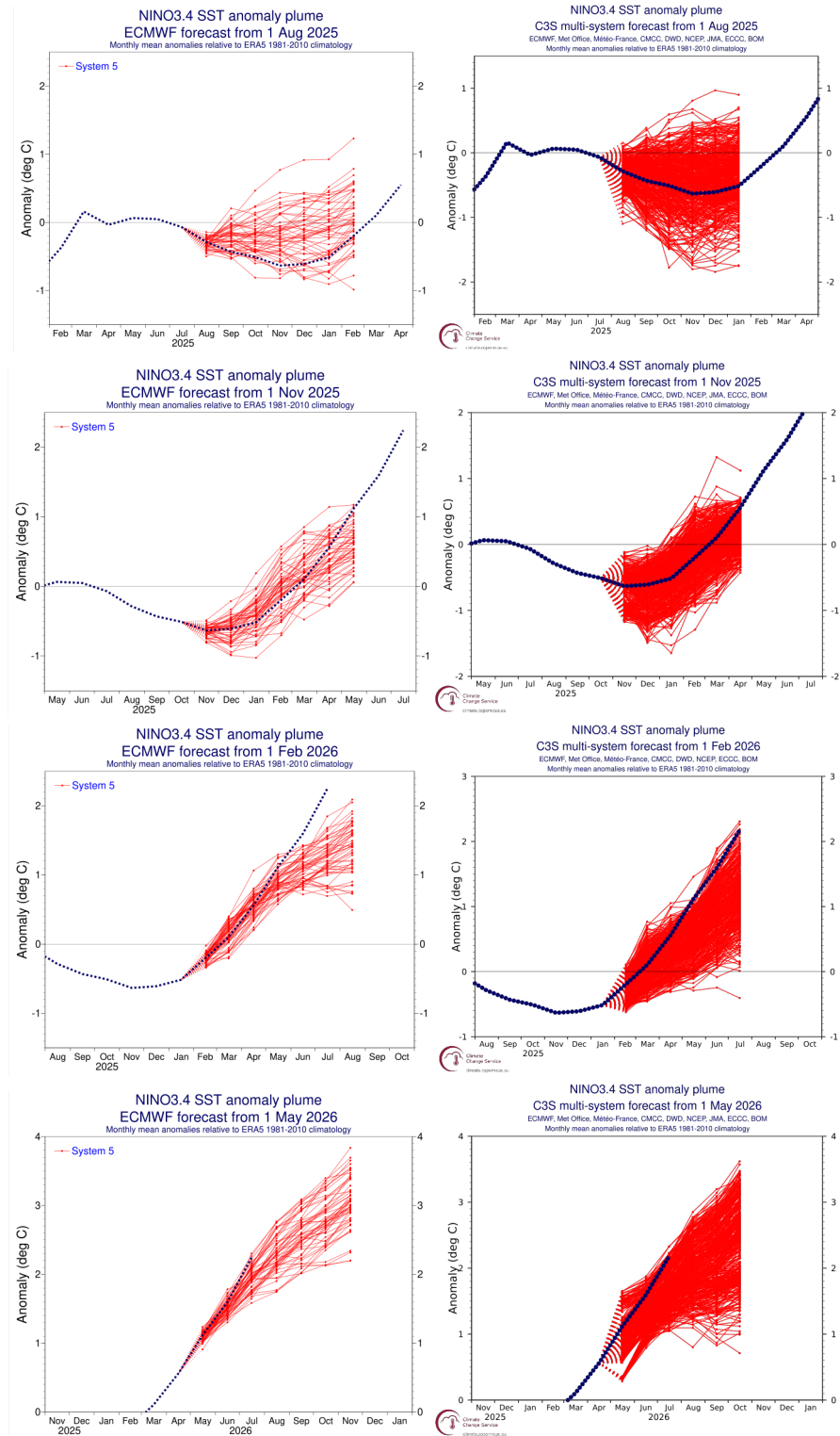


Figure 49: ECMWF Seasonal Forecasting System 5 (SEAS5, left column), and Copernicus Climate Change Service multi-model (right column) seasonal forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from (top to bottom rows) August 2025, November 2025, February 2026, and May 2026. 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 ECCO.

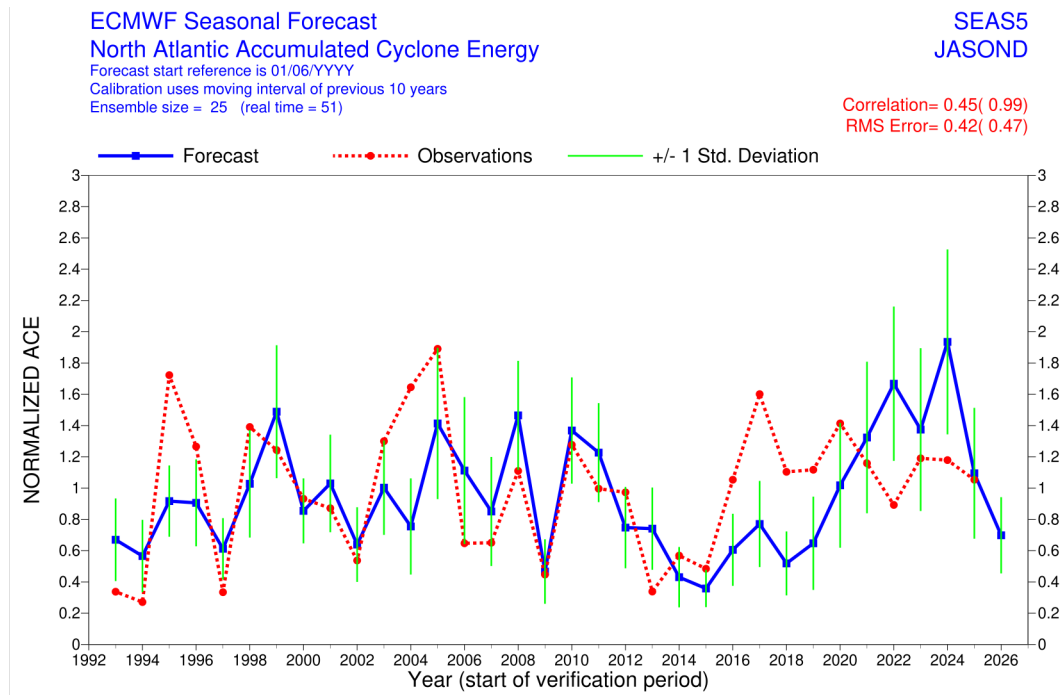


Figure 50: Time series of accumulated cyclone energy (ACE) for the Atlantic tropical storm seasons July–December 1993 to July–December 2025. Blue line indicates the ensemble mean forecasts and green bars show the associated uncertainty (± 1 standard deviation); red dotted line shows observations (up to the 2024 season). Forecasts are from SEAS5 of the seasonal component of the IFS: these are based on the 25-member re-forecasts; from 2017 onwards, they are from the operational 51-member seasonal forecast ensemble. Start date of the forecast is 1 June.

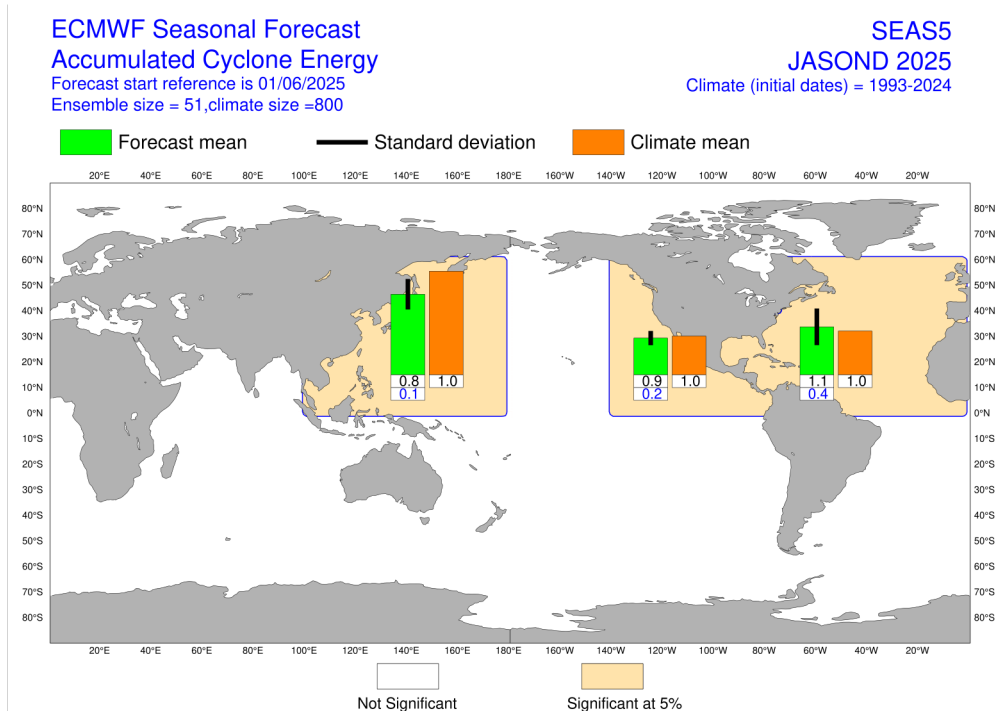


Figure 51: Forecast of tropical storm accumulated cyclone energy (ACE, normalized) issued in June 2024 for the six-month period July–December 2024. Green bars represent the forecast ACE in each ocean basin (ensemble mean); orange bars represent climatology. The values of each bar are written in black 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.

ECMWF Seasonal Forecast

Mean 2m temperature anomaly

Forecast start is 01/11/25, climate period is 1993-2016

Ensemble size = 51, climate size = 600

System 5

DJF 2025/26

Shaded areas significant at 10% level

Solid contour at 1% level

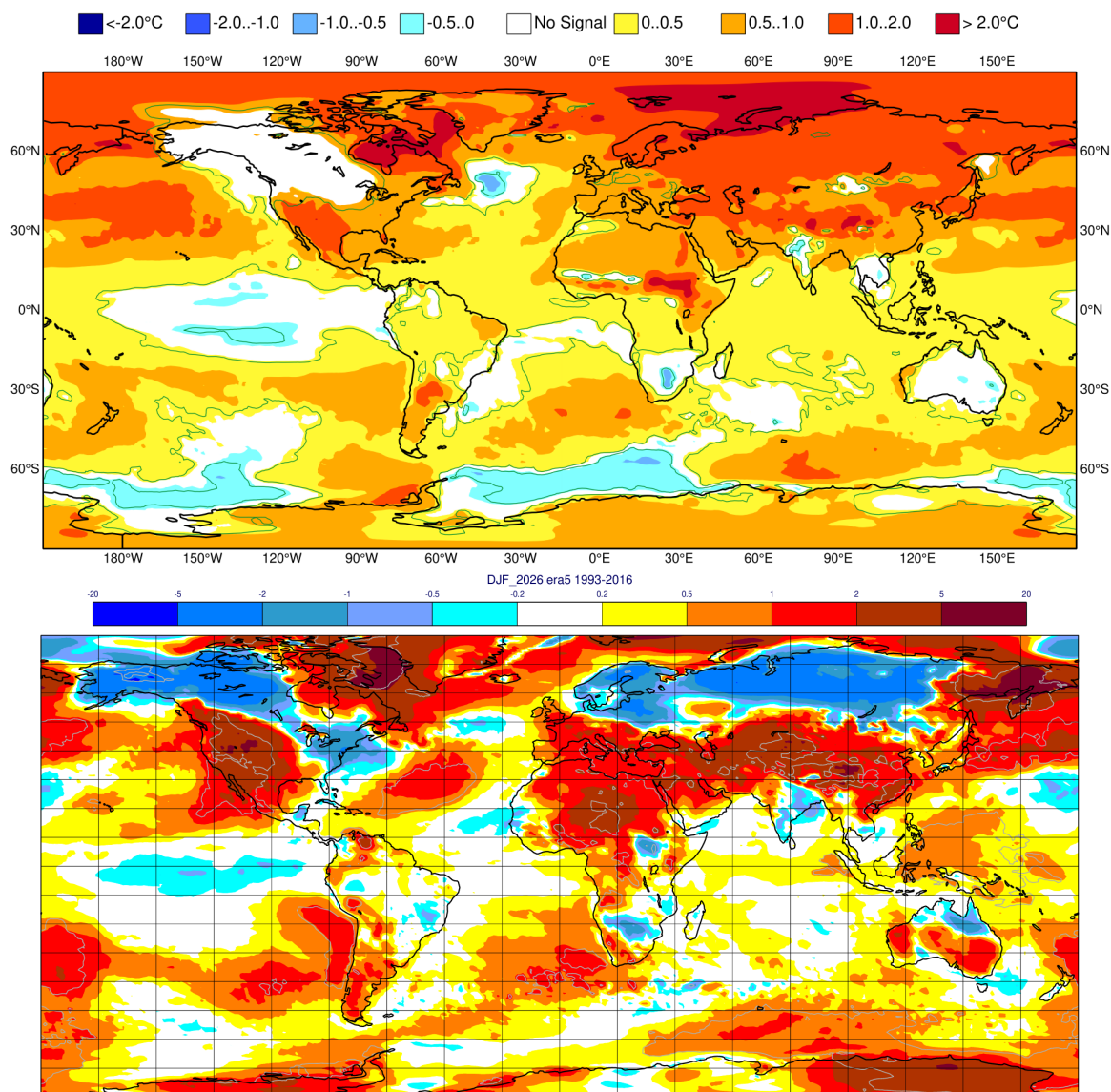


Figure 52: Anomaly of 2 m temperature as predicted by the seasonal forecast from November 2025 for DJF 2025-26 (upper panel) and verifying ERA5 analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

ECMWF Seasonal Forecast

Mean 2m temperature anomaly

Forecast start is 01/05/26, climate period is 1993-2016
Ensemble size = 51, climate size = 600

System 5

JJA 2026

Shaded areas significant at 10% level
Solid contour at 1% level

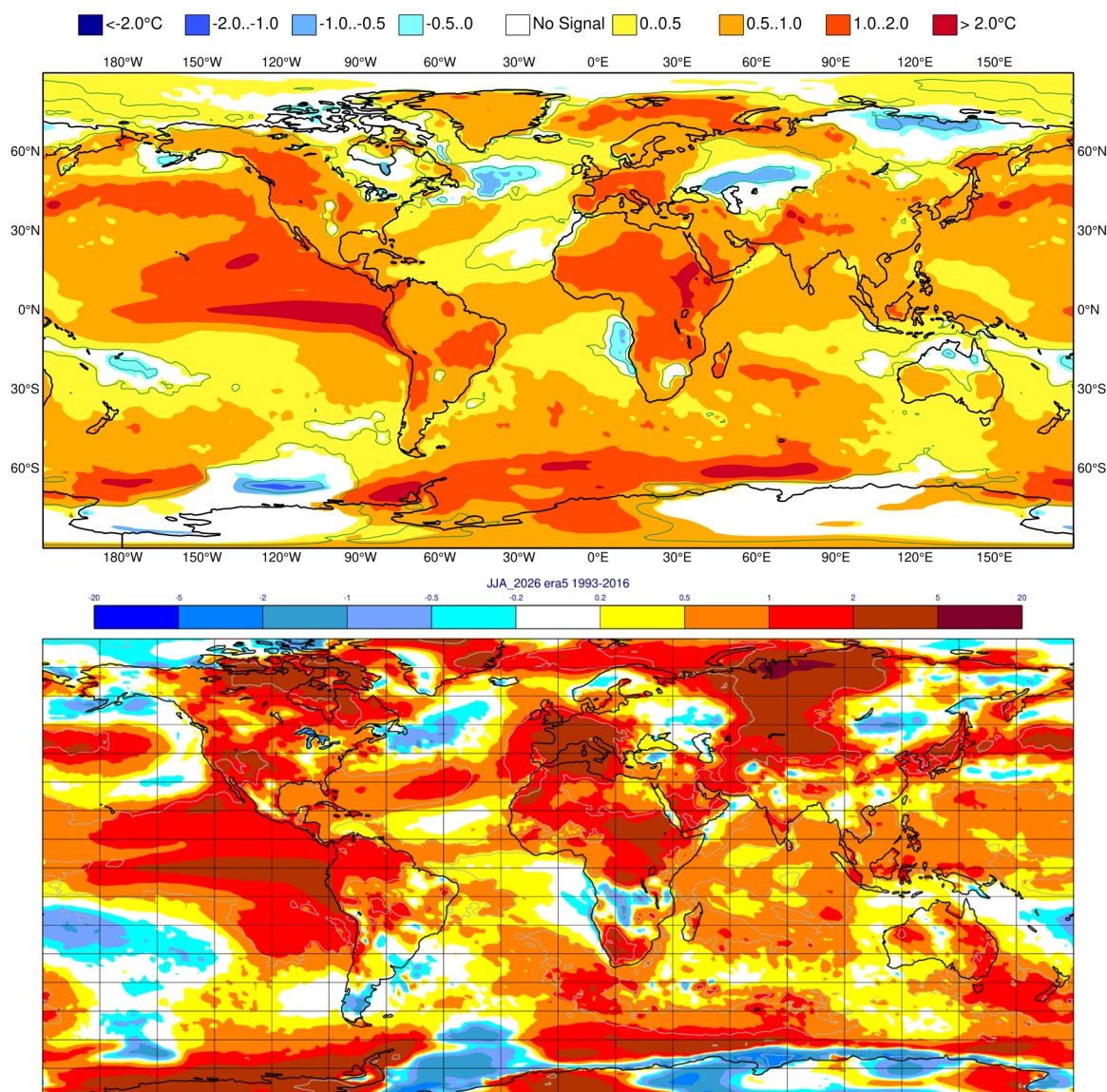


Figure 53: Anomaly of 2 m temperature as predicted by the seasonal forecast from May 2026 for JJA 2026 (upper panel) and verifying ERA5 analysis (lower panel). Grey contours in the analysis indicate regions where anomalies exceed 1.5 standard deviations.

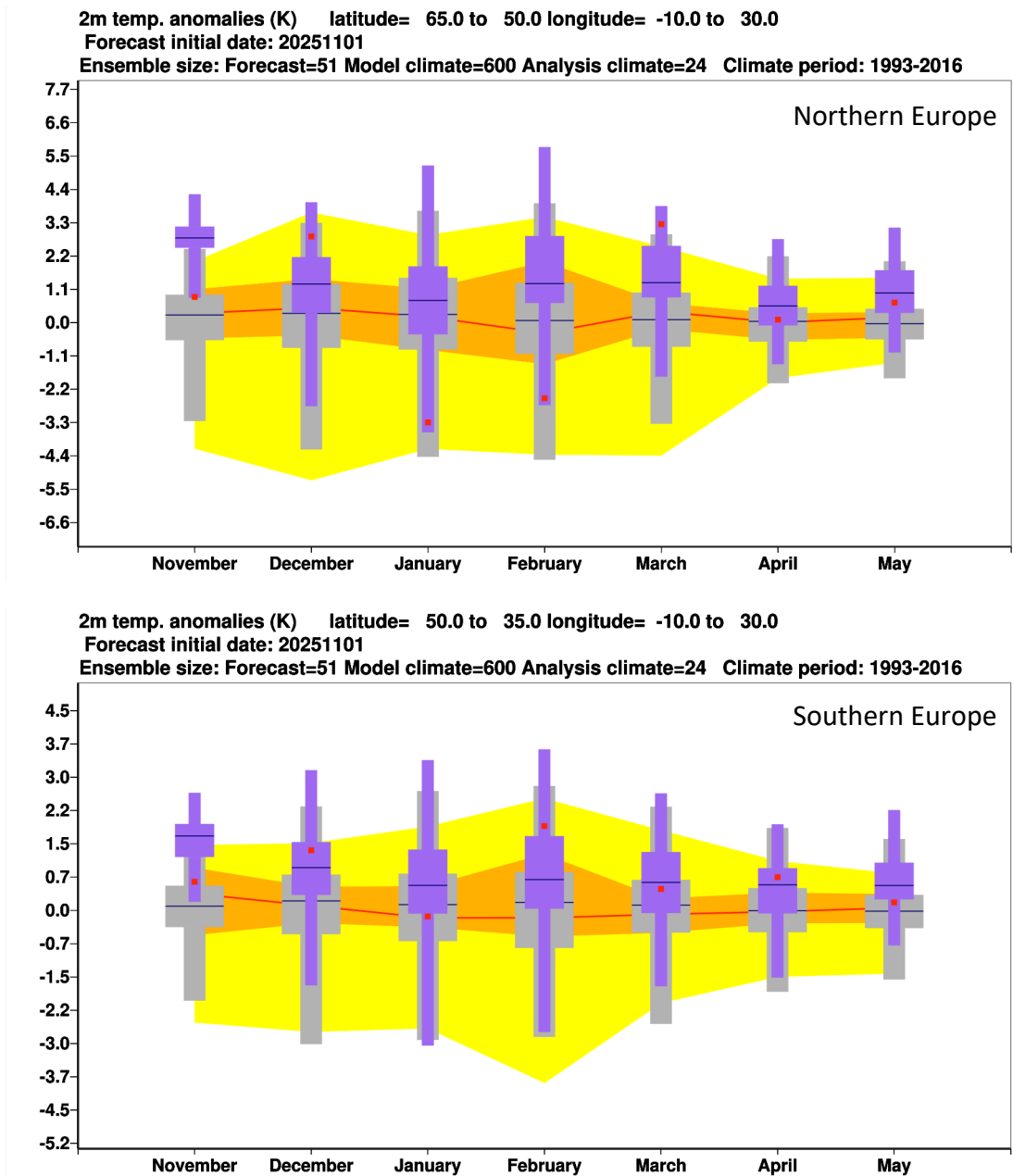


Figure 54: Verification of long-range forecasts of 2 m temperature anomalies (K) from November 2024 for DJF 2024–25 for northern (top) and southern Europe (bottom). The forecast is shown in purple, the model climatology derived from the System-5 hindcasts is shown in grey, and the analysis in the 24-year 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.

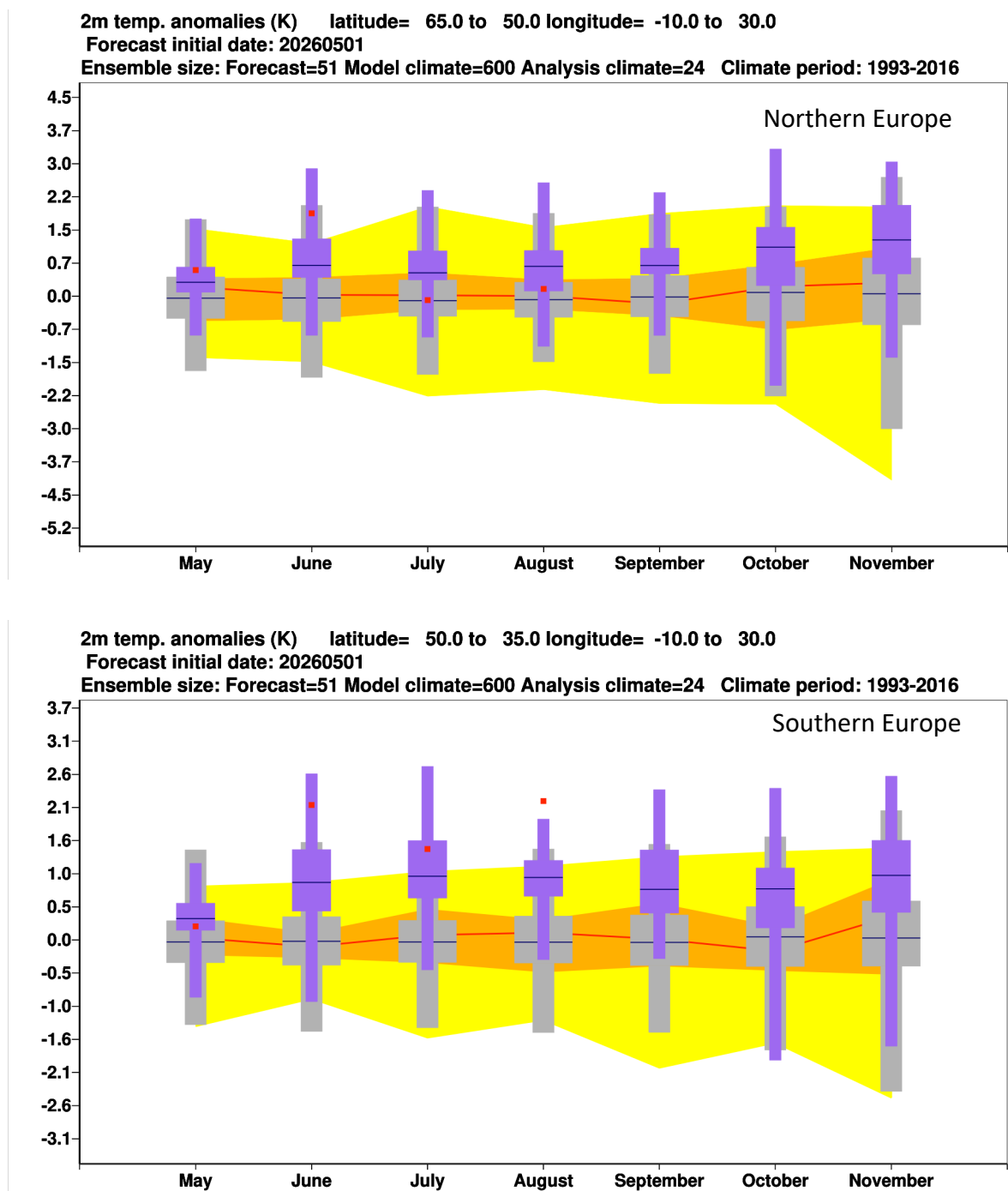
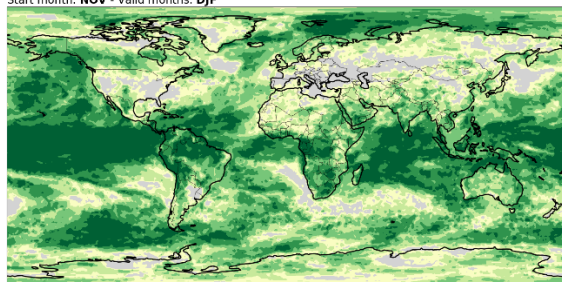
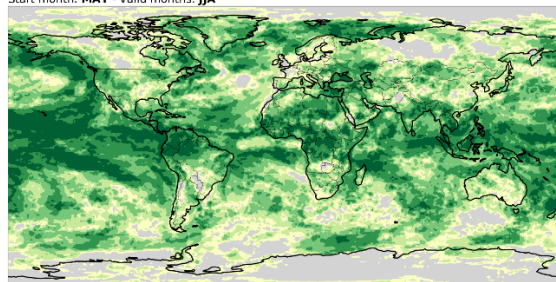


Figure 55: Same as Figure 54 but for forecasts from May 2026 for JJA 2026.

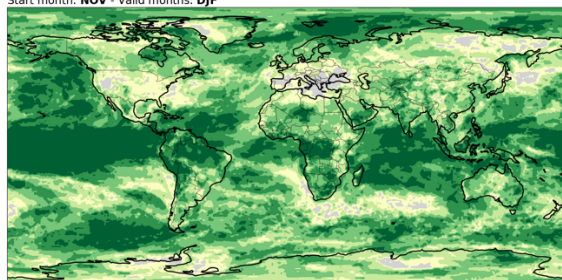
ECMWF SEAS5 (C35v5.1) **2-metre temperature** (lower tercile)
Start month: **NOV** - Valid months: **DJF**



ECMWF SEAS5 (C35v5.1) **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**

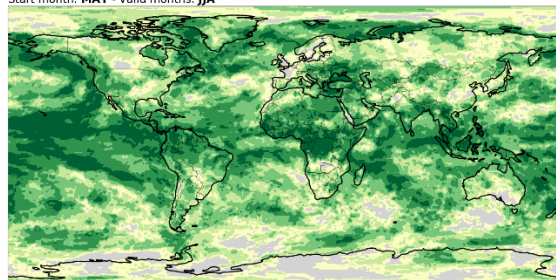
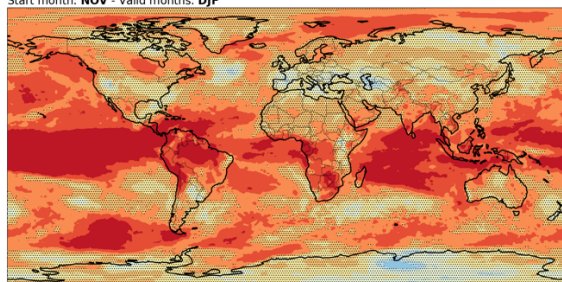
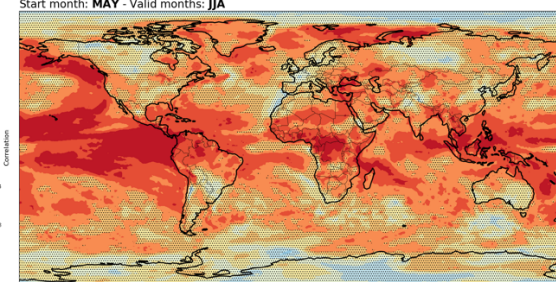


Figure 56: ROC area of seasonal forecasts of 2m temperature in DJF from forecasts initialized in November (left panels), and in JJA from forecasts initialized in May (right panels) over the period 1993-2016. Top panels: ECMWF forecasts, bottom panels: C3S multi-model forecasts. Verification is against ERA5 analyses.

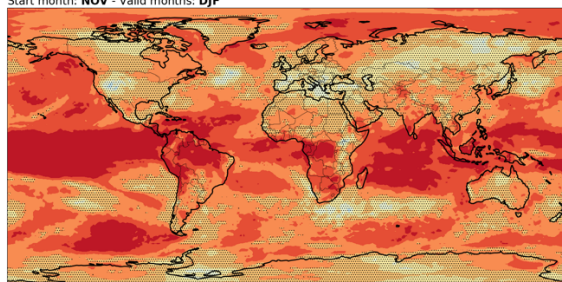
ECMWF SEAS5 (C35v5.1) **2-metre temperature** (stippling where significance below 95%)
Start month: **NOV** - Valid months: **DJF**



ECMWF SEAS5 (C35v5.1) **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**

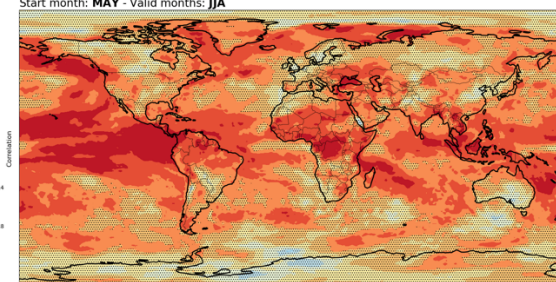


Figure 57: As in Figure 56 but for linear temporal (Spearman) correlation. Stippling indicates significance below 95%.

A short note on scores used in this report

A.1 Deterministic upper-air forecasts

Most of the medium-range verification methodology used follows WMO (WIPPS) recommendations. Scores are computed from forecasts on a standard 1.5×1.5 grid (computed from spectral fields with T120 truncation) and aggregated over standard domains. When other centres' scores are shown, they have mostly been provided as part of the WIPPS exchange of scores among centres or, for some plots, been computed by ECMWF based on an exchange of forecast fields (e.g. through TIGGE).

The root mean square error (RMSE) is the square root of the domain 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 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 some error metric, e.g. RMSE, achieved by the forecast with respect to a reference forecast, such as climatology, persistence, or the ERA5 forecast (Figures 12 and 32):

$$SS = 100 * \left(1 - \frac{RMSE^2}{RMSE_{ref}^2} \right)$$

When anomaly correlation is shown (e.g. Figure 3), it is a spatial correlation between the forecast anomaly and the verifying analysis anomaly (previously relative to ERA-Interim, now relative to ERA5).

A.2 Probabilistic forecasts

Probabilistic skill is evaluated in this report mostly using the continuous ranked probability skill score (CRPSS, e.g. Figures 8, 22, 27) and the area under relative operating characteristic (ROC) curve (e.g. Figures 43 and 45).

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

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

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 ERA5 analysis or observed climatology). CRPSS is used to measure the long-term evolution of skill of the IFS ensemble for upper-air and surface variables.

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. There is also 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), see Figure 45.

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).

The comparison of spread and skill 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 deterministic precipitation forecasts makes use of the SEEPS score (Rodwell et al., 2010), see Figures 25 and 26. 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 as this provides a positively oriented skill score.

Ensemble precipitation forecasts are evaluated with the CRPSS (Figures 25 and 27). Verification is against the same set of SYNOP observations as used for the deterministic forecast. For weather parameters in general, forecast fields 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. In the case of precipitation, the nearest gridpoint is used instead of the bilinear interpolation, and no limit is put on the elevation difference. A gross quality control is applied to SYNOP data (maximum departure from the model forecast must 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 constant lapse rate of 0.0065 K/m.

References

- Benedictow, A., and Co-authors, 2026: Validation report of the CAMS near-real-time global atmospheric composition service: December 2025 - February 2026. Copernicus Atmosphere Monitoring Service (CAMS) report, 64,141.
- Ben-Bouallègue, Z., 2020: Accounting for representativeness in the verification of ensemble forecasts. ECMWF Tech. Memo. 865. <https://doi.org/10.21957/5z6esc7wr>
- Hersbach, H., 2000: Decomposition of the Continuous Ranked Probability Score for Ensemble Prediction System. *Wea. Forecasting*, 15, 559–570.
- Lang, S., and co-authors, 2024: AIFS – ECMWF’s data-driven forecasting system. [arXiv:2406.01465](https://arxiv.org/abs/2406.01465) [physics.ao-ph], <https://doi.org/10.48550/arXiv.2406.01465>
- Polichtchouk, I., S. Massart, and Z. Kipling, 2025: Upgrade to IFS Cycle 50r1. ECMWF Newsletter No. 185, 25-33.
- Presseau, C., D. Kornic, S. J. Peng, S. Ménard, R. Munoz-Alpizar, C. Stroud, V. Savic-Jovicic, and A. Lupu, 2026: Air-Quality Multi-Model Verification for North America 2026-01 to 2026-03. *Environment and Climate Change Canada*, 7-8.
- 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.