TECHNICAL MEMORANDUM

606

Verification statistics and evaluations of ECMWF forecasts in 2008-2009

D.S. Richardson, J. Bidlot, L. Ferranti, A. Ghelli, C. Gibert, T. Hewson, M. Janousek, F. Prates and F. Vitart

Operations Department

October 2009

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

Series: ECMWF Technical Memoranda

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

Contact: library@ecmwf.int

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

1. Introduction

This document presents recent verification statistics and evaluations of ECMWF forecasts. Recent changes to the data assimilation/forecasting and post-processing system are summarised in Section 2. Verification results of the ECMWF medium-range free atmosphere forecasts are presented in Section 3, including, when available, a comparison of ECMWF forecast performance with that of other global forecasting centres. Section 4 deals with the verification of ECMWF forecasts of weather parameters and ocean waves, while severe weather events are addressed in Section 5. Finally, Section 6 provides insights into the performance of monthly and seasonal forecast systems. A short technical note describing the scores used in this report is given in the annex to this document.

In order to aid comparison from year to year, the set of verification scores shown here is mainly consistent with that of previous years (ECMWF Tech. Memos. 346, 414, 432, 463, 501, 504, 547, 578).

Verification pages have been created on the ECMWF web server and are regularly updated. Currently they are accessible at the following addresses:

http://www.ecmwf.int/products/forecasts/d/charts/medium/verification/ (medium-range)

http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/ (monthly range)

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/verification/ (seasonal range)

2. Changes to the data assimilation/forecasting/post-processing system

The changes to the system since the preparation of the last report are summarised below.

30 September 2008: Cycle 35r1, including the following main changes:

- Use of OSTIA high-resolution sea surface temperature produced by the Met Office and the corresponding sea ice analysis product provided by the EUMETSAT ocean and sea ice SAF
- Conserving interpolation scheme for trajectory fields in 4D-Var
- New variational bias correction (VarBC) bias predictors to allow correction of IR shortwave channels affected by solar effects (day/night variations)
- Changes to physics for melting of falling snow, albedo of permanent snow cover (e.g. over Antarctica), diurnal variation of sea surface temperature, and linear parametrization schemes
- Convective contribution added to wind gusts in post-processing

10 March 2009: Cycle 35r2, including the following main changes:

- Revised snow scheme including diagnostic liquid water storage and a new density formulation
- Revised ozone chemistry
- Active assimilation of IASI humidity channels and consistent use of AIRS and IASI humidity channels
- Direct all-sky 4D-Var assimilation of microwave imagers
- Increase of the weight assigned to GPS radio occultation data above 26 km and use of the data up to 50 km

- Satellite-related modifications, including activation of version 9 of the radiative transfer software package RTTOV-9 (developed by NWP-SAF) and revised HIRS cloud detection
- Optimization of the longwave and shortwave radiation schemes
- Extension the domain of the wave model from 81º to 90º N
- Use of ERA-interim analyses for the reforecasts used in the EFI and monthly forecast

12 May 2009: Operational assimilation of temperatures from five Indian radiosonde stations which had upgraded to use French Modem MK2K GPSonde (previously all Indian radiosonde temperatures were blacklisted because of their inconsistency and relatively low quality). Five more sites have since been upgraded and are now also assimilated.

Note: All forecasting system cycle changes since 1985 are described and updated in real-time at:
http://www.ecmwf.int/products/data/operational_system/index.html

3. Verification for free atmosphere medium-range forecasts

3.1. ECMWF scores

3.1.1. Extratropics

Figure 1 shows the evolution of the skill of the deterministic forecast of 500 hPa height over Europe and the extra-tropical northern hemisphere since 1980. Each curve is a 12-month moving average of root mean square error, normalised with reference to a forecast that persists initial conditions into the future. The last month included in the statistics is July 2009. For both regions skill has been consistently high relative to persistence, with skill increasing in particular over Europe at days 4-6. Figure 2 shows the evolution of performance using the anomaly correlation, where reference is to climatology instead of persistence. The figure shows rather consistent scores from month to month over the last year. The average score over the latest 12 months (red line) is slightly lower than the previous year over Europe, mainly because of fewer very high monthly-averaged scores, rather than any increase in months with poor forecasts (e.g. the exceptional predictability for some months in winter 2007-08 (and also 2006-07) was not repeated this year). Figure 2 can be compared with the results for day 8 (T+192 hours) in Figure 1 (day 8 is the approximate time at which the forecast anomaly correlation drops to 60%). Although in Figure 1 the increase in skill over Europe is perhaps rather less at day 8 than at earlier ranges, the skill relative to persistence does not show the same drop in skill as the anomaly correlation. Figure 3 shows that synoptic activity was exceptionally high over Europe in the past year. This is consistent with the results shown in Figure 1 and Figure 2: although predictability has been lower over the last year (resulting in fewer very high anomaly correlation scores), the model performance relative to that of persistence has been maintained and perhaps increased.

Figure 4 illustrates the forecast performance for 850 hPa temperature over Europe. The distribution of daily anomaly correlation scores for day 7 forecasts is shown for each winter (December to February) and summer (June to August) season since winter 1997-98. There have been fewer exceptionally good (anomaly correlation above 80% at day 7) and more forecasts of more moderate skill (50-70%) this summer compared to last year. Most of the lower skill forecasts occurred in June. June 2009 was much more variable over

Europe that in 2007 and 2008, with substantially larger persistence errors; this is probably the reason for the smaller number of very skilful forecasts for this season. The distribution for winter has been consistent for the last 3 years.

Figure 5 shows the time series of the average RMS difference between 4 and 3 day (and 6 and 5 day) forecasts from consecutive days of 500 hPa forecasts over Europe and the northern extratropics. This illustrates the consistency between successive 12 UTC forecasts for the same verification time; the general downward trend indicates that there is less 'jumpiness' in the forecast from day to day. There was a small increase in this measure following the introduction of model cycle 32r3 in November 2007, consistent with the increase in model activity in that cycle. Previous cycles underestimated activity slightly in mid-latitudes and more significantly in the tropics. Changes to the physical parametrizations in 32r3 addressed these deficiencies. The level of consistency between consecutive forecasts has been maintained since this model change.

The quality of ECMWF forecasts for the upper atmosphere in the extratropics is shown through the time series of wind scores at 50 hPa in Figure 6. In both hemispheres, scores for the last year are similar to those for the previous year.

The trend in EPS performance is illustrated in Figure 7, which shows the evolution of the Ranked Probability Skill Score (RPSS) for 850 hPa temperature over Europe and the northern hemisphere. As for the deterministic forecast, the EPS skill was good over the last year, but shows a small drop overall for Europe. The RPSS measures skill relative to climatology (a constant forecast of the climatological probability for each category) and so this drop is consistent with that of the anomaly correlation for the deterministic forecast associated with a more active (less persistent) and probably less predictable flow over Europe. The EPS performance benefited substantially from the increase in resolution in February 2006 (T255 to T399). This is apparent especially in the day 5 and day 7 scores over Europe in Figure 7: despite the small drop in scores, the higher overall skill level since 2006 has been maintained throughout the past year.

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 extra-tropical northern hemisphere for the last three winters are shown in Figure 8. The increase in model activity in Cycle 32r3 (introduced in November 2007) resulted in a significant increase in ensemble spread. The amplitude of the initial perturbations was therefore reduced (by 30%) to maintain the agreement between spread and error. This change resulted in an improved match between spread and error for 500 hPa height; in particular the substantial over-dispersion of the EPS in the early forecast range in previous cycles is no longer apparent. There is now more under-dispersion of the EPS for temperature at 850 hPa to around day seven. However, uncertainty in the verifying analysis should be taken into account when considering the relationship between spread and error in the first few days.

Figure 9 shows the skill of the EPS using RPSS for days 1 to 15 for winter over the extra-tropical northern hemisphere. In November 2006 the EPS was extended to 15 days, at reduced horizontal resolution beyond day 10. Skill in the extended range has been consistent for the three winter seasons since this extension, confirming the positive skill at this forecast range.

3.1.2. Tropics

The forecast performance over the tropics, as measured by root mean square vector errors of the wind forecast with respect to the analysis, is shown in Figure 10. Recent model changes have led to continued improvements, especially in the upper tropospheric winds. The increase in error at 850 hPa at the end of

2007 is associated with the introduction of cycle 32r3. Changes to the physical parametrizations in this cycle increased model activity to higher but more realistic levels, especially in the tropics. The 850 hPa wind error decreased slightly over the last year.

3.2. ECMWF vs other NWP centres

The common ground for comparison is the regular exchange of scores between WMO designated global data-processing and forecasting system (GDPFS) centres under WMO/CBS auspices, following agreed standards of verification. Figure 11 shows time series of such scores over the northern extratropics for both 500 hPa height and mean sea level pressure (MSLP). For 500 hPa height, errors for all models were generally higher in winter 2008-09 than in winter 2007-08, while for MSLP errors were generally similar for both winters. For both parameters, scores for early summer 2009 are similar to 2008 values. ECMWF continues to maintain its lead over the other centres at a similar level to last year; the gap in winter has been reduced since 2007, following improvements to both the UK Met Office and NCEP models. For MSLP the Canadian forecasts show a substantial improvement during 2009 at T+48; this is at least partially a result of a change to their verification procedure made in March 2009. In general, the ECMWF lead has been greater in the southern hemisphere extratropics (Figure 12). However, improvements in the Met Office forecasts have reduced the overall gap compared to previous years.

WMO exchanged scores also include verification against radiosondes over regions such as Europe. Figure 13, showing both 500 hPa height and 850 hPa wind errors averaged over the past 12 months, confirms the good performance of the ECMWF forecasts using this alternative reference relative to the other centres.

The comparison for the tropics is summarised in Figure 14. Since mid-2005, the UK Met Office has had the lowest short-range errors, while at day 5 ECMWF and the UK Met Office performance is similar. The errors of the JMA forecast system have steadily decreased over several years and are now comparable with those of the Met Office model at both short and medium ranges. The large increase in 850 hPa wind error of the Canadian forecasts from 2006 and sudden drop in early 2009 is related to the verification procedure and does not reflect differences in model performance. This does, however, demonstrate the importance of consistent verification methodology, when comparing forecasts from different centres. This matter will be addressed by a new WMO CBS Co-ordination Group on forecast verification that will review the current procedures used for these WMO standard scores; the group is scheduled to meet in autumn 2009.

4. Weather parameters and ocean waves

4.1. Weather parameters - deterministic and EPS

Long-term trends in mean error and standard deviation of error for 2-metre temperature, specific humidity, total cloud cover and 10 m wind speed forecasts over Europe are shown in Figure 15 to Figure 18. Verification is against synoptic observations available on the GTS. A correction for the difference between model orography and station height was applied to the temperature forecasts, but no other post-processing has been applied to the model output. In general the performance over the past year follows the trend of previous years and there is no adverse effect on the scores from the increase in atmospheric activity introduced in model cycle 32r3 in November 2007.

Both winter 2007-08 and 2008-09 had significant negative night-time temperature biases (Figure 15). This was noted by users who, in particular, reported cases of large errors on specific occasions; these were often associated with errors in both snow and cloud cover that resulted in much colder forecast temperatures than

subsequently observed. Changes to the snow parametrization in cycles 35r1 and 35r2 have made some improvements to the forecast snow cover (noted by users at the 2009 Forecast Products Users Meeting). However, this does not account for the overall negative bias which is still under investigation. Figure 15 confirms that although the night-time bias was large last winter, the error standard deviation was not strongly affected.

While specific humidity bias has remained fairly constant over recent years, the summer error standard deviation has consistently decreased. The negative daytime bias in cloud cover has reduced substantially since 2005 (Figure 17) and is now close to zero for both daytime and night-time; error standard deviation is also decreasing. Physics changes introduced in model cycle 31r1 (September 2006) increased 10 m wind speeds globally, generally improving negative biases in many regions. Over Europe this resulted in a change from negative to positive overall bias for daytime forecasts (Figure 18), but did not adversely affect the error standard deviation. This positive bias has been reduced over the past year.

The trend in precipitation skill for Europe is shown in Figure 19, using the True Skill Score (or Pierce's Skill Score) for thresholds of 10 mm and 20 mm per day. As noted in the past report there has been a consistent improvement since the introduction of cycle 31r1 in September 2006. For both 10 mm/day and the higher threshold of 20 mm/day, the exceptional performance over 2007-08 was matched in 2008-09. The same overall trend can be seen in the scores for the EPS probability forecasts shown in Figure 20.

4.2. Ocean waves

The quality of the ocean wave model analysis is shown in the comparison with independent ocean buoy observations in Figure 21. In general the errors in 2008-09 are similar to those of 2007-08. The improvement in the analysis since the introduction of JASON altimeter data in February 2006 is clear. Figure 21 also shows a time series of the analysis error for the 10 metre wind over maritime regions using the wind observations from the same set of buoys. The error has steadily decreased since 1998, providing better quality winds for the forcing of the ocean wave model and this year has been similar to last year.

The good performance of the wave model forecasts is confirmed this year, as shown in Figure 22 and Figure 23. This is particularly noticeable in the verification against observations and comparison with other wave models, as shown in Figure 24. The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of ocean buoys (mainly located in the northern hemisphere).

A comprehensive set of wave verification charts is now available on the ECMWF web site, including the figures shown in this report: http://www.ecmwf.int/products/forecasts/wavecharts/

5. Severe weather

5.1. Extreme Forecast Index (EFI)

The Extreme Forecast Index (EFI) was developed at ECMWF as a tool to provide some general guidance on potential extreme events. By comparing the EPS distribution of a chosen weather parameter to the model's climatological distribution, the EFI indicates occasions when there is an increased risk of an extreme event occurring. Verification of the EFI has been performed using synoptic observations over Europe from the GTS. An extreme event is judged to have occurred, if the observation exceeds the 95th percentile of the observed climate for that station (calculated from a 15 year sample, 1993-2007). The ability of the EFI to detect extreme events is assessed using the Relative Operating Characteristic (ROC). Results are presented in Figure 25 for each season from winter 2004-05 to spring 2009 for precipitation, 10 m wind speed and 2 m

temperature. For all parameters, there is a clear improvement during this period, especially for the 5-day forecast.

5.2. Tropical cyclones

The 2008 North Atlantic hurricane season was very active, with 16 tropical storms including 8 hurricanes. This was well predicted by the seasonal forecast system: the forecast from June 2008 predicted 15 tropical storms and 9 hurricanes in the North Atlantic.

Average position and intensity errors for the deterministic medium-range forecasts of all tropical cyclones (all ocean basins) over the last seven 12-month periods are shown in Figure 26. A significant reduction in intensity errors was reported for 2007-08, compared with the previous periods. This improved performance is confirmed for the most recent period: the intensity errors are similar to (or even slightly lower) than those for 2007-08. This is the case for both the mean intensity error (bottom left panel of Figure 26) and the mean absolute error of the TC intensity (bottom right panel of Figure 26). Position errors have gradually decreased over recent years, although the initial errors (the analysis errors) have not changed significantly. There is a clear tendency in the forecast for tropical cyclones to move too slowly (negative along-track error), despite some improvements over the last few years.

The EPS tropical cyclone forecast is presented on the ECMWF web site as a strike probability: the probability at any location that a reported tropical cyclone will pass within 120 km during the next 120 hours. Verification of these probabilistic forecasts for the three latest 12-month periods is shown in Figure 27. Results show an over-confidence for the three periods, with small variations from year to year. The signal detection capability (as indicated by ROC) has improved in both the last two years. This is particularly evident in the modified ROC which uses the false alarm ratio instead of the false alarm rate on the horizontal axis (this removes the reference to the non-events in the sample and shows more clearly the reduction in false alarms in those cases where the event is forecast).

6. Monthly and seasonal forecasts

6.1. Monthly forecast verification statistics and performance

The monthly forecasting system has been integrated with the medium-range Ensemble Prediction System (EPS) since March 2008. The new, combined system enables users to be provided with EPS output uniformly up to 32 days ahead, once a week. It also introduced a coupled ocean-atmosphere model for the forecast range day 10 to 15 for the forecast started from the 00 UTC analysis, on a daily basis.

Figure 28 shows the ROC area score computed over each grid point for the 2 m temperature monthly forecast anomalies at two forecast ranges: days 12-18 and days 19-25. All the real-time monthly forecasts since 7 October 2004 have been used in this calculation. The red colours correspond to ROC scores higher than 0.5 (the monthly forecast has more skill than climatology) and the blue colours correspond to ROC scores below 0.5 (the monthly forecast has less skill than climatology). Currently the anomalies are relative to the past 18-year model climatology. The monthly forecasts are verified against the ERA40 reanalysis or the operational analysis, when ERA40 is not available.

Although these scores are strongly subject to sampling, they provide the user with a first estimate of the forecast skill's spatial distribution.

Comprehensive verification for the monthly forecasts is available on the ECMWF website at: http://www.ecmwf.int/products/forecasts/d/charts/mofc/verification/.

6.1.1. Monthly forecasts' performance 2008-2009

Figure 29 shows the probabilistic performance of the monthly forecast over each individual season since September 2005 for the time ranges days 12-18 and days 19-32. The figure shows the ROC scores for the probability that the 2 metre temperature is in the upper third of the climate distribution over the extra-tropical northern hemisphere.

The monthly forecast system has continued to perform well at forecast range 12-18 days. At days 19-32 the skill for the northern hemisphere over the winter months was not as good as in previous years; however, this may be explained solely by the year-to-year variability, since the skill of the persistence forecast was also poor during the most recent winter season. The monthly forecast system has consistently outperformed persistence by a substantial margin for each season over the past year. Overall, the cold conditions observed over Europe between December 2008 and January 2009 were well predicted (not shown).

6.2. The 2008-2009 El Niño forecasts

After the peak of the recent La Niña event in February 2008 the sea-surface temperature (SST) anomalies across the central and eastern Pacific declined and by July 2008 SSTs were close to their climatological mean values. During autumn 2008 temperatures started to cool again, giving rise by January 2009 to another La Niña event but with more moderate amplitudes. During the first part of 2009, consistent with La Niña conditions, the atmospheric circulation showed enhanced trade winds across the central and western Pacific. The Equatorial Southern Oscillation index SOI was positive between November 2008 and May 2009.

For the seasonal forecast system, the return of La Niña conditions in September 2008 was a rather difficult event to predict. The majority of international dynamical models, including ECMWF's, predicted a continuation of neutral conditions for the tropical Pacific until at least mid-winter, while some models predicted the development of El Niño conditions. Given the wide range of predicted outlooks, the forecast uncertainty was very high. The EUROSIP predictions issued in August 2008 (Figure 30) presented a large spread, with the most probable outcome being neutral El Niño-Southern Oscillation conditions: anomalies close to zero or small negative anomalies. It is interesting to note that, although the observed sea surface temperature (SST) anomalies are colder than most ensemble members, they remain within the predicted range. Seasonal predictions for the recent seasons (December 2008 to February 2009 and March to May 2009) were realistic over the tropics and, to some extent, over the Pacific sector. This is consistent with the spatial distribution of the seasonal forecast skill based on the 25 years of past performance currently available on the ECMWF website.

Since the beginning of the year, the ECMWF forecasts have been consistently forecasting a change to El Niño conditions during 2009, with the El Niño continuing into 2010, as shown in the annual range forecasts started in February and May 2009 (Figure 31). There has been some uncertainty over the amplitude of the El Niño: the latest forecasts suggest that the SST anomalies in the Niño 3.4 region of the Pacific will more probably be around 1°C, while some earlier forecasts suggested a stronger warming.

6.3. Seasonal forecast performance for the tropical SST

Verification of the tropical SST seasonal forecasts is presented alongside the forecast products on the ECMWF web site, for example:

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/seasonal_range_forecast/group/seasonal_charts_sst/

These verification statistics are based on the hindcast integrations run over the period 1987 to 2005.

To complement this information, new skill estimates have recently been computed to establish the skill ranges of a persistence forecast and of a "perfect model" forecast. Figure 32 shows the temporal anomaly correlation coefficient (ACC) of the forecasts with the verifying analysis, calculated at each grid point, for seasonal mean forecasts with a one month lead time. The ACC for the ensemble-mean System 3 operational forecasts (S3) can be compared to that obtained by persisting the observed monthly mean SST anomaly from the month immediately prior to the start of the forecast. The "perfect model" score is obtained by replacing the verifying analysis by one of the ensemble members (as a proxy for "truth") in calculating the ACC for the ensemble mean. This perfect model ACC represents the upper limit of the score that could be achieved with a perfect model, if the initial perturbations of the S3 ensemble correctly represent the uncertainty in the initial conditions and if real world has predictability characteristics similar to those of the S3 model.

Figure 32 shows results for the June-August (JJA) season. The operational S3 forecasts (middle panel) have substantially higher skill than the persistence forecasts (top panel) over most of the tropics. However, the perfect model results (bottom panel) show that performance of the current system is still substantially below the upper limit of what might one day be reached. At longer lead times (not shown), persistence forecasts generally lose skill more rapidly than the operational system, enabling the S3 model to perform substantially better than persistence in a number of regions, but, again, the performance is a long way behind the estimated upper limit. The advantage of S3 predictions over persistence for the tropical Pacific is highest for the JJA season.

Figure 32 shows very high potential predictability of SSTs in the southern hemisphere winter. However, the S3 forecasts have low skill in the southern hemisphere, in many places below that of persistence. This may be because the ocean initial states in these latitudes during the 1981-2005 period contained little useful information - perhaps not surprising, given the almost complete lack of in-situ ocean data. It is expected that when the ARGO array is in place, SST prediction in this part of the ocean will improve, although this has yet to be confirmed.

In assessing Figure 32 it is important to remember that the main source of seasonal forecast predictability is associated with the SST predictions over the tropical oceans, and more particularly over those regions where small local changes in SST can lead to large spatial shifts in atmospheric deep convection.

6.4. Seasonal forecast performance for the global domain

A set of verification statistics based on the hindcast integrations (1981-2005) from the operational System 3 has been produced and is also available on the ECMWF website at:

http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/seasonal_range_forecast/group/seasonal_charts_2tm/

A set of verification statistics based on the hindcast integrations (1987-2005) from the operational EUROSIP multi-model is under development. The skill measures are the same used to evaluate the ECMWF seasonal forecast system.

A multi-model forecast system might give better information on the uncertainty of the forecast than a single-model system, especially where there are substantial model uncertainties, as is typical for seasonal systems. A forecast Probability Density Function (PDF) derived from a multi-model combination will typically be

broader than one derived from a single-model because the multi-model PDF naturally takes account of model uncertainty. The broader PDFs of multi-model forecasts may increase their reliability and, therefore, improve the probabilistic performance compared to the individual single model systems. In addition, the EUROSIP system has three times as many ensemble members as the individual contributing systems. This large increase in ensemble size should also help to provide more reliable probability estimates.

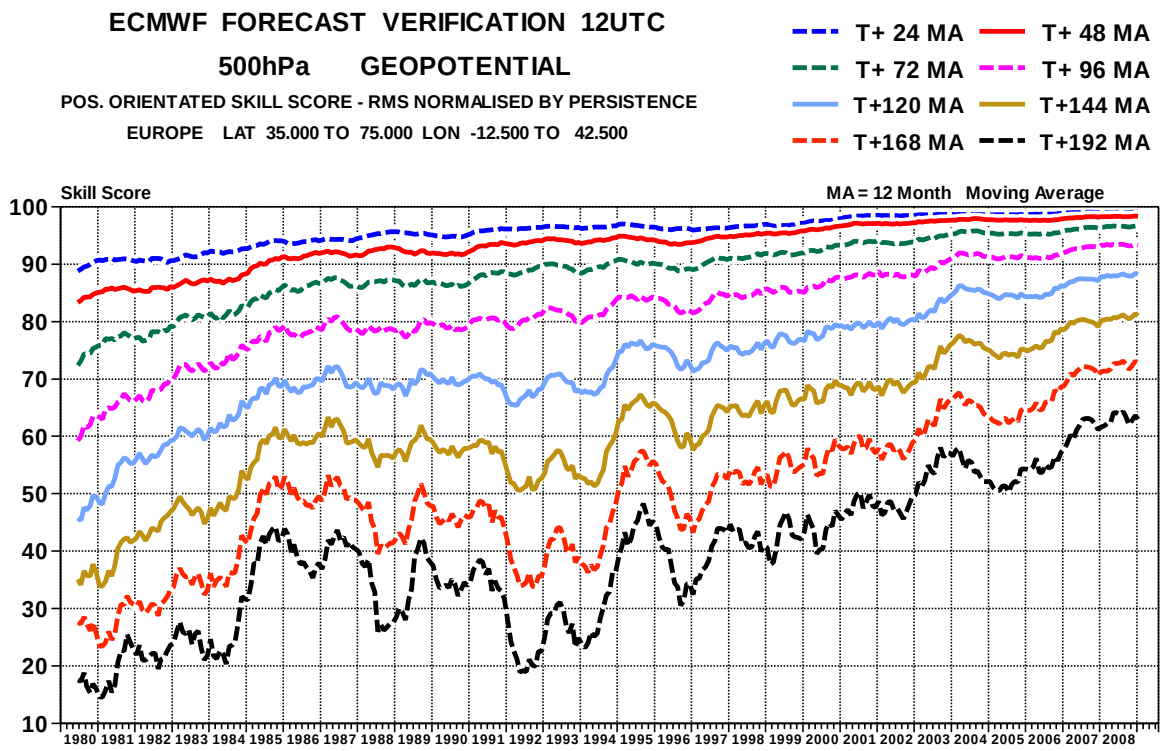
Figure 33 shows reliability diagrams for the EUROSIP multi-model system and the single-model ECMWF system and demonstrates the benefit of the EUROSIP system for the reliability of probability forecasts. The upper panel shows the reliability of probability forecasts from the ECMWF model for seasonal mean 2 m temperature being in the bottom third of the climate distribution for December-February for forecasts in the northern hemisphere extra-tropics. For a perfectly reliable set of forecasts, the observed frequency of occurrence would match the forecast probability i.e. the points would be on the diagonal. Although the ECMWF forecasts have some ability to discriminate between different likelihoods of a cold winter, they are clearly some way from being completely reliable. The bottom panel shows the result from the EUROSIP forecasts. The result is still not perfectly reliable, but is a substantial improvement on the ECMWF result: note, for example, the change in the forecasts of very low probability of a cold winter. The EUROSIP combination makes such a prediction less often (the size of the plotted circle represents the frequency with which the forecast probability is issued), but when such a forecast is made, it is much more reliable.

7. References

Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. *ECMWF Tech. Memo* **430**.

Vitart, F., S.J. Woolnough, M.A. Balmaseda and A. Tompkins, 2007: Monthly forecast of the Madden-Oscillation using a coupled GCM. *Monthly Weather Review*, **135**, 2700-2715.

List of Figures



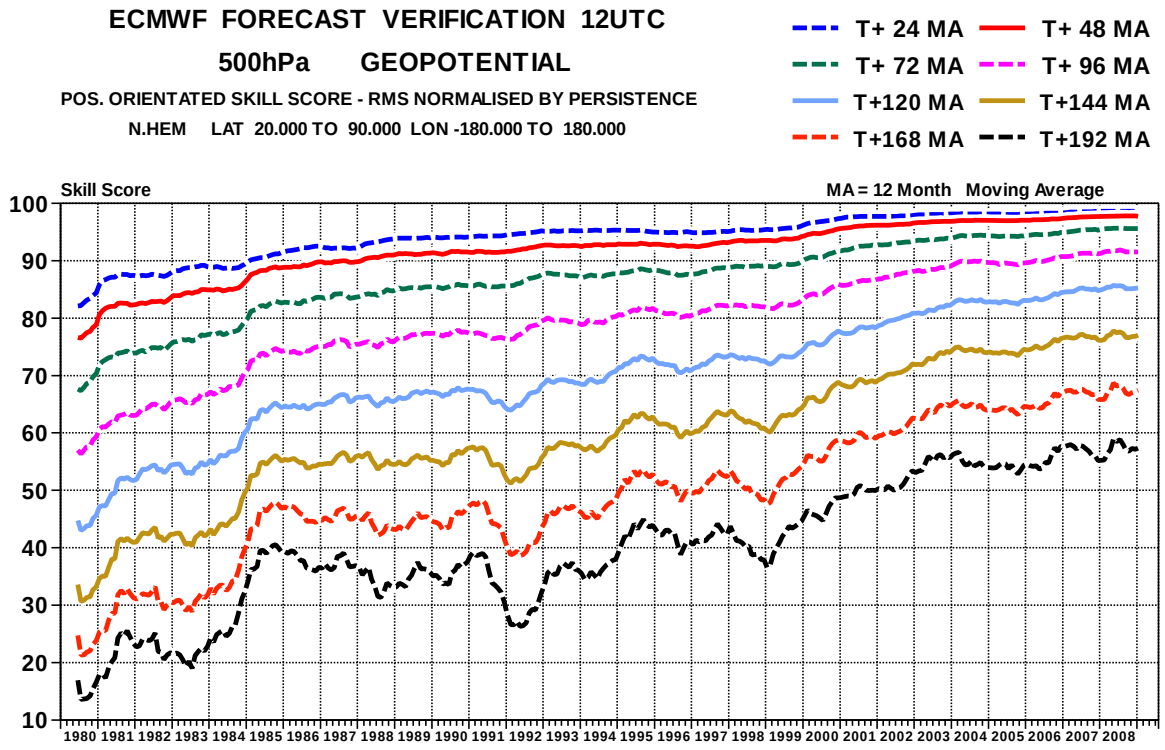


Figure 1: 500 hPa height skill score for Europe (top) and the northern hemisphere extra-tropics (bottom), 12-month moving averages, forecast ranges from 24 to 192 hours. The last point on each curve is for the 12-month period August 2008 - July 2009.

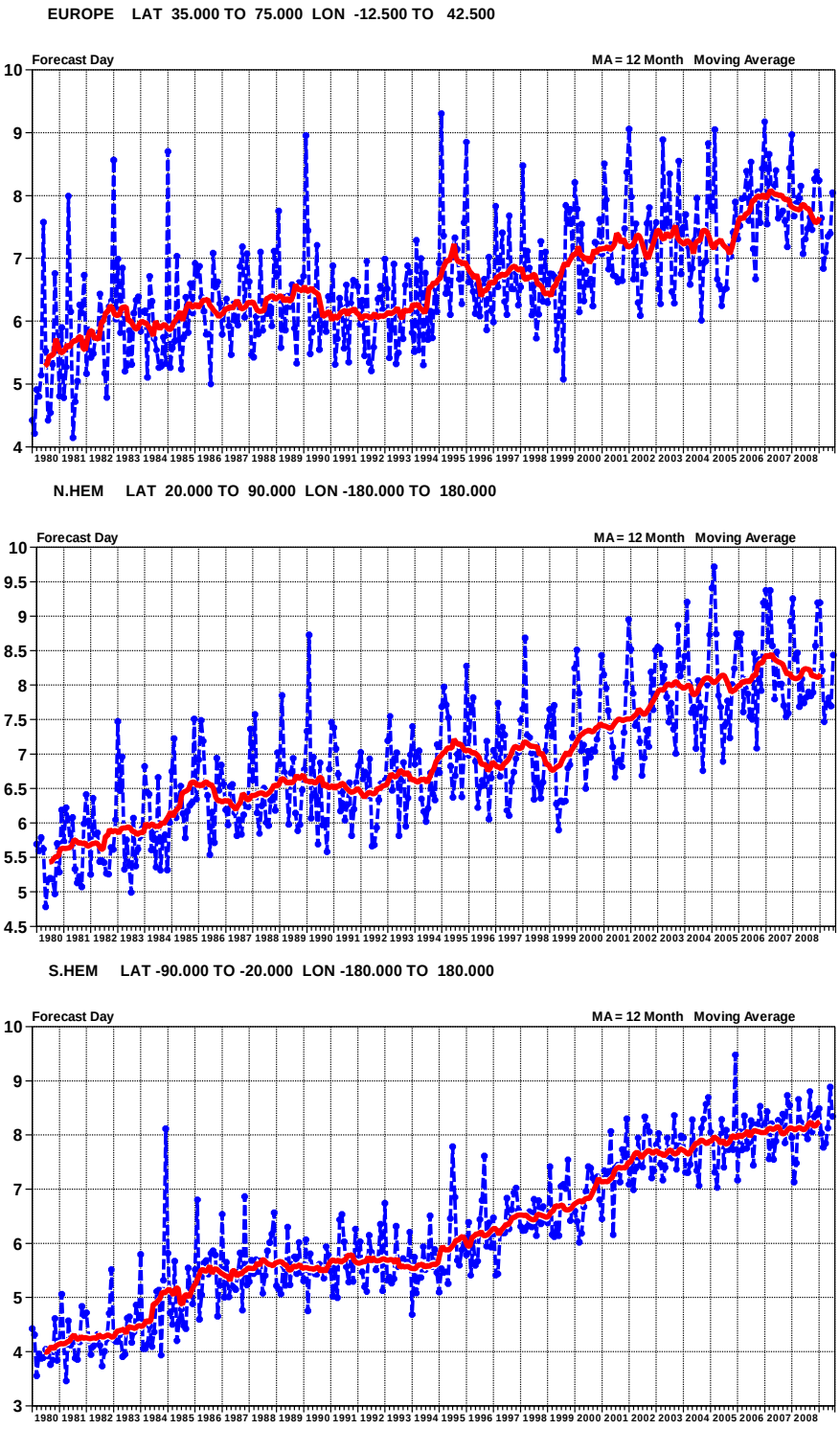


Figure 2: Evolution with time of the 500 hPa height forecast performance – each point on the blue curves is the forecast range at which the monthly average of the forecast anomaly correlation with the verifying analysis falls below 60% for Europe, northern and southern extratropics (the red curve is the 12-month moving average).

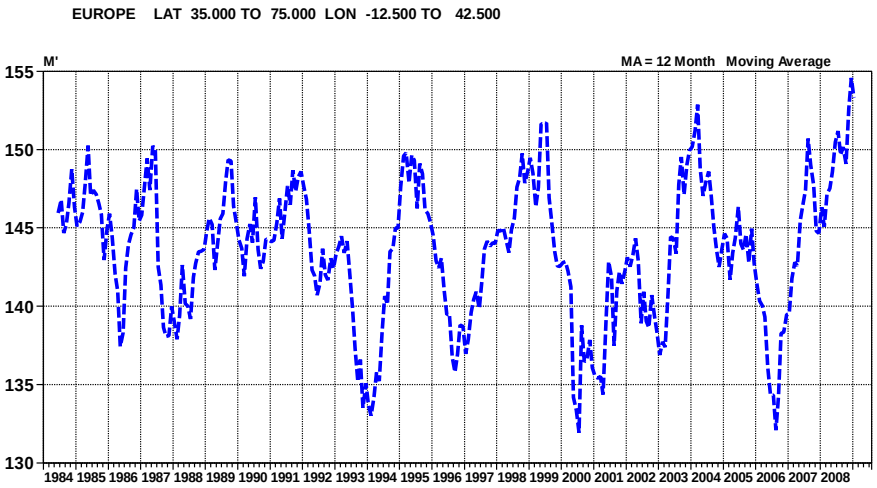


Figure 3: Root Mean Square Error of forecast made by persisting the analysis over 168 h and verifying it as a forecast for 500 hPa geopotential height over Europe. The 12-month moving average is plotted; the last point on the curve is for the 12-month period August 2008 - July 2009.

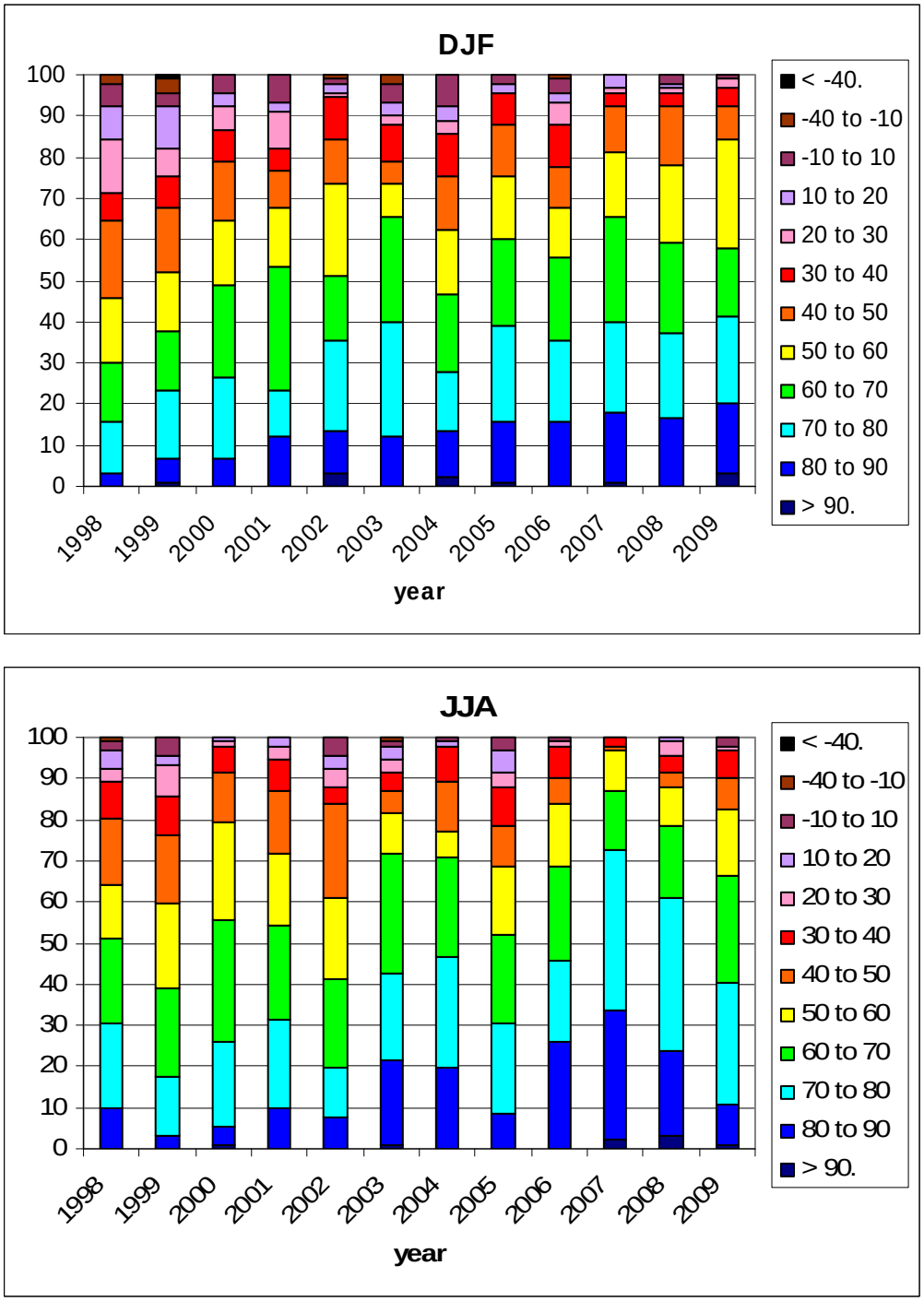


Figure 4: Distribution of Anomaly Correlation of the Day 7 850 hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA, bottom) since 1997-1998.

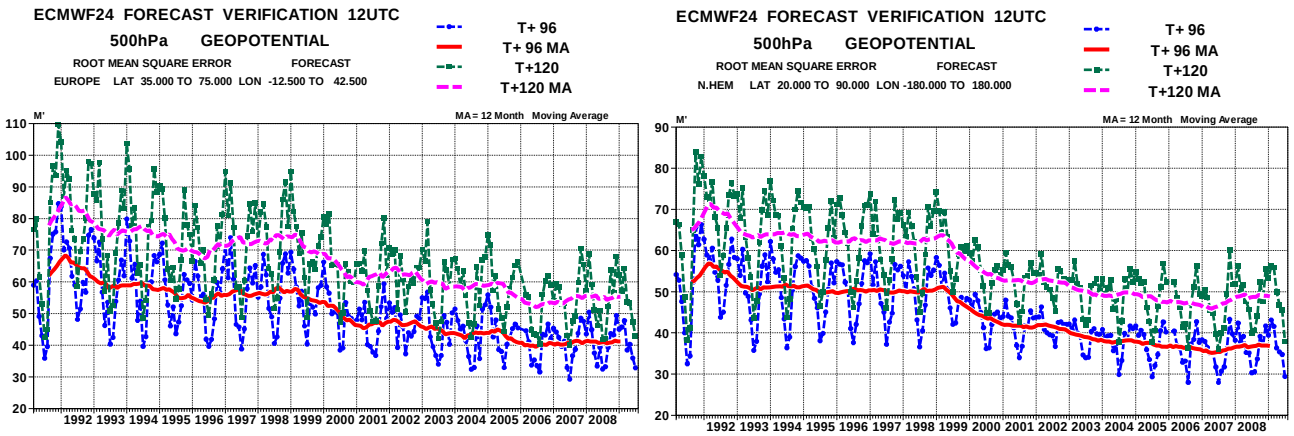


Figure 5: Consistency of the 500 hPa height forecasts over Europe (left panel) and northern extratropics (right panel). Curves show the monthly average RMS difference between forecasts for the same verification time but initialised 24 h apart, for 96-120 h (blue) and 120-144 h (green). 12-month moving average scores are also shown.

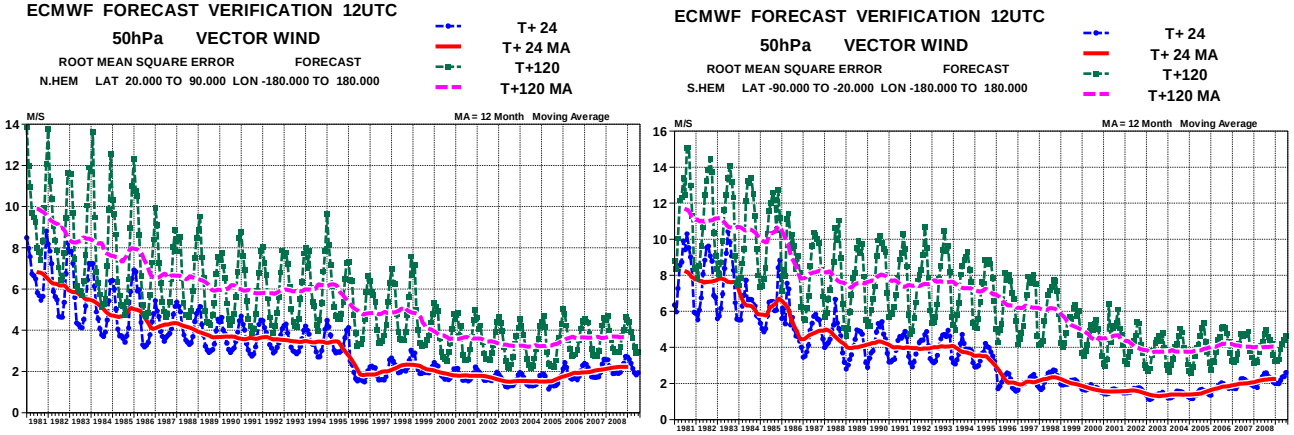


Figure 6: Model scores in the extra-tropical northern (left) and southern (right) hemisphere stratosphere (RMS vector wind error at 50 hPa for 1-day and 5-day forecasts).

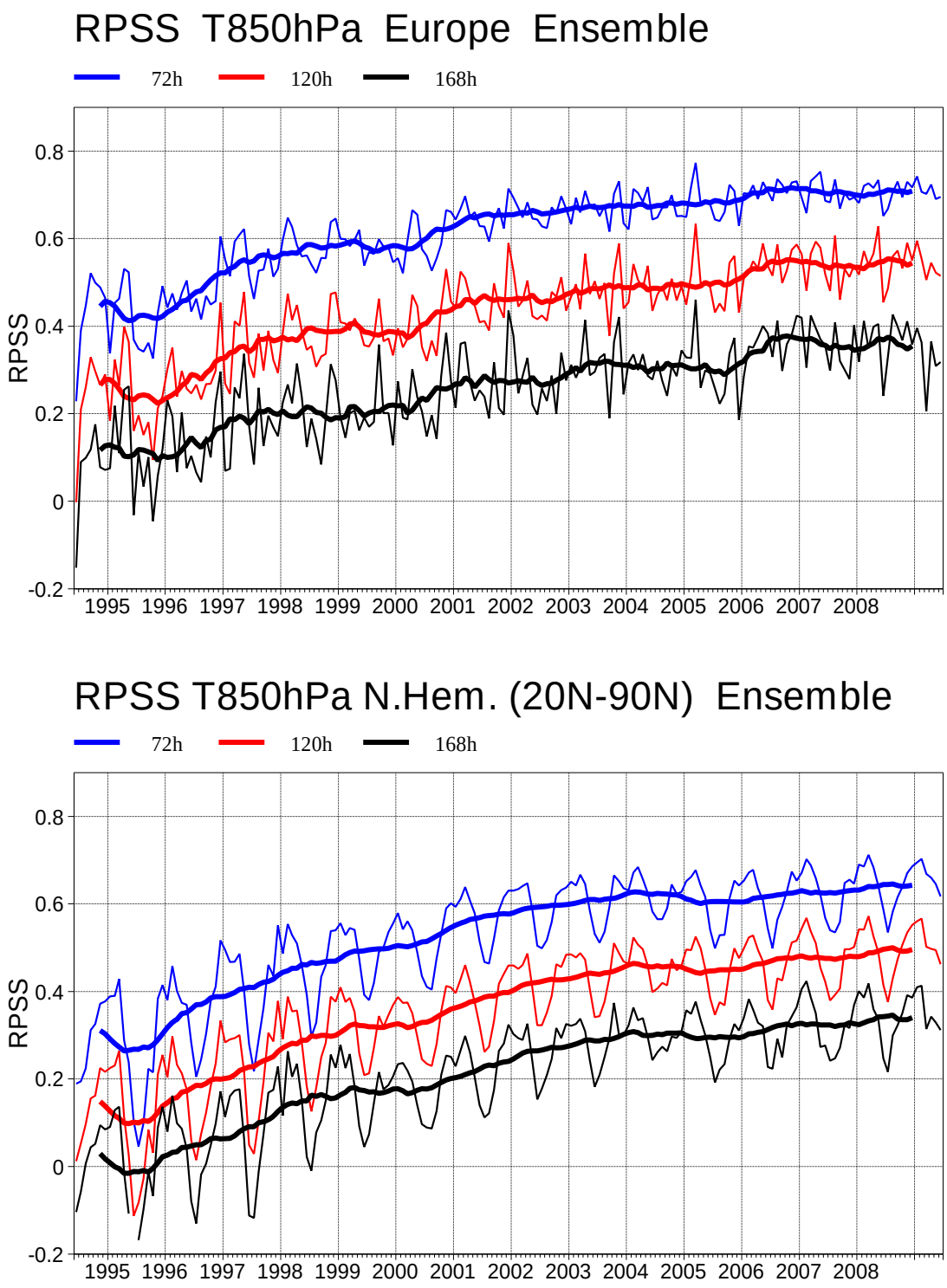


Figure 7: Monthly score and 12-month running mean (bold) of Ranked Probability Skill Score for EPS forecasts of 850 hPa temperature at day 3 (blue), 5 (red) and 7 (black) for Europe (top) and the northern hemisphere extratropics (bottom).

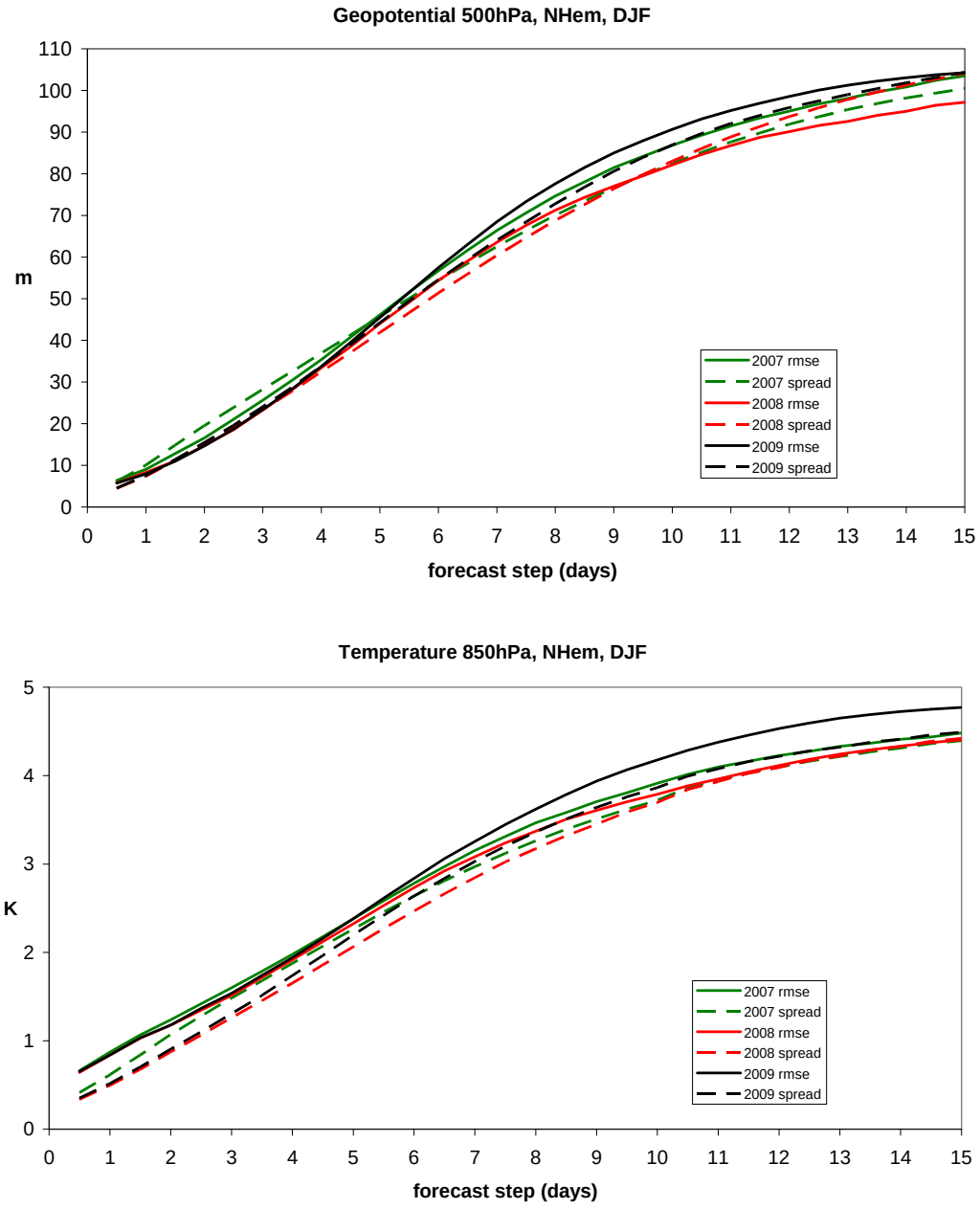


Figure 8: Ensemble spread (standard deviation, dashed lines) and root mean square error of ensemble-mean (solid lines) for 500 hPa geopotential (top) and 850 hPa temperature (bottom) for winter 2008-09 (black), 2007-08 (red) and 2006-07 (green) over the extra-tropical northern hemisphere.

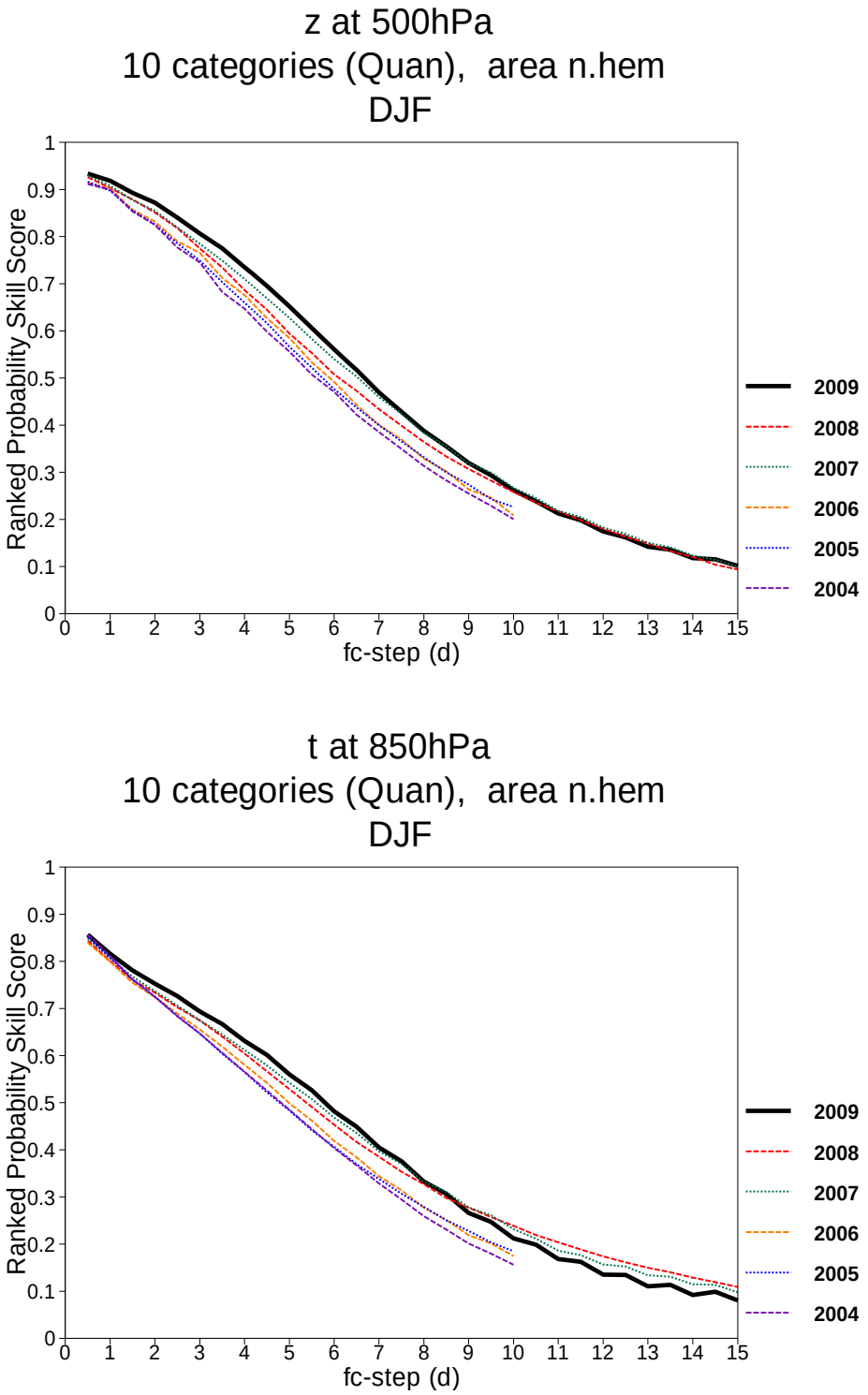
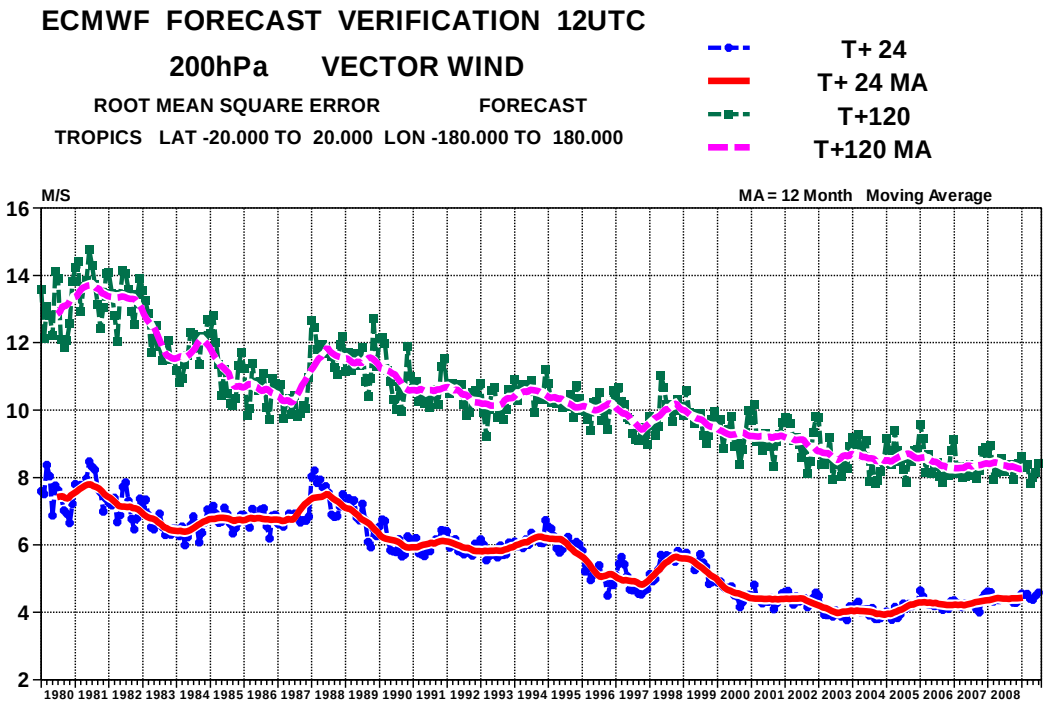


Figure 9: Ranked probability skill score for 500 hPa height (top) and 850 hPa temperature (bottom) EPS forecasts for winter (December-February) over the extra-tropical northern hemisphere. Skill from the EPS day 1-15 forecasts is shown for winter 2008-09 (black), 2007-08 (red) and 2006-07 (green); the EPS only ran to 10 days in previous years.



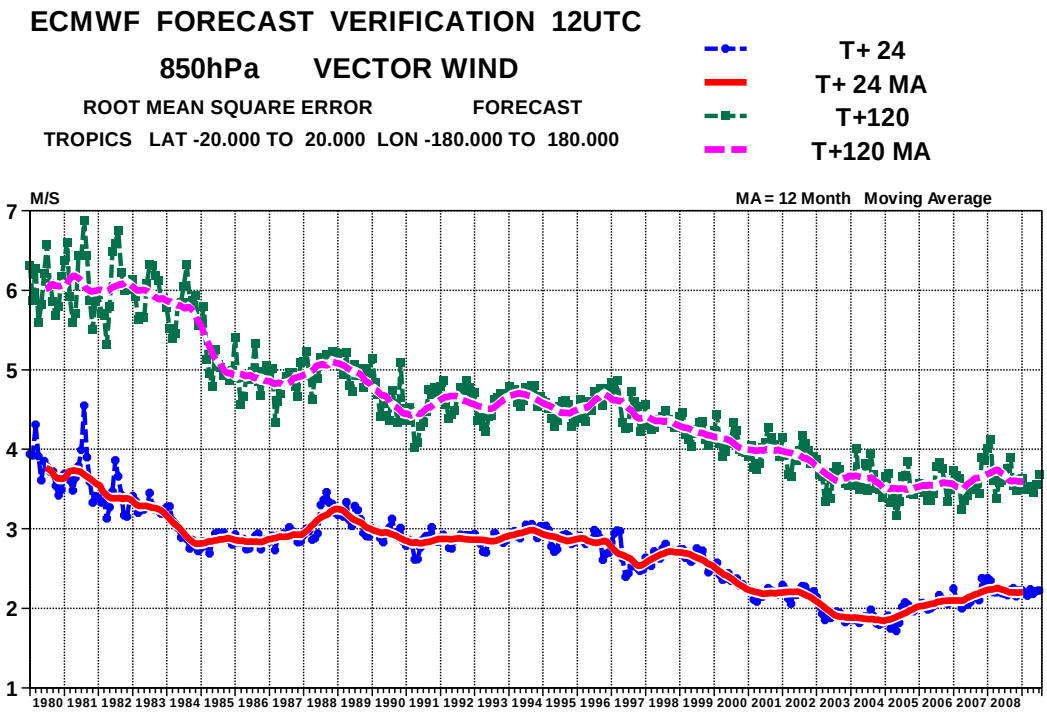


Figure 10: Model scores in the tropics (root mean square vector wind errors at 200 hPa and 850 hPa for 1-day and 5-day forecasts). Monthly mean and 12-month running mean.

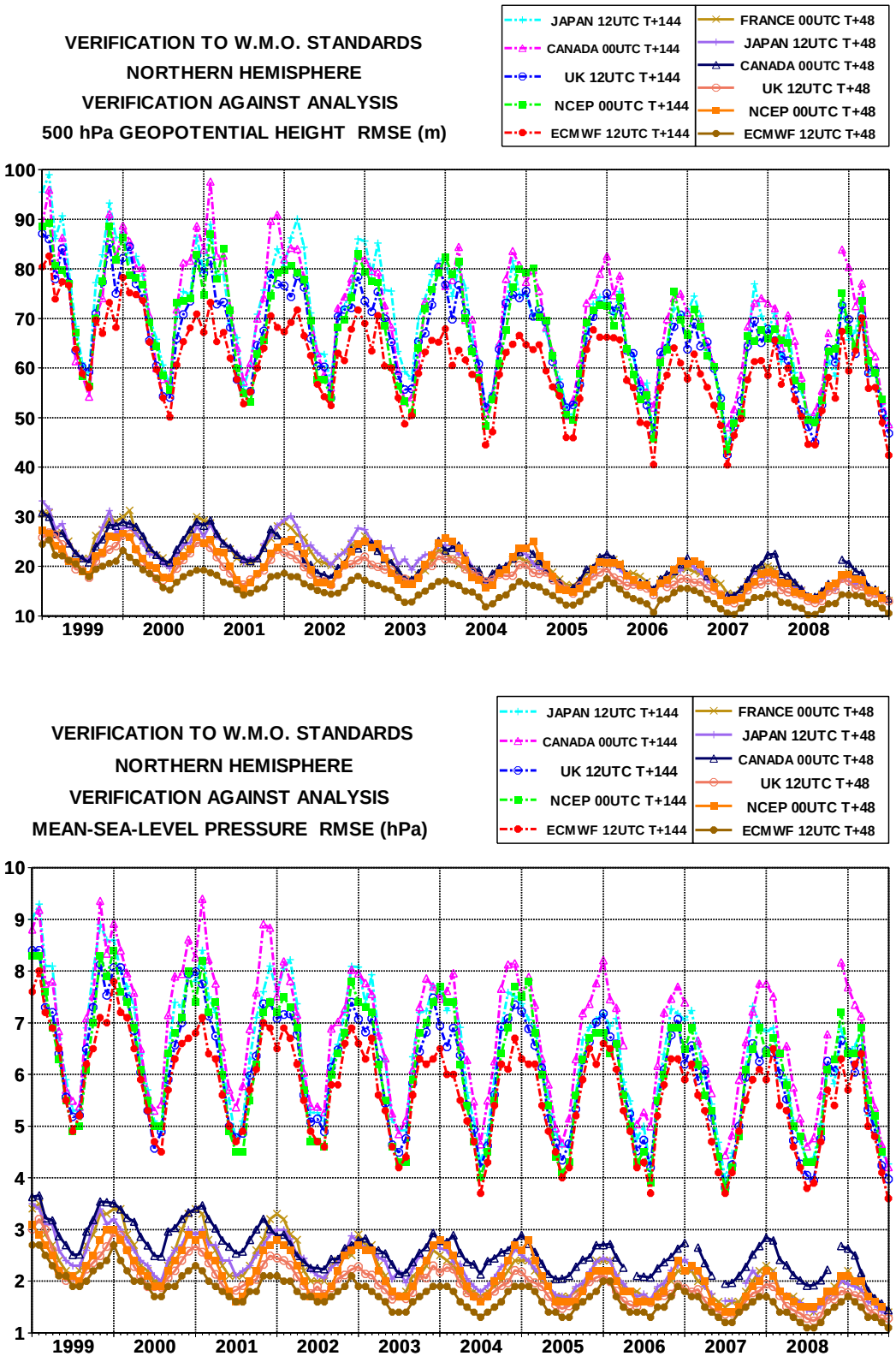
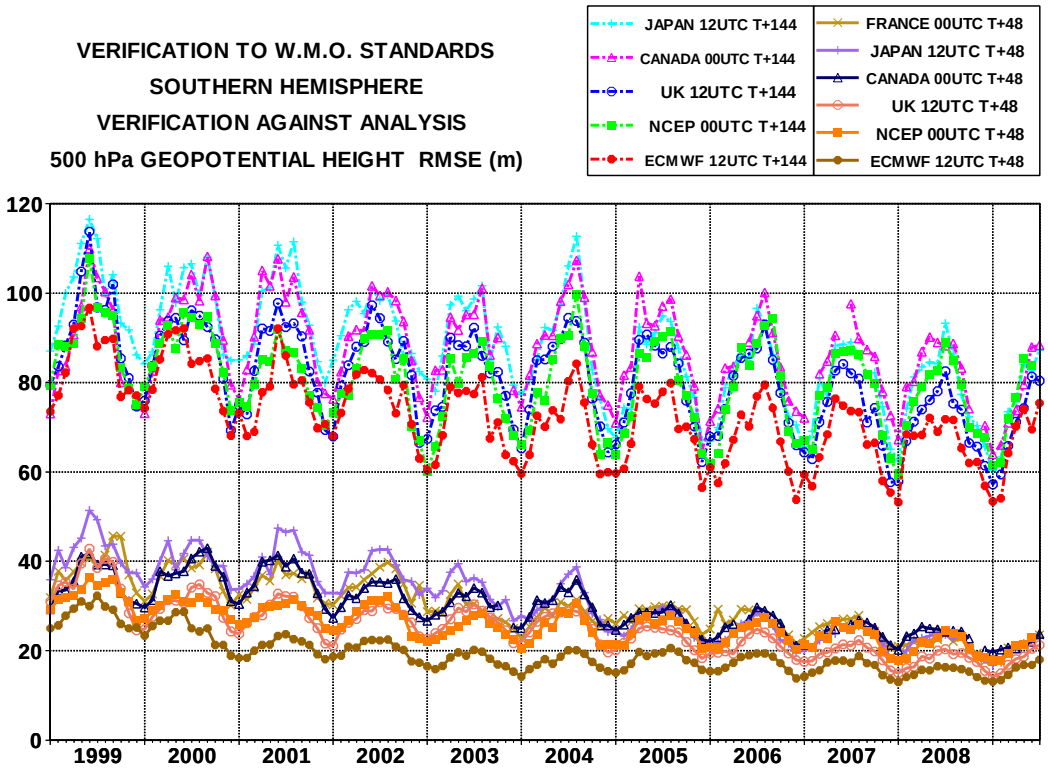


Figure 11: WMO/CBS exchanged scores (RMS error over northern extratropics, 500 hPa geopotential height and MSLP for 2-day and 6-day forecasts).



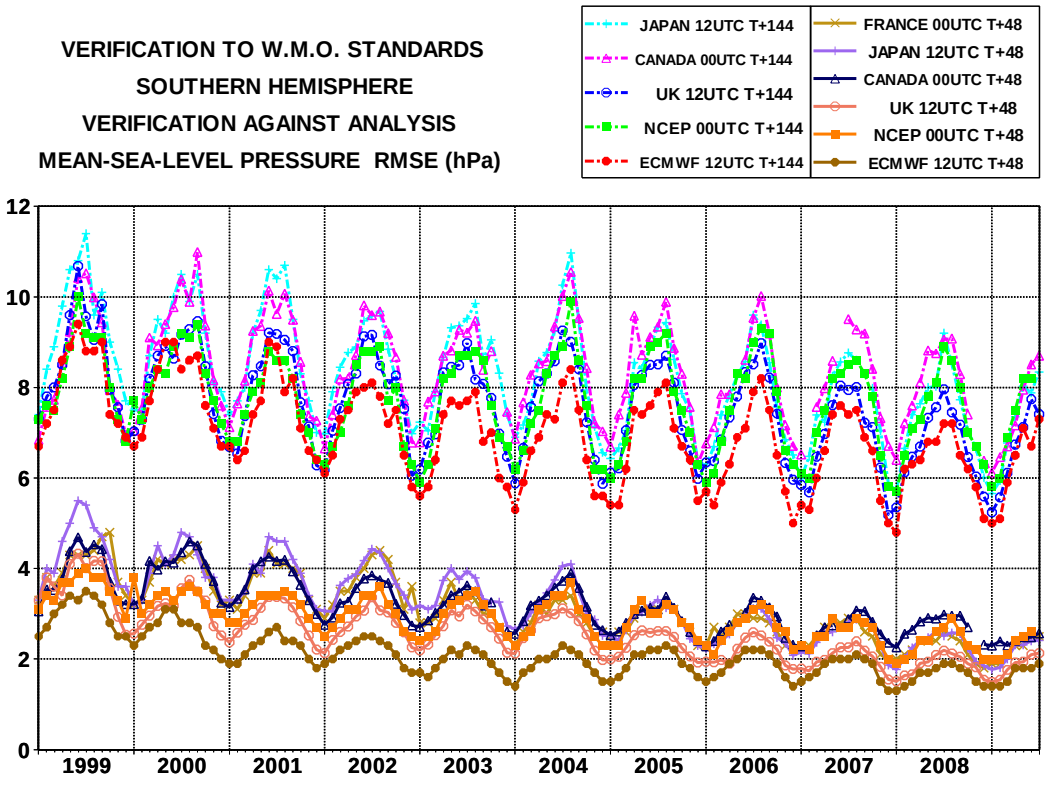
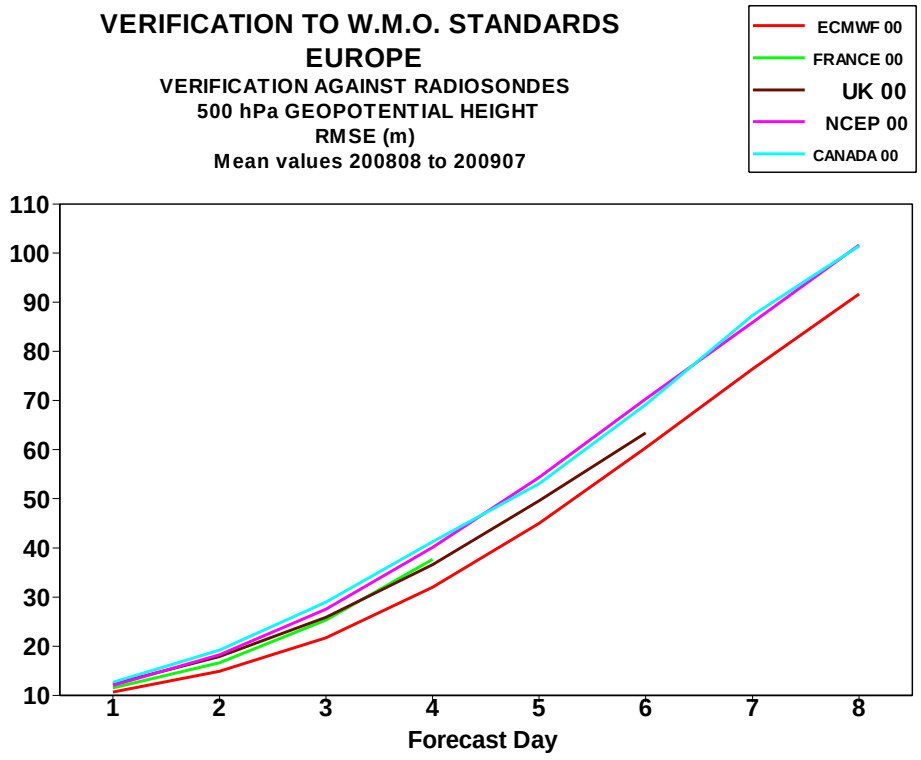


Figure 12: WMO/CBS exchanged scores (RMS error over southern extratropics, 500 hPa geopotential height and MSLP for 2-day and 6-day forecasts).



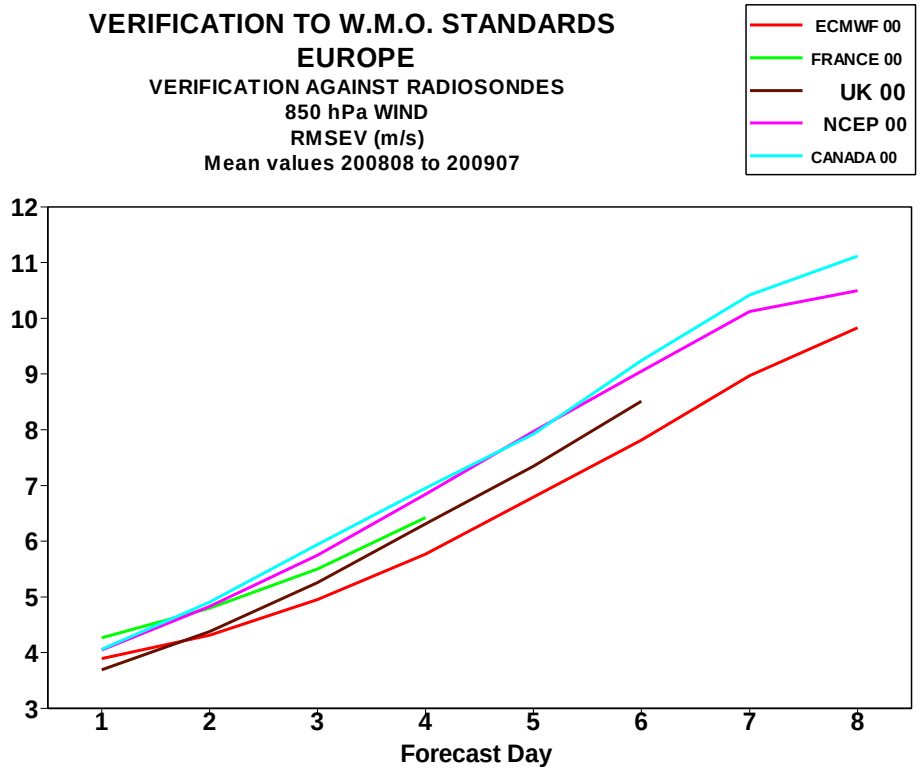


Figure 13: WMO/CBS exchanged scores using radiosondes: 500 hPa height and 850 hPa wind RMS error over Europe (annual mean).

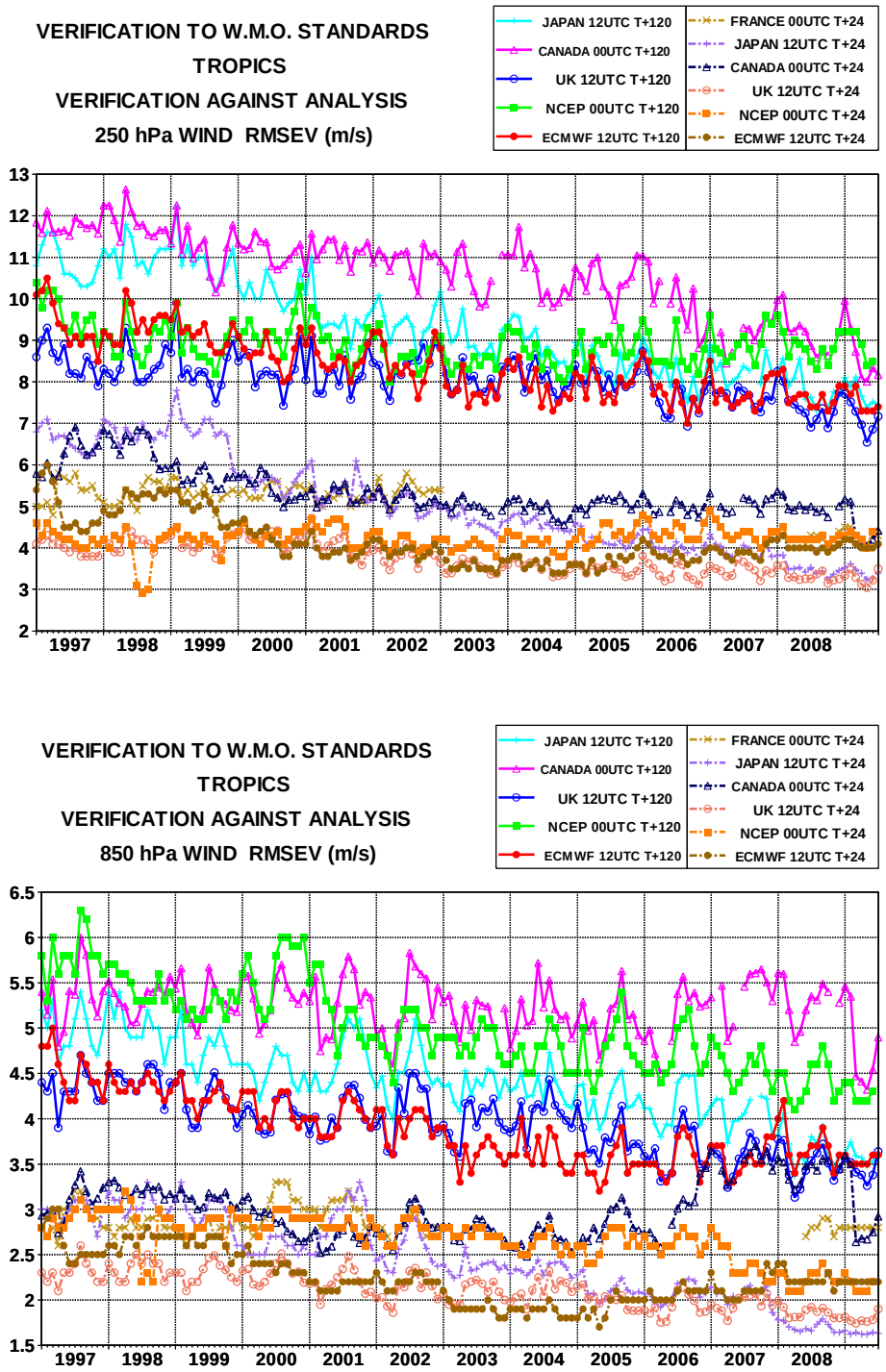


Figure 14: WMO/CBS exchanged scores (RMS vector error over the tropics, 250 hPa and 850 hPa wind forecast for day 1 and day 5).

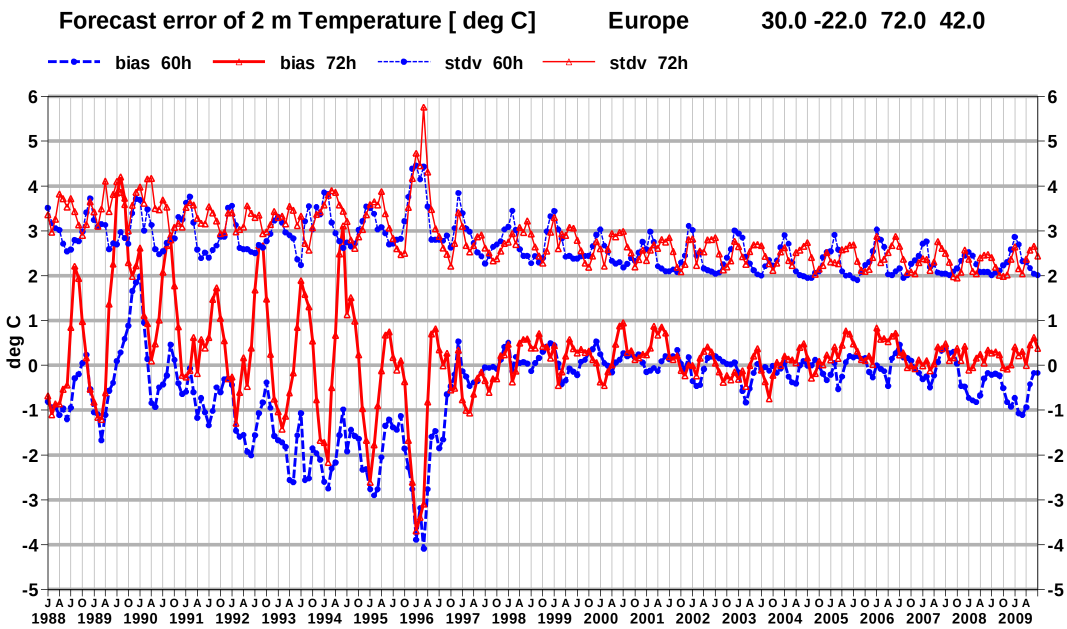


Figure 15: Verification of 2 metre temperature forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves are bias, upper curves are standard deviation of error.

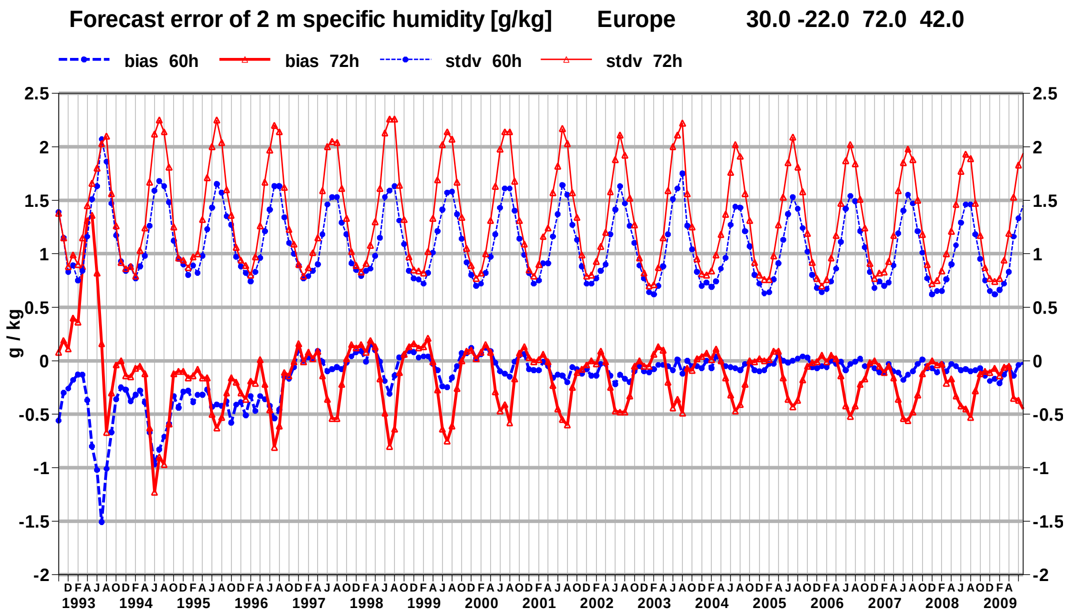


Figure 16: Verification of 2 metre specific humidity forecasts against European SYNOP data on the GTS for 60-hour (night-time) and 72-hour (daytime) forecasts. Lower pair of curves is bias, upper curves are standard deviation of error.

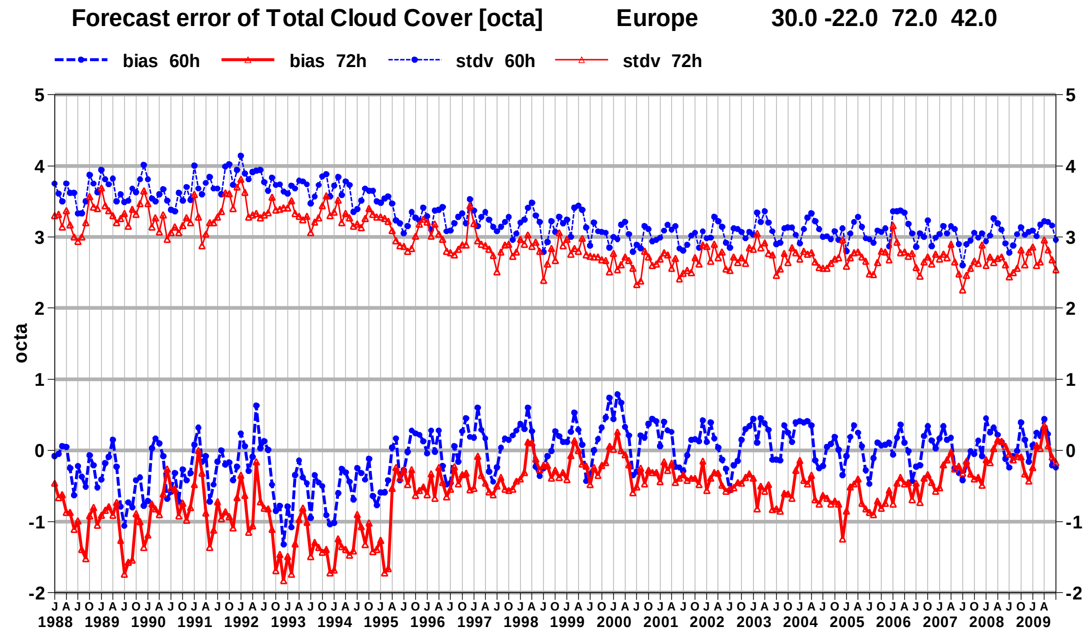


Figure 17: Verification of total cloud cover forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves is bias, upper curves are standard deviation of error.

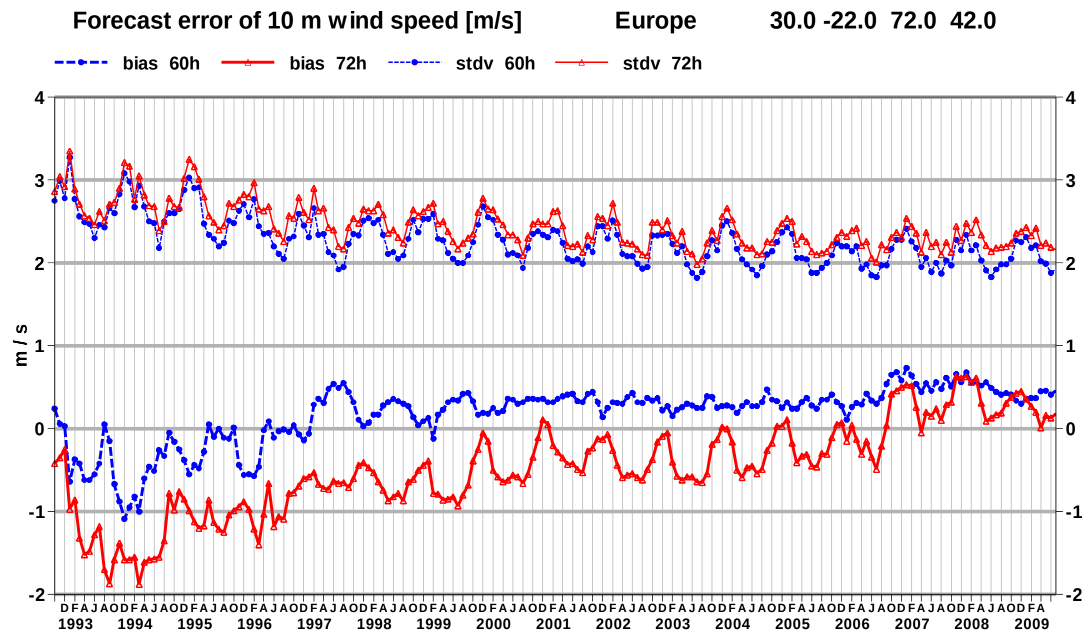
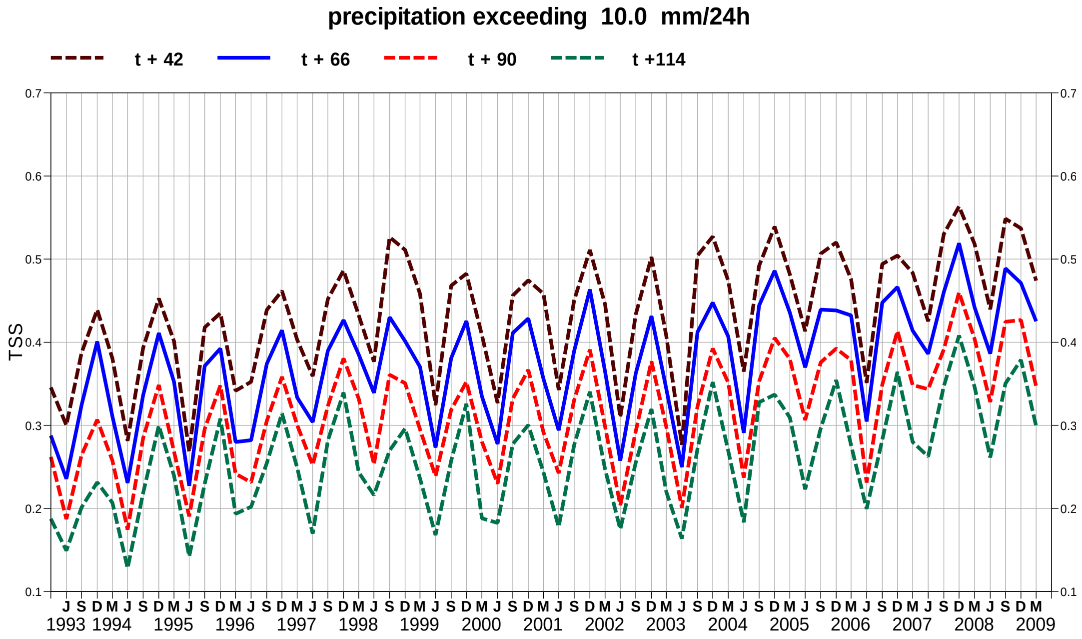


Figure 18: Verification of 10 metre wind speed forecasts against European SYNOP data on the GTS for 60 hour (night-time) and 72 hour (daytime) forecasts. Lower pair of curves is bias, upper curves are standard deviation of error.



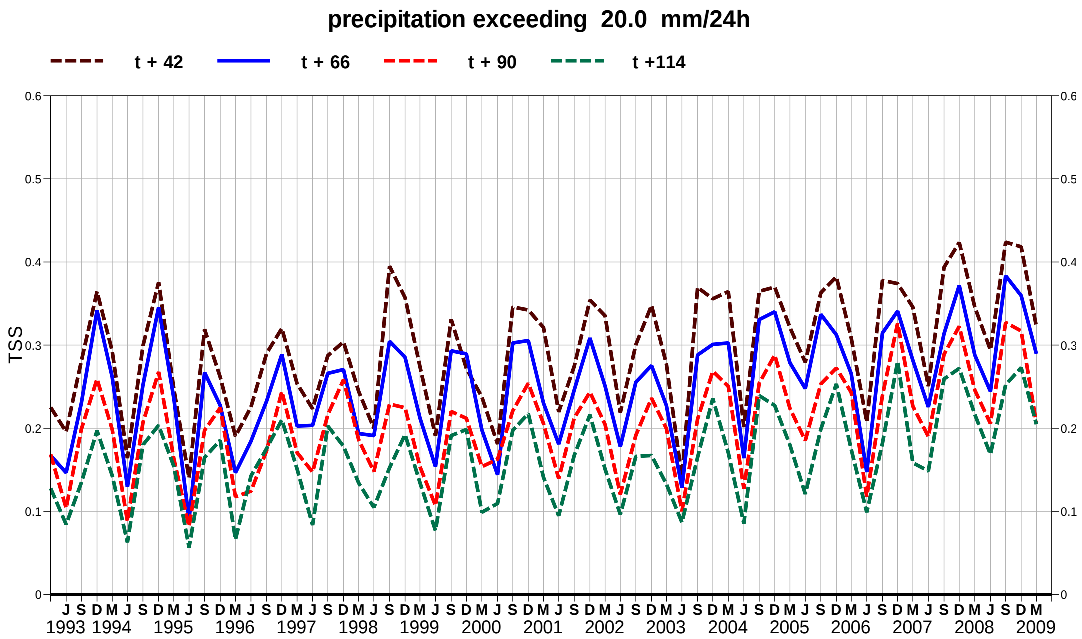


Figure 19: TSS time series for precipitation forecasts exceeding 10 mm/day (top) and 20 mm/day (bottom) verified against SYNOP data on the GTS for Europe. Curves are shown for the 24 hour accumulations up to 42, 66, 90, and 114 hours (from the forecasts starting at 12 UTC). 3 month mean scores (last point is March-May 2009).

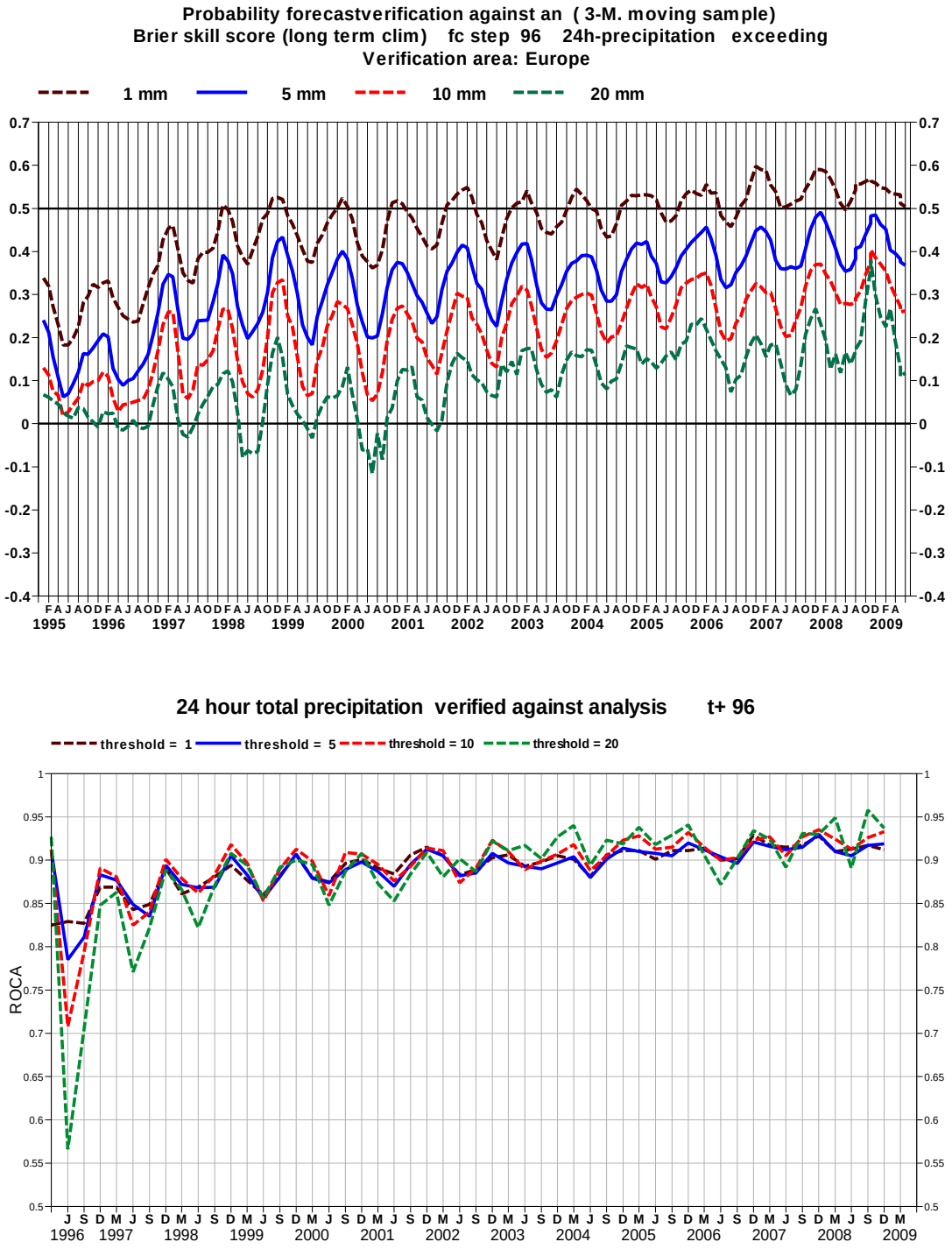


Figure 20: Time series of Brier Skill Score (top) and Relative Operating Characteristic Area (ROCA) for EPS probability forecasts of precipitation over Europe exceeding thresholds of 1, 5, 10 and 20 mm/day at day 4. The skill score is calculated for three-month running periods.

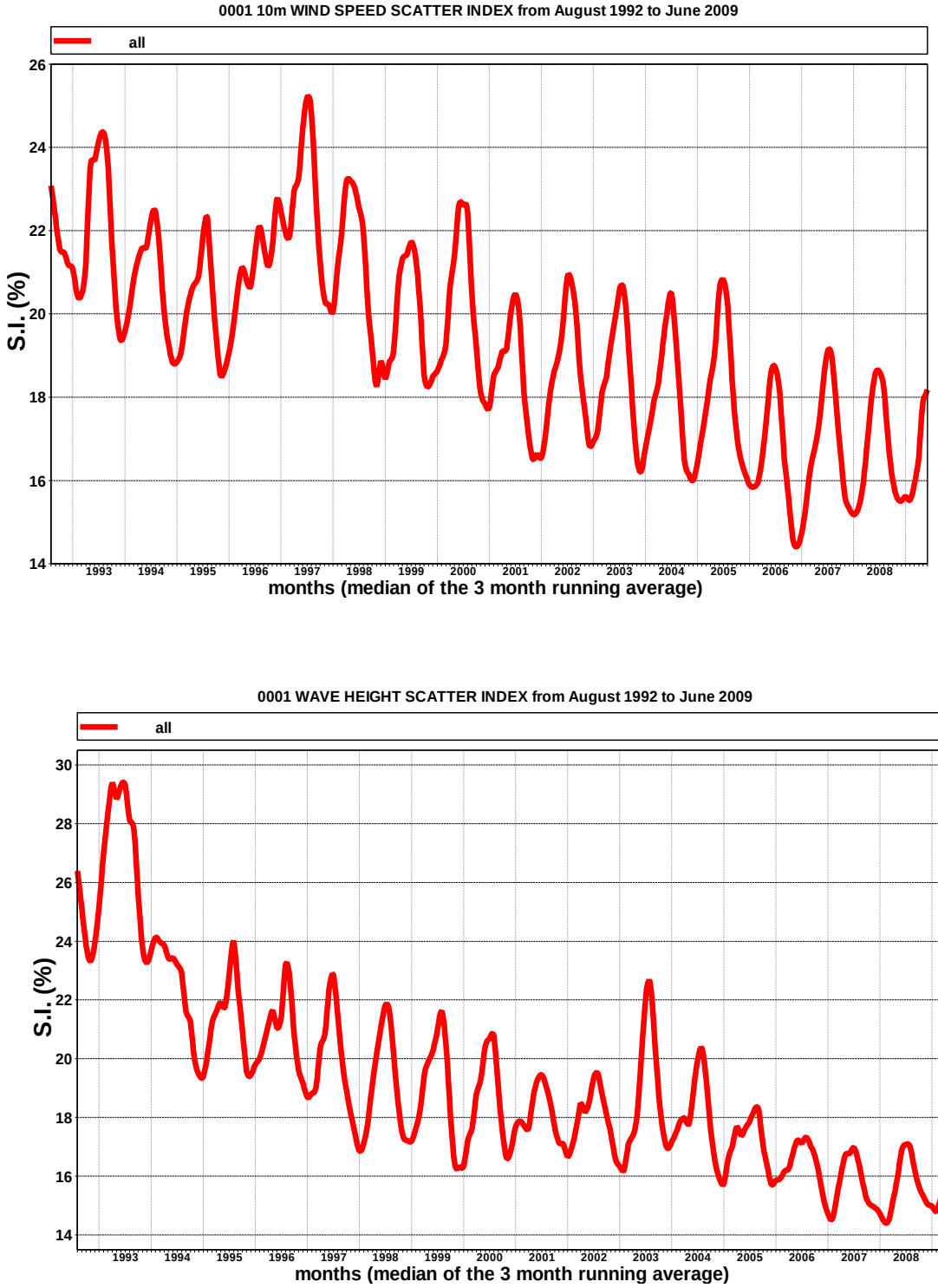
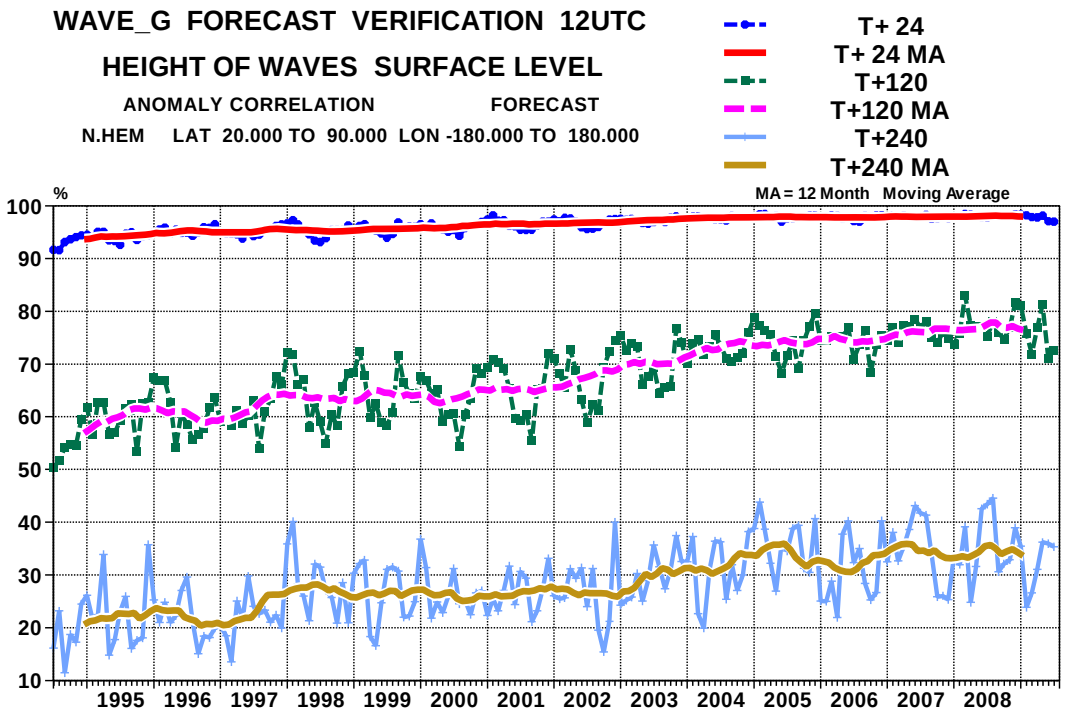


Figure 21: Time series of verification of the ECMWF 10 metre wind analysis and wave model analysis (wave height) verified against northern hemisphere buoy observations. The scatter index is the error standard deviation normalised by the mean observed value; a three-month running mean is used.



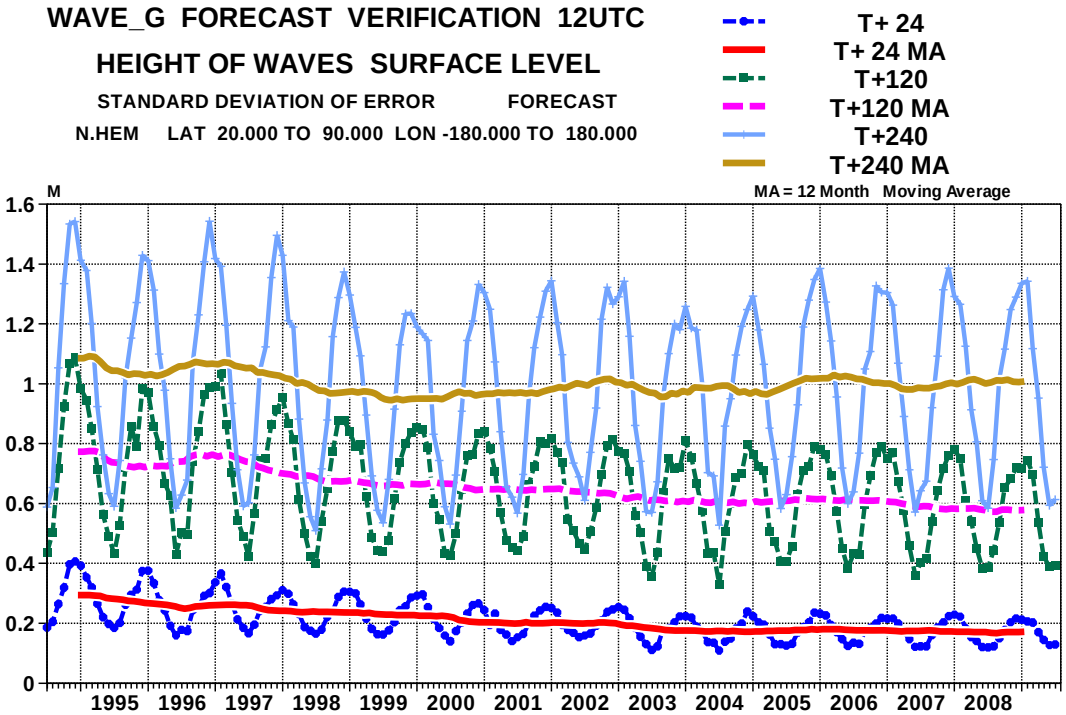
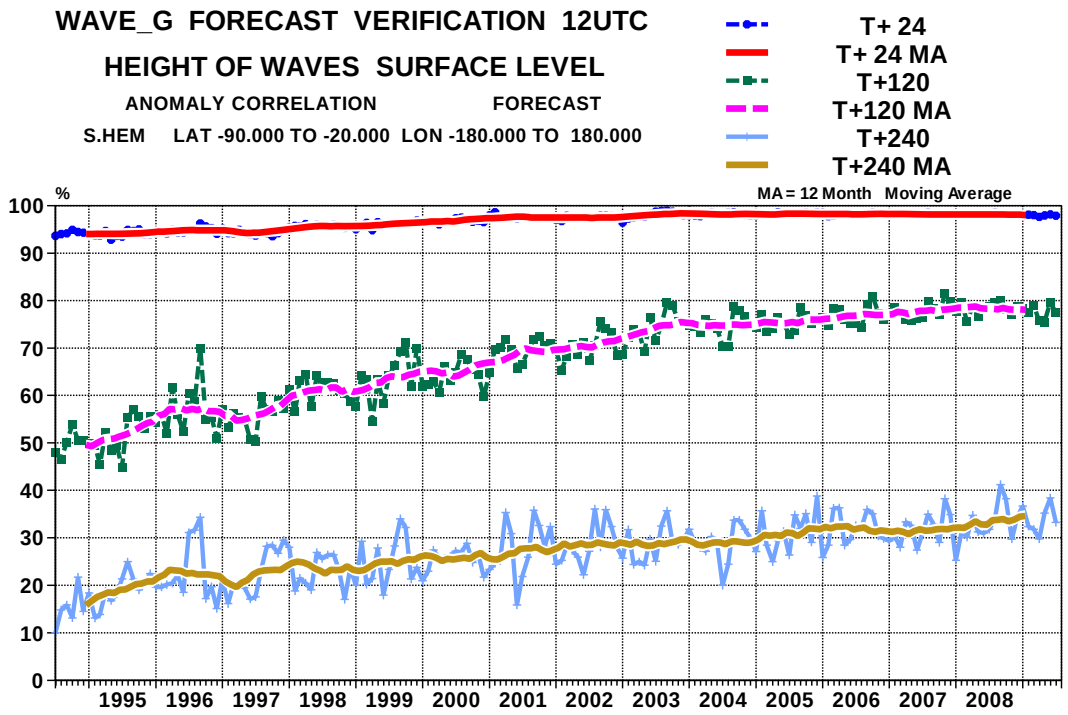


Figure 22: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (northern extratropics).



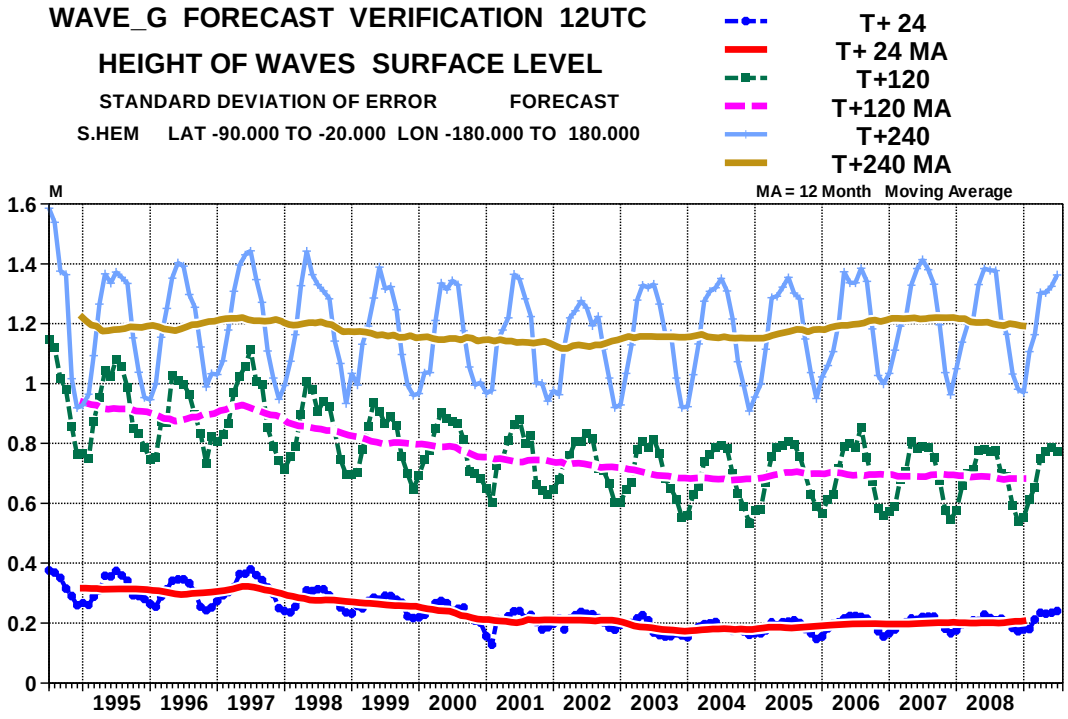


Figure 23: Scores (anomaly correlation and error standard deviation) of ocean wave heights verified against the analysis (southern extratropics).

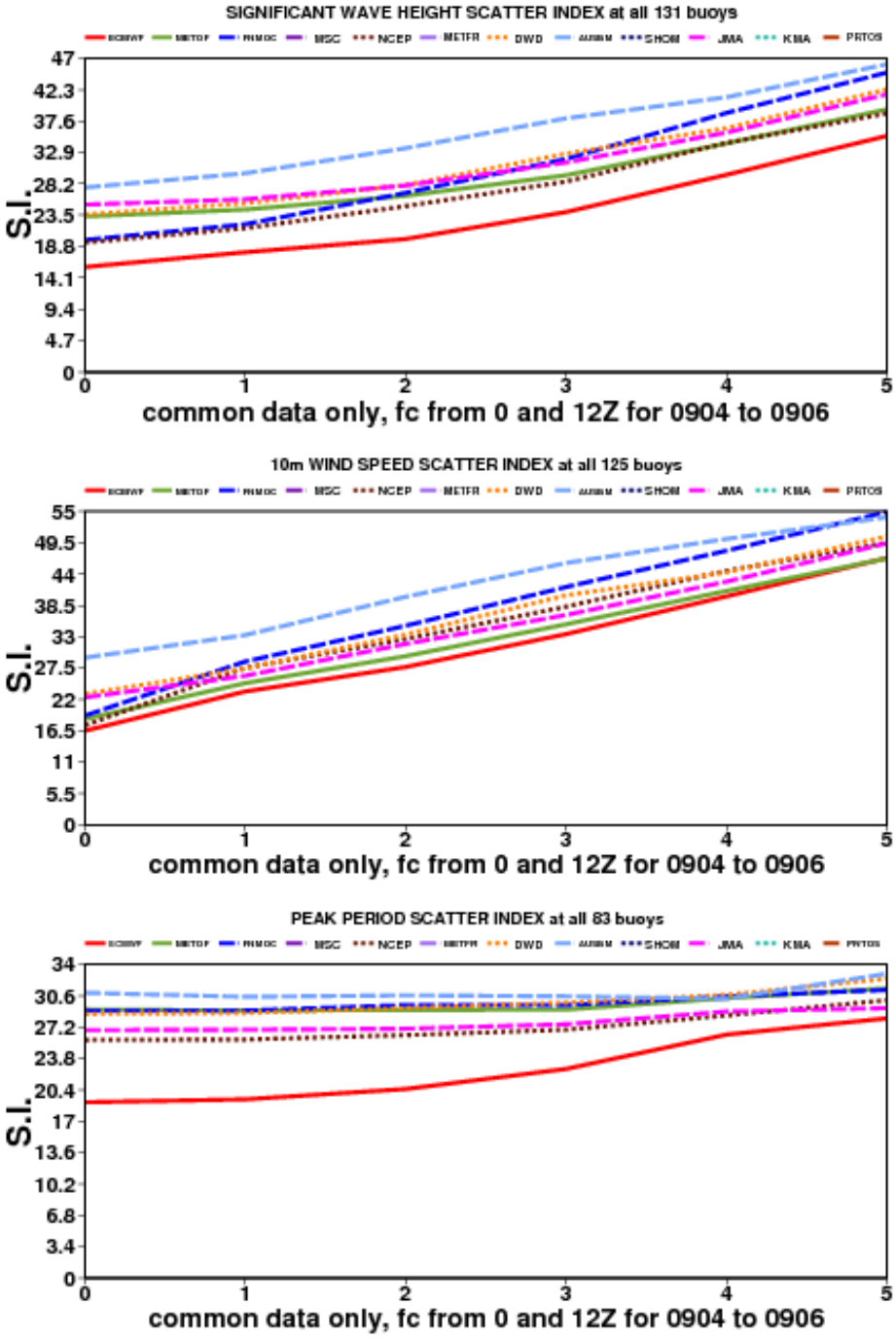
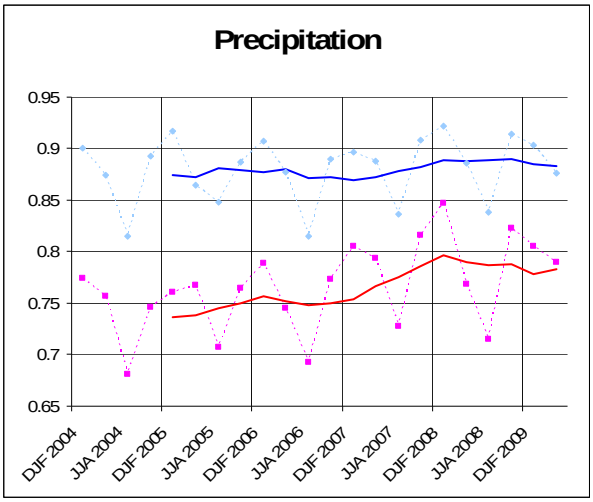
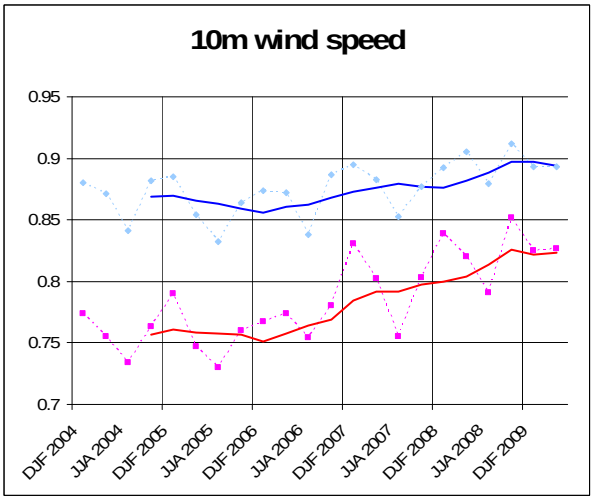


Figure 24: Verification of different model forecasts of wave height, 10 m wind speed and peak wave period using a consistent set of observations from wave buoys. The scatter index (SI) is the standard deviation of error normalised by the mean observed value; plots show the SI for the 3-month period April-June 2009. The x-axis shows the forecast range in days from analysis (step 0) to day 5. METOF: the Met Office, UK; FNMOC: Fleet Numerical Meteorology and Oceanography Centre, USA; MSC: Meteorological Service of Canada; NCEP: National Centers for Environmental Prediction, USA; METFR: Météo France; DWD: Deutscher Wetterdienst, AUSBM: Bureau of Meteorology, Australia; SHOM: Service Hydrographique et Océanographique de la Marine, France; JMA: Japan Meteorological Agency; KMA: Korea Meteorological Administration; PRTOS: Puertos del Estado, Spain.





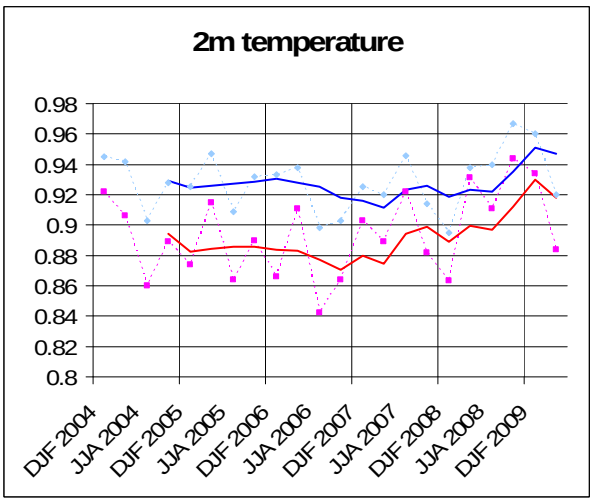


Figure 25: Verification of Extreme Forecast Index (EFI) for precipitation, 10 m wind speed and 2 m temperature over Europe. Extreme event is taken as an observation exceeding 95th percentile of station climate. Hit rates and false alarm rates are calculated for EFI exceeding different thresholds. Curves show the ROC area calculated for each 3-month season from winter (December-February, DJF) 2004 - 2005 to spring (March-May) 2009 for day 2 (light blue dashed) and day 5 (magenta dashed). Solid lines show running mean of seasonal scores averaged over 4 seasons for: day 2 (blue) and day 5 (red); last point is for average from summer (JJA) 2008 to spring 2009.

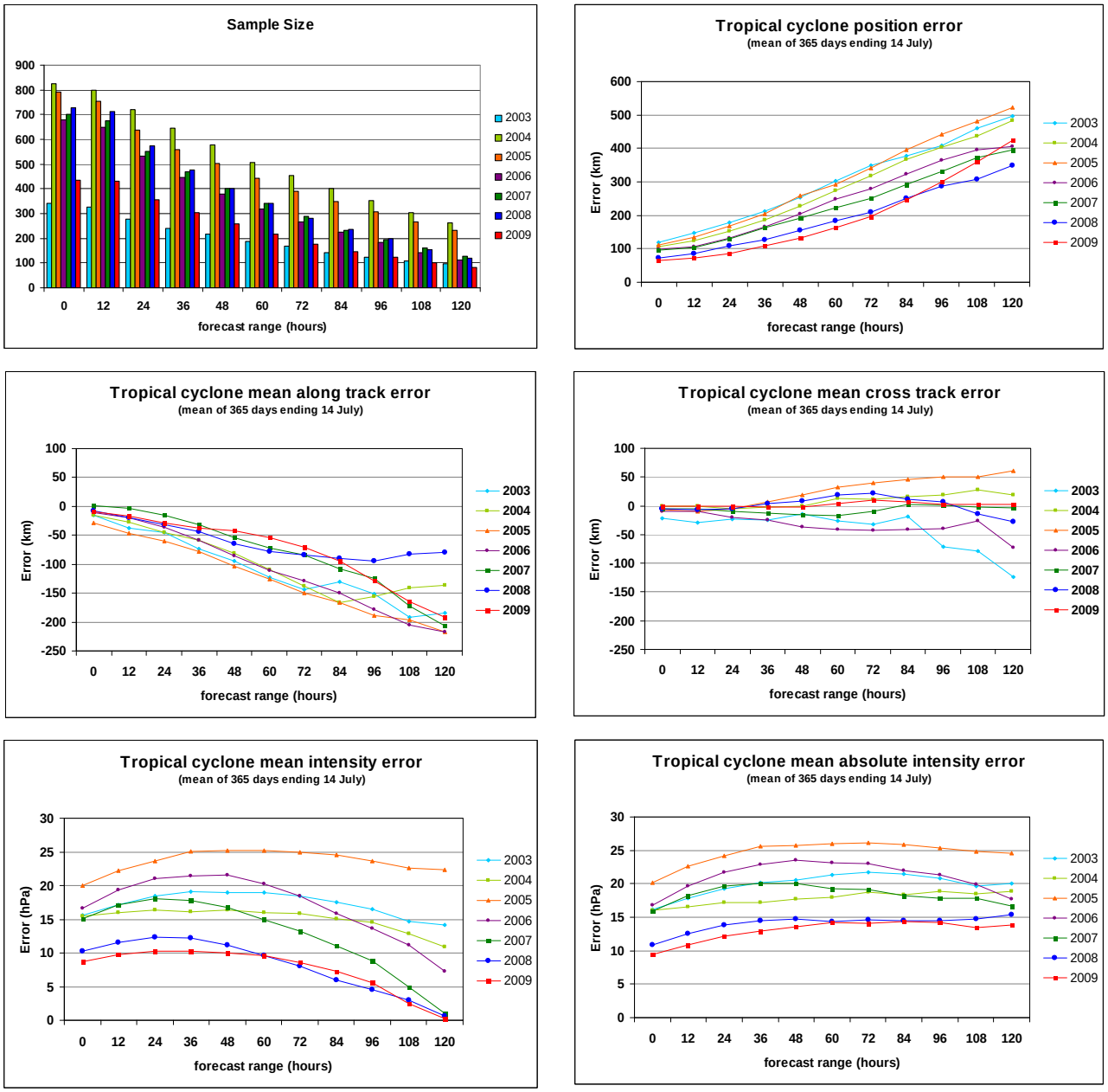


Figure 26: Verification of tropical cyclone predictions from the operational deterministic forecast. Results are shown for 12-month periods ending on 14 July. The latest period, 15 July 2008 to 14 July 2009, is shown in red. Verification is against the observed position reported in real-time via the GTS. The top right panel shows the mean position error (average over all cases of the distance between forecast and observed position; always positive). The middle panels show the mean error (bias) in the forecast cyclone position in the direction of travel of the cyclone (along track error; negative values indicate slow bias; left panel) and the mean error (bias) in the forecast cyclone position at right-angles to the direction of travel (cross track error; right panel). The bottom left panel shows the 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). The bottom right panel shows the mean absolute error of the intensity. The sample size at each forecast step for each year is shown in the top left panel: there are substantially fewer events at later forecast steps than earlier in the forecast and hence there will be greater uncertainty in the scores at the later ranges.

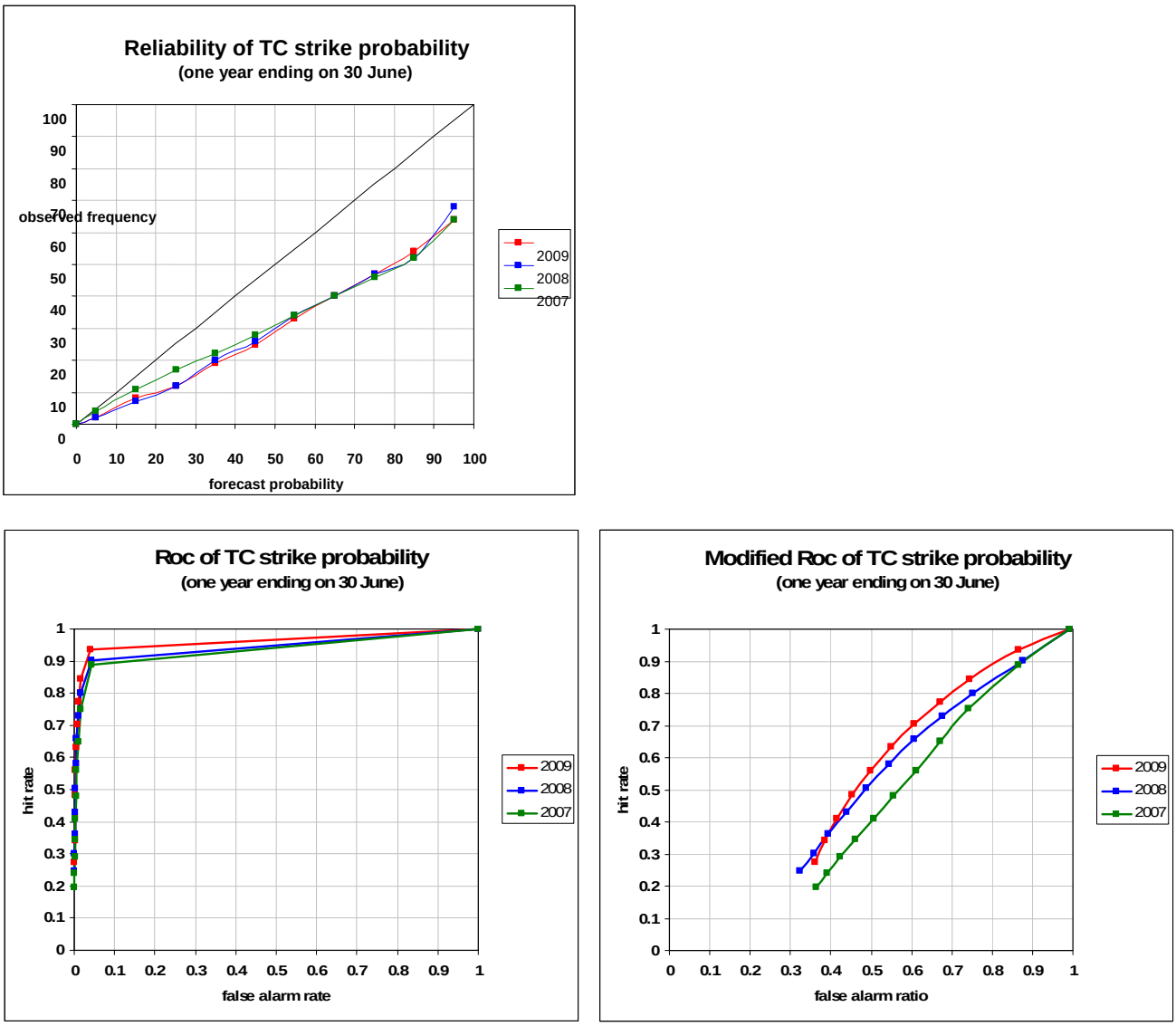


Figure 27: Probabilistic verification of EPS tropical cyclone forecasts for three 12-month periods: July 2006 - June 2007 (green), July 2007 - June 2008 (blue) and July 2008 - June 2009 (red). Upper panel shows reliability diagram (the closer to the diagonal, the better). The lower panel shows (left) the ROC diagram (the closer to the upper left corner, the better) and the modified ROC, where the false alarm ratio is used instead of the false alarm rate in the standard ROC.

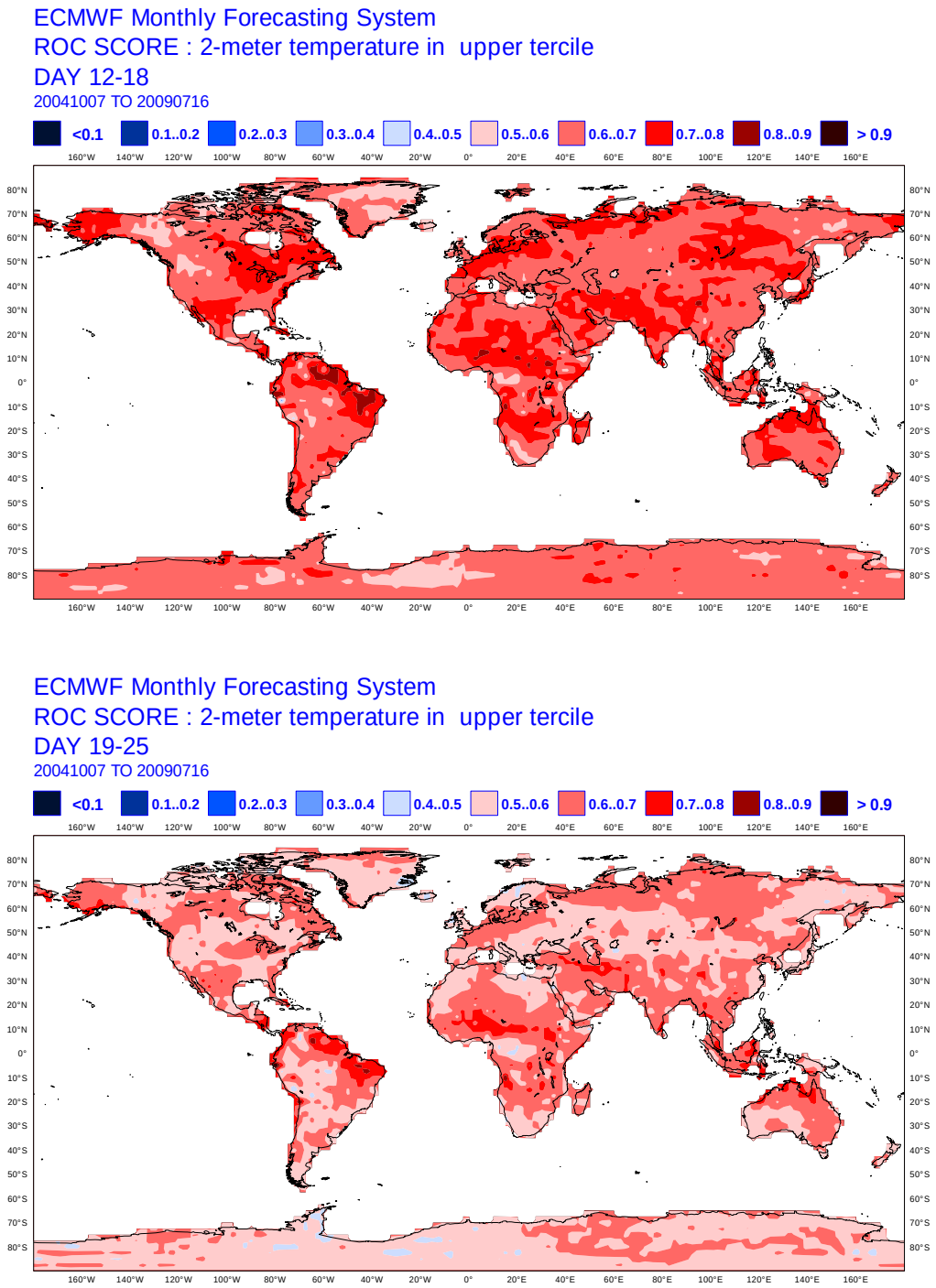


Figure 28: Spatial distribution of ROC area scores for the probability of 2 m temperature anomalies being in the upper third of the climatological distribution. The sample comprises all forecasts issued between 7 October 2004 and 16 July 2009 for two 7-day forecast ranges: days 12-18 (top) and days 19-25 (bottom). Red shading indicates positive skill compared to climate.

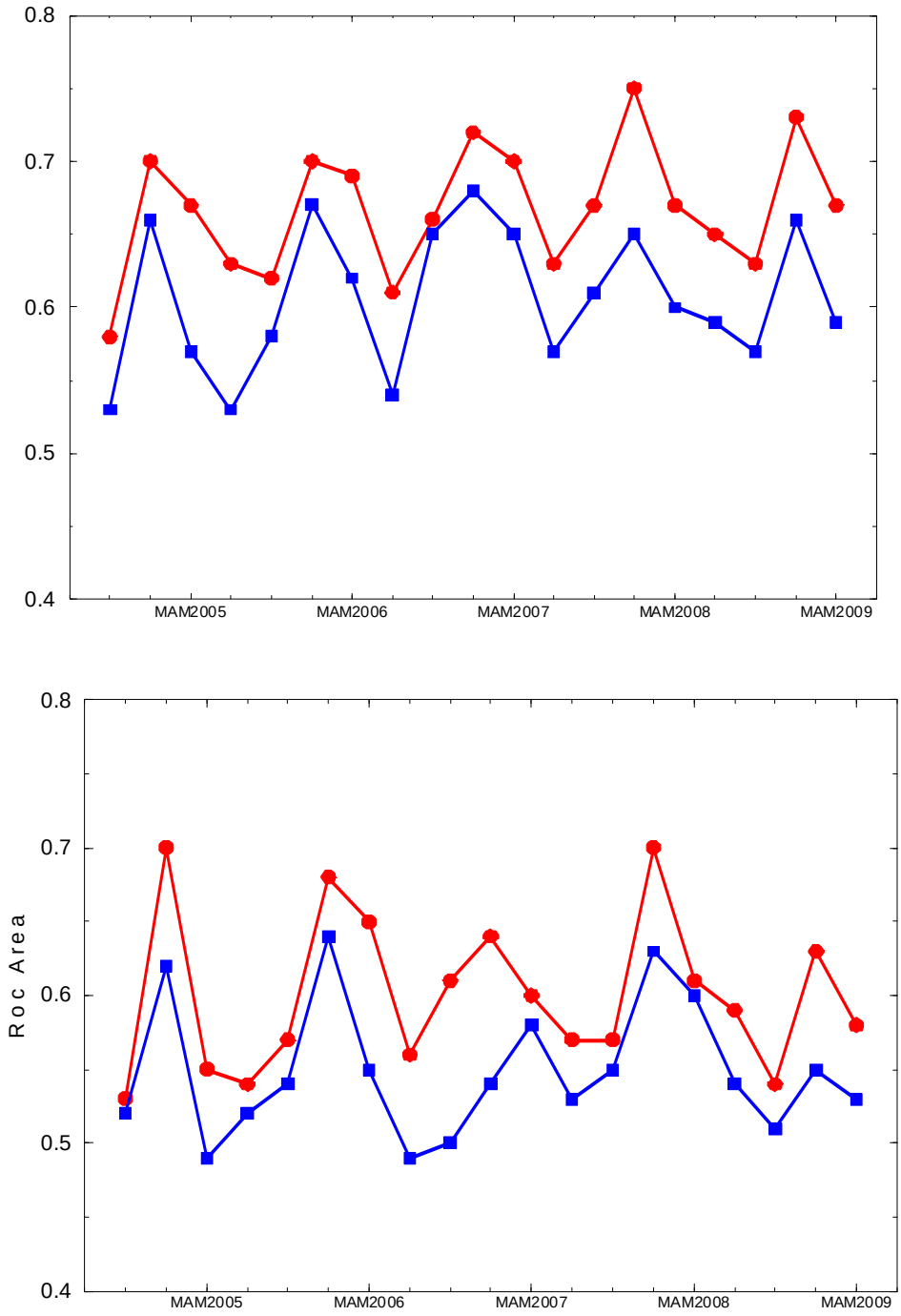


Figure 29: Area under ROC for the probability that 2 metre temperature is in the upper third of the climate distribution. Scores are calculated for each 3 month season since autumn (September-November) 2004 for all land points in the extra-tropical northern hemisphere. The red line shows the score of the operational monthly forecasting system for forecast days 12-18 (7-day mean) (top panel) and 19-32 (14-day mean) (bottom panel). As a comparison, the blue line shows the score using persistence of the preceding 7-day or 14-day period of the forecast.

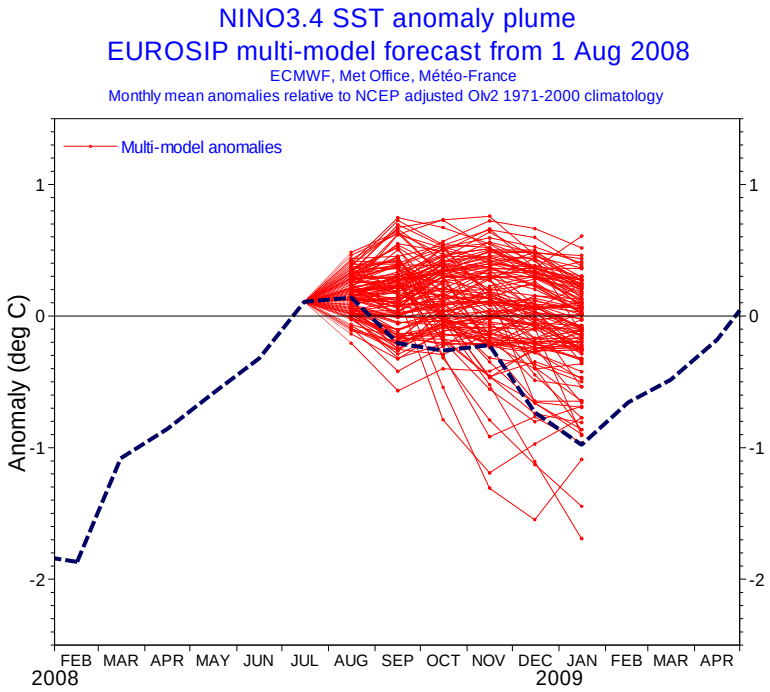


Figure 30: Plot of EUROSIP forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from August 2008. The red lines represent the 40 ensemble members; dashed blue lines show the subsequent verification.

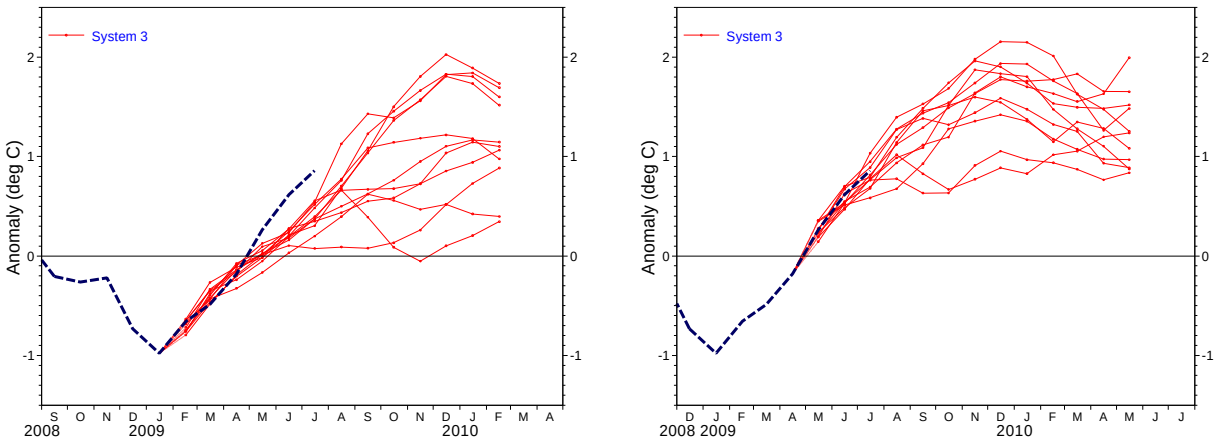


Figure 31: Plot of ECMWF forecasts of SST anomalies over the NINO 3.4 region of the tropical Pacific from February (left) and May (right) 2009. The red lines represent the 40 ensemble members; dashed blue lines show the subsequent verification.

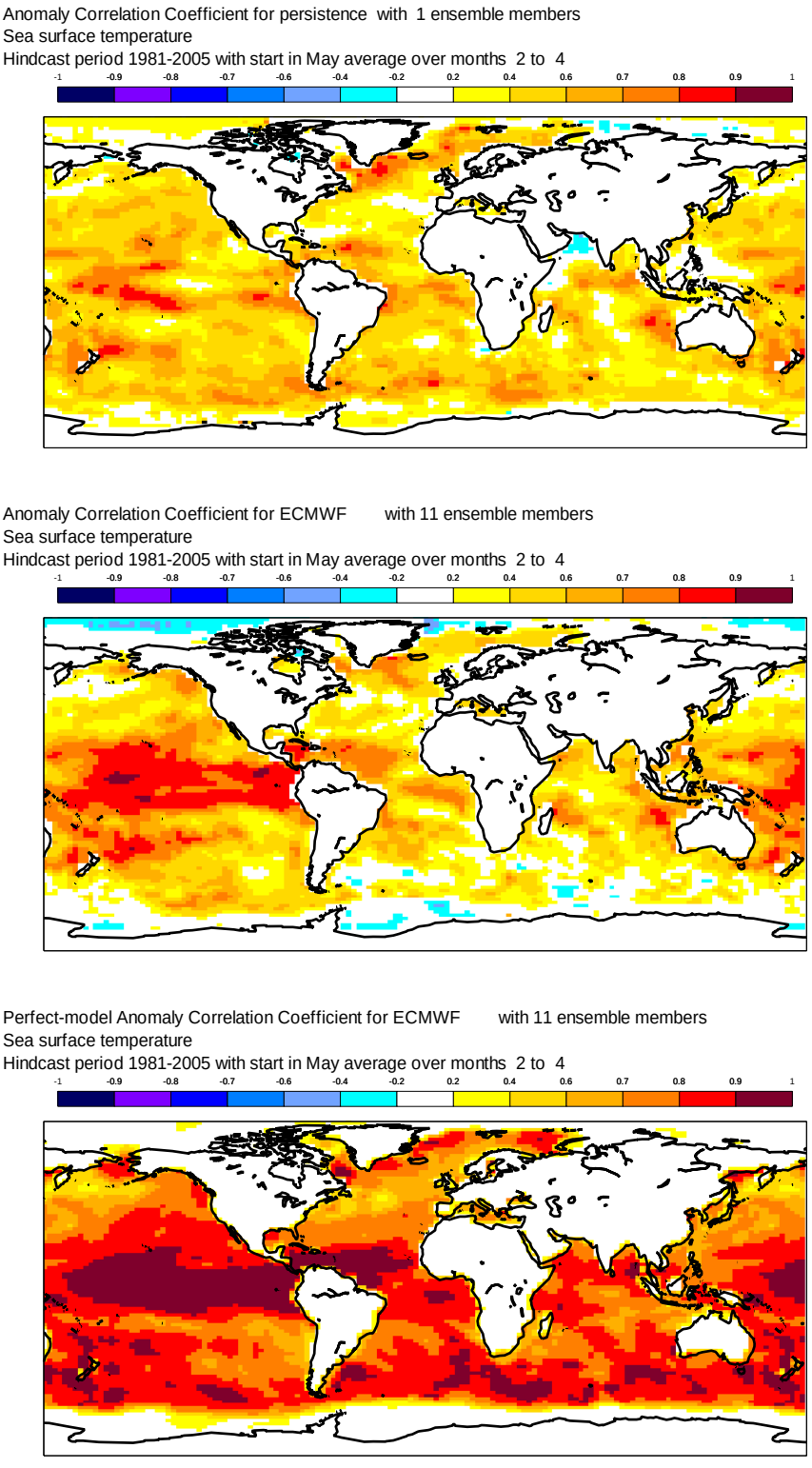


Figure 32: Spatial maps of the anomaly correlation (%) of seasonal forecasts for June-August (initialised on 1 May) from three forecast systems: persistence of the analysed April SST anomaly (top); operational System 3 ensemble-mean forecast (middle); "perfect model" forecast (see text for details, bottom). The correlations are computed using the hindcast integrations covering the period 1981-2005.

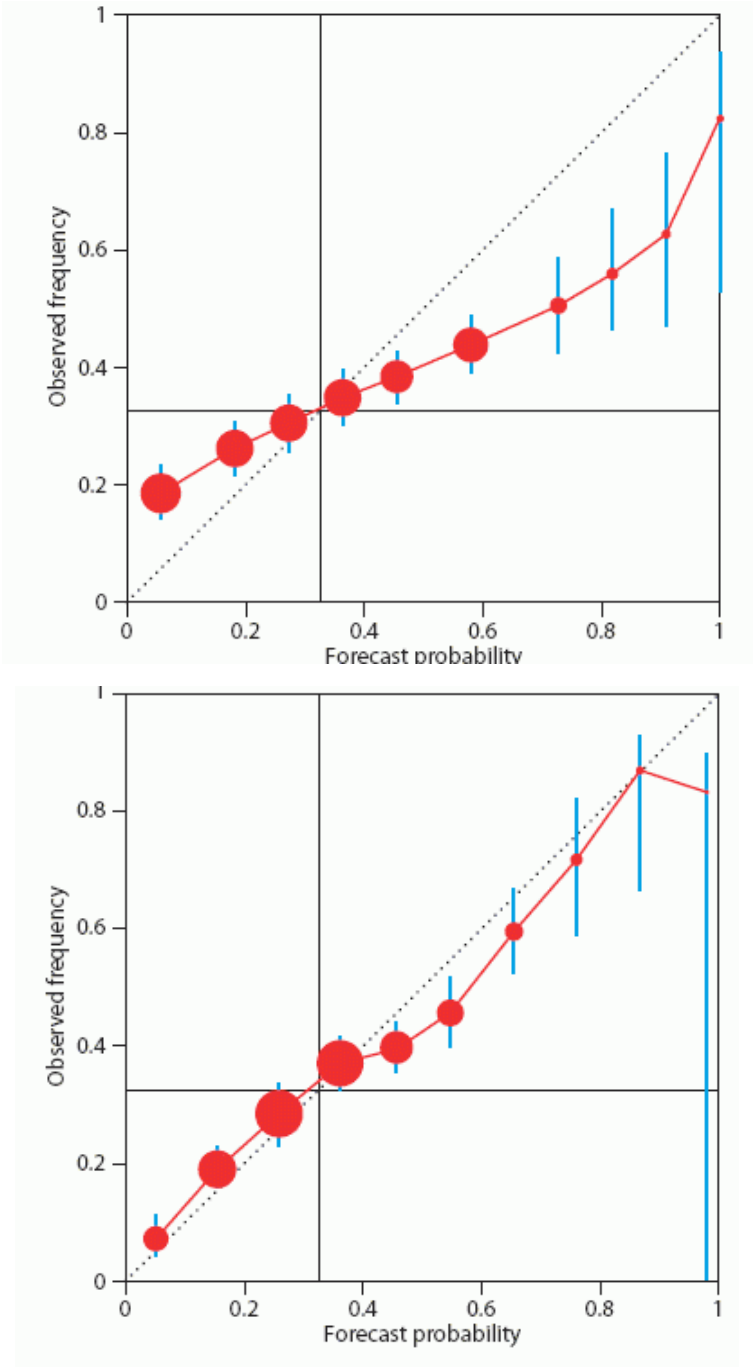


Figure 33: Reliability diagrams for the ECMWF model (top) and the EUROSIP system (bottom) for the December-February seasonal mean 2-metre temperature being in the lower third of the climate distribution. Forecasts are the hindcast integrations initialised on 1 November for the years 1987–2005. The sample is binned according to the forecast probability of the event (horizontal axis) and the red dots show the observed frequency of the event for each forecast probability bin. The size of the red dots indicates the relative number of cases included in each bin; the blue error bars show the effect of sampling uncertainty: probabilities higher than 0.6 are not often forecast in this sample (small dots) and the uncertainty is correspondingly larger than for lower probabilities (blue lines). The black lines show the climatological frequency of the event. For a reliable system, the observed frequency should match the forecast probability in each bin and the red dots would lie on the diagonal (dotted line). The EUROSIP system has a substantially higher reliability than the single model.

Annex I: A short note on scores used in this report

A. 1 Deterministic upper-air forecasts

The verifications used follow WMO/CBS recommendations as closely as possible. Scores are computed from forecasts on a standard 2.5 x 2.5 grid limited to standard domains (bounding co-ordinates are reproduced in the figure inner captions), as this is the resolution used for most products exchanged on the GTS. When other centres' scores are produced, they have been provided as part of the WMO/CBS exchange of scores among GDPS centres, unless stated otherwise - e.g. when verification scores are computed using radiosonde data (Figure 13), the sondes have been selected following an agreement reached by data monitoring centres and published in WMO/WWW Operational Newsletter.

Root Mean Square Errors (RMSE) are the square root of the geographical average of the squared differences between the forecast field and the analysis valid for the same time. When models are compared, each model uses its own analysis for verification; RMSE for winds (Figure 13, Figure 14) are computed by taking the root of the sums of the mean squared errors for the two components of the wind independently.

Skill scores (Figure 1) are computed as the reduction in Mean Square Error achieved by the model with respect to persistence (forecast obtained by persisting the initial analysis over the forecast range); in mathematical terms:

$$SS = 100 * \left(1 - \frac{RMSE_f^2}{RMSE_p^2} \right)$$

Figure 2 and Figure 4 show correlations in space between the forecast anomaly and the verifying analysis anomaly. Anomalies with respect to NMC Washington climate are available at ECMWF from the start of its operational activities in the late 1970s. For ocean waves (Figure 22, Figure 23) the climate has been derived from the ECMWF analysis.

A. 2 Probabilistic forecasts

Events for the verification of medium-range probabilistic forecasts are usually defined as anomalies with reference to a 10-year model climatology (1984-1993). This climatology is often referred to as the long-term climatology, as opposed to the sample climatology, which is simply the collation of the events occurring during the period considered for verification. Probabilistic skill is illustrated and measured in this report in the form of Brier Skill Scores (BSS), Ranked Probability Skill Scores (RPSS), and the area under Relative Operating Characteristic (ROC) curves.

The Brier Score (BS) is a measure of the distance between forecast probabilities p and the verifying observations o (which, as for any deterministic system, take only 0 or 1 as values). For a single event, it can be written as:

$$BS = \left(p - o \right)^2$$

As for any probabilistic score, however, the BS only becomes significant when results are averaged over a large sample of independent events. Its value ranges from zero (perfect deterministic forecast) to 1 (consistently wrong deterministic forecast). The Brier Skill Score is defined as:

$$BSS = \left(1 - \frac{BS}{BS_{cl}}\right)$$

Where BS_{cl} is the Brier Score for a climate forecast (forecast probability is constant and equal to the climatological probability of the event). Time series of the Brier Skill Scores can be found in Figure 20.

For multiple-category events, the Ranked Probability Score (RPS) is used. The RPS measures the distance between cumulative probabilities over the set of (k) events.

$$RPS = \frac{1}{k-1}\sum_{k}\left(\sum_{j \le k} p_j - \sum_{j \le k} o_j\right)^2$$

The RPS is equivalent to the average of the Brier Scores for exceeding the thresholds that separate the categories. The Ranked Probability Skill Score (RPSS) is defined similarly to the BSS, with the reference score being the RPS for a constant forecast of the climatological probability for each category. For the EPS upper-air verification, the climatology is based on ERA-40 analyses for 1979-2001. The RPS uses 10 climatologically equally-likely categories, so is equal to the average of BS for exceeding 10, 20, 30, …, 90 % of the climate distribution. The RPSS thus gives an overall measure of the probabilistic skill of the EPS at predicting a range of events.

There are four possible outcomes for a deterministic forecast of a dichotomous (yes/no) event: the event is forecast correctly (hit, H); the event is forecast and does not occur (False alarm, F); the event is correctly forecast not to occur (correct rejection, CR); or the event occurs but is not forecast (miss, M). The following measures are defined over a large sample:

Hit rate or probability of detection (POD) = H/(H+M)

False alarm rate = F/(F+CR)

False alarm ratio = F/(H+F)

Relative Operating Characteristic 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 one 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 will be issued (Figure 27). Figure 27 also shows a ‘modified ROC’ plot of hit rate against false alarm ratio.

Since the closer to the upper left corner (0 false alarm, 100% hits) the better, the area under the ROC curve (ROCA) is a good indication of the forecast skill (0.5 is no skill, 1 is perfect detection). Time series of the ROCA are shown in Figure 20 and Figure 29.

A. 3 Weather parameters (Section 4)

Verification data are European 6-hourly SYNOP data (area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the four closest grid points, provided the difference between the model and true orography is less than 500 m. A crude quality control is applied to SYNOP data (maximum departure from the model forecast has to be less than 100 mm, 25 K, 20 g/kg or 15 m/s for precipitation, temperature, specific humidity and wind speed respectively). 2 m temperatures are corrected for model/true orography differences, using a crude constant lapse rate assumption, provided the correction is less than 4 K amplitude (data are otherwise rejected).

For verification of EPS precipitation forecasts against analysis, the 0-24 h model forecast is used as a proxy for a model-scale analysis. A better alternative is to use an analysis derived from high-resolution networks upscaled to the model resolution. Although such data are not available in real time, ECMWF gets access to most networks in Europe and uses such analyses for internal purposes.