TECHNICAL MEMORANDUM

501

Verification statistics and evaluations of ECMWF forecasts in 2004-2005

F. Lalaurette, J. Bidlot, L. Ferranti, A. Ghelli, F. Grazzini, M. Leutbecher, D. Richardson, G. van der Grijn

Operations/Research Department

August 2006

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.html

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 medium-range free atmosphere ECMWF 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, oceanic waves and severe weather events. Finally, Section 5 provides insights into the performance of extended range forecast systems. A short technical note describing the scores used in this report is given in Annex A.

The set of verification scores shown here is mainly consistent with that of previous years, in order to aid comparison from year to year (ECMWF Tech. Memos. 346, 414, 432, 463).

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/seasonal/verification/ (seasonal range).

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

The list of changes to the system between September 2004 and August 2005 is as follows:

- 28 September 2004: Cycle 28r3
  - Revised convection scheme numerics and calling of the cloud scheme; use of tangent linear and adjoint of vertical diffusion in the 1st minimization of 4D-Var; Reduction of radiation frequency to 1 hour in the high-resolution forecasts; Improved numerics of surface tile coupling;
  - New Radiative Transfer code (RTTOV-8); minor revisions to ATOVS and AIRS usage; assimilation of MSG clear-sky radiances, GOES 9 AMVs and SCIAMACHY ozone products from KNMI; correction of error in AMSU-B usage over land; activation of EARS data;
  - Blacklisting of SYNOP humidity data at local night time; increased use of radiosonde humidity: use of RS90 to -80C, RS80 to -60C, other sensors to -40 C; proper cycling of the assimilation of the wave altimeter and land surface data (FG moved from 00 and 12 UTC to 06 and 18 UTC);
  - EPS Gaussian sampling for extra-tropical singular vectors, instead of selection and rotation; revision of tropical cyclones (TC) perturbations (extension from 25 S 25 N to 40 S 40 N, orthogonalisation with respect to extra-tropical singular vectors, optimisation regions based on predicted TC tracks from previous EPS run);
- 7 October 2004: Monthly forecasts are run operationally on a weekly basis (Thursdays);
- 10 October 2004: Harmonized bias correction for AMSU and HIRS radiances (called cy28r4 in Research documentation);
- 9 November 2004: All four BC-project analyses use background fields generated from the latest operational 4D-Var analysis;

- 16 March 2005: Tropical Cyclone tracks from the Deterministic and EPS forecasts are disseminated on the GTS in BUFR format;
- 5 April 2005: Cy29r1
  - New moist boundary layer scheme (more stratocumulus clouds in subtropical highs and more low level clouds in some extratropical winter conditions);
  - Wavelet formulation introduced in the representation of background error statistics (Jb), the tuning of which has been based on ensemble data assimilation runs;
  - Revised use of surface pressure observations: METAR are activated; all surface pressure data are subject to an adaptive bias correction; observation errors are reduced; PAOB observations are no longer used;
  - MODIS winds from AQUA are activated (previously only TERRA); AQUA winds observation error is reduced; 10 AIRS channels are blacklisted to address a problem in stratospheric humidity;
  - Bugfix in the first time-step of the semi-Lagrangian treatment of the physics; snow cover tile coupling revised; a new dissipation source function introduced for the oceanic wave modelisation;
  - Several technical implementations for the assimilation of rain and GPS data and for the variational bias correction of satellite radiances introduced but not activated;
  - Nonlinear Normal Mode Initialisation for the EPS was discontinued.
- 28 June 2005: Cy29r2 is implemented with the following changes:
  - Assimilation of rain-affected SSM/I radiances
  - Set of changes to AIRS assimilation
  - Use of Meteosat-8 (MSG) winds
  - Jb statistics from latest ensemble data assimilation
  - Refinement of surface-pressure bias correction
  - Humidity analysis changes affecting spin-up & stratosphere
  - Use of SMHI Baltic sea-ice analysis
  - Convection changes affecting mainly low level winds
  - Revised Gaussian sampling for EPS perturbations.

Note: All model 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 gives the evolution of the average forecast skill valid for consecutive 12-month periods since 1980. The forecast parameter is the 500 hPa height over the Northern Hemisphere (Extratropics only) and Europe, while the scoring method is 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 2005. The trend for a marked improvement in the quality of the forecasts observed over the last few years has reached a plateau this year, with even some degradation in the later stages of the forecast. However, if reference to climatology rather than persistence is used and errors are measured by anomaly correlation rather than root mean square differences (Figure 2), it is confirmed that for the first seven days performance over Europe and the Northern Extratropics is still at the very high level reached in the last few years. In particular there was a series of very good forecasts in the winter. In the Southern Extratropics, a further stage seems to have been reached with record-breaking scores during the last months of 2004. These two sets of figures give a somewhat different signal, which can partially be explained by the fact that the circulation in winter 2004-2005 was less active over Europe than the previous one, making persistence a more difficult forecast to improve upon (Figure 3). By contrast, summer 2004 was a very active one, third only to 1995 and 1999 in the last fifteen years. Full results for summer 2005 were not available at the time of writing, but July 2005 on this figure is also a very active one over Europe. This relatively high level of summer synoptic activity has, however, not resulted in the same degradation in performance as in 1999, which is further evidence that the efforts to address problems in the representation of physical processes at that time of year have yielded benefits.

The forecast of temperature anomalies in the medium range was made easier this winter by these relatively calmer conditions and this probably accounts for some of the improvement shown in Figure 4. It is interesting to see, however, that the EPS ensemble means (upper right panel) provide consistently more reliable forecasts at this forecast range (168h), whatever the conditions. Results for last summer confirm that, despite a more active warm season, performance has not been degraded, as it was in 1999.

One of the noteworthy results over the past few years has been that the improvement of the deterministic forecast quality has translated into improved consistency in the forecasts valid for the same date from one day to the next. This level of consistency has stayed very high again this year, despite the highly variable summer conditions, as can be seen from Figure 5, showing the time series of the average RMS difference between consecutive forecasts over Europe and the Northern Extratropics. Good consistency between consecutive forecasts is usually a feature that increases the confidence of forecasters in numerical forecasts - rightly so in this case, as the increase in forecast consistency is accompanied by a reduction in the size of the errors.

The quality of ECMWF analyses and forecasts for the upper atmosphere has been recognised by several institutions. Among them is WMO/AREP, which uses them when preparing the bi-weekly WMO Antartic Ozone Bulletins. The time series of wind scores at level 50hPa in the Extratropics is shown in Figure 6, confirming that this year, once again, the very good performance reached over the past couple of years was achieved.

The provision of EPS verification scores to JMA, initiated last year, has continued in 2005. This is ultimately intended to provide WMO/CBS members with a comparison of EPS scores. There is, so far, only a small

number of participants (JMA, KMA and ECMWF), so the exchange is still considered to be pre-operational and no comparisons can be made. Data are exchanged in the form of contingency tables, from which a variety of scores can be derived. As an example, Brier skill scores and ROC areas for the Northern Extratropics and Europe are shown in Figure 7. Some positive trend can be detected on these curves, but the level of noise attached to probabilistic verification scores when applied to a small sample makes it difficult to quickly detect major improvements or problems. Verifications of more basic parameters, such as the ensemble spread, react more quickly to problems. An example of this is shown in Figure 8, when the system suffered from a technical error, resulting in reduced spread. Corrective action was quickly taken and a correction implemented on 13 June.

3.1.2 Tropics

The skill over the Tropics, as measured by root mean square vector errors of the wind forecast with respect to the model analysis, is shown in Figure 9. The reduction in errors that followed the major model changes (both in the physics and data assimilation) early in 2003 (Cy25r4) is confirmed this year.

3.2 ECMWF vs other NWP centres

3.2.1 Deterministic (T511) model

The common ground for such a comparison is the regular exchange of scores between GDPFS centres under WMO/CBS auspices, following agreed standards of verification. Figure 10 shows time series of such scores over the Northern Extratropics for both 500hPa height and Mean Sea Level Pressure. These curves confirm our forecasts' very good performance, with our lead increased once again this winter, despite other centres' recent upgrades. Our error level early this summer is, however, much closer to those of the UK Met Office and US NCEP as in previous years. The gap is, in general, bigger in the Southern Hemisphere (Figure 11), even if the introduction of 4D-Var at the UK Met Office has clearly narrowed the gap during the cold season (April-July).

WMO exchanged scores also include verification against radiosondes over smaller areas such as Europe. Figure 12, showing both mass (Z500) and wind fields (850 hPa), confirms the good performance of our forecasts using this alternative reference.

The situation in the Tropics is summarised in Figure 13. The marked improvement in these scores since 2003 compared to other centres has been consolidated, with ECMWF 850hPa tropical winds now showing the smallest errors of all the centres, while at 250hPa there is only a more modest lead over the Met Office (UK).

4. Weather parameters and oceanic waves

4.1 Deterministic forecast

4.1.1 2 metre temperature (2mT)

The skill of the deterministic 2 metre temperature forecast is assessed against persistence error. Figure 14 shows a filtered time series of skill for four different forecast ranges verifying at 12 and 00 UTC for the European area. A seasonal trend in the performance is evident: late autumn to early spring forecasts perform better than summers, with the best during the winter months. Winter 2004-2005 is the exception, due to the frequent occurrence of anticyclonic and quiet weather conditions over Europe, characterised by a number of cold spells across the European continent, which led to an increase in cloud and temperature biases.

The skill for North America shows a positive trend, culminating with a maximum this winter, the 48 hour forecast being over 45% better than persistence (Figure 15). This value is an improvement with respect to the same period in previous years. As noted earlier for Europe, the performance shows a seasonal behaviour, with forecasts for late autumn to early spring performing better than those for the summer months.

4.1.2 2 metre specific humidity (2mq)

Figure 16 shows the filtered time series of 2 metre specific humidity skill (Mean Absolute Error assessed against persistence error) for the period January 1998 to July 2005 in Europe. The signal indicates a consolidation of the previous year's performance, with the best performance during the winter. It is interesting to note that in the last two winters the forecasts for t+60 range have reached the same skill as the t+48 forecast range.

4.1.3 Total Cloud Cover (TCC)

The Mean Absolute Errors over Europe reach their maximum values during the spring months for the forecast ranges verifying at 00UTC and minimum in the winter, as depicted in Figure 17. The time series (filtered) shows a reduction in the maximum value for spring 2005, which suggests that the changes introduced in the last model cycle have been beneficial.

4.1.4 Wind speed

Figure 18 shows a filtered time series of wind speed skill (Mean Absolute Error assessed against persistence error) for Europe. As already noted for other weather parameters, the performance has a marked seasonal behaviour. The trend suggests a consolidation of the previous years' behaviour.

4.1.5 Precipitation

The trend in precipitation skill score for the European area is shown in Figure 19. The time series of True Skill Score (or Pierce's Skill Score) is plotted, together with confidence intervals, for precipitation amounts above 10 mm/24h. Confidence intervals have been introduced to give a flavour of the uncertainty in the scores themselves: the larger the interval, the more uncertain the measure is. The last two rainy seasons (October to May) have been particularly well forecast by the model for thresholds up to 15 mm/24h. For higher thresholds the trends are not clear, as the confidence intervals widen considerably, due to a decrease in the sample size.

Nevertheless, the slow but constant improvement of precipitation forecasts has led to a one day gain in predictability over the last decade.

If the precipitation forecasts are stratified by season, it can be shown that winters have improved the most (Figure 20), while the summer precipitation forecasts do not confirm this positive trend (Figure 21).

4.1.6 Oceanic waves

Figure 22 shows a time series of the analysis error for the 10 m wind over maritime regions. The analysed wind is compared to the 10 metre wind observed by buoys in the oceans. The RMSE has steadily decreased since 1998, providing better quality winds for the forcing of the ocean wave model. The RMSE of the forecast wind and the wave height also decreased, as depicted in Figures 23 and 24, with excellent results for the last two winters. Better quality input and improvements in the ocean model itself have led to smaller RMSEs for the wave height.

The good performance of the system is confirmed this year, with a further reduction in errors when compared to the analyses (Figures 25 and 26). Comparisons with other models using buoy observations support this positive assessment. In Figure 27, the forecasts from models are verified against an independent dataset of 38 wave buoys in the northern hemisphere storm tracks. The various forecast centres contribute to this comparison by providing their forecasts at the locations of the agreed subset of buoys. Comparisons of the wave height RMSE of 5 day forecasts for the period February to April 2005 are shown in Figure 28. The ECMWF wave model has lower RMSE at analysis time and an advantage of about 2.5 days over the other models, except NCEP, which uses the independent buoy dataset for bogusing.

4.2 EPS forecasts

The European EPS performance for 4-day precipitation forecasts is summarised in Figures 29 and 30. The forecast is verified against observations stratified by rainfall amounts. The thresholds are for 1, 5, 10 and 20 mm/24 hours. The time series of Brier Skill Scores (reference system: sample climate) has been filtered using a 3-month running mean and shows a slightly decreased skill for the 2004-2005 European rainy season (October to April). Spring 2005 was characterised by below average rainfall in parts of Europe, this has affected the sample size, reducing the number of events in all classes but most likely in the higher thresholds. Figure 30 shows the time series of ROC Area calculated for 3-month periods. The decrease in the score value for the 20 mm/24h threshold resembles the loss of skill observed during summers, when the sample size for the higher thresholds decreases considerably. Less extreme thresholds confirm the previous years' performance of the EPS system.

4.3 Severe weather verifications

Verification of severe weather is complex. One of the issues that need to be taken into consideration is sample size. Many of the scores used for verification purposes are affected by the sample size. To overcome this deficiency, it is appropriate to introduce confidence intervals, which give an indication of the uncertainty associated with the score. A study by ECMWF of severe precipitation events has addressed the issue of the confidence assigned to a score when verifying forecasts. Confidence intervals have been constructed by using bootstrapping with replacements.

The case study included a set of severe precipitation events (about 100) which occurred during the period January 2001 to February 2002 in the United Kingdom. Figure 31 shows the True Skill Score for 3 thresholds as a function of forecast step. Forecasts were verified against the precipitation analysis computed for the T511 Gaussian grid, using the high resolution network data provided by the Member and Co-operating States. It is a gridpoint-to-gridpoint verification, whereby each gridpoint forecast is verified against its analysed value. The True Skill Score value for the 15 and 25 mm/24h rates decreases when the forecast range lengthens, as expected. The higher threshold (40 mm/24h) shows odd behaviour: the t+90 forecast range performs better than the shorter range forecasts. The introduction of confidence intervals for the TSS (Figure 32) illustrates that the previous conclusion is not appropriate, as the uncertainty in the 40 mm/24h score is quite large.

The issue of sample size in the verification of extreme events introduces complexity into the verification process. An approach widely used to increase sample size is to pull together cases occurring in different parts of the world or at different times of the year, which loosely corresponds to area or time averaging. Another approach is to relax the gridpoint-to-gridpoint verification requirement, that is, we consider a forecast to be correct, if at least one instance verifies within a defined area. For the purpose of this case study, the area was

considered to be the whole of the UK. The corresponding TSS score, shown in Figure 33, indicates better forecasting system performance when using this relaxed verification method. It is understood that the definition of the area is a key issue and sensitivity studies need to be carried out before any conclusions can be drawn.

4.4 Tropical cyclones

Verification of Tropical Cyclone (TC) forecasts is a routine activity that has recently been given a higher profile, due to the release of real-time TC forecast products on the GTS in BUFR format. Several hundreds have been tracked since 2002 (Figure 34), making it a suitable sample for both deterministic and probabilistic verification. 2004 was a very active summer, both in the North-West Pacific and the Northern Atlantic. June and July 2005 saw early activity in the Atlantic, with the number of cyclone-days above what was observed in 2003 and 2004.

A direct comparison between the deterministic T511 and T255 (EPS control) forecast TC track and the best track reported by WMO RSMCs is shown in Figure 35. To these deterministic forecasts, a “consensus” EPS forecast is added; the consensus being the average position of all EPS ensemble members that have successfully tracked a TC (Ensemble Mean). Results from the verification over 12 months (May 2004 - April 2005) are shown. Clearly the T511 forecast is providing the best results, most notably in terms of core pressure. When comparing these results to those gathered for the same period one year ago (Figure 36), it can be seen that the gap between the higher and lower resolution models has been reduced - a fact that is probably to be attributed to the several model changes aimed at reducing some of the excessive resolution dependence of the physical parametrizations. The direct comparison of this season’s scores with those of the previous one shows a degradation in the results. This, however, is not consistent with feedback we received from users of our TC forecasts. In the past, however, it has been seen that TC verification scores can vary considerably from year to year in both directions due to inhomogeneities in the observation sample. This area will therefore remain under careful scrutiny in the coming months.

A probabilistic verification of the “Strike probability” product offered on the web (the probability at any geographical point that a reported TC will get closer than 120km within the next 120h) is shown in Figure 37. Although an improvement in these scores was reported last year compared to the previous year, the trend seems to be reversed this year. It is difficult, however, to attribute the poorer deterministic verification to a degradation of the system, since all EPS changes introduced this year were carefully evaluated, including for EPS probabilistic TC forecast performance. Careful monitoring of the system will be needed next year, to ensure that this is more a sampling effect than a system degradation.

4.5 Other severe events

The Climate Atlas of Europe newly released by Météo-France as part of a EUMETNET/ECSN co-ordinated action has made it possible this year to lay the foundations of what could become a European scale verification system, using the ranking of observations within local climate records. Figure 38 shows the Relative Operating Characteristics of the EFI in predicting extreme precipitation events, while Figure 39 provides the same information, using the Probability of exceeding 20mm/day as a criterion to warn against the same extreme events. Clearly the EFI brings some refinement, allowing the detection of more events in the early medium range, even if this is at the expense of relatively high false alarm rates (left) and false alarm ratios (right).

5. Monthly and Seasonal forecasts

5.1 The 2004-2005 El Niño forecasts

During the first months of 2004, the sea surface temperature anomalies over the equatorial Pacific were close to neutral conditions. From July 2004, they became warmer (Niño 3.4 >0.7) and extended further over the Eastern Pacific (see Figure 40a). Several Kelvin waves associated with the western Pacific westerly wind events were observed during this period. One reached the South American Coast in January 2004, a second in June and a third in August. The moderate, warm conditions which developed in late 2004 can to some extent be attributed to the surfacing of the sub-surface temperature anomalies associated with those Kelvin waves.

From the end of January 2005, the warm anomalies near the date-line declined and the anomalies east of about 150W became noticeably negative. A westerly wind event developed in January over the Indian Ocean and it continued through February farther east near the dateline (not shown). This event was strong enough to induce an intense Kelvin wave that reached the South American coast by the beginning of May. By March the subsurface anomalies were between +4.5 and 5°C at about 130m dept, considering the up-welling of the very warm sub-surface anomalies associated with the Kelvin wave. By April its impacts became evident, with a significant warming not only along the South American coast but also westward over the Nino 3 area (see Figure 40b). However, by June, the sea surface temperature over the Equatorial Pacific returned to close to neutral conditions.

Figure 41 shows Nino-3.4 predictions throughout the year with subsequent verification (heavy blue dashed line). In general, forecasts over the Nino areas verified well for this year but we know that the system is, in general, sometimes over confident. Insufficient spread seems particularly evident for the forecasts over the Nino-4 area (not shown). Nino-3 predictions initiated in February and March both showed indications of the warming observed over this area during April (not shown). The latest El Niño forecasts present a relatively large spread and indicate that a more likely scenario is for SSTs to be close to neutral conditions.

5.2 Seasonal Forecast performance during 2004-2005

Despite the warm oceanic conditions of late 2004 over the Central Pacific, enhanced convection over the same area has been quite limited. The Southern Oscillation Index (SOI) was negative, but its anomalies were rather weak (see Figure 42). The model represented the extent of the SOI anomalies well. During this period the warm SST anomalies influenced only the local climate patterns, while the remote effect associated with ENSO teleconnections was not observed.

During spring/summer 2004, tropical storm activity over the west Pacific was higher than normal. The forecast consistently predicted an enhanced tropical storm frequency over that area. On the other hand, the forecast was not as good in reproducing the above normal tropical storm activity observed in summer over the Atlantic.

For the late spring and summer 2005 quite consistently warm anomalies were predicted over South-western Europe. Considering that Portugal and the South of Spain have suffered extremely dry conditions since last winter, it is quite possible that the warm predictions are mainly associated with the soil's dry initial conditions over those areas.

Although results are not discussed in this report, verification of the real-time multi-model seasonal system is in progress. Verification of each individual multi-model component (Météo-France and Met Office) is complete and is available on the internal website.

5.3 Monthly Forecast

The first operational monthly forecast was issued on 7 October 2004. Since then, the suite has been running weekly with Thursday at 00Z as the starting time. In this report, a few aspects of the monthly forecast performance during the recent period are discussed. For a description of the system and a comprehensive evaluation of the monthly forecast skill see Vitart 2003 (Monthly Forecasting System: ECMWF Technical Memorandum 424).

For summer 2004 the monthly forecasting system predicted consistently more tropical storm activity over the western North Pacific and over the North Atlantic than observed.

Over the Hudson Bay the monthly forecast system presented a surface temperature systematically warmer than the verifying analysis. Such a bias was related to a systematic under-estimation of the sea-ice initial conditions over that area. In some cases monthly predictions also exhibited unrealistic day to day evolution of sea-ice cover. A new treatment of sea-ice was implemented in the monthly system in June 2005 and it will be applied in the next seasonal forecast system.

By the end of January 2005 a long lasting blocking pattern developed, with a large-amplitude ridge over the central North Atlantic and flanking upper-level troughs over both the western North Atlantic and southern Europe. As a consequence, below average temperatures were observed in February over Southern Europe. The onset and demise of such a cold spell were well predicted by the monthly forecast up to 20 days in advance.

The predictions of precipitation from the monthly system during the Asian summer monsoon were evaluated by using 27 real-time cases covering the period May-June-July-August 2002, 2003 and 2004. Probabilistic scores for India (60E-85E, 10N-20N) and South East Asia (not shown) were computed. Figure 43 shows the ROC curves for the probability of precipitation to be in the above normal category (upper tercile) for different time ranges. It is interesting to see that the monthly forecast out-performs persistence at any forecast range and its skill becomes limited just beyond day 19-25. Over South-East Asia (not shown), the value of the ROC area remains above 0.5 even at day 25-32.

In addition, the monthly forecasting system has been particularly successful in predicting the monsoon onset. For example the 2005 late Indian monsoon onset was well predicted more than 20 days in advance, as well as the heat wave over India related to it.

Annex A : 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 12), 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 12, Figure 13) 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. Only for ocean waves (Figures 25, 26) has the climate 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 and the area under Relative Operating Characteristic (ROC) curves.

The Brier Score (BS) is a measure of the distance between forecast probabilities and the verifying observations (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:

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

Time series of the Brier Skill Scores can be found in Figure 7.

There are four possible outcomes for a deterministic forecast of a dichotomous (yes/no) event: the event if 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. Figures 37, 38 and 39 also show a "modified ROC" plot of hit rate against false alarm ratio.

Because 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 7 and Figure 30.

A. 3 Weather parameters (Section 4)

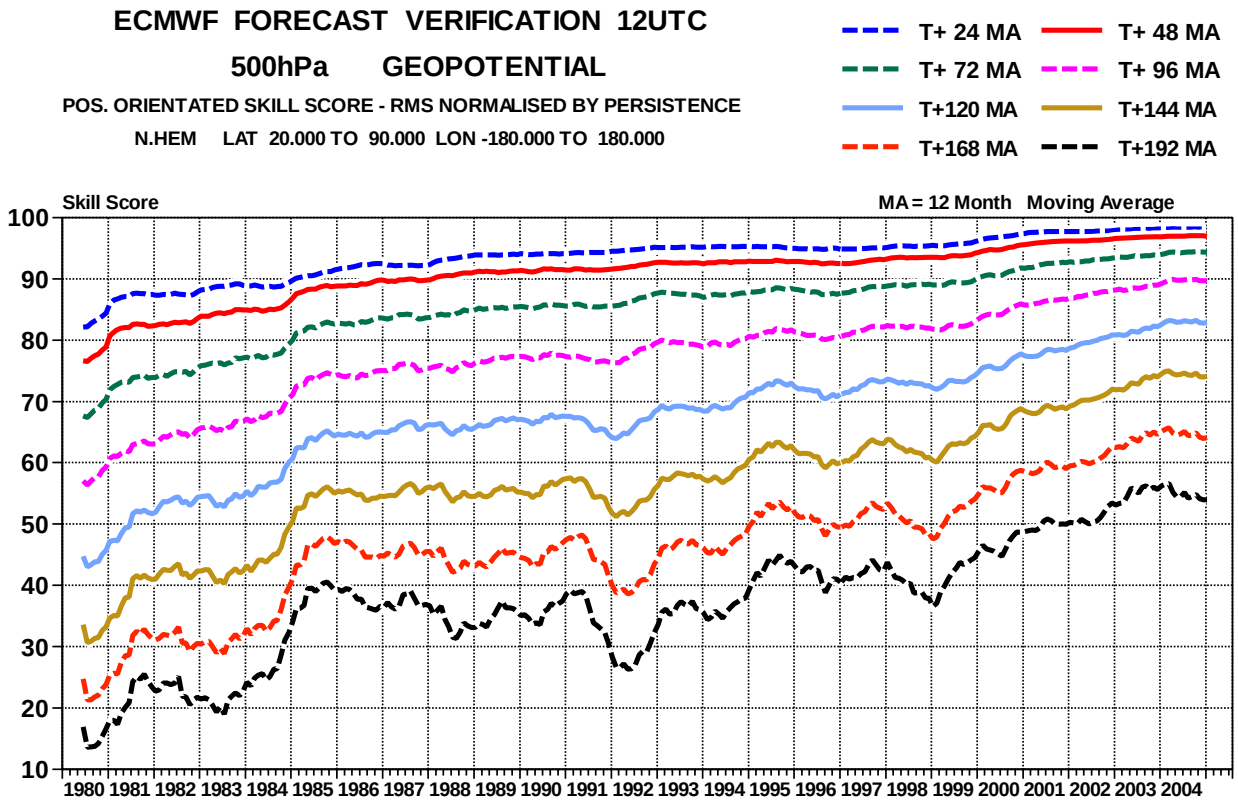
Verification data are European 6-hourly SYNOP data (limiting area boundaries are reported as part of the figure captions). Model data are interpolated to station locations using bi-linear interpolation of the 4 closest grid points, provided the difference between the model and true orography is less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

When verification against analyses for EPS forecasts of rainfall amounts is mentioned, the 0-24h-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 (e.g. Figure 31).

References

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

List of Figures



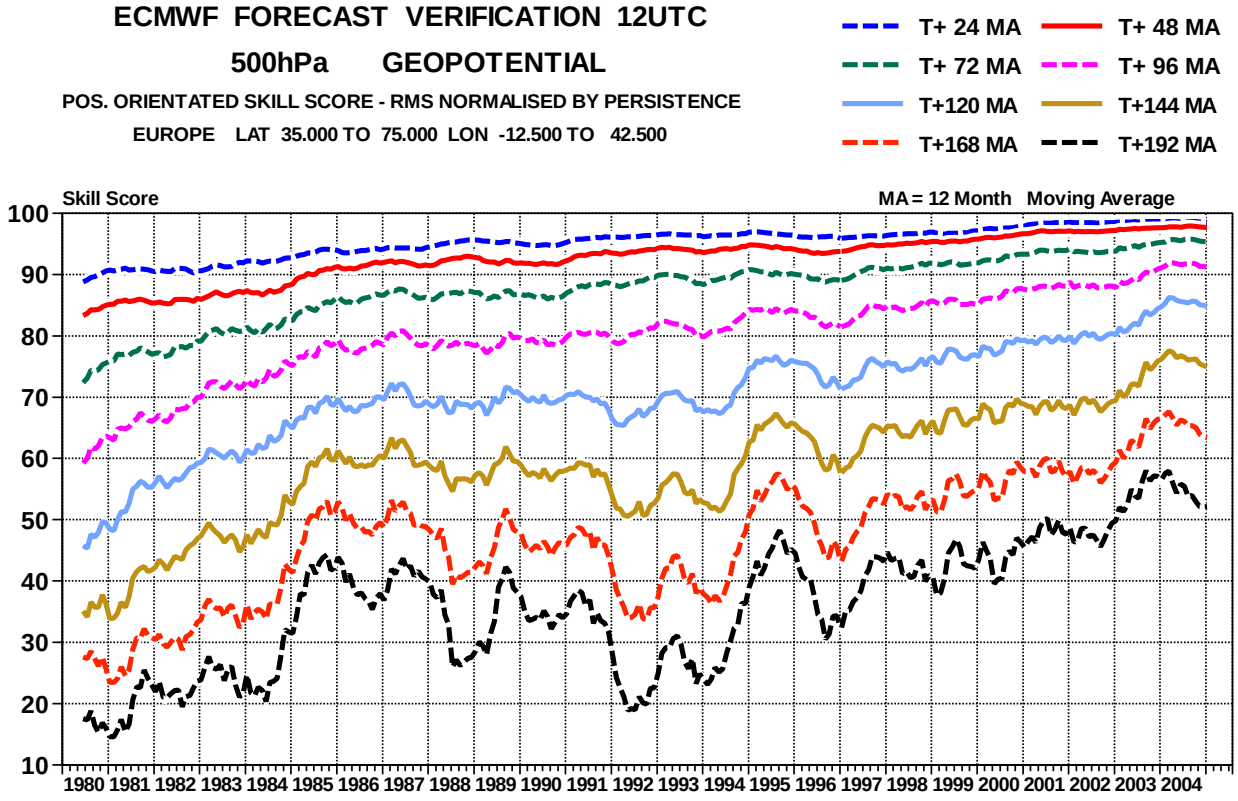


Figure 1: 500hPa height skill score (N. Hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)

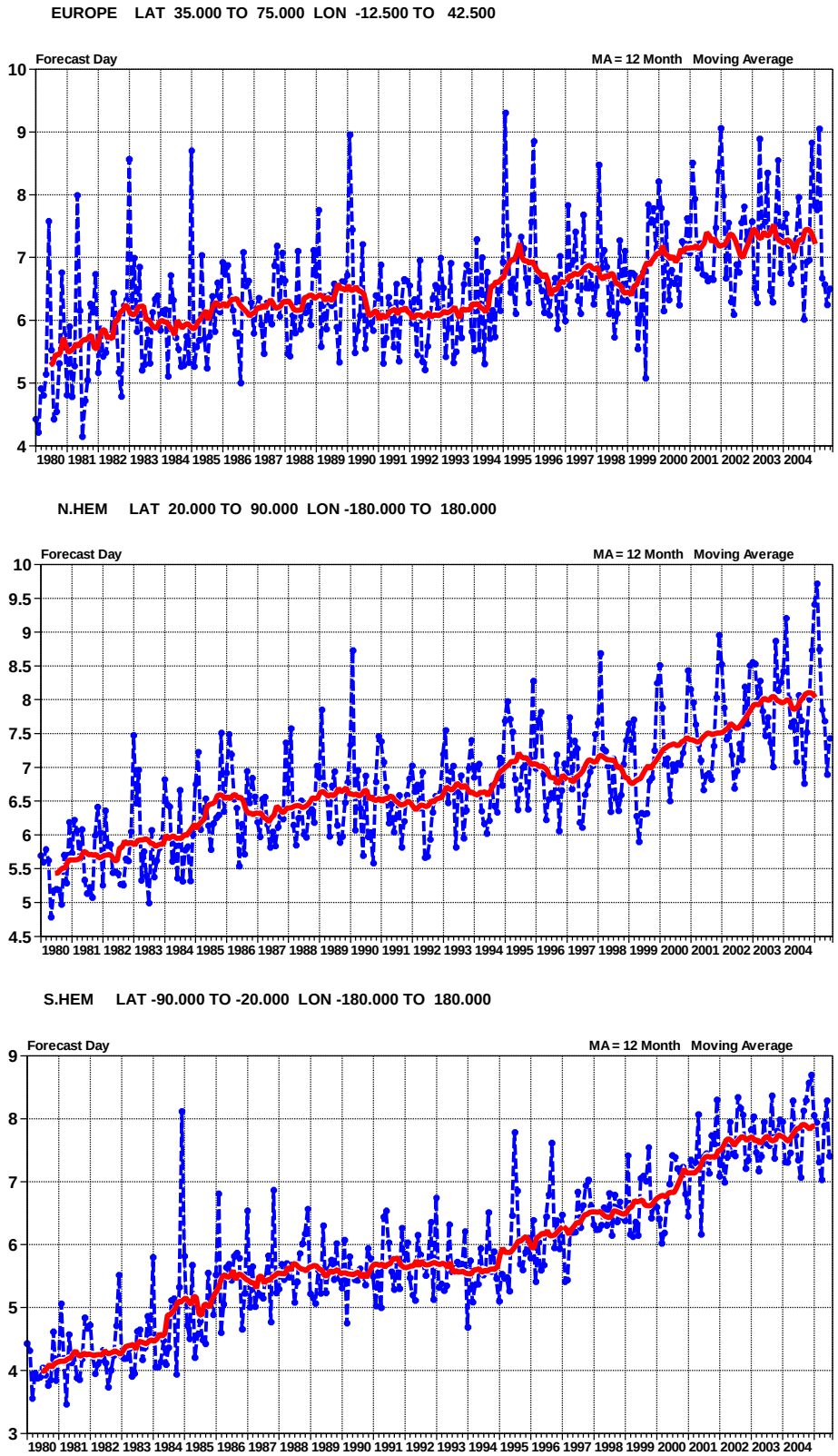


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the blue curves is the forecast range at which the monthly average of the daily forecast anomaly correlation with the verifying analysis falls below 60% for Europe, Northern and Southern Extratropics (the red curve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

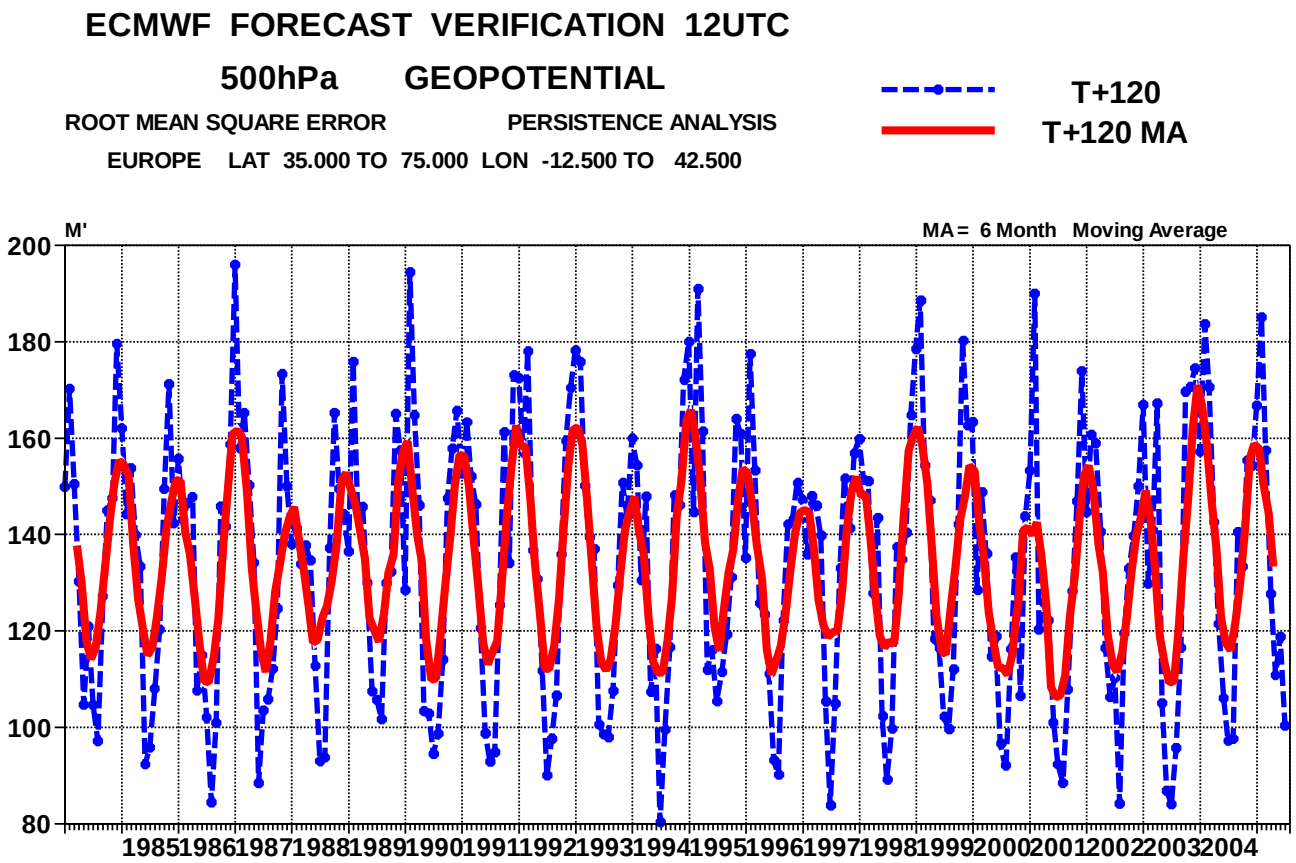


Figure 3: Root Mean Square Error made by persisting the analysis over 120h and verifying it as a forecast, monthly averages in blue, six-monthly moving averages in red.

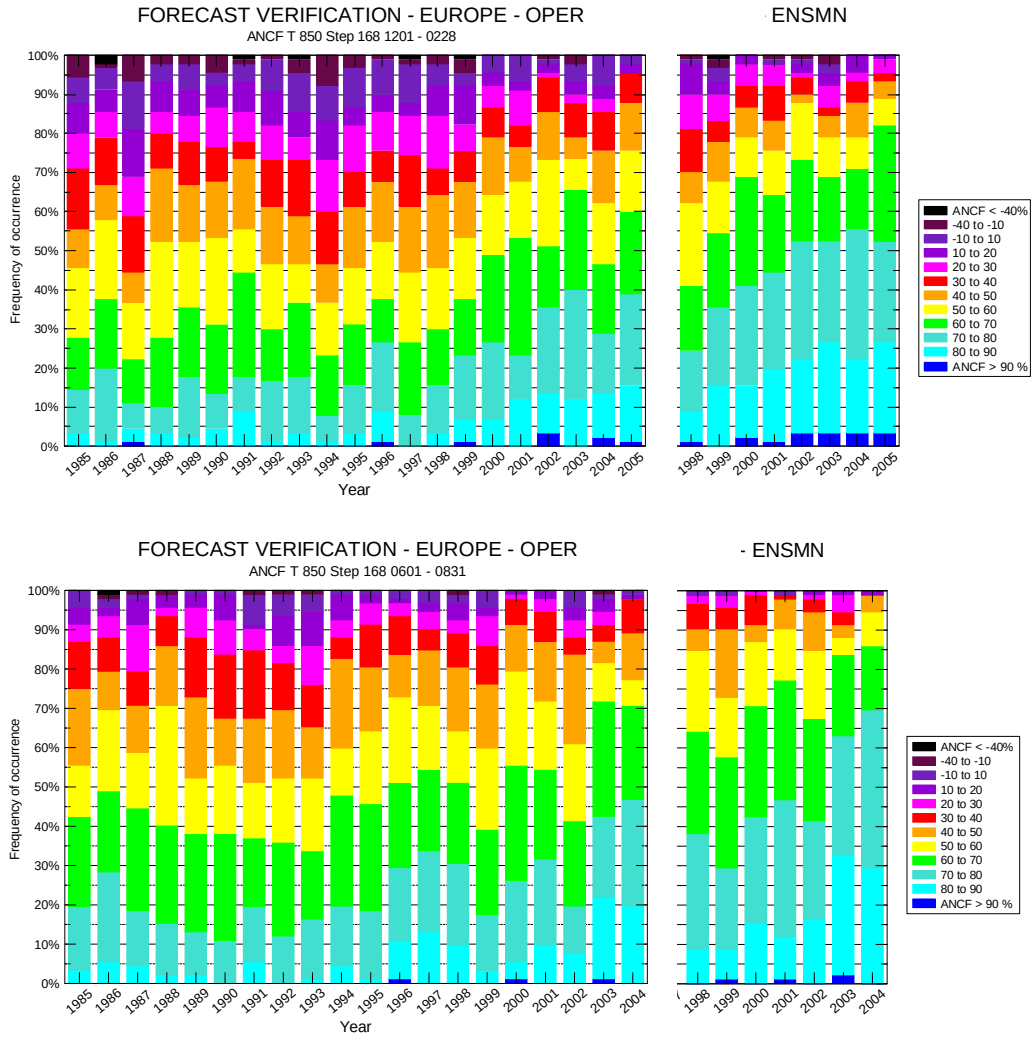


Figure 4: Cumulative distribution of Anomaly Correlation of the Day 7 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA., bottom) since 1984-85 for the deterministic, high resolution forecasts (left panels) and since 1997-98 for the EPS Ensemble mean (right panels).

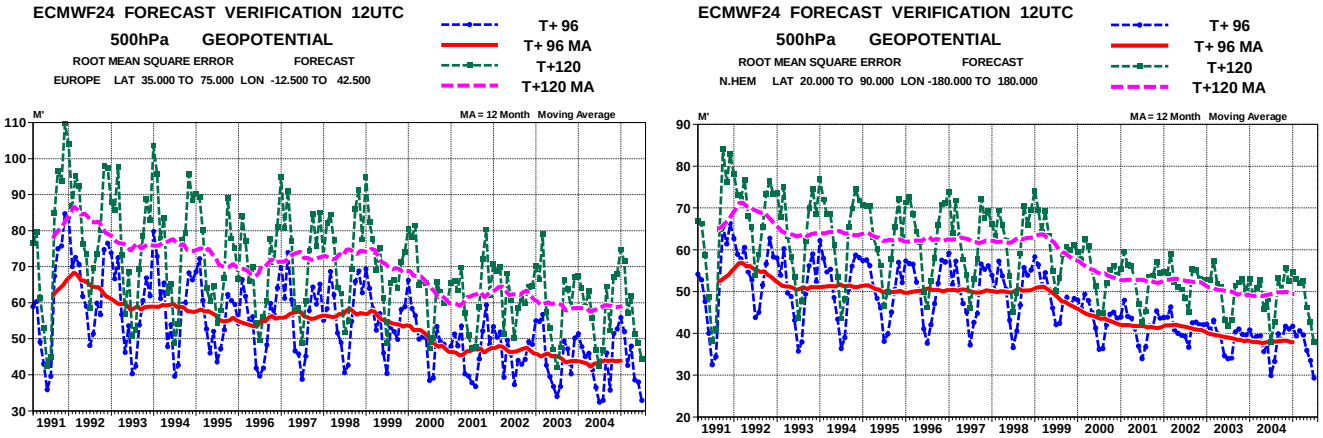


Figure 5: RMS difference between 24h-consecutive 500hPa height forecasts verifying on the same day over Europe (left panel) and Northern Extratropics (right panel). The last month plotted is July 2005.

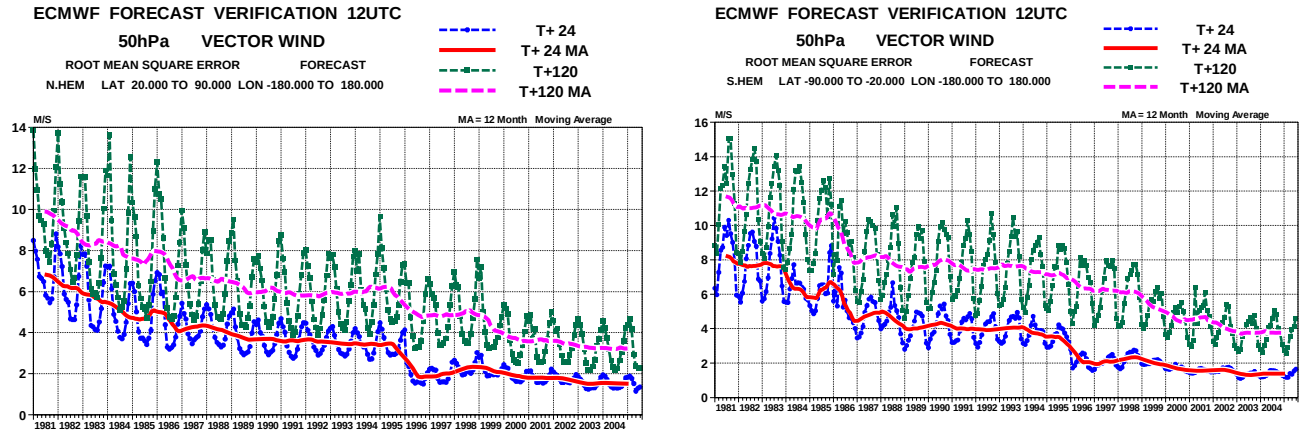


Figure 6: Model scores in the extratropical Northern (left) and Southern (right) Hemisphere stratosphere (RMS vector wind error at 50hPa for 1-day and 5-day forecasts)

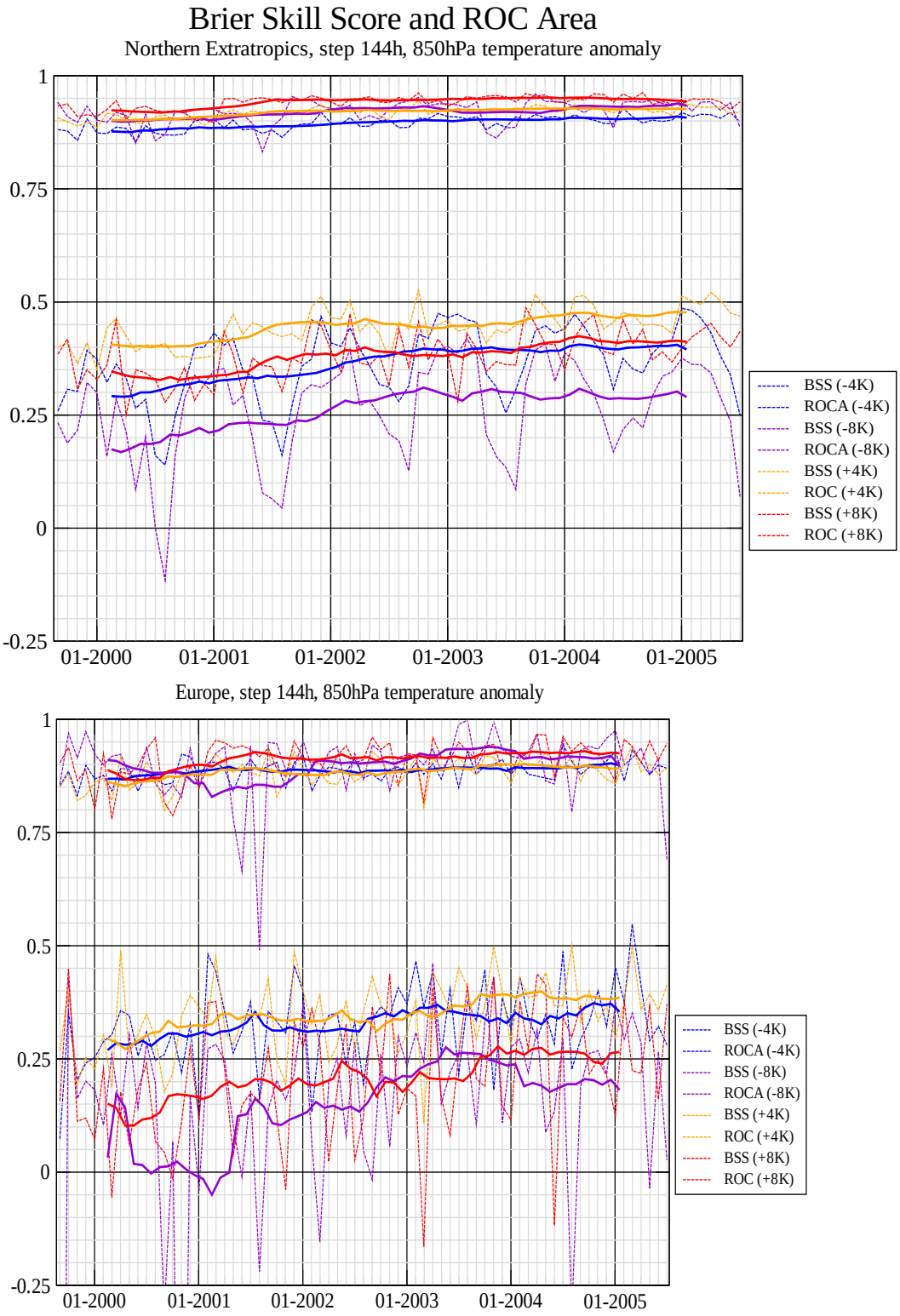


Figure 7: Time series of the skill of 144h EPS probability forecasts of 850 HPa temperature anomalies exceeding a series of thresholds for the northern hemisphere (upper panel) and Europe (lower panel). In each panel, the upper set of curves are the Relative Operating Characteristics Area (ROCA) and the lower set of curves are the Brier Skill Score; both monthly averages (dashed lines) and 12-month averages are shown.

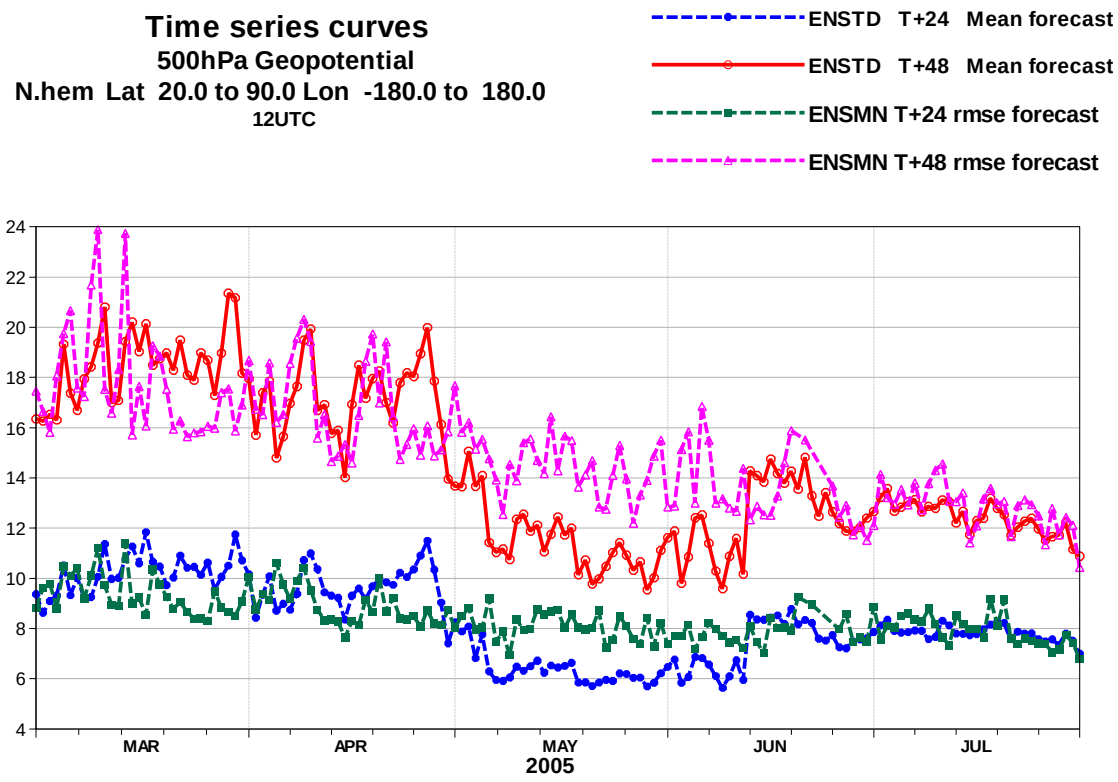


Figure 8: Daily time series of EPS spread and Ensemble Mean error at Day 1 and 2 over the Northern Extratropics. The reduction in spread that was introduced by mistake in early May can clearly be seen. The situation was restored to normal on 13 June 2005.

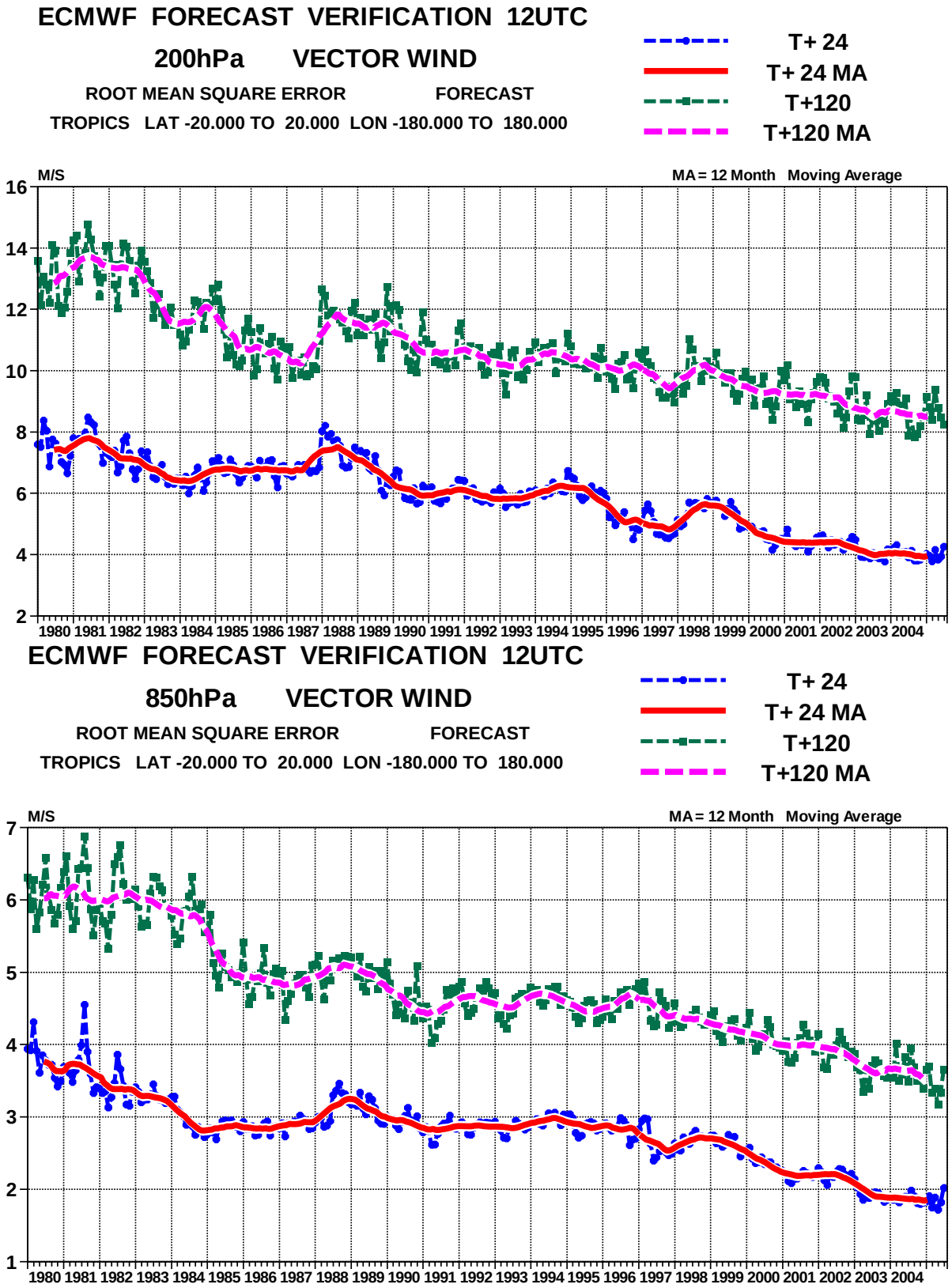


Figure 9: Model scores in the Tropics (root mean square vector wind errors at 200hPa and 850hPa for 1-day and 5-day forecasts; monthly mean and 12-month running mean scores)

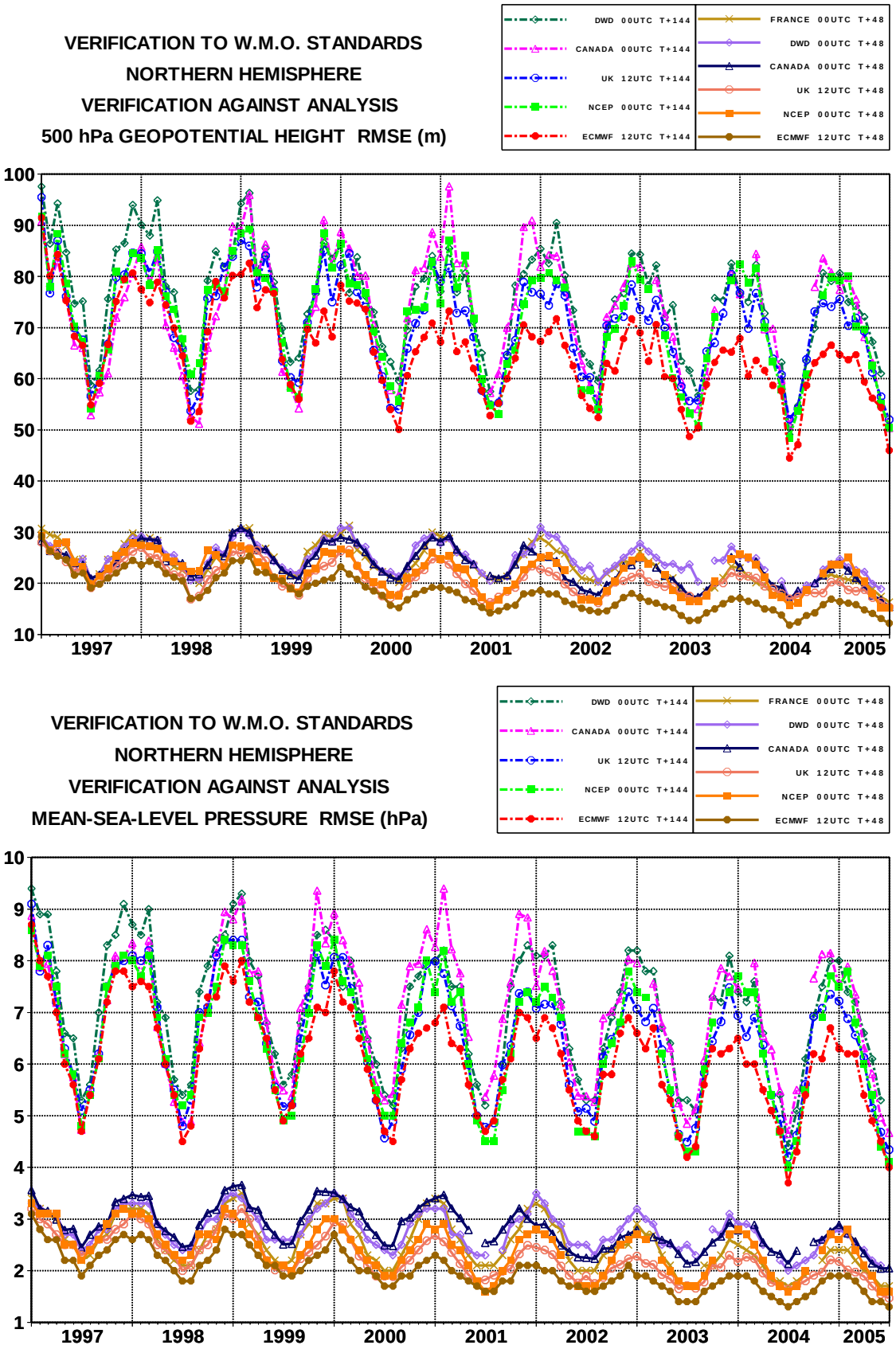


Figure 10: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500hPa and MSLP for D+2 and D+6)

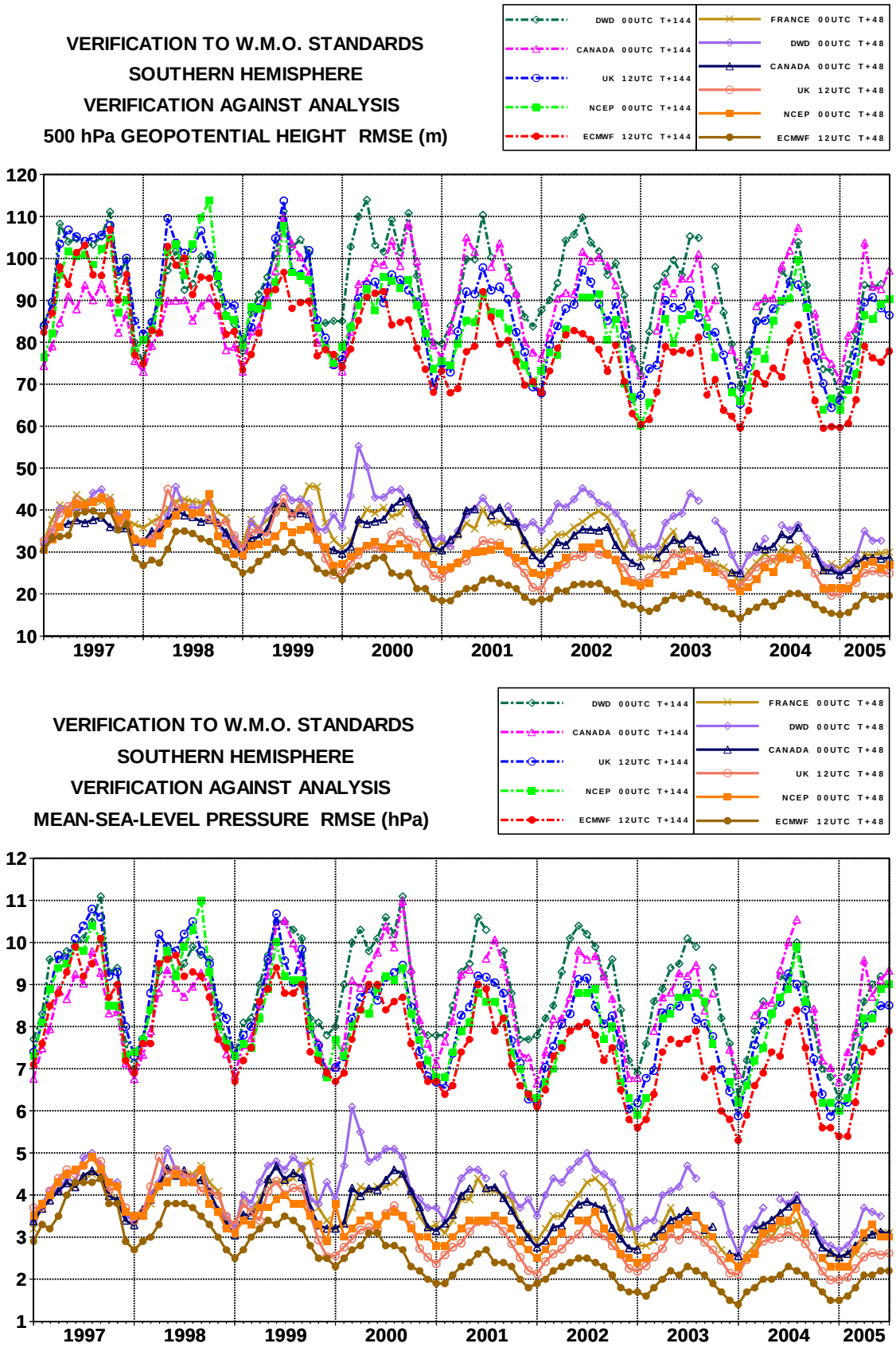


Figure 11: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500hPa and MSLP for D+2 and D+6)

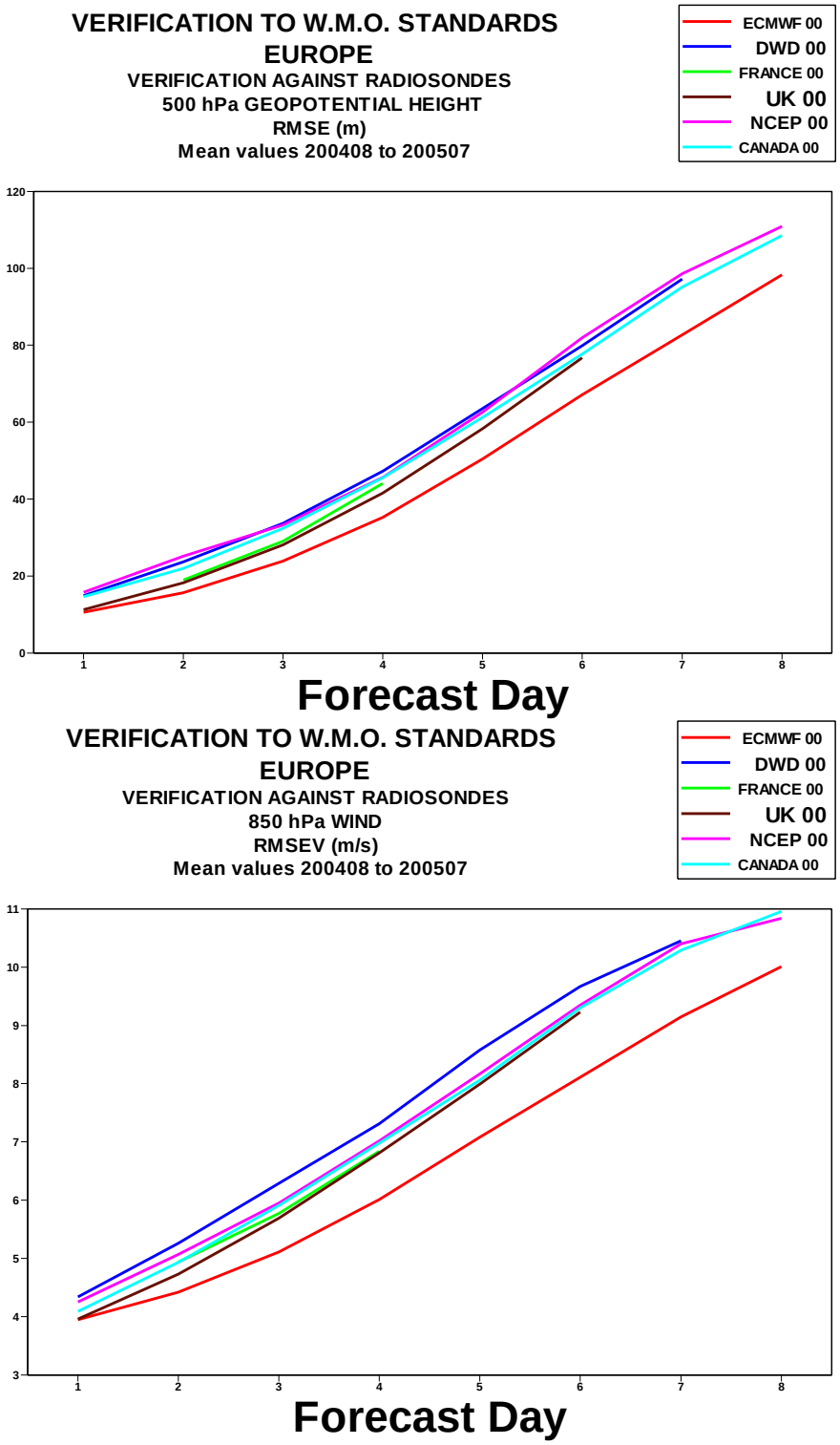
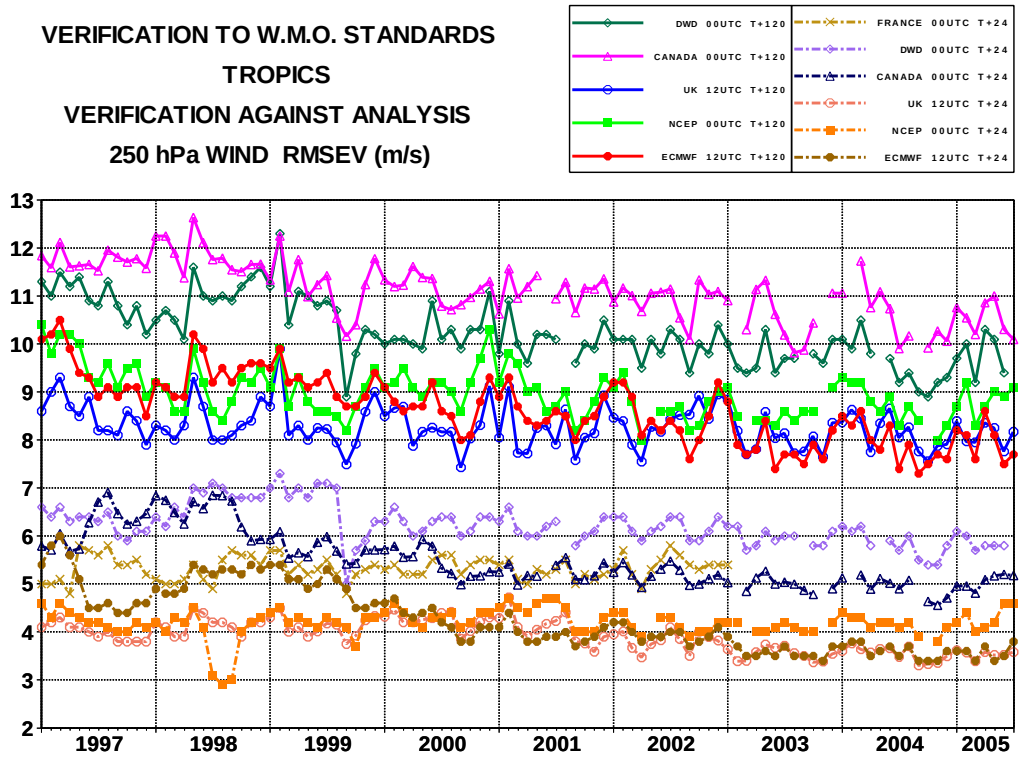


Figure 12: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)



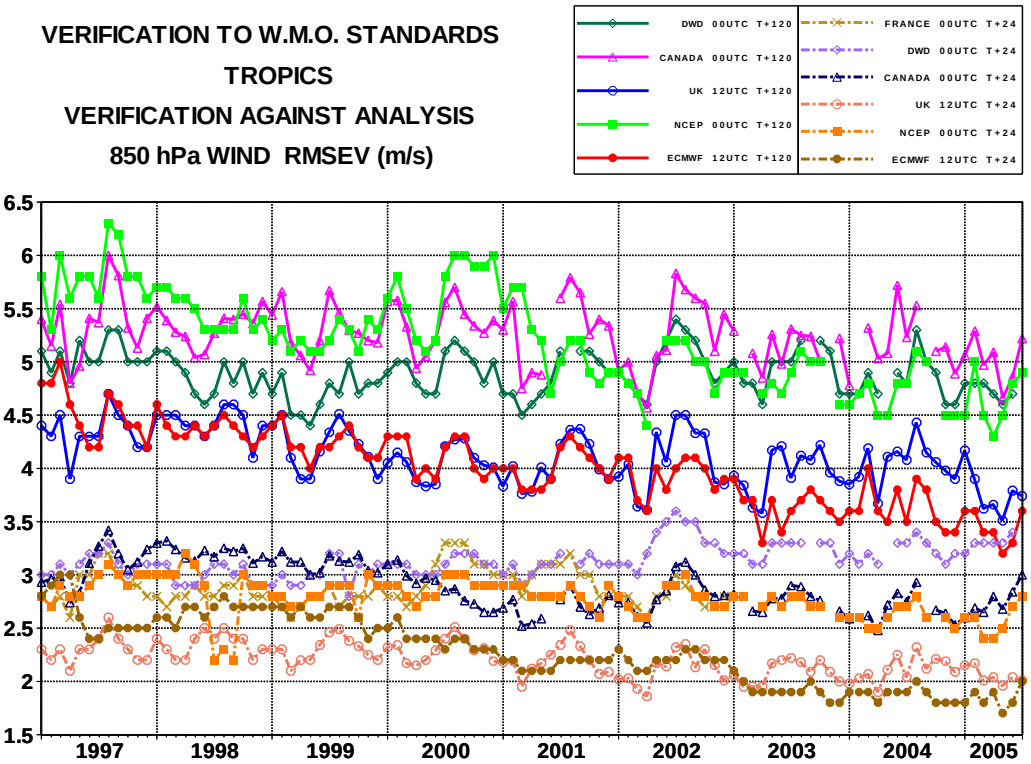


Figure 13: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250hPa and 850hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis

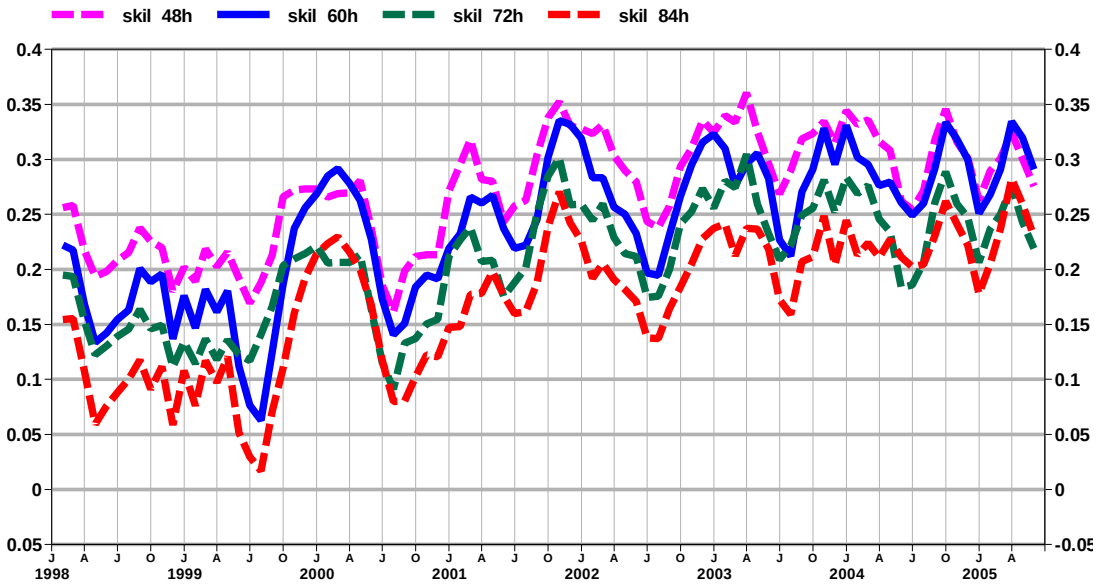


Figure 14: Skill (RMSE assessed against persistence error) of 2 metre temperature forecasts verified against SYNOP data on the GTS for Europe. A three-month filter has been applied.

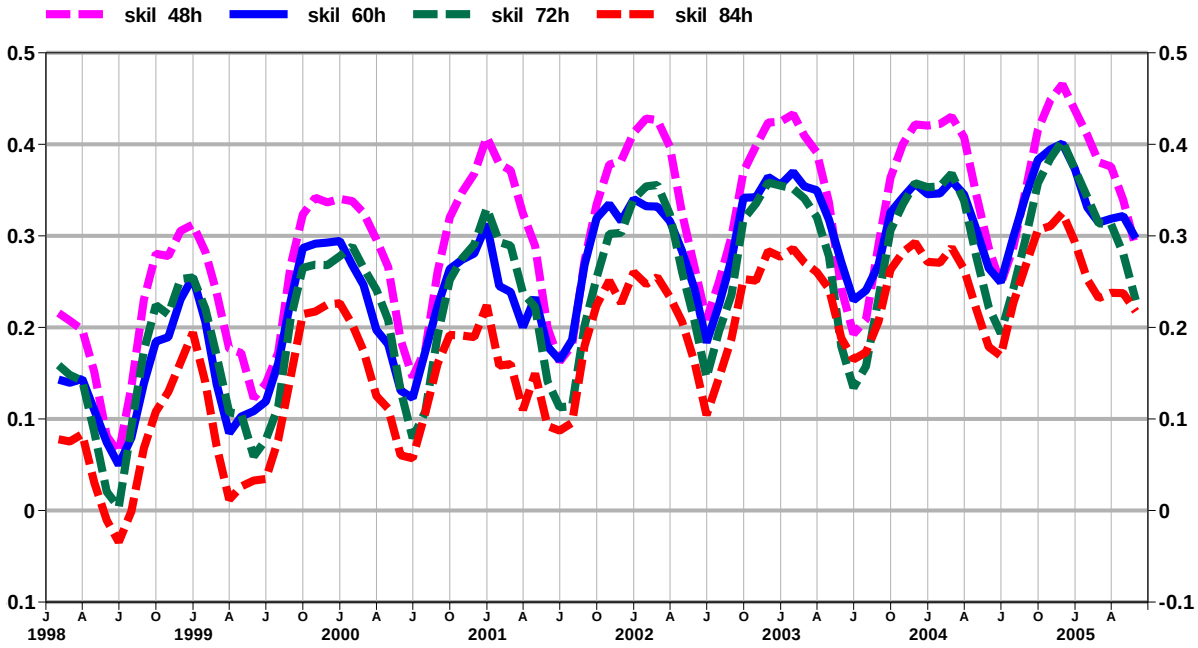


Figure 15: Skill (RMSE assessed against persistence error) of 2 metre temperature forecasts verified against SYNOP data on the GTS for North America. A three-month filter has been applied.

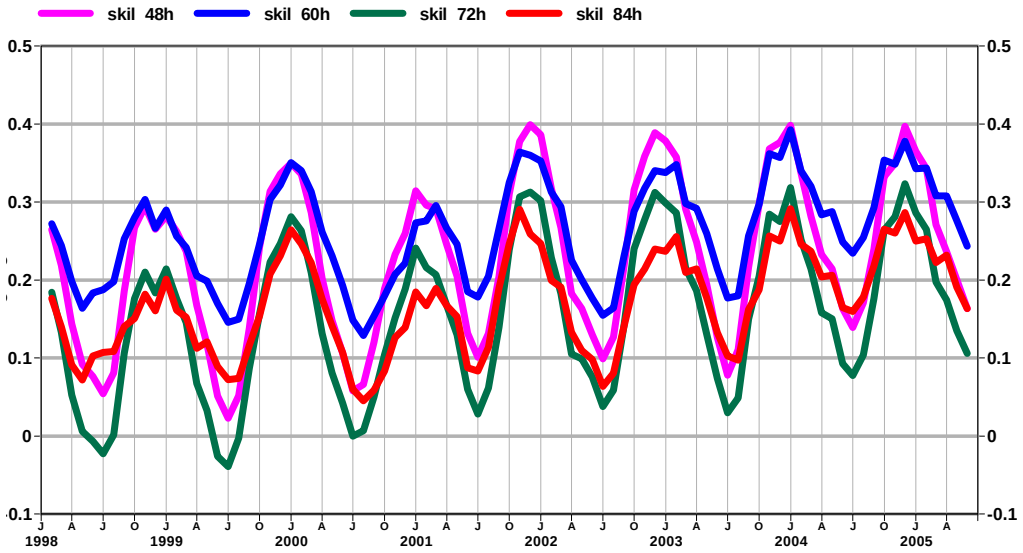


Figure 16: Skill (MAE assessed against persistence error) of 2 metre specific humidity forecasts verified against SYNOP data on the GTS for Europe. A three-month filter has been applied.

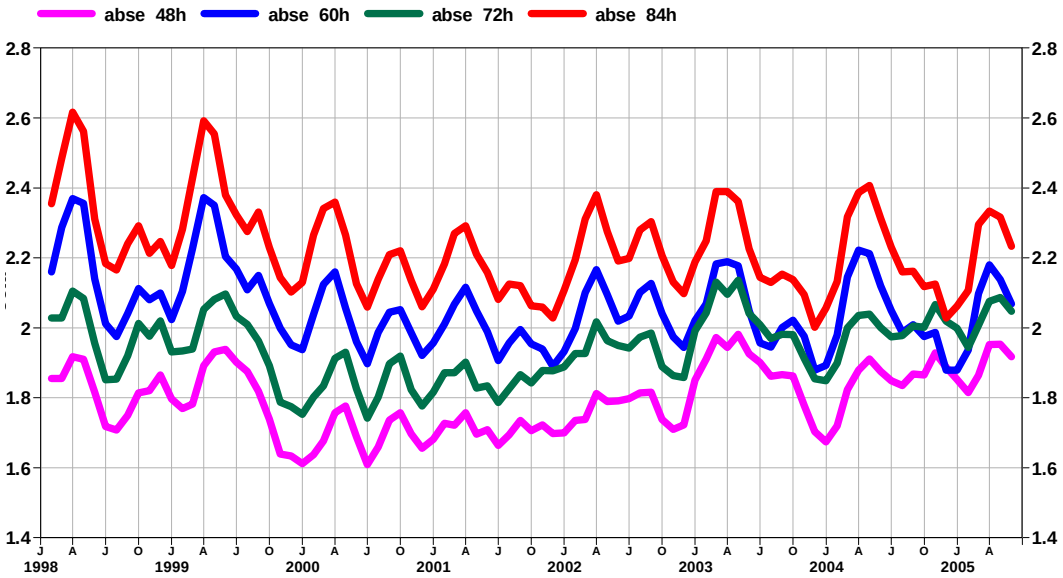


Figure 17: Mean Absolute Error (oktas) for the total cloud cover forecasts verified against SYNOP data on the GTS for Europe. A three-month filter has been applied.

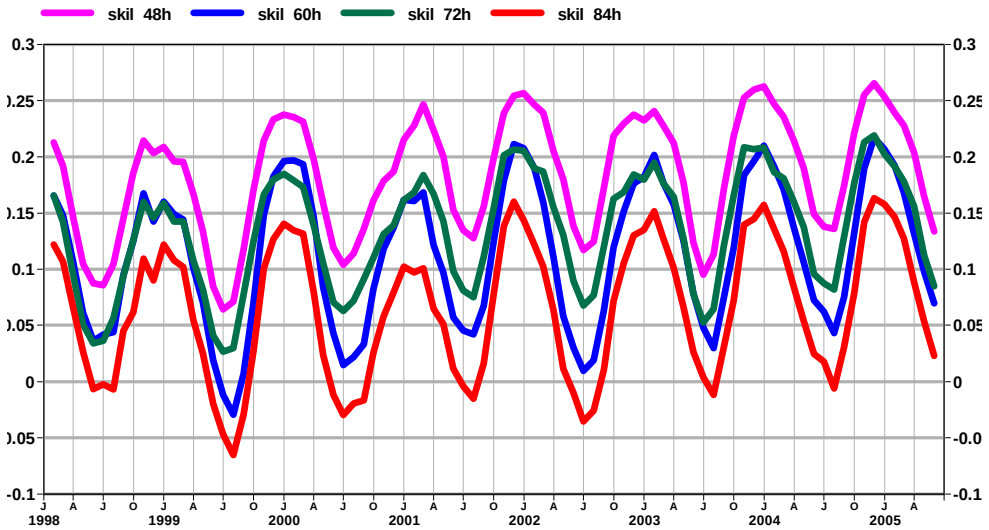


Figure 18: Skill (MAE assessed against persistence error) of 10 metre wind forecasts verified against SYNOP data on the GTS for Europe. A three-month filter has been applied to the 4 time series representing four forecast steps.

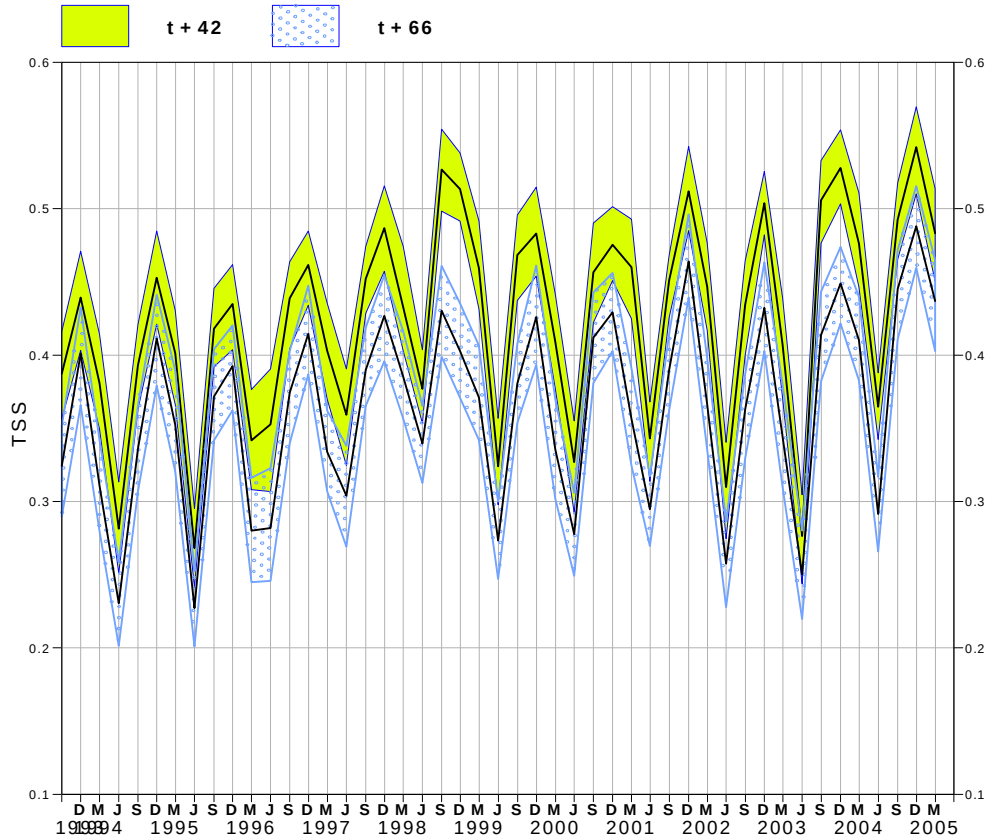


Figure 19: TSS time series (three-monthly values) for precipitation forecasts exceeding 10mm/24h verified against SYNOP data on the GTS for Europe. The TSS values (black dashed line) for the t+42 and t+66 forecast ranges are plotted together with the 5% and 95% confidence intervals (green solid shade for t+42 and blue dotted shade for t+66).

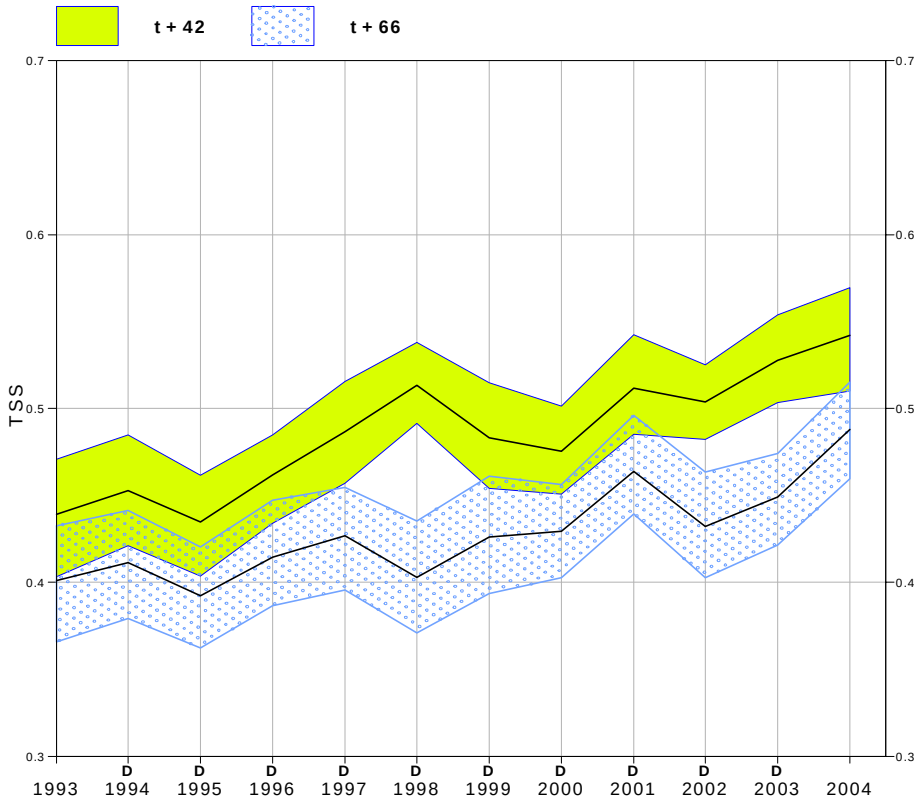


Figure 20: TSS time series for winter cases only. The precipitation forecasts exceeding 10mm/24hare verified against SYNOP data on the GTS for Europe. The TSS values (black dashed line) for the t+42 and t+66 forecast ranges are plotted together with the confidence intervals (green solid shade for t+42 and blue dotted shade for t+66).

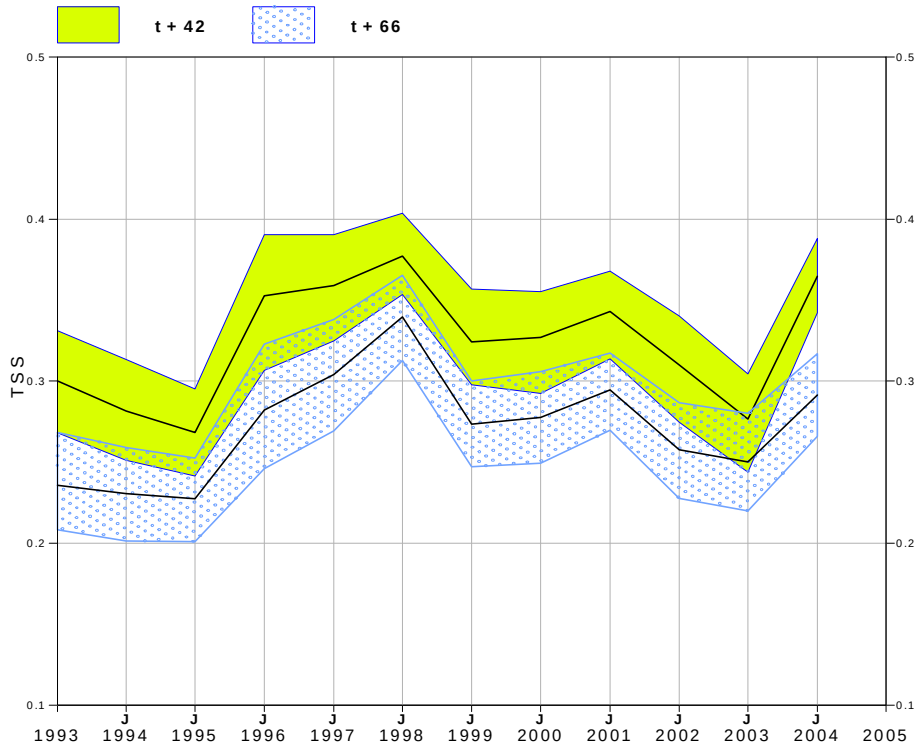


Figure 21: Same as before but for summer cases only.

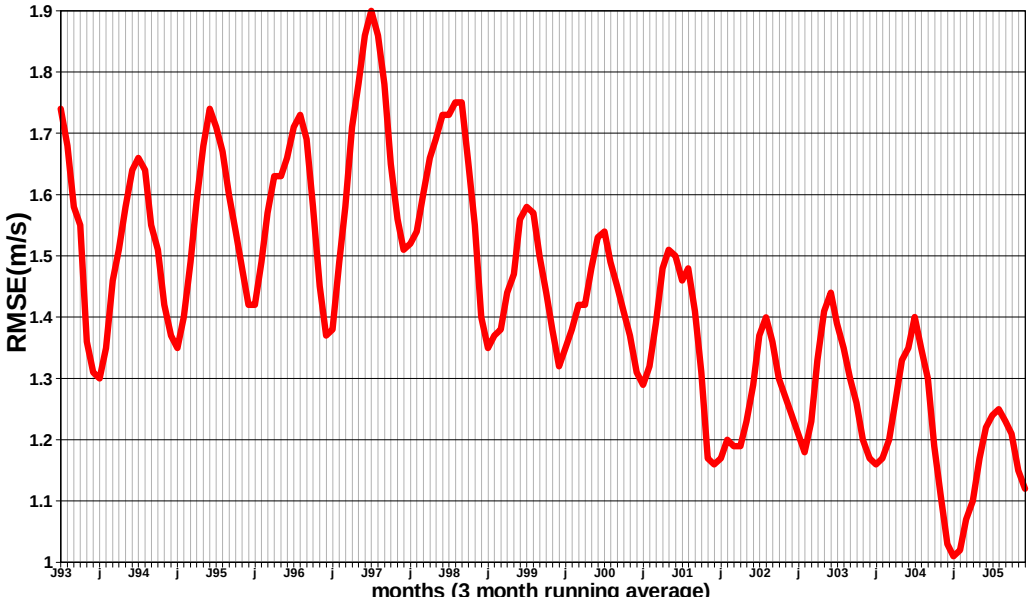


Figure 22: time series of RMSE for the ECMWF 10 metre wind analysis verified against northern hemisphere buoy observations. A three-month running mean is used.

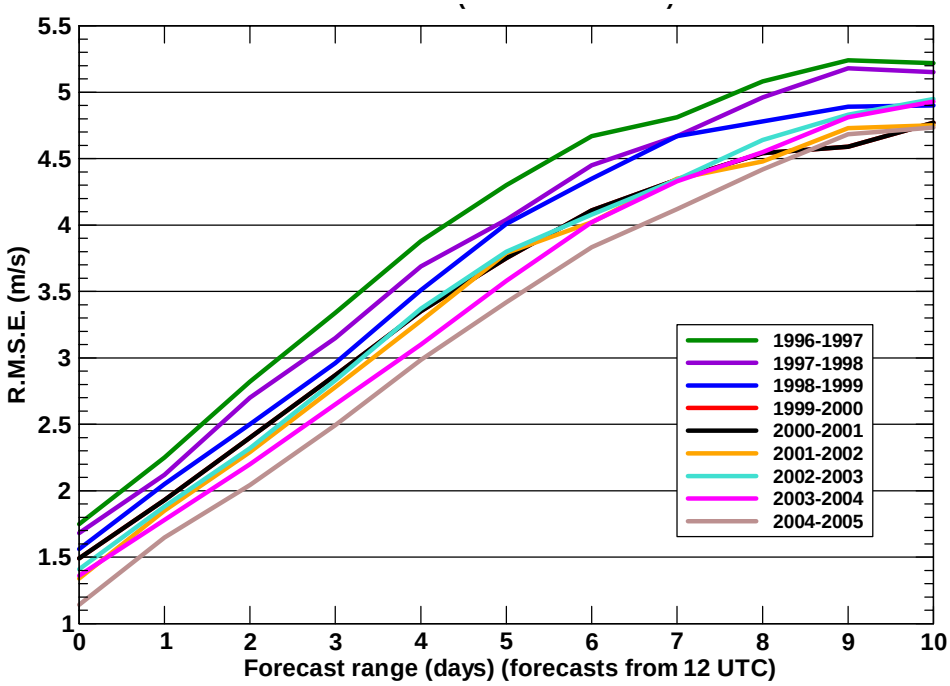


Figure 23: RMSE for the ECMWF 10 metre wind forecast verified against northern hemisphere buoy observations. The RMSE is plotted against the forecast range for a set of consecutive winters (October to March).

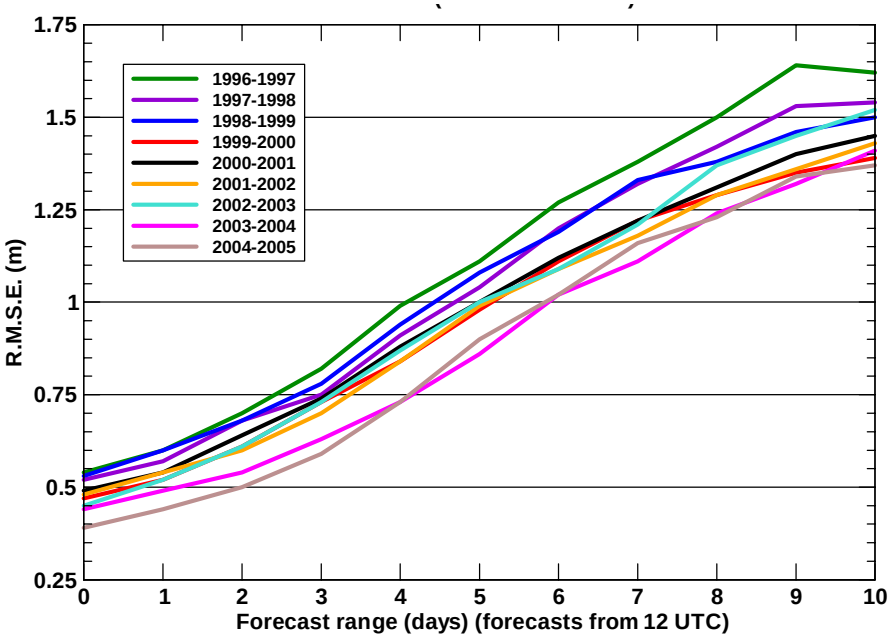
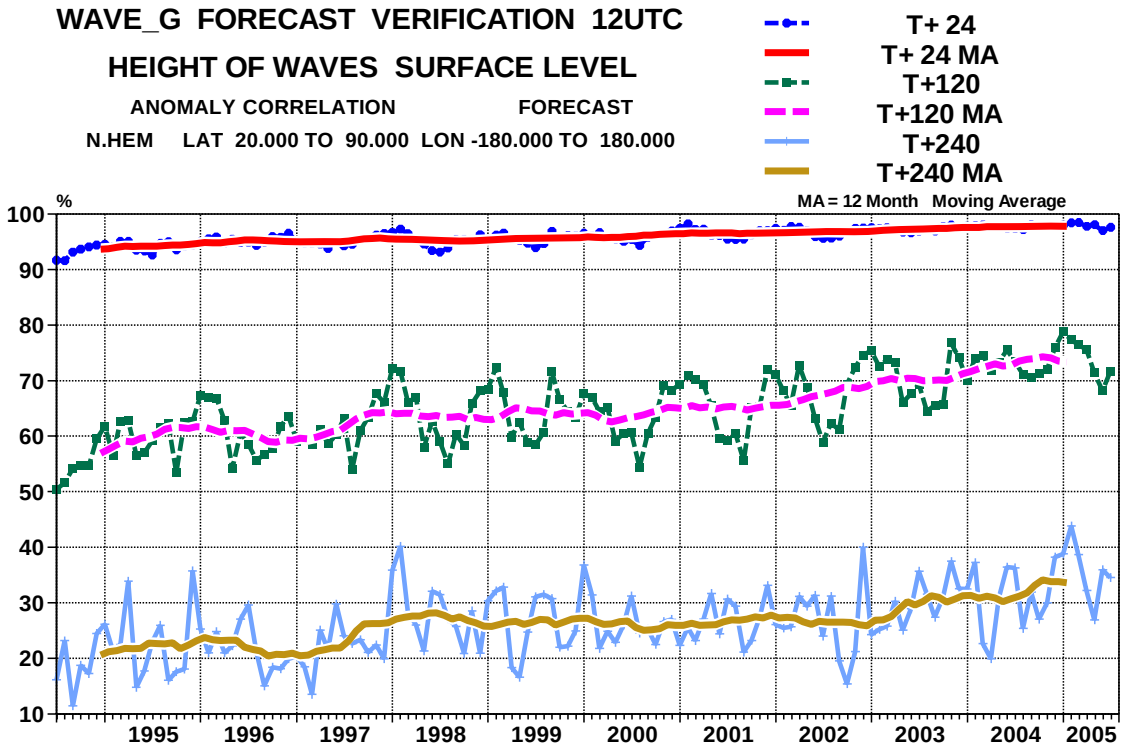


Figure 24: RMSE for the wave height forecast verified against northern hemisphere buoy observations. The RMSE is plotted against the forecast range for a set of consecutive winters (October to March).



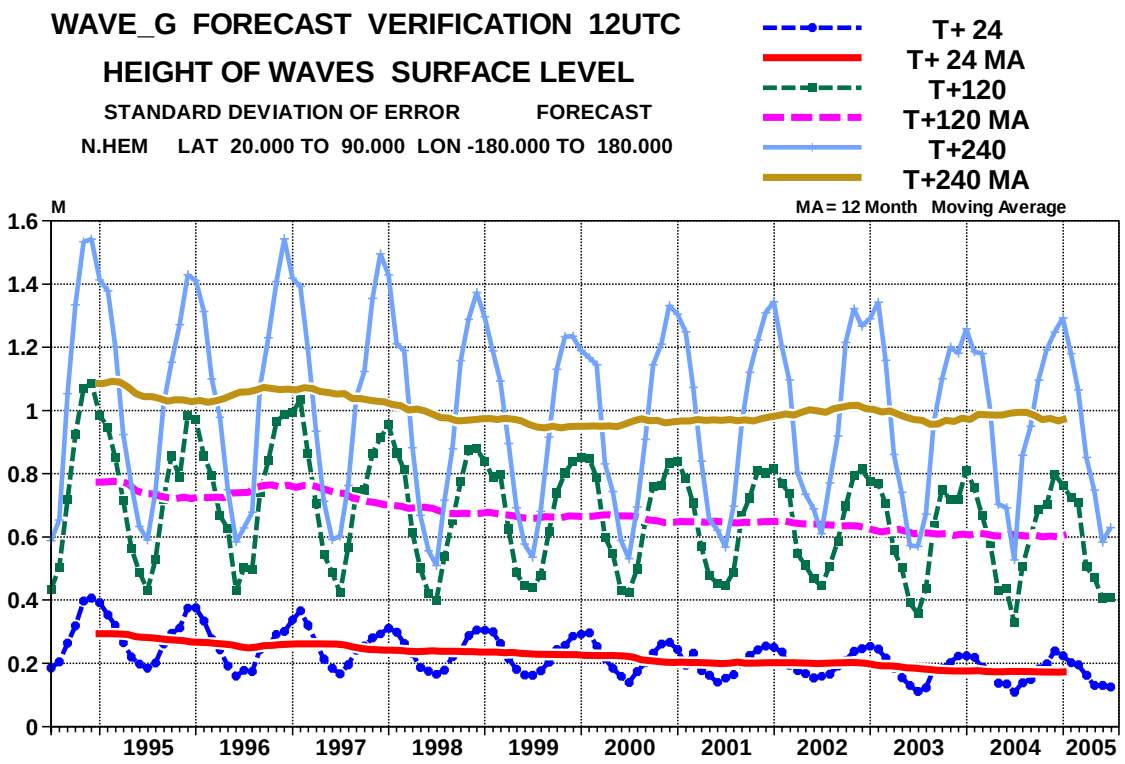
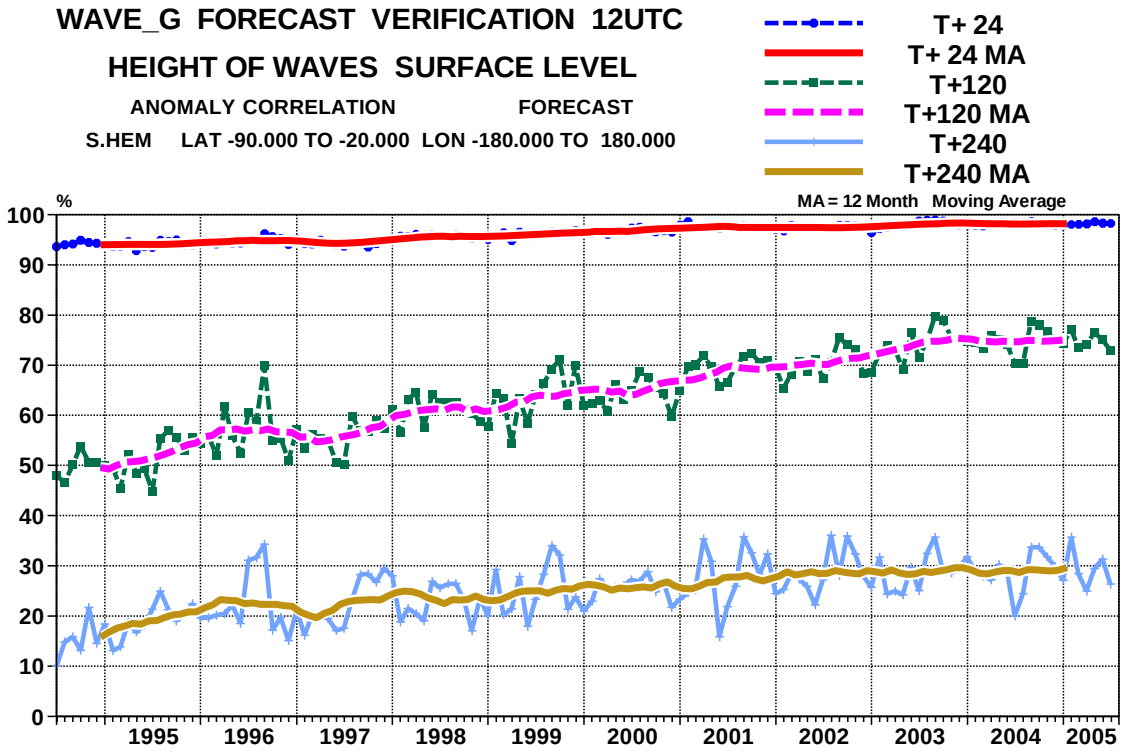


Figure 25: Scores (anomaly correlation and standard deviation) of ocean wave heights verified against the analysis (Northern Extratropics)



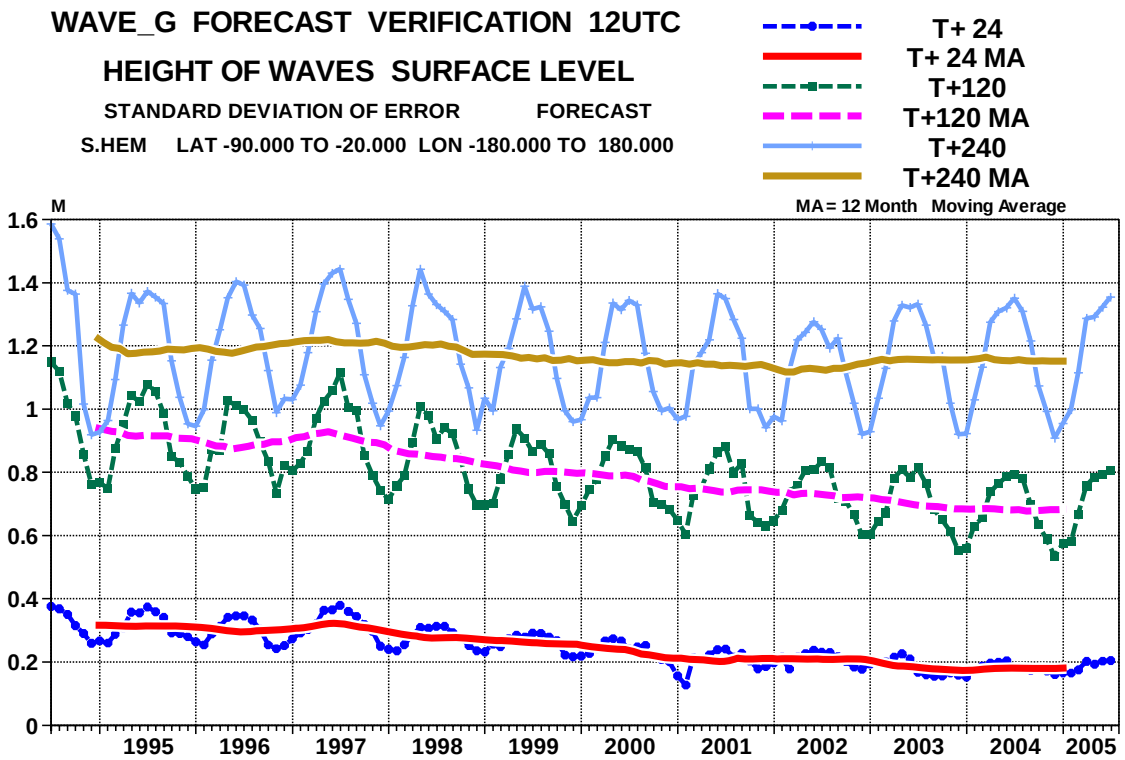


Figure 26: Scores (anomaly correlation and standard deviation) of oceanic wave heights verified against the analysis (Southern Extratropics)

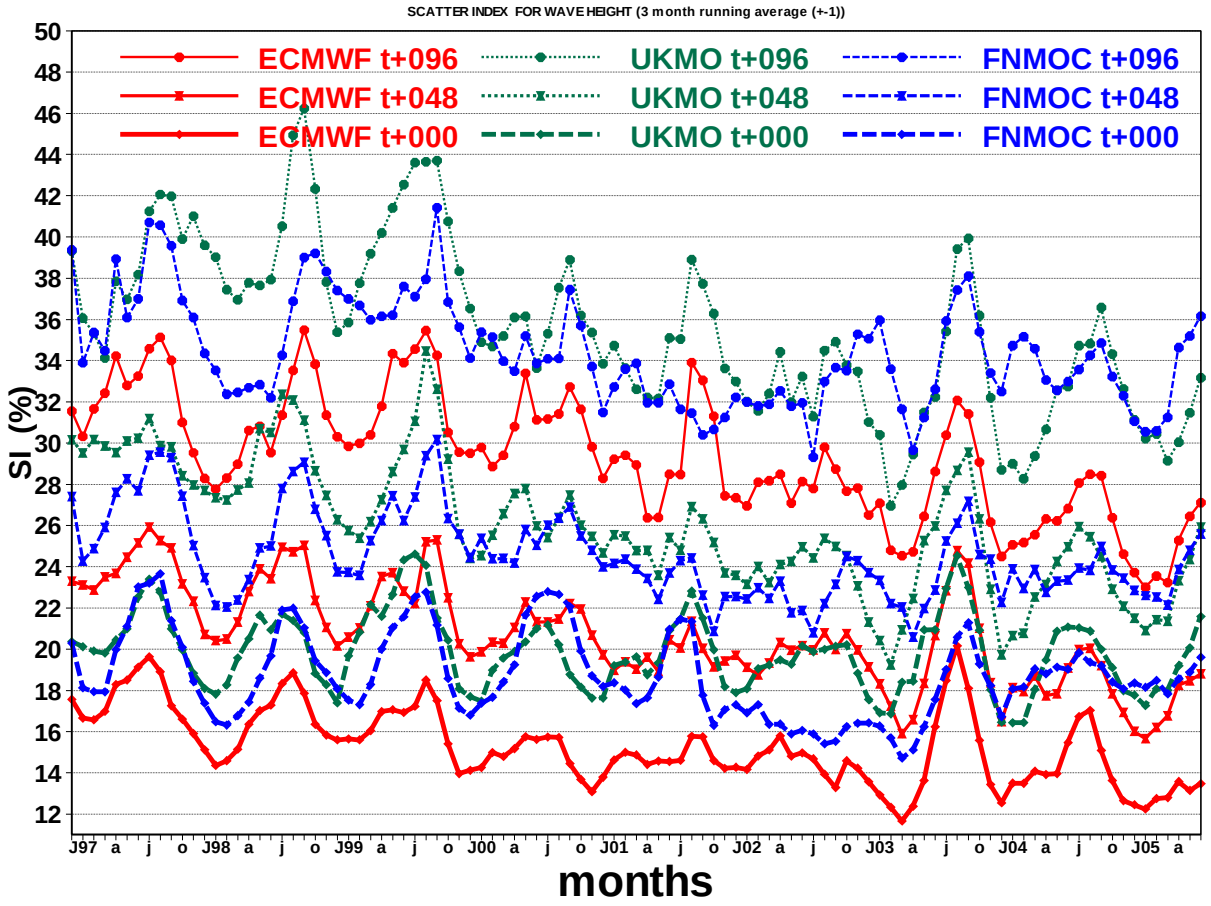


Figure 27: Verification of different model wave height forecasts using an independent set of observations from 38 wave buoys. The scatter index (SI) is the standard deviation of errors normalised by the mean observed value.

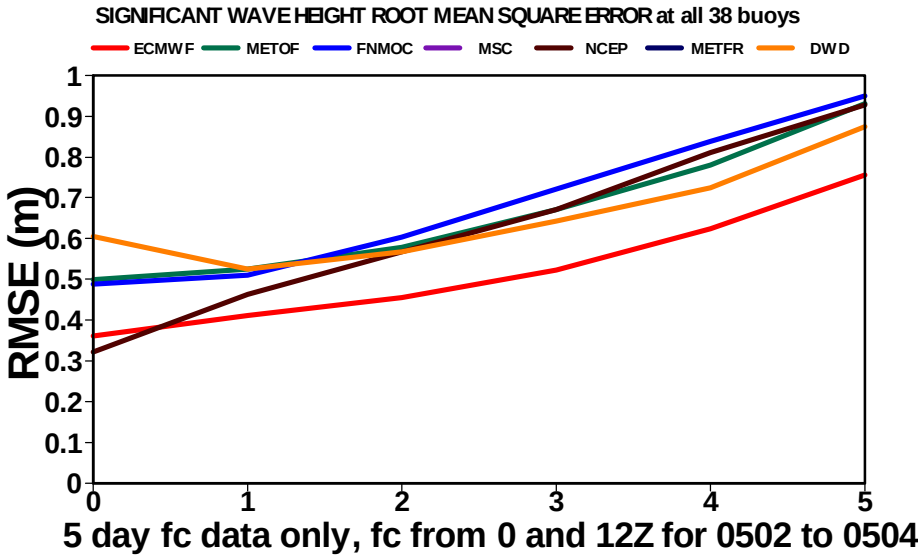


Figure 28: RMSE for the 5-day wave height forecast verified against a selection of 38 buoy observations in the northern hemisphere. The RMSE is plotted against the forecast range for the period February 2005 to April 2005. The curves refer to different forecast systems (see legend).

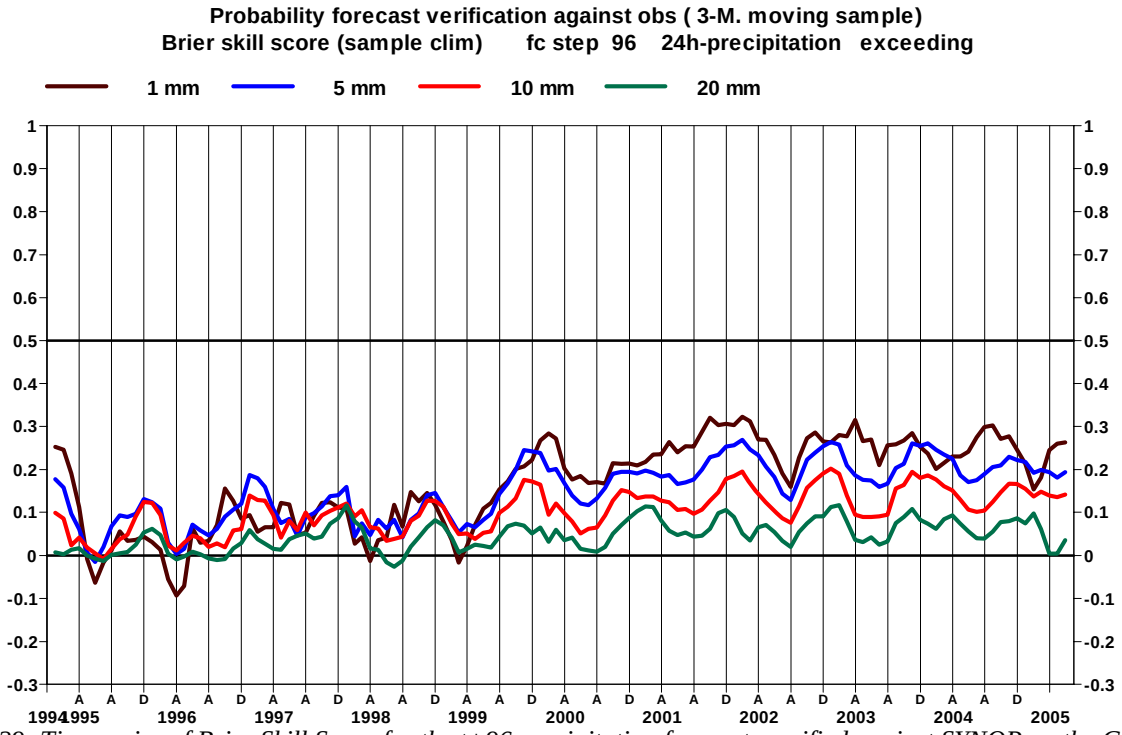


Figure 29: Time series of Brier Skill Score for the t+96 precipitation forecasts verified against SYNOP on the GTS over the European area. The skill score is calculated for three-month running periods. The reference is the sample climate.

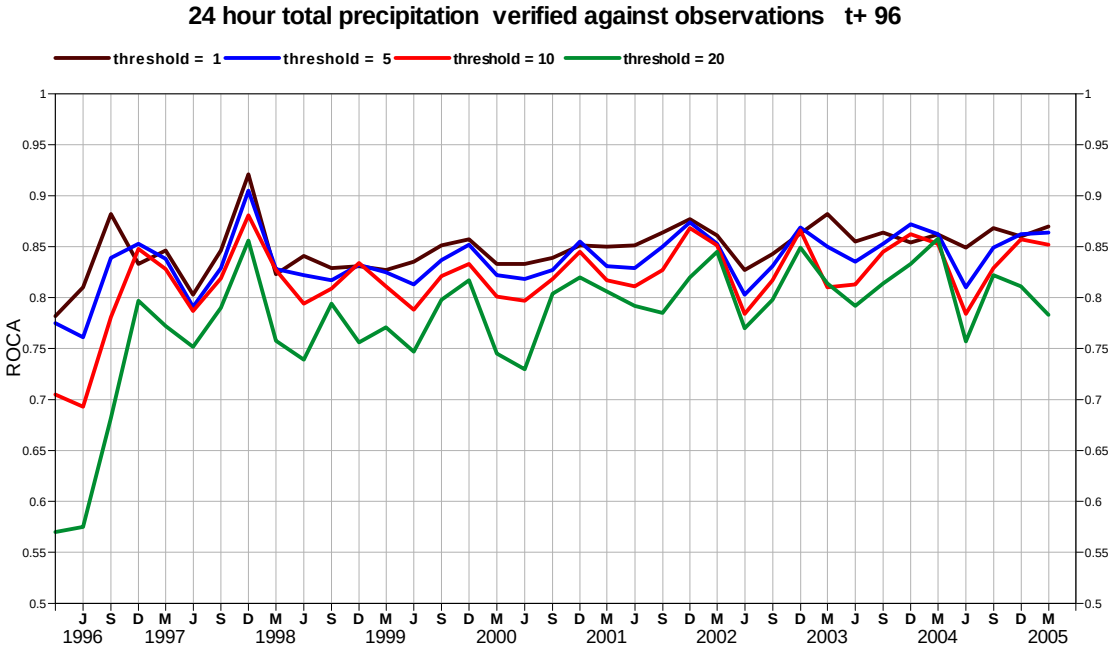


Figure 30: Time series of ROC Area for the t+96 precipitation forecasts verified against SYNOP on the GTS on the European area. A ROC area of 0.5 means no skill, a ROC Area of 1 shows a perfect system.

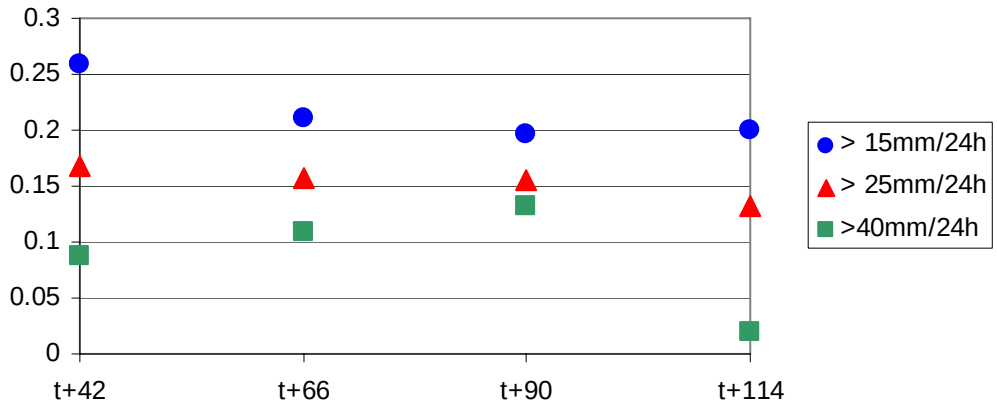


Figure 31: TSS value for a sample of extreme events which happened in the UK between 2002 and 2004. Three precipitation thresholds are shown: 15mm/24h (blue dot), 25mm/24h (red triangle) and 40 mm/24h (green square).

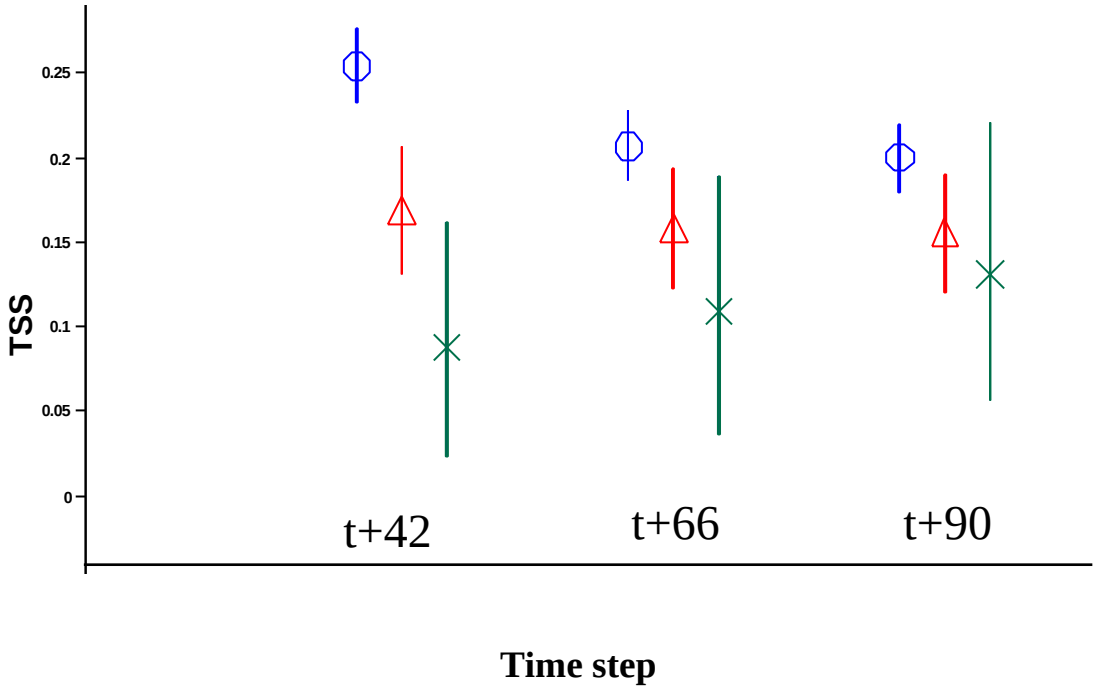


Figure 32: TSS value for a sample of extreme events which happened in the UK between 2002 and 2004. Three precipitation thresholds are shown: 15mm/24h (blue circle), 25mm/24h (red triangle) and 40 mm/24h (green cross). Vertical line across symbols indicate confidence intervals (0.05, 0.95 percentile)

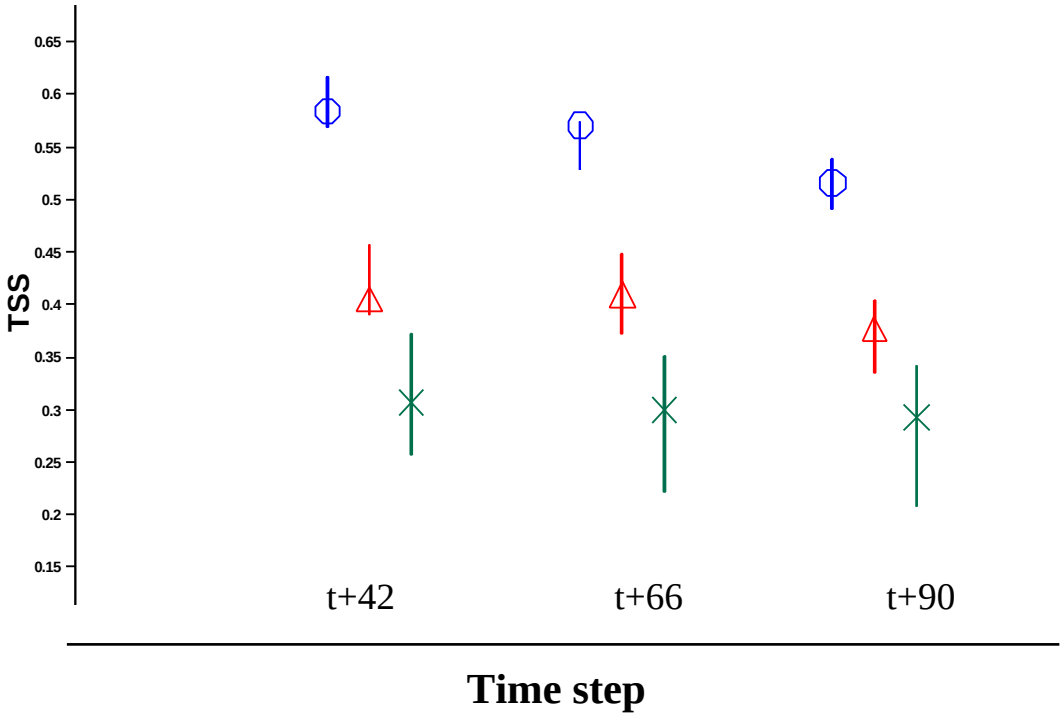


Figure 33: TSS value for a sample of extreme events happened in the UK between 2002 and 2004. The condition of gridpoint-to-gridpoint correspondence has been relaxed in these set of verifications. Three precipitation thresholds are shown: 15mm/24h (blue circle), 25mm/24h (red triangle) and 40 mm/24h (green cross). Vertical line across symbols indicate confidence intervals (0.05, 0.95 percentile)

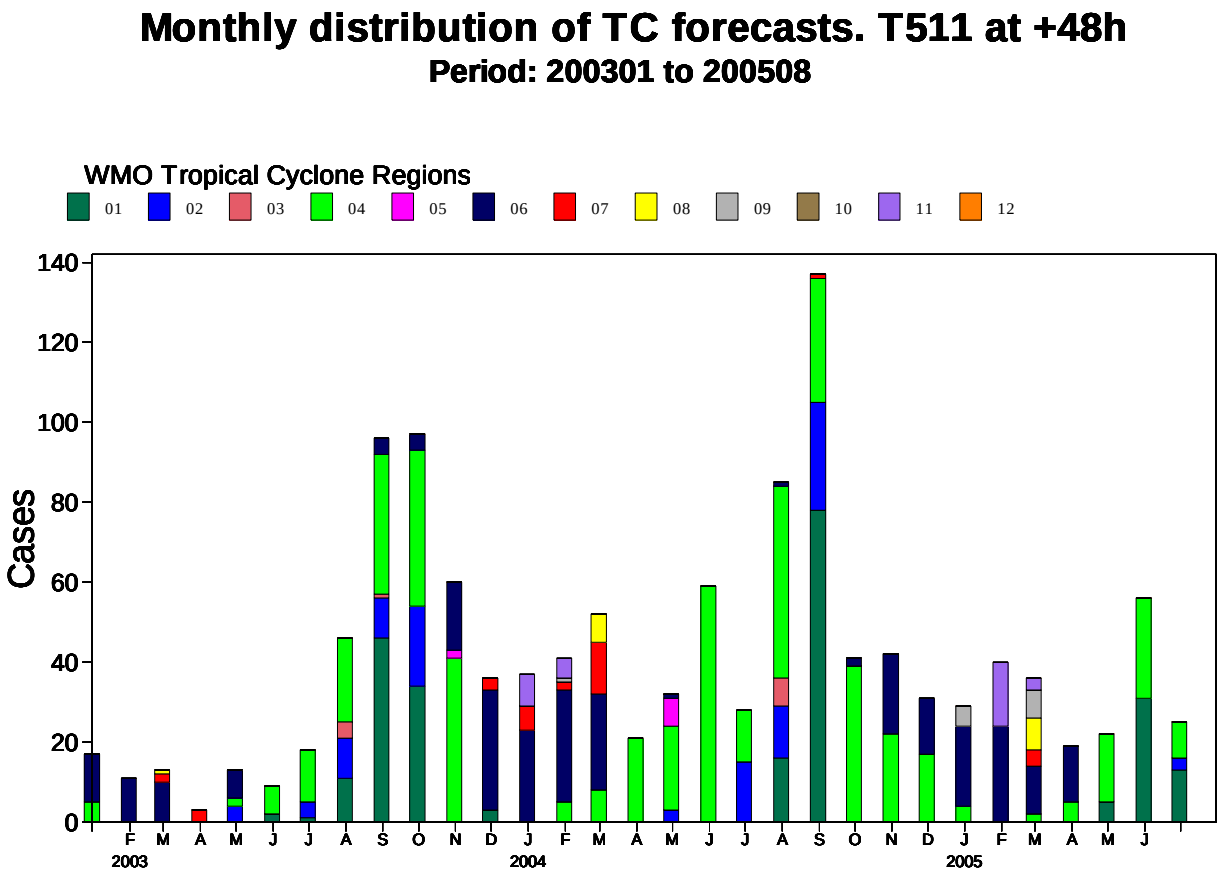


Figure 34: Number of Tropical Cyclone tracked by the determinist,T511 day 2 forecast from January 2003 to July 2005. For each month, the number is split per WMO Tropical Cyclone region (1=NW Atlantic; 2=.NE Pacific; 3=N Pacific; 4=NW Pacific; 5=N. Indian; 6= SW Indian; 7=SE Indian; 8/9/10=SW Pacific; 11/12=S. Pacific). Both 00 and 12UTC forecasts are tracked.

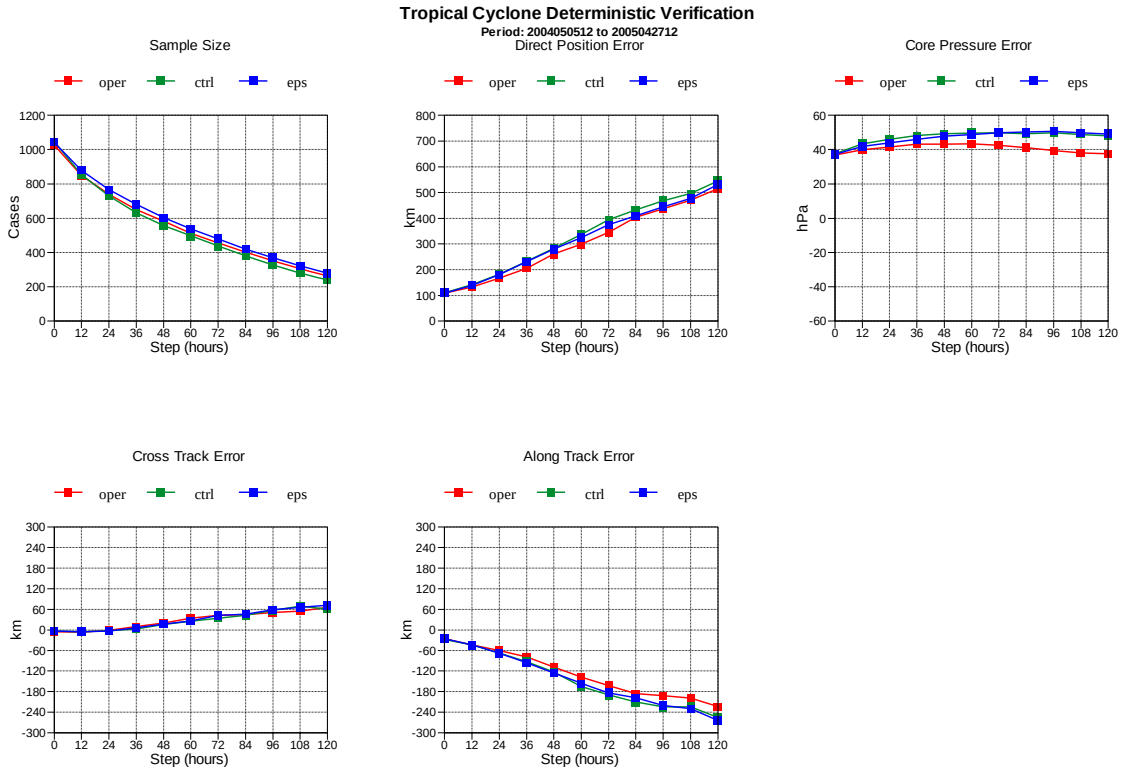


Figure 35: Verification of Tropical Cyclone forecasts from the deterministic, T511 forecast (red), EPS T255 Control (green) and mean position/ intensity averaged among all cyclones tracked in each member of the ensemble forecast (blue) for the period May 2004 to April 2005.

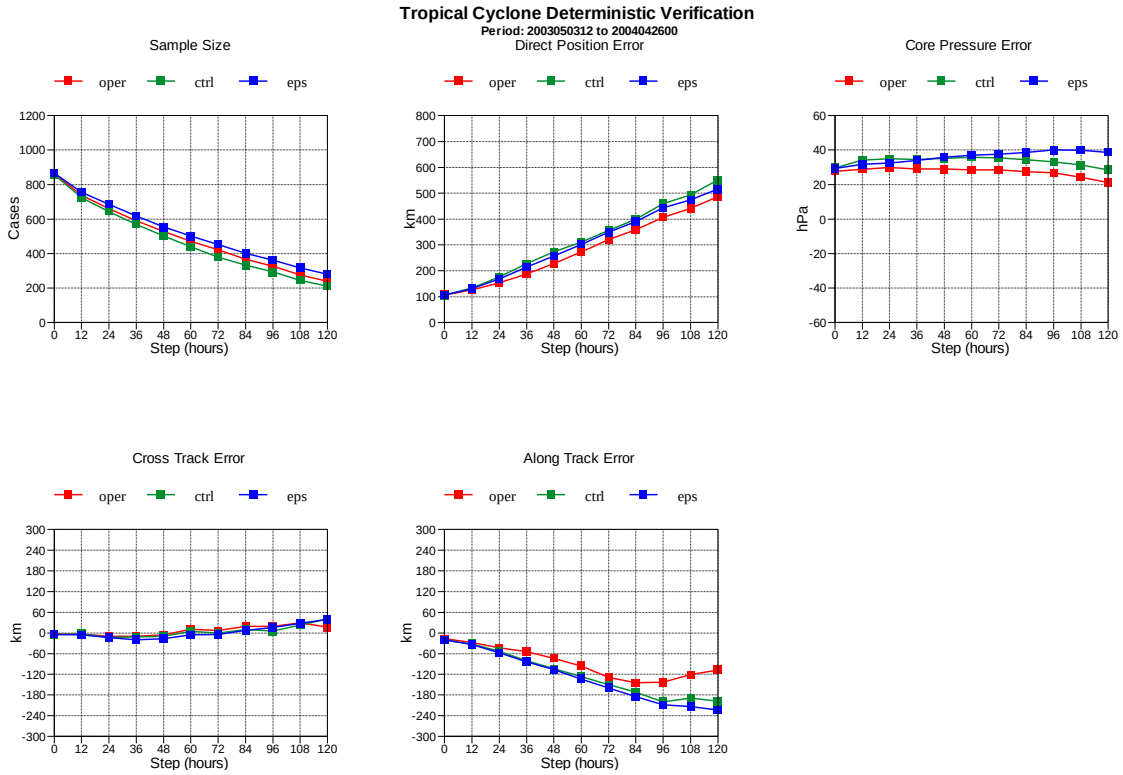


Figure 36: Same as Figure 35, but for the same period in 2003-2004.

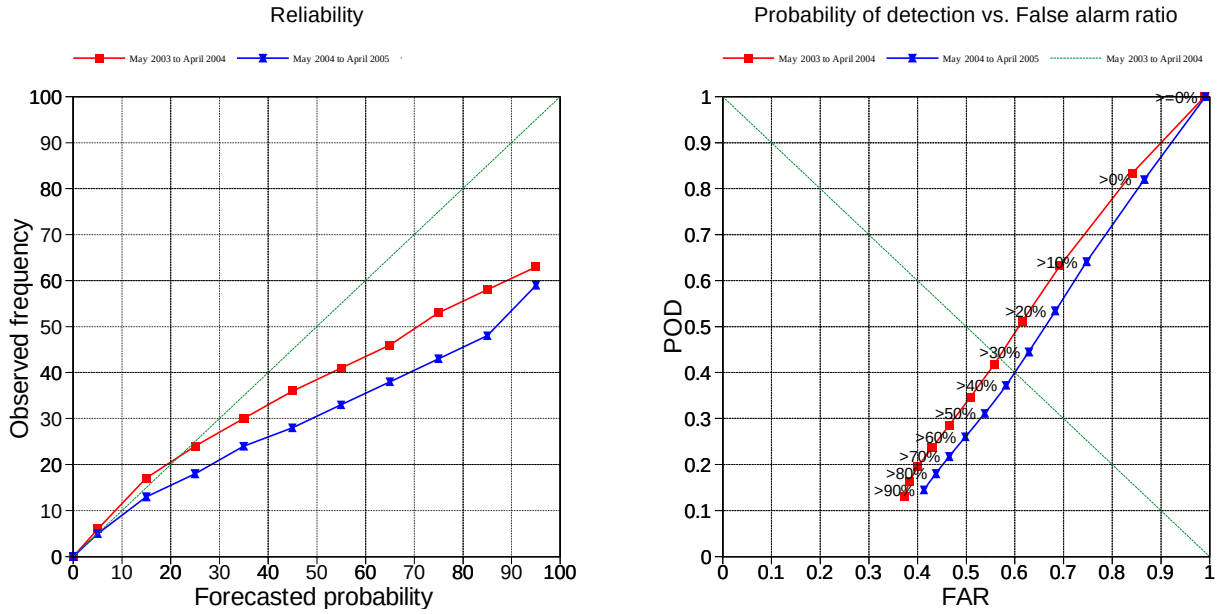


Figure 37: Probabilistic verification of TC strike probabilities for the 12-month period May-April in 2004-05 (blue) and 2003-04 (red). Left: reliability diagram (the closer to the diagonal the better); right: probability of detection (H/H+M)/ False Alarm Ratio (F/F+H) diagrams (the closer to the upper left corner the better). In both cases, the different points are for different probability thresholds.

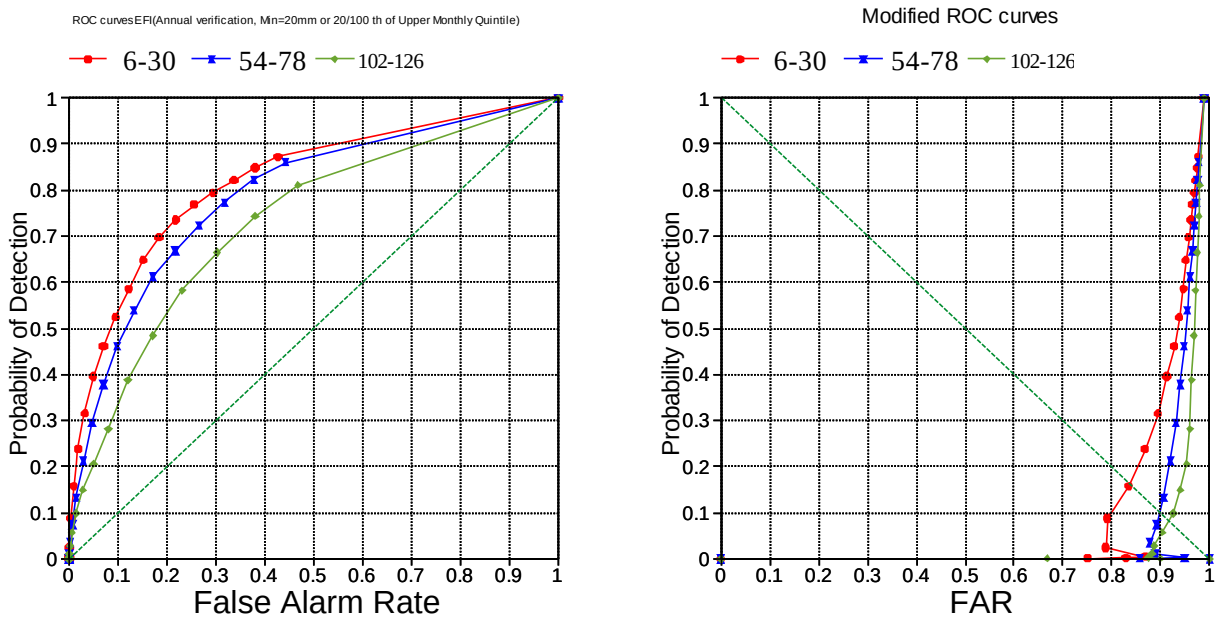


Figure 38: Verification of daily rainfall events exceeding the 99% threshold of the EUMETNET/ECSN station climatology using the EFI. Each curve is for a different forecast range, while each point of the curve is for a different EFI threshold. Sample contains 2622 events over the period October 2003 to May 2005.

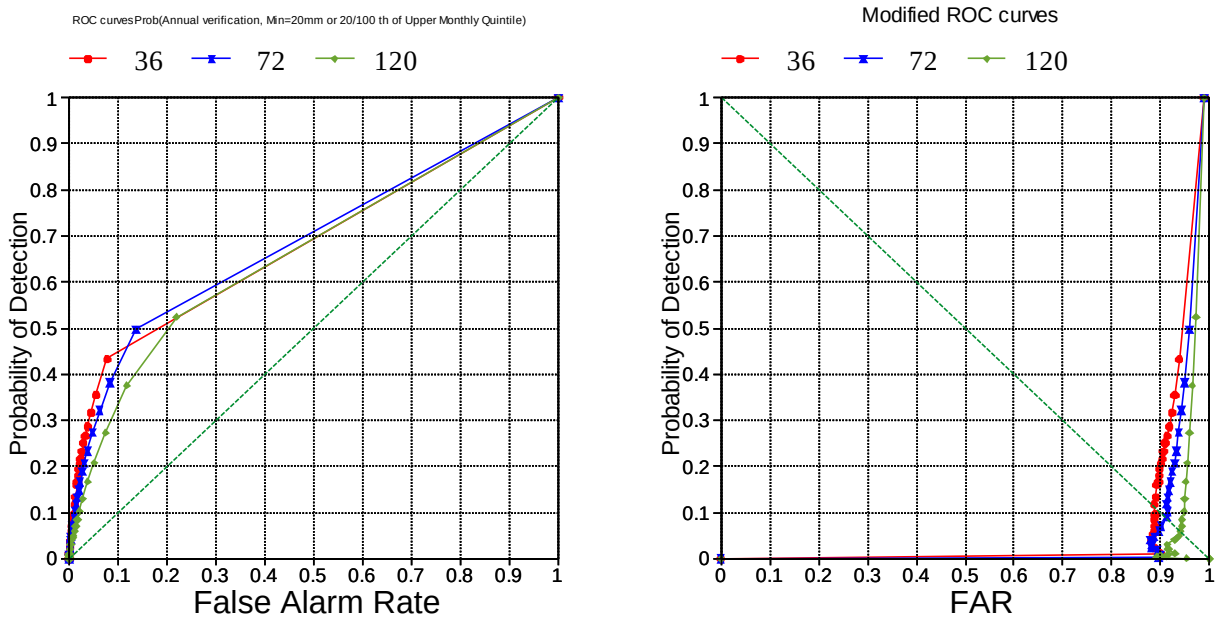
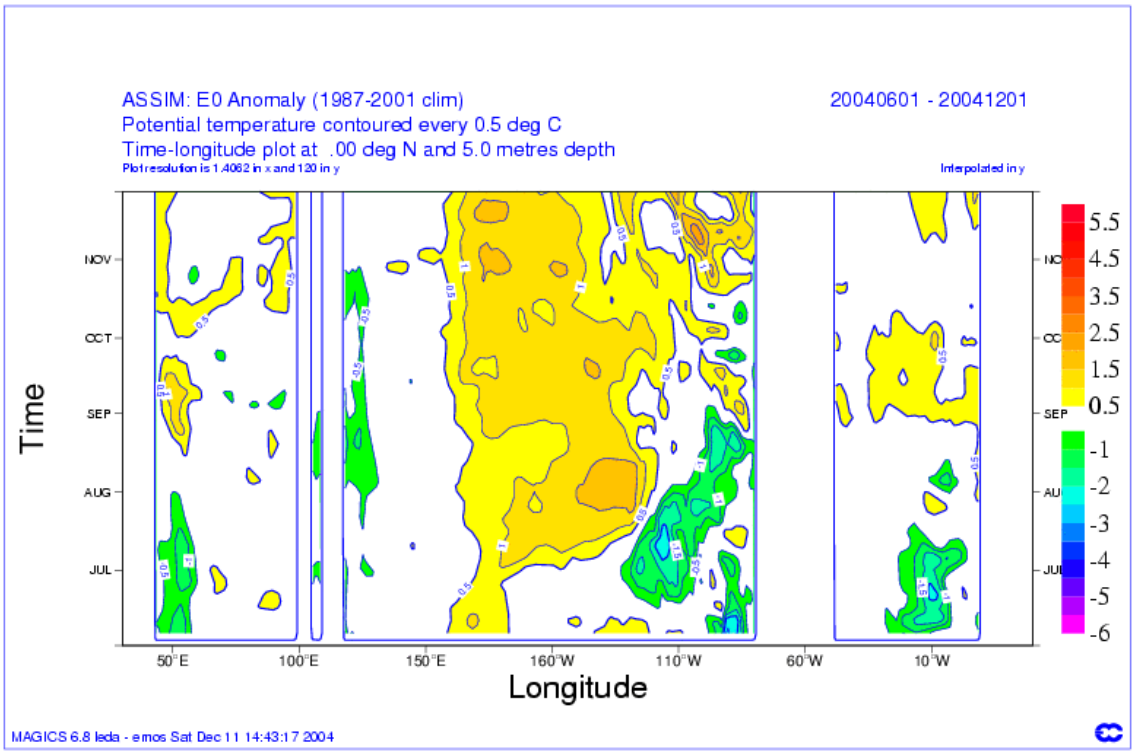


Figure 39: Same as Figure 38, but using different numbers of EPS members exceeding 20mm/day as a decision criterion to forecast the occurrence of the 99% threshold.

a)



b)

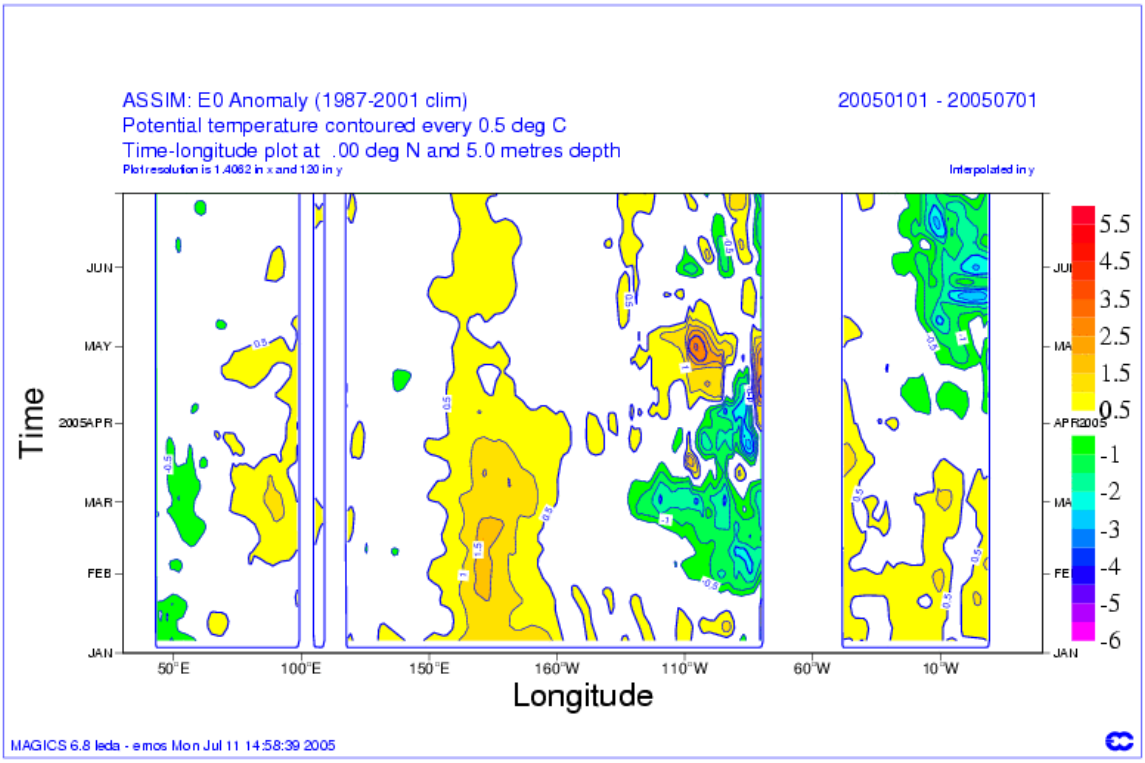


Figure 40: Time-longitude section of the sea surface temperature anomalies at the equator. Longitudes are represented by the X-axis. Time is represented by the Y-axis and it ranges: a) from June 2004 to November 2004; b) from January 2005 to July 2005.

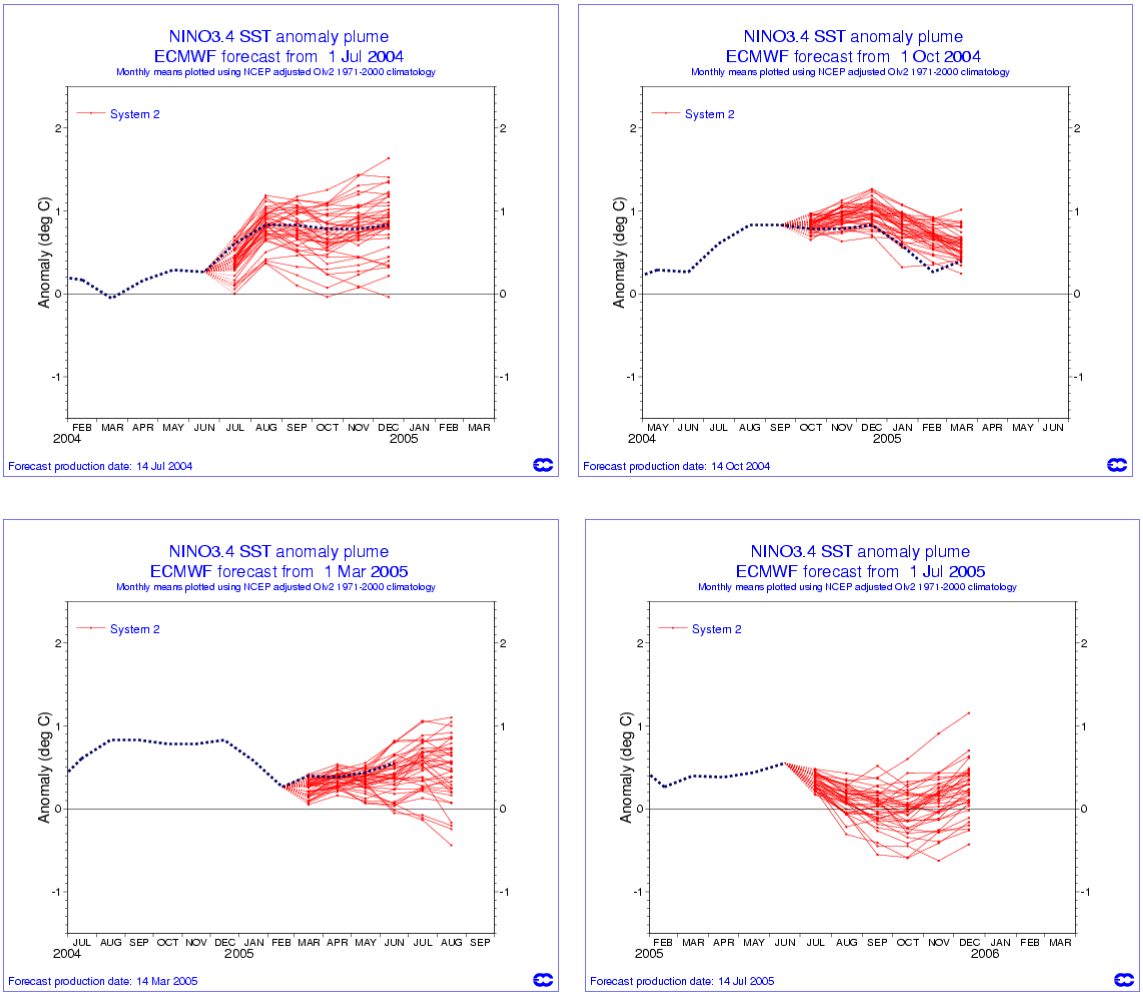


Figure 41: Plot of forecasts of Nino-3.4 at four start dates July, October 2004 March and July 2005. The red lines represent the 40 ensemble members. The heavy dashed line represents subsequent verification.

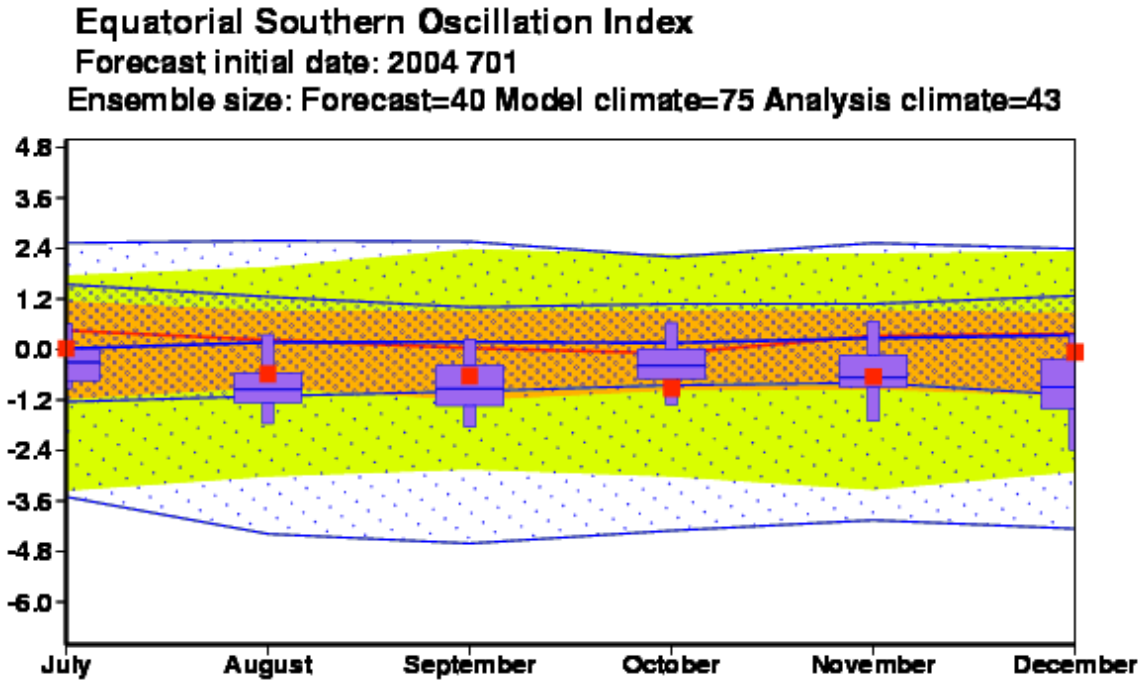


Figure 42: Equatorial Southern Oscillation prediction from the ensemble started in July 2004. The distribution of predicted monthly mean anomalies is represented by boxes (25-75 percentiles), median represented by the blue line and whiskers (5-95 percentiles). Red squares represent the observed anomalies for 2004. The distribution in the ERA-40 climate is represented by the orange (25-75 percentiles) and green (5-95 percentiles) bands; median is the red line. The model climate distribution is represented by dotted bands with higher and lower density for 25-75 and 5-95 percentiles, respectively; median is the blue tick line.

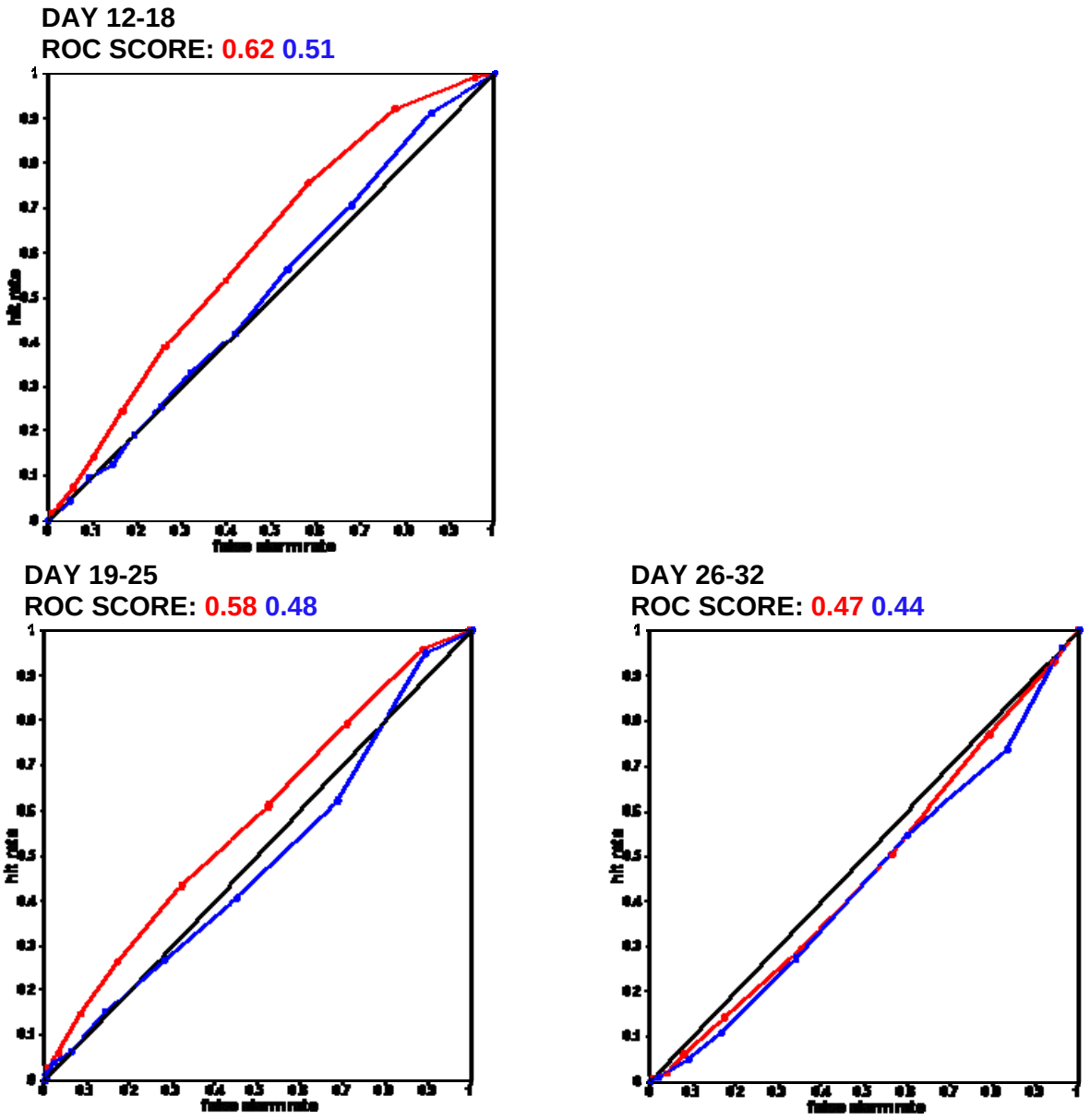


Figure 43: ROC curves verifying the probabilities from the monthly prediction system that the weekly mean of precipitation is in the upper tercile over India. The sample comprises 27 cases during the May to August period in the years 2003 to 2004. Forecast verification is shown in red, persistence corresponds to the probability predicted for the previous week and is shown in blue.