346

Verification statistics and evaluations of ECMWF forecasts in 2000-2001

F Lalaurette and L Ferranti

Operations Department

September 2001

For additional copies please contact

The Library
ECMWF
Shinfield Park
Reading, Berks RG2 9AX
library@ecmwf.int

Series: ECMWF Technical Memoranda

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

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

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

1. Introduction

This document summarises the recent changes to the data assimilation/forecasting system (Section 2). Verification results of the deterministic forecast produced at ECMWF are then presented for the free atmosphere, with additional information on the day to day consistency of the forecasts. A large part of this Section 3 is devoted to a comprehensive comparison with other centres providing medium range forecasts. Weather parameters and oceanic wave forecasts are also evaluated. Finally, section 4 deals with the verification of ECMWF ensemble forecasts (both into the medium range - EPS - and the seasonal range) while section 5 is a short technical note describing the scores used in this report.

The set of verification scores selected is kept consistent with previous reports, in order to facilitate year-to-year comparison, as requested by TAC in 1997. New aspects related to the verification of forecasts of severe weather events are dealt with in a separate document.

Please note that since 29 September 2000, verification pages have been created on the Member State web server with the following address:

http://wms.ecmwf.int/forecasts/medium/verification.html

Scores on these pages are updated on a monthly basis.

2. Changes to the data assimilation/forecasting system

The list of changes to the system over the period of reference for this report (September 2000 - August 2001) is as follows:

- 12 September 2000: 12 hour 4D-Var data assimilation is implemented with model cycle 23r1; 4D-Var now processes the observations in 12-hour sets, spanning 03 UTC - 15 UTC for the 12 UTC analysis, and 15 UTC - 03 UTC for the 00 UTC analysis. Surface analyses still run every 6 hours. Analysis fields are still archived every 6 hours. Other changes introduced with cycle 23r1 are:
    - Use of more accurate background trajectory in 4D-Var thanks to an improved interpolation procedure, use of the prognostic cloud scheme and first-guess cloud fields.
    - Production of cloud analyses from 4D-Var (the former practice was to start forecasts with clouds from the previous forecast).
    - Change to the 4D-Var incremental formulation by which the low-resolution increment is added to the high- resolution trajectory at analysis time (00 UTC and 12 UTC), instead of the start of the 4D-Var window.
    - New quality control step that prevents the use of observations which the incremental formulation of 4D-Var cannot handle correctly.
    - Resetting of the stratospheric ozone and switching off of the multivariate coupling between ozone and vorticity.
    - Monitoring of TOVS radiances in cloudy areas.
- 8 November 2000: ATOVS data from NOAA-16 are activated.
- 21 November 2000: the new higher resolution forecasting system is implemented with model cycle 23r3. This includes:
    - a T_L511L60 (40km) resolution determinist model used for 4D-Var outer loops and 10-day deterministic forecast to replace the T_L319L60 (60km) version; the oceanic wave component of the model continues

to run at approximately 55 km horizontal resolution with an increase in spectral information from 12 to 24 directions and from 25 to 30 frequencies; the additional limited area version (European waters model) continues to run at approximately 28 km with an increase in spectral information from 25 to 30 frequencies, while the number of directions remains at 24.

  - a T_L159L60 (120km) resolution tangent linear and adjoint codes used for 4D-var incremental minimizations (inner loops) to replace the T63L60 version;
  - a T_L255L40 (80km) resolution 51-member ensemble forecast model (EPS) to replace the T_L159L40 (120km) version; the wave model in the EPS runs in shallow water mode with an increased horizontal resolution at approximately 110 km with 12 directions and 25 frequencies (no change).

- 19 December 2000: bug fix applied to the stochastic physics formulation of the EPS; the bug was introduced with the new physics-dynamics coupling (13 July 1999, cycle 21r2);
- 6 February 2001: EPS perturbations are rescaled; this is to correct the underestimation of their amplitude that followed the re-tuning of data assimilation statistics introduced on 27 June 2000 (cycle 22r3);
- 28 March 2001: 00UTC-EPS starts running daily as a test system aimed at improving the forecasts of severe weather events in the early medium-range;
- 12 June 2001: several changes in the use of satellite data over land and sea ice (more AMSU data used over land and sea ice, use of skin temperature rather than upper-layer soil temperature in processing radiances); in addition, radiation parametrization routines are called on an hourly (instead of 3-hourly) basis during the data assimilation; the time integration of the surface skin temperature over land has been improved; the horizontal diffusion on vorticity has been increased to bring it in line with that applied to divergence and temperature without any noticeable impact on the forecast fields; an improved ozone model developed by Météo-France has been introduced with a new parametrization of the destruction of ozone by heterogeneous chemical processes.

3. Verification of the deterministic (reference) model

3.1 Verification for the free atmosphere

3.1.1 ECMWF scores

Figure 1 gives the evolution of forecast skill as measured by the 12-month moving average from 1980 to 2000 of the 500 hPa height error normalised by persistence, for the Northern Hemisphere and the European area. The last month included in the statistics is July 2001. Monthly and annual mean values of forecast ranges beyond 60% anomaly correlation can be found in Figure 2. These figures show that the rate of progress, already quite high in 1999-2000, has kept its pace this year: scores have reached unprecedented values in 12-month averages both over the extratropical Northern and Southern Hemispheres, while over Europe the very high performance achieved in 1999-2000 has been consolidated. It can be seen that this has been achieved among other factors by an improved performance in spring/summer compared to the last two years. Figure 3 pictures the distribution of errors for 1000-hPa surfaces (which can be considered as very similar to pressure fields at sea level) this winter compared to the previous two. They show a very clear improvement over the first four forecast days: nearly 60% of last winter forecasts of pressure anomalies correlated better than 90% with the verifying analysis over Europe at the 96h range. Subsequent ranges show slightly less good forecasts than last year, but are still better than the year before (1998-1999).

As usual, it is difficult to isolate in these scores the contribution made by intrinsic model performance from that of flow dependent predictability. Based on the assumption that unpredictability should affect all numerical models equally, ECMWF model daily errors have been substracted from a “reference” taken as the mean error of a selection of global weather forecast models (USA, Germany, UK and Canada) for the same day. Results are shown in Figure 4 for Europe (Day 3 and Day 5 forecast range) both in 2000-2001 and 1999-2000. The poor performance of the ECMWF model over Europe in August 1999 is clearly seen in the lower panel red curve, a signature that was not reproduced in August 2000. Moreover the average “gain” accumulated over 12 months is significantly larger this year, the gap widening between day 3 and day 5 while it was saturating by day 3 in 2000. It is noteworthy that the introduction of 12-h 4Dvar in September 2000 more or less coincides with the steep accumulation of “gains” by day 5, while the high resolution in November 2000 seems to be detectable in the day 3 curve. More comparisons with other models will be detailed in Section 3.1.2 below.

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 5. It indicates that at both lower and upper-tropospheric levels, the performance has also improved this year in these tropical areas to produce the lowest average error values ever reached by ECMWF model.

In the same Figure 5, information is provided regarding the quality of the stratospheric forecast (50-hPa level). The trend to improvement is very strong at these levels, most notably since the increase in vertical resolution introduced in March 1999 (cy19r2).

One of the concerns expressed by some users prior to the horizontal resolution change was related to the risk of the model becoming overactive and inconsistent from one day to the next. It can be seen on Figure 6 that this has certainly not been the case following the introduction of TL511 this autumn. Indeed the RMS difference between consecutive forecasts valid for the same day is falling sharply for the third consecutive year, showing a tendency to provide much more consistent forecasts than was the case some years ago.

3.1.2 Comparison with other centres

The basic common ground for such an intercomparison is the regular exchange of scores between GDPS centres under WMO/CBS auspices following agreed standards of verification. Figure 7 shows time series of such scores over Northern Extratropics for both 500-hPa height and Mean Sea Level Pressure. It can be seen that the reduction of ECMWF errors observed at all ranges this winter is not a feature that has been common to all models, which confirms that it has to be accredited to improvements of the forecasting system and not to a less active winter than usual. The gap from other models has now reached an amplitude by day 6 that was not seen before, at least not since the exchange of scores was set up in 1992. Although this improved performance is not found in summer, when there is not much difference between ECMWF and other centres, the annual means shows a substantial net benefit (Figure 8). The difference is even larger in the Southern Extratropics, where it is can be seen throughout the year, including the warm season (Figures 9 and 10).

The situation in the Tropics is summarised in Figures 11 and 12. The striking feature there is that both UKMO and ECMWF 5-day forecasts of low level (850-hPa) winds are notably better than any other centre’s product. At upper levels, ECMWF forecasts are also now of a quality similar to those of the UKMO, which was not the case two years ago and must be attributed to the series of model changes (particularly in the convective scheme) and data assimilation improvements in these tropical areas.

Finally scores are presented using radiosondes as the verifying dataset, both in Europe and in the Tropics (Figure 13). They confirm the conclusions drawn previously from the field verification against each model's own analysis, although it can be seen that ECMWF low level forecasts in the Tropics compare better to the radiosondes.

3.2 Weather parameters

Fig. 14 shows the monthly mean and standard deviation of the 2m temperature and specific humidity errors over Europe up to July 2000, verified against synoptic observations (a correction for the difference between model orography and real orography was applied to the temperature forecast error). Although the trend is mainly neutral, it should be noted that the standard deviation of nocturnal 2m-temperature has reached an all time low this spring/early summer (only slightly above 2K). Several Member States have commented positively on these improved 2m-temperature forecasts in the "2001 Report on Verification of ECMWF products in Member States and Co-operating States" (Green Book). The improvement seems to have been seen most clearly over Scandinavia and is associated with the improved land-surface parametrization scheme introduced in June 2000 (cy 22r3). On the negative side, a small warm bias shows up in daytime that was not there previously and is slightly stronger over N. America and East Asia (not shown). This is currently being investigated. In the same Figure 14, the specific humidity biases show that the lack of moisture in daytime is still there, although it has been reduced over the last two summers. Figure 15 shows the same comparison for total cloud cover and 10m wind speed forecasts. The reduction of errors in total cloud cover is confirmed again this year - and also has been commented on by several Member States in the Green Book. The reduction of the negative daytime wind speed biases that was noticed for the first time last year, following the introduction of increased resolution of the PBL (and the improved post-processing introduced on this occasion), is confirmed. In contrast to other improvements in the representation of weather parameters by the forecast, this is not something that is usually perceived by the users and is likely to be due to the large representativity limitations of model winds (biases are only a very small component of the forecast error).

The monthly mean error of the precipitation forecasts at day 3 over Europe is shown in Figure 16, for 00, 06, 12 and 18 UTC. The reduction in the diurnal cycle of convective precipitation errors (too early release of convective rain in spring/early summer) is confirmed, although the errors are slightly larger this year. In order to document the impact of higher resolution on the distribution of high precipitation events, a comprehensive dataset collected from the GTS (1531 Northern Extratropical SYNOP stations totalling 127000 daily observations) has been analysed for winter 2000/2001 and compared to similar data in winter 1999/2000. Results are summarised in Figure 17. It can be seen that although there is a large representativity difference between model grid forecasts and SYNOP observations, the higher resolution model run this winter brings the distribution of large precipitation events (black curve) much closer to observations (red curve) than was the case last year (magenta and green curves). Frequency Bias Indices are also shown for the two seasons in Figure 17. They are both decreasing sharply towards zero for large precipitation amounts due to the model not generating at the grid scale the amounts of precipitation which are observed locally[1]. However, the increase in resolution is found to have brought the index significantly closer to one for a wide range of precipitation events, indicating that the gain in resolution has been adequately transferred into the precipitation distributions.

[1] This was the very reason why additional datasets for generating super observations at the model resolution were requested in February this year, following both SAC and TAC recommendations in 2000.

3.3 Oceanic waves

Verification scores from the global oceanic wave products are shown in Fig. 18 and 19. Northern Extratropical scores show that the good level of performance previously reached has been maintained, with some reduction in the standard deviation of wave height errors.

4. Verification of ensemble forecasts

4.1 Ensemble Prediction System (EPS)

Although the performance of the new TL255L40 EPS was evaluated very positively following extensive pre-operational testing in 2000, the level of performance achieved in operations this winter has not met these expectations. This has been the subject of extensive investigations during the course of the winter. Indeed, two bugs were corrected on 19 December and 6 February (see section 2 above). The latter was probably the one affecting performance most dramatically, as it resulted in an underestimation of the EPS spread by roughly 30% by day 6, together with a large proportion of days when the clustering and tubing classification algorithms provided no clear alternative to the deterministic forecast (only one central cluster). That the rescaling of initial perturbations in February kept its impact into the medium range can be seen from the time series of ensemble spread over Europe at Day 6 (Figure 20): clearly the spread level has been brought to a level much closer to the error after the rescaling happened on 5 February (blue curve) than before (magenta). As expected, the average error also correlates better with different categories of forecast spread (Figure 20, lower panel). Only results following the rescaling of EPS perturbations are shown hereafter.

Reliability diagrams for the verification of 850 hPa temperature anomalies and 24-h accumulated precipitation over Europe are shown in Fig. 21 and 22. The reliability curves for the same periods last year are also reported for reference. It can be seen that the reliability has improved significantly since last year for the EPS forecast of temperature anomalies. The only exception is for very warm (>8K) anomalies. The diagrams for precipitation (Fig. 22) do not show the same improvement. This is partly related to the reference used in this case, which is the 0-24h precipitation field from the Control. The reference therefore is at higher resolution in 2001 than in 2000, which is likely to offset the expected benefit in terms of forecast performance. It should be noted at this point that new verification procedures for precipitation are being set up which will rely on high resolution datasets provided by Member States and Co-operating states. Gridded precipitation fluxes will be produced at the model scale which will provide a model-independent reference for verification.

The quality of any forecast system is its resolution, that is, its capability to forecast whether an event will occur or not. In the case of a probabilistic system, this can be quantified by how different from its climatological value the frequency of occurrence of an event is, when the forecast probability falls in a given category. This is known as the Resolution component of the Brier Skill Score and ranges from zero (climatology forecast) to one (perfect deterministic forecast). The variations of this score with forecast range and for different events are depicted in Figure 23 (850-hPa temperature anomalies) and Figure 24 (24h-accumulated precipitation, verified using SYNOP observations). For all of these events, the improvement achieved in 2001 with respect to 2000 is quite remarkable and confirm the results gathered from research experimentation on the T255 EPS configuration in 2000.

The time series of the (Total) Brier Skill Scores are shown in Figures 25 to 27. As expected they usually do not confirm the very high level of performance reached by the EPS system in winter 1999-2000, notably in terms of precipitation forecast. This is expected to a large extent to be the result of the two bugs that affected the EPS system in 2000/2001; it can be seen that, for most events, the most recent scores again show a positive trend compared to previous years.

One of the important projects initiated this year is the 00UTC EPS run that was started to investigate configurations of the forecast system which can add value to the forecast of severe weather events in the early medium-range. A preliminary result taken from the ongoing evaluation of this project is shown in Figure 28. Not surprisingly, it shows that the 00UTC EPS run has a 12h gain in predictability over the (12h older) 12UTC run. With the large drop in predictability observed for severe weather events in the 96-120h forecast range, this 12h gain may prove crucial. On this same Figure 28, it is also shown that for reasonably extreme events (in this case, more than 20mm of rain in 24h), it can prove useful to enlarge the ensemble size by keeping some memory of the previous forecast, which has been achieved in this case by giving 25% weight to 12h-old EPS members when computing probabilities.

4.2 Seasonal forecasts

The development of seasonal forecast verification has continued to progress. Grid point correlation and an extensive set of rainfall indices have been computed for all seasons. The seasonal predictions of inter-annual variations of the Northern hemisphere weather regimes, such as Pacific North American pattern and North Atlantic, have been analysed and compared with the performance of uncoupled atmopsheric simulations forced by observed sea surface temperatures. A large part of this validation material will be made available on the web within the next few months. Statistics available on the web since early 2000 (http://wms.ecmwf.int/ecmwf/seasonal/verstats/nino) indicate a rather high skill for seasonal predictions of tropical Pacific Sea Surface Temperature (SST) anomalies. Although the statistics are based on a relatively short time period dominated by the single large El Niño event of 1997/98 and may therefore not be highly representative of other periods, they indicate that the skill is substantial and is also well above the skill of persistence.

El Niño forecasts, issued monthly on the web site (http://wms.ecmwf.int/ecmwf/seasonal/forecast) are represented in the form of SST anomalies averaged over the Nino-3 region (5N-5S, 150W-90W).

Nino-3 predictions for spring 2000 up to summer 2001 are shown in Fig.29. Forecasts initiated in September, November 1999 and January 2000 correctly predicted a return to normal conditions after a period of la Niña conditions. Predictions initiated in March and May 2000, indicating that the spring warming would not lead to El Niño conditions but rather to neutral conditions, were also successful. Forecasts started in summer 2000 presented weak positive SST anomalies, whereas weak negative anomalies were observed. By the end of 2000 and at the beginning of 2001, predictions suggested a possible development of El Niño conditions for the following summer, which was not observed. The indication of a SST warming had a certain level of uncertainty considering the spread of the ensemble and that the signal was found in the later part of the forecast.

Fluctuations in ocean temperatures during El Niño and La Niña are accompanied by even larger-scale fluctuations in surface pressure known as the Southern Oscillation. The Southern Oscillation is a see-saw in surface pressure between the Pacific and Indian Oceans. The Equatorial Southern Oscillation Index (EQ SOI) is used to measure the differences in surface pressure between a region located over the eastern Pacific (5N-5S, 80W-130W) and a region located over Indonesia (5N-5S, 90E-140E). Fig. 30 shows EQ SOI predictions three months ahead. The difference between the surface pressure monthly anomalies is normalized. Monthly anomalies are departures for the 1990-1996 base period means and are normalized by their standard deviation. A three-month running mean is applied to the EQ SOI values.

Historically, there is considerable variability in the ENSO (El Niño / Southern Oscillation) cycle from one decade to the next. The 1990s featured a very active ENSO cycle, with three El Niño episodes (1991-1993, 1994/95, and 1997/98) and two La Niña episodes (1995/96, 1999/2000). This period also featured one of the strongest El Niño episodes of the century (1997/98), as well as two consecutive periods of El Niño conditions during 1991-1995 without an intervening cold episode. SOI seasonal predictions (3 months ahead) are quite close to the analysis for the period 1991-1996. The sudden reduction in the low level easterly flow related to the sudden warming of the sea surface temperature over the East equatorial Pacific during the first part of 1997 was predicted with some delay and its maximum amplitude was underestimated. However the transition from El Niño to La Niña was well captured. During the second half of 1998 the intensity of the Walker circulation was over-forecasted.

The seasonal forecast skill is compared with the skill that would have been obtained by a simple persistence of the observed anomaly. Persistence is generally hard to beat in the short range (less then 3 months), but forecast schemes should be better than persistence in the 3-6 month range and beyond. Correlation between forecast and analysis (0.84) is higher than the correlation between persistence and the analysis (0.77) indicating that SOI forecasts 3 months ahead have a higher predictive skill than persistence.

5. A short note on scores used in this report

At its 26th session (September 1998), the Technical Advisory Committee had, amongst its recommendations, that a short note on the verification scores used in the report should be attached. Notes hereafter follow the same layout as the report itself.

5.1 Verification of the deterministic (reference) model

5.1.1 Verification for the free atmosphere

The verifications used follow WMO/CBS recommendations as closely as possible (last update adopted by CBS-Ext(98) is available as http://www.wmo.ch/web/www/reports/CBS-Ext98.pdf). 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); when other centres' scores are produced, they have been provided as part of the WMO/CBS exchange of scores among GDPS centres*;* 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 geographical average of the squared differences between the forecast and the analysis valid for the same time; when models are compared, each model uses its own analysis for verification; RMSE for winds (Figures 5, 11, 12 and 13) root the sum of the mean squared errors for the two components of the wind independently.

Skill scores (Fig. 1) are computed as the reduction of the RMSE which the model achieves with respect to persistence (forecast obtained by persisting the initial analysis over the forecast range); in mathematical terms:

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

Anomaly correlation scores are spatial correlations between the forecast anomaly and the verifying analysis anomaly; anomalies with respect to NMC climate are available at ECMWF from the start of its operational activities in the late 1970s; they show for each month the average of the ranges at which the daily forecast was dropping below 60% of anomaly correlation.

The “gain” curves (Fig. 4) are an accumulation with time of differences between the average RMSE from N global forecast (M_n) retrieved daily from the GTS and ECMWF (M_0):

$$Gain(t_0 \to t) = \sum_{t_i = t_0}^{t} \left\{ \left[\frac{1}{N} \sum_{n=1}^{N} Rmse(M_n, t_i) \right] - Rmse(M_0, t_i) \right\}$$

5.1.2 Weather parameters

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^{-1} or 15m.s^{-1} 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).

In Figure 17, the precipitation data have been collected from the full Northern Extratropics (north of 30^{o}N) both from the model (18-42h forecast accumulation) and SYNOP observations (06-06UTC). The empirical distributions have then been obtained simply by ordering the data from the smallest to the largest amount. The stepwise signature for observations in Figure 17 is simply a feature from observations that most of the time only report full integer values (e.g. 10, 11 or 12mm). The Frequency Bias Index also shown on this Figure is the ratio of the total number of forecasts to observations exceeding a given threshold.

5.1.3 Oceanic waves.

Anomalies are derived as for free atmosphere verification, with the exception that the reference climate has been computed from the model (period is 1987-1993).

5.2 Verification of ensemble forecasts

5.2.1 Generalities

Events usually defined for the verification of probabilistic forecasts are 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. In order to have an evaluation that is not affected by representativity (scale) effects, results are shown with the model analysis at the same scale as the EPS forecast. For precipitation, the accumulation by the model over 24-h following the analysis is taken as the reference. It can therefore be argued that the verification results produced are more a diagnostic of the EPS performance than a verification of the value of the forecasts from the user's point of view. Verification of the EPS with direct reference to the observations is, however, part of the quarterly report issued to the EPS Contact Points in Member States. An example of such direct verification for precipitation is shown in Figures 24 and 27.

5.2.2 Brier Score (BS)

The BS is a measure of the distance between forecast probabilities and the verifying observations (which, as any deterministic system, takes only 0 or 1 as values). For a single event, it can be written as $BS = (p-o)^2$. As any probabilistic score, however, the BS only becomes significant when results are averaged over a large sample of independent events. Then its values range from zero (perfect, deterministic forecast) to 1 (consistently wrong, deterministic forecast).

The BS can be split into the sample climate uncertainty, the forecast reliability (BS_REL), and the forecast resolution (BS_RSL):

- uncertainty varies from 0 to 0.25 and indicates how close to 50% the occurrence of the event was during the sample period (uncertainty is 0.25 when the event is split equally into occurrence and non-occurrence);
- reliability tells how close the frequencies of observed occurrences are from the forecasted probabilities (on average, when an event is forecast with probability p, it should occur with the same frequency p);
- resolution tells how informative the probabilistic forecast is; it varies from zero for a system for which all forecasted probabilities verify with the same frequency of occurrence to the sample uncertainty for a system for which the frequency of verifying occurrences takes only values 0 or 100% (such a system resolves perfectly the forecast between occurring and non-occurring events).

From these components, skill scores can be derived:

- the Brier Skill Score (BSS, Figures 25, 26 and 27) is computed by reference to the BS of a probabilistic forecast that would consistently forecast the climate distribution

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

- the Resolution Skill Score (BSS_RSL) is the ratio of resolution by uncertainty (this component is the one depicted in Figures 23 and 24);
- the Reliability Skill Score is defined as:

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

6. List of figures

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

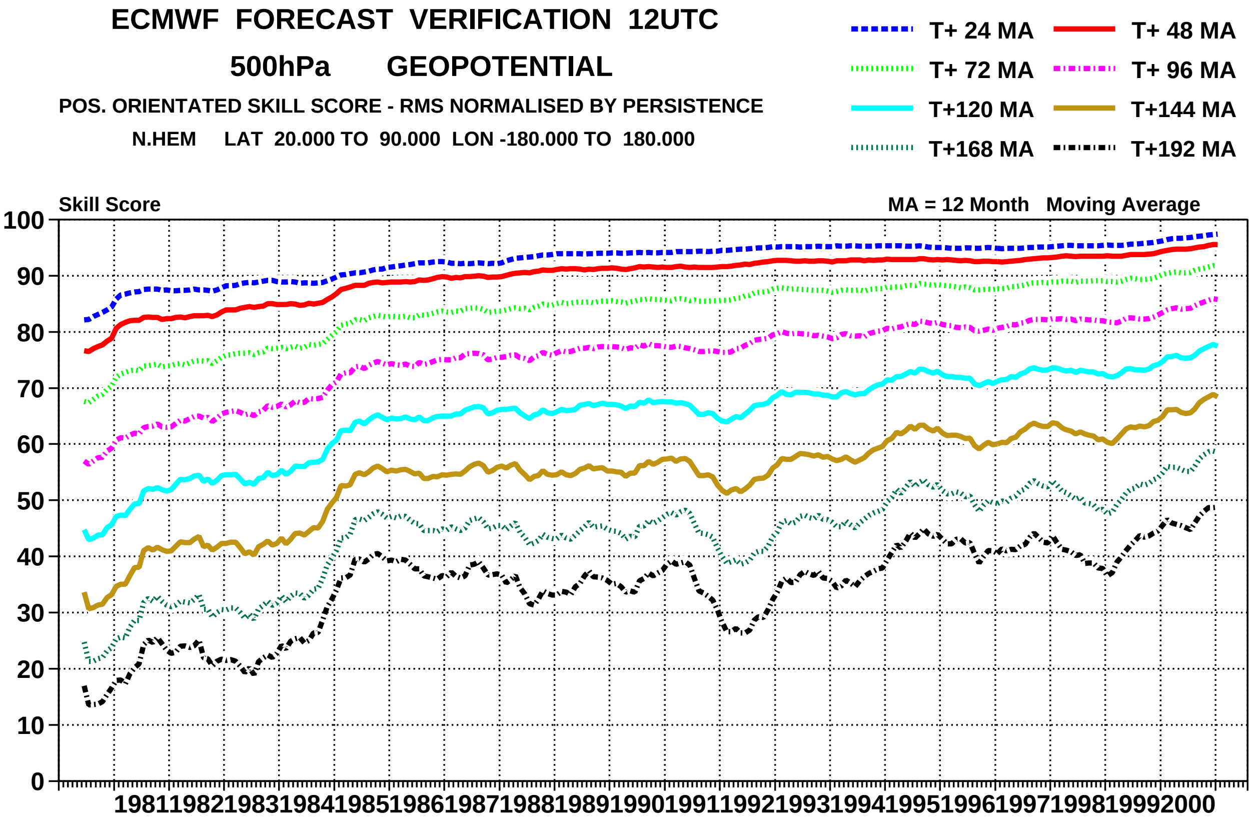


Figure 1: 500-hPa height skill score (N. Hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)

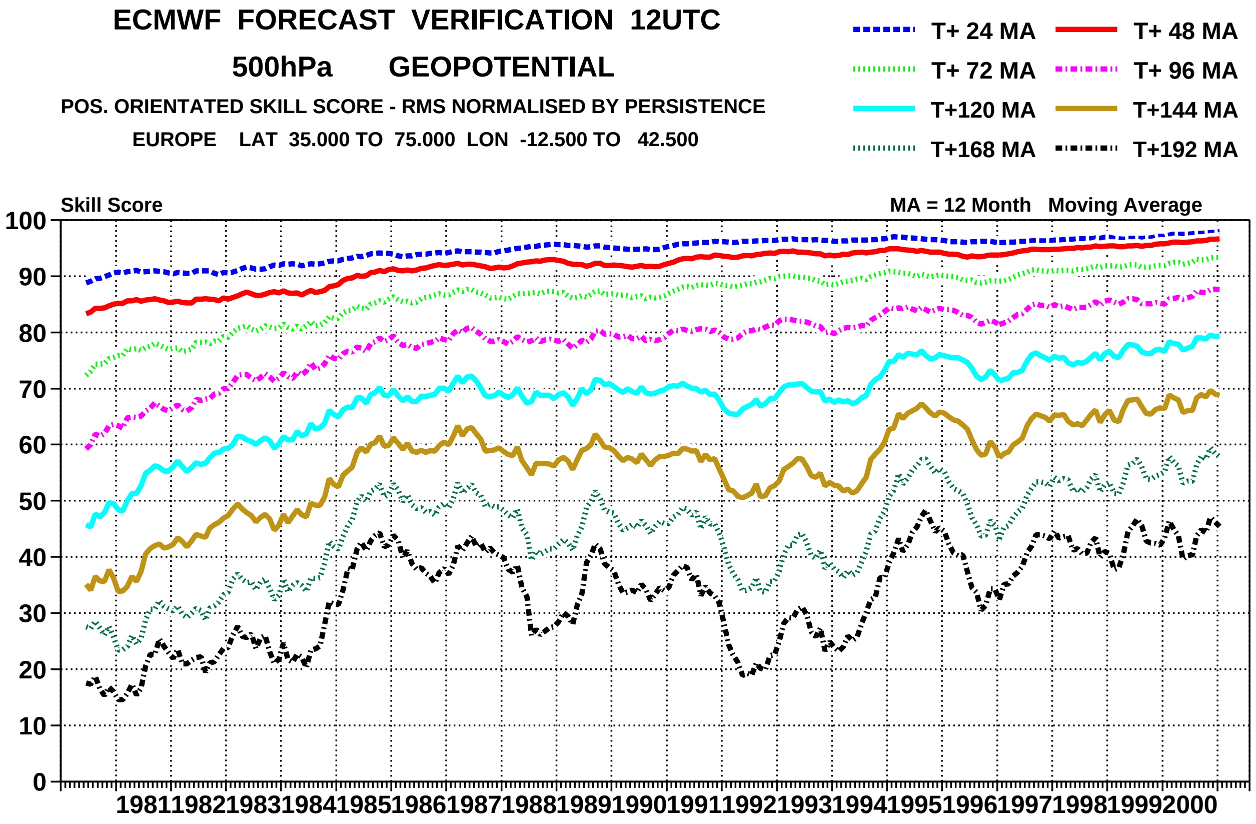


Figure 2: 500-hPa height monthly and annual means of the forecast range when the anomaly correlation is falling below 60% for Europe, Northern and Southern Extratropics

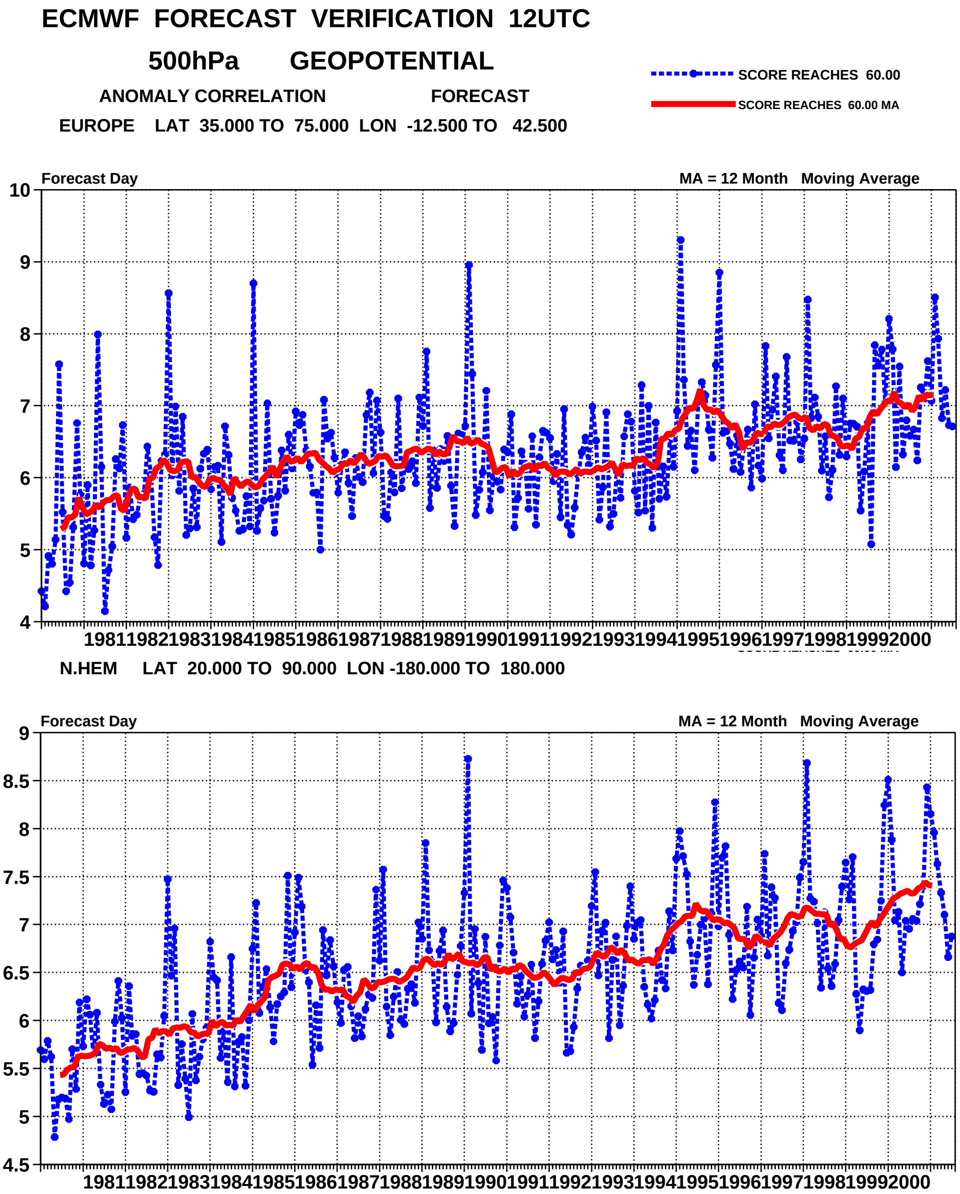


Figure 2: *500-hPa height monthly and annual means of the forecast range when the anomaly correlation is falling below 60% for Europe, Northern and Southern Extratropics*

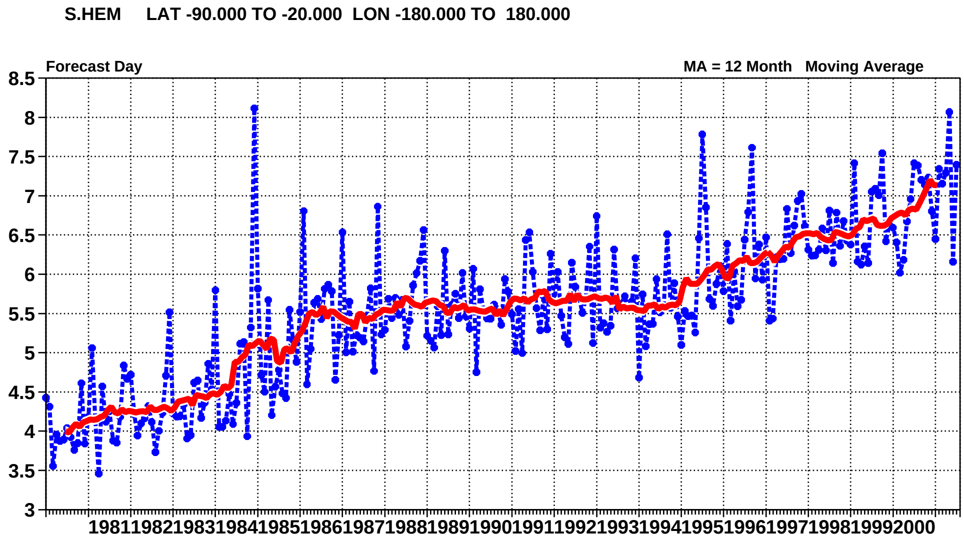


Figure 3: *Cumulative frequency distribution - 1000-hPa Anomaly Correlation over Europe, last three winters: from top to bottom, DJF 2000-01, 1999-2000 and 1998-99*

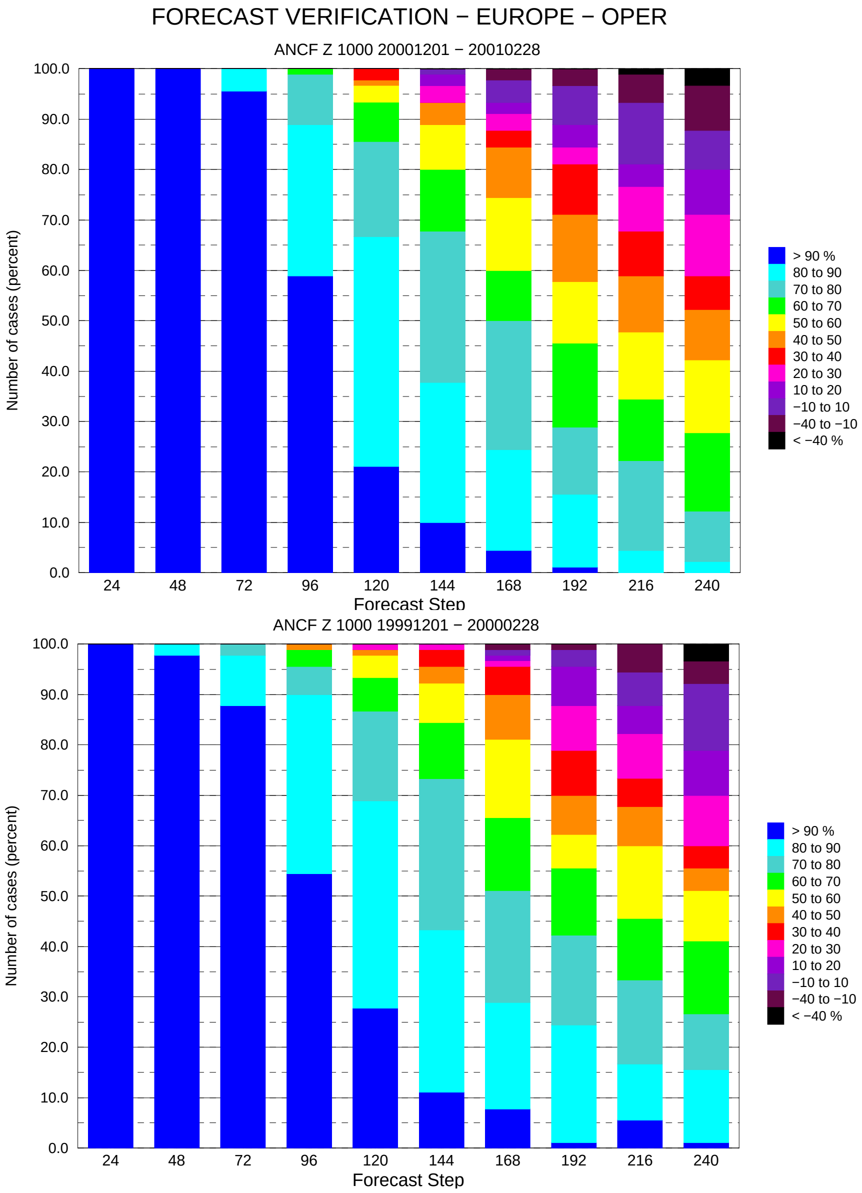


Figure 3: Cumulative frequency distribution - 1000-hPa Anomaly Correlation over Europe, last three winters: from top to bottom, DJF 2000-01, 1999-2000 and 1998-99

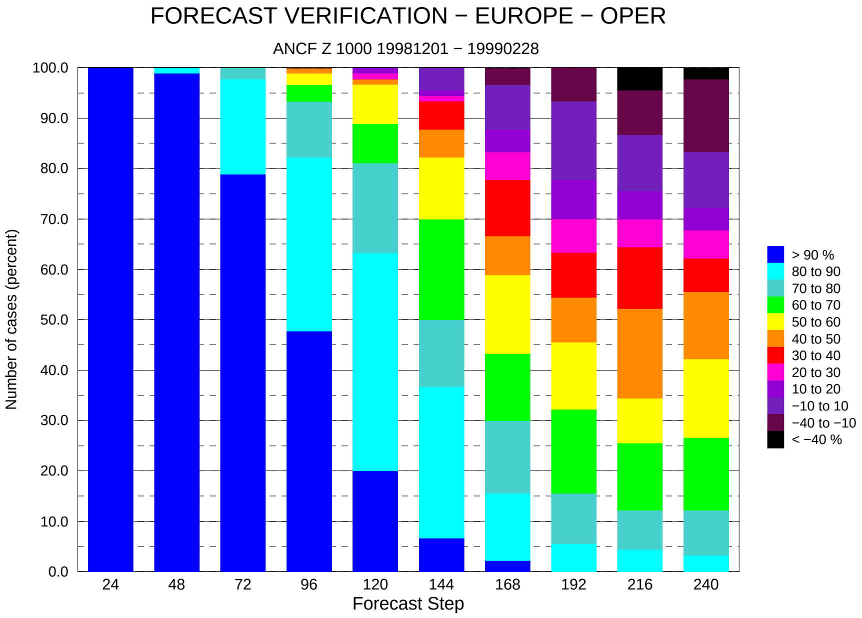


Figure 4: Gain curves: upward (downward) trends highlight periods when ECMWF 500-hPa RMSE over Europe were smaller (larger) than the mean error from NCEP, DWD, UKMO and CMC models; the same period in 1999-2000 is shown in the lower panel. Blue: 72h; red: 120h forecast.

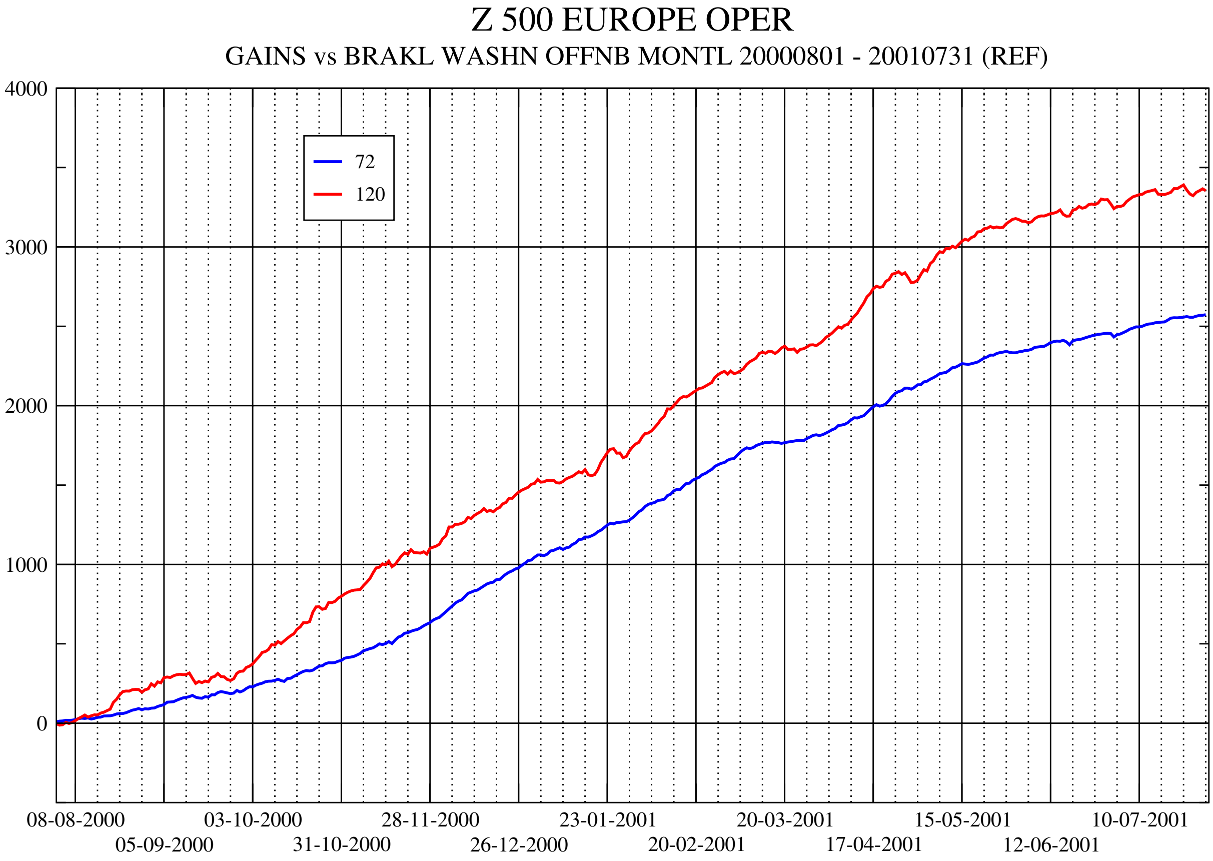


Figure 4: Gain curves: upward (downward) trends highlight periods when ECMWF 500-hPa RMSE over Europe were smaller (larger) than the mean error from NCEP, DWD, UKMO and CMC models; the same period in 1999-2000 is shown in the lower panel. Blue: 72h; red: 120h forecast.

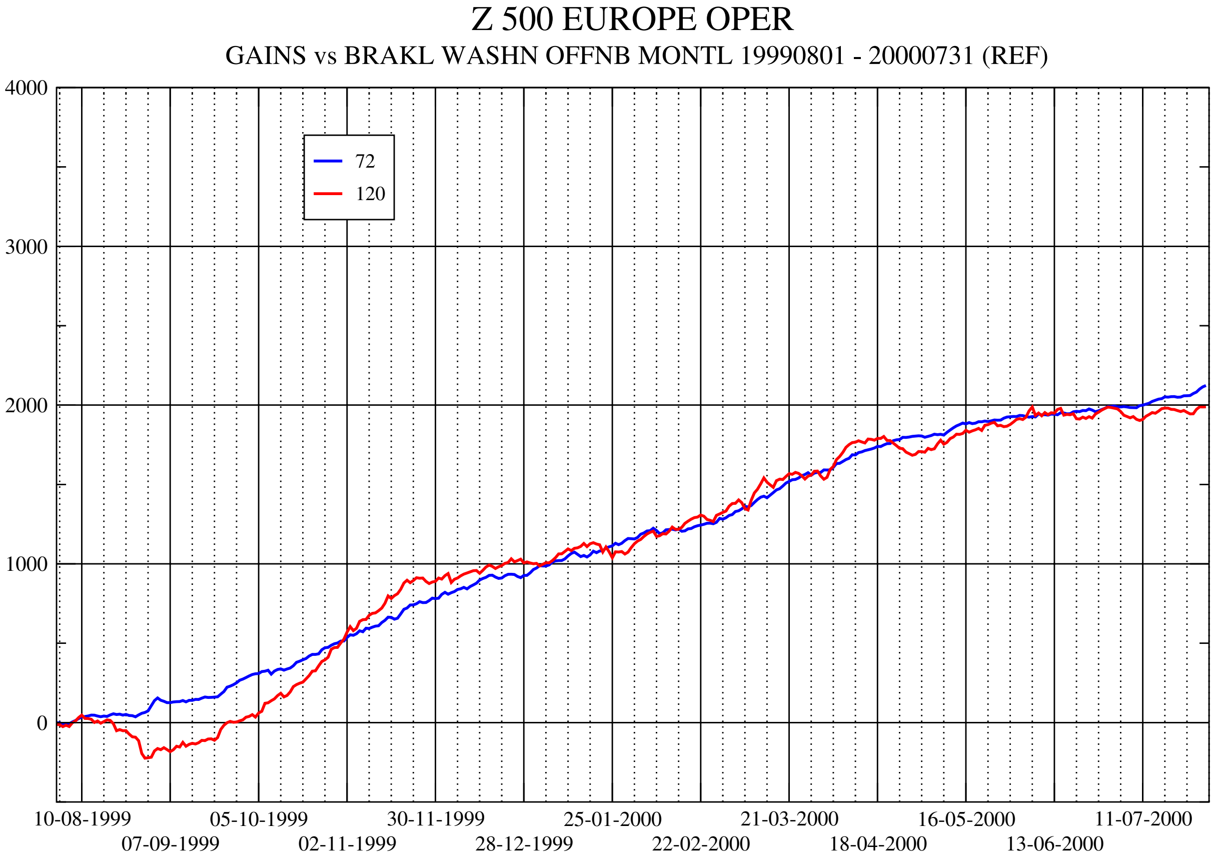
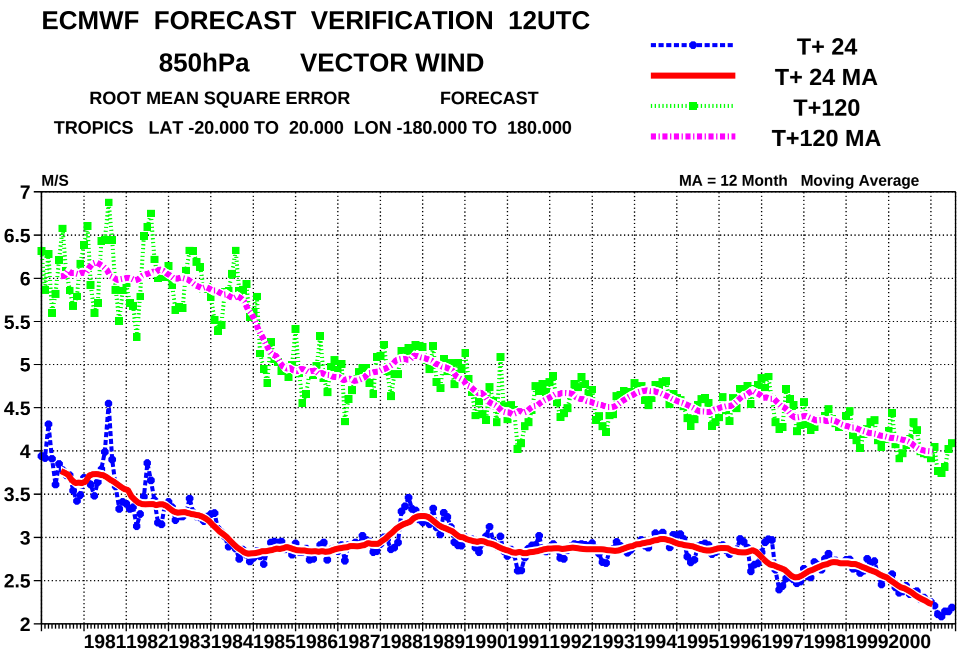


Figure 5: Model scores in the Tropics (root mean square errors against 850hPa and 200hPa wind analysis) and Northern Hemisphere stratosphere (same score, level 50hPa);



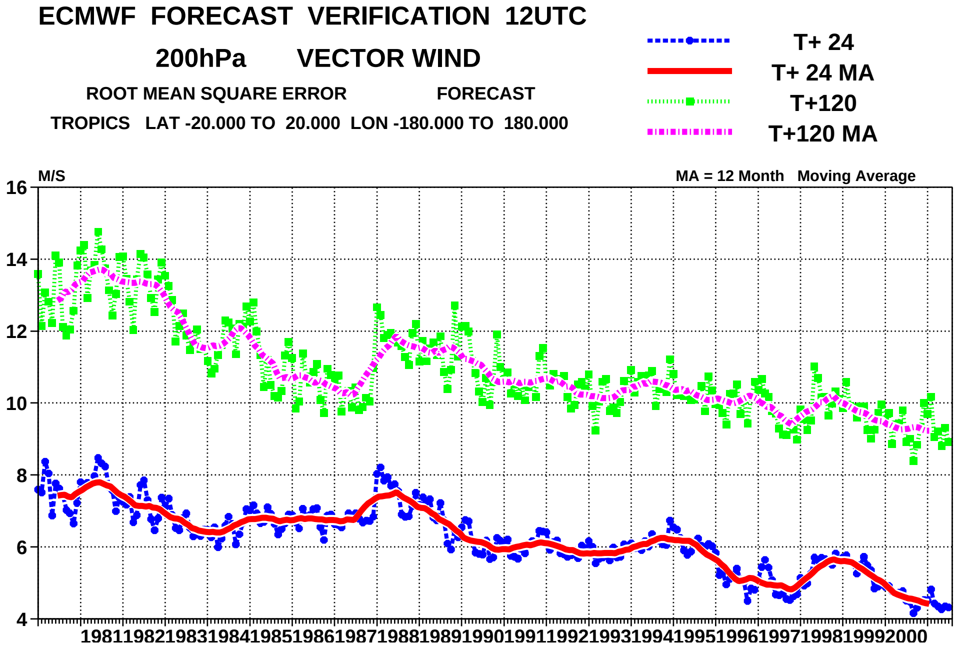


Figure 5: Model scores in the Tropics (root mean square errors against 850hPa and 200hPa wind analysis) and Northern Hemisphere stratosphere (same score, level 50hPa);

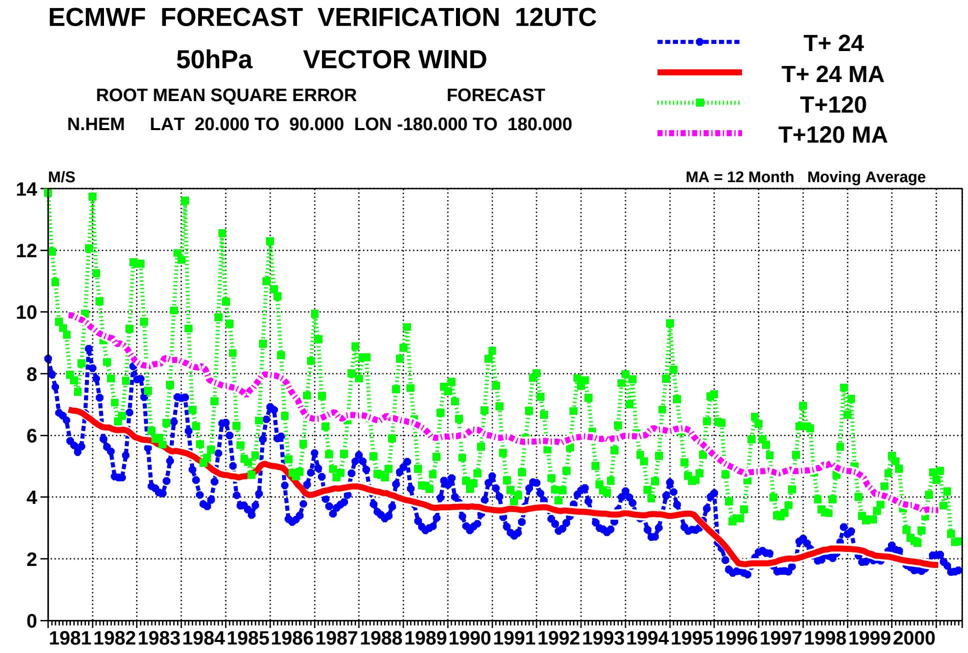


Figure 6: RMS of the difference between 24h-consecutive 500-hPa height forecasts verifying the same day: Europe (upper) and Extratropical Northern Hemisphere (lower)

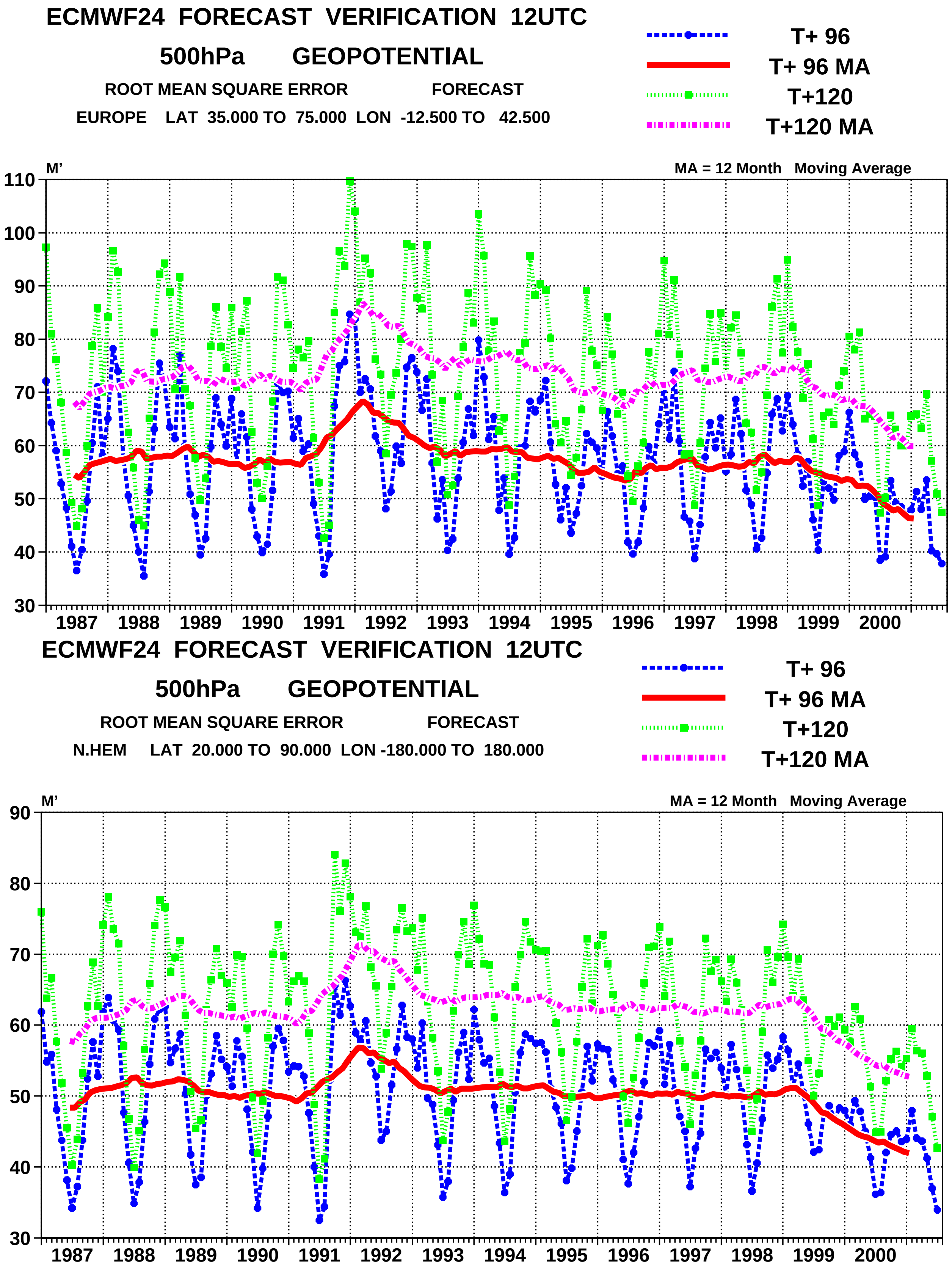


Figure 7: *WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)*

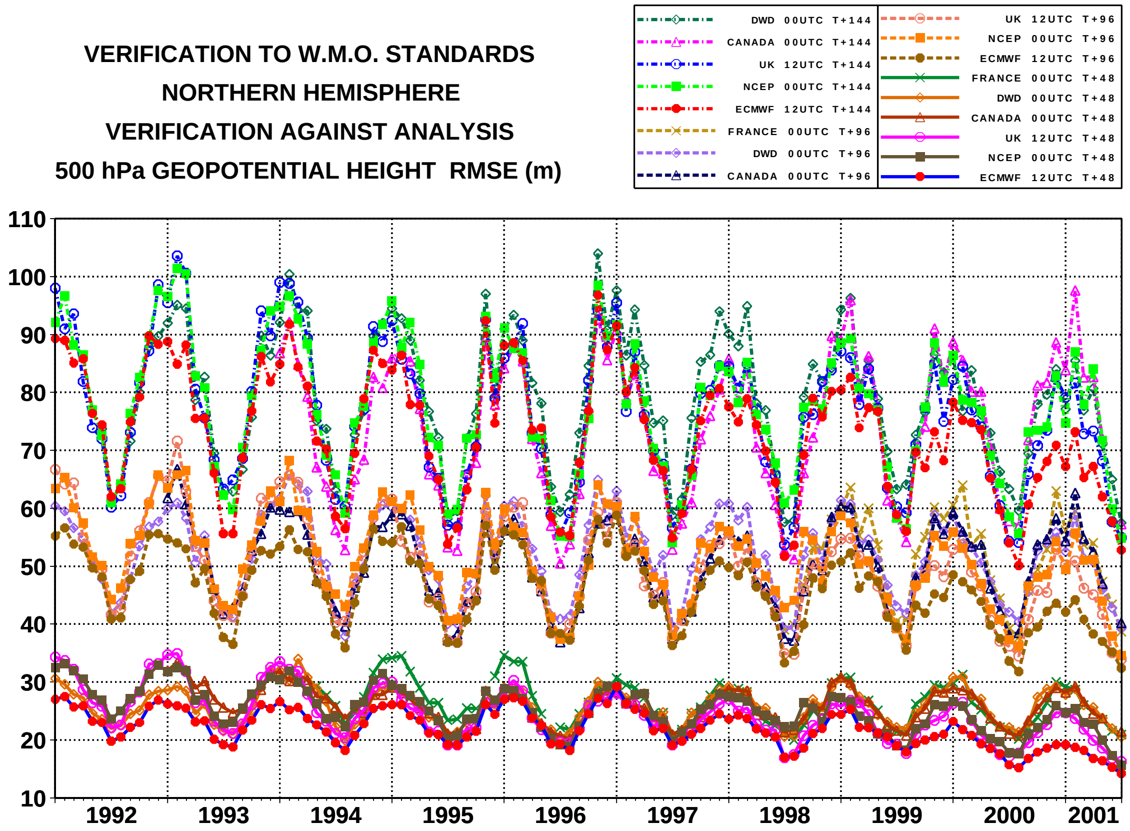


Figure 7: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)

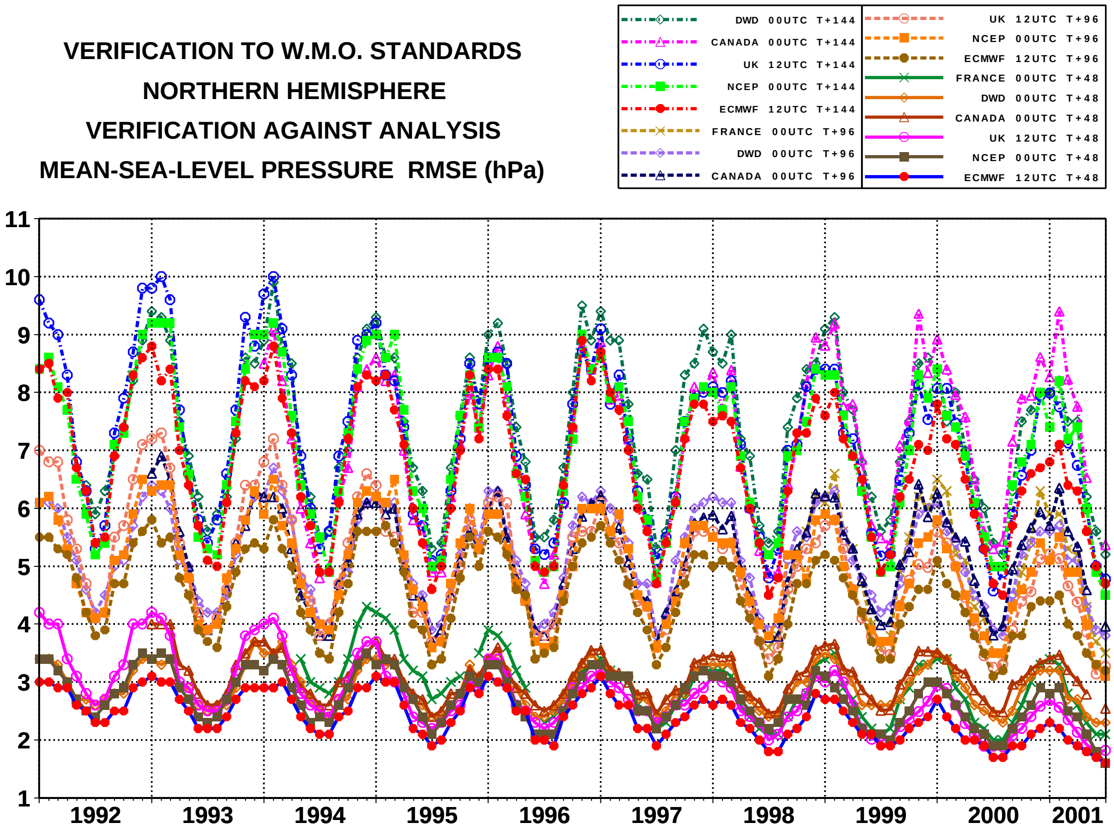


Figure 8: *WMO/CBS exchanged scores in the Northern Extratropics (12-months average, 500-hPa and mean sea-level pressure); reference for verification is taken from each centre's own analysis*

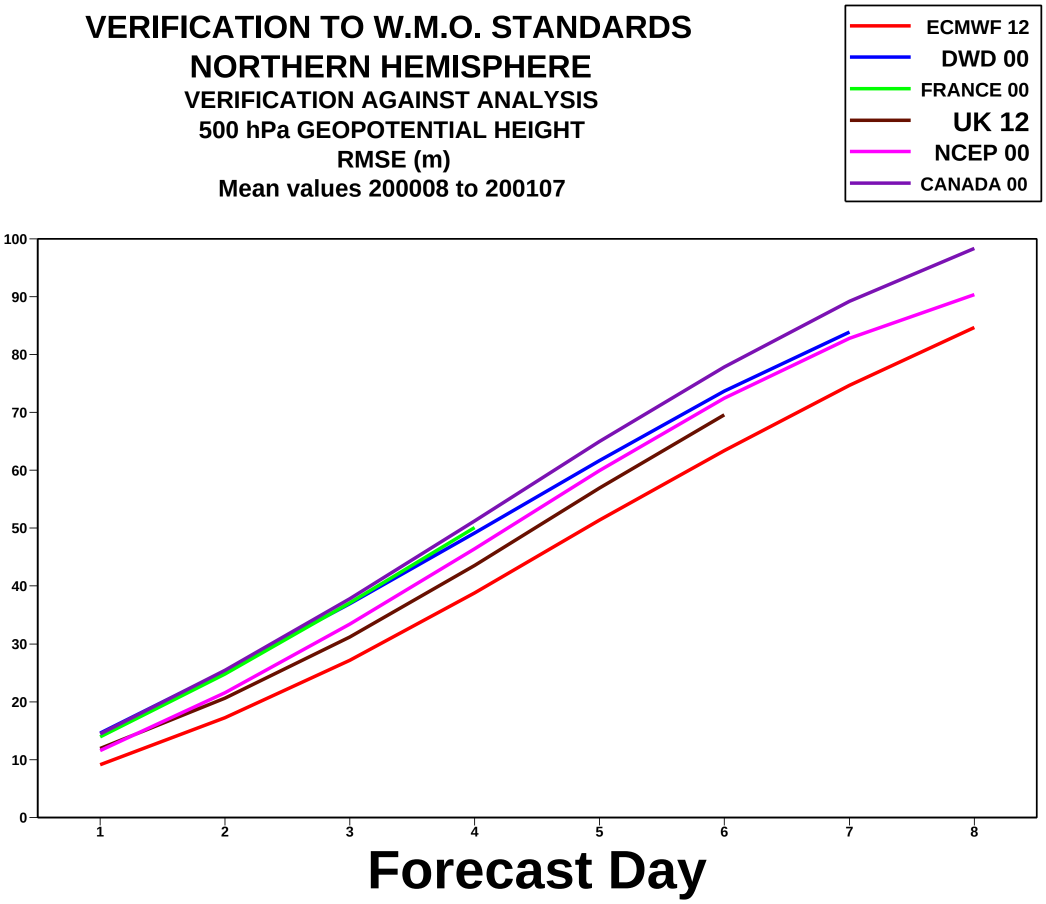


Figure 8: WMO/CBS exchanged scores in the Northern Extratropics (12-months average, 500-hPa and mean sea-level pressure); reference for verification is taken from each centre's own analysis

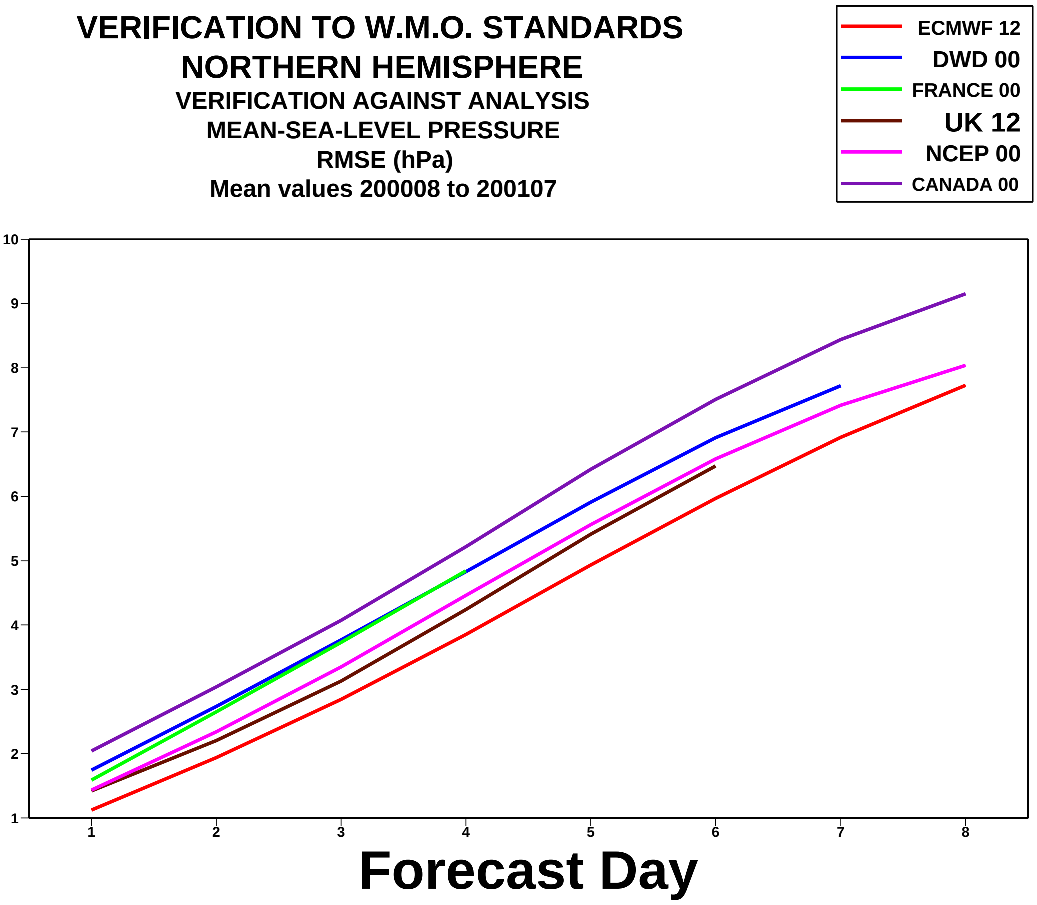


Figure 9: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)

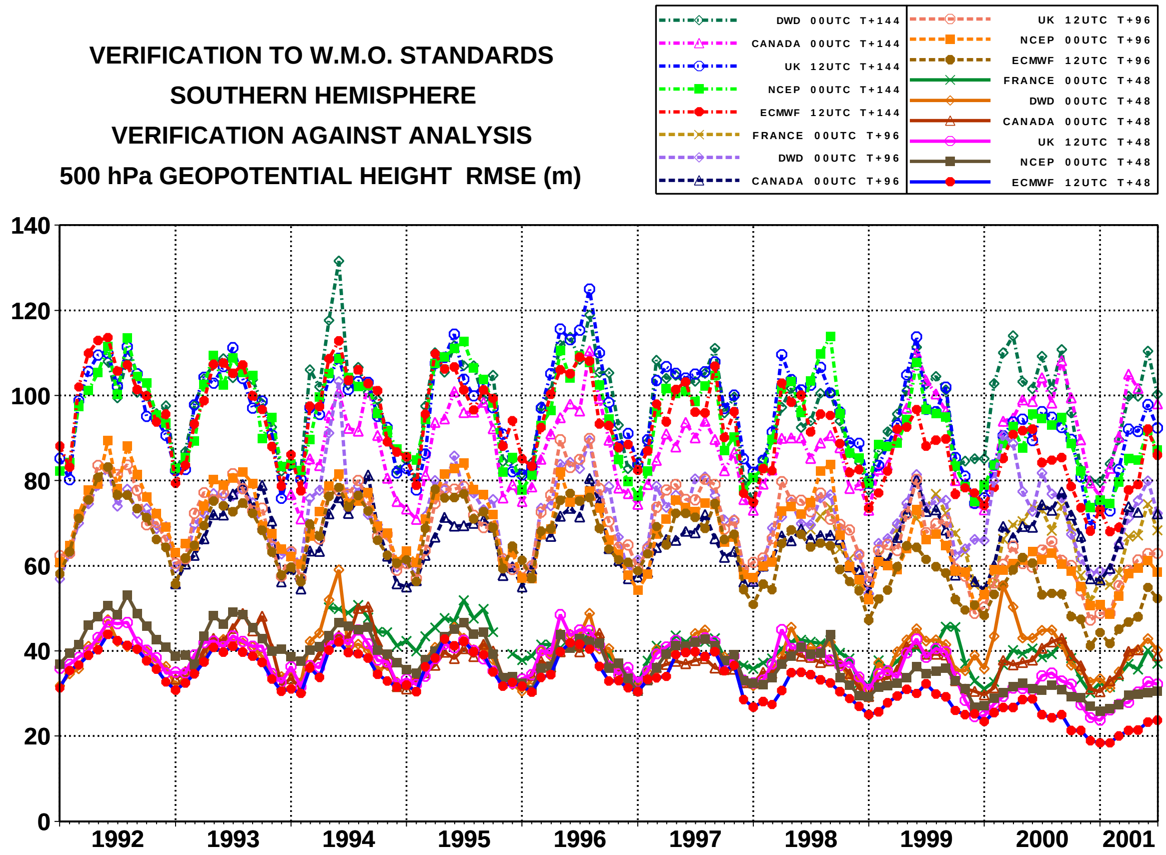


Figure 9: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500-hPa and MSLP for D+2, D+4 and D+6)

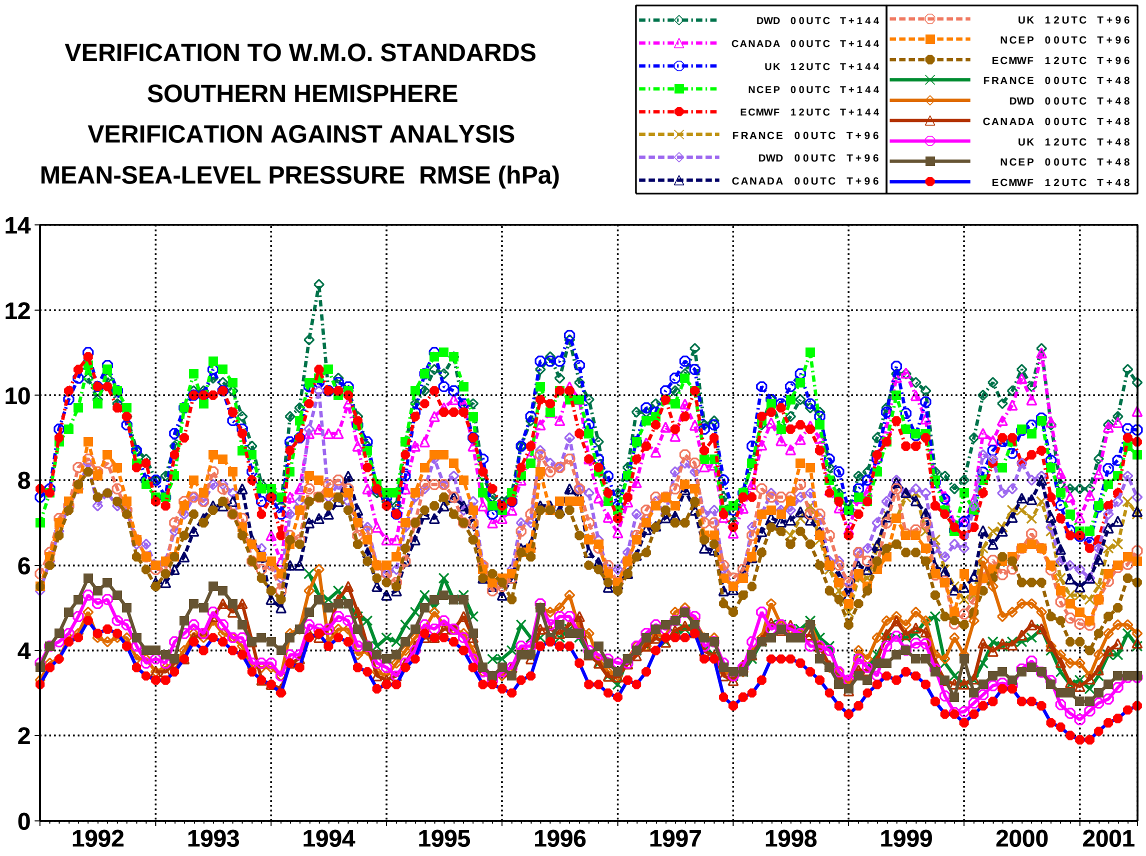


Figure 10: WMO/CBS exchanged scores in the Southern Extratropics (12-months average, 500-hPa and MSLP); reference for verification is taken from each centre's own analysis

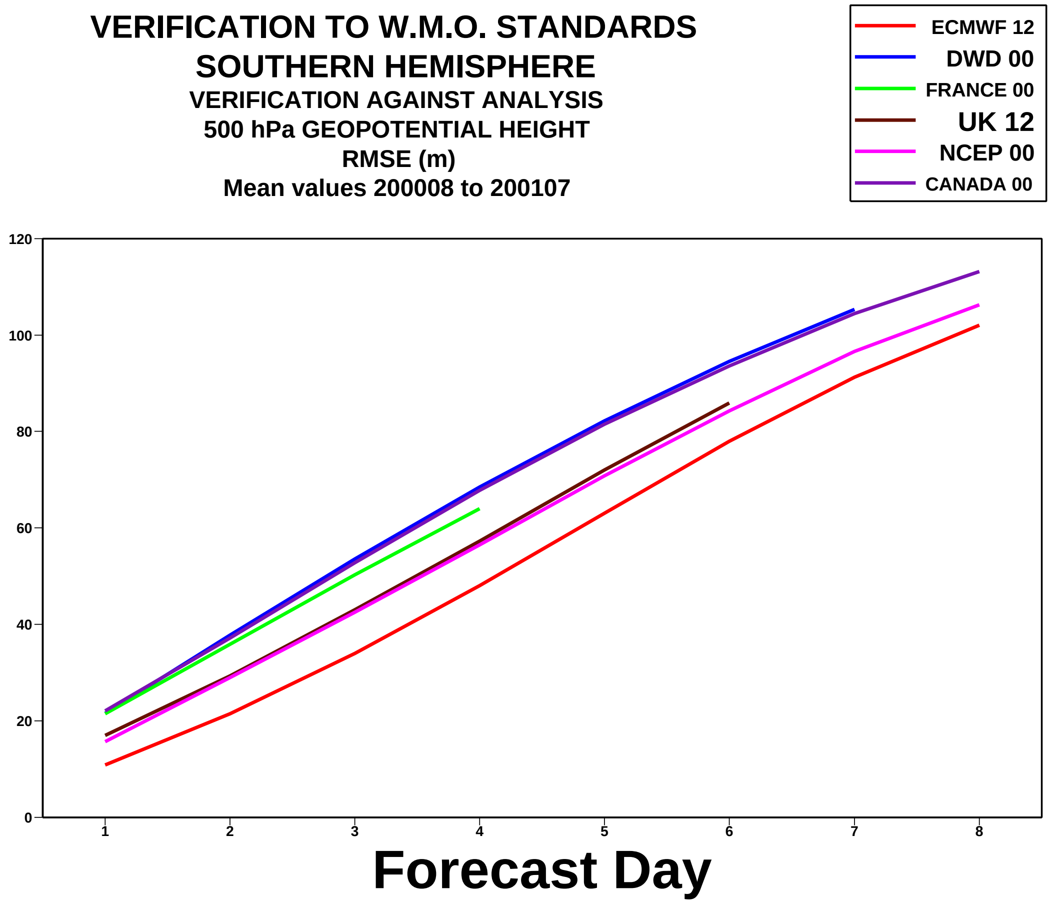


Figure 10: WMO/CBS exchanged scores in the Southern Extratropics (12-months average, 500-hPa and MSLP); reference for verification is taken from each centre's own analysis

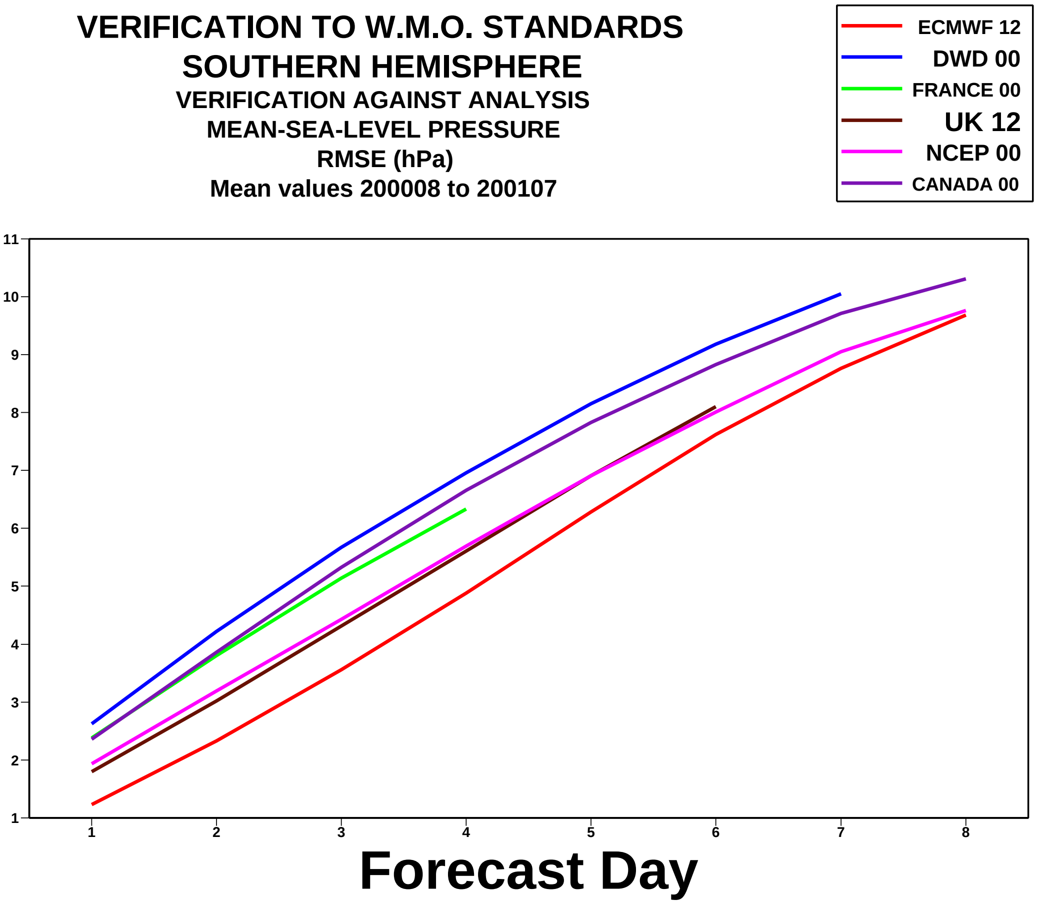


Figure 11: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250-hPa and 850-hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis

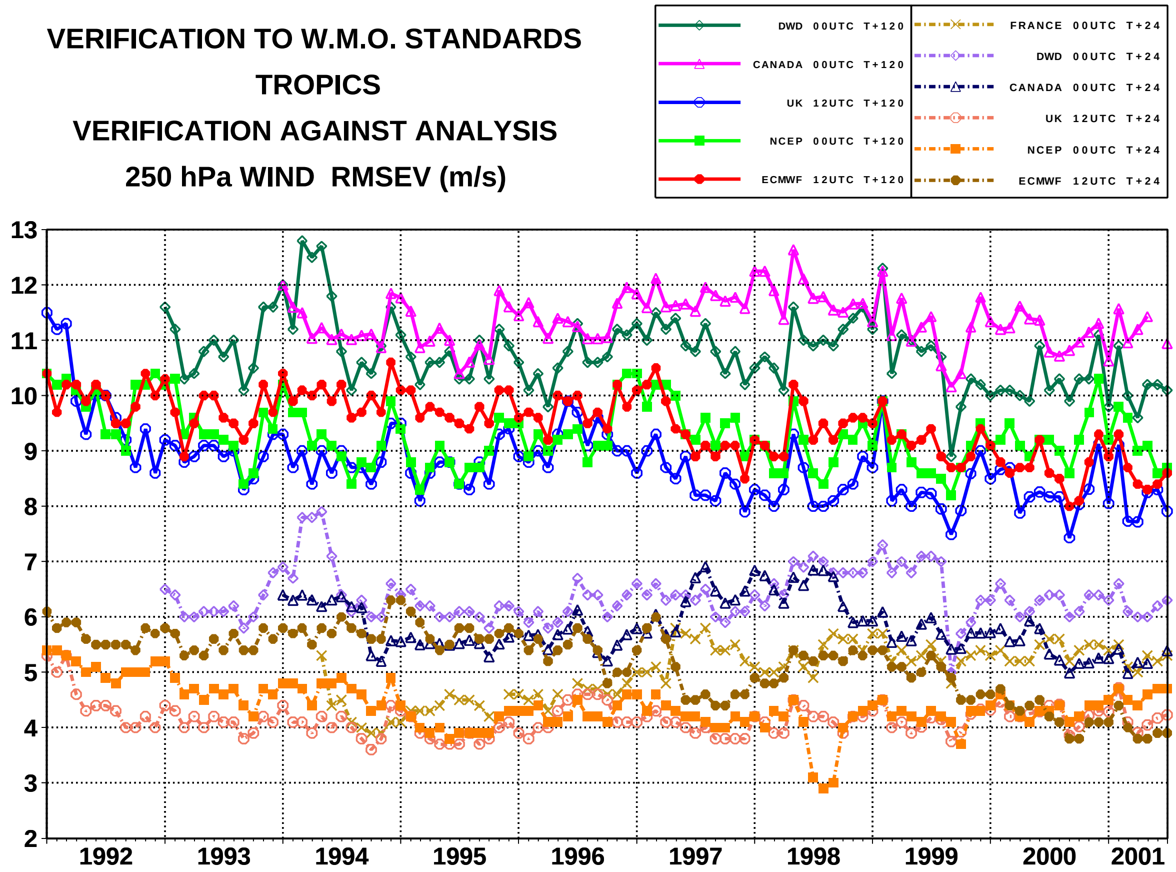


Figure 11: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250-hPa and 850-hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis

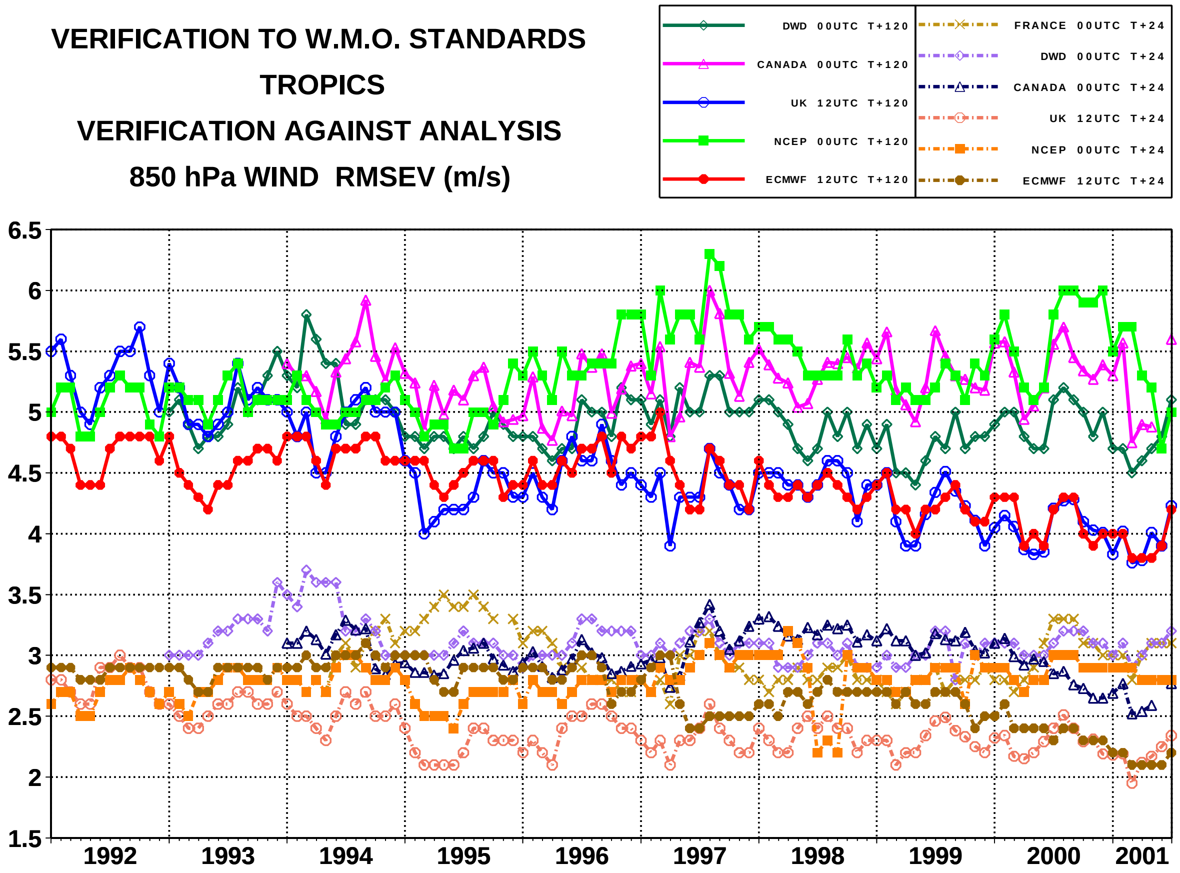


Figure 12: WMO/CBS exchanged scores in the Tropics (12-months average, 250- and 850-hPa); reference for verification is taken from each centre's own analysis

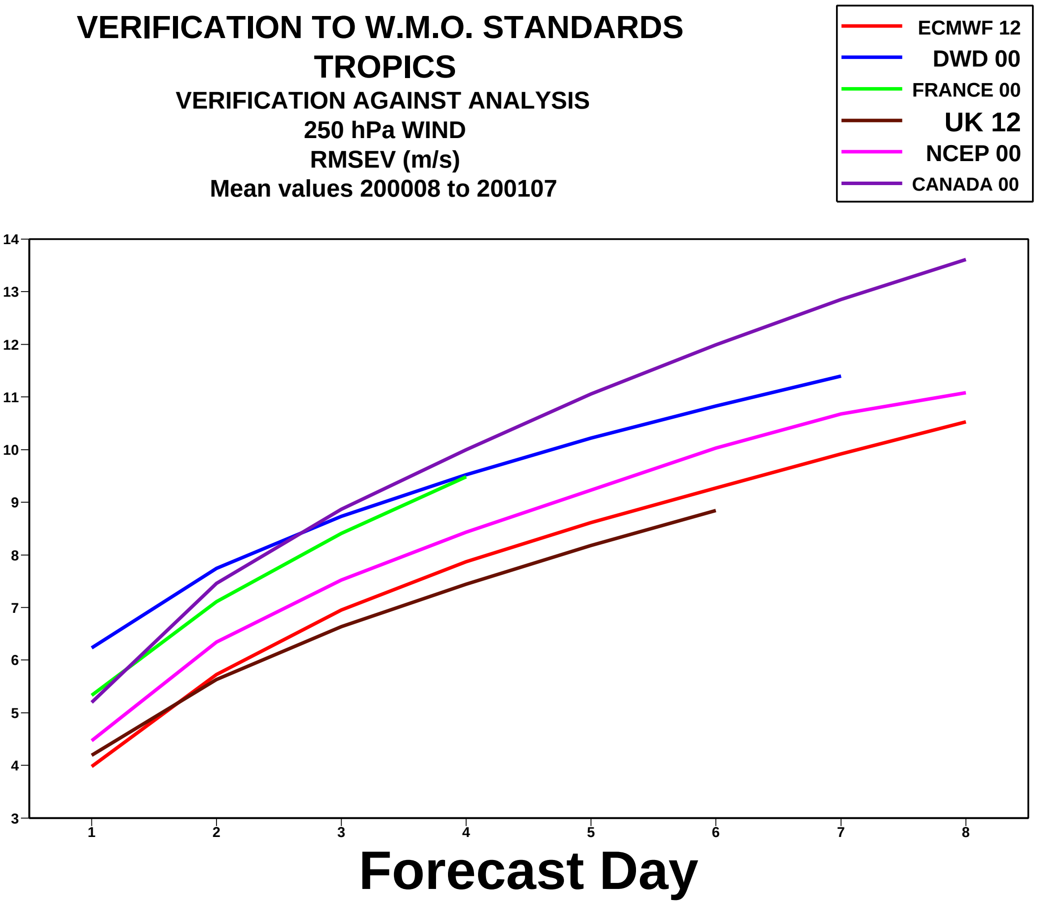


Figure 12: WMO/CBS exchanged scores in the Tropics (12-months average, 250- and 850-hPa); reference for verification is taken from each centre's own analysis

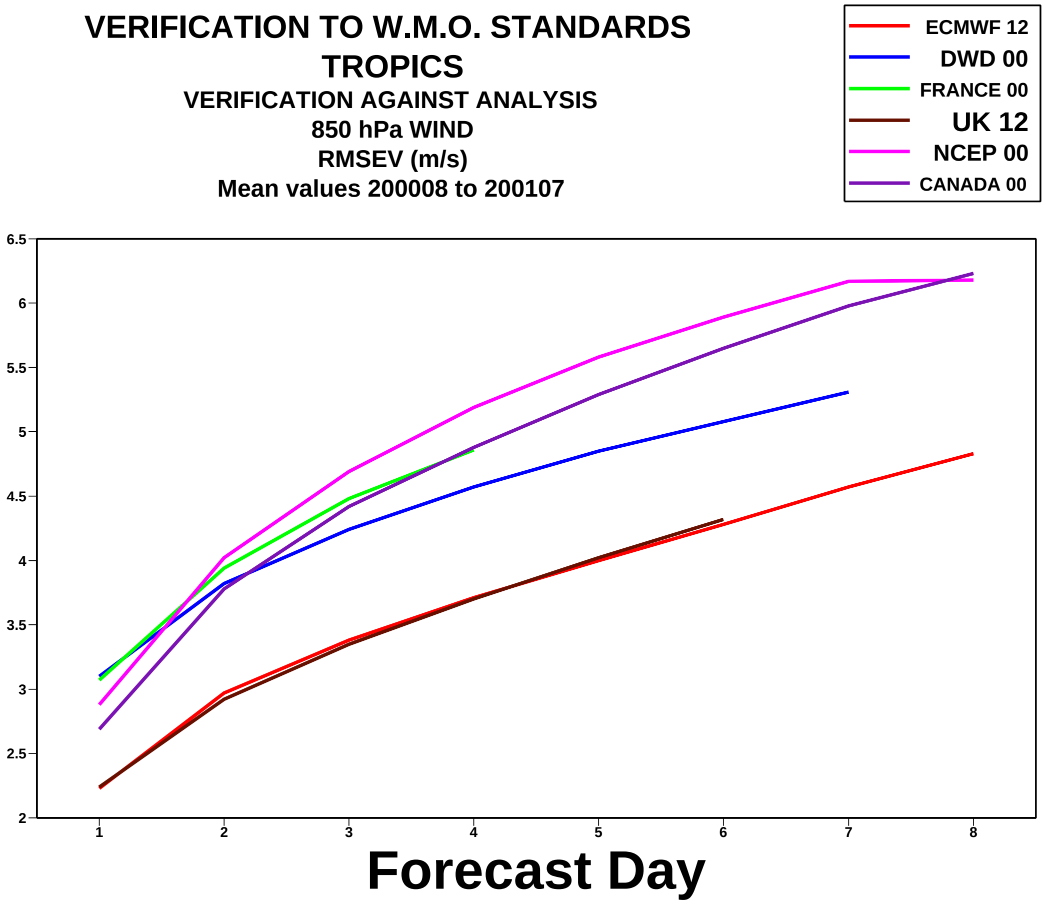


Figure 13: WMO/CBS exchanged scores using radiosondes: 500-hPa RMS error over Europe and 850-hPa wind errors in the Tropics (annual mean)

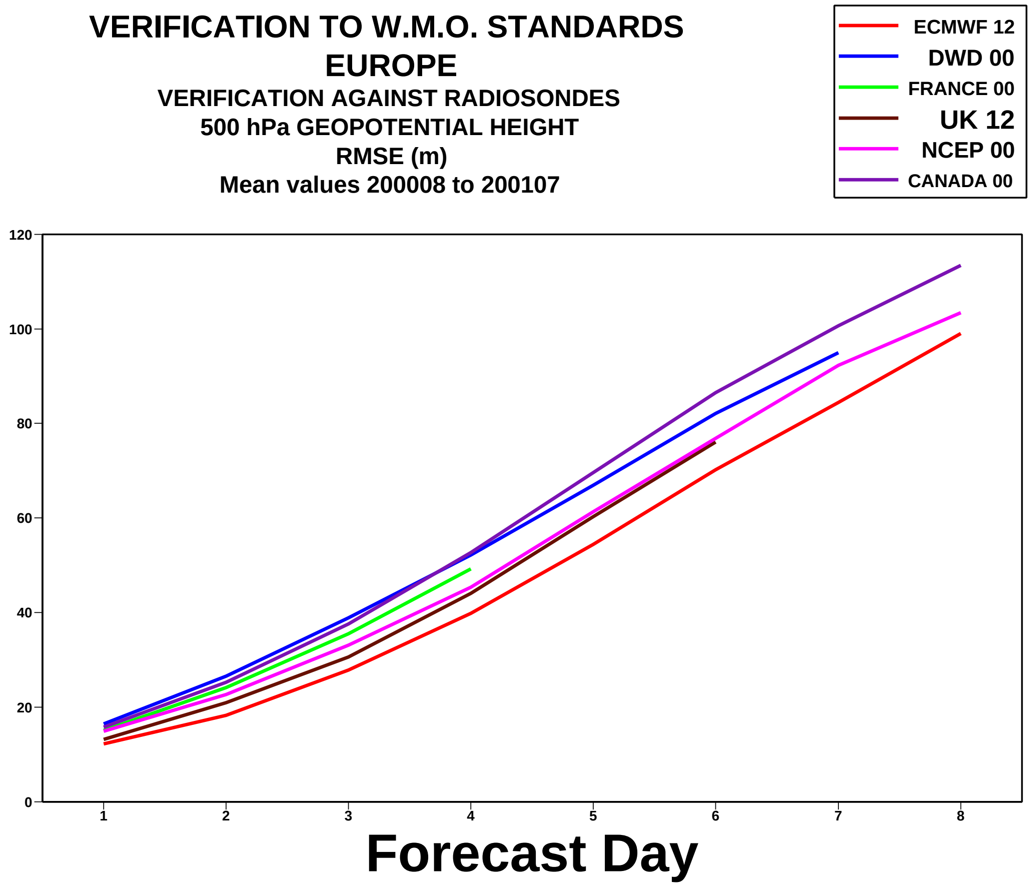


Figure 13: WMO/CBS exchanged scores using radiosondes: 500-hPa RMS error over Europe and 850-hPa wind errors in the Tropics (annual mean)

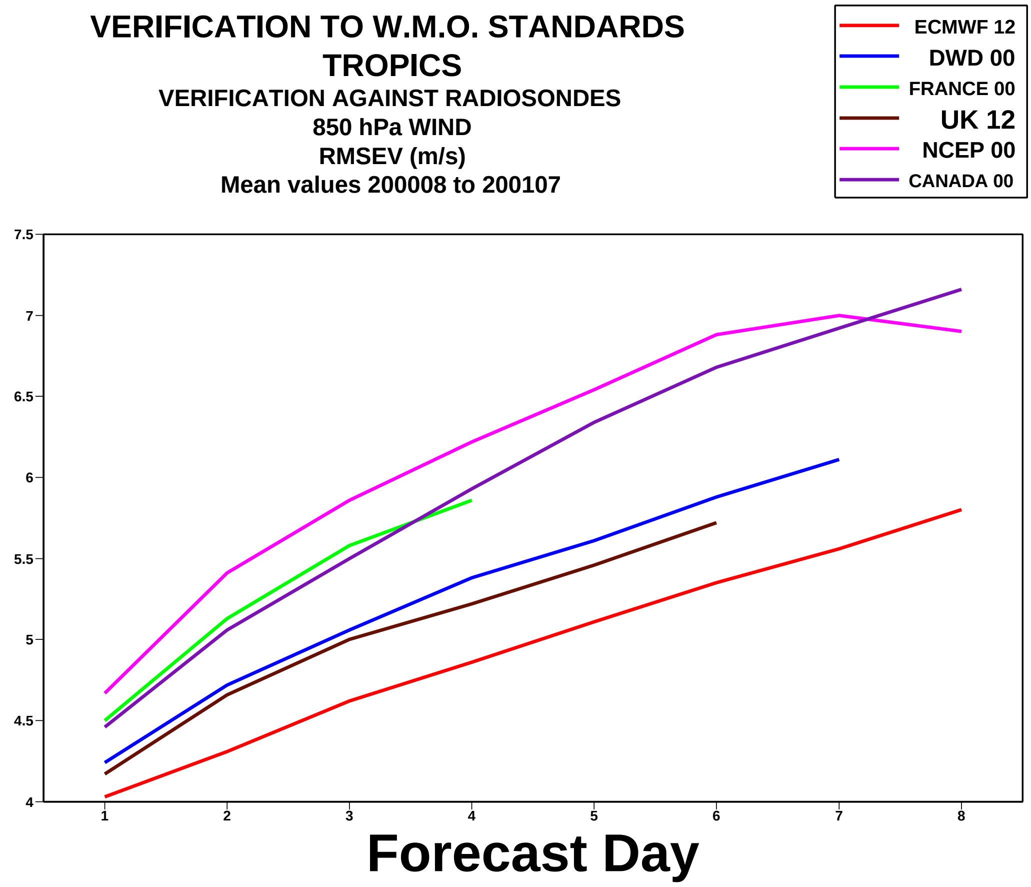


Figure 14: Scores against European SYNOPs of 2m-Temperature and specific humidity forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)

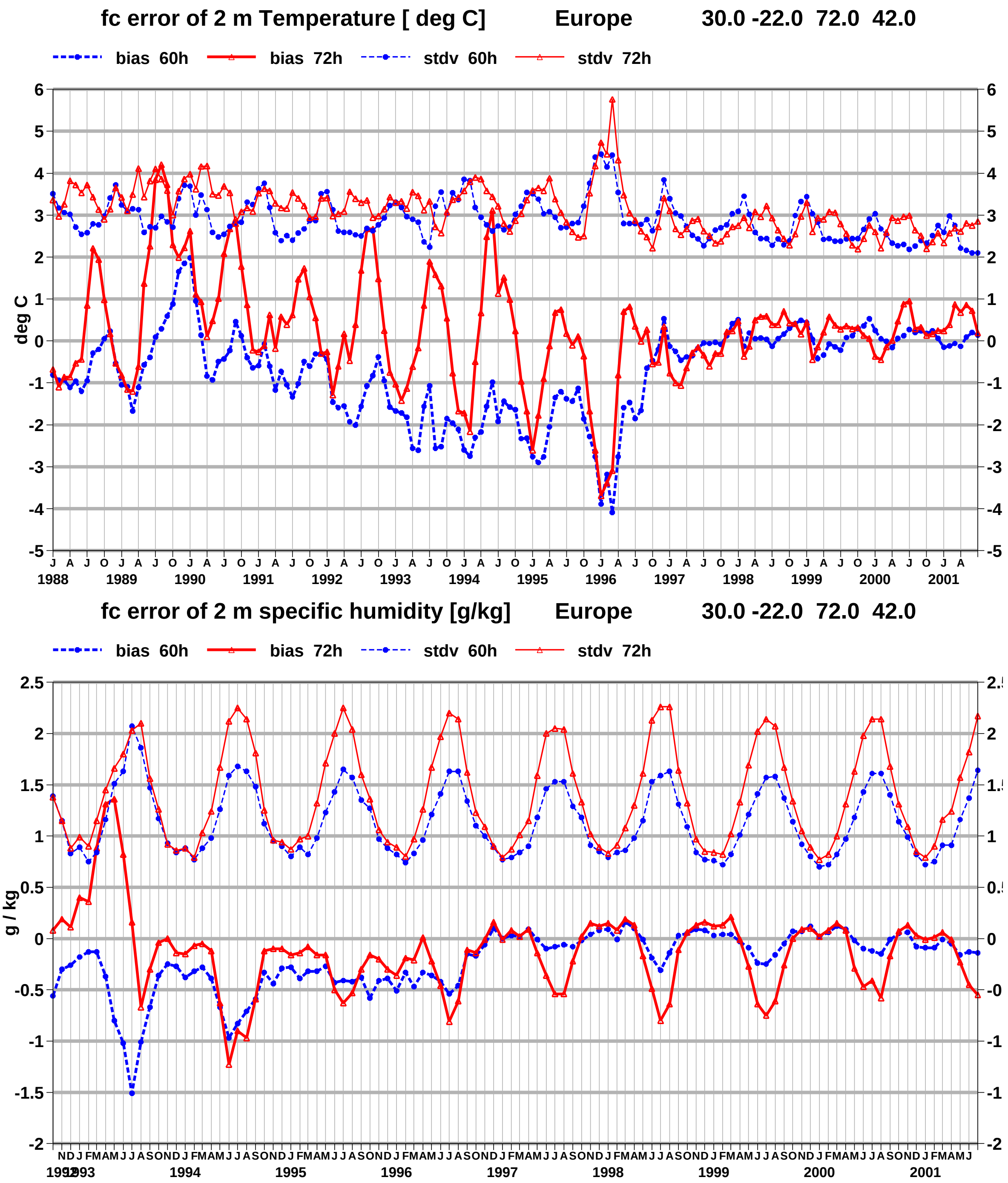


Figure 15: Scores against European SYNOPs of total cloud cover and 10m wind speed forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)

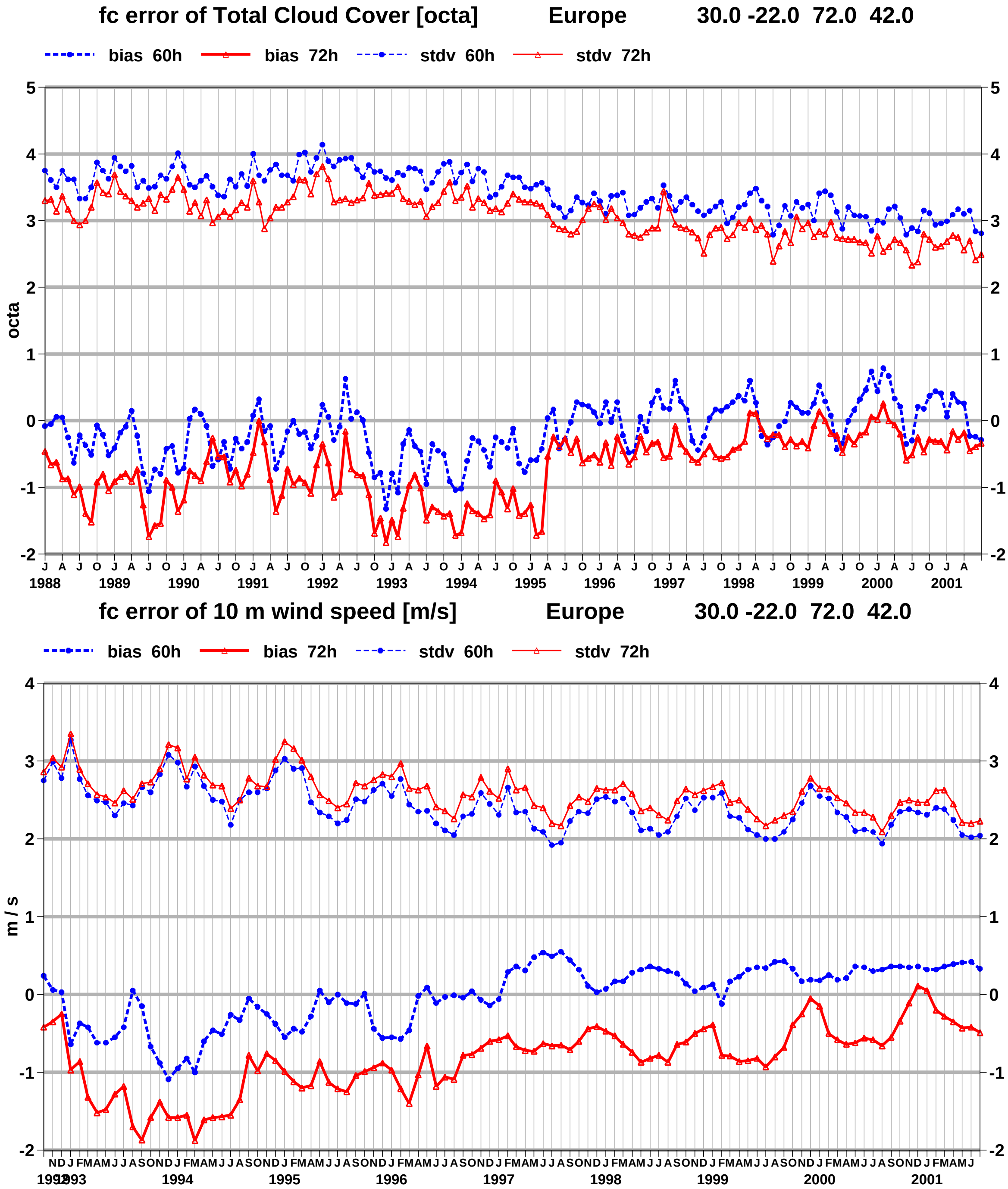


Figure 16: 6h-accumulated precipitation forecasts biases (T+54/60/66/72h) with respect to SYNOP

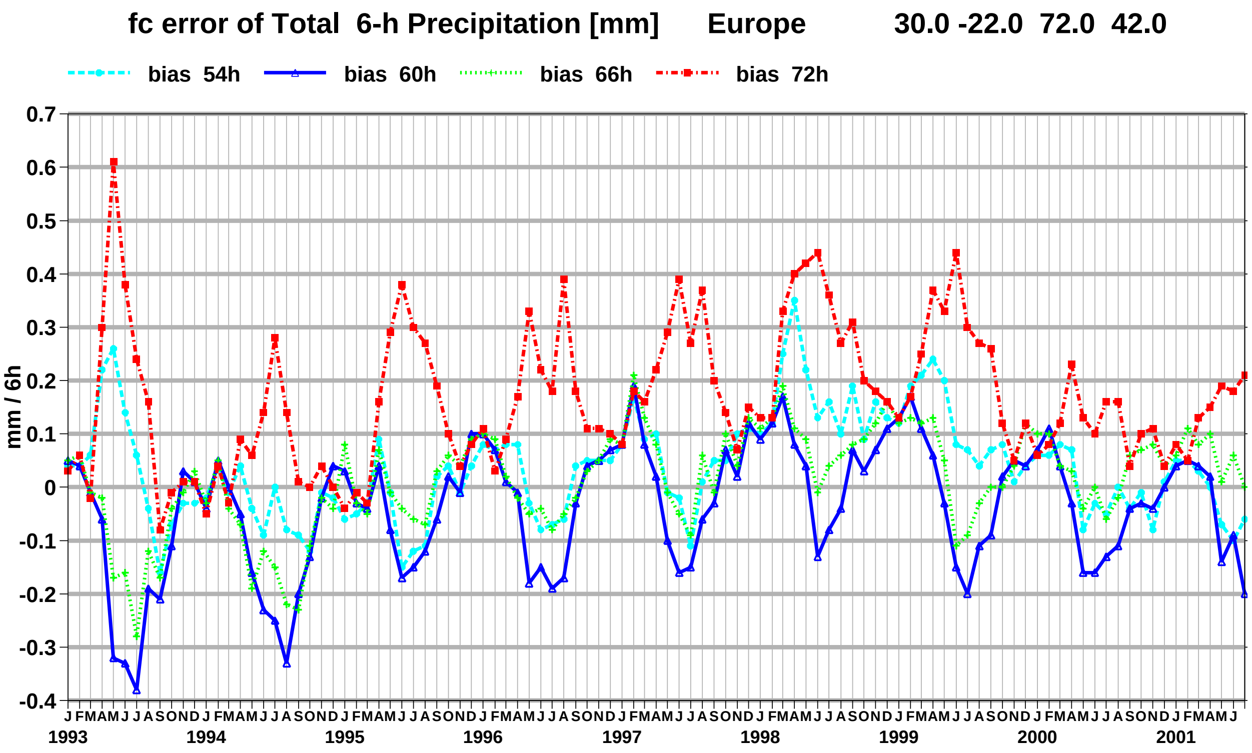


Figure 17: Empirical distributions of heavy daily precipitation: ECMWF forecast (magenta, DJF 2000; black: DJF 2001) and SYNOP observations (green: DJF 2000 and red: DJF2001); frequency biases are also shown; only events with more than 10mm of precipitation are shown; horizontal scale is logarithmic (ticks every 10mm from 10 to 100mm/day)

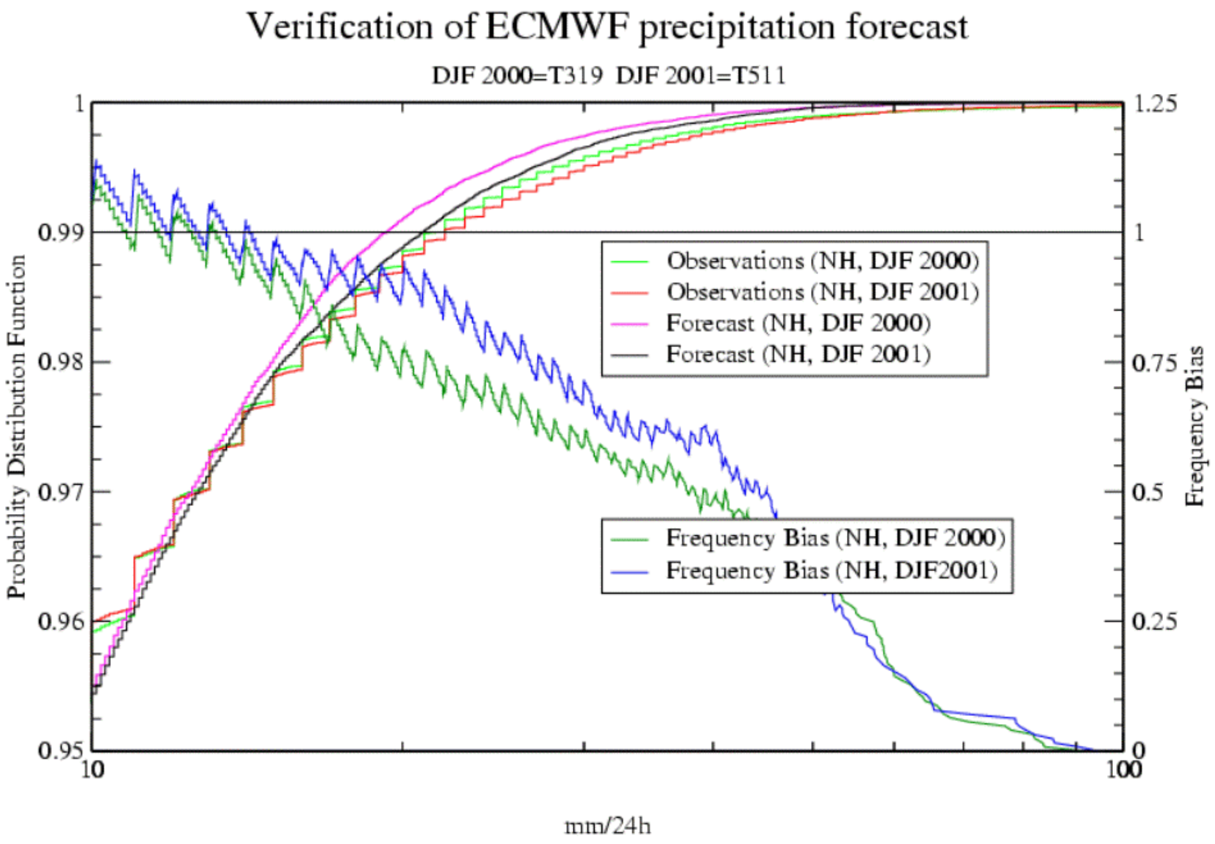
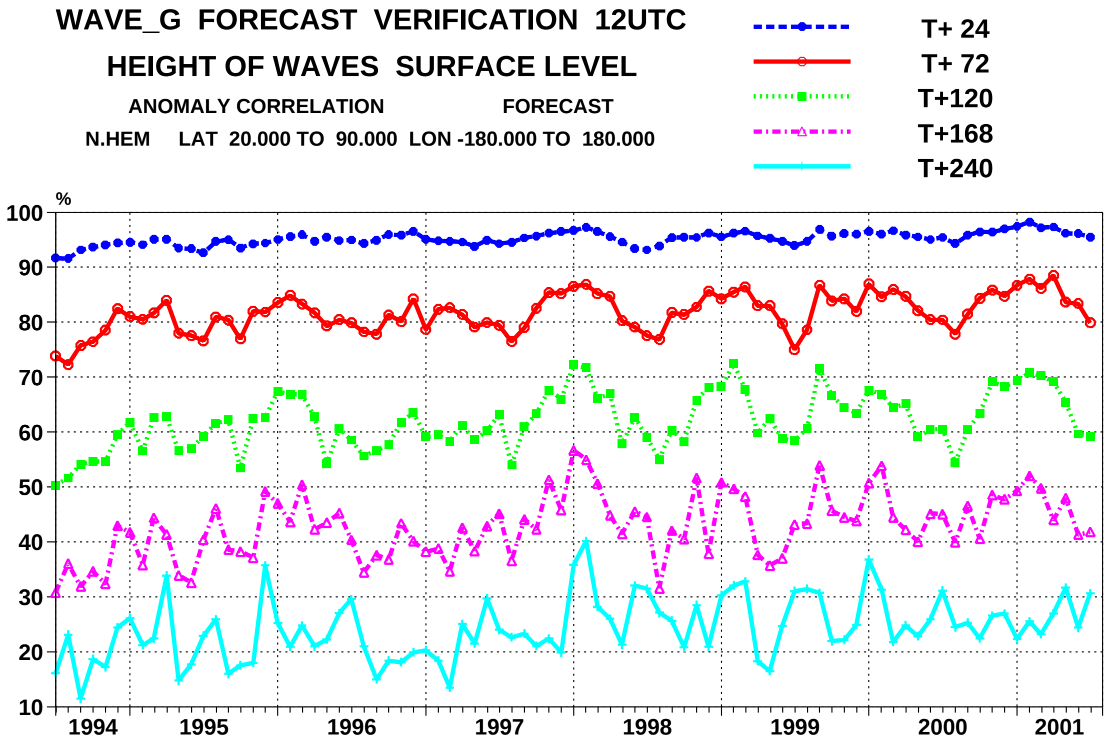


Figure 18: Scores (rms and anomaly correlation) of oceanic wave heights verified against the analysis (Northern Extratropics)



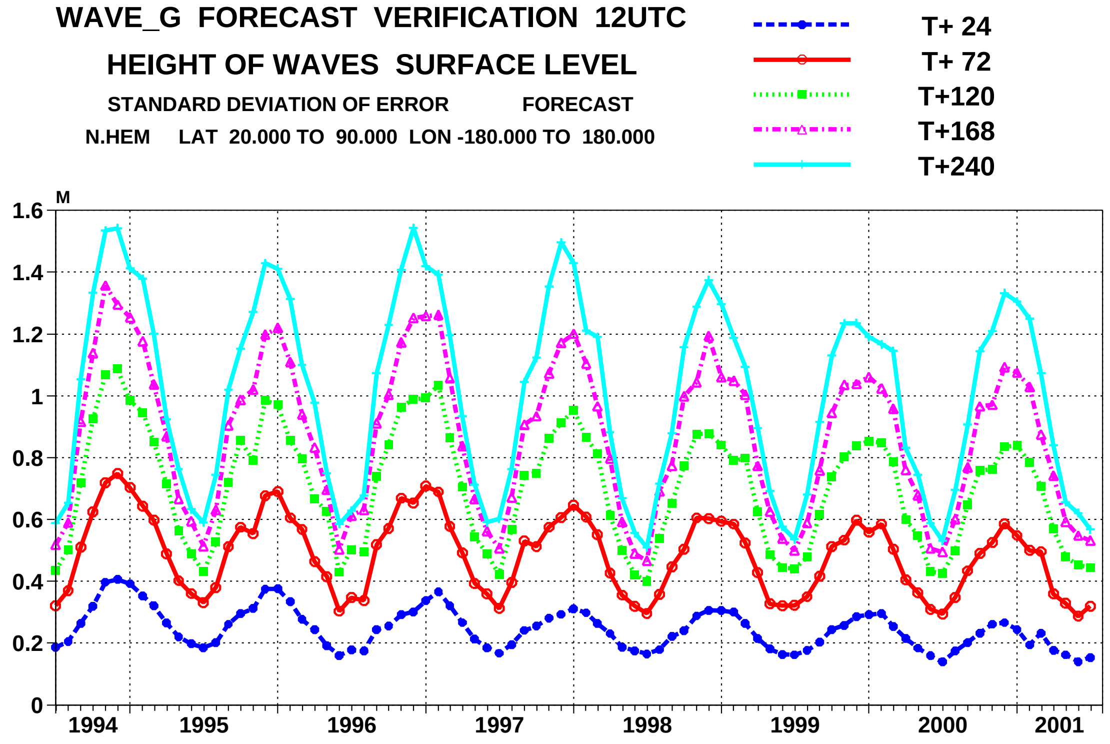
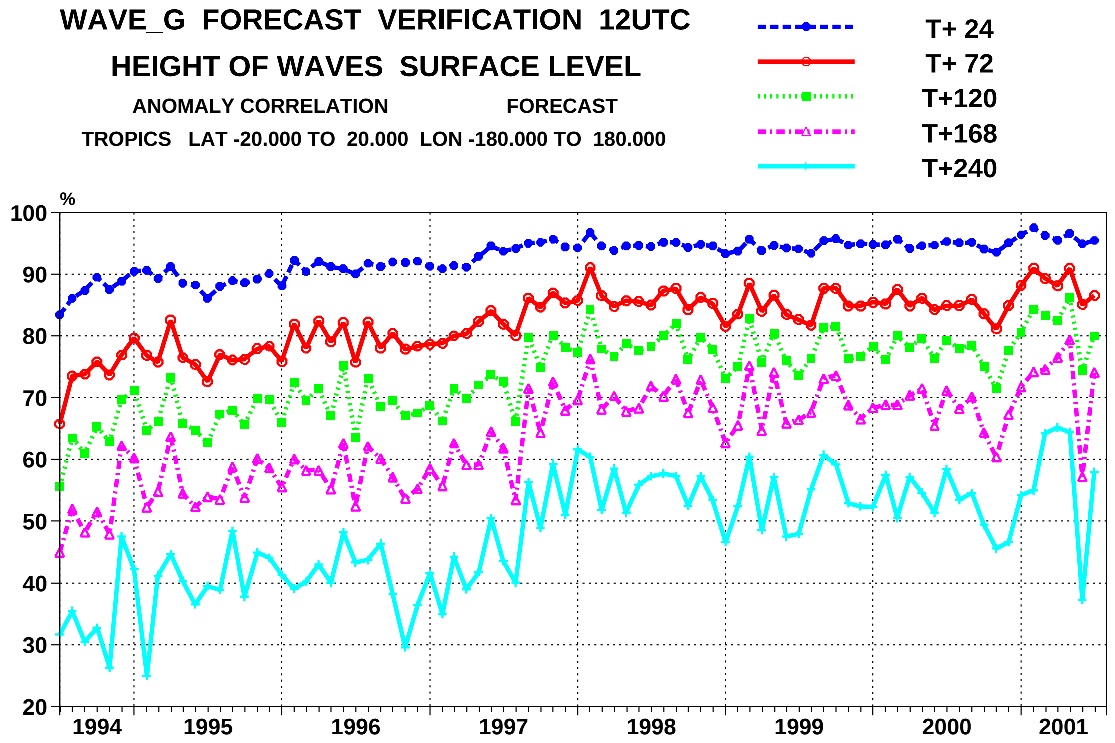


Figure 19: Scores (rms and anomaly correlation) of oceanic wave heights verified against the analysis (Tropics)



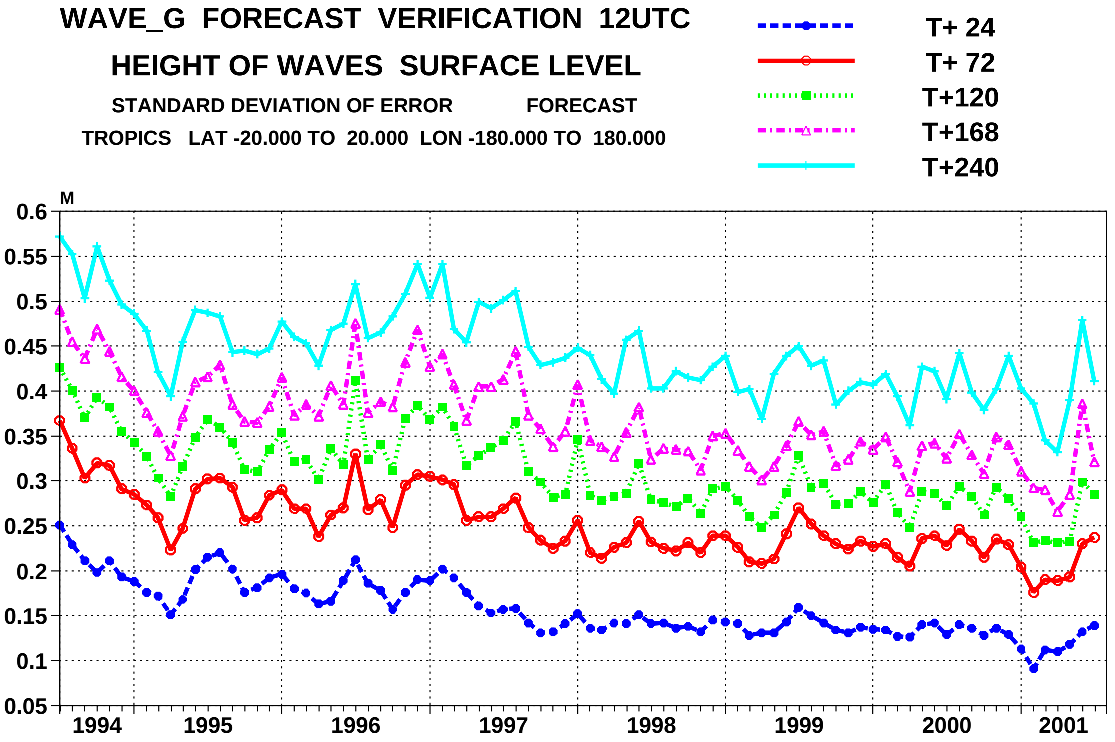


Figure 20: Spread-skill diagrams covering the six months period centred on the EPS perturbations rescaling date (5 February 2001); upper panel: time series; lower panel: scatter diagram (each point is an average of 15 days with similar spread); similar periods in 1999/2000 are also shown for reference; forecasts are for 500-hPa height over Europe at Day 6.

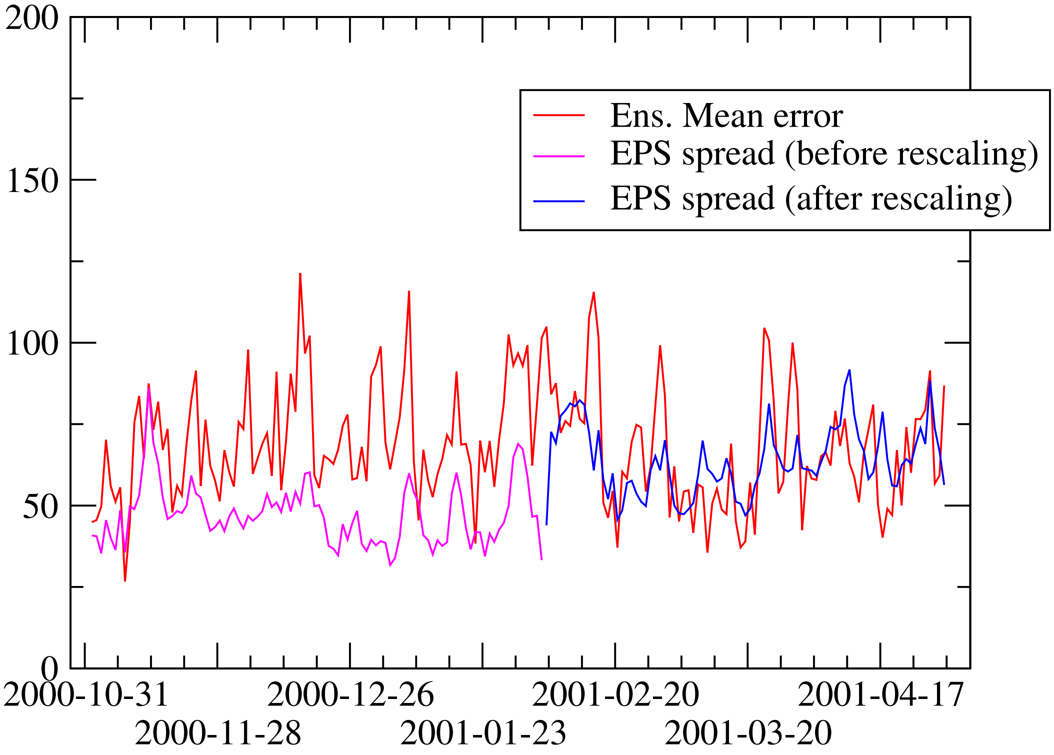


Figure 20: *Spread-skill diagrams covering the six months period centred on the EPS perturbations rescaling date (5 February 2001); upper panel: time series; lower panel: scatter diagram (each point is an average of 15 days with similar spread); similar periods in 1999/2000 are also shown for reference; forecasts are for 500-hPa height over Europe at Day 6.*

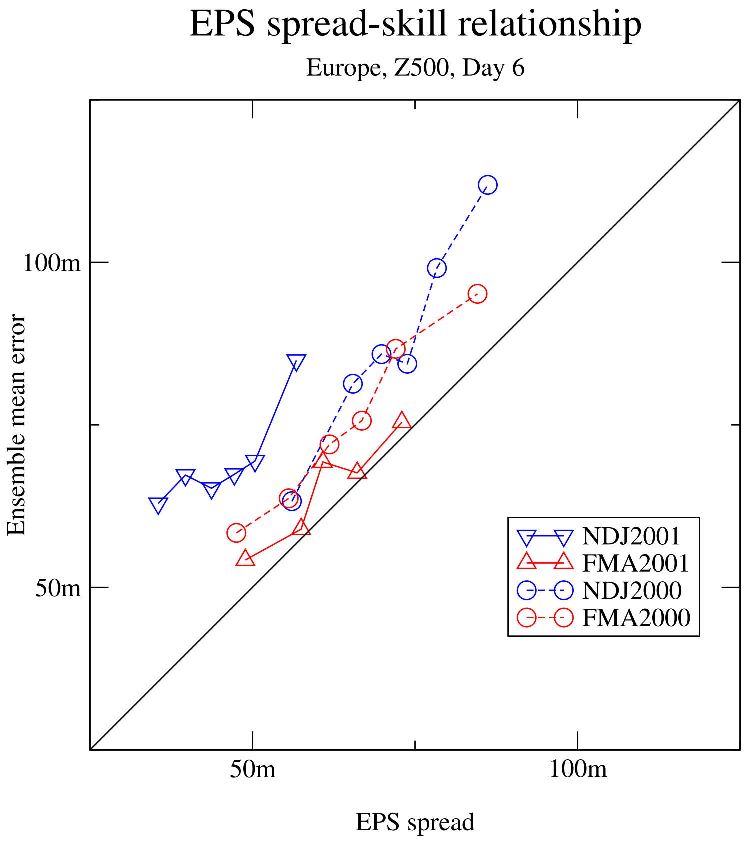


Figure 21: Reliability diagrams for D+6 forecasts of 850-hPa temperature anomalies over Europe

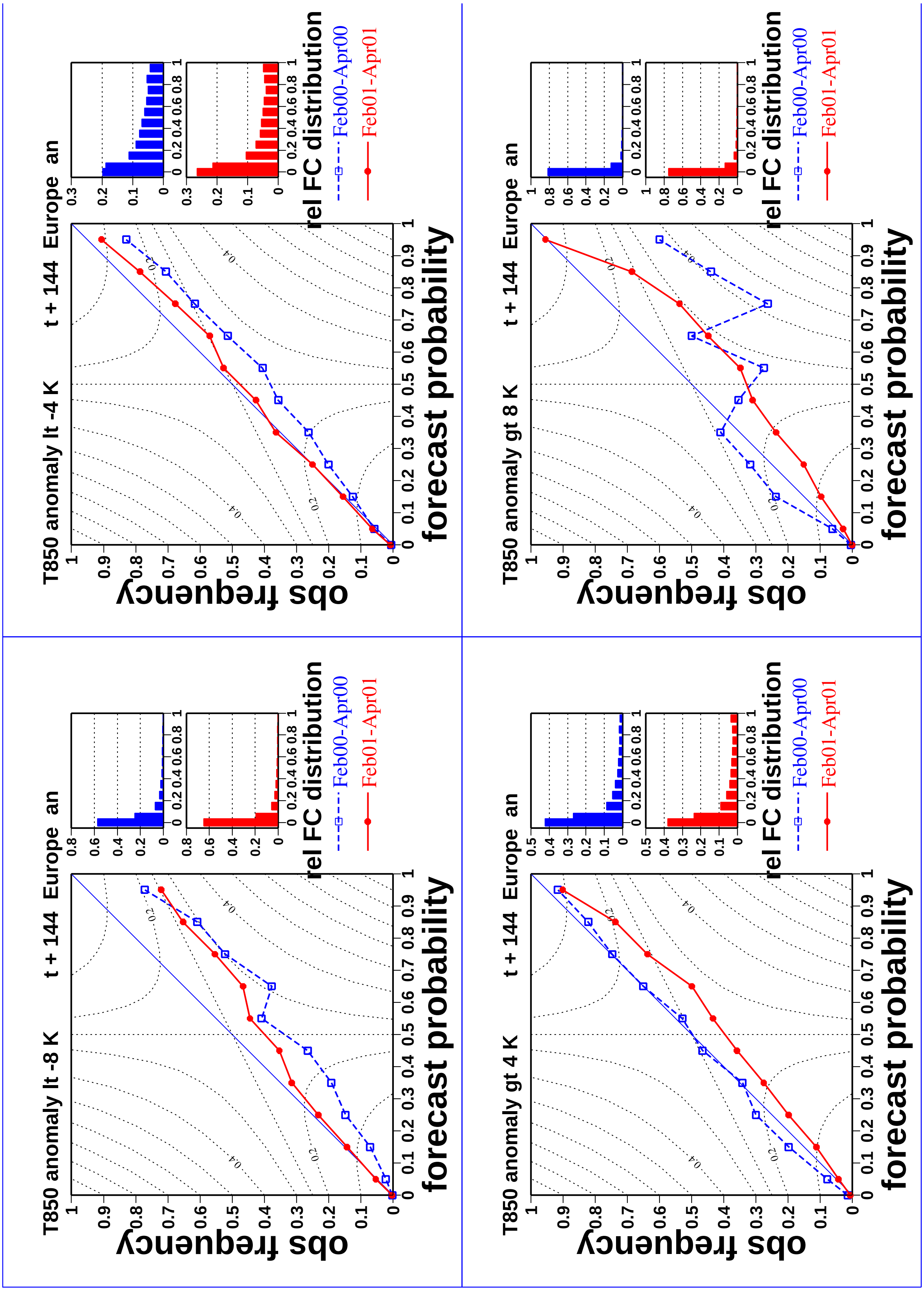


Figure 22: Reliability diagrams for D+6 forecasts of 24-h accumulated precipitation over Europe

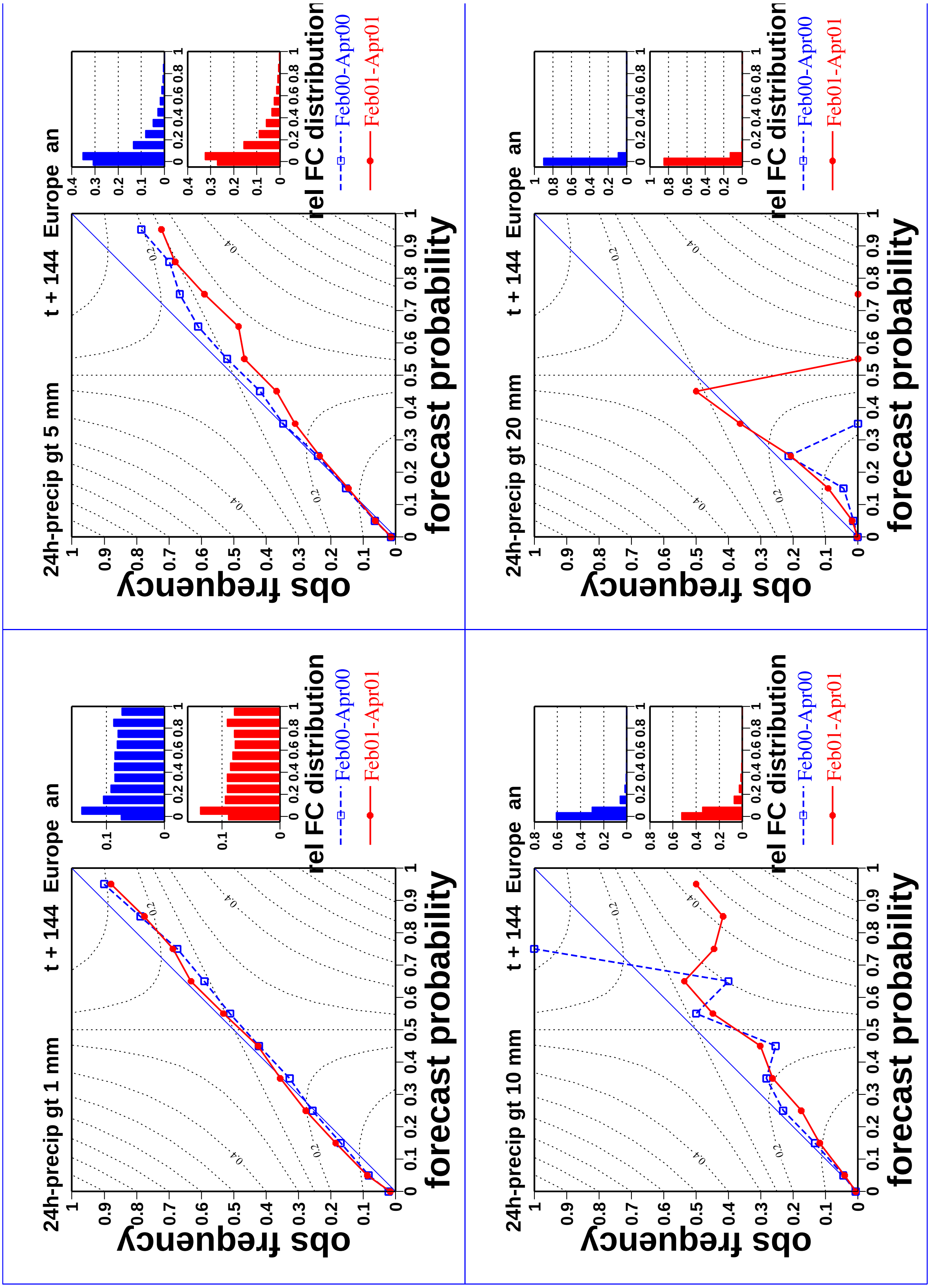


Figure 23: EPS forecast resolution (850-hPa temperature anomalies over Europe)

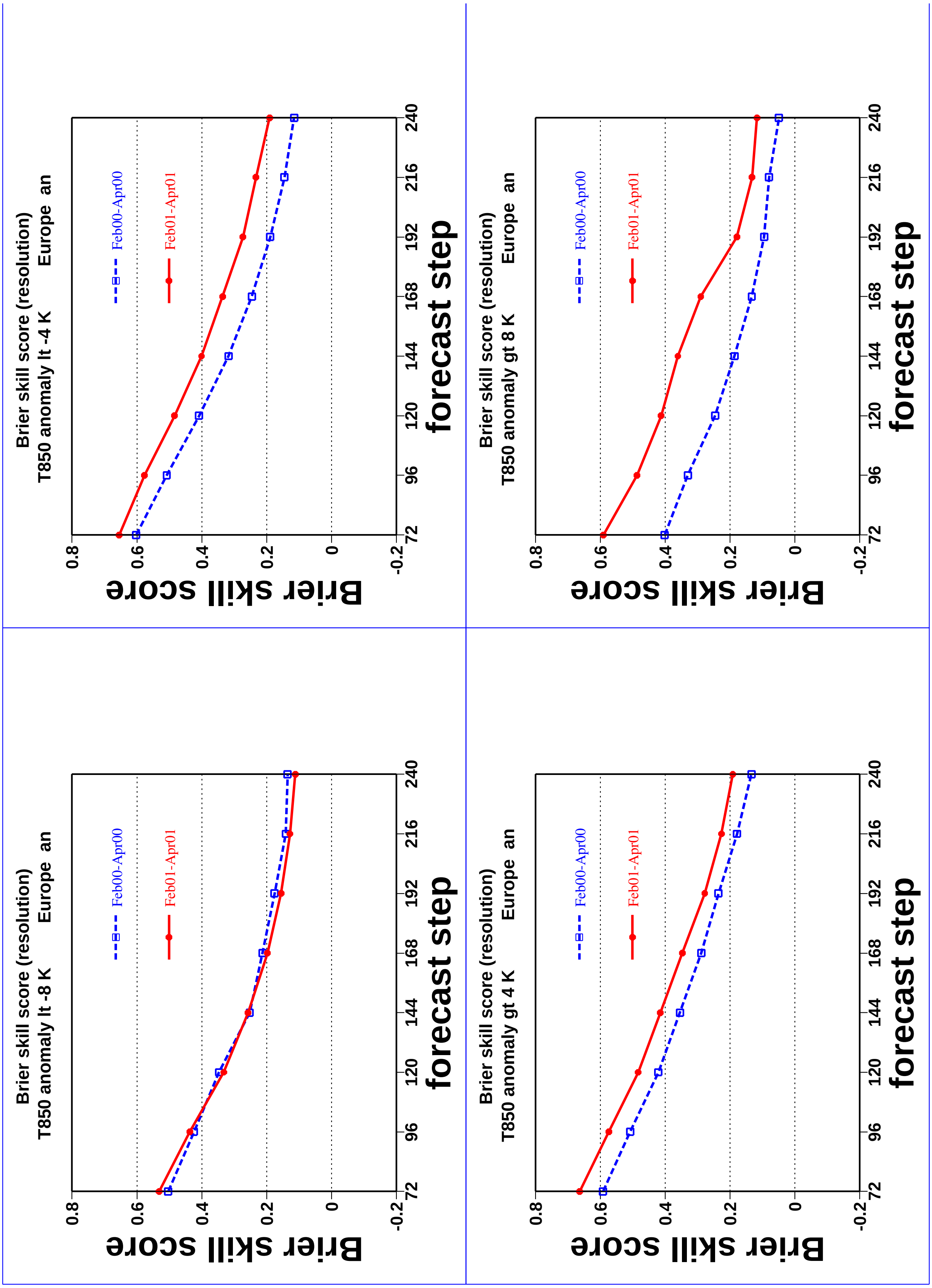


Figure 24: EPS forecast resolution (24h-accumulated precipitation verified with European SYNOP)

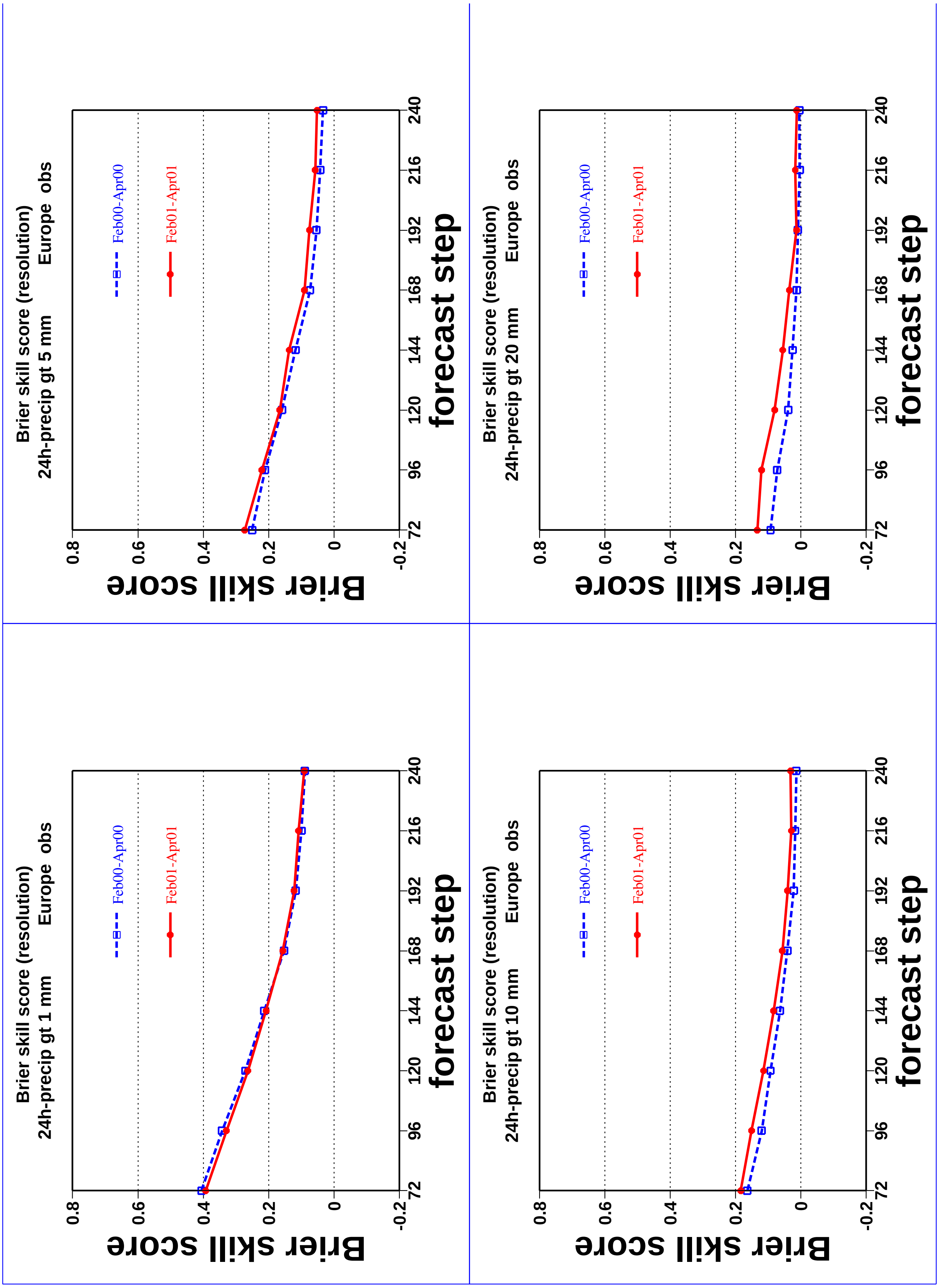
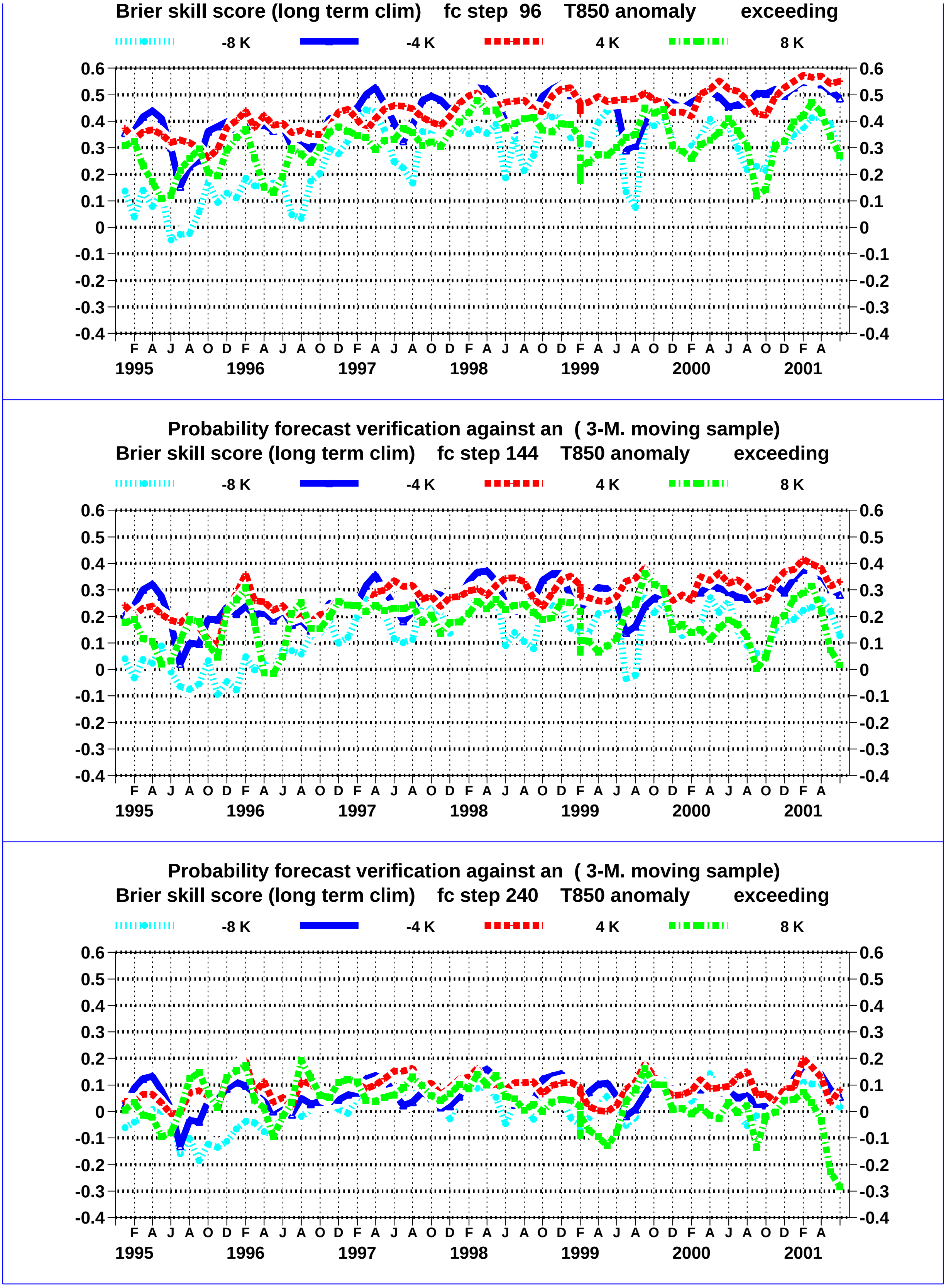


Figure 25: Time series of Brier Skill Scores for EPS 850-hPa temperature anomalies over Europe (data are sampled by 3-month moving intervals);



Brier skill score (long term clim) fc step 96 T850 anomaly exceeding

Probability forecast verification against an (3-M. moving sample)
Brier skill score (long term clim) fc step 144 T850 anomaly exceeding

Probability forecast verification against an (3-M. moving sample)
Brier skill score (long term clim) fc step 240 T850 anomaly exceeding

Figure 26: Time series of Brier Skill Scores for EPS 24h-accumulated precipitation (data are sampled by 3-month moving intervals); verification is the 0-24h precipitation from the EPS control

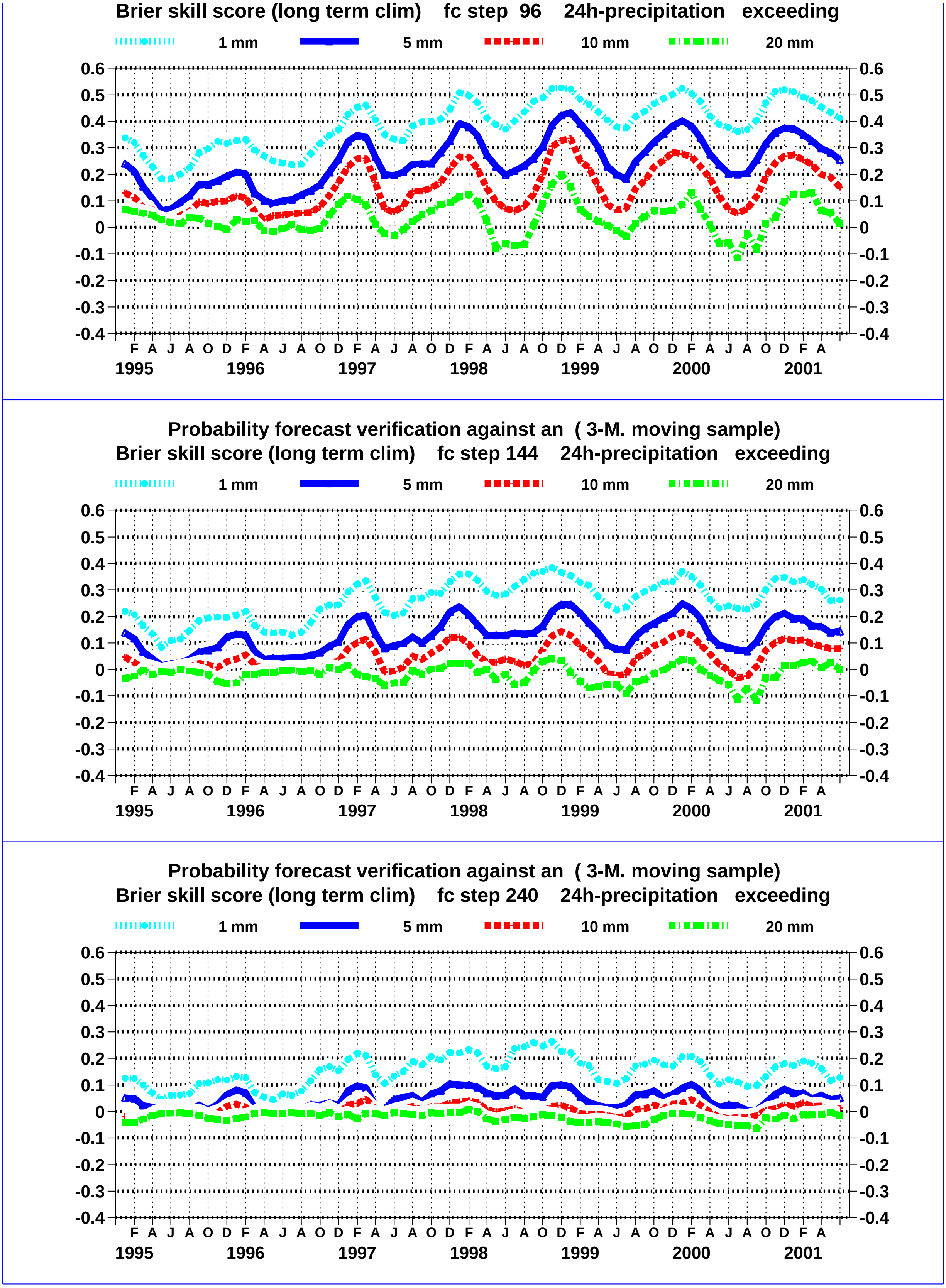
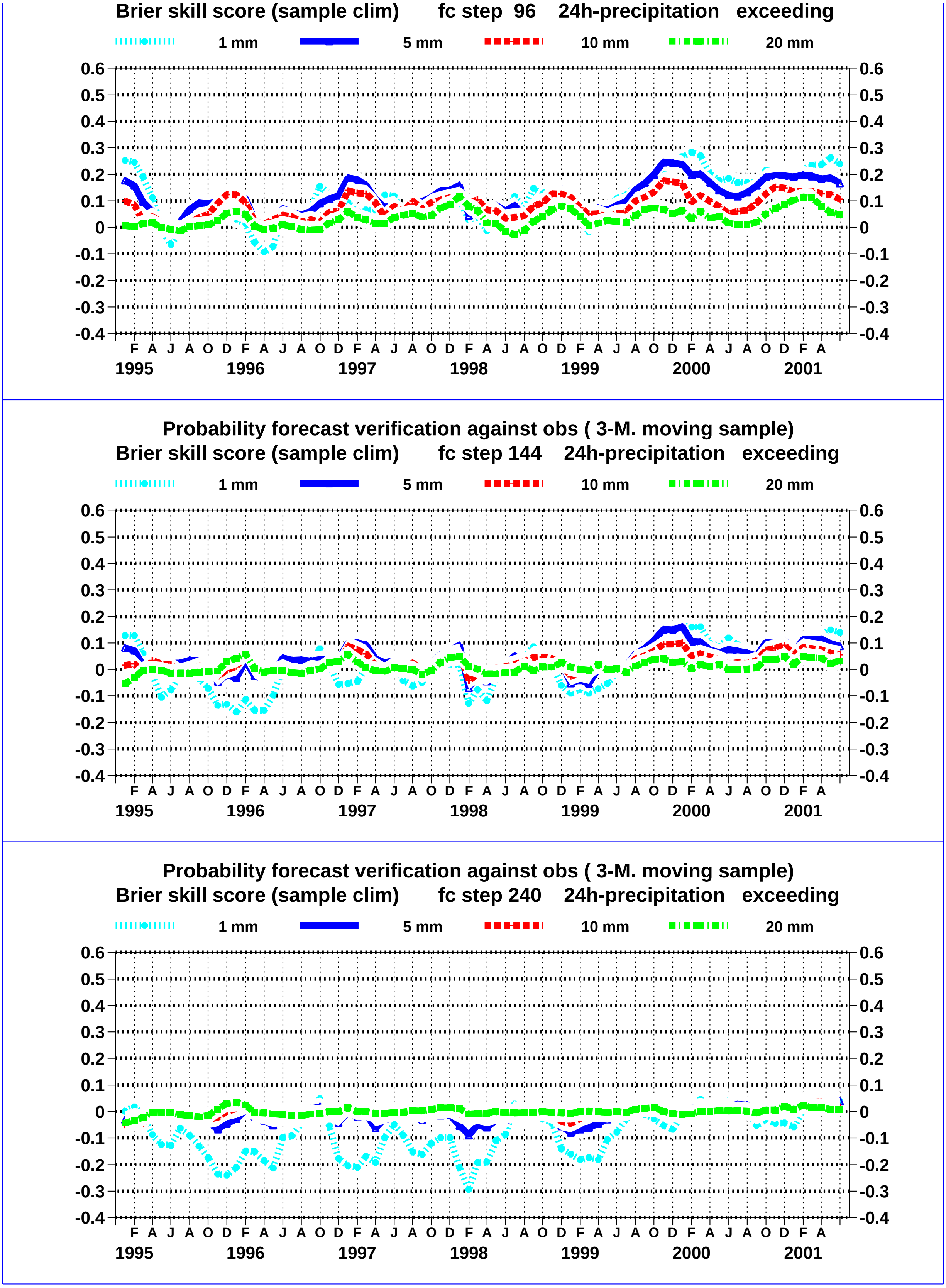


Figure 27: Time series of Brier Skill Scores for EPS 24h-accumulated precipitation (data are sampled by 3-month moving intervals); verification is using European SYNOP observations



Brier skill score (sample clim) fc step 96 24h-precipitation exceeding

Probability forecast verification against obs (3-M. moving sample)
Brier skill score (sample clim) fc step 144 24h-precipitation exceeding

Probability forecast verification against obs (3-M. moving sample)
Brier skill score (sample clim) fc step 240 24h-precipitation exceeding

Figure 28: Brier Skill Scores for Precipitation > 20mm; Red: reference 12UTC run; Black: combining 12UTC run (3/4) and 00UTC run from the same day (1/4); blue: using the 00UTC run from the next day; reference period is April-May 2001, area in NH (North of 30^{o}N)

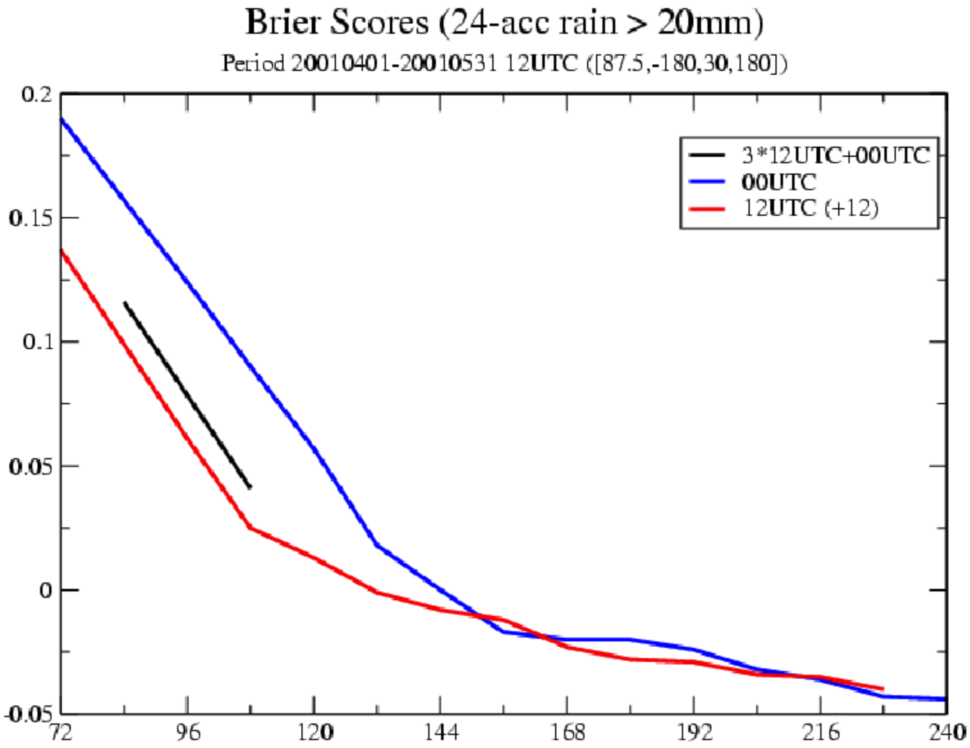


Figure 29: Plume diagrams of Niño-3 SST anomaly for several ensemble forecasts. a) forecast values of monthly mean anomalies from individual members (red curves), b) Reynolds' verifying analysis where available (dashed blue curve)

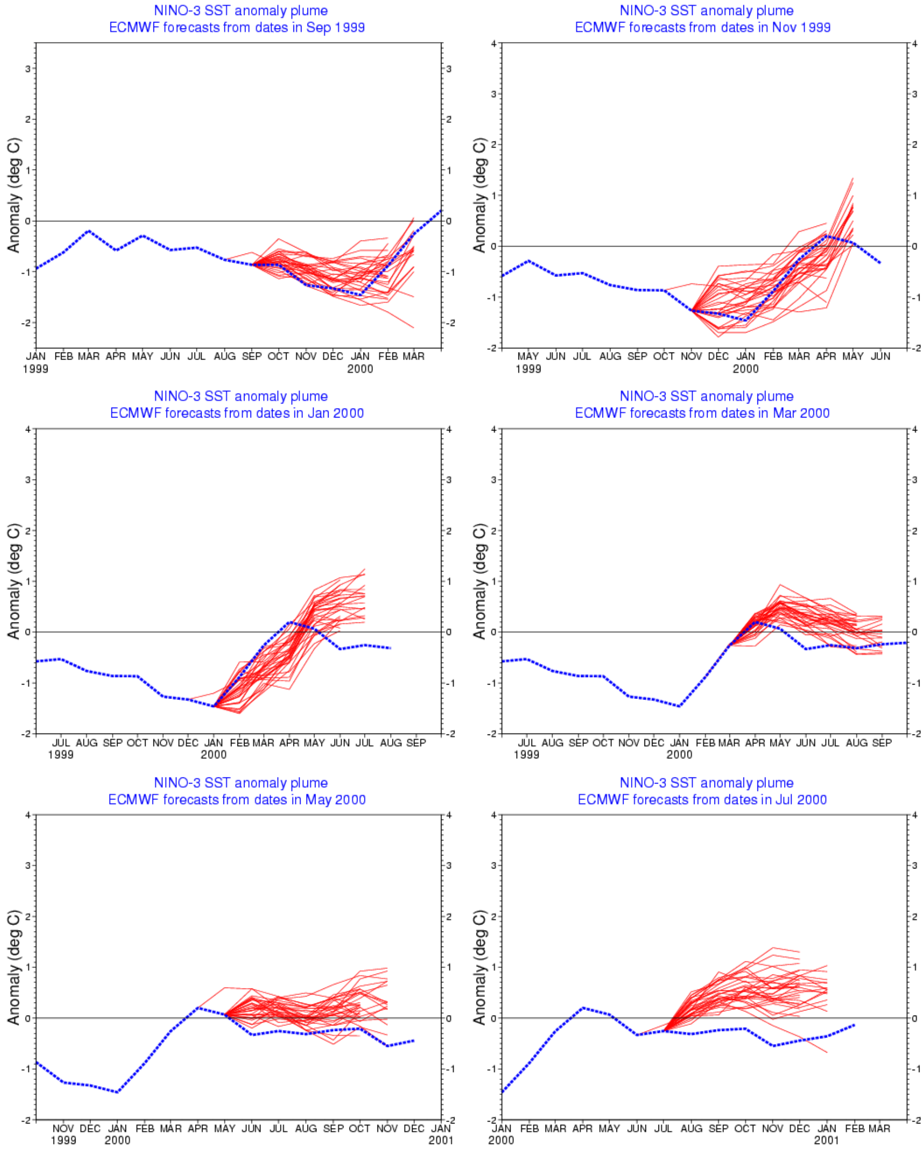


Figure 29 (continued): Plume diagrams of Niño-3 SST anomaly for several ensemble forecasts. a) forecast values of monthly mean anomalies from individual members (red curves), b) Reynolds' verifying analysis where available (dashed blue curve)

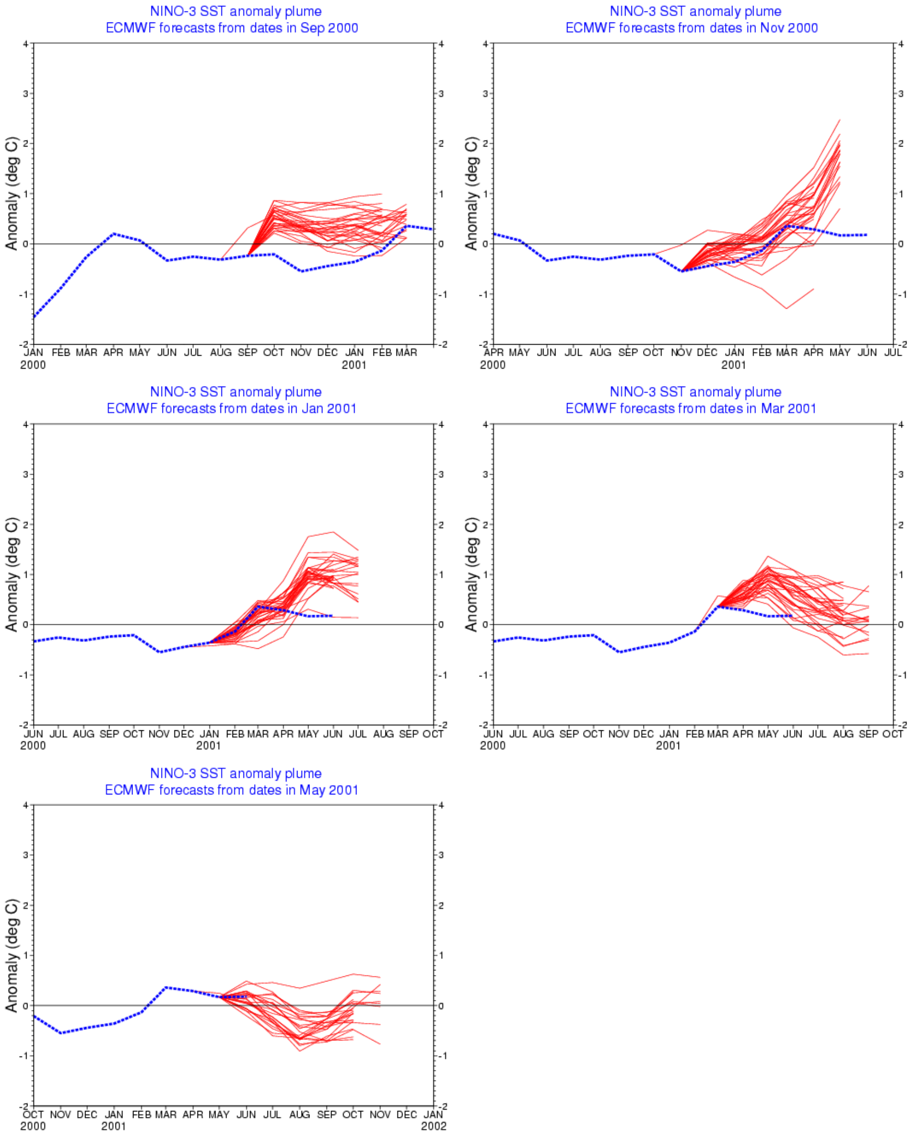
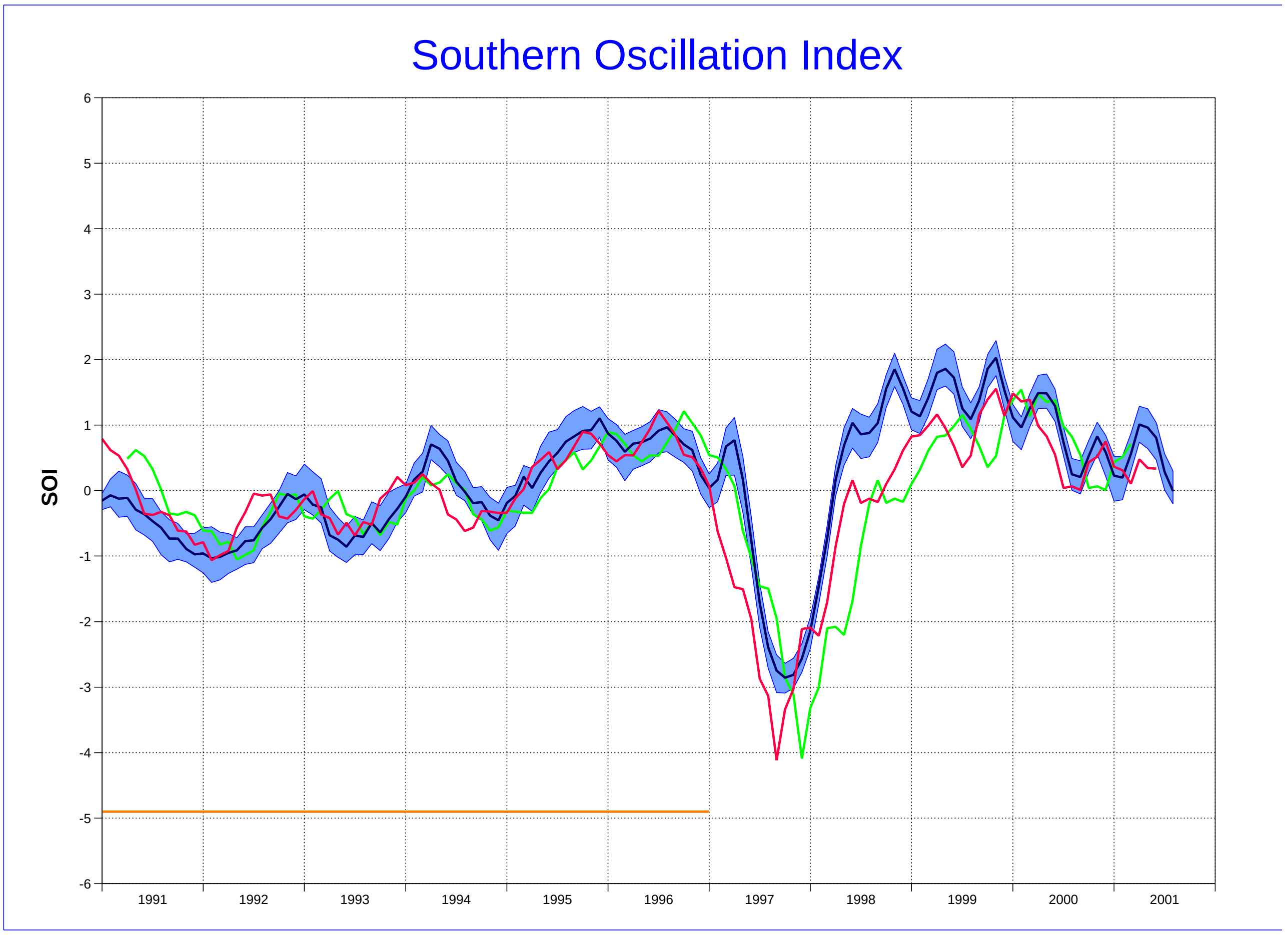


Figure 30: Three-month running mean of Equatorial Southern Oscillation Index (SOI): ECMWF operational analysis (red line), median (blue line) and intervals between 0.25 and 0.75 quartile for each ensemble forecast (light blue band), analysis values persisted for 3 months (green line). The forecast anomalies and the verifying anomalies are all computed with respect to model climate and observed climate respectively for the 1991-1996 base period..


Southern Oscillation Index
SOI