TECHNICAL MEMORANDUM

463

Verification statistics and evaluations of ECMWF forecasts in 2003 - 2004

F. Lalaurette, J. Bidlot, L. Ferranti, A. Ghelli, F. Grazzini, M. Leutbecher, J.-E. Paulsen and P. Viterbo

Operations/Research Department

May 2005

This paper has not been published and should be regarded as an Internal Report from ECMWF.
Permission to quote from it should be obtained from the ECMWF.

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

© Copyright 2005

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

Literary and scientific copyrights belong to ECMWF and are reserved in all countries. This publication is not to be reprinted or translated in whole or in part without the written permission of the Director. Appropriate non-commercial use will normally be granted under the condition that reference is made to ECMWF.

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 presents recent verification statistics and evaluation 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 2, including, when available, a comparison of our forecast performance with other global forecasting centres. Section 3 deals with the verification of ECMWF weather parameters and oceanic wave forecasts. Section 4 has been added this year to describe two "severe weather" studies - one on the forecast performance when dealing with weather patterns associated with severe floods over the southern alpine region, and the other on tropical cyclone forecast accuracy. Finally, Section 5 provides insights into the performance of the seasonal 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 help compare the performance year by year (ECMWF Tech. Memos. 346, 414, 432). Aspects related to experimental products are treated in separate documents.

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 since the preparation of documents for the last meeting of the Committee is as follows:

- **7 October 2003: Cy26r3, a major model upgrade**
  - new humidity analysis; new data streams (AIRS from Aqua, AMSU-B, AMSU-A from Aqua, Japanese wind profilers, Meteosat5, GOES9 and GOES12 WV CSR, GOES12 winds and MIPAS ozone-profile retrievals; ENVISAT global altimeter data for wave assimilation to replace ERS2, the coverage of which has reduced since July 2003);
  - new linear radiation scheme in 4D-var; a new radiation sampling (HALO) and a new aerosol climatology; a relaxation of the convective mass flux limiter for long time steps (used for the EPS and monthly forecasts);
  - monitoring of ENVISAT data: SCIAMACHY, GOMOS, and MIPAS; new model parameters (UVB, CAPE, photosynthetically active radiation, freak waves).

- **20 Oct: The Extreme Forecast Index (EFI) was added to the list of archived and disseminated products;**

- **28 Oct: NOAA-17 AMSU-A instrument operations stopped.**

- **17 Dec: Re-introduction of Met-7 CSR after a processing change at EUMETSAT had created an interruption;**

- **18 Feb: Blacklisted NOAA-15 AMSU-A channel 6, due to instrumental drift;**

- **9 Mar: A new version of ECMWF model, Cycle 28r1, was implemented, involving several, mainly minor, modifications:**
    - Data assimilation: new snow analysis using NESDIS snow cover product; improved use of GOES BUFR winds; improved clouds in 4D-Var minimisation; partial re-introduction of ERS-2 scatterometer;
    - Numerics: Semi Lagrangian fix for polar vortex instabilities; several code modifications to prepare the L91 version;
    - Physics: convection clean-up; optimisation of linearised physics and more optimisations of physics code;
    - Oceanic Waves: introduction of subgrid scale (unresolved) bathymetry effects; fix to the EPS wave-model interface (Charnock variable) ;

- **30 Mar: Changes and corrections to the dissemination of weather parameters and forecast probabilities (added wind gusts and corrected snowfall descripton);**

- **25 May: Blacklisted NOAA-16 HIRS radiances;**

- **15 June: a major upgrade of web services, with a substantial increase in forecast products:**
    - 00 and 12UTC forecasts (whichever is the most recent) are given the same status;
    - a 10 day archive of graphical products became accessible online;
    - Tropical Cyclones forecasts and verifying observations were made available to Member States with an archive going back to January 2003;
    - Seasonal forecast Climagrams were introduced for a range of parameters.

- **29 June: The early delivery suite (Cy28r2) was implemented: by shifting the 12h 4Dvar data assimilation window by 6h and running an early additional, uncycled 6h-4Dvar, operational products are now disseminated around 4h earlier without any noticeable impact on the forecast quality.**

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

Verification for free atmosphere medium range forecasts

ECMWF scores

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 2004. The trend for a marked improvement in the quality of the forecasts observed over the last few years is seen again this year. If reference to climatology rather than persistence is used and errors are measured by anomaly correlation rather than root mean square differences (Figure 2) the picture is apparently not so bright. A plateau seems to have been reached, keeping the level of performance to the very high level it reached in 2002/2003, but without much progress beyond that. The fact that these two sets of figures show a somewhat different signal is related to differences in the weather patterns observed in 2003-2004, which had more variability than usual, compared to previous years. Figure 3 shows the evolution of the mean root mean square error over

Europe, made by persisting the analysis as a Day 5 forecast. It can be seen that the cold season 2003-2004 was particularly active, in that weather patterns were more quickly varying over a five day period than in previous years. The last winter with a comparable (but still lower) level of activity over Europe was 1998-99 This was accompanied by a remarkable apparent degradation in the quality of the forecast, when measured by anomaly correlations. This year, despite the large increase in weather activity, no major degradation was found. This will be confirmed later in this report, when comparisons with other forecast centres performance are made (Section 2.2).

Not surprisingly in such active weather conditions, the forecast of temperature anomalies in the medium range was more difficult than during previous winters. Figure 4 illustrates the distribution of the anomaly correlation of the day 6 forecasts of 850hPa temperature with verifying analyses over Europe. The upper left panel of this figure confirms that the deterministic, high resolution forecasts at this range suffered from the large variability of the weather and that the number of good forecasts had decreased for most of the quality thresholds. It is interesting to see, however, that in such difficult situations, the EPS ensemble means (upper right panel) continued to provide increasingly reliable forecasts. Indeed, one of the main reasons for running an ensemble is to provide better forecasts in uncertain environments. It seems that this was the case this winter, even when using such a crude estimator as the ensemble mean. Early results for the summer show that the improvements found in 2003 and attributed at the time to several important model changes have been confirmed and amplified in 2004. The former Achilles' heel of our forecast system, our summer performance (with 1999 having been the worse case), seems to have been strengthened for good.

It was stressed last year that the performance of the EPS suffered from its lower time and space resolution compared to the higher resolution, T511 model. This was most often the case when intense developments occurred, with an important triggering role possibly played by convection. Some of this merely reflects the benefit of higher resolution, but some areas, where the numerics were unnecessarily sensitive to longer time steps, have been identified for improvement. Not all of the proposed changes have been implemented at the time of writing (some should go into operation with Cycle 28r3 this autumn). Although we have seen that this year was more active than the previous one, the gap in performance between the two model versions seems to have been reduced (Figure 5). Case studies to monitor the performance of the two systems did not indicate such systematic differences in behaviour as were reported last year (e.g. forecast of the 27 October 2002 storm). Figure 5 also shows that 00 and 12UTC forecasts, as expected, have very similar levels of quality.

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 conditions, as can be seen from Figure 6, 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 for the upper atmosphere has been recognised by several institutions. Among them WMO/AREP is a keen user of these analyses, which form an important part of the information they use when preparing their bi-weekly WMO Antartic Ozone Bulletins. Following the improved quality of the analyses, over the last few years we have seen a step-by-step improvement in the forecasts. The time series of scores computed as part of the routine evaluation of the system for 50hPa height in the Northern extratropics is shown, for example, in Figure 7 - the record breaking level of performance reached in 2003 has almost been equalled this year.

During the course of this year the first steps towards a significant upgrade of the verification software in use at ECMWF for both deterministic and probabilistic applications have been taken. The aim is to have a multi-purpose, open design system, where contributions from research and operations, both from in-house and possibly beyond, in the Member States, will be easy to integrate. One of the first applications was the generation of EPS verifications, requested as part of the recently agreed exchange of EPS verification scores among WMO/CBS producers (see section 2.2.2 below). Data for this purpose have been generated on more domains and at higher resolutions than before. As an example, time series of Brier Skill scores and ROC areas over the full Northern Extratropics at full EPS resolution (50 members instead of the previous 10 probability sub-categories) are shown in Figure 8. The first signal clearly visible on these curves is a general trend towards improvement over the last five years. Using a larger domain for verification (Northern Extratropics instead of Europe) helps to give a steadier signal. The very clear improvement in 2001 compared to 2000 corresponds to the correction in early 2001 of an underscaling of EPS perturbations that had been introduced in June 2000. Beyond this period, the steady improvement is to be found both in Brier Skill Scores and ROC areas. In both cases, though, forecasts for cold anomalies are less successful than for warm ones.

In order to offer verification in a domain of particular interest for most applications in our Member States, Figure 9 shows the EPS skill at forecasting large anomalies (8K and more) over Europe in the last few years. Results are mixed: while the forecasts for warm summer anomalies seem to have improved, 2004 having the highest scores of the last four years, cold winter anomalies seem to have deteriorated, results that seem broadly consistent with the time series of Day 6 scores over the Northern Extratropics (Figure 8). It is worth noting that May-July 2004 was unusual in that the frequency of very warm anomalies over Europe was much lower than usual, the frequency in 2003, by contrast, being much higher (2.2% in 2004, 4.5% in 2003)[1].

2.1.1 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 10. 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.

2.2 ECMWF vs other NWP centres

2.2.1 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 11 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 over last year's. In July 2004 errors were so small as to break records at both mean sea level and mid-tropospheric (500 hPa) level. The fact that other centres also broke their own records for the same month is, however, an indication that at least part of the credit is due to weather conditions favourable to these particular measures of performance. It should be remembered, however, that it is only recent ECMWF forecasts that have maintained our lead over the summer months - indeed summer 2003 was the first when our errors were the smallest on a month-per-month

[1] For such infrequent events, the frequency (*o<<1*) is in good approximation with the uncertainty *o(1-o)* - a value that can be found for each year in the caption of Figure 9

basis. Results for summer 2004 seem to confirm this outstanding result. The gap is even more striking in the Southern Hemisphere (Figure 12): that the curves for different forecast steps overlap each other simply reflects the fact that ECMWF forecasts can be compared to forecasts originating 48 hours later in some other centres.

WMO exchanged scores also include verification against radiosondes over smaller areas such as Europe. Figure 13, 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 14. The marked improvement in these scores compared to other centres last year has been consolidated, with ECMWF 850hPa tropical winds now showing the smallest errors of all centres, while at 250hPa the forecasts share the lead with the Met Office (UK).

2.2.2 Ensemble Prediction System

The regular exchange of scores that has operated under WMO/CBS co-ordination for deterministic models since the early 1990s started this year for EPS forecasts. The Japan Meteorological Agency (JMA) has offered to collect the verification data from participating centres on their ftp site and publish them on the web, under password protection for the time being. ECMWF is the first centre to join JMA in the comparison, sending data back to January 2003. An example of how these data are processed and published in real time is given in Figure 15. We hope that other centres will join soon, so that more can be learned from comparing the performance of operational ensemble systems.

3 Weather parameters and oceanic waves

3.1 Deterministic forecast

Figure 16 shows the monthly mean and standard deviation of the 2m temperature and specific humidity errors over Europe to July 2004, verified against synoptic observations (a correction for the difference between model and true orography was applied to the temperature forecast error). The trend towards a reduction in nighttime temperature errors has been followed again this year , while the negative bias that was noted last year in spring has not recurred. A measure of skill that uses the persistence forecast as a reference is presented in Figure 17. It appears that the skill has increased again this year, although only by a small amount and not including July. Figure 18 shows monthly bias and standard deviations from observations for total cloud cover and 10m wind speed forecasts. Here again the main signal seems to be a consolidation of the good performance last year.

Monthly mean error scores for precipitation forecasts at day 3 over Europe are shown in Figure 19, for 00, 06, 12 and 18 UTC. Although smaller than a few years ago, it seems that the summer daytime bias is on the increase. This may be related to several changes that have been introduced to the convection scheme over the past couple of years. These changes have been very beneficial to the system, in particular significantly reducing the sort of imbalances during the data assimilation that led to some very bad forecasts over the summer. Skill scores for precipitation have a positive trend overall, (Figure 20), autumn 2003 and spring 2004 having scored very well for moderately strong events - i.e. beyond 5mm (not shown) and 10mm/day. Summer 2003, though, was slightly worse than in previous years - something that was quite difficult to identify as a forecast problem, due to the highly unusual, dry conditions that have prevailed over much of Europe at that time. The slight overestimation of the summer daytime convection is, however, something that must be investigated.

One way to make a firmer statement is to use a more elaborate verification method than the rather crude comparison to isolated weather station reports. A WGNE initiative, aimed at providing consistent comparison of precipitation forecasts has now reached the stage at which several meteorological services are providing validation of various models (including ECMWF), using their high resolution observation network upscaled to a scale relevant for global forecast models. As an example,

Figure 21 shows results from the comparison over France during the warm season 2003 for a wide range of precipitation thresholds (from 0.1 mm/day up to 16 mm/day) with respect to the probability of detection (number of good positive forecasts over total number of occurrences of the event) and the false alarm ratio (number of wrong positive forecasts over total number of positive forecasts). Good forecasts are close to the upper left corner of the diagram, while the distance from the diagonal indicates the frequency bias. ECMWF compares extremely well with other centres on this graph and there is no sign of any significant overestimation of the number of rain events in the model (positive frequency bias). In order to get similar results over a larger domain such as Europe, ECMWF will consider joining the WGNE comparison, which will allow direct access to the other model forecasts that will then be evaluated, using the data already provided by our Member States and used for our own verification purposes.

Finally, verification scores from the global oceanic wave products are shown in Figure 22 and Figure 23. The trend for improvement is very clear, as a result of progress in the quality of the winds used as a forcing and improvements made to the oceanic wave model itself. Comparison to the analysis for oceanic waves is, however, likely to give an overly optimistic picture, when major sources of observations are lost, as was the case from July to October 2003, with the loss of global altimeter coverage from ERS2. The unusually small errors in the 24h forecast compared to the analysis Figure 22 (NH RMSE) during that period are likely to be related to this lack of independent observations. Further evidence of this can be found from direct comparison of the analysis to independent buoy observations (Figure 24). Clearly the departure from buoy data has been larger during summer 2003, when satellite altimeter observations were missing. This affected our forecasts as well (Figure 25). The comparison with other models in this latter figure is, however, favourable to ECMWF throughout the period. The global altimeter coverage resumed in October 2003, thanks to the use of ENVISAT data, and model improvements in March 2004 are expected to have contributed to a further reduction in errors since then.

3.2 EPS forecasts

In last year’s report (ECMWF Tech. Memo 432), a new set of diagrams for the verification of the relation between spread (as seen, for example, on an EPSgram) and skill was produced. The idea is that as an “EPSgram box” (second and third quartiles of the ensemble distribution) highlights the range of values taken by 50% of the ensemble members, a perfect ensemble would be one where verification lies in the box with the same 50% proportion. Making an assumption, that on average the second and third quartiles are symmetrical and equal to half the size of the (Q75-Q25) “box”, the absolute difference between the ensemble median (Q50) and the verifying observation should then exceed (Q75-Q25)/2 in exactly 50% of the cases (Figure 26).

Eight weather stations that have a reasonably similar climate have been selected over Europe (Figure 27). The scatter diagram for the EPS spread and absolute errors at these stations is given in Figure 28 (left panels) for Day 6 forecasts of 2m temperature and daily rainfall last winter. Conclusions are very similar to last year’s: there is remarkably good agreement between spread and skill. Large day-to-day variations happen, of course, within this relation, but the statistical relationship that should exist, when gathering a large sample of cases with a similar spread, holds. If spread categories populated with a reasonable number of cases are defined, then the distribution of errors within each spread category is centred almost exactly around the expected value (scatter diagram diagonal). The spread in temperature at Day 6 is, however, slightly underestimated for most spread categories, with the exception of the largest spread categories, that are slightly overestimated - but in that case, the sampling is probably not sufficient to obtain a robust result (too widely differing spread values gathered in a single category)

The assumption that positive and negative error distributions are similar is probably not well founded in this verification. In order to relax it, scatter diagrams in which positive and negative errors for rainfall are treated separately are shown in the right panel of Figure 28. For positive errors (observations below the EPS median), the distribution should be centred around Q50-Q25, while for negative errors (observations above the median), it should be centred around -(Q75-Q50). Once again, this is the case, in very good approximation (see both diagonals on the diagram). One of the possible remaining drawbacks in this comparison is that a strong link between the forecast value and the spread already exists in the error statistics - meaning that simple Model Output Statistics (MOS) might well be able to show good verification properties, based on a single deterministic model, instead of a full dynamical ensemble. More work clearly needs to be done to clarify this point. Although results in winter look quite good, it was noted in last year's report (ECMWF Tech. Memo 432) that the distribution for summer amounts of rain showed an overestimation of the spread. This was attributed then to possibly overactive stochastic physics. Research this year has investigated this issue further. By comparing the distribution of daily precipitation, with and without stochastic physics, using upscaled high-resolution network observations as a reference (Figure 29), it has been possible to confirm that most of the overestimation is indeed associated with stochastic physics. A refined formulation of the stochastic physics is currently under testing. Results from one of the possible configurations considered for operational implementation are shown in Figure 29. It seems to provide more realistic results for daily amounts of rain, while keeping most of the other desired properties, such as the representation of the uncertainty associated with model errors (not shown). Finally, Figure 30 provides a probabilistic evaluation of EPS forecasts of precipitation in the usual form, using Brier skill scores and ROC area time series over Europe. The last 12 months show a good performance, most notably for precipitation thresholds 5mm and higher, for which both the summer and winter scores have been the best on record (with the exception of winter 1997-98).

4 Severe weather

4.1 Weather patterns associated with severe Alpine floods

The main motivation for this study was the common criticism made when statistics such as those produced in this report are given: does it make sense to provide verification statistics that mix weather situations having little impact on human activities with the few that lead to severe weather?

One of the easiest weather scenarios to identify as leading to severe weather over Europe is the southerly flow advecting warm and humid air from the Mediterranean Sea in the autumn. When such air masses are lifted over the Alps, severe convection that leads to heavy rain and torrential floods can occur.

Since the weather pattern associated with such events is remarkably consistent, it is relatively easy to generate a composite from significant cases that retain the main synoptic ingredients associated with the event. Such a composite, built from identified cases in 1993 to 2003, is shown in Figure 31. The next step was to identify, in the ERA40 archive, those cases that correlate best, both for anomaly correlation and RMSE , with such a pattern. Inspection both case-by-case and of composite precipitation charts (Figure 32) reveals that the weather pattern that was retained is one that is frequently associated with severe weather over the region.

The next step was to question the performance of the ECMWF operational model at forecasting these events, compared to “average weather”. The results in Figure 33 show that such situations are certainly not those when the forecast performs worst. Moreover, the trend for improvement in severe weather forecasts was bigger than for the average ones, both for RMSE and anomaly correlation (with the exception, in the latter case, of forecasts beyond 7 days). Although this is certainly not the final word on the controversial issue of the assessment of severe weather forecasts, it was considered a useful contribution to the debate on the impact of NWP developments in this area.

4.2 Tropical cyclones

Verification of Tropical Cyclone (TC) forecasts is now a routine activity that has recently been given a higher profile due to the release of real-time TC forecast products to our Member States on our website. Several hundreds of them have been tracked since 2002 (Figure 34), making it a suitable sample for both deterministic and probabilistic verification.

The first type of verification is a direct comparison between the deterministic T511 and T255 (EPS control) forecast TC track and the best track reported by WMO RSMCs. To these deterministic forecasts, a “consensus” 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 2003-April 2004) are shown in Figure 35. Clearly the T511 forecast is providing the best results, which can be attributed to the impact of higher resolution for such strong events. More surprising, perhaps, is that the consensus EPS forecast is performing worse than the Control at the same resolution; this may be attributed, at least partly, to the fact that there are several cases when the EPS forecast exhibits multi-modal scenarios, for which the ensemble mean can only perform badly.

A probabilistic verification of the “Strike probability” product offered on the web (probability at any geographical point that a reported TC will get closer than 120km within the next 120h) is shown in Figure 36. While the forecasts remain overconfident for large probability thresholds, as already reported last year, the forecasts have been more reliable overall. More importantly, their resolution (capability to detect a large number of events with a relatively moderate number of false alarms) has also improved. Many changes that can claim some responsibility for this improved performance have been introduced into the data assimilation/ forecast system over the last couple of years, although it has not been possible to isolate any one in particular as the most significant.

5 Seasonal forecasts

5.1 The 2003-2004 El Nino forecasts

During the early months of 2003 the warm sea surface temperature anomalies over the equatorial Pacific steadily decreased. Since April 2003 oceanic conditions have been near to normal. Figure 37 shows Nino-3.4 predictions throughout the year with subsequent verification (heavy blue dashed line). In general, the forecast over the Nino areas verified well. For December and March the observations are outside the predicted range, indicating that the system was overconfident on that occasion. Insufficient spread was confirmed over the Nino-4 area (not shown). The latest, yet to be verified, El Nino forecasts present a relatively large spread. All members, however, forecast warm SST anomalies. Typically, the El Nino onset is observed during the first months of the year. However, in the past, a number of moderately warm episodes

(Nino3.4>0.5K) began in late summer, well after the northern spring barrier - which is what the latest forecasts support as a scenario for this year.

Since December 2003 tropical intraseasonal variability has been intense. Strong westerly wind anomalies, propagating from the Indian Ocean to the Pacific, were observed in December 2003, March and May 2004. Initiated by westerly anomalies, a number of oceanic Kelvin waves propagated eastward (see Figure 38). Although El Nino events have, in the past, intensified in response to such events, this was not the case last year.

5.2 Seasonal Forecast performance during 2003-2004

Summer 2003 over Europe was one of the hottest on record (Schär *et al*. 2004; Grazzini *et al.*, 2003) and it is therefore of great interest to document the seasonal forecast predictions for such an extreme event. In a large area, mean summer temperatures exceeded the 1958-2001 mean by ~3C, corresponding to an excess of up to 4 standard deviations (Figure 39, upper panel). The lower panels show the probability given by two successive forecasts that 2m-temperature will be above normal during the summer (upper tercile of the climate distribution). While many of the probabilities over France lie in the range of 50-60% during the May forecast (left), this signal was not there a month earlier (right). During the last 2 weeks of April the Mediterranean basin warmed quite rapidly. It is possible that the May forecast, by persisting this SST anomaly, produced a better signal. However, the warm conditions over the Mediterranean did not help the forecast initiated in June to make realistic predictions for the July to September period either (not shown).

It is important to note that SST predictions were generally realistic in persisting the warm SST anomaly over the Atlantic Ocean. The North Atlantic SSTs have been considerably above average during the past year. Since April they have remained above 2 standard deviations across the high latitudes and also across large portions of the Subtropics. This warm condition seems to be associated with an ongoing warm phase of the Atlantic multi-decadal mode. Predictions for DJF 2003/04 successfully reproduced the ridge and warm anomalies over the North Atlantic, probably due to these long standing warm SST anomalies.

On the other hand, over the Indian Ocean, positive SST anomalies in late spring and summer 2003 were underpredicted. In this area of warm waters, relatively small anomalies (about +0.5 degree) can have a significant impact on the monsoon circulation and, in turn, affect the summer circulation over the Mediterranean basin.

It is difficult to establish to what extent the poor seasonal predictions for the European hot summer are due to model errors or are related to the 'true' low predictability level of this event. Results from an ensemble of simulations with an atmospheric model forced by observed SST conditions (see Figure 40) indicate that even with prescribed oceanic conditions, the event was difficult to predict.

Since ECMWF seasonal predictions will come from a multi-model, ensemble-based seasonal forecast system in the near future, it is interesting to study the performance of the other models. Figure 41 shows the probability pattern for the upper tercile of 2 m temperature from the UKMO ensemble of forecasts started in May 2003, forecasting June-July-August. The warm signal over France is broadly comparable with the one in Figure 39, while the warm anomalies over the Mediterranean are somewhat underestimated. UKMO predictions initiated in June, like the ECMWF ones, did not show the warm signal. Similar inconsistence between predictions initiated in May and those initiated in June was also found in the Météo-France forecasts (André *et al*. 2004).

Considering that the spring and summer of 2003 were very dry, it is possible that the lack of soil moisture contributed to the local heating. A study (Ferranti and Viterbo, 2004) was carried out to establish the soil water conditions and evaluate the extent of the surface feedback and its contribution to the predictability. A brief description of the results obtained is given below.

The typical seasonal and interannual fluctuations of soil water averaged over Central Europe were estimated by using the ERA-40 data. Figure 42 shows that the operational soil water analysis for the period from March to September 2003 was extremely dry in comparison to ERA-40 records. August 2003 was drier than any of the months in ERA-40. Despite the dearth of soil water observations, there was evidence that the ERA-40 annual cycle of soil moisture is too small. In fact, the soil water analysis increments show a systematic wetting in summer. This reduces the annual cycle and makes the soil overly moist in summer.

Since large uncertainties in the analysed soil moisture values can have an impact on the seasonal forecast, particularly on the predictions initiated in spring, numerical experimentation has been used to document the model sensitivity to the soil moisture initial conditions. Several 9-member ensembles of 4-month atmospheric integrations, forced with observed sea surface temperature (SST), were performed. Each ensemble had initial soil moisture between the surface and a depth of one metre set to prescribed uniform values in a large European area. The prescribed soil moisture values ranged from very dry , effectively shutting off model evaporation (soil moisture index, SMI=0), to very wet (SMI=100).

For example, Figure 43 shows the ensemble mean 2m temperature differences between simulations with soil wetness initial conditions prescribed to a value of SMI=25 and to a value of SMI=75. The differences are averaged over the second month of the integrations. The impact of drier soil initial conditions is mainly local and highly significant, even after 2 months. Such a response remains significant in the temperature at 850 hPa and, although over a smaller area, in the geopotential height at 500 hPa.

Due to the lack of soil water measurements, it has so far been impossible to compare the various values of soil wetness used as initial conditions with the ones that were actually present in June 2003. It has therefore been difficult to quantify the real contribution of the surface conditions to the high temperature anomalies observed.

Nevertheless the extensive experimentation has shown that the atmospheric response to large soil moisture initial perturbations extends up to month 2 and is non-linear. The response is larger for drier regimes. Extending the perturbations to the soil below the root zone (to a depth of 2.89 m) increases the atmospheric response and its memory up to 3 months, if the anomalies are large.

In conclusion, it can be said that the anomalous hot European summer of 2003 was difficult to predict more than one month in advance. In fact, it is not yet clear which forcing - if any - was instrumental in maintaining the large-scale, anti-cyclonic circulation for longer than a season. However, the dry soil conditions certainly contributed to amplifying the local temperature anomalies. The large uncertainties in the soil moisture analysis and the atmospheric response to soil water conditions, documented in this study, suggest that further work needs to be done:

i) to improve soil moisture assimilation;

ii) to account for the uncertainty in the initial state of soil water content by introducing properly scaled initial perturbations into the initial conditions.

6 Summary

The forecasting system has again reached very high levels of skill in many areas this year, both on the large scale (500hPa scores over Northern and Southern Extratropics, tropical wind errors at 850hPa and 200hPa) and in terms of weather parameters (good verification of moderately strong - 5 and 10 mm/day - precipitation events) and of tropical cyclone forecasts. Scores for EPS forecasts have started to be exchanged under WMO/CBS co-ordination, using a new package for verification. These results, generated at a higher probability resolution and on larger domains, have shown a remarkable trend of improvement for the EPS forecasts over the last few years, although year to year fluctuations are still to be found.

A study has shown that forecasts of the weather patterns associated with severe convective precipitation over the Alps have improved faster than was the case for “average” situations over the last few years. Verification of the relation between the spread in EPS forecasts for daily rainfall and 12 UTC 2m-temperature and the distribution of errors has shown very good results. It has, however, not yet been possible to establish whether this relation is value added by running a dynamical ensemble, or whether it could be built from simple training with a single deterministic forecast. Some tendency to overestimate convective precipitation has been found this year, although this was not confirmed by independent WGNE verification. For EPS forecasts, it has been possible to link the excessive rainfall to the current stochastic physics configuration, with a hint that the new scheme under evaluation might reduce this problem.

Finally, the performance of the seasonal forecast system has been good in the tropical Pacific. Some more work has been carried out to investigate the possible impact of land surface conditions on the development of the heat wave of summer 2003. Although it is not possible to draw any definite conclusions regarding the predictability of such events a few months in advance, it seems that a better account of uncertainties related to the initial soil water conditions might have helped, at least to avoid overconfident forecasting in such cases.

References

André J-C, M Déqué P Rogel and S Planton, 2004: La vague de chaleur de l’été 2003 et sa prévision saisonnière. *C.R Acad. Sciences* vol 336, **6**, pp491-503

Black, E., M. Blackburn, G. Harrison, B. Hoskins and J. Methuen, 2004: Factors contributing to the summer 2003 European Heat Wave. *Weather,* 59, p217-

Ferranti, L. and P. Viterbo, 2005: The European summer of 2003: sensitivity to soil water initial conditions. *J. Climate* (submitted).

Grazzini, F., L. Ferranti, F. Lalaurette and F.Vitart, 2003: The exceptional warm anomalies of summer 2003. *ECMWF Newsletter No.* **99**, pp2-8

Lalaurette, F., L. Ferranti, A. Ghelli, and G. van der Grijn, 2003: Verification statistics and evaluations of ECMWF forecasts in 2002-2003. *ECMWF Tech. Memo*. **432**

Schär C., P.L. Vidale, D. Luthi, C. Frei, C Harbeli, M.A.Liniger and C. Appenzeller, 2004: The role ofincreasing temperature variability in European summer heatwaves. *Nature* .**427**, pp332-336

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 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 (Figure 14, Figure 13) root the sum of the mean squared errors for the two components of the wind independently;

Skill scores (Figure 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:

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

Figure 2, Figure 4, Figure 5 are correlations in space 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. Only for oceanic waves (Figure 22and Figure 23) has the climate been derived from the ECMWF analysis.

A.2 Probabilistic forecasts

Events usually defined for the verification of medium-range 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. Probabilistic skill is illustrated and measured in this report in the form of Brier Skill Scores and the area under Relative Operating Characteristics (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, takes only 0 or 1 as values). For a single event, it can be written as:

$$BS = (p - o)^2$$

- As for 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 Brier Skill Score is defined as:

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

Time series of the Brier Skill Scores can be found in Figure 8 and Figure 30, while variations with forecast range are shown in Figure 9. The Brier score can be split between the uncertainty (0 if the event occurs with frequency 0 or 100%, 0.25 if the event occurs with frequency 50%), reliability (how close the conditional frequencies of occurrence are from probabilities) and resolution (how different the conditional frequencies of occurrence are).

Relative Operating Characteristics 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. 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 8 and Figure 30. Its variation with forecast range is shown in Figure 9.

A.3 Weather parameters (Section 3)

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).

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 29).

Reference

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).....17

Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the curves is the forecast range when the monthly average of the daily forecast anomaly correlation with observation (analysis) is falling below 60% for Europe, Northern and Southern Extratropics (the red curve is the 12-month moving average) 18

Figure 3: Root Mean Square Error made by persisting the analysis over 120h and verifying it as a forecast. 6-monthly averages (red curve) confirm that the last cold season (October to March) has been unusually active over Europe..... 19

Figure 4: Cumulative distribution of Anomaly Correlation of the Day 6, 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (1-May-10 Aug., bottom) since 1984-85 for the deterministic, high resolution forecasts (left panels) and since 1997-98 for the EPS Ensemble mean (right panels). 19

Figure 5: Comparison between near surface pressure height (1000-hPa) scores from T511 (ECMWF, 12UTC in red, 00UTC in brown), T255 (CNTRL, blue) and EPS ensemble mean (green) during the cold season (15 Oct.-15 Apr.) over Europe. Left: 2002-2003; Right: 2003-2004..... 20

Figure 6: RMS of the difference between 24h-consecutive 500hPa height forecasts verifying the same day over Europe (left panel) and Northern Extratropics (right panel)..... 21

Figure 7: Model scores in the extratropical Northern Hemisphere stratosphere (50hPa height Day 1 and Day 5 forecasts RMSE) 21

Figure 8: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics Area (ROCA, lower panel), the later showing the skill shown by the EPS at detecting a signal verified by the analysis out of the Day 6 probability forecast of 850hPa anomaly temperature (0.5 is no skill, 1 is a perfect detection). Both panels show monthly scores (thin dotted lines) and 12-months moving averages (full heavy lines). Verifying area is the Northern Extratropics (20^{o}N and beyond) 22

Figure 9: EPS skill at forecasting strong temperature anomalies (8K or more) over Europe. Upper panel: cold winter anomalies. Lower panel: warm summer anomalies...... 22

Figure 10: Model scores in the Tropics (root mean square errors for 200hPa and 850hPa wind)..... 23

Figure 11: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)..... 24

Figure 12: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)..... 25

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 27

Figure 14: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)..... 26

Figure 15: Example of ECMWF EPS verification results as shown on JMA EPS intercomparison website. Upper panels are for forecasts of very cold (by more than 8K with respect to the climate) anomalies in January 2004. Lower panels are for very warm (by more than 8K with respect to the climate) anomalies in July 2004...... 28

Figure 16: Verification against European SYNOP observations of 2m Temperature and specific humidity (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)..... 29

Figure 17: RMSE skill with respect to the persistence forecast for 2m temperature forecasts over Europe during night time (00UTC, top) and daytime (12UTC, bottom) for different forecast ranges. .. 30

Figure 18: Scores against European SYNOPs of total cloud cover and 10m wind speed forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC). 31

Figure 19: 6h-accumulated precipitation forecasts biases (T+54/60/66/72h) with respect to SYNOP..... 31

Figure 20: Time series of Equitable Threat Scores for the forecast of daily precipitation verified using SYNOP reports over Europe; Top: threshold 1mm, bottom: 10mm...... 32

Figure 21: Verification of 54-78h precipitation 00UTC-based forecasts against upscaled observations from the French high-resolution network (2003 warm season); ECMWF is in green. The diagonal is a line of no frequency bias (above is over- , below under- forecasting). The better the forecast, the closest to the upper left corner(all events are detected, all positive forecasts are correct). For each forecast centre, the higher the detection threshold, the poorer the forecast (courtesy from WGNE/Meteo-France) 33
Figure 22: Scores (anomaly correlation and standard deviation) of oceanic wave heights verified against the analysis (Northern Extratropics) 34
Figure 23: Scores (anomaly correlation and standard deviation) of oceanic wave heights verified against the analysis (Southern Extratropics) 35
Figure 24: Comparison of ECMWF wave model analysis with bouy observations (wave height).The Scatter Index (S.I.) is the standard deviation of the errors normalised by the observationscatter........ 36
Figure 25: Verification of different model wave height forecasts using a consistent set of observations.... 36
Figure 26: Schematic description of the spread skill relation that should be found in a perfect probabilistic forecast. 37
Figure 27: Stations used for the verification of the EPS spread (Figure 28) 37
Figure 28: Scatter plot distribution of errors of the EPS median as a function of the spread. Left: Mean Absolute Error compared to the interquartile EPS distance (upper: 2m-temperature; below: daily rainfall). Right: Discriminating Between positive and negative errors for daily rainfall. (Day 6 Forecasts during winter 2003-2004 from 8 stations in the Northern European Plain - see map in Figure 27). 38
Figure 29: Distribution of daily rainfall events over Europe as observed (bold black line, upscaled from high resolution data) and forecasted in the operational EPS (dashed blue), EPS without stochastic physics (full blue) and with a revised stochastic physics scheme (dashed green). The effect of reducing the time step (from 2700s to 900s) is also shown in red. 38
Figure 30: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics curve Area (ROCA, lower panel), the later showing the skill shown by the EPS at detecting a signal out of the 120-144h probability forecast for rain (0.5 is no skill, 1 is a perfect detection). 39
Figure 31: Composite500hPa flow built from 6 severe flood events over the southern Alpine region from 1993 to 2003. 40
Figure 32: Mean precipitation anomaly for the 100 cases correlating best with the composite pattern from Figure 31 in October/November during the ERA40 period. Red contours are scaled with respect to the standard deviation of the climatology 40
Figure 33 : Score comparison between 1980-1992 (blue)and 1993-2003 (red); full lines are regardless of the flow pattern, while dashed lines are for the cases correlating best with thecomposite shown in Figure 31; Upper: Anomaly correlation; Lower: RMSE 41
Figure 34: Number of Tropical Cyclone tracked by the determinist,T511 day 2 forecast from April 2002 to May 2004. For each month, the number is split per WMO Tropical Cyclone region according to the colour code above. Both 00 and 12UTC forecasts are tracked. 41
Figure 35:Verification of Tropical Cyclone forecasts from the deterministic, T511 forecast (blue), EPS T255 Control (red) and mean position/ intensity averaged among all cyclones tracked in each member of the ensemble forecast (green)........ 42
Figure 36: Probabilistic verification of TC strike probabilities: left are reliability diagram (the closer to the diagonal the better), right are 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. The improvement in the forecast quality from 2002/2003 to 2003/2004 can easily be seen there. 42
Figure 37: Plot of forecasts of Nino-3.4 at four start dates September, December 2003 March and July 2004. The red lines represent the 40 ensemble members. The heavy dashed line represents subsequent verification. 43
Figure 38: Time-longitude section of the 20C isotherm depth anomalies at the equator. Longitudes are represented by the X-axis. Time is represented by the Y-axis and it ranges from January 2004 to June 2004. 43

Figure 39: Upper: Summer (JJA) 2003 2m temperature anomalies with respect to ERA-40 (1958-2001) (bold black contour show the anomaly normalised by the climate variability). Lower: Forecast probability of exceeding the upper tercile in the model climate distribution for 2m temperature in the same JJA 2003 period. Seasonal forecast starting date is 1 May 2003 (left) and 1 April 2003 (right). .. 44

Figure 40:Probability of exceeding the upper tercile of 2m temperature, in the model climate distribution, during June-July-August 2003 given by an ensemble of atmospheric simulations forced by observed SST. The starting date is 1 May 2003. 45

Figure 41:Probability of exceeding the upper tercile of 2m temperature, in the model climate distribution, during June-July-August 2003 given by the UKMO seasonal forecasting system. The forecast ensemble starting date is 1 May 2003. 45

Figure 42: Root layer soil moisture index (SMI) averaged over 50-40N, 0-15E. The blue circles represent the ERA-40 monthly means from 1959 to 2001; the orange triangles represent the operational analysis monthly values for 2001 and 2002 and the red triangles the operational analysis for 2003. 46

Figure 43:2m temperature ensemble mean differences. The first ensemble has soil wetness initial conditions prescribed to a uniform value of SMI=25 and the second one to a value of SMI=75. The soil wetness initial conditions are prescribed only over most of the European continent (37-60 N; 10W-30E). Black solid contour indicate areas with a level of significance higher than 90% (thin line), 95% (medium line) and 99% (thick line). 46

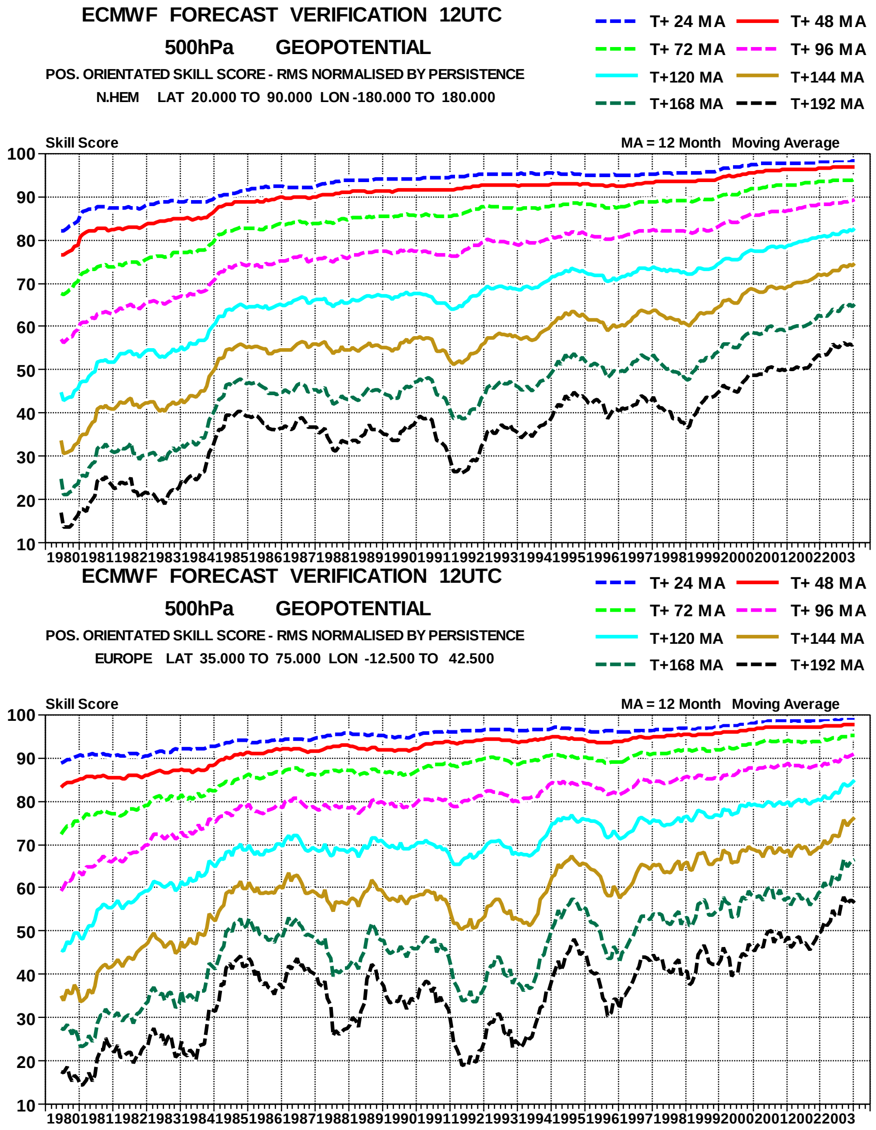


Figure 1: 500hPa height skill score (N. Hemisphere and Europe, 12-month moving averages, forecast ranges from 24 to 192 hours)

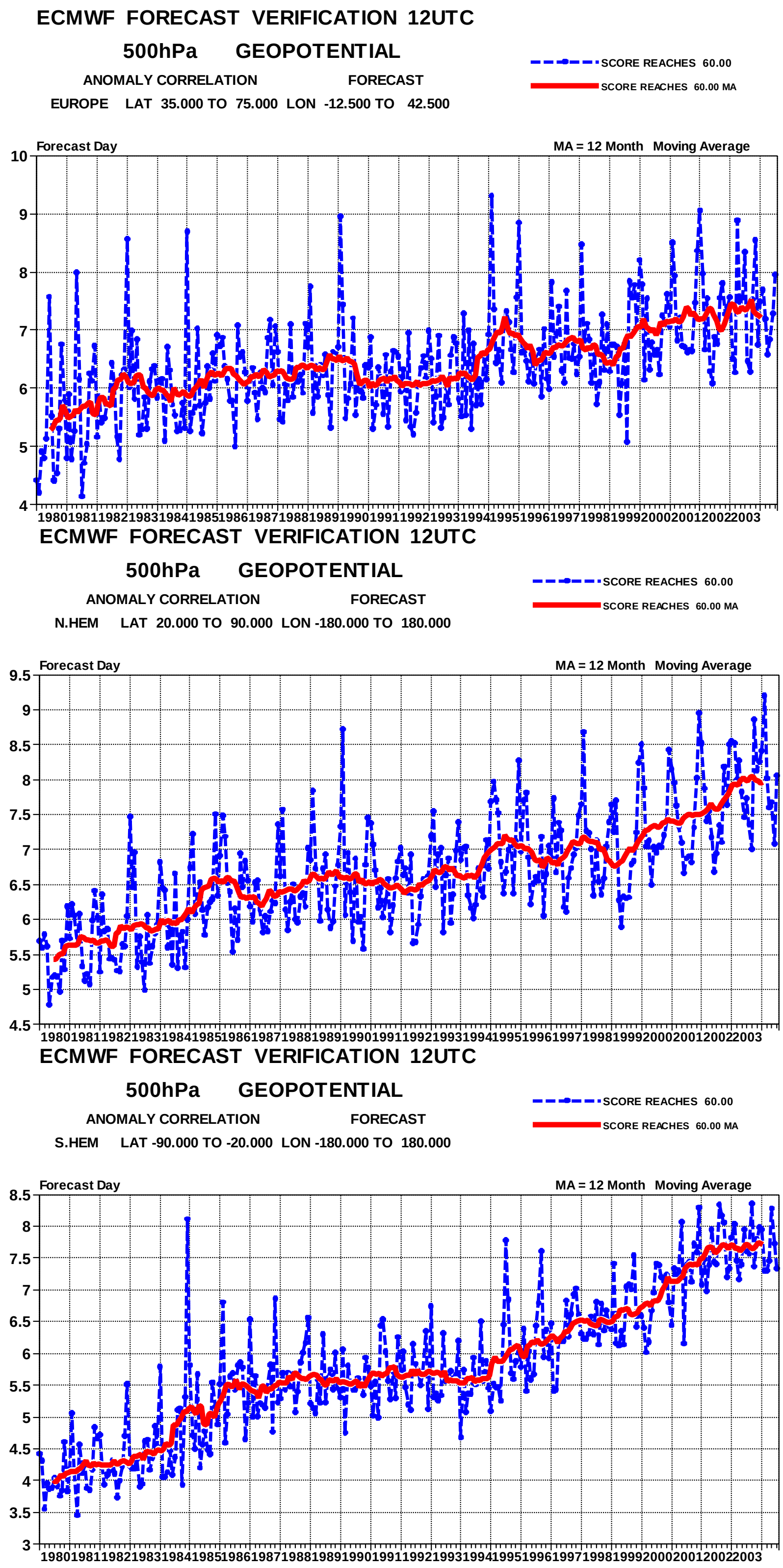


Figure 2: Evolution with time of the 500hPa height forecast performance – each point on the curves is the forecast range when the monthly average of the daily forecast anomaly correlation with observation (analysis) is falling below 60% for Europe, Northern and Southern Extratropics (the red curve is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

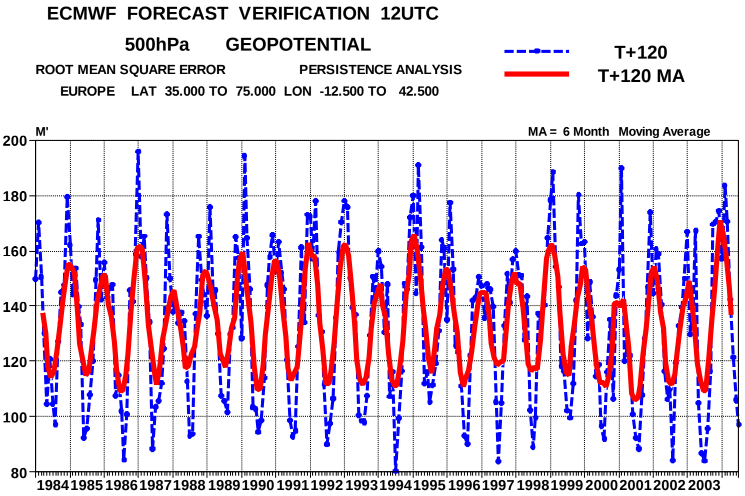


Figure 3: Root Mean Square Error made by persisting the analysis over 120h and verifying it as a forecast. 6-monthly averages (red curve) confirm that the last cold season (October to March) has been unusually active over Europe

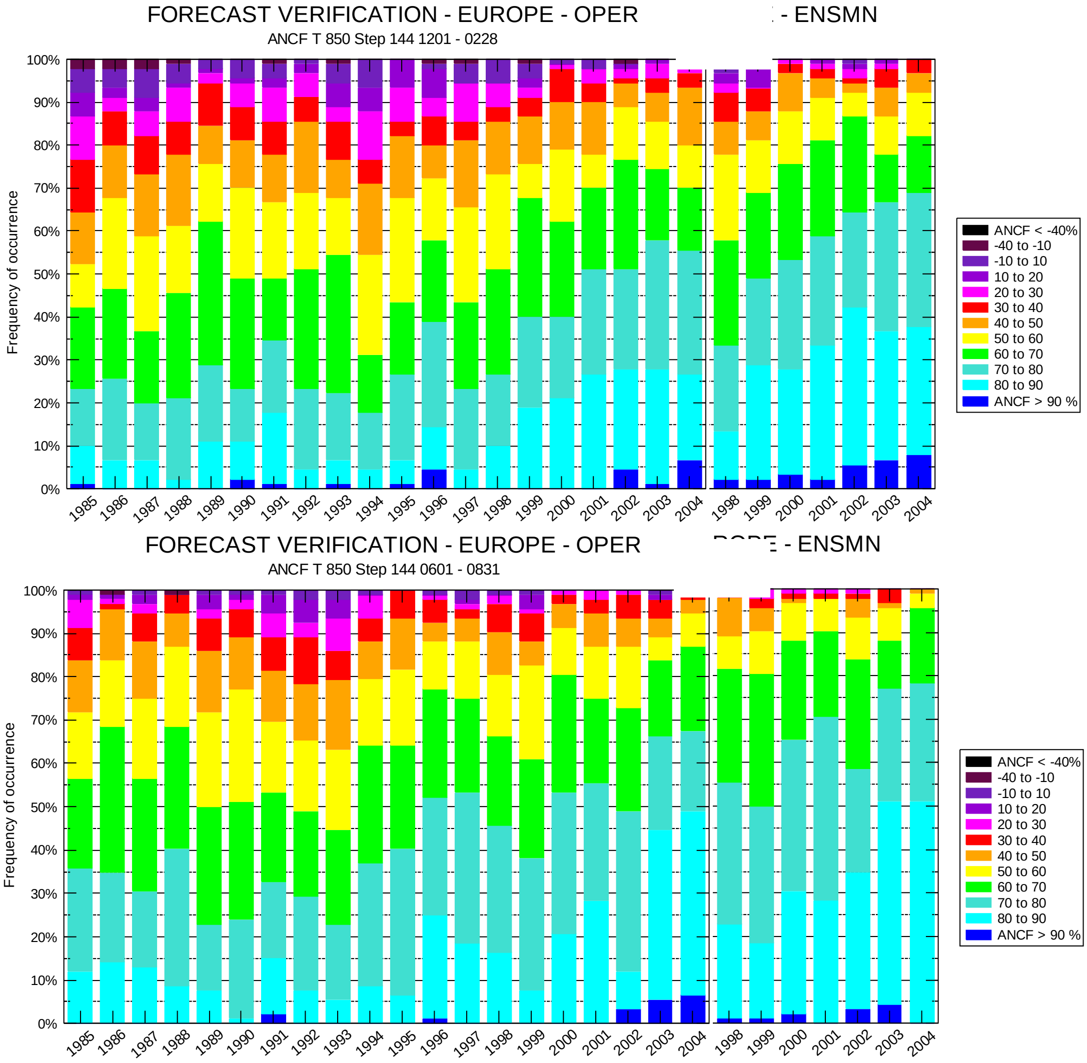


Figure 4: Cumulative distribution of Anomaly Correlation of the Day 6, 850hPa temperature forecasts with verifying analyses over Europe in winter (DJF, top) and summer (JJA., bottom) since 1984-85 for the deterministic, high resolution forecasts (left panels) and since 1997-98 for the EPS Ensemble mean (right panels).

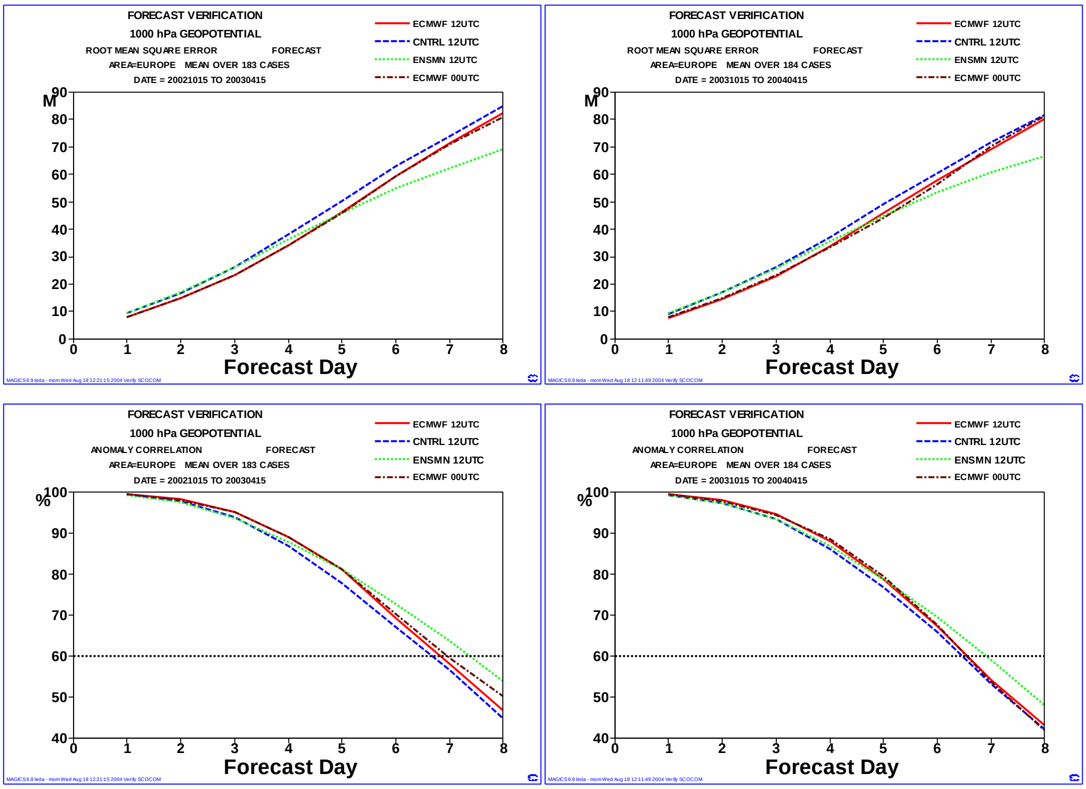


Figure 5: Comparison between near surface pressure height (1000-hPa) scores from T511 (ECMWF, 12UTC in red, 00UTC in brown), T255 (CNTRL, blue) and EPS ensemble mean (green) during the cold season (15 Oct.-15 Apr.) over Europe. Left: 2002-2003; Right: 2003-2004.

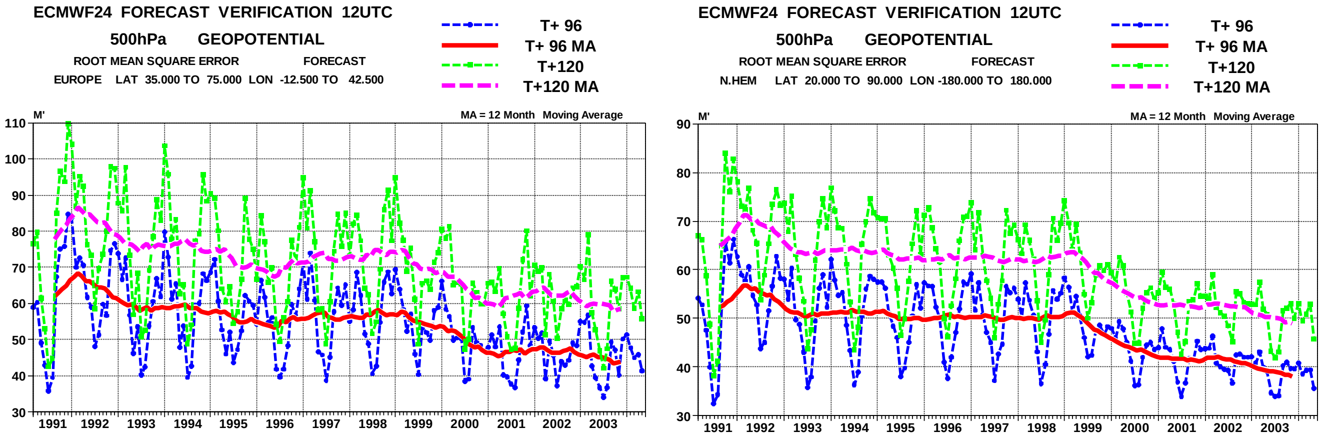


Figure 6: RMS of the difference between 24h-consecutive 500hPa height forecasts verifying the same day over Europe (left panel) and Northern Extratropics (right panel).

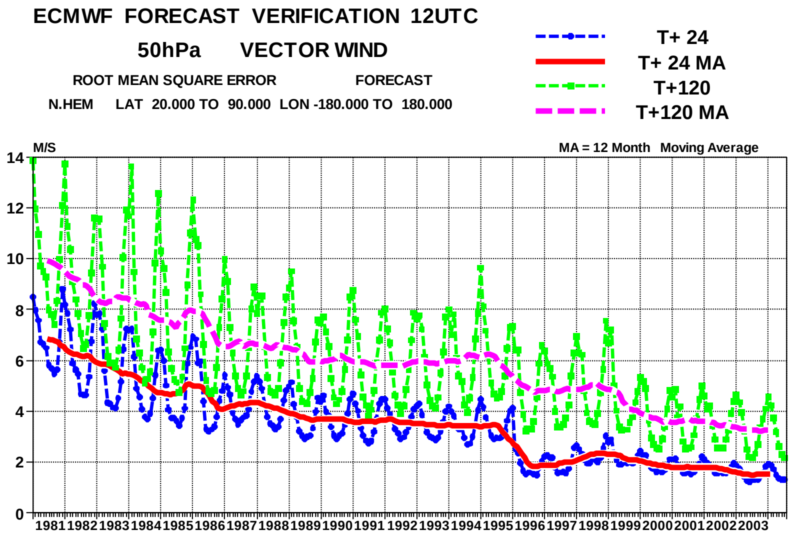


Figure 7: Model scores in the extratropical Northern Hemisphere stratosphere (50hPa height Day 1 and Day 5 forecasts RMSE)

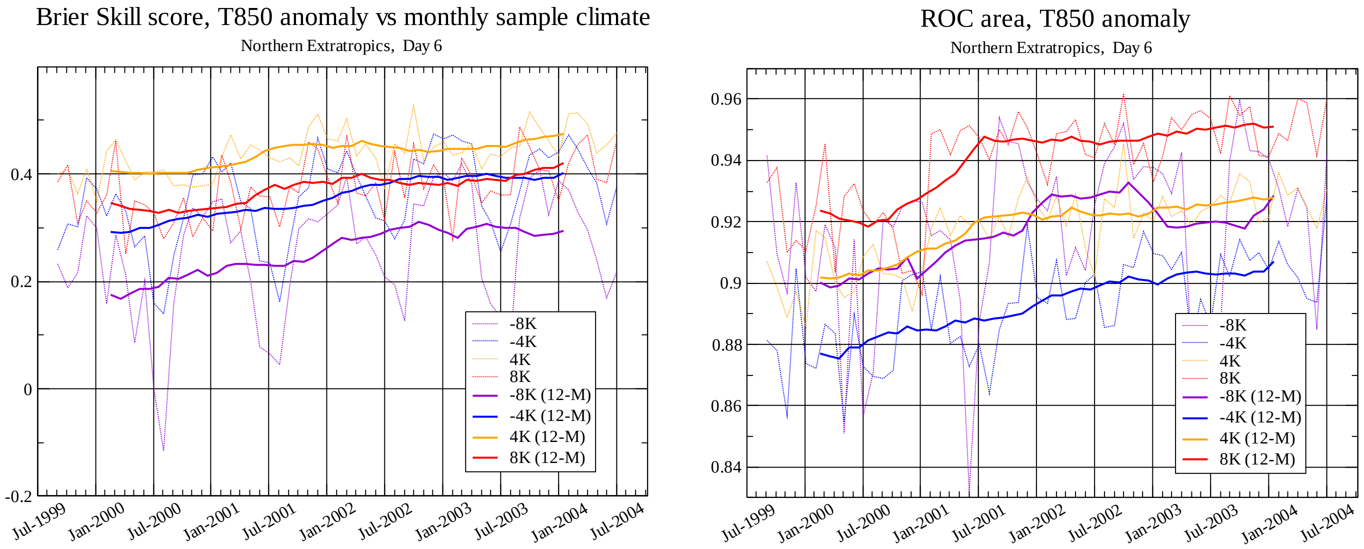


Figure 8: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics Area (ROCA, lower panel), the latter showing the skill shown by the EPS at detecting a signal verified by the analysis out of the Day 6 probability forecast of 850hPa anomaly temperature (0.5 is no skill, 1 is a perfect detection). Both panels show monthly scores (thin dotted lines) and 12-months moving averages (full heavy lines). Verifying area is the Northern Extratropics (20^{o}N and beyond)

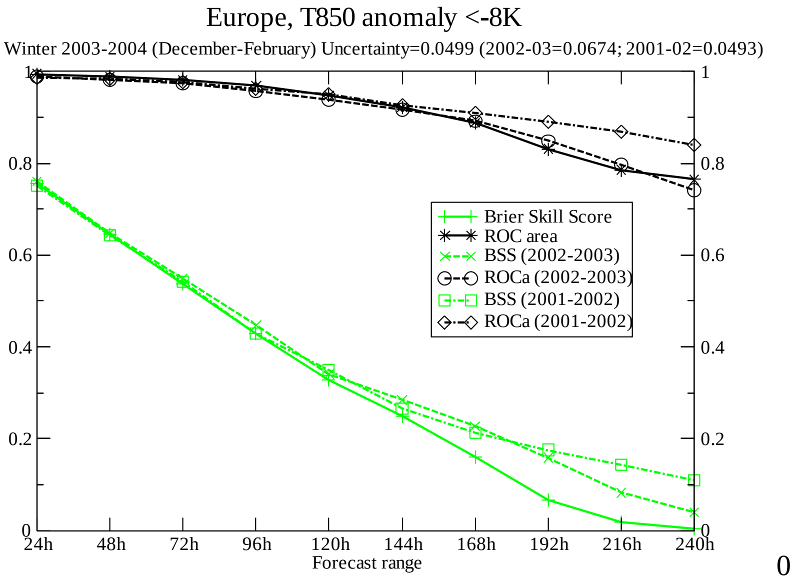


Figure 9: EPS skill at forecasting strong temperature anomalies (8K or more) over Europe. Upper panel: cold winter anomalies. Lower panel: warm summer anomalies.

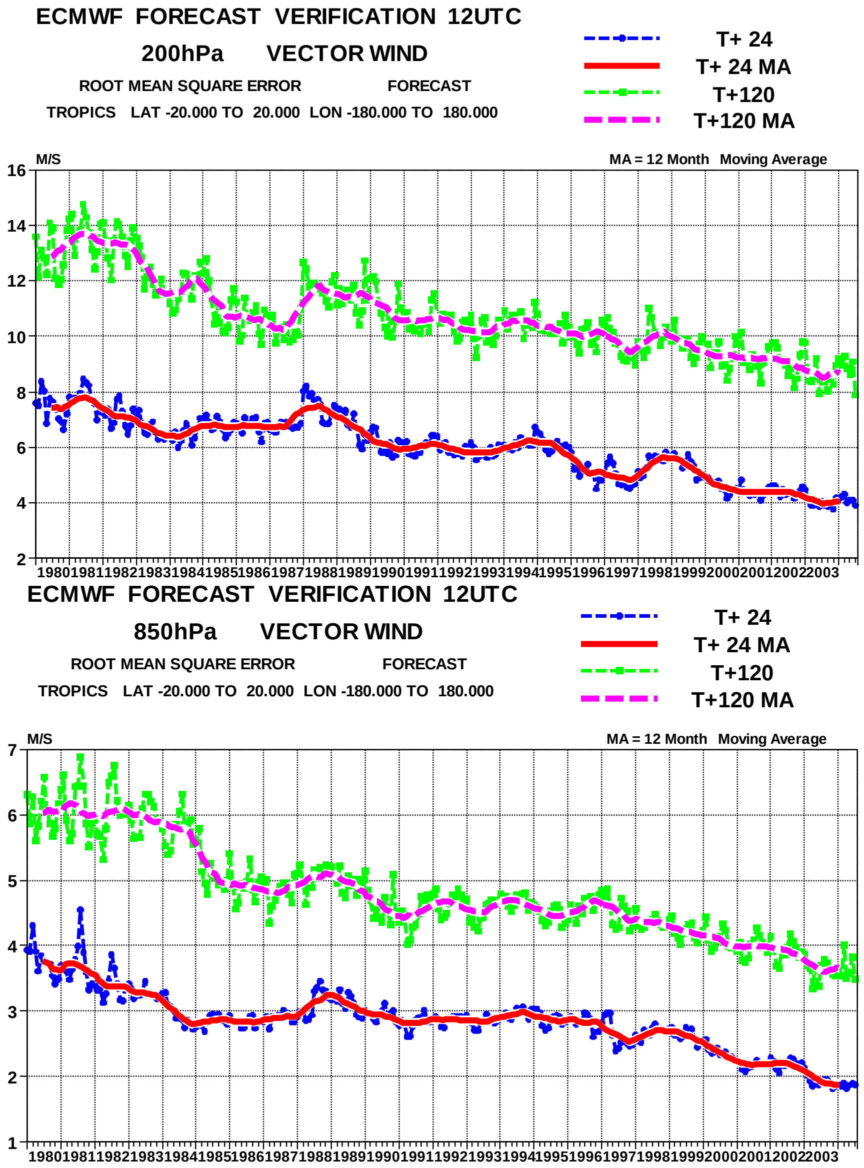
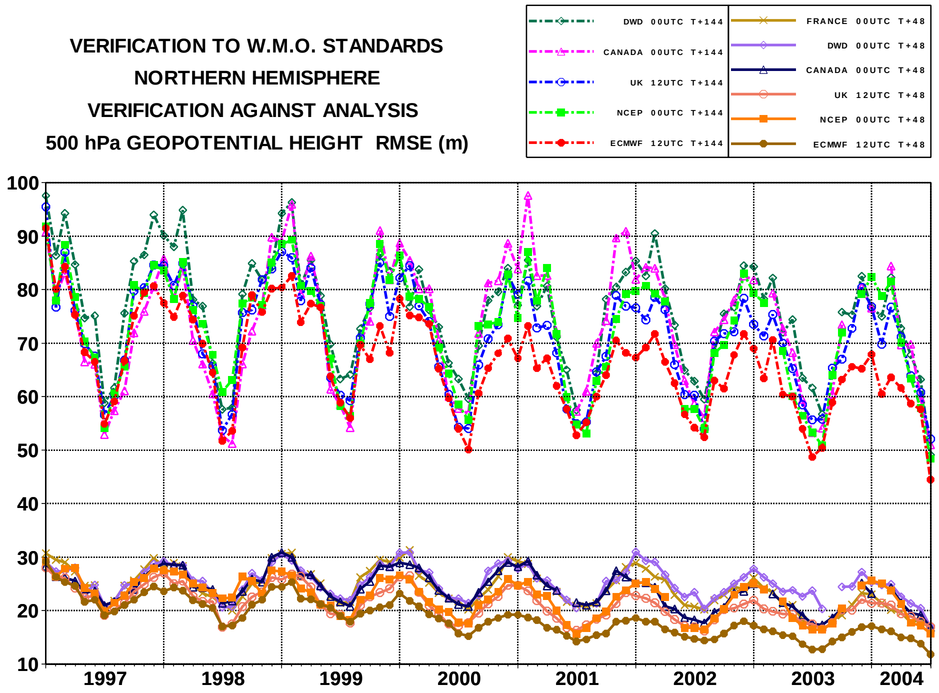


Figure 10: Model scores in the Tropics (root mean square errors for 200hPa and 850hPa wind)



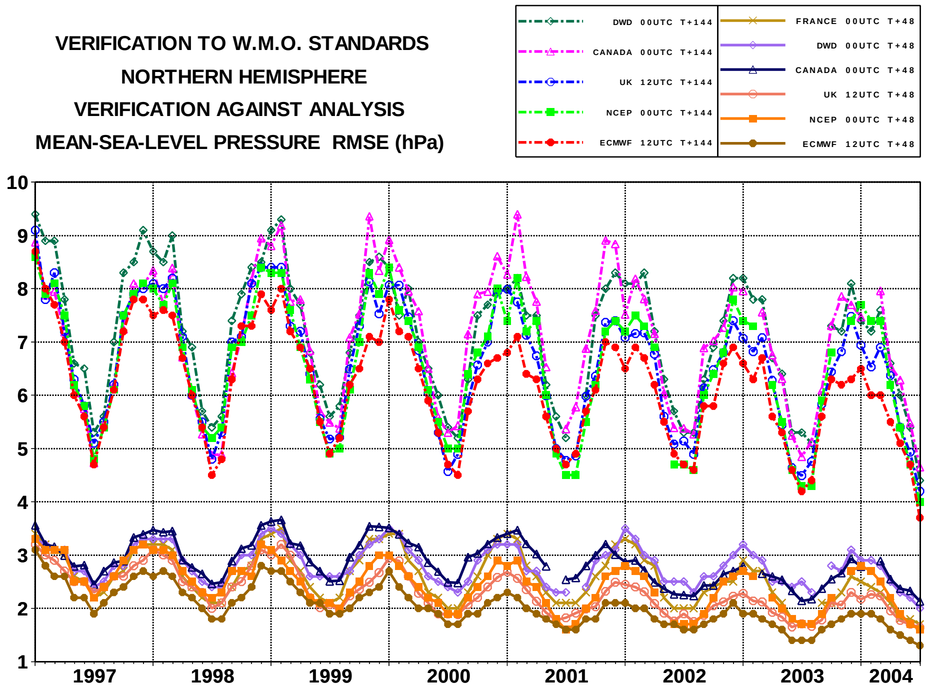


Figure 11: WMO/CBS exchanged scores (RMS error over Northern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)

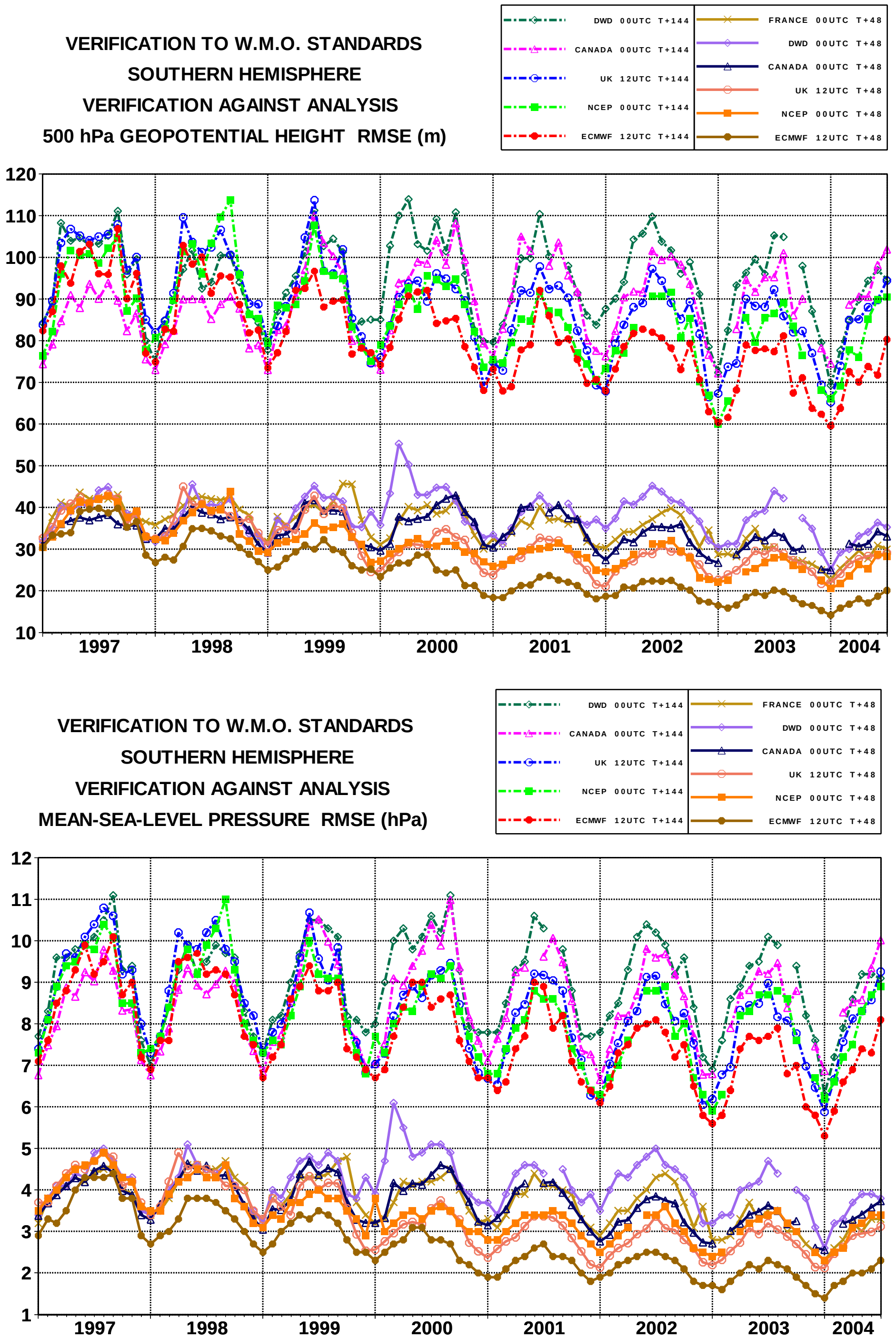


Figure 12: WMO/CBS exchanged scores (RMS error over Southern Extratropics, 500hPa and MSLP for D+2, D+4 and D+6)

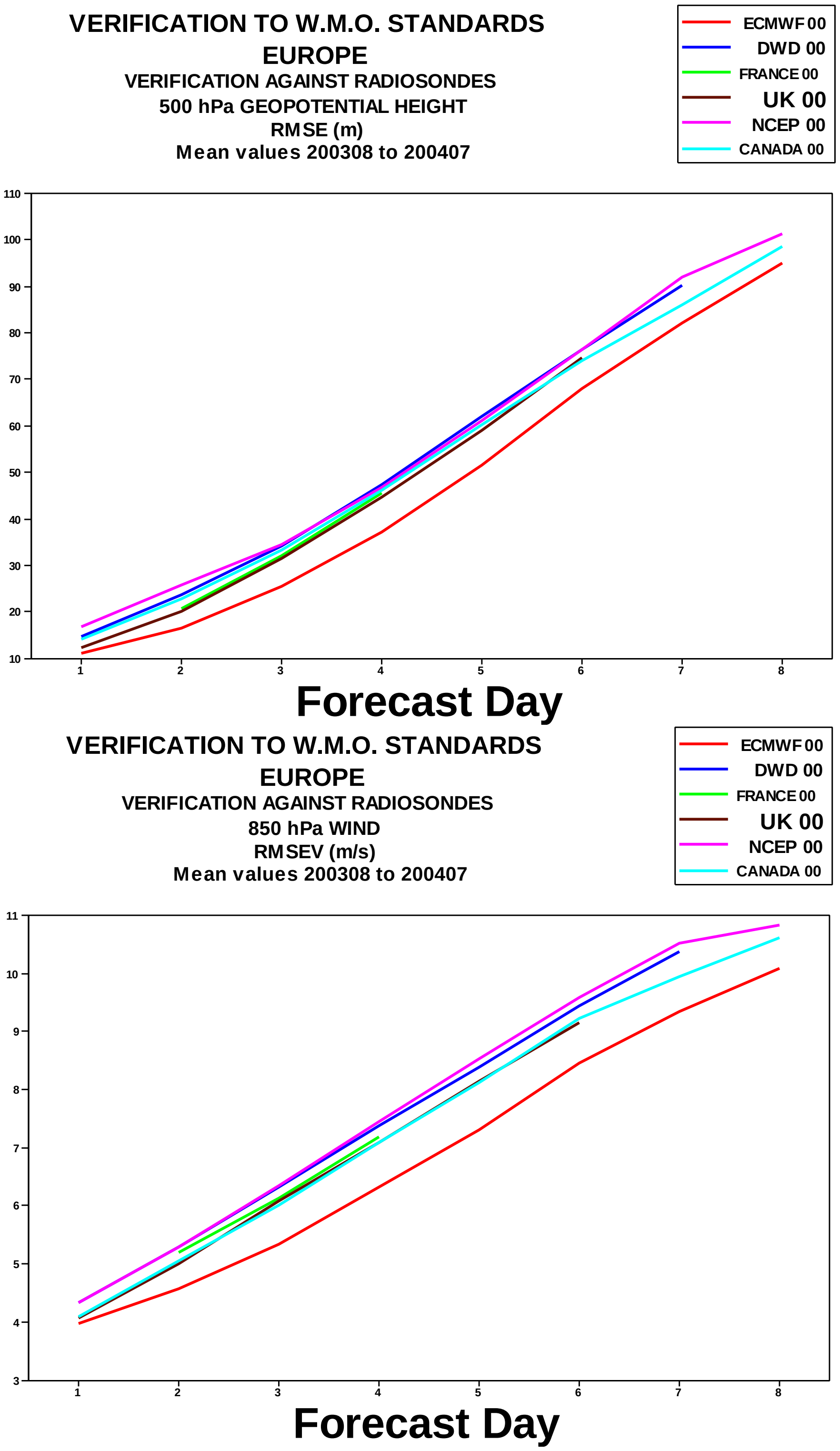


Figure 13: WMO/CBS exchanged scores using radiosondes: 500hPa height and 850hPa wind RMS error over Europe (annual mean)

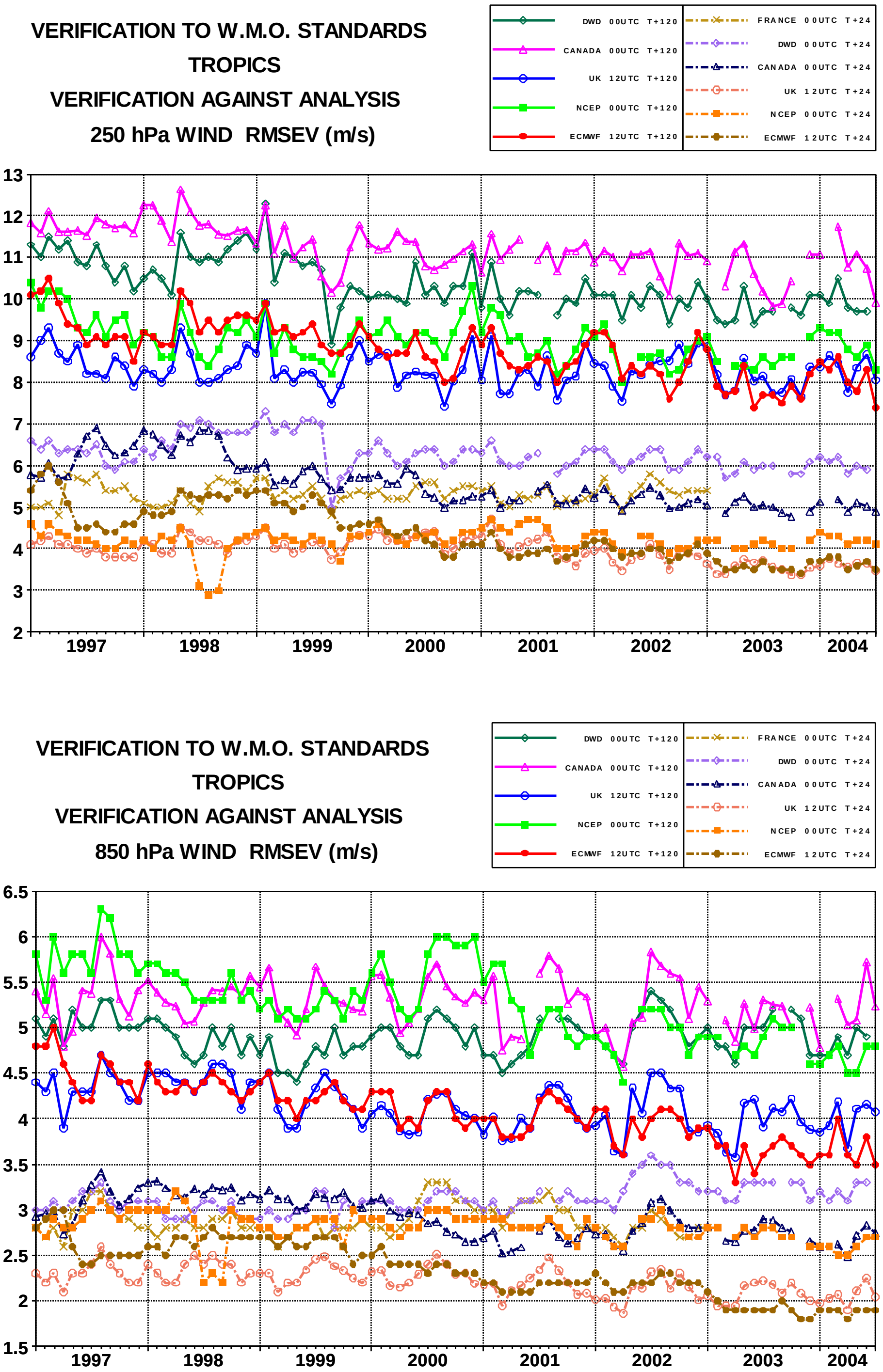


Figure 14: WMO/CBS exchanged scores (RMS vector error over the Tropics, 250hPa and 850hPa wind forecast for D+1 and D+5); reference for verification is each centre's own analysis

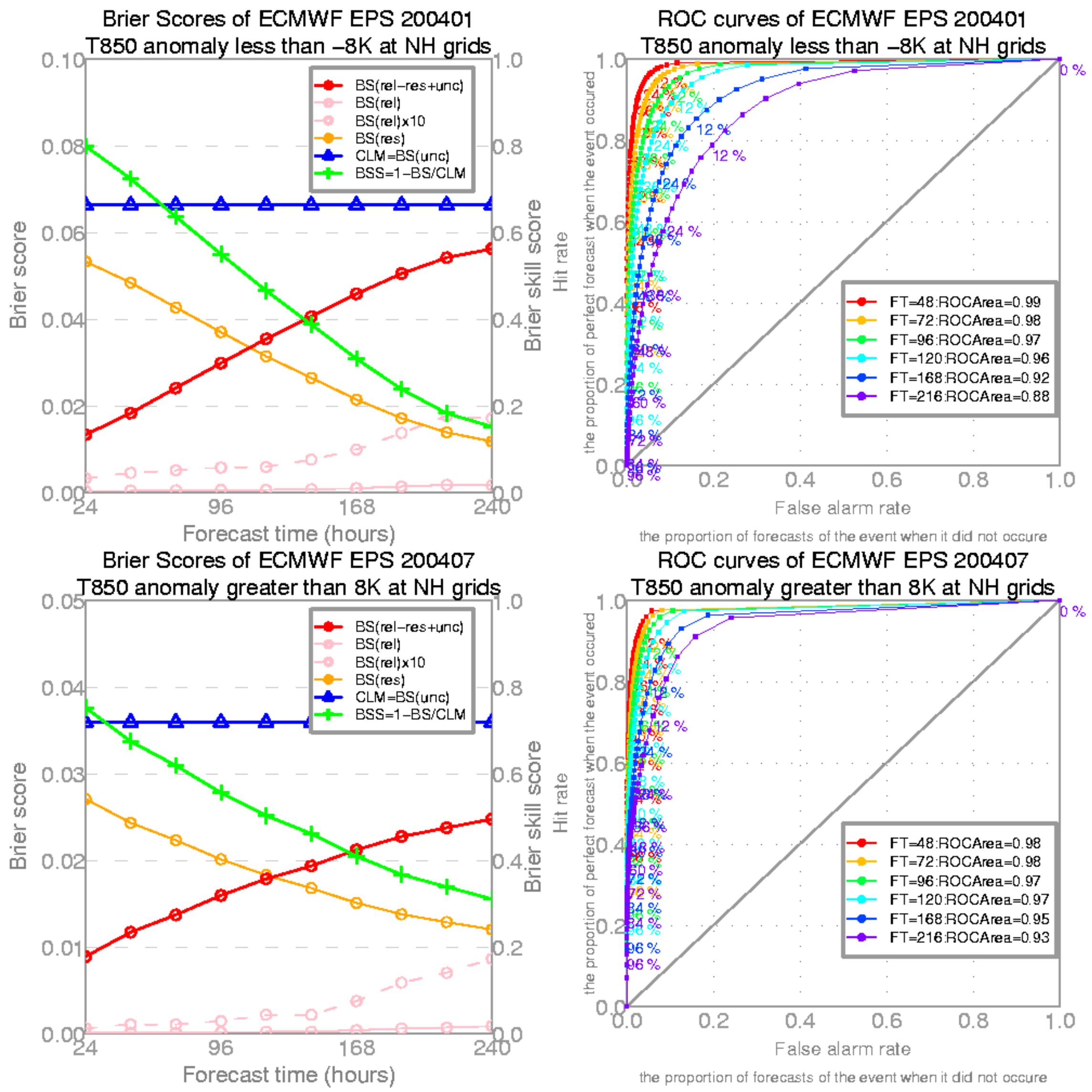
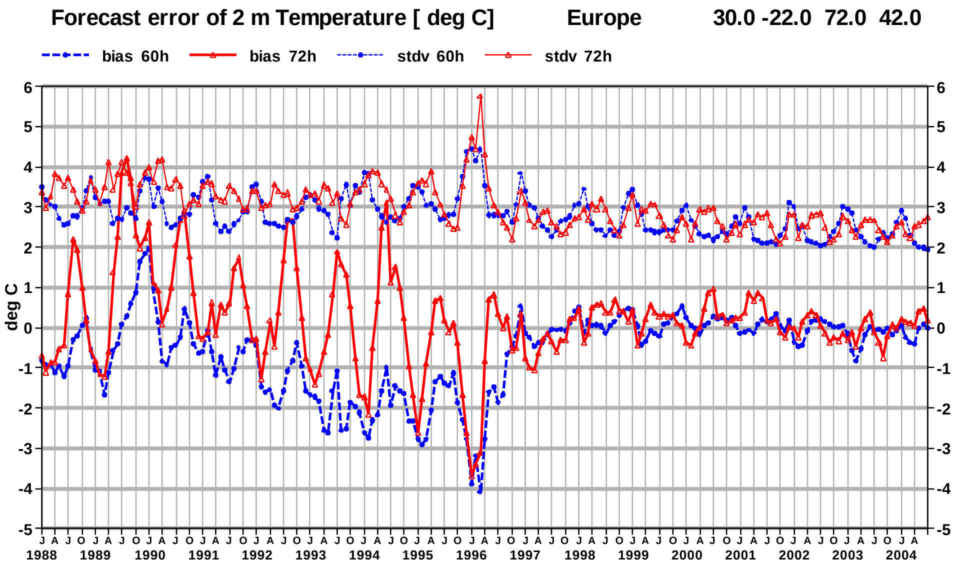


Figure 15: Example of ECMWF EPS verification results as shown on JMA EPS intercomparison website. Upper panels are for forecasts of very cold (by more than 8K with respect to the climate) anomalies in January 2004. Lower panels are for very warm (by more than 8K with respect to the climate) anomalies in July 2004.



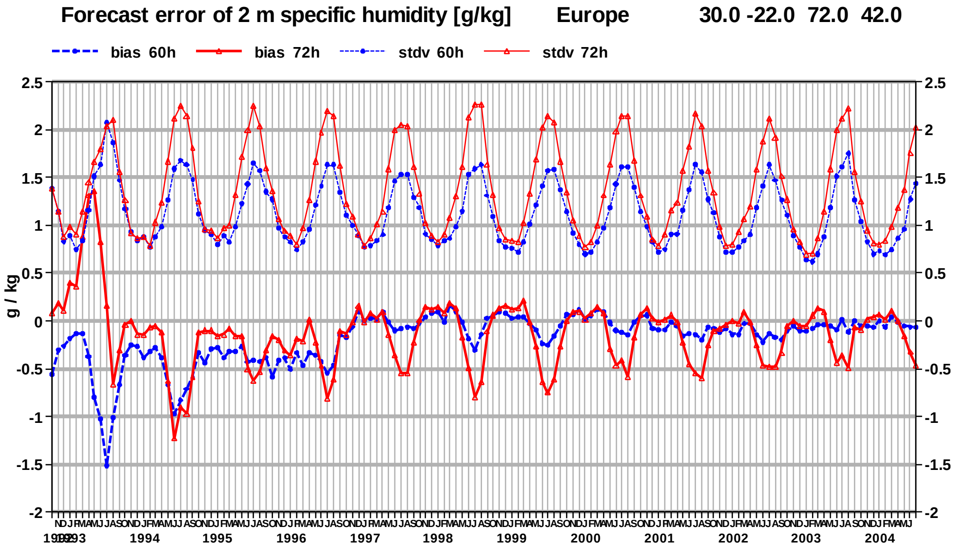
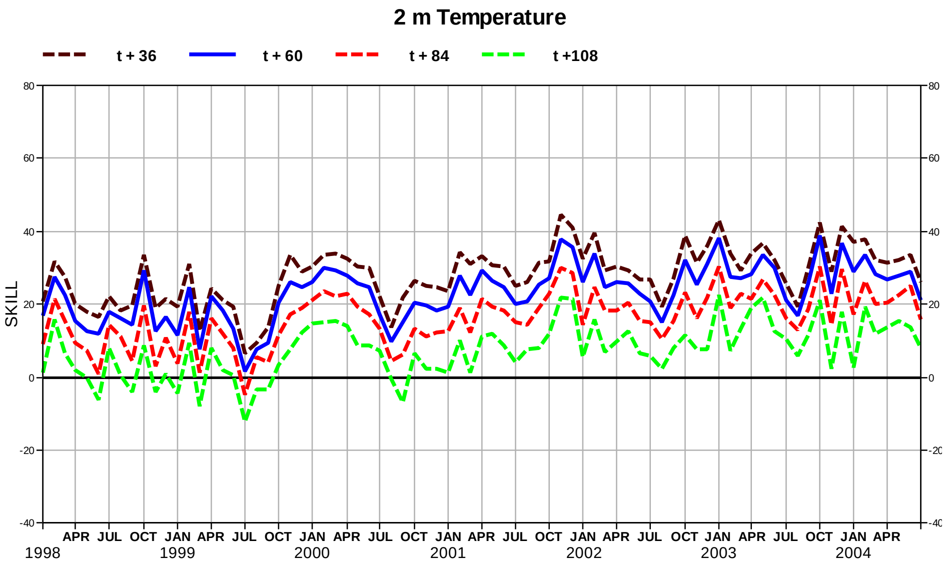


Figure 16: Verification against European SYNOP observations of 2m Temperature and specific humidity (bias and standard deviation, T+60h -00UTC- and +72h -12UTC)



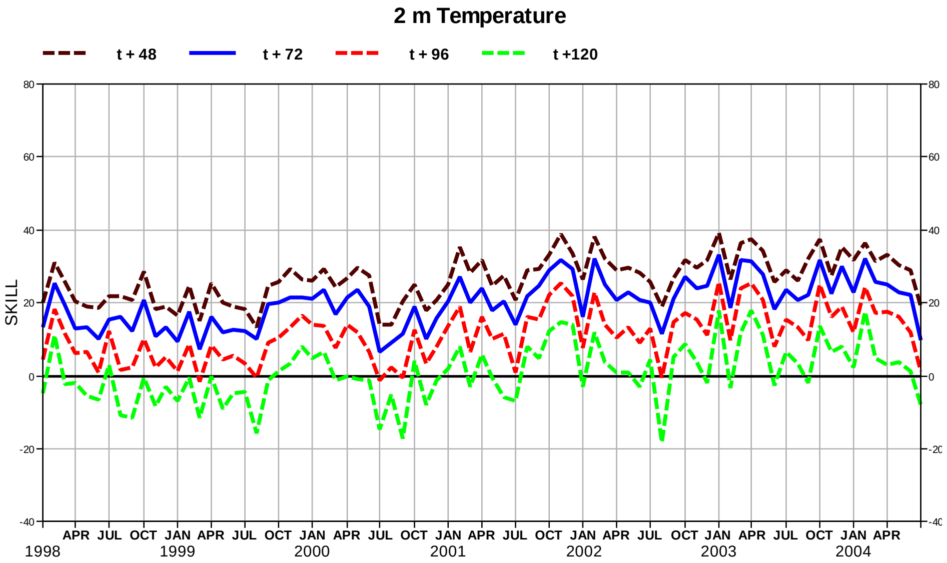


Figure 17: RMSE skill with respect to the persistence forecast for 2m temperature forecasts over Europe during night time (00UTC, top) and daytime (12UTC, bottom) for different forecast ranges.

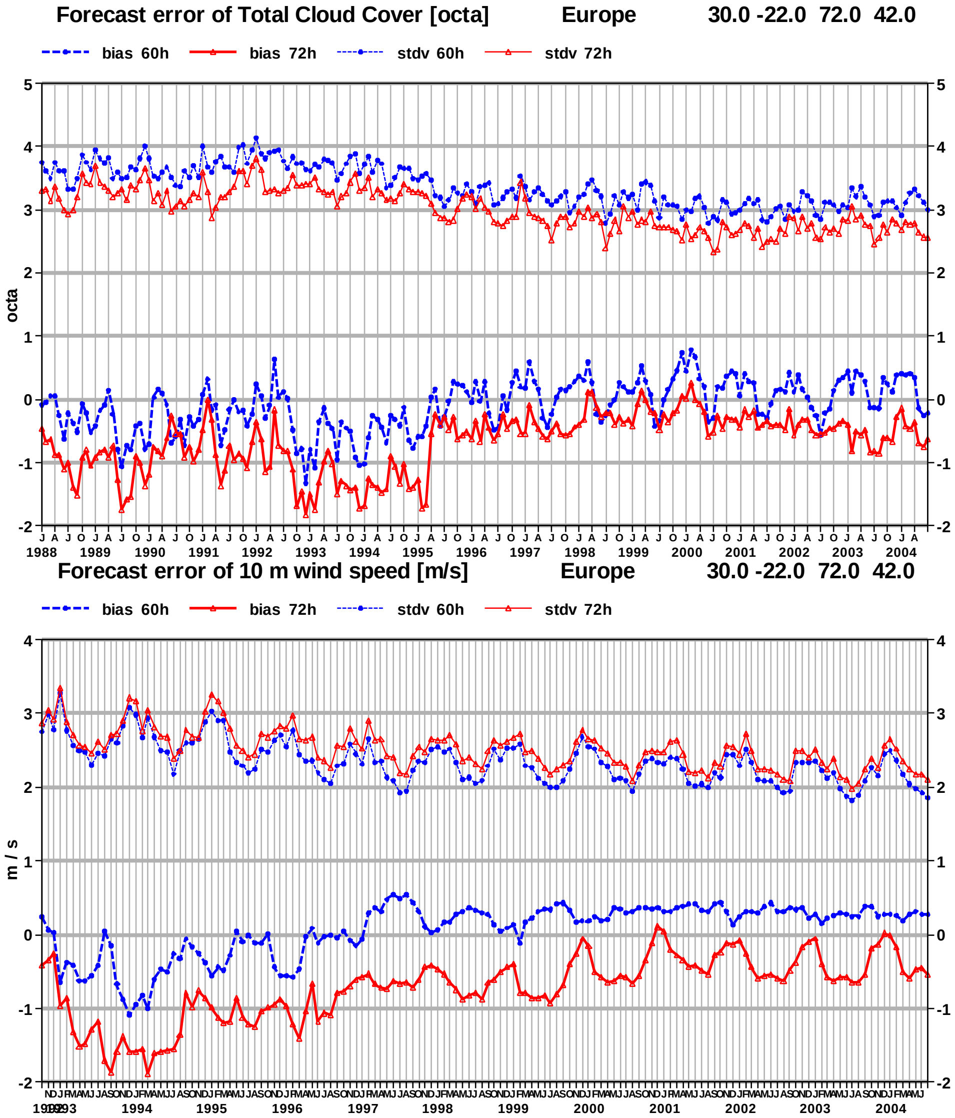


Figure 18: Scores against European SYNOPs of total cloud cover and 10m wind speed forecasts (bias and standard deviation, T+60h -00UTC- and +72h -12UTC).

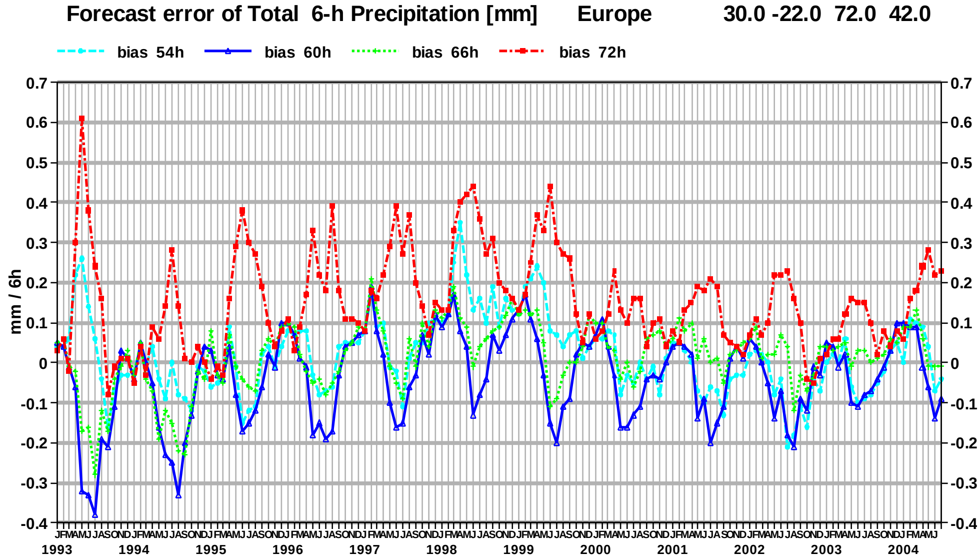


Figure 19: 6h-accumulated precipitation forecasts biases (T+54/60/66/72h) with respect to SYNOP

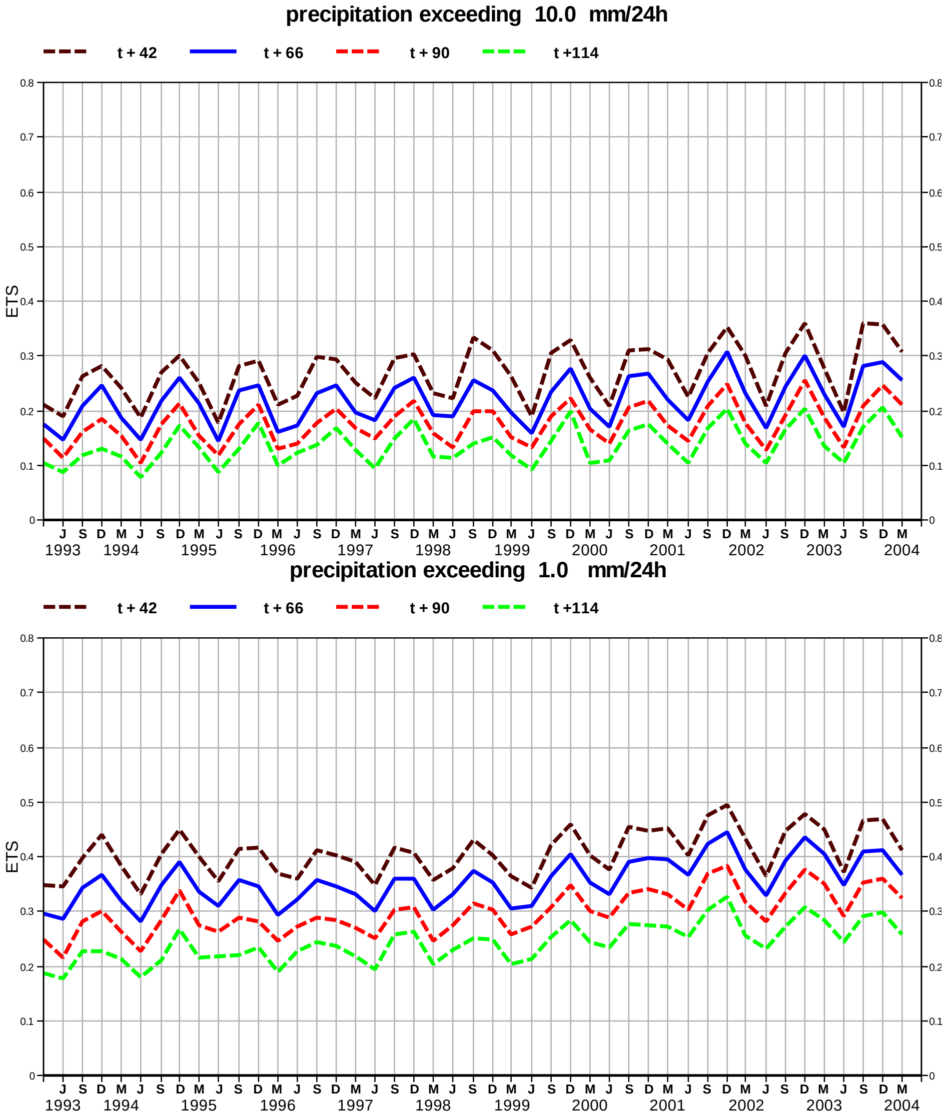


Figure 20: Time series of Equitable Threat Scores for the forecast of daily precipitation verified using SYNOP reports over Europe; Top: threshold 10mm, bottom: 1mm.

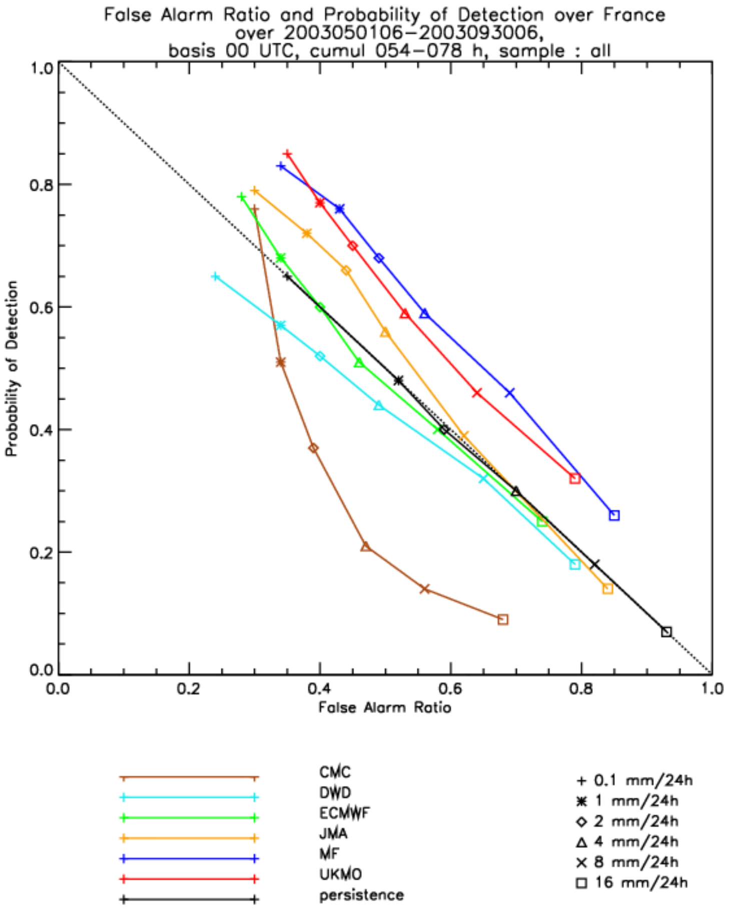


Figure 21: Verification of 54-78h precipitation 00UTC-based forecasts against upscaled observations from the French high-resolution network (2003 warm season); ECMWF is in green. The diagonal is a line of no frequency bias (above is over- , below under- forecasting). The better the forecast, the closest to the upper left corner(all events are detected, all positive forecasts are correct). For each forecast centre, the higher the detection threshold, the poorer the forecast (courtesy from WGNE/Meteo-France)

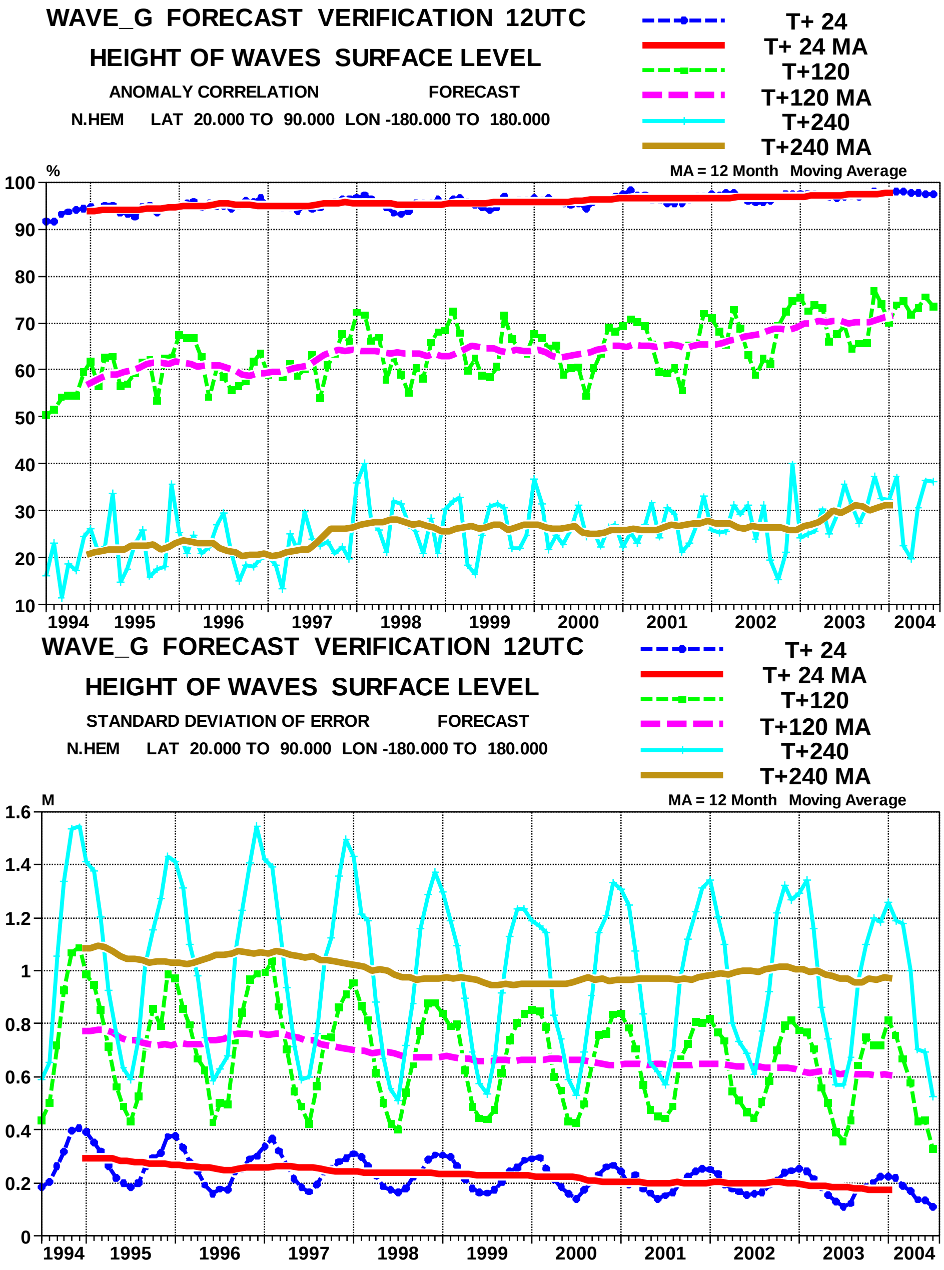


Figure 22: Scores (anomaly correlation and standard deviation) of oceanic wave heights verified against the analysis (Northern Extratropics)

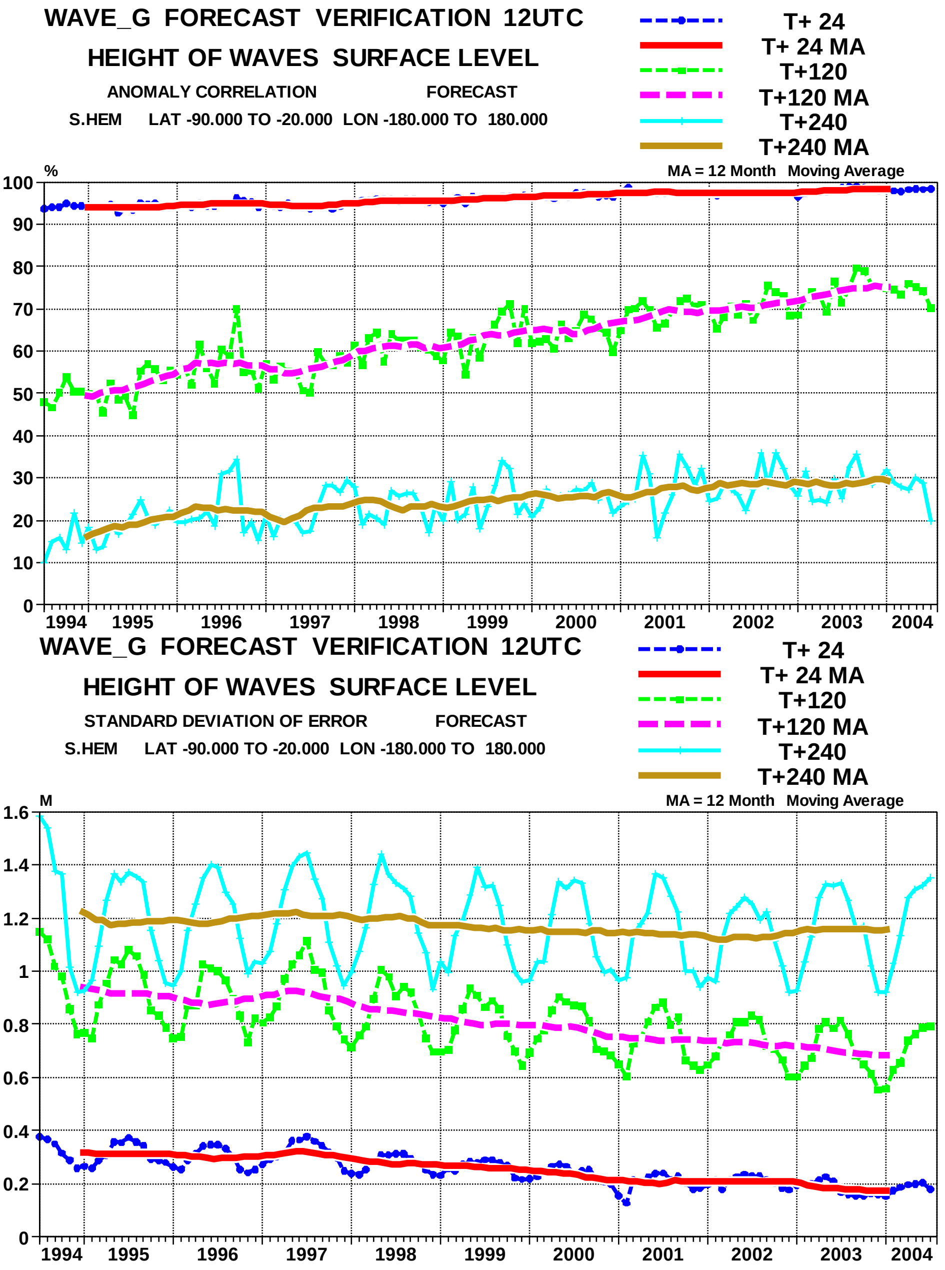


Figure 23: Scores (anomaly correlation and standard deviation) of oceanic wave heights verified against the analysis (Southern Extratropics)

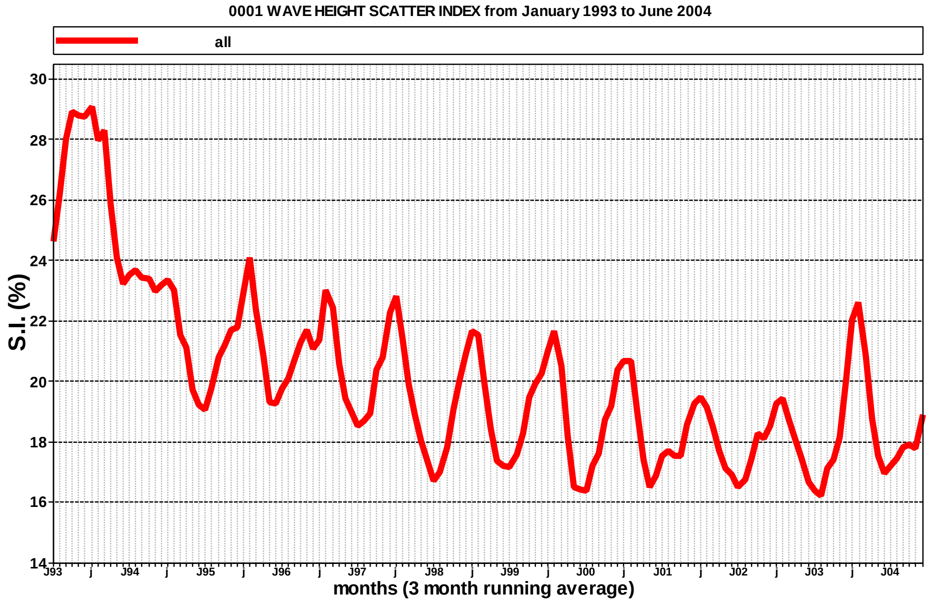


Figure 24: Comparison of ECMWF wave model analysis with bouy observations (wave height).The Scatter Index (S.I.) is the error standard deviation normalised by the mean value of the observations

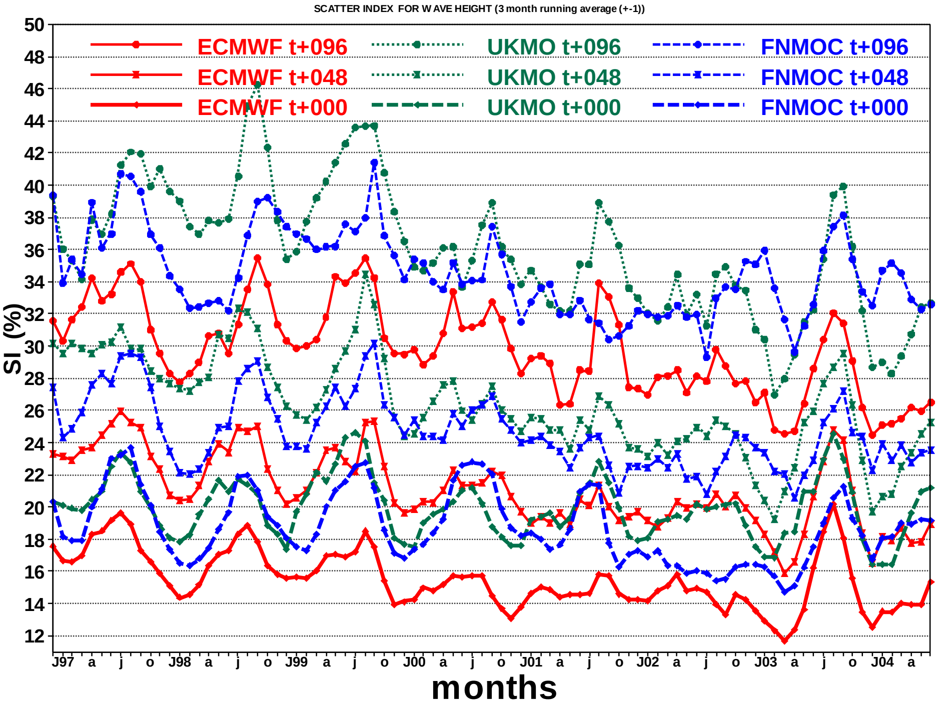
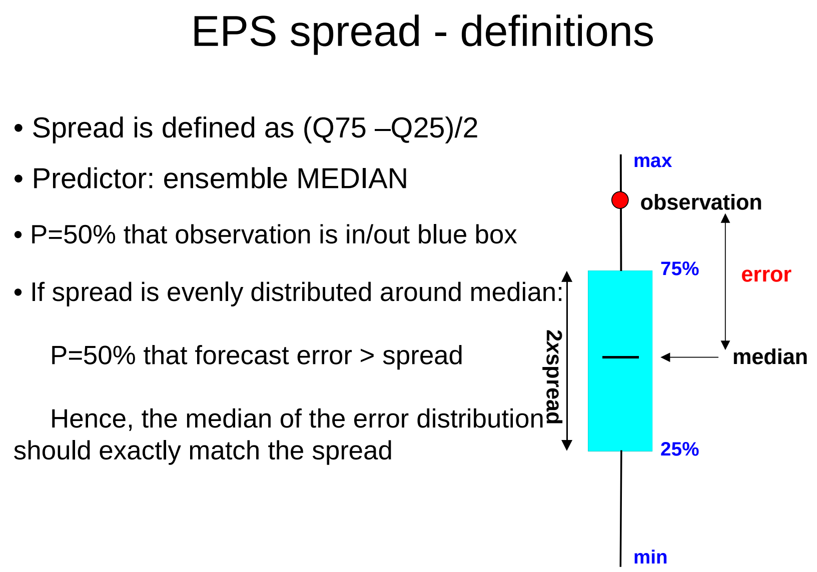


Figure 25: Verification of different model wave height forecasts using a consistent set of observations



EPS spread - definitions

- Spread is defined as (Q75 –Q25)/2
- Predictor: ensemble MEDIAN
- P=50% that observation is in/out blue box
- If spread is evenly distributed around median:

P=50% that forecast error > spread

Hence, the median of the error distribution should exactly match the spread

Figure 26: Schematic description of the spread skill relation that should be found in a perfect probabilistic forecast.

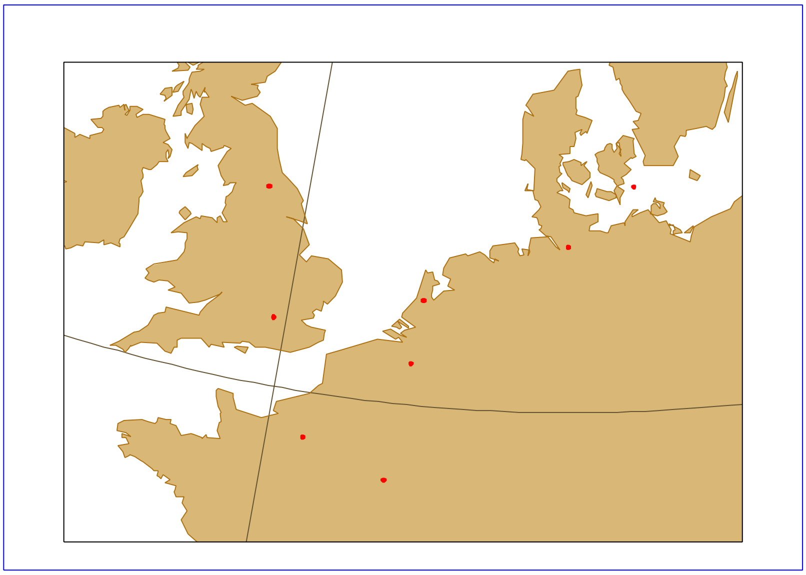


Figure 27: Stations used for the verification of the EPS spread (Figure 28)

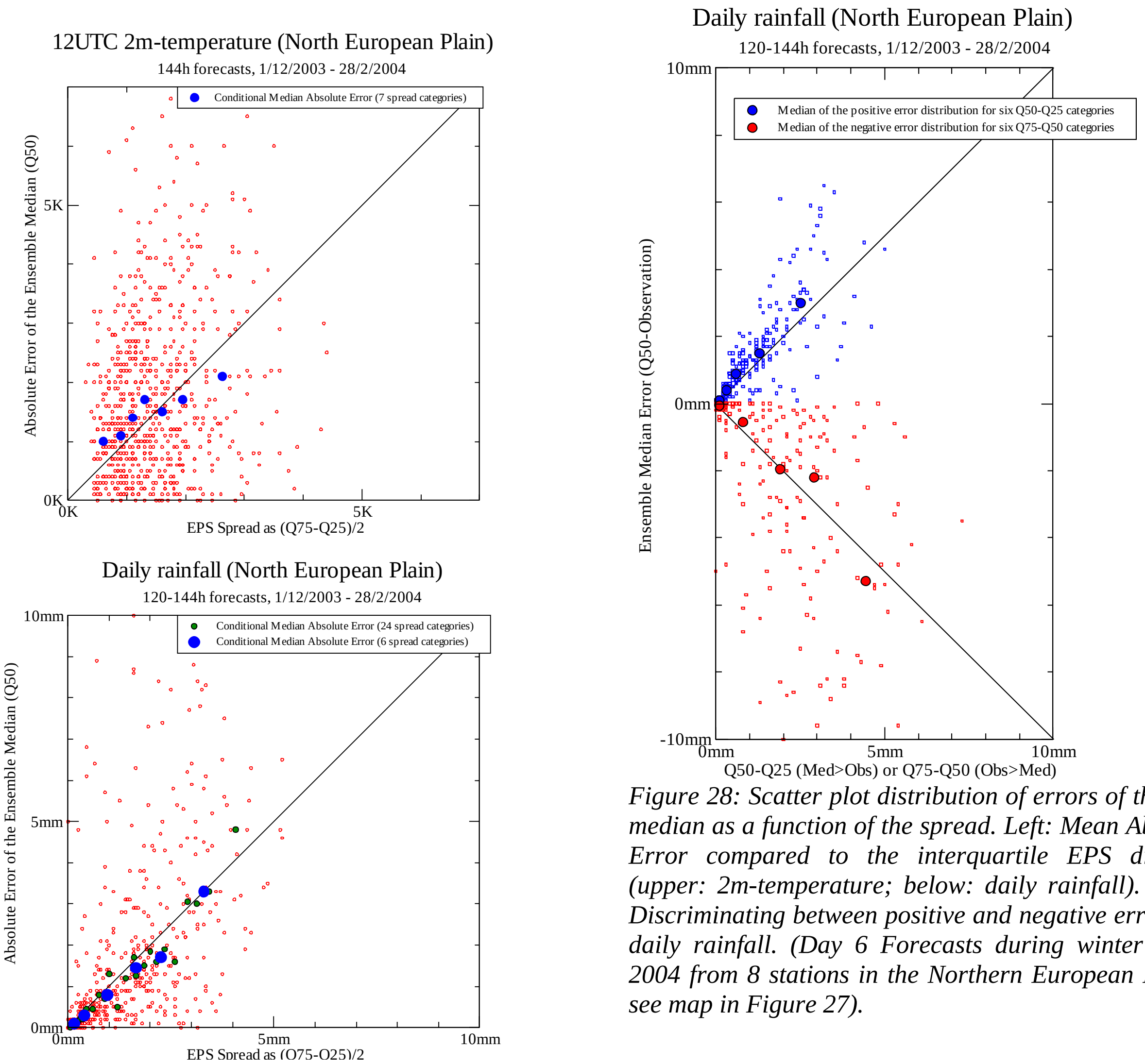


Figure 28: Scatter plot distribution of errors of the EPS median as a function of the spread. Left: Mean Absolute Error compared to the interquartile EPS distance (upper: 2m-temperature; below: daily rainfall). Right: Discriminating between positive and negative errors for daily rainfall. (Day 6 Forecasts during winter 2003-2004 from 8 stations in the Northern European Plain - see map in Figure 27).

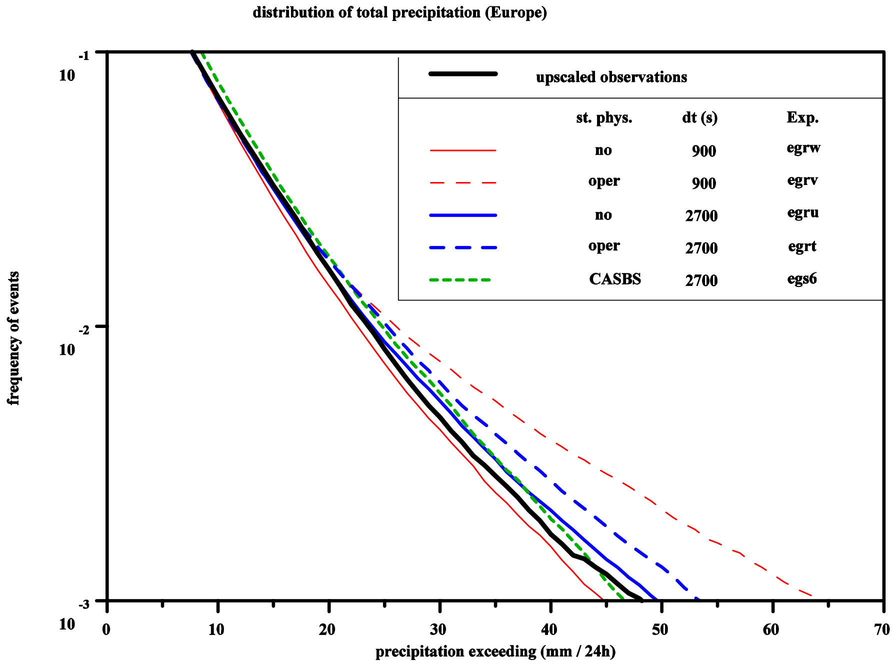


Figure 29: Distribution of daily rainfall events over Europe as observed (bold black line, upscaled from high resolution data) and forecasted in the operational EPS (dashed blue), EPS without stochastic physics (full blue) and with a revised stochastic physics scheme (dashed green). The effect of reducing the time step (from 2700s to 900s) is also shown in red.

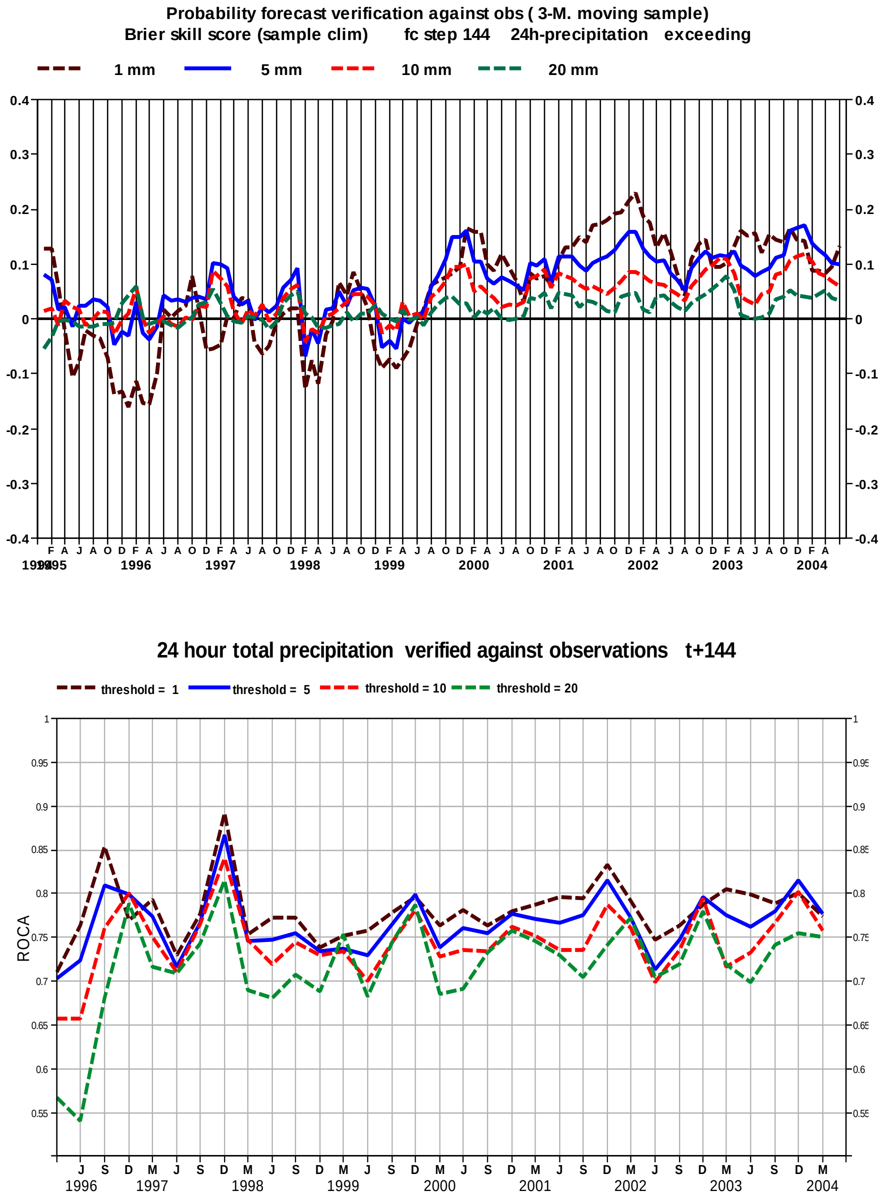


Figure 30: Time series of the Brier Skill Score (upper panel) and Relative Operating Characteristics curve Area (ROCA, lower panel), the later showing the skill shown by the EPS at detecting a signal out of the 120-144h probability forecast for rain (0.5 is no skill, 1 is a perfect detection).

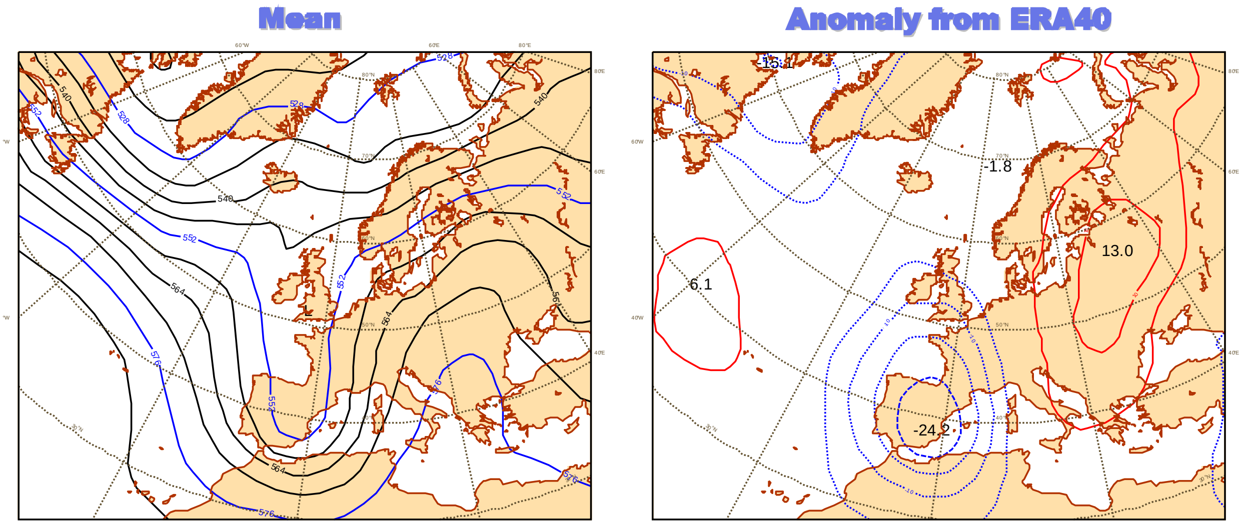


Figure 31: Composite500hPa flow built from 6 severe flood events over the southern Alpine region from 1993 to 2003.

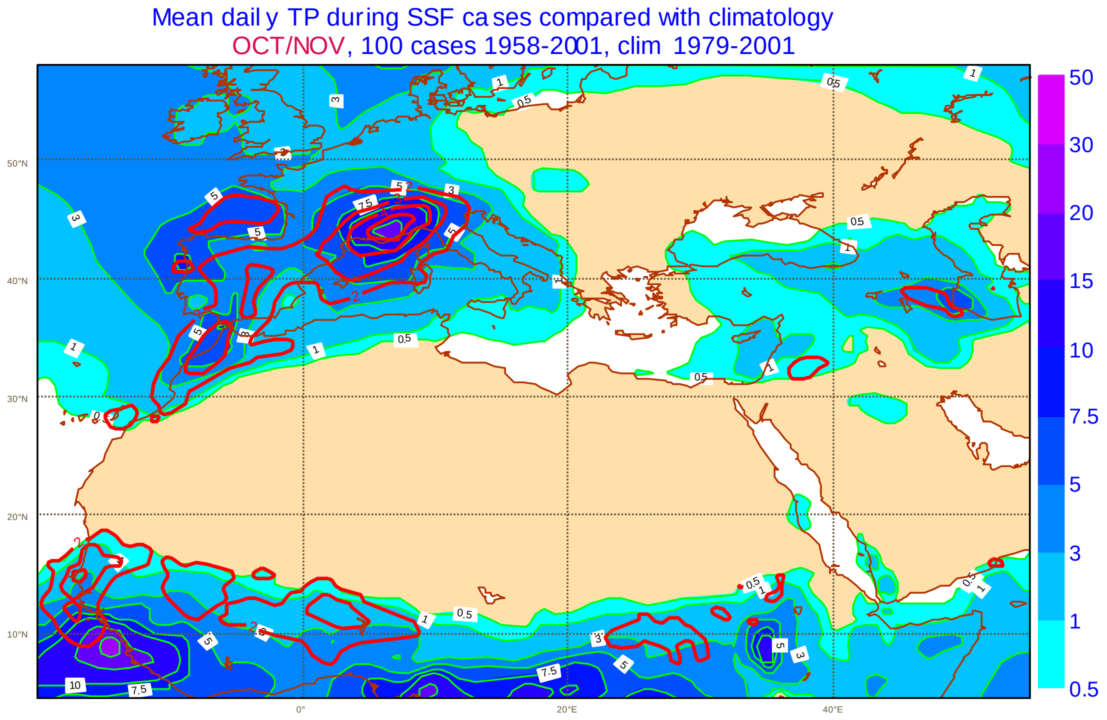

Mean dail y TP dur ing SSF ca ses compared with climatology
OCT/NOV, 100 cases 1958-2001, clim 1979-2001
50°N
40°N
30°N
20°N
10°N
0°
20°E
40°E
50
30
20
15
10
7.5
5
3
1
0.5

Figure 32: Mean precipitation anomaly for the 100 cases correlating best with the composite pattern from Figure 31 in October/November during the ERA40 period. Red contours are scaled with respect to the standard deviation of the climatology

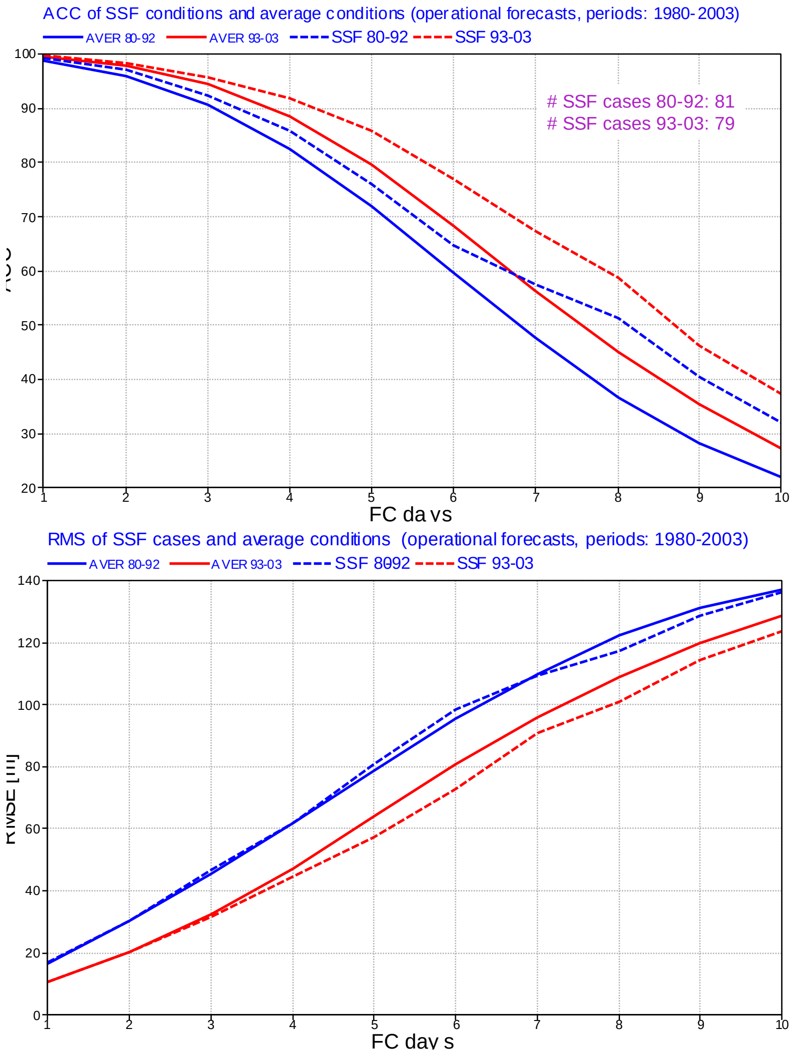


Figure 33 : Score comparison between 1980-1992 (blue)and 1993-2003 (red); full lines are regardless of the flow pattern, while dashed lines are for the cases correlating best with thecomposite shown in Figure 31; Upper: Anomaly correlation; Lower: RMSE

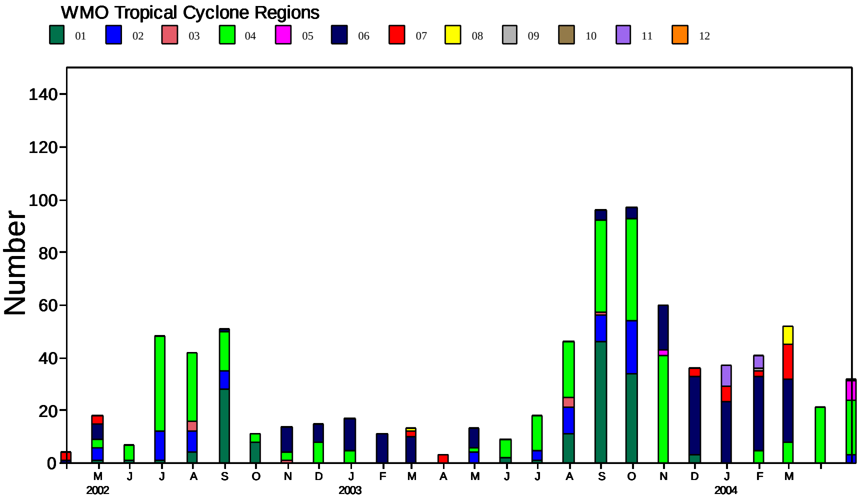


Figure 34: Number of Tropical Cyclone tracked by the determinist,T511 day 2 forecast from April 2002 to May 2004. For each month, the number is split per WMO Tropical Cyclone region (1=NW Atlantic; 2=.NE Pacific; 3=N Pacific; 4=NW Pacific; 5=N. Indian; 6= SW Indian; 7=SE Indian; 8/9/10=SW Pacific; 11/12=S. Pacific). Both 00 and 12UTC forecasts are tracked.

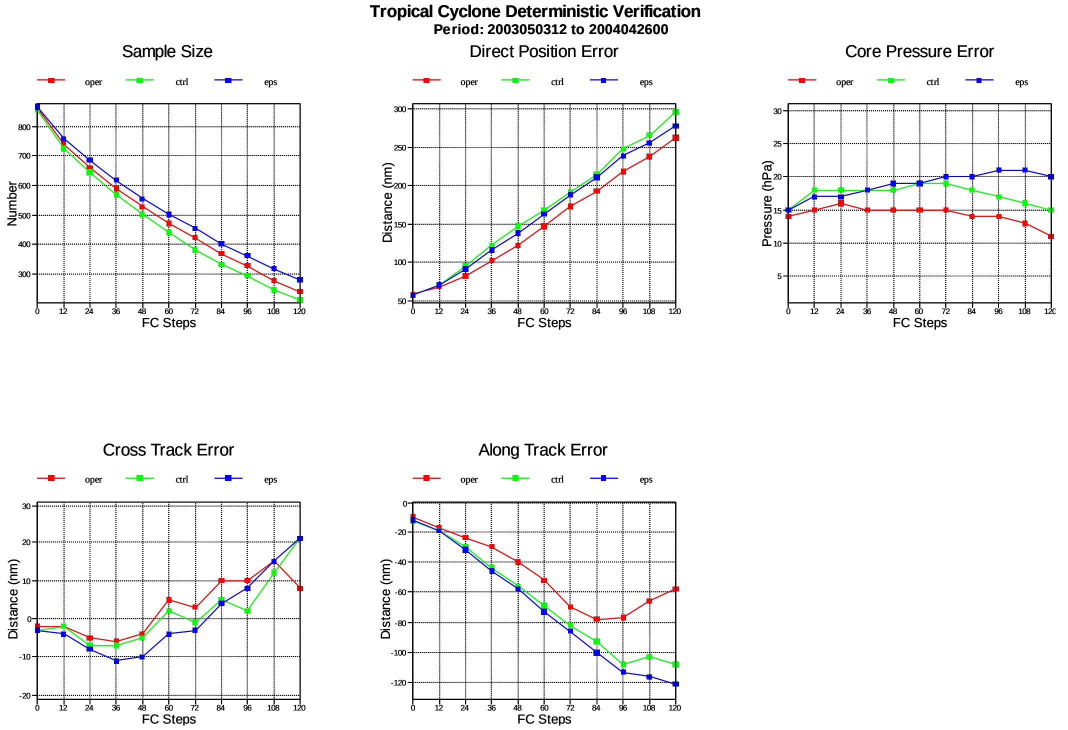


Figure 35:Verification of Tropical Cyclone forecasts from the deterministic, T511 forecast (blue), EPS T255 Control (red) and mean position/ intensity averaged among all cyclones tracked in each member of the ensemble forecast (green).

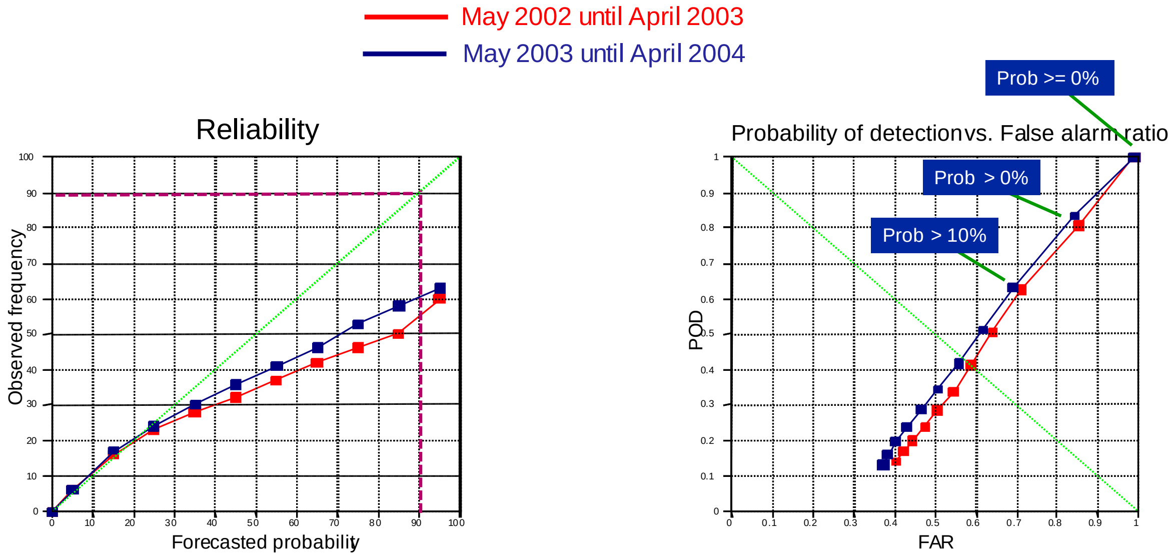


Figure 36: Probabilistic verification of TC strike probabilities: left are reliability diagram (the closer to the diagonal the better), right are Probability of detection (H/H+M)/ False Alarm Ratio (F/F+H)diagrams (the closer to the upper left corner the better). In both cases, the different points are for different probability thresholds. The improvement in the forecast quality from 2002/2003 to 2003/2004 can easily be seen there.

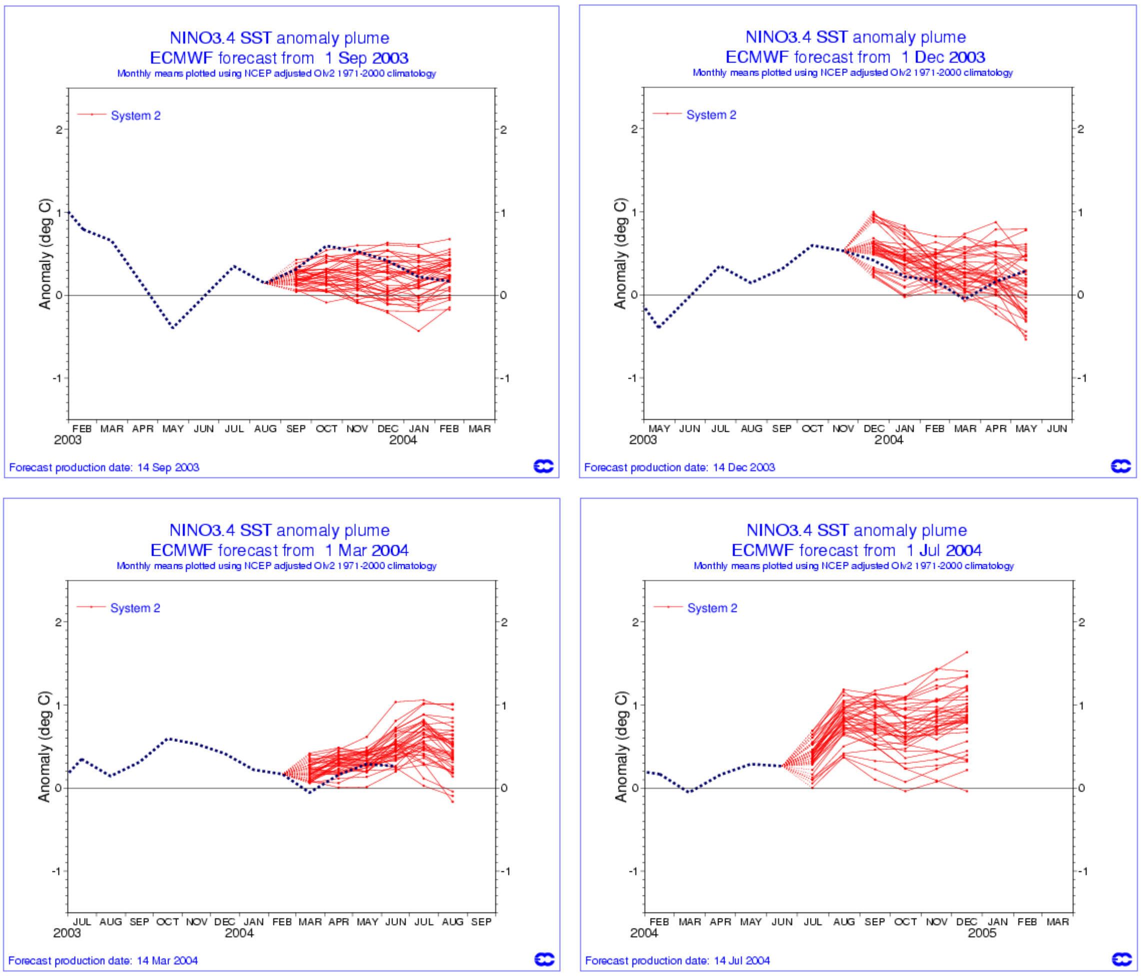


Figure 37: Plot of forecasts of Nino-3.4 at four start dates September, December 2003 March and July 2004. The red lines represent the 40 ensemble members. The heavy dashed line represents subsequent verification.

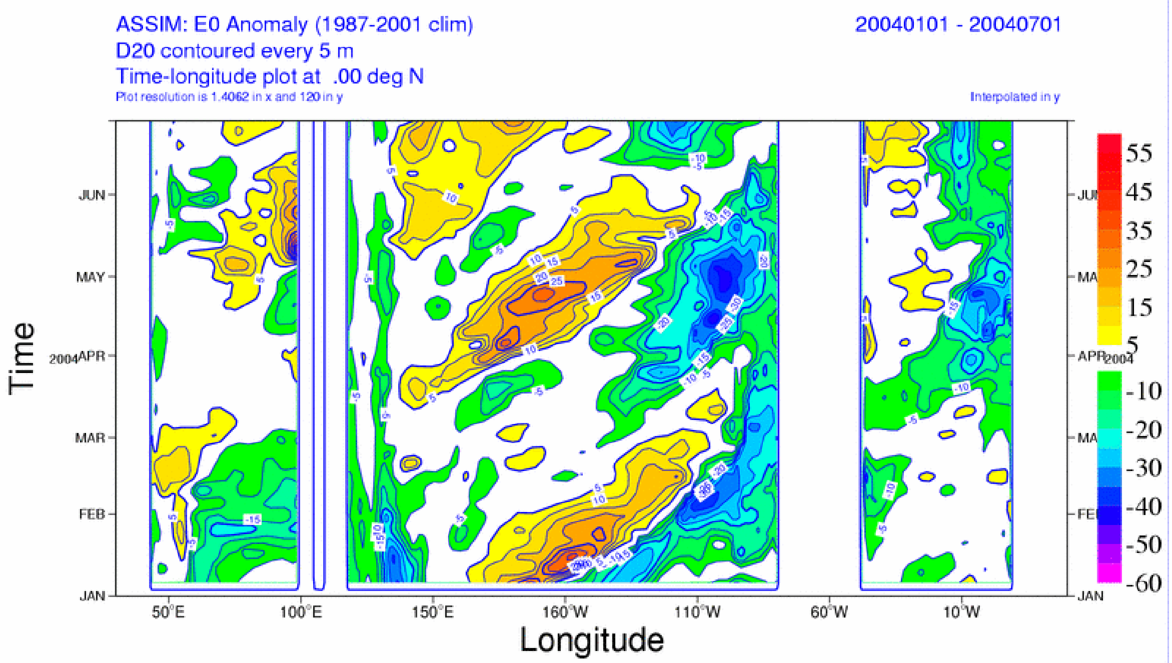


Figure 38: Time-longitude section of the 20C isotherm depth anomalies at the equator. Longitudes are represented by the X-axis. Time is represented by the Y-axis and it ranges from January 2004 to June 2004.

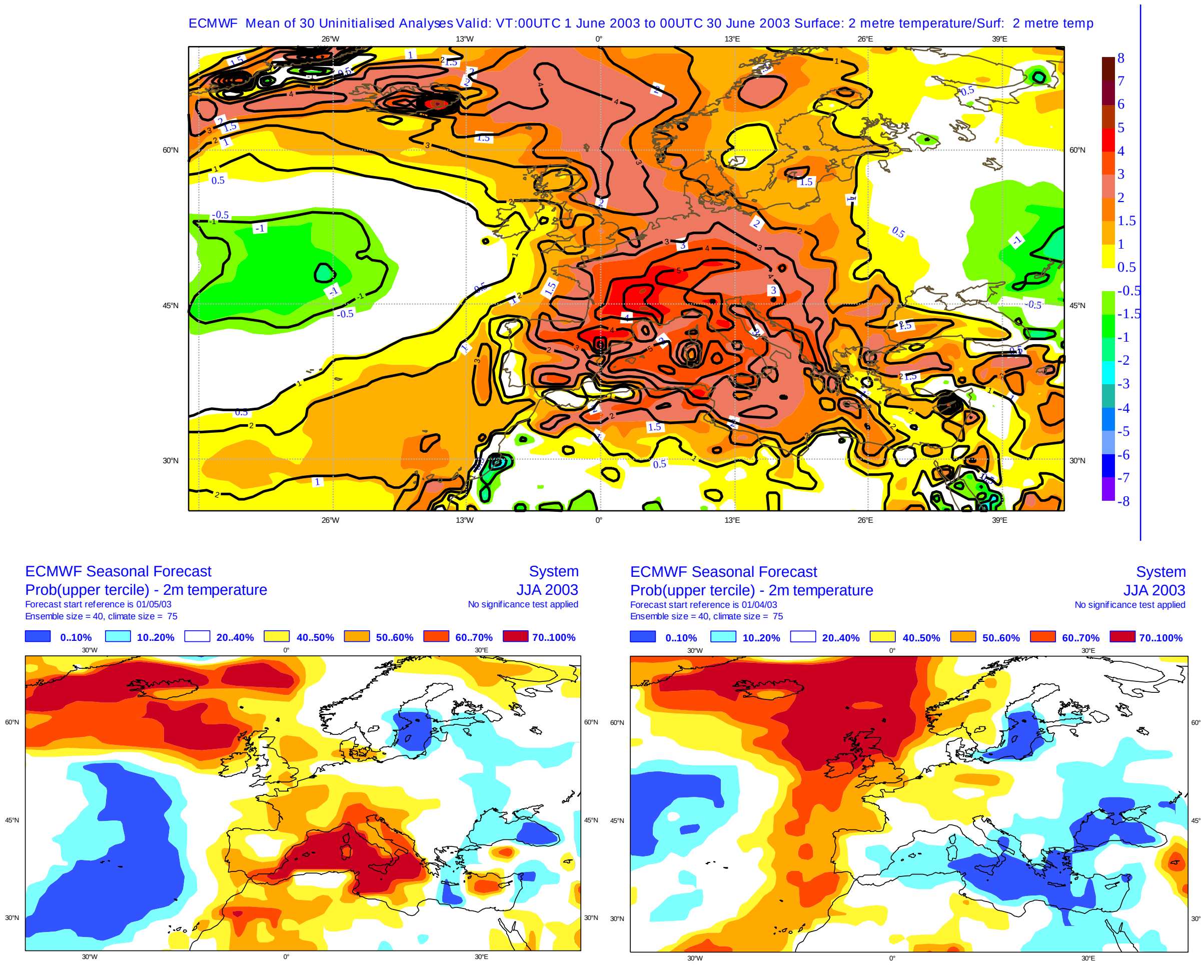

ECMWF Mean of 30 Uninitialised Analyses Valid: VT:00UTC 1 June 2003 to 00UTC 30 June 2003 Surface: 2 metre temperature/Surf: 2 metre temp
26°W
13°W
0°
13°E
26°E
39°E
60°N
45°N
30°N
8
7
6
5
4
3
2
1.5
1
0.5
-0.5
-1.5
-1
-2
-3
-4
-5
-6
-7
-8
ECMWF Seasonal Forecast
Prob(upper tercile) - 2m temperature
Forecast start reference is 01/05/03
Ensemble size = 40, climate size = 75
System
JJA 2003
No significance test applied
0..10%
10..20%
20..40%
40..50%
50..60%
60..70%
70..100%
30°W
0°
30°E
ECMWF Seasonal Forecast
Prob(upper tercile) - 2m temperature
Forecast start reference is 01/04/03
Ensemble size = 40, climate size = 75
System
JJA 2003
No significance test applied
0..10%
10..20%
20..40%
40..50%
50..60%
60..70%
70..100%

Figure 39: Upper: Summer (JJA) 2003 2m temperature anomalies with respect to ERA-40 (1958-2001) (bold black contour show the anomaly normalised by the climate variability). Lower: Forecast probability of exceeding the upper tercile in the model climate distribution for 2m temperature in the same JJA 2003 period. Seasonal forecast starting date is 1 May 2003 (left) and 1 April 2003 (right).

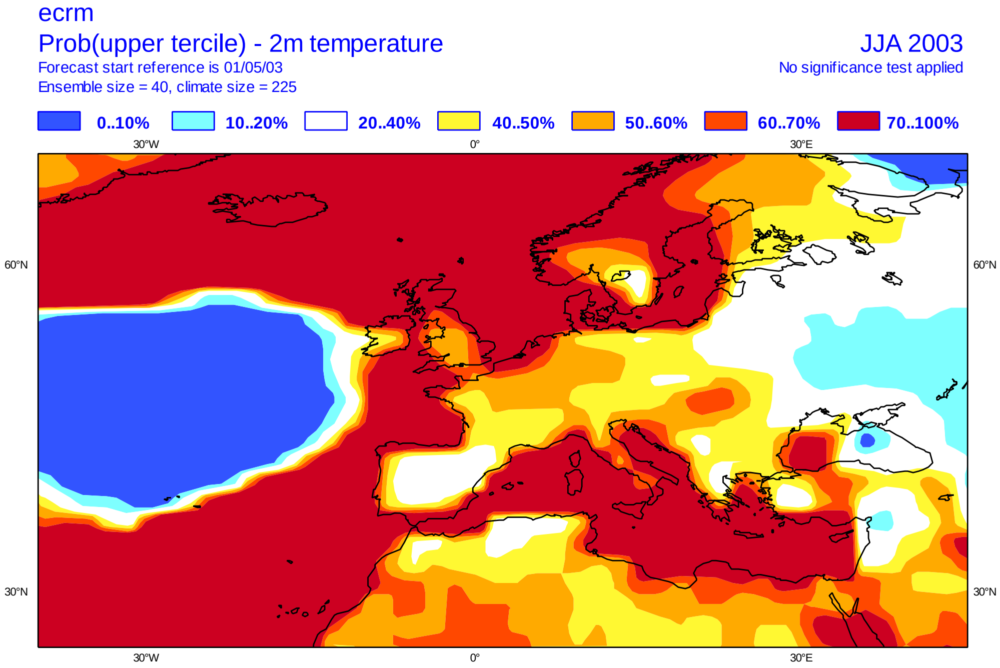

ecrm
Prob(upper tercile) - 2m temperature
JJA 2003
Forecast start reference is 01/05/03
No significance test applied
Ensemble size = 40, climate size = 225
0..10%
10..20%
20..40%
40..50%
50..60%
60..70%
70..100%
30°W
0°
30°E
60°N
30°N

Figure 40:Probability of exceeding the upper tercile of 2m temperature, in the model climate distribution, during June-July-August 2003 given by an ensemble of atmospheric simulations forced by observed SST. The starting date is 1 May 2003.

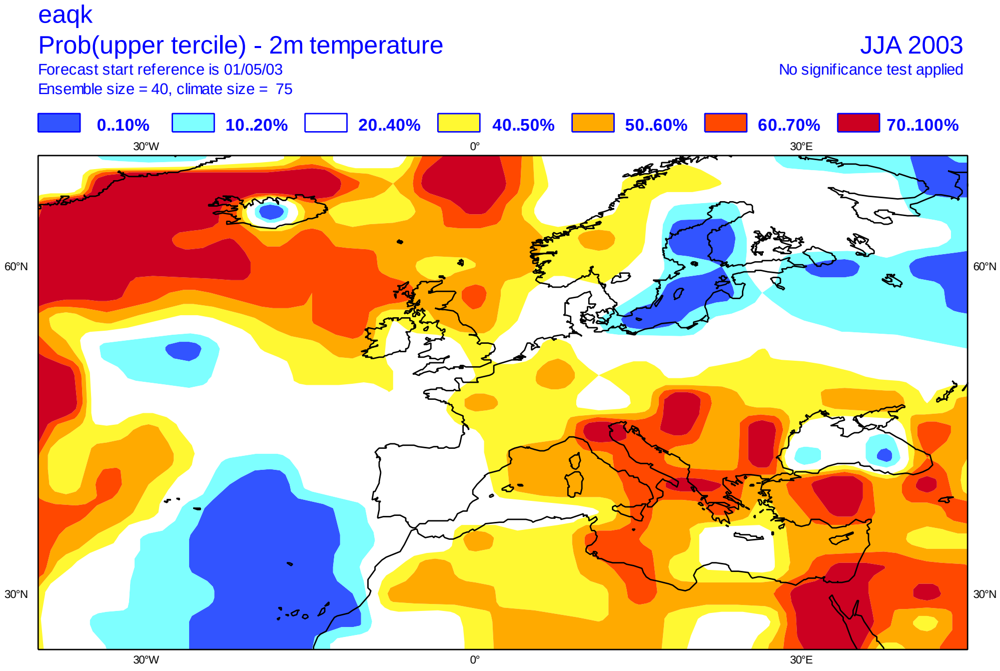

eaqk
Prob(upper tercile) - 2m temperature
JJA 2003
Forecast start reference is 01/05/03
No significance test applied
Ensemble size = 40, climate size = 75
0..10%
10..20%
20..40%
40..50%
50..60%
60..70%
70..100%
30°W
0°
30°E
60°N
30°N

Figure 41:Probability of exceeding the upper tercile of 2m temperature, in the model climate distribution, during June-July-August 2003 given by the UKMO seasonal forecasting system. The forecast ensemble starting date is 1 May 2003.

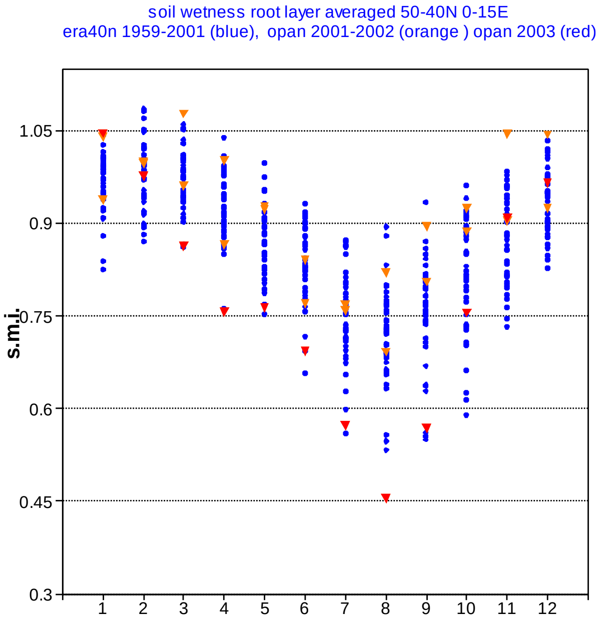


Figure 42: Root layer soil moisture index (SMI) averaged over 50-40N, 0-15E. The blue circles represent the ERA-40 monthly means from 1959 to 2001; the orange triangles represent the operational analysis monthly values for 2001 and 2002 and the red triangles the operational analysis for 2003.

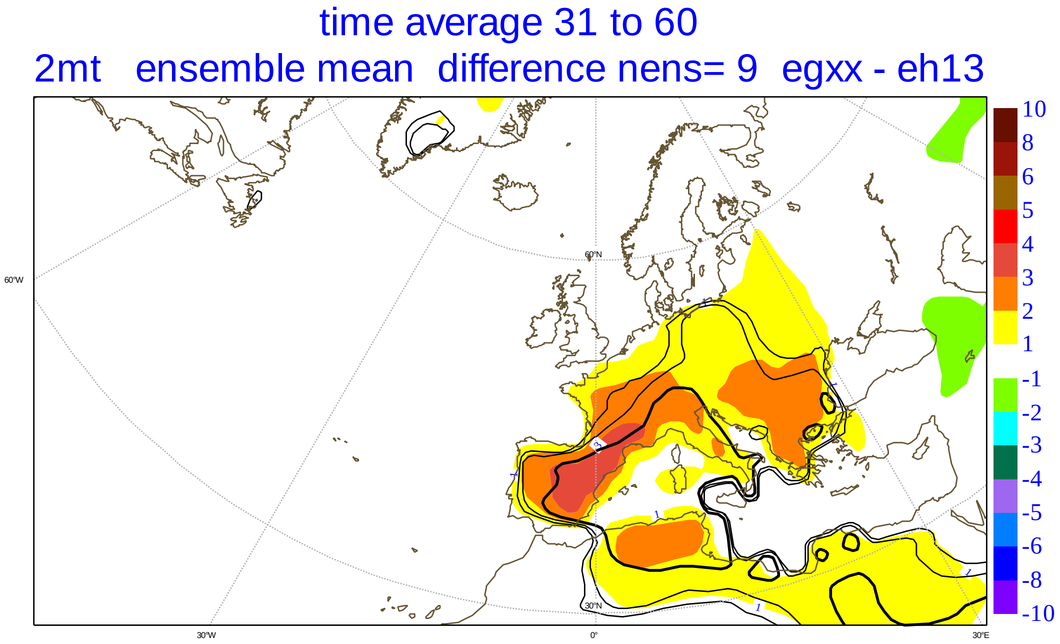


Figure 43:2m temperature ensemble mean differences. The first ensemble has soil wetness initial conditions prescribed to a uniform value of SMI=25 and the second one to a value of SMI=75. The soil wetness initial conditions are prescribed only over most of the European continent (37-60 N; 10W-30E). Black solid contour indicate areas with a level of significance higher than 90% (thin line), 95% (medium line) and 99% (thick line).