



Calibration of EPS derived probabilities

C.J. Kok

Scientific report = Wetenschappelijk Rapport; WR 2001-04

De Bilt, 2001

PO Box 201, 3730 AE De Bilt

The Netherlands

Wilhelminalaan 10

<http://www.knmi.nl>

Telephone +31 30 22 06 911

Telefax +31 30 22 10 407

Author: C.J. Kok

UDC: 551.509.51

551.509.313

(4)

ISSN: 0169-1651

ISBN: 90-369-2201-1



Calibration of EPS derived probabilities

C. J. Kok

Contents

1.	Introduction	3
2.	Data	6
3.	Evaluation tool	8
A.	Talagrand Diagrams or Rank Histograms	
B.	Cumulative Rank Histograms (C.R.H.s)	
C.	Note on the significance of the C.R.H. results	
4.	Impact of system changes	13
5.	Results over 1999 and 2000	32
6.	Calibration	35
7.	Concluding remarks	39
	Acknowledgement	40
	References	41
	Appendix	42

1. Introduction

In medium range forecasting EPS is one of the most important tools. Not only its deterministic information is used (the ensemble mean provides excellent forecasts from day 5 or so), but in particular in probabilistic sense it is one of the main guidelines in operational forecasting. The way in which probabilistic information is usually extracted from EPS is simply by counting the fraction of ensemble members that exceed a particular threshold of interest and then interpreting this fraction as the probability that this threshold will be reached. In this paper we will refer to these probabilities as *EPS probabilities*. Usually, this information is very qualitatively assessed from the presentation of the EPS output in the form of the so-called plume plots. These plots are available for a fixed set of meteorological quantities (predictands) and locations. For The Netherlands this is at the moment in the order of 6 locations over land and 3 at sea. The predictands are temperature, precipitation, wind and cloudiness for the land stations, and mainly wave height at the sea stations.

As is well known, the plumes are constructed from deterministic model output from 51 individual runs of a larger scale version of the “operational” model. Their values are representative for grid boxes. It is common practice, however, to interpret the EPS probabilities as being valid for specific locations inside those grid boxes. Depending on the correlation radius of the particular predictand and the characteristics of that location this can be a good or a rather poor assumption. One of the main objectives of this report is to analyze and quantify the validity of this assumption as a function of lead time for all available predictands at all stations for which plume plots are available.

There are many more ways in which probabilistic information obtained from EPS can be used. For instance, EPS yields a probability distribution from which the forecaster can deduce the likelihood of different “scenarios”. There can even be an indication of a bifurcation in the solutions of EPS. In that case a deterministically best forecast could be a value in the least likely part of the probability distribution (e.g. between the two maxima), whereas the forecaster could express the probability of the two distinct scenarios. This is only appropriate, of course, if EPS predictions of bifurcations are skilful. These matters will not be studied explicitly in this report.

Another important probabilistic application of the plumes may be in the field of predicting (more or less) extreme events. Sometimes EPS produces a so-called outlier which is climatologically rare. An outlier is a value of the predictand which is more or less isolated from the rest of the plume. It is sometimes believed that this indicates that there is indeed a small possibility for the occurrence of an “extreme” event. In the analysis of EPS probabilities that we present in this report the subject of extremes will not be addressed explicitly. All probability statements are analyzed together. More explicitly, only the relative order of the values of the ensemble members is considered and whether or not they are isolated from other members is not taken into account. The same is true for bifurcations or any other specific feature in the pdf.

An important requirement probability forecasts have to meet is their *reliability* or *statistical consistency*. By this we mean that a probability statement, of say 10% that an event will occur, will verify in exactly 10% of the cases. This is not only intuitively reasonable but is also an absolute constraint if one wants, for instance, to use these probabilities to assess the economic value of the forecasting system (see e.g. Richardson, 2000). The reliability can be quantified, which is also called reliability, and this is an important term of the decomposition of the Brier Score (Murphy, 1973; and Wilks, 1995). The Brier Score (Brier, 1950) is the mean square error of the probability forecasts. However, perfect reliability does not necessarily

mean, of course, that the forecasts are good or that they perform better than an unskilful reference forecast system. For instance, climatology forecasts, i.e. forecasting the event with its climatological probability, are usually almost perfectly reliable but are not regarded as having skill in any definition of the word.

As mentioned above, the evaluation of the statistical consistency of the EPS probabilities verified against single station observations is the primary goal of this study. It will be done with essentially one tool. That is the Talagrand Diagram or Rank Histogram. However, we have adopted a different way of presenting this histogram that makes it easier to assess its general characteristics. It will be called the *cumulative rank histogram* (C.R.H.). Consistent differences between the EPS probabilities and the observed frequencies will then be used to match, or calibrate, the two. Since the EPS probabilities are given in steps of (almost) 2 percent we are able to arrive at a nearly continuous calibrated pdf.

Calibration of probabilistic forecasts has been the topic of a number of studies in recent years. In all cases only precipitation forecasts were considered. In a series of papers Krzysztofowicz and Sigrest (1999a, 1999b) discussed the degree of statistical consistency of probabilistic quantitative precipitation forecasts (PQPFs) for two river basins made by forecasters of the Weather Service Forecast Office in Pittsburgh in the beginning of the 90s. They proposed to perform a “recalibration” of the forecasts as well as individualized training of the forecasters.

Hamill and Colucci (1997, 1998) investigated the statistical consistency of short-range PQPFs from the experimental Eta-Regional Spectral model ensemble. They analysed a limited number of cases and found that the ensemble consisting of 15 members was underdispersive. Their verification rank histogram was constructed by averaging the results of a few hundred stations. They concluded that the precipitation forecasts could be post-processed rather simply to correct for their undervariability.

Finally, Eckel and Walters (1998) calibrated PQPFs based on the NCEP medium-range forecast ensemble. The forecasts were verified against grid box representative precipitation amounts constructed by averaging the observations of many rain gauges within the 2.5 by 2.5 degrees grid boxes. They too found that calibration enhanced the performance of probabilistic precipitation forecasts. Both Eckel and Walters as well as Hamill and Colucci stratified the data into categories with different standard deviations of the ensemble forecasts in order to arrive at different calibrations for each category.

The present study primarily differs from the ones above in that it focuses on the statistical consistency (and subsequent calibration) of individual stations (instead of pooling a large number of stations) and also in the fact that many more predictands other than precipitation are considered.

In section 2 a description of the data set that is used in this study is given as well as the meteorological quantities and stations for which the evaluation has been performed. Next the principal evaluation tool is extensively described. The Talagrand Diagram is only briefly discussed but the advantages and disadvantages of its cumulative version are described much more elaborately. The results are presented mainly in the form of this new tool. In order to establish a more or less homogeneous (in the sense of statistical properties) set on which the calibration can be performed, the impact of two major system changes to EPS will be discussed. This is done in section 4. Also the reasons for stratifying the data into warm and cold 6 months periods are given. This lead to the conclusion that only the remaining two years of data after these two changes should be used to calibrate the ensemble probabilities. The behaviour of these two years in terms of the C.R.H. is described in section 5. On this data the calibration has been based which is described in section 6. This calibration can be used to

construct confidence intervals in the daily plume plots. Examples of this are also given in this section. The paper is concluded by a brief summary and conclusions.

2. Data

The EPS probabilities have been investigated on a data set running from 17 December 1996 until 1 October 2000. The start date has been chosen because of the major changes that were introduced to EPS on 10 December 1996. These changes included the introduction of model cycle 15R7 in EPS as well as in the deterministic model, with, among others, changes in snow albedo and humidity analysis. But moreover, on that date EPS was upgraded to run at T_L159 resolution with 31 levels in the vertical and the number of perturbed ensemble members was increased from 32 to 50 members. The control run was also made at T_L159L31 (with physics on a N80 Gaussian grid).

However, because of the introduction of a number of major changes in the EPS system we will first investigate the impact of these changes on the statistical consistency of the probabilities. This was done to determine the period on which the calibration had to be performed. Here we will only consider two major updates of the EPS system. The first is the one that was implemented on 25 March 1998. On that date the method used for the computation of initial perturbations of EPS was changed significantly. From that day, the initial perturbations for day D are a combination of the fastest growing modes between day D-2 and day D (Evolved Singular Vectors) and the fastest growing modes between day D and day D+2 (ECMWF Newsletter number 79, spring 1998). The second major change in the operational EPS system whose impact will be investigated in this report is the introduction of the so-called stochastic physics which was implemented on 21 October 1998.

These changes in the operational forecasting system were expected to have a more or less significant meteorological impact in the sense that the spread of the ensemble would show an increase in the early medium range. Since the spread is one of the main topics of interest in this calibration study we will first investigate the consequences of these changes as a function of forecast range. It turned out that the period after the two system changes showed remarkably different characteristics with respect to the earlier part of the data set for most of the meteorological quantities we looked at. This will be demonstrated in section 4. The final calibration which is intended to be used operationally is therefore based on the period from 21st of October 1998 until the first of October 2000.

The data are stratified into two 'seasons': a hereafter called 'summer' running from 1 April until 1 October, and a 'winter' season consisting of the remaining 6 months. The reasons for this will be outlined in section 4.

The evaluation has been performed on the data which are plotted on a daily basis in the form of the well-known plume plots. These data are disseminated for the requested predictands and stations in so-called weather parameter files. The data for the surface fields are constructed from the basic model grid by bi-linear interpolation from the four nearest original filed grid points. The surface fields of the EPS T159 forecasts are represented in N80 reduced Gaussian grid. We have only used values with 12 hours interval for the whole forecast range of EPS (analysis time excluded). All 51 members are used, so data of the control run are included in our analysis. The plumes are available for six synoptic land stations (Fig.1). The meteorological quantities that are examined are two-meter temperature (T), rainfall amount (RR), cloud cover (N) and 10-meter wind speed (ff). These are "verified" against station observations. We have to stress that no area average values for the quantities have been calculated, so care must be taken in interpreting the results.

In addition, for three North Sea platforms the same analysis has been performed for significant wave heights (H_s) derived from plume information from the wave-EPS. These three stations are Europlatform (or EURO), K13 and Meetpost Noordwijk (MPN). Their location is

given in Fig. 1. The wave height plumes have been constructed from field information on a 0.5x0.5 degrees grid retrieved from the MARS-system of ECMWF. To obtain time series for K13 we have interpolated the 4 surrounding grid point values, whereas for the other two stations we have taken the nearest grid point at sea. For details about the wave-EPS the reader is referred to Vogelesang and Kok (1999). For MPN data are only available since 1 March 1999, for the other two stations since 1 October 1998. Despite a number of small changes in model configuration in this period the data will be considered as a homogeneous set. Only winter seasons (1 October until 1 April) will be investigated for the sea locations. A summary of the stations and predictands which are evaluated in this report is given in Table 1.

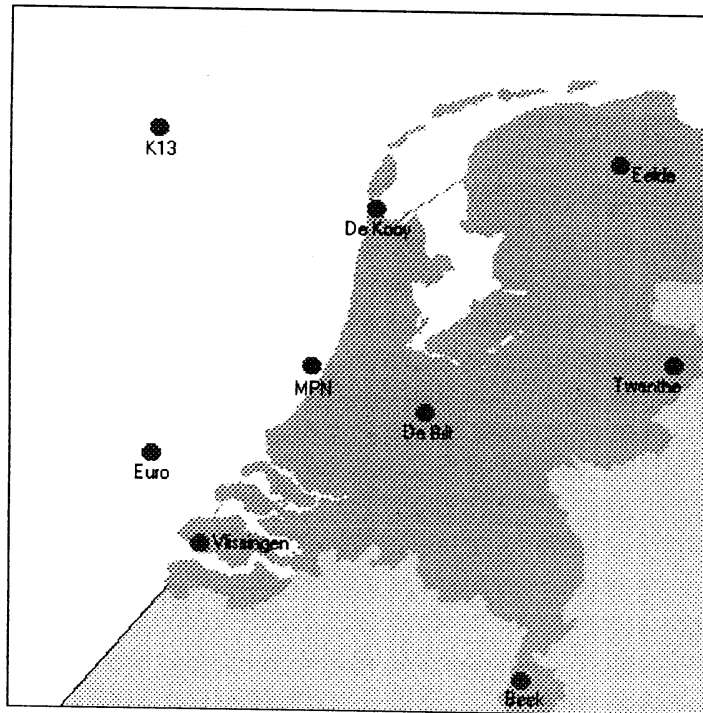


Fig. 1. Location of the stations for which the calibration has been performed.

Tabel 1. Stations (and their WMO-identification) for which the evaluation has been performed as well as the predictands considered.

De Kooy	06235	T, RR, ff, N
De Bilt	06260	T, RR, ff, N
Eelde	06280	T, RR, ff, N
Twente	06290	T, RR, ff, N
Vlissingen	06310	T, RR, ff, N
Beek	06380	T, RR, ff, N
K13	06252	Hs
Meetpost Noordwijk (MPN)	06254	Hs
Europlatform (EURO)	06321	Hs

3. Evaluation tool

The analysis of the probabilities derived from the plume information of EPS has been performed using essentially only one tool, but in two appearances. This is the so-called Talagrand diagram or Rank Histogram. First its construction and advantages as an evaluation tool will be shortly highlighted. For a more elaborate discussion the reader is referred to e.g. Talagrand et al. (1997), Anderson (1996) or Hamill and Colucci (1998). Also the integrated or cumulative Talagrand diagram will be introduced in this section. Its main characteristics and interpretation will be discussed as well as the advantages and disadvantages with respect to the Talagrand diagram.

A. *Talagrand diagrams or Rank Histograms*

Rank Histograms are computed to evaluate differences between forecast and expected probability density functions (pdfs). In the case of EPS the predicted pdf for temperature for instance, is defined by the 51 values produced by the ensemble forecasts (the control run included). When we rank these values in increasing order we get 52 intervals. If the verifying observation is an independent realization of the same pdf which has produced the ensemble temperatures then the observation is statistically undistinguishable from the ensemble temperatures. This is true for a hypothetical “perfect” ensemble prediction system in which all forecast errors are due to errors in the initial state and from which the initial perturbations are randomly selected. Also, the forecast members are assumed to have independent and identically distributed errors (Hamill and Colucci, 1998). In that case the observation will fall with equal frequency of $1/52$ in each of the intervals (Mullen and Buizza, 2001; Talagrand et al., 1997). Therefore, the histogram of the position of the observed temperatures with respect to the ensemble temperatures defines a measure of the statistical consistency of the EPS temperature forecasts. A perfectly consistent ensemble system will produce flat histograms. The deviation of the resulting observed histograms from a flat one can be quantified, for instance by summing the squared differences, and this turns out to be a very good and convenient measure of the statistical consistency (Kruizinga, 1999). Note however, that statistical consistency is by no means sufficient to render skilful forecasts.

Very frequently the observation will be equal to one or even more ensemble members. In those cases the rank that has to be assigned to the observation is not trivial. This situation occurs of course on an appreciable number of occasions for precipitation and cloudiness (for instance a number of forecast values as well as the observed value are equal to zero), but we cannot disregard the amount of cases for the other meteorological quantities either. Note that we have rounded off all quantities, e.g. wind speed in 0.1 m/s, temperatures in 0.1 °C and wave height in 0.1 m. In order to assign the ranks without introducing biases in the results additional rules must be specified. We have applied the following procedure which ensures that the observation will be assigned with equal probability to one of the ranks. For each case in which the observation equals one or more, say N , ensemble members we take N randomly chosen numbers (Press et al., 1994). This comprises $N+1$ ranks. Furthermore, we take one additional random number representing the observation. We then determine the position of this last number with respect to the ordered set of N random numbers. This position is used to determine the appropriate rank. In this way the observation is randomly assigned to one of the $N+1$ ranks.

A similar approach was used by Hamill and Colucci (1998) who generated a set of very small uniform random deviates and added those to the ensemble members that are equal and also added another random number to the observation. In this way the observation can uniquely be assigned to a rank.

A few examples of rank histograms are given in Fig. 2 (left column). In the top panel the rank histogram is shown for the +48 EPS 2meter temperatures for station Eelde during the summer six months of 1999 and those of 2000 combined (indicated by SUM 99 00). The rank distribution is clearly not uniform. There is a definite overpopulation of both end ranks (U-shape) indicating that the variability of the ensemble is not large enough: the ensemble is under-dispersive. Vice versa, if the verifying observation falls in one of the end ranks less frequently with respect to chance, then the spread of the forecast system may be too wide. A L-shaped histogram, with an overpopulation of only one end rank, can be a manifestation of a clear bias, i.e. a systematic under- or overforecasting of the particular quantity. In that case it is difficult to make any definite inference about the difference in spread between the two pdfs. An example of this is given in the middle panel of Fig. 2 for +168 wave height forecasts for platform Meetpost Noordwijk (06254) for the combined winter seasons of 1999 and 2000. The wave heights were in general highly exaggerated, or in other words, the probabilities of wave heights not exceeding certain thresholds was highly underestimated. Finally, in the bottom left panel of Fig. 2 the histogram for the +240 precipitation forecasts for station Beek is shown. It is based on only a few summer months of the years 1999 and 2000. In contrast to the two other panels this one is not used in the calibration but merely serves as an example of a rank histogram that exhibits a double maximum. There is an overpopulation of the center ranks (bell-shaped) but also near one of the ends, thus resembling characteristics of both over- and underdispersion.

B. *Cumulative Rank Histograms (C.R.H.s)*

In this report we have applied a different way of presenting the Rank Histograms. We have accumulated the ranks from left to right (from lower to higher values), but without the highest bin. In this way we get a monotonously increasing (or rather non-descending) curve starting at a level of the percentage of observations below the lowest ensemble member, i.e. the left end rank, and ending at a hundred percent minus the percentage of observations which is contained in the last bin, i.e. the cases in which the observation was higher than the highest ensemble member. If a section of the curve is horizontal it means that the corresponding bins are empty whereas if it is parallel to the diagonal this means that those bins are populated exactly according to its expected frequency, i.e. 1 over 52. The local derivative equals the bin frequency. If the slope is steeper than 1 (or the increments are higher than the expected value per bin) then the bins are overpopulated, and vice versa. So, the cumulative rank histogram contains exactly the same information as the traditional Rank Histogram but it offers, as we will see, a much more convenient tool to highlight some of its general characteristics and therefore to assess the validity of the EPS probabilities. Needless to say that in the C.R.H.s the same rank assignment procedure as before is included for the cases in which the observation equals one or more ensemble members.

A few examples are given in the right column of Fig. 2. These examples are the cumulative version of the cases for which the Talagrand Diagrams are given in the left column. The U-shaped Talagrand Diagram results in a C.R.H. (top right) with an average slope very much less than one. The highly overestimated wave heights for Meetpost Noordwijk yields a curve much above the diagonal (middle panel). Likewise, an underestimation gives a curve

below the diagonal (not shown). Finally, in the lowest panel the consequence of a double maximum is shown. A bell-shaped histogram with smaller than expected frequencies in both end ranks results in a sigmoid shaped C.R.H. (not shown).

A number of features of the Talagrand Diagram can also be inferred easily from the cumulative diagram. As mentioned above, the slope of the curve is a direct measure of the width of the predicted distribution with respect to the observed one. In fact, the difference between the highest and lowest value of the curve is the percentage of observations that fall between the highest and lowest ensemble values. Its complement equals, therefore, the total frequency of outliers on both sides. Also the asymmetry in the number of upper versus lower outliers can easily be assessed. If the average slope is less than 1 and the 'median' (or rather the 50 percent percentile) is more or less correct, then the simulated distribution is too small and the Talagrand Diagram is U-shaped; vice versa, higher than 1 slopes indicate a too wide distribution and a bell-shaped R.H. Trivially, a curve exactly on the diagonal, i.e. a perfectly flat R.H., yields statistical consistency. As mentioned before, this is not a sufficient requirement for a forecasting system to have skill. For instance, forecasting with the (sample) climatological distribution also yields a cumulative R.H. along the diagonal.

The C.R.H. gives also direct information about the shift or displacement of the predicted distribution. The integral of the curve with respect to the diagonal, or the integral of the curve minus one half (if the dimension is probability instead of percentage), can be regarded as a direct measure of this shift. Positive values correspond to shifts to the right, i.e. to higher values, and vice versa. In fact, this integral is equal to the average of the offsets of all 51 (independent) EPS probability forecasts. By first glance already a fairly good estimate can be obtained. If the medians of the two distributions are the same, then, of course, the curve intersects the diagonal at an EPS probability of 50 percent (see e.g. the top panel of Fig. 2). The differences in the position of the median can very easily be quantified from the C.R.H.. This is much more difficult to see from the rank histograms.

One of the most prominent advantages of the C.R.H. in the context of this paper is the fact that it is very straightforward to assess the long-term observed frequency for arbitrary values of the EPS probabilities. Therefore, it can be used to correct (or calibrate) the consistency 'errors' in the EPS probabilities. For example, from the middle right panel in Fig. 2 it can be seen that if 70% of the ensemble members are below a certain wave height this should be interpreted as a 85% probability.

The C.R.H. can also be used very easily to obtain calibrated EPS probabilities *exceeding* certain (probability) thresholds. This can simply be done by taking the complements in the C.R.H. diagrams: if 30% of the ensemble members exceed a certain wave height then the panel in Fig.2 implies that there is a only a 15% probability for the observed wave height to exceed that threshold.

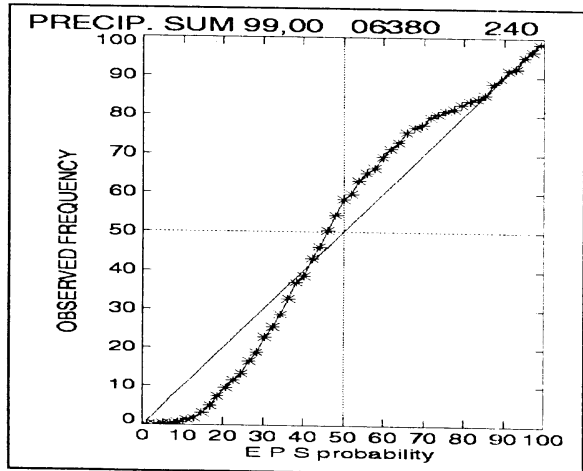
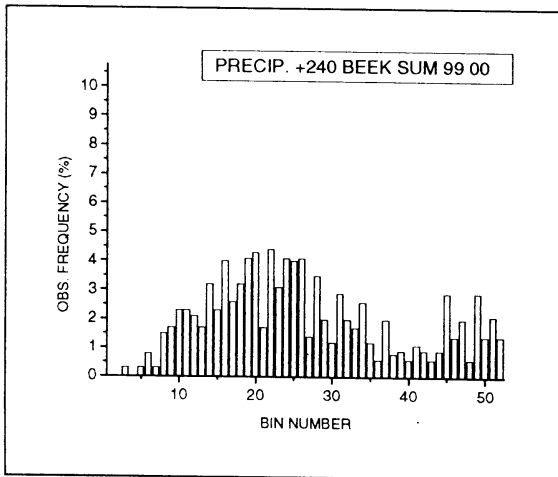
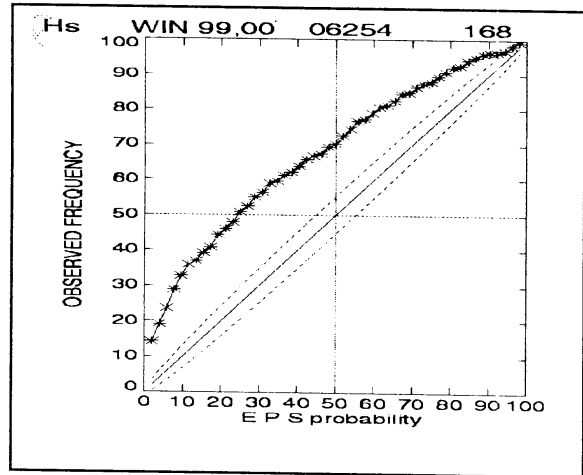
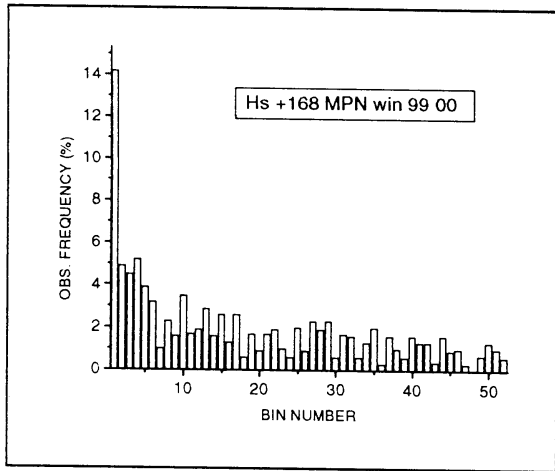
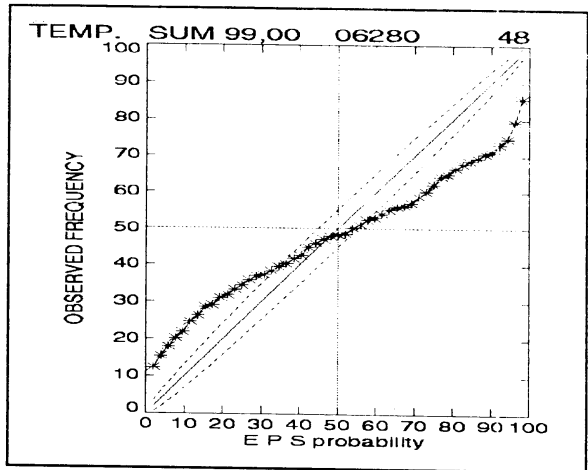
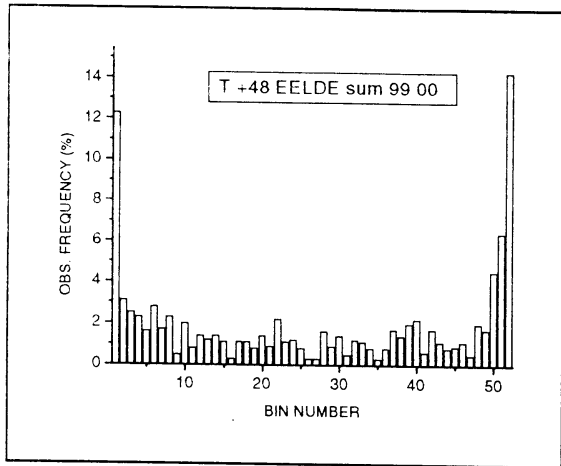


Fig. 2. Examples of rank histograms (left column) and the corresponding cumulative rank histograms (right column). See text for more details.

C. Note on the significance of the C.R.H. results

Although it is not the intention of this work to determine the degree of significance of the results, it may be interesting to see whether the deviations from the diagonal in the C.R.H.s are significant. If they are not then calibration on the basis of the C.R.H. can not be expected to improve the skill of the probability forecasts. In this section two calculations are given which may serve as a reference to interpret the significance of the C.R.H. results.

The first approach is to see whether the deviations between the observed C.R.H. curve and the diagonal can be considered to be likely or not under the perfect ensemble assumption as outlined in subsection A. The goodness of fit can be approximated using the Kolmogorov-Smirnov (K-S) test. As the theoretical cumulative distribution function the diagonal is taken. Note that differences may arise merely because the fact that we compare grid box averaged predictions to single station observed frequencies. Depending on the correlation radius of the predictand at hand the deviations can be expected to be quite large. Nevertheless it is interesting to take the uniform distribution over the ranks as a reference.

The K-S test statistic looks at the maximum value of the absolute difference between the empirical and fitted cumulative distribution functions. If this distance is sufficiently large then the null hypothesis that the observed data were drawn from the theoretical distribution will be rejected. In other words, we can test the hypothesis that this distribution is the true distribution on a predetermined significance level. The C.R.H.s are based on two summer or two winter half years, therefore consisting of around 360 cases. This implies that if the maximum deviation between the sample cumulative distribution function and the theoretical distribution is more than 7 percent (8.5 percent) then they are significantly different from each other on a significance level of 95% (99%). As we will see in the next chapters it appears that for many (if not most) of the C.R.H.s the perfect ensemble assumption is not met.

A second view that is particularly relevant with respect to the calibration for fixed probability thresholds may come from a comparison of the results with the probability that the same results can be obtained by mere chance. This is done once again under the assumption that we randomly draw from the uniform distribution. As mentioned before, the calibration results are based on around 360 cases. The expected number of occurrences for each of the 52 bins is therefore almost exactly equal to 7. The C.R.H. curve can be regarded as the evaluation tool in terms of observed frequencies for all 51 independent probability forecasts that can be obtained from EPS. For each ensemble probability p the chance of the predictand not to exceed the corresponding value can be regarded as a Bernoulli trial and therefore the number of occurrences follows a binomial distribution. Then for each probability statement p_k , with $p_k = k/52$ and $k=1, \dots, 51$ the expected number of occurrences is equal to $np_k \approx 7k$. The expected frequency is therefore p_k , i.e. exactly on the diagonal. The standard deviation is equal to the square root of $np_k(1-p_k)$. This reaches its maximum for $k=26$ with a value of approximately 9.5. This is equal to about 2.6 percent of the total number of cases. In the top two C.R.H.s of Fig.2 the values of $p_k \pm 2 \text{ sd}$ (with sd the standard deviation) are indicated by dashed lines. For many of the C.R.H. curves presented in the next two chapters it follows that the calibrated probabilities deviate significantly from the EPS probabilities.

4. Impact of system changes

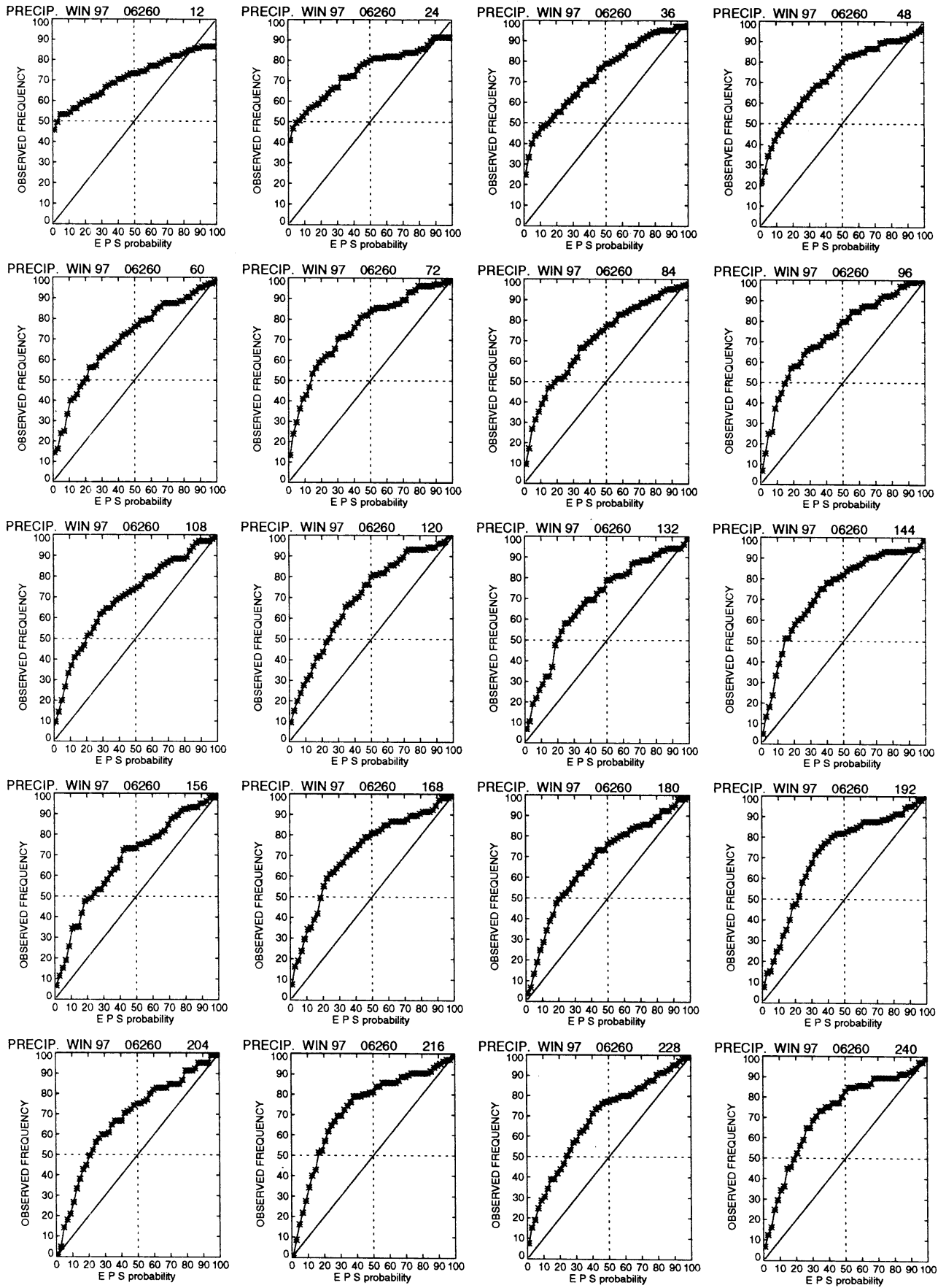
In this section we will analyze the statistical consistency of the EPS probabilities with respect to single station observed frequencies. This will be done mostly in terms of the C.R.H.s as described in section 3, and performed on the data which is stratified into two half-year periods. These periods will be referred to as winters ('WIN' in the figures), ranging from Oktober to March inclusive, and summers ('SUM') from April to September inclusive. The justification for this stratification will be studied as well. Only precipitation and temperature for a single station (De Bilt, 06260) will be discussed in this section. Since we consider EPS plumes from December 1996 onwards we have 4 winter and 4 summer 'seasons'. All seasons therefore consist of about 180 days, except for the winter of '96/'97 which contains only 110 days. The two major model changes (25 March and 21 October 1998), whose impact we will study in this section, are more or less at the beginning and end of the summer half year of 1998 respectively. This enables us to assess this impact by looking at differences in characteristics in the C.R.H. curves between consecutive seasons.

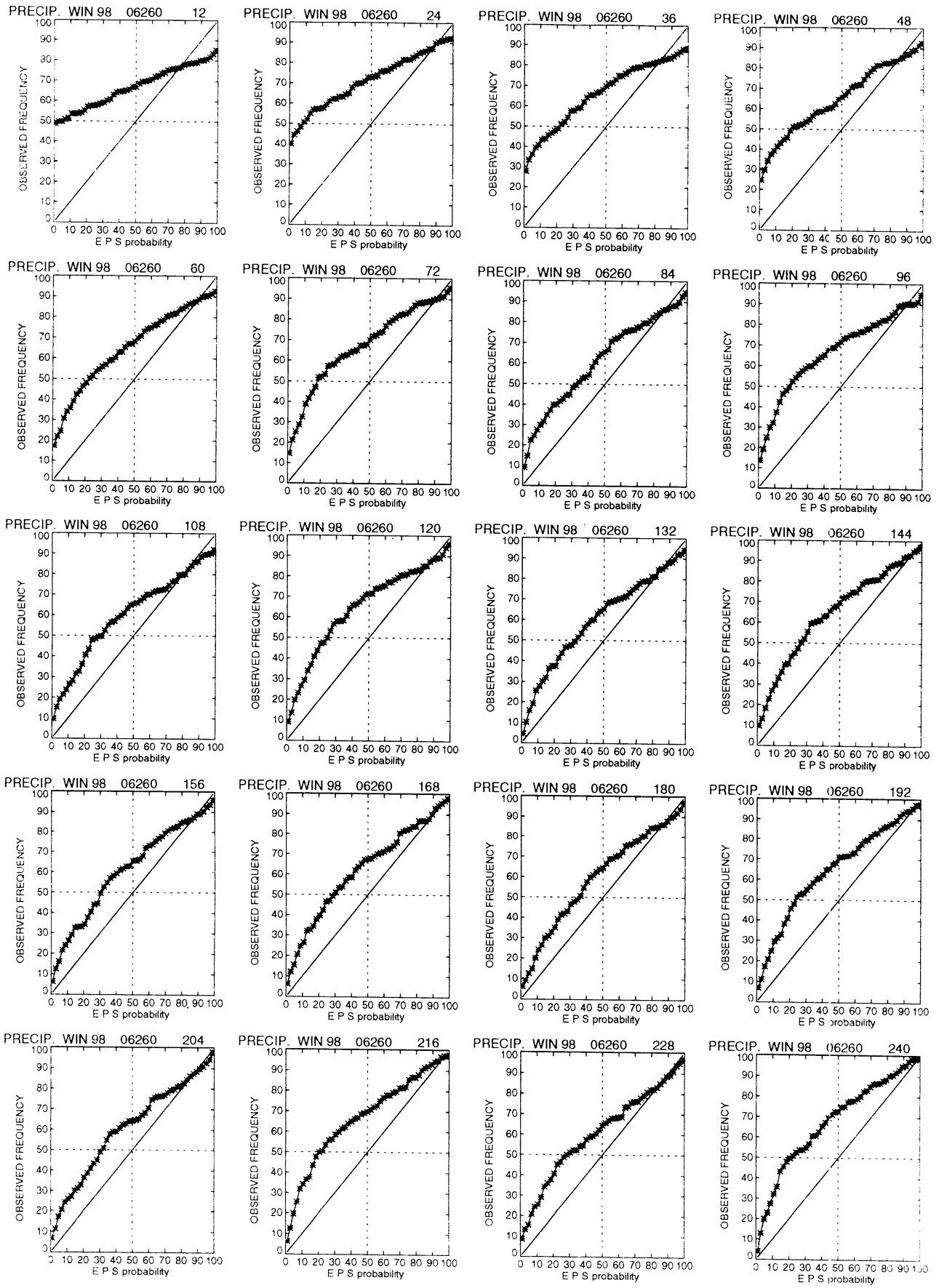
The results for the EPS probabilities of *precipitation* for grid box De Bilt (06260) verified against observed precipitation distributions for station De Bilt are shown in Fig. 3 for the four consecutive winters and summers. The winters of 1996-1997 and 1997-1998 (WIN 97 and WIN 98 respectively), which are both before the two major model changes, show a large underestimation of the spread in the earlier forecast ranges. But the spread is rapidly increasing until at least day 3. There appears to be a large asymmetry in the population of the ranks. The lower ranks are highly overpopulated whereas the ranks for relatively higher precipitation amounts are almost empty. This is true for most projection times equal to or higher than day 3 for both winter seasons. This is to a large part due to a systematic underestimation of the number of cases with little or no precipitation; e.g. too many ensemble members give small amounts of precipitation. The 'median' of the predicted pdf is typically in the order of 20 to 30% too high in the winter of '97 and only a little less in the next winter.

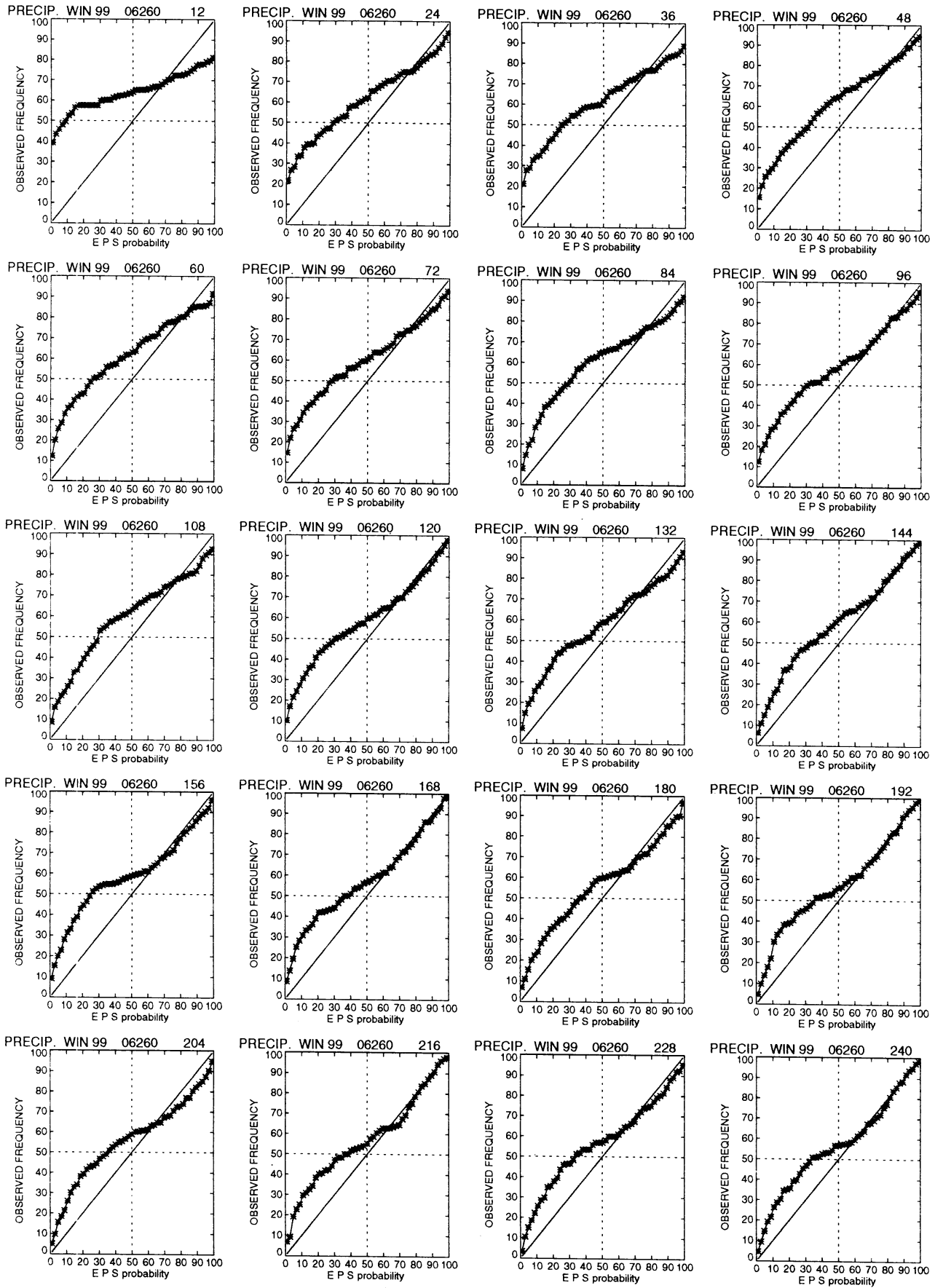
In both years there are systematic differences for the 12-hour period from noon to midnight versus the period from midnight to noon, i.e. +12, +36, ... , +228 (columns 1 and 3 in the figures) versus +24, +48, ... , +240 (columns 2 and 4). The bias is somewhat smaller in the 'afternoons'. Differences between 'afternoon' and 'morning' precipitation appeared to be present in almost all years and seasons. This might be partly related to the fact that the small daily cycle which is present in the observed precipitation amounts (even in our winters) is completely absent in the ensemble mean of EPS. We will come back to this later.

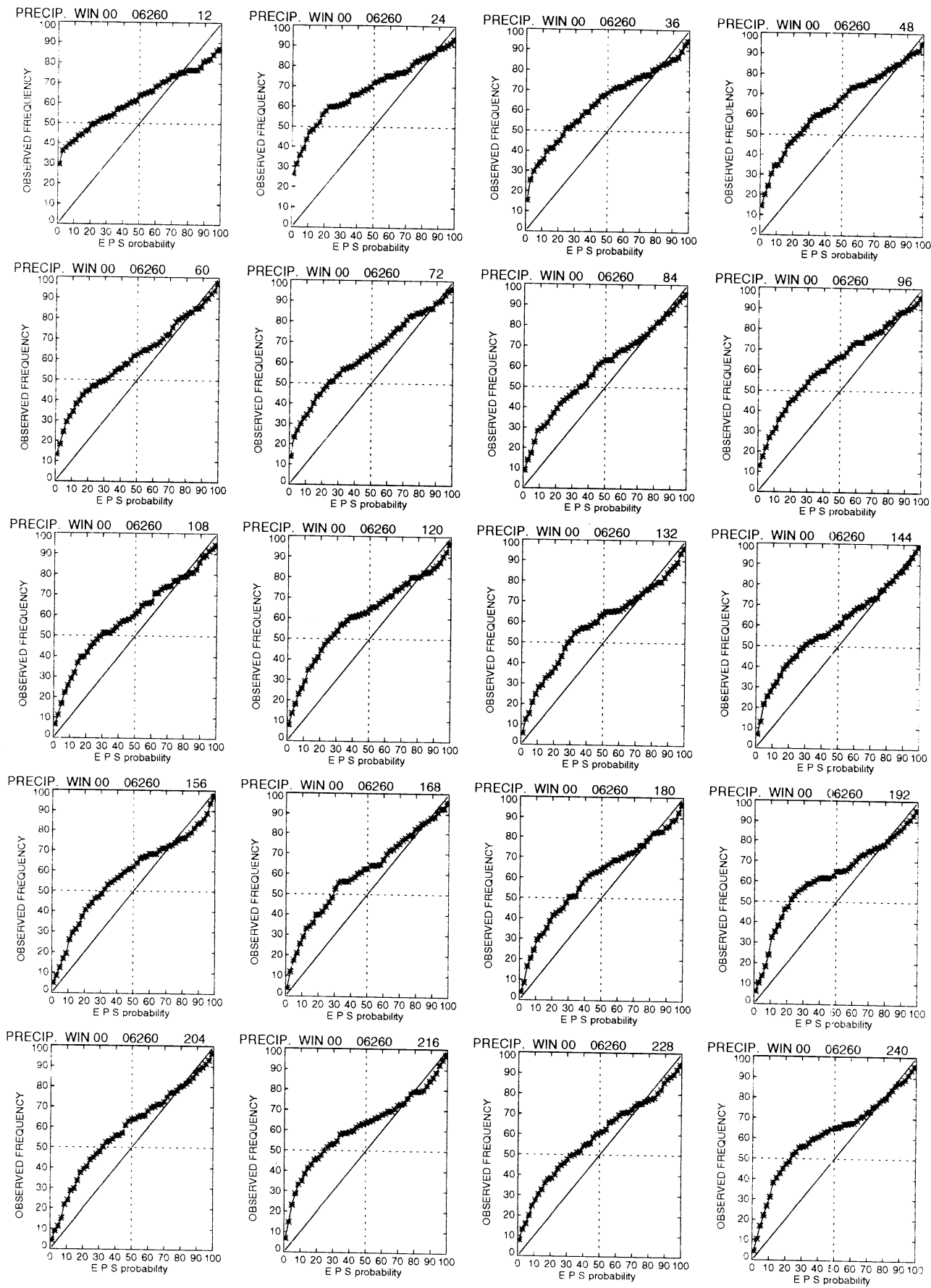
The last two winter periods (WIN 99 and WIN 00, i.e. the cold seasons of 1998-1999 and 1999-2000), show a substantial increase in forecast spread in the forecast range until +60 with respect to the two earlier years. Since these periods are both after the two model changes we cannot say what change is (mostly) responsible for the increased spread. The wet bias has been reduced but remains present. The medians of both distributions are closer to one another, especially in WIN 99. The same but small difference in C.R.H. between morning and afternoon predictions remain. Another prominent change in the last two winters with respect to the first two is the fact that the higher ranks are populated much closer to the expected values. For EPS probabilities higher than, say, 70 percent the C.R.H. curves are very close to the diagonal. This

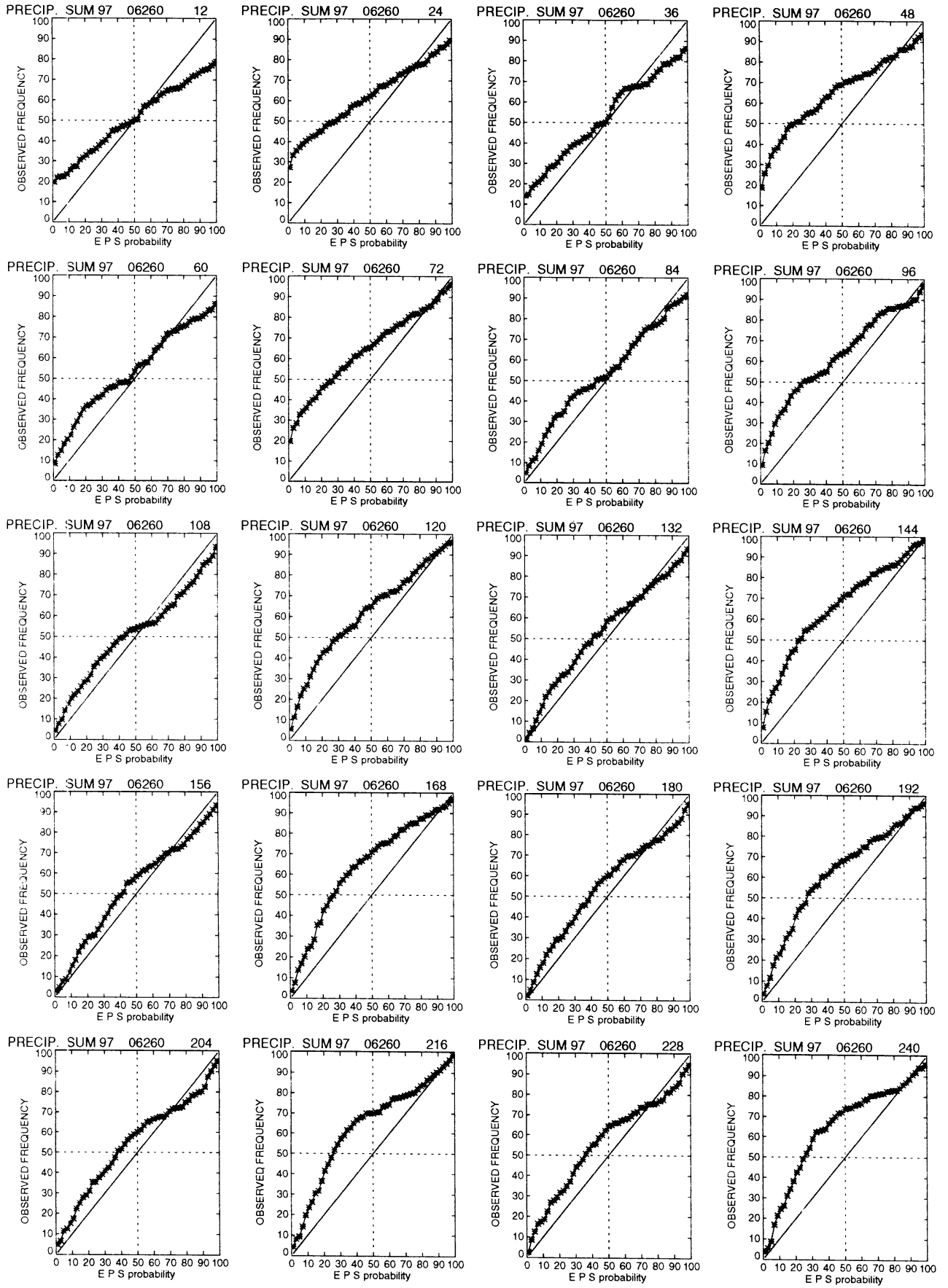
Fig.3 On the following pages the C.R.H.s are shown for all forecasts of precipitation and temperature for location De Bilt separated in winter and summer 'seasons' of the years 1997 until 2000. For more details see the text.

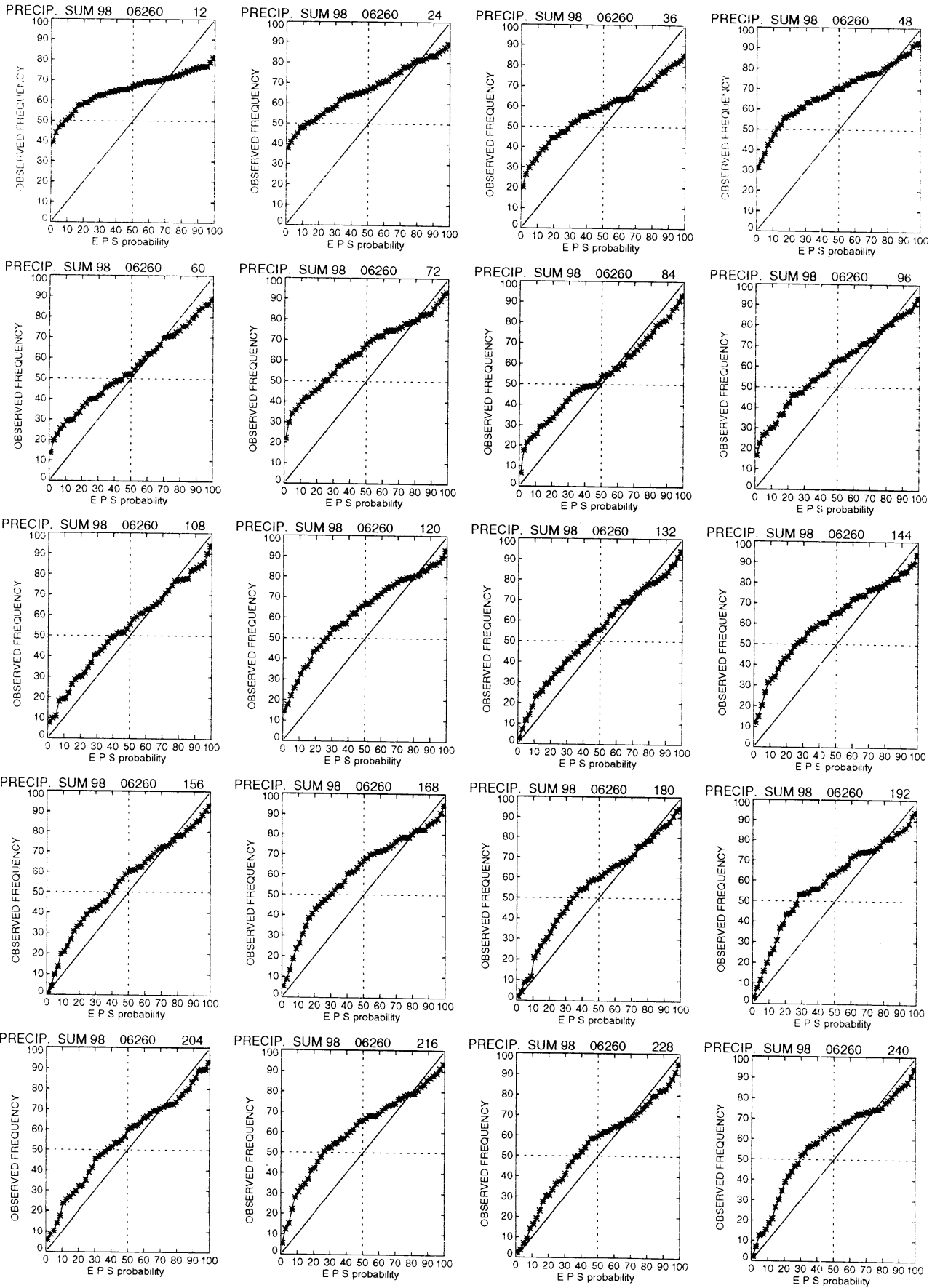


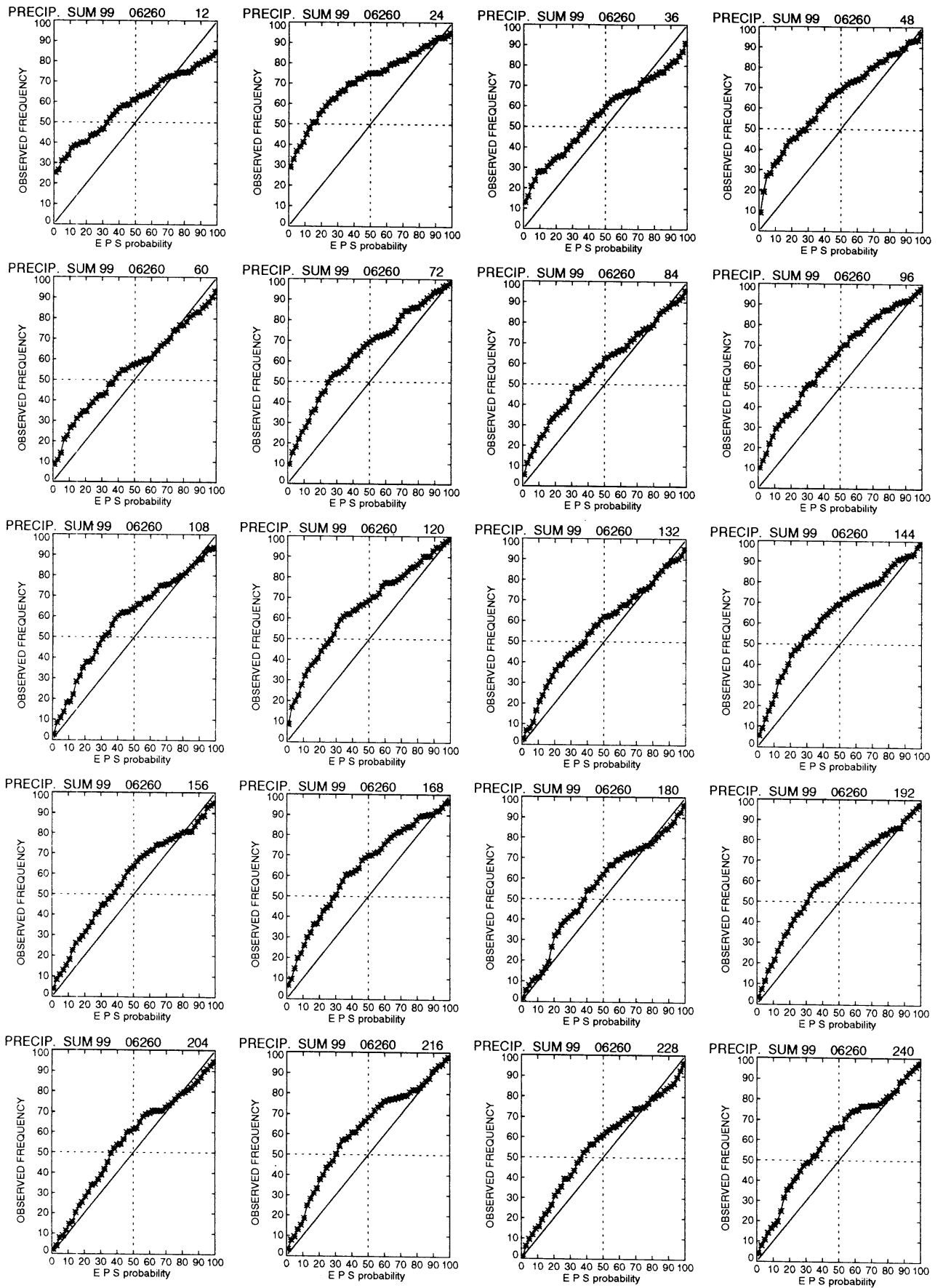


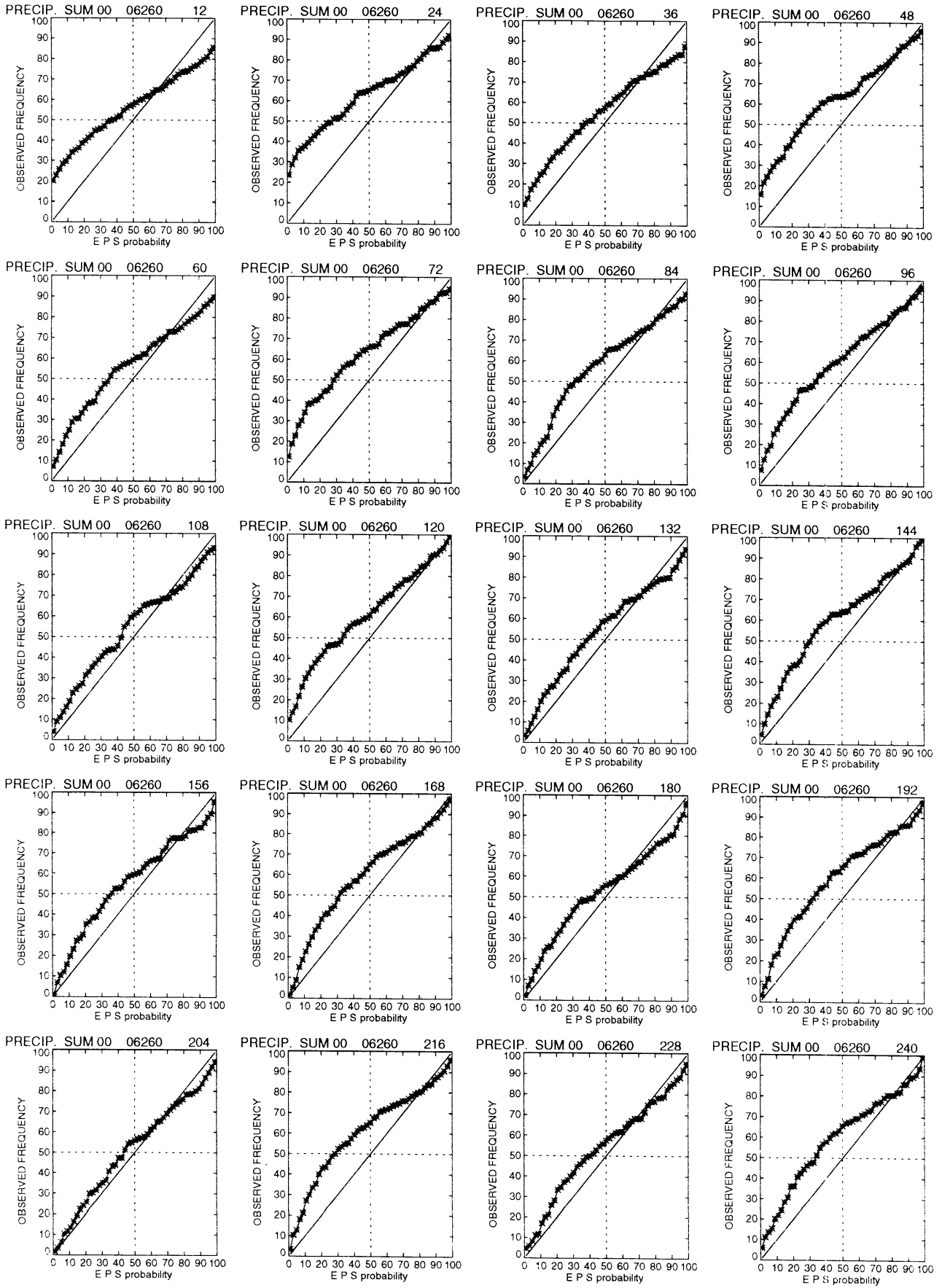


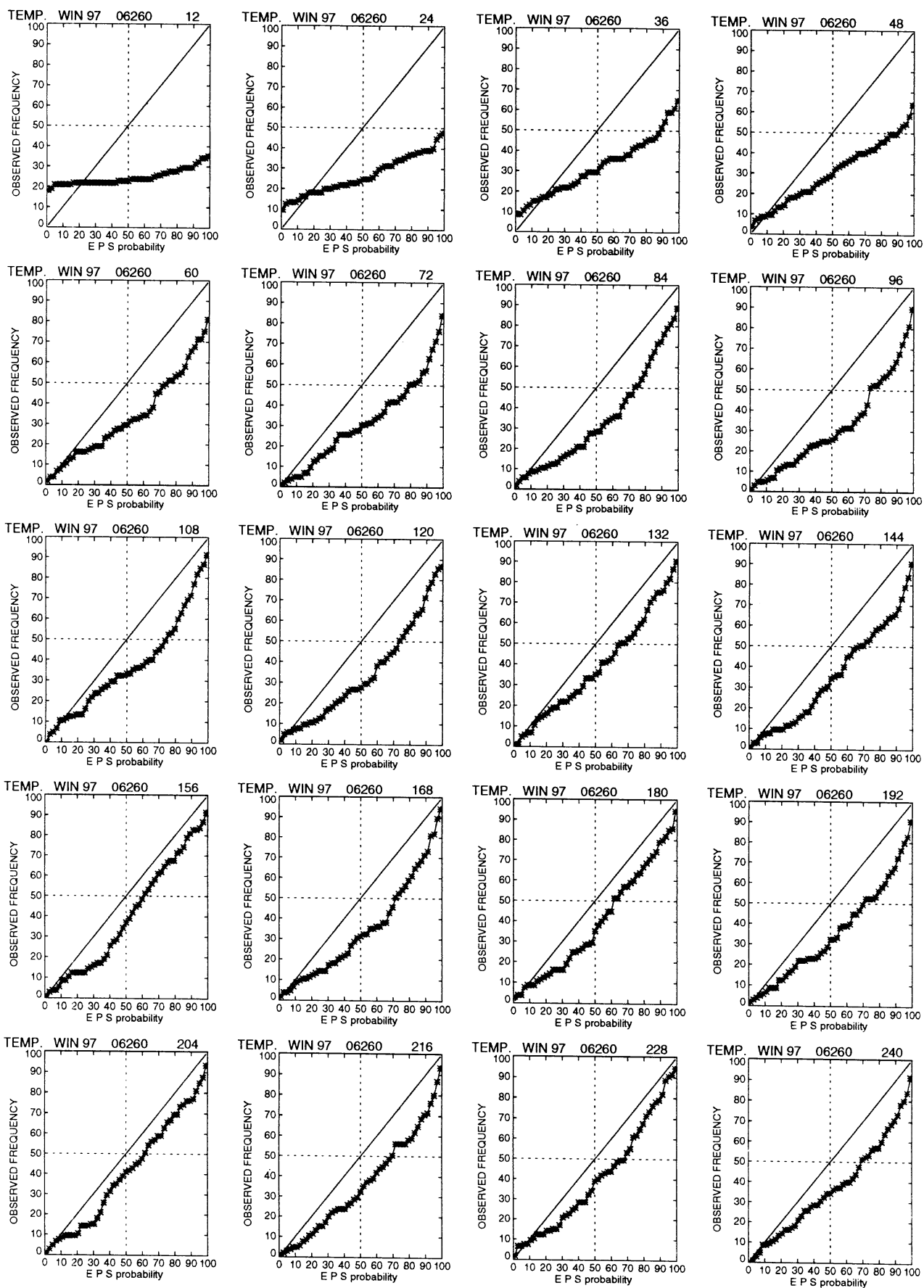


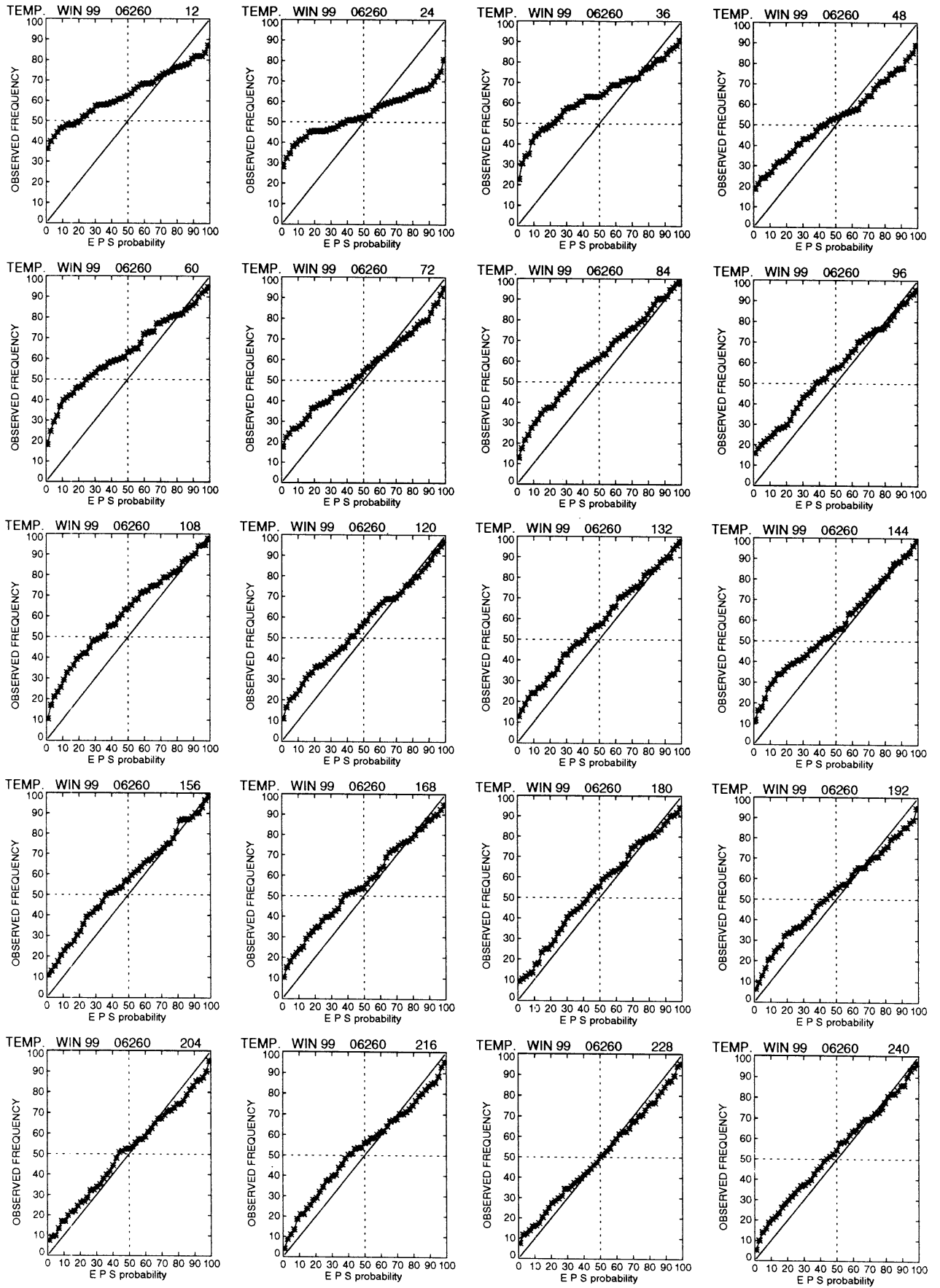


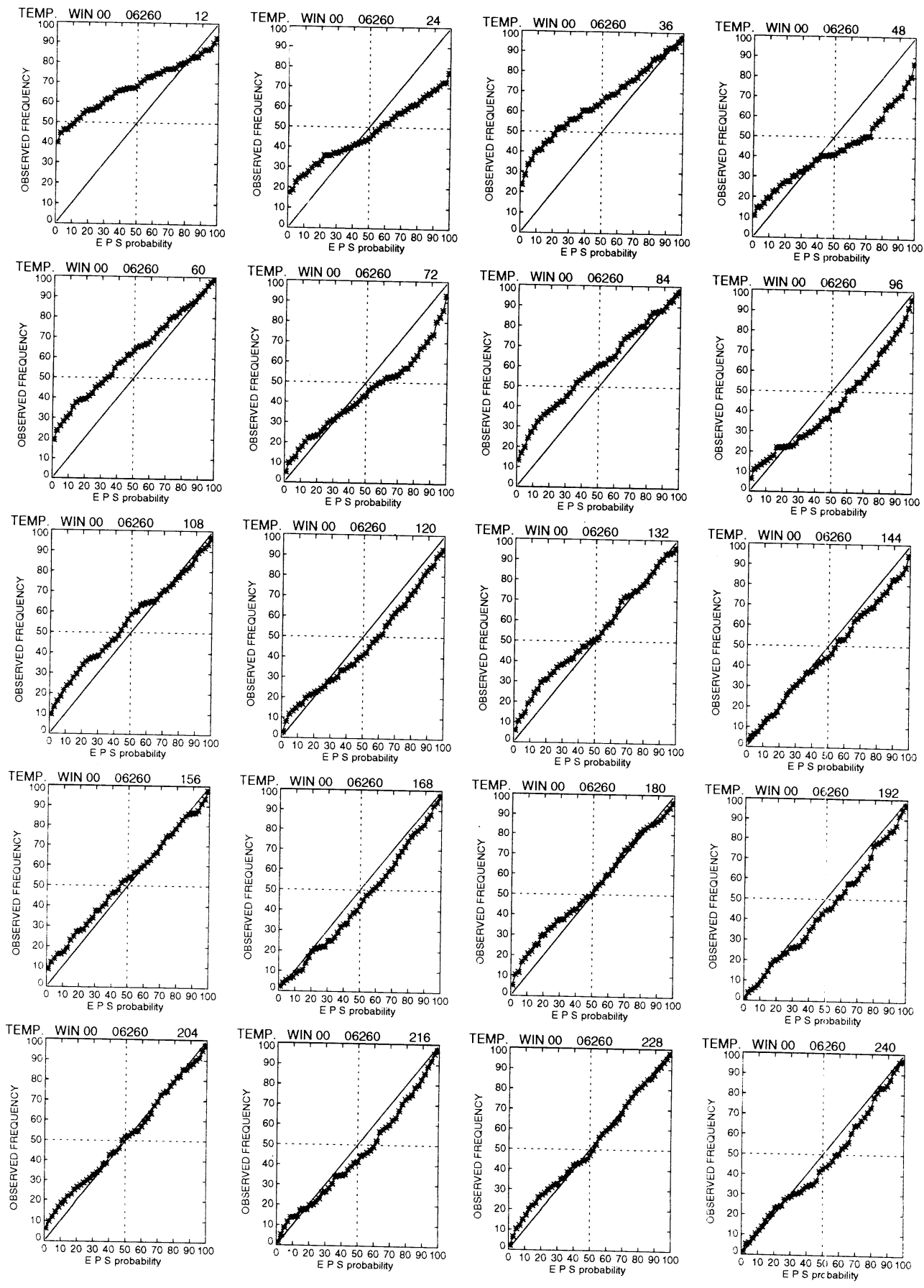


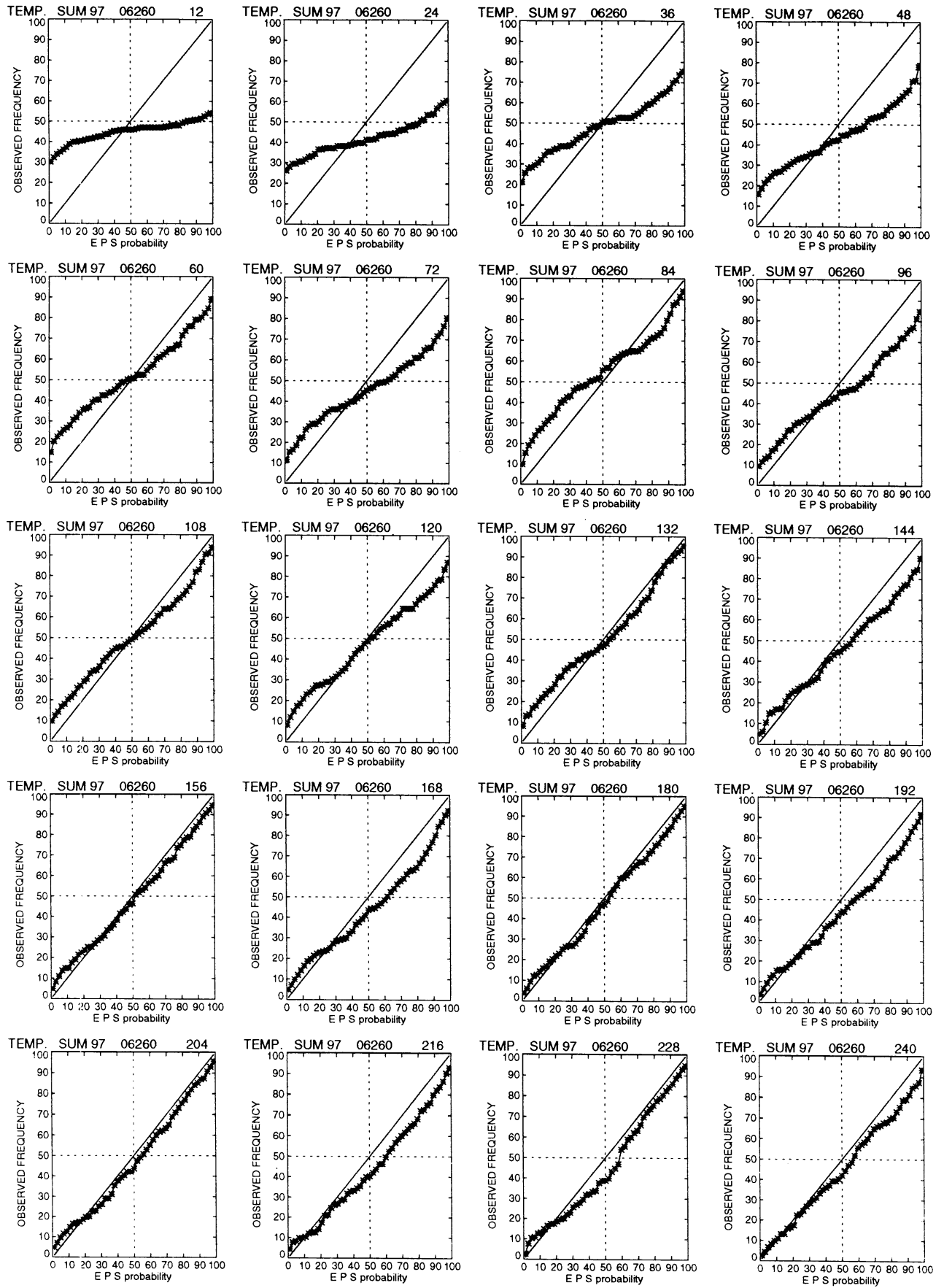


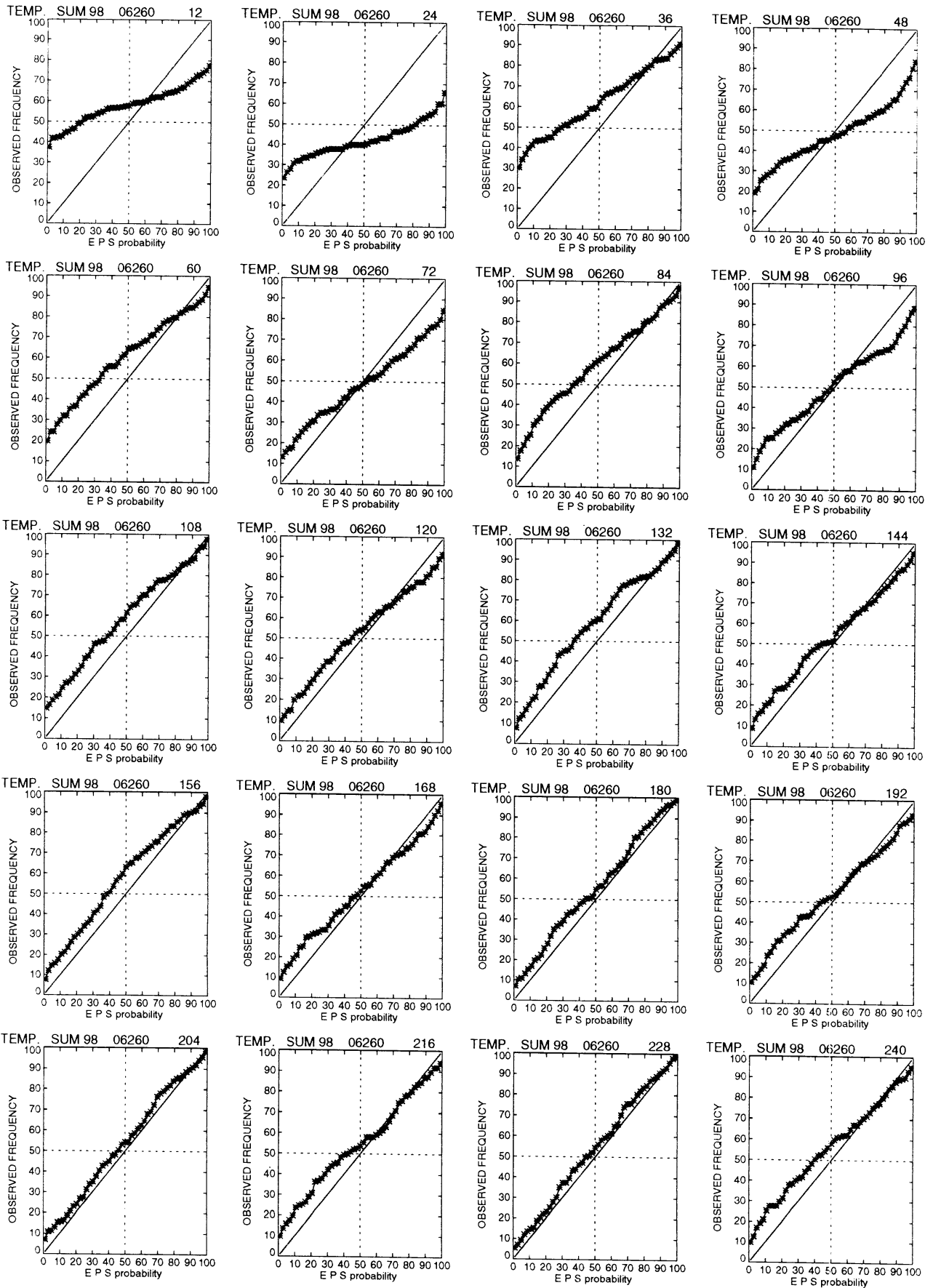


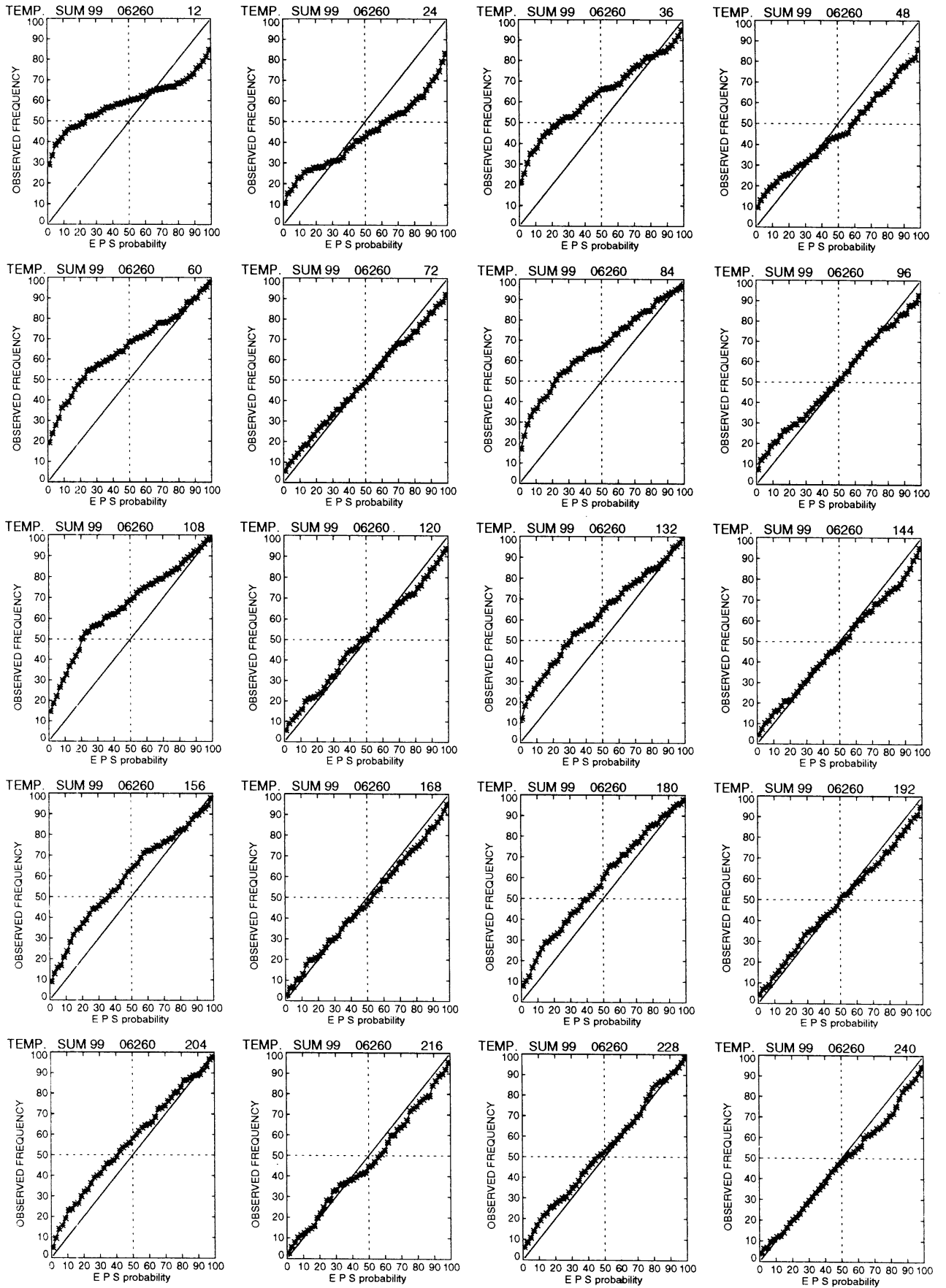












means that exceedance probabilities (i.e. probabilities of precipitation higher than certain thresholds) up to 30% are highly statistically consistent. Since this represents the ensemble members with the higher precipitation amounts this is a very important improvement. However, in our analysis it is not possible to say anything about the statistical consistency of the really extreme precipitation events.

The impact of the introduction of the evolved singular vectors in March 1998 can best be assessed by comparing the C.R.H. curves of the warm seasons of 1997 and 1998: i.e. SUM 97 vs SUM 98. The most conspicuous difference in the first forecast days are the much higher curves in '98 due to a large overestimation of precipitation probabilities. The spread seems to have diminished but this may be an obscuring effect due to this larger bias. The 'median' of the predicted pdf remains too high especially for morning predictions for longer lead times. The difference in skill between morning and afternoon precipitation probabilities is somewhat larger in the summer seasons than it is in the winters. This is true for all four years. This is probably related to the fact that there is a large daily cycle in the summer observed precipitation amounts which is virtually nonexistent in the ensemble mean of EPS. This difference in skill of the atmospheric model in simulating convective and large scale precipitation processes also affects the quality of the EPS probabilities.

The increase in spread in the early forecast ranges is only slight in the last two warm seasons and may, therefore, be attributed to the introduction of stochastic physics. There seem to be no significant changes in the predicted pdfs for day 3 onwards. We believe that the differences in C.R.H.s are simply indicative of the year-to-year variability of the observed and simulated precipitation frequencies and cannot be attributed to any model change. For all summer and winter seasons there is a large asymmetry in correspondence of the pdfs for higher and lower probabilities, the latter being far inferior. It is believed that this is related to the fact that EPS has a large overforecasting of the very small precipitation amounts. For the right part of the distribution, i.e. the part that includes statements about higher precipitation amounts, the EPS forecasts are much more consistent.

In summary, we can say that the impact of the evolved singular vectors on the predicted pdf for precipitation is rather small. This is in agreement with Barkmeijer et. al. (1999). The introduction of stochastic physics had a somewhat larger impact, especially in terms of the spread in the first three forecast days. This increase in spread in the last two cool seasons is therefore believed to be mainly the result of the introduction of stochastic physics. These results are in agreement with the findings of Buizza et. al. (1999), who compared ensembles with and without stochastic physics on a set of 14 cases, and also to Mullen and Buizza (2001) who performed an extensive evaluation of EPS against area observations over the United States over the period from January 1997 until January 1999 inclusive. These latter authors used a stratification into two seasons as well but excluded the transition months April and October.

Now the results for the 2meter *temperatures* at 00 and 12 hour UTC for De Bilt (06260) will be discussed. Despite the fact that the winter 'seasons' of 1996-1997 and 1997-1998 (designated in the figures by WIN 97 and WIN 98) are both before the first major system change which is considered here, there is a large difference in C.R.H. curves for these periods (Fig. 3). In WIN 97 the C.R.H. curves were much too low, meaning that the EPS temperature plumes indicated much too low temperatures: the ensemble mean was more than one degree too low for all lead times. This resulted in an underpopulation of the lower bins and an overpopulation of the higher bins. The difference between the 'median' of the predicted and observed pdfs is around 15 to 20 percent over the entire forecast range. For WIN 98, on the other hand, there is hardly any difference. The predicted distribution is very good from +84.

In the cold seasons of 1998-1999 and 1999-2000 the spread is, as was the case with precipitation, much wider for the first couple of forecast days. WIN 99 shows curves which are somewhat high, i.e. an overestimation of the temperatures, whereas WIN 00 shows an alternating behaviour depending on the time of day, in the sense that the daily cycle is slightly underestimated.

The effect of the introduction of the evolved singular vectors on temperature can be assessed from the differences in the C.R.H. curve between the summers of 1997 and 1998. In contrast to the case of precipitation a slight increase in spread for the first two forecast days can be noted. The spread in the predicted daytime temperatures remains too low over the entire forecast range. The spread for 00 hour is a little wider compared to the observed distribution, but then the lowest end rank is still overpopulated due to a small bias. The stochastic physics has, as we have seen before, an even larger effect on the spread in the beginning of the forecast range. In SUM 99 the positive bias in the night-time temperatures is very apparent. In SUM 00 it still exists but there it is smaller. In this year the large asymmetry between the temperatures at midnight and noon is very conspicuous. Even at the last projection days the population of the left end rank at midnight and the right end rank at noon is around 10 percent. Both 1999 and 2000 summers show an underestimation of the daily cycle in the ensemble mean: midnight temperatures are somewhat too high whereas noon temperatures are too low on average.

The general characteristics of the C.R.H.s show more agreement between the last two summers and between the last two winters than between the summers and winters, especially for precipitation. Therefore we think that stratifying the predictions into 2 “seasons” seems appropriate. Also the fact that the daily cycle in precipitation in summer is almost completely ignored by EPS is a strong reason. Although we have not presented a full analysis over all predictands and stations we will discuss in the next sections only results for stratified data. We will see many more examples of large differences between the two seasons in section 5.

For all values of the EPS probability the differences between 1999 and 2000 are less than 10% , which for many lead times is much smaller than the amount of calibration that is needed to obtain statistical consistency. This is true for both seasons and both predictands. Also the general characteristics of the C.R.H.s for corresponding lead times are very much the same. These are strong arguments to believe that a calibration method may be a definite improvement over directly using EPS (DMO) probabilities. We will pursue this further in section 6.

5. Results over 1999 and 2000

In the previous chapter we have discussed the impact of two major system changes on the statistical consistency of the EPS probabilities. On the basis of those results we have decided that the calibration of EPS probabilities will be based only on data after these changes. So in all of these data the impact of both evolved singular vectors as well as stochastic physics is incorporated. The data set used for further analysis and calibration consists therefore of the period ranging from 21 October 1998 until 1 October 2000. This period is further stratified into 'summer' and 'winter' half years (designated in the figures by 'SUM 99 00' and 'WIN 99 00' respectively). Both these subperiods consist of approximately 360 cases. Station numbers and locations are given in Fig. 1.

In this section the statistical consistency will be briefly reviewed of four meteorological quantities for the 6 land stations as well as for the significant wave height for the three North Sea platforms. Once again this will be done in terms of the cumulative rank histograms (C.R.H.). The plots of all diagrams are given in the Appendix. For station De Bilt (06260) the results for temperature and precipitation in both seasons are a combination of the two years (99 and 00 respectively) we have given already in Fig. 3. Note once again that the EPS probabilities are compared with single station observed frequencies.

In the following discussion the characteristics of the curves for the first couple of forecast days are omitted. This is done because of the common feature that for all predictands and all stations the plumes do not exhibit enough spread.

Temperature

In the C.R.H. curves for the warm seasons ('SUM') there is a large difference between night-time and day-time consistency of the EPS temperature probabilities for all land stations. For the four northernmost stations the plumes were in general too warm at midnight (first and third column) and too cold at noon (second and fourth column), implying an underestimation of the daily cycle. The alternating overpopulation of the lower and higher end ranks in consecutive plots is in agreement with this. But, surprisingly, the differences between predicted and observed pdfs seem to decrease with lead time; the improvement continues until day 10. In the two stations in the southern half of the Netherlands (Vlissingen, 06310, and Beek, 06380) both summer night-time and day-time temperatures are underestimated, at noon somewhat less than at midnight. This is an indication that there may be a slight overestimation of the daily cycle by EPS for these stations. So the over- or underestimation of the daily cycle does not appear to be directly related to the distance to the coast of the observing station.

In winter ('WIN') the agreement between the two pdfs from day 5 or so is quite good for 260, 280 and 310. There is a slight underestimation of the temperatures in 290 and 380, whereas an overestimation occurs at De Kooy. Only this last station exhibits a large difference between day and night. Probabilities of the relatively lower temperatures are highly underestimated. The top right parts of the plots indicate, however, that exceedance probabilities of the order of 10 to 20 percent are statistically consistent for the higher lead times.

Precipitation

For all stations and all forecast times the median of the predicted pdf is too high compared with the observed rain frequency. This is true for both seasons. As we have seen already this is to a large part due to the fact that the probability of small precipitation amounts is highly exaggerated by EPS. The form of the curves does not change much over the forecast times. In the summers there appears to be a significant difference between 'morning' (columns 2 and 4) and 'afternoon' (columns 1 and 3) results, the bias for morning probabilities being larger. This is more pronounced for the typical land stations than for the two stations near to the coast. For some stations the afternoon C.R.H.s develop a sigmoid tendency. In the colder season the daily cycle in C.R.H. is almost completely absent.

A very important feature in the C.R.H. curves for all stations in both seasons is the fact that small exceedance probabilities, say in the order of 10 to 20 percent, appear to be highly statistically consistent after day 3. This is more so in winter than in summer. This might indicate that predictions of the larger precipitation amounts are likely to be statistically consistent.

Wind speed

In summer wind speed predictions for midnight exhibit a large positive bias for the four northernmost stations but a negative one for the two stations in the southern part of the country. For the predictions valid for 12UTC the overestimation is always less than at midnight; in the case of Eelde (06280) and Twente (06290) it is even around zero. Apparently, EPS winds seem to miss the proper diurnal cycle, probably because the small scale vertical structure is not well represented. There is no apparent reason for the difference between stations in the northern and southern part.

The situation is more or less the same for most stations in the last two winters. Here, the positive bias is also restricted to the four stations in the north and central part of the country. Again the remaining two stations show a negative bias. The difference in bias for midnight and noon is also present but the difference is generally smaller than in summer.

A feature in the C.R.H. curves that is not present for the previously discussed meteorological quantities, is the fact that the consistency is in general improving with lead time until day 10. This is true for both observation times and for both seasons. This better correspondence between the pdfs does not (necessarily) mean that the skill of the EPS probabilities is improving. Again, the higher part of the predicted pdf is much more consistent than the lower part.

Cloud cover

When verified against station observations EPS underestimates cloud amounts for all stations and all forecast times in both seasons. The only possible exception may be De Kooy in winter. In summer the mismatch between the predicted and observed pdf is much larger at 12UTC ('medians' differ 10 to 25%) than it is at 00UTC (5 to 15%). In winter the 'medians' of the two pdfs are much closer to each other.

On a number of occasions, especially in winter at lead times longer than day 7, the cloud cover curves exhibit indications that there is a relative abundance of the ranks near the

center. The reasons for this are not known. Note, however, that the curves may be influenced in a systematic way by the fact that the observations are available in octas.

Significant wave height

We have restricted the analysis for significant wave height to the winter seasons (April to October). Only three stations are available: K13 (06252) on more or less open sea, Europlatform (06321) and Meetpost Noordwijk (06254) close to the Dutch coast. See Fig. 1 for their location. The resulting C.R.H. curves are given in the Appendix and are labeled on top of each plot by Hs.

The C.R.H. curves for platform K13 (06252) exhibit of course a much too small spread for the earlier forecast days. Until day 5 there is an asymmetry in the sense that the lowest end rank is strongly overpopulated whereas the right end rank is more or less what should be expected. In other words, the percentage of outliers is much larger for lower wave heights than for higher wave heights. This is due to a more general overforecasting which seems to be slowly decreasing with projection time until it is almost absent after seven days. For longer lead times the match between predicted and observed pdf is almost perfect. Moreover, exceedance probabilities of the order of 10 to 20 percent (top right in the C.R.H.s) are highly consistent from day 3 on already.

The two stations near the coast are very similar to one another. They both exhibit a large overforecasting. The difference in 'median' between the predicted and observed pdf is only slowly decreasing with projection from around 30% at days 1 and 2 to around 15% at the end of the forecast range for Meetpost Noordwijk to slightly less than that for Europlatform. This is presumably due to the closer proximity to the coast of the former station. It also results in an asymmetry in the population of the end ranks. The apparant deficiencies of the rather coarse grid wave model for the predictions of waves for locations very close to the coast are more elaborately discussed in Vogelesang and Kok (1999).

The spread of the wave height plumes is not wide enough over the entire forecast range for Meetpost Noordwijk and at least until day 8 for Europlatform. Especially the lower wave heights (the lower part of the predicted pdf) are poorly represented. Again, for all three stations the statistical consistency generally improves until day 10.

6. Calibration

In the previous chapter we have discussed the (more or less) systematic discrepancies between the EPS probabilities and the empirical frequencies of the events for individual stations. These differences will now be used to make the EPS probabilities statistically consistent. This will be done for all predictands and stations on the basis of the data presented in the previous section. So, only the years 1999 and 2000 are used and the stratification into two six month periods ('summers' and 'winters') is used.

The matching procedure is very straightforward. The cumulative rank histograms given in the Appendix allow us to make an almost continuous translation between predicted and observed pdfs. More specifically, in steps of 1.96% (i.e. 1/51) EPS probabilities for predictand values below a certain value can be matched, on the basis of historical data, to the observed frequency of that particular predictand and station. A lack in correspondence, caused by a "bias" or a too wide or narrow distribution (or both), can thus be accounted for.

The calibration can now be applied to individual forecasts. A convenient way of presenting the calibrated EPS probabilities is in terms of lines of fixed (prechosen) percentiles. A few examples are given in Fig. 4. In Fig. 4a an example is given for the ECMWF temperature forecast of the 28th of August 2000 for station De Bilt. Here deterministic and probabilistic information is merged. Since the temperatures are expressed in absolute numbers and not for instance with respect to climatology, we have separated the results for 00UTC (top panel) from those for 12UTC (bottom panel). The forecasts of all 51 members are given in thin lines, whereas the 3 thick solid lines represent (from bottom to top) the 25, 50 and 75% percentiles. The 5 and 95% percentiles (if present) are given in thick dashed lines. These percentiles were calculated on the basis of the summer calibration for this station (see the Appendix for the corresponding C.R.H.). As can be seen the percentiles are gradually dispersing with increasing forecast time. The temperature range between the 25 and 75% percentile, which can be interpreted as a 50% *confidence interval*, increases in this example from about 1.5 degrees at +36 to about 3.5 degrees at +228. The 5% percentile in the top panel and the 95% percentile in the bottom panel are absent. The former is due to the fact that averaged over all forecasts for midnight temperature, more than 5 percent of the observations were below the lowest ensemble member. Likewise, the underforecasting of midday temperatures (see also chapter 5) resulted in a more than 5% population of the highest bin in the corresponding Talagrand diagram. For this reason the respective percentiles cannot be calculated. However, this does not mean that we have no additional probabilistic information for the tails of the distribution. For instance, although the exact value of the 5% percentile in the top panel of Fig. 4a is not known, it has to be located below the lowest ensemble member. Similarly, in the bottom panel of Fig. 4a the 95% percentile lies (somewhere) above the highest ensemble member. Note that for the +12 forecast even the 25% percentile cannot be calculated.

A second example is given in Fig. 4b. Here the EPS wave forecast of 21 December 1999 is shown for two stations on the North Sea: one close to the Dutch coast (Meetpost Noordwijk) and one on more or less open sea (K13). This situation was chosen because of the high EPS probability of significant waves around the 25th of December.

For K13 (top panel) the match between the predicted pdf and the observed frequency distribution, as indicated by the C.R.H.s shown in the Appendix, was very good for the relatively high waves, even at day 2 already, and remained so for the rest of the forecast range. This results in the fact that the 95% percentile (top dashed line) can be given already at +48 and, moreover, that there are exactly two ensemble members located above the 95%

percentile. For the lower wave heights the statistical consistency is much worse, but is slowly improving out to day 10. This culminates in the 5% percentile from day 8. Here only one ensemble member is located below the line.

The situation is quite different for MPN (lower panel). This is of course due to the large overestimation of the wave heights in the vicinity of the coast. This results for instance, in about 4 ensemble members above the 95% percentile (instead of 2 when statistically consistent). The inconsistency of the EPS probabilities is even more conspicuous for lower wave heights. About 70 percent of the ensemble members lie above the median (the second thick line from above). The median is only present from day 2, the 25% percentile from +132, whereas the 5% line is missing altogether. After the 25th of December the spread is decreasing; the 50 percent confidence interval is more or less constant until day 10. This is true for both stations.

As mentioned above, in many cases the extreme percentiles cannot be calculated. For some predictands it is possible to overcome this situation. For instance for precipitation it has been suggested by Hamill and Colucci (1998) to linearly interpolate between zero and the value of the lowest ensemble member. A similar approach might be tested for wind speed and cloudiness. This has not been done in the present study.

The calibrated EPS probabilities have not yet been verified on independent data using e.g. the Brier score or reliability diagram.

The presentation of the calibrated probabilities as given in this section is of course only one of the many possible forms. For predictands which exhibit a large daily cycle it may be more appropriate to express the forecasts as anomalies with respect to climatology. Also, a presentation in terms of stacked bars, as used by Kok and Voegelzang (1999), may be an appealing alternative. Finally, box-and-whiskers plots are frequently used to summarize the predicted distribution. The choice of percentiles is of course arbitrary and may be adapted to the wishes of the user. The degree in which the pdfs differ may yield restrictions to what is still possible in terms of conveying the probabilistic information. In the examples in Fig. 4 a choice of 10% and 90% (instead of 5 and 95%) would possibly be more appropriate.

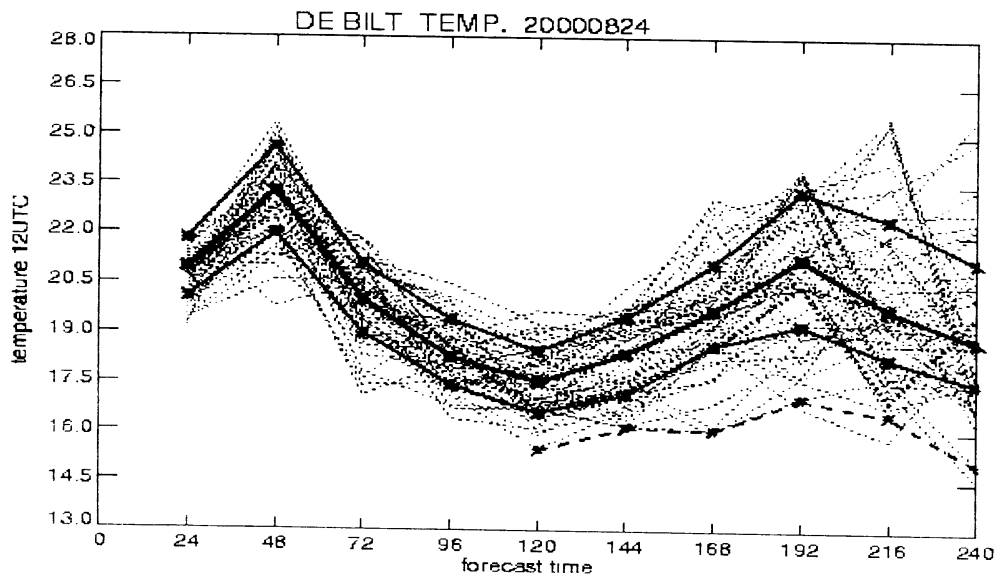
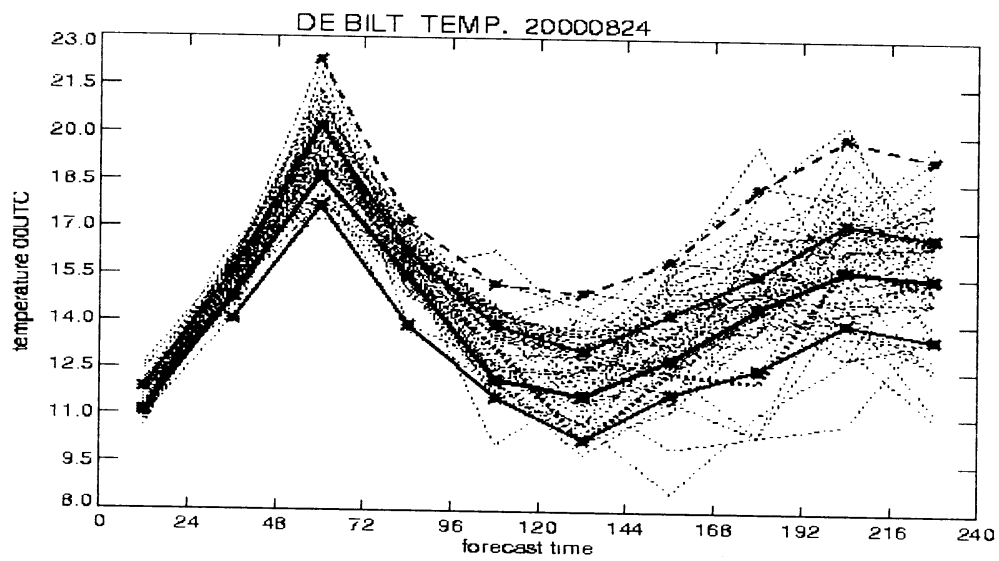


Fig. 4a. Example of an EPS forecast for 00 UTC (top) and 12 UTC temperatures (bottom) for De Bilt together with the 25, 50 and 75% percentiles (thick lines) as well as (if present) the 5 and 95% percentiles (dashed lines).

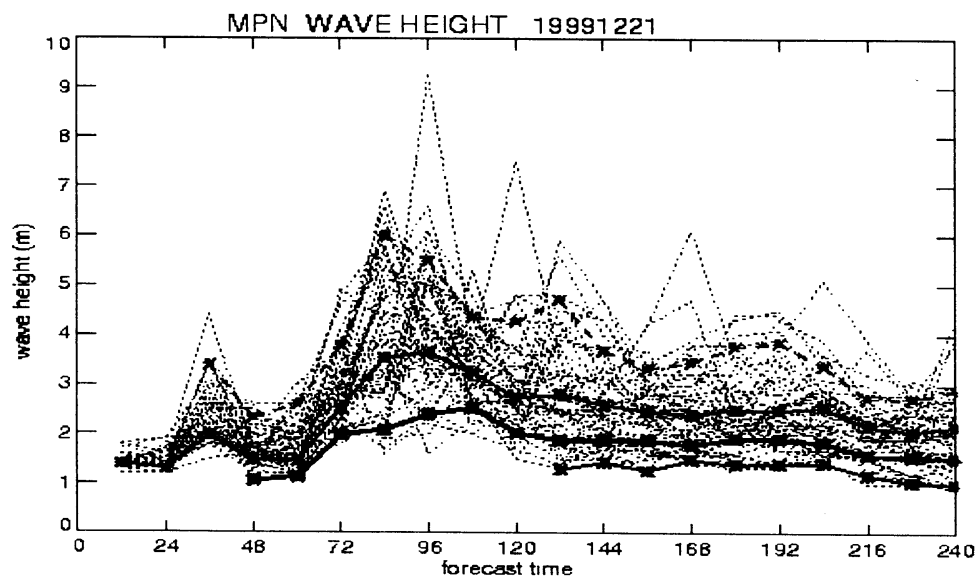
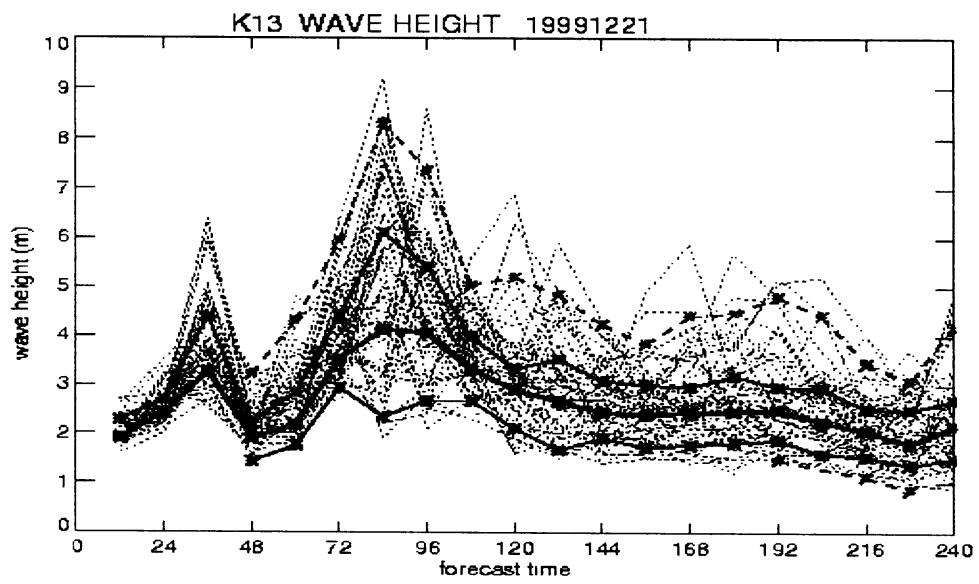


Fig 4b. Example of a wave-EPs forecast for K13 (top) and Meetpost Noordwijk (bottom) together with the percentiles as given in Fig. 4a.

7. Concluding remarks

In operational practice EPS probabilities derived for instance from the so-called plume plots are very frequently used to assess for a number of stations the probability for particular events to occur. To estimate and quantify the validity of this we have compared the EPS probabilities for a number of predictands to station observed frequencies. The main objective of this paper was to match, or calibrate, the predicted probabilities to the empirical observed frequencies.

To define the calibration data set first the impact of two recent major system changes had to be studied. These were the introduction of evolved singular vectors in March 1998 and of stochastic physics in October of that same year. This led to the conclusion that only data later than October 1998 should be used. The remaining two years of data, stratified into warm and cold 6 months periods, proved to be consistent enough in time to make it possible to develop a tentative system to calibrate individual forecasts. This has been done for all lead times for all predictands and stations which are currently used at KNMI on an operational basis.

The evaluation was done using essentially one tool. This is the Talagrand Diagram or rank histogram. But we have introduced a new way of presenting the information contained in this diagram; the ranks of the Talagrand diagram are summed from left to right (or low to high). This cumulative rank histogram (C.R.H.) enables us to quickly assess the main characteristics of the differences between the predicted and observed pdfs. Especially for the purpose of calibration it proved to be very helpful.

Since we have not matched the model forecasts to area averaged observations we cannot conclude whether or not the perfect model assumptions are met. No results are available to indicate whether the insufficient variability was due to model errors, the selection of initial perturbations, or both. The excessive population of the end ranks which sustained over the entire forecast range for many of the predictands is also present in a study by Mullen and Buizza who evaluated EPS precipitation probabilities against area observations over the United States (Mullen and Buizza, 2001). This was attributed to a model error "whose cause is not yet known". This may play a role in our analysis also.

However, in interpreting the C.R.H.s as we have presented in this report, one should be aware that they only yield a general evaluation of the combination of all probability statements. Therefore no definite conclusions can be made about the statistical consistency of probability statements for the exceedance of fixed thresholds (e.g. for the probability of precipitation amounts of more than 20 mm). The same is true for the skill of the system with respect to outliers. However, in this context it may be a good sign that for a number of predictands which have a clear asymmetric distribution, due, for instance, to a fixed lower limit (e.g. wind speed, precipitation amount) the statistical consistency proved to be better in general for the higher extremes than for the lower.

A conspicuous feature is the fact that for some predictands there seems to exist a growing agreement between the pdfs with increasing forecast time, not only for the first couple of forecast days but all the way to day 10. This is especially true for wave heights and cloudiness where the highest consistency is obtained for days 9 and 10. This may be somewhat surprising considering the fact that in general the skill of the forecasts deteriorates with increasing lead time.

The calibration results can be regarded as a first order improvement with respect to the EPS probabilities. Although we have not verified the calibrated probabilities on independent data it can be anticipated that the skill may be enhanced to some extent by resolving the major systematic discrepancies in "bias" and in width of the distribution. A further improvement may

be obtained by separating the ensemble into different categories depending on the spread of the ensemble. This was done by Hamill and Colucci (1997, 1998) who constructed three rank histograms based on high, medium and low ensemble variability. Eckel and Walters (1998) even used 16 different variability categories. They found that there were indeed significant differences in calibration. These authors have pooled the stations to arrive at combined rank histograms whereas we treated all stations separately. Also we stratified our data into 'summer' and 'winter' seasons. Nevertheless we believe that a more specific calibration than the one we used is possible. This will be explored in a future study.

For predictive purposes EPS probabilities can be much further improved when more sophisticated post-processing techniques are used. For instance, Kok and Voogelezang (1999) performed an experiment in which probabilistic temperature forecasts from three different sources were intercompared. It was shown that a statistical guidance (multiple linear regression) model using predictors obtained from the operational deterministic model only outscored the EPS probabilities by far. Including also EPS predictors yielded only a slight additional improvement. However, the ECMWF ensemble system has improved considerably since then.

In the context of calibration the presented C.R.H.s can only be used for individual stations and not for areas or regions surrounding the stations. To do so new C.R.H.s have to be derived. Depending on the correlation radius of the particular predictand the properties may or may not change considerably. This may be one of the reasons of the stronger inconsistency of precipitation forecasts compared to those of for instance temperature.

As mentioned above, the calibration has been applied to station data only. However, the method can (and should) be used for area averaged observations as well. Moreover, the method can also easily be applied to predictands that are highly correlated to the EPS predictand at hand. For instance, 12UTC plume information can easily be matched to observed local maximum temperatures, and 00UTC to minimum temperatures.

Acknowledgement

Martin Stam and Ilja Smits are greatly acknowledged for their help in the construction of figure 1. Daan Voogelezang and Jan Fokke Meirink are thanked for their help in some of the details of pvwave. Finally, the careful reading of this report by Seijo Kruizinga, his many suggestions and the fruitful discussions are greatly appreciated.

References

- Anderson, J. L. (1996). A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate*, **9**, 1518-1530.
- Barkmeijer, J., R. Buizza and T. N. Palmer (1999). 3D-Var Hessian singular vectors and their potential use in the ECMWF Ensemble Prediction System. *Quart. J. Roy. Meteor. Soc.*, **125**, 2333-2351.
- Brier, G. W. (1950). Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.*, **78**, 1-3.
- Buizza, R., M. Miller and T. N. Palmer (1999). Stochastic simulation of model uncertainties. *Quart. J. Roy. Meteor. Soc.*, **125**, 2887-2908.
- Eckel, F. A. and M. K. Walters (1998). Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. and Forecasting*, **13**, 1132-1147.
- Hamill, T. M. and S. J. Colucci (1997). Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.*, **125**, 1312-1327.
- Hamill, T. M. and S. J. Colucci (1998). Evaluation of Eta-RSM Ensemble Probabilistic Precipitation Forecasts. *Mon. Wea. Rev.*, **126**, 711-724.
- Kok, C. J. and D. H. P. Vogelesang (1999). Statistical post-processing on EPS. *ECMWF Expert meeting on Ensemble Prediction System*. 17-18 June 1999.
- Kruizinga, S. (1999). Verification of EPS output — characteristics of some verification scores. *Proceedings of the 7th workshop on Meteorological Operational Systems*. ECMWF 15-19 Nov. 1999.
- Krzysztofowicz, R. and A. A. Sigrest (1999a). Calibration of probabilistic quantitative precipitation forecasts. *Wea. and Forecasting*, **14**, 427-442.
- Krzysztofowicz, R. and A. A. Sigrest (1999b). Comparative verification of guidance and local quantitative precipitation forecasts: calibration analyses. *Wea. and Forecasting*, **14**, 443-454.
- Mullen, S. L. and R. Buizza (2001). Quantitative precipitation forecasts over the United States by the ECMWF Ensemble Prediction System. *Mon. Wea. Rev.*, **129**, 638-663.
- Murphy, A. H. (1973). A new vector partition of the probability score. *J. Appl. Meteorol.*, **12**, 595-600.
- Press, W. H., S. A. Teukolsky, W. T. Vetterling and B. P. Flannery (1994). *Numerical Recipes in Fortran*. 2nd Ed. Cambridge University Press, 963pp.
- Richardson, D. (2000). Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.*, **126**, 649-668.
- Talagrand, O., R. Vautard and B. Strauss (1997). Evaluation of probabilistic prediction systems. *Proceedings of the ECMWF Workshop on predictability*, Reading, England 20-22 October 1997, 1-25.
- Vogelesang, D. H. P. and C. J. Kok (1999). Golfhoogteverwachtingen voor de Zuidelijke Noordzee. KNMI Technical Report, TR-223.
- Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences*. Academic Press. 464pp.

Appendix.

In this appendix the C.R.H.s are given for all stations, predictands and forecast ranges, calculated over approximately two years: from 21 October 1998 until 1 October 2000. This period is indicated on top of each graph by '99 00' and is the same for all plots. The period is stratified into a summer and a winter 'season' of 6 months. This is indicated by 'SUM' and 'WIN' respectively. The forecast time is given on the top right of each plot. The locations of the stations is given in Fig.1 and the predictands in Table 1. For more details the reader is referred to the text.

The stations are indicated by their WMO identification:

De Kooy	06235
De Bilt	06260
Eelde	06280
Twente	06290
Vlissingen	06310
Beek	06380
K13	06252
Meetpost Noordwijk (MPN)	06254
Europlatform (EURO)	06321

First the results for the land stations are given in the following order:

- temperature (TEMP.) on pages A.1 – A.12
- precipitation (PRECIP.) A.13 – A.24
- wind speed (FF) A.25 – A.36
- cloudiness (CLOUD.) A.37 – A.48

For each predictand first the summer (SUM) results are shown for all stations followed by those for winter (WIN).

Finally, for the three sea stations the results for significant wave height (Hs) in the winter is given (on pages A.49 – A.51).

