Increasing the Reliability of Reliability Diagrams

Jochen Bröcker[1,*] and Leonard A. Smith[1,2]

[1]Centre for the Analysis of Time Series
London School of Economics
London, UK

[2]Pembroke College
Oxford University
Oxford, UK

[*]Corresponding Author, cats@lse.ac.uk

November 9, 2006

Abstract

The reliability diagram is a common diagnostic graph used to summarise and evaluate probabilistic forecasts. Its strengths lie in the ease with which it is produced and the transparency of its definition. While visually appealing, major long noted shortcomings lie in the difficulty of interpreting the graph visually; for the most part, ambiguities arise from variation in the distribution of forecast probabilities and from various binning procedures (Murphy and Winkler, 1977; Smith, 1997). A resampling method for assigning *consistency bars* to the observed frequencies is introduced which allows immediate visual evaluation as to just how likely the observed relative frequencies are under the assumption that the predicted probabilities are reliable. Further, an alternative presentation of the same information on probability paper easies quantitative evaluation and comparison. Both presentations can easily be employed for any method of binning. Code to implement this approach is available at `www.lsecats.org`.

1 Introduction

Reliability diagrams are common aids for illustrating the properties of probabilistic forecast systems. They consist of a plot of observed relative frequency against predicted probability, providing a quick visual inter-comparison when tuning probabilistic forecast systems, as well as documenting the performance of the final product, see for example Murphy and Winkler (1977, 1987); Atger (2004, 2003); Jolliffe and Stephenson (2003); Wilks (1995). Yet the visual impression of the reliability diagram can be misleading. Even a perfectly reliable forecast system is not expected to have an exactly diagonal reliability diagram due to limited counting statistics (Jolliffe and Stephenson, 2003). To evaluate a forecast system requires some idea as to how far the observed relative frequencies of that forecast system are expected to be from the diagonal if it *was* reliable. This paper provides two methods to visualise this expected deviation from the diagonal, thereby allowing the forecaster to see directly whether the observed relative frequencies fall within the variations to be expected even from a perfectly reliable forecast system.

In the first section, we revisit how reliability diagrams are computed, and explain in detail why limited counting statistics cause even perfectly reliable forecast systems to exhibit deviations from the diagonal. The next two subsections present two alternative approaches towards visualizing this information: the first is a revised set of consistency bars (Smith, 1997) computed through a consistency resampling technique; the second is to re-plot the same information in reliability diagrams on probability paper, providing a rather blunt presentation of the quality of the forecast system. Both methods aim to increase the reliability of interpretations of reliability diagrams. We demonstrate the benefit of both approaches with synthetic datasets and show an application to London Heathrow temperature anomaly forecasts.

2 How to make a reliability diagram

This section explains briefly how reliability diagrams are computed (for an excellent explanation and connections to various other statistics see Wilks (1995)). The main aim is to introduce the necessary terms and notation in order to facilitate the later discussion on shortcomings of a simple reliability diagram.

The reliability diagram is a diagnostic for probabilistic forecasts. In this paper, we will describe the occurrence or non–occurrence of the event under concern by a variable Y, which is equal to one (event does happen) or to zero (event does not happen). The variable Y is called the *verification*. Let

$Y_i, i = 1 \ldots N$ be a data set of verifications. For each i we also have a *forecast value* X_i, a number between zero and one, representing a forecast probability that the corresponding Y_i will be equal to one. The forecast value X_i need *not* assumed to be a probability in a frequentist's sense (see Wilks (1995), pp.9 for a discussion).

Reliability diagrams provide a diagnostic to check whether the forecast value X_i is *reliable*. Roughly speaking, a probability forecast is reliable[1], if the event actually happens with an observed relative frequency consistent with the forecast value. More specifically, considering only instances i for which $X_i = x$ for a certain value x, the event happens with an observed relative frequency equal to x. This definition implicitly assumes that the forecast values X_i can assume only a finite number of values, for example $[0, 0.1, 0.2 \ldots 1]$, and there is an obvious problem with this definition when X_i can assume any value between zero and one. In that case, the event $\{X_i = x\}$ is unlikely to happen more than once for any x, rendering the computation of observed relative frequencies impossible. In order to be able to compute any nontrivial observed relative frequencies, the forecast values X_i are collected into a number of representative bins. The above definitions are slightly altered thus: A forecast is reliable if the relative frequency of the event $Y_i = 1$, when computed over all i for which X_i falls into a small interval B, must be equal to the mean of X_i over that interval.

Reliability diagrams reveal reliability by plotting the observed relative frequencies versus the forecast values. If the bins are small, then in the limit of infinitely many forecast values these observed relative frequencies would fall along the diagonal for a reliable forecast. The remainder of the present subsection explains how basic reliability diagrams are computed.

First, the forecast values are partitioned into *bins* $B_k, k = 1 \ldots K$ (which form a partition of the unit interval into non overlapping exhaustive subintervals). The B_k are often taken to be of equal width, but if the distribution of the forecast values is non–uniform, then choosing the bins so that they are equally populated is an attractive alternative.

Next, for each i, it is established which of the K bins the forecast value X_i falls into. For each bin B_k, let I_k be the collection of all indices i's for which X_i falls into bin B_k, that is

$$I_k := \{i; X_i \in B_k\}. \tag{1}$$

The corresponding *observed relative frequency* f_k is the number of times the event happens, given that $X_i \in B_k$, divided by the total number of forecast

[1]Sometimes the expression "calibrated" is used instead of "reliable"

values $X_i \in B_k$. This can be expressed as

$$f_k = \frac{\sum_{i \in I_k} Y_i}{\# I_k}, \tag{2}$$

where $\# I_k$ denotes the number of elements in I_k.

Each bin B_k is represented by a single "typical" forecast probability r_k. Although the arithmetic centre of the bin is often used to represent the forecast values in that bin, this method has a clear disadvantage: If the forecast is reliable, the observed relative frequency for a given bin B_k is expected to coincide with the average of the forecast values over that bin B_k, rather than with the arithmetic centre of the bin. Plotting the observed relative frequency over the arithmetic centre can cause even a perfect reliability diagram to be off the diagonal by up to half the width of a bin. In this paper, observed relative frequencies for a bin B_k are plotted versus the average of the forecast values over bin B_k. This average, denoted by r_k, is:

$$r_k := \frac{\sum_{i \in I_k} X_i}{\# I_k}, \tag{3}$$

The reliability diagram comprises a plot of f_k versus r_k for all bins B_k.

2.1 Reliable Reliability Diagrams

The observed relative frequencies f_k for a given bin B_k fluctuate for several reasons. First, if we fix the forecast values falling into bin B_k, then under the hypothesis of reliability, the observed frequencies follow a binomial distribution with parameters I_k and r_k (i.e. the number of forecast values falling into bin B_k and the average of the forecast values over bin B_k respectively, see Equations 1,3). Second, these two parameters fluctuate as well, with I_k being of larger impact than r_k, especially in bins already containing relatively few samples.

Several approaches to visualise these effects quantitatively have been suggested. Commonly, a small viewgraph is plotted overlaying the reliability diagram, showing the distribution of the forecast values X_i. This kind of plot is also known as the calibration diagram. Although this pair of plots conveys all relevant information, it is difficult to mentally integrate the two graphs to estimate possible variations of the observed relative frequencies; no direct quantitative consistency check is available. In the reliability diagrams of the ECMWF (e.g. (Hagedorn et al., 2004)), the size of the symbol is often used to reflect I_k, the population of the bin. This is similar to the approach taken by Murphy and Winkler (1977), where I_k is plotted next to the symbol.

Although the information is visually displayed in these approaches, neither provides a measure of quantitative agreement with the hypothesis of reliability. In Smith (1997), the expected fluctuations in the observed frequency f_k for each bin are computed using the binomial density, but the r_k (see Equation 3) as well as the bin population I_k are assumed to be fixed. This is obviously an idealisation, especially if the number of forecast values falling into bin B_k is small or if they are not uniformly distributed.

Our approach, which can be seen as an extension to Smith (1997), is simply to compute the variations of the observed relative frequencies over a set of reliable forecasts generated by resampling technique referred to as *consistency resampling*. It computes the fluctuations of the observed relative frequencies f_k taking into account uncertainties arising due to varying bin means r_k as well as bin populations I_k. Let $(X_i, Y_i), i = 1 \ldots N$ be the data set consisting of forecast–verification pairs. A single resampling cycle consists of the following steps: We draw N times[2] with replacement from the set $X_i, i = 1 \ldots N$, obtaining a set of surrogate forecasts $\hat{X}_i, i = 1 \ldots N$. We then create surrogate observations $\hat{Y}_i, i = 1 \ldots N$ by means of

$$\hat{Y}_i = 1 \text{ if } Z_i < \hat{X}_i \text{ and } 0 \text{ else,}$$

where Z_i is a series of independent uniformly distributed random variables. Note that $\hat{X}_i$ is by construction a *reliable forecast* for $\hat{Y}_i$ (see Appendix I). A reliability diagram is computed using the surrogate data set comprising $\hat{X}$ and $\hat{Y}$. The resulting vector of surrogate observed relative frequencies is recorded. This completes a single resampling step.

The resampling step is repeated N_{boot} times, yielding N_{boot} surrogate observed relative frequencies $\hat{f}_B$ for each bin B_k. We plot the range of the surrogates for each bin as a vertical bar over r_k (the average of the forecast values over bin B_k) in the reliability diagram (See Figure 1). The bars extend from the 5% to the 95% quantiles, indicated by dashes. Henceforth, these bars will be referred to as *consistency bars*. Consistency bars, along with the observed relative frequencies of the original data set, allow an immediate visual interpretation of the quality of the probabilistic forecast system. The extent to which the system is calibrated is reflected by where the observed relative frequencies fall within the consistency bars, not their "distance from the diagonal". In a bin with a large number of forecast values, the observed relative frequency may be quite close to the diagonal in terms of linear distance, but quite far in terms of probability. In this case, the consistency bars reflect the expected distances, and will clearly indicate the failure of the forecast system. The benefit of the method is illustrated by comparing

[2] Recall that N is the total length of the data set

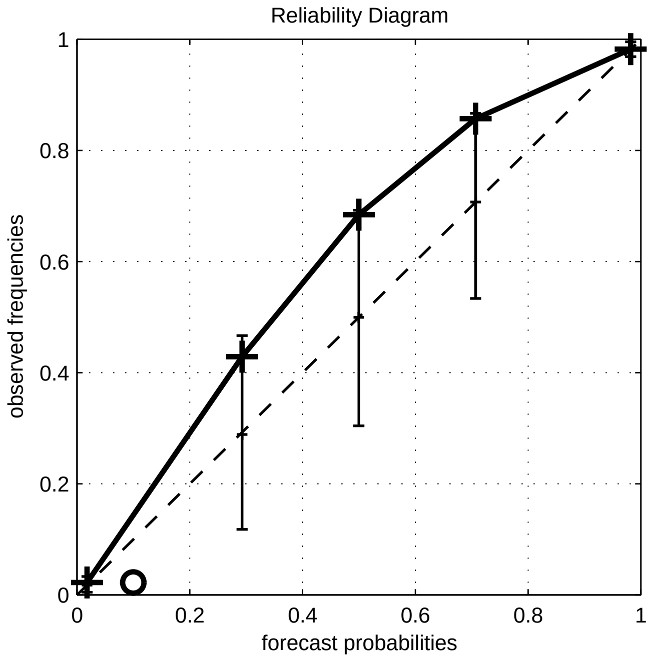


Figure 1: Reliability diagram for data set I using consistency bars. The observed relative frequencies fall all within the 5%-95% quantiles (indicated by vertical bars). Although the observed relative frequencies do not fall onto the diagonal, the deviation is still consistent with reliability. The bin boundaries were taken as $[0, 0.2, 0.4, 0.6, 0.8, 1]$. The observed frequencies are plotted versus r_k (as defined in Equation 3). Plotting versus the bin centres would have caused substantial deviations from the diagonal (circle).

two synthetically generated data sets. These two datasets were constructed to illustrate a case where closeness of observed frequencies to the diagonal does *not* necessarily mean greater consistency with the null hypothesis of the data being reliable; both data sets are slightly unreliable by design. In Figure 1, a reliability diagram with consistency bars for data set I is shown. All observed relative frequencies (marked with (+)) fall within the 5%-95% consistency bars. Although the observed relative frequencies are obviously not on the diagonal, the deviation is not inconsistent with what would be expected if the forecast was reliable. Figure 2 shows a reliability diagram with consistency bars for data set II. The observed relative frequencies (marked with ($\times$)) are closer to the diagonal than in Figure 1, but the observed relative frequencies lie further outside the 5%-95% consistency bars. Figure 3 shows again the reliability diagrams for both data set I (+) and data set II

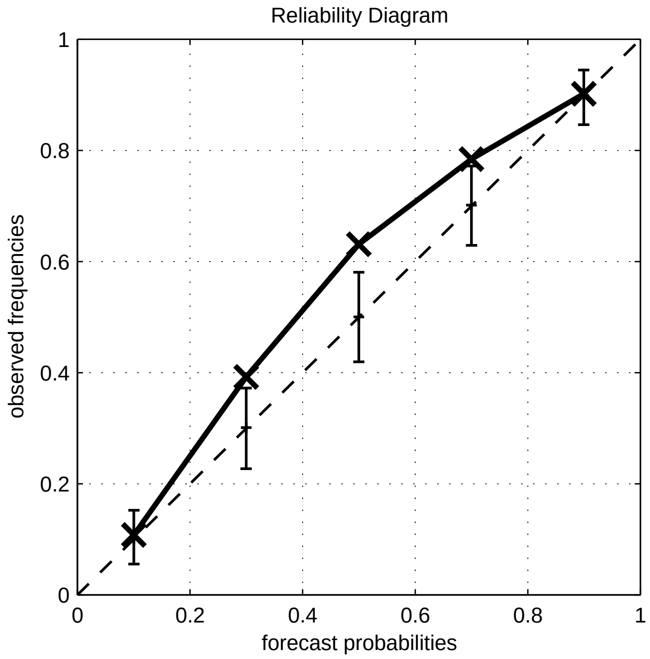


Figure 2: Reliability diagram for data set II using consistency bars. Although the observed relative frequencies are closer to the diagonal than in Figure 1, there are observed relative frequencies that do *not* fall within the 5%-95% bootstrap limits. The bin boundaries were taken as $[0, 0.2, 0.4, 0.6, 0.8, 1]$. The observed relative frequencies are plotted versus r_k (as defined in Equation 3).

($\times$), now overlaid in one viewgraph and without consistency bars. Since the observed relative frequencies of data set II do indeed ($\times$) lie closer to the diagonal, Figure 3 alone might lead to the false conclusion that data set II ($\times$) is a more reliable forecast. Another way to see this is by looking at the variance of the observed frequencies for the individual bins. For data set I, this variance is larger, thus the deviations from the diagonal are consistent with sampling errors. For data set II though, the deviations, albeit smaller than for data set I, are not consistent with sampling errors, since the variance of the observed frequencies is smaller as well.

Note also that in Figures 1, 2, and 3, the observed relative frequencies have been plotted versus r_k (see Equation 3) rather than the arithmetic centres of the bins. As has been explained in Section 2, the observed relative frequencies of a reliable forecast system are expected to be equal to r_k, the average over the forecast values in the bins, not the arithmetic center. The impact of

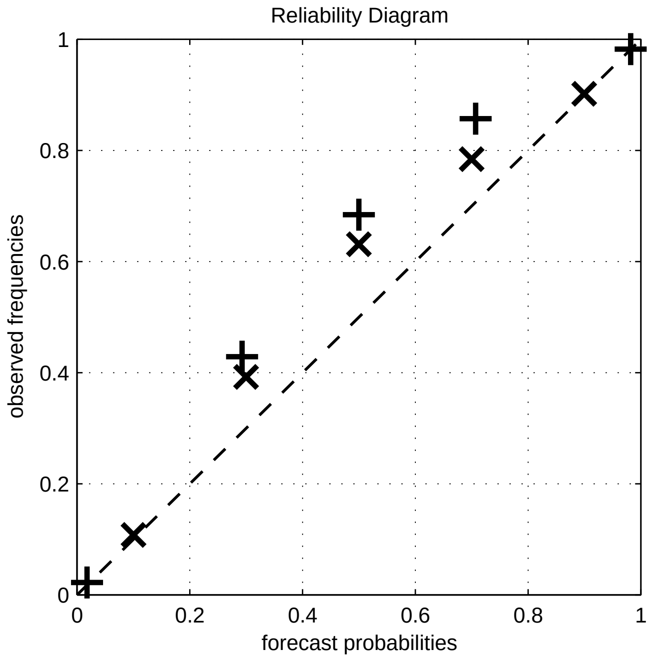


Figure 3: Reliability diagram without consistency resampling bars for data set I (+) and data set II (×) used in Figures 1 and 2 respectively. This plot gives the wrong impression that data set II (×) represents a more reliable forecast than data set I (+).

this effect is demonstrated in Figure 1 for the lowest forecast bin (stretching from 0 to 0.2). By design, the forecast values for this bin are reliable. The distribution of forecast values in this bin is, however, very uneven. Plotting the observed relative frequency versus the arithmetic center of the bin would have caused the observed relative frequency to be off the diagonal (indicated by a circle (◦)), giving the false impression that the forecast is unreliable. As an alternative, the consistency bars could be plotted at the arithmetic center of the bin as well. In this case, both the consistency bar and the observed relative frequency for the lowest forecast bin would be off the diagonal, but the observed relative frequency would again fall into the consistency bar, thereby correctly indicating reliability.

2.2 Reliability Diagrams on Probability Paper

Employing the consistency bars to indicate the distance of the observed relative frequencies from the diagonal *in probability* suggests a new graph, containing essentially the same information as the reliability diagram but plotted differently. In this graph, the x–axis still represents forecast values. The y–axis however, instead of showing the observed relative frequency directly, represents the probability that the observed relative frequency would have been closer to the diagonal than the actual observed relative frequency if the forecast was reliable (see Figure 4). In other words, the y–axis gives the likelihood of the observed frequency if the forecast was reliable, rather than the actual value of the observed frequency. This graph, showing the distance in probability of the observed relative frequencies from that expected for a reliable forecast system, will be referred to as a reliability diagram on *probability paper*. If the observed relative frequency fell exactly on top of the consistency bar, for example, its value on probability paper would be 0.9, since there is a 90% chance of the observed relative frequency to be within the range of the consistency bar if the forecast was reliable. The reliability diagrams on probability paper are mirrored vertically along the diagonal. Observed frequencies falling above the diagonal are plotted onto the upper panel, observed relative frequencies falling below the diagonal are plotted onto the lower panel.

Strictly speaking, the y–axis in these plots represents the distance in probability from the 50% quantile rather than from the diagonal. These two would coincide if the chance of the observed relative frequency falling either above or below the diagonal were exactly 50%. Although this is not quite the case, we found the difference to be very small (the true position of the diagonal is indicated on probability paper by a dash–dot line).

In principle, reliability diagrams on probability paper could be computed by the same resampling technique used to compute reliability diagrams with consistency bars (see Subsection 2.1), but plotting the fraction of *surrogate* observed frequencies closer to the diagonal than the *actual* observed frequency, rather than plotting the observed frequencies directly. Since reliability diagrams on probability paper require a high resolution at quantiles close to zero and close to one due to the logarithmic y–axis, the consistency resampling as presented in Subsection 2.1 was slightly modified: we used the same resampling to obtain surrogate bin populations, but rather than creating surrogate observed frequencies directly, we determined the value on the y–axis by employing the binomial distribution. In detail, this is done as follows. As in Subsection 2.1, surrogate forecasts are created, from which we obtain surrogate bin populations $\hat{I}_k$ and surrogate representative forecasts $\hat{r}_k$. If the forecast were reliable, the number of surrogate events in each bin would follow a binomial distribution with parameters $\hat{I}_k$ and $\hat{r}_k$. Consequently, the fraction z_k of surrogate observed frequencies smaller than the actual observed frequency f_k is given by

$$z_k = \mathcal{B}([f_k \hat{I}_k]; \hat{I}_k, \hat{r}_k)$$

where $\mathcal{B}$ is the cumulative Binomial distribution and $[f_k \hat{I}_k]$ is $f_k \hat{I}_k$, rounded to the nearest integer. The reliability diagram on probability paper comprises a plot of z_k versus f_k.

Figure 4 shows reliability diagrams for the two synthetic datasets considered in the previous section, but plotted on probability paper. Again, data set I is plotted with (+), the data set II is plotted using (×). Data set II, seemingly closer to the diagonal in the standard reliability diagram Figure 3, is clearly further away in probability for three out of five bins, another bin effectively being a tie. This indicates that data set II is actually less reliable than data set I. The dash–dot line represents the exact position of the diagonal, which usually falls close to the zero line.

Some general hints as to how reliability diagrams on probability paper should be read seem to be in order. For a given bin, there is a 90% chance the observed relative frequency falls within the range labelled 0.9 on the y–axis. Likewise there is a 99% chance the observed relative frequency falls within the range labelled 0.99 etc. Thus the chance of being outside the 0.9–band in any one bin is 0.1. Of course the chance of seeing *some* points of the entire plot falling outside the 0.9–band is larger. For example, the chance of all points on a reliability diagram with 6 bins falling inside the 0.9–band is only $0.9^6 \cong 0.53$. A band that would encompass *all points* with a 90% chance is indicated by the dashed line[3]. In other words, if the forecast is reliable, then there is a 90% chance that the entire diagram falls within the *dashed* line.

3 Relation of Consistency Resampling and the Consistency Bars With the Common Bootstrap

Consistency resampling is related to, but distinct from, bootstrap resampling in statistics (Efron and Tibshirani, 1993). Both are resampling techniques, but the consistency resampling differs from the common bootstrap of statistics in that the latter bootstrap resamples the data to extract an estimate of the uncertainty in the statistic of interest (in this case this would be the observed relative frequencies), while the consistency resampling quantifies

[3]This line indicates the Bonferroni corrected 90% level (see Bonferroni (1936)). Independence of the individual bins was assumed for this calculation.

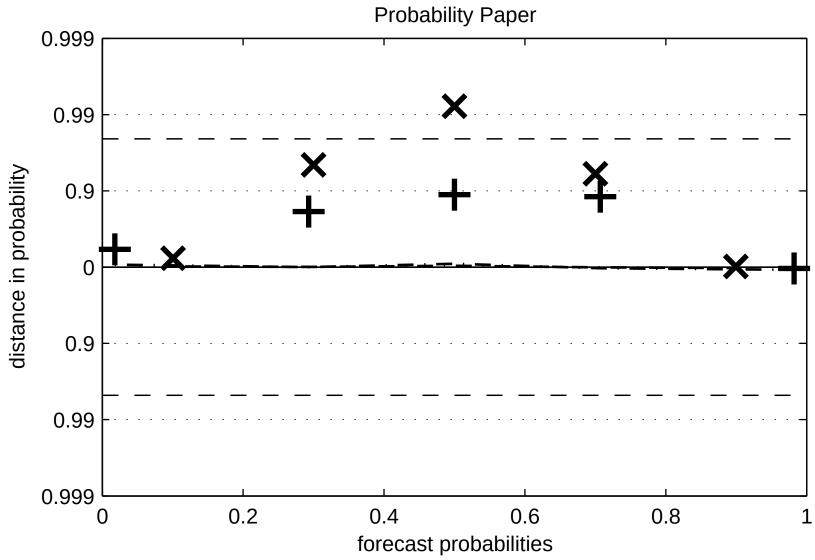


Figure 4: Reliability diagrams on probability paper for the data sets I (+) and II (×). Some observed relative frequencies that seemed to be close to the diagonal in Figure 3 are clearly further away in probability. The dash–dot line represents the exact position of the diagonal, which usually fall close to the zero–line. For a reliable forecast, we would expect the entire diagram with a 90% chance to fall within the dashed lines.

the range of results expected if the forecast values were in fact correct probabilities. While the traditional bootstrap would resample the forecast values, thereby quantifying the expected variation in the observed relative frequencies, the consistency resampling quantifies the range of relative frequencies which would be expected if the predicted probabilities were, in fact, reliable. This also differs from the common use of surrogate data in geophysics, where a null test is set up in the hope that it will be rejected (for a discussion, see Smith and Allen (1996)). Both consistency resampling and bootstrap resampling are common in statistics (see Efron and Tibshirani (1993)), and both are confusingly referred to as "the bootstrap". The two techniques address different questions and yield different information, so distinguishing them is important.

Two very different approaches to bootstrap resampling can be applied to the reliability diagram, one resampling by-bin, the other resampling the entire diagram. Resampling within each bin yields the sampling uncertainty in the relative frequency of that bin; this is easily displayed by either form of the diagram and the resample realisations in different bins are independent. Alternatively, resampling from the entire-diagram will alter the number of forecasts in each bin (or even the bins themselves) and introduces an inter-dependency between the bootstrap realizations across the diagram. This inter-dependency makes the entire-diagram alternative more difficult to evaluate visually, suggesting that it is best displayed in a manner that evaluates the diagram as a whole, and while both alternatives have their uses we consider only the by-bin method in the remainder of this paper.

The presentation of the reliability diagram on probability paper opens the possibility to use both the consistency resampling and the common bootstrap in parallel. After plotting the reliability diagram on probability paper, common bootstrap resamples for each bin can be added to the plot to give an immediate visual impression of the sampling uncertainty in each frequency. Examples including this addition are discussed in the end of the following section.

4 Numerical Examples

Consider daily forecasts of the two metre temperature at London Heathrow weather station taken at 12 noon, specifically whether this value exceeded the monthly average computed for the previous 21 years of data (1980-2000). The forecast was a 51 member ensemble, provided by downscaling output from ECMWF's global ensemble forecasting system (see Appendix II for details of this procedure). The forecast value X_i was taken as the fraction of ensemble members exceeding the current monthly average.

Reliability diagrams where plotted for London Heathrow, contrasting the years 2003 and 2005 (containing 365 forecast instances each) for lead times 1, 3, and 10 days. Figures 5, 6, and 7 show reliability diagrams on probability paper, while Figures 8, 9, and 10 show conventional reliability diagrams with consistency bars. The overall impression is that, as far as reliability is concerned, the forecasts have improved.

For lead time 1 day (Figures 5 and 8), the reliability of 2003 and 2005 is generally comparable. For forecast probabilities between 0.4 and 0.8, the observed relative frequencies fall within the 90% range for both years. It seems though that in 2003, the forecast system struggled to get the lower probabilities right (there were generally more events than the forecast probabilities would suggest), while in 2005 the system forecasts the high probabilities poorly. The confidence bars for the extreme events are very tight. This is a typical situation where the forecast is shown to be unreliable, although the observed relative frequencies appear "close to the diagonal" in terms of usual distance.

In 2003, there were considerably more days where the temperature exceeded the monthly average than in 2005 (247 vs 203), the year 2003 being one of the hottest on record in Europe. Therefore, a system usually overestimating the frequency of days hotter than normal would have scored better in 2003 than in 2005. It is not clear though whether high or low forecast probabilities would have been affected more. For lead time 3 days (Figures 6 and 9), the forecast was better in 2005 in every single bin and is accepted as reliable at a 0.9 significance level. In fact, it is accepted even at a significance level of $0.9^6 \cong 0.53$, since all points fall within the 0.9 confidence band (see discussion at the end of Subsection 2.2). For lead time 10 days (Figures 6 and 9), the forecast system seems to have given probabilities which were systematically too low in 2003. Since this bias is fairly uniform over different bins, simply subtracting a constant offset would have improved the performance. The forecast system obviously improved in 2005, although the overall reliability appears to be not as good as for lead time 3 days.

Using the same three day lead time data as in Figures 6 and 9, Figure 11 shows the traditional reliability diagram with by-bin bootstrap-resampled observed frequencies in addition to the consistency bars. Bootstrap resamples are shown as dots while the consistency bars for the original forecasts is shown as before. Technically, each bootstrap resample has its own corresponding consistency bar; the dots here only reflect sampling uncertainty for each bin: for instance, is resampling likely to yield a value "near" the diagonal? This is, in fact, much more common in the 2005 forecasts than the 2003 forecasts. Note that variations in the variance of the dots indicate bins with small populations.

Figure 12 shows the same analysis on probability paper, in effect adding the bootstrap resample dots to Figure 6. This figure is constructed to be consistent with Figure 11: each dot here is the resampled frequency in terms of the probability defined by the true-sample frequency; an interesting alternative (not shown) is to place each dot at the individual probability implied by that bootstrap resample. The choice of which resampling technique to apply is determined by the question to be resolved; to avoid confusion the details of the bootstrap scheme should be stated in every case.

Figures 11 and 12 add confidence to the conclusions drawn from Figures 6 and 9, namely that the forecast system is more reliable in 2005 than in 2003. This is clear after taking into account uncertainties in the observed frequencies (to the extent that this is simulated by resampling the original data). The resampled observed frequencies are much more often within the confidence bars for the 2005 data than for the 2003 data (see Figure 11). Similarly on probability paper, the resampled observed frequencies fall within higher probability regions for the 2005 data than for the 2003 data (note that in the 2003 data, dots even fall beyond the range of the probability graph in three of the bins).

5 Conclusion

The reliability diagram is a common diagnostic aid for quickly evaluating the reliability of probabilistic forecast systems. A reliable forecast should have a reliability diagram close to the diagonal, but how close exactly? Both forecasts and observed relative frequencies fluctuate, and to interprete a reliability diagram correctly, these deviations have to be taken into account. We have introduced a consistency resampling method for assigning consistency bars to the diagonal of the reliability diagram, indicating the region where a reliable forecast would fall into, or in other words, how likely the observed relative frequencies are under the assumption that the predicted probabilities are accurate.

We have shown a numerical example using synthetic data where two forecasts are compared using the proposed technique. One of the forecasts, although seemingly closer to the diagonal than the other, turns out to be further away in probability and therefore constitutes a less reliable forecasts. The method was also applied to anomaly data for London Heathrow temperature and ECMWF ensemble forecasts from 2003 and 2005. The overall reliability appears to have improved. We argue that the more reliable reliability diagrams as introduced in this paper add more credibility to this finding than conventional reliability diagrams.

Two approaches to make the quantification of just how reliable the forecast system is more visually accessible have been suggested; each contains the same reliability information as the traditional diagram. Examining a variety of meteorological forecast systems suggests that in practice diagrams which look “good” in the traditional presentation often have many points *well beyond* the .90 probability threshold (suggesting the forecast system is not reliable). We hope these new diagrams make reliability diagrams even more valuable in practice.

Acknowledgements

Support for this work was provided by the DIME EPSRC/DTI Faraday Partnership. We would like to thank European Centre for Medium Range Weather Forecasting (ECMWF) and especially R. Hagedorn and F. Doblas Reyes for provision of the DEMETER data. We also acknowledge fruitful discus-

sions with the members of the CATS centre as well as Antje Weisheimer, ECMWF.

Appendix I: Generating Data with Specified Reliability Diagram

This Appendix discusses how to create artificial forecast verification pairs with a given reliability diagram and Brier Skill Score. Let $p_i \in [0,1], i = 1 \dots N$ be an iid random variable with a probability density function $g(p)$. In order to generate a random variable $Y_i \in [0,1], i = 1 \dots N$ for which p_i is a reliable forecast, draw another random variable R_i, independent from p_i, from a uniform distribution on the unit interval and then set

$$Y_i = \begin{cases} 1 \text{ if } R_i < p_i \\ 0 \text{ else} \end{cases} \tag{4}$$

To see that this forecast is reliable, note that

$$\begin{aligned} & P(Y_i = 1 | p_i = z) \\ & \quad = \quad P(R_i < p_i | p_i = z) \\ & \quad = \quad P(R_i < z | p_i = z) \\ & \quad = \quad P(R_i < z) = z \end{aligned}$$

where the equality signs follow (in that order) from the definition of Y_i, the properties of conditional probabilities, the fact that R_i is independent of p_i and from its uniform distribution. This technique is used to generate "fake" verifications consistent with a given set of forecast values to draw consistency resampling bars (see Section 2.1). Draws from an unreliable forecasts with a specified reliability graph (that is with a specified reliability diagram in the large sample limit), can be generated as well. If $r(p)$ is the desired reliability graph, generate Y_i by applying Equation (4) but using $r(p_i)$ instead of p_i. Then the limiting observed relative frequencies corresponding to p_i are $r(p_i)$, as desired.

Appendix II: Downscaling the ECMWF data

This appendix explains the method used to downscale the ECMWF data. ECMWF publishes the output of their global model on a certain grid. This grid lacks a point *exactly* at London Heathrow, but even if it did, we would not expect the forecast at that grid point to be all the model has to say about temperatures around London Heathrow. Furthermore, ECMWF also publishes a forecast for London Heathrow explicitely, which is interpolated from several neighbouring grid points. This interpolation is done according to an interpolation scheme which does not involve any fitting to actual station data (this is why we use the word "interpolation" rather than "fitting"). Using these five ensemble forecasts (the four neighboring gridpoints and ECMWF's interpolated forecast), featuring 51 ensemble members each, we produce a forecast for London Heathrow using a linear fit as follows: Letting $x_1 \dots x_5$ denote the mean values of the five ensemble forecasts, we find coefficients $a_0 \dots a_5$ by fitting

$$z = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + a_4 x_4 + a_5 x_5 \tag{5}$$

to the observations at London Heathrow using a least squares error criterion. Then Equation 5 is applied to the entire ensembles (rather than just the means) to find the final ensemble. The linear fit is carried out using one years worth of data (2001). The same procedure is applied to each lead time individually. Note that the rest of the analysis described in this paper is carried out on different data, namely on the years 2002-2005.

References

Frédéric Atger. Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review*, 131:1509–1523, 2003.

Frédéric Atger. Estimation of the reliability of ensemble based probabilistic forecasts. *Quaterly Journal of the Royal Meteorological Society*, 130:627–646, 2004.

Carlo Emilio Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, first edition, 1993.

Renate Hagedorn, Francisco R. Doblas Reyes, and Tim Palmer. The rationale behind the success of multi-model ensembles in seasonal forecasting. part i: Basic concept. *Tellus*, 2004.

Ian T. Jolliffe and David B. Stephenson. *Forecast Verification*. Wiley, 2003.

A. H. Murphy and R. L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115(7):1330–1338, Jul 1987.

Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Applied Statistics*, 26(1):41–47, 1977.

L.A. Smith and M.R. Allen. Monte carlo ssa: Detecting irregular oscillations in the presence of coloured noise. *J. Climate*, (9):3373–3404, 1996.

Leonard A. Smith. The maintenance of uncertainty. *Proceedings of the International School of Physics Enrico Fermi*, 133:177–246, 1997.

Daniel S. Wilks. *Statistical Methods in the Athmospheric Sciences*, volume 59 of *International Geophysics Series*. Academic Press, first edition, 1995.

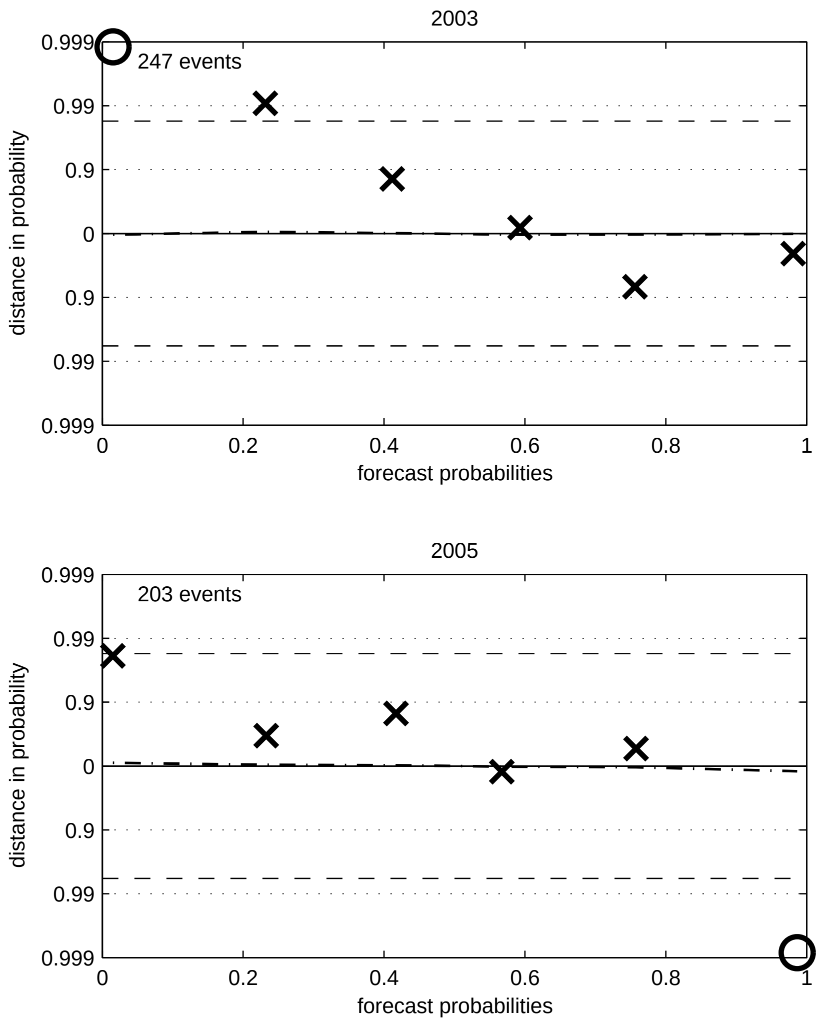


Figure 5: Reliability diagrams on probability paper for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 1 day. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).

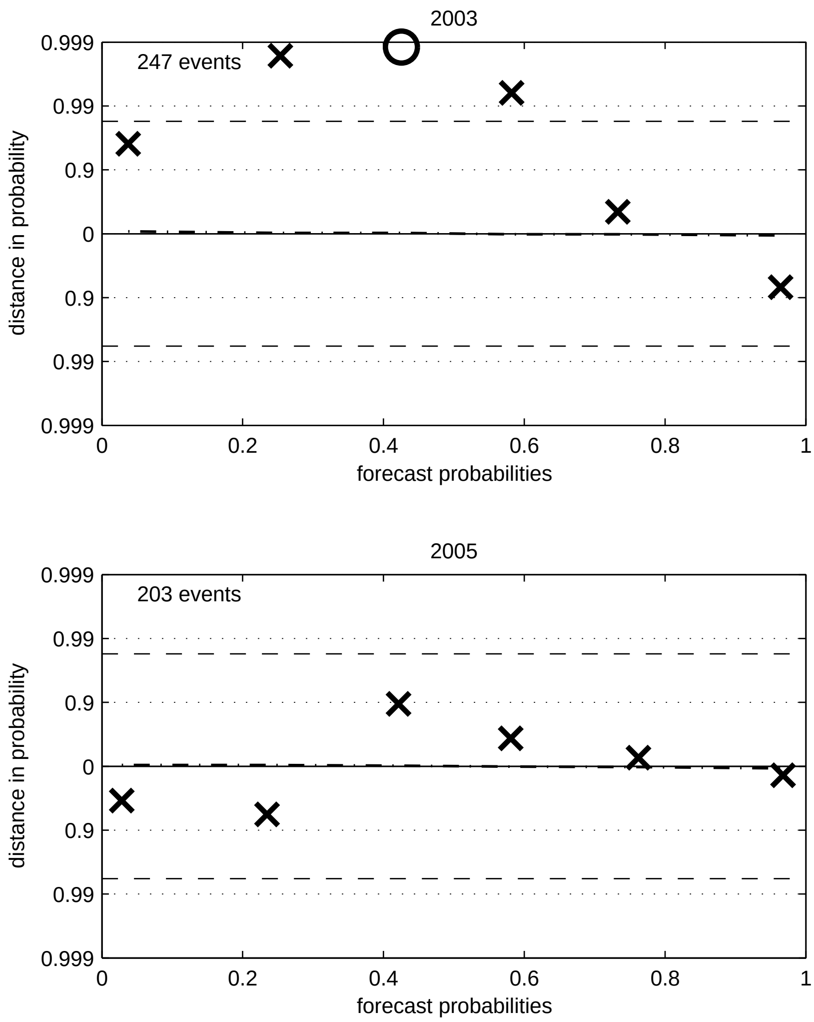


Figure 6: Reliability diagrams on probability paper for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 3 days. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).

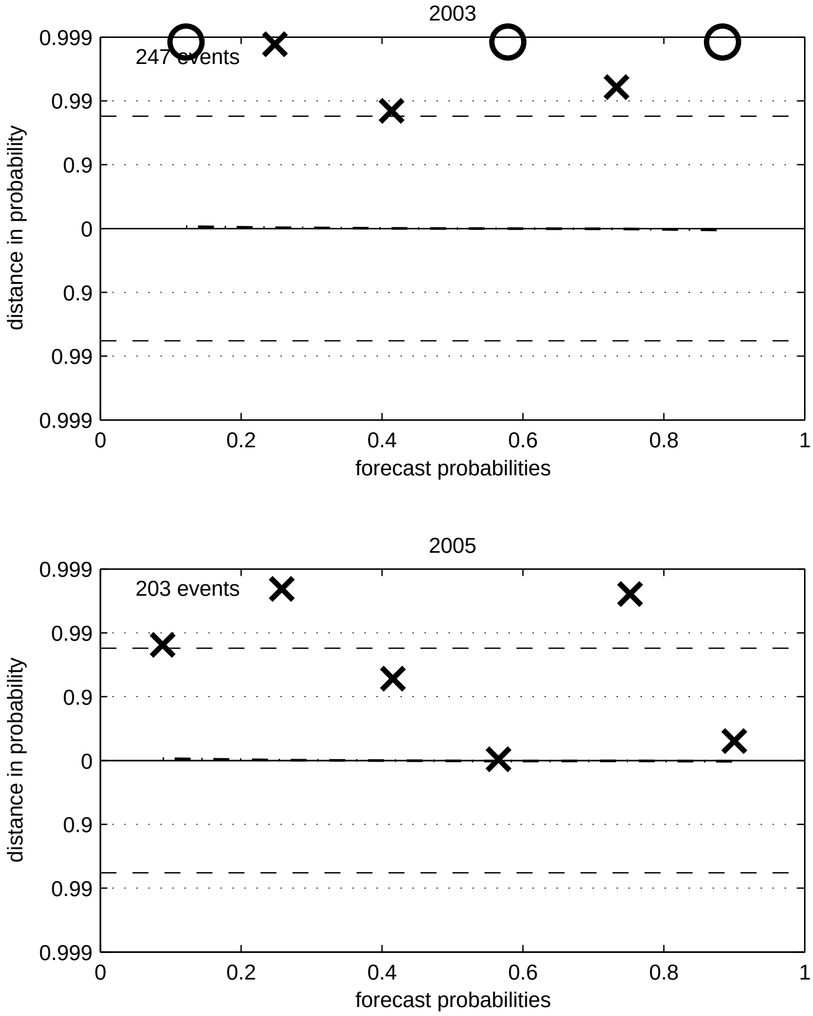


Figure 7: Reliability diagrams on probability paper for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 10 days. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).

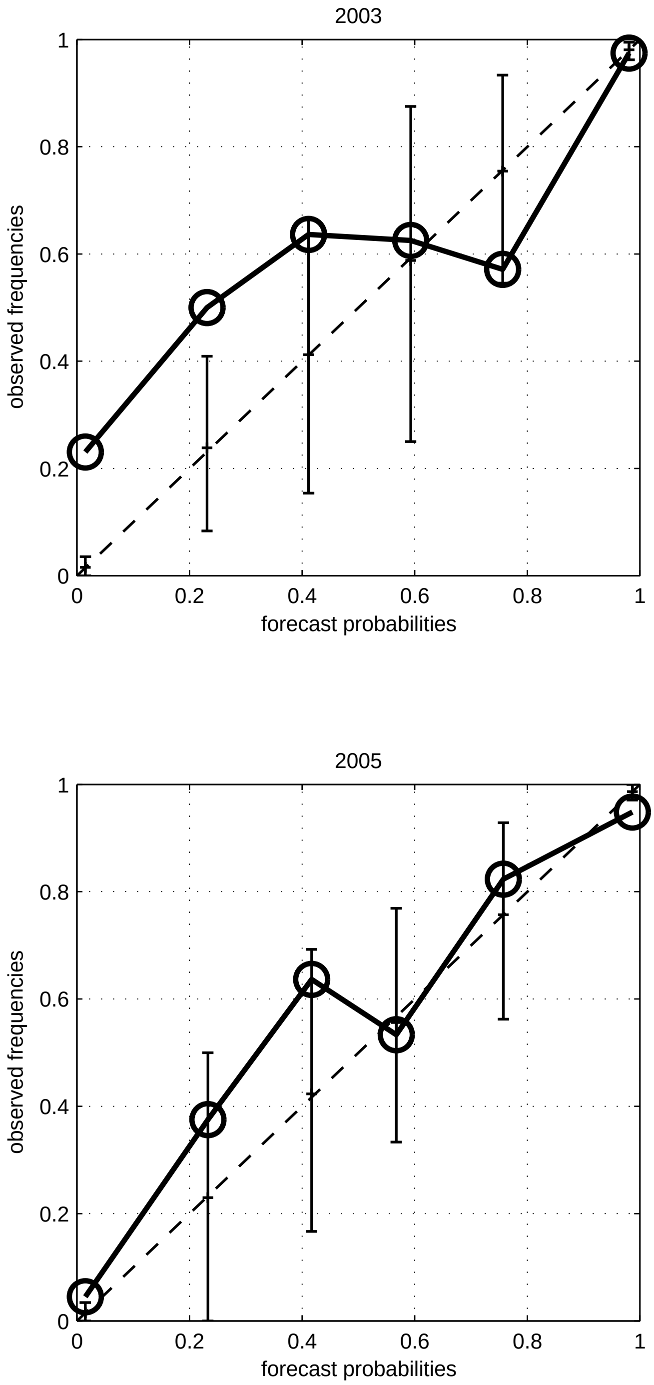


Figure 8: Reliability diagrams with consistency bars for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 1 day. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).

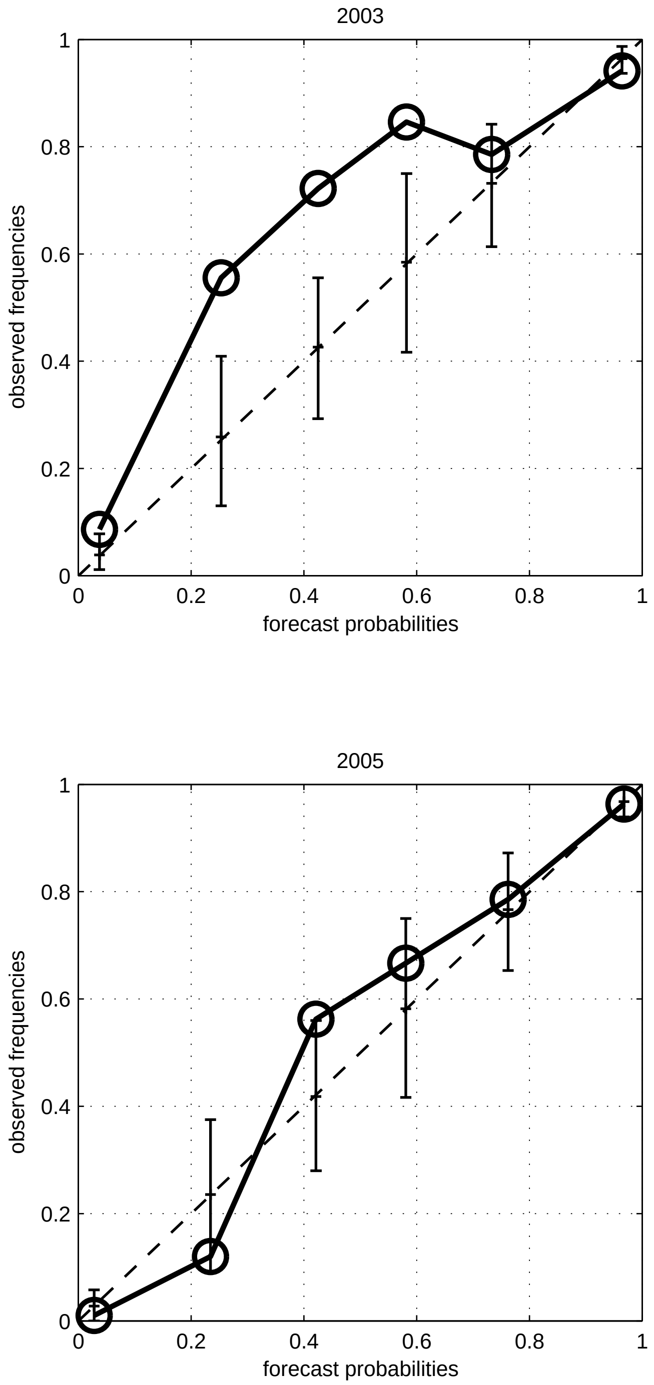


Figure 9: Reliability diagrams with consistency bars for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 3 days. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).

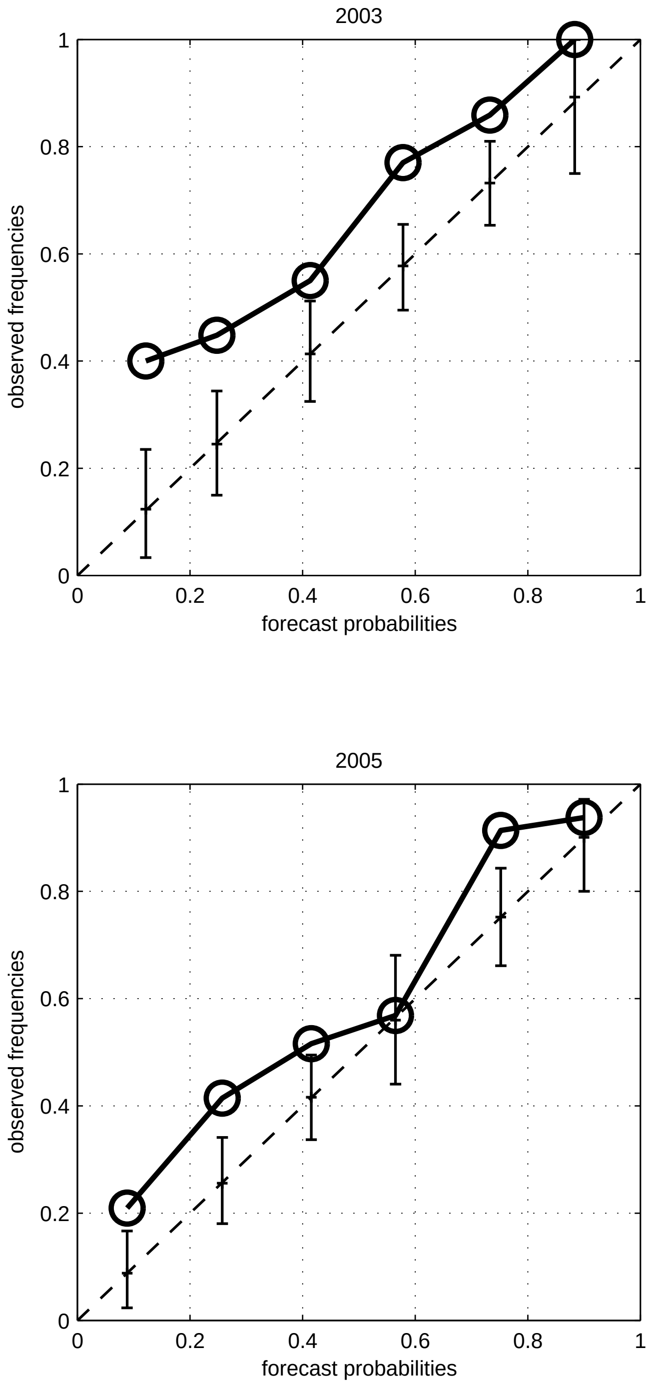
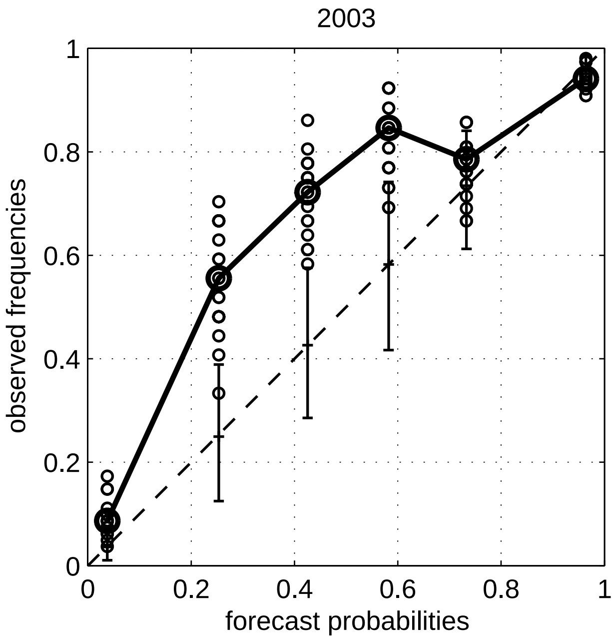


Figure 10: Reliability diagrams with consistency bars for two metre temperature forecasts (above/below monthly average) at London Heathrow. Forecasts were provided by ECMWF's ensemble forecasts. The lead time is 10 days. The upper (resp. lower) panel shows the performance for 2003 (resp. 2005).



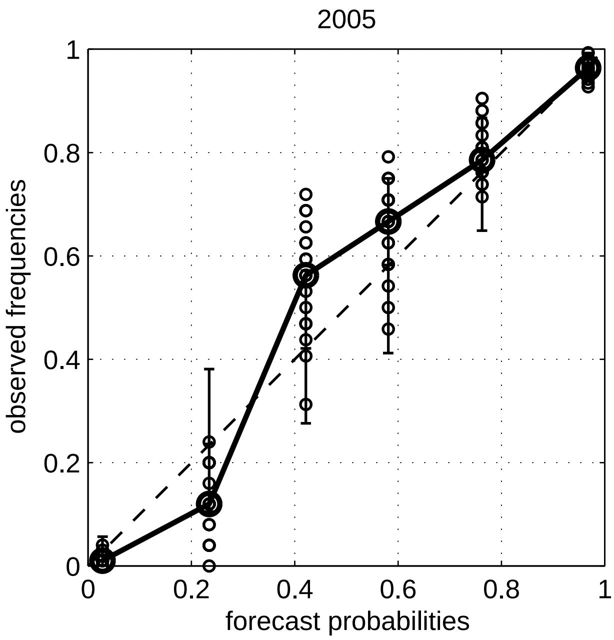


Figure 11: Reliability diagrams with consistency bars and resampled observed frequencies (small circles). The data is the same as for Figure 9 and was resampled 20 times. The actual observed frequencies are indicated with a cross, as before.

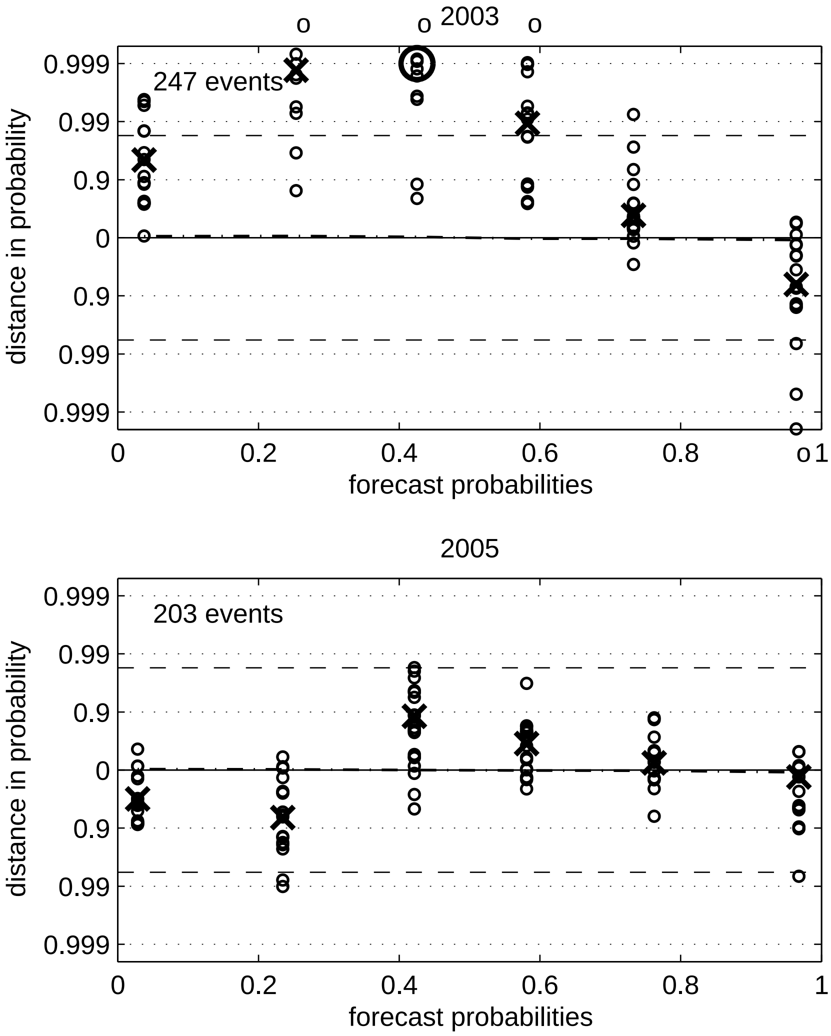


Figure 12: Reliability diagram on probability paper with resampled observed frequencies (small circles). The data is the same as for Figure 6 and was resampled 20 times. The actual observed frequencies are indicated with a cross, as before. The little "o"–marks either above or below the panel indicate that some resampled observed frequencies are outside the range of the y–axis.