Fairplay in the verification of operational quantitative precipitation forecasts

M. Göber*[1], C.A. Wilson*, S.F. Milton*, D. B. Stephenson+

*Met Office, Bracknell, U.K.

+Dept. of Meteorology, University of Reading, U.K.

Submitted to Journal of Hydrology

(Manuscript received 30 January 2003, revised 11th June 2003)

[1]Corresponding author. Deutscher Wetterdienst, Kaiserleistr. 42, 63067 Offenbach, Germany, Fax: 0049 69 80623676, e-mail: martin.goeber@dwd.de

Abstract

The accuracy of weather forecasts is not only influenced by the skill of the forecasting system, but also by the weather itself. Here we propose the use of the odds ratio (benefit), ORB, as a measure which is not influenced by the base-rate of the event and thus enables a fair comparison of categorical forecasts for different years, regions, events, etc. . The ORB has a simple interpretation and it permits a split of forecasting skill into contributions from forecasting the event and the non-event.

Applying this measure to operational quantitative precipitation forecasts reveals that forecasts of more extreme (rare) events have more skill than forecasts for more "normal" events which is contrary to the results typically obtained with other categorical measures traditionally used in meteorology and also to subjective perception. Both of the latter can be interpreted as a delusive consequence of the "neglect of the base-rate" effect. Further consequences are described for the composition of model trials and for the verification of forecast warnings.

Over recent years there have been trends showing a small improvement in skill and a large reduction in model bias for forecasts of slight precipitation.

Keywords: Quantitative precipitation forecast; Bayesian statistics; verification; skill; extreme events

1 Motivation

One objective in the verification of a weather forecasting system is the assessment of the skill of the *system which produces the forecast*, e.g. a numerical model. This must not be equated to the assessment of the accuracy of the *forecast* since the latter is also influenced by the weather itself, e.g. by the variable amplitude and predictability of the event to be forecast. For continuous variables, the reduction in the Mean Squared forecast Error (MSE) over a reference forecast is often used to define a skill score (SS), i.e.:

$$SS = \frac{MSE_f - MSE_{pers}}{MSE_{perf} - MSE_{pers}} , \tag{1}$$

where $MSE_{f,pers,perf}$ are the mean squared errors of the forecasts, persistence forecasts and perfect forecasts, respectively (Wilks (1995)). Following Murphy (1988), the MSE of forecasts f and observations o can be decomposed into the bias $B = E(f) - E(o)$, the variance of the observations s_o^2 or forecasts s_f^2 and the correlation between forecasts and observations, r_{of} or $r_{o\,pers}$:

$$SS = 1 - \frac{B^2 + s_o^2 + s_f^2 - 2 s_o s_f r_{of}}{2 s_o^2 (1 - r_{o\,pers})} . \tag{2}$$

This decomposition of the skill score reveals that for a given MSE_f the skill of, or improvement due to, the forecasting system is rated higher when the variance of the weather process is larger or when the predictability of the weather was lower. Here the correlation of the persistence forecast with observations acts as a surrogate measure of predictabil-

ity[1]. In other words, a forecast is less penalised when it "risks" larger errors by going far from the mean in a situation where the variance is high. Conversely, when the variance of the weather process is low, only a bad forecast has large errors or, similarly, only a bad forecast would miss predictable components like the daily cycle or a constant weather in a blocking situation.

Generally speaking, a skill score should provide a fair assessment of a forecasting system by isolating the accuracy of the forecasting system from the accuracy which a user has "for free", for instance given by knowledge of mean climatology and/or current conditions. Forecast systems for different seasons, regions, events and years can only be compared fairly by using a skill score which at least partly accounts for the varying "difficulty" to forecast the weather. A desirable property is the ability to compare skill from year to year, since it is important to highlight trends in skill over time to administrators and the public. Assuming a normal-linear model, Krzysztofowicz (1992) derives a measure of correlation for categorical forecasts of continuous predictands between forecasts and observations which allows a meaningful comparison of forecasts taking into account prior knowledge.

Let us look at the annual cycle of verification of northern hemispheric 500 hPa geopotential height forecasts (Fig. 1) as an example of this difference between the accuracy of the forecast as measured by the root of the MSE_f and the skill of the forecasting system as measured by the skill score. The MSE_f is highest in winter but this is largely due to the high variance in winter. This can be seen in the skill score which shows the same annual cycle, i.e. according to this measure the *forecasting system* is most beneficial in

[1]Measuring predictability is a complicated issue and focus of much current research. The skill of persistence forecasts is only a very crude, but handy measure of predictability. For the sake of brevity we will further refer to it as the predictability indicator.

winter. The use of such skill scores for forecasts of *continuous* variables has long been accepted practise. But how should one proceed for forecasts of *categorical* variables?

Binary yes-no forecasts of an event (often defined by the exceedance above some threshold) are the simplest form of a categorical forecast. The joint distribution of forecasts f and observations o (see Table 1 for definitions) can be documented in a 2x2 contingency table. A widely used measure to summarise the skill in this situation is the Equitable Threat Score (ETS, or Gilbert skill score in Schaefer (1990)) which is given as the improvement over chance of the probability for a hit relative to the probability for a threat which would not have been foreseen by chance (an overbar denotes the complement of an event, i.e. a non-event; see Table 1 for further definitions):

$$ETS = \frac{p(f,o) - p(f)\,p(o)}{p(f,o) + p(f,\bar{o}) + p(\bar{f},o) - p(f)\,p(o)} \tag{3}$$

Thus the ETS contains some aspect of a skill score by comparing the forecast hits against those obtained by chance with a system with the same $p(f)$. But it is explicitly dependent on the sample climate. Furthermore, there seems to be no obvious probabilistic interpretation of the ETS or indeed the difference or ratio of the ETS's of two forecasts. Thus it is not clear how to properly compare the ETS of a forecast to a persistence forecast and ultimately achieve a fair comparison of categorical forecasts.

In the present paper, we will show that the odds ratio discussed by Stephenson (2000) can be used as a straightforward measure to fairly compare the skill of categorical forecasts. Section 2 derives the odds ratio from Bayesian statistics, i.e. we show how the use of a forecast reduces the uncertainty about the future from the uncertainty in prior, climatological knowledge. Section 3 illustrates this concept using operational Met Office quantitative precipitation forecasts (QPF's). Section 4 gives a summary of the main results and suggestions for further applications of the concepts presented here. Note, that

we will solely focus on measuring the skill of the forecast and not the value of a forecast to a user (e.g. Thornes and Stephenson (2001)), which has to be treated separately (Murphy (1993)).

2 A Bayesian view of verification

Since Finley (1884), numerous measures of the accuracy of categorical forecasts have been proposed and discussed in the meteorological literature (see Marzban (1998) and Jolliffe and Stephenson (2003) for recent reviews). Murphy and Winkler (1987) established the assessment of the joint distribution of forecasts and observations as a general framework for forecast verification (as applied for instance by Brooks and Doswell III (1996)). Mason (1982) pioneered the use of the relative operating characteristic (ROC) in meteorology, a graph of the hit rate versus the false alarm rate which is based on signal detection theory. Recently, Stephenson (2000) brought the odds ratio to attention in meteorology, a measure which is widely used in the medical and social sciences (Agresti (1996)). The odds Ω or risk of an event is the ratio of the probability p that the event occurs to the probability $1-p$ that the event does not occur. The odds ratio θ compares the conditional odds of making a good forecast (a hit) to the odds of making a bad forecast (a false detection), i.e.:

$$\theta = \frac{\Omega(f|o)}{\Omega(f|\bar{o})} = \frac{p(f,o)}{p(\bar{f},o)} \frac{p(\bar{f},\bar{o})}{p(f,\bar{o})} \,. \tag{4}$$

For example, an odds ratio of five means that the odds for correctly detecting precipitation are five times higher than the odds for wrongly forecasting precipitation occurence. The forecasts have positive skill if the odds ratio is significantly larger than one and negative skill (worse than pure chance) when it is significantly smaller than one (see Stephenson (2000) for signifcance tests of the odds ratio). In the following we will derive the odds

ratio in a way that leads to a slightly different interpretation, namely as a product of the odds for forecasting the event correctly and the odds for forecasting the non-event correctly.

We start by noting that a joint probability of forecasts and observations can be viewed from two perspectives (Murphy and Winkler (1987), e.g.:

$$p(f,o) = p(f)\,p(o|f) = p(o)\,p(f|o) \tag{5}$$

The first factorisation splits the joint distribution into a measure $p(f)$ of the *refinement* (or sharpness) of the forecast and a measure $p(o|f)$ of the *calibration* (or reliability) of the forecast. The second view splits the joint distribution into a measure $p(o)$ of the forecasting situation called *base rate* (in meteorology its usually the sample climatology, i.e. the relative frequency of an event) and a measure $p(f|o)$ called the *likelihood* of the forecast, which indicates how well the forecast discriminates between events and non-events.

We now adopt the view of a user of a forecast (e.g. a gardener thinking about watering the plants) who gets a forecast for precipitation and might wonder what the probability is that the precipitation forecast is correct. Thus we rearrange eqn. (5) to obtain Bayes theorem:

$$p(o|f) = \frac{p(o)}{p(f)}\,p(f|o)\ . \tag{6}$$

Thus, only if the frequency bias:

$$FB = \frac{p(f)}{p(o)} \tag{7}$$

equals one, i.e. when the event was forecast as often as observed, then the user can expect the event to occur with the same probability $p(o|f)$ given the forecast of the event as the user knows from past verification $p(f|o)$. If this is not the case a recalibration of the

forecast can be performed (Jolliffe and Stephenson (2003)). The conditional probability that the precipitation forecast is wrong is:

$$p(\bar{o}|f) = \frac{p(\bar{o})}{p(f)}\, p(f|\bar{o}) \,. \tag{8}$$

Note, that for a rare event and an unbiased forecast this conditional probability is much higher than the conditional probability $p(f|\bar{o})$ for the forecast to fail to detect dry weather. In other words, eqn. (8) says that for a rare event it is very likely that a rain forecast goes wrong just because there are only a few occasions to be right anyway, whereas the forecast is very unlikely to fail to detect dry weather since there is almost always dry weather. Notice the potential for confusion when one unknowingly equates the two conditional probabilities $p(\bar{o}|f)$ and $p(f|\bar{o})$ in situations of rare or frequent events. It leads to numerous failures of judgement and decision making in everyday life (Plous (1993)), one of which is called "prosecutor's fallacy" in legal and statistical circles.

The odds for a correct precipitation forecast are:

$$\Omega(o|f) = \frac{p(o|f)}{p(\bar{o}|f)} = \frac{p(f,o)}{p(f,\bar{o})} \,, \tag{9}$$

which can be rewritten as

$$\underbrace{\Omega(o|f)}_{\text{posterior}} = \underbrace{\Omega(o)}_{\text{prior}} \underbrace{\frac{p(f|o)}{p(f|\bar{o})}}_{\text{likelihood ratio}} \,, \tag{10}$$

where

$$\Omega(o) = \frac{p(o)}{p(\bar{o})} \,. \tag{11}$$

A Bayesian interpretation of eqn. (10) reads that the *posterior odds* $\Omega(o|f)$ to observe precipitation given a precipitation forecast are equal to the *prior*[2] *odds* $\Omega(o)$ of precipi-

[2]In order to use the Bayesian terminology we assume the climatology to be stationary, i.e. the sample climatology equals the climatology and thus we can call the climatology *prior* to the forecast when in reality we know it only *posterior* to the observation.

tation times the likelihood ratio $p(f|o)/p(f|\bar{o})$ for a correct precipitation forecast. Thus the posterior "perception" $\Omega(o|f)$ of the quality of the forecast by the user is not only determined by the quality of the model to discriminate between precipitation and no-precipitation situations ($p(f|o)/p(f|\bar{o})$), but it is also explicitly determined by "nature" ($\Omega(o)$), which is not under the forecasters control. For instance, for a rare event the odds for getting the precipitation forecast right might be low, but just because there are only a few occasions in the first place to get the forecast right relative to those many situations where there is no event and the precipitation forecast could thus turn out to be a false alarm. Furthermore, the prior odds may vary substantially between "dry" and "wet" years and consequentially the quality of the forecast might appear to vary in different circumstances just because of this statistical base rate effect. The posterior odds for a correct precipitation forecast are the quantity that interests the user, but they are not a good measure to judge fairly the quality of the forecasting system. It is the likelihood ratio which measures the contribution of the forecasting system to the overall quality of the forecast.

Murphy (1991) discusses a proposal to include the likelihood ratio of the event forecast into a probabilistic forecast. This could highlight to the user the deviation of the odds for a rare event given the forecast from the prior odds for the event.

Matthews (1996a,b) describes how the neglect of the base-rate gives the false impression that precipitation forecasts have low skill. But meteorology is not the only subject to suffer from this "unfair" appreciation. More serious errors of judgement due to base-rate neglect occur in fields like cancer screening or DNA profiling (e.g. Gigerenzer (2002)).

Similarly, the odds for a correct no-precipitation forecast are given by:

$$\Omega(\bar{o}|\bar{f}) = \frac{p(\bar{o}|\bar{f})}{p(o|\bar{f})}$$

$$
\begin{aligned}
&= \frac{p(\bar{f},\bar{o})}{p(\bar{f},o)} \\
&= \Omega(\bar{o})\,\frac{p(\bar{f}|\bar{o})}{p(\bar{f}|o)}
\end{aligned}
\tag{12}
$$

For example, this number tells a cyclist what the odds are to arrive in dry weather given the forecast was for no-precipitation. Note that the prior odds for no-precipitation $\Omega(\bar{o})$ are just the inverse of the prior odds for precipitation.

The odds ratio (eqn. (4)) can be obtained by multiplying the posterior odds for correct precipitation and correct no-precipitation forecasts:

$$
\begin{aligned}
\theta &= \Omega(o|f)\;\Omega(\bar{o}|\bar{f}) \\
\theta &= \frac{p(o|f)}{p(\bar{o}|f)}\;\frac{p(\bar{o}|\bar{f})}{p(o|\bar{f})}
\end{aligned}
\tag{13}
$$

Thus, by explicitly accounting for the two intertwined tasks of the forecasting problem, i.e. the forecast of the event and the non-event, it is possible to remove the two equal, but inverse influences of nature represented by the prior odds for precipitation and no-precipitation. We are now left with a measure which is not explicitly influenced by the base rate of the event, and so fairly reflects the overall skill of the forecasting system.

In the derivation of the odds ratio above we have adopted a forecast-oriented view in that we looked at the conditional probability for the observations given a particular forecast. A similar derivation adopting an observation-oriented view gives the same result (13) as above, i.e. the odds ratio is symmetric with respect to the conditioning which is different to, for instance, the ETS.

Note, that a forecast using only climatology (e.g. $p(f,o) = (p(o))^2$ etc.) has an odds ratio of one, i.e. the climatological forecast has no skill and thus serves automatically as a reference base line and θ can be interpreted as the skill with respect to climatology. We can further raise the baseline by asking how much better the forecasting system is than

some other reference forecast, e.g. persistence, by forming the ratio of all probabilities involved to their equivalent probabilities of the reference forecast. The added benefit from the forecasting system can then be expressed in the odds ratio benefit (ORB) θ^+_{ref}:

$$\theta^+_{ref} = \frac{\theta}{\theta_{ref}} \tag{14}$$

Note, that for illustrative purposes it might be useful to take the logarithm of eqn. 14, since then all terms are additive and can be plotted on the same scale:

$$\begin{aligned} \ln \theta^+_{ref} &= (\ln p(f,o) - \ln p(f_{ref},o)) + (\ln p(\bar{f},\bar{o}) - \ln p(\bar{f}_{ref},\bar{o})) \\ &- (\ln p(f,\bar{o}) - \ln p(f_{ref},\bar{o})) - (\ln p(\bar{f},o) - \ln p(\bar{f}_{ref},o)) \end{aligned} \tag{15}$$

Furthermore, this split of the odds ratio benefit can be helpful in getting to the source of forecast differences because one might have a physical idea why one forecast has for instance produced more hits. It is more difficult to find out why an overall skill score is better or worse, since it is the balance between the entries in the contingency table which make up the total score. In the log odds all contributions have the same weight and thus can be compared, whereas in most other scores (e.g. ETS, eqn. (3)) the contributions are nonlinearly combined and thus their balance is not so clear.

3 Verification of operational precipitation forecasts

Eight years (1995-2002) of 4 times daily, six-hourly accumulated precipitation forecasts from the operational Met Office (U.K.) mesoscale model are verified against 42 rain gauge accumulations in the U.K.. Thus the daily cycle is sampled evenly. Because of representativity problems gauges are not optimal for estimating areal precipitation required for the verification of model grid box average precipitation, especially for single cases or in convective situations. Yet here we also want to compare monthly verification results over

a long time period. Thus it is more important that the selected set of stations is fairly well distributed over the U.K. and that it was available most of the time. Model forecasts are compared against six-hourly observed accumulations persisted for 24 hours.

Monthly and overall contingency tables are formed by computing the relative number of joint forecast and observed events and non-events (Table (1)). An event is defined by the exceedance of a threshold of accumulation amount, e.g. $0.1, 0.2, 0.5, 1, 2, 4, 6, 10$ mm per 6 hours in our analysis.

a Long term means

We start by analysing the contingency table for the whole eight year period. Figs. 2a-f present an illustration with operational data of the derivation of the odds ratio in section 2. The product of the first two plots in each column gives the third plot (cf. eqns. 10 and 12). The posterior odds for a correct precipitation forecast are on the bottom of the left column(Fig. 2c). Here the odds for getting the precipitation forecast right are about even for slight precipitation (0.1 mm/6h) and they drop with increasing amounts such that the forecast for heavy precipitation (10 mm/6h) is more likely to be a false alarm than a correctly detected event. Yet we see that the source of this drop in perceived quality of the forecast is solely due to the sharp drop in the prior odds for precipitation events with large amounts (Fig. 2a), i.e. the (few) precipitation forecasts go wrong relatively often since there are only a few occasions to be right anyway. Indeed, the likelihood ratio (Fig. 2b and eqn. 10) reveals that the (largely unbiased) model has more skill when it forecasts precipitation in distinguishing between events and non-events for heavy precipitation than for slight precipitation. The model raises the odds for observing slight precipitation from about 1:3 prior to the forecast to 2:1 when the forecast is for precipitation. For heavy

precipitation, the increase in the odds is even higher, i.e. from 1:125 prior to 2:5 after the forecast. Note also that the benefit from using model forecasts instead of persistence forecasts also increases with precipitation amount.

In other words, when one removes the effect of the low prior odds given by nature to get the precipitation forecast right then we see that the model actually has more skill at predicting heavy precipitation than slight precipitation. We suspect the reason for this behaviour lies in clearer precursors for a more extreme event than for slight precipitation. For instance, the initial state may consist of an already developed strong low which is well observed or of high available potential energy or high convective available potential energy, i.e. most extreme events do not happen "out of the blue" while drizzle forecasts have less obvious precursors. At least from the perspective of seasonal variations, continuous skill scores also show that more extreme winter weather forecasts have more skill than summer forecasts (Fig. 1). Van Den Dool and Toth (1991) cite numerous observations particularly from long range forecasting that forecasts of "near-normal" situations often fail.

The no-precipitation forecasts (right column of Fig. 2) have a prior odds which is just the inverse of the prior odds for precipitation (Fig. 2d). Thus it is no surprise that the posterior odds for a correct no-precipitation forecast are high and strongly increase with amount (Fig. 2f). In this situation many people know intuitively that the forecaster should not get the credit (e.g. by saying "no wonder that you get the no-rain forecast right in Southern California, because it never rains in Southern California"[3]). Yet for this no-precipitation part of the forecasting problem the skill of the model is not as high as for the precipitation forecast and it is slightly decreasing with amount (Fig. 2e), i.e. the

[3]In regions of a high base rate, i.e. of more wet than dry events, people rightly use the opposite argument, e.g. when they say "no wonder that you get the rain forecast right in Scotland, because it always rains in Scotland".

miss rate rises slightly stronger with amount than the correct rejection rate increases. The model raises the odds for observing no-precipitation from about 3:1 for slight precipitation prior to the forecast to 8:1 after the forecast was for no-precipitation and for no-heavy precipitation from 125:1 prior to 190:1 after the forecast.

Figure 3 shows that combining the skill for precipitation and no-precipitation forecasts using the odds ratio (eqn. 13) gives a clear increase in skill for more extreme events and relative to persistence forecasts. The odds for making a hit are 10 times larger than the odds for making a false alarm for slight precipitation and 60 times larger for heavy precipitation. Furthermore, the odds ratio is 5 times larger than for persistence forecasts for slight precipitation and it is 60 times larger for heavy precipitation. The increase in skill for more extreme events is generally opposite to what many other traditional categorical "skill" scores indicate (e.g. Ebert *et al* (2003)). However, the majority of these traditional scores are explicitely dependent on the base rate of the event, i.e. the base rate can not be factored out and thus the scores decrease as the base rate decreases (see chapter 3 in Jolliffe and Stephenson (2003)).

Fig. 4 presents an illustration of eqn. (15). We see that the major difference between model and persistence forecasts comes from a higher probability for a hit $p(f, o)$ achieved by the model and the difference also increases with amount. This underlines the assumption that with increasing severity of the event there are stronger precursors in the atmosphere which make the occurence of an extreme event less of a 'surprise' for the model compared to a persistence forecast.

It should be noted that there is hardly any difference between model and persistence forecasts for the successful forecast of the non-event ($p(\bar{f}, \bar{o})$). While the absolute change in this cell of the contingency table is large, it is the relative change which is important

in the odds ratio. However, this relative change is usually small and so is its influence on the change in odds ratio with threshold or in a comparison of models, i.e. the change in odds ratio is hardly influenced by the no-forecast, no-observation cell counts.

The model is doing slightly better than persistence in terms of false alarms ($p(f, \bar{o})$) and specifically in reducing the number of misses ($p(\bar{f}, o)$). Having more false alarms than missed events also shows that there is a positive bias in the model forecasts whereas the persistence forecasts are unbiased by construction.

b Time evolution of forecast skill

Having shown a fair comparison of forecast skill for different events we now turn to the comparison of forecast skill over time. The model formulation has been changed over the 8 years to improve temperature and wind forecasts. Here we look to see if there is also improvement in precipitation forecasting. Fig. 5 displays the time evolution of monthly values of the traditionally used ETS. We might be satistfied to identify an annual cycle as an interpretable feature (predominantly convective precipitation in summer is less predictable than large scale precipitation in winter), but this does not really tell us anything new about the model performance but just that we should not really compare seasonally varying performance in order to find out about the model performance. We note an increase in performance over the time period which is small compared to the variability.

Fig. 6 shows that the variability of the prior odds for slight precipitation (>0.2mm/6h) is larger than the variability of the likelihood ratios and the odds ratio which represent the model performance. Furthermore, the prior odds show an annual cycle whereas the model measures do not. Thus the "perception" of the variability of the accuracy of the forecast

as represented by the posterior odds would be strongly influenced by the variability of the base-rate (not shown). The variability of the likelihood ratio for the no-precipitation forecast is smallest and thus the variability of the odds ratio is mainly influenced by the variability of the model skill to forecast the event. No significant trends can be detected.

The time evolution of the model and persistence forecast skill as measured by the odds ratio and the ORB is given in Fig. 7 for heavy precipitation (>4mm/6h). No significant improvement can be seen for model forecasts yet the skill of persistence forecasts seems to have decreased slightly over this time period. Thus the ORB shows a small increase. To summarise, from all the measures presented here no big increase in skill of precipitation forecast can be detected over the last eight years, an observation which has been made world wide and is the subject of much research (Ebert *et al* (2003)).

However, a clear improvement in the model bias can be seen in Fig. 8. The frequency bias for slight precipitation events has decreased substantially from about 30% over-estimation to 10% over-estimation. This mesoscale model no longer produces predominantly large areas of slight precipitation but more realistic, concentrated areas of higher precipitation amounts. Yet the accurate placement of these features remains a problem.

4 Concluding remarks

It is a challenging task to fairly split forecast accuracy into a part which is due to the skill of the forecasting system and a noise part which is due to the natural variability[4] and predictability of the weather. It is only possible to fairly compare forecasts from

[4]The observations of the weather contain noise as well, which is especially important in precipitation estimation, a topic which is dealt with elsewhere in this special issue.

different times, regions, events etc. when we separate out the skill of the forecasting system. Skill scores for continuous variables largely achieve this goal but categorical skill scores traditionally used in meteorology fail to do so, because they depend explicitly on the base rate of the event. In order to play fair in the comparison of categorical forecasting systems, we propose the use of the odds ratio (benefit) which:

- is independent from the base rate of the event;
- has a simple interpretation as the ratio of the odds for making a hit to the odds for making a false alarm;
- enables an interpretable comparison of different forecasting systems;
- permits a split of the forecast accuracy into accuracy of forecasting the event and the non-event, which can be a useful diagnostic tool for forecasters, users and model developers.

One major result of this study has been that on average more extreme event precipitation forecasts have more skill than "normal event" forecasts, provided the rarity of this kind of event is properly taken into account as done in the odds ratio. We suspect this may be due to clearer precursor signals in the initial state. It would be interesting to similarly de-compose continuous skill scores (eqn. 1) into contributions from more extreme and "normal" cases using the amplitude of the anomaly as an indicator.

Given that more extreme precipitation events are more skillfully predicted, we question the wisdom of a widespread habit for model trials composed of a majority of extreme events. At other times the selection is for those less predictable extreme events which lead to major operational forecast failures. Overemphasis on extreme cases, predictable or less predictable, should be avoided and trials should consist of a proper selection of all kinds

of cases, ideally as frequent in the sample as observed climatologically. An obviously silly example would be to tune a model only on cases when it rained, thus probably degrading the performance in no-rain cases. Careful trial composition is especially an issue for mesoscale models, where the sample of different weather situations is small, whereas over the globe there is usually already a good distribution of different situations. Recently, Doswell *et al* (2002) discussed the underemphasis in the literature on the most frequent case in operational forecasting, i.e. the forecast of the non-occurence of extreme weather. Another application of the concepts presented here is now opening up in the verification of the spatial aspects of precipitation forecasts (Casati *et al* (2003)).

Extreme events are often treated as a special forecasting task and there are separate groups or even organisations who just issue *warnings* for extreme events but no forecasts for "normal" events. Many publications deal with the specific problem of how to verify warnings because they represent a degenerate case of the contingency table in that the no-forecast, no-observation element $p(\bar{f}, \bar{o})$ does not exist (e.g. Mason (1989), Schaefer (1990), Brown *et al* (1997)). This case of "posterior censoring" seems a bit odd because it is as if the forecasting task is changed after the forecast: before the forecast our eyes are shut and we pretend that there is no threat (thus no warning), but after the event our eyes are opened and it is sometimes realised that there was an extreme event and now this is counted as a miss. Furthermore, the forecast of no threat, i.e. no warning, can be very important for some users. A way out of this muddled situation is offered by the use of the odds ratio benefit (14). Here the normal 2x2 contingency table is used which now contains a high probability $p(\bar{f}, \bar{o})$ for both the forecast and the reference forecast. However, it is their ratio which enters the ORB which will be very close to one and thus does not influence the skill measure. The no-forecast, no-event element can

be measured by defining a regular sampling interval and then count the contributions to the contingency table as usual. Thus the ORB is a measure which is hardly influenced by the "easy" cases of no-forecast, no-observation. However, the verification of warnings still represents a particular challenge because the observations might be biased towards observing the event, e.g. pilot reports of in-flight icing (Brown *et al* (1997)).

The Bayesian approach to categorical forecast verification applied in this study complements the traditional quest for the answer to 'how good are the forecasts' by measuring 'how well do we forecast'.

Acknowledgments

The authors wish to thank Barbara Casati and Glen Harris for discussions during the course of this work and the reviewer Beth Ebert for suggesting very helpful clarifications in the article.

List of tables

List of figures

and their linear trends.

5. Time series of prior odds for precipitation (dash-dot-dot-dot line), likelihood ratio for precipitation forecast (solid line), likelihood ratio for no-precipitation forecast (dotted line) and odds ratio (dashed line) for slight precipitation (>0.2mm/6h) forecast. Note the logarithmic scale which allows a visual comparison of the variabilities of the measures (see eqn. 15).

6. Odds ratio for model forecasts (dashed line) and persistence forecasts (dotted line) and odds ratio benefit (eqn. 14, solid line) for heavy precipitation (>4mm/6h) for 12-18 hour forecast range accumulation model forecasts and their linear trends.

7. Frequency bias FB for 12-18 hour accumulation model forecasts of slight precipitation (>0.2mm/6h, solid line) and heavy precipitation (>4mm/6h, dashed lines) and their linear trends.

	Event observed		
Forecast	Yes	No	Total
Yes	$p(f,o)$	$p(f,\bar{o})$	$p(f)$
No	$p(\bar{f},o)$	$p(\bar{f},\bar{o})$	$p(\bar{f})$
Total	$p(o)$	$p(\bar{o})$	1

Table 1: Joint probability distribution $p(.,.)$ of forecasts f and observations o for a binary event as well as marginal distributions $p(.)$. The non-event is denoted by an overbar $\bar{.}$. The joint distribution can be estimated from the relative frequencies of the occurence of the joint events. A conditional probability is denoted by, e.g. $p(f|o)$, which reads "probability for f *given* o".

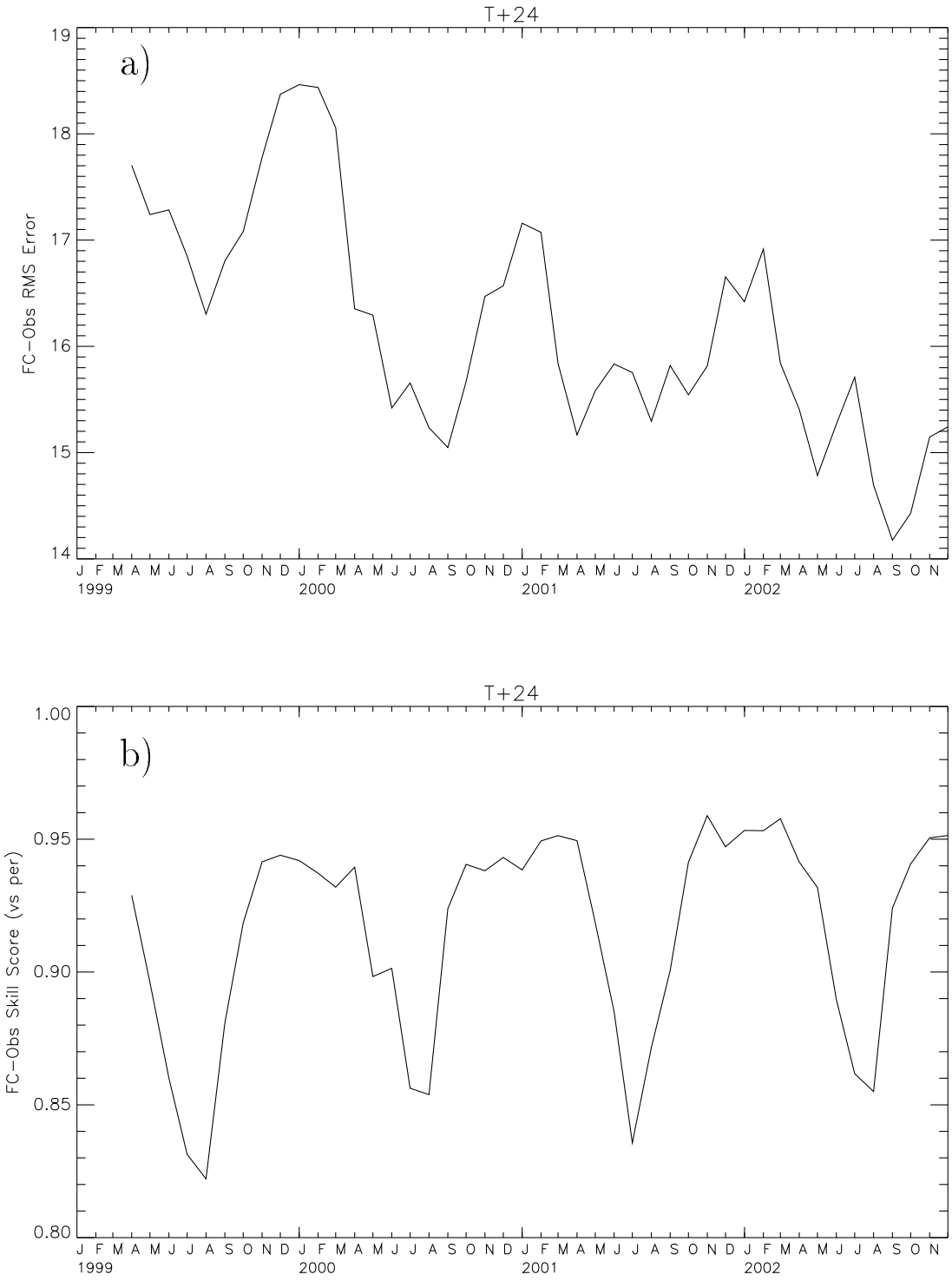

Figure 1: a) Monthly root mean square error ($RMSE$) of Northern hemispheric 500 hPa geopotential height T+24 hour forecasts from the Met Office Global Model; b) skill score (eqn. (1)).

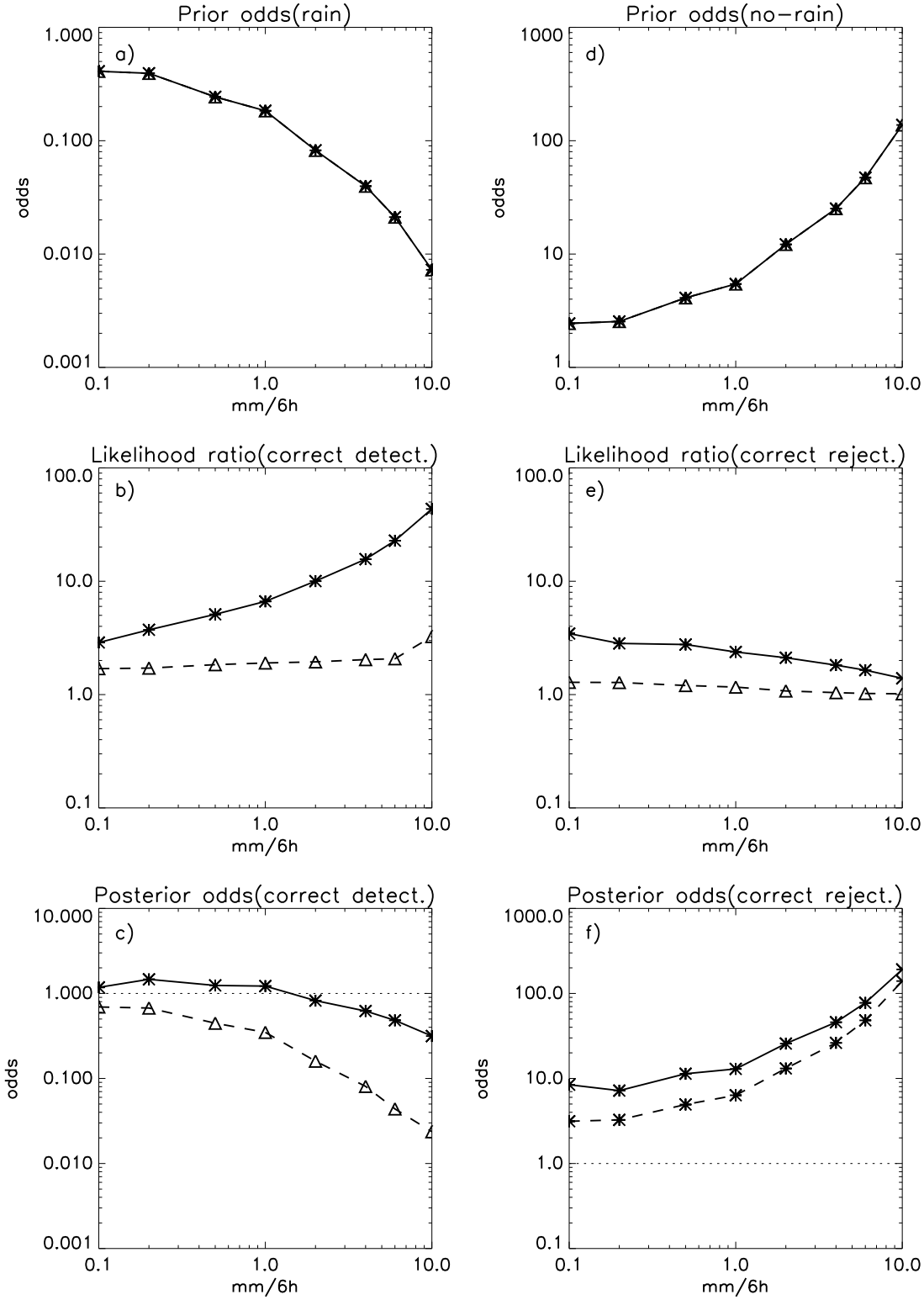

Figure 2: Contributions to the odds ratio as a function of accumulated precipitation threshold. Left column: forecast correctly detected precipitation amount; right column: forecast correctly rejected precipitation amount. 1st row: prior odds; 2nd row: likelihood ratios; 3rd row: posterior odds. 12-18 hour forecast range accumulation model forecasts (stars and solid line) and persistence forecasts (triangles and dashed line). Dotted line for odds equal one ("evens") . Prior odds are the same for model and persistence forecasts, because they reflect the base rate of the same observations.

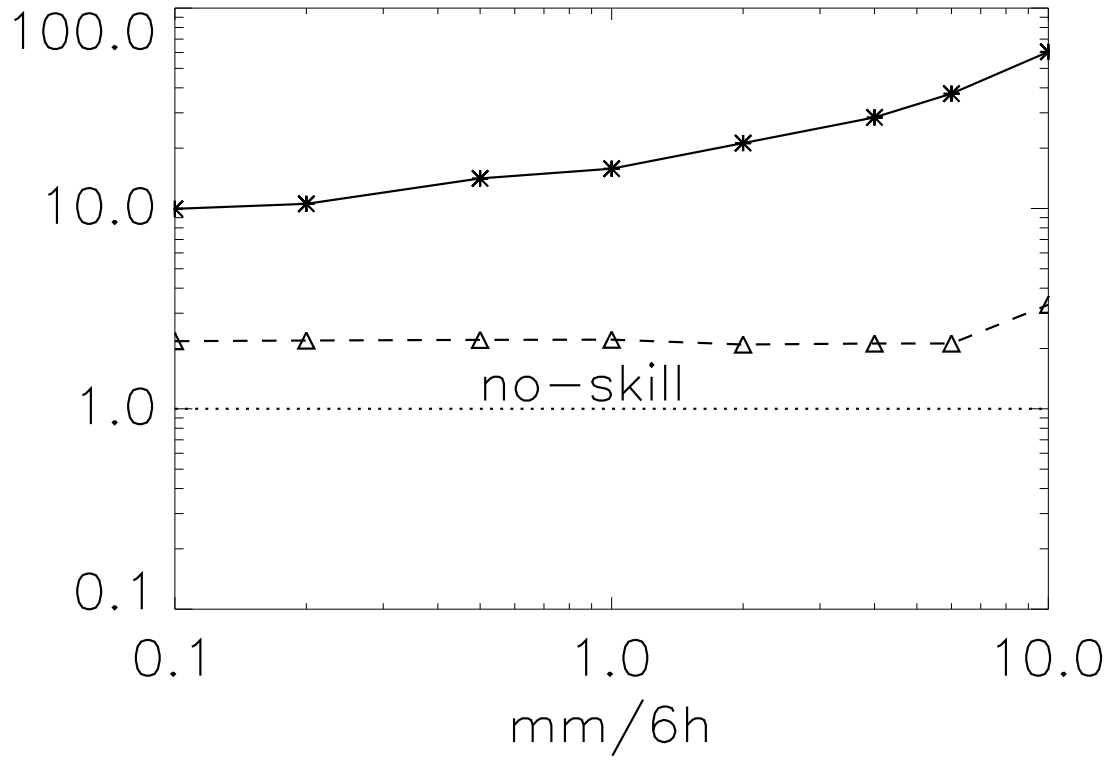

Figure 3: Odds ratio as a function of threshold of accumulation. 12-18 hour forecast range accumulation model forecasts (stars and solid line) and persistence forecasts (triangles and dashed line). Dotted line for an odds ratio of one, i.e. no-skill.

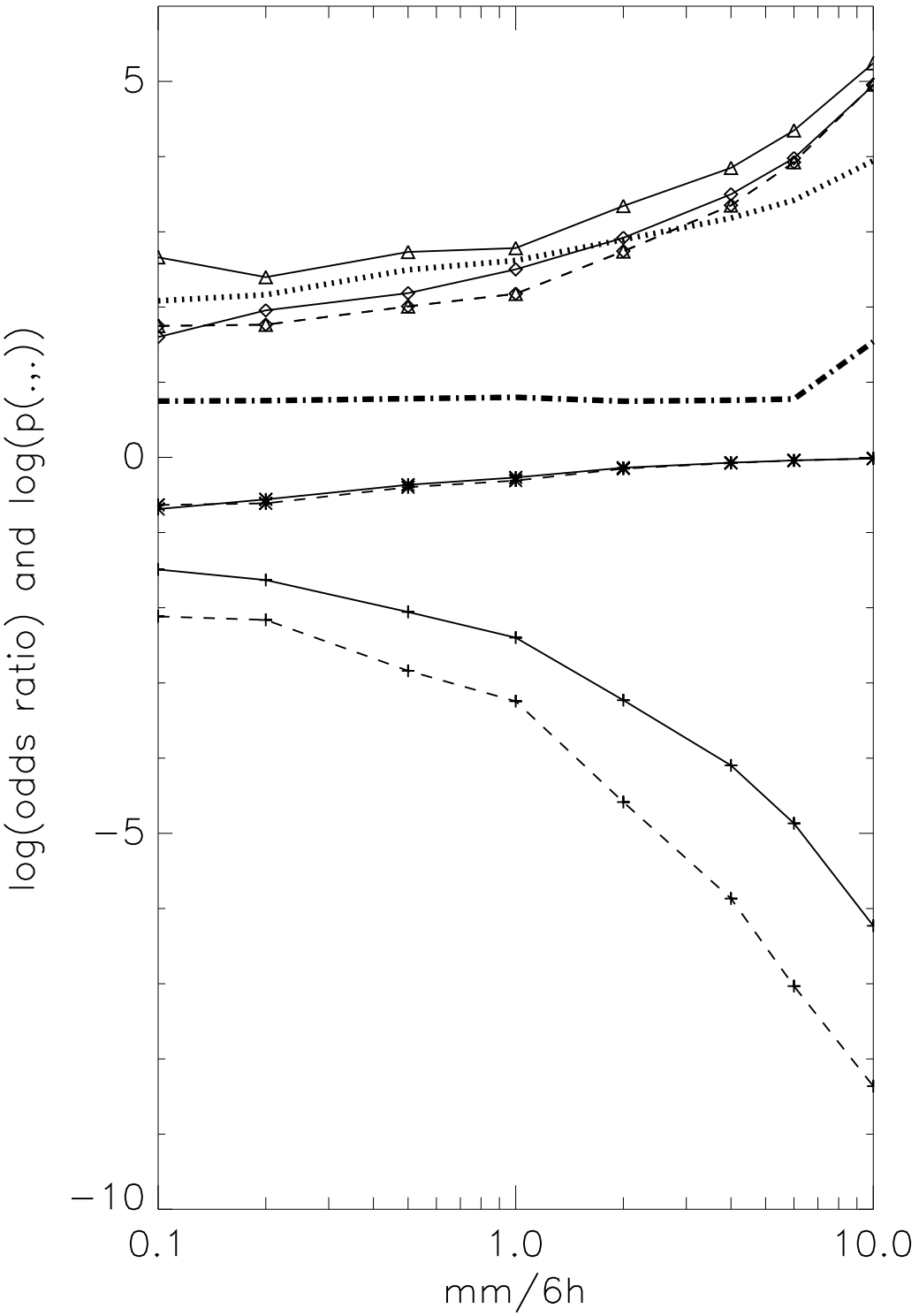

Figure 4: Contributions to Eqn. (15). Bold dotted line: model $\ln\theta$, bold dash-dotted line: persistence $\ln\theta_{pers}$. RHS of eqn. (15): model forecasts (solid lines connecting symbols); persistence (dashed lines connecting symbols). Symbols: $+\ \ln p(f,o)$; $*\ln p(\bar{f},\bar{o})$; $\diamond -(\ln p(f,\bar{o})$; $\triangle -(\ln p(\bar{f},o))$. Note, that false alarms and misses have been plotted with their negative sign from Eqn. (15), allowing to mentally add all curves from the RHS of Eqn. (15) to arrive at the LHS.

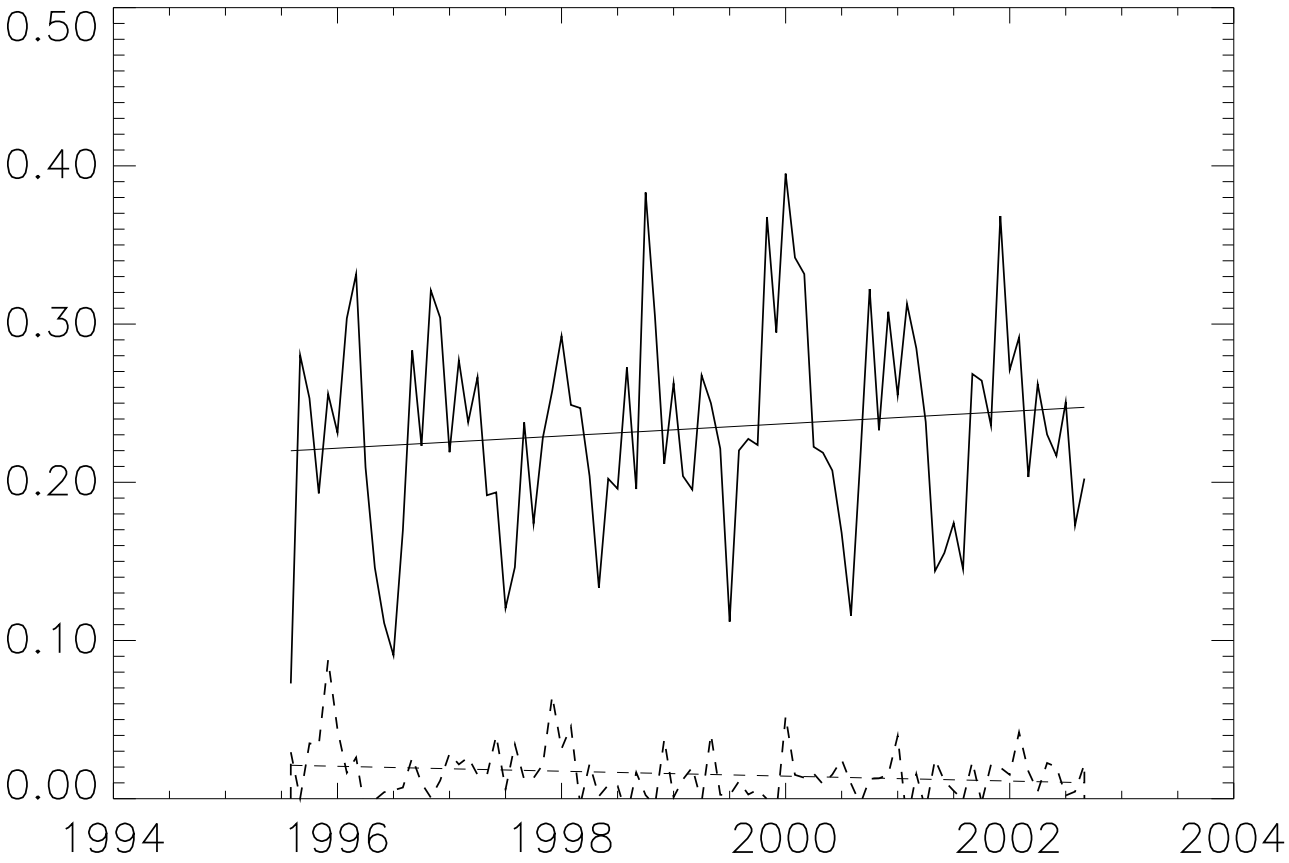

Figure 5: Equitable Threat Score (ETS) for heavy precipitation (>4mm/6h) for 12-18 hour accumulation model forecasts (solid line) and persistence forecasts (dashed lines) and their linear trends.

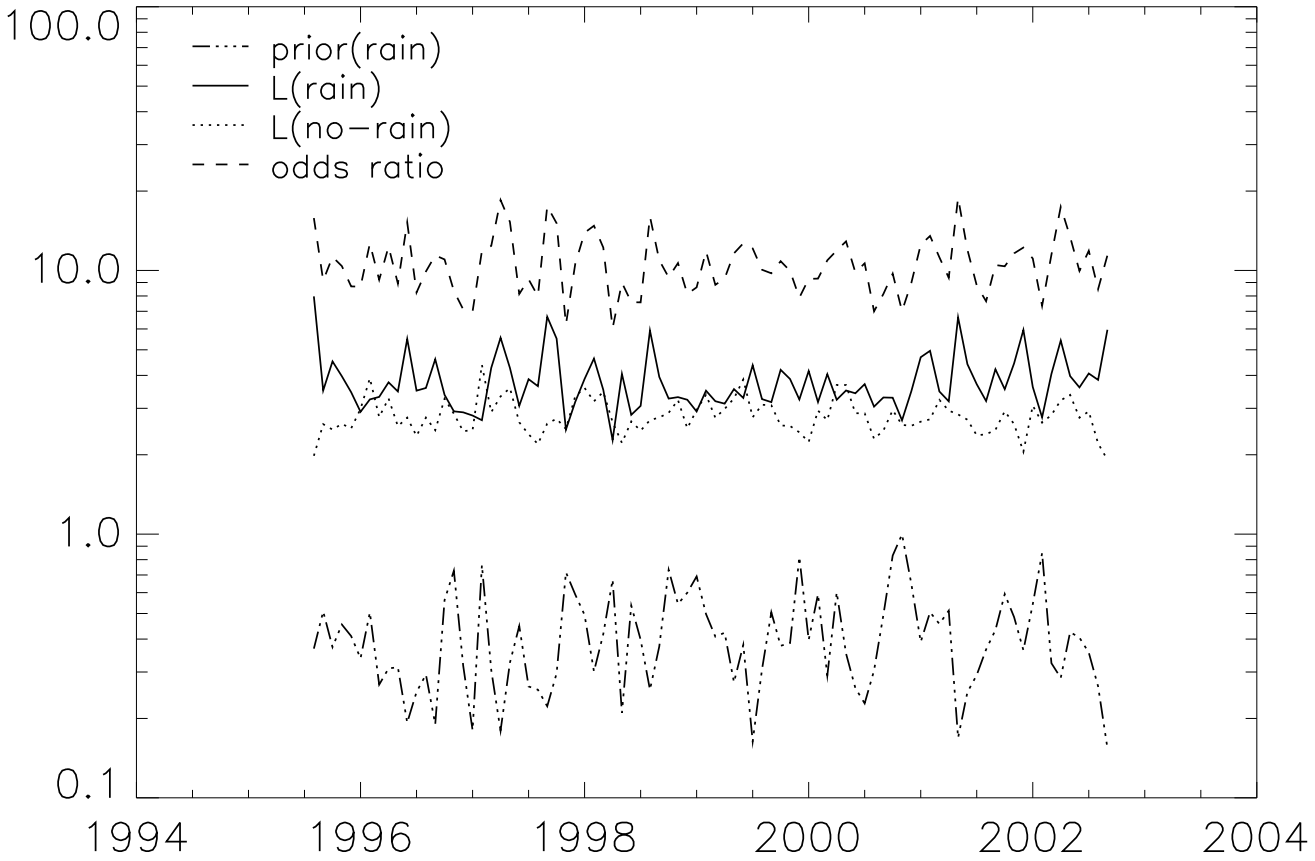

Figure 6: Time series of prior odds for precipitation (dash-dot-dot-dot line), likelihood ratio for precipitation forecast (solid line), likelihood ratio for no-precipitation forecast (dotted line) and odds ratio (dashed line) for slight precipitation (>0.2mm/6h) forecast. Note the logarithmic scale which allows a visual comparison of the variabilities of the measures (see eqn. 15).

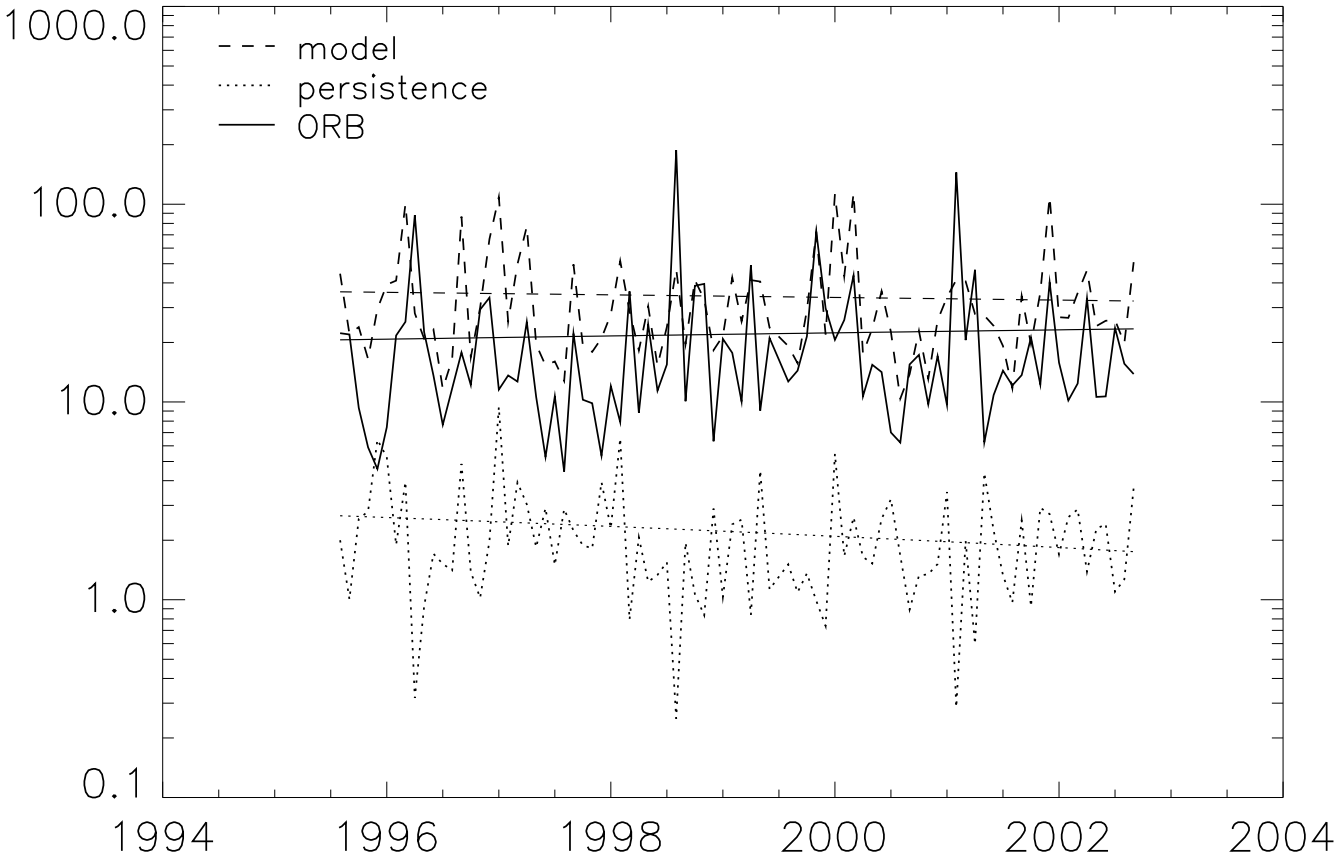

Figure 7: Odds ratio for model forecasts (dashed line) and persistence forecasts (dotted line) and odds ratio benefit (eqn. 14, solid line) for heavy precipitation (>4mm/6h) for 12-18 hour forecast range accumulation model forecasts and their linear trends.

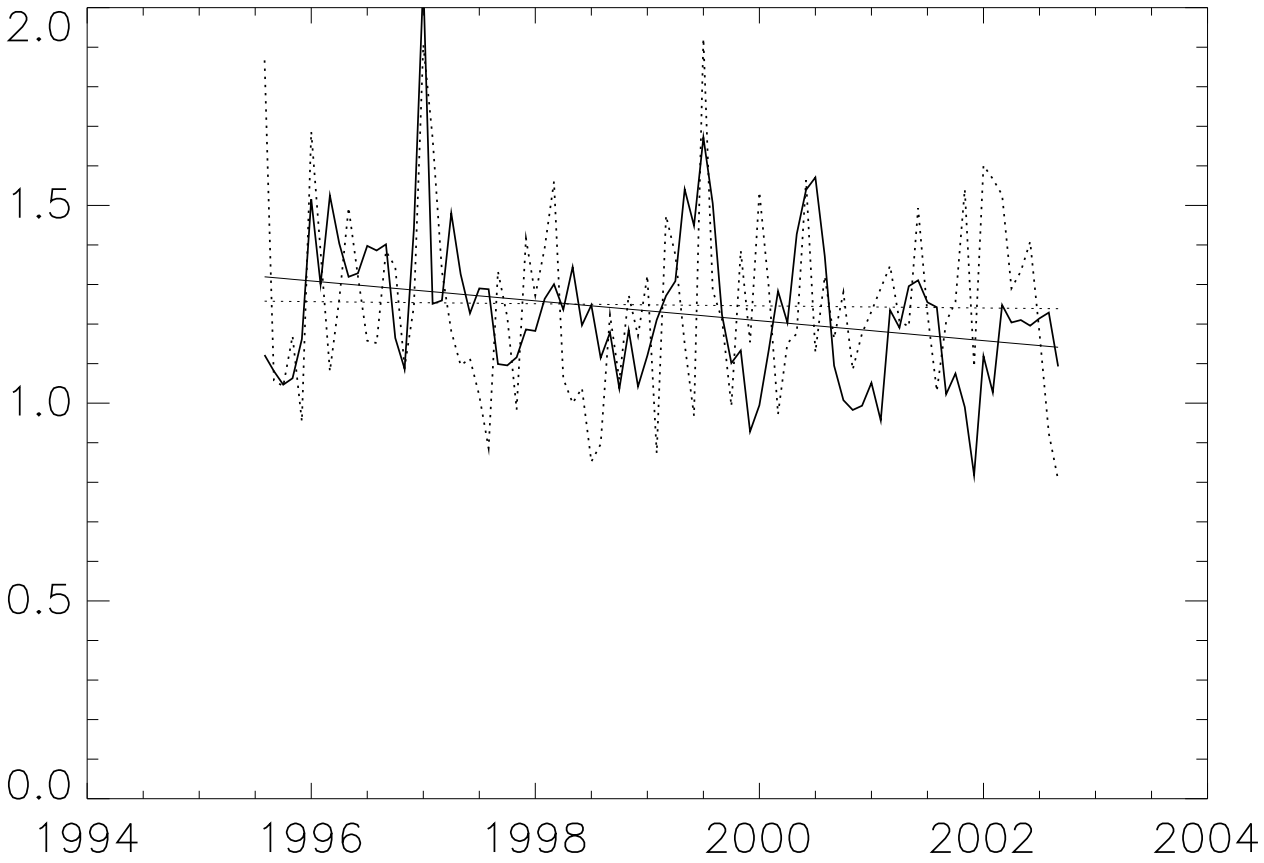

Figure 8: Frequency bias FB for 12-18 hour accumulation model forecasts of slight precipitation (>0.2mm/6h, solid line) and heavy precipitation (>4mm/6h, dashed lines) and their linear trends.

References

Agresti, A., 1996: *An Introduction to Categorical Data Analysis. Number 290 pp. John Wiley and Sons, 290 pp.*

Brooks, H. E., and C. A. Doswell III, 1996: A comparison of measures-oriented and distributions-oriented approaches to forecast verification. *Wea. Forecasting*, **11**, 288–303.

Brown, B. G., G. Thompson, R. T. Bruintjes, R. Bullock, and T. Kane, 1997: Intercomparison of in-flight icing algorithms. part ii. statistical verification results. *Wea. Forecasting*, **12**, 890–914.

Casati, B., G. Ross, and D. B. Stephenson, 2003: A new intensity-scale approach for the verification of spatial precipitation forecasts. *To be submitted to Meteorol. Appl.*

Doswell, C. A., D. V. Baker, and C. A. Liles, 2002: Recognition of negative mesoscale factors for severe-weather potential: a case study. *Wea. Forecasting*, **17**, 937–954.

Ebert, E. E., U. Damrath, W. Wergen, and M. E. Baldwin, 2003: The WGNE Assessment of Short-term Quantitative Precipitation Forecasts. *Bull. Am. Meteorol. Soc.*, **84**, 481–492.

Finley, J. P., 1884: Tornado predictions. *Amer. Meteor. J.*, **1**, 85–88.

Gigerenzer, G., 2002: *Reckoning with risk: Learning to live with uncertainty. Allen Lane the Penguin Press, 320 pp.*

Jolliffe, I., and D. Stephenson, 2003: *Forecast Verification: A Practioner's Guide in Atmospheric Science. John Wiley and Sons, Ltd., 240 pages.*

Krzysztofowicz, R., 1992: Bayesian correlation score: a utilitarian measure of forecast skill. *Mon. Wea. Rev.*, **120**, 208–219.

Marzban, C., 1998: Scalar measures of performance in rare-event situations. *Wea. Forecasting*, **13**, 753–763.

Mason, I., 1982: A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291–303.

——, 1989: Dependence of the critical success index on sample climate and threshold probability. *Aust. Met. Mag.*, **37**, 75–81.

Matthews, R. A. J., 1996a: Base-rate errors and rain forecasts. *Nature*, **382**, 766.

——, 1996b: Why are weather forecasts still under a cloud? *Mathematics Today*, **32**, 168–170.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.*, **115**, 1330–1338.

——, 1988: Skill scores based on the mean square error and their relationships to the correlation coefficient. *Mon. Wea. Rev.*, **116**, 2417–2424.

——, 1991: Probabilities, odds, and forecasts of rare events. *Wea. Forecasting*, **6**, 302–307.

——, 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. *Wea. Forecasting*, **8**, 281–293.

Plous, S., 1993: *Psychology of Judgement and Decision Making.* McGraw-Hill Education, 250 pp.

Schaefer, J. T., 1990: The critical success index as an indicator of warning skill. *Wea. Forecasting*, **5**, 570–575.

Stephenson, D. B., 2000: Use of the 'odds ratio' for diagnosing forecast skill. *Wea. Forecasting*, **15**, 221–232.

Thornes, J. E., and D. B. Stephenson, 2001: How to judge the quality and value of weather forecast products. *Meteorol. Appl.*, **8**, 307–314.

Van Den Dool, H. M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting*, **6**, 76–85.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences*. Academic Press, 467 pp.