Scientific report; WR-2000-04

KNMI

On the behaviour of a few popular verification scores in yes/no forecasting

C.J. Kok

De Bilt, 2000

Scientific report = Wetenschappelijk rapport; WR 2000-04

De Bilt, 2000

PO Box 201
3730 AE De Bilt
Wilhelminalaan 10
Telephone +31 30 220 69 11
Telefax +31 30 221 04 07

Author: C.J. Kok

UDC: 551.509.3
551.509.51

ISSN: 0169-1651

ISBN: 90-369-2178-3

On the behaviour of a few popular verification scores in yes/no forecasting

C. J. Kok

Contents

1. Introduction

The main objective of verification in meteorology is to assess and quantify the quality of the forecasts. It offers a means to compare results of different models, methods or forecasters. In addition, their performance over different periods or years can be established in order to assess whether there is a general improvement or not. If, for instance, the results of two models or of a number of forecasters are compared for the same location and time period then it is fairly easy to judge which one performs best in terms of a given verification score. (This is by no means the same as telling which forecasting system is the best one.) If, on the other hand, the verification is done for different times or stations then it is extremely important to be aware of the fact that the behaviour of all scores that are used in meteorological verification is highly dependent on the statistical properties of the predicted quantity at hand.

In general three different types of forecasts can be identified: deterministic, probabilistic and categorical ones. A categorical forecast is a forecast in which a statement is made that one and only one of a set of possible events will occur. In probabilistic forecasting probabilities are assigned to the event(s). This report focuses entirely on the issue of verification of categorical forecasts in which the predictand is binary and also the forecasts are issued in binary form. The results can therefore be summarised in a 2x2 contingency table. One can think of yes/no forecasts of, for instance, extreme events like tornadoes, but also of the occurrence of fog, thunderstorms or precipitation.

In verifying forecasting systems it is often desirable to focus on a particular part of the spectrum of possible occurrences of the phenomenon to be predicted. For instance, to establish the skill or accuracy of a forecasting system in predicting the occurrence or non-occurrence of tornadoes it is less important what the percentage of correct forecasts is, but much more important is the percentage of correctly forecast tornadoes. To be able to take into account or to emphasize special characteristics of the meteorological quantity there is a large number of verification scores in use. In this report we give an overview of some of the most frequently used scores which are used in the verification of categorical forecasts. Only the two-category situation is discussed, but many of the scores can also be defined for more categories.

Usually not one single verification measure is applied but instead a few scores are used simultaneously. This is done because for many scores it is fairly easy to adapt the forecasting system in such a way that better scoring results are obtained. For instance, some scores award overforecasting. In fact, in using the percentage correct as verification measure in rare event forecasting, always predicting the event to occur may turn out to be the best strategy to apply. But this is obviously not considered to be a good forecasting system. A good score or scoring system should penalise strategies like this and should encourage the forecaster to make forecasts according to his / her true beliefs. This is by no means the case for all scores that are frequently used.

Sometimes a predetermined target is set which the forecaster has to meet. For the above-mentioned reason this is usually done for more than one score at the same time. For instance, not only the percentage of correct forecasts should be higher than a certain value, but also the total number of false alarms should not be too high. This procedure was followed for instance in the case of forecasting the minimum road temperature in the British Isles in the beginning of the 1990s. An extensive verification is given by Halsey (1995). Obviously, the forecaster should not "miss" too many frost occurrences. On the other hand, because of costly salting of roads erroneously forecasting below freezing temperatures should be avoided as much as possible. In Fig. 1 the verification of the forecasts in the British Isles is shown in terms

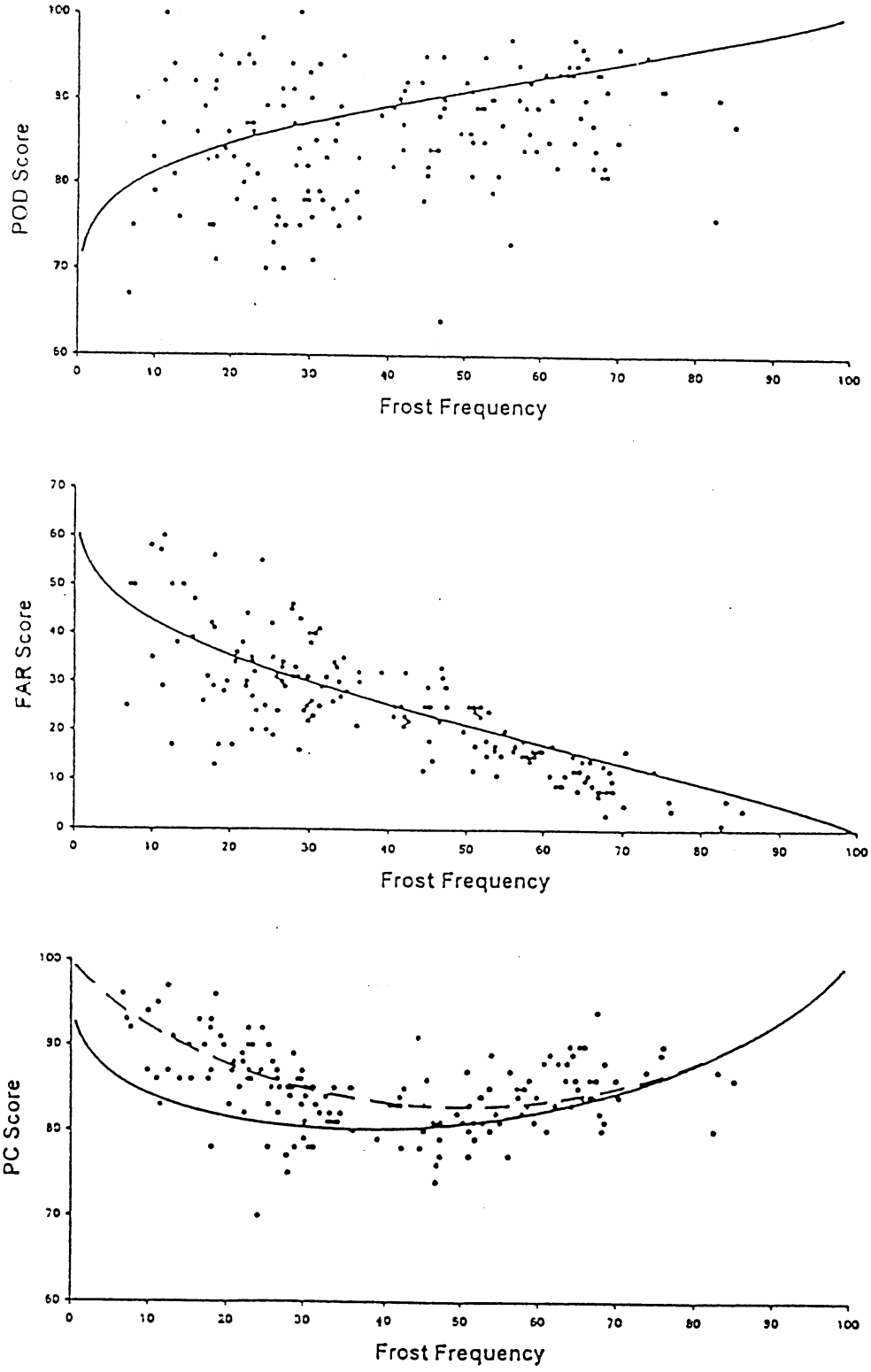


Fig.1. (taken from Halsey, 1995). Actual verification scores obtained by Met. Office forecasters across the British Isles for the winter of 1993/'94. Shown are the Probability of Detection (POD), False Alarm Ratio (FAR) and Percentage Correct (PC) as a function of observed frost frequency. Also plotted are the theoretical target values.

of three scores: the probability of detection, the false alarm ratio and the percentage correct. These scores will be discussed in the next chapter. Theoretical target values are also given in Fig. 1. The targets as well as the results appeared to be strongly dependent on the observed frost frequency. Apparently it is much more difficult to obtain a small false alarm ratio for rare events of frost than it is when frost occurs more frequently. The opposite is true for the probability of detection. The percentage correct shows a parabolic dependence on the sample climatology. The nature of these dependences upon the observed frequency is one of the main topics of this report. The paper by Halsey motivated the writing of this report.

In chapter II an overview of a number of popular verification scores in yes/no forecasting is given together with their general characteristics. Also some of their main differences are highlighted. It is not the intention of this report to provide guidelines as to what score should be preferred under different circumstances. Only some general considerations are discussed which may help the reader to determine which one(s) to use. Already in the case of two categories the scores can have many different appearances. Therefore many of the mathematical manipulations in that chapter (as well as in the following ones) are given very extensively.

In chapter III the behaviour of some of the scores discussed in chapter II is shown when certain simple assumptions are made. These assumptions or constraints include randomness of the forecasts and also a fixed skill with respect to climatological forecasts. In these and other "models" discussed in this chapter the scores are presented as a function of the sample climatology. In this way this may offer a qualitative assessment of differences in the behaviour of the scores and may help to interpret the differences in results (in terms of the chosen scores) for different operational forecasts or forecasting systems for different locations or time periods. It also offers a framework for the tightness of the targets that can be set for forecasting systems for different locations (i.e. climatologies) and time periods.

Categorical forecasts contain no information about the forecaster's uncertainty associated with the occurrence of the event. No distinction is made between cases in which the forecaster is absolutely sure about a particular category (or event) and when he / she is merely guessing. One way of incorporating this uncertainty is to make the decision criterion dependent on the level of uncertainty. Choosing different thresholds will result in different contingency tables. An elegant way of verifying these categorical forecasts for various thresholds is the relative operating characteristic (ROC). A brief discussion on this subject is given in chapter IV.

Finally, in chapter V. some concluding remarks are given.

II.1 Verification scores

In this section a number of frequently used verification measures for categorical forecasts are summarised. The definitions of the scores are given in terms of the cells of a 2x2 contingency table of forecast (FC) versus observed (OBS) categories as defined below. In this report the forecast and observed categories "yes" and "no" are mostly given in terms of relative frequencies (or probabilities) and not in number of occurrences. The relative frequencies in cells a, b, c and d therefore add up to one. In other papers these are usually referred to as p_{11} , p_{12} , p_{21} and p_{22} respectively. This so-called joint distribution of forecasts and observations is represented by an, in this case, 2x2 *performance matrix* (Murphy and Winkler, 1987). The relative frequency in which the event was forecast is defined as p_f , the observed frequency as p_o . These last two frequencies together with their complements are called the marginal distributions (or frequencies) of the forecasts and the observations respectively. Expressed in number of occurrences (simply obtained by multiplying all frequencies by the total sample size) they are called marginal totals or simply marginals; and in this case the contents of the cells are denoted in this paper in capitals. The respective contingency tables are given below.

		FC: yes	FC: no	
OBS	yes	a	b	p_o
OBS	no	c	d	$1 - p_o$
		p_f	$1 - p_f$	1

		FC: yes	FC: no	
OBS	yes	A hits	B misses	A+B
OBS	no	C false alarms	D correct rejections	C+D
		A+C	B+D	N

Although the scores that will be discussed in this report are formulated in terms of probabilities (or frequencies) most of them have the same appearance when expressed in number of

occurrences, since all the terms in the respective formulations have the same dimension. When this is not the case, as for instance with (one of the formulations of) the Heidke skill score or the Equitable Threat Score, we will give the expressions in absolute as well as in probabilistic formulation.

Many different scores have been proposed in the literature. A large number is given in e.g. Woodcock (1976), Mason (1979), Daan (1984), Murphy and Daan (1985), Flueck (1987), Stanski et al.(1989) and Marzban (1998). Verification scores which will be discussed in this report are given in section A of this paragraph, so-called skill scores in part B. A few other related scores which are not discussed at length in this report are briefly summarised in part C.

A. Scores

a) *Fraction Correct*

The *fraction correct* (FC), or the *percentage correct* (PC), is of course the sum of the diagonal elements of the contingency table. Sometimes it is also called the *Hit Rate* (HR). We will not use this name here to avoid misunderstanding since the hit rate with a completely different definition is much more frequently used in the context of the ROC score (see chapter IV).

$$FC = a + d$$

(In this paper FC is also used in the contingency tables to indicate the forecast category. We believe this will not lead to misunderstandings.) The *percentage correct* is, of course, the Fraction Correct times 100.

It was first used in a famous paper on the verification of tornado forecasts by Finley in 1884. The "serious fallacies" of using this score, as reported by a number of authors shortly afterwards, result from an important property of the Fraction Correct score and that is that it credits correct "yes" and "no" forecasts equally, and also penalises wrong forecasts for both categories by the same amount. In many cases this is not a desirable property, especially in rare event forecasting. Before the year was over a number of new verification scores had been proposed. Since then it is not used anymore as a sole verification measure.

b) *Probability of Detection* (POD)

This is defined as

$$POD = \frac{a}{a + b}$$

It is the fraction of the events that are correctly predicted (POD = a / p_o). Therefore it provides an estimate of the probability that an event will be forewarned. Expressed in probabilities:

$$P(\text{forecast} = \text{yes} \mid \text{observed} = \text{yes})$$

POD is also called *prefigurance* . In the context of the ROC score this quantity is called the *Hit Rate* (see chapter IV).

c) *False Alarm Ratio* (FAR)

The FAR is defined as the fraction of predicted events that did not occur:

$$FAR = \frac{c}{a + c}$$

or FAR = c / p_f . Expressed in probabilities:

$$P(\text{observed} = \text{no} \mid \text{forecast} = \text{yes})$$

It is sometimes also called the False Alarm *Rate*. But usually this last name is reserved for the fraction of wrong forecasts given that the event did not occur (i.e. c / (1-p_o)). See chapter IV for more information about this.

d) *Critical Success Index* (CSI)

The CSI , also called *Threat Score* (TS) , is defined as

$$CSI = \frac{a}{a + b + c}$$

It is the number of correct "yes" forecasts divided by the total number of occasions that the event was predicted and/or observed. It can be regarded as the fraction correct after removing the correct "no" forecasts. As opposed to Fraction Correct all correct forecasts are not equally weighted anymore. The independence of d does not imply that the correct "no" forecasts are not incorporated. This can be realised from the fact that c = (1 - p_o) - d , in which the observed frequency of the nonevents is fixed.

This score is frequently used in the verification of rare events. It was originally defined by Gilbert in 1884 in response to the use of the Percentage Correct by Finley (1884) in his verification study on tornado forecasts. It was rediscovered first by Palmer and Allen (1949), who called it the Threat Score, and also by Donaldson et al. (1975), who proposed the name Critical Success Index. Elaborate discussion of this score is given in Donaldson et al. (1975), Mason (1989), Doswell et al. (1990) and Schaefer (1990).

Its main disadvantage is that it is strongly dependent on the predicted frequency of the event. It appears that overforecasting is rewarded (see e.g. chapter III.1)
An alternative form in which the CSI can be encountered is

$$CSI = \frac{C}{F + O - C}, \quad \text{in which}$$

C is the number of correct forecasts of the event
(in the notation of this report: A)
F the number of forecasts of the event (A + C)
O the number of observed events (A + B)

e) *Bias (ratio)*

$$Bias = \frac{p_f}{p_o}$$

or in other words, the number of forecasts of occurrences per actual occurrence. It measures how well the model predicted the frequency of occurrence but it provides no information on the accuracy of the forecasts. If the bias is greater than 1 the events are being overforecasted, smaller than one underforecasting is occurring. When the bias ratio is equal to 1 there is no bias.

B. Skill Scores

It is often desirable to express the quality or skill of the forecasts in terms of the relative improvement over some set of reference forecasts. Two widely used reference systems are climatology and persistence (Daan, 1984; Murphy and Daan, 1985). But one can also use random guessing as reference or an older version of the forecasting system (model). The quantity that expresses the quality of the foreasts in terms of percentage improvement over the reference forecasts is called the *Skill Score* (SS).

$$SS = \frac{score_{fc} - score_{ref}}{score_{perfect} - score_{ref}} \times 100$$

Here *score* is a certain accuracy measure, e.g. *mean absolute error* (MAE), *mean square error* (MSE) , *fraction correct* (FC). The $score_{fc}$ is the score obtained in the verification of the forecast system under consideration. The $score_{ref}$ is the reference score and $score_{perfect}$ is the score that would be achieved by perfect forecasts. This will be zero if MAE or MSE is used as accuracy measure and is equal to one if FC is used.

The skill score is sometimes expressed as fractional improvement. Then the above definition applies without the factor 100. This is the formulation which is used in this report. The maximum value of SS is 1 in the case the forecasts are perfect. SS will be zero if the forecasts have the same score as the reference system. Its theoretical minimum value is minus infinity.

The skill score is a good relative quantity but its absolute value is very much dependent on the accuracy measure that is used. But the most important disadvantage is that it is difficult to compare the skill scores over different time periods. If, for instance, climatology is used as reference, then it is very hard to beat the reference forecasts in verification periods in which the weather was close to normal. Then the skill scores will be close to zero, regardless of the accuracy of the forecasts themselves.

In this section we will discuss four examples of skill scores which are frequently used in the verification of categorical forecasts. Most skill scores use percentage correct (or fraction correct) as accuracy score. Note that there is no fundamental difference between skill scores and the scores discussed under A of this paragraph, the only difference being that the former can be regarded as relative improvement over a particular reference forecast. This also means that not all skill scores can be expressed in the above form. The only requirement is that it yields a value of zero for a particular no-skill forecast (usually random forecasts) and a value of one if the forecasts are perfect.

f) *Hanssen-Kuipers Score* (HKS)

The original definition of the Hanssen-Kuipers Score (Hanssen and Kuipers, 1965), although the expression is not quite in agreement with the definition of skill scores as given in the beginning of section II.1.B., is as follows

$$HKS = \frac{FC_f - FC_r}{1 - FC_c}$$

in which

FC_f is the fraction correct of the forecasts
FC_r is the fraction correct under the assumption that there is no relation between forecasts and observations (subscript r stands for random)
FC_c is the fraction correct for unbiased climatological forecasts

However, usually this score is used in a slightly different way; the reference system is based on the sample climatology instead of the long-term climatology. This is also the way in which we shall use it throughout this report. The general form of the HKS written in probabilities looks like (in the case of yes/no forecasts):

$$HKS = \frac{a}{a+b} + \frac{d}{c+d} - 1$$

The first two terms are the correctly predicted fractions of the occurrences of the event and nonevent respectively.
The above expression is equal to:

$$\frac{a}{a+b} - \frac{c}{c+d}$$

In this latter form it is sometimes called the *likelihood difference* (LD). The second term in this expression is sometimes referred to as the *Probability of False Detection* (POFD); the first term is the probability of detection (POD).
Another frequently used form is :

$$HKS = \frac{ad - bc}{(a+b)(c+d)},$$

equivalent to

$$HKS = \frac{ad - bc}{p_o(1 - p_o)}$$

The numerator can be further evaluated using the marginal distributions:

$$ad - bc = a(1 - p_f - b) - b(p_f - a) =$$
$$= a - ap_f - bp_f = a - p_o p_f$$

This is equal to the value of cell a minus what can be expected for this cell for random forecasts given the observed frequency (see chapter III). It will be shown in the discussion of the correlation coefficient (and in Appendix A.1) that this is equal to the covariance between the forecasts and observations.
The result is a formulation of the HKS expressed in the frequency of hits and the marginal frequencies only (see also Wessels, 1993):

$$HKS = \frac{a - p_o p_f}{p_o(1 - p_o)}$$

We will see in part C. of this chapter and in Appendix A.1 that this is equal to the covariance between observations and forecasts divided by the variance of the observations. One other interesting feature is found by taking the least-squares linear regression fit to the forecast and observed occurrences. It turns out that the slope of the regression line is equal to HKS. This will also be discussed in Appendix A.1.

The Hanssen-Kuipers Score, introduced in The Netherlands in 1954 (Daan, 1984), is found in the literature under a large number of names, e.g. the *Kuipers Performance Index* (KPI), *Kuipers skill score* and the *Hanssen-Kuipers Discriminant*. In the above applied convention of using sample climatology instead of long-term climatology the HKS is identical to the score that was first proposed by Peirce in 1884. A skill score which is a linear transformation of HKS is the *Gringorten's skill score* (1967). The HKS was also independently derived by Dobryshman (1972) and is claimed by many others. The *True Skill Statistic* (TSS) (Flueck, 1987) is also identical to the score formulated by Peirce.

For arbitrary number of categories the general form of HKS (with again using sample climatology) in which the scores are written in probabilities looks like:

$$HKS = \frac{\Sigma\, p(f_i, o_i) - \Sigma\, p(f_i)\, p(o_i)}{1 - \Sigma\, p(o_i)^2}$$

Here the summations are over the number of classes of the forecasts f_i and observations o_i.

We now give the derivation for the case of two classes. Expressing the summations in terms of our notations gives

$$HKS = \frac{(a + d) - (1 - p_o)(1 - p_f) - p_o p_f}{1 - \{ p_o^2 + (1 - p_o)^2 \}}$$

We elaborate the numerator and denominator separately. The last two terms of the numerator yield

$$\begin{aligned} & p_o\, p_f + (1 - p_o)(1 - p_f) = \\ &= (a + b)(a + c) + (c + d)(b + d) = \\ &= a(a + b + c) + bc + d(b + c + d) + bc = \\ &= a(1 - d) + d(1 - a) + 2bc = \\ &= a + d - 2(ad - bc) \end{aligned}$$

This leads to $2(ad - bc)$ for the numerator.
The denominator:

$$\begin{aligned} & 1 - \{ p_o^2 + (1 - p_o)^2 \} \\ &= 1 - \{(a + b)(1 - (c + d)) + (c + d)(1 - (a + b))\} \\ &= 1 - \{(a + b) - (a + b)(c + d) + (c + d) - (a + b)(c + d)\} \\ &= 1 - \{ 1 - 2(a + b)(c + d) \\ &= 2(a + b)(c + d) \end{aligned}$$

Dividing these two expressions gives the above expression for HKS. □

This skill score as well as the Heidke skill score which is discussed next, are used very often. In fact, there seems to be some controversy in the literature about what score should be preferred. One of the arguments in that discussion is that there is a tendency for the HKS to reward overforecasting with respect to the Heidke Skill Score. Indeed, in the case of rare events, when in general the number of correct rejections is much larger than the number of correct hits, i.e. $d >> a$, the HKS is almost completely determined by the correctly predicted fraction of the event. Since it is much more easy to correctly predict the nonevent this contribution to the HKS is generally close to one and doesn't change much by a few individual forecasts. In other words, a correct hit is much more rewarded than a correct rejection. Also a false alarm (predicted but not occurred) is less penalized than a miss (occurred but not predicted). Therefore a forecaster may be tempted to overforecast the phenomenon. We will come back to this in sections II.2 and II.3.

g) *Heidke Skill Score* (HSS)

The Heidke skill score was first proposed by Heidke (1926) and was also derived by Panofski & Brier (1958). However, the two-class version of this score was already formulated by Doolittle in 1888 (Murphy, 1996).

In skill score notation it is defined as

$$HSS = \frac{FC_f - FC_r}{1 - FC_r}$$

with

FC_f is the fraction correct of the forecasts

FC_r is the fraction correct under the assumption that there is no relation between forecasts and observations (subscript r stands for random)

For arbitrary number of classes the general form of the HSS with the scores expressed in probabilities is

$$HSS = \frac{\Sigma\, p(f_i, o_i) - \Sigma\, p(f_i)\, p(o_i)}{1 - \Sigma\, p(f_i)\, p(o_i)}$$

in which the summations are over the number of classes. In the case of two classes this leads to

$$HSS = \frac{ad - bc}{(ad - bc) + \tfrac{1}{2}(b + c)}$$

The derivation is as follows:

$FC_f = a + d$

$FC_r = p_o\, p_f + (1 - p_o)(1 - p_f)$

$= a + d - 2(ad - bc)$

(This is given in the calculation of the numerator of the HKS)

$\Rightarrow FC_f - FC_r = 2(ad - bc)$

$1 - FC_r = 1 - a - d + 2(ad - bc)$

$= b + c + 2(ad - bc)$ □

In applying the HSS the following must be noted. In all the above-mentioned (skill) scores the formulations are such that the dimensions of the terms in the numerator and the denominator are the same. Therefore we can use the same formulas using probabilities or absolute numbers. The only two exceptions sofar are the Percentage Correct (by definition) and some of the formulations of the HSS. In the former the sum of the diagonal elements must be divided by the sample total. The expression of HSS changes in the following way. Writing the numbers in the cells of the contingency table in capitals A, B, C and D, and the total number as N then HSS looks like (starting from the skill score definition of HSS)

$$HSS = \frac{(A + D) - \{(A+B)(A+C) + (C+D)(B+D)\}/N}{N - \{(A+B)(A+C) + (C+D)(B+D)\}/N}$$

(In other words, HSS = (C - E) / (N - E) , with now C = the number of correct forecasts and E = the expected number of correct forecasts if random. This expected number is equal to { (O.F) - (N-O).(N-F) } / N , with O the total number of observed events and F of the forecast events)

The term between accolades leads to

$$A(A+B+C) + 2BC + D(B+C+D) =$$
$$AN - AD + 2BC + DN - AD =$$
$$N(A + D) - 2(AD - BC)$$

$$\Rightarrow HSS = \frac{2(AD - BC)/N}{N - (A+D) + 2(AD - BC)/N}$$

$$= \frac{AD - BC}{(AD - BC) + \frac{1}{2} N (B+C)}$$

This is the same formula as on the previous page but now with the total number of cases included in the denominator. This can also directly be inferred from dimension considerations.

An expression for the HSS which can be used on tables expressed in probabilities as well as in number of cases is the following:

$$HSS = \frac{2(AD - BC)}{(A+B)(B+D) + (A+C)(C+D)}$$

An alternative form, making use explicitly of the probability notation, is

$$HSS = \frac{2(ad - bc)}{p_o (1 - p_f) + p_f (1 - p_o)} \quad \text{or} \quad HSS = \frac{2(a - p_o p_f)}{p_o (1 - p_f) + p_f (1 - p_o)}$$

(For the last step see under HKS)

If $p_f = p_o$, i.e. for unbiased forecasts, then HSS = HKS, in accordance with their skill score definitions.

h) *Equitable Threat Score* (ETS)

A modification of the Critical Success Index (sometimes called Threat Score) is the Equitable Threat Score (Hamill, 1998). Another name is the *Gilbert Skill Score*, proposed by Schaefer (1990) in honour of Gilbert's work on the CSI. It is used mainly in the verification of forecasting precipitation amounts. It is defined as the number of correct forecasts in excess to those that would verify by chance, divided by the number of cases when there was a threat that would not have been foreseen by chance. It can be regarded as a skill corrected CSI:

$$ETS = \frac{a - p_o p_f}{a + b + c - p_o p_f}$$

The term $p_o p_f$ is the expected frequency of correct forecasts of the event in a random forecast (see chapter III.1). Random forecasts therefore yield ETS to be zero, whereas its maximum is one (for $b = c = 0$).
Since ETS is equal to the Critical Success Index but with the same (positive) quantity subtracted from both the numerator and the denominator it follows that

$$ETS \leq CSI$$

Their difference is determined by the sample climatology. For extremely rare events ETS approaches CSI.
Making use of the marginal frequencies as much as possible yields

$$ETS = \frac{a - p_o p_f}{p_o + p_f - p_o p_f - a}$$

It can also be written as

$$ETS = \frac{ad - bc}{(ad - bc) + (b + c)}$$

It closely resembles the expression for the Heidke Skill Score. In fact, it is easy to see that

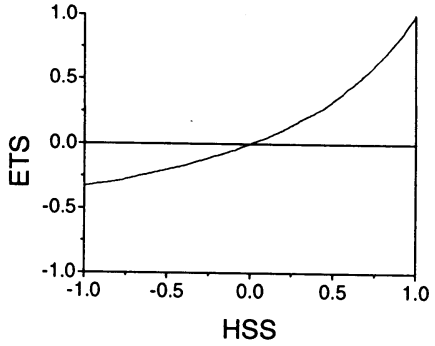
$$ETS = HSS / (2 - HSS)$$

Therefore, it shares many of the features of HSS that are discussed in this section as well as in the two remaining sections of this chapter. The different denominators of HSS and ETS imply that there is a stronger dependence on d (not predicted not observed) in HSS in rare event forecasting.
It follows that

$$ETS = 0 \Leftrightarrow HSS = 0$$

and

$$|ETS| \leq |HSS|$$



Of course, care must be taken in using the above expressions when the performance matrix is written in number of hits, misses, etc. instead of in frequencies. In that case ETS should be written, for instance, as

$$ETS = \frac{AD - BC}{(AD - BC) + N(B + C)}$$

The minimum value of ETS is $-\frac{1}{3}$ for HSS = -1. This will be reached only if two conditions are met. First of all the Forecast Correct should be zero and, secondly, the forecast frequency should be equal to the observed frequency. We will come back to this in section II.2. The main disadvantage of ETS is its dependence on the bias. Typically, in the verification of two competing models, the model with the larger bias (e.g. the wetter forecasts in case of precipitation) tends to have a higher ETS than if the two models had the same bias (Mason, 1989). Therefore, just as for the CSI overforecasting is rewarded. For this reason, when differences between two models are evaluated in terms of ETS also differences in bias should be taken into account. For an extensive discussion on this subject the reader is referred to Hamill (1999).

i) *Rousseau Skill Score* (RSS)

In the three previously discussed skill scores (i.e. HKS, HSS and ETS) random forecasts as well as systematic forecasts of the event or the nonevent all give a value of zero (see section II.2). No distinction can be made between these three strategies in this respect. It seems more logical, however, to classify forecasting systems in the following order. Random forecasts which are unbiased should be rated higher than, for instance, systematic forecasts of the most frequent event, and this in turn should be rated higher than systematic forecasts of the most unlikely event.

Rousseau (1980) developed a skill score which does exactly this, while retaining many of the convenient properties of other skill scores. It is defined by

$$RSS = \frac{4ad - (b + c)^2}{(2a + b + c)(2d + b + c)}$$

This can be transformed in such a way that only the correctly predicted fractions of the events and of the nonevents are used:

$$RSS = \frac{4ad - (1 - a - d)^2}{(1 + a - d)(1 - a + d)}$$

Much more convenient is expressing RSS in the marginal distributions as much as possible. Then it reads

$$RSS = \frac{4a - (p_o + p_f)^2}{(p_o + p_f)\{(1 - p_o) + (1 - p_f)\}}$$

This can be reduced even more by expressing it into the average of the predicted and observed frequencies M, i.e. $M = \frac{1}{2}(p_o + p_f)$, leading to

$$RSS = \frac{a - M^2}{M - M^2}$$

which is a much more simple expression than the one proposed by Rousseau. The maximum value is one for perfect forecasts, since $a = M \Rightarrow a = p_o = p_f$ and $b = c = 0$. Random forecasts yield a score of zero only if $p_o = p_f$, otherwise it's negative. We will come back to this in section III.1.B.

This last expression for the RSS corresponds well with the general formulation of the skill score definition given in the beginning of this section (II.1.B) and therefore a simple interpretation can be given. The accuracy measure of this skill score is now simply the relative frequency of hits (relative to the total number of cases): $score_{fc} = a$. But now the scale is not given in absolute numbers (between 0 and 1) but is expressed as a function of the sample climatology, or rather, as $\frac{1}{2}(p_o + p_f)$, which is one only if $p_o = p_f = 1$. The maximum value of a is therefore equal to M, i.e. $score_{perf} = M$. In RSS it is assumed that the reference forecasts are random unbiased forecasts. In that case the value for cell a becomes p_o^2 (see section III.1), which is equal to M^2. This means that $score_{ref} = M^2$.

The other two "no-skill" strategies, i.e. constant "yes" and constant "no" forecasts, give negative values of RSS. The systematic forecast of the (rare) event gives a lower skill score than forecasting the nonevent:

$$\frac{-(1 - p_o)}{(1 + p_o)} \quad \text{and} \quad \frac{-p_o}{2 - p_o} \quad \text{respectively.}$$

These terms are equal for $p_o = \frac{1}{2}$.

The Rousseau skill score is hardly ever used.

C. A few additional scores

j) There are six more ratios (apart from POD and FAR) that can be formed in which the elements of the contingency table are divided by their associated marginals. These have been given the following names (Doswell et al., 1990) :

post-agreement , also called *frequency of hits* (FOH)

$$FOH = 1 - FAR = \frac{a}{a+c}$$

probability of false detection

$$POFD = \frac{c}{c+d}$$

probability of a null event $$PON = \frac{d}{c+d}$$

frequency of misses $$FOM = \frac{b}{a+b}$$

frequency of correct null forecasts $$FOCN = \frac{d}{b+d}$$

detection failure ratio $$DFR = \frac{b}{b+d}$$

The relations between these scores are

1 - FAR = FOH
1 - POFD = PON
1 - POD = FOM
1 - DFR = FOCN

Of the above scores only POFD is occasionally discussed in the next chapters because it is one of the components of the ROC score which will be discussed in chapter IV (although there POFD is called the 'false alarm rate'). A relation which we encounter is

POD - POFD = HKS

k) *correlation coefficient*

Although not widely used, the correlation coefficient can be calculated from contingency tables as well. Let (f_i, o_i) be pairs of individual forecasts and observations where each forecast or observed occurrence is given a value of 1 and each non-occurrence a value of 0. Then the linear correlation coefficient

$$r = \frac{cov(o_i, f_i)}{\sqrt{\{var(o_i)\ var(f_i)\}}} =$$

$$= \frac{N \Sigma f_i o_i - \Sigma f_i . \Sigma o_i}{\sqrt{\{ [N \Sigma(f_i^2) - (\Sigma f_i)^2] [N \Sigma(o_i^2) - (\Sigma o_i)^2] \}}}$$

reduces to

$$r = \frac{ad - bc}{\sqrt{\{ (a+b)(a+c)(c+d)(b+d) \}}}$$

$$\text{or} \quad r = \frac{ad - bc}{\sqrt{\{ p_o (1 - p_o)\ p_f\ (1 - p_f) \}}}$$

Further details about the derivation can be found in Appendix A.1. Once again, the numerator can be expressed in the marginal frequencies using $ad - bc = a - p_o p_f$.

It can easily be seen that if the forecasts are unbiased the correlation coefficient is equal to HKS and also to HSS. We will come back to this in chapter II.2. If $a = d$ then $r = HKS$ (with a value of $(ad - bc)/p_o p_f$) but the correlation coefficient is generally not equal to HSS anymore.

l) *χ^2- statistic*

The χ^2- test can be used to test the significance of a categorical forecast system. In a 2x2 contingency table the expression reads:

$$\chi^2 = \frac{N (ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$

It follows that:

$\chi^2 = N r^2$, with r the correlation coefficient.

m) *Brier Score* (BS)

A score frequently used in the verification of probability forecasts is the Brier Score

$$BS = \frac{1}{N} \sum_{i=1}^{N} (f_i - o_i)^2$$

in which f_i and o_i are the individual forecasts and observations, N is the total number of cases. This score is also called the *mean square error*. The original formulation by Brier (1950) was twice that given in the above expression. Applying this score to the yes/no situation yields

$$BS = b + c$$

Therefore BS = 1 - Fraction Correct

The Brier score is hardly ever used in the context of categorical forecasting.

II.2. A few general remarks

All scores given in section II.1A. (except for Bias) lie between zero and one:

$$0 \leq \text{FC, POD, FAR, CSI} \leq 1.$$

All except the FAR have a positive orientation. So high values of PC, POD, CSI and low values of FAR are preferred.

It can easily be seen that

$$\text{CSI} \leq \text{POD}$$

Only in cases when $c = 0$ or $a = 0$ the equal sign is valid.

A less frequently used equation expressing the relation between some of the above scores is :

$$\text{CSI} = [\,(\text{POD})^{-1} + (1 - \text{FAR})^{-1} - 1\,]^{-1}$$

There is no bias if the bias ratio is equal to 1. This is the case when $b = c$, i.e. the contingency table is symmetric. If $b > c$ then $p_f < p_o$ (underforecasting); if $b < c$ the opposite is true.

FAR + POD = 1 if there is no bias (b = c).
> 1 if b < c
< 1 if b > c

This can be seen from looking at FAR + POD - 1 :

$$\frac{c}{a+c} + \frac{a}{a+b} - 1$$

$$= \frac{c\,(a+b) + a\,(a+c)}{(a+c)\,(a+b)} - 1$$

Taking these two terms together and elaborating gives for the numerator

$$ac + bc + a^2 + ac - (a^2 + ac + bc + ab) = a\,(c - b).$$

The sign of $c - b$ determines therefore whether FAR + POD is smaller or greater than one.

□

The ranges of the Kuipers, Heidke and Rousseau skill scores are

$$-1 \leq \text{HKS, HSS, RSS} \leq 1$$

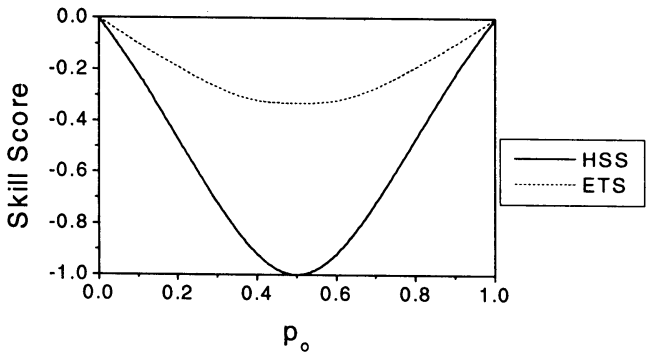
Note that these three skill scores cannot have values below -1, whereas the definition of skill scores allows values much smaller than that. Their lowest value is obtained when Fraction Correct = 0. In that case always HKS = - 1 and RSS = - 1, but this is not necessarily true for HSS. If FC = 0 the matrix elements are completely determined by the observed frequency p_o: $b = p_o$ and $c = 1 - p_o = p_f$. FC = 0 yields for the Heidke Skill Score:

$$\text{HSS} = \frac{-2bc}{(b^2+c^2)} = \frac{-2bc}{1-2bc},$$

with a minimum of -1 only for b = c, i.e. $p_o = p_f$, which can only be true for $p_o = ½$. (see the figure below)
For the (highly similar) ETS we get when FC = 0

$$\text{ETS} = \frac{-bc}{-bc+(b+c)} = \frac{-bc}{1-bc},$$

which has a minimum of $-⅓$ only for $p_o = p_f = ½$ as well (see also the figure below).



So for both HSS and ETS the mimimum skill depends on the sample climatology. Moreover, the lowest skill score values are reached for sample frequencies for which it is relatively "easy" to obtain no hits and no correct rejections. This is not true for HKS and RSS.
In the case of rare events the minimum value for HSS and ETS will be only slightly negative.
Summarising:

for FC = 0	if $p_o = ½$	HSS = -1	ETS = $-⅓$
	if $p_o \downarrow 0$ (or $p_o \uparrow 1$)	HSS $\uparrow 0$	ETS $\uparrow 0$
	regardless of p_o	HKS = -1	RSS = -1

Not surprisingly, the expression ad – bc or $a - p_o p_f$ appears in the numerators of many of the (skill) scores discussed in the previous paragraph. This is true for the Hanssen-Kuipers Score, Heidke Skill Score, correlation coefficient, χ^2- statistic and Equitable Threat Score (or Gilbert Skill Score). It is equal to the value of cell a minus the expected value for

this cell given the observed and predicted frequencies (see chapter III.1). Therefore, all these scores will be zero in the case of random forecasts. Furthermore, they all are zero if the event always occurs or never occurs and if the event is always predicted or never predicted.
Thus, if $p_o = 0$, $p_o = 1$, $p_f = 0$ or $p_f = 1$, or with random forecasts

$$\Rightarrow \quad HKS = HSS = r = \chi^2 = ETS = 0$$

As we have seen this is not the case for RSS. We will come back to this more elaborately in chapter III.1.

A short summary of some of the features of a number of (skill) scores is given in Table II.2.1. First of all the ranges of the scores are given between square brackets (indicating that their limit values can be reached under certain conditions). Also the values or the value ranges of the scores are given in case of the complete absence of the event or of the nonevent and if all forecasts are wrong and if they are all correct, i.e. the Percentage Correct equals zero and one hundred percent, respectively. Finally, this is also done for the three no-skill forecasting strategies of never c.q. always predicting the event to occur and for randomly forecasting with a predetermined (fixed) frequency. The results for the latter strategy depend for some scores on the observed frequency of the event as well as on the difference between this frequency and the predicted one. These relations are determined in chapter III.1; here only the ranges are given. Note that the numbers that are given in the table only hold for general conditions. Many solutions degenerate for specific combinations of observed and forecast frequencies. For instance, many of the values given for $p_o = 0$ only hold if $p_f \neq 0$ and for $p_o = 1$ if $p_f \neq 1$. The same is true for $p_f = 0$ and $p_f = 1$. These combinations should also be excluded for the other columns. For the four fixed frequencies the range of the RSS is between –1 and 0 (-1 included and 0 not included). The lowest value is reached for the observed frequency is 0 while the forecast frequency is 1, and vice cersa. Its upper limit, which is in all cases zero, is approached for very smal differences between observed and forecast frequency. If the forecasts are exactly unbiased then the RSS will be undefined in these cases.

	Range	PC=0	PC=100	$p_o = 0$	$p_o = 1$	$p_f = 0$	$p_f = 1$	random
HKS	[-1,1]	-1	1	undef	undef	0	0	0
HSS	[-1,1]	[-1,0]	1	0	0	0	0	0
RSS	[-1,1]	-1	1	[-1,0)	[-1,0)	[-1,0)	[-1,0)	[-1,0]
ETS	[-⅓,1]	[-⅓,0]	1	0	0	0	0	0
r	[-1,1]	-1	1	undef	undef	undef	undef	0
FC	[0,1]	0	1	$1-p_f$	p_f	$1-p_o$	p_o	[0,1]
POD	[0,1]	0	1	undef	p_f	0	1	[0,1]
FAR	[0,1]	1	0	1	0	undef	$1-p_o$	[0,1]
CSI	[0,1]	0	1	0	p_f	0	p_o	[0,1]
BS	[0,1]	1	0	p_f	$1-p_f$	p_o	$1-p_o$	[0,1]

Table II.2.1 Summary of the (ranges of) values for a number of (skill) scores under a few conditions. These conditions are: Percentage Correct (PC) is zero or hundred percent, absence of the event or nonevent, forecast distribution is zero, one or random.

Many forecasting systems yield predicted frequencies of a particular event, which are close to or even equal to the observed frequency. Let's compare therefore the expressions of a few (skill) scores under this assumption. Let's first look at forecasting systems in which the **forecasts are unbiased**. In this case the scores get a very convenient form. The probabilities of each cell of the performance matrix can then be expressed as a function of the probability of cell a and the observed frequency:

		FC: yes	FC: no	
OBS	yes	a	$p_o - a$	p_o
OBS	no	$p_o - a$	$1 - 2p_o + a$	$1 - p_o$
		p_o	$1 - p_o$	

It is easy to see from their definitions that HKS, HSS and the correlation coefficient r are all equal to

$$\frac{a - p_o^2}{p_o - p_o^2}$$

But also the Rousseau Skill Score is equal to this expression for unbiased forecasts. This can be inferred from one of the last expressions for RSS in paragraph II.B.i. with $M = \frac{1}{2}(p_o + p_f) = p_o$. Therefore:

$$\text{if } p_o = p_f \Rightarrow HKS = HSS = RSS = r = \frac{a - p_o^2}{p_o - p_o^2}$$

It follows that (for p_o not equal to 0 or 1) these four scores are

1 if $a = p_o$ (i.e. if Fraction Correct = 1)

0 if $a = p_o^2$ (e.g.if the (unbiased) forecasts are random, see chapter III.1)

Their minimum value is reached for a = 0:

$$\frac{-p_o}{1 - p_o}$$

But, according to the performance matrix for unbiased forecasts, also $p_o - a \leq 1 - p_o$, implying $p_o \leq \frac{1}{2}$ (for a = 0). Therefore the minimum value is limited to -1, which can only be reached for $p_o = \frac{1}{2}$. In that case also FC = 0. In the case of unbiased forecasts of a predictand which is relatively rare the four above scores can be only slightly negative.
The ETS can be expressed in the unbiased case as

$$\frac{a - p_o^2}{p_o - p_o^2 + (p_o - a)}$$

and is therefore always smaller (in absolute sense) than the above scores except in the perfect case that $a = p_o$.

Finally, let's look at the situation in which the **forecasts are biased**. In this case the general expressions of the scores hold as given before, but we will now express them as a function of the difference between observed and predicted frequency.
Suppose $p_f = p_o + \Delta$ with $-p_o \leq \Delta \leq 1 - p_o$. Δ is the amount, in terms of the fraction of the total number of cases, of over- or underforecasting. Δ is equal to (c - b). The expressions of the above 4 scores can then be written as follows (manipulations are not shown):

$$HKS = \frac{a - p_o^2 - \Delta p_o}{p_o - p_o^2}$$

$$HSS = \frac{a - p_o^2 - \Delta p_o}{p_o - p_o^2 - \Delta p_o + \frac{1}{2}\Delta}$$

$$RSS = \frac{a - p_o^2 - \Delta p_o - \frac{1}{4}\Delta^2}{p_o - p_o^2 - \Delta p_o - \frac{1}{4}\Delta^2 + \frac{1}{2}\Delta} = \frac{a - (p_o + \frac{1}{2}\Delta)^2}{(p_o + \frac{1}{2}\Delta) - (p_o + \frac{1}{2}\Delta)^2}$$

$$r = \frac{a - p_o^2 - \Delta p_o}{\sqrt{\{ (p_o - p_o^2)^2 + \Delta p_o(1 - p_o)[1 - 2p_o - \Delta] \}}}$$

For completeness, also ETS is given:

$$ETS = \frac{a - p_o^2 - \Delta p_o}{p_o - p_o^2 - \Delta p_o + \Delta + (p_o - a)}$$

This set of equations can be regarded as a set of alternative definitions of the particular (skill) scores. With $\Delta = 0$ all scores, except ETS, become the same. These expressions can be used to assess the relative dependence on Δ as a function of the sample climatology. This can be illustrated by comparing the two probably most frequently used skill scores HKS and HSS.
We have already seen that $HKS = 0 \Leftrightarrow HSS = 0$.
The numerators of HKS and HSS are the same. The denominators, on the other hand, differ only by the term $-\Delta(p_o - \frac{1}{2})$, with Δ equal to (c – b). It follows that HSS = HKS whenever the forecasts are unbiased, as we have seen before, and also when $p_o = \frac{1}{2}$. This latter case is much less obvious. It can also be obtained from the definitions of the scores expressed in summations over the classes in the previous section. In chapter III we will give a number of examples (or models) that give rise to unbiased forecasts. If $(p_o - \frac{1}{2})(b - c)$ has a positive sign then $|HSS| < |HKS|$; a negative sign yields $|HSS| > |HKS|$. The difference between b and c determines which of the two scores gives the highest value in (more or less) rare events, i.e.

when $p_o < ½$. So in skilful forecasting systems overforecasting yields higher values of HKS. More about this subject follows in the next paragraph.
The above conclusions can also be inferred from the following previously seen formulations of the two skill scores:

$$HKS = \frac{ad - bc}{p_o(1 - p_o)} \quad \text{and} \quad HSS = \frac{2(ad - bc)}{p_o(1 - p_f) + p_f(1 - p_o)}$$

Geometric interpretation of the contingency table

As we have indicated already in the discussion of the correlation coefficient (in II.1), the contingency table can also be regarded as scatter plot between observed and forecast cases. By assigning a value of one to a yes forecast and to an observed event and zero otherwise you get four possible combinations:

(obs,fc) = {(1,1), (1,0), (0,1), (0,0)}

The numbers of each of these pairs represent the elements of the 2x2 contingency table; in terms of the definition of chapter II they are A, B, C and D, respectively. Note that the positions of the elements of the contingency table are not the same as in the scatter plot.

By taking the regression lines of the forecasts upon the observations and of the observations upon the forecasts and intersecting these lines with obs = 0, obs = 1, fc = 0 and fc = 1 a number of scores defined in chapter II can be depicted. Details of this geometrical interpretation of these scores associated with the 2x2 contingency table are given in Appendix A.1.

II.3. Considerations on which score(s) to use

An elaborate discussion which score or scores can be used best under what conditions is outside the scope of this report. Only a few general remarks are given.

As mentioned before, hardly ever a single verification score is enough to get a full picture of the performance of a forecasting system. Almost always more scores are evaluated at the same time. If only one of the scores of the previous chapters is used it is fairly easy for a forecaster to adapt his / her forecasts in order to improve the results.

An example of the use of a single score in a verification study is presented in the paper by Finley in 1884. In this paper Finley reported on an experimental tornado forecasting system and obtained an extremely high Percentage Correct. However, this is a very inappropriate measure in rare event forecasting, as was argued in a number of subsequent papers later in 1884. It is easy to see that in case the number of false alarms exceeds the number of hits (i.e. in the notation of the performance matrix: $C > A$), which is usually the case with prediction systems for extremely rare phenomena, then always predicting the event not to occur simply leads to a higher percentage correct. This was indeed the case with the tornado data, as was pointed out by Gilbert (1884). Never predicting a tornado would have beaten the existing forecasting system by far in terms of percentage correct.

Another obvious example of the danger of using only one score can be inferred from the verification in terms of the probability of detection. It is always possible to obtain very high POD scores but only at the expense of a high FAR and vice versa. A good set of forecasts must therefore achieve fixed verification targets for both scores. For instance in the evaluation of the minimum road temperature forecasts in the British Isles, described by Halsey (1995), simultaneous targets were set for POD and FAR and also for Fraction Correct (see Fig.1). Because of the fact that a categorical forecast should have a high POD as well as a low FAR it is not surprising that a score like the relative operating characteristic (ROC) has grown so popular. This score makes use of two similar scores, although the definitions of those 'POD' and 'FAR' are somewhat different from the ones discussed so far. A more elaborate discussion of the ROC score is presented in chapter IV of this report.

A general requirement concerning the properties of a scoring system is that it should not offer the forecaster the possibility to "*hedge*". In other words, it should not be possible "playing the score" in order to obtain the best possible score. The forecasts should reflect the true beliefs of the forecaster and should not depend on the scoring system which is employed to evaluate the forecasts. That is, a good scoring system should encourage the forecaster to make forecasts corresponding to his / her true beliefs. Or even stonger, the best value of the score should be expected if and only if this is the case. This property is also called *strictly proper*. Hedging may lead to forecasts that exhibit undesirable characteristics, like systematic biases. The subject of hedging was first introduced by Epstein in 1966 in the context of the evaluation of probability forecasts. Further discussion of the topic of hedging in probabilistic forecasting is given by Murphy and Epstein (1967), Murphy (1973) and also by Wilks (1995).

A topic strongly related to hedging is the *equitability* of the skill scores which are used in categorical forecasting. In many skill scores constant forecasts of one event lead to better scores than constant forecasts of the other event. The use of these so-called inequitable skill scores may encourage the forecasters to favour that one event at the expense of the other. This will be illustrated in the following example.

Example.

Consider a performance matrix of precipitation forecasts of the following form:

OBS \ FC	yes	no	
yes	2	3	5
no	1	9	10
	3	12	

HKS = 3/10 HSS = 1/3
(ETS = 1/5 RSS = 0.318)

This is an example which clearly exhibits forecasting skill. The number of hits as well as the number of correct rejections is more than would be expected from pure chance. Therefore the Hanssen-Kuipers Score and Heidke Skill Score are both positive: HKS = 3/10 and HSS = 1/3. Because of the underforecasting HKS < HSS. Now suppose a forecaster has to make three more forecasts but doesn't know for meteorological reasons which category to "choose". He or she may be tempted to base his / her choice on which one will result in the best score. We furthermore suppose that these three new events do not alter the sample climatology, i.e. in one of the cases it rained and in the other two it remained dry. (Three is chosen just because we want to deal with integer numbers in this example). By doing this it is implicitely assumed that the sample climatology is equal to or at least very close to the climatological values. By "always" predicting rain the performance matrix will become as shown below on the left. The result of "always" predicting dry weather is shown on the right. Both skill scores are also given. In the case of 'rain' forecasts the performance matrix becomes unbiased and so the scores are equal. It appears that the HKS is in both cases the same. No matter which category has been chosen the skill score deteriorates by the same amount. This is a consequence of the equitability of the HKS.

OBS \ FC	yes	no	
yes	3	3	6
no	3	9	12
	6	12	18

HKS = 1/4 HSS = 1/4
(ETS = 1/7 RSS = 1/4)

OBS \ FC	yes	no	
yes	2	4	6
no	1	11	12
	3	15	18

HKS = 1/4 HSS = 2/7
(ETS = 1/6 RSS = 0.259)

The HSS, on the other hand, deteriorates less in this example if the forecaster predicts the dry category. This inequitability of the HSS is clearly an undesirable characteristic. It

encourages the forecaster to choose the category not on purely meteorological grounds. In general, if there is no a priori information whatsoever in rare event forecasting and we start with a positive value of the HSS so far, then it is beneficial in terms of this score to always predict the event not to occur.

This can be made clear in the following way. It can be proven first of all that under the above assumptions (no a priori information and no change in the sample climatology) always predicting either of the two categories yields the same numerator. This can be seen by looking at the change in AD - BC after a number of, say α, "yes" forecasts versus the change after the same number of "no" forecasts. The respective performance matrices look as follows

OBS \ FC	yes	no	
yes	$A + \alpha p_o$	B	$A+B+\alpha p_o$
no	$C+\alpha(1-p_o)$	D	$C+D+\alpha(1-p_o)$
	$A+C+\alpha$	B+D	$N + \alpha$

OBS \ FC	yes	no	
yes	A	$B+\alpha p_o$	$A+B+\alpha p_o$
no	C	$D+\alpha(1-p_o)$	$C+D+\alpha(1-p_o)$
	A+C	$B+D+\alpha$	$N + \alpha$

In the case with the additional "yes" forecasts AD - BC becomes

$$(A + \alpha p_o)D - B(C + \alpha (1 - p_o))$$

and in the second case

$$A(D + \alpha(1 - p_o)) - (B + \alpha p_o)C.$$

The change with respect to the initial value of AD - BC is

$$\alpha\{p_oD - (1 - p_o)B\} \text{ and } \alpha\{(1 - p_o)A - p_oC\} \text{ respectively.}$$

Substitution of $A = Np_o - B$ and $C = N(1 - p_o) - D$, with N the number of original cases, leads to the same expressions between the accolades, and therefore the numerators will be the same for the two strategies. (This is by the way the proof that the two strategies lead to the same HKS since also the denominator changes by the same amount for the two strategies.)

The denominator of the HSS, on the other hand, expressed for instance in relative frequencies, i.e. $p_o(1 - p_f) + p_f(1 - p_o)$, shows different behaviour for the two strategies. It is minimal if p_f is as small as possible, yielding maximum values of $|HSS|$. In other words, under the above assumptions and assuming positive skill HSS rewards underforecasting. The same conclusion can be obtained from looking at one of the alternative formulations of the denominator of HSS: (AD - BC) + ½ N (B+C). Always predicting "no" gives smaller changes in B than predicting "yes" gives in C, due to the imposed fixed observed relative frequency. More specifically, the term (B+C) changes into $\{B + \alpha p_o + C\}$ and $\{B + C + \alpha(1 - p_o)\}$ respectively. The first expression is always smaller, and therefore $|HSS|$ is higher, if $p_o < ½$, and vice versa.

□

Exactly the same arguments hold for the Equitable Threat Score, since it is composed of exactly the same components. Therefore, also the Equitable Threat Score is not an equitable skill score. Only in the case the observed frequency is exactly 50 percent then ETS as well as HSS are equitable. Also for ETS it follows that the best strategy in the case that we have no a priori information and the sample climatology is the same as (or at least very close to) what can be expected on the basis of climatological information, is that we should always predict "no" if $p_o < ½$ and always "yes" if $p_o > ½$. . In other words, one should always forecast the event with the highest climatological probability. (Under the assumption once again that we started off with a positive value of the score).

In fact, it can be proven that the HKS is the only equitable skill score in the two-category case (Gandin and Murphy, 1992, and Gerrity, 1992). All skill scores for categorical forecasts are based on so-called *scoring matrices*. These are matrices that assign scores (or weights) to each of the four possible combinations of forecasts and observations. Here they are referred to as s_{11} , s_{12} , s_{21} and s_{22} according to their position in the matrix. Correctly predicted rare events could have been assigned very high scores whereas correctly predicted non-occurrences could have a low score. Wrong forecasts of either event will have a negative score (weight) in most skill scores. And, just as a final example, in the case of a 3-category problem, predictions that are two classes off will in general have lower scores than if they are only one class off.

Equitable skill scores must assign, by definition, the same score to forecasts of both events (or event and nonevent). This means that the expected score of a constant event one forecast will be

$$p_o s_{11} + (1 - p_o)\, s_{21} = 0 \quad \text{and for event two:}$$
$$p_o s_{12} + (1 - p_o)\, s_{22} = 0 .$$

For a perfect forecast the expected score will be

$$p_o s_{11} + (1 - p_o)\, s_{22} = 1 .$$

The choice of zero and one in the right hand sides of these equations is in fact arbitrary. Together with $s_{12} = s_{21}$ this leads to

$$s_{11} = (1 - p_o) / p_o$$
$$s_{22} = p_o / (1 - p_o)$$
$$s_{12} = s_{21} = -1$$

An arbitrary forecast will lead therefore to an expected score of

$$a \frac{(1 - p_o)}{p_o} - b - c + d \frac{p_o}{(1 - p_o)} .$$

This is (with $b + c = 1 - a - d$)

$$\frac{a}{p_o} + \frac{d}{1 - p_o} - 1 ,$$

which is exactly equal to the Hanssen-Kuipers Score. So every equitable skill score in the two-category case can be written as a linear transformation of the HKS.

A useful consequence of the imposed equitability is that a random forecast will also have an expected score equal to zero. Namely, random forecasts render an expected score of

$$p_o\, p_f\, s_{11} + p_o(1 - p_f)\, s_{12} + (1 - p_o)\, p_f\, s_{21} + (1 - p_o)\,(1 - p_f)\, s_{22}$$

Using the above expressions for the expected scores for constant event one and constant event two predictions this is equal to zero. (Note that the assumption of symmetry of the scoring matrix is not needed for this). So random forecasts as well as constant forecasts of one event will give zero scores. An important additional feature is that the scores assigned to correct forecasts of an event increase as the climatological probability of the event decreases. Further details are given in Gandin and Murphy (1992).

□

Above we discussed the different behaviour of two skill scores (HKS and HSS) for constant event one or two forecasts given that the sample climatology remains unchanged. The outcome of this experiment can be used, for instance, for large enough data sets in which the observed frequency is equal to or close to the long-term climatological frequency. In that case the relative frequency of the event is not likely to change. It was found that if the verification score were HSS then, starting with a positive value of HSS, the best strategy would be to predict the (rare) event not to occur. If the score up to that moment was negative the opposite strategy should be preferred. If the initial value of HSS is equal to zero then there's no preference for either strategy (and HSS remains zero). The same is true for ETS. Due to the equitability of HKS no strategy can be favoured if HKS is used, regardless of its initial value. In other words, the probability of random succes is constant regardless of which category is forecast.

A next step could be to examine situations in which the observed frequency is (substantially) different from its (known) climatological mean. In that case one could consider the consequences of the two above strategies in terms of the particular skill score that is used, under the condition that the only available information of the phenomenon to be predicted is its climatological occurrence. Here only the HKS is discussed.

As before, let's look at the numerator of this skill score expressed in number of hits, misses, etc. Now the change in AD - BC will be the result of α new cases in which the event occurs according to its climatological frequency. And, as said before, this frequency may be different from the frequency of observed events up to that moment. The change of the numerator with respect to the initial value of AD - BC is for constant forecasts of the event

$$\alpha\{\, p_c D - (1 - p_c)B \,\}$$

and for constant forecasts of the nonevent

$$\alpha\{\, (1 - p_c)A - p_c C \,\}$$

Substituting now $A = Np_o - B$ and $C = N(1 - p_o) - D$, with N the number of original cases, leads to different values for the two strategies. The change in the numerator in the case of "no" forecasts is

$$\alpha\{ (1 - p_c)A - p_c C \} = \alpha\{ (1 - p_c)Np_o - (1 - p_c)B - p_c N (1 - p_o) + p_c D \} =$$
$$= \alpha\{ p_c D - (1 - p_c)B \} + \alpha\{ (1 - p_c)Np_o - (1 - p_o)Np_c \} =$$
$$= \alpha\{ p_c D - (1 - p_c)B \} + \alpha N (p_o - p_c)$$

This is not equal anymore to the change resulting from "yes" forecasts. This is due to the additional term which is proportional to the difference between the climatological and sample frequencies. So, if $p_o - p_c > 0$, i.e. in the original sample the event has occurred more than could be expected on the basis of its climatology, then the "no" strategy yields higher numerators of the HKS than the "yes" strategy does. Since the denominator of the HKS is only dependent on the observed frequency and not on the predicted one, this results in a very simple policy when the HKS is used as verification tool. Namely, if there is no other information about the predicted phenomenon than its climatological frequency of occurrence then it is beneficial in terms of the HKS verification score to keep the predicted frequency close to the climatological value. If there is more information available (e.g. from a deterministic or probabilistic model) indicating for instance only a slight increase of the probability of the event to occur then one could resort to arguments as given in chapter II.1.B.f. There it was stated that because correctly forecasting the event is rewarded much higher than a correct rejection, and a false alarm is hardly penalized it might be tempting in case of doubt to forecast the event to occur. In fact, it can be shown that if the probability of the event is larger than its climatological probability then one should forecast the event (Hanssen and Kuipers, 1965). This translation of probabilistic information into a categorical statement is outside the scope of this report. Further details can be found in Murphy and Katz (1985), Glahn et al. (1991) and Wilks (1995).

Above the effect of certain strategies has been shown on a few skill scores. It appears that some scores may be played with more than others. This doesn't necessarily mean, however, that the use of those particular scores should be avoided. There are many other arguments that may play a role.

III. Behaviour under different model assumptions

In the previous part of this report we have discussed a lot of (more or less) well-known (skill) scores that can be defined on 2x2 performance matrices together with some of their properties. In this part we will look at the behaviour of only a few of these scores under different assumptions or forecasting strategies. The expected value of the scores will be presented as a function of the observed frequency of the event. This is the part that the forecaster doesn't know beforehand but which has a great impact on the results. We shall see that small differences in sample climatology sometimes give rise to relatively large differences in the value of the scores.

The first "models" that are discussed in this chapter consist of random forecasts. The expected performance matrix and the resulting expressions of the scores will be shown. The values of the scores in dependence on the occurrence of the event can be regarded as reference level for these scores. Next a similar analysis is done under the assumption that a fixed fraction of the predicted events and nonevents are predicted correctly. Finally, the effects of a fixed skill over climatology are presented. All three models may offer a reference to the forecaster to interpret observed verification scores or to estimate the expected value of the scores. Moreover, they may offer a means of intercomparing verification scores obtained for forecasts valid for different locations where the predictands have different statistical properties. Maybe these models can even be used to help defining target values which different forecasters have to meet in dependence on the local climatologies of the predictands.

III.1 Random forecasts

Let's first consider the case in which the forecaster has no model information whatsoever to base his / her judgement on. In that case the forecaster has to resort to "random guessing" (or making random forecasts). But, since there usually is some understanding of the rareness of the event at hand this can be used to make a first guess of the frequency in which the event will occur. The next step is, using this frequency, to randomly forecast an event or nonevent. This is of course equivalent to repeatedly making probabilistic forecasts with the climatological probability. This can be looked at as the reference forecast system that the forecaster should beat in order to have any skill.

In part A. random guessing is considered under the additional constraint that the forecast frequency is equal to the sample climatology (i.e. the frequency was "guessed" right). It may be applied, for instance, to situations in which the sample is large enough for the observed frequency to be a good "approximation" of the climatological frequency, or rather, in which the climatological frequency is expected to be a good forecast of the observed frequency over a period of time. In part B. the situation is discussed for arbitrary predicted frequencies.

A. Unbiased random forecasts

This can be regarded as the situation in which the forecaster knows the observed frequency beforehand and that he / she thinks that it will be advantageous, in terms of the chosen verification score(s), to keep the predicted frequency of the event as close as possible to the observed one. It may serve as a reference to the forecaster of what might already be obtained without any prior knowledge of the phenomenon other than its (in most cases climatological) frequency.

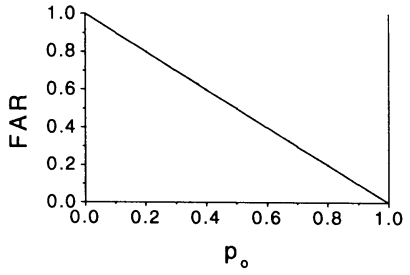
The expected 2x2 contingency table is in this case:

		FC: yes	FC: no	
OBS	yes	p_o^2	$p_o (1 - p_o)$	p_o
OBS	no	$p_o (1 - p_o)$	$(1 - p_o)^2$	$1 - p_o$
		p_o	$1 - p_o$	

The expressions for a number of scores are given next.

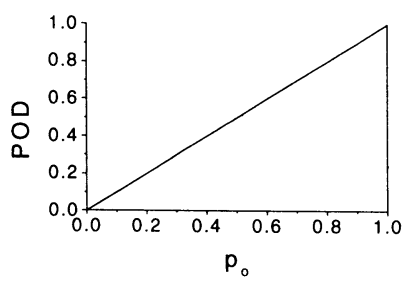
Bias ratio = 1

$$FAR = \frac{p_o (1 - p_o)}{p_o} = 1 - p_o$$



The FAR is always high for rare events, even in case of unbiased forecasts.

$$POD = \frac{p_o^2}{p_o} = p_o$$



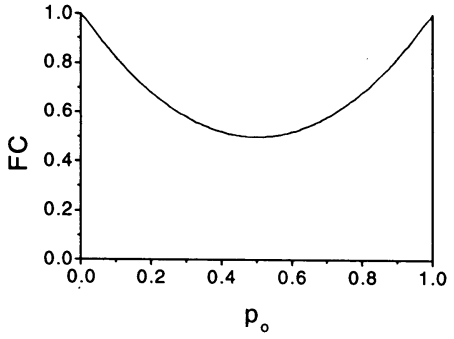
Since there is no bias: POD + FAR = 1

$$\text{POFD} = \frac{p_o (1 - p_o)}{1 - p_o} = p_o$$

So POD = POFD.

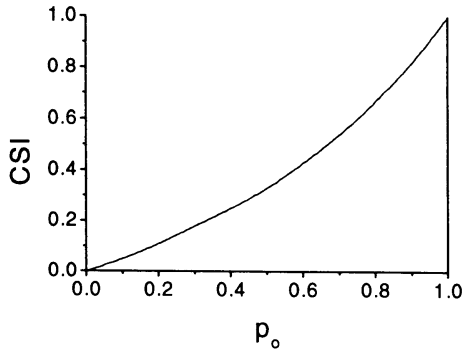
$$\text{FC} = p_o^2 + (1 - p_o)^2$$

which has a minimum of 50% at $p_o = ½$.



The general characteristics of the above three scores as a function of the observed frequency resemble those shown by Halsey (1995) and presented in Fig. 1. An increasing probability of detection with increasing number of occurrences of the event is apparently an intrinsic feature, which exists already in the no skill situation. The same is true for a decreasing False Alarm Ratio and a parabolically shaped Fraction Correct. The target values forecasters or forecast systems have to meet should therefore have a similar dependence on the expected frost frequency. And, of course, they should be tighter than the values of the scores in the random case to have skill. The frost forecasts used in Fig. 1 undoubtedly exhibit skill: Fractions Correct and Probabilities of Detection are higher and False Alarm Ratios are lower than in the case of random forecasts. In sections III.2 and III.3 we will give a few examples which enable us to judge the sensitivity of the scores to differences in the degree of skill (or accuracy) of the forecasts.

$$\text{CSI} = \frac{p_o^2}{p_o^2 + 2p_o - 2p_o^2} = \frac{p_o}{2 - p_o}$$



By definition is HKS = 0 and HSS = 0 .
This can also be inferred from the formulations of both scores, most easily from the expression with the numerators in the form $a - p_o p_f$. Therefore also ETS and the correlation coefficient r are zero. And, as we have seen in chapter II.2, also RSS = 0.

The above expressions of the scores are only valid for situations in which the event and nonevent are not completely absent in the considered data sample. In the case that $p_o = 0$ (and therefore $p_f = 0$) the Bias ratio, POD, FAR and CSI are undefined. If $p_o = 1$ the same is true for POFD. The skill scores (HKS, HSS, ETS and RSS) as well as the correlation coefficient are undefined for both extreme cases.

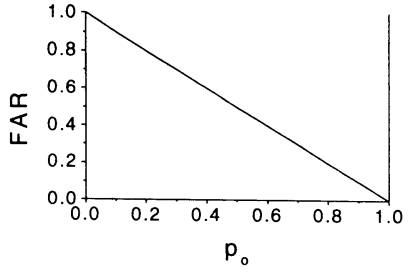
B. Random forecasts

In this subsection random forecasts are considered without additional constraints, i.e. in general $p_f \neq p_o$. This can be regarded as random guessing without any prior knowledge. Subsection A. is a special case of this. The performance matrix looks like this:

		FC: yes	FC: no	
OBS	yes	$p_o p_f$	$p_o (1 - p_f)$	p_o
OBS	no	$(1 - p_o) p_f$	$(1 - p_o)(1 - p_f)$	$1 - p_o$
		p_f	$1 - p_f$	

$FAR = 1 - p_o$

This expression is the same as in the case of unbiased random forecasts (see subsection A.). So randomly over- or underforecasting does not affect the False Alarm Ratio.



POD = p_f , so there is no explicit relation with the observed frequency. In general, however, p_f will be close to p_o .
The same is true for the probability of false detection: POFD = p_f.

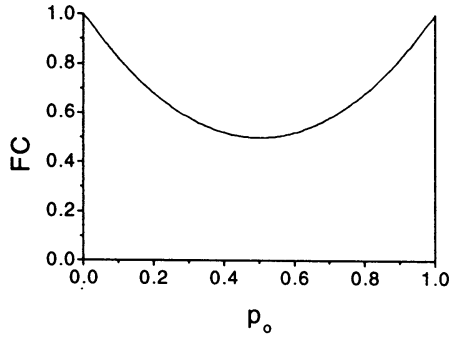
Since there is no bias: POD + FAR = 1

$$\text{POFD} = \frac{p_o (1 - p_o)}{1 - p_o} = p_o$$

So POD = POFD.

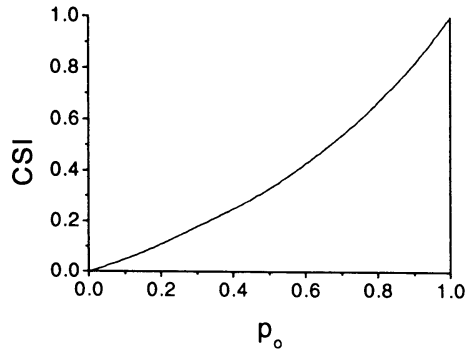
$$\text{FC} = p_o^2 + (1 - p_o)^2$$

which has a minimum of 50% at $p_o = ½$.



The general characteristics of the above three scores as a function of the observed frequency resemble those shown by Halsey (1995) and presented in Fig. 1. An increasing probability of detection with increasing number of occurrences of the event is apparently an intrinsic feature, which exists already in the no skill situation. The same is true for a decreasing False Alarm Ratio and a parabolically shaped Fraction Correct. The target values forecasters or forecast systems have to meet should therefore have a similar dependence on the expected frost frequency. And, of course, they should be tighter than the values of the scores in the random case to have skill. The frost forecasts used in Fig. 1 undoubtedly exhibit skill: Fractions Correct and Probabilities of Detection are higher and False Alarm Ratios are lower than in the case of random forecasts. In sections III.2 and III.3 we will give a few examples which enable us to judge the sensitivity of the scores to differences in the degree of skill (or accuracy) of the forecasts.

$$\text{CSI} = \frac{p_o^2}{p_o^2 + 2p_o - 2p_o^2} = \frac{p_o}{2 - p_o}$$



By definition is HKS = 0 and HSS = 0.
This can also be inferred from the formulations of both scores, most easily from the expression with the numerators in the form $a - p_o p_f$. Therefore also ETS and the correlation coefficient r are zero. And, as we have seen in chapter II.2, also RSS = 0.

The above expressions of the scores are only valid for situations in which the event and nonevent are not completely absent in the considered data sample. In the case that $p_o = 0$ (and therefore $p_f = 0$) the Bias ratio, POD, FAR and CSI are undefined. If $p_o = 1$ the same is true for POFD. The skill scores (HKS, HSS, ETS and RSS) as well as the correlation coefficient are undefined for both extreme cases.

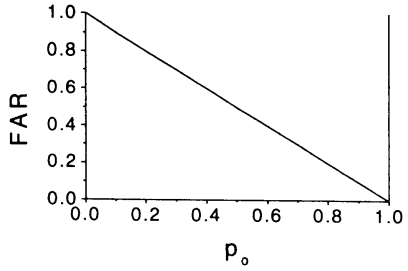
B. Random forecasts

In this subsection random forecasts are considered without additional constraints, i.e. in general $p_f \neq p_o$. This can be regarded as random guessing without any prior knowledge. Subsection A. is a special case of this. The performance matrix looks like this:

OBS \ FC	yes	no	
yes	$p_o p_f$	$p_o (1 - p_f)$	p_o
no	$(1 - p_o) p_f$	$(1 - p_o)(1 - p_f)$	$1 - p_o$
	p_f	$1 - p_f$	

$FAR = 1 - p_o$

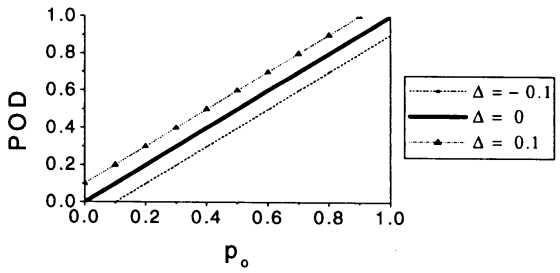
This expression is the same as in the case of unbiased random forecasts (see subsection A.). So randomly over- or underforecasting does not affect the False Alarm Ratio.



POD = p_f , so there is no explicit relation with the observed frequency. In general, however, p_f will be close to p_o.
The same is true for the probability of false detection: POFD = p_f.

In other words, if we have random forecasts then always POD = POFD.

Suppose $p_f = p_o + \Delta$ with $-p_o \leq \Delta \leq 1 - p_o$. Δ can therefore be considered to be the amount of over- or underforecasting. The POD graph will resemble the one of the previous section very closely. For different values of Δ we get a set of lines parallel to the 45° line. By simply overestimating the frequency of occurrence of the event a high value of the POD can be obtained. In fact, (randomly) overforecasting with a certain percentage results in an increase of the POD with the same amount. The same is true for POFD.



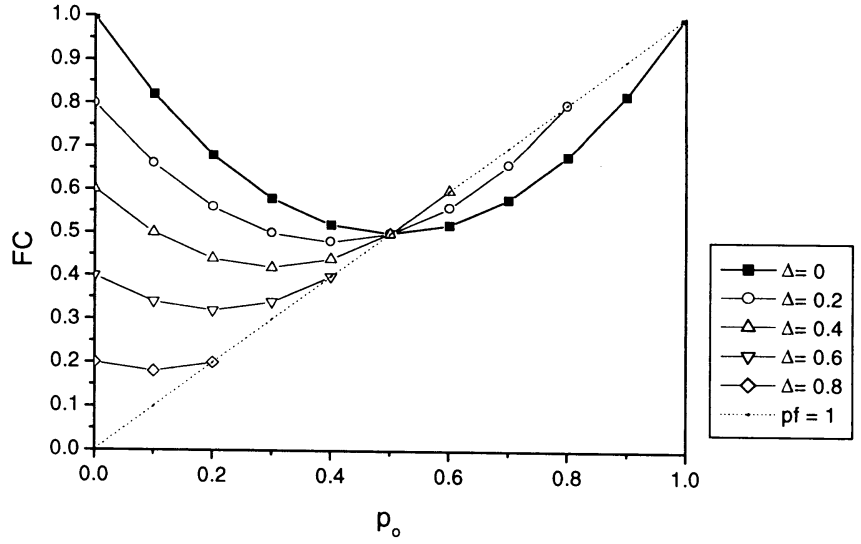
$FC = p_o p_f + (1 - p_o)(1 - p_f)$

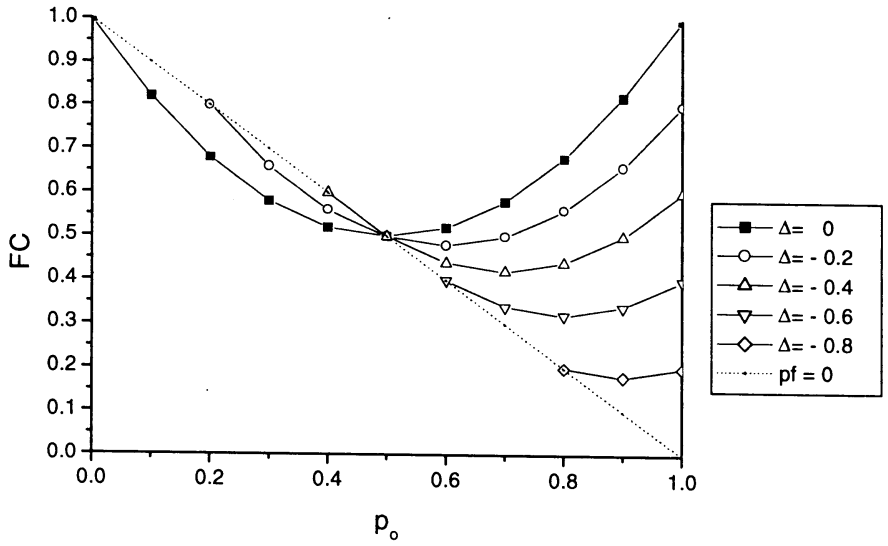
Define $p_f = p_o + \Delta$ and consider first $p_f \geq p_o$, i.e. $0 \leq \Delta \leq 1 - p_o$.
Δ is the degree of overforecasting.

then
$$FC = p_o^2 + p_o \Delta + 1 - p_o - p_o (1 - p_o) - \Delta (1 - p_o)$$
$$= p_o^2 + (1 - p_o)^2 + \Delta (2p_o - 1)$$
$$= 1 - 2p_o (1 - p_o) + \Delta (2p_o - 1)$$

Taking $\Delta = 0$ we get the unbiased case discussed under A.
The relation with the sample climatology is given in the figure below (top). Overforecasting leads to a decreasing Fraction Correct if $p_o < ½$ and an increasing FC for larger p_o. For an observed frequency of the event of exactly 50% overforecasting (as well as underforecasting) doesn't affect FC. For the maximum value of Δ , i.e. if $p_f = 1$, the equation degenerates to the straight line $FC = p_o$.
In case of underforecasting, i.e. $-p_o \leq \Delta \leq 0$ (bottom figure) we get the mirror image of the top figure mirrored at $p_o = ½$. Taking the lower limit, $\Delta = -p_o$, i.e. $p_f = 0$, results in
$FC = 1 - p_o$.

Forecasting systems for hazardous weather phenomena often exhibit a certain degree of overforecasting. This is mainly due to the fact that it is regarded to be much worse to fail to forecast the event than it is to have a few additional false alarms. So in case of doubt a warning will be issued. (That is, in categorical forecasting, the event will be predicted). This is also the reason why the targets set in Fig.1 for the Fraction Correct are somewhat lower for low than for high frost frequencies. The same asymmetry seems to be present for the observations. The target line is not symmetric with respect to $p_o = ½$ and is similar in shape (but on a much higher level) to the random case in which some degree of overforecasting is imposed (see the top figure below).





Similar manipulations can be performed for the Critical Success Index.

$$CSI = \frac{p_o\, p_f}{1 - (1 - p_o)(1 - p_f)} = \frac{p_o\, p_f}{p_o + p_f - p_o\, p_f}$$

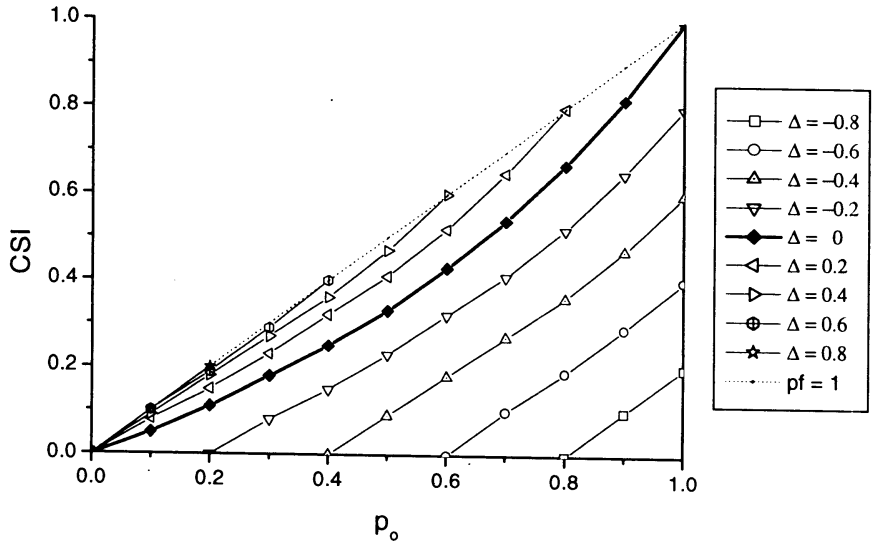
Suppose again that $p_f = p_o + \Delta$ with first $0 \le \Delta \le 1 - p_o$
then CSI becomes

$$CSI = \frac{p_o\,(p_o + \Delta)}{2\,p_o + \Delta - p_o\,(p_o + \Delta)} = \frac{p_o + \Delta}{2 - p_o - \Delta\,(1 - 1/p_o)}$$

This is always (except for $p_o = 0$) larger than $\frac{p_o}{2 - p_o}$, the expression we have seen earlier.

In the limit case with $\Delta = 1 - p_o$, i.e. $p_f = 1$, then $b = d = 0$ and $CSI = p_o$. This can also be seen directly from the performance matrix. With $\Delta < 0$ (and $-p_o \leq \Delta$), on the other hand, the CSI is always smaller than in the unbiased case. So underforecasting decreases CSI while overforecasting increases CSI. This possibility to manipulate the score is clearly an undesirable feature.

In the figure below a number of examples are given for various degrees of under- and overforecasting. The results for overforecasting, in steps of 20%, all originate at the origin and are confined to the area between the 45° line and the thick line for the random unbiased case. The CSI lines for underforecasting, also in steps of 20%, are parallel to their counterparts of opposite sign and are shifted to the right by $|\Delta|$. Therefore, relatively large penalties will be the result for underestimating the event (compared to overestimating), especially for small observed frequencies.



Here also

HKS = 0 and HSS = 0 (by definition), as well as the Equitable Threat Score (ETS) and the correlation coefficient. This can also easily be seen from the numerators of these scores (for instance, in the form $a - p_o p_f$).

The Rousseau Skill Score, on the other hand, is not always zero for arbitrary random forecasts. Substituting $a = p_o p_f$ yields

$$RSS = \frac{4 p_o p_f - (p_o + p_f)^2}{2(p_o + p_f) - (p_o + p_f)^2} = \frac{-(p_o - p_f)^2}{2(p_o + p_f) - (p_o + p_f)^2}$$

It follows therefore that $RSS \leq 0$ for random forecasts; only if $p_o = p_f$ $RSS = 0$. So predicting the observed frequency correctly is rewarded over predicting randomly with a prechosen frequency which is not correct.

Looking at biased forecasts we take again $p_f = p_o + \Delta$ with $-p_o \leq \Delta \leq 1 - p_o$. This leads to

$$RSS = \frac{-\Delta^2}{2(2p_o+\Delta)-(2p_o+\Delta)^2} = \frac{-\frac{1}{4}\Delta^2}{(p_o+\frac{1}{2}\Delta)-(p_o+\frac{1}{2}\Delta)^2}$$

which has a maximum of $-\Delta^2$ for $p_o+\frac{1}{2}\Delta = \frac{1}{2}$, i.e. for $\Delta = 1 - 2p_o$; this in turn equals to $p_o + p_f = 1$. In other words, the larger the amount of over- or underforecasting the lower the maximum obtainable value of the skill score. Moreover, these maximum scores are obtained only for situations in which the observed and predicted frequencies add up to one, but with the exception of unbiased forecasts in which the maximum RSS is always zero regardless of the sample climatology. In the figure below the expected Rousseau Skill Scores are given for a number of random forecasts with different amounts of over- and underforecasting. For unbiased forecasts ($\Delta = 0$) RSS is equal to zero for all values of the sample climatology. For increasing amounts of fixed underforecasting (in steps of 20 percent) RSS decreases with local maxima equal to minus the square of the underforecasting. These maxima are situated on the dashed parabola. The solutions in the diagram are of course confined to the right of the line for which $p_f = 0$. The figure is symmetric with respect to $p_o = \frac{1}{2}$. In this case the lines for fixed overforecasting are shown; these are confined to the left of $p_f = 1$. Minimum values of RSS of minus one are therefore reached if the predicted occurrence was 100% and the phenomenon never occurred and vice versa.

As can be seen from the figure, for fixed values of Δ minimum values of RSS are obtained for the highest possible over- or underforecasting, i.e. $\Delta = 1 - p_o$ and $\Delta = -p_o$ respectively. The values of RSS for these two cases are considered now in more detail.

If $\Delta = 1 - p_o$ (i.e. $p_f = 1$) :

$$RSS = \frac{-(1-p_o)^2}{2(1+p_o)-(1+p_o)^2} = \frac{-(1-p_o)^2}{(1+p_o)(1-p_o)}$$

$$= \frac{-(1-p_o)}{(1+p_o)} \quad (\text{with } p_o \neq 1)$$

For $p_o \uparrow 1$: $RSS \uparrow 0$.

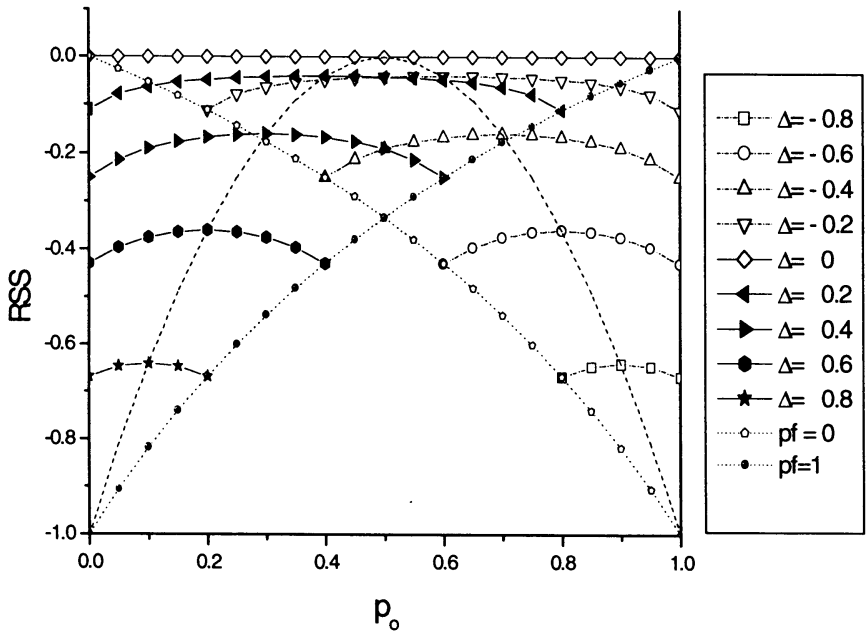
If $\Delta = -p_o$ (i.e. $p_f = 0$) :

$$RSS = \frac{-p_o^2}{2p_o - p_o^2} = \frac{-p_o}{2-p_o} \quad (\text{with } p_o \neq 0)$$

For $p_o \downarrow 0$: $RSS \uparrow 0$.

These two expressions, given by the dotted lines in the figure below, are symmetric with respect to $p_o = \frac{1}{2}$, as it should be. In the case of rare events the first expression of RSS is much smaller than the second one. The consequences in terms of RSS of a biased forecasting system therefore strongly depend on the sample climatology. For instance, in the case of low observed frequencies predicting the event always to occur is penalised much more severely

than never predicting it. Moreover, (randomly) overpredicting a rare event yields a higher, but still negative, RSS than (randomly) underpredicting it with the same amount. These are all very appealing properties of the RSS.



The model described in this subsection holds for arbitrary (categorical) forecast distributions. Random guessing with this pre-chosen p_f results in the above expressions of the scores. A frequently used reference forecast is the climatological distribution of the event, i.e. $p_f = p_{clim}$. By replacing p_f by p_{clim} in the above equations we find the typical dependence of the scores on the sample climatology in this no skill situation.

III.2. Fixed fractions predicted correctly

In the previous section we discussed the general features of a number of (skill) scores as a function of the observed frequency of the event in case there is no additional information available and the forecaster has to resort to random guessing after choosing a fixed frequency. This fixed frequency will usually be based on the forecaster's knowledge of the climatology of the particular predictand. In general this frequency will be (slightly) biased. Most of the scores show characteristic dependencies on the sample climatology which are also imminent in observed scores (see Fig. 1). In this section (as well as in section III.3) deviations from these features are examined more closely. We do that by introducing a few "model" assumptions and regard the relative changes with respect to the random case. Observed deviations can therefore be interpreted or expressed in terms of the model assumptions we discuss in this and the next section.

In the model in this section we assume that the correctly predicted events and nonevents are fixed fractions of the observed frequency. We examine the behaviour of the different scores as a function of this fraction. This fraction is denoted by p_m. It is sometimes called the *accuracy* of the forecasts.

In part A. we discuss the situation where the accuracy is the same for predicting the event and nonevent. This rather stringent condition may lead to evidently unrealistic results near the limits of the sample frequency. In part B. the more general case will be discussed.

A. Equal Fractions Correct for events and nonevents

Let us first consider the consequences of the assumption that the correctly predicted fraction p_m is the same for the occurrences and the nonoccurrences. Then the performance matrix can be written in the following form.

		FC: yes	FC: no	
O	yes	$p_o\, p_m$	$p_o\,(1 - p_m)$	p_o
B	no	$(1 - p_o)(1 - p_m)$	$(1 - p_o)\, p_m$	$1 - p_o$
S		p_f	$1 - p_f$	

The expression for the cells a and d follow from the definition and cells b and c follow then from the marginal frequencies. In general $b \neq c$ and therefore $p_f \neq p_o$. Note that given the sample climatology this system has only one degree of freedom.

FC = p_m (in accordance with the model assumption)

Bias ratio:

A direct consequence of this model is that (supposing $p_m \neq 1$) if $p_o > ½$ there is always underforecasting ($p_f < p_o$) of the event and similarly always overforecasting if $p_o < ½$. This can easily be seen by comparing cells b and c. Only for $p_o = ½$ and for $p_m = 1$ the bias ratio is equal to 1.

$$FAR = \frac{(1 - p_o)(1 - p_m)}{(1 - p_o)(1 - p_m) + p_o p_m}$$

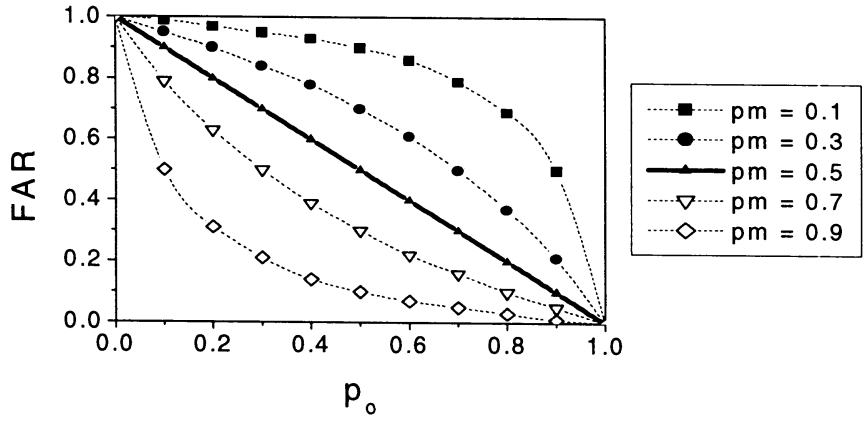
The behaviour of the FAR in dependence of the observed frequency of the particular event for different values of the correctly forecast fraction can be established more easily by dividing the numerator and denominator by $(1 - p_m)$, assuming $p_m \neq 1$,

$$FAR = \frac{1 - p_o}{1 - p_o\{1 - p_m/(1 - p_m)\}}$$

$$= \frac{1 - p_o}{1 - \alpha p_o}$$

$$\text{with} \quad \alpha = \frac{1 - 2p_m}{1 - p_m}$$

Alpha is a monotonously decreasing function of p_m running from 1 for $p_m = 0$ to $-\infty$ for p_m approaching 1. This yields for different values of p_m (or α) the following figure. For rare events the False Alarm Ratios are close to one for all values of the accuracy (see the examples below). For $p_m = 1$ the FAR is equal to zero (except when there are no events at all, in which case the FAR is undefined).



If the forecasts are correct in 50% of the cases (i.e. FC = $p_m = ½$) the FAR corresponds to the no skill (random) cases of the previous section, i.e. FAR = $1 - p_o$, regardless of the bias ratio. For higher correctly predicted fractions the curves are closer to the bottom left corner,

reminiscent to the observed case of the road temperature forecasts in the British Isles as reported by Halsey (1995): see Fig.1.
Since $p_m = ½$ is a very special case we will look at it a little more closely. The contingency matrix transforms to

		FC: yes	FC: no	
OBS	yes	$p_o \cdot ½$	$p_o \cdot ½$	p_o
OBS	no	$(1 - p_o) \cdot ½$	$(1 - p_o) \cdot ½$	$1 - p_o$
		½	½	

In other words, a 50% fraction correct in this model means that the forecast frequency is always 50%. From the contingency table it can also be inferred that the cells are the same as in a random case with this predicted frequency. This implies that the HKS, HSS and ETS are equal to zero. □

POD = p_m . Since we have (unrealistically) assumed complete independence of the correctly predicted fraction on the sample climatology we find no explicit relation with p_o.

POFD = $1 - p_m$

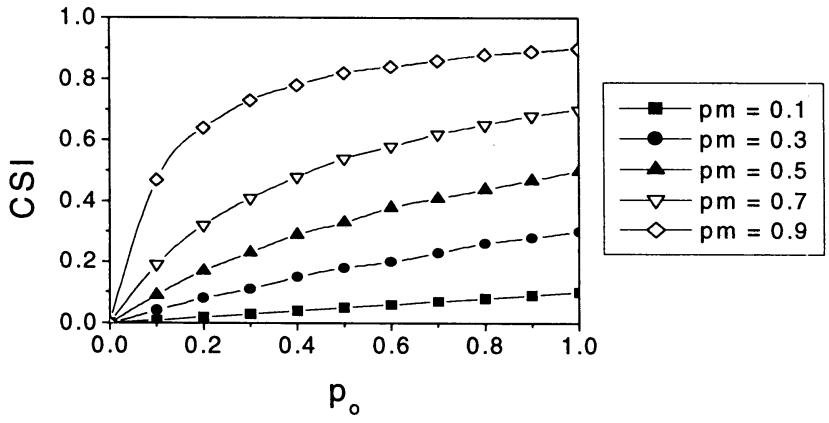
In other words, POD + POFD = 1.

$$CSI = \frac{p_o p_m}{(1 - p_o)(1 - p_m) + p_o} =$$

$$= \frac{p_o p_m}{1 - p_m + p_o p_m} = \frac{p_o}{p_o - 1 + 1/p_m} = \frac{p_o}{p_o + \alpha}$$

with $\alpha = \frac{1 - p_m}{p_m}$ (and assuming $p_m \neq 0$)

Alpha lies between zero and infinity, its limits are reached for $p_m = 1$ and $p_m = 0$, respectively. The dependence of CSI on the sample climatology is given for several values of p_m in the figure below. For $p_o = 1$ CSI = p_m, while for $p_m = 0$ and $p_m = 1$ CSI yields zero and one respectively, regardless of the sample climatology.



$HKS = 2p_m - 1$

with again $-1 \leq HKS \leq +1$, as it should be. A 50% accuracy of the forecasts yields a HKS of exactly zero. The value of the HKS is independent of p_o.

Since $a\,d - b\,c = p_o (1 - p_o) \{ p_m^2 - (1 - p_m)^2 \} = p_o (1 - p_o) (2p_m - 1)$ the Heidke Skill Score becomes

$$HSS = \frac{p_o (1 - p_o) (2p_m - 1)}{p_o (1 - p_o) (2p_m - 1) + \frac{1}{2} (1 - p_m)}$$

This is 1 if and only if $p_m = 1$, as it should. Not surprising, furthermore, is that HSS = 0 only if $p_m = \frac{1}{2}$ (apart from the trivial cases). The same is true for HKS. In comparing HKS and HSS it is helpful to examine the difference of the denominator of the two scores:

$$\begin{aligned} & p_o (1 - p_o) - p_o (1 - p_o) (2p_m - 1) - \tfrac{1}{2} (1 - p_m) \\ &= - p_o (1 - p_o) (2p_m - 2) - \tfrac{1}{2} (1 - p_m) \\ &= - 2p_o (1 - p_o) (p_m - 1) + \tfrac{1}{2} (p_m - 1) \\ &= (p_m - 1) \{ -2p_o (1 - p_o) + \tfrac{1}{2} \} \\ &= 2 (p_m - 1) (p_o - \tfrac{1}{2})^2 \end{aligned}$$

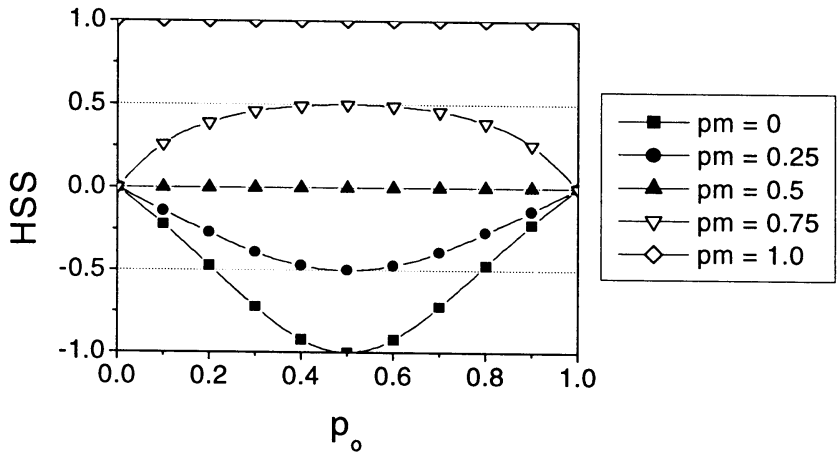
This is always < 0, except in the case $p_o = \frac{1}{2}$ (or when $p_m = 1$). Then it is equal to zero (i.e. HKS = HSS); see also chapter II.2. For $p_o = \frac{1}{2}$ also another condition under which the two scores are equal, is met. In that case the forecasts are always unbiased (b = c). Therefore in this model always $|HKS| \geq |HSS|$. The equal sign only holds in the trivial cases:

$p_m = 1$, then HKS = HSS = 1
(then also ETS = RSS = 1)

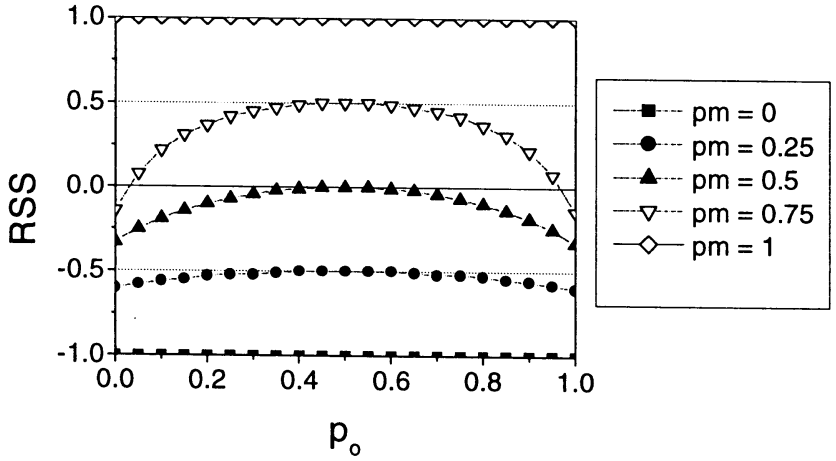
and if $p_o = \frac{1}{2}$ (i.e. b = c), then HKS = HSS = $2p_m - 1$.
(and also RSS = $2p_m - 1$)

At $p_o = 0$ or $p_o = 1$ then HSS = 0 except when $p_m = 1$. Therefore also ETS = 0 whereas RSS = $(p_m - 1) / (p_m + 1)$. For several values of the p_m the Heidke Skill Score is depicted in the

following diagram, once again as a function of the sample climatology. Remember that p_m is equal to FC in this model. The curve for FC = 0 we have seen already in section II.2. As stated above, for the same values of p_m the HKS are horizontal lines (independent of p_o) intersecting the HSS curves only at $p_o = ½$. These lines are also indicated in this figure.



For the same values of p_m the results for RSS are given.



Example 1.

This model has numerous applications. Especially in the case of extreme events this model has at first glimpse surprising consequences. Suppose for instance that 80 percent of the precipitation forecasts are correct and suppose precipitation occurs in 10% of the cases. We also assume that our forecasting accuracy is the same for dry and wet cases. By this we mean that our percentage correct is 80% for dry as well as for wet cases. In that case it follows that in 69% of the cases precipitation was predicted it remained dry. This can easily be inferred from the contingency table below.

The fraction of the rain forecasts in which it remained dry (the FAR) is 0.18/0.26 , i.e. 69%, despite the fact that 80% of the forecasts was correct. This is in correspondence with the FAR figure.

OBS \ FC	wet	dry	
wet	0.08	0.02	0.1
dry	0.18	0.72	0.9
	0.26	0.74	

□

Example 2.

A more extreme example is the following. Suppose that a HIV test is said to be 99% accurate, meaning that of all the HIV positive as well as the HIV negative people that are tested 99% is diagnosed correctly. Furthermore, suppose also that 0.1% of the entire population is HIV positive.
The expected contingency table looks like this:

OBS \ TEST	positive	negative	
positive	0.00099	0.00001	0.001
negative	0.00999	0.98901	0.999
	0.01098	0.98902	

It follows that a person who is found positive has a chance of around 91% of not being infected. This high value of the False Alarm Ratio also follows from the diagram.

□

The above two examples can also be regarded as applications of the two-event form of Bayes' theorem (Olson, 1965):

$$P(A \mid B) = P(A \cap B) / P(B)$$

(Here P stands for probability and A and B are arbitrary events)
In our case,

$$FAR = P(\text{observed} = \text{no} \mid \text{forecast} = \text{yes}) = \frac{P(\text{observed} = \text{no and forecast} = \text{yes})}{P(\text{forecast} = \text{yes})}$$

The examples show that an excellent forecast service in terms of fraction correct can be provided despite the fact that there is a large number of false alarms.

B. Fixed fractions correct

In this subsection there are no additional constraints, so these fractions may be different for the two categories.
The correctly predicted fractions for the event and nonevent are denoted by p_{m1} and p_{m2} respectively. According to the definition the probabilities in cells a and d are $p_o\, p_{m1}$ and $(1 - p_o)\, p_{m2}$ respectively. This leads to the following performance matrix:

		FC: yes	FC: no	
OBS	yes	$p_o\, p_{m1}$	$p_o\,(1 - p_{m1})$	p_o
	no	$(1 - p_o)(1 - p_{m2})$	$(1 - p_o)\, p_{m2}$	$1 - p_o$
		p_f	$1 - p_f$	

$POD = p_{m1}$ (see remarks under A.)

$POFD = 1 - p_{m2}$

$FC = p_{m2} + p_o\,(p_{m1} - p_{m2}) = p_o\, p_{m1} + (1 - p_o)\, p_{m2}$

These are straight lines crossing the left y-axis at p_{m2} and the right y-axis at p_{m1}. These straight lines only hold when the accuracy of the forecasts is not dependent on the sample climatology. But in general it is more realistic that it is (and also that it is dependent on the forecast frequency). In the case of rare events usually p_{m2} will be larger than p_{m1}, and smaller for very frequent events. This effect will induce a parabolically shaped FC-curve.

$HKS = p_{m1} + p_{m2} - 1$
With p_m is the average of p_{m1} and p_{m2} HKS becomes equal to $2p_m - 1$, which is equal to the expression found in subsection A. It's not dependent on p_o by assumption. Therefore this model can also be interpreted as the constraint of having a constant HKS (i.e. regardless of the sample climatology).

$$FAR = \frac{(1 - p_o)(1 - p_{m2})}{(1 - p_o)(1 - p_{m2}) + p_o\, p_{m1}}$$

Just as under A. we can obtain a more comprehensive form of the FAR by dividing numerator and denominator by $(1 - p_{m2})$, assuming $p_{m2} \neq 1$,

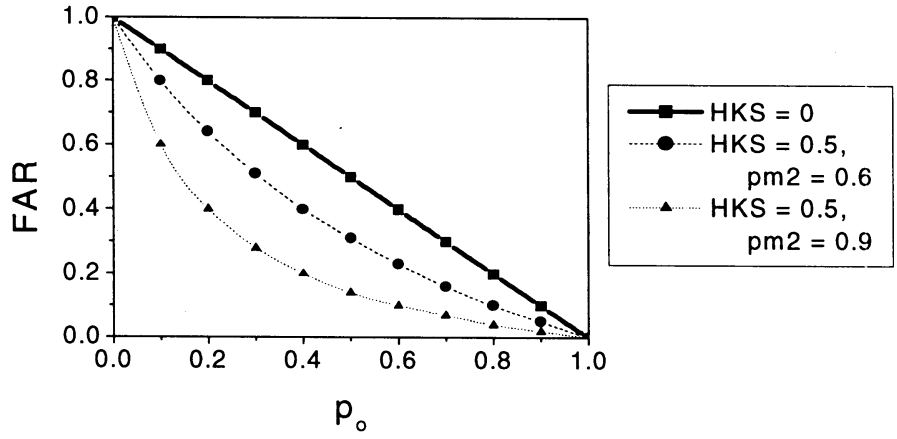
$$FAR = \frac{1 - p_o}{1 - p_o\{1 - p_{m1}/(1 - p_{m2})\}}$$

$$= \frac{1 - p_o}{1 - \alpha p_o}$$

This is the same as under A., but now with

$$\alpha = \frac{1 - p_{m1} - p_{m2}}{1 - p_{m2}}$$

The numerator is equal to minus HKS. Two lines for HKS = 0.5, i.e. $p_{m1} + p_{m2} = 1.5$, are shown in the figure below, one for which $p_{m2} = 0.6$ and $p_{m1} = 0.9$ and one the other way around. Obviously, it follows that the best False Alarm Ratios are obtained for the higher fraction of correctly predicted nonevents (i.e. a relatively small number of false alarms).



As mentioned before, in reality p_{m1} and p_{m2} will be functions of p_o in particular near the limits of p_o.

For completeness also the expression for CSI is given.

$$CSI = \frac{p_o p_{m1}}{1 - p_{m2} + p_o p_{m2}} = \frac{p_o}{(p_{m2}/p_{m1})\, p_o + (1 - p_{m2})/p_{m1}}$$

III.3. Skill over climatology

To assess the skill of a forecasting system it is customary to compare its results with the performance of some sort of reference system. One of the most frequently used references is climatology, since this is usually well-known for almost all weather phenomena. Using climatology as forecasts means that the event is forecast with climatological probability of occurrence, which in categorical forecasting is equivalent to (randomly) forecasting the event with a frequency which is equal to the climatological frequency. Since the observed number of events will generally not be the same as its expected number based on for instance the 30-year average, this usually results in a certain degree of under- or overforecasting. The effects on the behaviour of the verification scores of this so-called random guessing with predetermined (climatological) frequency have been discussed in section III.1.B. Most operational forecasting systems perform better than climatology and the verification scores sometimes look much better than those presented in that section.

To be able to quantify this improvement over climatology we will make use in this section of the skill score concept discussed in chapter II. There, skill was defined as the fractional improvement with respect to a reference forecasting system relative to what can be obtained in the case of perfect forecasts. This will be applied here to the events and nonevents separately. Therefore the diagonal elements of the performance matrix will be expressed as deviations from the expected values if for both events the forecast frequency is equal to the climatological one. These expected values are denoted by a_{clim} and d_{clim}. This leads to

$$a = a_{clim} + s_1 (p_o - a_{clim})$$
$$d = d_{clim} + s_2 (1 - p_o - d_{clim})$$

The respective forecasting skills for event and nonevent, s_1 and s_2, which may be different for the two categories, are

$$s_1 = \frac{a - a_{clim}}{p_o - a_{clim}}$$

$$s_2 = \frac{d - d_{clim}}{(1 - p_o) - d_{clim}}$$

If $s_1 = s_2 = 0$ then $a = a_{clim}$, $d = d_{clim}$ and $p_f = p_{clim}$, with p_{clim} defined as the climatological frequency. Perfect skills, i.e. $s_1 = 1$ and $s_2 = 1$, gives $a = p_o$ and $d = 1 - p_o$ (and $b = c = 0$), as it should be. In principle, s_1 and s_2 can be much smaller than minus one.

This section may help to interpret observed verification scores in terms of obtained skill over climatology for the event(s) (since after the fact p_o and p_{clim} are known). Vice versa, it may help to calculate the expected verification scores given the forecasting skills for the event(s) as a function of the observed and climatological frequencies. Again two cases will be considered. First the one that the observed frequency appears to be the same as the climatological frequency. This may be appropriate particularly when we deal with large sample sizes. Next (under B.) the more general case for arbitrary observed deviations from climatology will be discussed.

A. Observed frequency is equal to the climatological frequency

In other words, in this section we consider $p_o = p_{clim}$.
This gives

$$a_{clim} = p_o^2$$
$$d_{clim} = (1 - p_o)^2$$

Let's first look at the definitions of s_1 and s_2. Their denominators appear to be the same:

$$p_o - a_{clim} = p_o - p_o^2 = p_o (1 - p_o)$$
$$(1 - p_o) - d_{clim} = (1 - p_o) - (1 - p_o)^2 = p_o (1 - p_o)$$

For the probabilities a and d, expressed with respect to climatological probabilities, we get

$$a = p_o^2 + s_1 (p_o - p_o^2) =$$
$$= p_o \{ p_o + s_1 (1 - p_o) \}$$

$$d = (1 - p_o)^2 + s_2 \{1 - p_o - (1 - p_o)^2 \} =$$
$$= (1 - p_o) \{ (1 - p_o) + s_2 p_o \}$$

This notation shows that a and d are proportional to the observed marginals. Therefore this model is identical to the one in the previous section (III.2B.):

		FC: yes	FC: no	
OBS	yes	$p_o p_{m1}$	$p_o (1 - p_{m1})$	p_o
OBS	no	$(1 - p_o)(1 - p_{m2})$	$(1 - p_o) p_{m2}$	$1 - p_o$
		p_f	$1 - p_f$	

but now p_{m1} and p_{m2}, the correctly predicted fractions of the two events, are expressed in their respective skills over climatology, s_1 en s_2. Moreover, p_{m1} and p_{m2} are explicit functions of the sample climatology.

$$p_{m1} = p_o + s_1 (1 - p_o)$$
$$p_{m2} = (1 - p_o) + s_2 p_o$$

Note that if $p_o = 0$ then always $p_{m2} = 1$; likewise if $p_o = 1$ always $p_{m1} = 1$. The model assumptions do not hold for these two extremes. Therefore these extremes will not be regarded in this section.
Expressed in skills over climatology, s_1 and s_2, the matrix has the following form

		FC: yes	FC: no	
OBS	yes	$p_o \{ p_o + s_1 (1 - p_o) \}$	$p_o (1 - p_o)(1 - s_1)$	p_o
	no	$p_o (1 - p_o)(1 - s_2)$	$(1 - p_o)\{1 - p_o (1 - s_2)\}$	$1 - p_o$
		p_f	$1 - p_f$	

Note that, given the observed frequency, the system has only two degrees of freedom.
The predicted frequency can be written as

$$p_f = p_o + (s_1 - s_2)\, p_o (1 - p_o)$$

If both skills are zero then $p_f = p_{clim}$, in agreement with the model assumption. The same is true if the skills are equal. Then the predicted frequency is equal to the climatological frequency, which by assumption equals the observed one. Since most forecasting systems, verified on large sample sizes, approach this situation in which the three frequencies are the same, this special case will be discussed in more detail at the end of this paragraph.

It follows also that the bias ratio $B = 1 + (s_1 - s_2)(1 - p_o)$ is equal to 1 only if the skills are equal (apart from the trivial cases).

$$HKS = p_{m1} + p_{m2} - 1 = s_1 (1 - p_o) + s_2 p_o = s_1 + (s_2 - s_1) p_o$$

Without proof also the expression for HSS is given

$$HSS = \frac{s_1 + (s_2 - s_1) p_o}{1 + (s_2 - s_1)(p_o - \tfrac{1}{2})}$$

This is equal to HKS only in the unbiased situation and if p_o is 50%. This is in agreement with the discussion in section II.2. It follows that if in rare event forecasting a higher skill over climatology is obtained for the event than for the nonevent the HKS rewards this higher than the HSS. However, this is at the expense of a positive bias.

Consider the **special case that** $s_1 = s_2 =: s$
(This is only possible if p_o is not too close to zero or one.)

As we have seen this additional requirement results in an unbiased model. This can also be seen from the performance matrix (b = c). It also follows that

$$a - a_{clim} = d - d_{clim}$$

or $$a = d - 1 + 2p_o$$

Furthermore, a natural consequence of this model is that HKS is exactly equal to the skill:

$$HKS = s$$

This is also true for the Heidke Skill Score:

$$ad - bc = p_o(1 - p_o)\{p_o + s(1 - p_o)\}\{1 - p_o(1 - s)\} - p_o^2(1 - p_o)^2(1 - s)^2 =$$
$$= p_o(1 - p_o)\{ s + p_o(1 - p_o)(1 - s)^2 \} - p_o^2(1 - p_o)^2(1 - s)^2 =$$
$$= p_o(1 - p_o)\{ s \}$$

$$\tfrac{1}{2}(b+c) = p_o(1 - p_o)(1 - s)$$

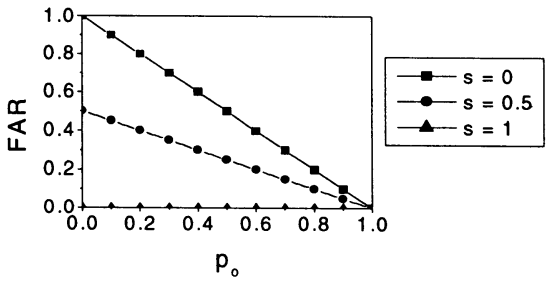
After dividing both numerator and denominator by $p_o(1 - p_o)$ the expression for HSS becomes:

$$HSS = \frac{s}{s + (1 - s)} = s$$

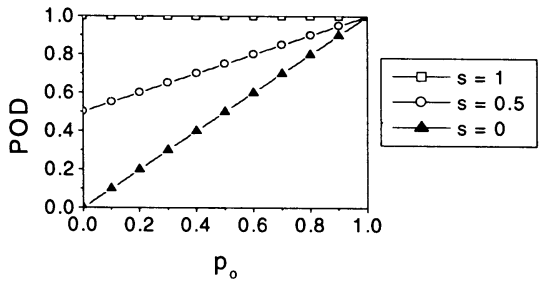
and therefore $ETS = s / (2 - s)$

So in this model always HSS = HKS. This is in agreement with the fact that for unbiased forecasts this is always true. Remember that in that case also the Rousseau Skill Score is equal to these scores.

$$FAR = \frac{p_o(1 - p_o)(1 - s)}{p_o} = (1 - p_o)(1 - s) = 1 - p_{ml}$$



$$POD = \frac{p_o\{p_o + s(1 - p_o)\}}{p_o} = p_o + s(1 - p_o) = p_{ml}$$

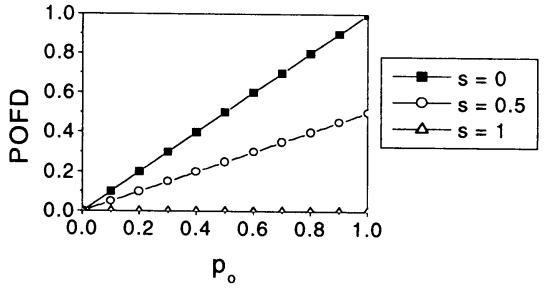


Comparing the above two figures with the results in Fig.1 it can be seen that for many stations or forecasters in the Britsh Isles the FAR corresponds with a skill over climatology of somewhere between 0.4 and 0.8 and for POD between 0.5 and 1.0. Note however, that there is

no information available from Fig.1 about the deviation of the local frost frequencies from their corresponding climatological values. The somewhat higher skill values for POD are in agreement with the fact that forecasters do not want to miss too many events and usually take a few extra false alarms for granted.

POD + FAR = 1 (always true for unbiased forecasts)

$POFD = p_o - s\,p_o = 1 - p_{m2}$



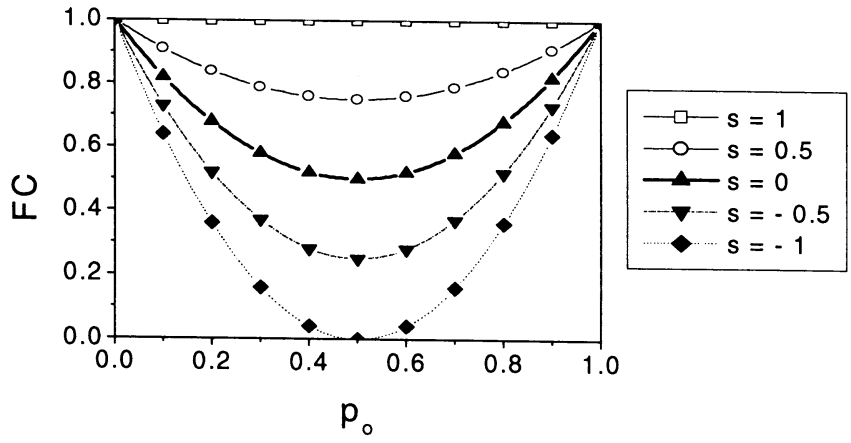
$POD = POFD + s$

$FC = p_o\{p_o + s(1 - p_o)\} + (1 - p_o)\{1 - p_o(1 - s)\}$

$= p_o^2 + (1 - p_o)^2 + 2\,s\,p_o(1 - p_o)$

$= 1 - 2\,p_o(1 - p_o)(1 - s)$

With $s = 0$ we get the random unbiased situation (chapter III.1A.), and for $s = 1$ it follows that $FC = 1$. The lowest possible value of s under the constraints in this section is minus one. This can only be reached for $p_o = 0.5$ (in which case $FC = 0$). For all other values of the sample climatology the minimum value of s is between 0 and –1. (In fact, this value is equal to the minimum of $-p_o / (1-p_o)$ and $-(1-p_o) / p_o$; not shown)



Comparing the above figure with the corresponding one in Fig.1 it appears that the forecasters' skill lies somewhere between 0.5 and 0.8.

B. Arbitrary observed frequencies

In general, the observed frequency is not equal to the climatological frequency in this subsection: $p_o \neq p_{clim}$. Therefore (with p_c short for p_{clim}),

$$a_{clim} = p_o\, p_c$$
$$d_{clim} = (1 - p_o)\,(1 - p_c)$$

Following the lines of the previous subsection we get for the probabilities a and d, expressed with respect to the climatological probabilities,

$$\begin{aligned} a &= p_o\, p_c + s_1\,(p_o - p_o\, p_c) = \\ &= p_o\,\{\, p_c + s_1\,(1 - p_c)\} \\ &= p_o\, p_{m1} \end{aligned}$$

$$\begin{aligned} d &= (1 - p_o)\,(1 - p_c) + s_2\,\{1 - p_o - (1 - p_o)\,(1 - p_c)\} = \\ &= (1 - p_o)\,\{\,(1 - p_c) + s_2\, p_c\,\} \\ &= (1 - p_o)\, p_{m2} \end{aligned}$$

$s_1 = s_2 = 1$ gives the maximum values again.
This makes the model analogous to the one in the previous subsection (and to the one in section III.2B.)

		FC: yes	FC: no	
OBS	yes	$p_o\, p_{m1}$	$p_o\,(1 - p_{m1})$	p_o
OBS	no	$(1 - p_o)(1 - p_{m2})$	$(1 - p_o)\, p_{m2}$	$1 - p_o$
		p_f	$1 - p_f$	

but now with slightly different correctly predicted fractions p_{m1} and p_{m2} . Expressed in their respective skills over climatology, s_1 en s_2 , they become:

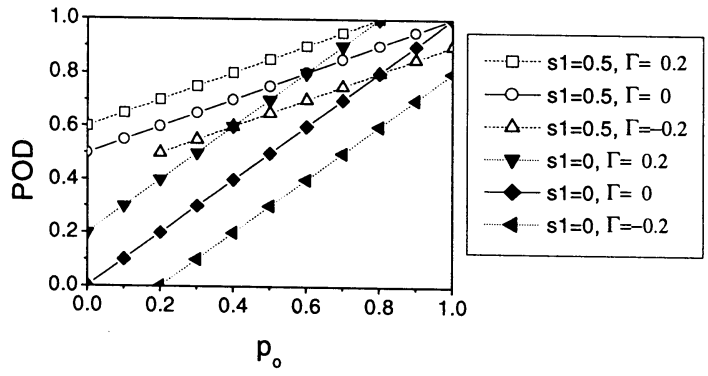
$$p_{m1} = p_c + s_1\,(1 - p_c)$$
$$p_{m2} = (1 - p_c) + s_2\, p_c$$

These expressions are exactly equal to the ones under A, but with p_o replaced by p_c . So there's no explicit relation anymore between these fractions and the sample climatology. But especially when the sample size is large, they will not differ much. Since the large similarity between this model and earlier examples only a few additional comments are given in this paragraph.
Expressed in skills over climatology, s_1 en s_2 , the performance matrix has the following form

		FC: yes	FC: no	
OBS	yes	$p_o \{ p_c + s_1 (1 - p_c) \}$	$p_o (1 - p_c) (1 - s_1)$	p_o
	no	$p_c (1 - p_o)(1 - s_2)$	$(1 - p_o)\{ (1 - p_c) + s_2 p_c \}$	$1 - p_o$
		p_f	$1 - p_f$	

$$POD = p_{m1} = p_c + s_1 (1 - p_c)$$

Suppose $p_c = p_o + \Gamma$, i.e. Γ is the deviation of the observed frequency from climatology.
In the figure below we see for two examples of the skill the POD for 3 values of Γ. The lines for $\Gamma = 0$ we have seen in section A. already. The effect of a large value of the climatological frequency, or rather, of a much below normal sample frequency, is quite pronounced and may obscure the differences in skill completely.



Similar arguments hold for the Probability of False Detection:
$POFD = 1 - p_{m2} = p_c (1 - s_2)$

$$\begin{aligned} HKS &= p_{m1} + p_{m2} - 1 \\ &= s_1 (1 - p_c) + s_2 p_c \\ &= s_1 + (s_2 - s_1) p_c \end{aligned}$$

Finally consider once again the **special case that $s_1 = s_2 := s$**

$$p_f = p_c + s (p_o - p_c)$$

$$HKS = s$$

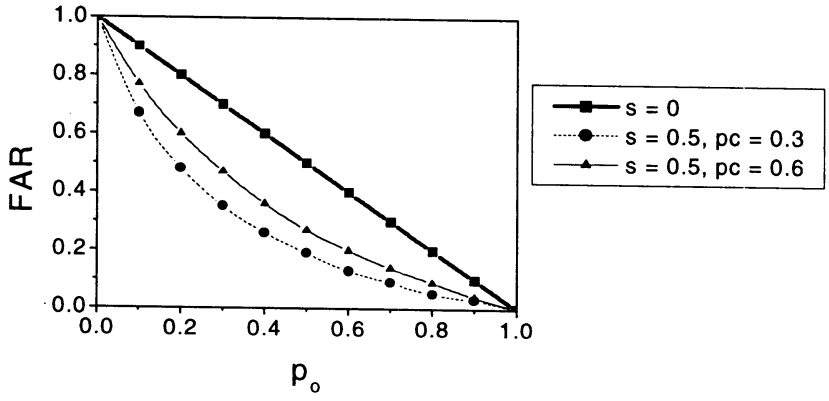
$$FAR = \frac{p_c\,(1 - p_o)\,(1 - s)}{p_c + s\,(p_o - p_c)} = \frac{p_c\,(1 - s)\,(1 - p_o)}{p_c\,(1 - s) + s\,p_o} = \frac{1 - p_o}{1 - \alpha\,p_o}$$

$$\text{with } \alpha = \frac{-\,s}{p_c\,(1 - s)}$$

The FAR has exactly the same form as in III.2A. and B. but with a different expression for α. The FAR $\rightarrow$ 1 as $p_o \downarrow 0$:

$$FAR \rightarrow \frac{p_c}{p_c\,(1 - s)}(1 - s) = 1$$

The last equation is true except for s = 1, in which case FAR = 0.
In the following figure for s = 0.5 the FAR is given for two examples of fixed p_c . The smaller this value the closer the FAR curve gets to the bottom left corner. Note however, that usually p_c will be rather close to p_o . For s = 0 the FAR is equal to 1 – p_o once again, regardless of the climatological frequency.



$$FC = p_o\,\{\,p_c + s\,(1 - p_c)\} + (1 - p_o)\,\{\,(1 - p_c) + s\,p_c\,\}$$

Suppose $p_c = p_o + \Gamma$
i.e. Γ is again the deviation of the observed frequency from climatology.

$$FC = p_o^2 + 2p_o\,\Gamma + 2s\,p_o + s\,\Gamma - 2s\,p_o^2 - 2s\,p_o\,\Gamma + (1 - p_o)^2 - \Gamma$$

$$= p_o^2 + (1 - p_o)^2 + 2\,s\,p_o\,(1 - p_o) + \Gamma\,(1 - s)\,(2p_o - 1)$$

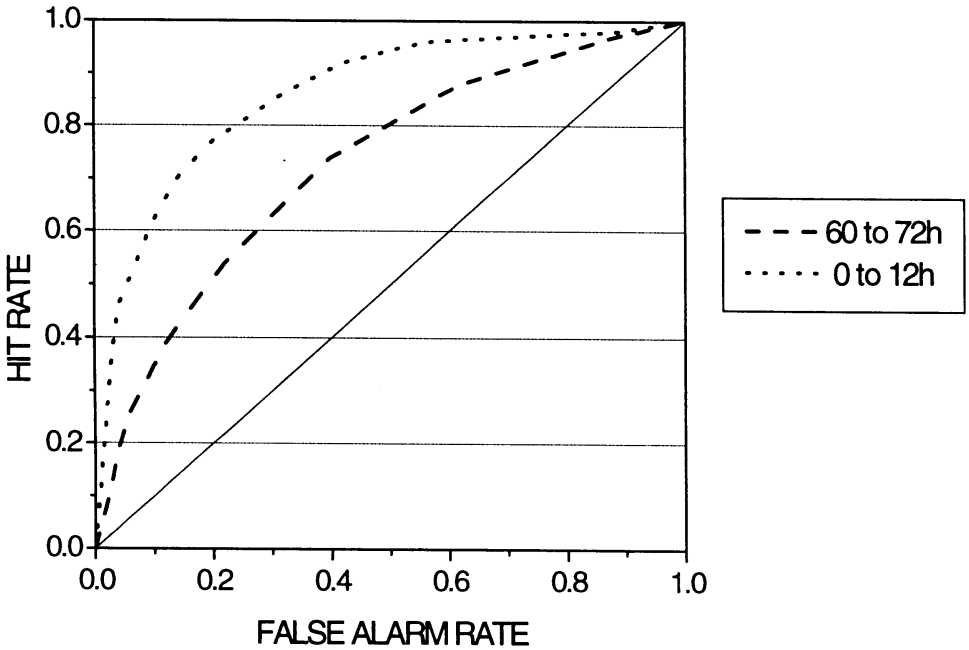
$$= 1 - 2\,p_o\,(1 - p_o)\,(1 - s) + \Gamma\,(1 - s)\,(2p_o - 1)$$

The first two terms of the last equation we have seen already under A. They are symmetric with respect to $p_o = ½$ and result in the figure presented in that section. The last term vanishes when $\Gamma = 0$ (case A.), s = 1, and also when $p_o = ½$. If $\Gamma > 0$ then this antisymmetric term

causes the lines of FC to tilt towards higher (lower) values for observed frequencies higher (lower) than 50%. If the phenomenon occurred more frequently than climatology it's the other way around. Note also the similarity of the above expression with the one for the random biased situation (section III.1.B), which gave rise to a similar behaviour of FC. As mentioned already, if $s = 1 \Rightarrow FC = 1$.

IV. Relative Operating Characteristic (ROC)

In the previous chapters we have looked at the behaviour of a number of verification scores mainly as a function of the observed frequency of the (binary) predictand. The last 15 years a new verification tool has been used which can be constructed from two of these scores. This new score is the Relative Operating Characteristic (ROC). It originates from the Signal Detection Theory (Swets and Pickett, 1982). It is sometimes also called the Receiver Operating Characteristic (see e.g. Swets, 1988) and was first introduced in meteorology by Mason (1980, 1982). This score is usually presented in the form of a diagram in which the 'Hit Rate' is plotted as a function of the 'False Alarm Rate'. An example, with scores given in percentages, is given in the figure below (adopted from Stanski et al., 1989). In this figure for two forecast periods (+12 and +72) a number of combinations of hit rates and false alarm rates are presented (and connected by straight lines). In this case each combination has been derived from a performance matrix which was determined by transforming probabilistic forecasts into categorical ones. Different categorical forecasts (and performance matrices) arise by applying different thresholds in the probabilistic forecasts associated with different degrees of uncertainty of the forecaster in predicting the occurrence of the event. In this way a 2xn contingency table, with n the number of forecast probabilities and 2 the yes or no occurrence of the predictand, can be reduced to a set of 2x2 contingency tables from which the scores can be calculated. We will come back to this later on.



For 0 to 12 h: ROC area 0.871
for 60 to 72 h: ROC area 0.728

The definitions of the hit rate and false alarm rate, however, are different from what is usually the case and which are used in the previous part of this report. To distinguish them from these previous scores they will be denoted between quotes in this chapter.

We use the same notations for the elements of the performance matrix as before:

		FC: yes	FC: no	
OBS	yes	a	b	p_o
	no	c	d	$1 - p_o$
		p_f	$1 - p_f$	

The 'Hit Rate' and 'False Alarm Rate' are defined as follows:

$$\text{'HR'} = \frac{a}{a+b} = \frac{a}{p_o}$$

$$\text{'FAR'} = \frac{c}{c+d} = \frac{c}{1-p_o}$$

These two measures imply a data stratification on the basis of the observations, in contrast to a stratification based on the forecasts in the earlier scores. I.e. the 'HR' is the fraction of correct forecasts of the event given that the event was observed. Similarly, the 'FAR' is the fraction of wrong forecasts of the event given that the event did not occur. So the hit rate as defined here ('HR') is the same as the probability of detection (POD) as it is normally used and as it is used in this paper. The 'FAR' was previously called (chapter II.1.C) the probability of false detection (POFD)[1]. Note that the FAR and 'FAR' can give rise to remarkably different behaviour under certain "model" assumptions as can be seen from the examples given in the previous chapter. But also with these "new" definitions of hit rate and false alarm rate one wants forecasting systems to yield hit rates as large as possible together with false alarm rates as small as possible. In other words, one would like to see values as close as possible to the top left of the diagram (see the previous page).

The diagonal in the ROC diagram is also called the *chance line* since random guessing yields equal hit and false alarm rates. This can easily be inferred from the performance matrix presented in section III.1.B. In this case both 'HR' and 'FAR' are equal to the forecast frequency.

Before going into more detail let's look at the expression 'HR' - 'FAR'.

[1] Some authors use false alarm *rate* when the stratification is on the basis of the observations and false alarm *ratio* when stratification is based on forecasts. We have adopted the same convention in this report.

'HR' - 'FAR' =

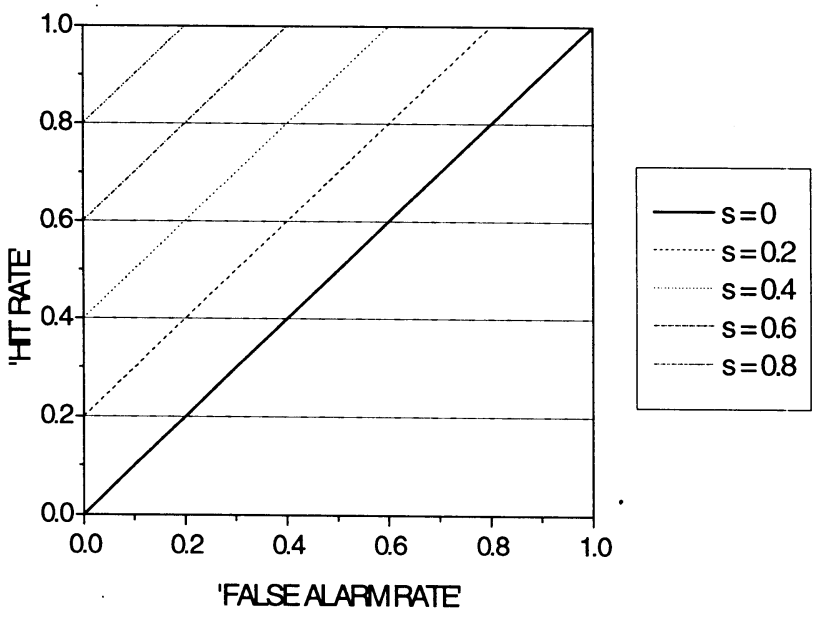
$$= \frac{a}{a+b} - \frac{c}{c+d} =$$

$$= \frac{a}{a+b} + \frac{d}{c+d} - 1 = \text{HKS}$$

with HKS the Hanssen-Kuipers score. In other words, the curves in the ROC diagrams follow the line

'HR' = 'FAR' + HKS

and the vertical distance from the curve to the diagonal is equal to HKS.
Because HKS is equal to 0 for forecasts without skill the diagonal in the ROC diagram can be regarded as the reference forecast with skill s = 0. For positive skill the points are above the diagonal, for negative skill the points are below the diagonal. If the elements of the performance matrix are expressed in terms of skill over climatology and the skill is the same for the cases that the event occurs and the event does not occur, i.e. the situation of the last section of the previous chapter, then HKS is exactly equal to that skill s . Lines with equal skill are indicated in the figure below. A value of s equal to 1 corresponds to the top left point of the diagram and s = -1 to the bottom right point (i.e. no hits and nothing but false alarms). Lines of negative HKS are not shown. If the performance matrix is not defined in terms of skill over climatology then the lines with equal HKS are not equal to s but, by definition, equal to $p_{m1} + p_{m2} - 1$, the sum of the correctly predicted fractions of the event and nonevent respectively, minus 1 (see chapter III.2).



Evaluation of categorical yes/no forecasts can be done on the basis of a 2x2 contingency table. Calculating the 'Hit Rate' and 'False Alarm Rate' gives only one point in the ROC diagram. The ROC diagram is used most frequently, however, in the analysis and verification of probabilistic forecasts. In order to do that a (generally large) series of probability forecasts has to be translated into categorical statements. The construction of a ROC diagram in the probabilistic case can best be illustrated by an example. Suppose you have a large set of forecasts for the probability of precipitation. Split these into 10 bins of equal width (0-10%, 10-20%, etc.). For each of the bins count the number of times precipitation occurred as well as the number of dry cases. This is similar to the construction of a reliability diagram. The translation from a probabilistic forecast into a categorical forecast is established by choosing a probability threshold below which you predict that the event does not occur and above which that it does occur. For instance, you could say that if the calculated probability of precipitation is 10% or less it remains dry, whereas for higher probabilities it will rain. For this particular threshold you can easily construct the appropriate 2x2 contingency table. The same procedure can be repeated for a threshold of, let's say, 20%, etc. In this way you arrive at 11 different contingency tables (the table for a threshold of 0% included), for which the hit rates and false alarm rates can be calculated. This gives 11 points in the ROC diagram, including the points (0,0) and (1,1) which are stemming from thresholds of predicted probabilities of precipitation of 100% and 0% respectively. The so-called ROC curve arises from connecting these points. By taking a larger number of thresholds off course a smoother curve will be obtained (assuming the data set is large enough). If a forecasting system yields a curve with points on or very close to the diagonal it means that the system cannot discriminate between occurrences and non-occurrences of the event. The higher the curve lies above the (no skill) diagonal, i.e. the higher the HKS, the better the forecasting system performs. A well-established measure of its performance is the area below the ROC curve.

By using this so-called *ROC area*, A_{ROC}, different forecast systems can be compared. We have seen already that this is true for categorical and probabilistic forecasts (leading to only one and to many points in the ROC diagram respectively). But the ROC area can also be used as a verification tool in deterministic forecasting. This is once again done by translating the deterministic forecast into a categorical statement. This can be quite straightforward when the predicted value can be identified uniquely with one of the categories. If, for instance, an atmospheric model gives .4 mm of precipitation then rain will be predicted, or if the predicted minimum temperature is 0.1°C then the categorical statement will be "no night frost". Verification of these essentially categorical forecasts, but now entirely based on a deterministic model, yields once again only one point in the ROC diagram. By connecting this point by straight lines to the corners (0,0 and (1,1), i.e. predicting rain or no rain regardless of the forecast, we can calculate the area under these lines. This area is directly related to HKS. In fact, the area under the curve but above the diagonal is equal to half the HKS. This can easily be inferred from geometrical inspection. In other words, $A_{ROC} = \frac{1}{2} + \frac{1}{2}$ HKS. Therefore, HKS can also be regarded as the relative improvement in A_{ROC} with respect to the area obtained in case of random guessing. In other words, the "ROC area skill score" is equal to HKS, i.e. in the definition of skill scores defined in chapter II:

$$HKS = \frac{A_{ROC}\,(\text{forecast}) - A_{ROC}\,(\text{random})}{A_{ROC}\,(\text{perfect}) - A_{ROC}\,(\text{random})} = 2\,A_{ROC} - 1$$

The direct relation between the ROC area and HKS is another reason for inserting equal HKS lines in the ROC diagram.[2]

In the two examples above (a precipitation of .4mm and a minimum temperature of 0.1°C), the deterministic forecasts were translated into categorical statements without incorporating the forecaster's faith in the deterministic model. It might be conceivable, however, that if the model gives a minimum temperature of 0.1°C the probability of night frost may be quite larger than in the case the forecast minimum temperature is, say, 10°C. This uncertainty can be incorporated in the translation from the deterministic statement into a categorical one. This can be done, analogous to what was done in the case of probabilistic forecasts, by choosing different thresholds to determine the choice of the category. These thresholds can be related to the degree of deviation from the deterministic value. For instance, a set of increasing amounts of precipitation can be used as thresholds above which rain will be predicted. Likewise, different thresholds for deterministic minimum temperatures can be applied: not only the zero degrees threshold but also for instance, +1, +2, etc. as well as -1, -2 degrees, etc. In this way the proximity of the deterministic forecast value to the value that divides the categories is used as a degree of the forecaster's uncertainty about the occurrence of the event. Therefore, similar to the probabilistic case, this leads again to different 2x2 contingency tables and, accordingly, to different points in the ROC diagram. In fact, for large enough numbers of cases a more or less continuous ROC curve can be obtained.

Although we have established that verification of deterministic, categorical and probabilistic forecasts can be expressed in the same quantity, A_{ROC}, one has to be extremely careful in interpreting the differences. One of the most obvious reasons is that the ROC area is rather sensitive to the number of thresholds that have been applied. Therefore, if translated to categorical statements in the usual way yielding only one point in the ROC diagram, a deterministic forecast system is likely to be underrated with respect to probabilistic forecasts. This can be overcome by applying the above procedure of incorporating the distance of the deterministic forecast to the boundary of the category. Due to the generally concave nature of the ROC curves for skilful forecasting systems all such systems, both deterministic or probabilistic, will benefit from taking as many thresholds as possible.

In rare event forecasting another undesirable property of the ROC diagram becomes apparent. In a large proportion of the cases it is very obvious that the event will not occur. Therefore, any forecasting system yields a high number of correct rejections (not forecast - not observed). This will be true for any set of thresholds, given the fact that bins of equal probability intervals are used. This means that a large number of ROC points lie in the left part of the ROC diagram: the more extreme the event the closer the points are to the bottom left corner. This means that the ROC area is to a relatively large extent determined by the point with the highest 'FAR' (disregarding point (1,1)). This makes the ROC area somewhat less appropriate for the verification of rare events.

A quantity that may partly overcome this disadvantage is related to the HKS. Instead of integrating HKS with respect to the 'FAR', yielding A_{ROC}, one can integrate HKS with respect to the threshold values. We will not elaborate on this.

---

[2] Moreover, in evaluating the economic value of a (probabilistic) forecast system HKS is also a very important quantity. It can be proven (Richardson, 1998) that the degree to which the mean expense of a user (which is dependent on his / her cost-loss ratio) can be minimized is related to the maximum value of HKS.

Reliable forecasts

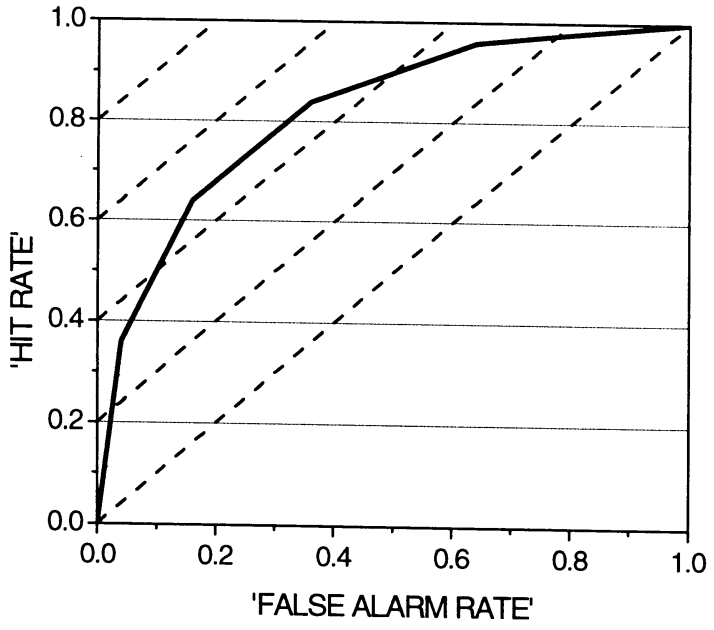
We have seen above what the relation can be between the position in the ROC diagram and the skill over climatology of a forecasting system. Another example which can serve as reference is the following. Suppose we have a set of perfectly reliable precipitation forecasts and, in addition, every forecast bin contains the same number of cases. Again we assume bins of equal width. These assumptions imply that the observed frequency of rain is 50%. By taking different thresholds and following the procedure outlined above we get the ROC points given in the figure below. These points, which are connected by straight lines, were calculated in this case by taking bins of 20%. It can be proven (see Appendix A.2) that under the above assumptions the points for arbitrary bins all lie on the following curve

$$\text{'HR'} = 1 - (1 - \sqrt{\text{'FAR'}})^2 \quad \text{or} \quad \text{'HR'} = 2\sqrt{\text{'FAR'}} - \text{'FAR'}$$

The area under this ROC curve is 5/6. The discretisation into bins of 20% leads in this case to a difference in ROC area with respect to the continuous curve of less than one percent.
Because 'HR' = 'FAR' + HKS

$$\text{HKS} = 2(\sqrt{\text{'FAR'}} - \text{'FAR'})$$

with a maximum value equal to 1/2 at a 'FAR' of 25%.
The area under the curve but above the diagonal line, which is equal to the integral of HKS over 'FAR', is 1/3.



Another interesting feature in this example follows from the expression for HKS. This reads (see Appendix A.2), with k the threshold probability ($0 \leq k \leq 1$):

$$\text{HKS} = (\tfrac{1}{2} - \tfrac{1}{2}k^2) / \tfrac{1}{2} + (k - \tfrac{1}{2}k^2) / \tfrac{1}{2} - 1 = 2k(1 - k)$$

(The derivation of this expression can be inferred from Appendix A.2. There, also a general expression for the performance matrix is given in the case that there are no constraints.)
Taking the integral of HKS over the entire range of thresholds leads to 1/3. This is exactly equal to the integral of HKS over 'FAR', as was mentioned above. So in this case the ROC area minus ½ , i.e. the area above the diagonal spanned by HKS, is equal to the area below HKS as a function of the threshold probability.

A more general example can be obtained by relaxing the above constraint that every bin has the same number of cases. It is more realistic to suppose a distribution over the bins in which the number of cases is a decaying function with increasing bin number. Suppose first a linear decline in the number of forecasts per probability interval, i.e. $N(k) \propto (1 - k)$, with N(k) the observed frequency as a function of bin number. This is again equal to the forecast frequency for each bin because of the reliability assumption. The observed frequency of the event is 1/6 in this example. The ROC area becomes 0.795.
If $N(k) \propto (1 - k)^2$, which is even more appropriate in the case of rare events, A_{ROC} is only 0.569 and the observed frequency becomes 1/12. In these two examples the integrals of HKS over the threshold probabilities is somewhat less than the corresponding integrals over 'FAR': 0.248 and 0.049 versus 0.295 and 0.069.

V. Final Remarks

In a paper by Finley in 1884 about the verification of an experimental tornado forecasting program he reported substantial skill in terms of the fraction correct of tornado versus no tornado forecasts. The inappropriateness of this measure was highlighted by many authors and led to a burst of verification-related papers in the last part of the nineteenth century. Many of the scores that were proposed in that time are still widely used; some of them have been "rediscovered", often more than once (Murphy, 1996). A number of these scores have been discussed in this report and some (by no means all) of their most important features have been reported. These features are extremely important in the interpretation of the verification results. It was not the intention to make a thorough comparison between the scores, nor to give guidelines as to which score(s) should be preferred in yes/no forecasting. Only a few general considerations have been given to make it easier for a researcher to base his / her choice upon.

The effects of a few "model" assumptions on a couple of widely used scores have been studied. First of all random forecasts with or without bias have been examined. The resulting values of the scores in dependence on the occurrence of the event can be regarded as reference level for these scores. Further model assumptions include the (prescribed) amount of skill over climatology and the relative frequency of correct forecasts of the event and / or the nonevent. The behaviour of the scores has been presented as a function of sample climatology. In this way observed scores, or a combination of scores, can be related to the outcome of these models for a given frequency of occurrences of the event in order to assess, for instance, the qualitative improvement with respect to climatology or with respect to random forecasts. Also the significance of differences in scores obtained at different stations (with different climatologies) or from different forecasting systems can be estimated. Therefore, the results found here for different model assumptions can be used to support the interpretation of verification results.

A more direct way of assessing the relative improvement with respect to some sort of reference forecast is the use of skill scores. Two of these scores, the Hanssen-Kuipers Score and the Heidke Skill Score, have been discussed more elaborately in this report, but also the Equitable Threat Score (mainly used in verification of precipitation forecasts) and the Rousseau Skill Score have been described. Again, no preference has been given but instead the consequences of several forecasting strategies have been outlined.

Categorical forecasts, the main subject of this report, can be obtained from for instance a deterministic atmospheric model. In that case uniquely one and just one category is chosen. But more often categorical forecasts are based on a decision process in which the forecaster arrives at his / her decision after a process of estimating the probability of the event of interest to occur. These estimates can be obtained with the help of for instance MOS equations or of an ensemble prediction system. Usually the critical threshold for translating judgements into categorical forecasts is simply the climatological probability, but also other values may be chosen. This is especially true in the prediction of very hazardous phenomena in which case the threshold should depend on the user's cost-loss ratio. When the loss incurred if no action has been taken is far greater than the costs that have to be made in order to prevent it, then the user will be inclined to take action already at a rather low probability of the event to occur, i.e. using a very low threshold, whereas the opposite is true when his / her cost-loss ratio is very high.

Applying these thresholds gives rise to yes/no forecasts which can be verified by using the 2x2 performance matrix. A natural extension can be obtained by taking a sequence of thresholds leading to different 2x2 performance matrices. By calculating the hit rate and false alarm rate for each of these matrices a so-called ROC curve can be constructed. The ROC area is a well-established verification measure with very appealing properties. The ROC area offers a tool to compare verification results of deterministic, probabilistic as well as categorical forecasts. For appropriate predictands the deterministic forecast yields very straightforwardly a categorical statement leading to a triangular area under the ROC curve. This means however, that the deterministic forecasting system is likely to be underrated with respect to a probabilistic one which yields many more points in the ROC diagram. Therefore it has been proposed here to choose a set of fixed distances to all deterministic values in the data set and calculate the hit and false alarm rates appropriate for each of these distances. This results in a larger number of ROC points. By applying this procedure the predictive skill of the probabilistic information which is contained in the deterministic model by means of the distance of the value of the predictand to the boundary of the category of interest (for instance night frost) is taken into account. This not only increases the ROC area for almost all deterministic forecasting systems it is also much more in agreement with the way deterministic models are generally used in practice.

Acknowledgements

I am indebted to Herman Wessels who pointed out to me the paper written by Halsey. Our discussions about the behaviour of the scores which were used in that paper to verify the minimum road temperature forecasts in the British Isles motivated the writing of this report. Seijo Kruizinga and Herman Wessels are thanked for their many helpful suggestions. They and also Hans Hersbach and Janet Wijngaard are thanked for their comments on earlier versions of this report.

Appendix A.1 Geometric interpretation of the contingency table

As we have mentioned in the discussion of the correlation coefficient, the contingency table can also be regarded as scatter plot between observed and forecast cases. By assigning a value of one to a yes forecast and to an observed event and zero otherwise you get four possible combinations:

(obs,fc) = {(1,1), (1,0), (0,1), (0,0)}

These pairs represent the elements of the 2x2 contingency table; in terms of the definition of chapter II.1 their numbers are A, B, C and D, respectively. Note that the positions of the elements of the contingency table are not the same in the scatter plot.

Now let's first take the linear regression line of the forecasts upon the observations. Following Hayes (1973) this is given by

$$\hat{F} = b_{F,O} (O - M_O) + M_F$$

in which O and F stand for observation and forecast, and M_O and M_F are their respective means. These means are equal to $p_o N$ and $p_f N$ respectively, in which N is the total number of cases. The slope of the regression line is

$$b_{F,O} = \frac{\mathrm{cov}(O,F)}{\mathrm{var}(O)},$$

Expressing the elements again in terms of fractions of the total number of cases, we get the notations which we apply in this report. The variance of the observations, var(O), is then

$$\mathrm{var}(O) = \Sigma (O_i - p_o)^2 = \underset{p_o}{\Sigma} (1 - p_o)^2 + \underset{1-p_o}{\Sigma} (- p_o)^2$$

$$= p_o (1 - p_o)^2 + (1 - p_o) p_o^2 = p_o (1 - p_o)$$

Similarly the covariance between observations and forecasts is

$$\mathrm{cov}(O,F) = \underset{a}{\Sigma} (1 - p_o) (1 - p_f) + \underset{b}{\Sigma} (1 - p_o) (- p_f) + \underset{c}{\Sigma} (- p_o) (1 - p_f) + \underset{d}{\Sigma} (- p_o) (- p_f)$$

where the summations are over the relative frequencies a, b, c and d of the contingency table. This gives

$$\begin{aligned}
\mathrm{cov}(O,F) &= a - a p_o - a p_f + a p_o p_f - b p_f + b p_o p_f - c p_o + c p_o p_f + d p_o p_f = \\
&= p_o p_f + a - (a + c) p_o - (a + b) p_f = \\
&= a - (a + b) (a + c) = \\
&= a - (a^2 + a b + a c + b c) = \\
&= a (a + b + c + d) - a (a + b + c) - b c = \\
&= a d - b c
\end{aligned}$$

Therefore the slope of the regression line is equal to the Hanssen-Kuipers Score. The linear regression of the forecasts upon the observations is in our notation

$$\hat{F} = HKS\,(O - p_o) + p_f$$

Similarly, the regression of the observations upon the forecasts looks like

$$\hat{O}_i = b_{O,F}\,(F - p_f) + p_o$$

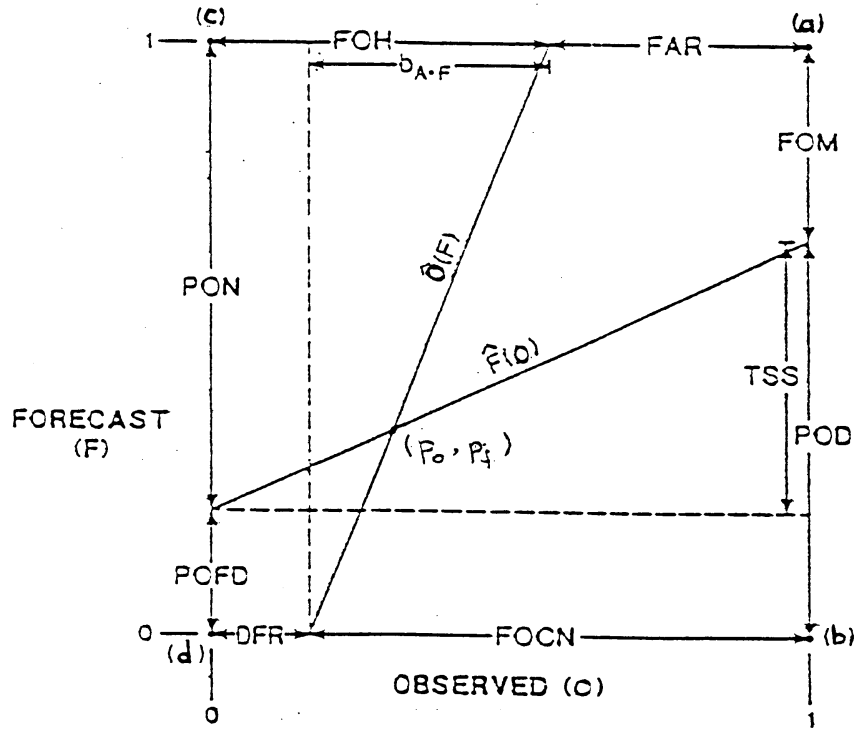
Here the slope is

$$b_{O,F} = \frac{cov(O,F)}{var(F)} = \frac{a\,d - b\,c}{p_f\,(1 - p_f)}$$

The correlation coefficient, which is equal to the covariance divided by the product of the standard deviations of the observations and the forecasts, therefore has the form which is given in chapter II.1.C:

$$r = \frac{a\,d - b\,c}{\sqrt{\{\,p_o\,(1 - p_o)\;p_f\;(1 - p_f)\,\}}}$$

The two regression lines are plotted in the scatter plot below. The intercepts of these lines with the OBS = (0,1) and FC = (0,1) axes define a number of quantities that are defined in chapter II.1. This figure is copied from Doswell et al.(1990).
(Remember that TSS, the True Skill Statistic, is equal to the Hanssen-Kuipers Score)



Appendix A.2 Proof of the expression of the ROC curve in the example in ch. IV.

Let's first consider the discrete case. Under the assumptions in this example (reliable, the same number of forecasts for all bins, equal bins of 20%) the performance matrix expressed in numbers instead of frequencies, looks like

		FC: yes	FC: no	
O	yes	$\sum_{i=k+1}^{5} p_i N_i$	$\sum_{i=0}^{k} p_i N_i$	$\sum_{i=0}^{5} p_i N_i$
B	no	$\sum_{i=k+1}^{5} (1-p_i) N_i$	$\sum_{i=0}^{k} (1-p_i) N_i$	$\sum_{i=0}^{5} (1-p_i) N_i$
S		$\sum_{i=k+1}^{5} N_i$	$\sum_{i=0}^{5} N_i$	$\sum_{i=0}^{5} N_i$

Here i is the bin number ranging from 1 to 5 and p_i is the relative frequency of the observed events for a particular 20% forecast interval. This frequency is completely determined by the reliability assumption. k is the threshold value, in terms of bin number, above which the event is predicted to occur. Finally, N_i is the total number of observations in bin i. In this example this is a constant: for every bin it is equal to 20% of the total number of cases N. We also allow k (and i) to be zero (with $N_0 = 0$) which corresponds to a threshold of 0%, i.e. always predicting the event to occur. The resulting six hit rates and false alarm rates, resulting from the six thresholds, yield the points in the ROC diagram which is shown in chapter IV.

In the continuous case and still making use of the reliability and the fact that we assume the same number of observations in each bin, the above matrix becomes as follows. Now k represents the threshold probability ($0 \leq k \leq 1$), which hereafter will be denoted as p_t.

		FC: yes	FC: no	
O	yes	$N\int_{p_t}^{1} pdp$	$N\int_{0}^{p_t} pdp$	$N\int_{0}^{1} pdp$
B	no	$N\int_{p_t}^{1} (1-p)dp$	$N\int_{0}^{p_t} (1-p)dp$	$N\int_{0}^{1} (1-p)dp$
S		$N(1-p_t)$	$N.p_t$	N

As was mentioned in chapter IV, the observed frequency of the events is 0.5. After evolving the matrix it looks as follows:

		FC: yes	FC: no	
O	yes	$N(\frac{1}{2}-\frac{1}{2}k^2)$	$N\frac{1}{2}k^2$	$\frac{1}{2}N$
B	no	$N(\frac{1}{2}-k+\frac{1}{2}k^2)$	$N(k-\frac{1}{2}k^2)$	$\frac{1}{2}N$
S		$N(1-k)$	$N.k$	N

The 'HR' and 'FAR' become 1 - k^2 and $(1 - k)^2$ respectively. Expressing the hit rate as a function of the false alarm rate gives therefore 'HR' = $2\sqrt{}$'FAR' - 'FAR' , which is the continuous curve in the ROC diagram.

□

Finally, for completeness, the general expressions of the elements of the performance matrix, so without the constraints of the example of chapter IV, are presented. They are similar to those of the performance matrix given at the previous page, but now the observed frequency N is a function of p and an arbitrary probability density function g(p) . This leads to the following matrix:

		FC: yes	FC: no	
O	yes	$\int_{p_t}^{1} N(p)g(p)dp$	$\int_{0}^{p_t} N(p)g(p)dp$	Np_o
B	no	$\int_{p_t}^{1} (N-N(p))g(p)dp$	$\int_{0}^{p_t} (N-N(p))g(p)dp$	$N(1-p_o)$
S		$N\int_{p_t}^{1} g(p)dp$	$N\int_{0}^{p_t} g(p)dp$	N

References

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1-3.

Daan, H. (1984). Scoring rules in forecast verification. Geneva, Switzerland, WMO Report, 60pp.

Dobryshman, E. M. (1972). Review of forecast verification techniques. WMO Tech. Note, No **120**, 17-20.

Donaldson, R. J., R. M. Dyer and M. J. Kraus (1975). An objective evaluator of techniques for predicting severe weather events. Preprints: *Ninth conference on severe local storms (Norman, OK)*, Amer. Meteor. Soc., Boston, 321-326.

Doolittle, M. H. (1888). Association ratios. *Bull. Philos. Soc.* Washington, **10**, 83-87, 94-96.

Doswell III, C. A., R. Davies-Jones and D. L. Keller (1990). On summary measures of skill in rare event forecasting based on contingency tables. *Wea. And Forecasting*. **5**, 576-585.

Epstein, E. S. (1966). Quality control for probability forecasts. *Mon. Wea. Rev.*, **94**, 487-494.

Finley, J. P. (1884). Tornado predictions. *Amer. Meteor. J.*, **1**, 85-88.

Flueck, J. A. (1987). A study of some measures of forecast verification. *Tenth Conference on Probability and Statistics in Atmospheric Sciences.* Oct. 6-8, 1987 Edmonton, Alta., Canada; Amer. Meteor. Soc. Boston, Mass.

Gandin, L. S., and A. H. Murphy (1992). Equitable skill scores for categorical forecasts. *Mon. Wea. Rev.*, **120**, 361-370.

Gerrity, J. P. jr. (1992). A note on Gandin and Murphy's equitable skill score. *Mon. Wea. Rev.*, **120**, 2709-2712.

Gilbert, G. F. (1884). Finley's tornado predictions. *Amer. Meteor. J.*, **1**, 166-172.

Glahn, H. R., A. H. Murphy, L. J. Wilson, J. S. Jensenius jr (1991). *Lectures presented at the WMO training workshop on the interpretation of NWP products in terms of local weather phenomena and their verification.* PSMP Reort no. 34.

Gringorten, I. I. (1967). Verification to determine and measure forecasting skill. *J. Appl. Meteorol.* **6**, 742-747.

Halsey, N. G. J. (1995). Setting verification targets for minimum road temperature forecasts. *Meteorol. Appl.* **2**, 193-197.

Hamill, T. M. (1998). Hypothesis tests for precipitation threat scores through resampling. *14th conference on probability and statistics in the atmospheric sciences.* Phoenix, Arizona, 11-16 Jan. 1998. p.17-19.

Hamill, T. M. (1999). Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. and Forecasting*. **14**, 155-167.

Hanssen, A. W., and J. A. Kuipers (1965). On the relationship between the frequency of rain and various meteorological parameters. *Mededelingen en verhandelingen* **81**. KNMI publ.

Harvey, L. O. jr., K. R. Hammond, C. M. Lusk, and E. F. Mross (1992). The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev.*, **120**, 863-883.

Hayes, W. L. (1973). Statistics for the social sciences. Holt, Rinehart, and Winston, Inc. 954pp.

Heidke, P. (1926). Berechnung des Erfolges und der Güte der Windstärkevorhersagen in Sturmwarnungsdienst. *Geogr. Ann.*, **8**, 310-349.

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

Mason, I. B. (1979). On reducing probability forecasts to yes/no forecasts. *Mon. Wea. Rev.*, **107**, 207-211.

Mason, I. B. (1980). Decision-theoretic evaluation of probabilistic forecasts using the relative operating characteristic. *WMO Symposium on Probabilistic and Statistical methods in weather forecasting.* Nice, September 8-12, 1980. 219-227.

Mason, I. B. (1982). A model for assessment of weather forecasts. *Aust. Met. Mag.*, **30**, 291-303.

Mason, I. B. (1989). Dependence of the Critical Success Index on sample climate and threshold probability. *Aust. Met. Mag.*, **37**, 75-81.

Murphy, A. H. (1973). Hedging and skill scores for probability forecasts. *J. Appl. Meteorol.* **12**, 215-223.

Murphy, A. H. (1996). The Finley Affair: a signal event in the history of forecast verification. *Wea. And Forecasting*, **11**, 3-20.

Murphy, A. H., and H. Daan (1985). Forecast evaluation. In: *Probability Statistics, and decision making in the atmospheric sciences*. Eds. A. H. Murphy and R. W. Katz, Westview Press, 379-437.

Murphy, A. H., and E. S. Epstein (1967). A note on probability forecasts and "hedging". *J. Appl. Meteorol.* **6**, 1002-1004.

Murphy, A. H., and R. W. Katz (1985). *Probability, Statistics and Decision making in the Atmospheric Sciences*. Westview Press, Boulder.

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

Olson, R. H. (1965). On the use of Bayes' theorem in estimating false alarm rates. *Mon. Wea. Rev.*, **93**, 557-558.

Palmer, W. C., and R. A. Allen (1949). Note on the accuracy of forecasts concerning the rain problem. U.S. Weather Bureau, 4pp.

Panofsky, H. A., and G. W. Brier (1958). Some applications of statistics to meteorology. Pennsylvania State University Park, 224pp.

Peirce, C. S. (1884). The numerical measure of success of predictions. *Science*, **4**, 453-454.

Richardson, D. S. (1998). Skill and relative economic value of the ECMWF Ensemble Prediction System. ECMWF Technical Memorandum No. 262.

Rousseau, D. (1980). A new skill score for the evaluation of yes/no forecasts. *WMO Symposium on Probabilistic and Statistical methods in weather forecasting*. Nice, September 8-12, 1980. 167-174.

Stanski, H. R., L. J. Wilson, and W. R. Burrows (1989). Survey of common verification methods in meteorology. Toronto, Ontario, Canada, Atmospheric Environment Service, Research Report No 89-5, 114pp.

Schaefer, J. T. (1990). The critical success index as an indicator of warning skill. *Wea. and Forecasting*. **5**, 570-575.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.

Swets, J. A., and R. M. Pickett (1982). Evaluation of diagnostic systems: Methods from signal detection theory. Academic Press, N. Y., 253pp.

Wessels, H. R. A. (1993). Meteorologische evaluatie van de zichtmetingen langs de A16. KNMI, Technische Rapporten TR-157.

Wilks, D. S. (1995). Statistical methods in the atmospheric sciences. Academic Press. 464pp.

Woodcock, F. (1976). The evaluation of yes/no forecasts for scientific and administrative purposes. *Mon. Wea. Rev.*, **104**, 1209-1214.