CHAPTER 11 — PROBABILITY FORECASTS

CHAPTER 11 — PROBABILITY FORECASTS

11.1 Basic concepts

Most forecasts give a categorical estimate of what various weather elements will be for a particular place/region and time/period. In reality uncertainty is inherent because:
(i) Observations do not provide a complete description of the state of the atmosphere.
(ii) Numerical models do not completely represent atmospheric processes (9.1).
(iii) Various assumptions are made in deriving expected weather from model forecasts.

The uncertainty can be implied by using such words as 'perhaps' (with the wide variations in meaning which can be attached to them) or expressed either qualitatively or quantitatively.

Probability forecasts are becoming more widely used because:
(i) They provide quantitative information for customers in uncertain situations.
(ii) They express inherent uncertainty in a precise and unambiguous manner.

11.1.1 Interpretation of probabilities

Probabilities can be interpreted in two ways:
(i) Relative frequency interpretation.
(ii) Subjective interpretation.

Thus consider a 'probability of precipitation (PoP) forecast of 30%':
(i) Relative interpretation: the present meteorological situation, observed on a large number of occasions, would give rise to precipitation on 30% of the time.
(ii) Subjective interpretation: the forecaster's judgement is that the odds against precipitation are 7 to 3 (odds against no precipitation being 3 to 7). Generally, if p is the probability, the odds against the event are:
$(1/p - 1)$ to 1.
(iii) The subjective interpretation gives a practical way of thinking about probabilities.

11.2 Types of probability measure

Three types of probability are in use:
(i) *Point probability*: probability that an event will occur at a particular point within a specified period of time.
(ii) *Average point probability*: the average point probability over a defined area.
(iii) *Area probability*: probability that the event will occur somewhere in the defined area within a specified period.

Point probability is easiest for interpretation and verification; average point probabilities, sometimes used for large areas by the media, can be misleading if there is a wide variation of point probability across the area.

The *area probability*, P_a, and *average point probability*, P_p, are related:

$$P_p = P_a\, a_c$$

where a_c is the proportion of the areal coverage if precipitation does occur.
(i) Note: $P_a \geq P_p$; and when precipitation is certain ($P_a = 1$), then $P_p = a_c$, i.e. the expected areal coverage of precipitation.
(ii) This area probability concept, P_a, can be very helpful when deriving P_p. Thus, if there is a 20% chance of precipitation reaching an area, $P_a = 0.2$, but if it does reach the area there will be precipitation everywhere (so that $a_c = 1$), then the average point probability, P_p, is 20%.
(iii) Similarly, if showers are certain in an area, $P_a = 1$, but if they do occur they will be scattered ($a_c = 0.2$ say), then again $P_p = 20\%$.

Conditional probabilities must be correctly identified and used.
(i) For example, if P(precip) is the probability of precipitation and P(precip|snow) is the conditional probability of snow (the probability of snow if precipitation occurs), then the probability of snow is given by:
P(snow) = P(precip) P(precip|snow).
(ii) The difference between P(snow) and P(precip|snow) is important; it is essential that the user knows which figure is being given.

11.3 Practical considerations

11.3.1 Determination of probabilities

Successful determination of probabilities depends upon the skill and experience of the forecaster. A few general points worth considering are given:

(i) Discussion between forecasters about probabilities is likely to be beneficial.

(ii) It may be useful to assess area probability of precipitation and conditional areal coverage separately before combining them to give the average point probability.

(iii) Although forecasters are likely to start with central guidance, in principle it is beneficial to make an independent assessment and then try to reconcile this with the guidance. In practice, there may not be sufficient time available to do this.

(iv) Most effort should be put into improving on the guidance for the early forecast period; later the value of local knowledge decreases rapidly.

11.3.2 Time period

A probability forecast must refer to a particular period. However, there are some pitfalls that must be avoided, as illustrated with the PoP forecasts:

(i) PoP can change discontinuously between periods so a continuous change should not be implied. Thus '80% chance of rain this evening *but only* a 30% chance tonight' is preferable to '80% chance of rain this evening *decreasing to* a 30% chance tonight'.

(ii) Periods should not be combined. Thus, '30% chance of rain today and tonight' is ambiguous. Does the 30% refer to each period separately or to them combined?

(iii) Do not use terms that leave the time period unclear, e.g. '20% chance of rain *by* this evening'.

(iv) Avoid using a period unhelpful or ambiguous to the user, e,g. 'late this evening'.

11.3.3 Incorporating probabilities into forecasts

The following are general guidelines for probability forecasts, especially PoP forecasts, to the general public (although not necessarily applicable to specialized services):

(i) Use 'chance' rather than 'probability' and avoid reference to 'threat of' or 'risk of'.

(ii) Give only one probability for each location. Thus '10% chance of showers this morning and a 60% chance of rain this evening' is to be avoided.

(iii) Do not combine probabilities about extent and duration. Thus, '30% of scattered showers' or '40% chance of occasional rain' should be avoided.

(iv) It is important that it is clear about the type of 'precipitation' to be expected.

(v) PoP should separate different types of precipitation, e.g. 'occasional rain with the possibility of an afternoon thunderstorm. 70% chance of rain'.

(vi) When a change of precipitation type is forecast, the PoP should refer to the chance of precipitation not the chance of the type changing, e.g. 'rain, possibly turning to snow this afternoon. 70% chance of precipitation'. Statements such as: '70% chance of rain turning to snow' should *not* be used.

11.3.4 Improving probability forecasts

Effective and timely feedback from a verification scheme can increase the reliability of probability forecasts. Common problems that arise when making probability forecasts are:

(i) Over-confidence, excessive use being made of very high/low probabilities.

(ii) Probabilities are not changed when a forecast is updated or even when developing as expected.

(iii) Range of probabilities available for a fixed period decreases with the length of the forecast. 100% PoP might be reasonable for day 1 of a forecast, but not for day 5.

(iv) Some probabilities are over- or under-used.

(v) There is a tendency to over-forecast probabilities for relatively infrequent events.

Reliability of forecasts (11.4.2.1) can be improved by identifying and remedying these problems. However, accuracy must not be compromised by artificially using 'under used' probabilities.

11.3.4.1 Characteristics of a reliable probability forecast

A reliable probability forecast is characterized by the following:

(i) When an event is infrequent at a particular location the probabilities tend to be lower than at locations where the event is frequent.

(ii) Probabilities tend to be lower the shorter the length of the period.

(iii) Probabilities tend to be less extreme as the lead time to the forecast period increases. For large lead times the range of probabilities reduces to the climatological frequency.

11.4 Verification

An assessment made of the extent of the agreement of a forecast with the actual state, using entirely *objective* methods, is termed an objective verification process.

11.4.1 Characteristics

The three characteristics of the verification process that are useful to assess are:
(i) Reliability.
(ii) Accuracy.
(iii) Skill.

(Another useful characteristic is 'factorization', discussed in 11.6).

However, before considering any summary measures from the verification scheme, it is important that the basic data are examined.

11.4.2 Display of information

Information about probability forecasts can be displayed in a contingency table. In the example a PoP of 0.4 was forecast on 264 occasions, with precipitation occurring 114 times and not occurring 150 times. The table also shows that a probability of 0.4 was forecast on 264 occasions out of a total of 8699 (this is referred to as the frequency of use).

Table 11.1. Contingency table

PoP	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Total
Precipitation	120	132	162	215	114	173	174	358	380	85	1172	3485
No precip	2619	739	492	398	150	168	126	211	132	102	77	5214
Total	2739	871	654	613	264	341	300	569	512	587	1249	8699

Often the climatological frequency of an event is not known. In this case the best estimate of the frequency can be derived from the contingency table. In the example the frequency of occurrence of precipitation was $P_o = 3485/8699 = 0.401$.

11.4.2.1 Reliability

The *reliability* is a measure of the degree of correspondence between the average of a set of forecasts and the corresponding average of a set of observations. The *bias* is one measure of reliability.

The forecasts are reliable (i.e. unbiased) if, for each forecast probability, the frequency with which an event occurs is the same as the forecast probability.

Consider the bias of the 0.4 probability forecasts. The contingency table shows that 0.4 was forecast on 264 occasions and of these precipitation was observed on 114 occasions. Therefore, on the occasions for which 0.4 was forecast, the observed frequency of precipitation was 114/264 = 0.432. This shows that there is a small bias in the 0.4 probability forecasts. The bias can be calculated in a similar way for each forecast probability.

A convenient way of displaying information about reliability is to use a reliability diagram (analogous to a scatter diagram) — a plot of frequency with which an event occurs when a particular probability is forecast against the forecast probability. Forecasts are perfectly reliable when points lie on a 45° line; points falling below/above the line indicate over-forecasting/under-forecasting **(Fig. 11.1(a))**

The frequency-of-use histogram is helpful in assessing whether the points on the reliability diagram are subject to significant sampling errors **(Fig. 11.1(b))**.

The frequency-of-use histogram is also useful in assessing whether the forecasts try to distinguish between different events — this property is called *sharpness.* The forecasts would be completely sharp if only probabilities of 0 and

1 were used. However, there would be no sharpness if a climatological probability was used because the forecasts would not distinguish between days with or without precipitation.

Note that if climatology is always used the forecasts would be perfectly reliable, but have no sharpness. In general forecasts should be reliable and quite sharp.

The reliability diagram is a very effective way of displaying information. However, it is convenient to be able to summarize the reliability with a single figure which represents the overall bias. There are two related measures that can be used:

$$\text{Bias} = [(\overline{P} - P_o)/P_o] \times 100\% \qquad \text{or Bias} = \overline{P}/P_o$$

where: $\overline{P}$ is the mean forecast probability and P_o is the frequency with which the event occurred.

(i) For the first measure, Bias = 0% for perfectly reliable forecasts, with positive values indicating over-forecasting and negative ones indicating under-forecasting.
(ii) For the second measure, Bias = 1.0 for perfectly reliable forecasts, with values greater than one indicating over-forecasting and values less than one indicating under-forecasting.
(iii) For the example, $\overline{P} = (2739 \times 0.0 + 871 \times 0.1 + ...)/8699 = 0.396$ and $P_o = 0.401$, giving a bias of -1.3% or 0.99 depending upon the measure used.

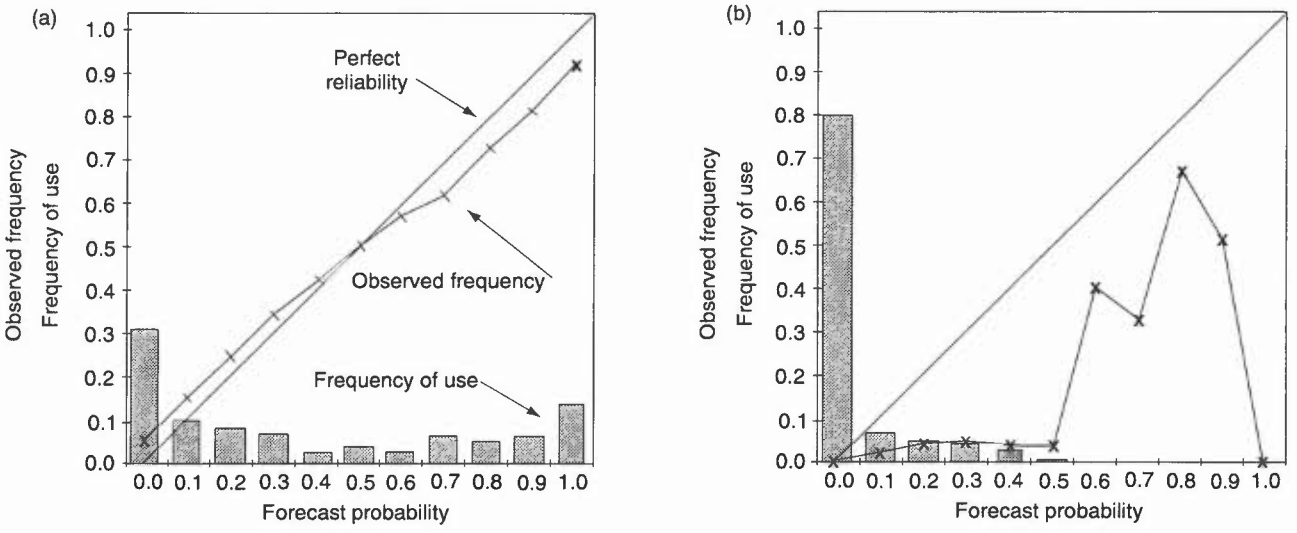

Figure 11.1. Reliability diagram and frequency of use histogram for (a) the PoP forecasts given in Table 1, and (b) for a set of forecasts of probability of fog.

11.4.2.2 Accuracy

Accuracy is defined as the average degree of correspondence between individual forecasts and what actually occurs.

The Brier Score is the most widely used measure of the accuracy of probability forecasts. It is a measure of the mean square probability error and is computed in essentially the same way as the Mean Square Error.

(i) First consider the case where the probability forecast refers to an event that falls into one of two categories (e.g. precipitation which either does or does not occur). PoP forecasts are of this type.
(ii) For a set of N forecasts, the Brier Score (BS) is given by:

$$BS = 1/N\,[\Sigma^N (F_i - O_i)^2]$$

where: F_i is the forecast probability of the event occurring.
O_i indicates whether the event occurred ($O_i = 1$ if it did and $O_i = 0$ if it did not).

With this scheme:
(i) $BS = O$ when all the forecasts are completely accurate (only probabilities 0 and 1 are forecast and $F_j = 0$ when $O_j = 0$ and $F_j = 1$ when $O_j = 1$).

(ii) $BS = 1$ when all the forecasts are completely wrong (only probabilities 0 and 1 are forecast and $F_i = 0$ when $O_i = 1$ and $F_i = 1$ when $O_i = 0$).

The Brier Score can easily be computed from the information used to produce the reliability diagram and the frequency of use histogram. The contribution to the Brier Score by forecasts of probability F is given by:

$$BS_F = \phi_F (F - 1)^2 + (1 - \phi_F)F^2$$

where: F is the forecast probability,
ϕ_F is the frequency the event occurred when probability F was forecast.

The overall Brier Score is then given by the sum of the individual contributions weighted by the frequency of use.

It should be noted that the Brier Score has three particularly desirable properties:
(i) Reliable forecasts are rewarded.
(ii) A willingness to discriminate between events is rewarded by penalizing forecasts in the mid-probability range (i.e. hedging).
(iii) The accuracy is maximized if and only if the forecaster makes a prediction that corresponds to his/her judgement about whether an event will occur.

The form of the Brier Score given above only applies when the event being forecast falls into one of two categories. If there are more than two categories (K say), the Brier Score for N sets of forecasts is given by:

$$BS = 1/2N \, [\Sigma^N \Sigma^K (F_{ij} - O_{ij})^2]$$

where: F_{ij} is the ith forecast for the jth category.
O_{ij} is the corresponding observation (1 if the event occurred and 0 if it did not).

When there are only two categories this expression reduces to the one given earlier.

The Brier Score has a negative orientation (the larger the score the lower the accuracy). If a positive orientation is required the Probability Score can be used:

$$PS = 1 - BS$$

11.4.2.3 Skill

The *skill* is a measure of the accuracy of a forecast relative to some reference such as persistence or climatology.

The *Probability Skill Score* (*PSS*) is based on the Brier Score of the forecasts (*BS*) relative to that based on climatology (BS_c)

$$PSS = (BS_c - BS)/BS_c$$

For a set of perfect forecasts $BS = 0$ giving $PSS = 1$, but if $BS = BS_c$ then $PSS = 0$.

One of the drawbacks of the *PSS* is its sensitivity to the accuracy of the climatological forecast. If the BS_c is small (i.e. the denominator in the expression for *PSS* is small) any difference between the forecast and climatological prediction is amplified. This makes the score rather unstable so it is essential that a large sample is used.

Ideally the value of the probability used as the climatological forecast should be the long-term frequency of the event. In reality this is often not available so the frequency of the event during the period of the forecasts is often used.

Suppose P_o is the observed frequency of an event. A forecast of P_o would lead to a contribution to the Brier Score of:
(i) $(1 - P_o)^2$ if the event occurred and the proportion of occasions on which this would happen is P_o.
(ii) $(0 - P_o)^2$ if the event did not occur and the proportion of occasions on which this would happen $(1 - P_o)$.
(iii) Therefore the reference Brier Score based on the observations from the period is given by:

$$BS_C = P_o (1 - P_o)^2 + (1 - P_o)P_o^2 = P_o (1 - P_o).$$

For the data in Table 11.1, $P_o = 0.401$, giving $BS_C = 0.240$. As $BS = 0.129$, the *PSS* is 0.463. Therefore the skill of the forecasts is 46.3% relative to climatology.

11.5 Making comparisons

Use of the Probability Skill Score is the most effective way of comparing two sets of probability forecasts. However, care still has to be taken in making such comparisons. For example, there is evidence that the skill of PoP forecasts sometimes depends upon the climatological frequency of precipitation, with the dependency being more marked in winter than in summer.

It has been argued that it would be expected the skill would be a maximum when the climatological frequency is about 50%, and approach zero as the climatological frequency approaches zero or 100%. These considerations suggest the following.

(i) Comparison of PoP forecasts from different stations should be avoided unless both sets of forecasts apply to regions with a similar frequency of precipitation.

(ii) Comparison for PoP forecasts produced by different forecasters at a station for the same location or area should produce meaningful results.

In both cases a large sample should be used in making a comparison. For example, for the comparison of individual forecasters about two years' worth of forecasts are required.

11.6 Factorization

11.6.1 Joint frequency distribution

Consider a set of PoP forecasts from the Central Forecasting Office. Information about the forecasts and observations ean be summarized in a contingency table.

Table 11.2. Contingency Table

PoP	0.0	0.1	0.2	0.3	0. 4	0.5	0.6	0.7	0.8	0.9	1.0	Total
Precip	120	132	162	215	114	173	174	358	380	85	1172	3485
No precip	2619	739	492	398	150	168	126	211	132	102	77	5214
Total	2739	871	654	613	264	341	300	569	512	587	1249	8699

The joint frequency distribution ean be derived by dividing each entry in the contingency table by the total number of forecasts. This distribution contains information about the forecasts, the observations and the relationship between them.

Table 11.3. Joint frequency distribution

PoP	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	Occurrence
Precip	0.014	0.015	0.019	0.025	0.013	0.020	0.020	0.041	0.044	0.056	0.135	0.401
No precip	0.301	0.085	0.057	0.046	0.017	0.019	0.014	0.024	0.015	0.012	0.090	0.599
Total	0.315	0.100	0.076	0.071	0.030	0.039	0.034	0.065	0.059	0.067	0.144	1.000

The joint frequency distribution can be factorized in two ways to give the conditional distributions.

11.6.2 Reliability factorization

Dividing each element in a column by the corresponding frequency of use gives the conditional distribution that indicates how often an event was observed when a particular probability was forecast. This is called the reliability factorization.

Table 11.4. Reliability factorization

PoP	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precip	0.044	0.152	0.248	0.351	0.432	0.507	0.580	0.629	0.712	0.826	0.938
No precip	0.956	0.848	0.752	0.649	0.568	0.493	0.420	0.371	0.258	0.174	0.062

This shows that on the occasions on which a probability of 0.4 was forecast, precipitation was observed with a frequency 0.432 (given by 0.013/0.030). This factorization is useful in assessing the reliability of forecasts — hence its name.

11.6.3 Likelihood factorization

Dividing each element in a row by the corresponding frequency of occurrence gives the conditional distribution that indicates how often an event was forecast with a particular probability when the event occurred (these are referred to as likelihoods). The whole process is called likelihood factorization.

Table 11.5. Likelihood factorization

PoP	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Precip	0.034	0.038	0.046	0.062	0.033	0.050	0.050	0.103	0.109	0.139	0.336
No precip	0.502	0.142	0.094	0.076	0.029	0.032	0.024	0.040	0.025	0.020	0.015

This shows that on the occasions on which precipitation was observed, a probability of 0.4 was forecast with a frequency of 0.033 (given by 0.013/0.401). This factorization is useful in assessing whether the forecasts discriminate between observed events (high values of forecast probability would be expected more often when an event occurs and low values when it does not).

Discrimination is the degree to which forecasts discriminate between occasions on which events occur. This can be assessed by examining the likelihood that particular probabilities were forecast when an event occurred. This information is provided by the likelihood factorization. A plot of the likelihood functions on a graph give a visual indication of the degree of discrimination.

For PoP forecasts there are two likelihood functions. The forecasts are not very discriminatory if, for each value of forecast probability, the likelihood functions are similar. However, if high probabilities are forecast when precipitation occurs and low values are forecast when it does not, there will be little overlap between the likelihood functions and the forecasts will be highly discriminatory — this is a desirable characteristic.

The likelihoods indicate the additional information provided by the forecast beyond that provided by a forecast based on climatology.

11.7 Ensemble forecasting and predictability

The current state of the atmosphere is never precisely known; the sensitivity of an NWP forecast to the adopted initial conditions varies from occasion to occasion. The *ensemble technique* involves running a large number of numerical forecasts (32 in the ECMWF scheme) to 10 or more days ahead, each with slightly differently perturbed initial conditions.

The perturbations are chosen to provide a good sampling of the major modes of error growth. The degree to which the individual members of the ensemble are consistent from day to day is, in turn, indicative of the amount of reliance that can be placed on the 'unperturbed' forecast.

The forecasts of 850 hPa temperature, total precipitation and 500 hPa height can be grouped into clusters, defined with respect to behaviour over selected regions and time intervals, to show the evolution of cluster average fields over 4 to 7 days. As a simple example of the application of the technique, the ensemble might divide into two clusters by day 6, which present in the one case (consisting of 60% of the whole) a strong likelihood of precipitation at day 6, the other group indicating no precipitation. The form of presentation of the forecast will depend on the customer and their requirements .

In one case the customer may only need to be told that there is a 'fair chance of rain towards the end of the week'; the other customer, with a specific interest in the outcome of rainfall, can be told that there is a 3 in 5 chance of rain at day 6.

Molteni, et al. (1996)

BIBLIOGRAPHY

CHAPTER 11 — PROBABILITY FORECASTS

Gordon, N., 1993: Verification of terminal forecasts. Am Meteorol Soc, Conference on Aviation Meteorology, Vienna (Virginia), USA.

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

Molteni, F., Buizza, R, Palmer, T.N. and Petroliagis, T., 1996: *QJR Meteorol Soc*, **122**, 73–119.

Murphy, A.H. and Katz, R.H., 1985: Probability, statistics and decision making in the atmospheric sciences. Westview Press (Boulder, Colorado).

Stanski, H.R., Wilson, L.J. and Burrows, W.R., 1989: Survey of Common Verification Methods in Meteorology. WMO. World Weather Watch Technical Report No. 8.

Wilks, D.S., 1995: Statistical methods in the atmospheric sciences. San Diego, Academic Press.