Diversity in Interpretations of Probability: Implications for Weather Forecasting

Ramon De Elia and Rene Laprise

Département des Sciences de la Terre et de l’Atmosphère, Université du Québec à Montréal, Montréal, Québec, Canada

(Manuscript received 11 February 2004, in final form 5 October 2004)

ABSTRACT

Over the last years, probability weather forecasts have become increasingly popular due in part to the development of ensemble forecast systems. Despite its widespread use in atmospheric sciences, probability forecasting remains a subtle and ambiguous way of representing the uncertainty related to a future meteorological situation. There are several schools of thought regarding the interpretation of probabilities, none of them without flaws, internal contradictions, or paradoxes. Usually, researchers tend to have personal views that are mostly based on intuition and follow a pragmatic approach.

These conceptual differences may not matter when accuracy of a probabilistic forecast is measured over a long period (e.g., through the use of Brier score), which may be useful for particular objectives such as cost/benefit decision making. However, when scientists wonder about the exact meaning of the probabilistic forecast in a single case (e.g., rare and extreme event), the differences of interpretation become important.

This work intends to describe this problem by first drawing attention to the more commonly accepted interpretations of probability, and then, the consequences of these assumptions are studied. Results suggest that without agreement on the interpretation, the usefulness of the probability forecast as a tool for single events—which include record-breaking events—remains unknown. An open discussion of this topic within the community would be useful to clarify the communication among researchers, with the public and with decision makers.

1. Introduction

Probability forecasting is today a common way of forecast delivery in major weather forecasting centers around the world. Its introduction, in a primitive way more than 100 yr ago, aimed to inform the public in a quantitative way of the forecaster’s uncertainty about the occurrence of a given event [see Murphy (1998) for an account of the early years]. Probabilistic forecasts are provided to users as a decision-making tool that can be used according to their needs. The uncertainty reflected in issuing forecasts in terms of probability instead of fact has proven to be both more scientifically sound—since it represents the actual state of the knowledge—and of more value for the users (Richardson 2000).

One of the disadvantages of switching from a deterministic to a probabilistic approach is in the difficulty of the interpretation of the latter. For example, what does it mean when a forecast states that there is an 80% probability of showers for tomorrow? The fact that this is not a trivial question is demonstrated by a series of studies carried out by A. H. Murphy, where he and other researchers wondered about the interpretations of probability arrived at by both forecasters and the public (Murphy and Winkler 1971a,b; Murphy et al. 1980). Results obtained by a questionnaire confirmed the lack of agreement among the public and among the forecasters themselves, and showed a variety of interpretations. They also found that probability forecasting usually involves ambiguity with respect to time and place as well as to the precise definition of the event, but these sources of misunderstanding, although important, will not be considered in this work.

Related to the problem of interpretation of probability forecasting is the problem of its verification against actual data. While a deterministic forecast for tomorrow at a given place may be easily verified, the verification of a probabilistic forecast represents a bigger challenge. For example, any person can provide a forecast for the probability of showers for tomorrow by just issuing a number between 0% and 100% without considering the characteristics of the atmosphere. This forecast, which is just a numerical representation of an unfounded opinion, is impossible to prove wrong. Therefore, in addition to the question of meaning, questions regarding quality of probabilistic forecasting are also relevant. This is a complex and important subject that has been the center of much research in the
last years (Katz and Murphy 1997; Jolliffe and Stephenson 2003).

These two main problems of probabilistic forecasting, that is, unclear meaning and complex verification, seem to be obstacles for its better understanding and its wider use. This work intends to highlight the richness of certain topics and to contribute to the clarification of specific issues, as well as to stimulate their discussion within the community. The way in which probability is understood may have consequences not only for the broadcasting of weather forecasts but also for the way the research is conducted.

Section 2 contains a discussion of the meaning of probability, where descriptions of each possible interpretation have been limited to the essential. Section 3 discusses some general properties of the Brier score, which has been chosen as the estimator of accuracy. Its properties are later used in section 4 to show that individual forecasts may differ even if produced by forecast systems of the same skill. The consequences of this in the interpretation of probability is extended to the probability distribution function (PDF) in section 5. In section 6 a simple numerical experiment illustrates some of the discussed topics and highlights the consequences of accepting the existence of a “true” probability distribution. A general discussion is presented in section 7.

2. The meaning of probability

The branch of mathematics usually called probability theory is characterized by a surprisingly late development in history. Its first numerical rudiments appeared in the late seventeenth century, although some claim that notions of probability were already present in the work of some ancient Greek philosophers and survived through the Middle Age in the work of some theologians (David 1962; Byrne 1968). At a time when other branches of mathematics already possessed a solid background, the basic theory of probabilities was in its infancy and resisted an unchallenged formalization and interpretation for years. Even today, a number of paradoxes in the theory, as well as highly contrasting interpretations, still exist. There are several hypotheses that intend to explain the relatively late development of probability theory as well as its resilience to a satisfactory interpretation. Despite the variety of existing hypotheses, it has become clear for most researchers that this is a particularly difficult concept to grasp (Franklin 2001).

Difficulties in the understanding of the concept of probability, however, did not stop its development and successful application in a vast variety of fields. The reason for this is that probability calculus is only a methodology for computing, from given probabilities, other probabilities that depend upon the former in a more or less complex manner. That is, the calculus of probabilities is just a manipulation of probabilities, whatever they mean, as long as the meaning of the prior and posterior probabilities remains the same. Poincaré put it this way: “Every probability problem involves two levels of study: the first—metaphysical, so to speak—justifies this or that convention; the second applies the rule of calculus to these conventions” (Byrne 1968).

There are several schools of thought regarding the interpretation of probabilities, and they intend to capture the meaning of probability in all its complexities. All of them, however, suffer flaws whose severity varies depending on the perspective of the individual users (Weatherford 1982). For this reason, there are authors who believe that probabilities cannot be defined in a unique way. They are of the opinion that, in the same way that the wave-particle duality exists in quantum mechanics, more than one interpretation of probability may have to be accepted simultaneously (Gillies 2000).

There are no standard classifications of probability interpretations, and even the more popular ones may suffer subtle variations from text to text. The limited list presented here has, therefore, a component of personal choice, but it is based on the belief that these have supporters in the atmospheric sciences. Some readers may find that their own view or understanding of probabilities does not exactly match any of the descriptions listed here. This may happen as a result of (i) the discretionary selection of the authors, (ii) the lack of discussion on the topic, which leaves room for intuitive and undisputed personal views, or (iii) the need for our community to develop its own interpretation of probabilities, given the particularities of the field.

In what follows, each interpretation is discussed in a context within weather forecasting that the authors think best facilitates its understanding (for readable, in-depth discussions, see the references mentioned herein). These interpretations are referred to in the upcoming sections, and much of the discussion relates to them. In addition, some of the developments challenge the solidity of these concepts.

a. Classical interpretation

The first rigorous research in the field of probability theory was carried out to understand the logic behind games of cards and dice. This now-called classical approach (Gillies 2000) is based on the “principle of indifference,” which proposes that each event is equiprobable (e.g., each face of a die due to symmetry conditions). The probabilities of the appearance of given combinations of numbers are estimated by the ratio of favorable cases against all possible cases (easily done now with combinatorial calculus). But what makes these computations useful is that the estimated probabilities coincide with frequencies of occurrence obtained by playing cards or dice in a large set of trials.

Among the fundamental limitations of this “classical” interpretation is the fact that, by just using an ir-
regular hexahedron instead of a cube as a die, the hypothesis of equiprobability among the faces becomes untenable. Assuming equiprobability would certainly lead to a biased estimate of the individual values. In addition, the classical interpretation suffers from what is usually known as “Bertrand’s paradox,” which can be described as a lack of uniqueness in probability values due to the fact that the equiprobability hypothesis is not always well posed (Weatherford 1982).

b. Frequentist interpretation

A way to solve the problem created by an irregular die is to estimate the probabilities of each face by throwing the die a large number of times and afterward studying the relative frequency of occurrence \( f \) of each face (that in a regular die would be \( f = \frac{1}{6} \) for each one).

For the holders of the frequentist interpretation, this is the only correct way of understanding a probability, since one cannot be sure of the fairness of a die until it is thrown a large number of times. That is, the computations of the classical theory should always await the confirmation of the experiment.

Hence, according to the frequentist viewpoint, the probability \( p \) is defined as

\[
p = \lim_{N \to \infty} \frac{N_f}{N},
\]

where \( N \) is the number of experiments, and \( N_f \) the number of times that the studied event did occur (Gillies 2000). This definition of probability is challenged by supporters of other interpretations as a confusion between a way of “measuring” a probability and the probability itself (Weatherford 1982).

The “frequentist” interpretation is very developed in sciences in general and a major player in the controversy about probability interpretation. In this way, the statement “there is 80% probability of showers for tomorrow” may be understood as “showers occur 80% of the times this is forecast” (which requires a large enough sample to obtain stable statistics). But actually, the definition of probability as frequency carries an ambiguity as to how the event is defined: the event could be “showers,” “summer showers,” “summer frontal showers,” “summer afternoon frontal showers,” etc. Hence the previous statement can be understood in a variety of ways. This is what is normally called the reference class problem, which says that there are several possible probability values for the same event, depending on the reference class in which the event is included. That is, various relative frequency of occurrence of showers can be obtained by using different subsets of the climatology [see also Simpson’s paradox in Suppes (1984)].

In addition to the reference class problem, two other characteristics of the frequentist interpretation usually come under attack. First, the infinite limit presented in (1) will never be reached, and even a finite large number would be problematic for nonstationary processes (Weatherford 1982). Second, probabilities defined in this way only allow for cases where the event is highly recurrent or has a large recorded history, leaving out a number of infrequent extreme cases.

c. Subjective interpretation

Although the frequentist interpretation has a large number of followers in applied sciences, many weather forecasters tend to think of probabilities in a “subjective” way. This view understands probability as a number that represents the degree of belief of a person about the occurrence of a given phenomena, implying that probability per se may not exist. Even though the subjective interpretation has a solid mathematical basis, it may not differ greatly from what some people call “opinion.”

Then, the statement “80% probability of showers” may be understood as saying that forecasters believe that there is high chance of rain. But why 80% and not 70%? One of the difficulties of this viewpoint is that there is no objective tool to measure the degree of belief of a person, and the quantification in numbers seems a bit arbitrary. Although controversial, many researchers hold the view that betting intentions are a way of measuring the degree of belief of a person in the occurrence of an event. This may be done by finding the probability at which the person will bet either in favor of or against the occurrence of an event with the same confidence. For example, if the forecaster believes that the probability of showers is 80% he should be as comfortable betting 4$ against 1$ that rain occurs as betting 15 against 4$ that rain does not occur. To find his true belief, the forecaster should find the number at which either bet would seem fair. Some researchers believe that this method of estimating probabilities is too sensitive to the character of the gambler (either conservative or risky) to be taken seriously, but nonetheless its use is widespread (Gillies 2000).

A distinctive characteristic of this interpretation of probability is that different persons may have conflicting beliefs about the likelihood of occurrence of an event, and all are equally valid in principle (although we will tend to believe whoever has the best record with respect to previous forecasts). It is also highly possible, as is the case with other subjective appreciations, that even the experts would be somewhat affected by non-rational behavior (Stewart 2001) and cognitive illusions (Nicholls 1999). The main advantage of this interpretation is the fact that it allows predictions of situations that have no previous record.

d. Propensity interpretation

In this approach, probabilities are thought of as an objective property of nature, a consequence of the physics of the problem under study. For example, the probabilities of showers for tomorrow may have an ob-
The propensity interpretation shares with the frequentist approach the idea that a probability exists, but it adds the interesting point that probabilities for single, individual events (“true” probabilities) exist as well. This means that the probability is not just a statistical property that is satisfied in the long term, but that it gives the best information regarding the likelihood for each individual case. It contrasts with the other interpretations discussed here for which probabilities are the result of either a pledge of uniform ignorance (classical), a statistical property (frequentist), or a subjective appreciation (subjective). “True” probabilities, if they exist, would provide a better decision-making tool, unrelated to intuition and to incomplete knowledge. This point of view seems to fit well the probabilities in the case of games of dice, but its appropriateness in weather forecasting seems unclear.

For some the “propensity” interpretation is linked with the concept of causality, and hence probabilities are just the degree to which an event is determined by its causes. Because of this, Gillies (2000) is of the opinion that “true” probabilities have less justification in applied sciences than in games of chance, the reason being that, the more complex the physical system is, the higher the possibility that extraneous factors may affect its value. This is not far from claiming that the “true” probability of an event is conditional to the state of the universe. For this reason some researchers believe that a “true” probability is a metaphysical concept devoid of meaning, since it can never be proved if it is attained or not (Gillies 2000).

This interpretation seems to be favored by a number of researchers in the atmospheric sciences. Several hints suggest that this understanding could have originated in the foundational texts of the ensemble forecasting community (e.g., Epstein 1969; Ehrendorfer 1994).

Table 1 summarizes the main characteristics of the theories discussed in this section.

<table>
<thead>
<tr>
<th></th>
<th>Classical</th>
<th>Frequentist</th>
<th>Subjective</th>
<th>Propensity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Main hypothesis</strong></td>
<td>Principle of indifference</td>
<td>Frequency of occurrence</td>
<td>Degree of belief</td>
<td>Degree of causal connection</td>
</tr>
<tr>
<td><strong>Conceptual basis</strong></td>
<td>Hypothetical symmetry</td>
<td>Past data and reference class</td>
<td>Knowledge and intuition</td>
<td>Present state of system</td>
</tr>
<tr>
<td><strong>Conceptual approach</strong></td>
<td>Conjectural</td>
<td>Empirical</td>
<td>Subjective</td>
<td>Metaphysical</td>
</tr>
<tr>
<td><strong>Single case possible</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Precise</strong></td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Problems</strong></td>
<td>ambiguity in principle of indifference</td>
<td>Reference class problem</td>
<td>Unchecked opinion</td>
<td>Disputed concept</td>
</tr>
</tbody>
</table>

3. The Brier score

In this section, the Brier score (Brier 1950) is developed and discussed from the point of view of some probability interpretations. The propensity interpretation, for which “true” probabilities exist for single-case events, is an interesting starting point for the development of the mathematical expressions.

As discussed earlier, the existence of a “true” probability of occurrence is by no means a proven fact. However, postulating the existence of the “true” probability $p_r$, may be useful for developing and understanding concepts of probability, in the same way that complex numbers are used fruitfully to solve equations in real numbers. Accepting the existence of a “true” probability $p_r$, allows the interpretation of any forecast probability $p$ as an estimation of $p_r$. In such a case, the usual tools for the study of random variables become available.

Figure 1a shows a hypothetical scattergram of $p_r$ against $p$. Under this condition, a perfect probabilistic forecast and verification system will be achieved when all the dots lie on the diagonal (to be distinguished from a perfect deterministic forecast system, which can be thought of as a probabilistic system that only produces $p_r = p = 0$ or $p_r = p = 1$).

a. Standard decomposition

A measure of the quality of the estimated probability $p$ can be obtained by using a mean square difference defined as

$$
\text{MSD} = \int_0^1 \int_0^1 (p - p_r)^2 Q(p, p_r) \, dp \, dp_r,
$$

(2)

where $Q(p, p_r)$ is the joint distribution of the estimated probability $p$ and the “true” probabilities $p_r$. This expression, using the definition of conditional probabilities (e.g., Papoulis 1991), can be rewritten as

$$
\text{MSD} = \int_0^1 \left[ p^2 - 2p \int_0^1 p, Q(p, p_r) \, dp_r \right] Q(p) \, dp + \int_0^1 p_r^2 Q(p, p_r) \, dp_r \, Q(p) \, dp.
$$

(3)
where $Q(p_i|p)$ is the distribution of “true” probabilities conditional to the estimated probabilities, and $Q(p)$ is the distribution of estimated probabilities. Since the variance of a variable $x$ can be expressed as

$$\sigma^2_x = \int x^2 Q(x) \, dx - \left[ \int x Q(x) \, dx \right]^2 = \bar{x}^2 - \bar{x}^2,$$  

(4)

and since the second and third integrals in (3) can be interpreted as conditional averages, the expression can be rewritten as

$$\text{MSD} = \int_0^1 [p^2 - 2p\bar{p}_{r|p} + \sigma^2_{p|r|p}] Q(p) \, dp,$$  

(5)

and finally as

$$\text{MSD} = \int_0^1 [(p - \bar{p}_{r|p})^2 + \sigma^2_{p|r|p}] Q(p) \, dp,$$  

(6)

where $\bar{p}_{r|p}$ is the mean of the “true” probability $p$, conditional to the estimated probability $p$, and $\sigma^2_{p|r|p}$ is the variance of the “true” probability conditional to the estimated probability $p$.

This expression could be thought of as an index of quality for probabilistic forecast, but in reality it is of limited practical use since $p_i$ is not known, assuming it even exists. The only information that may be available is the outcome of the events that are being forecast. In this case, $p_r$ would become either 1 or 0 according to whether or not the event has occurred after the prediction time is over. Figure 1b illustrates the collapse of the “true” probability on the scattergram of Fig. 1a after the information relative to the occurrence or not of the events has been received. Given the fact that under this condition $p_r$ may only take values 0 or 1, $p_r^2$ can be replaced by $p_r$ in the third term in (3), and after a few algebraic manipulations

$$\text{BS} = \int_0^1 [(p - \bar{p}_{r|p})^2 + \bar{p}_{r|p}(1 - \bar{p}_{r|p})] Q(p) \, dp,$$  

(7)

where now BS refers to the Brier score. This final expression has been derived from (2) assuming continuity for the variable $p$, and assuming that $p_i$ is either zero or one. Therefore, it can be thought of as a continuous version of the discrete Brier score that is normally defined as

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

(8)

where $N$ is the number of forecast events, $p_i$ the estimated forecast probability of each event $i$, and $o_i$ a function that takes the value $o_i = 1$ or $o_i = 0$ depending on the occurrence or not of the mentioned event, respectively.

It can be shown by comparing (6) and (7) that $\text{MSD} \leq \text{BS}$, the equality being only possible when the problem is thought of as deterministic ($p_r = 0$ or $p_r = 1$). In the propensity interpretation of probability, MSD represents the error of the forecasting system while BS includes an additional contribution for the uncertainty that is intrinsic to the weather dynamics. That is, for those holding that “true” single-case probabilities exist, and that they are different from 0 and 1, the BS overestimates the error of the forecast system as a tool. For other interpretations that do not postulate the existence of $p_r$, the BS represents the uncertainty of the forecast system (joining together predictability issues and model errors).
The expression $\tilde{p}_{\text{pr}}$, in (7), can be interpreted as the relative frequency of occurrence of events $f_p$ for a given forecast probability $p$ (i.e., the number of times that the event actually took place over the number of cases predicted with probability $p$). To simplify the notation, expression (7) may be rewritten as

$$BS = \int_0^1 (p - f_p)^2 Q(p) \, dp + \int_0^1 f_p(1 - f_p)Q(p) \, dp.$$  (9)

This decomposition of the Brier score was first introduced in discrete form by Murphy (1972), with a different approach from the one presented here.

The first term in (9), usually called reliability, measures the difference between the forecast probability $p$ and the relative frequency $f_p$. A forecasting system is said to be reliable if this term is nil or calibrated if this term is forced to be zero at a postprocessing stage, by changing the actual forecast probability value $p$ to $f_p$. In such a system, a given event with a forecast probability $p$ of occurring will eventually occur over a large number of forecasts with a frequency $f_p \sim p$ (i.e., when reliable forecasters predict 80% probability of showers, they will occur in around 80% of the cases). An introductory discussion of this topic and related references can be found in Wilks (1995).

Recalling the discussion of section 2b, this property is the basis of the definition of probability in the frequentist viewpoint. It is worth mentioning that the definition of the Brier score does not presuppose that probabilities are reliable, and hence its use permits different interpretations of probability.

When probabilities are reliable or calibrated in post-processing, (9) becomes simply

$$BS_c = \int_0^1 p(1 - p)Q(p) \, dp$$  (10)

or

$$BS_c = \bar{\tilde{p}}(1 - \bar{\tilde{p}}) - \sigma_f^2,$$  (11)

where $\bar{\tilde{p}}$ is the average relative frequency of occurrence and is defined as

$$\bar{\tilde{p}} = \int_0^1 f_p Q(p) \, dp,$$  (12)

or for discrete values can be expressed as

$$\bar{\tilde{p}} = \frac{1}{N} \sum_{i=1}^{N} \tilde{p}_i$$  (13)

The term $\sigma_f^2$ is the variance of the frequency of occurrence and is defined as

$$\sigma_f^2 = \int_0^1 (f_p - \bar{\tilde{p}})^2 Q(p) \, dp.$$  (14)

This shows that in the Brier score the quality of a calibrated probability forecast is a simple linear function of the variance of the calibrated probability forecast distribution. It can be seen that the larger the variance, the lower the Brier score, regardless of the probability interpretation or the method used to obtain it. The lowest accuracy level is achieved when the variance vanishes, which only occurs in a calibrated forecast when the climatological probability is used as the forecast at all times. The maximum accuracy level $\sigma_f^2 = \bar{\tilde{p}} (1 - \bar{\tilde{p}})$ is obtained by a perfect deterministic forecast, which is when a forecast system produces accurate predictions composed of 0% and 100% probabilities only. Then, a reliable forecast is to be judged only by the departure of its probability forecast with respect to the climatological value: the farther it is, the better.

In previous paragraphs it was shown that the Brier score is an upper limit for the mean square difference (MSD) between $p_r$ and $p$. The Brier score cannot distinguish between uncertainty due to the forecast system error and uncertainty due to the weather situation. Paradoxically, in the event that the “true” probability was identical to the frequency of occurrence $p_r = \tilde{p}$, no gain in the BS sense would be obtained.

b. Conditional decomposition

Another expression for the Brier score is obtained by following the same procedure as in (3), but exchanging the conditioning of the probabilities [to $Q(p,p_r)$, instead of $Q(p,p)$]. From this point of view the distribution of probabilities is thought of as two distinct distributions: one counting the episodes in which the event occurred, the other counting those episodes in which the event did not occur. This yields

$$BS = \bar{\tilde{p}} (\sigma_r^2 + (1 - \bar{\tilde{p}})\tilde{p}_0^2) + (1 - \bar{\tilde{p}})\sigma_r^2 + \sigma_0^2,$$  (15)

where $\bar{\tilde{p}}$ is the relative frequency of occurrence of the event; $\sigma_r^2$ and $\sigma_0^2$ are the variances of the forecast probabilities when the event occurred and did not occur, respectively; and the terms $\tilde{p}_1$ and $\tilde{p}_0$ are the average values of the probability forecasts when the event occurred and did not occur, respectively.

This expression of the Brier score sheds additional light on its interpretation. To obtain low BS values, the average conditional forecasts should be close to the deterministic values ($\tilde{p}_1 \sim 1$ and $\tilde{p}_0 \sim 0$), and the variances $\sigma_r^2$ and $\sigma_0^2$, should be small. This decomposition will be used extensively in the next section.

4. Forecast system equivalence

In this section, we discuss the possibility that a large number of forecast systems may perform identically in a Brier score sense despite having different individual forecasts. The consequences of this in the interpretation of probability are highlighted.
a. Number of equivalent forecasts

Let us imagine a series of forecasts \( p_i \) and verifications \( o_i \) where \( i \) indicates the time evolution. The Brier score of (8) can be rewritten as

\[
BS = \frac{1}{N} \left( \sum_{i=1}^{M} (p_i - o_i)^2 + \sum_{i=1}^{L} p_i^2 \right),
\]

where the sum on the first term is performed counting the \( M \) cases in which the event occurred, and the second counting the \( L \) cases in which the event did not occur. It can be easily seen that permuting the value of any \( p_k \) by \( p_i \) inside either sum will not alter the BS. [Note that with this rearrangement, permutation of \( p_k \) for \( p_i \) is equivalent to exchanging the entire term \((p_k - 1)^2\) for \((p_i - 1)^2\). This would not be the case if the permutation were done between \((p_k - 1)^2\) and \((p_i - 0)^2\), because they belong to different outcomes of the event.] This can be understood, for example, as switching the “80% probabilities of precipitation” forecast for one day with the 20% given for another day, provided that both were verified as rainy days. Different time series of forecasts with similar Brier scores can then be created by permuting some values as described in Fig. 2.

Assuming that probabilities are given with a certain precision, it is possible to divide them in \( K \) subclasses (e.g., \( 0 \leq p^{(1)} < 0.1, 0.1 \leq p^{(2)} < 0.2 \ldots \), when given with only one significant figure, in which case \( K = 10 \)). Then, it is easy to compute the number \( S \) of rearranged series of forecasts that perform similarly in their Brier score but differ in their time sequence in at least two data points (through one permutation, like that suggested in Fig. 2). By combinatorial calculus we obtain

\[
S = \frac{M! L!}{n_1! n_2! \ldots n_K!}, \quad M, L \geq 1, \quad n_k \geq 0, \quad k = 1, \ldots, K
\]

where \( M \) is the total number of forecast cases where the event occurred, \( L \) is the number of days where the event did not occur, \( n_k \) is the number of forecast cases of subclass \( k \) where the event occurred, and \( n_k^* \) is the number of forecast cases of subclass \( k \) where the event did not occur.

A unique forecast series \((S = 1)\) is obtained under two circumstances. One is when forecasts are deterministic and correct. This can be seen by replacing \( n_1 = M, n_1^* = 0, n_K = 0, n_K^* = L \), and for the remaining classes \( n_k = n_k^* = 0 \) (this case would yield \( BS = 0 \)). The other way is when the same value (the climatological value, if calibrated) is forecast at all times.

Expression (17) also shows that many equivalent forecasts are expected for different forecast systems with low accuracy, and that this number eventually diminishes as accuracy increases. This may be restated as follows: There is only one way of being extremely cautious (do nothing, forecast climatology), and only one way of being exactly right (perfect deterministic forecast), but there are many ways of doing a fair job (to have some moderate skill in a probabilistic prediction). This implies that different individual probabilistic forecasts may be produced for a given situation by different systems performing identically in the Brier score. In addition, the dispersion among forecasts can be thought of as a symptom of the low skill of the systems.

This result indicates that even if the “true” probabilities of a sequence of events are known, another sequence with permuted values (and hence, with the “wrong” individual values) would obtain the same Brier score. This seems to reduce the relevance of the “true” probability. However, the forecast based upon permutations can only be done a posteriori, that is, once the events are verified. The question remains whether the sequence of forecasts equivalent to the “true” one can in fact ever be produced a priori.

It is worth mentioning that forecast system equivalence is defined here in terms of the Brier score, and hence these results are valid only with respect to that score.

b. Distribution of equivalent forecasts

As shown in the previous section, when the forecasts for two events that are verified in the same category are permuted, the BS remains unchanged. Another way to obtain the same result is by producing forecasts randomly generated from the conditional distributions that appear in (15); it is easy to see that forecasting systems that are described by identical parameters \( \sigma_1^2, \sigma_0^2, p_1, \) and \( p_0 \) will perform similarly in a Brier score sense. These forecast systems would be equivalent in this sense and indistinguishable except for the individual values (see the right-hand side of Fig. 2).

Interestingly, the variety of values that can be obtained for a given case by different forecast systems are also described by these parameters. For example, on a
day in which the event under study occurred, the variance of the possible probability values produced by equivalent forecasts system would be $\sigma^2$.

c. Division in classes

In the previous sections it is assumed that probabilistic forecasts can be permuted freely with the aim of creating alternative solutions with the same Brier score. This may not be a suitable assumption for those who believe that weather situations have distinct individual characteristics. For example, is it reasonable to exchange any pair of forecasts that were verified as rainy days, although they are associated with different synoptic situations?

The concept that one probability value can be more adequate than another seems to relate to the idea of propensity, where a “true” probability value exists. It also can be related to the frequentist interpretation of probability, and its subdivision in different classes. The expression of the Brier score (15) can be developed further for the situation in which distinct cases are classified and evaluated separately. For example, precipitation can be classified as convective or stratiform, or according to season, or weather pattern, etc. Then (15) becomes

$$BS = \sum_{k=1}^{C} \{ \tilde{T}_k \sigma^2_{T_k} + (1 - \tilde{T}_k)(1 - \tilde{T}_k) \sigma^2_{\tilde{T}_k} \},$$

(18)

where $C$ is the number of classes, and the index $k$ denotes the class to which the parameter belongs. With this decomposition more parameters are needed to define a system. In addition, the number of equivalent forecasts $S$ is reduced with class separation, as there are fewer possible permutations when performed within each class.

The separation of past events into classes is a standard procedure of the frequentist interpretation of probability. When this concept is stretched to the limit, as in the “weather analogue” case, it suggests that each weather pattern may have properties of its own. Each class would contribute its share in the total Brier score, but parameters may differ among them. One of the objectives of this decomposition is undoubtedly to reduce the variance and bias of each individual class. Re-arranging terms in (18) we obtain

$$BS = \sum_{k=1}^{C} \{ \tilde{T}_k \sigma^2_{T_k} + (1 - \tilde{T}_k)(\sigma^2_{\tilde{T}_k} + \tilde{p}_0^2) \},$$

(19)

If the separation in classes succeeds, as $C$ becomes large the variances $\sigma^2_{T_k}$ and $\sigma^2_{\tilde{T}_k}$ should be reduced significantly (and thus the first summation becomes negligible), and if the forecast is calibrated, the probabilities $\tilde{p}_{1k}$ and $\tilde{p}_{0k}$ should become close to each other and to $\tilde{T}_k$. Then, (19) can be rewritten as

$$BS = \tilde{T}_k (1 - \tilde{T}_k) - \sigma_p^2,$$

(20)

where now $\sigma_p^2$ is understood as the variance of probabilities due to the variety of classes studied. Although the variance of forecasts within each class is small, the combination of classes generates the variability. Expression (20) is identical to the one obtained for calibrated forecasts in (11).

The assumption that the forecast probability variance is reduced by separation in classes is not far from acknowledging the existence of a “true” probability. Whether this existence is a consequence of the atmospheric dynamics or of the handling of this dynamics by the forecast system is immaterial for the moment. What matters is that single-case probabilities (as in the propensity interpretation) may have some conceptual interest. This view would allow us to think of individual cases as belonging to a class that a numerical forecast model is capable of predicting.

5. Relation between dichotomous probability and PDF

Until this point the discussion of probability has been focused on the dichotomous case, that is, when the probability refers to an event that may either occur or not. We can extend this definition to include a PDF by defining an event as a variable exceeding a given threshold. For example, the range of a variable $T$ can be divided into a dichotomous event by choosing the case when either $T \leq T_0$ or when $T > T_0$. Then, the probability of $T \leq T_0$ can be rewritten as

$$p(T \leq T_0) = \int_{-\infty}^{T_0} N(T) \, dT,$$

(21)

where $N(T)$ is a PDF of $T$. This relation extends to PDFs the interpretation of probability $p$, and hence, the PDF is also subject to be interpreted in different ways as discussed in previous sections.

6. Simple numerical experiments

It was shown that the Brier score is not altered either by permuting probability forecasts in the way described in section 4a, or by producing random forecasts derived from the appropriate distributions as in section 4b. In this section the effect of small errors in probability forecasts is studied by simulating an ensemble forecast approach. The errors, hence, will not be introduced in the estimation of the probabilities per se but in the estimation of the forecast probability distribution function, as is the case in ensemble forecasting.

The experiment consists of the comparison of two simple models. One represents the “true” uncertainty
of weather patterns some time in the future (uncertainty associated with predictability limits). The other represents the modeling of that reality, a forecast model that contains modeling errors as well as the uncertainty due to predictability limits. Both models are used to estimate the probability of a given event in the future; one representing the “true” probability \( p_r \), the other the forecast probability \( p \).

These numerical representations do not intend to capture the complexities of atmospheric phenomena and forecasting. In spite of their simplicity, however, they retain the paradigmatic behavior of probabilistic forecasting in a setting familiar to meteorologists.

a. Description of the experiment

A simple function is defined to represent the time evolution of the “true” uncertainty of surface temperature at some location:

\[
T_r = A \sin(wt) + B + T_r^*,
\]

where \( A \) is the amplitude, \( B \) a reference value, \( t \) the time, and \( w \) the frequency of the oscillation. The term \( T_r^* \) is a stochastic variation generated by a normal distribution with zero average and standard deviation \( \sigma_{T^*} \). The standard deviation \( \sigma_{T^*} \) is taken to vary randomly between events, among values uniformly distributed in the interval \( [\sigma_m/2, \sigma_m] \), where \( \sigma_m \) is the maximum standard deviation allowed. The term \( T_r^* \) can be thought of simply as a PDF whose mean value is subject to a cycle and whose standard deviation \( \sigma_{T^*} \) varies from event to event (see Fig. 3). In this case it is supposed that these events are taken some days apart, so they are not correlated; the variation of \( \sigma_{T^*} \) between events intends to describe changes in the uncertainty inherent to different meteorological situations.

The forecast model used to estimate \( T_r \) is similar to the definition of \( T_r \) itself, as presented in (22), but includes sources of error affecting the estimation of the deterministic and stochastic terms. The estimation \( T \) is defined as

\[
T = A \sin(wt) + B + \epsilon + \delta + T^*,
\]

where \( A, B, w, \) and \( t \) have the same meaning as in (22). The term \( \epsilon \) is a random variable normally distributed with zero mean and standard deviation \( \sigma_{\epsilon} \), while \( \delta \) is a constant. The term \( T^* \) is normally distributed with zero average and standard deviation \( \sigma_{T^*} \). The standard deviation \( \sigma_{T^*} \) varies randomly between forecasts among values uniformly distributed in the interval \( (\alpha \sigma_m/2, \alpha \sigma_m) \), where \( \sigma_m \) is the maximum standard deviation allowed for the “true” temperature, and \( \alpha \) is a measure of compression (or expansion) of the standard deviation.

This definition of \( T \) allows for the study of errors of different kinds. The term \( \epsilon \) can be thought of as random error in the determination of the deterministic part (this will be called “mean noise”), while \( \delta \) can be thought of as a systematic bias (this will be called “mean bias”). Errors in the stochastic term deserve a more detailed explanation. As stated above, the standard deviation \( \sigma_{T^*} \) represents the uncertainty inherent in a given meteorological situation, and this varies with time. This variation in time represents changes due to predictability considerations. In the forecast model it is assumed that the prediction of the standard deviation...
\( \sigma_p \) may be systematically biased by a factor \( \alpha \) (this will be called “sigma bias”). In addition, since the standard deviations of the stochastic terms in (22) and (23), \( \sigma'_p \) and \( \sigma_{p*} \), are random variables, it is important to establish the correlation \( \rho \) between them. If they are perfectly correlated (\( \rho = 1 \)), it means that the “true” uncertainty and that of the forecast model are in perfect synchronicity, predicting a small variance when there is a small one, or predicting a large one when there is a large one. On the other hand the standard deviations may not be correlated in time (\( \rho = 0 \)), that is, having both parameters generated by the same distribution, but with different time series (this will be called “sigma noise”). Other intermediate values of \( \rho \) could have been chosen, but these are preferred for representing extreme conditions.

As in the previous section, we want to estimate the probability that \( T_r \approx T_0 \) where \( T_0 \) is a given threshold for both the “true” and the forecast case. The time series of random parameters (\( \varepsilon, \sigma'_p \), and \( \sigma_{p*} \)) from (22) and (23) were built using computer-generated time series with the appropriate distributions. In addition, for each forecast case \( i \) a realization of the event \( T_r \leq T_0 \) or \( T_r > T_0 \) was associated by using a random generator with the “true” distribution. If the equality \( T_r \approx T_0 \) was true \( f_i = 1 \), if not \( f_i = 0 \). This was done in order to be able to construct relative frequencies of the occurrence of the event. In total 100,000 different forecasts were produced.

The variable \( T_r \) has been assigned values that are representative of continental midlatitude surface temperature. Hence, the amplitude of the cycle was chosen to be \( A = 4^\circ C, B = 0^\circ C \), and a maximum standard deviation for a given event \( \sigma_m = 3^\circ C \). The threshold was established at \( T_0 = 0^\circ \), corresponding to frost conditions. The sources of error were varied among realistic values, but only a set of them is presented here: the unbiased error \( \varepsilon \) with standard deviation \( \sigma_\varepsilon = 1^\circ C \), the bias \( \delta = -1^\circ C \), and the compression factor \( \alpha = 0.8 \).

b. Results

Figure 4 depicts six uncalibrated scattergrams where the “true” probability \( p \) is plotted against the forecast probability \( p \). The thick solid line represents the frequency of occurrence \( f_p \) of the event. The area highlighted by the parallel horizontal lines indicates the region that contributes positively to forecast skill (see Wilks 1995). In (a) only “mean noise” is introduced while the other sources of error are set to zero; in (b), only “mean bias”; in (c), only “sigma noise”; in (d), only “sigma bias,” in (e), the combined effect of noise in mean and in sigma, and in (f) all the sources of error together (see Table 2 for a brief description of each experiment).

For a perfect probabilistic forecast all the dots should lie on the dashed diagonal line of the diagrams. All individual probabilistic forecasts contain some error with respect to the “true” probabilities, as is shown by the cloud of points present in most diagrams (Fig. 4d is the exception, suffering loss of reliability but not dispersion). Panels on the left-hand side are only affected by noise and are quite reliable, with the exception of the extremes (the frequency of occurrence, depicted by the thick solid line, is very close to the diagonal).

Figure 4e shows that even in a reliable forecast, the difference between the “true” probability \( p \) and the forecast probability \( p \) can be quite large (e.g., for a forecast probability \( p = 0.4 \), the “true” probabilities may lay in a very large interval. This happens in spite of the fact that the predictive model is of good quality, as evaluated by the Brier skill score \( SS \) [defined as \( SS = 1 - BS(1 - \bar{p})^{-1} \) and presented in Table 2; see Wilks (1995) for additional information]. The overall quality arises from the fact that large errors do not occur in most cases, as can be seen in the clustering of dots near the extremes of the diagonal of Fig. 4e. This seems to indicate that reliability diagrams are of little use in detecting the presence of these kinds of errors. Despite the fact that individual probabilities are affected, it may not leave a clear trace in the frequency of occurrence.

This example serves to illustrate the fact that a reliable forecast does not come as a result of the coincidence between the “true” and forecast probability for each individual case, but as a statistical property. The average value \( f_{p*} \) is obtained by a succession of compensating underestimations and overestimations of the “true” probabilities of events. It is worth noting that near the extremes a symmetrical compensation is impossible because of the finite extension of the range, and this introduces a curve in \( f_{p*} \).

Figure 4c suggests that, even when the “true” variance of a given day is uncorrelated with that of the forecast model, the quality of the prediction can still be quite high (see Table 2). Figure 4b shows that a negative bias in the mean introduces concave curvature (when the bias is positive, the curvature is convex), while Fig. 4d shows that the compression of the variance introduces an S-shaped curvature (when the variance is expanded instead of compressed, the S shape is reversed). These deformations can be explained as follows: A bias in the mean between two identical PDFs forces one of them to have higher probability values than the other, while a compression of the variance introduces higher values for the PDF having more spread to the left of the mean, and lower values to the right of the mean. The combination of errors displayed in Fig. 4f look realistic if compared to results of actual weather forecast ensembles (e.g., Hamill and Colucci 1997; Pernson 2001). This resemblance, however, may be unrelated with the real causes behind the appearances of the actual curves.

Figure 5 depicts the distribution for probability forecasts of frost associated with Fig. 4, as defined in section 3. The solid line depicts the distribution \( Q(p) \) that can be interpreted as the distribution of “true” probabilities; the dotted line shows \( Q(p) \), the distribution of
Fig. 4. (a)–(f) Scattergrams between the “true” probability $p_r$ and the forecast probability $p$ for different sources of error in the estimation model. Both probabilities represent the chance of $T < 0°C$. In (a) the error is introduced only in the mean with an unbiased noise ($\sigma_y = 1°C$); in (b), a bias in the mean ($\delta = -1°C$); in (c), noise in the variance by decorrelating it with respect to the “true” one; in (d), a compression in the variance ($\alpha = 0.8$); in (e), a combination of noise in both the mean and the variance; and in (f), all errors acting simultaneously. The solid lines depict the relative frequency of occurrence $f_p$; the vertical and the horizontal dashed lines depict the average forecast and “true” probability, respectively. The shaded areas indicate the region that contributes positively to forecast skill.
forecast probabilities; and the dashed line depicts the distribution of forecast probabilities after calibration $Q(f_p)$. As in Figs. 4a, 4c, and 4e, all cases affected by noise only show little difference between the three distributions and hence little difference in Brier score and Brier skill score (see Table 2). On the contrary, when biases are present, the three distributions differ, the calibrated one becoming highly asymmetric but substantially improving the score (see Table 2).

These results complement those of section 4, which showed that several different probability forecasts are possible for low-skill forecasts systems with identical Brier score. Here it was shown that even when the spread of the model for individual forecasts is unrelated with that of the “true” uncertainty, the solution may not be much degraded. In addition, it has been shown that skillful probability forecasting can be obtained by compensating errors through balancing overestimations and underestimations of the “true” probabilities.

7. Discussion and conclusions

The main objective of this work is to discuss some issues regarding the interpretation of probabilistic forecasting at a moment when its use is becoming increasingly popular due to progress made in ensemble forecasting in weather and climate prediction. Part of this discussion concentrates on elementary concepts of probabilities that may still need further debate in the atmospheric sciences. The reality is that probability theory is an unusual branch of mathematics whose foundations have been rewritten several times in the last hundred years, and it has resisted and continues to resist a unified and simple interpretation. Although probability calculus possesses solid roots, probabilities themselves need to be interpreted carefully.

The calculus of probabilities consists of the handling of probabilities of certain events (prior probabilities) with the aim of obtaining the probabilities of other events (posterior probabilities) that relate to the former in a more or less complex manner. The interpretation of the meaning of probabilities is beyond the interest of probability calculus and this is why most textbooks devote little or no space to this discussion.

Some of the most popular interpretations of probabilities were briefly discussed in section 2, and the following sections rely on these descriptions. The fact that the meaning of probabilities is still an open debate should make us aware of the consequences that this may carry in our research, our predictions, and our communications among scientist and general users. We hope that this work contributes to this awareness.

The Brier score has been chosen in this work as the measure of quality of probabilistic forecasting and was presented in two different formats, one of them explicitly dependent on reliability. Users may eventually interpret probability values in a “frequentist” framework regardless of the view of the producer of those probabilities. The expectation of users that the relative frequency of events be similar to the forecast probability value may have implications in the way probabilistic forecasts are to be generated. Proper calibration of forecast probabilities requires a large database of observations and forecast probabilities; the latter can only be obtained by performing hindcasts on large historical databases (e.g., Derome et al. 2001; Hamill et al. 2004).

In section 4, the equivalence of forecast systems with respect to the Brier score was defined. It was shown that several forecasts may have the same Brier score and that the number increases for lower skill forecasts. This can be summarized by saying that there is only one way of doing a perfect forecast (deterministic) and many ways of doing a poorer job.

This equivalence between forecast systems also implies the existence of a sequence of forecasts as accurate as the one produced using “true” probabilities. This may seem to be a blow to the propensity interpretation, which proposes the existence of a “true” probability. However, it is important to remember that the conditioning of the Brier score implies that the verification is already known. That is, these alternative forecasts are obtained posteriori and hence only exist hypothetically. Someone convinced of the existence of “true” probabilities could argue that either the Brier score is unable to capture this distinction, or that this hypothetical forecast may never be achieved in practice.

As shown in section 4b, the variety of probability values obtainable with equivalent forecast systems are
FIG. 5. The corresponding distributions of probability from Fig. 4. Solid lines depict the distribution of the "true" probabilities, dotted lines depict the distribution of the forecast probabilities, and the dashed line depicts the distribution of the forecast probabilities after calibration.
described by the parameters present in the conditional expression of the Brier score (15). Hence, each probability forecast can be thought of as being a random variable with a given distribution, which implies that a probability forecast is just a “realization” of several possibilities. If we do not accept the idea that there is a “true” probability, we are forced to accept probabilities not only as a measure of uncertainty but also as an “inherently uncertain” measure of uncertainty. This fact has been rather traumatic for probability theory and has stimulated the development of several new fields that have tried to capture this idea [e.g., see in Smithson (1989) possibility theory, fuzzy sets theory, uncertainty factors theory, Dempster-Schafer theory, and qualitative probability theory]. Concepts such as “the probability of a probability,” “imprecise probabilities,” “second-order probabilities,” and “metaprobability,” developed in the last decades (Smithson 1989), reflect that it is impossible to provide a fair sense of uncertainty with one single value.

For example, even if in a given situation a forecast given by climatology and an ensemble of the most advanced numerical models coincide in a numerical value for the probability of rain (e.g., 30%); the second one is considered to be less uncertain. By this it is meant that a vast amount of information (scientific knowledge and data) was used for the production of the second, and that further usage of the available information may not substantially change the probability value. This complex relation between uncertainty and probability is ignored most of the time, opening the door to criticism of the kind that subjective interpretation of probability is the center: its lack of precision and of methodology to obtain a probability value (section 2c).

The existence of equivalent forecasts, as discussed previously, results from the fact that events that after verification fall in the same category (e.g., no rain) can be exchanged without modifying the Brier score. This permutation assumes that the distribution of probabilities for a given day is the same as that of any other day that verifies with the same outcome. This assumption denies the possibility that different individual weather situations may have different characteristics regarding probability distributions.

The conditional version of the Brier score was extended by restricting the exchange of probability forecasts to similar meteorological cases. This separation in “weather analogues” allows for the interpretation that probability distributions are dependent on weather dynamics, hence leading to the concept of objective, single-case probabilities. This point of view suits well both the propensity and the frequentist interpretation, the latter thinking of the “weather analogue” as a class with its own frequency of occurrence (which may be difficult to determine quantitatively because of the short sample of historic atmospheric data).

Further consequences of the interpretation of probability were found when this problematic was extended to probability distributions in section 5. This relation establishes that the interpretations given to probabilities of dichotomous cases apply to PDFs too. This suggests that several expressions common in the literature such as “the true PDF,” “the PDF,” “the real PDF,” “the theoretical PDF,” “the ideal PDF,” “the unknown PDF,” etc. may lead to confusion. In some cases, it may be understood that the PDF is defined under certain conditions, while in others the impression given is that the researcher believes that a “true PDF” exists.

The numerical experiment presented in section 6 illustrates how a reliable, skillful probability forecast may be obtained by compensating errors balancing large overestimations and underestimations of the “true” probabilities. These errors do not have much effect on the Brier score because they do not occur frequently. It is interesting to note that even when the “predicted” variance of the uncertainty is decorrelated with the “real” one, a rather good skill is obtained. The reliability in this forecast is achieved because a PDF with a variance narrower than the “true” PDF is eventually compensated with a wider one. Similar compensating problems were found by Hamill (2001) in the interpretation of rank histograms. A related problem was also found by Atger (2003): the reliability of a probabilistic forecast in a large area may be due to the contribution of compensating errors in different subregions. The fundamental question may perhaps be reduced to the problem of existence–meaning–usefulness of a single-case probability distribution. Unless a “true” probability distribution is believed to exist, the meaning of the ensemble distribution for a single case remains unclear. The consequences that low-skill forecast systems have a large number of equivalent solutions implies that probability distributions obtained by ensemble forecast may differ largely from system to system and prompt different actions in users.

The problem of the interpretation of a PDF in weather forecasting seems to be how to deal with a probabilistic prediction that may be valid in the long term, but uncertain or ill defined on a single case. Unless we can somehow convince ourselves that the forecast PDF is either always very close to the “true” PDF (for which we do not have any proof, neither of the closeness, nor of the existence of the “true” PDF) or very close to the deterministic solution (which is another way of being close to the “truth”), there seems to be no straightforward interpretation of the probabilistic information generated by an ensemble forecast system for single cases. This has the consequence that the predictions of the most dramatic cases—those whose occurrences in a lifetime are uncommon—are ignored.

All these considerations suggest that more effort needs to be devoted to the understanding of probabilities and the techniques for estimating them; otherwise, we may run the risk of perfecting methodologies without a clear aim.
Acknowledgments. We are very grateful to the scientific and technical staff of the Canadian Regional Climate Model Group (CRCM) at UQAM for their assistance, and to Mr. Claude Desrochers for maintaining an efficient computing environment for the CRCM group. In addition, we are indebted to the Consortium Ouranos for the facilities and work environment it provided. The first author would particularly like to thank the fruitful discussions with Jillian Tomm and with Drs. Martin Charron, Pascale Martineau, and Arturo Quintanar. In the same way, comments and suggestions by Prof. David B. Stephenson during his stay in our institution were greatly appreciated. The feedback obtained as a consequence of the peer-review process has proven to be invaluable; the comments of the editor of this journal, Dr. David Schultz; four anonymous reviewers; and Dr. Tom Hamill were instrumental in stimulating the authors and helping to focus the scope of this work. This work has been financially supported by Canadian NSERC, through a strategic grant to the Climate Variability (CliVar) Network.

REFERENCES


