Improved Seasonal Probability Forecasts

VIATCHESLAV V. KHRIN AND FRANCIS W. ZWIERS

Canadian Centre for Climate Modelling and Analysis, Meteorological Service of Canada, Victoria, British Columbia, Canada

(Manuscript received 8 May 2002, in final form 31 October 2002)

ABSTRACT

A simple statistical model of seasonal variability is used to explore the properties of probability forecasts and their accuracy measures. Two methods of estimating probabilistic information from an ensemble of deterministic forecasts are discussed. The estimators considered are the straightforward nonparametric estimator defined as the relative number of the ensemble members in an event category, and a parametric Gaussian estimator derived from a fitted Gaussian distribution. The parametric Gaussian estimator is superior to the standard nonparametric estimator on seasonal timescales. A statistical skill improvement technique is proposed and applied to a collection of 24-member ensemble seasonal hindcasts of northern winter 700-hPa temperature ($T_{700}$) and 500-hPa height ($Z_{500}$). The improvement technique is moderately successful for $T_{700}$ but fails to improve Brier skill scores of the already relatively reliable raw $Z_{500}$ probability forecasts.

1. Introduction

The chaotic nature of the climate system implies that predictions of future states are inherently uncertain. This is evident from short-term forecasts that are sensitive to small differences in initial conditions and have predictability limits of the order of two weeks. Nonetheless, experience had demonstrated that longer-term forecast of, for example, seasonal statistics (such as the seasonal mean) are possible with moderate levels of skill. The main source of this extended predictability on seasonal and longer timescales is thought to be lower boundary forcing associated with the distribution of persistent sea surface temperature and sea ice anomalies.

The utility of a single forecast of an atmospheric quantity is limited because the population of forecasts that are consistent with uncertainty in the initial conditions has a large spread at long lead times. Thus, forecast uncertainty must be quantified and communicated in order to extract useful information from long-lead forecasts. A common approach used in seasonal forecasting is to estimate the likelihoods of a small number of mutually exclusive events. Typically three equiprobable categories "below normal," "near normal," and "above normal" are considered. Forecast uncertainty is characterized by the discrete probability distribution of the three outcomes. This forecast format is motivated by the simplicity of the forecast presentation and is used by many operational centers that make seasonal forecasts.

The present study examines some aspects of seasonal probability forecasts in a very simple framework that assumes that seasonal variations are made up of a potentially predictable signal and unpredictable noise. This approach is not new and has been used in numerous studies in the past (e.g., Leith 1973; Madden 1976; Zwiers 1996). More recently it was used by Kharin and Zwiers (2001) and Kharin et al. (2001) to investigate some properties of deterministic seasonal forecasts.

Two aspects of probability forecast estimation are considered here. First, a parametric approach to deriving probabilistic information from an ensemble of deterministic forecasts is compared to the straightforward nonparametric approach based on the relative number of the ensemble members in a category of interest. The parametric probability estimator exploits explicitly the properties of the assumed underlying distribution of seasonal variations. Second, a statistical skill improvement technique for biased probability forecasts is proposed. The statistical procedure adjusts the signal-to-noise ratio in model forecasts to minimize the mean-square error of probability forecasts. The probability forecasts methods are tested on 24-member ensembles of northern winter seasonal hindcasts of 700-hPa temperature and 500-hPa height produced with the second-generation general circulation model of the Canadian Centre for Climate Modelling and Analysis (CCCma).

The paper is organized as follows. A simple statistical model of seasonal variability is introduced in section 2. This model is used to examine some probabilistic aspects of seasonal variations. Some of the properties of
the Brier score, the mean-square error of probability forecasts, are discussed in section 3 in the context of the introduced simple model. Readers who are interested in more practical aspects may skip this section and go directly to section 4 where the probability forecast estimation problem is addressed and a statistical skill improvement technique is introduced. The results for CCCma seasonal hindcasts are reported in section 5. The paper is completed with a summary in section 6.

2. Probability in a simple seasonal variability model

In this section we use a very simple statistical model of atmospheric variations on seasonal timescales to examine some probabilistic aspects of observed seasonal variability.

We assume that seasonal variations of a scalar quantity $X$ can be represented as a sum of a potentially predictable signal $\beta$ and nonpredictable variability that will be treated as stochastic noise $\epsilon$, that is,

$$X = \beta + \epsilon.$$  \hfill (1)

The term $\beta$ comprises all potentially predictable signals arising from external and internal sources. On seasonal and longer timescales, the primary predictability source is thought to be the atmospheric response to slowly varying lower boundary conditions such as sea surface temperature. The effect of the initial conditions is felt primarily in the first few weeks of model integrations. The noise term $\epsilon$ represents the unpredictable effects of day-to-day weather. It is assumed that all terms in (1) are stochastic processes with the zero mean, that is, $\mathbb{E}(\beta) = \mathbb{E}(\epsilon) = 0$, where the symbol $\mathbb{E}$ denotes expectation, or the mean of a random variable.

The ratio of the variance of the potentially predictable signal $\sigma_\beta^2$ to the total observed variance $\sigma_X^2$ is often referred to as the potential predictability (Madden 1976; Zwiers 1996; Rowell 1998). This ratio is equal to the squared correlation between the potentially predictable signal $\beta$ and the predictand $X$. Therefore, we will use the notation $\rho^2_{pot} = \sigma_\beta^2/\sigma_X^2$ to denote the deterministic potential predictability of the system (1).

a. Probability of an “event”

We define an event $\Omega$ as an occurrence of $X$ in an interval $(x_1, x_2)$. Let $F_X(x \mid \beta)$ be the (cumulative) distribution of the predictand $X$ conditional on a given value of $\beta$. Then the probability that $X$ lies in an interval $(x_1, x_2)$ conditional on $\beta$ is given by

$$P_X(\Omega \mid \beta) = \text{Prob}[X \in (x_1, x_2) \mid \beta] = F_X(x_2 \mid \beta) - F_X(x_1 \mid \beta).$$  \hfill (2)

If the noise term $\epsilon$ is Gaussian, this conditional probability may be written as

$$P_X(\Omega \mid \beta; \sigma_\epsilon) = \mathcal{N}(x_2 - \beta/\sigma_\epsilon) - \mathcal{N}(x_1 - \beta/\sigma_\epsilon),$$

where $\mathcal{N}$ is the distribution function of the standard normal distribution. The probability depends both on the value of $\beta$ and the standard deviation of $\epsilon$.

In seasonal forecasting it is common to define three equiprobable mutually exclusive and collectively exhaustive categories: below normal (B), near normal (N), and above normal (A). The boundaries $x_a$ and $x_b$ between the categories are usually defined in terms of the terciles of the normal distribution

$$x_a = -x_b = x_{1/3} \sigma_X,$$

where $x_{1/3} = \Phi^{-1}(1/3) \approx 0.43$. The corresponding conditional probabilities of the below-normal, near-normal, and above-normal events are given by

$$P_X(B \mid \beta; \sigma_\epsilon) = \mathcal{N}(\beta \mid 0, \sigma_\beta/\sigma_\epsilon),$$

$$P_X(A \mid \beta; \sigma_\epsilon) = \mathcal{N}(\beta \mid 0, \sigma_\beta/\sigma_\epsilon),$$

$$P_X(N \mid \beta; \sigma_\epsilon) = 1 - P_X(B \mid \beta; \sigma_\epsilon) - P_X(A \mid \beta; \sigma_\epsilon).$$  \hfill (4)

These probabilities depend on the standardized signal perturbations $\beta/\sigma_\beta$ and on the potential predictability of the system. The probability of the near-normal category is symmetric with respect to the sign of the potentially predictable signal perturbation, that is, $P_X(N \mid \beta; \sigma_\epsilon) = P_X(N \mid -\beta; \sigma_\epsilon)$.

Figure 1 illustrates the probabilities of observing $X$ in one of the three equiprobable categories conditional upon $\beta = +\sigma_\beta$ (right panel) and $\beta = -0.5\sigma_\beta$ (left panel). The potential predictability in these examples is set to $\rho^2_{pot} = 0.3$, which is within the range of typical values for the proportion of the boundary forced seasonal variability of 500-hPa geopotential height over North America in winter (Madden 1976; Zwiers 1996). The probabilities of the below-, near-, and above-normal categories are equal to the corresponding areas under the probability density curve of the noise as indicated by different shading intensities. Note that there is only weak dependence of $P_X(N \mid \beta)$ upon $\beta$ in these examples, but rather strong dependence of $P_X(A \mid \beta)$ and $P_X(B \mid \beta)$ upon $\beta$.

The dependence of conditional probabilities $P_X(B \mid \beta)$, $P_X(N \mid \beta)$, and $P_X(A \mid \beta)$ in (4) on $\beta$ is more fully illustrated in Fig. 2 for potential predictabilities $\rho^2_{pot} = 0.15$, 0.3, and 0.6. The latter value is typical for seasonal
atmospheric variations in the Tropics, while the former two values are typical of northern midlatitudes. A similar diagram is presented in Leith (1973). The \( \beta \) values on the \( x \) axis are expressed in terms of the signal standard deviation \( \sigma \). The probability of the above-/below-normal event depends monotonically on the signal. The greater the potential predictability, the steeper the probability curves, which become, asymptotically, step functions in a fully deterministic system with probability values of either 0 or 1. The \( P_X(\beta|b) \) curves are flat at the origin, and thus the conditional probability of the near-normal category is relatively insensitive to small signal perturbations. This is particularly evident in low or moderate potential predictability situations for which the near-normal probabilities are close to the climatological probability over a wide range of signal values. Thus knowing the predictable part of \( X \) generally does not help to greatly sharpen forecasts of the likelihood of the near-normal category beyond the climatology forecast. This is the main reason for the widely experienced seasonal forecasting phenomenon of near-normal forecasts that are not much more skillful than the climatology (e.g., Van den Dool and Toth 1991).

The insensitivity of the probability of the near-normal category to signal perturbations has led the U.S. National Weather Service (Epstein 1988) to always assign the climatological probability to the near-normal category. This restriction has allowed them to reduce the presentation of the three-category long-range probability outlooks to a single map that displays the forecast probabilities for the more likely of the two extreme categories.

**b. Skill score of probability forecasts**

Let \( P \) be a probability forecast, and let \( E \) be a random event variable, or predictand, that takes the value 1 when a forecast event occurs and 0 otherwise. A standard measure of the accuracy of probability forecasts is the mean-square error, which is normally referred to as the Brier score (\( B \)) in probabilistic forecasting. The Brier score is defined as

\[
B = \mathcal{E}(P - E)^2. \tag{5}
\]

The range of \( B \) is the closed interval \([0, 1]\). The Brier score is a negatively oriented accuracy measure meaning that smaller \( B \) values indicate better forecasts, and vice versa. The corresponding Brier skill score (BSS) is defined as

\[
\text{BSS} = 1 - \frac{\mathcal{E}(P - E)^2}{\sigma^2}, \tag{6}
\]

where \( \sigma^2 = P_c(1 - P_c) \) is the variance of \( E \), and \( P_c = \mathcal{E}(E) \) is the climatological frequency of the forecast event. Note that the variance \( \sigma^2 \) is equal to the Brier score of the probability forecast that always predicts the climatological probability \( P_c \). The range of the BSS is the closed interval \([1 - 1/(P_c(1 - P_c)), 1]\). For example, in the case of the three-event equiprobable partition of the observed value space, the climatological frequency of every category is \( P_c = 1/3 \) so that \( B_c = 2/9 \), and the range of the Brier skill score is the interval \([-7/2, 1]\). The BSS equals 1 for a perfect probability forecast \( P = E \), 0 for the climatological probability forecast \( P = \mathcal{E}(E) \), and is negative for forecasts that have larger Brier scores than that of the climatological probability forecast. Further discussion of this and other skill measures of a probability forecast can be found, for example, in Wilks (1995) and Pan and Van den Dool (1998).

### 3. A probabilistic interpretation of potential predictability

In this section we describe a measure of probabilistic potential predictability in the context of the simple statistical model described in section 2. This measure is
derived from the Brier score decomposition of Murphy (1973) and Murphy and Winkler (1987).

Formally, the Brier score in (5) can be rewritten in terms of the joint probability density function \( f(E, P) \) of the predictand \( E \) and its forecast \( P \) as follows:

\[
B = \int_P \int_E f(E, P) (P - E)^2 \, dE \, dP.
\]

Based on the calibration-refinement factorization of the joint distribution \( f(E, P) = f(E|P)f(P) \) (Murphy and Winkler 1987), where \( f(E|P) \) is the conditional probability distribution of observations \( E \) for each possible forecast \( P \) and \( f(P) \) is the marginal probability density function of the forecasts, the Brier score can be decomposed into three nonnegative terms:

\[
B = \int_P \left[ \frac{f(P|E) - E[E|P]]^2 \, dP}{\text{reliability}} \right] + \int_P \left[ f(P)[E(E|P) - P_C]^2 \, dP + P_C - P_C^* \right. \left. \text{resolution, uncertainty} \right] = B_{\text{rel}} - B_{\text{res}} + B_{\text{unc}} \tag{7}
\]

(see appendix A for derivation details). The three terms are known as the reliability, the resolution, and the uncertainty. The reliability term \( B_{\text{rel}} \) summarizes the calibration, or the conditional bias, of the forecasts and is equal to the weighted average of squared differences between the forecast and conditional observed probabilities. The resolution term \( B_{\text{res}} \) characterizes the amplitude of the deviations of the conditional probability from the climatological frequency \( P_C \). The uncertainty term \( B_{\text{unc}} \) is equal to the variance of \( E \) and as such characterizes the properties of the observed system only. The corresponding BSS can be written in terms of these three components as

\[
\text{BSS} = \frac{B_{\text{rel}}}{B_{\text{unc}}} - \frac{B_{\text{res}}}{B_{\text{unc}}}, \tag{8}
\]

Below, the Brier score in (5), corresponding BSS in (8), and the terms of the joint distribution are broken out into three terms as follows:

\[
B = \int_{-\infty}^{\infty} f_\beta(\beta) [P(\beta) - P_x(\beta)]^2 \, d\beta \Bigg|_{B_{\text{rel}}} - \int_{-\infty}^{\infty} f_\beta(\beta) [P_x(\beta) - P_C] \, d\beta \Bigg|_{B_{\text{res}}} + P_C - P_C^*. \tag{9}
\]

Note that under the assumptions made above the only term in (9) that depends explicitly on the probability forecast is the nonnegative reliability term \( B_{\text{rel}} \). The other two terms are functions only of the properties of the observed system.

The best probability forecast that minimizes the Brier score is given by \( P(\beta) = P_x(\beta) = P_{\text{pot}} \). This forecast can be seen as the probabilistic analog of the best deterministic forecast in system (1), which is given by \( \beta \). The corresponding \( \text{BSS}_{\text{pot}} \) is given by

\[
\text{BSS}_{\text{pot}} = \frac{B_{\text{rel}}}{B_{\text{unc}}} = \frac{\sigma_r^2}{\sigma_b^2}. \tag{10}
\]

Analogous to the deterministic potential predictability \( \rho_{\text{pot}}^2 \), the \( \text{BSS}_{\text{pot}} \) (10) may be referred to as the potential predictability in the probabilistic formulation.

Figure 3 shows \( \text{BSS}_{\text{pot}} \) for the above-below-normal and near-normal categories as a function of the deterministic potential predictability \( \rho_{\text{pot}}^2 \) for the system (1) in the Gaussian setting, that is, when both \( \beta \) and \( \epsilon \) are normally distributed. The \( \text{BSS}_{\text{pot}} \) in this case is given by

\[
\text{BSS}_{\text{pot}} = \frac{9}{2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{1}{2} \left( \frac{\beta}{\sigma_b} \right)^2 \right] \times P_x(\beta) [1 - P_x(\beta)] \left( \frac{d\beta}{\sigma_b} \right), \tag{11}
\]

where \( P_x(\beta) \) is defined in (4). The expression on the right-hand side of (11) was evaluated with the aid of an algebraic manipulator (Maple V, see Char et al. 1991). Initially, \( \text{BSS}_{\text{pot}} \) increases more slowly than \( \rho_{\text{pot}}^2 \). The dependence on \( \rho_{\text{pot}}^2 \) is particularly weak for the near-normal category for small to moderate levels of potential predictability. This occurs because the probability of the near-normal category depends only weakly on the amplitude of the potentially predictable signal. The diagram demonstrates that there is a one-to-one correspondence between the potential predictability in the deterministic sense and its counterpart in the probabilistic formulation. It also indicates that this relationship depends on the event definition. A given level of potential predictability in the deterministic sense corresponds to different levels of “probabilistic” potential.
4. Probability forecasts and their improvement

In this section we describe two methods for producing a probability forecast from an ensemble of deterministic forecasts and evaluate the skill of these forecasts in the idealized setting described in section 2. We also describe a method for the statistical improvement of probability forecasts.

a. Probability estimators

We begin by assuming that \( \{ F_i, i = 1, \ldots, N \} \) is an ensemble of \( N \) forecasts produced with a perfect forecast system that behaves as (1).

In these circumstances a simple and straightforward estimator of the event probability \( P_X \) in (2) is obtained as

\[
\hat{P}_X(\Omega) = n/N,
\]

where \( n \) is the number of ensemble members in the event category \( \Omega = B, N, \) or \( A \). The subscript \( B \) is used to indicate the fact that the number \( n \) is distributed according to the binomial distribution. For a given event probability \( P_X \), the mean and the variance of the binomial random variable \( n \) are given by

\[
E(n) = NP_X,
\]
\[
\text{var}(n) = NP_X(1 - P_X),
\]

from which it immediately follows that

\[
E[\hat{P}_X(\Omega)] = P_X,
\]
\[
\text{var}[\hat{P}_X(\Omega)] = P_X(1 - P_X)/N.
\]

Thus, \( \hat{P}_X(\Omega) \) is an unbiased and consistent estimator of the true event probability \( P_X \).

The nonparametric estimator (12) is very general as no specific assumptions are made about the underlying distribution of \( X \). Another estimate of \( P_X \) may be obtained by fitting a suitable distribution to the available data sample and then estimating the event probability by computing it from the fitted distribution. In particular, when \( \varepsilon \) in (1) behaves as a Gaussian random variable, the probabilities of the below-normal, near-normal, and above-normal categories are given by (4). Thus, given estimates of \( \beta \) and \( \sigma_\varepsilon \), the probabilities can be estimated as
where the subscript $g$ is used to indicate the Gaussian nature of the underlying statistical model. Given $N$ independent deterministic forecasts $\{F_i, i = 1, \ldots, N\}$ produced by a perfect forecast system, unbiased estimators of $\beta$ and $\sigma_x^2$ are given by

\[
\hat{\beta} = \frac{1}{N} \sum_{i=1}^{N} F_i, \\
\hat{\sigma}_x^2 = \frac{1}{N-1} \sum_{i=1}^{N} (F_i - \hat{\beta})^2.
\]

When more than one sample of ensemble forecasts is available, for example, for different realizations of the potentially predictable signal $\{\beta_t, t = 1, \ldots, T\}$, and assuming that the statistical properties of the noise term are independent of $\beta$, all samples can be utilized to estimate $\sigma_x^2$ by applying a standard analysis of variance (e.g., von Storch and Zwiers 1999). The unbiased estimator of $\sigma_x^2$ is then given by

\[
\hat{\sigma}_x^2 = \frac{1}{(N-1)T} \sum_{i=1}^{T} \sum_{i=1}^{N} (F_{i,t} - \hat{\beta}_i)^2,
\]

where $\{F_{i,t}, i = 1, \ldots, N\}$ is the $i$th ensemble forecast and $\hat{\beta}_i = 1/N \sum_{i=1}^{N} F_{i,t}$. Obviously, estimator (15) is obtained from (16) for $T = 1$.

Variance estimator (16) is useful when historical ensemble forecasts are available, and when the dependence of $\sigma_x^2$ on $\beta$ may be neglected. Since the overall sample size is larger in this case, one expects a more accurate estimate of $\sigma_x^2$. On the other hand, the assumptions required for estimator (15) are less stringent so it can be used even when $\sigma_x^2$ is thought to be dependent upon $\beta$.

In appendix B it is shown that, as expected, the Gaussian estimator is preferable to the nonparametric estimator when the data are consistent with the system (1) in the Gaussian setting. The Gaussian estimator has smaller rms errors in these circumstances. It is slightly biased for small ensemble sizes, but the bias is small relative to the corresponding rms errors.

b. Skill scores of biased probability forecasts

Real world forecasts are always subject to errors and biases that result from model errors and errors in specifying the initial and boundary conditions. In its simplest form, the effect of biases and errors can be represented by a forecast of the form

\[
F = a \beta + b e',
\]

where $\sigma_b^2 = \sigma_e^2$. That is, the potentially predictable signal that is contained in the forecast differs from the observed counterpart by the factor $a$, and the amplitude of the unpredictable variations in the forecast differs from that of the observed counterpart by the factor $b$.

In the following we consider how such model biases affect three-category probability forecasts and their skill scores.

The first step in constructing probability forecasts is often to ensure that the climatological frequency of the forecast categories is equal to that of their observed counterparts. This can be achieved by adjusting the boundaries $x_-$ and $x_+$ between the categories for the model forecasts. In the Gaussian setting this is equivalent to standardizing, or equalizing, the observed and model total variances. This step is sometimes referred to as the variance inflation (von Storch 1999). Without this step, the climatological frequency of the near-normal category would be overestimated at the expense of the below- and above-normal categories when the model variance is underestimated, and vice versa when it is overestimated. Thus in the following we require that

\[
\sigma_x^2 = \sigma_b^2 + \sigma_e^2 = a^2 \sigma_b^2 + b^2 \sigma_e^2 = \sigma_x^2.
\]

With this assumption the noise bias factor $b$ is related to the signal bias factor $a$ as

\[
b^2 = \frac{1 - a^2 \rho_{oot}^2}{1 - \rho_{oot}^2}.
\]

The parameter $a$, which is an indicator of the strength of the signal-to-noise ratio in forecasts relative to that in the observations, can take values in the range $[-1/\rho_{oot}, 1/\rho_{oot}]$. The lower and upper bounds correspond to forecasts with no unpredictable component, while the value $a = 0$ corresponds to no skill forecasts. Negative values of $a$ indicate that the sign of the forecast signal is opposite to that observed. The perfect forecast system is obtained for $a = 1$.

The conditional probabilities of the three equiprobable categories obtained from the biased forecast system (17) in the Gaussian setting are given by

\[
P_F(B|\beta) = \int a \beta - x_e \bar{b} \sigma_e \frac{\phi}{\phi - a \beta - x_e} = \frac{x_{1/3} - a \rho_{oot} \phi \beta_0}{\sqrt{1 - a^2 \rho_{oot}^2}},
\]

\[
P_F(A|\beta) = \int a \beta - x_e \bar{b} \sigma_e \frac{\phi}{\phi + a \beta - x_e} = \frac{x_{1/3} + a \rho_{oot} \phi \beta_0}{\sqrt{1 - a^2 \rho_{oot}^2}},
\]

\[
P_F(N|\beta) = 1 - P_F(B|\beta) - P_F(A|\beta).
\]

Figure 4 illustrates the behavior of the Brier skill score of biased probability forecasts (20) as a function of the model bias $a$ for $\rho_{oot} = 0.3$ and 0.6. The corresponding calculations were made by substituting $P_F(B|\beta)$ in (11) with the biased probability forecasts (20) and evaluating the resulting expressions with Maple V software. The Brier skill score for the near-normal category varies symmetrically about the origin and has two max-
ima at $a = 1$ and $a = -1$. This is because the probability of the near-normal category is symmetric with respect to the sign of the potentially predictable signal, and it therefore depends only on the absolute value of the signal but not on its sign. The skill of the near-normal forecast is rather insensitive to the bias and takes values that are only slightly above 0 for a wide range of positive and negative values of $a$. The Brier skill score for the above- and below-normal categories is more sensitive to the model errors. The maximum score is attained for $a = 1$ and the score is negative when $a$ is negative.

Note that the above results are obtained for the case when $\beta$ and $\sigma$, and the corresponding conditional probabilities (20) are known exactly. In practice, however, all estimates are subject to sampling variability due to finite sample sizes. Therefore, the dependence of the Brier skill score on the forecast bias $a$ obtained for sampled probability forecasts may differ somewhat from that in Fig. 4. In particular, since the uncertainty in estimating the potentially predictable signal effectively acts as an additional noise, the maximum Brier skill score is achieved for $a > 1$ for finite sample probability forecasts, that is, when the amplitude of the potentially predictable signal is over-simulated in a forecast system.

c. Improved probability forecasts

Pan and Van den Dool (1998) discuss a procedure that attempts to improve probability forecasts based on the count method (12) by combining ensemble probabilities with a single forecast probability that is derived from the model forecast history. Here we consider a somewhat different approach.

Figure 4 suggests that it may be possible to improve biased probability forecasts by using the Gaussian estimator and adjusting the estimates of the forecast signal $\hat{\beta}'$ and standard deviation $\hat{\sigma}_x$ with factors $\hat{\alpha}$ and $\hat{\beta}$, respectively, to maximize the Brier skill score. The resulting improved probability estimators have the form

$$\hat{P}_i(B) = \frac{\hat{\alpha} \hat{\beta}' - x_i}{b \hat{\sigma}_x},$$

$$\hat{P}_i(A) = \frac{\hat{\alpha} \hat{\beta}' - x_i}{b \hat{\sigma}_x},$$

$$\hat{P}_i(N) = 1 - \hat{P}_i(B) - \hat{P}_i(A).$$

The adjusting factors $\hat{\alpha}$ and $\hat{\beta}$ are chosen to satisfy relationship (18), that is, $\hat{\alpha} \hat{\beta}^2 \hat{\sigma}_x^2 + \hat{\beta}^2 \hat{\sigma}_x^2 = \hat{\sigma}_x^2$.

We use a leave-one-out cross-validation procedure (e.g., Barnston 1992) to validate the corrected seasonal probability hindcasts that are obtained with (21). The collection of hindcasts $\{F_{i,t}, i = 1, \ldots, N\}$, $t = 1, \ldots, T$ and verifying observations $\{X_t, t = 1, \ldots, T\}$ for a given season and years $t = 1, \ldots, T$ are repeatedly divided into a $T - 1$ yr training dataset and a 1-yr validation dataset. The boundaries between the forecast categories are determined as in (3) where the observed standard deviation is estimated from the training period. The adjusting factors $\hat{\alpha}$ and $\hat{\beta}$ are selected to minimize the Brier score sum $B[\hat{P}_i(B)] + B[\hat{P}_i(N)] + B[\hat{P}_i(A)]$ in the training data. In principle, separate factors $\hat{\alpha}$ and

![Figure 4](image_url)
\( \hat{b} \) could be determined for each category. However, we expect that the optimal values of \( \hat{a} \) and \( \hat{b} \) are similar for all categories. Our past experience dictates the use of very few parameters in statistical improvement schemes (Kharin and Zwiers 2002). We have therefore reduced the problem to one of estimating a single pair of parameters that minimizes the three-category Brier score sum in the training data. These factors are then used to adjust the values of \( \hat{b}' \) (14) and \( \sigma_y \) (16) to make the probability hindcasts for the withheld period. The whole procedure is repeated for each year. Finally, skill scores are calculated for the resulting collection of the corrected hindcasts.

A technical concern is that the Brier score does not have a parabolic dependence on the rescaling factor \( \hat{a} \). Consequently, the Brier score minimization problem cannot be reduced to that of solving a linear equation as in a standard regression analysis. In this study we applied a golden section search algorithm (Press et al. 1992) to the above minimization problem. Convergence problems were not encountered.

Appendix C summarizes the results of a number of Monte Carlo experiments designed to explore the performance of the proposed statistical skill improvement technique for the forecast system (17) in the Gaussian setting. The statistical method was able to improve the Brier score when biases in the signal-to-noise ratio in the forecast system are sufficiently large. However, the quality of probability forecasts produced by a nearly perfect forecast system is degraded by sampling errors that are the consequence of finite ensemble sizes and sample lengths, a result that is not unexpected given our experience with deterministic statistical forecast improvement.
5. Example—HFP seasonal hindcasts

We used the methods discussed above to analyze 24-member ensemble hindcasts of 26 northern winters [December–January–February (DJF)] produced for the 1969–95 period with the second-generation general circulation model of the Canadian Centre for Climate Modelling and Analysis (CCC AGCM2; McFarlane et al. 1992). These integrations were performed as the part of the Canadian Historical Forecast Project (HFP; Derome et al. 2001).

CCC AGCM2 is a spectral model with a T32 truncation and 10 vertical levels. It contains a comprehensive suite of physical parameterizations of subgrid-scale processes. The model has been used in a number of numerical integrations, several transient climate change simulations (Boer et al. 2000a,b), and the Atmospheric Model Intercomparison Project–like integrations with prescribed observed lower boundary conditions (Zwiers et al. 2000).

The HFP experimental design generally follows that suggested by Boer (1993) for monthly forecasts. The approach for the choice of initial conditions is known as lagged-average forecasting (Hoffman and Kalnay 1983). Each ensemble member is initialized from the National Centers for Environmental Prediction–National Center for Atmospheric Research (NCEP–NCAR) reanalyzed fields (Kalnay et al. 1996) lagged at 6-h intervals prior to the forecast season. That is, the first member is initiated just 6 h before the forecast season, the second member is initialized 12 h before the forecast season, and the remaining members are initialized every 6 h until the end of the forecast season.
hindcasts indicate the regional BSS of the Gaussian probability shading. The next two more lightly shaded bars in-ability hindcastously. The regional BSS of the nonparametric probability hindcasts made with the four forecast variants described previ-
tained for probability hindcasts of DJF region and then substituting the averaged value into (6).

The statistical improvement technique results in better cross-validated regional Brier skill scores for DJF $T_{700}$. However, slightly worse scores are obtained for the DJF $Z_{500}$ probability hindcasts. We will show that the unadjusted probability hindcasts for $T_{700}$ are less reliable than those for $Z_{500}$. Consequently, the temperature hindcasts benefit from the improvement technique, while the more reliable geopotential height hindcasts are slightly degraded because of sampling errors.

The geographical distribution of the Brier skill scores of the nonparametric probability hindcasts $\hat{P}_e$ for DJF $T_{700}$ and the corresponding improved probability hindcasts $\hat{P}$ are shown in Fig. 6. High skill values are confined mainly to the Tropics. There are extensive areas of negative skill for the nonparametric hindcasts in the extratropics. While the statistical improvement technique is generally successful in improving negative skill scores, it has either only modest success or fails to improve positive skill scores.

The probability hindcasts for the near-normal category outperform the climatological forecast in only a few tropical areas. Elsewhere, hindcast skill in this category is near or below 0. The below-normal category is more skillful than its above-normal counterpart. The near-normal category is less skillful than the other two categories.

The geographical distribution of the Brier skill scores of the nonparametric probability hindcasts $\hat{P}_e$ for DJF $T_{700}$ and the corresponding improved probability hindcasts $\hat{P}$ are shown in Fig. 6. High skill values are confined mainly to the Tropics. There are extensive areas of negative skill for the nonparametric hindcasts in the extratropics. While the statistical improvement technique is generally successful in improving negative skill scores, it has either only modest success or fails to improve positive skill scores.

The probability hindcasts for the near-normal category outperform the climatological forecast in only a few tropical areas. Elsewhere, hindcast skill in this category is near or below 0. The below-normal category is more skillful than the above-normal category in both the Tropics and in the subtropical regions. This asymmetry is apparently related to the asymmetry in the lower boundary forcing in these areas, mainly associated with differences in the frequency of El Niño–La Niña events during the HFP period, and to asymmetry in the atmospheric response to the lower boundary forcing. For example, Fig. 7 shows time series of standardized DJF $T_{700}$ anomalies in the

- The Gaussian hindcast $\hat{P}_g$ results in slightly but consistently better Brier scores than the nonparametric hindcast $\hat{P}_e$. Apparently, the Gaussian assumption is an acceptable approximation for the unpredictable variability of $T_{700}$ and $Z_{500}$ on seasonal timescales.
- The Gaussian hindcast $\hat{P}_g$ that uses all available training data to estimate the unpredictable variance ($T = 26$) generally results in slightly better regional Brier scores than the Gaussian hindcast that uses only the current ensemble ($T = 1$) to estimate the unpredictable variance. The additional information obtained by using all available training data apparently overcomes the possible effects of dependence of the noise distribution on the potentially predictable signal.
- The below-normal category for temperature and height is consistently more skillful than its above-normal counterpart. The near-normal category is less skillful than the other two categories.

The geographical distribution of the Brier skill scores of the nonparametric probability hindcasts $\hat{P}_e$ for DJF $T_{700}$ and the corresponding improved probability hindcasts $\hat{P}$ are shown in Fig. 6. High skill values are confined mainly to the Tropics. There are extensive areas of negative skill for the nonparametric hindcasts in the extratropics. While the statistical improvement technique is generally successful in improving negative skill scores, it has either only modest success or fails to improve positive skill scores.

The probability hindcasts for the near-normal category outperform the climatological forecast in only a few tropical areas. Elsewhere, hindcast skill in this category is near or below 0. The below-normal category is more skillful than the above-normal category in both the Tropics and in the subtropical regions. This asymmetry is apparently related to the asymmetry in the lower boundary forcing in these areas, mainly associated with differences in the frequency of El Niño–La Niña events during the HFP period, and to asymmetry in the atmospheric response to the lower boundary forcing. For example, Fig. 7 shows time series of standardized DJF $T_{700}$ anomalies in the

**Fig. 7.** Time series of standardized DJF $T_{700}$ anomalies at 5.6°N, 86.25°W in 1970–95 from NCEP–NCAR reanalysis (thick solid line) and the 24-member ensemble of HFP seasonal hindcasts (dotted lines). Horizontal dashed lines indicate the boundaries between the below-, near-, and above-normal categories. The BSS of the three categories are indicated at the top.

The last, 24th member is initialized 6 days before the forecast season. The monthly mean sea surface temperature anomalies observed in the month prior to the forecast period are persisted throughout the forecast season. These anomalies are obtained from the Global Sea Ice and Sea Surface Temperature dataset (GISST version 2.2; Rayner et al. 1996). Sea ice extent is specified from climatology. The initial snow dataset (GISST version 2.2; Rayner et al. 1996). Sea surface temperature anomalies observed in the month prior to the forecast period are persisted through-
The magnitude of the signal scaling factor $\alpha$ is found to be less than 1, while that of the corresponding noise scaling factor $\beta$ is found to be greater than 1 over most of the globe (not shown). This is not an unexpected result, and it does not necessarily imply that the signal-to-noise ratio is undersimulated in the model compared to that in the observations. The amplitude of the scaling factor $\alpha$ is also affected by the relationship between the model-simulated and observed potentially predictable signals. Indeed, consider a forecast

$$F' = \beta' + \epsilon',$$

with $\sigma_\beta = \sigma_{\beta'}$ and $\sigma_\epsilon = \sigma_\epsilon'$ but correlation $\rho_{\beta,\beta'} < 1$.

That is, the typical amplitude of the predictable and unpredictable components are correctly simulated, but the actual signal anomalies are not perfectly reproduced by the forecast model. In this case, only a fraction $(\rho_{\beta,\beta'}^2)$ of the observed signal variability can be predicted linearly via the relationship:

$$\beta = \rho_{\beta,\beta'} \beta' + \epsilon_\beta.$$

The term $\epsilon_\beta$ includes the portion of the observed signal

![Figure 8](image-url)
that cannot be predicted in this way. Thus, we can rewrite (1) as $X = \rho_{\beta \beta'} \beta' + \epsilon_1$, where $\epsilon_1 = \epsilon + \epsilon'$. Therefore, the forecast signal $\beta'$ must be scaled by $\rho_{\beta \beta'}$, while $\sigma_\epsilon$ must be scaled up correspondingly to obtain a forecast with an unbiased estimate of the predictable signal fraction and unpredictable variance. The lower the correlation between the observed and model-simulated signals, the smaller the scaling factor $\hat{\alpha}$ becomes. As a result, the corrected forecasts have a reduced signal-to-noise ratio. Thus, the effect of the improvement technique is similar to the forecast “randomization” approach (adding additional noise to the forecast) advocated by von Storch (1999).

Figure 8 displays the regional Brier skill scores of the HFP probability hindcasts in the Tropics and in the North American sector as a function of the ensemble size. The nonparametric hindcast ($\tilde{F}_N$, dotted lines) is substantially affected by sampling errors when ensembles are small. The dashed and solid line curves indicate the Brier skill scores of the Gaussian hindcasts $\tilde{F}_G$ for $T = 26$, and improved hindcasts $\tilde{F}_I$, respectively. The performance gain is substantial for small ensemble sizes.
in all three categories and is particularly large for the near-normal category. The improved \( T_{700} \) hindcasts are more skillful than the other two unadjusted hindcasts for all ensemble sizes. Evidence of performance differences between \( \hat{P}_g \) and \( \hat{P}_i \) is less pronounced for \( Z_{500} \). The unadjusted Gaussian hindcast \( \hat{P}_g \) for \( Z_{500} \) marginally outperforms the statistically improved hindcast \( \hat{P}_i \) for ensembles larger than approximately \( N = 10 \).

**b. Probability hindcast attributes**

Attributes diagrams (Hsu and Murphy 1986; Wilks 1995) are another way to represent the quality of probabilistic forecasts. Essentially an attributes diagram displays the dependence of the relative frequency of an observed event on the forecast probability. This is done by categorizing all probability forecasts in a number of bins and estimating the relative frequency of forecasts and the corresponding observed event relative frequency for each bin.

Attributes diagrams for the HFP \( T_{700} \) and \( Z_{500} \) probability hindcasts in the Tropics are shown in Figs. 9 and 10. The points on the diagrams are the area-averaged observed event relative frequencies plotted as a function of forecast probabilities. These points would lie on the diagonal connecting the points (0, 0) and (1, 1) for perfectly reliable forecasts. The attributes diagrams provide a geometrical interpretation of the Brier score decom-
Several approaches for deriving probabilistic information from an ensemble of deterministic forecasts are discussed and their sampling properties compared. When the noise is approximately normally distributed, the probability forecasts obtained with the fitted Gaussian distribution have better sampling properties than those obtained by counting the number of the ensemble members in the event category. Using the Gaussian probability estimator, a statistical skill improvement technique is proposed to reduce biases in probability forecasts. The approach is based on the minimization of the Brier score (mean-square error of the probability forecasts) by adjusting the amplitude of the model-simulated potentially predictable signal and noise variance estimates.

The proposed probability estimators and statistical improvement technique are tested on the 24-member ensemble of northern winter seasonal hindcasts of 700-hPa temperature and 500-hPa height produced with the second generation CCCma general circulation model. The Gaussian hindcasts result in somewhat better Brier scores than the nonparametric hindcasts. Apparently, the assumption of Gaussian noise is an acceptable approximation for $T_{700}$ and $Z_{500}$ on seasonal timescales. Also, the Gaussian hindcasts that use all available data to estimate the noise variance have slightly better skill scores.

The improvement technique is moderately successful for $T_{700}$ but generally fails to improve the Brier skill scores of unadjusted $Z_{500}$ probability hindcasts. We found that the unadjusted probability hindcasts for $T_{700}$ are less reliable than those for $Z_{500}$. Consequently, the temperature hindcasts benefit from the improvement technique, while the more reliable geopotential height hindcasts are slightly degraded because of sampling errors.

Acknowledgments. We thank George Boer and two anonymous reviewers for careful reading of an earlier version of the manuscript and providing us with many constructive and helpful comments. We also thank Normand Gagnon for producing the 24-member hindcasts employed in this study.

APPENDIX A

Brier Score Decomposition

Given the joint probability density distribution $f(E, P)$ of the predictand $E$ and its forecast $P$ the Brier score can formally be written as

$$B = \mathbb{E}(P - E)^2 = \int_P \int_E f(E, P)(P - E)^2 \, dE \, dP.$$
\[ f(E, P) = f(E|P)f_P(P), \]
where \( f(E|P) \) is the conditional probability distribution of the event \( E \) for each possible forecast \( P \) and \( f_P(P) \) is the marginal probability density function of the forecasts. Because \( E \) is a discrete random variable, the conditional probability density function \( f(E|P) \) must be expressed as a weighted sum of Dirac delta functions. In the case of a dichotomous event, \( E \) takes only values 1 or 0 depending upon whether the event occurred. In this case the conditional distribution \( f(E|P) \) is completely described by a single parameter, the conditional probability of the event occurrence \( E \). Also notice that

\[ \int_P f_P(P)f(E|P)\,dP = E(E) = P_c. \]

Using these expressions we have

\[
B = \int_P \int_E f(E, P)(P - E)^2\,dE\,dP = \int_P f(E = 1, P)(P - 1)^2\,dP + \int_P f(E = 0, P)(P - 0)^2\,dP,
\]

\[
= \int_P f_P(P)[E(E|P) - P_c]^2\,dP + P_c(1 - P_c). \]

It is readily shown that the uncertainty term above is equal to the variance of \( E \).

\[ \text{APPENDIX B} \]

\[ \text{Probability Forecast Sampling Properties} \]

In this section we briefly discuss the sampling properties of the probability forecasts introduced in section 4a. We consider three estimators: the nonparametric forecast \( \hat{P}_a \) (12), the Gaussian forecast \( \hat{P}_g \) (13), where \( \hat{B} \) and \( \hat{\sigma}_e \) are obtained from (14) and (15), and the Gaussian forecast \( \hat{P}_g \) where \( \hat{\sigma}_e \) is obtained from (16) for \( T = 25 \).

Figure B1 illustrates several sampling properties of these forecasts. In the case of \( \hat{P}_a \), the properties are derived directly from the corresponding binomial distribution, while those of \( \hat{P}_g \) are obtained from Monte Carlo simulations that produced 10 000 realizations of \( \hat{P}_g \) with \( T = 1 \), and 1000 realizations of \( \hat{P}_g \) with \( T = 25 \).

As we noted earlier, nonparametric forecast \( \hat{P}_a \) is unbiased. The “sawtooth” curves that describe its 80% confidence intervals are the result of the discrete domain range with values that are multiples of \( 1/N \). The bias of \( \hat{P}_g \) is generally small for ensemble sizes larger than 5. The forecast with \( T = 25 \) has a noticeable bias for small ensembles, which is nonetheless small relative to the rms errors. As expected, \( \hat{P}_g \) with \( T = 25 \) has the smallest sampling errors. Forecasts of the near-normal category made with \( \hat{P}_g \) benefit substantially from larger ensembles. The rms errors of forecasts of the above-normal category made with \( \hat{P}_g \) for \( T = 1 \) and \( T = 25 \) become nearly the same for ensembles with more than 10 members.
APPENDIX C

Sampling Properties of Statistically Improved Probability Forecasts

Monte Carlo experiments were also used to explore the performance of the statistical improvement technique proposed in section 4c. Some results of Monte Carlo simulations are summarized in Figs. C1 and C2. Here we discuss the results only for the below (above) normal category.

Figure C1 shows the Brier skill score of the statistically improved probability forecasts as function of the ensemble size on the horizontal axis and the model bias on the vertical axis for the three values of the potential predictability $\rho_{pot}^2 = 0.15, 0.3, \text{ and } 0.6$, with training period length $T = 25$. The model bias is expressed in terms of the adjusting factor $a$. The extreme values of $a$ on the plot indicate a fully deterministic model forecast system. Negative values of $a$ indicate that the simulated potentially predictable signal is of opposite sign. A no-skill forecast system is obtained for $a = 0$. A system with the perfect signal-to-noise ratio ($a = 1$) is indicated by the horizontal dashed lines. The shaded areas indicate the parameter space for which the statistically improved probability forecasts have better Brier skill scores than the unadjusted Gaussian forecasts.

As one would expect, the performance of the statistical technique improves with the longer training periods (not shown) and larger ensembles. The quality of the “improved” forecasts is worse than that of the unadjusted forecasts in areas around the “perfect-model” line and near the “no-skill” line. It is understandable...
that when a forecast system is nearly perfect, that is, when the signal-to-noise ratio in the model is close to
that in the observations, sampling errors involved in a statistical improvement technique, by trying to adjust this ratio, will only degrade the quality of already good forecasts. Sampling errors become particularly dominant in a forecast system with a small signal-to-noise ratio.

Figure C2 presents the results of the Monte Carlo simulations in a slightly different format. It shows the Brier skill score of the unadjusted and improved probability forecasts as a function of the model bias for the potential predictability $\rho_{\text{pot}}^2 = 0.3$ and ensemble sizes $N = 6, 24, \text{ and } 256$. The training length in all cases is $T = 25$. The main source of sampling errors in large ensembles such as $N = 256$ is associated with the finite length of the training periods. While it is unlikely that the training period can be substantially extended in the near future, there is no fundamental barrier to increasing the ensemble size, apart from the obvious cost and performance issues. Thus this exercise allows us to explore the gains in skill that might be achieved by the statistical technique in the best-case scenario of a very large ensemble.

The skill score of the unadjusted Gaussian forecasts is indicated by the solid line, while the dashed line indicates the skill score of the improved probability forecasts. The horizontal dotted line indicates the potential Brier skill score $\text{BSS}_{\text{pot}}$ that can be achieved for the specified potential predictability. Vertical horizontal
lines indicate a forecast system with a perfect signal-to-noise ratio. Negative bias $a$ indicates that the sign of the simulated potentially predictable signal is reversed. The statistical method performs rather well when the amplitude of the signal in the model is oversimulated, or is of opposite sign. However, the statistically improved forecast is often inferior to the raw forecast in a wide range of positive values between 0 and a point just above the perfect signal-to-noise ratio. Increasing the ensemble size leads not only to better skill scores of the unadjusted forecasts but also to the better improvements in skill from the statistical improvement scheme, especially when the signal amplitude is undersimulated in the model.

REFERENCES