# NOTES AND CORRESPONDENCE

## Calibration of Probabilistic Forecasts of Binary Events

CRISTINA PRIMO

*European Centre for Medium-Range Weather Forecasts, Reading, United Kingdom*

CHRISTOPHER A. T. FERRO, IAN T. JOLLIFFE, AND DAVID B. STEPHENSON

*School of Engineering, Computing and Mathematics, University of Exeter, Exeter, United Kingdom*

(Manuscript received 4 March 2008, in final form 2 October 2008)

ABSTRACT

Probabilistic forecasts of atmospheric variables are often given as relative frequencies obtained from ensembles of deterministic forecasts. The detrimental effects of imperfect models and initial conditions on the quality of such forecasts can be mitigated by calibration. This paper shows that Bayesian methods currently used to incorporate prior information can be written as special cases of a beta-binomial model and correspond to a linear calibration of the relative frequencies. These methods are compared with a nonlinear calibration technique (i.e., logistic regression) using real precipitation forecasts. Calibration is found to be advantageous in all cases considered, and logistic regression is preferable to linear methods.

### 1. Introduction

Probabilistic forecasts represent the uncertainty in a prediction by a probability distribution for the predictand. This distribution may be derived from historical errors of deterministic forecasts or from ensemble forecasts (see Leith 1974; Ehrendorfer 1997; Stephenson and Doblas-Reyes 2000, and references therein). In the latter case, probabilistic forecasts for binary events are often obtained as the relative frequency with which the event occurs in the ensemble. For perfect forecasting models and perfect ensembles, observations behave like draws from the ensemble distribution and relative frequencies will make good forecasts. In practice, however, models are imperfect (Ferranti et al. 2002) and ensemble generation techniques do not sample randomly from the probability distribution of initial-condition uncertainty (Hamill et al. 2000, 2003; Wang and Bishop 2003).

Various techniques have therefore been proposed for improving such probabilistic forecasts. One approach is to combine model forecasts with a prior belief about the value of the predictand (Robertson et al. 2004; Rajagopalan et al. 2002). For example, Bayesian techniques model the prior belief that the event happens (often using past data) and update it using the new information available from the numerical model via Bayes's theorem. Another approach is calibration (Gneiting et al. 2007), which adjusts forecasts based on past performance. Most seasonal forecasting centers calibrate their forecasts simply by adding or scaling by constants to correct biases in means and variances (Stephenson 2008). This simple procedure is based solely on the mean and variance of past forecasts and past observations and ignores other information about the joint distribution of past observations and forecasts (e.g., the skill of the forecasts). Recalibrating the forecasts with a regression model of past forecasts on past observations is often superior (Stephenson 2008).

Note that these combination methods may lead to linear transformations of the original forecasts (some examples will be presented in section 3). If the parameters of the linear transformation are chosen to optimize some measure of past performance, these methods can also be presented as linear calibration techniques.

Calibration can improve forecast skill by improving reliability (Murphy 1973), even though reliability is

*Corresponding author address:* Dr. Cristina Primo, European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading RG2 9AX, United Kingdom.
E-mail: cristina.primo@ecmwf.int

DOI: 10.1175/2008MWR2579.1

not a necessary condition for skill [see Jolliffe and Stephenson (2005) and Glahn (2004) for further discussion]. When the interest focuses on binary events, the binary nature makes common statistical calibration methods (Gneiting et al. 2005; Wilson and Valle 2002) inappropriate, and new methods that take into account the discreteness are needed.

This study reviews different calibration methods for probabilistic forecasts of a binary event. We show how some Bayesian methods, used to incorporate prior belief about the occurrence of the binary event, can be written as special cases of a beta-binomial model and correspond to a linear calibration of the relative frequencies. The common framework underlying these apparently different methods has not been noted before. A nonlinear calibration method in which past data are used to model the relationship between observations and forecasts is also reviewed and is compared with the combination methods. We find that calibration improves probabilistic forecasts of binary events presented in this study by improving reliability. Nonlinear calibration does as well as or better than other methods. However, both linear and nonlinear calibration techniques have consequences for warning systems because of a possible reduction of the range of issued forecasts.

This paper is organized as follows. Section 2 introduces notation and uncalibrated probabilistic forecasts of a binary event. Section 3 identifies some Bayesian methods as special cases of a beta-binomial model. Section 4 reviews a nonlinear way to calibrate probabilistic forecasts. Section 5 illustrates all the previous calibration methods using a real weather example. Last, section 6 gives some concluding remarks and caveats.

## 2. Uncalibrated probabilistic forecasts of binary events

This paper is focused on forecasting the future state of an observable binary event. Let $Y_t$ be the binary variable representing the observation of the event at time $t$: $Y_t$ is 1 when the event happens and is 0 otherwise. The observations are taken at a discrete set of time points. The index $t = 1, 2, \ldots$, denotes the index of times on which the event is observed, so $\{Y_1, \ldots, Y_T\}$ represents the set of past observations until the present time $T$. Uncertainty about $Y_t$ at future time $t$ can be represented by a Bernoulli distribution with probability $p_t$. The aim of probabilistic forecasting is to provide the best $p_t$ for the observable $Y_t$ for the time of interest ($t > T$).

Numerical models provide an ensemble of forecasts: $\{X_{ti}: i = 1, \ldots, m\}$ at times $t = 1, 2, \ldots$, where $m$ denotes the number of ensemble members. The $i$th forecast $X_{ti}$ is also a binary variable: $X_{ti} = 1$ if the event is forecast and is 0 otherwise. The ensemble forecasts at any time $t$ are assumed to be a set of independent identically distributed Bernoulli variables $X_{ti} \sim \text{Ber}(q_t)$. The probability $p_t$ may differ from $q_t$. The simplest approach is to assume that the model is perfect so that the probability that an ensemble member forecasts the event is the same as the probability that the event happens. In this case, the natural most frequent estimate of $p_t$ is the relative frequency of occurrence, $\hat{p}_t = n_t/m$, where $n_t$ is the number of members that forecast the event:

$$n_t = \sum_{i=1}^{m} X_{ti}.$$

This probability estimator is easy to obtain but has disadvantages. The probabilities can take only a finite set of discrete values, the probabilities can be 0 or 1, and there is no estimate of the uncertainty on the predicted probability.

We have assumed that the probability that an ensemble member forecasts the event is the same as the probability that the event happens. However, in practice, models are not perfect and so this assumption might not be true. One can try to overcome model imperfections by calibrating the original forecasts.

## 3. Linear calibration: Beta-binomial framework

The Bayesian framework is consistent with the fact that many users are aware of the uncertainty inherent in a limited ensemble size. For example, Katz and Ehrendorfer (2006) use a Bayesian approach with the beta distribution as a prior distribution to introduce such uncertainty into the decision process. This allows them to take into account uncertainty in estimating a forecast probability from a limited number of ensemble members. The choice of the prior distribution plays an essential role. The prior distribution $f(p_t)$ can be conveniently modeled using the beta distribution, $p_t \sim \text{beta}(\alpha, \beta)$, where $\alpha$ and $\beta > 0$ (appendix A in Epstein 1985; chapter 4 in Wilks 2006b). This distribution, defined for values between 0 and 1, is flexible, with a density that can be either convex or concave and skew or symmetric.

For binary events, it is convenient to assume the prior distribution to be the beta distribution since it has the advantage of being conjugate when combined with a binomial likelihood (Epstein 1985), so that a beta posterior distribution is obtained:

$$\begin{aligned} f(p_t|n_t) \propto f(p_t) f(n_t|p_t) &= \text{beta}(p_t; \alpha, \beta)\, \text{bin}(n_t; m, p_t) \\ &\propto \text{beta}(\alpha + n_t, \beta + m - n_t). \end{aligned} \tag{1}$$

The forecaster is not primarily interested in the posterior distribution of $p_t$, but in the predictive distribution of $Y_t$. The predictive distribution of $Y_t$ is the conditional distribution of $Y_t$ given the ensemble forecast. This can be written as an average with respect to the posterior distribution $\pi$ for $p_t$ as follows: $\Pr(Y_t = 1|n_t) = \int \Pr(Y_t = 1|p_t)\pi(p_t|n_t)\, dp_t = \int p_t \pi(p_t|n_t)\, dp_t = E(p_t|n_t)$. Thus, our prediction for the event $\{Y_t = 1\}$ is the posterior mean for $p_t$. Since the posterior distribution of the probability that the event happens is a beta distribution with parameters $\alpha + n_t$ and $\beta + m - n_t$, the posterior expectation is

$$E(p_t|n_t) = \frac{\alpha + n_t}{\alpha + \beta + m} = \upsilon \frac{n_t}{m} + (1 - \upsilon)\frac{\alpha}{\alpha + \beta}, \qquad (2)$$

where $\upsilon = m/(\alpha + \beta + m)$. Thus, this estimate is a weighted average of the relative frequencies and the prior mean $\alpha/(\alpha + \beta)$ with relative weights proportional to sample size $m$ and what can be thought of as the "effective" prior sample size $(\alpha + \beta)$.

Since the parameters that model our prior belief ($\alpha$ and $\beta$) are constants, the beta-binomial approach leads to a linear transformation of the relative frequencies:

$$E(p_t|n_t) = \gamma + \delta(n_t/m), \qquad (3)$$

where $\gamma = \alpha/(\alpha + \beta + m)$ and $\delta = m/(\alpha + \beta + m)$. The beta-binomial approach is for combining model forecasts with a prior belief about the value of the predict and; however, if $\gamma$ and $\delta$ are chosen to optimize some measure of past performance, then the beta-binomial approach can also be considered to be linear calibration of the relative frequencies.

Many current methods used by the climatological community to produce probabilistic forecasts of a binary event can be written as special cases of a beta-binomial model, and therefore they linearly calibrate the relative frequencies. These methods are different from each other in the choice of both $\alpha$ and $\beta$ parameters, as follows.

If the prior distribution is assumed to be beta(0, 0), defined as the limit as $\alpha$ and $\beta$ tend to 0, this corresponds to a prior distribution with probability mass function at 0 and 1. In this case, the posterior probability converges to the relative frequency as $\alpha$ and $\beta$ converge to 0:

$$\lim_{\alpha,\beta \to 0} E(p_t|n_t) = n_t/m. \qquad (4)$$

Instead of choosing both $\alpha$ and $\beta$ parameters directly, one can choose a central point (mean or mode) and some measure of the spread for the prior distribution. For the prior mean one can consider the climatological mean, that is, the long-term frequency of the observed event of interest: set $\alpha/(\alpha + \beta) = \bar{y}$, where

$$\bar{y} = \frac{1}{T-1}\sum_{t=1}^{T-1} Y_t.$$

For binary variables, this sample mean coincides with the frequency $p$ that the event happens in the sample ($\bar{y} = p$). The choice of the spread is more difficult since sample variance of the observations is not a good choice, especially if the prior distribution is highly skewed. An alternative is to use how many extra ensemble members $m'$ the prior information is worth. This number can then be equated to $m' = \alpha + \beta$. The $\alpha$ and $\beta$ parameters are then given by

$$\alpha = m'p \quad \text{and} \quad \beta = m'(1 - p). \qquad (5)$$

The probability estimate is then obtained by

$$E(p_t|n_t) = \frac{\widehat{m'}p + n_t}{\widehat{m'} + m}. \qquad (6)$$

However, there is no objective criterion to estimate $m'$.

Rajagopalan et al. (2002) and Robertson et al. (2004), introduced a different, but related, approach to estimate both $\alpha$ and $\beta$ parameters. They argue that climatological information given by past observations is combined with general circulation model forecasts and therefore a weight may be introduced to give different importance to both prior belief and model forecasts. They use a prior beta distribution whose parameters depend on a weight $w$ as follows: $\alpha = w^{-1}Tp$ and $\beta = w^{-1}T(1 - p)$, where $p$ is the frequency with which the event happens in the sample of past observations and $T$ is the sample size of climatology (number of past observations). The quantities $p$ and $T$ are called $P_k(x)$ and $n$, respectively, in Rajagopalan et al. (2002), and Robertson et al. (2004). The estimate of $p_t$ given the ensemble of forecasts is then given by

$$E(p_t|n_t) = \frac{Tp + wn_t}{T + wm} = w_1 \frac{n_t}{m} + (1 - w_1)p, \qquad (7)$$

where $w_1 = wm/(T + wm)$. The choice of the weight $w$ is found by maximizing what the authors refer to as the "posterior likelihood" function (Rajagopalan et al. 2002). This choice is equivalent to minimizing the logarithmic score (a popular verification score; Winkler 1968). One could also imagine optimizing other scores such as the quadratic Brier score (Brier 1950). Note that this method was developed for multicategory forecasts, but in this work we focus on binary events. Inspection of Eq. (7) shows that $E(p_t|n_t) = p$ when $n_t/m = p$ and, since Rajagopalan et al. (2002) restrict $w$ to be positive,

$0 < w_1 < 1$. Therefore, this method corresponds to a linear calibration curve that passes through the point $(p, p)$ and has a slope in the interval (0, 1).

The latter two methods estimate alpha and beta from data and so are forms of empirical Bayes approaches.

Table 1 summarizes the different choices of the prior distribution, the corresponding probabilistic forecasts for a future time $t > T$, and both the intercept and the slope of the different linear calibration techniques [$\gamma$ and $\delta$ in Eq. (3), respectively].

## 4. Nonlinear calibration: Logistic regression

When the interest focuses on binary events it is inappropriate to use a model based on the regression of binary forecasts on binary observations because predictions can be made that are impossible. In particular, if the mean of the dependent variable $Y_t$, which is the probability $p_t$, is modeled as a linear function of predictors, the predicted value may lie outside the range from 0 to 1. To overcome this problem, generalized linear models (GLMs; McCullagh and Nelder 1989) provides a suitable framework. As well as allowing nonlinearity, using a so-called link function, these models incorporate a range of distributions, including Bernouilli, for the dependent variable, whereas much of linear regression assumes a Gaussian distribution.

An example of a GLM is the logistic regression model (Collett 2002), in which the dependent variable is assumed to be binomial and the link function is the logit transformation:

$$Y_t|p_t \sim \text{Ber}(p_t) \quad \text{and} \tag{8}$$

$$\text{logit}(p_t) = \log[p_t/(1-p_t)] = \beta_0 + \beta_1 h_{t1} + \cdots + \beta_q h_{tq}, \tag{9}$$

where $\beta_0, \beta_1, \ldots, \beta_q$ are the regression coefficients and $h_{t1}, \ldots, h_{tq}$ are the predictor variables. Note that $\text{logit}(p_t)$ can theoretically assume any value between minus and plus infinity. It is easy to check that the predicted values of $p_t$ are in the range of 0–1. One can use maximum likelihood estimation to fit the logistic regression model (Collett 2002). Logistic regression is widely used by the statistical community and has recently been used for probability-of-precipitation forecasts (Wilks 2006a; Wilks and Hamill 2007).

The predictor variables will be a function of the ensemble forecasts. We assume that the members within any particular ensemble are independent and identically distributed. For a single ensemble, all members should be equally weighted, and therefore instead of choosing each $X_{ti}$ as a predictor variable only one predictor variable is chosen: $h_t$, and it will be a function of the symmetric combination $n_t$. Among the possible choices for the predictor, the natural choice would be to use the relative frequencies: $h_t = n_t/m$. However, it might be useful to relate $\text{logit}(p_t)$ to the logit transformation of the forecasts to have both predictands and predictors on the same scale. Furthermore, prior information could be combined by using $h_t = \text{logit}[E(p_t|n_t)]$ for $\alpha$ and $\beta$ not equal to 0. The prior parameters can be chosen to minimize a score, for instance the logarithmic score: $\alpha = w^{-1}Tp$ and $\beta = w^{-1}T(1-p)$ (Rajagopalan et al. 2002). There is no objective rule as to how best to choose the predictors, though the choice might be based on a diagnostic measure, such as minimizing the errors of the fit. In addition, a check of the residuals of the fit (e.g., deviance residuals) should be done to ensure that the model assumptions are valid and the model is not inappropriate. For simplicity, we will just illustrate the logistic regression method with the Rajagopalan et al. (2002) predictor.

## 5. Meteorological example

In this section we illustrate the methods described in sections 2–4. Our data are daily total precipitation forecasts generated by the Ensemble Prediction System (EPS) of the European Centre for Medium-Range Weather Forecasts (Molteni et al. 1996). Each ensemble has 50 members. We assume that the members are independent and identically distributed. Forecasts are daily data in the 3-month period of December–February from 1997 to 2006, at a single grid point (51°N, 1°W) near Reading in the United Kingdom. The first 4 yr define the training period, and the last 5 yr define the verification period. We use 72-h-ahead forecasts in the expectation that calibration will have a significant impact on the quality of the forecasts at this lead time. These data are verified with observations of daily precipitation in Reading at the University of Reading atmospheric observatory.

This example is focused on showing whether calibration improves forecasts of wet days in Reading. In this work a wet day is defined by precipitation exceeding 0.1 mm. Thus, the binary event is defined as $Y_t = 1$ when observed precipitation is above 0.1 mm and $Y_t = 0$ otherwise; $X_{ti} = 1$ if the $i$th forecast ensemble member is above 0.1 mm and $X_{ti} = 0$ otherwise.

Reports from 361 observations and 18 050 forecasts (50 members and 361 daily data) of wet days in Reading were considered for the training period. The model forecast a wet day 14 637 times (81% of the forecasts). However, wet days were observed on 9850 occasions (54% of the observations). Hence, the model clearly

TABLE 1. Parameters of the beta ($\alpha$, $\beta$) prior distribution, the corresponding probabilistic forecasts for a future time $t > T$, and coefficients for the linear calibration.

| Approach | $\alpha$ | $\beta$ | $\hat{p}_t \mid n_t$ | Intercept ($\gamma$) | Slope ($\delta$) |
|---|---|---|---|---|---|
| Relative frequencies | 0 | 0 | $n_t/m$ | 0 | 1 |
| Central point and spread | $m'p$ | $m'(1-p)$ | $(m'p + n_t)/(m' + m)$ | $m'p/(m' + m)$ | $m/(m' + m)$ |
| RLZ | $w^{-1}Tp$ | $w^{-1}T(1-p)$ | $(Tp + wn_t)/(T + wm)$ | $Tp/(T + wm)$ | $mw/(T + wm)$ |

overforecast the event. There were cases in which the event did not happen but all members forecast it. Similar results were found for the verification period, during which wet days were forecast 17 740 times but were observed on 11 700 occasions out of 22 550 forecasts.

Figure 1a shows the probabilities estimated by four different calibration methods against naive relative frequencies during the verification period: the methods are climatology (which is a horizontal line, $p = 0.54$), the relative frequencies (diagonal), probabilities obtained by the method described in Rajagopalan et al. (2002; hereinafter RLZ), and logistic regression using the logit transformation of RLZ probabilities as predictors. According to Eq. (7), RLZ probabilities are a linear combination of the relative frequencies with slope always positive (the method forces the weights to be always positive) and less than 1. In this example the slope is $w_1 = 0.48$. In addition, when $n_t = pm$ the probabilities are equal to climatology, and therefore this method is a linear transformation of the probabilities forced to pass through the climatological point ($p$, $p$). Hence, RLZ probabilities are always represented by a straight line through the point ($p$, $p$) and are obtained by rotating the diagonal (relative frequencies) toward the horizontal line (climatology). For example, it cannot simultaneously shift all values upward or all values downward. Frequencies less than climatology are calibrated upward and so they cannot take smaller values than climatology after calibration, whereas those greater than climatology are calibrated downward and so they cannot exceed climatology after calibration. This makes this method inappropriate to correct bias in means of the forecasts. Thus, this method always reduces the maximum forecast probability and the range of calibrated forecasts is bounded away from 0 and 1. Logistic regression does not have this problem because it transforms the relative frequencies in a nonlinear way, so that they are nonlinear curves. Thus, logistic regression lets the minimum forecast be close to 0. Since the model forecasts overforecast the observation, it is easy to see how logistic regression corrects relative frequencies to reduce them. Simulations not included in this work have shown that the RLZ method does not distinguish whether the model is overforecasting or underforecasting, and it provides similar probabilistic forecasts in both cases. Nevertheless, logistic regression also reduces the range of issued forecasts at the higher values.

To check whether calibration improves forecasts, the Brier score BS and reliability terms of the forecasts have been calculated (Brier 1950). All of the calibration approaches provide better probabilistic forecasts than do naive relative frequencies, but it is nonlinear calibration that achieves the greatest reduction:

$$\begin{aligned} \mathrm{BS}_{\mathrm{RelFreq}} = 0.308 > \mathrm{BS}_{\mathrm{clim}} = 0.252 > \mathrm{BS}_{\mathrm{RLZ}} \\ = 0.240 > \mathrm{BS}_{\mathrm{LogReg}} = 0.209. \end{aligned} \tag{10}$$

The probability score decomposition proposed by Murphy (1973) has been considered to distinguish the two main aspects of the forecast performance: reliability and resolution. Murphy's decomposition consists of three terms:

$$\sum_{k=1}^{N} \frac{N_k}{M}(P_k - \overline{o_k})^2 - \sum_{k=1}^{N} \frac{N_k}{M}(\overline{o_k} - \bar{o})^2 + \bar{o}(1-\bar{o}), \tag{11}$$

when a sample of $M$ forecasts has been divided into $N$ categories, each comprising $N_k$ forecasts of a probability $P_k$, $\overline{o_k}$ being the observed frequency when the forecast was lying in that category and $\bar{o}$ being the observed frequency in the whole sample. The first term is the reliability, the second is the resolution, and the third term is the uncertainty. Since all of the calibration techniques produce calibration curves that are strictly monotonic functions of the relative frequencies, the resolution and the uncertainty of all the methods are equal. Thus, the improvement of the Brier score when forecasts are calibrated is due to an improvement of the reliability component:

$$\begin{aligned} \mathrm{Rel}_{\mathrm{RelFreq}} = 0.131 > \mathrm{Rel}_{\mathrm{RLZ}} = 0.064 > \mathrm{Rel}_{\mathrm{LogReg}} \\ = 0.032 > \mathrm{Rel}_{\mathrm{clim}} = 0.006. \end{aligned} \tag{12}$$

When forecasts are equal to climatology and the frequency of the event does not change between the training period and the verification period, then the reliability component is close to 0 (note that the climatology is computed in the training period). However, although climatology may have near perfect reliability, it may

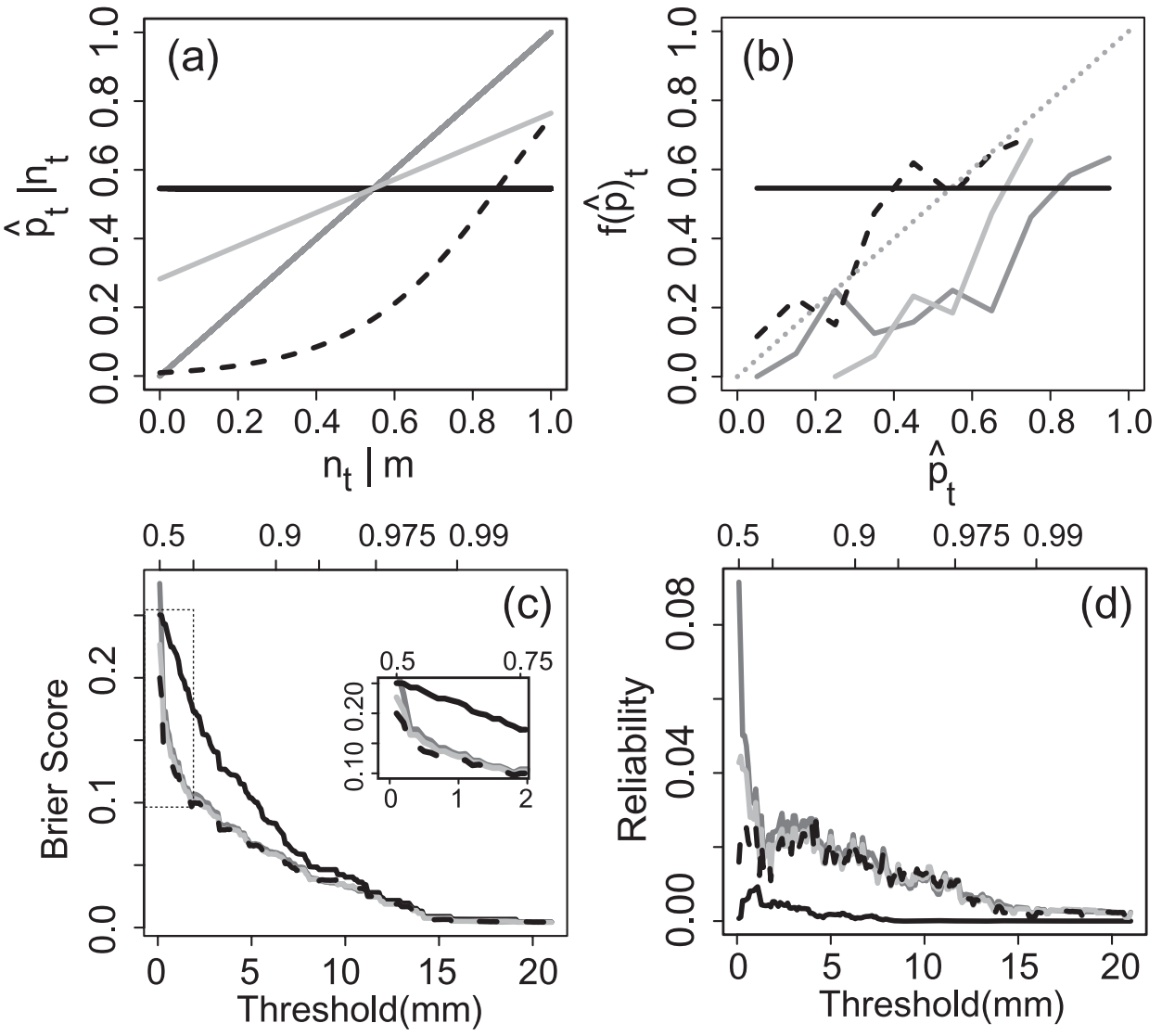


FIG. 1. (a) Calibration of relative frequencies for the precipitation in Reading: climatology (black), relative frequencies (dark gray), RLZ method (light gray), and logistic regression using the logit transformation of RLZ probabilities as predictors (dashed black line); (b) reliability diagram of the probabilistic forecasts; (c) Brier score; and (d) reliability term of precipitation in Reading forecasts for different thresholds and calibration techniques. The upper horizontal axis represents the percentiles. Inset in (c) zooms in on the lowest threshold values.

have the worst Brier score because climatology is like a flat calibration curve, which is not strictly monotonic and makes the resolution poor.

Figure 1b plots the stratified observed frequency against the forecast probability (reliability diagram). The range of forecast probabilities has been divided into 10 bins (each of width 0.1). The diagonal line indicates perfect reliability (average observed frequency equal to predicted probability for each category), and the horizontal line represents the climatological frequency. Nonlinear calibration is closer to the diagonal than relative frequencies or linear calibration.

In this particular example, results have been illustrated for a commonly occurring event. However, these methods can also be applied to rarer events above larger thresholds. Figure 1c shows how the Brier score evolves depending on the threshold. The upper horizontal axis represents the percentiles. The inset zooms in on the lowest threshold values. Using linear calibration, the Brier score decreases, but nonlinear calibration always provides the best forecasts. This improvement decreases for extreme events since the Brier score tends to 0 as the event becomes increasingly rare. Figure 1d shows the evolution of the reliability term.

## 6. Conclusions

Different ways of calibrating forecasts of binary events with past data have been considered. Currently the most common way to obtain probabilistic forecasts from an ensemble of forecasts is by using relative frequencies, but numerical models are not perfect and so calibrating them using past data can improve forecasts. This paper presents some Bayesian approaches used by the climatological community to produce probabilistic forecasts of binary events as special cases of a beta-binomial model, and so they are equivalent to a linear calibration. However, this calibration differs from the standard calibration techniques in that the relationship is not provided by the joint distribution of past observations and forecasts but from modeling the probability that the event occurs. Different choices of the parameters of the prior distribution representing the prior belief will determine the intercept and slope in the linear calibration.

The approach described by RLZ that is currently used by the meteorological community has been presented as a linear calibration technique that minimizes the logarithmic score for the particular case of binary events. In this approach the climatological value is never changed and relative frequencies are calibrated toward the climatology. For example, it cannot simultaneously shift all values upward or all values downward. This makes this method inappropriate to correct bias in means of the forecasts. In addition, the range of calibrated forecasts is bounded away from 0 and 1.

This work also presents a nonlinear calibration technique already widely used by the statistical community and recently used for probability of precipitation forecasts, namely, logistic regression. Logistic regression calibrates forecasts using a nonlinear curve and is a flexible method able to correct bias in the forecasts.

Bayesian methods and nonlinear calibration can be used together by a two-step approach. First, prior belief may be modeled and updated with ensemble predictions to obtain a posterior distribution. A first estimate of the probabilistic forecasts can be obtained from this posterior distribution. Then, forecasts might be recalibrated by a nonlinear method (logistic regression). This work has investigated one version of this two-step approach, namely, by calibrating the forecasts provided by RLZ.

These calibration techniques have been illustrated by a real meteorological case. The example shows that both calibration methods investigated always improve the Brier score relative to the probabilistic forecasts given by the relative frequencies. The improvement of the Brier score when the forecasts are calibrated is due to an improvement of the reliability term. Calibration does not affect resolution or uncertainty. For these examples, the relative-frequencies technique always has larger Brier scores because of poor reliability terms, and nonlinear calibration always improves linear calibration.

Forecasters interested in issuing warnings need to bear in mind that whenever the forecasts are calibrated the range of the obtained probabilistic forecasts is reduced. After calibration they do not range from 0 to 1 anymore. The range reduction depends on how poor the model was in the past. If the model overforecast the event, the calibration tends to reduce high probabilities and will never provide us with probabilities close to 1. Conversely, if the model underforecasts the event, then calibration tends to increase the probabilistic forecasts, avoiding probabilities close to 0. Thus, if the warning system relies on the forecast probability of an event exceeding a threshold and the threshold exceeds the upper bound for the forecast, then the warning will never be issued.

Logistic regression is a nonlinear method that calibrates probabilistic forecasts in a more flexible way than does the RLZ method, but if the prediction model has been very poor in the past then the probability forecast's range is still bounded away from 0 and 1. More work is needed to address this range issue.

*Acknowledgments.* Partial support for this work was provided by the Spanish government. We thank the European Centre for Medium-Range Weather Forecasts for providing the EPS data and the Department of Meteorology of the University of Reading for providing the rainfall observations. We are grateful to the editor and two anonymous referees for comments that helped to improve the paper.

## REFERENCES

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

Collett, D., 2002: *Modelling Binary Data*. 2nd ed. Chapman and Hall, 408 pp.

Ehrendorfer, M., 1997: Predicting the uncertainty of numerical weather forecasts: A review. *Meteor. Z.,* **6,** 147–183.

Epstein, E. S., 1985: *Statistical Inference and Prediction in Climatology: A Bayesian Approach. Meteor. Monogr.*, No. 42, Amer. Meteor. Soc., 199 pp.

Ferranti, L., E. Klinker, A. Hollingsworth, and B. J. Hoskins, 2002: Diagnosis of systematic forecast errors dependent on flow pattern. *Quart. J. Roy. Meteor. Soc.,* **128,** 1623–1640.

Glahn, H. R., 2004: Discussion of verification concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science. Wea. Forecasting,* **19,** 769–775.

Gneiting, T., A. E. Raftery, A. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.,* **133,** 1098–1118.

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

Hamill, T. M., C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

——, ——, and J. S. Whitaker, 2003: Ensemble forecasts and the properties of flow-dependent analysis-error covariance singular vectors. *Mon. Wea. Rev.,* **131,** 1741–1758.

Jolliffe, I. T., and D. B. Stephenson, 2005: Comments on ''Discussion of verification concepts in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*.'' *Wea. Forecasting,* **20,** 796–800.

Katz, R. W., and M. Ehrendorfer, 2006: Bayesian approach to decision making using ensemble weather forecasts. *Wea. Forecasting,* **21,** 220–231.

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

McCullagh, P., and J. Nelder, 1989: *Generalized Linear Models*. 2nd ed. Chapman and Hall, 532 pp.

Molteni, E., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles. *Mon. Wea. Rev.,* **130,** 1792–1811.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.,* **132,** 2732–2744.

Stephenson, D. B., 2008: An introduction to probability forecasting. *Seasonal Climate: Forecasting and Managing Risk*, A. Trocolli et al., Eds., Nato Science Series, Vol. 82, Springer Academic, 235–258.

——, and F. J. Doblas-Reyes, 2000: Statistical methods for interpreting Monte Carlo ensemble forecasts. *Tellus,* **52A,** 300–322.

Wang, X., and C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.,* **60,** 1140–1158.

Wilks, D. S., 2006a: Comparison of ensemble-MOS methods in the Lorenz '96 settings. *Meteor. Appl.,* **78,** 2851–2857.

——, 2006b: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysical Series, Vol. 91, Academic Press, 648 pp.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Wilson, L. J., and M. Valle, 2002: The Canadian Updateable Model Output Statistics (UMOS) system: Design and development tests. *Wea. Forecasting,* **17,** 206–222.

Winkler, R. L., 1968: ''Good'' probability assessors. *J. Appl. Meteor.,* **7,** 751–758.