School of Mathematical
and Physical Sciences

University of Reading

Department of Mathematics and Statistics

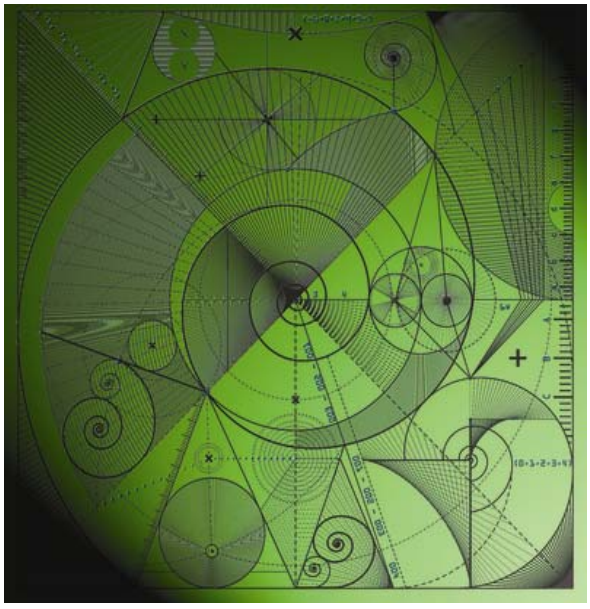
Preprint MPS_2010-34

29 November 2010

Early Warning with Calibrated and Sharper Probabilistic Forecasts

by

Reason L. Machete



Early Warning with Calibrated and Sharper Probabilistic Forecasts

Machete, R. L.†

†Department of Mathematics and Statistics, University of Reading, United Kingdom

r.l.machete@reading.ac.uk

Abstract

Given a nonlinear deterministic model, a density forecast is obtained by evolving forward an ensemble of starting values and doing density estimation with the final ensemble. The density forecasts will inevitably be downgraded by model misspecification. To mitigate model misspecification and enhance the quality of the predictive densities, one can mix them with the system's climatology. This paper examines the effect of including the climatology on the sharpness and calibration of density forecasts at various time horizons. The density forecasts are estimated using a non-parametric approach. The findings have positive implications for issuing early warnings in different disciplines including economic applications and weather forecasting, but a non-linear electronic circuit is used as a test bed.

Keywords: Calibration; Density forecasts; Nonlinear time series; Scoring rule; Uncertainty

1 Introduction

Brier [1950] was one of the first to highlight the importance of consistency between forecast probabilities and observed relative frequencies. This forecast attribute was later termed *validity* by Bross [1953] and *reliability* by Saunders [1958]. Currently it is commonly known as *calibration* (e.g. in Gneiting [2008], Lawrence *et al.* [2006]), although the weather community also uses the term reliability. While much of the discussion on calibration of probabilistic forecasts has centred on categorical events, Dawid [1984] is notable for proposing the use of *probability integral transforms* (PITs) to assess the calibration of density forecasts. A PIT is obtained by plugging an observation into the cumulative predictive distribution function. His proposed test included the additional condition that the PITs should be independent and identically distributed. Diebold *et al.* [1998] then showed that if density forecasts coincide with the ideal forecasts, then the PITs are independent and identically uniformly distributed (iid $U[0,1]$). Indeed iid $U[0,1]$ of PITs is a necessary and sufficient condition for the density forecasts to coincide with the ideal forecasts.

Unfortunately, ideal forecasts may be unattainable in practice (e. g. weather forecasting). It is, therefore, understandable that Gneiting *et al.* [2007] broke down calibration into three categories or modes: *probabilistic calibration, exceedance calibration* and

marginal calibration, which need not all hold at the same time. Density forecasts are said to be probabilistically calibrated if and only if the PITs are uniformly distributed. Marginal calibration refers to the case when the time average of all density forecasts is equal to that of ideal forecasts. There is no empirical way of assessing exceedance calibration.

Gneiting *et al.* [2007] then conjectured that when a subset of these modes of calibration holds, then the predictive distributions are at least as spread out as the ideal forecasts, which conjecture they termed a *sharpness principle*. The term *sharpness* seems to have been coined by Bross [1953], and it is a measure of how concentrated probabilistic forecasts are and a property of the forecasts only (Gneiting [2008], Gneiting *et al.* [2007], Wilks [2006]). Sharpness of predictive distributions has traditionally been measured by variance (e. g. Gneiting [2008], Gneiting *et al.* [2007]) and confidence intervals (e. g. in Gneiting *et al.* [2007]), even though Hirschman [1957] argued for the use of entropy. It has further been argued that the goal of probabilistic forecasting is to maximise sharpness subject to calibration (Gneiting [2008], Gneiting *et al.* [2007]). This so called *paradigm* (Gneiting [2008]) depends on the aforementioned conjecture, which we shall revisit later and present (with proof) a relevant theorem (or proposition).

When forecasting complex nonlinear systems like weather, model misspecification is inevitable due to simplifications and approximations involved. Even though the underlying system might be deterministic, a point forecast would be pointless. There may also be noise on the observations, increasing uncertainty in the forecasts. To account for model misspecification and observational uncertainty, a distribution of point forecasts is often issued at a given forecast lead time. To treat this ensemble of point forecasts as a probabilistic forecast would be naive. Each ensemble forecast can then be converted into a density forecast by assimilating some data to estimate kernel parameters as discussed in Broecker & Smith [2008], minimising a logarithmic scoring rule (Gneiting & Raftery [2007]). The scoring rule used is essentially the Kullback-Leibler divergence (Kullback & Leibler [1951]) and is discussed in § 2. Broecker & Smith [2008] pointed out that the density forecasts obtained as explained above can be improved by using what they called *affine functions* and mixing the densities with the unconditioned distribution of the data or *climatology* (Also called *time series density* in Hall & Mitchell [2007]). Their main aim was to circumvent the problem of large variance of parameters. The resulting density forecasts may be considered an example of mixture models common in Statistics.

As a dynamical system evolves in a bounded domain, it induces a probability density function according to relative frequencies with which it visits the different regions of state space. With no breaks nor drift in the dynamics, the arising probability distribution is called climatology. Thus the unconditioned distribution of time series from a stationary system provides an estimate of the system's climatology.

From the results of Broecker & Smith [2008], it is evident that mixing with climatology circumvents the problem of large variance of parameters. Including the climatology was also found by Hall & Mitchell [2007] to improve performance in terms of the logarithmic scoring rule. Except for mixing with climatology, Hall & Mitchell [2007] used a parametric density estimation approach. The approach here is non-parametric. It would be interesting to determine if improvement was due to an increase in sharpness at the expense of calibration. As a test bed, we use an electronic circuit to assess what happens

to these attributes when climatology is included. The modes of calibration assessed are probabilistic and marginal calibration, while it is emphasised that entropy should be used to measure sharpness. The new proposition that addresses Gneiting *et al.* [2007]'s conjecture helps explain the findings. Whereas Broecker & Smith [2008]'s criticism to not including affine functions in the density estimation was its guaranteed increase in variance of the predictive distribution compared to the raw ensemble, we demonstrate that the use of variance to measure sharpness could be misleading.

This paper is organised as follows: The next section discusses forecast qualities that are cumulatively measured by the logarithmic scoring rule. In particular, a decomposition of this scoring rule is presented. The sharpness principle conjectured by Gneiting *et al.* [2007] is discussed and a relevant proposition presented in § 3. In § 4, the methodology employed to produce density forecasts is outlined. Results concerning density forecasts obtained via the logarithmic scoring rule with respect to a nonlinear electronic circuit are discussed in § 5. Section 6 gives concluding remarks. Appendices A and B contain the proof of the proposition concerning the sharpness principle and appendix C contains proofs for the rest of the propositions.

2 Probabilistic-Forecast Quality

Model misspecification places limitations on the value of probabilistic forecasts. On the other hand, consumers of forecasts may demand predictive distributions that are both *calibrated* and *sharp*. If such forecasts are issued at long time horizons, then early warning is afforded. We suggest that these qualities can be cumulatively quantified by the logarithmic scoring rule proposed by Good [1952]. There are other scoring rules available for selection (see Gneiting & Raftery [2007]). For instance, there is the Brier score (Brier [1950]). This, however, decomposes into many terms (Murphy [1993]), some of which are not relevant to our discussion and it is suitable for categorical events. A generalisation of the Brier score to density forecasts is the *continuous rank probability score* Gneiting & Raftery [2007], but it lacks a clear interpretation. Indeed traditional decompositions of scoring rules do not contain a sharpness term. There is also the mean square error loss function (Corradi & Swanson [2006]), which is also irrelevant to the qualities of interest. Suffice it to say, the logarithmic scoring rule is preferred over others for its appeal to information theory concepts (see Roulston & Smith [2002]), which can be traced back to Shannon [1948, 1949]. Information theory has a strong hold on uncertainty, a concept equivalent to sharpness.

2.1 Logarithmic Scoring Rule

Consider a density forecast $f(x)$ and a target probability density function $g(x)$. If we think of X as a random variable, then the foregoing notation says that the true distribution of X is $g(x)$. With this notation, the information based scoring rule used in this paper is

$$\mathbb{E}[\mathrm{IGN}(f, X)] = -\int_{-\infty}^{\infty} g(x) \log f(x) \mathrm{d}x, \tag{1}$$

where $\mathrm{IGN}(f, X) = -\log f(X)$, proposed by Good [1952] and termed *Ignorance* in Roulston & Smith [2002] and *predictive deviance* in Knorr-Held & Rainer [2001]. Hence, (1) is the expected Ignorance. It is related to the Kullback-Leibler divergence (Kullback & Leibler [1951]),

$$D_{\mathrm{KL}}(g||f) = \int_{-\infty}^{\infty} g(x) \log\left(\frac{g(x)}{f(x)}\right) \mathrm{d}x$$

by

$$D_{\mathrm{KL}}(g||f) = \mathbb{E}[\mathrm{IGN}(f, X)] + \int_{-\infty}^{\infty} g(x) \log g(x) \mathrm{d}x.$$

It follows that the f that minimises $D_{\mathrm{KL}}(g||f)$ also minimises $\mathbb{E}[\mathrm{IGN}(f, X)]$. The expected Ignorance is the infamous cross entropy $H(g, f)$. The Ignorance score is especially relevant when one evaluates the performance of density forecasts given time series only, with no access to $g(x)$. An important property of the Ignorance score is that it attains the minimum if and only if $f(x) = g(x)$ (Broecker & Smith [2008], Gneiting & Raftery [2007]), meaning it is *strictly proper*.

Traditionally, the only score that has been decomposed into constituent terms is the Brier score: the infamous *reliability-resolution* decomposition (Murphy [1993], Wilks [2006]), after removing the uncertainty term. Broecker [2009] extended the decomposition to general scores, but in the context of categorical forecasts. Unlike sharpness, resolution is not a property of the forecasts only. Therefore, we introduce a decomposition of (1) as

$$\mathbb{E}[\mathrm{IGN}(f, X)] = -\int_{-\infty}^{\infty} f(x) \log f(x) \mathrm{d}x - \int_{-\infty}^{\infty} [g(x) - f(x)] \log f(x) \mathrm{d}x.$$

In this decomposition of expected Ignorance, the first term is *sharpness* and the second is *calibration*. Notice that the sharpness term is simply the density entropy $H(f)$, a property of the density forecast only. It is desirable for this term to be as negative as possible, effectively expressing more certainty about what is likely to happen. Since calibration is a statistical property of the forecasting system, it cannot be assessed based on one forecast only. For a time series of forecasts, we want each $f(x)$ to be close to $g(x)$ in some way. One is never furnished with $g(x)$ to aid assessment of calibration in an operational setup, but there are time series approaches to address this.

2.2 Sharpness

One way to quantify sharpness is to use the variance (e.g. Gneiting *et al.* [2007]). We emphasise that sharpness should be quantified by entropy, which "is a measure of concentration" of the distribution "on a set of small measure", a small value of entropy corresponding to a "high degree of concentration" (Hirschman [1957]). The entropy of a distribution $f(x)$ of variance σ^2 satisfies the inequality (Shannon [1948, 1949])

$$-\int_{-\infty}^{\infty} f(x) \log f(x) \mathrm{d}x \leq \frac{1}{2} \log\left(2\pi e \sigma^2\right).$$

Hence, a smaller variance guarantees lower entropy but not vice versa. Indeed two distributions with the same variances can have unequal entropies. For instance, a mixture

of two Gaussians will have lower entropy than a single Gaussian distribution of the same variance. Much more, a distribution of a higher variance can have a lower entropy than that of lower variance.

Sharpness has also been quantified by confidence intervals (Raftery *et al.* [2005], Gneiting *et al.* [2007]). Confidence intervals share a similar weakness to variance in the sense that a bimodal distribution that is fairly concentrated on the two modes can have larger confidence intervals than a unimodal distribution that is fairly spread out. Also, given two non-symmetric distributions, which of them is deemed sharper could depend on what the confidence level is.

2.3 Calibration

The calibration of density forecasts is a well trodden subject. Much of the literature takes the stand that a calibrated forecasting system is tantamount to a correctly specified model. Corradi & Swanson [2006] provide a comprehensive survey of formal statistical techniques for assessing calibration of density forecasts to determine if the underlying model is correctly specified. The work of Gneiting *et al.* [2007] strikes a discord by providing a calibration framework that accommodates model misspecification. They broke down calibration into three modes, each of which could be assessed separately.

Suppose a probability forecasting system issues predictive distributions $\{F_t(x)\}_{t=1}^T$, while the data-generating process issues ideal forecasts $\{G_t(x)\}_{t=1}^T$. Gneiting *et al.* [2007] then defined the following modes of calibration:

- The sequence $\{F_t(x)\}_{t=1}^T$ is *probabilistically calibrated* relative to $\{G_t(x)\}_{t=1}^T$ if

$$\frac{1}{T}\sum_{t=1}^{T} G_t\{F_t^{-1}(p)\} = p, \quad p \in (0,1). \tag{2}$$

- The sequence $\{F_t(x)\}_{t=1}^T$ is *exceedance calibrated* relative to $\{G_t(x)\}_{t=1}^T$ if

$$\frac{1}{T}\sum_{t=1}^{T} G_t^{-1}\{F_t(x)\} = x, \quad x \in \Re.$$

- The forecaster is *marginally calibrated* if

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} F_t(x) = \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} G_t(x).$$

If we have a time series of observations x_t, then $z_t = F_t(x_t)$ is a *probability integral transform* (PIT) (Corradi & Swanson [2006], Diebold *et al.* [1998]). Uniformity of the PITs is equivalent to probabilistic calibration (Gneiting *et al.* [2007]). A visual inspection of PIT histograms would reveal obvious departures from uniformity. The underlying model is correctly specified if and only if $z_t \sim$ iid $U[0,1]$.

Suppose we have a time series of density forecasts, $\{f_t(x)\}_{t\geq 1}$. Then define the forecaster's climatology as

$$\tilde{\rho}_T(x) = \frac{1}{T}\sum_{t=1}^{T} f_t(x).$$

We define a forecaster who issues

$$F_t(x) = \overline{G}_T(x) = \frac{1}{T}\sum_{t=1}^{T} G_t(x)$$

for all $t \in \{1\ldots,T\}$ to be the *finite climatological forecaster*. If $F_t(x) = \lim_{T\to\infty}\overline{G}_T(x)$, then we have the *climatological forecaster*. A forecaster is *finite marginally calibrated* if $\tilde{\rho}_T(x) = \overline{G}'_T(x)$.

For all practical purposes, T is finite and we have no access to the $G_t(x)$'s. Hence it is difficult to assess finite marginal calibration. If $d(\tilde{\rho}_T, \tilde{\rho}_{2T}) \approx 0$, where d is some metric, then we can take T to be large enough to evaluate marginal calibration. To this end, we can use the Hellinger distance (Pollard [2002]) and compute

$$h(\tilde{\rho}_T, \rho_c) = \frac{1}{2}\int \left[\sqrt{\tilde{\rho}_T(x)} - \sqrt{\rho_c(x)}\right]^2 \mathrm{d}x,$$

where $\rho_c(x) = \lim_{T\to\infty}\overline{G}'_T(x)$ is the system's climatology. It is useful to note that $0 \leq h(\cdot,\cdot) \leq 1$, assuming the value of 0 when the two distributions are identical and 1 when they do not overlap. This procedure for assessing marginal calibration is an alternative to the graphical tests performed in Gneiting *et al.* [2007]. It is expected to be more robust to finite sample effects.

3 The Sharpness Principle and Early Warning

Murphy & Wilks [1998] highlighted that forecasts need to be calibrated before one worries about sharpness. Much earlier, Bross [1953] argued that a forecaster with a sharper, but less calibrated probability forecasting system (PFS) could make a lot of money over one with a more calibrated but less sharp PFS. Recently, Gneiting *et al.* [2007] adopted a paradigm of maximising sharpness subject to calibration. They then conjectured that the goal to obtain ideal forecasts and of maximising sharpness subject to calibration are equivalent, which is the *sharpness principle*. A weaker alternative states that any sufficiently calibrated forecaster is at least as spread out as the ideal forecaster (Gneiting *et al.* [2007]). It has been demonstrated by counter examples that none of the individual modes of calibration alone is sufficient for the weaker conjecture to hold (Gneiting *et al.* [2007]). Since Pal [2009] admittedly did not satisfactorily address this conjecture, it is revisited.

It is noteworthy that Gneiting *et al.* [2007] could not find a counter example to disprove that a forecaster who is both probabilistically and marginally calibrated is at least as spread out as the ideal forecaster. The following proposition addresses this in the case of finite marginal calibration, which means:

$$\frac{1}{T}\sum_{t=1}^{T} F_t(x) = \frac{1}{T}\sum_{t=1}^{T} G_t(x). \tag{3}$$

PROPOSITION 1. *Suppose* $\{G_t\}_{t=1}^T$ *is a sequence of continuous and strictly increasing distribution functions (ideal forecasts). Then a forecaster who is both probabilistically and finite marginally calibrated has either issued ideal forecasts* $\{G_t\}_{t=1}^T$ *or is the finite climatological forecaster.*

Including exceedance calibration in the hypotheses of the above proposition would rule out the finite climatological forecaster. The proof for the above proposition is split into two parts and is given in appendix A and B. Even though this proposition does not deal with the case when T approaches infinity, it is operationally useful. Indeed the graphical tests for marginal calibration discussed in Gneiting *et al.* [2007] deal with *finite* marginal calibration. The implications (of the proposition) to the sharpness principle are that the level of expectation with regard to the two modes of calibration needs to be scaled down when the underlying model is misspecified. Without a correctly specified model, one cannot have both perfect probabilistic and finite marginal calibration unless he is the finite climatological forecaster. While marginal calibration might be easier to satisfy, probabilistic calibration may be harder to achieve. From this proposition, we note that when the underlying model is misspecified there will be no bound on the sharpness of predictive distributions. The forecaster should merely aim to maximise sharpness subject to some level of calibration. A forecaster who is both probabilistically and finite marginally calibrated affords early warning if he is sharper than climatology at long time horizons.

4 Density-Forecast Estimation

Suppose we have some data point s_t, at time t, and we want to know the future state at time $t+\tau$. We call τ the *forecast lead time*. If we acknowledge both model misspecification and noise in the data, then it makes no sense to issue a point forecast. To express uncertainty in the forecasts, we issue a density forecast. The first step to obtaining a density forecast is to generate many points in the neighbourhood of s_t and iterate each of the points forward with the model to obtain an ensemble of forecasts $\boldsymbol{X}^{(t+\tau)} = \left\{X_i^{(t+\tau)}\right\}_{i=1}^N$. If the model is stochastic, it may suffice to iterate it forward several times to generate the ensemble. This section is concerned with converting the ensemble into a density forecast. In all our computations, we used the Gaussian kernel function,

$$K(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\xi^2/2\right).$$

4.1 Single Model

One way to convert a forecast ensemble into a density forecast would be to perform density estimation according to Parzen [1962] and Silverman [1986]. The fundamental weakness of this approach is that it inherently assumes that the ensemble is a draw from the true distribution. In view of this, Roulston & Smith [2002] suggested taking into account how the model has performed in the past. A similar approach is followed by Hall & Mitchell [2007], who use past forecast errors to obtain density forecasts. Therefore,

we can form density forecast estimates of the form:

$$\rho^{(t)}(x) = \frac{1}{\sigma N} \sum_{i=1}^{N} K\left\{ \left(x - X_i^{(t)} - \mu \right) / \sigma \right\}, \tag{4}$$

where σ and μ are respective bandwidth and offset parameters chosen according to past performance and $K(\cdot)$ is the kernel function. The density forecast in (4) differs from the traditional Parzen [1962] estimates by the offset parameter. It is similar to the Bayesian Model Average proposed by Raftery *et al.* [2005] with a uniform bias correction, μ and equal weights. Here, the ensemble members are exchangeable and do not represent distinct models. Selecting σ using Silverman [1986] does not account for model misspecification.

To account for model misspecification, let us first denote a record of past time series and corresponding ensemble forecasts by $\mathcal{V}_T = \{(s_t, \boldsymbol{X}^{(t)})\}_{t=1}^{T}$. Then the density forecasts whose parameters, μ and σ, are selected by taking into account past performance may be denoted by $\rho^{(t)}(x|\mathcal{V}_T)$. While $\rho^{(t)}(x|\mathcal{V}_T)$ has the same form as in (4), its parameters are selected by doing the minimisation

$$\min_{\sigma>0,\mu} \left\{ -\frac{1}{T} \sum_{t=1}^{T} \log \rho^{(t)}(s_t|\mathcal{V}_T) \right\}. \tag{5}$$

Under certain assumptions, doing the minimisation in (5) is tantamount to minimising either the average cross entropy or the average Kullback-Leibler divergence. Without making any assumptions, the term in (5) should be called average Ignorance, $\langle \text{IGN} \rangle$. Minimising (5) is equivalent to maximum likelihood under the assumption of independence of forecast errors (Raftery *et al.* [2005]). Moreover, it is equivalent to *quasi maximum likelihood* (QML) under model misspecification with independent conditional forecasts as discussed by White [1994]. Interestingly, White [1982] called the QML estimator the 'minimum ignorance' estimator, arguing that it minimises our ignorance about the correct model structure.

4.2 Mixture Model

Broecker & Smith [2008] noted that, when doing the minimisation in (5), some of the $\boldsymbol{X}^{(t)}$ may be far from the corresponding s_t, which could result in choices of σ that were too big. Hence, the parameter estimates would not be robust. These short comings could largely be due to model misspecification. To circumvent these, they proposed a mixture model of the climatology, $\rho_c(x)$, and $\rho^{(t)}(x|\mathcal{V}_T)$:

$$f^{(t)}(x|\mathcal{V}_T) = \alpha \rho^{(t)}(x|\mathcal{V}_T) + (1-\alpha)\rho_c(x), \tag{6}$$

where the mixture parameter, $\alpha \in [0, 1]$. All the three parameters are fitted simultaneously by minimising average Ignorance. The system's climatology, $\rho_c(x)$, is estimated from data via

$$\rho_c(x) = \frac{1}{\sigma_c T} \sum_{t=1}^{T} K\left\{ (x - s_t - \mu_c) / \sigma_c \right\},$$

and the parameters σ_c and μ_c are then selected as proposed in Broecker & Smith [2008].

If we let $r_t = \rho^{(t)}(s_t)/\rho_c(s_t)$, then we can state the following proposition,

PROPOSITION 2. *For a given set of parameters μ and σ, the necessary and sufficient conditions for improvement from including the climatology in the sense of the logarithmic scoring rule are that*

$$\frac{1}{T}\sum_{t=1}^{T} r_t > 1 \quad \textit{and} \quad \frac{1}{T}\sum_{t=1}^{T} \frac{1}{r_t} > 1.$$

The proof for this proposition is given in appendix C. The ratio r_t may be interpreted as the *return ratio* on some invested capital in a Kelly betting scenario (Kelly [1956]) with no track take. The proposition states how the model and climatology are to outperform each other in order for the mixture model to provide additional value.

In order to capture the effect of including the climatology on the kernel width, we consider the case when $N = 1$ with $\mu = 0$. When there is no climatology included, minimising the logarithmic score yields,

$$\sigma_o^2 = \frac{1}{T}\sum_{t=1}^{T} \left\{ s_t - X_1^{(t)} \right\}^2 . \tag{7}$$

Let us write a time series version of the logarithmic scoring rule as

$$\langle \mathrm{IGN} \rangle = -\frac{1}{T}\sum_{t=1}^{T} \log f^{(t)}(s_t|\mathcal{V}_T). \tag{8}$$

PROPOSITION 3. *Suppose the score given by equation (8) assumes a minimum at parameter values (σ_*, α_*), then the following equation holds:*

$$\sigma_*^2 = \frac{1}{T}\sum_{t=1}^{T} \left\{ s_t - X_1^{(t)} \right\}^2 \frac{\rho^{(t)}(s_t|\mathcal{V}_T)}{f^{(t)}(s_t|\mathcal{V}_T)}. \tag{9}$$

See appendix C for the proof. For illustrative purposes, suppose that the kth forecast is far from the corresponding observation in the sense that

$$\left| s_k - X_1^{(k)} \right| \gg \max \left\{ \left| s_t - X_1^{(t)} \right| \right\}_{t \neq k} .$$

As a result, the kernel width in (7) would be inflated. Equation (9) provides a way to discount the contributions of a few bad forecasts on the kernel width. In this case, (σ_*, α_*) would be chosen such that

$$\frac{\rho^{(k)}(s_k|\mathcal{V}_T)}{f^{(k)}(s_k|\mathcal{V}_T)} \ll 1.$$

This is especially valuable when T is small, which is the case in typical time series. The idea is that a reduction in kernel width is necessary for the entropy of $f^{(t)}(x|\mathcal{V}_T)$ to decrease even when $N > 1$, but it is easier to explain how the reduction is achieved

when $N = 1$. Despite this reduction, some mixture forecasts may still be less sharp than climatology in the sense of entropy. A straight forward application of the Kullback-Leibler and Jensen's inequalities leads to the relations

$$\alpha H\left\{\rho^{(t)}\right\} + (1-\alpha)H(\rho_c) \leq H\left\{f^{(t)}\right\} \leq \alpha^2 H\left\{\rho^{(t)}\right\} + \alpha(1-\alpha)H\left\{\rho^{(t)}, \rho_c\right\} + \ldots \\ (1-\alpha)\alpha H\left\{\rho_c, \rho^{(t)}\right\} + (1-\alpha)^2 H(\rho_c),$$

where $H(f) = -\int f(x)\log f(x)\mathrm{d}x$ and $H(f,g) = -\int f(x)\log g(x)\mathrm{d}x$ are the entropy and cross entropy respectively. Therefore, the necessary and sufficient conditions for $H\left\{f^{(t)}\right\} \geq H(\rho_c)$ to hold are that $H(\rho_c) < H\left\{\rho^{(t)}, \rho_c\right\}$ and $H(\rho_c) < H\left\{\rho^{(t)}\right\}$ respectively. Whenever the climatology is sharper than the mixture forecast, it should be issued as the forecast instead of the mixture.

It is not obvious what the effect of the mixture is on calibration, except that including climatology improves the KL distance from the ideal forecasts. Nevertheless, the mixture parameter that minimises the logarithmic score yields the equation

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\rho_c(s_t)}{f^{(t)}(s_t|\mathcal{V}_T)} = 1.$$

On the other hand, we note that equation (2) is equivalent to

$$\frac{1}{T}\sum_{t=1}^{T}\frac{g_t(s_t)}{f_t(s_t)} = 1 \tag{10}$$

The two preceding equations are similar with the ρ_c replacing g_t in (10). What happens to calibration due the mixture will be explored by way of example in the next section.

5 Results and Discussion

This section presents the results that highlight the effects, on sharpness, calibration and the time horizon over which density forecasts are useful, of introducing the climatology to form the density forecasts. The system considered is a non-linear, chaotic, electronic circuit constructed in a Physics laboratory at the University of Oxford. The signal recorded consisted of voltages at some points on the circuit. The circuit was forecast using a data based, deterministic, non-linear model. A portion of 2^{10} data points was used to select the density forecast parameters as discussed in § 4. An out of sample evaluation of density forecasts was then performed.

The first quality considered was sharpness. A sample of density forecasts from two ensemble forecasts at a forecast lead time of 6.4 ms is shown in figure 1. On the left are density forecasts resulting from estimation without the climatology and those on the right were obtained by mixing with the climatology. It is evident, by visual inspection, that including climatology resulted in predictive distributions that were sharper (narrower). Note that all the predictive distributions shown in figure 1 are clearly sharper than the climatology, which is shown in figure 2.

Sharpness was assessed further by computing the corresponding variances and density entropies (see figures 3). The graph on the left shows a scatter plot of the variances

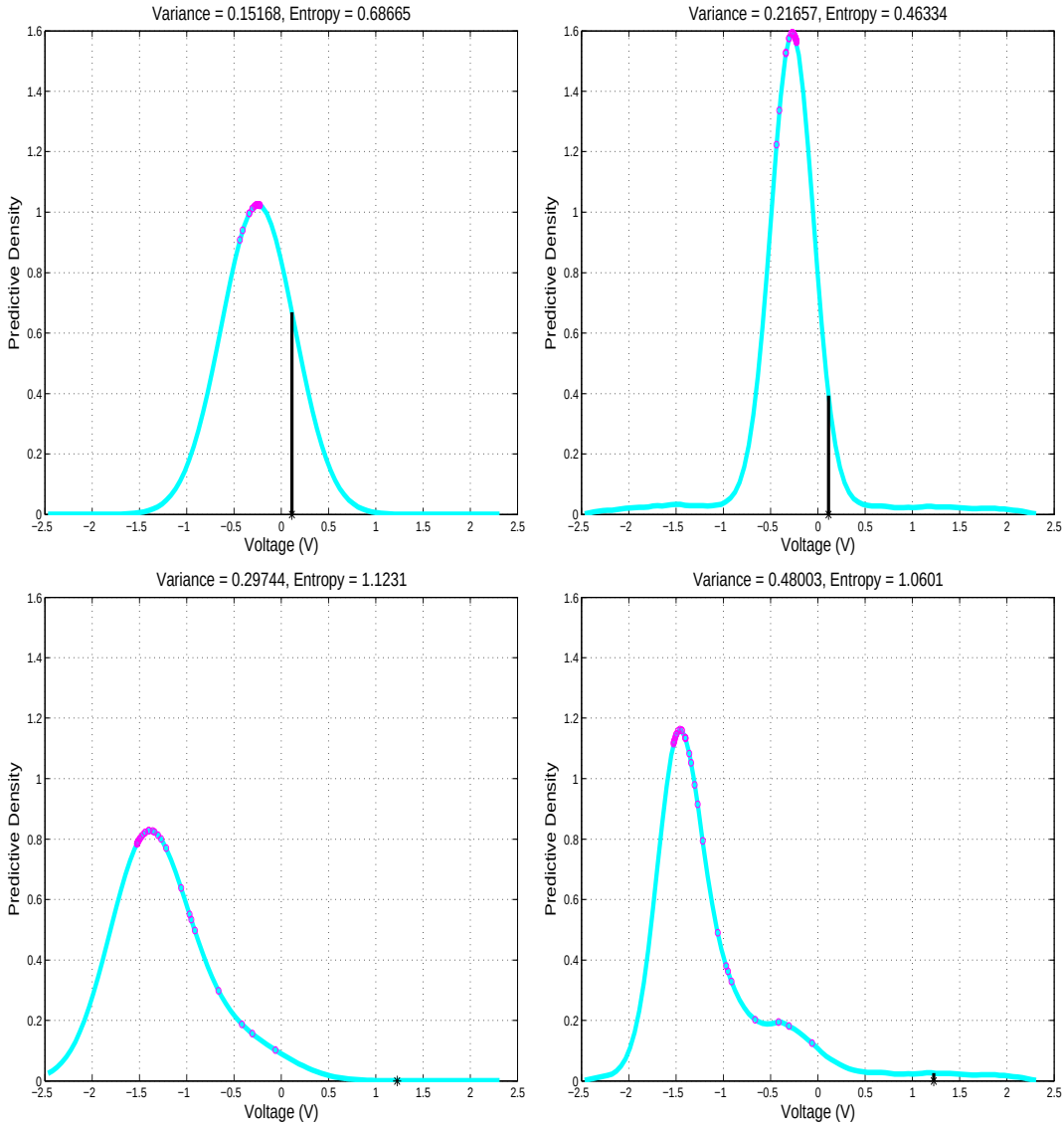


Figure 1: *Graphs of density forecasts for the circuit at a forecast time of 6.4 ms. Figures on the same row correspond to the same ensemble. The density forecasts on the left were obtained by estimation without the climatology and those on the right were obtained by including the climatology. Notice that including the climatology resulted in narrower distributions. Clearly, the lower entropies for the graphs on the right are a reflection of the noticeable increase in sharpness.*

of the predictive distributions containing the climatology against those without it. Only 6% of the predictive distributions mixed with the climatology resulted in variance reduction. Based on this graph, one could conclude that including the climatology resulted in predictive distributions that were more spread out. We contend that it is better to use entropy to measure sharpness. On the scatter plot of density entropies on the right hand side of figure 3, 77% of the points lie below the line $y = x$, implying that including the climatology generally yielded sharper predictive distributions. This example illustrates that one's conclusions can vary depending on whether they use entropy or variance to

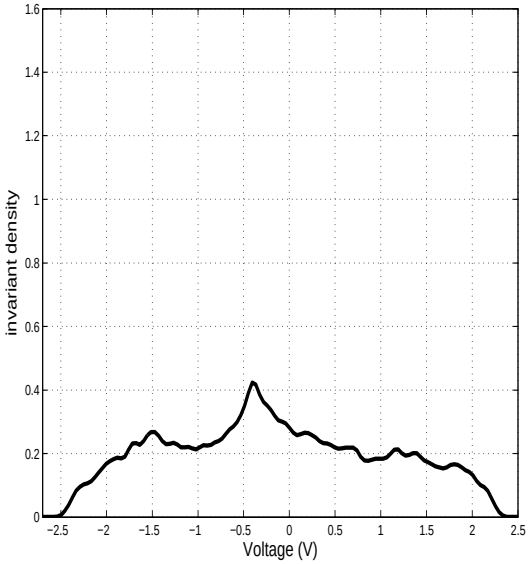


Figure 2: *Graph of the climatology of the circuit estimated from data. Its entropy is 2.15, which is greater than the entropies shown in figure 1.*

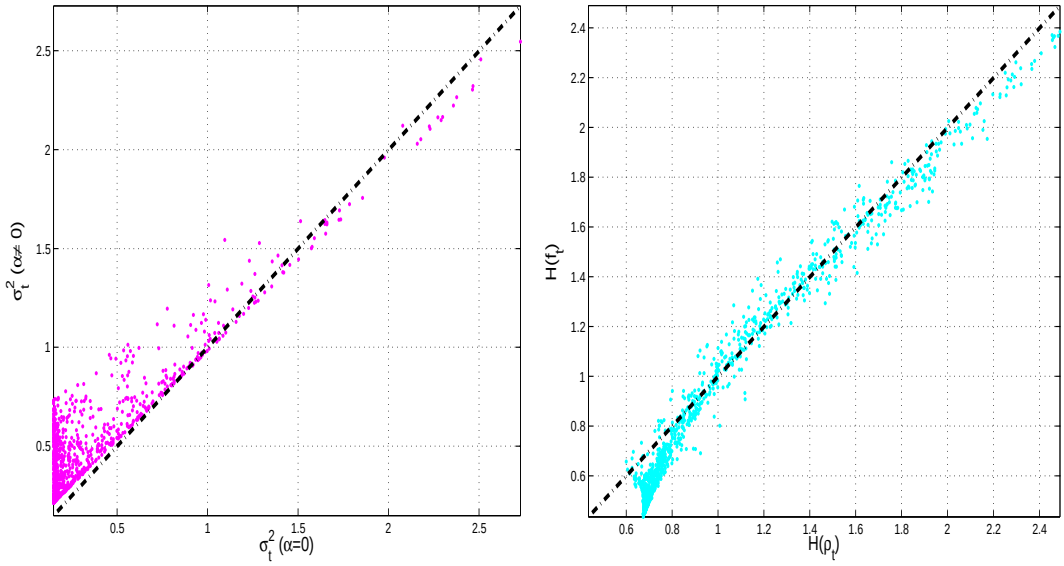


Figure 3: *Scatter plots of the variances (left) and entropies (right) of density forecasts containing climatology versus those without. On the left picture, only 6% of the points are below the line $y = x$, indicating that including the climatology resulted in bigger variances. On the contrary, the picture on the right indicates that 77% of the points are above the line $y = x$, hence including the climatology resulted in lower entropy.*

measure sharpness. At forecast time of 12.8 ms, 61% of predictive distributions containing the climatology had smaller variance whilst 71% of them had lower entropy. In that instance, both measures concurred that including the climatology tended to sharpen the predictive distributions.

We would also like to draw attention to the fact that reduction in the kernel band-

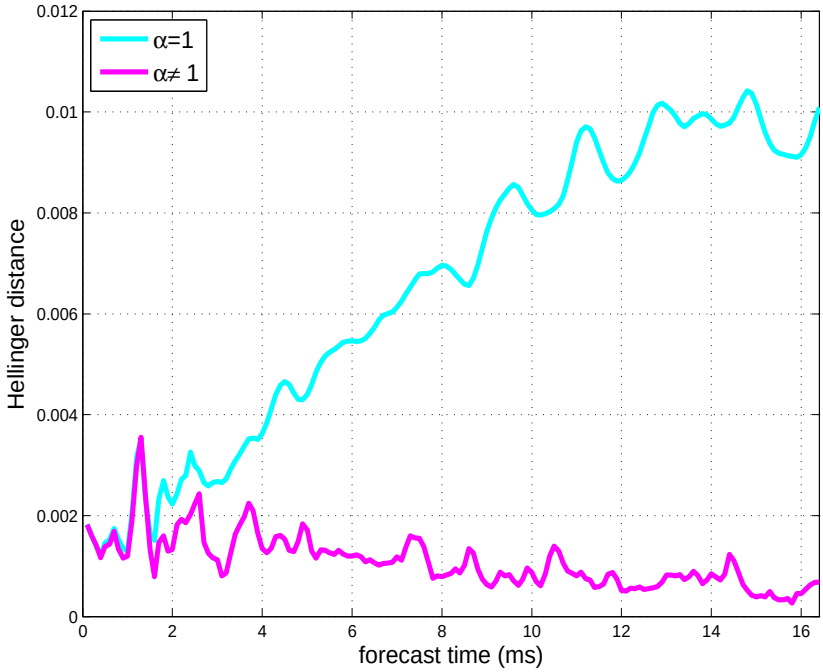


Figure 4: Graphs of the Hellinger distance of forecast climatologies from the system's climatology versus lead time.

width does not necessarily lead to reduction in variance. At a forecast lead time of 6.4 ms, we noted that including the climatology generally resulted in variance increase. However, the bandwidth was found to be 0.386 when the climatology was not included and 0.220 when it was. The graphs of bandwidth versus forecast time are shown in figure 5. The mixture distribution does indeed turn out to have smaller bandwidth, in agreement with analytic considerations of the previous section.

To assess marginal calibration, we compare forecast climatology with the system's climatology. At the forecast lead time of 6.4 ms, the average of density forecasts containing no climatology differed with the climatology by a Hellinger distance of 0.0056 while including the climatology resulted in a Hellinger distance of 0.0013. In both cases we can conclude that the density forecasts are marginally calibrated. Obviously, including the climatology should tend to improve marginal calibration. Graphs of the Hellinger distance of forecast climatologies from the system's climatology versus forecast lead time shown in figure 4 support this claim.

A sample of PITs is shown in figure 6. A visual inspection of the PITs indicates that including the climatology did not degrade probabilistic calibration. At a lead time of 5.6 ms, PITs for both versions of density forecasts appear uniformly distributed, a signature of probabilistic calibration. Actually, climatology appeared to improve probabilistic calibration at some higher lead times. For lead times up to about 5.6 ms, there was no noticeable difference between the PITs, yet density forecasts containing climatology scored better as shown in figure 7. Improvement in the score must be due to improvement in sharpness without compromise on probabilistic and marginal calibration.

Given that a PFS is sufficiently calibrated, a given density forecast is of value if it is

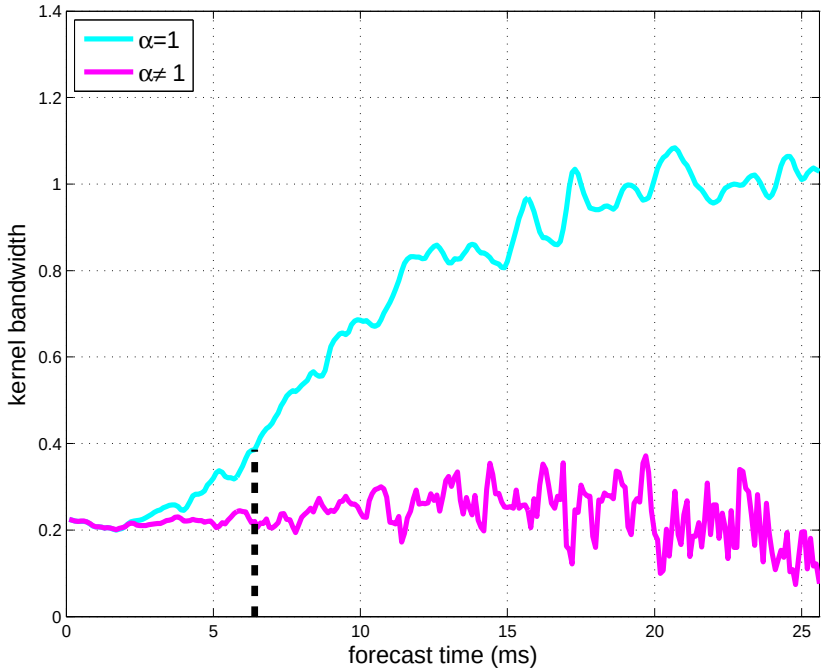


Figure 5: *Graphs of kernel bandwidth versus forecast lead time for density forecasts. The dotted vertical line corresponds to the forecast time of 6.4 ms. Evidently, including the climatology resulted in smaller kernel bandwidth.*

sharper than climatology. Entropy distributions for density forecasts at lead times of 6.4 ms, 9.6 ms and 12.8 ms are shown in figure 8. It is evident from the graphs that most of the predictive distributions are sharper than climatology at each of the lead times. At the forecast lead times of 12.8 ms, 89.8% of predictive distributions are sharper than climatology. Similarly, 95% and 98% of predictive distributions at lead times of 9.6 ms and 6.4 ms are respectively sharper than climatology. On this evidence, predictability is retained at least up to lead times of 12.8 ms. Whenever a predictive distribution is less sharp than climatology, climatology should be issued in its stead. That our forecasts are calibrated and yet generally sharper than climatology implies that early warning is afforded. The fact that climatology is sharper than some predictive distributions at each of the lead times is testimony to model misspecification. Bearing in mind the main proposition (Proposition 1) of this paper, it is not unexpected when we do not have perfect probabilistic and marginal calibration.

6 Conclusions

This paper discussed a way toward achieving the goal of probabilistic forecasting, which is to maximise sharpness subject to calibration. To this end, it considered the effect of including the climatology on sharpness and calibration when forming density forecasts from a discrete ensemble of model runs. The conjecture of Gneiting *et al.* [2007] was also revisited and a proposition concerning the sharpness principle proven. It turned out that one cannot have both probabilistic and marginal calibration hold when the

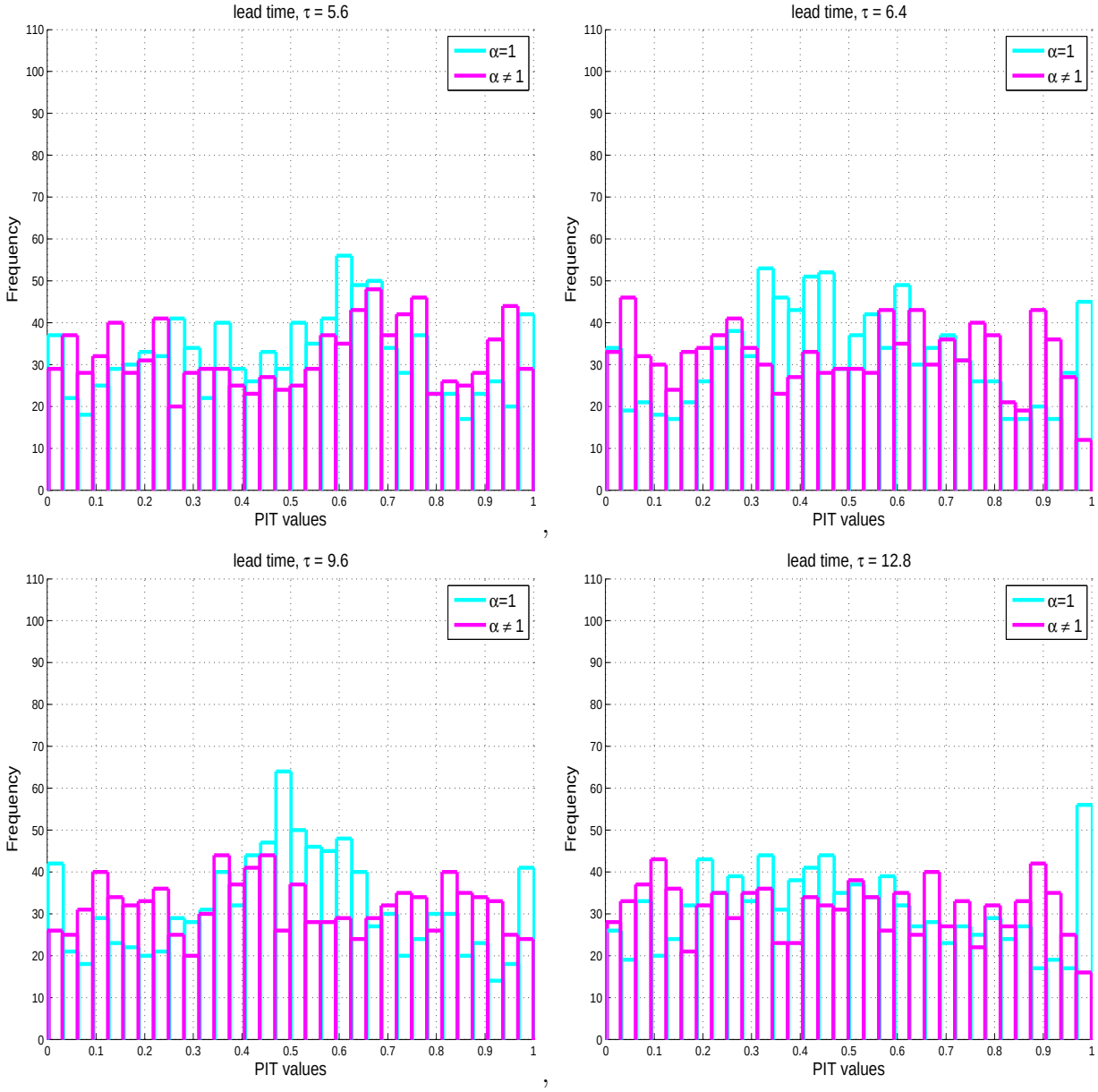


Figure 6: *Top: Graphs of probability integral transforms for predictive distributions at various lead times. A visual inspection suggests that including climatology via the logarithmic scoring rule tends to improve probabilistic calibration.*

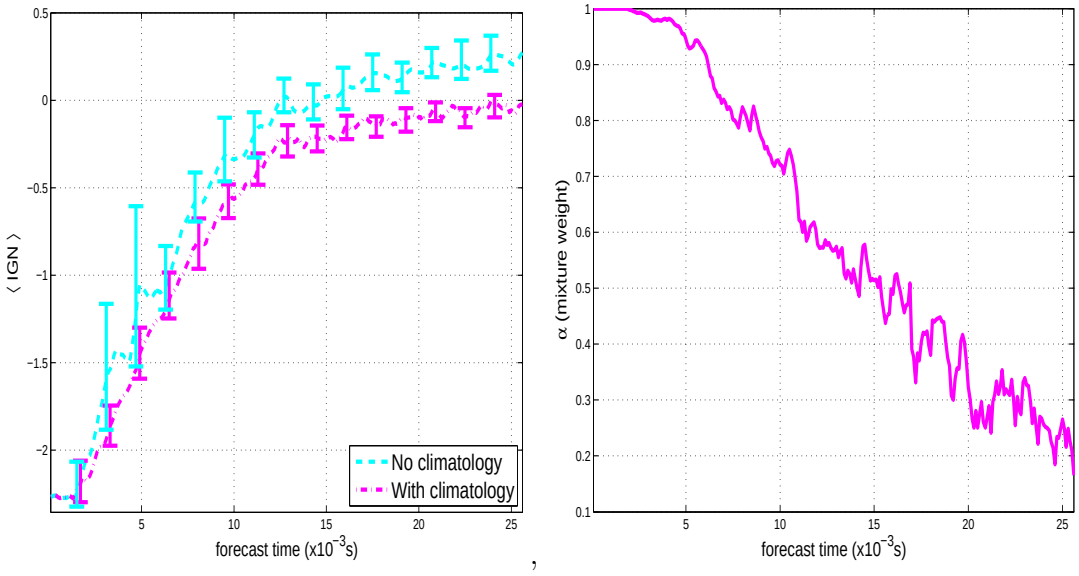


Figure 7: *Left: Graphs of out-of-sample average Ignorance, with the associated 95% confidence intervals, versus forecast time for density forecasts with and without the climatology. The average Ignorance is given relative to the entropy of the climatology. Right: Graph of the mixture weight versus forecast time. Notice from the left graph that if we do not include the climatology, we lose predictability after about 13 ms. On the other hand, including the climatology affords better performance up to 20 ms.*

underlying model is misspecified unless they settle for the climatological forecaster. In light of this fact, it has been suggested that one should scale down their calibration expectations when facing model misspecification.

It was found that including the climatology via the logarithmic scoring rule tended to improve marginal and probabilistic calibration. This was accompanied by a corresponding increase in sharpness as measured by entropy. Improvement in marginal calibration increased with lead time with no compromise to probabilistic calibration. It has also been argued that sharpness is better captured by entropy as opposed to variance. Fairly calibrated predictive distributions at higher lead times were found to be generally sharper than climatology, thus affording early warning. Crucially, though, some of the density forecasts may have larger entropy than the climatology. Such forecasts have to be rejected in favour of the climatology, which is sharper. Even though these observations were made on a nonlinear system, they may be useful in linear time series analysis as well, and/or when the model is stochastic. An open problem is to determine analytically why including the climatology via the logarithmic scoring rule tends to maintain or improve probabilistic calibration.

Acknowledgements

The author would like to acknowledge useful discussions with members of the CATS group at LSE and the Applied Dynamical Systems and Inverse Problems group at Oxford, and to thank David Allwright for his great insights especially pointing out Proposition 4. This work was supported by the RCUK Digital Economy Programme.

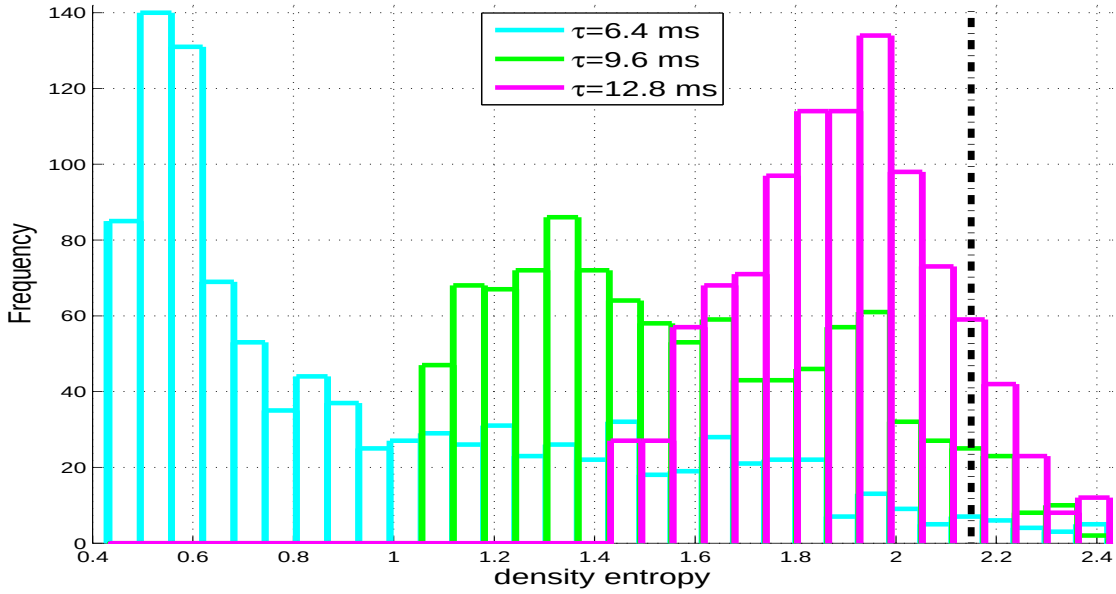


Figure 8: *Entropy distributions for density forecasts containing climatology at three different forecast lead times. The dash-dotted line corresponds to the entropy of the climatology. Density forecasts whose entropies exceed that of climatology should be rejected in its favour.*

References

Brier GW, 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review* **78**:1–3.

Broecker J, 2009. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society* **135**:1512–1519.

Broecker J, Smith LA, 2008. From Ensemble Forecasting to Predictive Distribution Functions. *Tellus A* **60**:663.

Bross IDJ, 1953. *Design for Decision: an introduction to statistical decision-making.* New York: Macmillan.

Corradi V, Swanson NR, 2006. *Predictive density evaluation, in Handbook of Economic Forecasting*, volume 1. North-Holland.

Dawid AP, 1984. Present position and potential developments: Some Personal Views: Statistical Theory: The Prequencial Approach. *J. R. Statist. Soc. A* **147**:278–292.

Diebold FX, Gunther TA, Tay AS, 1998. Evaluating density forecasts with application to financial risk management. *International Economics Review* **39**:863–883.

Gneiting T, 2008. Probabilistic forecasting. *J. R. Statist. Soc A* **171**:319–321.

Gneiting T, Balabdaoui F, Raftery AE, 2007. Probabilistic forecasts, calibration and sharpness. *J. R. Statist. Soc. B* **69**:243–268.

Gneiting T, Raftery AE, 2007. Strictly proper scoring rules, prediction and estimation. *J. Amer. Math. Soc.* **102**:359–378.

Good IJ, 1952. Rational decisions. *Journal of the Royal Statistical Society. Series B (Methodological)* **14**:107–114.

Hall SG, Mitchell J, 2007. Combining density forecasts. *International Journal of Forecasting* **23**:1–13.

Hirschman II, 1957. A note on entropy. *American Journal of Mathematics* **79**:152–156.

Kelly, 1956. A new interpretation of information rate. *Bell Systems Technical Journal* **35**:916–926.

Knorr-Held L, Rainer E, 2001. Projections of lung cancer in West Germany: A case study in bayesian prediction. *Biostatistics* **2**:109–129.

Kullback S, Leibler RA, 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics* **22**:79–86.

Lawrence M, Goodwin P, Marcus O, Onkal D, 2006. Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting* **22**:493–518.

Murphy AH, 1993. What is a good forecast? An essay on the nature and goodness in weather forecasting. *Weather and Forecasting* **8**:281–293.

Murphy AH, Wilks DS, 1998. A case study of the use of statistical models in the forecast verification: Precipitation probability forecasts. *Weather and Forecasting* **13**:795–810.

Pal S, 2009. A note on a conjectured sharpness principle for probabilistic forecasting with calibration. *Biometrika* **96**:1019–1023.

Parzen E, 1962. On the Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics* **33**:1065–1076.

Pollard D, 2002. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press.

Raftery AE, Gneiting T, Balabdaoui F, Polakowski M, 2005. Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review* **133**:1155–1174.

Roulston MS, Smith LA, 2002. Evaluating Probabilistic Forecasts Using Information Theory. *Monthly Weather Review* **130**:1653–1660.

Saunders F, 1958. The evaluation of subjective probability forecasts. Cambridge, Massachusetts Institute of Technology, Department of Meteorology, Contact AF 19(604)-1305, Sci. Rept. 5.

Shannon CE, 1948. A Mathematical theory of communication. *Bell Systems Technology Journal* **27**:379–423,623–656.

Shannon CE, editor, 1949. *Communication in the presence of noise*, volume 37. Pro. Institute of Radio Engineers.

Silverman BW, 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, first edition.

White H, 1982. Maximum likelihood estimation of misspecified models. *Econometrica* **50**:1–25.

White H, 1994. *Estimation, Inference and Specification Analysis*. New York: Cambridge University Press.

Wilks DS, 2006. *Statistical Methods in the Atmospheric Sciences*. Academic Press, 2nd ed edition.

A Generalised Construction of Probabilistically Calibrated Forecasts

PROPOSITION 4. *Suppose that G_t is a continuous strictly increasing distribution function on an interval I_t. Let I be any interval and choose for each t a strictly increasing continuous map $h_t : I \to I_t$. A probabilistically calibrated forecast distribution function precisely takes the form*[1]

$$F_s(x_s) = \frac{1}{T}\sum_{t=1}^{T} G_t\left[h_t\left\{h_s^{-1}(x_s)\right\}\right]. \tag{11}$$

Equation (11) is just Gneiting *et al.* [2007]'s construction in § 2.4 except that T is general rather than 2 and the linear maps x and x/a that Gneiting *et al.* [2007] use are replaced by the nonlinear maps h_t. Note that each F_s is a strictly increasing continuous distribution on I_s and they are probabilistically calibrated forecasts of the G_t's, because given $0 < p < 1$ there is some x in I with

$$p = \frac{1}{T}\sum_{t=1}^{T} G_t\left\{h_t(x)\right\} = F_s\left\{h_s(x)\right\},$$

whence $F_t^{-1}(p) = h_t(x)$ and

$$\frac{1}{T}\sum_{t=1}^{T} G_t\left\{F_t^{-1}(p)\right\} = \frac{1}{T}\sum_{t=1}^{T} G_t\left\{h_t(x)\right\} = p.$$

Moreover, any probabilistically calibrated forecast of G_t takes exactly this form. To see this, let I be any interval and h_1 be any suitable map from I onto I_1 and then define

[1]This proposition and its proof were kindly worked out and privately communicated to me by David Allwright who is at the Oxford Centre for Industrial and Applied Mathematics.

$h_t(x) = F_t^{-1}[F_1\{h_1(x)\}]$. It then follows that

$$\frac{1}{T}\sum_{t=1}^{T} G_t\left[h_t\left\{h_s^{-1}(x_s)\right\}\right] = \frac{1}{T}\sum_{t=1}^{T} G_t\left[F_t^{-1}\left\{F_s(x_s)\right\}\right] = F_s(x_s).$$

The first equality follows by definition of the h_t functions and the next by the probabilistic calibration property. Hence the F_t's have exactly the form of the construction.

B Proof of Proposition 1

Using proposition 4, probabilistic calibration implies that F_t takes precisely the form given in (11). If the sequence $\{F_t\}$ is also finite marginally calibrated, we can substitute (11) into (3) to obtain

$$\frac{1}{T(T-1)}\sum_{t=1}^{T}\sum_{s=1,s\neq t}^{T} G_t\left[h_t\left\{h_s^{-1}(x)\right\}\right] = \frac{1}{T}\sum_{t=1}^{T} G_t(x), \quad T \geq 2.$$

It is, therefore, required that $G_t\left[h_t\left\{h_s^{-1}(x)\right\}\right] = G_i(x)$, where $i \in \{1, .., T\}$. If $G_t[h_t\{h_s^{-1}(x)\}] = G_s(x)$ for any s, then the forecasts $\{F_t\}$ are ideal. On the other hand, if $G_t[h_t\{h_s^{-1}(x)\}] = G_t(x)$, then we have the finite climatological forecaster.

We now wish to show that a non-climatological forecaster who is both probabilistically and marginally calibrated is precisely the ideal forecaster. Consider $F_s(x)$ as defined by equation (11) for a given s. Suppose there exists q such that

$$G_t[h_t\{h_s^{-1}(x)\}] = G_s(x), \quad \text{for all} \quad t \leq q \tag{12}$$

and

$$G_t[h_t\{h_s^{-1}(x)] = G_t(x) \quad \text{for all} \quad t > q. \tag{13}$$

Equation (12) implies that $G_s[h_s\{h_t^{-1}(x)\}] = G_t(x)$ for all $t \leq q$ while (13) implies that $h_t(x) = h_s(x)$ for all $t > q$. $F_s(x)$ contains q counts of $G_s(x)$. Each

$$F_i(x) = \frac{1}{T}\sum_{t=1}^{T} G_t\left[h_t\left\{h_i^{-1}(x)\right\}\right],$$

$i \neq s$, contains 0 counts of $G_s(x)$ if $i \leq q$. If $i > q$, we get

$$G_t[h_t\{h_i^{-1}(x)\}] = G_t[h_t\{h_s^{-1}(x)\}] = G_s(x),$$

for all $t \leq q$. The first equality follows from noting that $h_i(x) = h_s(x)$ and the second from applying (12). Hence each $F_i(x)$ contains q counts of $G_s(x)$. Therefore, all the summations on the right hand side of the forecasters contain $q + (T-q)q$ counts of $G_s(x)$. Finite marginal calibration imposes the requirement that $q + (T-q)q = T$, which holds if and only if $q = T$. But $q = T$ implies that we have ideal forecasts.

More generally, the sequence $\{G_t[h_t\{h_s^{-1}(x)\}]\}_{t>q}$ may contain multiplicities of the $G_t(x)$ terms. This means that, for a given $t = r > q$ for which $G_r[h_r\{h_s^{-1}(x)\}] = G_r(x)$,

there may be at least another $p \neq r$ and $p > q$ such that $G_p[h_p\{h_s^{-1}(x)\}] = G_r(x)$. Let j be the number of all p's for all r's as defined above. Then the total number of $G_s(x)$ terms over the right hand sides of all $F_i(x)$ and $F_s(x)$ is $q + (T - q - j)q$. Marginal calibration imposes the condition that

$$q + (T - q - j)q = T \quad \Rightarrow \quad q^2 - q(T - j + 1) + T = 0.$$

For the above quadratic equation to have an integer solution in q, the discriminant must be a perfect square, which happens if and only if $j = 0$. Hence a non-climatological forecaster who is both finite marginally and probabilistically calibrated must have issued ideal forecasts.

C Proofs of Propositions 2 and 3

Proof of proposition 2: The second partial derivative of equation (8) with respect to the mixture parameter α yields

$$\frac{\partial^2 \langle \text{IGN} \rangle}{\partial \alpha^2} = \frac{1}{T} \sum_{t=1}^{T} \left\{ \frac{\rho^{(t)}(s_t|\mathcal{V}_T) - \rho_c(s_t)}{f^{(t)}(s_t|\mathcal{V}_T)} \right\}^2.$$

Hence the first derivative of $\langle \text{IGN} \rangle$ with respect to α is an increasing function of α. It follows that the first derivative will have a zero at some $\alpha = \alpha_* \in (0, 1)$ if and only if

$$\left. \frac{\partial \langle \text{IGN} \rangle}{\partial \alpha} \right|_{\alpha=0} < 0 \quad \text{and} \quad \left. \frac{\partial \langle \text{IGN} \rangle}{\partial \alpha} \right|_{\alpha=1} > 0.$$

These are essentially the inequalities in the proposition. The second derivative implies that α_* is a global minimiser of the score.

Proof of proposition 3: $\partial \langle \text{IGN} \rangle / \partial \sigma = 0$ implies that

$$\sigma_*^2 \frac{1}{T} \sum_{t=1}^{T} \frac{\rho^{(t)}(s_t|\mathcal{V}_T)}{f^{(t)}(s_t|\mathcal{V}_T)} = \frac{1}{T} \sum_{t=1}^{T} \left\{ s_t - X_1^{(t)} \right\}^2 \frac{\rho^{(t)}(s_t|\mathcal{V}_T)}{f^{(t)}(s_t|\mathcal{V}_T)}.$$

But $\partial \langle \text{IGN} \rangle / \partial \alpha = 0$ implies that

$$\frac{1}{T} \sum_{t=1}^{T} \frac{\rho^{(t)}(s_t|\mathcal{V}_T)}{f^{(t)}(s_t|\mathcal{V}_T)} = 1,$$

which may be plugged into the left hand side of the previous equation to complete the proof.