Combining Forecasts – Forty Years Later

Kenneth F. Wallis

Department of Economics
University of Warwick
Coventry CV4 7AL, UK
[K.F.Wallis@warwick.ac.uk]

For a special issue of *Applied Financial Economics* in memory of C.W.J. Granger

Headnote This article is dedicated to the memory of Clive Granger, a founding editor of this journal. Its title echoes the title of his invited review article in a special issue of the *Journal of Forecasting* in 1989. That issue marked the twentieth anniversary of the publication of his article with John Bates, which is widely regarded as the seminal article in the field of forecast combination. This article returns to two of the topics in 'Combining forecasts – twenty years later' that are of much current interest, namely the impact of forecasters' different information sets on the original point forecast combination result, and properties of different methods of combining density forecasts. A parallel result to his inefficiency-of-mean-forecasts result for point forecasts is seen to apply to density forecasts, where logarithmic combination is shown to have some advantage over linear combination.

Keywords Point forecast combination; Density forecast combination; Forecast efficiency; Calibration; Linear pool; Logarithmic pool

JEL codes C22, C53, E17

Acknowledgments I am grateful to Christopher Crowe for drawing attention to the work of Kim, Lim and Shaw (2001), to James Mitchell for running additional simulations on the model of Mitchell and Wallis (2010), and to Peter Hammond for helpful discussion.

1. Introduction

Clive Granger told in his Nobel autobiography how he came to forecasting as a line of research that 'had great potential' by reading an advance copy of Box and Jenkins' book in 1968. Two important articles quickly appeared, both in the British OR journal. First was an article on forecasting with generalised cost functions, which has fewer citations than the second article but is currently enjoying a resurgence of interest, as research finds increasing evidence that responses to forecast surveys in several countries exhibit important departures from the quadratic loss function that is found so convenient in theoretical work. Second was the seminal article with John Bates on forecast combination, which opened up a whole new sub-field in forecasting. Although Bates and Granger (1969) cited an earlier article by Barnard (1963) which contained an empirical example in which a simple average of two forecasts had smaller Mean Square Error (MSE) than either of the individual forecasts, theirs was the first work to develop a general analysis of the point forecast combination problem. Twenty years after its publication, Clemen (1989) provided a review of the new field and an annotated bibliography containing over 200 items, which he described as 'an explosion in the number of articles on the combination of forecasts.' Clemen's article in the *International Journal of Forecasting* was accompanied by several commentaries and reflections by leading researchers, while the *Journal of Forecasting* simultaneously published a special issue on combining forecasts, in which the first article was 'Combining forecasts – twenty years later' (Granger, 1989). This article returns to two of the topics in that article, forty years after Bates and Granger's work.

Clive Granger's remarkably long list of publications includes several articles in which he discussed current developments and future prospects in particular fields of time-series econometrics. They are full of his thoughts on where more research was needed and his ideas on how that research could be developed. They are more forward-looking than backward-looking, containing very little by way of literature review. In the present case, although the *Journal of Forecasting* described his article as an 'invited review article', he explicitly excluded an 'attempt to survey the literature on combining, which is now voluminous and rather repetitive.' More often, such an exclusion was implicit rather than explicit, without any excuse being offered, and the pages were devoted to an unceasing flow of new ideas and creative suggestions, which became an invaluable resource for other researchers. Not all of these ideas have stood the test of time, but his success rate is notably high.

A major theme of ‘Twenty years later’ is the relevance of information sets, discussion of which ‘is an essential feature of all forecasting situations, including combining.’ Section 2 of this article revisits its analysis of how the original treatment of point forecast combination is affected when competing forecasters have different information sets, and all of them use their own information efficiently. Explicit expressions are presented for the forecast MSEs that support its main result, referred to as ‘the inefficiency of mean forecasts’ in independent work by a later group of authors. The simple average of several available forecasts has been a popular summary device for as long as competing forecasts have been available, and how it might be improved, in similarly simple fashion, remains a topic of interest.

The final ‘extensions’ considered in ‘Twenty years later’ concern the combination of uncertainty measures such as quantiles, which ‘is a step towards the larger question of how we would combine or aggregate distributions.’ As I have observed (Wallis, 2005), distributions have been aggregated ever since Zarnowitz (1969) published averages of the density forecasts of inflation and GNP growth supplied in histogram form by respondents to the ASA-NBER survey of forecasts by economic statisticians (continued since 1990 as the Survey of Professional Forecasters). I noted that the finite mixture distribution is an appropriate statistical model for such combined density forecasts; the same algebraic form appears in statistical decision theory as the linear opinion pool, and the two terms have come to be used interchangeably. On the other hand the quantile-based methods of combining forecast distributions proposed in ‘Twenty years later’ have not been taken up.

A popular choice of functional form for interval and density forecasts is the normal distribution. When combining density forecasts it might be desired for presentational reasons and/or analytical convenience that the combined forecast retains the same functional form as the component forecasts, but in the normal case linear combination does not deliver this. Section 3 of this article considers the properties of the logarithmic opinion pool or geometric average of component normal density forecasts, which does preserve this distributional form, as an alternative to linear combination. An example is constructed as an additional variant forecast in the Monte Carlo experiment of Mitchell and Wallis (2010). With respect to the probabilistic calibration and goodness-of-fit of combined density forecasts, some parallels with the inefficiency of mean (point) forecasts discussed in Section 2 are observed.

2. Point forecasts with different information sets

The basic optimality result of Bates and Granger (1969) is that a linear combination of two competing point forecasts using the optimal (variance minimising) weight in general has a smaller forecast MSE than either of the two competing forecasts. It might have seemed counter intuitive that combining with an inferior forecast could improve matters, relative to the MSE measure, but combining using the weight that is optimal with respect to the same measure is seen to achieve this. The only case in which no improvement is possible is that in which one forecast is already the optimal (minimum MSE) forecast; its optimal weight is then 1, and there is no gain in combining with an inferior forecast, whose weight is correspondingly 0. Twenty years later Granger (1989, p.168) noted that in this original setting the two forecasts were implicitly based on the same information set, and 'For combining to produce a superior forecast, both component forecasts clearly had to be suboptimal. It is more usual for combining to produce a better forecast when the individual forecasts are based on different information sets, and each may be optimal for their particular set.'

A stylised representation of this more usual situation is obtained by assuming that each forecaster's information set has two components, one available to all *N* forecasters and the other available only to the specific individual, with the various components assumed to be mutually independent. All the information available to all the forecasters is referred to as the universal information set. The target variable is denoted y_t, and the contents of the public and private information sets are the current and past values of the variables z_t and x_{jt}, $j = 1,...,N$, respectively. These could be vector series, but are taken to be univariate for convenience. If the universal information set U_t were known, the optimum linear least squares one-step-ahead forecast would be

$$\begin{aligned} E_t &= \mathrm{E}\left(y_{t+1} \middle| U_t\right) \\ &= \alpha(B) z_t + \sum_{j=1}^{N} \beta_j(B) x_{jt} \end{aligned} \tag{1}$$

where $\alpha(B), \beta_j(B)$ are lag operator polynomials. This is equation (1) of Granger (1989), with subscript *t* replacing his subscript *n*. He continues 'With the assumptions made, the optimum forecast of the *j*th person is

$$E_{jt} = \alpha(B)z_t + \beta_j(B)x_{jt}\,. \tag{2}$$

Forming a simple average of these individual forecasts gives

$$\bar{E}_t = \frac{1}{N}\sum_{j=1}^{N} E_{jt}$$

$$= \alpha(B)z_t + \frac{1}{N}\sum_{j=1}^{N}\beta_j(B)x_{jt}\,.' \tag{3}$$

Clearly, neither this equally-weighted combination of the individual forecasts nor any other combination with weights adding to one achieves the optimum forecast E_t. 'It is seen that aggregating forecasts is not the same as aggregating information sets. $\bar{E}_t$ is based on all the information, but is not equal to E_t, as the information is not being used efficiently.'

However, if yet another forecast is constructed, using only the public information, namely

$$E_{0t} = \alpha(B)z_t \tag{4}$$

'then it is seen immediately that the optimal forecast E_t can be achieved by

$$E_t = \sum_{j=1}^{N} E_{jt} - (N-1)E_{0t}\,.' \tag{5}$$

No direct comparisons of forecast efficiency are given in this analysis, and we pursue these in a slightly modified setting. We redefine z_t and x_{jt}, $j = 1,...,N$, as the components of the forecasts of y_{t+1} based on public and private information respectively, and denote the respective MSEs of z_t and x_{jt} as σ_0^2 and σ_1^2, constant over j. Then the individual forecasts constructed as the optimally 'inverse MSE' weighted combination of the two components, and their MSEs, are

$$E_{jt} = \frac{\sigma_1^2 z_t + \sigma_0^2 x_{jt}}{\sigma_1^2 + \sigma_0^2}, \tag{2'}$$

$$\mathrm{MSE}\left(E_{jt}\right) = \frac{\sigma_0^2\sigma_1^2}{\sigma_1^2 + \sigma_0^2},\ j = 1,...,N\,.$$

We note that this alternative definition of z_t and x_{jt} does not affect the structure of the forecast E_{jt}: the variance weights in equation (2') are incorporated in the least squares

coefficients $\alpha(B)$ and $\beta_j(B)$ of the originally defined variables in equation (2). The corresponding mean of the *N* individual forecasts, and its MSE, is

$$\bar{E}_t = \frac{\sigma_1^2 z_t + \frac{1}{N}\sum_{j=1}^{N} \sigma_0^2 x_{jt}}{\sigma_1^2 + \sigma_0^2}, \tag{3'}$$

$$\mathrm{MSE}\left(\bar{E}_t\right) = \frac{\sigma_0^2 \sigma_1^2 \left(\sigma_1^2 + \frac{\sigma_0^2}{N}\right)}{\left(\sigma_1^2 + \sigma_0^2\right)^2}.$$

It is clear that the MSE of the mean forecast is less than the MSE of each individual forecast, increasingly so as *N* increases. The optimal forecast given the universal information set is a similarly modified version of equation (1); together with its MSE we have

$$E_t = \frac{\sigma_1^2 z_t + \sum_{j=1}^{N} \sigma_0^2 x_{jt}}{\sigma_1^2 + N\sigma_0^2}, \tag{1'}$$

$$\mathrm{MSE}\left(E_t\right) = \frac{\sigma_0^2 \sigma_1^2}{\left(\sigma_1^2 + N\sigma_0^2\right)}.$$

The inefficiency of the mean forecast, that is, $\mathrm{MSE}\left(\bar{E}_t\right) > \mathrm{MSE}\left(E_t\right)$ follows on observing that

$$\left(\sigma_1^2 + \frac{\sigma_0^2}{N}\right)\left(\sigma_1^2 + N\sigma_0^2\right) > \left(\sigma_1^2 + \sigma_0^2\right)^2 \quad \text{if } N>1.$$

Finally, we see that the optimal forecast can be achieved as a combination of the mean forecast and the public forecast, as above. Noting that equation (5) gives the optimal forecast as the combination $k\bar{E}_t + (1-k)E_{0t}$ with weight $k = N$, the effect of the modifications in the present paragraph is to require

$$k = \frac{N\left(\sigma_1^2 + \sigma_0^2\right)}{\sigma_1^2 + N\sigma_0^2}$$

in order to obtain the corresponding result.

Equivalent results to these modified results, except the last one, are given completely independently by Kim, Lim and Shaw (2001), working in accounting and finance and making no reference to the statistical forecasting literature. They consider several financial analysts

forecasting a firm's forthcoming earnings. Each analyst possesses both common and private information, in the form of a public and a private signal, each equal to the forthcoming earnings plus a zero-mean random error. The assumed source of random variation – noise in the signals – differs from that assumed in the forecasting literature – the innovation in the data-generating process – but the main result, referred to as 'the inefficiency of mean forecasts' and expressed in terms of the precision of the signals rather than forecast MSEs, is the same. However Kim, Lim and Shaw do not consider the possibility of constructing a fully efficient forecast, with no equivalent to equation (5) above, since 'neither analysts' common nor private information is separately observable' (2001, p.330). Instead they consider the exploitation of a sequence of signals and associated forecast revisions in a fixed-event forecast context. Models in which agents receive noisy public and private signals about the underlying state have a long history in economics, serving to relax the Walrasian postulate of complete information: see, for example, the references given by Morris and Shin (2002) prior to their analysis of the social value of public information. As with the forecasting literature, however, Kim, Lim and Shaw's work is independent of this.

Crowe (2010) similarly considers the fixed-event context of the forecasts of economic growth collected by *Consensus Economics* in several countries: forecasts for the current year and the next year are collected monthly. The source of the inefficiency in the mean or 'consensus' forecast is interpreted as its overweighting of public information – compare equations (1) and (3) or (1') and (3'). In fixed-event forecasting the public information includes previously published mean forecasts of the same target variable, and hence the successive revisions to those mean forecasts. The manifestation of the inefficiency is (negative) correlation between the mean forecast and its forecast error, the latter being a function of all past forecast revisions. Consideration of this relationship allows an estimate of the efficient forecast to be obtained via an adjustment to the current mean forecast which depends on the latest forecast revision. With the model estimated for 38 countries over the period 1989-2006, evaluation of out-of-sample current-year forecasts for 2007 and 2008 shows that efficiency gains of some 5% root mean square error could have been achieved.

3. Linear and logarithmic combination of density forecasts

We consider *N* individual density forecasts of a random variable *Y* at some future time, denoted $f_j(y),\ j=1,...,N,$ with means and variances $\mu_j,\ \sigma_j^2.$ For economy of notation time subscripts and references to the information sets on which the forecasts are conditioned are suppressed. The linear combination or finite mixture distribution is

$$f_C(y)=\sum_{j=1}^{N} w_j f_j(y), \tag{6}$$

with weights $w_j \geq 0,\ j=1,...,N,\ \Sigma w_j = 1.$ The combined density has mean and variance

$$\mu_C = \sum_{j=1}^{N} w_j \mu_j\,, \tag{7}$$

$$\sigma_C^2 = \sum_{j=1}^{N} w_j \sigma_j^2 + \sum_{j=1}^{N} w_j \left(\mu_j - \mu_C\right)^2\,. \tag{8}$$

If the density forecast means are equal to the point forecasts, and equal weights $w_j = 1/N$ are used, then equation (7) gives the simple average point forecast, and equation (8) gives the variance of the combined density as the average individual variance plus a measure of the dispersion of the individual point forecasts. This decomposition of the combined variance into measures of uncertainty and disagreement is employed in recent studies of survey forecasts in the UK (Boero, Smith and Wallis, 2008) and the US (Rich and Tracy, 2010). As noted above, the linear combination of normal densities is not a normal density, indeed it may be multimodal; it was the consideration of a mixture of two normal distributions by Pearson (1894) that initiated the finite mixture distribution literature.

The logarithmic opinion pool first appeared, according to Genest and Zidek's (1986) survey of combining probability distributions, in an unpublished 1972 manuscript by Michael Bacharach, who in turn, they say, attributed it to Peter Hammond, but there appears to be no surviving copy of either Bacharach's manuscript or Hammond's communication. A further remark by Genest and Zidek (1986, p.119) suggests that Winkler (1968) earlier referred to the logarithmic opinion pool, and later authors, for example Kascha and Ravazzolo (2010, pp.233, 237), similarly reference Winkler, but this is not correct: Winkler (1968) does not mention the logarithmic opinion pool. For Genest and Zidek, the most compelling reason for using the logarithmic opinion pool is that the processes of pooling and updating then

commute. That is, suppose (a) individual distributions are first combined, then the combined distribution is updated following Bayes' rule using some later observation, or (b) individual distributions are first updated using the later (common) observation, then combined: with logarithmic pooling both approaches give the same result, but not with linear pooling. For this reason the logarithmic opinion pool is sometimes said to be 'externally Bayesian' (see, for example, Faria and Smith, 1997), since to an external observer the combined forecast looks like that of a single Bayesian forecaster.

The logarithmic combination is usually written using the geometric form, as

$$f_G(y) = \frac{\prod f_j^{w_j}(y)}{\int \prod f_j^{w_j}(y)dy}, \tag{9}$$

where the denominator is a constant that ensures that $f_G(y)$ is a proper density. Our main interest lies in the logarithmic combination's preservation of the normality of component distributions; more generally, if the component distributions are from the regular exponential family, then the combined distribution is from the same family and hence, in particular, is unimodal (Faria and Mubwandarikwa, 2008). Formally, for the normal case, we have that if $f_j(y) = \mathcal{N}\left(\mu_j, \sigma_j^2\right),\ j = 1,...,N,$ then $f_G(y) = \mathcal{N}\left(\mu_G, \sigma_G^2\right)$ where

$$\frac{\mu_G}{\sigma_G^2} = \sum_{j=1}^{N} w_j \frac{\mu_j}{\sigma_j^2}, \tag{10}$$

$$\frac{1}{\sigma_G^2} = \sum_{j=1}^{N} w_j \frac{1}{\sigma_j^2}. \tag{11}$$

With equal weights $w_j = 1/N$ on the densities, equation (11) gives σ_G^2 as the harmonic mean of the individual variances; this is less than their arithmetic mean, which is less than the finite mixture variance given in equation (8). Substituting into equation (10) then gives the combined mean or point forecast as a linear combination of the component means with inverse variance weights. Thus logarithmic combination with equal weights delivers a combined point forecast that is more efficient than the linear combination with equal weights, unless the component forecast variances are equal, in which case the two combined means coincide. In practice it is often found that competing forecast variances are not very different from one another, in which case it may be more efficient to impose equal weights than to

estimate the weights – the squared bias is less than the estimation variance (Smith and Wallis, 2009).

Example

We add a logarithmic combination to some of the competing forecasts that appear in the simulation study of Mitchell and Wallis (2010), which extends the example given by Wallis (2005). The data-generating process is the Gaussian second-order autoregression

$$Y_t = \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t, \;\; \varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right).$$

The universal information set comprises the observations y_{t-1} and y_{t-2}, the model and its parameter values, hence the important practical effects of model uncertainty and parameter estimation error are neglected. Then the optimal or 'ideal' density forecast of y_t is

$$f_o\left(y_t\right) = \mathcal{N}\left(\phi_1 y_{t-1} + \phi_2 y_{t-2}, \sigma_\varepsilon^2\right).$$

The first individual forecaster (labelled AR1 below) assumes that the data are generated by a first-order autoregression and so issues the forecast

$$f_1\left(y_t\right) \;= \mathcal{N}\left(\rho_1 y_{t-1}, \sigma_1^2\right),$$

while the second individual (AR2) similarly uses only a single observation but is subject to a one-period data delay, so the forecast is

$$f_2\left(y_t\right) = \mathcal{N}\left(\rho_2 y_{t-2}, \sigma_2^2\right),$$

where $\sigma_1^2 = \left(1-\rho_1^2\right)\sigma_y^2$, $\sigma_2^2 = \left(1-\rho_2^2\right)\sigma_y^2$, $\sigma_\varepsilon^2 = \left(1-\phi_1\rho_1-\phi_2\rho_2\right)\sigma_y^2$ and ρ_i, $i = 1, 2$, are the autocorrelation coefficients:

$$\rho_1 = \phi_1 / \left(1-\phi_2\right), \;\; \rho_2 = \phi_1\rho_1 + \phi_2 .$$

The linear combination of these two individual forecasts, with equal weights, is the finite mixture distribution

$$f_C\left(y_t\right) = 0.5\mathcal{N}\left(\rho_1 y_{t-1}, \sigma_1^2\right) + 0.5\mathcal{N}\left(\rho_2 y_{t-2}, \sigma_2^2\right),$$

with mean and variance obtained from equations (7) and (8). The logarithmic combination is a normal distribution with mean and variance given as

$$\mu_G = \frac{\sigma_2^2\rho_1 y_{t-1} + \sigma_1^2\rho_2 y_{t-2}}{\sigma_1^2 + \sigma_2^2},$$

$$\sigma_G^2 = \frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}.$$

The composite information set for the combined density forecasts includes the same two observations as the information set of the optimal forecast. However the combined forecasts use this information inefficiently, relative to the optimal forecast. This parallels the inefficiency of the mean (point) forecast discussed in Section 2.

A standard tool for checking the distributional form of density forecasts is the sequence of probability integral transform (PIT) values of the outcomes in the forecast distributions

$$p_t = F_t(y_t),$$

where $F_t(.)$ is the forecast cumulative distribution function. For the ideal forecast, the p_ts are independent uniform $U(0,1)$ variables. If the forecast is the correct conditional distribution with respect to its specific information set, then the p_ts are uniformly but in general not independently distributed: such a forecast satisfies 'probabilistic calibration' but not 'complete calibration'. Uniformity is often assessed in an exploratory manner, by inspection of histograms of PIT values, for example, and goodness-of-fit tests are also available, although their performance is affected by departures from independence. For a joint test of complete calibration we consider the likelihood ratio test of Berkowitz (2001), based on the inverse normal transformation of the p_ts, namely $z_t = \Phi^{-1}(p_t)$, where $\Phi(\cdot)$ is the standard normal distribution function. If p_t is iid$U(0,1)$, then z_t is iid $\mathcal{N}(0,1)$. Under a maintained hypothesis of normality, the joint null hypothesis of correct mean and variance ('goodness-of-fit') and independence is tested against a first-order autoregressive alternative with mean and variance possibly different from (0,1), via a likelihood ratio test with three degrees of freedom.

Comparisons of forecast performance can be based on the logarithmic score, defined for forecast density f_{jt} as

$$\log S_j(y_t) = \log f_{jt}(y_t).$$

To a Bayesian the logarithmic score is the logarithm of the predictive likelihood, and if two forecasts are being compared, the log Bayes factor is the difference in their logarithmic

scores. If one of the forecasts is the ideal forecast, then the expected difference in their logarithmic scores is the Kullback-Leibler information criterion (KLIC) or distance measure

$$\text{KLIC}_{jt} = E_o\left\{\log f_{ot}(y_t) - \log f_{jt}(y_t)\right\}.$$

If the ideal forecast is known, as in a simulation experiment, competing forecasts can be tested against it via their average logarithmic scores, using the transformed variables z_t (Mitchell and Hall, 2005). For density forecasts based on normal distributions we have (Mitchell and Wallis, 2010), subscripting parameters appropriately,

$$\text{KLIC}_{jt} = -\tfrac{1}{2} - \tfrac{1}{2}\log\left(\frac{\sigma_o^2}{\sigma_j^2}\right) + \tfrac{1}{2}\frac{\sigma_o^2}{\sigma_j^2} + \frac{\left(\mu_o - \mu_j\right)^2}{2\sigma_j^2}.$$

This has a minimum at zero: the sum of the first three terms on the right-hand side is non-negative, as is the fourth term. Thus a positive KLIC may result from departures in mean and/or variance in either direction, and additional investigation, via the PIT histogram, for example, is needed to discover the direction of any departure. For example the competing forecast may be too dispersed or not dispersed enough, indicated by a hump-shaped or U-shaped PIT histogram, respectively.

The relative performance of the two combination methods is assessed by adding the logarithmic combination to the AR1, AR2 and linear combination forecasts that appear (among others) in the simulation study of Mitchell and Wallis (2010), retaining all other features of their experimental design, with 500 replications of samples of 150 observations. The four pairs of values of the autoregressive parameters ϕ_1 and ϕ_2 used are shown, together with the corresponding first- and second-order autocorrelation coefficients and the associated values of the density forecast variances, in Table 1. Case (1) represents a relatively persistent series, whereas case (2) exhibits rather less persistence than is observed in inflation and GDP growth, for example: it is tending towards white noise, in which case all these forecasts would be identical. The structure of case (3) is such that the AR1 forecast coincides with the unconditional forecast, hence its variance is equal to the variance of the observed series, while the AR2 forecast coincides with the ideal forecast, with variance equal to the innovation variance. Case (4) represents a rather unusual oscillating form. It is seen that, in all cases, $\sigma_G^2 < \sigma_C^2$, as anticipated in the discussion following equation (11) above. The difference is greatest in case (3), where the component variances are most unequal and the

Table 1. Simulation design and density forecast variances

Case	Parameter		Autocorrelation		Density forecast variance			
	ϕ_1	ϕ_2	ρ_1	ρ_2	σ_1^2	σ_2^2	σ_C^2	σ_G^2
(1)	1.5	–0.6	0.94	0.81	1.56	4.52	3.40	2.32
(2)	0.15	0.2	0.19	0.23	1.04	1.02	1.05	1.03
(3)	0	0.95	0	0.95	10.26	1	7.94	1.82
(4)	–0.5	0.3	–0.71	0.66	1.10	1.27	1.34	1.18

Note: $\sigma_\varepsilon^2 = 1$ in all cases

expected disagreement term in equation (8) also contributes strongly to σ_C^2. In cases (2) and (4) σ_C^2 exceeds both the component variances, whereas σ_G^2 always lies between σ_1^2 and σ_2^2, of course.

For initial diagnostic checking we consider histograms of PIT values, to allow an informal assessment of their uniformity and hence of the probabilistic calibration of the forecasts. This is expected to hold for the two separate components of the combinations, that is, the AR1 and AR2 forecasts, since each of these represents the correct conditional distribution in respect of its normality and its first two moments conditional on the past data being used. The results presented in Figure 1 entirely meet these expectations, with all the histograms in the first two columns being essentially uniform. Despite this, the PIT histograms for the combinations of these forecasts in the third and fourth columns show departures from uniformity, most dramatically in cases (3) and (4). The histograms are approximately symmetric but hump-shaped, thus the dominant effect is the increased variance of these forecasts relative to the correct conditional forecast for the combined information set: too few observations lie in the tails of these forecast densities. In the rows with the greater departures from uniformity, comparison of the third and fourth columns indicates that the logarithmic combination is performing better than the linear combination, having a smaller variance and the correct distributional form, but the relative contributions of these two factors are not distinguishable.

Table 2. Tests of complete calibration and forecast performance: rejection percentages at nominal 5% level

	Case (1)		Case (2)		Case (3)		Case (4)	
Forecast	Bk	KLIC	Bk	KLIC	Bk	KLIC	Bk	KLIC
AR1	100	98	17	25	99	100	62	46
AR2	100	100	30	15	5.6	n.a.	97	93
Lin combn	100	100	14	10	100	100	62	55
Log combn	100	100	12	7	99	84	37	25

Note: Bk, the likelihood ratio test of Berkowitz (2001); KLIC, the test of KLIC differences vs. the ideal forecast of Mitchell and Hall (2005)

The results of formal tests of complete calibration and forecast performance are reported in Table 2. The AR1 and AR2 forecasts are probabilistically calibrated, as shown in Figure 1, so the Berkowitz test is essentially a test of the independence of the z_ts, and indirectly a test of the independence of the p_ts. It is seen that the test has good power: in case (1), a relatively persistent series, there are no Type 2 errors in the 500 replications; the rejection rate is smallest in case (2), which is closest to a white noise series. In case (3) the AR2 forecast is the same as the ideal forecast, and the rejection rate is not significantly different from the nominal size of the test. Otherwise the test gives a good indication that the two individual forecasts do not satisfy complete calibration. The results of the KLIC-based test of forecast performance against the ideal forecast give very much the same finding.

The combination forecasts satisfy neither probabilistic calibration nor independence, and can be expected to be rejected against the ideal forecast. The main interest lies in the possibility of discriminating between the linear and logarithmic combinations. In case (1), no discrimination is possible, since there is no failure to reject either forecast using either test. In the remaining cases the results favour the logarithmic combination, which is rejected less frequently by either test than the linear combination. Finally, with respect to these measures some gains from combining can be observed, with the logarithmic combination being rejected less frequently than either of the two component forecasts in cases (2) and (4).

Discussion

The combination density forecasts in our example are constructed using equal weights, whereas there is an emerging literature on the problem of estimating the weights in practice. It seems to be agreed to base the weights on past forecast performance as measured by the average logarithmic score, possibly allowing for time variation in the weights (Hall and Mitchell, 2007; Geweke and Amisano, 2009; Jore, Mitchell and Vahey, 2010; Kascha and Ravazzolo, 2010). The first three groups of authors only consider linear combinations, whereas the last-named also include logarithmic combinations. Using the same performance measure they find no clear ranking of the two forms of combination of the various (linear) forecasting models of inflation they construct for four countries, but no information on their comparative calibration characteristics is given. However a similar exercise conducted at the same institution (Norges Bank), studying combinations of forecasting models for Norwegian GDP growth and inflation, presents PIT histograms for linear and logarithmic pools (Bjornland *et al*., 2009, Figures 15-18). For both variables the logarithmic combination appears to be better calibrated, with the linear combination having too few observations in the tails of the histograms, indicating excessive dispersion in the combined forecast densities, as in our example above.

The weak statistical evidence favouring the logarithmic over the linear combination of density forecasts may be augmented in practice by the advantages of a combined forecast that retains the distributional form of its components. A committee of individual forecasters, each producing a density forecast according to a given specification, and charged with producing a composite forecast, will find it more convenient analytically and easier to communicate the results if the composite forecast has the same form, which the logarithmic combination is more likely to deliver. When the normal distribution is used, this applies exactly.

The result that a combination of two probabilistically calibrated density forecasts is not itself probabilistically calibrated parallels the inefficiency of mean point forecasts discussed in Section 2. In each case the component forecasts are the correct conditional statements with respect to their specific information sets, but their combination, with respect to the combined information set, is not. The same result also obtains for combinations of event probability forecasts (Hora, 2004; Ranjan and Gneiting, 2010). In these broader

circumstances, it remains the case that 'aggregating forecasts is not the same as aggregating information sets' (Granger, 1989, p.169), a remark that, like so many others of Clive Granger's, requires our continuing attention.

References

Barnard, G.A. (1963). New methods of quality control. *Journal of the Royal Statistical Society A*, 126, 255-258.

Bates, J.M. and Granger, C.W.J. (1969). The combination of forecasts. *Operational Research Quarterly*, 20, 451-468.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business and Economic Statistics*, 19, 465-474.

Bjornland, H., Gerdrup, K., Jore, A.S., Smith, C. and Thorsrud, L.A. (2009). There is more than one weight to skin a cat: combining densities at Norges Bank. Conference on Forecasting and Monetary Policy, Deutsche Bundesbank, Berlin, March 2009.

Boero, G., Smith, J. and Wallis, K.F. (2008). Uncertainty and disagreement in economic prediction: the Bank of England Survey of External Forecasters. *Economic Journal*, 118, 1107-1127.

Box, G.E.P and Jenkins, G.M. (1970). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day

Clemen, R.T. (1989). Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting*, 5, 559-583.

Crowe, C. (2010). Consensus forecasts and inefficient information aggregation. Working Paper No.10/178, International Monetary Fund, Washington DC.

Faria, A.E.and Mubwandarikwa, E. (2008). The geometric combination of Bayesian forecasting models. *Journal of Forecasting*, 27, 519-535.

Faria, A.E. and Smith, J.Q. (1997). Conditionally externally Bayesian pooling operators in chain graphs. *Annals of Statistics*, 25, 1740-1761.

Genest, C. and Zidek, J.V. (1986). Combining probability distributions: a critique and an annotated bibliography (with discussion). *Statistical Science*, 1, 114-148.

Geweke, J. and Amisano, G. (2009). Optimal prediction pools. Working Paper No.1017, European Central Bank, Frankfurt.

Granger, C.W.J. (1969). Prediction with a generalized cost of error function. *Operational Research Quarterly*, 20, 199-207.

Granger, C.W.J. (1989). Combining forecasts – twenty years later. *Journal of Forecasting*, 8, 167-173.

Hall, S.G. and Mitchell, J. (2007). Combining density forecasts. *International Journal of Forecasting*, 23, 1-13.

Hora, S.C. (2004). Probability judgements for continuous quantities: linear combinations and calibration. *Management Science*, 50, 597-604.

Jore, A.S., Mitchell, J. and Vahey, S.P. (2010). Combining forecast densities from VARs with uncertain instabilities. *Journal of Applied Econometrics*, 25, 621-634.

Kascha, C. and Ravazzolo, F. (2010). Combining inflation density forecasts. *Journal of Forecasting*, 29, 231-250.

Kim, O., Lim, S.C. and Shaw, K.W. (2001). The inefficiency of the mean analyst forecast as a summary forecast of earnings. *Journal of Accounting Research*, 39, 329-336.

Mitchell, J. and Hall, S.G. (2005). Evaluating, comparing and combining density forecasts using the KLIC with an application to the Bank of England and NIESR 'fan' charts of inflation. *Oxford Bulletin of Economics and Statisics*, 67, 995-1033.

Mitchell, J. and Wallis, K.F. (2010). Evaluating density forecasts: forecast combinations, model mixtures, calibration and sharpness. *Journal of Applied Econometrics*, in press.

Morris, S. and Shin, H.S. (2002). Social value of public information. *American Economic Review*, 92, 1521-1534.

Pearson, K. (1894). Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London*, A, 185, 71-110.

Ranjan, R. and Gneiting, T. (2010). Combining probability forecasts. *Journal of the Royal Statistical Society* B, 72, 71-91.

Rich, R. and Tracy, J. (2010). The relationships among expected inflation, disagreement, and uncertainty: evidence from matched point and density forecasts. *Review of Economics and Statistics*, 92, 200-207.

Smith, J. and Wallis, K.F. (2009). A simple explanation of the forecast combination puzzle. *Oxford Bulletin of Economics and Statistics*, 71, 331-355.

Wallis, K.F. (2005). Combining interval and density forecasts: a modest proposal. *Oxford Bulletin of Economics and Statistics*, 67, 983-994.

Winkler, R.L. (1968). The consensus of subjective probability distributions. *Management Science*, 15, B61-B75.

Zarnowitz, V. (1969). The new ASA-NBER survey of forecasts by economic statisticians. *American Statistician*, 23(1), 12-16.

Figure 1. PIT histograms for the autoregressive example. Rows: cases (1)-(4) as defined in Table 1. Columns: forecasts, respectively AR1, AR2, their linear combination and their logarithmic combination