Density Forecast Combination

Stephen G. Hall and James Mitchell*
Imperial College, London and NIESR

November 8, 2004

Abstract

In this paper we investigate whether and how far density forecasts sensibly can be combined to produce a "better" pooled density forecast. In so doing we bring together two important but hitherto largely unrelated areas of the forecasting literature in economics, density forecasting and forecast combination. We provide simple Bayesian methods of pooling information across alternative density forecasts. We illustrate the proposed techniques in an application to two widely used published density forecasts for U.K. inflation. We examine whether in practice improved density forecasts for inflation, one year ahead, might have been obtained if one had combined the Bank of England and NIESR density forecasts or "fan charts".

1 Introduction

Forecasts of the future values of economic variables are used widely in decision making. For example in the U.K., inflation forecasts are central to the setting of monetary policy by the Monetary Policy Committee (MPC) at the Bank of England. Recently there has been increased attention given to providing measures of uncertainty associated with these forecasts. Measures of uncertainty surrounding a "central tendency" (the point forecast) can enhance the usefulness of the forecast; e.g. see the discussion in Garratt et al. (2003). So called "density" forecasts are being used increasingly in economics and finance since they provide commentators with a full impression of the uncertainty associated with the forecasts; see Tay & Wallis (2000) for a review. More formally, density forecasts of inflation provide an estimate of the probability distribution of its possible future values. In contrast

*Address for correspondence: James Mitchell, National Institute of Economic and Social Research, 2 Dean Trench Street, Smith Square, London, SW1P 3HE, U.K. Tel: +44 (0) 207 654 1926. Fax: +44 (0) 207 654 1900. E-Mail: j.mitchell@niesr.ac.uk. Thanks to Ray Barrell, Robert Metz and Rebecca Riley for their advice, in particular, about NIESR forecasts, Nigel Meade, Kostas Mouratidis and Martin Weale for helpful comments and Yongmiao Hong for making available GAUSS code to implement his test referred to below. Mitchell gratefully acknowledges financial support from the ESRC (Award Reference: RES-000-22-0610). Computations were performed using the GAUSS and Ox programming languages.

to interval forecasts, that state the probability that the outcome will fall within a stated interval such as inflation falling between 1% and 3%, density forecasts provide a complete description of the uncertainty associated with a forecast; they can be seen to provide information on all possible intervals.

Density forecasts of inflation in the U.K., for example, are now published each quarter both by the Bank of England in its "fan" chart and the National Institute of Economic and Social Research (NIESR) in its quarterly forecast, and have been for the last ten years. Density forecasts inform the user of the forecast about the risks involved in using the forecast for decision making. Indeed, interest may lie in the dispersion or tails of the density itself; for example inflation targets often focus the attention of monetary authorities to the probability of future inflation falling within some pre-defined target range while users of growth forecasts may be concerned about the probability of recession. Moreover, volatility forecasts, as measured by the variance, and other measures of risk and uncertainty, can be extracted from the density forecast.

It is well established that combining competing individual point forecasts of the same event can deliver more accurate forecasts, in the sense of a lower root mean squared error (RMSE); see Bates & Granger (1969) and Stock & Watson (2004).[1] The success of combination follows from the fact that individual forecasts may be based on misspecified models, poor estimation or non-stationarities. Indeed recent work, see e.g. Hendry & Clements (2004), has begun to explore further why point forecast combination works through analytical and Monte-Carlo investigation.

In this paper we take the natural next step of considering density forecast combination, to-date a relatively unexplored area, investigating whether and how far density forecasts sensibly can be combined to produce a "better" pooled density forecast. In so doing we bring together two important but hitherto largely unrelated areas of the forecasting literature in economics, density forecasting and forecast combination.[2]

However, a body of experience about combining probability distributions has accumulated in the management science literature. One popular approach is for a Bayesian motivation for the combination of density forecasts; typically this in the context of a 'decision maker' aggregating the probability forecasts of a group of 'experts'. This literature provides a useful stylised framework in which to consider density forecast combination and serves as the starting point in this paper. Accordingly, we review this literature in Section 2 and then in Section 3 explain how an extension is required to the basic Bayesian model when interested in forecasting other characteristics than the mean of the density. One interpretation of this extension is it accommodates the fact that in economics it is

[1] Analogously, it is well established that it is often better to invest in a portfolio of assets than just a single asset.

[2] Related work has considered the combination of event and quantile forecasts; see Clements (2002) and Granger et al. (1989). These inevitably involve a loss of information compared with consideration of the 'whole' density; e.g. only as the number of quantiles examined reaches infinity is no information about the density lost. Garratt et al. (2003) and Pesaran & Zaffaroni (2004) consider the combination of probability forecasts based on Bayesian model averaging. In contrast, the simple approaches to density combination suggested in this paper are not predicated on estimation of a statistical model; they are operational both with model-based and subjective (e.g. survey-based) density forecasts.

widely acknowledged that data are inherently uncertain, given our belief that data are realisations from stochastic data-generating-processes. In Section 4 we then make the distinction between combining experts' forecasts of various moments of the forecast density and directly combining the individual densities themselves. It is explained how this distinction can prove important in affecting the shape of the combined density forecast. Consistent with the Bayesian paradigm, we derive an analytical expression for the shape of the directly combined density forecast, defined as the posterior distribution of the variable we are seeking to forecast conditional on the information supplied by the experts. The techniques that we consider for the combination of density forecasts offer users of density forecasts useful and easy-to-use means of pooling information across alternative density forecasts, irrespective of whether the density forecasts are model-based or subjective (e.g. survey-based). Section 5 reviews the statistical tests proposed to evaluate the accuracy of density forecasts *ex post*. Section 6 then illustrates the proposed techniques in an application to two widely used density forecasts for U.K. inflation, that have both been published each quarter in real-time for ten years. It examines whether in practice improved density forecasts for inflation, one year ahead, might have been obtained if one had combined the Bank of England and NIESR density forecasts or "fan charts". Concluding comments are made in Section 7.

2 Combining density forecasts

It is common practice to seek the advice of more than one expert before making a decision. The benefits of doing so derive from the assumption that a set of experts provide more information than a single expert. While the benefits of combining information about point forecasts are well appreciated in economics, less attention has been paid to the aggregation of probability distributions. However, this has received considerable attention within many management science and risk analysis journals; for reviews see Genest & Zidek (1986) and Clemen & Winkler (1999). First, it is useful to provide a selective review of this work before considering an extension.

Clemen & Winkler (1999) distinguish behavioural and mathematical approaches to combination. The behavioural approach seeks to combine experts' opinions by letting the experts interact in some manner to reach some collective opinion. This approach is not considered further in this paper as one can imagine many situations in economic forecasting when it would not be feasible practically; e.g. in the context of the application in this paper the Bank of England and NIESR would have to sit down together and agree about the shape of their combined fan chart. While clearly possibly within an institution, indeed the Bank's fan chart is, to a degree, based on the nine members of the MPC interacting and reaching agreement, the behavioural approach is arguably less appropriate when combining forecasts across different, even rival, institutions.

By contrast mathematical approaches combine the information across experts by using some rule or model. Early work focused on combination rules that satisfied certain properties or axioms. Two common axiomatic approaches are the "linear opinion pool" and the "logarithmic opinion pool". Let us consider them in turn.

Consider N forecasts made by expert i $(i = 1, ..., N)$ of a variable y, assumed to be real-valued. These N forecasts, denoted g_i, are density forecasts, assumed continuous. The linear opinion pool takes a weighted linear combination of the experts' probabilities; the combined density is defined as the finite mixture:

$$p(y) = \sum_{i=1}^{N} w_i g_i(y), \tag{1}$$

where w_i are a set of non-negative weights that sum to unity. This combined density satisfies certain properties such as the "unanimity" property (if all experts agree on a probability then the combined probability agrees also); for further discussion, and consideration of other properties see Genest & Zidek (1986) and Clemen & Winkler (1999).[3] Further descriptive properties of mixture distributions are summarised in Everitt & Hand (1981).

Inspection of (1) also reveals that taking a weighted linear combination of the experts' densities can generate a combined density with characteristics quite distinct from those of the experts. For example, if all the experts' densities are normal, but with different means and variances, then the combined density will be mixture normal.[4] Mixture normal distributions can have heavier tails than normal distributions, and can therefore potentially accommodate skewness and kurtosis. If the true (population) density is non-normal we can begin to appreciate why combining individual density forecasts, that are normal, may mitigate misspecification of the individual densities.

The logarithmic opinion pool is defined as:

$$p(y) = k \prod_{i=1}^{N} g_i(y)^{w_i}, \tag{2}$$

where k is a normalising constant. When $w_i = (1/N)$, $p(y)$ is proportional to the geometric mean of the experts' distributions.

2.1 The Bayesian approach

More recently the Bayesian approach has received attention. The experts' densities are combined by a "decision maker" who views them as data. Following Morris (1974, 1977),

[3] The key practical issue is how to determine w_i. Hall & Mitchell (2004*b*) suggest a data-driven approach. Alternatively, model averaging has been suggested. Granger & Jeon (2004) suggest a thick-modelling approach, based on trimming to eliminate the $k\%$ worst performing forecasts and then taking a simple average of the remaining forecasts. Bayesian model averaging has been suggested also; e.g. see Garratt et al. (2003) and Pesaran & Zaffaroni (2004). This provides a means of weighting alternative model based density forecasts according to their respective posterior probabilities. These probabilities are often proxied by some measure of the relative statistical in-sample fit of the model, as measured by, say, the Schwartz information criterion. This approach cannot combine density forecasts that do not rely on estimation of some statistical model. Most simply, equal weights, $w_i = 1/N$, have been advocated; e.g. see Hendry & Clements (2004).

[4] For related analysis in the context of Bayesian model averaging see Pesaran & Zaffaroni (2004).

Bayes' Theorem is used to update the decision maker's prior distribution, $h(y)$, in the light of these data from the experts that takes the form of the joint density, or likelihood, derived from their N densities. The difficulty faced by the decision maker is deciding upon the form of the likelihood function; the likelihood must capture the bias and precision of the experts' densities as well as their dependence. In this paper we follow the spirit of the popular approach of Winkler (1981).

Consider the mean, m_i, and variance, v_i, of expert i's distribution:

$$m_i = \int_{-\infty}^{\infty} y g_i(y) dy, \tag{3}$$

$$v_i = \int_{-\infty}^{\infty} (y - m_i)^2 g_i(y) dy, \tag{4}$$

$(i = 1, ..., N)$. It is assumed that m_i and v_i are the information reported by expert i available to the decision-maker.[5] In this paper we focus on the case when expert i reports these two moments only; they are assumed to summarise their density $g_i(y)$. Appendix A summarises how the approach can be extended to higher moments.

m_i is considered as the point forecast of y, and v_i as a measure of uncertainty.[6] The forecasting error for expert i is:

$$s_i = m_i - y, \tag{5}$$

and expert i's density for s_i takes the form $g_i(m_i - y)$. Let $\mathbf{m} = (m_1, ..., m_N)'$ denote the N-vector of reported means and $\mathbf{v} = (v_1, ..., v_N)'$ the N-vector of reported variances. Let the N-stacked vector $\mathbf{s} = (s_1, s_2, ..., s_N)'$ have covariance matrix $\mathbf{\Sigma}$. Note that any dependence among experts is modelled here by the dependence among their forecasting errors.[7]

The decision maker aggregates the N experts' forecasts as follows. *Via* Bayes' Theorem the consensus distribution is given by the posterior distribution:[8]

$$\begin{aligned} h(y|\mathbf{m}, \mathbf{v}) &= \frac{h(\mathbf{m}|y).h(\mathbf{v}|\mathbf{m},y).h(y)}{h(\mathbf{m}, \mathbf{v})}, & (6) \\ &\propto h(\mathbf{m}|y).h(\mathbf{v}|\mathbf{m},y)h(y), & (7) \end{aligned}$$

so that, assuming (i) $h(y)$ is uniform and (ii) $h(\mathbf{v}|\mathbf{m},y)$ does not depend on $\mathbf{m}$ or y, an assumption which is later relaxed in Section 3.1 and 4.1,

$$h(y|\mathbf{m}, \mathbf{v}) \propto h(\mathbf{m}|y), \tag{8}$$

[5] Alternatively, we could think of m_i and v_i as the median and interquartile range.

[6] Given non-quadratic loss functions it is well known that the "optimal" central estimate may not be the mean.

[7] Alternatively, copulas could be used; see Jouini & Clemen (1996). They allow the consideration of more general multivariate distributions than the normal.

[8] Note we assume $h(\mathbf{m}|y, \mathbf{v}) = h(\mathbf{m}|y)$.

where h denotes a probability density.

We assume $\mathbf{s} = (m_1 - y, ..., m_N - y) \sim N(\mathbf{0}, \mathbf{\Sigma})$; i.e. $h(\mathbf{m}|y) \propto \exp\left[-\frac{1}{2}(\mathbf{m} - y\mathbf{e})'\mathbf{\Sigma}^{-1}(\mathbf{m} - y\mathbf{e})\right]$, where $\mathbf{e}$ is an N-vector of ones. This assumption implies that the means reported by the experts offer unbiased estimates for y. Calibration means this is plausible - we should expect forecasters *via* model improvement etc. to eliminate any systematic bias that their forecasts may have. Alternatively one could allow for biases in experts' announcements at the cost of introducing extra parameters to model the nature of the biases; see Lindley (1983). Normality of the errors s_i does not imply that the experts' densities g_i are themselves normal.

The posterior density for y, the combined density, is then given by:

$$h(y|\mathbf{m}) \propto \phi\left[(y - m^*)/\sigma_m^*\right], \tag{9}$$

where ϕ is the standard normal density function and

$$\begin{aligned} m^* &= \mathbf{e}'\mathbf{\Sigma}^{-1}\mathbf{m}/\mathbf{e}'\mathbf{\Sigma}^{-1}\mathbf{e}, & (10) \\ \sigma_m^{*2} &= 1/\mathbf{e}'\mathbf{\Sigma}^{-1}\mathbf{e}. & (11) \end{aligned}$$

m^* can also be interpreted as the maximum likelihood (ML) estimator.[9] In this sense m^* is optimal. The weights, m^*, are familiar to economists, even if this particular derivation is not.

This familiarity is seen easily by considering the simple case of combining 2 forecasts (when we can easily analytically invert the covariance matrix). Let the elements of $\mathbf{\Sigma} = (\sigma_{ij})$. Then (10) implies

$$m^* = \frac{(\sigma_{22} - \sigma_{12})m_1 + (\sigma_{11} - \sigma_{12})m_2}{\sigma_{11} + \sigma_{22} - 2\sigma_{12}}, \tag{12}$$

which is equivalent to both the minimum variance solution [see Bates & Granger (1969)] and the regression method [see Granger & Ramanathan (1984)]; also see Clements & Hendry (1998), pp. 229-230.[10] However, these two approaches do not require normality. The expression (12) is interesting as it clearly shows that weights can be negative.

At a practical level, to operationalise this approach to combination, the decision maker needs to know $\mathbf{\Sigma}$. Winkler suggests using the experts' reported variances on the main diagonal of $\mathbf{\Sigma}$ and then estimating the $N(N-1)/2$ non-diagonal elements; e.g. see Winkler (1981, p. 482 and p. 484). Typically estimation of these elements is based on historical data, namely the track-record of the N experts in predicting y. In fact, as we see below, the experts' reported variances should not be used to estimate $\mathbf{\Sigma}$. Rather the historical accuracy of the point estimates, which need not be the same, should be used. If $\mathbf{\Sigma}$ is

[9] The likelihood $h(\mathbf{m}|y)$ implies $\mathbf{m}$ are N random variables sampled from a normal distribution with common mean y and variance-covariance matrix $\mathbf{\Sigma}$. Treating $\mathbf{\Sigma}$ as fixed and known, differentiation of the likelihood with respect to y reveals that the ML estimator m^*: $\sqrt{T}(m^* - y) \overset{d}{\to} N(0, (\mathbf{e}'\mathbf{\Sigma}^{-1}\mathbf{e})^{-1})$. See also Halperin (1961).

[10] Nonlinear and time-varying combination methods can be considered too; e.g. see Deutsch et al. (1994) and Diebold & Pauly (1987).

unknown a prior can be adopted; see Winkler (1981). The combined forecasts then have a t-distribution; see also Lindley (1983).

σ_m^{*2} is the variance of the combined forecast. We see from (11) that $\sigma_m^{*2} \rightarrow 0$ as the number of forecasters N tends to infinity, even when they are making errors, so long as all of the $N(N-1)/2$ covariances are either zero or greater than zero. We might expect positive covariances when experts have access to common information sets etc.. In other words, σ_m^{*2} can tend to zero even when each expert remains uncertain about their forecasts ($v_i > 0$). In the following section we extend the focus of the Bayesian approach to higher moments than the mean. This is important when data are realisations from stochastic data generating processes (DGPs).

3 Forecasting the variance

Consistent with the spirit of the Bayesian paradigm introduced in Section 2.1, this section extends the focus to forecast other characteristics than the mean of the density, specifically the variance. The variance is an indicator of "uncertainty" or volatility. It is important to define precisely what we mean by uncertainty, and explain how the measure we focus on, average individual uncertainty, relates to alternative measures.

Let the time-series y_t be a realisation from the sequence of DGPs $f_t(y_t)$ $(t = 1, ..., T)$.[11] When the DGP is stochastic we consider realisations from the DGP at a given point in time, t, to have an associated variance, denoted σ_t^2. σ_t^2 can vary across time, reflecting the fact that for many time-series historically we can distinguish periods of varying volatility. When $\sigma_t^2 = 0$, y_t is a realisation from a deterministic process. $f_t(y_t)$ might be some linear or nonlinear, static or dynamic, process that is subject to random disturbances (shocks) across time. This is consistent with many explanations of the behaviour of aggregate inflation or output growth. Note that the process $f_t(y_t)$ $(t = 1, ..., T)$ may, or may not, be covariance-stationary.

Define μ_t and σ_t^2 as the mean and variance of the distribution of y_t:

$$\mu_t = E(y_t) = \int_{-\infty}^{\infty} y_t f_t(y_t) dy_t, \tag{13}$$

$$\sigma_t^2 = Var(y_t) = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_t(y_t) dy_t, \tag{14}$$

and let m_{it} and v_{it}, see (3) and (4), denote the mean and variance of the density forecasts announced or reported by expert i $(i = 1, ..., N)$ at time t $(t = 1, ..., T)$. Consistent with Section 2, it is assumed $E(m_{it}) = \mu_t$. Of course, v_i can be greater than zero even if expert i correctly forecasts the mean. Again let us assume for now that the density forecasts are

[11] For notational convenience we do not distinguish between random variables and their realisations.

summarised by their first two moments. No distributional assumptions need to be made about the density forecasts.

Decomposing expert i's point forecasting error $(y_t - m_{it})$:

$$\begin{aligned} s_{it} &= y_t - m_{it} = y_t - \mu_t + \mu_t - m_{it}, & (15) \\ s_{it} &= e_t + (\mu_t - m_{it}), & (16) \end{aligned}$$

and taking the conditional expectation of the average of the square of (16) across i, we see that the uncertainty around a (randomly drawn) individual forecast, $\overline{\sigma}_t^2$,

$$\overline{\sigma}_t^2 = \frac{1}{N}\sum_{i=1}^{N} s_{it}^2 = \sigma_t^2 + \sigma_{\mu t}^2, \tag{17}$$

equals aggregate uncertainty σ_t^2, explained by the stochastic nature of y_t, plus $\sigma_{\mu t}^2$, where $\sigma_{\mu t}^2 = \frac{1}{N}\sum_{i=1}^{N}(\mu_t - m_{it})^2$.[12] $\sigma_{\mu t}^2$ reflects experts' uncertainty, disagreement, about the true mean.

We will relate $\overline{\sigma}_t^2$ to the averaged or mean variance of the individual distributions (at a given point in time), namely $E(v_{it})$ - the mean of the reported variances. This has been considered, for example by Zarnowitz & Lambros (1987), as an explicit measure of uncertainty. In Section 3.1 we turn to how one might optimally combine these variances, rather than simply take unweighted averages. This is analogous to our discussion of the mean; experts' announced variances are combined according to their accuracy.

Alternatively, see (11), uncertainty can be defined with respect to the mean (consensus) forecast m_t^*, rather than the individual forecast m_{it}.[13] This measure of uncertainty, σ_m^{*2}, must have a lower variance than a randomly drawn individual forecast. For equal weights this is seen as follows. $Var\left(y_t - m_t^*\right) = \frac{1}{N^2}Var\left(\sum_{i=1}^{N} s_{it}\right) = \frac{1}{N}\overline{\sigma}_t^2 + \frac{(N-1)}{N}\overline{\rho}$, where $\overline{\rho}$ is the average covariance between all pairs of forecast errors. Since $\overline{\sigma}_t^2 \geq \overline{\rho}$ it follows that $Var\left(y_t - m_t^*\right) \leq \overline{\sigma}_t^2$; see also Batchelor & Dua (1995).

Average individual uncertainty $\overline{\sigma}_t^2$ also relates to the uncertainty of the aggregate distribution, derived *via* the linear opinion pool. The mean and variance of (1), now including time subscripts, are given by:[14]

[12] The cross-product term disappears on the assumption that $E(2e_t(\mu_t - m_{it})) = 0$, implying the true disturbance e_t is uncorrelated with experts' individual forecasts.

[13] Both the individual and mean measures of uncertainty rely on experts reporting information about their probability distributions, namely m_i and v_i. In the absence of such information, time-series methods (ARCH models etc.) are used widely to measure uncertainty; for an empirical comparison of alternative measures of uncertainty using the Survey of Professional Forecasters see Giordani & Söderlind (2003).

[14] Related expressions decomposing the aggregate density (1), based on the 'law of conditional variances', are seen in Giordani & Söderlind (2003). This law states that for the random variables y_t and i: $V(y_t) = E[V(y_t|i)] + V[E(y_t|i)]$.

$$
\begin{aligned}
E\left[p_t(y_t)\right] &= m_t^* = \sum_{i=1}^{N} w_{it} m_{it}, && (18) \\
V\left[p_t(y_t)\right] &= \sum_{i=1}^{N} w_{it} v_{it} + \sum_{i=1}^{N} w_{it} \left\{m_{it} - m_t^*\right\}^2 . && (19)
\end{aligned}
$$

(19) indicates that the variance of the aggregate distribution equals average individual uncertainty plus disagreement.[15]

3.1 Combining the variances

Consider N sample-based estimates of the population variance $\overline{\sigma}_t^2$, denoted s_{it}^2, where:

$$s_{it}^2 = (y_t - m_{it})^2; \; (i = 1, ..., N). \tag{20}$$

These N alternative estimates of the uncertainty measure $\overline{\sigma}_t^2$ differ across experts, i, when expert's point estimates m_{it} differ.

We assume that $E(s_{it}^2) = \overline{\sigma}_t^2$ for all i; so, for example, both s_{1t}^2 and s_{2t}^2 are unbiased estimates of the true variance $\overline{\sigma}_t^2$.[16] Of course, this is not inconsistent with there being statistical evidence for bias between s_{1t}^2 and s_{2t}^2 in finite samples. Then consider the difference between s_{it}^2 and the experts' reported variances v_{it}:

$$v_{it} - s_{it}^2 = u_{it}. \tag{21}$$

It is further assumed that $E(v_{it}) = \overline{\sigma}_t^2$, $\forall_i$; i.e. experts' reported forecasts of the variance offer unbiased estimates of the true variance.[17] Again in finite samples this may not hold, but we assume that the forecasters calibrate their forecasts to ensure unbiasedness of their reported variances in the long-run.

This assumption implies that the expert tries to forecast not just the variance of y_t, σ_t^2, but the variance allowing for its uncertainty about the mean forecast, $\sigma_{\mu t}^2$ (i.e. the expert adds in the extra uncertainty $\sigma_{\mu t}^2$ due to the fact that the mean of their distribution may not equal the true mean, in other words $(\mu_t - m_{it})$ may not equal zero). This is conceptually important; we are saying, for example, that the variance of the Bank of England's fan-chart seeks to forecast not just aggregate uncertainty, σ_t^2, but the uncertainty associated with the fact that the Bank has estimated the mean of its fan-chart.

[15] For further discussion of the relationship, if any, between dispersion/disagreement and individual uncertainty see Bomberger (1996).

[16] This assumption is made commonly when evaluating conditional volatility forecasts; see Taylor (1999). Since volatility, the outturn, is not observed directly the square of the observed error term is used as a proxy. This is not unreasonable. For example, if we consider the ARCH process $s_t = \sqrt{\overline{\sigma}_t}\xi_t$ where $\{\xi_t\}$ is an *i.i.d.* sequence with mean zero and unit variance and $\overline{\sigma}_t$ denotes latent volatility (that evolves according to some process), then if the model is correctly specified $E(s_t^2|t-1) = \overline{\sigma}_t^2$; see also Andersen & Bollerslev (1998).

[17] An equivalent assumption has already been made about the mean; $\mathbf{s} \sim N(\mathbf{0}, \mathbf{\Sigma})$.

Given these assumptions it follows that $E(u_{it}) = 0$, $\forall_i$. We further assume that the u_{it}'s are jointly normally distributed with variance-covariance matrix $\mathbf{\Omega}_t$. Although, of course, any variance can only be greater than or equal to zero, no such restriction need hold on the difference between two variance terms. Then we can use $\mathbf{\Omega}_t$ to find the consensus distribution for the variances following our consideration of the mean. This is achieved as follows.

Assuming $\mathbf{u}_t = (u_{1t}, ..., u_{Nt}) \sim N(\mathbf{0}, \mathbf{\Omega}_t)$, i.e. $h(\mathbf{v}_t | \mathbf{m}_t, y_t) \propto \exp\left[-\frac{1}{2}\mathbf{u}_t{}'\mathbf{\Omega}_t^{-1}\mathbf{u}_t\right] = \exp\left[-\frac{1}{2}\left(\mathbf{v}_t - \overline{\sigma}_t^2\mathbf{e}\right)'\mathbf{\Omega}_t^{-1}\left(\mathbf{v}_t - \overline{\sigma}_t^2\mathbf{e}\right)\right]$, where $\mathbf{v}_t = (v_{1t}, ..., v_{Nt})$ and $\overline{\sigma}_t^2\mathbf{e} = (y_t\mathbf{e} - \mathbf{m}_t)^2$, the posterior density for $\overline{\sigma}_t^2$ conditional on the experts' reported variances is:

$$h(\overline{\sigma}_t^2|\mathbf{v}) \propto \phi\left[(\overline{\sigma}_t^2 - v_t^*)/\sigma_v^*\right], \tag{22}$$

where

$$\begin{aligned} v_t^* &= \mathbf{e}'\mathbf{\Omega}_t^{-1}\mathbf{v}_t/\mathbf{e}'\mathbf{\Omega}_t^{-1}\mathbf{e}, & (23)\\ \sigma_{vt}^{*2} &= 1/\mathbf{e}'\mathbf{\Omega}_t^{-1}\mathbf{e}. & (24)\end{aligned}$$

v_t^* is the "optimally" combined variance, in the sense that it is the ML estimator of $\overline{\sigma}_t^2$.[18]

Again at a practical level the decision maker needs to know $\mathbf{\Omega}_t$. $\mathbf{\Omega}_t$ can be estimated based on historical data, namely the track-record of the N experts in predicting s_{it}^2. Let us suppose we have information on their track-record over the last Q periods, $\mathbf{u}_t$ ($t = 1, ..., Q$; where $Q > N$). A convenient estimator for $\mathbf{\Omega}_t$ is then given by $\frac{1}{Q}\sum_{t=1}^{Q}\mathbf{u}_t\mathbf{u}_t'$. Consistency of this estimator requires ergodicity of $\{\mathbf{u}_t\}$ in second moments so that:

$$\frac{1}{Q}\sum_{t=1}^{Q}\mathbf{u}_t\mathbf{u}_t' \xrightarrow{p} E(\mathbf{u}_t\mathbf{u}_t') = \mathbf{\Omega}_t, \tag{25}$$

where $\mathbf{\Omega}_t = \mathbf{\Omega}$ for all t.

3.2 Combining the mean having combined the variances

Let us re-consider the weights used to derive the combined mean forecast, m_t^*, that consistent with the discussion above are now time-varying. As indicated above, Winkler suggested estimating $\mathbf{\Sigma}_t$ based on using the experts' reported variances on the main diagonal and estimating the off diagonal elements based on the historical correlation between their errors in forecasting the mean. As intimated above, in fact it is sensible that the diagonal elements of $\mathbf{\Sigma}_t$ also should be estimated based on the historical accuracy of experts' estimates of y_t, namely their errors s_{it}, rather than their reported variances v_{it}.

[18] In theory, v_t^* can be negative; although all the elements of $\mathbf{v}_t$ are strictly positive the weights can be negative; see (12). This case is slightly pathological, however. It would require some experts to report quite odd variance forecasts, v_{it}. In any case, we can interpret v_t^* as a quasi-ML estimator.

Otherwise expert i's forecast may receive a large weight in the combined forecast just because expert i (perhaps incorrectly) reports a low variance v_{it}, irrespective of how well their estimates m_{it} of y_t may have performed historically.

However, not just should $\mathbf{\Sigma}_t$ be estimated solely based on the errors s_{it} but these errors should be weighted according to the volatility at the time, $\overline{\sigma}_t^2$. The point forecasting errors of a given expert, s_{it}, need to be weighted to reflect the fact that a small error at a time of considerable uncertainty (large $\overline{\sigma}_t^2$) is quite different from a small error at a time of relative certainty (small $\overline{\sigma}_t^2$). It is appropriate not to weight only when it is assumed, implicitly in Section 2.1, that $\{s_{it}\}$ is constant. In practice we may not wish to impose such an assumption; we might expect experts' errors s_{it} to depend on the volatility of y_t - e.g. currently experts may well be making historically small errors in forecasting inflation in the U.K., but we might expect this to be explained in part, at least, by the fact that inflation in the U.K. is currently stable compared with its historical behaviour and can be thought of as in some type of low volatility "regime".

We require $\{\mathbf{s}_t\}$, scaled by v_t^*, to be ergodic in second moments so that:

$$\frac{1}{Q}\sum_{t=1}^{Q}\frac{1}{\sqrt{v_t^*}}\mathbf{s}_t\mathbf{s}_t' \xrightarrow{p} E(\mathbf{s}_t\mathbf{s}_t') = \mathbf{\Sigma}_t, \tag{26}$$

where $\mathbf{\Sigma}_t = \mathbf{\Sigma}$ for all t. In contrast Section 2.1 implicitly assumed that $\{\mathbf{s}_t\}$ itself was ergodic in second moments; this is inconsistent with the view that experts' forecasting errors about the mean depend on the volatility of the process they are trying to forecast. Use of the estimator $\frac{1}{Q}\sum\limits_{t=1}^{Q}\mathbf{s}_t\mathbf{s}_t'$ will deliver biased estimates unless $\overline{\sigma}_t^2 = \overline{\sigma}^2$.

4 The combined density forecast: indirect and direct methods

As well as wishing to aggregate the N experts' predictions of the mean and variance separately, the decision maker may want an aggregated forecast of the density itself. It is important to distinguish two alternative processes by which the decision maker might seek to derive this combined forecast of the 'whole' density: an indirect and a direct process. The process chosen can affect the shape of the combined density.

The indirect approach derives the combined density by firstly combining the experts' moments. Following the discussion above, assume the decision maker has N experts' unbiased estimates of the (unknown) population's mean and variance. The decision maker then obtains optimal combined estimates of the mean and variance following Section 3. The decision maker can then translate these estimates into a combined density once she has made an assumption about the nature of the population density. For example, suppose the decision maker believes the population density to be normal. Then the combined density is $N(m_t^*, v_t^*)$.[19] It should be noted that the normality assumption for the pop-

[19] Similarly Hendry & Clements (2004) in their Monte-Carlo experiments consider a combined density

ulation density is not completely *ad hoc*. First, as ML estimates, both m_t^* and v_t^* are asymptotically normally distributed. Secondly, since we have assumed that all experts' forecasts have the same expectation of the mean and variance, if all experts' forecast densities are themselves normal then the combined density will be normal too. It is only when the individual forecast densities have different means and/or variances that the combined density will be mixture normal. Also if the individual densities are not normal then the combined density will not be normal.

The second approach directly combines the experts' densities. We have already seen one simple method of doing this, the linear opinion pool; see (1). Here we suggest an alternative, consistent with the Bayesian paradigm introduced above. We define the combined density forecast as the posterior distribution of the variable we are seeking to forecast conditional on the information supplied by the experts. This information is again assumed to be summarised by the moments announced by the experts. We continue to assume, without loss of generality, that the decision maker only has information on the forecast means and variances announced by the experts.[20] In contrast to the indirect approach, the decision maker no longer needs to take a view regarding the shape of the population distribution. As we see below, the direct approach can deliver a combined density that dramatically departs from normality. It follows that the direct approach is inappropriate if the decision maker knows the population distribution is normal since, in contrast to using the indirect approach, the combined density will not in general be normal. Let us consider this second approach in more detail.

forecast that is normal with mean equal to m_t^* and variance estimated from the historical (in their case in-sample) mean errors, defined as $(y_t - m_t^*)$. This approach to measuring the variance, as discussed below in the context of NIESR forecasts, runs into problems at times of change; the past error variance is then a poor indicator of future error variances. However, there may be an attraction to assuming a known distributional form for the combined density. Consider the case where the time-series $\{y_t\}$ is generated according to $y_t = w_1 x_{1t} + (1-w_1)x_{2t}$, where $\{x_{1t}\}$ and $\{x_{2t}\}$ are othogonal normally distributed processes. Then if we consider two misspecified forecasting models, the first of which considers only x_{1t} and the second only x_{2t}, then the combined density based on directly combining the two misspecified density forecasts, say using (1), will not be normal despite the fact that $\{y_t\}$ is. As discussed, this remains so even when the two misspecified density forecasts are normal, assuming the two forecasts are not identical, unless $w_1 = 0$ or 1.

[20] However, the analysis can be straightforwardly extended to higher moments. In so doing we can condition on the information provided when one expert, like the Bank of England when forecasting inflation for example, views the density as two-piece normal. We might view the two-piece normal assumption of the Bank as relevant information which we wish to condition on when deriving the combined density forecast. We can condition on this by noting that the two-piece normal distribution requires us to consider higher moments than the first and second. This can be done by introducing a third moment, skewness. See Appendix A.

4.1 Directly combining experts' density forecasts: a Bayesian approach

The form of the combined density follows directly from Bayes' Theorem, see (6), with time subscripts now added:[21]

$$h(y_t|\mathbf{m}_t, \mathbf{v}_t) \propto h(\mathbf{m}_t|y_t).h(\mathbf{v}_t|\mathbf{m}_t,y_t).h(y_t), \tag{27}$$

where $h(\mathbf{v}_t|\mathbf{m}_t,y_t)$ is such that $\mathbf{v}_t$ is allowed to depend on $\mathbf{m}_t$ and y_t; this is consistent, for example, with the view, often supported empirically, that there is a relationship between the level and variability of inflation; see Demetriades (1989). Since we retain the assumption that $h(y_t)$ is uniform, it follows that:

$$h(y_t|\mathbf{m}_t, \mathbf{v}_t) \propto h(\mathbf{m}_t|y_t).h(\mathbf{v}_t|\mathbf{m}_t,y_t). \tag{28}$$

Since we know the form of $h(\mathbf{m}_t|y_t)$ and $h(\mathbf{v}_t|\mathbf{m}_t,y_t)$ then:

$$\begin{aligned} h(y_t|\mathbf{m}_t, \mathbf{v}_t) &\propto \exp\left[-\frac{1}{2}(\mathbf{m}_t - y_t\mathbf{e})'\boldsymbol{\Sigma}_t^{-1}(\mathbf{m}_t - y_t\mathbf{e}) - \frac{1}{2}\mathbf{u}_t'\boldsymbol{\Omega}_t^{-1}\mathbf{u}_t\right] \text{ or} &(29) \\ h(y_t|\mathbf{m}_t, \mathbf{v}_t) &\propto \exp\left[\begin{array}{c} -\frac{1}{2}(\mathbf{m}_t - y_t\mathbf{e})'\boldsymbol{\Sigma}_t^{-1}(\mathbf{m}_t - y_t\mathbf{e}) - \\ \frac{1}{2}\left(\mathbf{v}_t - (y_t\mathbf{e} - \mathbf{m}_t)^2\right)'\boldsymbol{\Omega}_t^{-1}\left(\mathbf{v}_t - (y_t\mathbf{e} - \mathbf{m}_t)^2\right)\end{array}\right] &(30)\end{aligned}$$

where above we saw that $\overline{\sigma}_t^2\mathbf{e} = E(y_t\mathbf{e} - \mathbf{m}_t)^2$.[22]

The individual densities are combined in (29) according to the reliability of their first two moments. Note that the assumption that the moments of the individual densities offer unbiased estimates of the population moments is maintained.[23]

Inspection of (29) reveals that the mean and variance of the combined density, derived by differentiation of (29) with respect to y_t, no longer equal m_t^* and v_t^*. While the product of two normal (scalar) random variables with zero means and different variances has a known distributional form, being based on the delta and modified Bessel functions, in our multivariate case we simply illustrate the properties of (29) *via* some simple numerical simulations. These are instructive in demonstrating that the combined density using the direct method can take on various shapes.

In our experiments attention is restricted to two density forecasts; denote the reported means provided by the two experts m_1 and m_2, and their reported variances v_1 and v_2.

[21] Appendix A notes how the decision maker could condition on higher moments announced by the experts to derive the posterior distribution of y conditional on not just the first two moments announced by the experts but also, say, reported skewness.

[22] Note that the reciprocal of the normalising constant in (29) is $\int h(\mathbf{m}_t|y_t)h(\mathbf{v}_t|\mathbf{m}_t, y_t)dy = \int \exp\left[-\frac{1}{2}(\mathbf{m}_t - y_t\mathbf{e})'\boldsymbol{\Sigma}_t^{-1}(\mathbf{m}_t - y_t\mathbf{e}) - \frac{1}{2}\mathbf{u}_t'\boldsymbol{\Omega}_t^{-1}\mathbf{u}_t\right] dy$.

[23] While we feel that in practice this is a reasonable assumption, see Section 2.1, it is useful to consider combination of two sample based estimates of the population density that may not be unbiased. We distinguish between the following cases: (i) when both are biased and (ii) when one is unbiased and one is biased. In case (i) density combination may help. In case (ii) combination will deliver worse results.

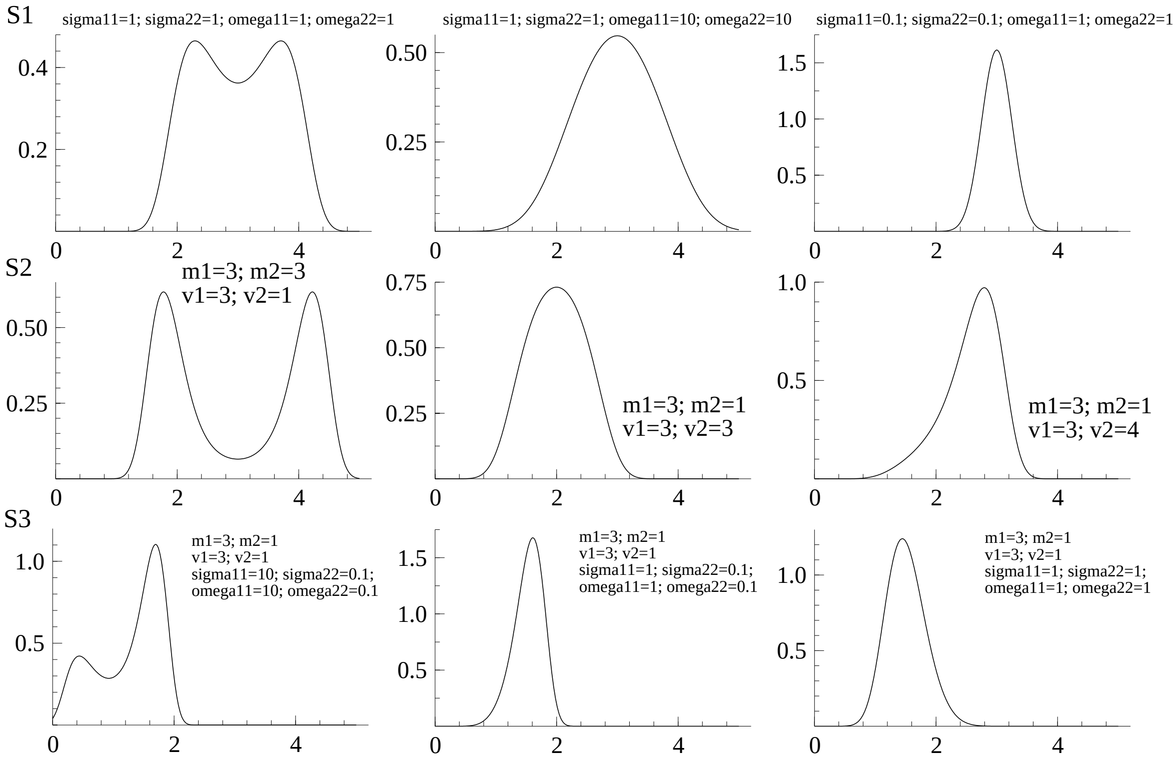

Figure 1: Exploring the form of the combined density forecast

We carry out three sets of illustrations. The first set considers the situation where both experts report the same mean and variance but the weight the decision maker places on the two experts' opinions varies. The second set examines when the experts disagree on the mean and variance and the decision maker places equal weight on their advice. The third set examines when the experts disagree on the mean and variance and the weight the decision maker places on the two expert varies too. The results are summarised in three panels (S1, S2 and S3) in Figure 1.

In S1 we explore the effect $\mathbf{\Sigma}_t$ and $\mathbf{\Omega}_t$ have on the shape of the combined density when the two forecasters agree on both the mean and the variance; $m_1 = m_2 = 3; v_1 = v_2 = 1$. $\mathbf{\Sigma}_t$ and $\mathbf{\Omega}_t$ are assumed diagonal; three sets of values for the elements on their main diagonals (denoted sigma11, sigma22, omega11 and omega22) are considered to reflect when the decision maker trusts their mean and variance forecasts equally and when they trust their mean predictions more than their variance predictions. Figure 1 shows that the combined density is, as we might expect, unimodal centered on the mean only when the unreliability associated with their mean forecasts is small compared with that associated with their variance forecasts. When the decision maker weights the means and variances similarly the combined density is bimodal. These shapes contrast the shape of the combined density that the indirect method would deliver; the indirect method forces a compromise; the combined density may then place a lot of weight in a region that neither of the experts thought likely.

In S2 we consider the situation where the decision maker trusts the experts' announcements equally; it is assumed $\mathbf{\Sigma}_t = \mathbf{\Omega}_t = \mathbf{I}_2$. Note this implies that the experts' forecasts of the mean and variance are unreliable (imperfect); they do not have zero variances. We see that the combined density, (29), takes on various shapes depending on $\{m_1, m_2, v_1, v_2\}$. Let us summarise three aspects of the figures in S2 : (i) when the experts report the same mean the combined density is symmetric, although bimodal in the case considered; (ii) when the reported means differ but the reported variances are the same the combined density is again symmetric but unimodal, although with far less weight in the tails than the normal distribution; (iii) when both the reported means and variances differ the combined density becomes asymmetric.

In S3 we consider the case when the decision maker believes one expert is more reliable than the other and the experts disagree about the mean and variance. We differentiate between three cases according to how much more accurate expert 2 is than expert 1. We see that as the decision maker trusts expert 2 more and more, the combined density moves increasingly in-line with this expert's announcements; the combined density becomes increasingly skewed to the left reflecting the lower mean (and variance) reported by expert 2.

5 Evaluation of density forecasts

While there exist well established techniques for the *ex post* evaluation of point forecasts, often based around the root mean squared error of the forecast relative to the subsequent outturn, only recently has the *ex post* evaluation of density forecasts attracted much attention. Currently, following Diebold et al. (1998), the most widespread approach is to evaluate density forecasts statistically using the probability integral transform, itself a well-established result.[24] Diebold et al. (1998) popularised the idea of evaluating a sample of density forecasts based on the idea that a density forecast can be considered "optimal" if the model for the density is correctly specified. One can then evaluate forecasts without the need to specify a loss function. This is attractive as it is often hard to define an appropriate general (economic) loss function. Alternatively, we could focus on a particular region of the density, such as the probability of inflation being in its target range; see Clements (2004).

[24] This methodology seeks to obtain the most "accurate" density forecast, in a statistical sense. It can be contrasted with economic approaches to evaluating forecasts that evaluate forecasts in terms of their implied economic value, which derives from postulating a specific (economic) loss function; see Granger & Pesaran (2000) and Clements (2004). Other work has evaluated density forecasts using scoring rules; e.g. see Giacomini (2002).

A sequence of estimated density forecasts, $\{p_t(y_t)\}_{t=1}^T$, for the realisations of the process $\{y_t\}_{t=1}^T$, coincides with the true densities $\{f_t(y_t)\}_{t=1}^T$ when the sequence of probability integral transforms (pit's), z_t, is independently and identically distributed (*i.i.d.*) with a uniform distribution, U(0,1), where,

$$z_t = \int_{-\infty}^{y_t} p_t(u)du; \; (t = 1, ..., T). \tag{31}$$

Density forecasts are optimal and capture all aspects of the distribution of y_t only when the $\{z_t\}$ are both *i.i.d.* and U(0,1). By taking the inverse normal cumulative density function (CDF) transformation of $\{z_t\}$ to give, say, $\{z_t^*\}$ the test for uniformity can be considered equivalent to one for normality on $\{z_t^*\}$; see Berkowitz (2001). This is useful as normality tests are widely seen to be more powerful than uniformity tests. However, testing is complicated by the fact that the impact of dependence on the tests for uniformity/normality is unknown, as is the impact of non-uniformity/normality on tests for dependence.

In the empirical application below we abstract from any perceived uncertainties regarding the appropriateness of specific tests; in total we consider ten alternative statistical tests for *i.i.d.* uniformity/normality. These tests have all been used, or proposed, for evaluation of density forecasts; see Appendix B for a review. The tests vary in what they test, as well as in terms of their properties/assumptions; some of the tests focus on the distribution, others focus on independence. We do supplement these with consideration of both a joint and portmanteau test for *i.i.d.* uniformity, namely the Hong and Thompson test, respectively. By considering a variety of tests we simply hope to provide robust evaluation; we look for consensus across the ten tests.

6 Combining the Bank of England and NIESR density forecasts of U.K. inflation

The application serves not just to illustrate the use of density forecast combination but is also an area of considerable interest *per se.* Forecasting inflation is of pivotal importance for central banks in an era of inflation targeting. We focus on quarterly forecasts of one-year ahead RPIX inflation (RPI excluding mortgage payments: ONS code CHMK), the principal monetary policy target over the sample period. The year ahead forecasts correspond to a five quarter ahead horizon. Strictly the forecasts are conditional on the assumption that nominal interest rates remain constant throughout the forecast period; however, following previous analysis, we regard the forecasts as unconditional on the (plausible) assumption that inflation does not react within a year to changes in interest rates; see Clements (2004) and Wallis (2004). As discussed by Hendry & Clements (2004), in any application the reasons for success or failure of combination can be multi-faceted. Our application is intended to illustrate the use of the proposed methods of combination, rather than explain why combination may, or may not, help. Moreover, just as there appears to be an empirical consensus that combination of point (mean) forecasts often works, this application can serve as an early indicator of whether combining density forecasts will prove as helpful.

6.1 Bank of England density forecasts

We consider the quarterly sequence of one-year ahead inflation forecasts published by the Bank. These forecasts are published in the *Inflation Report* in February, May, August and November, which we correspond to q1, q2, q3 and q4, respectively. The Bank of England has published density forecasts for RPIX inflation from 1993q1. Up until 1995q4 these took the form of charts showing the central projection, together with an estimate of uncertainty based on the historical mean absolute error. At this stage the Bank did not quantify a skew so that the mode, median and mean projections are equal; the density forecast is (implicitly) assumed normal.[25] From 1996q1 the Bank has published the so-called "fan" chart, that allows for skewness. From 1997q3 these charts have been based on the deliberations of the MPC.[26] The final projection for RPIX inflation, prior to the new target for inflation announced by the Chancellor in December 2003, was published in the February 2004 *Inflation Report.*

The fan chart is based analytically on the two-piece normal distribution; for details see Wallis (2004). The Bank publishes the parameter values underlying each published fan chart by supplying *via* its spreadsheets information on the following five statistics: the mode (μ^d), median, mean ($E(Y)$), uncertainty (σ) and skew. The uncertainty statistic is a parameter of the two-piece normal distribution; see Wallis (2004), Box A, for details - following Wallis note that we correct earlier confusion about what the uncertainty measure published by the Bank represents. The skew statistic is defined as the mean minus the mode. Given these parameters, following Wallis, we can back-out the standard errors σ_1 and σ_2 of the two normal distributions on which the two-piece normal distribution is based. Then [see also Clements (2004)] we can compute the $\{z\}$ as follows:

$$P(Y < y) = \left\{ \begin{array}{l} \frac{2\sigma_1}{\sigma_1+\sigma_2}\Phi\left(\frac{y-\mu^d}{\sigma_1}\right) \text{ for } y < \mu^d \\ \left(\frac{\sigma_1-\sigma_2}{\sigma_1+\sigma_2}\right) + \frac{2\sigma_2}{\sigma_1+\sigma_2}\Phi\left(\frac{y-\mu^d}{\sigma_2}\right) \text{ for } y > \mu^d \end{array} \right\}, \tag{32}$$

where Φ is the standard normal CDF. Using actual inflation data up to 2004q2 this gives us 42 density forecasts in total to compare with the subsequent outturn for RPIX inflation from 1994q1-2004q2.

[25] In 1995q1 uncertainty is not recorded; we simply assume the value from the previous *Inflation Report.* This seems a reasonable assumption given that uncertainty is being quantified based on historical RMSE which should not be expected, at least in large-samples, to change much quarter from quarter.

[26] The density forecasts from 1993q1-1997q2 are available at: http://www.bankofengland.co.uk/inflationreport/historicalforecastdata.xls. From 1997q3 they are available at http://www.bankofengland.co.uk/inflationreport/rpixinternet.xls.

6.2 NIESR density forecasts

We consider the quarterly forecasts of annual RPIX inflation as published in the *National Institute Economic Review.*[27] Since 1992q3 NIESR has, in a sense implicitly, published probability forecasts for inflation, in that the *Review* contained the table "Average Absolute Errors". This table indicated the historical accuracy of NIESR forecasts by reporting the mean absolute error.[28] Since 1996q1 NIESR has explicitly published probability forecasts for inflation. These have taken the form of (tabular) histograms, indicating the probability of inflation falling within a band, although these bands have changed periodically. These probability forecasts are centered on the point forecast published in the *Review.* This point forecast is produced by NiGEM, a large-scale macroeconometric model, subject as is usual in models of this type to the judgement of the forecasters. In deriving the density forecasts, normality is assumed. This is because earlier work that analysed the historical errors (from 1984-1995) made in forecasting RPI inflation could not reject normality; nor indeed could they reject unbiasedness (in fact rationality); see Poulizac et al. (1996). The variance of the density forecast is then set equal to the variance of the historical forecast error.[29] Given the backward looking and mechanistic nature to this method of determining the variance, it is important, as we see in Section 6.3, what historical sample period is chosen to estimate the variance.

The *Review* focuses on forecasting inflation in the fourth quarter of the current year and the fourth quarter of the next year; therefore only the q4 publication offers a one-year head forecast.[30] While we can extract from back-issues of the *Review* one-year ahead point forecasts for the other quarters, published uncertainty estimates are only available for q4. Therefore, we have to make an assumption in order to infer uncertainty estimates for the other quarters.[31] For further analysis of NIESR density forecasts see Hall & Mitchell

[27] The Review is currently published in January, April, July and October. Prior to 1996 the publication timetable was slighly different. In any case we refer to the four publications of the Review each year as q1, q2, q3 and q4. Given our interest in one-year ahead forecasts it does not seem unreasonable to ignore these changes to the publication timetable since the information set is little different and there is still one year's worth of shocks.

[28] The mean absolute error (or deviation about zero) is defined as $m_d = \int |y| \, f(y)dy$. For the standard normal distribution where $f(y) = \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}y^2}$, then $m_d = \int_{-\infty}^{\infty} |y| \frac{1}{\sqrt{2\pi}} \exp^{-\frac{1}{2}y^2} dy = \sqrt{\frac{2}{\pi}} \int_{0}^{\infty} y \exp^{-\frac{1}{2}y^2} dy = \sqrt{\frac{2}{\pi}} = 0.79788$. Accordingly, assuming normality, a 58% confidence interval around the point forecasts corresponds to the point estimate $\pm m_d$.

[29] Past forecast errors are commonly used as a practical way of forecasting future errors; e.g. see Wallis (1989), pp. 55-56.

[30] Consistent with NIESR's emphasis on inflation forecasts for Q4, H.M. Treasury in its regular comparison of independent forecasts focuses on Q4 when considering inflation; e.g. see Treasury (2004)

[31] In contrast, previous work evaluating NIESR density forecasts, notably Wallis (2004), has analysed only the q4 forecasts. Naturally this leads to a very small available sample. However the following caveat should be made when evaluating, as in this paper, NIESR density forecasts for all four quarters. Since NIESR has only (explicitly at least) published density forecasts of inflation for the fourth quarter it is in this sense unfair to evaluate NIESR on its performance for quarters q1-q3. Nevertheless, we feel that the measures of uncertainty used to calibrate the density forecasts in q1-q3 follow closely the spirit of

(2004*a*).

Here we simply assume the density forecast is normal with standard deviation equal across the four quarters in a year. This assumption is sensible if we believe NIESR only re-calibrated their forecast variances once a year. More specifically, from 2000q4 we use the standard deviation explicitly published in q4 of the previous year. From 1996q1, following Wallis (2004), the implied values for the standard deviation are backed out of the published histogram. From 1993q1-1995q4 the standard deviation is set at 2. This seems reasonable for the following reasons: (i) this was the value in 1996q1; (ii) analysis of the historical forecasting errors for forecasts made in q1, q2, q3 and q4 suggests a standard deviation around two, except for forecasts made in q2; (iii) the mean absolute error for RPI (not RPIX) presented in "Average Absolute Errors" was unchanged in these three years suggesting that NIESR viewed uncertainty to have remained constant.

Table 1 summarises the properties of the Bank of England and NIESR density forecasts, while Figures 2-4 provide a visual impression. Interestingly, these figures clearly show that for much of the sample-period NIESR placed far more weight in the tails of their density forecast than the Bank of England.

6.3 Combination: In-sample and Recursive Out-of-Sample Results

We compare the performance of Bank of England, NIESR and combined density forecasts both in-sample and using recursive out-of-sample experiments, designed to mimic 'real-time' application of the proposed density forecast combination techniques. In-sample we estimate $\mathbf{\Sigma}_t$ and $\mathbf{\Omega}_t$ using all of the 42 observations; these are then used to construct combined density forecasts using both the indirect and direct methods. The out-of-sample analysis is designed to simulate whether in practice, in real-time, the decision maker could have pooled the Bank of England and NIESR density forecasts to obtain 'better' forecasts. Accordingly, we split the sample period in two. From 1997q3 recursively, quarter by quarter, we re-estimate $\mathbf{\Sigma}_t$ and $\mathbf{\Omega}_t$ using data available up to period $(t-5)$. This acknowledges the fact that one has to wait five quarters to evaluate the performance of a given (year-ahead) forecast. These recursively updated estimates for $\mathbf{\Sigma}_t$ and $\mathbf{\Omega}_t$ are then used to produce a series of combined density forecasts from 1997q4 to 2004q2. These combined forecasts are computed used information available in real-time to the decision maker. Note that our out-of-sample period corresponds to the period post Bank of England operational independence. Following Clements (2004) we also consider a benchmark density forecast. It is assumed Gaussian with mean equal to actual inflation five quarters previously and variance equal to that estimated from the available sample for actual inflation.

Table 2 summarises the results of the forecast competition. For interest, as well as the density forecast evaluation results we report two summary statistics widely used to evaluate point forecasts, bias and RMSE. Statistical tests (not reported) reject the

NIESR's approach.

significance of the bias component, using a robust estimator of its standard error.

Let us consider the in-sample results first, although of course the Bank of England and NIESR results can also be seen as out-of-sample. The Bank of England density forecasts only fail the tests for independence. They pass the distributional tests.[32] Of course, we should expect serial correlation in $\{z_t\}$ since the forecast horizon is longer than the periodicity of the data; we expect $MA(4)$ auto-correlation. Accordingly we also test for dependence in $\{z_t\}$ having split them into four groups corresponding to those forecasts made in q1, q2, q3 and q4. Under the assumption that the density forecasts are adequate, each of the four groups should exhibit independence. As suggested by Diebold et al. (1998), we use a Bonferroni correction to mitigate the effects of an inflated Type 1 error when testing across the four groups. This test supports the $MA(4)$ assumption and so lends further support to the view that the Bank density forecasts are correctly specified.[33] Interestingly, if we impose normality on the Bank density forecasts, with mean and variance as before, results are virtually unchanged; the Bank's assumption of a two-piece normal distribution since 1996q1, empirically at least, makes little practical difference.

In contrast NIESR density forecasts fail many of the distributional tests, as well as the independence test. This distributional failure was expected; see Figures 2-4. NIESR clearly over-estimated the degree of uncertainty associated with its point forecast for the period 1994-2002.[34] This is explained by their reliance on a mechanical examination of historical forecast errors too far back into the past. With the advantage of hindsight, we can see that by considering forecast errors back until 1982, NIESR were basing their uncertainty forecasts on their track-record across two different inflation 'regimes', the recent regime (post 1992/3) characterised by lower volatility; see Hall & Mitchell (2004*a*) for further discussion. This serves as a timely reminder to forecasters that just as with point forecasts, basing density forecasts on past experience can lead to misleading forecasts. From 2002 NIESR considered errors from 1993 only and we see in Figure 4 the variance of

[32] These results are consistent with Clements (2004) who evaluates Bank density forecasts of year-ahead inflation, using a restricted set of statistical tests, over the shorter sample period, 1997q3-2002q1.

[33] While we focus on evaluation of density forecasts using statistical tests, it should be noted, particularly given the small-sample size available, that exploratory data analysis based on examination of the plot against the uniform distribution and the auto-correlation functions [see Figures 8 and 9] does suggest the Bank (albeit, if our results are reliable, in a statistically insignificant manner) over-estimated the degree of uncertainty. Figure 8 reveals that the distribution function for the Bank is S-shaped. This indicates that they placed too much weight in the tails. This is consistent with the findings of Wallis (2004).

[34] This will not surprise NIESR; they corrected their variance estimates in 2002. In fact, to quote from NIESR themselves, Poulizac et al. (1996) p. 62, "Both our inflation forecast and the reliability of this forecast must depend on the seriousness with which the government approaches inflation targetting. It is not clear that past experience is a good guide to this... and, in turn, [this] probably implies that the error variances... overstate the current uncertainty associated with the inflation rate". NIESR, see Blake (1996), have considered how stochastic simulation can be used as an alternative to historical errors to measure the uncertainty associated with the inflation rate. It is explained that this is expected to deliver a better measure of uncertainty if a new policy regime (say a new target for inflation) has been adopted. Using a coherent policy structure with interest rate setting determined by a monetary policy rule, Blake found that stochastic simulation suggested a smaller inflation standard error.

the density drop. We note that the Bank of England and NIESR mean (variance) forecast errors have a correlation coefficient of 0.73 (-0.33).

The benchmark density forecast passes all of the distributional tests, that do not also test independence in some manner, aside from the AD test. Like the Bank of England and NIESR forecasts the benchmark density does exhibit dependence due to the over-lapping nature of the forecasts.

But does combination help? Consider the indirect combination method first (column Ind) in Table 2. Combining the Bank of England and NIESR density forecasts using this method does not appear to deliver improved forecasts. While better than NIESR forecasts, the combined forecasts appear worse than those of the Bank of England.[35] Indeed, examination of the weights attributed to the Bank of England we see that the NIESR forecasts, using this method, offer little information compared to those of the Bank. The Bank has a weight of 0.9321 on the mean, and 0.8455 on the variance. Interestingly, using the indirect method to combine the Bank of England and benchmark densities does appear to help, with the majority of the tests suggesting an improvement in forecast accuracy.

However, and encouragingly, the direct method to combination suggests there are clear gains associated with combining Bank of England and NIESR density forecasts; see column Dir in Table 2. Not just do the distributional tests suggest improvement (in all cases the test statistic is lower, or the p-value is higher, relative to the individual density forecasts), but despite the over-lapping nature of the forecasts there is no longer evidence for serial correlation.

These results extend out-of-sample. The statistical tests suggest that the direct method of combination delivers more accurate density forecasts than use of Bank of England or NIESR forecasts alone. These improvements involve elimination of dependence. Combining competing density forecasts therefore appears to be a promising line of research. We note that using the indirect combination method to pool the Bank and NIESR forecasts there is little fluctuation in the weights attached to their mean and variance announcements across the recursive samples; the Bank has a minimum weight on the mean (variance) of 0.8353 (0.8132) and a maximum of 1.032 (0.8392).

7 Conclusion

This paper suggests techniques for the combination of density forecasts that have a Bayesian motivation. An application to U.K. inflation suggests that pooling information across density forecasts can deliver empirical gains. This is consistent with previous findings about point forecasts. In future work we hope to add further to the tool-kit of the applied econometrician interested in assessing density forecasts, by suggesting statistical tests that enable one to test statistically for the superiority of one density forecast over another. We might think of these tests as extensions to the Diebold-Mariano tests routinely used to compare statistically competing point forecasts. A start has already

[35] It is worth noting that, as expected, the RMSE is lower using the indirect combination method when the weights used on the point estimates are not re-computed following Section 3.2.

been made by Giacomini (2002) who proposes formal tests that can be used to rank alternative density forecasts based on their evaluation using scoring rules. It would also be interesting to compare methods of density forecast combination with event and quantile based combination techniques.

A Combining the reported measures of skewness

This section extends the discussion in Sections 3 and 4 to consider how the decision maker can use available information from the experts on higher moments than the mean and variance when combining their density forecasts. In a symmetric distribution the mean, median and mode coincide; but many experts such as the Bank of England, for example, view it as important to allow for asymmetries in their forecast densities; see Britton et al. (1998). The Bank of England does this by supplying information on the mean and mode of its density forecast. Standardised skewness, s_i^k, is then defined for expert i as the scaled (by uncertainty) difference between the mean and mode: $\left[E(y) - \mu^d\right]/\sigma$.

The decision maker when deriving the combined density can then condition on higher moments announced by the experts (such as standardised skewness $\mathbf{s}^k = (s_1^k, ..., s_N^k)$) by straightforwardly augmenting Bayes' Theorem to deliver the posterior distribution of y conditional on not just the first two moments announced by the experts, but also reported skewness: $h(y|\mathbf{m}, \mathbf{v}, \mathbf{s}^k) \propto h(\mathbf{m}|y).h(\mathbf{v}|\mathbf{m},y).h(\mathbf{s}^k|\mathbf{m}, \mathbf{v},y).h(y)$. To make the indirect and direct combination methods operational we need to define the likelihood $h(\mathbf{s}^k|\mathbf{m}, \mathbf{v},y)$. This is is achieved as follows; it is based on the difference between experts' announced skewness and actual skewness.

In measuring actual standardised skewness we confront the problem of determining the mode. However, a convenient measure of standardised skewness can be estimated from the sample by recalling that for the Pearson class of distributions standardised skewness may be expressed in terms of the first four moments; e.g. see Kendall & Stuart (1963), p.85. Specifically actual skewness is proxied for expert i by S_{ik}, where:

$$S_{ik} = \frac{\sqrt{\beta_{1i}}(\beta_{2i} + 3)}{2(5\beta_{2i} - 6\beta_{1i} - 9)}, \tag{33}$$

where

$$\beta_{1i} = \frac{\mu_{3i}^2}{\mu_{2i}^3} \text{ and } \beta_{2i} = \frac{\mu_{4i}}{\mu_{2i}^2}, \tag{34}$$

and $\mu_{ri} = s_{it}^r = (y_t - m_{it})^r$ denotes the r-th central moment for expert i.

Then consider the difference between actual skewness and reported skewness:

$$s_i^k - S_{ik} = u_{2it}, \tag{35}$$

where consistent with the discussion above it is assumed that experts' forecasts of skewness are unbiased in the long-run so that $\mathbf{u}_{2t} = (u_{2,1t}, ..., u_{2,Nt}) \sim N(\mathbf{0}, \Xi_t)$. Therefore $h(\mathbf{s}^k|\mathbf{m}, \mathbf{v},y) \propto \exp(-\frac{1}{2}(\mathbf{u}_{2t}'\Xi_t^{-1}\mathbf{u}_{2t}))$.

B Statistical tests for the adequacy of density forecasts: testing $i.i.d.$ $U(0,1)/N(0,1)$

This section reviews the approach taken in this paper to test the statistical adequacy of density forecasts. An eclectic approach to testing *i.i.d.* uniformity/normality is followed.

We consider a range of statistical tests that have been used in empirical studies.[36] The tests are used to detect misspecification in the mean, variance, skewness and/or kurtosis of the forecasts. But the tests differ not just in terms of their motivation, which we do not discuss here, but with respect to what they test. Some test for misspecification in all of the first four moments, while others focus on specific moments. Below we summarise relevant aspects of the tests.[37]

1. Kolmogorov-Smirnov (KS) test for uniformity of $\{z_t\}$. The KS test relies on random sampling. As noted, for example, by Diebold et al. (1999) the effect of dependence on the distribution of the KS test statistic is unknown.

2. Anderson and Darling (AD) test for uniformity of $\{z_t\}$. Using Monte-Carlo Noceti et al. (2003) found the AD test to have more power to detect misspecification than the KS test (and related distributional tests).

3. Berkovitz's parametric test of *i.i.d.* N(0,1); see Berkowitz (2001). We consider the three degrees of freedom test of zero mean, unit variance and independence against an AR(1). This test only has power to detect non-normality through the first two moments.

4. The Jarque-Bera (JB) test for normality of $\{z_t^*\}$; see Jarque & Bera (1980). JB test for normality looking at the coefficients of skewness and kurtosis. This test has no power to detect misspecification in the first two moments.

5. The Doornik-Hansen (DH) test for normality of $\{z_t^*\}$; see Doornik & Hansen (1994). DH develop a test for normality, again based on skewness and kurtosis, that seeks to overcome the perceived problem with JB-type tests that they are unreliable except for very large T. Based on a transformation, DH propose a small-sample test. The test is based on random sampling, and has no power to detect misspecification in the first two moments.

6. The Bai and Ng (BN) 'robust' version of the JB test on $\{z_t^*\}$; see Bai & Ng (2003). Bai and Ng extend the JB test, designed for *i.i.d.* data, to weakly dependent data. Essentially, any serial dependence is taken into account by consistently estimating the long-run variance (the spectral density at frequency zero) using a HAC estimator that is robust to serial dependence (and heteroscedasticity).

7. A robust test for mean zero and variance one (DM) of $\{z_t^*\}$. Along the lines of the Diebold & Mariano (1995) test, the joint hypothesis of a zero mean and a variance of unity for $\{z_t^*\}$ is tested by computing p-values for the statistic D using the standard normal CDF, where $D = \left(\frac{1}{T}\sum_{t=1}^{T}(z_t^*)^2 - 1\right)(\sigma^*/\sqrt{T})^{-1}$ and σ^* is an estimator

[36] Alternatively, graphical means of exploratory data analysis are often used to examine the quality of density forecasts; see Diebold et al. (1998) and Diebold et al. (1999).

[37] See Mitchell (2004) for further details and a Monte-Carlo based investigation of the size and power of these tests.

for the standard deviation of $\{(z_t^*)^2\}$. We use a HAC estimator for this standard deviation to be robust against serial correlation (and heteroscedasticity). This test ignores the third and fourth moments.

8. A Ljung-Box (LB) test of independence of $\{z_t\}$. To test for independence of the $\{z_t\}$ series we use the Ljung-Box test for auto-correlation; see Harvey (1989), p. 259. Since dependence may occur in higher moments we consider $(z_t - \overline{z})^j$ for $j = 1, 2, 3$; results are denoted LB1, LB2 and LB3, respectively.

9. The Hong joint test for uniformity and independence applied to $\{z_t\}$; see Hong (2002). This test is theoretically attractive as it offers a joint test. This means one can control the size of the test, something that cannot easily be done using separate tests for uniformity/normality and independence. Hong's joint test for *i.i.d.* U(0,1) is based on generalised spectral analysis. We estimate the spectrum nonparametrically using the Bartlett kernel and, following Hong, use a data-driven approach to determine the bandwidth. This requires us to choose a preliminary bandwidth. Hong proposes two test statistics, M_0 and M_1. Hong recommends $M1$ in small samples; accordingly we consider this. Hong also proposes a test for independence robust to non-uniformity.

10. A portmanteau test of uniformity and independence of $\{z_t\}$ based on summing alternative test statistics; see Thompson (2002). This test has been used by Clements (2004) to evaluate the Bank of England density forecasts of U.K. inflation. Thompson proposes the test statistic Q_T for uniformity and independence which is based on summing, assuming equal weights, the Cramer-von-Mises test for uniformity and tests for uncorrelatedness of $\{z_t\}$ and $\left\{(z_t - \overline{z})^2\right\}$ [see Durlauf (1991) and Fong & Ouliaris (1999)] based on the cumulative periodogram:

$$Q_T = \int_0^1 \widehat{F}^2(s)ds + \int_0^1 \widehat{P}_z^2(s)ds + \int_0^1 \widehat{P_{z_2}}^2(s)ds, \tag{36}$$

where $\int_0^1 \widehat{F}^2(s)ds$ is the Cramer-von-Mises, or integrated-squared, distance between $\widehat{F}(s)$ and zero measuring the goodness-of-fit relative to the uniform CDF[38], $\widehat{P}_z(s) = \frac{\sqrt{2}}{\pi}\sum_{j=1}^{T-1} T^{1/2}\widehat{\rho}_z(j)\frac{\sin j\pi s}{j}$, where $s \in [0, 1]$ and $\widehat{\rho}_z(j)$ is the j-th sample autocorrelation of $\{z_t\}$ $(j = 1, 2, 3...)$, where $\widehat{P}_z(s)$ is the basis for the spectral based test for the martingale hypothesis for z_t considered by Durlauf (1991) and Fong & Ouliaris (1999), and $\widehat{P}_{z_2}(s)$ is analogous to $\widehat{P}_z(s)$ but based on $\{z_{2t}\} = \{2\,|z_t - 0.5|\}$; this is analogous to in the time-domain testing for correlation in $\left\{(z_t - \overline{z})^2\right\}$. Finite sample critical values for Q_T, and its components, are derived by examining the quantiles of the test statistics based on simulation of uniform random variables with $T = 42$.

[38] One could consider the KS or AD test, but Thompson finds it easier to work with the Cramer-von-Mises test.

Aside from the Hong test, all of the above tests have been used by time-series analysts when evaluating density forecasts *ex post*. Selected examination of the empirical literature, with a bias towards macroeconomic applications, suggests that the KS/AD, DH and LB tests are most commonly used.

Table 1: Bank of England and NIESR Density Forecasts of Inflation: Summary Properties

	BANK					NIESR			
	mode	mean	s.d.	outturn	z	mean	s.d.	outturn	z
199301	3.4	3.4	1.8	2.675	0.3436	4.785	2	2.675	0.1458
199302	3.5	3.5	1.5	2.413	0.2343	4.942	2	2.413	0.103
199303	3.3	3.3	1.5	2.195	0.2308	4.497	2	2.195	0.125
199304	3.4	3.4	1.5	2.26	0.2236	4.202	2	2.26	0.1657
199401	3.4	3.4	0.9	2.746	0.2339	4.117	2	2.746	0.2466
199402	3.4	3.4	1	2.703	0.2428	3.204	2	2.703	0.401
199403	3.2	3.2	0.9	2.911	0.3739	3.707	2	2.911	0.3453
199404	2.4	2.4	0.9	2.901	0.711	2.79	2	2.901	0.5221
199501	2.7	2.7	0.9	2.879	0.5787	3.361	2	2.879	0.4046
199502	3.8	3.8	1	2.834	0.167	2.636	2	2.834	0.5394
199503	3.4	3.4	0.9	2.896	0.2876	2.547	2	2.896	0.5691
199504	3	3	0.9	3.221	0.5972	3.025	2	3.221	0.5391
199601	2.3	2.1	0.929	2.865	0.7924	3.023	2	2.865	0.4685
199602	2.2	2.3	0.9075	2.559	0.6199	2.404	2	2.559	0.5309
199603	2.4	2.6	0.929	2.814	0.6063	2.56	2	2.814	0.5506
199604	2.4	2.4	0.9	2.796	0.67	2.627	2	2.796	0.5336
199701	2.3	2.4	0.9075	2.591	0.5915	1.817	2	2.591	0.6506
199702	2.2	2.4	0.8324	2.943	0.7491	2.103	2	2.943	0.6628
199703	1.99	2.2	0.7877	2.546	0.6829	2.103	2	2.546	0.5877
199704	2.19	2.72	0.7503	2.53	0.4543	2.394	2	2.53	0.5272
199801	2.44	2.53	0.5009	2.525	0.5114	2.308	2	2.525	0.5433
199802	2.37	2.15	0.6592	2.3	0.5626	2.504	2	2.3	0.4593
199803	2.86	3	0.6213	2.173	0.08376	2.5	2	2.173	0.435
199804	2.59	2.72	0.6379	2.159	0.1904	2.47	1.8	2.159	0.4314
199901	2.52	2.58	0.6239	2.094	0.2193	2.372	1.8	2.094	0.4385
199902	2.23	2.34	0.6035	2.066	0.3354	2.24	1.8	2.066	0.4614
199903	1.88	2.03	0.5858	2.126	0.5844	2.78	1.8	2.126	0.3582
199904	1.84	1.79	0.5531	2.114	0.7171	2.324	1.67	2.114	0.45
200001	2.32	2.42	0.5718	1.87	0.1669	2.406	1.67	1.87	0.374
200002	2.47	2.52	0.5531	2.262	0.3253	2.107	1.67	2.262	0.5369
200003	2.48	2.48	0.54	2.38	0.4262	2.371	1.67	2.38	0.5022
200004	2.19	2.24	0.563	1.952	0.3086	2.016	1.67	1.952	0.4847
200101	2.09	2.04	0.5531	2.368	0.7201	1.662	1.67	2.368	0.6638
200102	1.94	1.89	0.5531	1.863	0.4727	2.168	1.67	1.863	0.4274
200103	1.96	1.96	0.55	1.976	0.5113	2.322	1.67	1.976	0.4179
200104	2.06	2.26	0.595	2.61	0.7325	1.795	1.67	2.61	0.6873
200201	2.13	2.33	0.5857	2.892	0.8298	2.15	1.67	2.892	0.6716
200202	2.05	2.05	0.52	2.914	0.9518	2.029	1.67	2.914	0.702
200203	2.31	2.31	0.51	2.849	0.8547	2.103	1.67	2.849	0.6725
200204	2.41	2.41	0.48	2.6	0.6541	2.351	0.71	2.6	0.6373
200301	2.7	2.7	0.56	2.305	0.2401	2.254	0.71	2.305	0.5286
200302	2.35	2.45	0.5229	2.165	0.3024	2.629	0.71	2.165	0.2567

Table 2: Evaluation of the Bank of England, NIESR and combined density forecasts using both the indirect (Ind) and direct (Dir) methods: In-sample and out-of-sample results

	In-sample: 93q1-03q2						Recursive Out-of-Sample: 97q3-03q2				
	Bank	NIESR	Bench	Ind	Ind2	Dir	Bank	NIESR	Bench	Ind	Dir
Bias	-0.104	-0.187	-0.067	-0.110	-0.086	-0.048	-0.020	0.053	-0.036	-0.011	-0.051
RMSE	0.5312	0.8384	0.4833	0.5389	0.4581	0.6386	0.4044	0.4543	0.458	0.4105	0.4744
KS	0.156	0.298	0.184	0.223	0.128	0.101	0.143	0.317	0.237	0.224	0.171
KS:cv	0.205	0.205	0.205	0.205	0.205	0.205	0.269	0.269	0.269	0.269	0.269
AD	1.723	5.374	2.545	3.670	0.887	0.344	0.675	4.181	1.760	2.122	0.999
DH	0.702	0.008	0.888	0.345	0.387	0.750	0.752	0.986	0.241	0.782	0.460
BN	0.316	0.508	0.603	0.175	0.494	0.830	0.653	0.861	0.327	0.491	0.619
Ber	0.000	0.000	0.000	0.000	0.000	0.302	0.007	0.000	0.002	0.000	0.022
JB	0.663	0.019	0.807	0.411	0.463	0.880	0.825	0.871	0.381	0.677	0.795
DM	0.002	0.000	0.072	0.000	0.202	0.246	0.056	0.000	0.195	0.000	0.009
mean	-0.057	-0.093	-0.190	-0.070	-0.087	-0.005	-0.009	0.028	-0.103	-0.017	-0.032
sd	0.661	0.440	1.377	0.488	0.837	0.887	0.721	0.305	1.325	0.484	0.683
Q_T	2.115	5.903	2.310	2.790	2.305	0.670	1.281	1.435	1.514	1.510	0.888
cvm	0.236	0.982	0.269	0.566	0.144	0.043	0.089	0.768	0.269	0.329	0.145
cvm1	1.467	2.601	1.712	1.628	1.302	0.367	0.677	0.450	0.924	0.636	0.424
cvm2	0.412	2.320	0.328	0.596	0.859	0.261	0.516	0.218	0.321	0.545	0.320
H: M1	0.002	0.005	0.011	0.004	0.005	0.000	0.001	0.002	0.010	0.001	0.001
H: iid	7.414	16.290	9.318	8.526	6.569	1.311	2.997	1.634	4.580	2.729	1.329
LB1	0.003	0.000	0.000	0.001	0.004	0.203	0.061	0.205	0.011	0.074	0.144
LB2	0.058	0.000	0.047	0.095	0.002	0.447	0.054	0.372	0.020	0.174	0.652
LB3	0.005	0.000	0.000	0.002	0.000	0.086	0.062	0.502	0.008	0.049	0.404
Bonf	x	x	x	x	x	x	x	x	x	x	x

Notes: Ind2 are the Bank and benchmark indirect combination; bias and RMSE summarise the performance of the point (mean) forecast; KS is the Kolmogorov-Smirnov statistic; KS:cv is the associated 95% critical value; AD is the Anderson-Darling statistic which has an associated 95% critical value of 2.502; DH is the p-value of the Doornik-Hansen test for normality; BN is the p-value of the Bai-Ng robust test for normality; Ber is the p-value of the Berkovitz test for iid normality; JB is the p-value of the Jarque-Bera test for normality; DM is the p-value of the Diebold-Mariano type test for mean and variance of zero and unity; mean and sd are the sample mean and standard deviation of $\{z_t^*\}$; Q_T is the Portmanteau test statistic of Thompson for uniformity and independence; cvm, cvm1 and cvm2 are the component test statistics for uniformity, first moment and second moment independence - the 95%critical values for these 4 statistics computed via simulation are 0.98, 0.45, 0.47 and 0.47. H: M1 is the Hong joint test statistic for uniformity and independence with associated 95% critical value of 0.052; H: iid is the Hong test statistic for independence that is robust to non-uniformity with associated 95% critical value of 1.96. LB1, LB2 and LB3 are the p-values for the Ljung-Box tests for serial correlation in the first, second and third power; Bonf is denoted x when there is no evidence of serial correlation in the q1, q2, q3 and q4 forecasts as judged using Bonferroni corrected critical values

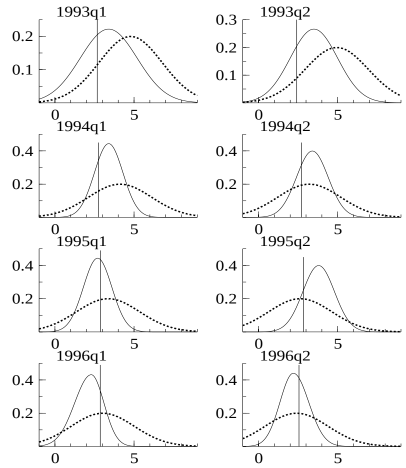

Figure 2: Bank of England (solid) and NIESR (dotted) density forecasts of inflation: subsequent outturn for inflation is marked with a vertical line

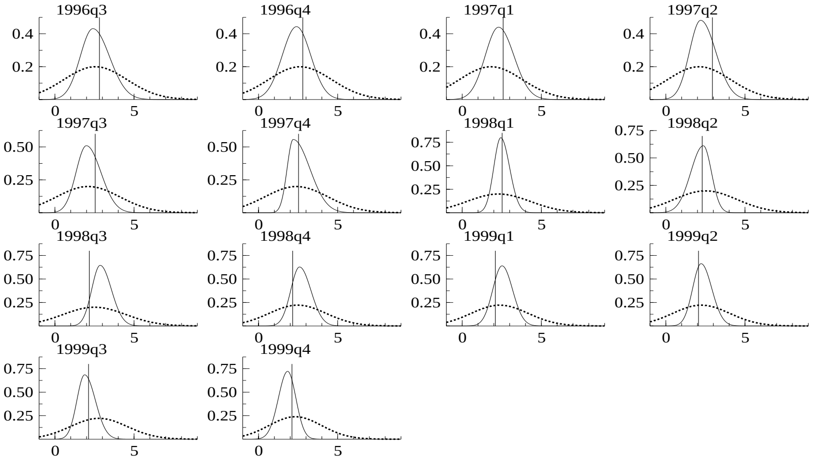

Figure 3: Bank of England (solid) and NIESR (dotted) density forecasts of inflation

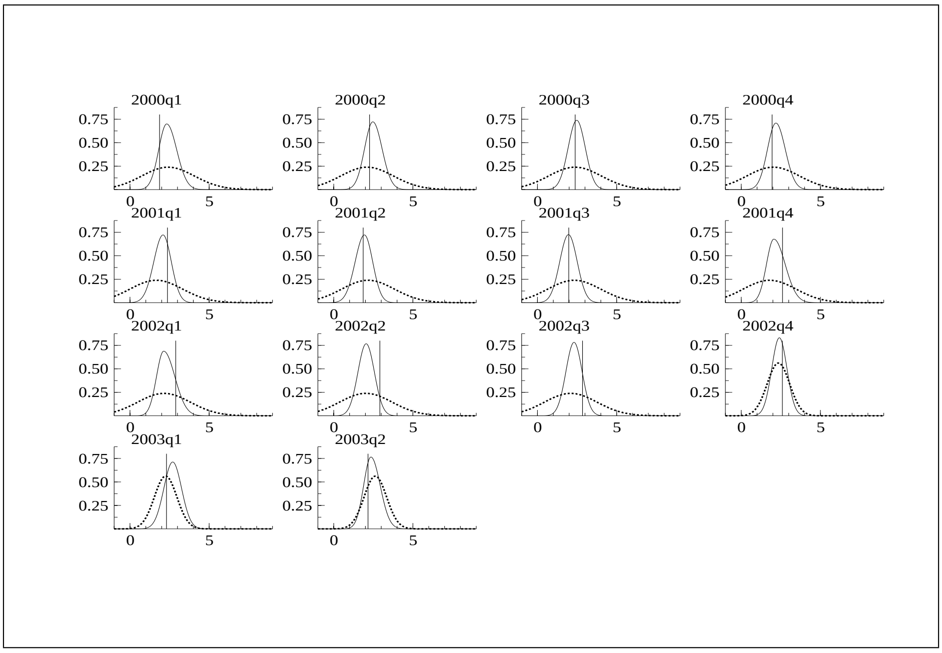

Figure 4: Bank of England (solid) and NIESR (dotted) density forecasts of inflation

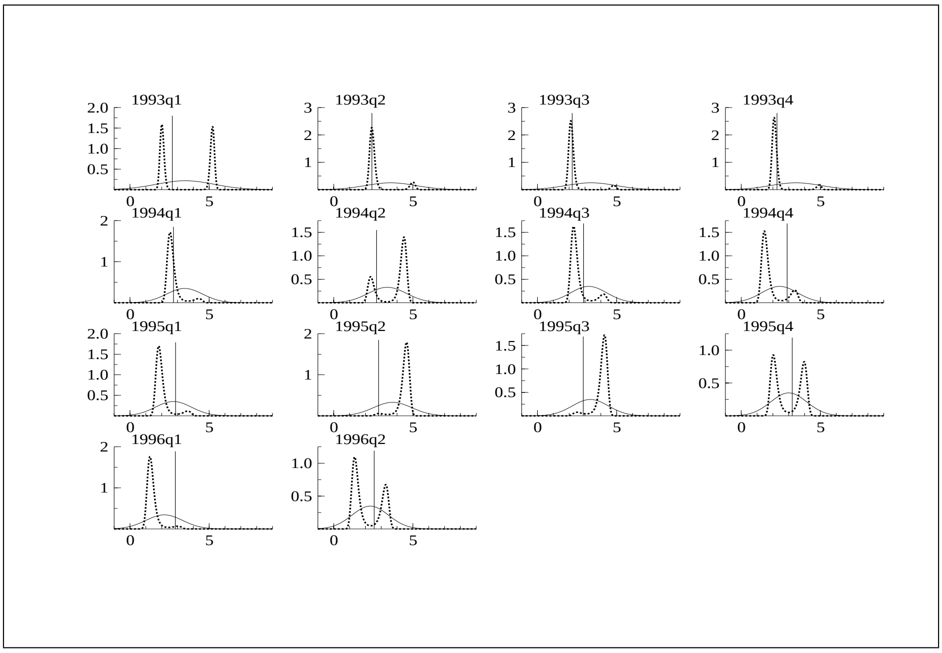

Figure 5: Indirect (solid) and direct (dotted) combined Bank of England and NIESR density forecasts of inflation

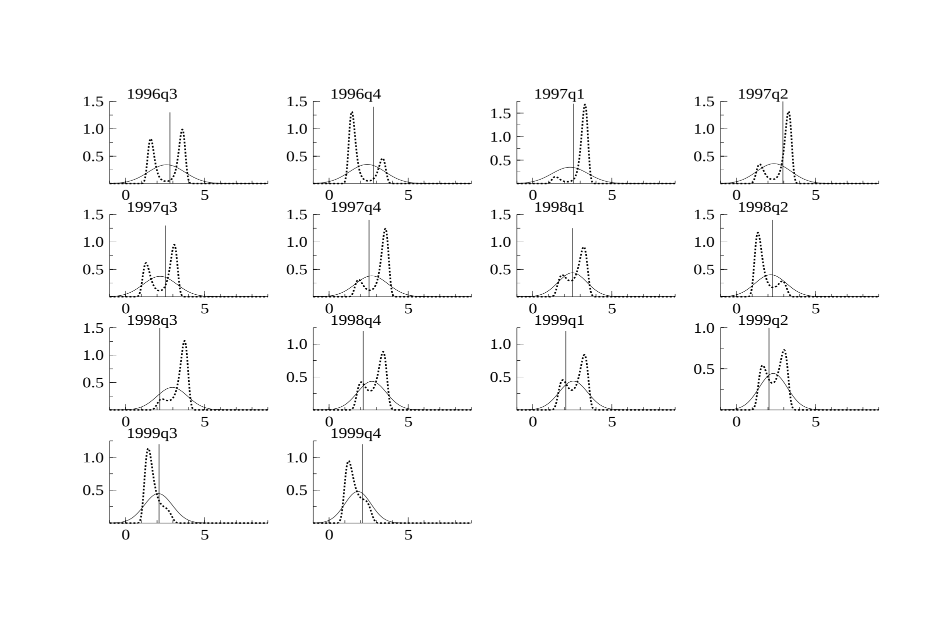

Figure 6: Indirect (solid) and direct (dotted) combined Bank of England and NIESR density forecasts of inflation

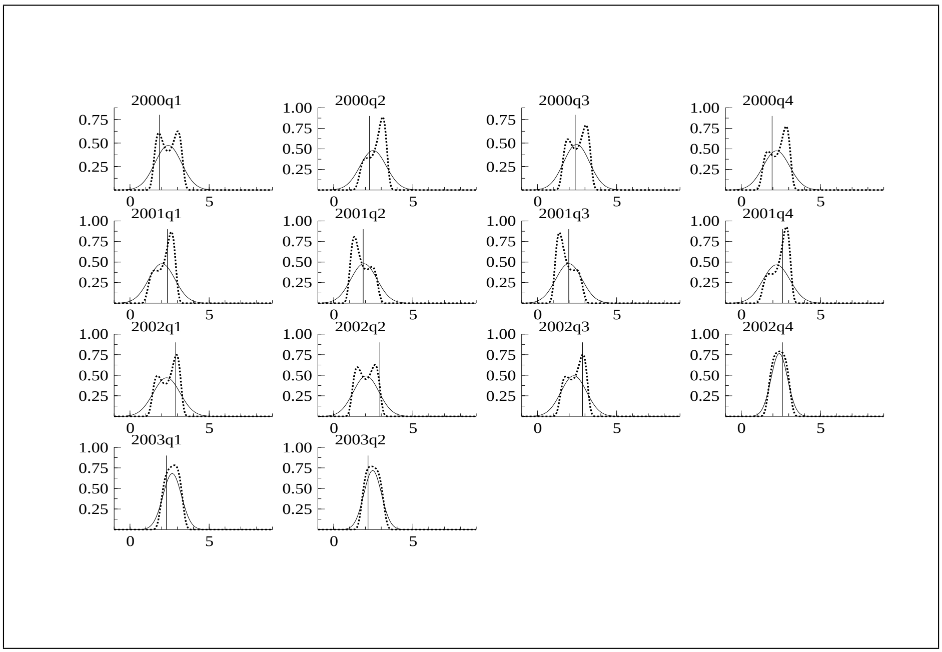

Figure 7: Indirect (solid) and direct (dotted) combined Bank of England and NIESR density forecasts of inflation

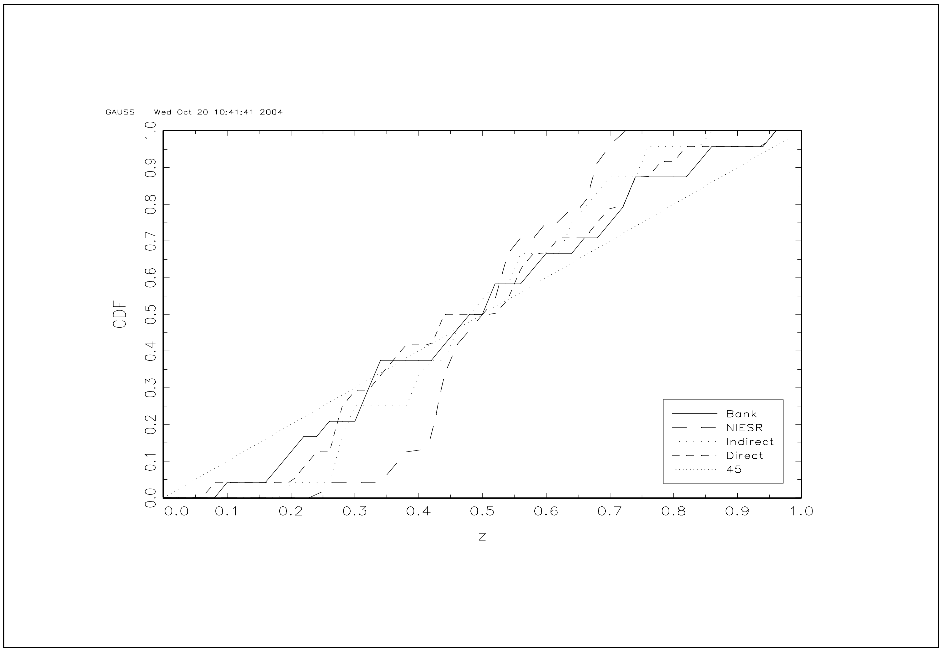

Figure 8: Cumulative Distribution Functions of $\{z_t)$ and uniform distribution for the out-of-sample period

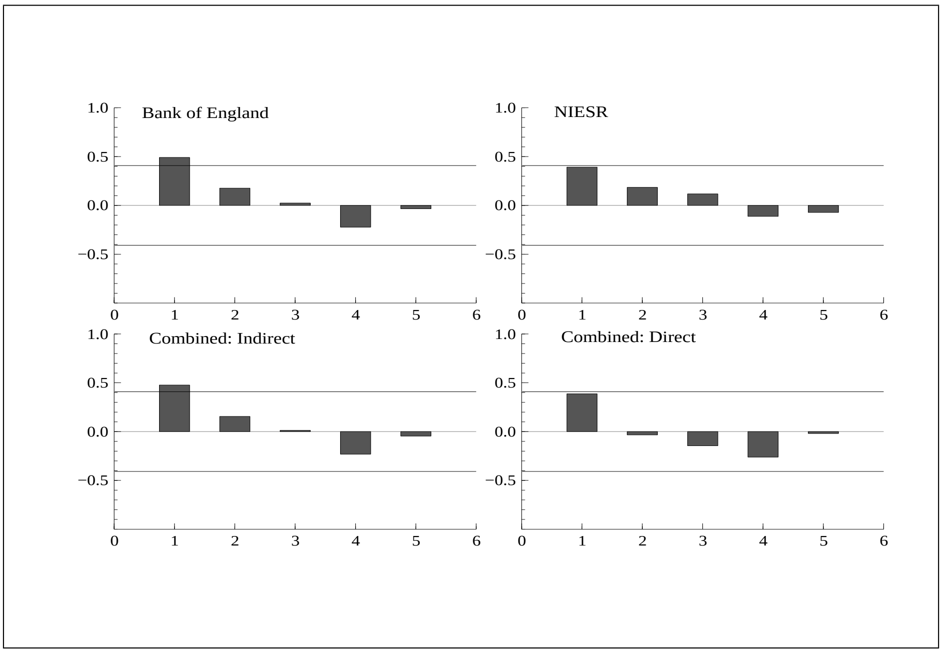

Figure 9: Sample auto-correlation functions of $\{z_t\}$ for the out-of-sample period

References

Andersen, T. & Bollerslev, T. (1998), 'Answering the skeptics: yes, standard volatility models do provide accurate forecasts', *International Economic Review* **39**, 885–905.

Bai, J. & Ng, S. (2003), 'Tests for skewness, kurtosis and normality for time series data', *Journal of Business and Economic Statistics* . Forthcoming.

Batchelor, R. & Dua, P. (1995), 'Forecaster diversity and the benefits of combining forecasts', *Management Science* **41**, 68–75.

Bates, J. M. & 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.

Blake, A. (1996), 'Forecast error bounds by stochastic simulation', *National Institute Economic Review* **156**, 72–79.

Bomberger, W. (1996), 'Disagreement as a measure of uncertainty', *Journal of Money, Credit and Banking* **28**, 381–392.

Britton, E., Fisher, P. & Whitley, J. (1998), 'The inflation report projections: understanding the fan chart', *Bank of England Quarterly Bulletin* **38**, 30–37.

Clemen, R. & Winkler, R. (1999), 'Combining probability distributions from experts in risk analysis', *Risk Analysis* **19**, 187–203.

Clements, M. P. (2002), 'An evaluation of the survey of professional forecasters probability distributions of expected inflation and output growth', *Warwick University Discussion Paper* .

Clements, M. P. (2004), 'Evaluating the Bank of England density forecasts of inflation', *Economic Journal* **114**, 844–866.

Clements, M. P. & Hendry, D. F. (1998), *Forecasting Economic Time Series*, Cambridge University Press: Cambridge.

Demetriades, P. O. (1989), 'The relationship between the level and variability of inflation: theory and evidence', *Journal of Applied Econometrics* **4**, 239–250.

Deutsch, M., Granger, C. W. J. & Terasvirta, T. (1994), 'The combination of forecasts using changing weights', *International Journal of Forecasting* **10**, 47–57.

Diebold, F. X., Gunther, A. & Tay, K. (1998), 'Evaluating density forecasts with application to financial risk management', *International Economic Review* **39**, 863–883.

Diebold, F. X. & Mariano, R. S. (1995), 'Comparing predictive accuracy', *Journal of Business and Economic Statistics* **13**, 253–263.

Diebold, F. X. & Pauly, P. (1987), 'Structural change and the combination of forecasts', *Journal of Forecasting* **6**, 21–40.

Diebold, F. X., Tay, A. S. & Wallis, K. F. (1999), Evaluating density forecasts of inflation: the Survey of Professional Forecasters, *in* R. Engle & H. White, eds, 'Cointegration, causality and forecasting: a festschrift in honour of Clive W. J. Granger', Oxford University Press.

Doornik, J. A. & Hansen, H. (1994), A practical test for univariate and multivariate normality. Discussion Paper, Nuffield College, Oxford.

Durlauf, S. (1991), 'Spectral based testing of the martingale hypothesis', *Journal of Econometrics* **50**, 355–37.

Everitt, B. S. & Hand, D. J. (1981), *Finite Mixture Distributions*, London: Chapman and Hall.

Fong, W. M. & Ouliaris, S. (1999), 'Spectral tests of the martingale hypothesis for exchange rates', *Journal of Applied Econometrics* **10**(3), 255–271.

Garratt, A., Lee, K., Pesaran, M. H. & Shin, Y. (2003), 'Forecast uncertainties in macroeconometric modelling: an application to the UK economy', *Journal of the American Statistical Association* **98**, 829–838.

Genest, C. & Zidek, J. (1986), 'Combining probability distributions: a critique and an annotated bibliography', *Statistical Science* **1**, 114–135.

Giacomini, R. (2002), Comparing density forecasts via weighted likelihood ratio tests: asymptotic and bootstrap methods. UCSD Discussion Paper 2002-12.

Giordani, P. & Söderlind, P. (2003), 'Inflation forecast uncertainty', *European Economic Review* **47**, 1037–1059.

Granger, C. W. J. & Jeon, Y. (2004), 'Thick modeling', *Economic Modelling* **21**, 323–343.

Granger, C. W. J. & Pesaran, M. H. (2000), 'Economic and statistical measures of forecast accuracy', *Journal of Forecasting* **19**, 537–560.

Granger, C. W. J. & Ramanathan, R. (1984), 'Improved methods of combining forecasts', *Journal of Forecasting* **3**, 197–204.

Granger, C. W. J., White, H. & Kamstra, M. (1989), 'Interval forecasting: an analysis based upon ARCH-quantile estimators', *Journal of Econometrics* **40**, 87–96.

Hall, S. G. & Mitchell, J. (2004*a*), An evaluation of NIESR's probability forecasts of inflation. mimeo, NIESR.

Hall, S. G. & Mitchell, J. (2004*b*), "Optimal" combination of density forecasts. National Institute of Economic and Social Research Discussion Paper No. 248.

Halperin, M. (1961), 'Almost linearly-optimum combination of unbiased estimates', *Journal of the American Statistical Association* **56**, 36–43.

Harvey, A. C. (1989), *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Hendry, D. F. & Clements, M. P. (2004), 'Pooling of forecasts', *Econometrics Journal* **7**, 1–31.

Hong, Y. (2002), Evaluation of out-of-sample probability density forecasts. Cornell University Discussion Paper.

Jarque, C. & Bera, A. (1980), 'Efficient tests for normality, homoskedasticity and serial independence of regression residuals', *Economics Letters* **12**, 255–259.

Jouini, M. N. & Clemen, R. T. (1996), 'Copula models for aggregating expert opinions', *Operations Research* **44**, 444–457.

Kendall, M. G. & Stuart, A. (1963), *The Advanced Theory of Statistics: Volume 1 (2nd Edition)*, Charles Griffin, London.

Lindley, D. (1983), 'Reconciliation of probability distributions', *Operations Research* **31**, 866–880.

Mitchell, J. (2004), A Monte-Carlo based comparison of alternative density forecast evaluation tests, with an application to the Bank of England's "fan" chart of inflation. mimeo, NIESR.

Morris, P. (1974), 'Decision analysis expert use', *Management Science* **20**, 1233–1241.

Morris, P. (1977), 'Combining expert judgments: A Bayesian approach', *Management Science* **23**, 679–693.

Noceti, P., Smith, J. & Hodges, S. (2003), 'An evaluation of tests of distributional forecasts', *Journal of Forecasting* **22**, 447–455.

Pesaran, M. H. & Zaffaroni, P. (2004), Model averaging and value-at-risk based evaluation of large multi asset volatility models for risk management. University of Cambridge.

Poulizac, D., Weale, M. & Young, G. (1996), 'The performance of National Institute economic forecasts', *National Institute Economic Review* **156**, 55–62.

Stock, J. & Watson, M. (2004), 'Combination forecasts of output growth in a seven-country data set', *Journal of Forecasting* **23**, 405–430.

Tay, A. S. & Wallis, K. F. (2000), 'Density forecasting: a survey', *Journal of Forecasting* **19**, 235–254.

Taylor, J. (1999), 'Evaluating volatility and interval forecasts', *Journal of Forecasting* **18**, 111–128.

Thompson, S. (2002), Evaluating the goodness of fit of conditional distributions, with an application to affine term structure models. manuscript Economics Department, Harvard University.

Treasury, H. M. (2004), 'Forecasts for the UK economy: a comparison of independent forecasts'. HM Treasury No. 201.

Wallis, K. F. (1989), 'Macroeconomic forecasting: a survey', *Economic Journal* **99**, 28–61.

Wallis, K. F. (2004), 'An assessment of Bank of England and National Institute inflation forecast uncertainties', *National Institute Economic Review* **189**, 64–71.

Winkler, R. (1981), 'Combining probability distributions from dependent information sources', *Management Science* **27**, 479–488.

Zarnowitz, V. & Lambros, L. (1987), 'Consensus and uncertainty in economic prediction', *Journal of Political Economy* **95**, 591–621.