International Journal of Forecasting 35 (2019) 474–484

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Combining forecasts: Performance and coherence

Mary E. Thomson [a,*], Andrew C. Pollock [b], Dilek Önkal [a], M. Sinan Gönül [a]

[a] Newcastle Business School, Northumbria University, City Campus East, Newcastle upon Tyne, NE1 8ST, UK
[b] 1 Pladda Avenue, Irvine, KA11 1DR, UK

ARTICLE INFO

Keywords:
Forecast
Accuracy
Coherence
Composite forecasts
Inflation

ABSTRACT

There is general agreement in many forecasting contexts that combining individual predictions leads to better final forecasts. However, the relative error reduction in a combined forecast depends upon the extent to which the component forecasts contain unique/independent information. Unfortunately, obtaining independent predictions is difficult in many situations, as these forecasts may be based on similar statistical models and/or overlapping information. The current study addresses this problem by incorporating a measure of coherence into an analytic evaluation framework so that the degree of independence between sets of forecasts can be identified easily. The framework also decomposes the performance and coherence measures in order to illustrate the underlying aspects that are responsible for error reduction. The framework is demonstrated using UK retail prices index inflation forecasts for the period 1998–2014, and implications for forecast users are discussed.

1. Introduction

Over recent decades, considerable research attention has been given to the combination of predictions (e.g., Armstrong, 2001; Bates & Granger, 1969; Clemen, 1989; De Menezes, Bunn, & Taylor, 2000; Soll & Larrick, 2009; Timmerman, 2006; Wallis, 2011) as a means of improving the forecasting accuracy relative to selecting the prediction of the best forecaster (e.g., Fifić & Gigerenzer, 2014). The seminal paper of Bates and Granger (1969) suggested that analysts should integrate multiple forecasts into a combined forecast by, for instance, constructing a weighted average of the independent forecasts via a relatively simple calculation for establishing effective weights. However, it is argued that weighting is only likely to be beneficial if the analyst has strong supporting evidence that some forecasts are likely to be more accurate than others (Armstrong, 2001); otherwise, it is difficult to outperform the simple average (Fischer & Harvey, 1999; Lawrence, Edmundson, & O'Connor, 1986; Leitner & Leopold-Wildburger, 2011; Soll & Larrick, 2009; Stock & Watson, 2004). Defining pivotal criteria for selecting which forecasts to combine remains a challenge for forecast-evaluation research.

Forecast evaluation involves comparing a set of predictions with their corresponding ex-post actual values. There are a variety of error measures that can be used for assessing forecast performances, including the mean squared error (MSE), the mean absolute error (MAE), the mean absolute percentage error (MAPE), and a range of related measures (Armstrong & Collopy, 1992; Hyndman & Koehler, 2006; Makridakis & Hibon, 2000). Unlike measures based on the absolute error, a major advantage of the MSE is that it can be decomposed into a number of underlying components that illustrate particular strengths and weaknesses in performance (as illustrated by Murphy, 1988; Pollock & Wilkie, 1996; Stewart & Lusk, 1994; Theil, 1966, for MSE; and by Murphy, 1973; Pollock, Macaulay, Thomson, & Önkal-Atay, 2005; Sanders, 1963, and Yates, 1982, for the mean probability score).

Based on a review of 30 studies that utilized such evaluation measures, Armstrong (2001) reports a reduction in

* Corresponding author.
E-mail addresses: mary.thomson@northumbria.ac.uk (M.E. Thomson), acpollock@virginmedia.com (A.C. Pollock), dilek.onkal@northumbria.ac.uk (D. Önkal), sinan.gonul@northumbria.ac.uk (M.S. Gönül).

https://doi.org/10.1016/j.ijforecast.2018.10.006
0169-2070/

ex-ante forecasting errors from combining forecasts and points out that the degree of improvement may rely upon the independent information content of individual forecasts. Bates and Granger (1969) emphasize that negatively-related forecast errors would present an ideal situation, as they would cancel each other out. Unfortunately, though, this condition rarely occurs in practice: forecast errors generally tend to be positively correlated, and often strongly so. Researchers attempt to accommodate this typical lack of independence by combining different types of forecasts (i.e., statistical and judgmental) based on a variety of forecasting methods and diverse information sets.

Unfortunately, it is often difficult to obtain independent forecasts, as the individual forecasters typically use comparable statistical models, or the individuals making judgmental forecasts use similar criteria and/or overlapping information. This paper addresses this problem directly by proposing measures of coherence that address consistency and dependence between forecasts and could be integrated effectively into the performance analysis of composite forecasts. The proposed framework demonstrates that composite forecasts will always have smaller MSEs than the average MSEs of the individual forecasts, due to diversity among the individual forecasts. The framework also illustrates that coherence measures can be used to quantify this improvement, such that the forecasts with the highest levels of independence can be identified easily.

This analytical framework makes an important contribution to the extant work by decomposing the performance and coherence using an integrated method in order to illustrate the underlying aspects that are responsible for the error reduction displayed in composite predictions. Specifically, the MSE is employed not only as a traditional measure of the accuracy or performance (denoted as the mean squared error for performance, MSEP), but also as a measure for examining the consistency between individual sets of paired forecasts, termed the mean squared error for coherence (MSEC). The MSEP is decomposed into component measures that involve bias, resolution and error variation, where bias is the difference between the mean of the forecasts and the mean of the actual values; resolution is effectively the slope coefficient of a linear relationship between the forecasts and the actual values; and error variation is the variation in the actual values that is not explained by variation in the forecasts. This decomposition is then extended to the second measure of MSEC. This integrated approach provides more in-depth information than performance analysis alone, and could be extremely useful for improving the forecast accuracy. Not only do the overall measures provide important insights into assessing groups of forecasters that use similar techniques, but the component measures can also be used to identify subtle differences in relative performance and coherence among these forecasters. At the same time, different aspects of the forecasting performance may be prioritized over others in specific situations (e.g. a low level of bias may be valued over a high resolution or vice versa), and being able to assess the components individually may be highly beneficial.

We demonstrate the application of the framework and how its measures may be interpreted by employing inflation rate data. Specifically, we use the retail prices index, RPI, with forecasts obtained through HM Treasury, for the UK economy (monthly), and compare inflation forecasts from the December editions of the years 1997–2013 for the fourth quarter (Q4) of the following year with actual RPI inflation rates for Q4 of years 1998–2014. Inflation forecasts are critical for decision/policy-makers across a wide range of domains, and hence provide an appropriate platform for examining various dimensions of performance and coherence across different forecast providers.

The remainder of the paper is set out as follows. Section 2 presents the proposed coherence and performance measures used in the framework. Section 3 describes an example of the practical application of the framework. Finally, Section 4 offers a concluding discussion and directions for future work.

2. Performance and coherence measures

A major advantage of the MSE (which examines squared errors) is that it can be decomposed into a number of underlying components of the accuracy more easily than can measures based on absolute errors. These decompositions are shown to yield unique aspects of performance, such as over/under-forecasting and discrimination skills (Pollock et al., 2005; Pollock & Wilkie, 1996; Theil, 1966; Thomson, Pollock, Henriksen, & Macaulay, 2004; Wilkie & Pollock, 1996; Yates, 1982). Furthermore, corrections can easily be made to obtain expected values for the MSE that incorporate noise when using simulated data, as demonstrated by Pollock, Macaulay, Önkal-Atay, and Thomson (1999). As has been mentioned, the MSE is used in this study not only in the context of a measure of the general accuracy (termed the mean squared error for performance, MSEP), but also as a measure of the coherence, reflecting the degree of consistency or agreement between sets of paired forecasts (termed the mean squared error for coherence, MSEC), expanding the consistency analysis set out by Thomson, Pollock, Gönül, and Önkal (2013). It is further illustrated that the MSEP for composite forecasts can be obtained statistically from the MSEP for individual forecasts and the MSEC between pairs of individual forecasts. This is extended to component performance measures of the bias squared (measuring under-/overestimation), resolution variation (measuring discrimination ability) and error variation (measuring variation in actual values not explained by variation in forecasts).

We use the following notation to facilitate the presentation of the proposed performance and coherence analysis. Forecasts are denoted by f_{ij}, where i denotes the forecaster ($i = 1,2,\ldots, n$) and j denotes the specific forecast period ($j = 1,2,\ldots k$). Composite forecasts (f_{mj}) for period j for all individuals ($i = 1,2,\ldots, n$) are obtained by taking the simple average of these individual forecasts for period j; that is, $f_{mj} = \frac{1}{n}\sum_{i=1}^{n} f_{ij}$, with the mean of the composite forecasts for all periods being given by $M(f_m) = \frac{1}{k}\sum_{j=1}^{k} f_{mj}$.

2.1. Performance measures

We evaluate performances by comparing forecasts with realized values. Ex-post, the actual measured value, denoted a_j, for the end of forecast period j ($j = 1,2,\ldots, k$)

will be known. The individual forecast f_{ij}, and its composite form f_{mj}, can be used in the performance analysis by comparing the values with this actual value, a_j, for period j.

When analyzing performance (and coherence), it is desirable to use hypothetical forecasters as standards of comparison. One such comparison benchmark is provided by the perfect forecaster (PF), who would make forecast changes that were precisely in line with the actual value, such that $f_{ij} = a_j$ for all j. Thus, it is not possible to perform better than the PF. Another benchmark is the constant value forecaster (CVF), who would make all predictions with a constant value, c_i, such that $f_{ij} = c_i$ for all j. The CVF would be appropriate for non-trending series and shows no variation in forecast values; thus, it displays no resolution in performance. These hypothetical forecasters are used in our discussion of the performance measures below.

2.1.1. The mean squared error for performance

Performance can be measured by the mean squared error for performance (MSEP), which is essentially the average of the squared forecast errors, where the forecast error is measured as the forecast value minus the actual value. This provides an overall performance measure for both individual forecasts (MSEPI_i) and composite forecasts (MSEPM over all j forecasts ($j = 1,2,\ldots, k$)), as defined in Eqs. (1a) and (1b) respectively:

$$MSEPI_i = \frac{1}{k}\sum_{j=1}^{k}(f_{ij} - a_j)^2 \tag{1a}$$

$$MSEPM = \frac{1}{k}\sum_{j=1}^{k}(f_{mj} - a_j)^2. \tag{1b}$$

A value of zero would imply that the forecast values are identical to the actual values (indicating perfect accuracy); hence, the higher the value of the MSEP, the poorer the forecast performance. In the case of the perfect forecaster, $f_{ij} = a_j$ for all j, such that $\text{MSEPI}_i = 0$. For the constant value forecaster, the predictions would be $f_{ij} = c_i$ for all j, and hence $\text{MSEPI}_i = V(a) + (c_i - M(a))^2$, where $M(a)$ and $V(a)$ denote the mean and variance of the actual values, respectively.

The MSEP is an overall performance measure which can be decomposed in order to identify specific components that reflect the multidimensional aspects of the accuracy. The decomposition that is used in the present study involves the bias squared for performance (BSP), the resolution variation for performance (RVP) and the error variation for performance (EVP). The MSEP decompositions are

$$MSEPI_i = BSPI_i + RVPI_i + EVPI_i \tag{2a}$$

$$MSEPM = BSPM + RVPM + EVPM \tag{2b}$$

These three components are discussed next.

2.1.2. Bias squared for performance

The bias (B) is measured by the difference between the mean of forecast values, $M(f)$, and the mean of the actual values, $M(a)$. Biases occur when the mean forecast value is either too low, reflecting underestimation of the average actual values ($B < 0$), or too high, reflecting overestimation of the average actual values ($B > 0$). The bias squared value for the individual forecaster is simply the square of the bias, and is a specific component of the MSEP decomposition. The perfect forecaster would have zero bias and the constant value forecaster would have a bias that was equal to the difference between the constant value (c) and the mean of the actual values. The bias squared value for the composite forecaster is the sum of the biases for individual forecasts squared, divided by the square of the number of forecasters.

The bias squared for performance (BSP) is defined as follows for the individual and composite forecasters, respectively:

$$BSPI_i = {B_i}^2 \tag{3a}$$

$$BSPM = {B_m}^2 = \frac{1}{n^2}(\sum_{i=1}^{n} B_i)^2, \tag{3b}$$

where

$$B_i = M(f_i) - M(a) \text{ and } B_m = M(f_m) - M(a),$$

$$M(a) = \frac{1}{k}\sum_{j=1}^{k} a_j \quad \text{and } M(f_i) = \frac{1}{k}\sum_{j=1}^{k} f_{ij} \quad \text{and}$$

$$M(f_m) = \frac{1}{n}\sum_{i=1}^{n} M(f_i).$$

2.1.3. Resolution variation for performance

The resolution variation component of performance (RVP) for individual and composite forecasters is related to the resolution or slope (SL_i or SL_m), which is a measure of discrimination that reflects the ability to detect a one-point change in the actual value and make an appropriate increment in the forecast. The resolution is measured by the slope coefficient from a linear relationship between the forecast and the actual values, and is of particular importance in situations where the forecaster needs to identify the direction of movement in the actual series under consideration correctly, and to discriminate between large and small movements. This is a critical aspect of performance that reveals a forecaster's level of expertise. The composite forecast resolution value is the average of the individual resolution values; hence, composite forecasts would not show any improvement in average resolution compared with the individual forecasts. Thus, the resolution, which is probably the most important aspect of performance, cannot be improved by averaging forecasts. This resolution (slope) term is unity for the perfect forecaster and zero for the constant value forecaster.

The resolution variation component of performance (RVP) is the square of unity minus the slope multiplied by the variance of the actual values, $V(a)$, and approaches zero as the resolution approaches unity. The resolution can be negative in cases where the forecaster's performance is worse than that of the constant value forecaster. The RVP terms for both the individual and composite forecasters are zero for the perfect forecaster and $V(a)$ for the constant value forecaster. The composite RVP is the square of unity

less the average of the individual slope terms, with the result being multiplied by the variance of the actual values.

The resolution variation for performance (RVP) is defined as follows for the individual and composite forecasters, respectively:

$$RVPI_i = (1 - SL_i)^2 V(a) \tag{4a}$$

$$RVPM = (1 - SL_m)^2 V(a), \tag{4b}$$

where

$SL_i = C(f_i, a)/V(a)$ and

$$SL_m = \frac{1}{n}\sum_{i=1}^{n} SL_i \quad V(a) = (\frac{1}{k}\sum_{j=1}^{k} a_j^2) - M(a)^2$$

$$C(f_i, a) = (\frac{1}{k}\sum_{j=1}^{k} f_{ij}a_j) - M(f_i)M(a).$$

2.1.4. Error variation for performance

The error variation for performance (EVP) is the variation in the forecast values that is not explained by the variation in the actual values. The error variation for individual forecasters is measured by the scatter term (SC_i) about the fitted simple regression of the forecast values (f_j) on a_j. Error variation can arise when forecasters use diverse strategies for forming their predictions or identify patterns in the series that are not relevant. The error variation is zero for both the perfect and constant value forecasters. The error variation for the composite forecaster is measured by the composite scatter (SC_m), which is the variance of the sum of the regression error terms divided by the square of n.

The error variation for performance (EVP) is defined as follows for the individual and composite forecasters:

$$EVPI_i = SC_i = V(u_i) \tag{5a}$$

$$EVPM = SC_m = \frac{1}{n^2} V(\sum_{i=1}^{n} u_i), \tag{5b}$$

where

$$u_{ij} = f_{ij} - A_i - SL_i a_j \quad \text{and } V(u_i) = V(f_i) - {SL_i}^2 V(a)$$

$$A_i = M(f_i) - SL_i M(a).$$

2.2. Measures of coherence

The forecasts of forecaster h ($h = 1,2,\ldots, n-1$), denoted by f_{hj}, can be compared with those of forecaster i ($i = 2,3,\ldots, n$, $i > h$), denoted by f_{ij}, in order to examine coherence, which is a measure of the consistency or agreement between the predictions. When all predictions are perfectly coherent for a specific period, j, the values should all be equal, i.e., $f_{hj} = f_{ij}$. The situations in which the values are not equal reflect a degree of diversity. The forecasts for all j periods can be compared for each pair of forecasters, h and i, in order to provide a measure for each pair of forecasts. There will be $n(n-1)/2$ sets of paired values in total for n forecasters. Coherence can be measured by a range of statistics that have forms similar to those used in the performance analysis.

2.2.1. The mean squared error for coherence

The mean squared error for coherence (MSEC) is an overall measure of the coherence obtained from two forecasts f_{hj} and f_{ij}. The MSEC between each pair of forecasters, h and i, over all j periods ($j = 1,2,\ldots, k$), is defined as

$$MSEC_{hi} = \frac{1}{k}\sum_{j=1}^{k} (f_{hj} - f_{ij})^2. \tag{6}$$

A value of zero would imply that the two individuals h and i have made identical predictions, and therefore are perfectly coherent.

As with MSEP, the MSEC can be decomposed so as to identify specific aspects of the coherence between two forecasters, h and i. This decomposition involves the bias squared for coherence (BSC), the resolution variation for coherence (RVC) and the error variation for coherence (EVC), and is written as

$$MSEC_{hi} = BSC_{hi} + RVC_{hi} + EVC_{hi}. \tag{7}$$

We discuss these three measures next.

2.2.2. Bias squared for coherence

The bias squared for coherence (BSC_{hi}) is the squared difference between the mean forecasts of the two forecasters (h and i). A value of zero for this measure indicates coherence in the means, ($M(f_h) - M(f_i)$). Bias (without squaring) is used as a measure of whether the forecasters give different mean predictions. For example, if $M(f_h)$ is less than $M(f_i)$, this indicates that forecaster h has generally given lower predicted values than forecaster i, which would indicate a negative coherence bias between h and i. On the other hand, if $M(f_h)$ is greater than $M(f_i)$, this indicates a positive coherence bias between h and i. The bias squared for coherence is the squared difference between the two forecasters on the bias for performance measure, which is a specific component of the MSEC decomposition.

The bias squared for coherence (BSC) is defined as

$$BSC_{hi} = [M(f_h) - M(f_i)]^2 = (B_h - B_i)^2, \tag{8}$$

where

$$M(f_h) = \frac{1}{k}\sum_{j=1}^{k} f_{hj} \quad \text{and} \quad B_h = M(f_h) - M(a).$$

2.2.3. Resolution variation for coherence

The resolution variation for coherence (RVC_{hi}) is the square of the difference between the two resolution terms (SL_h and SL_i) of two forecasters (h and i), multiplied by the variance of the actual values $V(a)$. A value of zero for this measure indicates coherence in resolution (slope). Nonzero values of this measure reflect a degree of diversity between the resolution performances of the two forecasters.

The resolution variation for coherence (RVC) is defined as

$$RVC_{hi} = (SL_h - SL_i)^2 V(a), \tag{9}$$

where

$$SL_h = C(f_h, a)/V(a).$$

2.2.4. Error variation for coherence

The error variation for coherence (EVC_{hi}) reflects the difference in the variances of the two dependent errors, or scatter ($SC_h + SC_i - 2SC_{hi}$), between two forecasters (h and i). In other words, the EVC_{hi} measures the coherence between the error variations for performance of the two forecasters. Higher values of the error variation for performance (scatter) lead to higher values of EVC_{hi}, depending on the offsetting effect of the scatter covariance. The error variation for coherence is defined as

$$EVC_{hi} = SC_h + SC_i - 2SC_{hi}, \tag{10}$$

where

$$SC_i = V(u_i) \text{ and } SC_h = V(u_h)$$

$$SC_{hi} = C(u_h, u_i) = (\frac{1}{k}\sum_{j=1}^{k} u_{hj}u_{ij}) = C(f_h, f_i) - SL_hSL_iV(a) \quad (\text{note:} M(u_h) = M(u_i) = 0)$$

$$u_{hj} = f_{hj} - A_h - SL_ha_h \text{ and } V(u_h) = V(f_h) - SL_h{}^2V(a)$$

$$u_{ij} = f_{ij} - A_i - SL_ia_j \text{ and } V(u_i) = V(f_i) - SL_i{}^2V(a)$$

$$A_h = M(f_h) - SL_hM(a) \text{ and } A_i = M(f_i) - SL_iM(a)$$

$$C(f_h, f_i) = (\frac{1}{k}\sum_{j=1}^{k} f_{hj}f_{ij}) - M(f_h)M(f_i).$$

2.3. Linking performance and coherence measures

A link exists between the performance measures for individual composite forecasts and the coherence measures between the individual pairs of forecasts. Performance measures for composite forecasts can be obtained directly from the performance measures of individual forecasters and the coherence measures, as follows:

$$MSEPM = \frac{1}{n}\sum_{i=1}^{n} MSPEI_i - \frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} MSEC_{hi}. \tag{11}$$

Eq. (11) shows that the MSE for performance for the composite forecaster (i.e., MSEPM) is the sum of the individual MSEs, $MSEPI_i$ for all i forecasters, divided by n less the sum of the MSE for coherence, $MSEC_{hi}$, for all pairs of forecasters (h and i), divided by the square of n. Thus, the composite MSEPM measure will be less than the sum of the individual MSEPI measures, provided that a degree of diversity exists, as measured by the MSEC for any pair of forecasters. Only in the case where all individual forecasters are perfectly coherent with each other will the MSEPs for composite and individual forecasters be the same.

Similarly formed equations apply to the components of the MSE for performance and coherence involving the squared bias, resolution variation and error variation:

$$BSPM = \frac{1}{n}\sum_{i=1}^{n} BSPI_i - \frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} BSC_{hi} \tag{12a}$$

$$RVPM = \frac{1}{n}\sum_{i=1}^{n} RVPI_i - \frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} RVC_{hi} \tag{12b}$$

$$EVPM = \frac{1}{n}\sum_{i=1}^{n} EVPI_i - \frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} EVC_{hi}. \tag{12c}$$

These equations show that the components of the MSE for performance for the composite forecaster, namely bias squared (BSPM), resolution variation (RVPM) and error variation (ERVPM), are the sum of the respective individual component values for all i forecasters (BSPI, RVPI and EVPI) divided by n less the sum of the paired component values for coherence (BSC, RVC and EVC) divided by the square of n. Thus, all three composite performance component measures will always be less than the sum of the respective performance component individual measures when any degree of diversity exists, as measured by the respective coherence measures for that component, between the pairs of forecasters.

2.4. The relative percentage improvement of composite forecasts

The relative improvement for the composite forecasters compared with the average of the individual forecasts can be obtained using the paired coherence measures. This is demonstrated using Eqs. (11), (12a), (12b) and (12c), by dividing the second term in each equation by the first term and then multiplying the result by 100 to give values in percentage terms. This gives the following relative measures:

$$RMSEPM = 100 * (\frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} MSEC_{hi}) / (\frac{1}{n}\sum_{i=1}^{n} MSEPI_i) \tag{13a}$$

$$RBSPM = 100 * (\frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} BSC_{hi}) / (\frac{1}{n}\sum_{i=1}^{n} BSPI_i) \tag{13b}$$

$$RRVPM = 100 * (\frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} RVC_{hi}) / (\frac{1}{n}\sum_{i=1}^{n} RVPI_i) \tag{13c}$$

$$REVPM = 100 * (\frac{1}{n^2}\sum_{h=1}^{n-1}\sum_{i=2,i>h}^{n} EVC_{hi}) / (\frac{1}{n}\sum_{i=1}^{n} EVPI_i), \tag{13d}$$

which show the relative percentage improvements of composite measures for the MSEP and its components, where RMSEPM is the composite forecasts' relative percentage MSE for performance, and RBSPM, RRVPM and REVPM give similar relative percentage performance values for the bias squared, resolution variation and error variation. These relative measures give the percentage improvements obtained by using composite forecasts relative to simply using the average of the individual forecasts. The higher the percentage value, the greater the improvement. Alternatively, these measures can be obtained from the values of the relevant MSEPI and the MSEPM, and from the values of their components.

Table 1
Actual and forecast yearly Q4 RPI percent inflation values and forecast errors.

Q4 Year	Actual	Forecasts				Forecast errors			
		F1	F2	F3	F4	F1	F2	F3	F4
1998	3.0	2.6	3.1	2.9	3.6	−0.4	0.1	−0.1	0.6
1999	1.5	1.5	1.5	1.3	1.6	0.0	0.0	−0.2	0.1
2000	3.1	2.9	2.5	3.1	3.4	−0.2	−0.6	0.0	0.3
2001	1.0	2.2	2.3	2.0	2.4	1.2	1.3	1.0	1.4
2002	2.5	2.4	2.5	2.6	2.1	−0.1	0.0	0.1	−0.4
2003	2.6	2.8	2.7	2.6	3.0	0.2	0.1	0.0	0.4
2004	3.4	3.4	2.7	2.5	3.0	0.0	−0.7	−0.9	−0.4
2005	2.4	2.9	2.6	2.8	2.0	0.5	0.2	0.4	−0.4
2006	4.0	2.9	1.9	1.5	2.5	−1.1	−2.1	−2.5	−1.5
2007	4.2	3.0	3.5	2.1	2.3	−1.2	−0.7	−2.1	−1.9
2008	2.7	2.1	2.1	1.7	2.2	−0.6	−0.6	−1.0	−0.5
2009	0.6	−0.4	−1.5	−1.5	−0.2	−1.0	−2.1	−2.1	−0.8
2010	4.7	3.4	2.9	4.7	2.8	−1.3	−1.8	0.0	−1.9
2011	5.1	3.9	3.5	2.8	3.5	−1.2	−1.6	−2.3	−1.6
2012	3.1	3.3	3.6	3.2	3.3	0.2	0.5	0.1	0.2
2013	2.6	3.0	3.3	2.5	2.3	0.4	0.7	−0.1	−0.3
2014	1.9	2.7	3.5	2.8	2.3	0.8	1.6	0.9	0.4
Mean	2.8	2.6	2.5	2.3	2.5	−0.2	−0.3	−0.5	−0.4
SD	1.2	1.0	1.2	1.3	0.9	0.8	1.1	1.1	0.9

3. A demonstration of the framework

We illustrate the above framework and the interpretation of its measures by presenting an analysis using published retail prices index (RPI) inflation forecasts that are compared with actual RPI values. The evaluation of inflation forecasts is crucial to the activities and decisions of governmental, commercial and financial institutions, and provides an appropriate context for demonstrating the proposed analytical framework.

3.1. The data

The forecast data used in this illustration involved Q4 RPI annual inflation rate forecasts, obtained from HM Treasury's forecasts for the UK economy;[1] specifically, four sets of City RPI forecasts, available in the December editions for Q4 of the following year. The editions used were dated December 1997 to December 2013, which provided forecasts for 1998Q4–2014Q4. The four City forecasters used were Barclays Capital, Credit Swiss, HSBC and Morgan Stanley, which provided forecasts for the whole period (i.e., 17 forecasts) and are denoted Forecasters 1 to 4 (F1 to F4) respectively in the analysis that follows. In two cases (F2 forecasts for 2004 and 2005), data were not available in the December edition, so data from the November and October editions were used instead. The forecasts were compared with the realized Q4 values of the retail prices index obtained from the Office of National Statistics (code CZBH).[2] The actual and forecast values for yearly Q4 RPI values are presented in Table 1.

[1] Forecasts for the UK economy, available monthly since September 1997, from http://www.hm-treasury.gov.uk/data_forecasts_index.html (accessed 24.05.2017).

[2] Retail prices index (code CZBH), from http://www.ons.gov.uk/ons/datasets-and-tables/data-selector.html?cdid=CZBH&dataset=mm23&table-id=1.2%20%20 (accessed 24.05.2017).

Table 2
Correlations between forecast errors.

Forecaster	F1	F2	F3	F4
F1	1			
F2	0.904	1		
F3	0.816	0.812	1	
F4	0.850	0.777	0.739	1

3.2. Characteristics of the data and forecast errors

The actual RPI inflation data for 1998Q4–2014Q4, presented in Table 1, had a range of values from 0.6 to 5.1, with a mean of 2.8 and sample standard deviation of 1.2. The forecasts from all four forecasters had slightly lower mean values, resulting in negative forecast errors, which reflects a general tendency to underestimate the inflation rate. The forecast sample standard deviations for all forecasters were generally in line with the actual sample standard deviations, and the forecast error sample standard deviations were relatively similar. Two forecasters, F1 and F4, had forecast errors that were less than 2% for all 17 years. However, there were negative forecast errors of over 2% for F2 and F3 in 2006 and 2009 and for F3 in 2007 and 2011, reflecting a considerable underestimation of inflation in these years.

The correlations in the forecast errors between pairs of forecasters are presented in Table 2. These values were positive and large, with F1 and F2 showing the highest correlations. The smallest correlation was between F3 and F4. These large correlation values, along with the differences between them, have implications for combining forecasts and emphasize the importance of performance and coherence analysis.

3.3. Interpretation of the results from the performance and coherence measures

The results presented in Tables 3–5 show the values calculated for the measures set out in Section 2 using the

Table 3
Performance measures: individual and composite for sets of forecasters.

Measure	MSEP	BiasSqP	Bias	Mean	ResVarP	Slope	ErrVarP	Variance
F1	0.584	0.050	−0.224	2.624	0.202	0.620	0.331	0.868
F2	1.257	0.112	−0.335	2.512	0.253	0.575	0.892	1.354
F3	1.425	0.268	−0.518	2.329	0.207	0.615	0.950	1.479
F4	0.957	0.137	−0.371	2.476	0.378	0.480	0.441	0.763
C1234	0.899	0.131	−0.362	2.485	0.256	0.572	0.512	0.970
C123	0.951	0.129	−0.359	2.488	0.220	0.603	0.602	1.110
C124	0.823	0.096	−0.310	2.537	0.273	0.558	0.454	0.889
C134	0.843	0.137	−0.371	2.476	0.257	0.571	0.449	0.906
C234	1.048	0.166	−0.408	2.439	0.275	0.557	0.607	1.040
C12	0.851	0.078	−0.279	2.568	0.227	0.597	0.546	1.045
C13	0.880	0.137	−0.371	2.476	0.205	0.617	0.538	1.071
C14	0.708	0.088	−0.297	2.550	0.284	0.550	0.336	0.758
C23	1.224	0.182	−0.426	2.421	0.229	0.595	0.813	1.308
C24	0.992	0.125	−0.353	2.494	0.312	0.527	0.555	0.944
C34	1.051	0.197	−0.444	2.403	0.286	0.547	0.567	0.986
PF	0.000	0.000	0.000	2.847	0.000	1.000	0.000	1.398
RWF	2.622	0.011	0.106	2.953	1.239	0.058	1.372	1.377
CVF(2)	2.115	0.718	−0.847	2.000	1.398	0.000	0.000	0.000
CVF(3)	1.421	0.023	0.153	3.000	1.398	0.000	0.000	0.000
CVF(4)	2.727	1.329	1.153	4.000	1.398	0.000	0.000	0.000

Table 4
Coherence measures between pairs of forecasts.

Measure	F1,F2	F1,F3	F1,F4	F2,F3	F2,F4	F3,F4
MSEC	0.277	0.495	0.251	0.466	0.460	0.559
BiasSqC	0.012	0.087	0.022	0.033	0.001	0.022
ResVarC	0.003	0.000	0.027	0.002	0.013	0.026
ErrVarC	0.262	0.409	0.202	0.431	0.446	0.512

Table 5
Relative percentage improvement of composite forecasts on performance.

Measure	RMSEP	RBiasSqP	RResVarP	RErrVarP
C1234	15	8	2	22
C123	13	10	0	17
C124	12	4	6	18
C134	15	10	2	22
C234	14	4	2	20
C12	8	4	0	11
C13	12	14	0	16
C14	8	6	2	13
C23	9	4	0	12
C24	10	0	1	17
C34	12	3	2	18
Mean	11	6	2	17

data presented in Table 1. This involves the four individual sets of forecasters (F1 to F4) and all possible composites (i.e., using all four forecasters [C1234], four sets of three forecasters [C123, C124, C134, C234], and six sets of two forecasters [C12, C13, C14, C23, C24, C34]). Table 3 presents the performance analysis results for the MSE (MSEP) and its components, the squared bias (BSP), resolution variation (RVP) and error variation (EVP), for the individual and combined forecasts. In addition, supplementary statistics are provided for the mean, variance, bias and resolution. Results are also given for the perfect forecaster (PF) and three constant value forecasters, reflecting constant inflation rate predictions of 2%, 3% and 4%, denoted CVF(2), CVF(3) and CVF(4), respectively. The choice of 2%, 3% and 4% is appropriate because only a small number of the forecasts for F1 to F4 over the 17-year period fell outside this range (two for F1, three for F2, five for F3 and two for F4). Table 3 also presents comparison results using the random walk forecaster (RWF), which simply uses the actual, Q4, value for the year prior to the forecast year. Table 4 presents the coherence results for the paired sets of forecasters using the MSE (MSEC) and its components, the squared bias (BSC), resolution variation (RVC) and error variation (EVC). Table 5 presents the relative percentage performance improvements of the composite forecasts compared with the average of the individual forecasts for the MSEP (RMSEP) and its components, denoted RBSP, RRVP and REVP. These results are discussed below.

3.3.1. Mean squared error

The MSE for performance (MSEP) measures the overall performance, enabling the best inflation forecasters over the period to be identified. Table 3 shows that the best values from the individual forecasters occurred for F1, followed by F4 and F2. The best value on this measure is zero, which is the value for the perfect forecaster. All four forecasters, and therefore all of the composite forecasters, had better values than the random walk forecaster and the three constant value forecasters, except for F3, which had a value that was slightly poorer than that of the 3% constant value forecaster. Table 4 shows that the overall paired coherence measure, MSEC, exhibited some diversity, although it was relatively low for F1 with respect to F2 and F4. Table 5 shows that the largest relative percentage improvement for composite forecasts occurred for combinations involving F3 and F4, with the composites C1234 (15%), C134 (15%), C234 (14%) and C34 (12%) indicating the relatively high level of diversity between F3 and F4. Overall, combining forecasts improved the performances, with the average improvement being 11%. As the MSE for performance shows clear differences between the four individual forecasters, the improvement in the composite forecasts resulted in only one case, C23, in which the composite improved on both of the relevant individual values, F2 and F3.

3.3.2. Bias squared

The bias squared and bias are important where one requires forecasts that do not show a general over-/under-forecasting of the rate of inflation over the longer term. This is particularly relevant for situations where long-term funding shortfalls could build up as a result of persistent under-/overestimation of yearly inflation rates. If the bias is small, any underfunding that may occur in some years will be offset by overfunding in other years. However, in this case, the bias was negative for all four forecasters, reflecting a general underestimation of inflation over the whole period. The best value under this measure is zero, which is the value for the perfect forecaster. Table 3 shows that the lowest bias (in absolute terms) occurred for F1, followed by F2 and F4, with the lowest bias for composite forecasts occurring for those that involved F1. The bias squared for performance gave a similar ordering, with all of the individual and composite forecasts being worse than both the random walk forecaster and the 3% constant value forecaster, but better than the 2% and 4% constant value forecasters. Table 4 shows that the paired bias squared for coherence values were relatively small, with the exception of the F1, F3 pair. Table 5 shows that the largest relative percentage improvement for composite forecasts (14%) occurred for the composite C13, indicating the relatively low coherence between F1 (which showed the lowest bias squared) and F3 (which showed the highest bias squared). This also partly explains the relatively large percentage improvements for the composite forecasts C123 (10%), C134 (10%) and C1234 (8%). Thus, the composite forecasts showed some improvement on the squared bias, with an average percentage improvement of 6%.

3.3.3. Resolution variation

In the present context, resolution variation and resolution are important where one requires forecasts that can successfully identify and distinguish between years when high or low inflation is likely to occur. This is relevant to situations where there is a need to identify years in which large changes in the inflation rate are likely, so that financial planning can be organized to account for the impact of these movements. With resolution, the best possible value (the case of the perfect forecaster) is unity, with all of the constant-value forecasters having a value of zero. Table 3 shows that the random walk forecaster performed considerably worse than all of the individual and composite forecasters. The best values on resolution occurred for F1, followed closely by F3, with F4 showing the poorest value. However, the resolution cannot be improved by taking combined forecasts, as composite resolution values are just the average of individual resolution values. Resolution is a key variable in the resolution variation for performance, which has a best possible value of zero, the value for the perfect forecaster, and its value for the three constant value forecasters is equal to the variance of the actual values ($V(a) = 1.398$). Table 4 shows that the paired values for the resolution variation for coherence were relatively small and almost zero for pairs that did not include F4, which had the poorest resolution. Table 5 shows that the relative percentage improvement in resolution variation for composite forecasts was almost zero for all composites that did not contain F4. Thus, composite forecasts showed only marginal improvements overall compared with individual forecasts, with average relative percentage improvements of 2%. For larger relative percentage improvements, it would be necessary to have considerable differences in the resolution values between individual forecasts.

3.3.4. Error variation

Error variation is important when forecasts with low levels of unexplained variation in the rate of inflation are required. The poorer the performance on the error variation, the greater the size of any financial provisions that may be necessary to compensate for unpredicted movements in the inflation rate. Table 3 shows that the lowest error variation for performance (EVP) for the individual forecasts occurred for F1, followed by F4. F2 and F3 showed relatively poor performances. All of the combined forecasts were worse than F1, but better than both F2 and F3. The best values occurred for composite forecasts involving F1 and F4. The best value on this measure is zero, the value for the perfect forecaster. All three constant value forecasters had values of zero, and the random walk forecaster had a much poorer performance than all of the other forecasters. Table 4 shows that the paired values for the error variation for coherence exhibit some diversity, which was relatively low for F1 with respect to F2 and F4. The paired values show that this component had a dominant effect on the MSE for coherence. Table 5 shows that the largest relative percentage improvement occurred for the composites C1234 (22%), C134 (22%) and C234 (20%), indicating the relatively high diversity between F2 and F4 (C24 (17%)) and F3 and F4 (C34 (18%)). Combining forecasts improved the overall performance, with the average improvement being 17%. In fact, this improvement level is better than those reported in the literature (Armstrong, 2001; Batchelor & Dua, 1995). Batchelor and Dua (1995) asserted that forecast combinations reduced the error by an average of between 9.2% (for combination across two forecasters) and 16.4% (for combination across 10 forecasters), while Armstrong (2001), in a review of 30 studies, found the reduction in ex-ante errors to average around 12.5%. The improvement in composite forecasts relative to the mean of individual forecasts resulted in only one case (C23) where the composite improved on the relevant individual error variation for performance values (F2 and F3).

3.3.5. Consolidation of the results

The improvement for combined forecasts relative to the mean of individual measures, as measured by the MSE for performance, was due mainly to lower values of the error variation and bias squared, with the resolution variation having a smaller effect. Forecaster F1 performed best on the performance measures for MSE and all of its components. In situations where there are clear differences between the values of the individual forecasts, the combined forecaster is unlikely to be better than every individual forecaster. Composite forecasts tend to give better performances than the individual forecasts when the individual forecast measures are reasonably close together and the respective coherence measures show reasonably high levels of diversity. However, composite forecasts may not improve on the best

individual forecast if an individual forecast is distinctly better than other individual forecasts on a given performance measure, particularly when there is a reasonable degree of coherence between them.

4. Discussion

This paper sets out an analytical framework that can be used for enhancing the assessment of forecast performances and guiding decisions as to which forecasters should be pooled in order to obtain an effective combined forecast. It is shown that composite forecasts (formed using a simple average) have lower mean squared errors than those of the averages of the individual forecasts. This result supports previous research (e.g., Armstrong, 2001; Clemen, 1989; De Menezes et al., 2000; Lawrence et al., 1986; Makridakis & Hibon, 2000; Schnaars, 1986; Soll & Larrick, 2009; Stock & Watson, 2004; Timmerman, 2006) and confirms the benefits of forecast combination. In expanding this work further, the framework makes a critical contribution in terms of its direct incorporation of a set of coherence measures that address the consistency and dependence among forecasts. The use of coherence enables analysts to combine forecasts, not only by using the best forecasters, but also by accounting for the levels of diversity among them, so as to achieve the best possible accuracy from a group of component forecasts. Coherence measures tend to be more stable over time than performance measures, as they do not involve any consideration of the actual values of the series under consideration, and therefore changes are likely to occur only when forecasters switch their forecasting models or procedures.

The framework also makes an important contribution to the research by decomposing the performance and coherence in an integrated framework in order to illustrate the underlying aspects that are responsible for the error-reduction that is evidenced in composite predictions. This integrated approach provides more information than performance analysis alone, and could be extremely useful for improving the forecast accuracy. Not only do the overall measures provide important insights into the assessment of groups of forecasters that share similar techniques, but also the component measures can be used to identify subtle differences in the levels of performance and coherence among these forecasters. This can be especially important when forecasts are based on judgmental inputs and forecasters show, for example, general overall coherence, but a degree of diversity on a specific element.

The paper consolidates the results from previous studies (e.g., Armstrong, 2001; Armstrong, Green, & Graefe, 2015; Goodwin, 2015; Graefe, Armstrong, Jones, & Cuzan, 2014; Green, Armstrong, & Graefe, 2015) and implies that the extent of the improvements may be related directly to the degree of diversity between pairs of individual predictions. The higher the diversity indicated between the individuals on each component, the greater the improvement on that component for combined forecasts.

The current results have important practical implications, as they suggest that composite forecasts can be improved by excluding forecasts, not only in terms of their relative performances, but also on the basis of relative coherence. The findings imply that composite forecasts should be obtained by pooling heterogeneous forecasters that show effective performances in specialized aspects that target a customized use of the pooled forecasts. This is important, as it is not always possible to control the coherence by selecting diverse forecasts. Thus, it is useful to incorporate measures of coherence explicitly into the selection and formation of composite forecasts. The findings also imply that if a weighting procedure is used, then the levels of coherence between individual forecasters need to be taken into account in addition to the performance, when setting the values for the relative weights.

The framework can be used to aid in the identification of good/superior forecasters who can be included in the formation of composite forecasts, while filtering out other (sub-standard/not-so-good) forecasters, as guided by coherence comparisons among the forecasters. This process can be particularly effective in conditions where forecasting procedures are not changing significantly over time and no significant structural changes are occurring. The coherence measures can be used to detect changes in forecast method use among forecasters effectively. As they do not require realized values in their computation, coherence measures will tend to be more stable over time than performance measures, with any changes being likely to occur only when forecasters switch their forecasting models or procedures. An awareness of multidimensional aspects of accuracy can also help to identify structural changes. For instance, higher error variation measures can be associated with a failure of forecasters to identify increases in volatility. The use of shorter rolling sample periods over the whole period can assist in the identification of these conditions, although this process would require a sufficiently large amount of data.

The multidimensional aspects of performance and coherence, identified by bias, resolution and error variation, have important implications for the subsequent use of forecasts. For instance, if the focus is on correctly identifying the direction of large movements in the relevant variable and distinguishing between large and small movements, it would be particularly important to include forecasters who display good resolution or resolution variation levels and exclude forecasters who show poor resolutions, directing less attention to other components. If we are interested in having forecasts that do not persistently over- or underestimate the variable under consideration, it is particularly important to include forecasters with low bias and bias squared values, excluding those with high values. If we are interested in having forecasts with less unexplained variability, then it is appropriate to include forecasters with low error variations and omit those with high error variations. However, when examining these aspects, it is desirable to have sufficiently large amounts of data to be able to discover variations in forecasters' methods and detect possible structural changes in the series being forecast. This could be extended to consider how experts could be replaced by statistical models or even machine learning algorithms.

Studies from the neural networks context further reinforce these arguments. In particular, Ueda and Nakano (1996) demonstrated that the generalization error in a

machine learning system can be reduced by a combination or ensemble of outputs from multiple estimators. For this purpose, they utilized a decomposition of the ensemble generalization error into a variety of components (bias, variance, covariance and noise variance). Their analysis of this decomposition showed that when two estimators were 'negatively correlated', their combination led to a lower generalization error, whereas a positive correlation between different estimators led to a higher generalization error. Thus, there exists a relationship between the diversity and ensemble/combination performances in a machine learning system (e.g., Du Jardin, 2016; Florez-Lopez & Ramon-Jeronimo, 2015; Yin, Huang, Hao, Iqbal, & Wang, 2014; Zhang & Zhou, 2013): the greater the diversity among the base samples, the lower the error variance of their combination. These findings are in direct agreement with the results of the current work.

It is worth noting that the proposed framework faces various potential limitations. Our study only addresses measures that are associated with quadratic loss functions of the MSE form. Further research could extend the analyses to, for example, absolute loss functions of the MAE and MAPE. In addition, while our paper contains directions for aiding the selection of appropriate individual forecasts in composite forecast construction, there is the potential to extend this process further. For example, future research could examine criteria values/thresholds for performance and consistency in a range of practical settings, which could be used in decisions involving the inclusion and exclusion of forecasts for combination. The framework and the explanatory example have only considered the application in a within-sample situation. It would be useful to examine the stability of these measures in out-of-sample or rolling sample applications. In addition, the framework could also be extended to target prediction intervals and probability forecasts, which in turn would provide an informative glimpse into the forecasters' uncertainties and their communication to forecast users. Through this potential, the framework may assist the existing research in the probabilistic forecasting domain (e.g. Broomell & Budescu, 2009; Budescu & Rantilla, 2000; Budescu, Rantilla, Yu, & Karelitz, 2003; Budescu & Yu, 2007) that has attempted to unveil the link between the aggregation of expert predictions and decision-maker confidence.

This paper illustrates how extended performance analysis may be employed to enhance our understanding of pooled forecasts. When composite forecasts are required, the framework can help to determine the appropriate error measures for addressing coherence and other aspects of performance. The framework can be applied to promote the efficient use of individual forecasts across a diverse portfolio of selection criteria, as well as to provide a basis for a focused selection of individual forecasters for use in composite predictions. Yearly UK RPI inflation forecasts (for 1998–2017) from four banking institutions were employed to illustrate the application of the framework and the interpretation of its statistical measures. This example was chosen so as to allow the study to be replicated easily with readily-available published data, while illustrating the framework in a context that is highly relevant in the business, economic and government policy landscapes. The framework's application to inflation forecasting showed that the composite forecast improvements resulted primarily from a lower error variation (with relatively high paired coherence values), and to a lesser extent from lower bias squared values, with small improvements in the resolution-adjusted variation. Although the demonstration of the framework employed inflation forecasts for illustrative purposes, the measures are applicable to a wide range of other contexts in the finance, economic and business arenas.

To summarize, the framework provides a powerful diagnostic toolbox that can be used in a multitude of practical forecasting situations for improving forecasts when individual predictions are to be combined into a single forecast that is to be communicated to decision/policy-makers. It makes an important contribution to our understanding of the role of coherence, and highlights the role of group heterogeneity, whereby the combined performance improves demonstrably when individual forecasters bring in asymmetric information and expertise from diverse fields. These results have direct repercussions for collaborative forecasts across a wide variety of focal business domains, and will undoubtedly provide a rich platform for promising forecasting applications.

References

Armstrong, J. S. (2001). Combining forecasts. In J. S. Armstrong (Ed.), *Principles of forecasting: a handbook for researchers and practitioners* (pp. 417–439). Norwell, MA: Kluwer Academic Publishers.

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecast methods: empirical comparisons. *International Journal of Forecasting*, *8*, 69–80.

Armstrong, J. S., Green, K. C., & Graefe, A. (2015). Golden rule of forecasting: be conservative. *Journal of Business Research*, *68*, 1717–1731.

Batchelor, R., & Dua, P. (1995). Forecast 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.

Broomell, S. B., & Budescu, D. V. (2009). Why are experts correlated? Decomposing correlations between judges. *Psychometrika*, *74*(3), 531–553.

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398.

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194.

Budescu, D. V., & Yu, H. T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177.

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

De Menezes, L. M., Bunn, D. W., & Taylor, J. W. (2000). Review of guidelines for the use of combined forecasts. *European Journal of Operational Research*, *120*, 190–204.

Du Jardin, P. (2016). A two-stage classification technique for bankruptcy prediction. *European Journal of Operational Research*, *254*, 236–252.

Fifić, M., & Gigerenzer, G. (2014). Are two interviewers better than one? *Journal of Business Research*, *67*(8), 1771–1779.

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*, 227–246.

Florez-Lopez, R., & Ramon-Jeronimo, J. M. (2015). Enhancing accuracy and interpretability of ensemble strategies in credit risk assessment: A correlated-adjusted decision forest proposal. *Expert Systems with Applications*, *42*, 5737–5753.

Goodwin, P. (2015). Is a more liberal approach to conservatism needed in forecasting? *Journal of Business Research*, *68*, 1753–1754.

Graefe, A., Armstrong, J. S., Jr., Jones, R. J., & Cuzan, A. G. (2014). Combining forecasts: an application to elections. *International Journal of Forecasting*, *30*, 43–54.
Green, K. C., Armstrong, J. S., & Graefe, A. (2015). Golden rule of forecasting rearticulated: Forecast unto others as you would have them forecast unto you. *Journal of Business Research*, *68*, 1768–1771.
Hyndman, R., & Koehler, A. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*, 679–688.
Lawrence, M., Edmundson, R. H., & O'Connor, M. J. (1986). The accuracy of combining judgmental and statistical forecasts. *Management Science*, *32*, 1521–1532.
Leitner, J., & Leopold-Wildburger, U. (2011). Experiments on forecasting behavior with several sources of information – a review of the literature. *European Journal of Operational Research*, *213*, 459–469.
Makridakis, S., & Hibon, M. (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting*, *16*, 451–476.
Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology*, *12*, 595–600.
Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*, 2417–2424.
Pollock, A. C., Macaulay, A., Önkal-Atay, D., & Thomson, M. E. (1999). Evaluating predictive performance of judgemental extrapolations from simulated currency series. *European Journal of Operational Research*, *114*, 281–293.
Pollock, A. C., Macaulay, A., Thomson, M. E., & Önkal-Atay, D. (2005). Performance evaluation of judgemental directional exchange rate predictions. *International Journal of Forecasting*, *21*, 473–489.
Pollock, A. C., & Wilkie, M. E. (1996). The quality of bank forecasts: The dollar-pound exchange rate, 1990-1993. *European Journal of Operational Research*, *91*, 306–314.
Sanders, F. (1963). On subjective probability forecasting. *Journal of Applied Meteorology*, *2*, 191–201.
Schnaars, S. P. (1986). An evaluation of rules for selecting an extrapolative model of yearly sales forecasts. *Interfaces*, *16*, 100–107.
Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology – Learning Memory and Cognition*, *35*(3), 780–805.
Stewart, T. R., & Lusk, C. M. (1994). Seven components of judgmental forecasting skill: implications for research and the improvement of forecasts. *Journal of Forecasting*, *13*, 579–599.
Stock, J. H., & Watson, M. W. (2004). Combining forecasts of output growth in a seven-country data set. *Journal of Forecasting*, *23*, 405–430.
Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North Holland.
Thomson, M. E., Pollock, A. C., Gönül, M. S., & Önkal, D. (2013). Effects of trend strength and direction on performance and consistency in judgemental exchange rate forecasting. *International Journal of Forecasting*, *29*, 337–353.
Thomson, M. E., Pollock, A. C., Henriksen, K. B., & Macaulay, A. (2004). The influence of forecast horizon on the currency predictions of experts, novices and statistical models. *European Journal of Finance*, *10*, 290–307.
Timmerman, A. (2006). Forecast combinations. In G. Elliot, C. W. J. Granger, & A. Timmerman (Eds.), *Handbook of economic forecasting, vol. 1*. Amsterdam: North Holland.
Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. In *Proceedings of the international conference on neural networks* (pp. 90–95). Washington, DC, USA: IEEE Computer Society.
Wallis, K. F. (2011). Combining forecasts – forty years on. *Applied Financial Economics*, *21*, 33–41.
Wilkie, M. E., & Pollock, A. C. (1996). An application of probability judgement accuracy measures to currency forecasting. *International Journal of Forecasting*, *12*, 91–118.
Yates, J. F. (1982). External correspondence: decompositions of the mean probability score. *Organisational Behavior and Human Performance*, *30*, 132–156.
Yin, X.-C., Huang, K., Hao, H.-W., Iqbal, K., & Wang, Z.-B. (2014). A novel classifier ensemble method with sparsity and diversity. *Neurocomputing*, *134*, 214–221.
Zhang, M.-L., & Zhou, Z.-H. (2013). Exploiting unlabeled data to enhance ensemble diversity. *Data Mining and Knowledge Discovery*, *26*, 98–129.

Mary E. Thomson is Professsor of Decsion Science at Newcastle Business School. Her research focuses on financial judgmental forecasting, expert judgment and decision making in a variet of contexts, risk perception and risk assessment. She completed her PhD in 1999 and has since published in a wide variety of book and journals, including the *European Journal of Operational Research, the International Journal of Forecasting, Risk Analysis, Decision Support Systems, Journal of Behavioural Decision Making, Journal of Empirical Legal Studies, the International Journal of Forensic Mental Health, Legal and Criminological Psychology, Journal of Customer Behavior, and the European Management Journal*.

Andrew C. Pollock is an independent consultant who was formally a Reader at Glasgow Caledonian University. He completed his Ph.D. in 1988 and since then has published numerous articles in a variety of books and journals. His research interests focus on integrating economics, finance, statistics and psychology in the formation and evaluation of point and probability forecasts of price movements in currency, equity and economic series using judgemental, time series and econometric techniques.

Dilek Önkal is Professor of Decision Sciences at the Newcastle Business School, Northumbria University. Her research focuses on judgmental forecasting, judgment and decision making, forecasting/decision support systems, risk perception and risk communication with a strong emphasis on multi-disciplinary interactions. Professor Önkal is a Co-Editor of the *International Journal of Forecasting and her work work has appeared in journals such as Organizational Behavior and Human Decision Processes, Decision Sciences Journal, Judgment and Decision Making, Risk Analysis, International Journal of Forecasting, Decision Support Systems, Journal of Behavioral Decision Making, Journal of Forecasting, Omega: The International Journal of Management Science, Technological Forecasting & Social Change, Frontiers in Finance and Economics, and European Journal of Operational Research*.

M. Sinan Gönül is a Senior Lecturer of Business Analytics in Newcastle Business School at Northumbria University, UK. He has carried out research in the areas of judgmental forecasting, judgment & decision making, decision/forecasting support systems and behavioral operations research. He has published in various journals including *Journal of Behavioral Decision Making, Decision Sciences Journal, Decision Support Systems, International Journal of Forecasting, Journal of Forecasting, Technological Forecasting & Social Change and Production and Operations Management*.