**Combining Forecasts: Performance and Coherence**

**Mary E. Thomson[[1]](#footnote-1)**

Newcastle Business School,   
Northumbria University, City Campus East, Newcastle upon Tyne, UK

NE1 8ST

Tel: +44 (0)754 025 8280

email: [mary.thomson@northumbria.ac.uk](mailto:mary.thomson@northumbria.ac.uk)

**Andrew C. Pollock**

1 Pladda Avenue, Irvine, UK  
KA11 1DR

Tel: +44 (0)129 421 6618

email: acpollock@virginmedia.com

**Dilek Onkal**

Newcastle Business School,   
Northumbria University, City Campus East, Newcastle upon Tyne, UK

NE1 8ST

email: dilekon@gmail.com

**M. Sinan Gonul**

Newcastle Business School,   
Northumbria University, City Campus East, Newcastle upon Tyne, UK

NE1 8ST

Tel: +44 (0)191 227 3296

email: [sinan.gonul@northumbria.ac.uk](mailto:sinan.gonul@northumbria.ac.uk)

**Combining Forecasts: Performance and Coherence**

**Abstract**

In many forecasting contexts, there is general agreement that combining individual predictions leads to better final forecasts. Yet, the relative error reduction in a combined forecast is dependent upon the degree to which the component forecasts contain unique/independent information. Unfortunately, in many situations, obtaining independent predictions is difficult 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 degree of independence between sets of forecasts can easily be identified. The framework also decomposes the performance and coherence measures 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.

KEYWORDS: FORECAST, ACCURACY, COHERENCE, COMPOSITE FORECASTS, INFLATION

**Combining Forecasts: Performance and Coherence**

1. ***Introduction***

Over recent decades, considerable research attention has been given to combining 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 forecasting accuracy over selecting the prediction of the best forecaster (e.g., Fifić & Gigerenzer, 2014). In their seminal paper, Bates and Granger (1969) suggest 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 to establish 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 the corresponding ex-post actual values. There are a variety of error measures that can be used for assessing forecast performance, 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 the measures based on absolute error, a major advantage of the MSEis 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 and Wilkie (1996), Stewart and Lusk (1994) and Theil (1996) for MSE*;* and by Murphy (1973), Pollock, Macaulay, Thomson and Önkal (2005), Sanders (1963) and Yates (1982) for the Mean Probability Score).

Based on a review of 30 studies utilizing such evaluation measures, Armstrong (2001) reports a reduction in 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, this condition rarely occurs in practice. Forecast errors generally tend to be positively correlated, and often strongly so. To accommodate this typical lack of independence, researchers attempt to combine different types of forecasts (i.e., statistical and judgmental) based on a variety of forecasting methods and diverse information sets.

Unfortunately, obtaining independent forecasts is often difficult as comparable statistical models are typically used by the individual forecasters, or judgmental forecasts are made by individuals who use similar criteria and/or overlapping information. To address this problem directly, this paper proposes measures of coherence addressing consistency and dependence between forecasts that could effectively be integrated into the performance analysis of composite forecasts. The proposed framework demonstrates that composite forecasts will always have a smaller MSE compared with the average of the MSEs of the individual forecasts due to diversities between individual forecasts. The framework also illustrates that coherence measures can be used to quantify this improvement such that those forecasts with the highest level of independence can easily be identified.

The analytical framework makes an important contribution to extant work by decomposing performance and coherence using an integrated method to illustrate the underlying aspects that are responsible for the error reduction evidenced in composite predictions. Specifically, the *MSE* is not only employed as a traditional measure of accuracy or performance (denoted as the *mean squared error for performance - MSEP)*, but also as a measure to examine 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. *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. *Error variation* is the variation of 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 in improving forecast accuracy. Not only will 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 between the relative performance and coherence among these forecasters. At the same time, different aspects of forecasting performance may be prioritized over the others in specific situations (e.g. low levels of bias may be valued over a high resolution or vice versa), and being able to assess the components individually may become a highly beneficial capability.

To demonstrate the application of the framework and how its measures may be interpreted, inflation rate data are employed. Specifically, we use the *Retail Prices Index, RPI,* with forecasts obtained through the HM Treasury, *for the UK Economy* (monthly). Precisely we compare inflation forecasts from the December editions from 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 to examine 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.

1. ***Performance and Coherence Measures***

A major advantage of the MSE (which examines squared errors)is that it can be more easily decomposed into a number of underlying components of accuracy 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 et al, 2004; Wilkie & Pollock,1996; Yates,1982). Furthermore, corrections can easily be made to obtain expected values for the MSEthat incorporate noise when using simulated data, as demonstrated by Pollock, Macaulay, Önkal-Atay and Thomson (1999). As mentioned previously, the *MSE* is used in this study not only in the context of a measure of general accuracy (termed the *mean squared error for performance, MSEP),* but also as a measure to examine 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 in Thomson, Pollock, Gönül and Önkal (2013). It is further illustrated that the *MSEP* for composite forecasts can be statistically obtained from the *MSEP* for individual forecasts and the *MSEC* between pairs of individual forecasts. This is extended to component performance measures of *bias squared* (measuring under/overestimation)*,* *resolution variation* (measuring discrimination ability) and *error variation* (measuring variation in actual values not explained by variation in forecasts).

To facilitate the presentation of the proposed performance and coherence analysis, the following notation is used. Forecasts are denoted as *fij*, where *i* denotes the forecaster *(i=1,2,…,n)* and *j* denotes the specific forecast period *(j=1,2,…k)*. Composite forecasts *(fmj)* for period *j* for all individuals *(i=1,2,…,n)* are obtained by taking the simple average of these individual forecasts for period *j*. That is: *fmj =* **with the mean of the composite forecasts for all periodsgiven by *M(fm)= *.

*2.1 Performance Measures*

To evaluate performance, forecasts can be compared with realized values. Ex-post, the actual measured value, denoted *aj*, for the end of the forecast period *j* (*j=1,2,…,k)* will be known. The individual forecast, *fij*, and its composite form, *fmj*, can be used in the performance analysis by comparing the values with this actual value, *aj*,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 precisely in line with theactual value such that *fij* = *aj* for *i* for all *j*. In that regard, 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, *ci*, for *i*, for all *j* periods such that that *fij* = *ci* for *i* for all *j*. The *CVF* would be appropriate for non-trending series and shows no variation between forecast values; therefore it displays no resolution in performance. These hypothetical forecasters are used in discussing 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 individual forecasts (*MSEPIi*), as well as for composite forecasts (*MSEPM* over all *j* forecasts *(j=1,2,…,k))*, as defined in Equations *(1a)* and *(1b)* respectively:

 *(1a)*

 *(1b)*

A value of zero would imply that forecast values are identical to actual values (indicating perfect accuracy); hence, higher the value of the *MSEP,* poorer the forecast performance. In the case of the *perfect forecaster*, *fij=aj* for all *j*, such that *MSEPIi= 0*. For the *constant value forecaster*, the predictions would be *fij =ci* for all *j*, hence *MSEPIi = V(a)+(ciM(a))2,* where *M(a)* and *V(a)* denotes the mean and variance of the actual values respectively.

The *MSEP* is an overall performance measure which can be decomposed to identify specific components that reflect the multidimensional aspects of accuracy. The decomposition used in the present study involves *bias squared for performance* (*BSP*), *resolution variation for performance* (*RVP*) and *error variation for performance* (*EVP*). The *MSEP* decompositions are presented in Equations *(2a)* and *(2b)*:

*MSEPIi = BSPIi + RVPIi + EVPIi (2a)*

*MSEPM= BSPM + RVPM + EVPM(2b)*

These three components are discussed next.

* + 1. *Bias Squared for Performance*

*Bias (B)* is measured by the difference in the mean of forecast values*, M(f)* and the mean of the actual values, *M(a)*. *Bias* occurs when the mean forecast value is either too low, reflecting underestimation *(B<0)*, or too high, reflecting overestimation *(B>0)*, of the average actual values. The *bias squared* value for the individual forecaster is simply the square of *bias* and is a specific component of the *MSEP* decomposition. The *perfect* forecaster would illustrate zero bias and the *constant value forecaster* would have *bias* that would equal 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 *bias* for individual forecasts squared and divided by the square of the number of forecasters.

*Bias squared for performance* (*BSP*) for the individual and composite forecasters is defined in Equations (3a) and (3b) below:

*BSPIi = Bi2 (3a)*

*BSPM= Bm2 =  (3b)*

where

*Bi= M(fi) - M(a)* and *Bm= M(fm) - M(a)*

*M(a) = *and *M(fi) = *and *M(fm)* = **

* + 1. *Resolution Variation* for *Performance*

*T*he *resolution variation* component of performance (*RVP)* for individual and composite forecasters is related to *resolution* or *slope (SLi or* *SLm ),* which is a measure of discrimination that reflects the ability to detect and make an appropriate increment in the forecast when a one-point change in the actual value occurs. *Resolution* is measured by the slope coefficient from a linear relationship between the forecast and the actual values. *Resolution* would particularly be important in situations where the forecaster needs to correctly identify the direction of movement in the actual series under consideration and to discriminate between large and small movements. This is a critical aspect of performance that reveals the 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. *Resolution*, probably the most important aspect of performance, therefore, 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)*. As *resolution* approaches unity, *RVP* approaches zero. *Resolution* can be negative in cases where the forecaster‘s performance is worse than the *constant value forecaster*. The *RVP* term for both the individual and composite forecaster is 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 multiplied by the variance of the actual values.

Resolution variation for performance (RVP) for the individual and composite forecasters is defined in Equations (4a) and (4b) below:

*RVPIi* = *(1  SLi)2V(a)* *(4a)*

*RVPM = * *(4b)*

where

*SLi =C(fi,a)/V(a)* and *SLm* =*V(a) = () – M(a)2*

C*(fi,a)= ()  M(fi)M(a)*

* + 1. *Error Variation for Performance*

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

*Error variation* for performance(*EVP*) for the individual and composite forecasters is defined in Equations (*5a*) and (*5b*) below:

*EVPIi = SCi = V(ui) (5a) EVPM= (5b)*

where

*uij = fij  Ai  SLi aj* and *V(ui)= V(fi) SLi2 V(a)*

*Ai =* *M(fi)SLi M(a)*

*2.2* *Measures of Coherence*

Forecasts of *forecaster* *h (h=1,2,…,n-1),* denoted *fhj*, can be compared with those of *forecaster* *i (i=2,3,…,n, i>h)*,denoted *fij*, to examine coherence , which is a measure of consistency or agreement between the predictions. When predictions are perfectly coherent for a specific period, *j*,these values should all be equal, i.e., *fhj =* *fij*. Those situations where 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,* to provide a measure for each set of paired forecasts. There will be *n(n-1)/2* sets of paired values for a total for *n* forecasters. Coherence can be measured by a range of statistics that have a similar form 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 coherence obtained from forecasts, *fhj* and *fij*. The *MSEC* between each pair of forecasters, *h* and *i*, over all *j* periods *(j=1,2,…,k)*, is defined in Equation *(6)*:

 *(6)*

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

As with *MSEP*, the *MSEC* can be decomposed to identify specific aspects of coherence between two forecasters, *h* and *i*. This decomposition involves *bias squared for coherence* (*BSC*), *resolution variation for coherence* (*RVC*) and *error variation for coherence* (*EVC*), and is presented in Equation *(7)*:

*MSEChi = BSChi + RVChi + EVChi  (7)*

These three measures are discussed next.

*2.2.2* *Bias Squared for Coherence*

*Bias squared for coherence (BSChi*) is the squared difference in the mean forecasts between the two forecasters *(h* and *i).* A zero on this measure indicates coherence in the means, *(M(fh)-M(fi)). Bias* (without squaring) is used as a measure to indicate if the forecasters give different mean predictions. For example, in a situation where *M(fh*) is less than *M(fi)*, this would indicate that forecaster *h* has generally given lower predicted values than forecaster *i,* which would indicate negative coherence bias between *h* and *i*. On the other hand, in a situation where *M(fh)* is greater than *M(fi)*, this would indicate positive coherence bias between *h* and *i*. *Bias squared* for coherence is the squared difference between the two forecasters on bias for performance measure, which is a specific component of the *MSEC* decomposition.

*Bias squared for coherence* (*BSC*) is defined in Equation (8) below:

*BSChi  [M(fh)  M(fi)]2 (Bh  Bi)2  (8)*

where

*M(fh) = *and *Bh = M(fh)  M(a)*

*2.2.3* *Resolution Variation for Coherence*

*Resolution variation for coherence* (*RVChi*) inis the square of the difference of the two *resolution terms (SLh* and *SLi*)between two forecasters *(h and i)* multiplied by the variance of the actual values *V(a).* A zero on this measure indicates coherence in *resolution* (*slope*). Non-zero values on this measure reflect a degree of diversity between resolution performances of the two forecasters.

Resolution variation for coherence *(RVC)* is defined in Equation (9) as follows:

*RVChi = (SLh SLi)2V(a)*  *(9)*

where

*SLi =C(fi,a)/V(a)*

*2.2.3* *Error Variation for Coherence*

Error variation for coherence (*EVChi)* reflects the difference in variance of the two dependent error or scatter (*SCh + SCi -2SChi)* between two forecasters (h and i). In other words, the *EVChi* measures coherence between the error variation for performance of the two forecasters. Higher values of error variation for performance (scatter) leads to higher values of *EVChi*, depending on the offsetting effect of scatter covariance. *Error variation for coherence* (*EVC*) is defined in Equation (10) below.

*EVChihhi= SCh + SCi  2SChi (10)*

where

*SCi = V(ui)* and *SCh = V(uh)*

*SChi = C(uh,ui) = ()=C(fh,fi)  SLh SLi V(a)* (note: *M(uh) = M(ui)* *=0*)

*uhj = fhj  Ah  SLh ah* and *V(uh)= V(fh) SLh2 V(a)*

*uij = fij  Ai  SLi aj* and *V(ui)= V(fi) SLi2 V(a)*

*Ah =* *M(fh)SLh M(a)* and *Ai =* *M(fi)SLi M(a)*

*C(fh,fi) = ()  M(fh)M(fi)*

*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 directly obtained from the performance measures of individual forecasters and the coherence measures. This is illustrated in Equation *(11)*:

*MSEPM =  (11)*

Equation *(11)* shows that the *MSE for performance* for the composite forecaster (i.e., *MSEPM)* is the sum of individual *MSEs,* *MSEPIi* for all *i* forecasters, divided by *n* less the sum of the *MSE for coherence*, *MSEChi,* for all pairs of forecasters (*h* and *i*) divided by the square of *n*. Therefore, the composite *MSEPM* measure will be less than the sum of the individual *MSEPI* measures, provided a degree of diversity exists, as measured by the *MSEC* for any the pairs of forecasters. Only in the case when all individual forecasters are perfectly coherent with each other will the *MSEP* for composite and individual forecasters be the same.

*S*imilarly formed equations apply to the components of the *MSE for performance and coherence* involving *squared bias*, *resolution variation* and *error variation***.** These are presented in Equations *(12a)*, *(12b)* and *(12c)* below:

*BSPM =  (12a)*

*RVPM =  (12b)*

*EVPM =  (12c)*

Equations *(12a)*, *(12b)* and *(12c)* show that the components of the *MSE for performance* for the composite forecaster, *bias squared (BSPM)*, *resolution variation* (*RVPM*) and *error variation (ERVPM*) are the sum of the respective individual component valuesfor 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*. Therefore, 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 Equations *(11), (12a), (12b)* and *(12c)* by dividing the second term in these equations by the first term and then multiplying the result by 100 to give a value in percentage terms. This gives the following relative measures set out below in Equations *(13a)* to *(13d)*.

*RMSEPM = 100 \*  (13a)*

*RBSPM = 100 \*  (13b) RRVPM = 100 \*  (13c)*

*REVPM = 100 \*  (13d)*

Equations *(13a)* to *(13d)* show the relative percentage improvement of composite measures for the *MSEP* and its components where: *RMSEPM* is the composite forecasts *relative percentage MSE for performance* with *RBSPM, RRVPM* and *REVPM* giving similar *relative percentage performance* values for *bias squared, resolution variation* and *error variation*. These relative measures give the percentage improvement made by using composite forecasts relative to simply using the average of the individual forecasts. The higher the percentage value, the greater the improvement. These measures can, alternatively, be obtained from the values of the relevant *MSEPI* and the *MSEPM* and from the values of their components.

1. ***A Demonstration of the Framework***

To illustrate the above framework and the interpretation of its measures, an analysis is presented 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 to demonstrate the proposed analytical framework.

*3.1 The Data*

The forecast data used in this illustration involved Q4 *RPI* annual inflation rate forecasts, obtained from the *HM Treasury – Forecasts for the UK Economy*, from four sets of *City RPI* forecasts, available in the December editions for Q4 the following year. The editions used dated from December 1997 to December 2013 which provided forecasts for Q4 1998 - Q4 2014. 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 following analysis. For two cases (i.e., *F2* forecasts for 2004 and 2005),data were not available in the December edition, and the data from the November and October editions were used. Forecasts were compared with the realized Q4 values of the Retail Prices Index obtained from the *Office of National Statistics (code CZBH)*.The actual and forecast values for yearly Q4 *RPI* values are presented in Table 1.

*3.2 Characteristics of the Data and Forecast Errors*

The actual *RPI* inflation data for Q4 1998 - Q4 2014 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*. Forecasts from all four forecasters had slightly lower mean values, resulting in negative forecast errors reflecting 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 forecast error sample standard deviations were relatively similar. Two forecasters, *F1* and *F4* had forecast errors that were less than 2% for all the 17 years. There were, however, negative forecast errors, above 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 for the forecast errors between pairs of forecasters are presented in Table 2. These values were positive and large, with *F1* and *F2* showingthe highest correlation*.* 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 calculated values for the measures set out in Section 2 using the data presented in Table 1. This involves the four individual sets of forecasters (*F1* to *F4*) and all possible composites (i.e., using all forecasters [*C1234*], for 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, *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*. The results are also given for the *perfect forecaster* *(PF)* and three *constant value* *forecasters*, reflecting a constant inflation rate prediction of 2%, 3% and 4%, denoted *CVF(2)*, *CVF(3)* and *CVF(4),* respectively. The choice of *2%*, *3%* and *4%* is appropriate as only a small number of the forecasts for *F1* to *F4* over the 17year period fell outside this range (i.e., 2 for *F1,* 3 for *F2*, 5 for *F3* and 2 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 components, *squared bias* (*BSC*), *resolution variation* (*RVC*) and *error* *variation* (*EVC*). Table 5 presents the *relative percentage* performance improvement 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 for the *MSE* and its components below.

*3.3.1 Mean Squared Error*

The *MSE for performance* (*MSEP*)measures overall performance so that the best inflation forecasters over the period can be identified. Table 3 shows that the best values from the individual forecasters occurred for *F1*, followed by *F4* and *F3*. The best value on this measure is zero which is the value for the *perfect forecaster*. All four forecasters and, therefore, all the composite forecasters had better values than the *random walk forecaster* and the three *constant value forecasters*, except for *F3*,which had a value slightly poorer than the *3%* *constant value forecaster.* Table 4 shows thatthe 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%)* and *C234 (14%)* with *C34 (12%)* indicating the relatively high diversity between *F3* and *F4*. Combining forecasts improved overall performance, 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 only resulted in one case, *C23*, where the composite improved on both of the relevant individual values, *F2* and *F3.*

*3.3.2 Bias Squared*

*Bias squared* and *bias* are important where forecasts are required that do not show general over/under-forecasting of the rate of inflation over the longer term. This is particularly relevant to situations when long-term funding shortfalls could build up arising from persistent under/overestimation of yearly inflation rates. If *bias* is small, any underfunding that occurred in some years would be offset by overfunding in other years. *Bias* was negative for all four forecasters, reflecting a general underestimation of inflation over the whole period. The best value on this measure is zero, the value for the *perfect forecaster*. Table 3 shows the lowest *bias* (in absolute terms) occurred for *F1*, followed by *F2* and *F4* with the lowest *bias* for composite forecasts occurring forthose that involved *F1. Bias squared for performance* gave a similar ordering with all the individual and composite forecasts being worse than 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 *F1,F3* pair.Table 5 shows that the largest relative percentage improvement for composite forecasts occurred for the composite *C13* of *14%* indicating the relatively low coherence, between *F1* (who showed the lowest *bias squared*) and *F3* (who showed the highest *bias squared*). This also partly explains the relatedly large percentage improvement for composite forecasts *C123 (10%), C134 (10%)* and *C1234 (8%)*.Composite forecasts, therefore, showed some improvement on *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 forecasts are required that successfully identify and distinguish between years when high or low inflation is likely to occur. This is relevant to situations when there is a need to identify years when 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 is unity, the case of the *perfect forecaster*, with all the *constant value forecasters* having a value of zero. Table 3 shows that the *random walk forecaster* was considerably poorer than the four individual and all composite forecasters. The best values on *resolution* occurred for *F1,* followed closely by *F3,* with *F4* showing the poorest value. *Resolution*, however, cannot be improved by taking combined forecasts as composite resolution values are just the average of individual resolutionvalues. *Resolution* is a key variable in *resolution variation for performance* which has a best possible value of zero, the value for the *perfect forecaster,* and a value equal to the variance of the actual values (*V(a)=1.398*)for the three *constant value forecasters*. Table 4 shows that the paired values for *resolution variation for coherence* were relatively small and almost zero for pairs that did not include *F4* which had the poorestresolution*.* Table 5 shows that the *relative percentage* improvement on *resolution variation* for composite forecasts was almost zero for all composites that did not contain *F4*.Composite forecasts, therefore, only showed marginal improvements compared with individual forecasts overall with an average *relative percentage* improvement 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 where it is required to have forecasts with low unexplained variation in the rate of inflation. The poorer the performance on *error variation,* the greater the size of any financial provisions 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 performance*.* All 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 a value of zero and the *random walk forecaster* had much poorer performance than all the other forecasters. Table 4 shows thatthe paired values for *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 composite *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 on overall performance with the average improvement being *17%*. This improvement level is in fact better than those reported in literature (Batchelor and Dua, 1995; Armstrong, 2001). Batchelor and Dua (1995) asserted that forecast combinations reduced error by an average of 9.2% (for combination across two forecasters) to 16.4% (for combination across 10 forecasters) while Armstrong (2001), in a review of 30 studies, found that the reduction of ex-ante errors averaged around 12.5%. The improvement in composite forecasts relative to the mean of individual forecasts only resulted in 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 of combined forecasts, as measured by *MSE for performance*,compared with the mean of individual measures was mainly due to lower *error variation* and *bias squared*,with *resolution variation* having a smaller effect. On the performance measures for *MSE* and all its components, forecaster *F1* performed best. 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 performance than all the individual forecasts when the individual forecast measures are reasonably close together and the respective coherence measures show reasonably high diversity. If, however, an individual forecast is distinctly better than other individual forecasts on a performance measure, particularly when there is a reasonable degree of coherence between them, then composite forecasts may not improve on the best individual forecast.

***4. Discussion***

This paper sets out an analytic framework that can be used to enhance the performance assessment of forecasts and guide decisions on which forecasters to pool for an effective combined forecast. It is shown that composite forecasts (formed using a simple average) have a lower *mean squared error* than that of the average of the individual forecasts. This result supports previous research (e.g., Armstrong, 2001; Clemen, 1989; De Menezes, Bunn & Taylor, 2000; Lawrence, Edmundson & O’Connor, 1986; Makridakis & Hibon, 2000; Schnaars, 1986; Soll & Larrick, 2009; Stock & Watson, 2004; Timmerman, 2006) and confirms 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 consistency and dependence between 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 between them, so as to achieve the best possible accuracy levels from a group of component forecasts. Coherence measures, as they do not involve consideration of the actual values of the series under consideration, will tend to be more stable over time than the performance measures, with any changes only likely to occur when forecasters switch their forecasting models or procedures.

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

The study consolidates the results from previous literature (e.g., Armstrong, 2001; Armstrong, et al, 2015; Goodwin, 2015; Graefe, et al, 2014; Green, et al, 2015) and implies that the extent of improvements may be directly related to the degree of diversity between pairs of individual predictions. The higher the diversity indicated between the individuals on each component, the better will be the improvement on that component for combined forecasts.

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 performance, but also on the basis of relative coherence. The findings imply that composite forecasts should be obtained by pooling heterogeneous forecasters that show effective performance in specialized aspects targeting customized use of the pooled forecasts. This is important as it is not always possible to control coherence by selecting diverse forecasts. It is useful, therefore, to explicitly incorporate measures of coherence into the selection and formation of composite forecasts. The findings also imply that if a weighting procedure is used, then levels of coherence between individual forecasters, as well as performance, need to be taken into account in setting the values for the relative weights.

The framework can be used to aid the identification of good /superior forecasters that can be included in the formation of composite forecasts while filtering 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 significant structural changes are not occurring. The coherence measures can be used to effectively detect changes in forecast method use among forecasters. 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 only likely to occur when forecasters switch their forecasting models or procedures. Awareness of multidimensional aspects of accuracy can also identify structural changes. For instance, higher error variation measures can be associated with a failure of forecasters to identify increased volatility. Shorter rolling sample periods over the whole period can be used to aid 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 in relation to 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* and exclude forecasters that show poor *resolution*, 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,* 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 variation* and omit those with high *error variation*. To examine these aspects, however, it is desirable to have sufficiently large data to consider variations in the methods used by forecasters 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 & 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 have utilized a decomposition of the ensemble generalization error into a variety of components (bias, variance, covariance and noise variance). Their analysis of this decomposition has shown that when two estimators were ‘negatively correlated’, their combination leads to a lower generalization error, whereas the generalization error would be greater in case of a positive correlation between different estimators. Thus, there exists a relationship between the diversity and ensemble/combination performance in a machine learning system (e.g., Du Jardin, 2016; Florez-Lopez & Ramon-Jeronimo, 2015; Yin et al., 2014; Zhang & Zhou, 2013) and that, the higher the diversity among 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 potential limitations. Our study only addresses measures 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 there are directions in our paper to aid the selection of appropriate individual forecasts in composite forecast construction, there is further potential for extending this process. 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 has 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. Additionally, the framework could be extended to target prediction intervals and probability forecasts, which would in turn 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 probabilistic forecasting domain (e.g. Broomell & Budescu, 2009; Budescu & Rantilla, 2000; Budescu, Rantilla, Yu & Karelitz, 2003; Budescu & Yu, 2007) that attempts to unveil the link between the aggregation of expert predictions and the 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 determine the appropriate error measures 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 to use in composite predictions. To illustrate the application of the framework and the interpretation of its statistical measures, yearly UK RPI inflation forecasts (for 1998-2017) from four banking institutions were employed. This example was chosen to allow the study to be easily replicated with readily available published data, while illustrating the framework in a context that is highly relevant in business, economic and government policy landscapes. The framework’s application to inflation forecasting showed that composite forecast improvements primarily resulted from lower error variation (that had relatively high paired coherence values) and to a lesser extent, from lower bias squared with small improvements in resolution adjusted variation. Although the demonstration of the framework employed inflation forecasts for illustration purposes, the measures are applicable to a wide range of other contexts in the financial economic and business arenas.

To summarize, the framework provides a powerful diagnostic toolbox that can be used in a multitude of practical forecasting situations to improve forecasts when individual predictions are to be combined into a single forecast to be communicated to decision/policy-makers. It makes an important contribution to understanding the role of coherence and highlights the role of group heterogeneity whereby combined performance demonstrably improves when individual forecasters bring in asymmetrical 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*. Norwell, M.A, Kluwer Academic Publishers, 417-439.

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., Jones Jr., 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.

HM Treasury. *Forecasts for the UK Economy.* Available monthly from September 1997. [*http://www.hm-treasury.gov.uk/data\_forecasts\_index.htm*](http://www.hm-treasury.gov.uk/data_forecasts_index.htm)Accessed 24.05.2017.

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*, 2417-24.

Office for National Statistics. *Retail Prices Index (code CZBH)*. <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.

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*, North Holland, Amsterdam.

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, Volume 1.* North Holland, Amsterdam.

Ueda, N., & Nakano, R. (1996). Generalization error of ensemble estimators. In *Proc. of Intern. Conf. 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*,.

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.

**Tables**

**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** |

**Table 2**

**Correlations between the 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** |

**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 Two Sets 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** |

**\***The lower the value the greater the coherence

**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** |

1. Corresponding author [↑](#footnote-ref-1)