![IMF Working Paper logo]

The Use of Encompassing Tests for Forecast Combinations

Turgut Kışınbay

INTERNATIONAL MONETARY FUND

IMF Working Paper

Monetary and Capital Markets Department

The Use of Encompassing Tests for Forecast Combinations

Prepared by Turgut Kışınbay[1]

Authorized for distribution by Karl Habermeier

November 2007

Abstract

The paper proposes an algorithm that uses forecast encompassing tests for combining forecasts. The algorithm excludes a forecast from the combination if it is encompassed by another forecast. To assess the usefulness of this approach, an extensive empirical analysis is undertaken using a U.S. macroecoomic data set. The results are encouraging as the algorithm forecasts outperform benchmark model forecasts, in a mean square error (MSE) sense, in a majority of cases.

Contents   Page

I. INTRODUCTION

The concepts of forecast combination and forecast encompassing are closely related. Often times, there are several competing forecasts of a target variable that come from different sources or models. One way to utilize all the competing forecasts is to merge them into a single, combined forecast. Forecast encompassing tests are used to determine whether one of a pair of forecasts contains all the useful information for prediction. If this is not the case and rather both models contain some incremental information, there is potential to form a combined forecast that blends the useful information of the two (or more) forecasts.

This paper proposes an algorithm that uses encompassing tests to combine forecasts. The algorithm is based on a simple idea: if a forecast is encompassed by another, it is excluded from the combination. Once all the encompassed forecasts are eliminated, the combined forecast is obtained by taking an arithmetic average. To assess the usefulness of this approach, an extensive empirical analysis is undertaken using a U.S. macroeconomic data set. There are 110 *target variables* to be forecast, and for each of the target variables, about 70 to 150 alternative forecasts to be used in combinations. For each of the 110 cases, the combined forecast obtained with the algorithm is compared with a benchmark model. The results are encouraging as the algorithm forecasts outperform benchmark model forecasts, in a MSE sense, in a majority of cases.

The approach proposed here differs from most of the literature on forecast combinations in that it suggests reducing the number of available forecasts before combining them. Much of the forecast combination literature concentrates on how to optimally combine a given set of forecasts.[2] A straightforward approach is to use linear regressions to obtain weights, as proposed by Granger and Ramanathan (1984). This could be done by regressing the target variable on the set of all available forecasts, and then using the regression coefficients as weights. Gupta and Wilton (1987) propose nonparametric methods for calculating weights. The possibility of structural change or state-dependent relationships led researchers to develop time-varying combination methods; see Diebold and Pauly (1987) and Zellner, Hong, and Min (1991). Finally, Diebold and Pauly (1990) propose Bayesian shrinkage methods. Most of these methods take the number of alternative forecasts as given, and then seek to find the best weights for combining them. Although weights of some forecasts can be (close to) zero, generally all forecasts enter the combination.

Methods that choose among alternative forecasts before combining them could enhance predictive power. Examples along this line of research include Schmittlein, Kim, and Morrison (1990) and Swanson and Zang (2001). Schmittlein, Kim, and Morrison propose using Akaike Information Criteria (AIC) for combining forecasts, and provide Monte Carlo evidence on the usefulness of their approach. Swanson and Zhang consider other model

---

[2] See surveys by Newbold and Harvey (2002) and Timmerman (2006).

selection approaches in addition to AIC, and also provide a real-time empirical analysis. They find that an approach based on the Schwartz Information Criterion (SIC) may provide a useful alternative to other forecast combination methods. Note that both papers examine model selection approaches to reduce the dimension of the forecast combination, rather than using statistical testing. Interestingly, although forecast combinations and forecast encompassing principle are closely related, the latter is rarely used for combining forecasts, but rather for evaluating them.

II. THE ENCOMPASSING PRINCIPLE

Let f_1 and f_2 be two alternative sets of forecasts of a variable. Assume that one of the two sets, say f_1, performs better by some criteria, say Root Mean Square Error (RMSE). The idea behind forecast combinations is that the poorly-performing forecast may provide some marginal information that is not contained in the better forecast. In such a case, the combined forecast will perform better then either forecast alone. However, if the poorly-performing f_2 contains no useful marginal information, than it is said that f_1 encompasses f_2.

To formally test for forecast encompassing, I will use the Harvey, Leybourne, and Newbold (HLN, 1998) test. The HLN (1998) test is based on the well-known forecast evaluation test developed by Diebold and Mariano (DM, 1995). The DM test is used to test for equal predictive ability of two competing forecasts. It considers a sample of loss differential series d_t, defined as $d_t = L(e_{1t}) - L(e_{2t})$; where $L(\ )$ is some arbitrary loss function, such as MSE; e_{it} is the τ-step-ahead forecast error of the model i; $i = 1, 2$; and $t = 1, \cdots, T$. Equal predictive accuracy amounts to $E(d_t) = 0$, and the test is based on the observed sample mean $\bar{d} = (1/T) \sum_{t=1}^{T} d_t$. Assuming covariance stationarity of the loss differential series, the DM test has an asymptotic standard normal distribution under the null hypothesis of equal predictive accuracy. The test statistic is as follows:

$$DM = \frac{\bar{d}}{\sqrt{\hat{V}(\bar{d})}} \tag{1}$$

Where $\hat{V}(\bar{d})$ is a consistent estimate of the asymptotic variance of $\bar{d}$, and assuming that τ-step-ahead forecasts exhibit dependence up to order $\tau - 1$, it is obtained as:

$$\hat{V}(\bar{d}) \approx \frac{1}{T}\left(\gamma_0 + 2\sum_{i=1}^{\tau-1}\gamma_i\right) \tag{2}$$

where γ_i is the ith autocovariance of $\bar{d}$, estimated by $\hat{\gamma}_i = T^{-1} \sum_{t=i+1}^{T} (d_t - \bar{d})(d_{t-i} - \bar{d})$.

HLN (1997) assess the performance of the DM test using Monte Carlo simulations. The authors recommend two simple modifications to the DM test that would improve the power of the test in small samples, resulting in the modified version of DM (MDM): (i) Compare the test statistic with critical values from the Student's t-distribution with $T-1$ degrees of freedom, instead of the standard normal, so as to reduce size distortions; and (ii) Modify the test statistic as $MDM = T^{-1/2} \left[T + 1 - 2\tau + T^{-1}\tau(\tau-1) \right]^{1/2} DM$.

The HLN forecast encompassing test is closely related to the DM test. It is simply obtained by modifying d_t to $d_t = (e_{i,t} - e_{j,t})e_{i,t}$. The null hypothesis is that model i forecast encompasses model j. That is, all the relevant information of model j is contained in model i.

III. DATA AND OUT-OF-SAMPLE FORECASTS

The data set used is from Marcellino, Stock and Watson (2006), and consists of 172 macroeconomic series from 1959:1 through 2002:4. (Most data are monthly, but converted to quarterly; the Data Appendix provides a detailed description.) Forecast combinations are produced for 110 of these series called *target variables*; series with a short time span are not forecast. For each variable to be forecast, other series in the data set are used as predictors. Simple models in the form of bivariate regressions are used to produce forecasts with each of these alternative predictors. Forecasts of the alternative models are then used in combinations.

The bivariate linear regressions used to obtain pseudo out-of-sample forecasts are of the form:

$$Y_t = \alpha + \sum_{i=0}^{q-1} \gamma_i Y_{t-i} + \sum_{i=0}^{p-1} \beta_i X_{t-i} + e_t \qquad (3)$$

where Y_t is the *target variable* to be forecast, X_i is a predictor, and e_t is an error term. The lags of the variables X and Y are chosen using the SIC, where $1 \le p \le 4$ and $0 \le q \le 4$. SIC is calculated as $SIC = n \ln \left(1/n \sum_{t=1}^{n} \hat{\varepsilon}_t^2 \right) + k \ln n$, where n is the number of observations, and k is the number of regressors including the constant. Regressions of this type are run for all available predictors X in the data set. In order to simulate real-time forecasting, models are re-estimated at each period.

The first estimation sample used to construct a set of forecasts is from the first quarter of 1959 to the last quarter of 1969, providing out-of-sample forecasts for the first quarter of

1970. I then expand the sample by one period, re-estimate the models, and produce forecasts for the second quarter of 1970, and so forth, ending with a final forecast for the fourth quarter of 2002. The maximum number of out-of-sample forecasts across time with this data set is 131.

To forecast price variables, all other series are used as predictors; however, real variables are not forecast using price variables. As a consequence, the number of forecasts produced that is used in the combination at a point in time depends on whether the target is a real or price variable. When the target is a price variable, the number of forecasts available at each period is close to 171, as all available predictors are used for forecasting price series. The exact number will be less than 171 as some series do not go back to 1959, and some were discontinued. Since price series are excluded from forecasting real series, the number of forecasts in that case is about 70.

To obtain the benchmark forecast, the simple arithmetic average of all available forecasts is taken. As demonstrated in many studies, simple averaging works very well in many applications and is extremely difficult to beat consistently even with sophisticated models; hence, it is a natural benchmark.[3] This benchmark forecast combination is called AVE forecast. The other forecast combination, also a simple average, but obtained after the number of forecasts in the combination is reduced by employing the encompassing algorithm, is called the EAL forecast (after the encompassing algorithm).

IV. THE ENCOMPASSING ALGORITHM

For each target variable, there are several alternative models, each corresponding to a different series in the data set. Each model produces a set of pseudo out-of-sample forecasts, starting from the first quarter of 1970 to the end of the sample. The algorithm uses pseudo out-of-sample forecasts as inputs. The idea is to compare all models available forecasts with each other using encompassing tests, eliminate those that are encompassed by others, and take the average of the remaining forecasts. The comparisons are done bilaterally, using the HLN (1998) forecast encompassing test. Once the pseudo out-of-sample forecasts are obtained, the following steps are taken to eliminate forecasts encompassed by others, and obtain the EAL combination.

Step 1. Start at time t and calculate the RMSE of the out-of-sample forecasts for each model using out-of-sample forecasts up to time $t-1$, and realized values. Rank the models according to their past performance based on RMSE.

Step 2. Pick the best model (i.e., model with the lowest RMSE), and test sequentially whether the best model forecast encompasses other models, using the HLN test. If the best

---

[3] On the forecasting performance of simple averaging, see the surveys by Clemens (1989), Newbold and Harvey (2002), and Timmerman (2006).

model encompasses the alternative model at some significance level α, delete the alternative model from the list of models.

Step 3. Repeat Step 2, with the second best model. Note that the list of models now contains only those that are not encompassed by the best model, and the best model.

Steps 4 and 4+ Continue with the third best model, and so on, until no encompassed model remains in the list.

Last step: Calculate the EAL forecast by taking the average of all remaining models' forecasts.

There are several issues to consider in applications. First, an initial set of out-of-sample forecasts is required in order to apply the HLN (1998) test. One option is to use all available forecasts prior to the date the forecast is being produced. An alternative would be to choose a rolling window of a fixed number of observations. I will present results for all available forecasts and the last 20 forecasts, denoted by $m = all, 20$, respectively. To start the applications, out-of-sample forecasts before 1980 are used as inputs, and then the sample is expanded by one-quarter to simulate real-time forecasting. Thus, the comparisons of forecast combinations are based on post-1980 data. Note that some series do not go back to 1959, and there is need to accumulate some data points before including the series in combinations. I impose the condition that there should be at least 30 observations for a variable to be included. So effectively, the window size is 30 in most cases. I then experiment with a window size of 40, which does not affect the conclusions of the paper.

The second concern is the choice of the significance level α of the HLN test. I will experiment with a range of significance levels ($\alpha = 0.01$, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45) to explore whether a subrange would appear as maximizing the forecasting power. Finally, there is the problem of outliers. To tackle this issue, forecasts that are more than five standard deviations away from the target variable are considered as extreme and removed from combinations.

The proposed algorithm has some desirable properties. A common problem facing modelers attempting to utilize large data sets is that in many cases the number of variables in the data set exceeds the number of observations. In those cases, it is difficult or impossible to use conventional model selection techniques, such as linear regression techniques. The method proposed in this study can easily deal with this problem. A large number of alternative forecasts should not pose any major problems, except for computing time. Another advantage of the algorithm is that it does not require knowledge of the models that produce alternative forecasts; all that is needed is the set of forecasts.

V. RESULTS

A. Does the Algorithm Work?

I compare the models by looking at the relative forecasting performance based on RMSE. The benchmark for each target variable is the AVE model. RMSE are calculated both for the benchmark model and the algorithm forecasts. The ratio $RMSE_{EAL}/RMSE_{AVE}$ gives a scale-free metric, where a ratio less than one indicates that the algorithm forecast outperforms simple averaging.

Table 2 reports the averages of the relative RMSEs for the 110 target variables for each significance level and two different training sample sizes. The results suggest that when $m = all$ and the significance level is greater then 0.15, the EAL forecasts provide some modest gains compared to the AVE forecasts.

Visual illustrations of the distribution of relative RMSE are provided in Figures 1a-d. The figures present results for $m = all$ and $\alpha = 0.15, 0.25, 0.35$, and 0.45, respectively. In the figures, the vertical axis shows the relative RMSEs, while the horizontal axis contains the 110 target variables to be forecast, sorted from the lowest relative RMSE to the highest. Thus the first variable in the horizontal axis is the target variable for which the EAL forecast provides the highest RMSE gains compared to the AVE method.

At $\alpha = 0.25$ and $\alpha = 0.35$, percentage RMSE gains arising from employing the EAL are over 10 percent for about half of the cases, with the highest gains of over 20 percent. When $\alpha = 0.45$, there are less gains, which is not surprising; at high significance levels, less variables are eliminated from the combination and the resulting combined forecasts are closer to the simple average forecast. In contrast, at the low significance levels, very few variables remain in the combination and the forecast benefits less from the advantages of combining. The next section analyzes this tradeoff and presents results on the choice of the significance level.

B. Which Significance Level to Use?

Further results for comparing the forecasting performance of the algorithm at different significance levels are provided in Table 2. I start by looking at simple rankings. First, for each target variable I rank the number of times for which the best results are obtained across the 10 significance levels examined. For example, if the forecast with the smallest RMSE is obtained using significance level 0.25, and the second smallest is 0.30, then using 0.25 and 0.30 are ranked as 1 and 2, respectively. This is done for all 110 models, then the ranks are summed up with the smallest sum suggesting the best performing significance level.

The results vary across the two training sample sizes. When a larger sample is used, the best results are obtained when the significance level is in the $0.25 - 0.40$ range. Although the

differences are marginal, 0.35 gives the best average result. When the training window size is 20, the best results are obtained at higher significance levels.

The average number of remaining variables is also presented in Table 2. As expected, the number of nonencompassed variables increases with the significance level of the HLN test. At the low significance level of 0.001, usually only one variable remains, which is the one with the best historical RMSE performance. With $\alpha = 0.10$ there are about seven nonencompassed variables on average and the mean relative RMSE is equal to one, suggesting that as few as seven variables can perform as well as the larger data set. The best performance is achieved when $\alpha = 0.35$ and the number of variables is 25. It is also interesting to note that with a fixed window of m=20, the mean relative RMSE less than or equal to one is achieved when the significance level is greater than 0.25, corresponding to six or more variables in the forecast combination.

VI. COMPARISONS WITH OTHER METHODS

This section compares the proposed method with other well-known forecast combination methods. The following models are used to generate alternative combined forecasts:

RMSE-weighted combinations:

Define the ith forecasting model's RMSE at time t as $RMSE_{t,i} = \left[(1/n) \sum_{1}^{n} e_{t-j,i}^2 \right]^{1/2}$ which is computed over a window of n observations. Then the RMSE weighted combined forecast is:

$$\hat{y}_t^c = \sum_{i=1}^{M} \hat{\omega}_{t,i} \hat{y}_{t,i}, \qquad \hat{\omega}_{t,i} = \frac{(1/RMSE_{t,i})}{\sum_{i=1}^{M} (1/RMSE_{t,i})} \qquad (4)$$

Rank-weighted combinations:

Define $R_{t,i}$ as the rank of the ith model based on its historical RMSE performance up to time t. As suggested in Aiolfi and Timmerman (2006), the weights of the combination can be calculated as:

$$\hat{\omega} = R_{t,i}^{-1} \Big/ \left(\sum_{i=1}^{M} R_{t,i}^{-1} \right) \qquad (5)$$

Combinations based on ranks have the advantage that ranks are not very sensitive to outliers so the estimated weights can be robust.

The thick-modeling approach:

Granger and Jeon (2004) advocate the thick modeling approach, where the top x percent of the best performers are kept in the forecast combination. There are no theory-based

guidelines on how to choose *x*. The authors state that thick modeling is more of a "pragmatic folk-view than anything based on clear theory." Researchers usually exclude an arbitrary portion of the worst performing forecasts. Despite that, the method works very well in empirical applications. In this study I have experimented with the following numbers for *x*: 1, 5, 10, 20, 30, ... , 90. The results not only allow for comparisons with the combination strategy proposed in this study, but also provide some empirical evidence on the choice of *x*.

Combinations before and after the algorithm:

The above mentioned combination techniques are used both before and after the algorithm is applied. Doing that can improve the performance as some of the problems associated with large data sets may be removed after EAL filtering. However, since comparisons with other models would be difficult to present for all significance levels, I will focus on the models that provide the best results in the previous section, that is $\alpha = 0.35$. This will also ensure that enough observations are included in the thick modeling approach, which cannot be meaningfully applied with only a few observations.

Factor models:

In addition to various forecast combination techniques, factor models are also included in forecast comparisons. These models, advocated by Stock and Watson in a series of papers in recent years,[4] are shown to be fairly successful models for forecasting with large data sets. An illustration of factor models is as follows. Assume at time *t* we have a large data set, denoted X_t, and there is some common information in this data set which can be represented by a few unobserved factors F_t. The dynamic factor representation is as follows:

$$X_t = \Lambda_t F_t + e_t \tag{6}$$

$$y_{t+1} = \beta_t F_t + \varepsilon_{t+1} \tag{7}$$

where Λ_t is the factor loading matrix, y_t is the stationary time series to be forecast, and β_t is a vector of parameters that defines the relationship between the factors and the variable to be forecast.

Stock and Watson (2002) show that the factors can be conveniently constructed as the principal components X_t. Once the factors are estimated, the forecasting exercise is similar to standard linear prediction. Coefficients β_t are estimated by regressing y_t onto F_t, and then the forecast is formed as $y_{t+1} = \hat{\beta}_t F_t$. One practical problem is the determination of the

---

[4] See Stock and Watson (2006) for a survey.

number of factors to use in the forecasting exercise. Stock and Watson (2002) propose using information criteria to choose the optimum number of factors. I will experiment with three different approaches: only using the first principal component, and choosing the number of factors using AIC and Bayesian Information Criterion, where the maximum number of factors considered is four.

Results:

First, I calculate combined forecasts for each of the 110 target variables using 35 alternative models. Altogether about 90 forecasts are produced with each model for each target variable, and the RMSE (mean absolute errors, MAD) is calculated relative to the benchmark model. As before, a ratio of less than one indicates that the alternative model provides forecast gains. Finally, I calculate the average relative RMSE (MAD) for each forecasting method. That is, I provide 110 different experiments with various models. In Table 3, the models are ranked from best to worst and the average relative RMSE and MAD are reported. As can be seen from the tables, in many cases the differences in RMSE and MAD between models are small, thus the comparisons are only suggestive. In some cases, however, the results are stronger, such as the poor performance of thick models with only a small percentage of the best performers, or the factor models.

For all models except for the thick modeling approach, there is evidence suggesting that post-algorithm forecasts are better then pre-algorithm forecasts. More specifically, this is true for the RMSE and rank weighted, mean, and median combinations. The evidence on the thick modeling approach is mixed. The best overall results are achieved by pre-algorithm thick models, where 20–40 percent of the best models are retained in the combinations. Best results for post-algorithm results are obtained with 80 to 100 percent of forecasts.

Combinations with only a few forecasts, such as the top 1–10 percent, do not perform well on average. Moreover, such strategies deliver a high variance in performance. For example, the highest overall RMSE gains are obtained when the strategy is to choose the historically best model, giving us a reduction in RMSE close to 20 percent. At the same time, the worst performance is also obtained using the same model. These results once again warn against relying too much on best performances in the past, which are not necessarily repeated in the future.

Forecast combinations generally outperform factor forecasts. Compared to simple averaging, the average RMSE loss from using factor models is about 4–6 percent, depending on the model selection mechanism used for factor models. A comparison of the performance of model choice strategies for factor models suggests that using a single factor with lags almost always outperforms models with more than one factor, but no lags.

VII. CONCLUSIONS

The algorithm proposed in this study is based on the link between forecast combinations and the encompassing principle. The aim is to reduce the dimension of the combination problem by testing before combining. The empirical analysis presented suggests that the algorithm provides forecast gains in a MSE sense, compared to the benchmark simple averaging method. The average gains arising from employing the algorithm amount to about a 1–2 percent reduction in RMSEs. Although the gains are not substantial, they should still be considered noteworthy, given the well-known difficulty of outperforming the simple averaging method in practice.

Perhaps more important than the marginal reductions in forecast errors is the ability of the algorithm to perform at least as well as the simple averaging method. In this way, the algorithm suggests a method to represent a larger data set with a smaller one. This could prove useful in some applications as it may be easier to monitor and understand the dynamics of a smaller number of variables. For example, the method could be used to identify leading indicators of a variable from a large data set.

Appendix I. Data

The following Appendix Table lists the time series used in the empirical analysis. The series were either taken directly from the DRI-McGraw Hill Basic Economics database, in which case the original mnemonics are used, or they were produced by the author's calculations based on data from that database, in which case the author's calculations and original DRI/McGraw series mnemonics are summarized in the data description field. Following the series name is a transformation code, the sample period for the data series, and a short data description. The transformations are (Lev) level of the series; (D) first difference; (Ln) logarithm of the series; (DLn) first difference of the logarithm. The following abbreviations appear in data descriptions: SA: seasonally adjusted; NSA: not seasonally adjusted; SAAR: seasonally adjusted at an annual rate; AC: authors' calculations.

Table 1. Description of Data

Series	Trans.	Sample Period	Description
(A) Income, output, sales, capacity utilization
Msmq	DLn	1967: 1–2001: 7	Sales, business—manufacturing (chained)
ips11	DLn	1959: 1–2002: 12	Industrial production index—products, total
ips299	DLn	1959: 1–2002: 12	Industrial production index—final products
ips12	DLn	1959: 1–2002: 12	Industrial production index—consumer goods
ips13	DLn	1959: 1–2002: 12	Industrial production index—durable consumer goods
ips18	DLn	1959: 1–2002: 12	Industrial production index—nondurable consumer goods
ips25	DLn	1959: 1–2002: 12	Industrial production index—business equipment
ipi	DLn	1959: 1–2002: 10	Industrial production: intermediate products (1992=100, sa)
ips32	DLn	1959: 1–2002: 12	Industrial production index—materials
ips34	DLn	1959: 1–2002: 12	Industrial production index—durable goods materials
ips38	DLn	1959: 1–2002: 12	Industrial production index—nondurable goods materials
ips43	DLn	1959: 1–2002: 12	Industrial production index—manufacturing (sic)
ipd	DLn	1959: 1–2002: 10	Industrial production: durable manufacturing (1992=100 sa)
ipn	DLn	1959: 1–2002: 10	Industrial production: nondurable manufacturing (1992 = 100, sa)
ipmin	DLn	1959: 1–2002: 10	Industrial production: mining (1992 = 100, sa)
iput	DLn	1959: 1–2002: 10	Industrial production: utilities (1992 = 100, sa)
utl10	Lev	1967: 1–2002: 12	Capacity utilization—total index
utl11	Lev	1959: 1–2002: 12	Capacity utilization—manufacturing (sic)
utl13	Lev	1967: 1–2002: 12	Capacity utilization—durable manufacturing (naics)
utl25	Lev	1967: 1–2002: 12	Capacity utilization—nondurable manufacturing (naics)
utl35	Lev	1967: 1–2002: 12	Capacity utilization—mining naics = 21
utl36	Lev	1967: 1–2002: 12	Capacity utilization—electric and gas utilities
gmpyq	DLn	1959: 1–2002: 12	Personal income (chained, series #52, bil 92$, saar)
gmyxpq	DLn	1959: 1–2002: 12	Personal income less transfer payments (chained, series #51, bil 92$, saar)
Gmcq	DLn	1967: 1–2002: 12	Personal consumption expenditure (chained)—total (bil 92$, saar)
gmcdq	DLn	1967: 1–2002: 12	Personal consumption expenditure (chained)—total durables (bil 1996$, saar)
gmcnq	DLn	1967: 1–2002: 12	Personal consumption expenditure (chained)—nondurables (bil 96$,saar)
gmcsq	DLn	1967: 1–2002: 12	Personal consumption expenditure(chained)—services (bil 92$, saar)
gmcanq	DLn	1967: 1–2002: 12	Personal consumption expenditure (chained)—new cars (bil 1996$, saar)
Wtq	DLn	1959: 1–2001: 7	Merch wholesalers: total (mil of chained 1996 dollars, sa)

Table 1. Description of Data

Series	Trans.	Sample Period	Description
Wtdq	DLn	1959: 1–2001: 7	Merch wholesalers: durable goods total (mil of chained 1996 dollars, sa)
Msdq	DLn	1959: 1–2001: 7	Mfg. and trade: mfg.; durable goods (mil of chained 1996 dollars, sa)
msmtq	DLn	1959: 1–2001: 7	Mfg. and trade: total (mil of chained 1996 dollars, sa)
msnq	DLn	1959: 1–2001: 7	Mfg. and trade: mfg.;nondurable goods (mil of chained 1996 dollars, sa)
wtnq	DLn	1959: 1–2001: 7	Merch. wholesalers: nondurable goods (mil of chained 1996 dollars, sa)
rtdrq	DLn	1967: 1–2001: 4	Retail sales durables, real (rtdr/pucd, AC)
rtnrq	DLn	1967: 1–2001: 4	Retail sales nondurables, real (rtnr/pu882, AC)
ips10	DLn	1959: 1–2002: 12	Industrial production index—total index
(B) Employment and unemployment			
lpnag	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: total (thous., sa)
lhu26	Lev	1959: 1–2002: 12	Unemployed by duration: persons unemployed 15–26 wks (thous., sa)
lpgd	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: goodsproducing (thous., sa)
lhu15	Lev	1959: 1–2002: 12	Unemployed by duration: persons unemployed 15 wks+(thous., sa)
lp	DLn	1959: 1–2002: 12	Employees on nonag payrolls: total, private (thous., sa)
lpcc	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: contract construction (thous., sa)
lhelx	Ln	1959: 1–2002: 12	Employment: ratio; help-wanted ads: no. unemployed clf
lhu5	Lev	1959: 1–2002: 12	Unemployed by duration: persons unemployed less than 5 wks (thous., sa)
lhu14	Lev	1959: 1–2002: 12	Unemployed by duration: persons unemployed 5–14 wks (thous., sa)
lpsp	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: serviceproducing (thous., sa)
lptu	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: trans. and public utilities (thous., sa)
lpt	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: wholesale and retail trade (thous., sa)
lpfr	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: finance, insur. and real estate (thous., sa)
lps	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: services (thous., sa)
lpgov	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: government (thous., sa)
lw	Dif	1964: 1–2002: 12	Avg. weekly hrs. of prod. wkrs.: total private (sa)
lphrm	Lev	1959: 1–2002: 12	Avg. weekly hrs. of production wkrs.: manufacturing (sa)
lpmosa	Lev	1959: 1–2002: 12	Avg. weekly hrs. of prod. wkrs.: mfg., overtime hrs. (sa)
lhu680	Lev	1959: 1–2002: 12	Unemployed by duration: average (mean) duration in weeks (sa)
lhur	Lev	1959: 1–2002: 12	Unemployment rate: all workers, 16 years and over (%, sa)
lpen	DLn	1959: 1–2002: 12	Employees on nonag. payrolls:nondurable goods (thous., sa)
lpem	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: manufacturing (thous., sa)
lhel	DLn	1959: 1–2002: 12	Index of help-wanted advertising in newspapers (1967 = 100; sa)
lped	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: durable goods (thous., sa)
lhem	DLn	1959: 1–2002: 12	Civilian labor force: employed, total (thous., sa)
lhnag	DLn	1959: 1–2002: 12	Civilian labor force: employed, nonagric. industries (thous., sa)
lpmi	DLn	1959: 1–2002: 12	Employees on nonag. payrolls: mining (thous., sa)
(C) Construction, inventories and orders			
hssou	Ln	1959: 1–2002: 12	Housing starts: south (thous.u., s.a.)
contc	DLn	1964: 1–2002: 12	Construct.put in place: total priv and public 1987 $ (mil$, saar)
conpc	DLn	1964: 1–2002: 12	Construct. put in place: total private 1987 $ (mil$, saar)
conqc	DLn	1964: 1–2002: 12	New construction put in place—public (c30)
condo9	Ln	1963: 1–2002: 12	Construct.contracts: comm'l and indus.bldgs (mil.sq.ft.floor sp.;sa)
hniv	Ln	1963: 1–2002: 12	New 1-family houses for sale at end of month (thous, sa)

Table 1. Description of Data

Series	Trans.	Sample Period	Description
hnr	Ln	1963: 1–2002: 12	New 1-family houses, month's supply @ current sales rate (ratio)
hns	Ln	1963: 1–2002: 12	New 1-family houses sold during month (thous, saar)
hsbr	Ln	1959: 1–2002: 12	Housing authorized: total new priv housing units (thous., saar)
hswst	Ln	1959: 1–2002: 12	Housing starts: west (thous.u.) s.a.
hmob	Ln	1959: 1–2002: 12	Mobile homes: manufacturers' shipments (thous.of units, saar)
hsmw	Ln	1959: 1–2002: 12	Housing starts: midwest (thous.u.) s.a.
hsne	Ln	1959: 1–2002: 12	Housing starts: northeast (thous.u.) s.a.
hsfr	Ln	1959: 1–2002: 12	Housing starts: nonfarm (1947–58);total farm and nonfarm (1959-, thous., sa)
ivmtq	DLn	1959: 1–2001: 7	Mfg. and trade inventories: total (mil of chained 1996, sa)
ivmfgq	DLn	1959: 1–2001: 7	Inventories, business, mfg. (mil of chained 1996 dollars, sa)
ivmfdq	DLn	1959: 1–2001: 7	Inventories, business durables (mil of chained 1996 dollars, sa)
ivmfnq	DLn	1959: 1–2001: 7	Inventories, business, nondurables (mil of chained 1996 dollars, sa)
ivwrq	DLn	1959: 1–2001: 7	Mfg. and trade inv: merchant wholesalers (mil of chained 1996 dollars, sa)
ivrrq	DLn	1959: 1–2001: 7	Mfg. and trade inv: retail trade (mil of chained 1996 dollars, sa)
ivsrq	DLn	1959: 1–2001: 7	Ratio for mfg. and trade: inventory/sales (chained 1996 dollars, sa)
ivsrmq	DLn	1959: 1–2001: 7	Ratio for mfg. and trade: mfg.;inventory/ sales (1996 $ s.a.)
ivsrwq	DLn	1959: 1–2001: 7	Ratio for mfg. and trade: wholesaler; inventory/sales (1996 $ s.a.)
ivsrrq	DLn	1959: 1–2001: 7	Ratio for mfg. and trade: retail trade;inventory/sales(1996 $ s.a.)
pmi	Lev	1959: 1–2002: 12	Purchasing managers' index (sa)
pmp	Lev	1959: 1–2002: 12	Napm production index (percent)
pmno	Lev	1959: 1–2002: 12	Napm new orders index (percent)
pmdel	Lev	1959: 1–2002: 12	Napm vendor deliveries index (percent)
pmnv	Lev	1959: 1–2002: 12	Napm inventories index (percent)
pmemp	Lev	1959: 1–2002: 12	Napm employment index (percent)
pmcp	Lev	1959: 1–2002: 12	Napm commodity prices index (percent)
mocmq	DLn	1959: 1–2002: 12	New orders (net)—consumer goods and materials, 1996 dollars (bci)
msondq	DLn	1959: 1–2002: 12	New orders, nondefense capital goods, in 1996 dollars (bci)
moq	DLn	1959: 1–2001: 5	Mfg. new orders: all manufacturing industries, total, real (mo/pwfsa, AC)
mdoq	DLn	1959: 1–2001: 5	Mfg. new orders: durable goods industries, total, real (mdo/pwfsa, AC)
muq	DLn	1959: 1–2001: 5	Mfg. unfilled orders: all manufacturing industries, total mu/pwfsa, AC)
mduq	DLn	1959: 1–2001: 5	Mfg. unfilled orders: durable goods industries, total (mdu/pwfsa, AC)
(D) Interest rates and asset prices			
fygt10	D	1959: 1–2002: 12	Interest rate: U.S. treasury const maturities, 10-yr.(% per ann, nsa)
fclnq	DLn	1959: 1–2002: 12	Commercial and industrial loans oustanding in 1996 dollars (bci)
fsncom	DLn	1959: 1–2002: 12	Nyse common stock price index: composite (12/31/65 = 50)
fsnin	DLn	1966: 1–2002: 12	Nyse common stock price index: industrial (12/31/65 = 50)
fsntr	DLn	1966: 1–2002: 12	Nyse common stock price index: transportation (12/31/65 = 50)
fsnut	DLn	1966: 1–2002: 12	Nyse common stock price index: utility (12/31/65 = 50)
fsnfi	DLn	1966: 1–2002: 12	Nyse common stock price index: finance (12/31/65 = 50)
fspcom	DLn	1959: 1–2002: 12	S&p's common stock price index: composite (1941-43 = 10)
fspin	DLn	1959: 1–2002: 12	S&p's common stock price index: industrials (1941-43 = 10)
fsdxp	Lev	1959: 1–2002: 12	S&p's composite common stock: dividend yield (% per annum)
fspxe	Lev	1959: 1–2002: 12	S&p's composite common stock: priceearnings ratio (%, nsa)
fyff	D	1959: 1–2002: 12	Interest rate: federal funds (effective) (% per annum, nsa)

Table 1. Description of Data

Series	Trans.	Sample Period	Description
fygm3	D	1959: 1–2002: 12	Interest rate: U.S. treasury bills, sec mkt, 3-mo. (% per ann, nsa)
fygm6	D	1959: 1–2002: 12	Interest rate: U.S. treasury bills, sec mkt, 6-mo. (% per ann, nsa)
fygt1	D	1959: 1–2002: 12	Interest rate: U.S. treasury const maturities, 1-yr. (% per ann, nsa)
fygt5	D	1959: 1–2002: 12	Interest rate: U.S. treasury const maturities, 5-yr. (% per ann, nsa)
fm2dq	DLn	1959: 1–2002: 2	Money supply—m2 in 1996 dollars (bci)
fyaaac	D	1959: 1–2002: 12	Bond yield: moody's aaa corporate (% per annum)
fybaac	D	1959: 1–2002: 12	Bond yield: moody's baa corporate (% per annum)
fymcle	D	1963: 1–2002: 12	Effective interest rate: conventional home mtge loans closed (%)
sfygm3	Lev	1959: 1–2002: 12	Fygm3-fyff (AC)
sfygm6	Lev	1959: 1–2002: 12	Fygm6-fyff (AC)
sfygt1	Lev	1959: 1–2002: 12	Fygt1-fyff (AC)
sfygt5	Lev	1959: 1–2002: 12	Fygt5-fyff (AC)
sfygt10	Lev	1959: 1–2002: 12	Fygt10-fyff (AC)
sfyaaac	Lev	1959: 1–2002: 12	Fyaaac-fyff (AC)
sfybaac	Lev	1959: 1–2002: 12	Fybaac-fyff (AC)
sfymcle	Lev	1963: 1–2002: 12	Fymcle-fyff (AC)
exrus	DLn	1959: 1–2002: 12	United states; effective exchange rate (merm, index no.)
exrsw	DLn	1959: 1–2002: 12	Foreign exchange rate: switzerland (swiss franc per U.S. $)
exrjan	DLn	1959: 1–2002: 12	Foreign exchange rate: japan (yen per U.S. $)
exruk	DLn	1959: 1–2002: 12	Foreign exchange rate: united kingdom (cents per pound)
exrcan	DLn	1959: 1–2002: 12	Foreign exchange rate: canada (canadian $ per U.S. $)
(E) Nominal prices, wages, and money			
fm1	DLn	1959: 1–2002: 12	Money stock: m1(curr, trav. cks, dem dep, other ck'able dep, bil$, sa)
fm2	DLn	1959: 1–2002: 12	Money stock: m2 (m1+o'nite rps, euro$, g/p&b/d mmmfs & sav & sm time dep, bil$)
fm3	DLn	1959: 1–2002: 12	Money stock: m3 (m2+lg time dep, term rp's & inst only mmmfs, bil$, sa)
fmfba	DLn	1959: 1–2002: 12	Monetary base, adj for reserve requirement changes (mil$, sa)
fmrra	DLn	1959: 1–2002: 12	Depository inst reserves: total, adj for reserve req chgs (mil$, sa)
leh	DLn	1964: 1–2002: 12	Avg hr earnings of prod. wkrs.: total private nonagric ($, sa)
lehcc	DLn	1959: 1–2002: 12	Avg hr earnings of constr. wkrs.: construction ($, sa)
lehm	DLn	1959: 1–2002: 12	Avg hr earnings of prod. wkrs.: manufacturing ($, sa)
lehtu	DLn	1964: 1–2002: 12	Avg hr earnings of nonsupv. wkrs.: trans & public util($, sa)
lehtt	DLn	1964: 1–2002: 12	Avg hr earnings of prod. wkrs.: wholesale & retail trade (sa)
lehfr	DLn	1964: 1–2002: 12	Avg hr earnings of nonsupv. wkrs.: finance, insur, real est ($, sa)
lehs	DLn	1964: 1–2002: 12	Avg hr earnings of nonsupv. wkrs.: services ($, sa)
pwfsa	DLn	1959: 1–2002: 12	Producer price index: finished goods (82 = 100, sa)
pwfcsa	DLn	1959: 1–2002: 12	Producer price index: finished consumer goods (82 = 100, sa)
pwimsa	DLn	1959: 1–2002: 12	Producer price index: intermed mat.supplies and components (82 = 100, sa)
Pwcmsa	DLn	1959: 1–2002: 12	Producer price index: crude materials (82 = 100, sa)
Pwfxsa	DLn	1967: 1–2002: 12	Producer price index: finished goods,excl. foods (82 = 100, sa)
psm99q	DLn	1959: 1–2002: 12	Index of sensitive materials prices (1990 = 100, bci-99a)
punew	DLn	1959: 1–2002: 12	Cpi-u: all items (82–84 = 100, sa)
pu81	DLn	1967: 1–2002: 12	Cpi-u: food and beverages (82–84 = 100, sa)
puh	DLn	1967: 1–2002: 12	Cpi-u: housing (82–84 = 100, sa)

Table 1. Description of Data

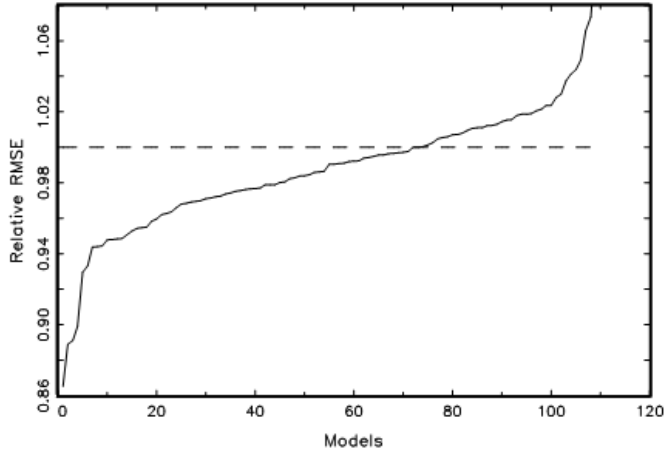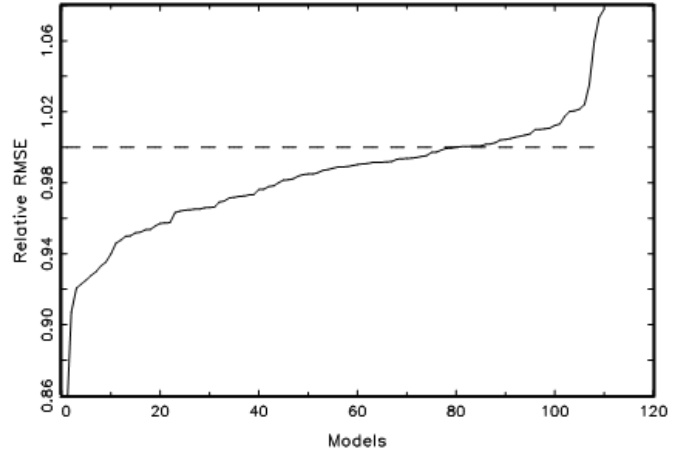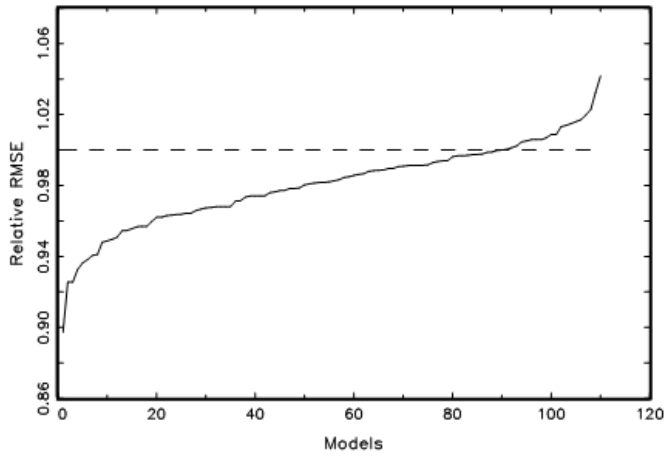| Series	Trans.	Sample Period	Description
pu83	DLn	1959: 1–2002: 12	Cpi-u: apparel and upkeep (82–84 = 100, sa)
pu84	DLn	1959: 1–2002: 12	Cpi-u: transportation (82–84 = 100, sa)
pu85	DLn	1959: 1–2002: 12	Cpi-u: medical care (82–84 = 100, sa)
pu882	DLn	1959: 1–2002: 12	Cpi-u: nondurables (1982–84 = 100, sa)
puc	DLn	1959: 1–2002: 12	Cpi-u: commodities (82–84 = 100, sa)
pucd	DLn	1959: 1–2002: 12	Cpi-u:durables (82–84 = 100, sa)
pus	DLn	1959: 1–2002: 12	Cpi-u: services (82–84 = 100, sa)
puxf	DLn	1959: 1–2002: 12	Cpi-u: all items less food (82–84 = 100, sa)
puxhs	DLn	1959: 1–2002: 12	Cpi-u: all items less shelter (82–84 = 100, sa)
puxm	DLn	1959: 1–2002: 12	Cpi-u: all items less midical care (82–84 = 100, sa)
gmdc	DLn	1959: 1–2002: 12	Pce, impl pr defl: pce (1987 = 100)
gmdcd	DLn	1959: 1–2002: 12	Pce, impl pr defl: pce; durables (1987 = 100)
gmdcn	DLn	1959: 1–2002: 12	Pce, impl pr defl: pce; nondurables (1996 = 100)
gmdcs	DLn	1959: 1–2002: 12	Pce, impl pr defl: pce; services (1987 = 100)

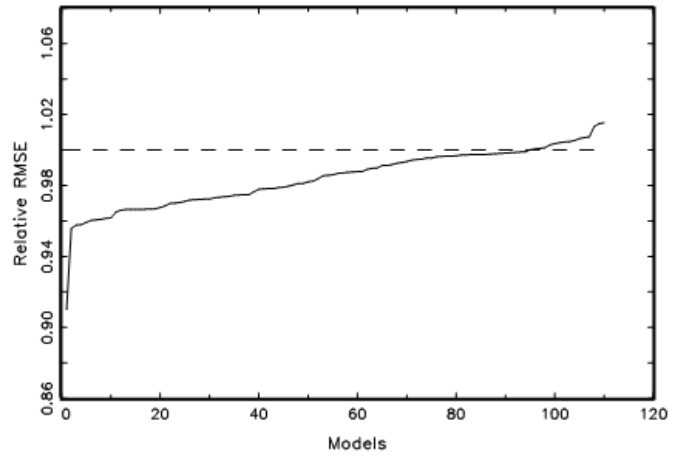**Table 2. Relative Performance of Algorithm Forecasts 1/**

	Significance level:									
	0.01	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45
Training sample size:										
All available										
Relative RMSE	1.09	1.01	1.00	0.99	0.98	0.98	0.98	0.98	0.98	0.98
Sum of ranks	976	770	673	582	518	498	498	448	497	590
Number of unencompassed	1	4	7	10	13	17	21	25	31	39
Last 20										
Relative RMSE	1.09	1.06	1.04	1.02	1.01	1.00	1.00	1.00	0.99	0.99
Sum of ranks	1,013	925	812	692	571	472	438	414	366	347
Number of unencompassed	1	2	3	4	5	6	8	11	15	24

1/ The relative RMSE indicates the ratio of MSE of algorithm forecasts to that of the simple averaging method. The table reports averages across 110 different experiments. The rankings compare the performance of different significance levels; the significance level that gives the best results is ranked one. The smallest sum of ranks across 110 models performs the best. The last rows in panels indicate the number of models that remain in the combination after the algorithm is applied.

Table 3. A Comparison of Forecast Combination Methods 1/

Rank	Mean of Relative MAD	Rank	Mean of Relative MSE
Top 30 PRE	0.981	Top 30 PRE	0.978
Top 80 POST	0.982	Top 80 POST	0.979
Top 90 POST	0.983	Top 90 POST	0.979
Top 70 POST	0.983	Top 20 PRE	0.980
Top 20 PRE	0.983	Top 70 POST	0.980
RMSE weighted POST	0.983	Top 40 PRE	0.980
Mean POST	0.984	RMSE weighted POST	0.981
Top 40 PRE	0.984	Mean POST	0.981
Top 60 POST	0.984	Top 60 POST	0.982
Top 50 POST	0.985	Top 50 POST	0.984
Top 50 PRE	0.988	Top 50 PRE	0.985
Top 40 POST	0.988	Median POST	0.988
Median POST	0.989	Top 40 POST	0.988
Top 60 PRE	0.991	Top 60 PRE	0.989
RANK weighted POST	0.993	RANK weighted POST	0.991
Top 30 POST	0.994	RANK weighted PRE	0.992
Top 70 PRE	0.994	Top 70 PRE	0.993
RANK weighted PRE	0.995	Top 30 POST	0.996
Top 80 PRE	0.997	Top 80 PRE	0.996
Top 90 PRE	0.998	Top 90 PRE	0.998
RMSE weighted PRE	0.999	top 10 PRE	0.998
top 10 PRE	1.000	RMSE weighted PRE	0.999
Top 20 POST	1.009	Top 20 POST	1.012
Median PRE	1.014	Median PRE	1.015
Top 5 PRE	1.029	Top 5 PRE	1.025
FAC BIC LAGS	1.048	FAC AIC LAGS	1.043
FAC AIC LAGS	1.050	FAC BIC LAGS	1.043
top 10 POST	1.052	FAC FIXED LAGS	1.044
FAC FIXED LAGS	1.054	FAC AIC FACTORS	1.049
FAC AIC FACTORS	1.062	top 10 POST	1.051
FAC BIC FACTORS	1.063	FAC BIC FACTORS	1.051
FAC FIXED FACTORS	1.067	FAC FIXED FACTORS	1.055
Top 5 POST	1.092	Top 5 POST	1.089
Top 1 PRE	1.098	Top 1 PRE	1.094
Best POST	1.104	Best PRE	1.100
Top 1 POST	1.104	Best POST	1.100
Best PRE	1.105	Top 1 POST	1.100

1/ The relative root mean absolute errors indicates the ratio of the MAD of a combination method to that of the simple averaging method. Same for the RMSE. Methods are ranked from the best performing to the worst.

Figure 1. Relative Performance of Algorithm Forecasts

Figure 1a: Results for sl=0.15

Figure 1b: Results for sl=0.25

Figure 1c: Results for sl=0.35

Figure 1c: Results for sl=0.45

REFERENCES

Aiolfi and A. Timmerman, 2006, "Persistence of Forecasting Performance and Combination Strategies," *Journal of Econometrics*, 135, 31–53.

Clemen, R.T., 1989, "Combining Economic Forecasts: A Review and Annotated Biography," *International Journal of Forecasting*, 5, 559–81.

Diebold, F. X., and R. Mariano, 1995, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics,* 13, 253–263.

Diebold, F.X., and P. Pauly, 1987, "Structural Change and the Combination of Forecasts," *Journal of Forecasting,* 6, 21–40.

Diebold, F.X., and P. Pauly, 1990, "The Use of Prior Information in Forecast Combinations," *International Journal of Forecasting*, 6, 503–508.

Granger, C.W.J., and Y. Jeon, 2004, "Thick Modeling." *Economic Modeling*, 21, 323–343.

Granger, C.W.J., and R. Ramanathan, 1984, "Improved Methods of Combining Forecasts," *Journal of Forecasting,* 3, 197–204.

Gupta, S., and P.C. Wilton, 1987, "Combination of Forecasts: An Extension," *Management Science,* 33, 356–372.

Harvey I. D., S. J. Leybourne, and P. Newbold, 1997, 'Testing the Equality of Prediction Mean Squared Errors," *International Journal of Forecasting,* 13, 281–91.

Harvey I. D., S. J. Leybourne, and P. Newbold, 1998, "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics,* 16, 254–59.

Newbold, P., and D.I. Harvey, 2002, "Forecast Combination and Encompassing," in Clements, M.P. and D.F. Hendry (eds), *A Companion to Economic Forecasting*. Oxford: Blackwells.

Marcellino, M., J.H. Stock, and M. Watson, 2006, "A Comparison of Direct and Iterated Multistep AR Methods for Forecasting Macroeconomic Time Series," *Journal of Econometrics*, 135, 499–526.

McCracken M. W., and K.D. West, 2002, "Inference About Predictive Ability," in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Oxford: Basil Blackwell, pp. 299–321.

Schmittlein, D.C., J. Kim, and D.G. Morrison, 1990, "Combining Forecasts: Operational Adjustments to Theoretically Optimal Rules," *Management Science,* 39, 1044–1056.

Stock, J. H., and M. Watson, 2002, "Forecasting Using Principal Components From a Large Number of Predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

Stock, J. H., and M. Watson, 2006, "Forecasting with Many Predictors,", in G. Elliott, C.W.J. Granger and A. Timmermann (eds) *Handbook of Economic Forecasting*, North-Holland: Elsevier. Pages 515–554.

Swanson, N.R., and T. Zang, 2001, "Choosing Among Competing Econometric Forecasts: Regression-based Forecast Combination Using Model Selection," *Journal of Forecasting,* 20, 425–440.

Timmerman A., 2006, "Forecast Combinations," in G. Elliott., C.W.J. Granger, and A. Timmerman (eds), *Handbook of Economic Forecasting*, North-Holland: Elsevier. Pages 135–196

Zellner, A., C. Hong, and C. Min, 1991, "Forecasting Turning Points in International Output Growth Rates Using Bayesian Exponentially Weighted Autoregression, Time-varying Parameter, and Pooling Techniques," *Journal of Econometrics,* 49, 275–304.