Journal of Forecasting
J. Forecast. **23**, 405–430 (2004)
Published online in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/for.928

Combination Forecasts of Output Growth in a Seven-Country Data Set

JAMES H. STOCK[1]* AND MARK W. WATSON[2]
[1] *Department of Economics, Harvard University and the National Bureau of Economic Research, USA*
[2] *Woodrow Wilson School and Department of Economics, Princeton University and the National Bureau of Economic Research, USA*

ABSTRACT

This paper uses forecast combination methods to forecast output growth in a seven-country quarterly economic data set covering 1959–1999, with up to 73 predictors per country. Although the forecasts based on individual predictors are unstable over time and across countries, and on average perform worse than an autoregressive benchmark, the combination forecasts often improve upon autoregressive forecasts. Despite the unstable performance of the constituent forecasts, the most successful combination forecasts, like the mean, are the least sensitive to the recent performance of the individual forecasts. While consistent with other evidence on the success of simple combination forecasts, this finding is difficult to explain using the theory of combination forecasting in a stationary environment.

KEY WORDS macroeconomic forecasting; high-dimensional forecasting; time-varying parameters; forecast pooling

INTRODUCTION

Historically, time series forecasting of economic variables has focused on low-dimensional models such as autoregressions, single-equation regressions using leading indicators as predictors, or vector autoregressions with perhaps a half-dozen or fewer variables. These low-dimensional models potentially omit information contained in the thousands of variables available to real-time economic forecasters. To forecast using many predictors, one needs to impose sufficient restrictions that the number of estimated parameters is kept small. One way to impose such restrictions on high-dimensional systems is to suppose that the variables have a dynamic factor structure, and recent research (e.g. Stock and Watson, 1999a, 2002a; Forni *et al*., 2000, 2001) suggests that there are potential gains from forecasting using high-dimensional dynamic factor models. There are, however, other ways to

* Correspondence to: James H. Stock, Department of Economics, Littauer Center, Harvard University, Cambridge, MA 02138-3001, USA. E-mail: james_stock@harvard.edu

impose structure on high-dimensional forecasting models, and one such way is to apply the methods of the forecast combining literature.[1]

This paper has two objectives. The first is to evaluate and compare the empirical performance of various combination forecasts of the growth rate of real output using a data set which covers seven OECD countries from 1959 to 1999 and, for each country, contains up to 73 recursively produced forecasts based on individual predictors. In previous work with this data set (Stock and Watson, 2003), we found that the performance of the individual forecasts was unstable; whether a predictor worked well depended on the current economic shocks and institutional and policy particulars. Surprisingly, however, a preliminary investigation found that some simple combination forecasts—the median and the trimmed mean of the panel of forecasts—were stable and reliably outperformed a univariate autoregressive benchmark forecast. Here, we extend that analysis to consider more sophisticated combination forecasts. The theory of combination forecasting suggests that methods that weight better-performing forecasts more heavily will perform better than simple combination forecasts, and that further gains might be obtained by introducing time variation in the weights or by discounting observations in the distant past. We find that most of the combination forecasts have lower mean squared forecast errors (MSFEs) than the benchmark autoregression. The combination methods with the lowest MSFEs are, intriguingly, the simplest, either with equal weights (the mean) or with weights that are very nearly equal and change little over time. The simple combination forecasts perform stably over time and across countries—much more stably than the individual forecasts constituting the panel.

The second objective of this paper is to compare combination forecasts to forecasts formed using a dynamic factor model, where the factors are estimated (country by country) using a panel of predictor series. We find that the combination forecasts generally outperform the forecasts produced using dynamic factor methods.

The data are described in the next section, and the combination forecast methods are described in the third section. Empirical results are presented in the fourth section, and a final section concludes.

THE SEVEN-COUNTRY DATA SET AND INDIVIDUAL FORECASTS

This section briefly summarizes the seven-country data set and the panel of forecasts constructed using the individual predictors in that data set.

The data

The seven-country data set is the same as used in Stock and Watson (2003). The data consist of up to 43 time series for each of seven developed economies (Canada, France, Germany, Italy, Japan, the UK and the USA) over 1959–1999 (some series are available only for a shorter period). The 43 series consist of various asset prices (including returns, interest rates and spreads); selected measures of real economic activity; wages and prices; and measures of the money stock. The list of series is given in Table Ia. All the analysis in this paper is done at quarterly frequency.

The data were subjected to five possible transformations, done in the following order. First, in a few cases the series contained a large outlier, such as spikes associated with strikes, and these

[1] For introductions to forecast combination methods and surveys of the large literature, see Diebold and Lopez (1996), Newbold and Harvey (2002) and Hendry and Clements (2002). Clemen (1989) provides a comprehensive survey of the literature through the late 1980s, and Makridakis and Hibon (2000) report recent results on combination forecasts.

Table Ia. Series in the seven-country data set

Series label	Sampling frequency	Description
		Asset prices
rovnght	M	Interest Rate: overnight
rtbill	M	Interest Rate: short term Gov. Bills
rbnds	M	Interest Rate: short term Gov. Bonds
rbndm	M	Interest Rate: medium term Gov. Bonds
rbndl	M	Interest Rate: long term Gov. Bonds
rrovnght	Q	Real overnight rate: rovnght – CPI Inflation
rrtbill	Q	Real short term bill rate: rtbill – CPI Inflation
rrbnds	Q	Real short term bond rate: rtbnds – CPI Inflation
rrbndm	Q	Real med. term bond rate: rtbndm – CPI Inflation
rrbndl	Q	Real long term bond rate: rtbndl – CPI Inflation
rspread	M	Term Spread: rbndl – rovnght
exrate	M	Nominal Exchange Rate
rexrate	M	Real Exchange Rate (exrate × relative CPIs)
stockp	M	Stock Price Index
rstockp	M	Real Stock Price Index: stockp/CPI
divpr	Q	Dividend Price Index
house	Q	House Price Index
rhouse	Q	Real House Price Index
gold	M	Gold Prices
rgold	M	Real Gold Prices
silver	M	Silver Prices
rsilver	M	Real Silver Prices
		Activity
rgdp	M	Real GDP
ip	M	Index of Industrial Production
capu	M&Q	Index of Capacity Utilization
emp	M&Q	Employment
unemp	M&Q	Unemployment Rate
pgdp	Q	GDP Deflator
cpi	M	Consumer Price Index
ppi	M	Producer Price Index
		Wages, goods and commodity prices
earn	M	Wages
commod	M	Commodity Price Index
oil	M	Oil Prices
roil	M	Real Oil Prices
rcommod	M	Real Commodity Price Index
		Money
m0	M	Money: M0 or Monetary Base
m1	M	Money: M1
m2	M	Money: M2
m3	M	Money: M3
rm0	M	Real Money: M0
rm1	M	Real Money: M1
rm2	M	Real Money: M2
rm3	M	Real Money: M3

Notes: M indicates that the original data are monthly, Q indicates that they are quarterly, M&Q indicates that monthly data were available for some countries but quarterly data were available for others. All forecasts and regressions use quarterly data, which were aggregated from monthly data by averaging (for CPI and IP) or by using the last monthly value (all other series).

outliers were replaced by interpolated values. Second, series that showed significant seasonal variation were seasonally adjusted using a linear approximation to X11. Third, when the data were available on a monthly basis, the data were aggregated to quarterly observations. Fourth, in some cases the data were transformed by taking logarithms. Fifth, the highly persistent or trending variables were differenced, second differenced, or computed as a 'gap', that is, a deviation from a stochastic trend. The gaps here were estimated using a one-sided Hodrick–Prescott (1981) filter, which maintains the temporal ordering of the series. For additional details, see Stock and Watson (2003).

In many cases we used more than one version (transformation) of a given series, for example, interest rates were used both in levels and in first differences. The series and transformations used in the full data set are listed by country in Table Ib. Counting all the constructed variables (like spreads) and different versions of the same variable that differ only in the transformation, the maximum number of series per country is 75 and the maximum number of predictors considered is 73 (75 minus the output measure being predicted and its associated output gap variable).

Some of the procedures considered in this paper require a forecasting track record to estimate forecast combining weights. Because the full data set contains some series that are available for short subsamples, we therefore also use two balanced panel subsets of this full data set. The first, the 'forecast combining balanced panel', includes between 27 and 66 series (and transformations) per country; these are the subset of series that are available since at least 1963:I. The second balanced panel, the 'dynamic factor model (dfm) balanced panel', is a subset of the first balanced panel, where the series in the dfm balanced panel were chosen to be approximately integrated of order zero, in keeping with the theoretical development of dynamic factor model forecasts in Stock and Watson (2002b). This subset contains between 9 and 23 series per country. Table Ib specifies the series in the two subsets.

Individual forecasts

The forecasts based on individual predictors are computed using h-step-ahead projections. Specifically, let $Y_t = \Delta \ln Q_t$, where Q_t is the level of output (either the level of real GDP or the Index of Industrial Production), and let X_t be a candidate predictor (e.g. the term spread). Let Y^h_{t+h} denote output growth over the next h quarters, expressed at an annual rate, that is, let $Y^h_{t+h} = (400/h)\ln(Q_{t+h}/Q_t)$. The forecasts of Y^h_{t+h} are made using the h-step-ahead regression model

$$Y^h_{t+h} = \beta_0 + \beta_1(L)X_t + \beta_2(L)Y_t + u^h_{t+h} \tag{1}$$

where u^h_{t+h} is an error term and β_1(L) and β_2(L) are lag polynomials. Forecasts are computed for $h = 2$, 4, 8-quarter horizons.

Model selection and coefficient estimation are done using pseudo out-of-sample methods. Specifically, the coefficients in (1) are estimated recursively using OLS, so that the forecast of Y^h_{t+h} made at date t with estimated coefficients, $\hat{Y}^h_{t+h|t}$, is entirely a function of data for dates 1, . . . , t. Lag lengths are determined recursively using the AIC with between one and four lags of X_t (we refer to X_t in (1) as the first lag because it is lagged relative to Y^h_{t+h}) and between zero and four lags of Y_t.

Two univariate benchmark forecasts are used. The first is a multistep autoregressive (AR) forecast, in which (1) is estimated recursively with no X_t predictor and the lag length is chosen recursively by AIC (between zero and four). The second is a recursive random walk forecast, in which $\hat{Y}^h_{t+h|t} = \hat{\mu}_t$, where $\hat{\mu}_t$ is the sample average of $400Y_s$, $s = 1, \ldots, t$.

Table Ib. Series in full data set and balanced panel subsets

Series	Transformation	Country						
		Canada	France	Germany	Italy	Japan	UK	USA
rovnght	level	c	c	b	c	b	c	b
rtbill	level	b	c	c	c		c	b
rbnds	level				c		c	b
rbndm	level				b			b
rbndl	level	b	b	b	b	c	b	b
rovnght	Δ	c	c	a	c	a	c	a
rtbill	Δ	a	c	c	c		c	a
rbnds	Δ				c		c	a
rbndm	Δ				a			a
rbndl	Δ	a	a	a	a	c	a	a
rrovnght	level	c	c	b	c	b	c	b
rrtbill	level	b	c	c	c		c	b
rrbnds	level				c		c	b
rrbndm	level				b			b
rrbndl	level	b	b	b	b	c	b	b
rrovnght	Δ	c	c	b	c	b	c	b
rrtbill	Δ	b	c	c	c		c	b
rrbnds	Δ				c		c	b
rrbndm	Δ				b			b
rrbndl	Δ	b	b	b	b	c	b	b
rspread	level	c	c	a	c	c	c	a
exrate	Δln	c	c	c	c	c	c	c
rexrate	Δln	c	c	c	c	c	c	c
stockp	Δln	b	b	b	b	b	b	b
rstockp	Δln	a	a	a	a	a	a	a
divpr	ln	c	c	c	c	c	c	b
house	Δln	c				c	c	c
rhouse	ln	c				c	c	c
rhouse	Δln	c				c	c	c
gold	Δln	b	b	b	b	b	b	b
gold	Δ^2ln	b	b	b	b	b	b	b
rgold	ln	b	b	b	b	b	b	b
rgold	Δln	a	a	a	a	a	a	a
silver	Δln	c	c	c	c	c	c	c
silver	Δ^2ln	c	c	c	c	c	c	c
rsilver	ln	c	c	c	c	c	c	c
rsilver	Δln	c	c	c	c	c	c	c
rgdp	Δln	a	c	a	a	a	a	a
rgdp	gap	b	c	b	b	b	b	b
ip	Δln	a	a	a	a	a	a	a
ip	gap	b	b	b	b	b	b	b
capu	level	b	c	c	c	c		a
emp	Δln	a	c	a		a	a	a
emp	gap	b	c	b		b	b	b
unemp	level	b	c	b	b	b	b	b
unemp	Δ	a	c	c	a	a	a	a
unemp	gap	b	c	b	b	b	b	b
pgdp	Δln	b	c	b	b	b	b	b
pgdp	Δ^2ln	a	c	a	a	a	a	a
cpi	Δln	b	b	b	b	b	b	b

Table Ib. *Continued*

Series	Transformation	Country						
		Canada	France	Germany	Italy	Japan	UK	USA
cpi	Δ^2ln	a	a	a	a	a	a	a
ppi	Δln	b		b	c	b	b	b
ppi	Δ^2ln	a		a	c	a	a	a
earn	Δln	b	b	c		b	c	b
earn	Δ^2ln	a	a	c		a	c	a
oil	Δln	b	b	b	b	b	b	b
oil	Δ^2ln	b	b	b	b	b	b	b
roil	ln	b	b	b	b	b	b	b
roil	Δln	a	a	a	a	a	a	a
commod	Δln	b	b	b	b	b	b	b
commod	Δ^2ln	b	b	b	b	b	b	b
rcommod	ln	b	b	b	b	b	b	b
rcommod	Δln	a	a	a	a	a	a	a
m0	Δln					c		a
m0	Δ^2ln					c		b
m1	Δln	a	c	a	c	c		a
m1	Δ^2ln	b	c	b	c	c		b
m2	Δln	c		a	c	c		a
m2	Δ^2ln	c		b	c	c		b
m3	Δln	c	a	c	c	c		a
m3	Δ^2ln	c	b	c	c	c		b
rm0	Δln					c		b
rm1	Δln	b	c	b	c	c		b
rm2	Δln	c		b	c	c		b
rm3	Δln	c	b	c	c	c		b
Number of series:								
dfm balanced panel (a)		15	9	15	11	13	12	23
combination forecast balanced panel (a and b)		43	27	43	33	35	33	66
full panel (a, b, c)		64	56	61	65	63	58	75

Notes: The 'dynamic factor model' data set consists of those series (and transformations) indicated by 'a'. The 'combination forecast balanced panel' data set consists of series marked 'a' or 'b'. The full (unbalanced panel) data set consists of series marked 'a,' 'b', or 'c.' The final rows give the total number of series contained in the various data sets, for forecasts made at the $h = 2$ horizon. In some cases, fewer series are available in the balanced panels for forecasts at the $h = 4$ and 8 horizons. In the transformations in the second column, 'level' means no transformation, 'gap' refers to one-sided HP detrending as discussed, Δ is the first-difference, Δln is the first-difference of the logarithm, and Δ^2ln is the second-difference of the logarithm.

All the individual-predictor forecasts considered in this paper are linear projections. There is evidence that combination forecasts that pool linear and nonlinear forecasts can outperform combination forecasts based solely on linear forecasts (e.g. Stock and Watson, 1999b; Blake and Kapetanios, 1999). Incorporating such nonlinear forecasts might improve upon the results reported here, but doing so would go beyond the linear framework of the dynamic factor model forecasts with which we wish to compare the combination forecasts.

COMBINATION FORECASTS AND FORECAST EVALUATION METHODS

Quite a few methods for pooling forecasts have been developed in the large literature on forecast combination. This section describes the combining methods studied in this paper and explains how they will be evaluated by comparing their pseudo out-of-sample forecasts.

Combination forecast methods

Five types of combination forecasts are considered in this paper: simple combination forecasts; discounted MSFE forecasts; shrinkage forecasts; factor model forecasts; and time-varying-parameter (TVP) combination forecasts. These methods differ in the way they use historical information to compute the combination forecast and in the extent to which the weight given an individual forecast is allowed to change over time. These methods, or closely related methods, have appeared previously in the forecast combining literature. Some standard methods for forecast combination, such as Granger–Ramanathan (1984) combining using regression weights, are inappropriate here, at least without some modifications, because of the large number of individual forecasts, relative to the sample size. The methods we use here are variants of linear forecast combinations; although there is evidence that nonlinear combination schemes can produce substantial gains (e.g. Deutsch *et al*., 1994), the number of constituent forecasts we consider arguably is too large for nonlinear combination methods to be effective.

Notation and estimation periods

Let $\hat{Y}^h_{i,t+h|t}$ denote the *i*th individual pseudo out-of-sample forecast of Y^h_{t+h}, computed at date t, that is, the *i*th forecast in the panel of forecasts for a given country. Most of the combination forecasts we consider are weighted averages of the individual forecasts (possibly with time-varying weights) and thus have the form

$$f_{t+h|t} = \sum_{i=1}^{n} w_{it} \hat{Y}^h_{i,t+h|t} \tag{2}$$

where $f_{t+h|t}$ is the combination forecast, w_{it} is the weight on the *i*th forecast in period t and n is the number of forecasts in the panel.

In general, the weights $\{w_{it}\}$ depend on the historical performance of the individual forecast. To evaluate this historical performance, we divide the sample into three periods. The observations prior to date T_0 are only used for estimation of the coefficients in the individual forecasting regression (1). The individual pseudo out-of-sample forecasts are computed starting in period T_0. The recursive MSFE of the *i*th individual forecast, computed from the start of the forecast period through date t, is

$$\text{MSFE}_{it} = \frac{1}{t - T_0 - 1} \sum_{s=T_0}^{t} \left(Y^h_{s+h} - \hat{Y}^h_{i,s+h|s}\right)^2 \tag{3}$$

The pseudo out-of-sample forecasts for the combination forecasts are computed over $t = T_1, \ldots, T_2$. For the empirical work reported in the next section, we used T_0 = 1973:I, T_1 = 1981:I + h and T_2 = min (1998:IV, $T_{\text{last}} - h$), where T_{last} is the end of the sample for that country.

Simple combination forecasts
The simple combination forecasts compute the combination forecast without regard to the historical performance of the individual forecasts in the panel. Three simple combination forecasts are used: the mean of the panel of forecasts (so $w_{it} = 1/n$ in (2)); the median; and the trimmed mean. The trimmed mean was computed with 5% symmetric trimming, subject to trimming at least one forecast.

Discounted MSFE forecasts
The discounted MSFE forecasts compute the combination forecast as a weighted average of the individual forecasts, where the weights depend inversely on the historical performance of each individual forecast (cf. Diebold and Pauly, 1987). Specifically, the discounted MSFE combination h-step-ahead forecast has the form (2), where the weights are

$$w_{it} = m_{it}^{-1} \Big/ \sum_{j=1}^{n} m_{jt}^{-1} \quad \text{where} \quad m_{it} = \sum_{s=T_0}^{t-h} \delta^{t-h-s} \left(Y_{s+h}^{h} - \hat{Y}_{i,s+h|s}^{h}\right)^2 \tag{4}$$

where δ is the discount factor.

The discounted MSFE forecasts are computed for three values of δ, $\delta = 1.0, 0.95, 0.9$. The case $\delta = 1$ (no discounting) corresponds to the Bates and Granger (1969) optimal weighting scheme when the individual forecasts are uncorrelated.

A related combination forecast is the 'most recently best', which as implemented here places all weight on the individual forecast that has the lowest average squared forecast error over the previous four periods.

Shrinkage forecasts
The shrinkage forecasts compute the weights as an average of the recursive OLS estimator of the weights (the Granger–Ramanathan, 1984 estimator, imposing an intercept of zero) and equal weighting. That is, the shrinkage forecasts have the form (2), where

$$w_{it} = \lambda \hat{\beta}_{it} + (1 - \lambda)(1/n) \tag{5}$$

where $\hat{\beta}_{it}$ is the ith estimated coefficient from a recursive OLS regression of Y_{s+h}^{h} on $\hat{Y}_{1,s+h|s}^{h}, \ldots, \hat{Y}_{n,s+h|s}^{h}$ for $s = T_0, \ldots, t-h$ (no intercept) and where $\lambda = \max\{0, 1 - \kappa[n/(t - h - T_0 - n)]\}$, where κ is a constant that controls the amount of shrinkage towards equal weighting. The shrinkage forecasts were evaluated for $\kappa = 0.25, 0.5, 1$, with larger values corresponding to more shrinkage towards equal weighting (smaller λ).

The shrinkage forecast based on (5) can be interpreted as a Bayes estimator (see Diebold and Pauly, 1990). In that context, the weight κ could be estimated using empirical Bayes methods, however we do not pursue that here because of the difficulties that arise when the number of individual forecasts n is large relative to $t - T_0$.

Principal component forecast combination
Principal component forecast combination entails (i) recursively computing the first few principal components of estimated common factors of the panel of forecasts, (ii) estimating a regression of $Y_{s+h|s}^{h}$ onto these principal components, and (iii) forming the forecast based on this

regression. Reduction of the many forecasts to a few principal components provides a convenient method for allowing some estimation of factor weights, yet reduces the number of weights that must be estimated. This method has been used by Figlewski (1983), Figlewski and Urich (1983) and Chan *et al*. (1999). One reason to think that this method might work well is that, as mentioned in the Introduction, recent work on large forecasting models suggests that large macroeconomic data sets are well described by a few common dynamic factors that are useful for forecasting, and that the common factors can be estimated by principal components (Forni *et al*., 2000, 2001; Stock and Watson, 1999a, 2002a). The forecast combining application here differs from the usual dynamic factor model approach, which is examined later, because the individual series are used first to compute a panel of forecasts, then static common factors are estimated from this panel of forecasts.

The principal component forecasts are constructed as follows. Let $\hat{F}^h_{1,s}, \ldots, \hat{F}^h_{m,s}$ denote the first m principal components of $\hat{Y}^h_{1,s+h|s}, \ldots, \hat{Y}^h_{n,s+h|s}$ for $s = T_0, \ldots, t$, computed as the first m principal components of the uncentred second moment matrix of the recursive forecasts over $s = T_0, \ldots, t$.[2] The principal component combination forecasts are computed using the regression

$$Y^h_{s+h} = \alpha_1 \hat{F}^h_{1,s} + \cdots + \alpha_m \hat{F}^h_{m,s} + v^h_{s+h} \tag{6}$$

where the regression coefficients $\alpha_1, \ldots, \alpha_m$ are estimated by OLS over the sample $s = T_0, \ldots, t - h$. The combined forecast is computed using the estimated weights, applied to $\hat{F}^h_{1,t}, \ldots, \hat{F}^h_{m,t}$.

Two versions of the principal component combination forecasts were computed, one with m chosen recursively by AIC, the other by BIC, where $1 \le m \le 4$.

Time-varying parameter forecasts

The TVP combination forecast uses the Kalman filter to estimate time-varying coefficients in the combining regression, where the coefficients are modelled as evolving according to a random walk. This method is used by Sessions and Chatterjee (1989) and by LeSage and Magura (1992). LeSage and Magura also extend it to mixture models of the errors, but that extension did not improve upon the simpler Kalman filter approach in their empirical application. Our implementation starts with the Granger–Ramanathan (1984) combining regression, modified to impose a zero intercept and extended to have time-varying parameters:

$$Y^h_{s+h} + \omega_{1t} \hat{Y}^h_{1,s+h|s} + \cdots \omega_{nt} \hat{Y}^h_{n,s+h|s} + \varepsilon^h_{s+h} \tag{7}$$

$\omega_{it} = \omega_{it-1} + \eta_{it}$, where η_{it} are serially uncorrelated, uncorrelated with ε^h_{s+h}, and uncorrelated across i. In principle, the relative variance $\text{var}(\eta_{it})/\text{var}(\varepsilon^h_{s+h})$ is estimable but with many forecasts its estimator could be quite unreliable, so instead we set the relative variance to $\text{var}(\eta_{it})/\text{var}(\varepsilon^h_{s+h}) = \phi^2/n^2$, where ϕ is a chosen parameter. Larger values of ϕ correspond to more time variation. The initial distribution of ω_{it} sets each weight to $1/n$ with zero variance; in the limit that $\phi = 0$, the TVP combination forecast thus reduces to the simple mean combination forecast. Three values of ϕ are investigated: $\phi = 0.1, 0.2, 0.4$. We found that performance of the TVP combination forecasts deteriorated sharply for larger values of ϕ than these.

[2] Because the forecasts are in the same units, the second moment matrix was computed without standardizing the individual forecasts, and the sample mean was not subtracted from the component forecasts.

Pseudo out-of-sample evaluation methods

The forecasting performance of a candidate combination forecast is evaluated by comparing its out-of-sample MSFE to the autoregressive benchmark. Specifically, let $\hat{Y}^h_{i,t+h|t}$ denote the pseudo out-of-sample forecast of Y^h_{t+h}, computed using data through time t, based on the ith combination forecast. Let $\hat{Y}^h_{0,t+h|t}$ denote the corresponding benchmark forecast made using the autoregression. Then the relative MSFE of the candidate combination forecast, relative to the benchmark forecast, is

$$\text{Relative MSFE} = \frac{\sum_{t=T_1}^{T_2} (Y^h_{t+h} - \hat{Y}^h_{i,t+h|t})^2}{\sum_{t=T_1}^{T_2} (Y^h_{t+h} - \hat{Y}^h_{0,t+h|t})^2} \tag{8}$$

where T_1 and T_2 are, respectively, the first and last dates over which the pseudo out-of-sample forecast is computed.

In principle, it is desirable to report standard errors for the relative MSFE (8), or to report p-values testing the null hypothesis that the relative MSFE is one. West (1996) obtained the null asymptotic distribution of (8) when the benchmark model 0 is not nested within the candidate forecast i. When the benchmark model is nested within the candidate model, the distribution of the relative MSFE, under the null hypothesis that $\beta_i(L) = 0$ in (1) and the other coefficients are constant, is non-standard and was obtained by Clark and McCracken (2001). In the analysis here, because of the recursive lag length selection, at some dates the two models are nested but at other dates they are not, and the null distribution of the relative MSFE is unknown. Moreover, it is not clear how applicable the West (1996) and Clark and McCracken (2001) distribution theory is when the parameter vector is very large, as is the case for the combination forecasts. For these reasons, in this paper we report relative MSFEs but not a measure of their statistical significance, leaving the latter to future work.[3]

EMPIRICAL RESULTS

This section examines the empirical performance of the combination forecasts constructed using the seven-country quarterly data set. We begin by briefly summarizing the performance of the individual forecasts that constitute the panel of forecasts.

Individual and simple combination forecasts

The individual forecasts for the seven-country data set are discussed and analysed in detail in Stock and Watson (2003). Consistent with the large literature on forecasting output growth using asset prices, some individual asset prices have predictive content for output in some time periods and in some countries. For example, the term spread (the yield on long-term government debt minus a short-term interest rate) was a potent predictor of output growth in the USA during the 1970s and early

[3] Clark–McCracken (2001) p-values are reported by Stock and Watson (2003) for fixed-lag versions (four lags) of the individual-indicator forecasts that constitute the panel of forecasts analysed here. The 5% critical value for the relative MSFEs typically range from 0.92 to 0.96 (the critical value depends on nuisance parameters and thus was computed on a series-by-series basis). By this gauge, many of the individual-indicator forecasts showed a significant improvement over the AR benchmark, at least in some periods and some countries.

1980s. There is, however, considerable instability in the performance of forecasts based on individual predictors: good performance in one period and country does not ensure good performance in another. Instead, performance of an individual predictor depends on the configuration of shocks hitting the economy, the current policy regime, and other institutional factors. For example, the term spread ceased to be a good predictor of output in the late 1980s and 1990s in the USA. As is discussed further in a later section, the individual forecasts, when used alone, perform worse on average than the AR.[4]

Comparison of alternative combination methods

The simple and recent best combination forecasts do not require an historical track record for the individual forecasts, and the discounted MSFE combination forecasts use only the past variances of the individual forecasts, not their covariances with the other forecasts in the panel. Thus these two methods are readily computed using the full data set, in which individual forecasts enter when there is enough data available on the predictor series to compute the forecasts.

The remaining combination methods require estimates of covariances among the panel of forecasts, so these are computed using the forecasting combination balanced panel subset of the full data set (see Table Ib). In addition, for comparability we also report the performance of the simple, recent best, and discounted MSFE combination forecasts, computed using the forecast combining balanced panel subset; doing so allows us to see whether there is a forecasting benefit associated with using the full, unbalanced data set, relative to the balanced subset.

The results for forecasts of real GDP growth over two, four and eight quarters are summarized in Tables II, III, and IV, respectively, and the results for IP growth over the three horizons are summarized in Tables V, VI and VII. In each of these tables, the entries for a candidate predictor (the row variable) are its MSFE for the forecast period (indicated in the first row), relative to the MSFE of the benchmark AR forecast. If the candidate predictor has a relative MSFE less than one, then it outperformed the AR benchmark over the forecast period in that country.

Several results emerge from Tables II–VII. First, many of the combination forecasts outperform the AR benchmark across countries, across horizons, and across the variable being forecasted.

Second, combination forecasts based on the full panel generally outperform their counterparts based on the balanced panel subset. Evidently the additional series in the full panel contain information useful for forecasting.

Third, although many of the improvements of the combination forecasts are modest, relative to the AR benchmark (relative MSFEs of 0.9 or 0.95), in some cases the gains are substantial (relative MSFEs of 0.85 or less).

Fourth, the simple combination forecasts show reliably good performance across different countries and horizons. Among the simple combination forecasts, there seems to be little difference between the mean and the trimmed mean. The median typically has somewhat higher relative MSFE than either the mean or trimmed mean.

Fifth, the shrinkage forecasts are not robust: for some countries and horizons they perform well, but for others they perform quite poorly. The less shrinkage, the less robust is the resulting combination forecast.

[4] The instability evident in the individual-predictor forecasts is consistent with other evidence of widespread instability in small econometric and time series models used for macroeconomic forecasting, see for example Stock and Watson (1996), Bernanke and Mihov (1998), Clements and Hendry (1999), Cogley and Sargent (2001, 2002), Sims and Zha (2002) and Marcellino (2002).

Table II. MSFEs of combination forecasts, relative to autoregression: forecasts of two-quarter growth of real GDP ($h = 2$)

Forecast period	Canada 81:III–98:IV	France –	Germany 81:III–98:IV	Italy 81:III–98:IV	Japan 81:III–98:IV	UK 81:III–98:IV	USA 81:III–98:IV
Univariate forecasts							
AR RMSFE	0.016	–	0.013	0.011	0.013	0.010	0.011
random walk	1.03	–	1.03	1.50	2.63	0.98	1.21
Combination forecasts, full panel							
median	0.90	–	0.97	0.92	0.95	0.98	0.99
mean	0.84	–	0.92	0.86	0.92	0.95	0.96
trimmed mean	0.86	–	0.93	0.87	0.93	0.96	0.97
disc. mse(0.9)	0.85	–	0.89	0.88	0.96	0.93	0.94
disc. mse(0.95)	0.85	–	0.90	0.89	0.96	0.94	0.94
disc. mse(1)	0.85	–	0.90	0.89	0.95	0.94	0.93
recent best	0.66	–	0.81	1.10	1.08	0.73	1.16
Combination forecasts, balanced panel subset							
median	0.96	–	0.97	1.01	1.00	0.97	0.98
mean	0.90	–	0.92	0.99	0.99	0.95	0.95
trimmed mean	0.92	–	0.93	1.01	1.00	0.95	0.96
disc. mse(0.9)	0.88	–	0.90	0.98	0.99	0.94	0.95
disc. mse(0.95)	0.88	–	0.91	0.99	0.99	0.94	0.94
disc. mse(1)	0.88	–	0.91	1.00	0.99	0.94	0.93
recent best	0.75	–	0.83	1.11	1.00	0.84	1.18
PC(BIC)	0.85	–	0.83	0.91	0.89	1.46	1.08
PC(AIC)	0.87	–	0.82	0.94	0.90	1.36	1.10
shrink(0.25)	0.82	–	1.87	1.39	0.95	1.54	0.96
shrink(0.5)	0.82	–	1.22	1.06	0.92	1.19	0.95
shrink(1)	0.89	–	0.93	0.95	0.96	0.95	0.95
tvp(0.1)	0.79	–	0.86	0.76	0.80	0.99	0.96
tvp(0.2)	0.78	–	0.86	0.70	0.81	1.04	0.99
tvp(0.4)	0.76	–	0.87	0.70	0.83	1.05	1.04

Notes: The entry in the row labelled AR RMSFE is the root mean squared forecast error of the benchmark autoregressive forecast (in decimal values of the h-period growth, i.e. not at an annual rate). The pseudo out-of-sample forecast period is given in the first row. The remaining entries are the MSFE of the forecast indicated in the first column, relative to the AR forecast. There are no entries for France because the GDP time series is too short. The forecast mnemonics are:

median	median of individual forecasts at date t
mean	average of individual forecasts at date t
trimmed mean	trimmed mean of individual forecasts at date t, 5% symmetric trimming
disc. mse(δ)	combining weights are inversely proportional to discounted forecast errors with discount factor δ
recent best	individual forecast with lowest average squared forecast error over past four quarters
PC(BIC), PC(AIC)	forecasts from regression onto principal components of the panel of forecasts; number of principal components determined by BIC or AIC
shrink(κ)	combining weights are linear combination of equal weighting and recursive OLS, with shrinkage weight $\max\{0, 1 - \kappa[n/(t - T_0 - n)]\}$
TVP(ϕ)	combining weights follow random walk, estimated by Kalman filter, with relative variance $\phi^2(t - T_0)/n^2$

Table III. MSFEs of combination forecasts, relative to autoregression: forecasts of four-quarter growth of real GDP ($h = 4$)

Forecast period	Canada 82:I–98:IV	France –	Germany 82:I–98:IV	Italy 82:I–98:IV	Japan 82:I–98:II	UK 82:I–98:IV	USA 82:I–98:IV
Univariate forecasts							
AR RMSE	0.025	–	0.018	0.019	0.023	0.018	0.016
random walk	0.99	–	1.05	1.31	2.97	0.96	1.04
Combination forecasts, full panel							
median	0.92	–	0.99	0.91	0.93	1.00	0.92
mean	0.88	–	1.00	0.82	0.88	0.98	0.90
trimmed mean	0.90	–	0.99	0.82	0.89	0.99	0.91
disc. mse(0.9)	0.90	–	0.98	0.85	0.93	0.94	0.90
disc. mse(0.95)	0.90	–	1.00	0.89	0.93	0.96	0.89
disc. mse(1)	0.91	–	1.00	0.91	0.92	0.97	0.87
recent best	0.85	–	1.26	0.71	0.97	0.80	1.67
Combination forecasts, balanced panel							
median	0.99	–	1.01	1.03	1.01	0.99	0.92
mean	0.96	–	1.05	1.06	0.98	0.94	0.89
trimmed mean	0.97	–	1.05	1.06	1.00	0.95	0.90
disc. mse(0.9)	0.94	–	1.03	1.01	0.97	0.93	0.91
disc. mse(0.95)	0.95	–	1.05	1.05	0.98	0.94	0.90
disc. mse(1)	0.95	–	1.05	1.08	0.98	0.94	0.88
recent best	0.89	–	1.23	0.98	0.91	1.11	1.69
PC(BIC)	0.82	–	1.05	0.76	0.68	1.43	0.95
PC(AIC)	0.75	–	1.11	0.77	0.67	1.39	0.98
shrink(0.25)	1.32	–	2.26	1.48	1.07	2.14	0.95
shrink(0.5)	1.09	–	1.37	1.20	1.02	1.49	0.88
shrink(1)	0.99	–	1.00	1.02	0.97	1.09	0.89
tvp(0.1)	0.85	–	0.91	0.55	0.63	1.11	0.98
tvp(0.2)	0.97	–	0.98	0.49	0.63	1.27	1.15
tvp(0.4)	1.07	–	1.11	0.52	0.65	1.33	1.41

Notes: See notes to Table II.

Sixth, the principal component (static factor regression) forecasts have quite variable performance, in some cases far outperforming the AR benchmark but in other cases performing much worse.

Seventh, the results for the methods designed to handle time variation are mixed. The TVP forecasts sometimes work well but sometimes work quite poorly, and in this sense are not robust; the larger is the amount of time variation, the less robust are the forecasts. Similarly, the discounted MSE forecasts with the most discounting ($\delta = 0.9$) are typically no better than, and sometimes worse than, their counterparts with less or no discounting ($\delta = 0.95$ or 1).

As discussed earlier, because the forecasting models are in some periods nested but in other periods non-nested, we have not performed formal tests of the null hypothesis that the combination forecasts provide no improvement over the AR benchmark. Thus the foregoing conclusions are based solely on the point estimate of the pseudo out-of-sample relative mean squared forecast error. An important but substantial remaining task is providing a measure of statistical precision for these relative MSFEs.

Table IV. MSFEs of combination forecasts, relative to autoregression: forecasts of eight-quarter growth of real GDP ($h = 8$)

Forecast period	Canada 83:I–97:IV	France –	Germany 83:I–97:IV	Italy 83:I–97:IV	Japan 83:I–97:II	UK 83:I–97:IV	USA 83:I–97:IV
Univariate forecasts							
AR RMSE	0.046	–	0.030	0.038	0.046	0.034	0.025
random walk	0.94	–	1.08	1.14	2.74	0.95	0.98
Combination forecasts, full panel							
median	0.95	–	0.98	0.82	0.95	1.04	0.99
mean	0.87	–	0.98	0.68	0.89	1.09	0.96
trimmed mean	0.89	–	0.96	0.71	0.90	1.07	0.98
disc. mse(0.9)	0.97	–	1.00	0.79	0.94	1.03	0.97
disc. mse(0.95)	0.96	–	1.00	0.81	0.93	1.03	0.96
disc. mse(1)	0.96	–	0.97	0.82	0.93	1.03	0.96
recent best	1.12	–	2.19	0.48	1.15	1.75	1.87
Combination forecasts, balanced panel subset							
median	1.00	–	1.01	1.02	1.01	1.01	1.00
mean	1.00	–	1.05	0.96	0.99	1.06	0.98
trimmed mean	0.99	–	1.04	1.02	1.00	1.03	0.99
disc. mse(0.9)	1.00	–	1.06	0.93	0.98	1.05	0.98
disc. mse(0.95)	0.99	–	1.06	0.95	0.99	1.05	0.97
disc. mse(1)	0.99	–	1.05	0.96	0.98	1.06	0.97
recent best	1.08	–	1.54	0.79	1.26	1.54	1.91
PC(BIC)	0.84	–	1.00	0.33	0.43	1.86	1.35
PC(AIC)	0.78	–	1.07	0.36	0.41	2.15	1.30
shrink(0.25)	1.68	–	1.58	1.20	0.44	3.65	1.15
shrink(0.5)	1.27	–	0.98	0.87	0.53	1.87	0.98
shrink(1)	1.00	–	1.02	0.78	0.78	1.17	0.98
tvp(0.1)	0.87	–	1.01	0.32	0.65	1.53	1.15
tvp(0.2)	1.08	–	1.24	0.37	0.69	1.94	1.41
tvp(0.4)	1.22	–	1.46	0.45	0.69	2.31	1.72

Notes: See notes to Table II.

Ranking combination forecasts by average loss

As a way to compare the combination methods, we computed an average estimated loss for each combination method, where the average is computed across all countries and across the two different measures of output. There are a total of 13 such cases (seven countries, two measures of output each, except for France for which the real GDP time series is too short). This average loss of a given combination forecast is computed as the weighted average of the MSFEs for the individual countries (where each MSFE is computed over $t = T_1, \ldots, T_2 - h$), where the country weights are the inverse of the full-sample ($t = 1, \ldots, T_2 - h$) variance of Y^h_{t+h}. Equivalently, the average loss of a given combination forecast is the unweighted average MSFE across countries, where each output measure is standardized to have a unit full-sample variance. One interpretation of this average loss is that it estimates the loss a forecaster would expect to have if she knew she would be forecasting output growth in a developed economy, but is not told which economy, which measure of output growth, or which horizon. The forecast that minimizes the population counterpart of this average loss is the forecast that has the lowest expected loss in the forecasting game in which the forecaster

Table V. MSFEs of combination forecasts, relative to autoregression: forecasts of two-quarter growth of IP ($h = 2$)

Forecast period	Canada 81:III–98:IV	France 81:III–98:IV	Germany 81:III–98:IV	Italy 81:III–98:II	Japan 81:III–98:IV	UK 81:III–98:IV	USA 81:III–98:IV
Univariate forecasts							
AR RMSE	0.031	0.018	0.026	0.028	0.026	0.018	0.019
random walk	1.17	1.20	1.00	1.07	2.35	1.00	1.30
Combination forecasts, full panel							
median	0.98	0.90	0.94	0.92	0.96	0.97	0.93
mean	0.92	0.89	0.90	0.90	0.94	0.97	0.88
trimmed mean	0.93	0.89	0.90	0.90	0.94	0.97	0.89
disc. mse(0.9)	0.92	0.91	0.89	0.93	0.96	0.96	0.88
disc. mse(0.95)	0.92	0.92	0.89	0.94	0.96	0.96	0.87
disc. mse(1)	0.92	0.92	0.88	0.93	0.96	0.96	0.85
recent best	0.77	1.16	1.01	1.32	1.05	1.23	1.17
Combination forecasts, balanced panel							
median	0.99	1.02	0.96	0.98	1.01	0.96	0.94
mean	0.93	1.08	0.90	1.00	1.02	0.98	0.89
trimmed mean	0.95	1.05	0.91	1.00	1.03	0.97	0.89
disc. mse(0.9)	0.93	1.03	0.90	0.98	1.01	0.97	0.89
disc. mse(0.95)	0.92	1.06	0.89	0.99	1.02	0.97	0.88
disc. mse(1)	0.92	1.08	0.89	1.00	1.02	0.98	0.86
recent best	0.85	1.12	1.06	1.21	1.07	1.31	1.14
PC(BIC)	0.99	1.05	0.87	0.83	1.00	1.05	0.85
PC(AIC)	0.92	1.17	0.85	0.84	1.00	1.05	0.82
shrink(0.25)	1.27	2.43	1.18	2.35	1.40	2.19	0.89
shrink(0.5)	1.04	1.63	1.02	1.16	1.16	1.45	0.89
shrink(1)	0.95	1.29	0.92	0.90	1.02	1.03	0.89
tvp(0.1)	0.92	0.97	0.88	0.93	0.92	0.97	0.89
tvp(0.2)	0.92	0.93	0.86	0.88	0.88	0.97	0.90
tvp(0.4)	0.94	0.96	0.84	0.87	0.90	0.98	0.91

Notes: See notes to Table II.

first chooses the combination method, then is assigned randomly a country, series and horizon to forecast.

The various combination forecasts (along with the random walk forecast), ranked by their average loss, are presented in Table VIII for all three horizons (39 cases averaged). The results are striking. When the loss is averaged over all countries, dependent variables and horizons, the best three combination forecasts are the TVP forecast with very little time variation, the simple mean and the trimmed mean; the performance of these three methods is very close numerically. (Recall that the TVP(0.1) forecast is nearly the simple mean combination forecast, with a small amount of time variation introduced.) These methods, and the other methods that do well, allow for little or no time variation in the weights applied to individual forecasts. In contrast, the combination methods that permit the greatest time variation in weights, or that rely the most on historical evidence to estimate the combination weights, exhibit the poorest performance, in some cases by a wide margin. These poorly performing combination methods include the shrinkage forecast with the least shrinkage, the

Table VI. MSFEs of combination forecasts, relative to autoregression: forecasts of four-quarter growth of IP ($h = 4$)

Forecast period	Canada 82:I–98:IV	France 82:I–98:IV	Germany 82:I–98:III	Italy 82:I–97:IV	Japan 82:I–98:IV	UK 82:I–98:IV	USA 82:I–98:IV
Univariate forecasts							
AR RMSE	0.047	0.031	0.037	0.041	0.052	0.026	0.029
random walk	0.96	1.05	1.06	1.11	1.95	1.01	1.09
Combination forecasts, full panel							
median	0.96	0.91	0.97	0.88	0.93	0.96	0.94
mean	0.90	0.91	0.95	0.84	0.88	0.93	0.85
trimmed mean	0.92	0.90	0.95	0.85	0.89	0.95	0.87
disc. mse(0.9)	0.95	0.93	0.95	0.90	0.90	0.93	0.86
disc. mse(0.95)	0.95	0.94	0.95	0.91	0.89	0.94	0.84
disc. mse(1)	0.95	0.93	0.93	0.91	0.86	0.94	0.83
recent best	1.40	1.04	1.09	0.85	1.05	1.03	2.33
Combination forecasts, balanced panel							
median	0.99	1.05	0.99	1.01	1.01	0.97	0.92
mean	0.96	1.16	0.98	1.03	1.02	0.95	0.86
trimmed mean	0.97	1.12	0.98	1.03	1.03	0.96	0.87
disc. mse(0.9)	0.96	1.08	0.97	0.99	1.01	0.94	0.88
disc. mse(0.95)	0.96	1.12	0.96	1.02	1.01	0.95	0.85
disc. mse(1)	0.96	1.15	0.96	1.04	1.02	0.95	0.84
recent best	1.35	1.67	1.08	1.05	1.29	1.15	2.36
PC(BIC)	0.89	0.99	0.88	1.01	0.77	1.07	0.77
PC(AIC)	0.89	0.98	0.91	1.00	0.77	1.04	0.80
shrink(0.25)	1.13	2.13	1.60	2.01	1.04	2.49	0.85
shrink(0.5)	1.01	1.55	1.22	1.32	1.06	1.61	0.86
shrink(1)	0.98	1.32	1.00	0.97	1.06	1.07	0.86
tvp(0.1)	0.95	0.92	0.93	0.85	0.83	0.95	0.88
tvp(0.2)	1.02	0.89	0.90	0.81	0.80	0.98	0.92
tvp(0.4)	1.17	0.95	0.91	0.86	0.85	1.04	1.02

Notes: See notes to Table II.

TVP forecast with the most time variation, and the recent best; a forecaster who uses these methods does worse than she would have done had she simply used the AR.

Comparison of combination and dynamic factor model forecasts

As discussed in the Introduction, an alternative way to forecast using many predictors is to compute forecasts based on a small number of estimated dynamic factors, where the dynamic factors are computed directly from the original (transformed) leading indicators, not (as is done in the PC combination method) from the forecasts $\{\hat{Y}^h_{i,t+h|t}\}$ based on these leading indicators. Success with this approach has been reported by Forni *et al*. (2000, 2001) and by Stock and Watson (1999a, 2002a). This subsection compares such dynamic factor model forecasts to combination forecasts.

Construction of dynamic factor model forecasts

The dynamic factor model-principal components (dfm-PC) forecasts were computed using the leading indicators in the dfm balanced panel subset of the data. Following Stock and Watson (1999a,

Table VII. MSFEs of combination forecasts, relative to autoregression: forecasts of eight-quarter growth of IP ($h = 8$)

Forecast period	Canada 83:I–97:IV	France 83:I–97:IV	Germany 83:I–97:III	Italy 83:I–96:IV	Japan 83:I–97:IV	UK 83:I–97:IV	US 83:I–97:IV
Univariate forecasts							
AR RMSE	0.070	0.050	0.054	0.059	0.111	0.041	0.042
random walk	0.99	1.12	1.18	1.23	1.50	1.00	1.00
Combination forecasts, full panel							
median	0.95	0.85	0.94	0.83	0.88	1.02	0.95
mean	0.89	0.83	0.93	0.77	0.79	1.04	0.87
trimmed mean	0.92	0.82	0.91	0.78	0.82	1.03	0.89
disc. mse(0.9)	1.08	0.91	0.94	0.88	0.84	0.97	0.90
disc. mse(0.95)	1.05	0.95	0.92	0.89	0.84	0.97	0.89
disc. mse(1)	1.04	0.98	0.90	0.90	0.81	0.98	0.89
recent best	1.80	0.89	2.36	1.13	0.93	1.77	2.21
Combination forecasts, balanced panel							
median	1.00	1.02	0.98	0.99	1.02	0.99	0.97
mean	1.00	1.04	0.99	0.98	0.99	1.03	0.89
trimmed mean	1.01	1.05	0.97	1.01	1.01	1.01	0.90
disc. mse(0.9)	1.05	0.95	0.96	0.96	0.96	0.99	0.90
disc. mse(0.95)	1.04	0.99	0.94	0.98	0.97	1.01	0.89
disc. mse(1)	1.04	1.02	0.93	0.99	0.98	1.02	0.88
recent best	1.71	1.31	2.29	1.16	1.01	1.48	2.13
PC(BIC)	0.92	1.01	0.85	0.71	0.56	1.53	1.06
PC(AIC)	0.92	1.08	0.84	0.71	0.56	1.79	0.98
shrink(0.25)	1.30	1.68	1.44	1.98	1.16	2.66	1.16
shrink(0.5)	1.07	1.26	0.98	1.11	0.98	1.18	0.90
shrink(1)	1.01	0.96	0.97	1.02	0.95	0.96	0.89
tvp(0.1)	1.00	0.82	0.90	0.68	0.59	1.13	0.93
tvp(0.2)	1.16	0.95	0.91	0.79	0.56	1.37	0.97
tvp(0.4)	1.39	1.10	0.96	1.00	0.62	1.85	1.04

Notes: See notes to Table II.

2002a), estimates of the dynamic factors were computed recursively as the first four principal components (ordered by the fraction of the variance explained), computed recursively from the recursive sample correlation matrix of the leading indicators labelled 'a' in Table Ib, country by country. The pseudo out-of-sample dynamic factor forecasts of output growth were then computed by regressing h-period output growth against the first m estimated dynamic factors and k lags of ΔY_t, that is, as in (6), except that the principal component predictors $\hat{F}^h_{1,s}, \ldots, \hat{F}^h_{m,s}$ (the estimated factors) were estimated using the original series (not the individual forecasts) and the regression also includes a constant and one or more lags of ΔY_t. The number of factors and the number of lags of ΔY_t both were chosen recursively by AIC (between one and four factors and between zero and four lags of ΔY_t). Also reported are the dfm-PC forecasts computed using a fixed number of factors (3) and lags (2). The details of the dfm-PC forecasting method (and an application to different data) are in Stock and Watson (1999a, 2002a).

Table VIII. Combination forecasts ranked by average losses: both output measures, all horizons (2, 4, 8-quarter growth)

Forecast	Average loss
tvp(0.1)	0.558
mean	0.560
trimmed mean	0.565
disc. mse(1)	0.575
disc. mse(0.9)	0.576
disc. mse(0.95)	0.577
median	0.585
PC(BIC)	0.595
disc. mse(0.9)—bal. panel	0.601
tvp(0.2)	0.603
PC(AIC)	0.603
disc. mse(0.95)—bal. panel	0.604
disc. mse(1)—bal. panel	0.605
mean—bal. panel	0.609
shrink(1)	0.612
trimmed mean—bal. panel	0.612
median—bal. panel	0.616
AR	0.621
tvp(0.4)	0.665
shrink(0.5)	0.709
random walk	0.745
recent best	0.747
recent best—bal. panel	0.769
shrink(0.25)	0.979

Notes: The average losses are weighted averages of the loss of the indicated combination forecast across countries, horizons and output measures, where the weighting is by the inverse of the full-sample standard deviation of the variable being forecasted. The average is over 13 sets of forecasts (six countries for real GDP, seven countries for IP) at three horizons, for a total of 39 cases.

Empirical results

The MSFEs of the dfm-PC forecasts, relative to the AR benchmark, are reported in Table IX. To save space, only the two best-performing combination methods from Table VIII, the mean and TVP(0.1), are reported in Table IX. (The entries for the mean and TVP(0.1) forecasts in Table IX differ from the corresponding entries in Tables II–VII because the results in Table IX were computed using the dfm balanced panel subset.) In some cases, such as forecasting Canadian GDP at the two- and four-quarter horizon, the dynamic factor model forecasts improve upon the AR benchmark by a considerable margin. In other cases, such as IP forecasts for Germany and the USA at the eight-quarter horizon, the dfm-PC forecasts are worse than the AR benchmark. For many countries and horizons the dynamic factor model forecasts have relative MSFEs near one. In seven of the 39 cases in Table IX, at least one of the dfm-PC forecasts outperforms both the mean and TVP(0.1). In most cases, however, the mean or TVP(0.1) forecasts outperform the dfm-PC forecast, sometimes by a wide margin.

Table X presents estimated losses and rankings of all the forecast combination methods plus the dfm-PC forecasts; for comparability, all forecasts in Table X were computed using the dfm balanced

Table IX. MSFEs of selected combination and dynamic factor model forecasts, relative to autoregression, using the dynamic factor model data set

	Canada	France	Germany	Italy	Japan	UK	USA
GDP, h = 2							
mean	0.91	–	0.92	1.03	0.99	0.96	0.90
tvp(0.1)	0.76	–	0.85	0.71	0.79	1.06	0.94
dfm-PC(AIC)	0.76	–	1.01	1.42	1.12	0.91	1.03
dfm-PC(2,3)	0.78	–	1.00	1.40	1.21	0.86	1.05
GDP, h = 4							
mean	0.95	–	1.03	1.14	1.00	0.90	0.77
tvp(0.1)	0.89	–	0.94	0.50	0.62	1.22	0.94
dfm-PC(AIC)	0.80	–	1.21	1.55	0.99	0.87	0.83
dfm-PC(2,3)	0.84	–	1.13	1.48	1.12	0.97	0.90
GDP, h = 8							
mean	0.98	–	1.04	1.05	1.01	1.01	0.90
tvp(0.1)	0.99	–	1.18	0.35	0.68	1.79	1.21
dfm-PC(AIC)	0.88	–	1.24	1.22	0.98	1.12	0.98
dfm-PC(2,3)	0.89	–	1.15	1.19	1.18	1.08	1.03
IP, h = 2							
mean	0.91	1.11	0.92	1.04	0.96	0.95	0.90
tvp(0.1)	0.88	0.94	0.88	0.91	0.86	0.95	0.91
dfm-PC(AIC)	0.85	1.16	0.98	1.22	1.18	1.19	1.00
dfm-PC(2,3)	0.82	1.07	0.92	1.09	1.10	1.02	0.99
IP, h = 4							
mean	0.88	1.23	1.00	1.10	0.94	0.93	0.83
tvp(0.1)	0.90	0.91	0.93	0.82	0.77	0.94	0.87
dfm-PC(AIC)	0.82	1.32	1.20	1.54	1.19	1.20	1.16
dfm-PC(2,3)	0.84	1.24	1.11	1.23	1.16	1.00	1.15
IP, h = 8							
mean	0.91	1.06	0.98	1.02	0.95	0.89	0.90
tvp(0.1)	1.01	0.92	0.92	0.73	0.54	1.17	0.94
dfm-PC(AIC)	1.02	1.22	1.31	1.17	0.95	1.25	1.41
dfm-PC(2,3)	1.03	1.20	1.15	1.16	0.98	1.11	1.35

Notes: All forecasts were computed using the dynamic factor model data set (the series denoted by 'a' in Table Ib). Forecasts were computed over the sample periods indicated in Tables II–VII, as appropriate for the series and horizon being forecasted. The forecasts are defined in the notes to Table II, with the addition of the two dynamic factor model forecasts:

dfm-PC(AIC) m principal components computed for the panel of individual leading indicators; forecasts computed from a dynamic regression including these principal components and p lags of y, where m and p are selected by AIC

dfm-PC(2,3) principal components computed using individual leading indicators; forecasts computed from a dynamic regression including three principal components and two lags of y

panel subset. In the overall sense of Table X, the dfm-PC forecasts perform slightly worse than the AR benchmark. Both the dfm-PC forecasts have substantially higher loss than the PC(AIC), PC(BIC) and TVP(0.1) forecasts, the best performers in Table X.

The dfm-PC forecasts reported in Tables IX–X generally provide substantially smaller improvements upon the AR benchmark than was found for the USA in Stock and Watson (2002a). Addi-

Table X. Dynamic factor model and combination forecasts ranked by average losses: both output measures, all horizons (2, 4, 8-quarter growth)

Forecast	Average loss
PC(AIC)	0.569
PC(BIC)	0.569
tvp(0.1)	0.572
shrink(1)	0.587
disc. mse(0.9)	0.594
disc. mse(0.95)	0.595
disc. mse(1)	0.596
mean	0.602
trimmed mean	0.603
median	0.610
AR	0.621
tvp(0.2)	0.637
dfm-PC(2,3)	0.646
shrink(0.5)	0.657
dfm-PC(AIC)	0.663
tvp(0.4)	0.708
shrink(0.25)	0.720
random walk	0.745
recent best	0.756

Notes: Entries are weighted averages of losses over the 39 cases described in the notes to Table VIII. All forecasts were computed using the dynamic factor model data set.

tional empirical work (not reported in the tables) suggests that one reason for the difference between these results and the more favourable results in Stock and Watson (2002a) is that their forecast sample included the 1970s, whereas the forecast period examined here commences in 1983. In addition, considerably fewer series are used here—for most countries, fewer than 20 series—than in other recent studies using dfm forecasts, and the asymptotic theory behind the dfm-PC forecasts relies on the number of series being large.

Forecast stability

So far the analysis has focused on average performance of the combination forecasts over the full forecast period, 1983–1999. Given the instability of the performance of individual forecasts making up the panel of forecasts, however, it is of interest to examine the stability of the high-dimensional forecasts. Accordingly, we divided the pseudo out-of-sample forecast period (the period in the first rows of Tables II–VII) in half and computed the MSFEs over the two periods of 1982:I–1990:II and 1990:III–1999:IV, where the earlier start date was used to increase the number of observations in the two subsamples. A stable and potent forecast would have population MSFEs less than the AR benchmark in both periods, whereas an unstably performing forecast would have a population relative MSFE less than one in one period but greater than one in the other period. Because of sampling variability, the sample MSFEs will differ from the population MSFEs, but even without a formal distribution theory for these relative MSFEs (for the reasons discussed earlier), examination of the relative MSFEs in the two subsamples can shed some light on the stability of the various forecasting methods.

As a basis for comparison, we first present a scatterplot of the logarithm of the relative MSFEs of the forecasts based on the individual indicators in the first versus the second subsample, for all predictors, horizons, countries and variables being forecasted. In this scatterplot, given in Figure 1, a point represents the pair of log relative MSFEs for a specific predictor, country, horizon and dependent variable (IP or real GDP); Figure 1 contains 2196 such points. If the forecasting relations were stable, then the points would be scattered around the 45° line, with a cluster around the origin for those predictors that have negligible marginal predictive content for output growth, above and beyond that in lags of output growth. But the points are neither scattered around the 45° line nor clustered around the origin; instead, there are many points far into the northwest and southeast quadrants, indicating relatively good performance in one period and relatively poor performance in the other. Indeed, there is little structure in this scatterplot, which suggests that performance of a randomly selected individual predictor/country/horizon/dependent variable combination in the first period is largely independent of its performance in the second period (for further discussion, see Stock and Watson, 2003).

The comparable scatterplot for the simple mean combination forecasts (using the full panel) is presented in Figure 2; each point represents a pair of log relative MSFEs for the simple mean forecast for a particular country, horizon and dependent variable. Evidently, the simple mean combination forecast shows considerable stability—especially when compared with the unstable performance of the constituent forecasts exhibited in Figure 1. Most of the points for the simple mean forecast are in the southwest quadrant, indicating an improvement over the AR benchmark in both periods.

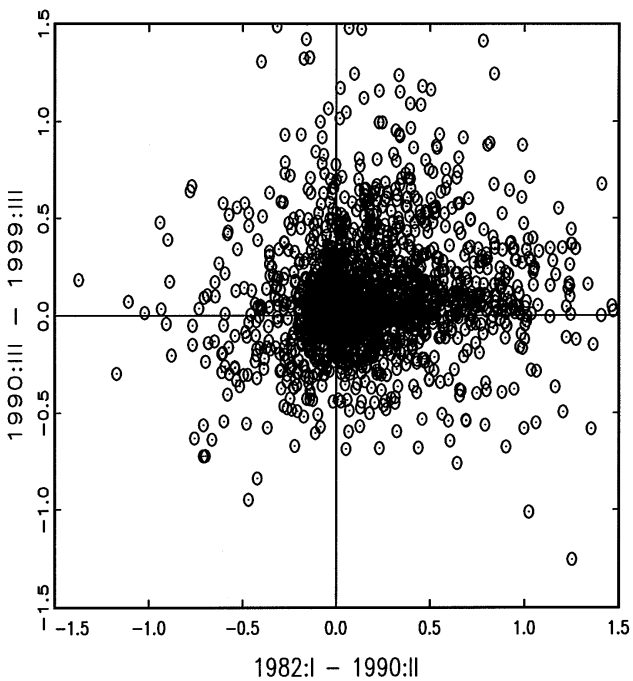

Figure 1. Relative MSFE (logarithm) of pseudo out-of-sample forecasts based on individual predictors

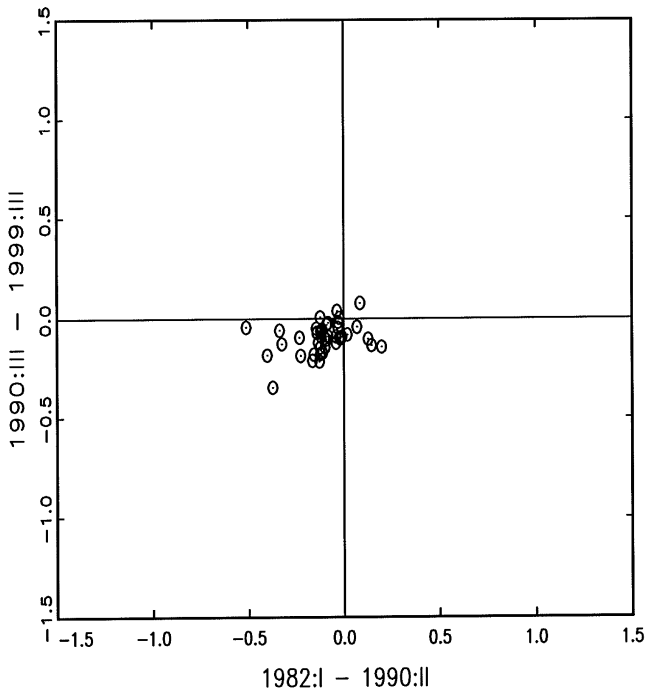

Figure 2. Relative MSFE (logarithm) of pseudo out-of-sample forecasts based on combined (mean) forecast

Additional measures of the stability of the various forecasts are summarized in Table XI, which reports the average relative MSFE for different categories of forecasts (averaged across countries, dependent variables and horizons) for each subperiod, as well as the average absolute difference between the relative MSFEs between the first and second periods.[5] On average, the forecasts based on the individual predictors do worse than the AR benchmark in both periods. Consistent with Figure 1, those forecasts also have, on average, a large absolute change of 0.40 in the relative MSFE between the two periods. In contrast, the simple mean combination forecast improves upon the AR benchmark on average, and has an average absolute change of only 0.08 between the two periods in the full panel (the average change of the median forecast is even less, 0.05). One striking result in Table XI is that the greater the amount of data adaptivity in the combination forecast, the less stably does it perform, with the recent best, PC, shrink(0.25), shrink(0.5) and tvp(0.4) combination forecasts all having large average changes in relative MSFEs between the first and second periods.

[5] The forecast combination results in Table XI are based on 38, 40 or 41 country/variable/horizon cases, depending on the forecasting method; 41 cases are plotted in the scatterplot in Figure 2. There are 42 possible method/country/horizons, and a forecasting method/country/horizon pair was included if it included at least 28 observations in each of the subsamples. Because our Italian GDP data end in 1998, there were insufficient observations for the eight-quarter horizon, which eliminated one pair for all methods. Because French GDP data were unavailable prior to 1970, this eliminated up to three additional method/horizon pairs.

Table XI. Stability of combination forecasts: average relative MSFEs in two subsamples

Forecast	Mean 82:I –90:II	Mean 90:III –99:IV	Mean absolute difference, 1st vs. 2nd period	*n*
Univariate forecasts				
individual forecasts	1.25	1.13	0.40	2196
random walk	1.23	1.30	0.15	41
Combination forecasts, full panel				
median	0.94	0.94	0.05	41
mean	0.91	0.91	0.08	41
trimmed mean	0.91	0.91	0.08	41
disc. mse(0.9)	0.94	0.92	0.09	38
disc. mse(0.95)	0.95	0.92	0.09	38
disc. mse(1)	0.95	0.91	0.11	38
recent best	1.29	1.30	0.32	40
Combination forecasts, balanced panel				
median	0.99	0.99	0.03	38
mean	0.99	0.98	0.06	38
trimmed mean	0.99	0.98	0.05	38
disc. mse(0.9)	0.97	0.97	0.07	38
disc. mse(0.95)	0.98	0.97	0.07	38
disc. mse(1)	0.99	0.97	0.08	38
recent best	1.32	1.56	0.77	38
PC(BIC)	1.26	0.88	0.61	38
PC(AIC)	1.28	0.89	0.64	38
shrink(0.25)	1.89	1.66	1.17	38
shrink(0.5)	1.15	1.29	0.48	38
shrink(1)	0.99	1.03	0.15	38
tvp(0.1)	0.93	0.91	0.19	38
tvp(0.2)	1.01	0.99	0.29	38
tvp(0.4)	1.13	1.08	0.35	38
Dynamic factor model data set				
mean	1.03	0.96	0.14	41
tvp(0.1)	0.97	0.97	0.33	38
dfm-PC(AIC)	1.31	1.04	0.40	41
dfm-PC(2,3)	1.24	1.03	0.35	41

Notes: The entries in the second and third columns are the average of the relative MSFEs for the class of forecasts indicated in the first column, over the 1982:I–1990:II (second column) and the 1990:III–1999:IV subsample (third column). The fourth column contains the average absolute difference between the relative MSFE in the first and second period, by forecasting method, averaged over the forecasting methods indicated in the first column. The final column reports the number of such methods included in the averages in columns 2, 3 and 4. The results in the final block were computed using the dynamic factor model subset of the data.

DISCUSSION AND CONCLUSIONS

The empirical analysis in this paper yields four main conclusions. First, some combination forecasts perform well, regularly having pseudo out-of-sample MSFEs less than the AR benchmark; in some cases, the improvements are quite substantial.

Second, the combination forecasts that perform best generally are those that have the least data adaptivity in their weighting schemes. Aggregated across all horizons, countries and dependent

variables, the forecasting methods with the lowest squared error loss were a time-varying parameter forecast with little time variation, the simple mean combination forecast and the trimmed mean. The best-performing TVP combination forecast has weights that are nearly equal to $1/n$, with a small amount of time variation, and the quantitative gain of this forecast over the simple mean was negligible. In contrast, sophisticated combination forecasts that heavily weight recent performance or allow for substantial time variation in the weights typically performed worse than—sometimes much worse than—the simple combination schemes.

Third, the combination forecasts performed well when compared to forecasts constructed using a dynamic factor model framework. This is interesting in light of recently reported good forecasting results for dynamic factor models. One possible explanation for the relatively poor performance of the dynamic factor model forecasts is that the number of series examined here is relatively small compared with those examined recently using dynamic factor models. In any event, this finding merits further study.

Fourth, the combination forecasts with the least adaptivity were also found to be the most stable when we divided the pseudo out-of-sample forecast period in half. This result is surprising. After all, the reason for introducing discounting and time-varying parameter combining regressions is to allow for instability in the performance of the constituent forecasts—which there clearly is—yet doing so worsens the performance of the resulting combination forecast.

An important caveat to these conclusions is that they are based on point estimates, specifically, the mean squared forecast errors of the combination forecasts relative to autoregressive benchmarks. For reasons discussed earlier, we did not compute measures of statistical precision associated with this forecast error reduction. A logical next step is to develop the asymptotic distribution theory for sample relative MSFEs for models that are sometimes nested and sometimes not and in which the number of parameters can be large, then to implement that theory numerically to provide a framework for formal tests of whether the measured improvements obtained using the combination forecasts are statistically significant. This step, while important, is sizeable and we leave it to future work.

Because of the substantial instability in the performance of the underlying individual forecasts, we consider it implausible that the classical explanation of the virtue of combination forecasts—the pooling of information in a stationary environment—can explain our results. Indeed, the mean of the contemporaneous forecasts has lower average loss than any of the more sophisticated combination forecasts, a finding consistent with other empirical investigations of combination forecasting. This 'forecast combination puzzle'—the repeated finding that simple combination forecasts outperform sophisticated adaptive combination methods in empirical applications—is, we think, more likely to be understood in the context of a model in which there is widespread instability in the performance of individual forecasts, but the instability is sufficiently idiosyncratic that the combination of these individually unstably performing forecasts can itself be stable.

ACKNOWLEDGEMENTS

We thank Fillipo Altissimo, Frank Diebold, Marcellino Massimilliano, two anonymous referees and participants at the Second Workshop on Forecasting Techniques at the European Central Bank for helpful comments and discussions. We especially thank Marcelle Chauvet and Simon Potter for pointing out an error in an earlier draft. An earlier version of this paper was circulated under the title

'Combination Forecasts of Output Growth and the 2001 U.S. Recession'. This research was funded in part by NSF grant SBR-0214131.

REFERENCES

Bates JM, Granger CWJ. 1969. The combination of forecasts. *Operations Research Quarterly* **20**: 451–468.

Bernanke B, Mihov I. 1998. Measuring monetary policy. *Quarterly Journal of Economics* **113**: 869–902.

Blake AP, Kapetanios G. 1999. Forecast combination and leading indicators: combining artificial neural network and autoregressive forecasts. Manuscript, National Institute of Economic and Social Research.

Chan L, Stock JH, Watson MW. 1999. A dynamic factor model framework for forecast combination. *Spanish Economic Review* **1**: 91–121.

Clark TE, McCracken MW. 2001. Tests of equal forecast accuracy and encompassing for nested models. *Journal of Econometrics* **105**: 85–100.

Clemen RT. 1989. Combining forecasts: a review and annotated bibliography. *International Journal of Forecasting* **5**: 559–583.

Clements MP, Hendry DF. 1999. *Forecasting Non-stationary Economic Time Series*. MIT Press: Cambridge, MA.

Cogley T, Sargent TJ. 2001. Evolving post World War II U.S. inflation dynamics. *NBER Macroeconomics Annual* **16**: 331–373.

Cogley T, Sargent TJ. 2002. Drifts and volatilities: monetary policies and outcomes in the post WWII U.S. Manuscript, New York University.

Deutsch M, Granger CWJ, Terasvirta T. 1994. The combination of forecasts using changing weights. *International Journal of Forecasting* **10**: 47–57.

Diebold FX, Lopez JA. 1996. Forecast evaluation and combination. In *Handbook of Statistics*, Maddala GS, Rao CR (eds). North-Holland: Amsterdam.

Diebold FX, Pauly P. 1987. Structural change and the combination of forecasts. *Journal of Forecasting* **6**: 21–40.

Diebold FX, Pauly P. 1990. The use of prior information in forecast combination. *International Journal of Forecasting* **6**: 503–508.

Figlewski S. 1983. Optimal price forecasting using survey data. *Review of Economics and Statistics* **65**: 813–836.

Figlewski S, Urich T. 1983. Optimal aggregation of money supply forecasts: accuracy, profitability and market efficiency. *The Journal of Finance* **28**: 695–710.

Forni M, Hallin M, Lippi M, Reichlin L. 2000. The generalized factor model: identification and estimation. *Review of Economics and Statistics* **82**: 540–554.

Forni M, Hallin M, Lippi M, Reichlin L. 2001. Do financial variables help forecasting inflation and real activity in the EURO area? CEPR Working Paper.

Granger CWJ, Ramanathan R. 1984. Improved methods of combining forecasting. *Journal of Forecasting* **3**: 197–204.

Hendry DF, Clements MP. 2002. Pooling of forecasts. *Econometrics Journal* **5**: 1–26.

Hodrick R, Prescott E. 1981. Post-war U.S. business cycles: an empirical investigation. Working paper, Carnegie-Mellon University; printed in *Journal of Money, Credit and Banking* **29** (1997): 1–16.

LeSage JP, Magura M. 1992. A mixture-model approach to combining forecasts. *Journal of Business and Economic Statistics* **3**: 445–452.

Makridakis S, Hibon M. 2000. The M3-competition: results, conclusions and implications. *International Journal of Forecasting* **16**: 451–476.

Marcellino M. 2002. Instability and non-linearity in the EMU. Manuscript, IEP–U, Bocconi.

Newbold P, Harvey DI. 2002. Forecast combination and encompassing. In *A Companion to Economic Forecasting*, Clements MP, Hendry DF (eds). Blackwell Press: Oxford; 268–283.

Sessions DN, Chatterjee S. 1989. The combining of forecasts using recursive techniques with non-stationary weights. *Journal of Forecasting* **8**: 239–251.

Sims CA, Zha T. 2002. Macroeconomic switching. Manuscript, Princeton University.

Stock JH, Watson MW. 1996. Evidence on structural instability in macroeconomic time series relations. *Journal of Business and Economic Statistics* **14**: 11–29.

Stock JH, Watson MW. 1999a. Forecasting inflation. *Journal of Monetary Economics* **44**: 293–335.
Stock JH, Watson MW. 1999b. A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series. In *Cointegration, Causality and Forecasting: A Festschrift for Clive W.J. Granger*, Engle R, White H (eds). Oxford University Press: Oxford; 1–44.
Stock JH, Watson MW. 2002a. Macroeconomic forecasting using diffusion indexes. *Journal of Business and Economic Statistics* **20**: 147–162.
Stock JH, Watson MW. 2002b. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association* **97**: 1167–1179.
Stock JH, Watson MW. 2003. Forecasting output and inflation: the role of asset prices. *Journal of Economic Perspectives* **41**: 788–829.
West KD. 1996. Asymptotic inference about predictive ability. *Econometrica* **64**: 1067–1084.

Authors' biographies:
James H. Stock is Professor of Economics at Harvard University, a Research Associate at the National Bureau of Economic Research, and a Fellow of the Econometric Society.

Mark W. Watson is Professor of Economics and Public Affairs at the Department of Economics and the Woodrow Wilson School at Princeton University, a Research Associate at the National Bureau of Economic Research, and a Fellow of the Econometric Society.

Authors' addresses:
James H. Stock, Department of Economics, Littauer Center, Harvard University, Cambridge, MA 02138-3001, USA.

Mark W. Watson, Department of Economics, Princeton University, Princeton, NJ, USA.

