EUROPEAN CENTRAL BANK
EUROSYSTEM

WORKING PAPER SERIES
NO 1277 / DECEMBER 2010

COMBINING THE FORECASTS IN THE ECB SURVEY OF PROFESSIONAL FORECASTERS

CAN ANYTHING BEAT THE SIMPLE AVERAGE?

by Véronique Genre, Geoff Kenny, Aidan Meyler and Allan Timmermann

EUROPEAN CENTRAL BANK

EUROSYSTEM

WORKING PAPER SERIES

NO 1277 / DECEMBER 2010

COMBINING THE FORECASTS IN THE ECB SURVEY OF PROFESSIONAL FORECASTERS

CAN ANYTHING BEAT THE SIMPLE AVERAGE?[1]

by Véronique Genre[2], Geoff Kenny[2,3], Aidan Meyler[2] and Allan Timmermann[4]

In 2010 all ECB publications feature a motif taken from the €500 banknote.

This paper can be downloaded without charge from http://www.ecb.europa.eu or from the Social Science Research Network electronic library at http://ssrn.com/abstract_id=1719622.

1 The authors would like to thank Tommy Kostka for providing excellent research support with the SPF data. Helpful comments and suggestions from participants in seminars at the ECB and the Central Bank of Ireland are also gratefully acknowledged. Any errors are the sole responsibility of the authors.

2 European Central Bank, Kaiserstrasse 29, D-60311 Frankfurt, Germany.

3 Corresponding author: DG Research, European Central Bank, Kaiserstrasse 29, D- 60311 Frankfurt, Germany; e-mail: Geoff.kenny@ecb.europa.eu

4 Rady School of Management and Department of Economics, University of California, San Diego, USA.

Address
Kaiserstrasse 29
60311 Frankfurt am Main, Germany

Postal address
Postfach 16 03 19
60066 Frankfurt am Main, Germany

Telephone
+49 69 1344 0

Internet
http://www.ecb.europa.eu

Fax
+49 69 1344 6000

Information on all of the papers published in the ECB Working Paper Series can be found on the ECB's website, http://www.ecb.europa.eu/pub/scientific/wps/date/html/index.en.html

ISSN 1725-2806 (online)

CONTENTS

Abstract
In this paper, we explore the potential gains from alternative combinations of the surveyed forecasts in the ECB Survey of Professional Forecasters. Our analysis encompasses a variety of methods including statistical combinations based on principal components analysis and trimmed means, performance-based weighting, least squares estimates of optimal weights as well as Bayesian shrinkage. We provide a pseudo real–time out-of-sample performance evaluation of these alternative combinations and check the sensitivity of the results to possible data-snooping bias. The latter robustness check is also informed using a novel real time meta selection procedure which is not subject to the data-snooping critique. For GDP growth and the unemployment rate, only few of the forecast combination schemes are able to outperform the simple equal-weighted average forecast. Conversely, for the inflation rate there is stronger evidence that more refined combinations can lead to improvement over this benchmark. In particular, for this variable, the relative improvement appears significant even controlling for data snooping bias.

Keywords: forecast combination, forecast evaluation, data snooping, real-time data, Survey of Professional Forecasters
JEL: C22, C53

Non-technical summary

In this paper, given the available sample of approximately 10 years of data that is now available from the ECB SPF, we explore alternative combinations of the SPF forecasts with a view to optimising the quality of information that is made available to decision makers and the public. The central focus of forecast combination is to reduce the information in a vector of forecasts to a single summary or combined forecast using an estimated set of combination weights. The optimal combination chooses the weights such that the expected loss of the combined forecast error is minimised. Such optimal weights will tend to be larger for more accurate forecasts particularly when such forecasts are less strongly correlated with other forecasts. Moreover, in a manner that is similar to the classical diversification gains in financial portfolio theory, the resulting combined forecast also offers potential improvements in forecasting performance by "averaging out" some of the error in individual forecasts.

Our analysis employs a wide array of forecast combination methods including trimming (i.e. exclusion of "extreme" forecasts), weighting based on historical performance, optimal weighting as well as Bayesian shrinkage of least squares weights toward equal weights. In estimating the optimal combination weights by linear projection, we propose using statistical techniques designed to reduce the high cross-sectional dimension of the SPF dataset by constructing sub-groups of forecasters (e.g. high, average and low performing) and estimate the combination weights conditional on this classification. We then compare the out-of-sample forecast performance of these alternative SPF combination strategies with the current practice of focussing on the equal weighted forecast. We also examine the robustness of our results to the vintage of the data used as the target variable and check the extent to which the performance of different forecast combinations may have changed during the period of exceptional macroeconomic volatility associated with the 2008-2009 financial crisis.

Over the sample period analysed, we demonstrate that the equal weighted combination sets a reasonably high benchmark in the sense that it is shown to be quite informative when measured against simple time series or other naïve forecasts. Notwithstanding the relatively good performance of this SPF benchmark, a number of

alternative combination strategies are shown to achieve quantitatively important and statistically significant gains relative to this benchmark in an out-of-sample "horse race" conducted over the five year period from 2004:q1 to 2008:q3. Looking across variables, the scope for improvements from alternative combinations is most significant for inflation with quantitatively smaller (and generally less significant) gains achievable for GDP and, especially, for the unemployment rate.

The above results refer to "normal times" as the sharp volatility and large forecast errors associated with the 2008-2009 financial crisis are excluded from the evaluation sample. Nonetheless, when the sample is extended to include the most recent period of large macroeconomic volatility, a number of the alternative combination strategies (least squares, Bayesian shrinkage and the recent best forecaster) continue to perform better than the equal weighted benchmark. During the crisis, it is noteworthy that a strategy of picking the recent best forecasters performed better than the benchmark for all three variables at the 1-year ahead horizon. Such a result points to the possible gains that may arise from placing all the weight on the forecaster adapting his/her outlook to the crisis environment and the loss in performance which may arise if positive weights continue to be assigned to forecasters who have not adapted.

Overall, we would conclude from this study that there exists a reasonable case to consider alternative combinations as means of more optimally summarising the information collected as part of the quarterly rounds of the ECB SPF. An important caveat applying to our analysis, however is the likely relevance of data snooping bias given the small sample size available for evaluation and the large specification search conducted across alternative combination strategies. Only for the inflation forecasts, is a "reality check" (which attempts to control for such bias) indicating a robust improvement of the best performing model relative to the benchmark at the 10% significance level. For the other variables and horizons, the reality check is more in line with the often reported result of being unable to outperform an equal weighted combination in practice. This tends to caution against any assumption that the identified improvements relative to the equal weighted benchmarks would necessarily persist in the future and argues for the reporting of a suite of alternative combinations rather than focussing on a particular single "best" combination method.

1. Introduction

Coinciding with the launch of the single currency in January 1999, the European Central Bank (ECB) started a Survey of Professional Forecasters as part of the information gathering and analysis of the macroeconomic outlook that was to be used as input for monetary policy decisions. Since then, the results of the ECB SPF have been communicated to policy makers on a quarterly basis and also have been regularly published in the ECB Monthly Bulletin and on its website (Garcia, 2003 and Bowles et al., 2007). Throughout this period, the forecast data collected in the SPF has normally been summarised by way of a simple average of the surveyed forecasts. Although a large literature exists on how to optimally combine forecasts (see Timmermann (2006), Newbold and Harvey (2002) and Clemen (1989)), such an approach was reasonable given the lack of any available track record among SPF panel members in forecasting euro area aggregates. Moreover, as discussed further below, various empirical studies have shown that such a simple equally weighted pooling of forecasts performs relatively well in practice compared with alternative approaches that rely on estimated combination weights and thus can be sensitive to parameter estimation error.

In this paper, given the available sample of approximately 10 years of data against which we can evaluate SPF forecasts, we explore alternative combinations of the SPF forecasts with a view to optimising the quality of SPF information that is made available to decision makers and the public. As discussed in Timmermann (2006), the central focus of forecast combination is to reduce the information in a vector of forecasts to a single summary or combined forecast using an estimated set of combination weights. The optimal combination chooses the weights such that the expected loss of the combined forecast errors is minimised. Such optimal weights will depend on the first two moments of the joint distribution of the vector of forecasts and the actual outcome and have an intuitive interpretation: They will tend to be larger for more accurate forecasts, particularly when such forecasts are less strongly correlated with other forecasts. In a manner that is similar to the classical diversification gains in financial portfolio theory, the resulting combined forecast offers potential improvements in forecasting performance by "averaging out" the idiosyncratic components in individual forecasts. Such idiosyncratic components in individual

forecasts may reflect misspecification in an individual forecasters' models, measurement error or divergence in the degree of adaptability of individual forecasts to new information, including information concerning possible structural breaks associated with technological innovation or institutional change impacting on the macro economy (see Diebold and Pauly (1987), Hendry and Clements (2002) and Aiolfi, Capistran and Timmermann (2010)). More generally, the potential gains from more optimal combination strategies may be particularly large during times of exceptional change or volatility in macroeconomic conditions. For example, combination methods which allow for time-variation in combination weights may be particularly suited to periods such as the "great recession" associated with the 2008 – 2009 financial crisis.

Our analysis encompasses a variety of methods that have been proposed in the literature including statistical combinations based on principal components analysis and trimmed means, performance-based weighting (Bates and Granger (1969)), optimal weighting as well as Bayesian shrinkage, advocated in Clemen and Winkler (1986), Diebold and Pauly (1990) and Stock and Watson (2004). In estimating the optimal combination weights by linear projection, we also follow the conditional combination strategy in Aiolfi and Timmerman (2006) and employ statistical techniques designed to reduce the cross-sectional dimensionality of the dataset by constructing sub-groups of forecasters and estimate optimal weights conditional on this sub-grouping. As a result, our mode of analysis aims at handling the relatively large cross sectional dimension of the SPF dataset with a view to maintaining a relatively parsimonious representation so as to minimise estimation error in the combination weights. In comparing the out-of-sample forecast performance of the above alternative combination strategies, we test the robustness of our findings along a number of dimensions that are judged to be important in practical applications: Most notably, given that we test a large number of combination methods using a single historical dataset, we employ the White (2000) reality check for data snooping as well as a novel "meta" selection procedure aimed at shedding light on the robustness of any identified gains from alternative combination methods. Also, given the significant revisions to euro area macroeconomic variables over our sample period, we examine the sensitivity of our results to the chosen vintage of the outcome for the forecast target variable against which forecast performance is assessed.

Finally, the sub-sample stability of the performance of alternative combinations is considered, in particular the extent to which the relative performance of different methods may have changed during the period of exceptional macroeconomic volatility associated with the 2008-2009 financial crisis.

Our main findings can be summarised as follows: Over the sample period analysed, the equal weighted combination sets a reasonably high benchmark in the sense that it is shown to be quite informative when measured against simple time series or other Naïve forecasts. Notwithstanding the relatively good performance of the SPF benchmarks, a number of alternative combination strategies are shown to achieve quantitatively important gains relative to this benchmark. Looking across variables, the scope for improvements from alternative combination strategies appears the most significant for inflation with smaller and less significant gains achievable for GDP and, especially, for the unemployment rate. In general, our results do not identify any single combination approach which appears to dominate across either variables or at different horizons. Instead, depending on the horizon and the variable, the best performing combination methods include least squares, Bayesian shrinkage as well as more simple strategies where the weighting is determined only by relative past performance.

The remainder of the paper is organised as follows. In Section 2 we review the main classes of combination methods we employ and explain how in practice we have applied them to the ECB SPF. Section 3 provides some information on the SPF dataset, focussing on the cross-sectional information that is available and some practical issues (such as the entry and exit of forecasters from the panel) that need to be overcome when implementing several of the forecast combination methods we apply. Section 4 presents our three main performance evaluation measures while in Section 5 we present the main out-of-sample forecast evaluation results for each of the three main forecast variables (inflation, GDP growth and the unemployment rate) over two different horizons (1- and 2-years ahead). This section also examines the overall robustness of our findings i) with respect to the data vintage used in the forecast evaluation, ii) in terms of overall sub-sample stability and iii) accounting for possible data snooping bias. Finally Section 6 concludes with a summary of our main findings.

2. Forecast Combination Methods

In this section, we review the main categories of approaches we apply for the estimation of combination weights and the alternative benchmarks against which they are evaluated. Let $\hat{y}_{i,t+h}$ be the *i*'th survey participant's forecast of the outcome in period *t*+*h*, based on the forecaster's information at time *t*. The main aim of forecast combination is to reduce the information in a vector of *N* forecasts ($\hat{y}_{i,t+h}$, *i* = 1, . . . , N) to a single summary or combined forecast $\hat{y}^{c}_{t+h}(\hat{y}_{i,t+h}, w)$ where *w* represents the N-dimensional vector of combination weights, $w_{i,t+h}$, i = 1, …, N. The optimal combination chooses *w* such that the conditional expected loss of the combined forecast errors is minimised, i.e. the optimal combination weights (w^{*}_{t+h}) solve the problem denoted by (2.1) below

$$w^{*}_{t+h} = \arg\min_{w_{i,t+h}} \; E[L(e^{c}_{t+h}, w_{i,t+h}) \,|\, \hat{y}_{i,t+h}] \qquad \forall\, i = 1, 2, \ldots, N \tag{2.1}$$

where e^{c}_{t+h} denotes the error of the combined forecast and $L(\,.\,)$ the representative decision maker's loss function. Assuming that the forecast is linear in the combination weights and that loss is of the MSE type, i.e. $L(e^{c}_{t+h}) = (e^{c}_{t+h})^{2}$, combination weights will depend on the first two moments of the joint distribution of the vector of forecasts and the actual outcome and can be estimated by linear projection of the individual forecasts on the target variable.[1] As discussed in Timmermann (2006), it can be shown that the equally weighted combination - which is currently the headline SPF indicator used by the ECB - is optimal only under strongly restrictive assumptions that one would not necessarily expect to hold *ex ante*, i.e. under the assumption that the forecasts all have the same variances and pair wise cross correlations. Such assumptions would be likely to hold only in the unlikely situation where forecasters all share a single common information set and the same model of the economy on the basis of which they report and update their forecasts. To the extent that forecasters hold differing views on the structure of the economy or adapt their views at different speeds in response to economic news, or have different

[1] The problem of finding the optimal combination weights is directly analogous to the well-known portfolio optimisation problem posed in finance.

information sets on which they condition their forecasts, there are potentially significant opportunities to better exploit the information content of the SPF through alternative more optimal combinations of the individual replies. Given the SPF's role as input included in the regular information set underpinning monetary policy decision-making in the euro area, together with its function of providing publicly available information on the macroeconomic outlook, such opportunities to improve forecast performance warrant a careful empirical investigation.

In practice, however, there may be some important limits to the gains from attempting to combine forecasts optimally. Optimal combination weights are obtained by linear projection of the target variable on each of the forecasts (Granger and Ramanathan (1984)) and will therefore be subject to sizeable estimation error particularly in situations where the number of individual forecasts is large relative to the number of time series observations, as is the case for the ECB SPF which was launched only in 1999 with a relatively large number of approximately 90 participants from across the EU. Such estimation error reflects the dependence of the optimal weights on the full conditional covariance matrix of forecasts which – when the number of forecasts is high – entails a large number of unknown parameters. Such estimation error is a commonly cited explanation for why more simple combination schemes – such as the equally weighted combination – have been shown to perform well in practices. For example, Stock and Watson (2004), Makridakis et al (1982), Makridakis and Winkler (1983), and Smith and Wallis (2009) are four notable studies highlighting the empirical success of the equal weighted combination. At the same time, some studies have suggested greater empirical success with more theoretically motivated combinations. For example, using a conditional combination strategy, Aiolfi and Timmermann (2006) report some out-of-sample improvement compared with simple equal weighted combinations or using the previous best model. More recently, Capistrán and Timmermann (2009) cite evidence in support of an alternative affine transformation of the equal weighted forecast as performing reasonably well in small samples.

Essentially there is a trade-off between the squared bias of the forecast – which generally is reduced by using more complex and flexible models – and the variance of the forecast error which comprises the effect of parameter estimation error and so

tends to be lower for simple combinations such as equal-weighting. Good forecasting methods exploit this trade-off in an optimal manner. The extent of this trade-off will depend on factors such as the cross-sectional and time-series dimensions of the data along with the (unknown) parameters of the data generating process and the joint distribution of the forecasts. These differ across variables, data sets, sample periods and forecast surveys and so it is difficult to come up with a universally appropriate strategy that uniformly performs well. In the remainder of this section, we discuss these and other simpler approaches to combining forecasts that are commonly adopted in practical applications. We restrict ourselves to the class of linear combinations and focus on those methods which emphasise parsimony with a view to minimising as much as possible estimation error.

2.1 Trimming and other statistical combinations

A first class of methods draws on statistical techniques in order to summarise the information in the distribution of individual forecasts. We consider the median as well as other trimmed mean measures which remove extreme values from the cross section of individual forecasts. Such combinations which assign zero weight to some forecasts and equal weights to all others can be motivated by the possibility of forecasts that are completely non-informative. To the extent that such forecasts represent "noise" their removal will improve the overall forecast accuracy of any combined forecast, including the equally weighted combination. Within the class of statistical approaches, Stock and Watson (2004) have suggested to use principal components analysis to estimate the static common factors from the panel of forecasts in order to derive the combined forecast. In order to implement this approach the first few principal components are computed recursively and regressed (again recursively) on the outcome of the target variable. Denoting $\hat{F}_{1,t}, \ldots, \hat{F}_{p,t}$ as the first p principal components of the panel of forecasters, the combined forecast is computed using the weights estimated using the OLS regression (2.2) below.

$$y^{c}_{t+h} = w_1 \hat{F}_{1,t} + \ldots + w_p \hat{F}_{p,t} + \varepsilon_{t+h} \tag{2.2}$$

In the practical application of this method, we consider the performance of the forecast combination for up to three principal components (p = 1, 2 and 3) and estimate all the combination weights recursively and allowing for the lags reflecting the publication delay of the outcome variable in order to preserve the "real time" character of the resulting combination.

2.2 Performance-based weighting schemes

Another class of combination methods is based on the intuitive idea of assigning higher weights to forecasts with a relatively good forecasting track record and lower weights to forecasts with a poor performance. The idea for such performance-based weighting was introduced by Bates and Granger (1969). Such combinations have been shown to perform reasonably well in practice (Newbold and Granger (1974)), a finding which is often interpreted in terms of robustness given high estimation uncertainty that plagues other approaches attempting to exploit forecast co-variances. Stock and Watson (2004) propose the following general representation for such a scheme which allows for an arbitrary discount function that can be applied to historical forecast errors (capturing the idea that past forecast performance has a smaller impact on current combination weights), i.e.

$$w_{it} = \frac{m_{it}^{-1}}{\sum_{j=1}^{n} m_{jt}^{-1}} \qquad m_{it} = \sum_{s=T_0}^{t-h} \delta^{t-h-s} (y_{s+h} - \hat{y}_{i,s+h})^2 \tag{2.3}$$

where δ is the discount factor and m_{it} represents the cumulative sum of past (discounted) forecast errors computed since the start of the sample (T_0). The case of δ = 1 (no discounting) corresponds to an optimal weighting scheme when the individual forecasts are uncorrelated (Bates and Granger (1969)). Hence this method essentially ignores any correlation in the errors of the individual forecasts. Setting values for δ below unity allows for higher (lower) weights to be assigned to more recent (distant) forecast errors in the calculation of the combination weights. In empirical applications of this approach, the weights can be computed either recursively using all available observations from the start of the sample (T_0) or over a

rolling window of a given length (v) to take account of possible time variation in relative forecast performance. The latter assigns zero weight to any past forecast errors occurring in periods prior to this rolling window. The shorter is v the more weight is put on the model's recent track record and the larger the part of historical performance that is discarded. Assuming no discounting of performance within the rolling window, the relevant combination can are derived according to (2.4) below.

$$w_{it} = \frac{m_{it}^{-1}}{\sum_{j=1}^{n} m_{jt}^{-1}} \qquad m_{it} = \sum_{s=t-h-v+1}^{t-h} (y_{s+h} - \hat{y}_{i,s+h})^2 \tag{2.4}$$

In the empirical implementation of (2.4), we assess the performance for v = 1, 4 and 8 quarters. Another commonly used and simple performance-based combination method is the Recent Best (RB) forecast As implemented in this study this assigns all weight to the individual forecasts with the lowest most recently observed squared forecast errors or the lowest average mean squared error over a rolling window of length v. In the absence of any strong prior information about the length of window necessary in order to identify the best forecaster, the empirical analysis considers window lengths of v = 1 and v = 4 quarters.

2.3 Least squares (optimal) combination weights

Under mean-squared loss, the optimal combination weights have a straightforward interpretation as the coefficients in a multiple regression of the observed outcome on the individual forecasts (Granger and Ramanathan (1984)). We consider the following four basic combination regressions:

$$y_{t+h}^{c} = w_{0,h} + \sum_{i=1}^{N} w_{i,h} \hat{y}_{i,t+h} + \varepsilon_{t+h} \tag{2.5}$$

$$y_{t+h}^{c} = \sum_{i=1}^{N} w_{i,h} \hat{y}_{i,t+h} + \varepsilon_{t+h} \tag{2.6}$$

$$y^{c}_{t+h} = \sum_{i=1}^{N} w_{i,h}\hat{y}_{i,t+h} + \varepsilon_{t+h} \quad \text{where} \quad \sum_{i=1}^{N} w_{i,h} = 1 \tag{2.7}$$

$$y^{c}_{t+h} = \sum_{i=1}^{N} w_{i,h}\hat{y}_{i,t+h} + \varepsilon_{t+h} \quad \text{where} \quad 0 \leq w_{i,h} \leq 1.0 \quad \forall\, i = 1,\ldots,N \tag{2.8}$$

Equation (2.5) is the most general regression allowing for an intercept term and estimating the weights on individual forecasts using unconstrained OLS. Hence it allows for a possible bias adjustment in the combined forecast which may adjust for any bias in the individual forecasts.[2] In contrast, equation (2.6) omits a constant in the combination regression and therefore does not include any correction for possible bias.[3] Equation (2.7) imposes the adding up constraint that the estimated weights sum to unity. The constraint ensures that the combined forecast will be unbiased if the individual forecasts are also unbiased. Lastly, equation (2.8) follows Granger and Newbold (1986) by ruling out negative weights and weights greater than unity in order to ensure that the combined forecast always lies within the range of the individual forecasts.[4] All of the above four regressions (2.5) – (2.8) are estimated recursively and the recursive weights are used to derive the combined forecasts in pseudo real time.[5] In practice, a key problem which arises in applying the above regression approaches to the SPF forecasts is the relatively large cross sectional dimension of available forecasts to be combined together with the relatively small time series dimension that can be used to estimate the combination regressions. Aiolfi and Timmermann (2006) have suggested the use of clustering techniques as a simple way of overcoming this problem. They apply the k-mean clustering algorithm to the panel of individual forecasts in order to identify the group structure of the dataset. The combination weights are then estimated by replacing the *N* individual forecasts in

[2] As a result the combined forecast estimated using equation 2.5 may be unbiased *even if* the individual forecasts are biased.
[3] As highlighted subsequently in Section 3 of this paper, the case for some bias adjustment exists for a number of variables in the ECB SPF.
[4] The convexity constraints are implemented using non-linear least squares. In principal, such a non-negativity constraint may be sub-optimal. However, in practice, such constraints may help improve forecast performance by helping to limit the impact of parameter estimation error on the combined forecast.
[5] Hence, although the notation in (2.4) to (2.8) suppresses any time subscript on the weights but rather emphasises their horizon dependence, the least squares combination weights will also vary over time reflecting the recursive estimation of the regression coefficients.

equations (2.5) to (2.8) with the mean of each cluster. In the empirical application we restrict ourselves to a maximum of three clusters given the restricted time series sample that is available for the estimation.

Capistrán and Timmermann (2009) propose an alternative least-squares combination approach which may be useful in practical situations where the dimension of the vector of forecasts is very large. Their combination is based on a simple linear projection of the target variable on the equally weighted forecast $\bar{y}_{t+h}$, i.e.

$$y^{c}_{t+h} = w_{0,h} + \overline{w}\ \bar{y}_{t+h} + \varepsilon_{t+h} \tag{2.9}$$

where $\overline{w}$ is the estimated slope parameter in the combination regression. This simple linear projection has the advantage of being relatively parsimonious, thus helping to limit the impact of parameter estimation error. Moreover, it provides a simple transformation of the equal weighted forecast and has been shown empirically to perform well in finite samples. It can also be implemented easily in cross-sectional panels like the SPF with frequent missing observations due to the entry, exit and possible re-entry of forecasters from the panel of respondents. As with the other approaches above, (2.9) can be estimated either with or without the bias adjustment parameter ($w_{0,h}$).

2.4 Shrinkage (Bayesian) Combinations

A fourth and related class of combination schemes is to calculate the combination weights as a weighted average of the weights from the least squares estimates and the weights of the equally weighted forecast combination. The combination weights are as a result shrunk toward an equally weighted prior, thus giving the resulting combination a Bayesian interpretation (Diebold and Pauly (1990)). As implemented in Stock and Watson (2004), our shrinkage weights take the form

$$w_{i,h} = \psi\ \hat{w}_{i,h} + (1-\psi)\ (1/N) \tag{2.10}$$

where $\psi = \max(0, 1 - \kappa N/(T-h-N-1))$ and the parameter κ governs the amount of shrinkage and T denotes the number of observations used in the least squares regressions. Hence, when the sample size is high relative to the number of forecasts used in the combination, the least squares weights will be given a higher weight. As with other constrained estimates of combination weights the intuition for the above approach is to limit the impact of the data on the estimated weighting scheme as a way to possibly reduce the associated parameter estimation error. In the practical application below, in line with the strategy adopted in Aiolfi and Timmermann (2006), we estimate the shrinkage combination using the same estimated clusters employed in the least squares combination and for alternative values of the shrinkage parameter. We vary the intensity of the shrinkage parameter, allow for the possibility of either 2 or 3 clusters, either with or without a bias adjustment, allowing for a total of eight possible shrinkage combinations.[6]

3. The SPF dataset

In this section we provide a brief overview of the ECB SPF dataset focusing on its cross-sectional dimension. We also highlight the extent of the entry and exit of forecasters in the ECB panel and, given that a number of combination methods require a panel without missing observations, we present a simple approach to create a fully balanced panel.

3.1 The ECB SPF: some key features

The SPF forecasts are described in some detail in previous studies such as Bowles et al. (2010) and Garcia (2003). A key aspect that warrants clarification is the definition and transformations of the variables being forecasted. For both the 1 and 2- year horizons, these refer to the annual change in the (level of) GDP and the (level of) the HICP in quarter *t+h* compared with quarter *t+h-4* and the *level* of the unemployment expressed as a percentage of the euro area labour force in quarter *t+h*.[7] To get a sense

[6] The prior mean of the bias adjustment parameter is set equal to zero. The choice of shrinkage parameter allows the weight on the prior mean to vary between 25% and 75% depending on the number of observations and clusters used.

[7] Recent literature has emphasised the possible impact of such transformations on forecast performance of model-based forecasts. In the present context, the evaluation is very much constrained by the

of the behaviour of these SPF forecasts over the sample period, Figure 1 plots the mean SPF forecasts for the three variables (GDP, inflation and the unemployment rate) and two horizons (1 and 2-year ahead forecasts) analysed in this study. Both the current (2009:q3) and the 1st vintages are shown for each of the outcomes together with the forecast errors for the equal weighted forecast calculated using the 1st vintage. The figure highlights the relatively sizeable and often persistent forecast errors from the SPF; the errors are particularly sizeable for the quarters starting in 2008:q3 reflecting the impact of the 2008-2009 financial crisis. The inflation and GDP forecast errors show a clearly one-side pattern while those for the unemployment rate are more two-sided.[8] This graphical presentation also highlights some difference between the longer 2-year horizon forecasts and the 1-year forecasts, in the sense that the former have tended to be much smoother and, hence, less correlated with the actual outcome.

To provide some graphical information on the heterogeneity embedded in the SPF panel, Figure 2 plots the histograms of the mean errors for each variable and horizon. In order to avoid any sampling distortions associated with the entry and exit of forecasters, the plots are based on a preliminary filtering of the data so as to include only those forecasters who have been contributing relatively frequently.[9] Without this pre-filtering, some forecasters would perform poorly (or reasonably well) in relative terms simply because they contributed to the survey during a period when the target variable exhibited above (below) average volatility. For example, in the case of GDP, the cohort of forecasters who entered the panel only in 2007 and 2008 performed particularly poorly in relative terms reflecting the exceptionally high volatility in the macroeconomic environment around this time. Table 1 provides further evidence on this, reporting the mean errors and RMSE over different sub samples for each variable and horizon. For example, the average RMSE on the 1-year ahead SPF forecasts for GDP rose from 1.1 when calculated over the period 1999:q1-

definitions of the variables used in the SPF questionnaire. A major virtue of the latter - from the perspective of empirical analysis - has been its relative stability in terms of structure and definitions of forecast variables. More generally, however, an important question for future research is the possible impact of survey design on actual forecast performance of the SPF.

[8] For a more comprehensive analysis of the properties of the SPF forecasts for growth and the unemployment rate see Bowles et al (2010).

[9] The filter is such that forecasters with more than four consecutive missing observations are excluded from the panel. The unbalanced nature of the SPF panel is discussed in more detail in Section 3.2 below.

2003:q4 to 1.8 when calculated over the period 2004:q1-2009:q3. This latter period includes the influential observations linked to the exceptional macroeconomic volatility associated with the 2008-2009 financial crisis. A similar, though less marked, deterioration in forecast performance is evident for the unemployment rate and inflation forecasts reflecting the impact of the financial crisis (see also Table 1).

A clear feature of the data evident from Figure 1 and Table 1 is the presence of possible bias in SPF forecasts. In the case of GDP and the unemployment rate, the bias has tended to be positive (i.e. as defined here the forecasted level for both variables has tended to be above the actual outcome). In Figure 2, μ denotes the average mean error across all forecasters, i.e. of the equal weighted forecast combination, and N the number of forecasters that "survive" in the filtered panel. In the case of the 1- and 2-year ahead unemployment rate forecast the average bias has nonetheless tended to be quite small (i.e. less than 0.2 percentage points), while in the case of GDP 1 year ahead it is larger (close to 0.6 percentage points). In the case of inflation, there is also evidence of possible negative bias (i.e. the forecasted level for inflation has tended to be below the actual outcome). Overall, this graphical analysis suggests that for some variables and some horizons, combination methods which allow for bias adjustment could yield superior out-of-sample performance over the evaluation period relative to combination methods that do not directly adjust for bias.[10]

Figure 3 shows the equivalent histograms for the RMSEs of each forecaster (again using the filtered dataset). From the plots, it is also clear that the SPF forecasts exhibit standard features one would expect to observe. In particular, with the exception of the inflation forecasts (see also Table 1), forecast performance as measured by the RMSE deteriorates with an increase in the length of the forecast horizon (see Patton and Timmermann (2010)). As with the mean error plot in Figure 2, the plots also highlight significant heterogeneity in forecasting performance across individuals and it is precisely this heterogeneity in the data that alternative combination methods seek to

[10] The findings in Croushore (2009) for the US SPF suggest less evidence of bias when surveyed forecasts are evaluated over significantly longer runs of data than the 10 years that is available for the ECB SPF. Most combinations we apply incorporate some form of bias adjustment via the constant in the combination regressions. A relatively large bias in individual forecasts will also tend to result in a relatively low weight when using the performance based weighting scheme discussed in Section 2.2. However, performance based combination does not include a bias adjustment at the aggregate level.

exploit. Another interesting feature that can be examined from these cross plots is the possible existence of a group structure in the data set. However, Figure 3 does not suggest a clear clustering of panel members into groups of low or high forecast accuracy. Nonetheless, in the following section we evaluate the performance of least squares combination schemes where the forecast are replaced by the simple mean of alternative clusters identified by applying the *K*-mean algorithm to the panel of squared forecast errors in the fist half of the sample (1999:q3- 2003:q4).

As a final graphical insight into the dataset, Figure 4 plots forecaster performance over the first part of the sample (X-axis) against the performance in the second part of the sample (Y-axis) for each variable and each horizon. In line with the data reported in Table 1, a clear feature evident in Figure 4 is the sharp deterioration in average forecast performance in the second half of the sample compared with the first half. High persistence in forecast performance would suggest a positive correlation between past and subsequent forecast performance.[11] Such a positive relation is indeed evident for some variables and horizons but it tends not to be statistically significant. However, in the case of unemployment, the graphical evidence is more suggestive of anti-persistence or "crossings" (see Aiolfi and Timmermann (2006)) whereby relatively good past performance in the first part of the sample is associated with a relatively poor forecast performance in the second half of the sample (and vice versa). Finally Figure 4 also highlights that the overall dispersion in forecaster accuracy is not time invariant. This is particularly evident for GDP forecasts (both horizons) where a much greater level of dispersion is evident in the second sub-sample, again most likely linked to the larger forecast errors made at the time of 2008-2009 financial crisis. From the perspective of combining such forecasts, both observations above (i.e. time varying dispersion in forecast accuracy and anti-persistence) would tend to highlight the possible gains from combination methods which allow for sufficient time variation in the combination weights. More generally, such instability in each forecaster's relative performance also highlights the possible gains from forecast combination as a way of hedging against instability in any particular forecaster's individual performance.

[11] We include a break between the two sub-samples to ensure that any observed correlation does not reflect the overlapping nature of the time series of forecasts.

3.2 Balancing the panel

As highlighted in the recent study by Capistrán and Timmermann (2009), a major practical challenge that arises in forecast surveys is the frequent and extensive 'entry' and 'exit' of participants from the SPF panel. Focussing on the filtered data (i.e. which includes only those regularly participating as defined above), depending on the variable or horizon, the employed filter yields a panel of (approximately) between 30 and 40 forecasts. However, even the filtered data involves several missing values and gaps reflecting the entry and exit of forecasters.[12] As these gaps need to be filled in order to implement several of the combination methods discussed in Section 2, we propose a simple panel regression approach to balance the panel and fill these gaps. Our simple approach, focuses on the dynamics of relative forecast performance using the panel regressions of the form:

$$\hat{y}_{i,t+h} - \bar{y}_{t+h} = \beta_i (\hat{y}_{i,t+h-1} - \bar{y}_{t+h-1}) + \varepsilon_{i,t+h} \tag{3.1}$$

(3.1) posits a simple AR(1) process, whereby the relative deviation of each forecaster to the simple average in period t is linked to its relative deviation in period $t-1$. When imposing $\beta_i = \beta = 1.0$, (3.1) implies missing observations for individual forecasts are set equal to the previously reported individual forecast updated with the change in the simple average of those forecasters who do respond. For $0 \leq \beta \leq 1.0$, the missing values for forecaster i in period t are replaced with the period t average forecast plus a fraction of the previously observed deviation from the average forecast. In implementing equation (3.1), β can be estimated recursively over the sample period to ensure that the method used to balance the panel preserves the pseudo real time nature of the resulting dataset. Alternatively, one could use a factor model and an EM algorithm to fill out missing observations, see Stock and Watson (2002).[13]

[12] In each filtered panel, the share of missing observations in the total panel is approximately 5%. For example, for the 1-year ahead GDP forecasts with N = 38 participants and 40 time series observations collected over the period 1999:q3-2009:q3, 79 of the 1520 panel observations were missing due to entry and exit of forecasters.

[13] An alternative even simpler - though possibly more controversial approach - would set the first missing observations for an individual's forecast equal to his/her previously reported forecast. This "*naïve updating*" can then be applied recursively through the sample to give a fully balanced panel. We have examined the sensitivity of the results to the choice of method used to balance the panel and find them to be overall insensitive to this choice.

3.3. Real time data issues

A key practical complication that arises in forecast combination and forecast evaluation relates to the impact of data revisions. Survey forecasts are by definition "real time" in the sense that they cannot use information that was unavailable at the time the survey was carried out and combinations of such forecasts also possess a corresponding real-time dimension. However, data revisions alter the estimate of the outcome for the forecast target variable and the evaluation of alternative combinations may therefore be sensitive to the choice of vintage of data used to define the target variable.[14]

To get a sense of the relevance of the role of such real time issues for the evaluation of SPF combinations, Figure 5 plots the difference between the 1st estimate provided by Eurostat for each of the three SPF variables and the corresponding "current" estimates available in 2009:Q3.[15] From the chart it is clear that substantial revisions in euro area data are apparent for both GDP and the unemployment rate. Compared with initial published results, the 2009:Q3 estimates of euro area GDP has been revised upward substantially over most of the period since 1999.[16] Similar sized revisions are evident for the unemployment rate (with downward revisions in the first half of the sample being followed by significant upward revisions subsequently). In the case of inflation, there have been more limited revisions overall and mainly in the early years of the sample.

Figure 5 would suggest a clear need to consider the possible impact of the vintage of data used for the target variable on the evaluation of alternative forecast combinations. *A priori*, however, there is no simple rule which could guide the choice

[14] It is not just the evaluation of alternative combinations but also the estimation of combination weights that will be sensitive to the data vintage used to derive the target variable. All estimated combinations discussed in Section 2 - with the exception of the simple trimming and the equal weighted combination - may be sensitive to the vintage of the data used to define the target variable. For example, under performance-based combination, historical revisions to the outcomes for the target variable may imply possible changes in the relative weight on any given forecaster. Similarly, the estimated coefficients in regression-based combinations may be sensitive to the vintage of the time series used as the dependent variable in the combination regression. The pseudo real time approach in this study discards the relevance of data revisions on estimated combinations for the sake of simplicity but this is clearly a relevant area for future research.

[15] Our real time data is fully consistent with the estimates in the real time database of the Euro Area Business Cycle Network as described in Giannone, Henry, Lalik and Modugno (2010).

[16] The importance of such data revisions for the euro area is similar to the evidence for the US which is reviewed in Croushore and Stark (2003).

of data vintage to use for the evaluation. On the one hand, to the extent that a new data release is pure "news", the associated revision should be completely unpredictable. This would suggest that revisions could be ignored in forecast evaluation and, in line with this news hypothesis, our baseline results use the 1st estimate for each variable in deriving forecast performance statistics. However, to the extent that measurement error (or "noise") partly accounts for subsequent data revisions, they may have a predictable component which would suggest a possible preference to focus on the revised vintages of data in the forecast evaluation.[17] Given this alternative hypothesis, we report our forecast evaluation results also using the current vintage of estimates for the three macro variables analysed and check the sensitivity of the performance of different combinations to this alternative choice.

4. Forecast Performance Measurement

In this section we briefly discuss the out-of-sample performance evaluation methods we use to assess the alternative combination strategies discussed in Section 2. The evaluation is presented in the form of the Mean Squared Error (MSE) of the alternative SPF combinations ($\hat{y}_{c,t+h}$) *relative* to the benchmark equal weighted combination, $\bar{y}_{t+h}$. Assuming our "holdout" sample (used for the out-of sample evaluation) runs from period T_1 to period T, our performance evaluation measure is given by:

$$\text{(Relative) MSE} = \frac{\sum_{t=T_1}^{T}[y_{t+h} - \hat{y}^{c}_{t+h}]^2}{\sum_{t=T_1}^{T}[y_{t+h} - \bar{y}_{t+h}]^2} \tag{4.1}$$

[17] Clark and McCracken (2008) analyse the impact of data revisions on forecast evaluation and show that such revisions can significantly impact the asymptotic behaviour of tests of equal predictive ability. In particular, they highlight that the result in West (1996) that parameter estimation error can be ignored in tests of predictive accuracy only holds under the "news" hypothesis. In contrast, in the presence of noisy data revisions, parameter estimation error contributes to the variance of the test statistic and cannot be ignored in inference.

and will be less than unity whenever the alternative combination performs better than the simple equal weighted combination. To help gauge the overall performance of the equal weighted benchmark, we also report the Relative MSE for three simple time series models. In particular, we consider a Naïve forecast which sets the projected level of the variable equal to its current level as known at the time of the survey and allowing for publication lags.[18] We also estimate a Random Walk with drift for the seasonally adjusted *level* of GDP and the *level* of the consumer price index (HICP).[19] Lastly, with a view to capturing any persistence in the dynamics of the three variables, we also estimate an AR(1) process for the log change in GDP and HICP and for the level of the unemployment rate. While these time series benchmarks are quite simple, they have proven to be quite difficult to beat in practice - particularly at horizons beyond 1 quarter ahead. They therefore provide a reasonably good time series benchmark against which to assess the performance of the SPF. In addition to examining Relative MSEs, it is useful to be able to assess the overall statistical significance of any observed difference in forecast performance. Therefore we also report the results of Diebold and Mariano (1995) test (DM) for the null hypothesis that a given combination fails to beat the equal-weighted benchmark. This test is particularly suited to the evaluation of multi-period forecasts as is the case here and where there is evidence of non-Gaussian forecast errors which are likely to be serially correlated (see Mariano (2002)).

The DM statistic provides statistical evidence about whether *a particular* combination performs better than the equal weighted benchmark. Ideally however we would like to assess the extent to which the evidence in favour of the *best* performing combinations is robust in a statistical sense. Given that we are evaluating a large set of potential combinations repeatedly using the same historical dataset, chance alone may be able to explain a statistically significant DM result for any given combination. Therefore, as a further test of the robustness of our findings, we also report the results of the

[18] Publication lags imply that the known current level for each forecast variable is approximately lagging the survey month by 1 month in the case of inflation, by 2 months in the case of the unemployment rate and by 1 quarter in the case of the annual GDP growth. This information on the level of each forecast variable that is known at the time of the survey is provided to ECB SPF participants when they receive the survey questionnaire.

[19] Given that it is not a clearly trending variable like GDP and the HICP, a random walk without drift would seem more appropriate for the level of the unemployment rate. This is, however, equivalent to the Naïve forecast for the level of the unemployment rate.

White (2000) reality check for data snooping. The reality check tests the null hypothesis that the expected performance of the best performing model is no better than the benchmark. As such the test provides useful information as to whether or not the identification of some improvement compared with the equal weighted combination is merely the result of data mining (and therefore perhaps less likely to persist over time). Denoting f_j as a measure of the out-of-sample forecasting performance of the j^{th} combination (j = 1, . . . , J) relative to the benchmark, e.g., $MSE_j - MSE_0$, where the benchmark (represented by model zero) is the equal-weighted combination, the White Reality Check (RC) test can be applied to the performance statistic:

$$\text{White RC} = \min_j \left\{ T^{\frac{1}{2}} \bar{f}_j \right\} \tag{4.2}$$

Where $\bar{f}_j$ is the sample mean of the forecasting performance of model j *measured relative to the performance of the benchmark*, i.e.

$$\bar{f}_j = T^{-1} \sum_{t=1}^{T} f_{t,j} = (\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_J) \tag{4.3}$$

While a closed form solution for the distribution of the minimum in (4.2) above is not available, it can be approximated using a bootstrap sampling procedure and the relevant P-values can then be reported for the null hypothesis that the expected performance of the best performing combination is no better than the equal weighted combination.[20] This provides some indication as to whether or not the findings are robust to possible "data snooping" bias.

One limitation to the reality check procedure is that it assesses the performance of the best combination scheme jointly with the performance of a large cross-section of competing specifications. Under such circumstances, the power of the test may be

[20] See White (2000), Sullivan, Timmermann and White (1999) and Qi and Wu (2006) for further information and practical applications of the reality check.

reduced due to the inclusion of poorly performing models.[21] We therefore consider one final alternative to help shed light on the overall robustness of the out-of-sample results. Suppose, for example, that the decision maker has some way of selecting in real time the best combination. This would happen if, e.g., the best combination scheme dominates other models from an early point in the sample onwards. We evaluate the performance of the combinations identified from such a 'meta' rule that recursively selects the top combination from among all specifications discussed in Section 2 against the benchmark over the out-of-sample evaluation period. As an example, subject to having a minimum initial track record, at each point in time one could choose that combination strategy which historically (up to that point in time) generates the smallest MSE-value. The identity of this model may change through time so we are effectively referring to the forecasting performance of the combination selection or 'search' rule. Such a meta selection procedure is not subject to the 'data snooping' criticism since it effectively only considers one combination strategy at each point in time.

5. Results

In this section we first discuss the performance of the equal weighted SPF combinations compared with alternative time series and other simple benchmarks. We then turn to a comparison of the alternative classes of combination methods with the equal weighted combination and conclude with an examination of the overall robustness of our findings in terms of their sensitivity to real time data issues, their stability over time and the reality check to assess the relevance of possible data snooping bias.

5.1 Comparison of SPF with statistical benchmarks

Tables 2, 3 and 4 report the baseline out-of-sample evaluation results for GDP growth, HICP inflation and the unemployment rate respectively for both 1- and 2-year ahead horizons. The results are reported in the form of the MSE for the various combinations relative to the MSE of the equal weighted combination. The performance statistics are calculated for the period 2004:q1 to 2008:q3, i.e. over

[21] Hansen (2005) has therefore suggested a modification to White's test aimed at reducing the impact of irrelevant alternatives using a studentized test statistic and incorporating additional sample information by means of a data dependent null distribution.

"normal" business cycle conditions and excluding the very large macroeconomic shocks associated with the 2008-2009 financial crisis which had a strong impact on all variables particularly from 2008:q4 onwards. In addition to the relative MSE, the P-value from the Diebold-Mariano test is reported for each combination, providing an indication of the likelihood that the equal weighted combination outperforms the alternative combinations.

One notable feature looking at the tables is the relatively good performance of (equally-weighted) SPF forecasts for the real variables (GDP and unemployment) relative to the simple time series models (Random Walk, Naïve or AR(1)). Only in the case of inflation, do Naïve time series predictors outperform the SPF average at both one and two-year ahead horizons. For example, the Random Walk with drift for the level of the price index significantly improves on the equal weighted average by over 20% at both the one- and two year ahead horizons. These results highlighting the relatively good performance of univariate time series models for inflation are consistent with previous studies of the SPF (e.g. Bowles et. al. (2007) and Bowles et al. (2010)) and euro area inflation forecasting (e.g. Benalal et al. (2004) and Giannone et al. (2010)). From the perspective of the ECB, in the case of inflation, the relatively poor performance of the equal weighted SPF forecast certainly motivates the case for examining the extent to which alternative SPF combinations might improve on the existing information extracted from the survey. As a final comment on the comparison with other simple benchmarks, it is notable that there are no significant gains from either trimming or focusing on the median SPF forecast for any variable or any horizon. This suggests very little evidence of "noisy" forecasters in the ECB SPF panel in line with the survey's policy of including only "professional" forecasters with a sound reputation and experience in producing euro area forecasts.

5.2 Relative performance of alternative combinations methods

To get a sense of the relative performance of alternative combinations, Figure 6 summarises the detailed results in Tables 2, 3 and 4 by displaying the relative MSEs for the *best* performing specification within 7 main combination categories for all variables and both horizons. The evaluation statistics are computed using the "real time" or 1st vintage of the target variable as the actual outcome in line with the hypothesis that subsequent revisions are pure news and therefore unpredictable (in

Section 5.3 below we assess the sensitivity of the results to the choice of vintage for the outcome series). When the performance gain from a combination method is statistically significant according to the Diebold-Mariano test, Figure 6 also reports the relevant P-value.

Looking first at the results for GDP in Figure 6(a), several alternative combinations outperform moderately (by approximately 10%) the equal weighted combination. At the 1-year horizon, the gains are strongest for least squares combinations and to a lesser extent combinations based on shrinkage. In the case of the 2-yr ahead horizon, the scope for improving on the performance of the equal weighted combination appears much smaller (see also the detailed results in Table 2 where for this horizon almost all specifications have relative MSEs that exceed unity). However, the performance based combination using a short rolling window (v = 1 quarter) also demonstrates a quantitatively noticeable 13% improvement which is also statistically significant according to the DM statistic. Given that just over 30 specifications for alternative combinations have been tried, this finding may reflect data snooping bias (see the discussion in Section 5.5 below).

Figure 6(b) also depicts the relative performance of alternative combination strategies for the SPF inflation forecasts. In the case of inflation, several of the alternative combinations outperform the equal weighted one. The gains are most clearly evident for regression based combinations such as the projection on the mean, principal components combination as well as least squares and shrinkage based combinations (see also the detailed results reported in Table 3). However, several of the performance based strategies also outperform the benchmark, most noticeably the strategy of using the recent best forecaster. In a number of cases the relative gains from alternative combination strategies are both quantitatively important and statistically significant according to the DM statistic. At the 1- and 2-year ahead horizons the best models for inflation are, respectively, the least squares and the projection on the mean, both without any bias adjustment. Compared with the equal weighted combination, the best performing models deliver a quantitative reduction in the MSE that is greater than 40%. The relatively good performance of alternative combinations for inflation may link to the very persistent downward bias in the equal weighted combination over the sample period analysed as well as some evidence of

positive persistence in individual forecaster performance shown in Figure 4. In particular, the alternative combinations may better adjust for bias and better exploit observed persistence in forecaster performance relative to the equal weighted benchmark combination. Finally, the fact that so many of the alternative specifications outperform the benchmark in the case of inflation may be indicative that the results are not just a reflection of data snooping bias although we provide a more rigorous analysis of this question in Section 5.5 below.

Finally, the corresponding results for the unemployment rate, in Figure 6(c), indicate a relatively low scope for achieving a quantitatively important improvement in forecast performance relative to the equal weighted combination. At both horizons only very few alternative combination methods have MSEs that are smaller than the equal weighted combinations. However, at the shorter horizon, the combinations based on shrinkage weights suggest some significant gains (see also Table 4).

5.3 Sensitivity to definition of target variable

Figure 7 provides the first of some robustness checks on the results discussed in Section 5.2 above. It reports the relative MSEs computed using the "current" (i.e. 2009:q3) vintage of outcome variables and compares them with those computed using the real-time or 1st vintages (as used in Figure 6). In general, the overall results do not appear excessively sensitive to the choice of target variable, although the performance of the alternative forecast combinations is generally slightly worse in the case of the GDP growth and unemployment rate forecasts when the more recent 2009:q3 data vintage is used. At the same time, the best performing combination methods (shrinkage, constrained least squares or performance-based depending on the variable and horizon) continue to perform best when evaluated against the current vintage of the outcomes. The scope for improving on the equal weighted combination remains most evident for the case of inflation; indeed the performance of the various inflation combinations is broadly unaffected given that inflation was hardly revised during the evaluation period (as also seen from Figure 5).

5.4 Sub-sample stability: 2008-2009 financial crisis effects

Given the exceptional impact of the financial crisis on forecast performance as reflected in the summary statistics reported in Table 1, it may be insightful to examine

the sensitivity of the performance of alternative combinations to the inclusion of the crisis period in the evaluation sample. In particular, this will allow some assessment of whether the results obtained during times of normal business cycle fluctuations are also applicable during periods of exceptional macroeconomic volatility. It may also help identify combination strategies that offer superior performance during periods of exceptional economic change.

Figure 8 summarises the results for the SPF sample which includes the last four observations from 2008:q3 to 2009:q4 and which is therefore strongly influenced by the financial crisis. Compared with Figure 6, it is indeed clear that the results are sensitive to the crisis period. One feature is that for GDP (at both horizons), there is some improvement in the relative performance of several alternative combinations when the sample is extended to include the crisis period. The best performing combination at the one year horizon is the recent best forecaster, although according to the DM test (P-value = 0.29) it is not significantly better than the benchmark. Several other regression based methods also improve moderately on the GDP benchmark at both 1 and 2-year ahead horizons. A similar picture emerges from the unemployment rate forecasts where, once again, relative performance improves compared with the results from the sample period excluding the extreme volatility of end 2008 and 2009. In the case of inflation, the performance improvement relative to the benchmark - while nonetheless remaining significant - is lowered as a result of the inclusion of the crisis period.

The above comparison of the results from the two samples tends to suggest that models which perform well during normal times may not be best suited to periods of exceptional macroeconomic volatility. This can be seen easily in Table 5a and 5b which reports the best performing combinations for each of the two samples. From the tables it can be seen that for no variable or horizon is it the case that the best performing specifications is unchanged when the sample period is extended to include the crisis. During the normal times the methods which tend to dominate are either constrained least squares, shrinkage or diversified performance based weighting. In contrast, during the crisis period, a number of the unconstrained least squares combinations as well as a strategy of picking the recent best forecaster tend to perform best. Such combinations in general allow more adaptability in the weights to

changing economic circumstances and this may partly explain their better performance during times of exceptional changes in the macroeconomic environment.[22] For the sample strongly impacted by the crisis errors, it is noteworthy that a strategy of picking the recent best forecasters performed better than the benchmark for all three variables at the 1-year ahead horizon (see Figure 8). Such a result points to the gains that may arise from placing all the weight on the forecaster adapting his/her outlook to the crisis environment (and the losses which may arise if positive weights continue to be given to forecasters who have not adapted their outlook to the rapidly changing environment).

5.5 The "reality check" for data snooping

Table 5a and 5b also reports the results of the "reality check" procedure described in Section 4 to assess the relevance of possible data snooping bias in the empirical results. The results are reported in the form of P-values which provide an estimate of the likelihood that the best performing model does *not* outperform the equal weighted combination. The White P-values are reported together with the "Nominal" P-values from the standard DM tests taken from Tables 2, 3 and 4. Given the role of influential observations linked to the macroeconomic effects of the 2008-2009 financial crisis, we report the reality check for both the sample excluding (Table 5a) and including (Table 5b) these observations.

In general, the reality check results highlight the major caveat applying to the apparent success of some models in forecast "horse races" of the type we have undertaken here, particularly in a context where the number of "horses" running in the race is quite large and the "course" (the sample period available) is relatively short. In particular, for both GDP and unemployment, the reality check suggests that significant gains identified by the DM test may be a reflection of data snooping bias. In the case of both variables the test indicates that it is more likely than not that the best performing models do not outperform the equal weighted combination. In the case of inflation, very much in line with the overall stronger evidence on the scope for improvements in relative performance as stressed in Section 5.2, the white P-values are considerably lower. Indeed for both the 1- and 2- year horizons, the reality check

[22] From a practical point of view, an important question concerns the length of time needed to identify the recent best forecaster. As seen in Table 5b, the model with a 4 quarter window is best for the sample including the crisis although the performance based on a shorter window is almost equally as good as the latter for all three variables.

tends to confirm that the best performing models are more likely than not to outperform the equal weighted combinations even when we control for the effects of possible data snooping bias. This is also consistent with the evidence on the inflation forecasts in Table 3 showing a much larger fraction of alternative combination strategies outperforming the benchmark. However, only for the 1-year ahead inflation forecasts, is the reality check indicating a significant improvement of the best performing model relative to the benchmark at the 10% significance level.

The above relatively strong result for inflation is only valid during "normal times", however. When the sample period is extended to include the period of high macroeconomic volatility from end 2008 onwards (Table 5b), the improvements identified for all alternative combinations appear to "fail" the reality check at standard levels of significance. These findings are therefore more in line with the often reported result in the combination literature on the difficulty of being able to outperform an equal weighted combination in practice. They also tend to caution against any tendency to take a relatively good past performance among the alternative combination strategies as a strong indication of a likely better performance in the future.

5.6 A recursive "meta" selection procedure

Table 6 reports the results from the evaluation of the meta selection procedure described in Section 4 for both the out-of-sample evaluation periods, i.e. excluding and including the financial crisis period. For the period which excludes the impact of the crisis, quantitative improvements are identified for inflation at both one and two year horizons and for GDP growth at the 2-year horizon. For GDP at the 1-year horizon and the unemployment rate at both horizons, the meta selection rule performs worse than the equal weighted combination out of sample. According to the P-values from the Diebold Mariano tests, some of the improvements are also statistically significant for inflation (in line with the reality check result above). When the sample is extended to include the crisis, the meta selection procedure generally performs worse, again in line with the reality check results. One exception is the inflation combination 1 year ahead which achieves a 5% improvement over the benchmark although this gain is not significant at the 10% level.

6. Concluding remarks

In this paper we have reviewed the potential for forecast performance improvements through the application of forecast combination methods to the ECB Survey of Professional Forecasters. Our analysis is based on a pseudo out-of-sample comparison of four broad classes of linear combination methods and just over 30 alternative combination specifications compared with the equal weighted benchmark which is currently the headline indicator from the SPF that is reported to policy makers and the public.

Our main findings can be summarised as follows: Over the sample period analysed, the equal weighted combination sets a reasonably high benchmark in the sense that it is shown to be quite informative when measured against other time series and Naïve forecasts. Notwithstanding the relatively good performance of the SPF benchmarks, a number of alternative combination strategies are shown to achieve quantitatively important gains relative to this benchmark in an out-of-sample "horse race" conducted over the five year period from 2004:q1 to end 2008:q3. Looking across variables, the scope for improvements from alternative combination strategies appears the most significant for inflation with smaller gains achievable for GDP and, especially, for the unemployment rate. The relatively good performance for inflation combinations links to a downward bias in the equal weighted combination as well as some evidence of positive persistence in individual forecaster performance which the alternative combination strategies are able to exploit. However, in general, our results do not identify any single combination approach which appears to dominate across either variables or at different horizons. Instead, depending on the horizon and the variable, the best performing combination methods include least squares, Bayesian shrinkage as well as more simple strategies where the weighting is determined only by relative past performance.

We have also examined the sensitivity of the above results across a number of key dimensions. Firstly, in general, the findings are insensitive to the chosen vintage of the target variable used in the forecast evaluation. Moreover, when the sample is extended to include the most recent period of large macroeconomic volatility associated with the 2008/2009 financial crisis, some of the alternative combination

strategies (least squares, Bayesian shrinkage and the recent best forecaster) continue to perform better than the equal weighted benchmark. However, the best performing model for each variable and horizon differs when comparing the results including the crisis period with those covering periods of more normal business cycle fluctuations. During the crisis, it is noteworthy that a strategy of picking the recent best forecasters performed better than the benchmark for all three variables at the 1-year ahead horizon. In addition unconstrained least squares methods generally perform better during the crisis. Such results point to the possible gains that may arise from combination methods which allow the weights adapt most quickly in a context where there are large and persistent shocks to the macro economic environment. In such a context, for example, there may be gains associated with placing all the weight on the forecaster adapting his/her outlook to the crisis environment and a corresponding loss in performance which may arise if positive weights continue to be assigned to forecasters who have not sufficiently adapted their forecasts to the changing economic situation.

Finally, we have also assessed the sensitivity of our analysis to possible data snooping bias using the reality check suggested by White (2000) together with a novel meta procedure rule that should be robust to data snooping critiques. In general, both of these "reality checks" highlight an important caveat applying to the apparent success of some models in forecast "horse races" of the type we have undertaken. In particular, given that we have applied a large number of combination models repeatedly on the same small evaluation sample, some of the quantitatively sizeable improvements identified relative to the SPF benchmark may be simply due to chance. In line with this, only for the 1-year ahead inflation forecasts, is the reality check indicating a robust improvement of the best performing model relative to the benchmark at the 10% significance level. For the other variables and horizons, the reality check highlights the difficulty of being able to statistically outperform an equal weighted combination in practice. Such results tend to caution against any assumption that the identified improvements relative to the equal weighted benchmarks would necessarily persist in the future.

Overall, we would conclude from this study that there exists a reasonably good case to consider alternative combinations as a means of more optimally summarising the

information collected as part of the regular quarterly rounds of the ECB SPF. Among the range of models considered, it is notable for example that the equal weighted model is never the best performing model in the out-of-sample evaluation. However, the variation in the best performing specification through time, across target variables and across horizons together with the likely role of chance in explaining the success of some models in our sample would caution against any temptation to try and pick out a preferred or best combination method. Rather, our results would argue in favour of reporting a suite of alternative combinations which forecast users could draw on taking into account the historical track record of individual combination methods and the prevailing economic context.

Bibliography

Aiolfi, M. and A. Timmerman (2006) "Persistence in forecasting performance and conditional combination strategies", Journal of Econometrics, 135, 31-53.

Aiolfi, M., Capistrán, C. and A. Timmermann (2010) "Forecast Combinations", Unpublished manuscript.

Bates, J.M. and C, W. J. Granger 1969. "The combination of Forecasts", Operations Research Quarterly 20, 451-468

Benalal N., J.L. Diaz del Hoyo, B. Landau, M. Roma and F. Skudelny (2004): "To aggregate or not to aggregate?" Euro Area Inflation Forecasting, ECB Working Paper series No. 374.

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler and T. Rautanen (2010), "An evaluation of the growth and unemployment rate forecasts in the ECB SPF", *forthcoming*, Journal of Business Cycle Measurement and Analysis.

Bowles, C., R. Friz, V. Genre, G. Kenny, A. Meyler and T. Rautanen (2007), "The ECB Survey of Professional Forecasters: A Review after eight years of experience", ECB Occasional Paper No. 59, April.

Capistrán, C. and A.Timmermann (2009) "Forecast Combination with Entry and Exit of Experts", Journal of Business and Economic Statistics, 27(4), 428-440.

Clark, T. E. and M. W. McCracken (2008) "Tests of Equal Predictive Ability with Real Time Data", Working Paper 2008-029A, August, Federal Reserve Bank of St. Louis

Clemen, Robert T (1989). "Combining Forecasts: A review and annotated bibliography", International Journal of Forecasting, 5, 559-583

Clemen, R. T. and R. L. Winkler (1986) "Combining economic forecasts", Journal of Business and Economic Statistics, 4, 39-46

Croushore, D. (2009) "Philadelphia Fed Forecasting Surveys: Their Value For Research", Paper presented to the International Workshop on Expectation Formation, Federal Reserve Bank of Philadelphia, 26-27 February.

Croushore, D and T. Stark (2003) "A Real Time Dataset for Macroeconomists: Does the data vintage matter", Review of Economics and Statistics, 85, 605-617

Diebold, F. X., and R. S. Mariano, "Comparing Predictive Accuracy", Journal of Business and Economic Statistics, v.13, no.3 (July 1995), pp. 253-63.

Diebold Francis X. and P. Pauly (1987), "Structural Change and the combination of forecasts", Journal of Forecasting, 6, 21-40.

Diebold Francis X. and P. Pauly (1990), The use of prior information in forecast combination" International Journal of Forecasting, 6, 503-508

Garcia, J. A. (2003), "An introduction to the ECB survey of professional forecasters", ECB Occasional Paper No. 8, September 2003.

Giannone, D, M. Lenza, D. Momferatou and L. Onorante (2010) "Short-term inflation projections: A Bayesian Vector Autoregressive approach", CEPR Discussion Paper, No. 7746, March 2010

Giannone, D., J. Henry, M. Lalik and M. Modugno 'An area-wide real time database for the euro area" ECB Working Paper No 1145, January 2010

Granger C. W. J. and P. Newbold (1986) Forecasting Economic Time Series, 2nd Edition, London, Academic Press

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

Hansen, Peter R. (2005) "A Test of Superior Predictive Ability", Journal of Business and Economic Statistics", October 2005, Vol 23, 4, 365-380

Hendry, D.F. and M.P. Clements (2002) "Pooling of Forecasts", Econometrics Journal, 5, 1-26.

Newbold P. and C. W. J. Granger (1974) "Experience with forecasting univariate time series and the combination of forecasts", Journal of the Royal Statistical Society Series A, 137, 131-46

Newbold, P. and David I. Harvey (2002) "Forecast Combination and Encompassing" Chapter 12 in A Companion to Economic Forecasting, Michael P. Clements and David F. Hendry (eds.) Blackwell, 2002.

Makridakis, S. A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parsen, and R. Winkler (1982) "The accuracy of extrapolation (time series) methods: results of a forecasting competition", Journal of Forecasting, 1, 111-153

Makridakis, S. and R. L Winkler (1983) "Averages of Forecasts: some empirical results", Management Science, 29, 987-996.

Mariano, R. S."Testing Forecast Accuracy" Chapter 13, Handbook of Forecasting, Michael P. Clements and David F. Hendry (eds.) Blackwell, 2002.

Patton, A. and A.Timmermann, 2010, "New Tests of Forecast Optimality Across Multiple Horizons". Manuscript, Duke and UCSD.

Qi, M. and Y. Wu (2006) "Technical Trading-Rule Profitability, Data Snooping, and Reality Check: Evidence from the Foreign Exchange Market" Journal of Money, Credit, and Banking Volume 38, Number 8, December 2006.

Smith, J. and K.F. Wallis (2009) "A Simple Explanation of the Forecast Combination Puzzle". Oxford Bulletin of Economics and Statistics, 71(3), 331-355.

Stock, J.H. and M.W. Watson (2002) "Macroeconomic Forecasting Using Diffusion Indexes", Journal of Business and Economic Statistics, 20, 147-162.

Stock, J.H. and M.W. Watson (2004) "Combining Forecasts of Output Growth in a Seven-Country Data Set", Journal of Forecasting, 405-430

Sullivan, R. A Timmermann, and H. White (1999) "Data snooping, technical trading rule performance and the bootstrap", Journal of Finance, 54, 1647-91

Timmermann, A. (2006) Forecast Combinations, Chapter 4, Vol 1, Handbook of Economic Forecasting, Graham Elliott, Clive William John Granger, Allan Timmermann (eds.), North Holland

West, K. (1996) "Asymptotic inference About Predictive Ability", Econometrica 64, 1067-1084

White, Halbert (2000) "A Reality Check for Data Snooping", Econometrica, Vol 68, No. 5, 1097-1126

Figure 1: SPF Panel – Equal weighted forecasts and outcomes (alternative vintages)

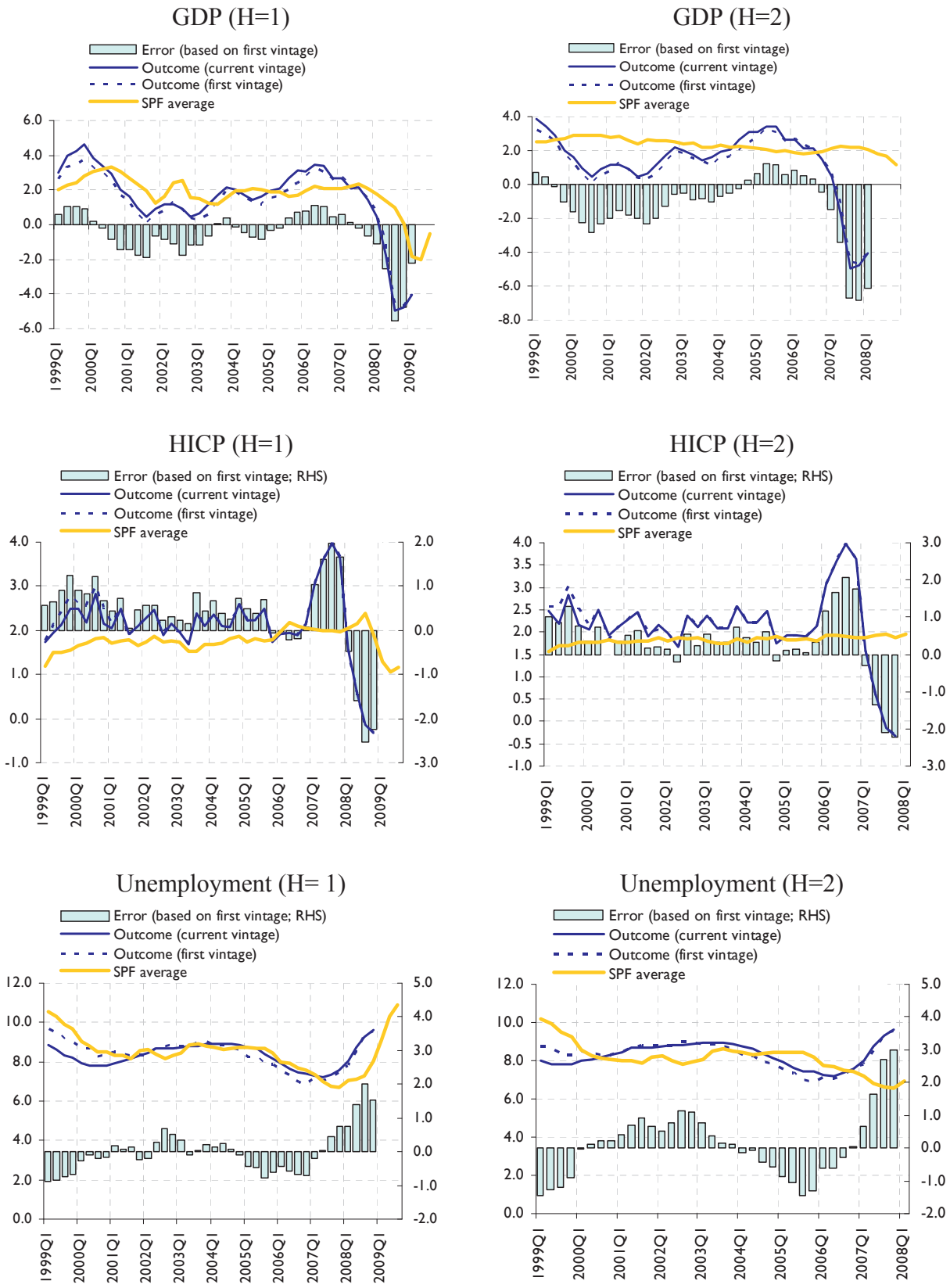
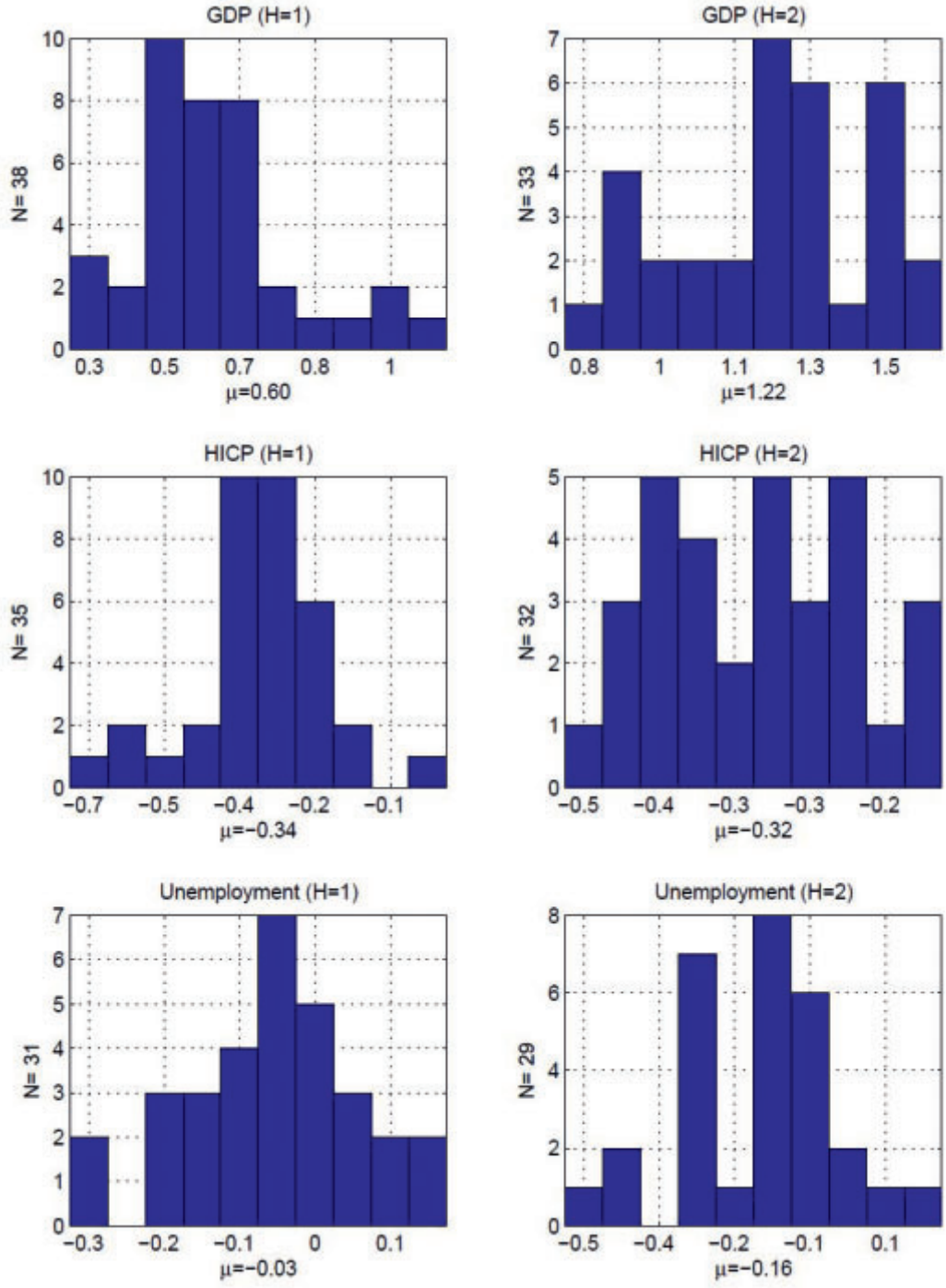
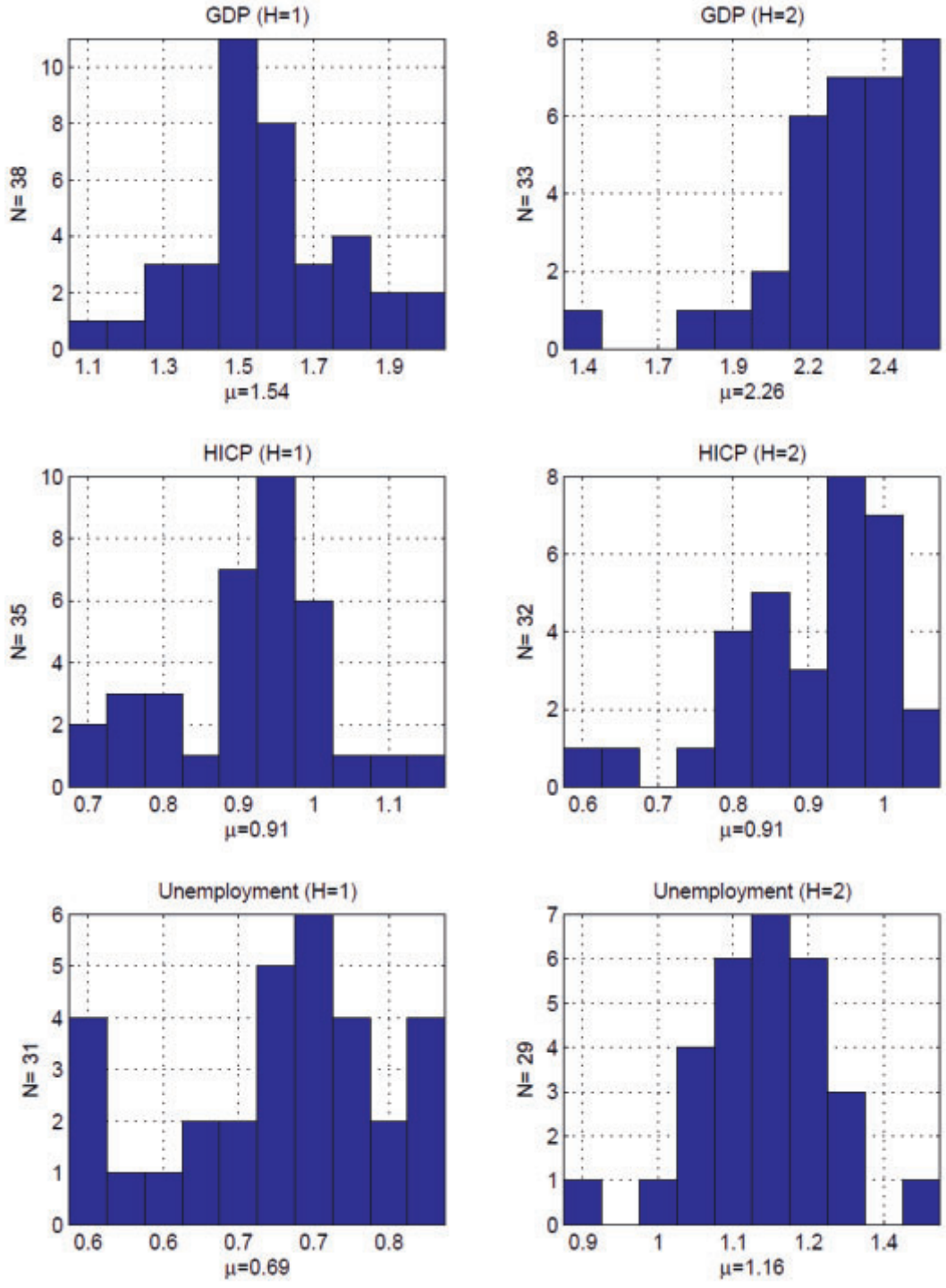


Figure 2: SPF Panel – Mean Forecast Errors (1999Q3-2009Q3)
(percentage points)



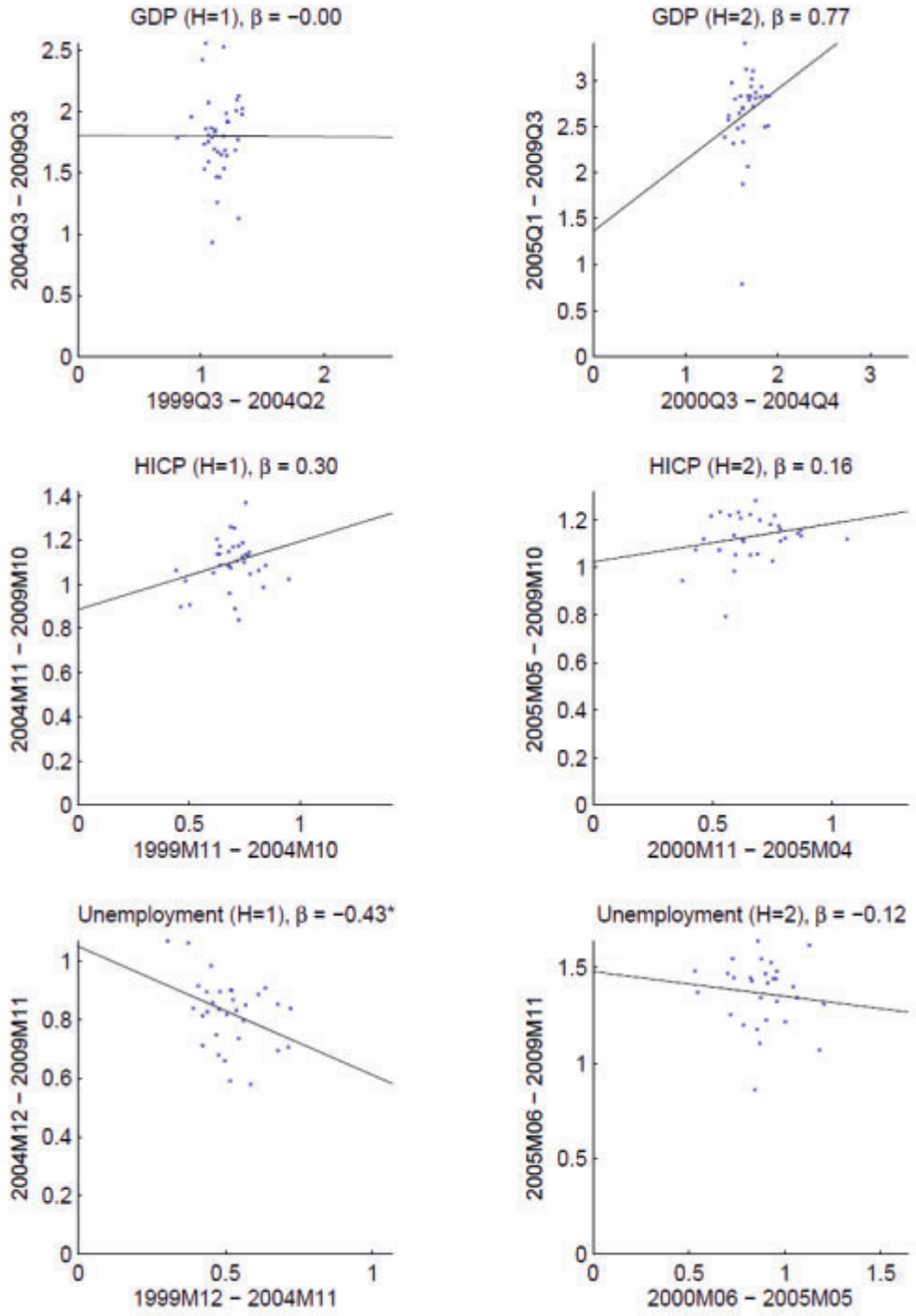
Note: The charts illustrate the distribution of mean errors across forecasters. Mean errors are calculated using the filtered SPF panel and the 1st estimate of the target variables as the actual outcome. μ denotes the average mean error across individual forecasters where the error is defined as forecast value less the actual outcome. H indicated the horizon (1 or 2 years) and N the number of forecasters included in the filtered dataset

Figure 3: SPF Panel – Root Mean Squared Forecast Errors (1999Q3-2009Q3)
(percentage points)



Note: The charts illustrate the distribution of root mean squared errors across forecasters (RMSEs). RMSEs are calculated using the filtered SPF panel and the 1st estimate of the target variables as the actual outcome. μ denotes the average RMSE across individual forecasters. H indicated the horizon (1 or 2 years) and N the total number of forecasters included in the filtered dataset Root MSE calculated using the filtered SPF panel and the 1st estimate of the target variables is the actual outcome. μ denotes the average Root MSE across individual forecasters.

Figure 4: SPF Panel: Persistence in Root Mean Squared Error across sub-samples
(percentage points)



*Note: The Root MSE calculated using the filtered SPF panel and the 1st estimate of the target variables as the actual outcome. β provides the estimate slope parameter from a regression (including a constant) of the RMSE in the first half of the sample on the RMSE in the second half of the sample. A * indicates that the estimate of β is statistically different from zero.*

Figure 5: Difference between 1st and 2009:Q3 vintages of outcomes for SPF variables
(percentage points)

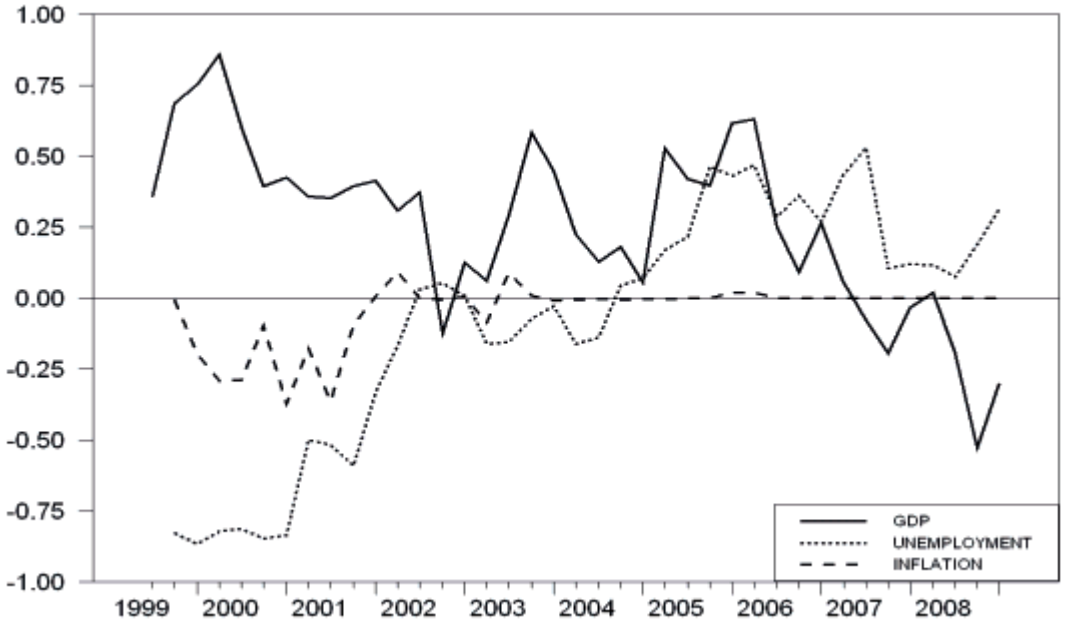


Figure 6: Comparison of forecast performance for alternative combinations
(Relative MSE; 1st vintage of target variable;2004:1 – 2008:Q3)

(a) GDP growth

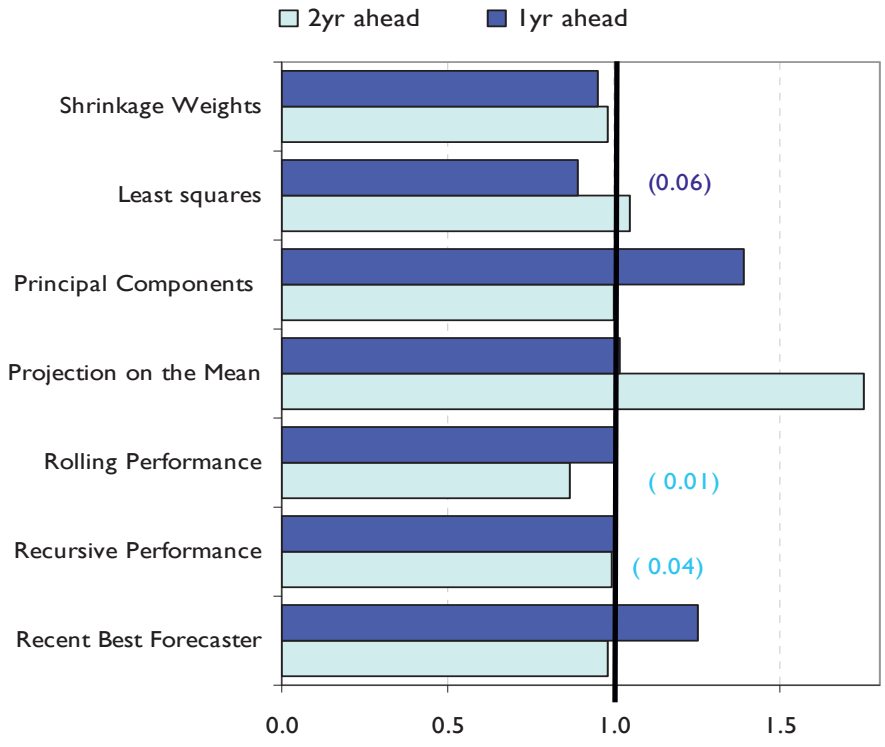
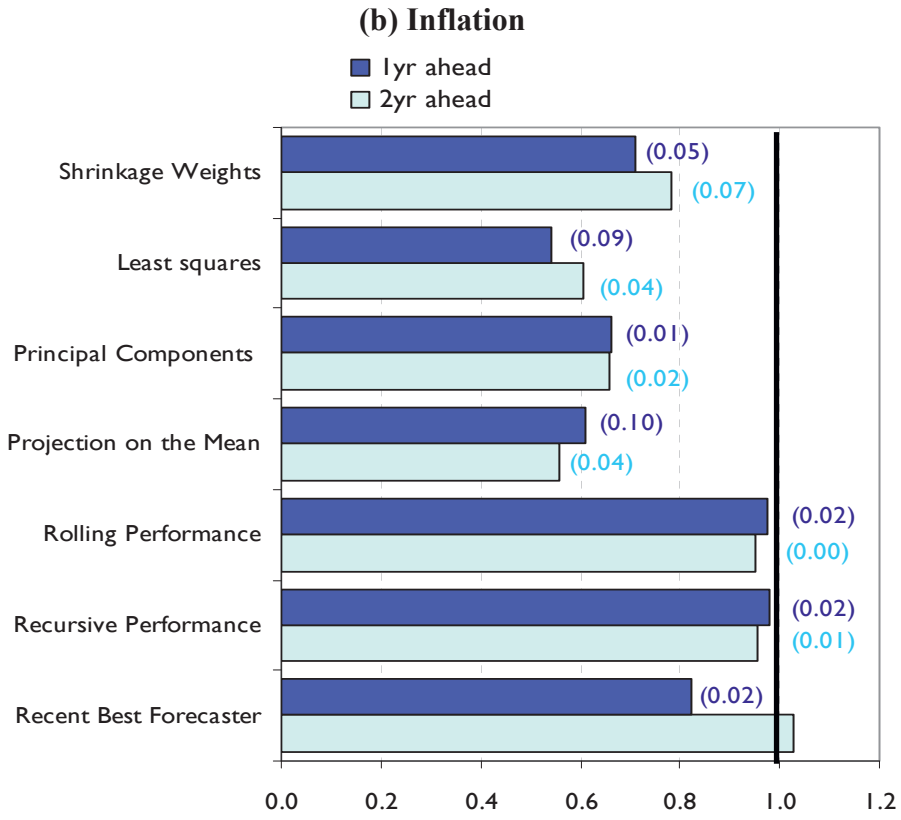


Figure 6: Comparison of forecast performance for alternative combinations
(Relative MSE; 1st vintage of target variable;2004:1 – 2008:Q3)



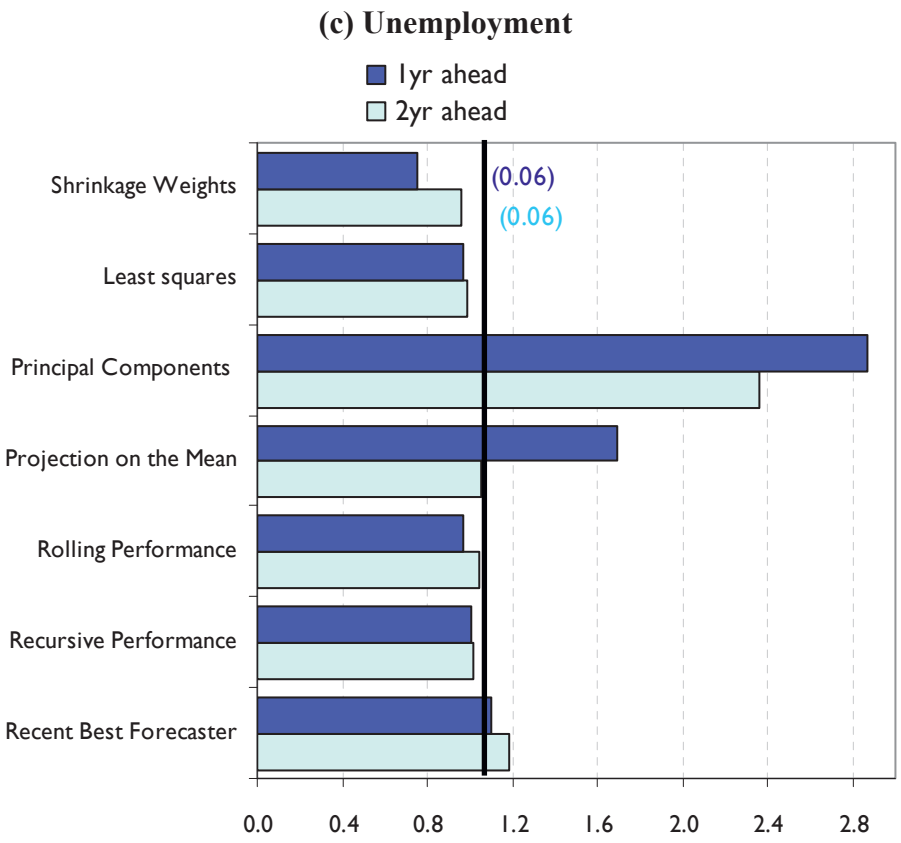
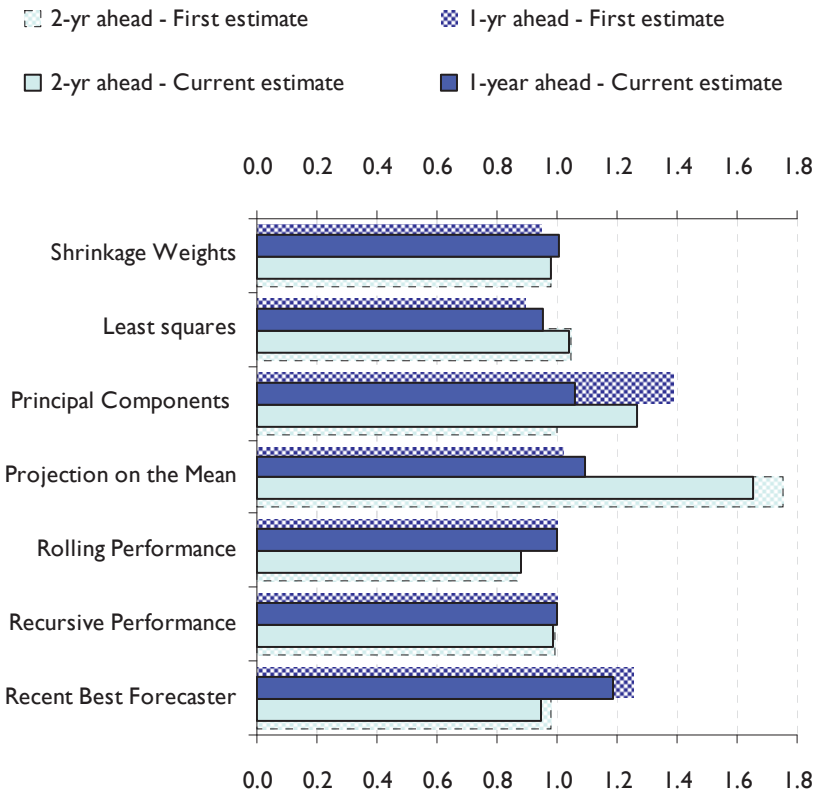


Figure 7: Alternative Combinations Relative MSE

(a) GDP growth



(b) Inflation

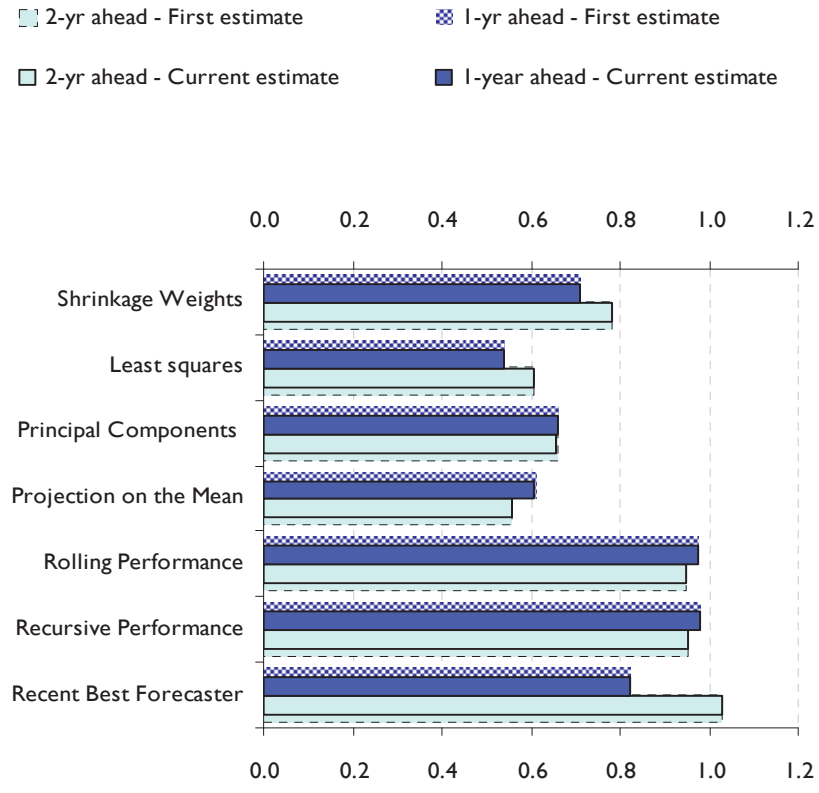


Figure 7: Alternative Combinations Relative MSE
(c) Unemployment

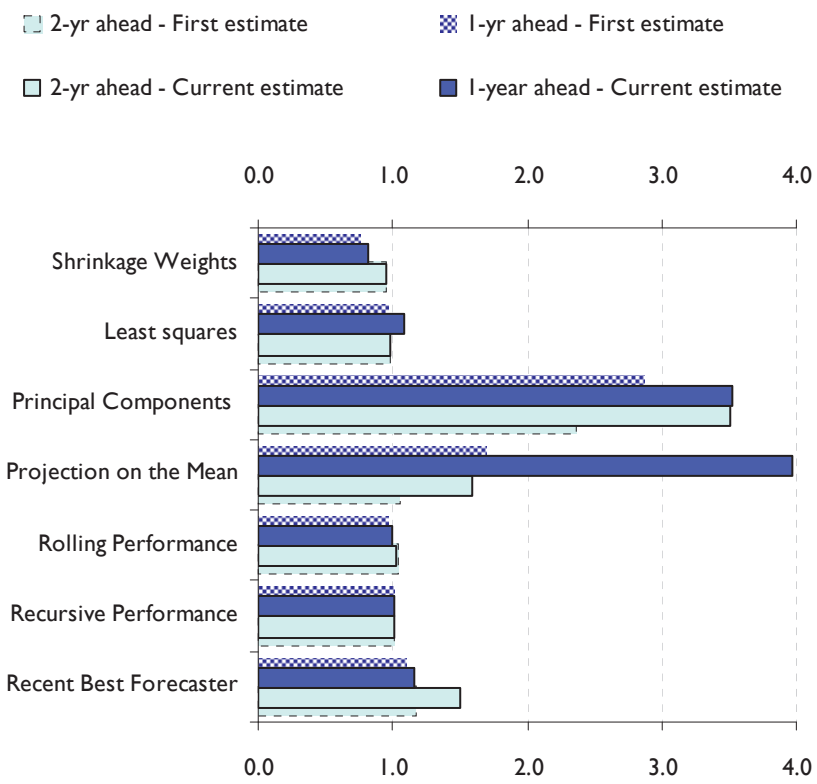


Figure 8: Alternative Combinations Relative MSE
(a) GDP growth

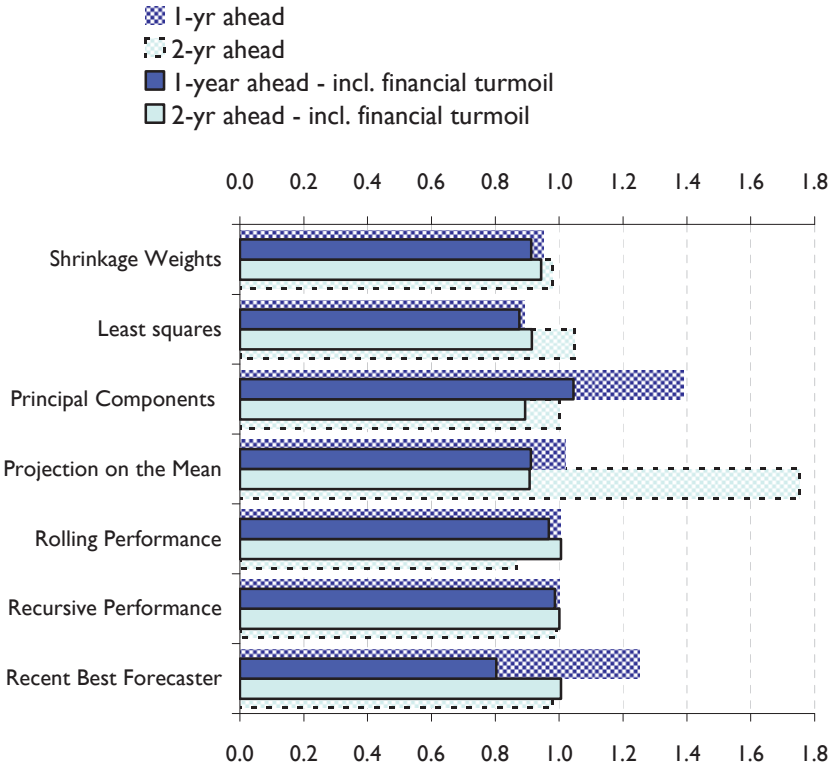
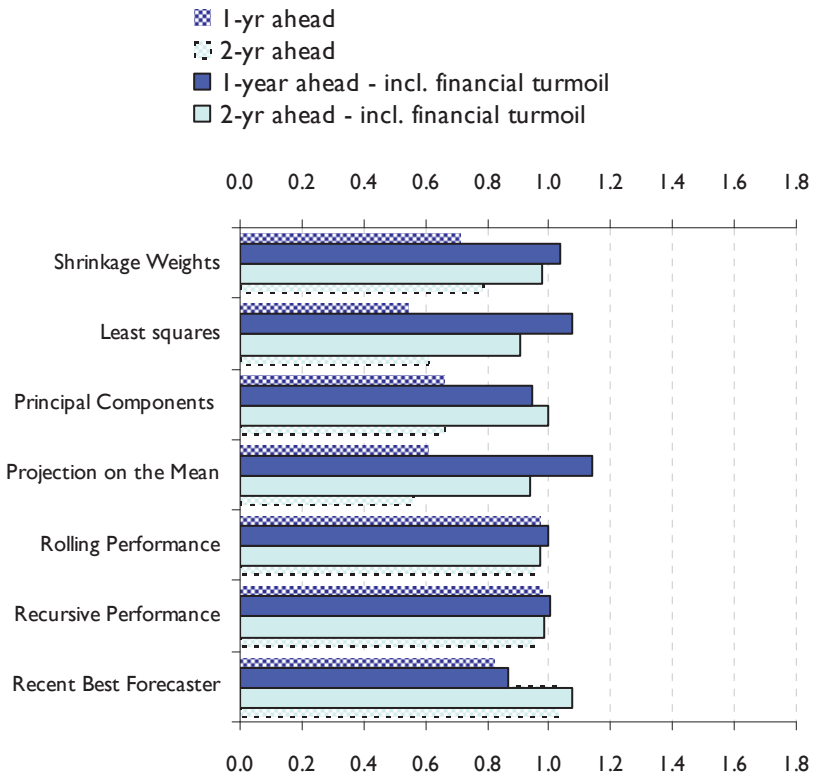


Figure 8: Alternative Combinations Relative MSE

(b) Inflation



(c) Unemployment

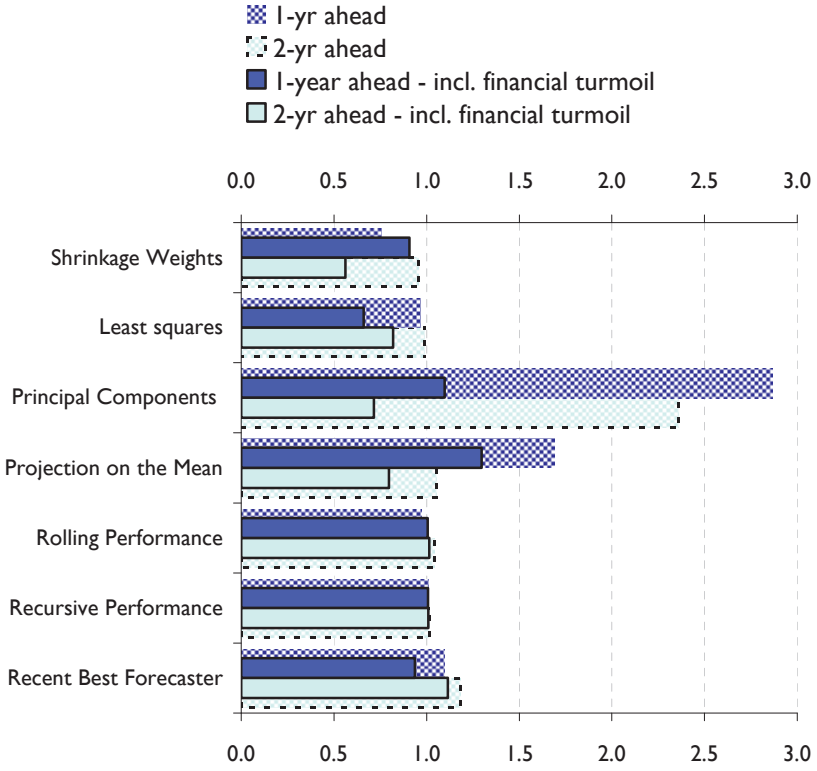


Table 1: Forecast performance statistics for the ECB SPF: Alternative samples

	Full sample (1999:3-2009:3)	Sample excluding financial turmoil (1999:3-2008:3)	First half of the sample (1999:1-2003:4)	2nd half of the sample (2004:1-2009:3)
GDP one-year ahead				
Mean forecast value	1.9	2.1	2.2	1.6
Mean error	-0.6	-0.3	-0.9	-0.7
RMSE	1.5	0.8	1.1	1.8
GDP two-year ahead				
Mean forecast value	2.4	2.4	2.7	2.4
Mean error	-1.3	-0.7	-1.9	-1.2
RMSE	2.3	1.8	2	2.7
Inflation one-year ahead				
Mean forecast value	1.8	1.8	1.7	1.9
Mean error	0.4	0.6	0.5	0.1
RMSE	0.9	0.6	0.6	1.2
Inflation two-year ahead				
Mean forecast value	1.8	1.8	1.8	1.9
Mean error	0.3	0.5	0.3	0.2
RMSE	0.9	0.6	0.4	1.1
Unemployment one-year				
Mean forecast value	8.4	8.4	8.9	7.9
Mean error	0	-0.1	0.1	0.1
RMSE	0.7	0.2	0.3	0.8
Unemployment two-year				
Mean forecast value	8.1	8.3	8.5	7.9
Mean error	0.1	-0.1	0.6	0.1
RMSE	1.2	0.6	0.7	1.2

Table 2: Evaluation of alternative SPF GDP combinations: 2004:Q1 – 2008:Q3

	H = 1 Year Ahead		H = 2 Years Ahead	
	MSE	P-Value	MSE	P-Value
Benchmarks				
Equal Weighted SPF	1.00	0.50	1.00	0.50
Random Walk	1.86	0.93	1.04	0.59
NAÏVE	1.95	0.95	1.75	0.89
AR(1)	1.80	0.95	3.35	0.96
Recent Best				
v = 1 quarter	1.25	0.97	0.98	0.42
v = 4 quarters	1.26	0.98	1.30	1.00
Trimmed means				
Symmetric Trim (5%)	1.03	0.78	0.98	0.24
Median (50%)	1.03	0.80	0.97	0.02
Recursive Performance (RP)				
$\delta = 1.0$	1.01	0.85	0.99	0.04
$\delta = 0.95$	1.00	0.81	0.99	0.07
$\delta = 0.85$	1.00	0.58	1.00	0.22
Rolling Performance (RP)				
v = 1 quarter	1.06	0.73	0.87	0.01
v = 4 quarters	1.00	0.58	1.00	0.68
v = 8 quarters	1.01	0.79	1.00	0.26
Projection on Mean (PM)				
PM	1.05	0.66	1.91	0.95
PM ($w_{0,h} = 0$)	1.02	0.61	1.75	0.84
Principal Components (PC)				
PC (p = 1)	1.42	0.87	1.00	0.50
PC (p = 2)	1.39	0.86	1.02	0.53
PC (p = 3)	1.79	0.96	1.03	0.55
Least Squares (LS)				
LS ($c = 2, w_{0,h} = 0$)	0.95	0.39	1.58	0.81
LS ($c = 2$)	0.90	0.26	5.80	0.96
LS ($c = 2, \ w_{0,h} = 0, \sum w_{i,h} = 1.0$)	0.89	0.06	1.09	0.72
LS ($c = 2, \ w_{0,h} = 0, \ 0 \leq w_{i,h} \leq 1.0$)	0.95	0.10	1.76	0.87
LS ($c = 3, w_{0,h} = 0$)	1.33	0.85	1.38	0.75
LS ($c = 3$)	1.47	0.99	3.25	0.96
LS ($c = 3, \ w_{0,h} = 0, \ \sum w_{i,h} = 1.0$)	1.30	0.95	1.05	0.68
LS ($c = 3, \ w_{0,h} = 0, 0 \leq w_{i,h} \leq 1.0$)	0.97	0.13	2.47	0.99
Shrinkage Weights (SW)				
SW ($w_{0,h} = 0, c = 2, \kappa = 4$)	0.98	0.40	1.25	0.86
SW ($w_{0,h} = 0, c = 3, \kappa = 4$)	1.18	0.97	1.03	0.71
SW ($w_{0,h} = 0, c = 2, \kappa = 6$)	1.02	0.62	1.11	0.87
SW ($w_{0,h} = 0, c = 3, \kappa = 6$)	1.08	1.00	0.98	0.04
SW ($c = 2, \kappa = 4$)	0.95	0.21	1.33	0.90
SW ($c = 3, \kappa = 4$)	1.16	0.98	1.03	0.63
SW ($c = 2, \kappa = 6$)	1.00	0.45	1.01	0.54
SW ($c = 3, \kappa = 6$)	1.07	0.99	1.02	0.78

Note: δ denotes the discount factor applied to past forecast errors; v the length of window used for rolling performance weighting; p the number of principal components; $w_{0,h}$ and $w_{i,h}$ the constant and slope parameters in the forecast regressions; c the number of clusters; κ denotes the degree of shrinkages.

Table 3: Evaluation of alternative SPF Inflation combinations: 2004:Q1 – 2008:Q3

	H = 1 Year Ahead		H = 2 Years Ahead	
	MSE	P-Value	MSE	P-Value
Benchmarks				
Equal Weighted SPF	1.00	0.50	1.00	0.50
Random Walk	0.76	0.00	0.78	0.01
NAÏVE	1.02	0.56	0.74	0.05
AR(1)	0.93	0.16	1.12	0.85
Recent Best				
v = 1 quarter	0.82	0.08	1.08	0.84
v = 4 quarters	0.85	0.02	1.03	0.89
Trimmed means				
Symmetric Trim (5%)	1.03	0.99	0.99	0.11
Median (50%)	1.02	0.96	1.01	0.83
Recursive Performance (RP)				
$\delta = 1.0$	0.98	0.00	0.96	0.01
$\delta = 0.95$	0.98	0.00	0.96	0.01
$\delta = 0.85$	0.98	0.01	0.95	0.01
Rolling Performance (RP)				
v = 1 quarter	1.04	0.85	0.98	0.27
v = 4 quarters	0.99	0.13	0.95	0.00
v = 8 quarters	0.97	0.02	0.95	0.00
Projection on Mean (PM)				
PM	0.71	0.04	0.78	0.01
PM ($w_{0,h} = 0$)	0.61	0.10	0.56	0.04
Principal Components (PC)				
PC (p = 1)	0.68	0.02	0.67	0.02
PC (p = 2)	0.67	0.01	0.67	0.02
PC (p = 3)	0.66	0.01	0.66	0.02
Least Squares (LS)				
LS ($c = 2, w_{0,h} = 0$)	0.65	0.10	0.62	0.05
LS ($c = 2$)	0.74	0.04	0.80	0.00
LS ($c = 2, \ w_{0,h} = 0, \sum w_{i,h} = 1.0$)	0.92	0.16	0.93	0.15
LS ($c = 2, \ w_{0,h} = 0, \ 0 \leq w_{i,h} \leq 1.0$)	0.62	0.10	0.89	0.40
LS ($c = 3, w_{0,h} = 0$)	0.54	0.09	0.61	0.04
LS ($c = 3$)	0.67	0.04	0.76	0.01
LS ($c = 3, \ w_{0,h} = 0, \ \sum w_{i,h} = 1.0$)	0.97	0.36	0.90	0.11
LS ($c = 3, \ w_{0,h} = 0, 0 \leq w_{i,h} \leq 1.0$)	0.61	0.09	0.67	0.09
Shrinkage Weights (SW)				
SW ($w_{0,h} = 0, c = 2, \kappa = 4$)	0.71	0.05	0.78	0.07
SW ($w_{0,h} = 0, c = 3, \kappa = 4$)	0.73	0.08	0.88	0.09
SW ($w_{0,h} = 0, c = 2, \kappa = 6$)	0.80	0.08	0.87	0.08
SW ($w_{0,h} = 0, c = 3, \kappa = 6$)	0.86	0.11	0.99	0.07
SW ($c = 2, \kappa = 4$)	0.80	0.02	0.91	0.04
SW ($c = 3, \kappa = 4$)	0.84	0.07	0.94	0.07
SW ($c = 2, \kappa = 6$)	0.88	0.05	0.95	0.07
SW ($c = 3, \kappa = 6$)	0.93	0.13	0.99	0.04

Note: δ denotes the discount factor applied to past forecast errors; v the length of window used for rolling performance weighting; p the number of principal components; $w_{0,h}$ and $w_{i,h}$ the constant and slope parameters in the forecast regressions; c the number of clusters; κ denotes the degree of shrinkages.

Table 4: Evaluation of alternative SPF Unemployment combinations: 2004:Q1 – 2008:Q3

	H = 1 Year Ahead		H = 2 Years Ahead	
	MSE	P-value	MSE	P-value
Benchmarks				
Equal Weighted SPF	1.00	0.50	1.00	0.50
NAÏVE	2.82	0.96	2.92	0.99
AR(1)	2.58	0.96	2.67	0.99
Recent Best				
v = 1 quarter	1.10	0.84	1.18	0.96
v = 4 quarters	1.25	0.86	1.40	0.93
Trimmed means				
Symmetric Trim (5%)	1.04	0.87	1.12	0.99
Median (50%)	1.06	0.96	1.07	0.98
Recursive Performance (RP)				
$\delta = 1.0$	1.01	0.81	1.02	0.92
$\delta = 0.95$	1.01	0.88	1.02	0.95
$\delta = 0.85$	1.03	0.92	1.05	0.97
Rolling Performance (RP)				
v = 1 quarter	0.97	0.39	1.26	1.00
v = 4 quarters	1.07	0.91	1.07	0.96
v = 8 quarters	1.05	0.88	1.04	0.96
Projection on Mean (PM)				
PM	2.84	0.98	2.64	0.95
PM ($w_{0,h} = 0$)	1.69	0.81	1.05	0.59
Principal Components (PC)				
PC (p = 1)	2.89	0.97	2.36	0.95
PC (p = 2)	2.87	0.97	2.40	0.94
PC (p = 3)	2.92	0.97	2.59	0.96
Least Squares (LS)				
LS ($c = 2, w_{0,h} = 0$)	1.64	0.79	1.94	1.00
LS ($c = 2$)	2.14	0.91	2.72	0.96
LS ($c = 2,\ w_{0,h} = 0, \sum w_{i,h} = 1.0$)	0.99	0.47	1.95	1.00
LS ($c = 2,\ w_{0,h} = 0,\ 0 \leq w_{i,h} \leq 1.0$)	1.66	0.80	1.06	0.96
LS ($c = 3, w_{0,h} = 0$)	1.60	0.78	2.48	0.99
LS ($c = 3$)	2.19	0.93	3.05	0.97
LS ($c = 3,\ w_{0,h} = 0, \sum w_{i,h} = 1.0$)	0.96	0.42	2.08	0.99
LS ($c = 3,\ w_{0,h} = 0, 0 \leq w_{i,h} \leq 1.0$)	1.66	0.80	0.99	0.00
Shrinkage Weights (SW)				
SW ($w_{0,h} = 0, c = 2, \kappa = 4$)	0.81	0.23	1.28	0.99
SW ($w_{0,h} = 0, c = 3, \kappa = 4$)	0.76	0.06	1.07	0.94
SW ($w_{0,h} = 0, c = 2, \kappa = 6$)	0.77	0.07	1.04	0.91
SW ($w_{0,h} = 0, c = 3, \kappa = 6$)	0.95	0.21	0.96	0.06
SW ($c = 2, \kappa = 4$)	1.98	0.89	1.60	0.93
SW ($c = 3, \kappa = 4$)	1.45	0.81	1.26	0.91
SW ($c = 2, \kappa = 6$)	1.55	0.83	1.24	0.89
SW ($c = 3, \kappa = 6$)	1.10	0.69	0.96	0.06

Note: **δ** denotes the discount factor applied to past forecast errors; ***v*** the length of window used for rolling performance weighting; ***m*** the number of principal components; $w_{0,h}$ and $w_{i,h}$ the constant and slope parameters in the forecast regressions; ***c*** the number of clusters**; κ** denotes the degree of shrinkages**.**

Table 5a): White "Reality Check"- Sample excluding the financial crisis (2004:Q1 – 2008:Q3)

Variable	Best Model	MSE	DM P-Value	White P-Value
GDP (H=1)	OW ($c = 2$, $w_{0,h} = 0, \sum w_{i,h} = 1.0$)	0.89	0.06	0.68
GDP (H=2)	RP (v = 1 quarter)	0.97	0.01	0.79
INF (H=1)	OW ($c = 3$, $w_{0,h} = 0$)	0.54	0.09	0.05
INF (H=2)	PM ($w_{0,h} = 0$)	0.56	0.04	0.25
Unemployment (H=1)	SW ($w_{0,h} = 0$, $c = 3$, $\kappa = 4$)	0.76	0.06	0.93
Unemployment (H=2)	SW ($c = 3$, $\kappa = 6$)	0.96	0.06	0.71

Table 5b): White "Reality Check" – Sample including the financial crisis (2004:Q1 – 2009:Q3)

Variable	Best Model	MSE	DM P-Value	White P-Value
GDP (H=1)	Recent Best ($v = 4$)	0.80	0.15	0.54
GDP (H=2)	PC (p = 3)	0.89	0.13	0.93
INF (H=1)	Recent Best ($v = 4$)	0.87	0.02	0.67
INF (H=2)	OW ($c = 3$)	0.91	0.04	0.33
Unemployment (H=1)	OW ($c = 2$)	0.66	0.31	0.96
Unemployment (H=2)	SW ($c = 2$, $\kappa = 6$)	0.56	0.15	0.45

Table 6: Evaluation based on "meta" selection of combination with best historical performance

	2004:1-2008:3		2004:1-2009:3	
	MSE	P-value	MSE	P-value
GDP (H=1)	1.06	0.81	1.01	0.60
GDP (H=2)	0.91	0.17	0.99	0.34
INF (H=1)	0.68	0.01	0.95	0.36
INF (H=2)	0.91	0.03	1.15	0.77
Unemployment (H=1)	2.06	0.98	1.39	0.98
Unemployment (H=2)	1.22	0.98	1.07	0.88