University Library

MINERVA ACCESS

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Aristidou, C;Lee, K;Shields, K

Title:
Evaluating the use of realtime data in forecasting output levels and recessionary events in the USA

Date:
2019-01-01

Citation:
Aristidou, C., Lee, K. & Shields, K. (2019). Evaluating the use of realtime data in forecasting output levels and recessionary events in the USA. Journal of the Royal Statistical Society Series A: Statistics in Society, 182 (1), pp.131-163. https://doi.org/10.1111/rssa.12366.

Persistent Link:
https://hdl.handle.net/11343/227025

Evaluating the Use of Real-Time Data in Forecasting Output Levels and Recessionary Events in the US*

by

Chrystalleni Aristidou,† Kevin Lee† and Kalvinder Shields††

Abstract

The paper proposes a modelling framework and evaluation procedure to judge the usefulness of real-time datasets incorporating past data vintages and survey expectations in forecasting. The analysis is based on 'meta models' obtained using model-averaging techniques and judged by various statistical and economic criteria, including a novel criterion based on a fair bet. Analysing US output data over 1968q4-2015q1, we find both elements of the real-time data are useful with their contributions varying over time. Revisions data are particularly valuable for point and density forecasts of growth but survey expectations are important in forecasting rare recessionary events.

Keywords: Real-Time Data, Revision, Survey, Forecasting, Model Averaging, Output, Recession.

JEL Classification: E52, E58

*†University of Nottingham, UK; ††University of Melbourne, Australia. Version dated January 2018. We have received helpful comments from the Referees and Associate Editor and from participants at the Computational Economics and Finance Conference, Canada, the RBNZ Nowcasting and Model Combination Workshop, a Philadelphia Fed Workshop on Real Time Data Analysis, the Conference of the Centre for International Research on Economic Tendency Surveys, Denmark, and a Reserve Bank of Australia seminar. We are grateful for funding received from the Australian Research Council (DP0988112). Corresponding author: Kalvinder K. Shields, Department of Economics, University of Melbourne, Victoria, 3010, Australia. E-mail: k.shields@unimelb.edu.au, tel: 00 613-83445300, fax: 00 613-83446899.

1 Introduction

Real-time datasets are now widely available covering macroeconomic variables for many countries. The datasets contain the available history of data vintages, showing the preliminary estimates of variables published at the earliest opportunity alongside their subsequent revision as more complete information becomes available. The datasets also often include direct measures of expectations as expressed in surveys published at the time, describing agents' beliefs on expected future values of the macroeconomic variables and the expected contemporaneous values when the first-release data are published with a delay. A substantial literature has now grown developing the methods required for the analysis of real-time datasets and their use in prescribing and evaluating policy; see, for example, the review in Croushore (2011).

One area in which real-time data are potentially important is in forecasting since the data provide a comprehensive description of the context in which forecasts and subsequent decisions are made. However, while real-time data are often employed in forecasting, there remains some scepticism about their usefulness and they do not figure in forecasting exercises as systematically as might be expected. For example, of the fifty-four papers concerned with forecasting output published over the three years 2014-2016 in this journal, the *Journal of Business and Economic Statistics*, *Journal of Forecasting*, *International Journal of Forecasting* and *Review of Economics and Statistics*, only six made use of the availability of both successive data vintages and survey data, and twenty-one papers made no use of real time data at all.

The purpose of this paper then is to judge the usefulness of real-time datasets in forecasting through an evaluation exercise that uses revisions data and survey data to forecast quarterly output growth and the occurrence of recessions in the US. The paper contributes to the discussion on the use of real-time data in at least three ways. *First*, it suggests a simple canonical modelling framework that can readily accommodate revisions and survey data alongside the most recent data measures to characterise the underlying data generating process of the variables of interest as well as the expectation formation and measurement processes. *Secondly*, it investigates the usefulness of revisions and survey

data by comparing the forecasting performances of models that make full use of the data with those of models that make only partial use of the information contained in real-time datasets. And *thirdly*, the paper considers various evaluation criteria to judge the usefulness of real-time data in forecasting, drawing a distinction between 'real-time forecast evaluation' and a 'final assessment' of forecast performance. The usefulness of real time data is judged according to statistical criteria, based on models' point forecasts and density forecasts, and according to economic criteria. In the latter case, we focus on forecasting the likely occurrence of a set of recessionary events and introduce a novel means of evaluating these probability forecasts, based on a fair bet, to investigate the role of real-time data in forecast-based decisions involving relatively rare events.

The three aspects of our modelling exercise are motivated by different strands of the literature. The early paper by Mankiw, Runkle and Shapiro (1984) was influential in generating scepticism over the use of revisions data, concluding that revisions are mainly 'news' (i.e. have no predictable content),[1] while Croushore (2010) notes that the inefficiencies and biases in expectational errors in surveys, as found in early studies, generated a long-lasting scepticism of the value of survey data too.[2] The first challenge for the paper then is to set out a modelling framework that can accommodate revisions data and survey data coherently alongside the first-release data. This can be used to build a picture of the information available to individuals at each time, and how it is used, to establish whether the scepticism found in some parts of the literature is warranted.

The second aspect of the paper focuses on the usefulness of revisions and survey data in forecasting and relates to the use of information when there are many potential predictor variables, as discussed in Clements and Hendry (2005) and Stock and Watson (2006) for

[1]Scepticism is found in Croushore and Stark (2003), Croushore (2006) and Koenig et al (2003). On the other hand, Patterson (2002), Arouba (2008), Garratt et al. (2008), Clements and Galvao (2010), Jacobs and van Norden (2011) and Kishor and Koenig (2012) all argue that revisions contain useful information.

[2]Croushore's own results, and those of Ang et al. (2007) and Aretz and Peel (2010) for example, show that survey expectations are often hard to beat in real time forecasting exercises. Similarly, Frale et al. (2010) and Banbura and Runstler (2011) show that survey data are useful in nowcasts from mixed-frequency models and Matheson et al (2010) show that survey data are useful in predicting actual series and their revisions.

example. This literature recognises that, with the samples of data typically available, parameter estimation error can dominate model's forecast performance. This means, for example, that adding a variable to a forecasting model can undermine its forecasting performance even if the variable is part of the true data generating process. One way to mitigate against this problem is to average across forecasts from different models (see, for example, Harvey and Newbold (2005) and Timmermann (2006) for discussion). This is the approach taken here, producing forecasts using various 'meta' models each constructed using model averaging techniques. The meta models are distinguished according to their use of the real time data (making use of vintage data only, survey data only or both). The averaging allows for time-varying weights and ensures that each meta model makes best use of the information available to it in forecasting. Comparison of the forecasts across the meta models then provides an assessment of the contributions of the different types of real-time data.

The third aspect of the paper relates to the ambiguity on the criteria to be used in forecast evaluation. This partly arises from an increasing awareness of the importance of properly characterising forecast uncertainties which has shifted attention from point forecasts to density forecasts, and evaluation criteria from models' root mean squared errors (RMSEs) to their probability integral transforms (PITs) and logarithmic scores; see, for example, the June 2010 Special Issue of *Journal of Applied Econometrics* for an overview. But there is also increasing interest in judging the economic value of a model's forecast, concentrating on the usefulness of the models in a specific decision-making context rather than on its statistical performance, as discussed in Granger and Machina (2006) for example. Certainly economic and statistical evaluation criteria highlight different features of the models and their forecast performance and so, in this paper, we judge the usefulness of real time data not just in terms of their use in generating point and density forecasts of output growth but also their role in forecasting the probability of relatively rare/extreme recessionary events.

The layout of the remainder of the paper is as follows. Section 2 outlines the methods employed in the paper, introducing our modelling framework and defining and explaining the construction of the meta models. Section 3 sets out the statistical and economic

criteria used in our forecast evaluation exercise, including a description of the evaluation based on a fair bet. Section 4 applies the methods to US data over $1968q4 - 2015q1$, including all the data vintages available for actual output and the expected output data from surveys over the period. As it turns out, we find that both elements of the real-time data, from data vintages and from surveys, are useful in forecasting, judged by statistical and economic criteria, with the contribution of the different elements varying over time. Revisions data are particularly valuable in producing point and density forecasts for output growth but the direct measures of expectations taken from surveys play an important role in forecasting rare recessionary events. Section 5 concludes.

2 A Modelling Framework to Accommodate Real-Time Information

2.1 The Basic VAR Model

Our interest in real time datasets revolves around the distinction between the actual and expected value of a variable measured at different times and so it is important to be clear about notation and terminology from the outset.[3] In what follows, ${}_t y_{t-s}$ is the measure of the (logarithm of the) variable y at time $t-s$ as released at time t , while ${}_t y^e_{t+s}$ is a direct measure of the expected value of the variable at $t+s$, with the expectation formed on the basis of information available at the time the measure is released, t. Throughout, we shall assume that data is published with a one period delay, and the time-t vintage of data is denoted $Y_t = \{{}_t y_1,\ {}_t y_2,\ ...,\ {}_t y_{t-2},\ {}_t y_{t-1},\ {}_t y^e_t,\ {}_t y^e_{t+1},...,\ {}_t y^e_{t+F}\}$ which includes the time-t measures of the actual variables at $t = 1, ..., t-1$ and the time-t measures of expected contemporaneous and future values of the variables published for up to F periods ahead. In our real-time forecast evaluation exercise, we denote the period in which decisions are made by τ for $\tau \leq T$ and Y_τ is termed 'the most recent data vintage' while Y_T is the 'final data vintage'. The information arriving between τ and $\tau + h$ is denoted, $\mathbf{Y}_{\tau,\tau+h} = \{Y_\tau,\ ..,\ Y_{\tau+h}\}$.

For the modelling exercise, we assume that revisions continue for no longer than R

[3] The modelling framework can be readily extended to accommodate data on revisions and surveys on more than one variable.

periods after the first-release (so that the true value of y_t is measured by its post revision measure ${}_{t+R}y_t$) and that the true value of the variable is difference-stationary.[4] If the surveys provide measures of the expected values of y for up to F periods ahead then, in making a decision at time τ, we can use a model that explains the following $n = 1 + R + F + 1$ easy-to-interpret series each published in time t:

$$
\begin{aligned}
g_t^y &= {}_t y_{t-1} - {}_{t-1}y_{t-2} : \text{growth in } t-1 \text{ as described by the first-release data;} \\
m_{r,t}^y &= {}_t y_{t-1-r} - {}_{t-1}y_{t-1-r} : r^{th} \text{ revision of } y_{t-1-r} \text{ updating previous measures, } r = 1, ..R; \\
e_{f,t}^y &= {}_t y_{t+f}^e - {}_t y_{t+f-1}^e : \text{expected contemporaneous and future growth of } y_{t+f},\ f = 0, ..F.
\end{aligned}
$$

The variable g_t^y is a linear combination of growth in the true value of y and revisions. If the true value of y is difference stationary and revisions $m_{r,t}^y$ are stationary, then g_t^y is stationary. Similarly $e_{f,t}^y$ is a linear combination of growth in the true value of y and expectational error. So if the true value of y is difference stationary and expectational errors are stationary, then $e_{f,t}^y$ is stationary. Stationarity in revisions is reasonable if the data reflects measurement error and abstracts from the effects of definitional or 'benchmark' changes. Stationary expectational errors mean that these errors cannot grow without bounds and are consistent with a wide range of hypotheses on expectation formation including, for example, the Rational Expectations Hypothesis.

The three types of measure of y (first-release, revised and expected) reflect the fact that three interrelated processes occur here: (i) 'behavioural' economic decisions are made by economic agents to determine the actual values of the variables at each time; (ii) surveys are published reporting the expectations formed on the variables by those same economic agents; and (iii) the economic outcomes are measured reflecting the data collection and survey practices of the statistical agencies. These processes occur simultaneously and in real time. For (i) and (ii), economic theory provides innumerable examples of intertemporal decision-making in which actual and expected future economic outcomes are determined simultaneously, driven by exogenous factors and influenced by lags through inertia and rigidities.[5] And for (iii), official statistics aim to quantify the outcomes as quickly

[4]For the (log of the) output level, for example, the assumption of difference stationarity is relatively uncontentious.

[5]The survey expectations provide direct insights on the expectation formation process although mea-

as possible but measures are revised as new source data becomes available - including information from late respondents and the replacement of judgemental or estimated inputs with firm data - and as errors in source data and computations are corrected. This means the extent of the revisions is likely to be determined endogenously and to be systematically related to the economic outcomes themselves.[6]

Any structural model capturing the contemporaneous and lagged interactions between g_t^y, $m_{r,t}^y$ and $e_{f,t}^y$ can be re-written as a simple reduced form p-order vector autoregressive model that explains these stationary series for $t = 1, ..., \tau$:

$$g_t^y = \alpha_{10} + \sum_{i=1}^{p} \left[\alpha_{11i}\ g_{t-i}^y + \sum_{r=1}^{R} \alpha_{12ri}\ m_{r,t-i}^y + \sum_{f=1}^{F} \alpha_{13fi}\ e_{f,t-i}^y \right] + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

$$m_{r,t}^y = \alpha_{20} + \sum_{i=1}^{p} \left[\alpha_{21i}\ g_{t-i}^y + \sum_{r=1}^{R} \alpha_{22ri}\ m_{r,t-i}^y + \sum_{f=1}^{F} \alpha_{23fi}\ e_{f,t-i}^y \right] + \boldsymbol{\varepsilon}_{rt}, , \quad r = 1, .., R, \tag{2.2}$$

$$e_{f,t}^y = \alpha_{30} + \sum_{i=1}^{p} \left[\alpha_{31i}\ g_{t-i}^y + \sum_{r=1}^{R} \alpha_{32ri}\ m_{r,t-i}^y + \sum_{f=1}^{F} \alpha_{33fi}\ e_{f,t-i}^y \right] + \boldsymbol{\varepsilon}_{ft}, , \quad f = 0, .., F, \tag{2.3}$$

where the α's are coefficients and $\boldsymbol{\varepsilon}$'s are vectors of mean-zero shocks. We denote this model by $M_{R,F,\tau}$ in what follows, with the subscript 'R, F, τ' highlighting that the estimated model will differ depending on the maximum number of revisions, the forecast horizon in the survey and on the estimation period involved.[7]

Noting that the variables g_t^y, $m_{r,t}^y$, and $e_{f,t}^y$ involve the y variables measured at t and earlier, the equations in (2.1)-(2.3) can be stacked and transformed to obtain the $(p+1)^{th}$-order autoregressive model

$$\mathbf{z}_t = \mathbf{A}_0 + \sum_{i=1}^{p+1} \mathbf{A}_{1i}\ \mathbf{z}_{t-1} + \boldsymbol{\varepsilon}_t, \qquad t = 1, ..., \tau \tag{2.4}$$

surement issues can arise when real-time data are involved because there may be ambiguity on whether respondents are predicting the actual values or the measured first-releases.

[6]See Jacobs and van Norden (2011) and Kishor and Koenig (2012) for further discusson of the sources of revision error in published data.

[7]The expectations series ${}_t y_{t+f}^e$ typically reflect the average of the survey respondents' expectations. A measure of the variation across respondents' expectations provides a useful direct measure of the uncertainty associated with these expectations which could be included in an extended version of the model at (2.1)-(2.3). This idea is pursued in Garratt et al. (2018).

where $\mathbf{z}_t = ({}_t y_{t-1},\ {}_t y_{t-2}, ...,{}_t\, y_{t-R+1},\ {}_t y^e_t, ...,\ {}_t y^e_{t+F})'$ is the $n \times 1$ vector of data published in time t, $\mathbf{A}_0$ is a $n \times 1$ vector of parameters derived from the α's and the $\mathbf{A}_{1i}$ are $n \times n$ matrices similarly derived. Note that only $g^y_t = {}_t y_{t-1} -\ {}_{t-1} y_{t-2} = \Delta\ {}_t y_{t-1}$ involves the difference operator Δ; $m^y_{r,t}$, and $e^y_{f,t}$ are 'quasi-differences' involving, for $m^y_{r,t}$, the variable dated in the same quarter but measured at different times and, for $e^y_{f,t}$, the variable dated at different quarters but measured at the same time. If (2.1)-(2.3) or (2.4) are re-written to explain $\Delta\mathbf{z}_t$, the model takes the form of a cointegrating VAR in which the parameters reflect the assumptions that revisions and expectational errors are stationary.[8] These parameter restrictions, whether transferred to the cointegrating VAR form or to the stacked form in (2.4), ensure that the various measures of the variables in y converge to the same values in the long run. While the form in (2.1)-(2.3) is the most natural one for estimation purposes, (2.4) is more convenient for describing simulation and forecasting exercises.

It is worth noting that the above transformation incorporates the assumption that the data is revised no more than R times. Taken literally, this means that the most recent vintage of data Y_τ describes the post-revision series for observations dated at $\tau - R$ and earlier (i.e. ${}_t y_{\tau-s} = {}_\tau y_{\tau-s},\ t = 1, .., \tau,\ s = R+1, ...$).[9] The model at (2.4) therefore encompasses the '*conventional model*' that would be estimated in real time based only on the most recent vintage of data. This latter model can be denoted $M_{0,-1,\tau}$ since it does not use either past vintages of data or the expectations of even the contemporaneous value of y as provided in surveys.

[8] In this form, we find there are R cointegrating relations between ${}_t y_{t-1}$ and each of the ${}_t y_{t-1-R}$, and $F+1$ cointegrating relations between ${}_t y_{t-1}$ and each of the ${}_t y^e_{t+F}$, all of the form $(1,\ -1)$. See Garratt et al. (2008) for details.

[9] The assumption that there are literally no more than R revisions means the $(R+1)^{th}$ diagonal in a standard 'data triangle' is equal to (and can be overwritten by) the observations in the most-recent vintage. The model in (2.4) then explains the observations on the lowest $(R+1)^{th}$ diagonals (plus the survey measures). In what follows, we call the $(R+1)^{th}$ observation the 'post-revision' observation although, in practice, small, unsystematic revisions may continue indefinitely.

2.2 Meta Models and Forecast Combinations

The VAR model of (2.1)-(2.4) provides a simple framework within which all the real-time data available can be accommodated in a coherent way. But such a model could be very highly parameterised, depending on the number of revisions available, the length of the survey horizon and the chosen order of the VAR, and this could undermine its value as a forecasting tool. Similarly, it is possible that different parts of the real-time data become more or less useful for forecasting at different times. For example, statistical agencies' procedures could mean that measurement errors contained in the first-release of data are more pronounced in times of very high or very low growth, making revisions data more useful. Or forecasters may watch incoming news more carefully at times of crisis so that survey data becomes more informative at these times.[10]

To mitigate against these problems, we estimate a set of models of the form in (2.1)-(2.4) and combine these using model averaging techniques into a 'meta' model. The weights used to combine the models, and their associated forecasts, can be chosen so that forecast performance is maximised and can change over time so that different parts of the real-time data can be used when they are helpful. Of course, the approach exploits the fact that forecast performance is typically improved through forecast combinations, as established by Harvey and Newbold (2005).

The meta model is obtained noting that forecasts at τ can be based on model $M_{R,F,\tau}$ or indeed any model $M_{r,f,\tau}$ for $r = 0, .., R$ and $f = -1, .., F$, or a weighted average of these models. We choose weights that reflect the relative forecasting performance over the recent past, where performance is judged according to the question of interest. So here, the weights could be based on the models' $(R+1)$-period-ahead forecasts of the (post-revision) measure of output ${}_{\tau+R+1}\, y_\tau$ if the 'true' output level is the variable of interest, or the weights could be based on the models' one-period-ahead forecast of the first-release measure ${}_{\tau+1}y_\tau$ if forecasts of this variable inform decisions. Given that some of the models with large R and F could be very heavily parameterised, the weights should also reflect the outcome of any shrinkage procedure employed to improve efficiency in estimation.

[10] See, Loungani et al. (2013), for example, for discussion on the changing impact of information rigidities on survey data at different points of the business cycle.

More specifically, if interest focuses on the one-step-ahead forecast of the first-release measure for example, the meta model that makes full use of the real-time data is obtained as follows:

- Split the currently available sample into two sub-samples: an estimation period $t = 1, ..., \tau - \Lambda$; and a 'training period' $t = \tau - \Lambda + 1, .., \tau$. Estimate model $M_{r,f,\tau-\Lambda}$ defined in (2.1)-(2.3) over $t = 1, ..., \tau - \Lambda$ and for $r = 0, .., R$ and $f = -1, .., F$, providing $(R + 1) \times (F + 2)$ alternative models;

- Undertake a specification search procedure for each model $M_{r,f,\tau-\Lambda}$ in which variables are dropped sequentially when the (absolute value of the) t-ratio on a coefficient is less than unity.[11] If all the lags of a long-horizon revision variable are dropped during the specification search, the model is eliminated from consideration as a candidate forecasting model (in favour of the smaller $M_{r-1,f,\tau-\Lambda}$ model). Similarly for long-horizon survey variables;

- Evaluate the forecast performance of all the individual models that remain with a chosen statistical criterion. If the question of interest relates to the point forecasts of the one-step-ahead forecast of the first-release measure, then the squared forecast error might be used: $s^p_{r,f,\tau-\Lambda} = ({}_{\tau-\Lambda+1}y_{\tau-\Lambda} - E\left[{}_{\tau-\Lambda+1}y_{\tau-\Lambda} \mid M_{r,f,\tau-\Lambda}\right])^2$. Or, if the whole density forecast is of interest, the log score can be used: $s^d_{r,f,\tau-\Lambda} = \ln({}_{g(\tau-\Lambda+1}y_{\tau-\Lambda} \mid M_{r,f,\tau-\Lambda})$ where $g({}_{\tau-\Lambda+1}y_{\tau-\Lambda} \mid M_{r,f,\tau-\Lambda})$ is the nowcast density for model $M_{r,f,\tau-\Lambda}$.

- Repeat this exercise for samples over the whole of the training period and calculate weights for the models, $w_{r,f,\tau}$ on the basis of the relative forecast performance of the individual models over the training period.[12] For example, forecast performance

[11] The unit threshold is chosen arbitrarily although Clements and Hendry (2005) note that, in the context of a univariate AR(1) model, this is the threshold below which the inclusion of a variable damages the RMSE of a one-step-ahead forecast, on average, even if the variable is part of the true data-generating process.

[12] A variety of weighting schemes have been proposed in the model averaging literature. Elliott et al. (2013) reviews some of these, noting that the appropriateness of a weighting scheme depends on various trade-offs including that between omitted variable bias and parameter estimation error.

might be judged by the mean of the squared forecast error $MSE_{r,f,\tau} = \frac{1}{\Lambda}\sum_{\lambda} s^{p}_{r,f,t-\lambda}$ if point forecasts are of interest, obtaining weights as follows:

$$w_{r,f,\tau} = \frac{(\sqrt{MSE_{r,f,\tau}})^{-1}}{\sum_r \sum_f (\sqrt{MSE_{r,f,\tau}})^{-1}}. \tag{2.5}$$

Alternatively, if density forecasts are of interest, performance might be judged by the average of the log scores, $MLS_{r,f,\tau} = \frac{1}{\Lambda}\sum_{\lambda} s^{d}_{r,f,t-\lambda}$, with weights given by

$$w_{r,f,\tau} = \exp(MLS_{r,f,\tau}) / \sum_r \sum_f \exp(MLS_{r,f,\tau}). \tag{2.6}$$

The 'meta model' that makes full use of the real-time data over the period $t = 1, ..., \tau$ consists of the individual estimated models and their weights and it is denoted by

$$\overline{M}_{R,F,\tau} = \{M_{r,f,\tau},\ w_{r,f,\tau} \quad \text{for } r = 0, .., R \quad \text{and} \quad f = -1, .., F\} \tag{2.7}$$

The meta model can be used to obtain point forecasts and density forecasts using the weighted average of the models' individual point forecasts and aggregating over the models' individual densities.[13]

3 Assessing Models' Forecasting Performance and Usefulness in Decision-Making

3.1 Real-Time and Final Forecast Assessment

A judgement of the usefulness of real-time data in forecasting and decision-making can be based on two complementary elements: a real-time assessment and a final assessment. The first of these elements is based on the weights found in the 'meta model' described above since these provide a straightforward summary of the usefulness of the revisions and survey data as it would be judged in real time. Specifically, the nature of the meta model $\overline{M}_{R,F,\tau}$ obtained at time τ can be summarised by the statistics

$$\mu^{r}_{\tau} = \sum_{r=0}^{R} r \times w_{r,f,\tau} \quad \text{and} \quad \mu^{f}_{\tau} = \sum_{f=-1}^{F} f \times w_{r,f,\tau} \quad , \tag{3.8}$$

[13] If interest focuses on the $(R+1)$-step-ahead forecast of the post-revision measure, say, the meta model is as at (2.7) but with weights based on the corresponding statistical criterion: e.g. $s^{p}_{r,f,\tau-\Lambda} = ({}_{\tau-\Lambda+1}y_{\tau-\Lambda-R} - E\left[{}_{\tau-\Lambda+1}y_{\tau-\Lambda-R} \mid M_{r,f,\tau-\Lambda-R}\right])^2$.

showing the weighted average of the models' revision length and forecast horizon. The μ_τ^r and μ_τ^f statistics capture the relative importance of the revision data and the survey data in defining the meta model at time τ.[14] If they deviate from 0 and -1 respectively, they show that the revisions data and the survey data would have made a contribution to an out-of-sample forecasting exercise if it had been conducted in real time. This provides a *real-time assessment* of the usefulness of the revisions and survey data in $\overline{M}_{R,F,\tau}$ therefore. Different values for the weights could be obtained for different τ, allowing the possibility that the usefulness of the revision and survey data changes over time.

The meta model $\overline{M}_{R,F,\tau}$ can be used to provide forecasts for any decision-date τ. If this exercise is repeated over the whole evaluation period ($\tau = \underline{\tau}, ..., T$, say), then the forecast criterion (squared error or log score, for example) can be calculated at each observation to obtain a measure of the overall performance of the models which, at least in principle make full use of R revisions and F survey forecasts. A *final assessment* of the usefulness of the revisions and survey data can be based on the performance of the meta model $\overline{M}_{R,F,\tau}$ over $\tau = \underline{\tau}, ..., T$ compared to that of alternative meta models in which

- no use is made of the survey data throughout (i.e. based on the meta model $\overline{M}_{R,-1,\tau}$ for $\tau = \underline{\tau}, ..., T$);
- no use is made of the revisions data throughout (i.e. based on the meta model $\overline{M}_{0,F,\tau}$ for $\tau = \underline{\tau}, ..., T$); and
- no use is made of either the revisions or survey data throughout (i.e. based on the conventional real time model $\overline{M}_{0,-1,\tau}$ for $\tau = \underline{\tau}, ..., T$).

These three models are nested within $\overline{M}_{R,F,\tau}$ and, in principle, could be chosen if zero weights are placed on the models involving revisions or surveys at all times when deriving $\overline{M}_{R,F,\tau}$. In practice, zero weights might be unlikely and so comparison of the forecast criteria obtained from the four models provides an overall assessment of the usefulness of the revisions and survey data taking into account that they might be more or less useful at different times.

[14] The statistics automatically capture the effect of the shrinkage specification search since zero weight is given to models in which all lags of a particular revision horizon or forecast horizon are dropped.

3.2 Event probability forecasts and economic evaluation criteria

The discussion above focuses on statistical criteria for judging models' point and density forecasts. However, recent years have seen a growing interest in a decision-based approach to the evaluation of forecasts with performance judged according to the economic value of forecasts in an explicit decision-making context.[15] The preponderance of studies employing this decision-based approach are in the area of applied finance where investment strategies and their outcomes are relatively straightforward to describe.[16] The decision-making context in macroeconomics is not so straightforward and there is no generally accepted decision-based criterion with which to judge models' forecasts of output fluctuations. However, we believe that a judgement on the usefulness of real time data in output forecasts should include an element that reflects the economic worth of the forecast and, to this end, we also consider models' abilities to forecast the likely occurrence of a recession (suitably defined). Given the interest shown by the media in whether the economy is or is not in recession, it appears that this dichotomous event is important in real-world decisions. The probability of a recession occurring at τ can be forecast using the meta models $\overline{M}_{R,F,\tau}$, $\overline{M}_{0,F,\tau}$, $\overline{M}_{R,-1,\tau}$ and $\overline{M}_{0,-1,\tau}$ and comparison of the probability forecasts again provides an indication of the usefulness of real-time data.

A straightforward statistical evaluation of a model's event probability forecast, π_τ, is obtained through a contingency table analysis. Here, the forecast probability is converted to a prediction on whether the event will happen or not ($r_t = 1$ or 0 respectively) depending on whether the probability is greater or less than 0.5. A two-by-two contingency table then shows the number of occasions a recession is correctly forecast to occur (YY, hits) over the $T - \underline{\tau} + 1$ observations of the evaluation period, when it is incorrectly forecast to occur (YN, false alarms), when it is incorrectly predicted that recession will not occur (misses, NY) and when it is correctly predicted that a recession will not occur (NN). The

[15] This recognises that the statistical criteria used to evaluate models, typically measured using MSE, provide information on the economic value of their forecasts only under certain conditions. See Granger and Pesaran (2000) for an overview of this discussion.

[16] See, for example, Leitch and Tanner (1991), Barberis (2000), Abhyanker et al. (2005) and Garratt and Lee (2010).

performance of the model can be described by the proportion of forecasts that are correct ($\frac{YY+NN}{T-\underline{\tau}+1}$) or the Kuipers score [KS] (a statistic that takes values between -1 and 1 and summarises the degree of correspondence between predictions and outcomes similar to a correlation coefficient).[17] Formal tests can also be applied against the null that there is no relationship between the outcome and the predictions.[18]

A more 'economic' evaluation might be based on an explicit investment scenario in which an investor bets on whether an event occurs or not and the model is judged according to the returns obtained using it. (See Johnstone *et. al.*, 2013, for a related approach). We can define a 'symmetric bet' to be where an investor pays a fixed charge each period to make a bet on whether the event will occur or not and receives a payment if the prediction turns out to be true. Alternatively, the bet could be defined as 'asymmetric' if the bet is made only when the investor believes the event will occur. In either scenario, the profits obtained from decision-making directly measure the economic value of the model over the evaluation period.

To formalise the ideas, and using y_t to denote the logarithm of output from now on, we note that any recessionary event defined at τ as a set of outcomes involving outputs up to h periods ahead can be written as $R(\mathbf{Y}_{1,\tau+h})$. This could be a complicated, possibly non-linear, function of output measures dated, and published in any data vintage, anytime up to $\tau+h$. For example, if recession at τ is defined to occur when the post-revision measure of output falls below its previous peak (which could have been one period ago or many periods earlier), then $R(\mathbf{Y}_{1,\tau+h})$ depends on the entire history of post-revision measures of output and their forecasts up to ${}_{\tau+1+R}y_{\tau}$. The probability that the event occurs is

$$\text{probability of recession} = \int_R \Pr(\mathbf{Y}_{\tau+1,\tau+h} |\ \mathbf{Y}_{1,\tau},\ \overline{M}_{R,F,\tau})\ \ \partial \mathbf{Y}_{\tau+1,\tau+h}, \tag{3.9}$$

where Pr(.) is the joint density forecast of output values observed in data vintages $\tau+1, ...\tau+h$ and (3.9) integrates over all the possible combinations of output outcomes that could define recession. This could be very difficult to evaluate analytically but is read-

[17] The KS focuses attention on the successful prediction of recession while also penalising false alarms ($= \frac{YY}{YY+NY} - \frac{YN}{YN+NN}$; i.e. the hit rate - false alarm rate)

[18] For example, Pesaran and Timmermann [PT] (2009) describe tests of the null that the model's predictions are no better than those achieved based only on the unconditional probability.

ily obtained through simulation. Specifically, abstracting from parameter uncertainty, one can use the estimated parameters and weights of (2.7) at time τ to generate, for example, 10000 replications of the future vintages of data $Y_{\tau+1}$ to $Y_{\tau+h}$. For this, we take each model $M_{r,f,\tau}$ contained in (2.7) in turn and, using random draws from the multivariate Gaussian distribution with the corresponding estimated variance-covariance matrix, generate a proportion of the 10000 replications in line with the associated weights. This set of simulated futures give directly the forecast densities of the first-release, expected and post-revision output series over the horizons up to $\tau + h$ based on $\overline{M}_{R,F,\tau}$, and simply counting the number of times an event occurs in these simulations provides a forecast of the probability that the event will occur based on this meta model. Point, density and probability forecasts associated with the simpler meta models $\overline{M}_{R,-1,\tau}$, $\overline{M}_{0,F,\tau}$ and $\overline{M}_{0,-1,\tau}$ can use the same set of simulations but aggregated in proportion to their different respective weights.[19]

In a decision-making context, where an individual's objective function $\nu(r_\tau, R(\mathbf{Y}_{\tau+1,\tau+h}))$ depends on the outcome of a choice variable r_τ and the occurrence of the recessionary event, the decision-maker's problem can be written as

$$\max_{r_\tau} \left\{ \int \nu(r_\tau,\ R(\mathbf{Y}_{\tau+1,\tau+h})) \Pr\left(\mathbf{Y}_{\tau+1,\tau+h} \mid \mathbf{Y}_{1,\tau},\ \overline{M}_{R,F,\tau}\right) \ d\mathbf{Y}_{\tau+1,\tau+h} \right\}. \tag{3.10}$$

In terms of the simulations, the decision involves simply choosing the value of r_τ that maximises the value of the objective function when averaging across the simulations. We can denote the optimal value of the choice variable chosen using model $\overline{M}_{R,F,\tau}$ by $r_{R,F,\tau}$. Pesaran and Skouras (2000) then suggest measuring the model's performance with the statistic

$$\overline{\Psi}_{R,F,T} = \frac{1}{T - \underline{\tau} + 1} \sum_{\tau=\underline{\tau}}^{T} \nu(r_{R,F,\tau},\ R(\mathbf{Y}_{\tau+1,\tau+h})), \tag{3.11}$$

calculated over the out-of-sample evaluation period $\underline{\tau}, ..., T$ and based around the values of $r_{R,F,\tau}$ chosen using model $\overline{M}_{R,F,\tau}$ in each period. Similar statistics can be calculated

[19] A more detailed discussion of simulation methods, including those that accommodate model uncertainty and parameter uncertainty as well as the stochastic uncertainty considered here, is given in Garratt et al. (2003).

for any other model, with associated optimal choice variable, and these provide the basis of a comparison of the forecast performance of the models on economic grounds.

The payout contingencies relating to the symmetric and asymmetric bets described above are summarised as:

Payout contingencies for outcomes of a symmetric fair bet

	Recession Occurs	
Recession Forecast	*Yes*	*No*
Yes	$s-1$	-1
No	-1	$s-1$

Payout contingencies for outcomes of an asymmetric fair bet

	Recession Occurs	
Recession Forecast	*Yes*	*No*
Yes	$s-1$	-1
No	0	0

The bet can be described as 'fair' if the payout, s, is chosen so that the investor would break even if her bet is based on the unconditional probability, p, that the event occurs. For the symmetric bet, this is where $s = \frac{1}{2p^2-2p+1}$ and it is where $s = \frac{1}{p}$ in the asymmetric case.[20] If the model's forecast probability is π and if the investor bets on recession when π exceeds some critical threshold value π^c, then the defining factor in the decision to bet on recession or not is the choice of the threshold value. In the symmetric case, the investor's expected end-of-forecast-period wealth corresponding to $\nu(r_\tau,\ R(\mathbf{Y}_{\tau+1,\tau+h}))$ in (3.10) is given by

$$E[W_{\tau+h}] = \begin{cases} (s-1)\pi - (1-\pi) & = \frac{\pi}{2p^2-2p+1} - 1 \qquad \text{if } \pi > \pi^c \Leftrightarrow r_{R,F,\tau} = 1 \\ (s-1)(1-\pi) - \pi & = \frac{1-\pi}{2p^2-2p+1} - 1 \qquad \text{if } \pi < \pi^c \Leftrightarrow r_{R,F,\tau} = 0 \end{cases}$$

and maximum expected wealth is achieved by choosing a threshold value of $\pi^c = 0.5$ since $\frac{\pi}{2p^2-2p+1} > \frac{1-\pi}{2p^2-2p+1}$ if $\pi > 0.5$ and vice versa if $\pi < 0.5$. In the asymmetric case, wealth is given by

$$E[W_{\tau+h}] = \begin{cases} (s-1)\pi - (1-\pi) = \frac{\pi}{p} - 1 \qquad \text{if } \pi > \pi^c \Leftrightarrow r_{R,F,\tau} = 1 \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad\quad \text{if } \pi < \pi^c \Leftrightarrow r_{R,F,\tau} = 0 \end{cases}$$

[20] The payout for a correct prediction is largest in a symmetric bet when $p = 0.5$ and increases monotonically as $p \to 0$ in the asymmentric case.

and maximum expected wealth is achieved by choosing a threshold value of $\pi^c = p$ since $\frac{\pi}{p} - 1 > 0$ if $\pi > p$. In both cases, model $\overline{M}_{R,F,\tau}$ can be used to predict the occurrence of a recession or not in each observation through the evaluation period and, depending on the actual outcome, this will generate a sequence of financial returns that can again be used to judge the model as in (3.11). Carrying out the same exercise for models $\overline{M}_{R,-1,\tau}$, $\overline{M}_{0,F,\tau}$ and $\overline{M}_{0,-1,\tau}$ provides a \$ value for each model which are comparable across models and which conveys directly the economic usefulness of each of these models (and of the different elements of the real-time data). Moreover, if the \$ value is expressed as a ratio to the return that would be achieved if a perfect forecaster was betting on the event (under the same conditions and with the same pay-out), this provides a measure that ranges across the interval $[0, 1]$ from entirely uninformed to perfect forecasters and which is comparable across events also.[21]

4 Forecasting Output and Recessions using US Real-Time Data

The empirical work of the paper considers nowcasts of output outcomes and recessionary events based on the first-release and revised measures of output and on the direct measures of output expectations for the U.S. These are obtained from the real-time datasets of the Federal Reserve Bank of Philadelphia available at *www.phil.frb.org/econ/forecast/*.

The officially-released backward-looking data consist of 172 quarterly vintages of data; the first was released in $1965q4$ and the final vintage used in this paper is dated $2015q1$. All vintages include observations dated back to 1947q1. In what follows, the analysis distinguishes between standard 'revisions' and once-and-for-all 'benchmark adjustments' arising from the re-definition or reclassification of a series. The latter are announced in advance and we assume in our work that these are entirely taken into account in forecasting and decision-making. To do this, we adjust the data by splicing the pre- and

[21] Woodcock (1981) shows that the KS can have a similar economic interpretation in some circumstances. Specifically, the KS shows the ratio of the economic gain achieved by the forecaster relative to that of a perfect forecaster in the special case where the cost of acting on the assumption that recession will occur relative to the cost of failing to act when recession occurs has adjusted to reflect the unconditional probability of recession.

post-benchmark-adjustment series to eliminate the effects prior to the analysis.[22]

The forward-looking data are the experts' forecasts on output provided in the *Survey of Professional Forecasters* (SPF) from $1968q4 - 2015q1$. The forecasts in the SPF are made around the mid-point of quarter t and include nowcasts of the current quarter and forecasts of up to four quarters ahead. In fact, the data on US macroeconomic aggregates in quarter $t-1$ are released for the first time at the end of the first month of quarter t so the first-release information on the previous quarter's output is available to the professional forecasters at the time they make their forecasts. Nevertheless, it is reasonable to assume ${}_{t}y_{t-1}$ and ${}_{t}y^{e}_{t+f}$, $f = 0, .., 4$ are determined simultaneously when working at the quarterly frequency.

Figure 1 illustrates the nature of the output series under investigation. Assuming for the moment that data is revised for three quarters, then actual quarter-on-quarter output growth at time t, can be measured by the post-revision series ${}_{t+4}y_t - {}_{t+4}y_{t-1}$. This series has an average annualised rate of 0.61% (with standard deviation of 0.83%) and is plotted in Figure 1a alongside the first-release and first-revision series. The size of revisions is small on average but the first and second revisions have a range of [-1.58%, 1.63%] and [-1.23%, 1.55%], and standard deviation of 0.41% and 0.34%, respectively and so are often of a similar order of magnitude to the actual growth figures themselves. Figure 1b plots the revisions directly showing that there are occasionally some very large revisions, with a relatively large number in excess of 0.5% occurring during the late seventies and mid-eighties and a large number less than -0.5% in the early eighties and after 2007. The fact that these episodes coincide with periods of unusually strong or weak growth suggests that the measurement errors are (understandably) related to business cycle conditions and suggests that revisions may be more or less useful in forecasting growth outcomes at different times.

The expectations series obtained from the SPF are shown in Figure 1c, again set against actual post-revision growth. This figure shows that the expectations series also display some volatility but they move more conservatively than the actual growth series

[22] Benchmark adjustments took place in 1976q1, 1981q1, 1986q1, 1992q1, 1996q1, 1999q4, 2004q1 and 2009q3.

itself. The conservatism becomes more pronounced as the forecast horizon grows so that four-period-ahead survey expectations rarely move outside the [0.5%, 1.0%] range, especially over the latter half of the sample. Defining expectational errors observed in the SPF series by ${}_{t+4}y_t - {}_{t-f}y_t^e$ for $f = 0, .., 4$, i.e. comparing the post-revision series to the survey forecasts for up to 4 quarters earlier,[23] Figure 1d plots the expectational errors directly, showing some very large errors in the four-period-ahead forecasts.

4.1 'Real-time' evaluation of point and density forecasts

The purpose of the empirical work is to find whether the information contained in the revision and survey data is useful in forecasting. All of the models that we estimate can be accommodated by the meta model $\overline{M}_{3,3,\tau}$ defined in (2.7), with $r = 0, 1, 2, 3$ and $f = -1, 0, 1, 2, 3$, and in (2.1)-(2.3). Hence, twenty versions of the model in (2.1)-(2.3) are estimated with the most general including three revisions and survey forecasts up to three quarters ahead in addition to the first release data, while the most simple version of the model is the 'conventional model' which uses the first-release data only.

The empirical exercise begins by estimating the meta model $\overline{M}_{3,3,1991q2}$ based on the real time data available for $1968q4 - 1991q2$, using the 80-quarter period $1968q4 - 1988q3$ in estimation and holding back the 12 quarters' data for $1988q3 - 1991q2$ for the training period. Each of the twenty underlying models are estimated and subjected to the specification search procedure described earlier in which variables are dropped sequentially if the (absolute) value of the t-ratio is less than unity. The models are then used to produce forecasts of the various measures of output, including, for example, the one-period-ahead forecast of the first-release measure of contemporaneous output, ${}_{1988q4}y_{1988q3}$, say, and the four-period-ahead forecast of the post-revision measure of contemporaneous output, ${}_{1989q3}y_{1988q3}$. For the purpose of obtaining the model weights, we focus here on the forecast of the first-release measure, comparing this to the first-release outcome observed during the training period.[24] The twenty models are then estimated over the 81-quarter period

[23] This is the appropriate measure of 'expectational error' only if the survey participants report predictions of actual, post-revision output in their returns, not the predictions of the first-release measure.

[24] If just one set of weights is discussed, the first-release series is the natural choice on which to base the

$1968q4 - 1988q4$, and the forecast of the first-release measure ${}_{1989q1}y_{1988q4}$ obtained and compared to the observed outcome. This is repeated twelve times, moving through the training period and judging the relative performance of the twenty models each time to obtain the set of weights $w_{r,f,1991q2}$ defined in (2.5) for the MSE (or (2.6) for the MLS) for $r = 0, .., 3$ and $f = -1, .., 3$.[25] Moving on one period, this entire exercise can then be repeated over the sample $1969q1 - 1991q3$, using $1988q4 - 1991q3$ as the training period, to derive the set of weights $w_{r,f,1991q3}$ and so on to the final vintage date $w_{r,f,2013q3}$.

Figure 2a shows the weighted average of the revision length of the models included in the meta model $\overline{M}_{3,3,\tau}$ for $\tau = 1991q1, ..., 2014q4$ based on the models' forecasts judged according to their MSE and according to their MLS; that is μ^r_τ defined in (3.8) and using alternative weights as in (2.5) and (2.6). These weights incorporate the effect of the specification search, placing a weight of zero on models which have 'collapsed' to simpler low-r and/or low-f models as variables are dropped.[26] The plot based on MSE shows a high degree of stability: the average revision horizon is around 1.5 and lies in the range [1.00, 2.00] for nearly all the sample. The statistics reflect the finding that, when using MSE as the criterion, many models appear to perform equally well so that the average of their revision lengths is mid-way between zero and three, the minimum and maximum values. In contrast, the plot based on log score weights is much more discerning, showing a relatively low average revision length -i.e. with few revisions used - during the first part of the evaluation period but rising to a value of 1.5 during the early 2000's and to close to 3.0 - making full use of revisions - over the evaluation period after 2007. As noted above, there were a number of large revisions in the output data released in the early 2000's and again in the years following the financial crisis and it appears that the meta model

weights since the associated forecast errors are likely to be relatively small and stable over time compared to longer horizon forecasts.

[25]The log scores are calculated using the simulation methods described previously but applied recursively in each quarter of the training period; i.e. in each quarter, the score is the value of the Gaussian distribution, with mean and variance obtained from the simulated density, evaluated at the observed first-release outcome.

[26]Averaging over time, the proportion of variables dropped following the specification search in the 8-variable VAR underlying $\overline{M}_{3,3,\tau}$ is 34% (i.e. 46/136). The proportion dropped in the smaller 7-, 6-, 5-, 4-, 3-, 2-order VARs was 30%, 26%, 20%, 15%, 8% and 5% respectively.

adjusts to exploit the extra information contained in the revisions at this time, placing more weight on models that include the long revisions.[27]

Figure 2b shows the equivalent plots to those in Figure 2a focusing now on the expectations horizon included in the meta model $\overline{M}_{3,3,\tau}$ for $\tau = 1991q2, ..., 2015q1$ based on the models' forecasts judged according to their MSE and according to their MLS; that is, μ_τ^f defined in (3.8). The weights based on MSE are again relatively uninformative: the average is broadly stable at around 1, mid-way between the minimum and maximum values of -1 and 3, once more reflecting the difficulty in discriminating between models according to their point forecasts. However, the weights based on log scores are again more informative, broadly rising from values close to -1 at the beginning of the forecast period to closer to +1 through the late nineties and early 2000's, dropping to make little use of surveys between 2005-2007 and then playing an important role again following the financial crisis. This pattern is less straightforward to interpret, although comparison with Figure 1d again suggests that the data becomes more useful at times of increased uncertainty as the surveys appear to play a greater role when expectational errors are largest. In any case, the real time evaluation exercise indicates that survey data showing expected outputs one or two periods ahead can be useful for forecasting but that their usefulness changes over time.[28]

Figures 3a and 3b provide some further insight on this shift in the weights over time, showing the observed first-release output series alongside the point forecasts and 5th/95th percentile of the forecast densities for the most general model $M_{3,3,\tau}$ and for the simplest model $M_{0,-1,\tau}$ during two illustrative episodes. Figure 3a, which relates to the period 2005q1-2007q4 just before the financial crisis, shows that the point forecasts of the two models are broadly the same. However, the forecasts density is rather narrower for the $M_{0,-1,\tau}$ model so that it outperforms the more complicated $M_{3,3,\tau}$ model in terms of log score. In contrast, over the period 2008q1-2010q4 when there were some large revisions

[27]The correlation between the time-t weights on revisions, based on MLS, and the size of the revisions (measured as a five-period moving average of the absolute value of the first revision) is statistically significant at 0.20.

[28]The correlation between the time-t weights on survey expectations, based on MLS, and the size of the contemporaneous expectational errors is statistically significant at 0.36.

and expectational errors in the data, the point forecasts of the $M_{3,3,\tau}$ model are closer to the actual than those from model $M_{0,-1,\tau}$ and the densities are also wider so that the observed outcome lies towards the centre of the forecast density much more often. This illustrates the idea that, by placing more weight on the models including long revisions and surveys during this time, the meta model adjusts to incorporate the information contained in these series during the periods when they become significant.

Following the suggestion of Pesaran and Timmermann (2007), the empirical exercise described above can also be extended to include additional models defined using different sample periods as well as using more or less of the real-time data. This allows the meta model to trade off the advantage of extra precision on parameter estimates gained from longer samples of data against the danger of using samples that include structural breaks. As reported in Aristidou (2015), it turns out that very little weight is given to models based on short samples when the extended exercise is carried out so that the meta models obtained allowing for the additional models based on short sample are very similar to those described above based on the longest possible sample in each period. This suggests the time-varying weights found in Table 2 arise because of changes in the trade-offs between parameter estimation uncertainties and the effects of omitting variables and are not the result of structural breaks.

4.2 'Final' evaluation of point and density forecasts

The shifting weights over time provide insights on the usefulness of revisions and survey data in forecasting as would be judged at the time. The 'final assessment' statistics of Table 1 judge their usefulness over the whole evaluation period by comparing the forecast performance of four alternative meta models which are more or less constrained in their use of the revisions and survey data. Specifically here, we compare the performance of: (i) the general meta model discussed above, $\overline{M}_{3,3,\tau}$ for $\tau = 1991q2, ..., 2013q3$, which uses the revisions and survey data as the estimated weights indicate; (ii) the meta model $\overline{M}_{3,-1,\tau}$, obtained choosing models of differing sample lengths with desired use of revisions but making no use of the survey data at all; (iii) $\overline{M}_{0,3,\tau}$ making no use of revisions; and (iv) the 'conventional' meta model, $\overline{M}_{0,-1,\tau}$, making no use of revisions data or survey data.

In principle, we could conduct a separate forecast evaluation at every forecast horizon and for each of our output measures (i.e. the first-release measure and various revisions and survey expectation measures at different future dates). In what follows, we focus on the four-period-ahead forecast of the post-revision measure of contemporaneous output which is a natural way of thinking of the 'nowcast of current actual output'.[29]

The results of the table show the 'conventional' meta model, $\overline{M}_{0,-1,\tau}$ has a RMSE of 1.16% when judged over the whole evaluation period 1991q2-2013q3. The three meta models $\overline{M}_{3,-1,\tau}$, $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$ all outperform the conventional model, with smaller RMSEs reported in each case and all three show a statistically significant improvement according to the Giacomini-White (2006) [GW] test of equal forecasting performance (where the GW tests are performed using Newey-West robust standard errors with automatic selection for bandwidth). When weights are chosen according to the log scores, models $\overline{M}_{3,-1,\tau}$ and $\overline{M}_{3,3,\tau}$, both of which make use of the revisions data, show a statistically significant improvement in the log-score over that of the conventional meta model while model $\overline{M}_{0,3,\tau}$, which uses surveys but not revision data, actually has a statistically-significant deterioration in forecast performance compared to the conventional model. Hence, the 'final evaluation' results suggest it is a good idea to include real-time data when forecasting output although the argument for the use of revisions is more compelling than for the use of surveys. In every case, it is a good idea to take into account the fact that the real-time data might be more or less useful at different times.

4.3 Evaluation of event probabilities and fair bet outcomes

We have argued that models' forecast performance might also be judged by their ability to predict recession and to enhance decision-making. We also consider the models against this criterion then, with predictions again based on density forecasts of the post-revision

[29] In principle, the criterion used to construct the weights employed in the meta model could be changed to match the criterion used in the final evaluation. In what follows, we report results using the weights based on first-release forecasts discussed above even though the final evaluation is concerned with the post-revision nowcast. Results were qualitatively unchanged when weights were based on the post-revision nowcasts.

output measures and outcomes measured by the realised post-revision series. In what follows, we use six definitions of recession that capture different recessionary features of the business cycle at time τ; namely:

- $DROP1 : \{ (\ {}_{\tau+1+R}y_{\tau} - {}_{\tau+R}y_{\tau-1} < 0) \}$; i.e. a nowcast of negative growth based on post-revision observation;

- $DROP2 : \{ (\ {}_{\tau+1+R}y_{\tau} - {}_{\tau+R}y_{\tau-1} < 0) \cap (\ {}_{\tau+R}y_{\tau-1} - {}_{\tau+R-1}y_{\tau-2} < 0) \}$; i.e. a nowcast of two consecutive periods of negative growth at τ;

- $DROP2^{+} : \{ (\ {}_{\widetilde{\tau}+1+R}y_{\widetilde{\tau}} - {}_{\widetilde{\tau}+R}y_{\widetilde{\tau}-1} < 0) \cap (\ {}_{\widetilde{\tau}+R}y_{\widetilde{\tau}-1} - {}_{\widetilde{\tau}+R-1}y_{\widetilde{\tau}-2} < 0) \}$ for any $\widetilde{\tau} \in [\tau - 2, \tau + 2]$; i.e. a nowcast of two consecutive periods of negative growth occurring during the five-period interval centred on τ;

- $DROP2^{++} : \{ (\ {}_{\widetilde{\tau}+1+R}y_{\widetilde{\tau}} - {}_{\widetilde{\tau}+R}y_{\widetilde{\tau}-1} < 0) \}$ for any two $\widetilde{\tau} \in [\tau - 2, \tau + 2]$; i.e. a nowcast of two periods of negative growth occurring anytime during the five-period interval centred on τ;

- $BPEAK : \{ \ {}_{\tau+1+R}y_{\tau} < \max(\ {}_{\tau+R}y_{\tau-1}, \ {}_{\tau+R-1}y_{\tau-2}, \ {}_{\tau+R-2}y_{\tau-3}, ...) \}$; i.e. period τ output lies below its previous peak level;

- $BTREND : \{ \ {}_{\tau+1+R}y_{\tau} < \overline{y}_{\tau} \}$,where $\overline{y}_{\tau} = \frac{1}{5}(\ {}_{\tau+R-1}y_{\tau-2} + \ {}_{\tau+R}y_{\tau-1} + \ {}_{\tau+1+R}y_{\tau} + \ {}_{\tau+2+R}y_{\tau+1} + \ {}_{\tau+3+R}y_{\tau+2})$; i.e. output lies below trend, defined as the centred five-period moving-average of output.

$DROP1$ captures a basic feature of recession while $DROP2$ is a frequently-used definition. Both are concerned with the contemporaneous experience of the event dated at τ but relate to post-revision measures and so rely on 4-quarter-ahead forecasts. Neither of the events occur very regularly in our sample, ($DROP1$ occurs on 7% of occasions and $DROP2$ on 4%) and $DROP2^{+}$ and $DROP2^{++}$ therefore broaden the definition to consider nowcasts of two periods of negative growth observed, respectively, consecutively or in any quarter during the five-periods centred around τ. These definitions of recession are suggested in Anderson and Vahid (2001) and occur more frequently (13% and 14% of occasions respectively in our sample) and persist over time, providing a more varied

forecasting challenge than looking at $DROP1$ and $DROP2$ alone. The final two events, $BPEAK$ and $BTREND$, are also often used to define recessionary times. These are both defined by complicated sets of forecast output outcomes, but probability forecasts of the events can be readily calculated based on the simulation methods described earlier. As it happens, $BPEAK$ occurs 22% of the time in our sample, while $BTREND$ is, by definition, likely to occur more or less half of the time. These events also provide good variation in the forecasting challenge to our models therefore.

Our forecast evaluation exercise straddles three periods of relatively low growth (at the beginning of the nineties, at the beginning of the 2000's, and during the financial crisis) and these may or may not be interpreted as 'recession' according to the various definitions we have proposed. In the absence of specified pay-out contingencies, a forecasted probability that exceeds 0.5 is interpreted as predicting recession will occur. Figures 4 and 5 illustrate the type of results obtained, showing the forecast probabilities of $DROP1$ and $DROP2^{+}$ based on the meta models using MSE and MLS weights. According to Figures 4a,b, the meta models that use revision and survey data ($\overline{M}_{3,-1,\tau}$, $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$), based on either MSE or MLS weights, appear to perform well at forecasting a drop in output despite the relative infrequency of the event. In Figures 5a,b, the higher frequency and greater persistence of the $DROP2^{+}$ recession is shown to occur during all three periods of low growth in our forecast evaluation period and again these are predicted well by the meta models that use the real-time data.

One interesting feature exposed by the two figures is the way in which the perceived occurrence of recession can itself change over time due to data revision. For example, in Figure 4, negative growth measured by ${}_{\tau+1+R}y_{\tau} - {}_{\tau+R}y_{\tau-1}$ was observed in $\tau = 2001q3$ but not in $\tau = 2001q2$, while Figure 5 shows that *two* consecutive periods of negative growth were experienced in $2001q3$. This apparent contradiction arises because data continues to be updated even beyond the three (systematic) revisions we have incorporated into our model and used to define 'post-revision' output. In the event, revisions in the measures of output in 2001 more than four quarters later mean that growth in 2001q2, calculated to be positive in 2002q1 (as reported in Figure 4), is calculated to be negative in 2002q2 and a $DROP2$ (and a $DROP2^{+}$ and a $DROP2^{++}$) recession is retrospectively defined to

occur in Figure 5. This feature of the data again emphasises the importance of forecasting using real-time data although it also highlights the complications in forecast evaluation exercises and in interpreting decisions made in real time.

Tables 2 and 3 provide more formal measures of the extent to which the models meet the challenges of forecasting recessions defined in the various ways. For Table 2, the 90 predictions and outcomes observed over 1991q2-2013q3 are arranged into a two-by-two contingency table. Table 2a shows the proportion of forecasts that are correct ($\frac{YY+NN}{90}$) for each model and Table 2b reports the Kuipers scores. Table 2b also reports, in parentheses, the results of two tests described in Pesaran and Timmermann (2009): a static χ^2 test of whether a model's forecast performance is any better than would have been achieved guessing randomly based only on the unconditional probability of the event p; and a dynamic version in which the random guess also takes account of the possibility that the event is known to occur in runs.

The events $DROP1$, 2, 2^+, 2^{++} and $BPEAK$ occur relatively infrequently and so the accuracy rates (proportion correct) of Table 2a - which treat correct predictions of no-recession in the same way as correct predictions of recession - are high across all models as would be expected. Nevertheless, it appears that the meta models $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$ perform best: model $\overline{M}_{0,3,\tau}$ shows the largest accuracy rates throughout. The dominance of the models $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$ is confirmed, and exaggerated, in the Kuipers scores of Table 2b which focus more on the models' ability to correctly predict the rare recession events. The χ^2 tests also confirm that the performance of $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$ is significantly better than would be achieved by chance, unlike the 'conventional' meta model, $\overline{M}_{0,-1,\tau}$ and in many cases unlike model $\overline{M}_{3,-1,\tau}$. The over-riding conclusion then is that the models provide a valuable tool for forecasting rare recessionary events and that the models which include the survey data typically perform best in predicting these rare

events.[30,31]

Table 3 provides the results of evaluating forecasting performance in the more sophisticated 'fair bet' decision-making context, reporting the maximum possible return that would be achieved by a perfect forecaster and the returns achieved by a forecaster using each of the models expressed relative to this maximum. Table 3a relates to the symmetric fair bet in which the investor bets every period, gains the same payout for YY as NN and compares the forecast recession probability against the same 0.5 threshold used in Table 2. Given that the set up is similar to that underlying the Kuipers score, it is not surprisingly that the results are very similar to those in Table 2b: the rank ordering of the models obtained for each event is the same with the models using survey data, $\overline{M}_{3,3,\tau}$ and especially $\overline{M}_{0,3,\tau}$ performing best; and the performance measures are high and broadly similar for $DROP1$, 2, 2^{+}, 2^{++} and $BPEAK$and lower for $BTREND$ in both tables. The results are similar in Table 3b, where the asymmetric setup delivers a higher payout on events that are more rare, although $\overline{M}_{0,3,\tau}$ and $\overline{M}_{3,3,\tau}$ share the honour of best performing model across the definitions of recession. Again though the over-riding conclusion is that models which include survey data perform best in predicting rare recessionary events.

5 Concluding Comments

The empirical exercise provides clear-cut evidence that forecasts of output growth and recession are enhanced through the use of real-time data. The 'real-time' and 'final' eval-

[30]The results show that none of the models perform well in predicting $BTREND$ recessions. This is perhaps unsurprising given the complexity of the event - involving a non-linear function of post-revision output levels over five quarters. Nevertheless, this illustrates the point that there are events that are difficult for any model to predict and that forecasters should consider when models are fit for purpose and when they are not.

[31]We also calculated the more general statistical tests of no forecasting power based on the Receiver Operating Characteristic (ROC) curve; see Hanley and McNeil (1982). These tests consider the predictive power of the models for all possible threshold probability values (not just 0.5 as above). In the event, the tests delivered broadly similar conclusions to those in Table 2: all the models were found to have significant forecasting power (except for $\overline{M}_{0,-1,\tau}$ for DROP1) and $\overline{M}_{3,3,\tau}$ produced the largest test statsistics under both weight schemes and for almost all events.

uations of the forecasts from the VAR models considered in the paper show that point forecasts and density forecasts are improved by using survey data on expected future output movements and by using first-release and revisions data. The exercise shows that this is especially true if, as here, the modelling takes into account that the data can be more or less helpful at different times, with the revisions data appearing to be particularly important during downturns when larger measurement errors appear in the first-release data and with survey data being more important at times of higher uncertainty. A final evaluation of forecast performance shows models that include real-time data show a statistically significant improvement over conventional models although, based on MSE and MLS, the argument for the use of revisions is more compelling than that for the use of surveys. On the other hand, it is the survey data which seems particularly important when forecasting the likelihood of recessionary events. These are relatively rare and extreme events which conventional linear forecasting models might struggle to accommodate but which are incorporated into professional forecasters' predictions reasonably quickly. Survey data therefore provides the means to quickly include this information in a time series model so that, again, forecast performance is improved by allowing the data to be used more or less intensively at different times.

Table 1: RMSE and Average Log Scores for Output Growth Nowcasts, First Release Data (1991q2-2013q3)

	Without Specification Search	
	RMSE	Log Score
$\overline{M}_{0,-1,\tau}$ *(no revisions, no survey data)*	*0.0116*	*-2.611*
$\overline{M}_{3,-1,\tau}$ *(no surveys, revisions only)*	-0.0012**	0.8453**
$\overline{M}_{0,3,\tau}$ *(no revisions, surveys only)*	-0.0016**	-0.0915
$\overline{M}_{3,3,\tau}$ *(full revisions and survey data)*	-0.0018**	0.8279**
	With Specification Search	
	RMSE	Log Score
$\overline{M}_{0,-1,\tau}$ *(no revisions, no survey data)*	0.0116	-2.596
$\overline{M}_{3,-1,\tau}$ *(no surveys, revisions only)*	-0.0013**	0.8509**
$\overline{M}_{0,3,\tau}$ *(no revisions, surveys only)*	-0.0015**	-1.199
$\overline{M}_{3,3,\tau}$ *(full revisions and survey data)*	-0.0018**	0.8203**

Notes: The meta models $\overline{M}_{R,F,\tau}$ are as defined in (2.7). Actual RMSE and average log scores are reported for model $\overline{M}_{0,-1,\tau}$, and scaled difference from model $\overline{M}_{0,-1,\tau}$ are reported for other models. A '$*$' denotes significance at the 10% level, '$**$' denotes significance at 5% level and '$***$' significance at the 1% level in the Giacomini-White (2006) test of equal forecast performance testing whether the RMSE and the log predictive score are significantly smaller or larger, respectively, than the corresponding statistics from model $\overline{M}_{0,-1,\tau}$.

Table 2a: Hit Rate, Post Revision Outcomes (1991q2-2013q3)

	p	RMSE Weights				Log Score Weights			
		$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$	$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$
Event									
DROP1	*7%*	0.922	0.911	**0.956**	**0.956**	0.933	0.933	**0.956**	0.944
DROP2	*4%*	0.956	0.956	**0.978**	0.967	0.956	0.956	**0.978**	**0.978**
DROP2+	*13%*	0.889	0.911	**0.944**	**0.944**	0.922	0.922	**0.944**	0.933
DROP2++	*14%*	0.867	0.844	**0.933**	0.911	0.922	0.867	**0.933**	0.911
BPEAK	*22%*	0.878	0.900	**0.911**	0.900	0.911	0.889	**0.922**	0.889
BTREND	*48%*	0.533	**0.556**	0.544	0.500	0.533	0.511	**0.544**	0.500

Table 2b: Kuipers Score, Post Revision Outcomes (1991q2-2013q3)

	p	RMSE Weights				Log Score Weights			
		$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$	$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$
Event									
DROP1	*7%*	−0.012 (−,−)	0.131 (*,−)	**0.488** (***,***)	**0.488** (***,***)	0.000 (−,−)	0.155 (**,−)	**0.488** (***,***)	**0.321** (***,*)
DROP2	*4%*	0.000 (−,−)	0.238 (***,−)	**0.738** (−,−)	0.488 (***,***)	0.000 (−,−)	0.238 (***,−)	**0.738** (−,−)	0.500 (***,***)
DROP2+	*13%*	0.237 (***,−)	0.474 (***,*)	**0.654** (***,***)	**0.654** (***,***)	0.487 (***,**)	0.558 (***,**)	**0.634** (***,***)	0.571 (***,**)
DROP2++	*14%*	0.141 (***,−)	0.499 (***,*)	0.602 (***,***)	**0.640** (−,−)	0.525 (***,***)	0.524 (***,***)	0.602 (***,***)	**0.640** (***,***)
BPEAK	*22%*	0.486 (***,***)	**0.693** (***,**)	0.636 (***,−)	0.621 (***,−)	0.636 (***,−)	0.643 (***,−)	**0.721** (***,**)	0.571 (***,−)
BTREND	*48%*	0.053 (−,−)	**0.105** (−,−)	0.094 (−,−)	−0.001 (−,−)	0.065 (−,−)	0.020 (−,−)	**0.088** (−,−)	−0.003 (−,−)

Note: The meta models $\overline{M}_{R,F,\tau}$ are as defined in (2.7). Event DROP1 is one-period negative output growth; DROP2 is two successive periods of negative output growth; DROP2+ is two successive periods of negative output growth over a centered 5-period interval; DROP2++ is at least two periods of negative output growth over the corresponding interval; BPEAK is output level below previous peak; BTREND is output level below 5-period moving average. p is the unconditional probability of the event 1991q2-2013q3. Emboldened figures show the largest hit rate and Kuipers scores (KS). The figures in parentheses (a,b) below the KS show the outcome of the static and dynamic versions of the Pesaran and Timmerman (2009) tests of no additional predictive power beyond that of the unconditional probability; a '***' indicates significance at 1% level, '**' indicates significance at 5% level, '*' indicates significance at 10% level, and '-' indicates no significance at 10% level.

Table 3a: Returns to Fair Bet with Symmetric Payoffs,

Post Revision Outcomes (1991q2-2013q3)

		RMSE Weights				Log Score Weights			
		$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$	$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$
Event	Max								
DROP1	*12.79*	0.38	0.29	**0.64**	**0.64**	0.46	0.46	**0.64**	0.55
DROP2	*8.35*	0.48	0.48	**0.74**	0.61	0.48	0.48	**0.74**	**0.74**
DROP2+	*27.05*	0.52	0.62	**0.76**	**0.76**	0.66	0.66	**0.76**	0.71
DROP2++	*29.54*	0.46	0.37	**0.73**	0.64	0.69	0.46	**0.73**	0.64
BPEAK	*47.55*	0.65	0.71	**0.74**	0.71	0.74	0.68	**0.77**	0.68
BTREND	*89.65*	0.06	**0.11**	0.09	0.00	**0.06**	0.02	0.09	0.00

Table 3b: Returns to Fair Bet with Asymmetric Payoffs, Post Revision Outcomes (1991q2-2013q3)

		RMSE Weights				Log Score Weights			
		$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$	$\overline{M}_{0,-1,\tau}$	$\overline{M}_{3,-1,\tau}$	$\overline{M}_{0,3,\tau}$	$\overline{M}_{3,3,\tau}$
Event	Max								
DROP1	*84*	-0.10	0.08	0.21	**0.24**	-0.13	0.04	**0.18**	0.12
DROP2	*86*	0.49	0.76	0.69	**0.85**	0.70	0.70	0.70	**0.74**
DROP2+	*78*	0.09	0.35	**0.50**	**0.50**	0.25	0.47	0.29	**0.56**
DROP2++	*77*	0.02	0.26	**0.54**	0.38	0.28	0.26	0.24	**0.36**
BPEAK	*70*	0.29	0.51	**0.63**	0.50	0.27	0.43	**0.49**	0.39
BTREND	*47*	0.04	0.02	**0.24**	0.05	**0.17**	0.00	0.09	-0.02

Note: The events are described in notes to Table 2. The maximum possible return, achieved by a perfect forecsater, is reported in italics and the returns achieved by a forecaster using the meta models are expressed relative to this maximum. Emboldened figures show the largest return.

References

Abhyankar, A., Sarno, L., and G. Valente (2005), "Exchange Rates and Fundamentals: Evidence on the Economic Value of Predictability", *Journal of International Economics*, 66, 325-348.

Anderson, H. and F. Vahid (2001), "Predicting the Probability of a Recession with Non-Linear Autoregressive Leading-Indicator Models", *Macroeconomic Dynamics*, 5, 482-505.

Ang, A., G. Bekaert and M. Wei (2007), "Do Macro Variables, Asset Markets or Surveys Forecast Inflation Better ?", *Journal of Monetary Economics*, 54, 1163-1212.

Aretz, K. and D. Peel (2010), "Spreads versus Professional Forecasters as Predictors of Future Output Change", *Journal of Forecasting*, 29, 517-522.

Aristidou, C. (2015), "Nowcast and Forecast Evaluation when Real-Time Data are Available", *Working Paper*, University of Nottingham.

Arouba, S.B. (2008), "Data Revisions are not Well Behaved", *Journal of Money, Credit and Banking*, 40, 319-340.

Barberis,N. (2000), "Investing for the Long Run when Returns Are Predictable", *Journal of Finance*, 55, 225-264.

Banbura, M. and G. Rünstler (2011),. "A Look into the Factor Model Black Box: Publication Lags and the Role of Hard and Soft Data in Forecasting GDP", *International Journal of Forecasting*, 27, 2, 333-346.

Clements, M.P. and A.B. Galvão (2013), "Forecasting with Vector Autoregressive Models of Data Vintages: US Output Growth and Inflation", *International Journal of Forecasting,* 29, 698–714.

Clements, M.P. and D.F. Hendry (2005), "Guest Editors' Introduction: Information in Economic Forecasting", *Oxford Bulletin of Economics and Statistics* (Supplement), 67, 713-753.

Croushore, D. (2006), "Forecasting with Real-Time Economic Data", in G. Elliott, C.W.J Granger and A. Timmermann (eds.) *Handbook of Economic Forecasting*, 961-982.

Croushore, D. (2010), "An Evaluation of Inflation Forecasts from Surveys using Real-Time Data", *BE Journal of Macroeconomics: Contributions*, 10,

Croushore, D. (2011), "Frontiers of Real-Time Data Analysis", *Journal of Economic Literature*, 49, 72-100.

Croushore, D. and T. Stark, (2003), "A Real-Time Data Set for Macroeconomists: Does the Data Vintage Matter?", *Review of Economics and Statistics*, 85, 605-617.

Elliott, G., A. Gargano and A. Timmermann, (2013), "Complete Subset Regressions", *Journal of Econometrics*, 177, 357-373.

Frale, C., M. Marcellino, G.L. Mazzi and T. Proieyyi, (2010), "Survey Data as Coincident or Leading Indicators", *Journal of Forecasting,* 29, 109-131.

Garratt, A. and K.C. Lee, (2010), "Investing Under Model Uncertainty: Decision Based Evaluation of Exchange Rate Forecasts in the US, UK and Japan", *Journal of International Money and Finance*, 29, 3.

Garratt, A., K.C. Lee, E. Mise and K. Shields, (2008), "Real Time Representations of the Output Gap", *Review of Economics and Statistics*, 2008, 90, 4, 792-804.

Garratt, A., K.C. Lee, M.H. Pesaran and Y. Shin, (2003), "Forecast Uncertainties in Macroeconometric Modelling: An Application to the UK Economy", *Journal of the American Statistical Association*, 98, 464, 829-838.

Garratt, A., K.C. Lee and K. Shields, (2018), "The Role of Uncertainty, Sentiment and Cross-Country Interactions in G7 Output Dynamics", (forthcoming) *Canadian Journal of Economics.*

Giacomini, R. and H. White (2006), "Tests of Conditional Predictive Ability", *Econometrica,* 74, 1545–1578.

Granger, C.W.J and M. Machina (2006), "Forecasting and Decision Theory", in G. Elliott, C.W.J. Granger and A.Timmermann (eds.), *Handbook of Economic Forecasting*, 81-98.

Granger, C.W.J. and M.H. Pesaran (2000), "Economic and Statistical Measures of Forecast Accuracy", *Journal of Forecasting*, 19, 537-560.

Hanley, J. and B.J. McNeil (1982) "The Meaning and Use of the Area Under the Receiver Operating Characteristic (ROC) Curve", *Radiology*, 143, 29-36

Harvey, D. and P. Newbold (2005), "Forecast Encompassing and Parameter Estimation", *Oxford Bulletin of Economics and Statistics* (Supplement), 67, 815–835.

Jacobs, J. and S. van Norden (2011), "Modeling Data Revisions: Measurement Error and Dynamics of 'True' Values", *Journal of Econometrics*, 161, 2, 101-109

Johnstone,.D.J., S. Jones, V.R.R. Jose, and M. Peat (2013), "Measures of the Economic Value of Probabilities of Bankruptcy", *Journal of the Royal Statistical Society A*, 176

Kishor and E.F. Koenig (2012), "VAR Estimation and Forecasting when Data are Subject to Revision", *Journal of Business and Economic Statistics*, 30, 181-190.

Koenig, E. F., S. Dolmas, and J. Piger (2003), "The use and abuse of real-time data in economic forecasting", *The Review of Economics and Statistics*, 85, 3, 618–628.

Leitch, G. and J.E. Tanner (1991), "Economic Forecasts Evaluation: Profits Versus the Conventional Measures", *American Economic Review*, 81, 580-590.

Loungani, P, H. Steklerb, and N. Tamirisaa, (2013), "Information rigidity in growth forecasts: Some cross-country evidence", *International Journal of Forecasting*, 29, 4, 605–621.

Mankiw, N.G., D.E. Runkle and M.D. Shapiro (1984), "Are preliminary announcements of the money stock rational forecasts?", *Journal of Monetary Economics*, 14, 1, 15-27

Matheson, T.D., J. Mitchell and B. Silverstone (2010), "Nowcasting and predicting data revisions using panel survey data", *Journal of Forecasting*, 29, 3, 313-330

Patterson, K.D. (2002), "The Data Measurement Process for UK GNP: Stochastic Trends, Long Memory and Unit Roots", *Journal of Forecasting*, 21, 245-264.

Pesaran, M.H. and S. Skouris (2002),"Decision-based Methods for Forecast Evaluation", in M.P. Clements and D.F. Hendry (eds.), *A Companion to Economic Forecasting*, Oxford: Blackwell.

Pesaran, M. H and A. Timmermann (2007), "Selection of Estimation Window in the Presence of Breaks", *Journal of Econometrics*, 137, 1, 134-161.

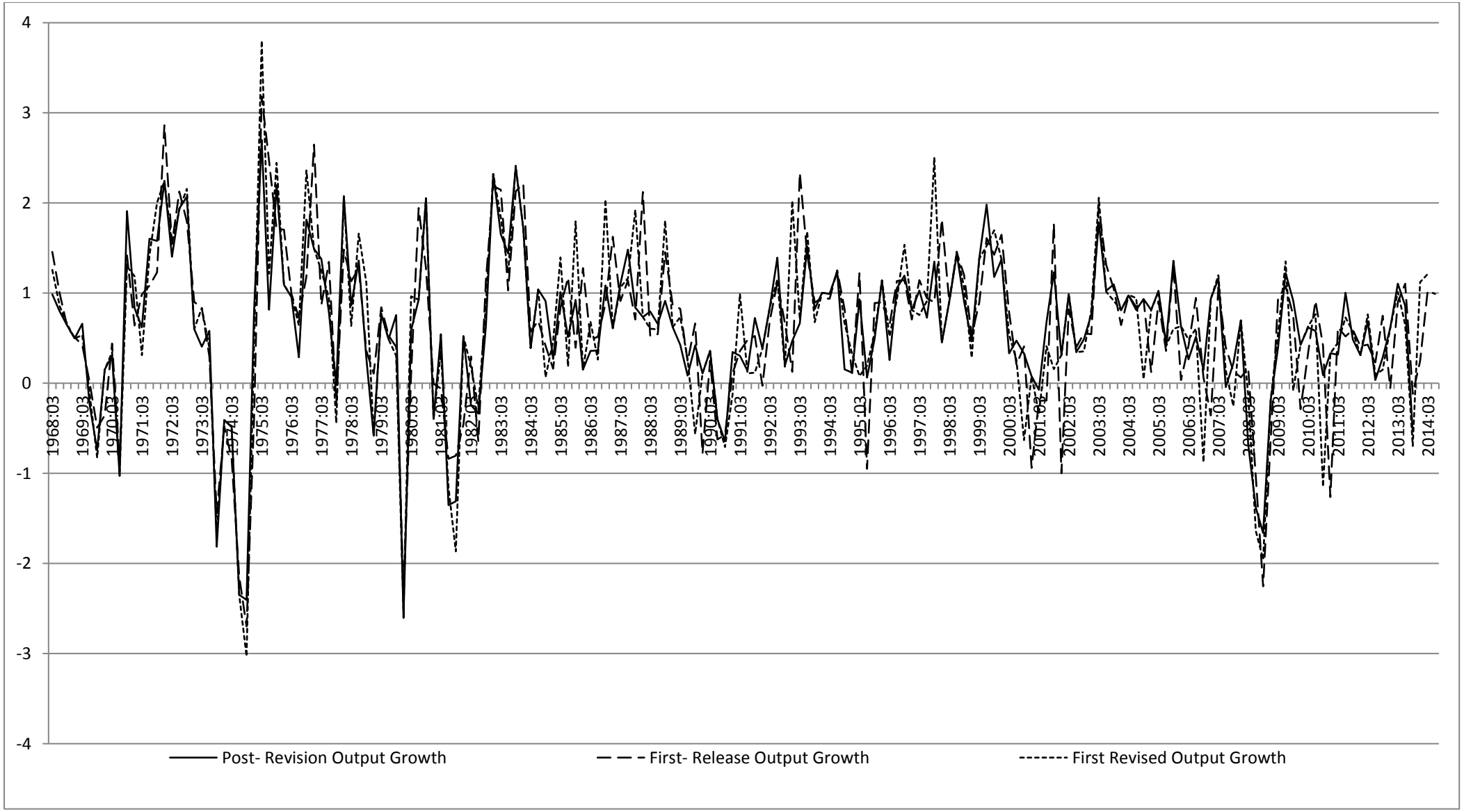
Pesaran, M. H and A. Timmermann (2009), "Testing Dependence Among Serially Correlated Multicategory Variables", *Journal of the American Statistical Association*, 104, 325-337.

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

Timmermann, A. (2006) "Forecast Combination", in in G. Elliott, C.W.J. Granger and A. Timmermann (eds.), *Handbook of Economic Forecasting*, Amsterdam: North Holland, 135-196.

Woodcock, F. (1981), "Hannsen and Kuipers Discriminant Related to the Utility of Yes/No Forecasts", *Monthly Weather Review: The Journal of the American Meterological Society*, 172-173.

Figure 1a: Post-Revision, First-Release and First-Revised Output Growth



Note: Post-Revision Output Growth defined as $_{t+4}y_t - _{t+4}y_{t-1}$; First-Release Output Growth defined as $_{t+1}y_t - _{t}y_{t-1}$; First-Revised Output Growth defined as $_{t+2}y_t - _{t+1}y_{t-1}$.

Figure 1b: First and Second Revision of the Output Series

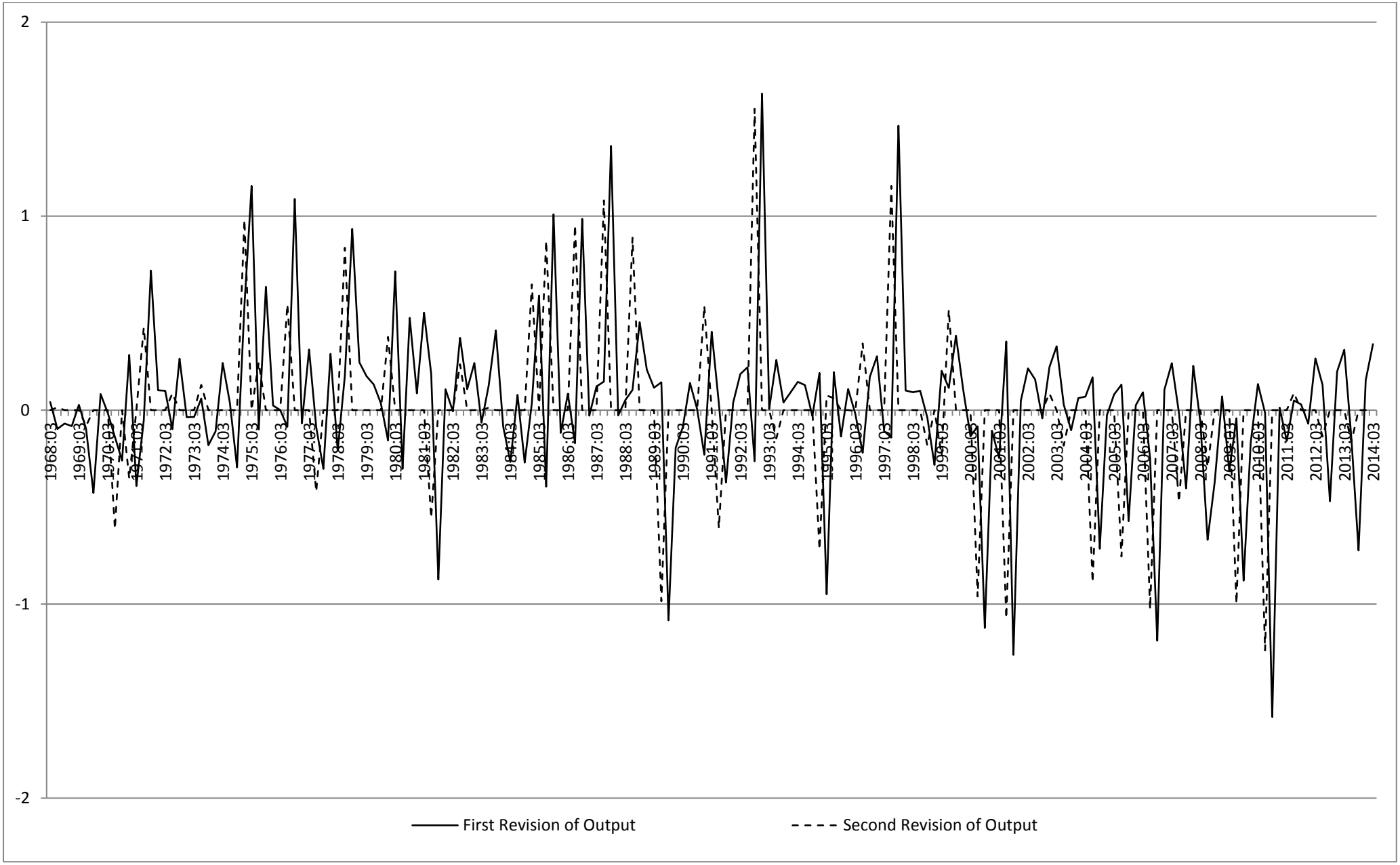
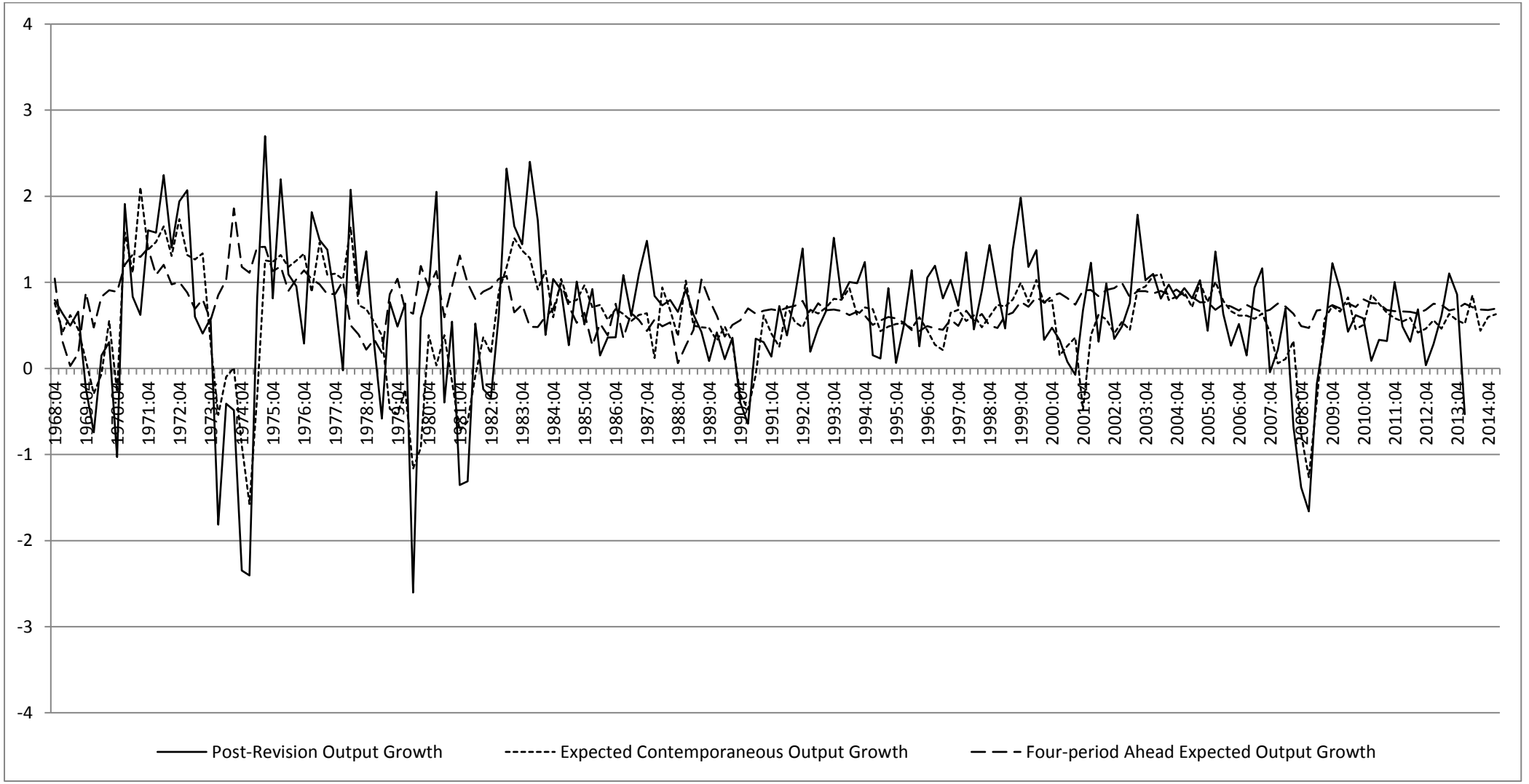


Figure 1c: Expected Contemporaneous Output Growth, Four-Period Ahead Expected Output Growth and Post-Revision Output Growth



Note: Expected Contemporaneous Output Growth defined as $_{t}y_{t} - _{t}y_{t-1}$; Four-Period Ahead Expected Output Growth defined as $_{t-4}y_{t} - _{t-4}y_{t-1}$; Post- Revision Output Growth defined $_{t+4}y_{t} - _{t+4}y_{t-1}$.

Figure 1d: Contemporaneous and Four-Period Ahead Expectational Errors

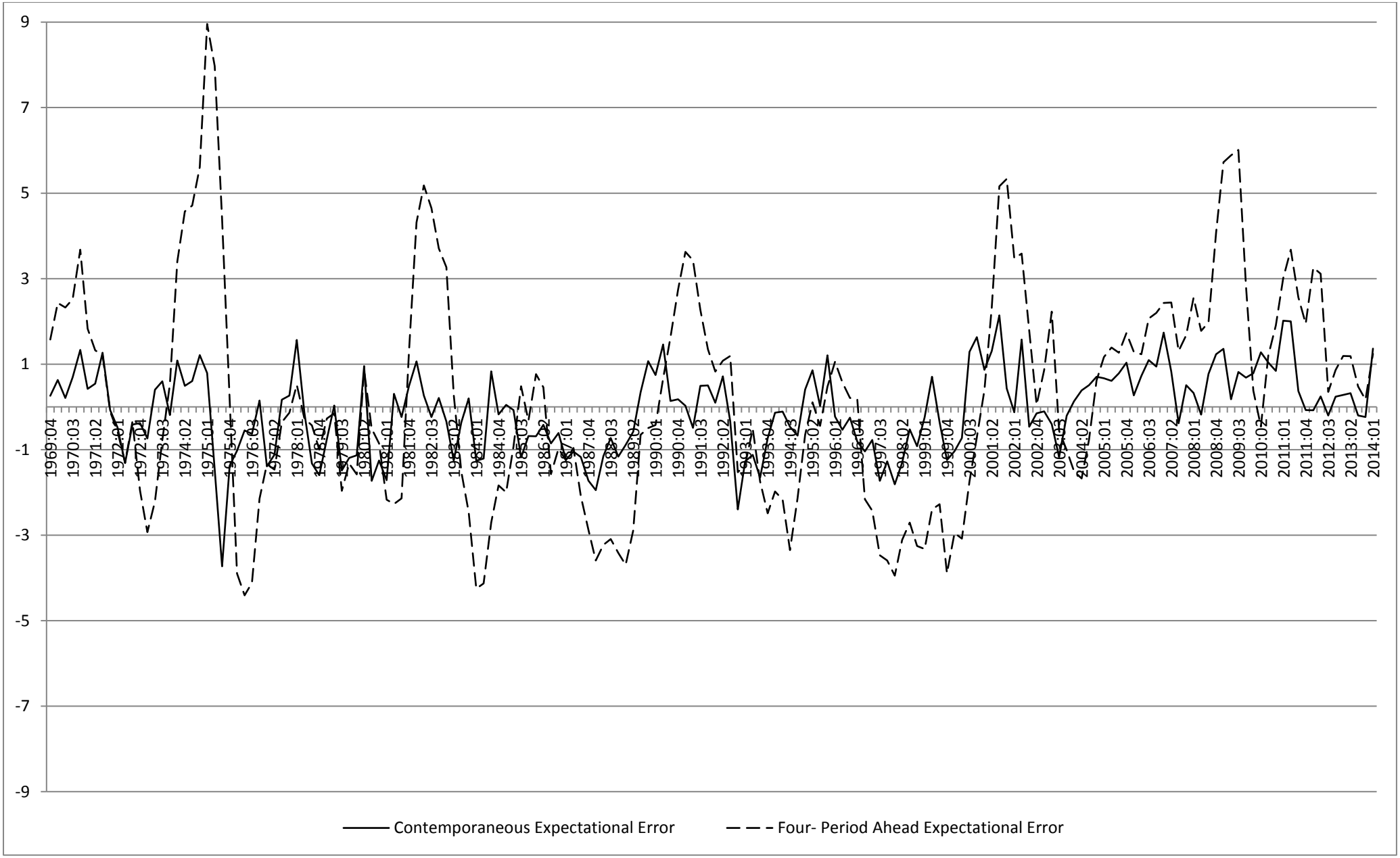


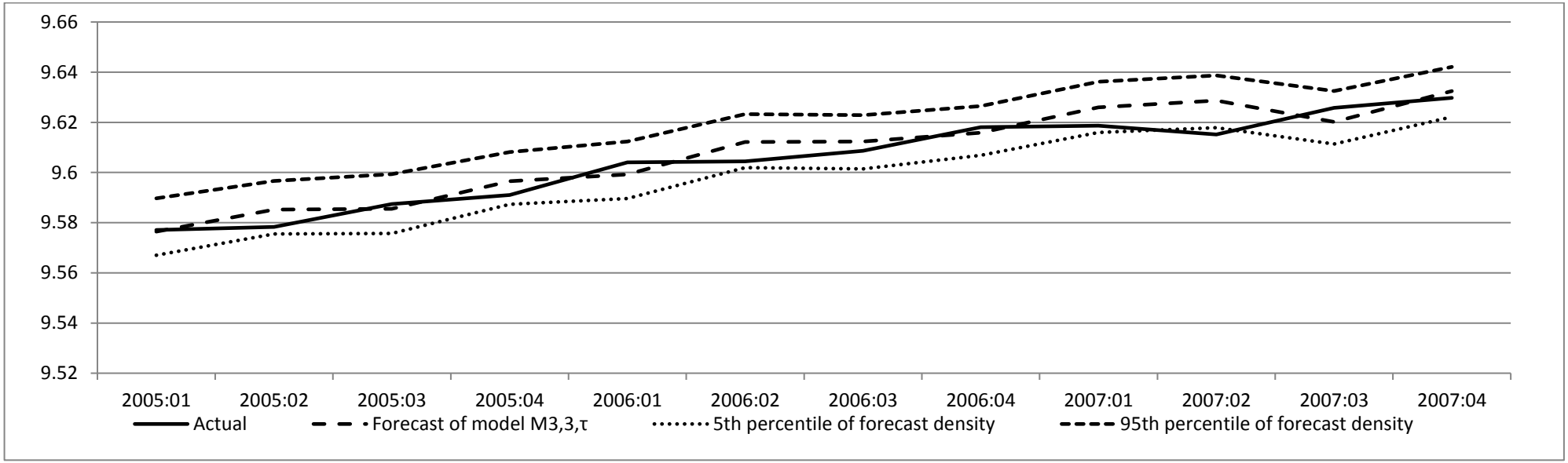
Figure 2a: Real Time Assessment of the Use of Revisions Data in Post-Revision Output Forecasts



Figure 2b: Real Time Assessment of the Use of Survey Data in Post-Revision Output Forecasts (With Specification Search)



Figure 3a: Mean, 5th percentile and 95th percentile of forecast density of models $\bar{M}$3,3,τ and $\bar{M}$0,-1,τ under the log score weight scheme over the period 2005:01- 2007:04, First Release Outcomes



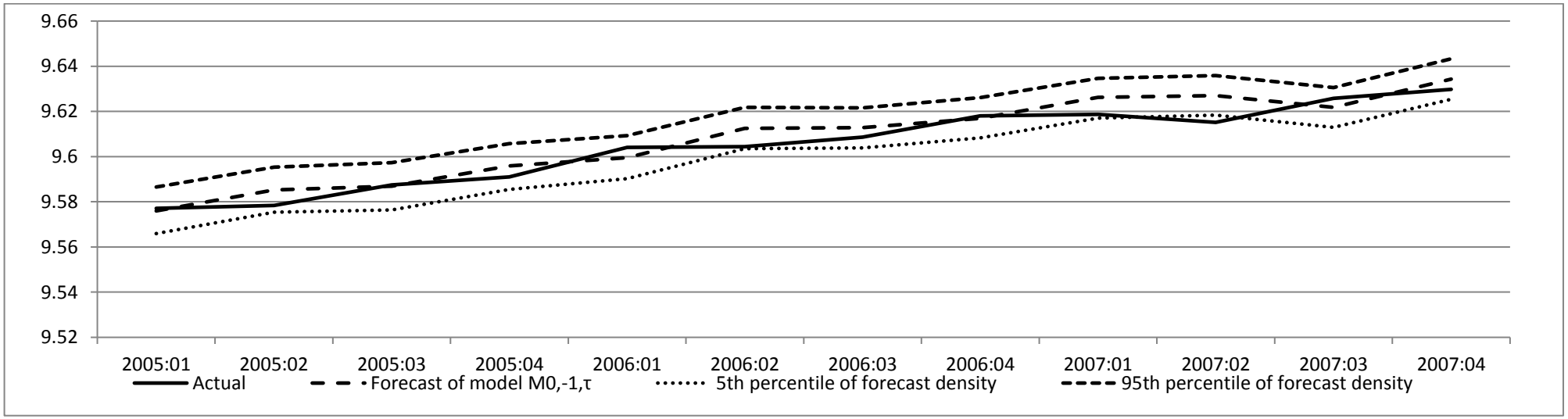
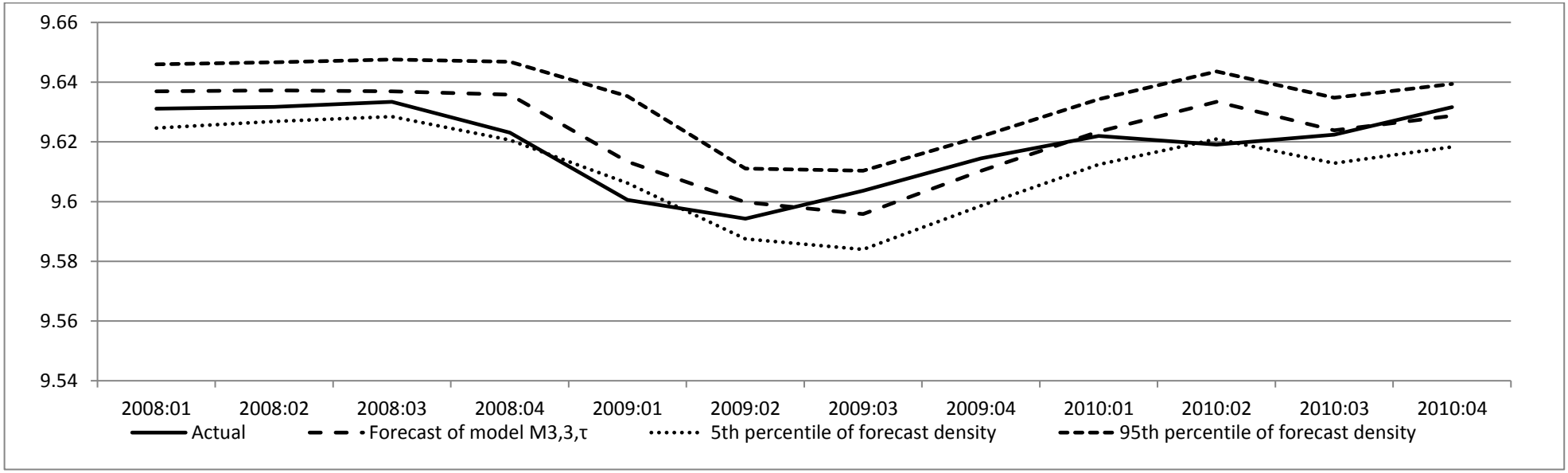


Figure 3b: Mean, 5th percentile and 95th percentile of forecast density of models $\bar{M}$3,3,τ and $\bar{M}$0,-1,τ under the log score weight scheme over the period 2008:01- 2010:04, First Release Outcomes



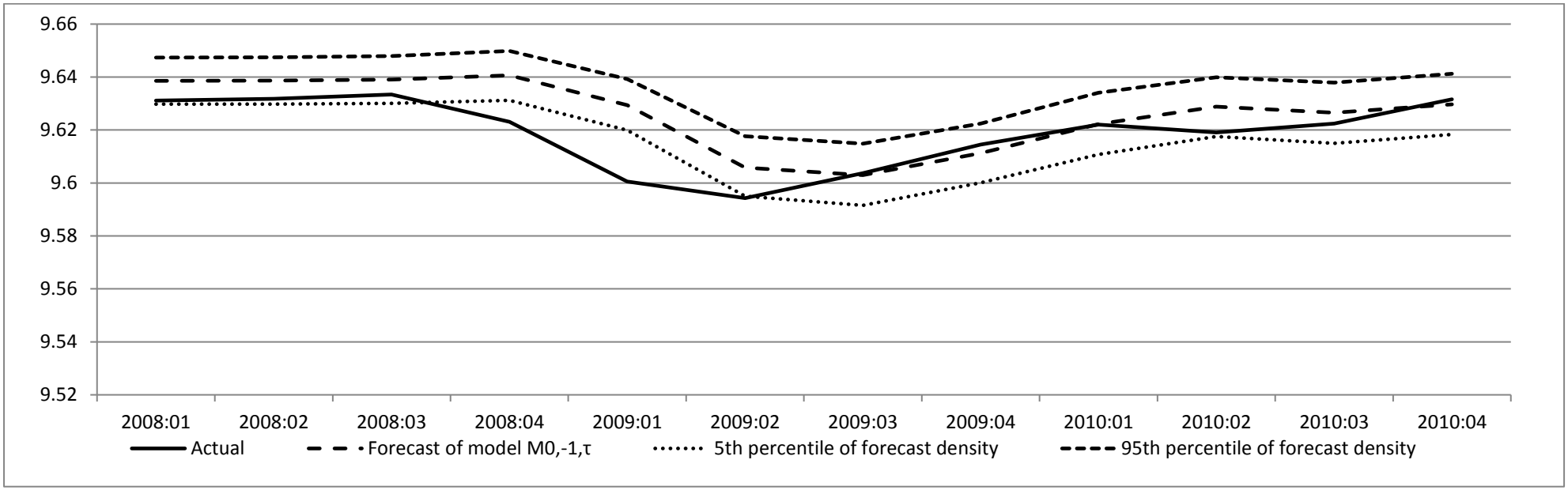


Figure 4a: Forecast probability of one period of negative output growth based on meta models constructed using RMSE weights

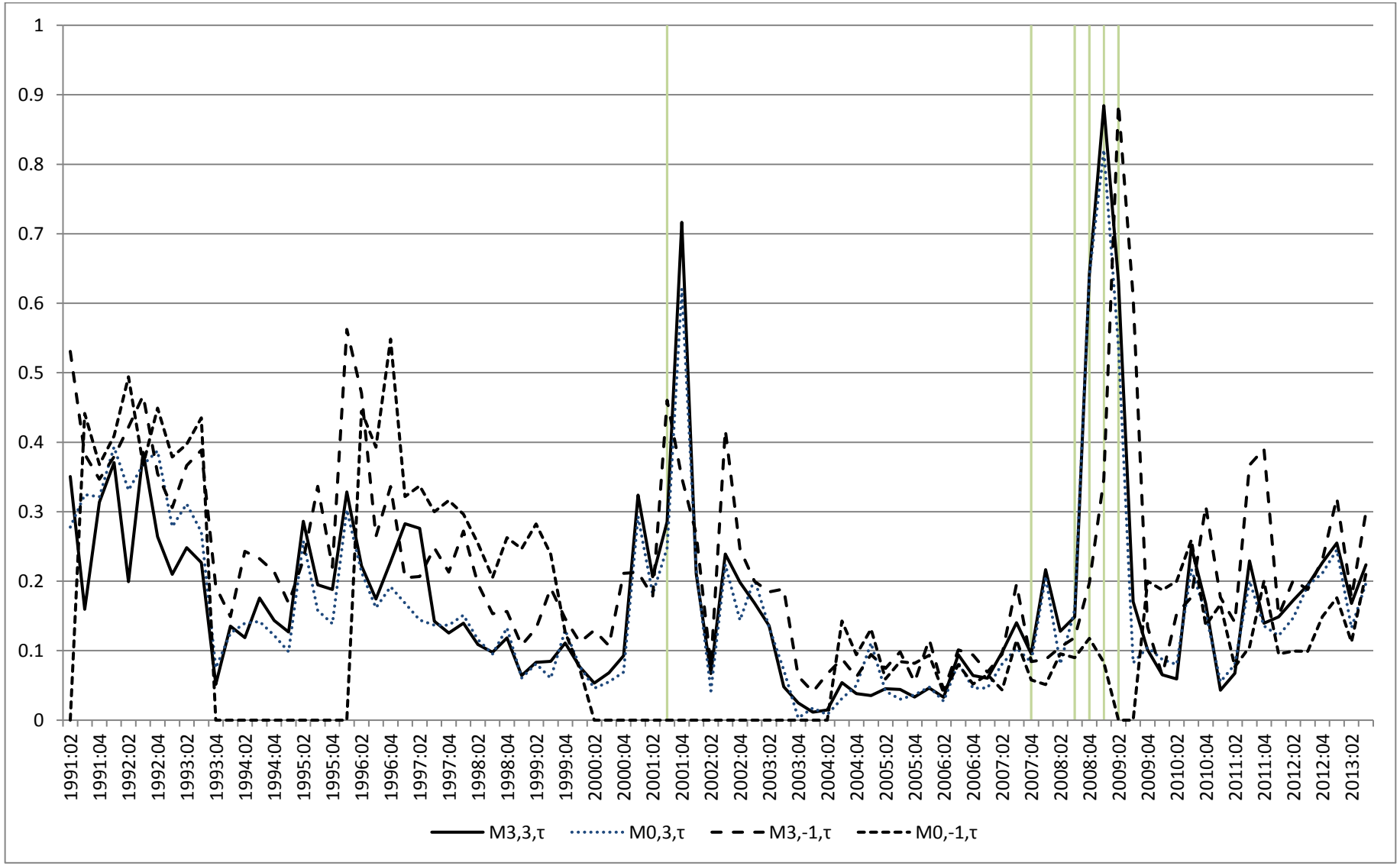
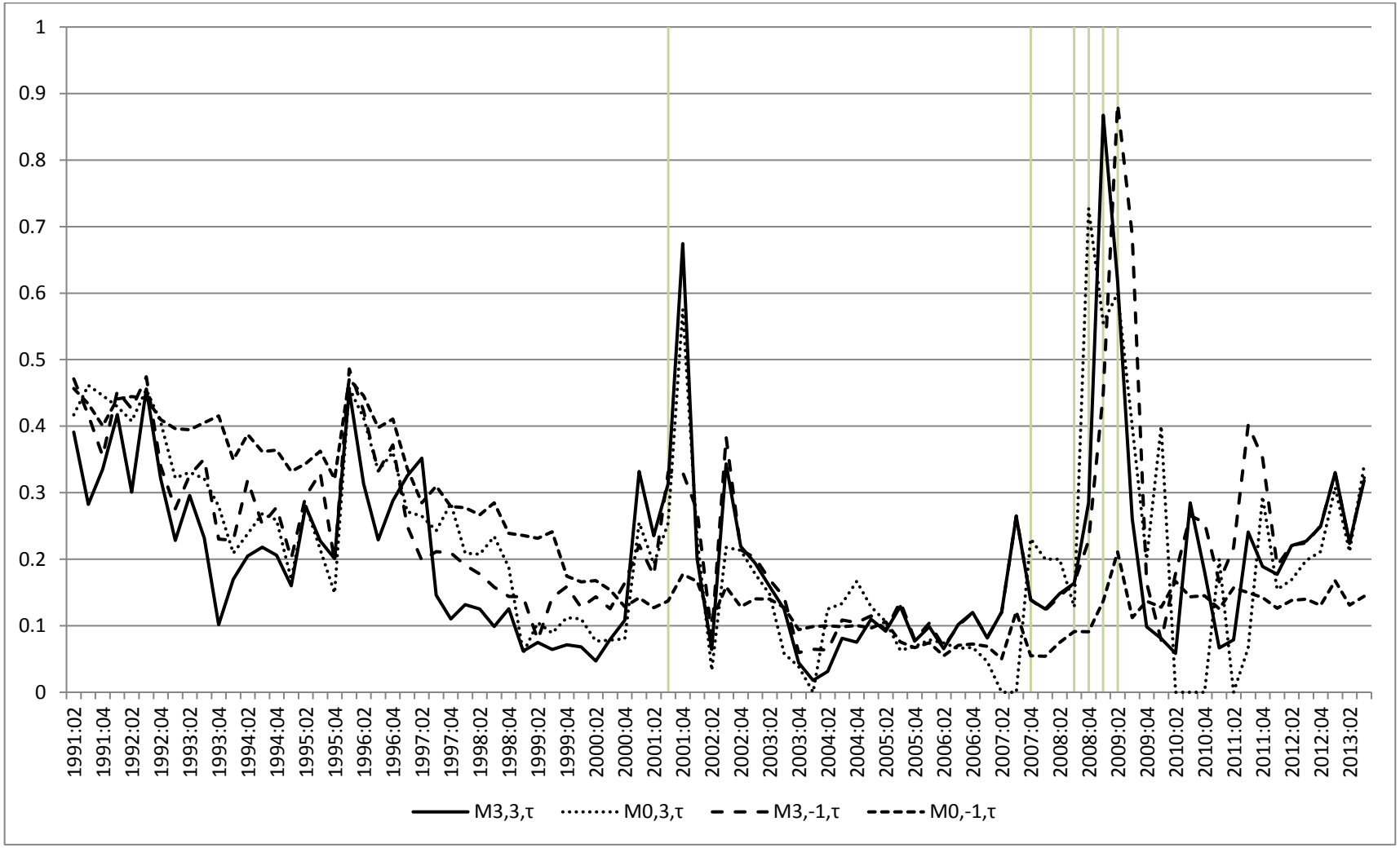


Figure 4b: Forecast probability of one period of negative output growth based on meta models constructed using log-score weights



Note: Vertical lines denote when the event has taken place (2001:03, 2007:04 and 2008:03-2009:02)

Figure 5a: Forecast probability of two consecutive periods of negative output growth over a five-period interval based on meta models constructed using RMSE weights

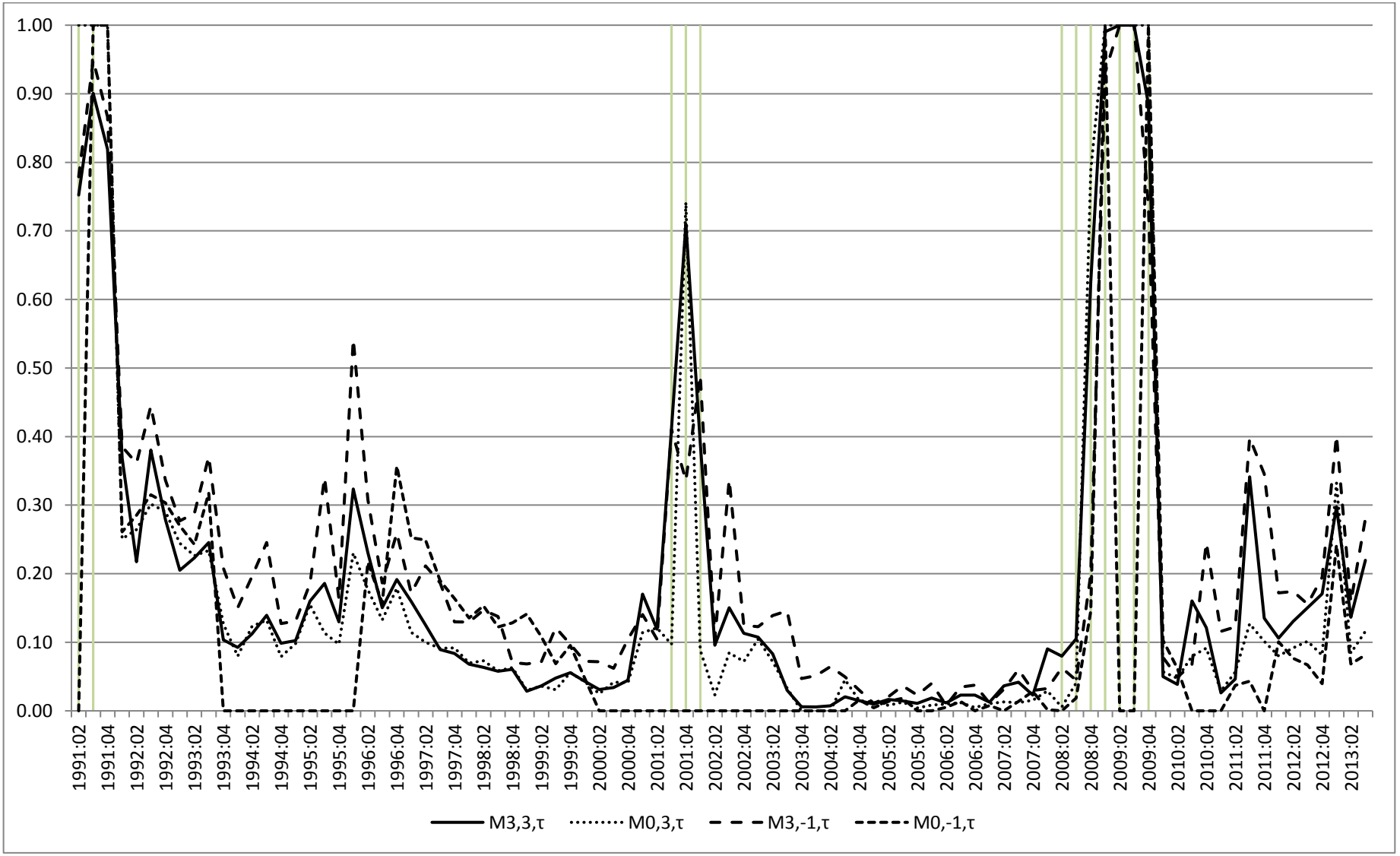
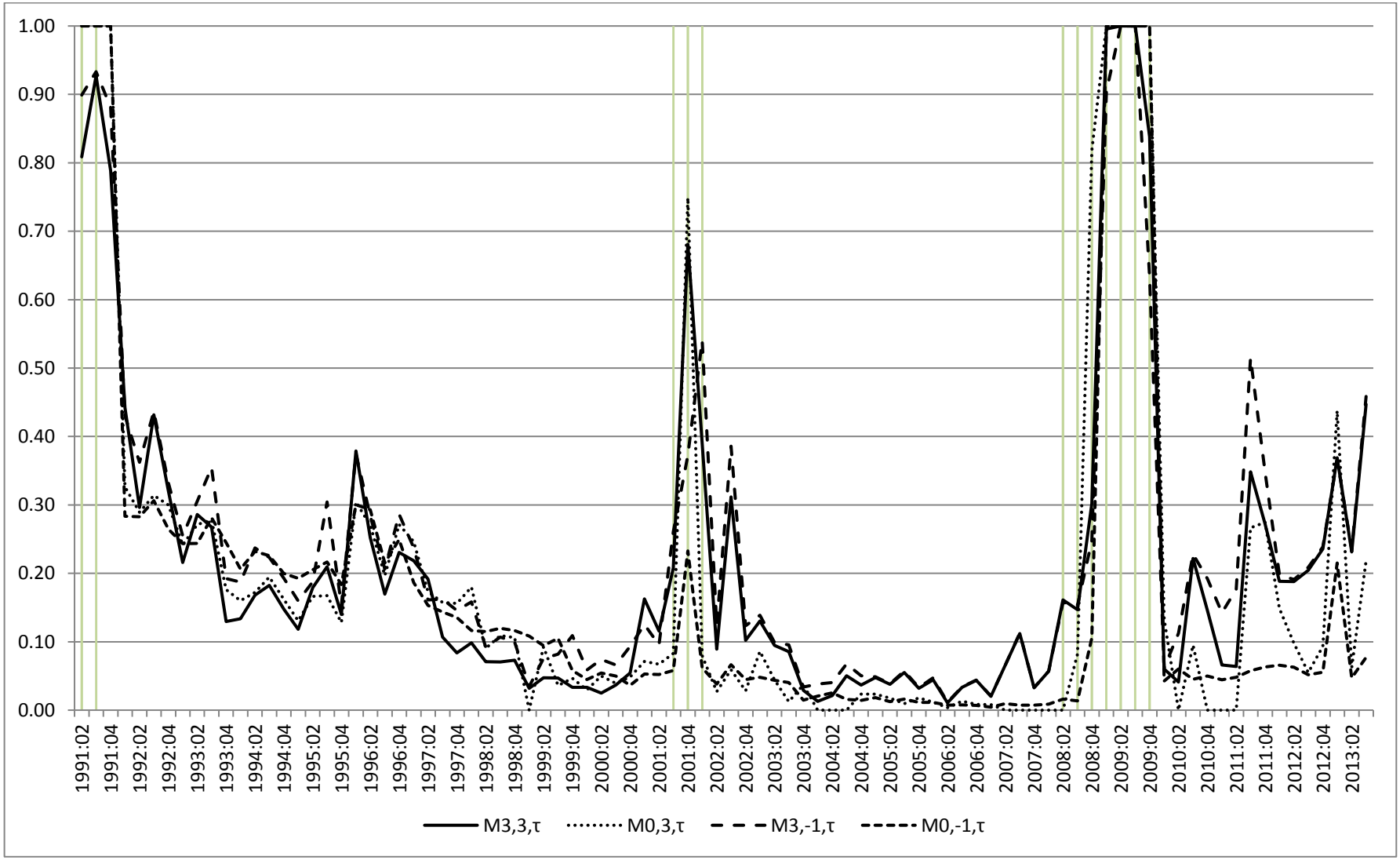


Figure 5b: Forecast probability of two consecutive periods of negative output growth over a five-period interval based on meta models constructed using log-score weights



Note: Vertical lines denote when the event has taken place (1991:02-1991:03, 2001:03-2002:01, 2008:02-2009:04)