Forecaster Efficiency, Accuracy and Disagreement: Evidence using Individual-Level Survey Data

Michael P. Clements*
ICMA Centre,
Henley Business School,
University of Reading,
Reading RG6 6BA
m.p.clements@reading.ac.uk.

March 10, 2019

Abstract

Recent theories of expectations which stress the role of information rigidities suppose that agents make efficient forecasts given their information sets. These theories have generally been tested on aggregate quantities, such as (cross-sectional) mean forecasts and mean forecast errors. We use individual-level data to consider whether there are systematic differences between forecasters in terms of their degrees of contrarianism, and the accuracy of their forecasts, and whether these are explicable by inefficiencies in the use of information. We find that forecaster inefficiency cannot the explain persistence in levels of disagreement across forecasters, but there is evidence that the inefficient use of information is responsible for persistent differences in accuracy across forecasters.

Keywords: Expectations formation, informational rigidities, disagreement, forecast efficiency. C53, E37.

*Very helpful comments from two referees of this journal are gratefully acknowledged, as is the guidance of the editor, Ken West. An earlier version of this paper was distributed as ICMA Centre Discussion Paper ICM-2016-08, *Are Macro-Forecasters Essentially The Same? An Analysis of Disagreement, Accuracy and Efficiency.*

1 Introduction

Recent years have seen much innovative work on expectations formation, and in particular on explaining why forecasters disagree. The full-information rational expectations (FIRE) model in which all agents know the true structure of the economy and have access to the same information set leaves no room for differences in expectations across agents. The FIRE assumption has often been replaced with some notion of 'bounded rationality' or adaptive learning, such that agents act rationally subject to certain constraints (see, e.g., Sargent (1999)). Informational rigidities (IR) have become prominent: forecasters form their expectations rationally subject to the information constraints they face. The two key models of informational rigidities are sticky information, and noisy information.[1] Sticky information assumes that in each period, each agent updates their information (relative to the previous period) with a given probability. When they do update, they acquire full information and act as FIRE agents. The noisy information model assumes agents base their forecasts on the latest information, but only ever observe noisy signals about economic fundamentals. However, they filter the signal optimally, and conform to the rational expectations hypothesis conditional on their information set. Under both models of expectations behaviour agents' forecasts are efficient in the sense of Mincer and Zarnowitz (1969): their forecasts are uncorrelated with their forecast errors.

The influential papers to date provide support for the macro-level implications of the baseline IR model of noisy information: see Coibion and Gorodnichenko (2012, 2015). The IR models predict departures from rationality at the aggregate level. These departures are an emergent property, in the sense that they hold when we consider mean forecast errors and the revision to mean forecasts, but are not evident at the level of the individual forecaster. The baseline noisy information model assumes the noise-variance contaminating agents' signals is equal across agents, and that agents share a common model of the economy, so that forecasters are effectively *identical* or *interchangeable*: this period Forecaster A may happen to receive a more accurate signal than Forecaster B, resulting in A's forecast being more accurate than B's, but next period B is just as likely to produce the superior forecast as A. Although not all these assumptions (e.g., homogenous signal-noise ratios) are necessarily fundamental to the noisy information model, they have been found by Coibion and Gorodnichenko (2012, 2015) to be consistent with the macro-level evidence.

We consider the micro-level evidence for whether individual forecasters are essentially the same in certain key respects. We consider whether there are persistent differences between forecasters in terms of their degree of non-conformity with the 'consensus' (that is, their degree of disagreement or contrarianism). We consider whether forecasters are identical in terms of

[1] See, *inter alia*, Mankiw and Reis (2002) and Mankiw, Reis and Wolfers (2003)) for sticky information, and e.g., Woodford (2002) and Sims (2003) for noisy ifnormation.

forecast accuracy, or whether there are systematic differences, and if so, whether differences in forecasting ability are persistent over time. Systematic differences between forecasters in terms of their degree of contrarianism, or their forecast accuracy, would go against the baseline IR models, although such models could be adapted to accommodate such heterogeneity without jettisoning rationality. For example, under the noisy information model one could allow for signal heterogeneity, and allow some forecasters to receive more informative signals than others.

A more fundamental challenge to IR models would be posed by the finding that forecasters are not rational, in the sense that their forecasts are not efficient. Weak efficiency is the requirement that an agent's forecasts and forecast errors are not systematically correlated. If they were, then the forecasts would not make efficient use of forecast-origin information because the resulting forecast errors would be predictable from the forecasts (which are of course a function of the forecast origin information). Stronger tests of efficiency would consider other subsets of the forecast-origin information. The advantage of the approach of Mincer and Zarnowitz (1969) is that there is no ambiguity as to what is known to the agent: the forecast is obviously known to the agent who made it.

In principle at least it is straightforward to test for forecast efficiency. However, a rejection of the null of forecast efficiency (of the sort proposed by Mincer and Zarnowitz (1969), or a related test) could be dismissed on the grounds that statistical significance does not necessarily mean the implied departure from rationality is of economic importance. To respond to this, a key innovation in our paper is to gauge the importance of the departures from efficiency in terms of the measurable characteristics of forecaster behaviour, such as contrarianism - that some forecasters systematically disagree with the consensus to a greater or lesser extent, and differences across forecasters in terms of forecast accuracy. Is it the case that the more accurate forecasters make more efficient use of their information, or are some forecasters inherently better than others (even when all are using their information efficiently)? Do different degrees of contrarianism across individuals reflect different qualities of signals, or a failure to process the signals rationally? If one were to find that forecast inefficiency accounted for persistent differences in forecast accuracy, or contrarianism, one might conclude that statistical rejections of efficiency were also of economic significance or importance. These questions go to the heart of the assumption that IR forecasters are rational given their information sets.

The paper asks whether forecaster heterogeneity can be explained by forecaster inefficiency. The intention is not to formally test theories of information frictions. Theories such as the baseline noisy information model assume efficiency, and then generate disagreement by supposing the forecasters receive noisy signals on the fundamentals. We explore the extent to which forecaster inefficiency can account for heterogeneity (persistent differences across forecasters in terms of accuracy or contrarianism) by correcting all the forecasts for inefficiency.

We use a multivariate disagreement measure to take into account a forecaster's beliefs

about the inter-dependencies between the variables being forecast. When the form of these inter-dependencies matches the consensus view, then the measure is reduced (compared to for a forecaster who does not share the consensus view about 'how the economy operates'). We consider US professional forecasters expectations of consumption, investment and output, because the growth rates of these variables move together, and a number of studies have considered whether there are constant long-run or equilibrium relationships between the log levels of these variables.[2] It seems reasonable to suppose that individuals 'disagree less' when they agree about the inter-dependencies. We explain more fully with an illustrative example in the main text.

We also consider the micro-level evidence for the assertion that individual forecasters are equally accurate. Under the standard versions of the IR models forecasters are essentially interchangeable. Under noisy information, for example, agents receive homogeneous signals, have the same model of the economy, and use that information efficiently to generate their expectations. We find evidence against the assumption of equal accuracy. We then consider the reasons for the rejection of equal forecast accuracy within the confines of the noisy information model. We consider whether the differences in accuracy are attributable to heterogeneous signals or differences in the efficiency with which agents generate their forecasts, given their information sets.

Finally, we use the relationship between contrarianism and accuracy at the individual level to consider whether some forecasters do receive superior signals, resulting in the more accurate forecasters tending to stand apart form the crowd.

The plan of the remainder of the paper is as follows. Section 2 describes the forecast data used throughout the paper. Section 3 considers the evidence for forecast efficiency. Section 4 describes the multivariate measures of disagreement, and presents our empirical findings on forecaster disagreement. Section 5 describes the assessment of individual-level forecast accuracy. Section 6 addresses the question of whether more accurate forecasters are more or less contrarian, on average. If some forecasters systematically received superior signals, we might expect to find the more accurate forecasters tend to stand apart from the consensus to a greater degree. In section 7 we consider whether correcting the individuals' forecasts for inefficiency accounts for differences in accuracy across forecasters, or whether some are the beneficiaries of superior signals. Although our focus is squarely on cross-sectional characteristics of the survey respondents, and how these characteristics depend on forecast efficiency, in section 8 we show the effect of efficiency correction for one forecaster, to illustrate the magnitude of the effects of correcting for inefficiency. We illustrate with the individual who responded to the most surveys over the sample period (98 of the possible 107 quarterly surveys between 1990:4 and 2017:2).

[2] King, Plosser, Stock and Watson (1991) found support for the 'great ratios' of Kosobud and Klein (1961) on data up to 1990, consistent with balanced growth paths (of the Solow-Ramsey model), whereas more recently two-sector models (such as, e.g., Whelan (2003)) predict that the key NIPA aggregates grow at constant but different rates.

Section 9 checks the robustness of our findings for a smaller sample of forecasters. Section 10 offers some concluding remarks.

2 Forecast Data: SPF Respondents' Forecasts

We use the US Survey of Professional Forecasters (SPF). The SPF is a quarterly survey of macroeconomic forecasters of the US economy that began in 1968, administered by the American Statistical Association (ASA) and the National Bureau of Economic Research (NBER). Since June 1990 it has been run by the Philadelphia Fed, renamed as the Survey of Professional Forecasters (SPF): see Croushore (1993). The SPF is made freely available by the Philadelphia Fed, allowing results to be readily reproduced and checked by other researchers. Its constant scrutiny is likely to minimize the impact of respondent reporting errors. An academic bibliography of the large number of published papers that use SPF data is maintained[3] and listed 101 papers as of January 2019.

We use the SPF multi-horizon forecasts of real GDP, consumption and investment from 1990:4 to 2017:2, i.e., from when it was administered by the Philadelphia Fed. It is tempting to use the earlier survey data, but the SPF documentation warns of its suspicion that the forecast identifiers may not have been uniquely assigned over the earlier period - newcomers may have been given the identifiers once associated with participants who have left the survey. Given our focus on individual behaviour, it seems preferable to forego the additional survey data.

Forecasts are made of the current quarter (i.e., the quarter in which the survey takes place), and of the quarterly values of the variables in each of the next four quarters, so that the longest-horizon quarterly forecast is of the same quarter of the year in the following year.

The latest survey we consider is 2017:2, so that the most recent target period we consider is 2018:2 (the four-quarter ahead forecast made in response to the 2017:2 survey). We stop here so that we have the vintage-values of all the actuals from two quarters after the reference quarters. (The last is the 2018:4 vintage data for reference quarter 2018:2).

In total we use 107 surveys from 1990:4 to 2017:2 inclusive. Table 1 provides details concerning the actual and forecast data. We consider the 50 individuals who made the most forecasts during this period. The average number of forecasts per person for this group was 55 (for each variable and at each forecast horizon, with a minimum of 31 and a maximum of 98). We could have widened our net to include more forecasters at the cost of including forecasters who made fewer forecasts, resulting in less precise estimates of the performance of these individuals.

The analysis of survey data at the individual level inevitably entails missing forecast data. We follow the literature in implicitly assuming that the data are missing 'at random', that

[3] http://www.phil.frb.org/research-and-data/real-time-center/survey-of-professional-forecasters/academic-bibliography.cfm.

is, 'that participation in the survey after recruitment is statistically independent of forecasters beliefs about inflation' (Engelberg, Manski and Williams (2011, p.1061)). [4] Because individuals are active respondents at different times, fair comparisons across forecasters in terms of accuracy or contrarianism require that we control for the different economic conditions prevalent at different times. Looking ahead, in section 4 we use measures of disagreement which control for the underlying level of variability (the measures given by (10) or (11), as opposed to (12)), and in calculating forecast accuracy in section 5, normalized forecast errors are used.

In the paper we report results for the current quarter ($h = 0$) and forecasts, and for the year-ahead quarter forecasts ($h = 4$). At the time the forecasts are filed - around the the middle of the quarter, respondents will have some information on the first month of the quarter, and the advanced estimates of the national accounts for the previous quarter will have been released. As emphasized by Lahiri and Sheng (2008) and Patton and Timmermann (2010) in their studies of the term-structure aspect of cross-sectional disagreement, we would expect the relative importance of information signals to diminish as the forecast horizon lengthens. As the horizon lengthens, the forecasts of stationary variables approach the long-run expectation. As a consequence, disagreement would be expected to lessen unless forecasters possess different priors about long-run means.

3 Forecaster Efficiency

3.1 Defining Forecaster Efficiency

We suppose that each forecaster i has an information set $\mathcal{F}_i$ where $\mathcal{F}_i \subseteq \mathcal{F}$, with $\mathcal{F}$ denoting all relevant information. Forecaster efficiency as used in this paper is due to Mincer and Zarnowitz (1969), and is related to the notion of calibration in the mathematical statistics literature, which has been discussed when there is diverse information by, e.g., Satopää, Pemantle and Ungar (2016) and Satopää (2018)). Forecaster i's prediction y_i is calibrated, or efficient, if:

$$y_i = E\left(y|\mathcal{F}_i\right). \tag{1}$$

That is, if the prediction is the conditional expectation of y given the forecaster's information set.[5] We do not know what information an individual has access to. To make (1) operational, we assume only that the forecaster knows her own forecast, by replacing $\mathcal{F}_i$ by y_i in (1). This

[4] Engelberg *et al.* (2011) argue that there is little evidence for this assumption, and argue that the assumption is also required for the analysis of aggregate (or consensus) forecasts.

[5] In our empirical work, we assume the survey forecasts are the conditional means of the respondents' underlying probability distributions. This assumption is standard in the literature. A number of authors have been able to consider the possibility that the respondents' point forecasts reflect other moments when histogram forecasts are also provided (see, e.g., Engelberg, Manski and Williams (2009), Clements (2009, 2010)).

is a conservative assumption, but satisfies the requirement that y_i is necessarily included in the forecaster's information set, $\mathcal{F}_i$.

This formulation makes clear that: i) we do not require that all forecasters have access to all relevant information, and ii) nor are the individuals' information sets necessarily the same.[6] We are interested in how they use their information sets: that is, whether (1) holds or not when $\mathcal{F}_i$ is specialized to $\mathcal{F}_i = y_i$.

As a simple illustration, suppose the data generating process is given by:

$$y_t = \alpha_1 y_{t-1} + \alpha_2 y_{t-2} + \varepsilon_t$$

that is, y_t is generated by a stationary autoregression of order 2, where ε_t is a white noise innovation on $\{\ldots, y_{t-2}, y_{t-1}\}$. But agent i's forecast of y_t is given by an AR(1) model $y_{i,t} = \gamma y_{t-1}$. When $\gamma = \gamma^* \equiv \gamma_1/\gamma_0$, the forecast is efficient or calibrated, where $\gamma_i = Cov\,(y_t, y_{t-i})$, so that γ^* is the first-order autocorrelation coefficient for an AR(2). In this illustration, the agent's information set for forecasting y_t is y_{t-1}, i.e., $\mathcal{F}_i = \{y_{t-1}\}$, is less than $\mathcal{F} = \{y_{t-1}, y_{t-2}\}$. The information set is used efficiently when $y_{i,t} = \gamma y_{t-1}$ and $\gamma = \gamma^*$, and inefficiently when $\gamma \neq \gamma^*$.[7]

A key question we address is the extent to which inefficient use of information by individual forecasters accounts for disagreement between forecasters, and differences in the accuracy of their forecasts. As stressed, we do not need to know what information an agent has access to.

3.2 Testing for Forecaster Efficiency

There is a large literature on testing forecaster rationality or efficiency, and the main approach is that of Mincer and Zarnowitz (1969) (MZ). For each individual i, we estimate the regression:

$$y_t = \delta_0 + \delta y_{i,t|t-h} + u_{i,t} \tag{2}$$

for a particular h, on all the forecasts made by i.[8] The MZ test of the null of optimality is a joint test that $\delta_0 = 0$ and $\delta = 1$. When $\delta = 1$ the correlation between the forecast error and

[6] For example, all forecasters may use a subset of publicly available information (one forecaster may use macro-indicators, another financial variables), or they may also have private information. As suggested by Satopää *et al.* (2016), differences in information sets (and therefore in y_i) across individuals may arise from differences in how individuals choose to use the information they have access to. This is perhaps the interpretation that best fits macro-forecasters, where most relevant information would appear to be 'public' and freely available (apart from the costs of processing/accessing, as stressed by the informational rigidities theories in the Introduction).

[7] Straightforward algebra shows the correlation between forecast $y_{i,t} = \gamma y_{t-1}$ and forecast error $e_t = y_t - \gamma y_{t-1}$ is zero when $\gamma = \gamma^*$.

[8] The dependence of δ_0 and δ on i and h is suppressed in the notation.

the forecast is zero:

$$Cov\left(y_t - y_{i,t|t-h}, y_{i,t|t-h}\right) = Cov\left(\left(\delta - 1\right) y_{i,t|t-h} + u_{it}, y_{i,t|t-h}\right).$$

Unless $\delta = 1$, the forecast and forecast error will be systematically related, and this correlation could be exploited to generate a superior forecast. For $\delta = 1$, the forecast error will be biased unless $\delta_0 = 0$. Note that $\delta_0 = 0$ and $\delta = 1$ is sufficient but not necessary for unbiasedness: $E\left(y_t - y_{i,t|t-h}\right) = 0$ when $E\left(y_{i,t|t-h}\right) = \delta_0 \left(1 - \delta\right)^{-1}$ (see, Holden and Peel (1990)).

From (2), $E\left(y_t | y_{i,t|t-h}\right) = \delta_0 + \delta y_{i,t|t-h} + E\left(u_{i,t} | y_{i,t|t-h}\right)$, where $E\left(u_{i,t} | y_{i,t|t-h}\right) = 0$, so that the MZ null that $\delta_0 = 0$ and $\delta = 1$ ensures calibration as given by (1): $E\left(y_t | y_{i,t|t-h}\right) = y_{i,t|t-h}$.

As noted by Bonham and Cohen (2001), pooled cross-section time-series regressions are sometimes used to improve the degrees of freedom and the power of the test. However, Bonham and Cohen (2001) show that pooling over individuals is invalid (following Zarnowitz (1985), and contrary to the claim made by Keane and Runkle (1990)), and for this reason we run individual regressions for each forecaster.

Patton and Timmermann (2012) suggest extending MZ to the optimal revision regression (henceforth ORR), when fixed-event forecasts (see, e.g., Nordhaus (1987) and Clements (1995)) are available, as here. Write the short horizon forecast (e.g., $h_1 = 1$) as:

$$y_{t|t-h_1} \equiv y_{t|t-h_H} + d_{t|h_1,h_2} + \ldots d_{t|h_{H-1},h_H} \tag{3}$$

where $h_1 < h_2 < \ldots < h_H$, with h_H the longest horizon forecast of the target y_t, and $d_{t|h_j,h_{j+1}} = y_{t|t-h_j} - y_{t|t-h_{j+1}}$. Then regressing y_t on $y_{t|t-h_1}$, where we replace $y_{t|t-h_1}$ by $y_{t|t-h_1} = y_{t|t-h_H} + \sum_{i=1}^{H-1} d_{t|h_i,h_{i+1}}$, and allowing a free coefficient on each of the components of $y_{t|t-h_1}$, results in:

$$y_t = \delta_0 + \delta_H y_{t|t-h_H} + \sum_{i=1}^{H-1} \delta_i d_{t|h_i,h_{i+1}} + u_t, \tag{4}$$

and the null hypothesis is that $\delta_0 = 0$ and $\delta_1 = \delta_2 = \ldots = \delta_H = 1$. Under the null, the error for the short-horizon forecast $y_{t|t-h_1}$ is uncorrelated with all forecasts of the target y_t made at earlier times (and hence on smaller information sets). Equation (4) becomes $y_t = y_{t|t-h_1} + u_t$ under the null hypothesis. Hence the ORR test has power to reject the null against the alternative that the short-horizon forecast error is systematically related to revisions in earlier forecasts of the target value.

A disadvantage of the ORR test arises when there are missing forecast observations. The failure to respond to a single survey will reduce the number of observations used to estimate (4) by 5 when $H = 4$, but only by 1 in the case of the MZ regression (2). For this reason we report MZ regression results for specific forecast horizons.

Patton and Timmermann (2012) propose a variant of the MZ test, which has the advantage

that one need not take a stand over the vintage of data being targeted: Keane and Runkle (1990) criticize the use of revised data, and suggest it may be responsible for the erroneous rejection of rationality. The actual value of y_t can be replaced by a short-horizon forecast, say, $y_{t|t-h_1}$:

$$y_{t|t-h_1} = \delta_0 + \delta y_{t|t-h_2} + u_t \tag{5}$$

where $h_2 > h_1$, or for ORR:

$$y_{t|t-h_1} = \delta_0 + \delta_H y_{t|t-h_H} + \sum_{i=2}^{h_H-1} \delta_i d_{t|h_i,h_{i+1}} + u_t \tag{6}$$

when $h_H > h_{H-1} > \ldots > h_1$. As noted by Patton and Timmermann (2012, p.6), (5) now tests the internal consistency of the forecasts $y_{t|t-h_1}$ and $\delta y_{t|t-h_2}$. We prefer instead to use actual values as the dependent variables in MZ and ORR regression tests, but to use real-time vintage estimates of the actual values.

Researchers often use vintages released soon after the reference quarter, rather than the latest-available vintage at the time of the investigation. This is because the 'fully-revised' data will typically include benchmark revisions, rebasings, and other methodological changes to the way the data are collected and measured, which could not have been foreseen when the forecast was made.[9] The Real Time Data Set for Macroeconomists (RTDSM) maintained by the Federal Reserve Bank of Philadelphia (see Croushore and Stark (2001)) greatly facilitates the use of real-time data in macro analysis and forecasting research. For the forecast efficiency tests the actual values are either the vintage-values published two quarters after the reference quarter, or the first estimates. The second quarterly estimates include more information than the initial 'advance estimates' (available one month after the reference quarter). But as explained below, the initial estimates allow an efficiency correction to be calculated in real time. As shown in section 3.4, the null of efficiency is often rejected whichever of the two vintages is used.

3.3 Efficiency-Corrected Forecasts

We can use the MZ-regression run on an individual respondent's forecasts to 'efficiency correct' (EC) those forecasts. The *in-sample* EC forecasts are given by the predicted values from (2), $\hat{y}_{i,t|t} = \hat{\delta}_0 + \hat{\delta} y_{i,t|t}$, when $h = 0$ and the forecast errors of the corrected forecasts are given by $\hat{u}_{i,t}$. By the properties of OLS, these forecast errors are orthogonal to the predicted values - the corrected forecasts. In this sense we have carried out a forecast-efficiency correction. By construction, the sum of squares of the residuals - the corrected forecast errors - is no larger

[9] See, e.g., the review articles by Croushore (2011a, 2011b) as well as Landefeld, Seskin and Fraumeni (2008) and Fixler, Greenaway-McGrevy and Grimm (2014).

than the sum of squared forecast errors of the reported forecasts. The EC-forecasts are more accurate on squared-error loss.[10]

However, the in-sample correction is not real-time, in the sense that the survey t forecast will be corrected using regression estimates calculated from a sample that includes future forecasts and actual values. This is sometimes referred to as 'look-forward' bias. We implement the correction in real time as follows (c.f., Arai (2014)). Let n^* denote a minimum number of observations used to generate initial estimates of (2). Then for $t \leq n^*$, $\hat{y}_{i,t|t} = \hat{\delta}_{0,n^*} + \hat{\delta}_{n^*} y_{i,t|t}$, that is, the coefficients are estimated on data up that available at time n^*, and the correction is in-sample. For $t > n^*$, we calculate the correction using only the sequence of forecasts and actual values available up to that point, $y_{\tau-1|\tau-1}$ and $y^{\tau}_{\tau-1}$, for $\tau = t_1, \ldots, t$, where $y_{\tau-1|\tau-1}$ is the forecast of y in period $\tau - 1$ made at time $\tau - 1$ (for $h = 0$), and $y^{\tau}_{\tau-1}$ is the value of $y_{\tau-1}$ available at time τ. At survey t, the latest available forecast and corresponding actual value are therefore $y_{t-1|t-1}$ and y^{t}_{t-1}. The MZ regression is then:

$$y^{\tau}_{\tau-1} = \delta_0 + \delta y_{\tau-1|\tau-1} + u_{\tau-1} \tag{7}$$

for $\tau = t_1, \ldots, t$. We calculate the efficiency-adjusted forecast of period t using the parameter estimates, as:

$$y_{t|t} = \hat{\delta}_{0,t} + \hat{\delta}_t y_{t|t} \tag{8}$$

where $\hat{\delta}_{0,t}$ and $\hat{\delta}_t$ are the estimates of (7) based on data available at survey t. We estimate (7) on an expanding window of data as t increases. (Alternatively, a rolling window of data could be used, discarding earlier forecasts and actual values.)

Our approach means that the efficiency correction for all but the first n^* forecasts is real time. A forecaster could have applied the correction to her forecasts at each point in time. We set $n^* = 10$, so that for an average forecaster (with over 50 forecasts) in excess of 80% of the efficiency corrections to the forecasts are real time.[11] Notice that the use of first-release data means that at time t we can use data up to an including last period's survey forecast to calculate the correction, because y^{t}_{t-1} is known. This would not be the case were we to use the second quarterly estimates, or more mature data.

We have described the efficiency correction for the current-quarter $h = 0$ forecasts. The correction is also applied to the year ahead $h = 4$ forecasts, but then the real-time implementation

[10] We assume squared-error loss throughout, although there is a literature suggesting forecasters loss functions may be asymmetric: see, e.g., Elliott, Komunjer and Timmermann (2005), Elliott, Komunjer and Timmermann (2008), Patton and Timmermann (2007) and Lahiri and Liu (2009). Asymmetric loss might be more natural for inflation forecasting, but in any case, it would not be straightforward to accommodate asymmetry in the analysis.

[11] In fact although we only use survey data from 1990:4 to derive the main results (on accuracy and contrarianism), we do use pre 1990:4 data, where available, to initialize the efficiency correction. Hence for respondents who make n^* or more forecasts prior to 1990:4, the correction is wholly real time.

requires that at survey time t the latest forecast and actual value pair available for estimating the equivalent of (7) are $y_{t-1|t-5}$ and y_{t-1}^{t}. That is, the $h = 4$ forecast of y_{t-1} made to the $t-5$ survey. To illustrate: for correcting the $h = 4$ forecast from the 1995:1 survey, the latest survey used to estimate the correction will be the 1993:4 survey. This will supply the $h = 4$ forecast of the 1994:4 target period.

Forecasts could be corrected for inefficiency based on other test regressions, such the ORR of Patton and Timmermann (2012), although we use the MZ regression, as in (7).

3.4 Empirical Findings

In table 2 we summarize the results of running (2) for each individual respondent for real-time actual values: both the initial 'advance' estimates, and the vintage available in the RTDSM (see table 1) two quarters after the reference quarter. As an example, in the second case, the 2010:1 value is taken from the 2010:3 data vintage. Using the second-quarterly release actual values, the null is rejected at the 5% level for over a half of forecasters for consumption at $h = 0$, and for around a quarter for investment and output. The rejection rates are well in excess of a half for all variables at $h = 4$. Using the advance estimates as the actual values, the evidence against the null of efficiency is strengthened, with rejections for higher proportions of respondents. Table 3 provides results for each individual respondent for $h = 0$ and $h = 4$ and for the three variables, and indicates the differences in the available samples of forecasts across individuals.

There is *prima facie* evidence that over a half of the individuals do not make efficient forecasts at $h = 4$, and this finding is not specific to a particular vintage of data, but holds for two reasonable choices of 'real-time' actual values.

As argued in the Introduction, rejections of forecast efficiency could be dismissed on the grounds that statistical significance does not necessarily mean the implied departure from rationality is of economic importance, and one could quibble at the use of the 5% significance level that underlies the calculation of the proportion for which we reject in table 2. In response to this, we focus on the importance of the departures from efficiency in terms of key characteristics of forecaster behaviour, such as whether inefficiency accounts for the findings of contrarianism and differences in forecast accuracy. The following sections consider the extent to which these inefficiencies are able to explain observed patterns of inter-forecaster disagreement and accuracy.

Before doing so, figure 1 provides an indication of the magnitude of the changes in the forecasts from the out-of-sample efficiency corrections. The figure reports the cross-sectional mean absolute change between the reported and corrected forecasts at each survey data, for $h = 0$ and $h = 4$.[12] Although the forecasts are annualized percentage growth rates, the changes

[12] To aid interpretability, some smoothing is undertaken: each point is a centred moving average with one lead

are sizeable. The average absolute change for output growth is around a quarter of a percentage point, and close to one percentage point for investment over the last 10 years or so. Generally the corrections are larger for the longer horizon forecasts, and although there is some business cycle variation it is not pronounced, and the corrections are a feature of the whole sample period.

4 Disagreement

There is a large literature on disagreement.[13] However with few exceptions each variable is considered in isolation. For our purpose the multivariate measure of disagreement of Banternghansa and McCracken (2009) is an attractive option. The multivariate measure takes into account the forecaster's beliefs about the inter-dependencies between the variables implicit in the vector of forecasts. Any such inter-dependencies are lost when the variables are considered in isolation. Suppose at time $t-h$ we have a set of forecasts of time t for individuals $i = 1, \ldots, N_{t,h}$. Banternghansa and McCracken (2009) define the cross-sectional forecast covariance matrix as:

$$S_{t|t-h} = N_{t,h}^{-1} \sum_{i=1}^{N_{t,h}} \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right) \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)' \tag{9}$$

where $y_{i,t|t-h}$ is the vector of forecasts made by i (at time $t-h$ for a target y_t), and $\overline{y}_{t|t-h} = N_{t,h}^{-1} \sum_{i=1}^{N_{t,h}} y_{i,t|t-h}$, the cross-sectional average. Then they define their multivariate disagreement measure for individual i forecasting the vector y_t at forecast origin $t-h$ as the Mahalanobis distance:

$$D_{i,t|t-h} = \sqrt{\left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)' S_{t|t-h}^{-1} \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)}. \tag{10}$$

When $S_{t|t-h}$ is restricted to being a diagonal matrix, with the diagonal consisting of the cross-sectional variances, $D_{i,t|t-h}$ simplifies to:

$$D_{i,t|t-h} = \sqrt{\sum_{j=1}^{n} \frac{\left(y_{j,i,t|t-h} - \overline{y}_{j,t|t-h} \right)^2}{S_{jj,t|t-h}}} \tag{11}$$

that is, it is the sum of agent i's squared deviations for each variable, where each is scaled by the cross-sectional variance. Here, j indexes the n variables, $y_{i,t|t-h} = \left[y_{1,i,t|t-h} \ \cdots y_{j,i,t|t-h} \cdots y_{n,i,t|t-h} \right]'$, and $S_{jj,t|t-h}$ is the j-th diagonal element of $S_{t|t-h}$. When S is diagonal, the cross-sectional covariances do not affect the calculation of disagreement.

and lag.

[13] See, *inter alia*, Zarnowitz and Lambros (1987), Bomberger (1996), Rich and Butler (1998), Capistrán and Timmermann (2009), Lahiri and Sheng (2008), Rich and Tracy (2010) and Patton and Timmermann (2010).

Finally, if we set S to the identity matrix:

$$D_{i,t|t-h} = \sqrt{\sum_{j=1}^{n} \left(y_{j,i,t|t-h} - \overline{y}_{j,t|t-h} \right)^2} \tag{12}$$

no allowance is made for some variables being inherently more difficult to forecast than others, or for the underlying variability to change over time. Both these effect are captured by (11) (or (10)). $S_{jj,t|t-h}$ will tend to exceed $S_{kk,t|t-h}$ if j denotes investment and k consumption, for example. At times of greater uncertainty, the deviation from the consensus will likely be larger than in more quiescent times. These larger deviations will be reduced by larger than average cross-sectional variances at those time. The use of $S_{t|t-h}$, calculated as in (9) ought to reduce distortions from respondents being active survey participants at different times. This is potentially important because quite different economic conditions prevailed over the period 1990 – 2017, and on average respondents filed returns to around half the possible surveys.

The above arguments suggest using either (10) or (11). The difference between the two can be illustrated with a simple example. Suppose y consists of just two variables, and for forecaster A at time $t-h$, $y_{A,t|t-h} - \overline{y}_{t|t-h} = (1,1)'$, so that this respondent's forecasts of both variables differ from the consensus forecasts by a positive amount (of 1 unit). For simplicity, suppose that the cross-sectional variances of the forecasts are one for both variables - S has ones on it's diagonal. Then the Euclidean measure of disagreement given by (11) is $\sqrt{1^2 + 1^2}$.

Suppose the diagonal elements of S are still unity, and the off-diagonal element is ρ. If $\rho = 0.9$, so the cross-sectional covariance between the other respondents' forecasts of the two variables (equivalently, forecast errors) is positive, then equation (10) for D gives $D = \sqrt{2/1.9}$, which is less than the $\sqrt{2}$ from using (11). This is because forecaster A agrees with the consensus view that the variables are positively correlated: she over-predicts both variables relative to the consensus (of course she would still agree with the consensus if she under-predicted both variables).

Suppose a second forecaster (B) disagrees with the consensus view regarding the relationship between the two variables, simultaneously over-predicting the first variable (relative to the consensus) and under-predicting the second, at odds with the consensus view that the variables are positively correlated ($\rho = 0.9$). For this forecaster, $y_{j,t|t-h} - \overline{y}_{t|t-h} = (1,-1)'$, say. Using S diagonal, Forecaster A and B disagree by the same amounts, because for Forecaster B we also have $D = \sqrt{2}$. But forecaster B is penalized using (10) (with a non-diagonal S) for being out of kilter with the consensus, and $D = \sqrt{20}$.

Standard measures of disagreement consider the forecasters *en masse*, and correspond to taking the square roots of the diagonal elements of (9) as the cross-sectional standard deviations, for example. However, our primary interest is not in measuring disagreement *en masse*, but calculating the extent to which each individual disagrees with the consensus. Equations (10) and

(11) provide two alternative measures of individual-level disagreement. As shown, in principle disagreement will be reduced for a forecaster if both her deviations are of the same sign when the consensus forecast covariance is positive - that is, if the individual shares the consensus view of how the variables are related. In practice, unless the cross-sectional forecast covariance is large, the two are unlikely to deliver similar results, and that transpires to be the case in our application.

4.1 Individual Multivariate Disagreement Estimates

We calculate the average disagreement for each individual (the average of eqn. (10)) across all the 107 surveys from 1990:4 and 2017:2 to which the individual responded, for $h = 0$ and $h = 4$, respectively. We calculate the multivariate disagreement measure which takes into account the correlations between variables, and we also calculate the measure assuming $S_{t|t-h}$ is diagonal, and so simply sums the scaled disagreement for each variable. We do not report detailed results for a diagonal $S_{t|t-h}$ because they qualitatively similar to using (10) with $S_{t|t-h}$ calculated as in (9).

We test whether differences across forecasters in terms of disagreement are systematic, in the sense that some respondents' forecasts tend to systematically differ by more or less from the consensus than those of others. The alternative would be that overall disagreement at any point in time is as likely to be due to any one forecaster disagreeing with the consensus as any other forecaster. Systematic differences between forecasters in terms of the extent to which they disagree with the consensus would count as evidence against the proposition that forecasters are identical/interchangeable. For the $h = 0$ forecasts we report a formal test of whether the population means of the $D_{i,h}$ differ across individuals, i.e., of the null that $H_0 : \mu_{i,h} = \mu_{m,h}$ versus $H_1 : \mu_{i,h} \neq \mu_{m,h}$ for individuals i, where m is the individual with the average level of disagreement at $h = 0$, and where $\mu_{i,h}$ denotes a population mean. The $\left\{D_{i,t|t-h}\right\}$ are regarded as realizations, and we calculate t-tests of the equality of two population means allowing the variances to be unequal.

For now, consider the reported (i.e., uncorrected forecasts). The left-hand-side of table 4 reports the average (across t) multivariate disagreement estimate (equation 10) for each respondent, for $h = 0$ (column 2), as well as the ranks for $h = 0$ and $h = 4$ (columns 4 and 5), and the p-value of each individual having the same (population) disagreement as the median forecaster (column 3). The p-value is calculated such that a larger value suggests larger than average disagreement, and a value close to zero indicates a smaller than average value of disagreement. In a two-sided test at the 10% level the null is rejected for a half the forecasters (p-values exceed 0.95 or are less than 0.05), and at the 5% significance level the null is still

rejected for nearly 40% of the respondents.[14] The test results show that the variation in the average measure across individuals, from a low of 0.896 to a high of 3.103, for $h = 0$, does constitute statistically significant differences. We also rank each forecaster for $h = 0$ and $h = 4$, to allow a comparison across horizons. Below we test whether the ranks are correlated.

A method of assessing the persistence in individual forecasting behaviour, which doesn't require pairwise comparisons of each individual to the average forecaster, is to compare the ranks of forecasters based on their average levels of multivariate disagreement in the first and second halves of the sample. We split the sample 1990:4 to 2017:2 in half, and refer to the first (or earlier) and second (or later) samples. When an individual makes too few forecasts in one of the two samples to reliably estimate disagreement, that individual is not included in the tests we report comparing the behaviour of individual forecasters across the two samples. The test of whether the rankings are the same over the two sub-samples is given by Spearman's rank correlation coefficient. This tests whether individual-level disagreement in the two samples is correlated or not without relying on there being a linear relationship between disagreement in the two periods.[15] Spearman's rank correlation r lies between -1 and 1, where 0 indicates no relationship, and is calculated as:

$$r = 1 - \frac{6R}{N\left(N^2 - 1\right)} \tag{13}$$

where R is the sum of squared differences between the ranks (e.g., of the forecasters in the first sample, and in the second sample). We follow the literature and calculate the Fisher transformation:

$$F\left(r\right) = \frac{1}{2} \ln \frac{1+r}{1-r}$$

such that $z = F\left(r\right) . \sqrt{\frac{N-3}{1.06}} \sim N\left(0,1\right)$ under the null of statistical independence.

In addition to comparing forecaster behaviour over time, in terms of disagreement, we also address the constancy of forecaster behaviour across horizon, and the effects on these comparisons of adopting a true multivariate measure as opposed to summing disagreement for the individual variables. Table 5 reports rank correlation tests of the null hypotheses that there is no relationship between forecaster disagreement: *i*) across time - between the earlier and later periods - for a given h ($h = 0, 4$): Panel 1A; *ii*) between short ($h = 0$) and long-horizon

[14] If instead of using (10) we use (11) the numbers of rejections are essentially unchanged.

[15] Here and elsewhere in the paper, when we test for persistence in differences in disagreement or accuracy across individuals we take the disagreement or accuracy estimates at face value. We do not attempt to make an allowance for the fact that the individual measures are estimates and in some cases rely on estimated efficiency corrections. In so doing we follow e.g., Boero, Smith and Wallis (2015) who consider the persistence of individual forecaster's relative uncertainty. It might be possible to allow for some of the sources of uncertainty using a bootstrap (see, e.g., Curran (2015)), but we do not attempt to do so here. It seems likely that some of the decisive rejections of the null we report would be unaffected but this is a conjecture.

forecasts ($h = 4$), across all surveys and in each of the two sub-periods: Panel 1B. We carry out *i*) and *ii*) for the multivariate disagreement measure (using 'S'), and for the sum of the individual variable measures (using 'Diag. S').

For each test, the table records the rank correlation coefficient (equation 13), as well as the probability of z being at least as large as we obtained if the null hypothesis (of a zero correlation) is true. Probabilities less than 0.025 or greater than 0.975 indicate rejections of the null in a two-sided test at the 5% level. (High probabilities suggest a negative relationship, and low probabilities a a positive relationship).

We reject the null of no relationship in the rankings of disagreement between the earlier and later sample periods for both forecast horizons, and for both disagreement measures (referred to as 'S' and 'Diag. S' in the table. Note that for the diagonal measure the probability of obtaining a z statistic at least as large is 0.029. The other rejections are at much more stringent levels.). In terms of short and long-horizon disagreement, we reject the null of no relationship in the rankings across forecasters for the whole forecast sample and both sub-samples, whether we use the multivariate or 'univariate' measure (i.e., the sum of the disagreement measures for each variable).

The micro-level evidence strongly suggests that forecasters are not interchangeable in terms of their degrees of conformity with the consensus. Moreover, these results are generally not sensitive to whether the disagreement measure is adjusted for the degree of agreement about how the economy operates. Allowing an offset to disagreement from agreement regarding how the economy operates is largely inconsequential for determining the degree of relative contrarianism of individual forecasters.

4.2 Efficiency-corrected Disagreement Estimates

To what extent does forecast disagreement reflect a failure of the assumption that forecasters make rational-expectations forecasts given their information sets? In the context of the noisy information model, for example, it might be the case that agents receive equally precise signals but produce forecasts which are efficient to varying degrees. Alternatively, disagreement might result from rational forecasters receiving heterogeneous signals.

To gauge the extent to which differences in individual agent contrarianism are attributable to heterogeneous signals or differences in the efficiency with which those agents generate their forecasts, we re-run the calculations in tables 4 and 5, having first corrected the forecasts using the real-time efficiency-correction procedure described in section 3.3. That is, each time a forecast is made, we efficiency-correct that forecast using that respondent's past history of forecasts and actual values.

The results in the right-hand panel of table 4, and the second panel of table 5, are based on the corrected forecasts. Consider table 4. Firstly, the proportion who differ from the median

forecaster is now a third (at the 10% level) compared to a half when the correction is not applied. Hence inefficient use of information explains some of the significant differences in contrarianism across individual respondents. Secondly, the rankings change, particularly at the $h = 4$ horizon. For example, the most contrarian at $h = 4$ (rank 50) becomes the second least contrarian (rank 2) after the efficiency correction has been applied.

Table 5 shows that the evidence of persistence in contrarianism across the two sample periods (see Panel 2A.) remains when the forecasts are efficiency corrected. The finding that more (less) contrarian forecasters in the first period remain so in the second period is not solely due to the inefficient use of information. In terms of the constancy of forecaster behaviour across horizon, efficiency correction breaks the link between the short and long-horizon forecasts. It is no longer the case that agents who make more contrarian forecasts at one horizon are more likely to do so at the other horizon, whether we consider the whole period, or either of the two sub-periods.

The findings do not depend on whether the disagreement measure is adjusted for the degree of agreement about how the economy operates, i.e., whether the measure is (10) or (11).

5 Forecast Accuracy

In this section we consider the micro-level evidence for the proposition that individuals' forecasts are equally accurate. The basic noisy information model suggests that forecasters are essentially identical: they receive homogeneous signals, have the same model of the economy, and use that information efficiently to generate their expectations. We find evidence against the assumption of equal accuracy for the short-horizon forecasts. To what can we attribute the differences in forecast accuracy? It does not appear to be the case that the differences in forecast accuracy are solely due to some forecasters using their information sets more efficiently than others: the differences are still apparent for the short-horizon forecasts when the forecast-efficiency correction is implemented.

Equal predictive accuracy is assessed in terms of whether the more (less) accurate forecasters over a given period remain the more (less) accurate over a subsequent period. The forecast accuracy measures are the trace and the determinant of the Mean-Squared Forecast-Error Matrices (MSFEMs) for $h = 0$ and $h = 4$ forecasts. The determinant is a multivariate measure, whereas the trace simply sums the individual-variable MSFEs. Having a single measure of forecast accuracy - as opposed to one for each variable - make the comparisons more manageable, and the multivariate aspect of the determinant measure is in tune with our approach to measuring disagreement. Clements and Hendry (1993) propose the determinant as an invariant measure of forecast accuracy for 1-step forecasts: it is invariant to forecasting linear transformations of the vector of variables. For $h = 4$ an invariant measure would be the Generalized Forecast Error Second Moment Matrix (GFESM), as discussed by Clements and Hendry (1993), although we have relatively small samples of forecasts at our disposal to calculate such a measure (but see

Hendry and Martinez (2017)). Komunjer and Owyang (2012) propose a multivariate loss function which allows for dependence between the different variables' forecast errors, and Sinclair, Stekler and Carnow (2015) also present a multivariate analysis (evaluating a vector of forecasts of a number of variables against a vector of outcomes by Mahalanobis distance).

We adjust for individuals forecasting during different economic conditions by controlling for differences over time in the average accuracy of all forecasters, following D'Agostino, McQuinn and Whelan (2012) and Clements (2014). Not controlling for the degree of difficulty in forecasting at time t might distort the inter-personal comparisons of forecast accuracy. As an extreme example, consider investment around the time of the recent Crisis. Investment fell by about 12% in 2009:1 relative to 2008:4 (not annualized). The magnitude of the fall was unforeseen, and those who happened to respond to the 2008:1 survey registered much larger 4-step ahead forecast errors than those made in response to any other survey.

Letting $e_{i,n,t+h|t}$ denote the forecast error made by individual i, for variable n, in response to forecast survey t, for period $t+h$, we calculate the normalized forecast errors as:

$$\widetilde{e}_{i,n,t+h|t} = \frac{e_{i,n,t+h|t}}{\sqrt{\frac{1}{N_{t,h}} \sum_{j=1}^{N_{t,h}} e^2_{j,n,t+h|t}}} \tag{14}$$

where $N_{t.h}$ is the number of respondents to survey t, so that the denominator is the cross-section RMSE. Then letting $\widetilde{e}'_{i,t+h|t} = \left[\widetilde{e}_{i,1,t+h|t}\ \widetilde{e}_{i,2,t+h|t}\ \widetilde{e}_{i,3,t+h|t}\right]$ denote the vector of normalized forecast errors results in the adjusted MSFE matrix for respondent i (at horizon h) of:

$$\frac{1}{n_i} \sum_{t \in N_i} \widetilde{e}_{i,t+h|t} \widetilde{e}'_{i,t+h|t} \tag{15}$$

where the summation is over all the surveys to which i responded, given by the set N_i, and n_i is the number of elements in N_i.

As for the forecast efficiency tests, the actual values used to calculate forecast errors are either the initial estimates or the vintage-values published two quarters after the reference quarter.

5.1 Forecast Accuracy Results

Table 6 reports Spearman rank tests of the null that the rankings across the two sub-samples are unrelated. As expected, normalizing the forecast errors using (14) to account for the forecasters being active survey participants during different economic conditions significantly affects the findings. Use of the 'raw' or un-normalized forecast errors to calculate the forecast accuracy measures (Panel A) suggests no evidence against the null of no persistence across time at the

5%, for both measures of accuracy, and for both horizons. Normalizing the forecast errors (Panel B) results in the clear rejection of the null for the short-horizon forecasts, and for $h = 4$ using the determinant (of equation (15)). We interpret this as suggesting the use of the raw forecast errors is likely to be misleading when forecasters face very different conditions, and that some forecasters generate more accurate forecasts than others.

Figure 2 shows how the difficulty in forecasting changes over time, by plotting the cross-sectional root mean squared forecasts errors (RMSFEs) for the three variables separately, and for the two horizons. These are the denominators of (14), except that we have averaged the survey quarter value over the previous and subsequent quarters to provide a smoother estimate. The RMSFEs are twice as large in some periods as in others, with the recent Crisis period exemplifying difficult conditions. (The spikes for $h = 4$ appear to lead those for $h = 0$ because the horizontal axes shows the survey quarter (and not the target period).

Of interest is whether the rejection of the null - of no persistence in the rankings across forecasters between the two periods - is due to an inefficient use of information. Using the (normalized) efficiency-corrected forecasts (Panel C) suggests no evidence of persistence in the year-ahead forecasts, and more nuanced findings for the $h = 0$ forecasts: we do not reject at the 5% level, but we do at less stringent levels, such as the 10% level.

In summary, testing using a 5% significance level suggests forecast inefficiency accounts for the persistence in the rankings of forecasters which we observe.

6 Are More Contrarian Forecasters Less Accurate Forecasters?

In this section we consider whether the more contrarian forecasters tend to be the more accurate forecasters. Such would be the case if some forecasters received superior signals (in terms of the noisy information model, for example) and so simultaneously distance themselves from the crowd and record more accurate forecasts.

Two measures of forecast accuracy are considered, the trace and determinant of the MSFEM for the three variables, based on forecast errors scaled by the estimated difficulty of forecasting. The multivariate disagreement measures are given by equation (10), and also make an allowance for some periods being inherently more difficult to forecast than others, as well as including an offset for agreement over how the economy operates (i.e., S is non-diagonal).

The Spearman rank correlation test results recorded in table 7 indicate a statistically positive relationship between disagreement and squared-error loss, for both horizons. This suggests that more contrarian forecasters make less accurate forecasts. This is at odds with the conjecture that superior private information would simultaneously distance some forecasters from the consensus and result in their forecasts being more accurate.

This finding holds up when the forecasts are corrected for inefficiency (as evident from the second panel of table 7).

7 Correcting Forecast Inefficiencies

In this section we again consider the relationship between the efficient use of information and forecast accuracy. In section 5 we found persistence in accuracy rankings of agents' short-horizon forecasts across the two sample periods. If all forecasts were efficiency corrected, but the evidence for persistence was considerably weakened: we do not reject the null of no persistence at the 5% level, but we do at the 10% level. At the more stringent 5% level, forecast inefficiencies explain differences in accuracy, without the need to assume some forecasters have better information or models.

In this section we approach the issue from a different angle. Instead of considering accuracy rankings across different time periods, we consider the relationship between the magnitude of improvement from correcting for forecast inefficiency and the accuracy of the reported forecasts. The finding of a negative correlation across individual forecasters, such that less accurate forecasters tend to benefit from larger improvements in accuracy from removing inefficiency, would suggest that differences in accuracy are attributable to some forecasters generating inefficient forecasts. On the other hand, no correlation between the two would suggest forecast efficiencies do not explain the differences in forecast accuracy, consistent with differences in accuracy across forecasters reflecting signal heterogeneity or different models. That is, if accurate forecasters do not use their information more efficiently, then more accurate forecasters are presumably the beneficiaries of more informative signals or use superior forecasting models or ways of forecasting.

The results described in table 8 use the real-time efficiency correction described in section 3.3, and used hitherto. The correlations reported in the table are negative for all three variables, and the null hypothesis of no relationship is clearly rejected. The improvement in accuracy from efficiency correction is statistically related to the inaccuracy of the reported forecasts. Forecast inefficiencies explain some of the inter-forecaster differences in forecast accuracy.

8 Illustrative Results for one Survey Respondent

Our primary focus is on the differences in forecast behaviour across individual respondents, and whether these differences in characteristics (forecast accuracy, contrarianism) depend on whether or not information is used efficiently. However, in this section we illustrate the effects of correcting for inefficiency using as an example the individual who responded to the most surveys over the sample period (98 of the possible 107 quarterly surveys between 1990:4 and 2017:2).

Firstly, figures 3 and 4 give the time series of the reported and real-time corrected forecasts for this individual for horizons $h = 0$ and $h = 4$ respectively. For the $h = 0$ forecasts the corrections are large around the time of the Crisis in 2009. For the year ahead forecasts the

effect of the correction is to reduce the variability of the forecasts. Because the forecasts are real time or out-of-sample, as explained in section 3.3, they need not necessarily improve the accuracy of the forecasts. Table 9 Panel A shows the corrected forecasts depicted in the figures are 5 to 9% more accurate on RMSFE than the reported forecasts for consumption and output growth for this individual. The table also records the markedly larger improvements resulting from an 'in-sample' correction (described in section 3.3). This is not feasible in real time, and the results reported elsewhere in this paper are all real time. Nevertheless, the in-sample correction reduces the RMSFE for consumption at both horizons, and for output at $h = 4$, by 12%. Calculating the correction in real-time tempers the improvements in accuracy for this individual, but does not remove them.

Panels B and C of table 9 show how the ranking of this individual changes as a result of efficiency correction, both in terms of (multivariate) accuracy and disagreement. In terms of accuracy (Panel B), this is one of the less accurate respondents, ranked as the 37th or 40th (out of 50) most accurate forecaster for $h = 0$, depending on the measure of accuracy. Similar ranks are found at $h = 4$. Efficiency correction has little affect on this forecaster in terms of short-horizon forecasts, but moves the forecaster to around the first quintile for $h = 4$. Efficiency correction has a similar effect reducing the long-horizon ($h = 4$) contrarianism of this respondent, but leaving the rank for the $h = 0$ largely unchanged.

This section illustrates the changes to one forecaster's performance from efficiency correcting the forecasts (of all the respondents). As stressed, the key results in the paper relate to the inter-forecaster comparisons.

9 Robustness

As explained in section 2, our results are for the 50 individuals who made the most forecasts in response to the 107 surveys from 1990:4 to 2017:2 (inclusive). Selecting the top 50 gives an average number of forecasts per person of 55, and a minimum of 31 and a maximum of 98. If we halve the number of forecasters the average per respondent rises to 71, and the minimum to 52. This ought to increase the reliability of the estimates of individual-level contrarianism and accuracy, especially when we consider sub-samples. On the downside we have only half the number of forecasters for the inter-forecaster comparisons.

The tables we re-calculate for the sample of 25 respondents are tables 5 and 6. Table 10 shows that the null of no relationship in the rankings of disagreement between the earlier and later sample periods is again rejected for both forecast horizons (compare to table 5), although now the rejection for $h = 4$ depends on the use of the of the disagreement measure with the non-diagonal 'S'. We suggested in section 4 that this might be a more meaningful measure of disagreement. Hence the micro-level evidence that forecasters are not interchangeable in terms of their degrees of conformity with the consensus holds for the sample of 25 forecasters.

Table 5 suggested the evidence of persistence in contrarianism across the two sample periods (see Panel 2A.) remained after efficiency correction. The same is true for the sample of 25, except that the null is not rejected at the 5% level for the $h = 0$ horizon, but it is at the 7% level (non-diagonal S), and the results are unchanged for $h = 4$.

As before, after efficiency correction, more contrarian forecasts at one horizon are not more or less likely to be so at the other horizon.

As to forecast accuracy, we still find that the null of no persistence in the rankings across forecasters is rejected once the forecast errors are normalized (compare table 11 for the 25 forecasters with the original table 6). For the 50 forecasters, we were unable to reject the null at the 5% level after the forecasts had been efficiency corrected. For the 25 forecasters we are unable to reject at any reasonable significance level.

In summary, reducing the number of forecasters leaves the results concerning multivariate disagreement essentially unchanged. The results for the 25 forecasters support the finding that efficiency correction accounts for the persistence in forecast accuracy rankings.

10 Conclusions

Models of expectations formation often assume that agents act rationally subject to certain constraints. For example, the IR literature stresses information rigidities, and assumes agents generate efficient (or rational) forecasts given their information sets. Moreover, the aggregate-level evidence on expectations formation of Coibion and Gorodnichenko (2012, 2015) is broadly consistent with the baseline version of the noisy information model. The baseline model supposes that forecasters are effectively identical or interchangeable. We find that the micro-level evidence suggests approximately a half of forecasters do not generate rational forecasts given their information sets. Nor are the forecasters essentially identical, either in terms of their degree of contrarianism, or predictive ability.

We found persistent differences across individuals in terms of their degree of contrarianism, that is, in terms of the extent to which they stand apart from the crowd. This suggests that at any point in time the level of the overall disagreement between forecasters is more likely to be due to a given set of forecasters, as opposed to any randomly-selected set of forecasters. Our findings are based on a multivariate measure that adjusts for the degree of difficulty in forecasting at each point in time, and also downweights an individual's disagreement measure when he/she is in agreement with the consensus view about how the economy operates, albeit not the magnitudes of the future values of the variables. Much of the literature considers disagreement between forecasters as a possible proxy for uncertainty (beginning with the seminal paper by Zarnowitz and Lambros (1987)), and considers how it varies with the state of the business cycle. However, individual forecasters are not identified, and the implicit assumption seems to be that any one forecaster is as likely to make the same contribution to overall disagreement at any point in

time as any other. The finding that more (less) contrarian forecasters in the first period remain so in the second period does not appear to be a manifestation of forecast inefficiency.

We also establish that there are systematic differences between forecasters in terms of accuracy. Interpreted within the context of the noisy information model, this could be because some forecasters (the more accurate set) are the beneficiaries of more informative signals, or there might be signal homogeneity, but some forecasters use that information less efficiently (the less accurate set).

A key focus is the extent to which the inefficient use of information explains the differences we observe between forecasters. Forecast inefficiency does not explain the persistence in contrarianism. In terms of accuracy, our results are less clear. Whether the persistence in accuracy rankings can be explained by forecast inefficiencies depends on the significance level we adopt. The null of no persistence in accuracy rankings of the corrected forecasts cannot be rejected at the 5% level, but it can at the 10% level. That is, there is some uncertainty as to whether the rejections of efficiency in the individual-level regressions do account for the substantive finding that forecasters differ in terms of accuracy.

When we reduced the number of forecasters to 25, as a robustness check, the findings relating to multivariate disagreement were largely unchanged, and for forecast accuracy suggested efficiency correction does account for the persistence in the forecast accuracy rankings.

Finally, we consider whether the more contrarian forecasters tend to be the more accurate forecasters. If so, we would conclude that such forecasters receive superior signals. That they stand apart from the crowd by virtue of receiving a different (superior) signal, and simultaneously report a more accurate forecast. The evidence strongly suggests more contrarian forecasters are less accurate. Forecasters who stand out from the crowd do not tend to produce more accurate forecasts.

The micro-level evidence suggests macro-forecasters are not 'essentially the same' as each other. The effect of inefficiency is somewhat nuanced - it does not explain why some forecasters appear to be systematically more contrarian than others at short term horizons, but forecast inefficiency may explain why some forecasters produce more or less accurate forecasts than others.

References

Arai, N. (2014). Using forecast evaluation to improve the accuracy of the Greenbook forecast. *International Journal of Forecasting*, *30*(1), 12–19.

Banternghansa, C., and McCracken, M. W. (2009). Forecast disagreement among FOMC members. Working papers 2009-059, Federal Reserve Bank of St. Louis.

Boero, G., Smith, J., and Wallis, K. F. (2015). The measurement and characteristics of professional forecasters' uncertainty. *Journal of Applied Econometrics*, **30(7)**, 1013–1234.

Bomberger, W. A. (1996). Disagreement as a measure of uncertainty. *Journal of Money, Credit and Banking*, **28**, 381–392.

Bonham, C., and Cohen, R. (2001). To aggregate, pool, or neither: Testing the rational expectations hypothesis using survey data. *Journal of Business and Economic Statistics*, **190**, 278–291.

Capistrán, C., and Timmermann, A. (2009). Disagreement and biases in inflation expectations. *Journal of Money, Credit and Banking*, **41**, 365–396.

Clements, M. P. (2009). Internal consistency of survey respondents' forecasts: Evidence based on the Survey of Professional Forecasters. In Castle, J. L., and Shephard, N. (eds.), *The Methodology and Practice of Econometrics. A Festschrift in Honour of David F. Hendry. Chapter 8*, pp. 206–226. Oxford: Oxford University Press.

Clements, M. P. (2010). Explanations of the Inconsistencies in Survey Respondents Forecasts. *European Economic Review*, **54(4)**, 536–549.

Clements, M. P. (2014). Forecast Uncertainty - Ex Ante and Ex Post: US Inflation and Output Growth. *Journal of Business & Economic Statistics*, **32(2)**, 206–216. DOI: 10.1080/07350015.2013.859618.

Clements, M. P., and Hendry, D. F. (1993). On the limitations of comparing mean squared forecast errors. *Journal of Forecasting*, **12**, 617–637. With discussion. Reprinted in Mills, T. C. (ed.) (1999), *Economic Forecasting. The International Library of Critical Writings in Economics*. Cheltenham: Edward Elgar.

Clements, M. P. (1995). Rationality and the role of judgement in macroeconomic forecasting. *Economic Journal*, **105**, 410–420.

Coibion, O., and Gorodnichenko, Y. (2012). What can survey forecasts tell us about information rigidities?. *Journal of Political Economy*, *120*(1), 116 – 159.

Coibion, O., and Gorodnichenko, Y. (2015). Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts. *American Economic Review*, *105*(8), 2644–78.

Croushore, D. (1993). Introducing: The Survey of Professional Forecasters. *Federal Reserve*

Bank of Philadelphia Business Review, November, 3–15.

Croushore, D. (2011a). Forecasting with real-time data vintages, chapter 9. In Clements, M. P., and Hendry, D. F. (eds.), *The Oxford Handbook of Economic Forecasting*, pp. 247–267: Oxford University Press.

Croushore, D. (2011b). Frontiers of real-time data analysis. *Journal of Economic Literature*, **49**, 72–100.

Croushore, D., and Stark, T. (2001). A real-time data set for macroeconomists. *Journal of Econometrics*, **105(1)**, 111–130.

Curran, P. A. (2015). Monte Carlo error analyses of Spearman's rank test. mimeo, International Centre for Radio Astronomy Research, Curtin University, Australia.

D'Agostino, A., McQuinn, K., and Whelan, K. (2012). Are some forecasters really better than others?. *Journal of Money, Credit and Banking*, **44(4)**, 715–732.

Elliott, G., Komunjer, I., and Timmermann, A. (2005). Estimation and testing of forecast rationality under flexible loss. *Review of Economic Studies*, **72**, 1107–1125.

Elliott, G., Komunjer, I., and Timmermann, A. (2008). Biases in macroeconomic forecasts: Irrationality or asymmetric loss. *Journal of the European Economic Association*, **6**, 122–157.

Engelberg, J., Manski, C. F., and Williams, J. (2009). Comparing the point predictions and subjective probability distributions of professional forecasters. *Journal of Business and Economic Statistics*, **27(1)**, 30–41.

Engelberg, J., Manski, C. F., and Williams, J. (2011). Assessing the temporal variation of macroeconomic forecasts by a panel of changing composition. *Journal of Applied Econometrics*, *26*(7), 1059–1078.

Fixler, D. J., Greenaway-McGrevy, R., and Grimm, B. T. (2014). The revisions to GDP, GDI, and their major components. *Survey of Current Business*, **August**, 1–23.

Hendry, D. F., and Martinez, A. B. (2017). Evaluating Multi-Step System Forecasts with Relatively Few Forecast-Error Observations. *International Journal of Forecasting*, **33(2)**, 359–372.

Holden, K., and Peel, D. A. (1990). On testing for unbiasedness and efficiency of forecasts. *The Manchester School*, **58**, 120–127.

Keane, M. P., and Runkle, D. E. (1990). Testing the rationality of price forecasts: new evidence from panel data. *American Economic Review*, **80(4)**, 714–735.

King, R. G., Plosser, C. I., Stock, J. H., and Watson, M. W. (1991). Stochastic trends and economic fluctuations. *American Economic Review*, **81**, 819–840.

Komunjer, I., and Owyang, M. T. (2012). Multivariate Forecast Evaluation and Rationality

Testing. *The Review of Economics and Statistics*, *94*(4), 1066–1080.

Kosobud, R., and Klein, L. (1961). Some econometrics of growth: Great ratios of economics. *Quarterly Journal of Economics*, **25**, 173–198.

Lahiri, K., and Sheng, X. (2008). Evolution of forecast disagreement in a Bayesian learning model. *Journal of Econometrics*, **144(2)**, 325–340.

Lahiri, K., and Liu, F. (2009). On the use of density forecasts to identify asymmetry in forecasters' loss function. *Business and Economic Statistics Section - JSM*, 2396–2408.

Landefeld, J. S., Seskin, E. P., and Fraumeni, B. M. (2008). Taking the pulse of the economy. *Journal of Economic Perspectives*, **22**, 193–216.

Mankiw, N. G., and Reis, R. (2002). Sticky information versus sticky prices: A proposal to replace the New Keynesian Phillips Curve. *Quarterly Journal of Economics*, **117**, 1295–1328.

Mankiw, N. G., Reis, R., and Wolfers, J. (2003). Disagreement about inflation expectations. mimeo, National Bureau of Economic Research, Cambridge MA.

Mincer, J., and Zarnowitz, V. (1969). The evaluation of economic forecasts. In Mincer, Jacob, A. (ed.), *Economic Forecasts and Expectations: Analysis of Forecasting Behavior and Performance*, pp. 3–46. New York: National Bureau of Economic Research.

Nordhaus, W. D. (1987). Forecasting efficiency: Concepts and applications. *Review of Economics and Statistics*, **69**, 667–674.

Patton, A. J., and Timmermann, A. (2007). Testing forecast optimality under unknown loss. *Journal of the American Statistical Association*, **102**, 1172–1184.

Patton, A. J., and Timmermann, A. (2010). Why do forecasters disagree? Lessons from the term structure of cross-sectional dispersion. *Journal of Monetary Economics*, **57(7)**, 803–820.

Patton, A. J., and Timmermann, A. (2012). Forecast rationality tests based on multi-horizon bounds. *Journal of Business & Economic Statistics*, *30*(1), 1–17.

Rich, R., and Tracy, J. (2010). The relationships among expected inflation, disagreement, and uncertainty: Evidence from matched point and density forecasts. *Review of Economics and Statistics*, **92(1)**, 200–207.

Rich, R. W., and Butler, J. S. (1998). Disagreement as a measure of uncertainty: A comment on Bomberger. *Journal of Money, Credit and Banking*, **30**, 411–419.

Sargent, T. J. (ed.)(1999). *The Conquest of American Inflation*: Princeton University Press.

Satopää, V. A. (2018). Combining information from multiple forecasters: Inefficiency of central tendency. mimeo, INSEAD, Technology ad Operations Management.

Satopää, V. A., Pemantle, R., and Ungar, L. H. (2016). Modeling probability forecasts via

information diversity. *Journal of the American Statistical Asociation*, **111(516)**, 1623–1633.

Sims, C. A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, **50**, 665–690.

Sinclair, T. M., Stekler, H., and Carnow, W. (2015). Evaluating a vector of the Fed's forecasts. *International Journal of Forecasting*, *31*(1), 157–164.

Whelan, K. (2003). A two-sector approach to modeling U.S. NIPA data. *Journal of Money, Credit and Banking*, *35*(4), 627–56.

Woodford, M. (2002). Imperfect common knowledge and the effects of monetary policy. In Aghion, P., Frydman, R., Stiglitz, J., and Woodford, M. (eds.), *Knowledge, Information, and Expectations in Modern Macroeconomics: In honor of Edmund Phelps*, pp. 25–58: Princeton University Press.

Zarnowitz, V. (1985). Rational expectations and macroeconomic forecasts. *Journal of Business and Economic Statistics*, **3(4)**, 293–311.

Zarnowitz, V., and Lambros, L. A. (1987). Consensus and uncertainty in economic prediction. *Journal of Political Economy*, **95(3)**, 591–621.

Table 1: Description of Forecast Data and Real-Time Data

Variable	SPF code	RTDSM code
Real GDP (GNP)	RGDP	ROUTPUT
Real personal consumption	RCONSUM	RCON
Real nonresidential fixed investment	RNRESIN	RINVBF
Real residential fixed investment	RRESINV	RINVRESID

The SPF data are from the Philadelphia Fed website `http://www.phil.frb.org/econ/spf/`.
For the investment series we used RNRESIN + RRESINV.
The real-time data were downloaded from:
`http://www.philadelphiafed.org/research-and-data/real-time-center/real-time-data/`.
Both the forecast data and real-time data were downloaded in December 2018.

Table 2: MZ Forecast Efficiency Tests: Summary

Vintage	Consumption		Investment		Output	
	$h = 0$	$h = 4$	$h = 0$	$h = 4$	$h = 0$	$h = 4$
Advance	0.560	0.760	0.400	0.900	0.420	0.940
2nd quarterly	0.540	0.700	0.260	0.640	0.220	0.800

The table reports the proportion of rejections across the 50 respondents of the null of forecast efficiency for $h = 0$ and $h = 4$, for each variable, based on equation (2), with HAC estimation of the variance-covariance matrix of the parameter estimates.. The test is run at the 5% level. The actual value is either the advance estimate, or the value available in the RTDSM two months after the reference quarter

Table 3: MZ Forecast Efficiency Tests: Individual Respondent Results

Respondent	No. Forecasts	Consumption		Investment		Output	
		$h=0$	$h=4$	$h=0$	$h=4$	$h=0$	$h=4$
20	61	0.000	0.000	0.000	0.000	0.000	0.000
40	45	0.144	0.376	0.001	0.001	0.111	0.173
84	57	0.065	0.727	0.143	0.446	0.062	0.006
94	33	0.071	0.005	0.335	0.000	0.062	0.000
99	47	0.000	0.000	0.160	0.616	0.019	0.007
404	31	0.506	0.002	0.033	0.000	0.089	0.000
405	31	0.030	0.000	0.020	0.000	0.167	0.000
407	75	0.006	0.000	0.006	0.000	0.041	0.000
411	80	0.105	0.000	0.817	0.000	0.539	0.000
414	32	0.011	0.000	0.298	0.000	0.024	0.000
420	72	0.001	0.003	0.625	0.026	0.470	0.004
421	90	0.075	0.000	0.629	0.000	0.292	0.000
422	48	0.161	0.018	0.000	0.000	0.000	0.000
423	47	0.073	0.000	0.299	0.098	0.009	0.000
424	35	0.819	0.546	0.037	0.030	0.036	0.045
426	94	0.001	0.000	0.052	0.000	0.002	0.000
428	91	0.000	0.000	0.298	0.000	0.366	0.000
429	69	0.025	0.000	0.059	0.002	0.007	0.001
431	65	0.008	0.033	0.076	0.029	0.001	0.001
433	86	0.004	0.006	0.756	0.014	0.246	0.000
439	38	0.011	0.000	0.001	0.000	0.139	0.000
446	82	0.011	0.002	0.415	0.000	0.184	0.000
456	64	0.000	0.000	0.075	0.000	0.000	0.000
463	76	0.047	0.000	0.181	0.005	0.044	0.000
472	61	0.186	0.000	0.717	0.000	0.706	0.000
483	56	0.004	0.072	0.149	0.004	0.010	0.000
484	75	0.000	0.000	0.926	0.000	0.005	0.000
498	31	0.001	0.962	0.449	0.000	0.000	0.016
504	64	0.295	0.001	0.000	0.000	0.000	0.000
506	57	0.026	0.000	0.143	0.000	0.229	0.000
507	62	0.013	0.000	0.684	0.000	0.380	0.000
508	53	0.016	0.000	0.005	0.058	0.377	0.001
510	67	0.081	0.000	0.748	0.055	0.069	0.008
512	53	0.002	0.000	0.000	0.000	0.532	0.000
516	40	0.176	0.079	0.460	0.000	0.693	0.057
518	57	0.016	0.004	0.867	0.014	0.085	0.011
520	53	0.290	0.005	0.119	0.005	0.300	0.015
524	48	0.004	0.010	0.000	0.000	0.759	0.000
526	33	0.006	0.464	0.002	0.000	0.166	0.000
527	42	0.064	0.000	0.188	0.000	0.104	0.000
528	31	0.627	0.472	0.388	0.019	0.417	0.154
535	40	0.067	0.000	0.005	+DEN	0.000	0.000
540	31	0.589	0.071	0.320	0.006	0.787	0.000
542	41	0.125	0.087	0.000	0.000	0.012	0.015
546	42	0.196	0.036	0.013	0.000	0.465	0.000
548	43	0.000	0.090	0.009	0.010	0.000	0.018
553	30	0.270	0.226	0.052	0.000	0.013	0.000
555	34	0.233	0.000	0.004	0.000	0.119	0.000
556	27	0.028	0.000	0.050	0.004	0.085	0.000
557	33	0.025	0.000	0.000	0.000	0.002	0.000

The table reports the p-values of the null hypothesis that $\delta_0 = 0$ and $\delta = 1$ in equation (2), using HAC estimation of the variance-covariance matrix of the parameter estimates. The actual value is the advance estimate.

Table 4: Multivariate Disagreement Statistics by Individual

id.	Reported Forecasts				Real-time Efficiency-Corrected Forecasts			
	Ave	Equal	Rank $h=0$	Rank $h=4$	Ave	Equal	Rank $h=0$	Rank $h=4$
1	2	3	4	5	6	7	8	9
20	2.868	1	50	50	2.079	0.990	43	2
40	1.696	0.957	36	29	1.643	0.282	19	1
84	1.328	0.330	22	37	1.320	0	5	5
94	1.323	0.322	21	28	1.797	0.671	31	3
99	2.356	1	49	48	2.530	0.999	50	4
404	1.247	0.195	14	9	2.255	0.989	49	18
405	1.755	0.975	39	43	2.086	0.944	44	16
407	1.206	0.094	12	27	2.143	0.990	46	10
411	1.083	0.006	5	11	1.434	0.004	7	23
414	1.448	0.662	28	45	1.616	0.237	17	36
420	1.295	0.246	20	30	1.641	0.246	18	11
421	1.481	0.778	31	36	1.674	0.316	22	17
422	1.832	0.998	41	31	1.259	0	1	40
423	2.047	1	47	41	1.914	0.859	38	26
424	1.293	0.264	19	7	1.663	0.315	20	9
426	1.899	1	43	34	1.881	0.901	36	12
428	1.281	0.215	17	32	1.664	0.300	21	6
429	1.335	0.352	23	23	1.552	0.067	13	20
431	1.109	0.016	6	4	1.536	0.070	11	30
433	0.879	0	2	6	1.277	0	2	8
439	1.620	0.927	34	22	1.698	0.420	23	28
446	1.141	0.032	10	8	1.494	0.030	8	15
456	1.269	0.183	16	25	1.732	0.502	26	33
463	1.470	0.736	29	21	1.731	0.500	25	27
472	1.290	0.242	18	24	1.827	0.769	33	19
483	1.851	0.999	42	33	1.515	0.036	9	22
484	1.481	0.758	32	16	1.993	0.966	42	31
498	1.235	0.140	13	3	1.769	0.592	29	41
504	2.006	1	44	49	1.852	0.826	34	38
506	1.134	0.027	8	17	1.586	0.109	16	29
507	1.633	0.961	35	46	1.944	0.921	40	14
508	1.474	0.732	30	38	2.109	0.995	45	24
510	1.814	0.998	40	44	1.796	0.692	30	32
512	2.024	1	45	47	2.174	0.995	47	37
516	1.423	0.603	27	19	1.746	0.540	27	42
518	1.403	0.556	26	13	1.422	0.008	6	25
520	0.994	0.001	4	1	1.278	0	3	13
524	1.716	0.980	37	20	1.713	0.449	24	21
526	1.566	0.857	33	18	2.228	0.979	48	43
527	1.337	0.367	24	39	1.754	0.558	28	34
528	1.135	0.067	9	10	1.815	0.677	32	49
535	1.258	0.166	15	15	1.543	0.076	12	44
540	1.156	0.049	11	14	1.555	0.131	14	48
542	1.115	0.017	7	12	1.901	0.832	37	45
546	0.925	0	3	5	1.314	0.001	4	47
548	2.234	1	48	26	1.880	0.818	35	50
553	0.859	0	1	2	1.518	0.073	10	35
555	1.384	0.500	25	35	1.579	0.123	15	46
556	1.746	0.973	38	42	1.933	0.860	39	7
557	2.040	1	46	40	1.948	0.862	41	39

The table reports results for the multivariate disagreement measure, where $S_{t|t-h}$ is calculated as in eqn. (10). The 2nd and 6th columns denote the mean (across surveys) value of $D_{i,t|t-h}$ (eqn. 10) for $h=0$. For $h=0$ we also report the p-values of testing the equality of means of each individual against the 'average' forecaster (precisely, the forecaster with the $N/2$ largest average disagreement), in column 2 (or 6). The tests are constructed such that a p-value greater than 0.95 (0.975) in column 3 (7) has a larger population D_i, and a p-value less that 0.05 (0.025) suggests a D_i significantly smaller than that of the average forecaster, in a two-sided test at the 10% (5%) level).
Columns 4 and 5 (and 8 and 9) rank the forecasters in terms of average $D_{i,t|t-h}$ for $h=0$ and $h=4$.

Table 5: Rank Correlation Tests of Multivariate Disagreement

Reported Forecasts			
Panel 1A. Earlier and later periods			
S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.547	0.557	0.517	0.365
0.001	0.001	0.002	0.029

Panel 1B. $h=0$ and $h=4$ forecasts					
S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
0.790	0.710	0.786	0.711	0.650	0.738
0.000	0.000	0.000	0.000	0.001	0.000

Panel 2A. Efficiency-Corrected Forecasts			
A. Earlier and later periods			
S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.533	0.378	0.439	0.410
0.002	0.024	0.010	0.016

Panel 2B. $h=0$ and $h=4$ forecasts					
S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
-0.047	-0.106	0.043	-0.028	0.104	-0.028
0.624	0.724	0.396	0.574	0.280	0.569

The Spearman rank correlation r lies between -1 and 1, where 0 indicates no relationship. For each test, there are two entries. The first row entry is the rank correlation given by:

$$r = 1 - \frac{6R}{N\left(N^2 - 1\right)}$$

where R is the sum of squared differences between the ranks (e.g., of the forecasters in the first sample, and in the second sample).

The second row entry is the probability of the test statistic being at least as large as we obtained if the null hypothesis (of a zero correlation) is true. Probabilities less than 0.025 or greater than 0.975 indicate rejections of the null in a two-sided test at the 5% level. (High probabilities suggest a negative relationship, and low probabilities a a positive relationship).

The probabilities we report are calculated for the Fisher transformation,

$$F\left(r\right) = \frac{1}{2}\ln\frac{1+r}{1-r}$$

such that $z = F\left(r\right).\sqrt{\frac{N-3}{1.06}} \sim N\left(0,1\right)$ under the null of statistical independence.

Table 6: Forecast Accuracy Rankings: Persistence Across Sub-samples

Panel A. Reported: Not normalized			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.245	0.350	0.104	0.182
0.103	0.033	0.299	0.177
Panel B. Reported: Normalized			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.522	0.455	0.258	0.505
0.002	0.007	0.092	0.002
Panel C. Efficiency-Corrected (and Normalized)			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.358	0.342	-0.412	-0.177
0.029	0.036	0.987	0.817

The forecast errors in panels B and C are normalized. Panel A reports accuracy measures based on the raw errors.

The table shows the Spearman test of no relationship in the accuracy ranks (either trace or determinant measure) between the earlier and later samples. The first value is the rank correlation r, and the second is the probability of observing a larger value: see notes to table 5 for an explanation.

Normalized denotes that the forecast errors have been adjusted for differences over time in average forecast accuracy.

Table 7: Rank Correlation Tests: Accuracy and Disagreement

Reported Forecasts			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det.
0.925	0.913	0.765	0.822
0	0	0	0
Efficiency-Corrected Forecasts			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det.
0.719	0.767	0.414	0.440
0	0	0.002	0.001

The table shows the Spearman test of no relationship in the accuracy ranks (either trace or determinant measure) and the disagreement ranks. The measures of accuracy are based on normalized forecasts.
The first value reported in the table is the rank correlation r, t statistic, and the second is the probability of observing a larger value: see notes to table 5 for an explanation.

Table 8: Relationship between Forecast Accuracy and the Gains/Losses from Real-time Efficiency Correction

Consumption		Investment		Output	
$h=0$	$h=4$	$h=0$	$h=4$	$h=0$	$h=4$
-0.452	-0.545	-0.363	-0.401	-0.421	-0.491
0.999	0.999	0.994	0.998	0.999	1.000

The table shows the Spearman test of no relationship between the ranks of ratio of the RMSE of the efficiency-corrected forecasts to the RMSE of the reported forecasts, and the RMSE of the (normalised) reported forecast.
The first value is the rank correlation r, t statistic, and the second is the probability of observing a larger value: see notes to table 5 for an explanation.

Table 9: Results for the Most Prolific Forecaster, 1990:4 to 2017:2

Panel A

Consumption		Investment		Output	
$h=0$	$h=4$	$h=0$	$h=4$	$h=0$	$h=4$
RMSE ratio of Real-Time Corrected to Reported					
0.918	0.922	0.980	0.959	0.947	0.913
RMSE ratio of In-Sample Corrected to Reported					
0.874	0.882	0.928	0.921	0.917	0.868

Panel B

Reported Forecasts, Accuracy Ranks				Corrected Forecast, Accuracy Ranks			
$h=0$		$h=4$		$h=0$		$h=4$	
TMSFE	Det	TMSFE	Det	TMSFE	Det	TMSFE	Det
37	40	33	33	34	36	9	12

Panel C

Reported Forecasts, Disag. Ranks		Corrected Forecast, Disag. Ranks	
$h=0$	$h=4$	$h=0$	$h=4$
40	31	37	13

The table shows illustrative results for a single forecaster (SPF identifier 426). Panel A indicates the improvements in forecast accuracy from real-time and in-sample correction.
Panel B the ranks of the forecaster using multivariate accuracy measures (trace and determinant of the mean squared forecast error matrix) for the reported and efficiency-corrected forecasts.
Panel C the ranks for the multivariate disagreement measure (non-diagonal S-matrix) for the reported and efficiency-corrected forecasts.
The actual values for the forecast accuracy RMSFE comparisons are the first-release advance estimates.

Table 10: Rank Correlation Tests of Multivariate Disagreement, Top 25 Forecasters

Reported Forecasts			
Panel 1A. Earlier and later periods			
S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.554	0.478	0.526	0.309
0.003	0.010	0.005	0.078

Panel 1B. $h=0$ and $h=4$ forecasts					
S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
0.808	0.642	0.838	0.758	0.617	0.802
0.000	0.000	0.000	0.000	0.001	0.000

Panel 2A. Efficiency-Corrected Forecasts			
A. Earlier and later periods			
S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.386	0.510	0.360	0.554
0.035	0.006	0.047	0.003

Panel 2B. $h=0$ and $h=4$ forecasts					
S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
0.132	0.096	-0.024	0.240	0.404	-0.041
0.273	0.335	0.543	0.132	0.028	0.574

The table is the same as table 5, but for the top 25 forecasters, rather than the top 50.
The table shows the Spearman test of no relationship in multivariate disagreement across time, and between the $h=0$ and $h=4$ forvarious sample periods.
The first value is the rank correlation r, and the second is the probability of observing a larger value: see notes to table 5 for an explanation.

Table 11: Forecast Accuracy Rankings: Persistence Across Sub-samples, Top 25 Forecasters

Panel A. Reported: Not normalized			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.307	0.286	0.326	0.347
0.074	0.090	0.062	0.050
Panel B. Reported: Normalized			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.421	0.389	0.266	0.438
0.020	0.031	0.107	0.016
Panel C. Efficiency-Corrected (and Normalized)			
$h = 0$		$h = 4$	
Tr.	Det.	Tr.	Det
0.208	0.215	-0.410	-0.145
0.168	0.159	0.976	0.746

The table is the same as table 6, but for the top 25 forecasters, rather than the top 50.
The table shows the Spearman test of no relationship in the accuracy ranks (either trace or determinant measure) between the earlier and later samples.
The first value is the rank correlation r, and the second is the probability of observing a larger value: see notes to table 5 for an explanation.

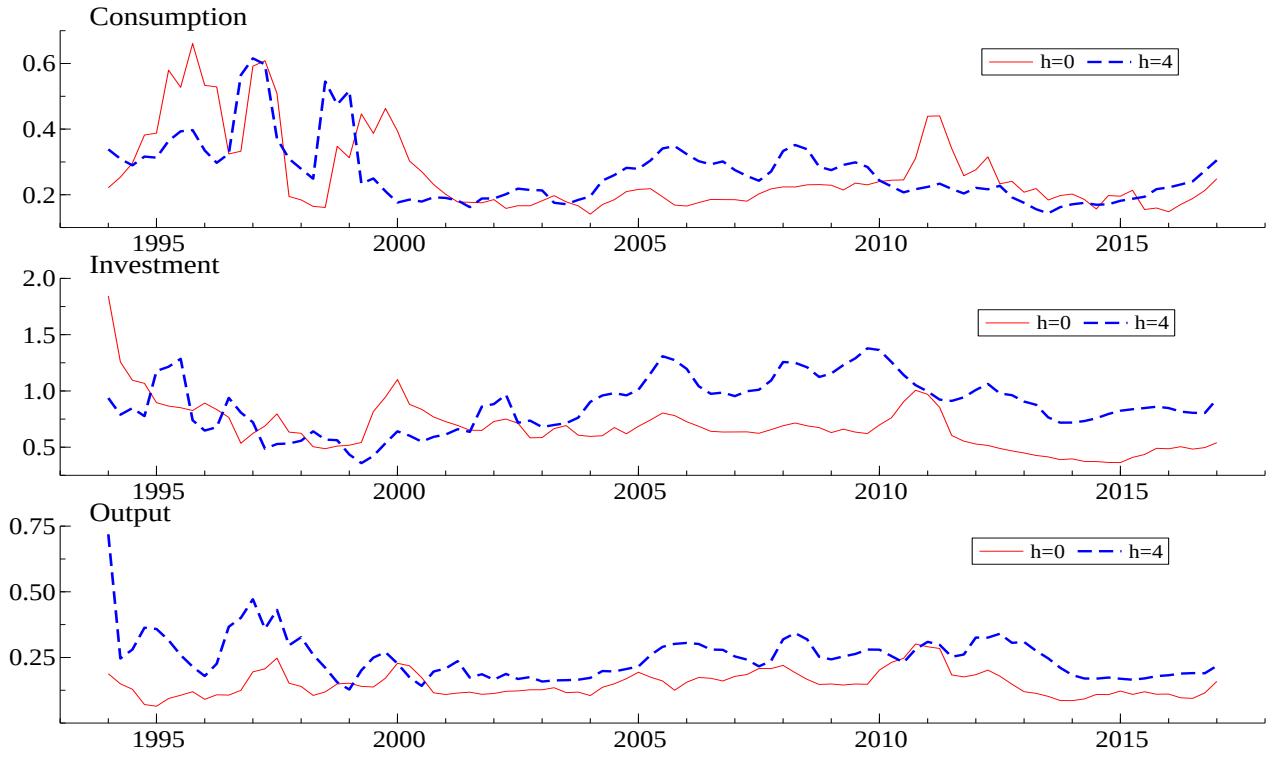

Figure 1: Mean absolute efficiency corrections at each survey date (with some smoothing, a centred moving average with leads and lags of 1).

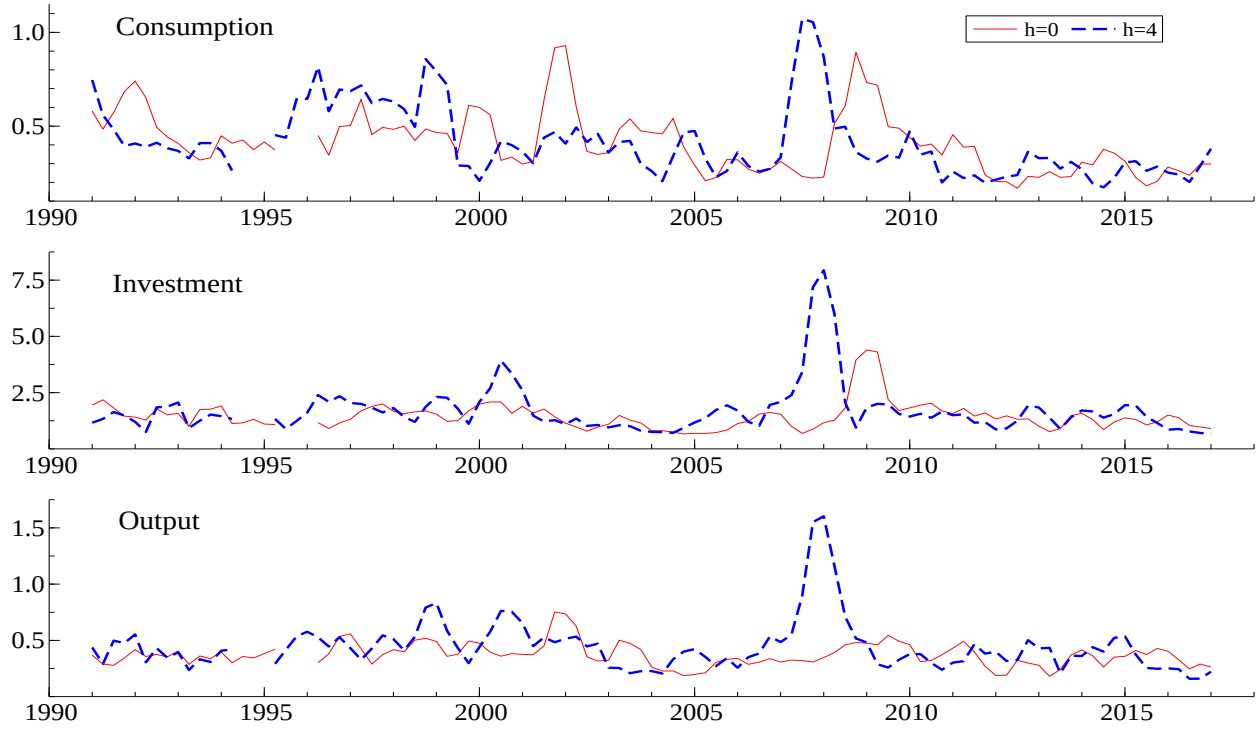

Figure 2: Cross-sectional root mean squared error at each survey date (with some smoothing, a centred moving average with leads and lags of 1).

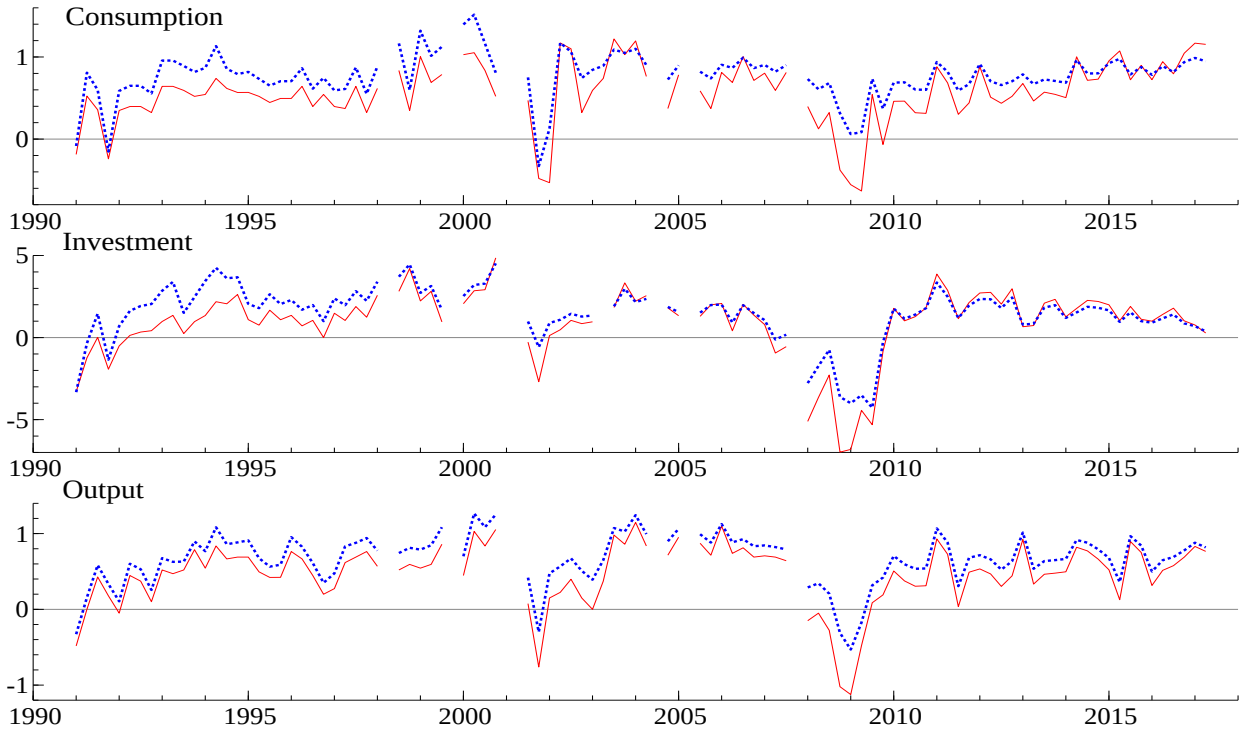

Figure 3: Reported (solid line) and efficiency-corrected (dotted line) current quarter forecasts for the most prolific forecaster (id 426). The forecasts are one hundred times the log differences of the annualized levels.

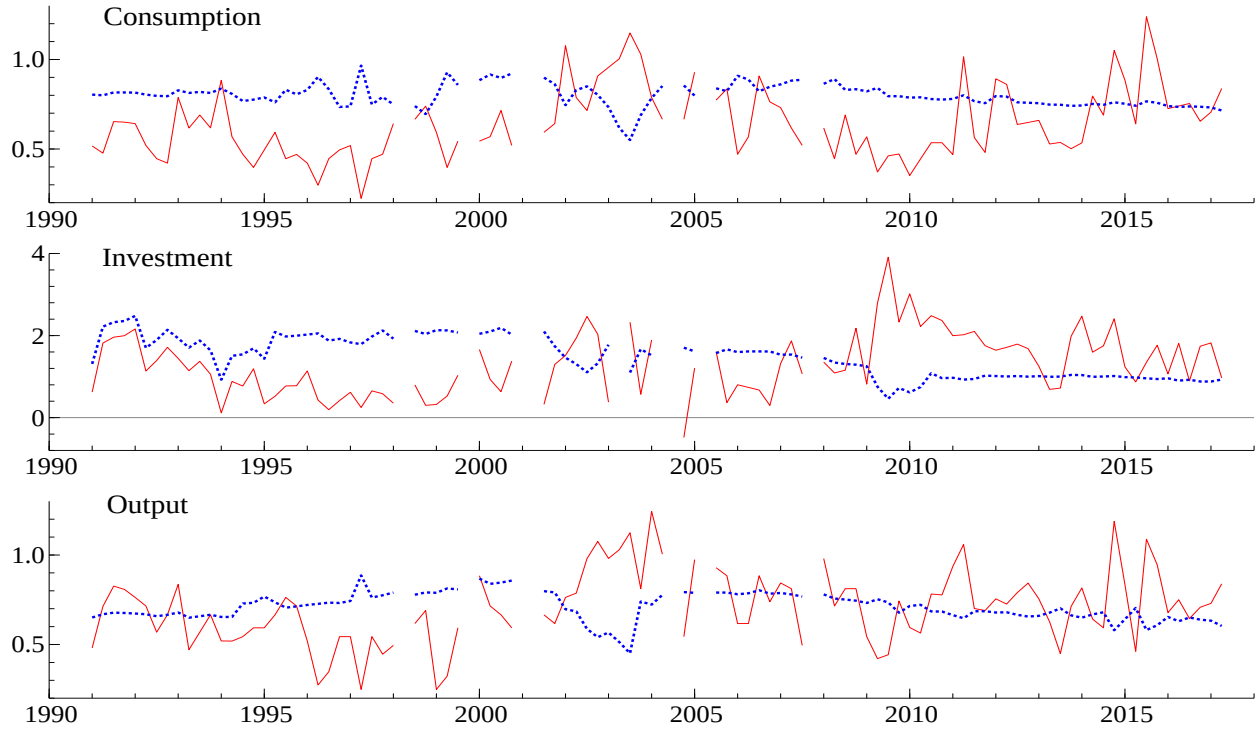

Figure 4: Reported (solid line) and efficiency-corrected (dotted line) $h = 4$ quarter forecasts for the most prolific forecaster (id 426). The forecasts are one hundred times the log differences of the annualized levels.