Forecaster efficiency, accuracy and disagreement: evidence using individual-level survey data

Article

Accepted Version

Clements, Mike ORCID logoORCID: https://orcid.org/0000-0001-6329-1341 (2022) Forecaster efficiency, accuracy and disagreement: evidence using individual-level survey data. Journal of Money, Credit and Banking, 54 (2-3). pp. 537-568. ISSN 1538-4616 doi: https://doi.org/10.1111/jmcb.12867 Available at https://centaur.reading.ac.uk/93811/

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.

October 14, 2020

Abstract

Theories of expectations formation sometimes suppose that agents make efficient forecasts given their information sets. We use individual-level data to test whether survey respondents' forecasts are efficient. We assess 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 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. Helpful comments were received from seminar participants at Reading, Queen's University, Belfast, Warwick Business School, and the International Symposium of Forecasting, Thessaloniki.

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] Under both models of expectations behaviour agents' forecasts are efficient in the sense of Mincer and Zarnowitz (1969): their forecasts are systematically uncorrelated with their forecast errors.

Influential papers by Coibion and Gorodnichenko (2012, 2015) test the macro-level implications of these models on aggregate quantities, such as the mean error and consensus forecasts. We consider the micro-level evidence for whether respondents' forecasts are efficient, and whether 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 also consider whether forecasters are identical in terms of forecast accuracy, or whether there are systematic differences, and if so, whether differences in forecasting ability are persistent over time.

The finding that forecasters are not rational, in the sense that their forecasts are not efficient, would pose a fundamental challenge to IR models. 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.

Interpreting the rejection of weak efficiency as suggesting forecasters are not rational is predicated on forecasters having symmetric loss functions, and not being motivated by possible strategic behaviour, such as (anti-) herding, for example (see, e.g., Bernhardt, Campello and Kutsoati (2006)). Asymmetric loss functions or strategic considerations may contribute to the finding of inefficiency. Nevertheless, our analysis still charts the impact of those inefficiencies

[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 information.

on forecaster disagreement and accuracy.

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 efficiency correction we implement is real time, and ensures that we are not simply capturing a 'look-ahead' bias. Instead, the forecaster could have made the corrections at the time each forecast was issued, since they use only their past history of forecasts and outcomes.

The paper asks whether various aspects of forecaster heterogeneity can be explained by an inefficient use of information. Although the finding of inefficiency is at odds with IR models, the intention is not to formally test theories of expectations formation. Recent papers including Fuhrer (2018), Broer and Kohlhas (2018) and Bordalo, Gennaioli, Ma and Shleifer (2018) also find forecaster inefficiency. For example, Broer and Kohlhas (2018) and Bordalo *et al.* (2018) suggest forecasters are over-confident, in the sense that they over-react to new information, inducing a negative correlation between their forecast revision and their forecast error. Broer and Kohlhas (2018) attribute the over-reaction to private information as being due to 'absolute' over-confidence, and the over-reaction to public information as being due to 'relative' over-confidence. Bordalo *et al.* (2018) explain the over-reaction with a model of 'diagnostic' expectations (following Bordalo, Gennaioli, Porta and Shleifer (2017)). By way of contrast, Fuhrer (2018) argues for 'intrinsic inflation persistence', that is, that individuals *under*-react to new information, smoothing their response to news. We do not attempt to choose between them, or to provide an explanation of our own. But we explore the practical relevance of forecaster inefficiency, in terms of the extent to which it accounts for heterogeneity (persistent differences across forecasters in terms of accuracy or contrarianism) by correcting all the forecasts for inefficiency. The evidence for over-reaction we have alluded to comes from regressions of individual forecast errors on forecast revisions. However, a rejection of the null

that the coefficient on the forecast revision is zero (in favour of it being negative, say) does not immediately indicate how important the inefficiency is.[2] We ask whether the inefficiencies are large enough to account for heterogeneity, by comparing the reported and efficiency-corrected forecasts in various dimensions.

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.[3] 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. Persistent differences over time between forecasters in terms of their degree of contrarianism are found not to be solely due to forecaster inefficiency.

We also consider the micro-level evidence for the assertion that individual forecasters are equally accurate. We find evidence against the assumption of equal accuracy. Once the forecasts are corrected for inefficiency, the evidence for persistent differences in forecast accuracy is considerably weakened.

The plan of the remainder of the paper is as follows. Section 2 describes the forecast data. Section 3 explains the notion of forecaster efficiency, and describes the empirical evidence. It also explains how the reported forecasts can be 'efficiency corrected', so that comparisons between individual forecasters based on their reported forecasts, compared to comparisons based on the corrected forecasts, serve to isolate the impact of the inefficient use of information. 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. Both sections 4 and 5 determine the impact of inefficiency on the corresponding attribute of forecaster performance: contrarianism, or accuracy. Sections 6 and 7 further explore the relationships between efficiency and accuracy, and between accuracy and contrarianism. Section 8 checks the robustness of our findings for a smaller sample of forecasters. Section 9 offers some

[2] Fuhrer (2018) argues that the finding of a negative coefficient in the regression of the forecast error $y_{t+1} - y_{t+1|t}$ on the revision $y_{t+1|t} - y_{t+1|t-1}$ (here $y_{t+1|t-1}$ is the forecast of period $t+1$ made at time $t-1$) does not suggest an over-response to the 'news' embodied in the revision, because he finds it is $y_{t+1|t}$ (rather than $y_{t+1|t} - y_{t+1|t-1}$) which has predictive power for the forecast error. The regression of the error on the forecast is the basis of the efficiency correction we use.

[3] 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.

concluding remarks.

A web-only Appendix contains additional results, which are summarized in the main paper. Although our focus is squarely on cross-sectional characteristics of the survey respondents, and how these characteristics depend on forecast efficiency, in the web-only Appendix we also illustrate the effect of efficiency correction for one forecaster. 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 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[4] 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

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

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 is, 'that participation in the survey after recruitment is statistically independent of forecasters beliefs about inflation' (Engelberg, Manski and Williams (2011, p.1061)). [5] 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 (6) or (7), as opposed to (8)), and in calculating forecast accuracy in section 5, normalized forecast errors are used.

In the paper we report results for quarterly growth rates - for the current quarter ($h = 0$) and for the year-ahead quarter ($h = 4$). The growth rates are calculated as (one hundred times) the difference of the logs of the levels. For $h = 0$, this is the growth rate between the current and previous quarters, and for $h = 4$, it is the growth rate between the same quarter next year, and the quarter one before that.

At the time the forecasts are filed - around 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

[5] However, as argued by Engelberg *et al.* (2011), there is little available evidence on whether this is a reasonable assumption. Note that the assumption is also required for the analysis of aggregate (or consensus) forecasts. For the European Central Bank's panel of forecasters, López-Pérez (2016) provides evidence against the 'missing at random' assumption.

(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.[6] We generally 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 is a conservative assumption, but satisfies the requirement that y_i is necessarily included in the forecaster's information set, $\mathcal{F}_i$. Hence in testing for forecast efficiency, we do not make make use of actual private information sets which are generally unobserved. Instead, we assume that all private and public information used by a particular forecaster is summarized or encapsulated in his/her own forecast y_i.

We can show that a forecaster can be efficient without having access to all relevant information, and that they may possess private information.[7] 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 $\gamma_s = Cov\left(y_t, y_{t-s}\right)$, 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 $\gamma \neq \gamma^*$.[8]

[6] 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)).

[7] Figlewski and Wachtel (1983) make this point in response to a comment by Dietrich and Joines (1983) on Figlewski and Wachtel (1981). It may be that all forecasters 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).

[8] Straightforward algebra shows that (1) holds when $\gamma = \gamma^*$. By definition,

$$E\left(y|y_i\right) = \frac{Cov(y, y_i)}{Var(y_i)} y_i$$

and $Cov(y, y_i) = Var(y_i)$ when $\gamma = \gamma^*$. Alternatively, the correlation between forecast $y_{i,t} = \gamma y_{t-1}$ and forecast

To illustrate the role of private information, we drop the t-subscripts, and suppose that forecaster i's prior belief of y is given by $y = \mu + v$, where $v \sim iid\left(0, \sigma_v^2\right)$, and i also receives a private signal x_i, $x_i = y + u + \varepsilon_i$, where $u \sim iid\left(0, \sigma_u^2\right)$, a common error to all forecasters, and where $\varepsilon_i \sim iid\left(0, \sigma_{\varepsilon_i}^2\right)$. If the precision of the signal does not vary over i, i.e., $\sigma_{\varepsilon_i}^2 = \sigma_\varepsilon^2$ for all i, then all forecasters are equally as good on average, but this issue does not bear on forecaster efficiency. It is assumed that u, ε_i and v are all uncorrelated. Forecaster i's optimal forecast (in terms of squared error loss) weights the private signal x_i and prior μ as $y_i = \omega_i x_i + (1 - \omega_i)\mu$, where the weight ω_i is given by $\omega_i = \sigma_v^2\left(\sigma_v^2 + \sigma_u^2 + \sigma_\varepsilon^2\right)^{-1}$. Simple algebra shows that such forecasters will produce efficient forecasts, as defined by (1), with the property that the forecast error $y - y_i$ and forecast y_i are uncorrelated.[9]

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, across t. Here $y_{i,t|t-h}$ denotes the forecast made by i at time $t-h$ of y_t.[10] The MZ test of the null of optimality is a joint test that $\delta_0 = 0$ and $\delta = 1$. Unless $\delta = 1$, the forecast and forecast error will be systematically related, and this correlation could be exploited to generate a superior forecast. 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 early as Zarnowitz (1985) it has been argued that pooled cross-section time-series regressions are not an appropriate vehicle for testing for rationality. That is, we should not estimate (2) by pooling over individuals i and time periods t. This is because for each t the dependent variable y takes a single value, which induces a negative cross-sectional covariance between the disturbance and the forecast, resulting in the slope parameter δ being biased towards zero. This happens irrespective of whether or not the model is estimated with individual-specific intercepts.[11] But notice this pre-supposes that there is a common slope parameter across indi-

error $e_t = y_t - \gamma y_{t-1}$ is zero when $\gamma = \gamma^*$.

[9] This example is used by Crowe (2010) to show that consensus forecasts will under-weight private information when individual forecasters have different information sets.

[10] The dependence of δ_0 and δ on i and h in (2) is suppressed in the notation.

[11] The literature disagrees as to whether the slope estimator is biased or inconsistent. Bonham and Cohen

viduals. This is often referred to as microhomogeneity. When the population parameters δ_0 and δ_1 differ across individuals, Zarnowitz (1985) and Bonham and Cohen (2001) argue that pooled regressions make little sense. Because we wish to allow for the possibility that respondents differ in terms of the use of their information sets, we estimate separate regressions for each individual. That is, we do not wish to impose microhomogeneity at the outset. This marks a major departure from the recent literature. In the web-only Appendix we show the variability in the estimates of δ across individuals.

Finally, Keane and Runkle (1990) criticize the use of revised data in regressions such as (2), and suggest it may be responsible for the erroneous rejection of rationality. We use real-time vintage estimates of the actual values. That is, vintages released soon after the reference quarter, rather than the latest-available vintage at the time of the investigation. The latest-available 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.[12] 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' those forecasts. The *in-sample* efficiency-corrected 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 than the sum of squared forecast errors of the reported forecasts. The efficiency corrected forecasts are necessarily more accurate on squared-error loss.[13]

(2001) argue that it is inconsistent, contrary to Keane and Runkle (1990), who suggest it is $\sqrt{T}$-consistent.

[12] 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).

[13] As mentioned in the Introduction, 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.

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. When the correction is implemented in real time, the corrected forecasts will not necessarily be more accurate than the reported forecasts. 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:

$$y^{\tau}_{\tau-1} = \delta_0 + \delta y_{\tau-1|\tau-1} + u_{\tau-1} \tag{3}$$

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

$$y_{t|t} = \hat{\delta}_{0,t} + \hat{\delta}_t y_{t|t} \tag{4}$$

where $\hat{\delta}_{0,t}$ and $\hat{\delta}_t$ are the estimates of (3) based on data available at survey t. We estimate (3) 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.[14] 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 requires that at survey time t the latest forecast and actual value pair available for estimating the equivalent of (3) are $y_{t-1|t-5}$ and y^{t}_{t-1}. 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

[14] 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.

of the 1994:4 target period.

Forecasts could be corrected for inefficiency based on other test regressions, such as the Optimal Revision Regression of Patton and Timmermann (2012), although we use the MZ regression, as in (3).

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. A web-only Appendix contains a Table providing 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 between individuals. 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, we consider the effects of the efficiency corrections on aggregate measures. Table 3 provides some summary statistics for the actual forecast errors and the corrected forecast errors. The ratio of the mean absolute errors to the mean actual values exceeds a half in all cases, and is in excess of one for investment, reflecting the greater difficulty of forecasting this variable. Figure 1 plots the time series of the changes in the forecasts from the out-of-sample efficiency corrections. These are the cross-sectional mean absolute change between the reported and corrected forecasts at each survey date, for $h = 0$ and $h = 4$.[15] Figure 2 shows that the

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

ratio of the cross-sectional means of the absolute efficiency corrections to the mean absolute errors varies considerably over time for the longer horizon forecasts of investment and output growth, with increases towards the end of the 2000's. This increase in the degree of inefficiency relative to the magnitude of the errors of the reported forecasts did not occur for the short horizon forecasts of investment and output. The efficiency corrections are sizeable. Although there is business cycle variation, the corrections are a feature of the whole sample period. Table 3 shows that the mean quarterly output growth over the period is around 0.6%, whereas the mean absolute correction for the current-quarter output forecasts is 0.14%, and for the 4-quarter ahead is 0.24% (see row 'Mean absolute correction: Always-Reported'). The corrections are larger (on average) for the longer horizon forecasts for all variables, and this also holds for most time periods for output and investment (see the figure).

The corrections discussed so far have been calculated for all forecasters. Instead, we can only correct the forecasts of respondents for whom we reject a test of forecast efficiency (for a given variable and horizon) at a given significance level. We choose a 5% significance level, and report the average absolute magnitude of the difference in forecasts between always correcting and correcting conditional on the pre-test (see row 'Mean absolute correction: Always-Conditional' in table 3). The differences between the two strategies are relatively small.

Whether we always correct or correct conditional on the pre-test makes little difference to forecast accuracy on average. From the table, the ratio of the average MAE of the corrected forecasts to the reported is 0.95, and 1% more accurate at 0.94 for the pre-test corrected forecasts. In terms of forecast disagreement, always correcting tends to have a larger effect than the pre-test correction: the ratios of the average disagreement from applying each of these two strategies to the disagreement of the reported forecasts are 0.82 and 0.87 for current-quarter output growth.

While the effects of efficiency correction on the aggregate measures serve as a useful summary and show the effects to be sizeable, our primary focus is on individual-level performance. Generally, the effects of the corrections on aggregate quantities are not overly sensitive to whether the correction is applied after pre-testing, and this will be shown to be true of the individual-level findings as well.

4 Disagreement

There is a large literature on disagreement.[16] However with few exceptions each variable is considered in isolation. For our purpose the multivariate measure of disagreement of Banternghansa

and lag.

[16] 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).

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:

$$S_{t|t-h} = N_{t,h}^{-1} \sum_{i=1}^{N_{t,h}} \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right) \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)' \tag{5}$$

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:

$$D_{i,t|t-h} = \sqrt{\left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)' S_{t|t-h}^{-1} \left(y_{i,t|t-h} - \overline{y}_{t|t-h} \right)}. \tag{6}$$

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:

$$D_{i,t|t-h} = \sqrt{\sum_{j=1}^{n} \frac{\left(y_{j,i,t|t-h} - \overline{y}_{j,t|t-h} \right)^2}{S_{jj,t|t-h}}} \tag{7}$$

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.

Finally, if we set S to the identity matrix:

$$D_{i,t|t-h} = \sqrt{\sum_{j=1}^{n} \left(y_{j,i,t|t-h} - \overline{y}_{j,t|t-h} \right)^2} \tag{8}$$

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 (7) (or (6)). $S_{jj,t|t-h}$ will tend to exceed $S_{kk,t|t-h}$ if j denotes investment and k consumption, for example, because of the greater volatility of investment relative to consumption. 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 times. The use of $S_{t|t-h}$, calculated as in (5) 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 (6) or (7). 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 (7) 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 (6) for D gives $D = \sqrt{2/1.9}$, which is less than the $\sqrt{2}$ from using (7). 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 (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 (6) (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 (5) 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 (6) and (7) provide two alternative measures of individual-level disagreement. 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 measures may 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. (6)) 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 are qualitatively similar to using (6) with $S_{t|t-h}$ calculated as in (5).

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.

The detailed results for each individual are reported in the web-only Appendix. In summary, we find that the null is rejected for half the forecasters at the 10% level, and is still rejected for nearly 40% of the respondents at the 5% significance level, for the current-quarter forecasts. For $h = 4$, the rejection rate is also nearly 40% at the 5% level.

An alternative method of assessing the persistence in individual forecasting behaviour 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. The test statistic is described in the notes to table 4.

Here and elsewhere in the paper, when we test for persistence in differences in disagreement (this section) or accuracy (next section) 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. By making our efficiency corrections out-of-sample we guard against overstating their importance, at least in terms of improving forecast accuracy.

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 4 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 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').

Generally we reject the null of no relationship in the rankings of disagreement between the earlier and later sample periods (for both forecast horizons), panel A; and between the two forecast horizons, panel B. That is, there is persistence in individual contrarianism across time, and across horizons. The findings are the same whether we suppose S is diagonal, or use our preferred non-diagonal S measure, except for panel 1A, $h = 4$. The probability of obtaining a z statistic at least as large as that obtained is 0.029, for the diagonal measure, so formally we do not reject the null (at the 5% level in a two-sided test). For the diagonal measure the null is rejected.

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? We re-run the calculations in table 4, 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. We also apply the correction conditional on the outcome of the test for forecast efficiency.

Table 4 Panel 2 reports results for the corrected forecasts. 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 pre-test correction (table 4 Panel 3) gives qualitatively the same findings as always correcting.

The results for each individual are shown in the web-only Appendix. The efficiency correction (without pre-testing) reduces the proportion who differ from the median forecaster for the current quarter forecasts to around 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. For the $h = 4$ forecasts the efficiency correction has little affect on the proportion of rejections.

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 (6) or (7).

5 Forecast Accuracy

In this section we consider the micro-level evidence for the proposition that individuals' forecasts are equally accurate. Under some models of expectations formation, forecasters are predicted to be essentially identical.[17] If there are differences between individuals in terms of forecast accuracy, to what extent are these attributable to some forecasters using their information sets more efficiently than others?

Equal predictive accuracy is assessed in two ways. The first asks 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.[18] Komunjer and Owyang (2012) propose a multivariate loss function which allows for dependence between the different variables' forecast errors, and Sinclair, Stekler

[17] One such model is the 'basic' noisy information model under which individual agents receive homogeneous signals, have the same model of the economy, and use that information efficiently to generate their expectations.

[18] Hendry and Martinez (2017) develop an approach that could be used when there are few forecast observations.

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).

The second way compares each individual to the forecaster with the average level of accuracy using a Diebold and Mariano (1995) test. The details are confined to a web-only Appendix.

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 2008 Financial 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:

$$\widetilde{e}_{i,n,t+h|t} = \frac{e_{i,n,t+h|t}}{\sqrt{\frac{1}{N_{t,h}} \sum_{j=1}^{N_{t,h}} e^2_{j,n,t+h|t}}} \tag{9}$$

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:

$$\frac{1}{n_i} \sum_{t \in N_i} \widetilde{e}_{i,t+h|t} \widetilde{e}'_{i,t+h|t} \tag{10}$$

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.

The actual values used to calculate forecast errors are again either the initial estimates or the vintage-values published two quarters after the reference quarter.

5.1 Forecast Accuracy Results

Table 5 reports Spearman rank tests of the null that the rankings across the two sub-samples are unrelated. As expected, normalizing the forecast errors using (9) 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 (10)). We interpret this as suggesting the use of the raw forecast errors is misleading when forecasters face very different conditions, and that some forecasters generate more accurate forecasts than others.

Figure 3 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 (9), 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, 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. The results for applying the correction based on the pre-test for efficiency (Panel D) are qualitatively the same. There is now no evidence against the null at conventional significance levels.

The table of results for each individual in the web-only Appendix indicates that we reject the null (of equal accuracy to the average forecaster) for 40% of the forecasters for $h = 0$, at the 10% level, which is reduced to 24% when the (non-pre-test) efficiency correction is applied. For the $h = 4$ horizon the null is rejected for 16% of the forecasters, and this is halved when the efficiency correction is applied.

In summary, testing using a 5% significance level suggests forecast inefficiency accounts for the persistence in the accuracy rankings of forecasters which we observe. Nevertheless, the pairwise tests against the average-accuracy forecaster suggests that around a quarter of forecasters differ in terms of accuracy.

6 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, the evidence for persistence was considerably weakened: we do not reject the null of no persistence at the 5% level. Forecast inefficiencies explain differences in accuracy, without the need to assume some forecasters have better information, or better models, or have different behavioural motivations (that run counter to making accurate forecasts judged by squared-error loss).

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.

The results described in table 6 use the real-time efficiency correction described in section 3.3, and used hitherto.[19] 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 have a role to play in explaining inter-forecaster differences in forecast accuracy, consistent with the findings of section 5.

7 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, for example, if some forecasters received superior signals, 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 (6), 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 some forecasters benefit from superior private information and are both more contrarian and more accurate as a result. It does not rule out heterogeneous signals, of course, because forecasters might have motives other than minimizing squared error loss.

This finding holds up when the forecasts are corrected for inefficiency (as evident from the second panel of table 7). Failure to find more contrarian forecasters are more accurate is not due to an inefficient use of information.

[19] Here and from now on we only consider the results for the efficiency correction applied without pre-testing.

8 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 4 and 5. Table 8 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 4), 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 4 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 9 for the 25 forecasters with the original table 5). 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.

9 Conclusions

Some models of expectations formation, such as models which stress information rigidities, assume agents act rationally subject to certain constraints. 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. Our micro-level evidence suggests that approxi-

mately a half of the forecasters do not make efficient forecasts, where efficiency is defined as orthogonality between the forecasts and forecast errors. Nor are the forecasters essentially identical, either in terms of their degree of contrarianism, or predictive ability.

We documented 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 will in part be determined by the particular set of forecasters who are active at that time. The literature which considers disagreement between forecasters as a possible proxy for uncertainty (beginning with the seminal paper by Zarnowitz and Lambros (1987)), typically does not identify individual forecasters, implicitly assuming that any one forecaster is as likely to make the same contribution to overall disagreement at any point in time as any other.

We also establish that there are systematic differences between forecasters' accuracy.

A key focus of our paper 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 - that more (less) contrarian forecasters in the first period remain so in the second period. 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. The inefficient use of information plays an important role in explaining 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. The evidence strongly suggests forecasters who stand out from the crowd do not tend to produce more accurate forecasts. More contrarian forecasters are not better informed.

The micro-level evidence suggests macro-forecasters are not 'essentially the same' as each other. The effect of inefficiency is 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.

A number of recent papers consider whether forecasters over- or under-react to new information. Such behaviour suggests forecaster inefficiency - a correlation between the forecast errors and forecasts. Evidence for over- or under-reaction is often garnered from regressions of forecast errors on forecast revisions. A finding of a non-zero coefficient indicates inefficiency. Our paper complements this literature by exploring the practical relevance of the rejections - evidence of forecaster inefficiency - in terms of the extent to which it accounts for forecaster heterogeneity (persistent differences across forecasters in terms of accuracy or contrarianism). This serves to

quantify the importance of those rejections. Our paper suggests the over-reaction to new information of Broer and Kohlhas (2018) and Bordalo *et al.* (2018), and the under-reaction found by Fuhrer (2018), explains an important part of the differences across forecasters in accuracy.

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.

Bernhardt, D., Campello, M., and Kutsoati, E. (2006). Who herds?. *Journal of Financial Economics*, *80*(3), 657–675.

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.

Bordalo, P., Gennaioli, N., Ma, Y., and Shleifer, A. (2018). Over-reaction in Macroeconomic Expectations. NBER Working Papers 24932, National Bureau of Economic Research, Inc.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2017). Diagnostic Expectations and Stock Returns. NBER Working Papers 23863.

Broer, T., and Kohlhas, A. (2018). Forecaster (Mis-)Behavior. Cepr discussion papers 12898, C.E.P.R. Discussion Papers.

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.

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.

Crowe, C. (2010). Consensus forecasts and inefficient information aggregation. *IMF Working Paper*, **WP/10/1787**.

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.

Diebold, F. X., and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **13**, 253–263.

Dietrich, J. K., and Joines, D. H. (1983). Rational Expectations, Informational Efficiency, and Tests Using Survey Data: A Comment. *The Review of Economics and Statistics*, *65*(3), 525–529.

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 Econo-*

metrics, *26*(7), 1059–1078.

Figlewski, S., and Wachtel, P. (1981). The Formation of Inflationary Expectations. *The Review of Economics and Statistics*, *63*(1), 1–10.

Figlewski, S., and Wachtel, P. (1983). Rational Expectations, Informational Efficiency, and Tests Using Survey Data: A Reply. *The Review of Economics and Statistics*, *65*(3), 529–531.

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.

Fuhrer, J. C. (2018). Intrinsic Expectations Persistence: Evidence from Professional and Household Survey Expectations. Working papers 18-9, Federal Reserve Bank of Boston.

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.

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.

López-Pérez, V. (2016). Does uncertainty affect participation in the European Central Bank's Survey of Professional Forecasters?. *Economics - The Open-Access, Open-Assessment E-Journal, Kiel Institute for the World Economy (IfW)*, **10**, 1–47.

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.

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: Forecast Data and the Effects of Efficiency Corrections: Aggregate Findings.

	$h=0$			$h=4$		
	Cons.	Invest.	Output	Cons.	Invest.	Output
Mean actual value	0.685	0.770	0.586	0.704	0.860	0.619
MAE reported forecasts	0.345	1.261	0.317	0.366	1.578	0.401
MAE/Actual value	0.505	1.637	0.541	0.520	1.835	0.647
Disagreement	0.235	0.819	0.216	0.192	0.654	0.198
Mean absolute correction:						
Always-Reported	0.197	0.593	0.144	0.216	1.068	0.238
Always-Conditional	0.052	0.327	0.056	0.042	0.096	0.013
MAE:						
Corrected/Reported	0.930	0.969	0.950	0.956	0.951	0.900
Condit. Corrected/Reported	0.930	0.950	0.940	0.934	0.946	0.898
Disagreement:						
Corrected/Reported	0.804	0.941	0.824	1.052	1.436	0.960
Condit. Corrected/Reported	0.834	0.965	0.870	0.943	1.433	0.949

The MAE (mean absolute error) of the reported forecasts is the average absolute error across respondents and surveys. Disagreement is the average over surveys of the cross-section standard deviations. The Mean absolute correction - Always vs Reported, is the average absolute correction across individuals and time periods. The Mean absolute correction - Always vs Conditional is the average absolute difference between the efficiency-corrected forecasts, and the efficiency-corrected forecasts conditional on a test for forecast efficiency, across individuals and time periods.

We then report the ratio of the MAE of the Corrected forecasts to that of the reported, and of the efficiency-corrected forecasts conditional on a test for forecast efficiency to that of the reported forecasts.

The final two rows are the ratio of disagreement of the Corrected forecasts to that of the reported, and of the efficiency-corrected forecasts conditional on a test for forecast efficiency to that of the reported forecasts.

Table 4: Rank Correlation Tests of Multivariate Disagreement

Panel 1. 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 2. Efficiency-Corrected Forecasts			
Panel 2A. 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

Panel 3. Conditional Efficiency-Corrected Forecasts			
Panel 3A. Earlier and later periods			
S		Diag. S	
$h=0$	$h=4$	$h=0$	$h=4$
0.454	0.401	0.417	0.471
0.008	0.018	0.014	0.006

Panel 3B. $h=0$ and $h=4$ forecasts					
S			Diag. S		
Whole	Earlier	Later	Whole	Earlier	Later
-0.025	0.129	-0.233	-0.016	0.077	-0.136
0.565	0.235	0.927	0.543	0.334	0.800

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 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 5: 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
Panel D. Correction based on Test for Efficiency			
$h=0$		$h=4$	
Tr.	Det.	Tr.	Det
0.234	0.290	-0.393	-0.049
0.114	0.066	0.982	0.598

The forecast errors in panels B, C and D are normalized. Panel A reports accuracy measures based on the raw errors. In Panel C all the forecasts have been corrected. In panel D only the forecasts of respondents for whom we reject forecast efficency (at the 5% level) are corrected.
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 4 for an explanation.
Normalized denotes that the forecast errors have been adjusted for differences over time in average forecast accuracy.

Table 6: 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, and the second is the probability of observing a larger value: see notes to table 4 for an explanation.

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, and the second is the probability of observing a larger value: see notes to table 4 for an explanation.

Table 8: 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 4, 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 4 for an explanation.

Table 9: 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 5, 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 4 for an explanation.

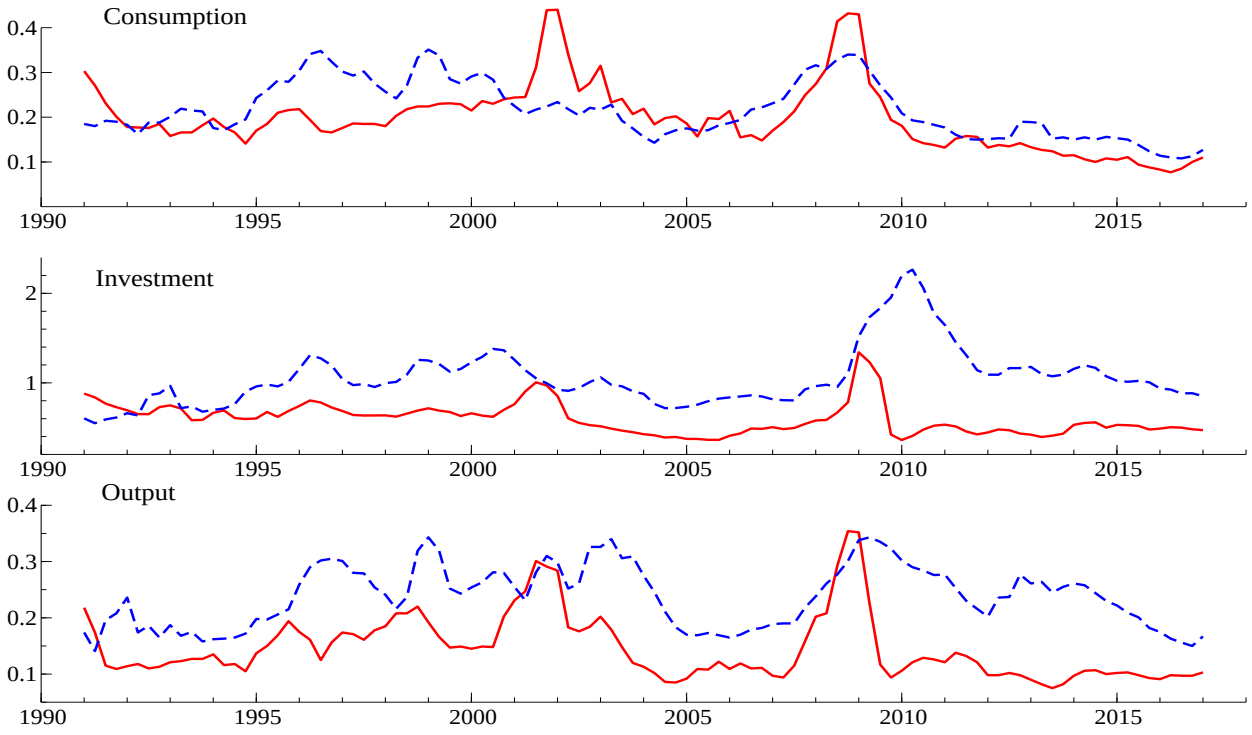

Figure 1: Cross-sectional means of the absolute efficiency corrections at each survey date (smoothed using a centred moving average with leads and lags of 1). $h = 0$ is the solid line. $h = 4$ is the dashed line

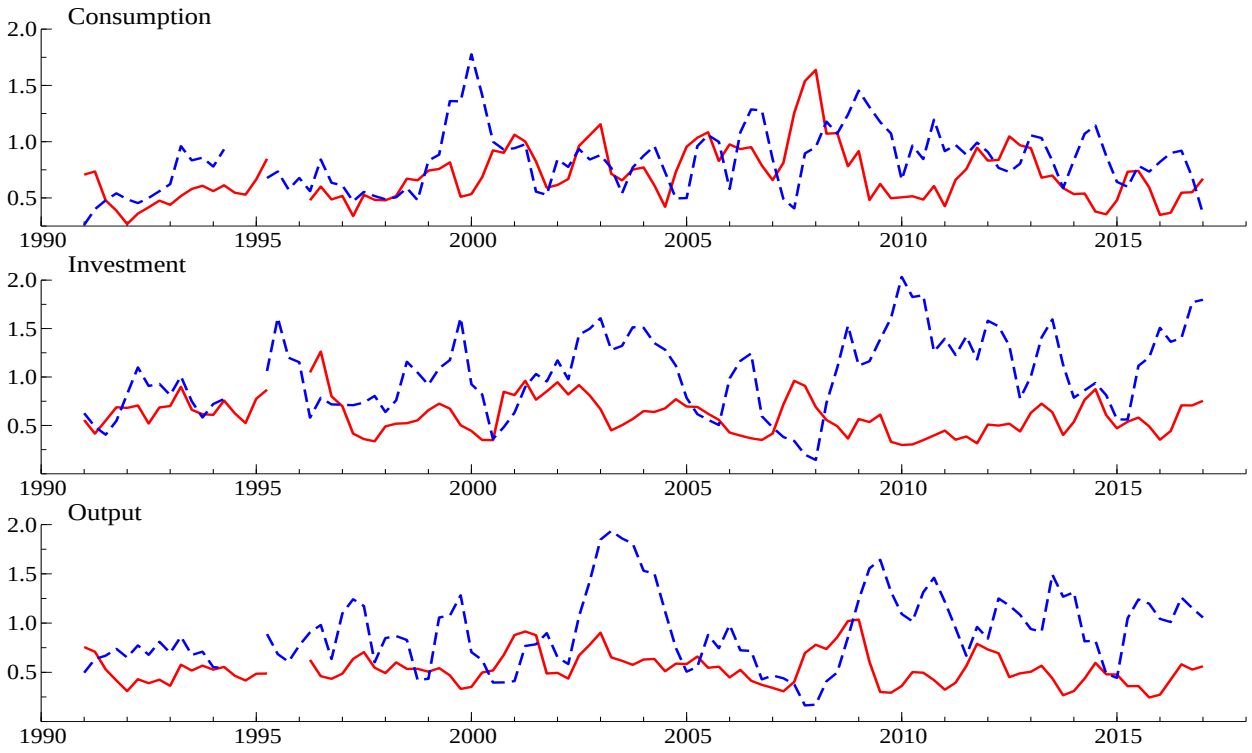

Figure 2: Ratio of the cross-sectional means of the absolute efficiency corrections to the cross-sectional averages of the absolute forecast errors (smoothed using a centred moving average with leads and lags of 1). $h = 0$ is the solid line. $h = 4$ is the dashed line

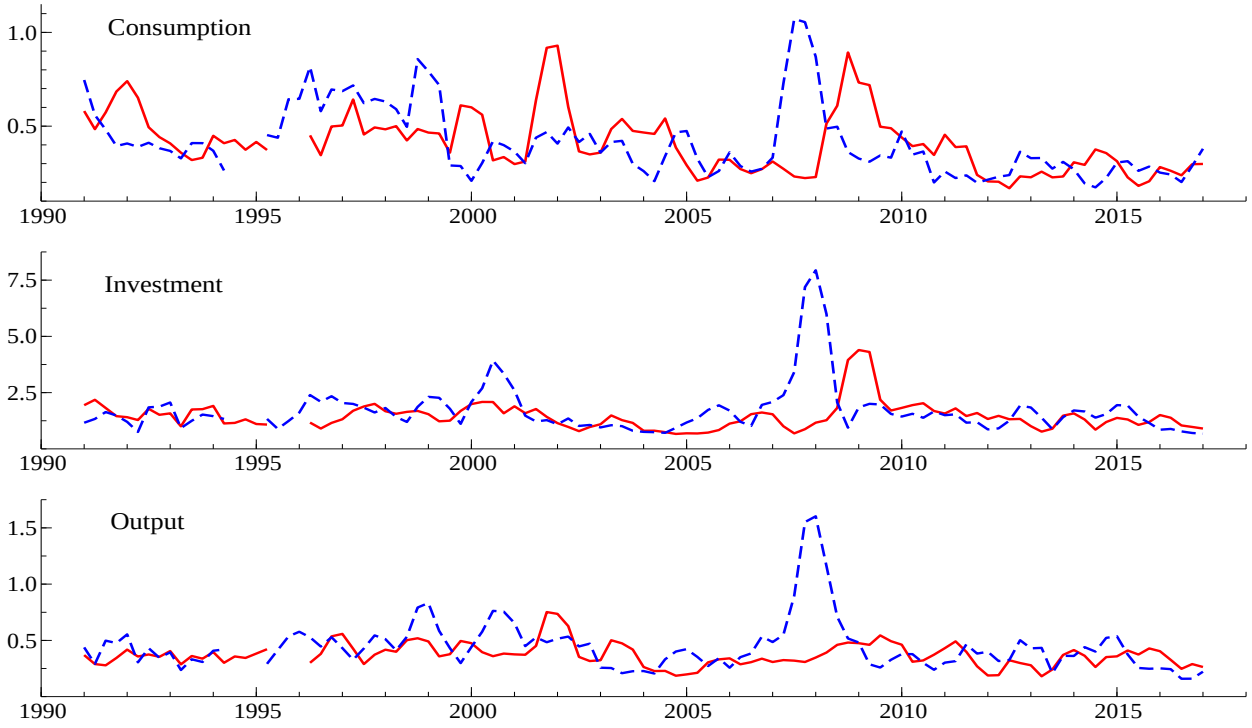

Figure 3: Cross-sectional root mean squared error at each survey date (smoothed using a centred moving average with leads and lags of 1). $h = 0$ is the solid line. $h = 4$ is the dashed line