International Journal of Forecasting

All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance

--Manuscript Draft--

Manuscript Number:	IJF-D-23-00164R3
Full Title:	All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance
Article Type:	Full Length Article
Keywords:	forecasting profession; surveys; Evaluating forecasts; Point forecasts; Density forecasts; persistent heterogeneity
Abstract:	Abstract: Expectations models incorporating information rigidities typically imply the absence of systematic patterns in individual forecast behavior. Motivated by this consideration, we examine the European Central Bank's Survey of Professional Forecasters to investigate if participants are interchangeable. The evidence indicates participants display distinguishing behaviors over time, both within and across target variables and horizons. Moreover, the systematic patterns in predictive performance are strongly linked to the degree of difficulty in the forecasting environment, which is a new finding in professional forecast surveys. Our results argue for the development of expectations models that can generate persistent heterogeneity of this form.

August 13, 2024

Professor Norman Swanson
Editor
International Journal of Forecasting

Dear Professor Swanson:

We would like to thank you, the Associate Editor, and the three referees for the many helpful comments and suggestions on our submission "All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance" (IJF-D-23-000164). As requested, we have included our responses to the referee's comments which are contained in the attached files.

We wanted to take this opportunity to provide some additional discussion of the summary of the Associate Editor and the referee reports.

The Associate Editor provided the following summary in the last report:

I am in a quandary. All three referees are experts in the area. Two of them see serious issues with the revised version of the paper and suggest the paper is beyond redemption. I see merits in their arguments. On the other hand, since the Hounyo-Lahiri (JMCB) paper came out after the authors first circulated their first draft, I am wondering if the authors can revise the paper to satisfy the two critical referees. Needless to say, we cannot assure that additional work will make the paper acceptable for publication in IJF. The revision will certainly be non-trivial. I am sorry that the situation turned out this way. I will leave the decision to resubmit to the discretion of the authors.

My co-author (Joe Tracy) and I have discussed the referees' comments from the two rounds of the review process, and we are unclear about the basis for the Associate Editor's view concerning the gravity and extent of the deficiencies of the paper. It now appears that this view is largely shaped by Reviewer 1 and Reviewer 3. Consequently, we would like to provide you with an overview of the referees' concerns and how we have addressed them in the latest version of the paper. This overview is intended to complement to the more detailed responses that we have provided to the referees.

Reviewer 1: Reviewer 1 has principally raised concerns about the contribution of the paper compared to the recent work of Hounyo and Lahiri (2023). We recognize that there is some overlap in methodologies pertaining to the unbalanced panel of the survey as well as the cross-sectional and time series dependence in the errors. However, we have argued that our work extends beyond Hounyo and Lahiri (2023) in several important ways. For example, we have recast the paper in terms of the interchangeability of forecasters which motivates us to consider more than just the issue of equal predictive

ability for our investigation into the properties of forecast behavior. In addition, we view the adoption of Pesaran's (2006) common correlated effects estimator and its application to the ECB-SPF as novel, with the resulting four-quadrant partition of the individual respondent's (α, λ) pairings as a particularly useful and insightful device to characterize various features of their predictive performance. Moreover, we would contend that our documentation of systematic patterns in predictive performance through inter- and intra-forecaster comparisons (the former of which now extends to first and second moments of predictive performance metrics in our test for 'distributional homogeneity' which is discussed further below for Reviewer 3), the linkage of the systematic patterns to variation in the difficulty of the forecasting environment, and evidence of instability in the rank ordering of forecasters are notable and significant contributions to the literature. We have provided a detailed listing in our response to the referee as well as rewritten the Introduction to reflect these same points to the reader.

Regarding the specific comments of the referee from the last round, we now provide an example (footnote 6) illustrating our claim that the normalized (squared) metric involves an asymmetric treatment of accuracy at the individual level versus the aggregate level.

Reviewer 2: We addressed the two minor points raised by the reviewer and we believe that there are no outstanding issues.

Reviewer 3: After careful review, we agree with the referee that the null hypothesis $\alpha_j = 0 \bigcap \lambda_j = 1$ for j=1, . . . N respondents is a joint test of equal prediction performance and equal variance of the performance metric. Drawing upon the referee's comments in the first and second rounds of the review process, the Appendix (Section A.1) now contains a formal proof for this claim and we refer to this property of the data as 'distributional homogeneity'. However, this change in the interpretation of our testing procedure is a benefit to the paper which we have recast in terms of the interchangeability of forecasters. As we note, the implications of IR models apply as much to the variance as the mean of the distribution of the predictive performance metric which the referee previously described as the conditions that $E(FP^{(1)}_{t+1|t}) = E(FP^{(2)}_{t+1|t})$ and $E(FP^{(1)}_{t+1|t})^2 = E(FP^{(2)}_{t+1|t})^2$

Because of the joint nature of the testing procedure, the referee expressed a concern that heterogeneity in the variance of the predictive performance metric could generate a rejection of the null hypothesis even if participants display equal predictive ability. To address this concern, we apply the testing procedure of Hounyo and Lahiri (2023) to the ECB-SPF and report results in the Appendix (Section A.2) that strongly reject the null hypothesis of equal predictive ability. Consequently, the rejection of the joint null hypothesis in Tables 1-2 does not mask equal predictive ability among the participants. We also note that the nature of the rejection of equal predictive ability for

the ECB-SPF is very similar to that reported by Hounyo and Lahiri (2023) for the US-SPF.

Thank you again for the time extensions that you afforded us for the revision process and please contact me if you have any questions or would like to discuss anything further.

Reviewer 1.

IJF-D-23-00164

"All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance"

We would like to thank the referee for your very helpful comment. Our response to your comment (stated in italics) is provided below:

1. *The current version of the manuscript represents a substantial improvement over its predecessor, although its contribution to the existing body of literature may not be deemed as overwhelming. I would like to offer the following comment:*
   On page 7, the authors assert, 'On a more general level, another aspect of Hounyo and Lahiri (2023) to consider centers on their use of the metric in equation (1). Specifically, the normalized metric in (1) involves an asymmetric treatment of accuracy at the individual level versus the aggregate level. In essence, a forecaster who makes a relatively large error when the average forecast error is small will incur a substantial penalty, whereas a forecaster who makes a relatively small error when the average forecast error is large will not benefit significantly.'

   This comparison may not be entirely equitable when evaluating Hounyo and Lahiri's (2023) approach. It is essential to note that the bootstrap-based testing procedure outlined by Hounyo and Lahiri (2023, cf. Algorithm 2) is versatile and can be applied using various measures of forecast error, including normalized squared error, absolute error, squared error, etc.
   Furthermore, could you kindly provide clarification regarding the statement 'whereas a forecaster who makes a relatively small error when the average forecast error is large will not benefit much'? Offering an example or justification for this claim would significantly enhance the understanding of this point.

   On a general level, we recognize that there is some overlap in Hounyo and Lahiri (2022) and our methodology regarding the unbalanced panel of the survey as well as cross-sectional and time series dependence in the errors. However, we would contend that our work extends far beyond Hounyo and Lahiri (2022) and we have rewritten the Introduction to clarify and highlight the contributions of the paper to the Literature which we briefly summarize here:
   1) While the evaluation of forecast performance is an issue of longstanding interest and importance, we further motivate the analysis by drawing upon expectations models featuring information rigidities and their implications for forecaster interchangeability.
   2) We find that survey participants display systematic patterns in their forecast behavior and document a new result that the patterns are strongly linked to the degree of difficulty in the forecasting environment.
      a. Like other studies, we make inter-forecaster comparisons of predictive performance within a target variable.
         i. We demonstrate that the restriction $(\alpha_j, \lambda_j) = (0,1)$ provides a new and very convenient/straightforward joint test of equal prediction

performance and equal variance of the performance metric (which we refer to as 'distributional homogeneity') which moves beyond conventional consideration of only the first moments of accuracy. Because of the joint nature of the test, we also apply the testing procedure of Hounyo and Lahiri (2022) to demonstrate that the rejection of the joint null hypothesis does not mask equal predictive performance across participants.

ii. While we reject equal predictive performance using the Hounyo and Lahiri (2022) testing procedure, we demonstrate that the rejection does not reflect a group of respondents with superior forecasting skills and another group of respondents with inferior forecasting skills. Rather, the source of the rejection stems from some forecasters being relatively more accurate in tranquil environments and other forecasters being relatively more accurate in volatile environments.

b. In contrast to most studies, we extend the analysis by also making intra-forecaster comparisons. Importantly, we find that the relationship between a participant's relative predictive performance and the nature of the forecasting environment is similar across variables and horizons. That is, respondents who are relatively more accurate in a tranquil (volatile) environment for a particular variable and horizon tend to be relatively more accurate in a tranquil (volatile) environment across different variables and horizons.

c. Taken together, we view the results in 2)a and 2)b as providing strong evidence that that the observed features of ECB-SPF data at the individual level are inconsistent with expectations models with information rigidities and their implication of forecaster interchangeability.

3) The link between the forecasting environment and predictive performance documented in our study also bears upon another important finding. In general, the rank orderings of participants based on predictive performance are not stable over time but instead vary with the difficulty of the forecasting environment. Consequently, the predictive performance profiles of participants typically display crossings with each other.

a. We illustrate this feature of forecast performance in two ways. First, Figure 8 depicts the predictive performance profile of 4 selected respondents who provide one-year-ahead point forecasts of GDP growth. Table 3 then extends the coverage to consider all respondents and provides a rank ordering evaluated at eight different values spanning the forecasting environment.

b. We view the evidence of notable time variation in the rank ordering of forecasters as a cautionary note for studies evaluating predictive performance from surveys. Specifically, our results suggests that conventional testing procedures may mask variation in participants' relative forecast accuracy, with the reliability of resulting inferences and conclusions sensitive to the prevalence of tranquil and volatile forecasting environments in the selected sample.

4) Our empirical framework has several attractive features:

a. The four-quadrant partition of the (α, λ) pairings provides a unique and extremely convenient device to characterize and illustrate the relationship

between a forecaster's predictive performance and the difficulty of the forecasting environment. The four-quadrant partition is not only useful for determining the relative prevalence of different types of forecasters within a target variable, but also for quantifying similarities in predictive performance of a forecaster across target variables.

   b. The cross-sectional average of forecast performance $(\overline{FP})$ facilitates our identification of the forecasting environment in terms of tranquil vs. volatile episodes.
   c. It allows us to nest and subsequently reject the adequacy of time effects and normalized predictive performance metrics to control for cross-sectional dependence in the data.

5) In contrast to Hounyo and Lahiri (2022) who examine point forecasts from the US-SPF, we examine the ECB-SPF and include density forecasts as a robustness check. While our evidence of systematic patterns in participants' predictive performance is robust to the application Hounyo and Lahiri's (2022) testing procedure, we think it is interesting and instructive to learn that the nature of our rejection of forecaster homogeneity in the ECB-SPF is similar to their rejection for the US-SPF.
6) The Appendix provides several additional discussion points that include a formal proof for our proposed test restriction for distributional homogeneity of the predictive performance metrics, a full reporting of the results using the testing procedure of Hounyo and Lahiri (2022), and an analysis of forecast combinations using weights that depend on a respondent's predictive performance across variation in the forecasting environment. Abstracting from the absence of real-time considerations in the selected weighting scheme, we nevertheless demonstrate that our forecast combination allows us to achieve the rare outcome of outperforming the simple average.

Turning to the referee's more specific comment, we agree that one of the attractive features of the approach of Hounyo and Lahiri (2023) is that it is versatile and can be applied to a range of forecast performance metrics. Our discussion point, however, was not directed at the approach of Hounyo and Lahiri (2023) but rather at a property of the normalized squared forecast error. As the referee has requested, the following example is intended to clarify our statement concerning the asymmetric treatment of accuracy at the individual level vs. the aggregate level.

- If an individual's forecast error is 0.5 percentage point and the average forecast error is 2 percentage points, then the individual's normalized forecast error decreases to 0.25 percentage point and results in a 'benefit' of 0.25 percentage point.
- If an individual's forecast error is 2 percentage points and the average forecast error is 0.5 percentage point, then the individual's normalized forecast error increases to 4 percentage points and results in a 'penalty' of 2 percentage points.

The two scenarios involve a situation where individual accuracy differs by 1.5 percentage points in either direction from average accuracy, but there is an asymmetric treatment of the adjustment. As noted in our paper, the linear framework of our regression model results in a

symmetric treatment of accuracy at the individual level vs. aggregate level. We have included this discussion in footnote 6 of the paper.

Reviewer 2.

IJF-D-23-00164

"All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance"

We would like to thank the referee for your very helpful comments. Our response to your comments (stated in italics) are provided below:

Reviewer 2: I only have two small points:

1. *On page 13, lines 27-30 it states that "the survey is fielded in February, May, August and November, with a little under 50 panelists on average responding per survey". I think these months are the US SPF and the ECB is January, April, July and October. Also I think the number of respondents in the ECB SPF is above 50, see https://www.ecb.europa.eu/pub/economic-bulletin/articles/2019/html/ecb.ebart201901_01~8300a24082.en.html*

   We would like to thank the referee for bringing both points to our attention as well as for providing the link to the document. The referee is correct that the ECB-SPF survey is fielded in January, April, July, and October and that approximately 55 responses are received per quarter. We have made these corrections in the paper.

2. *Despite the title change "systematic" instead of "persistent", the term "persistent" is still used a number of times in the paper (including the abstract) when it is not necessarily persistence. It would only be persistent to the extent that the "degree of difficulty in the forecasting environment" is persistent.*

   We agree with the referee and we have largely modified the text accordingly. There are, however, a couple places in the text where we use the phrase "persistent performance heterogeneity" which is borrowed directly from Hounyo and Lahiri (2023). The authors use this phrase to describe their results that reject equal forecast ability for the US-SPF.

Reviewer 3.

IJF-D-23-00164

"All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance"

We would like to thank the referee for the time and effort put into previous reviews, with the many thoughtful and helpful comments helping us to improve the paper. We have two responses concerning questions about the formulation of the joint null hypothesis $\alpha_j = 0$ and $\lambda_j = 1 \quad \forall\ j$ to test for equal predictive ability.

1. As we have noted in our reply to the Editor, we found after careful review that the referee was correct and that the null hypothesis $\alpha_j = 0 \bigcap \lambda_j = 1$ for j=1, . . . N respondents is a joint test of equal prediction performance and equal variance of the performance metric. Drawing upon the referee's comments, the Appendix (Section A.1) now contains a formal proof for this claim which we refer to as 'distributional homogeneity' of the predictive performance metrics. As a point of emphasis, the proof incorporates the following point made by the referee in the first round of review: "*If* $E(FP^{j}_{t+h|t})$ *is the same for all j, then this means that that* $\alpha_j + \lambda_j E(\overline{FP_{t+h|t}})$ *does not depend on j*". We use this point to impose the condition that $\alpha_i = \alpha_k = \alpha$ and $\lambda_i = \lambda_k = \lambda$.
However, we would argue that the change in the interpretation of our testing procedure is a benefit to the paper which we have recast in terms of the interchangeability of forecasters. As we note, the implications of IR models apply as much to the variance as the mean of the distribution of the predictive performance metric which the referee previously described as the conditions that $E(FP^{(1)}_{t+1|t}) = E(FP^{(2)}_{t+1|t})$ and $E(FP^{(1)}_{t+1|t})^2 = E(FP^{(2)}_{t+1|t})^2$

2. With regard to the joint nature of the testing procedure, the referee also expressed a concern that heterogeneity in the variance of the predictive performance metric could generate a rejection of the null hypothesis even if participants display equal predictive ability. To address this concern, we apply the testing procedure of Hounyo and Lahiri (2023) to the ECB-SPF and strongly reject the null hypothesis of equal predictive ability. Consequently, the rejection of the joint null hypothesis in Tables 1-2 do not mask equal predictive ability among the participants. We also note that the nature of the rejection of equal predictive ability for the ECB-SPF is very similar to that reported by Hounyo and Lahiri (2023) for the US-SPF.

All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance

January 28, 2025

Abstract: Are all forecasters the same? Expectations models incorporating information rigidities typically imply forecasters are interchangeable which predicts an absence of systematic patterns in individual forecast behavior. Motivated by this prediction, we examine the European Central Bank's Survey of Professional Forecasters and find, in contrast, that participants display systematic patterns in predictive performance both within and across target variables. Moreover, we document a new result from professional forecast surveys which is that inter- and intra-forecaster relative predictive performance are strongly linked to the degree of difficulty in the forecasting environment. This insight can inform the ongoing development of expectations models.

Keywords: forecasting profession; surveys; evaluating forecasts; point forecasts; density forecasts; heterogeneity

I. Introduction

Expectations are important for understanding the decision-making of households and firms, as well as for explaining movements in economic and financial variables. Early work on the formation of beliefs posited that agents form their expectations in a static or adaptive manner. However, these models eventually drew criticism because of their restrictions on agents' information sets and for allowing agents to make systematic forecast errors. In response, the full-information rational expectations (FIRE) model was developed which assumes that all agents know the true structure of the economy and have access to the same information set.

While the FIRE model remains the main paradigm for the formation of expectations, it implies that agents display identical forecast behavior and, therefore, cannot generate the type of dispersion in agents' expectations – that is, disagreement – observed in surveys or financial markets. Consequently, in recent years the FIRE model has been replaced with a weaker form of rational expectations in which agents use available information efficiently subject to certain constraints. A prominent feature of these models is the presence of informational rigidities (IR) either in the form of sticky information [Mankiw and Reis (2002); Mankiw, Reis, and Wolfers (2003)] or noisy information [Woodford (2003); Sims (2003); Mackowiak and Wiederholt (2009)].

While IR models can generate disagreement, a key, but largely overlooked, implication of almost all these models is that heterogeneity in individual forecast behavior should not display systematic patterns.[1] This is because variation in forecast behavior only arises from randomness either in the updating of individual information sets or in the configuration of shocks faced by individuals. Agents can display differences in their forecast behavior at a point in time, but their forecast behavior should be the same on average over time. Consequently, forecasters should be viewed as interchangeable with no distinguishing patterns in their average observed behavior.[2]

Motivated by this consideration, this study uses data from the European Central Bank's Survey of Professional Forecasters (ECB-SPF) to explore the implications of interchangeability across three aspects of forecast behavior. The first is whether the predictive performance metrics of participants display "distributional homogeneity" within a target variable – that is, do the mean and

[1] Coibion and Gorodnichenko (2012, 2015) test the predictions of the sticky information and the noisy information models for various aspects of forecast behavior at the aggregate level, but they do not consider this implication of IR models at the individual level.

[2] Clements (2022) makes this same observation to motivate his analysis of the US Survey of Professional Forecasters. The average observed behavior of forecasters refers to a period long enough to allow individuals to update their information sets on a comparable basis in the case of the sticky information model, or to be subject to a comparable set of shocks in the case of the noisy information model.

variation over time in a participant's accuracy align with those of others on average.[3] The second is whether the relative accuracy of forecasters displays systematic patterns within a target variable. The third is whether individual forecasters display similar behavior across target variables. While these empirical features are of general interest for expectations models, we have noted their relevance for IR models.

We conduct the analysis within a panel data framework using the common correlated effects (CCE) estimator of Pesaran (2006). The CCE modeling strategy is attractive for several reasons. First, it allows for a broad identification of heterogeneity and correlation patterns in participants' predictive performance. Second, the inclusion in the individual regressions of a time-specific cross-sectional average of predictive performance controls for aggregate shocks that can generate dependence across participants which is a typical concern in panel data models. Last, the average predictive performance variable also provides a natural basis to identify tranquil/volatile forecast episodes that play a critical role in the analysis.

The results provide strong evidence that ECB-SPF participants are not interchangeable. Our tests reject the property of distributional homogeneity which indicates there are significant differences across forecasters in the mean and variance of their predictive performance metrics within a target variable. There are also systematic patterns in participants' relative predictive performance over time, both within and across target variables. Moreover, we document a new finding that these systematic patterns are strongly linked to the degree of difficulty in the forecasting environment. Within a target variable, some participants display higher relative accuracy in tranquil episodes, while other participants display higher relative accuracy in volatile episodes. Across target variables, we find that participants who display higher (lower) relative accuracy in tranquil/volatile environments for one target variable tend to display the same behavior for the other target variables. These results pose a challenge to IR models.

Our results are consistent with recent work by Hounyo and Lahiri (2023) who test for equal predictive ability among participants in the US Survey of Professional Forecasters (US-SPF) and report evidence of "persistent performance heterogeneity". Hounyo and Lahiri (2023) improve upon the bootstrap technique of D'Agostino, McQuinn and Whelan (2012) by allowing for cross-sectional and serial correlation in the forecast errors that can otherwise lead to incorrect inference. Taken together, the findings in our study and Hounyo and Lahiri (2023) contrast with previous evidence

[3] We use the term "target variable" to denote the combination of an outcome variable (e.g., GDP growth) and forecast horizon (e.g., one-year-ahead horizon).

presented by Kenny, Kostka and Masera (2014) and Meyler (2020) for the ECB-SPF and D'Agostino, McQuinn, and Whelan (2012) for the US-SPF.

There are, however, several important differences between our study and Hounyo and Lahiri (2023). Here we provide a short comparison and defer a more detailed discussion until Section II. Beyond methodology and data sets, one difference pertains to focus. The empirical framework of Hounyo and Lahiri (2023) is designed to test for equal predictive performance within a target variable. In contrast, we develop and conduct a more stringent test of comparable forecast behavior by considering both first and second moments of predictive performance. The ability to discriminate between participants who may display equal accuracy but differ along other dimensions of forecast behavior is an important extension in the evaluation of expectations models. Our investigation also extends beyond the inter-forecaster comparisons in Hounyo and Lahiri (2023) by considering intra-forecaster comparisons across target variables as an additional basis to explore the issue of the interchangeability of forecasters.

Another difference with Hounyo and Lahiri (2023) concerns the rank ordering of participants. While their approach allows for the identification and ranking of participants with superior or inferior forecasting skills during the sample period, it is silent on whether the ordering is stable over time. In contrast, our approach allows for a deeper investigation into the behavior of the rank orderings. We find that the bulk of the rank ordering of participants changes with variation in the forecasting environment. This result suggests that forecaster comparisons can be sensitive to the relative prevalence of tranquil and volatile episodes in a selected sample period and offers a cautionary note for studies that assume rank orderings are largely stable.

We conclude that models featuring information rigidities and their implication for forecaster interchangeability are not consistent with observed features of the ECB-SPF. The mean and variation over time in participants' accuracy do not align with each other on average. In addition, we document systematic patterns in individual predictive performance that are strongly linked to the degree of difficulty in the forecasting environment. While such behavior could reflect heterogeneity in participants' loss functions or the use of different models, a deeper exploration into this line of research is beyond the scope of this paper.[4] We do, however, investigate the possibility of non-uniform processing capacity across forecasters that is, in turn, related to differential private information [Clements (2022)]. Taken together, our study principally contributes to a large literature

[4] While strategic behavior could offer another explanation for these features, the anonymity of the ECB-SPF forecasters would likely rule out this explanation.

that uses survey data to inform the ongoing development of models of expectations formation, with particular focus on uncovering new facts about predictive performance.

The paper is organized as follows. The next section discusses the modeling strategy and estimation framework used for the empirical analysis. Section III provides a summary of the literature evaluating various aspects of professional forecasters' predictive performance. Section IV describes the ECB-SPF data. Section V reports the estimation results and documents systematic patterns in participants' relative predictive performance and how it varies with the difficulty of the forecasting environment. This section also explores the behavior of the rank ordering of forecasters. Section VI concludes by discussing the implications of our findings.

II. Modeling Strategy and Estimation Framework

Our modeling strategy and estimation framework are motivated by the survey-based predictive performance metric introduced by D'Agostino, McQuinn, and Whelan (2012) for the US-SPF and adopted by Hounyo and Lahiri (2023). A key aspect of both analyses concerns the challenge of evaluating predictive performance when there is time variation in the forecasting environment. Specifically, participants generating the same prediction error in different periods will not reflect equal predictive ability if forecasting in some periods is easier/more difficult as compared to others. The evaluation of predictive performance is further complicated in an unbalanced panel setting due to the entry and exit of participants either on an intermittent or permanent basis.

To account for both considerations, D'Agostino, McQuinn, and Whelan (2012) originally proposed an adjustment to conventional predictive performance metrics. Their approach begins by constructing a normalized forecast error statistic for each variable, period, and participant. While we discuss their methodology within the context of point forecasts, it can also be applied to density forecasts. Abstracting from details related to data and survey features, the normalized squared error statistic for participant *j*, $(e^{j}_{t+h|t})^2$, is given by:

$$(1) \qquad (e^{j}_{t+h|t})^2 = \frac{(e^{j}_{t+h||t})^2}{(1/N_t)\sum_{i=1}^{N_t}(e^{i}_{t+h||t})^2} = \frac{(e^{j}_{t+h||t})^2}{\overline{(e_{t+h|t})^2}}$$

where $e^{j}_{t+h|t}$ is participant *j*'s forecast error associated with the survey point prediction in period *t* and the realization of the target variable in period *t+h,* and $\overline{(e_{t+h|t})^2}$ is a measure of average forecast performance defined over the N_t survey participants in period *t.* Importantly, the metric in (1)

depends on participant *j*'s forecast performance relative to the other forecasters. Consequently, the normalization is designed to control for changes in the forecasting environment by generating, for a given value of $(e^{j}_{t+h|t})^2$, a value of $(e^{j}_{t+h|t})^2$ that is lower (higher) when forecasters are collectively less (more) accurate compared to periods when they are more (less) accurate.

For each forecaster, we can calculate a score by taking an average of the normalized squared error statistics. Letting T^j denote the total number of surveys in which participant *j* appears and T denote the total number of surveys conducted, the score of participant *j* is defined as:

$$S^j = \left(1/T^j\right)\sum_{t=1}^{T}(\tilde{e}^{j}_{t+h|t})^2 \quad , \tag{2}$$

where $(e^{j}_{t+h|t})^2$ is set to zero if participant *j* did not respond to that survey. Because the performance score in (2) is calculated as an average, it can account for a participant entering or exiting a survey.

D'Agostino, McQuinn, and Whelan (2012) derive a historical distribution of forecast performance using the score in (2) and the associated rank ordering of all participants. A test for equal ability proceeds by randomly reshuffling and reassigning individual forecasts of a given variable for a particular survey. The same procedure is applied to each survey, resulting in a new sequence of forecasts for each participant that can be used to calculate an overall score from (2) and construct a rank ordering. The process is repeated many times to generate a large number of simulated distributions of forecaster performance, with the test for equal ability comparing the historical distribution of forecast performance to the simulated distributions. Under the null hypothesis of equal ability, the historical distribution of forecast performance should lie within selected percentiles of the simulated distribution that serve as confidence intervals. D'Agostino, McQuinn, and Whelan (2012) find little evidence that the best forecasters are significantly better than others, although there is a relatively small group of forecasters that perform very poorly.

While the approach in D'Agostino, McQuinn, and Whelan (2012) is attractive, Hounyo and Lahiri (2023) argue that the independent nature of the resampling method used to generate the simulated distribution is problematic for two reasons. First, common aggregate shocks can generate cross-sectional dependence across participants. Second, overlapping forecast horizons can generate time-series dependence across participants. As Hounyo and Lahiri (2023) note, D'Agostino, McQuinn, and Whelan (2012) independently resample observations from one forecaster to another within the same survey which ignores possible cross-sectional dependence in participants' forecast errors. They also note that D'Agostino, McQuinn, and Whelan (2012) independently resample

observations from one period to another which rules out possible serial correlation in a participant's forecast errors even though two of the four target variables involve overlapping forecast horizons.

To address these two concerns, Hounyo and Lahiri (2023) propose an alternative to the bootstrap procedure of D'Agostino, McQuinn, and Whelan (2012). Their method applies a wild bootstrap to the vector containing all the individual forecast errors at each point in time. Their approach accounts for any cross-sectional and serial correlation in participants' forecast errors while preserving the unbalanced nature of the panel.[5] Importantly, the application of their testing procedure to forecasts of GDP growth and inflation from the US-SPF strongly rejects the null hypothesis of equal predictive ability, overturning the findings of D'Agostino, McQuinn, and Whelan (2012). In particular, the results indicate there are systematic differences in forecasters' ability that extend beyond the best forecasters and include forecasters across all percentiles of the distribution of predictive performance.

While the issue of equal predictive ability in Hounyo and Lahiri (2023) has relevance for our investigation, there is scope for further exploration into the properties of individual forecast behavior. For example, the implications of IR models for forecaster interchangeability apply equally to the mean and variation over time in accuracy and, therefore, would argue for also taking higher moments into consideration. In addition, the presence of systematic patterns in predictive performance raises questions about the source(s) for this feature of the data as well as the possible impact of these patterns on the rank ordering of forecasters. Hounyo and Lahiri (2023) rank forecasters based on average predictive performance, but it is not clear whether this ranking is stable over time.

On a more general level, another aspect of Hounyo and Lahiri (2023) to consider centers on their use of the metric in (1). Specifically, the normalized metric in (1) involves an asymmetric treatment of accuracy at the individual level versus the aggregate level. That is, a forecaster who makes a relatively large error when the average forecast error is small will incur a large penalty, whereas a forecaster who makes a relatively small error when the average forecast error is large will not benefit much.[6]

[5] See Hounyo and Lahiri (2023) for a more detailed discussion.

[6] For example, if an individual's forecast error is 0.5 percentage point and the average forecast error is 2 percentage points, then the individual's normalized forecast error decreases to 0.25 percentage point and there is a 'benefit' of 0.25 percentage point. On the other hand, if an individual's forecast error is 2 percentage points and the average forecast error is 0.5 percentage point, then the individual's normalized forecast error increases to 4 percentage points and there is a 'penalty' of 2 percentage points.

Drawing upon the previous discussion, we propose a heterogeneous panel data model to describe the predictive performance of survey participants. The empirical analysis is based on the following specification for the forecast performance of each participant *j* and survey in period *t:*

$$FP^{j}_{t+h|t} = \alpha_j + \lambda_j\left(\overline{FP_{t+h|t}}\right) + \varepsilon^{j}_{t+h|t} \tag{3}$$

where $FP^{j}_{t+h|t}$ and $\overline{FP_{t+h|t}}$ denote a forecast performance (FP) metric at the individual and cross-sectional average level, respectively, and $\varepsilon^{j}_{t+h|t}$ is a mean-zero error term. For the moment, we only note that lower (higher) values of the FP and $\overline{FP}$ measures denote higher (lower) forecast accuracy and defer a more detailed discussion of the specific metrics until Section IV.

There is a close parallel between the specifications in (3) and (1) such that the panel data model can be viewed as a linear regression-based analogue to the normalized metric used in D'Agostino, McQuinn, and Whelan (2012) and Hounyo and Lahiri (2023). There are, however, several advantages to our empirical framework. First, the participant-specific intercept and slope allow for a deeper exploration into the nature of heterogeneity. Specifically, we can evaluate individual forecast performance through two channels: an individual fixed effect α -- which captures the component of a forecaster's performance that is time-invariant -- and a time-varying component $\lambda(\overline{FP})$ -- which captures the component that depends on the degree of difficulty in the forecast environment. In addition, the linear specification in (3) does not maintain the asymmetric treatment of forecast accuracy in (1) at the individual level versus the aggregate level.

Our empirical framework also accounts for cross-sectional dependence in the data. Specifically, the inclusion of the cross-sectional average of predictive performance $\left(\overline{FP}\right)$ in the individual regressions in (3) allows us to interpret our empirical framework within the context of the common-correlated effects (CCE) estimator of Pesaran (2006). As shown by Pesaran (2006), averaging the dependent variables in a panel data model at a point in time yields a proxy for an unobserved common component that can control for cross-sectional correlation across units.[7] In the context of the ECB-SPF, the movements in $\overline{FP}$ capture the effect of aggregate shocks that generate higher or lower accuracy across participants in a period and thereby also provide a very natural way to describe time variation in the difficulty of the forecasting environment.

[7] See Pesaran (2006) for a more detailed discussion.

We calculate robust standard errors for the estimated parameters in (3) by applying the Newey-West (1987) covariance matrix modified for use in a panel setting to account for autocorrelation and conditional heteroscedasticity in the data. As previously discussed, the issue of time series dependence arises when the data involve overlapping forecast horizons which is relevant for our analysis.

The heterogeneity admitted by the specification in (3) also allows for a more detailed comparison of the statistical features of participants' accuracy and an evaluation of their alignment. Specifically, we can consider both first and second moments of accuracy as a basis to investigate the issue of comparable forecast behavior. As shown in the Appendix, the restriction $\alpha_j = 0$ and $\lambda_j = 1$ provides a test for distributional homogeneity which involves a joint test of equal predictive performance and equal variance of the predictive performance metric across participants.[8] The testing procedure formalizes the idea that if forecasters are interchangeable, then over time their observed behavior should be indistinguishable from that of the consensus forecast.

In addition to the test for distributional homogeneity, another attractive feature of our empirical framework is that we can examine the estimated parameter pairings $(\hat{\alpha}_j, \hat{\lambda}_j)$ for evidence of other distinguishing patterns in participants' predictive performance within a target variable. A property of the regression equation (3) is that the estimated values for α and λ across participants will be centered around 0 and 1, respectively. As shown in Figure 1, we can partition the parameter space into four quadrants which affords an extremely intuitive way to visualize features of participants' predictive performance. If the estimated parameter pairings are not distributed randomly across the quadrants in Figure 1, then one possibility is that a scatterplot of the estimated parameter pairings principally run from the lower-left $(\alpha < 0, \lambda < 1)$ quadrant up through the upper-right $(\alpha > 0, \lambda > 1)$ quadrant. In this configuration, the lower-left quadrant would identify participants who are more accurate on average than their peers irrespective of the forecasting environment, while the upper-right quadrant would identify participants who are less accurate on average than their peers irrespective of the forecasting environment.[9]

A second possibility is that the estimated parameter pairings principally run from the upper-left $(\alpha < 0, \lambda > 1)$ quadrant down through the lower-right $(\alpha > 0, \lambda < 1)$ quadrant, implying that relative predictive performance varies with the forecasting environment. Specifically, participants in

[8] We would like to thank an anonymous referee for bringing this point to our attention.

[9] Recall that lower (higher) $\boldsymbol{FP}$ values are associated with higher (lower) individual forecast accuracy and lower (higher) $\overline{\boldsymbol{FP}}$ values are associated with tranquil (volatile) forecasting environments.

the upper-left quadrant are relatively more accurate in a tranquil environment and then relatively less accurate as the environment becomes more volatile. The opposite holds for participants in the lower-right quadrant. Additionally, the quadrant scatterplot is informative about the dispersion of the estimated parameter pairings and their correspondence with the nature of the forecasting environment.

Estimation of (3) also allows us to generate a forecast performance profile for each participant based on the predicted values of the individual regressions $\left(\boldsymbol{FP}\right)$ where:

$$FP^{j}_{t+h|t} = \hat{\alpha}_j + \hat{\lambda}_j \left(\overline{FP_{t+h|t}}\right) \tag{4}$$

The behavior of the performance profiles depends on the forecasting environment and the quadrant location of the estimated parameter pairings. As we vary the value of $\overline{FP}$, the performance profile of participants located in the lower-left quadrant will run below the profile of those located in the upper-right quadrant without crossing. In the case of participants in the upper-left quadrant and lower-right quadrant, however, changes in the forecasting environment can cause their profiles to cross and generate variation in the rank orderings of predicted forecast accuracy. Our ability to observe movements in participants' rank orderings and gauge their stability over time provides a deeper insight into forecaster performance compared to other studies and highlights another important contribution of the analysis.

Our empirical framework also lends itself to making intrapersonal comparisons of the participants. Earlier discussion of IR models highlighted the implication that individuals should not display systematic patterns in their forecast behavior across target variables. To investigate this issue, we consider one approach that focuses on the quadrant location of a participant's estimated parameter pairing. Specifically, Section V describes a simulation exercise to assess if the quadrant locations for a participant's estimated parameter pairings are similar across target variables.

We also consider a second approach that examines the relationship between a participant's overall forecast performance relative to the consensus across target variables. Specifically, we construct the following metric for individual *j* for each target variable:

$$\left(\overline{FP^{j} - \overline{FP}}\right) = \left(1/T^{j}\right) \sum_{t=1}^{T} \left(FP^{j}_{t+h|t} - \overline{FP_{t+h|t}}\right) \tag{5}$$

where $FP^{j}_{t+h|t}$ is set to $\overline{FP_{t+h|t}}$ if participant *j* did not respond to that survey. The metric in (5) is similar to the score described in (2) and provides an assessment of a participant's average relative forecast performance. That is, it indicates how a participant's predictive performance compares to the cross-sectional average over time, with negative (positive) values associated with higher (lower) overall accuracy. The findings that a participant's estimated parameter pairings tend to locate in the same quadrant and that the metric from (5) is correlated across target variables would indicate commonalities in a participant's forecast behavior and offer evidence of deviations from IR models.

One last consideration is that the specification in (3) nests two alternative approaches previously used to capture the effects of aggregate shocks. Specifically, the normalization procedure proposed by D'Agostino, McQuinn, Whelan (2012) and maintained by Meyler (2020) and Hounyo and Lahiri (2023) corresponds to the restriction that the $\alpha_j's$ are jointly equal to zero, while the application of time fixed effects adopted by Kenny, Kostka, and Masera (2014) corresponds to the restriction that all of the $\lambda_j's$ are equal. To preview our findings, the data reject both alternative approaches designed to control for variation in the forecasting environment.

Taken together, our modeling strategy provides a unified empirical framework to analyze the predictive performance of ECB-SPF participants and to inform models of the expectations formation process. Regarding IR models, our approach affords several avenues to analyze and characterize patterns in predictive performance and to determine if those patterns are consistent with interchangeable behavior on the part of participants. Moreover, our approach uses conventional estimation and testing procedures, as well as accounts for a range of econometric issues arising from the nature of the survey instrument and data. Importantly, changes in the predictability of a target variable do not present a challenge or require the adoption of some type of sub-sample analysis. Rather, time variation in the forecasting environment is an integral element in our methodology and plays a central role in our ability to compare and to contrast various features of participants' predictive performance.

III. Literature Review

Our findings make several contributions to the existing literature on the expectations formation process. One area of interest focuses on the use of a panel data framework to explore different aspects of the predictive performance of ECB-SPF participants. Kenny, Kostka, and Masera (2014) compare the predictive performance of individual-level ECB-SPF density forecasts to density forecasts from a set of simple alternative benchmark models. Their results indicate

significant time variation in the forecast accuracy of participants relative to the benchmark models. Examining the link between the moments of density forecasts and density forecast performance, Kenny, Kostka, and Masera (2015a) find that forecast performance could be improved if participants corrected a downward bias in their reported variances. Kenny, Kostka, and Masera (2015b) report that predictive performance differs across forecasting tasks, with surveyed densities being much more informative about direction-of-change predictions than high and low outcome events. We also find significant time variation in the relative performance of both point and density forecasts. Moreover, we extend previous work on the determinants of differential predictive performance of ECB-SPF participants by identifying changes in the forecast environment as a new and important channel of influence that induces time-variation in the rank orderings of the panel.

Meyler (2020) applies the bootstrapping and Monte Carlo simulation techniques of D'Agostino, McQuinn, and Whelan (2012) to examine the issue of equal predictive performance for ECB-SPF point forecasts. He correctly notes that the testing procedure relies on participants' forecast errors being uncorrelated across periods. When the data in (1) involve overlapping forecast horizons ($h>1$), the conventional application of the testing procedure is not valid because of autocorrelation in the forecast errors. To remedy this situation, Meyler (2020) proposes separating the data across nonoverlapping forecast horizons. A drawback of this approach is that it restricts his analysis to SPF rounds that are four quarters apart which dramatically lowers the power of the testing procedure because of the reduced time series dimension of the data. An attractive feature of our approach is that it does not require the panel to be separated into sub-samples, thereby allowing us to exploit efficiency gains from using all the information on participants in a collective manner.

Another area of interest is heterogeneity in the forecast features of the ECB-SPF. Kenny, Kostka, and Masera (2014, 2015a, 2015b) find considerable heterogeneity in the performance of the surveyed densities, while Meyler (2020) finds little evidence of participants who perform significantly better or worse than their peers in terms of point forecasts. Our empirical framework provides an extremely flexible approach to investigate heterogeneity in predictive performance across multiple dimensions such as target variables and the quadrant location of a participant's estimated $(\hat{\alpha}_j, \hat{\lambda}_j)$ parameter pairings. Moreover, we apply our empirical framework to both point and density forecasts as a robustness check and find that predictive performance displays stronger correlation patterns for the surveyed density forecasts than the point forecasts. Abstracting from other considerations, this result may help to explain some of the conflicting evidence reported in Kenny, Kostka, and Masera (2014, 2015a, 2015b) and Meyler (2020).

Our paper also contributes to a related literature that focuses more broadly on systematic patterns in survey-based forecasts. Bruine de Bruin et al. (2011) report strong evidence of persistence in individual participants' relative levels of uncertainty from the Federal Reserve Bank of New York Survey of Consumer Expectations. Boero, Smith, and Wallis (2015) examine the Bank of England Survey of External Forecasters and find significant persistence in the relative levels of point forecasts and uncertainty. Clements (2022) documents persistence in the relative levels of accuracy and disagreement of point forecasts for the US-SPF, while Rich and Tracy (2021) document persistence in the relative levels of disagreement and uncertainty for the ECB-SPF. While our study shares a similar motivation to Clements (2022), there are notable differences. For example, Clements (2022) restricts his analysis to point forecasts from the US-SPF and relies on a rank correlation test applied to two sub-samples to assess systematic patterns in individual forecast behavior.

Finally, our modeling strategy is closely related to the work of Qu, Timmermann, and Zhu (2019, 2021) that uses a panel data framework to analyze forecast accuracy. They consider various approaches to separate the importance of common shocks from idiosyncratic, individual-specific shocks. Our analysis differs in two important respects. The first is in terms of dimensions of the data. The modeling framework and testing procedures in Qu, Timmermann, and Zhu (2021) are designed for a large cross-section. However, our examination of the ECB-SPF only includes three variables – real GDP growth, inflation, and unemployment – which is too small for applying their methods. The second is in terms of focus. A key issue of interest in Qu, Timmermann, and Zhu (2019) is the identification of participants with superior forecasting skills. Consequently, their methodology involves the consideration of predictive performance across multiple dimensions and the assessment of a very large number of pairwise comparisons. In contrast, our interest is not in a detailed exploration aimed at an overall ranking of forecasters. Rather, the inter- and intrapersonal comparisons in our analysis are much more limited in scope and are more narrowly directed at assessing their consistency with expectations models that predict the absence of systematic patterns at the individual level.

IV. The European Central Bank Survey of Professional Forecasters

The ECB-SPF began in January 1999 and provides a quarterly survey of euro area forecasts. The survey draws its pool of panelists from both financial and nonfinancial institutions, with most, but not all, located in the euro area. Meyler (2020) notes that the principal aim of the survey is to solicit expectations about real GDP growth, inflation, and unemployment, although the questionnaire also contains a noncompulsory section asking participants for their expectations of

other variables and to provide qualitative comments that inform their quantitative forecasts.[10] The ECB-SPF asks panelists for forecasts at short-, medium- and longer-term horizons, including both "rolling" and "calendar year" variants. The survey is fielded in January, April, July, and October, with approximately 55 panelists on average responding per survey. For additional details about the ECB-SPF, see Garcia (2003) and Bowles et al. (2007).

We examine forecasts for real GDP growth, HICP inflation, and the unemployment rate. This choice partly reflects the structure of the survey instrument that asks respondents to submit both point- and density-based forecasts for these three macroeconomic variables.[11] Because Kenny, Kostka, and Masera (2014, 2015a, 2015b) restrict their analyses to surveyed density forecasts and Meyler (2020) restricts his analysis to surveyed point forecasts, our inclusion of both types of forecasts offers an important robustness check. For the density forecasts, participants report their subjective probability distribution of forecasted outcomes as a histogram using a set of intervals provided in the survey. While the ECB-SPF occasionally changes the number of closed intervals for the histogram, it has essentially maintained a common bin width for the closed intervals throughout its history.[12]

Regarding forecast horizons, we examine point and density forecasts that involve rolling one-year-ahead and one-year/one-year-forward horizons. Compared to calendar year horizons, an advantage of the rolling horizons is that the horizon length remains constant through time and allows us to treat the data as quarterly observations on a set of individually homogeneous series. As Garcia (2003) notes, there is a temporal misalignment between the target variables because of differences in the data frequency and publication lags of the variables. Specifically, real GDP growth is published quarterly with a two-quarter lag, while HICP inflation and the unemployment rate are published monthly with a one-month and a two-month lag, respectively.[13]

Our study analyzes surveys conducted from 1999:Q1–2018:Q3, with forecast evaluation for all series ending in 2019:Q3. The ECB-SPF, like other surveys, has experienced entry and exit of

[10] The additional expectations are for variables such as wage growth, the price of oil, and the exchange rate.
[11] The ECB-SPF is among a small but growing number of surveys that solicit both point and density forecasts. Other notable surveys include the US-SPF (published by the Federal Reserve Bank of Philadelphia), the Bank of England Survey of External Forecasters, and Federal Reserve Bank of New York Survey of Consumer Expectations.
[12] The only deviation in this design started with the 2020:Q2 survey in response to the COVID-19 outbreak. The (nearly) constant interval width of the ECB-SPF density forecasts contrasts with the US-SPF density forecasts, which have experienced periodic changes in interval widths.
[13] For example, the 2010:Q1 survey questionnaire asks respondents to forecast one-year-ahead output growth from 2009:Q3–2010:Q3. For HICP inflation, the corresponding forecast horizon is December 2009–December 2010. For the unemployment rate, the corresponding forecast is for November 2010.

respondents over time. In addition, occasionally participants do not respond to a questionnaire or to individual items within the questionnaire. As noted by Meyler (2020), participants provide the highest number of forecasts for HICP inflation and the lowest number for unemployment, with the number of forecasts at the one-year-ahead horizon exceeding that at the one-year/one-year-forward horizon. Participants also report more point forecasts than density forecasts. Given the unbalanced panel structure of the ECB-SPF, we only include participants at each individual target variable/horizon who provide at least 50 forecasts.[14] Further, we only consider matched point and density forecasts to maintain comparability across the types of forecasts. Consequently, the number of participants varies from 34 (HICP inflation at the one-year-ahead horizon) to 21 (unemployment rate at the one-year/one-year-forward horizon).

An important issue for the assessment of predictive performance is the choice of data vintage used to construct realizations of the target variables. As is the case for most macroeconomic data for most countries, euro-area macroeconomic statistics tend to be revised from preliminary releases. Consequently, a choice must be made about the relevant release associated with a participant's forecast. Following Meyler (2020), we construct realizations of the target variables for HICP inflation and the unemployment rate using monthly data from the first full release.[15] For real GDP growth, we construct realizations of the target variables using quarterly data from the second estimate. We have considered other approaches to construct realizations of the target variables as additional robustness checks.[16]

Another important issue for the assessment of predictive performance is the choice of point and density forecast accuracy measures. For the point forecasts, we adopt the absolute error as the metric:

$$^{POINT}FP^{j}_{t+h|t} = \left| X_{t+h} - E^{j}_{t}\left[X_{t+h}\right] \right| \tag{6}$$

where X_{t+h} denotes the realized value of the relevant ECB-SPF target variable in period $t+h$ and $E^{j}_{t}[X_{t+h}]$ denotes the reported point forecast from participant j in the survey at date t.

[14] We have also experimented with a lower threshold of 40 participants and obtained similar results.
[15] For example, if the target variable is one-year-ahead HICP inflation, we use the first full release reporting the value of the price index in month $t+12$. The same release is used to obtain the value of the price index in month t.
[16] We used current vintage data as one robustness check. As another robustness check, we construct growth rates using the first full release to obtain the value of the price index in month t and the second estimate for the level of real GDP in quarter t. The results changed very little using these alternative approaches.

For the density-based accuracy measure, we adopt the absolute rank probability score (ARPS) as the metric:

$$(7) \qquad {}^{DENSITY}FP^{j}_{t+h|t} = \frac{1}{k_t - 1}\sum_{i=1}^{k_t}\left|\sum_{g=1}^{i} {}_g p_t^j - \sum_{g=1}^{i} {}_g I_{t+h}\right|$$

where we assume there are k_t bins associated with the histogram for the survey at date t, ${}_g p_t^j$ is the probability assigned by respondent j to the g^{th} bin, and ${}_g I_{t+h}$ denotes an indicator variable that takes a value of one if the actual outcome in period $t+h$ is in the g^{th} interval of the histogram from the survey at date t. The ARPS has the property that a participant receives "credit" by assigning probability in bins close to the bin containing the actual outcome.[17]

The evaluation of the ECB-SPF density forecasts requires additional discussion beyond the selections of data vintage and metrics. To the extent that respondents place any probability in either open interval, the manner chosen to close off the open intervals will affect the value of the forecast performance metric in (7). We follow a common—although ad hoc—assumption and close the exterior open intervals by assigning them twice the width of the interior closed intervals. We also need to address the issue of the location of probability mass associated with the density forecasts. We again draw upon common practices and assume that the probability mass is distributed uniformly within each bin of the histogram. Finally, we exclude the 2009:Q1 one-year-ahead real GDP growth density forecast data because many respondents placed significant probability in the lower open interval of the histogram in this survey.[18]

V. Empirical Results

We begin by examining the behavior of the (cross-sectional) average forecast performance metrics $\left(\overline{FP}\right)$ to compare predictability across the target variables as well as to identify tranquil and volatile episodes. Figure 2 plots the movements of $\left(\overline{FP}\right)$ for the point forecasts and density

[17] The squared norm is used in Meyler (2020) for point forecasts and in Kenny, Kostka, and Masera (2014, 2015a) for density forecasts. Compared to the absolute value norm in (6) and (7), the squared norm is more sensitive to outliers and the manner used to close the exterior open intervals for density forecasts. For robustness, we also used the squared norm and found similar results.

[18] For this survey, the significant probability mass at the lower open interval corresponded to a growth rate of "-1 percent or less" and was due to the survey design of the density forecasts and its inability to provide sufficient coverage for the pessimistic point predictions of output growth. For individuals who either reported point predictions below -1 percent or wanted to indicate significant downside risk, they assigned most of their probability to the open-ended interval. See Abel et al. (2016) for further discussion.

forecasts of real GDP growth, HICP inflation, and the unemployment rate at the one-year-ahead horizon, while Figure 3 provides the corresponding information at the one-year/one-year-forward horizon. The metrics are plotted based on the realization of the target variable, with gray bars indicating recessions as determined by the Euro Area Business Cycle Dating Committee of the Center for Economic Policy Research.[19]

As shown, there is generally a close correspondence between the point and density forecast performance metrics for the same target variable. While the difficulty of forecasting outcomes around the time of the global financial crisis and the euro-area debt crisis is evident, the data indicate other episodes associated with sizable forecast errors that are not uniform in their timing across the target variables. Consequently, there is sufficient variability in the forecasting environments to mitigate concerns that our results may be largely driven by just a few events.

As for the pattern of the forecast errors, they are highest for real GDP growth around the time of the global financial crisis. For HICP inflation, they are also highest around the time of the global financial crisis as well as elevated toward the beginning of the sample and during the middle of the last decade. For the unemployment rate, the forecast errors are again largest around the time of the global financial crisis, although they are also elevated at the beginning of the sample and around the time of the euro-area debt crisis.

Interpersonal Comparisons of Predictive Performance

We estimate the parameters in (3) using ordinary least squares (OLS), with standard errors computed using the Newey-West (1987) covariance matrix estimator modified for use in a panel data set.[20] Column 1 in Tables 1-2 presents formal tests for distributional homogeneity of the predictive performance metrics. Letting $\hat{\theta} = \left[(\hat{\alpha}_1, \hat{\lambda}_1), (\hat{\alpha}_2, \hat{\lambda}_2), ..., (\hat{\alpha}_N, \hat{\lambda}_N) \right]$ denote the vector of estimated parameters of the model, we construct the following Wald test statistic for the joint null hypothesis that $\alpha_j = 0 \cap \lambda_j = 1$ for $j = 1, \ldots, N$ participants in the panel for a specific target variable:

$$W = \left(\hat{\theta} - \theta_0\right)' \left[\text{var}\left(\hat{\theta}\right) \right]^{-1} \left(\hat{\theta} - \theta_0\right) \tag{8}$$

[19] For example, the metric associated with the forecasts of HICP inflation from 2015:Q1-2016:Q1 is plotted at 2016:Q1. While Figure 2 plots the value for the one-year-ahead point forecasts of real GDP growth in 2009:Q1, recall that the analysis does not include these data due to the exclusion of the matched density forecasts. Unlike the absolute error metric, the ARPS metric is restricted to fall in the range between 0 and 1.

[20] We allow the error terms to follow a fourth-order moving average process to account for the overlap of forecast horizons.

The values of the test statistic indicate strong evidence of systematic differences in the mean and variance of participants' forecast accuracy as we reject the null hypothesis at the 1 percent significance level in all cases except for the point forecasts of inflation at the one-year horizon.[21]

The test of distributional homogeneity involves a joint test of equal predictive performance and equal variance of the predictive performance metric. Because almost all investigations into forecast performance have focused exclusively on equal predictive ability, it would be of interest to explore how this issue may bear upon the reported rejections. Even if participants display equal predictive ability, our more restrictive testing procedure could lead to a rejection of the null hypothesis due to heterogeneity in the variance of the performance metric. To investigate this possibility, we apply the testing procedure of Hounyo and Lahiri (2023) to the ECB-SPF data and report the findings in the Appendix.

As shown, the evidence strongly rejects the null hypothesis of equal predictive ability at conventional significance levels.[22] Consequently, the rejection of distributional homogeneity does not mask equal predictive ability among forecasters. It is also interesting to note that the pattern of rejections of equal predictive ability is remarkably similar to that documented by Hounyo and Lahiri (2023) for the US-SPF where the best forecasters as well as forecasters across the other percentiles of the distribution of predictive performance are more accurate than what would be expected by random chance using the bootstrap procedure.

Because our empirical framework nests two common approaches to control for variability in the forecasting environment, we also construct Wald test statistics for the validity of the normalized predictive performance metric $(\alpha_1 = \alpha_2 = \ldots = \alpha_N = 0)$ and for the use of time fixed effects $(\lambda_1 = \lambda_2 = \ldots = \lambda_N)$. Except for the point forecasts of inflation at the one-year horizon, the results in column 2 and column 3 strongly reject the normalized predictive performance metric and the use of time fixed effects to control for the effects of aggregate shocks, respectively. These findings offer an additional reason why our evidence of systematic differences in the predictive performance of ECB participants contrasts with the analyses of Kenny, Kostka, and Masera (2014) and Meyler (2020). Specifically, Kenny, Kostka and Masera (2014) use time fixed effects to control for changes in the forecasting environment., while Meyler (2020) adopts the normalization procedure.[23]

---

[21] The different degrees of freedom reflect the varying number of respondents meeting the participation restriction for the various target variables.

[22] Following Hounyo and Lahiri (2023), we also consider the case of excluding forecasters who scored worse than the 80th percentile and found similar results. These findings are also reported in the Appendix.

[23] As previously discussed, the more general nature and greater parameter flexibility of our empirical framework allows the model estimates to account for changing relative performance rankings. The inability of

To gain insight into the relative accuracy of participants, Figures 4-5 display scatterplots and correlation coefficients (r) for the individual estimated parameter pairings $\left(\hat{\alpha}_j, \hat{\lambda}_j\right)$ for the point forecasts and density forecasts, respectively. The patterns are striking in their similarity across target variables. Because few of the estimated parameter pairings fall in the lower-left $(\alpha < 0, \lambda < 1)$ and upper-right $(\alpha > 0, \lambda > 1)$ quadrants, the visual evidence does not support the view that the ECB-SPF panel is comprised of participants who remain relatively more accurate and other participants who remain relatively less accurate across all forecasting environments. Instead, evidence of the estimated parameter pairings largely falling in the upper-left $(\alpha < 0, \lambda > 1)$ and lower-right $(\alpha > 0, \lambda < 1)$ quadrants indicates that participants' relative accuracy varies with the forecasting environment. Moreover, the patterns do not suggest clustering or that the negative relationship reflects the behavior of a few participants. Rather, the observations are dispersed within each of the two quadrants and display a comparable count across the two quadrants. We also observe the correlations are larger in absolute value for the density forecasts compared to the point forecasts.

Given the evidence documenting a strong link between a participant's predictive performance and the difficulty of the forecasting environment, it is natural to ask what might be driving this result. Here we consider one possible explanation that draws upon Clements (2022) and can be easily incorporated within our empirical framework. Specifically, Clements (2022) examines the US-SPF and investigates whether systematic differences in forecast accuracy are related to systematic differences between forecasters in their degree of contrarianism. Such would be the case if some forecasters receive superior private information, resulting in predictions that display greater contrarianism but also higher accuracy.

We investigate the relationship between forecast accuracy and contrarianism by pairing a participant's average relative forecast performance metric in (5) with an analogue for disagreement. Specifically, we construct the following measure of average relative disagreement:

$$\left(\overline{D^j - \overline{D}}\right) = \left(1/T^j\right) \sum_{t=1}^{T} \left(D^j_{t+h|t} - \overline{D_{t+h|t}}\right) \tag{9}$$

where

$$D^j_{t+h|t} = \left| E^j_t[X_{t+h}] - \overline{E_t[X_{t+h}]} \right| \tag{10}$$

the specifications in Kenny, Kostka and Masera (2014) and Meyler (2020) to capture this feature of the data may be another factor explaining why our conclusions differ.

and where $D^{j}_{t+h|t}$ is set to $\overline{D_{t+h|t}}$ if participant *j* did not respond to that survey. The metric in (9) indicates how a participant's disagreement compares to average disagreement across the responding survey group over time, where individual disagreement is described in (10) and is measured by the absolute value of the deviation between a participant's forecast and the consensus forecast.[24] Because a positive (negative) value in (9) reflects higher (lower) relative disagreement, we would expect a negative association with average relative forecast performance if superior information is the source for the persistent performance heterogeneity among forecasters.

Figure 6 displays scatterplots and correlation coefficients (r) for the individual pairings of disagreement and predictive performance for the point forecasts. As shown, there is a positive relationship of varying strength between disagreement and accuracy across target variables which suggests that more contrarian forecasters on average make less accurate forecasts. This finding is consistent with evidence in Clements (2022) for the US-SPF and does not support the conjecture that heterogeneity in forecast accuracy reflects some participants benefiting from superior private information.

Time-variation in Forecast Performance Profiles

An attractive feature of our empirical framework is that we can examine the implications of the estimation results for the forecast performance profiles of participants. To illustrate, we will initially select a participant from each of the four quadrants associated with a target variable. While any scatterplot can be used for the exercise, we select the one-year-ahead point forecasts of GDP growth because it is likely to be of particular interest.[25] Figure 7 depicts the data and the estimated regression line for each participant identified by the color-coded circles for the one-year-ahead GDP growth rate in the upper-left panel in Figure 4. As shown, the estimated regression lines display a very high fit to the data and suggest little reason to depart from the linear specification in (3).[26]

Figure 8 plots the predicted forecast performance profiles of the same four participants and provides a visual investigation into their behavior as well as the incidence and nature of crossings that bear upon the issue of the stability of rank orderings. Using the estimated parameter pairing for

---

[24] Similar to the inclusion of $\overline{FP}$ in (5), the inclusion of average disagreement $(\overline{D_{t+h|t}})$ in (9) is consistent with Clements (2022) who cites the importance of controlling for variation in the extent of disagreement over time.

[25] While the one-year-ahead point forecasts of inflation would also be of interest, recall that we do not reject the property of distributional homogeneity for this series.

[26] There is an outlier observation for three of the four participants associated either with realized GDP growth in 2008:Q3 or 2008:Q4. We exclude the relevant observation from the scatterplots (but not the reported R^2 values) in Figure 7 to enhance presentation of the data. The scatterplots and regression lines including all observations are provided in the Appendix.

each participant, we vary the average forecast performance metric $\left(\overline{FP}\right)$ to trace out the performance profiles. Figure 8 also includes a 45-degree line indicating where individual predicted forecast performance $\left(FP\right)$ equals the cross-sectional average of forecast performance $\left(\overline{FP}\right)$.

The resulting performance profiles closely align with our expected behaviors.[27] If participants were principally located in the lower-left (purple) and the upper-right (blue) quadrants, then this configuration would produce relatively stable rankings over time. This is shown by the purple and blue lines not crossing, with increases in the difficulty of the forecasting environment only acting to widen the gap between them. Because participants in the lower-left quadrant are predicted to be systematically more accurate than the average, the purple line always lies below the 45-degree line. The opposite holds for the participant in the upper-right quadrant. It is important to note that our illustration does not claim that the performance profiles cannot display crossings, but it does indicate that the crossings can only occur among participants located in the same quadrant.

As shown in Figures 4 and 5, most participants are located in the upper-left (yellow) and lower-right (red) quadrants. In contrast to the previous configuration, this configuration will produce highly variable rankings over time as performance profiles will display crossings beyond those involving participants located in the same quadrant. This is illustrated on a general level by the yellow and red lines crossing the 45-degree line which indicates a switch in the forecast accuracy of the participants relative to the cross-sectional average. Focusing on our selected participants, we see the yellow and red lines cross at 0.83 (the 59th percentile of $\overline{FP}$) as well as crossings with the participant from the lower-left quadrant at 0.41 (29th percentile) and 1.45 (90th percentile), respectively.[28] As shown, these crossings are associated with changes in rankings of participants as the forecasting environment evolves from low difficulty to extreme difficulty.[29]

---

[27] Similar to Figure 7, there is a corresponding outlier value of $\overline{FP}$ that we elect to exclude from the plots in Figure 8 to enhance presentation of the performance profiles. The performance profiles values using the full range of $\overline{FP}$ values are also provided in the Appendix.

[28] While participants in the lower-left quadrant display forecasts that are systematically more accurate than the average, this does not imply that their forecasts always outperform those of individuals in other quadrants. Consequently, there is no inconsistency with the figure displaying the crossings of the performance profiles by the participants in the upper-left and lower-right quadrants. A similar point holds for participants in the upper-right quadrant.

[29] As shown in the corresponding figure in the Appendix, there is an additional crossing of the yellow and blue lines near the upper range of the $\overline{FP}$ values. While a crossing point can always be calculated between the 45-degree line and the performance profile of a participant in the upper-left or lower-right quadrant, this may occur outside the range of $\overline{FP}$ values in the sample.

To gain a better appreciation of the extent of this variability, we now consider all 31 participants associated with the one-year-ahead point forecasts of GDP growth. We construct rank orderings based on participants' forecast accuracy evaluated at eight values spanning the range of $\overline{FP}$ values for this target variable. Table 3 reports the results, where the first column lists the forecaster IDs and the remaining columns moving from left to right indicate the rank ordering of each forecaster as the forecasting environment becomes more difficult. For ease of comparison, the order of the ID numbers is based on the initial ranking of forecasters at the lowest value of $\overline{FP} = 0.25$.

As shown in Table 3, the pattern of the rank orderings is consistent with the evidence from the scatterplot of the estimated parameter pairings in the upper-left panel of Figure 4. Some respondents largely maintain similar rankings either because they tend to be highly accurate (#22), highly inaccurate (#36), or close to the cross-sectional average most of the time (#54, #16). For most respondents, however, their rank orderings vary over the forecast environment. While this variation can reflect dramatic improvements (#94, #52) or dramatic declines (#37, #39) in predictive performance, it is more typical to observe individuals who become relatively more accurate (#24, #93) or relatively less accurate (#54, #98) as the forecasting environment turns more challenging.

There are two key takeaways that emerge from Figure 8 and Table 3. The first is that forecaster evaluations and comparisons may not be invariant to the relative prevalence of tranquil and volatile episodes in a selected sample period. The second is that the evidence may provide one explanation for the finding that it is difficult, ex ante, to devise forecast combination methods that beat a simple average.[30] While our analysis links predictive performance to variation in the forecasting environment, this feature may not be easily exploitable because of the inherent difficulty of predicting tranquil/volatile episodes in real time. If, however, we were to allow for the availability of some information on an ex post basis, then the Appendix describes a "performance weighting" combination scheme that consistently and significantly outperforms the equally-weighted consensus forecast.[31]

Intrapersonal Comparisons of Predictive Performance

The analysis up to this point has examined forecast data for the target variables in isolation. However, we can also investigate if there are commonalities in individual predictive performance across target variables. While the scatterplots in Figure 4 and Figure 5 show that the estimated parameter pairings principally lie in the $(\alpha < 0, \lambda > 1)$ and $(\alpha > 0, \lambda < 1)$ quadrants, they do not

[30] See Timmermann (2006) and Genre et al. (2013).

[31] We would like to thank an anonymous referee for suggesting this exercise.

indicate the extent to which the pairings for a participant tend to locate in the same quadrant across target variables. Another consideration is the extent to which a participant's predictive performance for a target variable correlates with performance for other target variables. Previous discussion has noted the relevance of these additional considerations for IR models.

Our investigation into the location of the parameter pairings for a participant requires more than simply focusing on the quadrant associated with the estimates. We must also account for the uncertainty associated with the estimates. Consequently, we adopt Monte Carlo simulation techniques and generate 1,000 draws of the parameter pairings vector for each target variable and horizon using the estimated joint normal distribution for $\hat{\theta} = \left[\left(\hat{\alpha}_1, \hat{\lambda}_1\right), \left(\hat{\alpha}_2, \hat{\lambda}_2\right), ..., \left(\hat{\alpha}_N, \hat{\lambda}_N\right)\right]$. We can then use the simulated distributions to calculate the percentage of simulated parameter pairings located in each quadrant for a participant.

Figure 9 and Figure 10 plot the distributions for the point forecasts and density forecasts, respectively, where the estimated pairings are color-coded in black and the simulated pairings are in gray. As shown, the distributions for the density forecasts are much tighter compared to the point forecasts.[32] While we only make note of this difference at present, it would be interesting for future research to explore the reasons for this feature of the data. For example, it is possible that respondents make less use of rounding and report less judgmental density forecasts compared to point forecasts.[33] Relatedly, Glas and Hartmann (2022) analyze the US-SPF and ECB-SPF and note that survey participants who report rounded point forecasts differ from respondents who round probabilities for density forecasts.

Figure 11 and Figure 12 focus on the quadrant location of the parameter pairings associated with participants' point forecasts and density forecasts, respectively. The values report the highest fraction of simulated parameter pairings for a participant that fall in the same quadrant based on the 6,000 simulations (1,000 simulations for each of the six target variables). For purposes of comparison, we present the results for the 23 participants included in all six combinations of target variables.[34] Overall, the evidence in Figure 11 provides general support for the idea that a participant's parameter pairings tend to locate in the same quadrant most of the histogram bars exceed 40 percent. Using 50 percent as an arbitrary threshold, the histogram bars show that about a third of the participants exceed the threshold criterion.

[32] The difference in precision may explain why rejections of various hypotheses in Tables 1-2 are stronger for the density forecast data.

[33] We would like to thank an anonymous referee for bringing this point to our attention.

[34] We have also extended the analysis to include the other participants in our study and the results are similar.

A very different picture emerges when we look at the density forecasts. There are now 16 participants who exceed the threshold criterion, with the calculated percentages notably higher than the 50 percent value in many cases. Particularly noteworthy are the two participants whose forecast behavior suggests their parameter pairings would almost always fall in the same quadrant across the six combinations of target variables. Compared to the point forecasts, the density forecasts indicate considerably more overlap in quadrant locations which is consistent with the evidence from Figure 9 and Figure 10 and provides another example of the different conclusions that can be drawn between the point and density forecasts.

We can also use the average relative forecast performance metrics in (5) to make various comparisons across the forecast data, where lower values again indicate better predictive performance. Because of the large number of comparisons, we only provide a summary of the results. Overall, we find that forecast performance correlates positively across horizons and outcome variables in almost all cases. There are, however, differences across some dimensions that are worth noting. One difference is that the density forecast data generate a much stronger association than the comparable point forecast data. The top panel in Figure 13 is representative of this finding and shows scatterplots of the average relative forecast performance metrics for inflation at the two forecast horizons for the point forecast data and density forecast data, respectively. While the point forecast data indicate a modest correlation of 0.43, the density forecast data indicate a correlation of 0.77 which is nearly twice as high. Looking across all pairwise combinations of target variables, the correlations for the point forecast data are typically in the 0.2-0.4 range, while the correlations for the density forecast data are in the 0.7-0.8 range.

Another feature of predictive performance that emerges is that the correlations are generally higher for the same target variable at different horizons than for different target variables at the same horizon. An ordered ranking of the correlations indicates that the lowest three values are associated with inflation and GDP growth at the two forecast horizons and GDP growth and unemployment at the one-year/one-year-forward horizon. In contrast, the highest three values are associated with unemployment (using both types of forecast data) and inflation at the two forecast horizons.

A further examination of forecast performance across the target variables reveals two other features. First, there tends to be a stronger correlation at the shorter horizon. The middle panel of Figure 13 shows scatterplots of the average relative forecast performance metrics for GDP growth and unemployment. Unlike the pattern at the one-year-ahead horizon, there is much less of a translation of forecast performance from unemployment into GDP growth at the one-year/one-

year-forward horizon. Second, there is less of a linkage between forecast performance for GDP growth and inflation than there is for GDP growth and unemployment. The bottom panel of Figure 13 shows the scatterplots of the corresponding average relative forecast performance metrics at the one-year-ahead horizon, where we again include the GDP growth/unemployment scatterplot to facilitate the comparison. For inflation and GDP growth, predictive performance shows a slightly negative relationship.[35] In the case of GDP growth and unemployment, however, there is a sufficiently meaningful positive association.

VI. Conclusion

This paper adopts the common correlated effects (CCE) estimator of Pesaran (2006) to investigate whether ECB-SPF participants can be viewed as interchangeable. While the behavior of professional forecasters is of interest by itself, our study draws further motivation from IR models and their implication that systematic patterns should not be evident in the forecast data. In addition to making comparisons of predictive performance across participants, we investigate the correlation patterns for an individual's predictive performance across parameter configurations and target variables. As a robustness check, we also consider the evidence from point forecasts and density forecasts.

Based on forecasts for output, inflation, and unemployment, we find strong evidence of systematic patterns in participants' predictive performance. Moreover, the patterns are not a consequence of differential innate ability, but instead are episodic in nature and directly linked to changes in the forecasting environment. By way of a simple narrative, our interpersonal analysis of predictive performance suggests that participants largely divide into two "camps": those who display relatively more accurate forecasts in low-variance times and those who do so in high-variance times. Consistent with this view, we find the rank orderings of participants shift over time and display considerable variability.

Our intrapersonal analysis of predictive performance indicates that the influence of the forecasting environment carries over to other features of the forecast profile of participants. Specifically, we find there are commonalities across parameter quadrants and target variables, with the density forecast data revealing greater similarities in individual forecast behavior. In terms of the narrative introduced above, participants tend to locate in the same "camp" which indicates that a

[35] This is the only instance where the average relative forecast performance metrics display a negative relationship.

participant's relative accuracy in a forecasting environment is positively correlated across target variables.

Overall, we conclude that the predictive performance of ECB-SPF participants reflects distinguishing behaviors that are inconsistent with the implications of IR models for interchangeability. The strong evidence of systematic patterns in predictive performance and their relationship to the nature of the forecasting environment is a new finding and reflects the capabilities and advantages of the CCE empirical framework.

It would be interesting and important to determine if these same empirical features are present in other long running panel survey data.[36] Our findings support further development of expectations models that can generate systematic patterns in key features of forecasters' behavior as well as account for the differential effects of the forecast environment on predictive performance. The opportunity to explore and identify the key underpinnings during such a development would serve as fertile ground for future research.

[36] The US-SPF would seem to be a natural candidate. It is unclear, however, if a parallel analysis can be conducted for the US-SPF because of differences in the survey instrument. Specifically, the US-SPF does not feature "rolling" forecast horizons. Such investigations, however, would not need to be restricted to professional forecasters and should be considered for surveys more generally.

Table 1

Comparison of Predictive Performance Behavior of ECB-SPF Participants

Point Forecasts

$$^{POINT}FP^{j}_{t+h|t} = \alpha_j + \lambda_j\left(\overline{FP_{t+h|t}}\right) + \varepsilon^{j}_{t+h|t}$$

Point Forecast Data	Distributional Homogeneity $H_0 : \alpha_j = 0 \cap \lambda_j = 1 \ \forall \ j$	Normalization Approach $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$	Time Fixed Effects $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_N$
GDP growth: one-year-ahead	$\chi^2(62) = 375.4^{**}$	$\chi^2(31) = 84.4^{**}$	$\chi^2(30) = 278.8^{**}$
GDP growth: one-year/one-year-forward	$\chi^2(58) = 244.7^{**}$	$\chi^2(29) = 68.1^{**}$	$\chi^2(28) = 143.9^{**}$
Inflation: one-year-ahead	$\chi^2(68) = 77.7$	$\chi^2(34) = 41.1$	$\chi^2(33) = 29.3$
Inflation: one-year/one-year-forward	$\chi^2(62) = 156.2^{**}$	$\chi^2(31) = 90.9^{**}$	$\chi^2(30) = 80.4^{**}$
Unemployment: one-year-ahead	$\chi^2(56) = 134.5^{**}$	$\chi^2(28) = 58.0^{**}$	$\chi^2(27) = 58.1^{**}$
Unemployment: one-year/one-year-forward	$\chi^2(48) = 225.8^{**}$	$\chi^2(24) = 42.0^{*}$	$\chi^2(23) = 100.1^{**}$

Note: Model parameters are estimated using ordinary least squares (OLS), with standard errors computed using the Newey-West (1987) covariance matrix estimator modified for use in a panel data set. The error terms to follow a fourth-order moving average process to account for the overlap of forecast horizons. Degrees of freedom are reported in parentheses.

** Significant at the 1% level

* Significant at the 5% level

Table 2

Comparison of Predictive Performance Behavior of ECB-SPF Participants

Density Forecasts

$$^{DENSITY}FP^{j}_{t+h|t} = \alpha_j + \lambda_j\left(\overline{FP_{t+h|t}}\right) + \varepsilon^{j}_{t+h|t}$$

Density Forecast Data	Distributional Homogeneity $H_0 : \alpha_j = 0 \bigcap \lambda_j = 1 \ \forall \ j$	Normalization Approach $H_0 : \alpha_1 = \alpha_2 = \ldots = \alpha_N = 0$	Time Fixed Effects $H_0 : \lambda_1 = \lambda_2 = \ldots = \lambda_N$
GDP growth: one-year-ahead	$\chi^2(62) = 337.4^{**}$	$\chi^2(31) = 235.5^{**}$	$\chi^2(30) = 160.6^{**}$
GDP growth: one-year/one-year-forward	$\chi^2(58) = 371.2^{**}$	$\chi^2(29) = 173.1^{**}$	$\chi^2(28) = 100.6^{**}$
Inflation: one-year-ahead	$\chi^2(68) = 352.6^{**}$	$\chi^2(34) = 208.8^{**}$	$\chi^2(33) = 103.2^{**}$
Inflation: one-year/one-year-forward	$\chi^2(62) = 345.6^{**}$	$\chi^2(31) = 273.5^{**}$	$\chi^2(30) = 134.3^{**}$
Unemployment: one-year-ahead	$\chi^2(56) = 324.2^{**}$	$\chi^2(28) = 141.1^{**}$	$\chi^2(27) = 106.0^{**}$
Unemployment: one-year/one-year-forward	$\chi^2(48) = 138.6^{**}$	$\chi^2(24) = 69.8^{**}$	$\chi^2(23) = 74.6^{**}$

Note: Model parameters are estimated using ordinary least squares (OLS), with standard errors computed using the Newey-West (1987) covariance matrix estimator modified for use in a panel data set. The error terms to follow a fourth-order moving average process to account for the overlap of forecast horizons. Degrees of freedom are reported in parentheses.

** Significant at the 1% level

* Significant at the 5% level

Table 3

Rank Orderings of Forecast Accuracy: Point Forecasts of One-Year-Ahead GDP Growth

Forecaster ID	$\overline{FP} = 0.25$	$\overline{FP} = 0.50$	$\overline{FP} = 0.75$	$\overline{FP} = 1.00$	$\overline{FP} = 1.50$	$\overline{FP} = 2.00$	$\overline{FP} = 4.00$	$\overline{FP} = 6.00$
37	1	4	4	10	18	21	23	26
39	2	3	3	4	11	14	20	22
42	3	2	2	2	5	7	10	11
22	4	1	1	1	2	2	4	4
61	5	5	6	12	14	17	19	19
95	6	6	9	13	13	16	18	18
4	7	7	10	14	12	13	16	17
5	8	10	17	20	24	24	26	27
88	9	8	5	6	8	10	9	9
23	10	9	8	7	7	9	8	8
89	11	16	23	25	25	25	24	23
54	12	12	15	17	17	15	14	14
38	13	11	11	11	9	8	7	7
16	14	14	16	16	15	12	12	13
15	15	24	27	27	27	27	28	29
98	16	19	24	24	23	23	22	21
56	17	23	26	26	26	26	25	25
31	18	20	21	22	22	22	21	20
33	19	26	28	28	28	28	29	28
24	20	18	18	19	19	19	15	15
26	21	13	12	8	6	5	5	5
93	22	21	22	21	21	20	17	16
29	23	29	29	29	29	29	30	30
20	24	22	19	18	16	11	11	10
1	25	27	25	23	20	18	13	12
96	26	17	13	5	3	3	2	3
85	27	15	7	3	1	1	1	1
94	28	25	14	9	4	4	3	2
52	29	28	20	15	10	6	6	6
2	30	30	30	30	31	31	31	31
36	31	31	31	31	30	30	27	24

Figure 1

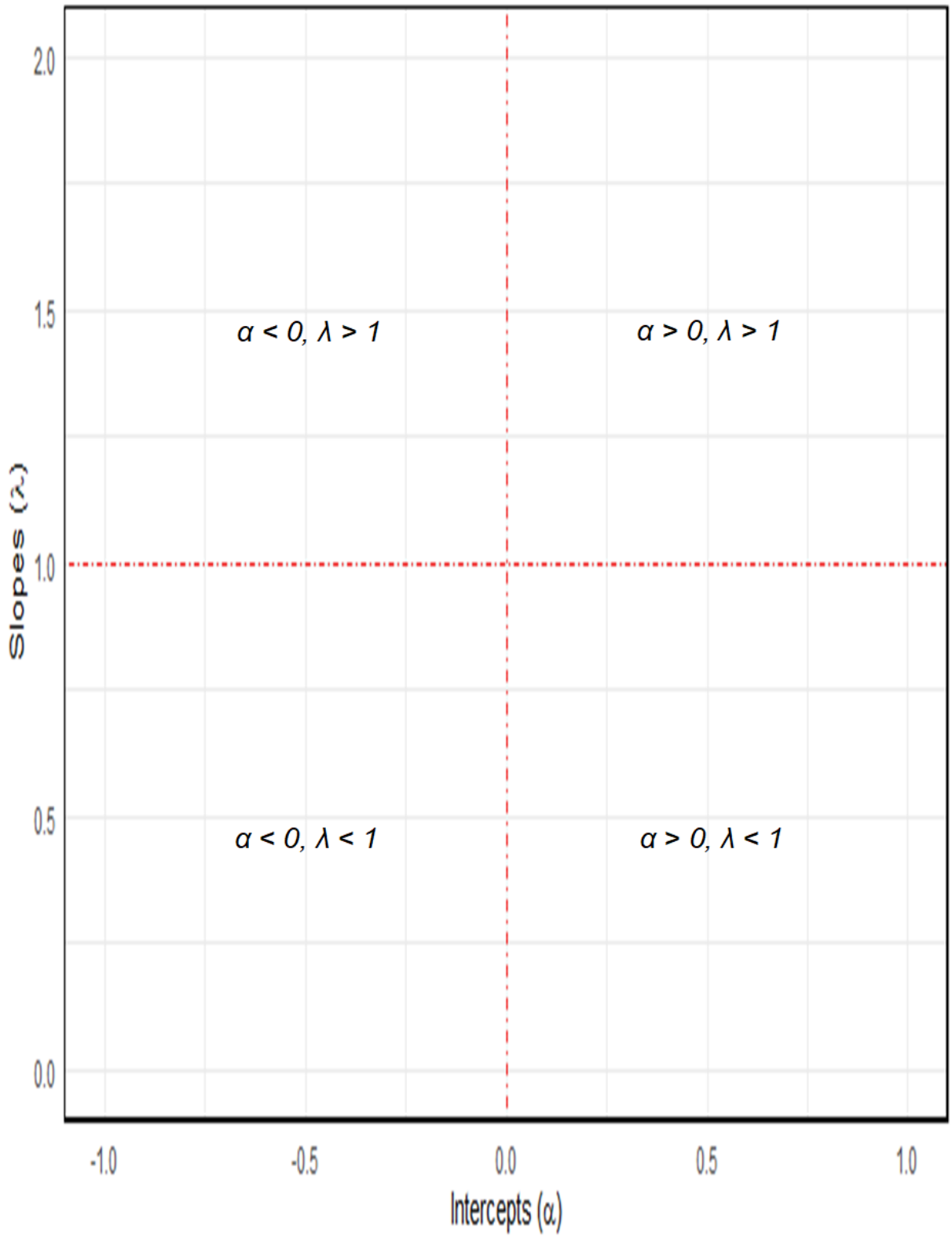

2.0
1.5
1.0
0.5
0.0
Slopes (λ)
α < 0, λ > 1
α > 0, λ > 1
α < 0, λ < 1
α > 0, λ < 1
-1.0
-0.5
0.0
0.5
1.0
Intercepts (α)

Figure 2

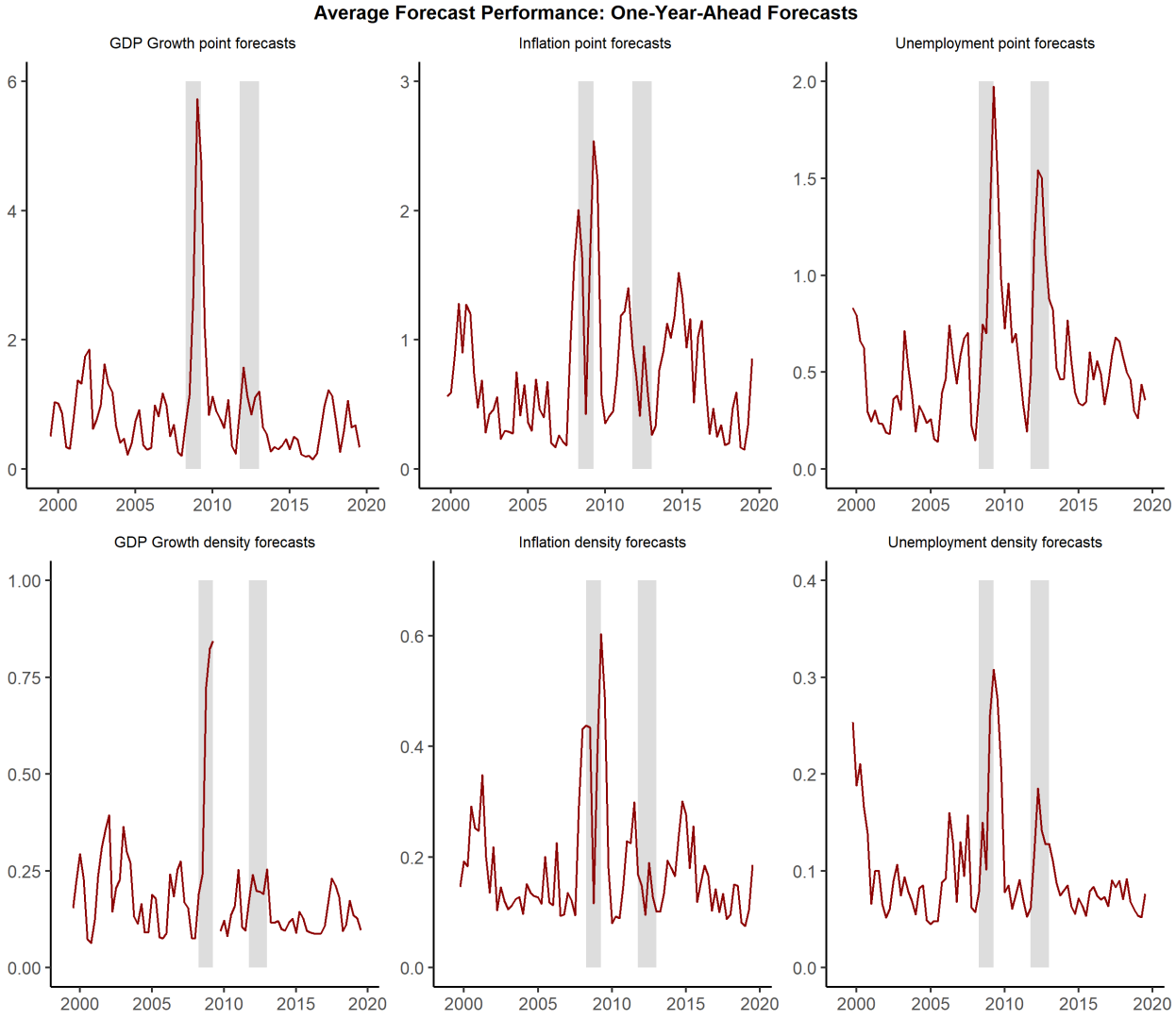

Average Forecast Performance: One-Year-Ahead Forecasts
GDP Growth point forecasts
Inflation point forecasts
Unemployment point forecasts
GDP Growth density forecasts
Inflation density forecasts
Unemployment density forecasts
2000
2005
2010
2015
2020

Figure 3

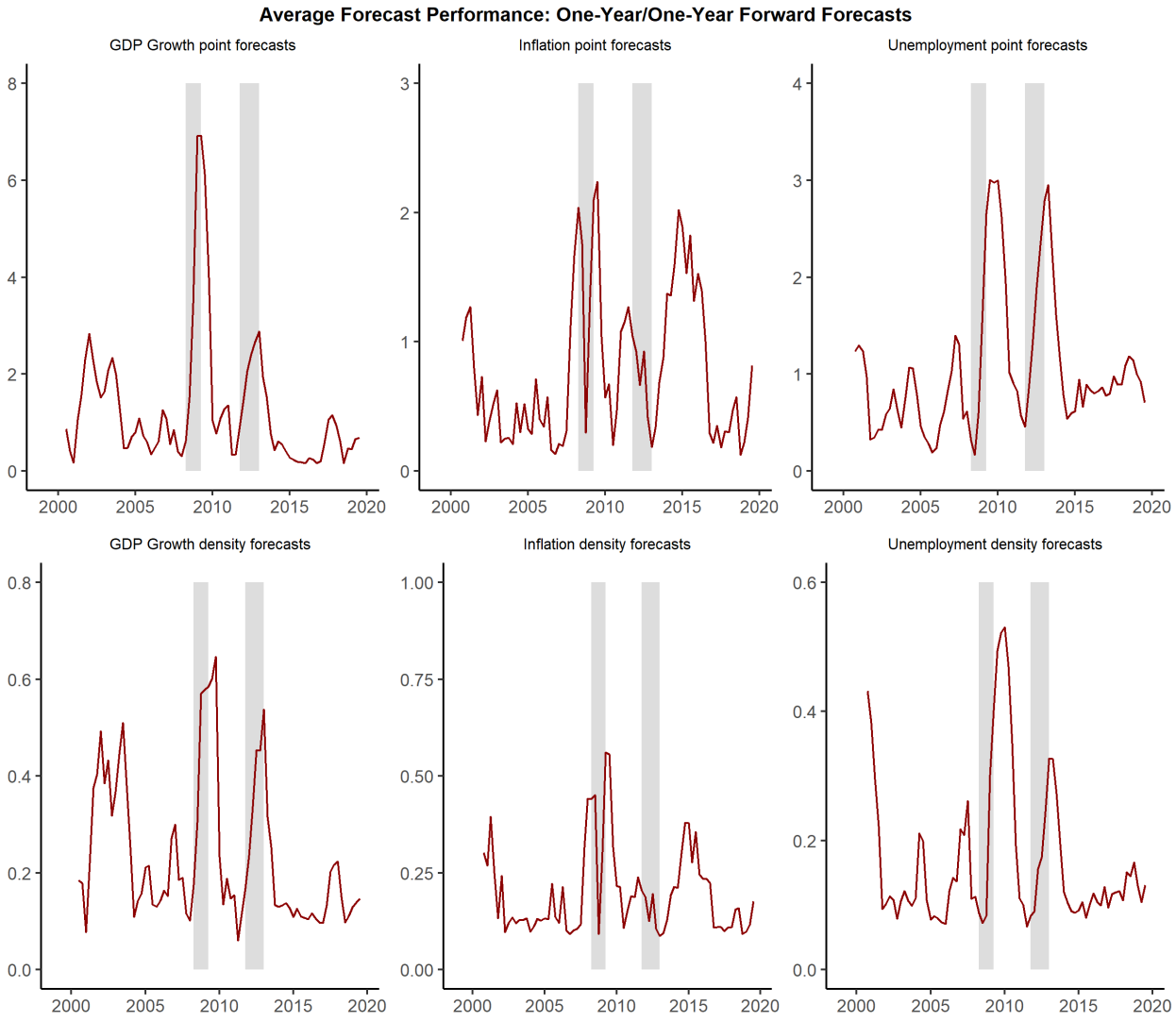

Average Forecast Performance: One-Year/One-Year Forward Forecasts
GDP Growth point forecasts
Inflation point forecasts
Unemployment point forecasts
GDP Growth density forecasts
Inflation density forecasts
Unemployment density forecasts
2000
2005
2010
2015
2020

Figure 4

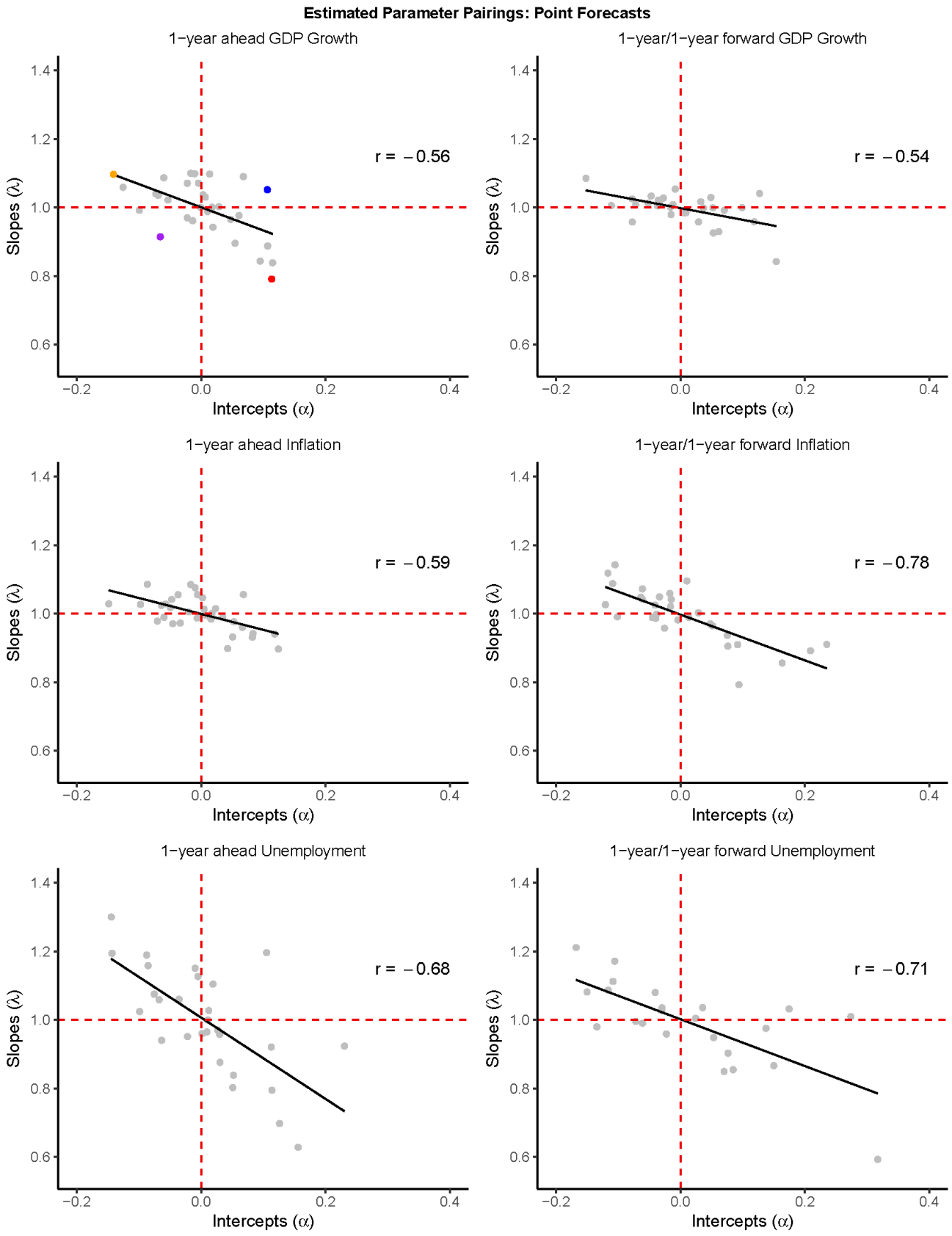

Estimated Parameter Pairings: Point Forecasts
1−year ahead GDP Growth
r = −0.56
1−year/1−year forward GDP Growth
r = −0.54
1−year ahead Inflation
r = −0.59
1−year/1−year forward Inflation
r = −0.78
1−year ahead Unemployment
r = −0.68
1−year/1−year forward Unemployment
r = −0.71
Slopes (λ)
Intercepts (α)
1.4
1.2
1.0
0.8
0.6
−0.2
0.0
0.2
0.4

Figure 5

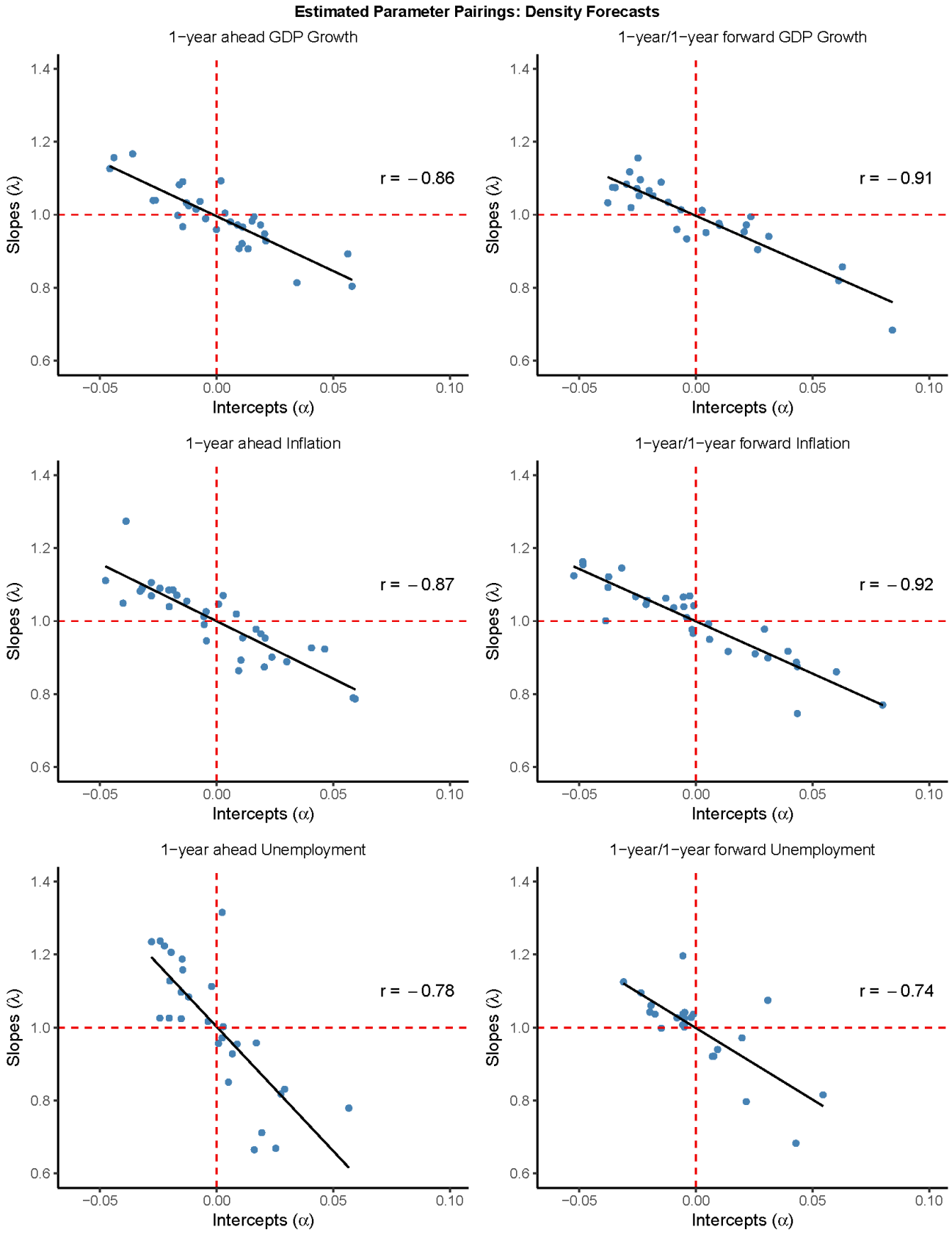

Estimated Parameter Pairings: Density Forecasts
1−year ahead GDP Growth
r = −0.86
1−year/1−year forward GDP Growth
r = −0.91
1−year ahead Inflation
r = −0.87
1−year/1−year forward Inflation
r = −0.92
1−year ahead Unemployment
r = −0.78
1−year/1−year forward Unemployment
r = −0.74
Slopes (λ)
Intercepts (α)
1.4
1.2
1.0
0.8
0.6
−0.05
0.00
0.05
0.10

Figure 6

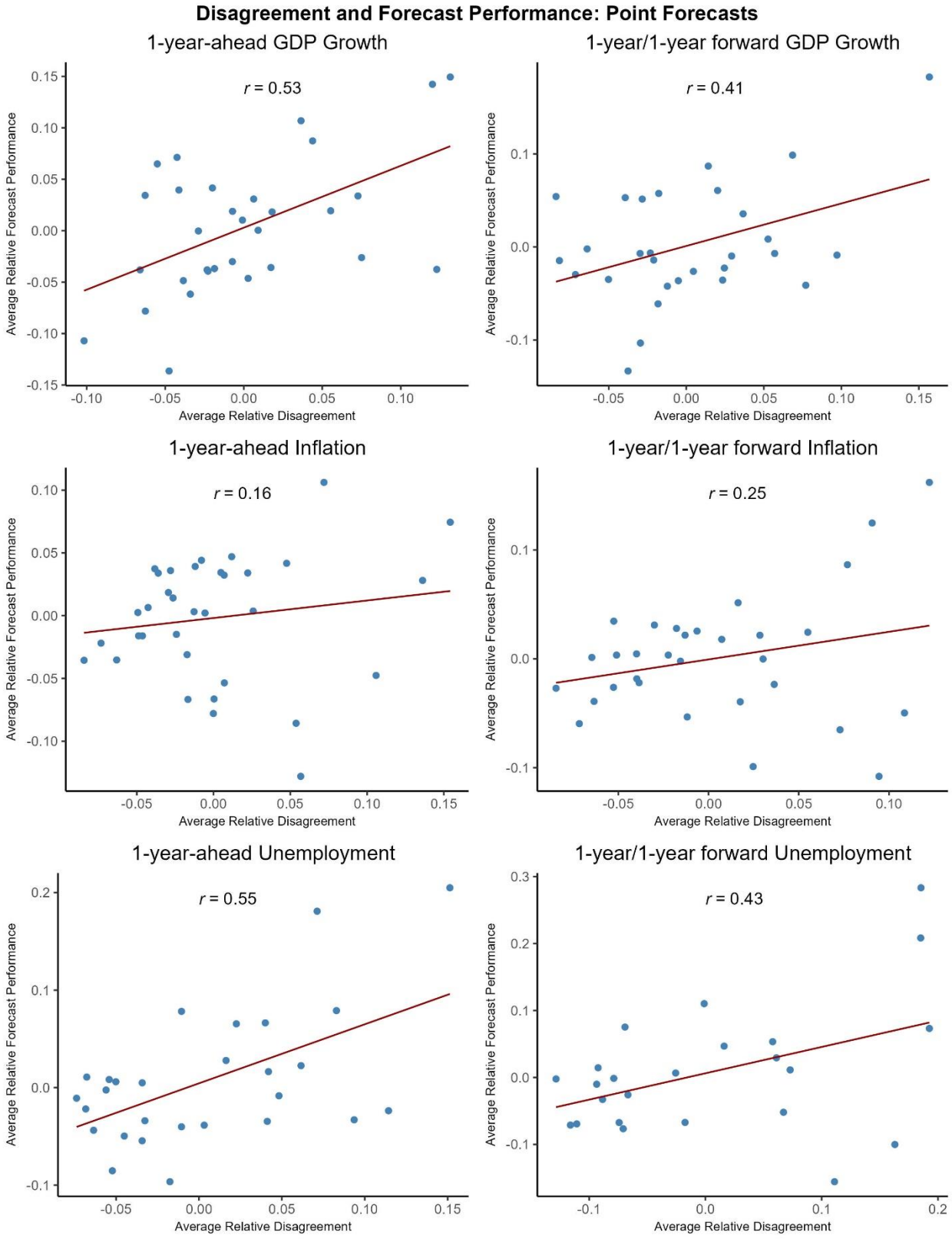

Disagreement and Forecast Performance: Point Forecasts
1-year-ahead GDP Growth
r = 0.53
1-year/1-year forward GDP Growth
r = 0.41
1-year-ahead Inflation
r = 0.16
1-year/1-year forward Inflation
r = 0.25
1-year-ahead Unemployment
r = 0.55
1-year/1-year forward Unemployment
r = 0.43
Average Relative Forecast Performance
Average Relative Disagreement

Figure 7

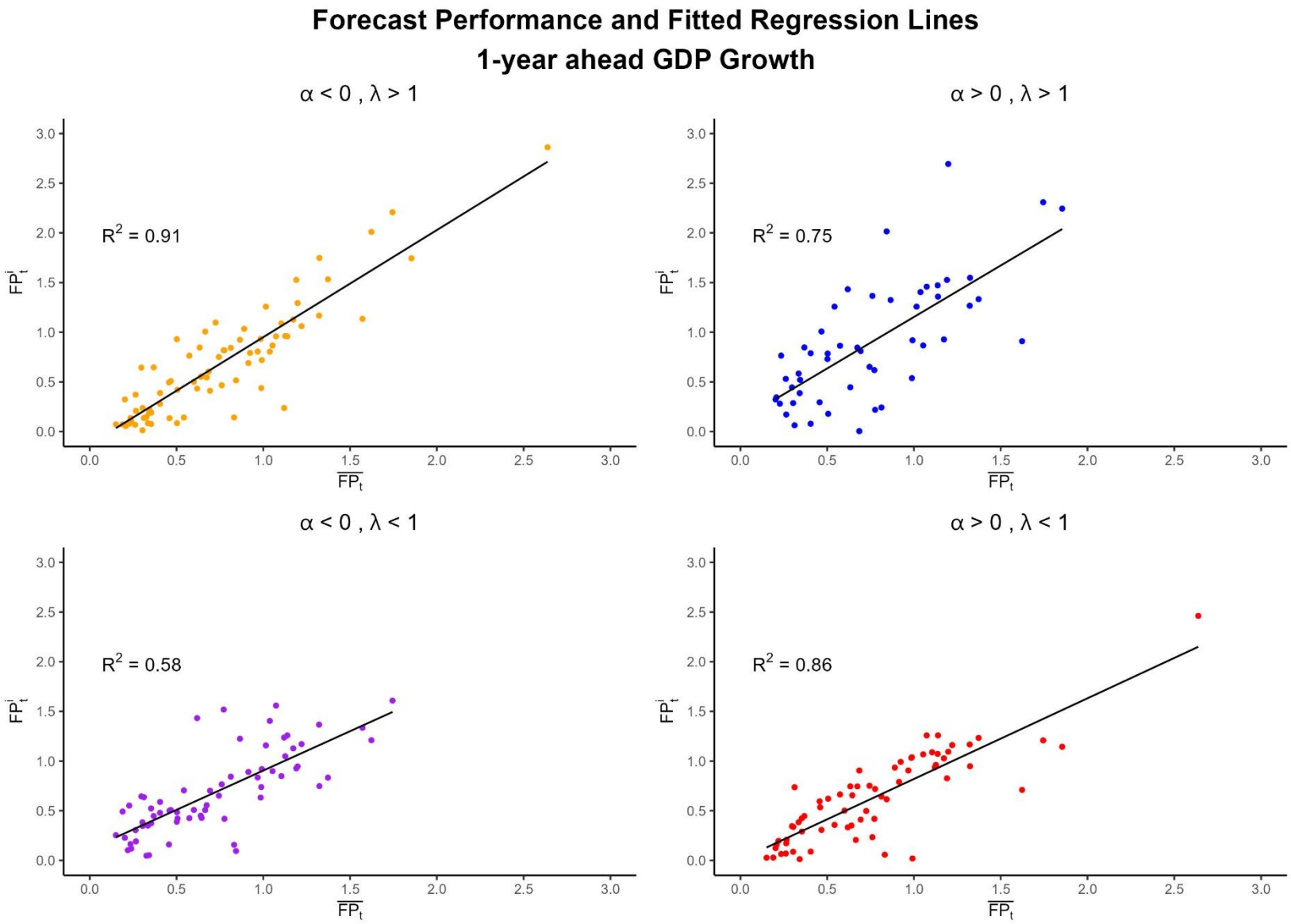
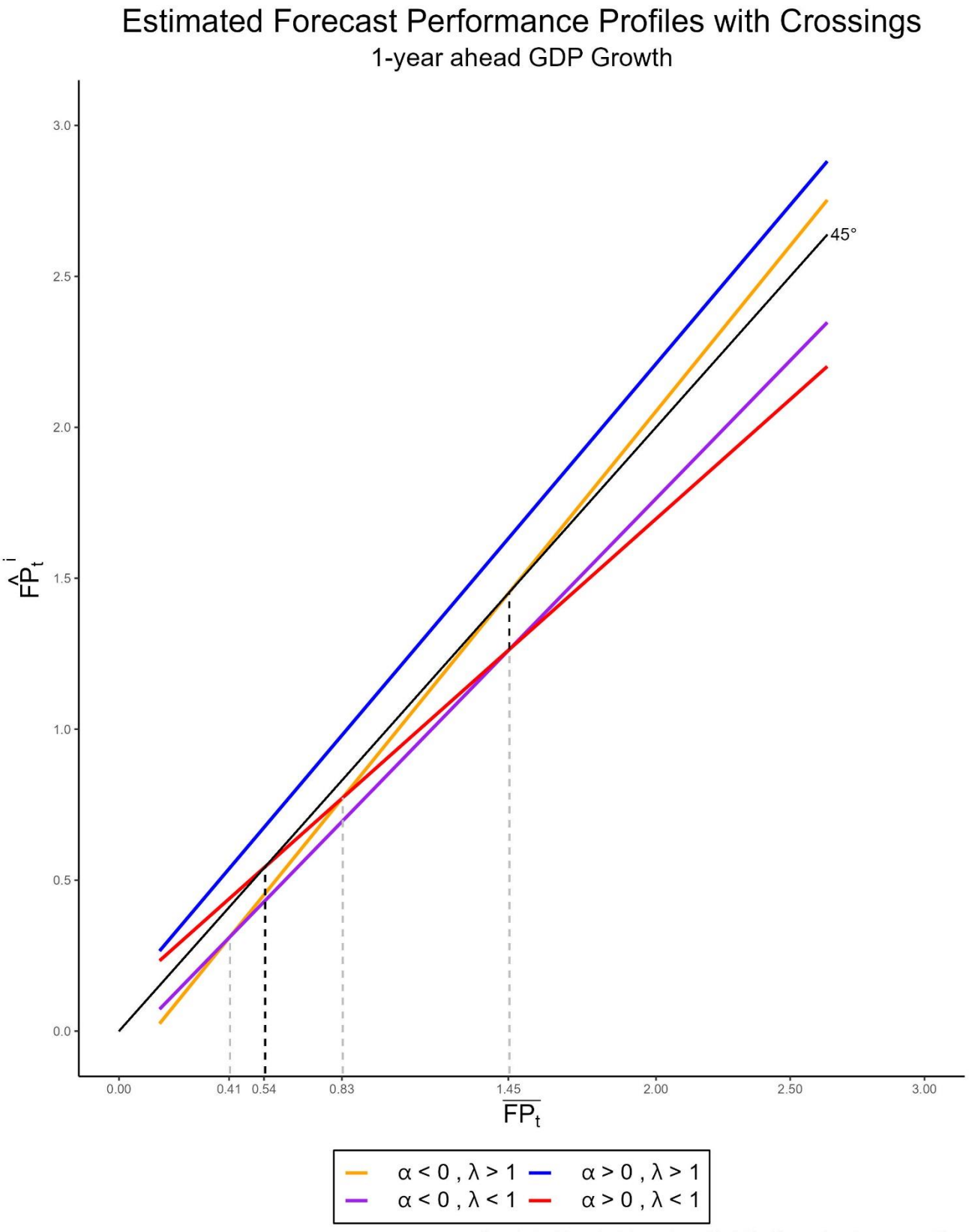
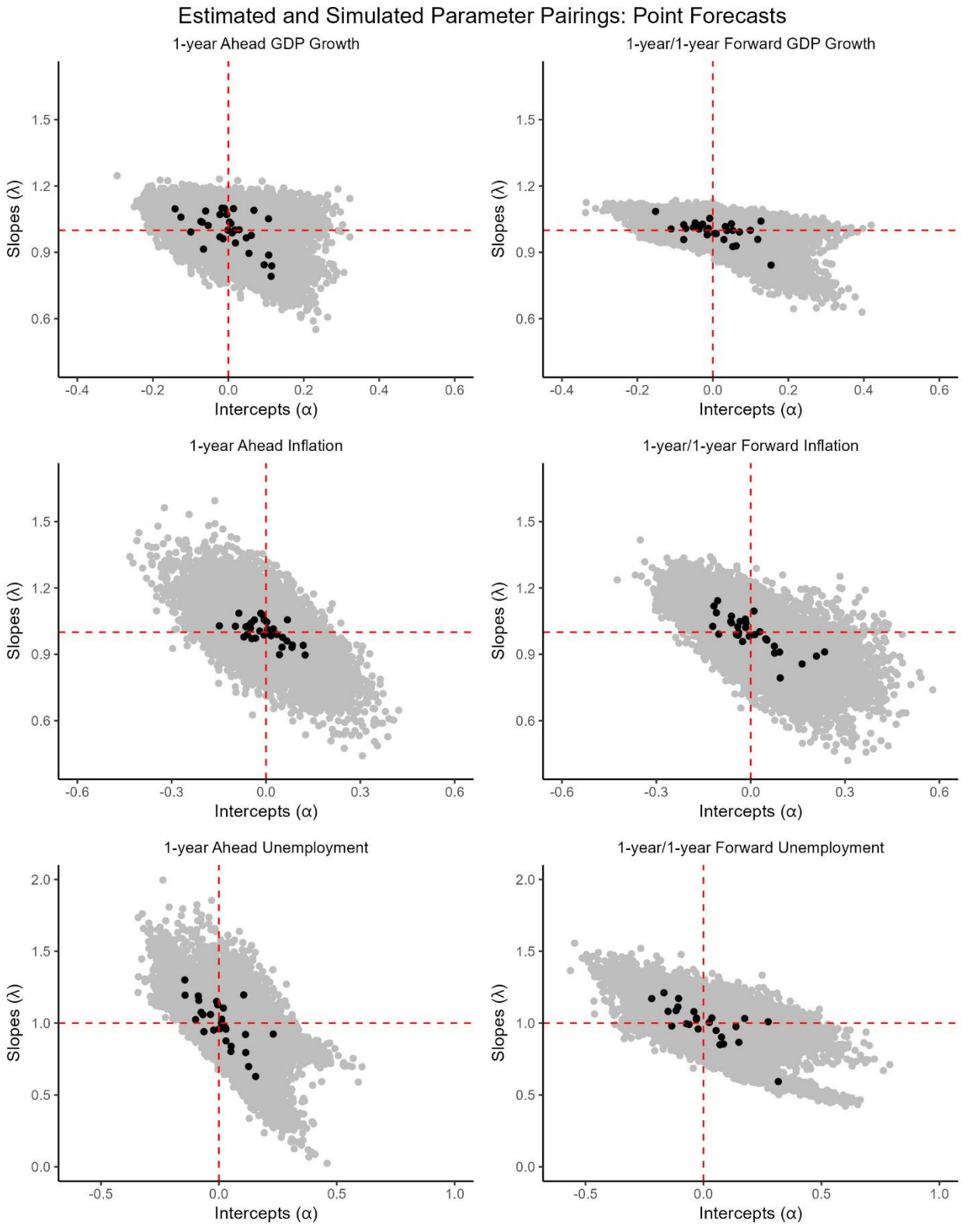

Forecast Performance and Fitted Regression Lines
1-year ahead GDP Growth
α < 0 , λ > 1
R2 = 0.91
α > 0 , λ > 1
R2 = 0.75
α < 0 , λ < 1
R2 = 0.58
α > 0 , λ < 1
R2 = 0.86
FPit
FPt

Figure 8



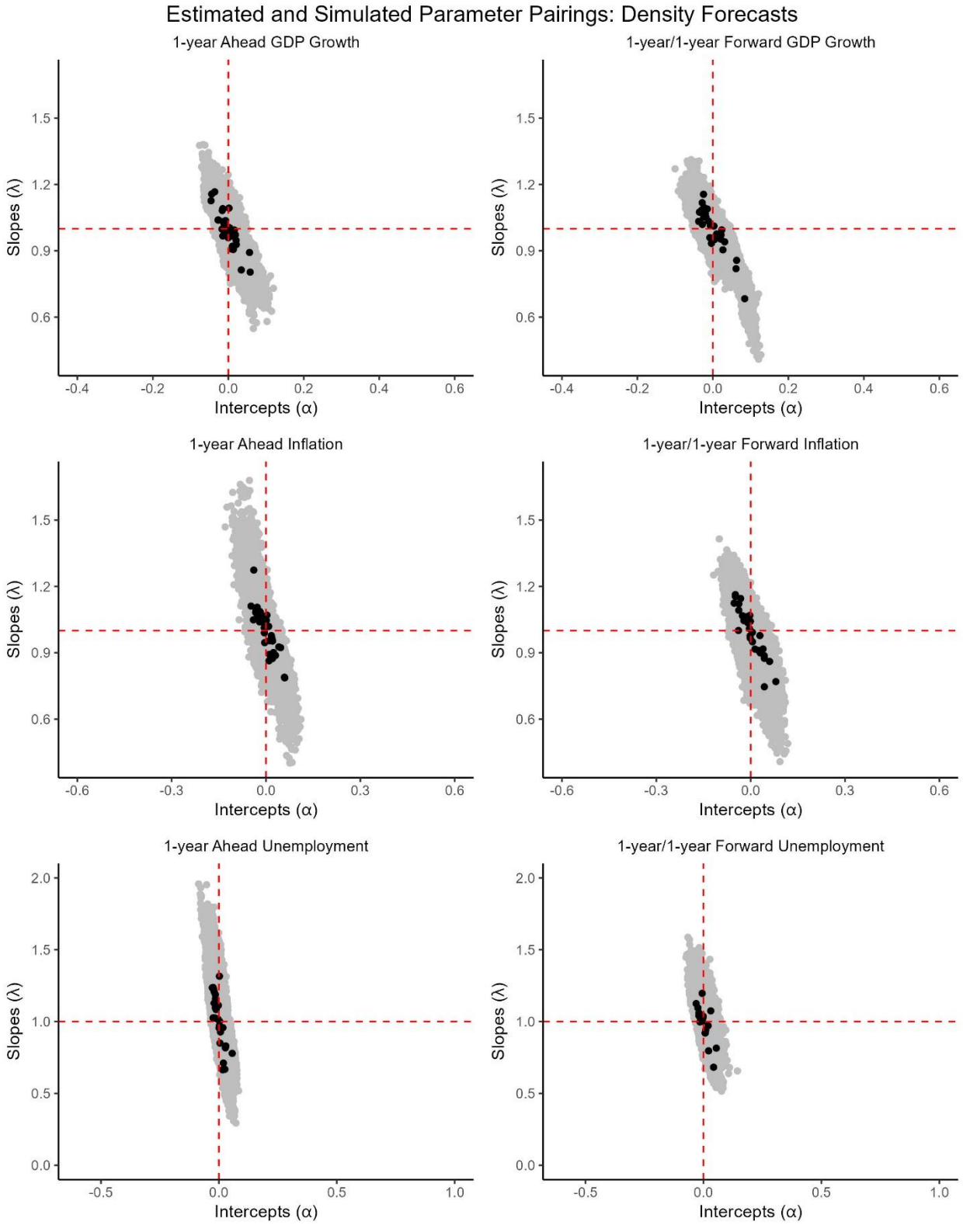
Grey dashed lines depict crossings of individual forecast performance profiles.
Black dashed lines depict crossings of individual forecast performance profiles with consensus forecast performance.

Figure 9



Note: Black dots are estimated values and grey dots are simulated values from the estimated joint distributions.

Figure 10



Note: Black dots are estimated values and grey dots are simulated values from the estimated joint distributions.

Figure 11

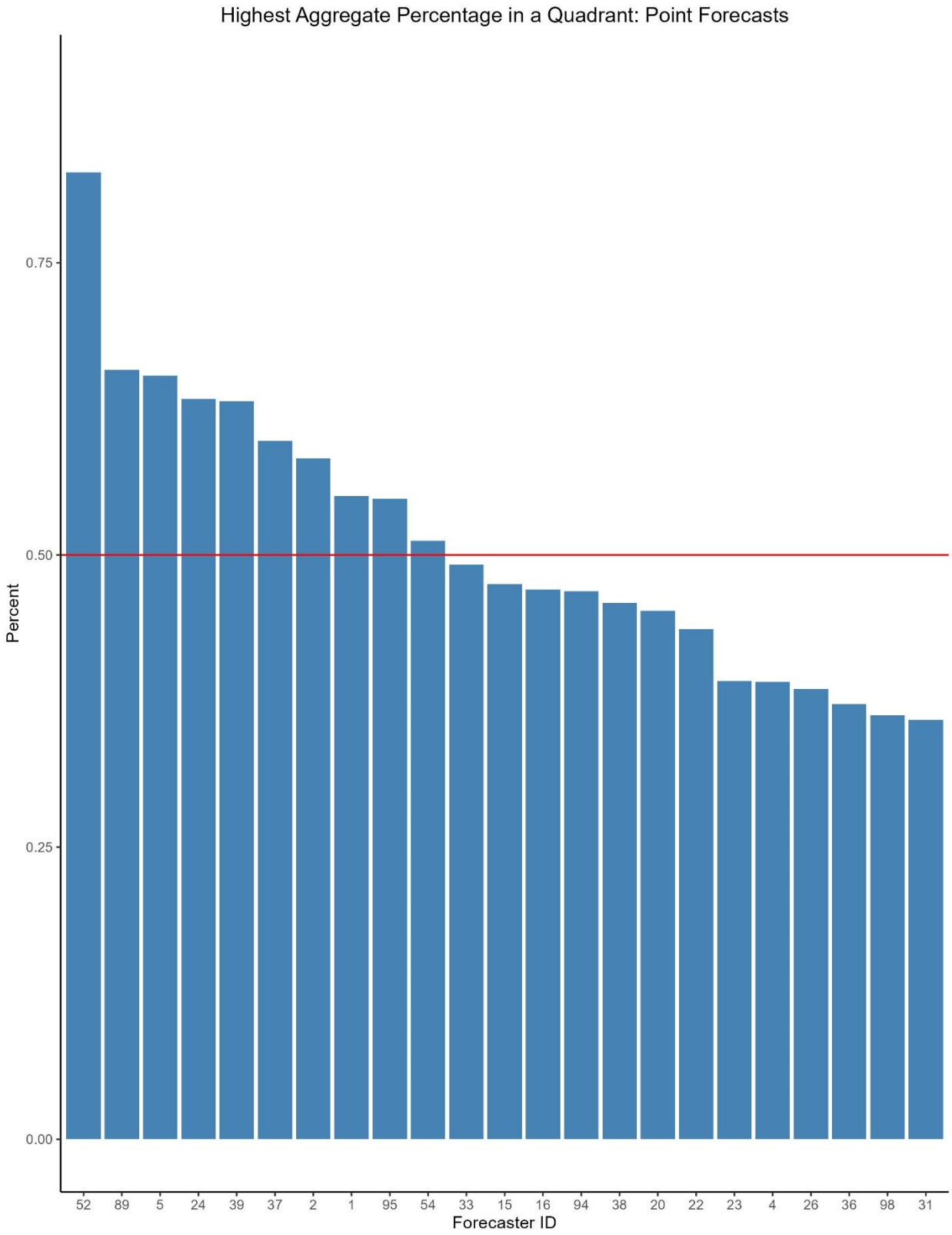

Highest Aggregate Percentage in a Quadrant: Point Forecasts
Percent
0.75
0.50
0.25
0.00
52 89 5 24 39 37 2 1 95 54 33 15 16 94 38 20 22 23 4 26 36 98 31
Forecaster ID

Figure 12

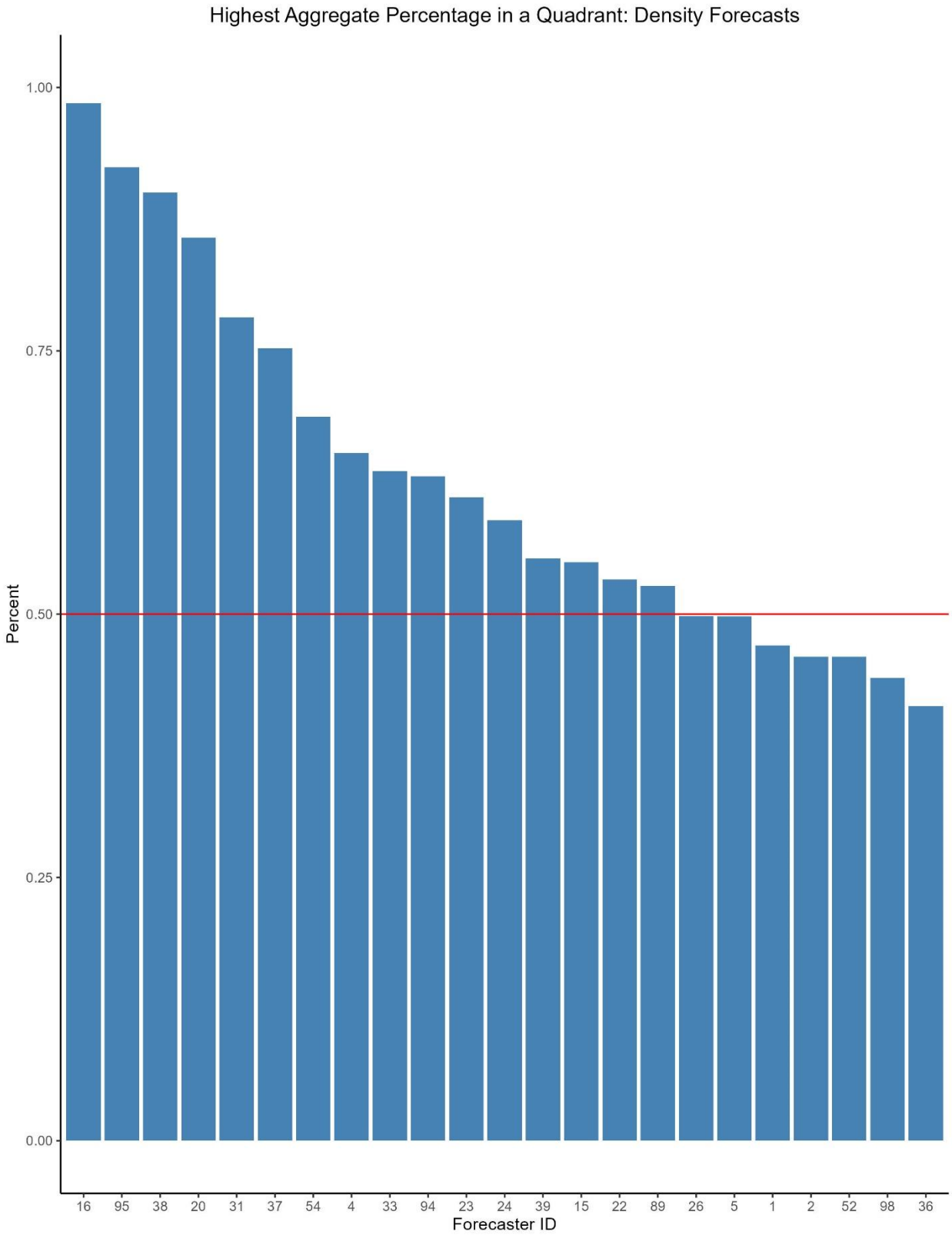
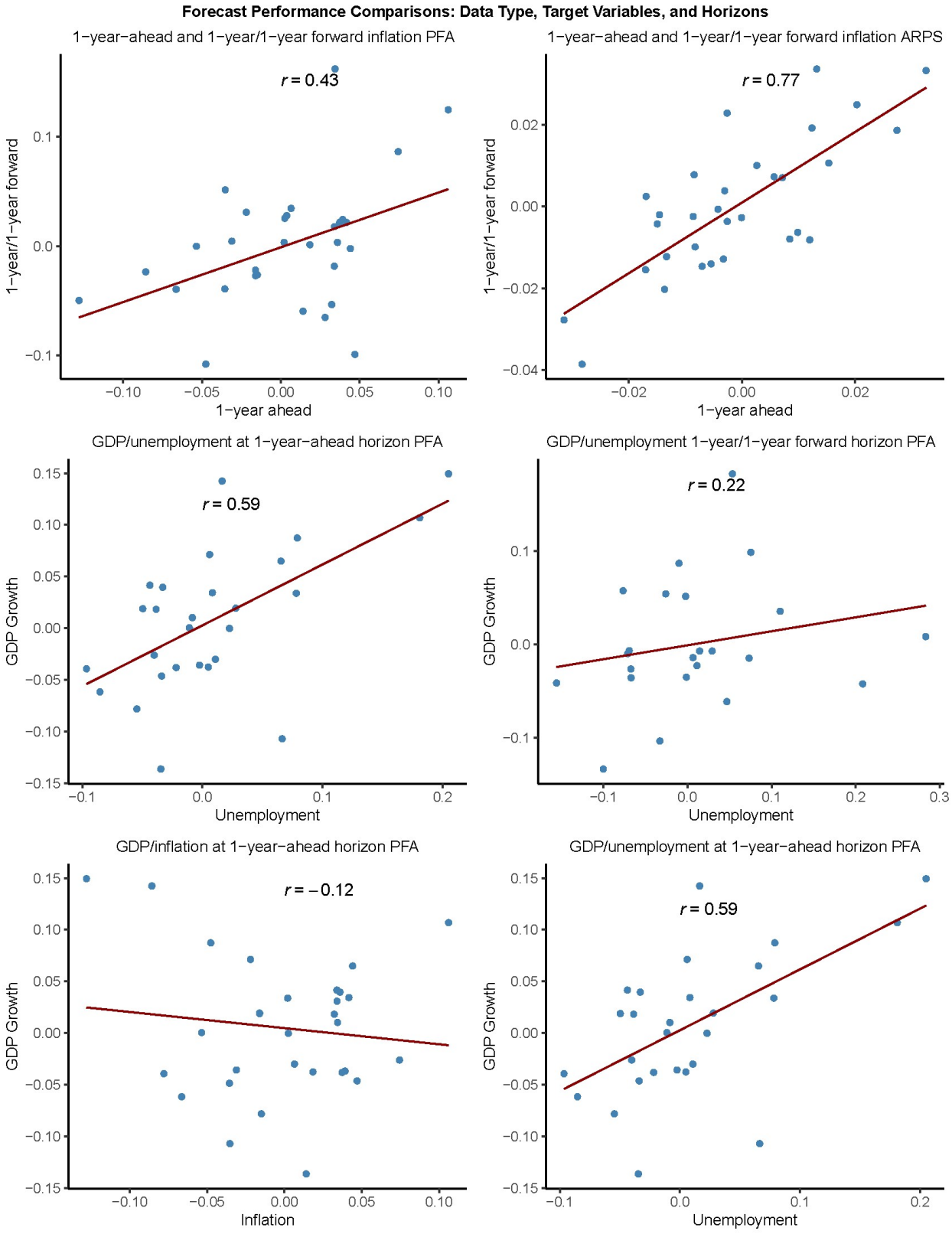

Highest Aggregate Percentage in a Quadrant: Density Forecasts
Percent
1.00
0.75
0.50
0.25
0.00
16 95 38 20 31 37 54 4 33 94 23 24 39 15 22 89 26 5 1 2 52 98 36
Forecaster ID

Figure 13


Forecast Performance Comparisons: Data Type, Target Variables, and Horizons
1−year−ahead and 1−year/1−year forward inflation PFA
r = 0.43
1−year/1−year forward
1−year ahead
1−year−ahead and 1−year/1−year forward inflation ARPS
r = 0.77
1−year/1−year forward
1−year ahead
GDP/unemployment at 1−year−ahead horizon PFA
r = 0.59
GDP Growth
Unemployment
GDP/unemployment 1−year/1−year forward horizon PFA
r = 0.22
GDP Growth
Unemployment
GDP/inflation at 1−year−ahead horizon PFA
r = −0.12
GDP Growth
Inflation
GDP/unemployment at 1−year−ahead horizon PFA
r = 0.59
GDP Growth
Unemployment

References

Boero, Gianna, Jeremy Smith, and Kenneth F. Wallis. "The Measurement and Characteristics of Professional Forecasters' Uncertainty." *Journal of Applied Econometrics* 30 (December 2015): 1029-1046.

Bowles, Carlos, Roberta Friz, Veronique Genre, Geoff Kenny, Aidan Meyler, and Tuomas Rautanen. "The ECB Survey of Professional Forecasters (SPF): A Review After Eight Years' Experience." ECB Occasional Paper No 59. European Central Bank, April 1, 2007.

Bruine de Bruin, Wandi, Charles F. Manski, Giorgio Topa, and Wilbert van der Klaauw. "Measuring Consumer Uncertainty About Future Inflation." *Journal of Applied Econometrics* 26 (May 2011): 454-478.

Clements, Michael P. "Forecaster Efficiency and Disagreement: Evidence Using Individual-Level Survey Data." *Journal of Money, Credit and Banking* 54 (April 2022): 537-568.

Coibion, Olivier, and Yuriy Gorodnichenko. "What Can Survey Forecasts Tell Us About Information Rigidities?" *Journal of Political Economy* 120 (February 2012): 116-159.

------, and Yuriy Gorodnichenko. "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts." *American Economic Review* 105 (August 2015): 2644-2678.

D'Agostino, Antonello, Kieran McQuinn, and Karl Whelan. "Are Some Forecasters Really Better Than Others?" *Journal of Money, Credit, and Banking* 44 (June 2012): 715-732.

Garcia, Juan A. "An Introduction to the ECB's Survey of Professional Forecasters." ECB Occasional Paper No 8. European Central Bank, 2003.

Genre, Veronique, Geoff Kenny, Adian Meyler, and Allan Timmermann. "Combining Expert Forecasts: Can Anything Beat the Simple Average?" *International Journal of Forecasting* 29 (March 2013): 108-121.

Glas, Alexander, and Matthias Hartman. "Uncertainty Measures from Partially Rounded Probabilistic Forecast Surveys." *Quantitative Economics* 13 (July 2022): 979-1022.

Hounyo, Ulrich, and Kajal Lahiri. "Are Some Forecasters Really Better Than Others? A Note." *Journal of Money, Credit and Banking* 55 (April 2023): 577-593.

Kenny, Geoff, Thomas Kostka, and Federico Masera. "How Informative are the Subjective Density Forecasts of Macroeconomists?" *Journal of Forecasting* 33 (April 2014): 163-185.

------, Thomas Kostka, and Federico Masera. "Can Macroeconomists Forecast Risk? Event-Based Evidence from the Euro-Area SPF." *International Journal of Central Banking* 11 (December 2015): 1-46.

------, Thomas Kostka, and Federico Masera. "Density Characteristics and Density Forecast Performance: A Panel Analysis." *Empirical Economics* 48 (May 2015): 1203-1231.

MacKowiak, Bartosz, and Mirko Wiederholt. "Optimal Sticky Prices Under Rational Inattention." *American Economic Review* 99 (June 2009): 769-803.

Mankiw, Gregory N., and Ricardo Reis. "Sticky Information Versus Sticky Prices: A Proposal to Replace the New Keynesian Phillips Curve." *Quarterly Journal of Economics* 117 (November 2002): 1295-1328.

------, Ricardo Reis, and Justin Wolfers. "Disagreement about Inflation Expectations." *NBER Macroeconomics Annual* 18 (2003): 209-248.

Meyler, Aidan. "Forecast Performance in the ECB SPF: Ability or Chance?" Working Paper Series. European Central Bank, 2020.

Newey, Whitney K., and Kenneth D. West. "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55 (May 1987): 703-708.

Pesaran, M. Hashem. "Estimation and Inference in Large Heterogeneous Panels with a Multifactor Error Structure." *Econometrica* 74 (July 2006): 967-1012.

Qu, Ritong, Allan Timmermann, and Yinchu Zhu. "Do Any Economists Have Superior Forecasting Skills?" Working Paper. University of California - San Diego, October 1, 2019.

------, Allan Timmermann, and Yinchu Zhu. "Comparing Forecasting Performance in Cross-Sections." *Journal of Econometrics* 237 (December 2021).

Rich, Robert W., and Joseph Tracy. "A Closer Look at the Behavior of Uncertainty and Disagreement: Micro Evidence from the Euro Area." *Journal of Money, Credit, and Banking* 53 (February 2021): 233-253.

Sims, Christopher A. "Implications of Rational Inattention." *Journal of Monetary Economics* 50 (April 2003): 665-690.

Timmerman, Allan. "Forecast Comparisons." In *Handbook of Economic Forecasting, Volume 1 Chapter 4*, edited by G. Elliott, C. Granger and A. Timmerman, 135-196. Elsevier, 2006.

Woodford, Michael. "Imperfect Common Knowledge and the Effects of Monetary Policy." In *Knowledge, Information, and Expectations in Modern Macroeconomics: In Honor of Edmund S. Phelps*, edited by Philippe Aghion, Roman Frydman, Joseph Stiglitz and Michael Woodford. Princeton, Princeton University Press, 2003.

To view the read-me file, code, data, and output of *All Forecasters Are Not the Same: Systematic Patterns in Predictive Performance*, please click on the Dropbox link below:

https://www.dropbox.com/scl/fi/6b4f7trpqdo0wtbhaae4n/REPLICATION-PACKAGE.zip?rlkey=tfkz4b8jtl8d6z85mdbtjtiba&st=8xd616ds&dl=0

When prompted to enter a passcode, please use the following string:

IJFForecasterHeterogeneity2025!