Subjective beliefs, disagreement, and market return predictability[*]

Felipe S. Iachan [†1] and **Raul Riva** [‡2]

[1]FGV EPGE

[2]Northwestern University

First Version: March 6, 2024

This Version: March, 22, 2024

Abstract

We jointly evaluate the ability of consensus beliefs about long-term earnings growth and analyst disagreement to predict stock market returns. We bridge the gap between the strands of literature testing theories of belief extrapolation and disagreement-fueled over-pricing. Using I/B/E/S analyst forecast data from 1982-2022 and conducting both in-sample and out-of-sample analyses, we find limited evidence that these measures of subjective beliefs predict returns. Our analyses across specifications and sample periods highlight the fragility of prior findings after including relevant controls or varying sample periods.

Keywords: Subjective beliefs; predictability; out-of-sample; earnings-per-share forecasts; disagreement; I/B/E/S;

JEL Classification: G12; G17; G40.

[*] We thank Torben Andersen, Aditya Chaudhry, Robert Korajczyk, Humberto Moreira, Viktor Todorov, and Beverly Walther for early comments and helpful discussions.

Raul Riva is grateful for the hospitality of FGV EPGE on multiple occasions, where a large portion of this work was conducted.

[†] https://sites.google.com/site/fiachan

[‡] https://rgriva.github.io/

1 Introduction

A recent surge of interest in measuring market participants' beliefs has emerged, with these measures being utilized to account for asset pricing data. Incorporating such beliefs leads to conclusions that substantially diverge from those derived from representative-agent rational expectation models. For instance, empirical subjective beliefs are volatile enough to account for the price (or price-earnings ratio) volatility seen in the data (De La O and Myers, 2021; Bordalo et al., 2023), and there is no evidence of a time-varying subjective risk premium (Nagel and Xu, 2023). These conclusions starkly contrast with the interpretation of evidence based on rational expectations (**?**).

This empirical behavior of subjective beliefs has called attention to theories of extrapolation and overpricing. These theories suggest that market returns can be predictable based on knowledge of these subjective beliefs. A current realization of high earnings growth leads to excessive optimism that fuels high prices first and disappointing returns later. Therefore, beliefs about long-term growth can forecast low future returns (La Porta, 1996; Bordalo et al., 2019, 2023).

In parallel, another important strand of literature deals with subjective beliefs, overpricing, and the empirical behavior of returns. It is focused on disagreement, which is also significant in survey data and abstracted from by rational-expectation-based models. The theory has an establishing contribution by Miller (1977). It has been taken to data by Yu (2011) and Hong and Sraer (2016), among others.[1] The theoretical mechanism of this strand can also be cast and tested in terms of overpricing and its anticipation of low returns. In the presence of short-selling constraints, high disagreement today means optimists bid up asset prices without the disciplining role of pessimists. Consequently, there is overpricing today and lower returns in the future.

These two strands suggest a central role for subjective beliefs in asset pricing. Tests of return predictability can serve as tests of the underlying theories. Empirically, both extrapolation and disagreement-fueled overpricing might happen. They are easy to confound, and their relative contributions to return predictability are unknown.

We intend to combine these two strands and assess return predictability based on average (consensus) beliefs about long-term earnings growth (`LTG`) and disagreement (`D`) among forecasters regarding that variable. We nest the main model specifications and sample periods from the previous literature, offering a joint evaluation of the roles played by consensus and disagreement in asset pricing. Our main data source is the Institutional Brokers' Estimate System (I/B/E/S) database of analyst forecasts of long-term earnings growth. We complement this data with CRSP for stock prices and market capitalization and Robert Shiller's cyclically adjusted price-earnings ratio of the S&P 500 index. We construct value-weighted measures of consensus long-term earnings growth forecasts and analyst disagreement using I/B/E/S data, following standard procedures in the literature. We focus on long-term growth for a few

[1] See, for instance, Li (2016), Gao et al. (2019), and Huang et al. (2021) for related work on disagreement and asset prices.

reasons: it is likely less contaminated by earnings guidance from firms (Yu, 2011; Hong and Sraer, 2016), it is offered in growth rates, which makes it directly comparable across firms and over time, and it is relatively standard in the empirical literature that this paper evaluates.

Our findings on market return predictability using subjective beliefs are dismal. In sample, market excess return predictions are unstable across model specifications and samples. Important results from both strands of the literature fail to hold with the inclusion of additional controls like the price-earnings ratio, which has stronger forecasting power than both variables, or with sample extensions to include data unavailable in the original analyses.[2] This is true across different forecasting horizons, ranging from 12 to 60 months ahead.

We complement the in-sample analysis with an investigation of out-of-sample predictability of aggregate returns using consensus long-term growth and disagreement. We design a recursive forecasting scheme using a linear model that nests specifications from previous in-sample studies. In the first part of our predictive analysis, we use an out-of-sample period ranging from July 2007 up to December 2022. This timeline ensures a rich learning phase, that includes events like the dot-com burst in early 2000, while also allowing models to be tested during challenging periods such as the financial crisis, the unconventional monetary policy period, and the coronavirus pandemic.

The second part of our out-of-sample investigation uses different starting points for our out-of-sample period and shows how our findings change depending on that choice. Our conclusions are: (i) LTG has some forecasting power for market returns, but only at short horizons; (ii) there is no strong evidence, however, that LTG brings additional forecasting power relative to the traditional price-earnings index. If anything, the predictability through LTG appears only at more recent periods; and (iii) disagreement does not offer any meaningful prediction power, either as an individual variable or combined with other analysis variables.

Relationship with literature. We contact at least two strands of literature in asset pricing and aim to improve their connection. The first is the literature on asset pricing under heterogeneous beliefs, building on seminal theoretical contributions from Miller (1977) and Jarrow (1980). More recently, the problem of price formation under disagreement was revisited by Atmaz and Basak (2018), whose model does not feature short-selling constraints.[3] Under some parameter restrictions in these models, disagreement is translated to higher stock prices (or lower returns). Diether et al. (2002), Park (2005), and Yu (2011) are empirical tests in favor of such an argument, although their analyses are not out-of-sample.[4] Hong and Sraer (2016) use analyst disagreement to explain a flatter capital market line. More recently, Huang et al. (2021) tackled the out-of-sample aggregate return prediction problem with different disagreement measures and showed that they all individually fail, but there is a disagreement index extracted from these

[2] We discuss the relationship between the price-earnings ratio and our main variables of study and the adequacy of its inclusion in predictive regressions in Section 4.2.

[3] See also related work by Iachan et al. (2021) and the survey by Simsek (2021).

[4] Additionally, we refer to Hong and Stein (2007) for an early review of the literature.

measures that has prediction power. Their analysis does not consider both first and second moments of subjective beliefs in prediction simultaneously, as we do. Goyal et al. (2021) revisit and update the contribution of Welch and Goyal (2008). They find similar unappealing out-of-sample prediction power for the belief disagreement measure, although they only study univariate models.

The second strand is associated with a step outside the rational expectations framework, taking seriously the evidence from survey data. For instance, Greenwood and Shleifer (2014) show that consumer subjective expected returns are not lower in good times, as most traditional asset pricing models imply. De La O and Myers (2021) explain much of the price-earnings ratio variability using subjective beliefs about dividends and earnings, even with constant discount rates. Nagel and Xu (2023) suggest a positive risk-return trade-off in subjective expectations. Giglio et al. (2021) document, among other findings, a trade-off between subjective expected returns and disaster risk and limited (but heterogeneous) sensitivity of portfolios to beliefs.

More closely related to us, La Porta (1996) and Bordalo et al. (2019) show how long-term earnings growth from the average forecaster from I/B/E/S can help price the cross-section of stocks, but they do not analyze aggregate returns. Bordalo et al. (2023) analyzes the relationship between this consensus measure and the aggregate market, but they do not have disagreement in their setup and focus only on in-sample analyses.

The remainder of the paper is organized as follows. Section 2 discusses our data sources, while Section 3 details how to construct both the consensus and disagreement measures. Section 4 discusses our in-sample results; out-of-sample analyses are discussed in Section 5. Section 6 concludes the paper.

2 Data

Our data come from three different sources. First, we use CRSP data to collect end-of-month closing prices and the number of outstanding shares. Our notion of market return here is also derived from the value-weighted index available through CRSP.[5] This data delivers the main target we try to predict and the respective market capitalization weights we use throughout the paper. As usual, we consider stocks that traded at the NYSE, AMEX, and NASDAQ. We also focus on common stock, using share codes 10 and 11 from CRSP. Our notion of asset returns is the arithmetic return, regardless of the horizon.

We also use data from I/B/E/S to measure analyst forecasts.[6] Since March 1976, I/B/E/S has collected professional analysts' forecasts for many different performance measures for firms, such as earnings per share, return on investment, return on equity, etc. The prime forecast variable in I/B/E/S is, however,

[5] The correlation with S&P500 returns is over 99% at the monthly horizon. We experimented using S&P500 returns, and the results were nearly identical. For the sake of space, we don't report those, although they are available upon request.

[6] For a recent discussion of I/B/E/S-related measures, see Adam and Nagel (2023).

earnings per share, and that's the one we concentrate on. This is the variable with the best coverage on I/B/E/S , both in terms of time series and cross-sectional availability. Also, it is less prone to endogenous choices from the firm's management side we do not seek to model, such as dividends per share. Analysts are asked to forecast firm earnings at different horizons on a rolling basis, ranging from the subsequent fiscal quarter up to a long-term earnings growth forecast. These forecasts are collected (and updated) on a monthly frequency.[7]

Using this long-term earnings growth forecast over other I/B/E/S measures has at least three advantages. First, analysts report a growth *rate* for the long-term forecast while they forecast *dollar amounts* in other cases. When using rates, we do not need to worry about a numeraire. In contrast, due to inflation, dollar amounts from different times require some normalization. These forecasts based on rates are, therefore, more comparable over time. Second, this long-term forecast, in theory, should be less influenced by short-term fluctuation in the business cycle and reflect a cleaner assessment of growth opportunities for the firms. Third, we use data that is comparable to previous literature.

We also acknowledge an important unfavorable aspect of long-term forecasts. Analysts are required to forecast an annual growth rate over a period that is, unfortunately, loosely defined. The precise definition of the rate they should forecast is *"... the expected annual increase in operating earnings over the company's next full business cycle. These forecasts refer to a period of between three to five years"*. We understand that different analysts might have slightly different horizons in mind when making those forecasts. To remedy this issue, aside from noting that analysts report annual rates anyway, we study forecasting performance for shorter and longer horizons and show that our results are consistent regardless of this choice. We also note that long-term forecasts are available to us starting in December 1981, and that's the starting point of our empirical analyses in the next sections. The last month included is December 2022. That amounts to 493 monthly observations.

The last source of data is Robert Shiller's website, from which we collect the cyclically adjusted price-over-earnings ratio (`CAPE` on his website) for the S&P500 index. That is a standard measure in asset pricing that compares, on average, how expensive equity is compared to earnings from the biggest American companies. It is also a common variable to include in forecasting regressions equity returns.[8]

[7] We note that we are silent about non-US firms. We believe, however, that investigating international evidence is an important step in this agenda.

[8] See, for example, **?**, La Porta (1996), Welch and Goyal (2008), and Yu (2011).

3 Constructing the Variables

3.1 From CRSP

We use the value-weighted US CRSP index as our notion of index prices to construct returns. We use end-of-month prices. We denote by $R_{t+h|t}$ the h-month arithmetic index return, whose price at time t is denoted by S_t:

$$R_{t+h|t} = \frac{S_{t+h} - S_t}{S_t}.$$

We work with this variable at a monthly frequency, although we study predictability for different values of $h > 1$. That implies we work with overlapping returns. We know this choice induces a moving-average component in linear regression error terms and try to remedy that using HAC-adjusted inference with different lag lengths below.

3.2 From I/B/E/S

All variables used in the analyses evolve at the monthly frequency. Our construction follows standard procedures along the lines of Yu (2011), Huang et al. (2021), and Bordalo et al. (2023), for example. In any given month, we consider firms from I/B/E/S for which we have at least three analysts reporting forecasts. Increasing the threshold for the number of analysts implies that averaging over different forecasts will be less noisy, but it does introduce survivorship bias. This bias occurs because a higher number of analysts typically corresponds with larger firms. Given that our analyses weight firms by market value, there's a deliberate choice to include a broader array of firms. We believe that the benefits of reducing noise in our data outweigh bias risks, preferring inclusivity to ensure a comprehensive market representation.

Then, for each firm i at month t, we track the average over forecasts, which we denote by $LTG_{i,t}$. We also track the standard deviation of these forecasts across analysts, which is a measure of the disagreement, denoted by $D_{i,t}$. We winsorize both measures for each t at the 1% and 99% quantiles to offset the influence of possible outliers. Moreover, we use CRSP to collect their closing price on the last trading day of the month t, denoted by $P_{i,t}$ and the number of outstanding shares $Q_{i,t}$. We then define the weights $w_{i,t}$ as

$$w_{i,t} \equiv \frac{P_{i,t} \cdot Q_{i,t}}{\sum_{j=1}^{N_t} P_{j,t} \cdot Q_{j,t}},$$

where N_t is the number of firms that survived the first stage. Finally, we define the long-term earnings growth consensus LTG_t and the disagreement about this measure D_t as

$$LTG_t \equiv \sum_{i=1}^{N_t} w_{i,t} \cdot LTG_{i,t} \qquad (1)$$

$$D_t \equiv \sum_{i=1}^{N_t} w_{i,t} \cdot D_{i,t} \qquad (2)$$

We do not modify the original data we download from Shiller's website.[9].

3.3 Summary Statistics and Time-Series Evolution

Now we analyze the basic features of our working variables. Throughout the paper, we use `LTG`, `D`, and `PE` to designate information related to our measure of consensus earnings growth, the associated disagreement, and the price-earnings ratio, respectively. Similarly, we reserve the notation R_h to denote information related to the h-month market index return.

Table 1 reports summary statistics for the full sample. The first four rows display results for market returns, while the bottom three are dedicated to `PE` and the variables we construct from subjective beliefs. It is useful to note that these variables have different scales. This will motivate some scaling in our empirical analysis to make results comparable across horizons and regression specifications.

Table 1: Summary statistics for the whole sample (December 1981 - December 2022). SD denotes the standard deviation of each variable while we reserve the notation "Px" for the x-th quantile. The last column reports the number of observations.

	Mean	SD	P5	P25	Median	P75	P95	N
R_6	0.06	0.11	-0.12	0.00	0.06	0.13	0.23	493
R_{12}	0.13	0.17	-0.18	0.05	0.14	0.23	0.39	493
R_{36}	0.44	0.35	-0.23	0.30	0.45	0.62	1.08	493
R_{60}	0.84	0.59	-0.07	0.44	0.90	1.17	1.96	493
PE	23.64	8.30	9.61	18.09	23.95	28.38	38.41	493
LTG	13.50	2.11	11.37	12.16	12.93	13.87	18.00	493
D	3.61	0.63	2.79	3.10	3.47	4.09	4.68	493

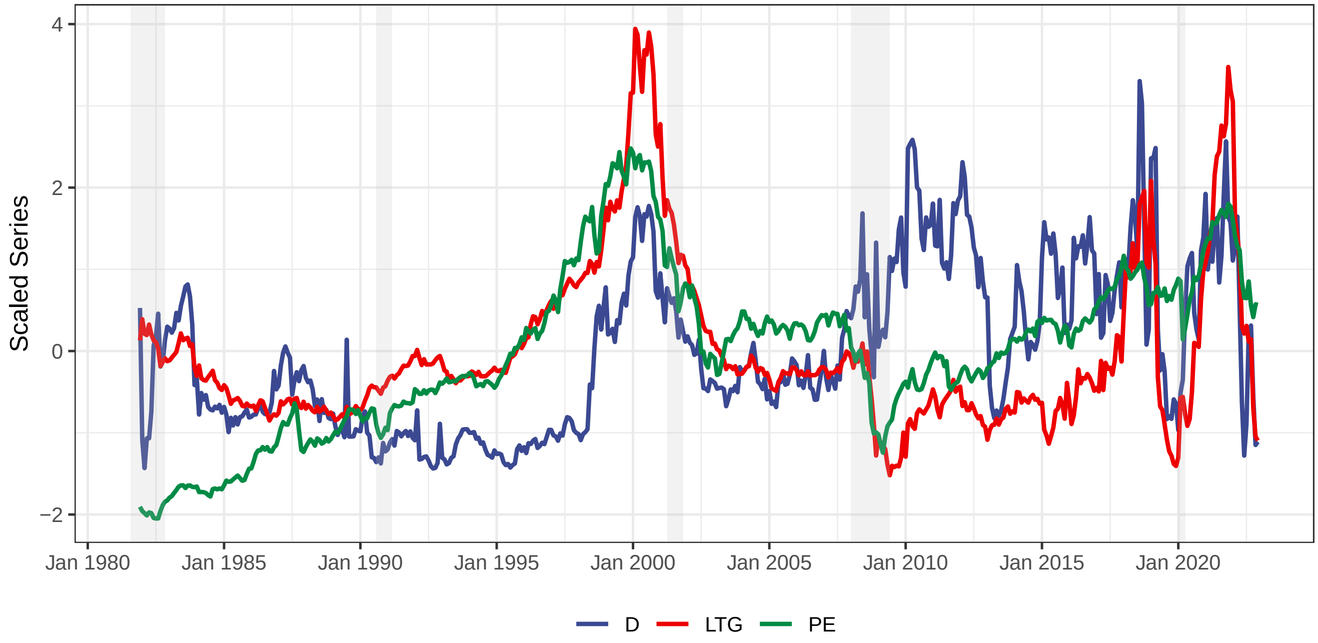
Figure 1 shows the evolution of `LTG`, `D`, and `PE` over time. We demean each variable and divide the resulting series by its standard deviation to have the three variables on the same scale.[10] The shaded bars denote NBER recessions. There are several interesting aspects of their joint evolution.

We discuss the evolution of `PE` and then focus on the subjective belief variables. There has been an almost monotonic increase since the early 1990's until its peak just before the dot-com bubble burst in early 2000. It then follows a downward trend, stabilizing its value around 2005, just before a marked dive during the global financial crisis in the late 2000s. The subsequent movement is a gradual recovery that does not seem too disturbed by the coronavirus pandemic. The very last part of the sample displays a decreasing profile for `PE`, likely driven by lower equity prices in response to the monetary cycle in the US.

[9] Available at `http://www.econ.yale.edu/~shiller/data.htm`

[10] In Appendix A, Figure A.3 plots all variables in individual panels at their original scales for convenience.

Figure 1: Time-series evolution for the consensus growth forecast `LTG`, the associated disagreement `D` and the price-over-earnings ration `PE`. All variables have been normalized to be plotted on the same scale. The sample starts in December 1981 and ends in December 2022. The shaded bars track NBER recessions.



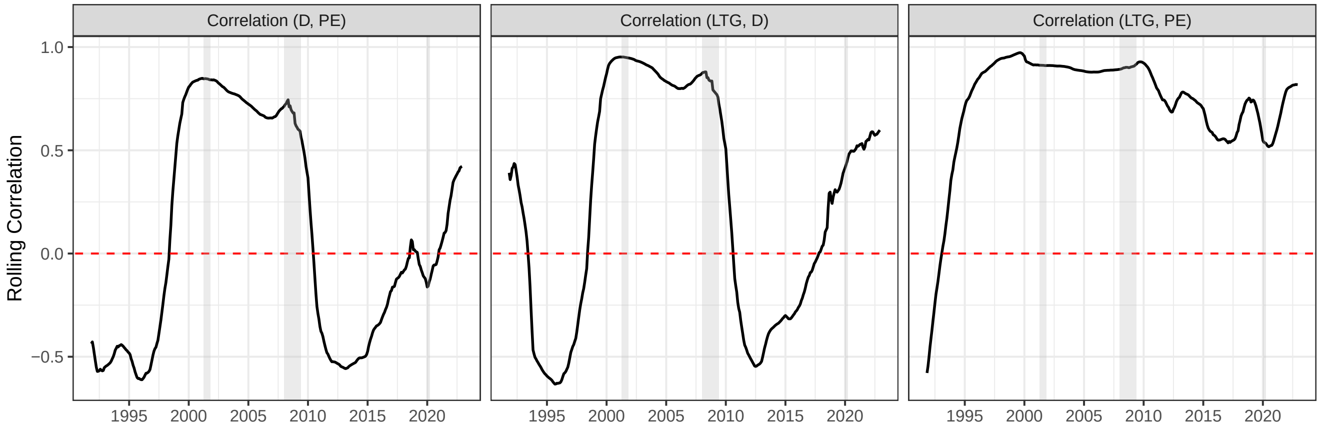
The evolution of `LTG` is similar, although with some important differences. It also peaked just before the dot-com bubble, after a decade of steady increase, followed by a fast decline up to stabilization around 2005. We also see a sharp decrease in 2008 and some recovery thereafter. On the other hand, during the first nine years of the sample, the consensus growth forecast decreased over time, while `PE` was increasing. During the very last part of the sample, `LTG` follows a jagged path with no clear trend.

Interestingly, `D` seems to follow a similar qualitative path until 2008 when we spot some stronger disconnection between `LTG` and `D`. Since then, there has been no clear trend for `D`. It starts moving in tandem with `LTG` once again towards the very last part of the sample. It is out of the scope of this paper to investigate why this disconnection happens. However, we believe we are the first to document it.

Figure 2 displays the 120-month rolling correlation between these measures. It makes our points about co-movement more formally. We index the rolling correlations by the final month included. Hence, the curves start in December 1991. The first panel studies the correlation between `D` and `PE`. It is initially negative and steadily increases, becoming strongly positive just before the dot-com burst. After 2008, the jagged path for `D` from Figure 1 is translated to a lower correlation with `PE`. The unconditional correlation between them is 0.36 over the full sample.

The second panel studies the correlation between `LTG` and `D` over time. We see a similar profile to the first panel, although the correlation starts positive. In both cases, it peaked just before the dot-com bubble burst. After 2008, the correlation changes sign for some time, which is another way of visualizing

Figure 2: We compute the 120-month rolling correlation between variables. The curves are indexed by the last month used in the computation, so they all start in December 1991. Shaded bars represent NBER recessions.



the disconnection between LTG and D after 2008. The unconditional correlation between them over the full sample is 0.30.

The last panel shows an important stylized fact. Aside from some initial rolling windows when correlation was negative, it is always positive for the (LTG, PE) pair, and strongly so. This can be seen in Figure 1 by how movements in the red line co-move with the green one. The correlation starts negative because, for the first ten years of data, the increase in PE happened at the same time when the consensus growth forecast was becoming more pessimistic. The correlation is reduced, although still positive, during the last part of the sample due to the jagged path followed by LTG, which was accompanied by a relatively smooth evolution of PE. The unconditional correlation between these two variables is 0.68.

4 In-Sample Analysis

We turn to our in-sample analysis now. To put our approach into perspective, we revisit some of the arguments previous literature has used to motivate the type of predictive regressions we present below. Seminal work from Miller (1977) and Jarrow (1980) suggested that when agents have different beliefs about the prospect of an asset's payoff stream, and at least some of them face short-selling constraints, price formation should depend on the level of disagreement. The essential argument is that pessimists would like to sell the asset to optimists, but such trade may face restrictions. In this context, Miller (1977) hypothesized that prices should increase with disagreement and that ensuing returns should decrease. A similar conclusion arises from the model in Jarrow (1980) under some conditions.[11]

Early empirical work from Chen et al. (2002) and Diether et al. (2002) found support for this predic-

[11] More recently, Atmaz and Basak (2018) revisited this question with a more sophisticated modeling framework and reached similar conclusions depending on conditions related to the parameters of their model.

tion in the cross-section of stocks. More recently, Yu (2011) tested this hypothesis for the market index, using a construction similar to ours, and found support for it once more. However, his sample stopped in 2007, right before the apparent change in behavior of `D` that can be seen in Figure 1. A natural question is how these results would change if we were to add fifteen years more of data, including two large recessions and a period of unconventional monetary policy in the middle.

Although naturally related to the literature on disagreement, an agenda on the dynamics of subjective growth expectations evolved in parallel fashion.[12] This literature has focused on stepping outside the rational expectations framework and understanding the implications. Arguably, it has devoted less attention to disagreement and focused, for example, on `LTG` as an index of optimism regarding the stock market performance both at the aggregate and firm levels.

Early empirical work from La Porta (1996) argued that these beliefs are systematically wrong due to a possible overreaction to news, leading to return predictability on the cross-section of stocks. Higher optimism would then lead to excessively high prices or, equivalently, lower returns. Bordalo et al. (2019) confirmed this claim on an extended sample. More recently, Bordalo et al. (2023) showed that `LTG` has predictive power for aggregate stock market indexes and provided a model based on belief extrapolation to justify their findings. Their framework, nonetheless, features a single agent, leaving no role for disagreement.

As demonstrated by Figures 1 and 2, the first and second moments of subjective beliefs co-move. Additionally, both co-move with the price-earnings ratio over the business cycle. Hence, the second question we investigate is whether `LTG` still has predictive power when considering the disagreement around this consensus growth forecast. We also control for a simple measure of stock market valuation in the form of `PE`. Perhaps more importantly, we repeat our analysis over three different sub-samples. The first one uses all data available to us, ranging from December 1981 to December 2022. The second one stops in December 2015, the same choice from Bordalo et al. (2023). The third one stops in December 2007, just before the apparent disconnect between `LTG` and `D`. This is the same period considered in Yu (2011).

We study predictive regressions of the following form:

$$R_{t+h|t} = \alpha + \beta_{LTG} \cdot LTG_t + \beta_D \cdot D_t + \beta_{PE} \cdot PE_t + \epsilon_{t+h} \tag{3}$$

where $R_{t+h|t}$ is the market return over h months. We focus on $h \in \{12, 36, 60\}$ for the in-sample results. When making inferences about estimated coefficients, we use an HAC-adjusted standard error estimator along the lines of Newey and West (1987), with h lags. Since these variables have different scales, as shown in Table 1, we demean each and divide them by their respective standard deviations. The coefficients then measure the effect on the dependent variable due to an increase of one standard deviation on each variable on the right-hand side, measured in standard deviations of the left-hand side variable.

[12] See, for example, Nagel and Xu (2023) for a recent account and the references therein.

4.1 Main Results

Table 2 summarizes our in-sample results. Each of the three panels concentrates on a different subsample. Each panel reports estimates for 18 regressions since there are six specifications for each horizon h. Regressions with the same dependent variable are grouped. In all cases, we report t-statistics in the parentheses. Stars measure statistical significance at 10%, 5%, and 1%, respectively.

In each of these groups, the first specification studies the effect of including only `LTG` to (3). The second specification includes only `D`. The third specification uses both the consensus growth forecast and the disagreement measure to predict returns. Its main goal is trying to disentangle the effects of these two variables. The fourth, fifth, and sixth specifications follow the same pattern, but all of them control for the price-earnings ratio as well. They are designed to assess the effects of the variables constructed from subjective beliefs on market returns *above and beyond* what can already be accounted for by a simple measure of stock market overvaluation.[13] The two different research agendas mentioned above would predict that coefficients on `LTG` and on `D` are negative and significant.

[13] See **?** for this interpretation regarding the price-earnings ratio.

Table 2: Each panel displays estimates for regression coefficients from (3). All variables are standardized before the regressions. Dependent variables are the 12-month, 36-month, and 60-month market index returns. Different panels cover different sub-samples. We report *t*-statistics inside the parentheses. Stars denote significance at 10%, 5%, and 1% confidence levels. Standard errors are computed with a HAC-adjusted estimator following Newey and West (1987) with *h* lags. The last rows of each table report the number of observations in each regression and the associated adjusted R^2. The last two rows report the number of observations for each regression and the adjusted R^2.

Panel A: Full-Sample (1981-2022)

	R_{12}						R_{36}						R_{60}					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
LTG	-0.34***		-0.27**	-0.19		-0.17	-0.37		-0.33	-0.01		-0.01	-0.45***		-0.42***	0.02		0.01
	(-2.97)		(-2.30)	(-0.90)		(-0.88)	(-1.56)		(-1.49)	(-0.03)		(-0.03)	(-5.71)		(-4.76)	(0.13)		(0.06)
D		-0.32**	-0.24**		-0.22*	-0.21*		-0.27	-0.20		-0.10	-0.10		-0.28	-0.22		-0.08	-0.08
		(-2.14)	(-2.07)		(-1.67)	(-1.75)		(-1.16)	(-0.91)		(-0.42)	(-0.42)		(-1.18)	(-0.94)		(-0.29)	(-0.27)
PE				-0.23	-0.28	-0.16				-0.55***	-0.52**	-0.52***				-0.70***	-0.66***	-0.66**
				(-0.85)	(-1.55)	(-0.66)				(-3.85)	(-2.29)	(-2.87)				(-3.12)	(-4.36)	(-2.34)
N	481	481	481	481	481	481	457	457	457	457	457	457	433	433	433	433	433	433
R^2	0.117	0.103	0.170	0.145	0.168	0.182	0.129	0.068	0.163	0.288	0.298	0.296	0.184	0.067	0.225	0.421	0.426	0.425

Panel B: Same sample as Bordalo et al. (2024) (1981-2015)

	R_{12}						R_{36}						R_{60}					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
LTG	-0.32**		-0.26**	-0.10		-0.09	-0.47***		-0.41***	-0.02		-0.01	-0.47***		-0.39***	0.17		0.22
	(-2.40)		(-2.12)	(-0.38)		(-0.40)	(-3.19)		(-3.47)	(-0.09)		(-0.04)	(-4.15)		(-2.89)	(0.51)		(0.60)
D		-0.35*	-0.30*		-0.27	-0.27		-0.36	-0.28		-0.22	-0.22		-0.38	-0.26		-0.19	-0.22
		(-1.70)	(-1.94)		(-1.52)	(-1.61)		(-1.03)	(-1.21)		(-0.74)	(-0.74)		(-1.40)	(-1.19)		(-0.68)	(-0.71)
PE				-0.31	-0.31	-0.25				-0.59**	-0.55***	-0.54*				-0.84***	-0.66***	-0.82***
				(-0.87)	(-1.65)	(-0.81)				(-2.10)	(-3.00)	(-1.70)				(-2.84)	(-4.12)	(-2.62)
N	397	397	397	397	397	397	373	373	373	373	373	373	349	349	349	349	349	349
R^2	0.103	0.121	0.187	0.147	0.211	0.213	0.198	0.120	0.265	0.339	0.382	0.380	0.199	0.130	0.248	0.465	0.484	0.501

Panel C: Same sample as Yu (2011) (1981-2007)

	R_{12}						R_{36}						R_{60}					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)
LTG	-0.41**		-0.07	-0.27		0.38	-0.65***		-0.33	-0.42		0.47**	-0.63***		-0.45	-0.13		0.68***
	(-2.52)		(-0.43)	(-0.56)		(1.18)	(-3.59)		(-1.27)	(-1.09)		(2.13)	(-3.06)		(-1.57)	(-0.20)		(4.34)
D		-0.53***	-0.48**		-0.44***	-0.61***		-0.68***	-0.45		-0.51*	-0.72**		-0.57***	-0.26		-0.35	-0.66***
		(-3.70)	(-2.56)		(-2.72)	(-3.00)		(-3.30)	(-1.38)		(-1.88)	(-2.22)		(-3.70)	(-0.96)		(-1.39)	(-2.80)
PE				-0.18	-0.21	-0.44				-0.28	-0.41**	-0.72***				-0.61	-0.58**	-1.03***
				(-0.30)	(-0.99)	(-1.33)				(-0.73)	(-2.14)	(-4.83)				(-0.98)	(-2.54)	(-4.94)
N	277	277	277	277	277	277	253	253	253	253	253	253	229	229	229	229	229	229
R^2	0.170	0.281	0.281	0.178	0.313	0.337	0.378	0.416	0.460	0.396	0.538	0.562	0.326	0.267	0.351	0.419	0.504	0.557

4.1.1 Full Sample (1981-2022)

We concentrate on Panel A. At the 12-month horizon, the coefficients on LTG and D have the sign one would expect from reading previous literature. Increases in these variables are associated with lower future returns. However, the effect through LTG is insignificant after controlling for the price-earnings ratio. The effect of D remains significant, although only at the 10% level, with a t-statistic of -1.75.

At the 36-month horizon, the individual coefficients have the expected sign, but none are significant. That happens both with and without the control provided by the price-earnings ratio. Only the coefficients on the price-earnings ratio are significant, with the correct expected negative sign. The inclusion of neither LTG nor D boosts the in-sample R^2 in a meaningful way.

The 60-month horizon follows the same pattern. Although LTG is still negative and significant after the inclusion of D , controlling for PE even makes the coefficient change sign, although we cannot reject it is different than zero. We also notice that the inclusion of PE makes the R^2 increase more than twofold, indicating a much stronger in-sample predictive power of PE than relative to the two survey belief moments.

4.1.2 The Sample Up to December 2015

We now shift our focus to Panel B. This panel mimics the sample from Bordalo et al. (2023), stopping in December 2015. We start emphasizing specifications 1, 7, and 13. These are the basic specifications reported in Bordalo et al. (2023), and we report that we can replicate their results qualitatively and almost match their estimates. Our estimates for β_{LTG} range from -0.32 to -0.47, while theirs go from -0.23 to -0.43. We also report similar R^2 measures, validating our construction.

In this sample, D has a significant coefficient only at the 12-month horizon, although it still has the expected sign over the other two horizons. When we include both LTG and D , without controlling for PE , the estimates for β_{LTG} generally decrease in absolute value, but such movement is not strong enough to hinder them insignificant.

The inclusion of PE in these regressions changes the results, however. When controlling for PE , no coefficient was significant at the 12-month horizon. When considering the other two horizons, only PE itself is significant. The coefficient on D remains negative across specifications, but we can never reject it is different than zero. At the 36-month horizon, the estimates for β_{LTG} are essentially zero, while they become positive at the 60-month horizon. Although we confirm the initial finding from Bordalo et al. (2023), we show that controlling for the cyclically adjusted price-earnings ratio makes estimates insignificant or even flips their sign.[14]

[14] In their analysis, Bordalo et al. (2023) never included D in their regressions, which makes sense since their empirical setup reflects the single-agent model they have. They do include a PE variable, which is not cyclically adjusted and is less standard

4.1.3 The Pre-Crisis Sample (1981-2007)

Finally, we analyze Panel C. The sample that ends in 2007 is special in two ways. First, it is the same time frame used by Yu (2011). This will exclude the Global Financial Crisis and the coronavirus pandemic from the analysis. Second, this is the period of highest co-movement between LTG and D , as seen in Figure 2.

In his empirical work, Yu (2011) studied specifications analogous to columns 6, 12, and 18. He included both the consensus growth forecast and PE in his analysis. We start stressing that we can replicate his main findings. The coefficient on D is negative and statistically significant even if we control for the two other variables. It is also significant when only D is included in the regressions, as demonstrated by columns 2, 8, and 14. The magnitude of these estimates is also similar across different horizons.[15]

Nonetheless, we have a more surprising result for $\widehat{\beta}_{LTG}$ in this sample. Even though it remains negative and significant when included alone in these regressions, its sign becomes positive across the three horizons when all variables are included. In fact, for the 36-month and the 60-month horizons, it is positive and statistically different than zero, contradicting the findings from a later sample from Bordalo et al. (2023). When LTG is included alongside D, its coefficient is still negative but never statistically different than zero.

4.1.4 Summary of Results

We can summarize our in-sample results in the following way:

1. The strength of the association between LTG and market index returns highly depends on whether the price-earnings ratio is included or not in these regressions. Even when it has the predicted negative sign, estimates are not statistically different than zero if PE is included. Moreover, when both PE and D are included, estimates for β_{LTG} are generally positive. They are, sometimes, statistically different than zero - but this is the opposite sign from the belief overreaction mechanism advocated in extant literature.

2. The robust negative effects on market returns documented by Yu (2011) can be replicated, but only on his sample. That effect seems to be much weaker when we consider an extended sample.

3. Controlling for D does not typically hinder the estimates on β_{LTG} insignificant, except for the earlier sample that stops in December 2007. However, it does decrease the absolute value of estimates across samples and horizons.

in the forecasting literature. We compare these two measures and some minor differences in implementation in Appendix **??**.

[15] Yu (2011) did not study the 60-month horizon originally. Therefore, we can provide a new finding and extend his result to a longer forecasting horizon.

4.2 Is PE a ”bad control”?

After inspecting our in-sample results, one could ask: should we include PE in these regressions after all? One reasonable concern is that PE might only mediate the underlying effect on returns of either the consensus growth forecast or the associated disagreement. More concretely, one could be worried that PE is solely a noisy proxy for either of the two variables constructed from subjective beliefs and brings no additional information. In that case, adding this control to regressions could crowd out the effect of either LTG or D and make us incapable of rejecting the null of zero effect directly through subjective beliefs, even when it is false. Borrowing language from the causal inference literature, should we be concerned with PE being a “bad control”?[16] This is an important issue both for our analysis and for the literature in general since regressions like the ones presented in Table 2 are usually taken as evidence in favor (or against) different theoretical models.

We believe that both Figure 1 and Figure 2 provide some evidence that the answer is “no”. An observer with only the first ten years of our data would likely conclude that the consensus growth forecast and PE move in different directions, at least at lower frequencies. She would draw the same conclusion regarding the disagreement measure as well. That is seen by the negative correlation between PE and these variables at the beginning of our sample. Another observer would conclude, however, that these measures are positively correlated strongly if she were given data from 1990 until the financial crisis - an entirely different assessment. A third observer, inspecting the 2010s, would lean on saying that they are positively correlated, although not as strongly as during the preceding twenty years. In summary, there are long-lasting lower frequency movements that could not take place if PE were a simple proxy of any of the other measures.

Table 3 quantifies the degree of linear association between PE and the measures constructed from subjective beliefs using our full sample. We consider the following contemporaneous specifications, both in levels and in 12-month differences:

$$PE_t = \alpha + \gamma \cdot LTG_t + \delta \cdot D_t + u_t \tag{4}$$

$$\Delta PE_t = \tilde{\alpha} + \tilde{\gamma} \cdot \Delta LTG_t + \tilde{\delta} \cdot \Delta D_t + \tilde{u}_t \tag{5}$$

where, for any variable x_t, we define $\Delta x_t \equiv x_t - x_{t-12}$. The first specification assesses the degree of linear relationship between PE and the variables constructed from subjective beliefs. The second specification, concentrating on 12-month changes, checks whether swings in LTG and D are linearly transmitted to PE. This last specification is also useful because these time series are fairly persistent at the monthly horizon, and one could be concerned about any spurious findings due to persistence.[17] Before running each

[16] See Angrist and Pischke (2008) and Cinelli et al. (2022) for treatments of bad controls in the standard context of causal inference with cross-sectional data.

[17] Figure A.1 in our Appendix A reports the autocorrelation functions of these variables at the monthly frequency, up to 36 months ahead. By far, PE is the most persistent, almost integrated. Figure A.2, also in Appendix A, shows the time-series evolution of these 12-month differences, all scaled before plotting to be shown in the same axis.

Table 3: The table reports coefficient estimates for the regressions in (4) and (5). We define $\Delta x_t \equiv x_t - x_{t-12}$ for any variable x_t. We display t-statistics in the parentheses. Standard errors are computed using the HAC estimator from Newey and West (1987). The sample used ranges from December 1981 to December 2022. Stars denote significance at 10%, 5%, and 1% confidence levels.

	PE_t				ΔPE_t		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
PE_{t-12}	0.85***						
	(3.64)						
LTG_t		0.68***		0.62***			
		(4.88)		(5.32)			
D_t			0.37*	0.18			
			(1.91)	(1.37)			
ΔLTG_t					0.63***		0.64***
					(4.34)		(4.00)
ΔD_t						0.39**	-0.02
						(2.45)	(-0.15)
N	481	493	493	493	481	481	481
R^2	0.785	0.455	0.132	0.483	0.400	0.151	0.399

regression, we also normalize all variables. The associated t-statistics are reported inside the parentheses. The first column reports the regression of PE_t on PE_{t-12} as another way to document the persistence in PE_t and as a reference for the R^2 levels. The point estimate of 0.85 is highly significant. The R^2 is above 0.78, even though we are conditioning on the price-earnings ratio lagged by a full year.

The second regression includes only LTG . Although the estimate is positive and significant at usual levels, the R^2 is around 45%, even though both PE and LTG are aligned in terms of time indexes, showing that the contemporaneous association is not that strong. The third regression uses only D, delivering a low R^2 and an estimate that is not statistically significant at the 5% level anymore. The fourth column uses both LTG and D and shows that estimates barely change for LTG compared to the second column. The coefficient on D is not significant anymore, and the R^2 of this regression is essentially the same as the one from column (2). The specifications that use the 12-month differences tell the same story. While these results do not rule out that PE is a noisy proxy of the other two variables, they do imply that it is *at most* a very noisy one.

Still under the hypothesis that PE is a noisy proxy of either LTG or D, we would expect that, in terms of forecasting power, conditioning only on PE would generate worse forecasts than using the other variables. After all, predicting with a cleaner signal from either LTG or D should be better than predicting with that signal plus noise. Our next findings provide evidence that this is not the case, which goes against the idea that PE is just a proxy for other variables.

5 Out-of-Sample Analysis

Our previous analyses replicated some of the findings from previous literature and showed that they are arguably dependent on the sample used and the controls included. Nonetheless, all these results happen in sample. We now develop an out-of-sample forecasting exercise to more closely examine whether these different variables have forecasting power. We see this exercise as a complement to our previous analysis. Importantly, this type of exercise will make clear *when* each variable (or combination of variables) performs the best. We also inspect the coefficient estimates from (3) over time.

5.1 Forecasting Methodology

We fixed the sample starting date to December 1981, and we consider a recursive forecasting scheme. This is traditional in the macro-finance forecasting literature.[18] It implies that, for each point in time, we estimate the specification in (3) and predict h periods ahead. We keep track of that forecast and advance one month in time. Then, we estimate the specification (3) and predict once again h periods ahead. We continue this until we exhaust our data.

These forecasts, done with data up to time $t-h$, generate forecast errors that are known at time t. We then use the R^2_{OOS} measure from Campbell and Thompson (2008) to assess the quality of these forecasts:

$$R^2_{OOS}(t_0,h) = 1 - \frac{\sum_{t=t_0}^{T}\left(R_{t|t-h} - \widehat{R}_{t|t-h}\right)^2}{\sum_{t=t_0}^{T}\left(R_{t|t-h} - \overline{R}_{t|t-h}\right)^2} \tag{6}$$

where t_0 indexes the first out-of-sample date, $\widehat{R}_{t|t-h}$ is a forecast of the market index return made at $t-h$ and $\overline{R}_{t|t-h}$ is a benchmark computed with information available up to $t-h$. In this setting, a natural benchmark is a model that only includes a constant. Hence, its forecasts equal the historical average of market returns for each considered horizon. This measure is also common in the forecasting literature, comparing the forecasting mean-squared error from both models. We notice that we index the out-of-sample R^2 by t_0, because it differs depending on where one starts the out-of-sample period, which we discuss below.

When this measure is above zero, it implies that the analyzed model can beat the benchmark. On the other hand, it is negative when the model fails against the proposed baseline forecasting method. We also highlight that the out-of-sample R^2 is bounded above by 1 but has no lower bound. It also does not

[18] Huang et al. (2021) and Goyal et al. (2021) also use recursive forecasting designs in a context similar to ours, as this is very common in the forecasting literature. For example, Gu et al. (2020) uses a recursive forecasting scheme to predict equity returns, while Bianchi et al. (2020) and Freire and Riva (2023) use this recursive estimation to forecast sovereign bond risk premia. For a general review, see Clark and McCracken (2013).

have a measurement unit since it is an index that can assess the quality of different forecasting methods based on a ratio of loss functions.

Aside from the benchmark, we consider seven models in total. Three models include only each of the individual variables LTG , D , and PE ; then, three models feature each possible combination of two variables; last, a model includes all three. They are all compared against the historical mean of returns.

Our approach is related to, albeit different from, Huang et al. (2021) and Goyal et al. (2021). The former considers out-of-sample predictability of market index returns based on different measures of disagreement, not only the subjective beliefs measure D we also use. However, they do not look into the predictability of first-moment measures such as the consensus growth forecast. The latter analyzes how D and several other measures predict market returns. However, they do not have variables such as LTG and never study regressions with two or more regressors. In this sense, we find additional support for some of their results and see our approach as complementary to theirs.

5.2 A Snapshot of Performance

In this section, we report a snapshot of each model's performance and leave the variation of relative performances over time for the next sections. As is common in forecasting exercises, choosing the out-of-sample period is not trivial. Given a fixed amount of data, a long out-of-sample period implies a richer testing site for each model, with the caveat that each of these specifications is estimated more imprecisely at the earlier part of the sample since there is less data to run regressions. Estimation uncertainty might then drive results. In contrast, a very short out-of-sample period allows for comparing models that have been more precisely estimated but can only be tested on a handful of data points.

We believe including the dot-com burst in any *in-sample* period is important. Otherwise, endowed only with the data from Figure 1, any model would take PE and LTG almost as ever-growing variables, at least at lower frequencies. Ideally, we want to include at least one large stock market crash in any in-sample period. After the stock market crash in 2000, a recession ensued. Following the official timing from the NBER, the American economy left the recessive period in November 2001. Since the longest forecasting horizon we study is five years, we would need to start the out-of-sample period at least in late 2006 to ensure all forecasts include these events in their estimation (or "training", borrowing language from the Machine Learning literature) sample.

However, an analyst in late 2001 would not know how firms navigated the recession since balance-sheet data usually takes a few months to be released, and there are different fiscal year-end dates. Considering all this and trying to maximize our out-of-sample period length, we choose July 2007 as the first out-of-sample date. In that way, even for the longest horizon, forecasts are produced with the knowledge of the crash and preliminary information on how firms did during the recession, even allowing for delays in balance sheet data releases for many firms. This sample start also implies that we consider

Table 4: We report $R^2_{OOS}(t_0, h)$ for different models and horizons, taking $t_0 =$ (July, 2007). We also report p-values in the parentheses that test the one-sided null hypothesis that each model is not better than a constant-only model. These are computed using the adjusted statistic from Clark and West (2007). We do not report p-values when the models fail to beat the benchmark.

	Forecast Horizon h (in months)			
Regressors	6	12	36	60
PE	0.04	0.08	0.12	0.36
	(0.002)	(0.000)	(0.000)	(0.000)
LTG	0.01	0.04	-0.51	-0.22
	(0.019)	(0.001)		
D	-0.09	-0.32	-1.51	-0.92
PE + LTG	0.03	0.06	-0.17	0.21
	(0.011)	(0.001)		(0.000)
PE + D	-0.06	-0.19	-0.77	0.08
				(0.000)
LTG + D	-0.09	-0.28	-1.35	-0.38
PE + LTG + D	-0.13	-0.48	-2.18	-1.98

the performance of different models during the global financial crisis, the recovery period, and the first years after the stock market crash in 2020 - a challenging range of different economic conditions.

Table 4 displays $R^2_{OOS}(t_0)$ with $t_0 =$ (July, 2007) for a range of models and horizons. Positive values imply that a certain model beats the historical average in terms of forecasting power. Negative values imply otherwise. Whenever a model beats the benchmark, we assess the statistical significance of this improvement using the adjusted statistic from Clark and West (2007) since these models are nested. We use recursive windows, which implies that the traditional approach from Diebold and Mariano (1995) does not provide correct inference and is severely undersized.[19] We report p-values inside parentheses, which test the one-sided null hypothesis that $R^2_{OOS}(t_0) \leq 0$. We only report such values when the out-of-sample R^2 is positive.

The first row displays results for a model that uses the price-earnings ratio to try to beat the benchmark. Interestingly, we find that `PE` beats the benchmark at all horizons. Such improvement is statistically significant at the usual levels. Compared to a naive guess based on the historical average, the performance of this model is especially strong at the 60-month horizon.

The second row considers a model that uses `LTG` to beat the constant-only model. There is a modest improvement at the 6-month horizon, which is not statistically significant at the 1% level, and a slightly stronger improvement at the 12-month horizon. However, this model performs much worse

[19] See Clark and McCracken (2013) for a review about these issues. Huang et al. (2021) and Goyal et al. (2021) follow the same approach as ours.

than the benchmark for the two longer horizons. This is perhaps surprising because I/B/E/S explicitly elicits beliefs regarding growth prospects for a more extended timeline. Nevertheless, such information, as reported in Table 4, seems useful to forecast market returns over shorter horizons. The third row concentrates on the forecasting power of D when used alone in these regressions. This model cannot beat the benchmark for any of the horizons we considered. We find strong evidence against using this disagreement measure to predict aggregate stock market returns.

The fourth row considers a model combining information from LTG and from PE to create forecasts. It beats the benchmark for all horizons except for the 36-month one. However, the improvement at the 6-month horizon is not statistically significant at the 1% level. We also note that this richer model cannot generate out-of-sample R^2 values higher than the one generated only with PE . Even though the latter specification nests the former, it has to estimate one more parameter at each point in time, which increases estimation uncertainty.

The fifth specification combines information both from PE and D . This model performs uniformly worse than the model with PE only, as we have seen that the information in D is not useful for the prediction task at hand. Hence, this model pays the price of one more parameter to be estimated without the benefit of extracting useful information. Conditioning on D does more harm than good.

The sixth specification excludes PE from the conditioning set and combines information from both LTG and D , focusing only on the two moments elicited from analyst surveys. We find no signs of predictability in that case. That is also the case for the last specification, using the three variables. Once more, even though it nests the first specification with only PE , it also stresses more the data due to the estimation of a larger system. Starting the out-of-sample period in July 2007, the best specification was an univariate regression using the price-earnings ratio as a single index return predictor.

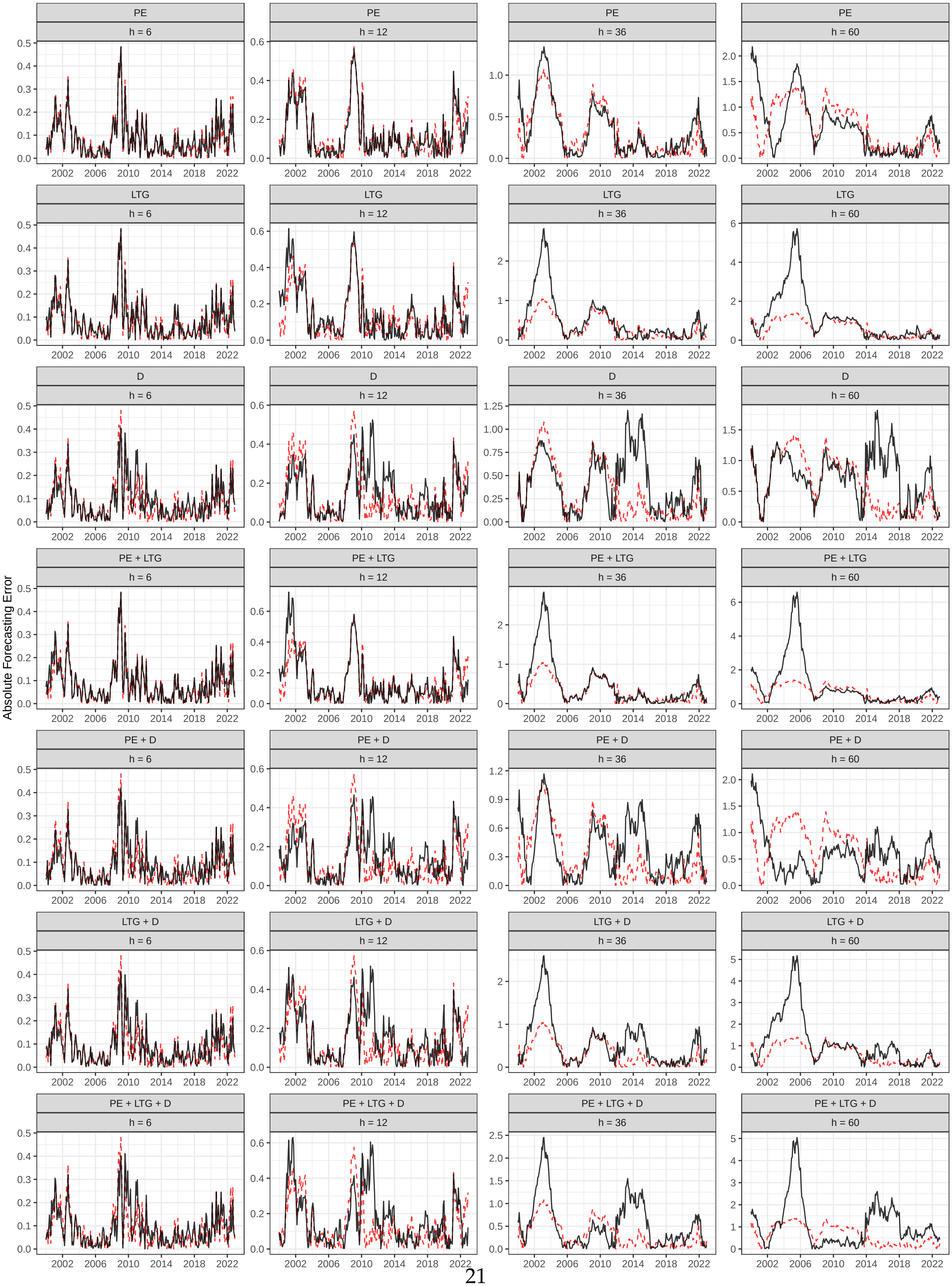
5.3 Evolution of Forecast Errors

We now return to the problem of choosing the starting point for the out-of-sample period. One may not agree with our principle of choosing July 2007 and may prefer an entirely different starting point. How can one be sure that the results from the previous section are free from "p-hacking", i.e., taking conclusions from data based on specific slices of our observations? This section shows the evolution of both absolute forecasting errors and $R^2_{OOS}(t_0)$ as a function of t_0 in an attempt to tie our hands as much as we can.

Starting in 2000 and ending in 2022, Figure 3 displays the evolution of absolute forecast errors. Each column concentrates on a different forecasting horizon. Each row is dedicated to a different forecasting model. There are 28 panels in total. For any model M, for a fixed horizon h, the forecasting error at time t is defined as

$$e_{t,h}^{(M)} \equiv \widehat{R}_{t|t-h}^{(M)} - R_{t|t-h} \tag{7}$$

Figure 3: Each plot depicts the evolution of $|e_{t,h}^{(M)}|$ for different models (rows) and different horizons (columns). The solid black line represents the absolute forecast error. The red dashed line represents the error from the constant-only benchmark model. The index convention is such that the points are plotted as they are realized. See the main text for details.



where $\widehat{R}^{(M)}_{t|t-h}$ is the forecast for market returns from model M, done with information up to time $t-h$. This random variable is only realized at time t. The different panels that compose Figure 3 display the evolution of $\left|e^{(M)}_{t,h}\right|$ as a function of t, plotted as a solid black line. The red dashed line represents the absolute forecast error for the benchmark, i.e., the constant-only model. The red line is the same across different rows since each column has a fixed forecasting horizon. The unit of the vertical axis is the same as market returns. The proposed models do better than the benchmark whenever the black line is below the red one. We intentionally started these plots before July 2007, allowing the reader to see what happens both before and after our chosen start date.

For the first column of panels, referring to the 6-month horizon, we see that the black and red lines move together. There is no single event or moment in which models get suddenly better (or worse) than the benchmark. As seen in Table 4, performance is relatively similar to the benchmark because of that. For the specification with the three variables, in the last row, we do see moments between 2010 and 2014 and then between 2019 and 2021, for which the benchmark did somewhat better than the full model. This gets reflected in a negative R^2 of -0.13 with we start in July 2007, for instance.

The second column of panels from Figure 3 starts shedding light on the poor performance of `D` as a predictor. All models that include that variable (third, fifth, sixth, and seventh columns) display large forecast errors between 2010 and 2014. We note that these forecast errors are due to predictions made with information available up to 2009 if we concentrate on data plotted in 2010, for example. This coincides in timing with the stronger disconnection between `D` and the two other variables documented in Figure 1. If this disconnection is due to some structural change, which is not modeled here, it is natural that it would generate large forecast errors since the model in (3) would have been estimated under a different regime.

The third column confirms our observation about `D` and brings an interesting perspective regarding the dot-com bubble. First, we notice that large errors take place roughly between late 2011 and 2015. For these plots, a forecast error realized, for example, in late 2011 was done with data up to late 2008, around the stronger disconnection between `D` and two other series. The location of these absolute error peaks gets shifted forward in time because we index these curves by the time they become realized, not the time at which forecasts are done. In any case, the same peaks that appeared with $h = 12$ now appear with $h = 36$. They are also present for $h = 60$ as long as `D` is included in the regression.

Second, we notice that models conditioning either on `PE` or `LTG` display absolute forecast error peaks in the early 2000s. That is due to the dot-com bubble burst. An observer with data available only up to early 2000 and making forecasts conditioning on these two variables, as an example, would have predicted high 36-month returns ahead. After all, the stock market was going up in the early 1990s, hand in hand with both `PE` and `LTG` - which were at their highest levels in that period. This observer would regret his forecast only when it is then revealed, around 2003. The same peaks appear when we concentrate on the last column of panels, which uses $h = 60$, taking place around 2005. Once again, the peaks are shifted forward in time due to our indexing conventions. When making forecasts five years ahead, the optimistic observer from late 2000 would regret her predictions only in 2005. The estimates $\widehat{\beta}_{LTG}$ were indeed positive for this earlier part of our data (as we investigate in greater detail below). When combined with values for `LTG` at their historical peak, this fact generated overly optimistic forecasts for market returns.

Figure A.4 in Appendix A supports the above interpretation. It plots the signed forecast error across models and horizons as a function of time. If we take the absolute value of the plotted curves, we recover Figure 3. The dotted line represents a reference at zero. Positive values denote forecast errors that are indicative of excessive optimism. One can see that models conditioning on either `D` or `PE` were caught off-guard by the dot-com burst because the stock market returns were expected to be much higher than they were.

Figure 3 also helps understand why the specification with the three variables fared so badly, even though it nests all models considered here. In some sense, it combines the worst of both cases described above: the conditioning on `PE` and `LTG` led to excessively high predictions in the earlier part of the sample, and the conditioning on `D` implied large errors after 2008 when some deeper modification on the dynamics of `D` might have happened.

5.3.1 Evolution of R^2_{OOS}

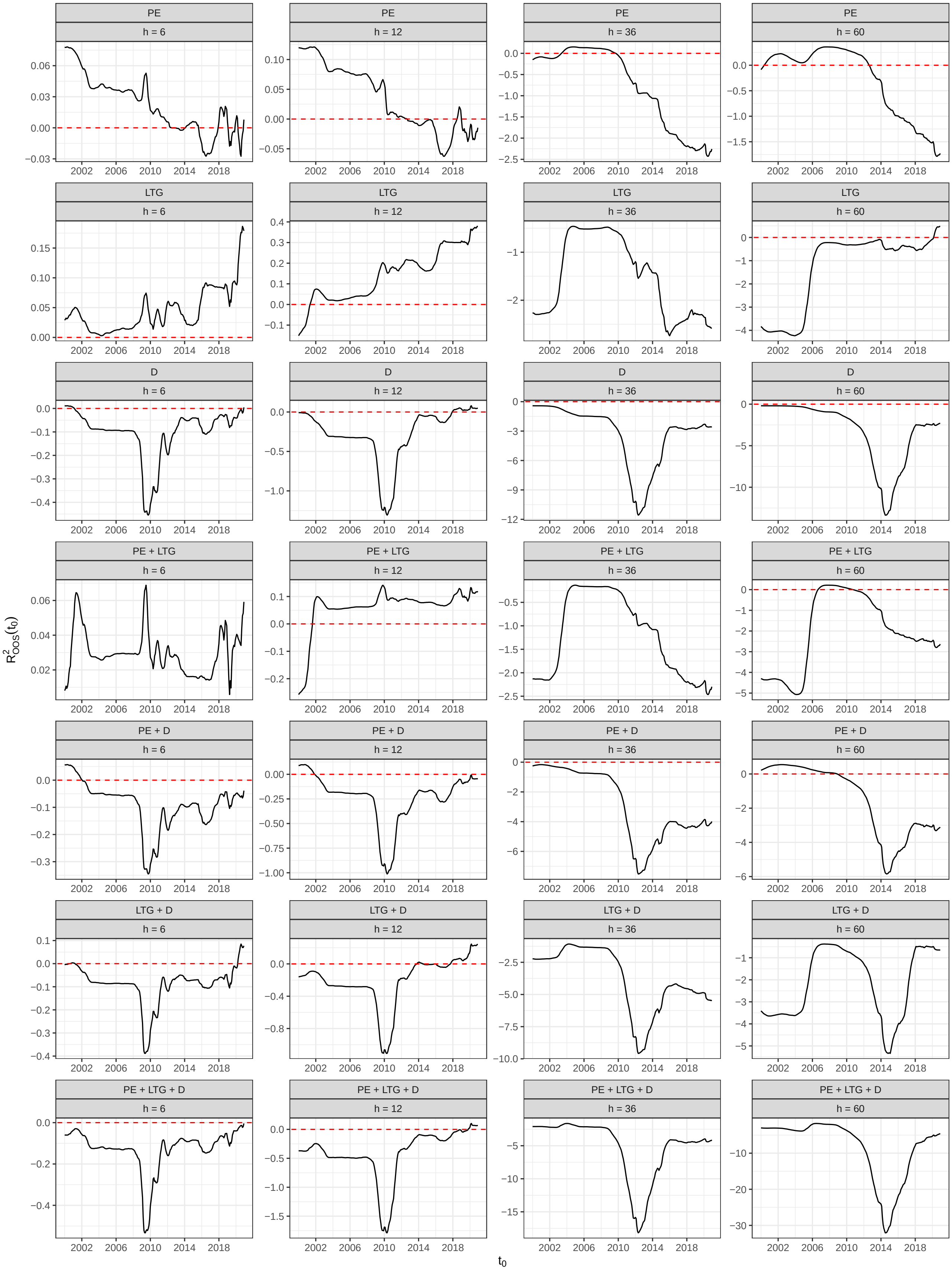
Another way of assessing the sensitivity of different models to the out-of-sample period is through the evolution of $R^2_{OOS}(t_0)$. Figure 4 follows the same scheme of plotting from the previous figure: each column is devoted to a forecasting horizon h while different models are displayed in different rows. For each point in time t we plot $R^2_{OOS}(t)$, which represents the number we would have found in a table like Table 4 if we were to start our out-of-sample period at t. In fact, at $t =$ (July 2007) these curves assume exactly the values displayed in that table.[20]

A reader who disagrees with our choice of out-of-sample period to present a snapshot of results can pick her favorite starting point and look up R^2 values in Figure 4. However, caution is needed as we approach the end of these curves. These last points are computed using far fewer observations, making the out-of-sample R^2 a noisier performance index as we move forward in the horizontal axis. We still believe this picture is useful as a call to caution when picking out-of-sample periods.

There are a few interesting lessons from Figure 4, providing support for our earlier interpretations. First, we note that all plots displaying models containing `D` have a type of "valley" of predictability, i.e., a sharp drop in out-of-sample R^2. As we have emphasized, the timing coincides with the moment in which `D` started displaying a more jagged path over time. As we move forward and start the sample later,

[20] We don't report p-values in the interest of clarity here. Most of the paths are frequently negative on most dates, anyway. Further analyses regarding p-values are available upon request.

Figure 4: We plot the evolution of $R^2_{OOS}(t_0, h)$ over time for different models (rows) and different forecasting horizons (columns). Positive values mean that a model would have been able to beat the benchmark if we had started the out-of-sample period at t_0. Negative values imply otherwise.



effectively excluding the earlier regime from the out-of-sample period, we see that the out-of-sample R^2 improves. Nonetheless, we rarely found a starting date in which models including D would consistently beat the benchmark. For instance, the third row displays this phenomenon for all horizons considered. Even if we tried our best to p-hack results, we would never find an out-of-sample period capable of generating positive R^2 for that model.

The inspection of the shorter horizons brings the second pattern to notice. The first row shows that the predictive power of PE alone decreases if we choose starting points later in the sample. Figure A.4 reveals that this happens because this model overestimated stock market performance slightly less than the sample average during the early 2000s. As we move forward the starting point of our out-of-sample period, these mistakes are no longer relevant for the computation of the performance metric, which then drifts down.

The exact reverse happens with LTG . Using later starting points, we would find that LTG is a stronger solo predictor. This is because this model had some out-of-sample overestimates in the early 2000s that exceeded the mistakes made by the historical average. The performance metric increases as we discard these earlier observations from the out-of-sample period moving the starting point further in time. When we combine both measures, we see no apparent drift in the performance metric, aside from an initial sharp increase of $h = 12$. To some extent, at these shorter horizons, the bivariate model showed it could use the best of both signals.

The last insight comes from $h = 36$ and $h = 60$. Results are dismal across columns and rows, possibly except for the model that uses only PE (first row). For all other models, once more, even if we tried our best to p-hack results, we would almost always find negative values for R^2_{OOS}. If we start the out-of-sample period in 2014, however, it holds without qualification that these profiles of out-of-sample R^2 are uniformly negative.

We finish this section by summarizing our findings regarding out-of-sample predictability with three major lessons:

1. LTG might have some predictive power for market returns, but only at shorter horizons. This is somewhat surprising because this measure is meant to gauge longer-term expectations;
2. It is unclear, however, if LTG brings additional forecasting power not already provided by a traditional financial multiple such as PE . If anything, the predictability through LTG appears only at more recent out-of-sample periods;
3. There is no point in using D as a predictor of stock market returns, regardless of the horizon or out-of-sample period.

5.4 Regression Coefficients

Our final analysis concentrates on the coefficient estimates from (3). Moving forward with our recursive forecasting scheme, we collect time series of coefficient estimates and associated standard errors.[21] Ideally, we would expect estimated coefficients to remain stable over time if the model is correctly specified. Additionally, we would also expect confidence intervals to shrink as we estimate the same regression with an increasing number of data points. In the interest of brevity, we focus only on the estimates of $\widehat{\beta}_{LTG}$ and $\widehat{\beta}_{D}$. We analyze how these estimates vary over time for each forecasting horizon and specification.

Figure 5 reports our results in three panels. The columns in each of these panels display results for different forecasting horizons. The first row is always dedicated to $\widehat{\beta}_{LTG}$, while the second one shows results for $\widehat{\beta}_{D}$. The first panel focuses on specifications where `LTG` and `D` were the only conditioning variables. The second panel reports coefficient estimates when both variables are included in the regression model. The last panel shows estimates when `PE` is included as well. The solid line is the time series of estimated coefficients in all subplots. We index these estimates according to the target date to keep our convention consistent. The curves plotted at date t represent the coefficients used to make a forecast for the random variable $R_{t|t-h}$ using information up to $t-h$. The darker shade represents a 1-standard-error symmetric band around the point estimate, while the lighter represents a 2-standard deviation band. The final value for each solid line corresponds to the full-sample estimates we presented in Table 2.

We see two main patterns in this figure. The first one is related to the behavior of $\widehat{\beta}_{LTG}$. When included alone or alongside `D`, the estimates start as positive and migrate to the negative plane with the arrival of more recent data. This pattern gets stronger as we forecast at longer horizons. If Bordalo et al. (2023) were to run the same regression they studied in Table 2 of their paper with data only up to 2003, for example, they would find a different sign for their estimates. We can't rule out that this is just estimation uncertainty getting reduced as more data comes in, but we believe it is important to document this change in sign since it has different implications for models of subjective belief formation.

These positive estimates are closely connected to overly optimistic forecasts generated by these models on the verge of the dot-com bubble burst. Positive estimates for the coefficient on `LTG` , coupled with an ever-rising consensus growth forecast, led to high estimates of market returns. The estimates for β_{LTG} start being negative only after the forecast errors due to the dot-com bubble are realized. Alternatively, when `PE` is also included, the estimates are either insignificant or have a positive sign all the time.

The second pattern comes from the evolution of $\widehat{\beta}_{D}$. We note that the point estimates are typically negative, as Yu (2011) found and as we confirmed using his sample. However, the associated light shades often include zero, suggesting imprecise estimation, in particular, as we accumulate more data.

[21] We use HAC-adjusted standard errors following Newey and West (1987) with h lags.

In light of the jagged behavior of D after 2008, one possible explanation is that the larger samples used to generate the final point estimates mix data from different regimes, making it impossible to estimate the desired parameter due to misspecification. The second explanation is that the coefficient is zero, but due to having a smaller and finite sample, Yu (2011) found negative and significant results. We cannot rule out any of these explanations.

Nevertheless, we notice some “pockets of significance”, i.e., data strips that would lead to significant results if a researcher were given only a truncated version of our sample. That can be seen, for example, on the second row of the first panel. In that case, D is included alone in these regressions. The estimates start as being positive, although insignificant. As more data comes in, they become negative. During that period, Yu (2011) was published. More data comes in, and significance disappears once again. A similar qualitative phenomenon happens when we control for LTG and PE.

6 Conclusion

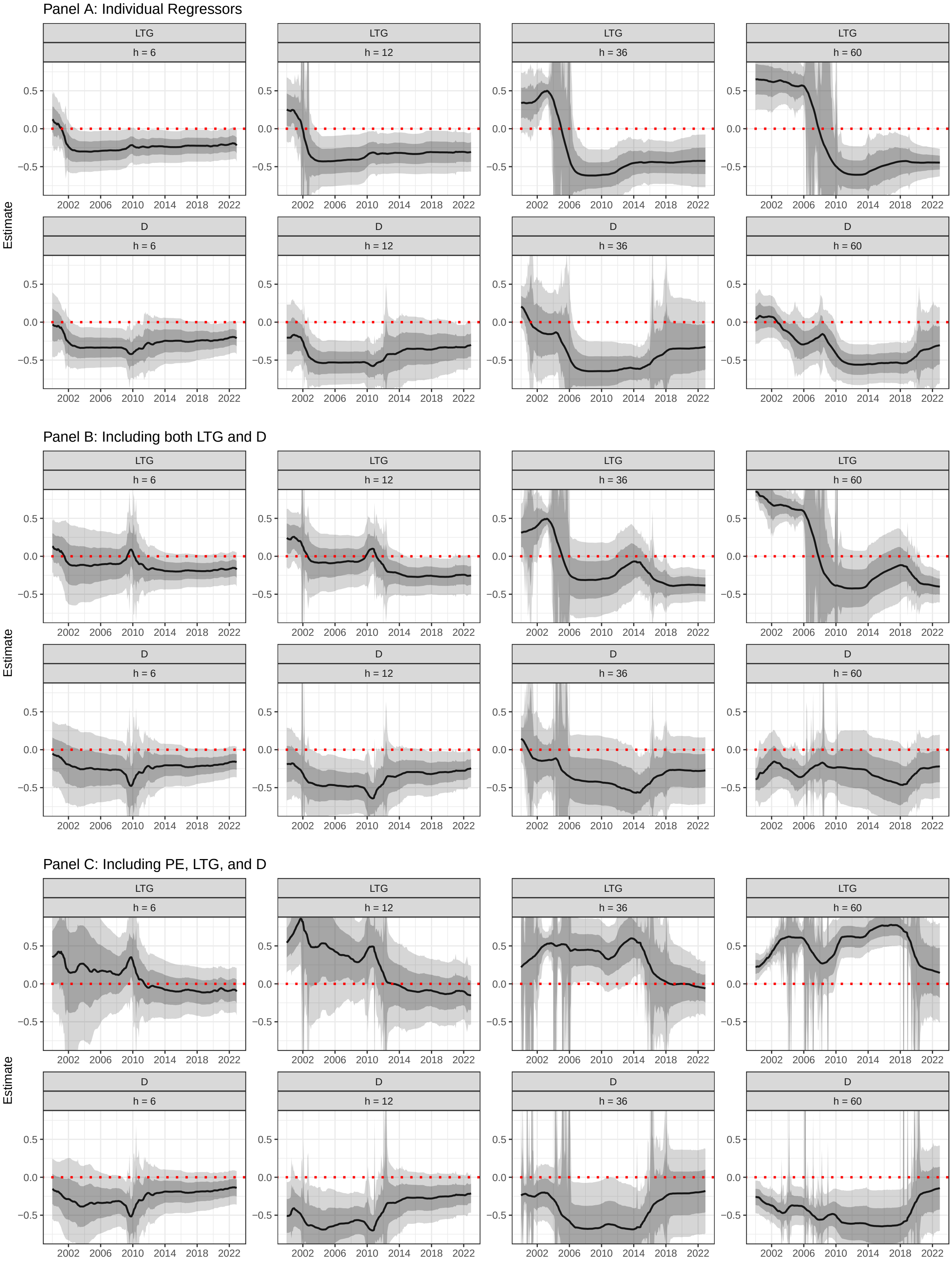
Eliciting and characterizing beliefs becomes crucial when departing from the rational expectations framework. Taking seriously contradictions between model-derived expectations about dividend and return processes and surveys is an important step in that direction.

In this paper, we aimed to narrow the gap between the literature on price formation under diagnostic expectations and the literature focused on disagreement. Our results highlight that consensus long-term growth forecasts have limited predictive power beyond what is already captured by the price-earnings ratio. Disentangling the forces that shape belief formation and characterizing their impact on asset prices and returns is an important but difficult task that merits further research.

Moreover, we document that analyst disagreement offers no meaningful predictive power after accounting for more recent samples - neither in-sample nor out-of-sample. Notably, we are the first to identify an apparent regime shift in how the disagreement series constructed from IBES data behaves after 2008.

We highlight two important caveats regarding our analyses. First, our study utilizes monthly data spanning from the early 1980s onward, which constitutes a relatively short sample period, especially for highly persistent processes and low-frequency phenomena. As demonstrated, researchers working with truncated subsets of this data would likely have reached different conclusions. This underscores the importance of using comprehensive samples to avoid drawing premature inferences. Leveraging cross-sectional evidence across firms may provide a valuable complementary approach to test the implications of different theories. While we are keen to explore the cross-sectional dimension, we defer an in-depth analysis to future research to maintain the focus of the current contribution. The heterogeneity offered by the cross-section of firms could shed new light on this problem.

Figure 5: Each panel shows the evolution of OLS estimates from regression (3). Panel A is dedicated to univariate models. Panel B studies the bivariate model that includes both `LTG` and `D`. Panel C shows the estimates one would get from including `PE` alongside the other two measures from I/B/E/S . Darker and lighter shades indicate symmetric 1 and 2 standard-deviation bands around point estimates, computed using the HAC estimator from Newey and West (1987) with *h* lags.



Second, we do not attempt to incorporate other disagreement measures as in Huang et al. (2021). Instead, we rely solely on IBES data to maintain a closer connection with the previous literature. Additionally, while newly designed surveys like Giglio et al. (2021) can be extremely valuable for providing new empirical facts, they are unlikely to offer insights into low-frequency phenomena like return predictability due to their necessarily short time series.

References

Adam, K. and Nagel, S. (2023). Expectations data in asset pricing. In *Handbook of Economic Expectations*, pages 477–506. Elsevier.

Angrist, J. D. and Pischke, J.-S. (2008). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Atmaz, A. and Basak, S. (2018). Belief dispersion in the stock market. *The Journal of Finance*, 73(3):1225–1279.

Bianchi, D., Büchner, M., and Tamoni, A. (2020). Bond risk premiums with machine learning. *The Review of Financial Studies*, 34(2):1046–1089.

Bordalo, P., Gennaioli, N., La Porta, R., and Shleifer, A. (2019). Diagnostic expectations and stock returns. *The Journal of Finance*, 74(6):2839–2874.

Bordalo, P., Gennaioli, N., LaPorta, R., and Shleifer, A. (2023). Belief overreaction and stock market puzzles. *Journal of Political Economy*.

Bordalo, P., Gennaioli, N., Porta, R. L., and Shleifer, A. (2024). Belief overreaction and stock market puzzles. *Journal of Political Economy*, 132(5):1450–1484.

Campbell, J. Y. and Thompson, S. B. (2008). Predicting excess stock returns out of sample: Can anything beat the historical average? *Review of Financial Studies*, 21(4):1509–1531.

Chen, J., Hong, H., and Stein, J. C. (2002). Breadth of ownership and stock returns. *Journal of Financial Economics*, 66(November-December):171–205.

Cinelli, C., Forney, A., and Pearl, J. (2022). A crash course in good and bad controls. *Sociological Methods & Research*, page 004912412210995.

Clark, T. and McCracken, M. (2013). *Advances in Forecast Evaluation*, page 1107–1201. Elsevier.

Clark, T. E. and West, K. D. (2007). Approximately normal tests for equal predictive accuracy in nested models. *Journal of Econometrics*, 138(1):291–311. 50th Anniversary Econometric Institute.

De La O, R. and Myers, S. (2021). Subjective cash flow and discount rate expectations. *The Journal of Finance*, 76(3):1339–1387.

Diebold, F. and Mariano, R. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3):253–63.

Diether, K. B., Malloy, C. J., and Scherbina, A. (2002). Differences of opinion and the cross section of stock returns. *The Journal of Finance*, 57(5):2113–2141.

Freire, G. and Riva, R. (2023). Asymmetric violations of the spanning hypothesis. *SSRN Electronic Journal*.

Gao, G. P., Lu, X., Song, Z., and Yan, H. (2019). Disagreement beta. *Journal of Monetary Economics*, 107:96–113.

Giglio, S., Maggiori, M., Stroebel, J., and Utkus, S. (2021). Five facts about beliefs and portfolios. *American Economic Review*, 111(5):1481–1522.

Goyal, A., Welch, I., and Zafirov, A. (2021). A comprehensive look at the empirical performance ofequity premium prediction ii. *SSRN Electronic Journal*.

Greenwood, R. and Shleifer, A. (2014). Expectations of returns and expected returns. *Review of Financial Studies*, 27(3):714–746.

Gu, S., Kelly, B., and Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Hong, H. and Sraer, D. A. (2016). Speculative betas. *The Journal of Finance*, 71(5):2095–2144.

Hong, H. and Stein, J. C. (2007). Disagreement and the stock market. *Journal of Economic Perspectives*, 21(2):109–128.

Huang, D., Li, J., and Wang, L. (2021). Are disagreements agreeable? evidence from information aggregation. *Journal of Financial Economics*, 141(1):83–101.

Iachan, F. S., Nenov, P. T., and Simsek, A. (2021). The choice channel of financial innovation. *American Economic Journal: Macroeconomics*, 13(2):333–372.

Jarrow, R. (1980). Heterogeneous expectations, restrictions on short sales, and equilibrium asset prices. *The Journal of Finance*, 35(5):1105–1113.

La Porta, R. (1996). Expectations and the cross-section of stock returns. *Journal of Finance*, 51(5):1715–42.

Li, F. W. (2016). Macro disagreement and the cross-section of stock returns. *The Review of Asset Pricing Studies*, 6(1):1–45.

Miller, E. M. (1977). Risk, uncertainty, and divergence of opinion. *The Journal of finance*, 32(4):1151–1168.

Nagel, S. and Xu, Z. (2023). Dynamics of subjective risk premia. *Journal of Financial Economics*, 150(2):103713.

Newey, W. K. and West, K. D. (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica*, 55(3):703.

Park, C. (2005). Stock return predictability and the dispersion in earnings forecasts*. *The Journal of Business*, 78(6):2351–2376.

Simsek, A. (2021). The macroeconomics of financial speculation. *Annual Review of Economics*, 13:335–369.

Welch, I. and Goyal, A. (2008). A comprehensive look at the empirical performance of equity premium prediction. *Review of Financial Studies*, 21(4):1455–1508.

Yu, J. (2011). Disagreement and return predictability of stock portfolios. *Journal of Financial Economics*, 99(1):162–183.

Supplementary Appendix

A Figures

Figure A.1: Autocorrelation function for the variables from I/B/E/S and for PE. We plot up to lag 36. We use the full sample for estimation (December 1981 - December 2022).

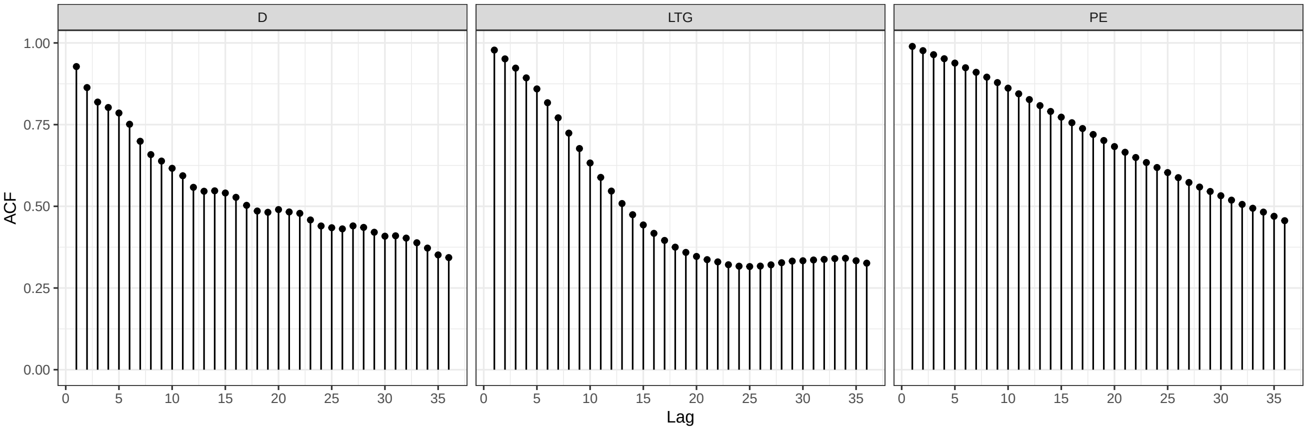


Figure A.2: We plot the evolution of $X_{t+12} - X_t$ for $X \in \{D, LTG, PE\}$. Data starts in December 1981 and ends in December 2022. Shaded areas represent NBER recessions. We normalized variables before plotting so they could be shown on the same scale.

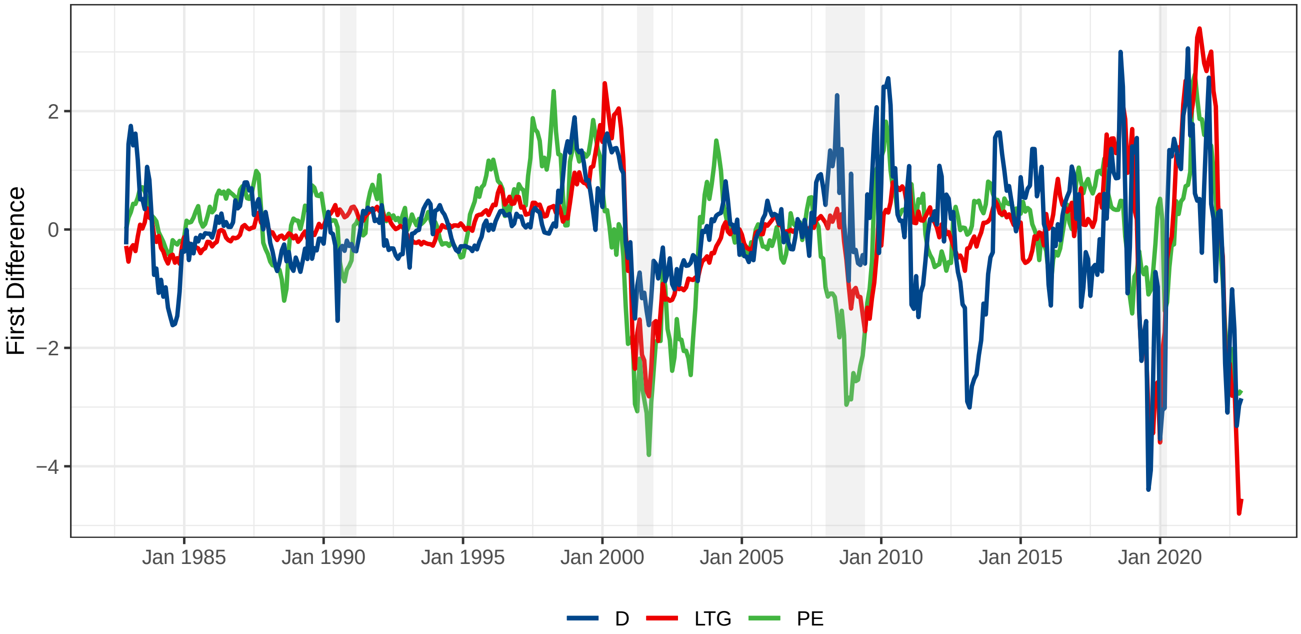


Figure A.3: We plot all the time series used in the analysis in their original scales. The sample starts in December 1981 and ends in December 2022. Shaded areas represent NBER recessions.

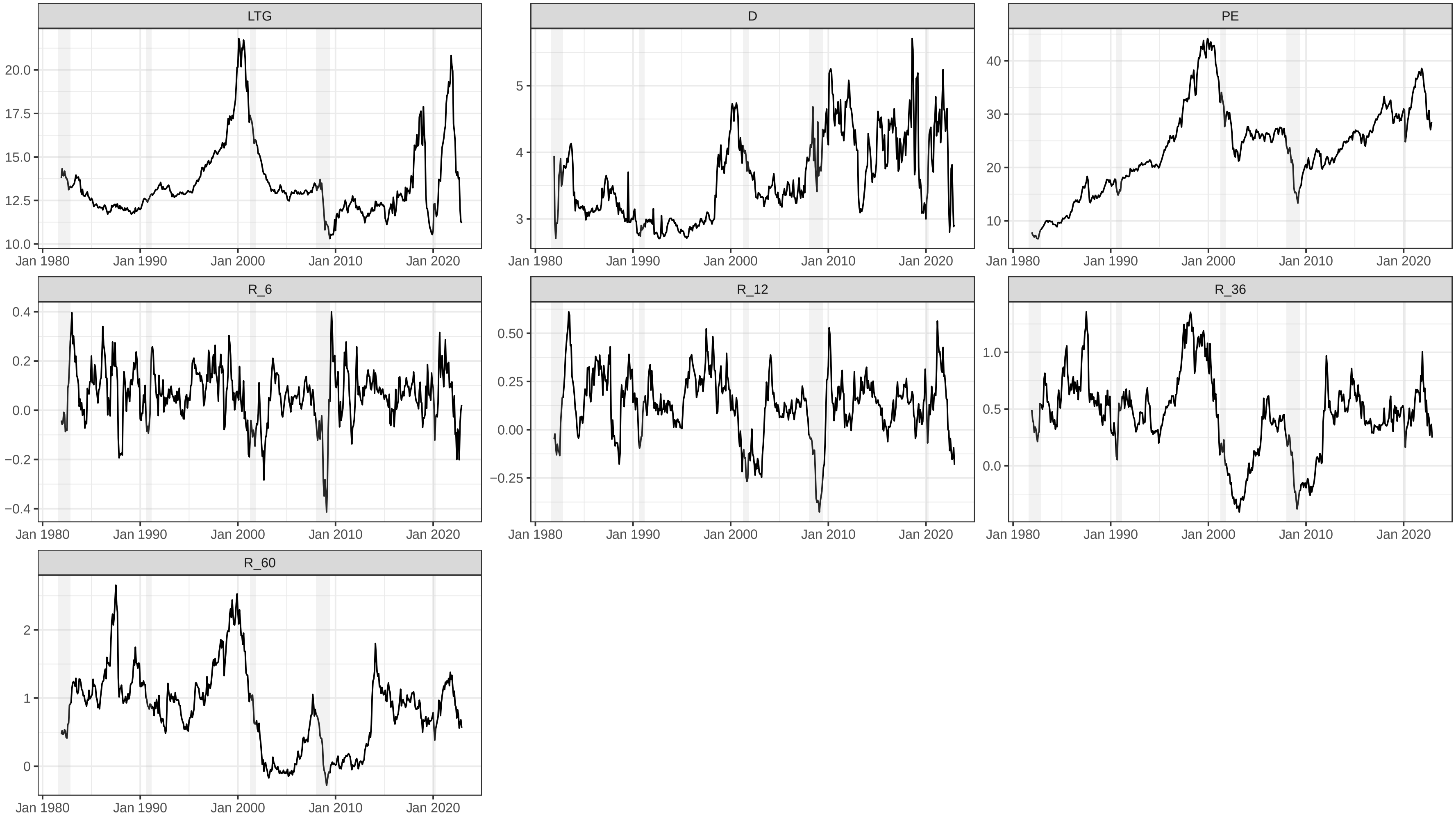
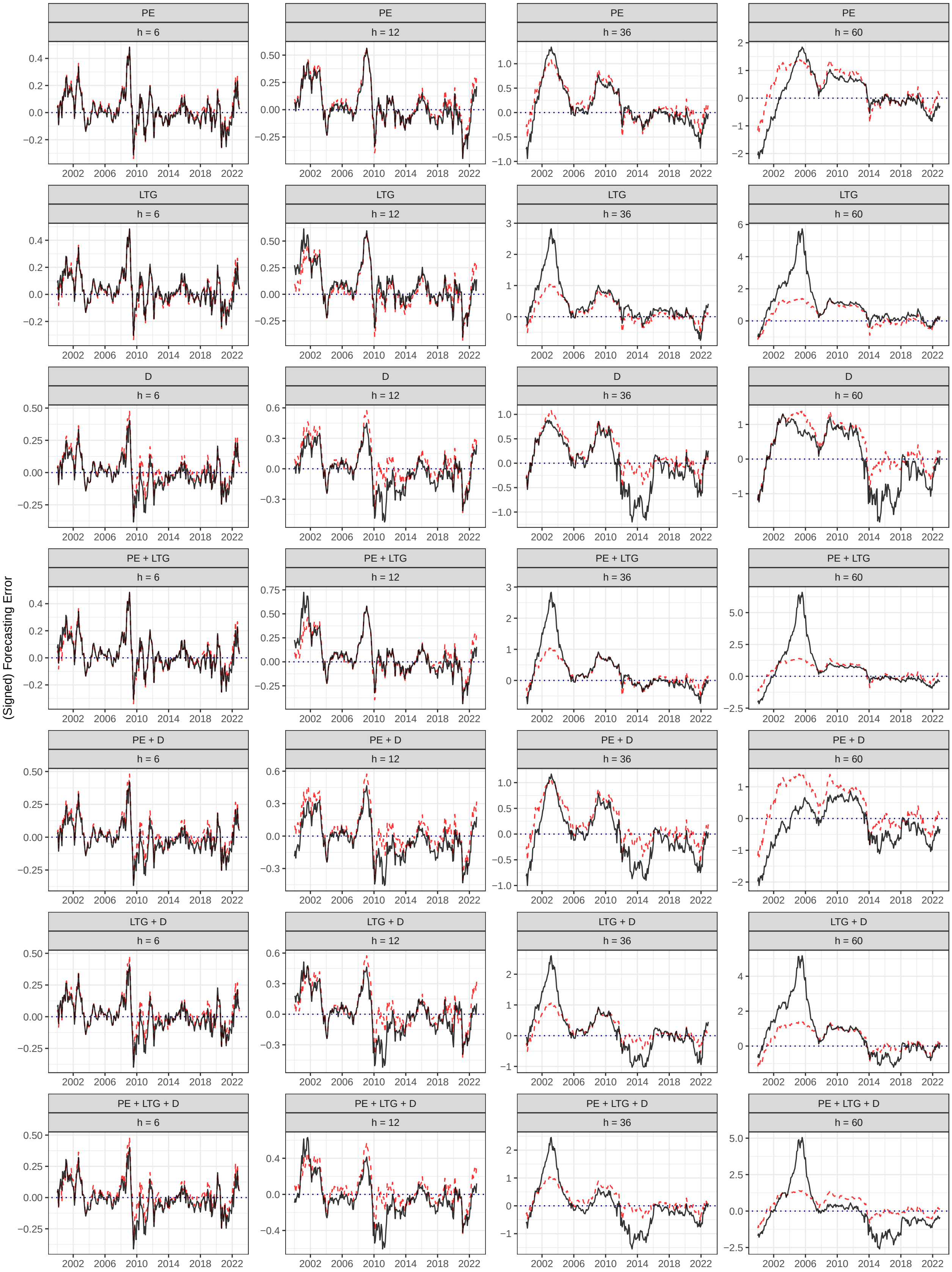


Figure A.4: We show the evolution of $e_{t,h}^{(M)}$. See Figure 3 for details. The only difference is that Figure 3 reports the absolute value of the measures shown here.


(Signed) Forecasting Error
PE
h = 6
PE
h = 12
PE
h = 36
PE
h = 60
LTG
h = 6
LTG
h = 12
LTG
h = 36
LTG
h = 60
D
h = 6
D
h = 12
D
h = 36
D
h = 60
PE + LTG
h = 6
PE + LTG
h = 12
PE + LTG
h = 36
PE + LTG
h = 60
PE + D
h = 6
PE + D
h = 12
PE + D
h = 36
PE + D
h = 60
LTG + D
h = 6
LTG + D
h = 12
LTG + D
h = 36
LTG + D
h = 60
PE + LTG + D
h = 6
PE + LTG + D
h = 12
PE + LTG + D
h = 36
PE + LTG + D
h = 60