Subjective Bond Returns and Belief Aggregation

Buraschi, Andrea; Piatti, Ilaria; Whelan, Paul

Document Version
Accepted author manuscript

Published in:
Review of Financial Studies

DOI:
10.1093/rfs/hhab115

Publication date:
2022

License
Unspecified

Citation for published version (APA):
Buraschi, A., Piatti, I., & Whelan, P. (2022). Subjective Bond Returns and Belief Aggregation. *Review of Financial Studies*, *35*(8), 3710-3741. https://doi.org/10.1093/rfs/hhab115

Subjective Bond Returns and Belief Aggregation

Andrea Buraschi Ilaria Piatti Paul Whelan

Abstract

The forecasting literature has presented overwhelming evidence that the aggregation of heterogeneous expectations leads to improvements in forecast accuracy; however, outperforming a simple equal weighting scheme has proved challenging. This paper proposes an aggregation scheme of subjective bond return expectations based on the historical accuracy of professional interest rate forecasters. Our aggregate belief proxy outperforms equal weight and median weight combinations and is comparable to statistical projections even if its dynamics are quite different. With this measure at hand, we study the relationship between quantities of risk and compensation for risk and demonstrate a strong link to subjective expectations even if this is difficult to detect using realized returns.

JEL classification: D9, E3, E4, G12

Keywords: Forecasting, Forecast Combinations, Heterogeneous Beliefs, Bond Returns.

This version: April 2021

Andrea Buraschi is Chair of Finance at Imperial College Business School; Ilaria Piatti is at the School of Economics and Finance, Queen Mary University of London; and Paul Whelan is at Copenhagen Business School. We thank Steven Baker, Anna Cieslak, Bernard Dumas, Christian Eyerdahl-Larsen, Marco Giacoletti, Erik Hjalmarsson, Philipp Illeditsch, Liangliang Jiang, Daniel Pesch, Zhan Shi, Kenneth Singleton, Gyuri Venter, Jonathan Wright, Ti Zhou and participants at the Cavalcade North America 2018, Winter Finance Summit 2018 St Moritz, Cavalcade Asia 2017, Adam Smith conference HEC-Paris 2017, SGF 2017, Interest rates after the financial crisis Örebro University 2017, AFA Chicago 2017, EFA Mannheim 2017, CICF Xiamen 2016, FMA Helsinki 2016, Empirical Asset Pricing Workshop at the University of Lancaster 2016, Gerzensee 2016, and the seminar participants of Vienna University, Green Templeton College, the University of Porto, Goethe University Frankfurt, Renmin University, Paris Dauphine University, and Bocconi University. Paul Whelan gratefully acknowledges acknowledges financial support from the Independent Research Fund Denmark (grant no. DFF-9037-00105B), from the FRIC Center for Financial Frictions (grant no. DNRF-102), and from Danish Finance Institute (DFI). The usual disclaimer applies. Correspondence should be sent to Paul Whelan, Copenhagen Business School, Solbjerg Plads 3, Room A5.11, Denmark, DK-2000, +4538152410, pawh.fi@cbs.dk.

The estimation of expected returns is a central tenet of empirical asset pricing. It is, therefore, not surprising that a variety of measurement approaches exist, ranging from simple historical averages to sophisticated time-series models. Surveys are an alternative measurement approach, whose availability has grown substantially over time, and are often available at the individual forecaster level. This greatly increases the information available for measurement but poses challenges in how to aggregate heterogeneous survey forecasts. The most common approach when using surveys is to focus on consensus beliefs, defined as an equally weighted average expectation, which is consistent with a conclusion from the forecast combination literature that shows simple average forecasts are hard to beat in practice.

Motivated by the idea that simple averages might overlook useful information in the cross-section of beliefs, this paper studies the aggregation of subjective bond return forecasts. Our main result demonstrates that persistence in accuracy can be exploited to construct a novel real-time weighting of individual beliefs, which outperforms common benchmarks, and is simple to implement: we discard historically bad forecasters and form an expectation from the remaining good forecasters' beliefs.

To establish this result, we estimate subjective bond return expectations from a monthly panel of professional market participants' beliefs about future U.S. Treasury yields. The availability of disaggregated survey data is important since it allows us to track the relative performance of forecasters over time, which is the key step in constructing a forecast combination that beats the simple average.

We begin by studying the cross-sectional properties of expectations about bond returns. At the one-year forecast horizon, we document large disagreement about future bond returns which is highly persistent. Forecasters in the top or bottom quartiles of the distribution of beliefs have around a 40% probability of remaining in that quartile in the following year.

The cross-sectional distribution of bond return forecast accuracies also displays large heterogeneity; moreover, it is skewed towards bad forecasters. For the 10-year bond, 37% of the forecasters statistically underperform the consensus at the 5% level, while 9% outperform the consensus. Most importantly,

accuracy is persistent: forecasters who are good tend to remain good and forecasters who are bad tend to remain bad. Forecasters in the top quartile of the cross-sectional distribution of accuracy have a 49% probability of being in the same quartile the following year, and they transition to the bottom quartile of the distribution with a probability of only 3%.

Persistence in accuracy allows us to identify the best and worst forecasters ex-ante. We propose and study a real-time measure of aggregate subjective expected excess bond returns that (a) accounts for agents' performance records; (b) does not rely on parameter estimation; and (c) accounts for time variation in forecasting environments. We denote this measure $EBR_t^\star$ which stands for 'Expected Bond Returns'. Each month we compute relative accuracies defined as a percentile position in the cross-sectional distribution of squared forecast errors. Then, we compute the median of the accuracy ranking percentiles over 5-year rolling windows ending at time t for each agent and sort them. Finally, we select the top quartile of agents and aggregate their beliefs with linear weights that are increasing in past accuracy. This measure is available for the cross section of U.S. long term bond maturities, although in the paper we focus on 5 and 10-years tenors, and is available in real time from January 1988 to January 2020 at monthly frequency.

The difference between $EBR_t^\star$ and the equally weighted consensus belief can be substantial, varying between -2% and +5% for 10-year bonds. Moreover, the mean squared forecast errors implied by $EBR_t^\star$ are smaller than those of equal weight or median forecasts and this difference is strongly statistically significant. The outperformance of $EBR_t^\star$ is quite striking given the conclusion from the forecast combination literature that, once estimated, theoretically optimal combinations struggle to beat an equally weighted consensus measure.

A decomposition of mean squared errors into squared mean errors and error variances shows the existence of a trade-off in the construction of an aggregate measure. On the one hand, including a large number of agents achieves greater forecast error diversification; on the other hand, this comes at the cost

of introducing the beliefs of agents with larger average forecast errors. We document the existence of an interior solution that balances out these two effects and maximizes overall accuracy. Selecting the 25% of agents who have been more accurate in the past, and assigning them weights that are increasing in historical performance, is an effective and simple way to address this trade-off. Moreover, we also find that while it is important for the econometrician to use information on agents' accuracy beyond 12-months, this information eventually becomes redundant and one should discount information from the distant past. We study the robustness of our measure along different dimensions and show that the outperformance of $EBR_t^\star$ with respect to the consensus holds in different forecasting environments and that the agents selected by our procedure tend to have both low mean errors and low error variance. This suggests $EBR_t^\star$ agents are genuinely skilful and were not just lucky to have been relatively pessimistic about future yields in a period of generally decreasing interest rates.

We provide an alternative statistical validation of our aggregation scheme using a Bayesian model averaging approach. An asset manager with the prior belief that all groups of forecasters have identical skill at the beginning of our sample would have learned to assign a weight of 100% to the quartile of most accurate agents half way through our sample, and zero to the remaining three quartiles.

Our aggregation approach contributes to the literature that explores optimal forecast combinations. Related to our work, Aiolfi and Timmermann (2006) show that competing forecasting models display persistent periods of out-performance but also switch, that is, at times the best models become the worst and vice versa. This is consistent with our selection approach that considers rolling relative accuracy as a sorting variable. Our approach also relates to Diebold and Shin (2019) who propose a LASSO-based procedure that sets some combining weights to zero and shrinks the survivors toward equality. Their results suggests that the vast majority of forecasters should be discarded, and the remainder should be averaged. A notable difference between our paper and existing literature is our focus on expected bond returns. In fact, the forecast combination literature mostly studies macroeconomic variables such as GDP

growth or inflation. This may be due to a belief that survey expectations of financial returns are not consistent with rational expectations models (see e.g. Greenwood and Schleifer (2014) for expected stock returns).

However, when we compare the properties of our survey-implied measure to predictions from benchmark equilibrium models which link variation in expected excess returns to conditional variances of state variables, we find a statistically strong correlation whose sign makes sense economically. The risk factor proxies we consider include economic uncertainty, dispersion in beliefs, consumption surplus, realized and risk-neutral volatilities. We find quantities of risk can explain subjective expected excess returns with R^2 ranging between 20% and 30% with highly significant regression coefficients. The loadings on the factors are consistent with rational expectation models that predict investors demand compensation for holding volatility risk. These findings provide a partial explanation for a long standing puzzle that fails to detect a relationship between quantities of risk and expected excess returns. We argue that empirical results based on projections on ex-post realizations can be quite different from ex-ante investors expectations and that our measure can capture the link between risk compensation and the quantity of risk more precisely.

The paper proceeds as follows. Section 1 presents the data, reports the properties of the cross section of subjective expected excess bond returns, and documents their degree of heterogeneity and persistence. Section 2 investigates the accuracy of the forecasters. Section 3 describes our aggregation approach and compares it to standard consensus measures and also to common statistical measures of expected excess bond returns. Section 4 discusses the implications when an econometrician uses our $EBR^{\star}$ measure to examine the link between bond expected excess returns and the risk factors emerging from structural models. Section 5 concludes. Finally, further empirical results and robustness checks are provided in an Online Appendix (OA).

1. The Cross Section of Beliefs

1.1. Data

BlueChip Financial Forecasts (BCFF) is a monthly survey providing extensive panel data on the expectations of professional economists working at leading financial institutions about all maturities of the U.S yield curve and economic fundamentals, such as GDP and inflation. We construct real-time measures of subjective expected excess bond returns from BCFF for the sample period January 1988 to January 2020.

Our empirical approach exploits institution specific forecasts for Treasury bills with maturities 3-months/6-months, Treasury notes with maturities 1, 2, 5, 7, 10-years, and the 30-year Treasury bond.[1] The contributors to BCFF are asked to provide point forecasts at horizons that range from the end of the current quarter to 5 quarters ahead (6 from January 1997). We restrict the panel such that agents must contribute at least 60 times with a minimum number of 4 interest rate projections per month which generates an unbalanced panel of 89 agents where the median number of contributions to the panel is around 140 months. To obtain expected zero coupon discount rates we estimate a Nelson-Siegel (NS) model on individual agent subjective par-yield forecasts. We calculate the term structures using all available maturities (including 30-year Treasury yield forecasts) and obtain a monthly panel data of expected constant time-to-maturity zero coupon (continuously compounded) discount rates.[2] In the following, we retain bond maturities evenly spaced between 1 and 10 years. One complication of BCFF is that while surveys are conducted on a monthly basis the projections are reported on a future quarterly calendar cycle so that the forecast horizon varies each month. To construct a j-quarter ahead constant maturity forecast we linearly interpolate along adjacent horizons.[3]

[1] Forecasters are identified by institution's name. For example, 'J.P. Morgan' or 'Goldman Sachs' or 'Fannie Mae'. In the paper we use the terms agent and institution interchangeably.

[2] We note that while these objects are actually estimated subjective expected zero-coupon yields, since they are not elicited directly, in what follows we drop the qualifier 'estimated' to save space.

[3] Data are available from 1983 but we show in Section 1 of the OA that the number of contributors and quality of the data improves substantially from 1988, which is where we begin our sample. The OA also reports a detailed discussion of the BCFF data and our construction of constant maturity zero-coupon yield expectations.

For realized bond data we use zero-coupon bond yields provided by Gürkaynak, Sack, and Wright (2006) for the sample period June 1961 to January 2021, which are available from the Federal Reserve website, and in robustness tests we use par yields reported in the Federal Reserve's H.15 release.

1.2. Subjective excess bond returns

Given information on expectations about the cross section of future interest rates, BCFF allows us to compute individual subjective expected excess bond returns as follows. Let p_t^n be the logarithm of the time-t price of a risk-free zero-coupon bond that pays one unit of the numeraire n-years in the future. Continuously compounded spot yields are then defined as $y_t^n = -\frac{p_t^n}{n}$. We can compute the implied cross section of expected excess bond returns (EBRs) with 1-year (12 months) forecast horizon as $erx_{i,t}^n = E_t^i\left[p_{t+12}^{n-1}\right] - p_t^n - y_t^1$ since from the surveys we directly observe yield forecasts $E_t^i\left[y_{t+12}^{n-1}\right]$:

$$erx_{i,t}^n = -(n-1) \times \underbrace{E_t^i\left[y_{t+12}^{n-1}\right]}_{\substack{\text{Survey Yield} \\ \text{Forecasts}}} + ny_t^n - y_t^1 \tag{1}$$

where t is measured in months, that is our observation frequency, and the bond maturity n is expressed in years.[4]

1.3. Belief heterogeneity and persistence

The top panel of Figure 1 displays the forecasts of the 10-year expected excess bond returns from each agent. The figure documents significant time-series variation of the entire distribution of beliefs and clarifies the extent of cross-sectional dispersion around the consensus. The plot suggests there are no outright '*crazy*' forecasts even though we do not winsorize. It also highlights that while on average subjective expected excess bond returns are positive, there are a significant number of periods in which

[4]The literature studying yield predictability often focuses on continuously compounded log yields and interprets expected excess log returns as risk premia. However, risk premia should really be measured from expected excess simple returns, which differ from expected excess log returns by a convexity term. While in realized data the difference between simple and log returns is known to be small. Section 1 in the OA shows this approximation is even tighter using surveys expectations.

they are negative and, indeed, at some points towards the end of the sample almost the entire distribution of beliefs shifts below zero.[5]

The bottom panel of Figure 1 highlights the time variation in heterogeneity by plotting the cross-sectional interquartile range of subjective excess bond returns for 5 and 10-year maturities. Disagreement about bond excess returns is clearly time-varying, large and persistent. For the 10-year bond, the average value of the interquartile range is around 5 percentage points, which is one order of magnitude larger than the median value of the expected 10-year bond excess return, that is around 0.6%.[6]

Given the relatively long forecast horizon (1-year) and the nature of the target variable (a multivariate function of the state of the economy) we believe it is unlikely that such disagreement originates from private information or differences in information sets. Instead, it is more likely due to heterogeneity in priors or models as argued by Patton and Timmermann (2010), or to information frictions as argued by Andrade, Crump, Eusepi, and Moench (2016).

[Insert Figure 1 here]

In order to quantify the persistence in beliefs, we compute annual transition probabilities between quartiles of the distribution of forecasts. Table 1 shows the probability of a forecaster transitioning from a given quartile of the cross-sectional distribution to another quartile in the following year (annual horizon). To assess statistical significance, we compute confidence intervals by simulation under the null of zero belief persistence, explicitly accounting for the unbalanced nature of our panel. Specifically, within each year we sample from the data those agents contributing to the panel and compute a random permutation of their beliefs with replacement so that every 12 periods the panel indices are randomized. Resampling from the data 1000 times we compute the distribution under the null. If beliefs were not persistent, the entries in these transition matrices should be approximately equal to 25%. Instead, we find that transitions in

[5] Figure 8 in the OA displays the fractions of positive versus negative forecasts at each point in time.

[6] Summary statistics for the quartiles of the distribution of subjective expected excess bond returns, for maturities of 5 and 10 years, are provided in Table 4 of the OA.

the outer quartiles is highly persistent and statistically higher than 25%. For example, for the 10-year bond, forecasters in the top quartile (Q_1) have a probability of 41% to remain in the same quartile of the distribution, while they have a probability of 14% of transitioning to the bottom quartile (Q_4). Belief persistence in the bottom quartile is quantitatively similar.

[Insert Table 1 here]

2. The Cross Section of Errors

2.1. Heterogeneity in accuracy

We now turn our attention to heterogeneity in forecasting ability. For each contributor i, we compute forecast errors ($FE^n_{i,t+12}$) and mean squared errors (MSE^n_i) at the one year horizon

$$FE^n_{i,t+12} = rx^n_{t+12} - erx^n_{i,t} \tag{2}$$

$$MSE^n_i = \frac{1}{T_i}\sum_{t=1}^{T_i}(FE^n_{i,t+12})^2 \tag{3}$$

where rx^n_{t+12} is the realized excess return on an n-period bond and T_i is the total number of contributors to the BCFF panel for agent i. In the following, we also study mean forecast errors $ME^n_i = \frac{1}{T_i}\sum_{t=1}^{T_i} FE^n_{i,t+12}$, squared mean errors $SME^n_i = (ME^n_i)^2$ and error variances $EVAR^n_i = MSE^n_i - SME^n_i$. Figure 2 summarizes the cross-sectional distribution of MSEs.[7] The figure demonstrates significant unconditional heterogeneity in accuracy for both bond maturities, and the 10-year bond MSE distribution is positively skewed. Indeed, a large fraction (∼30%) of the distribution have *root*-mean-square errors in excess of 10%.

[Insert Figure 2 here]

[7] Our panel is unbalanced so individual forecast errors are computed over different samples. Section 2 of the OA explores in greater detail the statistical properties of heterogeneity in accuracy. In Section 3 that follows, we explore conditional properties of heterogeneity in accuracy and the decomposition of MSEs into SMEs and $EVAR$s.

Figure 3 plots the cross section of Diebold and Mariano (1995) (DM) test statistics, computed at the individual agent level for each of our 89 forecasters, under the null hypothesis that MSEs are equal to the consensus MSE, against the one-sided alternative. Throughout the paper, we calculate DM test statistics and associated p-values as in Harvey, Leybourne, and Newbold (1997). There are two learning points to be extracted. Firstly, the cross-sectional distribution of DM test statistics is not only disperse but strongly skewed towards bad forecasters. Secondly, a large fraction of forecaster accuracy is statistically different than the consensus. For the 10-year bond, 37% of the forecasters statistically underperform, while 9% outperform the consensus, at the 5% level.

[Insert Figure 3 here]

2.2. Persistence in accuracy

A natural question that arises at this point is whether heterogeneity in accuracy is persistent. To address this question, we compute annual transition probabilities on MSEs, defined as the probability that forecasters in a given quartile of the MSE distribution at time t stay in that quartile the following year or move to a different quartile of the distribution. We find strong evidence of persistence in accuracy, especially in the outer quartiles, which are all significantly different than 25% at the 1% level. Statistical inference is conducted under the null of no persistence by resampling randomly from the data as in Section 1.3. For example, for the 10-year bond, a good forecaster (Q_1) has a probability of 49% of being a good forecaster the following year. The probability that a bad forecaster (Q_4) remains a bad forecaster is even higher, equal to 63%. Transitioning from Q_1 to Q_4, or from Q_4 to Q_1 is highly unlikely, with probabilities less than or equal to 3%. Interestingly, comparing Tables 1 and 2, we can see that the persistence in accuracy is even stronger than the persistence in beliefs.

[Insert Table 2 here]

3. A Real Time Measure of Subjective Bond Returns

Given the large heterogeneity in subjective bond return expectations, it is not surprising that one can, *ex-post*, identify forecasters that are significantly better than others, but this does not imply that 'good' forecasters are identifiable *ex-ante*. While the persistence in forecast accuracy suggests that this might be possible, it is an open question whether there exists an aggregation method that would allow a decision maker to combine beliefs in real time to improve forecast accuracy relative to a simple aggregation, such as an equally weighted consensus measure. This is a non-trivial task, since a large literature has shown that while simple averages are not theoretically optimal under mean-square-error loss criteria, empirically they often outperform implementations of theoretically optimal weights ('the equal weights puzzle').

To highlight this point, using our dataset and for our sample period, Table 3 reports the forecasting performance of two empirical implementations of the theoretically optimal mean-square-error loss weights vis-à-vis the equally weighted consensus forecast. In the first row of each panel we estimate the optimal weights of Bates and Granger (1969) (BG), which depend on error variances and covariances. Empirically, estimating the off-diagonal elements of the covariance matrix is numerically challenging so we follow common advice and set them to zero. The second row of each panel uses the optimal weights implementation of Granger and Ramanathan (1984) (GR), which is the outcome of a restricted least squares regression. Here, we follow these authors and relax the unbiasedness assumption by including a constant with unrestricted weights. The BG and GR weights are estimated in real-time using 5-year rolling windows. Summarizing, once estimated, optimal weights do not outperform simple averages. Focusing on the 10-year bond, the BG weights generate $MSEs$ that are larger than the consensus, while the GR weights generate smaller MSEs but not significantly so. Section 3 in the OA reports the forecasting performance of a wider set of existing aggregations and shows that, amongst the alternatives we consider, only the median forecast beats the equally weighted average.

[Insert Table 3 here]

3.1. The aggregation scheme

The results of the paper so far suggest that deviating from an equally weighted forecast combination might lead to an improvement in the accuracy of the survey-implied expected excess bond returns. However, we have just shown that traditional aggregation methodologies do not work well in our setting. Motivated by these findings, we propose and test a novel aggregation scheme.

The proposed aggregation combines three objectives. First, it does not require the estimation of a model. Second, it gives greater weight to agents with better historical accuracy. Third, it accounts for differences in forecasting environments over time and for the unbalanced nature of survey forecasts. For each time period t:

1. We compute a panel of squared forecast errors over rolling windows $[t-N, t]$ and translate this to a panel of percentile rankings $\mathcal{R}_{it} \in (0, 1)$. An agent with a low ranking corresponds to a low squared error relative to their cohort.

2. For each agent present at date t, we compute their median percentile ranking over the rolling window $[t-N, t]$.

3. We select agents in the top Q^{th} percentile of the median percentile ranking distribution, setting the weights on complimentary agents to zero.

4. We aggregate the expectations of the selected agents using linear weights that are increasing in past historical rank accuracy, i.e. linearly decreasing in median $\mathcal{R}_{it}$.

Note that our aggregation approach can be applied to settings with unbalanced panels, as is often the case with surveys. Moreover, measuring accuracy in terms of historical rankings, instead of the level of average squared errors, addresses concerns related to heteroskedasticity since we do not favour agents who

are only present when the level of interest rate volatility happened to be low or, more generally, when the level of forecast errors is on average smaller.

We label this conditional measure of subjective n-year bond returns $EBR_t^{\star,n}(N,Q)$. We denote consensus expectations computed from equal weights as EBR^c. We employ expectations running from 1988.1 to 2020.1 and realizations running from 1989.1 to 2021.1, and compute measures in real time using only current and past information. We focus on the 5 and 10 year maturity bonds and consider the average percentile rankings $\mathcal{R}_{it}$ across these two maturities in the construction of $EBR_t^{\star,n}(N,Q)$.[8]

3.2. $EBR^\star$ vs EBR^c

The most important choice in the construction of $EBR_t^{\star,n}(N,Q)$ is the trimming parameter Q which determines which agents to discard and which agents to keep. Figure 4 summarizes how the forecasting properties of $EBR_t^{\star,n}$ change as a function of Q, for n equal to 5 (left panels) and 10 years (right panels), and a look-back parameter N equal to 60 months. The top panels show the ratio of MSEs defined as $\frac{MSE(EBR_t^c)}{MSE(EBR_t^\star)}$, when Q is allowed to vary between 0.05 and 1.

[Insert Figure 4 here]

A number of learning points emerge. Firstly, the MSE ratio is a humped shaped function of Q. When Q is chosen to be either very small or very large, our forecast combination has a large MSE compared to that of the consensus. This is clearly visible as we move from the left of the distribution, where we combine the beliefs of a small number of historically good forecasters, to the right of the distribution where the combination tends towards placing a linear weight across all agents. The lowest MSE is attained for $Q \sim 0.25$, i.e., when a quarter of the top performing agents is included in $EBR_t^\star$. The finding that

[8] Constructing $EBR^{\star,n}$ based on maturity specific forecast errors produces quantitatively similar results. Indeed, there is substantial overlap in the selected forecasters and their weights computing accuracy at the forecaster level or forecaster-maturity level. For example, building weights based on maturity specific errors for the 5 and 10-year maturities, and then computing the number of times agents appear in each measure as a fraction of times they appear in the dataset, we find that the correlation between the two frequency of occurrence vectors is 77%.

a relatively small subset of forecasters should be included implies that the majority of the cross-section provide limited useful marginal information.

To study the statistical significance of these results, the bottom panels of Figure 4 plot the p-values of a DM test for different values of Q. We test the null hypothesis of equality of MSEs against the alternative that the MSE of $EBR_t^\star$ is lower than the consensus MSE. When $EBR_t^{\star,n}(N,Q)$ is constructed with $0.20 \leq Q \leq 0.50$, the p-values of the DM test statistics reject the null hypothesis of equality in forecast accuracy. For the 10 year bond, when $EBR_t^{\star,10}$ is based on a small set of most accurate agents, i.e. $Q = 0.05$, the p-value of the DM test statistic is equal to 0.60 and above $Q = 0.55$ the p-values quickly rise above the 10% significance level. When $Q = 1$ the DM statistics (not shown) are negative, i.e. the MSE of the consensus is smaller than the MSE of $EBR^\star$. For $Q = 1$, both $EBR_t^\star$ and EBR_t^c use the same sample of agents, but the former uses aggregation weights that are linearly increasing in past accuracy while the latter uses equal weights.

To gain further insight, Figure 5 decomposes the MSEs of each forecast into the sum

$$MSE(N,Q) = SME(N,Q) + EVAR(N,Q), \tag{4}$$

The top two panels show that SMEs are increasing for $Q > 0.25$ but that the variance of the forecast errors are increasing for $Q < 0.25$. These two effects highlight the trade-off between the benefit of including a larger number of agents to obtain greater forecast error diversification and the cost of introducing a larger average forecast error. As a result, there exists an interior value of Q that balances out these two effects and minimizes MSE. In what follows, we will focus on this case, namely $Q = 0.25$.

[Insert Figure 5 here]

The second choice in the construction of $EBR_t^{\star,n}(N,Q)$ is N, which defines the length of the rolling window used to estimate the median percentile rankings. Table 4 shows the sensitivity of the MSE

to varying N, holding constant $Q = 0.25$. We find that for both bond maturities, the MSE is higher when using a rolling window either of $N = 24$ or 120 months, compared to $N = 60$ months. This is consistent with the empirical fact that while accuracy is persistent, so that the econometrician wants to track the accuracy of an agents beyond 12 months, it has limited memory meaning that when computing the time-varying weights one should discount or discard information from the distant past.

[Insert Table 4 here]

Next, we study alternative definitions of historical good performance. In addition to our benchmark historical rankings based on squared errors ($EBR_t^\star$), we also consider historical rankings based on mean errors and error variances and we denote the corresponding aggregate measures EBR_t^{ME} and EBR_t^{EV}, respectively. For EBR_t^{ME} we follow exactly the same procedure as for $EBR_t^\star$ but with 'forecast errors' instead of 'squared forecast errors' in the first step of the algorithm in Section 3.1. For EBR_t^{EV} we compute the variance of the forecast errors over the rolling windows $[t-N, t]$ and linearly weight the quartile of agents with the lowest variance. The results are summarized in Table 5. We find that for our benchmark measure and for EBR_t^{ME}, p-values easily reject the null for both maturities, while EBR_t^{EV} is not superior to the consensus, at least statistically. Expanding on this finding, we measure the similarity between the groups of agents selected by the different accuracy metrics. In order to do this, for each metric, we compute how frequently agents appear in the set of good forecasters, relative to the number of times they appear in the dataset, and then we compute the correlation between vectors of occurrence frequencies. Table 6 shows that, for the 10 year bond, the similarity between the agents in $EBR_t^\star$ and EBR_t^{ME} is 64% while the correlation for $EBR_t^\star$ and EBR_t^{EV} is 21%, and there is 29% correlation between the agents in EBR_t^{ME} and EBR_t^{EV}. These results suggest that agents with the smallest mean errors also tend to have low error variances, even if sorting directly on error variances does not produce significantly lower MSE than the consensus. This could be because forecasters display more heterogeneity in mean errors than in errors variances. In fact, unconditionally, the error variance is not significantly different

across quartiles of the cross-sectional distribution of forecasters (see Section 2 and Table 6 of the OA). We expand on the properties of the first two moments of the cross-sectional distribution of errors and on the characteristics of the $EBR_t^\star$ agents in Section 3.3 below.

[Insert Tables 5 and 6 here]

We study the robustness of $EBR_t^\star$'s outperformance along three dimensions.[9] As noted above, in our data, the median forecast (EBR_t^{Med}) is statistically more accurate than an equally weighted average forecast (EBR_t^c). A natural question is whether $EBR_t^\star$ outperforms EBR_t^{Med}. Table 7 addresses this question by repeating the DM tests from above with the alternative median benchmark. We find that for both bond maturities $EBR_t^\star$ produces statistically lower MSEs with a large degree of statistical confidence. The following row of Table 7 asks whether an alternative version of $EBR_t^\star$ computed from equal weights, as opposed to linear weights, outperforms EBR_t^{Med}. Comparing rows two and three, we see that an equally weighted $EBR_t^\star$ generates larger MSEs than a linearly weighted $EBR_t^\star$. However, the equally weighted version still outperforms EBR_t^{Med} and for the 10-year bond the difference is statistically significant at the 10% level.[10] The final rows in the table questions the importance of using a relative versus an absolute measure of past accuracy, by computing a 'non-ranked $EBR_t^\star$' (EBR_t^{NR}), where the weights are linearly increasing in the median of the level of past squared errors rather than the median squared error percentiles. We show that EBR_t^{NR} produces clearly larger MSEs than $EBR_t^\star$ for both bond maturities, that are not significantly different from the benchmark EBR_t^{Med}.

Taken together, these findings suggest that post trimming decision makers should overweight forecasters with historically good relative performance records but should retain a degree of diversification in aggregation.

[Insert Table 7 here]

[9] Additional robustness tests are provided in Section 4.1. of the OA.

[10] The equally weighted version of $EBR_t^\star$ is statistically more accurate than the equally weighted consensus at the 5% level or below. Results available on request.

Panel (a) of Figure 6 displays the dynamics of $EBR^{\star}$ for 5-year and 10-year bond maturities. Eyeballing the dynamics, we see that subjective expected excess bond returns appear countercyclical. In Section 4 that follows we study this claim more formally. Panel (b) of Figure 6 displays the difference between $EBR_t^{\star,n}$ and EBR^c. The figure shows that the spread between the two measures can be very large. In late 1999, immediately before the burst of the internet bubble, the spread exceeded 200 basis points for the 10 year bond. We observe similarly large positive spikes throughout the sample. Large negative spreads are less frequent and, on average, the subjective bond return implied by $EBR_t^{\star}$ is larger than that implied by the consensus.

[Insert Figure 6 here]

In summary, we have demonstrated that a simple approach that gives greater weight to historically good forecasters generates a subjective bond return expectation which is economically different and significantly improves on the simple equally weighted average forecast.

3.3. Are $EBR^{\star}$ agents skillful or lucky?

Our sample period is characterized by a downward trend in interest rates and persistent periods of high and low interest rate volatility (see Figure 14 in the OA). Therefore, one could be suspicious of the fact that the outperformance of our $EBR_t^{\star}$ with respect to the consensus is an artefact of the sample and that the $EBR_t^{\star}$ agents were just lucky rather than skillful.

A common approach used in the fund management literature to distinguish skill from luck is to study persistence in funds α (see, for example, Fama and French (2021)). Our dataset allows to address the luck versus skill question in an analogous fashion. We define a forecaster as 'skilful' if the accuracy in their forecasts is persistent and regime independent. In Section 2.2, we presented a test of the null of 'no-skill' based on persistence in accuracy (annual mean squared forecast errors) and Table 2 demonstrated a strong rejection. Unfortunately, it is extremely challenging to design a direct formal test of skill versus

all plausible different permutations of luck.[11] Instead, we follow an alternative approach that exploits the disaggregated nature of our dataset. We decompose the MSEs of agents selected in $EBR_t^\star$ in mean prediction errors and error variance. Figure 7 shows the cross-sectional distribution of mean prediction errors (top panel) and error variances (bottom panel) over time, for the 10-year bond excess return.[12] The mean and variances are computed over rolling periods of 5 years, which are equivalent to the look back periods used to select the $EBR_t^\star$ agents, denoted by red dots in the figure and in the x-axis we report the timing of the realization of the forecast errors. Focusing on the top panel we note that, in the second part of the sample, when the cross section of mean errors is almost completely positive, $EBR_t^\star$ selects agents with forecast errors that are persistently closer to zero on average. In fact, most red dots lie below the solid blue line, which denotes the cross-sectional mean error of the consensus.

Over the full sample, more than 78% of the $EBR_t^\star$ agents produce lower mean errors than the consensus. When we split the sample before and after January 2006, we find that the fraction of $EBR_t^\star$ agents with a mean error below the consensus is on average 90% after January 2006 (see also Figure 12 in the OA). Before January 2006, the distribution of mean errors for the $EBR_t^\star$ agents is slightly wider and a few of these agents are above the cross-sectional mean. Nonetheless, the outperformance is persistent over time. Moreover, we find that over the full sample $EBR_t^\star$ does not necessarily select the agents who are most pessimistic about yields.

The bottom panel of Figure 7 shows that $EBR_t^\star$ agents also tend to have persistently lower error variance, i.e. higher forecast precision, than the consensus. The fraction of $EBR_t^\star$ agents with an error variance below the consensus is on average 72%. This effect is particularly strong in the first and in the last few years of the sample, when almost all red dots in the plot lie below the blue line and in fact this fraction is close to 85%. In the central part of the sample, between 2009 and 2011, the cross-sectional distribution of error variance for all agents is very narrow and most agents which are part of the survey

[11]We thank an anonymous referee for helpful comments.
[12]The corresponding figure for the 5-year bond is displayed in the OA and shows similar results.

produce low error variance. During this subsample, mean squared forecast errors are mainly driven by squared mean errors, instead of error variance, and our aggregation approach rightly selects agents with the lowest mean error in expected yields. Overall, these findings suggest the agents selection approach which is part the construction of $EBR_t^\star$ identifies agents that are persistently more accurate than others (relatively skilled).

[Insert Figure 7 here]

3.4. A Bayesian approach to forecast aggregation

As an alternative to statistical inference based on DM test statistics, we consider how a Bayesian decision maker might have exploited persistence in past accuracy. Suppose there are two types of forecasters, $i = a$ and b and an asset manager who observes their forecasts and forms beliefs about future bond returns to decide her asset allocation. Let us assume that type a has shown better (past) accuracy in predicting bond returns than type b but the prior beliefs of the manager is that the two forecasters are equivalent in terms of skill. Given the observed accuracy, the posterior probabilities can deviate from 50%. Thus, in aggregating the beliefs of the forecasters, the asset manager may assign a larger weight to the beliefs of agent a. This approach of weighting different forecasting models by their posterior probabilities is called Bayesian model averaging (BMA) and is used in a finance setting for example by Avramov (2002) to predict stock returns and by Della Corte, Sarno, and Tsiakas (2009) to empirically evaluate the predictive ability of exchange rate models.[13]

Following this methodology we split the sample of professional forecasters in $N = 4$ sets based on past accuracy. Group Q_1 corresponds to the quartile of agents who have been more accurate in the past, group Q_2 is the second quartile, Q_3 is the third, and group Q_4 is the quartile of agents who have been less accurate. In order to be consistent with our $EBR^\star$ construction, accurate agents are identified based

[13]Baks, Metrick, and Wachter (2001) use a similar Bayesian argument to estimate the expected alpha (managerial skills) of mutual fund managers. Section 4.2. of the OA provides more details on the BMA approach.

on their ranked accuracy percentiles, over five years rolling windows, and linearly weighted. Therefore group Q_1 corresponds to $EBR^\star$. For the prior distribution, we assume that all groups have identical skill and assign them equal weights $1/N = 25\%$. We then compute the Bayesian posterior probability of the groups, i.e. their BMA weights, given their observed forecast errors as follows:

$$p(Q_i|data) \approx \frac{\exp\left(-0.5BIC_i\right)}{\sum_{j=1}^{N} \exp\left(-0.5BIC_j\right)}, \tag{5}$$

where the Bayesian information criterion (BIC) for each set of forecasters is only a function of their prediction errors.

Figure 8 displays Bayesian weights (the posterior probabilities) assigned to each quartile. We find the asset manager would have assigned weights of approximately 25% to each quarter for the first 5-years in our sample. Beyond this date, she sharply increased her weight on the Q_2 and Q_3 forecasts, while holding the weight on the Q_1 forecast relatively constant and low. In other words, a Bayesian would have first learned to discard the worst performing forecasters. Post dot-com bubble the manager was assigning around 50% of their weight on the Q_2 forecasters who historically were outperforming the median and this weight increased to around 90% in 2005. Beyond that point the weight on $Q_1 = EBR^\star$ jumped above 80% and beyond 2010 remained close to 100% until the end of our sample.

This Bayesian approach provides an alternative statistical validation of our aggregation procedure and provides additional evidence that heterogeneity and persistence in accuracy is strong enough to optimally deviate from 1/N weights, i.e. the consensus approach. We also note how our approach relates to existing contributions. First, Diebold and Pauly (1990), in the spirit of this subsection, propose improving on consensus forecasts by estimating a Bayesian weighted average of an equal weights prior and least squares projection weights. Second, Aiolfi and Timmermann (2006) show that competing forecasting models display persistent periods of out-performance but also switch, that is, at times the best models become

the worst and vice versa. This is related to the persistence in accuracy documented in Section 2 and the selection approach of this section that considers rolling relative accuracy as a sorting variable. Third, a decision to assign a weight of zero to the least accurate forecasts is consistent with Diebold and Shin (2019), who study a LASSO-based procedure which suggests the vast majority of forecasters should be discarded and the remainder should be averaged.

[Insert Figure 8 here]

3.5. How different is $EBR^{\star}$ from statistical measures?

We now compare the dynamics of $EBR^{\star}$ to a set of statistical models of expected excess returns that are used in the literature. We consider four alternative specifications: a random walk model for yields (RW);[14] a historical mean expected excess bond return (μ^H); a slope of the yield curve implied forecast ($SLOPE$), defined as the spread between the n-year and the 1-year yield. This forecast requires computing the factor loading on the slope, which we estimate in real-time using only historical information and an expanding window which starts in June 1961, to avoid any look-ahead bias. The same expanding windows are used to compute the average excess return for μ^H. The fourth benchmark is developed by Bianchi, Büchner, and Tamoni (2020) and based on a non-linear predictive methodology that uses machine-learning (ML). Their expected bond returns are based on the predictions implied by a neural network with one hidden layer and three nodes using only forward rates ("NN 1 Layer Group Ensem + fwd rate net", in their classification).[15]

Table 8 shows that survey-implied expected returns and statistical models have different properties.[16] The first difference is persistence. The autocorrelation coefficient of the 10-year expected excess returns is 0.98 for the random walk and slope model and 0.95 for the machine learning model, while it is only

[14] The expected excess bond return under this model is obtained as in Equation (1) but with $E_t^i\left[y_{t+12}^{n-1}\right]$ substituted with y_t^{n-1}, the current yield.

[15] We thank Andrea Tamoni for providing us with the machine learning forecasts.

[16] The sample period considered is 1993.12 - 2017.1 for expectations and 1994.12 - 2018.1 for realizations which is determined by the availability of the ML forecasts.

0.73 for $EBR_t^{\star,n}$. The second difference relates to their dynamics. The correlation between $EBR^\star$ and the statistical models is positive but modest. For the 10-year bond, the correlation is between 0.24 (ML) and 0.45 (RW). To the extent that the $SLOPE$ is a counter-cyclical risk premium proxy, this implies that $EBR^\star$ is also counter-cyclical. Moreover, the correlation with the $SLOPE$, which is equal to 0.34 for the 10-year bond, is relatively large but far from perfect, which is consistent with the idea that survey forecasters exploit information beyond a simple slope projection and potentially base their forecasts on information beyond the term structure. The third most noticeable difference is the mean excess return. For the 10-year bond the average $EBR^\star$ is close to zero, equal to 23 basis points, while the statistical projections imply much larger mean excess returns ranging from 2.96% (ML) to 5.50% ($SLOPE$). Thus, while $EBR^\star$ significantly outperforms the consensus in predicting future realized excess bond returns, it still displays a positive mean error. Indeed, even the most accurate agents were not able to fully capture the extent of the decline in yields that we have witnessed over our sample period.

[Insert Table 8 here]

With the exception of the random walk forecast, the statistical models require parameter estimation including a constant which can capture a trend in interest rates. In comparing accuracy of statistical models versus surveys, we follow the forecast aggregation literature in relaxing the unbiasedness assumption (see e.g. Granger and Ramanathan (1984)) and consider an extended version of $EBR_t^\star$ that includes a constant. In order to avoid any look-ahead bias, the constant correction is computed in real time. Namely, the extended version of $EBR_t^\star$ is equal to our benchmark $EBR_t^{\star,n}$ plus its average forecast error up to time t, starting with an initial window of 5 years and expanding. We denote the extended version $EBR_t^{\star,n}$-ext.[17] The two panels in Figure 9 display the time series of our benchmark $EBR_t^{\star,n}$ versus $EBR_t^{\star,n}$-ext. Consistent with the trend in yields, the constant correction is small initially but quickly widens and stabilizes to around two percentage points. The correction leads to a significant improvement

[17] Figure 11 in the OA compares the dynamics of $EBR_t^{\star,n}$-ext versus the statistical benchmarks.

in performance. Over the overlapping sample, i.e. November 1999 to January 2020, the MSEs of $EBR_t^\star$-ext are 15.38 and 51.18 for the 5 and 10-year bond maturities, respectively, against 20.97 and 68.30 for $EBR_t^\star$.

[Insert Figure 9 here]

Table 9 compares the predictive performance of the alternative models and shows that the MSE of $EBR^\star$-ext is approximately equal to the MSE of the SLOPE, slightly lower than the MSE of μ^H and slightly higher than ML. There are no statistically significant differences in accuracy. However, we note that the MSE of the RW forecast is lower than the MSE of $EBR^\star$-ext for both maturities. Interestingly, the best performer for 5-year bonds is the RW forecast, while the ML forecast is the best performer for 10-year bonds.

[Insert Table 9 here]

4. What Explains Subjective Expectations of Excess Bond Returns?

In this section we investigate the question of whether time variation in subjective expected excess bond returns can be understood in terms of time variation in compensation for risk.

4.1. Countercyclical subjective bond returns

Figure 6 suggests that $EBR_t^\star$ increases in periods of high market risk. This is particularly evident for the 10-year bond, which spikes in the aftermath of the Russian crisis and the collapse of Long-Term Capital Management, the 2001-2002 recession that followed the bursting of the dot-com bubble, and the 2008/2009 recession that followed the subprime mortgage crisis.

In order to evaluate more formally the countercyclicality of the $EBR_t^\star$ dynamics, we compute the correlation between $EBR_t^{\star,10}$ and the Chicago Fed National Activity Indicator ($CFNAI_t$), which is -0.18.

Moreover, a regression of $EBR_t^{\star,10}$ on CFNAI yields a significantly negative regression coefficient and R-squared of 3.5%. Considering survey expectations of GDP growth the correlation between $EBR_t^{\star,10}$ and an equally weighted consensus expectation $E_t[GDP_{t+12}]$ is -0.16, and a regression of $EBR_t^{\star,10}$ on $E_t[GDP_{t+12}]$ yields a significantly negative regression coefficient and R-squared of 2.5%. These observations are in line with investors requiring higher returns in bad states of the world.

4.2. Structural models

A long standing puzzle in asset pricing is that equilibrium models predict a mapping between variation in expected excess returns and quantities of risk. However, the relationship has proved difficult to detect empirically. We revisit this prediction by examining the link between $EBR^{\star}$ and a set of proxies for risk factors that arise in equilibrium models that generate time-varying bond risk premia:

$$EBR_t^{\star,n} = \alpha + \beta^{\top}\mathbf{X}_t + \varepsilon_{n,t}. \tag{6}$$

We consider four types of risk factor proxies to explain the dynamics of $EBR^{\star}$: (i) differences in beliefs; (ii) consumption surplus; (iii) economic uncertainty and (iv) bond volatility.[18] The empirical construction of $\mathbf{X}_t$ follows existing literature; our innovation rests on the alternative specification for the subjective expected excess bond return as the dependent variable.

Our sample for $EBR_{n,t}^{\star}$ is from December 1993 to January 2020, and for realized bond excess returns is from January 1994 to January 2021. To assess economic importance we standardize left and right hand variables, so that a 1-standard deviation change in the right hand variables implies a β-standard deviation in the dependent variable. Standard errors are reported in parentheses below the point estimates and calculated using a block bootstrap where the optimal block length is chosen in a data driven way

[18] Section 5 in the OA discusses the estimation of the factors, plots their time series and discusses their interpretation.

following Patton, Politis, and White (2009).[19] The top panel of Table 10 reports point estimates and test statistics.

[Insert Table 10 here]

The specification in row (i) studies differences in belief. We proxy for real disagreement ($DiB(g)$) and nominal disagreement ($DiB(\pi)$) using the 4-quarter ahead cross-sectional inter-quartile range in GDP and CPI forecasts from our survey dataset.[20] Consistent with the prediction of heterogeneous beliefs models, the slope coefficients of $DiB(g)$ and $DiB(\pi)$ are positive and significant at the 1% and 5% level, respectively, with an $\overline{R}^2$ of 24%. The positive sign of the slope coefficient supports the prediction of equilibrium models in which heterogeneous agents optimally trade on the basis of their beliefs: the greater the disagreement, the greater the trading among agents and therefore the quantity of risk that each agents holds in equilibrium. This induces a larger risk premium.

When agents have habit preferences, the price of risk is state-dependent and negatively related to the consumption surplus ratio.[21] To assess this link, in specification (ii) we follow Wachter (2006) and calculate consumption surplus ($Surp$) using a weighted average of 10 years of monthly consumption growth rates: $Surp = \sum_{j=1}^{120} \phi^j \Delta c_{t-j}$, where the weight is set to $\phi = 0.97^{1/3}$ to match the quarterly autocorrelation of the price-dividend ratio in the data. We find that the slope coefficient in this regression does have the correct sign but has a very low statistical significance and the $\overline{R}^2$ in the regression is also small.

Specification (iii) focuses on the significance of proxies of economic growth and inflation uncertainty, $UnC(g)$ and $UnC(\pi)$, as suggested by long-run risk models. To obtain a proxy for economic uncertainty we adapt the procedure of Bansal and Shaliastovich (2013). First, we use our survey data on consensus expectation of 4-quarter GDP growth and inflation and fit a bivariate VAR(1). In a second step, we compute a GARCH(1,1) process on the VAR residuals to estimate the conditional variance of expected

[19] The code for the automatic block selection is kindly provided Andrew Patton.

[20] For predictions linking differences in belief to bond markets, see Xiong and Yan (2010), Ehling, Gallmeyer, Heyerdahl-Larsen, and Illeditsch (2018) and Buraschi and Whelan (2020)

[21] Related studies include Buraschi and Jiltsov (2007), Campbell and Cochrane (1999) and Wachter (2006).

real growth and expected inflation. The loading on $UnC(g)$ is statistically significant at the 1% level and the $\overline{R}^2$ is 20%. The positive slope coefficient shows that subjective bond excess returns *increase* with real economic uncertainty. The coefficient on inflation uncertainty ($UnC(\pi)$) is not statistically significant.[22]

Specification (iv) studies the link between bond volatility and subjective bond expected excess returns. We study this link using two proxies for interest rate volatility: the intra-month sum of squared bond returns on a constant maturity n-year zero-coupon bond, which we denote as $\sigma_B^{(n)}$, and the 1-month implied 10-year maturity bond risk neutral volatility published by the CME (TYVIX). The regression results show that the quantity of risk channel is strongly positively related to $EBR_t^{\star,10}$: the R^2 is large, around 30%, and the coefficients on both physical and risk neutral volatility are significant at the 10% and 1% level, respectively. Importantly, the point estimates are also positive, consistent with the prediction that investors demand compensation for holding volatility risk.

To highlight the difficulty in detecting a relationship between expected excess returns and quantities of risk using traditional projection methods, the bottom panel of Table 10 shows regressions results using realized future returns $hprx_{t,t+12}$ as the dependent variable. The R^2 of these regressions are much smaller, and only one of the explanatory variables, TYVIX, is statistically significant. Moreover, the estimated coefficient on realized bond return volatility, $\sigma_B^{(n)}$, is not only insignificant but has the wrong sign.

A possible explanation for these findings is that projections based on future realizations can be quite different from survey-based expected returns. Finally, we note that the link between subjective expected bond returns and factors that proxy for discount variation is different to what has been found in equity markets, where financial ratios that proxy for risk compensation are negatively correlated with subjective expected returns (Greenwood and Schleifer (2014)).

[22]For related work on the link between macroeconomic volatility and bond risk premia see Gomez-Cram and Yaron (2021)

5. Conclusion

This paper studies aggregation of expectations from a cross-section of individual agent subjective yield curve forecasts from financial institutions that allows the estimation of heterogeneous subjective expected excess bond returns.

Belief heterogeneity about bond returns translates into a large dispersion in the distribution of mean squared forecast errors and we show a significant fraction of the cross-section outperform (underperform) consensus forecasts at conventional significance levels. More importantly, we show that heterogeneity in forecasting ability is persistent: forecasters who are good tend to remain good and forecasters who are bad tend to remain bad.

We show that persistence in accuracy can be exploited to construct a novel real-time weighting of individual agents beliefs, which outperforms the equally weighted consensus belief, a benchmark the literature has found difficult to beat. Our approach is simple to implement: we remove historically bad forecasters and form expectations from the remaining set of good forecasters.

Studying the properties of our aggregate subjective measure, we find support for rational determinants of expected bond returns. In particular, we find a strong correlation between quantity of risk factors and subjective bond returns, even if this relationship is difficult to detect using ex-post realizations.

Finally, we note the forecast combination scheme we propose can be applied to subjective expectations of macro quantities such as GDP growth or inflation. Comparing such measures within the same dataset one could ask questions related to whether being good at forecasting fundamentals translates to being good at forecasting asset prices. We leave this question for future research.

References

Aiolfi, M., and A. Timmermann (2006): "Persistence in forecasting performance and conditional combination strategies," *Journal of Econometrics*, 135(1-2), 31–53.

Andrade, P., R. K. Crump, S. Eusepi, and E. Moench (2016): "Fundamental disagreement," *Journal of Monetary Economics*, 83, 106–128.

Avramov, D. (2002): "Stock Return Predictability and Model Uncertainty," *Journal of Financial Economics*, 64, 423–458.

Baks, K., A. Metrick, and J. Wachter (2001): "Should investors avoid all actively managed mutual funds? A study in Bayesian performance evaluation," *Journal of Finance*, 56, 45–85.

Bansal, R., and I. Shaliastovich (2013): "A long-run risks explanation of predictability puzzles in bond and currency markets," *Review of Financial Studies*, 26, 1–33.

Bates, J. M., and C. W. Granger (1969): "The combination of forecasts," *Journal of the Operational Research Society*, 20(4), 451–468.

Bianchi, D., M. Büchner, and A. Tamoni (2020): "Bond risk premia with machine learning," *Review of Financial Studies (forthcoming)*.

Buraschi, A., and A. Jiltsov (2007): "Habit formation and macroeconomic models of the term structure of interest rates," *Journal of Finance*, 62, 3009 – 3063.

Buraschi, A., and P. Whelan (2020): "Sentiment, Speculation, and Interest Rates," *Management Science (forthcoming)*.

Campbell, J., and J. Cochrane (1999): "By force of habit: A consumption-based explanation of aggregate stock market behavior," *Journal of political Economy*, 107, 205–251.

Della Corte, P., L. Sarno, and I. Tsiakas (2009): "An Economic Evaluation of Empirical Exchange Rate Models," *The Review of Financial Studies*, 22, 3491–3530.

Diebold, F. X., and R. S. Mariano (1995): "Comparing predictive accuracy," *Journal of Business & economic statistics*, 20(1), 253–263.

Diebold, F. X., and P. Pauly (1990): "The use of prior information in forecast combination," *International Journal of Forecasting*, 6(4), 503–508.

Diebold, F. X., and M. Shin (2019): "Machine learning for regularized survey forecast combination: partially-egalitarian LASSO and its derivatives," *International Journal of Forecasting*, 35, 1679–1691.

Ehling, P., M. Gallmeyer, C. Heyerdahl-Larsen, and P. Illeditsch (2018): "Disagreement about inflation and the yield curve," *Journal of Financial Economics*.

Fama, E. F., and K. R. French (2021): "Luck versus skill in the cross-section of mutual fund returns," in *The Fama Portfolio*, pp. 261–300. University of Chicago Press.

Gomez-Cram, R., and A. Yaron (2021): "How important are inflation expectations for the nominal yield curve?," *The Review of Financial Studies*, 34(2), 985–1045.

Granger, C. W., and R. Ramanathan (1984): "Improved methods of combining forecasts," *Journal of forecasting*, 3(2), 197–204.

Greenwood, R., and A. Schleifer (2014): "Expectations of Returns and Expected Returns," *Review of Financial Studies*, 27(3), 714–746.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2006): "The U.S. Treasury Yield Curve: 1961 to the Present," *Federal Reserve Board Working Paper Series*.

Harvey, D., S. Leybourne, and P. Newbold (1997): "Testing the equality of prediction mean squared errors," *International Journal of Forecasting*, 13, 281–291.

Patton, A., D. N. Politis, and H. White (2009): "Correction to "Automatic block-length selection for the dependent bootstrap" by D. Politis and H. White," *Econometric Reviews*, 28(4), 372–375.

Patton, A. J., and A. Timmermann (2010): "Why do Forecasters Disagree? Lessons from the Term Structure of Cross-Sectional Dispersion," *Journal of Monetary Economics*, 57, 803–820.

Wachter, J. A. (2006): "A consumption-based model of the term structure of interest rates," *Journal of Financial Economics*, 79, 365–399.

Xiong, W., and H. Yan (2010): "Heterogeneous expectations and bond markets," *Review of Financial Studies*, 23, 1433–1466.

6. Tables

	Q_1	Q_2	Q_3	Q_4
		5-year Bonds		
Q_1	40^{***}	27^{*}	19^{***}	14^{***}
Q_2	28^{**}	29^{**}	26	17^{***}
Q_3	21^{**}	26	28^{**}	24
Q_4	17^{***}	19^{***}	26	38^{***}
		10-year Bonds		
Q_1	41^{***}	26	19^{***}	14^{***}
Q_2	27^{*}	30^{**}	25	18^{***}
Q_3	20^{**}	26	30^{**}	24
Q_4	17^{***}	19^{***}	26	38^{***}

Table 1. Belief transition probabilities

This table reports the year-on-year probability of a forecaster transitioning between the quartiles of the cross-sectional distribution of excess bond return forecasts. Units are percentages. Statistical significance is assessed under the null $Q_{ij} = 25\%$ against the alternative $Q_{ij} \neq 25\%$. One, two or three stars indicate significance at the 10% , 5%, and 1% levels, respectively. Sample period is 1988.1 - 2020.1.

	Q_1	Q_2	Q_3	Q_4
		5-year bond		
Q_1	51***	36***	12***	1***
Q_2	35***	36***	23	6***
Q_3	11***	25	41***	23
Q_4	2***	5***	26	67***
		10-year bond		
Q_1	49***	32	15***	3***
Q_2	31**	32**	26	11***
Q_3	15***	25	35***	24
Q_4	2***	10***	25	63***

Table 2. Accuracy transition probabilities

This table reports the year-on-year probability of a forecaster transitioning between the quartiles of the cross-sectional distribution of mean-square-errors. Units are percentages. Statistical significance is assessed under the null $Q_{ij} = 25\%$ against the alternative $Q_{ij} \neq 25\%$. One, two or three stars indicate significance at the 10% , 5%, and 1% levels, respectively. Sample period is 1988.1 - 2020.1.

	MSE	DM	p-value (%)
		5-year Bonds	
EBR^c	22.84		
BG	23.77	-1.28	89.85
GR	24.08	-0.34	63.41
		10-year Bonds	
EBR^c	80.25		
BG	82.14	-1.00	84.11
GR	80.17	0.01	49.66

Table 3. Forecast evaluation of MSE loss optimal weights

This table reports the mean-square-error (MSE), Diebold-Mariano (DM) forecast evaluation test statistic and the corresponding p-value testing that the MSE of the consensus forecast EBR^c is equal to the MSE of a forecast using optimal MSE loss weights, against the 1-sided alternative that the EBR^c is less accurate. The optimal weights are either computed using the diagonal covariance matrix version of Bates and Granger (1969) (BG) or from a least-squares projection as in Granger and Ramanathan (1984) (GR). Details of the numerical implementation are discussed in the OA. Sample period is 1993.12 - 2020.1 for expectations and 1994.12 - 2021.1 for realizations

	MSE	DM	p-value (%)
		5-year Bonds	
EBR^c	23.60		
$N = 24$	22.62	1.20	11.54
$N = 60$	21.35	2.83	0.25
$N = 120$	22.11	1.88	3.05
		10-year Bonds	
EBR^c	80.49		
$N = 24$	75.11	1.73	4.20
$N = 60$	70.97	3.33	0.05
$N = 120$	74.64	1.70	4.56

Table 4. $EBR^\star$ **vs** EBR^c **varying N**

This table displays mean square prediction errors (MSE), Diebold-Mariano (DM) forecast evaluation test statistics and corresponding p-values testing for equality of MSEs of the consensus forecast EBR^c versus $EBR^\star$ (N,Q) for N = 24, 60 and 120, and Q = 0.25. Sample period is 1993.12 - 2020.1 for expectations and 1994.12 - 2021.1 for realizations.

	MSE	DM	p-value (%)
		5-year Bonds	
EBR^{c}	22.84		
$EBR^{\star}$	21.20	2.34	0.99
EBR^{ME}	21.22	2.13	1.72
EBR^{EV}	22.90	-0.09	53.65
		10-year Bonds	
EBR^{c}	80.25		
$EBR^{\star}$	73.57	2.49	0.67
EBR^{ME}	72.67	2.44	0.76
EBR^{EV}	77.33	1.09	13.92

Table 5. Forecast evaluation tests of $EBR^{\star}$ computed from alternative accuracy metrics
EBR(N,Q) is computed for N = 60 and Q = 0.25 by sorting on past ranked (*i*) squared errors ($EBR^{\star}$), (*ii*) mean errors (EBR^{ME}), or (*iii*) error variances (EBR^{EV}). This table then reports the mean-square-error (MSE), Diebold-Mariano (DM) forecast evaluation test statistic and the corresponding p-value testing that the MSE of the consensus forecast EBR^{c} is equal to the MSE of $EBR^{\star}$, EBR^{ME} and EBR^{EV}, respectively, against the 1-sided alternative that the EBR^{c} is less accurate. Sample period is 1993.12 - 2020.1 for expectations and 1994.12 - 2021.1 for realizations.

	$EBR^{\star}$	EBR^{ME}	EBR^{EV}
$EBR^{\star}$	1.00	0.64	0.21
EBR^{ME}	0.64	1.00	0.29
EBR^{EV}	0.21	0.29	1.00

Table 6. Similarity of forecasters in $EBR^{\star}$ computed from alternative accuracy metrics
$EBR^{\star}$(N,Q) is computed for N = 60 and Q = 0.25 by sorting on past ranked (*i*) squared errors (benchmark $EBR^{\star}$), (*ii*) mean errors (EBR^{ME}), or (*iii*) error variances (EBR^{EV}). A frequency of occurrence for each agent is then computed from the number of times they appear in the EBR measure as a fraction of the number of times they appear in the dataset. This table reports the correlations between the vectors of frequency of occurrence for (*i*),(*ii*), and (*iii*). Sample period is 1993.12 - 2020.1 for expectations and 1994.12 - 2021.1 for realizations.

	MSE	DM	p-value (%)
		5-year Bonds	
EBR^{Med}	22.50		
$EBR^{\star}$	21.20	1.93	2.76
EBR^{EW}	21.66	1.13	12.94
EBR^{NR}	22.12	0.51	30.62
		10-year Bonds	
EBR^{Med}	78.78		
$EBR^{\star}$	73.57	2.18	1.51
EBR^{EW}	74.37	1.62	5.34
EBR^{NR}	77.47	0.45	32.50

Table 7. Forecast evaluation robustness

$EBR^{\star}$(N,Q) is computed for N = 60 and Q = 0.25 by sorting on past ranked MSEs. This table then reports the mean-square-error (MSE), Diebold-Mariano (DM) forecast evaluation test statistic and the corresponding p-value testing that the MSE of $EBR^{\star}$ is equal to the benchmark, against the 1-sided alternative that the MSE of $EBR^{\star}$ is smaller than the benchmark, where the benchmark is computed from the *median* forecast (EBR_t^{Med}). We also report this test for an alternative 'Equally weighted $EBR^{\star}$' (EBR_t^{EW}), which is based on the same sort as $EBR^{\star}$(N,Q) but with *equal weights* instead of linear weights, and a 'non-ranked $EBR^{\star}$' (EBR_t^{NR}), where the weights linearly increasing in the median of the level of past squared errors. Sample period is 1993.12 - 2020.1 for expectations and 1994.12 - 2021.1 for realizations.

	Mean	Std Dev	AR(1)	ρ
		$\underline{EBR^{\star}}$		
5-year	-0.13	1.29	0.61	
10-year	0.23	3.38	0.73	
		random walk		
5-year	1.91	1.32	0.97	0.45
10-year	3.03	1.74	0.98	0.37
		SLOPE		
5-year	3.04	1.84	0.98	0.42
10-year	5.50	4.08	0.98	0.34
		Machine Learning		
5-year	1.89	1.64	0.94	0.38
10-year	2.96	3.46	0.95	0.24

Table 8. Summary Statistics

Summary statistics of bond expected excess returns are computed from $EBR^{\star}$, from a random walk forecast, from a slope of the yield curve forecast and from the yields only machine learning forecast of Bianchi, Büchner, and Tamoni (2020). The sample period is considered is 1993.12 - 2017.1 for expectations and 1994.12 - 2018.1 for realizations which is determined by the availability of the ML forecasts.

	MSE	DM	p-value (%)
		5-year Bonds	
$EBR^{\star}$-ext	15.05		
RW	13.24	1.18	11.93
μ^H	15.11	0.03	48.74
$SLOPE$	14.76	0.11	45.50
ML	13.65	0.64	26.06
		10-year Bonds	
$EBR^{\star}$-ext	45.88		
RW	42.15	0.92	17.86
μ^H	52.75	1.51	6.65
$SLOPE$	40.85	0.53	29.78
ML	40.15	0.74	23.05

Table 9. $EBR^{\star}$-ext forecast accuracy comparisons against statistical projections
This table reports the mean-square-error (MSE), Diebold-Mariano (DM) forecast evaluation test statistic and the corresponding p-value testing that the MSE of the extended $EBR^{\star}$-ext is equal to the MSE of a given statistical model against the 1-sided alternative that the EBR^{c} is less accurate. We consider 4 real-time statistical projections: a random walk (RW), a constant expected return forecast (μ^H), a slope of the yield curve forecast ($SLOPE$), and the yields only machine learning forecast (ML) of Bianchi, Büchner, and Tamoni (2020). Sample period is 1998.12 - 2017.1 for expectations and 1999.12 - 2018.1 for realizations.

	$DiB(g)$	$DiB(\pi)$	$Surp$	$UnC(g)$	$UnC(\pi)$	$\sigma_B^{(n)}$	$TYVIX$	$\overline{R}^2$ (%)
				Panel A: 10-year $EBR_t^{\star,10}$				
(i)	0.32^{***}	0.25^{**}						24.03
	(0.10)	(0.11)						
(ii)			-0.17					2.66
			(0.21)					
(iii)				0.46^{***}	0.01			20.47
				(0.19)	(0.22)			
(iv)						0.17^{*}	0.46^{***}	30.21
						(0.09)	(0.13)	
				Panel B: 10-year $hprx_{t,t+12}$				
(i)	-0.06	0.18						2.72
	(0.12)	(0.13)						
(ii)			-0.16					2.17
			(0.14)					
(iii)				0.19	0.13			8.11
				(0.18)	(0.16)			
(iv)						-0.00	0.21	3.93
						(0.14)	(0.18)	

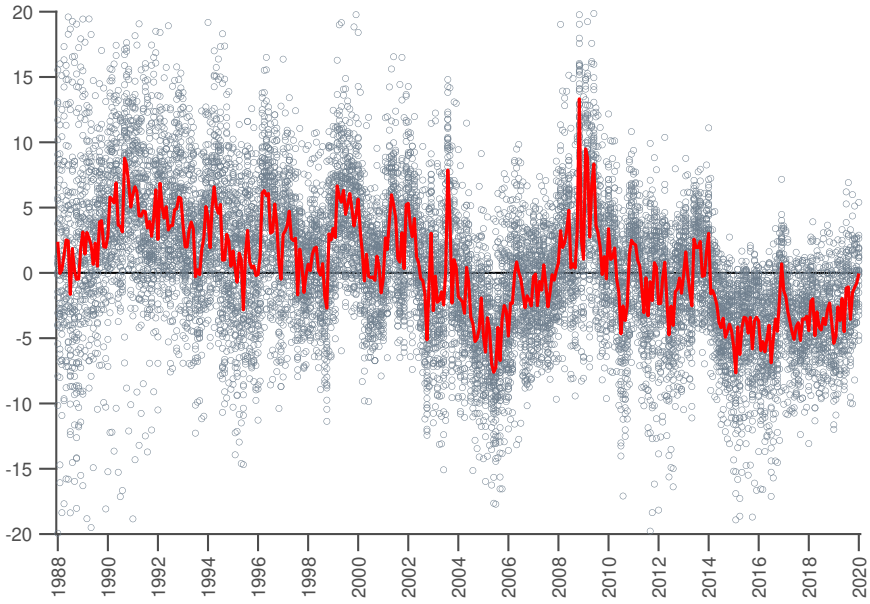
Table 10. Compensation for Risk

Panel A shows estimates from regressions of the subjective expected excess returns on 10-year bonds on a set of explanatory variables:

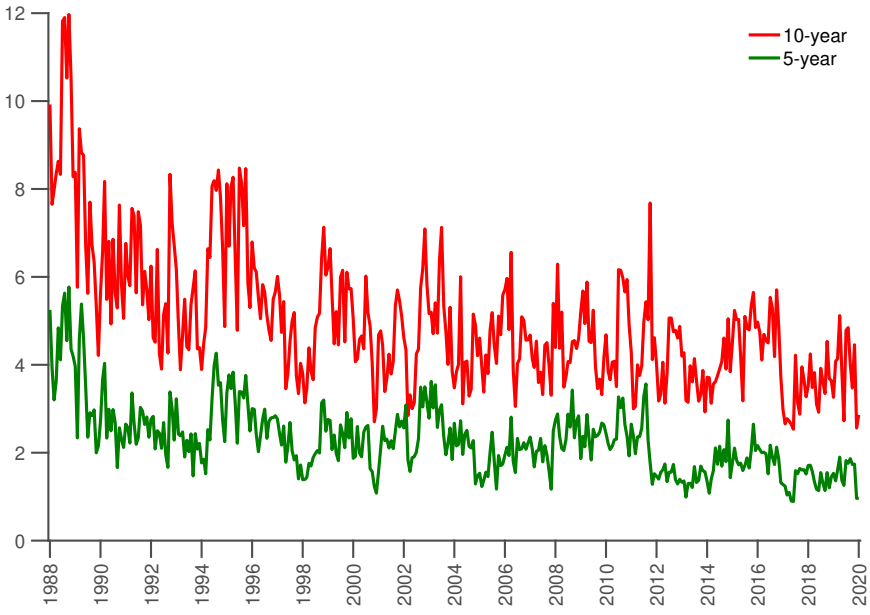
$$EBR_{10,t}^{\star} = \alpha + \beta^{\top}\mathbf{X}_t + \epsilon_{10,t}.$$

These factors are discussed in detail in the main body of the paper and all variables are standardized. Panel B shows the results of the corresponding regressions using the ex post realized excess return, $hprx_{t,t+12}$, as dependent variable. Standard errors are reported in parentheses below the point estimates and are calculated using a block bootstrap. Superscripts $*$, $**$ and $***$ denote statistical significance at 90%, 95% and 99%, respectively, based on block bootstrap confidence intervals. Adjusted R-squared of the regressions are reported in the last column. Sample period is 1993.12 - 2020.1

7. Figures



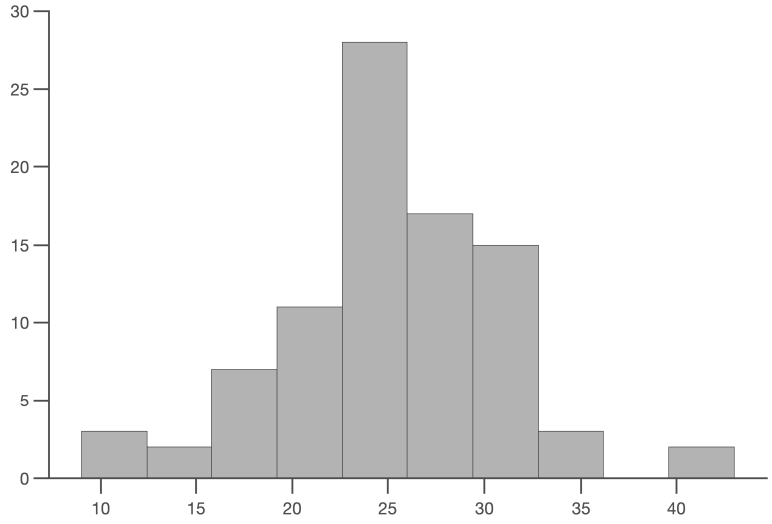
(a) 10-year excess bond return forecasts



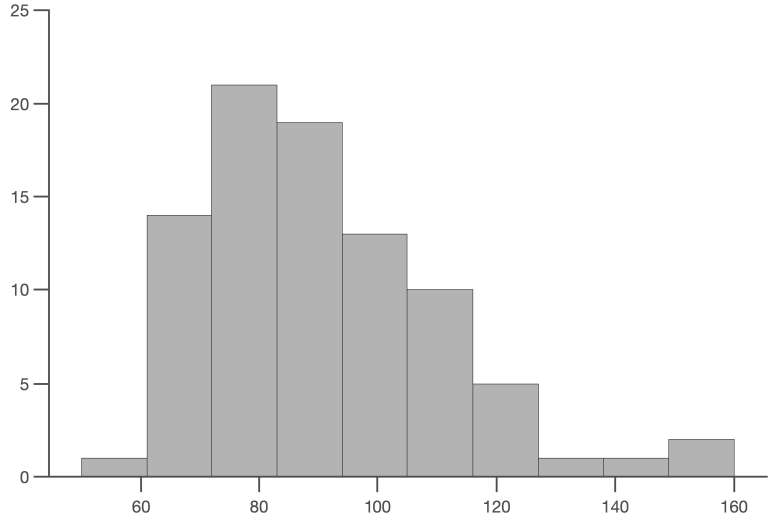
(b) disagreement

Figure 1. Subjective expectations

The top panel displays the cross-sectional distribution of one year subjective excess bond returns for 10-year zero coupon bonds. The red solid line indicates the equally weighted mean (consensus) forecast. The bottom panel plots disagreement returns defined as the cross-sectional interquartile range in subjective expectations. Sample period is 1988.1 - 2020.1



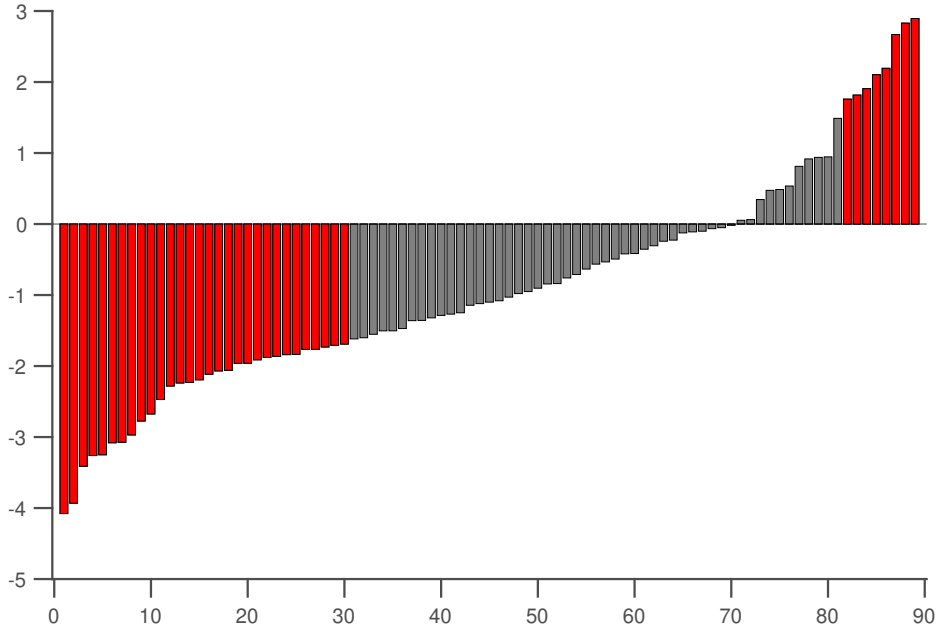
(a) 5-year



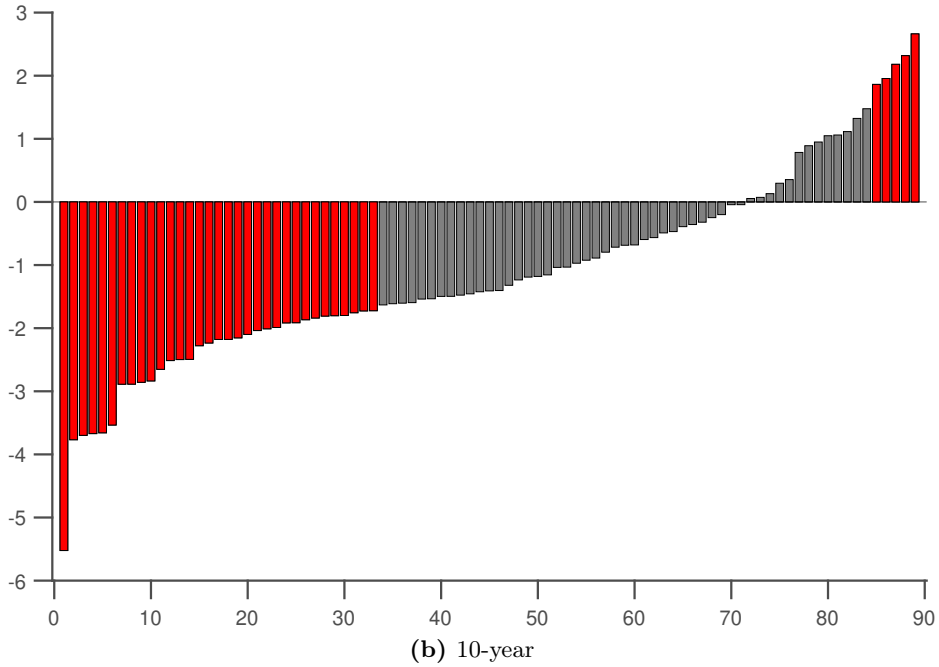
(b) 10-year

Figure 2. Distribution of MSEs

Histograms of individual forecaster mean squared errors. We consider only the contributors with at least 60 months of forecasts. Note that since our panel of forecasters is unbalanced errors are realized over different sample periods. Sample period is 1988.1 - 2020.1 for the expectations and 1989.1 - 2021.1 for subsequent realizations.



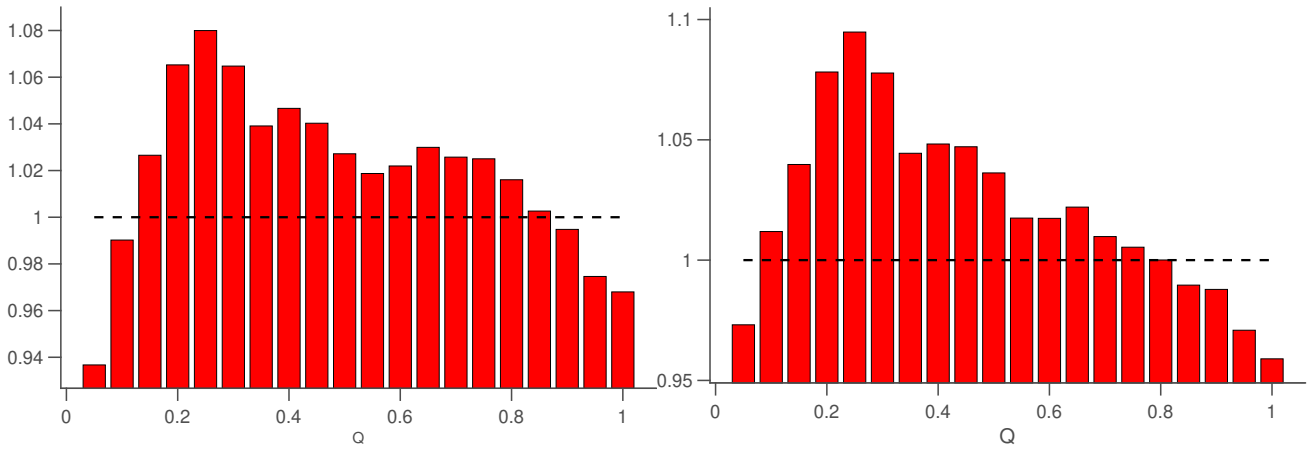
(a) 5-year



(b) 10-year

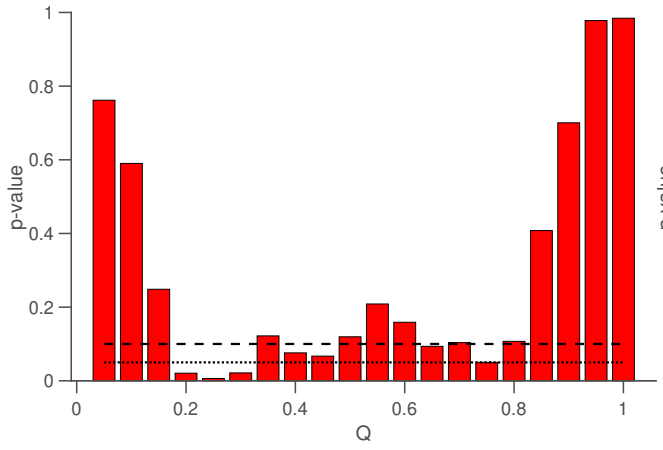
Figure 3. Cross-section of DM statistics

These figures plot the cross-section of individual forecaster DM test statistics. Positive values indicate outperformance with respect to the consensus. Red bars indicate statistically significance at the 5% level where the variance of the loss differential is calculated taking into account autocorrelation in the errors induced by the 11 overlapping forecasts. Sample period is 1988.1 - 2020.1 for the expectations and 1989.1 - 2021.1 for subsequent realizations.

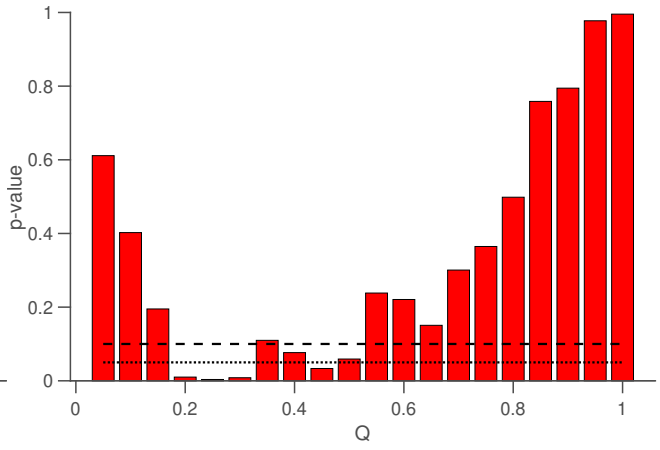


(a) 5-year Ratio of MSEs

(b) 10-year Ratio of MSEs



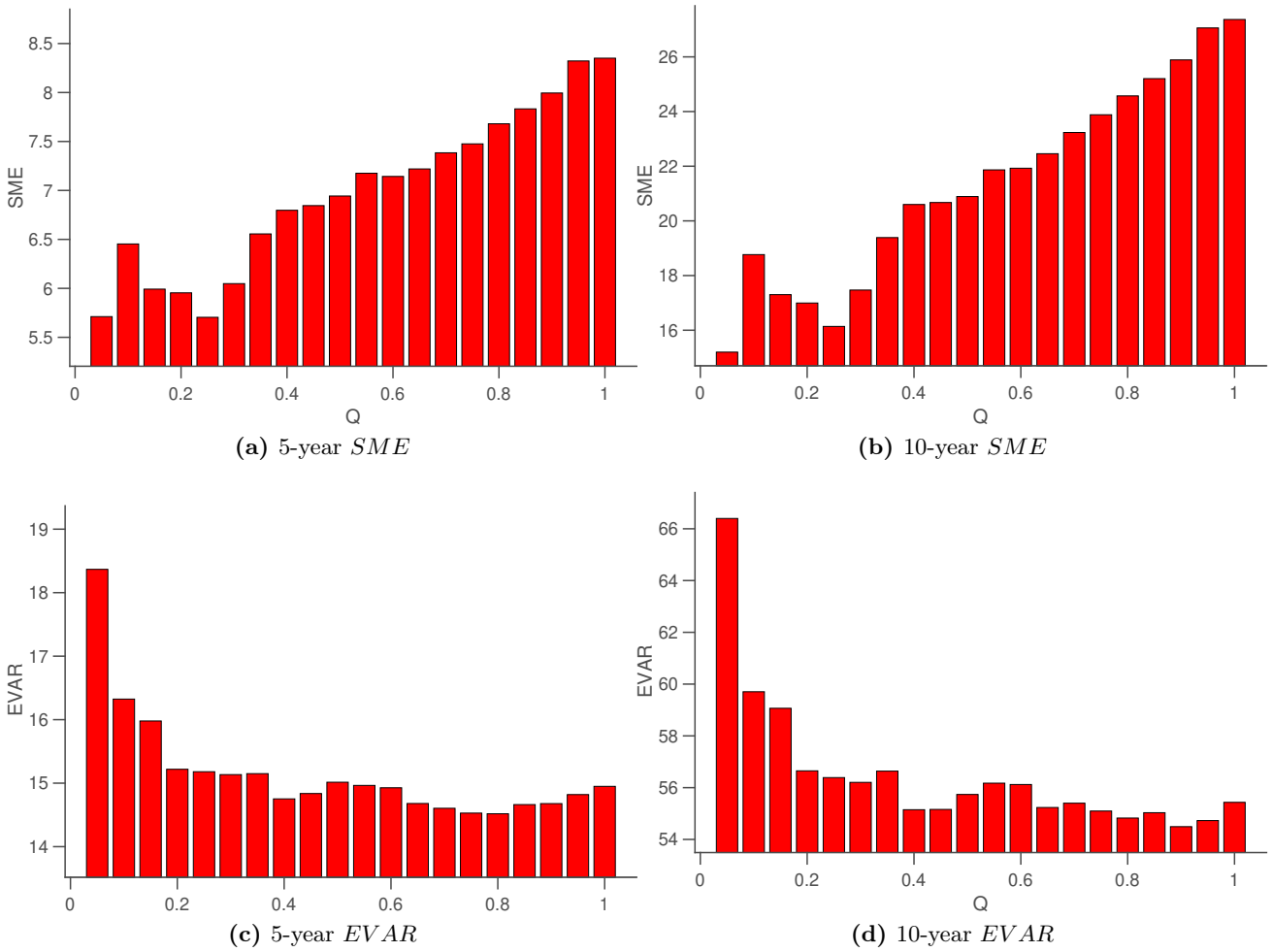
(c) 5-year p-value



(d) 10-year p-value

Figure 4. $EBR^{\star}$ **varying** Q**: MSE ratio and DM p-values**

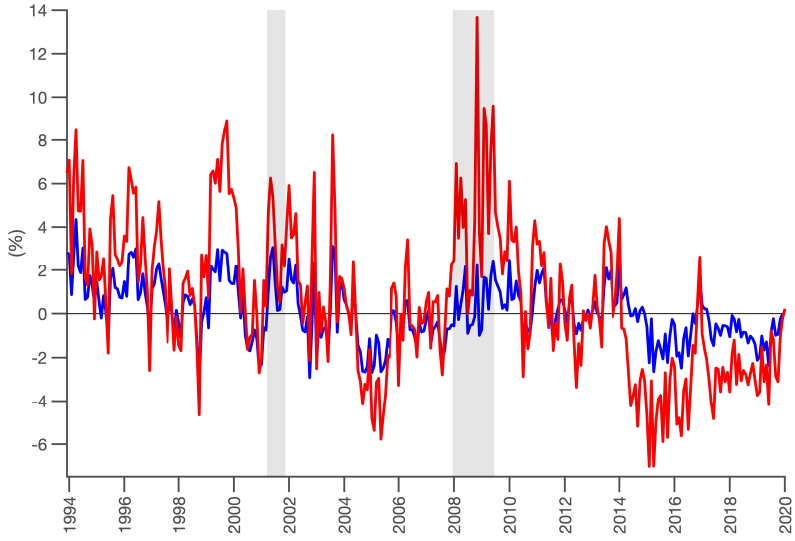
We compute $EBR^{\star}(N, Q)$ fixing $N = 60$ and varying $Q \in [0.05\ ,\ 1.00]$. The top panels plot the ratio of the consensus (EBR^{c}) MSE to the MSE of $EBR^{\star}$ and the bottom panels plot p-value of the Diebold-Mariano (DM) forecast evaluation test statistic testing for equality of MSE's of the consensus forecast EBR^{c} versus $EBR^{\star}$ against the one-sided alternative that $EBR^{\star}$ is more accurate. Dashed lines in the bottom plot indicate the 5% and 10% levels. Sample period is 1993.12 - 2020.1 for the expectations and 1994.12 - 2021.1 for subsequent realizations.



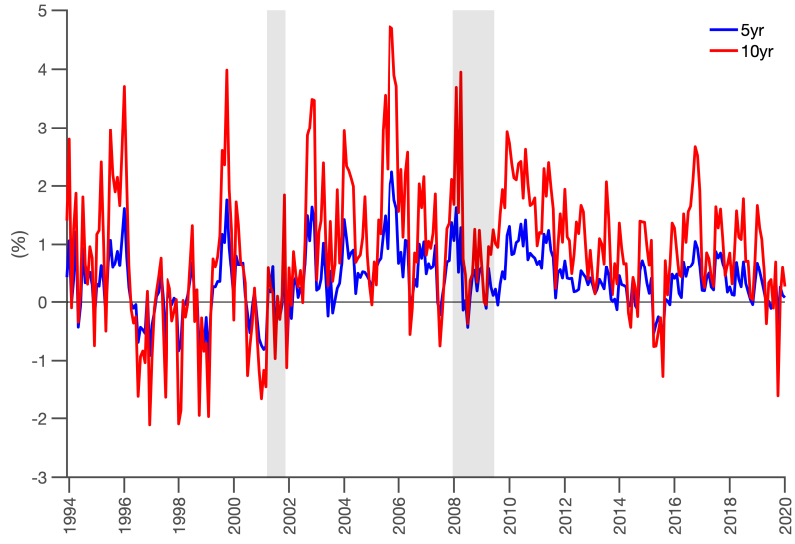
(a) 5-year SME **(b)** 10-year SME

(c) 5-year $EVAR$ **(d)** 10-year $EVAR$

Figure 5. $EBR^{\star}$ **varying** Q**: SMEs and EVARs**
We compute $EBR^{\star}(N, Q)$ fixing $N = 60$ and varying $Q \in [0.05\ ,\ 1.00]$. For each value of Q, the top panels plot associated squared mean errors and the bottom panels plot error variances for 5-year and 10-year bond maturities. Sample period is 1993.12 - 2020.1 for the expectations and 1994.12 - 2021.1 for subsequent realizations.



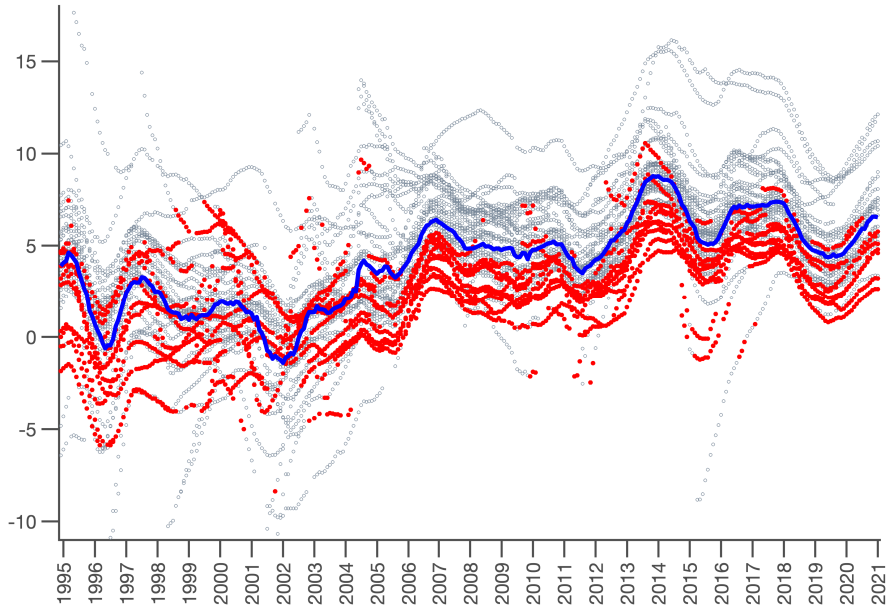
(a)



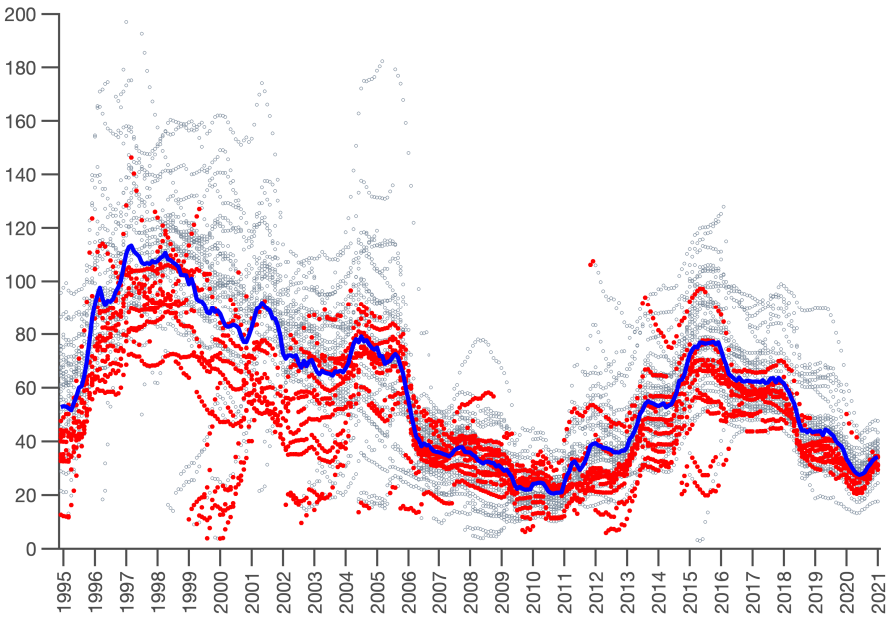
(b)

Figure 6. $EBR^{\star}$ vs EBR^{c}

The top panel plots $EBR^{\star,n}(N,Q)$ for $Q = 0.25$ and $N = 60$ for 5-year and 10-year bonds. The bottom plot shows the difference between the subjective expected excess bond returns of $EBR^{\star}(N,Q)$ and consensus expectations (EBR^{c}), for bond maturities of 5 and 10 years. Sample period is 1993.12 - 2020.1 for the expectations.



(a)



(b)

Figure 7. Cross-Sectional distribution of mean and variance of errors: 10-year bond
This figure plots the cross-sectional distribution of mean-errors (panel a) and error variances (panel b) computed on 5-year rolling windows. The solid blue line in each plot is the equally weighted average (consensus) at each point in time and the red-dots represent $EBR^{\star}$ agents. The figure plots the results for the the 10-year bond (the corresponding figure for the 5-year bond is displayed in the OA). On the x-axis we report the time of the realization of forecast errors.

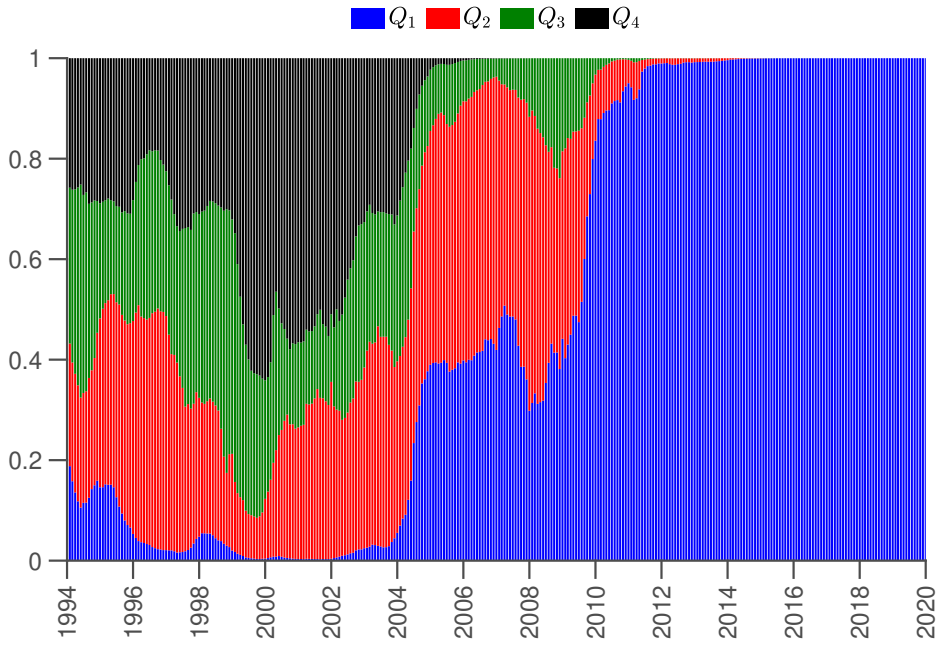
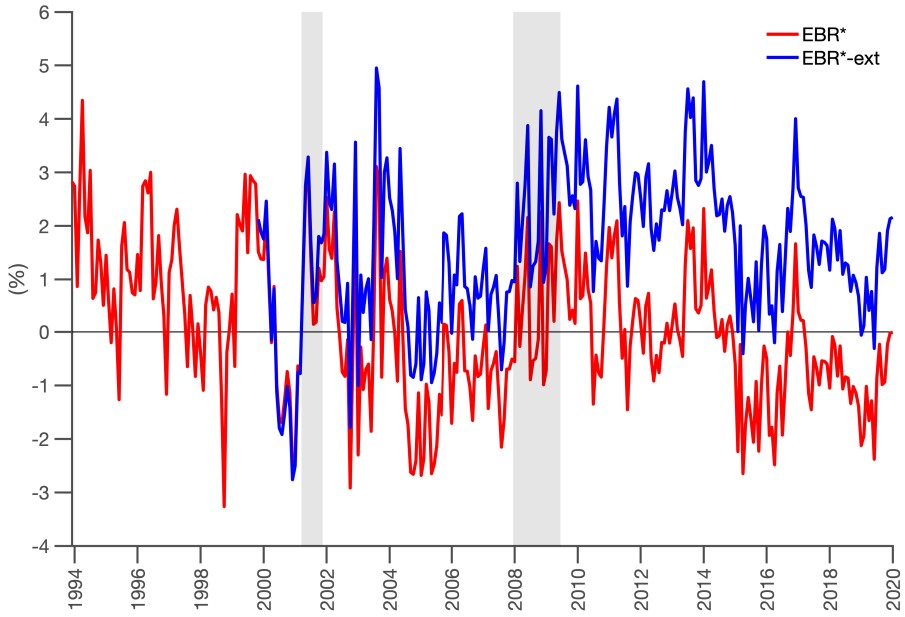
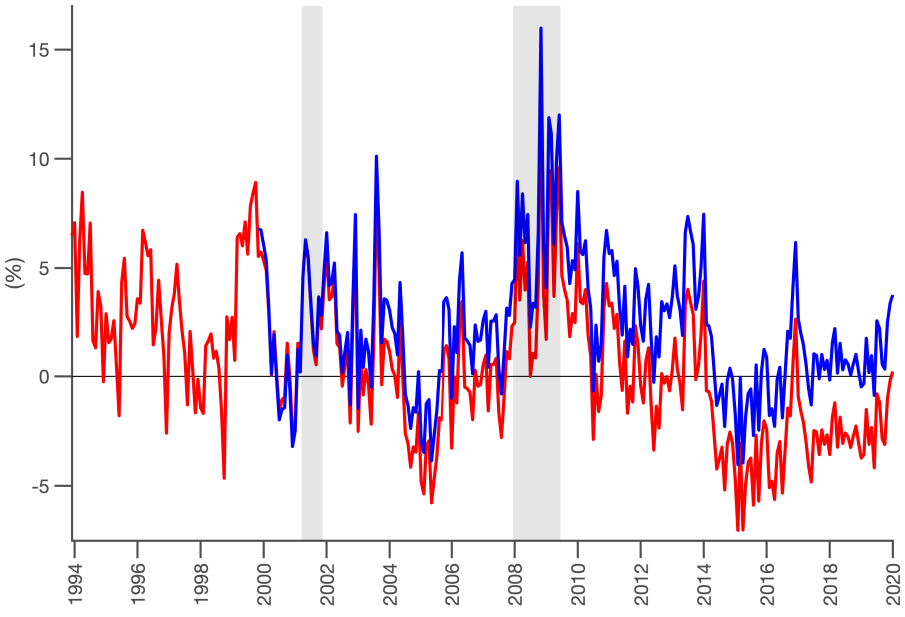


Figure 8. Bayesian exercise

This figure displays the time-series of Bayesian posterior weights assigned to quartiles of linearly weighted (so that weights sum to one within each quartile) beliefs which are formed from past historical mean-square errors. Group $Q1$ corresponds to the quartile of agents who have been more accurate in the past, group $Q2$ is the second quartile, $Q3$ is the third, and group $Q4$ is the quartile of agents who have been least accurate. The Bayesian updating is discussed in the main body of the text and additional detailed are reported in the OA. Sample period is 1993.12 - 2020.1 for the expectations and 1994.12 - 2021.1 for subsequent realizations.



(a) 5-year Bond



(b) 10-year Bond

Figure 9. $EBR^{\star}$ vs $EBR^{\star}$ extended

We propose an alternative version of $EBR_t^{\star}$ that takes into account the positive sample average forecast errors in individual forecasts by including a time-varying constant adjustment computed in real-time. The extended version of $EBR_t^{\star}$ is equal to our benchmark $EBR_t^{\star}$ plus its average forecast error up to time t, starting with an initial window of 5 years and expanding. This figure displays the time series of our benchmark $EBR_t^{\star}$ versus $EBR_t^{\star}$-extended, for a 5 and 10-year bond, respectively.