Predicting bond return predictability[*]

Daniel Borup[†] Jonas N. Eriksen[‡] Mads M. Kjær[§] Martin Thyrsgaard[‖]

This version: January 6, 2020

[*] We thank Francis X. Diebold, Anders B. Kock, Bent Jesper Christensen, Kim Christensen, Christian M. Dahl, Tom Engsted, Stig V. Møller, Johan S. Jakobsen, and participants at the Econometrics Workshop at University of Pennsylvania (2017) and research seminars at Aarhus University (2019) for many constructive comments and suggestions. Jonas N. Eriksen acknowledges support from the Danish Council of Independent Research (DFF 7024–00020B). Part of this paper was previously circulated under the title "Statistical tests for equal predictive ability across multiple forecasting methods".

[†] CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark, and the Danish Finance Institute (DFI). Email: dborup@econ.au.dk. Corresponding author.

[‡] CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark, and the Danish Finance Institute (DFI). Email: jeriksen@econ.au.dk.

[§] CREATES, Department of Economics and Business Economics, Aarhus University, Fuglesangs Allé 4, DK-8210 Aarhus V, Denmark. Email: mads.markvart@econ.au.dk.

[‖] Kellogg School of Management, Northwestern University, 2211 Campus dr, Evanston, IL 60208, and CREATES, Aarhus University. Email: martin.thyrsgaard@kellogg.northwestern.edu.

Predicting bond return predictability

Abstract

We document predictable shifts in bond return predictability related to economic activity and uncertainty in the U.S. Treasury bond market using standard bond excess return predictors. Bond returns are predictable in high (low) economic activity (uncertainty) states, but not in others. We develop a new test for equal conditional predictive ability among two or more forecasting methods and show that relative performances are predictable and exploitable in a real-time forecasting setting. Using a novel forecast combination scheme with dynamic trimming based on predicted forecasting performance leads to strongly countercyclical out-of-sample risk premia estimates and substantial gains in predictive accuracy and economic value.

Keywords: Bond excess returns, forecasting, state-dependencies, multivariate test, equal conditional predictive ability

JEL Classification: C12, C52, E44, G12

First version: January 4, 2020

This version: January 6, 2020

1. Introduction

We study time variation in bond return predictability and document predictable shifts related to economic activity and uncertainty. Existing evidence on bond return predictability has mostly been established using linear predictive regressions designed to assess whether bond excess returns are predictable *on average* using time series that potentially span many diverse states of nature.[1] If predictability shifts over time, however, then an unconditional approach may be misleading and yield unstable conclusions. The continued discussion of the degree of predictability in U.S. Treasury bonds is indicative of such instabilities. In-sample evidence frequently points to predictability by means of variables such as yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), linear combinations of forward rates (Cochrane and Piazzesi, 2005), and macroeconomic variables (Cooper and Priestley, 2009, Ludvigson and Ng, 2009, Cieslak and Povala, 2015, Eriksen, 2017), but out-of-sample exercises often fail to deliver consistent evidence of predictability and statistical and economic evaluations often disagree.[2] Della Corte, Sarno, and Thornton (2008), Thornton and Valente (2012) and Sarno, Schneider, and Wagner (2016), for instance, fail to find economic value of statistical bond predictability.

In this paper, we address this issue by developing a new method that is able to assess *conditional* predictive ability among two or more forecasting methods using observable state variables and identify methods anticipated to be informative of future relative forecast performance.[3] Our contributions are fourfold. First, we provide new empirical evidence on predictable state-dependencies in bond return predictability. We document that bond return predictability shifts over time for a set of predictors well-known to the literature.

[1] Early studies include Fama and Bliss (1987), Keim and Stambaugh (1986), Fama and French (1989), and Campbell and Shiller (1991). More recent studies of bond return predictability includes Cochrane and Piazzesi (2005), Cooper and Priestley (2009), Ludvigson and Ng (2009), Cieslak and Povala (2015), Eriksen (2017), Ghysels, Horan, and Moench (2018), Berardi, Markovich, Plazzi, and Tamoni (2019), Bianchi, Büchner, and Tamoni (2019), and Gargano, Pettenuzzo, and Timmermann (2019).

[2] Bauer and Hamilton (2018) even challenge in-sample predictability by pointing out that standard regressions are subject to serious small-sample distortions when using overlapping returns. A related point is made by Wei and Wright (2013).

[3] Being able to anticipate future relative forecast performance is also relevant viewed in the light of the numerous studies that provide empirical evidence of model instabilities in predictive models. Prominent examples include Pesaran, Pettenuzzo, and Timmermann (2006), Giacomini and Rossi (2009, 2010), Pettenuzzo and Timmermann (2011), Rossi (2013), and Pettenuzzo and Timmermann (2017).

In particular, we consider yield spreads, forward spreads, yield curve factors, forward rates, and macroeconomic factors. We begin with a standard evaluation of out-of-sample forecasts generated using a rolling window scheme and find that none of the predictors are able to reliably outperform the expectations hypothesis (EH) when considering traditional *unconditional* predictive ability tests. However, this does not exclude the possibility that a given method works well in certain states of the world. To facilitate a conditional, state-dependent view of bond return predictability, we therefore develop a new statistical test for equal (un)conditional predictive ability among two or more forecasting methods. The test is a multivariate generalization of the test presented in Giacomini and White (2006) that enables us to identify forecasting methods anticipated to be informative of future (relative) predictability.[4] As such, our test is well-suited to study state-dependencies and shifts in predictability as it is directly designed to compare two or more competing forecast methods and reveal differences in relative conditional predictive ability that would otherwise be hidden in standard unconditional tests of equal predictive ability. We then employ our test to examine differences in *conditional* predictive abilities and find overwhelming evidence favoring state-dependencies in bond return predictability.

Second, we document that these shifts are related to economic activity and uncertainty measured using the Purchasing Manager's Index (PMI) (see, e.g. Berge and Jordà (2011) and Christiansen, Eriksen, and Møller (2014)) and the index ($\mathfrak{U}$) proposed in Jurado, Ludvigson, and Ng (2015), respectively. We uncover a striking pattern in bond return predictability across states related to these variables. More specifically, interpreting the expectations hypothesis (EH) as a no-predictability benchmark, we find that bond risk premia are predictable in high (low) economic activity (uncertainty) states. Conversely, the EH implication of constant risk premia provides a reasonable approximation in low (high) economic activity (uncertainty) states. Consistent with this, we find that out-of-sample R^2s (Campbell and Thompson, 2008) for individual predictors are mostly negative in low (high) economic activity (uncertainty) states and positive in high (low) activity (uncertainty)

[4] The test further extends the (unconditional) multivariate Diebold-Mariano statistic (Diebold and Mariano, 1995) proposed in Mariano and Preve (2012) by allowing for comparison of a mixture of nested and non-nested models.

states. In short, albeit several predictors fail to provide valuable information *on average*, many outperform the EH *conditional* on the state of the world.

Third, we show that the predictable state-dependencies in bond return predictability are exploitable for real-time forecasting purposes. In particular, we document sizable gains in predictive accuracy when evaluated using both standard statistical criteria and when measuring the economic value from the viewpoint of a mean-variance investor that trades in the Treasury bond market. To facilitate this analysis, we device a simple and intuitive dynamic ranking rule for identifying the set of forecasting methods with indistinguishable conditional predictive ability in real-time. The dynamic ranking rule is inspired by the Model Confidence Set (Hansen, Lunde, and Nason, 2011) (MCS) for ranking a set of forecasting methods. A rejection of the null hypothesis of equal conditional predictive ability implies that one or more methods display superior predictive ability in some or all states. The rule enables us to predict relative forecasting performance using simple least squares and, subsequently, to rank the methods according to their predictive performance. If a single method is selected, then this method constitutes the forecast. If several methods with equal conditional predictabive ability are identified, then we perform equal-weighted forecast combination (Bates and Granger, 1969) among the selected methods. We refer to this strategy as a *dynamic forecast combination* strategy. It is well established in this literature that a simple equal-weighted combination forecast is hard to beat (Timmermann, 2006). Yet, as pointed out by Aiolfi, Capistrán, and Timmermann (2011), little attention has been paid to determining the optimal set of models to combine given a potential pool of candidate predictors. We argue that our dynamic forecast combination schemes provides a simple and intuitive way to dynamically trim the candidate set of predictors prior to combination.[5] Our dynamic forecast combination strategy can thus be viewed as a dynamic trimming strategy (as opposed to the static version considered in, among others, Rapach et al. (2010)), where we only combine across forecasts from models anticipated to display superior predictive ability in the current state.

[5] A large empirical literature documents gains from (statically) trimming forecasts prior to averaging. Notable examples include Aiolfi and Favero (2005), Aiolfi and Timmermann (2006), Timmermann (2006), Stock and Watson (2004), Rapach, Strauss, and Zhou (2010), Bjørnland, Gerdrup, Jore, Smith, and Thorsrud (2012), and Genre, Kenny, Meyler, and Timmermann (2013).

Fourth, we document that our dynamic forecast combination scheme generates out-of-sample risk premia estimates that are strongly countercyclical and spikes in recessions. This is important as nearly all individual predictors (except the Ludvigson and Ng (2009) macro factor) generates procyclical risk premia estimates. The latter (former) is (in)consistent with standard finance theory, where risk premia are expected to be high in bad times due to heightened risk aversion (Campbell and Cochrane, 1999, Wachter, 2006, Joslin, Priebsch, and Singleton, 2015, Cochrane, 2017). The equal-weighted combination schemes, on the other hand, generates acyclical forecasts that display no relation to the real economy. The fact that our dynamic forecast combination scheme delivers strongly countercyclical out-of-sample risk premia forecasts that improve overall predictive accuracy and economic value strongly supports our conclusion that our test is able to correctly identify and exploit shifts in bond returns predictability.

In sum, we provide new empirical evidence of predictable state-dependencies in bond return predictability that are linked to economic activity and uncertainty. We document that these predictability shifts are exploitable in real-time and delivers sizable gains in both predictive accuracy and economic value. The gains originate from our method's ability to correctly predict relative forecasting performance and that this leads to better and economically meaningful out-of-sample bond risk premia estimates.

Related literature Our paper is related to two broad strands of literatures. First, there is an extensive literature that studies the predictability of Treasury bond excess returns. Most of the literature focuses on unconditional predictive ability relative to the EH. Yet, a recent literature has started to document state-dependencies and differences in inference between statistical and economic evaluations. Della Corte et al. (2008), Thornton and Valente (2012) and Sarno et al. (2016) find that high statistical predictability does not translate into economic gains for mean-variance investors in out-of-sample tests. Gargano et al. (2019) reconcile the seemingly contradictory evidence on the statistical and economic value of bond prediction models by incorporating stochastic volatility and time-varying parameters into the predictive regression. They further find that bond return predictability is significantly stronger in recessions than in expansions. Related in-sample evidence for time-varying predictive performance is found in Andreasen, Engsted, Møller, and Sander (2018) and Andreasen, Jørgensen, and Meldrum (2019). Specifically, Andreasen et al. (2018) find that bond risk premia are positively (negatively) related to yield spreads in expansions (recessions) and Andreasen et al. (2019) argue that there is a significantly stronger relation between yield spreads and bond risk premia during the zero lower bound period. We contribute to this literature by uncovering novel evidence on predictable time-variations in forecasting performance for a broad set of well known bond predictors. We directly test for conditional predictive ability and documents that predictability itself varies over time and that it is predictable and exploitable. We further contribute to the understanding of bond market dynamics by demonstrating that relative performance is closely related to economic activity and macroeconomic uncertainty and that bond risk premia are predictable in times of high (low) economic activity (uncertainty) states, whereas the EH provides a reasonable anchor in low (high) economic activity (uncertainty) states. We also find that our out-of-sample forecasts are consistent with bond risk premia being high in bad times and spiking in recessions (Campbell and Cochrane, 1999, Wachter, 2006).

Our paper further contributes to a large and active literature on forecasting and forecast evaluations. First, we provide the first multivariate test for equal conditional predictability ability. Our multivariate generalization of the Giacomini and White (2006) test provides forecasters with the opportunity to test equal (un)conditional predictive ability among many forecast methods without having to rely on multiple testing adjustments, which would otherwise be appropriate if testing many models against each other on a pairwise basis (Hubrich and West, 2010). Second, we facilitate easy testing of equal un(conditional) predictive ability as all our proposed tests are simple Wald statistics with chi-squared limited distribution as opposed to the non-standard and context-specific distribution often found in the literature (Clark and McCracken, 2001, McCracken, 2007, Clark and McCracken, 2012, Gonçalves, McCracken, and Perron, 2017).[6] Third, and in contrast to

[6] Moreover, our tests are generally invariant to any reordering of the forecasting methods under comparison, ensuring that conclusions drawn from a single test is unaltered by any permutation of the ordering of the forecasting methods. This is important as it alleviates the need for incorporating multiple testing adjustments.

Hubrich and West (2010), Mariano and Preve (2012), and Clark and McCracken (2012), the proposed tests are applicable to a mixture of both nested and non-nested models, hold for a general loss function, and allow for non-stationarity in the data. Last, we allow for comparison of a wider class of forecasting methods not considered in the application of this paper, including linear, non-linear, Bayesian, and non-parametric methods, something that is not allowed in the methods proposed in Clark and McCracken (2012), Granziera, Hubrich, and Moon (2014), and Gonçalves et al. (2017) that apply to linear models only. We further contribute to a literature that studies the impact of trimming forecasts prior to combination. Makridakis and Winkler (1983) show that the marginal impact of including an additional method decreases as the number of methods increases. Similarly, Jose and Winkler (2008) document that trimming or winsorizing improve forecast accuracy and reduce the risk of large errors. Samuels and Sekkel (2017) find that using the (unconditional) MCS as a trimming device prior to constructing combined forecasts can greatly improve accuracy and Diebold and Shin (2018) propose a LASSO-based procedure that sets some combining weights to zero and shrinks the survivors toward equality. Our approach differs from theirs by being rooted in a formal multivariate test of equal conditional predictive ability and by focusing in predicted performance rather than past performance. For comparison, we implement a version of the unconditional trimming rule (Samuels and Sekkel, 2017) and find that our conditional trimming provides superior predictive ability. Finally, our work is related to recent papers studying the predictability of relative forecast performance (Timmermann and Zhu, 2017, Granziera and Sekhposyan, 2019).

The remainder of the paper proceeds as follows. Section 2 outlines our data and state variables. Section 3 develops our multivariate statistical tests for equal (un)conditional predictive ability and introduces our dynamic ranking rule. Section 4 present our main empirical results on state-dependencies in bond return predictability and Section 5 examines the sources of conditional predictability. Section 6 examines the link between our out-of-sample risk premia estimates and the real economy. Section 7 examines the economic value attainable for a mean-variance investor. Finally, Section 8 provides concluding remarks.

2. Bond return predictability

This section discusses our setting and describes the construction of monthly bond excess returns and provides summary statistics. We then outline the set of bond return predictors used in our empirical analysis and their construction and, last, discuss the state variables used to assess state-dependencies in bond excess return predictability.

2.1. Predictive regression for bond returns

To motivate our study, consider a classic predictive regression model for bond risk premia of the form

$$rx^{(k)}_{t+\tau} = \ {}^{(k)} + \ {}^{(k)}\boldsymbol{x}_t + \varepsilon^{(k)}_{t+\tau}, \tag{1}$$

where $rx^{(k)}_{t+\tau} = p^{(k-\tau)}_{t+\tau} - p^{(k)}_t - p^{(\tau)}_t$ denotes the τ-month log excess holding period return on a k-month zero-coupon Treasury bond and $p^{(k)}_t$ is the time t log price of a bond with k months to maturity. We are interested in determining whether a set of predictors $\boldsymbol{x}_t$ can predict bond excess returns, where a natural benchmark is the expectations hypothesis that implies $\ {}^{(k)} = \mathbf{0}$ (i.e. no predictability). Our empirical analysis focuses on monthly U.S. Treasury bond excess returns ($\tau = 1$) over the period 1962 to 2018 constructed using the Gürkaynak, Sack, and Wright (2007) dataset and a one-month Treasury bill obtained from the Center for Research in Security Prices (CRSP) as in Gargano et al. (2019).[7] The use of a monthly holding period returns avoids the many issues with persistence induced from using annual overlapping returns for conducting inference (Bauer and Hamilton, 2018) and may better facilitate the capture of short-lived dynamics in bond excess returns across economic states (Farmer, Schmidt, and Timmermann, 2019, Gargano et al., 2019).

[Insert Figure 1 About Here]

Figure 1 plots time series of excess returns for bonds with two, three, four, and five years to maturity, respectively. The same set of maturities are considered in, e.g., Fama and Bliss (1987) and Gargano et al. (2019). Bond excess returns are notably more volatile

[7] We detail the construction of monthly log yields and bond prices in the Internet Appendix. The data are available at https://www.federalreserve.gov/data/nominal-yield-curve.htm.

during the early 1980s and more calm in the late 2010s. The magnitude of bond risk premia also appears to narrow towards the end of our sample period.

[Insert Table 1 About Here]

Panel A of Table 1 presents descriptive statistics for our monthly bond excess return series. We see that longer maturity bonds are more volatile and earn higher excess returns on average. The Sharpe ratios are generally high and range between 0.46 for the two-year bond and 0.35 for the five-year bond. Morover, short-term bonds display higher skewness, kurtosis, and have slightly more persistent excess returns. However, the persistence in these monthly bond excess return series are substantially lower compared to those typically observed in studies using annual overlapping bond excess returns (e.g. Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009)) and the first-order autocorrelation coefficient never exceeds 0.17 across the maturity spectrum. Panel B of Table 1 provides contemporaneous bond excess return correlation across maturities and confirms the well known observation that bond excess returns are highly cross-sectionally correlated across maturities. Correlation coefficients range from 0.99 to 0.93, where bonds closest to each other in the maturity spectrum obtain the highest contemporaneous correlations.

2.2. Predictor variables

We consider a set of standard bond predictors from the extant literature. In particular, we consider yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), forward rates (Cochrane and Piazzesi, 2005), and macroeconomic factors (Ludvigson and Ng, 2009).

In particular, the Campbell-Shiller (CS) yield spreads are computed as

$$ys_t^{(k)} = y_t^{(k)} - y_t^{(1)}, \tag{2}$$

where $y_t^{(k)}$ denotes the time t log yield on a bond with k periods to maturity and $y_t^{(1)}$ denotes the safe one-period return measured using the implied yield on a one-month Treasury bill obtained from CRSP. The Fama-Bliss (FB) forward spreads are computed similarly as

$$fs_t^{(k)} = f_t^{(k)} - y_t^{(1)}, \tag{3}$$

where $f_t^{(k)}$ denotes the forward rate for loans between $t + k - 1$ and $t + k$. The principal component (PC) of yields are computed from the set of 12-, 24-, 36-, 48-, and 60-month maturity yields and we focus on the first three components often referred to as level, slope, and curvature. These components account for almost all of the variation in yields. The Cochrane-Piazzesi (CP) single factor is formed from a linear combination of forward rates using the projection

$$\overline{rx}_{t+1} = \delta + \ {}_1 f_t^{(12)} + \ {}_2 f_t^{(24)} + \ {}_3 f_t^{(36)} + \ {}_4 f_t^{(48)} + \ {}_5 f_t^{(60)} + \varepsilon_{t+1}, \tag{4}$$

where $\overline{rx}_{t+1} = \frac{1}{4} \sum_{i=2}^{5} rx_{t+1}^{(i \times 12)}$ can be viewed as the excess return on a portfolio of Treasury bonds with different maturities. The CP factor is then obtained as $\mathrm{CP}_t = \widehat{\delta} + \hat{}\,\boldsymbol{f}_t$, with $\hat{} = (\hat{}_1, \hat{}_2, \hat{}_3, \hat{}_4, \hat{}_5)$ and $\boldsymbol{f}_t = (f_t^{(12)}, f_t^{(24)}, f_t^{(36)}, f_t^{(48)}, f_t^{(60)})'$. Last, the Ludvigson-Ng (LN) factor is based on a $T \times M$ panel of macroeconomic variables, x, that we assume can be adequately described by a static factor model, i.e.

$$x_{i,t} = \kappa_i g_t + \nu_{i,t,}, \tag{5}$$

where g_t is an $s \times 1$ vector of common factors with $s \ll M$ that we estimate using principal component analysis. We use the dataset from McCracken and Ng (2016). Following Ludvigson and Ng (2009), we build a single factor as a linear combination of a subset of the principal components. We determine the subset using the BIC and obtain the factor from a projection of $\overline{rx}_{t+1}$ onto the set of selected macroeconomic factors.

[Insert Table 2 About Here]

Table 2 presents descriptive statistics for the set of predictors (Panel A) along with contemporaneous correlations (Panel B). All variables are constructed using the full range of available observations here, but are constructed recursively in the out-of-sample exercise. Yield spreads and forward spreads are fairly persistent with first-order autocorrelations between 0.82 and 0.92 and are heavily cross-correlated. Unsurprisingly, PC2 — the slope component of the yield curve — is strongly related to both yield and forward spreads. CP and LN are similarly positively correlated with the spread variables and also positively correlated with each other. Last, we note that CP and LN are relatively less persistent compared to the remaining variables.

2.3. State variables

Conventional tests of equal predictive ability gauge if forecasts are equally accurate *on average*, not *if and when* predictors exhibit predictive ability. We are interested in this latter question and below we develop a new test to address this question in a multivariate setting. The basic premise of the test rests on the intuition that even if a given predictor does not display unconditional predictive ability, it may display superior predictive ability conditional on some states of the world. To identify these states, we need to identity state variables that are likely to capture fluctuations in forecast losses. We consider two state variables well-known for their ability to capture salient features of the state and properties of the business cycle. We use the Purchasing Managers' Index (PMI) published by the Institute of Supply Management and the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015).

2.3.1. Purchasing managers' index The PMI is an index constructed from a survey of the manufacturing sector that ranges from 0 to 100 and is released on the first business day of every month. The index is specifically designed to capture the state of the economy with values below 50 indicating a recession in the manufacturing economy and is regarded as a prime leading indicator of the business cycle (Berge and Jordà, 2011, Christiansen et al., 2014). Using a variable that tracks business cycle fluctuations to assess state-dependencies in bond predictability is motivated by a recent literature that documents stark differences in predictive performance for asset returns across different phases of the business cycle (Henkel, Martin, and Nardari, 2011, Dangl and Halling, 2012, Andreasen et al., 2018, Eriksen, 2017, Gargano et al., 2019, Farmer et al., 2019).

2.3.2. Macroeconomic uncertainty $\mathcal{U}$ measures a common component in the time-varying volatilities of h-step ahead forecast errors across a large number of macroeconomic series that include categories such as real activity, prices, and financial assets.[8] The index is therefore associated with the variance of the unpredictable components of macroeconomic variables.[9] Macroeconomic uncertainty has recently been identified as an important contributor to business cycle fluctuations (Bloom, 2009, Ludvigson, Ma, and Ng, 2019) and asset prices (Drechsler, 2013, Bali, Brown, and Tang, 2017, Borup and Schütte, 2019). Moreover, it has recently been been used to study state-dependent performance of affine term structure models (Sarno et al., 2016). Last, uncertainty is likely to be linked to risk aversion (Bekaert, Engstrom, and Xu, 2019), which bears direct influence on the required compensation for bearing interest rate risk.

[Insert Figure 2 About Here]

Figure 2 displays the evolution of the two state variables over time. Green (yellow) shaded ares represent periods of (high) low activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. PMI and $\mathcal{U}$ are both persistent series with first-order autoregressive coefficients of 0.94 and 0.99, respectively. PMI ($\mathcal{U}$) mostly takes on (low) high values in bad times and the two series realize a full sample correlation of -0.48, suggesting that the series are related, but not perfect substitutes. For our purpose, we remain agnostic about the lead-lag relation between uncertainty and the macroeconomy, but note that Ludvigson et al. (2019) provide evidence that higher macroeconomic uncertainty in recessions arises as an endogenous response to output shocks (see also Andreasen (2019)).

[8] We focus on the index associated with $h = 1$ step ahead forecast errors to match the holding period of the bond as well as the data frequency in general.

[9] An alternative is the macroeconomic uncertainty index proposed in Rossi and Sekhposyan (2015), although its quarterly frequency puts it at a disadvantage compared to the monthly frequency of the Jurado et al. (2015) index.

3. Multivariate tests for equal predictive ability

This section introduces our econometric methodology. We develop a multivariate test for equal conditional predictive ability, present our main forecasting methods and hypotheses, and discuss applications within dynamic forecast selection and combination.

3.1. Notation

To introduce a general notation, let $\boldsymbol{w}_t \equiv (y_t, \boldsymbol{x}_t)'$ be an observed vector defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, where y_t is the target object of interest and $\boldsymbol{x}_t$ is a vector of predictors. We consider a setting where $p+1$, $p \geq 1$, methods are available for forecasting τ periods into the future. We denote the forecast of $y_{t+\tau}$ originating a time t by $\hat{f}^i_{t+\tau} = f^i \left(\boldsymbol{w}_t, \boldsymbol{w}_{t-1}, \ldots, \boldsymbol{w}_{t-m^i+1}; \hat{\boldsymbol{\theta}}^i_{t,m^i} \right)$ $\Big($for $i = 1, \ldots, p+1$, where f^i is a measurable forecast function. $\hat{\boldsymbol{\theta}}^i_{t,m^i}$ denotes the parameter estimates used to construct the forecast for the ith forecasting methods obtained using observations from the m^i most recent periods in the past. For ease of exposition and along the lines of Giacomini and White (2006), we define $m = \max\{m^1, \ldots, m^{p+1}\}$ and require that $m < \infty$. This excludes expanding window forecast schemes from our test, but allows for rolling window estimators. The number of out-of-sample forecasts is $T = N - (m + \tau - 1)$ with a total sample size of N (time series) observations. In order to assess the forecasting ability of each forecasting method, we use a real-valued loss function $L_{t+\tau}\left(Y_{t+\tau}, \hat{f}^i_{t+\tau}\right)$.$\Big($ Important examples of L include economic measures such as utility or profits (Granger and Machina, 2006) and statistical measures such as the square or absolute value of the forecast errors (West, 2006), where forecast errors are given by $e^i_{t+\tau} = \hat{f}^i_{t+\tau} - y_{t+\tau}$. To ease the notational burden, we suppress the arguments of L and write the ith loss function as $L^i_{t+\tau}$ for the remainder the of the paper.

3.2. Rolling window forecasts

Our out-of-sample analysis is based on conventional predictive regression models of the form presented in (1), which is arguably the most common methodology on forecasting bond risk premia (see, e.g., Gargano et al. (2019)). We note, however, that our econometric framework is not limited to such regressions, but naturally extends to a broad array of parametric, non-parametric, and Bayesian methods. We consider a set of p methods, indexed by i, defined by the set of predictors outlined in Section 2.2 in addition to the natural EH benchmark. The predictive regression models will be estimated by a rolling window OLS scheme, in accordance with our and Giacomini and White (2006) assumptions, and forecasts generated at time t according to (suppressing maturity-dependence for notational simplicity)

$$\hat{f}^i_{t+\tau} = \hat{}^{\,i}_t + \hat{}^{\,i}_t \boldsymbol{x}^i_t, \tag{6}$$

for $i = 1, \ldots, p+1$ with $\hat{\boldsymbol{\theta}}^i_{t,m^i} = \left(\left({}^i_t, \hat{}^{\,i}_t\right)\right)'$. The benchmark EH forecast naturally includes no predictors and is simply defined as $\hat{f}^i_{t+\tau} = \hat{}^{\,i}_t$, which is consistent with a no-predictability interpretation as implied by financial theory.

3.3. The hypothesis of equal conditional predictive ability

We are interested in formally evaluating whether a set of $p+1$ forecasting methods display equal conditional predictive ability using some σ-field (information set), $\mathcal{G}_t$. That is, we want to test the hypothesis that

$$\mathbb{H}_0\colon \mathbb{E}\left[\left(L^i_{t+\tau}|\mathcal{G}_t\right] = \mathbb{E}\left[\left(L^{i+1}_{t+\tau}|\mathcal{G}_t\right]\left(\quad i = 1, \ldots, p, \right. \tag{7}$$

or equivalently that

$$\mathbb{H}_0\colon \mathbb{E}\left[\Delta \boldsymbol{L}_{t+\tau}|\mathcal{G}_t\right] = \boldsymbol{0}, \tag{8}$$

where $\Delta \boldsymbol{L}_{t+\tau} = \left(\Delta L^1_{t+\tau}, \ldots, \Delta L^p_{t+\tau}\right)'$ and $\Delta L^j_{t+\tau} = L^j_{t+\tau} - L^{j+1}_{t+\tau}$ for $j = 1, \ldots, p$ and where $L^i_{t+\tau}$ is the loss function for the ith method. This null hypothesis offers three main advantages. First, it allows us to study conditional predictive abilities and identify *if and when* there are differences in the competing models' conditional predictive accuracy.

This is distinctly different from testing whether methods have equal predictive accuracy *on average*. Indeed, a given forecasting method can display superior predictive ability in certain states of the world as captured by $\mathcal{G}_t$, yet still perform poorly on average. In other words, the null hypothesis implies that $\mathcal{G}_t$ is uninformative about the relative predictive accuracy of one or more forecasting methods when forecasting the object of interest τ periods into the future. A rejection of the null hypothesis, conversely, implies that relative predictive accuracy is predictable by $\mathcal{G}_t$ and that this may be exploited to improve forecasts. Second, if $\mathcal{G}_t$ is set to the trivial σ-field, $\mathcal{G}_t = \{\emptyset, \quad\}$, then the null hypothesis becomes unconditional and, as such, comparable to the one considered in Mariano and Preve (2012). In this case, the hypothesis test provides information about the average predictive ability of the forecasting methods as in Diebold and Mariano (1995) and West (1996). Third, the loss functions depend explicitly on the parameter estimates and not on their probability limits, leading to a test statistic that takes into account estimation uncertainty. Importantly, by allowing for asymptotically non-vanishing estimation uncertainty, the test can accommodate the empirically relevant case of inclusion of nested models in the set of forecasting methods which is a feature that the (unconditional) multivariate test in Mariano and Preve (2012) cannot handle.[10] This is particularly important in our context as the EH model is nested within every competing forecasting model coming from (1).

3.4. The multivariate test statistic

The null hypothesis in (8) is equivalent to stating that

$$\mathbb{H}_0\colon \mathbb{E}\left[\tilde{h}_t \Delta \boldsymbol{L}_{t+\tau}\right] \Big(= \mathbf{0} \tag{9}$$

for all $\mathcal{G}_t$-measurable functions $\tilde{h}_t$. We restrict attention to a subset of these functions that we gather in the q-dimensional vector $\boldsymbol{h}_t = \left(\tilde{h}_t^{(1)}, \ldots, \tilde{h}_t^{(q)}\right)'$. We refer to this vector as the state function. For this choice of state function, we can reformulate the multivariate null

[10] Technically, with $\mathcal{G}_t = \{\emptyset, \quad\}$ and asymptotically vanishing estimation uncertainty the standard errors of differences in forecast performance between a set of nested models will equal zero, leading to non-standard limiting distributions of the test statistics.

hypothesis of equal conditional predictive ability as follows

$$\mathbb{H}_{0,h}\colon \mathbb{E}\left[\boldsymbol{h}_t \quad \Delta\boldsymbol{L}_{t+\tau}\right] = \boldsymbol{0}, \tag{10}$$

where the subscript h indicates the dependence on the state function and denotes the Kronecker product. The specification in (10) is a natural multivariate extension of the null hypothesis in Giacomini and White (2006). Indeed, we obtain their econometric framework as a special case when $p = 1$.

Our empirical analysis focuses on one-step ahead forecasting, $\tau = 1$, as is common in the bond return predictability literature and we consider an information set $\mathcal{G}_t$, $\mathcal{F}_t \subseteq \mathcal{G}_t$, containing the state variables discussed in Section 2.3. We view this setting our leading example, but provide theoretical results for multi-step ahead forecasting, i.e. $\tau > 1$, in the Internet Appendix along with our assumptions that are identical to those of Giacomini and White (2006). Finally, let $\boldsymbol{d}_{t+1} = \boldsymbol{h}_t \quad \Delta\boldsymbol{L}_{t+1}$. We then consider the following quadratic statistic

$$S_h = T\bar{\boldsymbol{d}}'\hat{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}}, \tag{11}$$

where $\bar{\boldsymbol{d}} \equiv T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{t+1}$, and $\hat{\boldsymbol{\Sigma}}_T \equiv T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{t+1}\boldsymbol{d}_{t+1}'$ is a $(qp \times qp)$ sample covariance matrix that consistently estimates the variance of $\boldsymbol{d}_{t+1}$.[11] That is, S_h is a natural Wald statistic constructed for testing whether $\bar{\boldsymbol{d}}$ is a zero vector. When formulating an alternative hypothesis, one must take into account the generality that data is allowed to exhibit non-stationarity. We provide a discussion in the Internet Appendix. For some $c > 0$, we formulate the alternative in line with Giacomini and White (2006) as

$$\mathbb{H}_{A,h}\colon \mathbb{E}\left[\bar{\boldsymbol{d}}'\right] (\mathbb{E}\left[\bar{\boldsymbol{d}}\right] \geq c, \tag{12}$$

for all T sufficiently large. Under stationarity, the null and alternative hypothesis are

[11] We note that for large values of q and/or p, the dimension of $\boldsymbol{\Sigma}_T$ and $\bar{\boldsymbol{d}}$ may become large, potentially leading to issues with statistical inferences in finite samples. We propose remedies in Borup and Thyrsgaard (2017), but note that our empirical analysis use single instruments together with $p = 6$, leading to reasonable dimensions.

exhaustive. Under non-stationarity, this may not necessarily be the case. If an important $\mathcal{G}_t$-measurable variable is omitted from the state function, it may happen that $\mathbb{E}\left[\bar{\boldsymbol{d}}'\right] \Big(\mathbb{E}\left[\bar{\boldsymbol{d}}\right] = 0$ for a particular sample size due to, for instance, shifting means without the null hypothesis being true. As an example, one could easily imagine a situation where one method outperforms (some of) the other methods in certain states, while it performs worse than the same methods in other states. Therefore, the test has little power against alternatives where the loss differentials are correlated with $\mathcal{G}_t$-measurable random variables not included in the state function. While this concern is important, it also highlights the flexibility of the test statistic. As mentioned above, the econometrician chooses the state function to include state variables relevant for disentangling the forecasting abilities of two or more forecasting methods. The test therefore only provides power in situations when this is possible. As a result, the test statistic changes with the choice of state function and the subscript in S_h in (11) emphasizes this.

The asymptotic properties of the test statistic are summarized in Theorem 1 and the proof can be found in the Internet Appendix.

Theorem 1 (**One-step multivariate conditional predictive ability test**)**.** *Suppose Giacomini and White (2006) type assumptions hold (Assumptions 1-3 in the Internet Appendix). Then the test statistic has the following properties.*

A. ***Asymptotic distribution under the null.*** *For forecast horizon $\tau = 1$, state function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$, and under $\mathbb{H}_0$ in (8),*

$$S_h \xrightarrow{d} \chi^2\left(qp\right), \quad \textit{as } T \to \infty. \tag{13}$$

B. ***Consistency under the alternative.*** *For any $c \in \mathbb{R}_+$ and under $\mathbb{H}_{A,h}$ in (12),*

$$\mathbb{P}\left[S_h > c\right] \to 1, \quad \textit{as } T \to \infty. \tag{14}$$

C. ***Permutation invariance.*** *Let $\boldsymbol{L}^*_{t+1}$ be an arbitrary permutation of the forecast* *losses, and define* $\Delta \boldsymbol{L}^*_{t+1} = \boldsymbol{D}\boldsymbol{L}^*_{t+1}$*, where*

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \dots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 & -1 \end{bmatrix} \tag{15}$$

is a $p \times (p+1)$ *matrix. Let* $\bar{\boldsymbol{d}}^* = T^{-1}\sum_{t=1}^{T} \boldsymbol{d}^*_{t+1}$ *with* $\boldsymbol{d}^*_{t+1} = \boldsymbol{h}_t \quad \Delta \boldsymbol{L}^*_{t+1}$ *and* $\hat{\boldsymbol{\Sigma}}^*_T \equiv \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{d}^*_{t+1}\boldsymbol{d}^{*\prime}_{t+1}$*. Then,*

$$S^*_h \equiv T\bar{\boldsymbol{d}}^{*\prime}_m \left(\hat{\boldsymbol{\Sigma}}^*_T\right)^{-1} \bar{\boldsymbol{d}}^*_m = S_h, \quad \forall T. \tag{16}$$

We provide a corresponding result for the unconditional, possibly multi-step, case, in the Internet Appendix. This case, where we compare the average performance of the methods over the out-of-sample window, is obtained by setting $\boldsymbol{h}_t = 1$ for all t. The limiting distribution is $\chi^2(p)$ for a test statistic that employs a HAC type covariance matrix estimator. In the case of the conditional test and multi-step forecast horizons an identical $\chi^2(qp)$ limiting distribution is obtained when using an appropriate HAC type covariance matrix estimator to capture arising serial dependence.[12]

Although any reordering of the forecasting methods alters the dynamics of $\boldsymbol{d}_{t+1}$, Theorem 1.C. shows that we obtain the same value of the test statistics and the same limiting distribution under the null hypothesis for each permutation (reordering) of the forecasting methods, irregardless of the null being true or not. This is important as it allows the researcher to perform just a single test.

3.5. Understanding the test

To provide an intuitive understanding of our test statistics, we consider the simplest case of $p = 1$, where the problem reduces to a comparison between a single forecasting method

[12] Borup and Thyrsgaard (2017) provide Monte Carlo evidence for all test statistics. They show that all tests display good size and power properties in dimensions similar to the ones considered in this paper.

and a benchmark. An unconditional test is equivalent to the regression

$$\Delta L_{t+1} = \varphi_0 + \eta_{t+1}, \tag{17}$$

where the null hypothesis that $\varphi_0 = 0$ can be testing using a standard t-test using an appropriate HAC type of covariance estimator. The conditional test augments the regression with a set of state variables. Our empirical study considers a single state variable (plus a constant) at a time to facilitate economic interpretation. Suppose, accordingly, that we have a single state variable $\tilde{h}_t$, then the conditional test amounts to running the extended regression

$$\Delta L_{t+1} = \boldsymbol{\varphi} \boldsymbol{h}_t + \eta_{t+1} = \varphi_0 + \varphi_1 \tilde{h}_t + \eta_{t+1}, \tag{18}$$

with $\boldsymbol{\varphi} = (\varphi_0, \varphi_1)$ and $\boldsymbol{h}_t = \left(1, \tilde{h}_t\right)'$ being the state function.[13] In this case, we are interested in testing jointly $\varphi_0 = 0$ and $\varphi_1 = 0$ using a Wald test and an appropriate estimator of the covariance matrix. The limiting distribution under the null hypothesis is equivalent to the ones provided in Theorem 1. From (18), it is clear that a rejection of $\varphi_1 = 0$ indicates that there is information in the state variable that informs about future relative predictability of the models under consideration. That is, there is evidence of state-dependency. Importantly, the expression in (18) is nothing more than a full sample predictive regression similar in spirit to (1) estimated over the out-of-sample window. The key difference being that (18) predicts the future relative predictive ability among the candidate forecasting methods using state variables whose values are observable at the construction of the forecast and are picked by the researcher. We refer to them as state regressions in the following. These ideas naturally extend to our case of $p > 1$, resembling a seemingly unrelated regression (SUR) type of interpretation of our test statistic. We will make use of this insight below when formulating a simple decision rule to exploit rejections of the null hypothesis to dynamically select or combine among forecasting methods with indistinguishable predicted performance.

[13] If one uses several state variables in addition to the constant, this amounts to a multiple regression and joint inference on all parameters.

3.6. Ranking of forecasting methods

Rejection of the null hypothesis suggests that one or more of the forecasting methods exhibit superior predictive ability in certain states. However, it provides no guidance towards which method(s) that causes the rejection and display(s) the strongest predictability. The identification of the method(s) is of both economic and practical interest. Central banks, international organizations (IMF, OECD, and the World Bank), and professional forecasters (SPF and Blue Chip) frequently generate forecasts that are widely followed by market participants and policy makers. Designing routines that can identify forecasts and/or forecasters that are predicted to do well in a given state of the world therefore seems worthwhile. To that end, we propose a simple and intuitive algorithm that ranks forecasting methods based on their predicted performance with respect to one or more state variables and identifies the set of best methods. This set may consists of a single model, all models, or any number of models in between. It depends on the ability of the state function to accurately inform us about any, possible time-varying, differences in predictive accuracy. This procedure reveals potential fluctuations in predictive ability over time, similar in spirit to the fluctuation test of Giacomini and Rossi (2010), but also suggests why these fluctuation occurs due to the use of state variables. The procedure can also be applied dynamically, at the forecast origin date, to select forecast methods that is expected (conditional on $\mathcal{G}_t$) to yield the lowest loss at a future time point and to conduct conditional combination techniques. In formulating the algorithm, we consider a MCS-type procedure (Hansen et al., 2011) to eliminate methods according to an elimination rule and rank forecasting methods into a best set whose elements have equal predicted conditional predictive ability.

3.6.1. Full out-of-sample ranking rule We first device a statistical algorithm for ranking all forecasting methods based on their predictive accuracy conditional on the state variable(s) over the fill sample. In line with our empirical analysis below, we will formulate the rule using a single state variable in addition to the constant, but note that it can be extended directly to a setting with several state variables. Since $\tilde{h}_t$ may be continuous, we assume that it can be classified into a finite set of $\mathcal{A}$ discrete, non-empty, states s_a, $a = 1, \ldots, \mathcal{A}$. For example, the state variable can be a measure of economic growth, which may be classified into recessionary or expansionary states, or a measure of macroeconomic uncertainty, which may be classified into low, medium, and high uncertainty states.

Let M^0 be the set of the $p+1$ forecasting methods under consideration and M_a^* a set of best forecasting methods in terms of some loss function within the ath state. We then consider the following three-step procedure.

Step 0: Set $M_a = M^0$ for $a = 1, \ldots, \mathcal{A}$. Estimate by OLS the regression model

$$\Delta L_{t+1}^j = \boldsymbol{\varphi}^j \boldsymbol{h}_t + \eta_{t+1} \tag{19}$$

for all pairwise combinations of forecasting methods, $j = 1, \ldots, p \times (p+1)/2$. The conditional expectation of the loss differentials within each state, $\mathbb{E}\left[\Delta L_{t+\tau}^j | s = s_a\right] \Big(= \varphi_0^j + \varphi_1^j \mathbb{E}\left[\tilde{h}_t | s = s_a\right], \Big(a = 1, \ldots, \mathcal{A}$, is approximated by $\hat{\varphi}_0^j + \hat{\varphi}_1^j \hat{\mu}_{\tilde{h}}^a$, where $\hat{\mu}_{\tilde{h}}^a$ is the sample average of the state variable $\tilde{h}_t$ in state s_a. Based on those estimated conditional means, rank all $p+1$ methods (using a normalization of one method) in all states. The forecasting method with lowest predicted loss across all pairwise combinations is ranked first and similarly the method with highest predicted value is ranked at last.

Step 1: Run the multivariate test for equal conditional predictive ability.

Step 2: If the test is not rejected, set $M_a^* = M_a$. Otherwise, eliminate the lowest ranked forecasting method from M_a based on the ranking that associates with state a. Iterate Steps 1–2 until the null is no longer rejected for all $\mathcal{A}$ states.

Concluding the algorithm leads to a set M_a^* for each state s_a that contains the best forecasting methods statistically indistinguishable in terms of predictive ability. A few remarks are worthwhile here. First, the ranking rule exploits the state regression interpretation of our test statistic and is, as such, strongly rooted in econometric theory. Second, since the elimination of models is based on state-specific ranking, it will capture the state-dependency of predictability over the full out-of-sample period which proves insightful in the empirical analysis below. Third, since the algorithm provides sets of equal predictive ability within each state it can be thought of as a version of a conditional MCS algorithm. Last, since the test is permutation invariant as per Theorem 1, we only need to run it once for each time Step 2 is conducted, even though the elimination of models alters the ordering of models. However, the ranking of all models in Step 0 is not permutation invariant and requires, as such, an examination of all combinations. Fortunately, this step is only conducted once and has little computational demands being based on least squares.

3.6.2. Dynamic ranking rule The above full out-of-sample ranking rule is not applicable for real-time forecasting as Step 0 depends on a regression over all out-of-sample periods. We therefore formulate a dynamic rule that enables researches to select and/or combine among methods conditional on the realization of the state variable at the time of the forecast. To that end, we divide the out-of-sample window into two parts. The first part is used for initially estimating the state regression and the second part for forecast selection and/or combination. Suppose that the first part has length T_1 and that the second part has length T_2 with $T_1 + T_2 = T$. We then propose the following three-step ranking algorithm at each time point $t = m + T_1, \ldots, N - 1$.

Step 0: Set $M_t = M_0$. Estimate by OLS the regression model

$$\Delta L_{t+1}^{j} = \boldsymbol{\varphi}^{j} \boldsymbol{h}_t + \eta_{t+1} \tag{20}$$

over a rolling window of length T_1 for all pairwise combinations of forecasting methods, $j = 1, \ldots, p \times (p+1)/2$. The conditional expectation $\mathbb{E}\left[\Delta L_{t+1}^{j} | \mathcal{G}_t\right]\Big($is estimated by $\hat{\boldsymbol{\varphi}}^{j} \boldsymbol{h}_t = \hat{\varphi}_0^{j} + \hat{\varphi}_1^{j} \tilde{h}_t$ which measures the time t prediction of the future j'th loss differential using current information in the state variable. Based on those predictions, rank all $p + 1$ methods (using a normalization of one method). The forecasting method with lowest predicted loss across all pairwise combinations is ranked first and similarly the method with highest predicted value is ranked at last.

Step 1: Run the multivariate test for equal conditional predictive ability.

Step 2: If the test is not rejected, set $M_t^* = M_t$. Otherwise, eliminate the lowest ranked forecasting method from M based on the ranking of predicted forecast losses. Iterate Steps 1–2 until the null is no longer rejected.

This algorithm is a real-time version of the full out-of-sample version above that allocates forecasting models at each time point $t = m + T_1, \ldots, N - 1$ into a set of the best models, M_t^*, with lowest expected forecast losses, using the current information in the state variable.[14] Since this ranking is conducted at the same time forecasts are generated, it provides valuable information about the usefulness of a given set of models to base current predictions upon.

3.7. Forecast combination

We then formulate a simple and natural procedure for exploiting this ranking of predicted performances at each time t. Let $\hat{f}_{t+1}^*$ denote a combination forecast given by

$$\hat{f}_{t+1}^* = \frac{1}{\#M_t^*} \sum_{i \in M_t^*} \Bigg(\hat{f}_{t+1}^i, \tag{21}$$

where $\#M_t^*$ denotes the cardinality of (number of elements in) M_t^*. If M_t^* consists of a single forecasting method, then we rely on that single method for forecasting. If M_t^* consists of more than one method, we perform forecast combination within the set of best models. To keep focus on the ability of our method to identify the best set of models, we consider the simplest possible combination scheme: equal-weighting.[15] The equal-weighted combination scheme has a long tradition in the forecasting literature and is empirically hard to beat as it involves no estimation error in weights (Timmermann, 2006, Rapach et al., 2010). Other combination schemes are naturally possible, e.g. using estimated least squares weights, possibly with shrinkage to equal weights (Bates and Granger, 1969,

[14] Note also that it does not require a categorization of the state variable into discrete states.

[15] While one could possibly increase forecast performance further by considering more complicated combination schemes, this is not the aim of our paper. Instead, we focus on the ability of our method to discriminate between forecasting methods that are predicted to perform well and those predicted to perform poorly and show that this does indeed lead to significant improvements.

Granger and Ramanathan, 1984, Zellner, 1986, Diebold and Pauly, 1987). Our proposed combination scheme is essentially an equal-weighting principle, but with the modification that we dynamically trim the set of models prior to combination, where the trimming is based on the predicted losses from our dynamic ranking rule. In the case of only two models, $p = 1$, this reduces to the switching rule provided in Giacomini and White (2006). Timmermann and Zhu (2017) formally show that forecast improvements are guaranteed when state variables are powerful and Granziera and Sekhposyan (2019) provide empirical evidence.

3.8. A check of size and power properties

To check the finite sample properties of our tests, we perform a Monte Carlo study. We focus on their size and power properties in settings corresponding to its application in both a full out-of-sample analysis and when used in the dynamic ranking rule.

We examine a situation where the forecasts have equal predictive ability unconditionally, but conditional on some state variable $\tilde{h}_t$ at least one of the forecasts are more (or less) accurate than the others. The data-generating process is set to

$$\Delta \boldsymbol{L}_{t+1} = \boldsymbol{\mu}(\tilde{h}_t - \varrho) + \boldsymbol{\varepsilon}_{t+1}, \tag{22}$$

where $\mathbb{P}[\tilde{h}_t = 1] = \varrho$ and $\mathbb{P}[\tilde{h}_t = 0] = 1 - \varrho$. To allow for the presence of estimation error (approximately) asymptotically, as delineated by our theoretical setting, we re-sample with replacement from de-meaned loss differentials from our empirical analysis when generating $\boldsymbol{\varepsilon}_{t+1}$. In this way they maintain every influence of the estimation coming the forecasting models as well as ensure simulated time series that exhibit realistic empirical behavior. Note also that $\mathbb{E}[\Delta \boldsymbol{L}_{t+1}] = \boldsymbol{0}$, together with $\mathbb{E}[\Delta \boldsymbol{L}_{t+1} | \tilde{h}_t = 1] = \boldsymbol{\mu}(1 - \varrho)$ and $\mathbb{E}[\Delta \boldsymbol{L}_{t+1} | \tilde{h}_t = 0] = -\boldsymbol{\mu}\varrho$. That is, the unconditional null hypothesis is true, whilst the conditional one is not necessarily so, depending on the value of (the elements in) $\boldsymbol{\mu}$ and ϱ.

We consider three sample sizes; a short, medium, and long length. The medium size equals the length of our full out-of-sample window, $T = 348$, the short size equals the sample length used in the dynamic ranking rule in the application, $T_1 = 120$, and the long size is set to $1,000$ observations. Consistent with our empirical analysis, we set $p = 5$ as the number of models under comparison less one due to the computation of loss differentials. Since our ranking rules eliminate a model sequentially until it no longer rejects, we consider the full range $p = 1, \ldots, 5$. When $p < 5$, we randomly sample without replacement (in any random order) among our full set of models and subsequently reconstruct loss differentials based on the selected models. Note that any reshuffling of the order of models has no influence on the test statistic due to its permutation invariance, presented in Theorem 1, such that it has no influence on the performance of the test statistic within a fixed p. To obtain (samples of) $\boldsymbol{\varepsilon}_{t+1}$, we consider two separate cases, using the empirical loss differentials resulting from forecasting each of the 2-year and 5-year bonds, respectively. We set $\varrho = 0.4$, since this links to our findings below that documents notable superior predictability of at least one model in each of the high and low economic activity or uncertainty states, and less differences in predictive accuracy within the normal state. We use 10,000 Monte Carlo replications.

3.8.1. Size properties To examine the size properties of our test, we set $\boldsymbol{\mu} = \mathbf{0}$ such that both the unconditional and conditional null hypothesis are true. We consider two implementations of the test. The first is unconditional and uses $\boldsymbol{h}_t = 1$ for all t, whereas a conditional implementation uses $\boldsymbol{h}_t = (1, \tilde{h}_t)'$. The results are reported in Table 3 for a significance level of 5%. Conclusions are identical using a 1% and 10% significance level, and relevant tables are available upon request.

It is clear that both the conditional and unconditional tests are well-sized, showing negligible deviations from the nominal significance level. Those minor deviations generally decrease in sample size and increase in number of models under comparison. It is comforting to note that the tests maintain good size properties for the short sample size used in the dynamic ranking rule. There is no notable difference when sampling from loss differentials associated with the 2-year or 5-year bonds, except from in the short sample case where the 5-year bond loss differentials lead to a slight undersizing.

3.8.2. Power properties To examine the power properties of our test, we let the first element of $\boldsymbol{\mu}$ deviate from zero, and set the remaining elements equal to zero in a similar style to Mariano and Preve (2012). Denote this first element by μ_1. The deviation is anchored in the empirical loss differentials, making it realistic in the context of the present paper. Specifically, we compute the average absolute loss differentials across all models within the low and high activity states defined in the empirical section below, denoting it by $\hat{\eta}$. We then set $\mu_1 = c\hat{\eta}$ where $c \in [0, 2.5]$.[16] Given the specification in (22) and $\varrho = 0.4$, this allows μ_1 to deviate at most 1.5 times the empirical value of average absolute loss differentials. We have also implemented a version that lets all elements of $\boldsymbol{\mu}$ deviate from zero with a fraction c of each respective element's average absolute loss differentials within the low and high activity states. The power is uniformly stronger in this case, and results are available upon request. Note also that, in both versions, the unconditional null hypothesis remains true. We therefore set $\boldsymbol{h}_t = (1, \tilde{h}_t)'$ and examine the power of the conditional version of our equal predictability test. The power curves for a 5% significance level are depicted in Figure 3. Conclusions are identical using a 1% and 10% significance level, and the results are available upon request.

In line with the theoretical power result in Theorem 1, the test is consistent under the (local) alternative considered, as power increases to unity for stronger deviations from the null. It correctly exhibits empirical rejections equal to the nominal size at $c = 0$. Power is stronger for less model comparisons, as expected, but the difference is not substantial. As was the case for the size properties, it is comforting that the test exhibits good power properties even for the relatively short sample length. To put this into context, for $c = 1/\varrho = 1.67$ we recover the empirical value of the mean absolute values of loss differentials obtained in the empirical analysis when using (22). In this case, the power exceeds 0.94 for the smallest sample size and $p = 5$, showing very desirable power properties. There is no notable difference when sampling from loss differentials associated with the 2-year or 5-year bonds.

[16] We also ran the simulation using $\mathcal{U}$ as state variable, yielding similar conclusions, yet somewhat stronger power.

4. State-dependencies in bond return predictability

This section discovers novel evidence on predictable state-dependencies in bond excess return predictability. We first compute standard out-of-sample forecasts using a rolling window and then conduct full sample tests for equal (un)conditional predictive ability among a standard set of bond predictors using state variables capturing economic activity and uncertainty. Last, we document substantial gains in forecast accuracy from using a simple dynamic decision rule that exploits predictable differences in relative forecast performance.

4.1. Out-of-sample predictability

We begin our empirical analysis by gauging the unconditional predictive ability of our predictors individually using a rolling window estimation scheme in which predictors and parameters are estimated recursively using information available at time t only. We use the period January 1962 to December 1989 as our initial estimation period, the period from January 1990 to December 1999 as initial our testing period, and the period from January 2000 to December 2018 as our evaluation period. We focus on U.S. Treasury bonds with $k = \{24, 36, 48, 60\}$ months to maturity and consider models based on the predictor variables outlined in Section 2.2.[17] To evaluate the out-of-sample performance of the predictive methods relative to the constant expected return benchmark implied by the EH, we compute the out-of-sample R^2 statistic proposed in Fama and French (1989) and Campbell and Thompson (2008)

$$\mathrm{R}^2_{OS,i,k} = 1 - \frac{\sum_{t=R+1}^{N} \left(rx_t^{(k)} - \widehat{rx}_{t,i}^{(k)} \right)^2}{\sum_{t=R+1}^{N} \left(rx_t^{(k)} - \widehat{rx}_{t,EH}^{(k)} \right)^2}, \tag{23}$$

where $\widehat{rx}_{t+1,i}^{(k)}$ and $\widehat{rx}_{t+1,EH}^{(k)}$ denote the forecast from the ith predictor model and the EH benchmark, respectively, $R = m + T_1$ denotes the end of the testing period, and N denotes the total number of observations. The R^2_{OS} statistic in (23) is thus equivalent to

[17] Our choice of k is motivated by previous research that similar focuses on these maturities, e.g. Fama and Bliss (1987), Cochrane and Piazzesi (2005), Ludvigson and Ng (2009), and Gargano et al. (2019).

one minus the ratio of mean squared prediction errors, i.e. $\mathrm{R}^2_{OS,i,k} = 1 - \frac{MSPE_i^{(k)}}{MSPE_{EH}^{(k)}}$. An $\mathrm{R}^2_{OS} > 0$ implies that the MSPE of the ith predictor model is lower than that of the EH benchmark model, indicating higher predictive accuracy. We interpret the EH model as a no-predictability benchmark and test the null of no predictability $\left(\mathrm{R}^2_{OS} \leq 0\right)$ $\Bigg($against the one-sided alternative of predictability by the ith predictor model $\left(\mathrm{R}^2_{OS} > 0\right)$ $\Bigg($using the Diebold and Mariano (1995) (DM) test for equal predictive ability.[18]

[Insert Table 4 About Here]

Table 4 reports R^2_{OS} values and DM p-values for our predictor models across the maturity spectrum. The key observation from this table is that no individual model is able to convincingly outperform the EH benchmark unconditionally for all maturities. Most models deliver negative R^2_{OS} values and those that are positive are far from being significant at any of the conventional levels.[19] These results are in line with Gargano et al. (2019), who similarly find few positive R^2_{OS} values for linear predictive models. Like us, they find forward spreads to consistently be among the best predictor of monthly bond excess returns for short maturities and LN the best for longer maturities. However, we find poorer performance for CP using rolling window regressions, indicating that bond return predictability is sensitive to the forecasting setup.[20] Last, we consider a simple equal-weighted forecast combination scheme (Bates and Granger, 1969, Timmermann, 2006, Rapach et al., 2010). We denote this combined forecast by EW. The combined forecast generates positive R^2_{OS} values from 6.08% for the two-year bond to 4.58% for the five-year bond. These values are all significant according to the DM p-value at the five percent level. That is, although no individual predictor is able to consistently outperform the EH, a simple equal-weighted average of the individual forecasts is.

[Insert Figure 4 About Here]

[18] Note that this is the unconditional version of the test statistic in Giacomini and White (2006) which is nested within our framework for $p = 1$.

[19] We provide in-sample predictive regression results in the Internet Appendix, where we show that our set of predictors are reliably related to bond risk premia when using the full range of available information.

[20] In unreported results, we indeed find that most of your R^2_{OS} values improve when considering a forecasting environment with an expanding window instead. However, the qualitative results and conclusions are very similar.

Figure 4 plots the cumulative difference in squared prediction errors (CDSPE) between the EH and the ith predictor model

$$\mathrm{CDSPE}_{t,i}^{(k)} = \sum_{l=R+1}^{t} \left(rx_l^{(k)} - \widehat{rx}_{l,EH}^{(k)} \right)^2 - \sum_{l=R+1}^{t} \left(rx_l^{(k)} - \widehat{rx}_{l,i}^{(k)} \right)^2, \tag{24}$$

where $R+1$ denotes the time of the first forecast and $\widehat{rx}_{t+1,i}^{(k)}$ and $\widehat{rx}_{t+1,EH}^{(k)}$ denote the forecast from the ith predictor model and the EH benchmark, respectively. This graphical device is suggested by Goyal and Welch (2008) as a way to assess relative performance over time (and is thus indirectly a visual inspection of state-dependencies). Figure 4 plots the CDSPEs against economic activity and uncertainty states identified using PMI and $\mathcal{U}$, respectively, to assess the relation between relative forecasting performance and our state variables. The plots supports the use of conditioning variables that tracks salient features of the business cycle and that these are related to relative predictive abilities. For instance, CS and FB derive a sizable portion of their overall positive performance from high (low) economic activity (uncertainty) period. This is consistent with Andreasen et al. (2018). Moreover, CS and FB appears to provide valuable information over the 2008 to 2018 periods, which is consistent with the stronger relationship between the slope of the yield curve and future excess bond returns documented in Andreasen et al. (2019). PC and CP are consistently poor, and especially so in low (high) economic activity (uncertainty) periods, whereas LN initially performs well, but particularly poorly at the end of the latest financial crisis. Consistent with the positive R_{OS}^2 values in Table 4, the equal-weighted forecast combination (EW) performs well over the entire evaluation period.

4.2. Testing for equal conditional predictive ability

The previous section establishes that linear predictive methods are unable to reliably beat the EH on average. However, this does not exclude the possibility that some methods provide significantly better forecast in certain states of the world. To investigate this hypothesis more formally, we consider our multivariate test for equal conditional predictive ability introduced in Section 3. A rejection of the null of equal conditional predictive ability implies that some methods are better than others and that relative forecasting performance is predictable by the state variable(s). If conditional forecast performance is predictable, then it may be possible to exploit this information to generate more informative forecasts. A natural way to do so, which we explore in more detail below, is to combine across forecasts methods with indistinguishable conditional predictive ability. Throughout the empirical analysis, we consider three specifications for the state regression. First, we consider the information in PMI to examine if predictive ability is related to economic activity and specify the state function as $\boldsymbol{h}_t = (1, \mathrm{PMI}_t)'$. Second, we specify $\boldsymbol{h}_t = (1, \mathcal{U}_t)'$ to study the effect of macroeconomic uncertainty. Last, we also consider an unconditional version of the multivariate test in which we set $\boldsymbol{h}_t = 1$ for all t. We denote this by NONE.

[Insert Table 5 About Here]

Table 5 reports test statistics and corresponding p-values for our multivariate test for equal (un)conditional predictive ability over the evaluation period using the three specifications for the state regression discussed above using: PMI, $\mathcal{U}$, and NONE. The implementation is based on a sample covariance matrix as dictated by theory (see Section 3 and the Internet Appendix).[21] We find strong rejections of the null hypothesis of equal conditional predictive ability for both specifications of $\boldsymbol{h}_t$ that uses conditioning information representing salient features of the business cycle across all maturities, indicating that there is substantial evidence favoring state-dependencies in bond excess return predictability. The unconditional test, on the other hand, fails to reject equal predictive abilities across all models and maturities. In other words, our choice of state variables enables the detection of conditional differences.

4.3. Full out-of-sample period ranking and elimination

Having established that bond return predictability is state-dependent and related to state variables tracking economic activity and uncertainty, we now turn to a more detailed analysis of this link. We first study the ranking and elimination of models over the full

[21] We note that NONE should, in theory, by evaluated using a HAC estimator, but we use a sample estimator here to ease comparison. However, results are both qualitative and quantitatively similar when employing a Newey and West (1987) estimator with a bandwidth of 12 lags.

out-of-sample period and, subsequently in Section 4.4, how to use the information in a real-time forecasting exercise.

[Insert Figure 5 About Here]

Our state variables, PMI and $\mathfrak{U}$, are continuous variables. To facilitate interpretation and later empirical analyses, we therefore classify our sample into low, normal, and high economic activity (uncertainty) periods using the 20% and 80% quantiles of the time series for PMI ($\mathfrak{U}$), similarly to Rapach et al. (2010). Figure 5 illustrates the full out-of-sample elimination order of the predictive models when conditioning on the low, normal, and high PMI and $\mathfrak{U}$ states, respectively, using a 10% significance level. Specifically, whenever we reject the null of equal predictive ability, we use the ranking rule discussed in Section 3, which determines the order of elimination and the best set of models within each state. The patterns that emerge are striking. First, the EH is always excluded in the high economic activity state across the entire maturity spectrum. If we interpret EH as a no-predictability benchmark, this implies that bond risk premia are predictable when economic activity is high. Conversely, the EH is always included in the best set of models in the low economic activity state, suggesting that bond risk premia are unpredictable when the economy is doing poorly. This is consistent with the in-sample result in Andreasen et al. (2018) which focus on yield curve slope risk only. LN, PC, FB, and CS are instead (mostly) included in (excluded from) the best set of methods in periods with high (low) economic activity. Using $\mathfrak{U}$ as our state variable produces similar results. The EH is always included in (excluded from) the best set of methods in high (low) uncertainty states. Last, the EH is usually included in the best of methods in normal times, where LN, CP, and PC are usually excluded.

Overall, we argue that our empirical results are consistent with, and clearly points to, state-dependencies in bond excess return predictability linked to economic activity and uncertainty. Bond excess returns are predictable in states with high (low) economic activity (uncertainty), whereas the EH serves as a reliable anchor in the remaining states of the world.

4.4. Dynamic forecast combination

Bond excess return predictability displays state-dependencies over the full out-of-sample period. As a natural next step, we investigate if they can be exploited to improve out-of-sample forecasts in real-time. As detailed in Section 3.6.2, we consider a dynamic rule that enables the identification at each point time of the best set of methods with indistinguishable conditional predictive ability. If the set consists of a single method, then we rely on the forecasts for that method. If the set consists of two or more models, we perform forecast combination within the set using equal weights. Forecast combination has since the seminal work of Bates and Granger (1969) been viewed as an elegant way to improve forecast accuracy and combinations of individual forecasts often deliver more accurate forecasts than using the single best model (Timmermann, 2006). However, as pointed out in Aiolfi et al. (2011), little focus has been put on determining the optimal set of models to combine given a potential pool of predictors. We view our procedure as a way to do exactly that. It identifies the best set of forecasting methods whose conditional predictive ability is indistinguishable.[22] We denote this set by M_t^* and note that its composition may vary over time and is identical to the standard equal-weighted combination forecasts when all models exhibit equal conditional predictability, whereas it collapses to dynamic method selection if the set is a singleton. For cases in between, we simply average across the selected forecasting methods in M_t^*.

Panel B of Table 4 presents the results for our dynamic forecast combination scheme using PMI and $\mathcal{U}$, respectively, as conditioning variables and using NONE as the unconditional alternative. This unconditional alternative is related to Samuels and Sekkel (2017) who suggest trimming a given set of models using a recursive implementation of the MCS. Our conditional alternative achieves trimming using a conditional MCS idea with the elimination based on the predictability of bond excess return predictability. One can view this as a dynamic extension of the trimming strategy considered in, among others, Rapach et al. (2010). Strikingly, this strategy delivers positive R^2_{OS} values relative to the EH across

[22] Recent alternative suggestions include determining the optimal set based on past performance (Aiolfi and Timmermann, 2006), the model confidence set (Samuels and Sekkel, 2017), and lasso-based procedures (Diebold and Shin, 2018).

all conditioning variables and bond maturities. R^2_{OS} values are economically large with values between 5.11% and 7.98% for PMI and between 4.98% and 9.86% for $\mathcal{U}$. Moreover, these values generally exceed even those of the EW strategy with some margin. All (most) are significant relative to the EH (EW) at conventional levels when using either PMI or $\mathcal{U}$, whereas NONE does not deliver significant improvements against the EW.

[Insert Figure 6 About Here]

Figure 6 plots the CDSPE for our two dynamic forecast combination strategies and the unconditional alternative NONE relative to the EH. Overall, we find that relative forecasting gains are mostly uniformly distributed across the out-of-sample evaluation period and that no particular event or period drive the positive results, although we do observe a particularly forceful increase during the latest recession relative to the EH benchmark for the five-year bond using PMI as the state variable.

[Insert Figure 7 About Here]

Figure 7 plots the CDSPE for our two dynamic forecast combination strategies and NONE relative to EW. As above, we find that that our dynamic forecast combination strategy always performs on par or better than EW. This is also reflected in Panel C of Table 4 in which we observe positive R^2_{OS} values that are of economically meaningful magnitudes and most are significant at conventional significance levels. These relative forecasting gains are concentrated in periods with low (high) economic activity (uncertainty). That is, our dynamic forecast combination scheme delivers improvements in forecast accuracy in periods of turmoil, exactly when investors and forecasters arguably needs it the most. Moreover, we see that trimming the set of candidate methods prior to combination using a dynamic rule rooted in our multivariate test for equal conditional predictive ability delivers sizable improvements.

In sum, our results establish that bond return predictability display predictable and exploitable state-dependencies in an out-of-sample forecasting exercise. Our results are further supportive of the notion that bond return predictability itself is linked to variables capturing economic activity and uncertainty.

5. Understanding the sources of conditional predictability

This section studies the underlying sources of conditional predictability and the sizable improvements in predictive accuracy established above. We address this in several steps. First, we compute inclusion frequencies for each forecasting method and conditioning variable using the low, normal, and economic activity and uncertainty regimes, respectively, identified earlier. We then study how the individual methods perform in each state and relate it to the overall performance. Third, we inspect the methods selected by the decision rule over time.

5.1. Inclusion frequencies

We compute inclusion frequencies for each forecasting method and state variable using the low, normal, and high states for economic activity (PMI) and uncertainty ($\mathcal{U}$), respectively, defined in Section 4.2. Within each state s_a, we then define the inclusion frequency of the ith forecasting method as the fraction of months the model is included in the best set relative to the total number of months in state a.

[Insert Table 6 About Here]

Table 6 reports the inclusion frequencies for bond return predictor models when conditioning on PMI and $\mathcal{U}$, respectively. These inclusion frequency largely mirror the image from the full sample elimination order in Figure 5. The EH is almost always included in the low activity state, whereas the inclusion frequencies are low for the high activity state. That is, bond excess returns are predictable in high economic activity states, but less so in other states. The EH, conversely, provides a reliable anchor in periods with low and normal economic activity. A similar conclusion is reached when conditioning on macroeconomic uncertainty. EH is almost always included in the high uncertainty state, but rarely in the low uncertainty state. PC, CP, and LN, on the other hand, is mostly included in high (low) economic activity (uncertainty) states.

5.2. State-dependent predictability

The inclusion frequencies are indicative of when certain models are predicted to do well. In this section, we ask whether the inclusion frequencies align with relative performance. That is, we ask whether the procedure correctly identifies methods with good and bad relative performance.

[Insert Table 7 About Here]

Table 7 reports state-specific R^2_{OS} values for the individual predictors relative to the EH. The results are supportive of the procedure correctly identifying methods that do well. We find that individual predictors are generally performing poorly ($\mathrm{R}^2_{OS} < 0$) in low (high) economic activity (uncertainty) states and well ($\mathrm{R}^2_{OS} > 0$) in high (low) economic activity (uncertainty) states. This is consistent with the inclusion frequencies of the EH. Specifically, the procedure appears to correctly anticipate periods in which the EH provides a reasonable anchor for expected bond excess returns and period in which bond risk premia are predictable. Moreover, there is also a close mapping between the inclusion frequencies and the magnitudes of the R^2_{OS} values, where models are more likely to be included (excluded) in a given state the higher (lower) its R^2_{OS}. That is, the gains in predictive accuracy are coming from the rule's ability to correctly predict predictability.

5.3. Decision rule and model selection

Figure 8 illustrates the models selected for the best set of models using the decision rule over time using PMI and $\mathfrak{U}$ as conditioning variables, respectively. Green (yellow) shaded aras indicate high (low) states identified using the 20% and 80% quantiles of the series. A "+" indicates inclusion.

[Insert Figure 8 About Here]

[Insert Figure 9 About Here]

Figures 9 illustrates the size of the set of best models selected over time using the decision rule using PMI and UNC as conditioning variables, respectively. We note that the best set of models varies considerably over time and includes situations in which the set include all models, leading to forecasts equal to EW, and situations with a singleton. That is, at times there is no need for trimming of the full set of models and at other times we should only use the forecasts from a single model. Importantly, this tells us that dynamically trimming leads to improvements over a simple, static trimming rule.

6. Links to the real economy

In this section, we examine the link between our out-of-sample bond risk premia forecasts and the real economy. Standard finance theory implies that investors demand a compensation for risks associated with recessions (or macroeconomic activity in general) due to heightened risk aversion, see, among many, Fama and French (1989), Campbell and Cochrane (1999), Wachter (2006), Cochrane (2017), and Bekaert et al. (2019). That is, bond risk premia ought to be countercyclical and spike in recessions (Ludvigson and Ng, 2009, Joslin et al., 2015, Andreasen et al., 2018).

[Insert Table 8 About Here]

We employ PMI as our measure of economic activity (Berge and Jordà, 2011) and report in Table 8 the contemporaneous correlation among PMI and the risk premia estimates from the set of individual models, EW, and the dynamic forecast combinations generated by PMI, $\mathcal{U}$, and NONE. The results offer two main insights. First, yield-based variables such as CP, FB, PC, and CP all deliver risk premia estimates that are significantly positively correlated with real economic activity. That is, these models imply procyclical risk premia, which sharply contrasts canonical theory. LN, on the other hand, obtains a significant negative correlation of about -38% across the maturity spectrum, which is consistent with countercyclical risk premia. Interestingly, the EW combination strategy produces risk premia estimates with almost identically zero correlation with the real economy. That is, even though the EW combination produces significantly more accurate forecasts, cf. Table 4, they are acyclical and unrelated to the state of the economy. The acyclicality is likely to be caused by the, apparently too crude, equal-weighting across counter and procyclical forecasts. Our dynamic combination strategy that selects individual methods for subsequent combination, based on information in the state variables, produces markedly negative and statistically significant correlations with the real economy. As such, our conditional view and associated trimming rule provides *both* economically meaningful risk premia estimates, through marked countercyclicality, and much stronger predictability.[23]

[Insert Figure 10 About Here]

Supporting this, Figure 10 depicts our dynamic combination forecast using PMI and $\mathcal{U}$ as state variables, along with NBER-dated recessions. We see a clear tendency for the risk premia estimates to increase during recessionary periods and decline during expansionary periods, resembling a countercyclicality in business cycles. These findings altogether demonstrate the importance of appropriately selecting among plausible models, as done in the present paper.

7. Economic value

This section measures the economic value of the strong predictive improvements established above for our dynamic forecast combination strategy. Specifically, we consider the asset allocation decision of an investor with mean-variance preferences and relative risk aversion

that chooses the weight $\omega_t^{(k)}$ to invest in a k-period bond and the weight $\left(1-\omega_t^{(k)}\right)$ to invest in a one-period safe bond (Marquering and Verbeek, 2004).[24] The resulting portfolio return is then

$$r_{p,t+1}^{(k)} = y_t^{(1)} + \omega_t^{(k)} rx_{t+1}^{(k)}, \tag{25}$$

[23] Other types of business cycle indicators can naturally be entertained. We report in the Internet Appendix contemporaneous correlations among generated forecasts and each of the macroeconomic uncertainty ($\mathcal{U}$), recession probabilities of Chauvet and Piger (2008), the Chicago Fed National Activity Index (CFNAI), and logarithmic growth rates to industrial production growth. It stands out that our dynamic forecasting combination technique leads to much stronger countercyclical bond risk premia than all yield-based variables and EW.

[24] Assuming that investors have mean-variance preferences in asset allocation exercises has a long tradition in predictability studies and similar approaches can be found in, among many, Campbell and Thompson (2008), Goyal and Welch (2008), Wachter and Warusawitharana (2009), Thornton and Valente (2012), Sarno et al. (2016), Eriksen (2017), Ghysels et al. (2018), and Gargano et al. (2019).

where $rx^{(k)}_{t+1}$ denotes monthly bond excess returns for a Treasury bond with k periods until maturity. We assume that the investor has a utility function, $U(r^{(k)}_{p,t+1})$, of the form

$$U(r^{(k)}_{p,t+1}) = \mathbb{E}_t\left[r^{(k)}_{p,t+1}\right] \Big(- \frac{1}{2} \ \mathrm{Var}_t\left[r^{(k)}_{p,t+1}\right], \tag{26}$$

where denotes the Pratt-Arrow measure of relative risk aversion. Solving the maximization problem yields the optimal portfolio weights

$$\omega^{(k)}_t = \frac{1}{\,} \frac{\mathbb{E}_t\left[rx^{(k)}_{t+1}\right]}{\mathrm{Var}_t\left[\Big(rx^{(k)}_{t+1}\right]}, \tag{27}$$

where $\mathbb{E}_t\left[rx^{(k)}_{t+1}\right]$ (is estimated using the ith predictive method and $\mathrm{Var}_t\left[rx^{(k)}_{t+1}\right]$ is computed using a rolling window of past bond excess return realizations.[25] We winzorize weights according to reasonable leverage and shorting constraints, similarly to Thornton and Valente (2012) and Gargano et al. (2019), such that $\omega^{(k)}_t \in [-1, 2]$ for all maturities. Using the sequence of portfolio weights, we can compute the average utility, or certainty equivalent return (CER), for each forecast method using (26). We similarly compute the CER for the EH benchmark prediction in lieu of the predictive models. The CER gain is then the difference between the CER for the predictive models and the CER for the EH benchmark. We annualize the CER gain so that it can be interpreted as the annual portfolio management fee that an investor would be willing to pay to have access to the information in the predictive forecast relative to the EH benchmark.[26] In this way, we measure directly the economic value of bond excess return predictability.

7.1. Certainty equivalent returns

Table 9 reports annualized CER gains for all individual bond predictors (for comparison) relative to the EH in Panel A and for our dynamic forecast combination strategy relative to the EH and the equal-weighted combination strategy in Panels B and C, respectively. In our main results, we set $= 10$ as in Eriksen (2017), but show in the Internet Appendix

[25] We always use the same variance estimated over the same period as the forecasts for all models so that the optimal portfolio weights only differ because of differences in the excess bond return forecast.

[26] Trading costs are generally small in U.S. Treasury bond markets (Adrian, Fleming, and Vogt, 2017).

that our results are almost identical for lower values of relative risk aversion, e.g. $= 5$. In order to evaluate the statistical significance of the CER gains, we follow Eriksen (2017) and Gargano et al. (2019) and conduct a conventional t-test on the mean of the time series of realized utility differences, evaluated using a Newey and West (1987) estimator for the standard errors.

[Insert Table 9 About Here]

Overall, we find little evidence of individual predictive models reliably generating economic value. The exception is LN that generally do remarkably well utility-wise, something that starkly contrast the statistical results. CS and FB do poorly for the two- and three-year maturities, but obtains positive CER gains for the four- and five-year maturities, albeit not significantly so. PC and CP are overall unable to deliver any economic value to an investor above that provided by the EH benchmark. LN, on the other hand, delivers positive and significant CER gains across the full maturity spectrum. Overall, we find little evidence that predictable deviations from the EH can be exploited to generate economic value on average when considering individual methods. EW, on the other hand, obtains positive CER gains for all maturities, indicating that combination forecasts may improve the economic value.

Panel B considers the CER gains for our dynamic forecast combination scheme for PMI, $\mathcal{U}$, and NONE. Consistent with our statistical results, we obtain positive CER gains in almost all instances and many are reliably different from zero. The PMI-based dynamic forecast scheme delivers positive CER gains between 0.39 and 1.43, which are significantly different from zero at the ten percent level for all maturities. The $\mathcal{U}$-based dynamic forecast scheme similarly delivers positive values that are significant for the longer maturity bonds. NONE is mostly delivering less economic value than PMI and $\mathcal{U}$. As such, the overall message is clearly supportive of the notion that taking state-dependencies in bond return predictability into account leads to substantial improvements in forecasting accuracy and that these improvement translates into better investment performance for a mean-variance investor that trades in the U.S. Treasury bond market.

Panel C mirrors this conclusion by documenting positive CER gains for the dynamic forecast combination strategies relative to EW. All PMI-based CER gains are statistically significant at the ten percent level and the $\mathcal{U}$-based CER gains are significant for the three- and four-year bonds. We argue that this strongly supports the idea that dynamically trimming the set of models prior to averaging can substantially improve forecast performance and the resulting economic value. That is, eliminating forecasting methods predicting to perform poorly and only maintaining methods with indistinguishable conditional predictive ability delivers both statistical as well as economic value.

[Insert Figure 11 About Here]

[Insert Figure 12 About Here]

Figures 11 and 12 plots the cumulative realized utilities for our dynamic forecast combination strategies relative to the EH and the EW, respectively. Overall, we note that utility gains are enjoyed uniformly over the out-of-sample period relative to the EH. This is remarkable as our approach is not designed to capture utility, but predictability.

7.2. State-dependent utility

Analogous to Section 5.2, we report in Table 10 the state-dependent CER gains for the individual predictors relative to the EH.

[Insert Table 10 About Here]

We find that individual predictors are generally delivering negative CER gains in low (high) economic activity (uncertainty) states and positive CER gains in high (low) economic activity (uncertainty) states. This is fully consistent with the results from the statistical evaluation and suggest that PMI and $\mathcal{U}$ predict not only statistical performance, but economic value as well. The only difference is LN, which generally delivers positive CER gains across all states and maturities. That is, although it looks poor overall from a statistical point of view, it is superior from an economic point of view.

8. Concluding remarks

We study predictable state-dependencies in bond return predictability and provide empirical evidence consistent with bond return predictability being state-dependent and closely related to economic activity and macroeconomic uncertainty. We show that bond risk premia are predictable in times of high (low) economic activity (uncertainty) states identified using the the Purchasing Managers' Index (PMI) and the uncertainty index proposed in Jurado et al. (2015), whereas the EH implication of constant risk premia (no-predictability) provides a reasonable anchor in low (high) economic activity (uncertainty) states. A dynamic forecast combination strategy that averages across forecasting methods predicted to do well delivers forecasts that are substantially more informative than a simple, static equal-weighted forecast combination scheme. This holds both across standard statistical evaluation metrics and when considering the economic value to a mean-variance investor that trades in the U.S. Treasury bond market. We provide evidence that the improved forecast performance originates from the state variables ability to correctly predict periods in which individual predictors are likely to perform well.

To facilitate our empirical analysis, and to explicitly take into account the fact that we have more than two forecasting methods to distinguish between, we develop a new multivariate statistical test for equal conditional and unconditional predictive ability. The test is a multivariate generalization of the test presented in Giacomini and White (2006) and therefore inherits the main properties of their test. Most importantly for our application, it allows for a mixture of nested and non-nested models. Our dynamic forecast combination strategy is rooted in this test and delivers a simple and intuitive way to trim the pool of candidate forecasting methods prior to averaging.

We end by emphasizing that our multivariate test of conditional predictive ability is not confined to studies of the Treasury bond market, but may find many and diverse applications across the fields of economics and finance. For instance, it would be natural to study the conditional predictive ability of, say, the Goyal and Welch (2008) set of predictors in a multivariate setting as a complement to the large literature on their unconditional performance. Indeed, recent studies suggest that state-dependencies are present in stock return predictability Henkel et al. (2011), Dangl and Halling (2012), Farmer et al. (2019). Similarly, the approach is likely to be useful in evaluating inflation predictability and identifying periods in which variables such as unemployment rates provides useful information. Finally, we also envision its use in comparing professional forecasters and, in particular, to determine if some forecasters are better than others conditional on being in a certain state. We leave these considerations for future research.

References

Adrian, T., M. Fleming, and E. Vogt (2017). An index of Treasury market liquidity: 1991-2017. Federal Reserve Bank of New York Staff Reports, no. 827.

Aiolfi, M., C. Capistrán, and A. Timmermann (2011). Forecast combination. In M. P. Clements and D. F. Hendry (Eds.), *Oxford Handbook of Economic Forecasting*, Chapter 12, pp. 355–388. Oxford University Press.

Aiolfi, M. and C. A. Favero (2005). Model uncertainty, thick modelling and the predictability of stock returns. *Journal of Forecasting 24*(4), 233–254.

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

Andreasen, M. M. (2019). Explaining bond return predictability in an estimated New Keynesian model. Working paper, Aarhus University.

Andreasen, M. M., T. Engsted, S. V. Møller, and M. Sander (2018). The yield spread and bond return predictability in expansion and recessions. Working paper, Aarhus University.

Andreasen, M. M., K. Jørgensen, and A. Meldrum (2019). Bond risk premiums at the zero lower bound. Working paper, Aarhus University.

Andrews, D. W. K. (1991). Heteroskedasticity and autocorrelation consistent covariance matrix estimation. *Econometrica 59*(3), 817–858.

Bali, T. G., S. J. Brown, and Y. Tang (2017). Is economic uncertainty priced in the cross-section of stock returns? *Journal of Financial Economics 126*, 471–489.

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

Bauer, M. D. and J. D. Hamilton (2018). Robust bond risk premia. *Review of Financial Studies 31*(2), 399–448.

Bekaert, G., E. C. Engstrom, and N. R. Xu (2019). The time variation in risk appetite and uncertainty. Working paper, Columbia University.

Berardi, A., M. Markovich, A. Plazzi, and A. Tamoni (2019). Mind the (covergence) gap: Bond predictability strikes back! Working paper.

Berge, T. J. and Ò. Jordà (2011). Evaluating the classification of economic activity into recessions and expansion. *American Economic Journal: Macroeconomics 3*, 246–277.

Bianchi, D., M. Büchner, and A. Tamoni (2019). Bond risk premia with machine learning. Working paper, Warwick Business School.

Bjørnland, H. C., K. Gerdrup, A. S. Jore, C. Smith, and L. A. Thorsrud (2012). Does forecast combination improve Norges bank inflation forecasts? *Oxford Bulletin of Economics and Statistics 74*(2), 163–179.

Bloom, N. (2009). The impact of uncertainty shocks. *Econometrica 77*(3), 623–685.

Borup, D. and E. C. M. Schütte (2019). Asset pricing with data revisions. Working paper.

Borup, D. and M. Thyrsgaard (2017). Statistical tests for equal predictive ability across multiple forecasting methods. Working paper, Aarhus University.

Campbell, J. Y. and J. H. Cochrane (1999). By force of habit: A consumption-based explanation of aggregate stock market behavior. *Journal of political Economy 107*(2), 205–251.

Campbell, J. Y. and R. J. Shiller (1991). Yield spreads and interest rate movements: A bird's eye view. *Review of Economic Studies 58*(3), 495–514.

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

Chauvet, M. and J. Piger (2008). A comparison of the real-time performance of business cycle dating methods. *Journal of Business & Economic Statistics 26*(1), 42–49.

Christiansen, C., J. N. Eriksen, and S. V. Møller (2014). Forecasting US recessions: The role of sentiment. *Journal of Banking and Finance 49*, 459–468.

Cieslak, A. and P. Povala (2015). Expected returns in treasury bonds. *Review of Financial Studies 28*(10), 2859–2901.

Clark, T. E. and M. McCracken (2013). Advances in forecast evaluation. *In Handbook of Economic Forecasting Volume 2, Elsevier B.V.*, 1107–1201.

Clark, T. E. and M. W. McCracken (2001). Tests of equal forecat accuracy and encompassing for nested models. *Journal of Econometrics 105*(1), 85–110.

Clark, T. E. and M. W. McCracken (2012). Reality checks and comparison of nested predictive models. *Journal of Business & Economic Statistics 30*(1), 53–66.

Cochrane, J. H. (2017). Macro-finance. *Review of Finance 21*(3), 945–985.

Cochrane, J. H. and M. Piazzesi (2005). Bond risk premia. *American Economic Review 95*(1), 138–160.

Cooper, I. and R. Priestley (2009). Time-varying risk premiums and the output gap. *Review of Financial Studies 22*(7), 2801–2833.

Dangl, T. and M. Halling (2012). Predictive regressions with time-varying coefficients. *Journal of Financial Economics 106*, 157–181.

Della Corte, P., L. Sarno, and D. L. Thornton (2008). The expectations hypothesis of the term structure of very short-term rates: Statistical tests and economic value. *Journal of Financial Economics 89*, 158–174.

Diebold, F. X. and R. S. Mariano (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics 13*(3), 134–144.

Diebold, F. X. and P. Pauly (1987). Structural change and the combination of forecasts. *Journal of Forecasting 6*(1), 21–40.

Diebold, F. X. and M. Shin (2018). Machine learning for regularized survey forecast combination: Partially-egalitarian lasso and its derivatives. *International Journal of Forecasting*, Forthcoming.

Drechsler, I. (2013). Uncertainty, time-varying fear, and asset prices. *Journal of Finance 68*(5), 1843–1889.

Eriksen, J. N. (2017). Expected business conditions and bond risk premia. *Journal of Financial and Quantitative Analysis 52*(4), 1667–1703.

Fama, E. and K. R. French (1997). Industry cost of equity. *Journal of Financial Economics 43*(2), 153–193.

Fama, E. F. (1984). Forward and spot exchange rates. *Journal of Monetary Economics 14*, 319–338.

Fama, E. F. and R. R. Bliss (1987). The information in long-maturity forward rates. *American Economic Review 77*(4), 680–692.

Fama, E. F. and K. R. French (1989). Business conditions and expected returns on stocks and bonds. *Journal of Financial Economics 25*, 23–49.

Farmer, L. E., L. Schmidt, and A. Timmermann (2019). Pockets of predictability. Working paper, University of Virginia.

Gargano, A., D. Pettenuzzo, and A. Timmermann (2019). Bond return predictability: Economic value and links to the macroeconomy. *Management Science 65*(2), 508–540.

Genre, V., G. Kenny, A. Meyler, and A. Timmermann (2013). Combining expert forecasts: Can anything beat the simple average? *International Journal of Forecasting 29*, 108–121.

Ghysels, E., C. Horan, and E. Moench (2018). Forecasting through the rearview mirror: Data revisions and bond return predictability. *Review of Financial Studies 31*(2), 678–714.

Giacomini, R. and B. Rossi (2009). Detecting and predicting forecast breakdowns. *Review of Economic Studies 76*(2), 669–705.

Giacomini, R. and B. Rossi (2010). Forecast comparisons in unstable environments. *Journal of Applied Econometrics 25*(4), 595–620.

Giacomini, R. and H. White (2006). Tests of conditional predictive ability. *Econometrica 74*(6), 1545–1578.

Gonçalves, S., M. W. McCracken, and B. Perron (2017). Tests of equal accuracy for nested models with estimated factors. *Journal of Econometrics 198*(2), 231–252.

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

Granger, C. W. J. and M. J. Machina (2006). Forecasting and decision theory. *In Handbook of Economic Forecasting, Elsevier B.V. 1*, 82–98.

Granger, C. W. J. and R. Ramanathan (1984). Improved method of combining forecasts. *Journal of Forecasting 3*(2), 197–204.

Granziera, E., K. Hubrich, and H. R. Moon (2014). A predictability test for a small number of nested models. *Journal of Econometrics 182*(1), 174–185.

Granziera, E. and T. Sekhposyan (2019). Predicting relative forecasting performance: An empirical investigation. *International Journal of Forecasting*, In Press.

Gürkaynak, R. S., B. Sack, and J. H. Wright (2007). The U.S. treasury yield curve: 1961 to the present. *Journal of Monetary Economics 54*(8), 2291–2304.

Hansen, P. R., A. Lunde, and J. M. Nason (2011). The model confidence set. *Econometrica 79*(2), 453–497.

Henkel, S. J., J. S. Martin, and F. Nardari (2011). Time-varying short-horizon predictability. *Journal of Financial Economics 99*, 560–580.

Hubrich, K. and K. D. West (2010). Forecast evaluation of small nested model sets. *Journal of Applied Econometrics 25*, 574–594.

Jose, V. R. R. and R. L. Winkler (2008). Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting 24*, 163–169.

Joslin, S., M. Priebsch, and K. J. Singleton (2015). Risk premiums in dynamic term structure models with unspanned macro risk. *Journal of Finance 69*(3), 1197–1233.

Jurado, K., S. C. Ludvigson, and S. Ng (2015). Measuring uncertainty. *American Economic Review 105*(3), 1177–1216.

Keim, D. B. and R. F. Stambaugh (1986). Predicting returns in the stock and bond markets. *Journal of Financial Economics 17*, 357–390.

Litterman, R. and J. Scheinkman (1991). Common factors affecting bond returns. *Journal of Fixed Income 1*(1), 54–61.

Ludvigson, S. C., S. Ma, and S. Ng (2019). Uncertainty and business cycles: Exogenous impulse or endogenous response? Working paper, New York University.

Ludvigson, S. C. and S. Ng (2009). Macro factors in bond risk premia. *Review of Financial Studies 22*(12), 5027–5067.

Makridakis, S. and R. L. Winkler (1983). Averages of forecasts: Some empirical results. *Management Science 29*(9), 987–1112.

Mariano, R. S. and D. Preve (2012). Statistical tests for multiple forecast comparison. *Journal of Econometrics 169*(1), 123–130.

Marquering, W. and M. Verbeek (2004). The economic value of predicting stock index returns and volatility. *Journal of Financial and Quantitative Analysis 39*(2), 407–429.

McCracken, M. W. (2007). Asymptotics for out of sample tests of Granger causality. *Journal of Econometrics 140*(2), 719–752.

McCracken, M. W. and S. Ng (2016). FRED-MD: A monthly database for macroeconomic research. *Journal of Business & Economic Statistics 34*(4), 574–589.

Nelson, C. R. and A. F. Siegel (1987). Parsimonious modeling of yield curves. *Journal of Business 60*(4), 473–489.

Newey, W. K. and K. D. West (1987). A simple, positive semi-definite, heteroskedasticity and autocorrelation consistent covariance matrix. *Econometrica 55*(3), 703–708.

Paye, B. S. and A. Timmermann (2006). Instability of return prediction models. *Journal of Empirical Finance 13*(3), 274–315.

Pesaran, M. H., D. Pettenuzzo, and A. Timmermann (2006). Forecasting time series subject to multiple structural breaks. *Review of Economic Studies 73*(4), 1057–1084.

Pettenuzzo, D. and A. Timmermann (2011). Predictability of stock returns and asset allocation under structural breaks. *Journal of Econometrics 164*, 60–78.

Pettenuzzo, D. and A. Timmermann (2017). Forecasting macroeconomic variables under model instability. *Journal of Business & Economic Statistics 35*(2), 183–201.

Rapach, D. E., J. K. Strauss, and G. Zhou (2010). Out-of-sample equity premium prediction: Combination forecasts and links to the real economy. *Review of Financial Studies 23*(2), 821–862.

Rossi, B. (2013). Advances in forecasting under instability. In G. Elliott and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 2, Part B, Chapter 21, pp. 1203–1324. Elsevier.

Rossi, B. and T. Sekhposyan (2015). Macroeconomic uncertainty indices based on nowcast and forecast error distributions. *American Economic Review 105*(5), 650–655.

Samuels, J. D. and R. M. Sekkel (2017). Model confidence sets and forecast combination. *International Journal of Forecasting 33*, 48–60.

Sarno, L., P. Schneider, and C. Wagner (2016). The economic value of predicting bond risk premia. *Journal of Empirical Finance 37*, 247–267.

Schrimpf, A. and Q. Wang (2010). A reappraisal of the leading indicator properties of the yield curve under structural instability. *International Journal of Forecasting 26*(4), 836–857.

Stock, J. H. and M. W. Watson (2003). Forecasting output and inflation: the role of asset prices. *Journal of Economic Literature 41*(3), 788–829.

Stock, J. H. and M. W. Watson (2004). Combination forecasts of output growth in a seven-country data set. *Journal of Forecasting 23*, 405–430.

Svensson, L. E. O. (1994). Estimating and interpreting forward interest rates: Sweden 1992-1994. NBER Working Paper No. 4871.

Thornton, D. L. and G. Valente (2012). Out-of-sample predictions of bond excess returns and forward rates: An asset allocation perspective. *The Review of Financial Studies 25*(10), 3141–3168.

Timmermann, A. (2006). Forecast combination. In G. Elliott, C. W. J. Granger, and A. Timmermann (Eds.), *Handbook of Economic Forecasting*, Volume 1, Chapter 4, pp. 135–196. Elsevier.

Timmermann, A. and Y. Zhu (2017). Monitoring forecast performance. Working paper, Rady School of Management.

Wachter, J. A. (2006). A consumption-based model of the term structure of interest rates. *Journal of Financial Economics 79*(2), 365–399.

Wachter, J. A. and M. Warusawitharana (2009). Predictable returns and asset allocation: Should a skeptical investor time the market? *Journal of Econometrics 148*, 162–178.

Wei, M. and J. H. Wright (2013). Reverse regressions and long-horizon forecasting. *Journal of Applied Econometrics 28*, 353–371.

West, K. D. (1996). Asymptotic inference about predictive ability. *Econometrica 64*(5), 1067–1084.

West, K. D. (2006). Forecast evaluation. *In Handbook of Economic Forecasting, Elsevier B.V. 1*, 100–134.

White, H. (1994). *Estimation, inference and specification analysis.* New York: Cambridge University Press.

White, H. (2001). *Asymptotic theory for econometricians.* San Diego: Academic Press.

Wooldridge, J. M. and H. White (1988). Some invaraince principles and central limit theorems for dependent heterogeneous processes. *Econometric Theory 4*(2), 210–230.

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with g -prior distributions. *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno De Finetti 6*, 233–43.

Table 1: Descriptive statistics

This table presents descriptive statistics for monthly excess bond returns. Panel A reports mean, standard deviation, skewness, kurtosis, Sharpe ratios, and first-order autocorrelation (AC(1)) of bond excess returns for two- to five-year bond maturities. Bond returns are in excess of the implied yield on a one-month Treasury bill. Gross returns do not subtract the one-month implied Treasury bill yield. Monthly bond excess returns are constructed using end-of-month Treasury yield data from Gürkaynak et al. (2007). Panel B reports contemporaneous correlations between the excess bond return series. The sample period is January 1962 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Descriptive statistics				
Mean	1.29	1.60	1.85	2.06
Mean (Gross)	5.73	6.04	6.29	6.50
Std. dev.	2.80	3.92	4.95	5.93
Skewness	0.57	0.25	0.08	0.03
Kurtosis	16.68	11.76	8.58	7.05
Sharpe ratio.	0.46	0.41	0.37	0.35
AR(1)	0.17	0.15	0.13	0.12
Panel B: Correlations				
2-year bond	1.00			
3-year bond	0.99	1.00		
4-year bond	0.96	0.99	1.00	
5-year bond	0.93	0.97	0.99	1.00

Table 2: Predictor variables

This table reports descriptive statistics (Panel A) for the predictor variables used in our empirical study and their contemporaneous correlations (Panel B). We report means, standard deviations, skewness, kurtosis, and first-order autocorrelation (AC(1)) of each predictor. The predictors include maturity-specific yield (Campbell and Shiller, 1991) and forward (Fama, 1984) spreads, the first three principal components of the yield curve (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) (CP) factor, and the Ludvigson et al. (2019) (LN) factor. PC1–3, CP, and LN are based on full sample estimates. The sample period is January 1962 to December 2018.

	Campbell-Shiller				Fama-Bliss								
	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year	PC1	PC2	PC3	CP	LN
						Panel A: Descriptive statistics							
Mean	0.07	0.09	0.10	0.11	0.11	0.13	0.15	0.16	12.20	0.95	-0.17	0.14	0.14
Std. dev.	0.07	0.08	0.08	0.09	0.09	0.11	0.12	0.13	7.03	0.59	0.10	0.20	0.32
Skewness	0.88	0.41	0.13	-0.03	0.06	-0.19	-0.23	-0.17	0.48	0.11	-0.79	0.71	1.52
Kurtosis	6.34	4.99	4.36	3.96	4.05	3.64	3.27	2.95	3.07	2.80	3.91	4.48	9.82
AR(1)	0.82	0.85	0.87	0.88	0.88	0.90	0.91	0.92	0.99	0.95	0.86	0.72	0.41
						Panel B: Correlation matrix							
CS-2	1.00												
CS-3	0.98	1.00											
CS-4	0.93	0.99	1.00										
CS-5	0.89	0.96	0.99	1.00									
FB-2	0.93	0.99	0.99	0.98	1.00								
FB-3	0.83	0.93	0.98	0.99	0.97	1.00							
FB-4	0.74	0.86	0.93	0.97	0.91	0.98	1.00						
FB-5	0.67	0.80	0.88	0.93	0.86	0.95	0.99	1.00					
PC1	0.34	0.21	0.10	0.01	0.12	-0.05	-0.17	-0.25	1.00				
PC2	0.65	0.79	0.86	0.89	0.86	0.93	0.93	0.91	-0.00	1.00			
PC3	0.33	0.28	0.21	0.14	0.29	0.13	0.01	-0.08	0.01	0.00	1.00		
CP	0.47	0.50	0.52	0.53	0.49	0.50	0.51	0.52	0.30	0.56	-0.23	1.00	
LN	0.11	0.14	0.17	0.19	0.14	0.19	0.22	0.25	-0.13	0.16	-0.14	0.20	1.00

Table 3: Empirical size properties

This table reports the rejection frequency (empirical size) of the multivariate test for equal predictive ability with a nominal size of 5%, data-generating process given by (22) with $\boldsymbol{\mu} = \mathbf{0}$, and 10,000 Monte Carlo replications. We implement an unconditional test that sets $\boldsymbol{h}_t = 1$ for all t and a conditional test that sets $\boldsymbol{h}_t = (1, \tilde{h}_t)'$, and use three samples sizes referred to as short (120 observations), medium (348 observations) and long (1,000 observations). Panel A (B) reports results where $\boldsymbol{\varepsilon}_{t+1}$ in (22) is sampled from the empirical loss differentials when forecasting the 2-year (5-year) bond. The value of p indicates the dimension of the test arising from the number of comparing models less one.

	Unconditional test ($\boldsymbol{h}_t = 1$)			Conditional test ($\boldsymbol{h}_t = (1, \tilde{h}_t)'$)		
	Short	Medium	Long	Short	Medium	Long
			Panel A: 2-year bond			
$p = 1$	5.22	5.08	4.84	5.48	4.82	5.08
$p = 2$	4.65	4.99	5.09	4.85	5.27	4.98
$p = 3$	4.82	5.34	4.88	5.33	5.38	5.17
$p = 4$	4.91	5.12	5.03	4.84	5.27	5.14
$p = 5$	5.22	4.61	4.76	4.62	5.10	5.29
			Panel B: 5-year bond			
$p = 1$	4.73	4.87	5.26	3.96	4.53	4.99
$p = 2$	4.36	4.56	4.99	4.07	4.48	4.92
$p = 3$	4.05	4.47	4.99	3.79	4.56	4.78
$p = 4$	4.38	4.24	4.96	3.69	4.34	5.11
$p = 5$	4.30	4.59	5.02	3.22	4.50	4.89

Table 4: Out-of-sample results

This table reports out-of-sample R^2_{OS} values for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the out-of-sample R^2 from Campbell and Thompson (2008) and the associated Diebold and Mariano (1995) *p*-value in parenthesis for the null of no predictability implied by the EH (Panels A and B) and a static forecast combination strategy (Panel C). PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathfrak{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
Panel A: Individual bond predictors against EH				
CS	-2.73	-0.53	0.67	1.38
	(0.70)	(0.56)	(0.40)	(0.27)
FB	-0.02	1.31	1.72	1.78
	(0.50)	(0.29)	(0.22)	(0.23)
PC	-9.86	-7.64	-5.91	-4.83
	(0.92)	(0.93)	(0.92)	(0.90)
CP	-6.63	-5.29	-4.27	-3.43
	(0.96)	(0.96)	(0.94)	(0.90)
LN	-7.61	-0.48	1.93	2.43
	(0.73)	(0.52)	(0.42)	(0.39)
EW	6.08	5.28	4.89	4.58
	(0.03)	(0.02)	(0.02)	(0.02)
Panel B: Dynamic forecast combination against EH				
PMI	7.98	5.64	5.11	6.16
	(0.01)	(0.02)	(0.02)	(0.01)
$\mathfrak{U}$	9.86	6.77	6.09	4.98
	(0.01)	(0.00)	(0.00)	(0.01)
NONE	6.66	5.31	5.25	4.81
	(0.02)	(0.02)	(0.01)	(0.02)
Panel C: Dynamic forecast combination against EW				
PMI	2.02	0.39	0.23	1.66
	(0.01)	(0.24)	(0.33)	(0.06)
$\mathfrak{U}$	4.02	1.58	1.26	0.42
	(0.02)	(0.00)	(0.00)	(0.22)
NONE	0.62	0.04	0.38	0.24
	(0.21)	(0.47)	(0.16)	(0.32)

Table 5: Testing for equal (un)conditional predictive ability

This table reports full sample multivariate test statistics for equal (un)conditional predictive ability using three different conditioning variables. PMI refers to the case of $\boldsymbol{h}_t = (1, \text{PMI}_t)'$ that is designed to capture business cycle fluctuations. $\mathcal{U}$ refers to the case of $\boldsymbol{h}_t = (1, \mathcal{U}_t)'$ that is chosen to study the effect of macroeconomic uncertainty. NONE refers to an unconditional version of the tests in which $\boldsymbol{h}_t = 1$ for all t. PMI is the Purchasing Managers' Index and UNC is the macroeconomic uncertainty index of Jurado et al. (2015). p-values are presented in parenthesis. The full sample test period runs from January 1990 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
PMI	31.36	34.65	29.76	26.73
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	27.03	27.95	26.16	26.22
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	8.07	5.68	5.10	5.77
	(0.15)	(0.34)	(0.40)	(0.33)

Table 6: Inclusion frequencies across states

This table reports the inclusion frequencies of the predictor models in three different states of the world identified using the 20% and 80% quantiles of the Purchasing Managers' Index (PMI). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. EH denotes the benchmark expectations hypothesis model. The out-of-sample evaluation periods runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	1.00	1.00	1.00	1.00	0.25	0.64	0.64	0.57
FB	0.94	1.00	0.88	0.76	0.25	0.64	0.64	0.57
PC	0.45	0.36	0.42	0.42	1.00	1.00	0.93	1.00
CP	0.88	0.91	1.00	1.00	0.25	0.57	0.54	0.46
LN	0.73	0.67	0.70	0.64	0.96	1.00	1.00	1.00
EH	0.97	1.00	1.00	1.00	0.18	0.50	0.57	0.46
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	0.86	0.90	0.94	0.98	0.69	0.88	0.86	0.84
FB	0.90	0.98	0.98	0.95	0.79	0.93	0.88	0.76
PC	0.51	0.41	0.47	0.50	0.43	0.53	0.54	0.58
CP	0.65	0.68	0.68	0.62	0.62	0.71	0.63	0.51
LN	0.71	0.79	0.87	0.86	0.83	0.85	0.90	0.92
EH	0.70	0.84	0.85	0.82	0.63	0.83	0.85	0.75
	Panel C: High activity				Panel F: High uncertainty			
CS	0.58	0.75	0.85	0.93	0.95	1.00	1.00	0.98
FB	0.65	0.80	0.88	0.83	0.95	1.00	0.98	0.95
PC	0.83	0.74	0.78	0.88	0.49	0.51	0.44	0.28
CP	0.53	0.53	0.45	0.08	0.95	1.00	0.98	0.98
LN	1.00	1.00	1.00	0.95	0.72	0.98	0.95	1.00
EH	0.33	0.45	0.45	0.48	1.00	1.00	1.00	0.98

Table 7: Out-of-sample R^2 across states

This table reports out-of-sample R^2_{OS} values for various linear predictive models for bond excess return conditional on states identified by the Purchasing Manager's Index (PMI) and the the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the out-of-sample R^2 from Campbell and Thompson (2008) relative to the expectations hypothesis. High (low) states are identified using the 80% (20%) quantiles of the time series of PMI and $\mathcal{U}$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	-19.73	-10.56	-6.41	-3.85	14.33	9.87	7.79	6.40
FB	-13.17	-6.09	-3.43	-2.22	15.19	8.57	5.00	3.03
PC	-37.08	-26.84	-20.98	-17.02	24.53	16.39	12.09	10.05
CP	-13.30	-7.27	-3.54	-0.83	-83.09	-42.16	-23.01	-12.28
LN	-23.94	-15.44	-12.03	-11.24	14.05	10.04	7.52	5.73
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	2.16	1.68	1.95	2.20	-2.64	-0.51	0.61	1.28
FB	4.26	3.25	2.90	2.62	1.09	1.78	2.00	2.00
PC	-5.04	-5.29	-4.19	-3.26	-4.86	-4.46	-3.44	-2.74
CP	-6.09	-6.42	6.06	-5.61	-1.03	-2.14	-2.56	-2.90
LN	-2.58	3.8	5.78	6.25	-2.15	4.31	6.44	7.16
	Panel C: High activity				Panel F: High uncertainty			
CS	6.86	5.32	5.38	5.78	-4.29	-2.03	-0.67	0.30
FB	4.44	3.79	3.83	4.02	-2.82	-0.55	0.50	0.98
PC	20.26	12.96	9.43	7.36	-19.49	-16.08	-13.96	-12.51
CP	9.32	7.30	6.43	6.00	-7.19	-4.90	-3.26	-1.81
LN	-8.41	-2.57	0.26	2.18	-18.80	-11.59	-8.99	-8.73

Table 8: Correlations between forecasts and economic activity

This table reports correlation coefficients between out-of-sample generated forecasts from individual bond predictors (Panel A) and the dynamic forecast strategy (Panel B) and economic activity as measured by the Purchasing Managers' Index (PMI). We report *p*-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Individual bond predictors				
CS	0.36	0.31	0.27	0.23
	(0.00)	(0.00)	(0.00)	(0.00)
FB	0.27	0.19	0.14	0.09
	(0.00)	(0.00)	(0.01)	(0.08)
PC	0.33	0.36	0.36	0.35
	(0.00)	(0.00)	(0.00)	(0.00)
CP	0.15	0.16	0.16	0.16
	(0.01)	(0.00)	(0.00)	(0.00)
LN	-0.38	-0.38	-0.38	-0.38
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.07	0.14	0.16	0.17
	(0.17)	(0.01)	(0.00)	(0.00)
EW	0.01	0.01	0.00	-0.01
	(0.88)	(0.87)	(0.98)	(0.90)
Panel B: Dynamic forecast combination				
PMI	-0.39	-0.40	-0.39	-0.35
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	-0.40	-0.36	-0.37	-0.39
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.29	-0.26	-0.27	-0.28
	(0.00)	(0.00)	(0.00)	(0.00)

Table 9: Economic Value

This table reports certainty equivalent return (CER) gains for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the CER gains relative to the expectations hypothesis (Panels A and B) and a static forecast combination strategy (Panel C). PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathfrak{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $= 10$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
Panel A: Individual bond predictors against EH				
CS	-0.64	-0.35	0.10	0.45
	(0.90)	(0.75)	(0.43)	(0.20)
FB	-0.43	-0.12	0.32	0.58
	(0.84)	(0.62)	(0.24)	(0.17)
PC	-1.65	-1.78	-1.65	-1.44
	(0.98)	(0.96)	(0.93)	(0.89)
CP	-0.66	-0.83	-0.76	-0.48
	(0.96)	(0.95)	(0.87)	(0.73)
LN	0.85	1.75	2.32	2.74
	(0.00)	(0.00)	(0.00)	(0.00)
EW	0.10	0.34	0.86	1.07
	(0.36)	(0.16)	(0.03)	(0.02)
Panel B: Dynamic forecast combination against EH				
PMI	0.39	0.59	1.05	1.43
	(0.08)	(0.06)	(0.02)	(0.00)
$\mathfrak{U}$	0.26	0.60	1.17	1.18
	(0.16)	(0.05)	(0.01)	(0.02)
NONE	0.17	0.33	0.92	1.16
	(0.26)	(0.19)	(0.03)	(0.02)
Panel C: Dynamic forecast combination against EW				
PMI	0.28	0.25	0.19	0.37
	(0.01)	(0.04)	(0.09)	(0.03)
$\mathfrak{U}$	0.16	0.25	0.31	0.12
	(0.06)	(0.02)	(0.01)	(0.27)
NONE	0.06	-0.02	0.06	0.09
	(0.13)	(0.56)	(0.24)	(0.26)

Table 10: CER gains across states

This table reports certainty equivalent return (CER) gains for various linear predictive models for bond excess return conditional on states identified by the Purchasing Manager's Index (PMI) and the the macroeconomic uncertainty index ($\mathcal{U}$) proposed in Jurado et al. (2015). We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the CER gain relative to the expectations hypothesis. High (low) states are identified using the 80% (20%) quantiles of the time series of PMI and $\mathcal{U}$. CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $= 10$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year	2-year	3-year	4-year	5-year
	Panel A: Low activity				Panel D: Low uncertainty			
CS	-3.22	-2.41	-2.04	-2.05	0.02	0.08	1.29	1.68
FB	-2.44	-1.82	-1.83	-2.70	0.01	0.16	0.98	0.86
PC	-7.17	-7.06	-6.92	-6.77	0.02	0.94	2.28	2.76
CP	-2.42	-1.85	-0.91	-0.05	0.08	0.78	2.13	2.70
LN	1.22	2.47	2.05	0.25	0.02	0.41	1.30	1.55
	Panel B: Normal activity				Panel E: Normal uncertainty			
CS	-0.19	0.01	0.32	0.48	-0.36	-0.21	0.03	0.24
FB	-0.10	0.04	0.13	-0.01	-0.32	-0.21	0.01	0.07
PC	-1.00	-1.20	-0.96	-0.72	-0.66	-0.72	-0.58	-0.49
CP	-0.35	-0.78	-1.03	-1.02	-0.58	-0.80	-0.68	-0.69
LN	0.82	1.32	1.50	1.78	0.90	1.47	1.72	1.94
	Panel C: High activity				Panel F: High uncertainty			
CS	0.51	0.57	1.43	1.82	-1.41	-0.57	-0.06	-0.18
FB	0.25	0.23	1.17	1.38	-0.82	-0.38	-0.55	-1.69
PC	1.75	2.44	2.84	2.56	-5.19	-5.62	-5.67	-5.60
CP	0.34	0.80	1.61	1.86	-0.74	-1.14	-1.98	-1.42
LN	0.59	1.34	1.99	1.88	1.16	2.30	1.71	0.29

Figure 1: Bond excess returns

This figure plots times series of monthly bond excess returns (in percentage) for Treasury bonds with maturities ranging from two to five years. Shaded areas represent NBER recession dates. Monthly bond returns are in excess of the implied yield on a one-month Treasury bill rate. Yield data are end-of-month and have been obtained from Gürkaynak et al. (2007) over the period January 1962 to December 2018.

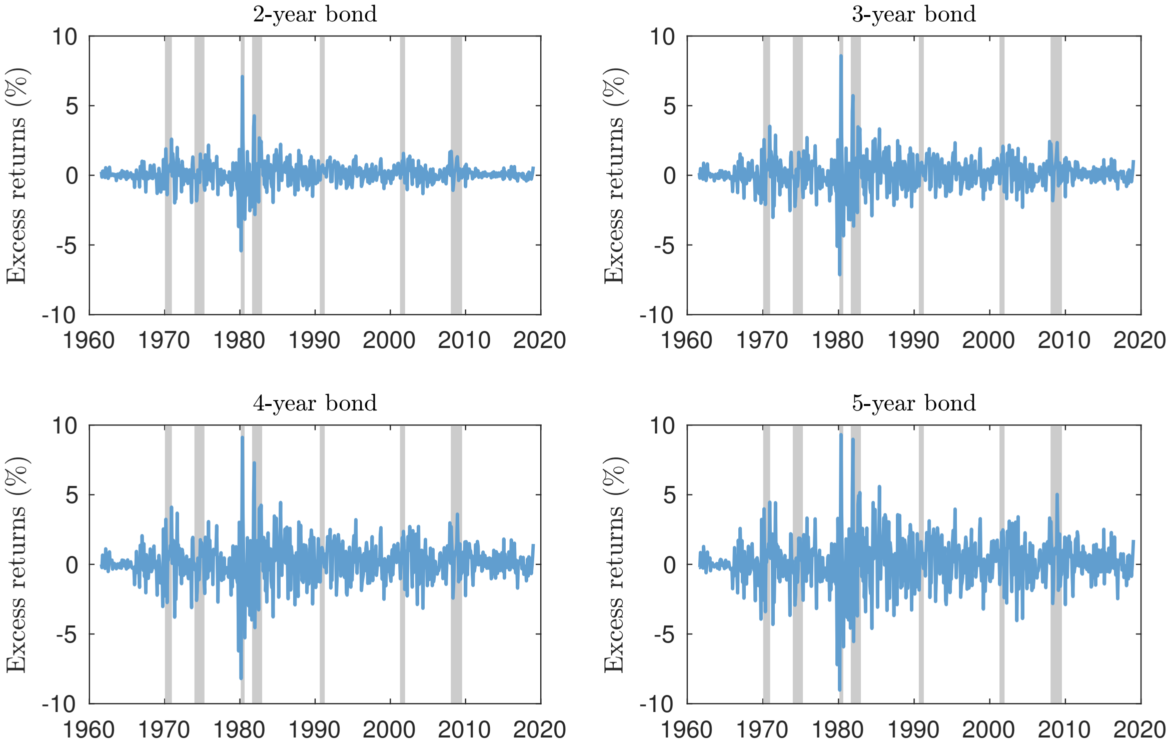


Figure 2: Conditioning variables

This figure shows times series of the Purchasing managers' index (PMI) published by the Institute for Supply Management and the macroeconomic uncertainty ($\mathcal{U}$) index from Jurado et al. (2015). Green (yellow) shaded ares represent periods of (high) low activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. The sample period covers January 1962 to December 2018.

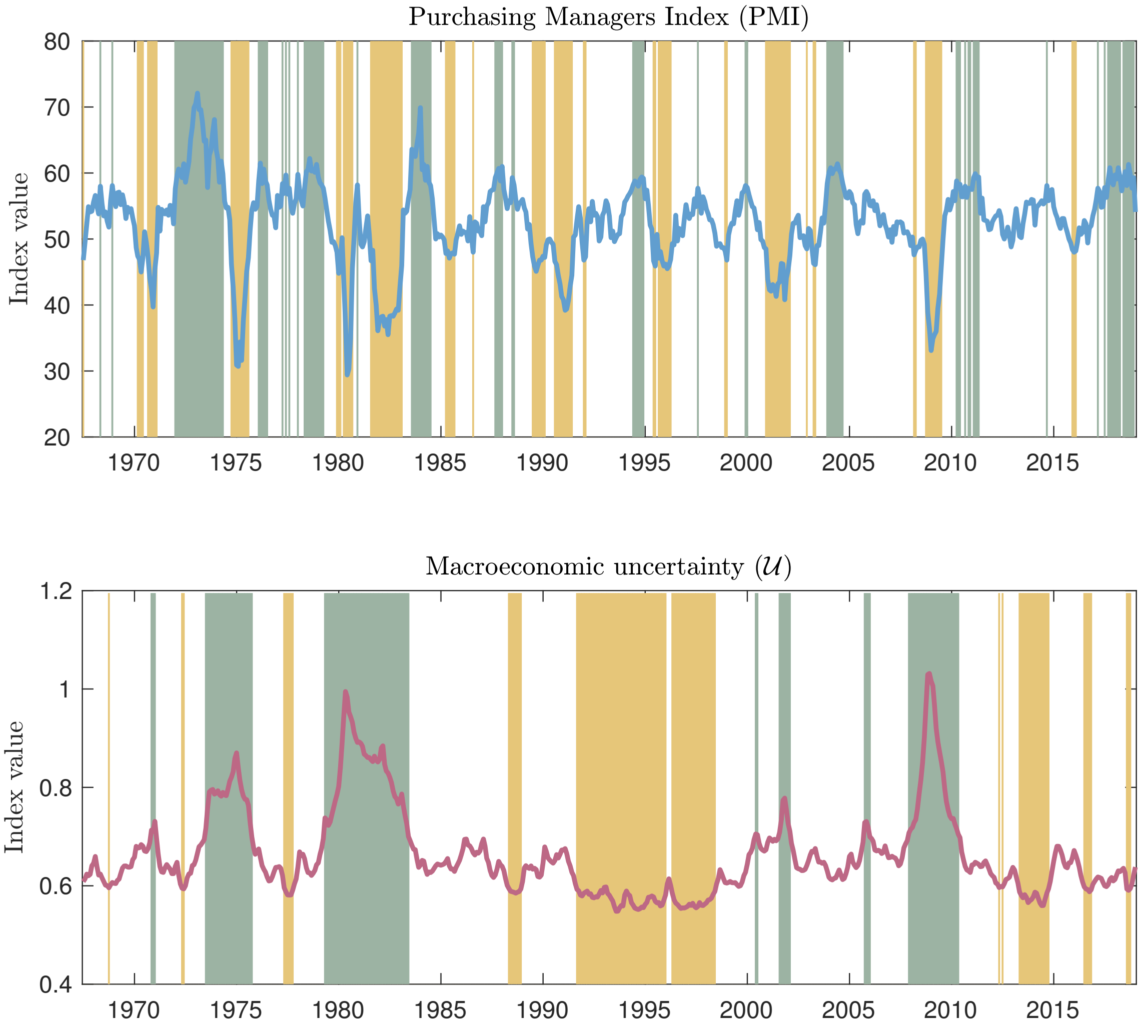


Figure 3: Empirical power curves

This figure shows the rejection frequency (empirical power) of the multivariate test for equal predictive ability with a nominal size of 5% and data-generating process given by (22) with the first element in $\boldsymbol{\mu}$ deviating and the remaining elements are set to zero. The first element of $\boldsymbol{\mu}$ is set to $c\hat{\eta}$ where $\hat{\eta}$ is the average absolute loss differentials across all models within the low and high economic activity states defined in the empirical section and $c \in [0, 2.5]$. We use 10,000 Monte Carlo replications. We implement a conditional test that sets $\boldsymbol{h}_t = (1, \tilde{h}_t)'$, and use three samples sizes referred to as short (120 observations), medium (348 observations) and long (1,000 observations). The left (right) panel depicts results where $\boldsymbol{\varepsilon}_{t+1}$ in (22) is sampled from the empirical loss differentials when forecasting the 2-year (5-year) bond. The value of p indicates the dimension of the test arising from the number of comparing models less one.

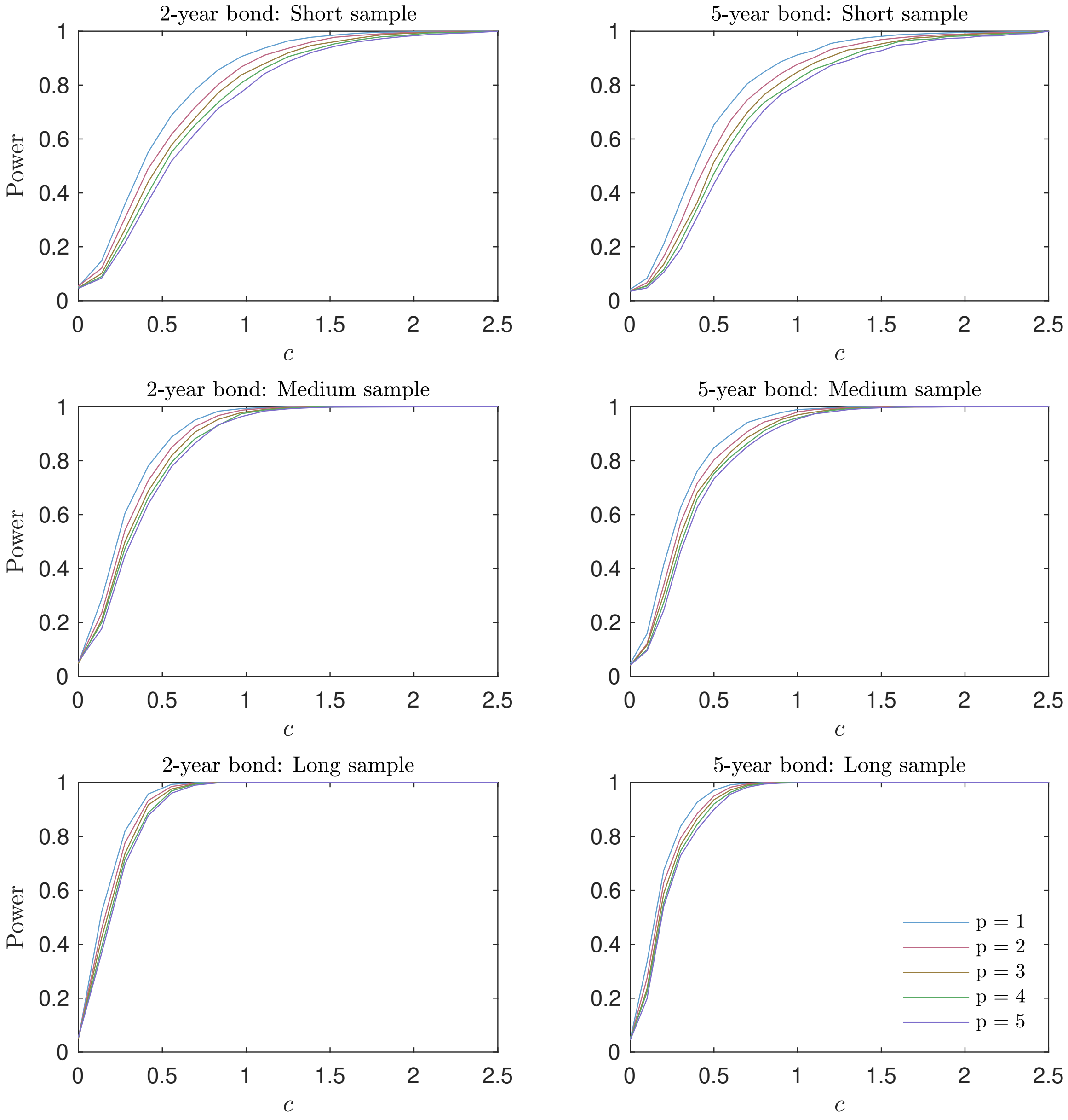


Figure 4: Relative forecasting performance

This figure plots the recursively updated cumulative difference in the squared prediction errors from the EH benchmark model and the ith predictor model over the out-of-sample evaluation period. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. We also consider a simple equal-weighted combination of the individual forecasts. A positive (negative) slope indicates that the predictive model delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Managers' Index (PMI) and uncertainty in the index developed by Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

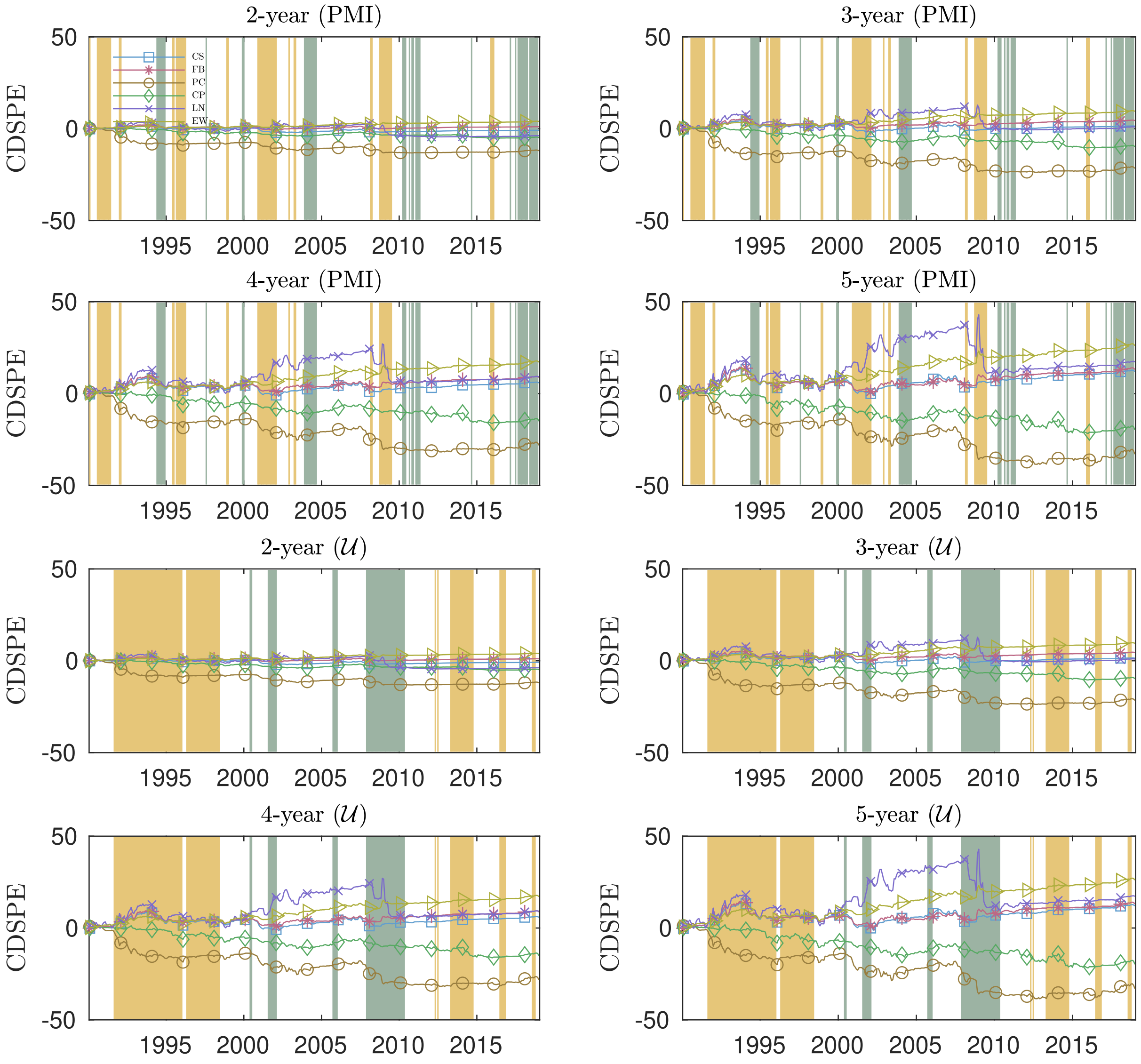
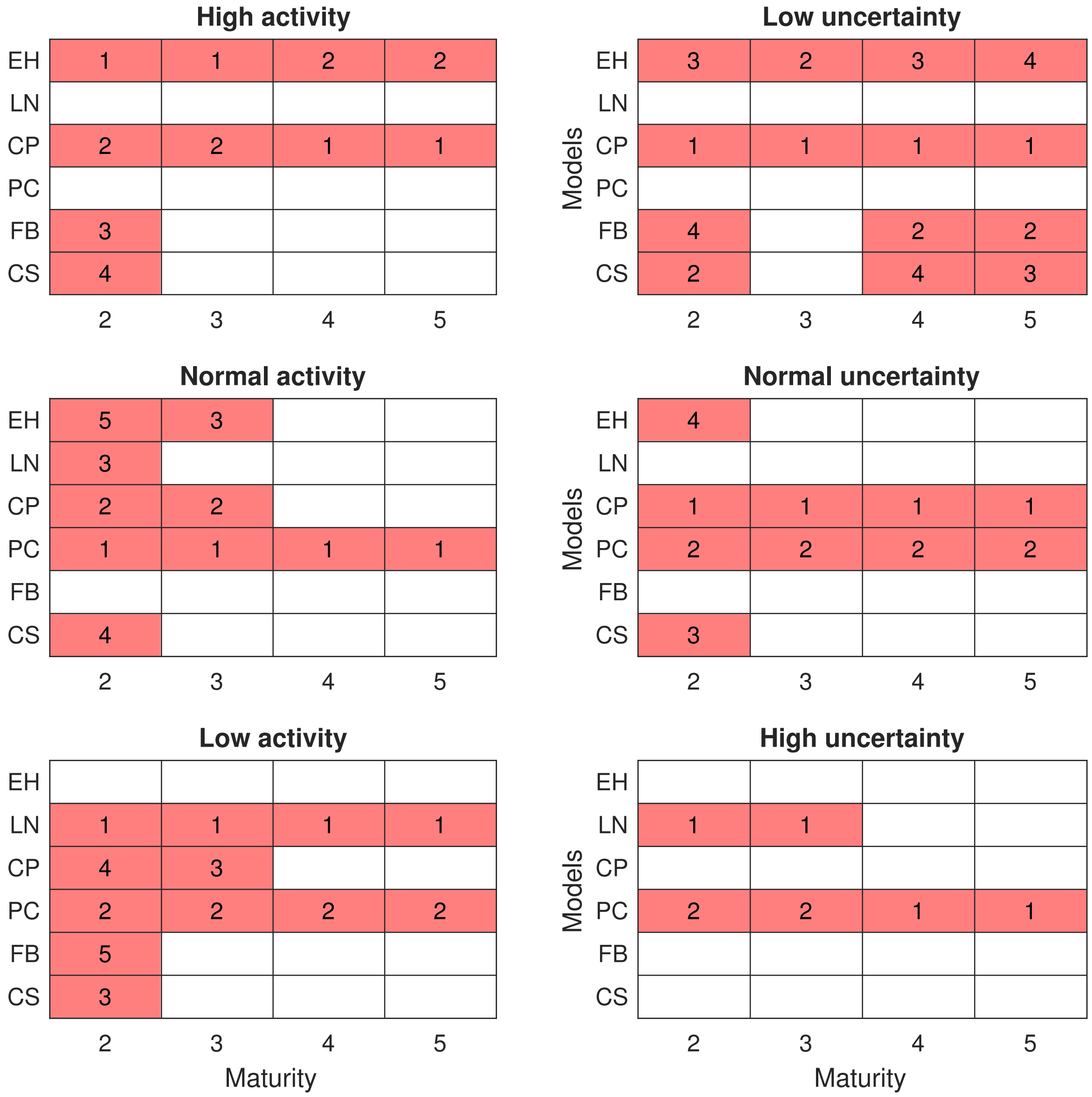


Figure 5: Full sample elimination order

This figure displays the full sample elimination order of predictive model in high, normal, and low states separately for the Purchasing Managers' Index (PMI) (left graphs) and the macroeconomic uncertainty index ($\mathcal{U}$) of Jurado et al. (2015) (right graphs) using the 20% and 80% quantiles of their time series. White squares denote models included in the best set of models and numbered tiles denotes eliminated models and their elimination order. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. The out-of-sample evaluation periods runs from January 2000 to December 2018.



High activity

Models	2	3	4	5
EH	1	1	2	2
LN				
CP	2	2	1	1
PC				
FB	3			
CS	4			

Low uncertainty

Models	2	3	4	5
EH	3	2	3	4
LN				
CP	1	1	1	1
PC				
FB	4		2	2
CS	2		4	3

Normal activity

Models	2	3	4	5
EH	5	3		
LN	3			
CP	2	2		
PC	1	1	1	1
FB				
CS	4			

Normal uncertainty

Models	2	3	4	5
EH	4			
LN				
CP	1	1	1	1
PC	2	2	2	2
FB				
CS	3			

Low activity

Models	2	3	4	5
EH				
LN	1	1	1	1
CP	4	3		
PC	2	2	2	2
FB	5			
CS	3			

High uncertainty

Models	2	3	4	5
EH				
LN	1	1		
CP				
PC	2	2	1	1
FB				
CS				

Figure 6: Dynamic forecast combinations

This figure plots the recursively updated cumulative difference in the squared prediction errors from the EH benchmark model and the dynamic forecast combination forecast for each of the tree conditioning cases. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

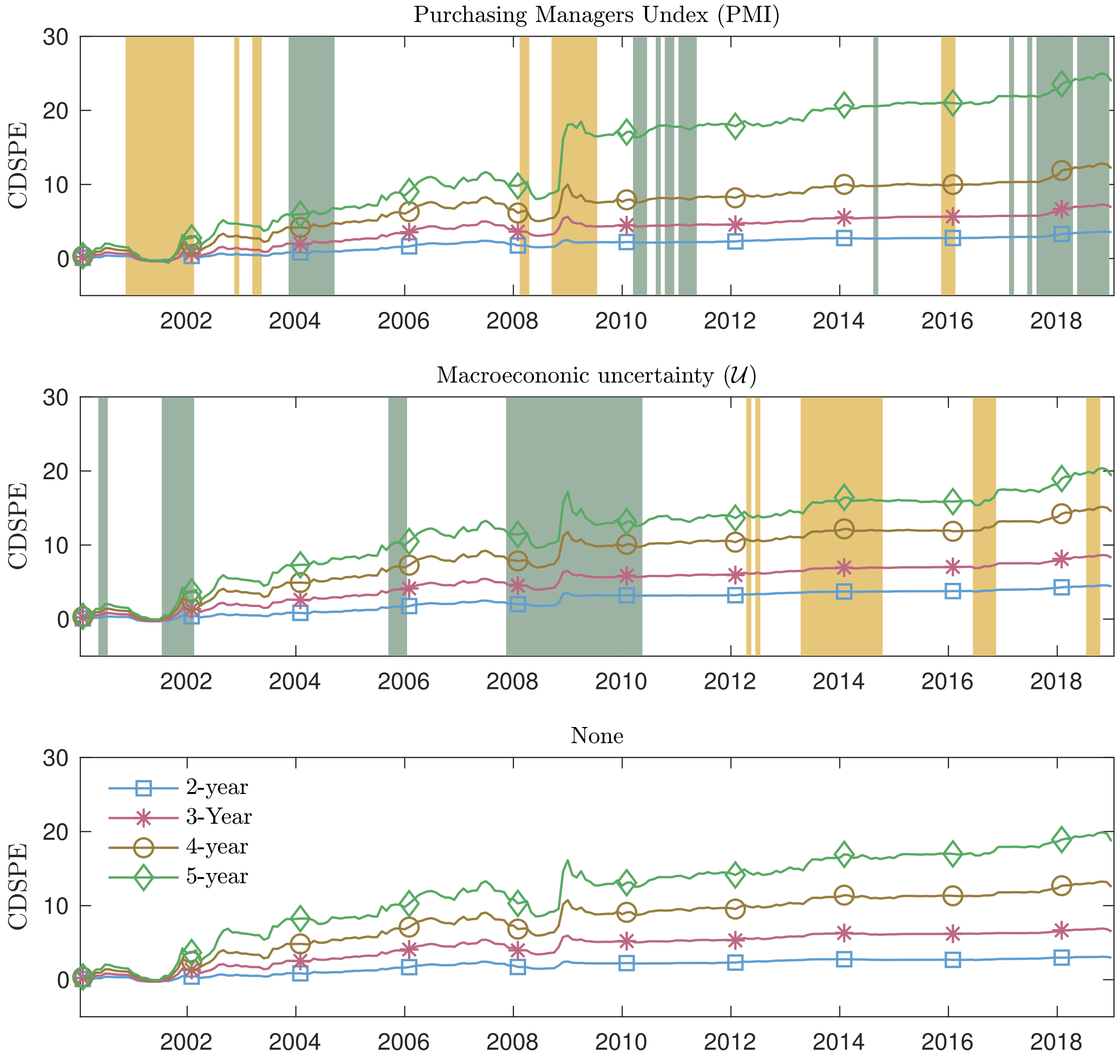
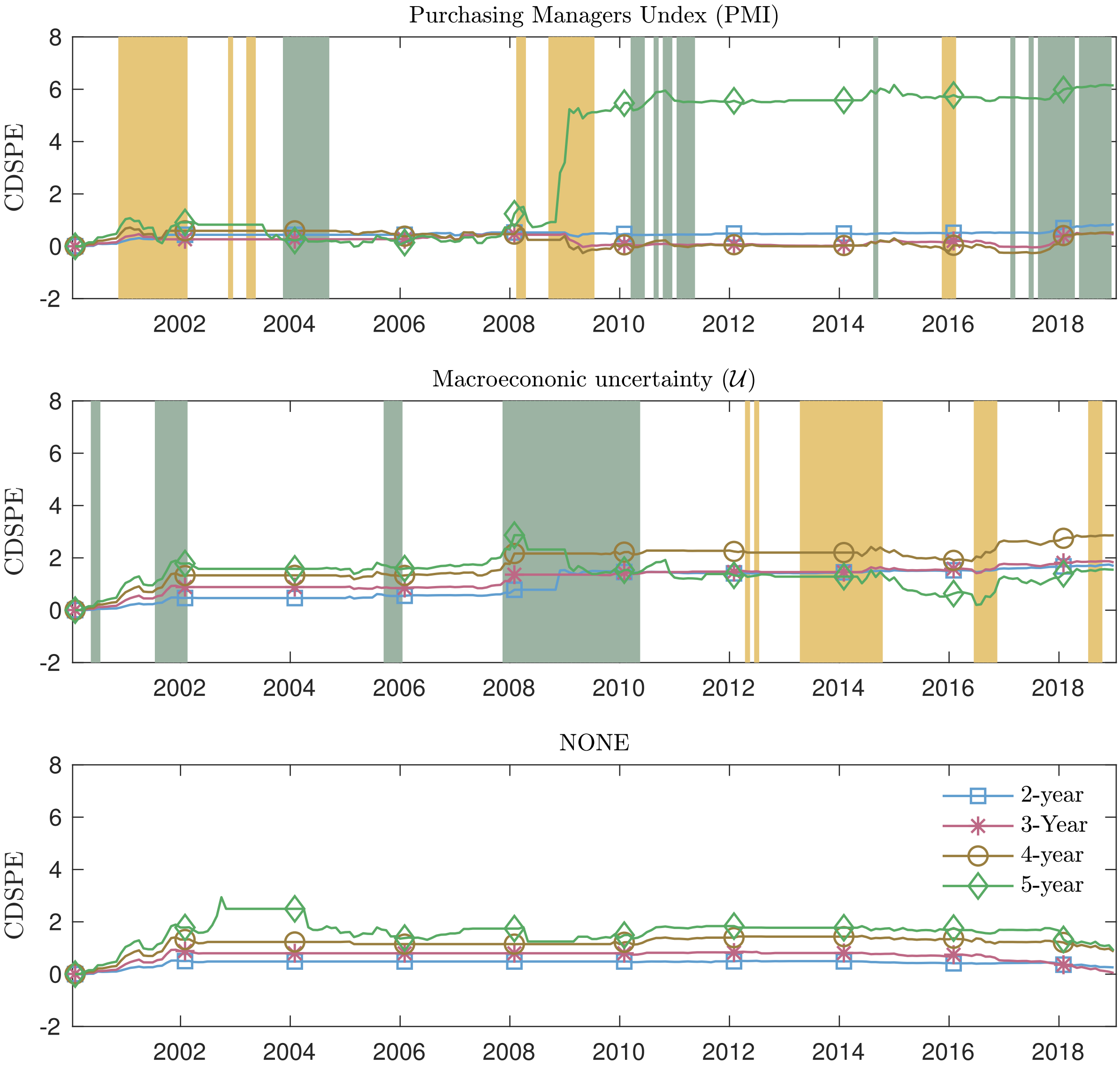


Figure 7: Dynamic versus static forecast combination

This figure plots the recursively updated cumulative difference in the squared prediction errors from a static equal-weighted forecast combination benchmark and the dynamic forecast combination forecast for each of the tree conditioning cases. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the static equal-weighted forecast combination benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.



Purchasing Managers Undex (PMI)

Macroecononic uncertainty ($\mathcal{U}$)

NONE

Figure 8: Inclusion plots across states

This figure displays the inclusion of each predictive model into the best set of models. Green (yellow) shaded ares represent periods of high (low) states of the Purchasing Managers' Index (PMI) (left) and the Jurado et al. (2015) macroeconomic uncertainty index ($\mathcal{U}$) (right) identified using the 20% and 80% quantiles of the series. White areas are normal times. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. EH denotes the benchmark expectations hypothesis model. Inclusion of a predictive model is marked with +. The out-of-sample evaluation periods runs from January 2000 to December 2018.

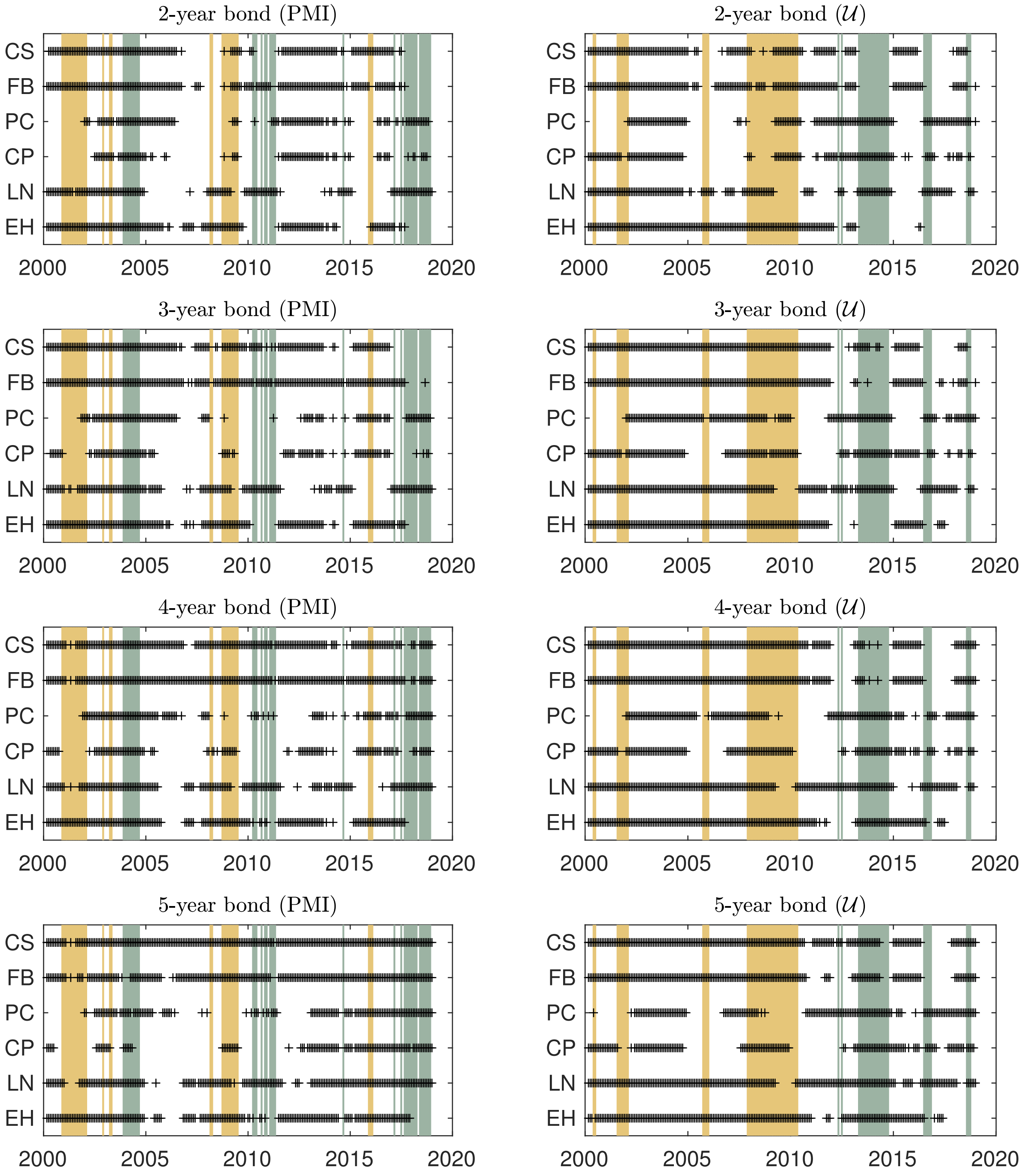
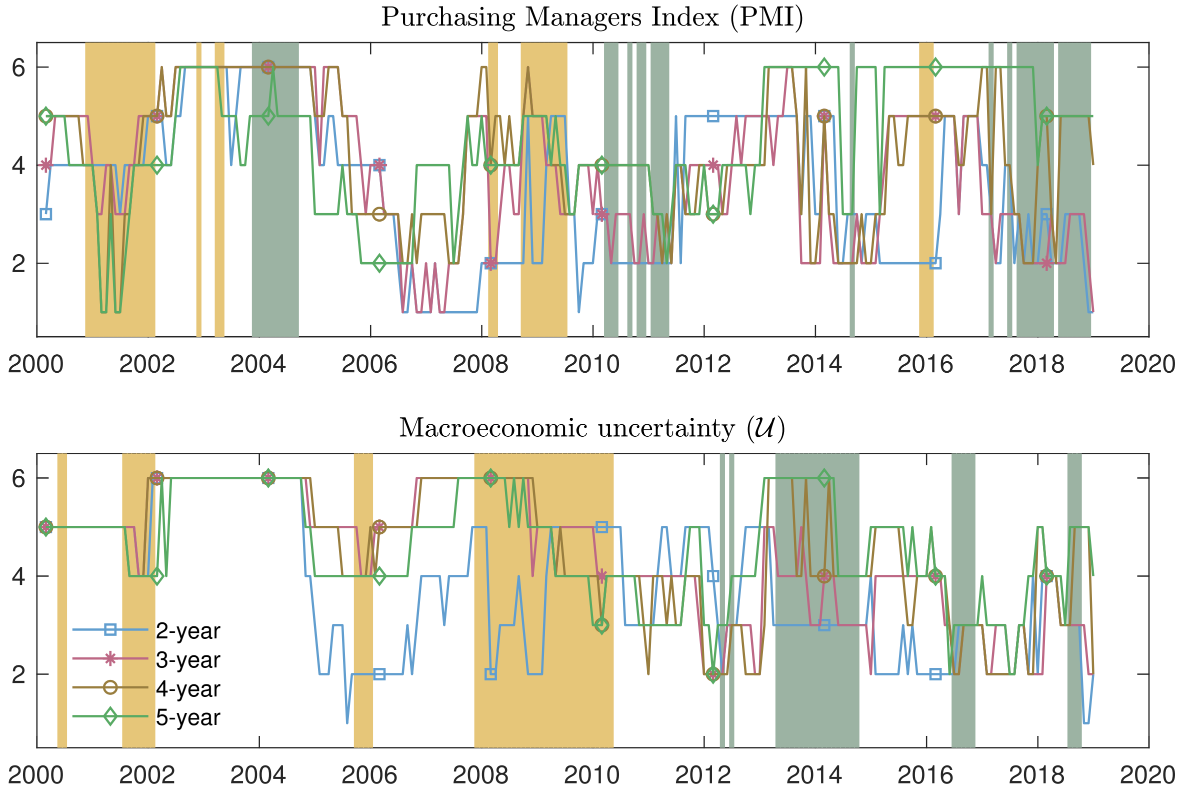


Figure 9: Size of the set of best models

This figure illustrates the size of the set of best predictive models for each of the four bond maturities and conditioning variables. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Manager's Index (PMI) published by the Institute for Supply Management and uncertainty is the macroeconomic uncertainty index ($\mathfrak{U}$) proposed in Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.



Purchasing Managers Index (PMI)

Macroeconomic uncertainty ($\mathcal{U}$)

Figure 10: Bond risk premia forecasts for dynamic combination strategy
This figure illustrates the time series behavior of bond risk premia forecasts originating from our dynamic forecast combination strategy. PMI is the Purchasing Managers' Index published by the Institute for Supply Management and $\mathcal{U}$ is the macroeconomic uncertainty index proposed in Jurado et al. (2015). The out-of-sample forecasting periods runs from January 2000 to December 2018.

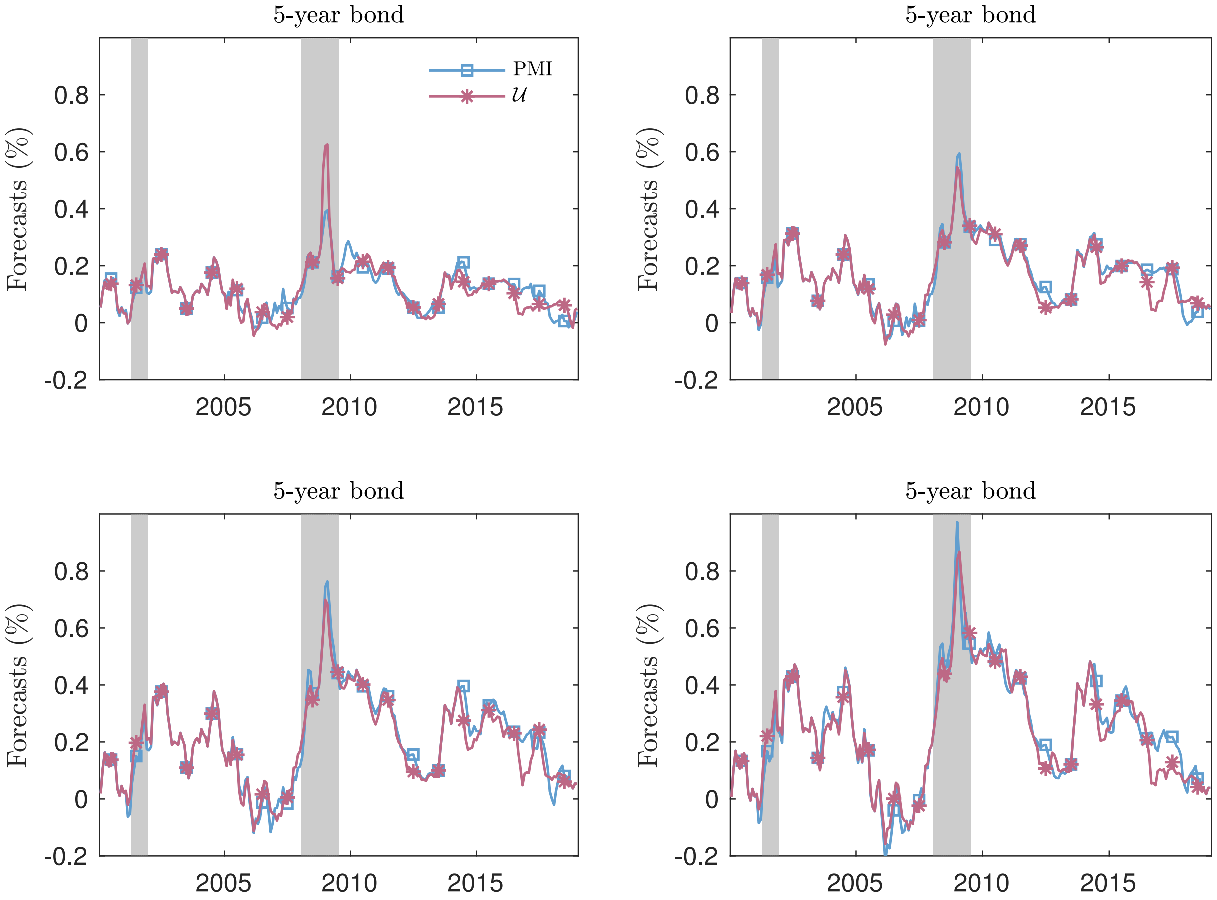


Figure 11: Dynamic forecast combinations: CER gains

This figure plots the recursively updated cumulative difference in realized utility from the dynamic forecast combination forecast for each of the tree conditioning cases and the EH benchmark model. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.

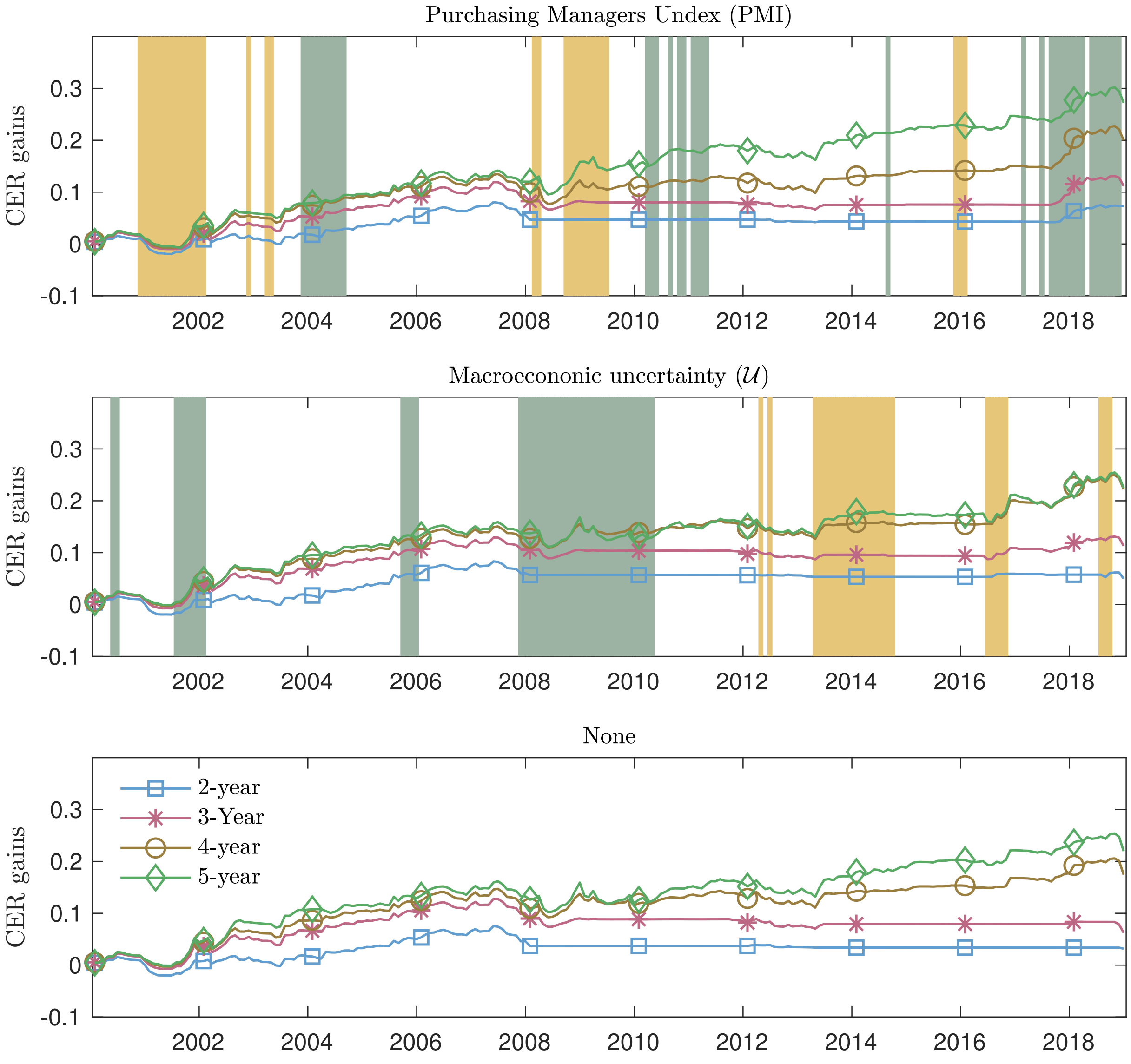
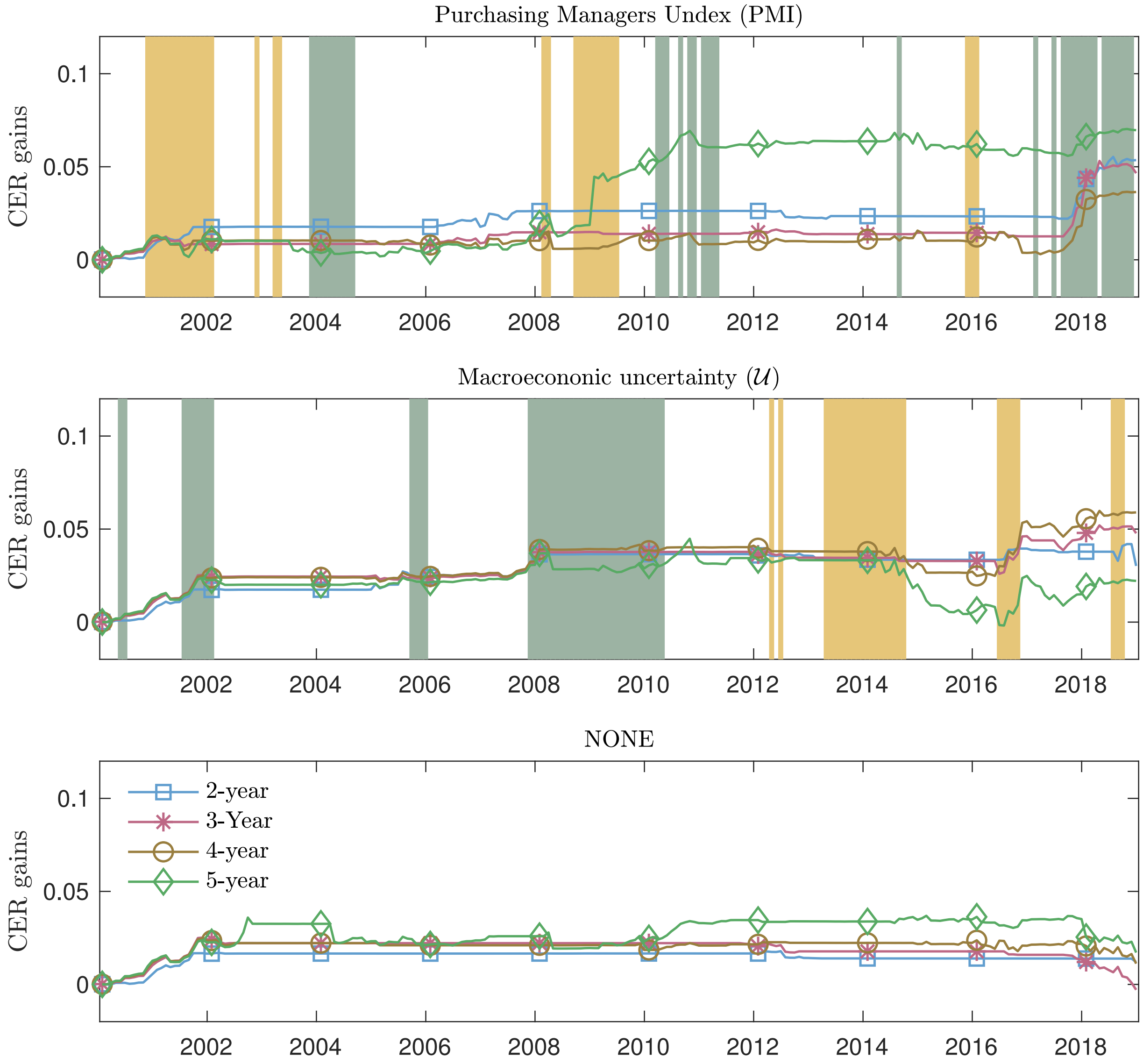


Figure 12: Dynamic versus static forecast combination: CER gains

This figure plots the recursively updated cumulative difference in the squared prediction errors from the dynamic forecast combination forecast for each of the tree conditioning cases and a static equal-weighted forecast combination benchmark. We consider the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index ($\mathfrak{U}$) from Jurado et al. (2015) as our conditioning variables along with an unconditional version labeled NONE. A positive (negative) slope indicates that the dynamic forecast combination delivers more (less) accurate forecasts than the static equal-weighted forecast combination benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where high (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.



Purchasing Managers Undex (PMI)

Macroecononic uncertainty ($\mathcal{U}$)

NONE

Internet Appendix for

Predicting bond return predictability

(not intended for publication)

IA.A. Theoretical results, assumptions, and proofs

This section explains the Giacomini and White (2006) assumptions used in Theorem 1 along with its proof. The outline of several of the proofs follow Giacomini and White (2006), making the necessary adjustments to account for the multivariate nature of our tests. We also provide theoretical results with associated proofs for the case of multi-step ahead forecasting, $\tau > 1$, and the unconditional case, $\mathcal{G}_t = \{\emptyset, \quad \}$.

IA.A.1. One-step ahead forecasting and Giacomini and White (2006) assumptions

In the one-step ahead case, $\tau = 1$, we impose the following assumptions that are adopted from Giacomini and White (2006).

Assumption 1. *$\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are ϕ-mixing with $\phi(t) = O\left(t^{-r/(2r-1)-\iota}\right)\Big($, $r \geq 1$, or -mixing with $(t) = O\left(t^{-\frac{r}{r-1}-\iota}\right)\Big($,$r > 1$, for some $\iota > 0$.*

Assumption 1 imposes relatively mild restrictions on the dependence structure and heterogeneity of data. We do not impose the stricter and common (covariance) stationarity assumption as used in for instance Diebold and Mariano (1995) and Mariano and Preve (2012). Specifically, data may exhibit arbitrary structural changes, which is a common feature found in many empirical studies within e.g. macroeconomic prediction (see e.g. Stock and Watson (2003) and Schrimpf and Wang (2010)), stock return prediction (see e.g. Fama and French (1997) and Paye and Timmermann (2006)), and exchange rate prediction (see e.g. Giacomini and Rossi (2010)) to name a few.

Assumption 2. *$\mathbb{E}[|\boldsymbol{d}_{t+1,i}|^{2(r+\delta)}] < \infty$ for some $\delta > 0$, $i = 1, \ldots, qp$, and for all t, where subscript i indicate the i'th element of $\boldsymbol{d}_{t+1}$.*

Assumption 3. *$\boldsymbol{\Sigma}_T \equiv T^{-1} \sum_{t=1}^{T} \mathbb{E}[\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1}]$ is uniformly positive definite.*

Assumptions 2-3 are mainly technical assumptions ensuring (uniformly) bounded moments of data and positive definiteness of the asymptotic variance. Both of these assumptions are common in the forecast evaluation literature.

IA.A.1.1. Proof of Theorem 1 The proof of part A. and B. adopts the necessary steps in Giacomini and White (2006). We start by proving part A. Let $\boldsymbol{d}_{t+1} = \boldsymbol{h}_t \quad \Delta \boldsymbol{L}_{+1}$ and write

$$\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1} = g(\boldsymbol{h}_t, \boldsymbol{w}_{t+1}, \ldots, \boldsymbol{w}_{t-m}) \tag{IA.A.1}$$

for some measurable function g. Since $m < \infty$, and $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are mixing of the same size according to Assumption 1, it follows from Theorem 3.49 in White (2001) that $\{\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1}\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$.

By Assumption 2 there exists a $\bar{C} \in \mathbb{R}_+$ and $\delta > 0$ such that $\mathbb{E}[|\boldsymbol{d}_{m,t+1,i}|^{2(r+\delta)}] < \bar{C} < \infty$ for $i = 1, \ldots, qp$ and for all t, where subscript i indicates the i'th element in $\boldsymbol{d}_{t+1}$. Hence, by the Cauchy-Schwartz inequality, one obtains

$$\mathbb{E}[|\boldsymbol{d}_{t+1,i}\boldsymbol{d}_{t+1,j}|^{r+\delta}] \leq \mathbb{E}[|\boldsymbol{d}^2_{t+1,i}|^{r+\delta}]^{1/2}\mathbb{E}[|\boldsymbol{d}^2_{t+1,j}|^{r+\delta}]^{1/2} < \bar{C} \tag{IA.A.2}$$

for $i, j = 1, \ldots, qp$ and for all t. By Corollary 3.48 in White (2001), it then follows that $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Furthermore, by Assumption 2 it follows that $\boldsymbol{\Sigma}_T$ is finite and by Assumption 3 it is uniformly positive definite.

Next, let $\boldsymbol{\lambda} \in \mathbb{R}^{qp}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider

$$\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}_{t+1} = T^{-1/2}\sum_{t=1}^{T-1}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}. \tag{IA.A.3}$$

Let $\tilde{\boldsymbol{\lambda}}_i$ denote the i'th element of the product $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}$, such that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1} = \sum_{i=1}^{qp}\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{t+1,i}$. Hence, under the null hypothesis

$$\mathbb{E}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}|\mathcal{G}_t] = \mathbb{E}\left[\sum_{i=1}^{qp}\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{t+1,i}|\mathcal{G}_t\right]\Bigg(= \sum_{i=1}^{qp}\Bigg(\tilde{\boldsymbol{\lambda}}_i\mathbb{E}[\boldsymbol{d}_{t+1,i}|\mathcal{G}_t] = 0, \tag{IA.A.4}$$

by measurability of $\tilde{\boldsymbol{\lambda}}_i$, such that the sequence $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}, \mathcal{G}_t\}$ is an MDS. The asymptotic variance is

$$\begin{aligned}\sigma_d^2 &= \mathrm{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}]\\ &= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\mathrm{Var}[\sqrt{T}\bar{\boldsymbol{d}}]\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda}\\ &= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda}\\ &= 1 \end{aligned} \tag{IA.A.5}$$

for sufficiently large T. Furthermore, since $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$ it follows by the Continuous Mapping Theorem that

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\Bigg(\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}'_{t+1}\boldsymbol{d}_{t+1}\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \sigma_d^2\\ &= \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} - \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} \xrightarrow{\mathbb{P}} 0.\end{aligned} \tag{IA.A.6}$$

Lastly, we need to check that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}$ has absolute $2+\delta$ moment. By Minkowski's inequality and Assumption 2 we obtain

$$\begin{aligned}\mathbb{E}[|\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+1}|^{2+\delta}] &= \mathbb{E}\left[\sum_{i=1}^{qp}\Bigg(\tilde{\boldsymbol{\lambda}}_i\boldsymbol{d}_{t+1,i}{}^{2+\delta}\right]\Bigg(\\ &\leq \sum_{i=1}^{qp}\tilde{\boldsymbol{\lambda}}_i\mathbb{E}\left[|\boldsymbol{d}_{t+1,i}|^{2+\delta}\right]^{1/(2+\delta)}\Bigg)^{2+\delta} < \infty.\end{aligned} \tag{IA.A.7}$$

Consequently, we can apply the CLT for MDS and deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0,1)$. By the Cramér-Wold device it then follows that

$$\boldsymbol{\Sigma}^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0, \boldsymbol{I}_{qp}). \tag{IA.A.8}$$

Since $\hat{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$, we deduce that

$$\sqrt{T}(\hat{\boldsymbol{\Sigma}}_T^{-1/2}\bar{\boldsymbol{d}})'\sqrt{T}\boldsymbol{\Sigma}_T^{-1/2}\bar{\boldsymbol{d}} = T\bar{\boldsymbol{d}}'\hat{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}} = S_h \xrightarrow{d} \chi^2(qp), \tag{IA.A.9}$$

as $T \to \infty$. □

We now prove part B. By the same arguments as in the proof for part B., it follows that the sequence $\{\boldsymbol{d}_{t+1}\}$ is mixing of the same size as $\{\boldsymbol{w}_t\}$ and $\{\boldsymbol{h}_t\}$. Furthermore, Assumption 2 ensures that each element of $\boldsymbol{d}_{t+1}$ is bounded uniformly in t, such that

$$\bar{\boldsymbol{d}} - \mathbb{E}[\bar{\boldsymbol{d}}] \xrightarrow{\mathbb{P}} 0 \tag{IA.A.10}$$

by Corollary 3.48 in White (2001). Under the alternative hypothesis there exists $\eta > 0$ such that $\mathbb{E}[\bar{\boldsymbol{d}}_m']\mathbb{E}[\bar{\boldsymbol{d}}_m] > 2\eta$ for T sufficiently large. It follows that

$$\begin{aligned} \mathbb{P}[\bar{\boldsymbol{d}}'\bar{\boldsymbol{d}} > \eta] &\geq \mathbb{P}[\bar{\boldsymbol{d}}'\bar{\boldsymbol{d}} - \mathbb{E}[\bar{\boldsymbol{d}}']\mathbb{E}[\bar{\boldsymbol{d}}] > -\eta] \\ &\geq \mathbb{P}[|\bar{\boldsymbol{d}}_m'\bar{\boldsymbol{d}} - \mathbb{E}[\bar{\boldsymbol{d}}']\mathbb{E}[\bar{\boldsymbol{d}}]| < \eta] \to 1, \end{aligned} \tag{IA.A.11}$$

where the convergence to unity is due to (IA.A.10). By identical arguments as the proof of part B., $\boldsymbol{d}_{t+1}'\boldsymbol{d}_{t+1}$ is mixing with the same size as $\{\boldsymbol{w}_t\}$ and each element is uniformly bounded in t. Corollary 3.48 in White (2001) can then be applied, and it follows that $\hat{\boldsymbol{\Sigma}}_T$ is a consistent estimator of $\boldsymbol{\Sigma}_T$. By Assumption 3, $\boldsymbol{\Sigma}_T$ is uniformly positive definite. Let $c \in \mathbb{R}_+$. It then follows from Theorem 8.13 in White (1994) that

$$\mathbb{P}[S_h > c] \to 1, \quad \text{as } T \to \infty. \tag{IA.A.12}$$

$\square$

Lastly, we prove part C. Let $\boldsymbol{L}_{t+1}^*$ be an arbitrary permutation of the forecasting losses, i.e. $\boldsymbol{L}_{t+1}^* = \boldsymbol{P}\boldsymbol{L}_{t+1}$, where $\boldsymbol{P}$ is a $(p+1)\times(p+1)$ permutation matrix and $\boldsymbol{L}_{t+1} = (L_{t+1}^1, \ldots, L_{t+1}^{p+1})'$. Define the $p \times (p+1)$ matrix $\boldsymbol{D}$ by

$$\boldsymbol{D} = \begin{bmatrix} 1 & -1 & 0 & \ldots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \ldots & 0 & 1 & -1 \end{bmatrix}$$

such that $\Delta \boldsymbol{L}^*_{t+1} = \boldsymbol{D}\boldsymbol{L}^*_{t+1} = \boldsymbol{D}\boldsymbol{P}\boldsymbol{L}_{t+1}$. In total, the number of permutations of the forecast losses at each point of time t is $(p+1)!$. Mariano and Preve (2012) show that there always exists a nonsingular matrix $\boldsymbol{B}$ of dimension $p \times p$ such that $\boldsymbol{B}\Delta\boldsymbol{L}_{t+1} = \Delta\boldsymbol{L}^*_{t+1}$. Consequently, define the $qp \times qp$ matrix $\boldsymbol{A} = (\boldsymbol{I}_q \quad \boldsymbol{B})$, where $\boldsymbol{I}_q$ is the $q \times q$ identity matrix. By standard properties of the Kronecker product $\boldsymbol{A}$ is nonsingular, and we have that

$$\boldsymbol{d}^*_{t+1} = \boldsymbol{h}_t \quad \Delta\boldsymbol{L}^*_{t+1} = (\boldsymbol{I}_q\boldsymbol{h}_t) \quad (\boldsymbol{B}\Delta\boldsymbol{L}_{t+1}) = (\boldsymbol{I}_q \quad \boldsymbol{B})(\boldsymbol{h}_t \quad \Delta\boldsymbol{L}_{t+1}) = \boldsymbol{A}\boldsymbol{d}_{t+1}. \tag{IA.A.13}$$

Since the null hypothesis implies that the asymptotic variance can be estimated consistently by the sample variance, it follows that

$$\hat{\boldsymbol{\Sigma}}^*_T \equiv \frac{1}{T}\sum_{t=1}^{T} \Bigg(\boldsymbol{d}^*_{t+1}\boldsymbol{d}^{*\prime}_{t+1} = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{A}\boldsymbol{d}_{t+1}\boldsymbol{d}'_{t+1}\boldsymbol{A}' = \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{A}'.$$

Due to the nonsingularity of $\boldsymbol{A}$ and $\hat{\boldsymbol{\Sigma}}_T$, it follows that

$$\begin{aligned} \bar{\boldsymbol{d}}^{*\prime}_{t+1}(\hat{\boldsymbol{\Sigma}}^*_T)^{-1}\bar{\boldsymbol{d}}^*_{t+1} &= \boldsymbol{d}'_{t+1}\boldsymbol{A}'(\boldsymbol{A}\hat{\boldsymbol{\Sigma}}_T\boldsymbol{A}')^{-1}\boldsymbol{A}\boldsymbol{d}_{t+1} \\ &= \boldsymbol{d}'_{t+1}\hat{\boldsymbol{\Sigma}}_T^{-1}\boldsymbol{d}_{t+1}, \end{aligned}$$

which shows that the test is invariant to a permutation of the ordering of the forecast losses. □

IA.A.2. Unconditional and multi-step predictive ability tests

In both the unconditional, $G_t = \{\emptyset, \quad\}$, and multistep conditional case the loss series are no longer martingale difference sequences under the null hypothesis. Thus, the sequence $\{\boldsymbol{h}_t \quad \Delta\boldsymbol{L}_{t+\tau}\}$ may be serially autocorrelated.[27] In the conditional setting, the null hypothesis imposes a particular structure on the serial correlation, namely that it can be at most order $\tau - 1$. However, in the unconditional case no such restriction exists. Consequently, we can no longer rely on the sample variance under the null for estimating

[27] Note that that in the unconditional case $\boldsymbol{h}_t = 1$ for all t.

the covariance matrix as was the case in the one-step conditional setting considered in the previous section. Instead, we consider a HAC-type estimator (see e.g. Newey and West (1987) and Andrews (1991)) with a bandwidth choice guided by the implications of the null hypothesis. The estimator is given by

$$\tilde{\boldsymbol{\Sigma}}_T = \frac{1}{T}\Big[\sum_{t=1}^{T} \boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau} + \sum_{j=1}^{b_T}\kappa(j,b_T)\sum_{t=1+j}^{T}\left(\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j} + \boldsymbol{d}_{t+\tau-j}\boldsymbol{d}'_{t+\tau}\right)\Big], \tag{IA.A.14}$$

where $\{b_T\}$ is an integer-valued truncation point sequence satisfying $b_T \to \infty$ as $T \to \infty$ and $b_T = o(T)$ (Newey and West, 1987) in the unconditional case, and $b_T = \tau - 1$ in the conditional case. Furthermore, $\kappa(\cdot,\cdot)$ is a real-valued kernel weight function satisfying the condition that $\kappa(j, b_T) \to 1$ as $T \to \infty$ for each $j = 1, \dots, b_T$ (Andrews, 1991), and $\kappa(j, b_T) = 0$ for $j > b_T$. For a review of data driven bandwidth selection methods see Clark and McCracken (2013).

Along the lines of the construction of the conditional test with $\tau = 1$, we construct the following Wald statistic which can be used in testing either unconditional or multi-step conditional equal predictive ability. The test statistic is given by

$$S_{h,\tau} = T\bar{\boldsymbol{d}}\tilde{\boldsymbol{\Sigma}}_T^{-1}\bar{\boldsymbol{d}}, \tag{IA.A.15}$$

where $\bar{\boldsymbol{d}} = T^{-1}\sum_{t=1}^{T}\boldsymbol{d}_{t+\tau}$. Before turning the properties of the proposed test statistic, we will need a slight modification of the assumptions from the previous section on one-step ahead forecasting.

Assumption 1*. *$\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ are $\phi-$mixing with $\phi(t) = O\left(t^{-r/(2r-2)-\iota}\right)$, $r \geq 2$, or $-$mixing with $(t) = O\left(t^{-\frac{r}{r-2}-\iota}\right)$, $r > 2$, for some $\iota > 0$.*

Assumption 2*. *$\mathbb{E}[|\boldsymbol{d}_{t+\tau,i}|^{r+\delta}] < \infty$ for some $\delta > 0$, $i = 1, \dots, qp$, and for all t, where subscript i indicate the i'th element of $\boldsymbol{d}_{t+1}$.*

Assumption 3*. $\boldsymbol{\Sigma}_T \equiv T^{-1}\sum_{t=1}^{T}\mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau}] + T^{-1}\sum_{j=1}^{b_T}\sum_{t=1+j}^{T}\left(\mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j}] + \mathbb{E}[\boldsymbol{d}_{t+\tau-j}\boldsymbol{d}'_{t+\tau}]\right)$ *is uniformly positive definite, where $b_T = \tau - 1$ in the conditional case* *and $b_T = T - 1$ in the unconditional case.*

Analogue to Theorem 1 $S_{h,\tau}$ is asymptotically chi-squared distributed with qp degrees of freedom under the null hypothesis, has power under the alternative hypothesis, and is permutation invariant. We summarize these results in Theorem 2 below.

Theorem 2 (Multistep multivariate predictive ability tests). *Suppose Assumptions 1*-3* hold.*

A. ***Asymptotic distribution under the null.*** *Suppose that either $\mathcal{G}_t = \{\emptyset, \quad\}$ and $\tau \geq 1$ or $\mathcal{F}_t \subseteq \mathcal{G}_t$ and $\tau > 1$. For any test function sequence $\{\boldsymbol{h}_t\}$, $m < \infty$, and under $\mathbb{H}_0$ in (8),*

$$S_{h,\tau} \xrightarrow{d} \chi^2(qp), \quad \textit{as } T \to \infty. \tag{IA.A.16}$$

B. ***Consistency under the alternative.*** *For any $c \in \mathbb{R}_+$ and under $\mathbb{H}_{A,h}$ in (12),*

$$\mathbb{P}[S_{h,\tau} > c] \to 1, \quad \textit{as } T \to \infty. \tag{IA.A.17}$$

C. ***Permutation invariance.*** *Let $\boldsymbol{L}^*_{t+\tau}$ be an arbitrary permutation of the forecast losses, and define $\Delta\boldsymbol{L}^*_{t+\tau} = \boldsymbol{D}\boldsymbol{L}^*_{t+\tau}$, $\bar{\boldsymbol{d}}^* = T^{-1}\sum_{t=1}^{T} \boldsymbol{d}^*_{t+\tau}$ with $\boldsymbol{d}^*_{t+\tau} = \boldsymbol{h}_t \quad \Delta\boldsymbol{L}^*_{t+\tau}$ and $\tilde{\boldsymbol{\Sigma}}^*_T$ be the associated covariance estimator defined in equation (IA.A.14). Then,*

$$S^*_{h,\tau} \equiv T\bar{\boldsymbol{d}}^{*'}_m(\tilde{\boldsymbol{\Sigma}}^*_T)^{-1}\bar{\boldsymbol{d}}^* = S_{h,\tau} \tag{IA.A.18}$$

for all T.

Consequently, a multivariate test for equal conditional multistep predictive ability or (multistep) unconditional predictive ability can be conducted by rejecting the null hypothesis whenever $S_{h,\tau} > z_{1-\alpha,qp}$. The permutation invariance result in Theorem 2 for the unconditional case is similar to Proposition 2 in Mariano and Preve (2012), but holds under the milder Assumptions 1*-3*, and hence also applies in a setting with non-stationary data, inclusion of nested models and explicit account of estimation uncertainty.

IA.A.2.1. Proof of Theorem 2 We start by proving part A. We proceed by a similar procedure as in the proof of Theorem 1, however with modifications due to the dependency in $\boldsymbol{d}_{t+\tau}$ under the null hypothesis. First, by Assumptions 2* and 3*, $\boldsymbol{\Sigma}_T$ is finite and uniformly positive definite. Let $\boldsymbol{\lambda} \in \mathbb{R}^{qp}$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ and consider $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} = T^{-1/2}\sum_{t=1}^{T}\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}$. Furthermore, identical arguments as in Theorem 1 imply that $\{\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}\}$ being mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$. Moreover, the asymptotic variance satisfies $\sigma_d^2 = \mathrm{Var}[\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}}] = \boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\Sigma}_T\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{\lambda} = 1$ for all T sufficiently large. By Minkowski's inequality and computations as in (IA.A.7), $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\boldsymbol{d}_{t+\tau}$ has absolute $2+\delta$ moment for some $\delta > 0$. Then, by Corollary 3.1 in Wooldridge and White (1988) we deduce that $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0,1)$. Hence, by the Cramér-Wold device it follows that $\boldsymbol{\Sigma}_T^{-1/2}\sqrt{T}\bar{\boldsymbol{d}} \xrightarrow{d} N(0, \boldsymbol{I}_{qp})$.

It remains to be shown that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$. Consider

$$\begin{aligned}\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T = &\frac{1}{T}\sum_{t=1}^{T}\Big(\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau} - \mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau}]\Big)\\ &+\frac{1}{T}\sum_{j=1}^{b_T}\Bigg(\kappa(j,b_T)\sum_{t=1+j}^{T}\Bigg(\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j} - \mathbb{E}[\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j}]\\ &+\boldsymbol{d}_{t+\tau-j}\boldsymbol{d}'_{t+\tau} - \mathbb{E}[\boldsymbol{d}_{t+\tau-j}\boldsymbol{d}'_{t+\tau}]\Big).\Bigg(\end{aligned} \tag{IA.A.19}$$

By Theorem 3.49 in White (2001), $\{\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j}\}$ is mixing of the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ for each $j = 0, \ldots, b_T$. Moreover, each of its elements are bounded uniformly in t by Assumption 2*. Hence, since $\kappa(j, b_T) \to 1$ as $T \to \infty$ and $\kappa(0, b_T) = 1$ it follows via Corollary 3.48 in White (2001) that

$$\frac{1}{T}\kappa(j,\tau)\sum_{t=1+j}^{T}\Bigg(\Big(\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-j} - \mathbb{E}[\boldsymbol{d}_{m,t+\tau}\boldsymbol{d}'_{t+\tau-j}]\Big) \xrightarrow{\mathbb{P}} 0,$$

for each $j = 0, \ldots, b_T$. Combined with equation IA.A.19, this implies that $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$ (see also Andrews (1991)). Hence, we can deduce via similar steps as in (IA.A.9) that $S_{h,\tau} \xrightarrow{d} \chi^2(qp)$ as $T \to \infty$. □

We now prove part B. The result follows by arguments similar to those in the proof of

Theorem 1. Hence, $\{\boldsymbol{d}_{t+\tau}\}$ is mixing with the same size as $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{w}_t\}$ and each element in $\boldsymbol{d}_{t+\tau}$ is bounded uniformly in t by Assumption 2*. Then it follows by Corollary 3.48 in White (2001) that $\bar{\boldsymbol{d}} - \mathbb{E}[\bar{\boldsymbol{d}}] \xrightarrow{\mathbb{P}} 0$, and consequently similar computations as in (IA.A.11) applies. By arguments identical to those in the proof of part A., $\tilde{\boldsymbol{\Sigma}}_T - \boldsymbol{\Sigma}_T \xrightarrow{\mathbb{P}} 0$, where $\boldsymbol{\Sigma}_T$ is positive definite by Assumption 3*. Theorem 8.13 in White (1994) then implies that under $\mathbb{H}_{A,h}$ in (12) and for any constant $c \in \mathbb{R}_+$, $\mathbb{P}[S_{h,\tau} > c] \to 1$ as $T \to \infty$. □

Lastly, we prove part C. Due the arguments in the proof of Theorem 1 it suffices to show that $\tilde{\boldsymbol{\Sigma}}_T * = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, where $\boldsymbol{A} = \boldsymbol{I}_q \quad \boldsymbol{B}$. Thus, let

$$\tilde{\boldsymbol{\Sigma}}_T(b) \equiv \frac{1}{T}\sum_{t=1+b}^{T}\Big(\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-b},$$

for $b = 0, 1, 2 \ldots$. It then follows that

$$\tilde{\boldsymbol{\Sigma}}_T(b)^* \equiv \frac{1}{T}\sum_{t=1+b}^{T}\Big(\boldsymbol{d}^*_{t+\tau}\boldsymbol{d}^{*\prime}_{t+\tau-b} = \frac{1}{T}\sum_{t=1+b}^{T}\boldsymbol{A}\boldsymbol{d}_{t+\tau}\boldsymbol{d}'_{t+\tau-b}\boldsymbol{A}' = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T(b)\boldsymbol{A}'.$$

Consequently, it follows that $\tilde{\boldsymbol{\Sigma}}^*_T = \boldsymbol{A}\tilde{\boldsymbol{\Sigma}}_T\boldsymbol{A}'$, which completes the proof. □

IA.B. Bond data

We use the Gürkaynak et al. (2007) dataset from 1962:M1 to 2018:M12. The time t log yield on a k-period bond is computed using the methods developed in Nelson and Siegel (1987) and Svensson (1994) as

$$\begin{aligned} y_t^{(k)} = {}_{0,t} + {}_{1,t}\frac{1-\exp\left(-\frac{n}{\kappa_{1,t}}\right)}{\frac{n}{\kappa_{1,t}}} &+ {}_{2,t}\left[\frac{1-\exp\left(-\frac{n}{\kappa_{1,t}}\right)}{\frac{n}{\kappa_{1,t}}}\left(-\exp\quad -\frac{n}{\kappa_{1,t}}\right)\right]\left(\right. \\ + {}_{3,t}\left[\frac{1-\exp\left(-\frac{n}{\kappa_{2,t}}\right)}{\frac{n}{\kappa_{2,t}}}\left(-\exp\quad -\frac{n}{\kappa_{2,t}}\right)\right]\left(\right. \end{aligned} \tag{IA.B.20}$$

where we use parentheses in the superscript to distinguish maturity from exponentiation and $n = \frac{k}{m}$ and m denotes, respectively, the bond maturity in years and the number of periods per year.

Let $p_t^{(k)} = -\left(\frac{k}{m}\right) y_t^{(k)}$ be the log price of a k-period bond at time t. The log forward rate at time t for loans between $t+k-1$ and $t+k$ is defined as

$$f_t^{(k)} = p_t^{(k-1)} - p_t^{(k)} = -\tfrac{k-1}{m} y_t^{(k-1)} + \tfrac{k}{m} y_t^{(k)}. \tag{IA.B.21}$$

The excess return to purchasing a k-period bond today and selling it as a $k-1$ period bond after one month is

$$rx_{t+1}^{(k)} = p_{t+1}^{(k-1)} - p_t^{(k)} - p_t^{(1)} = -\tfrac{k-1}{m} y_{t+1}^{(k-1)} + \tfrac{k}{m} y_t^{(k)} - \tfrac{1}{m} y_t^{(1)}, \tag{IA.B.22}$$

where $y_t^{(1)}$ denotes the risk-free one-period rate that we proxy using the implied yield on a one-month Treasury bill obtained from the Center for Research in Security Prices (CRSP) as in Gargano et al. (2019).[28]

[28] For $k = 1$, we have that $f_t^{(1)} = y_t^{(1)}$ and that $y_t^{(k-1)} = y_t^{(0)} = 0$ due to $p_t^{(0)}$ being zero (log of terminal payoff of one is zero).

IA.C. Additional empirical results

IA.C.1. Descriptive statistics for state variables

Table IA.1 presents full sample descriptive statistics for our two state variables that captures economic activity and uncertainty, respectively: the Purchasing Managers' Index (PMI) and the macroeconomic uncertainty index of Jurado et al. (2015).

[Insert Table IA.1 About Here]

The series are both highly persistent with autocorrelation coefficients well above 0.9. Most importantly, we note that the series obtains a negative contemporaneous correlation of -0.48 in the data, suggesting that they capture part of the some features, but are not perfect substitutes.

IA.C.2. In-sample predictive regressions

Table IA.2 presents full sample least squares estimation results to facilitate comparison with the extant literature. Specifically, we estimate predictive regressions of the form presented in (1) with the risk premium on a Treasury bond with k-periods to maturity $rx_{t+1}^{(k)}$ as the dependent variable. We focus on bonds with $k = \{24, 36, 48, 60\}$ months to maturity and consider models based on the predictor variables outlined in Section 2.2. We stress that these results are not available to a real-time investor, but they are useful for gauging the mechanisms of the predictive models.

[Insert Table IA.2 About Here]

The slope coefficients for CS and FB are all positive and increasing with maturity and are all statistically significant at conventional levels.[29] We note that these positive slope coefficients imply negative slopes for the companion regression of yield or forward spreads on future yield changes as documented in Campbell and Shiller (1991). Thus, both yield

[29] Bauer and Hamilton (2018) show that statistical test of predictive regression in full sample analyses are subject to serious small sample distortions when using 12-month overlapping returns. However, we use one-month non-overlapping returns here and are therefore not affected by their results. See also the discussion in Gargano et al. (2019).

and forward spreads contain useful information about future bond excess returns over the full range of available observations. Turning to the principal components, we find that PC1 has a constant slope coefficient across maturities, PC2 increased monotonically, and PC3 displays an inverse U-shape. P1 and PC3 are mostly insignificant, whereas PC2 is significant for the longer maturities. This mirrors the results for CS, but shows that maturity-specific spreads are more informative than the common slope factor. Last, CP and LN both display monotonically increasing slope coefficients that are highly significant. Of all the models, LN appears to explain the largest fraction of bond risk premia, closely followed by CP and yield spreads. Overall, in-sample results points to predictive relation between all our candidate predictors.

IA.C.3. Links to uncertainty

Table IA.4 presents contemporaneous correlations among $\mathcal{U}$ and the risk premia estimates from the set of individual models, EW, and the dynamic forecast combinations generated by PMI, $\mathcal{U}$, and NONE.

[Insert Table IA.4 About Here]

We find that most forecasts are positively correlated with uncertainty, implying that investors higher risk premia in periods with heightened uncertainty. The exception is CS and FB for the shorter maturities, where we observe negative correlations. As for our main results concerning the relation to economic activity (see Table 8), we find that LN displays the highest correlation with $\mathcal{U}$ among the individual predictors and EW. Turning to the dynamic forecast combination estimates in Panel B, we find that both PMI and $\mathcal{U}$ trimming delivers forecasts that are tightly linked to uncertainty. That is, not only do they produce countercyclical risk premia estimates, they only procedure forecasts closely linked to uncertainty.

IA.C.4. Additional results for economic value

Figure IA.1 plots the cumulative CER gains for the individual predictor variables along with the equal-weighted forecast (EW). The results largely mirrors those in Table 9 in the main paper and illustrate that most individual predictors fail to deliver economic value on a consistent basis. The exception being LN.

[Insert Figure IA.1 About Here]

[Insert Table IA.5 About Here]

Table IA.5 reconstructs the results from Table 9 in the main paper using instead a coefficient of relative risk aversion of $= 5$ to verify that our results are robust to other, and lower, values of risk aversion. The table clearly demonstrates that this is the case.

Table IA.1: Conditioning variables

This table presents descriptive statistics for the state variables used in the empirical analysis. PMI is the Purchasing Managers' Index published by the Institute for Supply Managers and $\mathcal{U}$ is the macroeconomic uncertainty index developed in Jurado et al. (2015). The table reports mean, standard deviation, skewness, kurtosis, and first-order autocorrelation (AC(1)) of each state variable. We also report the contemporaneous correlation between the variables. The sample period is January 1962 to December 2018.

	PMI	$\mathcal{U}$
Mean	52.61	0.66
Std. dev.	6.37	0.09
Skewness	-0.61	1.63
Kurtosis	4.37	5.79
AR(1)	0.94	0.99
Correlation	-0.48	

Table IA.2: In-sample regressions

This table reports full sample least squares estimates of the slope coefficients for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor computed from a projection of average excess bond returns on two-, three-, four-, and five-year forward rates, and the Ludvigson and Ng (2009) macroeconomic factor computed as a projection of average excess bond returns on factors obtained from a large panel of macroeconomic variables. For each model, we report slope coefficients, Newey and West (1987) *t*-statistics using a bandwidth of twelve lags in parenthesis, and the adjusted R^2 in square brackets. The sample period is January 1962 to December 2018.

	2-year	3-year	4-year	5-year
	Panel A: Campbell-Shiller			
CS	2.02	2.36	2.75	3.15
	(2.67)	(2.64)	(2.85)	(3.17)
	[2.55]	[2.32]	[2.42]	[2.61]
	Panel B: Fama-Bliss			
FB	1.20	1.41	1.69	1.99
	(2.20)	(2.30)	(2.79)	(3.38)
	[1.80]	[1.68]	[1.90]	[2.14]
	Panel C: Principal components			
PC1	0.01	0.01	0.01	0.01
	(1.43)	(1.04)	(0.76)	(0.56)
PC2	0.13	0.21	0.29	0.37
	(1.72)	(2.10)	(2.46)	(2.77)
PC3	0.23	0.31	0.24	0.09
	(0.66)	(0.63)	(0.39)	(0.13)
	[1.05]	[1.09]	[1.19]	[1.30]
	Panel D: Cochrane-Piazzesi			
CP	0.65	0.88	1.11	1.36
	(4.60)	(4.30)	(4.12)	(4.08)
	[2.37]	[2.16]	[2.17]	[2.30]
	Panel E: Ludvigson-Ng			
LN	0.65	0.90	1.12	1.33
	(3.68)	(3.96)	(4.25)	(4.46)
	[6.62]	[6.47]	[6.33]	[6.15]

Table IA.3: Correlations between forecasts and macroeconomic uncertainty
This table reports correlation coefficients between out-of-sample generated forecasts from individual bond predictors (Panel A) and the dynamic forecast strategy (Panel B) and economic uncertainty as measured by the the macroeconomic uncertainty index ($\mathcal{U}$) from Jurado et al. (2015). We report p-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
	Panel A: Individual bond predictors			
CS	-0.09	-0.04	0.01	0.05
	(0.10)	(0.41)	(0.92)	(0.39)
FB	-0.04	0.07	0.12	0.15
	(0.49)	(0.21)	(0.02)	(0.00)
PC	0.03	0.04	0.05	0.06
	(0.57)	(0.5)	(0.34)	(0.23)
CP	0.12	0.11	0.10	0.10
	(0.02)	(0.04)	(0.06)	(0.07)
LN	0.44	0.46	0.47	0.48
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.43	0.38	0.34	0.32
	(0.00)	(0.00)	(0.00)	(0.00)
EW	0.31	0.34	0.35	0.35
	(0.00)	(0.00)	(0.00)	(0.00)
	Panel B: Dynamic forecast combination			
PMI	0.54	0.53	0.50	0.50
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathcal{U}$	0.59	0.56	0.54	0.55
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	0.54	0.47	0.46	0.47
	(0.00)	(0.00)	(0.00)	(0.00)

Table IA.4: Alternative proxies for economic activity

This table reports correlation coefficients between forecasts and alternative proxies for economic activity. We use the Chicago Fed National Activity Index (Panel A), recession probabilities from Chauvet and Piger (2008) (Panel B), and log growth rates to industrial production (Panel C). We report p-values for the null of no correlation in parenthesis. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year bond	3-year bond	4-year bond	5-year bond
Panel A: Chicago Fed National Activity Index (CFNAI)				
CS	0.10	0.04	-0.01	-0.05
	(0.07)	(0.40)	(0.89)	(0.35)
FB	0.04	-0.06	-0.13	-0.16
	(0.46)	(0.25)	(0.02)	(0.00)
PC	0.19	0.18	0.15	0.12
	(0.00)	(0.00)	(0.01)	(0.03)
CP	-0.10	-0.09	-0.08	-0.07
	(0.05)	(0.10)	(0.15)	(0.17)
LN	-0.48	-0.49	-0.50	-0.51
	(0.00)	(0.00)	(0.00)	(0.00)
EH	-0.20	-0.17	-0.16	-0.15
	(0.00)	(0.00)	(0.00)	(0.01)
EW	-0.26	-0.29	-0.30	-0.32
	(0.00)	(0.00)	(0.00)	(0.00)
PMI	-0.51	-0.55	-0.54	-0.51
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathfrak{U}$	-0.56	-0.54	-0.54	-0.57
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.53	-0.49	-0.49	-0.50
	(0.00)	(0.00)	(0.00)	(0.00)
Panel B: Recession probabilities (Chauvet and Piger, 2008)				
CS	-0.01	0.02	0.05	0.08
	(0.89)	(0.72)	(0.33)	(0.13)
FB	0.03	0.09	0.14	0.16
	(0.64)	(0.09)	(0.01)	(0.00)
PC	-0.05	-0.05	-0.03	-0.01
	(0.37)	(0.35)	(0.57)	(0.86)
CP	0.10	0.08	0.06	0.05
	(0.08)	(0.16)	(0.26)	(0.32)
LN	0.56	0.57	0.58	0.59
	(0.00)	(0.00)	(0.00)	(0.00)
EH	0.18	0.13	0.11	0.09
	(0.00)	(0.01)	(0.05)	(0.09)
EW	0.37	0.38	0.38	0.38
	(0.00)	(0.00)	(0.00)	(0.00)
PMI	0.51	0.54	0.53	0.53
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathfrak{U}$	0.55	0.56	0.54	0.55
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	0.58	0.53	0.51	0.52
	(0.00)	(0.00)	(0.00)	(0.00)
Panel C: Log industrial production growth				
CS	0.07	0.06	0.03	0.01
	(0.16)	(0.28)	(0.55)	(0.86)
FB	0.07	0.01	-0.03	-0.05
	(0.16)	(0.79)	(0.61)	(0.33)
PC	0.16	0.15	0.14	0.13
	(0.00)	(0.00)	(0.01)	(0.02)
CP	-0.08	-0.07	-0.07	-0.07
	(0.16)	(0.17)	(0.18)	(0.18)
LN	-0.26	-0.27	-0.28	-0.28
	(0.00)	(0.00)	(0.00)	(0.00)
EH	-0.09	-0.10	-0.10	-0.10
	(0.11)	(0.08)	(0.07)	(0.06)
EW	-0.12	-0.14	-0.15	-0.16
	(0.03)	(0.01)	(0.01)	(0.00)
PMI	-0.23	-0.25	-0.27	-0.23
	(0.00)	(0.00)	(0.00)	(0.00)
$\mathfrak{U}$	-0.28	-0.25	-0.25	-0.27
	(0.00)	(0.00)	(0.00)	(0.00)
NONE	-0.25	-0.22	-0.21	-0.21
	(0.00)	(0.00)	(0.01)	(0.01)

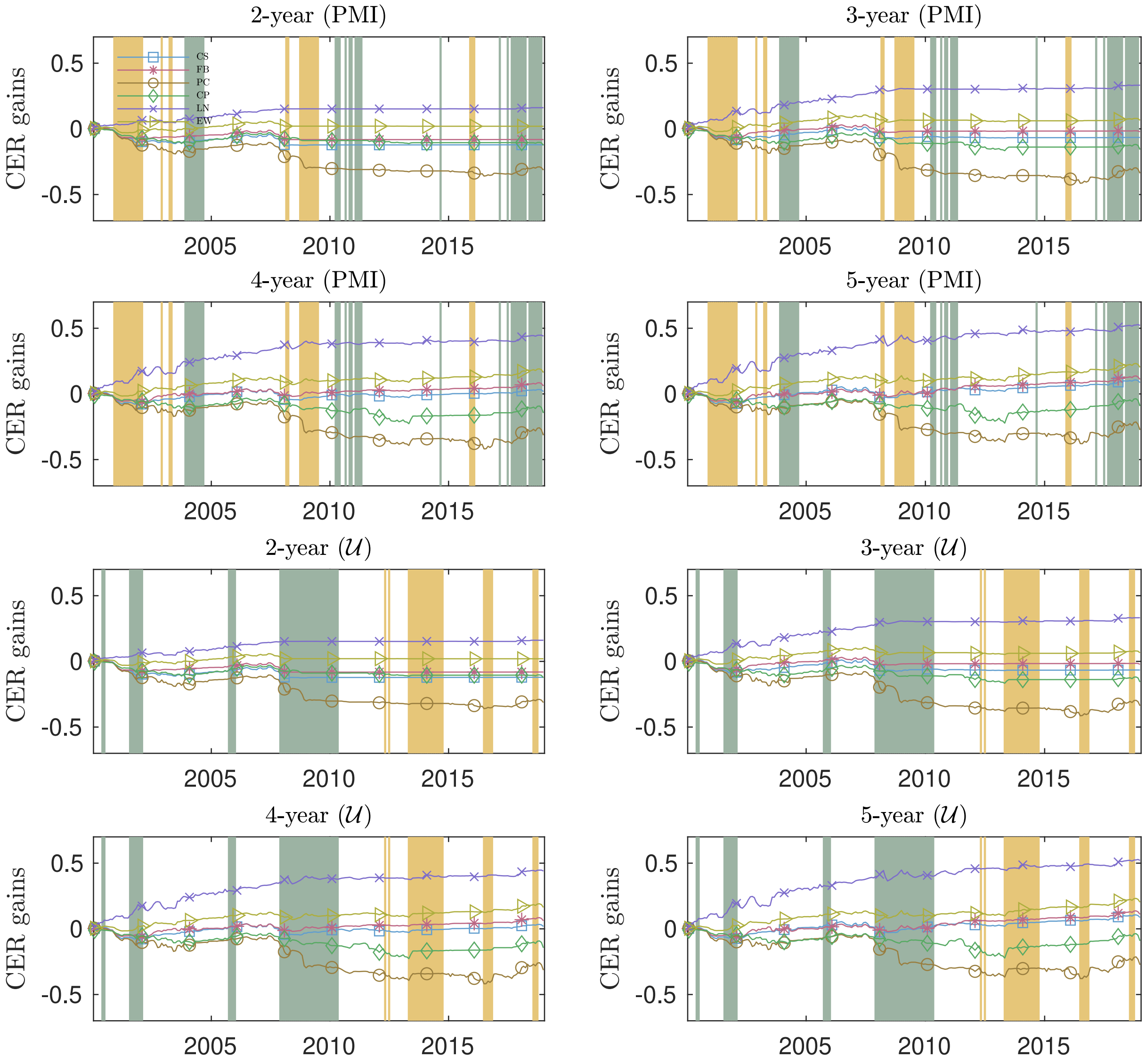
Table IA.5: Economic Value: $= 5$

This table reports certainty equivalent return (CER) gains for various linear predictive models for bond excess return. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. For each model, we report the CER gains relative to the expectations hypothesis (Panels A and B) and a static forecast combination strategy (Panel C). PMI denotes the Purchasing Managers Index published by the Institute for Supply Management and $\mathfrak{U}$ is the macroeconomic uncertainty index from Jurado et al. (2015). CER gains are based on an investor with mean-variance preferences and a relative risk aversion of $= 5$. The out-of-sample evaluation period runs from January 2000 to December 2018.

	2-year	3-year	4-year	5-year
	Panel A: Individual bond predictors against EH			
CS	-0.91	-0.88	-0.52	-0.25
	(0.94)	(0.87)	(0.74)	(0.62)
FB	-0.62	-0.68	-0.55	-0.34
	(0.88)	(0.86)	(0.80)	(0.67)
PC	-2.06	-2.46	-2.41	-2.36
	(0.99)	(0.96)	(0.93)	(0.9)
CP	-0.80	-1.20	-1.31	-1.36
	(0.96)	(0.94)	(0.91)	(0.87)
LN	0.61	1.39	2.41	3.24
	(0.01)	(0.01)	(0.00)	(0.00)
EW	0.03	0.25	0.70	1.08
	(0.46)	(0.32)	(0.13)	(0.07)
	Panel B: Dynamic forecast combination against EH			
PMI	0.28	0.59	1.07	1.47
	(0.19)	(0.14)	(0.05)	(0.02)
$\mathfrak{U}$	0.19	0.53	1.22	1.60
	(0.27)	(0.14)	(0.02)	(0.01)
NONE	0.12	0.30	0.76	1.07
	(0.34)	(0.28)	(0.10)	(0.07)
	Panel C: Dynamic forecast combination against EW			
PMI	0.25	0.34	0.37	0.39
	(0.02)	(0.04)	(0.05)	(0.03)
$\mathfrak{U}$	0.16	0.28	0.52	0.52
	(0.08)	(0.02)	(0.00)	(0.02)
NONE	0.09	0.04	0.06	-0.01
	(0.15)	(0.35)	(0.28)	(0.52)

Figure IA.1: Relative certainty equivalent returns

This figure plots the recursively updated cumulative difference in realized utility from the EH benchmark model and the ith predictor model over the out-of-sample evaluation period. We consider five different predictors: yield spreads (Campbell and Shiller, 1991), forward spreads (Fama and Bliss, 1987), principal components of yields (Litterman and Scheinkman, 1991), the Cochrane and Piazzesi (2005) forward rate factor, and the Ludvigson and Ng (2009) macroeconomic factor. We also consider a simple equal-weighted combination of the individual forecasts. A positive (negative) slope indicates that the predictive model delivers more (less) accurate forecasts than the EH benchmark. Green (yellow) shaded ares represent periods of high (low) activity and uncertainty, respectively, where activity is measured using the Purchasing Managers' Index (PMI) and uncertainty in the index developed by Jurado et al. (2015). High (low) episodes are identified using the 80% (20%) quantiles of their time series. White areas are normal times. The out-of-sample evaluation periods runs from January 2000 to December 2018.


2-year (PMI)
3-year (PMI)
4-year (PMI)
5-year (PMI)
2-year ($\mathcal{U}$)
3-year ($\mathcal{U}$)
4-year ($\mathcal{U}$)
5-year ($\mathcal{U}$)
CER gains
CS
FB
PC
CP
LN
EW