DISCUSSION PAPER SERIES

DP16357
(v. 2)

Equity premium predictability over the business cycle

Emanuel Moench and Tobias Stein

FINANCIAL ECONOMICS

ISSN 0265-8003

Equity premium predictability over the business cycle

Emanuel Moench and Tobias Stein

Discussion Paper DP16357
First Published 13 July 2021
This Revision 10 September 2021

Centre for Economic Policy Research
33 Great Sutton Street, London EC1V 0DX, UK
Tel: +44 (0)20 7183 8801
www.cepr.org

Equity premium predictability over the business cycle

Abstract

The equity premium follows a pronounced v-shape pattern around the beginning of recessions. It sharply drops into negative territory just before business cycle peaks and then strongly recovers as the recession unfolds. Recessions are preceded by an inverted yield curve. Thus probit models using the term spread as predictor time the beginning of recessions well. We show that such model-implied recession probabilities strongly improve equity premium prediction out-of-sample. We document a structural break in the mean of the term spread in 1982. When correcting for this break, the forecast performance further strengthens, outperforming other recently proposed benchmark predictors.

JEL Classification: E32, E37, C53, G11, G17

Keywords: Recession predictability, return predictability, Business cycle, Probit Model, term spread

Emanuel Moench - emanuel.moench@bundesbank.de
Deutsche Bundesbank, Goethe University Frankfurt and CEPR

Tobias Stein - tobi.stein2007@web.de
Deutsche Bundesbank

Acknowledgements
We thank Arturo Estrella, Roberto Gomez-Cram, Amit Goyal, Mete Kilic, Stijn Van Nieuwerburgh, and an anonymous referee for helpful comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect those of the Deutsche Bundesbank or the Eurosystem.

Equity premium predictability over the business cycle[*]

Emanuel Moench[†] Tobias Stein[‡]

September 7, 2021

Abstract

The equity premium follows a pronounced v-shape pattern around the beginning of recessions. It sharply drops into negative territory just before business cycle peaks and then strongly recovers as the recession unfolds. Recessions are preceded by an inverted yield curve. Thus probit models using the term spread as predictor time the beginning of recessions well. We show that such model-implied recession probabilities strongly improve equity premium prediction out-of-sample. We document a structural break in the mean of the term spread in 1982. When correcting for this break, the forecast performance further strengthens, outperforming other recently proposed benchmark predictors.

JEL classification: E32, E37, C53, G11, G17

Keywords: Recession predictability, return predictability, business cycle, probit model, term spread

[*]We thank Arturo Estrella, Roberto Gómez-Cram, Amit Goyal, Mete Kilic, Stijn Van Nieuwerburgh, and an anonymous referee for helpful comments and suggestions. The views expressed in this paper are those of the authors and do not necessarily reflect those of Deutsche Bundesbank or the Eurosystem.

[†]Deutsche Bundesbank, Goethe University Frankfurt, CEPR (Email: emanuel.moench@bundesbank.de)

[‡]Deutsche Bundesbank, Goethe University Frankfurt (Email: tobias.stein@bundesbank.de)

1 Introduction

Whether the equity premium is predictable has been the subject of long debate in finance. A large literature documents in-sample predictability using a host of financial and economic variables such as valuation ratios, the default spread or the consumption-wealth ratio as predictors (see, e.g. Fama and French (1988), Campbell and Shiller (1988), Lettau and Ludvigson (2001)). However, in an influential paper Welch and Goyal (2008) show that none of the proposed predictors would have consistently outperformed a simple historical average return out-of-sample. Since then, a growing literature proposes alternative predictors and forecasting methods which appear to provide superior statistical predictability relative to the historical average benchmark, see Rapach and Zhou (2013) for an overview.[1] A common finding is that predictability primarily arises around recessions.

This is consistent with a related literature suggesting that expected equity returns vary over the business cycle (e.g. Fama and French (1989), Ferson and Harvey (1991), Cochrane (2007), Campbell and Diebold (2009)). Lustig and Verdelhan (2012) document significant variation in realized excess returns around recessions, showing that they are negative at the business cycle peak and then sharply rise over following quarters. This is confirmed by Figure 1 which depicts the forward-looking arithmetic mean and median of the U.S. log equity premium over different time windows around the ten NBER recessions for the period from March 1951 to December 2019. The equity premium is mostly negative but relatively volatile for the one-month and three-month window around the beginning of recessions. However, a clear v-shape emerges for the cumulative six- and twelve-month ahead horizons, highlighting that equity returns are sharply negative around business cycle peaks but strongly recover thereafter.

Table 1 presents moments of the annualized log equity premium over the business cycle for the same 70-year period. The total annual equity premium was 6.3% with a standard deviation of 14.3%, implying a Sharpe ratio of 0.4. Focusing only on NBER expansions, the equity premium amounted to 8.4% with an annualized Sharpe ratio of 0.6. In recessions, it was negative at -5.9%. Zooming in around business cycle peaks, we see that the equity premium tended

[1] Among others, recently proposed predictors include the output gap (Cooper and Priestley, 2009), short interest (Rapach et al., 2016), industrial electricity usage (Da et al., 2017), gold-to-platinum ratio (Huang and Kilic, 2019), variance risk premium (Pyun, 2019), and investor attention (Chen et al., 2020). Methodological contributions include non-negativity constraints (Campbell and Thompson, 2008), combination forecasts (Rapach et al., 2010), time-varying coefficient models (Dangl and Halling, 2012), principal component analysis (Neely et al., 2014), economic constraints (Pettenuzzo et al., 2014), and machine learning techniques (Gu et al., 2020).

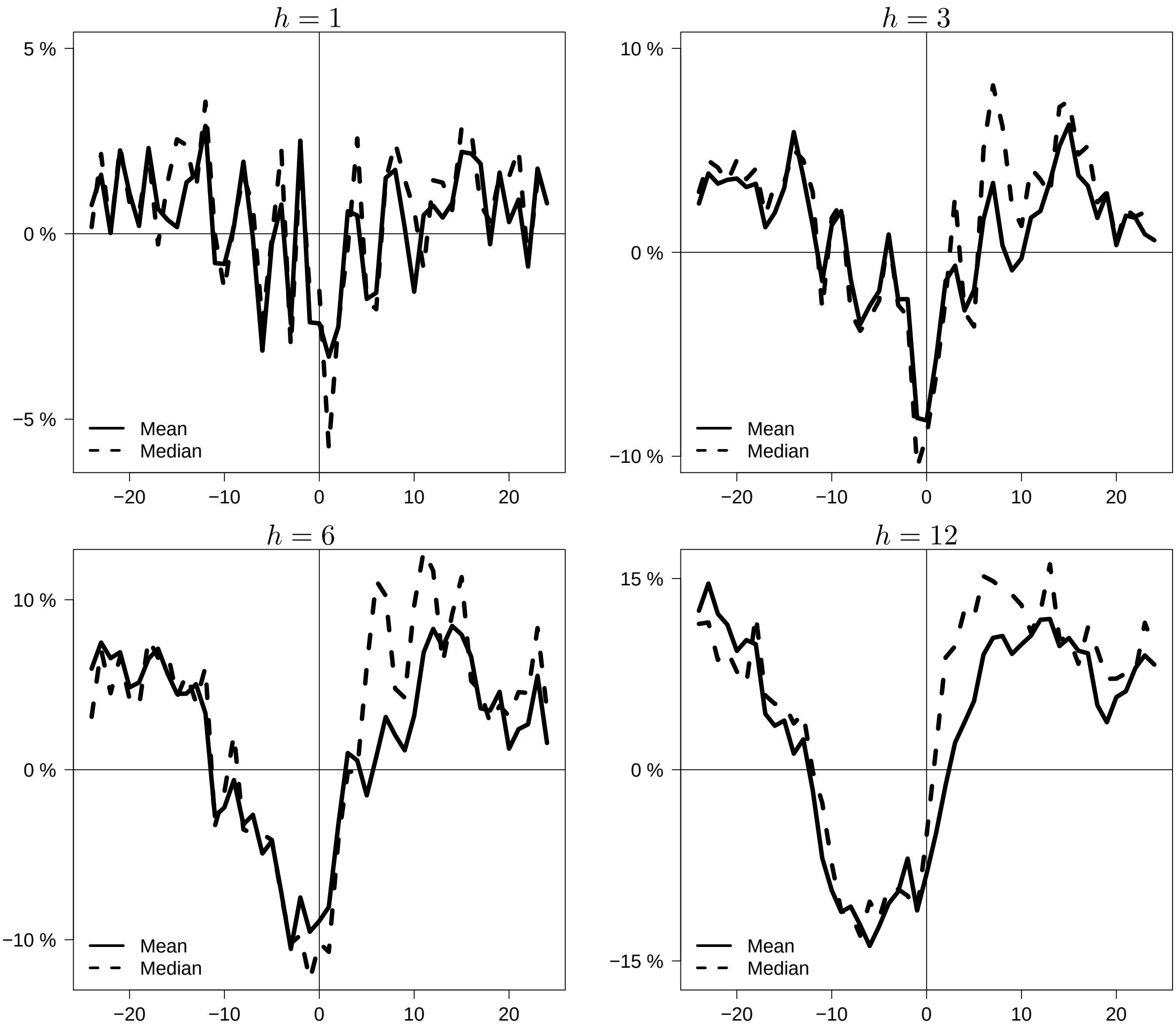

Figure 1

Log equity premium around business cycle peaks

This figure presents the arithmetic average and median of the (cumulative) log equity premium around the 10 recessions in the sample from 1951:3 to 2019:12. The equity premium is the difference between value-weighted returns on the S&P 500 index (including dividends) and the Treasury bill rate. The vertical axis depicts $\sum_{j=0}^{h-1} r_{t+j}$ for $t = -24, \ldots, -1, 0, 1, \ldots, 24$, whereby r_{t+j} is the log equity premium in month $t + j$. The horizontal axis displays the 24 months before and after a business cycle peak - with $t = 0$ referring to the first month of a NBER-dated recession. Results are shown for $h = 1, 3, 6, 12$.

to be strongly negative during the six months before and after the business cycle peak, with annualized values of almost -10% and -18%, respectively. Hence, the stock market on average incurs large losses in the one-year window around the beginning of recessions. While it tends to recover in the subsequent months, on average it only gains an annualized 1.5% six to eleven months after the peak. The last two rows in Table 1 show the annualized log equity premium for samples that exclude the 12 months and the 24 months around the beginning of recessions, respectively. When excluding two years (one year) of observations around each peak, the average equity premium and Sharpe ratio rise to 11.1% and 0.8 (9.7% and 0.7), compared to 6.3% and 0.4 for the full sample. This evidence strongly suggests that, to the extent that one can

Table 1
Log equity premium over the business cycle

This table reports the annualized mean, median, standard deviation, and Sharpe ratio of the monthly U.S. log equity premium. The equity premium is the difference between value-weighted returns on the S&P 500 index (including dividends) and the Treasury bill rate. $\rho(1)$ (N) denotes the first order serial correlation (number of observations). The descriptive statistics are presented for the full sample from March 1951 to December 2019, as well as separately for recessions and expansions. Further statistics are provided for sub-samples before and after business cycle peaks. Peak refers to the peak month of NBER-dated business cycle contractions (first month of a recession). The total number of recessions in the sample is 10.

	1951:3 to 2019:12							
log equity premium	Mean	Median	Std. dev.	Sharpe ratio	Skewness	Kurtosis	$\rho(1)$	N
Full sample	6.31	10.9	14.38	0.44	-0.67	5.41	0.04	826
Recessions	-5.90	-3.75	19.83	-0.30	-0.26	3.01	0.18	121
Expansions	8.40	11.25	13.14	0.64	-0.70	6.20	-0.06	705
Before the peak:								
peak−12 to peak−7	6.64	7.99	12.25	0.54	0.39	2.61	-0.10	60
peak−6 to peak−1	-9.85	-7.84	13.11	-0.75	-0.27	2.80	-0.14	60
At/after the peak:								
peak to peak+5	-17.80	-12.36	16.01	-1.11	-0.29	2.24	0.03	60
peak+6 to peak+11	1.50	15.20	19.89	0.08	-0.65	3.95	0.32	60
Excl. peak−6 to peak+5	9.73	12.31	14.11	0.69	-0.74	6.21	0.02	706
Excl. peak−12 to peak+11	11.07	12.95	13.60	0.81	-0.75	6.53	-0.02	592

predict the beginning of recessions, one should be able to time the market.

It is well documented that the term spread is a robust predictor of recessions for horizons of one year ahead and longer (see, e.g., Estrella and Mishkin (1998)). In the post-war period, every single U.S. recession was preceded by an inverted yield curve. In this paper, we show that recession probabilities derived from the term spread are strong predictors of the equity premium. The forecasts from probit models in the spirit of Estrella and Hardouvelis (1991) significantly forecast the cumulative equity premium over horizons from 1 to 12 months out-of-sample. Liu and Moench (2016) document that the additional incorporation of lagged observations of the term spread further improves short-horizon recession forecasts. We build on this finding and document that a backward-looking three-year moving average of the term spread substantially strengthens the recession classification ability of the term spread by reducing false positives and better timing the beginning of recessions. The improved recession predictability goes hand in hand with strong improvements in equity premium forecasts: especially at short horizons, equity premium predictability is substantially enhanced by including a backward-looking moving average of the term spread into the probit models. Our results highlight the close link between recession expectations and equity market returns. Specifically, we show that the v-shaped pattern of excess returns around business cycle peaks is well captured by real-time recession probability forecasts based on information in the yield curve. Our findings are thus in line with Rapach et al. (2010) and Dangl and Halling (2012) who find that the predictive power of combination forecasts and time-varying coefficient models primarily arises from business-cycle variation in the equity premium. Importantly, however, equity premium forecasts based on recession probabilities outperform the historical average benchmark also in expansions. The reason is that by correctly anticipating low equity market returns heading into recessions, they also correctly predict higher returns in business cycle booms.

Several authors have argued that the probit model for forecasting recessions suffers from a structural break (see, e.g., Chauvet and Potter (2002, 2010)). We indeed document strong evidence for a structural break in the mean of the term spread in 1982 and show that it would have been possible for investors to identify this break in real-time a few years after it occurred. We follow Lettau and Van Nieuwerburgh (2008) and Pesaran and Timmermann (2007) and apply four different methods to correct for the break in the term spread. All further improve the out-of-sample R^2 for forecasting the equity premium. This improvement is partly due to the fact that the real-time break-corrected recession probabilities better predict the beginning of the 2001 and 2008-2009 recessions.

In terms of predictive ability our approach outperforms other recently proposed predictors including "short interest" of Rapach et al. (2016) and the "gold-to-platinum" ratio of Huang and Kilic (2019). The out-of-sample R^2 is as high as 3% for monthly forecasts and often higher than 10% for cumulative one-year ahead forecasts. Moreover, we perform an asset allocation exercise for a mean-variance investor who invests in the equity market and the risk-free rate. This exercise reveals an excellent market timing ability of recession probability forecasts, which is even more pronounced for the break-correction methods. The models signal to run down equity exposure before the onset of recessions when the yield curve flattens and to re-enter the market toward the end of recessions when the yield curve steepens. An investor who forecasts with (break-corrected) recession probabilities increases the Sharpe ratio to around 0.85 compared to 0.50 for the buy-and-hold investor. Using a VAR-based decomposition in the spirit of Campbell (1991), we find that the predictability is driven by both higher anticipated discount rates and lower expected future dividends, consistent with countercyclical risk premia. We also show that our results are robust to taking into account transaction costs and that model-implied recession probabilities predict the equity premium in other countries.

In a related recent paper, Gómez-Cram (2021) studies one-month ahead equity premium predictability over the business cycle. Consistent with his results, we find that stock returns are negative at the beginning of recessions and that business cycle variables help to time these periods. Despite this broad similarity, there are a number of important differences between our and his paper. First, while Gómez-Cram (2021) studies only one-month ahead forecasts, we predict equity returns also over longer horizons. Second and more importantly, we combine the recession and equity premium prediction literatures by directly using recession probability forecasts to forecast equity returns, while Gómez-Cram (2021) uses a state-space model to link expected excess equity returns to the business cycle. He estimates a common growth component from real-time data of nine U.S. macroeconomic variables related to real output, income, employment, consumer spending, investment, and inflation. We instead confirm that the term spread is a robust leading indicator of recessions and strongly improves equity premium forecasts. Related to our finding Andreasen et al. (2021) show that the yield spread better predicts bond risk premiums when conditioning on the business cycle.

The paper proceeds as follows. Section 2 summarizes the data used. Section 3 presents the main results, focusing first on the recession probability forecasts and then on the equity premium prediction using model-implied recession probabilities. Section 4 provides additional robustness checks and Section 5 concludes.

2 Data

We obtain data on the equity premium and term spread from Amit Goyal's homepage.[2] The equity premium is computed as the continuously compounded log return on the S&P 500 index, including dividends, minus the Treasury bill rate (Welch and Goyal, 2008). The term spread (TMS) is calculated as the difference between the long-term government bond yield and the Treasury bill rate. The yields are taken from Ibbotson's *Stocks, Bonds, Bills, and Inflation Yearbook* and have a maturity of approximately 20 years (Ibbotson and Sinquefield, 1976). Our data set consists of monthly observations from March 1951 to December 2019. We start our analysis in March 1951 after the Treasury-Federal Reserve Accord - which laid the foundation

[2] http://www.hec.unil.ch/agoyal/

for an independent monetary policy (Lacker, 2001). During World War II and the six years afterwards the Fed was tasked to support the financing requirements of the Treasury by stabilizing long-term interest rates (Eichengreen and Garber, 1991; Carlson and Wheelock, 2014). Hence, we begin our analysis after this extraordinary period of pegged interest rates. The business cycle chronology with classifications into expansions and recessions is taken from the National Bureau of Economic Research (NBER). A business cycle peak is defined to be the first month of a recession. We start our pseudo out-of-sample forecasting exercise in 1980 when the Business Cycle Dating Committee of the NBER began to release timely announcements of its business cycle classifications.

3 Empirical results

This section provides our empirical results. In Sections 3.1 and 3.2, we first confirm the ability of probit models along the lines of Estrella and Mishkin (1998) to predict NBER recessions. We show that the forecast performance of the standard probit model using the term spread as the only explanatory variable strongly improves when lagged information about the term spread is added. We then document in Section 3.3 that the recession probability forecasts implied by probit models have strong predictive power for the U.S. equity premium. In Section 3.4, we compare the results to OLS regressions based on term spread information. In Section 3.5, we provide evidence for a structural break in the mean of the term spread in 1982, and show that this break could have been identified in real-time. We then show in Sections 3.6 and 3.7 that break-correction methods can further improve recession and equity premium forecasts. In Section 3.8, we perform an asset allocation exercise showing that the recession forecasts based on information in the yield curve significantly improve market timing. In Section 3.9, we extend our analysis to characteristics portfolios and in Section 3.10 we provide additional international evidence. Finally, in Section 3.11 we show that estimated recession probabilities forecast the equity premium by predicting higher discount rates and lower future dividends.

3.1 Predicting recessions

In this section we are interested in predicting U.S. recessions as classified by the NBER Business Cycle Dating Committee. The literature typically distinguishes between the probability of a recession in exactly h months and the probability of a recession within the next h months. Here, we focus on the latter, as we aim to forecast cumulative log equity premiums over the next h

months in later sections; for similar definitions see Wright (2006) and Johansson (2018). More precisely, let $Y_{t+1:t+h} = 1$ if the NBER has classified at least one month between $t+1$ and $t+h$ as a recession. We follow common practice and assume that $Y_{t+1:t+h}$ is based on a latent variable $Y^*_{t+1:t+h}$ where $Y_{t+1:t+h} = 1$ for $Y^*_{t+1:t+h} \geq 0$ and $Y_{t+1:t+h} = 0$ for $Y^*_{t+1:t+h} < 0$. The latent variable is assumed to follow a (multivariate) linear regression model:

$$Y^*_{t+1:t+h} = X_t^{'}\beta + \epsilon_{t+1:t+h}, \tag{1}$$

$$\Pr\big[Y_{t+1:t+h} = 1 | X_t\big] = \Phi\big[X_t^{'}\beta\big], \tag{2}$$

where $X_t^{'} = (1, x_{1,t}, \ldots, x_{p,t})^{'}$ is the $1 \times (p+1)$ vector of predictor variables including the intercept, β is the $(p+1) \times 1$ vector of coefficients, and $\epsilon_{t+1:t+h}$ is the error term. Further, $\Phi\big[\,\cdot\,\big]$ is the cumulative distribution function of the standard normal distribution and Pr denotes probability. Let $\hat{p}_{t+1:t+h}$ be the out-of-sample forecast for $\Pr\big[Y_{t+1:t+h} = 1\big]$ based on information contained in X_t. We follow Jacobsen et al. (2019) and account for the fact that the NBER typically publishes business cycle classifications with a substantial delay by estimating the β coefficients with information up to $t-24$. This is a conservative choice as other authors (see, e.g., Kauppi and Saikkonen (2008)) only account for a delay of one year.[3] The sample with T observations is split into an in-sample estimation period of M months and an out-of-sample period of $T-M$ months. We only use data that are available in real-time mimicking as closely as possible the information an investor would have had.

In their seminal paper Estrella and Hardouvelis (1991) show that an inverted yield curve is a strong predictor of recessions and future real economic activity. They find that a decline of the term spread is associated with an increase in the probability of a recession four quarters ahead, and that this predictability is not incorporated in other variables such as lagged inflation, lagged real output growth, and survey data. Estrella and Mishkin (1998) complement this finding by comparing the predictive power of the term spread with financial variables such as stock prices and other spreads, as well as monetary aggregates. While some alternative predictors are useful over one- to three-quarter horizons, it is the term spread that predicts best over horizons of one-year and longer. Moreover, the binary models for recessions are found to be more stable than continuous models for economic growth (Estrella et al., 2003), and the

[3]The longest delay in our sample was 21 months: the NBER announced the November 2001 business cycle trough on July 17, 2003. Results are very similar when we use information up to $t-12$.

relation is also present in other countries including Germany, Japan, and the U.K. (Bernard and Gerlach, 1998). Wheelock and Wohar (2009) provide a comprehensive survey of the ability of the term spread to predict recessions.[4]

The observed predictability of an inverted or flat yield curve for future recessions is in line with counter-cyclical monetary policy: if the central bank tightens monetary policy by raising short-term interest rates, the yield curve tends to flatten and the economy slows down (Estrella et al., 2003). Estrella (2005) demonstrates this mechanism in a rational expectations model where the monetary policy regime and the reaction function are key determinants for the predictive power of the term spread. More recently, Adrian et al. (2019) argue that the predictive power of the term spread for future economic activity may result from active balance sheet management of financial intermediaries. If the liabilities of intermediaries are of shorter maturity than their loans, then a narrowing of the interest margin has a negative effect on the profitability of the marginal loan. As a result, they reduce the supply of credit. Following this logic, a decrease in the term spread has a negative effect on real activity and thus offers a causal mechanism for its strong forecasting power.

In what follows, we focus on the simple probit model using (transformations of) the term spread as input. The first model only includes a constant and the term spread as predictors - this is the model of Estrella and Hardouvelis (1991). It is well known that this model performs well for forecast horizons of one to two years, with only weak predictability for shorter horizons. More recently, Liu and Moench (2016) show that the short-horizon forecasts substantially improve when adding lagged observations of the term spread. Building on this, the second model includes the term spread lagged by six months as an additional predictor. As we will see below, the recession probabilities implied by the two models closely track the dynamics of the term spread and thus tend to be quite volatile. This implies a number of false positive signals about impending recessions. To address this issue, we also consider a specification that includes a constant, the term spread, as well as a backward-looking three-year moving average of the term spread (MA-TMS) which we construct as $\frac{1}{36}\sum_{j=0}^{35}\mathrm{TMS}_{t-j}$. We show in the Online Appendix

[4] Additionally, the literature has studied the predictive power of other variables and methods in detail. Fornaro (2016) analyzes the performance of a probit model with large-dimensional datasets and Bayesian shrinkage. The model performs well over short-horizons and generates smoother forecasts than benchmark models. Several articles document forecasting gains by including credit data like credit spreads, credit growth, and illiquidity measures (Chen et al., 2016; Ponka, 2017; Mihai, 2020), as well as by including sentiment variables (Christiansen et al., 2014). Similarly, there is increasing interest in whether daily and weekly data can improve monthly recession forecasts - for example in MIDAS regressions (Galvão and Owyang, 2020).

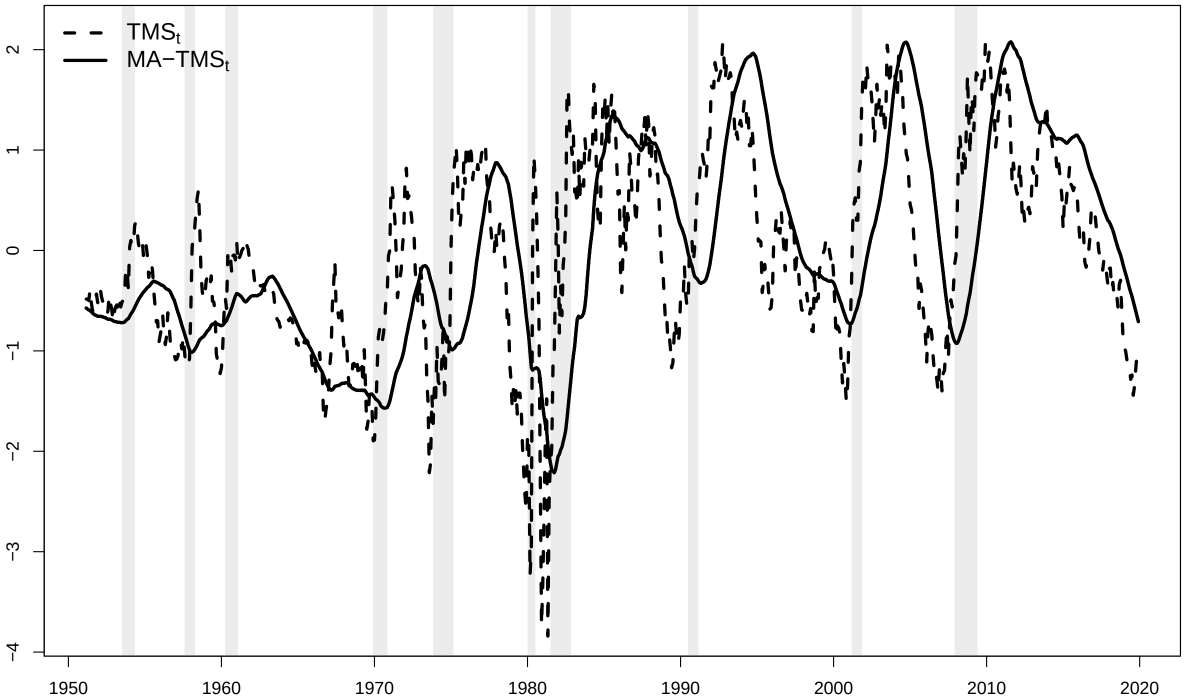

Figure 2
Term spread and moving average term spread
This figure presents the term spread (dashed line) and the moving average term spread (solid line), whereby the latter is the moving average of past three-year observations of the term spread. The time series are normalized to a mean of zero and a standard deviation of one. The sample is 1951:3 to 2019:12 and shaded areas indicate NBER-dated recession periods.

B that our results are robust to the length of the moving average window.

Figure 2 superimposes TMS and MA-TMS. While the local minima of the term spread usually lead the beginning of recessions by several months, the moving average term spread reaches its troughs often just before the onset of recessions. Moreover, the smoothed series averages out some local minima that are not followed by recessions. The smoothing thus emphasizes lower frequency components of the term spread which appear to be more relevant for signaling recessions. We will indeed show below that the incorporation of lagged and averaged term spread information into the probit model significantly enhances short horizon forecasts.

Figure 3 shows the out-of-sample recession probability forecasts for the three models from 1980:1 to 2019:12. Several points are noteworthy. First, the model with only a constant and the term spread performs relatively poorly for $h = 1$ and $h = 3$, with several false positives and no pronounced differences between expansions and recessions since the mid-1980s. The performance for this model gradually increases in the forecast horizon. This is consistent with the prior literature which has documented an improved recession prediction with the term spread for horizons beyond six months. Second, the models adding lagged term spread information perform substantially better for short-horizon forecasts, where the model with the moving average term spread implies substantially smoother recession probabilities. This is in line with Rudebusch et al. (2007) who show that the one-year lagged term premium predicts future GDP growth, and that differences rather than levels of the expectations component and term premium matter more for forecasting real output growth. The finding that lagged observations of the term spread improve recession predictability is consistent with monetary policy affecting the economy with a delay of a few quarters (Rudebusch and Williams, 2009).

Third, the recession probabilities in the 1990s and 2000s are less pronounced compared to the probabilities in the early 1980s - with values rarely exceeding 50% even in recessions. Similarly, Estrella et al. (2003) document that the signal of the models was weaker in the 1990-91 recession compared to the early 1980s. Kauppi and Saikkonen (2008) also document - using probit models with lagged dependent variables - that the 1990-91 and 2001 recessions were difficult to predict. This pattern is not unique to models using the term spread as predictor, the lack of predictability is also documented for models with larger sets of predictors (Hamilton, 2011; Fornaro, 2016). We show in Section 3.5 that the weaker recession signals result from a structural break in the mean of the term spread in the early 1980s. The probabilities are considerably stronger when this break is accounted for.

3.2 Forecast evaluation

We follow the recession prediction literature and use the quadratic probability score (QPS), the logarithm score (LS) and the diagonal elementary score (DES) to formally evaluate the accuracy of recession probability forecasts.[5] Perfect classification ability results in values of zero for all three scores; otherwise they have positive values, with higher values indicating poorer forecast performance, see, e.g., Nyberg (2013); Christiansen et al. (2014); Fornaro (2016) for recent applications. The loss function of LS penalizes large forecast errors more heavily than the loss function of QPS (Diebold and Rudebusch, 1989). Galvão and Owyang (2020) argue that the loss functions of LS and DES are more suitable in the context of rare events such as recessions. We further calculate the out-of-sample pseudo R^2 (Estrella, 1998) and the area under the receiver operating characteristic curve (AUROC). While the QPS, LS and DES metrics evaluate the model accuracy, the AUROC is a metric to evaluate the classification ability - here

[5]Details on the estimation of all statistics in this section are provided in the Online Appendix.

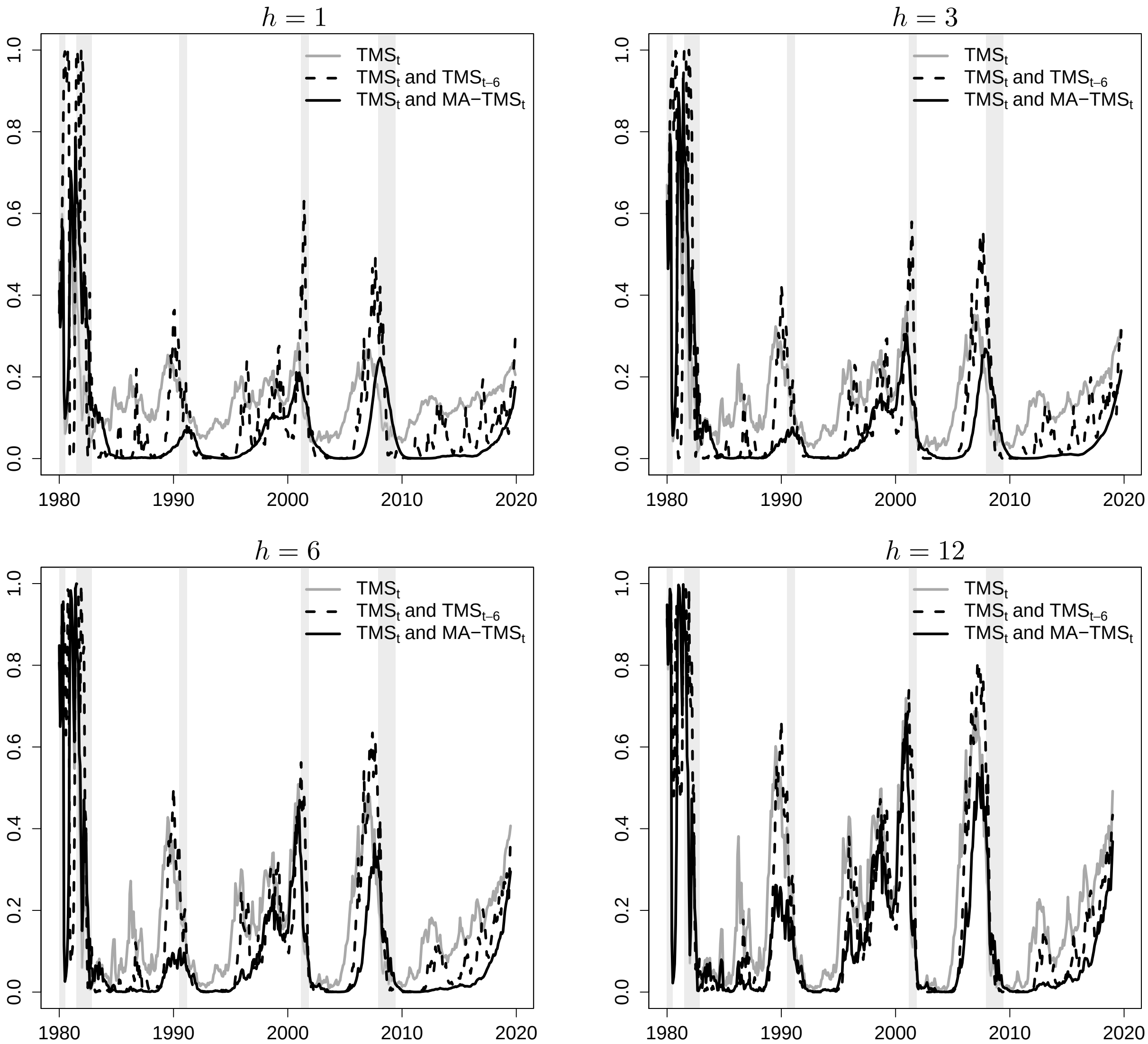

Figure 3
Out-of-sample recession probability forecasts
This figure shows the out-of-sample recession probability forecasts at the 1-, 3-, 6-, 12-month horizons. Results are presented for three different forecasting models: the solid gray line depicts forecasts from a model with the term spread as a predictor, whereas the solid (dashed) black line denotes a model with the term spread and the moving average of the past three years (six month lagged value) of the term spread as predictors. Gray bars denote NBER-dated recession periods and the out-of-sample period is 1980:1 to 2019:12.

into recessions and expansions - of a forecasting model. A perfect classifier has an AUROC of one and a coin-toss classifier has an AUROC of 0.5; for further details see Berge and Jordà (2011).

Table 2 presents values of the QPS, LS, DES, pseudo R^2 and AUROC for the three different models for horizons $h = 1, 3, 6, 12$ months ahead. The model with a constant and the term spread performs worst, with an AUROC close to 0.5 and a negative pseudo R^2 for $h = 1$ and $h = 3$. The performance improves with the forecast horizon, with an AUROC of 0.73 and a pseudo R^2 of 0.18 for $h = 12$, reflecting the well known finding that the term spread has a lead time of about four to six quarters (Estrella and Mishkin, 1998). When adding the six-months lagged term spread and the moving average term spread as predictors the model accuracy and classification ability substantially increase, especially for short-horizon forecasts. Each of the evaluation statistics improves for these more sophisticated probit models. At the one-month ahead horizon the AUROC jumps from 0.47 for the model with only the term spread to 0.81 for the model adding the lagged term spread and to 0.92 for the model adding MA-TMS. These differences in AUROC values, shown in the second to last column of the table, are highly statistically significant according to the test of Hanley and McNeil (1983). While the marginal improvement of predicting recessions by adding TMS_{t-6} or MA-TMS declines with the forecast horizon, it is statistically significant at the 1% level at all horizons. This highlights that adding lagged term spread information increases the recession classification precision of the probit models.[6]

The last column of Table 2 provides the correlation between the implied recession probability forecasts and the (cumulative) log equity premium over the next h months (ρ). The negative figures indicate that the equity premium tends to decrease when the recession probability rises, and that this correlation pattern is more pronounced for the models including lagged term spread information and for longer forecast horizons. At the one-year ahead horizon, the model using the backward-looking moving average term spread as additional regressor features a sizable negative 37% correlation of the implied recession probability with the cumulative equity premium over the next year.

The classification into expansions and recessions based on model-implied probabilities requires one to define a threshold level above which one calls a recession. Hence, the proportion of correctly predicted recessions (percentage of true positives, PTP) and the proportion of falsely predicted recessions (percentage of false positives, PFP) are functions of this threshold. The receiver operating characteristic (ROC) curve traces all combinations of PTP and PFP for different thresholds in the unit box. The diagonal line represents uninformative forecasts (PTP = PFP). Curves above the diagonal line depict informative forecasts and the ROC curve of a perfect classifier "will hug the north-west border of the positive unit quadrant" (Berge and

[6] We also formally test the null hypothesis that the AUROC value for the model with TMS and MA-TMS is equal to the AUROC value for the model with TMS and lagged TMS against the one-sided alternative that the former is statistically significantly larger than the latter (Hanley and McNeil, 1983). While the null is rejected at the 5% level for $h = 1, 3, 6$, it cannot be rejected for $h = 12$.

Table 2

Out-of-sample performance of probit models: Forecast evaluation statistics

This table presents five forecast evaluation statistics for the out-of-sample performance of three different probit models, as well as the correlation between the probability forecasts and the (cumulative) log equity premium (ρ). The statistics are the quadratic probability score (QPS), logarithm score (LS), diagonal elementary score (DES), pseudo R^2, as well as the area under the receiver operating characteristic curve (AUROC). The predictor variables are the term spread (TMS) and lagged and averaged variants of the term spread. MA-TMS$_t$ refers to the backward-looking three-year moving average of the term spread, and "Historical average" depicts forecasts from a probit model with only a constant. The recession probability forecasts refer to the probability that a recession occurs within the next h months. Results are shown for $h = 1, 3, 6, 12$ and the out-of-sample period is 1980:1 to 2019:12. We test the null hypothesis that AUROC = 0.5 (random classification) against the two-sided alternative (Hanley and McNeil, 1982); asterisks for this test are provided next to the AUROC value. ΔAUROC shows the gains relative to the probit model with TMS$_t$ only and asterisks denote that the AUROC of the respective bivariate model is significantly larger than the AUROC of the probit model with TMS$_t$ only based on the test of Hanley and McNeil (1983). *, **, and *** denote significance at the 10%, 5%, and 1% significance levels.

	1980:1 to 2019:12						
Variables in probit model	QPS	LS	DES	pseudo R^2	AUROC	ΔAUROC	ρ
Panel A: h = 1							
TMS$_t$	0.23	0.40	0.12	-0.01	0.47		-0.04
TMS$_t$, TMS$_{t-6}$	0.17	0.34	0.05	0.12	0.81***	0.34***	-0.07
TMS$_t$, MA-TMS$_t$	0.17	0.27	0.05	0.27	0.92***	0.45***	-0.14
Historical average	0.23	0.39	0.11	0.00	0.43**		0.05
Panel B: h = 3							
TMS$_t$	0.25	0.43	0.12	-0.01	0.54		-0.08
TMS$_t$, TMS$_{t-6}$	0.18	0.36	0.06	0.15	0.82***	0.28***	-0.15
TMS$_t$, MA-TMS$_t$	0.20	0.31	0.07	0.25	0.90***	0.36***	-0.22
Historical average	0.26	0.43	0.12	0.00	0.43**		0.09
Panel C: h = 6							
TMS$_t$	0.27	0.46	0.12	0.03	0.63***		-0.11
TMS$_t$, TMS$_{t-6}$	0.21	0.40	0.08	0.17	0.81***	0.18***	-0.19
TMS$_t$, MA-TMS$_t$	0.22	0.36	0.09	0.25	0.87***	0.24***	-0.25
Historical average	0.30	0.48	0.14	0.00	0.43*		0.11
Panel D: h = 12							
TMS$_t$	0.29	0.48	0.12	0.18	0.73***		-0.24
TMS$_t$, TMS$_{t-6}$	0.21	0.39	0.07	0.36	0.86***	0.13***	-0.33
TMS$_t$, MA-TMS$_t$	0.25	0.41	0.10	0.32	0.86***	0.13***	-0.37
Historical average	0.38	0.57	0.17	0.00	0.47		0.12

Jordà, 2011).

Figure 4 presents ROC curves for the three different forecasting models. The model with a constant and the term spread (solid gray line) is close to the diagonal line for $h = 1$ and $h = 3$ but shifts toward the north-west corner for $h = 6$ and $h = 12$. Hence, the model is relatively uninformative for short-horizons but gains predictive power with increasing h. Adding the lagged term spread and the moving average component helps to predict recessions especially

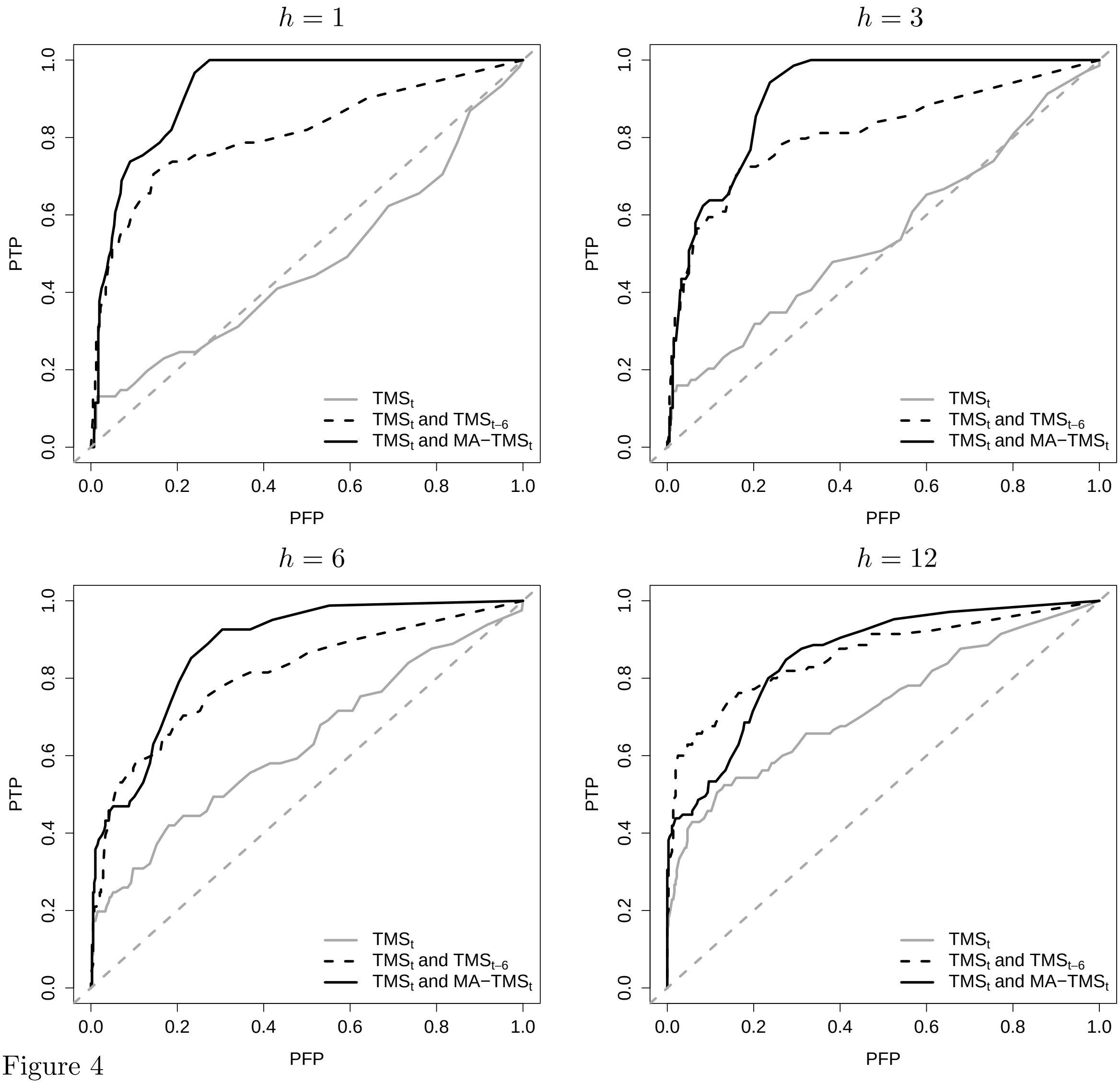

Figure 4

Out-of-sample performance: ROC curves

This figure shows the receiver operating characteristic (ROC) curve for three different probit models. The solid gray line depicts a model with a constant and the term spread, and the solid (dashed) black line presents the performance of a model when the moving average term spread (six-month lagged term spread) is added as a predictor. The dashed gray line is the 45 degree line. Results are shown for the out-of-sample period from 1980:1 to 2019:12. The vertical axis depicts the percentage of true positives (PTP) and the horizontal axis depicts the percentage of false positives (PFP). Predictions are made for a recession starting within the next $h = 1, 3, 6, 12$ months.

at these shorter horizons. The ROC curves substantially shift toward the north-west and the improvements for the moving average component are highest for $h = 1, 3, 6$. This is consistent with Figure 3: the first and second model have relatively high recession probabilities during the 1990s and 2010s, thus generating some false positives for low threshold levels. Overall, the ROC curves of the models with lagged and averaged term spread information lie well above the ROC curve of the spread only model for any forecast horizon and any threshold value. Thus, adding lagged term spread information strongly improves recession prediction.

3.3 Forecasting the equity premium with recession probability forecasts

Having shown that equity premiums are particularly low around business cycle peaks and that the onset of recessions can be well predicted using yield curve information, we next assess the usefulness of implied recession probabilities to forecast the equity premium out-of-sample. We use the standard linear predictive regression model:

$$r_{t+1:t+h} = \alpha_{t+h} + \beta_{t+h}\hat{p}_{t+1:t+h} + \epsilon_{t+1:t+h}, \tag{3}$$

where $r_{t+1:t+h} = \frac{1}{h}\sum_{j=1}^{h} r_{t+j}$ is the average of the cumulative log equity premium between $t+1$ and $t+h$, α_{t+h} and β_{t+h} are coefficients, and $\epsilon_{t+1:t+h}$ is the error term. Here, we use the recession probability forecasts $\hat{p}_{t+1:t+h}$ as predictor variables. It is important to note that we only use information that is available to investors in real-time. Suppose that we are interested in forecasting $r_{t+1:t+h}$ at time t. First, we estimate the coefficients of the probit model with information up to time $t-24$ to account for the fact that the NBER calls recessions typically with a few months delay. We then combine these estimated coefficients with the values of the term spread and its backward-looking moving average up to month t into the implied recession probabilities $\hat{p}_{t+1:t+h}$. Second, we regress the log equity premium until time t on a constant and the estimated in-sample recession probabilities until $\hat{p}_{t-h+1:t}$. Third, we use the estimated coefficients $\hat{\alpha}_t$ and $\hat{\beta}_t$ and the out-of-sample recession probability forecast $\hat{p}_{t+1:t+h}$ to predict $r_{t+1:t+h}$. Thus, the log equity premium forecast is $\hat{r}_{t+1:t+h} = \hat{\alpha}_t + \hat{\beta}_t\hat{p}_{t+1:t+h}$. We recursively re-estimate the coefficients of the probit model and the linear predictive regression model and real-time forecasts for each month over the period from 1980:1 to 2019:12.

We follow the convention in the literature and evaluate the forecast performance based on the out-of-sample R^2 of Campbell and Thompson (2008). The R^2_{OS} statistic measures the proportional reduction in mean squared forecast error (MSFE) relative to the benchmark model with only a constant ($\beta_t = 0$), see, among others, Rapach et al. (2010); Jiang et al. (2019):

$$R^2_{\text{OS}} = 1 - \frac{\sum_{j=M}^{T-h}(r_{j+1:j+h} - \hat{r}_{j+1:j+h})^2}{\sum_{j=M}^{T-h}(r_{j+1:j+h} - \bar{r}_{j+h})^2}, \tag{4}$$

where $\bar{r}_{j+h}$ is the prevailing mean with information up to period t. Welch and Goyal (2008) show that none of the theoretically motivated predictors such as the dividend-price ratio, term spread or book-to-market ratio can consistently outperform this naive benchmark. In fact, they find that the predictive power is mainly driven by the 1973-1975 oil shock, and that the period from 1975 to 2005 is characterized by "30 years of poor performanc" (Welch and Goyal, 2008, page 1504). We test the null hypothesis of a lower or equal MSFE from forecasts of the historical average benchmark ($R^2_{\text{OS}} \leq 0$) against the alternative that forecasts from the models using recession probabilities as predictors have a lower MSFE ($R^2_{\text{OS}} > 0$) using the MSFE-adjusted statistic of Clark and West (2007), which corrects for the fact that the Diebold and Mariano (1995) statistic follows a non-standard distribution for nested models. We account for serial correlation in the residuals by estimating Newey and West (1987) standard errors with lag lengths of h months.

Panel A in Table 3 presents the R^2_{OS} statistics (in %) when using the same three probit models as in the previous sections to derive recession probability forecasts. The following results are worth noting. First, while the R^2_{OS} statistic for TMS is negative (-0.83%) and insignificant for $h = 1$, it is positive (4.21%) and significant at the 5% level for $h = 12$. Hence, the forecasts from the standard probit model with the term spread as explanatory variable are helpful in predicting cumulative log equity premiums over the next year, although the reduction in MSFE relative to the historical average is below 5%. Second, the models with lagged term spread information have consistently positive R^2_{OS} values. The improvements for $h = 1$ are relatively small with values of 0.23% and 1.11%, respectively, but increase to almost 10% for $h = 12$. While these gains may appear small, Campbell and Thompson (2008) show that a monthly R^2_{OS} of only 0.50% can already translate into significant economic gains for an investor. Third, the best performing model uses the term spread and the moving average term spread to predict recessions. The R^2_{OS} values are significant at the 5% level for each forecast horizon and monotonically increase in h. In terms of magnitude, the R^2_{OS} statistics of these simple models are comparable to or larger than those of other predictors for longer horizons but are somewhat smaller for shorter forecast horizons (Huang et al., 2015; Chen et al., 2020). As noted by Rapach and Zhou (2013), the forecast performance heavily depends on the data set and on the state of economy. A thorough comparison with two recently proposed benchmark predictors follows in Section 3.7.

3.4 A comparison to the standard OLS approach

So far, we have shown that recession probabilities derived from probit models using the term spread as predictor significantly outperform the historical average benchmark. This contrasts

Table 3
Out-of-sample $\mathbf{R^2}$ statistics for log equity premium forecasts

This table reports R^2_{OS} statistics in % for the out-of-sample predictability of (cumulative) log excess returns on the S&P 500 index at the h-month ahead horizon relative to forecasts from the historical average. Forecasts are based on the linear predictive regression model with a constant and one predictor variable. Panel A shows results when forecasting with model-implied recession probabilities. The recession probability forecasts are derived by three different probit models: the first model only includes a constant and the term spread, whereas the second and third model add either the term spread lagged by six-months (TMS_{t-6}) or the three-year moving average of the term spread (MA-TMS_t) as additional predictors. Panel B shows results when the term spread variables are directly used as predictors in the OLS regression. CF-MEAN refers to an equally-weighted average of forecasts from univariate regressions with TMS_t, TMS_{t-6}, and MA-TMS_t, respectively. *, **, and *** indicate significance at the 10%, 5%, and 1% levels according to the Clark and West (2007) MSFE-adjusted statistic. The null hypothesis is equal MSFE and the alternative is that the more sophisticated model has smaller MSFE than the historical average benchmark. The out-of-sample period runs from 1980:1 to 2019:12.

(1)	(2)	(3)	(4)	(5)
Variable	$h=1$	$h=3$	$h=6$	$h=12$
		Panel A: Probit model		
TMS_t	-0.83	-1.38	-1.68	4.21**
TMS_t, TMS_{t-6}	0.23	1.58**	1.98**	8.23***
TMS_t, MA-TMS_t	1.11**	3.09***	3.83**	9.27***
		Panel B: OLS model		
TMS_t	-0.99	-2.17	-3.23	2.04**
TMS_{t-6}	-0.05	1.02**	2.40**	4.80***
MA-TMS_t	0.39*	1.22**	2.18**	4.26**
CF-MEAN	0.21	1.10*	1.96**	5.87***
TMS_t, MA-TMS_t	-1.18	-2.61	-4.18	0.24**

the common finding that forecasts from a regression of the equity premium on the term spread perform poorly for short horizons (Rapach and Zhou, 2013). In this section we compare forecasts from model-implied recession probabilities to forecasts from linear predictive regressions with TMS and MA-TMS as predictors. Panel B in Table 3 shows the R^2_{OS} statistics when directly forecasting the equity premium with the three variants of the term spread. This corresponds to Equation (3) replacing $\hat{p}_{t+1:t+h}$ with TMS_t, TMS_{t-6}, and MA-TMS_t, respectively. We further show results for a simple combination forecast that takes the average of the three individual OLS regression forecasts (CF-MEAN). While TMS only has mild predictive power over one-year ahead forecast horizons, TMS_{t-6} and MA-TMS_t also significantly predict the equity premium for $h=3$ and $h=6$. MA-TMS_t further outperforms the historical average for $h=1$, although with only a R^2_{OS} of 0.39%.

Our finding that adding lagged information significantly improves equity premium forecasts is surprising in light of the efficient market hypothesis. However, it is consistent with Gómez-Cram (2021) who documents that analysts only sluggishly revise their expectations downward and that stock prices do not fully reflect publicly available information on turning points. The last row in Panel B shows the performance of a joint OLS regression with TMS_t and MA-TMS_t. The OLS regression analogue to our probit model performs worst, in line with the common finding that multivariate regression models with several parameters often underperform the historical average. Importantly, at all forecast horizons even the best linear models are substantially outperformed by the equity premium forecasts based on recession probabilities.

The upper panel of Figure 5 superimposes one-month ahead equity premium forecasts from the recession probability based on TMS_t and MA-TMS_t (solid black line), the analogue OLS model with the two predictors (dashed black line), and the historical average benchmark (solid gray line). While the OLS model generates forecasts that are volatile both in recessions and expansions, the probit model forecasts are relatively stable in expansions and markedly higher than the historical average. As the implied recession probabilities are high just before the 1981-82 recession, the implied equity premium forecasts are sharply negative around that time. This effect, although substantially less pronounced, is also visible around the business cycle peaks in 2001 and 2007.

Welch and Goyal (2008) have popularized a simple way to visualize the relative forecast performance of different prediction models over time. The lower panel in Figure 5 follows their approach and plots the difference in cumulative squared forecast errors (CSFE) for the historical average and the CSFE for two different models: the solid black line depicts equity premium forecasts based on the implied recession probability using TMS and MA-TMS, whereas the dashed black line shows the OLS model with both TMS and MA-TMS. An increasing curve indicates superior performance relative to the naive benchmark. We can see that the curve for the OLS model is decreasing over most of the sample, with reversed trends only around the 1981-82 and 2001 recessions. In contrast, the curve for the probit model forecasts is rising over most of the sample, indicating that the recession probability forecast of the equity premium consistently outperforms the historical average.

This superior performance is driven by two distinct effects. First, and similar to the OLS model, the model significantly predicts the negative excess returns in the one-year window around the peak in 1981. This is consistent with Table 1 and with the notion that term spread information anticipates recessions. Second, and more importantly, the model-implied recession probabilities outperform the naive benchmark also in expansions. This contrasts the "no predictability in good times puzzle" that is often documented in related articles (Huang et al., 2017). The explanation is simple: Table 1 shows that the annualized equity premium averages 6.31% in our sample, but is even higher at an annualized 8.40% in expansions. While the historical average closely tracks the full-sample average, our recession probability-based forecast corrects for negative values around business cycle peaks and thus correctly predicts a higher equity premium in expansions.

3.5 Structural break in the mean of the term spread

We have seen above that the implied recession probabilities are muted and rarely exceed 50% after the mid-1980s. This is consistent with e.g. Chauvet and Potter (2002) who find evidence for a structural break in the probit model based on the term spread but argue that the exact date of the break is difficult to localize. When fixing the break to 1984, they show that the model is able to predict the recession in 2001 with probabilities as high as 90% for the 12 month ahead forecast horizon. Galvão (2006) proposes a structural break threshold-VAR (SBTVAR) model that allows for non-linearities and breaks in the link between the term spread and U.S. output growth and identifies a break in 1985:2. She further shows that the SBTVAR model correctly anticipates the 2001 recession in real-time. Other papers also document breaks in the dynamic relationship between the term spread and real growth. Chauvet and Potter (2010) find evidence for recurrent breaks in a probit model with industrial production, sales, personal income, and employment. Schrimpf and Wang (2010) analyze whether the predictive power of the term spread on output growth suffers from structural breaks. They allow for multiple breaks using the test procedure of Bai and Perron (1998, 2003) and find evidence for breaks in Germany, Canada, U.K., and the U.S.

To summarize, there is ample prior evidence for a structural break in the link between the term spread and output growth, as well as for a break in the estimated recession probabilities from the standard probit model. This suggests that accounting for such a break may improve the recession probability forecasts and hence the equity premium predictability. In what follows, we provide further evidence for a structural break. Instead of focusing on a break in the estimated relation between the term spread and future recessions, we focus our attention on a break in the mean of the term spread. The reason is that a break in the parameters of the probit model is generally difficult to identify and even more so difficult to narrow down to an exact point in time - not least because of the infrequent occurrence of recessions (Wright, 2006). In contrast,

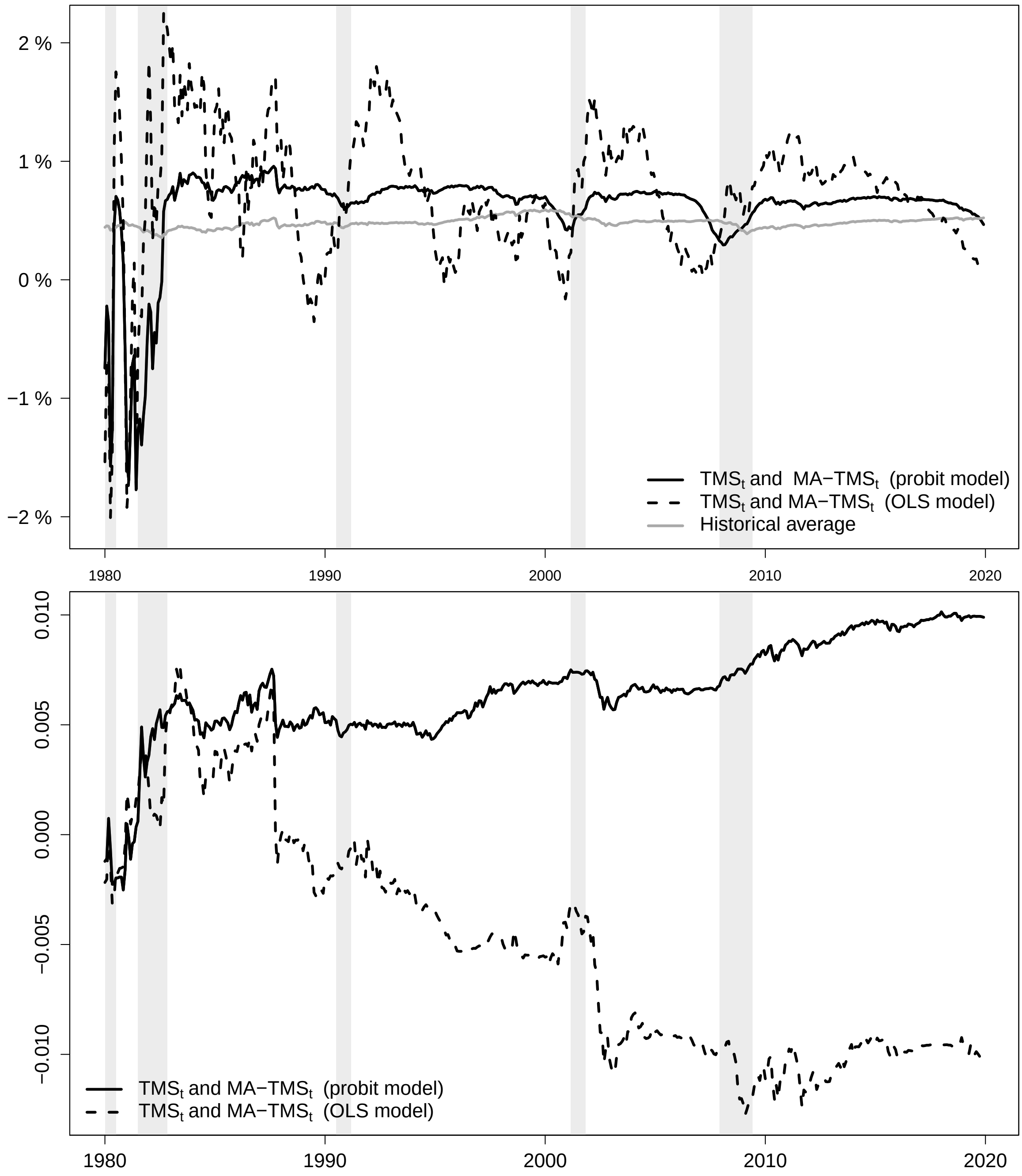

Figure 5

Out-of-sample forecasts and performance over time

This figure shows one-month ahead forecasts of the log equity premium for three different models (upper panel). The solid black line depicts forecasts from model-implied recession probability forecasts of a probit model with the term spread (TMS) and the backward-looking three-year moving average of the term spread (MA-TMS). The dashed black line presents forecasts from a standard linear predictive regression model with TMS and MA-TMS as predictors, and the solid gray line denotes the historical average. The lower panel shows the difference between cumulative squared forecast errors (CSFE) of the historical average and the CSFE of the probit model forecasts (solid black line) and the OLS model forecasts (dashed black line). All forecasts are estimated with a recursively expanding information set that mimics the real-time situation of an investor. The out-of-sample period is 1980:1 to 2019:12.

as we will see below, the shift in the mean of the term spread can be identified in a more timely manner regardless of the state of economy. Figure 2 plots the normalized term spread from 1951:3 to 2019:12. Eyeballing this time series shows that the mean of the term spread has shifted upwards in the early 1980s. This is most visible when comparing the period from 1965 to 1982 with the period from 1983 to 2019: the mean of the former period is -0.61 whereas it is 0.46 for the latter period. In what follows, we formally test the hypothesis of a structural break in the mean of the term spread.

The classical break test for coefficients in linear regression models goes back to Chow (1960). A critical limitation of the Chow-test is that the break date has to be known a priori. Here, we treat the break date as unknown and perform break tests over a grid of candidate values – namely on a fraction of the sample between $[\tau_1, \tau_2]$ with $\tau_1 = \pi_\tau T$ and $\tau_2 = (1 - \pi_\tau)T$. We refer to π_τ as the trimming value. When performing the Chow test on a sequence of dates the standard chi-square critical values are not applicable; for a discussion of this point see Hansen (2001). We estimate the following model for all z values between $[\tau_1, \tau_2]$:

$$\text{TMS}_t = \beta_1 \mathbf{I}\{t \leq z\} + \beta_2 \mathbf{I}\{t > z\} + \epsilon_t, \tag{5}$$

where $\mathbf{I}\{t \leq z\}$ ($\mathbf{I}\{t > z\}$) is an indicator function that equals one for $t \leq z$ ($t > z$). The coefficients β_1 and β_2 are re-estimated for a grid of z values and the SSE values are saved for each of these grid points. If there is no structural break in the coefficients then the SSE values vary randomly over time. However, if there is a unique structural break, then the time series will have a well-defined global minimum near the true break date (Hansen, 2001).

The upper panel in Figure 6 presents the SSE as a function of z with a trimming value of $\pi_\tau = 0.15$. The SSE is thus calculated for potential breaks from 1961:6 to 2009:8. The resulting SSE clearly shows a strong v-shape, indicating a well defined and unique break point. The break date corresponds to the month with the lowest sum of squared errors (Bai, 1997). This global minimum is in 1982:5. We formally test the null hypothesis of no structural break by using the Sup-F, Ave-F, and Exp-F statistics of Andrews (1993) and Andrews and Ploberger (1994). The variance-covariance matrix is estimated according to Newey and West (1994) and the p-values are computed following Hansen (1997). The null hypothesis of no structural break is rejected at the 1% significance level for Sup-F, Ave-F, and Exp-F and is robust to changes in the trimming value and pre-whitening of the residuals; details are provided in the Online Appendix. The middle panel in Figure 6 presents the estimated sub-sample means in the term spread with full sample information. The upward shift in the mean is consistent with attenuated recession probabilities after the break that we have observed above.

The previous test results are based on full sample information. Would an investor have been able to identify the break in real-time? To answer this question we estimate the p-values with a recursively expanding sample from 1980:1 onwards and re-estimate the p-values each month until 2019:12. The lower panel in Figure 6 presents the resulting series of p-values. The null hypothesis of no structural break is first rejected at the 10% critical value by the Sup-F, Ave-F, and Exp-F tests in 1986:7, 1987:3, and 1986:9. Since then, the p-values have consistently remained below 5%, providing strong evidence that the break in the mean of the term spread could have been identified in real-time as early as the mid 1980s. From 1995:1 to 2019:12 the null hypothesis is always rejected at the 1% level for each of the test statistics. The estimated break date is identical for the different test statistics as it is simply localized at the global minimum of the sum of squared errors (Bai, 1997).[7] Overall, the results are in line with other evidence of structural breaks in the standard probit model and the estimated break date is close to the structural change in variance of U.S. GDP growth in the early 1980s, see e.g. Kim and Nelson (1999), McConnell and Perez-Quiros (2000), and Pettenuzzo and Timmermann (2017).[8]

3.6 Forecasting in the presence of structural breaks

We have documented in the previous section that the term spread suffers from a structural break in the mean in 1982. We now show how to adjust the recession prediction and equity premium forecasts in the presence of this break.

First, we apply the procedure by Lettau and Van Nieuwerburgh (2008). They find evidence in favor of multiple shifts in the mean of the dividend-price ratio. To correct for these breaks, they create a break-adjusted time series by subtracting the sub-sample means from the dividend-

[7]Figure A.2 in the Online Appendix displays recursively estimated break dates from 1980:1 to 2019:12. This shows that the identified break dates are very stable around the full-sample break date 1982:5.

[8]We also test for multiple breaks in the mean of the term spread by applying the methods proposed in Bai and Perron (1998, 2003). The sequential tests do not provide evidence in favor of multiple breaks. Moreover, we apply the sequential tests to a longer sample that starts in 1933:4, just after the Great Depression. Interestingly, we find evidence for another break in the mean in 1947:6, in addition to the one in 1982:5. The additional break aligns well with the Treasury-Federal Reserve Accord, indicating a change in monetary policy after longer-term interest rates were pegged during wartime (Eichengreen and Garber, 1991; Carlson and Wheelock, 2014).

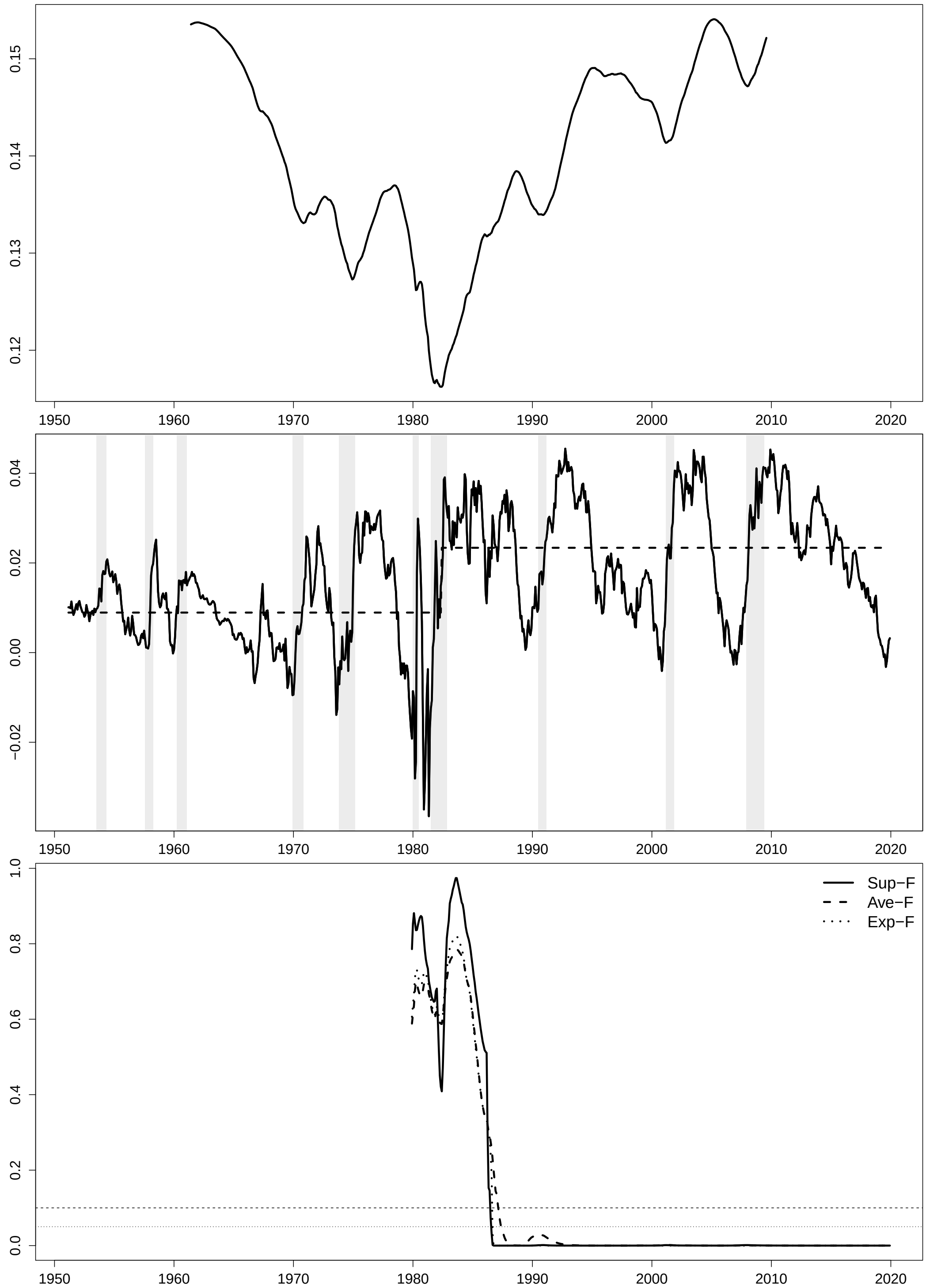

Figure 6

SSE as a function of the break date and real-time detection of the break

This figure presents the sum of squared errors (SSE) when testing for a structural beak in the mean of the term spread (upper panel). The change point is allowed to lie between 1961:6 and 2009:8, which corresponds to a trimming value of 15%. The analysis is based on full sample information and the coefficients and SSE are re-estimated for each potential break date. The middle panel shows the term spread from 1951:3 to 2019:12 (solid line) and the sub-sample means from 1951:3 to 1982:5 and from 1982:6 to 2019:12 (dashed line). The estimated break date is in 1982:5 and corresponds to the global minimum in SSE (Bai, 1997). The lower panel reports the recursively estimated p-values of the null hypothesis of no structural break (Hansen, 1997). This analysis is feasible in real-time and recursively expands the information set. Sup-F, Ave-F, and Exp-F refer to the Wald-type statistics of Andrews (1993) and Andrews and Ploberger (1994). The dashed (dotted) horizontal line shows the 10% (5%) level. The estimation of p-values is carried out from 1980:1 to 2019:12.

price ratio. They then show that the adjusted time series has robust in-sample predictive power for the equity premium but fails to beat the historical average out-of-sample.[9] Specifically, we carry out the following steps: we first test for a break in the mean of the term spread by estimating the Sup-F statistic and by using a significance level of 10% and a trimming value of 15%. Only real-time information is used to mimic the situation of an investor. Second, if the null hypothesis of no break is rejected, we estimate the two sub-sample means and subtract them from the term spread to create a break-adjusted term spread, denoted by $\text{TMS}^{\text{break}}$. Finally, we estimate the probit model with this adjusted time series and generate out-of-sample forecasts for recession probabilities and the log equity premium. If the null hypothesis is not rejected, we predict with the unadjusted term spread.

Alternative approaches have been proposed by Pesaran and Timmermann (2007). They present methods to determine the optimal estimation window in the presence of structural breaks. These methods are based on the insight that if the pre-break data follow a data generating process that is different from the one characterizing the post-break data, then the coefficient estimates are biased when using all data. Pesaran and Timmermann (2007) show in a simulation study that the MSFE can be significantly reduced by combining different forecasts from the same model when a structural break is present. The individual forecasts only differ in their estimation window.

We implement the combination of different forecasts from our probit models as follows. If the null hypothesis of no break is rejected, we estimate the probit model over an equally-spaced grid of starting values. This grid covers the beginning of our data set until the estimated break date. To reduce computing time, we only estimate models at annual increments in the starting date. This provides us with multiple recession forecasts and equity premium forecasts that only differ in the start date of the estimation window. Then, our pooled forecast is simply the average of the individual forecasts over the grid of start values, denoted as "Pooling (average)".[10]

[9] Lettau and Van Nieuwerburgh (2008) argue that the break dates can be estimated in real-time but that the uncertainty about the shift in the mean prevents significant forecasting gains. Smith and Timmermann (2021) present a method that uses cross-sectional information and economically motivated priors to (i) better detect breaks in real-time and to (ii) estimate parameters more accurately. The latter point is especially relevant when only few post-break observations are available.

[10] Additionally, we have estimated forecasts from weighted pooling. As the differences are negligible, we only show results for the equal-weighted combination of forecasts. We demand at least 15 years of data in the probit model to guarantee reliable coefficient estimates. Suppose that the Sup-F test first identifies in 1986:7 that a break has occurred in 1981:3. Then, the shortest estimation window uses data starting in 1969:7. This is because we have to lag the data by 17 years (not 15 years) to account for the delay in NBER announcements. Hence, the probit model is estimated over the grid of start dates from 1969:7,1968:7,... until the beginning of the forecast sample.

A disadvantage of the pooling approach is that one includes many forecasts with a large fraction of pre-break data when the sample is long or when the break occurs relatively late in the sample. Alternatively, one can only choose the best performing grid point. We implement this approach by performing a pseudo out-of-sample exercise over the most recent five years of data. Then we evaluate all start dates over this holdout period and select the start date that minimizes the MSFE for the equity premium. The selection of the grid point is chosen based on forecasts of the log equity premium and not on forecasts of the probit model as the former is the main purpose of this paper. Forecasts from this approach are denoted as "Cross-validation".

Finally, we consider one estimation strategy that only uses post-break data to estimate parameters of the probit model and to forecast recession probabilities and the equity premium. We denote this strategy as "Post-break window".[11] If the null hypothesis of no structural break is not rejected, then the forecasts of the break-correction methods are identical to forecasts by the unadjusted probit model. In what follows, we apply these break-adjustment methods to the probit model with TMS and MA-TMS as predictors.

Figure 7 shows recession probability forecasts for cross-validation and for post-break window, as well as for the standard probit model with the term spread. The estimated probabilities for cross-validation remain at fairly low levels during the 1990-91 and 2001 recessions for short-horizon forecasts. However, the probabilities for the 2008-09 recession increase substantially relative to the unadjusted full-sample models. They rise above 60% for cross-validation and are thus considerably higher than the probabilities in Figure 3 which are less than 30% for $h = 1$ and unadjusted models. Moreover, for $h = 12$ we see a substantially improved forecast performance with estimated probabilities as high as 90% for the 2001 and 2008-09 recessions. Post-break window generates out-of-sample probabilities as high as 90% for the 2001 recession, however, this approach also gives rise to a false positive in the late 1990s. Overall, we confirm previous findings that the implied recession probabilities are more pronounced when properly correcting for instabilities and breaks (Galvão, 2006; Chauvet and Potter, 2010). We provide forecast evaluation statistics for the break-correction methods in the Online Appendix.

[11] We use at least 15 years of data. Hence, while the sample of post-break data is shorter than 15 years we use the most recent 15 years of data for estimation and forecasting.

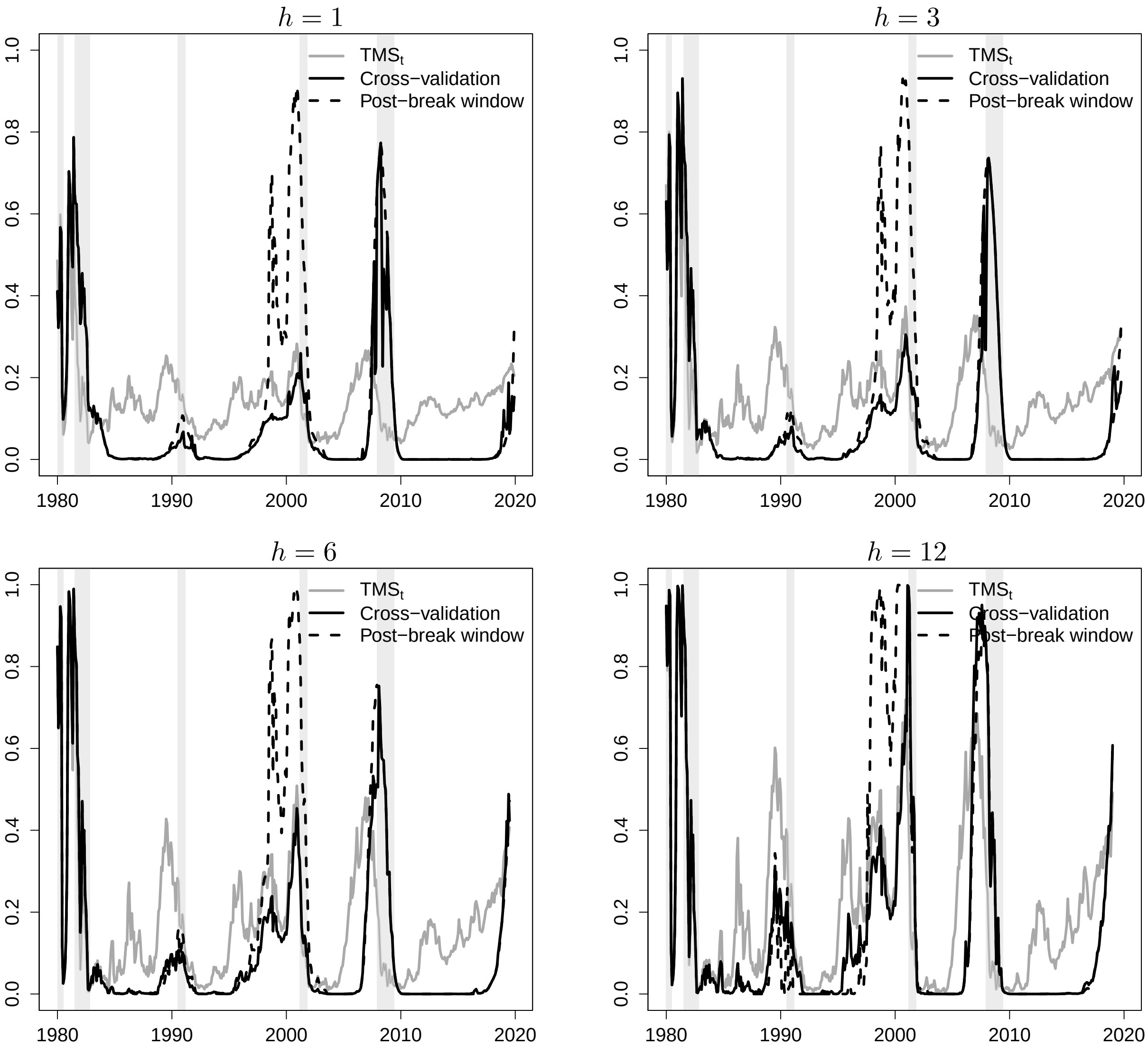

Figure 7
Out-of-sample performance: (break-corrected) recession probabilities
This figure presents out-of-sample recession probability forecasts for four different forecast horizons and three different models. The forecasts denote the probability of a recession within the next $h = 1, 3, 6, 12$ months. The solid gray line shows forecasts from the standard probit model with the term spread as the only predictor (TMS_t). The solid black line depicts forecasts from a probit model with the term spread and the moving average component (MA-TMS_t), where the optimal estimation window is determined by cross-validation over a holdout period of 60 months. The dashed black line presents forecasts from a probit model with TMS and MA-TMS that only uses post-break data for coefficient estimation. Out-of-sample forecasts are recursively estimated for the sample from 1980:1 to 2019:12.

3.7 Correcting for structural breaks: Out-of-sample equity premium prediction

We have seen in the previous section that the recession probability forecasts are substantially higher in the second part of the sample when using the cross-validation and post-break window approaches to correct for the break in the mean of the term spread in 1982. In this section we compare the equity premium forecasts from recession probabilities of the probit model with the unadjusted term spread and moving average term spread with forecasts from the same model when applying the break-correction methods. We will see that the break adjustment substantially improves the equity premium predictability of the implied recession probabilities.

Table 4 provides the R^2_{OS} statistics for the out-of-sample equity premium prediction using the break-correction methods to adjust the term spread relative to the historical average. Panel A shows that the one month ahead forecasts improve for each of the four methods, and that the R^2_{OS} is as high as 3.2% for post-break window for the prediction sample from 1980:1 to 2019:12, shown in column (1). Columns (2) to (4) show the R^2_{OS} values for different subsamples. While the gains in statistical predictability relative to the historical average are significant also from 1980-1999 (column (2)), they are much stronger for the period since 2000 (column (3)). The R^2_{OS} statistics in this sub-period exceed 3% for break-adjusted term spread, cross-validation, and post-break window. It is worth noting that the estimated predictive coefficients (not shown) are consistently negative and highlight that the superior performance relative to the historical average is driven by negative equity premium forecasts during recessions. This is in contrast with Campbell and Thompson (2008) who argue that imposing non-negativity constraints on the equity premium can improve performance.

The results are qualitatively the same for longer forecast horizons, shown in panels B to D. The R^2_{OS} statistics are consistently positive for the different sub-samples and gradually increase in the forecast horizon h - with R^2_{OS} values above 10% for cumulative one year ahead equity premiums. The only exception is the probit model with the break-adjusted term spread which has negative values between 1980:1 to 1999:12, resulting from a poor performance during the 1990s. Comparing the different break-adjustment methods of Pesaran and Timmermann (2007), we observe a clear pattern. The approach that only uses post-break data performs best, closely followed by cross-validation, and with some distance followed by pooling. In our setting, the break in the mean is relatively sizable and the break-date can be estimated accurately in real-time. Therefore, the improvements from pooling are smallest, due to the large impact of pre-break information. Cross-validation seems to be the most robust choice both for improving recession and equity premium forecasts. The post-break window also works well in our application but may be more risky from an ex ante perspective, putting all weight on post-break forecasts. Hence, our preferred break-correction method is cross-validation.[12]

[12]We show the recursively selected estimation windows for one-year ahead forecasts from cross-validation in the Online Appendix.

Table 4
Out-of-sample $\mathbf{R^2}$ statistics when correcting for a structural break

This table presents R^2_{OS} statistics (in %) for forecasts of the $h = 1, 3, 6, 12$ months ahead (cumulative) log equity premium. This statistic measures the reduction in MSFE relative to forecasts from the historical average. Results are shown for a probit model with the term spread and the moving average component as predictors, and for four methods that correct for a structural break in the mean of the term spread. The first method forecasts with a break-adjusted term spread ($\mathrm{TMS}_t^{\mathrm{break}}$, $\frac{1}{36}\sum_{j=0}^{35}\mathrm{TMS}_{t-j}^{\mathrm{break}}$), see Lettau and Van Nieuwerburgh (2008). Cross-validation selects an optimal estimation window over a holdout period, whereas pooling combines forecasts from several models with a grid of different starting values. Post-break window refers to forecasts from a model that only uses post-break data. For further details see Section 3.6 and Pesaran and Timmermann (2007). Short interest and gold-to-platinum ratio are the predictors of Rapach et al. (2016) and Huang and Kilic (2019), and perfect classifier refers to a simple two-state model that can perfectly anticipate NBER-dated recessions. *, **, *** denote significance at the 10%, 5%, and 1% significance levels according to the Clark and West (2007) MSFE-adjusted statistic. Columns (1) to (4) report results for different sub-samples, and panels A to D present results for different forecasting horizons.

Variable	(1) 1980:1-2019:12	(2) 1980:1-1999:12	(3) 2000:1-2019:12	(4) 1990:1-2013:12
Panel A: h = 1				
TMS_t, MA-TMS_t	1.11**	1.52**	0.67**	0.74**
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	1.60***	0.03	3.29***	2.17**
Cross-validation	2.27***	1.30**	3.31**	2.94**
Pooling (average)	1.45***	1.33**	1.59***	1.40***
Post-break window	3.21***	1.64**	4.89***	4.57***
Perfect classifier	1.06*	0.10	2.09*	1.74*
Short interest				1.67**
Gold-to-platinum ratio				1.15**
Panel B: h = 3				
TMS_t, MA-TMS_t	3.09***	4.40**	1.84**	2.17**
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	3.41***	-1.11	7.80***	4.56**
Cross-validation	7.25***	3.70**	10.76**	9.25**
Pooling (average)	4.10***	3.95**	4.26***	3.78***
Post-break window	8.78***	4.43**	13.12***	11.69***
Perfect classifier	2.50*	0.27	4.71*	3.74*
Short interest				5.46***
Gold-to-platinum ratio				4.54**
Panel C: h = 6				
TMS_t, MA-TMS_t	3.83**	5.04**	2.91*	3.54**
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	3.02**	-7.34	10.94***	5.91**
Cross-validation	10.53***	3.16*	16.04**	14.25**
Pooling (average)	5.54***	4.10*	6.62***	6.01***
Post-break window	11.36***	5.45**	15.78**	15.49**
Perfect classifier	3.63*	-0.13	6.43*	5.49*
Short interest				9.44**
Gold-to-platinum ratio				8.52**
Panel D: h = 12				
TMS_t, MA-TMS_t	9.27***	10.10**	8.87**	8.58***
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	8.08**	-8.75	18.74***	8.17*
Cross-validation	12.28***	7.76*	15.27***	14.13**
Pooling (average)	11.89***	7.73*	14.56***	11.38***
Post-break window	16.67***	13.22**	19.10**	18.45**
Perfect classifier	10.23***	3.70*	14.52***	12.55**
Short interest				7.85*
Gold-to-platinum ratio				9.55**

To better understand the strong predictive power of recession probabilities for the equity premium, we provide separate R^2_{OS} statistics for expansions and recessions in Table 5. Strikingly,

Table 5

Out-of-sample $\mathbf{R^2}$ statistics in recessions and expansions

This table presents R^2_{OS} statistics (in %) for forecasts of the $h = 1, 3, 6, 12$ months ahead (cumulative) log equity premium. Results are shown for a probit model with the term spread (TMS) and the moving average component (MA-TMS) as predictors, and for four methods that correct for a structural break in the steady state mean of the term spread. For further details see Section 3.6. The R^2_{OS} statistics are displayed separately for recessions and expansions. *, **, *** denote significance at the 10%, 5%, and 1% significance levels according to the Clark and West (2007) MSFE-adjusted statistic. Column (1) reports the respective forecasting model, and columns (2) to (5) present results for different forecasting horizons. The out-of-sample period is 1980:1 to 2019:12.

(1)	(2)	(3)	(4)	(5)
Variable	$h = 1$	$h = 3$	$h = 6$	$h = 12$
		Panel A: Recessions		
TMS_t, MA-TMS_t	1.83*	2.59*	-1.08	-1.20
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	5.94**	9.00**	5.95**	9.44**
Cross-validation	6.72**	14.15**	15.11**	8.33***
Pooling (average)	3.40**	5.41**	2.74**	4.77***
Post-break window	9.84***	17.62***	18.20**	17.72**
		Panel B: Expansions		
TMS_t, MA-TMS_t	0.86**	3.35***	7.03**	15.31***
$\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$	0.11	0.54**	1.11*	7.29**
Cross-validation	0.75**	3.70***	7.56***	14.56***
Pooling (average)	0.79**	3.43***	7.37***	16.00***
Post-break window	0.93**	4.22***	6.91***	16.07***

we see that for essentially all considered models and forecast horizons, there is an economically and statistically significant improvement over the historical average benchmark in both recessions *and* expansions. As discussed above, the reason is the following. Recession probabilities correctly predict low equity premiums in recessions. In addition, by adjusting for low equity premiums in recessions they correctly predict higher equity premiums than the historical average benchmark in expansions.[13] That said, the break-correction methods improve the performance primarily in recessions, as the adjusted recession probabilities predict a highly negative equity premium around the peak in 2007. For example, the R^2_{OS} for the one-month ahead forecasts improves from 1.8% to 6.7% and 9.8% for cross-validation and post-break window.[14]

[13] Note that the R^2_{OS} statistic for the unadjusted probit model turns negative in recessions for six- and twelve-month ahead forecasts and becomes highly significant in expansions. This reflects the fact that for longer horizons the recession forecasts substantially decrease prior to the peak and, therefore, have the strongest predictive power already before the beginning of the recession, see Figure 1.

[14] We perform encompassing tests to study whether the forecasting models provide distinct information. Specifically, we apply the ENC-T statistic of Harvey et al. (1998) to test whether post-break window encompasses the information of the other four forecasting models. Formally, we test the null hypothesis that $\lambda = 0$ in a convex combination of forecasts, $\hat{r}^c_{t+1:t+h} = (1-\lambda)\hat{r}^{\text{post-break window}}_{t+1:t+h} + \lambda \hat{r}^j_{t+1:t+h}$ with j being equal to one of the four alternative models (TMS_t, MA-TMS_t; $\mathrm{TMS}_t^{\mathrm{break}}$, MA-$\mathrm{TMS}_t^{\mathrm{break}}$; Cross-validation; Pooling (average)). We cannot reject the null hypothesis that post-break window encompasses the other forecasts. It may, however, still be possible that the other models provide equally good forecasts. We thus test whether the other forecasts encompass the forecasts from post-break window ($\lambda = 1$). For $h = 1$ and $h = 3$ we reject the null hypothesis for all models. However, for $h = 6$ and $h = 12$ we cannot reject the null that cross-validation encompasses post-break window. For $h = 12$ we can likewise not reject the null that TMS_t, MA-TMS_t and Pooling (average) encompass post-break window. In sum, we conclude that post-break window encompasses the other forecasts for $h = 1$ and

To assess the relative performance of our approach, we compare our results with three additional benchmark models. First, we generate forecasts from a simple two-state model that can perfectly foresee NBER recessions. We recursively estimate the following regression:

$$r_t = \alpha + \beta \times \mathbf{I}\{\text{NBER}_t = 1\} + \epsilon_t, \tag{6}$$

where $\mathbf{I}\{\text{NBER}_t = 1\}$ is an indicator function that equals one in recessions. If there is a recession in the next month then the forecast equals the average log equity premium in past recessions and, vice versa, if the next month is in an expansion then the forecast equals the average during past expansions. We implement this benchmark to see if the inverse v-shape in estimated recession probabilities has any value above and beyond simply classifying periods into expansions and recessions. The second benchmark predictor is the short interest variable of Rapach et al. (2016), which they characterize as "*the* strongest known predictor of aggregate stock returns". Short interest is calculated as the log of the equally-weighted mean of short interest across publicly listed stocks on U.S. exchanges. The series shows a strong linear upward trend and is therefore recursively detrended. Our third benchmark is the (log) gold-to-platinum ratio of Huang and Kilic (2019), constructed as the log of the ratio of gold to platinum prices, which the authors show to perform particularly well out-of-sample over longer forecast horizons.[15] Due to data availability, we can only evaluate the latter two predictors over a shorter sample from 1990:1 to 2013:12.

We see in Column (4) of Table 4 that cross-validation and post-break window consistently outperform the three benchmark models for each forecast horizon. The gains in reduced MSFEs are substantial, with an R^2_{OS} statistic that is often twice as large. The R^2_{OS} of cross-validation is 2.94% for $h = 1$, compared to 1.67% for short interest, 1.15% for gold-to-platinum ratio, and 1.74% for a perfect recession classifier.[16] At the one-year ahead horizon, cross-validation delivers an R^2_{OS} statistic of more than 14%, compared to 7.85% for short interest and 9.55% for the gold-to-platinum ratio.

$h = 3$ but that other models contain similar information for longer horizons. We therefore report results for all three break-adjustment methods.

[15]We thank Dave Rapach for making the data publicly available on his homepage, and Darien Huang and Mete Kilic for kindly sharing their data with us.

[16]The results differ slightly compared to those published in the papers because our historical average uses data from 1951:3 onwards. We re-estimate the predictive coefficients recursively using expanding data available in real-time.

Figure 8 plots the difference in CSFE for the historical average and CSFE for different models using recession probabilities: the unadjusted probit model, as well as cross-validation, pooling, and post-break window. An increasing line indicates that the historical mean is outperformed during this period, whereas the opposite is true for a decreasing line. The upper panel presents results for one month ahead forecasts. The solid gray line shows the relative performance for the unadjusted probit model. It performs very well during the 1981-82 recession but does only increase slightly compared to the other models thereafter. For cross-validation and pooling we see a similar pattern until 2007, which is followed by a positive jump during the Great Recession. The post-break window performs best, with a positive trend since the mid 1990s, and the sharpest rise during the 2008-09 recession. The chart thus shows that the predictability is strongest during recessions, but still weakly present during some expansionary periods. This is consistent with previous evidence of time-varying predictive power (Rapach and Zhou, 2013; Jiang et al., 2019).

The lower panel depicts the performance for one year ahead forecasts. While the picture is qualitatively the same as for the one month ahead forecasts, the basic model without break corrections shows a strong upward trend since 1995, which is not exclusively driven by recessions. The model clearly outperforms the historical average since 2009.[17] Nonetheless, the break-correction approaches still perform better, and the ordering remains unchanged: the post-break window performs best, followed by cross-validation and pooling. The distance between the latter two approaches substantially shrinks compared to the one month ahead forecasts. This can be explained by the observation that the unadjusted model performs better for $h = 12$ than for $h = 1$, and, as pooling includes many forecasts that are similar to this model, pooling also performs comparatively well.

Figure 9 compares the relative performance of cross-validation and post-break window with the performance of short interest and gold-to-platinum ratio. The upper (lower) panel shows results for one month (cumulative one year) ahead forecasts. None of the depicted models strongly outperforms the historical average between 1990 and the onset of the Great Recession. Even though the gold-to-platinum ratio performs well from 2001 to 2003 the gains are offset between 2004 to 2007. The models have in common that the positive R^2_{OS} statistics are the

[17]The range of the vertical axis is smaller for $h = 1$ than for $h = 12$. This simply reflects that averaging of 12 month ahead log equity premiums generates a substantially smoother time-series. This smoothed series can better be predicted, resulting in smaller squared forecast errors than for the one month ahead log equity premium.

result of superior predictability between 2007 to 2013. The picture is similar for one-year ahead forecasts: the curves for short interest and gold-to-platinum ratio have negative values prior to the 2008-09 crisis and strongly outperform the historical average in the subsequent years. The CSFE curves for cross-validation and post-break window are solidly above the curves for the two benchmark predictors for both horizons.

3.8 Implications for asset allocation

In this section we analyze the economic value of the improved equity premium predictions. Cenesizoglu and Timmermann (2012) show that the correlation between statistical and economic measures of forecast performance is positive but typically of low magnitude. Specifically, many models produce negative R^2_{OS} values while still providing investors with improved Sharpe ratios and gains in the certainty equivalent return. Interestingly, the reverse – positive R^2_{OS} values and negative economic gains – is observed less often. We follow Dangl and Halling (2012) and Rapach et al. (2016) and others and consider a mean-variance investor who allocates funds across the equity market portfolio and the risk-free rate. At the end of period t, the investor optimally invests a share ω_t in the risky asset:

$$\omega_t = \frac{1}{\gamma}\frac{\hat{r}_{t+1}}{\hat{\sigma}^2_{t+1}}, \tag{7}$$

where γ is the coefficient of relative risk aversion and $\hat{r}_{t+1}$ is a forecast of the equity premium.[18] Similar to Rapach et al. (2016), we estimate the variance of excess returns, $\hat{\sigma}^2_{t+1}$, as a 10-year rolling window of past data. Thus, ω_t only differs because of the different equity premium forecasts implied by the various models, $\hat{r}_{t+1}$. The realized portfolio return, r^P_{t+1}, is:

$$r^P_{t+1} = \omega_t r_{t+1} + r^f_{t+1}, \tag{8}$$

where r_{t+1} is the realized excess equity market return in period $t+1$ and r^f_{t+1} is the risk-free rate between period t and $t+1$. Furthermore, the certainty equivalent return (CER) can be calculated as:

$$\text{CER}_P = \hat{\mu}_P - \frac{\gamma}{2}\hat{\sigma}^2_P, \tag{9}$$

[18] Forecasts in this section are based on excess returns rather than log excess returns.

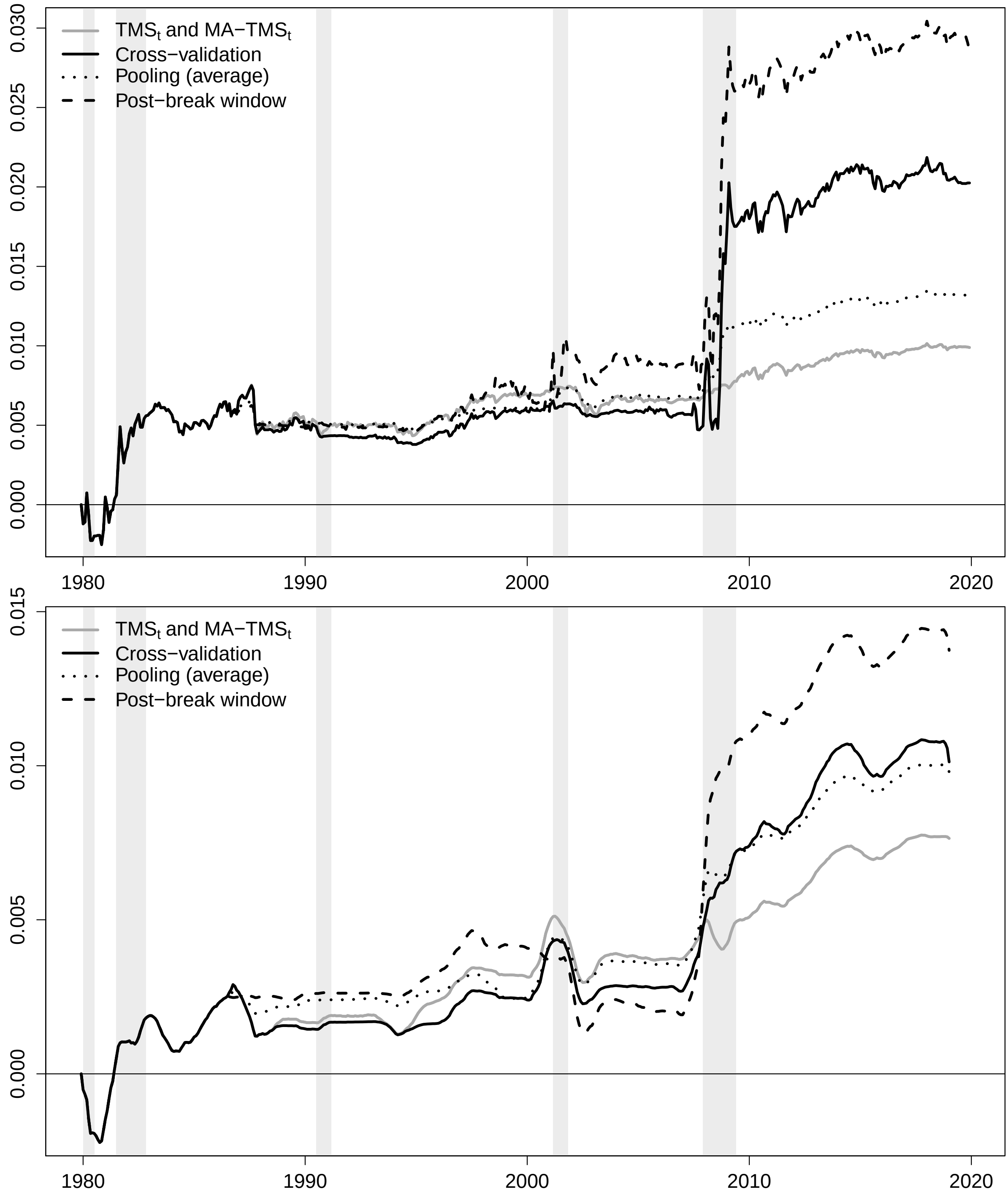

Figure 8

Out-of-sample performance over time

This figure presents the difference between cumulative squared forecast errors (CSFE) of the historical average and the CSFE of four different forecasting models. The upper panel (lower panel) shows results for forecasts of the one month ahead (cumulative one year ahead) log equity premium. Recession probabilities are derived by the probit model with TMS and MA-TMS (solid gray line), and for three methods that correct for a structural break in the term spread. These methods are cross-validation (solid black line), pooling (dotted black line), and post-break window (dashed black line). The forecasting period is 1980:1 to 2019:12 and vertical bars denote recessions.

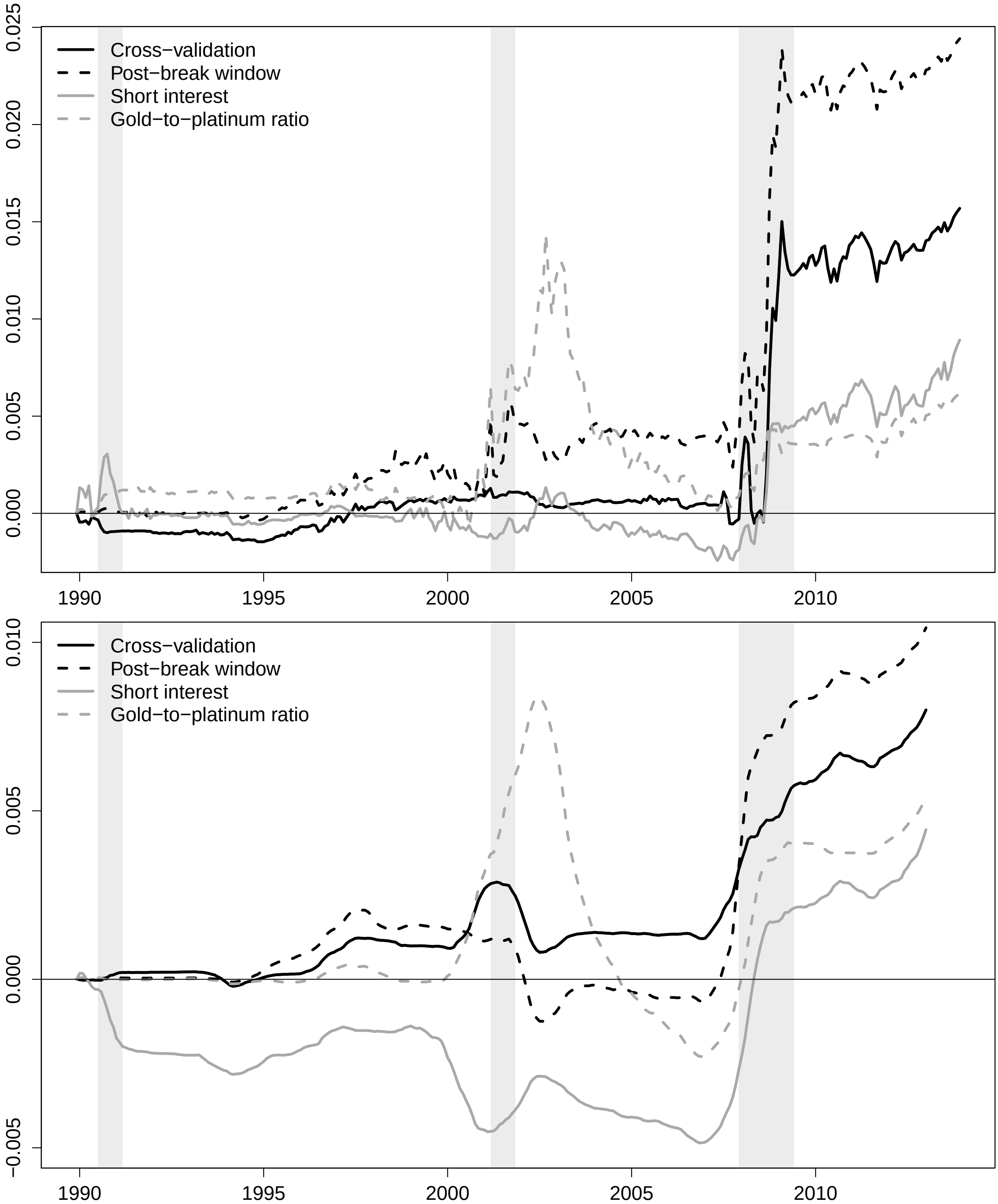

Figure 9

Out-of-sample performance relative to two benchmark predictors

This figure presents the difference between cumulative squared forecast errors (CSFE) of the historical average and the CSFE of four different forecasting models. The upper panel (lower panel) shows results for forecasts of the one month ahead (cumulative one year ahead) log equity premium. The solid (dashed) black line presents results for cross-validation (post-break window) and the solid (dashed) gray line shows the performance of short interest (gold-to-platinum ratio). The latter two predictors are suggested by Rapach et al. (2016) and Huang and Kilic (2019). The sample is 1990:1 to 2013:12 and vertical bars denote recessions.

where $\hat{\mu}_P$ and $\hat{\sigma}_P^2$ are the sample mean and variance of the portfolio over the out-of-sample period. We multiply the CER by 12 to interpret it as the annual risk-free rate that an investor would be willing to accept to not hold the risky portfolio (Chen et al., 2020). The difference in CER of two models – also known as utility gain – can be interpreted as the annual portfolio management fee that an investor would be willing to pay to have access to the alternative forecasting model (Ferreira and Santa-Clara, 2011). Additionally, we calculate the annualized Sharpe ratio (SR) to evaluate the risk-return profile of the chosen portfolio allocations.

Table 6 presents ΔCER and ΔSR for monthly re-balancing for the sample periods 1980-2019 (Panel A) and 1990-2013 (Panel B).[19] The CER and SR values for the prevailing mean model are shown in the first row of each panel. All subsequent rows then provide differences of CER and SR with respect to that benchmark. We implement a bootstrap approach similar to DeMiguel et al. (2013) to evaluate the statistical significance.[20] We consider different specifications for the coefficient of relative risk aversion (γ) and for leverage and short-selling constraints (range of ω). Over the baseline sample period from 1980-2019, the probit model with TMS and MA-TMS provides utility gains in the range of 1.4% to 2.7%. Hence, an investor would be willing to pay between 140 to 270 basis points annually - depending on risk preferences and constraints - to have access to the equity premium forecasts of this model. The gains are highest when γ is three and when leveraging and short-selling up to 50% is allowed. Nonetheless, even an investor with ω between zero and one would be willing to pay a portfolio management fee above 100 basis points annually. These gains further increase when correcting for the structural break in the mean of the term spread. We see a similar pattern as for the statistical predictability: the post-break window performs best, followed by cross-validation and the break-adjusted term spread, while gains for pooling are smallest. That said, any of these strategies outperforms a simple buy-and-hold strategy shown in the last row – often the gains more than triple or quadruple. Similarly, the annualized Sharpe ratio rises from around 0.50 for the historical average to between 0.65 to 0.85 for the break-correction methods. Panel B shows that the gains are comparable with and often substantially better than those for short interest and the gold-to-platinum ratio.

The upper panel in Figure 10 displays the optimal share in risky assets over time for the post-

[19]We show in the Online Appendix that the results remain qualitatively the same when re-balancing in 3-, 6-, 12-month intervals.

[20]We set the average block length to three months (Politis and Romano, 1994).

break window and pooling approaches relative to the historical average; γ is set to three and ω is allowed to vary between -0.5 and 1.5. Naturally, equity premium forecasts based on the historical average are continuously positive and change only slowly, hence the weights are very persistent and always above 0.5. In contrast, the portfolio weights for pooling are most often above those for the historical average in expansions and below in recessions. This is most salient in the early 1980s, where the share in risky assets runs down to -0.5 at the beginning of the recession and then quickly reverses to more than 1.0 toward the end of the recession. Similar patterns – even though less pronounced – are observed for the 2001 and 2008-09 recessions. The post-break window signals for four of five recessions in the sample to short the equity market prior to the beginning of a recession when the yield curve has been flattening for some time and to move back aggressively into the market towards the end of the recession when the yield curve is steepening again. Interestingly, both break-adjustment models signal to run down equity exposure at the end of our sample in 2019 due to rising recession probability forecasts.

The lower panel in Figure 10 shows the log cumulative wealth for five portfolio allocations: post-break window, cross-validation, pooling, historical average, and for the buy-and-hold strategy. While the buy-and-hold and historical average portfolios suffer severe losses in the Global Financial Crisis of 2008-2009, this plunge is less pronounced for pooling and even reversed for post-break window and cross-validation. Post-break window performs best, as it further signals to go short prior to the 2001 recession. An investor that would have started with $1 in 1979:12, reinvests all proceeds, and forecasts according to post-break window (cross-validation) would have earned $957 ($528) in 2019:12, compared to $90 for the buy-and-hold strategy. Overall, the results in this section show that the break-correction methods outperform the naive benchmarks, and generate utility gains that are mostly superior to those of other recently proposed predictors.

3.9 Forecasting characteristics portfolios

In this section, we extend our analysis to a rich set of equity portfolios. Specifically, we follow Huang et al. (2015) and analyze the predictive power of recession probability forecasts for portfolios sorted on different characteristics. We focus on 10 industry portfolios, 10 momentum portfolios, 10 size portfolios, and 10 book-to-market portfolios, all obtained from Kenneth French's homepage.[21] We predict the log excess return for these portfolios with the same re-

[21] The returns are value-weighted and include dividends.

Table 6
Asset allocation exercise

This table reports the annualized ΔCER and the annualized ΔSR for a mean-variance investor relative to forecasts from the historical average. The investor can invest in the S&P 500 index and the risk-free rate. Results are shown for one month ahead forecasts of the equity premium and different values for the coefficient of relative risk aversion (γ), and different restrictions on the equity weights (ω). The "Prevailing mean" shows the CER and SR values, whereas all other values denote the improvements relative to this benchmark. Panel A (Panel B) shows results for the out-of-sample period from 1980:1 to 2019:12 (1990:1 to 2013:12). *, **, *** indicate significantly improved performance relative to the prevailing mean benchmark at the 10%, 5%, and 1% significance level. The p-values are obtained by using a bootstrap approach similar to DeMiguel et al. (2013) with the average block length set to three months (Politis and Romano, 1994).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: 1980:1 to 2019:12								
	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	[−0.5, 1.5]	[−0.5, 1.5]	[0, 1.5]	[0, 1]	[−0.5, 1.5]	[−0.5, 1.5]	[0, 1.5]	[0, 1]
Prevailing mean	8.30	6.59	8.30	8.69	0.49	0.48	0.49	0.53
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	2.73***	1.49**	2.41***	1.39***	0.15***	0.14**	0.13***	0.10***
TMS_t^{break}, MA-TMS_t^{break}	4.07**	2.62**	3.33**	1.91*	0.30**	0.31**	0.25**	0.23**
Cross-validation	4.18***	2.77**	3.39***	2.29***	0.23***	0.24***	0.19***	0.18***
Pooling (average)	3.44***	2.10***	3.13***	2.12***	0.20***	0.20***	0.18***	0.16***
Post-break window	6.02***	4.20***	4.83***	3.01***	0.35***	0.35***	0.28***	0.25***
Buy-and-hold	0.84*	0.36	0.84*	0.45*	0.06***	0.07**	0.06*	0.02*
Panel B: 1990:1 to 2013:12								
	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	[−0.5, 1.5]	[−0.5, 1.5]	[0, 1.5]	[0, 1]	[−0.5, 1.5]	[−0.5, 1.5]	[0, 1.5]	[0, 1]
Prevailing mean	5.80	4.61	5.80	6.56	0.41	0.39	0.41	0.45
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	2.05**	1.24**	2.05**	0.72**	0.12***	0.13***	0.12***	0.05**
TMS_t^{break}, MA-TMS_t^{break}	5.68**	3.71**	4.71**	2.68	0.45**	0.47**	0.38**	0.32**
Cross-validation	4.44**	3.29**	3.66**	2.23*	0.26**	0.29**	0.21**	0.17*
Pooling (average)	3.56***	2.16**	3.56***	2.04***	0.21***	0.21***	0.21***	0.15***
Post-break window	8.11***	5.69**	6.59***	3.89***	0.48***	0.49***	0.39***	0.32***
Short interest	3.58*	2.40*	3.24**	1.71	0.22*	0.22*	0.20**	0.16
Gold-to-platinum ratio	4.03*	2.93*	4.11**	2.67*	0.28*	0.31*	0.30**	0.27**
Buy-and-hold	1.23*	0.23	1.23*	0.48*	0.08**	0.09*	0.08**	0.03*

cession probability forecasts as in the previous sections. Results are shown in Table 7. We find that durable, manufacturing, energy, technology, and telecom portfolios are most predictable, whereas health, utility and nondurables are not significantly predictable. This is intuitive as the former sectors are more exposed to business cycle variation.[22] Interestingly, all of the momentum portfolios – independently of the break-correction method – have positive and significant

[22]It is also consistent with Da et al. (2017), who find that electricity usage better forecasts excess returns of capital-intensive producers that are more exposed to fluctuations in the business cycle and have higher operating leverage.

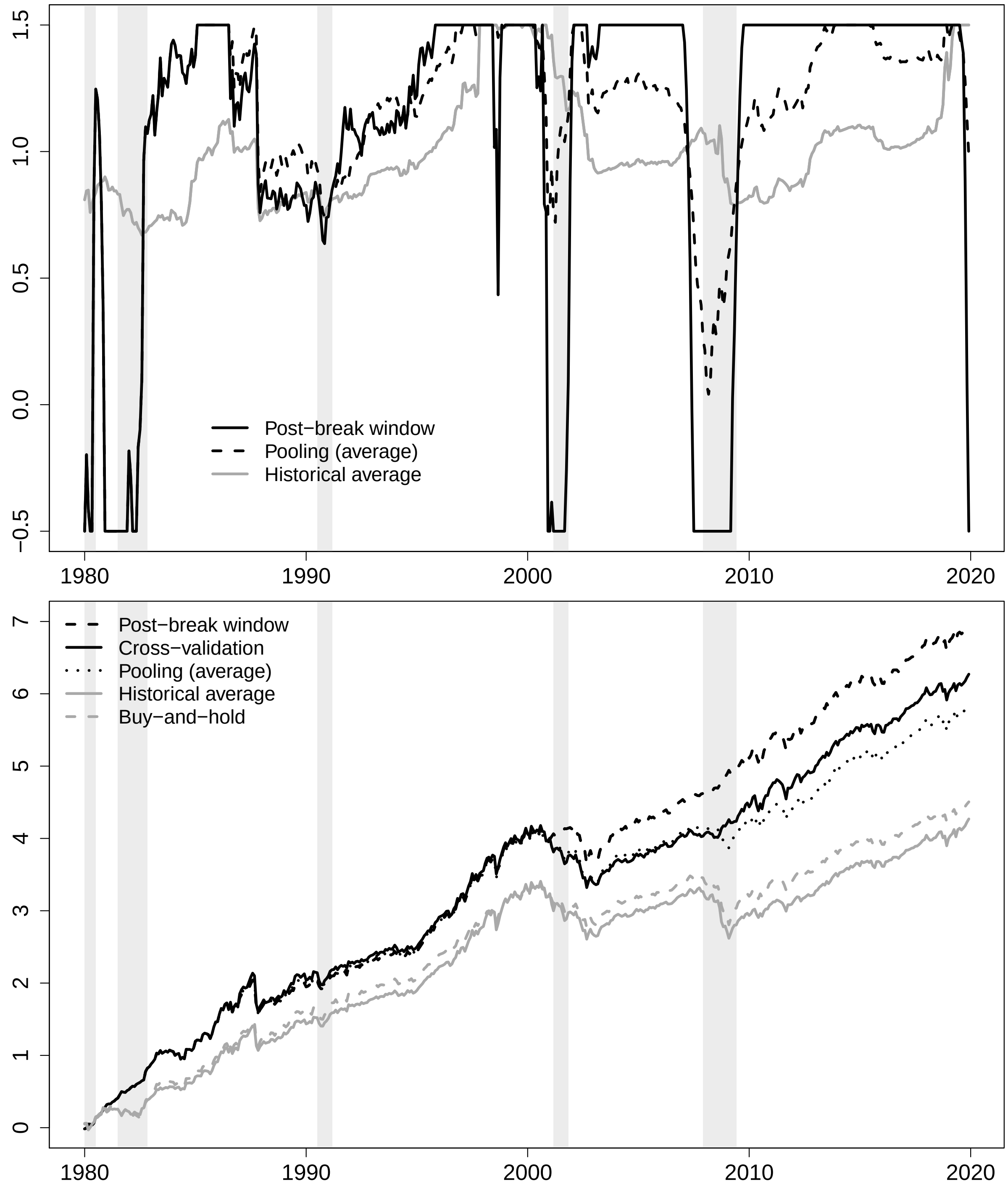

Figure 10

Share in risky assets and log cumulative wealth

The upper panel presents the optimal share in risky assets over time for three different forecasting models. The solid (dashed) black line shows the equity weights for post-break window (pooling) and the solid gray line depicts the equity weights for the historical average. The coefficient of relative risk aversion is three and ω is restricted to lie between -0.5 and 1.5. The lower panel plots the log cumulative wealth for an investor that starts with \$1 in 1979:12 and reinvests all proceeds. Results are shown for three break-correction portfolios, as well as for the historical average and the buy-and-hold strategy. Vertical bars denote recessions and the forecasting period is 1980:1 to 2019:12.

Table 7

Forecasting characteristics portfolios with recession probability forecasts

This table presents R^2_{OS} statistics (in %) for one month ahead equity premium forecasts of 10 industry portfolios (Panel A), 10 momentum portfolios (Panel B), 10 size portfolios (Panel C), and 10 book-to-market portfolios (Panel D). Forecasts are derived from the linear predictive regression model with recession probability forecasts as a predictor variable. These recession probabilities are estimated by a probit model with the term spread (TMS) and the backward-looking three-year moving average of the term spread (MA-TMS) - results are shown in column (2) - as well as for four break-correction methods - columns (3) to (6). These correction methods include cross-validation, pooling, post-break window, and forecasts from a break-adjusted term spread series (Pesaran and Timmermann, 2007; Lettau and Van Nieuwerburgh, 2008). The out-of-sample period is 1980:1 to 2019:12. *, **, *** denote significance at the 10%, 5%, and 1% significance levels according to the Clark and West (2007) MSFE-adjusted statistic.

(1) Portfolio	(2) TMS_t, $\mathrm{MA\text{-}TMS}_t$	(3) Cross-validation	(4) Pooling (average)	(5) Post-break window	(6) $\mathrm{TMS}_t^{\mathrm{break}}$, $\mathrm{MA\text{-}TMS}_t^{\mathrm{break}}$
Panel A: Industry portfolios					
Nondurable	-0.28	0.07	-0.28	-0.61	-0.07
Durable	0.16*	1.65**	0.37*	1.64**	0.86**
Manufacture	1.09**	1.57**	1.22***	2.11***	1.12**
Energy	0.51*	0.50*	0.63*	0.40*	0.46**
Technology	0.68**	1.01***	0.83**	0.68*	0.96**
Telecom	0.21	0.85**	0.67**	2.99***	1.31**
Shop	-0.68	0.24	-0.63	0.03	-0.50
Health	-0.05	0.15	-0.01	-0.28	0.03
Utility	-0.01	0.24	0.08	-0.09	0.03
Other	0.51*	2.78***	0.87**	2.63**	1.26**
Panel B: Momentum portfolios					
Loser	0.18	2.05**	0.67**	2.56**	1.33**
2	0.21*	1.85**	0.37*	1.25*	0.28
3	0.28*	1.75**	0.62**	2.22**	0.66*
4	0.46**	1.68**	0.57**	1.29**	0.48*
5	0.90**	1.43**	1.10***	1.80***	0.98**
6	0.96**	1.48**	1.29**	2.28***	1.35**
7	0.88*	1.26**	1.09**	1.99***	1.18**
8	1.17**	1.69**	1.36**	2.15***	1.21**
9	0.67*	2.15**	1.15**	3.24***	2.00***
Winner	0.71*	1.01**	1.03**	1.05**	1.12**
Panel C: Size portfolios					
Small	0.23	1.12**	0.40*	0.17	1.08*
2	0.13	0.30	0.20	-0.32	0.52*
3	0.20	0.73*	0.33	0.49	0.86**
4	0.08	0.32	0.21	0.50*	0.76**
5	0.19	0.58*	0.34	0.96**	0.75**
6	0.29	0.69*	0.45*	1.06**	0.95**
7	0.49*	1.66**	0.68**	1.81**	1.04**
8	0.37	1.41**	0.54*	1.34**	0.78**
9	0.94**	2.21**	1.19***	2.38***	1.33**
Large	1.33***	2.36***	1.69***	3.35***	1.62***
Panel D: Book-to-market portfolios					
Growth	1.32**	2.03***	1.59***	2.20***	1.24**
2	0.80**	0.94**	0.83**	1.43***	0.52*
3	1.02**	2.08***	1.18**	2.24***	1.06**
4	0.96**	1.51**	1.09**	1.77***	1.08**
5	1.40**	2.35***	1.71***	2.41***	1.59***
6	0.54*	1.69**	0.74**	1.52**	1.18**
7	0.39*	4.06**	0.84**	3.76**	1.76**
8	-0.30	1.10*	0.03	0.91*	0.32*
9	-0.23	0.74*	-0.08	0.31	0.35*
Value	0.08	1.23*	0.15	0.93*	0.75**

R^2_{OS} statistics. Additionally, there is a tendency toward higher predictive power for portfolios formed on high market equity (large stocks) and low book-to-market ratios (growth stocks).

3.10 International evidence

In this section, we provide results on equity premium predictability using recession probabilities for four additional countries: Germany, France, Canada, and the United Kingdom. We obtain data for long-term interest rates and short-term interest rates from the OECD.[23] The long-term interest rates are for ten-year government bonds, the short-term interest rates are "based on three-month money market rates where available". Data are available from 1970:1 to 2019:12.[24] The term spread is the difference between the long-term and short-term interest rates. Recession data are from the Economic Cycle Research Institute. We download end-of-month data for the DAX 30, CAC 40, S&P/TSX Composite index, and the FTSE 100 from Yahoo Finance and calculate monthly log excess return series. The data for Canada are available from 1979:7 onward, while the series for Germany, France and the U.K. start in 1988:1, 1990:4, and 1984:1, respectively. We create a sample for the log equity premium from 1980 onward by combining these series with data from Rapach et al. (2013) who provide equity premia for several developed countries starting in 1980.[25]

Figure 11 shows the evolution of the cumulative one-year ahead log equity premium ($\sum_{j=0}^{11} r_{t+j}$) for the four countries around recessions. We see a clear v-shape pattern similar to the one for the U.S. The solid gray lines depict the individual recessions, which show pronounced variation both in terms of magnitude and timing. Next, we perform a pseudo out-of-sample exercise for forecasting the log equity premium from 1990:1 to 2019:12. In line with the U.S. results, we compare the performance of three models: (1) TMS_t, (2) TMS_t and TMS_{t-6}, (3) TMS_t and MA-TMS_t. Table 8 shows the R^2_{OS} statistics for the log equity premium forecasts relative to a country-specific historical average. The values for Germany, France, and Canada are positive and statistically significant for almost all horizons when forecasting with the recession probabilities derived from the probit model with only the term spread. The R^2_{OS} statistics for the U.K. are positive only for $h = 6$ and $h = 12$. In contrast to the U.S., we do not see

[23] `https://data.oecd.org/interest/long-term-interest-rates.htm`

[24] The data for the UK short-term interest rate are downloaded from FRED. We use the three month Treasury yield series until 2017:6 (`https://fred.stlouisfed.org/series/IR3TTS01GBM156N`) and the LIBOR (`https://fred.stlouisfed.org/series/USD3MTD156N`) from 2017:7 to 2019:12 (corrected for a minor difference in sample means).

[25] The correlation of overlapping return data from Yahoo Finance and Rapach et al. (2013) is above 0.98 for Germany, France and the U.K. We can thus splice both data sets to extend the sample back until 1980. The overlapping data have minor differences in their means; we correct for this by adjusting the series to the mean of Rapach et al. (2013). The results with and without this mean correction are essentially identical.

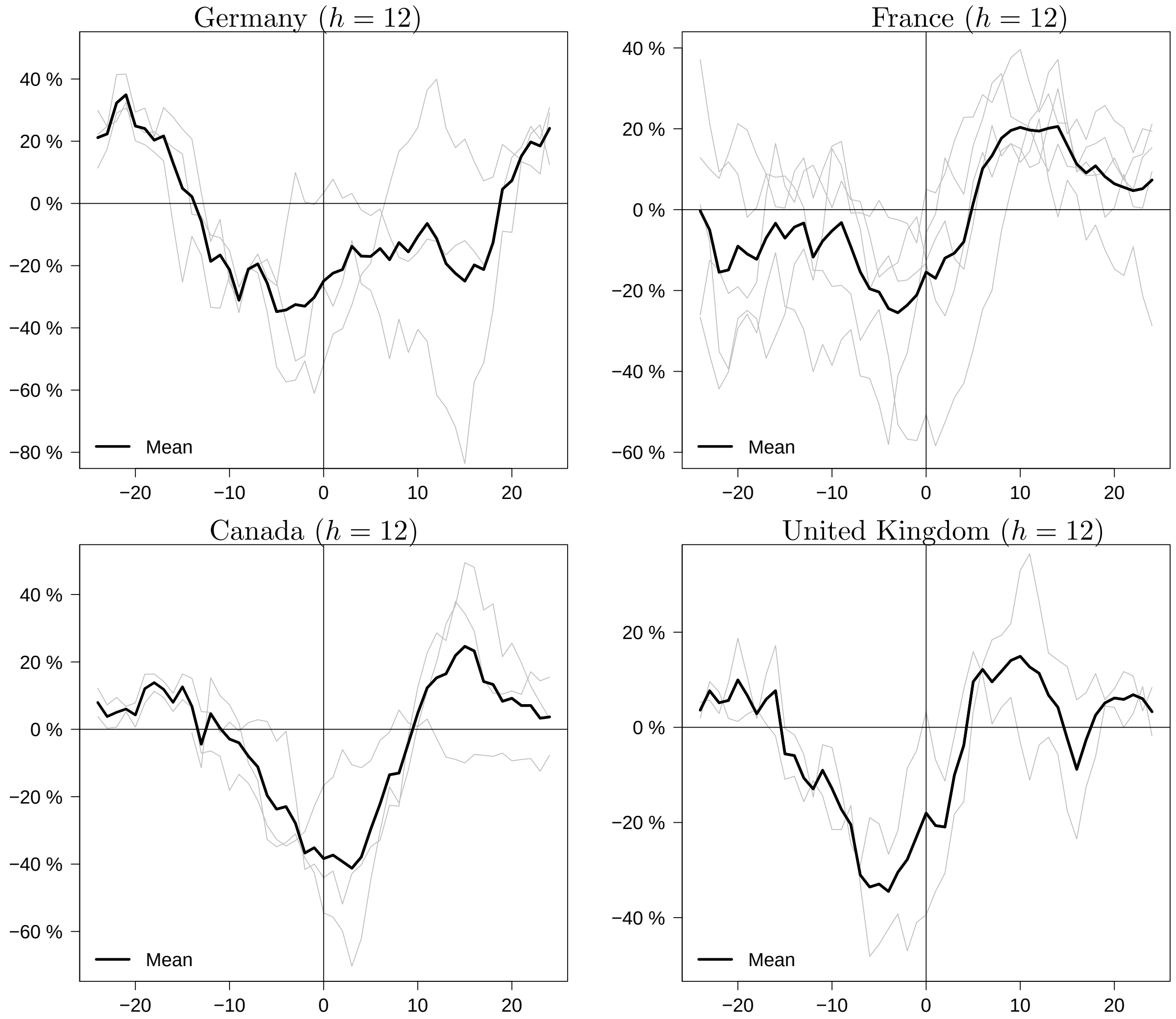

Figure 11

Log equity premium around business cycle peaks - International evidence

This figure presents the arithmetic average (solid black line) of the cumulative one-year ahead log equity premium around the recessions in the sample from 1980:2 to 2019:12. The solid gray lines depict the cumulative log equity premium of the individual recessions. The equity premium is the difference between the country-specific end-of-month index return and the country-specific short-term interest rate. The vertical axis depicts $\sum_{j=0}^{11} r_{t+j}$ for $t = -24, \ldots, -1, 0, 1, \ldots, 24$, whereby r_{t+j} is the log equity premium in month $t+j$. The horizontal axis displays the 24 months before and after a business cycle peak - with $t = 0$ referring to the first month of a recession. Recession indicators are taken from the Economic Cycle Research Institute and results are shown for Germany, France, Canada, and the U.K.

significant improvements by adding lagged and averaged term spread information to the probit model. The reason is that the recession probabilities for models (2) and (3) do not predict the beginning of recessions better than the simple model (1). Overall, however, the international data provide additional support for our main finding that recession probabilities derived from the term spread help to time the equity market. Interestingly, we further find that the recession probabilities estimated from U.S. data predict the equity premiums in those four countries as well. Table 8 shows that the U.S. recession probabilities from cross-validation with TMS_t and MA-TMS_t often perform similarly to those of the best country-specific recession probabilities.

Table 8
Out-of-sample performance: Germany, France, Canada, United Kingdom

This table reports R^2_{OS} statistics in % for the out-of-sample predictability of (cumulative) log excess returns at the h-month ahead horizon relative to forecasts from the historical average. Results are shown for Germany, France, Canada, and the U.K. Forecasts are based on the linear predictive regression model with a constant and model-implied recession probabilities as a predictor variable. The recession probability forecasts are derived by three different probit models: the first model only includes a constant and the term spread, whereas the second and third model add either the term spread lagged by six-months (TMS_{t-6}) or the three-year moving average of the term spread ($\mathrm{MA\text{-}TMS}_t$) as additional predictors. Cross-validation (U.S. data) refers to the probability forecasts from the U.S. data with TMS_t and $\mathrm{MA\text{-}TMS}_t$ and cross-validation as the break-correction method. *, **, and *** indicate significance at the 10%, 5%, and 1% levels according to the Clark and West (2007) MSFE-adjusted statistic. The null hypothesis is equal MSFE and the alternative is that the more sophisticated model has smaller MSFE than the historical average benchmark. The out-of-sample period is 1990:1 to 2019:12.

Variables in probit model	h=1	h=3	h=6	h=12
Germany:				
TMS_t	0.74**	1.93**	3.57*	7.35**
TMS_t, TMS_{t-6}	0.48	1.64*	3.57*	7.90**
TMS_t, $\mathrm{MA\text{-}TMS}_t$	0.54*	1.44*	2.84*	7.77**
Cross-validation (U.S. data)	0.80*	2.14*	3.03*	9.63**
France:				
TMS_t	0.16	2.25**	3.90**	6.41***
TMS_t, TMS_{t-6}	0.53*	1.07**	3.46**	5.69***
TMS_t, $\mathrm{MA\text{-}TMS}_t$	0.10	1.44**	2.75**	-2.49
Cross-validation (U.S. data)	1.12*	3.54**	5.01*	8.90*
Canada:				
TMS_t	2.32***	4.66***	8.53***	15.47***
TMS_t, TMS_{t-6}	2.24***	4.87***	7.45***	11.28***
TMS_t, $\mathrm{MA\text{-}TMS}_t$	-0.43	0.25	3.22**	7.22**
Cross-validation (U.S. data)	1.47**	5.29**	10.87**	12.58***
United Kingdom:				
TMS_t	-0.44	-0.05	0.32	4.11**
TMS_t, TMS_{t-6}	0.19	2.12**	2.92**	5.45***
TMS_t, $\mathrm{MA\text{-}TMS}_t$	-0.10	0.10*	0.08*	0.57
Cross-validation (U.S. data)	0.70*	2.08**	2.53*	11.82**

3.11 Dissecting the sources of predictability

In this section, we aim to shed light on the question whether the predictability of excess equity returns using recession probabilities derives from the cash flow or the discount rate channel. To this end, we follow Campbell (1991) and Campbell and Ammer (1993) and apply a VAR decomposition to U.S. equity market returns. According to Campbell and Shiller (1988) the log return r_{t+1} can be rewritten as:

$$r_{t+1} = E_t[r_{t+1}] + N_{t+1}^{CF} - N_{t+1}^{DR}, \tag{10}$$

where $E_t[r_{t+1}]$ is the conditional expectation of the log return at time t, and N_{t+1}^{CF} and N_{t+1}^{DR} represent cash flow and discount rate news, respectively. We use a first-order VAR model to estimate the three components:

$$z_{t+1} = \Gamma z_t + u_{t+1}. \tag{11}$$

Here, z_t is a k-dimensional vector of state variables, Γ is a $k \times k$ matrix of parameters, and u_{t+1} is k-dimensional vector of innovations. We demean the variables in the VAR model and, therefore, can omit an intercept. The cash flow news and discount rate news can be estimated as:

$$N^{CF}_{t+1} = (e1' + e1'\lambda)u_{t+1}, \tag{12}$$

$$N^{DR}_{t+1} = e1'\lambda u_{t+1}, \tag{13}$$

where $e1$ is a k-vector with a one in the first cell and a zero in all remaining cells, and $\lambda \equiv \rho\Gamma(\mathbf{I} - \rho\Gamma)^{-1}$. We set $\rho = 0.95^{\frac{1}{12}}$, which corresponds to approximately 5% consumption of total wealth per year (Campbell and Vuolteenaho, 2004; Maio, 2013).

The state vector is assumed to capture the dynamics of log equity market returns. However, the estimated news components may be sensitive to the variables included in the VAR (Chen and Zhao, 2009). Therefore, we follow Rapach et al. (2016) and estimate a series of trivariate VAR models including the log return series, the dividend-price ratio, and one of the 14 commonly used Welch and Goyal (2008) predictors at a time.[26] Additionally, we show results for a VAR model, where we add the first three principal components of the predictors of Welch and Goyal (2008) to the log return and the log dividend-price ratio (Rapach et al., 2016). We use the OLS estimates of Γ and u_{t+1} to calculate $\hat{E}_t[r_{t+1}]$, $\hat{N}^{CF}_{t+1}$, and $\hat{N}^{DR}_{t+1}$. Then, we run the following regressions to estimate the effect of our model-implied recession probabilities on the components:

$$\hat{E}_t[r_{t+1}] = \alpha^E + \beta^E \times \hat{p}_{t+1} + \epsilon^E_{t+1} \tag{14}$$

$$\hat{N}^{CF}_{t+1} = \alpha^{CF} + \beta^{CF} \times \hat{p}_{t+1} + \epsilon^{CF}_{t+1} \tag{15}$$

$$\hat{N}^{DR}_{t+1} = \alpha^{DR} + \beta^{DR} \times \hat{p}_{t+1} + \epsilon^{DR}_{t+1} \tag{16}$$

[26]The 14 predictors are the log dividend-price ratio (DP), log dividend yield (DY), log earnings-price ratio (EP), log dividend-payout ratio (DE), excess stock return volatility (RVOL), book-to-market ratio (BM), net equity expansion (NTIS), Treasury bill rate (TBL), long-term yield (LTY), long-term return (LTR), term spread (TMS), default yield spread (DFY), default return spread (DFR), and inflation (INFL).

While the return decomposition is based on the VAR estimates over the full sample from 1951:3 to 2019:12, we estimate the recession probabilities $\hat{p}_{t+1}$ with recursively expanding information sets to rule out any look-ahead bias. Table 9 presents the coefficients $\hat{\beta}^{E}$, $\hat{\beta}^{CF}$, and $\hat{\beta}^{DR}$ estimated using data from 1980:1 to 2019:12 which corresponds to our out-of-sample forecasting period. As predictor we use the recession probabilities estimated by cross-validation, but the results are essentially unchanged for the alternative break-correction methods. The estimates show that the recession probability has significant predictive power for both the cash flow and the discount rate news component in most VAR specifications. Specifically, increased recession probabilities signal low future dividends and high discount rates. The returns recover as the recession unfolds and recession probabilities approach zero.

Our results are consistent with limited information processing of investors to macroeconomic news. The term spread is readily available to investors and stock markets should be forward-looking. This implies that news about the business-cycle embedded in the term spread should be fully incorporated in stock prices. In contrast, our results show that recession probabilities using the term spread as input provide valuable information to time the market. This is in line with studies documenting information rigidities and sluggishness in forecast revisions. Rudebusch and Williams (2009) document that a simple probit model that is based on the yield curve performs substantially better than professional forecasters in predicting recessions - even though evidence in favor of the former model has been known for years. Coibion and Gorodnichenko (2015) show for inflation forecasts of professional forecasters that the null hypothesis of consistency with full-information rational expectations models can primarily be rejected due to deviations from full-information. Their findings suggest that agents on average update their information sets every six to seven months. An alternative interpretation of these results is that agents put a relatively small weight on incoming information. Consistent with such a view, Loungani et al. (2013) find evidence of information rigidities in growth forecasts for a cross-section of countries, with an average period of 4 to 6 months until new information is fully incorporated.

Information rigidities are not restricted to professional forecasters, however. Bouchaud et al. (2019) document that analysts who are less experienced and follow more industries have stickier beliefs - consistent with information processing constraints (Peng, 2005). Several articles

Table 9
Predictive regressions for stock market return components

This table presents slope coefficients for regressions of log stock market return components on model-implied recession probabilities. The stock market return is decomposed into the conditional return expectation ($\hat{E}_t[r_{t+1}]$), a cash flow news component ($\hat{N}^{CF}_{t+1}$), and a discount rate news component ($\hat{N}^{DR}_{t+1}$). The decomposition is based on the VAR approach of Campbell (1991) and Campbell and Ammer (1993) and includes as states the variables in columns (1) and (5). The log return on the S&P 500 index (r) and the log dividend-price ratio (DP) are included in each of the VAR models. PC denotes the first three principal components of the 14 popular predictors of Welch and Goyal (2008). The three return components are separately regressed on a constant and model-implied recession probabilities from cross-validation. The probabilities are recursively estimated and identical to those used in the out-of-sample exercises in the previous sections. The regressions are based on data from 1980:1 to 2019:12. The t-statistics for the slope coefficients are shown in the brackets below and the standard errors are HAC-robust (Andrews, 1991). *, **, *** denote significance at the 10%, 5%, and 1% significance levels.

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
VAR variables	$\hat{\beta}^E$	$\hat{\beta}^{CF}$	$\hat{\beta}^{DR}$	VAR variables	$\hat{\beta}^E$	$\hat{\beta}^{CF}$	$\hat{\beta}^{DR}$
r, DP	0.04	-0.17**	0.30***	r, DP, LTY	0.00	-0.17**	0.26**
	[0.65]	[-2.50]	[2.76]		[0.10]	[-2.59]	[2.36]
r, DP, DY	0.03	-0.17**	0.29***	r, DP, LTR	0.06*	-0.18***	0.32***
	[0.58]	[-2.49]	[2.68]		[1.77]	[-2.62]	[2.79]
r, DP, EP	0.04	-0.76**	-0.29	r, DP, TMS	-0.04	-0.15**	0.24**
	[0.60]	[-2.02]	[-1.16]		[-0.85]	[-2.16]	[2.10]
r, DP, DE	0.04	-0.76**	-0.29	r, DP, DFY	0.07	-0.22***	0.28***
	[0.60]	[-2.02]	[-1.16]		[1.09]	[-2.68]	[2.73]
r, DP, RVOL	0.08	-0.20**	0.31***	r, DP, DFR	0.03	-0.17**	0.30***
	[1.14]	[-2.55]	[3.02]		[0.53]	[-2.49]	[2.71]
r, DP, BM	-0.01	-0.19**	0.23**	r, DP, INFL	-0.00	-0.17**	0.26**
	[-0.27]	[-2.54]	[2.05]		[-0.03]	[-2.51]	[2.07]
r, DP, NTIS	0.07	-0.15	0.35***	r, DP, PC	-0.00	-0.30***	0.13
	[1.61]	[-1.49]	[2.77]		[-0.05]	[-3.09]	[1.17]
r, DP, TBL	-0.06*	-0.14*	0.24**				
	[-1.86]	[-1.95]	[2.00]				

show at the firm level that even though information is publicly available, it is often not fully incorporated into stock prices, resulting in delayed price reaction to news (see, e.g., Bernard and Thomas (1989); Huberman and Regev (2001); Hirshleifer et al. (2009)). Chen et al. (2020) find that aggregate investor attention - extracted from 12 different attention measures - has predictive power for the aggregate stock market that cannot be explained by macroeconomic fundamentals. Moreover, Gómez-Cram (2021) shows that news about impending recessions are incorporated into stock prices only with some delay. Analysts' earnings forecasts are too optimistic when the excess return forecast is negative.

4 Robustness

In this section we show the robustness of our results along several dimensions. The results are provided in the Online Appendix.

First, we provide R^2_{OS} statistics for different averaging windows to construct MA-TMS and show that the results are robust to the exact choice of moving average. In the main text we have focused on a moving average of three years for the US. The long-horizon forecasts for $h = 3, 6, 12$ are statistically significant also for moving averages between one to five years, whereas the one-month ahead forecasts perform best for averages between three to five years.

Second, we present out-of-sample R^2_{OS} values for log raw returns (without subtracting the short rate). The previous results carry over to raw returns, however, the predictive power decreases slightly for cumulative six- and twelve-month ahead forecasts.

Third, we provide the certainty equivalent returns and Sharpe ratios when correcting for proportional transaction costs of 50 basis points per transaction (Balduzzi and Lynch, 1999). The gains relative to the historical average and the buy-and-hold strategy remain sizable.

5 Conclusion

Excess equity market returns are negative around business cycle peaks, and strongly recover during recessions. In this paper, we have shown that probit models which predict recessions using the term spread also have strong predictive power for the U.S. equity premium. The gains are statistically and economically significant and further improve when adding a backward-looking moving average of the term spread to the probit model. Equity premium forecasts based on recession probabilities correctly anticipate negative market returns heading into recessions and positive returns in expansions. We provide evidence for a structural break in the mean of the term spread in 1982. When correcting for this structural break, both recession and equity premium forecasts further improve. Our paper thus provides further evidence for the strong link between the business cycle and the equity premium. More specifically, it shows that information in the yield curve can be used to time the equity market.

References

Adrian, T., A. Estrella, and H. S. Shin, "Risk-taking channel of monetary policy," *Financial Management*, 2019, *48*, 725–738.

Andreasen, M. M., T. Engsted, S. V. Møller, and M. Sander, "The Yield Spread and Bond Return Predictability in Expansions and Recessions," *Review of Financial Studies*, 2021, *34*, 2773–2812.

Andrews, D. W. K., "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 1991, *59* (3), 817–858.

_ , "Tests for Parameter Instability and Structural Change With Unknown Change Point," *Econometrica*, 1993, *61*, 821–856.

_ **and W. Ploberger**, "Optimal Tests when a Nuisance Parameter is Present Only Under the Alternative," *Econometrica*, 1994, *62*, 1383–1414.

Bai, J., "Estimation of a change point in multiple regression models," *Review of Economics and Statistics*, 1997, *79*, 551–563.

_ **and P. Perron**, "Estimating and Testing Linear Models with Multiple Structural Changes," *Econometrica*, 1998, *66* (1), 47–78.

_ **and** _ , "Computation and analysis of multiple structural change models," *Journal of Applied Econometrics*, 2003, *18* (1), 1–22.

Balduzzi, P. and A. W. Lynch, "Transaction costs and predictability: some utility cost calculations," *Journal of Financial Economics*, 1999, *52*, 47–78.

Berge, T. J. and Ò. Jordà, "Evaluating the Classification of Economic Activity into Recessions and Expansions," *American Economic Journal: Macroeconomics*, 2011, *3* (2), 246–277.

Bernard, H. and S. Gerlach, "Does the Term Structure Predict Recessions? The International Evidence," *International Journal of Finance and Economics*, 1998, *3*, 195–215.

Bernard, V. L. and J. K. Thomas, "Post-Earnings-Announcement Drift: Delayed Price Response or Risk Premium?," *Journal of Accounting Research*, 1989, *27*, 1–36.

Bouallègue, Z. B., L. Magnusson, T. Haiden, and D. S. Richardson, "Monitoring trends in ensemble forecast performance focusing on surface variables and high-impact events," *Quarterly Journal of the Royal Meteorological Society*, 2019, *145*, 1741–1755.

Bouchaud, J.-P., P. Krüger, A. Landier, and D. Thesmar, "Sticky Expectations and the Profitability Anomaly," *Journal of Finance*, 2019, *74* (2), 639–674.

Campbell, J. Y., "A Variance Decomposition for Stock Returns," *Economic Journal*, 1991, *101*, 157–179.

_ **and J. Ammer**, "What Moves the Stock and Bond Market? A Variance Decomposition for Long-Term Asset Returns," *Journal of Finance*, 1993, *48*, 3–37.

_ **and R. J. Shiller**, "The Dividend-Price Ratio and Expectations of Future Dividends and Discount Factors," *Review of Financial Studies*, 1988, *1* (3), 195–228.

_ **and S. Thompson**, "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?," *Review of Financial Studies*, 2008, *21*, 1509–1531.

— **and T. Vuolteenaho**, "Bad Beta, Good Beta," *American Economic Review*, 2004, *94* (5), 1249–1275.

Campbell, S. D. and F. X. Diebold, "Stock Returns and Expected Business Conditions: Half a Century of Direct Evidence," *Journal of Business and Economic Statistics*, 2009, *27*, 266–278.

Carlson, M. A. and D. C. Wheelock, "Navigating Constraints: The Evolution of Federal Reserve Monetary Policy, 1935-1959," Federal Reserve Bank of St. Louis, Working Papers 2014-13 2014.

Cenesizoglu, T. and A. Timmermann, "Do return prediction models add economic value?," *Journal of Banking & Finance*, 2012, *36*, 2974–2987.

Chauvet, M. and S. Potter, "Predicting a recession: evidence from the yield curve in the presence of structural breaks," *Economics Letters*, 2002, *77*, 245–253.

— **and** — , "Business cycle monitoring with structural changes," *International Journal of Forecasting*, 2010, *26*, 777–793.

Chen, J., G. Tang, J. Yao, and G. Zhou, "Investor Attention and Stock Returns," Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3194387 2020.

Chen, L. and X. Zhao, "Return Decomposition," *Review of Financial Studies*, 2009, *22* (12), 5213–5249.

Chen, S.-S., Y.-H. Chou, and C.-Y. Yen, "Predicting US recessions with stock market illiquidity," *B.E. Journal of Macroeconomics*, 2016, *16*, 93–123.

Chow, G. C., "Tests of Equality Between Sets of Coefficients in Two Linear Regressions," *Econometrica*, 1960, *28*, 591–605.

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

Clark, T. E. and K. D. West, "Approximately normal tests for equal predictive accuracy in nested models," *Journal of Econometrics*, 2007, *138*, 291–311.

Cochrane, J. H., "Financial Markets and the Real Economy," in R. Mehra, ed., *Handbook of the Equity Risk Premium*, Elsevier, 2007, chapter 7, pp. 237–325.

Coibion, O. and Y. Gorodnichenko, "Information Rigidity and the Expectations Formation Process: A Simple Framework and New Facts," *American Economic Review*, 2015, *105* (8), 2644–2678.

Cooper, I. and R. Priestley, "Time-Varying Risk Premiums and the Output Gap," *Review of Financial Studies*, 2009, *22*, 2801–2833.

Da, Z., D. Huang, and H. Yun, "Industrial Electricity Usage and Stock Returns," *Journal of Financial and Quantitative Analysis*, 2017, *52* (1), 37–69.

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

DeMiguel, V., Y. Plyakha, R. Uppal, and G. Vilkov, "Improving Portfolio Selection Using Option-Implied Volatility and Skewness," *Journal of Financial and Quantitative Analysis*, 2013, *48* (6), 1813–1845.

Diebold, F. X. and G. D. Rudebusch, "Scoring the Leading Indicators," *Journal of Business*, 1989, *62*, 369–391.

_ **and R. S. Mariano**, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 1995, *13*, 253–263.

Eichengreen, B. and P. Garber, "Before the Accord: U.S. Monetary-Financial Policy, 1945-1951," in "Financial Markets and Financial Crises," National Bureau of Economic Research, Inc, 1991, pp. 175–206.

Estrella, A., "A New Measure of Fit for Equations With Dichotomous Dependent Variables," *Journal of Business and Economic Statistics*, 1998, *16*, 198–205.

_ , "Why does the yield curve predict output and inflation?," *Economic Journal*, July 2005, pp. 722–744.

_ **, A. P. Rodrigues, and S. Schich**, "How stable is the predictive power of the yield curve? Evidence from Germany and the United States," *Review of Economics and Statistics*, 2003, *85*, 629–644.

_ **and F. S. Mishkin**, "Predicting U.S. recessions: financial variables as leading indicators," *Review of Economics and Statistics*, 1998, *80* (1), 45–61.

_ **and G. A. Hardouvelis**, "The Term Structure as a Predictor of Real Economic Activity," *Journal of Finance*, 1991, *46*, 555–576.

Fama, E. F. and K. R. French, "Dividend yields and expected stock returns," *Journal of Financial Economics*, 1988, *22*, 3–25.

_ **and** _ , "Business conditions and expected returns on stocks and bonds," *Journal of Financial Economics*, 1989, *25* (1), 23–49.

Ferreira, M. A. and P. Santa-Clara, "Forecasting stock market returns: The sum of the parts is more than the whole," *Journal of Financial Economics*, 2011, *100*, 514–537.

Ferson, W. E. and C. R. Harvey, "The Variation of Economic Risk Premiums," *Journal of Political Economy*, 1991, *99*, 385–415.

Fornaro, P., "Forecasting US Recessions with a Large Set of Predictors," *Journal of Forecasting*, 2016, *35*, 477–492.

Galvão, A. B., "Structural Break Threshold VARs for Predicting US Recessions using the Spread," *Journal of Applied Econometrics*, 2006, *21*, 463–487.

_ **and M. T. Owyang**, "Forecasting Low Frequency Macroeconomic Events with High Frequency Data," FRB St. Louis Working Paper No. 2020-028 2020.

Gómez-Cram, R., "Late to Recessions: Stocks and the Business Cycle," Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3671346 2021.

Gu, S., B. Kelly, and D. Xiu, "Empirical Asset Pricing via Machine Learning," *Review of Financial Studies*, 2020, *33*, 2223–2273.

Hamilton, J. D., "Calling recessions in real time," *International Journal of Forecasting*, 2011, *27*, 1006–1026.

Hanley, J. A. and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, 1982, *143*, 29–36.

_ **and** _ , "A Method of Comparing the Areas under Receiver Operating Characteristic Curves Derived from the Same Cases," *Radiology*, 1983, *148*, 839–843.

Hansen, B. E., "Approximate asymptotic p-values for structural change tests," *Journal of Business and Economic Statistics*, 1997, *15*, 60–67.

_ , "The New Econometrics of Structural Change: Dating Breaks in U.S. Labor Productivity," *Journal of Economic Perspectives*, 2001, *15* (4), 117–128.

Harvey, D. I., S. J. Leybourne, and P. Newbold, "Tests for Forecast Encompassing," *Journal of Business and Economic Statistics*, 1998, *16*, 254–259.

Hirshleifer, D., S. S. Lim, and S. H. Teoh, "Driven to Distraction: Extraneous Events and Underreaction to Earnings News," *Journal of Finance*, 2009, *64* (5), 2289–2325.

Huang, D. and M. Kilic, "Gold, platinum, and expected stock returns," *Journal of Financial Economics*, 2019, *132*, 50–75.

_ , F. Jiang, J. Tu, and G. Zhou, "Investor Sentiment Aligned: A Powerful Predictor of Stock Returns," *Review of Financial Studies*, 2015, *28* (3), 791–837.

_ , _ , _ , and _ , "Forecasting stock returns in good and bad times: The role of market states," Available at SSRN: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2188989 2017.

Huberman, G. and T. Regev, "Contagious Speculation and a Cure for Cancer: A Nonevent that Made Stock Prices Soar," *Journal of Finance*, 2001, *56* (1), 387–396.

Ibbotson, R. G. and R. A. Sinquefield, "Stocks, Bonds, Bills, and Inflation: Year-by-Year Historical Returns (1926-1974)," *Journal of Business*, 1976, *49*, 11–47.

Jacobsen, B., B. R. Marshall, and N. Visaltanachoti, "Stock Market Predictability and Industrial Metal Returns," *Management Science*, 2019, *65*, 2947–3448.

Jiang, F., J. Lee, X. Martin, and G. Zhou, "Manager sentiment and stock returns," *Journal of Financial Economics*, 2019, *132*, 126–149.

Johansson, P., "Predicting Recession Probabilities Using the Slope of the Yield Curve," FEDS Notes. Washington: Board of Governors of the Federal Reserve System, March 1 2018.

Kauppi, H. and P. Saikkonen, "Predicting U.S. recessions with dynamic binary response models," *Review of Economics and Statistics*, 2008, *90* (4), 777–791.

Kim, C.-J. and C. R. Nelson, "Has the U.S. Economy Become More Stable? A Bayesian Approach based on a Markov-Switching Model of the Business Cycle," *Review of Economics and Statistics*, 1999, *81*, 608–616.

Lacker, J. M., "Introduction to Special Issue on the 50th anniversary of the Treasury-Federal Reserve Accord," Federal Reserve Bank of Richmond *Economic Quarterly*, Vol. 87, 2001.

Lettau, M. and S. Ludvigson, "Consumption, Aggregate Wealth, and Expected Stock Returns," *Journal of Finance*, 2001, *56*, 815–849.

_ **and S. Van Nieuwerburgh**, "Reconciling the Return Predictability Evidence," *Review of Financial Studies*, 2008, *21* (4), 1607–1652.

Liu, W. and E. Moench, "What predicts US recessions?," *International Journal of Forecasting*, 2016, *32*, 1138–1150.

Loungani, P., H. Stekler, and N. Tamirisa, "Information rigidity in growth forecasts: Some cross-country evidence," *International Journal of Forecasting*, 2013, *29*, 605–621.

Lustig, H. and A. Verdelhan, "Business cycle variation in the risk-return trade-off," *Journal of Monetary Economics*, 2012, *58*, 535–549.

Maio, P., "Return decomposition and the Intertemporal CAPM," *Journal of Banking and Finance*, 2013, *37*, 4958–4972.

McConnell, M. M. and G. Perez-Quiros, "Output Fluctuations in the United States: What Has Changed Since the Early 1980's?," *American Economic Review*, 2000, *90*, 1464–1476.

Mihai, M. M., "Do credit booms predict US recessions?," *Journal of Forecasting*, 2020, *39*, 887–910.

Neely, C. J., D. E. Rapach, J. Tu, and G. Zhou, "Forecasting the Equity Risk Premium: The Role of Technical Indicators," *Management Science*, 2014, *60* (7), 1772–1791.

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

_ and _ , "Automatic Lag Selection in Covariance Matrix Estimation," *Review of Economic Studies*, 1994, *61*, 631–653.

Nyberg, H., "Predicting bear and bull stock markets with dynamic binary time series models," *Journal of Banking and Finance*, 2013, *37*, 3351–3363.

Peng, L., "Learning with Information Capacity Constraints," *Journal of Financial and Quantitative Analysis*, 2005, *40* (2), 307–329.

Pesaran, M. H. and A. Timmermann, "Selection of estimation window in the presence of breaks," *Journal of Econometrics*, 2007, (137), 134–161.

Pettenuzzo, D., A. Timmermann, and R. Valkanov, "Forecasting stock returns under economic constraints," *Journal of Financial Economics*, 2014, *114*, 517–553.

_ and _ , "Forecasting Macroeconomic Variables Under Model Instability," *Journal of Business and Economic Statistics*, 2017, *35*, 183–201.

Politis, D. N. and J. P. Romano, "The Stationary Bootstrap," *Journal of the American Statistical Association*, 1994, *89* (428), 1303–1313.

Ponka, H., "The Role of Credit in Predicting US Recessions," *Journal of Forecasting*, 2017, *36*, 469–482.

Pyun, S., "Variance risk in aggregate stock returns and time-varying return predictability," *Journal of Financial Economics*, 2019, *132*, 150–174.

Rapach, D. E. and G. Zhou, "Forecasting Stock Returns," in G. Elliott and A. Timmermann, eds., *Handbook of Economic Forecasting*, Vol. 2A, Elsevier, Amsterdam, 2013, pp. 329–383.

_ , J. K. Strauss, and G. Zhou, "Out-of-Sample Equity Premium Prediction: Combination Forecasts and Links to the Real Economy," *Review of Financial Studies*, 2010, *23* (2), 821–862.

_ , _ , and _ , "International Stock Return Predictability: What Is the Role of the United States?," *Journal of Finance*, 2013, *68* (4), 1633–1662.

— **, M. C. Ringgenberg, and G. Zhou**, “Short interest and aggregate stock returns,” *Journal of Financial Economics*, 2016, *121*, 46–65.

Rudebusch, G. D. and J. C. Williams, “Forecasting Recessions: The Puzzle of the Enduring Power of the Yield Curve,” *Journal of Business and Economic Statistics*, 2009, *27*, 492–503.

— **, B. P. Sack, and E. T. Swanson**, “Macroeconomic Implications of Changes in the Term Premium,” Federal Reserve Bank of St. Louis *Review* 2007.

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

Smith, S. C. and A. Timmermann, “Break Risk,” *Review of Financial Studies*, 2021, *34*, 2045–2100.

Welch, I. and A. Goyal, “A Comprehensive Look at The Empirical Performance of Equity Premium Prediction,” *Review of Financial Studies*, 2008, *21*, 1455–1508.

Wheelock, D. C. and M. E. Wohar, “Can the Term Spread Predict Output Growth and Recessions? A Survey of the Literature,” Federal Reserve Bank of St. Louis Review, September/October, Part 1 2009.

Wright, J. H., “The Yield Curve and Predicting Recessions,” FEDS Working Paper No. 2006-7 2006.

Online Appendix for

“Equity premium predictability over the business cycle”

Not for Publication

A Forecast evaluation

The quadratic probability score (QPS), the logarithm score (LS), and the diagonal elementary score (DES) are given by:

$$\mathrm{QPS} = \frac{2}{\tau} \sum_{j=M+1}^{T-h+1} \left[Y_{j:j+h-1} - \hat{p}_{j:j+h-1}\right]^2 \tag{A.1}$$

$$\mathrm{LS} = \frac{1}{\tau} \sum_{j=M+1}^{T-h+1} -\ln\left[\left|1 - Y_{j:j+h-1} - \hat{p}_{j:j+h-1}\right|\right] \tag{A.2}$$

$$\mathrm{DES} = \frac{1}{\tau} \sum_{j=M+1}^{T-h+1} \pi \mathbb{I}\left[\hat{p}_{j:j+h-1} > \pi\right](1 - Y_{j:j+h-1}) + (1-\pi)\mathbb{I}\left[\hat{p}_{j:j+h-1} \leq \pi\right] Y_{j:j+h-1} \tag{A.3}$$

where $|\cdot|$ refers to the absolute value, $\mathbb{I}[\,\cdot\,]$ is an indicator function, π equals the unconditional probability of $Y_{j:j+h-1} = 1$ in the evaluation period, and $\tau = T - M - h + 1$. QPS assigns values between 0 and 2 and can be seen as a counterpart to the mean squared forecast error, and LS ranges between 0 to ∞.

The out-of-sample pseudo R^2 and the area under the receiver operating characteristic (AUROC) curve are:

$$\text{pseudo } R^2 = 1 - \left[\frac{\ln L_u}{\ln L_c}\right]^{\frac{-2\ln L_c}{\tau}}, \tag{A.4}$$

$$\mathrm{AUROC} = \frac{1}{n_0 n_1} \sum_{i=1}^{n_0} \sum_{j=1}^{n_1} A(\hat{p}_{i:i+h-1}, \hat{p}_{j:j+h-1} | Y_{j:j+h-1} > Y_{i:i+h-1}), \tag{A.5}$$

where $\ln L_u$ ($\ln L_c$) denotes the unconstrained (constrained) log likelihood with out-of-sample forecasts from the probit model (Estrella, 1998; Chen et al., 2016). The number of expansions (recessions) in the out-of-sample period is denoted by n_0 (n_1) and A assigns values similar to the Mann-Whitney U statistic (Bouallègue et al., 2019).[1]

[1] More precisely, we have $\ln L_u = \sum_{j=M+1}^{T-h+1} Y_{j:j+h-1} \ln\left[\hat{p}_{j:j+h-1}\right] + (1 - Y_{j:j+h-1}) \ln\left[1 - \hat{p}_{j:j+h-1}\right]$ and $\ln L_c = \sum_{j=M+1}^{T-h+1} Y_{j:j+h-1} \ln\left[\bar{p}_{j:j+h-1}\right] + (1 - Y_{j:j+h-1}) \ln\left[1 - \bar{p}_{j:j+h-1}\right]$, where $\bar{p}_{j:j+h-1}$ denotes recession forecasts from a model with only a constant. A assigns values as follows:

$$A(\hat{p}_{i:i+h-1}, \hat{p}_{j:j+h-1} | Y_{j:j+h-1} > Y_{i:i+h-1}) = \begin{cases} 0 & \text{if } \hat{p}_{j:j+h-1} < \hat{p}_{i:i+h-1},\ Y_{j:j+h-1} > Y_{i:i+h-1}, \\ 0.5 & \text{if } \hat{p}_{j:j+h-1} = \hat{p}_{i:i+h-1},\ Y_{j:j+h-1} > Y_{i:i+h-1}, \\ 1 & \text{if } \hat{p}_{j:j+h-1} > \hat{p}_{i:i+h-1},\ Y_{j:j+h-1} > Y_{i:i+h-1}. \end{cases}$$

B Different lengths for moving average

Figure A.1 presents R^2_{OS} statistics (in %) for forecasts of the log equity premium. The predictor variable is the recession probability forecast from a probit model with TMS_t and $\frac{1}{l}\sum_{j=0}^{l-1}\text{TMS}_{t-j}$. We let the length of the moving average component, denoted by l, vary between values of 2 and 60. The panels on the left of Figure A.1 present the R^2_{OS} values for $h = 1, 6, 12$ and the panels on the right depict the Clark and West (2007) statistics. The vertical axis denotes the R^2_{OS} values and the MSFE-adjusted statistics and the horizontal axis denotes different values of l. We can see for $h = 1$ that the R^2_{OS} is above 0.50% for any backward-looking moving average between 30 to 60 months and that the MSFE-adjusted statistic is significant at the 5% level for these values. A shorter moving average of one year also has a R^2_{OS} above 0.50% and is significant at the 10%. The R^2_{OS} for one-month ahead forecasts is highest for moving averages between three to five years. For cumulative six- and twelve-month ahead forecasts the R^2_{OS} is always positive for moving averages between 2 to 60 months. A moving average between one to five years has statistically significant R^2_{OS} values above 2.50% (6%) for the six-month (twelve-month) horizon.

Table A.1 presents results for the break-correction methods when using values of $l = 12, 24, 48, 60$. The picture is similar to Figure A.1: the longer moving averages perform better for short horizon forecasts, whereas the exact choice of moving average is less important for long horizon forecasts. The R^2_{OS} values are positive for all break-correction methods and moving averages. Furthermore, the results are robust to weighted pooling (Pooling (weighted)) and to a minimum window length of 20 years for post-break window (Post-break window (20 years)).

C Forecasting log raw returns

Table A.2 presents the R^2_{OS} statistics when forecasting the log raw returns instead of the log equity premium (without subtracting the short rate). The results show that recession probabilities have strong predictive power for equity market returns not only in excess of the risk-free rate.

D Transaction costs

Table A.3 shows the gains in certainty equivalent return and Sharpe ratio relative to the prevailing mean when correcting for proportional transaction costs of 50 basis points per transaction.

This is similar to the base case in Balduzzi and Lynch (1999), which assumes proportional costs rather than fixed costs per transaction. This specification has often been used in related articles to account for trading costs; see, for example, Neely et al. (2014); Jiang et al. (2019). The results show that even when proportional trading costs are taken into account, asset allocation based on equity premium forecasts using recession probabilities provides investors with an economically and statistically significant advantage relative to forecasts based on the historical average.

E Asset allocation exercise for lower frequency re-balancing

Table A.4 shows the gains in certainty equivalent return and Sharpe ratio when an investor only re-balances the portfolio at the same frequency as the forecast horizons of $h = 3, 6, 12$ months. Hence, non-overlapping forecasts are used in this exercise; see Rapach et al. (2016) for details.

F Forecast evaluation: probit model with break-corrections and alternative lengths of the moving average

Table A.5 presents the forecast evaluation statistics for the probit model with TMS_t and $\frac{1}{36}\sum_{j=0}^{35}\text{TMS}_{t-j}$ as predictors, and for five break-corrected versions of this model. We can see that the break-adjusted model ($\text{TMS}_t^{\text{break}}$, $\frac{1}{36}\sum_{j=0}^{35}\text{TMS}_{t-j}^{\text{break}}$) does not improve the forecast performance. Contrarily, for $h = 1$ the other methods - cross-validation, pooling, and post-break window - have smaller (or identical) values for QPS, LS, and DES. We have seen in the main text that post-break window with at least 15 years of data in the probit model generates a false positive prior to 2000. This is canceled out when setting the minimum number of observations to 20 years. Thus, post-break window with 20 years performs even better for $h = 1, 3, 6, 12$ compared to post-break window with 15 years. Overall, cross-validation is most reliable in improving the performance relative to the unadjusted model; the statistics improve for each forecast horizon and each forecast evaluation statistic.

Table A.6 presents the forecast evaluation statistics for the probit model with TMS_t and $\frac{1}{l}\sum_{j=0}^{l-1}\text{TMS}_{t-j}$ for values of $l = 12, 24, 36, 48, 60$. The statistics often improve further for $l = 12$ and $l = 24$ compared to $l = 36$ but worsen for $l = 48$ and $l = 60$. Nonetheless, the latter two values still provide superior values compared to the probit model without the moving average component. This is most salient for short-horizon forecasts and AUROC values.

G Recursively estimated break dates and optimal starting points for cross-validation

The Sup-F, Ave-F, and Exp-F statistics are estimated as:

$$\text{Sup-F} = \sup_{\tau_1 \leq \tau \leq \tau_2} F_T(\tau), \tag{A.6}$$

$$\text{Ave-F} = \frac{1}{\tau_2 - \tau_1 + 1} \sum_{\tau=\tau_1}^{\tau_2} F_T(\tau), \tag{A.7}$$

$$\text{Exp-F} = \ln \Big[\frac{1}{\tau_2 - \tau_1 + 1} \sum_{\tau=\tau_1}^{\tau_2} \exp F_T(\tau)\Big], \tag{A.8}$$

where $F_T(\tau)$ refers to the Wald statistic for testing $\hat{\beta}_2 = \hat{\beta}_1$ and $\tau = \tau_1, \ldots, \tau_2$.

The recursively estimated break dates for the Sup-F test are shown in the upper panel of Figure A.2. The estimated location of the break is consistently between 1981 and 1983.

The lower panel of Figure A.2 presents the selected start values from cross-validation for one-year ahead log equity premium forecasts. At each point in time we estimate the probit model with different estimation windows, and evaluate forecasts from these models over a pseudo out-of-sample period consisting of the most recent five years of data. The optimal start value refers to the probit model with the smallest MSFE for the cumulative log equity premium with $h = 12$. It is important to mention that this analysis is feasible in real-time without any look-ahead bias in the data. Our most recently available observation at time t and $h = 12$ is $r_{t-11:t}$. The evaluation period runs from $r_{t-70:t-59}, \ldots, r_{t-11:t}$.

H Break tests with pre-whitening of standard errors

Table A.7 provides p-values for the Ave-F, Exp-F, and Sup-F statistics when applying pre-whitening to the standard errors. The null of no structural break is rejected for no pre-whitening, as well as for AR(1), AR(3), AR(6), and AR(12) pre-whitening. Furthermore, the results are robust to the choice of trimming value.

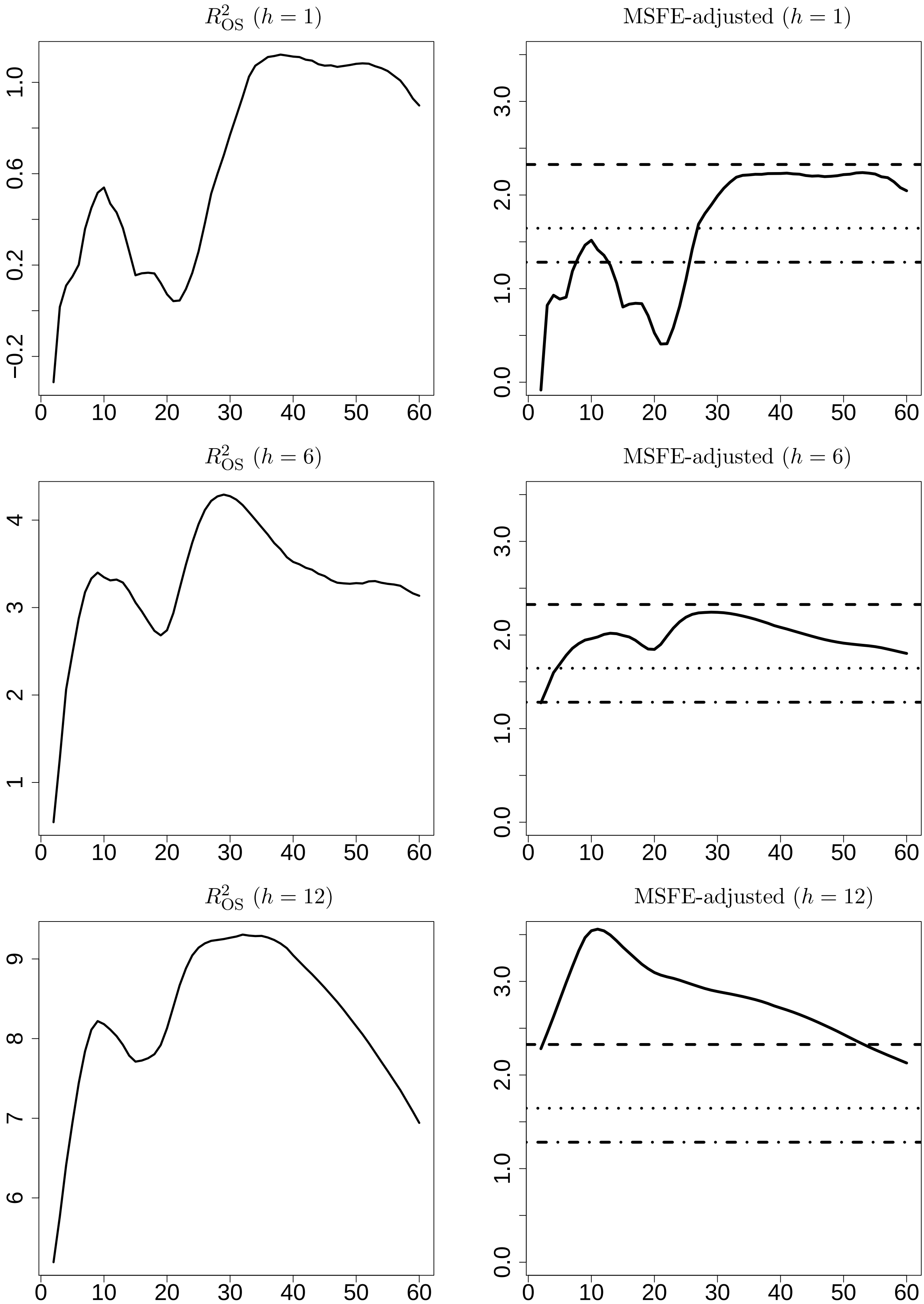

Figure A.1

Equity premium forecasts with different values of l in the probit model

This figure presents R^2_{OS} values (in %) for the log equity premium relative to the historical average and MSFE-adjusted statistics. Forecasts are based on recession probability forecasts from a probit model with TMS_t and $\frac{1}{l}\sum_{j=0}^{l-1}\mathrm{TMS}_{t-j}$, for l values between 2 and 60 months. The horizontal axis denotes the value of l and the vertical axis denotes the R^2_{OS} statistic (left panels) and the MSFE-adjusted statistic of Clark and West (2007) (right panels). Results are shown for out-of-sample forecasts from 1980:1 to 2019:12 and for three forecast horizons, namely $h = 1, 6, 12$. The black dashed line (dotted line) denotes the 1% (5%) critical value, and the dash-dotted line depicts the 10% critical value of the MSFE-adjusted statistic.

Table A.1

Out-of-sample R^2: Alternative lengths of moving average component

This table shows R^2_{OS} statistics (in %) for forecasts of the log equity premium relative to the historical average benchmark. The forecasts are based on a linear predictive regression model with a constant and recession probability forecasts as variables. Hereby, the probability forecasts are derived from a probit model with TMS_t and $\frac{1}{l}\sum_{j=0}^{l-1}\text{TMS}_{t-j}$, and for different break-correction methods of this probit model. Results are shown for different lengths of the moving average component l, namely for $l = 12, 24, 48, 60$. The out-of-sample period is 1980:1-2019:12 and $h = 1, 3, 6, 12$ depicts the forecast horizon.

(1)	(2)	(3)	(4)	(5)
Variable	l=12	l=24	l=48	l=60
Panel A: h = 1				
TMS_t, MA-TMS_t	0.43*	0.17	1.07**	0.90**
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.30	0.69**	1.25**	0.82**
Cross-validation	0.10	0.29	1.94**	1.04**
Pooling (weighted)	0.35	0.12	1.33***	1.02**
Pooling (average)	0.32	0.13	1.32***	1.04**
Post-break window (15 years)	0.10	0.13	2.43***	1.45***
Post-break window (20 years)	0.36*	0.10	1.94**	1.01**
Panel B: h = 3				
TMS_t, MA-TMS_t	1.60**	1.76**	2.63***	2.37**
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.63**	2.86**	2.30**	1.40**
Cross-validation	1.17**	2.58***	4.69***	2.67***
Pooling (weighted)	1.46**	1.81**	3.35***	2.73***
Pooling (average)	1.37**	1.86**	3.33***	2.81***
Post-break window (15 years)	1.24**	4.68***	6.18***	4.27***
Post-break window (20 years)	1.51**	2.76***	4.74***	2.69***
Panel C: h = 6				
TMS_t, MA-TMS_t	3.32**	3.74**	3.28**	3.13**
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	1.56**	4.77**	1.76**	0.94**
Cross-validation	1.96**	7.05***	5.90***	2.60**
Pooling (weighted)	3.25**	4.31***	4.42**	3.69**
Pooling (average)	3.12**	4.08**	4.31**	3.73**
Post-break window (15 years)	3.91***	7.57***	7.57***	4.95***
Post-break window (20 years)	3.42**	7.32***	6.31***	3.57**
Panel D: h = 12				
TMS_t, MA-TMS_t	8.03***	9.04***	8.36***	6.94**
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	6.93**	9.24***	7.63**	7.12**
Cross-validation	7.67***	9.50***	8.43***	5.98**
Pooling (weighted)	8.64***	10.92***	10.06***	8.05**
Pooling (average)	8.31***	10.54***	10.12***	8.29***
Post-break window (15 years)	9.53***	14.65***	14.24***	10.17***
Post-break window (20 years)	9.54***	14.81***	13.09***	9.87***

Table A.2

Out-of-sample $\mathbf{R^2}$ statistics for log raw returns

This table presents R^2_{OS} statistics (in %) for forecasts of log raw returns relative to the historical average benchmark. In contrast to the main text, we do not subtract the short rate from the continuously compounded returns on the S&P 500 index. The forecasts are based on a linear predictive regression model with a constant and recession probability forecasts as variables. Hereby, the probability forecasts are derived from a probit model with TMS_t and $\frac{1}{36}\sum_{j=0}^{35}\text{TMS}_{t-j}$, and for four different break-correction methods of this probit model. Results are shown for four different out-of-sample periods, and $h = 1, 3, 6, 12$ depicts the forecast horizon. *, **, *** denote significance at the 10%, 5%, and 1% significance levels according to the Clark and West (2007) MSFE-adjusted statistic. "Short interest" and "Gold-to-platinum ratio" refer to the predictors of Rapach et al. (2016) and Huang and Kilic (2019).

(1)	(2)	(3)	(4)	(5)
Variable	1980:1-2019:12	1980:1-1999:12	2000:1-2019:12	1990:1-2013:12
Panel A: h = 1				
TMS_t, MA-TMS_t	0.33*	0.60	0.05	0.28
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.97**	-0.31	2.34**	1.80**
Cross-validation	1.95**	0.63	3.36*	3.30**
Pooling (average)	0.58**	0.51	0.65*	0.80**
Post-break window	2.54***	0.83*	4.38**	4.51***
Short interest				1.19**
Gold-to-platinum ratio				1.56**
Panel B: h = 3				
TMS_t, MA-TMS_t	1.02**	2.29**	-0.18	0.70
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	2.04**	-1.31	5.25**	3.62**
Cross-validation	4.91**	2.06**	7.71*	7.14*
Pooling (average)	1.70**	2.10**	1.35*	1.92**
Post-break window	6.96***	2.67**	11.22**	11.25***
Short interest				4.09**
Gold-to-platinum ratio				6.12***
Panel C: h = 6				
TMS_t, MA-TMS_t	0.61*	2.53*	-0.82	0.79
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	1.39*	-5.75	6.71**	4.25**
Cross-validation	5.44*	1.28	8.48	8.34*
Pooling (average)	1.60*	2.20*	1.13	2.40*
Post-break window	8.19**	4.04**	11.27**	13.85**
Short interest				6.93**
Gold-to-platinum ratio				11.93***
Panel D: h = 12				
TMS_t, MA-TMS_t	3.38**	6.77**	1.40	3.18*
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	4.53**	-6.29	11.11**	5.05*
Cross-validation	0.39*	4.99*	-2.16	-0.10
Pooling (average)	4.45**	6.46*	3.34	4.47*
Post-break window	10.50**	12.57***	9.74	14.50*
Short interest				4.26*
Gold-to-platinum ratio				15.82***

Table A.3

Asset allocation exercise with proportional transaction costs

This table reports the annualized ΔCER and the annualized ΔSR for a mean-variance investor relative to forecasts from the historical average. The investor can invest in the S&P 500 index and the risk-free rate. The gains are corrected for proportional transaction costs of 50 basis points per transaction. Results are shown for one month ahead forecasts of the equity premium and different values for the coefficient of relative risk aversion (γ), and different restrictions on the equity weights (ω). The out-of-sample period is 1980:1 to 2019:12. *, **, *** indicate significantly improved performance relative to the prevailing mean benchmark at the 10%, 5%, and 1% significance level. The p-values are obtained by using a bootstrap approach similar to DeMiguel et al. (2013) with the average block length set to three months (Politis and Romano, 1994).

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: 1980:1 to 2019:12							
	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	2.57***	1.37*	2.29**	1.38***	0.14***	0.13**	0.13***	0.10***
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	3.80**	2.41**	3.16**	1.77	0.28**	0.28**	0.24**	0.22**
Cross-validation	3.79***	2.42**	3.09**	2.20***	0.21**	0.21**	0.17***	0.17**
Pooling (average)	3.23***	1.95***	2.96**	2.07***	0.19***	0.19***	0.17***	0.16***
Post-break window	5.69***	3.88***	4.58**	2.91***	0.33***	0.33***	0.27***	0.25***
Buy-and-hold	0.95**	0.42	0.95**	0.49**	0.07**	0.07**	0.07**	0.03*

Table A.4

Asset allocation exercise with lower-frequency re-balancing

This table reports the annualized ΔCER and the annualized ΔSR for a mean-variance investor relative to forecasts from the historical average. The investor can invest in the S&P 500 index and the risk-free rate. Results are shown for forecast horizons of $h = 3, 6, 12$ months of the equity premium, where an investor re-balances at the same frequency as the forecast horizon (Rapach et al., 2016). The coefficient of relative risk aversion is denoted by γ, and ω states restrictions on the weights in the risky asset. The "Prevailing mean" shows the CER and SR values, whereas all other values denote the improvements relative to this benchmark. Results are shown for the out-of-sample period from 1980:1 to 2019:12.

Panel A: $h = 3$

	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$
Prevailing mean	7.47	5.92	7.47	8.26	0.44	0.43	0.44	0.49
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	3.00	1.21	2.84	1.48	0.17	0.14	0.16	0.10
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	3.01	2.03	3.06	1.97	0.19	0.20	0.19	0.20
Cross-validation	5.05	3.20	4.28	2.92	0.27	0.27	0.23	0.21
Pooling (weighted)	3.68	2.05	3.52	2.32	0.21	0.19	0.20	0.16
Pooling (average)	3.49	1.94	3.34	2.24	0.20	0.18	0.19	0.16
Post-break window	6.02	3.93	5.07	3.26	0.33	0.33	0.28	0.25
Buy-and-hold	1.49	0.58	1.49	0.71	0.09	0.11	0.09	0.04

Panel B: $h = 6$

	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$
Prevailing mean	8.04	6.29	8.04	8.49	0.48	0.46	0.48	0.52
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	2.40	1.29	2.20	1.38	0.14	0.14	0.13	0.09
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	1.64	1.08	1.72	1.10	0.11	0.12	0.12	0.10
Cross-validation	4.55	2.97	3.98	2.50	0.26	0.27	0.22	0.19
Pooling (weighted)	3.14	1.91	2.94	1.90	0.19	0.19	0.17	0.14
Pooling (average)	2.93	1.78	2.74	1.74	0.18	0.18	0.16	0.12
Post-break window	5.24	3.19	4.60	2.43	0.30	0.28	0.27	0.19
Buy-and-hold	1.23	0.71	1.23	0.78	0.08	0.10	0.08	0.04

Panel C: $h = 12$

	ΔCER				ΔSR			
γ	3	5	3	3	3	5	3	3
ω	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$	$[-0.5, 1.5]$	$[-0.5, 1.5]$	$[0, 1.5]$	$[0, 1]$
Prevailing mean	7.66	5.59	7.66	8.19	0.46	0.41	0.46	0.48
Gains relative to prevailing mean:								
TMS_t, MA-TMS_t	2.85	1.26	2.97	1.99	0.16	0.13	0.16	0.14
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.99	1.03	1.49	1.04	0.06	0.08	0.09	0.11
Cross-validation	2.39	0.99	2.50	1.76	0.13	0.10	0.13	0.13
Pooling (weighted)	2.83	1.76	2.95	2.35	0.16	0.15	0.17	0.18
Pooling (average)	2.68	1.60	2.79	2.24	0.15	0.14	0.16	0.17
Post-break window	2.60	0.75	2.79	2.04	0.14	0.09	0.15	0.16
Buy-and-hold	1.49	0.85	1.49	0.96	0.08	0.12	0.08	0.05

Table A.5

Out-of-sample forecasting performance - Probit model

This table presents five forecast evaluation statistics for the out-of-sample recession probability forecasts, as well as the correlation between the probability forecasts and the (cumulative) log equity premium (ρ). The statistics are the quadratic probability score (QPS), logarithm score (LS), diagonal elementary score (DES), pseudo R^2, and the area under the receiver operating characteristic curve (AUROC). Results are shown for the probit model with TMS_t and $\frac{1}{36}\sum_{j=0}^{35}\text{TMS}_{t-j}$ as predictors, and for five break-corrected versions of this probit model (Pesaran and Timmermann, 2007; Lettau and Van Nieuwerburgh, 2008). The recession probability forecasts refer to the probability that a recession occurs within the next h months. Results are shown for $h = 1, 3, 6, 12$ and the out-of-sample period is 1980:1 to 2019:12.

	1980:1 to 2019:12					
Variables in probit model	QPS	LS	DES	pseudo R^2	AUROC	ρ
Panel A: h = 1						
TMS_t, MA-TMS_t	0.17	0.27	0.05	0.27	0.92	-0.14
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.20	0.32	0.06	0.15	0.84	-0.13
Cross-validation	0.15	0.23	0.03	0.35	0.94	-0.16
Pooling (average)	0.17	0.26	0.04	0.29	0.92	-0.15
Post-break window (15 years)	0.17	0.25	0.04	0.30	0.91	-0.18
Post-break window (20 years)	0.14	0.22	0.04	0.37	0.94	-0.19
Panel B: h = 3						
TMS_t, MA-TMS_t	0.20	0.31	0.07	0.25	0.90	-0.22
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.24	0.37	0.07	0.11	0.81	-0.20
Cross-validation	0.16	0.25	0.04	0.37	0.93	-0.29
Pooling (average)	0.19	0.29	0.06	0.28	0.90	-0.25
Post-break window (15 years)	0.19	0.28	0.05	0.31	0.91	-0.29
Post-break window (20 years)	0.16	0.25	0.05	0.38	0.93	-0.31
Panel C: h = 6						
TMS_t, MA-TMS_t	0.22	0.36	0.09	0.25	0.87	-0.25
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.29	0.44	0.08	0.08	0.79	-0.22
Cross-validation	0.19	0.30	0.07	0.37	0.92	-0.35
Pooling (average)	0.21	0.34	0.08	0.29	0.88	-0.29
Post-break window (15 years)	0.21	0.33	0.06	0.31	0.90	-0.34
Post-break window (20 years)	0.18	0.29	0.06	0.39	0.91	-0.38
Panel D: h = 12						
TMS_t, MA-TMS_t	0.25	0.41	0.10	0.32	0.86	-0.37
$\text{TMS}_t^{\text{break}}$, MA-$\text{TMS}_t^{\text{break}}$	0.34	0.52	0.11	0.11	0.80	-0.32
Cross-validation	0.21	0.34	0.08	0.45	0.90	-0.48
Pooling (average)	0.24	0.39	0.09	0.35	0.87	-0.43
Post-break window (15 years)	0.29	0.51	0.09	0.12	0.87	-0.43
Post-break window (20 years)	0.25	0.39	0.08	0.35	0.89	-0.46

Table A.6

Out-of-sample forecasting performance - Different lengths of the moving average component

This table presents five forecast evaluation statistics for the out-of-sample performance of the probit model with the term spread and the moving average term spread as variables, as well as the correlation between the probability forecasts and the (cumulative) log equity premium (ρ). The statistics are the quadratic probability score (QPS), logarithm score (LS), diagonal elementary score (DES), pseudo R^2, as well as the area under the receiver operating characteristic curve (AUROC). Results are shown for five different lengths of the moving average term spread, $\frac{1}{l}\sum_{j=0}^{l-1}\text{TMS}_{t-j}$, namely $l = 12, 24, 36, 48, 60$. The recession probability forecasts refer to the probability that a recession occurs within the next h months. Results are shown for $h = 1, 3, 6, 12$ and the out-of-sample period is 1980:1 to 2019:12.

	1980:1 to 2019:12					
Variables in probit model	QPS	LS	DES	pseudo R^2	AUROC	ρ
Panel A: h = 1						
$l = 12$	0.15	0.30	0.04	0.19	0.87	-0.08
$l = 24$	0.15	0.23	0.02	0.34	0.93	-0.09
$l = 36$	0.17	0.27	0.05	0.27	0.92	-0.14
$l = 48$	0.19	0.31	0.07	0.16	0.87	-0.12
$l = 60$	0.21	0.37	0.08	0.05	0.79	-0.12
Panel B: h = 3						
$l = 12$	0.16	0.33	0.05	0.21	0.86	-0.15
$l = 24$	0.17	0.26	0.03	0.35	0.93	-0.19
$l = 36$	0.20	0.31	0.07	0.25	0.90	-0.22
$l = 48$	0.22	0.36	0.09	0.15	0.85	-0.19
$l = 60$	0.23	0.41	0.09	0.05	0.76	-0.18
Panel C: h = 6						
$l = 12$	0.18	0.35	0.06	0.27	0.87	-0.22
$l = 24$	0.19	0.30	0.05	0.37	0.91	-0.25
$l = 36$	0.22	0.36	0.09	0.25	0.87	-0.25
$l = 48$	0.24	0.40	0.09	0.16	0.83	-0.22
$l = 60$	0.25	0.45	0.10	0.07	0.74	-0.20
Panel D: h = 12						
$l = 12$	0.18	0.35	0.06	0.44	0.90	-0.37
$l = 24$	0.21	0.35	0.07	0.44	0.90	-0.40
$l = 36$	0.25	0.41	0.10	0.32	0.86	-0.37
$l = 48$	0.26	0.43	0.10	0.28	0.84	-0.33
$l = 60$	0.27	0.46	0.11	0.21	0.78	-0.29

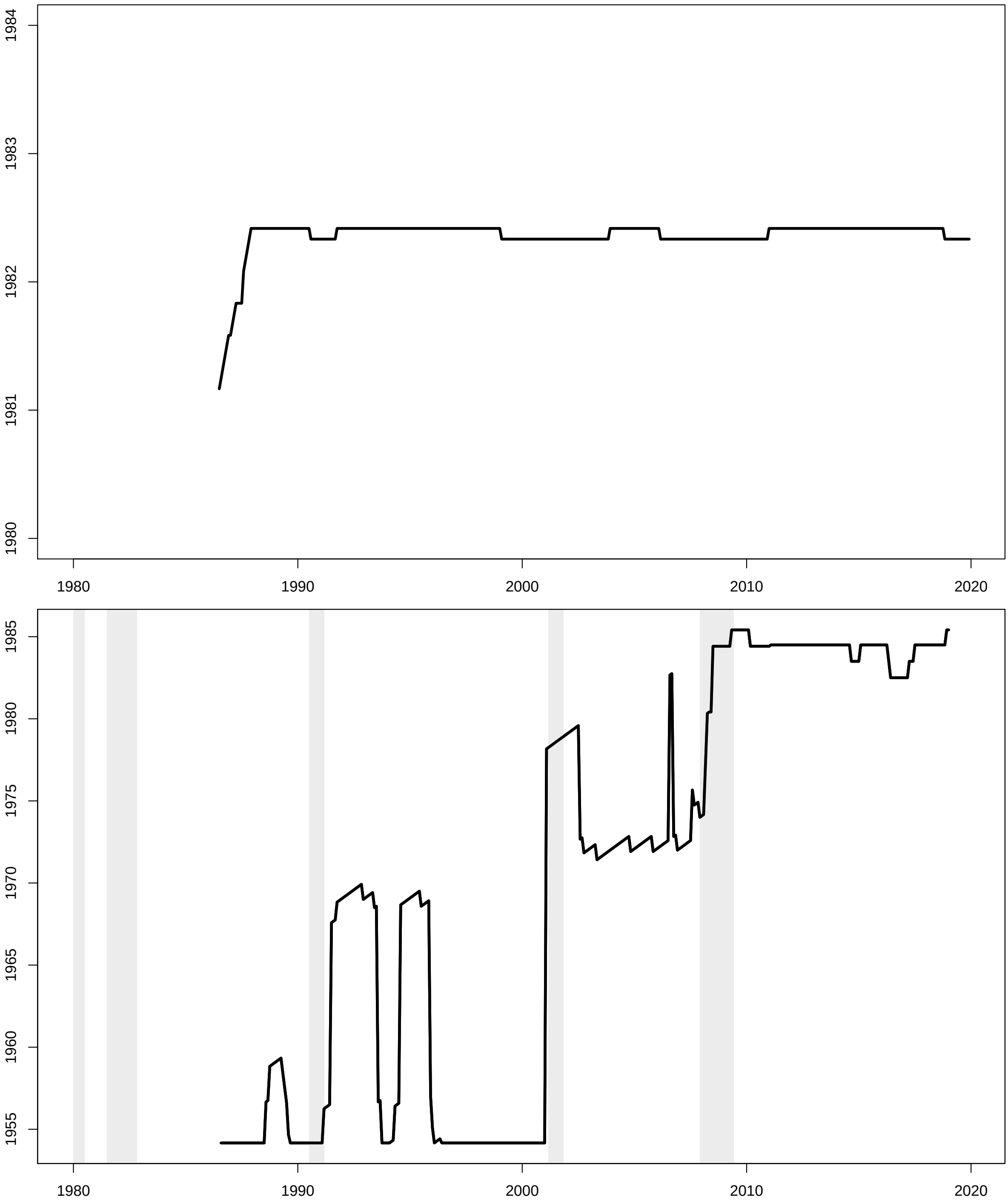

Figure A.2

Recursively estimated break dates and optimal starting points for cross-validation

This figure presents the recursively estimated break dates for the Sup-F test (upper panel) and the optimal start values of the estimation window for cross-validation for one-year ahead equity premium forecasts (lower panel). The horizontal axis denotes the time of estimation and the vertical axis denotes the respective break date and start value for the probit model. As an example, from 1990 onward the estimated break date was consistently between 1982 and 1983. The optimal start value for cross-validation is estimated by performing a pseudo out-of-sample exercise over the most recent five years of data, whereby the data are recursively expanding. The selected start date refers to the probit model with the lowest MSFE for the equity premium over the holdout period.

Table A.7

Estimated p-values with pre-whitening of standard errors

This table reports p-values for the null hypothesis of no structural break in the mean of the term spread. Ave-F, Exp-F, and Sup-F refer to the test statistics of Andrews (1993) and Andrews and Ploberger (1994), and p-values are estimated by Hansen (1997). The standard errors are Newey and West (1994) with an automatic bandwidth selection and different versions of pre-whitening. Results are shown for no pre-whitening of standard errors and for AR(1), AR(3), AR(6), and AR(12) pre-whitening of standard errors. The tests are carried out for trimming values of 15% (Panel A) and 5% (Panel B). The estimations are based on data from 1951:3 to 2019:12 and the break date refers to the global minimum in the sum of squared errors (Bai, 1997).

(1)	(2)	(3)	(4)	(5)
	Ave-F	Exp-F	Sup-F	break date
		Panel A: 15% trimming		
No pre-whitening	0.00	0.00	0.00	1982:5
AR(1)	0.06	0.04	0.01	
AR(3)	0.03	0.02	0.01	
AR(6)	0.01	0.01	0.01	
AR(12)	0.00	0.00	0.00	
		Panel B: 5% trimming		
No pre-whitening	0.00	0.00	0.00	1982:5
AR(1)	0.07	0.05	0.02	
AR(3)	0.03	0.02	0.01	
AR(6)	0.01	0.01	0.02	
AR(12)	0.00	0.00	0.00	