12-2022

# Forecasting Variance Swap Payoffs

Jonathan Dark
*University of Melbourne*

Xin Gao
*Sacred Heart University*

Thijs van der Heijden
*University of Melbourne*

Federico Nardari
*University of Melbourne*



## Recommended Citation

Dark, J., Gao, X., van der Heijden, T., & Nardari, F. (2022). Forecasting variance swap payoffs. *Journal of Futures Markets, 42*(12), 2135-2164. Doi:10.1002/fut.22371

Received: 23 February 2022 | Accepted: 7 July 2022

DOI: 10.1002/fut.22371

RESEARCH ARTICLE

FUTURES MARKETS WILEY

# Forecasting variance swap payoffs

**Jonathan Dark**[1] | **Xin Gao**[2] | **Thijs van der Heijden**[1] | **Federico Nardari**[1]

[1]Department of Finance, University of Melbourne, Parkville, Victoria, Australia

[2]Department of Finance, Sacred Heart University, Fairfield, Connecticut, USA

**Correspondence**
Jonathan Dark, Department of Finance, University of Melbourne, Parkville, Level 11, 198 Berkeley street, VIC 3010, Australia.
Email: jdark@unimelb.edu.au

**Abstract**

We investigate the predictability of payoffs from selling variance swaps on the S&P500, US 10-year treasuries, gold, and crude oil. In-sample analysis shows that structural breaks are an important feature when modeling payoffs, and hence the ex post variance risk premium. Out-of-sample tests, on the other hand, reveal that structural break models do not improve forecast performance relative to simpler linear (or state invariant) models. We show that a host of variables that had previously been shown to forecast excess returns for the four asset classes, contain predictive power for ex post realizations of the respective variance risk premia as well. We also find that models fit directly to payoffs perform as well or better than models that combine the current variance swap rate with a realized variance forecast. These novel findings have important implications for variance swap sellers, and investors seeking to include volatility as an asset in their portfolio.

**KEYWORDS**

forecasting, model selection, variance risk premium, variance swaps

**JEL CLASSIFICATION**

G17, C52, C53, C58

## 1 | INTRODUCTION

Financial markets have long recognized that asset's volatility is stochastic, making investors exposed to variance risk. In a seminal contribution, Carr and Wu (2009) show that variance risk is significantly priced. In investment practice, volatility trading and hedging have grown tremendously over the past two decades. Variance swaps, in particular, have become popular vehicles for investors to achieve their desired volatility exposures and to possibly harvest the variance risk premium (VRP). In this paper we seek to forecast the payoffs from selling variance swaps, as those payoffs represent ex post realizations of the VRP.[1]

Existing research investigates the size and dynamics of the VRP (e.g., Bakshi & Kapadia, 2003; Carr & Wu, 2009; Eraker & Wu, 2017; Li & Zinna, 2018, among many others), while several studies, starting with Bollerslev et al. (2009), focus on the predictive power of the VRP for the equity risk premium. Other contributions focus on the effects of

[1]See, for example, Carr and Wu (2009) and Sections 2 and 3. Our exercise is, therefore, about forecasting the ex post VRP.

including variance swaps or other volatility products within an asset allocation framework (important examples are Alexander et al., 2016; Brière et al., 2010; Chen et al., 2011; Doran, 2020; Egloff et al., 2010; Nieto et al., 2014; Warren, 2012).

Despite this growing literature, the number of papers that look specifically at the *out-of-sample* (OOS) forecasting of the VRP is rather limited: most of the studies only present in-sample analysis, typically relying on a few predictive variables.[2] The literature on return predictability has long recognized that a large set of predictors may be useful in forecasting excess returns across assets.[3] These variables have been shown to track changes in macroeconomic and business conditions, and thus to anticipate fluctuations in risk premia for the respective asset classes. Consequently, it seems natural to ask whether they are also able to forecast the VRP. This is because the latter is indeed a compensation for bearing systematic risk, and may therefore display time-variation originating from changes in the economic or market environment. We address the issue by employing a host of predictors from the return predictability literature.[4]

Another drawback of VRP forecasting studies is the focus on a single model specification, typically a linear multiple regression or a vector autoregression (VAR). Given the need for more complex dynamics, such as regime switching, emphasized by recent literature (Rombouts et al., 2020), and given the more general concern of model uncertainty, it seems relevant to explore a variety of model structures within a unified framework for assessing forecasting ability.[5] We therefore consider a large set of forecasting models and rely on (a) the Model Confidence Set (MCS) approach of Hansen et al. (2011) to evaluate OOS forecasting performance and (b) forecast combination (or model averaging) to alleviate issues arising from model misspecification.

Finally, the few papers that look at OOS forecasting of the VRP focus almost exclusively on the equity VRP. We also investigate the predictability of the VRP for government bonds, gold, and crude oil.[6]

To the best of our knowledge, this is the first study on the OOS predictability of the VRP that entertains: (a) a variety of model specifications, linear as well as nonlinear; (b) a large set of predictive variables; and (c) four major asset classes.

Our in-sample analysis shows that for all four assets, the dynamics of the realized (or ex post) VRP is characterized by marked regime changes (or structural breaks). Longer periods of positive, less volatile, and more persistent payoffs for variance swap sellers, alternate with shorter stretches of negative (typically large), more volatile, and less predictable payoffs. The OOS analysis generates five main conclusions. First, nonlinear specifications that include structural breaks do not generate superior forecasts relative to linear models. Second, the variables previously utilized in the literature to predict excess returns also contain predictive power for the respective ex post VRP realizations as they outperform a no-predictability (or prevailing mean) benchmark. For equities and treasuries, the specifications that utilize predictors (specifically, by extracting their principal components [PCs]) outperform pure time-series models (e.g., the autoregressive moving average [ARMA] class of processes); while for gold and crude oil, the time-series models are not outperformed by those that include predictors. Third, for equities, gold, and oil, relatively simple benchmarks (namely, for gold and oil, a linear autoregressive [AR] specification; for equities, a linear model containing only the lagged first PC of predictors) perform as well or better than richer models or model combinations. For treasuries, on the other hand, a combination of PC models and a combination of MIDAS models are the best performers. Fourth, models fit directly to variance swap payoffs generally perform as well or better than models that combine a realized variance forecast with the current variance swap rate. Fifth, in terms of economic significance, the predictors for

[2] Feunou et al. (2014) consider a comprehensive set of predictors for the equity VRP, but their analysis is in-sample and relies exclusively on linear models.

[3] Rapach and Zhou (2013) provide a survey of the literature for equities. See Gao and Nardari (2018) for applications to Treasury bonds and to commodities.

[4] In Section 3.1, we discuss the conceptual foundations that link time-variation in the VRP with the macro and business environment.

[5] Konstantinidi and Skiadopoulos (2016) forecast the return from a position on a variance swap on the S&P500 using a range of predictors. Whilst a number of variables are considered, their models are confined to simple linear regressions. Andreou and Ghysels (2021) construct various factors (short-run funding spreads, and long-run corporate and government bond spreads) to predict the Volatility Index (VIX), realized volatility (RV), and VRP in the equity market. The factors work well when forecasting the VIX and RV OOS. However the VRP results are not reported: according to the authors (p. 23, footnote 18) "The OOS results for the VRP are also relatively more weak than those of the VIX and RV and therefore are not reported." Further, Andreou and Ghysels (2021) only consider simple linear regression as well as Mixed-Data Sampling (MIDAS) regression.

[6] Prokopczuk et al. (2011) analyze the VRP for many commodities using synthetic variance swaps. Choi et al. (2017) comprehensively study the VRP in the US Treasury market and evaluate trading strategies that rely on variance swaps. Neither study, though, investigates the predictability of the VRP.

equity, gold, and oil VRPs appear to improve trading performance relative to strategies based on a prevailing mean forecast and to strategies based on pure time-series models. That is not the case, however, for the predictors of the bond VRP. Our novel evidence should be of relevance for variance swap sellers, hedgers as well as portfolio managers.

The paper proceeds as follows. Section 2 outlines the methodology, which includes an overview of the models considered, the MCS approach, and model averaging. Section 3 discusses the data, namely, the predictors and the method used to extract swap payoffs. Empirical results are presented in Section 4, which examines in-sample fit, OOS forecast performance, and some simple trading strategies. Section 5 concludes.

## 2 | METHODOLOGY

A variance swap entered at time $t$ and held to maturity, $t + T$, is a zero-cost contract that pays the difference between the physical, realized variance over the period, $\mathrm{RV}_{t,T}$, and the agreed upon (i.e., fixed) variance swap rate at the time the contract is entered into, $\mathrm{VS}_{t,T}$. The difference is paid per dollar of notional amount. The variance swap rate is set as the expectation, under the risk-neutral measure, of the realized variance over the life of the swap, $\mathrm{VS}_{t,T} = E_t^Q(\mathrm{RV}_{t,T})$, so that investors are indifferent between the fixed and the floating legs of the swap. The ex ante VRP is then defined as

$$\mathrm{VRP}_{t,T} = E_t^P(\mathrm{RV}_{t,T}) - \mathrm{VS}_{t,T}. \tag{1}$$

The payoff (per unit of notional) on a long variance swap *contract* entered at time $t$ and held to maturity $T$ is computed as

$$\mathrm{PL}_{t,T} = \mathrm{RV}_{t,T} - \mathrm{VS}_{t,T}. \tag{2}$$

It can be shown that this payoff equals the return in excess of the risk-free rate of a fully collateralized long *position* on the variance swap.[7]

At the end of each month $t$, we seek to forecast the payoff (PL) over the forthcoming month of a long position on a variance swap with a 1-month maturity:

$$\mathrm{PL}_{t,t+1} = \mathrm{RV}_{t,t+1} - \mathrm{VS}_{t,t+1}, \tag{3}$$

where $\mathrm{RV}_{t,t+1}$ is the realized variance over the month, and $\mathrm{VS}_{t,t+1}$, the variance swap rate for the forthcoming month, is observable at time $t$. The variance swap payoff we aim to forecast (Equation 3), therefore represents an ex post realization of the VRP defined in Equation (1). If $E_t^P(\mathrm{RV}_{t,T}) = \mathrm{RV}_{t,T}$, as per Bollerslev et al. (2011) the ex ante VRP is equivalent to the ex post VRP. In what follows we use the terms ex post VRP and variance swap payoff interchangeably.[8]

We analyze the returns on fully collateralized positions as that is customary in the literature when looking at the performance of zero-investment assets, such as futures and swaps. In addition, the asset allocation literature that

[7]The return on a long variance swap *position* with a required collateral is the sum of the return on the contract and the return on the collateral. Let $\mathrm{rf}_{t,T}$ be the net return on a risk-free asset between $t$ and $T$ and cf be the amount of collateral required per dollar notional in the swap, which we term the "collateralization factor." Then

$$r_{t,T}^{\text{swap, long}} \equiv \left(\frac{1}{\mathrm{cf}}\tilde{r}_{t,T}^{\text{swap}} + \mathrm{rf}_{t,T}\right).$$

It follows that the *excess* return on a long swap position $\mathrm{xr}_{t,T}^{\text{swap, long}}$ equals the scaled payoff of the swap contract:

$$\mathrm{xr}_{t,T}^{\text{swap, long}} = \frac{1}{\mathrm{cf}}\tilde{r}_{t,T}^{\text{swap}}.$$

With full collateralization, $\mathrm{cf} = 1$, this excess return equals the payoff in Equation (2).

[8]This is consistent with the equity premium forecasting literature where realized returns are often used as a noisy proxy for expected returns.

investigates the value of adding pure variance exposure to equity and/or fixed income, typically relies on fully collateralized positions as well.[9,10]

We consider two alternative forecasting approaches. The first approach fits models directly to the payoff series and uses the model to make a one-step-ahead forecast. Given that PL is not observed until the end of the month, this approach ignores the most recent VS rate at time $t$. We refer to this first approach as a *direct* forecast. The second approach forecasts RV conditional on the information available at time $t$. The RV forecast is then combined with the VS rate at time $t$ to form the payoff forecast for $t + 1$. The second approach therefore combines the VS rate with a model-based forecast (RV), and is referred to as a *hybrid* forecast. The RV models are estimated using daily data and the monthly variance forecast obtained via aggregation. An examination into the relative performance of these two approaches has not been considered in the literature.

OOS forecasts are generated using a recursive scheme on an expanding window. Namely, the models are first estimated using information available up to time $T_0$ and then used to forecast the payoff at the end of period $T_0 + 1$ for a 1-month variance swap sold at $T_0$. At $T_0 + 1$, the realized payoff on the swap is observed. The models are then re-estimated using the additional information available and the forecasts reperformed for $T_0 + 2$. This is continued until the end of the data set.

Forecast performance is then evaluated using the Mean-Squared Error (MSE), where the actual is the realized payoff. We employ a large set of predictors and forecast models which may raise concerns regarding multiple-testing. The MCS approach of Hansen et al. (2011) is designed to handle this situation, so we use it to rank forecast performance and arrive at a final set of models with equal predictive ability (EPA). Let $M^0$ represent the initial set of models. The MCS identifies a subset of models, $M^*$ which are superior in predictive ability with respect to all other models in $M^0$.[11]

## 2.1 | Models

### 2.1.1 | Direct swap payoff forecasts

For the first approach, the direct forecast of payoffs, we consider pure time-series models as well as models with predictors (i.e., models that include independent lagged forecasting variables in addition to lags of the dependent variable). For the pure time-series models we consider the following specifications: ARMA, breakpoint regression, threshold regression, and Markov switching (MS) regressions. For the models with predictors, we augment all the pure time-series specifications with predictors. We also consider reduced form VAR, linear regression against lagged PCs extracted from the predictors, and MIDAS regressions. As the ARMA, reduced form VAR and linear regression against lagged PCs specifications are well understood, in what follows we only briefly detail the remaining processes.

The breakpoint regression is a standard linear regression model with say $T$ periods and $m$ breaks (resulting in $m + 1$ regimes). For the observations in regime $j$

$$y_t = X_t'\beta + Z_t'\delta_j + \sigma_j\epsilon_t \tag{4}$$

[9]See, for instance, C. Alexander et al. (2016), Brière et al. (2010), Doran (2020), Egloff et al. (2010), Nieto et al. (2014), Warren (2012), and Chen et al. (2011).

[10]Alternatively, one may consider returns on variance swap positions with *no collateral*. The return in such case is computed by scaling PL with the agreed upon swap rate, $\mathrm{VS}_{t,T}$, as shown, for instance, in Konstantinidi and Skiadopoulos (2016). As discussed in Carr and Wu (2009), this notion of return aligns with an alternative way of defining the VRP. Namely,

$$\mathrm{VRP}_{t,T} = \frac{1}{\mathrm{VS}_{t,T}}\left[E_t^P(\mathrm{RV}_{t,T} - \mathrm{VS}_{t,T})\right].$$

In their analysis of ex post returns based on this definition of VRP, Carr and Wu (2009) actually use $\ln(\mathrm{RV}/\mathrm{VS})$ and interpret it as the continuously compounded excess return to going long a variance swap and holding it until maturity. However in their study $\mathrm{VS}_{t,T}$ is seen as the initial investment, whereas (2) and (3) regard a \$1 notional as the initial investment.

[11]The starting hypothesis is that all models have EPA which is measured by the loss function $L_{i,j}$. Define the loss differential between two models as $d_{i,j} = L_{i,t} - L_{j,t}$ for all $i \neq j \in M_0$. Let $\bar{d}_{ij} = n^{-1}\sum_{t=1}^{n} d_{ij,t}$ be the relative performance of models $i$ and $j$ and $\bar{d}_i = (m-1)^{-1}\sum_{j\in M}\bar{d}_{i,j}$ denotes model $i$'s performance relative to the average of the models in $M$. The null hypothesis is $H_{0,M^0}: E(d_{ij}) = 0, \forall\ i \neq j \in M^0$. The tests employ the range $T_R$ statistic, $T_R = \max_{i,j\in M}\frac{|\bar{d}_{ij}|}{\sqrt{\mathrm{var}(\bar{d}_{ij})}}$, where the sum is over all models in $M$ and $\mathrm{var}(\bar{d}_{ij})$ is a bootstrap estimate of the variance of $\bar{d}_{ij}$. To control for any serial correlation, a block bootstrap is employed with 10,000 replications. If the null is rejected, the model with the highest standardized loss, $\mathrm{argmax}_{i\in M}\bar{d}_i/\sqrt{\mathrm{var}(\bar{d}_i)}$ is removed ($\mathrm{var}(\bar{d}_i)$ is also a bootstrap estimate). This is repeated until the null can no longer be rejected for a given level of confidence. The surviving models are the MCS.

for regimes $j = 0, 1, ..., m$, where the $\beta$ coefficients on the $X$ variables do not vary across regimes, while the $Z$ variables have coefficients that are regime specific. The variance of the errors may also be regime dependent ($\sigma_0 \neq \cdots \sigma_m$). In all our specifications $Z_t$ contains the lagged $y_t$ while additional predictors may be either in $Z_t$ or in $X_t$, as detailed below.

The threshold regression has the same linear specification for each of the $j = 0, 1, ..., m$ regimes. The difference is that threshold regression switches between regimes according to the values of an observable variable $q_t$. Consider strictly increasing threshold values $\gamma_1 < \gamma_2 < \cdots < \gamma_m$ so that the process is in regime $j$ if $\gamma_j \leq q_t < \gamma_{j+1}$, where $\gamma_0 = -\infty$ and $\gamma_{m+1} = \infty$. If $q_t$ is the $d$th lag of $y_t$, the model is a self-exciting model with delay $d$. If the model is self-exciting with only a constant and AR component in each regime, the model is a self-exciting threshold autoregression. Given the large number of predictors considered below, to avoid model proliferation all threshold regressions we consider are self-exciting.

Breakpoint regression is therefore equivalent to a threshold regression with time as the threshold variable. This means that threshold regression can re-enter a regime after exiting, however, breakpoint regression cannot. For both models, the breakpoints are determined using sequential tests of $l + 1$ versus $l$ breakpoints. This is applied sequentially until the null of $l$ breakpoints is no longer rejected.

MS models assume that $y_t$ follows a process that depends on a latent state variable $s_t$. For each of the $m$ regimes

$$y_t = X_t'\beta + Z_t'\delta_j + \sigma_j \epsilon_t, \tag{5}$$

where $\epsilon_t$ is i.i.d. standard normal and $\sigma_j$ is state dependent. Again, $Z_t$ contains the lagged $y_t$ while additional predictors may be either in $Z_t$ or in $X_t$. A first-order Markov process is used for the regime probabilities where

$$P(s_t = j/s_{t-1} = i) = p_{ij}, \tag{6}$$

where $p_{ij}$ is the probability of transitioning from regime $i$ to regime $j$ in period $t$.

The PC models we consider can also be cast within Equation (4) where $X_t$ contains the lagged PCs extracted from the predictors and $Z_t$ contains the lagged $y_t$. Therefore, we consider PC models with breakpoint, threshold, and MS specifications for the lagged dependent variable. For all these models, lagged PCs are state invariant and are therefore in $X_t$. State dependence is captured via the constant and AR(1) term (i.e., via $Z_t$).

The above models require all variables to be measured at the same frequency. The MIDAS model (Ghysels et al., 2004) enables variables with different frequencies to enter the regression. The most common specification regresses a low-frequency dependent variable against higher-frequency regressors. The MIDAS model for a single explanatory variable and $h$-step ahead forecasting is

$$y_t = \beta_0 + \beta_1 B(L^{1/m}; \theta) x_{t-h}^{(m)} + \varepsilon_t, \tag{7}$$

where $B(L^{1/m}; \theta) = \Sigma_{j=1}^{K} b(j; \theta) L^{(j-1)/m}$, $K$ is the lag length, and $L^{s/m} x_{t-1}^{(m)} = x_{t-1-s/m}^{(m)}$. Here $t$ is the low-frequency time unit (say monthly) and $m$ is the higher sampling frequency (say daily), where $L^{1/m}$ operates at the higher frequency. This has been extended to the $p$-order AR-MIDAS model

$$y_t = \beta_0 + \sum_{i=1}^{p} \lambda_i y_{t-h-i+1} + \beta_1 B(L^{1/m}; \theta) x_{t-h}^{(m)} + \varepsilon_t. \tag{8}$$

A common polynomial choice is the exponential Almon lag which parameterizes $b(j; \theta)$ as

$$b(j; \theta) = \frac{\exp(\theta_1 j + \theta_2 j^2)}{\Sigma_{j=1}^{K} \exp(\theta_1 j + \theta_2 j^2)}. \tag{9}$$

The model requires estimation via nonlinear least squares. Parameter estimates can be sensitive to starting values and convergence issues often result. This is particularly problematic when estimating a large number of models with rolling or expanding windows, so we also consider analytically estimated MIDAS models. The first model is the polynomial distributed lag (PDL) which lets

$$b(j; \theta) = \sum_{k=0}^{Q} \theta_k j^k, \tag{10}$$

however, there are significant costs if an incorrect polynomial order ($Q$) or lag length is imposed. An alternative model employs stepwise weights (Forsberg & Ghysels, 2007) which require determination of the number of steps (or step size)

$$b(j;\theta) = \theta_1 I_{j\in[a_0,a_1]} + \sum_{p=2}^{R} \theta_P I_{j\in[a_{p-1},a_p]}. \tag{11}$$

The $p+1$ parameters $a_0, \ldots, a_p$ are thresholds with $a_0 = 1 < a_1 < \cdots < a_p = K$ and $I_{j\in[a_{p-1},a_p]}$ is an indicator equal to 1 if $a_{p-1} \leq j < a_p$, otherwise zero. With the polynomial specification in (11), Equation (8) becomes a STEP AR ($p$) MIDAS model.

### 2.1.2 | Hybrid forecasting models

For the second (or hybrid) forecasting approach we fit ARFIMA$(1, d, 1)$ and Heterogeneous Autoregressive (HAR) models to daily RV and daily log(RV) conditional on the information available at time $t$. The second approach, therefore, utilizes data at a much higher frequency and hence relies on considerably more observations.

The HAR–RV model is

$$\mathrm{RV}_{t+1}^{(D)} = c + \beta^{(D)}\mathrm{RV}_t^{(D)} + \beta^{(W)}\mathrm{RV}_t^{(W)} + \beta^{(M)}\mathrm{RV}_t^{(M)} + \varepsilon_{t+1}, \tag{12}$$

where $\mathrm{RV}_t^{(D)}$ denotes the realized variance for day $t$, and $\mathrm{RV}_t^{(W)}$ and $\mathrm{RV}_t^{(M)}$ denotes the weekly and monthly averages, respectively. For example,

$$\mathrm{RV}_t^{(W)} = \frac{1}{5}\sum_{i=0}^{4} \mathrm{RV}_{t-i}. \tag{13}$$

The ARFIMA$(1, d, 1)$–RV model is specified as

$$(1-\phi L)(1-L)^d(\mathrm{RV}_t - \mu) = (1+\theta L)\varepsilon_t, \tag{14}$$

where $(1-L)^d = 1 - dL - d(1-d)L^2/2! - d(1-d)(2-d)L^3/3! - \ldots$ and the roots of the AR and moving average polynomials lie outside the unit circle. Daily RV forecasts over the coming month are generated recursively and then aggregated to form the monthly RV. This is combined with the monthly VS rate at time $t$ to create the VRP forecast via Equation (1).

### 2.1.3 | Model averaging

Finally, we construct model average forecasts separately for each of the model classes detailed above. Namely, (i) pure time-series models (i.e., models with no predictors), (ii) models for each monthly predictor, (iii) PC models, and (iv) MIDAS models.[12] Only models deemed to have EPA via the MCS will be included in the respective averages. Given the conservative nature of the MCS, we consider model averages using 10% and 25% levels of significance. The model average forecasts for each model class are computed by equally weighting the individual one-step ahead forecasts from the models in the MCS for that class.[13]

[12]Initially, we set out to compute the MCS and hence a model-averaged forecast for the hybrid models presented in Section 2.1.2. However, as it will be apparent below, there was never a case where more than one hybrid model was in the MCS. Consequently, a model average forecast was not constructed for the hybrid models. Instead, their individual forecast performance is compared with the model combinations from the other model classes.

[13]Timmermann (2006) and D. E. Rapach et al. (2010) demonstrate the benefits of using equally weighted forecast combinations for predicting the equity risk premium. Intuitively, using equal weights avoids an additional layer of estimation, which could potentially induce estimation error and thus deteriorate the OOS performance.

## 3 | DATA

There is no publicly available source of variance swaps data as they are OTC products. Fortunately, absent data on traded swaps, the variance swap rate, $VS_{t,T}$, can be estimated by any measure of expected variance under $Q$. A large portion of the literature, starting with Carr and Wu (2009), relies on option-implied variance, $IV_t^T$. We follow this literature and extract synthetic variance swap rates from options data.

Specifically, we employ options written against: the S&P500 index, US T-bond futures, gold futures, and West Texas Intermediate crude oil futures. We employ the midpoint of the bid and ask for options that span a 30-day maturity. To overcome the absence of a continuum of options with strikes from 0 to ∞, the observed option prices are translated into implied volatilities via the Black–Scholes–Merton (BSM) model. To create a finer equally spaced grid, a cubic spline is then fit to implied BSM volatilities with an extrapolation beyond the endpoints assuming constant endpoint volatility. We then interpolate the implied volatilities for both maturities in the strike dimension. The interpolated implied volatilities are then back transformed into dollar option prices, retaining put prices for strikes below the current futures price $K_0$, and call prices for strikes above the current futures price. A smoothed version of the Bakshi and Kapadia (2003) methodology is used to compute the price of two log-return contracts. The annualized variance swap rate is equal to

$$VS_{t,T} = \frac{V - U^2}{T}, \tag{15}$$

where the first of the two contracts pays $V$, the squared log return on the underlying asset (spot for the S&P500 or the futures for the other assets) and the second contract pays $U$, the log return on the underlying asset.[14] Finally, we calculate realized variances $RV_{t,T}$ as the sum of squared 5 min returns using the S&P500 spot index and the nearby futures contract for the other assets. The final data set commences January 1, 1996 and ends July 31, 2018 (271 monthly observations).[15]

Figure 1 plots the payoffs (PL) for the four asset classes from January 1996 to July 2018. Payoffs behave very differently from traditional financial market returns. Payoffs appear to be characterized by regime switches between low volatility periods with a consistently negative payoff (or positive payoff for the swap seller) and high volatility episodes where the payoff can spike significantly—with the positive spikes representing a significant loss or negative tail event for the swap seller. These visual inspections motivate our use of the breakpoint, threshold regression, and MS models in addition to linear specifications.

Descriptive statistics in Table 1 show that even without conditioning on the market state, swap sellers generally make significant risk-adjusted payoffs, as measured by their Sharpe ratios (SRs). The ability to forecast swap payoffs and switches into and out of the high volatility states could significantly improve these risk-adjusted payoffs and facilitate better hedging and asset allocation decisions. These features help motivate the empirical investigation below.

[14]Specifically,

$$\begin{aligned} V = E_t^Q\left[\log^2\left(\frac{S}{S_0}\right)\right] &= \log^2\left(\frac{K_0}{S_0}\right) + 2\log\left(\frac{K_0}{S_0}\right)\left(\frac{F_0 - K_0}{K_0}\right) \\ &+ 2e^{rT}\int_0^{K_0} \frac{1 - \log\left(\frac{K}{S_0}\right)}{K^2} P(K)\,dK + 2e^{rT}\int_{K_0}^{\infty} \left|\frac{1 - \log\left(\frac{K}{S_0}\right)}{K^2}\right| C(K)\,dK \end{aligned} \tag{16}$$

and

$$\begin{aligned} U = E_t^Q\left[\log\left(\frac{S}{S_0}\right)\right] &= \log\left(\frac{K_0}{S_0}\right) + \frac{F_0 - K_0}{K_0} \\ &- e^{rT}\int_0^{K_0} \frac{1}{K^2} P(K)\,dK - e^{rT}\int_{K_0}^{\infty} \frac{1}{K^2} C(K)\,dK. \end{aligned} \tag{17}$$

[15]For robustness, we also consider the alternative formula for option-implied variance of Britten-Jones and Neuberger (2000). This is because the Bakshi, Kapadia and Madan risk-neutral variance is the expected holding period variance (Du & Kapadia, 2012), not the variance of the realized 1-month return, which is unobservable (Andersen et al., 2015). Implied variance measures for all assets are virtually identical, with correlations exceeding 0.998.

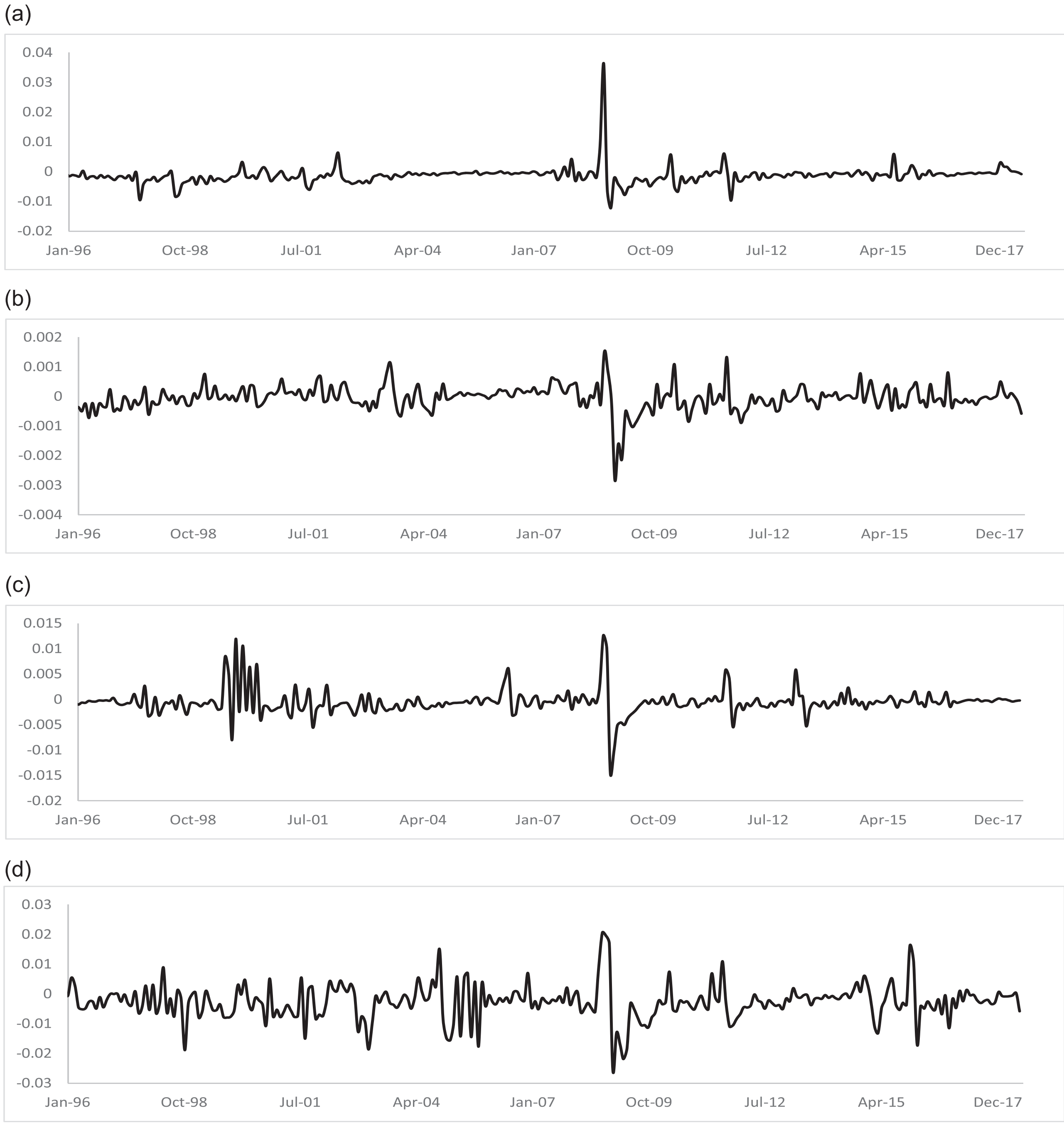


**FIGURE 1** Monthly payoffs: (a) S&P500, (b) bonds, (c) gold, and (d) oil

## 3.1 | Predictors

We collect a comprehensive set of predictive variables relied upon by the literature. These variables have typically been utilized to forecast assets excess returns: as such, they have been shown to track fluctuations in systematic risk for the respective asset classes.[16] Consequently, they appear to be natural candidates for predicting the VRP (and hence

[16]The predictors we use for the equity VRP have become standard in the equity premium predictability literature since the work of Welch and Goyal (2008). The predictors for the bond VRP follow the bond risk premia predictability literature, such as Cochrane and Piazzesi (2005) and Ludvigson and Ng (2009). The predictors for gold VRP and oil VRP relate to the literature on commodity excess return predictability. See, among others, Bakshi et al. (2011), Chen et al. (2010), Gao and Nardari (2018), and Gargano and Timmermann (2014).

**TABLE 1** Descriptive statistics

| | Monthly swap payoffs | | | |
|---|---|---|---|---|
| | **S&P500** | **Bond** | **Gold** | **Oil** |
| Mean | −0.0013 | −0.0001 | −0.0005 | −0.0027 |
| Median | −0.0012 | −0.0001 | −0.0007 | −0.0026 |
| Maximum | 0.0363 | 0.0015 | 0.0127 | 0.0206 |
| Minimum | −0.0123 | −0.0028 | −0.0146 | −0.0256 |
| Standard deviation | 0.0032 | 0.0004 | 0.0025 | 0.0060 |
| Skewness | 6.063 | −1.093 | 1.023 | 0.060 |
| Kurtosis | 76.258 | 11.898 | 13.889 | 6.021 |
| Jarque–Bera | 62,259*** | 948*** | 1386*** | 103*** |
| Sharpe ratio | −0.418 | −0.133 | −0.208 | −0.451 |

*Note*: Reports descriptive statistics for the monthly payoffs from January 1996 to July 2018. J.B. denotes the Jarque–Bera test for normality. *** denotes significance at the 1% level (Jarque–Bera test only).

variance swap payoffs), as the VRP is compensation for bearing systematic risk and, as discussed below, it may be characterized by time-variation and countercyclicality.

### 3.1.1 | Economic justification

Several contributions rationalize the links of the VRP with macroeconomic and business conditions. For instance, Bollerslev et al. (2011) consider an economy where asset prices follow affine stochastic volatility models, the volatility risk premium is linear in the level of volatility and a representative agent has constant relative risk aversion utility. They show that if in equilibrium the representative agent holds the market portfolio, then the volatility risk premium is directly proportional to the agent's relative risk aversion coefficient. The authors proceed to argue that the relation still applies (in an approximate sense) if risk aversion displays time-variation, possibly induced by state-dependent preferences as in the framework of Gordon and St-Amour (2004). The latter study proposes a modeling development that leads to countercyclicality in risk aversion and provides empirical evidence supporting the theoretical prediction. Kim (2014) also finds strong empirical support for countercyclical risk aversion, although within a recursive utility framework.

A separate stream of literature identifies the economic underpinnings of time-variation and countercyclicality in aggregate market variance. As summarized by Paye (2012), those features originate from either: (a) uncertainty about future economic prospects (the true economic states), induced by countercyclical shocks to fundamentals as in long-run risk models[17] or by learning (Timmermann, 1993, 1996); or (b) uncertainty about expected returns as in Mele (2007).[18]

It remains an empirical question whether the countercyclicality of the VRP is sufficiently pronounced to be captured and anticipated by variables that proxy for uncertainty on the state of the economy and/or on expected returns. It is also an empirical question whether the noise that separates ex ante and ex post VRP masks the predictive power of the chosen variables. In our investigation we aim to tackle those empirical challenges.

### 3.1.2 | Selection of variables

We attempt to be as comprehensive as possible within the set of predictive variables, rather than arbitrarily selecting a few predictors from previous studies. Together with the forecast combination method detailed above, these choices aim

[17]Long-run risk models assume that the volatilities of dividends and consumption growth are countercyclical. It follows that market volatility also displays countercyclical variations.

[18]In Mele's (2007) framework, expected returns in bad times are relatively more sensitive to fluctuations in a state variable that captures the state of the economy than in good times. If the asymmetry is sufficiently pronounced, the price dividend ratio is increasing and concave in the state variable, and thus market volatility increases relatively more on the downside.

to limit the impact of model misspecification and the concerns of data mining. The predictive variables and their sources are listed in Appendix A.

For each predictor, ADF and KPSS tests were used to determine their order of integration. When the two tests provided inconsistent evidence, specifications using the predictor in both levels and first differences were employed.

Given the large number of predictors under consideration, we consider model specifications that include only one predictor at a time, except for the models that rely on PCs (please see below). To exemplify, for the equity VRP we rely on 14 predictors: consequently, we estimate 14 different bivariate VAR specifications, each containing the swap payoff and one of the predictors. Similarly for the other models (breakpoint, threshold, MS, and MIDAS).

While a number of these variables are available daily, several others are only provided monthly and reported with a lag. For example, many US macroeconomic variables for a given month are not available until the middle of the next month. We therefore separate the predictor variables according to whether they are available daily and those available monthly (with a lag). To illustrate, consider a swap seller at the end of February 2000, seeking to predict the payoff on the swap (or the VRP) over the coming month. The daily variables will be available up to the end of February 2000 and so they can be used to create a monthly predictor as at the end of February for the swap payoff in March. On the other hand, for the predictors available only at a monthly frequency (e.g., US unemployment rate), the investor will use their values as of the end of January 2000 to generate swap payoff forecasts for March 2000.

We follow a similar approach for models using PCs. Namely, we construct two separate sets of PCs: one extracted from all the monthly time series of predictors constructed using daily data and the other extracted from the time series that are only available monthly with a lag. For the first set of PCs, the forecast as of the end of February 2000 will condition on the PCs extracted as at the end of February. For the second set of PCs, the forecast as at the end of February 2000 will condition on the PCs as of the end of January.

## 4 | EMPIRICAL RESULTS

### 4.1 | In-sample fit

Before conducting the OOS forecasting exercise, we aim for a first model characterization of the swap payoff dynamics by estimating, over the entire sample, the pure time-series models illustrated in Section 3. In the interests of space, we do not report estimates for all models which (including models with predictors) number several hundreds across the four assets. Details for all models are available on request.

Two main results arise from the in-sample estimation. First, there is strong evidence supporting structural breaks in payoffs: the breakpoint, threshold regression, and MS models systematically show a better fit than linear ARMA or PC models.

Second, among the models with structural breaks, the two-regime MS model consistently provides the lowest Akaike information criterion (AIC) and Schwarz information criterion (SIC) for all four asset classes.[19,20] We therefore present detailed results for these MS models in Table 2.

MS models identify significant differences in the AR coefficients between high ($\delta_{0,2}$) and low volatility ($\delta_{1,2}$) states. The low volatility states all have positive AR coefficients. The high volatility states tend to have either statistically insignificant (S&P500, gold), or positive but lower magnitude AR coefficients (oil). Bonds are the exception with a larger positive AR coefficient in the high volatility state. The transition probabilities indicate that, with the possible exception of the oil VRP, the low volatility state is much more persistent than the high volatility state ($p11 > p22$).

To examine the reasonableness of the regimes, Figure 2 plots the smoothed probabilities of being in the low volatility states (regime 1). Results suggest that variance swap payoffs go through sustained periods of time where

[19]Three state MS models were also estimated. Testing the number of regimes via MS models is complicated given the presence of unidentified nuisance parameters under the null. This may be overcome via a Davies procedure. Instead we rely on the AIC to assess the optimal number of regimes (Psaradakis & Spagnolo, 2003).

[20]When estimating the three-state models many different starting values were required as a number of models failed to converge. These problems were exacerbated over the shorter estimation windows required for the OOS analysis, especially for the bond and gold VRPs. These issues were also evident when the three state MS models were extended to include predictor variables. Given our desire to re-estimate these models a large number of times (namely, 136 times as detailed below) with an expanding window, and given the information provided by the AIC, we decided to proceed with the two-state MS models.

TABLE 2 Markov switching models

| | S&P500 | | Bonds | | Gold | | Oil | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error |
| *Regime 1: Low volatility* | | | | | | | | |
| $\delta_{0,1}$ | −0.001*** | 0.000 | 0.000 | 0.000 | −0.001*** | 0.000 | −0.001*** | 0.000 |
| $\delta_{0,2}$ | 0.512*** | 0.049 | 0.318*** | 0.077 | 0.227*** | 0.064 | 0.606*** | 0.127 |
| $\log(\sigma_1)$ | −6.975*** | 0.063 | −8.186*** | 0.064 | −7.172*** | 0.083 | −6.215*** | 0.138 |
| Uncond mean | $-1.24e^{-03}$ | | $-3.14e^{-05}$ | | $-6.55e^{-04}$ | | $-2.49e^{-03}$ | |
| *Regime 2: High volatility* | | | | | | | | |
| $\delta_{1,1}$ | −0.001 | 0.001 | 0.000 | 0.000 | 0.000 | 0.001 | −0.002*** | 0.001 |
| $\delta_{1,2}$ | 0.148 | 0.163 | 0.448** | 0.185 | 0.009 | 0.137 | 0.209** | 0.095 |
| $\log(\sigma_2)$ | −4.908*** | 0.123 | −7.028*** | 0.194 | −5.279*** | 0.106 | −4.811*** | 0.083 |
| Uncond mean | $-1.18e^{-03}$ | | $-2.66e^{-04}$ | | $-2.87e^{-05}$ | | $-3.824e^{-04}$ | |
| *Transition matrix parameters* | | | | | | | | |
| *p*11 | 0.942 | | 0.974 | | 0.939 | | 0.874 | |
| *p*22 | 0.686 | | 0.770 | | 0.787 | | 0.848 | |
| *Diagnostics* | | | | | | | | |
| LL | 1357.2 | | 1769.5 | | 1381.6 | | 1066.9 | |
| SIC | −9.924 | | −12.990 | | −10.106 | | −7.766 | |
| AIC | −10.031 | | −13.096 | | −10.213 | | −7.873 | |

*Note*: Reports two-state Markov switching models fit to monthly variance swap payoffs from January 1996 to June 2018 (the last model in the expanding estimation window). $\delta_{j,1}$ and $\delta_{j,2}$ denote the constant and first-order autoregressive coefficient for each state, respectively. Uncond mean is the unconditional mean estimate implied by the parameter estimates. AIC, Akaike information criterion; LL, log-likelihood; SIC, Schwarz information criterion. *** and ** denote significance at the 1% and 5% levels, respectively.

payoffs are stable and highly persistent, suggesting that the VRP could be harvested. This is consistent with the reported unconditional mean estimates implied by the parameter estimates for each regime. For all assets (excluding bonds), the unconditional mean is more negative in the low volatility state than the high volatility state. This suggests that on average, higher returns from selling swaps are achieved when VRP volatility is low.[21]

For comparative purposes we report selected PC model estimates in Table 3. Models using the first PC of monthly predictors available at the daily and monthly frequencies are reported. We present AR(1) versions of the models as they provide a better fit than otherwise equivalent models without AR dynamics. $\delta_0$, $\delta_1$, and PC1(−1) denote the constant, AR(1), and lagged first PC coefficients, respectively.

With the exclusion of bonds, the adjusted $R^2$ statistics are modest ranging from 1% to 9%. The AIC and SIC for each model are also worse than the MS models reported in Table 2. It is well understood that MS models superior in-sample fit do not always translate into superior OOS forecast performance. This is largely due to the difficulty associated with forecasting regime switches (Dacco & Satchell, 1999). We leave this model comparison for Section 4.2.

[21]Breakpoint models consistently identify multiple breakpoints with significant variation in the coefficient estimates through time. AR coefficients demonstrate significant periods of time where the payoffs have high levels of persistence (i.e., highly positive and statistically significant autoregressive coefficient estimates). This is, again, consistent with long periods where the payoff is predictable. Other periods are also identified where the AR coefficients are highly negative, statistically significant, and indicate predictable variations in the payoffs.

Two-state threshold regression models find that when the payoff is negative the AR term is positive and highly persistent. When the payoff is positive the AR term is negative and much less persistent. Like the breakpoint models, this is consistent with long periods where the payoff is predictable and the risk premium can be harvested. Threshold regression models with >2 states tend to find a region where the payoff is approximately zero (or just negative) and the AR term is significantly negative. Outside of this region the AR terms are high and positive or high and negative, again suggesting high degrees of predictability.

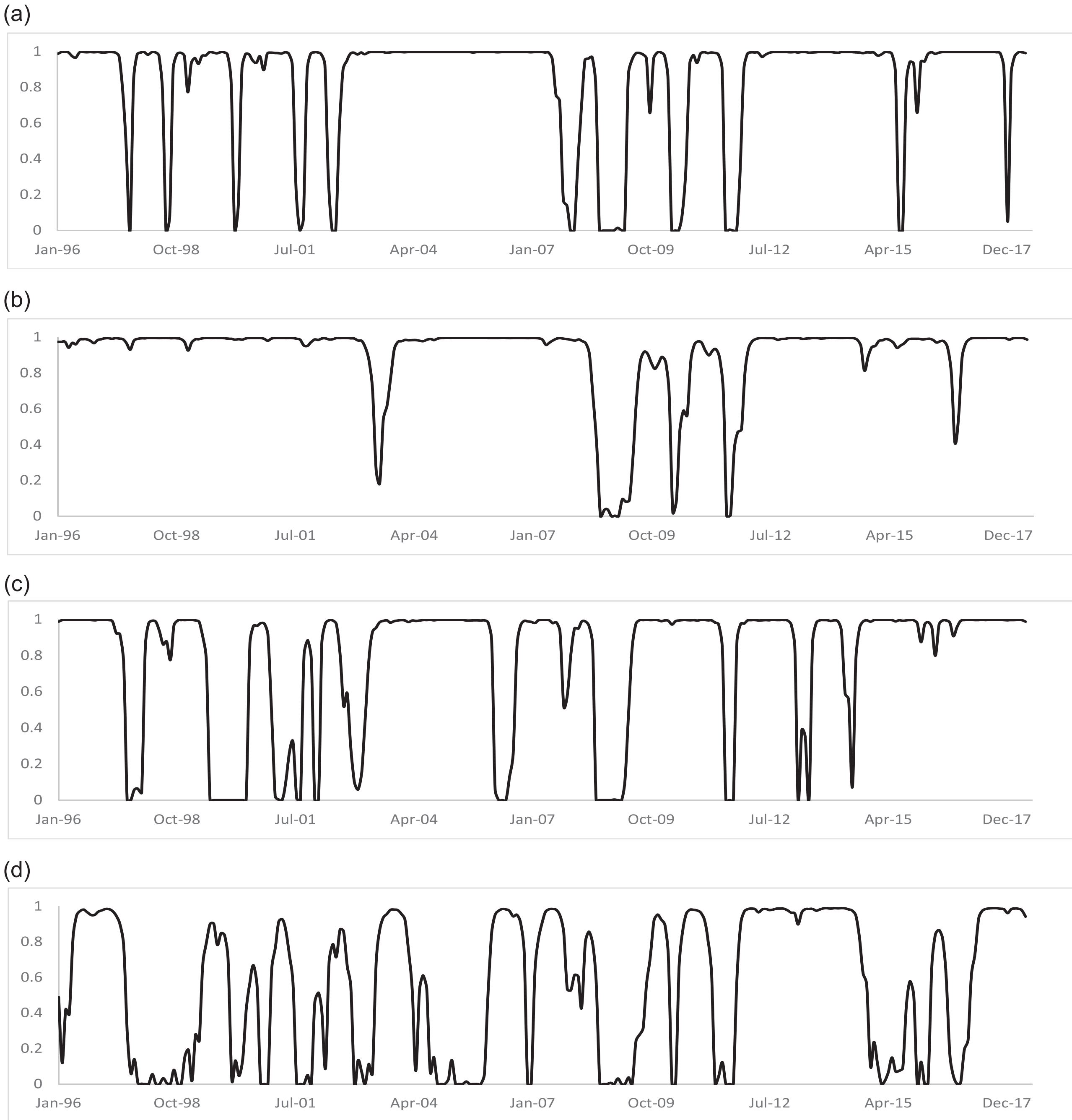


**FIGURE 2** Markov switching model probability of being in the low volatility state: (a) S&P500, (b) bonds, (c) gold, and (d) oil

Finally, we also report the ARFIMA(1, $d$, 1) and HAR models fit to the daily log realized variances. For a given time series, both models report similar goodness of fit (adjusted $R^2$, AIC, and SIC). The ARFIMA estimates of $d$ are consistent with the literature and support the presence of long memory. HAR estimates are also consistent with expectations with all RV coefficient estimates positive and statistically significant (Tables 4 and 5).[22]

[22]The link between regime switching and long memory is well recognized (Banerjee & Urga, 2005). Both processes can be easily confused, with more recent literature developing tests for long memory versus breaks (Qu, 2011). We employ the MS models for the direct modeling of the VRP given the limited number of observations (271 monthly observations). When forecasting, the first estimation window only has 135 observations, which is inadequate when estimating a long memory model. In contrast, we employ the ARFIMA and HAR models for the daily RVs, given that long memory processes often forecast as well or better than MS models because they do not require regime forecasts (Diebold & Inoue, 2001).

**TABLE 3** PC1-AR models

| | SP500 | | Bonds | | Gold | | Oil | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error |
| *Daily* | | | | | | | | |
| $\delta_0$ | 0.0004 | 0.0004 | 0.0000 | 0.0000 | −0.0005*** | 0.0002 | −0.0022*** | 0.0004 |
| $\delta_1$ | 0.3397*** | 0.0657 | 0.3870*** | 0.0563 | 0.0859 | 0.0626 | 0.2122*** | 0.0619 |
| PC1(−1) | $-3.8e^{-04}$*** | $9.7e^{-05}$ | $-2.4e^{-03}$* | $1.4e^{-03}$ | $3.5e^{-03}$** | $1.8e^{-03}$ | $-1.1e^{-02}$*** | $4.2e^{-03}$ |
| *Diagnostics* | | | | | | | | |
| Adjusted $R^2$ | 0.0943 | | 0.1564 | | 0.0100 | | 0.0892 | |
| AIC | −8.7528 | | −12.8384 | | −9.1036 | | −7.4795 | |
| SIC | −8.7126 | | −12.7982 | | −9.0635 | | −7.4394 | |
| *Monthly* | | | | | | | | |
| $\delta_0$ | 0.0010* | 0.0006 | 0.0002*** | 0.0001 | 0.0013*** | 0.0005 | −0.0005 | 0.0011 |
| $\delta_1$ | 0.2102*** | 0.0584 | 0.3549*** | 0.0564 | 0.0210 | 0.0602 | 0.2634*** | 0.0591 |
| PC(−1) | $-1.7e^{-05}$*** | $4.5e^{-06}$ | $5.0e^{-06}$*** | $1.5e^{-06}$ | $-1.6e^{-05}$*** | $4.0e^{-06}$ | $-1.3e^{-05}$ | $9.3e^{-06}$ |
| *Diagnostics* | | | | | | | | |
| Adjusted $R^2$ | 0.0915 | | 0.1825 | | 0.0535 | | 0.0711 | |
| AIC | −8.7497 | | −12.8698 | | −9.1449 | | −7.4565 | |
| SIC | −8.7095 | | −12.8296 | | −9.1047 | | −7.4163 | |

*Note*: Reports model estimates for the PC(1) model with an AR(1) coefficient fit to monthly variance swap payoffs from January 1996 to June 2018 (the last model in the expanding estimation window). The daily models are fit to the first PC extracted using data that are available daily. The monthly models are fit to the first PC using data that are available monthly and are reported with a 1-month lag. $\delta_0$ and $\delta_1$ denote the constant and first-order autoregressive coefficients. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Abbreviations: AIC, Akaike information criterion; AR, autoregressive; PC, principal component; SIC, Schwarz information criterion.

## 4.2 | OOS results

In this section we first determine the MCS for each of the model classes introduced in Section 2.1. Next, we construct model average forecasts separately for each model class, as illustrated in Section 2.1.3. We then compare the model-averaged forecasts with a few simple benchmarks. Finally, we consider the economic relevance for the various forecasts by analyzing some simple variance swap trading strategies.

In all cases, OOS forecasts are generated using an expanding window. We use the first half of the sample to estimate the models conditional on the information available on March 31, 2007. The models are then used to forecast the payoff for a 1-month ahead variance swap sold as of March 31, 2007. At the end of the month, the payoff is observed. The models are then re-estimated using the additional information available and the forecasts reperformed. This is continued until the end of the data set, generating 136 OOS observations.[23]

### 4.2.1 | MCS across model classes

We determine the MCS for the four separate model classes introduced earlier: (i) pure time-series models (i.e., models with no predictors), (ii) models for each monthly predictor, (iii) PC models, and (iv) univariate MIDAS models. Given

[23]Given that the in-sample analysis reveals the importance of structural breaks, we seek to evaluate the forecast performance of this model class (breakpoint, threshold, and MS) relative to linear or state invariant models. Given our limited number of total observations (271) and the length of the low volatility regimes, rolling windows may not identify regime switches over short time spans. Under these circumstances expanding windows are the most logical way to proceed.

**TABLE 4** ARFIMA(1, *d*, 1) estimates

| | SP500 | | Bonds | | Gold | | Oil | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error |
| $c$ | −9.897*** | 1.149 | −10.768*** | 0.773 | −10.315* | 5.400 | −8.518 | 5.824 |
| $d$ | 0.466*** | 0.023 | 0.476*** | 0.027 | 0.494*** | 0.015 | 0.495*** | 0.011 |
| $\phi$ | 0.485*** | 0.087 | 0.477*** | 0.021 | 0.347*** | 0.024 | 0.470*** | 0.018 |
| $\theta$ | −0.392*** | 0.079 | −0.812*** | 0.015 | −0.687*** | 0.022 | −0.804*** | 0.013 |
| $\sigma^2$ | 0.327*** | 0.003 | 1.290*** | 0.015 | 1.191*** | 0.017 | 1.369*** | 0.017 |
| *Diagnostics* | | | | | | | | |
| Adjusted $R^2$ | 0.731 | | 0.148 | | 0.354 | | 0.204 | |
| AIC | 1.722 | | 3.094 | | 3.015 | | 3.154 | |
| SIC | 1.726 | | 3.098 | | 3.020 | | 3.158 | |

*Note*: Reports model estimates for the ARFIMA(1, *d*, 1) model fit to daily realized volatility (in logs) from January 4, 1996 to June 30, 2018. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Abbreviations: AIC, Akaike information criterion; SIC, Schwarz information criterion.

**TABLE 5** HAR model estimates

| | S&P500 | | Bonds | | Gold | | Oil | |
|---|---|---|---|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error | Coefficient | Standard error |
| $c$ | −0.598*** | 0.067 | −2.638*** | 0.263 | −1.518*** | 0.155 | −1.987*** | 0.151 |
| $\beta^{(D)}$ | 0.564*** | 0.012 | 0.151*** | 0.012 | 0.137*** | 0.014 | 0.151*** | 0.012 |
| $\beta^{(W)}$ | 0.152*** | 0.018 | 0.185*** | 0.032 | 0.291*** | 0.028 | 0.120*** | 0.030 |
| $\beta^{(M)}$ | 0.229*** | 0.015 | 0.437*** | 0.039 | 0.453*** | 0.029 | 0.527*** | 0.032 |
| *Diagnostics* | | | | | | | | |
| Adjusted $R^2$ | 0.726 | | 0.123 | | 0.325 | | 0.193 | |
| AIC | 1.741 | | 3.122 | | 3.059 | | 3.166 | |
| SIC | 1.745 | | 3.126 | | 3.063 | | 3.170 | |

*Note*: Reports model estimates for the HAR model fit to daily realized volatility (in logs) from January 4, 1996 to June 30, 2018. ***, **, and * denote statistical significance at the 1%, 5%, and 10% levels, respectively.

Abbreviations: AIC, Akaike information criterion; HAR, heterogeneous autoregressive; SIC, Schwarz information criterion.

the conservative nature of the MCS, we consider model averages using 10% and 25% levels of significance as recommended by Hansen et al. (2011).

(i) *Pure time series*

Table 6 reports losses and MCS results for the eight pure time-series models considered.[24] Across assets, no model is clearly dominant as most, if not all models are consistently in the MCS. At the 25% confidence level (and hence the less restrictive 10% level), all models appear to have equal predictive abilities for the bond and oil VRPs. For the S&P500 and gold, all models are also in the MCS at the 10% level. For gold, even though not all models are in the MCS at the 25% level, the linear ARMA(1, 1) specification is, meaning that it has EPA with models that allow for breaks and different regimes. Overall, this evidence leads to a first important conclusion. Namely, while nonlinear specifications with

[24]Please note the null hypothesis is that the remaining models have EPA. Therefore, failure to reject means the remaining models are in the MCS. A model in the MCS at 10% (25%) is included if the $p$ value for the model is ≥10% (≥25%). Models with one star therefore have $p$ values ≥10% but <25%. Models with two stars have $p$ values ≥25%. Models at the 25% level are therefore a subset of the models included at the 10% level.

**TABLE 6** Out-of-sample MSE: Pure time-series models

| Model | S&P500 | Bonds | Gold | Oil |
|---|---|---|---|---|
| ARMA(1, 1) | 0.750* | 0.0024** | 0.074** | 0.374** |
| Breakpoint ($\sigma_0 = \cdots \sigma_m$) | 0.744* | 0.0024** | 0.078** | 0.447** |
| Breakpoint ($\sigma_0 \neq \cdots \sigma_m$) | 0.197** | 0.0024** | 0.080* | 0.441** |
| Threshold ($d \in R : 1 \leq d \leq 5, \sigma_0 = \cdots \sigma_m$) | 0.757* | 0.0023** | 0.124* | 0.338** |
| Threshold ($d = 1, \sigma_0 = \cdots \sigma_m$) | 0.203** | 0.0023** | 0.076** | 0.400** |
| Threshold ($d \in R : 1 \leq d \leq 5, \sigma_0 \neq \cdots \sigma_m$) | 0.270* | 0.0023** | 0.122* | 0.362** |
| Threshold ($d = 1, \sigma_0 \neq \cdots \sigma_m$) | 0.278* | 0.0023** | 0.074** | 0.367** |
| MS ($\sigma_0 \neq \sigma_1$) | 0.278* | 0.0024** | 0.105* | 0.377** |

*Note*: Reports out-of-sample MSE × 10,000 from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. Note $M^*_{0.25} \subset M^*_{0.10}$. Breakpoint and threshold regression determines breakpoints using a sequential test of $l + 1$ versus $l$ breaks. MS denotes a two-state Markov switching model.

Abbreviations: ARMA, autoregressive moving average; MCS, Model Confidence Set; MSE, Mean-Squared Error.

**TABLE 7** Initial models estimated for each monthly predictor

| Model type | Model number | State dependent ($Z_t$) | State invariant ($X_t$) | Other |
|---|---|---|---|---|
| VAR(1) | 1 | – | vrp(−1), x(−1) | – |
| Breakpoint | 2 | vrp(−1), x(−1) | – | $\sigma_0 = \cdots \sigma_m$ |
| | 3 | vrp(−1) | x(−1) | $\sigma_0 = \cdots \sigma_m$ |
| | 4 | vrp(−1), x(−1) | – | $\sigma_0 \neq \cdots \sigma_m$ |
| | 5 | vrp(−1) | x(−1) | $\sigma_0 \neq \cdots \sigma_m$ |
| Threshold | 6 | vrp(−1), x(−1) | – | $d \in R : 1 \leq d \leq 5, \sigma_0 = \cdots \sigma_m$ |
| | 7 | vrp(−1) | x(−1) | $d \in R : 1 \leq d \leq 5, \sigma_0 = \cdots \sigma_m$ |
| | 8 | vrp(−1), x(−1) | – | $d = 1, \sigma_0 = \cdots \sigma_m$ |
| | 9 | vrp(−1) | x(−1) | $d = 1, \sigma_0 = \cdots \sigma_m$ |
| | 10 | vrp(−1), x(−1) | – | $d \in R : 1 \leq d \leq 5, \sigma_0 \neq \cdots \sigma_m$ |
| | 11 | vrp(−1) | x(−1) | $d \in R : 1 \leq d \leq 5, \sigma_0 \neq \cdots \sigma_m$ |
| | 12 | vrp(−1), x(−1) | – | $d = 1, \sigma_0 \neq \cdots \sigma_m$ |
| | 13 | vrp(−1) | x(−1) | $d = 1, \sigma_0 \neq \cdots \sigma_m$ |
| MS | 14 | vrp(−1), x(−1) | – | 2 states, $\sigma_0 = \sigma_1$ |
| | 15 | vrp(−1) | x(−1) | 2 states, $\sigma_0 \neq \sigma_1$ |

*Note*: vrp(−1) denotes the payoff (PL) on the variance swap lagged 1 month, x(−1) denotes the predictor lagged 1 month. Breakpoint and threshold regression determines breakpoints using a sequential test of $l + 1$ versus $l$ breaks. The following models were selected to evaluate the Model Confidence Set (MCS) for all predictors: VAR(1), Model 1; Breakpoint regression, Model 5; Threshold regression, Model 13; Markov switching (MS) regression, Model 15.

structural breaks provide a better characterization of the VRP dynamics in the sample, this does not translate into an improvement in OOS forecasting performance relative to a simple ARMA(1, 1) process.

(ii) *Models with predictors*

Table 7 summarizes the 15 models initially considered for each monthly predictor. All models regress the payoff against the variables which are separated according to whether they have coefficients that are state dependent or time invariant.

On considering the MCS across the 15 models for each predictor separately, state-dependent variances ($\sigma_0 \neq \cdots \sigma_m$) generally provided better or comparable forecasts to state invariant variances ($\sigma_0 = \cdots \sigma_m$). State-dependent parameters for the predictors provided no forecasting improvement. For the threshold regressions, a delay of one ($d = 1$) was generally comparable or better than an approach that selected the delay by minimizing the sum of squared residuals. The following models for each predictor were therefore considered for inclusion in the MCS: (i) VAR(1) (Model 1), (ii) Breakpoint regression (Model 5), (iii) Threshold regression (Model 13), and (iv) MS (Model 15).

Tables 8 and 9 report MCS results for the S&P500, bond, gold, and oil VRP payoffs for each of the four models selected in Table 7. Table 8 reports results for the predictors in levels, and Table 9 reports results for the variables in first differences. Gold and oil threshold regressions did not identify any breaks, so they were removed from the analysis as forecasts were identical to VAR(1) forecasts. The MCS for each asset is calculated at the 10% and 25% levels using the models in both tables.

At the 10% level of significance, all models are in the MCS for the S&P500, gold, and oil, indicating that (a) no single predictor possesses superior forecasting ability; and (b) the linear VAR has similar forecasting performance to models that allow for structural breaks.[25] This is not the case for bonds, where only equity market volatility (EMV) and a few of the term structure factors generate forecasts in the MCS.

At the 25% level though, a few important differences emerge. For the S&P500, four out of the five models in the MCS allow for structural breaks, with the dividend payout ratio, the inflation rate, economic policy uncertainty (epu-3), and EMV, displaying better predictive performance than the remaining predictors. For bonds, five out of the six models in the MCS are MS models. The remaining model is a breakpoint regression. For gold, the breakpoint models are dominant appearing in the MCS 18 times, while the only other model to be included is the VAR, which was included 12 times. For oil, only the breakpoint and MS models are in the MCS. Like gold, the breakpoint model is dominant as it is in the MCS 18 times with the MS model only included five times.

Overall, the more restrictive test supports the structural break models with predictors relative to their linear counterparts. It remains to be seen whether this holds after combining forecasts of the models in the MCS.

(iii) *PC models*

Table 10 reports the results for the PCs models. Results with PCs calculated using the predictors that are available daily are reported in the *Daily* columns.[26] Results with the PCs calculated using the predictors available monthly (and thus reported with a 1-month lag) are reported in the *Monthly* columns. The MCS for each asset is calculated using the results in both columns.

The following notation is used to characterize the different PC specifications, detailed in Section 2.1.1. PC($k$) for $k = 1, \ldots, 6$ indicates a regression that includes all PCs from PC1 to PC($k$). Only 1 lag of each PC is employed. PC($k$)-AR is the same as PC($k$) however it also includes a lagged dependent variable. PC($k$)-brk is a breakpoint regression that includes all PCs from lag 1 to $k$ plus a lagged dependent variable. The specification allows for state-dependent AR coefficients and variances. The PC coefficients are state invariant and the maximum number of breaks is 5. PC($k$)-thr is a threshold regression that includes a lagged dependent variable plus all PCs from lag 1 to $k$. The AR terms are state dependent but the PC coefficients are state invariant. There is a maximum of five breaks with a trimming percentage of 15%. PC($k$)-ms is a two-state MS regression that includes all PCs from lag 1 to $k$ plus a lagged dependent variable. The model also allows for state-dependent variances and AR parameters but state invariant PC coefficients.

At the 10% level all models are in the MCS for the S&P500, bonds, and gold. This result obtains whether one uses only the predictors available daily or those available monthly (with a lag). Approximately 50% of the models are, on the other hand, included in the MCS for oil. Among those, there are linear specifications (e.g., PC1 and PC2) as well as specifications with breaks (e.g., PC1-AR, PC1-brk, PC1-thr, and PC1-ms).

At the 25% level, all models remain in the MCS for the S&P500. Gold and oil have a much smaller number of models in the MCS. For gold all the included models are linear (with the exception of PC2-brk), and without any AR terms. For oil almost all the better forecasting models are still linear but with an AR(1) component. The only exceptions are PC2-thr and PC1-ms. Like the S&P500, a simple linear model that employs the first PC performs well. The forecasting performance for bonds is somewhat different, as the only model in the 25% MCS employs the first PC but includes breaks in the constant and AR term.

Overall, the OOS analysis of the PC models delivers two consistent and important messages across all four asset classes: First, adding PCs beyond the first, provides no statistically significant improvement in forecasting power. Second, a simple linear model that regresses the VRP against the lagged first PC (constructed using either the daily or monthly data sets) appears to be sufficient in terms of forecasting accuracy as structural breaks, much like higher-order PCs, contain no additional predictive power.

[25]This result may also reflect the conservative nature of the MCS as well as the power of MSE being compromised in the presence of outliers.

[26]Given the limited number of bonds predictors available daily, we only extract one or two PCs from those variables.

**TABLE 8** Out-of-sample MSE: Monthly predictors (levels)

| | S&P500 | | Bonds | | | | Gold | Oil |
|---|---|---|---|---|---|---|---|---|
| | Predictor | Loss | Predictor | Loss | | Predictor | Loss | Loss |
| VAR | ret | 0.265* | ret | 0.0024 | VAR | aud | 0.082** | 0.450* |
| Breakpoint | | 0.778* | | 0.0024 | Breakpoint | | 0.078** | 0.374** |
| Threshold | | 0.265* | | 0.0024 | MS | | 0.110* | 0.414* |
| MS | | 0.284* | | 0.0023 | VAR | bdi | 0.081** | 0.429* |
| VAR | dp | 0.254* | emv | 0.0023 | Breakpoint | | 0.075** | 0.360** |
| Breakpoint | | 0.744* | | 0.0023** | MS | | 0.103* | 0.379** |
| Threshold | | 0.271* | | 0.0023 | VAR | ltr | 0.083** | 0.424* |
| MS | | 0.283* | | 0.0025 | Breakpoint | | 0.080** | 0.350** |
| VAR | dy | 0.256* | epu-3 | 0.0024 | MS | | 0.104* | 0.385* |
| Breakpoint | | 0.743* | | 0.0023 | VAR | cad | 0.086* | 0.426* |
| Threshold | | 0.257* | | 0.0023 | Breakpoint | | 0.084** | 0.351** |
| MS | | 0.281* | | 0.0022 | MS | | 0.086* | 0.402* |
| VAR | ep | 0.258* | epu-news | 0.0023 | VAR | chp | 0.087** | 0.417* |
| Breakpoint | | 0.743* | | 0.0023 | Breakpoint | | 0.082** | 0.347** |
| Threshold | | 0.276* | | 0.0024 | MS | | 0.109* | 0.396* |
| MS | | 0.305* | | 0.0023 | VAR | ret | 0.073** | 0.449* |
| VAR | de | 0.248** | *f*1 | 0.0023 | Breakpoint | | 0.077** | 0.373** |
| Breakpoint | | 0.749* | | 0.0023 | MS | | 0.106* | 0.392** |
| Threshold | | 0.259* | | 0.0024 | VAR | ds | 0.086** | 0.437* |
| MS | | 0.291* | | 0.0027 | Breakpoint | | 0.091** | 0.361** |
| VAR | tms | 0.251* | *f*1–3 | 0.0026 | MS | | 0.087* | 0.369** |
| Breakpoint | | 0.747* | | 0.0027 | VAR | emv | 0.093* | 0.453* |
| Threshold | | 0.264* | | 0.0028 | Breakpoint | | 0.083* | 0.386** |
| MS | | 0.310* | | 0.0021** | MS | | 0.113* | 0.420* |
| VAR | infl | 0.308* | *f*2 | 0.0024 | VAR | epu-3 | 0.080** | 0.440* |
| Breakpoint | | 0.804* | | 0.0024 | Breakpoint | | 0.125* | 0.374** |
| Threshold | | 0.322* | | 0.0023 | MS | | 0.097* | 0.424* |
| MS | | 0.231** | | 0.0023 | VAR | epu-news | 0.069** | 0.446* |
| VAR | ltr | 0.253* | *f*3 | 0.0023 | Breakpoint | | 0.071** | 0.375** |
| Breakpoint | | 0.756* | | 0.0024 | MS | | 0.086* | 0.374** |
| Threshold | | 0.266* | | 0.0023 | VAR | inr | 0.091* | 0.427* |
| MS | | 0.275* | | 0.0025 | Breakpoint | | 0.085** | 0.353** |
| VAR | dfr | 0.481* | *f*4 | 0.0024 | MS | | 0.107* | 0.406* |
| Breakpoint | | 1.086* | | 0.0024 | VAR | ip | 0.085* | 0.489* |
| Threshold | | 0.494* | | 0.0025 | Breakpoint | | 0.091* | 0.418* |
| MS | | 0.287* | | 0.0023 | MS | | 0.108* | 0.420** |
| VAR | epu-3 | 0.897* | *f*6 | 0.0024 | VAR | ms | 0.081** | 0.488* |
| Breakpoint | | 0.281* | | 0.0024 | Breakpoint | | 0.082** | 0.401** |

(Continues)

**TABLE 8** (Continued)

| | S&P500 | | Bonds | | | | Gold | Oil |
|---|---|---|---|---|---|---|---|---|
| | **Predictor** | **Loss** | **Predictor** | **Loss** | | **Predictor** | **Loss** | **Loss** |
| Threshold | | 0.407* | | 0.0024 | MS | | 0.105* | 0.427* |
| MS | | 0.264* | | 0.0023** | VAR | nzd | 0.083* | 0.452* |
| VAR | epu-news | 0.883* | $f7$ | 0.0024 | Breakpoint | | 0.078** | 0.376** |
| Breakpoint | | 0.261* | | 0.0024 | MS | | 0.114* | 0.398* |
| Threshold | | 0.366* | | 0.0023 | VAR | real-un | 0.087* | 0.457* |
| MS | | 0.283* | | 0.0023 | Breakpoint | | 0.079** | 0.391** |
| VAR | emv | 0.827* | $f8$ | 0.0024 | MS | | 0.105* | 0.403* |
| Breakpoint | | 0.374* | | 0.0024 | VAR | sar | 0.083** | 0.404* |
| Threshold | | 0.312* | | 0.0024 | Breakpoint | | 0.076** | 0.346** |
| MS | | 2.639* | | 0.0023 | MS | | 0.086* | 0.409* |
| VAR | real-un | 2.422* | real-un | 0.0025 | VAR | urate | 0.090* | 0.449* |
| Breakpoint | | 3.758* | | 0.0024 | Breakpoint | | 0.081** | 0.377** |
| Threshold | | 2.435* | | 0.0025 | MS | | 0.095* | 0.383** |
| MS | | 13.223* | | 0.0025 | | | | |
| VAR | vix | 0.289* | | | | | | |
| Breakpoint | | 0.801* | | | | | | |
| Threshold | | 0.303* | | | | | | |
| MS | | 0.429* | | | | | | |

*Note*: Reports out-of-sample MSE × 10,000 from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. The MCS is calculated for the above models plus the models in Table 9. Note $M^*_{0.25} \subset M^*_{0.10}$. Model details for each specification are provided in Table 7, where Model 1 (VAR), 5 (breakpoint), 13 (threshold), and 15 (Markov switching) are employed. Breakpoint and threshold regression determines breakpoints using a sequential test of $l+1$ versus $l$ breaks. MS denotes a two-state Markov switching model.

Abbreviations: MS, Markov switching; MSE, Mean-Squared Error; VAR, vector autoregression.

(iv) *MIDAS models*

The final set of models directly forecasting the VRP considers the MIDAS specification, where the monthly VRP is regressed against daily regressors that commence with a 1-month lag. Preliminary investigations found that, across all estimation windows, exponential Almon estimates were often unstable, sensitive to starting values, and did not converge. We therefore considered four analytically estimated MIDAS specifications. The first two are Step AR(1)–MIDAS models with step sizes of 15 and 25 lags. The third and fourth models are PDL AR(1)–MIDAS models with $P = 2$ and 3. All models select the optimal lag ($K$) for each estimation window by minimizing the residual sum of squares. Residual diagnostics demonstrate that the AR(1) process captures any remaining serial correlation.

Results generally found that for a given predictor, the null of equal OOS predictive ability across the four MIDAS models could not be rejected. If, for a given predictor, there was a difference between model forecasts, it was usually in favor of Step AR(1)–MIDAS with a step size of 25. For each daily predictor, we therefore report the MCS results for the Step AR(1)–MIDAS model with $P = 25$.[27]

Table 11 reports the MIDAS model results. Bonds, gold, and oil include virtually all models in the MCS at both the 10% and 25% levels, indicating that all daily predictors possess very similar predictive power for the respective VRP. The S&P500 is the exception as three daily predictors stand above all others: the dividend payout ratio (de), the long-term government bond yield (lty), and the S&P RV.

[27]As results from the other three specifications are very similar, we do not report them to conserve space and avoid duplications.

TABLE 9 Out-of-sample MSE: Monthly predictors (first differences)

| | S&P500 | | Bonds | | | | Gold | Oil |
|---|---|---|---|---|---|---|---|---|
| | **Predictor** | **Loss** | **Predictor** | **Loss** | | **Predictor** | **Loss** | **Loss** |
| VAR | *d*(dp) | 0.270* | *d*(emv) | 0.0023 | VAR | *d*(ec-act) | 0.077** | 0.452* |
| Breakpoint | | 0.779* | | 0.0024 | Breakpoint | | 0.074** | 0.367** |
| Threshold | | 0.271* | | 0.0023 | MS | | 0.106* | 0.379** |
| MS | | 0.285* | | 0.0026 | VAR | *d*(emv) | 0.080** | 0.449* |
| VAR | *d*(dy) | 0.252* | $f(f-12)$ | 0.0023 | Breakpoint | | 0.074** | 0.365** |
| Breakpoint | | 0.737* | | 0.0024 | MS | | 0.106* | 0.369** |
| Threshold | | 0.268* | | 0.0023 | VAR | *d*(fin-un) | 0.093* | 0.485* |
| MS | | 0.279* | | 0.0025 | Breakpoint | | 0.083** | 0.396* |
| VAR | *d*(ep) | 0.279* | $d(f-23)$ | 0.0023 | MS | | 0.108* | 0.389* |
| Breakpoint | | 0.729* | | 0.0023 | VAR | *d*(mac-un) | 0.094* | 0.474* |
| Threshold | | 0.279* | | 0.0023 | Breakpoint | | 0.083** | 0.390* |
| MS | | 0.282* | | 0.0022** | MS | | 0.107* | 0.382* |
| VAR | *d*(de) | 0.251* | $d(f-34)$ | 0.0023 | VAR | *d*(u-rate) | 0.083* | 0.448* |
| Breakpoint | | 0.725* | | 0.0023 | Breakpoint | | 0.075** | 0.372** |
| Threshold | | 0.250** | | 0.0023 | MS | | 0.105* | 0.384** |
| MS | | 0.275* | | 0.0023 | | | | |
| VAR | *d*(bm) | 0.252* | $d(f-45)$ | 0.0023 | | | | |
| Breakpoint | | 0.738* | | 0.0023 | | | | |
| Threshold | | 0.251* | | 0.0023 | | | | |
| MS | | 0.276* | | 0.0022** | | | | |
| VAR | *d*(tbl) | 0.256* | $d(f1)$ | 0.0024 | | | | |
| Breakpoint | | 0.732* | | 0.0024 | | | | |
| Threshold | | 0.256* | | 0.0023 | | | | |
| MS | | 0.264* | | 0.0023 | | | | |
| VAR | *d*(dfy) | 0.253* | $d(f5)$ | 0.0024 | | | | |
| Breakpoint | | 0.713* | | 0.0024 | | | | |
| Threshold | | 0.254* | | 0.0024 | | | | |
| MS | | 0.263* | | 0.0025 | | | | |
| VAR | *d*(lty) | 0.254* | $d(f7)$ | 0.0024 | | | | |
| Breakpoint | | 0.721* | | 0.0024 | | | | |
| Threshold | | 0.253* | | 0.0023 | | | | |
| MS | | 0.264* | | 0.0022** | | | | |
| VAR | *d*(tms) | 0.252* | $d(f8)$ | 0.0024 | | | | |
| Breakpoint | | 0.710* | | 0.0024 | | | | |
| Threshold | | 0.252* | | 0.0023 | | | | |
| MS | | 0.257* | | 0.0024 | | | | |
| VAR | *d*(ntis) | 0.251* | *d*(f-unc) | 0.0026 | | | | |
| Breakpoint | | 0.712* | | 0.0027 | | | | |

(Continues)

**TABLE 9** (Continued)

| | S&P500 | | Bonds | | | Gold | Oil |
|---|---|---|---|---|---|---|---|
| | Predictor | Loss | Predictor | Loss | Predictor | Loss | Loss |
| Threshold | | 0.249* | | 0.0027 | | | |
| MS | | 0.261* | | 0.0026 | | | |
| VAR | $d$(emv) | 0.766* | $d$(fwd-sprd) | 0.0023 | | | |
| Breakpoint | | 0.189** | | 0.0023 | | | |
| Threshold | | 0.239** | | 0.0023 | | | |
| MS | | 0.261* | | 0.0023 | | | |
| VAR | $d$(fin-un) | 0.795* | $d$(macr-unc) | 0.0027 | | | |
| Breakpoint | | 0.620* | | 0.0027 | | | |
| Threshold | | 0.265* | | 0.0027 | | | |
| MS | | 0.426* | | 0.0025 | | | |
| VAR | $d$(mac-un) | 0.426* | $d(y-1)$ | 0.0023 | | | |
| Breakpoint | | 1.039* | | 0.0023 | | | |
| Threshold | | 0.426* | | 0.0023 | | | |
| MS | | 0.281* | | 0.0025 | | | |
| VAR | $d$(vix) | 0.281* | $d$(yld-sprd) | 0.0023 | | | |
| Breakpoint | | 0.840* | | 0.0023 | | | |
| Threshold | | 0.277* | | 0.0023 | | | |
| MS | | 0.324* | | 0.0023 | | | |

*Note*: Reports out-of-sample MSE × 10,000 from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. The MCS is calculated for the above models plus the models in Table 8. Note $M^*_{0.25} \subset M^*_{0.10}$. Model details for each specification are provided in Table 7, where Model 1 (VAR), 5 (breakpoint), 13 (threshold), and 15 (Markov switching) are employed. Breakpoint and threshold regression determine breakpoints using a sequential test of $l+1$ versus $l$ breaks. MS denotes a two-state Markov switching model.

Abbreviations: MCS, Model Confidence Set; MS, Markov switching; MSE, Mean-Squared Error; VAR, vector autoregression.

### 4.2.2 | Model averaging and final comparisons of modeling approaches

Having determined the MCS for the different model classes, we now construct model-averaged forecasts for each class. Only models deemed to have EPA via the MCS will be included in the respective averages.

The model combinations are compared with hybrid forecasts as well as four simple benchmarks. Namely, (1) a linear AR(1) model fitted on monthly variance swap payoffs[28]; (2) a naive (or prevailing mean) model, which sets the forecast at the end of month $t$ equal to the average variance payoff from the beginning of the sample to the end of month $t$[29,30]; (3) a PC model that employs a linear predictor at the end of each month equal to the lagged first PC of all predictors that are available daily; and (4) the same as (3) but with the first PC extracted from the predictors that are available only monthly (with a lag). Note that the third and fourth benchmark models are also reported in Table 10 as PC1.

[28]The AR(1) specification is based on autocorrelation and partial autocorrelation functions.

[29]The prevailing mean forecast is regularly used as benchmark in the literature on risk-premia predictability. See, for example, the influential Welch and Goyal (2008) study. Given that the VRP is more persistent than excess returns, we acknowledge that an AR specification may be a more suitable benchmark than the prevailing mean (Hollstein et al., 2019).

[30]Although more persistent than realized excess returns, the realized VRP is much less persistent than volatility. The autocorrelation function for the first lag of the VRP is 0.220 (S&P), 0.392 (bond), 0.057 (gold), and 0.266 (oil). The 2nd lag of the partial autocorrelation function is insignificant. On the other hand, the autocorrelation function of volatility is in the 0.60 to 0.90 range at the first lag and decays rather slowly.

**TABLE 10** Out-of-sample MSE: Principal component models

| | S&P500 | | Bonds | | Gold | | Oil | |
|---|---|---|---|---|---|---|---|---|
| | Daily | Monthly | Daily | Monthly | Daily | Monthly | Daily | Monthly |
| PC6 | 0.197** | 0.171** | – | 0.003* | 0.081* | 0.071** | 0.042 | 0.040* |
| PC5 | 0.184** | 0.155** | – | 0.003* | 0.081* | 0.069** | 0.041 | 0.040* |
| PC4 | 0.180** | 0.159** | – | 0.003* | 0.080* | 0.065** | 0.040* | 0.041 |
| PC3 | 0.175** | 0.155** | – | 0.003* | 0.077* | 0.065** | 0.040* | 0.040* |
| PC2 | 0.169** | 0.154** | 0.003* | 0.003* | 0.079* | 0.063** | 0.039* | 0.038* |
| PC1 | 0.183** | 0.161** | 0.003* | 0.003* | 0.078* | 0.062** | 0.038* | 0.041* |
| PC6-AR | 0.324** | 0.243** | – | 0.002* | 0.090* | 0.084* | 0.039* | 0.037** |
| PC5-AR | 0.297** | 0.234** | – | 0.002* | 0.090* | 0.079* | 0.039* | 0.037** |
| PC4-AR | 0.307** | 0.235** | – | 0.002* | 0.089* | 0.075* | 0.038* | 0.038* |
| PC3-AR | 0.303** | 0.255** | – | 0.002* | 0.086* | 0.075* | 0.038* | 0.037** |
| PC2-AR | 0.290** | 0.244** | 0.002* | 0.002* | 0.086* | 0.072* | 0.037** | 0.035** |
| PC1-AR | 0.259** | 0.266** | 0.002* | 0.002* | 0.085* | 0.071* | 0.036** | 0.037** |
| PC6-brk | 0.783** | 0.250** | – | 0.002* | 0.079* | 0.091* | 0.049 | 0.044 |
| PC5-brk | 0.765** | 0.664* | – | 0.002* | 0.078* | 0.078* | 0.048 | 0.043 |
| PC4-brk | 0.777** | 0.668** | – | 0.002* | 0.075* | 0.086* | 0.046 | 0.044 |
| PC3-brk | 0.762** | 0.700** | – | 0.002* | 0.077* | 0.088* | 0.046 | 0.043 |
| PC2-brk | 0.759** | 0.697** | 0.002* | 0.002* | 0.071** | 0.071** | 0.046 | 0.042 |
| PC1-brk | 0.780** | 0.729* | 0.002* | 0.002** | 0.073* | 0.070* | 0.044* | 0.045* |
| PC6-thr | 0.791** | 0.654** | – | 0.002* | 0.090* | 0.133* | 0.041 | 0.039* |
| PC5-thr | 0.766** | 0.670** | – | 0.002* | 0.090* | 0.079* | 0.041 | 0.039* |
| PC4-thr | 0.790** | 0.675** | – | 0.002* | 0.089* | 0.117* | 0.040* | 0.040 |
| PC3-thr | 0.782** | 0.725** | – | 0.002* | 0.086* | 0.117* | 0.040 | 0.039* |
| PC2-thr | 0.782** | 0.720** | 0.002* | 0.002* | 0.086* | 0.071* | 0.039* | 0.036** |
| PC1-thr | 0.789** | 0.753** | 0.002* | 0.002* | 0.085* | 0.071* | 0.037* | 0.040* |
| PC3-ms | 0.240** | – | – | 0.003* | 0.105* | 0.084* | 0.035** | 0.038* |
| PC2-ms | 0.270** | 0.262** | 0.002* | 0.003* | 0.105* | 0.082* | 0.037** | 0.038* |
| PC1-ms | 0.287** | 0.265** | 0.002* | 0.003* | 0.105* | 0.080* | 0.039* | 0.037** |

*Note*: Reports out-of-sample MSE ×10,000 from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. Note $M^*_{0.25} \subset M^*_{0.10}$. The MCS for each asset is determined for each asset by considering the losses for the daily and monthly variables. The number after PC denotes the total number of PCs included as regressors, for example, PC3 regresses $Y_t$ against lagged first, second, and third PCs. AR includes an autoregressive term. The remaining models allow for structural breaks via breakpoint (brk), self-exciting threshold (thr), and two-state Markov switching models.

Abbreviations: AR, autoregressive; MCS, Model Confidence Set; MSE, Mean-Squared Error; PC, principal component.

Table 12 reports OOS MSE and MCS results for all assets. As we determine the MCS for both the 10% and the 25% significance levels, we compute model averages and hence their MSE for each significance level. As a result, for each of the four model classes (Pure Time Series, Monthly Predictors, PC, and MIDAS), the table reports two MSEs: MSE (10%) and MSE(25%).

At the 10% level, the simple AR benchmark is in the MCS for all four asset classes. The same is true for the benchmark PC1 daily and PC1 monthly. For bonds, gold, and oil, all model combinations are in the MCS, while for equity VRP only the PC combination (from the MCS at the 25% level) and the MIDAS combination are in the MCS. At the more restrictive 25% confidence level, the benchmark AR and pure time-series combination remain in the MCS for

**TABLE 11** Out-of-sample MSE: MIDAS models

| S&P500 | | Bonds | | | Gold | Oil |
|---|---|---|---|---|---|---|
| **Predictor** | **Loss** | **Predictor** | **Loss** | **Predictor** | **Loss** | **Loss** |
| de | 0.235** | $f-12$ | 0.0024** | bdi | 0.081** | 0.453* |
| dfy | 0.288 | $y1$ | 0.0024** | cad | 0.082** | 0.430** |
| dp | 0.255 | yield | 0.0024** | chp | 0.078** | 0.359** |
| dy | 0.263 | $f-23$ | 0.0024** | dfr | 0.084** | 0.425** |
| ep | 0.256 | fwd-sprd | 0.0023** | ret | 0.085** | 0.366** |
| lty | 0.246** | | | inr | 0.082** | 0.409** |
| tbl | 0.267 | | | ltr | 0.081** | 0.374** |
| tms | 0.266 | | | sar | 0.080** | 0.376** |
| dfr | 0.280 | | | aud | 0.083** | 0.389** |
| ltr | 0.249 | | | nzd | 0.085** | 0.425** |
| RV | 0.249** | | | | | |
| sp | 0.256 | | | | | |
| vix | 0.255 | | | | | |

*Note*: Reports out-of-sample MSE ×10,000 from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. Note $M^*_{0.25} \subset M^*_{0.10}$.

Abbreviations: MCS, Model Confidence Set; MIDAS, Mixed-Data Sampling; MSE, Mean-Squared Error.

gold and oil. For the S&P500, the benchmark PC1 daily and PC1 monthly are in the MCS, as well as the combination of PC models and the MIDAS combination. For bonds only the PC combination and the MIDAS combination remain in the MCS. Hybrid models rarely appear in the MCS and never by themselves.

This set of results leads us to four main takeaways. First, in forecasting the S&P, gold, and oil VRPs, several of the simple benchmarks perform as well as the richer models and model combinations. In particular, for gold and oil, a linear AR model is never significantly outperformed by models containing predictors, and/or by models containing structural breaks. Second, the predictors do lead to better forecasting performance for the S&P500 VRP and for the bond market VRP compared with pure time-series models. While for the S&P500 a simple linear PC1 model has equal forecasting ability as any other model containing predictors, for bonds a combination of PC models and a combination of MIDAS models are the best performers. Third, for all four assets, the inclusion of structural breaks does not improve OOS forecasting performance relative to one-regime (or state invariant) specifications. Fourth, for all four assets, the hybrid models do not lead to improved OOS performance. Therefore, rather than forecasting directly the monthly swap payoff, there seems to be no advantage in using higher-frequency data to separately forecast RV and then combine it with the observed market swap rate.

### 4.2.3 | Trading strategies

We now consider the economic significance of the various forecasts via two simple trading strategies. The first strategy shorts a variance swap at the end of the month if the VRP forecast for the forthcoming month is negative and buys a variance swap if the VRP forecast is positive. We label this strategy as *symmetric*. The second strategy takes into account the strength of the signal by shorting a swap only if the VRP forecast is below a certain percentile of the unconditional VRP distribution. If a negative VRP forecast is not below the percentile, the investor does not take a position in variance swaps and earns the risk-free rate. If the VRP forecast is positive, a swap is bought. The second strategy accounts for the large negative asymmetry in the payoff distribution. Table 13 reports unconditional VRP percentiles that inform the second strategy. We consider percentiles of 20%, 30%, and 40%. We label these strategies as *asymmetric*. For all strategies, the swap position is held to maturity, that is, to the end of the following month as we are using 30-day swaps. Only models included in the MCS in Table 12 are evaluated for economic significance.

TABLE 12 Out-of-sample MSE: VRP payoff forecasts

| | S&P500 | Bond | Gold | Oil |
|---|---|---|---|---|
| *Benchmarks* | | | | |
| AR | 0.250* | 0.0024* | 0.074** | 0.367** |
| Naïve | 0.271 | 0.0033 | 0.086* | 0.442* |
| PC1 daily | 0.183** | 0.0028* | 0.078** | 0.375** |
| PC1 monthly | <u>0.161</u>** | 0.0026* | 0.062** | 0.414* |
| *Combination: Monthly predictors* | | | | |
| MSE (10%) | 0.353 | 0.0024* | 0.082** | 0.384** |
| MSE (25%) | 0.324 | – | 0.077** | 0.407* |
| *Combination: Time series* | | | | |
| MSE (10%) | 0.359 | 0.0023* | 0.086** | 0.371** |
| MSE (25%) | 0.747 | – | 0.073** | – |
| *Combination: PC* | | | | |
| MSE (10%) | 0.360 | 0.0023** | 0.069** | 0.369** |
| MSE (25%) | 0.191** | <u>0.0022</u>** | <u>0.061</u>** | <u>0.353</u>** |
| *Combination: MIDAS* | | | | |
| MSE (10%) | 0.231** | 0.0024** | 0.077** | 0.359** |
| MSE (25%) | – | – | – | 0.357** |
| *Hybrid* | | | | |
| HAR | 0.265 | 0.0033 | 0.081** | 0.725 |
| ln HAR | 0.165** | 0.0063 | 0.136* | 1.284* |
| ARFIMA | 0.226 | 0.0041 | 0.114* | 1.255 |
| ln ARFIMA | 0.239 | 0.0058 | 0.134* | 1.386* |

*Note*: The lowest MSE is underlined for each asset. The absence of a result at the 25% level is denoted by "-". This occurs when the MCS at 10% contains the same predictors as the MCS at 25%. Reports out-of-sample MSE × 10,000 for VRP payoffs from April 2007 to July 2018. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. Note $M^*_{0.25} \subset M^*_{0.10}$. Benchmark models are naive, AR(2) for S&P500 and bond VRP and AR(1) for gold and oil VRP, PC1 daily is a linear regression of the VRP against the first principal component (daily variables), PC1 monthly is a comparable model against monthly predictors. Hybrid employs models that are fit to daily realized variances which are then combined with implied volatility to form the VRP forecast—HAR is the Heterogeneous Autoregressive model and ARFIMA is an ARFIMA$(1, d, 1)$ model.

Abbreviations: AR, autoregressive; MCS, Model Confidence Set; MIDAS, Mixed-Data Sampling; MSE, Mean-Squared Error; PC, principal component; VRP, variance risk premium.

TABLE 13 Percentiles of the VRP distribution

| Percentiles (%) | S&P500 | Bonds | Gold | Oil |
|---|---|---|---|---|
| 10 | −0.369 | −0.048 | −0.231 | −0.952 |
| 20 | −0.272 | −0.034 | −0.163 | −0.588 |
| 30 | −0.195 | −0.021 | −0.122 | −0.468 |
| 40 | −0.160 | −0.012 | −0.085 | −0.329 |
| 50 | −0.124 | −0.007 | −0.066 | −0.259 |
| 60 | −0.086 | 0.002 | −0.044 | −0.169 |
| 70 | −0.056 | 0.010 | −0.021 | −0.063 |
| 80 | −0.035 | 0.022 | 0.011 | 0.066 |
| 90 | 0.008 | 0.041 | 0.105 | 0.391 |

*Note*: Reports percentiles of the variance swap payoff distribution × 100 for the period from January 1996 to July 2018.

Abbreviation: VRP, variance risk premium.

**TABLE 14** VRP trading

| | Returns (%) | Sharpe ratio | Max drawdown (%) |
|---|---|---|---|
| S&P500 | | | |
| *Symmetric strategy* | | | |
| AR | 0.118* | 0.294 | 1.072 |
| Naïve | 0.119* | 0.298 | 1.336 |
| PC1 daily | 0.084* | 0.205* | 5.038 |
| PC1 monthly | 0.100* | 0.246* | 4.450 |
| Combination: PC | 0.091* | 0.224* | 5.038 |
| Combination: MIDAS | 0.131** | 0.330* | 0.926 |
| ln HAR | 0.095* | 0.235* | 4.450 |
| *30% strategy* | | | |
| AR | 0.102* | 0.276 | 0.589 |
| Naïve | 0.114* | 0.311 | 0.608 |
| PC1 daily | 0.052 | 0.146 | 4.214 |
| PC1 monthly | 0.137** | 0.589** | 0.825 |
| Combination: PC | 0.116* | 0.540** | 0.589 |
| Combination: MIDAS | 0.128** | 0.350 | 0.589 |
| ln HAR | 0.123** | 0.593** | 0.243 |
| Bond | | | |
| *Symmetric strategy* | | | |
| AR | 0.007 | 0.128 | 0.432 |
| Naïve | 0.014** | 0.277** | 0.258 |
| PC1 daily | 0.005* | 0.087* | 0.379 |
| PC1 monthly | 0.013** | 0.259** | 0.293 |
| Combination: Monthly | 0.010** | 0.200** | 0.411 |
| Combination: Time series | 0.008* | 0.146* | 0.440 |
| Combination: PC | 0.010* | 0.195* | 0.411 |
| Combination: MIDAS | 0.012* | 0.237** | 0.258 |
| *30% strategy* | | | |
| AR | 0.027** | 0.479** | 0.293 |
| Naïve | 0.021** | 0.444** | 0.179 |
| PC1 daily | 0.032** | 0.417** | 0.442 |
| PC1 monthly | 0.039** | 0.473** | 0.293 |
| Combination: Monthly | 0.027** | 0.485** | 0.293 |
| Combination: Time series | 0.025** | 0.461** | 0.293 |
| Combination: PC | 0.028** | 0.496** | 0.293 |
| Combination: MIDAS | 0.027** | 0.551** | 0.139 |
| Gold | | | |

TABLE 14 (Continued)

| | Returns (%) | Sharpe ratio | Max drawdown (%) |
|---|---|---|---|
| *Symmetric strategy* | | | |
| AR | 0.037* | 2.010 | 2.504 |
| Naïve | 0.033 | 0.125 | 1.457 |
| PC1 daily | 0.030* | 0.114 | 1.987 |
| PC1 monthly | 0.051** | 0.197** | 2.504 |
| Combination: Monthly (25%) | 0.028* | 0.108 | 3.025 |
| Combination: Time series | 0.015 | 0.057 | 4.999 |
| Combination: PC | 0.041** | 0.158** | 1.987 |
| Combination: MIDAS | 0.028* | 0.108 | 2.504 |
| HAR | 0.032* | 0.123 | 2.504 |
| *30% strategy* | | | |
| AR | 0.037* | 0.235** | 2.010 |
| Naïve | 0.062* | 0.249 | 1.457 |
| PC1 daily | 0.045* | 0.238 | 2.127 |
| PC1 monthly | 0.098** | 0.485** | 1.031 |
| Combination: Monthly (25%) | 0.055** | 0.250 | 2.495 |
| Combination: Time series | 0.033 | 0.173 | 2.495 |
| Combination: PC | 0.085** | 0.401** | 1.457 |
| Combination: MIDAS | 0.028 | 0.142 | 3.408 |
| HAR | 0.025 | 0.100 | 2.504 |
| Oil | | | |
| *Symmetric strategy* | | | |
| AR | 0.230* | 0.356* | 2.559 |
| Naïve | 0.275** | 0.437** | 2.559 |
| PC1 daily | 0.265** | 0.418** | 2.912 |
| PC1 monthly | 0.257** | 0.404** | 6.562 |
| Combination: Monthly | 0.248** | 0.388** | 2.912 |
| Combination: Time series | 0.234* | 0.363* | 5.737 |
| Combination: PC | 0.264** | 0.417** | 2.912 |
| Combination: MIDAS (25%) | 0.259** | 0.407** | 2.912 |
| ln HAR | 0.265** | 0.418** | 6.562 |
| ln ARFIMA | 0.221* | 0.340* | 6.562 |
| *30% strategy* | | | |
| AR | 0.132 | 0.282 | 2.559 |
| Naïve | 0.196* | 0.341* | 2.559 |
| PC1 daily | 0.120 | 0.364** | 1.091 |
| PC1 monthly | 0.052 | 0.592** | 0.892 |
| Combination: Monthly | 0.131 | 0.265 | 2.559 |
| Combination: Time series | 0.146 | 0.302 | 2.559 |

(Continues)

**TABLE 14** (Continued)

| | Returns (%) | Sharpe ratio | Max drawdown (%) |
|---|---|---|---|
| Combination: PC | 0.171 | 0.329* | 2.559 |
| Combination: MIDAS (25%) | 0.141 | 0.287 | 2.559 |
| ln HAR | 0.223* | 0.362* | 6.562 |
| ln ARFIMA | 0.177* | 0.284* | 6.562 |

*Note*: Reports out-of-sample returns, Sharpe ratios (SR), and Maximum Drawdown (MDD) from April 2007 to July 2018 for portfolio strategies with monthly rebalancing. Only the models in the MCS in Table 12 are evaluated for trading purposes. The symmetric strategy sells vol if the VRP forecast is negative and buys vol if the VRP forecast is positive. The remaining strategies only sell vol if the VRP forecast is below the stated percentile of the unconditional VRP distribution. If a negative VRP forecast is not below the percentile, the investor does not take a variance swap position for that month and earns the risk-free rate. ** and * denote inclusion in the MCS at the 25% ($M^*_{0.25}$) and 10% ($M^*_{0.10}$) levels, respectively. Note $M^*_{0.25} \subset M^*_{0.10}$.

Abbreviations: AR, autoregressive; HAR, heterogeneous autoregressive; MIDAS, Mixed-Data Sampling; MSE, Mean-Squared Error; PC, principal component; VRP, variance risk premium.

For each strategy and for each considered forecasting model, we compute three performance measures: the average excess return, the SR, and the Maximum Drawdown (MDD). We compute the SR as it is a standard measure of risk-adjusted performance both in the academic literature and in industry practice. The MDD measures the single largest peak to trough return over the considered OOS period; it therefore represents the worst loss resulted from a fully collateralized variance swap position over such period (see, e.g., G. Alexander & Baptista, 2006). The MDD is a useful diagnostic as it focuses on downside risk and, hence, takes into account the large negative asymmetry in variance swap payoffs illustrated above.[31] As we did when comparing forecasting accuracy, we compare trading performance (or economic significance) by computing the MCS for average returns and SRs within a given strategy.[32] Namely, we start with the null hypothesis that all the, for example, SRs generated by the alternative forecasting models are equal. Then, we progressively eliminate models until the null can no longer be rejected.

Table 14 reports the OOS performance measures. The average return and SR figures are monthly while the MDD numbers represent the worst cumulative loss for a given strategy. Results for the asymmetric strategies are similar for each of the three thresholds, so we only report the results using the 30th percentile. Two main considerations emerge. First, strategies based on the naive (i.e., prevailing mean) forecast are outperformed by those based on predictors for the S&P500, gold, and oil, but not for Treasuries. For the latter, the naive strategy is in the MCS, for both symmetric and asymmetric strategies, in terms of average returns as well as SRs. The MDD measure does not appear to discriminate across forecasting models either, confirming that the naive strategy is not outperformed by the alternatives. These findings for bonds are reminiscent of the often reported disconnect between forecasting accuracy and trading performance in the equity premium prediction literature.[33] For equities and gold, the naive strategy is never in the MCS, while for oil it is in the MCS of the symmetric strategies but not in the MCS at the 25% level of the asymmetric ones. The improvements in SRs from using the predictors rather than the prevailing mean are economically tangible, especially from the asymmetric strategies. For equities, the SR from the PC monthly and the Combination PC models are between 0.54 and 0.59/month (roughly, 1.90 on an annual basis), compared with 0.31 (about 1.07 annualized) for the naive strategy. Comparably large improvements occur for gold, where the asymmetric strategy based on monthly PCs almost doubles the SR of the naive strategy or of the AR1-based strategy.[34]

Second, the simple PC1 model with monthly predictors appears to be the overall best performer in terms of SR, as it is in the MCS in all instances and, in the majority of cases (again outside of the bonds case), leads to sizeable gains not only relative to the naive forecast but also compared with the other specifications. The trading performance of the PC1 monthly model matches the findings on forecasting accuracy. MDDs suggest that the overall improvement in trading

[31]The MDD statistic is different from other metrics, such as volatility, and downside measures like skewness or semivariance in that it crucially depends on the order in which the returns occur.

[32]As it is not obvious how to apply the MCS framework to MDD, we do not conduct formal tests for differences in such measures.

[33]See, for example, Cenesizoglu and Timmermann (2012) for a thorough illustration.

[34]We also test whether the SR generated by a given strategy is statistically different from the SR generated by the naive strategy. Following, for example, DeMiguel et al. (2009), we use the method suggested by Jobson and Korkie (1981) with the correction proposed by Memmel (2003). We find that the broad conclusion drawn from the MCS tests holds and no additional insights arise. Consequently, we choose not to report these results.

performance with the PC1 model does not come at the cost of facing large negative returns. For gold and oil, MDD for the PC1 model compares quite favorably with those generated by other predictive models, in particular among the asymmetric strategies.

## 5 | CONCLUSION

In this paper we investigate the OOS predictability of the realized (ex post) VRP across four major asset classes (US equities, US government bonds, gold, and crude oil) using a large collection of forecasting models. To our knowledge such a comprehensive forecasting exercise is currently absent from the literature.

Our empirical results convey the following main messages. First, while in-sample nonlinear specifications that include structural breaks provide a better characterization of the VRP dynamics, this does not translate into an improvement in OOS forecasting performance relative to simpler linear (or state invariant) models. Second, variables that have been shown to forecast excess returns for the four asset classes, contain predictive power for the respective VRP as well, as they outperform a naive no-predictability (or prevailing mean) specification. Third, while for the S&P500 VRP and for the Treasury bond VRP the predictors also outperform pure time-series models, for gold and oil a linear AR model is never significantly outperformed by models containing predictors, and/or by models containing structural breaks. For the S&P500, a linear PC model forecasts as well as any other model containing predictors. While for the government bonds VRP, a forecast combination of PCs models and a combination of MIDAS models are the best performers. Fourth, there seems to be no advantage in using higher-frequency data to separately forecast RV and combine the forecast with the observed market swap rate rather than forecasting directly the monthly swap payoff. Fifth, with the exception of the treasuries VRP, the predictors appear to improve trading performance relative to strategies based on a prevailing mean forecast and to strategies based on pure time-series models.

Our novel evidence should be of relevance to variance swap sellers, hedgers, and portfolio managers. Future research could therefore consider an asset allocation exercise with variance swaps, and the impact of different VRP models on portfolio performance. Other extensions could consider VRP forecasts and asset allocation decisions using different swap rate maturities over various horizons.

### ACKNOWLEDGEMENT

Open access publishing facilitated by The University of Melbourne, as part of the Wiley - The University of Melbourne agreement via the Council of Australian University Librarians.

### DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from multiple sources many of which are public and detailed in Appendix A. Restrictions apply to the availability of data from Thomson Reuters Tick History, Datastream and CRSP, which were used under license for this study. Data are available at https://doi.org/10.26188/62142d7356d1d with permission from the data providers.

### ORCID

*Jonathan Dark* http://orcid.org/0000-0002-7688-0193

### REFERENCES

Alexander, C., Korovilas, D., & Kapraun, J. (2016). Diversification with volatility products. *Journal of International Money and Finance*, *65*, 213–235.

Alexander, G., & Baptista, A. M. (2006). Portfolio selection with a drawdown constraint. *Journal of Banking and Finance*, *30*(11), 3171–3189.

Andersen, T. G., Bondarenko, O., & Gonzalez-Perez, M. T. (2015). Exploring return dynamics via corridor implied volatility. *The Review of Financial Studies*, *28*(10), 2902–2945.

Andreou, E., & Ghysels, E. (2021). Predicting the VIX and the volatility risk premium: The role of short-run funding spreads volatility factors. *Journal of Econometrics*, *220*(2), 366–398.

Bakshi, G., & Kapadia, N. (2003). Cdelta-hedged gains and the negative market volatility risk premium. *The Review of Financial Studies*, *16*(2), 527–566.

Bakshi, G., Panayotov, G., & Skoulakis, G. (2011). *The Baltic Dry Index as a predictor of global stock returns, commodity returns, and global economic activity* [AFA 2012 Chicago Meetings Paper].

Banerjee, A., & Urga, G. (2005). Modelling structural breaks, long memory and stock market volatility: An overview. *Journal of Econometrics*, *129*, 1–34.

Bollerslev, T., Tauchen, G., & Zhou, H. (2009). Expected stock returns and variance risk premia. *The Review of Financial Studies*, *22*(11), 4463–4492.

Bollerslev, T., Gibson, N., & Zhou, H. (2011). Dynamic estimation of volatility risk premia and investor risk aversion from option-implied and realized volatilities. *Journal of Econometrics*, *160*(1), 235–245.

Brière, M., Burgues, A., & Signori, O. (2010). Volatility exposure for strategic asset allocation. *Journal of Portfolio Management*, *36*(3), 105–116.

Britten-Jones, M., & Neuberger, A. (2000). Option prices, implied price processes, and stochastic volatility. *The Journal of Finance*, *55*(2), 839–866.

Carr, P., & Wu, L. (2009). Variance risk premiums. *Review of Financial Studies*, *22*(3), 1311–1341.

Cenesizoglu, T., & Timmermann, A. (2012). Do return prediction models add economic value? *Journal of Banking & Finance*, *36*(11), 2974–2987.

Chen, H.-C., Chung, S.-L., & Ho, K.-Y. (2011). The diversification effects of volatility-related assets. *Journal of Banking and Finance*, *35*, 1179–1189.

Chen, Y.-C., Rogoff, K. S., & Rossi, B. (2010). Can exchange rates forecast commodity prices? *The Quarterly Journal of Economics*, *125*(3), 1145–1194.

Choi, H., Mueller, P., & Vedolin, A. (2017). Bond variance risk premiums. *Review of Finance*, *21*(3), 987–1022.

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

Dacco, R., & Satchell, S. (1999). Why do regime switching models perform so badly? *Journal of Forecasting*, *18*, 1–16.

DeMiguel, V., Garlappi, L., & Uppal, R. (2009). Optimal versus naive diversification: How inefficient is the 1/n portfolio strategy? *The Review of Financial Studies*, *22*(5), 1915–1953.

Diebold, F., & Inoue, A. (2001). Long memory and regime switching. *Journal of Econometrics*, *105*, 131–159.

Doran, J. S. (2020). Volatility as an asset class: Holding VIX in a portfolio. *Journal of Futures Markets*, *40*, 841–859.

Du, J., & Kapadia, N. (2012). *The tail in the volatility index* [Amherst Working paper]. University of Massachusetts.

Egloff, D., Leippold, M., & Wu, L., et al (2010). The term structure of variance swap rates and optimal variance swap investments. *Journal of Financial and Quantitative Analysis*, *45*(5), 1279.

Eraker, B., & Wu, Y. (2017). Explaining the negative returns to volatility claims: An equilibrium approach. *Journal of Financial Economics*, *125*, 72–98.

Feunou, B., Fontaine, J.-S., Taamouti, A., & Tedongap, R. (2014). Risk premium, variance premium, and the maturity structure of uncertainty. *Review of Finance*, *18*, 219–269.

Forsberg, L., & Ghysels, E. (2007). Why do absolute returns predict volatility so well? *Journal of Financial Econometrics*, *5*(1), 31–67.

Gao, X., & Nardari, F. (2018). Do commodities add economic value in asset allocation? New evidence from time-varying moments. *Journal of Financial and Quantitative Analysis*, *53*(1), 1–29.

Gargano, A., & Timmermann, A. (2014). Forecasting commodity price indexes using macroeconomic and financial predictors. *International Journal of Forecasting*, *30*(3), 825–843.

Ghysels, E., Santa-Clara, P., & Valkanov, R. (2004). *The MIDAS touch: Mixed data sampling regression models*. Working paper, UNC and UCLA.

Gordon, S., & St-Amour, P. (2004). Asset returns and state-dependent risk preferences. *Journal of Business and Economic Statistics*, *22*(3), 241–252.

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

Hollstein, F., Prokopczuk, M., Tharann, B., & WeseSimen, C. (2019). Predicting the equity market with option-implied variables. *The European Journal of Finance*, *25*(10), 937–965.

Jobson, J. D., & Korkie, B. M. (1981). Performance hypothesis testing with the Sharpe and Treynor measures. *The Journal of Finance*, *36*, 889–908.

Kim, K. (2014). Counter-cyclical risk aversion. *Journal of Empirical Finance*, *29*, 384–401.

Konstantinidi, E., & Skiadopoulos, G. (2016). How does the market variance risk premium vary over time? Evidence from S&P 500 variance swap investment returns. *Journal of Banking & Finance*, *62*, 62–75.

Li, J., & Zinna, G. (2018). The variance risk premium: Components, term structures, and stock return predictability. *Journal of Business and Economic Statistics*, *36*(3), 411–425.

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

Mele, A. (2007). Asymmetric stock market volatility and the cyclical behavior of expected returns. *Journal of Financial Economics*, *86*(2), 446–478.

Memmel, C. (2003). Performance hypothesis testing with the Sharpe ratio. *Finance Letters*, *22*, 21–23.

Nieto, B., Novales, A., & Rubio, G. (2014). Variance swaps, non-normality and macroeconomic and financial risks. *The Quarterly Review of Economics and Finance*, *54*(2), 257–270.

Paye, B. (2012). Deja vol: Predictive regressions for aggregate stock market volatility using macroeconomic variables. *Journal of Financial Economics*, *106*(3), 527–546.

Prokopczuk, M., Symeonidis, L., & Simen, C. W. (2011). Variance risk in commodity markets. *Journal of Banking and Finance*, *81*, 136–149.

Psaradakis, Z., & Spagnolo, N. (2003). On the determination of the number of regimes in Markov-switching autoregressive models. *Journal of Time Series Analysis*, *24*(2), 237–252.

Qu, Z. (2011). A test against spurious long memory. *Journal of Business and Economic Statistics*, *29*, 423–438.

Rapach, D., & Zhou, G. (2013). Forecasting stock returns. In G. Elliott, & A. Timmermann (Eds.), *Handbook of economic forecasting 2A* (pp. 329–383). Elsevier.

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

Rombouts, J. V., Stentoft, L., & Violante, F. (2020). Variance swap payoffs, risk premia and extreme market conditions. *Econometrics and Statistics*, *13*, 106–124.

Timmermann, A. (1993). How learning in financial markets generates excess volatility and predictability in stock prices. *Quarterly Journal of Economics*, *108*(4), 1135–1145.

Timmermann, A. (1996). Excess volatility and return predictability of stock returns in autoregressive dividend models with learning. *Review of Economic Studies*, *63*(4), 523–577.

Timmermann, A. (2006). Forecast combination. In G. Elliott & A. Timmermann (Eds.), *Handbook of economic forecasting 2A* (pp. 329–383).

VRP forecasting. (2022). https://doi.org/10.26188/62142d7356d1d with permission from the data providers.

Warren, G. J. (2012). Can investing in volatility help meet your portfolio objectives? *Journal of Portfolio Management Winter*, *38*(2), 82–98.

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

**How to cite this article:** Dark, J., Gao, X., van der Heijden, T., & Nardari, F. (2022). Forecasting variance swap payoffs. *Journal of Futures Markets*, 1–30. https://doi.org/10.1002/fut.22371

## APPENDIX A: DATA APPENDIX (TABLE A1)

**TABLE A1** Data description

| Label | Time series | Frequency | Source |
|---|---|---|---|
| S&P500 | S&P500 Total Return Index | Monthly, daily, and intradaily | Thomson Reuters Tick History |
| Bond | US Treasury Bond futures | Monthly, daily, and intradaily | Thomson Reuters Tick History |
| Gold | Gold futures | Monthly, daily, and intradaily | Thomson Reuters Tick History |
| Oil | WTI crude oil futures | Monthly, daily, and intradaily | Thomson Reuters Tick History |
| DP | log of Dividend Price ratio | Monthly and daily | Datastream |
| DY | log of Dividend Yield | Monthly and daily | Datastream |
| EP | log of Earnings to Price ratio | Monthly and daily | Datastream |
| DE | log of Dividend Payout ratio | Monthly and daily | Datastream |
| BM | Book-to-Market ratio | Monthly | GW2012 |
| TBL | 3-Month US T-bill | Monthly and daily | FRED |
| DFY | Default Yield Spread (AAA–BAA) | Monthly and daily | FRED |
| LTY | Long-term government bond yield | Monthly and daily | FRED |
| TMS | Government bond term spread | Monthly and daily | FRED |
| NTIS | Net Equity Expansion | Monthly and daily | Amit Goyal website |
| INFL | Inflation rate | Monthly | GW2012 |
| LTR | Long-term government bond return | Monthly and daily | FRED |
| DFR | Default Return Spread | Monthly and daily | FRED |
| epu-3comp | US Economic policy uncertainty index (three components) | Monthly | Baker, Bloom and Davis website |

(Continues)

**TABLE A1** (Continued)

| Label | Time series | Frequency | Source |
|---|---|---|---|
| epu-news | US Economic policy uncertainty index (news based) | Monthly | Baker, Bloom and Davis website |
| emv | US equity market volatility index | Monthly | Baker, Bloom and Davis website |
| real-un | Real uncertainty index | Monthly | Sidney Ludvigson's website |
| fin-un | Financial uncertainty index | Monthly | Sidney Ludvigson's website |
| mac-un | Macroeconomic uncertainty index | Monthly | Sidney Ludvigson's website |
| u-rate | US Unemployment Rate | Monthly | FRED |
| ec-act | Kilian's Real Economic Activity Index | Monthly | Lutz Kilian's website |
| VIX | CBOE Volatility Index (VIX) | Monthly and daily | CBOE |
| $f1, ..., f8$ | Ludvigson-Ng Term Structure factors | Monthly | Sidney Ludvigson's website |
| $y_1$ | log of 1-year bond yield | Monthly and daily | CRSP |
| yld-sprd | Yield Term Spread (5–1 year yield) | Monthly and daily | CRSP |
| $f_{ij}$ | Forward rate between years $j$ and $j$ | Monthly | CRSP (Fama–Bliss files) |
| fwd-sprd | Forward Spread $(f_{12} - y_1)$ | Monthly and daily | CRSP |
| AUD | Australian Dollar to 1 USD FX return | Monthly and daily | Datastream |
| CAD | Canadian Dollar to 1 USD FX return | Monthly and daily | Datastream |
| CHP | Chilean Peso to 1 USD FX return | Monthly and daily | Datastream |
| INR | Indian Rupee to 1 USD FX return | Monthly and daily | Datastream |
| NZD | New Zealand Dollar to 1 USD FX return | Monthly and daily | Datastream |
| SAR | South African Rand to 1 USD FX return | Monthly and daily | Datastream |
| BDI | Baltic Exchange Dry Index (growth rate) | Monthly and daily | Datastream |
| IP | US Industrial Production (growth rate) | Monthly | Datastream |
| MS | US Money Supply M1 (growth rate) | Monthly | Datastream |

*Note*: GW2012 denotes the data set in Welch and Goyal (2008), which is extended by the authors and available at http://www.hec.unil.ch/agoyal/; CBOE denotes the website of Chicago Board Options Exchange, http://www.cboe.com/micro/VIX/historical.aspx; FRED denotes the Federal Reserve Economic Data, the link to Lutz Kilian's website is http://www-personal.umich.edu/~lkilian/; the link to Baker, Bloom, and Davis website is https://www.policyuncertainty.com/all_country_data.html; CRSP, Center for Research in Security Prices; FX, exchange.