Received: 11 January 2022 | Revised: 22 November 2023 | Accepted: 25 April 2024

DOI: 10.1002/for.3152

RESEARCH ARTICLE

WILEY

Regime-dependent commodity price dynamics: A predictive analysis

Jesus Crespo Cuaresma[1,2,3] | **Ines Fortin**[4] | **Jaroslava Hlouskova**[2,4,5] | **Michael Obersteiner**[2,6]

[1]Department of Economics, Vienna University of Economics and Business, Vienna, Austria

[2]International Institute of Applied Systems Analysis (IIASA), Laxenburg, Austria

[3]Wittgenstein Center for Demography and Global Human Capital (IIASA,VID/OEAW,UniVie) Austrian Institute of Economic Research (WIFO), Vienna, Austria

[4]Macroeconomics and Economic Policy, Institute for Advanced Studies, Vienna, Austria

[5]Department of Economics, University of Economics in Bratislava, Bratislava, Slovakia

[6]Environmental Change Institute, University of Oxford, Oxford, UK

Correspondence
Jesus Crespo Cuaresma, Vienna University of Economics and Business, Vienna, Austria.
Email: jcrespo@wu.ac.at

Funding information
Austrian Science Fund

Abstract

We develop an econometric modelling framework to forecast commodity prices taking into account potentially different dynamics and linkages existing at different states of the world and using different performance measures to validate the predictions. We assess the extent to which the quality of the forecasts can be improved by entertaining different regime-dependent threshold models considering different threshold variables. We evaluate prediction quality using both loss minimization and profit maximization measures based on directional accuracy, directional value, the ability to predict turning points, and the returns implied by a simple trading strategy. Our analysis provides overwhelming evidence that allowing for regime-dependent dynamics leads to improvements in predictive ability for the Goldman Sachs Commodity Index, as well as for its five sub-indices (energy, industrial metals, precious metals, agriculture, and livestock). Our results suggest the existence of a trade-off between predictive ability based on loss and profit measures, which implies that the particular aim of the prediction exercise carried out plays a very important role in terms of defining which set of models is the best to use.

KEYWORDS

commodity prices, forecast performance, forecasting, states of economy, threshold models

1 | INTRODUCTION

This study aims at creating an econometric modelling framework to forecast commodity prices, taking explicitly into account the potentially different dynamics and linkages existing in different states of the world and using different performance measures to validate the predictions. The literature on commodity price forecasts can be categorized into two broad groups depending on the approach they take. While some studies use asset prices as predictors of commodity prices, a more agnostic approach exploits statistical methods to search for the most effective set of predictors of commodity price changes. The more common approach based on asset prices, routinely used by central banks, creates predictions of commodity prices using futures prices. Recently, some authors argue that such a forecasting method rather provides noisy signals about future spot prices (see Gorton & Rouwenhorst, 2006; Groen & Pesenti, 2011; Hong & Yogo, 2012).

The early literature on commodity price modelling and forecasting builds upon large macroeconometric specifications (Just & Rausser, 1981), while modern methods rely on univariate and multivariate time series modelling which jointly assess the dynamics of macroeconomic variables and commodity prices (see, e.g., Ahumada & Cornejo, 2015, 2016). Groen and Pesenti (2011) and Gargano and Timmermann (2014) provide relevant examples of the more agnostic and flexible approach to model building in the context of commodity price forecasting. In both studies, the authors assess whether forecasts of commodity prices based on a large pool of macroeconomic predictors can systematically improve upon naive benchmarks. Groen and Pesenti (2011) study the predictability of 10 commodity indices in an out-of-sample forecasting experiment. They conclude that neither commodity exchange rates nor a broad cross-section of macroeconomic variables produce overwhelmingly strong evidence of spot price predictability when compared with random walk or autoregressive benchmarks. Gargano and Timmermann (2014), on the other hand, examine the out-of-sample predictability of seven commodity indices over the period 1947–2010, using macroeconomic and financial variables. They find that commodity currencies have some predictive power at short (monthly and quarterly) forecast horizons, while growth in industrial production and the investment-capital ratio have some predictive power at longer (yearly) horizons, a result that resembles that by Chen et al. (2010). Other modelling frameworks aimed at forecasting short-term changes in agricultural commodity prices are employed in more recent contributions to the literature, such as those by Xu ((2017), (2018), (2020)). In parallel, efforts to improve forecasts of commodity prices by explicitly modelling their volatility have also been carried out (see, e.g., Bernard et al., 2008; Ramirez & Fadiga, 2003; or the recent contribution by ; Degiannakis et al., 2020).

In striving for modelling frameworks with good predictive accuracy for commodity prices, in this contribution, we assess the extent to which the quality of the forecasts depends on the state of the economy. Issues related to optimizing out-of-sample prediction in the presence of structural breaks and parameter instability have been particularly prevalent in the modern forecasting literature (see, e.g., Giacomini & Rossi, 2010). We aim at assessing whether, for example, models tend to provide more accurate predictions of commodity prices in calm than in turbulent times. First findings in this direction were provided by Gargano and Timmermann (2014), who observe that commodity price predictability is better during recessions than during expansions. In stock and bond markets, the importance of models that account for regime-dependent parameters has often been acknowledged. Recent studies (e.g., Guidolin & Timmermann, 2005; for excess stock and bond returns or Guidolin & Timmermann, 2009; for short-term interest rates) have found that regime switching models may prove extremely useful to forecast over intermediate horizons, using monthly data. Guidolin and Ono () find overwhelming evidence of regime switching in the joint process for asset prices and macroeconomic variables. They also find that modelling explicitly the presence of such regimes improves considerably the out-of-sample performance of a model of the linkages between asset prices and the macroeconomy. Guidolin and Pedio (2021) forecast commodity futures returns using a Markov-switching model that identifies different volatility regimes and maps the observations into high-volatility and low-volatility states. In addition, they find that the models that outperform under a statistical loss function are not necessarily the best when an economic loss function is used to evaluate the predictive performance of the different models. Jacobsen et al. (2019) investigate stock return predictability and find a strong positive relation between industrial metals and equity returns in times of recessions and a negative relation during expansions. In this study, we entertain different regime-dependent models (threshold models), considering different threshold variables to capture states of the world.

In addition, we assess the quality of commodity forecasts not only with the mean squared error (MSE), the traditional forecast performance measure used in many studies including Gargano and Timmermann (2014) but also with measures that evaluate directional accuracy (DA), directional value (DV), the ability to predict price movements when large swings take place, and returns implied by a trading strategy based on commodity price forecasts. These additional measures (profit measures, as opposed to the loss measures like mean-squared error or mean absolute error [MAE]) do not directly assess forecast accuracy but relate to other dimensions of forecasting quality and may be more relevant than accuracy for particular applications in policy and applied work.

We create models to predict commodity price dynamics as captured by the changes in an overall commodity price index, as well as in five subindices (energy, industrial metals, precious metals, agriculture, and livestock), for short- and long-term forecast horizons, using monthly observations in the period 1980–2018. Our forecast models include threshold models that are based on different threshold variables, and we consider the various performance measures discussed above. For the multivariate threshold models, we use the following variables:

composite leading indicator for the USA and real effective exchange rate of US dollar (macroeconomic variables), world stock market index (financial variable), and stock-to-use ratio[1] (fundamental variable). Based on the extensive empirical evidence, we find overwhelming evidence that allowing for regime-dependent dynamics leads to improvements in predictive ability for commodity prices. This is the case because the characteristics of the dynamics and the interactions with other variables are not constant over time but differ depending on particular phenomena (e.g., periods of high and low volatility, good and bad economic times, times of high/low interest rates or inflation). To the extent that the estimated models lead to stable dynamics, modelling the interactions in a regime-specific fashion allows for better predictions of commodity price changes. However, the nature of these improvements also differs across predictive measures and sectors.

Our results show that an interesting trade-off appears between loss and profit measures, which implies that the particular aim of the prediction exercise carried out plays a very important role in terms of defining which set of models is the best to use. Our results indicate a systematic correlation between loss-based and profit-based predictive error measures that suggests that correctly predicted directions of change tend to happen in periods where MSEs are particularly large. The optimal specifications for applications where the metrics for success are related to systematically predicting the direction of change of commodity prices accurately may thus be systematically different from those aimed at providing point predictions with an absolute minimal distance to the realized values. In the context of the existing literature, we employ a relatively large model space in terms of potential covariates and threshold variables, which can explain the differences in results as compared to other studies where the predictive performance of nonlinear models is humble compared with that of simpler linear specifications.

The paper is structured as follows. In Section 2, we present the forecast models, where we describe the class of threshold models, which are our main focus, in more detail. In Section 3, we introduce the commodity price data and the explanatory and threshold variables. We present forecast performance measures, including traditional and new measures, in Section 4. The following section presents and discusses the empirical results, and Section 6 concludes the study.

2 | METHODOLOGY

In order to address our research question, which deals with how different states of the economy (like recessions/expansions, high/low volatility, high/low inflation, high/low interest rates, market sentiment, ...) affect the price forecasting performance of different commodity classes, we assess threshold models (both univariate and multivariate). These types of models allow their parameters to change in different regimes (states of the world), whose occurrence depends on the value of a given threshold variable. In principle, there is a large universe of potential threshold variables that could be used as a trigger quantity which determines the regime where the process resides at a given moment. It has often been suggested, for example, that variables may behave differently in booming and declining markets. Hence, indicators describing different stages of the business cycle (e.g., business cycle indicators, economic sentiment indicators, inflation, spreads between long- and short-term interest rates) may prove useful in defining the corresponding states of the economy. On the other hand, the behaviour of economic variables may vary in periods of high and low risk, which are usually identified by a high or low volatility in the equity markets. The level of oil price inflation may also induce different types of dynamics in commodity prices. We also examine whether the use of threshold variables based on the rolling correlation between stock and government bond markets, as well as the correlation between stock and oil markets (which are both relevant in portfolio diversification) may lead to differences in the quality of commodity price forecasting models. Finally, we are interested in whether the level of the target variable itself, that is, the commodity index, may be useful to define different states of the world.

In our application, the set of variables that are assessed as potential drivers of the threshold-nonlinearity and thus define the states of the economy is given by te following: the composite leading indicator for the USA (CLI), the consumer confidence indicator for the USA (CCI), the USA inflation rate (INF), the spread between long-term and short-term US interest rates (spread), the volatility of the US stock market (VOLA), oil price inflation (Δoil), the correlation between the US stock and government bond markets based on a 6-month rolling window (COR), the correlation between the world stock market and the oil price based on a 6-month rolling window (COR-oil), the S&P Coldman Sachs commodity index (GSCI), and its subindices. For more details, see Table A2 in Appendix A.

As the set of potential specifications aimed at forecasting commodity prices, we consider a large battery of model classes, including autoregressive models, Bayesian vector autoregressive models, GARCH models, and vector error correction models. In addition to these specifications, which do not allow for threshold effects, we consider univariate and multivariate two-regime threshold models. In a preliminary analysis, we recursively tested for the optimal number of regimes in different threshold specifications making use of the test by Bai and Perron (2003). The data appear to strongly support two-regime models against threshold specifications with a higher number of regimes, which leads us to fix the number of thresholds to one throughout the study, thus reducing the computational costs involved in the analysis.[2] All the models employed are listed in Table 1. The simplest threshold model is the threshold autoregression in levels with p lags and with k lags in the threshold variable, TAR(p,k),

$$y_t = \begin{cases} \phi_{01} + \sum_{i=1}^{p} \phi_{i1} y_{t-i} + \varepsilon_t, & \text{for} z_{t-k} \leq \gamma_\phi \\ \phi_{02} + \sum_{i=1}^{p} \phi_{i2} y_{t-i} + \varepsilon_t, & \text{for} z_{t-k} > \gamma_\phi \end{cases} \tag{1}$$

where y_t is the log of the Goldman Sachs commodity index (or its subindex) at time t, $z \in \mathbf{Z}$, with $\mathbf{Z}$ being the set of above mentioned threshold variables, namely, $\mathbf{Z} = \{y$, CLI, CCI, INF, spread, VOLA, Δoil, COR, COR-oil$\}$. Finally, $\varepsilon_t \sim \text{NID}(0, \sigma_\varepsilon^2)$. The estimator of γ_ϕ is the value of z that minimizes the sum of squared residuals in the nonlinear regression (1), that is,

$$\hat{\gamma}_\phi = \arg\min_z \left\{ \sum \hat{\varepsilon}(z)^2 \right\}. \tag{2}$$

Once the estimator of γ_ϕ is found, (1) can be estimated in a straightforward manner making use of OLS.

Given that the objective of the analysis is to assess the relative performance exclusively in terms of out-of-sample predictive power and exploiting a large space of specifications, we entertain both models with variables in first differences and models where the variables are included in levels. We also consider threshold autoregressions in first differences with p lags and with a k-th lag in threshold variable, TDAR(p,k)

$$\Delta y_t = \begin{cases} \theta_{01} + \sum_{i=1}^{p} \theta_{i1} \Delta y_{t-i} + \epsilon_t, & \text{for} z_{t-k} \leq \gamma_\theta \\ \theta_{02} + \sum_{i=1}^{p} \theta_{i2} \Delta y_{t-i} + \epsilon_t, & \text{for} z_{t-k} > \gamma_\theta \end{cases} \tag{3}$$

TABLE 1 Model description.

Abbreviations	Model description
AR(p)	Autoregression in levels with p lags
DAR(p)	Autoregression in first differences with p lags
s-AR(p)	Subset autoregression in levels with p lags
s-DAR(p)	Subset autoregression in first differences with p lags
ARCH(p,q)	Autoregression conditional heteroskedasticity in levels with p lags in mean equation and q lags in variance equation
DARCH(p,q)	Autoregression conditional heteroskedasticity in first differences with p lags in mean equation and q lags in variance equation
GARCH(p,q)	Generalized autoregression conditional heteroskedasticity in levels with p lags in mean equation and q lags in variance equation
DGARCH(p,q)	Generalized autoregression conditional heteroskedasticity in first differences with p lags in mean equation and q lags in variance equation
TAR(p,k)	Threshold autoregression in levels with p lags and with k-th lag in threshold variable
TDAR(p,k)	Threshold autoregression in first differences with p lags and with k-th lag in the threshold variable
VAR(p)	Vector autoregression in levels with p lags
DVAR(p)	Vector autoregression in first differences with p lags
VEC(p, c)	Vector error correction model with p lags and c cointegration relationships
s-VAR(p)	Subset vector autoregression in levels with p lags
s-DVAR(p)	Subset vector autoregression in first differences with p lags
BDVAR(p)	Bayesian vector autoregression in first differences with p lags
TVAR(p,k)	Threshold vector autoregression in levels with p lags and with k-th lag in threshold variable
TDVAR(p,k)	Threshold vector autoregression in first differences with p lags and with k-th lag in threshold variable
RW	Random walk

where $\epsilon_t \sim \mathrm{NID}(0,\sigma_\epsilon^2)$, $\hat{\gamma}_\theta = \arg\min_z\left\{\sum \hat{\epsilon}(z)^2\right\}$ and $z \in \mathbf{Z}$.

In addition to univariate threshold models, we entertain multivariate threshold models, which generalize the class of threshold vector autoregression in levels with p lags and with a k-th lag threshold variable, TVAR (p,k). Let x_t be an N-dimensional vector, then the model under consideration is

$$x_t = \begin{cases} \Psi_{01} + \sum_{l=1}^{p} \Psi_{l1} x_{t-l} + \mu_t, & \text{for} z_{t-k} \leq \gamma_\Psi \\ \Psi_{02} + \sum_{l=1}^{p} \Psi_{l2} x_{t-l} + \mu_t, & \text{for} z_{t-k} > \gamma_\Psi \end{cases} \tag{4}$$

where Ψ_{01} and Ψ_{02} are N-dimensional column vectors, Ψ_{l1} and Ψ_{l2} are $N \times N$ matrices, $\mu_t \sim \mathrm{NID}(0,\Sigma_\mu)$, the S&P GS commodity index (or its sub-index) is the first element of x_t, that is, $x_{t1} = y_t = \log(GSCI_t)$ and $z \in \mathbf{Z}$. Finally, γ_Ψ is estimated such that

$$\hat{\gamma}_\Psi = \arg\min_z \left\{\sum \hat{\mu}_1(z)^2\right\} \tag{5}$$

thus, the estimator of γ_Ψ is the value of z that minimizes the sum of squared residuals corresponding to the first equation in (4), that is, the residuals corresponding to the commodity index. Vector x_t consists of the following macroeconomic and financial variables: the US composite leading indicator (CLI), the real effective exchange rate with respect to the US dollar (REER), the world stock market index (stock), stock-to-use ratios,[3] and additionally the S&P Goldman Sachs commodity index (GSCI) if the dependent variable is a commodity subindex. With the use of these variables, the aim is to quantitatively approximate shifts in the commodity demand and supply curves and to incorporate changes in expectations for the global economic situation. Similar variables are employed in Crespo Cuaresma et al. (2021) to forecast agricultural commodity prices. All variables are logged, with the exception of the stock-to-use ratios.

Finally, we consider also a variation of threshold vector autoregression in first differences with p lags and with k-th lag in threshold variable, TDVAR(p,k) such as

$$\Delta x_t = \begin{cases} \chi_{01} + \sum_{l=1}^{p} \chi_{l1} \Delta x_{t-l} + u_t, & \text{for} z_{t-k} \leq \gamma_\chi \\ \chi_{02} + \sum_{l=1}^{p} \chi_{l2} \Delta x_{t-l} + u_t, & \text{for} z_{t-k} > \gamma_\chi \end{cases} \tag{6}$$

with parameter vectors and matrices defined analogously to those in the model above and $u_t \sim \mathrm{NID}(0,\Sigma_u)$, $z \in \mathbf{Z}$. The threshold value γ_χ is estimated such that

$$\hat{\gamma}_\chi = \arg\min_z \left\{\sum \hat{u}_1(z)^2\right\} \tag{7}$$

Thus, the estimator of γ_χ is the value of z that minimizes the sum of squared residuals corresponding to the first equation in (6), that is, the residuals corresponding to the commodity index in first differences ΔGSCI. As in (4), the regimes are implied by the first equation and taken as given for the remaining equations in (6). With the choice of a threshold value that minimizes the sum of squared residuals of the commodity price regression equation, we aim at optimizing predictive ability for our objective variable and ensure that the nonlinearities identified are related to the dynamics of commodity prices.[4]

In our empirical analysis, when we compare threshold and linear models, we consider up to three lags of the variables (with $p = 3$ being the maximum lag length) and up to 12 lags for the threshold variable under consideration (with $k = 12$ being the maximum lag length). Models are compared and selected according to out-of-sample performance measures. We explicitly consider all combinations of explanatory variables and all lags of the explanatory and threshold variables up to the specified maximal lag lengths and choose the best model according to the given forecast performance measure. It should be noted that the space of models we address implies that we are agnostic about the time series properties of commodity prices, with particular specifications building upon the assumption of mean reversion, while others assume nonstationary behaviour of the commodity price variable. Since we address different predictive measures and use a rolling window design for the forecast validation, exploiting short-term mean reverting dynamics may actually lead to satisfactory predictions in particular periods. Such an approach makes it particularly difficult for nonlinear models to achieve superior predictive ability in a systematic manner.

3 | DATA

We use the family of S&P GSCI (Standard & Poors Goldman Sachs Commodity Index) indices to measure commodity prices. We employ both the total aggregate commodity index (S&P GSCI) and five subindices that reflect the developments of certain components of the index, namely energy (with a weight in the total commodity index of 63%), industrial metals (with a weight of 11%), precious metals (with a weight of 4%), agriculture (with a weight of 15%), and livestock (with a weight of 7%). The S&P GSCI is regarded as a benchmark for investment in commodity markets and is designed to be a tradable index. It is calculated using a world productionweighted basis and includes physical commodities that are traded in liquid futures markets. The criteria for inclusion into the index are based on trading volume. In addition, the contracts must be denominated in US dollars and traded in an OECD country or on a trading facility that has its principal place of business in an OECD country. The current S&P GSCI comprises 24 commodities from all commodity sectors with a high exposure to energy. These energy contracts include crude oil, heating oil, and gasoline traded in the US, as well as crude oil and gasoil traded in Europe. Table A1 in Appendix A lists all contracts included in the S&P GSCI and their respective weights and trading places. We consider the class of total return indices.[5] For more information on the S&P GSCI, see S&P Dow Jones (2019). Some descriptive statistics related to the indices are given in Table 2. Price developments are quite heterogeneous across indices, with only the overall and the energy indices displaying rather similar dynamics. The volatility in returns varies considerably as well, which has a direct impact on the forecasting accuracy of econometric models. The monthly returns of the energy index, for example, show a standard deviation of 7.7% over the total data sample (1980–2018), while the corresponding value for the livestock index is only 3.5%. Overall, the correlations between different commodity sector returns are low (with the exception of the overall index and the energy index), which reinforces the need to analyze the different sectors separately.

As macroeconomic and finance variables in our models, we take the composite leading indicator for the USA (CLI), the real effective exchange rate related to the US dollar (REER), and the world stock market indicator (stock).[6] In addition, we employ fundamental variables summarizing the forces in the commodity market: stock-to-use ratios (stu) for oil (worldwide), wheat (USA), and meat (USA). More precisely, we use the worldwide oil stock-to-use ratio for the aggregate index and for the subindices energy, industrial metals, and precious metals, we use the US wheat stock-to-use ratio for the subindex agriculture, and we use the US meat stock-to-use ratio for the subindex livestock. In those cases where we model commodity subindices, we also use the total commodity index as an additional variable. As threshold variables, in addition to lagged values of the modelled index itself, we use the composite leading indicator for the USA (CLI), the consumer sentiment indicator for the US (CCI), the US inflation measured by the consumer price index (INF), the spread between long-term and short-term US interest rates (spread), the volatility of the S&P 500 (VOLA), the oil price inflation (Δoil), the correlation between the US stock and government bond markets (COR), and the correlation between the global stock market and the oil market (COR-oil). The correlations are calculated between daily returns in the respective markets, over a rolling window of 130 trading days (i.e., approximately 6 months), recorded at the end of a given month. For details on all the data we use, see Table A2.

The data sample covers monthly observations for the period ranging from January 1980 through December 2018. We consider rolling-window estimation for our analysis, that is, we keep the size of the estimation sample constant and equal to 20 years and move forward the sample by one month while re-estimating the model parameters. The use of a rolling window for the predictive assessment of the models allows our class of threshold models to better identify changes in regimes if they

TABLE 2 Summary statistics for commodity returns.

	All	Energy	Ind. met.	Prec. met.	Agriculture	Livestock	Stock
Descriptive statistics							
Mean (%)	0.3771	0.6246	0.5491	0.1750	−0.0300	0.3361	0.6706
Std. (%)	4.8416	7.6745	5.4517	4.3801	4.3028	3.5431	3.6639
Skew.	−0.6128	0.1151	0.1907	0.0174	0.5327	0.0220	−0.8491
Kurt.	5.6013	5.7475	6.6736	6.1019	6.5898	3.7059	7.3947
Correlation matrix							
Energy	0.9406	1					
Ind. met.	0.3947	0.2819	1				
Prec. met.	0.2031	0.1398	0.2738	1			
Agriculture	0.3274	0.1419	0.2362	0.1843	1		
Livestock	0.1953	0.0780	0.0672	−0.0534	0.0611	1	
Stock	0.2712	0.1922	0.3039	0.1306	0.1888	0.1180	1

Note. The table reports the mean, standard deviation, skewness, and kurtosis for monthly commodity returns over the sample period from January 1980 to December 2018. Commodity returns are computed from the S&P GSCI commodity indices. The last column shows returns of the world stock market index.

happen at the end of the in-sample period, which could prove important to preserve predictive ability. The rolling-window design should thus avoid that the thresholds identified are exclusively driven by nonlinear behaviour at the beginning of the available sample. The out-of-sample period used to evaluate the forecast performance spans from January 2005 to December 2018.[7] Note that "best" models are chosen based on the forecast performance of the individual models for all lags (up to specified maximum lags) and all combinations of variables under consideration.

4 | FORECAST EVALUATION

The evaluation of different commodity price forecasts are carried out employing not only traditional loss measures, like MAE and MSE, but also profit-based measures like DA, DV, and DV of turning points (TP). The latter might be more relevant in situations where getting the right future value of commodity prices may be of lesser importance than predicting their direction of change, in particular if the change in prices is large. The DA indicator, or hit rate, is a binary variable measuring whether the direction of a price change was correctly forecast. The DV incorporates the economic value of directional forecasts by assigning to each correctly predicted change its magnitude. The DA of TPs describes the ability to predict TPs in commodity price dynamics.[8]

The loss-based and profit-based performance measures are formally defined as follows:

$$\begin{aligned}
AE_{t+h,h} &= \left|\log\hat{P}_{t+h|t} - \log P_{t+h}\right| \\
SE_{t+h,h} &= \left(\log\hat{P}_{t+h|t} - \log P_{t+h}\right)^2 \\
DA_{t+h,h} &= I\left(\operatorname{sgn}(P_{t+h} - P_t) = \operatorname{sgn}(\hat{P}_{t+h|t} - P_t)\right) \\
DV_{t+h,h} &= |P_{t+h} - P_t|\ DA_{t+h,h} \\
TP_{t+h,h} &= \begin{cases} DA_{t+h,h} & \text{if } \operatorname{sgn}(P_{t+h} - P_t) \times \operatorname{sgn}(P_t - P_{t-h}) = -1 \\ 0 & \text{otherwise} \end{cases}
\end{aligned} \tag{8}$$

where P_t is the price of the commodity index at time t, $\hat{P}_{t+h|t}$ is the forecast of the price of the commodity index for time $t+h$ conditional on the information available at time t, that is, h is the forecast horizon, and $I(\cdot)$ is the indicator function.

In addition, we consider forecast ability measures based on the returns implied by predicting commodity prices and using a simple "buy low, sell high" trading strategy. This strategy is based on buying the commodity index if its price is forecast to rise and selling it when its price is forecast to fall. This strategy is described (for exchange rates), for example, in Gençay (1998) and will be used under the assumption of no transaction costs.[9] Predicted upward movements of the commodity index with respect to the actual value (positive returns) are executed as long positions, while predicted downward movements (negative returns) are executed as short positions. The following discrete return $r_{t+h,h}$ is implied by the "buy low, sell high" trading strategy,

$$r_{t+h,h} = \begin{cases} \dfrac{1}{P_t}(P_t - P_{t+h}) = 1 - \dfrac{P_{t+h}}{P_t}, \\ \qquad\qquad \text{if } \underset{t+h|t}{P} < P_t \\ \text{commodity index is bought at } t+h \\ \dfrac{1}{P_t}(P_{t+h} - P_t) = \dfrac{P_{t+h}}{P_t} - 1, \\ \qquad\qquad \text{if } \underset{t+h|t}{P} > P_t \\ \text{commodity index is sold at } t+h \end{cases} \tag{9}$$

Later on, we will sometimes refer to the return implied by this trading strategy simply as the return.

The aggregate performance measures for each model are calculated over the out-of-sample period for a given forecasting horizon as follows:

$$\begin{aligned}
MSE_h &= \sum_{j=0}^{T_2-T_1} \frac{SE_{T_1+j,h}}{T_2 - T_1 + 1} \\
MAE_h &= \sum_{j=0}^{T_2-T_1} \frac{AE_{T_1+j,h}}{T_2 - T_1 + 1} \\
DA_h &= 100 \sum_{j=0}^{T_2-T_1} \frac{DA_{T_1+j,h}}{T_2 - T_1 + 1} \\
DV_h &= 100 \frac{\sum_{j=0}^{T_2-T_1} DV_{T_1+j,h}}{\sum_{j=0}^{T_2-T_1} |S_{T_1+j} - S_{T_1+j-h}|} \\
&= 100 \frac{\sum_{j=0}^{T_2-T_1} |\hat{S}_{T_1+j} - S_{T_1+j-h}| DA_{T_1+j,h}}{\sum_{j=0}^{T_2-T_1} |S_{T_1+j} - S_{T_1+j-h}|} \\
TP_h &= 100 \frac{\sum_{j=0}^{T_2-T_1} TP_{T_1+j,h}}{\sum_{j=0}^{T_2-T_1} TP^{actual}_{T_1+j,h}}
\end{aligned}$$

where

$$TP^{actual}_{t+h,h} = \begin{cases} 1 & \text{if } \operatorname{sgn}(P_{t+h} - P_t) \times \operatorname{sgn}(P_t - P_{t-h}) = -1 \\ 0 & \text{otherwise} \end{cases}$$

$$R_h = 100 \left[\left(\sum_{j=0}^{T_2-T_1} \frac{r_{T_1+j,h}}{T_2 - T_1 + 1} + 1 \right)^{12/h} - 1 \right]$$

where $T_0 =$ January 1980, $T_1 =$ January 2005, and $T_2 =$ December 2018.

The aim of our analysis is to evaluate the potential improvement in out-of-sample predictive ability for commodity prices that can be obtained by entertaining different regime-dependent threshold models (i.e., models where threshold effects are triggered by different variables). In this respect, the linear alternative plays the role of a general benchmark, so as to answer the question: Can threshold models improve predictions compared to models that do not include regime dependence? In addition, individual threshold models also appear as a benchmark reference in our comparisons when we look for an answer to the question whether particular threshold variables lead to better predictive performance than others.

5 | RESULTS

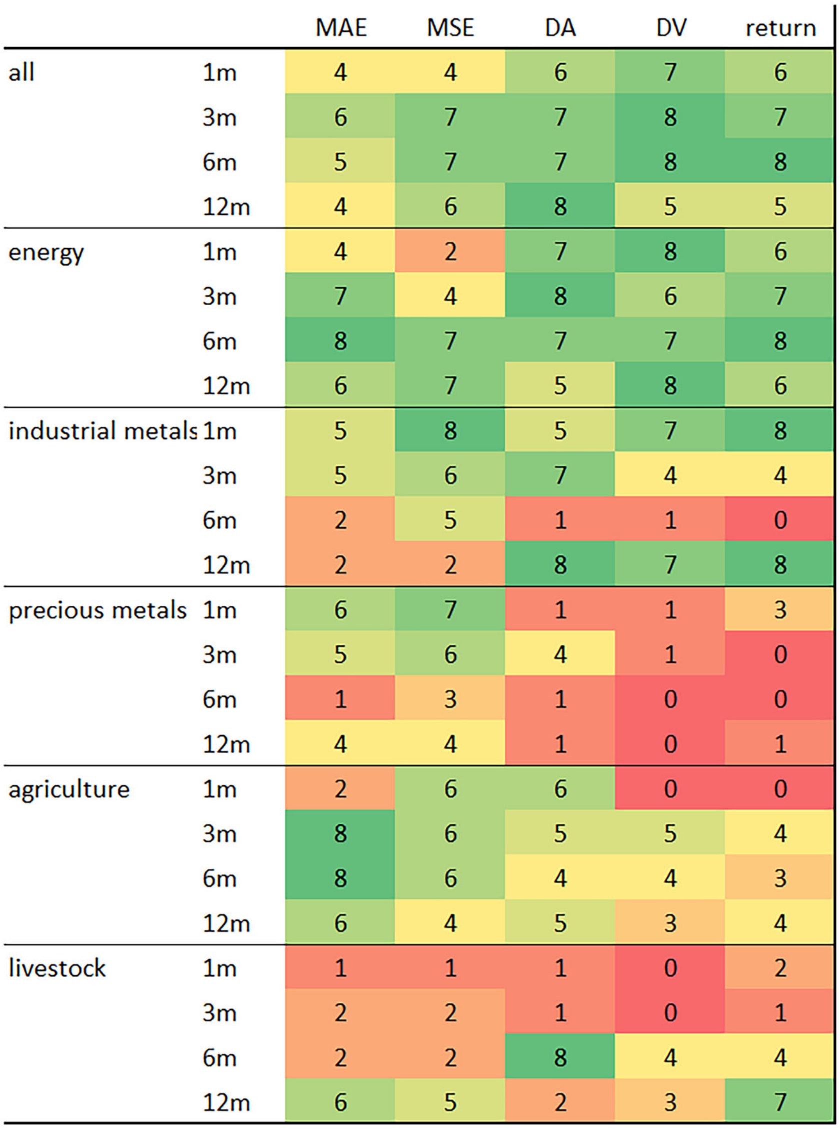
When analyzing the precision of threshold models in commodity price forecasting, we focus on different metrics. At first, we compare threshold models with linear models, which is the most natural benchmark to find out about the value of threshold models as predictive instrument. In this context, we also analyze the differences across threshold models implied by the use of different threshold variables. We employ different performance metrics to evaluate the forecasting performance and consider both total and regime-specific accuracy measures. In addition to assessing the models in terms of predictive power, we also examine the nature of the threshold variables and selected explanatory variables in the best threshold models and also discuss the pattern of forecasting performance for the two regimes. Furthermore, we look at the sector-specific performance of best threshold models. Finally, we compare threshold models with a larger set of models to find out whether threshold models tend to outperform specifications created out of this expanded set of covariates and consider the additional performance measure related to forecasting TPs.

5.1 | Threshold models and linear models

Our primary focus is to compare the performance of best threshold models (for a given threshold variable) with the performance of linear models. The threshold models entertained contain (vector) autoregression threshold models in levels and differences (TAR, TDAR, TVAR, and TDVAR), including self-exciting threshold autoregressive models, and linear models, that is, (vector) autoregressive specifications in levels and differences (AR, DAR, VAR, and DVAR), as described in Table 1. We examine threshold models where the threshold variable presents stationary behaviour and thus restrict the following variables to act as threshold variables: the lagged value of the dependent variable, the composite leading indicator for the US (CLI), the consumer sentiment indicator for the US (CCI), the US inflation measured by the consumer price index (INF), the spread between long-term and short-term US interest rates (spread), the volatility of the S&P 500 (VOLA), oil price inflation (Δoil), the correlation between the US stock and government bond markets (COR), and the correlation between the global stock market and the oil market (COR-oil).

Before we examine the relative performance of threshold versus linear models, we investigate the performance of threshold variables other than the dependent variable itself and examine whether different threshold variables imply large differences in the forecasting performance of their corresponding specifications. We therefore compare the performance of the best self-exciting threshold model with that of the best threshold model when the threshold variable is one of the other eight threshold variables listed in Table A2. With this exercise, we assess whether states of the world defined by the commodity price itself are informative enough to capture the economic environment implied by various other threshold variables. Figure 1 shows how many threshold models (from the maximum number of eight threshold variables) outperform the self-exciting model, for different performance measures, different forecast horizons, and the various commodity sectors. Our results suggest that the use of other threshold variables different from the overall commodity price index adds predictive information to our models. The self-exciting model is only better than any other threshold model for the index corresponding to precious metals, agriculture, and livestock when considering profit measures (DV and return). For the overall GSCI index, at least half of the threshold models outperform the self-exciting specification for all forecast horizons, irrespective of which performance measure used. This implies that explicitly acknowledging information like economic sentiment, uncertainty, interest rate spread, oil prices, or correlation can help to improve commodity price forecasting. Results are somewhat less clear-cut for energy, industrial metals and agriculture, and they are the least strong for precious metals and livestock. Even in these two sectors, however, in most cases, the best threshold models in terms of forecasting ability are not the self-exciting ones.

We turn to comparing the performance of the best threshold model with the performance of the linear counterpart that uses the same variables and lag structure. We compare the predictive performance over the whole out-of-sample period, as well as in the two regimes implied by the threshold model separately. The best threshold



		MAE	MSE	DA	DV	return
all	1m	4	4	6	7	6
	3m	6	7	7	8	7
	6m	5	7	7	8	8
	12m	4	6	8	5	5
energy	1m	4	2	7	8	6
	3m	7	4	8	6	7
	6m	8	7	7	7	8
	12m	6	7	5	8	6
industrial metals	1m	5	8	5	7	8
	3m	5	6	7	4	4
	6m	2	5	1	1	0
	12m	2	2	8	7	8
precious metals	1m	6	7	1	1	3
	3m	5	6	4	1	0
	6m	1	3	1	0	0
	12m	4	4	1	0	1
agriculture	1m	2	6	6	0	0
	3m	8	6	5	5	4
	6m	8	6	4	4	3
	12m	6	4	5	3	4
livestock	1m	1	1	1	0	2
	3m	2	2	1	0	1
	6m	2	2	8	4	4
	12m	6	5	2	3	7

FIGURE 1 Number of threshold models outperforming the self-exciting threshold specification. *Note*: The heatmap shows the result of comparing the best threshold model for a given threshold variable other than the dependent variable with the best threshold model for the threshold variable being the dependent variable. The numbers indicate the number of threshold variables where the best model outperforms the best self-exciting threshold model. Eight different threshold models are employed.

models with respect to specific threshold variables mostly outperform the corresponding linear models and also the best linear models.[10] In addition, threshold models outperform the corresponding linear specifications in at least one regime, mostly, however, in both regimes. In Table 3, we show the performance of the best threshold model and the performance of the corresponding linear model for the aggregate GSCI, for the threshold variable “spread” (difference between long and short-term US interest rates), as a representative example of the results obtained. This particular class of models was chosen based on the best short-term forecasting performance (MSE, 1-month ahead) for the overall GSCI index. For horizons of 1, 3, 6, and 12 months ahead, the total performance of the best threshold model is better than that of the corresponding linear model. When comparing the regime-based performance of the best threshold model and the regime-based performance of the corresponding linear model, the best threshold model outperforms the corresponding linear model in both regimes in most of the cases (17 out of 20 cases), and the threshold model is never outperformed by the corresponding linear specification in both regimes.[11] In addition, we also present the results of the Diebold–Mariano test of equal forecasting accuracy of the best threshold model against the corresponding linear model (Diebold & Mariano, 1995), which indicate statistically significant differences in predictive performance for many of the forecast error measures, in particular, for longer term forecasting.[12]

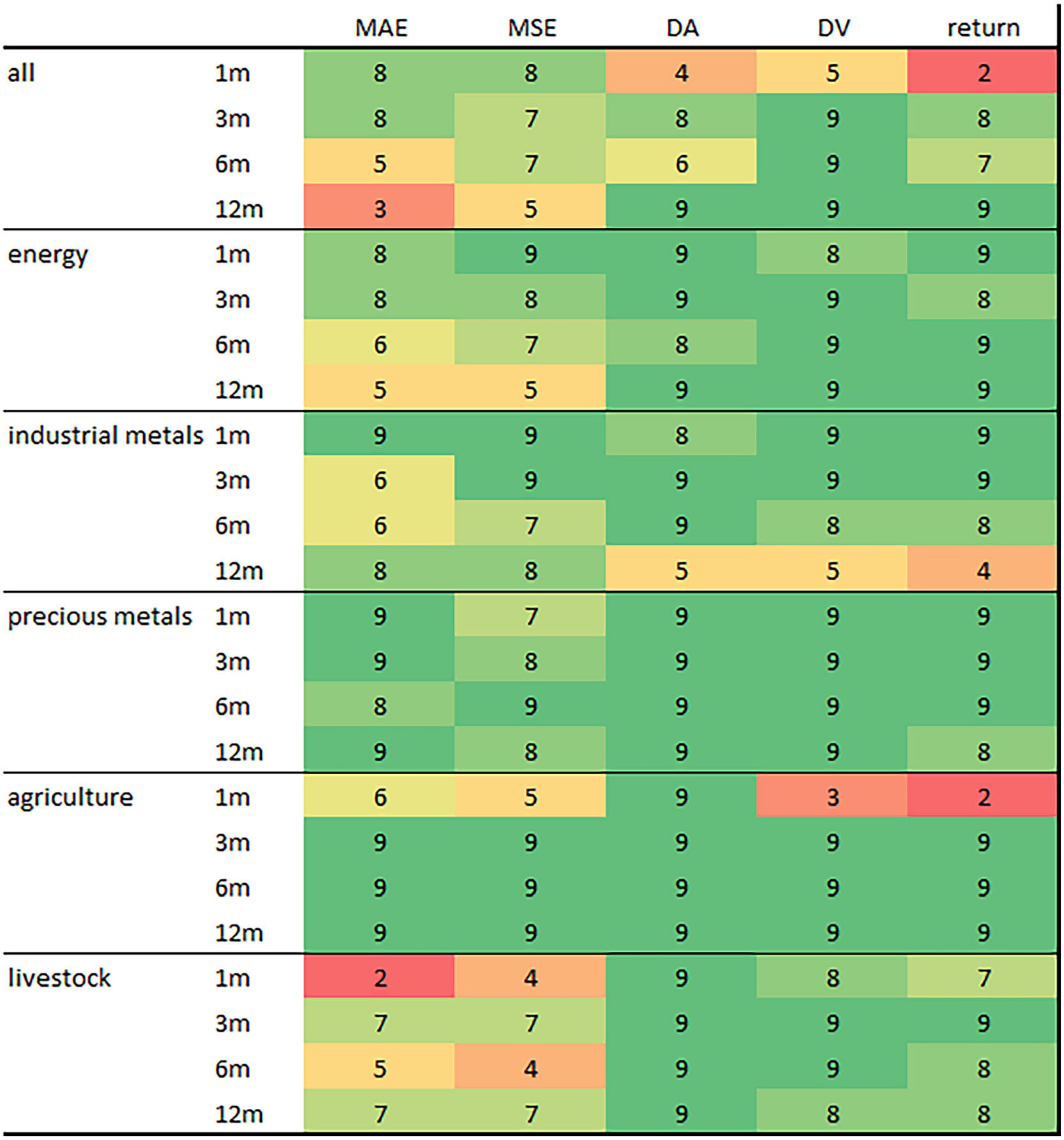
As a next step, we evaluate whether the total performance of the best threshold model is better than the total performance of the best linear model (out of *all* possible linear models, not just those including similar variables). The best threshold model (across all threshold variables) always outperforms the best linear model if we consider mean values of the performance criteria over the full out-

TABLE 3 Performance of best threshold model for the spread as a threshold variable against the corresponding linear model for the aggregate commodity price index.

		MAE	MSE	DA	DV	Return
1-month horizon		TDVAR(1,8)	TDVAR(1,5)	TDVAR(3,2)	TDVAR(1,1)	TDVAR(1,1)
		1100	1110	1000	1100	1100
Threshold	Total	4.06	0.28	68.45*	76.85*	29.75*
	Regime 1	4.59	0.23	68.97	80.42*	34.98*
	Regime 2	3.84	0.29	68.18	51.11	2.06
Linear	Total	4.21	0.30	63.69	68.19	20.97
	Regime 1	4.83	0.27	65.52	70.72	24.85
	Regime 2	3.94	0.31	62.73	49.91	-0.07
3-month horizon		TDVAR(1,12)	TDVAR(1,12)	TDVAR(2,1)	TDVAR(2,5)	TDVAR(2,5)
		1000	1000	1110	1110	1110
Threshold	Total	8.70	1.46	67.86*	78.34	20.70
	Regime 1	15.16	4.58	66.67*	72.50	17.98
	Regime 2	7.83	1.04	75.00	80.20	21.34
Linear	Total	9.22	1.71	61.90	56.16	9.72
	Regime 1	16.74	6.18	59.03	61.31	9.51
	Regime 2	8.20	1.11	79.17	54.52	9.77
6-month horizon		TDVAR(3,2)	TVAR(2,12)	TDVAR(2,2)	TDVAR(2,2)	TDVAR(2,2)
		1100	0010	1010	1010	1010
Threshold	Total	13.27**	4.24	67.26	77.16	13.87
	Regime 1	15.07	11.94	65.71	69.22	10.74
	Regime 2	12.30***	3.25*	67.67	80.08	14.71
Linear	Total	14.46	4.76	64.88	63.94	10.27
	Regime 1	15.22	11.48	62.86	51.55	4.41
	Regime 2	14.05	3.90	65.41	68.51	11.84
12-month horizon		TVAR(2,10)	TVAR(2,10)	TVAR(2,1)	TVAR(2,11)	TVAR(2,1)
		1100	1100	1010	1100	1100
Threshold	Total	20.21**	7.85*	68.45*	74.60**	8.87**
	Regime 1	18.27*	6.81	66.67*	82.68*	9.67***
	Regime 2	20.74	8.13	79.17	71.00	4.04
Linear	Total	25.00	10.82	54.76	46.91	−0.27
	Regime 1	28.02	11.14	51.39	29.36	−1.18
	Regime 2	24.17	10.73	75.00	54.76	5.13

Note: * (**/***) Indicates rejection of the null hypothesis of equal forecasting accuracy between the best threshold model and the corresponding linear model at 10% (5%/1%). The four-digit combination of ones and zeros below the model shows the inclusion (1) or exclusion (0) of the explanatory variables CLI, REER, stock market index, and oil stock-to-use ratio. Petrol shading indicates that the best threshold model outperforms the best linear model. Light petrol shading shows better total performance between best threshold model and corresponding linear model. Red shading indicates better regime-based performance between best threshold model and corresponding linear model. Regime 1 is defined by spread$_{t-k} < \gamma$, while regime 2 is defined by spread$_{t-k} > \gamma$.
Abbreviations: DA, directional accuracy; DV, directional value; MAE, mean absolute error; MSE, mean squared error.

of-sample period. Furthermore, in virtually all cases, the best threshold model for *any* given threshold variable outperforms the best linear model. Figure 2 presents the results of the analysis by showing the number of threshold models that outperform the best linear models for the different error measures, horizons, and commodity price subindices. The superiority of threshold models is systematic across all dimensions and can be observed when considering the distribution of the difference in squared prediction errors between the best linear and and the best threshold models. Figure 3 presents boxplots of these differences across all commodity sectors for forecast



		MAE	MSE	DA	DV	return
all	1m	8	8	4	5	2
	3m	8	7	8	9	8
	6m	5	7	6	9	7
	12m	3	5	9	9	9
energy	1m	8	9	9	8	9
	3m	8	8	9	9	8
	6m	6	7	8	9	9
	12m	5	5	9	9	9
industrial metals	1m	9	9	8	9	9
	3m	6	9	9	9	9
	6m	6	7	9	8	8
	12m	8	8	5	5	4
precious metals	1m	9	7	9	9	9
	3m	9	8	9	9	9
	6m	8	9	9	9	9
	12m	9	8	9	9	8
agriculture	1m	6	5	9	3	2
	3m	9	9	9	9	9
	6m	9	9	9	9	9
	12m	9	9	9	9	9
livestock	1m	2	4	9	8	7
	3m	7	7	9	9	9
	6m	5	4	9	9	8
	12m	7	7	9	8	8

FIGURE 2 Number of threshold models outperforming the best linear model. *Note*: The graphs show a comparison of the best threshold model (for a given threshold variable, including the dependent variable) with the best linear model. The numbers indicate how many of the best threshold models outperform the best linear model. The maximum possible number is nine.

horizons of one and twelve months. The average difference is always positive, and the support of the distribution varies substantially across the different commodity sectors. The pattern observed is relatively similar for forecast horizons of one and twelve months: The interquartile range is largest in the energy sector and smallest in the livestock and precious metals sectors for both forecast horizons.

5.2 | Threshold and explanatory variables

Analyzing the best performing threshold models with respect to threshold variables across commodity sectors, a pattern can be extracted (see Figure 4, which presents the ranking of models by threshold variable). For most of the commodity price indices, as well as for the general index, the threshold variables which tend to systematically appear in the best forecasting models in terms of MSE are the spread, the correlation between the US stock and government bond markets, and the composite leading indicator and inflation. The results indicate that capturing the dynamics of particular commodity markets requires different threshold variables. For example, while the correlation between stock and bond markets appears as a good predictor of regime changes in industrial metals, precious metals, agriculture and livestock, it performs weakly in the energy sector. The differences between loss and profit measures of predictive error are

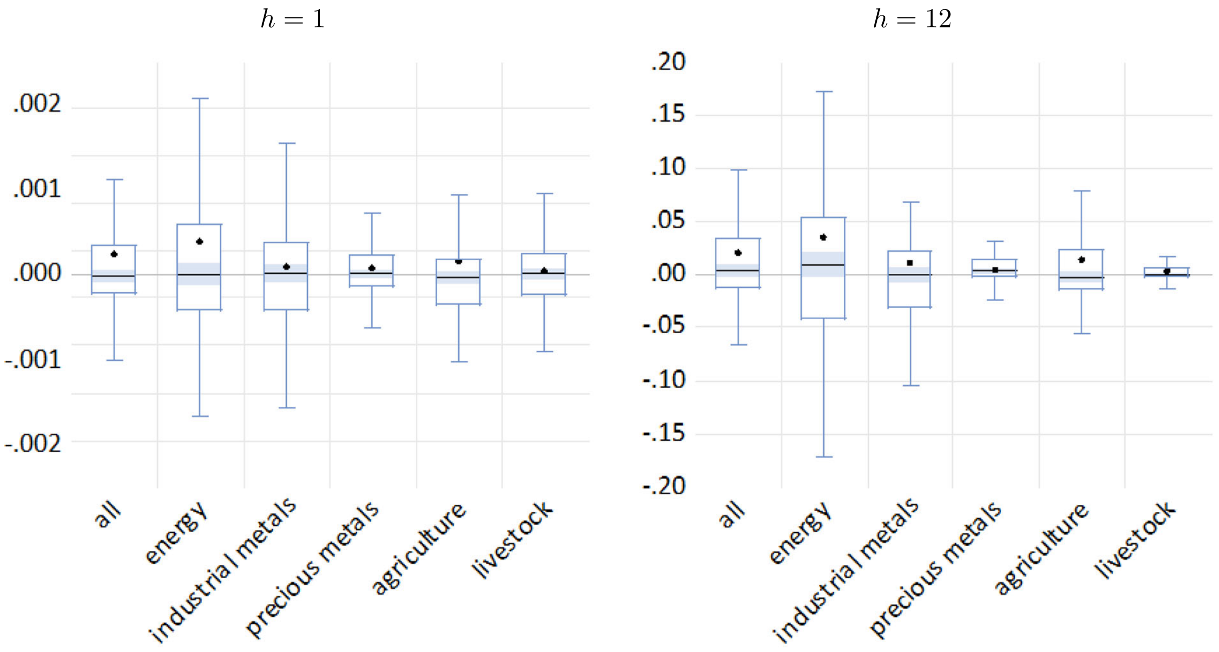
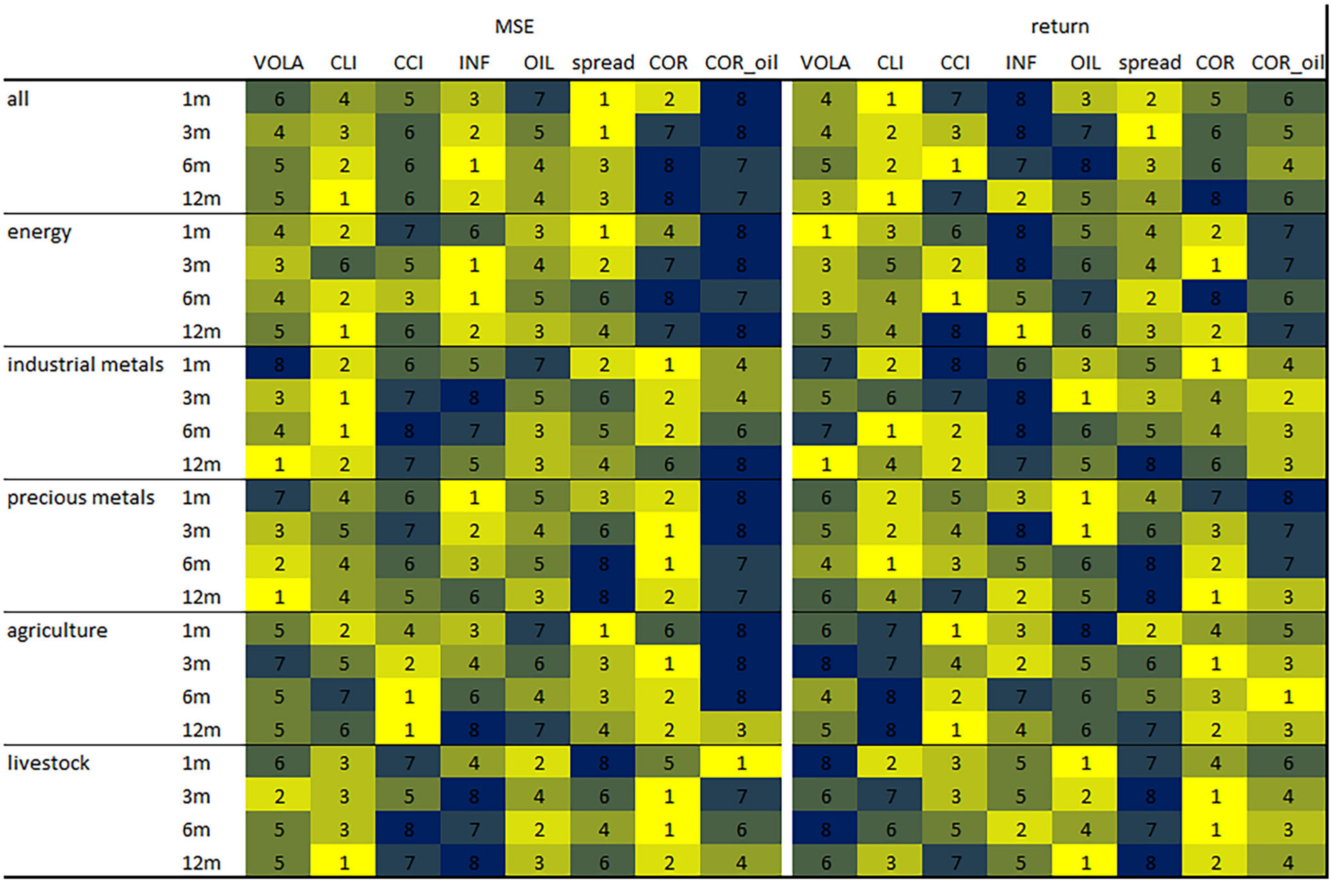


FIGURE 3 Difference between squared errors for best linear and best threshold models. *Note*: The graphs show boxplots of the differences between the squared errors for the best linear model and the squared errors for the best threshold model, for forecast horizons of one (left) and twelve (right) months. The differences are taken such that a higher mass in the positive region (or a positive mean) indicates a better performance of the threshold model.

remarkable: while using the correlation between stocks and bond markets as a threshold variable leads also to clear return predictive gains in industrial metals, precious metals, agriculture and livestock, in other sectors, the best performing threshold variable changes depending on the predictive error measure used.

In general, the forecasting performance of different best threshold models (implied by the different threshold variables) does not vary substantially. Table 4 provides some information on the variability of predictive performance across best threshold models, as compared with that of the best linear models. In particular, the table reports the average deviation of the predictive error of the best threshold models for a certain threshold variable from the *best overall* threshold model ("average deviation"), in proportion to the deviation of the best linear model from the best threshold model ("linear deviation"). Note that the best threshold model is always better than the best linear model; the "average" threshold model, however, may be worse than the best linear model (implied by a ratio larger than one in the table). The latter is rarely the case. In almost all cases (111 out of 120), the average deviation is smaller than the linear deviation (reflected by a number in the table that is smaller than one) and often to a very large extent. In a clear majority of all cases, the average deviation is less than half the linear deviation, implying that, in general, threshold models seem to perform (similarly) well and considerably better than the best linear model.

We turn to examining the nature of the variables included in the set of best threshold models, so as to assess the relative importance of different theoretical drivers of commodity price dynamics. Within the group of best linear models, one group of commodity indices can be found whose explanatory factors are similar among themselves but different from those of other indices. This group includes the aggregate sector, the energy subsector, and the industrial metals subsector. Best models in the remaining indices (precious metals, agriculture, livestock) tend to contain determinants different from this group and also different from each other. In this (first) group, the CLI indicator appears particularly important for prediction, while information on the oil stock-to-use ratio does not seem to systematically improve forecasting. By contrast, the importance of the real effective exchange rate (REER), the world stock market index and the aggregate GSCI index (for the subsectors) depends on the forecast horizon and performance criterion used. For the best threshold models, the pattern is relatively similar to that for best linear models. For the aggregate sector, energy, and industrial metals, the CLI is an important predictor, the oil stock-to-use ratio is not particularly important, and the real effective exchange rate, the world stock market index, and the GSCI



		MSE								return							
		VOLA	CLI	CCI	INF	OIL	spread	COR	COR_oil	VOLA	CLI	CCI	INF	OIL	spread	COR	COR_oil
all	1m	6	4	5	3	7	1	2	8	4	1	7	8	3	2	5	6
	3m	4	3	6	2	5	1	7	8	4	2	3	8	7	1	6	5
	6m	5	2	6	1	4	3	8	7	5	2	1	7	8	3	6	4
	12m	5	1	6	2	4	3	8	7	3	1	7	2	5	4	8	6
energy	1m	4	2	7	6	3	1	4	8	1	3	6	8	5	4	2	7
	3m	3	6	5	1	4	2	7	8	3	5	2	8	6	4	1	7
	6m	4	2	3	1	5	6	8	7	3	4	1	5	7	2	8	6
	12m	5	1	6	2	3	4	7	8	5	4	8	1	6	3	2	7
industrial metals	1m	8	2	6	5	7	2	1	4	7	2	8	6	3	5	1	4
	3m	3	1	7	8	5	6	2	4	5	6	7	8	1	3	4	2
	6m	4	1	8	7	3	5	2	6	7	1	2	8	6	5	4	3
	12m	1	2	7	5	3	4	6	8	1	4	2	7	5	8	6	3
precious metals	1m	7	4	6	1	5	3	2	8	6	2	5	3	1	4	7	8
	3m	3	5	7	2	4	6	1	8	5	2	4	8	1	6	3	7
	6m	2	4	6	3	5	8	1	7	4	1	3	5	6	8	2	7
	12m	1	4	5	6	3	8	2	7	6	4	7	2	5	8	1	3
agriculture	1m	5	2	4	3	7	1	6	8	6	7	1	3	8	2	4	5
	3m	7	5	2	4	6	3	1	8	8	7	4	2	5	6	1	3
	6m	5	7	1	6	4	3	2	8	4	8	2	7	6	5	3	1
	12m	5	6	1	8	7	4	2	3	5	8	1	4	6	7	2	3
livestock	1m	6	3	7	4	2	8	5	1	8	2	3	5	1	7	4	6
	3m	2	3	5	8	4	6	1	7	6	7	3	5	2	8	1	4
	6m	5	3	8	7	2	4	1	6	8	6	5	2	4	7	1	3
	12m	5	1	7	8	3	6	2	4	6	3	7	5	1	8	2	4

FIGURE 4 Best threshold variables according to mean squared error (MSE) and return. *Note*: The graph indicates which threshold variables yield the best (1), second best (2), ... , to the worst (8) performance according to MSE and return.

aggregate index (for sub-sectors) are sometimes included in the best predictive specifications but not systematically so. Figure 5 shows how often a given explanatory variable is included in the best threshold model, considering the total of nine best threshold models (one for each threshold variable under consideration: commodity price, VOLA, CLI, CCI, INF, Δoil, spread, COR, COR-oil). For the precious metals sector, the most important variable appears to be the REER, while for the sectors agriculture and livestock, the most important variable is the world stock market index, followed by the CLI. These results emphasize the need to assess sectoral dynamics differently in commodity markets in order to optimize the predictive power of multivariate time series models.

5.3 | Threshold models and performance criteria

In a next step, we investigate patterns concerning the performance of threshold models across predictive criteria. In some situations, loss measures (MAE, MSE) and profit-based measures (DA, DV, return) behave differently when comparing predictive accuracy between regimes. For instance, threshold models with stock market volatility as the threshold variable perform systematically better in times of low volatility than in times of high volatility in terms of loss measures (MAE, MSE), while they perform better in times of high volatility than in times of low volatility in terms of profit-based measures (DA, DV, return). Table 5 presents the forecasting results of the aggregate GSCI with the threshold variable being the US stock market volatility. In the table, shading indicates better performance across the two regimes implied by the threshold model. The results suggest that, for all forecast horizons, commodity prices can be forecast more accurately in times of low volatility than in times of high volatility, but DA, DV, and the returns of a simple trading strategy (i.e., all profit measures) are higher in times of high volatility. While the first observation can probably be explained through lower price variability and thus better forecasting ability in times of low uncertainty, the second observation may be related to the chances of making more profits in large volatility markets when the direction of price change is forecast correctly.

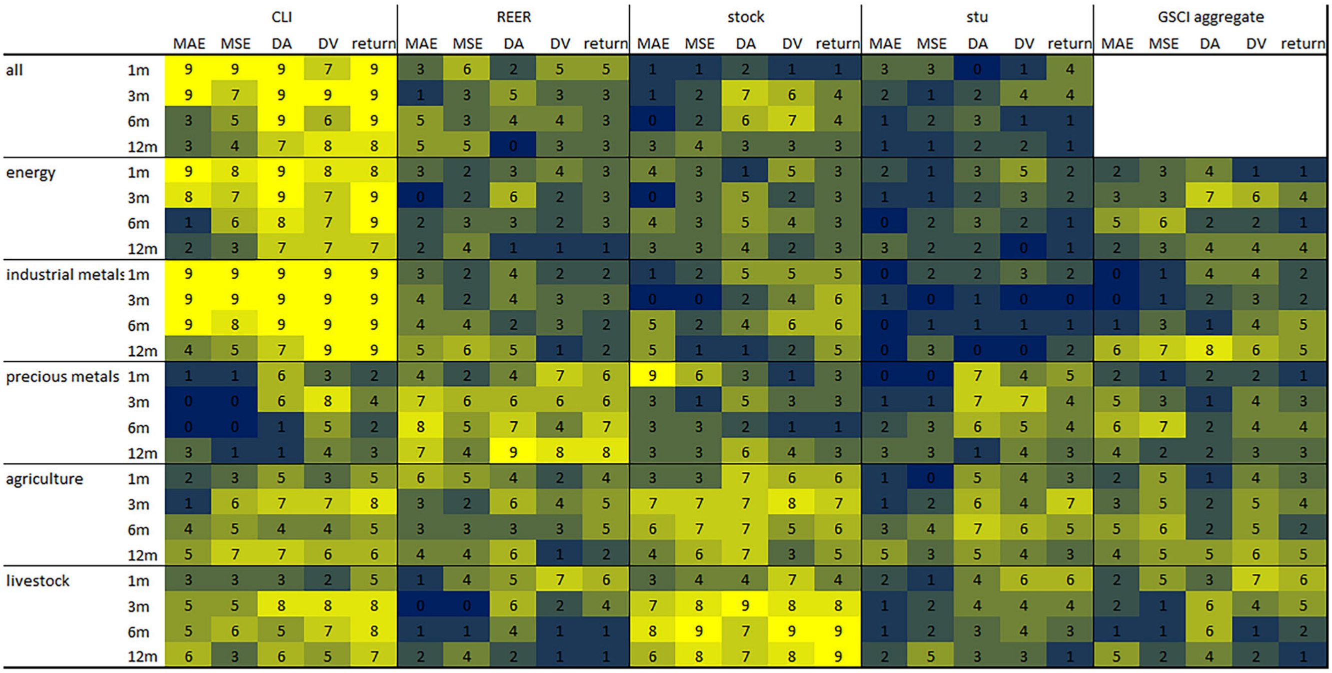
TABLE 4 Deviation of average performance of threshold models from performance of the best threshold model divided by deviation of best linear model from best threshold model.

		MAE	MSE	DA	DV	return
1-month horizon	All	0.48	0.51	1.12	0.80	1.43
	Energy	0.40	0.54	0.42	0.70	0.59
	Industrial metals	0.58	0.40	0.66	0.45	0.47
	Precious metals	0.55	0.71	0.52	0.47	0.45
	Agriculture	0.79	0.89	0.49	1.54	1.45
	Livestock	4.07	1.04	0.55	0.71	0.70
3-month horizon	All	0.58	0.69	0.45	0.44	0.61
	Energy	0.39	0.50	0.51	0.29	0.50
	Industrial metals	0.98	0.54	0.60	0.59	0.59
	Precious metals	0.78	0.65	0.50	0.36	0.44
	Agriculture	0.65	0.49	0.33	0.40	0.34
	Livestock	0.82	0.81	0.46	0.52	0.45
6-month horizon	All	0.98	0.82	1.25	0.54	0.69
	Energy	0.90	0.67	0.65	0.41	0.59
	Industrial metals	0.96	0.78	0.59	0.43	0.64
	Precious metals	0.54	0.43	0.46	0.49	0.63
	Agriculture	0.80	0.63	0.54	0.45	0.42
	Livestock	0.94	0.99	0.72	0.36	0.50
12-month horizon	All	1.15	0.89	0.35	0.50	0.54
	Energy	0.94	0.90	0.36	0.31	0.27
	Industrial metals	0.72	0.47	1.33	1.01	0.99
	Precious metals	0.51	0.55	0.58	0.41	0.79
	Agriculture	0.55	0.51	0.47	0.39	0.53
	Livestock	0.77	0.91	0.50	0.55	0.64

Note: Each figure is calculated as the average deviation of performance of a threshold model (across different threshold variables) from the performance of the best threshold model divided by the deviation of the best linear model from the best threshold model. Deviations are taken in absolute values, so the numbers are always positive. Note that the best threshold model is always better than the best linear model; the average threshold model, however, may be worse than the best linear model (implied by a ratio larger than one). The smaller the ratio, the better the average threshold model compared with the best linear model. Light petrol shading indicates smallest deviation of average threshold model compared with best linear model; red shading indicates largest deviation, across commodity sectors.

Abbreviations: DA, directional accuracy; DV, directional value; MAE, mean absolute error; MSE, mean squared error.

An analysis of the forecast errors over the period confirms that loss and profit measures tend to be positively correlated over the out-of-sample period for threshold models, with high forecast errors occurring in times when the direction of change was nevertheless correctly predicted. Such a behavior can be observed by comparing the MSE with profit measures over the out-of-sample period. Figure 6 shows the two-year rolling average of MSEs, returns, and DVs measured over the out-of-sample period for the threshold model using the stock market volatility as a threshold variable, when forecasting aggregate GSCI one month ahead. These results indicate that the financial instability in the aftermath of the financial crisis of 2008, which led to large increases in commodity prices, caused large prediction errors in terms of MSEs. However, threshold models based on stock market volatility (and other threshold variables) were able to predict direction of change in such times of high uncertainty and large price changes systematically better than their linear counterparts. The same phenomenon can be observed (albeit in smaller magnitude) for the generalized drop in commodity prices that started in 2015. The correlation between the 2-year rolling-averaged MSE and the return is 0.88, and for DV, it is 0.77. These results give evidence that threshold models, if specified efficiently, show a high degree of flexibility in adapting to structural changes in the dynamics of commodity prices and are able to achieve systematic gains in predictive ability for



		CLI					REER					stock					stu					GSCI aggregate				
		MAE	MSE	DA	DV	return	MAE	MSE	DA	DV	return	MAE	MSE	DA	DV	return	MAE	MSE	DA	DV	return	MAE	MSE	DA	DV	return
all	1m	9	9	9	7	9	3	6	2	5	5	1	1	2	1	1	3	3	0	1	4					
	3m	9	7	9	9	9	1	3	5	3	3	1	2	7	6	4	2	1	2	4	4					
	6m	3	5	9	6	9	5	3	4	4	3	0	2	6	7	4	1	2	3	1	1					
	12m	3	4	7	8	8	5	5	0	3	3	3	4	3	3	3	1	1	2	2	1					
energy	1m	9	8	9	8	8	3	2	3	4	3	4	3	1	5	3	2	1	3	5	2	2	3	4	1	1
	3m	8	7	9	7	9	0	2	6	2	3	0	3	5	2	3	1	1	2	3	2	3	3	7	6	4
	6m	1	6	8	7	9	2	3	3	2	3	4	3	5	4	3	0	2	3	2	1	5	6	2	2	1
	12m	2	3	7	7	7	2	4	1	1	1	3	3	4	2	3	3	2	2	0	1	2	3	4	4	4
industrial metals	1m	9	9	9	9	9	3	2	4	2	2	1	2	5	5	5	0	2	2	3	2	0	1	4	4	2
	3m	9	9	9	9	9	4	2	4	3	3	0	0	2	4	6	1	0	1	0	0	0	1	2	3	2
	6m	9	8	9	9	9	4	4	2	3	2	5	2	4	6	6	0	1	1	1	1	1	3	1	4	5
	12m	4	5	7	9	9	5	6	5	1	2	5	1	1	2	5	0	3	0	0	2	6	7	8	6	5
precious metals	1m	1	1	6	3	2	4	2	4	7	6	9	6	3	1	3	0	0	7	4	5	2	1	2	2	1
	3m	0	0	6	8	4	7	6	6	6	6	3	1	5	3	3	1	1	7	7	4	5	3	1	4	3
	6m	0	0	1	5	2	8	5	7	4	7	3	3	2	1	1	2	3	6	5	4	6	7	2	4	4
	12m	3	1	1	4	3	7	4	9	8	8	3	3	6	4	3	3	3	1	4	3	4	2	2	3	3
agriculture	1m	2	3	5	3	5	6	5	4	2	4	3	3	7	6	6	1	0	5	4	3	2	5	1	4	3
	3m	1	6	7	7	8	3	2	6	4	5	7	7	7	8	7	1	2	6	4	7	3	5	2	5	4
	6m	4	5	4	4	5	3	3	3	3	5	6	7	7	5	6	3	4	7	6	5	5	6	2	5	2
	12m	5	7	7	6	6	4	4	6	1	2	4	6	7	3	5	5	3	5	4	3	4	5	5	6	5
livestock	1m	3	3	3	2	5	1	4	5	7	6	3	4	4	7	4	2	1	4	6	6	2	5	3	7	6
	3m	5	5	8	8	8	0	0	6	2	4	7	8	9	8	8	1	2	4	4	4	2	1	6	4	5
	6m	5	6	5	7	8	1	1	4	1	1	8	9	7	9	9	1	2	3	4	3	1	1	6	1	2
	12m	6	3	6	5	7	2	4	2	1	1	6	8	7	8	9	2	5	3	3	1	5	2	4	2	1

FIGURE 5 Inclusion of explanatory variables in best threshold model. *Note*: The graph shows the number of times a given explanatory variable (CLI, REER, stock, stu, GSCI aggregate) is included in the best threshold model (aggregated over the nine different threshold variables). The maximum number possible is nine.

directional change. In this respect, our results add to a growing literature comparing statistical and economic approaches to measure predictive loss and profit and that find conflicting evidence of predictive power depending on the measure employed (see Dal Pra et al., 2018, for instance).

5.4 | Threshold models across sectors

Figure 7 presents MSEs and returns of the set of best threshold models for the different commodity sectors. Prices of livestock, precious metals, and agricultural commodities can be predicted comparatively well compared to the rest of the sectors and the aggregate index, while they tend to lead to low returns. On the other hand, prices of energy and industrial metals lead to the highest prediction errors but yield the largest returns.[13] This observation holds over all forecast horizons and can be explained due to the fact that larger deviations of the forecasts from its realizations are needed in order to increase the implied profit. To a lower extent, this pattern persists also for the other profit-based measures. DA and directional deviation appear higher for commodity sectors which are harder to predict in terms of MSE. An overview of all performance measures across all sectors and forecast horizons is presented in Figure 8.

Considering best threshold models, both loss measures, MAE and MSE, and the return display a clear structure relating to the forecast horizon. The loss measures increase, that is, forecast accuracy decreases, with an increasing forecast horizon. For example, the MSE when forecasting aggregate commodity prices increases from 0.28% when forecasting 1 month ahead to 6.83% when forecasting 12 months ahead. Using the return as a predictive ability measure, we observe the best performance for the shortest forecast horizon, with decreasing performance for increasing forecast horizons. While the return implied by a simple trading strategy for the aggregate commodity index is 31.46% when forecasting 1 month ahead, the corresponding return is only 12.41% when forecasting 12 months ahead.[14] The observed patterns (for MAE, MSE and return) with respect to the forecast horizon hold for all commodity sectors (Table 6). For the other two profit-based measures (DA, DV), the behavior with respect to the forecast horizon is not similar across sectors. While for precious metals and agriculture, the DA and DV grow with increasing forecast horizons, the picture is mixed for energy, industrial metals, livestock and for the aggregate sector. In most cases, however, the DA and value statistics are largest when forecasting twelve months ahead. See Table 7 and Figure 8.

Table 6 indicates that the commodity sector whose returns dominate those of the others is most of

TABLE 5 Performance of best threshold model (threshold variable = volatility) and of corresponding linear model for the aggregate index.

		MAE	MSE	DA	DV	return
1-month horizon		TDVAR(1,2)	TDVAR(1,4)	TDVAR(3,4)	TVAR(3,4)	TDVAR(3,4)
		1001	1100	1100	1100	1100
Threshold	Total	4.10	0.29	69.05	75.86*	28.63
	Regime 1	3.87	0.26	68.49	75.73*	25.83
	Regime 2	4.57	0.50	72.73	76.49	48.66
Linear	Total	4.23	0.30	64.88	67.60	24.44
	Regime 1	3.82	0.29	65.07	65.26	22.59
	Regime 2	5.07	0.42*	63.64	78.52	37.40
3-month horizon		TDVAR(2,4)	TDVAR(2,9)	TDVAR(3,12)	TDVAR(1,4)	TDVAR(1,4)
		1000	1011	1000	1010	1010
Threshold	Total	8.88	1.55	66.67	75.20*	19.05**
	Regime 1	8.63	1.23	65.52	71.83*	15.63**
	Regime 2	10.64	3.74	73.91*	91.35	45.15
Linear	Total	9.15	1.85	63.10	65.01	13.33
	Regime 1	8.88	1.32	64.14	59.17	8.94
	Regime 2	11.00	5.54	56.52	92.99	47.90
6-month horizon		TDVAR(3,10)	TDVAR(2,9)	TDVAR(3,4)	TDVAR(2,11)	TDVAR(2,4)
		1100	1011	1000	1110	1100
Threshold	Total	14.23	4.43	67.86	74.11	11.27
	Regime 1	13.11	3.82	65.99	68.15	6.51
	Regime 2	19.17	8.68	80.95	96.80	45.51
Linear	Total	14.46	4.92	66.67	64.02	9.67
	Regime 1	13.09	4.22	65.99	66.70	7.37
	Regime 2	20.54	9.83	71.43	53.80	25.53
12-month horizon		TDVAR(3,6)	TVAR(3,11)	TVAR(2,4)	TVAR(2,4)	TVAR(2,4)
		1000	0101	1010	1010	1010
Threshold	Total	21.14	8.36	69.64**	75.42	9.34
	Regime 1	18.10	7.47	67.35**	68.84	6.95
	Regime 2	39.35	11.41*	85.71	88.44	26.03
Linear	Total	21.65	12.70	54.76	58.38	3.77
	Regime 1	18.08	8.46	52.38	54.45	2.51
	Regime 2	43.06	27.20	71.43	66.15	12.58

Note. * (**/***) Indicates rejection of the null hypothesis of equal forecasting accuracy between the best threshold model and the corresponding linear model at 10% (5%/1%). The four-digit combination of ones and zeros below the model shows the inclusion (1) or the exclusion (0) of the explanatory variables CLI, REER, stock market index, and oil stock-to-use ratio. Petrol shading indicates that the best threshold model outperforms the best linear model. Light petrol shading shows better total performance between best threshold model and corresponding linear model. Grey shading indicates better performance between the two regimes for the best threshold model. Regime 1 is defined by $\mathrm{VOLA}_{t-k} \leq \gamma$, while regime 2 is defined by $\mathrm{VOLA}_{t-k} > \gamma$.
Abbreviations: DA, directional accuracy; DV, directional value; MAE, mean absolute error; MSE, mean squared error.

the time the industrial metals sector. Exceptions are the energy sector for return and DV for $h = 1$ and the sector of agriculture for DA and DV for $h = 12$. The sector with the best loss-based performance is livestock. The smallest loss-based performance occurs for livestock in case of one month forecast horizon, namely, 2.42% for MAE and 0.1% for MSE, and the largest profit-based performance occurs for agriculture for 12 months forecast horizon, namely, 80.95% for DA and 89.41% for DV and for energy sector where the return of 46.74% occurs in the case of 1 month forecast horizon.

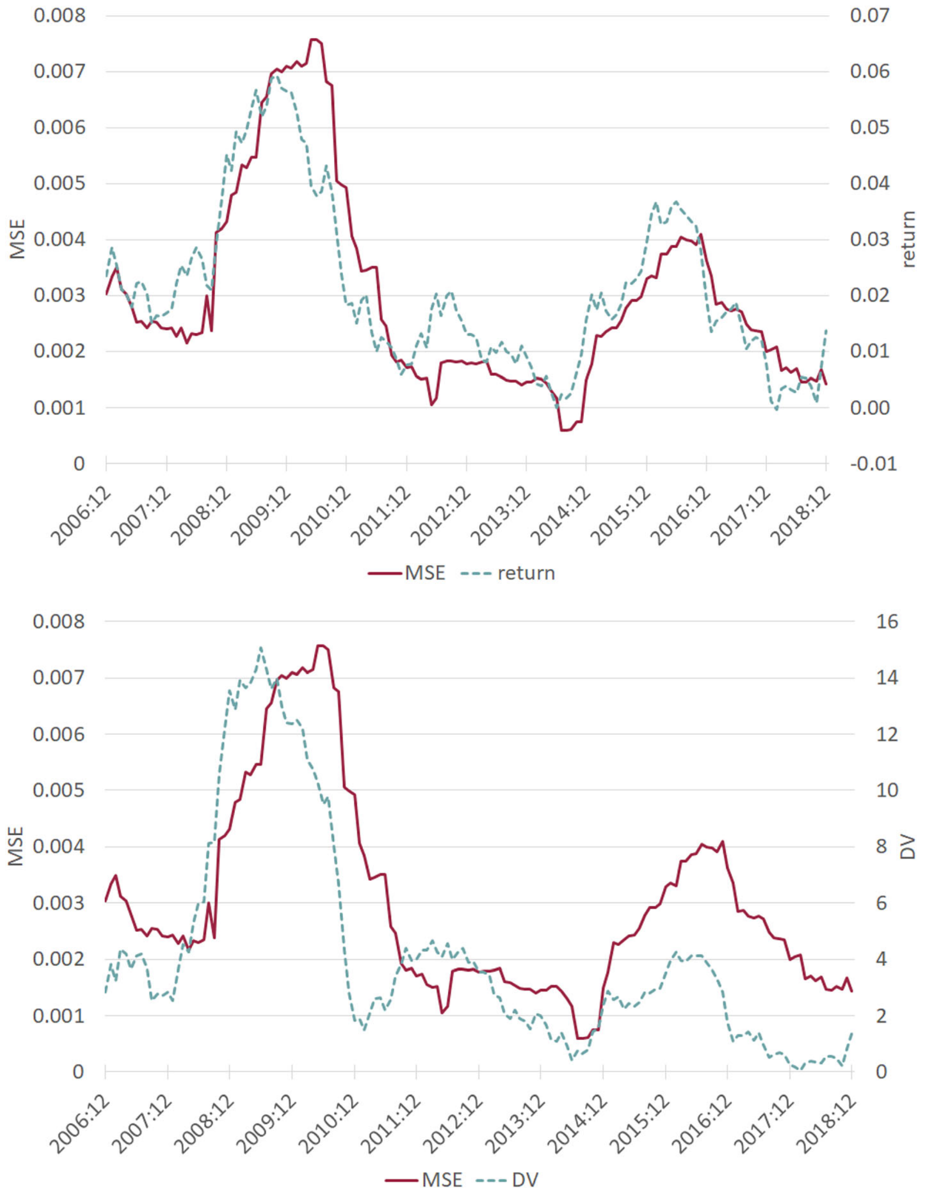


FIGURE 6 Mean-squared error, return and directional value, 2-year rolling average (threshold model with stock market volatility as the threshold variable, aggregate GSCI, 1-month forecast horizon)

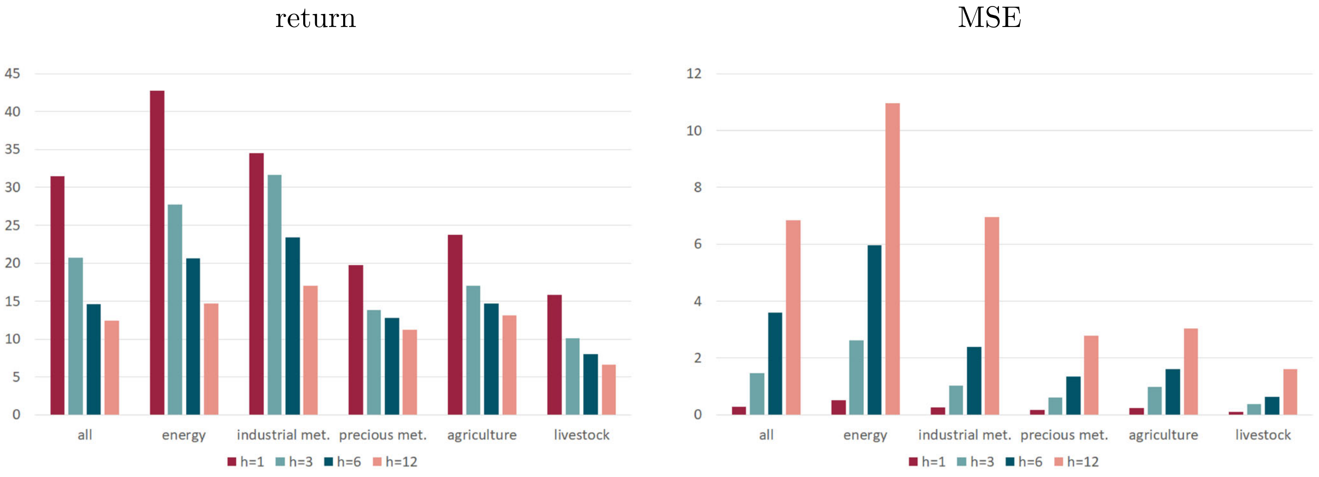


FIGURE 7 Returns and mean squared error (MSE) of best threshold models for different GSCI sectors. *Note*: The graph shows the returns (left) and MSE (right) of best threshold models for different GSCI sectors and different forecast horizons.

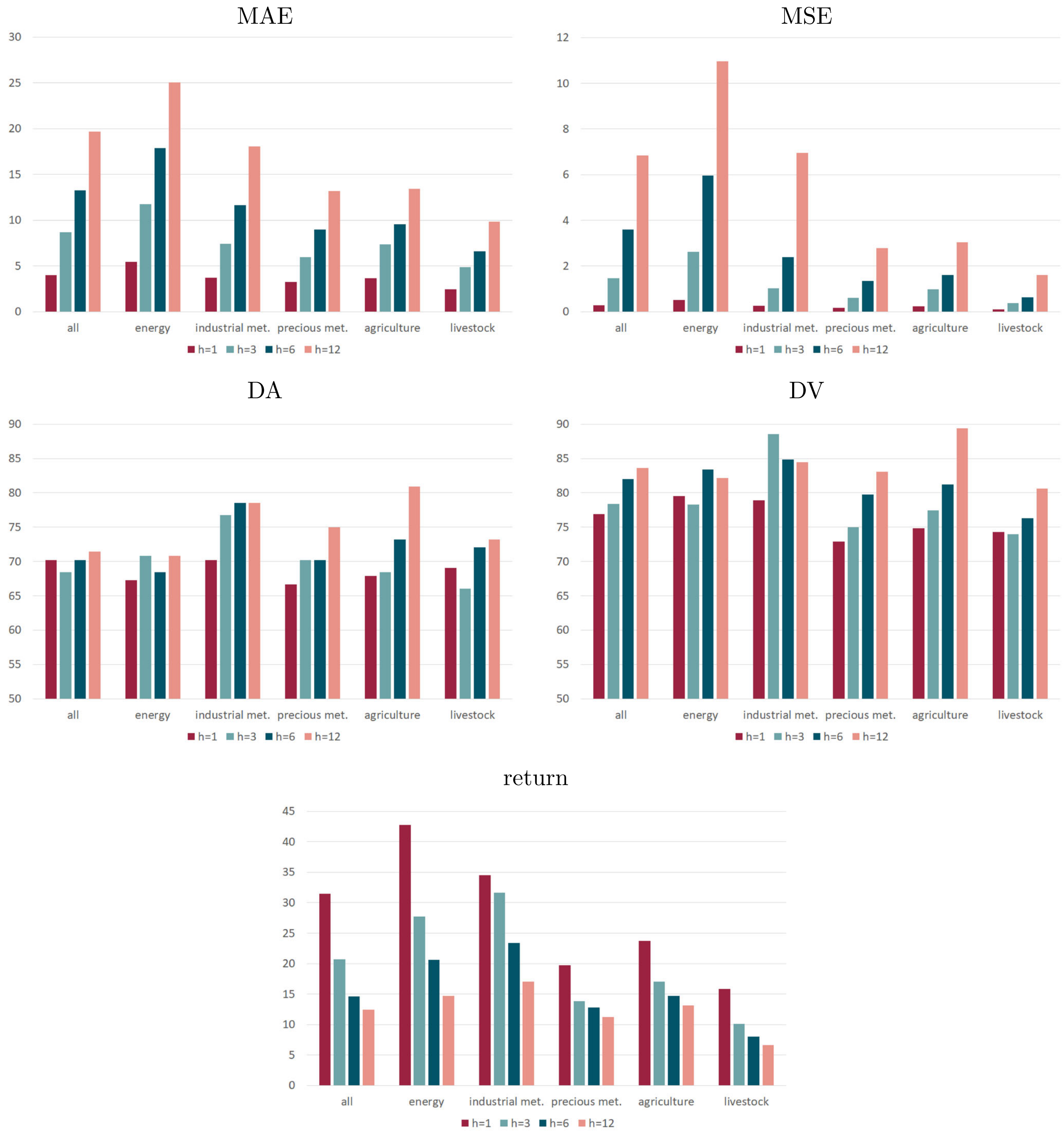


FIGURE 8 Loss and profit measures for different S&P Coldman Sachs commodity index (GSCI) sectors. *Note*: The graphs show MAE, MSE, DA, DV and return for different GSCI sectors and different forecast horizons.

5.5 | Threshold models and larger class of models

In addition to standard linear vector autoregressive models, we also consider a much larger class of models in order to find out whether threshold models also outperform other specifications. This class includes different univariate GARCH models, vector error correction models and Bayesian VAR models (see Table 1). For this larger class of models, we choose the lag structure based on in-sample model selection based on optimizing the Akaike information criterion.[15] We also use an additional performance measure, namely, the proportion of correctly forecast TPs.[16]

Our results show that threshold models have the best predictive performance in the vast majority of cases (see

TABLE 6 Performance of best threshold model across GSCI sectors.

		MAE	MSE	DA	DV	Return
$h=1$	Best TM	2.45	0.10	70.24	79.53	42.76
	TV	dep	dep	CLI	VOLA	VOLA
	Sector	Livestock	Livestock	Aggregate	Energy	Energy
$h=3$	Best TM	4.84	0.38	76.79	88.57	31.67
	TV	COR	COR	Δoil	Δoil	Δoil
	Sector	Livestock	Livestock	Industrial met.	Industrial met.	Industrial met.
$h=6$	Best TM	6.58	0.64	78.57	84.83	23.43
	TV	COR	COR	CLI	COR	dep
	Sector	Livestock	Livestock	Industrial met.	Industrial met.	Industrial met.
$h=12$	Best TM	9.84	1.61	80.95	89.41	17.01
	TV	CLI	CLI	CCI	CCI	VOLA
	Sector	Livestock	Livestock	Agriculture	Agriculture	Industrial met.

Abbreviations: DA, directional accuracy; DV, directional value; MAE, mean absolute error; MSE, mean squared error.

TABLE 7 Performance of best threshold models for different GSCI sectors.

		MAE	MSE	DA	DV	Return
1-month horizon	All	4.00	0.28	70.24	76.90	31.46
	Energy	5.45	0.50	67.26	79.53	42.76
	Industrial metals	3.72	0.26	70.24	78.89	34.47
	Precious metals	3.23	0.17	66.67	72.90	19.75
	Agriculture	3.64	0.23	67.86	74.82	23.75
	Livestock	2.45	0.10	69.05	74.27	15.79
3-month horizon	All	8.70	1.46	68.45	78.34	20.70
	Energy	11.77	2.63	70.83	78.31	27.71
	Industrial metals	7.42	1.03	76.79	88.57	31.67
	Precious metals	5.93	0.61	70.24	75.00	13.82
	Agriculture	7.36	0.97	68.45	77.47	17.07
	Livestock	4.84	0.38	66.07	74.01	10.10
6-month horizon	All	13.27	3.59	70.24	82.01	14.58
	Energy	17.90	5.97	68.45	83.41	20.58
	Industrial metals	11.65	2.40	78.57	84.83	23.43
	Precious metals	8.94	1.35	70.24	79.80	12.83
	Agriculture	9.56	1.60	73.21	81.26	14.66
	Livestock	6.58	0.64	72.02	76.31	7.99
12-month horizon	All	19.64	6.83	71.43	83.62	12.41
	Energy	25.02	10.97	70.83	82.14	14.67
	Industrial metals	18.05	6.95	78.57	84.50	17.01
	Precious metals	13.18	2.79	75.00	83.11	11.19
	Agriculture	13.41	3.04	80.95	89.41	13.12
	Livestock	9.84	1.61	73.21	80.60	6.64

Notes. The table shows the forecast performance of best threshold models for different GSCI sectors and different forecast horizons. Light petrol shading indicates best performance across GSCI sectors, red shading indicates worst performance.

Abbreviations: DA, directional accuracy; DV, directional value; MAE, mean absolute error; MSE, mean squared error.

TABLE 8 Best models in smaller and larger class of models.

		Smaller class of models					Larger class of models					
		MAE	MSE	DA	DV	return	MAE	MSE	DA	DV	return	TP
$h=1$	all	4.00	0.28	70.24	76.90	31.46	4.07	0.28	66.67	74.52	27.44	20.25
	energy	5.45	0.50	67.26	79.53	42.76	5.64	0.55	64.88	75.08	37.62	24.38
	industrial met.	3.72	0.26	70.24	78.89	34.47	3.81	0.27	69.05	77.85	32.23	21.48
	precious met.	3.23	0.17	66.67	72.90	19.75	3.38	0.18	60.71	66.96	14.59	33.74
	agriculture	3.64	0.23	67.86	74.82	23.75	3.70	0.25	63.69	70.28	18.70	19.38
	livestock	2.45	0.10	69.05	74.27	15.79	2.50	0.11	65.48	73.06	14.68	31.51
$h=3$	all	8.70	1.46	68.45	78.34	20.70	8.70	1.46	66.07	78.34	20.70	20.48
	energy	11.77	2.63	70.83	78.31	27.71	12.03	2.73	67.26	75.30	24.69	25.84
	industrial met.	7.42	1.03	76.79	88.57	31.67	7.76	1.20	72.02	80.89	25.13	28.38
	precious met.	5.93	0.61	70.24	75.00	13.82	6.39	0.65	64.88	69.48	10.82	23.47
	agriculture	7.36	0.97	68.45	77.47	17.07	7.77	1.06	64.29	72.28	14.29	22.62
	livestock	4.84	0.38	66.07	74.01	10.10	5.30	0.46	63.69	70.07	8.29	29.35
$h=6$	all	13.27	3.59	70.24	82.01	14.58	13.99	4.27	67.26	76.63	12.18	20.97
	energy	17.90	5.97	68.45	83.41	20.58	18.33	7.59	67.86	76.91	16.57	18.42
	industrial met.	11.65	2.40	78.57	84.83	23.43	12.67	3.25	73.81	80.16	19.91	25.58
	precious met.	8.94	1.35	70.24	79.80	12.83	9.37	1.45	65.48	73.13	10.37	30.16
	agriculture	9.56	1.60	73.21	81.26	14.66	10.88	2.08	67.86	75.56	12.36	28.07
	livestock	6.58	0.64	72.02	76.31	7.99	7.46	0.88	63.69	68.86	5.76	23.33
$h=12$	all	19.64	6.83	71.43	83.62	12.41	20.44	8.42	60.71	73.30	6.75	35.00
	energy	25.02	10.97	70.83	82.14	14.67	26.92	13.09	65.48	72.75	11.75	36.36
	industrial met.	18.05	6.95	78.57	84.50	17.01	21.82	7.97	70.24	76.81	13.39	39.39
	precious met.	13.18	2.79	75.00	83.11	11.19	14.57	3.09	63.69	75.50	9.57	16.28
	agriculture	13.41	3.04	80.95	89.41	13.12	15.58	3.87	75.00	79.12	9.71	29.55
	livestock	9.84	1.61	73.21	80.60	6.64	10.54	1.79	66.07	72.78	5.17	26.32

Notes. The table shows the performance criteria of best models in "Smaller class of models" and of best models in "Larger class of models" for different GSCI sectors and different forecast horizons. The best model in the smaller class of models (left panel) is always better than, or at least as good as, the best model in the larger class of models (right panel). In the smaller class of models best models are always threshold models, in the larger class of models, in 25 out of the total of 144 cases the best model is not a threshold model. Light petrol shading indicates the cases when the best model is not a threshold model.

Table 8). In only 25 out of a total of 144 cases (six commodity sectors, six performance measures and four forecast horizons), the threshold model is outperformed by a different specification.

None of the best models in this expanded specification set can keep up with the best prediction models found in the smaller set used before. In all cases without any exception, the best model determined in our previous analysis, which is always a threshold model, outperforms the best model found now, including the cases when the best model now is not a threshold model (see Table 8).[17] As best threshold models for different threshold variables do often perform similarly (well), as found in our previous analysis, not only the best threshold model but often also other threshold models (with different threshold variables) outperform the corresponding best model found now.

The performance of best models with respect to both loss measures and the return show a clear pattern with respect to the forecast horizon: Forecast accuracy decreases with an increasing forecast horizon and so does the return. The proportion of correctly forecast TPs, which was not analysed before, does not show a uniform pattern with respect to the forecast horizon. However, it is largest for the 12-month forecast horizon for the total commodity index, for energy, and industrial metals, while it is largest for the 1-month forecast horizon for the remaining sectors (precious metals, agriculture, livestock). When forecasting 12 months ahead, the overall index and energy are actually among the best (ranking third and second) according to TP (see Table 8).

The vast majority (all but one) of the best performing threshold models with respect to correctly forecasting

TPs for the aggregate index and the energy sector rely on a threshold variable that is connected to oil (Δoil or COR-oil). All models for the aggregate index and for energy, except for one case, contain the oil stock-to-use ratio as a determinant. Best models for precious metals according to TP are either based on a threshold variable related to oil or have the oil stock-to-use ratio among the explanatory variables. The same holds for industrial metals and livestock. For all indices, best models according to TP rely on an oil related threshold variable for a forecast horizon of 12 months. For all indices (but agriculture), the REER is included in the best model (according to TP) for a 12-month forecast horizon.[18]

6 | CONCLUSIONS

In this paper, we present overwhelming evidence that allowing for regime-dependent dynamics in models for commodity prices leads to improvements in predictive ability. This follows from the fact that the characteristics of the dynamics of commodity prices and their interactions with other variables are not constant over time but differ depending on particular phenomena (e.g., periods of high and low volatility in the equity markets, good and bad economic times or the level of inflation). If these regimes can be properly defined out of the data, the stability of dynamics and interactions within particular regimes allow for better predictions. However, the nature of these improvements also differs across predictive measures and sectors.

We assess the quality of commodity forecasts with a variety of different performance measures. In addition to the MSE, the traditional forecast performance measure used in many studies, we also consider measures that evaluate DA, DV, the ability to predict TPs, and the returns implied by a simple trading strategy based on commodity price forecasts. These additional profit-based measures do not directly assess forecast accuracy but relate to other dimensions of forecasting quality and may be more relevant for particular applications in policy and applied work. We create an econometric modeling framework to predict commodity price dynamics as captured by the changes in an overall commodity price index, the S&P Goldman Sachs Commodity Index, as well as in five sub-indices (energy, industrial metals, precious metals, agriculture, livestock). We consider short-term and long-term forecast horizons (ranging from one month to twelve months) and use monthly observations in the period 1980–2018. Our forecast models include threshold models that are based on different threshold variables.

We provide a rich set of empirical results. In addition to the forecast performance comparison of threshold and linear models we investigate the threshold variables and explanatory variables that imply "best" models, the structural pattern of evaluation criteria across different regimes, and best sector-specific forecast performance. We observe that threshold models with volatility in equity markets defining the states of the economy seem to perform better in times of low volatility than in times of high volatility with respect to loss measures, while, on the other hand, they seem to perform better in times of high volatility than in times of low volatility with respect to profit-based measures. Our results suggest that an interesting trade-off appears between loss and profit measures, which implies that the particular aim of the prediction exercise carried out plays a very important role in terms of defining which set of models is the best to use. The optimal specifications for applications where the metrics for success are related to systematically predicting the direction of change of commodity prices accurately may thus be systematically different from those aimed at providing point predictions with an absolute minimal distance to the realized values. In addition, the positive results found in the paper for threshold models (as compared to part of the literature) are also related to the fact that we exploit a large specification space as compared with other studies, both in terms of potential covariates and threshold variables.

The importance of the oil market as a determinant of commodity price dynamics is reflected in the results of our analysis, with oil related variables appearing in the best forecasting models for TPs (either as a covariate or a threshold variable) in the aggregate GSCI, energy, and precious metals models. This result indicates that particular oil price dynamics may act as a leading indicator of changes in trends in commodity prices, and its inclusion in econometric specifications aimed at predicting TP probabilities may lead to significant improvements in forecasting ability.

Exploiting the potential for improving predictive ability in order to refine the specification and estimation of models may be a potentially fruitful avenue of future research. In particular, entertaining estimation methods that differ from least squares (and thus do not build on the minimization of in-sample squared errors) or Bayesian methods with suitable prior specifications could lead to further improvements in the prediction of commodity prices. Enlarging the set of possible models to account for nonlinearities to include smooth transition in the parameters appears also as a natural next step that builds upon the results presented in this study, as does the implementation of threshold dynamic factor models in the spirit of the specifications in Massacci (2017).

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support from the Austrian Science Fund FWF (project number P 30915-G27) and helpful comments by an anonymous referee and Robert Kunst that improved the manuscript considerably

CONFLICT OF INTEREST STATEMENT

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Jesus Crespo Cuaresma https://orcid.org/0000-0003-3244-6560
Ines Fortin https://orcid.org/0000-0003-4517-455X

ENDNOTES

[1] Type of stock-to-use ratio depends on the class of commodity index that is forecast.

[2] See Table A3 in Appendix A, where we present a summary of such a testing exercise.

[3] The type of stock-to-use ratio (namely for oil, wheat and meat) depends on the specification of the commodity class under consideration.

[4] In principle, the threshold estimation could have been based on optimizing an overall model selection criterion of the full system of equations given by the multivariate model, but such an estimation procedure could result in identifying threshold dynamics related to other endogenous variables of the vector autoregressive model, thus complicating the identification of the source of potential improvements in predictive ability of commodity prices using such specifications.

[5] The S&P GSCI total return indices reflect the performance of a total return investment in commodities, that is, the contract daily return plus the daily interest on the funds hypothetically committed to the investment.

[6] In alternative modelling exercises, we also included the industrial production index as an additional variable but removed it from the list of variables as it is heavily correlated with the CLI and did not help to improve the forecast performance substantially.

[7] As a robustness check, we also performed the analysis over the out-of-sample period January 2001 to December 2018 and obtained similar results.

[8] Note, however, that it is difficult to compute a reliable regime-specific value of this measure. Three consecutive time points are needed to calculate the TP, there are usually not that many TPs in general, and there tend to be even less in each regime. It may easily happen that the three consecutive time points required to calculate the measure are not in the same regime. Therefore, we do not use TPs in the comparison of threshold and linear models, where the analysis of regime-based performance is essential. However, we use this performance measure when analyzing overall performance differences.

[9] Notice that while the "buy low, sell high" trading strategy is not a feasible trading strategy for physical commodities, as it would require calculating spot returns net of the cost of carry such as storage costs and insurance, it may well be implemented for investable indices like the GSCI indices. See Miffre (2016) for an overview on strategies in commodity markets.

[10] The best linear models with respect to a certain performance measure were chosen according to all possible combinations of lag lengths as well as all combinations of explanatory variables (in case of multivariate models) such that the performance measure under consideration is maximized.

[11] A similar conclusion follows from Table 5, where we present the results for the aggregate GSCI with the threshold variable being volatility of the US stock market. These results are discussed below in more detail when analysing a different aspect of the forecasting exercise.

[12] The use of the Diebold-Mariano test in the case of nested models is known to be modestly conservative (see the Monte Carlo evidence in Clark & McCracken, 2013), that is, to have size slightly below nominal size. Since the tests of interest in our exercise have the linear model as the null hypothesis, the potential bias plays against the threshold model and thus provides particularly credible evidence for the nonlinear models if we observe a rejection.

[13] Note that the different forecasting accuracy across commodity sectors corresponds to different variability in returns, as suggested above. Commodity sectors with smaller variability are easier to predict than those with larger variability; see Table 2 and Figures 7 and 8.

[14] Notice that the return we report does not account for potential trading costs. The returns from actual trading strategies related to different forecast horizons which include trading costs may be different, and the current pattern with respect to the forecast horizon may not be preserved.

[15] Note that compared to the setting discussed before, where we choose specifications based on out-of-sample validation, the in-sample lag determination is in some sense more restrictive and may provide (slightly) inferior forecast performance of the linear and threshold models used before. This different model selection approach when considering this larger set of specifications is required due to the expensive computational price of carrying out out-of-sample validation in the larger model space employed here.

[16] As discussed before, see Section 4, this measure cannot be reliably computed for individual regimes in threshold models separately and thus has not been used in the previous analysis.

[17] This comparison does not include best models with respect to the proportion of correct TPs, as this measure was not used before.

[18] More detailed tables presenting the forecasting performance of best models when a larger class of models is included can be obtained from the authors upon request.

REFERENCES

Ahumada, H., & Cornejo, M. (2015). Explaining commodity prices by a cointegrated time series-cross section model. *Empirical Economics*, *48*, 1667–1690.

Ahumada, H., & Cornejo, M. (2016). Forecasting food prices: the case of corn, soybeans and wheat. *International Journal of Forecasting*, *32*, 838–848.

Bai, J., & Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, *18*, 1–22.

Bernard, J. T., Khalaf, L., Kichian, M., & McMahon, S. (2008). Forecasting commodity prices: GARCH, jumps, and mean reversion. *Journal of Forecasting*, *27*, 279–291.

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

Clark, T., & McCracken, M. (2013). Advances in forecast evaluation. *Handbook of Economic Forecasting*, *2*, 1107–1201.

Crespo Cuaresma, J., Hlouskova, J., & Obersteiner, M. (2021). Agricultural commodity price dynamics and their determinants: a comprehensive econometric approach. *Journal of Forecasting*, *40*, 1245–1273.

Dal Pra, G., Guidolin, M., Pedio, M., & Vasile, F. (2018). Regime shifts in excess stock return predictability: an out-of-sample portfolio analysis. *Journal of Portfolio Management*, *44*, 10–24.

Degiannakis, S., Filis, G., Klein, T., & Walther, T. (2020). *Forecasting realized volatility of agricultural commodities*, Vol. 38, pp. 74–96.

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, *133*, 253–263.

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

Gençay, R. (1998). Optimization of technical trading strategies and the profitability in security markets. *Economics Letters*, *59*, 249–254.

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

Gorton, G., & Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, *62*, 47–68.

Groen, J. J. J., & Pesenti, P. A. (2011). Commodity prices, commodity currencies, and global economic developments. In Ito, T., & Rose, A. (Eds.), *Commodity Prices and Markets, NBER East Asia Seminar on Economics*, Vol. 20: Chicago University Press, pp. 15–42.

Guidolin, M., & Ono, S. (2006). Are the dynamic linkages between the macroeconomy and asset prices time-varying? *Journal of Economics and Business*, *58*(5-6), 480–518.

Guidolin, M., & Pedio, M. (2021). Forecasting commodity futures returns with stepwise regressions: do commodity-specific factors help? *Annals of Operations Research*, *299*, 1317–1356.

Guidolin, M., & Timmermann, A. (2005). Economic implications of bull and bear regimes in UK stock and bond returns. *Economic Journal*, *115*, 111–143.

Guidolin, M., & Timmermann, A. (2009). Forecasts of US short-term interest rates: a flexible forecast combination approach. *Journal of Econometrics*, *150*, 297–319.

Hong, H., & Yogo, M. (2012). What does futures market interest tell us about the macroeconomy and asset prices? *Journal of Financial Economics*, *150*, 473–490.

Jacobsen, B., Marshall, B. R., & Visaltanachoti, N. (2019). Stock market predictability and industrial metal returns. *Management Science*, *65*, 3026–3042.

Just, R. E., & Rausser, G. C. (1981). Commodity price forecasting with large-scale econometric models and the futures market. *American Journal of Agricultural Economics*, *63*, 197–208.

Massacci, D. (2017). Least squares estimation of large dimensional threshold factor models. *Journal of Econometrics*, *197*, 101–129.

Miffre, J. (2016). Long-short commodity investing: a review of the literature. *Journal of Commodity Markets*, *1*, 3–13.

Ramirez, O. A., & Fadiga, M. (2003). Forecasting agricultural commodity prices with asymmetric-error GARCH models. *Journal of Agricultural and Resource Economics*, *28*, 71–85.

S&P Dow Jones (2019). S&P GSCI methodology. S&P Dow Jones indices: index methodology.

Xu, X. (2017). Short-run price forecast performance of individual and composite models for 496 corn cash markets. *Journal of Applied Statistics*, *44*, 2593–2620.

Xu, X. (2018). Using local information to improve short-run corn price forecasts. *Journal of Agricultural & Food Industrial Organization*, *16*, 1–15.

Xu, X. (2020). Corn cash price forecasting. *American Journal of Agricultural Economics*, *102*, 1297–1320.

AUTHOR BIOGRAPHIES

Jesus Crespo Cuaresma is a professor of economics at the Vienna university of Economics and Business and research scholar at the International Institute for Applied Systems Analysis, as well as a scientific consultant to the Austrian Institute of Economic Research. He has published extensively to topics related to applied econometrics, economic growth, and forecasting.

Ines Fortin is a senior researcher at the Institute for Advanced Studies (IHS), Vienna, and holds a PhD in Economics from the University of Vienna. She is part of the economic forecast group at IHS and her research interests include behavioral finance and economics and forecasting.

Jaroslava Hlouskova is a senior researcher at the Institute for Advanced Studies, Vienna, and holds a PhD in Mathematics from the Comenius University in Bratislava. Her research interests are in the fields of forecasting, behavioral economics and finance, and quantitative finance.

Michael Obersteiner is the director of the Environmental Change Institute, University of Oxford and a professor in Global Change and Sustainability. He joined the International Institute for Applied Systems Analysis (IIASA), where he was the director of the Ecosystems Services and Management (ESM) Program. His scientific interest stretches from global terrestrial ecosystems science to economics.

How to cite this article: Crespo Cuaresma, J., Fortin, I., Hlouskova, J., & Obersteiner, M. (2024). Regime-dependent commodity price dynamics: A predictive analysis. *Journal of Forecasting*, 1–26. https://doi.org/10.1002/for.3152

APPENDIX A: DATA DESCRIPTION

TABLE A1 Contracts included in the S&P GSCI in 2019 (RPDW = reference percentage dollar weight; see S&P Dow Jones, 2019).

Commodity	Trading facility	2019 RPDW	Sector
Chicago wheat	CBT	2.77%	Agriculture
Kansas wheat	KBT	1.15%	Aagriculture
Soybeans	CBT	3.14%	Agriculture
Coffee	ICE - US	0.72%	Agriculture
Sugar	ICE - US	1.54%	Agriculture
Cocoa	ICE - US	0.32%	Agriculture
Cotton	ICE - US	1.41%	Agriculture
Lean hogs	CME	1.91%	Agriculture
Live cattle	CME	3.48%	Agriculture
Feeder cattle	CME	1.27%	Agriculture
WTI crude oil	NYM / ICE	26.42%	Energy
Heating oil	NYM	4.45%	Energy
RBOB gasoline	NYM	4.48%	Energy
Brent crude oil	ICE - UK	18.61%	Energy
Gas oil	ICE - UK	5.56%	Energy
Natural gas	NYM / ICE	3.11%	Industrial metals
Aluminum	LME	3.89%	Industrial metals
Copper	LME	4.45%	Industrial metals
Nickel	LME	0.76%	Industrial metals
Lead	LME	0.78%	Industrial metals
Zinc	LME	1.28%	Industrial metals
Gold	CMX	3.72%	Precious metals
Silver	CMX	0.42%	Precious metals

TABLE A2 Data description and sources.

Abbreviation	Variable	Unit	Note	Source	Code	Start date	Frequency
Commodity							
GSCI	S&P GSCI	Index	Total return index	Ref DS: S&P	GSCITOT	1980:1	m
GSCI-energy	S&P GSCI Energy	Index	Total return index	Ref DS: S&P	GSENTOT	1982:12	m
GSCI-industrial	S&P GSCI Industrial Metals	Index	Total return index	Ref DS: S&P	GSINTOT	1980:1	m
GSCI-precious	S&P GSCI Precious Metals	Index	Total return index	Ref DS: S&P	GSPMTOT	1980:1	m
GSCI-agri	S&P GSCI Agriculture	Index	Total return index	Ref DS: S&P	GSAGTOT	1980:1	m
GSCI-live	S&P GSCI Livestock	Index	Total return index	Ref DS: S&P	GSLVTOT	1980:1	m
Explanatory variables: macro/finance							
CLI	US Composite Leading Indicator	Index	Amplitude adjusted, seasonally adjusted	Ref DS: OECD	USOL2000Q	1980:1	m
REER	US real effective exchange rate	Index		Ref DS: OECD	USOCC011	1980:1	m
stock	world stock market index	Index		Ref DS: DS	TOTMKWD	1980:1:1	d
Explanatory variables: fundamental							
oOl-stu	Oil stock-to-use ratio, total world	ratio	linear interpolation from annual	own calc., OPEC		1980:1	m (a)
Wheat-stu	US wheat stock-to-use ratio	%	linear interpolation from annual	USDA (FAS)		1980:1	a
Meat-stu	US meat stock-to-use ratio	%	lin interp from annual, meat: beef & veal	USDA (FAS)		1980:1	a
Threshold variables							
CLI	US Composite Leading Indicator	Index	Amplitude adjusted, seasonally adjusted	Ref DS: OECD	USOL2000Q	1980:1	m
CCI	US consumer confidence index	Index	Seasonally adjusted	Ref DS: Conference Board	USCNFCONQ	1980:1	m
VOLA	US Stock Market Volatility	%	SD of daily stock market ret. in one month, ann	own calc., Ref DS		1980:1	m
COR	Cor betw. US stock & bond markets, 6m	Cor	Correlation between stock and bond, 6m	own calc., Ref DS		1980:6	m
COR-oil	Cor betw. world stock & oil markets, 6m	Cor	Correlation between stock and oil, 6m	own calc., Ref DS		1980:6	m
Oil	Oil price (Brent)	USD/b	Crude Oil BFO M1 Europe FOB $/BBl, Brent	Ref DS: Ref	OILBREN	1980:1	m
INF	US inflation (consumer price index)	%	All urban sample: all items	Ref DS: BLS	USCPANNL	1980:1	m
Spread	diff. betw. long- and short-term US int. rates	pp	IR-long minus IR	own calc., Ref DS		1980:1	m
Auxiliary variables							

(Continues)

TABLE A2 (Continued)

Abbreviation	Variable	Unit	Note	Source	Code	Start date	Frequency
Stock-us	US stock market index	Index	S&P 500	Ref DS: S&P	S&PCOMP	1980:1:1	d
Bond	US government bond market index	Index	US tracker all Lives DS government index	Ref DS: DS	TUSGVAL (RI)	1980:1:1	d
Stock	World stock market index	Index		Ref DS: DS	TOTMKWD	1980:1:1	d
IR	US interbank rate, 3 months	%		TR DS: Reuters	USINTER3	1980:1	m
IR-long	US treasury constant maturity, 10 years	%		Ref DS: US Fed	FRTCM10	1980:1:1	d

Note: All variables with the unit "index" are indexed at 2000:1=100. The volatility is calculated as the standard deviation of the daily returns in a given month, annualized. The correlations are calculated between returns in the respective markets over the last 65 days (~ 3 months) or over the last 130 days (~ 6 months), recorded at the end of a given month. If daily data are available for a given variable the monthly values are computed as the averages of the daily values in a given month.

Abbreviations: b, barrel; BLS, Bureau of Labor Statistics; d, day; DS, datastream; GSCI, goldman sachs commodity index; own calc., own calculations; pp, percentage points; Ref, refinitiv; SD, standard deviation; S&P, standard and poors.

TABLE A3 Results of regime testing over the out-of-sample period.

Threshold variable	Mean number of regimes	Median number of regimes
Long- and short-term US int. rate spread	2.05	2.00
US stock market volatility	1.42	1.00
US consumer confidence index	2.05	2.00
Cor. betw. US stock & bond markets, 6m	1.82	2.00
Cor. betw. world stock & oil markets, 6m	1.57	2.00
Inflation	1.71	2.00
Oil price change	1.37	1.00
US composite leading indicator	1.75	2.00

Note. The table presents the mean and median number of regimes chosen by recursive testing based on a Bai-Perron test (Bai & Perron, 2003).