Journal of Applied Econometrics

WILEY

Journal of APPLIED ECONOMETRICS

RESEARCH ARTICLE OPEN ACCESS

Forecasting Natural Gas Prices in Real Time

Christiane Baumeister[1,2,3] | Florian Huber[4,5] | Thomas K. Lee[6] | Francesco Ravazzolo[7,8]

[1]University of Notre Dame, Notre Dame, Indiana, USA | [2]National Bureau of Economic Research, Cambridge, Massachusetts, USA | [3]Centre for Economic Policy Research, London, England, UK | [4]University of Salzburg, Salzburg, Austria | [5]International Institute for Applied Systems Analysis, Laxenburg, Austria | [6]U.S. Energy Information Administration, Washington, DC, USA | [7]BI Norwegian Business School, Oslo, Norway | [8]Free University of Bozen-Bolzano, Faculty of Economics and Management, Bolzano, South Tyrol, Italy

Correspondence: Christiane Baumeister (cjsbaumeister@gmail.com)

Received: 15 December 2024 | **Revised:** 30 June 2025 | **Accepted:** 8 September 2025

Funding: The work of Christiane Baumeister was supported by grant DE-EI0003268 from the Department of Energy. Francesco Ravazzolo acknowledges financial support from the NextGenerationEU program (Piano Nazionale di Ripresa e Resilienza (PNRR) Missione 4 Componente 2, Investimento 1.5 D.D. 1058 23/06/2022, ECS-00000043).

Keywords: expert forecasts | model confidence set | natural gas futures | out-of-sample forecasting | real-time data vintages | temperature anomalies

ABSTRACT

This paper provides a comprehensive analysis of the forecastability of the real price of natural gas in the United States at the monthly frequency considering a universe of models that differ in complexity and economic content. We find that considerable reductions in mean-squared prediction error relative to a no-change benchmark can be achieved in real time for horizons of up to 2 years. A particularly promising model is a vector autoregressive (VAR) model that includes the fundamental determinants of supply and demand for natural gas. To capture real-time data constraints of these and other predictors, we assemble a rich database of historical vintages from multiple sources. We also compare our model-based forecasts to model-free forecasts provided by experts and futures markets. Given that no single forecasting method dominates, we show that combining forecasts from individual models selected in real time using the model confidence set as a novel criterion for dynamic model selection delivers the most accurate forecasts.

JEL Classification: C11, C32, C52, Q41, Q47

1 | Introduction

Natural gas is an important primary source of energy that is used across all sectors in the economy in varying amounts: It is one of the most popular fuels for residential and commercial heating, it plays a major role in electricity generation, and it has a myriad of industrial uses such as feedstock for fertilizer, hydrogen, and other petrochemical products, and as heat source for steel, glass, and paper manufacturing. Natural gas is also considered the "cleanest" among the traditional fossil fuels given that it produces less carbon dioxide when it is burned compared to coal or petroleum, and thus, it takes center stage in reducing pollution and fostering the energy transition. Given that natural gas plays such a critical role in today's energy system, price developments in the natural gas market matter significantly for a range of energy and environmental policies since they directly impact the costs of power generation and input factors firms face and the electricity and gas bills households pay, which in turn will influence firms' and consumers' energy choices. They also drive policy decisions regarding the energy mix, the promotion of renewables, the supply of affordable energy, infrastructure development, tax subsidies, and energy security. Thus, to design and implement such policies, it is key for federal and state governments as well as other stakeholders such as environmental organizations, utility and clean power companies, regulatory agencies, and natural gas producers to have access to reliable forecasts of natural gas prices.

There exist few sources that can be tapped for obtaining regular forecasts of the price of natural gas. The US Energy Information Administration (EIA) releases monthly forecasts of nominal natural gas prices in its *Short-Term Energy Outlook* (*STEO*) for horizons varying between 1 and 2 years ahead. Besides these institutional forecasts, market-based forecasts can be constructed from daily trading prices of futures contracts that deliver natural gas for several years into the future. Yet another strategy is to assume that the price over the forecast horizon is the same as the one observed today. However, little is known about the accuracy of these readily available forecasts.

What is even more surprising is that despite the growing importance of natural gas as a transition fuel, there have been limited efforts by researchers to date to explore alternative, model-based forecasting approaches. One exception is Ferrari et al. (2021) who forecast a nominal natural gas price index, which is a weighted average of natural gas prices in the United States, Europe, and Japan, at the quarterly frequency using a dataset of around 200 variables for 33 economies whose information is summarized by a sparse dynamic factor model. They document some out-of-sample forecasting success relative to a random walk (RW). Another contribution is Gao et al. (2021) who examine the forecasting ability of monthly autoregressive models with time-varying parameters and stochastic volatility (SV), Markov-switching dynamics, and a hybrid of the two for nominal natural gas prices across regional markets; none of these models consistently outperforms their autoregressive benchmark.

In this paper, we provide a systematic evaluation of the forecastability of the real price of natural gas in the United States against a common benchmark, considering a broad range of modeling approaches that exploit insights from economic theory, market structure, and time-series properties and that differ in complexity and information content. We focus on the United States given that competitive spot markets have been in place since the early 1990s, while elsewhere natural gas prices are determined by long-term contracts linked to oil or other non–market-pricing mechanisms (e.g., Halser et al. 2023; Hupka et al. 2023; Stock and Zaragoza-Watkins 2024).[1] Our goal is to forecast the monthly average level of the real price of natural gas for horizons up to 24 months in a real-time setting.

We consider four model classes. First, we investigate the predictive content of standard univariate models where forecasts depend only on past price dynamics. In particular, we study the accuracy of autoregressive, autoregressive-moving average, and exponential smoothing forecasts. Second, motivated by recently developed structural models for the US natural gas market (see Arora and Lieskovsky 2014; Wiggins and Etienne 2017; Winkler 2023), we explore the predictive power of various supply-side and demand-side determinants of the real price of natural gas which we combine in the form of a vector autoregressive (VAR) model. Relevant fundamental drivers include natural gas production and consumption, the number of active drilling rigs for gas wells, the amount of gas in underground storage, measures of meteorological conditions, and domestic economic activity indicators. Given that most of these predictor variables are subject to real-time data constraints in the form of publication lags and revisions to preliminary data releases, one of the contributions of this paper is to assemble an extensive real-time database of monthly vintages from multiple sources. Accounting for these real-time aspects is crucial to avoid look-ahead bias when assessing the forecasting performance of economic models. Mindful that the domestic natural gas market might be subject to unusual episodes such as extreme weather events and infrastructure bottlenecks, we examine whether introducing time-varying volatility impacts the accuracy of natural gas price forecasts. Third, we consider several energy price models that are derived from the interaction between natural gas and oil markets. Because of fuel substitutability in the power sector, the US natural gas prices are closely tied to the evolution of oil prices, with both prices moving together in the long run (see Brown and Yücel 2008; Hartley et al. 2008). We quantify the predictive ability of price spread models as well as bivariate VAR models with and without cointegration restrictions imposed. Given that over time the oil-gas nexus has been affected by technological advances in power generation, the shale gas revolution, and shortages of transportation facilities (see Ramberg and Parsons 2012; Stock and Zaragoza-Watkins 2024), we also examine the benefit of allowing for regime switches in the relationship between the two fuel prices. Fourth, we propose a hybrid model that accommodates arbitrage between natural gas and oil prices as well as fundamental drivers of the natural gas market. This integrated-markets model is the most comprehensive model since it contains both own-market and cross-market determinants of natural gas prices.

Our analysis reveals that several forecasting methods perform quite well in real time compared to the no-change benchmark for horizons up to 24 months. At the nearest horizon, using the most recent daily observation of the natural gas price to forecast the average price next month delivers the most accurate forecast, but this approach is superseded at the 3- and 6-month horizons by more precise forecasts based on futures prices and various economic models of the natural gas market that differ in predictor variables but have in common parsimonious dynamics of order one. Forecasting the real price of natural gas at intermediate horizons, from 9 to 15 months, is most successful with futures prices, exponential smoothing, and an economic model that includes the full set of fundamental drivers. Exponential smoothing remains competitive at the longest horizons, but adding SV to some of the economic models also achieves substantial improvements in forecasting performance. While these are the most promising models across horizons, the differences in average performance with the next best tier of models are often small.

This summary highlights that there is no clear winner in this forecasting horserace which begs the question of which model to rely on to produce accurate real-time out-of-sample forecasts. Since the predictive content of individual models changes considerably over time, we propose a novel criterion that selects the best-performing models dynamically in real time based on the model confidence set (MCS) of Hansen et al. (2011). The benefit of MCS in the current application is that, given a loss function, it can be applied to *any* forecast independent of how it has been generated. We show that the MCS procedure is successful at picking the relevant models in real time since their pooled forecasts yield impressive accuracy gains at all horizons dominating all other individual forecasting approaches.

The plan for the paper is as follows. Section 2 describes the forecasting environment and evaluates the performance of three model-free forecasting approaches. Section 3 proposes a diverse set of forecasting models and explores their relative merits. Section 4 conducts a joint assessment of the predictive accuracy of the entire model space, introduces MCS as a new, versatile tool for dynamic model selection, and examines its promise for forecast combinations. Section 5 offers some concluding remarks.

2 | The Forecasting Environment

2.1 | Forecast Target and Evaluation Metric

Our objective is to forecast the average monthly value of the real Henry Hub spot price of natural gas which serves as the benchmark price for the entire North American natural gas market and parts of the global liquefied natural gas (LNG) market. Daily and monthly spot prices have been published by the EIA since 1997M1. Monthly data for Henry Hub natural gas prices going back to 1993M11 were released by the *Wall Street Journal* (WSJ), but their publication was discontinued in 2014M3. This series is made available in the FRED database of the Federal Reserve Bank of St. Louis. Before that, the natural gas wellhead price, which is the average sales price across several trading hubs in the Southwestern United States from which Henry Hub emerged as the leading pricing point in the early 1990s when it became the official delivery location for natural gas futures contracts traded on the NYMEX, was used as the reference price. This monthly series is provided by the EIA starting in 1976M1 but ceased to be reported at the end of 2012. We construct a historical natural gas price series based on the wellhead price for the period 1976M1 to 1993M10, the WSJ Henry Hub price for 1993M11 to 1996M12, and the EIA spot price from 1997M1 to 2024M2.[2]

Throughout the analysis, we mimic as closely as possible the situation of a real-life forecaster who can rely only on the information available at the point in time the forecast is generated. This means working with preliminary data that are subject to revisions later on and taking delays in data releases into account to accurately reflect real-time data constraints when assessing the out-of-sample performance of forecasting methods. Since our focus is on the real price of natural gas, we deflate the price by the US CPI in real time. For this purpose, we update the real-time vintages of the monthly seasonally adjusted US CPI originally compiled by Baumeister and Kilian (2012) up to February 2024.[3] While the nominal spot price is available in real time, CPI data are published with a 1-month lag; we nowcast the missing observation using the past average inflation rate.

We produce out-of-sample forecasts for monthly horizons h of up to 2 years. Model-based forecasts are obtained by recursively re-estimating the model at each forecast origin t based on data contained in the real-time vintage t. The evaluation period starts in February 1997 determined by the EIA's earliest reporting of the Henry Hub spot price. Thus, the initial estimation window runs from 1976M1 to 1997M1 using data from the January 1997 vintage. After generating h-step-ahead forecasts, we move to the February 1997 vintage which adds one more observation. We repeat estimation and forecasting for all vintages until February 2024 which is the last available vintage and thus the endpoint of our evaluation period.

Since our interest centers on point forecasts, we use the mean-squared prediction error (MSPE) as our metric to assess the forecasting ability of alternative models. The real-time forecasts are evaluated against the final release of the real price of natural gas which we take to be the values in the May 2024 vintage. The MSPE results of all candidate models are normalized relative to the monthly RW, the established benchmark in the energy price forecasting literature (e.g., Hamilton 2009; Baumeister et al. 2017, 2022, 2024; Ferrari et al. 2021). An MSPE ratio below one indicates an improvement in accuracy, while a value above one indicates a deterioration in accuracy.

2.2 | Model-Free Forecasts

A question that has been raised recently is whether the monthly RW is the appropriate benchmark for forecast comparisons of energy prices. In the context of oil price forecasting, Ellwanger and Snudden (2023) argue that the end-of-month price is a stricter benchmark than the average price over the month. This issue was first examined by Baumeister and Kilian (2014) in relation to forecasting the quarterly price of oil where they show that the most recent monthly observation is indeed tougher to beat than the quarterly average. Ellwanger and Snudden (2023) arrive at the same conclusion when comparing the last-day-of-the-month RW forecast with the monthly average.

We start by addressing this question for natural gas price forecasts. The first column of Table 1 presents the average MSPE ratios for forecasts generated using the closing price on the last trading day of each month deflated by the nowcasted CPI relative to the monthly average price deflated by the same value for selected horizons. Using end-of-month prices yields a substantial gain in forecast accuracy at the 1-month horizon with an MSPE reduction of 33%. This forecasting success is not entirely surprising given that the last observed daily price contains more recent information and thus is "closer" to the monthly average of the subsequent month which explains its superior performance one month ahead; however, this informational advantage quickly dissipates. While there is still a small improvement of 2% at the 3-month horizon, from $h = 6$ onward, the end-of-month no-change forecast is dominated by the monthly no-change benchmark. In sum, for the forecast horizons we are interested in, the monthly RW remains the relevant benchmark.

A potentially valuable source for forecasting natural gas prices is the futures market. Since futures contracts make it possible to lock in a price today for delivery of a fixed quantity of natural gas at a specified date in the future, the market price of futures contracts with different maturities can be viewed as the aggregated expected value of the spot price at expiry (see Baumeister 2023). Thus, futures prices should have predictive content for future spot prices. As in Baumeister and Kilian (2012, 2014), we produce forecasts using the following futures-spot spread relation: $R^{HH}_{t+h|t} = R^{HH}_{t}\,(1 + f^{HH,h}_{t} - s^{HH}_{t} - E_t(\pi_{t+h}))$, where R^{HH}_{t} denotes the average level of the real natural

TABLE 1 | Average MSPE ratios relative to the no-change forecast of the real natural gas spot price for model-free forecasting approaches.

	End-of-month No-change forecasts		Futures-market-based forecasts		US EIA expert forecasts
Monthly horizon	**(1) 1997.2–2024.2**	**(2) 2005.9–2024.2**	**(3) 1997.2–2024.2**	**(4) 2005.9–2024.2**	**(5) 2005.9–2024.2**
1	**0.671***	**0.439****	**0.897****	**0.891****	1.631
3	**0.976**	**0.852**	**0.978**	1.068	1.136
6	1.051	1.044	**0.848***	**0.941**	**0.796**
9	1.046	1.040	**0.792****	**0.849**	**0.645****
12	1.051	1.065	**0.758****	**0.802***	**0.615****
15	1.054	1.070	**0.811****	**0.849**	—
18	1.072	1.073	**0.838***	**0.851**	—
21	1.065	1.050	**0.859**	**0.813**	—
24	1.076	1.050	**0.892**	**0.799**	—

Note: Boldface indicates improvements relative to the monthly no-change forecast. Significant at **5% and *10% based on the Diebold and Mariano (1995) test. Expert forecasts for the nominal Henry Hub spot price are reported in the US EIA's *Short-Term Energy Outlook* from August 2005 onward and consistently cover a maximum horizon of 12 months ahead.

gas price in month t, $f_t^{HH,h}$ denotes the log of the current price for a futures contract deliverable at Henry Hub with maturity h, s_t^{HH} denotes the log of the Henry Hub spot price, and $E_t(\pi_{t+h})$ denotes the expected inflation rate over the next h months which we approximate with average inflation over a rolling window of 10 years.[4] Column 3 of Table 1 shows that market-based forecasts beat the monthly no-change forecast at all horizons. Futures perform best in the medium term with accuracy gains between 19% and 24% for $h = 9, 12$, and 15, while MSPE reductions for short and long horizons range from 10% to 16%. For six out of the nine horizons, improvements are statistically significant according to the Diebold and Mariano (1995) test for equal MSPE.

Another model-free approach is to rely on expert forecasts. The EIA regularly publishes monthly forecasts of nominal natural gas prices in its *STEO* report. We compiled the real-time forecasts for the Henry Hub price from past issues of *STEO* which were first reported in 2005M8. The forecast horizon varies from a minimum of 12 to a maximum of 23 months ahead. These features restrict both the sample and horizons over which we can evaluate these expert forecasts. We deflate the nominal price forecasts with the same measure of expected inflation as for the futures-spot spread model. The last column of Table 1 reveals that expert forecasts result in considerable losses compared to the RW one and three months ahead, but they greatly outperform the benchmark for subsequent horizons with statistically significant gains in accuracy of 35% at $h = 9$ and 38% at $h = 12$.

To enable a direct comparison with the EIA expert forecasts, we also report the MSPE ratios for the end-of-month RW and the market-based forecasts for the same shorter evaluation period in Columns 2 and 4, respectively. As before, accuracy gains of the last-day-of-the-month no-change forecast do not extend beyond 3 months. While the short-run MSPE reductions are even more impressive, these improvements vanish quickly as the horizon lengthens. The futures market offers better forecasts at the longest horizons with additional gains of 4% at $h = 21$ and of 10% at $h = 24$ but loses its edge at intermediate horizons with a relative deterioration of 5% on average.

The evidence so far suggests that the simple end-of-month RW is only useful for very near-term forecasts, and market-based forecasts are in the lead for long horizons, whereas the EIA experts have a clear advantage at medium-term horizons. While all these model-free approaches yield readily available forecasts of the real price of natural gas that can be used in policy- and decision-making, it remains unclear what drives those forecasts. This begs the question of whether we can improve upon existing market-based and expert forecasts by designing alternative forecasting models for which we choose the determinants.

3 | A Universe of Forecasting Models

We explore the relative forecasting performance of a diverse set of candidate models that differ in information set and economic content. In Section 3.1, we examine several simple time-series models that are solely based on past price dynamics. In Section 3.2, we turn to economic models of the natural gas market that feature the key determinants of the real price of natural gas on both the supply and demand side which we measure drawing on our newly-assembled real-time database. In Section 3.3, we consider forecasting approaches based on the link between natural gas and crude oil markets that finds its origin in the (imperfect) substitutability between these two energy sources. In Section 3.4, we jointly model own- and cross-market fundamentals. The regression models include the real natural gas price either in logs or growth rates for estimation. Given that our loss function is specified in levels, we exponentiate the log forecasts and accumulate and exponentiate the growth forecasts to get forecasts of the level of the real price of natural gas for evaluation purposes.

3.1 | Models of Past Price Dynamics

It is well-known that parsimonious time-series models that exploit the dynamic relationship of the current price with its own past often deliver quite accurate forecasts. A useful starting point is an ARMA$(1, 1)$ model for the real price of natural gas in logs which we estimate numerically by Gaussian maximum likelihood methods. Conditional on the estimated coefficients and the most recent observations, we iterate the model forward to produce h-step-ahead forecasts for the log of the real price and then convert it to levels. Column 1 of Table 2 shows that the ARMA$(1, 1)$ model outperforms the RW at all horizons except for the shortest one but yields lower MSPE reductions compared to the futures-spread-based forecast, in particular for intermediate horizons. Since we cannot rule out a priori that the log of the real price of natural gas is a unit root process, we consider an MA(1) model in percent changes, thus an IMA(1), as an alternative. Column 2 reveals that imposing a unit root on the ARMA$(1, 1)$ process produces MSPE ratios well above one across all horizons, indicating that the process is not well described by a (near) unit root. Relaxing this constraint by estimating an ARIMA$(1, 1)$ does not alter the message (see Column 3).

Another standard choice is to use purely autoregressive AR(p) forecasting models which can alternatively be estimated by unrestricted least squares (LS) or Bayesian methods that shrink unconstrained models toward a parsimonious benchmark. The latter have been shown to pay off in highly parameterized models where overfitting is a concern (see, e.g., Doan et al. 1984; Huber and Feldkircher 2019). As in Baumeister and Kilian (2012), we rely on the data-driven approach of Giannone et al. (2015) for selecting the optimal degree of shrinkage in real time based on the marginal data density. We set the lag length $p = 12$ as is common for monthly data. Column 4 shows that the AR(12) model produces hardly any reductions in MSPEs compared to the RW except at horizon 24 where we find a marginal improvement. Interestingly, while Bayesian shrinkage achieves some accuracy gains over the RW from horizon $h = 6$ onward (see Column 5), the BAR(12) model is not competitive with the ARMA$(1, 1)$ at any horizon. An alternative strategy to determine the lag order is to rely on the Akaike Information Criterion (see, e.g., Marcellino et al. 2006), allowing the number of lags to be optimally chosen for each vintage. To foster parsimony, we impose an upper limit of six autoregressive lags. Columns 6 and 7 indicate that this approach improves considerably on the (B)AR models with more lags and makes the forecasts competitive with the ARMA$(1, 1)$ model for horizons up to 1 year and even better thereafter with MSPE reductions of 15%–17%. Once we use the AIC to select the lag length, there is little difference between the AR(AIC) and BAR(AIC) forecasts. Thus, the Bayesian shrinkage does not play much of a role. Reducing the lag length to $p = 1$ helps outperform the no-change forecast at short horizons but at the cost of losing some forecast accuracy at longer horizons relative to their AIC counterparts (see Columns 8 and 9).

A very different forecasting method is recursive exponential smoothing which uses a weighted average of past realizations to predict future values with exponentially decaying weights for observations further in the past. This approach is suitable for data that are not trending. Given that the log level of the real price of natural gas has no discernible trend, it is an obvious candidate for applying exponential smoothing. Following Baumeister et al. (2017), we set the smoothing parameter to 0.8 which implies a moderate degree of smoothing in line with macroeconomic data (e.g., Faust and Wright 2013). The last column shows that exponential smoothing performs poorly at horizons 1 and 3 but yields large accuracy gains as the horizon lengthens. Beyond the 1-year horizon, it dominates market-based forecasts with statistically significant MSPE reductions of around 20%.

3.2 | Economic Models of the Natural Gas Market

The price formation in the natural gas market results from the interaction of supply and demand forces. This suggests moving beyond modeling the dynamics of the equilibrium outcome since additional information can be derived from the fundamental determinants of the real price of natural gas suggested by economic theory and market structure. In Section 3.2.1, we develop a new real-time dataset that covers economically motivated predictors that provide the basis for a rich economic model of the natural gas market, which in Section 3.2.2, we estimate in the form of a VAR and evaluate for the purpose of out-of-sample forecasting.

3.2.1 | A Real-Time Dataset of Economic Determinants of US Natural Gas Prices

We construct a comprehensive real-time database that contains variables that according to economic theory should have predictive power for the future path of natural gas prices and that allows us to account for publication delays as well as subsequent data revisions when assessing the accuracy of our forecasting models. The dataset is at the monthly frequency and has been assembled from a variety of sources. It consists of vintages from 1991M1 to 2024M2, each covering data going back to 1973M1 where available.[5] Table 1A in the Supporting Information Appendix gives an overview of the components of the real-time database, data sources, start dates of vintages and data as well as lags in releases and nowcasting rules to fill those gaps.

Natural Gas Market Fundamentals. The amount of natural gas produced domestically is one of the fundamental factors for price setting in the national natural gas market. The standard measure is dry natural gas production which is the quantity of natural gas extracted from gas and oil wells purified from other compounds and of high enough quality to be eligible for pipeline transportation. An alternative measure is marketed natural gas withdrawals which contain not only the consumer-grade dry natural gas but also other liquids. These “wet” components tend to be sold separately which is why we consider this second production concept less suitable to represent the supply side of the market, but the quantitative difference is small. US natural gas rig counts which measure the number of active drilling rigs employed in the search for gas are an indicator for the intensity of resource development and thus trends in future production. It stands to reason that movements in rig counts affect expectations about the future balance of supply and demand thereby inducing price swings. Another important price determinant is the volume of natural gas held in underground storage facilities since injections into and withdrawals from stockpiles allow to balance the more seasonal demand with the more steady supply. Natural gas inventories are measured as working natural gas in underground storage which is the flow component that adjusts to market conditions and seasonal factors. Price developments will also be influenced by the demand behavior of multiple end users. The quantity of natural gas delivered to households, manufacturing industries, electric power plants, commercial entities, and the transportation sector is measured by total natural gas consumption. We also consider residential consumption separately since households' demand patterns differ from those of other sectors given their limited substitution possibilities which can exacerbate price changes.

Weather Conditions. A more direct measure for consumption are variations in temperature and other weather phenomena since they are important drivers of heating demand and demand for air conditioning in residential homes and commercial buildings which account for roughly one third of natural gas use. For example, colder than usual winters typically lead to increased heating demand and low winds during the summer lead to the substitution from wind-generated to gas-generated energy, exerting upward pressure on natural gas prices. However, severe weather not only affects the demand side but also interferes with natural gas supply by disrupting production or transportation (e.g., hurricanes, winter storms, and freezes). We consider several measures to capture the impact of meteorological conditions on price developments. Two widely used measures are heating and cooling degree days which are based on outdoor air temperature recorded daily at weather stations nationwide and converted into a quantitative index of energy requirements to heat or cool spaces weighted by population. A related indicator constructed by the National Oceanic and Atmospheric Administration (NOAA) is the Residential Energy Demand Temperature Index (REDTI) which uses heating and cooling degree days as input to determine residential energy needs. A very different measure is NOAA's national index for temperature anomalies which focuses on periods of extreme heat or cold snaps computed as deviations from long-term averages.

Macroeconomic Environment. Another key factor of demand for natural gas that has a direct impact on the price is the state of the domestic economy. Growth-driven increases in natural gas consumption can be particularly strong in the industrial sector, which uses natural gas as a fuel and a feedstock in the production process to make goods such as fertilizer and pharmaceuticals. This suggests using US industrial production (IP) or the related capacity utilization rates as cyclical indicators since both measures cover manufacturing, mining, and electric and gas utilities. A broader measure that is routinely used to gauge overall economic activity is the Chicago Fed National Activity Index (CFNAI). Among the three proxies for the US business cycle, the degree of capacity utilization seems to be the least affected by the sharp COVID-19 contraction, which makes this our first choice.

Backcasting. Given that historical data for the natural gas market fundamentals as well as the heating and cooling degree days are only available in print format in published issues of the *Monthly Energy Review* (*MER*), they cover at most a period of 54 months. We therefore assume that observations that are dropped from the historical editions are no longer revised, and we use the last record in the construction of subsequent vintages. To backcast the data all the way to 1973M1, we collect data from issues going back to 1976M9 where possible. For those series that were not reported in all the past issues of the *MER*, we approximate the missing observations in the early part of the sample using data from the most recent vintage. In this way, all variables from the *MER* can be extended back in time to at least 1976M1 except for natural gas drilling rigs which started to be measured separately only in 1987M8. We extrapolate the data before that back to 1973M1 using the growth rate of crude oil and natural gas rig counts. This is facilitated by the fact that these two series are not revised. NOAA does not provide vintages for REDTI and the index of temperature anomalies, so we assume that these data are not revised. We account for the publication lag of 1 month in REDTI when compiling pseudo vintages. While we have a complete set of vintages for IP, the first vintage for capacity utilization is available in 1996M11 and for CFNAI in 2001M1, both containing data back to 1973M1. Earlier vintages are based on data in these first actual vintages adjusted for the 1-month delay in their release.

TABLE 2 | Average MSPE ratios relative to the no-change forecast of the real natural gas spot price for a set of univariate models.

Monthly horizon	ARMA(1,1) (1)	IMA(1) (2)	ARIMA(1,1) (3)	AR(12) (4)	BAR(12) (5)	AR(AIC) (6)	BAR(AIC) (7)	AR(1) (8)	BAR(1) (9)	Exponential smoothing (10)
1	1.019	1.049	1.049	1.076	1.059	1.043	1.033	**0.987**	**0.987**	2.120
3	**0.962**	1.042	1.038	1.071	1.038	**0.978**	**0.976**	**0.956**	**0.956**	1.257
6	**0.895**	1.055	1.040	1.031	**0.987**	**0.891**	**0.891**	**0.900**	**0.899**	**0.941**
9	**0.854**	1.083	1.060	**0.999**	**0.946**	**0.849**	**0.849**	**0.855**	**0.854**	**0.819***
12	**0.843**	1.099	1.073	1.001	**0.937**	**0.827**	**0.828**	**0.837**	**0.837**	**0.798****
15	**0.857**	1.122	1.088	**0.980**	**0.926**	**0.825**	**0.828**	**0.836**	**0.836**	**0.802****
18	**0.874**	1.141	1.106	**0.987**	**0.934**	**0.830**	**0.834**	**0.843**	**0.843**	**0.808****
21	**0.888**	1.163	1.141	**0.995**	**0.940**	**0.836**	**0.841**	**0.853**	**0.854**	**0.792****
24	**0.915**	1.190	1.175	**0.966**	**0.923**	**0.844**	**0.850**	**0.871**	**0.871**	**0.782****

Note: Evaluation period: 1997.2–2024.2. Boldface indicates improvements relative to the no-change forecast. BAR refers to AR models estimated using the Bayesian method of Giannone et al. (2015). The AIC lag order estimates are based on an upper bound of six lags for parsimony. The exponential smoothing forecasts are based on a weight of 0.8. The statistical significance of MSPE reductions is only assessed for the exponential smoothing forecasts based on the Diebold–Mariano (1995) test with significance at **5% and *10%. All forecasts are generated recursively from data subject to real-time data constraints.

Nowcasting. To obtain a balanced dataset, we fill in missing observations at the end of each vintage by devising a set of nowcasting rules that are informed by the time-series properties of our variables, such as seasonal patterns or trend behavior. As shown by Baumeister and Kilian (2012) for oil prices and by Baumeister et al. (2017) for gasoline prices, simple, variable-specific rules work particularly well in energy price forecasting models. For natural gas production, which features between one and three missing values depending on the vintage, we use the annual log difference from the preceding year to take seasonality into account when filling the gaps. To accommodate the seasonal pattern of working gas underground inventories, we apply the month-on-month change 1 year prior to the last observation to nowcast the most recent values. For natural gas consumption, we assume that the same amount is consumed as in the same month the year before. Given the seasonal behavior of the three temperature series that are updated with a delay, a sensible nowcasting rule is to rely on the average value for the same month over the 10 most recent years to account for slow-moving climatic changes. We extrapolate the one missing observation for IP based on its average monthly growth rate. Nowcasts for the current values of capacity utilization and CFNAI are obtained with exponential smoothing using a parameter of 0.95, which is the same that Faust and Wright (2013) applied to the unemployment rate.

3.2.2 | Forecasting With VAR Models

We model the dynamic relationship between the real price of natural gas and its economic determinants as a reduced-form VAR, which can be written as $\boldsymbol{y}_t = \boldsymbol{c} + \boldsymbol{\Phi}_1 \boldsymbol{y}_{t-1} + \cdots + \boldsymbol{\Phi}_p \boldsymbol{y}_{t-p} + \boldsymbol{\epsilon}_t$, where $\boldsymbol{y}_t$ is a $n \times 1$ vector of monthly data, $\boldsymbol{c}$ is a $n \times 1$ vector of intercepts, $\boldsymbol{\Phi}_i, i = 1, \dots, p$, are $n \times n$ coefficient matrices with p being the number of lags, and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. In addition to the log of the real price of natural gas, in our baseline model, $\boldsymbol{y}_t$ includes dry natural gas production, natural gas rig counts, working gas inventories, capacity utilization, and the REDTI as predictors with $n = 6$. For the three natural gas fundamentals, we consider transformations to log-levels and to monthly growth rates by taking the first difference of the natural logarithm. We also examine the sensitivity of our results to alternative choices for the last two variables and explore other model specifications that have been proposed for structural analysis of natural gas price fluctuations.

As in the case of AR models, we estimate the VAR with unrestricted LS as well as Bayesian shrinkage methods starting with a fixed lag order of $p = 12$. Columns 1 and 7 of Table 3 show that the VAR(12) is unsuccessful in improving upon the RW independent of the transformation of the gas market fundamentals, with the MSPE ratios exceeding 1 for all horizons. This does not mean that the economic variables have no predictive power but rather that the VAR(12) is overparameterized which hurts its out-of-sample forecasting performance. In fact, applying Bayesian shrinkage yields more precise forecasts; so much so that the BVAR(12) beats the RW at all horizons except for $h = 1$ but only for the growth-rate specification (Column 2). Reducing the lag length to six or less by means of the AIC leads to more substantial accuracy gains for variables in log-levels and changes. The size of the MSPE reductions for both BVAR(AIC) models is comparable to the BAR(1) from horizon 3 onward. Columns 5 and

TABLE 3 | Average MSPE ratios relative to the no-change forecast of the real natural gas spot price based on an economic model of the US natural gas market: The role of lag length and variable transformation.

Monthly horizon	(1) VAR(12)	(2) BVAR(12)	(3) VAR(AIC)	(4) BVAR(AIC)	(5) VAR(1)	(6) BVAR(1)	(7) VAR(12)	(8) BVAR(12)	(9) VAR(AIC)	(10) BVAR(AIC)	(11) VAR(1)	(12) BVAR(1)
	Predictors in monthly growth rates						Predictors in log-levels					
1	1.179	1.028	1.126	1.015	**0.968**	**0.969**	1.169	1.035	1.110	1.006	**0.970**	**0.969**
3	1.054	**0.950**	**0.995**	**0.954**	**0.935**	**0.935**	1.078	1.029	**1.000**	**0.972**	**0.926**	**0.926**
6	1.089	**0.923**	**0.943**	**0.900**	**0.863**	**0.865**	1.305	1.122	**0.952**	**0.898**	**0.845**	**0.852**
9	1.082	**0.873**	**0.872**	**0.836**	**0.822**	**0.824**	1.525	1.198	**0.881**	**0.827**	**0.794**	**0.804**
12	1.213	**0.863**	**0.864**	**0.825**	**0.799**	**0.801**	1.857	1.228	**0.870**	**0.810**	**0.758**	**0.766**
15	1.175	**0.858**	**0.859**	**0.824**	**0.798**	**0.798**	1.777	1.255	**0.887**	**0.827**	**0.762**	**0.767**
18	1.144	**0.871**	**0.865**	**0.834**	**0.798**	**0.797**	1.677	1.282	**0.906**	**0.852**	**0.783**	**0.785**
21	1.187	**0.871**	**0.876**	**0.847**	**0.811**	**0.810**	1.450	1.293	**0.894**	**0.856**	**0.817**	**0.819**
24	1.106	**0.874**	**0.893**	**0.867**	**0.836**	**0.833**	1.432	1.343	**0.918**	**0.883**	**0.851**	**0.850**

Note: Evaluation period: 1997.2–2024.2. Boldface indicates improvements relative to the no-change forecast. BVAR refers to VARs estimated with the Bayesian method of Giannone et al. (2015). The AIC is based on an upper bound of six lags. All forecasts are generated recursively from data subject to real-time data constraints.

6 show that further restricting the number of lags to 1 delivers the best performance with little to choose between the VAR(1) and the BVAR(1). The only model that produces larger MSPE reductions at horizons 21 and 24 is exponential smoothing, but this method records big losses at horizons 1 and 3. The log-level specification also dominates the futures-based forecasts at most horizons. We conclude that the economic drivers of natural gas prices greatly enhance the forecast accuracy if the dynamics are kept parsimonious.

Table 2A in the Supporting Information Appendix compares the performance of our baseline six-variable BVAR(1) model with predictors in logs (Column 1) to models where we replace one variable at a time with alternative indicators for the US meteorological and macroeconomic conditions. Columns 2–4 reveal that using an index for temperature anomalies or heating and cooling degree days in deviations from the historical average in the same month instead of the REDTI delivers equally accurate forecasts. While CFNAI achieves similar reductions in MSPE as capacity utilization across all horizons, IP does slightly better in the near term, but its accuracy decreases as the horizon lengthens with losses as high as 7% 2-years-ahead relative to the other business cycle indicators (Columns 5 and 6).[6]

Table 4 reports results for a set of other models that have not been designed for forecasting but for studying the structural dynamics of the US natural gas market. Given that any structural model is associated with a reduced form, it is useful to evaluate the success of these models in real-time out-of-sample forecasting. The first model is the four-variable VAR of Arora and Lieskovsky (2014) that includes marketed natural gas production on the supply side and IP and residential natural gas consumption on the demand side with variables transformed to annual growth rates to remove seasonality and trends.[7] Column 3 shows that while this smaller model beats the RW across all horizons, it has on average 6% higher MSPE ratios than the more comprehensive baseline model (Column 1). The second small-scale model is from Wiggins and Etienne (2017) who use seasonally adjusted data for production, underground inventories, and IP in monthly growth rates. Column 5 reveals that this model also does not outperform the baseline BVAR, but their choice of variables leads to lower MSPE ratios at longer horizons compared to the other four-variable model. Model 3 is of the same size as the baseline model but features a different selection of variables proposed by Winkler (2023). Active gas drilling rigs and marketed production describe the supply side, while IP is the main demand-side determinant of prices; we add heating and cooling degree days to the demand side which Winkler (2023) includes as exogenous variables. The MSPE ratios in Column 7 indicate that this model is a serious competitor for horizons up to $h = 6$. Model 4 is a modified version of model 3 that does not distinguish between the sources of demand but lumps them all together via total natural gas consumption. This model yields MSPE reductions similar to the baseline up to 9 months out; beyond that, there is a gap of about 3%, but overall, it performs best among the alternative models (Column 9).

3.2.3 | The Role of Time-Varying Volatility

Over the decades, the US natural gas market has experienced important changes in its regulatory and market structure which can impact the forecasting performance of our economic models. Hupka et al. (2023) document that structural shifts often happen quickly and for a variety of reasons. Prominent examples include abrupt production disruptions due to winter storms and hurricanes and consumption spikes due to other extreme weather events like hard freezes and heat waves. Technological advances in drilling and delivery methods or shifts in consumer preferences toward cleaner fuels can also trigger sudden changes in

TABLE 4 | Average MSPE ratios relative to the no-change forecast of the real natural gas spot price for baseline and other BVAR(1) models of the US natural gas market with and without stochastic volatility.

	Baseline		Model 1		Model 2		Model 3		Model 4	
Monthly horizon	**(1) BVAR**	**(2) BVAR-FSV**	**(3) BVAR**	**(4) BVAR-FSV**	**(5) BVAR**	**(6) BVAR-FSV**	**(7) BVAR**	**(8) BVAR-FSV**	**(9) BVAR**	**(10) BVAR-FSV**
1	**0.969**	**0.978**	**0.981**	**0.981**	1.016	**0.985**	**0.960**	**0.989**	**0.967**	**0.984**
3	**0.926**	**0.953**	**0.950**	**0.950**	**0.961**	**0.954**	**0.909**	**0.966**	**0.920**	**0.950**
6	**0.852**	**0.909**	**0.903**	**0.913**	**0.904**	**0.920**	**0.861**	**0.936**	**0.853**	**0.914**
9	**0.804**	**0.876**	**0.866**	**0.874**	**0.858**	**0.887**	**0.827**	**0.914**	**0.813**	**0.882**
12	**0.766**	**0.858**	**0.854**	**0.851**	**0.844**	**0.865**	**0.820**	**0.903**	**0.798**	**0.864**
15	**0.767**	**0.857**	**0.862**	**0.832**	**0.842**	**0.841**	**0.828**	**0.906**	**0.800**	**0.861**
18	**0.785**	**0.870**	**0.875**	**0.816**	**0.849**	**0.830**	**0.850**	**0.913**	**0.811**	**0.866**
21	**0.819**	**0.891**	**0.887**	**0.796**	**0.854**	**0.807**	**0.872**	**0.937**	**0.839**	**0.897**
24	**0.850**	**0.944**	**0.906**	**0.777**	**0.870**	**0.792**	**0.906**	**0.979**	**0.880**	**0.951**

Note: Evaluation period: 1997.2–2024.2. Boldface indicates improvements relative to the no-change forecast. All models contain the real price of natural gas in logs. The additional predictor variables are as follows: (a) Baseline model: log of dry natural gas production, log of working gas inventories, log of natural gas rig counts, capacity utilization, and REDTI; (b) Model 1: annual growth rates of marketed natural gas production, of residential natural gas consumption, and of industrial production; (c) Model 2: monthly growth rates of dry natural gas production, of working gas inventories, and of industrial production, all three deseasonalized; (d) Model 3: log of marketed natural gas production, log of natural gas rig counts, log of industrial production, heating and cooling degree days in deviations from their historical average; (e) Model 4: log of marketed natural gas production, log of natural gas rig counts, and log of total natural gas consumption. All forecasts are generated recursively from data subject to real-time data constraints.

the energy mix. Another possible sources of periodic instabilities are infrastructure bottlenecks in transportation or capacity constraints in underground storage. All these unexpected events might cause an increase in price volatility that interferes with the models' predictive ability.

It is by now well-established that introducing nonlinearities in the form of SV greatly enhances the accuracy of point forecasts of a range of macroeconomic variables and other energy prices compared to models that impose homoskedasticity (see, e.g., Clark and Ravazzolo 2015; Baumeister et al. 2022). In particular, Chan and Grant (2016) show that SV dominates GARCH for modeling energy price dynamics. As noted by Primiceri (2005) and Carriero et al. (2019), allowing for time variation in error variances not only captures jumps in volatility but also interacts with model dynamics in a way that might further improve the precision of point forecasts. We investigate the benefit of modeling time variation in the volatilities of shocks to $\boldsymbol{y}_t$ by postulating the following specification for the VAR residuals as in Kastner and Huber (2020), $\boldsymbol{\epsilon}_t = \boldsymbol{\Lambda} \boldsymbol{q}_t + \boldsymbol{\nu}_t$, where $\boldsymbol{\Lambda}$ denotes an $n \times m$ matrix of factor loadings with m referring to a small number of latent factors in $\boldsymbol{q}_t$, which are conditionally heteroskedastic with a time-varying diagonal covariance matrix $\boldsymbol{H}_t = diag(e^{h_{1t}}, \dots, e^{h_{qt}})$, and $\boldsymbol{\nu}_t$ are measurement errors with a time-varying diagonal covariance matrix $\boldsymbol{\Omega}_t = diag(e^{h_{q+1,t}}, \dots, e^{h_{q+n,t}})$. The law of motion for the $m+n$ logarithmic volatilities h_{it} is given by independent AR(1) processes of the following form, $h_{it} = \mu_{hi} + \rho_{hi}(h_{it-1} - \mu_{hi}) + \sigma_{hi}\eta_{it}$ for $i = 1, \dots, m+n$, with μ_{hi} the unconditional mean of the log volatilities, ρ_{hi} their persistence parameter, σ_{hi} their standard deviation, and $\eta_{it} \sim \mathcal{N}(0,1)$. The time-varying variance of $\boldsymbol{\epsilon}_t$ can thus be written as $\boldsymbol{\Sigma}_t = \boldsymbol{\Lambda} \boldsymbol{H}_t \boldsymbol{\Lambda} + \Omega_t$ which amounts to the factor stochastic volatility (FSV) model proposed by Pitt and Shephard (1999). This specification has several attractive features. First, in contrast to popular alternatives, the FSV model is order-invariant. Second, the factor structure reduces the number of free parameters if $m < n$.[8] Third, despite its parsimonious nature, this model offers rich dynamics by capturing both common and idiosyncratic sources of volatility. Details can be found in the Supporting Information Appendix A.1.

Table 4 compares the forecast accuracy of the economic models with and without FSV. While all the models with time-varying error variances outperform the no-change forecast at all horizons, only Models 1 and 2 improve upon their constant-variance counterparts for horizons beyond 12 months, with the largest MSPE reductions of around 20% 2 years out (Columns 4 and 6). Given that these two models are smaller in size compared to the baseline and Model 3, it might be the case that the FSV model favorably influences the model dynamics to produce better point forecasts at longer horizons, whereas adding more variables compensates for the lack of time-varying volatility. As can be seen in Column 10, this is not borne out by Model 4 which contains the same number of variables as Models 1 and 2 but displays a weaker performance than the same model without FSV. Another difference is that the predictors in Models 1 and 2 are included in growth rates, while all other models are specified in log-levels, suggesting that FSV has more bite for models estimated in first differences. Thus, while yielding accuracy gains for some models, BVAR-FSV models only do better than the baseline BVAR at horizons $h = 21$ and $h = 24$.[9]

3.3 | Energy Price Models: A Tale of Two Markets

So far, we considered economic fundamentals within the domestic natural gas market as the main predictors of natural gas price dynamics, but natural gas prices also have close ties with other energy prices, in particular crude oil because of the possibility of fuel switching. In fact, historically, natural gas was often considered a by-product of oil production, and thus, production was not necessarily determined by market forces. Instead, the ability of large-volume fuel consumers such as power plants and iron, steel, and paper mills to switch between natural gas and petroleum as a function of their cost establishes a link between the two markets which suggests to jointly model natural gas and oil prices giving rise to bivariate energy price models; yet, the two fuels are not perfect substitutes due to differences in heat content, transportation costs, and market structure. In contrast to other countries where oil indexation is the dominant pricing mechanism for natural gas, energy markets in the United States are more competitive and relative price movements determine the marginal market (Hupka et al. 2023). This means that there is no mechanical relationship between these two prices, and the forecasting content of bivariate fuel-pricing models is an empirical question.

Figure 1A in the Supporting Information Appendix shows that there is strong comovement between the Henry Hub natural gas and West Texas Intermediate (WTI) oil spot prices from 1993 to 2010 but no clear pattern thereafter.[10] For this first period, Brown and Yücel (2008) and Hartley et al. (2008) find a cointegration relationship between natural gas and oil prices and use an error correction model to study the short-run and long-run interaction effects among both markets. Bachmeier and Griffin (2006) also provide empirical evidence for energy market integration. However, the stability of this long-run relationship has been questioned by Ramberg and Parsons (2012) who show that the price of oil has only weak explanatory power for short-term natural gas price fluctuations. The advent of the shale gas revolution in 2006 marks a major turning point that first weakened, then severed the linkages between the global oil market and the US natural gas market. The steep increase in gas production from fracking led to the decoupling of natural gas prices from oil prices given that shale gas was landlocked (due to limited pipeline capabilities) until 2016 when LNG export terminals enabled shipments initiating the recoupling of the US natural gas prices to international energy price developments (see Stock and Zaragoza-Watkins 2024).

Against this background, we investigate the promise of two types of models that explicitly or implicitly feature a cointegrating relationship. Specifically, we explore the predictive power of price spread models in Section 3.3.1 and of bivariate VAR models in Section 3.3.2.

3.3.1 | Price Spread Models

This forecasting model is based on the assumption that the nominal spot prices of natural gas and oil are cointegrated. In that case, current deviations of the spot price of natural gas from the spot price of oil would be expected to have predictive power for cumulative changes in the nominal spot price of natural gas: $\Delta s^{h,HH}_{t+h|t} = \alpha + \beta(s^{HH}_t - s^{WTI}_t) + \varepsilon_{t+h}$, where s^{HH}_t is the log of the nominal spot price of natural gas, s^{WTI}_t is the log of the nominal spot price of WTI crude oil, and $\Delta s^{h,HH}_{t+h|t}$ denotes the cumulative change in s^{HH}_t over the next h months.[11] Similar to Baumeister et al. (2017), we map recursive estimates of this relationship into a forecast for the real price of natural gas as follows: $R^{HH}_{t+h|t} = R^{HH}_t \exp\left[\hat{\alpha} + \hat{\beta}(s^{HH}_t - s^{WTI}_t) - E_t(\pi_{t+h})\right]$, where $E_t(\pi_{t+h})$ denotes expected inflation over the next h months approximated as before.

Columns 1 and 2 in Table 5 report the average MSPE ratios for two alternative specifications of the price spread models, an unrestricted variant, and a restricted variant where, in the interest of parsimony, α is set to zero. Both models display a dismal forecasting performance with MSPE ratios as high as 1.5. Only at horizon 1 do price spread models marginally improve on the RW. This result suggests that the maintained hypothesis of cointegration likely fails to hold for the entire evaluation period in line with evidence provided by Stock and Zaragoza-Watkins (2024). To get a better sense of the evolution of the forecasting performance over time, Figure 2A in the Supporting Information Appendix presents recursively updated MSPE ratios for selected horizons. The message that emerges from these plots is that price spread models delivered accurate forecasts until the breakdown of the cointegrating relationship dated January 2009 by statistical tests (see Table 1 in Stock and Zaragoza-Watkins, 2024). Thus, the shift to a regime without cointegration due to the shale gas revolution explains the sudden deterioration in the forecasting ability of these models.

This evidence raises the question of whether a model that explicitly allows for a switch in regime would have been able to detect this change in real time thereby preventing the massive forecast errors in the post-2009 period. To examine this question empirically, we allow the parameters in the price spread model to be driven by a Markov-switching (MS) process as follows: $\Delta s^{h,HH}_{t+h|t} = \alpha_{s_{t+h}} + \beta_{s_{t+h}}(s^{HH}_t - s^{WTI}_t) + \varepsilon_{t+h}$, where $\alpha_{s_{t+h}}$ is the MS intercept, $\beta_{s_{t+h}}$ is the MS spread coefficient, $\varepsilon_{t+h} \sim \mathcal{N}(0, \sigma^2_{s_{t+h}})$ with $\sigma_{s_{t+h}}$ the MS volatility, and $\{s_{t+h}\}$ follows an m-states ergodic and aperiodic Markov-chain process that determines the regime. This process is an unobservable variable which takes integer values $s_{t+h} \in \{1, \dots, m\}$ and has transition probabilities $\mathbb{P}(s_{t+h} = j | s_{t+h-1} = i) = p_{ij}$, with $i, j \in \{1, \dots, m\}$. We set $m = 2$ postulating a switch between a cointegration and a noncointegration regime. We apply a Bayesian approach to estimation (see Supporting Information Appendix A.2). The third column of Table 5 shows that accounting for the possibility of a shift in regime over the evaluation period improves upon the constant-coefficient counterparts at most horizons but only improves upon the RW benchmark at $h = 3$ and $h = 6$.

3.3.2 | Bivariate VAR Models

Given the economic interaction between natural gas and oil markets, an alternative modeling approach that only implicitly allows for a long-term relationship between natural gas and oil prices is a bivariate VAR(p) with both prices included in log-levels. Thus, in the VAR, $\boldsymbol{y}_t = [r^{HH}_t, r^{WTI}_t]'$ is now a 2×1 vector that contains the log prices of Henry Hub natural gas and WTI crude oil, both deflated in real time with the CPI. Columns 4–9 of

TABLE 5 | Average MSPE ratios relative to the no-change forecast for energy price models.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Price spread models			**Bivariate VAR models**					
Monthly	**Constant coefficient**		**Markov switching**						
horizon	$\hat{\alpha}, \hat{\beta}$	$\alpha = 0, \hat{\beta}$	$\hat{\alpha}^{MS}, \hat{\beta}^{MS}$	**VAR(12)**	**BVAR(12)**	**VAR(AIC)**	**BVAR(AIC)**	**VAR(1)**	**BVAR(1)**
1	**0.987**	**0.992**	1.270	1.033	**0.991**	1.031	**0.986**	**0.985**	**0.998**
3	1.057	1.032	**0.956**	1.006	**0.947**	**0.964**	**0.951**	**0.962**	**0.994**
6	1.175	1.117	**0.988**	**0.971**	**0.891**	**0.923**	**0.914**	**0.925**	**0.957**
9	1.369	1.233	1.042	**0.968**	**0.871**	**0.921**	**0.908**	**0.915**	**0.944**
12	1.516	1.337	1.093	1.010	**0.899**	**0.957**	**0.944**	**0.949**	**0.960**
15	1.547	1.382	1.176	1.033	**0.941**	1.010	**0.998**	1.003	**0.993**
18	1.483	1.371	1.290	1.062	**0.978**	1.048	1.042	1.048	1.031
21	1.363	1.302	1.357	1.074	**0.998**	1.065	1.066	1.082	1.059
24	1.332	1.281	1.290	1.081	1.040	1.107	1.116	1.141	1.108

Note: See Table 3.

Table 5 examine the role of different lag orders for the forecasting performance of this model when estimated by unrestricted LS and the Bayesian method of Giannone et al. (2015). Column 4 shows that even in a simple bivariate model, 12 lags prevent the LS estimator at most horizons from delivering forecasts that are more accurate than the RW, whereas Bayesian shrinkage yields MSPE ratios below 1 except at $h = 24$. The largest gains of around 10% are found for medium-term horizons of 6–12 months ahead. Determining the lag length recursively based on the AIC or setting $p = 1$ helps the VAR outperform the RW more often but hurts the performance of the BVAR both in terms of the size of MSPE reductions and the number of horizons for which it outperforms the no-change forecast. It is also interesting to note that the VAR(1) does better than the BVAR(1) for horizons up to 1 year (Columns 8 and 9). Comparing the results to the (B)AR(p) models in Table 2, which are nested in the bivariate (B)VAR(p) models, suggests that information about past oil price dynamics is not particularly useful for obtaining more accurate forecasts except maybe in the short run.

While the VAR model estimated in log-levels remains agnostic about the existence of a cointegrating relationship, it suffers an efficiency loss should the variables be cointegrated. Despite the evidence derived for the price spread models and the narrative about the decoupling of natural gas and oil prices, we investigate the usefulness of imposing cointegration restrictions on the bivariate VARs for forecasting. The vector error correction (VEC) representation is implied by the VAR specification with $\boldsymbol{y}_t = [\Delta r_t^{HH}, r_t^{HH} - r_t^{WTI}]'$. Table 8A confirms our earlier results that enforcing cointegration is detrimental to the forecasting performance. Except for some small accuracy gains at horizons 3–9, VEC(p) models perform poorly independent of lag order and estimation method.

3.4 | Integrating Own- and Cross-Market Fundamentals

A natural question is whether jointly modeling natural gas fundamentals and oil prices can further improve forecast accuracy. We augment our baseline economic model with the log of the real WTI price which results in a model of size $n = 7$. The first three columns of Table 6 show that among the BVAR models with constant variance, the model where the AIC determines the lag length recursively performs best, beating the RW at all horizons. This suggests that the inclusion of oil prices leverages past dynamics to yield more precise forecasts despite the higher dimensionality of the model. Allowing for FSV leads to relative improvements for the model with $p = 12$ but worsens the performance of the model with $p = 1$ (Columns 4 and 6); neither of them outperforms the BVAR(AIC) model except for $h = 1$. Only the BVAR-FSV model with $p = 6$ is competitive at long horizons. However, none of the integrated models that account for arbitrage across fuels does better than the baseline model alone.[12]

4 | Model Comparison and Forecast Pooling

4.1 | Joint Assessment Based on the MCS

The preceding analysis highlights that while there are quite a few models that beat the monthly no-change forecast, there is no single forecasting method that, based on pairwise comparisons, dominates all the others across all horizons, and thus, it is unclear which model to choose for forecasting the real price of natural gas.[13] With such a rich assortment of competitive candidate models, it is not easy to get a good sense of whether some alternatives are more useful than others for out-of-sample forecasting. We shed light on this question with the help of the MCS procedure proposed by Hansen et al. (2011) that allows us to jointly assess the entire model space for a fixed horizon h to decide which models should be included in the MCS based on their predictive power. Specifically, the approach selects the subset of models $\hat{M}^*_{1-\alpha}$ that form the MCS such that it contains the best model with probability $(1 - \alpha)$. The pruning of the initial model space is data-driven, and the number of surviving models reflects the informativeness of the sample.

Table 10A reports the MCS p-values for the entire collection of models evaluated previously, including the RW benchmark, and summarizes the number of models that enter the MCS at each horizon h for confidence levels $\alpha = 0.1$ and $\alpha = 0.25$ based on their squared forecasting errors over the full out-of-sample period.[14] Starting with an initial set of 60 individual models, we

TABLE 6 | Average MSPE ratios relative to the no-change forecast for the baseline economic model augmented with the real WTI oil price with and without factor stochastic volatility.

Monthly horizon	(1) BVAR(12)	(2) BVAR(AIC)	(3) BVAR(1)	(4) BVAR(12)-FSV	(5) BVAR(6)-FSV	(6) BVAR(1)-FSV
1	1.002	**0.992**	**0.971**	**0.980**	**0.984**	**0.988**
3	**0.948**	**0.927**	**0.936**	**0.938**	**0.938**	**0.967**
6	**0.970**	**0.868**	**0.873**	**0.904**	**0.888**	**0.936**
9	**0.950**	**0.823**	**0.844**	**0.893**	**0.866**	**0.917**
12	**0.935**	**0.820**	**0.839**	**0.891**	**0.863**	**0.913**
15	**0.933**	**0.853**	**0.880**	**0.909**	**0.876**	**0.927**
18	**0.978**	**0.899**	**0.931**	**0.933**	**0.888**	**0.955**
21	1.010	**0.921**	**0.987**	**0.968**	**0.909**	**0.992**
24	1.078	**0.973**	1.055	1.012	**0.963**	1.075

Note: See Table 3.

find that for horizons up to 9 months, the MCS basically covers the complete model space. The fact that no models are discarded suggests that the information in the data cannot discriminate between good and bad models; that is, the MCS procedure finds the forecasting ability of any one model not to be significantly worse than any other and thus keeps them all. As emphasized by Hansen et al. (2011), the MCS approach tends to err on the side of caution and does not shrink the initial set if the data provide insufficient information about the relative performance of the forecasting models. As the forecast horizon lengthens, some competitors are eliminated; for example, for $h = 18$ and $h = 24$, only 72% of the models pass the threshold to enter $\hat{M}^*_{75\%}$. But even for longer horizons, the set of superior models remains large.

In addition to determining the size of the set for a given level of confidence, the MCS p-values imply a ranking where models with larger p-values are more likely to be among the best-performing models. The top-5 models are marked in red in Table 10A. What stands out is that the economic models of the natural gas market whose dynamics are summarized by VARs of order one always show up at the head of the pack. At intermediate horizons from $h = 6$ to $h = 18$, the VAR(1) with the baseline predictors entering the model in log-levels receives a p-value of one three times. At long horizons, some models that feature SV start appearing in the upper echelon with one of the alternative economic models being assigned a p-value of one at $h = 24$. Exponential smoothing and simple AR models with moderate lag length are also among the most promising models further out, while most energy price models are removed from $\hat{M}^*_{75\%}$ at longer horizons. Most other univariate models make a poor showing, especially at short horizons. The forecasting ability of models that combine own- and cross-market fundamentals is rated as comparable to economic models of the natural gas market with alternative predictors at the shortest horizons, irrespective of their lag order and volatility component, but these models are ranked considerably lower at horizons beyond one year. Not surprisingly, for $h = 1$, the end-of-month no-change forecast takes the lead with a p-value of one, but its relevance diminishes quickly, and, at horizons 12 and 18, it is even kicked out of $\hat{M}^*_{75\%}$. The futures-market forecast is a strong contender across all horizons and even makes first place at $h = 9$. While the benchmark is always included in $\hat{M}^*_{75\%}$, it never ends up among the top models.

4.2 | Tracking Forecasting Performance Over Time

The evidence presented so far focuses on the average forecast quality of the entire set of models over the full evaluation period. However, as already discussed in Section 3.3.1 for the case of constant-coefficient price spread models, the forecasting performance is not necessarily stable over time and might depend on the prevailing market structure.

To get a better idea to what extent the predictive ability varies also for other models across the out-of-sample period, Figure 1 displays the evolution of the cumulative MSPE ratios relative to the RW for a subset of representative models from each class and two different forecast horizons. The key takeaway from these plots is that there is substantial heterogeneity in the forecast accuracy across models and horizons and that the relative performance of some models changes considerably over time. For example, at the 3-month horizon, the end-of-period RW starts out poorly, while the futures-based forecast is quite successful, but from 2010 onward, they converge and end up with the same average performance with an MSPE ratio of 0.98. For 24-month-ahead forecasts, the BVAR(1) model with predictors in logs is one of the most promising candidates early on but then experiences a notable deterioration in the interim only to reclaim its spot among the best models at the end of the evaluation sample. A similar pattern can be observed for the simple AR(AIC) model. Instead, the accuracy gains from exponential smoothing are relatively constant throughout.[15]

We complement this illustrative evidence with a more formal assessment of changes in the forecasting performance of the entire universe of models in the Supporting Information Appendix C by splitting the evaluation period into four subsamples guided by the institutional regimes defined by Stock and Zaragoza-Watkins (2024). The graphical and subsample analyses reveal some striking changes in the predictive power and relevance of different models in our set both over time and across horizons. Given this variation in performance and the difficulty of settling on a single model for out-of-sample forecasting, we next investigate the promise of combining forecasts, not only to possibly enhance the forecast accuracy further but also to guard against forecast failures due to structural change and model misspecification (see, e.g., Granger and Jeon 2004; Timmermann 2006).

4.3 | Model Selection and Forecast Combination

To explore the benefits of aggregating forecasts from different models, we consider a real-life forecaster who had good reasons for compiling the entire universe of models in January 1997 based on the insights discussed above but without any knowledge of their out-of-sample forecasting performance. During the first year, he pools the forecasts from all 60 models by taking the arithmetic average but then employs the MCS approach to dynamically select the set of competing models that contains the best model with a specified level of confidence α. An attractive feature of the MCS for real-time model selection is that it explicitly acknowledges the fact that more than one model can be the best (see also Granger and Jeon 2004). Given that the procedure is tied to the informativeness in the data, the set of superior models evolves dynamically and thus differs at each forecast origin with models exiting and (re-)entering. We then combine the forecasts of the best-performing models where "best-performing" is defined in two different ways. First, we pool the forecasts from all the models above the threshold for inclusion in the MCS which we set at the $\alpha = 25\%$ significance level. Second, we exploit the ranking of models implicit in the MCS testing procedure and only retain models that yield the highest MCS p-values within $\hat{M}^*_{75\%}$ which essentially means that we pool forecasts from the models with the highest and second highest p-values given that these are associated with the smallest sample loss of all forecasts.[16] The sample loss based on squared forecast errors is determined either recursively or rolling, where, in the former case, the real-life forecaster uses the evidence on the performance of each model since

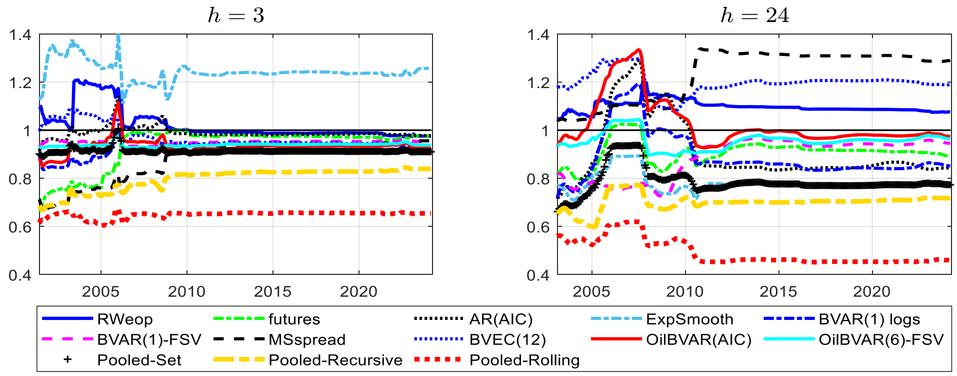

FIGURE 1 | Real-time cumulative MSPE ratios relative to the monthly no-change forecast for a subset of models and model combinations at selected horizons. *Note:* A ratio below 1 indicates that model-based forecasts outperform the no-change forecast. To allow the MSPE ratio to stabilize, we skip the first 50 forecast periods. The three pooled forecasts refer to the equal-weighted forecast combinations.

TABLE 7 | Real-time accuracy of forecast combinations.

					Models with the highest *p*-value:			
	All models		Models in $\hat{M}^*_{75\%}$ set		Recursive		Rolling	
Monthly horizon	Equal weights (1)	Inverse MSPE weights (2)	Equal weights (3)	Inverse MSPE weights (4)	Equal weights (5)	Inverse MSPE weights (6)	Equal weights (7)	Inverse MSPE weights (8)
1	**0.956**	**0.949**	**0.944**	**0.936**	**0.831**	**0.814**	**0.491**	**0.488**
3	**0.913**	**0.908**	**0.910**	**0.905**	**0.839**	**0.838**	**0.655**	**0.654**
6	**0.867**	**0.856**	**0.857**	**0.847**	**0.760**	**0.759**	**0.583**	**0.583**
9	**0.837**	**0.822**	**0.813**	**0.801**	**0.711**	**0.711**	**0.531**	**0.531**
12	**0.833**	**0.814**	**0.766**	**0.758**	**0.699**	**0.698**	**0.522**	**0.522**
15	**0.843**	**0.822**	**0.753**	**0.746**	**0.724**	**0.724**	**0.498**	**0.498**
18	**0.856**	**0.833**	**0.749**	**0.742**	**0.713**	**0.713**	**0.476**	**0.476**
21	**0.866**	**0.840**	**0.770**	**0.764**	**0.756**	**0.755**	**0.473**	**0.473**
24	**0.881**	**0.852**	**0.773**	**0.766**	**0.718**	**0.717**	**0.459**	**0.459**

Note: Evaluation period: 1997.2–2024.2. Average MSPE ratios of pooled forecasts relative to the no-change forecast of the real natural gas spot price. Boldface indicates improvements relative to the no-change forecast. Models that enter the forecast combinations are selected in real time by the MCS procedure with 10,000 block bootstrap replications using a block size of 12 from the entire universe of 60 individual models.

the beginning, while, in the latter case, he only relies on model accuracy observed over the past year.

Table 7 presents the MSPE ratios for these alternative forecast combinations relative to the RW. We start with including all 60 models in the forecast combination in line with the thick modeling approach of Granger and Jeon (2004) which serves as a useful reference. Column 1 shows that these pooled forecasts beat the no-change forecast across all horizons, but none of them is more accurate than the best individual model by horizon. If we restrict the forecast combination to models that end up in $\hat{M}^*_{75\%}$ in real time, we find substantial MSPE reductions of around 24% from $h = 12$ onward which are either tied with or outperform individual models (see Column 3). For horizons up to 9 months, improvements are minor relative to pooling all models. This difference between short and long horizons can be traced back to the weaker sample information at short horizons, which results in a large number of surviving models closer to the "all-models" case.[17] Since it apparently pays to be selective, we now limit pooling to the models ranked first and second according to the MCS *p*-values. Column 5 reveals that using this stricter criterion results in fewer but more promising models yielding further accuracy gains with MSPE reductions ranging from 16% at the 3-month horizon to 30% at the 12-month horizon. The combined forecasts now perform better than individual models at all horizons except for $h = 1$ where the end-of-month RW still dominates. Applying the same selection criterion but on a rolling basis leads to even more impressive improvements in MSPE ratios with the lowest value being a staggering 0.46 (Column 7). Thus, narrowing the evaluation of losses to a short window increases the power of the MCS procedure to make a precise selection of the relevant models whose forecasts when pooled achieve the largest MSPE reductions overall. Figure 1 includes the cumulative MSPE ratios for the MCS-based forecast combinations, highlighting that being increasingly selective yields more effective pooling not just on average but over the entire evaluation period.[18] These results are based on weighing all models equally when aggregating their

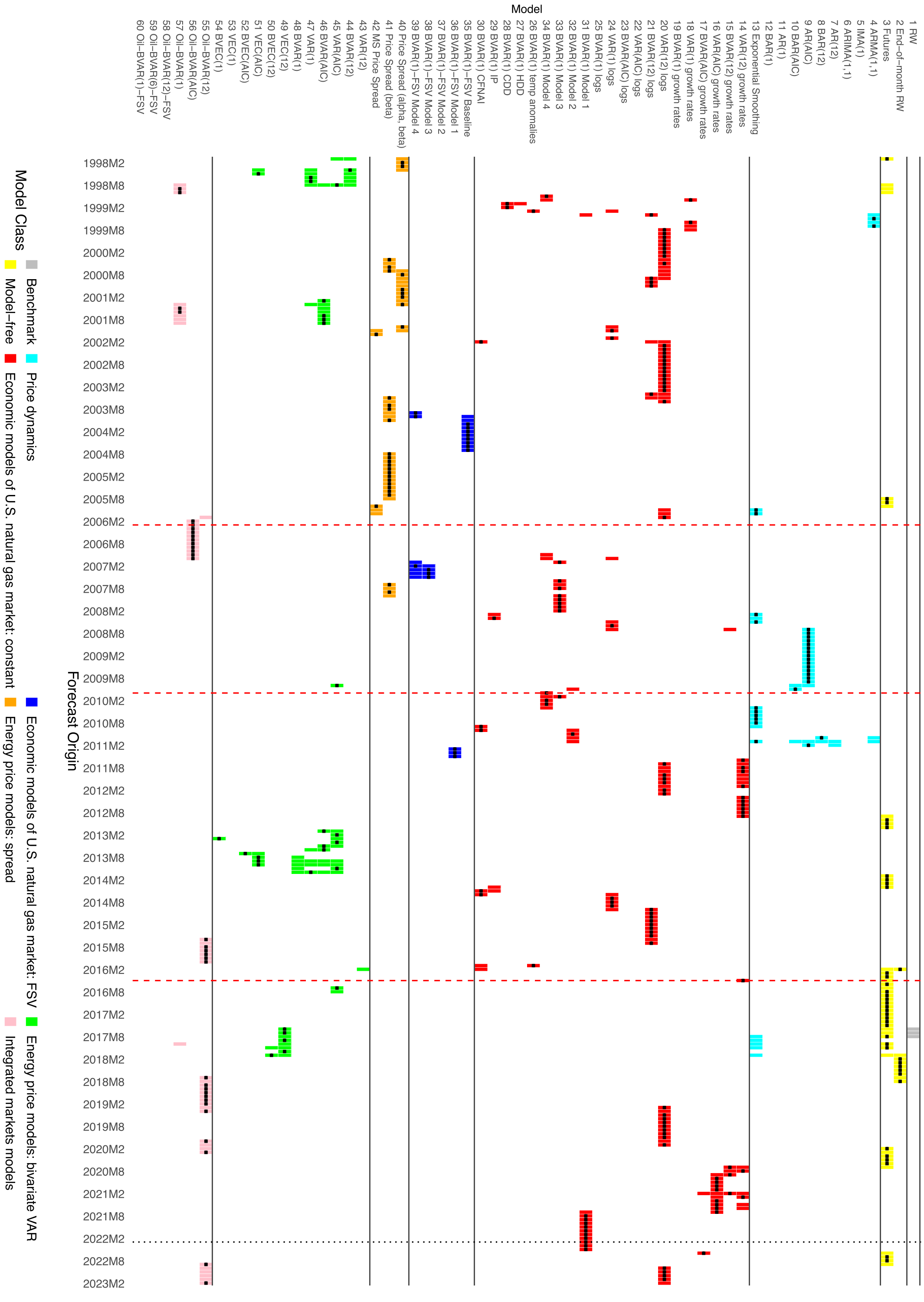

FIGURE 2 | Dot plot of real-time dynamic MCS model selection for $h = 12$. *Note:* The black dots indicate the model with a p-value of 1, while the other entries pertain to the models with the second highest p-values. The red dashed lines refer to the subperiods of Stock and Zaragoza-Watkins (2024). The black dotted line marks the beginning of the last 12 months.

forecasts. As can be seen from Columns 2, 4, 6, and 8, using inverse MSPE weights instead makes no material difference for the accuracy of pooled forecasts; all that matters is how the models for the forecast combination are chosen.

Figure 2 reports the models that rank highest at each forecast origin. For illustration, we focus on $h = 12$. Further evidence and discussion for other horizons is in the Supporting Information Appendix D. The spread price models are frequently selected in the first subperiod when natural gas and oil prices co-move closely but disappear after the decoupling of energy prices. Bivariate VARs, which belong to the same model class, also occasionally enter the pool, especially in later periods, hinting at price recoupling. Constant-volatility economic models of the natural gas market are consistently part of the forecast combinations with the VAR(12) based on fundamentals in logs being the dominant choice. Hybrid oil price-economic models are selected for consecutive periods at the onset of the transition to shale gas and appear again in the model set toward the end of the "shut-in fracking" period and regularly thereafter. While futures do not play any role in the first half of the evaluation period, they gain prominence in the last subperiod. Economic models with FSV show up a few times early on but then lose their informational edge and never return. Thus, the MCS flexibly adapts to changes in the predictive content across models in line with the historical narrative.

In sum, this analysis provides practitioners with two options. One is to update all the models each month to select the ones with the highest MCS *p*-values for the purpose of pooling, which delivers the best forecast of the real price of natural gas throughout. Alternatively, to reduce the burden of regular updates, forecasters could rely on the models selected in the recent past and only recalibrate the mix of models once a year. For more practical guidance, see Supporting Information Appendix D.

5 | Conclusions

Accurate forecasts of the natural gas price inform the design of a range of policies on climate change, energy affordability, incentives for fuel shifting, and the use of public lands for mineral exploration and extraction. In this paper, we evaluated the usefulness of a variety of candidate models that find their origin in economic theory, cross-market dynamics, and long-run relationships, in terms of their out-of-sample forecasting performance for the monthly real price of natural gas for horizons up to 24 months. The assessment was conducted in a real-time setting that accounted for delays in the release of predictor variables and subsequent revisions. For this purpose, we compiled a rich database of key determinants of the real price of natural gas from multiple sources. It consists of vintages from 1991M1 to 2024M2, each covering data back to 1973M1, that report only the information that was available to a real-life forecaster at the time the forecasts were generated.

Our analysis offers several key takeaways. First, at short horizons, model-free forecasts based on futures prices and last-day-of-the-month prices alongside forecasts from economic models of the natural gas market with a selective choice of fundamental drivers perform best. Second, at longer horizons, forecasts derived from exponential smoothing, futures prices, and various economic models with and without SV display the most success. Third, given that no single model wins across all horizons and that the relative forecasting performance of individual models changes over time, we propose a real-time MCS-based model selection criterion and show that pooled forecasts from individual models that are ranked highest by this criterion that flexibly adapts to each models' most recent performance achieve the largest reductions in MSPE and thus promise to deliver reliable forecasts going forward. Alternatively, to keep regular forecast updates feasible, the set of recently selected models can be used for pooling with revisions to the model space only once a year.

The real-time database together with the large universe of economically-motivated forecasting models offers promising avenues for future research. First, our analysis has focused on point forecasts because a systematic forecast evaluation for natural gas prices was missing in the literature; it would be straightforward to investigate the usefulness of these models for real-time density forecasting which could be used to quantify uncertainty, study tail behavior, construct risk measures and other real-time market monitoring tools. Second, while we have for the first time relied on MCS for real-time dynamic model selection, an interesting direction is to explore alternative methods for forecast aggregation such as Bayesian model averaging, prediction pools, Bayesian predictive synthesis, among others. Third, since our forecasts are targeted at informing economic decisions such as the residential energy mix, energy-saving investments, and government policies, the appropriate forecast frequency is monthly at which predictors of market fundamentals are available. Developing forecasting approaches for daily and weekly frequencies that could be put to use by financial investors, traders, and market analysts would be useful, but they require a different set of models, predictor variables, and evaluation criteria and therefore is left for future work.

Acknowledgments

We thank the editor, Mike McCracken, and two anonymous referees as well as Joe Bozeman, Pedro Gete, Jim Hamilton, and seminar participants at the EIA and the Central Bank of Colombia for helpful suggestions. Ray Bae and Pascal Frank provided excellent research assistance. The work of Christiane Baumeister was supported by grant DE-EI0003268 from the Department of Energy. Francesco Ravazzolo acknowledges financial support from the NextGenerationEU program (Piano Nazionale di Ripresa e Resilienza (PNRR) Missione 4 Componente 2, Investimento 1.5 D.D. 1058 23/06/2022, ECS-00000043). The views expressed in this paper are solely the responsibility of the authors and should not be interpreted as reflecting the views of the US Energy Information Administration.

Data Availability Statement

The data that support the findings of this study are openly available in the Journal Data Archive at https://doi.org/10.15456/jae.2025266.1900967125.

Endnotes

[1] For example, in 2005, only 15% of the natural gas purchased in Europe on wholesale markets was determined by spot prices according to the International Gas Union. While this share increased to 64% by 2015 and to 81% by 2020, oil-indexed contracts are still a distinctive feature of the European gas market. By contrast, 99% of natural gas in the United States is sold at spot prices.

[2] The wellhead price is quoted in dollars per thousand cubic feet, whereas the Henry Hub price is quoted in dollars per million Btu. We convert the wellhead price to dollars per million Btu by dividing it by 1.038.

[3] These real-time data are obtained from *Economic Indicators* published by the Council of Economic Advisers and made available in the FRASER database of the Federal Reserve Bank of St. Louis.

[4] We use the monthly average of daily prices for natural gas futures obtained from Bloomberg.

[5] While we only use vintages from January 1997 onward, the entire dataset can be found in the JAE Data Archive and on Christiane Baumeister's webpage. The date of January 1991 for the first vintage is a natural starting point since the publication of the *Monthly Energy Review* (*MER*) was temporarily suspended between October and December 1990. The vintages up to December 2023 were

hand-collected from electronic copies of historical issues of the *MER* available on the EIA's website. In January 2024, the EIA started to release real-time vintages in electronic format for all the data contained in the *MER*. Thus, our historical real-time database for variables related to the US natural gas market can now be easily extended.

[6] Table 3A in the Supporting Information Appendix shows that this finding is not due to the inclusion of the COVID-19 observations but is a genuine pattern of IP. In fact, the MSPE ratios are robust to the pandemic period.

[7] While the authors set the lag length to four months based on the AIC, we stick with one lag since we have already shown that for forecasting natural gas prices parsimony is key. Selecting the lag length recursively with the AIC does not yield any systematic improvements in accuracy (see Table 4A in the Supporting Information Appendix).

[8] We select m according to the Ledermann bound which determines the largest number of factors that implies a unique decomposition of the covariance matrix $\boldsymbol{\Sigma}_t$ and is derived by solving $(n-m)^2 \geq n+m$.

[9] In the Supporting Information Appendix, we compare BVAR models with homoskedastic and heteroskedastic factor structure on the covariance matrix in Table 5A, with different priors on the reduced-form dynamics in Table 6A, and with FSV and traditional SV in Table 7A. Our results are robust to these changes.

[10] Monthly data for the WTI spot oil price are taken from FRED (WTISPLC). We convert the oil price which is quoted in dollars per barrel to dollars per million Btu by dividing it by 5.8 such that it is in the same units as the Henry Hub spot price.

[11] We estimate the spread model starting in 1993.1 because the Federal Energy Regulatory Commission Order 636 issued in 1992 marks the completion of the deregulation process of wellhead prices (Joskow 2013).

[12] We show that this finding is robust to different prior choices and modeling of SV in Table 9A.

[13] Upon the request of two referees, we complement our analysis with a systematic evaluation of the predictive ability of our proposed set of models for the nominal price of natural gas in the Supporting Information Appendix B.

[14] Following Hansen et al. (2011), we compute the MCS p-values with 10,000 block bootstrap replications using a block size of 12.

[15] Additional results for other forecast horizons can be found in Figure 3A in the Supporting Information Appendix.

[16] The highest p-value is always 1, whereas the second highest p-value can either be above or below the threshold for inclusion in the MCS. Thus, the minimum is one selected model.

[17] For example, at $h=3$, the minimum number of models included is 4, the maximum is 60, and the average is 57, as opposed to a minimum of 1, a maximum of 46, and an average of 35 selected models at $h=24$.

[18] See also Figures 3A and 4A for other forecast horizons that tell the same story.

References

Arora, V., and J. Lieskovsky. 2014. "Natural Gas and U.S. Economic Activity." *Energy Journal* 35: 167–182.

Bachmeier, L. J., and J. M. Griffin. 2006. "Testing for Market Integration: Crude Oil, Coal, and Natural Gas." *Energy Journal* 27: 55–71.

Baumeister, C. 2023. "Measuring Market Expectations." In *Handbook of Economic Expectations*, edited by R. Bachmann, G. Topa, and W. van der Klaauw, 413–442. Academic Press.

Baumeister, C., F. Huber, and M. Marcellino. 2024. Risky Oil: It's All in the Tails, NBER WP 32524.

Baumeister, C., and L. Kilian. 2012. "Real-Time Forecasts of the Real Price of Oil." *Journal of Business and Economic Statistics* 30, no. 2: 326–336.

Baumeister, C., and L. Kilian. 2014. "What Central Bankers Need to Know About Forecasting Oil Prices." *International Economic Review* 55, no. 3: 869–889.

Baumeister, C., L. Kilian, and T. K. Lee. 2017. "Inside the Crystal Ball: New Approaches to Predicting the Gasoline Price at the Pump." *Journal of Applied Econometrics* 32, no. 2: 275–295.

Baumeister, C., D. Korobilis, and T. K. Lee. 2022. "Energy Markets and Global Economic Conditions." *Review of Economics and Statistics* 104, no. 4: 828–844.

Brown, S. P. A., and M. K. Yücel. 2008. "What Drives Natural Gas Prices?" *Energy Journal* 29, no. 2: 45–60.

Carriero, A., T. E. Clark, and M. Marcellino. 2019. "Large Vector Autoregressions with Stochastic Volatility and Non-Conjugate Priors." *Journal of Econometrics* 212, no. 1: 137–154.

Chan, J. C. C., and A. L. Grant. 2016. "Modeling Energy Price Dynamics: GARCH versus Stochastic Volatility." *Energy Economics* 54: 182–189.

Clark, T. E., and F. Ravazzolo. 2015. "Macroeconomic Forecasting Performance Under Alternative Specifications of Time-Varying Volatility." *Journal of Applied Econometrics* 30, no. 4: 551–575.

Diebold, F. X., and R. S. Mariano. 1995. "Comparing Predictive Accuracy." *Journal of Business and Economic Statistics* 13, no. 3: 253–263.

Doan, T., R. B. Litterman, and C. A. Sims. 1984. "Forecasting and Conditional Projection Using Realistic Prior Distributions." *Econometric Reviews* 3, no. 1: 1–100.

Ellwanger, R., and S. Snudden. 2023. "Forecasts of the Real Price of Oil Revisited: Do They Beat the Random Walk?" *Journal of Banking and Finance* 154: 106962.

Faust, J., and J. H. Wright. 2013. "Forecasting Inflation." In *Handbook of Economic Forecasting, 2A*, edited by G. Elliott and A. Timmermann, 2–56. North-Holland.

Ferrari, D., F. Ravazzolo, and J. Vespignani. 2021. "Forecasting Energy Commodity Prices: A Large Global Dataset Sparse Approach." *Energy Economics* 98: 105268.

Gao, S., C. Hou, and B. H. Nguyen. 2021. "Forecasting Natural Gas Prices Using Highly Flexible Time-Varying Parameter Models." *Economic Modelling* 105: 105652.

Giannone, D., M. Lenza, and G. Primiceri. 2015. "Prior Selection for Vector Autoregressions." *Review of Economics and Statistics* 97, no. 2: 436–451.

Granger, C. W. J., and Y. Jeon. 2004. "Thick Modeling." *Economic Modelling* 21: 323–343.

Halser, C., F. Paraschiv, and M. Russo. 2023. "Oil-Gas Price Relationships on Three Continents: Disruptions and Equilibria." *Journal of Commodity Markets* 31: 100347.

Hamilton, J. D. 2009. "Understanding Crude Oil Prices." *Energy Journal* 30, no. 2: 179–206.

Hansen, P. R., A. Lunde, and J. M. Nason. 2011. "The Model Confidence Set." *Econometrica* 79, no. 2: 453–497.

Hartley, P. R., K. B. Medlock, and J. E. Rosthal. 2008. "The Relationship of Natural Gas to Oil Prices." *Energy Journal* 29, no. 3: 47–66.

Huber, F., and M. Feldkircher. 2019. "Adaptive Shrinkage in Bayesian Vector Autoregressive Models." *Journal of Business and Economic Statistics* 37: 27–39.

Hupka, Y., I. Popova, B. Simkins, and T. K. Lee. 2023. "A Review of the Literature on LNG: Hubs Development, Market Integration, and Price Discovery." *Journal of Commodity Markets* 31: 100349.

Joskow, P. L. 2013. "Natural Gas: From Shortages to Abundance in the United States." *American Economic Review* 103, no. 3: 338–343.

Kastner, G., and F. Huber. 2020. “Sparse Bayesian Vector Autoregressions in Huge Dimensions.” *Journal of Forecasting* 39, no. 7: 1142–1165.

Marcellino, M., H. Stock, and M. W. Watson. 2006. “A Comparison of Direct and Iterated Multistep AR methods for Forecasting Macroeconomic Time Series.” *Journal of Econometrics* 135, no. 1–2: 499–526.

Pitt, M. K., and N. Shephard. 1999. “Time-Varying Covariances: A Factor Stochastic Volatility Approach.” In *Bayesian Statistics–Proceedings of the Sixth Valencia International Meeting*, edited by J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 547–570. Oxford University Press.

Primiceri, G. 2005. “Time Varying Structural Vector Autoregressions and Monetary Policy.” *Review of Economic Studies* 72, no. 3: 821–852.

Ramberg, D. J., and J. E. Parsons. 2012. “The Weak Tie Between Natural Gas and Oil Prices.” *Energy Journal* 33, no. 2: 13–35.

Stock, J. H., and M. Zaragoza-Watkins. 2024. The Market and Climate Implications of U.S. LNG Exports, NBER WP 32228.

Timmermann, A. 2006. “Forecast Combinations.” In *Handbook of Economic Forecasting, 1*, edited by G. Elliott, C. W. J. Granger, and A. Timmermann, 135–196. North-Holland.

Wiggins, S., and X. L. Etienne. 2017. “Turbulent Times: Uncovering the Origins of US Natural Gas Price Fluctuations Since Deregulation.” *Energy Economics* 64: 196–205.

Winkler, V. 2023. What Drives Drilling Up and Prices Down? IMK WP 217.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. Online Appendix.