School of Economics and Management

Price forecasting for Henry Hub natural gas market using statistical and machine learning methods and their combinations

Master Thesis Finance

Author: Kristóf Bartha

SNR: 2147136

ANR: 277551

Supervisor: prof. dr. Lieven Baele

Second reader: Gazi Kabaş

Date: 26 June 2025

Abstract

In recent decades, the significance of natural gas forecasting has grown within the domain of energy economics, gaining particular attention in the wake of recent geopolitical developments. As a critical component of energy supply and economic progression, natural gas prices are influenced by a diverse range of factors. Although various studies have examined this topic, notable gaps persist – especially concerning the integration of multiple economic and energy-related indicators and the comparative performance of different modeling methodologies. This thesis evaluates eleven predictors to enhance forecasting accuracy and introduces five modeling approaches: one traditional econometric model (SARIMAX), two machine learning models (Random Forest and LSTM), and two hybrid combinations (SARIMAX-RF and SARIMAX-LSTM). For the machine learning models, forecasting is conducted using rolling windows of 12, 24, and 36 months. In the hybrid frameworks, the SARIMAX model is first applied to extract residuals, which are then used to train the machine learning models to capture non-linear patterns. Model performance is assessed using four evaluation metrics. The standalone LSTM model consistently achieved the highest accuracy across all measures, surpassing both SARIMAX and Random Forest. However, neither hybrid model demonstrates improvements over the standalone LSTM. This forecasting framework contributes to a deeper understanding of the natural gas market and lays the groundwork for future investigations.

Table of Contents

List of Tables

List of Figures

1. Introduction

1.1 Motivation

During my master's program, a variety of compelling research topics were introduced, several of which captured my interest. Ultimately, I chose to pursue a self-developed topic, influenced not only by my prior experience in the energy sector but also by a strong desire to apply the theoretical and technical skills – particularly in programming and statistical analysis using R – that I acquired during my studies. My undergraduate involvement in the energy industry also reinforced this interest, motivating me to select a thesis topic that integrates my academic knowledge with practical experience.

The specific research question emerged during the proposal phase of the thesis process. In our coursework, we frequently applied various forecasting models to price data, prompting me to investigate how these models perform in the volatile environment of today's energy markets. My goal is to assess their accuracy and determine which model offers the most robust predictions.

"American gas is 3-4 times cheaper on the domestic market than the price at which they offer it to Europeans. These are double standards."

– stated by French President Emmanuel Macron after an EU summit ("Macron Calls out US for Double Standards on Gas Prices to EU," 2022). This sentiment reflects growing European concerns about energy supply security, especially given the U.S. market's rising importance as an alternative to Russian gas.

Thus, I focus on the American natural gas market due to its global relevance and the availability of high-quality data. While the academic literature on energy markets is extensive, most studies prioritize oil, relegating gas to a secondary role. However, given the geological link between oil and gas – often co-located during extraction – and recent geopolitical developments, gas has assumed a more prominent role, particularly in Europe following the Russia-Ukraine war. This shift has elevated the importance of indices like the Dutch TTF and increased European dependence on U.S. LNG as a strategic alternative.

Natural gas also plays a crucial role in energy transition efforts aimed at reducing emissions through the so-called "oil-to-gas switch" – a trend that has waned recently due to soaring gas prices. As of this writing, the price of Dutch TTF is more than three-times higher than U.S. Henry Hub, highlighting Europe's competitive disadvantage and dependence on American exports. Extreme volatility – that was experienced in the recent years, sending TTF prices above 300 euros – also increases the unpredictability of prices, leading this commodity to the focus of the society and policymakers. Thus, accurate forecasting of Henry Hub prices is essential for European industry leaders and policymakers. Given this context of geopolitical instability and market volatility, my thesis aims to identify the most effective forecasting model for the U.S. natural gas benchmark.

1.2 Research Focus and Questions

The main research question of this thesis is whether hybrid models – those combining classical statistical techniques with machine learning algorithms – can enhance the accuracy of natural gas price forecasts. Specifically, the thesis examines one traditional model (SARIMAX) and two machine learning approaches (Random Forest and LSTM), followed by hybrid versions of SARIMAX with each of the latter.

The objective is to determine which model yields the most accurate forecasts, evaluated using RMSE, MAE, SMAPE, and DA. Although the main focus lies in testing the efficiency of hybrid models, individual models are first assessed independently. This leads to two subquestions. First, whether Random Forest can outperform SARIMAX, and second, whether LSTM independently delivers the most precise forecasts. Based on a review of the literature and the theoretical design of hybrid models – which aim to capture both linear and non-linear patterns – I assume that machine learning models will outperform traditional models, and their combination will further enhance predictive accuracy.

1.3 Data Resources and Methodology

Natural gas prices are influenced by numerous factors, many of which – such as political statements or investor sentiment – are difficult to quantify, but can fundamentally shake the stock exchanges, causing above-average volatility and market shocks. While these elements introduce volatility, they fall outside the scope of this research. Instead, this study focuses on forecasting accuracy using observable and accessible variables, acknowledging that uncertainty will persist.

Data is sourced primarily from the U.S. Energy Information Administration (EIA), including consumption (residential and industrial), storage levels, dry production (which represents the amount of gas intended for end users), export-import balances (including LNG and pipeline flows), rig counts (which indicates the number of active gas wells in a given period), and front-month gas and oil prices (West Texas Intermediate, or WTI). Additional variables include the

Federal Reserve's interest rate and weather indicators such as heating and cooling degree days from NOAA, which indicate temperature changes and thus explicit measures of heating and cooling demand by citizens. To capture for geopolitical risks and uncertainty, the Global Political Risk (GPR) index by Caldara and Iacoviello (2022) is also included, downloaded from the publishers' website. A dummy variable is added for hurricane events affecting Louisiana – the physical location of Henry Hub – using Wikipedia-sourced data via web scraping. Natural disasters are common in the Gulf of Mexico area during the late summer and early autumn periods and can also strongly influence gas prices. All data series are initially downloaded on a monthly basis, but where only daily data is available, the monthly data is calculated as the average of the daily values.

The methodology unfolds in two main phases. First, I estimate forecasts using SARIMAX, Random Forest, and LSTM models, applying four evaluation metrics. Second, I construct hybrid models by using SARIMAX residuals as inputs to Random Forest and LSTM algorithms. Each model is tested using rolling windows of 12, 24, and 36 months, indicating 1-year, 2-year and 3-year periods. To validate time-series assumptions, I apply the Ljung-Box, ADF, and KPSS tests. Data is split into 70% training and 30% testing sets, with out-of-sample forecasts produced based on learned patterns.

1.4 Results

The comparative performance metrics for all models are summarized in Table 1. Among the Random Forest configurations, the 36-month lagged model, RF(36) demonstrates the best performance, outperforming SARIMAX in terms of magnitude-based error metrics – reducing RMSE by 11%, MAE by 9%, and lowering SMAPE by 23 percentage points. However, the LSTM models exhibit clear superiority across all metrics. The LSTM(12) model achieves a 67% reduction in both RMSE and MAE compared to RF(36), and a 71% reduction relative to SARIMAX. It also attains SMAPE values around 12%, which corresponds to approximately one-third of the value observed in RF(36) and one-fifth of that in SARIMAX. Directional Accuracy (DA) further highlights LSTM's robustness in capturing price movements, including abrupt increases and declines.

For the SARIMAX-RF combination, performance deteriorates significantly – RMSE and MAE increased by 14–37% in the best-case scenario (lag 24), compared to standalone SARIMAX and RF(36). Additionally, SMAPE rises to a range of 66–75%, with the lowest SMAPE still double that of RF(36) and 10 percentage points higher than SARIMAX. DA is dropped to around 4% or lags 12 and 36, with only lag 24 marginally surpassing the standalone models.

The SARIMAX-LSTM hybrid performs even more poorly. RMSE and MAE values are several times higher than those of the standalone LSTM, and around 50% worse than SARIMAX. SMAPE exceeds 100%, and DA falls below 10%, indicating an inability of the model to capture any meaningful patterns. In conclusion, the standalone LSTM model, the most complex of those examined, decisively outperforms all other models – both individually and in combination. The most accurate forecasts are achieved when the model utilized the recent 12 months of data to predict the next value.

	Model	RMSE	MAE	SMAPE (%)	DA (%)
1	SARIMAX	2.0960	1.4582	56.3758	9.3023
2	RF(12)	1.9096	1.5413	43.1809	2.4096
3	RF(24)	2.5159	2.0265	50.3773	5.0633
4	RF(36)	1.8559	1.3270	33.4656	7.8947
5	LSTM(12)	0.6097	0.4291	11.8458	24.6575
6	LSTM(24)	0.6331	0.4359	11.8425	18.0328
7	LSTM(36)	0.7064	0.4963	12.3208	16.3265
8	SAR-RF(12)	2.6072	1.8859	75.9386	4.0541
9	SAR-RF(24)	2.3976	1.8255	71.3846	9.6774
10	SAR-RF(36)	2.6433	2.0346	66.6249	4.0000
11	SAR-LSTM(12)	3.4132	2.4466	102.3196	5.4054
12	SAR-LSTM(24)	3.2462	2.4512	103.4218	1.6129
13	SAR-LSTM(36)	3.1925	2.4145	88.9851	8.0000

Table 1: Final results

1.5 Literature Review

Ślepaczuk and Kashif explored the effectiveness of hybrid LSTM-ARIMA models across three major stock indices – S&P 500, CAC 40, and FTSE 100 – spanning nearly 25 years. Their research also incorporated Random Forest models, both independently and in combination with ARIMA. Unlike the approach in this thesis, their ARIMA models did not utilize exogenous variables. Additionally, their evaluation focused on investment strategy outcomes such as return and volatility, rather than statistical forecast accuracy indicators like RMSE or MAE (Kashif & Ślepaczuk, 2025).

Their findings were partially aligned with this thesis. Both studies confirmed the superiority of LSTM over Random Forest and ARIMA when used in isolation. However, while Ślepaczuk and Kashif found the LSTM-ARIMA hybrid often outperformed all other models, this thesis finds contrary results: the hybrid combinations underperformed in the examined time frame and market context (Kashif & Ślepaczuk, 2025).

Ray and colleagues investigated hybrid LSTM-ARIMA models using three agricultural commodity indices over a relatively short time span (324 data points, comparable to the 288 used here). They excluded standalone Random Forest models but did assess ARIMA-LSTM hybrids. As in this thesis, they observed that LSTM outperformed ARIMA in isolation. However, in contrast, they concluded that the hybrid model yielded the best predictive accuracy across all cases (Ray et al., 2023).

1.6 Conclusions

This thesis evaluates both standalone and hybrid models to forecast natural gas prices, testing three key hypotheses. The first hypothesis – that Random Forest would consistently outperform SARIMAX – is rejected, as Random Forest do not demonstrate superior accuracy in all cases. The second hypothesis, which proposed that LSTM would yield the most accurate forecasts across all evaluation metrics and time frames, is approved. LSTM consistently outperforms both Random Forest and SARIMAX models. The third hypothesis, anticipating that hybrid models would outperform their individual counterparts, is also rejected. In fact, the combined models produce inferior results compared to the standalone LSTM.

Overall, under the specific conditions of this study – monthly data, selected exogenous variables, and the defined time period – the LSTM model on its own emerges as the most effective forecasting tool for predicting Henry Hub front-month gas prices. These findings suggest that further refinement, such as hyperparameter optimization, alternative explanatory variables, or extended time frames, could potentially enhance the performance of future models.

1.7 Roadmap

The rest of this thesis is structured as follows: Section 2 presents a comprehensive review of the most pertinent literature, offering a broader view of commodity markets. Section 3 details the data, including explanations of each variable and the preprocessing steps undertaken. In Section 4, the modeling methodology is explained, including descriptions of the forecasting models and performance evaluation metrics. Section 5 outlines the hypotheses. Sections 6 and 7 present the empirical results and conclusions, respectively, based on the findings of the study. Sections 8 highlights the use of AI tools in this thesis.

2. Literature review

Numerous studies have addressed the forecasting of commodity prices, with a predominant focus on the oil market, which remains the world's most significant energy commodity. In contrast, research specifically dedicated to the natural gas market remains relatively limited, despite the increasing strategic importance of gas amid recent global conflicts.

Initially, this review surveyed the academic literature to identify the principal determinants of natural gas price fluctuations. Serletis and Herbert (1999) noted a fundamental transformation in the natural gas industry, shifting from a tightly regulated regime to a market-driven structure. This shift implies a growing responsiveness of gas prices to external market forces such as production volumes and storage levels. Several studies have underscored the relationship between oil and gas prices. For instance, Brown and Yücel (2008) and Panagiotidis and Rutledge (2007) both provided empirical evidence of a strong correlation between these two commodities.

However, the strength of this linkage has diminished over time. Erdős (2012) found that the U.S. natural gas market had largely decoupled from crude oil prices, in contrast to the UK market, where the connection remained strong. Similarly, Brigida (2014) reported a weakening relationship between oil and gas, providing further justification for treating them as partially independent markets in empirical models. In light of these findings, this thesis incorporates several exogenous variables beyond oil prices to capture the unique dynamics of the gas market.

Weather-related indicators, such as Heating Degree Days (HDD) and Cooling Degree Days (CDD), are commonly employed in gas price modeling to reflect seasonal variations in energy demand. Prior studies have shown that these factors, alongside production and storage levels, significantly influence gas prices (Geng et al., 2016; Ji et al., 2018; Mu, 2007). In addition, Roberts (2019) emphasized the importance of accounting for extreme weather events, such as hurricanes in the Gulf of Mexico, which can severely disrupt gas production and supply chains. His findings supported the inclusion of a control variable for hurricanes in forecasting models.

From a methodological perspective, Armstrong (2001) analyzed a wide range of forecasting models, concluding that no single model consistently outperforms others in all contexts. He argued that the effectiveness of a given method often depends on the specific structure and dynamics of the dataset. The SARIMA framework, and its variant SARIMAX (which includes exogenous variables), is one of the most used models for time-series forecasting due to its ability to capture seasonality and linear patterns in economic and energy-related data. Numerous studies have applied SARIMA-type models to energy market forecasts (Jia & Yan, 2023; McHugh et al., 2019; Sheng & Jia, 2020). However, while SARIMAX is adept at modeling linear components, it lacks the capacity to effectively capture complex nonlinear dependencies – thus making it suitable for combination with machine learning approaches.

Random Forest (RF) and Long Short-Term Memory (LSTM) neural networks have emerged as powerful alternatives in the machine learning domain. Siami-Namini et al. (2018) conducted a comparative study on 11 global stock indices and found that LSTM reduced RMSE by 82–87% compared to ARIMA models, highlighting its potential for superior forecasting accuracy.

In contrast, Sarangi et al. (2025) applied both Random Forest and LSTM models to forecast stock prices for two Indian firms – one of which operated in the natural gas sector. Using an 80/20 training-testing split, they observed that Random Forest outperformed LSTM in terms of RMSE, MAE, and MAPE, suggesting that model performance is context-dependent and not universally predictable.

Further support for the robustness of Random Forest was provided by Herrera et al. (2019), who examined six commodities – including three oil types, two gas indices, and coal – using RF, Artificial Neural Networks (ANN), and a hybrid econometric model. Their results indicated that RF consistently achieved the lowest RMSE and MAPE across all cases, outperforming both ANN and hybrid approaches.

Xu et al. (2025) applied a diverse range of individual and hybrid models to forecast crude oil prices. While ARIMAX outperformed simple LSTM in their study, the introduction of a hybrid LSTM model substantially improved forecasting accuracy, as evidenced by significant reductions in MAE and MAPE values. This finding emphasized the strength of hybrid configurations in capturing both linear and nonlinear patterns.

Seba and Belaide (2024) conducted a study forecasting COVID-19 fatality rates using various models, including ARIMA, Random Forest. They demonstrated that the RF-ARIMA hybrid yielded the most accurate predictions over two different forecast horizons (60 and 120 days), as reflected by the lowest RMSE, MSE, and MAE values.

Regarding hybrid models, several scholars have investigated combinations of Random Forest or LSTM with ARIMA or SARIMAX. Kashif and Ślepaczuk (2025) applied LSTM-ARIMA and RF-ARIMA models to forecast three major stock indices (S&P 500, CAC 40, FTSE 100) across a 25-year period. Their evaluation focused on investment strategies – specifically long-short and short-long positions – measuring returns and volatility using metrics such as the Maximum Drawdown (MD) and Information Ratio (IR). The standalone LSTM model consistently outperformed ARIMA in most cases, a result attributed to its architectural complexity. Considering the combined models, the dominance of LSTM-ARIMA is unquestionable: the LSTM-ARIMA hybrid model generally delivered the highest accuracy, although the RF-ARIMA variant occasionally outperformed it in specific scenarios.

Ray et al. (2023) examined agricultural price indices using a hybrid LSTM-ARIMA model and also included an ARIMA-GARCH framework in their study. The researchers excluded the Random Forest model on its own from the analysis but considered it in the hybrid LSTM-ARIMA, in which the number of lags was decided by this method. In case of simple models (when only LSTM and ARIMA were compared), the machine-learning model provided better results over the examined period. Their findings also revealed that the LSTM-ARIMA hybrid consistently achieved the lowest RMSE, MAPE, and MASE across all tested series, highlighting the effectiveness of combining neural networks with traditional time-series models.

In conclusion, while commodity price forecasting has received substantial academic attention, most studies prioritize crude oil, with natural gas often examined secondarily. Furthermore, LSTM-ARIMA hybrids dominate the literature, whereas empirical investigations involving RF-ARIMA hybrids and SARIMAX-based combinations remain scarce. The findings across these studies are often contradictory, suggesting that model effectiveness is highly dependent on the time frame, dataset characteristics, and specific commodity in question. In most cases, the dominance of combined models can be observed due to their complexity and better capturing performance. This thesis aims to contribute to the existing literature by offering a detailed evaluation of U.S. natural gas price forecasting using SARIMAX, Random Forest, and LSTM models, along with their hybrid forms.

3. Data

In this section, I present the data employed in the analysis. Prior to initiating the research, I reviewed existing literature and drew upon my prior experience to assemble a dataset composed of variables widely recognized as key indicators of natural gas prices. These variables encompass critical components of the gas market such as consumption, storage levels, and production, as well as external indicators like weather conditions and geopolitical risks. In total, 11 explanatory variables are included.

3.1 Variables

- **Price_fm** (USD/MMBtu): This dependent variable reflects the settlement price of Henry Hub natural gas contracts for delivery in the next month, traded on the New York Mercantile Exchange (NYMEX). Prices are expressed in USD per million British thermal units (MMBtu), differing from European standards, which use megawatt-hours (MWh). Contracts expire three business days before the first calendar day of the delivery month. Data is sourced from the EIA.
- **natgas_dryproduction** (million cubic feet): Represents the volume of dry natural gas available for end users. It is the total production (gross withdrawals) minus the natural gas used for repressuring the system and maintaining operations. In cases of extraction of unwanted gases, like alkane hydrocarbons (namely methane), they will be vented or flared at the spot (resulting in the total marketed production). Moreover, nonhydrocarbon gases (like water vapor, carbon dioxide) are removed from the gas stream, as well as the gas converted to liquid form like lease condensate and plant liquids – this is called extraction loss. The gas available for end users is the marketed production minus extraction loss. Data is sourced from the EIA.
- **workinggas_in_storage** (million cubic feet): Denotes the extractable portion of gas stored in underground facilities. There are theoretically two types of gas in storage – cushion gas and working gas – though they have the same characteristics. Cushion gas is the volume needed to maintain adequate reservoir pressure and cannot be withdrawn, while working gas is the natural gas available for withdrawal, hence the portion that can be used by consumers. Cushion gas is indispensable for maintaining reservoir operations, as only working gas is considered in the model. Data is sourced from the EIA.
- **total_consumption** (million cubic feet): Aggregated monthly consumption from residential, industrial, and commercial sectors. Residential usage covers household heating and cooking; industrial includes energy use in manufacturing and related fields; and commercial comprises non-manufacturing entities like businesses and public buildings. For the model, the sum of the above three is taken into account. Data is sourced from the EIA.
- **rigs_count** (number): Reflects the number of active natural gas drilling rigs across the U.S., excluding oil-specific rigs, even though gas and oil often co-occur. Oil extraction is usually followed by substantial amounts of so-called associated petroleum gas (APG),

while gas extraction is followed by condensate oil. For simplicity, only active gas rigs are considered, and the APG amount is omitted from the data. Data is sourced from the EIA.

- **Interest_rate** (%): The Federal Reserve's target federal funds rate, used as the benchmark interest rate in the U.S. financial system. Data is sourced from FRED.
- **Oil_futures** (USD/barrel): The front-month futures prices for West Texas Intermediate (WTI) crude oil traded on NYMEX. WTI is produced in Texas and southern Oklahoma, serving as a reference for pricing a number of other crude streams, and is traded at Cushing, Oklahoma. Please note that this price refers to crude oil, which is a mixture of hydrocarbons that exists in liquid phase in natural underground reservoirs and remains liquid at atmospheric pressure after passing through surface separating facilities. Lease condensate recovered as a liquid from natural gas wells in lease or field separation facilities and later mixed into the crude stream is also included. Liquids produced at natural gas processing plants are excluded from crude oil withdrawals. Data is sourced from the EIA.
- **HDD and CDD** (number of days): Degree days indicate how cold or hot a location is compared to 65°F, which is a standard value across the U.S. The more extreme the outside temperature, the higher the number of degree days. Heating Degree Days (HDD) measure how much (and for how long) outside temperatures are below 65°F, while Cooling Degree Days (CDD) measure how much outside temperatures exceed 65°F on average. For example, a day with a mean temperature of 30°F has 35 HDDs, while a day with a mean of 80°F has 15 CDDs. Data is provided by NOAA and sourced from EIA.
- **GPR** (index): This is the only variable not limited to the U.S. but global in scope. It captures worldwide geopolitical risk and is developed by researchers Dario Caldara and Matteo Iacoviello. *"Higher geopolitical risk foreshadows lower investment and employment and is associated with higher disaster probability and larger downside risks."* (Caldara & Iacoviello, 2022, pp. 1194). The thesis uses the Recent GPR Index, which is constructed from ten newspapers. The index is calculated by counting the number of articles related to adverse geopolitical events in each newspaper for each month (as a share of the total number of news articles). In the period 2001–2024, several spikes were examined, such as the 9/11 terrorist attack in 2001, the breakout of the Iraqi (2003) and Ukrainian (2022) wars, and the COVID-19 situation in 2020.

- **EXIM** (million cubic feet): Constructed from total U.S. natural gas exports and imports (including both pipeline and LNG). Due to its geographic position, the US focuses more on LNG exports. Based on data from the Global Energy Monitor, the US has the largest LNG export capacity in the world (Global Energy Monitor, 2024). According to EIA, the US remained the top LNG exporter in 2024, with a daily capacity of 11.9 billion cubic feet, ahead of Australia and Qatar (Zaretskaya, 2025). For Europe, US exports are especially significant, as the continent is the primary market for American LNG. Data is sourced from EIA.
- **hurricanes** (dummy): A binary variable capturing whether a hurricane hit Louisiana (Henry Hub's location) during a given month. Data is downloaded by web scraping from Wikipedia, taking into account the month in which the hurricane reached the state. If an event occurs in a given month, 1 is assigned and 0 otherwise. In total, 13 hurricanes hit the state between 2001 and 2024, and the strongest ones (Category 4), Ida and Laura, caused several problems and damages to power and gas plant sites, leading prices upward.

3.2 Data Preprocessing – Pearson correlation

Pairwise correlation (Table 2) analysis revealed a strong linear dependence (0.98) between the EXIM and production variables, indicating almost complete explanatory overlap, while the relationships between other variables remain in the moderate range. By addressing their collinearity, EXIM is regressed on production, and the residuals are used as a substitute variable. Since the coefficient of the exim variable was originally positive, and the regression between the two variables also yields a positive coefficient, excluding the exim variable results in a (small) positive bias in the production variable (value approximately 0.0000016).

After the substitution, correlation is checked again (Table 3). Surprisingly, the correlation matrix also showed a negative relationship (-0.39) between CDD and total consumption, while HDD showed a positive correlation (0.68). It means – in contrast of the economic explanation, stating higher degree days lead to higher demand – an increase in CDD will decrease the total consumption. One explanation is that cooling (air conditioning) demand is largely met by nuclear or renewable energy sources, rather than natural gas.

The value between CDD and HDD indicates a strong relationship (0.83), which is entirely understandable given the nature of the two variables: when the daily average temperature is 16 above the threshold (65°F), CDD is positive while HDD is zero, and vice versa below the threshold.

Another noteworthy result is the negative correlation (-0.82) between production and rigs, which contradicts economic logic, as one would expect a positive relationship, as the more rigs in operation the more gas is extracted. This anomaly may be explained by technological advancements in drilling, whereby fewer rigs can yield higher production volumes.

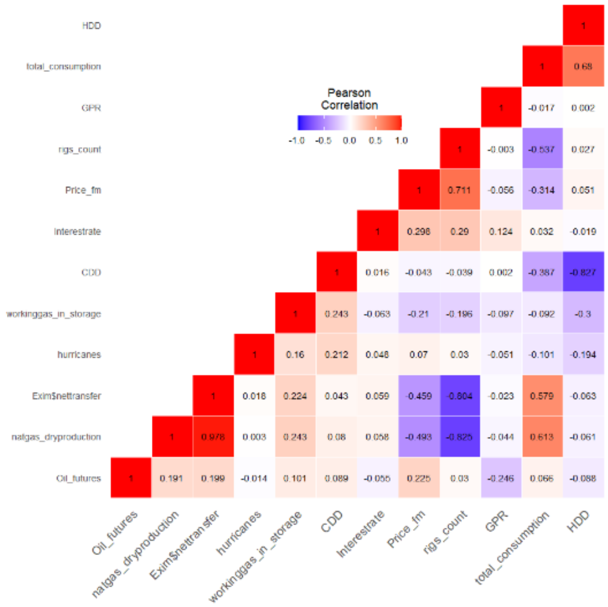


Table 2: Pearson correlation matrix with original variables

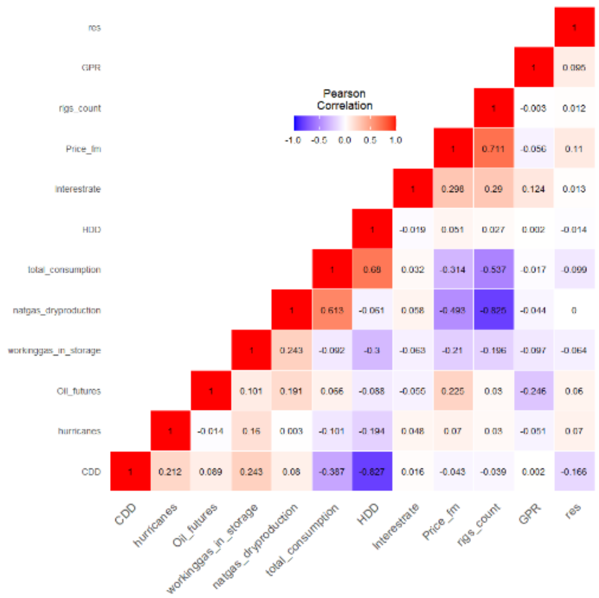


Table 3: Pearson correlation matrix after residuals are included

3.3 Data Preprocessing – sample selection and multicollinearity

As a next step, the dataset is split for later model validation: 70% of the data is used for training and the remaining 30% for testing, resulting in 202 datapoints for training and 86 for testing. Afterward, I observe stationarity of the target variable, as prices are typically not stable over time. While natural gas prices do not show the same continuous growth as stock prices, both the ADF and KPSS test indicate that prices are non-stationary. Since this is a required condition for both SARIMAX and LSTM models, first-order differencing is applied. Post-differencing tests now confirms stationarity, though autocorrelation function (ACF) plots still display some residual lags (Tables 7&8, please find them in the Appendix).

Considering the model, I cannot ignore multicollinearity when analysing the data (Table 4). Multicollinearity analysis identifies four variables with Variance Inflation Factors (VIFs) exceeding 10, confirming redundancy in the correlation matrix: HDD and CDD are highly correlated with each other, while production and consumption also have a strong mutual relationship.

	Multicollinearity
workinggas_in_storage	1.255
total_consumption	39.171
natgas_dryproduction	17.232
rigs_count	7.058
Interestrate	1.954
Oil_futures	1.443
HDD	52.59
CDD	13.016
GPR	1.115
hurricanes	1.103
res	1.162

Table 4: Multicollinearity in the model

Despite this, none of the variables are excluded due to their economic significance and the model's emphasis on explanatory power. As multicollinearity increases standard errors but does not introduce bias, retaining these variables is deemed appropriate. For example, high CDD or HDD values typically result in increased demand, which in turn leads to higher consumption and production. To sum up, despite VIFs are above 10 in these cases, variables are retained due to their theoretical importance.

Latest, additional tests are conducted to confirm seasonality, as SARIMAX places importance on it. Table 9 (can be found in the Appendix) summarizes all the variables, each plotted separately on its own scale before normalization. The charts support the test results: production, consumption, gas storage, imports, and both CDD and HDD exhibit seasonality – as this logically follows from their nature. However, prices do not exhibit seasonality – despite the fact there is a so-called summer-winter spread in the gas industry, referring price differences between the two seasons – supporting the choice to model them using non-seasonal techniques.

Table 5 represent a descriptive statistic of the variables before scaling. Jarque-Bera test for normality are also included. Please note that the respective critical values at 5% level are 5.991 for Jarque-Bera (with degrees of freedom equals to 2), -2.87 for ADF and 0.463 for KPSS. Most variables are non-normally distributed, and several (including price and consumption) exhibit unit-root behaviour.

	n	mean	sd	median	skew	kurtosis	se	JB_test	ADF_test	KPSS_test
Price_fm	288	4.41	2.22	3.78	1.45	2.4	0.13	173.69	-3.65	1.66
workinggas_in_storage	288	2621468.36	783447.63	2654397	-0.28	-0.77	46165.09	10.47	-5.94	0.81
total_consumption	288	2194922.39	498469.7	2150220.5	0.54	-0.37	29372.61	15.6	-12.73	3.15
natgas_dryproduction	288	2142425.95	562287	2020872.5	0.56	-1.07	33133.08	28.76	-2.64	4.61
rigs_count	288	598.44	475.04	424.5	0.5	-1.14	27.99	27.36	-1.81	3.77
Interestrate	288	1.74	1.86	1.09	0.87	-0.69	0.11	42.35	-3.03	0.56
Oil_futures	288	65.01	24.84	63.89	0.17	-0.69	1.46	6.86	-2.53	0.87
HDD	288	349.61	308.26	282	0.39	-1.29	18.16	27.18	-17.05	0.04
CDD	288	118.08	125.22	51	0.88	-0.71	7.38	43.15	-19.23	0.07
GPR	288	108.86	51.34	94.84	4.31	26.35	3.02	9361.86	-4.93	0.62
Exim$nettransfer	288	-62683.5	237497.46	-120962.5	0.73	-0.83	13994.67	34.07	-1.62	4.48

Table 5: Descriptive statistics of the variables

4. Methodology

This thesis investigates five distinct forecasting models, including two hybrid variations. The following section provides a concise yet comprehensive overview of each model employed.

4.1 An introduction to SARIMAX model

The Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX) model is a widely used statistical approach for time-series forecasting, particularly within finance and economics. It effectively captures both trend and seasonal components of historical data while incorporating external variables to enhance predictive accuracy. Given its methodological credibility and adaptability, SARIMAX is utilized here as the benchmark model for performance comparison.

4.2 The construction of SARIMAX model

The SARIMAX model builds upon the ARIMA framework, which was originally introduced by Box and Jenkins in their work in 1976 and subsequently expanded upon by Wilson (2016).

The ARIMA model comprises the following components:

- **Autoregression (AR)**: Incorporates lagged values of the time-series to forecast the present. Denoted by order *p*, this component performs a linear regression on the previous *p* observations:

$$y_t = \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \cdots + \Phi_p y_{t-p} + \epsilon_t$$

, where y_t is the time-series value at time t, ε is the error term, and Φ is to capture the relationship between the current and past observations at a lag of p.

- **Moving Average (MA)**: Accounts for past forecast errors, allowing y_t to depend on the current and previous error terms. It is characterized by an order generally noted q. The moving average part involves running a linear regression on the last q error terms to predict the current value, expressed as follows:

$$y_t = \mu + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

, where y_t is the time-series value at time t, μ is the mean of the given time-series, ε_t is the error term at time t, and θ is the respective weight for each error term from t-1 until t-q.

By combining the two aforementioned, the ARMA (p,q) process will look like as the follows (Brooks, 2019):

$$y_t = \mu + \Phi_1 y_{t-1} + \cdots + \Phi_p y_{t-p} + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} + \epsilon_t$$

However, this model is effective on stationary time-series. To apply this to any timeseries, one must use the Integration (I) component, generally noted *d*.

- **Integration (I):** Transforms non-stationary time-series into stationary ones by differencing the data *d* times. In some cases, only the first-order difference could be enough, but some time-series can be made stationary after a number of differentiations.

$$y_t^d = (1 - B)^d y_t$$

, where B is the backshift operator notation, and *d* is the number of times the time-series is differenced.

The complete ARIMA (p,d,q) model then looks like as the follows:

$$(1 - B)^d y_t = \mu + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

Given the dynamics of gas prices, seasonality is incorporated into the model to capture recurring patterns that emerge at regular intervals. Price movements often exhibit seasonal behaviour, largely because one of the primary drivers – natural gas consumption – is closely tied to weather conditions and heating or cooling demand. In addition to seasonal effects, gas prices are influenced by various external factors such as storage levels, production rates, and other macroeconomic variables. To account for these influences, I introduce exogenous variables that help capture additional variance and enhance the model's predictive accuracy. The SARIMAX framework allows for the integration of both seasonal components and exogenous predictors. Here, P, D, and Q – the uppercase counterparts of the standard ARIMA parameters – denote the seasonal autoregressive order, the degree of seasonal differencing, and the seasonal moving average order, while β_k represents the coefficient of the exogenous variable X_k.

$$\left(1 - \sum_{i=1}^{p} \phi_i B^i\right)\left(1 - \sum_{I=1}^{P} \phi_I B^{sI}\right)(1 - B)^d (1 - B^s)^D y_t = \left(1 - \sum_{j=1}^{q} \theta_j B^j\right)\left(1 - \sum_{J=1}^{Q} \theta_J B^{sJ}\right)\epsilon_t + \sum_{k=1}^{m} \beta_k X_{k,t}$$

In this study, the *auto.arima* function in R is used to select the optimal model configuration, automatically handling differencing and detecting seasonality. Non-stationary data is fed into the function, which scales the series prior to estimation to avoid standard error inconsistencies. The best-fitting model is chosen based on the Akaike Information Criterion (AIC).

4.3 Random Forest model

The second modeling technique employed in this thesis is the Random Forest (RF) algorithm, well-suited for both classification and regression tasks. It constructs by building an ensemble of decision trees using the bagging method, where each tree is trained on a random subset of data (the so-called bootstrap sample), with a random selection (replacement) of features considered at each split.

Within each training iteration, a portion of the sample is excluded to form the out-of-bag (OOB) sample, which is subsequently used for cross-validation. This randomness of selections introduces diversity among the trees, reducing the correlation between decision trees and enhancing generalization. Final predictions are obtained by aggregating outputs, through majority voting in case of classification or averaging in case of regression, as is done here (Kashif & Ślepaczuk, 2025).

Random Forest is especially advantageous for its ability to handle cross-sectional and hand-dimensional datasets, observe feature importance, and provide robust performance with hyperparameter tuning. The three main hyperparameters to be set prior to training are: the number of trees, node size, and *m*, which denotes the number of features considered at each split (Breiman, 2001).

In this study, 3 different Random Forest models are developed using lag structures of 12, 24, and 36 months, corresponding to one-, two-, and three-year time windows, respectively. The same proportion of splitting is applied to the training and test sets (70-30) in each case. The number of trees is fixed at 500, with 10 iterations for robustness. The *mtry* parameter is optimized via the *tuneRF* function for each lag specification, selecting the configuration that minimized OOB error. Predictions, mean squared errors are extracted from the model, and variable importances are also checked as will be presented in Section 6. The models are trained on raw data, without differencing or scaling, given that Random Forest algorithms do not require stationarity or normalized inputs. Figure 1 visualizes the model.

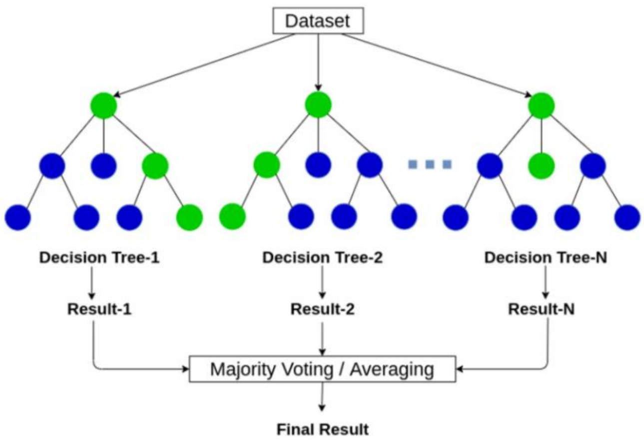


Figure 1: Composition of a Random Forest model (Omarzai, 2014)

4.4 An introduction to Long-Short Term Memory model

Long-Short Term Memory model is a variant of neural networks, that was first introduced by Sepp Hochreiter and Jürgen Schmidhuber in 1997, and later became very popular especially to improve time-series prediction accuracy, and solving more complex predictive tasks (Hochreiter & Schmidhuber, 1997). LSTM represents an advanced form of Recurrent Neural Networks (RNN), that is itself also a developed version of Artificial Neural Networks (ANN).

Neural networks typically comprise three layers: input, hidden, and output. Thus, neural networks learn past information to produce future predictions similarly to humans. However, basic neural networks operate with fixed-size inputs and outputs, which constrains their ability to generalize to unseen data or longer sequences (Kaur & Mohta, 2019).

To handle this drawback, RNNs were developed to detect and learn patterns in sequential data, thereby enhancing their estimation capability. Moreover, RNN adds additional loops to the architecture, meaning the nodes of the hidden layers are interconnected, and the output of the previous hidden node (e.g., h1) will be an input of the next node (e.g., h2) in parallel with new input values (see Figure 2). It means the output depends on a sequence of data (Manaswi, 2018). RNNs can be categorized by the number of inputs and outputs. In this study, a sequence-to-one format is used, whereby the model uses a sequence of previous observations to predict the next point in time.

A major limitation of RNNs lies in their computational behaviour. First, during forward propagation, hidden states are passed forward, second, during backpropagation, gradients are computed across all time steps to update network weights. Due to the chain rule in differentiation, these gradients can either diminish to near-zero (values below 1 become even smaller, called vanishing gradients) or increase uncontrollably (values over 1 become even larger, called exploding gradients). Small gradients will cause an early stop in effectively updating previous layers, while too large weights will cause instability and possible NaN errors (Sherstinsky, 2020). Although RNNs possess memory through feedback loops, their ability to retain information over long sequences is compromised. This loss of long-term memory – exacerbated by vanishing gradients – prevents the model from effectively learning long-term dependencies (Hochreiter et al., 2001).

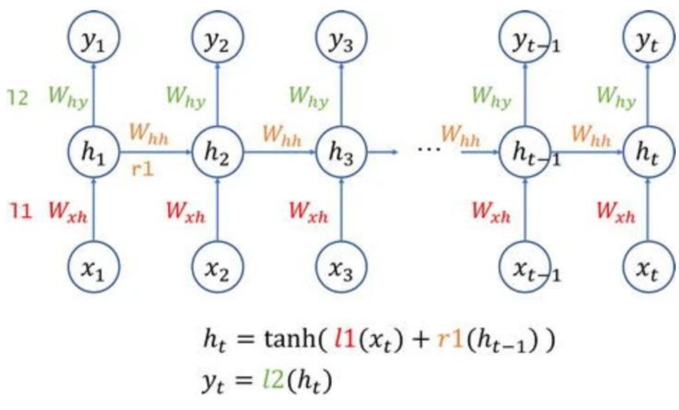


Figure 2: Illustration of a Recurrent Neural Network according to Manaswi (2018). A hidden node (h2) not only depends on the input node (x2), but also on the previous hidden node (h1)

4.5 The construction of Long-Short Term Memory model

To address the aforementioned challenges, a modified RNN architecture – Long Short-Term Memory (LSTM) – is introduced. LSTM tackles the issue of exploding or vanishing gradients: the model addresses this challenge by introducing gated units and memory cells into the architecture. Specifically, memory cells maintain a persistent cell state, functioning as internal long-term memory that preserves relevant information across time steps. Unlike traditional RNNs, these memory cells are not influenced by trainable weights or biases, thereby circumventing gradient instability. For short-term memory, a hidden state is utilized, which is directly updated via learned weight matrices. When new input data is introduced, the LSTM model leverages both the current input and existing cell state to determine what to retain, update, or delete (Istiake Sunny et al., 2020).

The model comprises 3 types of gates: input, forget and output. Firstly, in the **forget gate**, the model determines – by a *sigmoid* function – the proportion of the long-term memory to be preserved, producing an updated long-term memory. Secondly, an **input gate** (sometimes referred to as **update gate**) is introduced to modify the long-term memory from the previous unit. Inside the update gate, a *tanh* function is to determine the potential long-term memory, while the previous short-term memory is also used as an input. By the *sigmoid* function, the gate determines the proportion of the long-term memory to be considered in the model. The result is an updated cell state formed by summing the retained memory (from the forget gate) and the new memory (from the input gate). Lastly, the **output gate** creates the potential short-term memory by a *tanh* function again. The input of this gate is the updated long-term memory from the previous unit, as well as the previous short-term memory. By the *sigmoid* function, the gate again determines the proportion of the potential short-term memory to be integrated. The output of this gate (and the overall step) is a new short-term memory, which is passed to the next time step (Olah, 2015; Starmer, 2022). The updated long-term memory is forwarded to the forget gate in the next iteration (see Figure 3 below).

To sum up, the output of LSTM is influenced by both the current cell state and the incoming information. Unlike traditional RNNs that struggle to retain long-term information, LSTMs are designed to naturally remember important information over extended sequences. Their design avoids the vanishing or exploding gradient issues, allowing for robust modeling of long-term patterns. In this study, a sequence-to-one architecture is applied using the same lag configurations as in the Random Forest models (12, 24, and 36 months). The input series is differenced to ensure stationarity and then normalized before being processed by the model.

$$i_t = \sigma(W_i x_t + W_{hi} h_{t-1} + b_i)$$

$$f_t = \sigma\left(W_f x_t + W_{hf} h_{t-1} + b_f\right)$$

$$o_t = \sigma(W_o x_t + W_{ho} h_{t-1} + b_o)$$

$$\widehat{c_t} = tanh(W_c x_t + W_{hc} h_{t-1} + b_c)$$

$$Long\ memory: C_t = f_t \cdot c_{t-1} + i_t \cdot \widehat{c_t}$$

$$Short\ memory: h_t = o_t \cdot tanh(c_t)$$

, where W stands for weights, b is for biases, x_t is the sequence of time t, f_t is the forget gate at time t, h_{t-1} is the hidden state at time t-1, i_t is the input gate at time t, $\widehat{c_t}$ is the change gate at time t, and o_t is the output gate at time t (Kashif & Ślepaczuk, 2025).

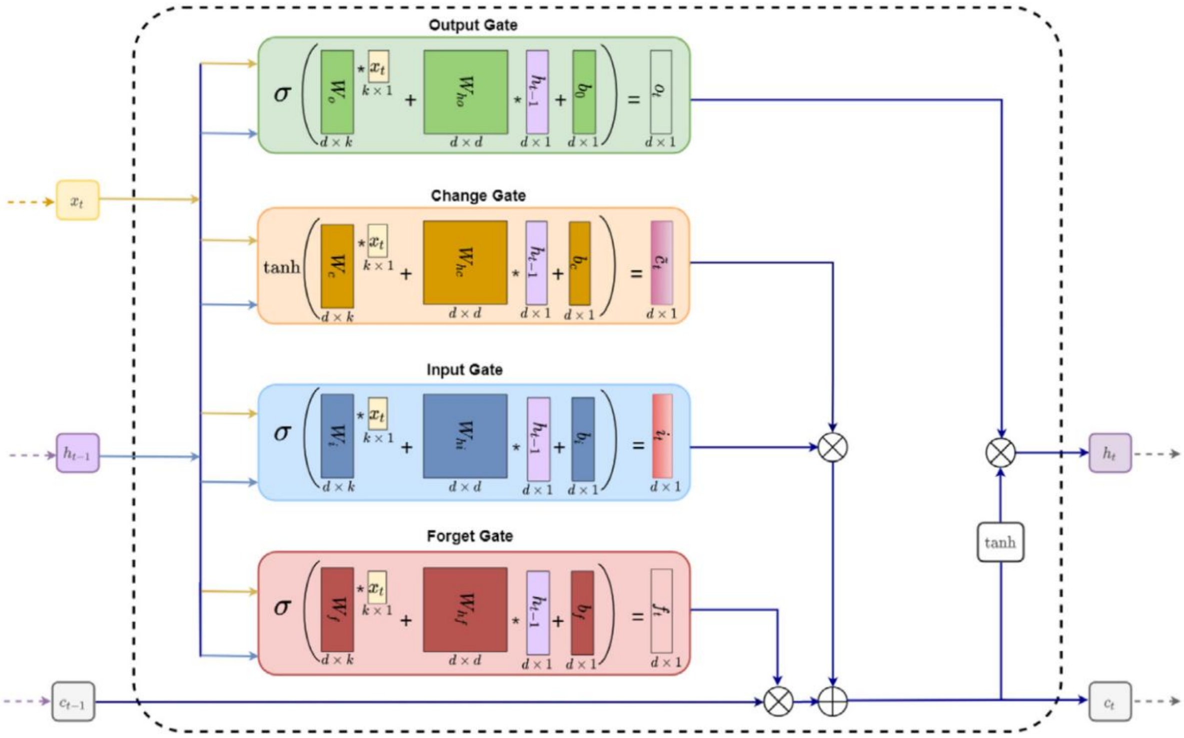


Figure 3: Illustration of an LSTM model according to Kashif & Ślepaczuk (2025). Please note that sometimes Input gate is split to input and change gates, where the previous refers to the one determines the proportion of data being considered, while the latter is to determine about the potential long-term memory

As machine learning models, Random Forest and LSTM do not require stationarity and normality, due to its ability to model complex temporal patterns (Abid et al., 2025), however some papers state that deep learning models, especially neural networks may face the challenge of catastrophic forgetting despite their ability to learn patterns accordingly (Ao & Fayek, 2023). For this reason, I feed LSTM with a differenced and scaled dataset.

4.6 The construction of SARIMAX-Random Forest hybrid model

In the first model, I combine the SARIMAX and Random Forest methods. While the previous is to capture linear trends, seasonality, and autoregressive relationships in time-series data, it struggles with complex nonlinear dependencies, interactions, or sudden shifts in patterns. However, Random Forest was introduced to capture this non-linear, handle complex feature relationships, and irregular structures in the data, as its non-parametric ensemble method based on decision trees (Topal et al., 2019).

For this model, the residuals of SARIMAX, in other words, those parts that the model is not able to predict properly, are used as the inputs to Random Forest. The hybrid models are evaluated using the same time lags (12, 24, and 36 months). Hyperparameters remains the same with the number of trees 500 and iterations 10, only m changes – it is set to 1, given the univariate nature of the residual. No more regressors are added to the model in this step. The final predictions are produced by added back the Random Forest predictions to the SARIMAX ones. SARIMAX predictions are truncated by 12, 24, and 36 to match with Random Forest's predictions. Figure 4 illustrates the process.

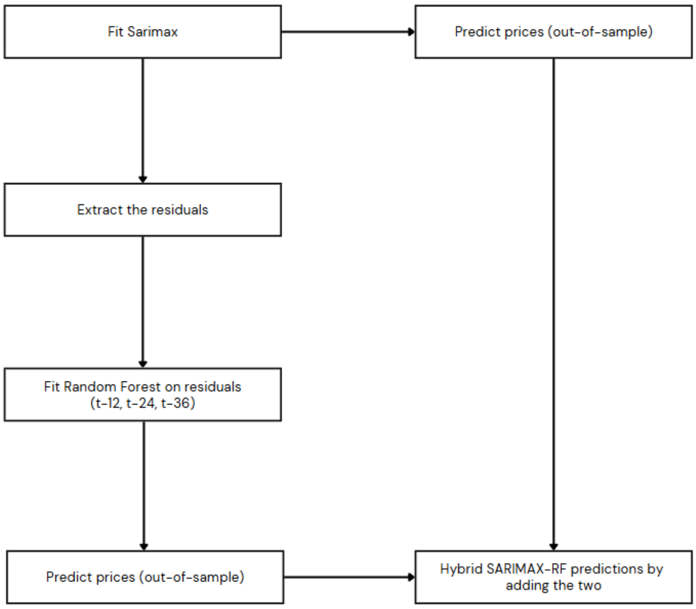


Figure 4: Flowchart for SARIMAX-RF model

4.7 The construction of SARIMAX-LSTM hybrid model

The second hybrid model combines SARIMAX and LSTM. As with the previous hybrid, SARIMAX is first used to estimate the linear trend and generate residuals, which are then used as input features for the LSTM model. Model trained on the same sequences (12, 24, and 36 months). The residuals are scaled before feeding into the model. This hybrid model holds the same perks as the previous one: as ARIMA stands for capturing linear trends and short-term dependencies, as well as autocorrelation, LSTM contributes by dealing with long-term dependencies and nonlinear relationships (Kashif & Ślepaczuk, 2025; Sheng & Jia, 2020). According to Jang, Seo and Lee (2025), the hybrid model can be written as the follows, where y_t is the time-series at time t, l_t is the linear component predicted by SARIMAX at t and n_t is the non-linear component predicted by LSTM at t:

$$y_t = l_t + n_t$$

Moreover, r_t represents the residuals derived from SARIMAX prediction at time t (the difference between actual and predicted value), expressing the non-linear component of the model:

$$r_t = y_t - \widehat{l_t} \sim n_t$$

Consequently, the final prediction made by the model equals the addition of the linear and nonlinear parts, as the follows:

$$\widehat{y_t} = \widehat{l_t} + \widehat{n_t}$$

This hybrid setup also includes lagged gas prices alongside the residuals to enrich the temporal context and potentially enhance the model's learning of dynamic patterns. The LSTM parameters (architecture and hyperparameters) remain consistent with those used in the standalone LSTM models. The final predictions are produced by added back the LSTM predictions to the SARIMAX's values. SARIMAX predictions are truncated by 12, 24, and 36 to match LSTM's output window. Figure 5 illustrates the process.

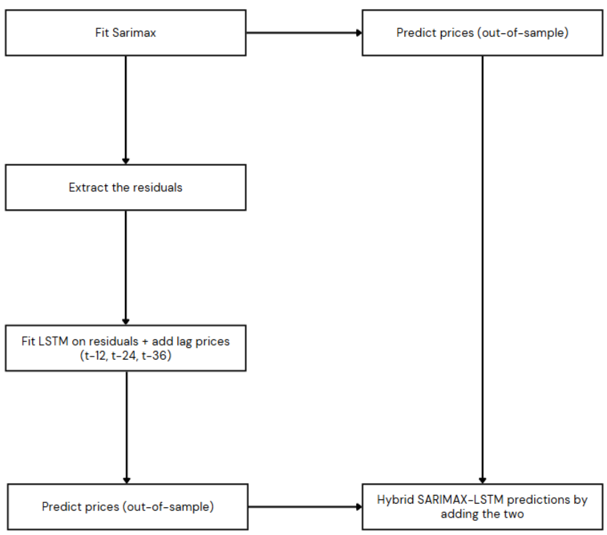


Figure 5: Flowchart for SARIMAX-LSTM model

4.8 Performance Evaluation Metrics

I have assessed for different metrics to evaluate the models' performance. Root Mean Squared Error, or RMSE, Mean Absolute Error, or MAE, Symmetric Mean Absolute Percentage Error, or SMAPE and Directional Accuracy, or DA are observed in this study. As previously mentioned, a 70–30 split is applied to divide the data into training and testing sets. The models are trained on the training set and subsequently evaluated on the test set, generating out-of-sample forecasts to assess predictive capability (Hyndman & Athanasopoulos, 2018).

RMSE, MAE, and DA are widely utilized performance indicators in time-series forecasting, especially for models involving numeric prediction tasks (Chai et al., 2018; Jang et al., 2025; Zhao et al., 2017). Additionally, MAPE is often regarded as a heteroscedasticity-adjusted MAE and have been widely popularized in time-series forecasting tasks (Wen et al., 2016; Yu et al., 2022). However, MAPE is not symmetric, meaning over- and under-predictions are penalized unequally, and becomes undefined or unstable when actual values approach zero due to division by zero or near-zero denominators.

To eliminate these drawbacks, I introduce Symmetric Mean Absolute Percentage Error, or SMAPE, that has gained popularity in the forecasting literature (Abdulrahim et al., 2025; Akshayaa et al., 2025; Dong et al., 2024). SMAPE is advantageous due to its balanced treatment of overestimation and underestimation, setting it apart from previously used metrics RMSE and MAE. This design makes it more robust to outliers and provides a fairer evaluation in forecasting tasks (Swanson & Tayman, 1999). Generally, lower values of RMSE, MAE, and SMAPE indicate higher predictive accuracy, whereas higher DA values reflect greater proficiency in capturing the correct direction of change in the target variable.

The mathematical formulas for each metric are as follows:

$$RMSE = \sqrt{\left\{\frac{1}{n}\right\} * \sum_{t=1}^{n} (y_t - \widehat{y_t})^2}$$

$$MAE = \left\{\frac{1}{n}\right\} * \sum_{t=1}^{n} |y_t - \widehat{y_t}|$$

$$SMAPE = \left\{\frac{100\%}{n}\right\} * \sum_{t=1}^{n} [\frac{|y_t - \widehat{y_t}|}{\left(\frac{(|y_t| + |\widehat{y_t}|)}{2}\right)}]$$

$$DA = \left(\frac{1}{(n-1)}\right) * \sum_{t=2}^{n} 1[\sin(y_t - y_{t-1}) = \sin(\widehat{y_t} - \widehat{y_{t-1}})]$$

5. Hypothesis

The central objective of this study is to evaluate the performance of machine learning models – specifically Random Forest and Long Short-Term Memory – in forecasting the front-month Henry Hub natural gas price, and to assess whether hybrid modeling approaches can further enhance predictive accuracy. Based on an extensive review of the relevant literature and preliminary data analysis, three core hypotheses are formulated:

- **Hypothesis 1 (subquestion):** Over the given time horizon and using the specified regressors, Random Forest model demonstrates superior predictive accuracy compared to the SARIMAX model across all metrics.

This hypothesis is based on the capacity of ensemble learning methods, like Random Forest, to capture complex, nonlinear relationships and patterns in data. While SARIMAX is effective in modeling seasonal and autoregressive structures, its linear formulation may be inefficient to observe patterns or structural breaks that characterize natural gas markets.

- **Hypothesis 2 (subquestion):** Over the given time horizon and using the specified regressors, LSTM consistently yields the most accurate forecasts across all metrics and time horizons.

This assumption is grounded in the architectural sophistication of LSTM networks, which are designed to retain long-term dependencies and capture temporal dynamics. Numerous prior studies in financial forecasting have reported superior results for LSTM compared to traditional econometric models and even other machine learning methods, particularly when dealing with sequential data.

- **Hypothesis 3 (main question):** Combined models can improve forecasting performance and clearly outperform their standalone counterparts across all metrics.

This hypothesis stays on the premise that hybrid models can capture both linear and nonlinear patterns of the values. By first modeling the linear structure via SARIMAX and then capturing the residual nonlinearities using RF or LSTM, the combined models are expected to provide improved generalization and lower forecast errors. This assumption is further supported by several empirical studies in both energy and financial forecasting domains, where hybrid approaches have been shown to outperform individual models (Kashif & Ślepaczuk, 2025; Ray et al., 2023; Seba & Belaide, 2024).

The validity of these hypotheses is evaluated based on the performance of each model using four key metrics – RMSE, MAE, SMAPE, and DA – applied to the out-of-sample test dataset.

6. Results

I would like to present my findings in the following section. First, I introduce SARIMAX results, move on to machine learning methods, and finish with combined models.

6.1 Hypothesis 1 – a comparison between SARIMAX and RF models

Based on the Akaike Information Criterion (AIC), the *auto.arima* function identifies the optimal model as SARIMAX (3,0,2)(0,0,0), which entails three autoregressive components, no differencing, and two moving average terms. Notably, there is no seasonal component, as the seasonal parameters – *P*, *D*, and *Q* – are all zero. This is intriguing, given that both the ADF and KPSS tests previously indicated that the target variable possesses one unit root. The explanation behind is the following: *auto.arima* relies on AIC minimization, thus selects the d=0 specification yielding an AIC of 91.3, compared to 101.9 under d=1.

Furthermore, the Ljung-Box test returns a p-value of 27.62%, suggesting that autocorrelation is not statistically significant in the residuals, and thus the null hypothesis of no autocorrelation cannot be rejected (see Table 10 in the Appendix). Model predictions are then rescaled to their original values, and evaluation metrics are computed accordingly. Table 6 below presents the estimated coefficients.

Variable	Estimate	Std_Error
ar1	-0.6330	0.1187
ar2	0.6786	0.0667
ar3	0.4990	0.1060
ma1	1.7251	0.0801
ma2	0.8528	0.0704
workinggas_in_storage	-0.0360	0.0549
total_consumption	-0.3876	0.1212
natgas_dryproduction	-0.1319	0.1229
rigs_count	0.1813	0.1782
Interestrate	0.3320	0.1483
Oil_futures	0.3611	0.0941
HDD	0.5940	0.1474
CDD	0.1949	0.0735
GPR	0.0151	0.0288
res	-0.0694	0.0360
hurricanes	-0.0481	0.0772

Table 6: SARIMAX model's estimates

As we can notice, all AR and MA coefficients are statistically significant. Economically, a -0.633 coefficient for AR1 indicates when controlling for the remaining AR and MA terms, a one-unit increase in the price last month is associated with a \$0.633 decline in the current price. Interestingly, this negative relationship contrasts with the positive signs observed for AR2 and AR3, which suggest that price movements two and three months ago have a stronger upward influence. This result is somewhat counterintuitive, as one might expect more recent months to exert a stronger impact. Additionally, it is noteworthy to have different signs in case of the latest 3 months.

MA1, with a value of 1.725, implies that a positive shock one month prior tends to elevate the current price by \$1.725, while MA2 retains a positive yet more modest effect. Among the exogenous regressors, five variables – consumption, oil futures, interest rate, HDD, and CDD – are statistically significant at the 5% level. A coefficient of 0.361 for oil futures suggests that a one-dollar increase in WTI prices leads to a \$0.361 rise in gas prices. Similarly, positive coefficients for HDD, CDD, and interest rates align with the notion that increased demand and economic slowdowns generally put upward pressure on prices.

Overall, SARIMAX delivers moderate forecast accuracy, yielding an RMSE of 2.096 and an MAE of 1.46. The SMAPE metric reflects the model's limited ability to capture peak and valley behaviour, while a Directional Accuracy (DA) of only 9.3% indicates that it correctly predicts the direction of price changes in less than one out of ten cases. The model's modest performance motivates the use of more sophisticated deep learning techniques for improving accuracy. Figure 11 in the Appendix displays the actual and fitted values. Evaluation results are reiterated in Table 1 below.

	Model	RMSE	MAE	SMAPE (%)	DA (%)
1	SARIMAX	2.0960	1.4582	56.3758	9.3023
2	RF(12)	1.9096	1.5413	43.1809	2.4096
3	RF(24)	2.5159	2.0265	50.3773	5.0633
4	RF(36)	1.8559	1.3270	33.4656	7.8947
5	LSTM(12)	0.6097	0.4291	11.8458	24.6575
6	LSTM(24)	0.6331	0.4359	11.8425	18.0328
7	LSTM(36)	0.7064	0.4963	12.3208	16.3265
8	SAR-RF(12)	2.6072	1.8859	75.9386	4.0541
9	SAR-RF(24)	2.3976	1.8255	71.3846	9.6774
10	SAR-RF(36)	2.6433	2.0346	66.6249	4.0000
11	SAR-LSTM(12)	3.4132	2.4466	102.3196	5.4054
12	SAR-LSTM(24)	3.2462	2.4512	103.4218	1.6129
13	SAR-LSTM(36)	3.1925	2.4145	88.9851	8.0000

Table 1: Final results

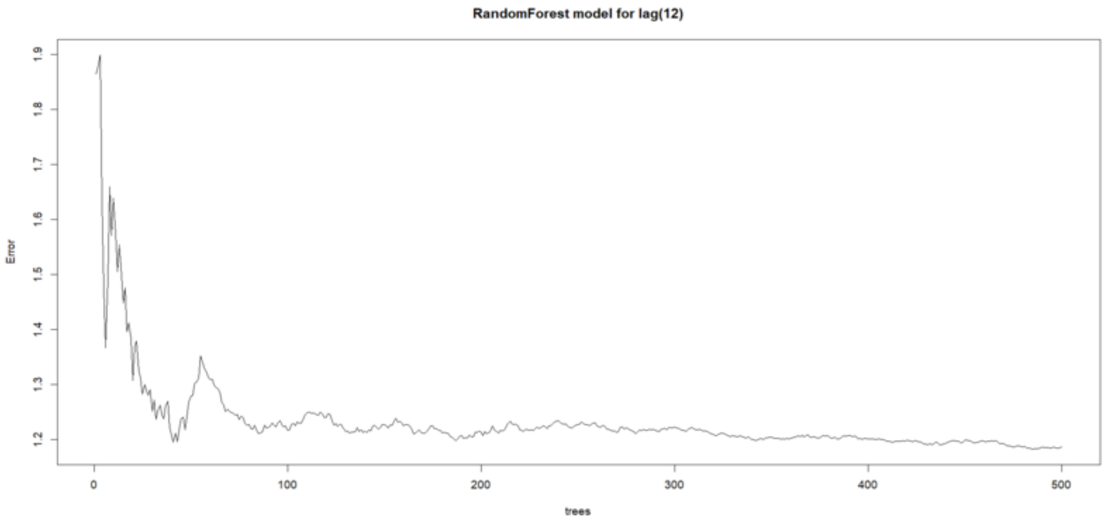
Random Forest, a tree-based ensemble learning method, does not require input data scaling and is not bound by stationarity assumptions. The incorporation of lagged variables enables the model to detect and learn temporal patterns. In this study, three lag configurations – 12, 24, and 36 months – are tested, each using a 70-30 train-test split. The number of trees is set at 500, with 10 iterations performed to ensure robustness and reproducibility. An optimal *mtry* value is determined via preliminary tuning (see Figure 12 in the Appendix), resulting in $m = 6$ for lag(12) and $m = 9$ for the 24- and 36-month models.

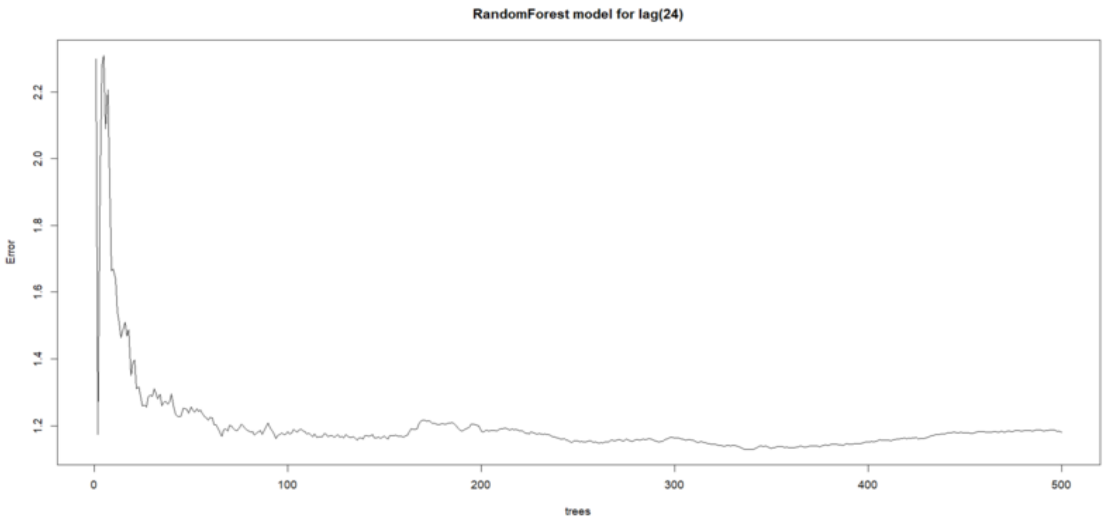
Figure 6 illustrates the relationship between the number of trees and associated prediction errors. For all lag configurations, the out-of-bag (OOB) error rapidly decreases until around 100 trees and stabilizes near 200, indicating model convergence. Based on this pattern, lag(12) appears to be the best-performing configuration. However, the final evaluation metrics gives controversial results.

The final predictions for each lag structure are obtained by averaging across all iterations. As reported in Table 1, the Random Forest model demonstrates moderate errors in RMSE, MAE, and SMAPE, but underperforms in terms of DA. The best-performing variant – RF(36) – achieves an 11% reduction in RMSE (from 2.096 to 1.85), a 9% reduction in MAE (from 1.46 to 1.33), and a striking 23 percentage point improvement in SMAPE (from 56.4% to 33.5%) relative to SARIMAX. However, RF(36) performs even worse than SARIMAX in directional accuracy, failing to capture upward or downward movements effectively. This smoothing tendency is visually evident in Figure 8.

The performance discrepancies across lags highlight the importance of window size, with RF(36) producing the most accurate forecasts. Interestingly, RF(24) underperforms RF(12), underscoring the potential value of further hyperparameter optimization. In summary, Random Forest substantially outperforms SARIMAX in terms of absolute error, though its limitations in directional forecasting indicate room for enhancement. Forecast plots for all RF models can be found in Figures 13, 14, and 15 in the Appendix.

Variable importance, illustrated in Figure 7, is assessed using the *IncNodePurity* metric. It corresponds to the tree's loss function, which guides split selection. For regression trees this is the mean squared error. The *IncNodePurity* score accumulates all impurity decreases attributed to a given variable – each weighted by the probability of reaching the split – and then averages that across all decision trees in the ensemble. A larger *IncNodePurity* value suggests the variable plays a more critical role in the model. (Islam et al., 2023).





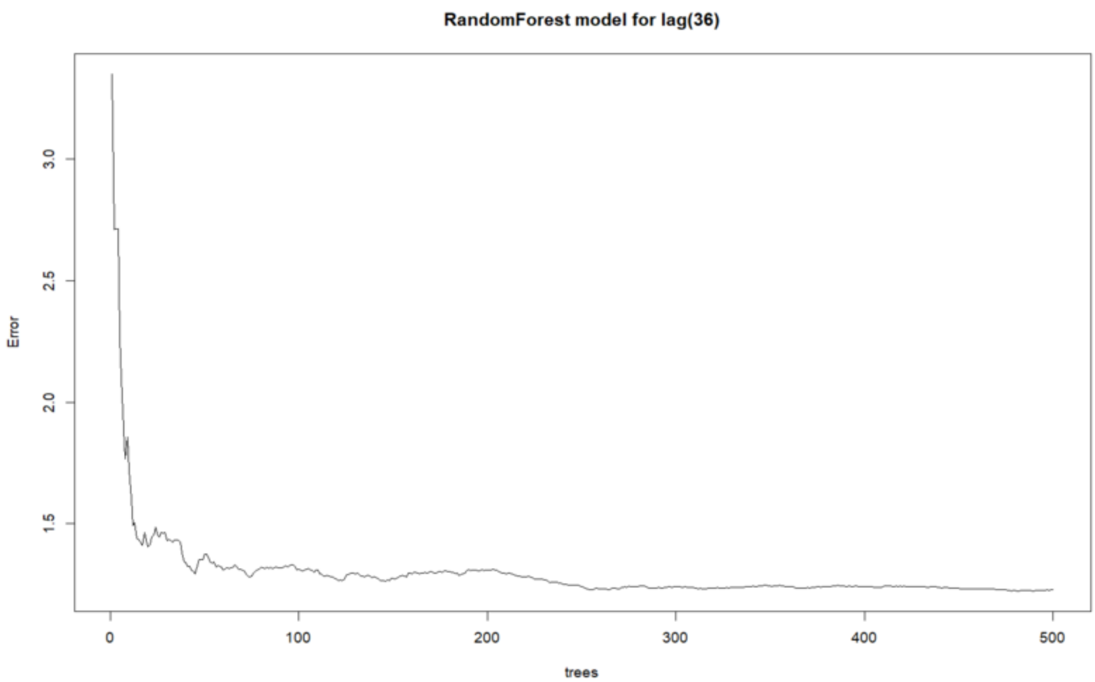


Figure 6: OOB errors for Random Forest models

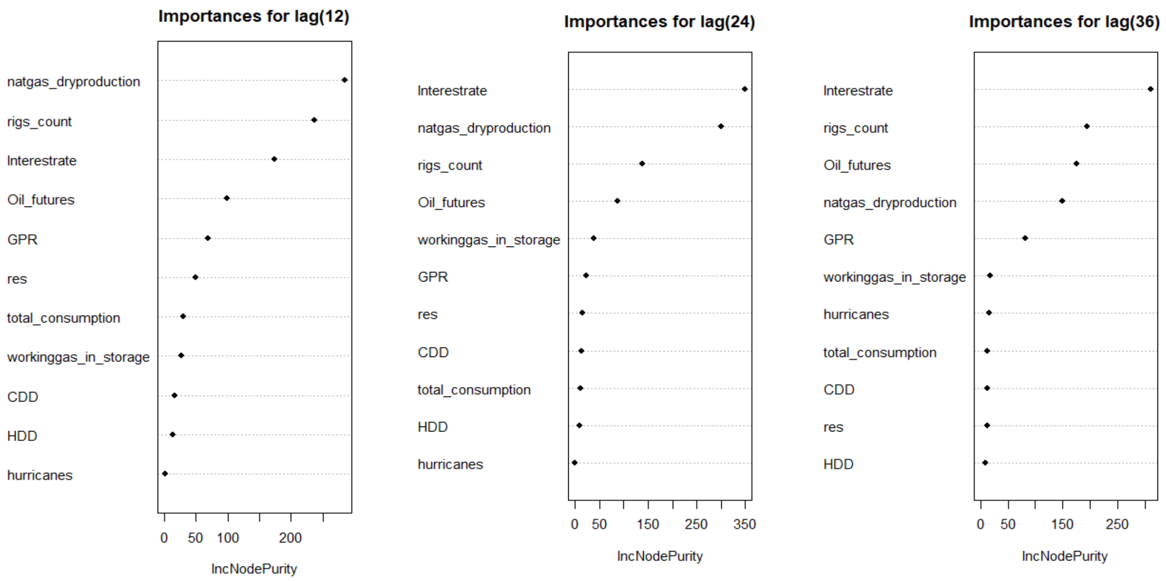


Figure 7: Variable importances for Random Forest models

According to this plot, interest rate is the most dominant variable, while oil futures, production and rigs count also contribute a lot to the explanatory power of the given models. In contrast, consumption and storage levels importance are lower than expected. Notably, the GPR performs better despite its insignificance in the SARIMAX model. Weather variables (CDD, HDD, hurricanes) contribute very little to predictive power despite those are all statistical and economically significant in case of SARIMAX. This pattern remains consistent across RF(24) and RF(36), suggesting that variable importance stabilizes after the 2-year window.

6.2 Hypothesis 2 – LSTM predictions and its comparison to previous models

As a machine learning model, LSTM is designed to identify and learn complex sequential patterns. However, if the input time series is non-stationary, the presence of trend or seasonality can obscure underlying dynamics, reducing model effectiveness and possibly introducing spurious patterns. Differencing helps mitigate these issues by transforming the series to highlight genuine fluctuations. Hence, a differenced and scaled dataset is used as input to the LSTM model.

Sequence-to-one configurations of 12, 24, and 36 months are applied, where the model uses a window of lagged values (e.g., months 1–12) to predict the subsequent value (e.g., month 13). In each case, an LSTM model with three layers is implemented using the *keras* package in R. The architecture consists of two LSTM layers – comprising 64 and 32 units respectively – with *tanh* activation and dropout rates of 0.2. In the first layer, I allow return sequences, while in the second case I left it as default, meaning the model returns only the last output of the sequence instead of one output for each time step. A final dense layer is added with one unit and a *linear* activation function. The model is trained using the Mean Squared Error loss function and the *adam* optimizer over 50 epochs, with a batch size of 32 and a 20% validation split.

The predictions are transformed back to their original scale by reversing both differencing and normalization. According to Table 1, LSTM models decisively outperform both Random Forest and SARIMAX. This aligns with expectations, as neural networks "have the ability to estimate a general level of function with a high degree of precision and to discover nonlinear patterns in input datasets without defaults" (Safari & Davallou, 2018, pp. 51). LSTM(12) achieves the best overall performance in three out of four metrics, while LSTM(24) slightly outperforms it in SMAPE.

	Model	RMSE	MAE	SMAPE (%)	DA (%)
1	SARIMAX	2.0960	1.4582	56.3758	9.3023
2	RF(12)	1.9096	1.5413	43.1809	2.4096
3	RF(24)	2.5159	2.0265	50.3773	5.0633
4	RF(36)	1.8559	1.3270	33.4656	7.8947
5	LSTM(12)	0.6097	0.4291	11.8458	24.6575
6	LSTM(24)	0.6331	0.4359	11.8425	18.0328
7	LSTM(36)	0.7064	0.4963	12.3208	16.3265
8	SAR-RF(12)	2.6072	1.8859	75.9386	4.0541
9	SAR-RF(24)	2.3976	1.8255	71.3846	9.6774
10	SAR-RF(36)	2.6433	2.0346	66.6249	4.0000
11	SAR-LSTM(12)	3.4132	2.4466	102.3196	5.4054
12	SAR-LSTM(24)	3.2462	2.4512	103.4218	1.6129
13	SAR-LSTM(36)	3.1925	2.4145	88.9851	8.0000

Table 1: Final results

Compared to RF(36), LSTM(12) reduces RMSE and MAE by 67%, and by 71% when compared to SARIMAX. LSTM models also achieve SMAPE values near 12%, approximately one-third of RF(36)'s and one-fifth of SARIMAX's values. Moreover, DA underscore LSTM's capacity to forecast directional shifts – particularly rapid spikes and sharp downturns – achieving nearly 25% directional accuracy in the best-case scenario.

Interestingly, the bigger window applied the poorer accuracy observed, suggesting that a longer lookback injects noise or overfitting. Thus, LSTM(12) appears to strike an optimal balance between historical context and model simplicity. To conclude, LSTM's dominance over RF and SARIMAX is unquestionable, and achieves much better results both in minimizing error and in predicting turning points. However, as longer windows experienced downgrade in performance, I would like to move on with hybrid models, as those can capture both linear and non-linear patterns, and maybe result further improvements. Plots for the actual and fitted values can be found in the Appendix (Figure 16,17,18).

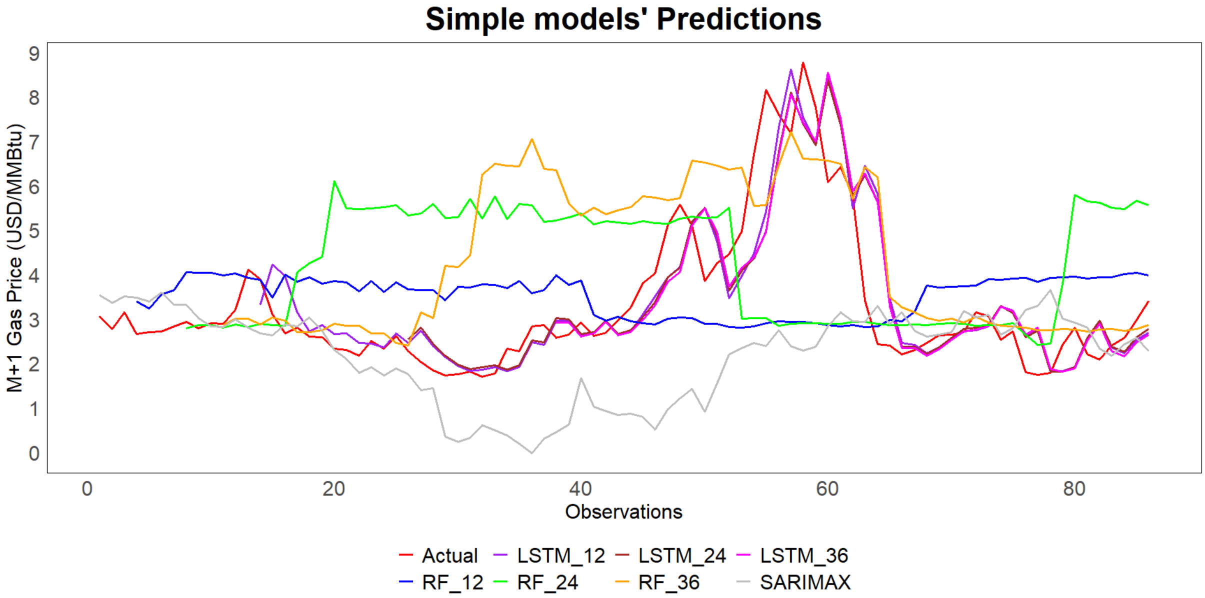


Figure 8: Simple models' predictions

6.3 Hypothesis 3 – Hybrid models' performance and comparison

As introduced in Section 4, two hybrid models are constructed. The first, SARIMAX-RF, uses the residuals from the SARIMAX model as the dependent variable for a new Random Forest regression. Since all other variables were already included in the SARIMAX model, only lagged residuals are used as predictors. Final forecasts are obtained by summing RF-predicted residuals with SARIMAX outputs, which are then rescaled.

Surprisingly, the hybrid models perform markedly worse than the standalone ones. In the best-case scenario (lag (24)), RMSE and MAE increase by 14–37% compared to either SARIMAX or RF(36), and results lag far behind LSTM. SMAPE ranges from 66% to 75%, with lag(36) performing the best – but still twice as high as RF(36) and 10 percentage points higher than SARIMAX. DA scatters around 4% for lag(12) and lag(36), with only lag(24) marginally surpassing SARIMAX and RF. These results suggest that combining SARIMAX with RF introduces additional complexity without improving predictive accuracy. Residuals may be primarily noise, leading to overfitting in the RF component. Furthermore, the lack of autocorrelation in SARIMAX residuals implies that RF has no meaningful patterns to learn. Plots for the actual and fitted values can be found in the Appendix (Figure 19,20,21).

The second hybrid model, SARIMAX-LSTM, follows a similar process. Residuals are scaled before being input into an LSTM with the same hyperparameters (units, activation functions, dropout rates, epochs, batch size and validation split) and configurations (12, 24, 36 months) to the standalone version. Lagged gas prices also included as a second feature. Predictions are then scaled back twice to restore the original level.

Strikingly, SARIMAX-LSTM provides an even worse performance than the previous hybrid model. MSE and MAE are five to six times higher than the standalone LSTM model and approximately 50% worse than SARIMAX. SMAPE exceeds 100%, and DA falls below 10%, suggesting that the residuals are essentially white noise, lacking any learnable structure. A plausible explanation is that SARIMAX already captures most of the explainable variance, leaving little for LSTM to learn. Thus, LSTM just introduce complexity to the model. Additionally, the limited sample size – 288 observations, reduced further by lagging – suggests LSTM may not have enough data to learn that can make its predictions worse and diminishing generalization on test data. Plots for the actual and fitted values can be found in the Appendix (Figure 22,23,24).

To sum up, combined models do not show any improvement in evaluation metrics over the simple methods, due to possibly overfitting, white noise error terms, already captured variance or limited sample size. Forecasts from the hybrid models are visualized in Figure 9 below.

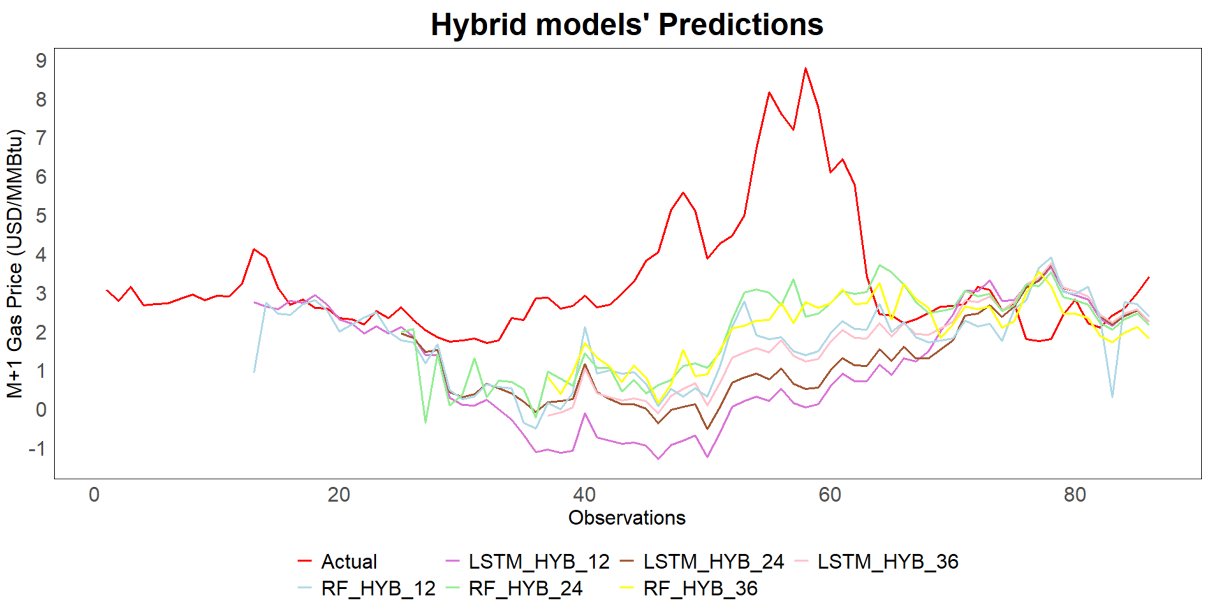


Figure 9: Hybrid models' predictions

7. Conclusion

The strategic importance of energy resources – particularly natural gas – has grown significantly in Europe in the aftermath of the Russian-Ukrainian war and the reemergence of conflict in the Middle East. This shift is primarily due to the continent's gradual disengagement from its reliance on Russian energy imports. Among the available alternatives, American LNG has emerged as the most viable – but more expensive – option. Notably, its price is strongly influenced by the U.S. benchmark, the Henry Hub (Günther & Nissen, 2021). Consequently, it has become increasingly essential for European policymakers, energy strategists, and industry stakeholders to monitor the U.S. gas market closely, especially as Europe is to become one of the largest global importers of American LNG.

In recent years, numerous studies have addressed price forecasting, using either simple statistical models like ARIMA, or more recently, machine learning models such as Random Forest and LSTM, thanks to the advancement of computer hardware and processing power. To enhance predictive accuracy, researchers have begun using hybrid models capable of capturing both linear and non-linear dependencies in the data.

This thesis introduced one econometric model and two machine learning approaches. The first is SARIMAX, which enhances the commonly used ARIMA model by including seasonality and 11 exogenous regressors. The Random Forest and LSTM models are applied under three configurations (12-, 24-, and 36-month lags). In the case of hybrid models (SARIMAX-RF and SARIMAX-LSTM), the machine learning models are trained on the residuals of the linear model to capture potential non-linearities. The thesis tests three hypotheses: the first evaluates the efficiency of Random Forest compared to SARIMAX; the second assesses the superiority of LSTM models; and the third examines the potential for hybrid models to improve forecast accuracy.

The empirical findings support the second hypothesis. LSTM models significantly outperform both SARIMAX and Random Forest across all lag horizons and evaluation metrics. In contrast, the first and third hypotheses are rejected. The Random Forest model does not consistently outperform SARIMAX, and the hybrid models, rather than enhancing predictive accuracy, resulted in greater forecasting errors. This suggests that SARIMAX may have already accounted for the majority of the data's variance, leaving little residual structure for machine learning models to exploit – leading instead to overfitting or unnecessary model complexity. Figure 10 below visualizes all model predictions.

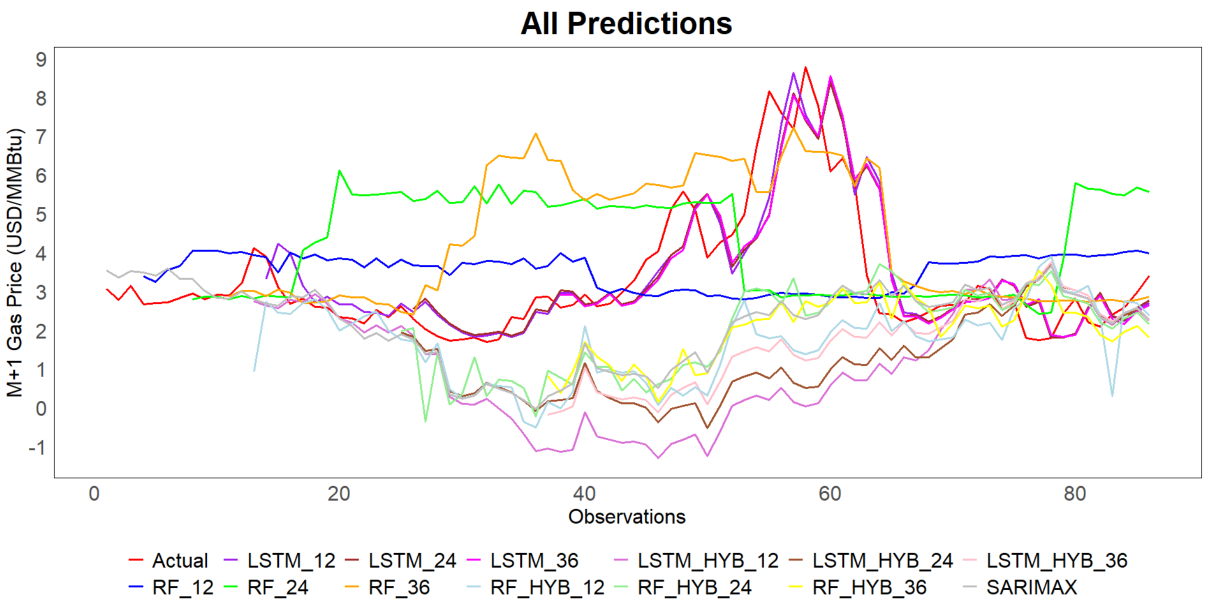


Figure 10: All models' predictions

While these results may seem counterintuitive in light of previous literature – where hybrid models often surpass simpler alternatives – it is important to recognize that the SARIMAX model used here is more advanced than a basic ARIMA, incorporating a rich set of exogenous variables. This likely improves its baseline performance, leaving less room for hybrid enhancements. Therefore, based on 24 years of monthly data and the parameters included, it can be concluded that combined models do not result in superior forecasts for Henry Hub front-month natural gas prices. Among all configurations, LSTM(12) achieves the lowest errors and the highest directional accuracy and is thus recommended as the optimal forecasting tool for gas prices.

This thesis aspires to contribute to the academic dialogue and provide a critical perspective on model selection in energy forecasting. Its limitations are acknowledged and may inspire further research. The performance of hybrid models could potentially be improved through more comprehensive hyperparameter tuning, which would require greater computational capacity. Moreover, the analysis is constrained using monthly data. By applying higher-frequency data (e.g., daily prices) and fewer exogenous variables could diminish the explanatory power of the SARIMAX model and allow machine learning models to capture more of the variance. Additional independent variables could also be incorporated in future work. Furthermore, as noted in Section 1, energy prices are heavily influenced by investor psychology, which is in turn shaped by geopolitical developments and political statements – factors that are inherently difficult to quantify and are beyond the scope of this study.

In summary, the hybrid models examined in this thesis do not demonstrate superior predictive performance relative to their standalone counterparts. However, considering the limitations of the thesis, this does not mean that the hybrid models presented here are ineffective – with alternative data structures, extended features, or revised parameters, hybrid models may indeed prove more effective in future research.

8. The Use of AI tools

During the writing of this thesis, large language models (LLMs) were used in the key areas described in this section. Among the language models, I primarily used OpenAI's ChatGPT 4o-mini model and Meta's Llama 3.1 8B model in LM studio.

The LLMs were applied in four main areas, as detailed below. First and most frequently, I used the LLMs to correct linguistic expressions, spelling, and grammatical errors, and to translate certain parts into a more academic or scientific tone whenever I felt that my own text could be improved in this sense. This helped enhance the overall clarity and quality of the thesis.

Secondly, they were used to write, correct, review, or improve R codes. I also often asked for explanations of specific lines, commands, or attributes within the code to deepen my understanding.

Thirdly, I used LLMs to help summarize the essence of the machine learning models presented in the thesis in order to ensure a clearer and more concise understanding. For more complex models – especially LSTM, which was not covered extensively during my studies – I found it useful to include a short and focused summary that complemented the knowledge I had acquired from various papers.

Lastly, following the literature review and the conceptual development of the thesis (e.g., in the application and structure of hybrid models), I used the LLMs to check and validate these concepts and assumptions. In certain cases – such as with the LSTM models – I also asked for suggestions on how to improve or extend the models, contributing to a more robust modeling strategy.

9. Acknowledgements

I would like to express my sincere gratitude, first and foremost, to my family, whose strong support has accompanied me throughout my academic journey. I am also deeply thankful to my supervisor, Mr. Lieven Baele, for his invaluable guidance during the development of this thesis and for providing me a deeper insight into this topic. Finally, I extend my appreciation to my university friends for their encouragement and the hours we spent together in the library.

10. Bibliography

- Abdulrahim, H., Alshibani, S. M., Ibrahim, O., & Elhag, A. A. (2025). Prediction OPEC oil price utilizing long short-term memory and multi-layer perceptron models. Alexandria Engineering Journal, 110, 607–612. https://doi.org/10.1016/j.aej.2024.10.057
- Abid, A., Souissi, N., Béjaoui, A., & Jeribi, A. (2025). Predicting Commodity Prices amidst Russia-Ukraine war: Some Evidences from Artificial Intelligence Models. Finance Research Open, 100023. https://doi.org/10.1016/j.finr.2025.10002
- Akshayaa, R. N., S A, S., & R, R. (2025). Session and energy forecasting for electric vehicle charging station using custom weighted ensemble machine learning. Results in Engineering, 26, 105104. https://doi.org/10.1016/j.rineng.2025.10510
- Ao, S.-I., & Fayek, H. (2023). Continual Deep Learning for Time Series Modeling. Sensors, 23(16), 7167. https://doi.org/10.3390/s2316716
- Armstrong, J. S. (Ed.). (2001). Principles of Forecasting: A Handbook for Researchers and Practitioners (Vol. 30). Springer US. https://doi.org/10.1007/978-0-306-47630-
- Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:101093340432
- Brigida, M. (2014). The switching relationship between natural gas and crude oil prices. Energy Economics, 43, 48–55. https://doi.org/10.1016/j.eneco.2014.01.01
- Brooks, C. (2019). Introductory Econometrics for Finance (4th ed.). Cambridge University Press. https://doi.org/10.1017/978110852487
- Brown, S. P. A., & Yttcel, M. K. (2008). What Drives Natural Gas Prices? The Energy Journal, 29(2), 45–60. https://doi.org/10.5547/ISSN0195-6574-EJ-Vol29-No2-
- Caldara, D., & Iacoviello, M. (2022). Measuring Geopolitical Risk. American Economic Review, 112(4), 1194–1225. https://doi.org/10.1257/aer.2019182

- Čeperić, E., Žiković, S., & Čeperić, V. (2017). Short-term forecasting of natural gas prices using machine learning and feature selection algorithms. Energy, 140, 893–900. https://doi.org/10.1016/j.energy.2017.09.02
- Chai, J., Xing, L.-M., Zhou, X.-Y., Zhang, Z. G., & Li, J.-X. (2018). Forecasting the WTI crude oil price by a hybrid-refined method. Energy Economics, 71, 114–127. https://doi.org/10.1016/j.eneco.2018.02.00
- Dickey, D. A., & Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. Journal of the American Statistical Association, 74(366a), 427–431
- Dong, X., Guo, W., Zhou, C., Luo, Y., Tian, Z., Zhang, L., Wu, X., & Liu, B. (2024). Hybrid model for robust and accurate forecasting building electricity demand combining physical and data-driven methods. Energy, 311, 133309. https://doi.org/10.1016/j.energy.2024.133309
- eia.gov
- Erdős, P. (2012). Have oil and gas prices got separated? Energy Policy, 49, 707–718. https://doi.org/10.1016/j.enpol.2012.07.022
- Geng, J.-B., Ji, Q., & Fan, Y. (2016). The behaviour mechanism analysis of regional natural gas prices: A multi-scale perspective. Energy, 101, 266–277. https://doi.org/10.1016/j.energy.2016.02.045
- globalenergymonitor.org
- Guan, K., & Gong, X. (2023). A new hybrid deep learning model for monthly oil prices forecasting. Energy Economics, 128, 107136. https://doi.org/10.1016/j.eneco.2023.107136
- Günther, M., & Nissen, V. (2021). Effects of the Henry Hub Price on U.S. LNG Exports and on Gas Flows in Western Europe. Gases, 1(2), 68–79. https://doi.org/10.3390/gases1020006
- Herrera, G. P., Constantino, M., Tabak, B. M., Pistori, H., Su, J.-J., & Naranpanawa, A. (2019). Long-term forecast of energy commodities price using machine learning. Energy, 179, 214–221. https://doi.org/10.1016/j.energy.2019.04.077
- Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. IEEE. https://www.researchgate.net/profile/Y-Bengio/publication/2839938_Gradient_Flow_in_Recurrent_Nets_the_Difficulty_of_L

earning_Long-Term_Dependencies/links/546cd26e0cf2193b94c577c2/Gradient-Flow-in-Recurrent-Nets-the-Difficulty-of-Learning-Long-Term-Dependencies.pdf

- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
- Hyndman, R., & Athanasopoulos, G. (2018). Forecasting: Principles and Practice (2nd ed.). OTexts. chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://d1wqtxts1xzle7.cloudfront.net/64659947/Athanasopoulos__George__Hyndman__Rob_J._-_Forecasting__Principles_and_Practice_%282018%29-libre.pdf?1602522707=&response-content-disposition=inline%3B+filename%3DForecasting_Principles_and_Practice.pdf&Expires=1750159165&Signature=V8dnN6bNI~8B2d6ehcf4DPMRfvSpz557-fSmE1e~rFMJkLIqhReS876HGShfWZJcpEe6xTaLovC1aKe9tDgP4ggc0XD29dtF3hf4~lCzJsJs7zOOHAolfzAZLfFi4VdFmKMz04gnCAWB~~tc0~5wytqzNEU9J1PmLlI4~6I9nAyy96OXmGbrD0npHBDerGj~5JJTpzsByL~DDeNv4wBQm7gJukiEQDWoTB4oRYBb5QR-XTUEJ9z4C4O4oNRRBNgcAx3ym3cb690JVLJHYmJ9TKfWUd8Re1ZZ89rNldMlkcL~zNYyU6Y~B0B~RymMzbKatPQ3VltCKYMHOjA9DeIdAQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA
- Islam, K. I., Elias, E., Carroll, K. C., & Brown, C. (2023). Exploring Random Forest Machine Learning and Remote Sensing Data for Streamflow Prediction: An Alternative Approach to a Process-Based Hydrologic Modeling in a Snowmelt-Driven Watershed. Remote Sensing, 15(16), 3999. https://doi.org/10.3390/rs15163999
- Istiake Sunny, Md. A., Maswood, M. M. S., & Alharbi, A. G. (2020). Deep Learning-Based Stock Price Prediction Using LSTM and Bi-Directional LSTM Model. 2020 2nd Novel Intelligent and Leading Emerging Sciences Conference (NILES), 87–92. https://doi.org/10.1109/NILES50944.2020.9257950
- Jang, G., Seo, J., & Lee, H. (2025). Analyzing the impact of COVID-19 on seasonal infectious disease outbreak detection using hybrid SARIMAX-LSTM model. Journal of Infection and Public Health, 18(7), 102772. https://doi.org/10.1016/j.jiph.2025.102772

- Ji, Q., Zhang, H.-Y., & Geng, J.-B. (2018). What drives natural gas prices in the United States? – A directed acyclic graph approach. Energy Economics, 69, 79–88. https://doi.org/10.1016/j.eneco.2017.11.002
- Jia, J., & Yan, R. (2023). Research on Short-Term Forecast of Product Demand Based on SARIMAX Model. 2023 International Conference on Applied Physics and Computing (ICAPC), 449–454. https://doi.org/10.1109/ICAPC61546.2023.00088
- Kashif, K., & Ślepaczuk, R. (2025). LSTM-ARIMA as a hybrid approach in algorithmic investment strategies. Knowledge-Based Systems, 320, 113563. https://doi.org/10.1016/j.knosys.2025.113563
- Kaur, M., & Mohta, A. (2019). A Review of Deep Learning with Recurrent Neural Network. 2019 International Conference on Smart Systems and Inventive Technology (ICSSIT), 460–465. https://doi.org/10.1109/ICSSIT46314.2019.8987837
- Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. Journal of Econometrics, 54(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y
- Lin, Y., Dai, D., Yu, Y., Li, Z., Huang, W., Zhao, L., & Xing, H. (2025). Forecasting natural gas prices using a novel hybrid model: Comparative study of different sliding windows. Energy, 329, 136607. https://doi.org/10.1016/j.energy.2025.136607
- Macron calls out US for double standards on gas prices to EU. (2022, October 22). https://english.almayadeen.net/news/politics/macron-calls-out-us-for-double-standards-on-gas-prices-to-eu?utm_source=chatgpt.com
- Manaswi, N. K. (2018). RNN and LSTM. In N. K. Manaswi, Deep Learning with Applications Using Python (pp. 115–126). Apress. https://doi.org/10.1007/978-1-4842-3516-4_9
- McHugh, C., Coleman, S., Kerr, D., & McGlynn, D. (2019). Forecasting Day-ahead Electricity Prices with A SARIMAX Model. 2019 IEEE Symposium Series on Computational Intelligence (SSCI), 1523–1529. https://doi.org/10.1109/SSCI44817.2019.9002930
- Mu, X. (2007). Weather, storage, and natural gas price dynamics: Fundamentals and volatility. Energy Economics, 29(1), 46–63. https://doi.org/10.1016/j.eneco.2006.04.003
- Olah, C. (2015). Understanding LSTM Networks. Colah's Blog. https://colah.github.io/posts/2015-08-Understanding-LSTMs/

- Omarzai, F. (2024). Random Forest In-Depth. Medium. https://medium.com/@fraidoonomarzai99/random-forest-in-depth-f0556817c40b
- Panagiotidis, T., & Rutledge, E. (2007). Oil and gas markets in the UK: Evidence from a cointegrating approach. Energy Economics, 29(2), 329–347. https://doi.org/10.1016/j.eneco.2006.10.013
- Ray, S., Lama, A., Mishra, P., Biswas, T., Sankar Das, S., & Gurung, B. (2023). An ARIMA-LSTM model for predicting volatile agricultural price series with random forest technique. Applied Soft Computing, 149, 110939. https://doi.org/10.1016/j.asoc.2023.110939
- Roberts, G. (2019). Revisiting the Drivers of Natural Gas Prices. A replication study of Brown & Yücel (The Energy Journal, 2008). International Journal for Re-Views in Empirical Economics, 20192. https://doi.org/10.18718/81781.11
- Safari, A., & Davallou, M. (2018). Oil price forecasting using a hybrid model. Energy, 148, 49–58. https://doi.org/10.1016/j.energy.2018.01.007
- Sarangi, P. K., Singh, E., Kaushal, A., Sharma, B., Dutta, M., & Dubey, V. P. (2025). Analysing the Fluctuations in Commodity Prices and Forecasting the Future Directions Using Machine Learning Techniques. Procedia Computer Science, 259, 1827–1836. https://doi.org/10.1016/j.procs.2025.04.138
- Seba, D., & Belaide, K. (2024). Forecasting infection fatality rate of COVID-19: Measuring the efficiency of several hybrid models. Russian Journal of Infection and Immunity, 14(2), 313–319. https://doi.org/10.15789/2220-7619-FIF-1754
- Serletis, A., & Herbert, J. (1999). The message in North American energy prices. Energy Economics, 21(5), 471–483. https://doi.org/10.1016/S0140-9883(99)00015-
- Sheng, F., & Jia, L. (2020). Short-Term Load Forecasting Based on SARIMAX-LSTM. 2020 5th International Conference on Power and Renewable Energy (ICPRE), 90–94. https://doi.org/10.1109/ICPRE51194.2020.923311
- Sherstinsky, A. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena, 404, 132306. https://doi.org/10.1016/j.physd.2019.132306
- Siami-Namini, S., Tavakoli, N., & Siami Namin, A. (2018). A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 1394–1401. https://doi.org/10.1109/ICMLA.2018.00227

- Stajic, L., Đorđević, B., Ilić, S., & Brkić, D. (2021). The volatility of natural gas prices – Structural shocks and influencing factors. Revista Internacional de Métodos Numéricos Para Cálculo y Diseño En Ingeniería, 37(4). https://doi.org/10.23967/j.rimni.2021.12.002
- Starmer, J. (Director). (2022). Long Short-Term Memory (LSTM), Clearly Explained [Mp4]. https://www.youtube.com/watch?v=YCzL96nL7j0
- Su, M., Zhang, Z., Zhu, Y., Zha, D., & Wen, W. (2019). Data Driven Natural Gas Spot Price Prediction Models Using Machine Learning Methods. Energies, 12(9), 1680. https://doi.org/10.3390/en12091680
- Swanson, D. A., & Tayman, J. (1999). Onthe validity of MAPE as a measure of population forecast accuracy. 18, 229–322.
- Topal, K. H., Kizilarslan, S., Bulbul, H., & Akay, E. Ç. (2019). Forecasting of Turkish housing price index: ARIMA, random forest, ARIMA-random forest. Pressacademia, 10(10), 7–11. https://doi.org/10.17261/Pressacademia.2019.1134
- Wen, F., Gong, X., & Cai, S. (2016). Forecasting the volatility of crude oil futures using HAR-type models with structural breaks. Energy Economics, 59, 400–413. https://doi.org/10.1016/j.eneco.2016.07.014
- What is random forest? (n.d.). IBM. https://www.ibm.com/think/topics/random-forest
- Wilson, G. T. (2016). Time Series Analysis: Forecasting and Control, 5th Edition, by George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel and Greta M. Ljung, 2015. Published by John Wiley and Sons Inc., Hoboken, New Jersey, pp. 712. ISBN: 978-1-118-67502-1. Journal of Time Series Analysis, 37(5), 709–711. https://doi.org/10.1111/jtsa.12194
- wikipedia.org/wiki/List_of_Louisiana_hurricanes_(2000-present)
- Wu, Y.-X., Wu, Q.-B., & Zhu, J.-Q. (2019). Improved EEMD-based crude oil price forecasting using LSTM networks. Physica A: Statistical Mechanics and Its Applications, 516, 114–124. https://doi.org/10.1016/j.physa.2018.09.120
- Xu, Y., Liu, T., Fang, Q., Du, P., & Wang, J. (2025). Crude oil price forecasting with multivariate selection, machine learning, and a nonlinear combination strategy. Engineering Applications of Artificial Intelligence, 139, 109510. https://doi.org/10.1016/j.engappai.2024.109510

- Yang, Y., Guo, J., Sun, S., & Li, Y. (2021). Forecasting crude oil price with a new hybrid approach and multi-source data. Engineering Applications of Artificial Intelligence, 101, 104217. https://doi.org/10.1016/j.engappai.2021.104217
- Yu, L., Liang, S., Chen, R., & Lai, K. K. (2022). Predicting monthly biofuel production using a hybrid ensemble forecasting methodology. International Journal of Forecasting, 38(1), 3–20. https://doi.org/10.1016/j.ijforecast.2019.08.014
- Zaretskaya, V. (2025). The United States remained the world's largest liquefied natural gas exporter in 2024. EIA. https://www.eia.gov/todayinenergy/detail.php?id=64844#
- Zhao, Y., Li, J., & Yu, L. (2017). A deep learning ensemble approach for crude oil price forecasting. Energy Economics, 66, 9–16. https://doi.org/10.1016/j.eneco.2017.05.023

11. Appendix

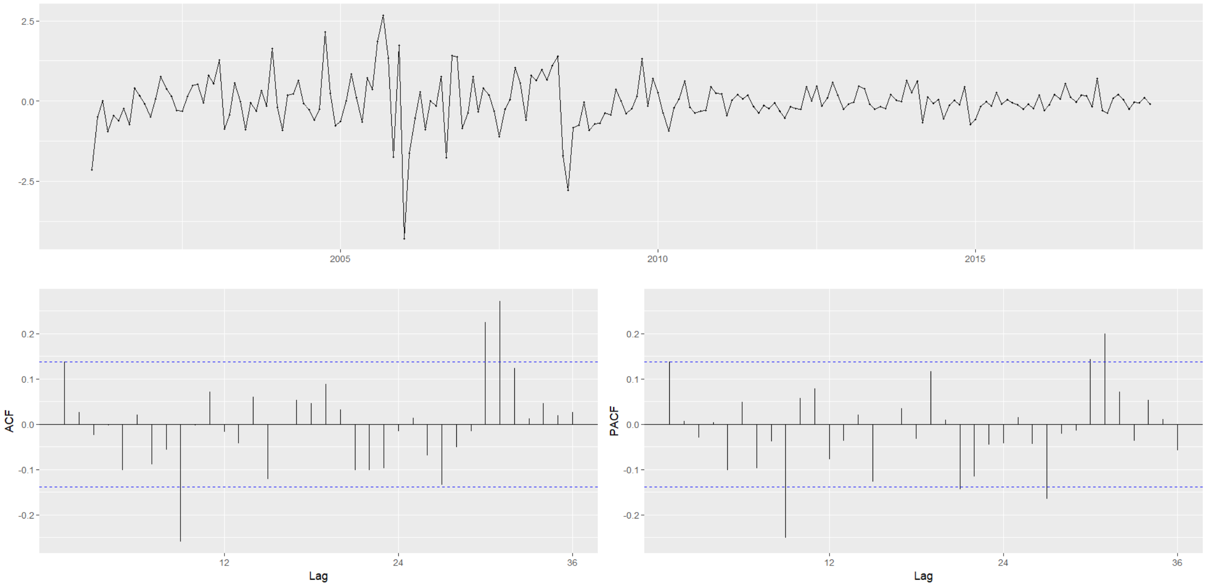


Table 7: ACF and PACF plot for y_train data

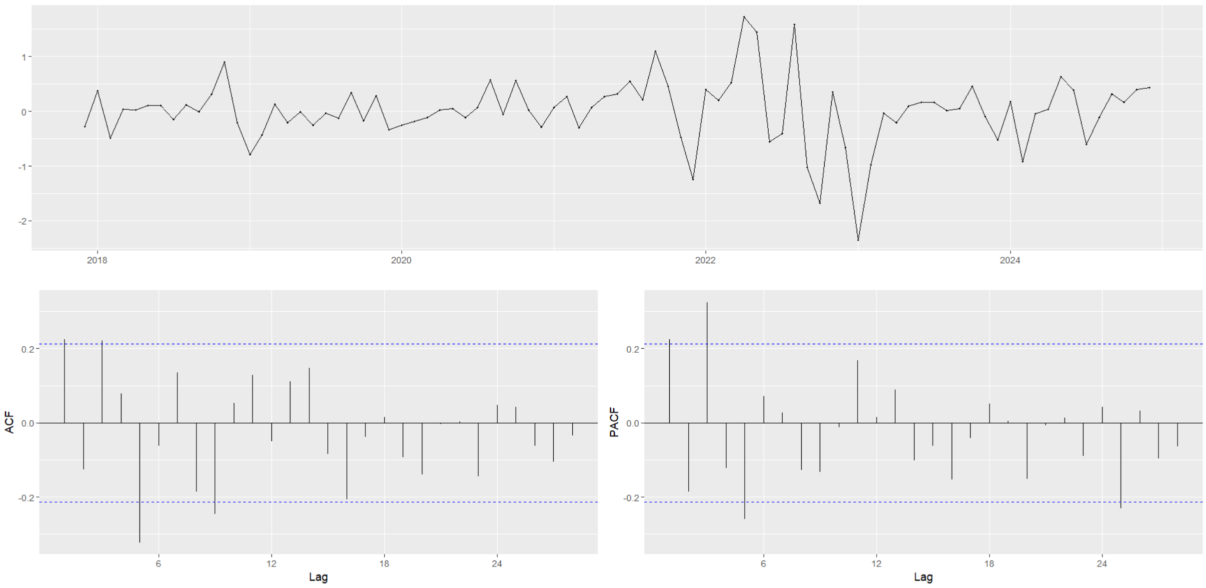


Table 8: ACF and PACF plot for y_test data

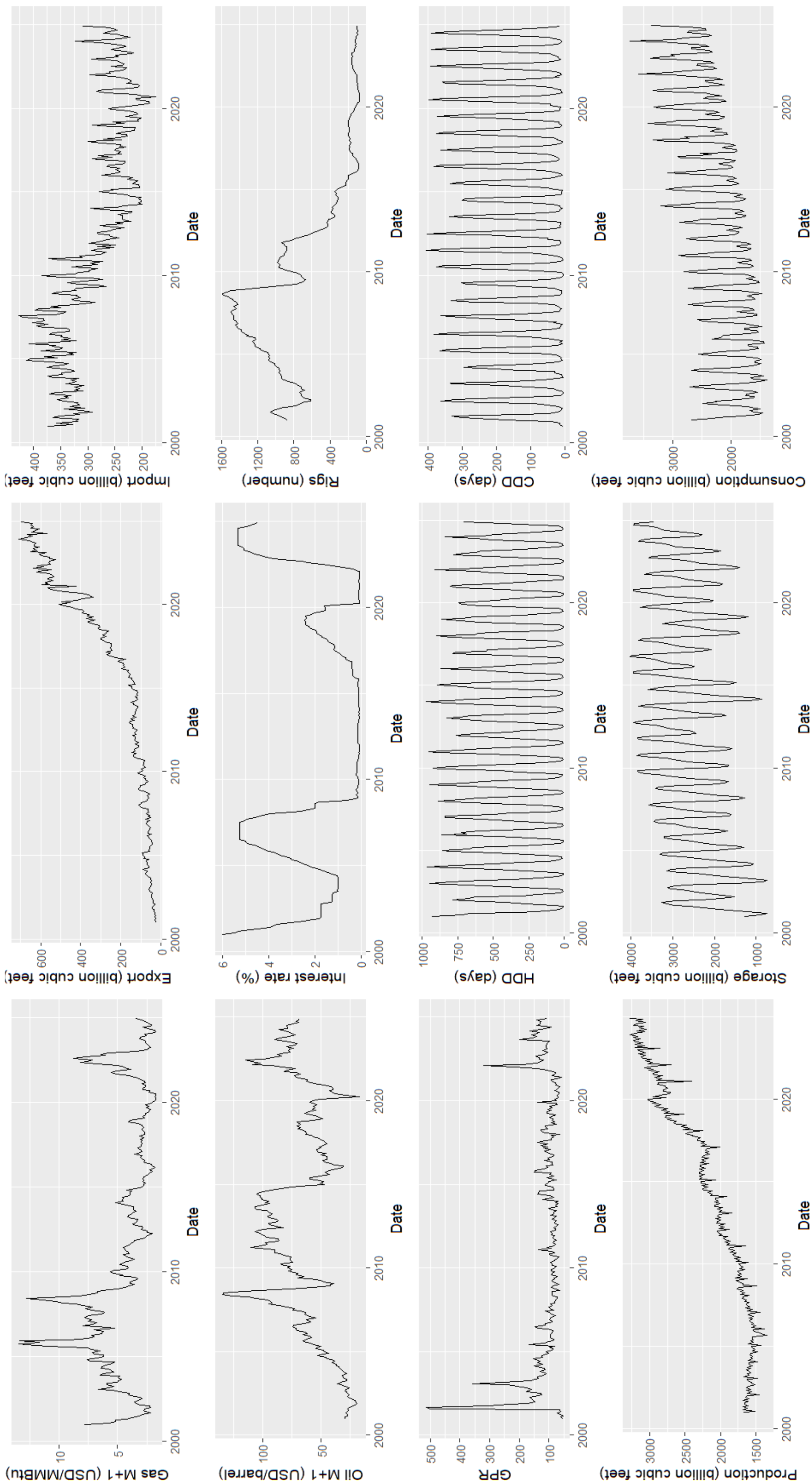


Table 9: Time-series plots of different variables. Please note, that export and import are considered separately here

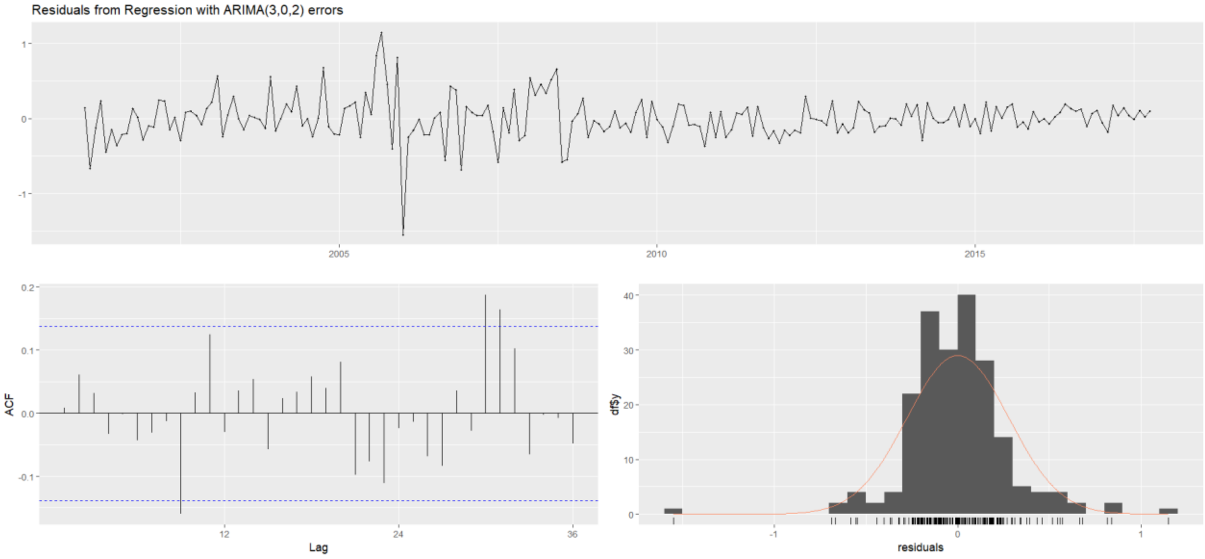


Table 10: Ljung-Box test statistics for SARIMAX-model Please note that certain spikes in ACF plot are due to the financial crisis in 2008, when commodity prices skyrocketed, and the Katrina and Rita hurricanes in 2005 that seriously hit the state and thus sent the prices above 10 USD for months long

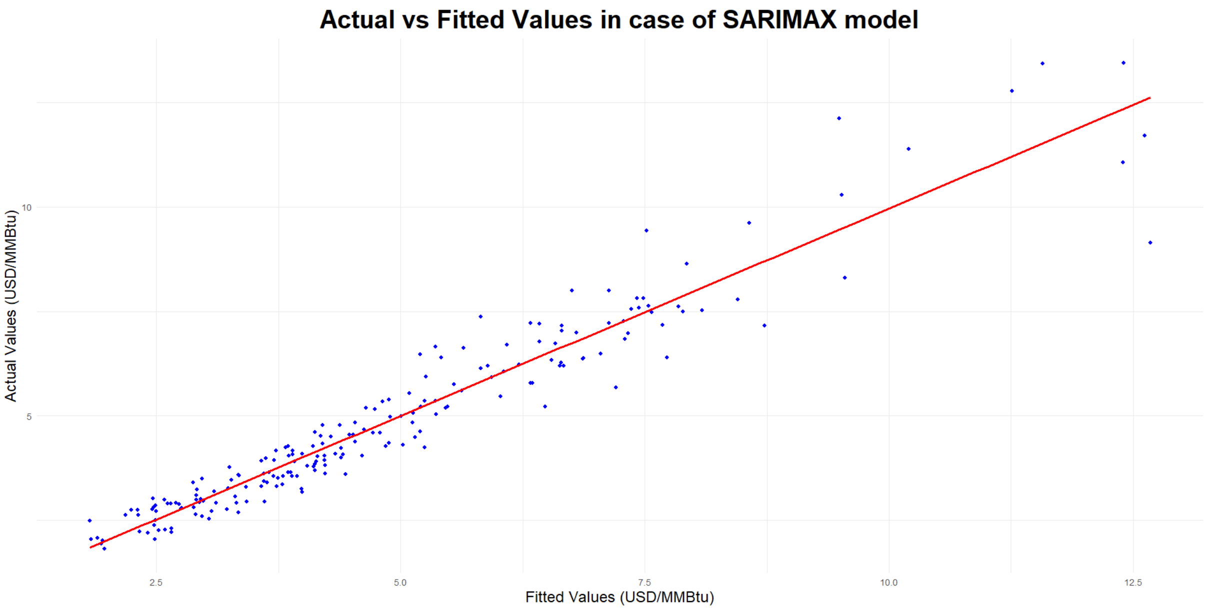


Figure 11: Actual and Fitted values in case of SARIMAX

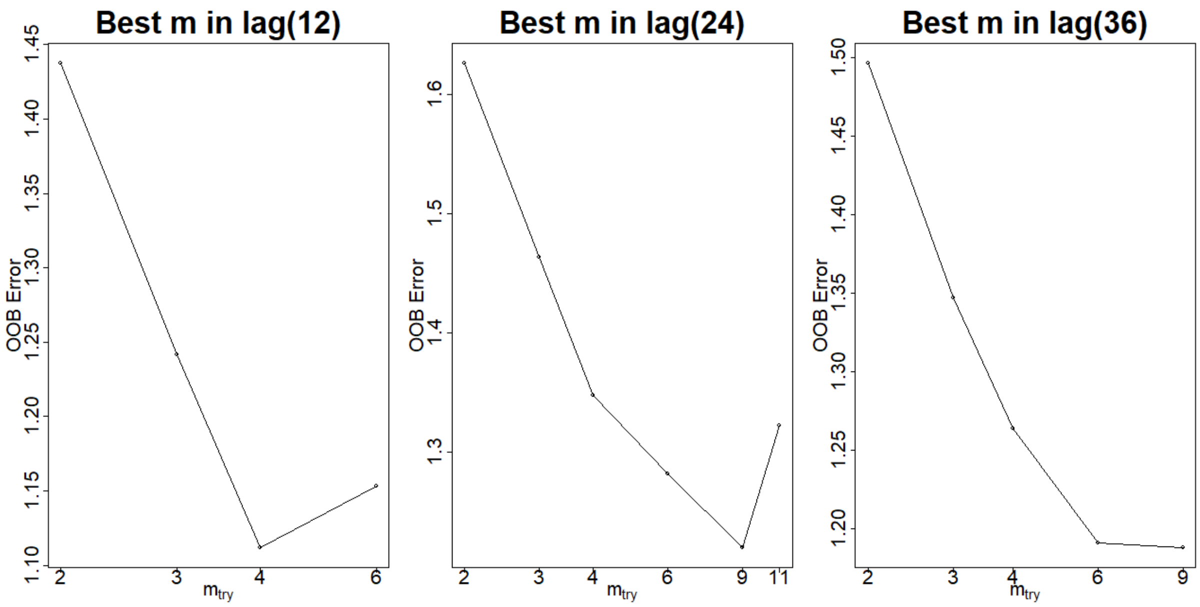


Figure 12: OOB errors and best m in case of Random Forest models

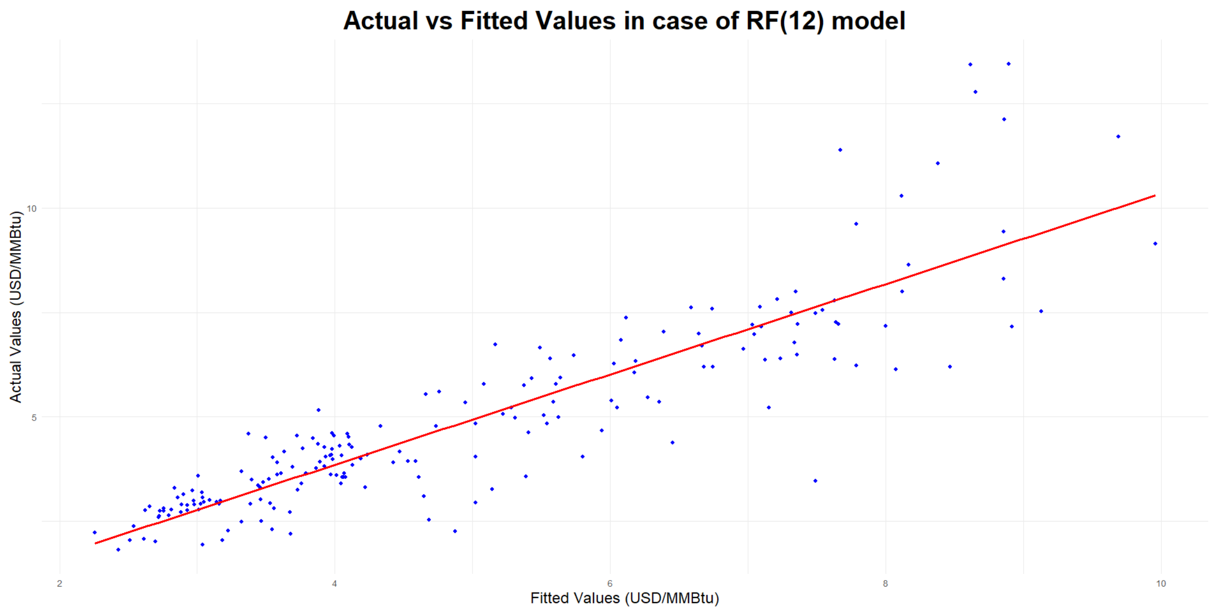


Figure 13: Actual and Fitted values in case of RF(12)

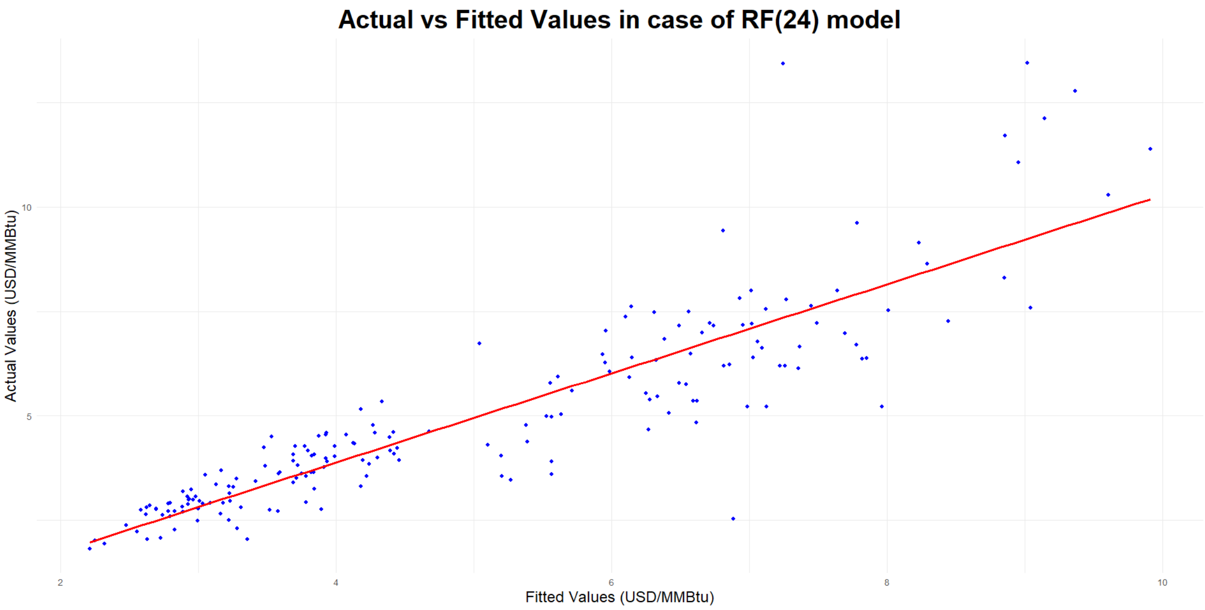


Figure 14: Actual and Fitted values in case of RF(24)

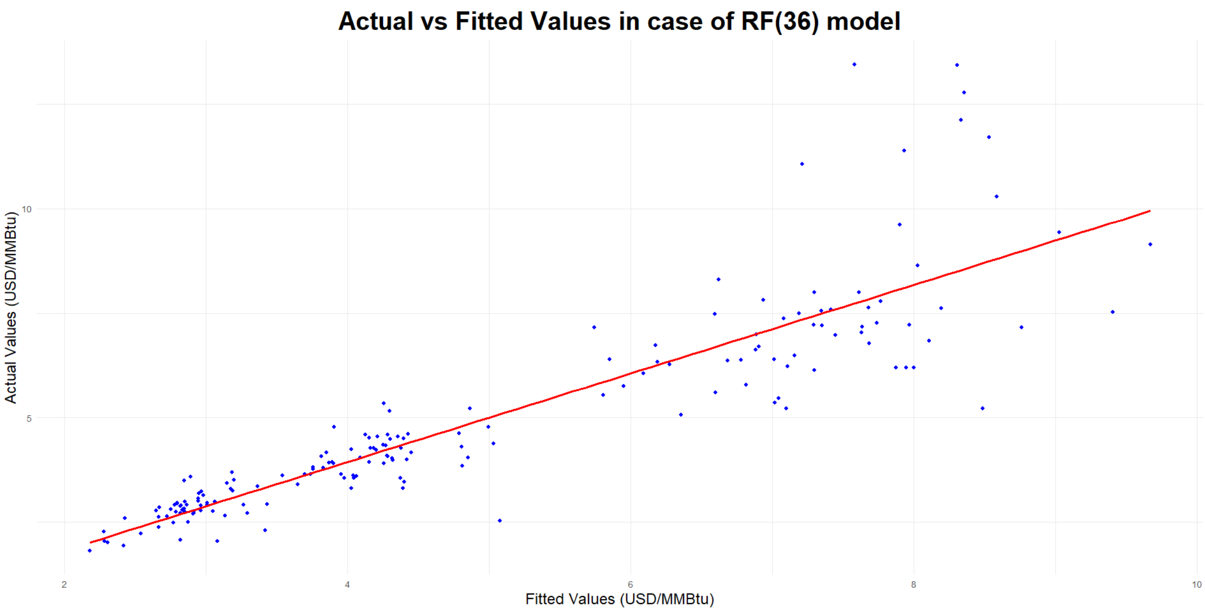


Figure 15: Actual and Fitted values in case of RF(36)

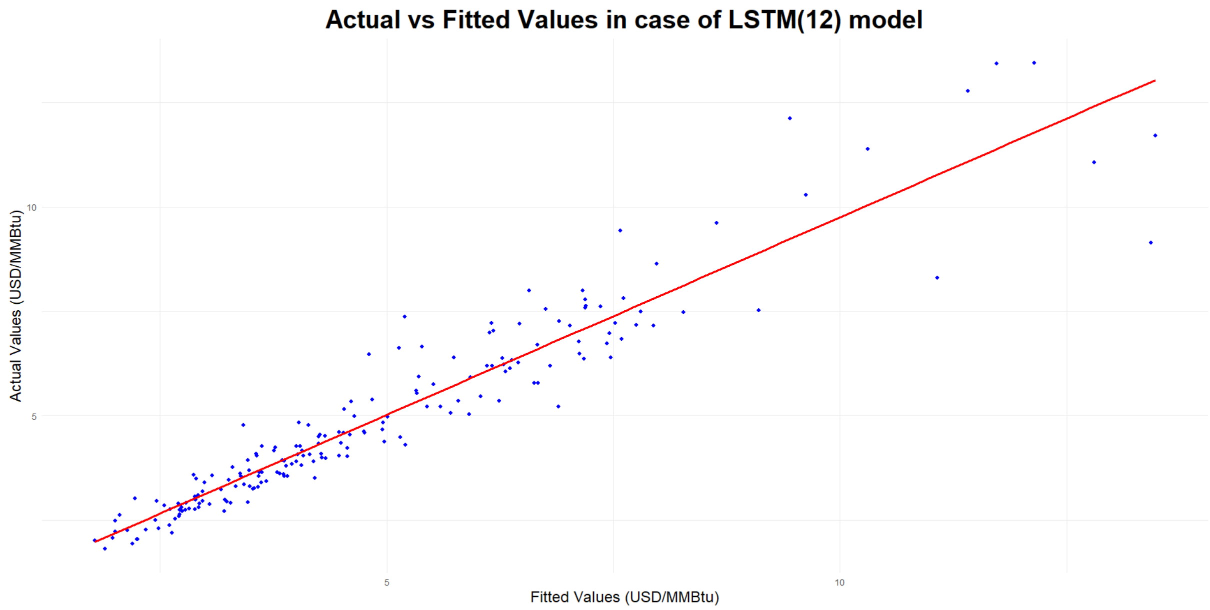


Figure 16: Actual and Fitted values in case of LSTM(12)

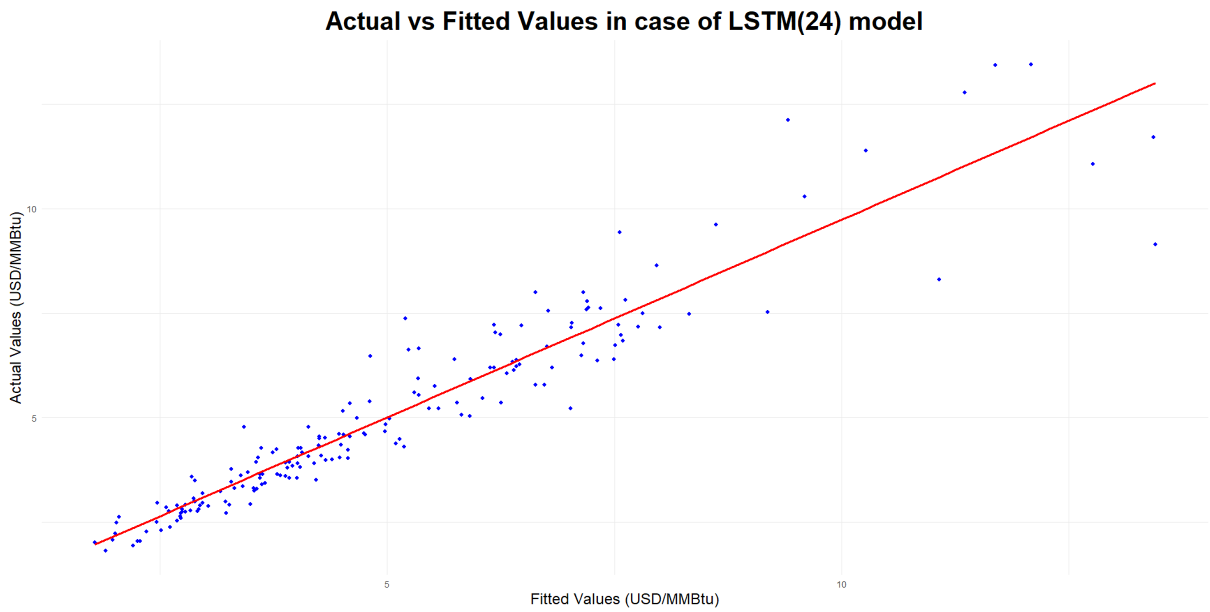


Figure 17: Actual and Fitted values in case of LSTM(24)

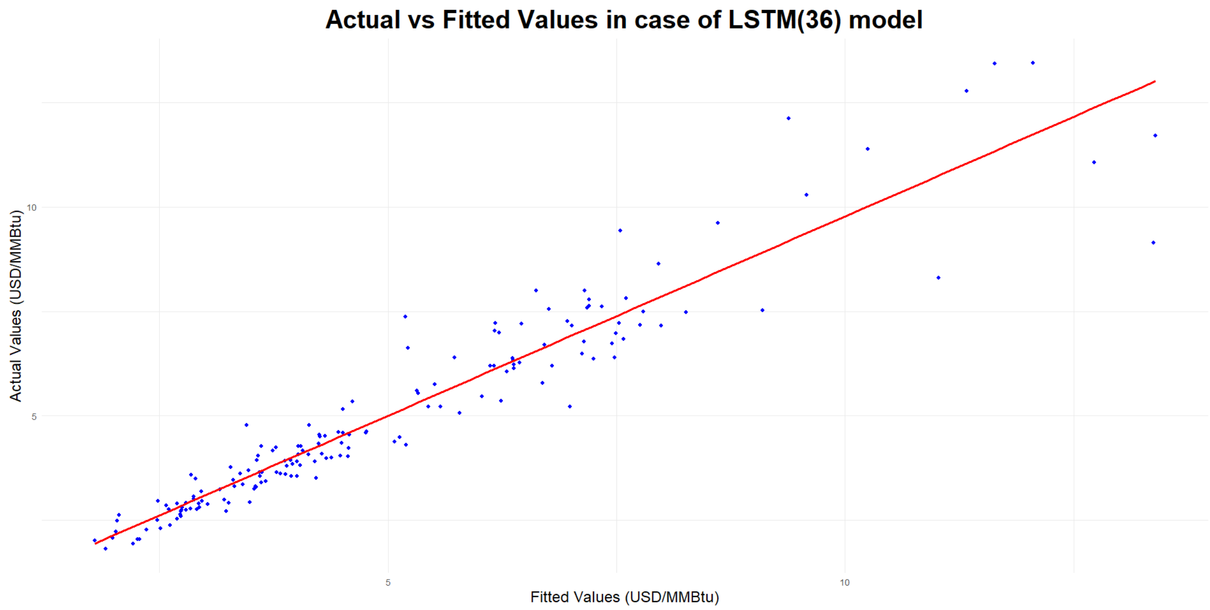


Figure 18: Actual and Fitted values in case of LSTM(36)

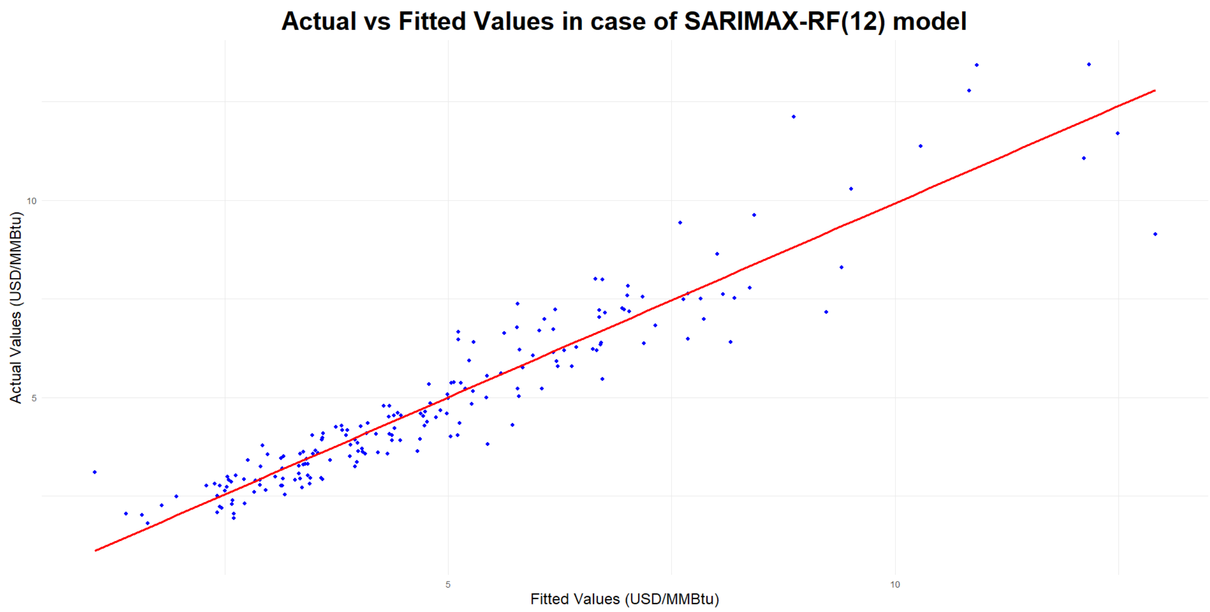


Figure 19: Actual and Fitted values in case of SARIMAX-RF(12)

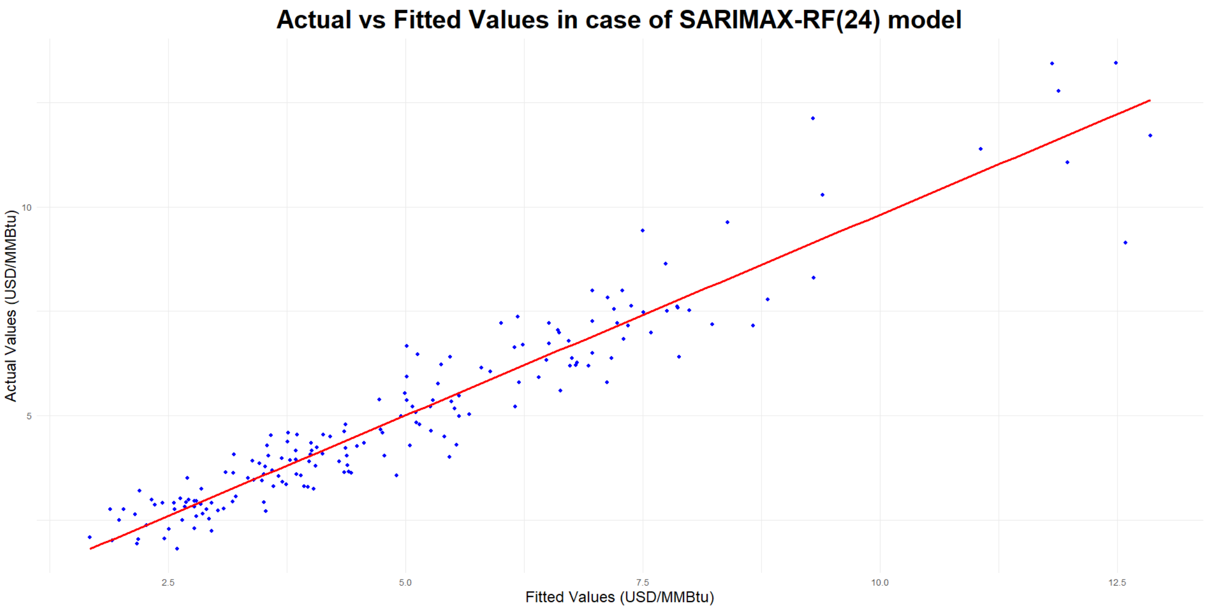


Figure 20: Actual and Fitted values in case of SARIMAX-RF(24)

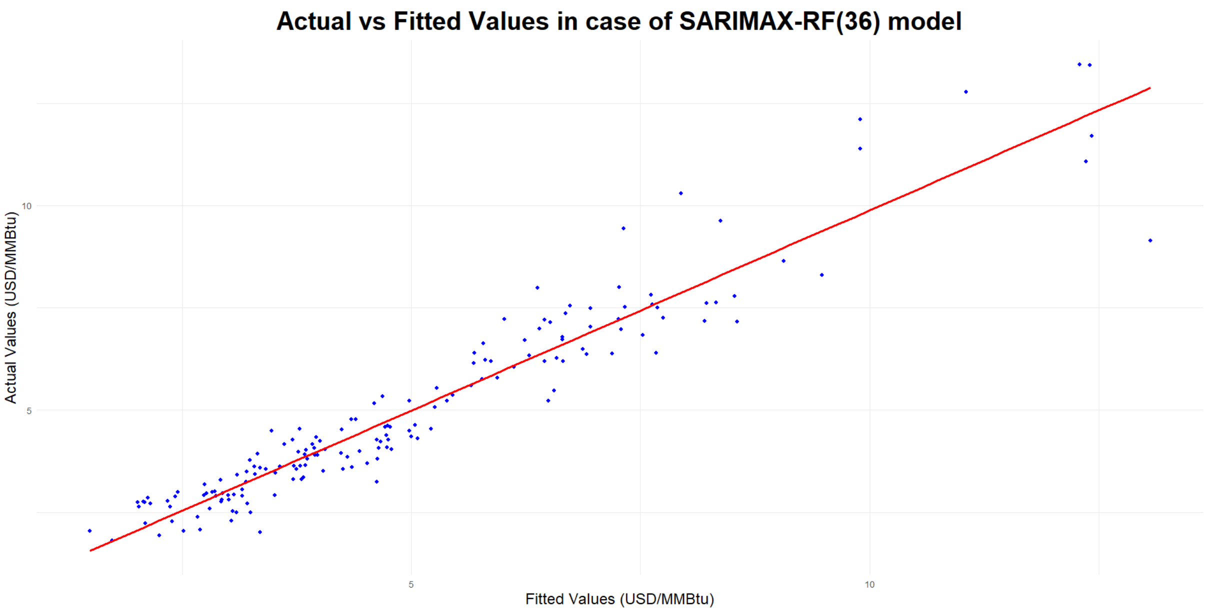


Figure 21: Actual and Fitted values in case of SARIMAX-RF(36)

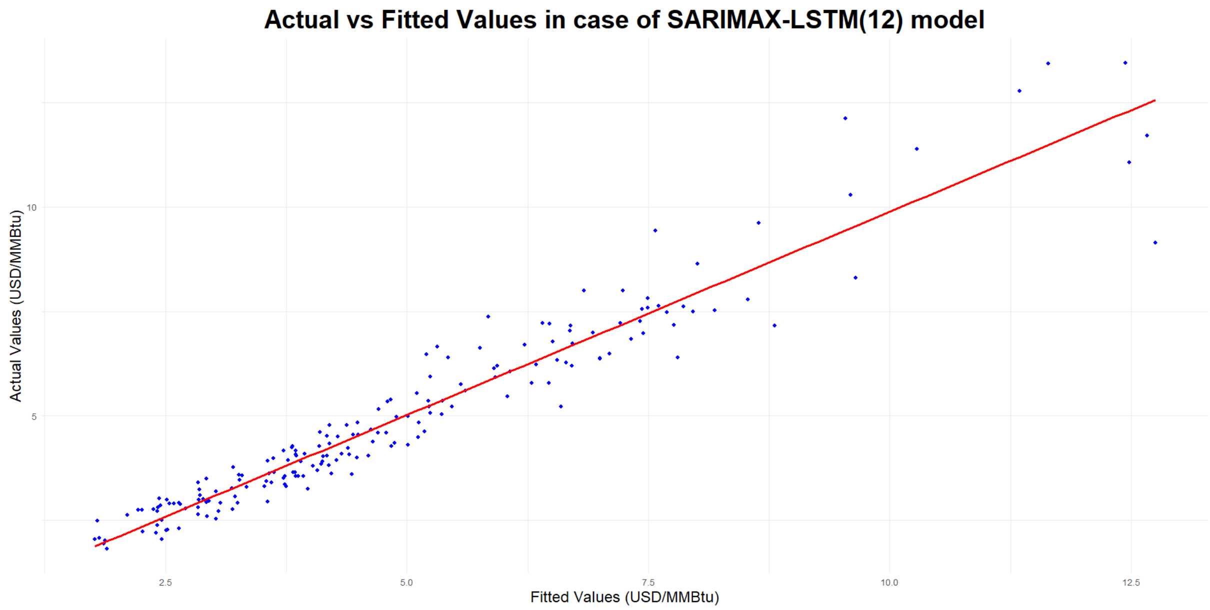


Figure 22: Actual and Fitted values in case of SARIMAX-LSTM(12)

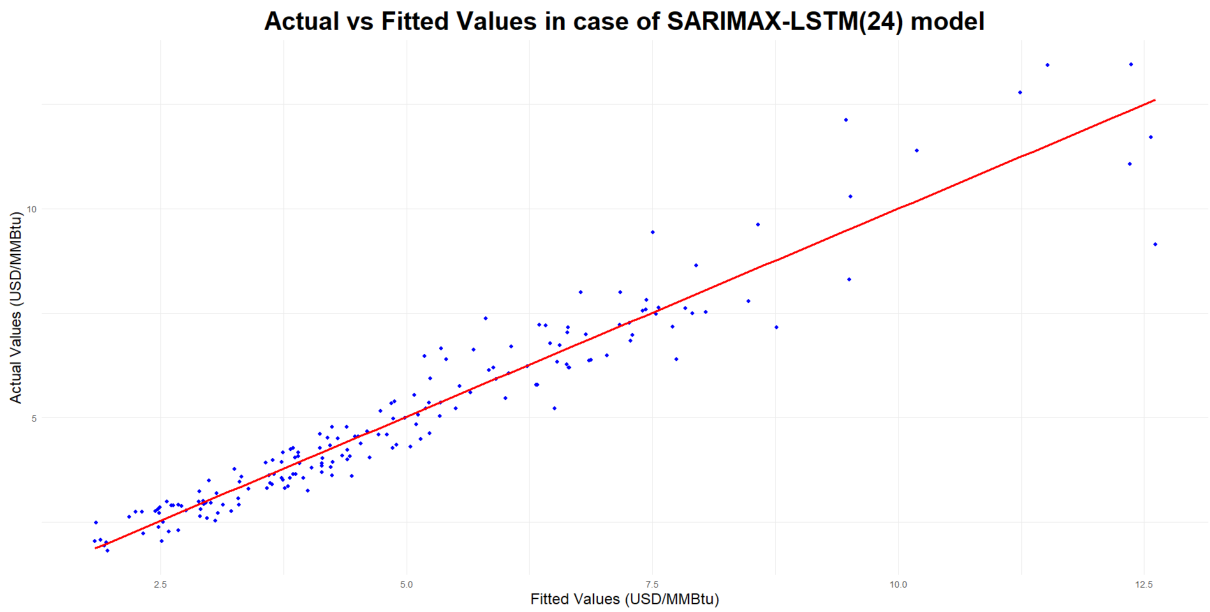


Figure 23: Actual and Fitted values in case of SARIMAX-LSTM(24)

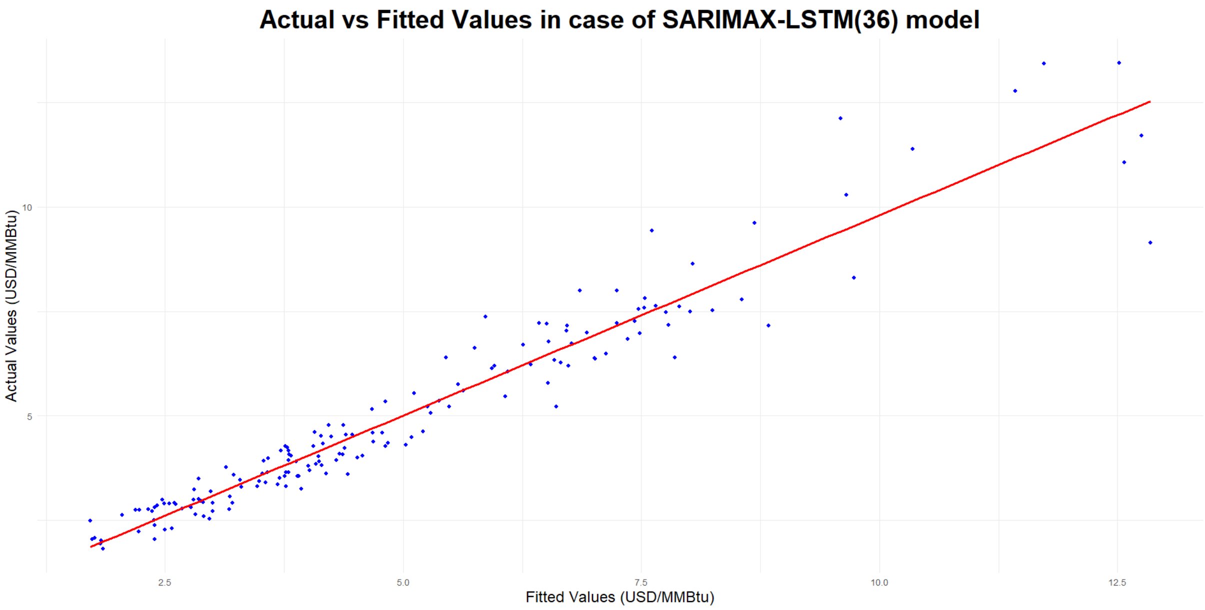


Figure 24: Actual and Fitted values in case SARIMAX-LSTM(36)