What Does the Commodity Market Tell Us about the Environment? A Case Study On Oil Futures, Gasoline, and Carbon Dioxide Concentration

Abstract--**Crude Oil is one of the most essential exhaustible energy that keeps the world functioning. The price movements in crude oil have great influence on various industries and international trades. In this paper, we propose approaches for gasoline price and Carbon Dioxide concentration predictions based on crude oil futures. To evaluate the forecasting ability, we compared several statistical and machine learning models and presented our findings in this study.**

Index terms--**Granger Causality, Cointegration Test, PACF, AR, VAR, Ridge Regression, Random Forest Regression, Gradient Boost Regression, LSTM**

I. INTRODUCTION

Historically, Crude Oil accounted for over one third of the world energy consumption. Today, Crude Oil remains the leading fuel used in the world. Agriculture, Shipping, Airline industries rely heavily on Crude Oil. Due to its importance, Crude Oil Futures is one of the most traded commodities. Besides the consumer gasoline prices, fluctuating oil prices affects the world's environment. When the price of oil decreases, consumers are more willing to use gasoline and less incentivised to choose clean or renewable energy. Consequently, the drop in oil prices could lead to the increase of carbon emission. In this paper, we proposed both statistical and machine learning techniques to investigate the questions that whether Crude Oil prices and futures contain predictive powers on retail gasoline prices and the concentration of carbon dioxide.

II. DATA SELECTION

A . Oil futures vs. Gasoline Price

In this study, we used weekly oil futures prices from Yahoo Finance, and weekly gasoline prices from the Energy Information Administration (EIA) dated from September 2000 to January 2021. We initially looked into the twenty years trend of both oil futures and gasoline prices. Because of the great value difference between oil futures and gasoline prices, logarithmic scaling is applied. The following graph shows timeline trends in 20 years after scaling.

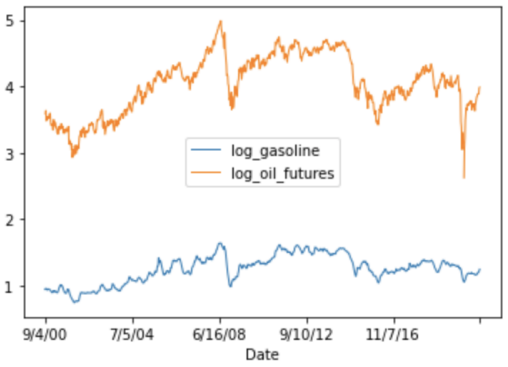

Fig. 1: Log Transformed Gas vs Oil Futures

To better explain the variation of gasoline price, we included additional variables such as seasonality, financial market uncertainty and inflation rate (represented by gold price and USD index collected on Yahoo Finance and weekly unemployment insurance claim from the U.S. Department of Labor), refined oil products (diesel, heating oil and jet fuel futures price from EIA and Yahoo Finance), and supply and demand (ending stock of gasoline, crude oil product, and refinery operating capacity from EIA).

B. Oil futures vs. CO2 concentration

We collected weekly carbon dioxide concentrations (in units of ppm) from the National Oceanic and Atmospheric Administration dated from September 2000 to January 2021. Combined with oil futures prices, we investigated whether oil futures is a good indicator of atmospheric carbon dioxide.

To improve our model, we included seasonality and natural gas futures to help explain the evolution of CO2 concentration. Data of weekly natural gas futures is collected from Yahoo Finance, with a time frame in accordance with that of crude oil futures prices and CO2 concentration.

III. Is it possible that futures prices of crude oil tell us something about future behavior of gasoline prices?

Empirical studies show that whether commodity futures is a good predictor of spot prices depends on the type of commodity and cost of carry. According to the Federal Reserve Bank of St. Louis, futures is

great in forecasting market expectations of non-storable and perishable commodities like dairy products and eggs. However, the performance of futures varies when commodity is storable and has cost of carry in the long term. For instance, soybean futures alone is not as reliable to forecast future soybean spot prices. Crude oil falls into the category of storable commodity with cost of carry, therefore, we are curious in finding whether crude oil futures prices tell us something about gasoline prices.

STATISTICAL ANALYSIS

A. Granger Causality

We first used Granger's Causality Test to see whether oil futures Granger causes gasoline prices and the causality is uni-directional or bi-directional. If oil futures granger causes gas prices, then lags of oil futures have the ability to predict future gasoline prices beyond the predictive power of the past gasoline prices.

Table I: Granger Causality Gas & Oil Futures

	Gasoline x	Crude Oil x
Gasoline Price_y	1.0	0.0
Crude Oil_y	0.0	1.0

In Table I, the rows are response variables whereas the columns are time series used to predict, and the entries of the matrix are p values. The 0.0 in row 1 column 2 means that the p-value of crude futures granger causing gasoline price is 0, which indicates oil futures is significant in predicting gasoline prices. Furthermore, the lower left entry is also 0, meaning that gasoline price is also significant as a predictor of crude futures, the causal relationship is bi-directional. As a result, we considered the Vector autoregression (VAR) model, a statistical model used to capture the relationship between multiple time series data as they change over time.

We then used the granger causality test to find the lag of influence of oil futures price movement on gasoline price movement. We set the maxlag=15 and found that the F statistics kept declining after the one-week lag.

Granger Causality	F test	p value
Lags 1	57.7199	0.0000
Lags 2	47.6009	0.0000
Lags 3	38.1912	0.0000

Table II: F-statistic for Oil Futures Lag

Hence, the lag effect should be at least within one week, which corresponds to our assumption that today's market is highly information-sensitive, the gasoline price will respond quickly to the changing oil futures and the lag effect should not be long-term.

B. Spearman Correlation and Cointegration Test

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d_i = difference between the two ranks of each observation

n = number of observations

Correlation: We applied Kolmogorov Smirnov Test (KS Test) and found that neither oil futures nor gasoline prices follow normal distribution. Spearman correlation between oil futures price and gasoline price has coefficient of 0.9447, showing that they are indeed highly correlated.

Cointegration Test: The Granger's Causality Test suggests that there is a bi-directional causal relationship between oil futures and gasoline prices. To further analyze their relationship we conducted a Cointegration test to help establish the presence of a statistically significant connection between two time series. The result is that the p-value of crude futures causing gasoline price is about 0.0005187, which is much smaller than the threshold 0.05. We reject the null hypothesis of no cointegration and conclude they are cointegrated. This result is consistent with Granger's Causality Test that both prove there is a significant connection between crude futures price and gasoline price.

C. OLS Results

For the optimal lag, we did the ordinary least square regression for crude oil lags from 0 to 4 weeks. The below table shows crude oil futures with lag=1 and lag=2 are both significant under 5% significance level. Because we wanted to match the previous result obtained by Granger Causality, we used 1 week lag of crude oil futures in our model.

Table III: Crude Oil Lags Significance

	coeff	Std err	P > | t |
Intercept	0.9280	0.019	0.000
lag(crude oil) 0	0.0075	0.002	0.000
lag(crude oil) 1	0.0067	0.002	0.007
lag(crude oil) 2	0.0053	0.002	0.034
lag(crude oil) 3	0.0030	0.002	0.219
lag(crude oil) 4	0.0042	0.002	0.026

D. AutoRegression

We used the Partial Autocorrelation Function (PACF) plot and Autocorrelation Function plot (ACF) with maxlag=10 to determine the order of auto-correlation of gasoline price.

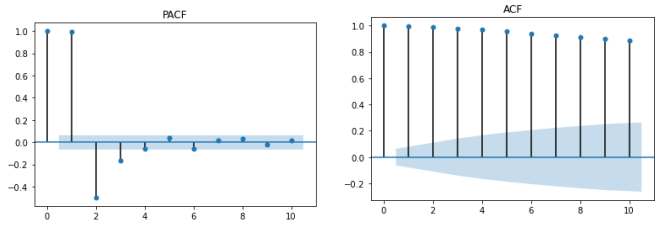

Fig. 2: PACF & ACF of Gasoline

According to the PACF plot, the first-order and second-order coefficients are high, and the coefficients after the second-order converge to 0 very quickly. According to the ACF plot, within 10 weeks, there is no obvious evidence showing that the coefficients converge to 0. Hence, we should use PACF plot to determine the order of auto-correlation of gasoline price, which is AR(2).

PREDICTION MODELS

In the following section, we present different econometric and time series models for predicting future behavior of gasoline prices using root mean squared error as a performance measure.

A. Model Section

(1) Vector autoregression (VAR): After splitting the data into training and testing sets, we first performed Augmented Dickey-Fuller (ADF) Test to test the Stationarity of both gasoline price and crude futures in our training set. However, the result showed that neither of these time series are stationary. We used the first difference transformation by taking their difference once. After the transformation, ADF Test showed that two time series became stationary. Another key step before building the model was to select the order for VAR. We relied on the AIC score in this case by choosing Order 1 through 9 to find the one which provided the smallest AIC score, and the result suggested lag order is 4. Hence, our final model is VAR(4). After fitting the model, we checked for serial correlation of errors which can be measured using the Durbin Watson's Statistic.

$$DW = \frac{\sum_{t=2}^{T}((e_t - e_{t-1})^2)}{\sum_{t=1}^{T} e_t^2}$$

This test statistic indicates there is no serial correlation of errors, which is a desired result. Finally, we present the performance of the VAR(4) model as follows.

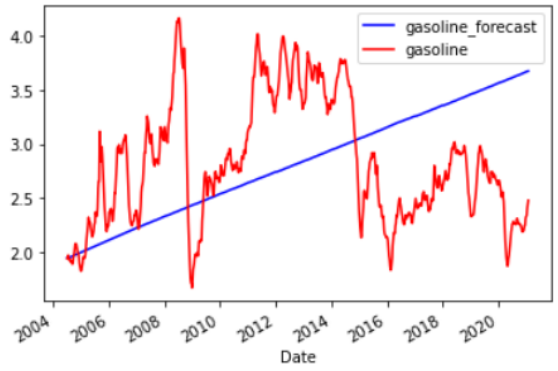

Fig. 3: VAR Model Forecast Performance

The graph shows that the blue line/gasoline forecast did not catch the up-and-down trend shown in actual gasoline prices in the past 4 years. Thus, we believe that VAR is not a good prediction model. The result given using Root Mean squared error (RMSE) score echoes that VAR is not good enough. In our case, RMSE=0.7725, however, RMSE heavily depends on the dependent variable (DV) which is gasoline price, so we calculated the range of DV which is 3.064. In this case, 0.7725 is very large compared with the range of DV, we then concluded that VAR with lag order 4 is not a good model.

(2) Ridge: After implementing the ADF test on first difference gasoline prices as well as the other dependent variables. All time series data are stationary and ready to be put into a model. We selected Ridge Regression Model as our baseline model.

(3) Random Forest Regression (RFR): As a regression model allowing multiple variables and is easy dealing with overfitting problems, Random Forest is an apparent choice. At inception, we only included the oil futures and its lag in the model. With a RMSE = 0.1786, our prediction follows the hold-out data roughly but having some volatility. We then decided to add more variables such as seasonality and the gasoline price 2 periods prior to the target gasoline price (the lag selection depends on the autoregression test stated above). The model with additional variables gives a better result with RMSE = 0.1768 and an almost perfect matching line

with the hold out data. We saw a similar result in the Gradient Boosting Regression model.

(4) Gradient Boosting Regression (GBR): GBR, an ensemble of weak decision tree models, often outperforms the RFR because it optimizes the model by itself. Following the steps above, we tried to insert different variables into GBR and check the results. Even though generating a slight difference compared with the RFR according to RMSE, GBR creates a less volatile prediction. Additionally, after adding those variables, the new model can follow the plummeted point, which means adding more variables makes the model more sensitive.

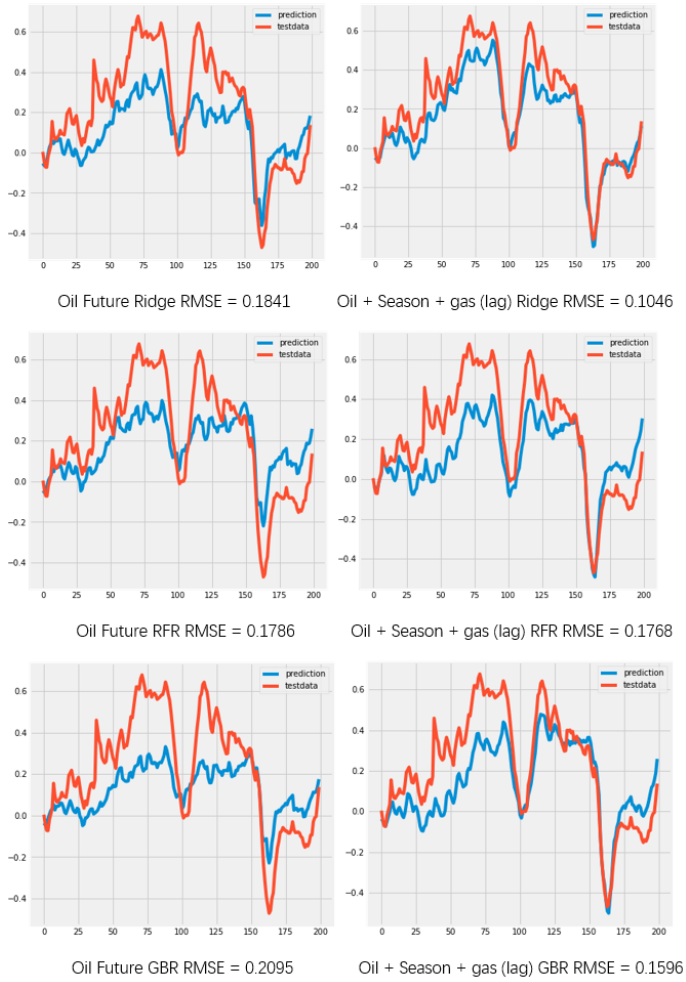

Fig. 4: GBR Model Forecast Performance

(5) Long short-term memory (LSTM): Since the gasoline price has autocorrelation, we tried the time series model LSTM to predict its future behavior. Data is splitted as 80% train set and 20% test set in chronological order. Features include crude oil futures in this week and last week since the lag effect of crude oil futures on gasoline price is one week, as mentioned before. In order to fit the three dimensional input requirement for Keras sequential model, we rescaled the data before putting them into the model. The model was built with a 1D convolutional layer with activation function ReLu. Then the output is flattened to feed into the two LSTM layers followed by a dense layer to provide the output. With 15 epochs, the RMSE of the test set is 0.0965.

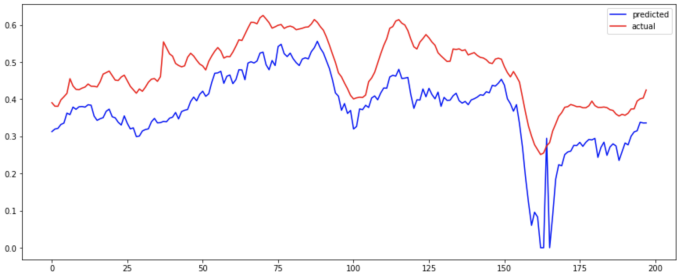

Fig. 5: Oil Futures LSTM Forecast

The graph above also shows that the predicted gasoline price (blue line) roughly matches the actual gasoline price (red line). Therefore, we conclude that LSTM is an effective model for gasoline price forecasting.

To improve the prediction, we add additional features including seasonal dummies, heating oil price, gold price, USD index, and refinery operation capacity. The model performance significantly rises with a RMSE equals 0.0328. The enhancement is also visible on the graph below.

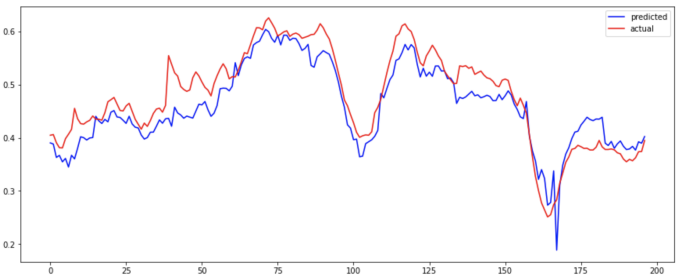

Fig. 6: Oil Futures + Additional LSTM Forecast

Interpretation:

RMSE of Models with additional variables:

Table IV: Model Performance (RMSE)

Model	Ridge	RFR	GBR	LSTM
Oil	0.1841	0.1786	0.2095	0.0965
Additional	0.1046	0.1768	0.1596	0.0328

After comparing above models, LSTM is the best model with lowest RMSE as well as a smoother and matching prediction line with the hold-out data. Also, including additional variables helps to improve the model performance. It is worth mentioning that

prediction power of Ridge Regression is better than the more advanced models, such as RFR and GBR.

B. Asymmetric Crude to Gas Price Transmission

According to the study by Borenstein et al in 1997, gasoline price was found to respond more quickly to increasing crude prices than decreasing crude prices. This asymmetric price adjustment is possibly caused by the adjustment lag in the gasoline distribution and production process as well as the market power of retail gasoline sellers. The following section of this study follows Borenstein et al. (1997)'s method in construction of a simple model to estimate the response of gasoline prices to oil price changes.

To begin with, we use R to represents the retail gasoline price/gallon and similarly C represents the crude oil price/gallon. Next, we denote the change in prices as $\Delta C_t = C_t - C_{t-1}$ and $\Delta R_t = R_t - R_{t-1}$. Since it takes multiple periods for gasoline to adjust to the crude prices, a simple model with asymmetric adjustment is constructed as follows.

Table V: Constructing Asymmetric Transmission

When $\Delta C_t > 0$,	When $\Delta C_t < 0$,
$\Delta R_t^t = \beta_0^+ \Delta C_t$	$\Delta R_t^t = \beta_0^- \Delta C_t$
$\Delta R_{t+1}^t = \beta_1^+ \Delta C_t$	$\Delta R_{t+1}^t = \beta_1^- \Delta C_t$
$\vdots$	$\vdots$
$\Delta R_{t+n}^t = \beta_n^+ \Delta C_t$	$\Delta R_{t+n}^t = \beta_n^- \Delta C_t$

where $\Delta C_t^+ = max\{\Delta C_t, 0\}$ $\Delta C_t^- = min\{\Delta C_t, 0\}$

The design of ΔC_t^+ and ΔC_t^- allows the model to show the directional difference in response rate to crude prices, combining this with the multi-period lag adjustment structure we have

$$\Delta R_t = \sum_{i=0}^{n} (\beta_i^+ \Delta C_{t-i}^+ + \beta_i^- \Delta C_{t-i}^-) + \varepsilon_t$$

According to the ordinary simple regression, the coefficients for crude price change are no longer significant after 6 periods. The cumulative response of gasoline price is calculated by using the coefficient of each period divided by the sum of all significant coefficients. Below is the graph that demonstrates the difference in response in gasoline prices.

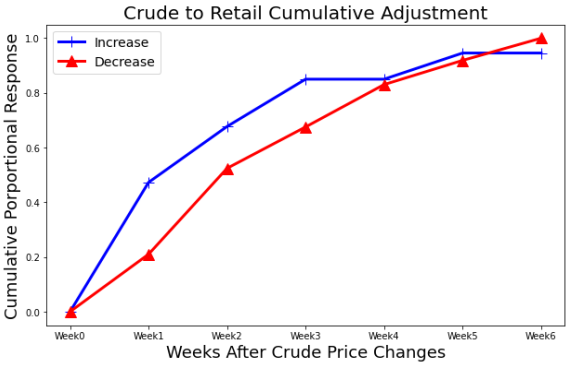

Fig. 7: Crude to Retail Cumulative Adjustment

The line with plus sign is the cumulative proportional adjustment to per unit increase in crude oil prices. Therefore, 55% of one unit increase in crude price was reflected in gasoline prices in the week after and another 15% of the adjustment happened in the next week, after 3 weeks more than 80% of the price increase in crude price has transferred on to the price of gasoline. The coefficients for positive crude price change in week 4 and 6 are insignificant and have been replaced by 0. From the graph above, we can see that the curve for crude oil price increase is higher than the curve for crude oil price decrease, showing the asymmetrical response to increase and decrease in crude prices.

ADDITIONAL FEATURES

According to the estimation of the Energy Information Administration, the major factor influencing gasoline prices is crude oil market's speculation. However, gasoline prices can fluctuate due to other reasons. For example, the consumer gasoline price is affected by refining cost, marketing and distribution, and taxes. On the other hand, the producer gasoline price can vary because of the supply and demand, crude oil price, and refining costs which are influenced by seasonal factors. In the following section, we examined the relationship between gasoline price and additional explanatory variables. We considered the additional variables from 4 aspects, seasonality, macroeconomic trend, related refined oil products, and the demand for refined products.

Seasonality: Season is likely to affect gasoline price, and it could be explained from two aspects. On one hand, road trips are more common in summer which lead to increased gasoline demand. To meet this high summer demand, refineries usually start preparing in the spring. On the other hand, refineries also make adjustments in their gas formulation in summer, to reduce global warming effects by replacing cheap yet easily evaporative elements with expensive and less evaporative ones, according to EIA. This becomes the second factor for the seasonality effect of gasoline price.

To see the seasonality effect clearly, we decomposed gasoline prices into trend, seasonal and residual factor.

$$y_{additive\ model} = trend + seasonality + residuals$$

$$y_{multiplicative\ model} = trend * seasonality * residu$$

We examined the gasoline prices using both the additive model and multiplicative model. The seasonality effect is quite evident in both models, as the seasonal chart shows periodical waves. However, the multiplicative model (right) shows quieter residuals, as opposed to the additive model (left).

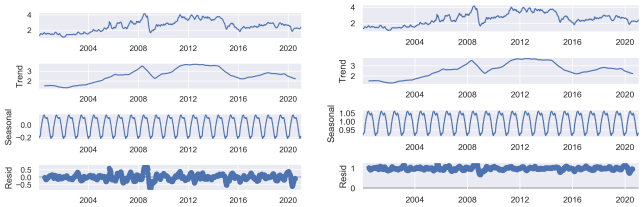

Fig. 8: Seasonality in Additive & Multiplicative Model

To investigate the effect more precisely, we used the multiplicative model, and took the first two years' data to zoom in. We observed that gasoline prices grow in spring, peak in summer and drop with fluctuation in fall and winter every year.

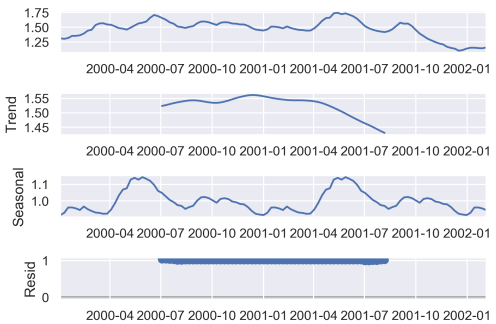

Fig. 9: A Closer Look at Seasonality in Gas Prices

Financial Market Uncertainty: The precious metal Gold, has long been regarded as a safe haven or hedge for the financial market uncertainty and inflation by many investors. Based on empirical analysis, crude oil, Gold, and DXY are connected and usually move in opposite directions. We further study whether this relationship exists between gasoline, Gold, and DXY. In addition, the global pandemic COVID-19 in early 2020 had great impact on oil futures, bringing oil futures to negative for the very first time; We picked U.S. unemployment insurance claims as a feature to see how the pandemic affected the retail gasoline prices.

Refined Oil Product: The production of gasoline is a joint process; the demand and price of other refined products is an important factor when we study the future behavior of gas prices. Heating oil, jet fuel, and retail diesel prices are selected as potential explanatory variables.

Demand for refined products: In search for data that captures the supply and demand, we found US refinery utilization capacity, gasoline inventory, and crude oil production to represent the market demand for gasoline.

The correlation matrix below granted us some insight on the relationship between gasoline and the three sections. On the left hand side of the matrix, the refined oil products are found strongly correlated. Just like crude prices are positively correlated with Gold price, retail gasoline price has similar positive correlation with Gold price as well. In the middle of the matrix, the features negatively correlated with gasoline price are unemployment insurance claims, DXY, and refinery operating capacity. When inflation rate rises the price of gasoline will follow, which justifies the negative correlation between DXY and gasoline. Lastly, the ending stock of gasoline and crude oil production have almost no correlation with gasoline price.

	Heating Oil Price	Diesel Retail Prices	Jet Fuel	Gold Price	Unemployment Insurance	USD_Index	Refinary Operation Capacity	Gasoline Inventory	Crude Oil Production	Gasoline
Heating Oil Price	1.00	0.97	0.99	0.64	-0.14	-0.77	-0.22	0.07	0.01	0.96
Diesel Retail Prices	0.97	1.00	0.96	0.74	-0.08	-0.72	-0.24	0.17	0.16	0.97
Jet Fuel	0.99	0.96	1.00	0.58	-0.17	-0.77	-0.23	-0.00	-0.07	0.96
Gold Price	0.64	0.74	0.58	1.00	0.13	-0.41	-0.32	0.58	0.56	0.69
Unemployment Insurance	-0.14	-0.08	-0.17	0.13	1.00	0.03	-0.44	0.19	0.09	-0.13
USD_Index	-0.77	-0.72	-0.77	-0.41	0.03	1.00	0.31	0.08	0.29	-0.75
Refinary Operation Capacity	-0.22	-0.24	-0.23	-0.32	-0.44	0.31	1.00	-0.14	0.11	-0.21
Gasoline Inventory	0.07	0.17	-0.00	0.58	0.19	0.08	-0.14	1.00	0.71	0.09
Crude Oil Production	0.01	0.16	-0.07	0.56	0.09	0.29	0.11	0.71	1.00	0.08
Gasoline	0.96	0.97	0.96	0.69	-0.13	-0.75	-0.21	0.09	0.08	1.00

Table VI: Additional Features Correlation Matrix

To fully understand whether the past data from additional time series cause movements in gasoline prices, Granger Causality test is employed. In order to proceed, these time series need to be transformed to stationary time series, a differencing function is applied on the training set. Following the ADF test confirming the stationarity after first difference transformation, we tested the causation using Granger's Causality Test. The following table is the result of the granger causality test based on the row and column pair relationship.

Table VII: Features Granger Causality Matrix

	Heating Oil Price_x	Diesel Retail Prices_x	Jet Fuel_x	Gold Price_x	Unemployment Insurance_x	USD_Index_x	Refinary Operation Capacity_x	Gasoline Inventory_x	Crude Oil Production_x	Gasoline_x
Heating Oil Price_y	1.0000	0.2025	0.0586	0.1601	0.0204	0.0013	0.8003	0.1092	0.1419	0.0555
Diesel Retail Prices_y	0.0000	1.0000	0.0000	0.0000	0.0113	0.0000	0.1630	0.0070	0.0253	0.0000
Jet Fuel_y	0.0000	0.0255	1.0000	0.0000	0.1345	0.0000	0.0002	0.0899	0.0223	0.0115
Gold Price_y	0.1436	0.2047	0.1503	1.0000	0.0100	0.6047	0.4015	0.0609	0.3768	0.1340
Unemployment Insurance_y	0.0019	0.0756	0.0000	0.0000	1.0000	0.0001	0.0006	0.0012	0.1805	0.0119
USD_Index_y	0.0271	0.0541	0.0456	0.2106	0.5874	1.0000	0.3008	0.8427	0.0665	0.1074
Refinary Operation Capacity_y	0.0898	0.0473	0.0000	0.0006	0.0000	0.0453	1.0000	0.0000	0.0117	0.0018
Gasoline Inventory_y	0.0369	0.3577	0.2711	0.0003	0.0000	0.4740	0.0000	1.0000	0.0000	0.2319
Crude Oil Production_y	0.0023	0.0201	0.0008	0.0149	0.0000	0.0373	0.0336	0.1673	1.0000	0.0873
Gasoline_y	0.0000	0.0000	0.0000	0.0001	0.1529	0.0000	0.0152	0.0000	0.0445	1.0000

According to the granger causality matrix, the prices of selected refined oil products have significant influence on gasoline price movements. Gasoline, along with diesel fuel, heating oil and jet fuel, are all products from the crude oil refining process. Because of the joint production, there are co-movements in refined oil products.

Moving on to the financial market uncertainty section, both the gold price and USD index cause price movements in gasoline, while gasoline price doesn't cause movement in neither financial market uncertainty nor inflation. Originally, the factor of unemployment insurance claim was designed to reflect the impact of economic recession and pandemic on gasoline prices. However, by test result gasoline price has statistical significant influence on unemployment insurance claims but not the other way around. The granger causality test is not significant, we failed to reject the null hypothesis where employment is affecting gas price. However, the gasoline price might be able to predict the future unemployment insurance claims.

Finally, all variables in the supply and demand section are significant. Meaning that refinery operating capacity, gasoline inventory, and crude oil production can be used in predicting the future trend of gasoline prices.

IV. Do crude prices contain predictive power for the concentration of carbon dioxide in the atmosphere?

STATISTICAL ANALYSIS

A. Granger Causality

To find the relationship between oil futures and CO2 concentration, we repeat the same steps. We first used Granger Causality Tests with maxlag=15 and found that p-values for all lags are greater than 0.05, which may imply that there is no correlation between crude futures and CO2 concentration. To confirm that, we presented Granger's Causality Matrix.

Table VIII: Crude Future - CO2 Granger Causality

	CO2_x	Crude Oil_x
CO2_y	1.0	0.0592
Crude Oil_y	0.4053	1.0

On the off-diagonal entries, because $0.0592 > 0.05$, we failed to reject the null hypothesis that past oil futures prices doesn't cause future movements in CO2 concentration at 5% significance level; Since $0.4053 > 0.05$, CO2 concentration does not cause movements in crude futures. At this point, our first impression is that oil futures might not be a good predictor for future atmospheric carbon dioxide concentration.

Table IX: Crude-CO2 Lags

Granger Causality	F-statistic	P value
lag 1	1.2081	0.2720
lag 2	0.4225	0.6555
lag 3	3.5513	0.0141
lag 4	2.1332	0.0747
lag 5	1.9892	0.0778

B. Spearman Correlation and Cointegration Test

Correlation: The result of KS test indicates both oil futures price and CO2 concentration are not following a normal distribution. Therefore, we used Spearman correlation that showed coefficient between the two is 0.2785 with p-value $2.092e^{-20}$, which proves that they have very low correlation.

Cointegration test: The p-value provided by Cointegration test is 0.2183 which is greater than 0.05, so we failed to reject the null hypothesis of no cointegration and concluded that these two time series do not cause each other.

C. OLS Results

TABLE X: Crude - CO2 Lags

	coeff	Std err	P > | t |
Intercept	0.1103	0.050	0.027
lag(close5) 0	-0.0116	0.006	0.055
lag(close5) 1	0.0103	0.008	0.225
lag(close5) 2	0.0163	0.008	0.054
lag(close5) 3	-0.0233	0.008	0.006
lag(close5) 4	0.0073	0.006	0.225

Since the Granger Causality test produced insignificant results, we used Ordinary Least Squares Regression to find the optimal lag. Based on the OLS results, only when lag=3, the p-value is 0.006 which is significant when the threshold=0.05. Hence, we add 3 week lags of crude oil futures in our model.

D. AutoRegression

Using the same approach as earlier, we plot PACF and ACF with maxlag=10 to determine the lag order of CO2 concentration itself and find it is AR(1). We will include 1 lag of CO2 concentration in our model.

PREDICTION MODELS

In this section, we experiment with different machine learning models to examine if crude futures contain predictive power for CO2 concentration and use RMSE as a performance measure.

A. Model Section

(1) Random Forest Regression (RFR): After the ADFuller test, we find that ppm of CO2 is not stationary, so we use the first time difference as a dependent variable. We introduced the grid search method to help us look for the best pair of parameters. Because of the poor performance of oil futures and the clear seasonal pattern of CO2 concentration, we included additional variables in our model. The seasonal variable showed quite a lot of improvement by letting the model follow the periodic principles. Natural gas futures, however, is not as good a predictor since it significantly dragged down the line after the first year. As for improvements, we could take a look into Gradient Boosting Regression.

(2) Gradient Boosting Regression (GBR): GBR provided a more accurate and smoother prediction. After each period, it can easily estimate and catch up the trend. More importantly, the prediction line generated from the holdout data by GBR has no gap with the CO2 concentration. Similarly, we saw crude oil futures and natural gas futures performed poorly. The trend became much clearer once we added the seasonal variables.

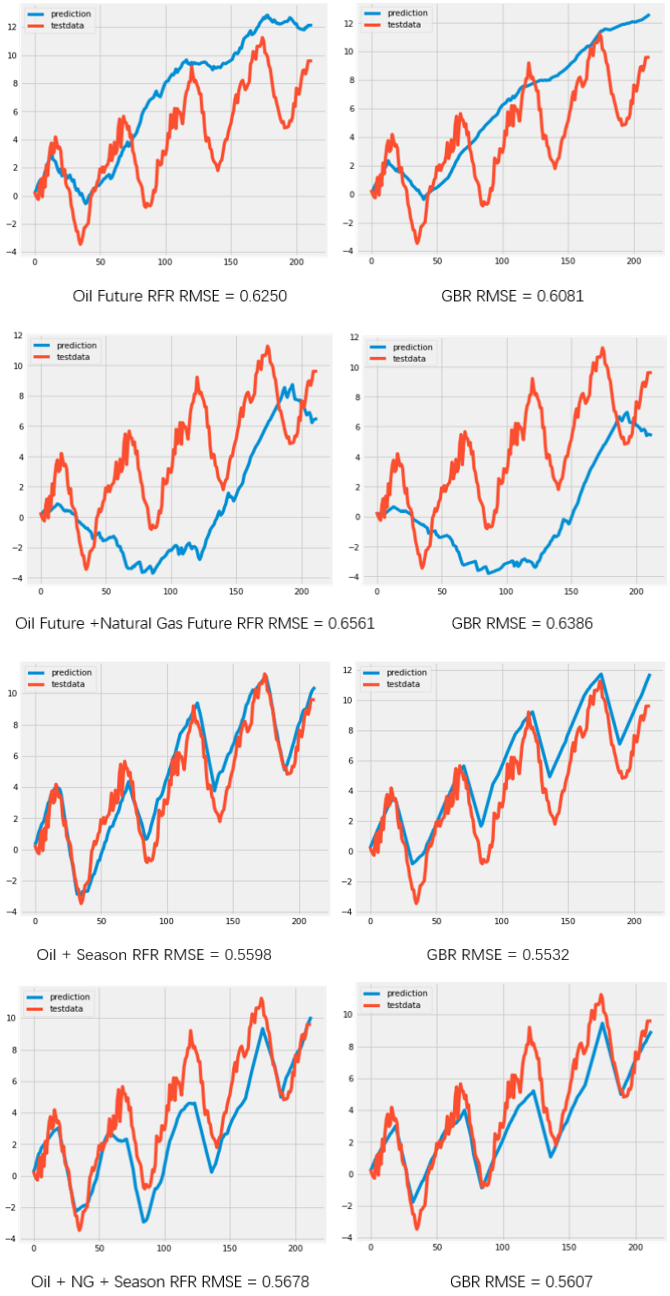

Fig. 10: Seasonality Improves Prediction Performance

(3) Long short-term memory (LSTM): We also tried a time series model to predict future CO2 concentration due to the existence of autocorrelation. With the first 80% samples as train set, we initially use oil futures for the past three weeks as our independent variables due to the three week lag effect. We then rescaled the data by minmaxscaler to fit them into the three dimensional sequential model. The model is a combination of one convolutional layer, two LSTM layers and a dense

layer. RMSE after 15-episode training is 0.5604, which shows poor performance.

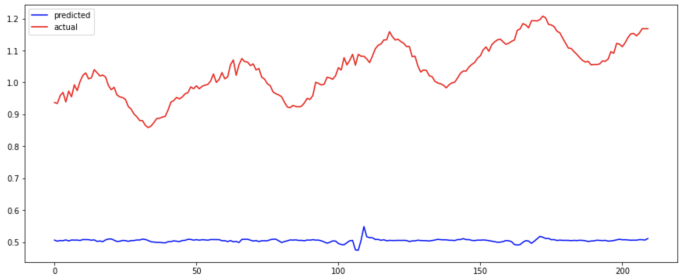

Fig. 11: LSTM Model Performance

Through the graph above, we noticed the regular fluctuation and made a conjecture about the effect of seasonality. After adding seasonal dummies, the RMSE is still 0.5738. But the model better captures the fluctuation of the CO2 concentration.

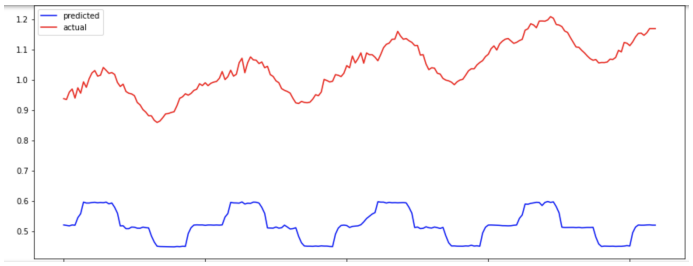

Fig. 12: LSTM with Oil futures+Seasonal dummies

We then added natural gas futures data and got a much better prediction with RMSE 0.3591.

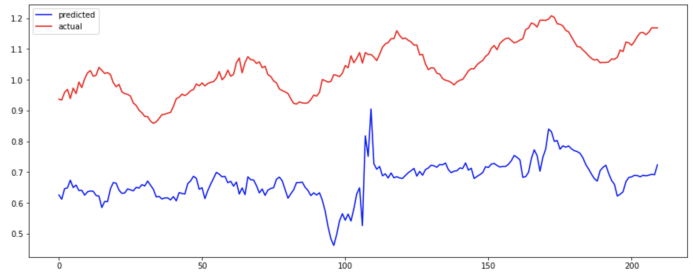

Fig. 13: LSTM with Oil futures+Natural gas futures

B. Implications

Table XI: Model Performance (RMSE)

Model	RFR	GBR	LSTM
Oil Futures	0.6250	0.6081	0.5604
Oil NG	0.6561	0.6386	0.3591
Oil Season	0.5598	0.5532	0.5738
Oil NG Season	0.5678	0.5607	0.3737

Based on the result, LSTM is the best model. Because LSTM is insensitive to seasonal dummies, including season made the model even worse. In contrast, natural gas futures itself significantly improved the model and gave a smaller score than adding both features. Compared to LSTM, RFR and GBR showed a different result: seasonality alone improved the model, but natural gas futures didn't. Although RMSE suggests that seasonality does not improve the LSTM, based on Fig. 12, adding seasonal dummies helps to capture the general trend of the actual CO2 concentration. In LSTM, we found that oil futures along with natural gas futures generate the smallest RMSE, which implies that natural gas futures is the additional feature that has strong predictive power of CO2 concentration.

ADDITIONAL FEATURES

A. Seasonality:

In 1958, climate scientist Charles David Keelings discovered that the concentration of CO2 in the atmosphere decreases in spring and summer due to photosynthesis, and increases during fall and winter. This well known cyclic pattern is the famous Keeling Curve. In Yuan et al (2017), the study shows that after 2000, the amplitude of seasonality effect increased greatly from global vegetation growth. However, according to the research, this seasonal variation is largely driven by plant respiration instead of carbon emission due to human activities, which explains why oil futures is a poor predictor.

To see the seasonality effect more clearly, we decompose CO2 concentration into trend, seasonal and residual factor, and interpolate by year.

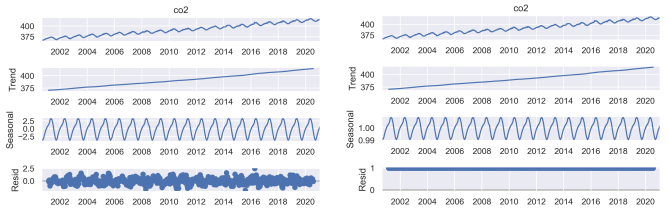

Fig. 14: Seasonality in Additive & Multiplicative Model

Multiplicative model (right) showed quieter residual, compared to the additive model (left). Both extract seasonality successfully, as the seasonal charts show periodical waves. Observe that CO2 concentration grows in mid-fall, stably climbs and peaks in spring, declines in summer and hits bottom in fall every year.

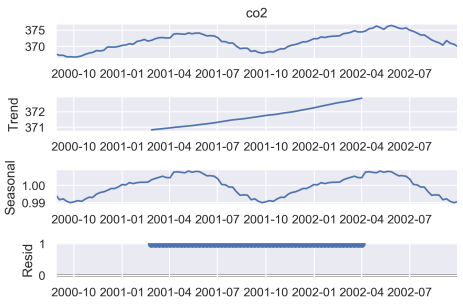

Fig. 15: A Closer Look at Seasonality in CO2

B. Natural Gas

Natural Gas, oil and coal are the top three energy sources that contribute to CO2 emissions, according to the International Energy Agency (IEA). Although less essential than the role of coal and oil, natural gas still consists of a considerable part in CO2 emission.

As an indicator for natural gas, we considered natural gas futures price and spot price. Our choice came down to natural oil futures, since futures prices reflect people's expectations, and would be more consistent with the adoption of crude oil futures prices as an explanatory variable.

The relationship between natural gas futures prices and CO2 concentration is tested using the Granger Causality Tests.

TABLE XII: CO2-Natural Gas Granger Causality

	CO2_x	Natural Gas Futures_x
CO2_y	1.0	0.0
Natural Gas Futures_y	0.0	1.0

The Granger Causality Test is carried out with a max lag of 15. As the matrix shows, p-value on the off-diagonal positions are 0, which implies that natural gas futures prices and CO2 concentration have bilateral causal relationship. This result is meaningful and can be used to refine our prediction model.

V. IMPROVEMENTS

A. Limitation on feature selection

Initially, we considered several factors that could be good explanatory variables. However, in practice the data frequency or time horizons are not ideal. We believe a better explanation could be generated if more data on Coal CO2 emission or price of clean energy are available.

B. Forecast Time Horizons: We acknowledge that certain models might have better performance in short term but underperform in longer term predictions. Without comparing the models in different time horizons it's unclear whether the best performing model selected in this paper is the best in all time horizons.

D. Model Selection: it is worth saying that extrapolating background noise from the data can be an improvement. Since the highly volatile market makes a deep impact on either the futures or the gasoline price. A decomposition process, to some extent, may generate significant results. It not only makes the prediction smoother, but also tells us more about the relationship between dependent and independent variables. We would like to continue digging deeper into this field using Empirical Mode Decomposition as well as Butterworth Filters.

VI. CONCLUSION

Making predictions with crude oil futures is a challenging task due to its volatile nature. In our research, we found that oil futures with one week lag is a reliable predictor for gasoline prices. Gasoline prices respond more quickly to rising oil futures prices. The forecast performance significantly improved with additional variables representing financial market uncertainty, seasonality, refinery operation capacity, etc. In contrast, oil futures alone is not a good predictor for CO2 concentration. By comparison, natural gas futures prices have strong predictive power for CO2 concentration. It is safe to conclude that oil futures isn't the driving factor in CO2 concentration. In both problems, the LSTM model is selected because it gives the most reliable prediction with the lowest root mean squared error among our models, which proves that it is an effective model for time series prediction problems.

REFERENCES

[1] Borenstein, S., Cameron, A. C., & Gilbert R. (1997). Do Gasoline Prices Respond Asymmetrically to Crude Oil Price Changes? *The Quarterly Journal of Economics, 112*(1), 305-339.

[2] Emmons, W. R., & Yeager, T. J. (2002). *The Futures Market as Forecasting Tool: An Imperfect Crystal Ball*. STL FED. https://www.stlouisfed.org/publications/regional-economist/january-2002/the-futures-market-as-forecasting-tool-an-imperfect-crystal-ball

[3] Fuller, W. A. (1976). Introduction to Statistical Time Series. New York: John Wiley and Sons. ISBN 0-471-28715-6.

[4] Gasoline price fluctuations - U.S. Energy Information Administration (EIA), last accessed Feb.10 2021. https://www.eia.gov/energyexplained/gasoline/price-fluctuations.php

[5] Granger, C. W. J. (1980). Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424-438.

[6] Yuan, W., et al. (2017) Influence of Vegetation Growth on the Enhanced Seasonality of Atmospheric CO2. *Global Biogeochemical Cycles, 32*(1), 32-41