Example_Project_A

March 1, 2022

CMSE 201 Final Semester Project - Honors

0.0.1 Connor Trask

Date: 12/3/20

1 Future of US Energy Production: 30-Year Output Forecast

1.1 Background and Motivation

As man-made climate change continues to intensify due to the burning of fossil fuels, attention is increasingly focused on the world's "energy mix", the composition of primary sources that fuel our society. As the second largest producer of energy in the world[1] and the country with the most robust data on energy production, the United States is a natural selection for detailed analysis. Due to technological developments and pressure from climate change, renewable energy is beginning to be implemented widely into existing power grids. Despite this progress, however, the United States remains incredibly dependent on fossil fuels[2], especially natural gas, the production of which has fallen dramatically due to advancements in extraction technology (fracking).

As the next 30 years are incredibly important to control climate change, it is equally important to predict how US energy production will develop over this timespan. Statistical analysis of trends in energy production levels will be conducted in order to model historical data and forecast forward to 2020 - 2050. **What will the US energy mix be in 2050, and how will it evolve to reach that point?**

1.2 Methodology

The primary source for US energy data is the US Energy Information Administration (EIA), and hence the source for the historical data surrounding US energy production[3]. The EIA's dataset contains production levels of coal, natural gas, crude oil, natural gas plant liquids (NGPL), nuclear, and renewable energy measured in quadrillions of BTU from 1950 - 2019.

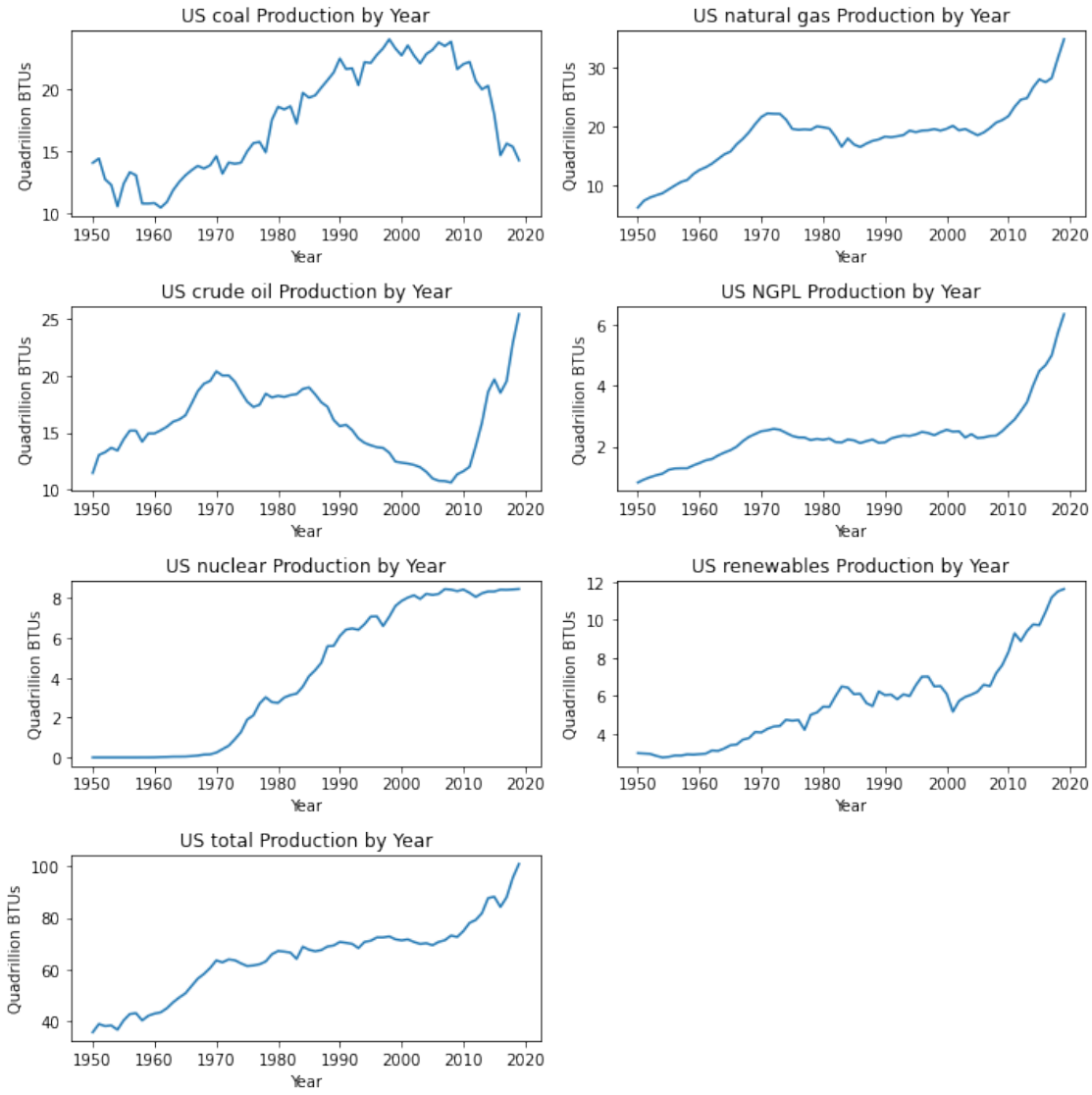
[1]:
```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
from scipy.optimize import curve_fit
import seaborn as sns
%matplotlib inline
```

```
#Import historical US energy production data
energy_history = pd.read_csv("US_Energy_Production_History.
 ↪csv",header=5,index_col=0)
energy_history["total"] = energy_history.sum(axis=1) #Add total column
energy_history.tail() #Check csv loaded properly
```

[1]:

	coal	natural gas	crude oil	NGPL	nuclear	renewables	total
2015	17.9461	28.0669	19.6959	4.4760	8.3369	9.7288	88.2506
2016	14.6671	27.5760	18.5117	4.6648	8.4268	10.4229	84.2693
2017	15.6254	28.2893	19.5350	4.9871	8.4190	11.1959	88.0517
2018	15.3634	31.6899	22.8897	5.7270	8.4381	11.5084	95.6165
2019	14.2681	34.8946	25.4389	6.3366	8.4624	11.6370	101.0376

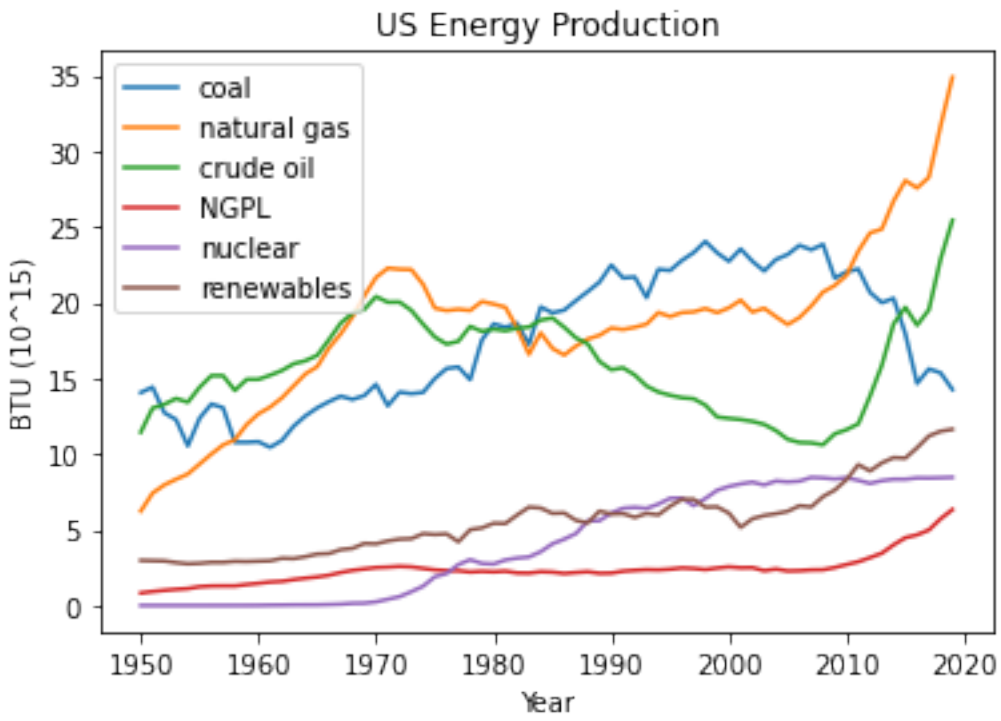
[2]:
```
i = 0
fig = plt.figure(figsize=(10,10)) #Set figure size (from class assignments)
for source in energy_history.columns:  #Plot each energy source individually
    i += 1
    plt.subplot(4,2,i)
    plt.plot(energy_history.index,energy_history[source])
    plt.xlabel("Year")
    plt.ylabel("Quadrillion BTUs")
    plt.title("US " + source + " Production by Year")
plt.tight_layout()
```



By viewing each primary source in its own subplot, their shape can be properly discerned, which is important in determining how best to model the data. Each source's shape will be discussed in further detail later. Before moving on to the individual sources, however, it is beneficial to appreciate their relationships in context to the US energy mix and to each other.

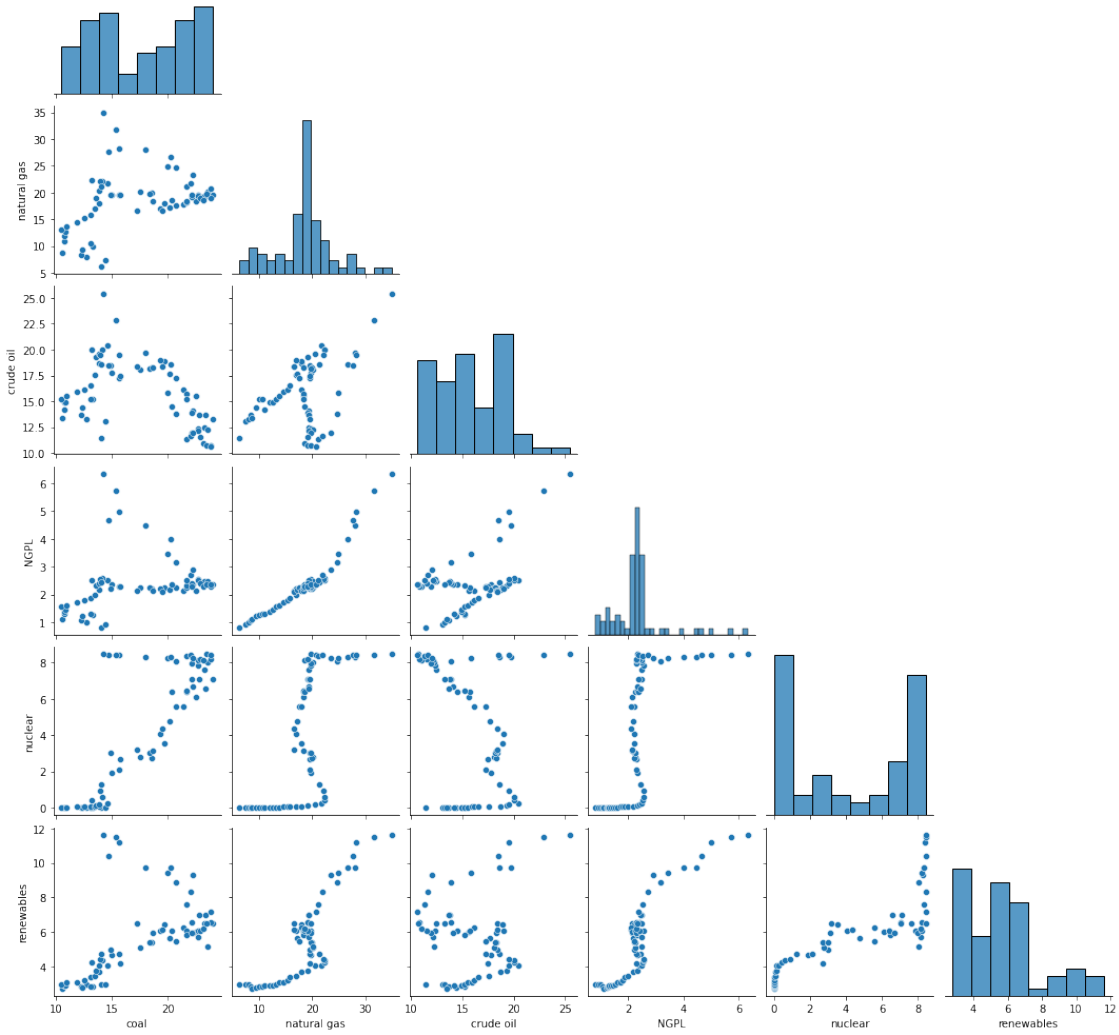
[3]:
```
energy_history.drop(columns="total").plot() #Plot all sources together, except␣
↪for the total
#Since inplace=False, this doesn't remove the total column from the dataframe
plt.title("US Energy Production")
plt.xlabel("Year")
plt.ylabel("BTU (10^15)")
```

[3]:
```
Text(0, 0.5, 'BTU (10^15)')
```



[4]:
```
#Use pairplot to scan for correlations between sources
sns.pairplot(energy_history.drop(columns="total"),corner=True)
```

[4]: <seaborn.axisgrid.PairGrid at 0x7facd05ddb20>

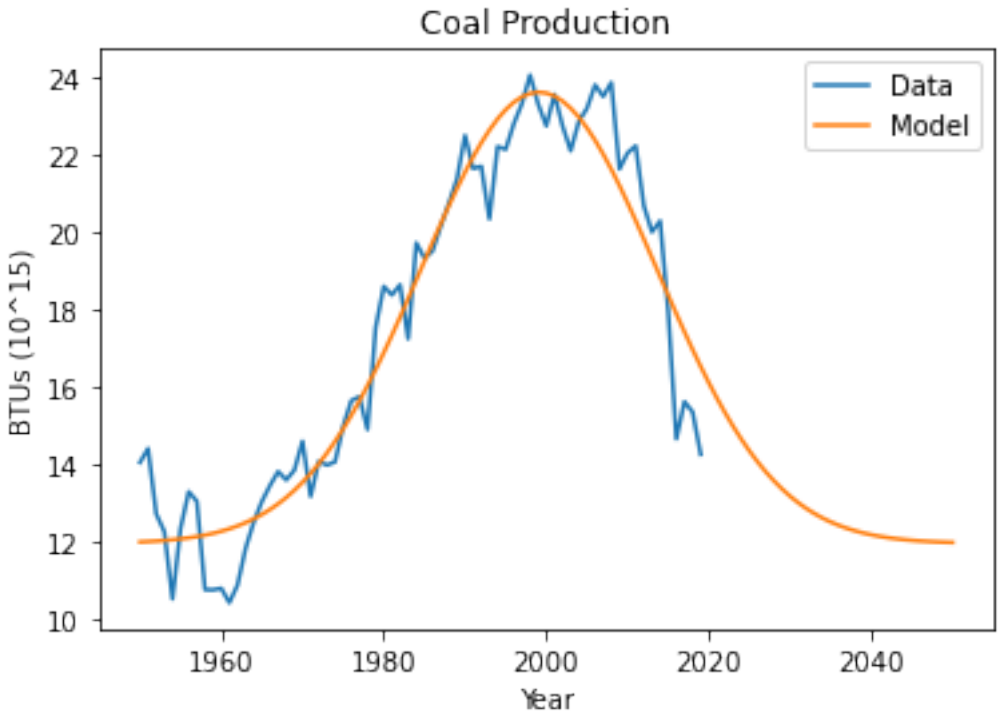


While the seaborn pairplot reveals that there is no strong correlation between the majority of the energy sources, it does identify a few interesting connections. Perhaps unsurprisingly there is a weak positive correlation between natural gas & crude oil, as well as crude oil & NGPL. Most importantly, however, is the incredibly strong correlation between natural gas and NGPL, although this is again expected due to the nature of these two fuels.

[5]:
```
def coal_production(year,a,b,c,t): #Utilize a normal distribution function for␣
↪coal production
    co_t = a*np.exp(-((year-t)**2)/b) + c #Rationale and parameters explained␣
↪below
    return co_t
#Convert pandas series to arrays
```

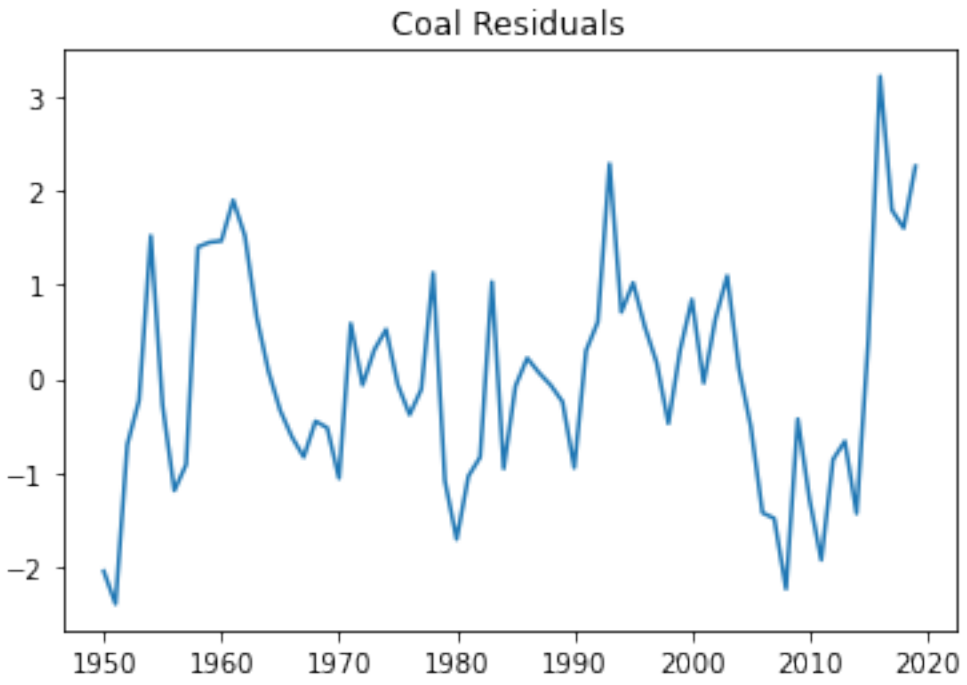
```
years = np.asarray(energy_history.index) #Years represents the range covered by␣
 ↪the data (1950 - 2019)
coal = np.asarray(energy_history["coal"])
#Calculate coal parameters, p0 found through experimentation
params_co,pcov = curve_fit(coal_production,years,coal,p0=[16,1000,10,2000])

future = np.arange(1950,2051,1) #Create future array, representing prediction␣
 ↪model's range (1950 - 2050)
forecast_co = coal_production(future,*params_co) #Forecast model for prediction
fit_co = coal_production(years,*params_co)          #Fit model for calculating␣
 ↪R-squared value (goodness of fit)
#Plot the data, model
plt.plot(years,coal,label="Data")
plt.plot(future,forecast_co,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Coal Production")
print("A:{:.4f}, B:{:.4f}, C:{:.4f}, T:{:.4f}".format(*params_co)) #Print␣
 ↪parameters
```

```
A:11.6313, B:423.3207, C:11.9728, T:1999.0914
```



[6]:
```
def r_squared(actual,model): #Define function to calculate residuals, R-squared value (formula from here: [4])
    residuals = model - actual #Calculate residuals
    RSS = np.sum(residuals**2) #Sum the square of all residuals (unexplained deviation)
    TSS = np.sum((actual - actual.mean())**2) #Sum the square of total variance
    R_sqr = 1 - RSS/TSS #Calculate R-squared value
    return residuals, R_sqr #Return residual array, R-squared value

#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_co, R_co = r_squared(coal,fit_co)
plt.plot(years,residuals_co)
plt.title("Coal Residuals")
print("R-Squared value: {:.4f}".format(R_co))
```

```
R-Squared value: 0.9308
```

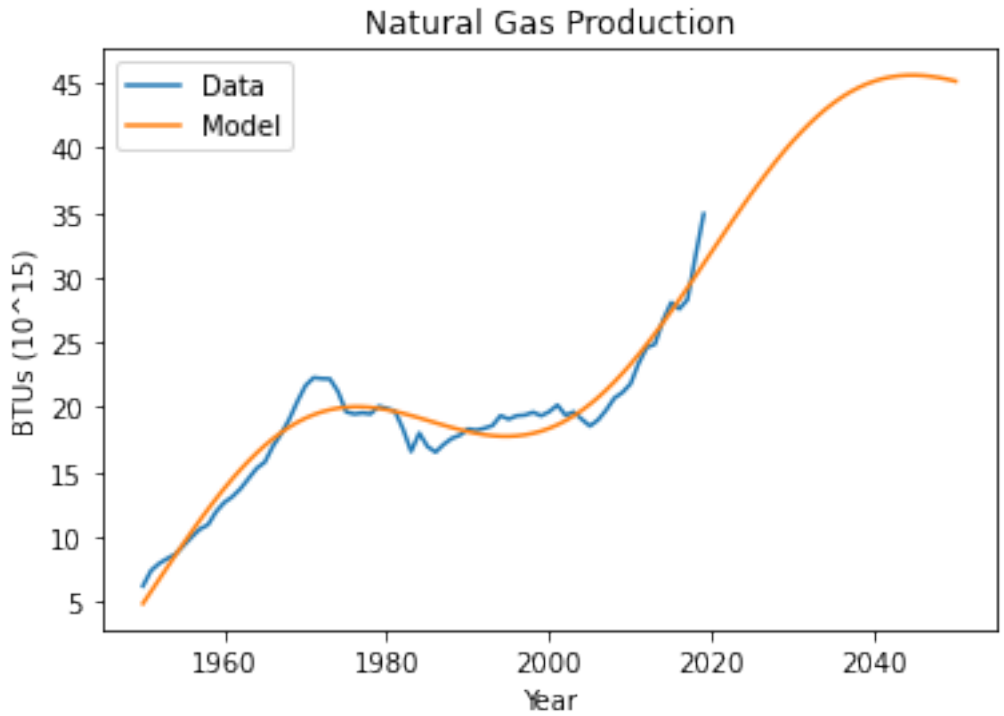


Coal Methodology US Coal production experienced a steady increase for the second half of the 20th century, eventually reaching its peak in the early 2000s. While coal had been preferred for power generation due to its cheap cost, the falling price of natural gas and renewables[5] combined with the decreasing energy value of coal[6] have caused production to plummet in recent years. As a result of this steep curve with a lower limit shape, a bell curve function was chosen to model the data. As the residuals plot and R2 value shows, this model provides a satisfactory fit to the data.

Parameters: * A: Controls the steepness of the curve * B: Controls the width of the curve * C: Controls the lower limit of the curve * t: Defines the year in which the curve peaks

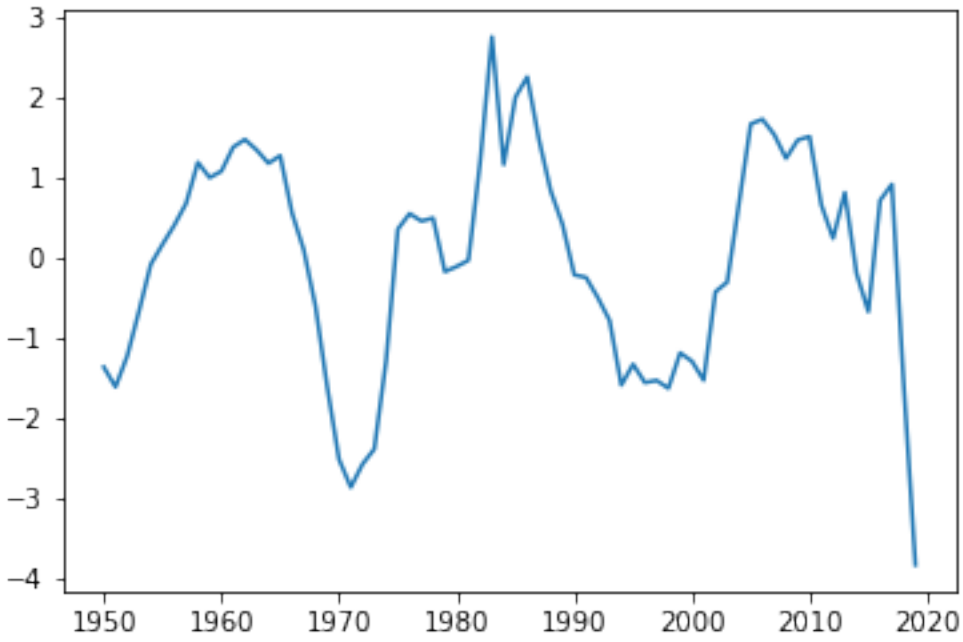
[7]:
```
#Identical format to coal production
def nat_gas_production(year,a,b,c,d,t): #Model for natural gas production
    ng_t = a*np.sin(b*(year-t)) + c*(year-t) + d #Rationale and parameters␣
 ↪explained below
    return ng_t

nat_gas = np.asarray(energy_history["natural gas"])

params_ng,pcov = curve_fit(nat_gas_production,years,nat_gas,p0=[6,0.07,0.
 ↪4,2,1945])

forecast_ng = nat_gas_production(future,*params_ng) #For forecasting future␣
 ↪values
fit_ng = nat_gas_production(years,*params_ng)    #For confirming goodness of fit

plt.plot(years,nat_gas,label="Data")
plt.plot(future,forecast_ng,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Natural Gas Production")
print("A:{:.4f}, B:{:.4f}, C:{:.4f}, D:{:.4f}, t:{:.4f}".format(*params_ng))
```

```
A:6.1187, B:0.0920, C:0.3741, D:6.1222, t:1951.3311
```



[8]:
```
#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_ng, R_ng = r_squared(nat_gas,fit_ng)
plt.plot(years,residuals_ng)
print("R-Squared value: {:.4f}".format(R_ng))
```

```
R-Squared value: 0.9371
```



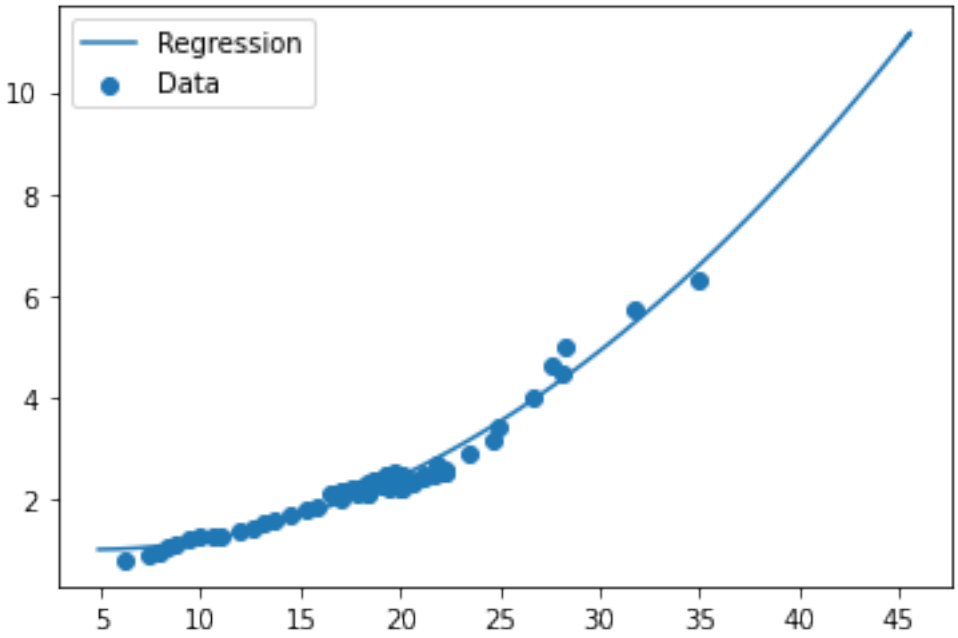
Natural Gas Methodology The natural gas curve is composed of two periods of rapid growth connected by a flat plateau where production is relatively constant. This period of stagnation can likely be attributed to the energy crisis of the 1970s, when petroleum and natural gas supply and production began to slow down[7]. The fresh growth in the 21st century can be attributed to developments in fracking, a method of extraction that can tap into the US's vast resovoirs of shale/tight natural gas[8]. While production is rising rapidly now, it can be expected that it will plateau in the future, either as a result of environmental regulations or due to decreasing supply. As a result, the model incorporates both a sinusoid function and a linear function, modeling the source's cyclic nature and the overall growth of production. As seen above, this model provides a satisfactory R2 value, with the residuals plot revealing much of the deviation is due to spikes in the early 70s and late 2010s.

Parameters: * a: Amplitude of the sine function * b: Frequency of the cycles * c: Slope of the linear growth * d: Baseline production level * t: x-adjustment to account for the domain beginning at 1950, not 0

[9]:
```
ngpl = np.asarray(energy_history["NGPL"])

params_ngpl = np.polyfit(nat_gas,ngpl,2) #Quadratic regression between natural␣
 ↪gas and NGPL
```

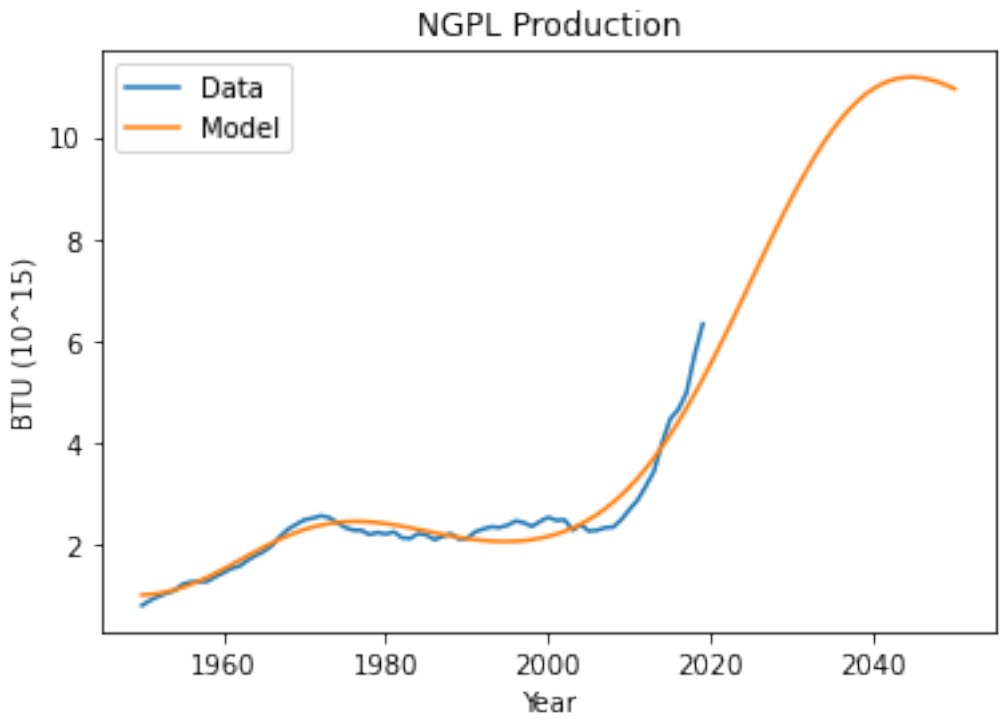
```
model_ngpl = np.poly1d(params_ngpl)
forecast_ngpl = model_ngpl(forecast_ng) #For future values
fit_ngpl = model_ngpl(nat_gas)      #For goodness of fit
#Plot regression
plt.scatter(nat_gas,ngpl,label="Data")
plt.plot(forecast_ng,forecast_ngpl,label="Regression")
plt.legend()
print(params_ngpl)
```

```
[ 0.00603156 -0.0554422   1.15938977]
```



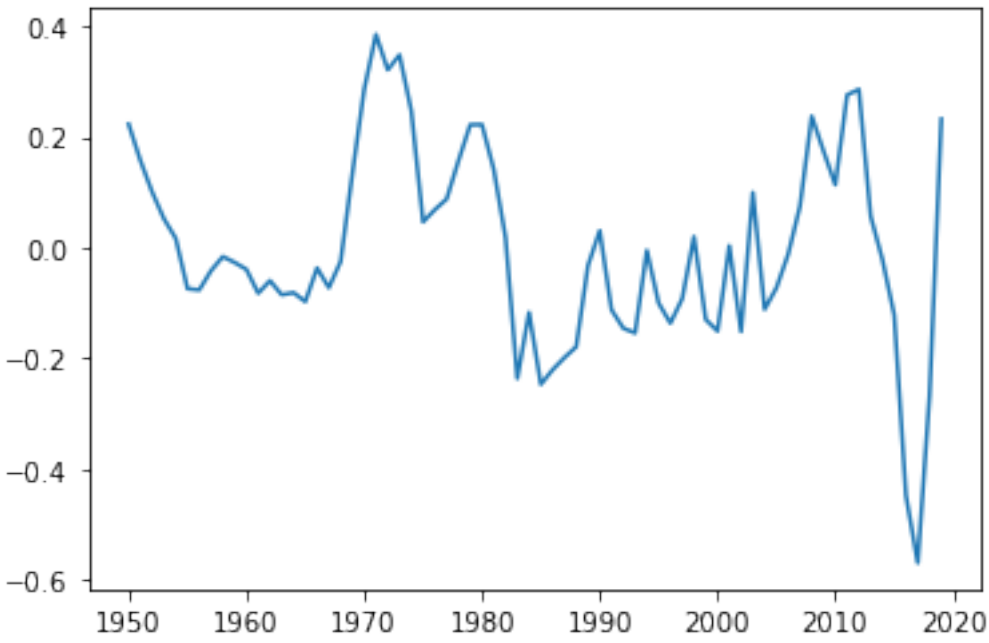
[10]:
```
#Plot the NGPL model vs data
plt.plot(years,ngpl,label="Data")
plt.plot(future,forecast_ngpl,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTU (10^15)")
plt.title("NGPL Production")
```

[10]: Text(0.5, 1.0, 'NGPL Production')



[11]:
```
#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_ngpl, R_ngpl = r_squared(ngpl,fit_ngpl)
plt.plot(years,residuals_ngpl)
print("R-Squared value: {:.4f}".format(R_ngpl))
```

```
R-Squared value: 0.9675
```

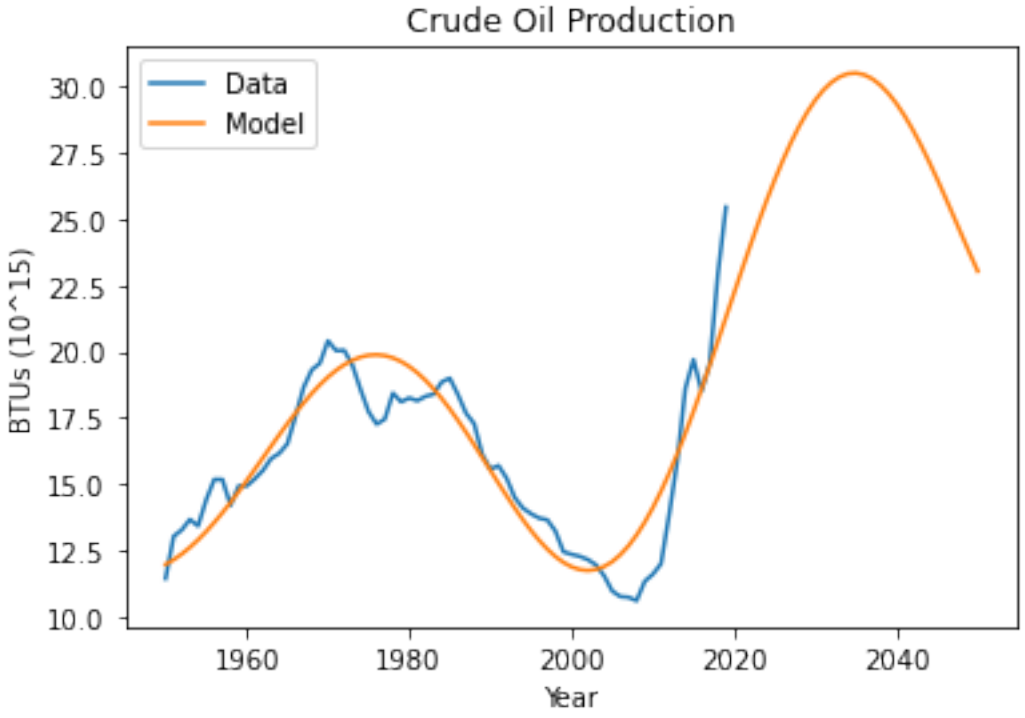


NGPL Methodology While the shape of NGPL and natural gas production is visually very similar, the difference in scale between the two makes it impossible for curve_fit to find an optimized solution with the natural_gas_production equation. However, there is a very strong correlation between the two curves, fitting a linear regression well and a quadratic regression nearly perfectly. As a result, this quadratic regression can be used to generate a model for NGPL production based on natural gas production instead of year, leading to the curve seen above. As natural gas plant liquids are byproducts of natural gas production, this methodology is very sound for predicting future NGPL production. Furthermore, the residuals plot confirms that this is a very accurate fit, with an R2 value higher than the natural gas curve and the majority of deviations coming from the current spike.

[12]:
```
def crude_production(year,a,b,c,d,t): #Define crude oil production formula
    cr_t = a*(year-t)*np.sin(b*(year - t)) + (c*(year-t))**3 + d #Rationale and␣
 ↪parameters below
    return cr_t

#Years already converted from previous cell
crude = np.asarray(energy_history["crude oil"])

params_cr,pcov = curve_fit(crude_production,years,crude,p0=[0.075,0.11,0.
 ↪02,11,1905])

#Future already created in previous cell
forecast_cr = crude_production(future,*params_cr)
fit_cr = crude_production(years,*params_cr)
```

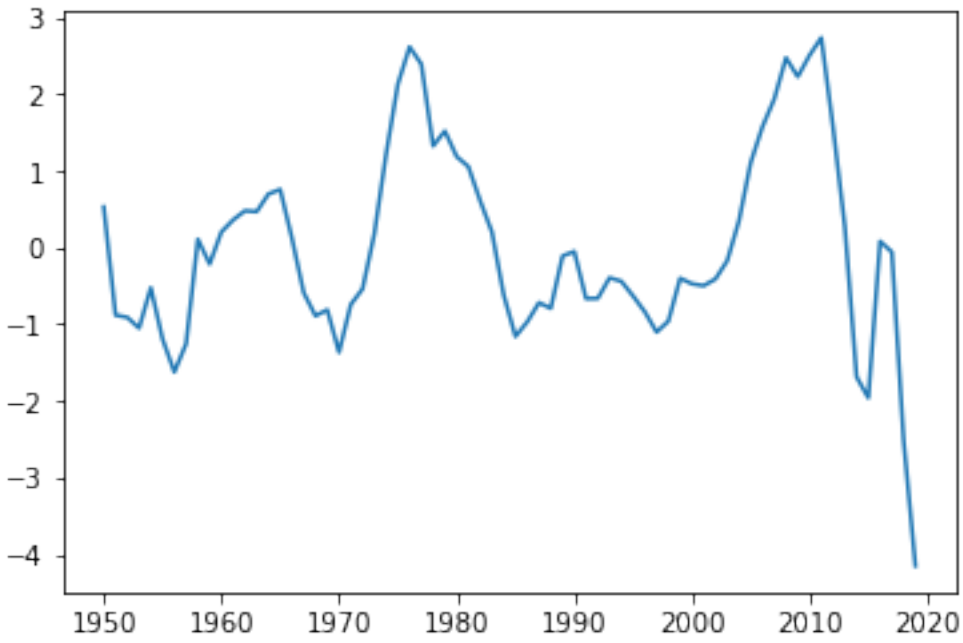
```
plt.plot(years,crude,label="Data")
plt.plot(future,forecast_cr,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Crude Oil Production")
print("A:{:.4f}, B:{:.4f}, C:{:.4f}, D:{:.4f}, t:{:.4f}".format(*params_cr))
```

```
A:0.0596, B:0.1075, C:0.0154, D:13.9482, t:1900.4441
```



[13]:
```
#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_cr, R_cr = r_squared(crude,fit_cr)
plt.plot(years,residuals_cr)
print("R-Squared value: {:.4f}".format(R_cr))
```

```
R-Squared value: 0.8342
```



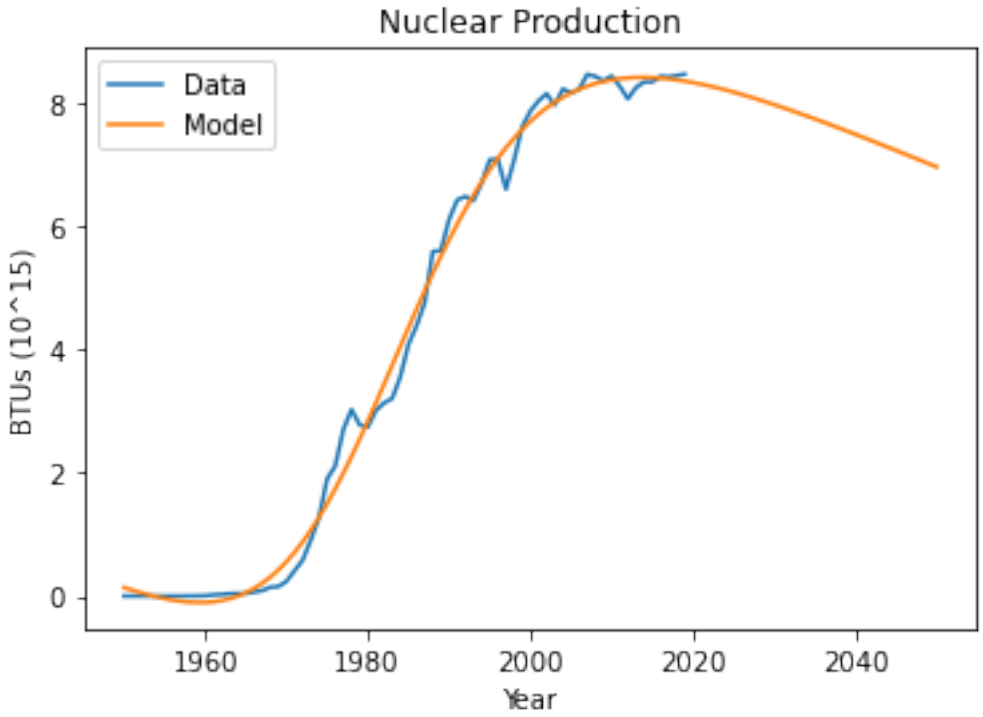
Crude Oil Methodology The crude oil production curve was the most difficult to fit out of all of the primary sources investigated in this analysis. The data has an undeniably sinusoidal shape, however, a simple sine function is unable to capture the spike and overall elevated level of crude production brought on by fracking[9]. Two additions were made to account for this development: the amplitude of the function was made dependent on the year, and a cubic function was added. By making the amplitude time-dependent, it would increase over time, capturing the recent spike. However, this also resulted in an unreasonably low value by 2050, which necessitated the addition of the cubic function to prop up the low end of the sinusoid. The messiness inherient in this process can be clearly seen in the residual plot and R2 value, the lowest of all primary sources. Despite this, however, the model clearly captures the trend of the data and produces a reasonable forecast.

Parameters: * A: Controls the amplitude of the sine function (time-dependent) * B: Controls the frequency of the sine function * C: Controls the growth rate of the cubic function * D: Controls the y-position of the curve * t: x-adjustment to account for the domain beginning at 1950, not 0

[14]:
```
def nuclear_production(year,A,K,B,v,Q,d,t): #Define generalised logistic␣
↪function with parameters [10]
    n_t = A + (K-A) / ((1 + Q*np.exp(-B*(year-t)))**(1/v)) - d*(year-t)
    return n_t

#Years already converted from previous cell
nuclear = np.asarray(energy_history["nuclear"])

params_nuc,pcov = curve_fit(nuclear_production,years,nuclear,p0=[0,9,0.1,1,1,0.
↪02,1985])
```

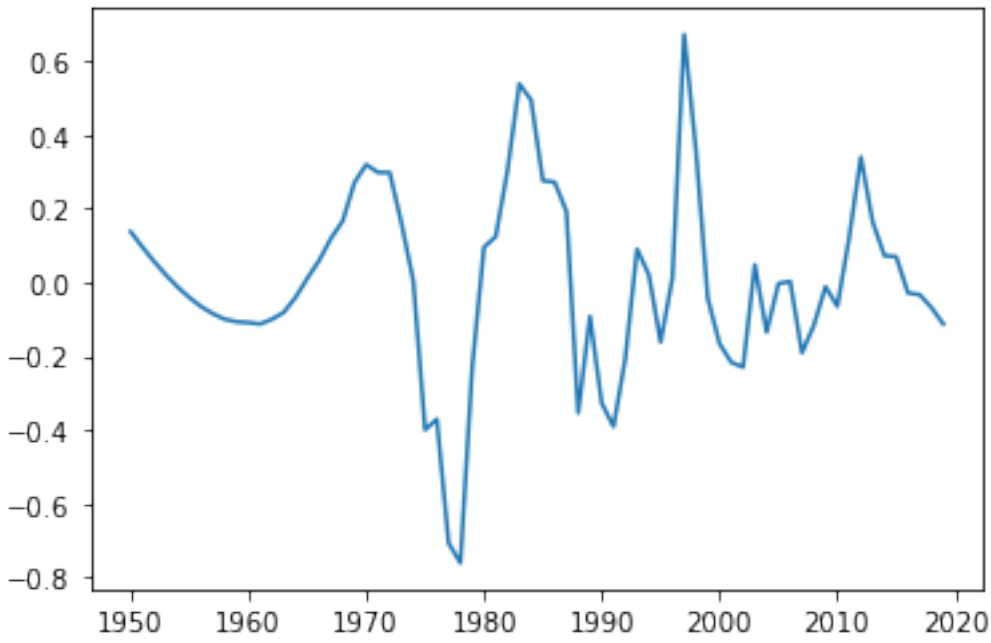
```
#Future already created in previous cell
forecast_nuc = nuclear_production(future,*params_nuc)
fit_nuc = nuclear_production(years,*params_nuc)

plt.plot(years,nuclear,label="Data")
plt.plot(future,forecast_nuc,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Nuclear Production")
```

[14]: Text(0.5, 1.0, 'Nuclear Production')



[15]:
```
#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_nuc, R_nuc = r_squared(nuclear,fit_nuc)
plt.plot(years,residuals_nuc)
print("R-Squared value: {:.4f}".format(R_nuc))
```

```
R-Squared value: 0.9949
```



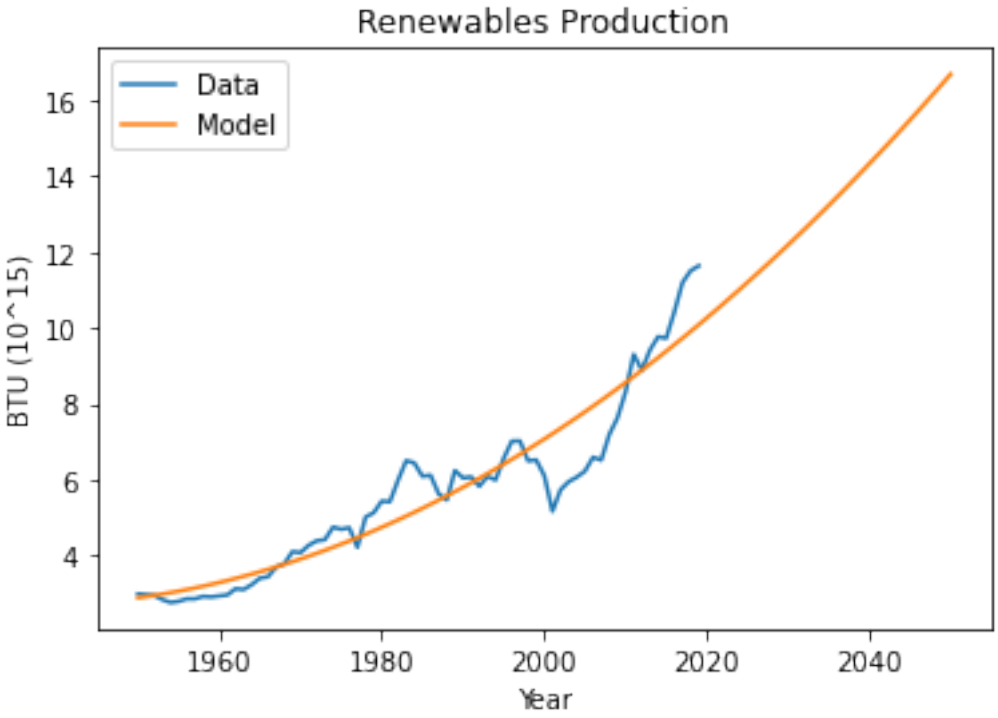
Description of Nuclear Methodology Upon first inspection, it becomes immediately obvious that US nuclear energy production is best modeled by a logistic curve, being nonexistent until ~1970, when it experienced a boom until the early 2000s, when nuclear energy fell out of favor and became stagnant. While the initial model for this data only used the generalised logistic function, due to nuclear energy's long lifespan and high cost of new construction, this resulted in an optimistic forecast, predicting no significant plant retirements in the next 30 years. In order to account for this retirement factor, a decay term was added to the model, representing old nuclear plants shutting down[11]. The robustness of this model is clearly apparent from the residuals plot and R2 value, the highest of any of the primary sources.

Parameters: * A: The lower asymptote of the logistic function * K: The upper asymptote of the logistic function * B: Slope of the logistic function (growth rate) * v: Where the maximum slope occurs * Q: Affects the value of f(0), which roughly corresponds to year 1985 * d: Rate of decay (nuclear plant retirement) * t: x-adjustment to account for the domain beginning at 1950, not 0

[16]:
```
renewables = np.asarray(energy_history["renewables"])

params_re = np.polyfit(years,renewables,2) #Quadratic regression
model_re = np.poly1d(params_re)
forecast_re = model_re(future)
fit_re = model_re(years)

plt.plot(years,renewables,label="Data")
plt.plot(future,forecast_re,label="Model")
plt.legend()
plt.xlabel("Year")
```

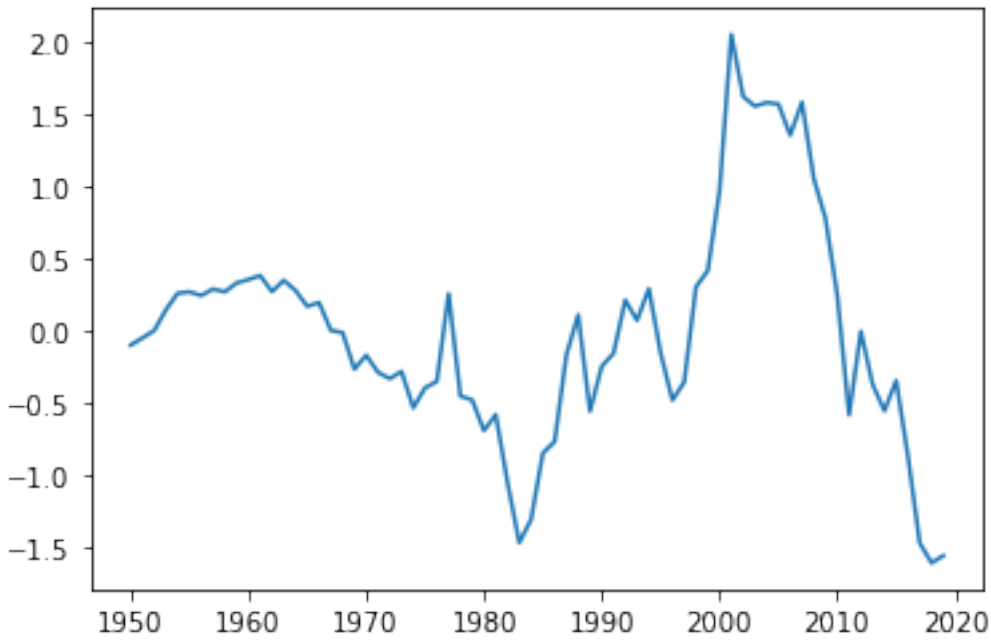
```
plt.ylabel("BTU (10^15)")
plt.title("Renewables Production")
```

[16]: Text(0.5, 1.0, 'Renewables Production')



[17]:
```
#Utilize Residuals and R-squared values to confirm goodness of fit
residuals_re, R_re = r_squared(renewables,fit_re)
plt.plot(years,residuals_re)
print("R-Squared value: {:.4f}".format(R_re))
```

```
R-Squared value: 0.8881
```

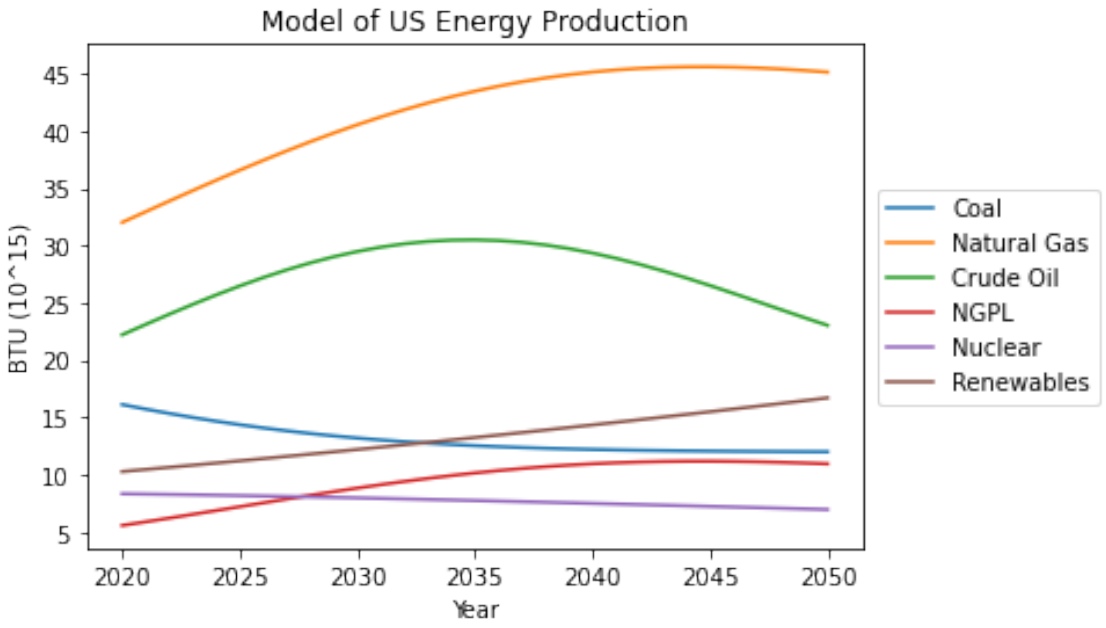


Renewables Methodology As US production of renewable energy has followed a relatively consistent upwards trend, a simple polynomial function was chosen to model future growth. However, to capture the accelerating nature of renewable energy production, a quadratic fit was chosen over a linear regression. While a more detailed model could be developed to attempt to capture all of the data's richness, the residuals plot reveals that the primary deviation from a pure quadratic curve occured in the early 2000s, when renewable energy production dropped sharply. As production reaccelerated around 2008, this drop in renewable energy production correlates strongly with the Bush Administration, however geopolitical influences are outside the scope of this investigation. As a result, the simple quadratic model is maintained.

1.3 Results

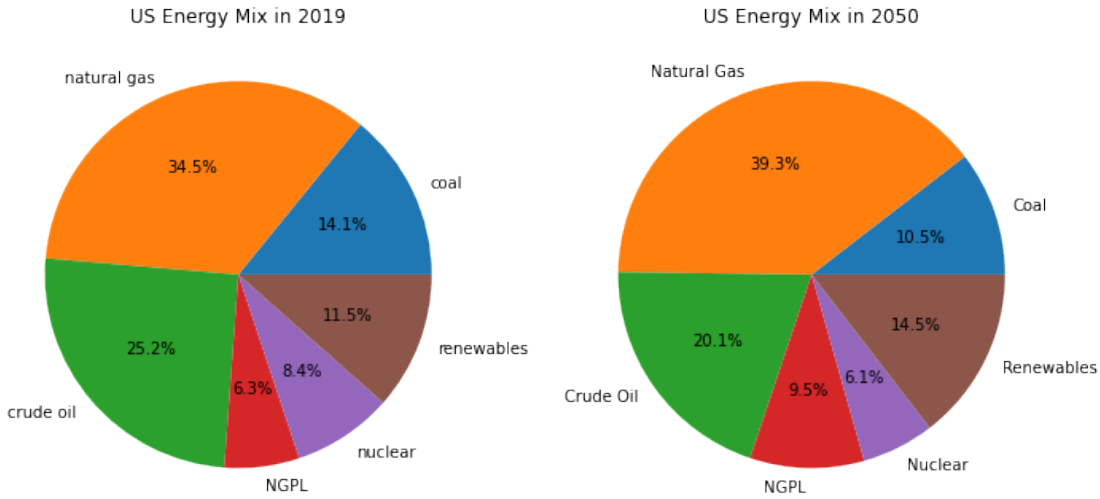
[18]:
```
forecast_data = pd.DataFrame()
forecast_data["Coal"] = forecast_co
forecast_data["Natural Gas"] = forecast_ng
forecast_data["Crude Oil"] = forecast_cr
forecast_data["NGPL"] = forecast_ngpl
forecast_data["Nuclear"] = forecast_nuc
forecast_data["Renewables"] = forecast_re
forecast_data.index = future
forecast_data.loc[2020:2050].plot() #Plot only the forecast years, 2020 - 2050
plt.xlabel("Year")
plt.ylabel("BTU (10^15)")
plt.title("Model of US Energy Production")
plt.legend(bbox_to_anchor=(1.32, .5), loc='center right') #Legend plot code [12]
```

[18]: <matplotlib.legend.Legend at 0x7facd77450d0>



Now that all of the primary energy sources for the US have been modeled, the results for the next 30 years can be analyzed in order to understand how the US energy mix will develop. Above is a chart showing the curves for each model from 2020 to 2050, illustrating both overall changes in energy production and changes in specific sources. This chart clearly illustrates that natural gas production will remain the primary source of US energy for the near future, steadily increasing until approximately 2040, at which point production begins to level off. Crude Oil experiences similar growth until 2035, at which point production begins to shrink, eventually settling near 2020 levels. Coal production decreases slightly over these 30 years, moving from 3rd place to 4th, as renewables increases steadily over the same timespan, potentially eclipsing crude oil by the 2060s. Nuclear energy experiences a less severe decline than coal, slowly fading into the smallest primary source. Finally, natural gas plant liquids benefit greatly from the increase in natural gas production, over doubling in production and rivaling coal by 2050.

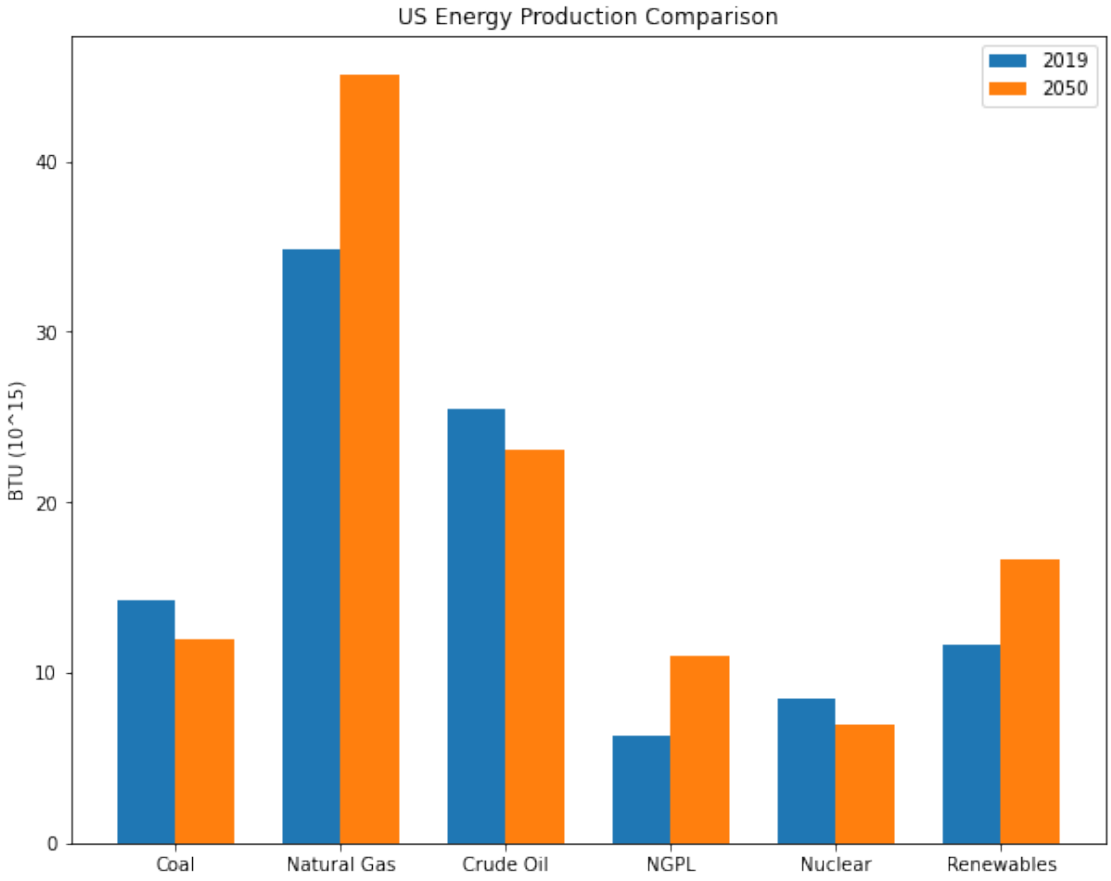
[19]:
```
fig = plt.figure(figsize=(10,10)) #Set figure size (from class assignments)
plt.subplot(1,2,1)
current_production = energy_history.drop(columns="total")
plt.pie(current_production.loc[2019],labels=current_production.
 ↪columns,autopct='%1.1f%%') #From matplotlib pyplot documentation
plt.title("US Energy Mix in 2019")
plt.subplot(1,2,2)
plt.pie(forecast_data.loc[2050],labels=forecast_data.columns,autopct='%1.1f%%')
plt.title("US Energy Mix in 2050")
plt.tight_layout()
```



By expressing each primary source as a percentage of the US's overall energy production, these developments across the next three decades are made even more clear. Natural gas cements itself as the primary source of US energy production, growing 4.8% while natural gas plant liquids increases by 3.2%. Surprisingly, much of this growth comes in the form of crowding out the other primary fossil fuels, as crude oil decreases by 5.1% and coal decreases by 3.6%. As a result of this growth and crowding out, the portion of non-fossil fuel energy production remains relatively constant, only increasing from 19.9% to 20.6%. Within this climate-friendly subset, renewables waxes while nuclear wanes, gaining 3% and losing 2.3% respectively.

[20]:
```
fig = plt.figure(figsize=(10,10)) #Set figure size (from class assignments)
labels = forecast_data.columns  #Grouped bar chart code: [13]
data_2019 = current_production.loc[2019]
data_2050 = forecast_data.loc[2050]

x = np.arange(len(labels))  # the label locations
width = 0.35  # the width of the bars

fig = plt.figure(figsize=(10,8)) #Set figure size (from class assignments)
plot_2019 = plt.bar(x - width/2, data_2019, width, label='2019')
plot_2050 = plt.bar(x + width/2, data_2050, width, label='2050')

plt.ylabel("BTU (10^15)")
plt.title("US Energy Production Comparison")
plt.xticks(x,labels)
plt.legend()
```

[20]:
```
<matplotlib.legend.Legend at 0x7facd7a52f40>

<Figure size 720x720 with 0 Axes>
```



[21]:
```
print(energy_history.tail(1))
print(forecast_data.tail(1))
print("2050 Total Production: {:.4f}".format(forecast_data.loc[2050].sum()))
```

```
         coal  natural gas  crude oil    NGPL  nuclear  renewables     total
2019  14.2681      34.8946    25.4389  6.3366   8.4624      11.637  101.0376
         Coal  Natural Gas  Crude Oil       NGPL   Nuclear  Renewables
2050  11.9983    45.126774   23.02776  10.940288  6.957552   16.695223
2050 Total Production: 114.7459
```

With the above chart and table, a very clear side-by-side comparison of 2019 energy production and predicted 2050 energy production can be performed. Natural gas, natural gas plant liquids, and renewables all experience significant growth in production, as resources increasingly shift away from coal, crude oil and nuclear. Additionally, overall US energy production is forecasted to grow by approximately 14 quadrillion BTU over the next 30 years, an increase of ~13.5%.

In order to evaluate the accuracy of these predictions, they will be compared to the forecasts compiled by the EIA in their 2020 Annual Energy Outlook[14]. While the EIA does not disclose the nature of their models, their forecast data is accessible by downloading the powerpoint presentation of the report, then opening the charts within it in Excel. For the simplicity of this notebook, the 2050 values have been transcribed below:

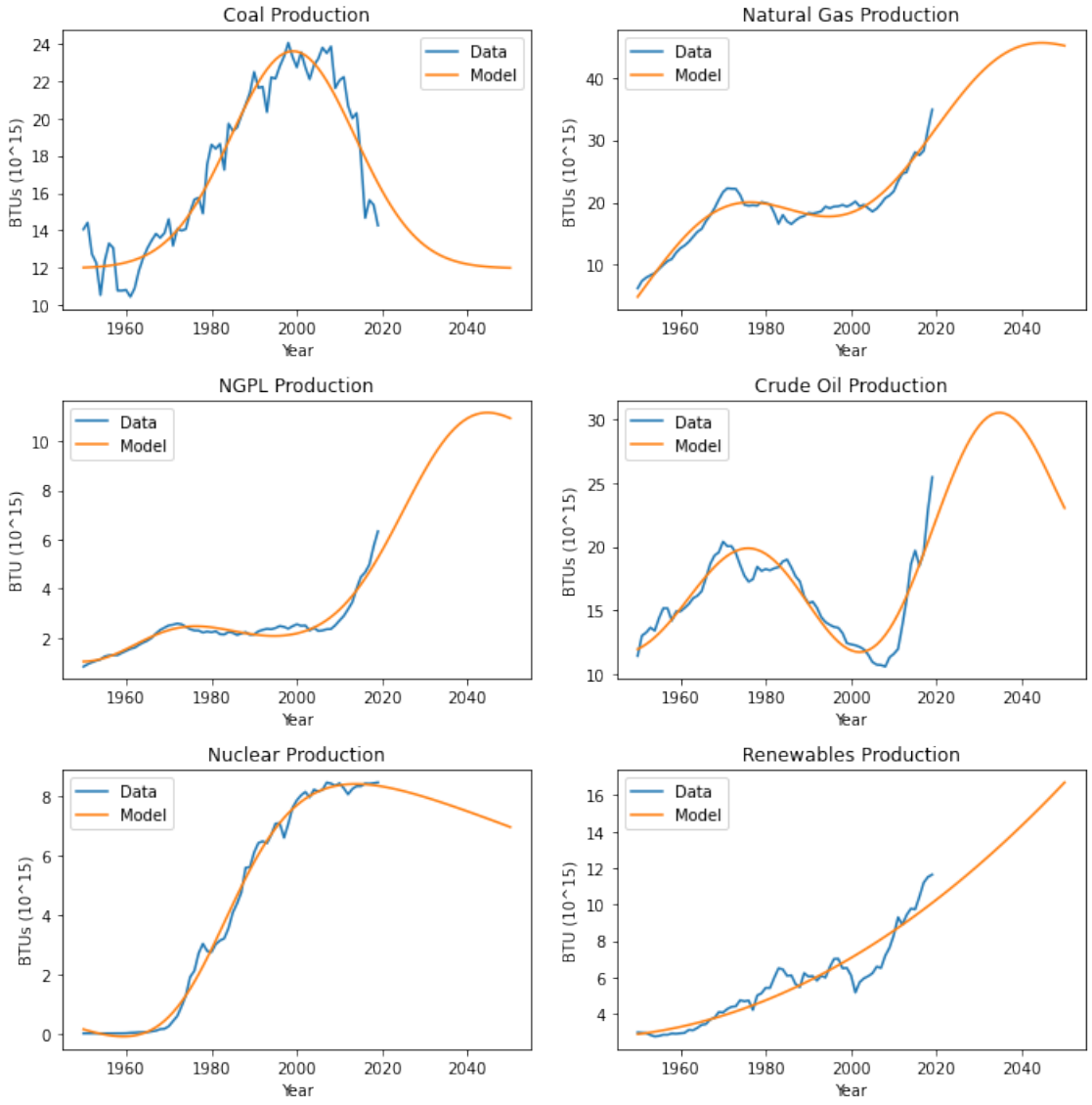
[22]:
```
model_values = forecast_data.loc[2050]
eia_cr = 24.79623   #Crude Oil
eia_ng = 46.616814  #Natural Gas
eia_re = 18.178046  #Renewables
eia_co = 10.675206  #Coal
eia_nuc = 6.716295  #Nuclear
eia_ngpl = 8.190884 #NGPL
eia_values = [eia_co,eia_ng,eia_cr,eia_ngpl,eia_nuc,eia_re]

def percent_error(model, eia):
    return (abs(model - eia) / eia) * 100

pct_error = []
for i in range(len(model_values)):
    pct_error.append(percent_error(model_values[i],eia_values[i]))

print("Coal error: {:.2f}%, Natural Gas error: {:.2f}%, Crude Oil error: {:.
 ↪2f}%, NGPL error: {:.2f}%, Nuclear: {:.2f}%, Renewables error: {:.2f}%".
 ↪format(*pct_error))
```

```
Coal error: 12.39%, Natural Gas error: 3.20%, Crude Oil error: 7.13%, NGPL
error: 33.57%, Nuclear: 3.59%, Renewables error: 8.16%
```

As can be seen above, the model's predictions for 2050 generally fall within 10% of EIA's estimates, with coal and natural gas plant liquids the two outliers. The deviation for coal is largely due to our model hitting a lower limit of ~12 quadrillion BTU, while in reality there is no such hard lower limit to coal production. Finally while natural gas plant liquid has a high percent error at 33.57%, this can be largely attributed to two main causes: low overall production and alternative uses. The actual deviation between the model and EIA's values for NGPL production is just over 2 quadrillion BTU, still the highest of any primary source, but comparable to the 1.8 quadrillion BTU deviation seen with crude oil. However, because NGPL production is generally less than 10 quadrillion BTU, this results in a much higher percent error. Secondly, the model for NGPL production was tied to natural gas production, as natural gas plant liquids are produced as a byproduct of natural gas extraction. This could result in an overestimate, since a significant portion of NGPL are not used for energy generation, but rather as building components in the petrochemical industry[15] The cell below will be used to store plots for use in the presentation, however as their material was comprehensively covered in the methodology section it will not be repeated here.

[23]:
```
fig = plt.figure(figsize=(10,10))
plt.subplot(3,2,1) #Coal Plot
plt.plot(years,coal,label="Data")
plt.plot(future,forecast_co,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Coal Production")
plt.subplot(3,2,2) #Natural Gas Plot
plt.plot(years,nat_gas,label="Data")
```

```
plt.plot(future,forecast_ng,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Natural Gas Production")
plt.subplot(3,2,3) #NGPL Plot
plt.plot(years,ngpl,label="Data")
plt.plot(future,forecast_ngpl,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTU (10^15)")
plt.title("NGPL Production")
plt.subplot(3,2,4) #Crude Oil Plot
plt.plot(years,crude,label="Data")
plt.plot(future,forecast_cr,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Crude Oil Production")
plt.subplot(3,2,5) #Nuclear Plot
plt.plot(years,nuclear,label="Data")
plt.plot(future,forecast_nuc,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTUs (10^15)")
plt.title("Nuclear Production")
plt.subplot(3,2,6) #Renewables Plot
plt.plot(years,renewables,label="Data")
plt.plot(future,forecast_re,label="Model")
plt.legend()
plt.xlabel("Year")
plt.ylabel("BTU (10^15)")
plt.title("Renewables Production")
plt.tight_layout()
```



1.4 Discussions and Conclusions

Overall, I would conclude that that a robust model of US energy production for the next 30 years has been achieved with relatively minimal data and some macro-level intuition. The curve-fitting method has delivered reasonable forecasts without accounting for complex market forces or geopolitical events. It is important to note, however, that as a result these forecasts do assume that no significant market changes or world-changing events occur in the next 30 years, something which is appearing increasingly unlikely.

The two most significant obstacles I encountered during the course of this investigation were curve_fit being unable to find an optimized solution and the fitted curve resulting in an unreasonable forecast. The first was initially discussed during the methodology for NGPL production, as despite having a very similar shape to natural gas production, the difference in scale between

the two resulted in curve_fit being unable to fit the NGPL curve. Thankfully the very strong correlation between these two variables meant that I could generate an accurate NGPL curve using said correlation and the natural gas forecast, though it took many frustrated attempts before that solution presented itself. A much more persistent issue was the wildly unreasonable forecasts often provided by the initial models. While I believe the rationale behind each function is solid, there certainly were adjustments made to bring their forecasts into plausible ranges. For example, the first function used to model natural gas production predicted natural gas output of ~75 quadrillion BTU in 2050, more than twice the production level in 2019. This went the other way as well, with quadratic and sinusoidal functions for coal production often predicting negative output.

When I initially began this project I planned to model both the prices of each primary source and the production levels, aiming to use the correlation between the two to create more robust models. This caused me a myriad of headaches and complications, as price data was much more difficult to come by, needed many unit conversions, and almost always exhibited weak to no correlation with production data. If I were to do this project again, I would drop the extraneous price aspects and focus on the core topic as soon as it became apparent that there was no significant connection between price and production levels.

According to the models, the US energy mix in 2050 will be 39.3% (43.15 quads) natural gas, 20.1% (23.03 quads) crude oil, 14.5% (16.70 quads) renewables, 10.5% (12 quads) coal, 9.5% (10.94 quads) natural gas plant liquids, and 6.1% (6.96 quads) nuclear power. This mix emerges as a result of natural gas production booming, which crowds out the other fossil fuels (crude oil and coal), while renewables grow and nuclear declines steadily.

1.5 Citations

[1] "BP Statistical Review of World Energy 2019" (PDF). https://www.bp.com/content/dam/bp/business-sites/en/global/corporate/pdfs/energy-economics/statistical-review/bp-stats-review-2019-full-report.pdf

[2] Sanchez, Bill. "Fossil Fuels Account for the Largest Share of U.S. Energy Production and Consumption." Today in Energy - U.S. Energy Information Administration (EIA), EIA, 14 Sept. 2020, www.eia.gov/todayinenergy/detail.php?id=45096.

[3] "Consumption & Production." U.S. Energy Facts Explained, U.S. Energy Information Administration (EIA), 7 May 2020, www.eia.gov/energyexplained/us-energy-facts/.

[4] R-Squared formula: https://www.investopedia.com/terms/r/r-squared.asp

[5] Rhodes, Joshua. "Is The US Coal Industry Completely Burned Out?" Forbes, 12 Feb. 2020, https://www.forbes.com/sites/joshuarhodes/2020/02/12/is-the-us-coal-industry-almost-completely-burned-out/?sh=3a146de9594f

[6] "Table A5 Approximate Heat Content of Coal and Coke Coal." U.S. Energy Information Administration (EIA), https://www.eia.gov/totalenergy/data/browser/index.php?tbl=TA5#/?f=A&start=1949&end=2

[7] "Table 6.2 Natural Gas Production, 1949 - 2011." Annual Energy Review, U.S. Energy Information Administration (EIA), 27 Sept. 2012, https://www.eia.gov/totalenergy/data/annual/showtext.php?t=ptb0602

[8] "Where our natural gas comes from." Natural Gas Explained, U.S. Energy Information Adminis-

tration (EIA), 1 Oct. 2020, https://www.eia.gov/energyexplained/natural-gas/where-our-natural-gas-comes-from.php

[9] Maverick, J.B. "How has fracking decreased U.S. dependence on foreign oil?" Investopedia, 19 July 2020, https://www.investopedia.com/ask/answers/012915/how-has-fracking-helped-us-decrease-dependence-foreign-oil.asp

[10] Generalised Logistic Formula: https://en.wikipedia.org/wiki/Generalised_logistic_function

[11] "U.S. nuclear industry." Nuclear explained, U.S. Energy Information Administration (EIA), 15 April 2020, https://www.eia.gov/energyexplained/nuclear/us-nuclear-industry.php

[12] Legend code: https://stackoverflow.com/questions/4700614/how-to-put-the-legend-out-of-the-plot

[13] Grouped bar chart code: https://matplotlib.org/3.1.1/gallery/lines_bars_and_markers/barchart.html#sphx-glr-gallery-lines-bars-and-markers-barchart-py

[14] "Annual Energy Outlook 2020." U.S. Energy Information Administration (EIA), 29 Jan. 2020, https://www.eia.gov/outlooks/aeo/

[15] "What are natural gas liquids and how are they used?" Today In Energy, U.S. Energy Information Administration (EIA), 20 April 2012, https://www.eia.gov/todayinenergy/detail.php?id=5930