Predicting Energy Consumption in California

Logan Clark [1] Sarah Kernal [2] Samantha Rehome [1] Scott Sprengel [1] Ahoora Tamizifar [3]

[1]California State University, Fullerton [2]Cypress College [3]University of California, Irvine

Abstract

With the growing demand for sustainable energy planning and resource optimization, accurate forecasting of energy consumption has become a critical task for policymakers in California. This project presents a comprehensive study on forecasting energy consumption in California using a data-driven approach that leverages historical energy usage data, climatic factors, and population. We develop and implement a class of time series models known as autoregressive integrated moving average (ARIMA) and autoregressive integrated moving average with exogenous variables (ARIMAX) to forecast consumption for different sectors and sources of energy in California. Our results reveal that ARIMA models are quite sensitive to capturing the nuances needed to forecast consumption and produce reliable prediction values. However, ARIMAX models enable us to incorporate exogenous data that can improve the accuracy of predictions. The models indicate an increasing growth in renewable energy usage, but we expect fossil fuel usage to remain constant.

Goals

Our project seeks to predict energy consumption in California in the following way:

- Predict fossil fuels and renewable energy consumption for the next ten years.
- Predict transportation and industrial sector energy consumption for the next ten years.

Dataset Description

Our data come from various public institutions, primarily the *Energy Information Administration*. The study is based on annual data from 1960/1970 to 2021 for California's energy usage, measured in trillion British thermal units (Btu). It includes data on energy sources (coal, natural gas, petroleum, biomass, geothermal, solar, wind, and hydroelectric) and sectors (commercial, industrial, residential, and transportation).

Exploratory Data Analysis

Analyzing the trends in energy consumption led us to appreciate the importance of historical events and policies.

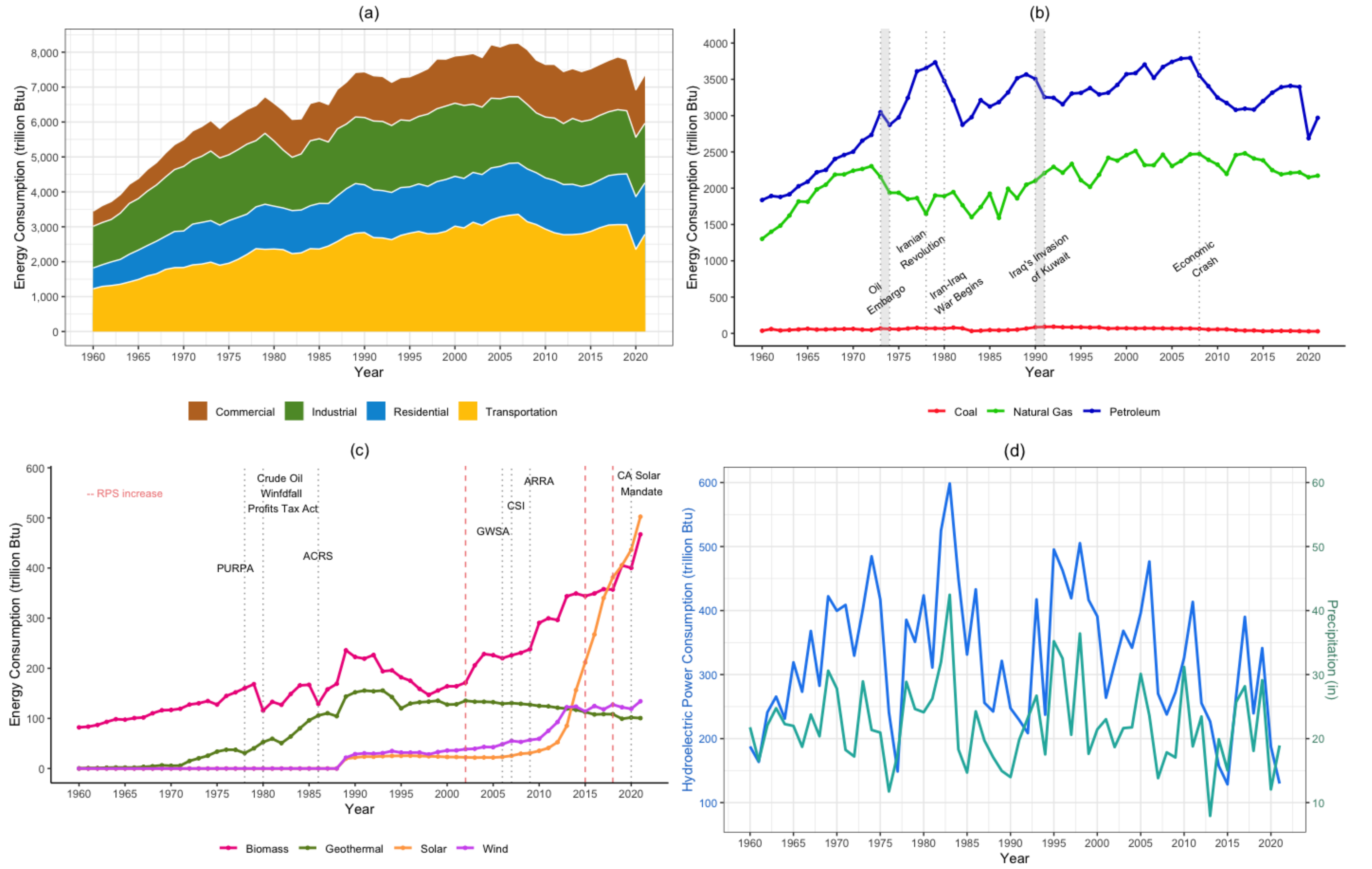


Figure 1. (a) Energy Consumption in California by Sector (b) Consumption for Different Types of Fossil Fuels in California (c) Consumption for Different Types of Renewable Energy in California (d) Yearly Hydroelectric Power Consumption vs. Yearly Precipitation

Model Implementation

The time-dependent nature of our data leads us to choose ARIMA to forecast future values. Autoregressive Integrated Moving Average (ARIMA) models consist of three parameters (p,d,q):

- p specifies the number of Autoregressive terms in the model
- d specifies the number of differencing applied on the time series values
- q specifies the number of Moving Average terms in the model

Let Y_t and ϵ_t be the differenced value and error at time t, and let μ, ϕ, θ be our intercept and coefficients. Then, the ARIMA model can be written as:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q}.$$

We also use ARIMAX models which include exogenous variables. ARIMAX can be written as:

$$Y_t = \mu + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + .. + \theta_q \epsilon_{t-q} + \langle \vec{\beta}, \vec{x_t} \rangle,$$

where $\vec{x_t}$ is our vector of exogenous predictors at time t and $\vec{\beta}$ is a coefficient vector.

Model Selection

We excluded the last 5-10 years of consumption values to use as a testing set and trained our model on the preceding values. We looked at different (p,d,q) parameters for ARIMA and also included various exogenous predictors to see which would obtain the lowest mean-square error (MSE) compared to the true testing data.

Predicting Fossil Fuel Consumption

Forecasting consumption of petroleum and natural gas utilized a (2,1,0) ARIMAX model that used population and production as exogenous predictors. Taking into consideration COVID's anomalous effect on energy consumption habits, we built another model on pre-COVID data and predicted the next 12 years.

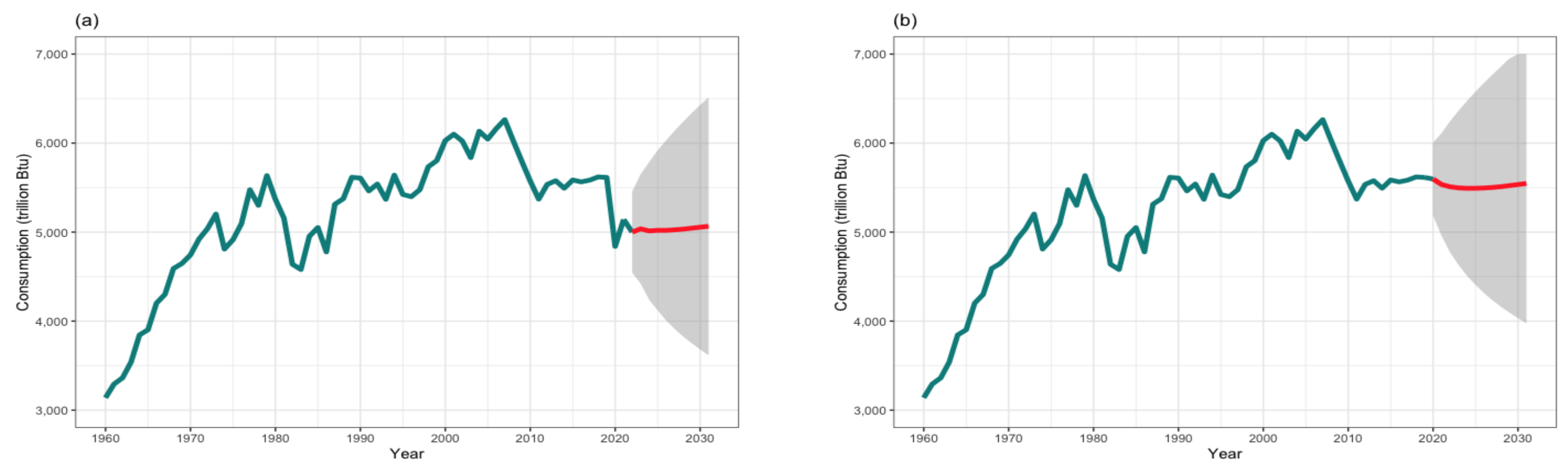
Renewable Energy Consumption

Forecasting consumption of solar, wind, biomass, and geothermal energy utilized a (2,1,2) ARIMA model, and forecasting hydroelectric consumption utilized a (1,1,1) ARIMAX model that used precipitation as an exogenous predictor.

Transportation and Industrial Sector Consumption

Forecasting consumption in transportation sector utilized a (2,1,2) ARIMAX model that used population and average energy price in that sector as exogenous predictors while forecasting consumption in industrial sector utilized a (2,1,2) ARIMAX model that used natural gas and petroleum prices as exogenous predictors.

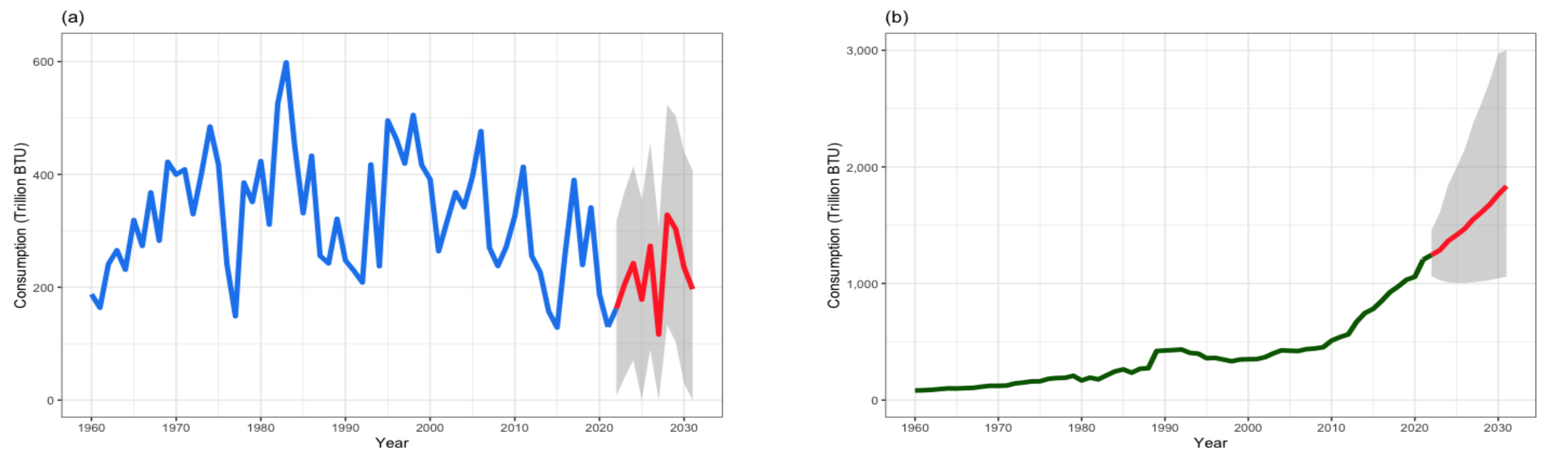
Results



(a) Including COVID Data (b) Excluding COVID Data

Figure 2. 10 Year Fossil Fuel Consumption Forecasts. Blue line represents observed values, red line represents forecast values, gray shaded area represents a 95% confidence interval for forecast values

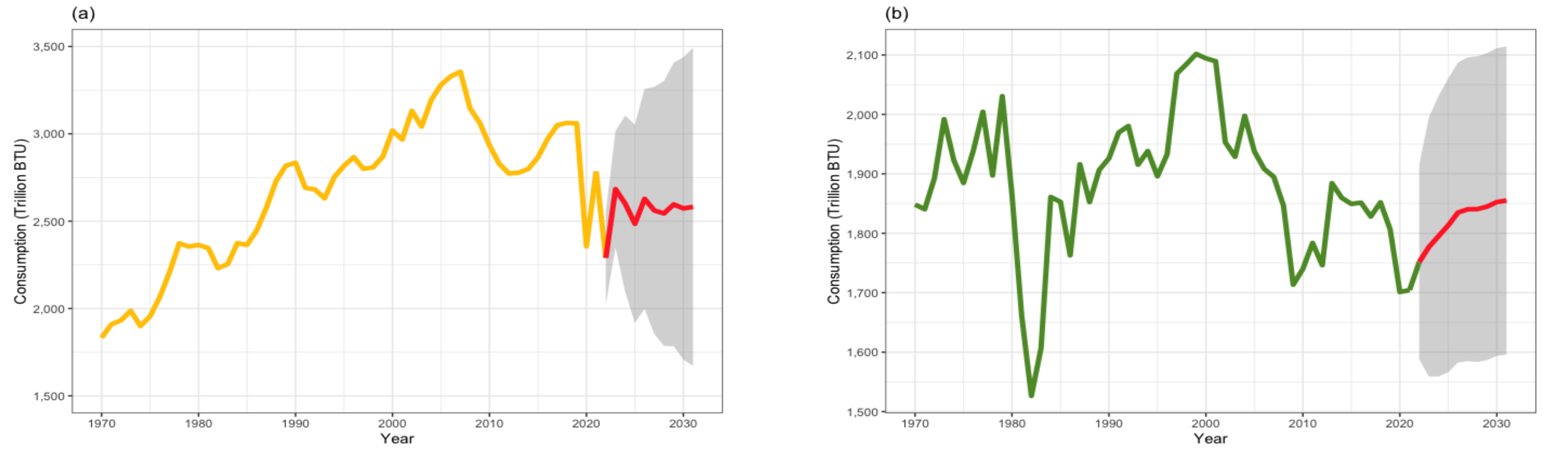
We predict fossil fuel usage will remain constant, but may decrease under new state policies. California's fossil fuel reduction efforts face challenges due to extensive transportation infrastructure and the ongoing demand for energy from various industries.



(a) Hydroelectric (b) Solar, Wind, Biomass, and Geothermal

Figure 3. 10 Year Renewable Energy Consumption Forecasts. Blue/Green line represents observed values, red line represents forecast values, gray shaded area represents a 95% confidence interval for forecast values

We expect the consumption of hydroelectric power to be volatile and weather-dependent in the future. As for the other renewable sources, *we anticipate a ceaseless rise in consumption in the next decade, led by solar and biomass energy*. California needs to be 60% reliant on renewable energy by 2030, and has committed to achieving 100% clean energy by 2045.



(a) Transportation Sector (b) Industrial Sector

Figure 4. 10 Year Sector Consumption Forecasts Yellow/green line represents observed values, red line represents forecast values, gray shaded area represents a 95% confidence interval for forecast values

Discussion: In all of our models, we assume economic stability throughout the forecast period and that no major natural catastrophes or wars will befall the state. Furthermore, ARIMA and ARIMAX models struggle to effectively adjust for the impact of COVID-19 on energy consumption due to the unprecedented disruptions caused by the pandemic. Lack of data for 2022, a year of near-normalcy, further exacerbates this issue. Additionally, some exogenous variables had less recorded years than the consumption variables, reducing the number of values our ARIMAX model was trained on. Despite all these difficulties, our models were able to perform quite well on the testing set, demonstrating the accuracy of ARIMA and ARIMAX models.

Future Work

The accuracy of our models could be further improved by using monthly data instead of annual data, allowing for the exploration of seasonal trends. Additionally, implementing Recurrent Neural Networks (RNN) could possibly yield significantly better accuracy with forecasting.

Acknowledgements

This project was funded by National Science Foundation HDR Data Science Corps Awards #2123366, #2123380, #2123384. We thank our wonderful faculty, Dr. Sam Behseta, Dr. Jessica Jaynes, Dr. Babak Shahbaba, Dr. Roberto Pelayo, Dr. Mine Dogucu, Professor Alma Castro, and teaching assistants, Brian Schetzsle and Zahra Moslemi, for their support during our research.

References

- U.S. Energy Information Administration (EIA); National Oceanic and Atmospheric Administration (NOAA)
- Methodology report: State of California, Department of Finance. Demographic Research Unit. Population Projections Methodology (2019 Baseline). Sacramento, California. July 2023.
- *Time Series Analysis: Forecasting and Control*, 5th Edition, by Box, Jenkins, Reinsel, and Ljung.