Technical Documentation

UNIVERSITY OF CALIFORNIA

Santa Barbara

Investigating the Social and Environmental Impacts of Supply Side Oil and Gas Policies in California

A Capstone Project submitted in partial satisfaction of the requirements for the degree of Master of Environmental Data Science
for the
Bren School of Environmental Science & Management

by

Mariam Garcia
Haejin Kim
Maxwell Patterson

Advisor: Dr. Paasha Mahdavi, UC Santa Barbara, Department of Political Science

Committee in charge:
Dr. Carmen Galaz García, Bren School of Environmental Science & Management
Dr. Paasha Mahdavi, 2035 Initiative

JUNE 2024

Investigating the Social and Environmental Impacts of Supply Side Oil and Gas Policies in California

As developers of this Capstone Project documentation, we archive this documentation on the Bren School's website such that the results of our research are available for all to read. Our signatures on the document signify our joint responsibility to fulfill the archiving standards set by the Bren School of Environmental Science & Management.

_____
Mariam Garcia

_____
Haejin Kim

_____
Maxwell Patterson

The Bren School of Environmental Science & Management produces professionals with unrivaled training in environmental science and management who will devote their unique skills to the diagnosis, assessment, mitigation, prevention, and remedy of the environmental problems of today and the future. A guiding principle of the School is that the analysis of environmental problems requires quantitative training in more than one discipline and an awareness of the physical, biological, social, political, and economic consequences that arise from scientific or technological decisions.

The Capstone Project is required of all students in the Master of Environmental Data Science (MEDS) Program. The project is a six-month-long activity in which small groups of students contribute to data science practices, products or analyses that address a challenge or need related to a specific environmental issue. This MEDS Capstone Project Technical Documentation is authored by MEDS students and has been reviewed and approved by:

_____
Dr. Paasha Mahdavi

_____
Dr. Carmen Galaz García

_____
DATE

Acknowledgements

This project could not have been completed without the guidance, support, and mentorship of several people at the Bren School at UC Santa Barbara, emLab, and The 2035 Initiative. A special thank you goes out to faculty advisors Dr. Paasha Mahdavi, Dr. Ranjit Deshmukh, Capstone advisor Dr. Carmen Galaz-Garcia, our clients Lucas Boys and Tracey Mangin, and Bren affiliates Samantha Csik and Kat Le. This study builds upon the findings from the *Equitable Low-Carbon Transition Pathways for California's Oil Extraction* study led by Dr. Ranijt Deshmukh and Dr. Paige Weber. We are grateful for the support and guidance that helped our team throughout this project.

Table of Contents

Abstract

This project examines how supply side oil and gas regulations impact greenhouse gas emissions, employment, and the health of communities in California living near oil wells. The project is especially focused on the implications of Senate Bill 1137 (SB1137), which would prohibit the construction of new oil and gas wells within 3,200 feet of schools, hospitals, and other sensitive receptors. This policy aims to mitigate the adverse effects of oil well pollution, which disproportionately harms disadvantaged communities throughout the state. The California public will vote on whether or not to implement SB1137 in a referendum vote in November 2024. Through the adaptation and extension of an existing workflow, the project statistically evaluates the environmental, health and labor effects of the 3,200 foot setback policy specified by SB 1137. Machine learning methods are incorporated to enhance the predictive accuracy of oil well operations and outcomes through 2045. An interactive dashboard is developed to present the findings in an accessible way to policymakers, advocates, and the public that will be voting on SB1137. Overall, this project describes the potential of supply side oil and gas regulations to reduce harmful emissions and health risks at the expense of fewer employment opportunities, equipping policymakers and the public with data-driven insights to support sustainable environmental practices.

1. Executive Summary

About 2.1 million Californians, predominantly from low-income and underrepresented communities, live within at least one mile of at least one active drilling well (Czolowski et al. 2017). Residents living near oil well activity are exposed to higher levels of air pollutants linked to asthma, cancer, cardiovascular diseases, preterm birth, and other long-term health effects (Zhang 2021). Supply side oil regulations, such as setback policies – in which oil and gas production is prohibited and or restricted, incorporating a given distance – have proven to be an effective way to reduce oil production and improve health outcomes of communities near producing wells (Lewis et al. 2018). One such policy is Senate Bill 1137 (SB 1137), passed in 2022 to prohibit new oil and gas wells within 3,200 feet of sensitive receptors, such as schools, hospitals, and residential communities (SB1137, 2022). Existing wells within this range are also subject to strict regulation under the bill, and future regulations can expand the impact of this policy to apply to existing wells (SB1137, 2022). SB 1137 would require existing facilities within a health protection zone, the area within 3,200 feet of sensitive receptors, to develop leak detection systems for harmful chemicals, include detailed spill response plans, and compliance with air district requirements . Moreover, existing facilities would adhere to California's Air Resource Board and Water Resource performance standards for their emissions detection system (SB1137, 2022).

Living near active and idle wells can increase the risk of harmful exposure to particulate matter concentration. Unsafe levels of PM 2.5 have been correlated with respiratory diseases and hospitalizations, placing the health of predominantly disadvantaged communities at elevated risk (Stanford University 2021). The setback mandate would reduce Particulate Matter PM2.5 exposure for communities in or near active oil drilling areas. An existing model developed by emLab simulates the emissions, health, and labor outcomes with respect to setback distances of 1000, 2500, and 5280 feet, but does not include results for the setback distance of 3200 feet. To gauge the impact of SB 1137, the existing model has been modified by adding an additional setback scenario to reflect and predict the impact of the Bill. This is all the more salient given a referendum on the November 2024 ballot to repeal SB 1137.

The project has three main objectives: (1) rerun the existing model while adding an additional scenario in order to calculate emission, employment, and health outcomes under a 3,200-foot setback scenario; (2) predict the number of new and idle wells in each oil field through 2045 by updating the entry and exit models from a Poisson specification to a machine learning approach using a Random Forest specification; and (3) produce a publicly-accessible online dashboard to make impacts and findings available to the public to inform Californians on the implications of SB 1137. This project seeks to bridge the gap between the previous work done by the clients and the need for accessible, public-facing material to inform Californians on the importance of SB 1137 in light of the upcoming referendum. While the Bill includes both oil and gas wells, this project focuses strictly on oil production.

The development of the public-facing dashboard is a key component of this project, as it will provide an interactive platform for users to digest the effects of the 3,200 foot setback at both the state and county levels through summary statistics, time series plots, and an interactive well map. The dashboard contains visualizations of the impacts of the 3,200 foot setback imposed on new wells from the forecasted period of 2020 to 2045 through plots that show production, worker compensation, and avoided mortality costs associated with an increase in health outcomes under a 3,200 foot setback scenario and a business-as-usual (BAU) scenario. These findings provide insights into potential implications of the Bill, with additional analysis done to understand the potential implications of the 3,200 foot setback on disadvantaged communities in California. The dashboard can increase awareness and understanding of the impact of the setback distance associated with SB 1137, which can help voters gain familiarity

with the Bill's implications. With a crucial referendum vote on this matter scheduled for November, which was pushed to a referendum due to lobbying efforts by the oil industry involving a $20 million campaign to gather signatures to push the Bill to a veto referendum, the findings of this project have the potential to influence public opinion and the outcome of this vote to uphold the Bill (CIPA, 2023). This project stands at the intersection of environmental science, public health, and data science, offering a unique opportunity to influence policy decisions and public opinion on one of the most pressing issues of our time.

2. Problem Statement

Under minimal change in supply-side policy adjustment, the effects of the oil industry will further harm Californians, especially underrepresented communities, with the continuation of oil extraction. Senate Bill 1137 is an important legislation to reduce emissions and pollution from oil production and protect the health of people living near active oil fields. As the setback distance increases, the reduction in production, and the subsequent health and environmental benefits, become more pronounced. Deshmukh et. al finds that in a no-supply-side policy business as usual scenario, greenhouse gas emissions decline by 53%, whereas a setback policy of one mile achieves a forecasted 75% reduction by 2045 (Deshmukh et al., 2023). A larger setback distance prevents more wells from being drilled near sensitive areas, thus limiting the areas where oil might be extracted. The reduction in oil production due to setbacks directly translates to a decrease in emissions and subsequently improved health outcomes due to lower PM2.5 exposure.

3. Specific Objectives

The workflow developed by the clients simulates the emissions, health, and labor implications of setback distances at 1,000 feet, 2,500 feet, and 5,280 feet. In order to understand the implications of SB 1137, a new scenario has been added to the workflow representing the 3,200 foot setback distance. Research on the implications of SB 1137 lacks public-facing material that can inform Californians on the importance of this vote. The goals of this project are as follows:

1. Updating the existing workflow to calculate the impacts on emissions, employment, and health of a 3,200-foot setback from 2020 through 2045.
2. Predict the number of new and idle wells in each field from 2020 through 2045 by updating the entry and exit models from a Poisson specification to a machine learning approach using a Random Forest specification
3. Produce a publicly accessible interactive online dashboard to display findings and provide voters with the opportunity to make an informed decision about the referendum on SB 1137.

4. Solution Design and Results

4.1: Overall strategy

The project is broken down into three phases: updating the workflow to incorporate the 3,200 foot setback (Phase 1), developing a predictive model using machine learning techniques to have more accurate projections of well entry and exit (Phase 2), and creating an interactive web dashboard to display the associated implications of SB 1137 to the general public (Phase 3).

4.2: Data and Metadata

The workflow incorporates many types of data. The Github repository contains the scripts used to execute the end-to-end workflow. The extraction model incorporates extraction and scenario factors, sourced from various categories such as well-field location and production data. It encompasses input, intermediate, and output data across both public and private subfolders.

Extraction datasets cover various aspects of oil and gas well operations in California, including geographical information, production, injection data, and cost-related metrics. The data utilizes structured formats like CSV files and Excel spreadsheets, offering comprehensive information on well locations, types, and administrative boundaries of oil fields. Historical data from 1977 to 2019 provides a longitudinal view of well-specific oil and gas production and injection activities. Data from Rystad details the financial aspects of the industry, including capital expenditures (CapEx), operational expenditures (OpEx), and oil price historical values and projections. Production and injection volumes are measured in barrels of oil (bbl) for liquid hydrocarbons and thousand of cubic feet (Mcf) for natural gas, with water injection volumes measured in gallons. Economic figures are presented in US dollars for CapEx, OpEx, and government amounts. Spatial data utilize latitude and longitude coordinates, with areas measured in square miles. Oil prices are documented in dollars per barrel for both West Texas intermediate (WTI) and Brent crude, allowing for an evaluation of operational efficiencies, financial planning, and environmental impacts in the industry. Data used in the health analysis include CalEnviroScreen 3.0 data, predicted annual GDP (Gross Domestic Product) and mortality rates by county. The CalEnviroScreen 3.0 variables are thoroughly explained in the metadata file. Variables of interest include a binary variable indicating whether a given census tract is disadvantaged or not, its associated pollution burden score, and particulate matter score which measures annual particulate matter concentrations in the census tract. Moreover, this subsection of data will also contain datasets with information on population demographics, health indicators, income levels, and disadvantaged populations. Population projections and predictions are based on county-level, and grouped by age. The time of reference for this population prediction spans out to 2057. The age groups dataset assigns a unique variable code to each age group.The data type is predominantly CSV files. Data used in the labor analysis are at the county level and are proprietary. However, data can be accessed through a license with IMPLAN. To construct its underlying data, IMPLAN draws on over 90 sources of information including the Quarterly Census of Employment and Wages (QCEW) from the Bureau of Labor Statistics, County Business Patterns from the Census Bureau, and the Regional Economic Accounts (REA) from the Bureau of Economic Analysis. Its output from the multiplier relies on inputs based on industry type, and this multiplier can help calculate the full-time employment calculation.

The README file of the project repository has been expanded from the original version to document the datasets that have been updated by the addition of the new setback scenario to the workflow.

4.3: Updating Workflow with New Setback Scenario

The first phase of the project, building off of the work done by emLab, involves rerunning the end-to-end workflow by restructuring and reading in data, adding in a 3,200 foot setback

distance layer to go along with the 1,000, 2500, and 5,280 foot buffer distances used in the original study, and confirming the validity of the results for the new setback distance. Since the data paths were defined as the paths on local machines for the majority of the scripts, a key part of building this reproducible workflow is storing and reading in data in a way that makes it easy for future users to run each of the scripts. This has been achieved by storing the data and updating their working directory or path to the repository where the data and scripts will be contained. For internal future work, the data folder can be downloaded and moved into the project repository for seamless integration.

The next step of Phase 1 involves adding the new setback distance of 3,200 feet into the workflow. The setback distances are constructed by creating buffers around the sensitive receptors with the R *sf* package. Those outputs are used in the workflow and to produce energy, labor, and health outputs. So far, the 3,200 foot setback outcomes in the figures recreated from the Nature Energy paper show appropriate results, with production, health, and labor implications falling between 2,500 feet and 5,280 feet scenarios (Deshmukh et al., 2023). Furthermore, the testing section looks into the amount of area covered by the 3,200 foot setback in oil well fields, with the amount of coverage lining up with expectations.

Since most of the code in the original project was completed by 2021, there have been several package updates since the time the workflow was originally created that required debugging efforts to connect scripts together to recreate the previous results and incorporate the new setback scenario. The *data.table* package is used extensively in the code since the library is effective at handling large data tables like the ones used in the project. There are five main updates to the code:

1. **Calling *dplyr* for select() and filter() operations**: In several scripts, such as health_data.R, clean_doc_prod.R, zero_prod.R, income_data.R, create_entry_econ_variables.R, and load_input_info.R, the *dplyr* package was called to perform select() and filter() operations. This update streamlines data manipulation and improves code readability, and was necessary to avoid using erroneous functions.
2. **Converting data frames to *data.table* objects**: In multiple scripts, including opgee-carb-results.R, predict_existing_production.R, and fun_extraction_model_targets.R, data frames were converted to data.table objects using setDT() or as.data.table() functions. This conversion was necessary to perform specific data manipulation tasks efficiently and fix errors from invalid *data.table* operations attempted on data frame objects, especially after melting or merging operations.
3. **Handling missing data with na.rm = TRUE**: In the ica_multiplier_process.R script and potentially others, the na.rm = TRUE argument was added to summarize() and sum() functions to remove missing values. This ensures that missing data is handled consistently across the code, preventing incorrect results due to NA values.
4. **Updating column names and data types**: Across several scripts, such as ica_multiplier_process.R, rystad_processing.R, and load_input_info.R, column names were updated to reflect the actual names in the input files. Additionally, data types were modified when reading in data files to ensure compatibility and consistency throughout the analysis.
5. **Replacing deprecated functions**: In multiple instances, deprecated functions were replaced with their updated counterparts. For example, in the social_cost_carbon.R script, the melt() function from the data.table package was replaced due to its deprecation. Similarly, in the load_input_info.R script, the r*ead.xslx* function was replaced with *readxl* or *read_excel* to read in data files.

Many adjustments and testing was done to the code in order for the end-to-end workflow to run effectively and recreate the results from the original project. Table 5, located in the Appendix, contains information on all of the updates made to the scripts to allow the workflow to

run effectively. Note that while these updates make it unlikely that new package-related issues will arise, it will be important to consider updates to the *data.table* and *sf* packages in particular as they are two of the main libraries utilized in the workflow. To avoid further package dependency issues for future users, the environment in which the code for this project has been updated is stored and can be easily reactivated.

4.4: Recreating Figures and DAC Investigation

Visuals in the original study have been recreated to ensure accuracy and validity of the results generated from incorporating the 3,200 foot setback distance. Since the project entails many input, processed (or intermediate) and output data, plots utilizing data from each of the three stages is presented. The plots created in this document rely on the Poisson models for well entry and exit, rather than the newly developed Random Forest models. This is expanded upon in Section 4.5.

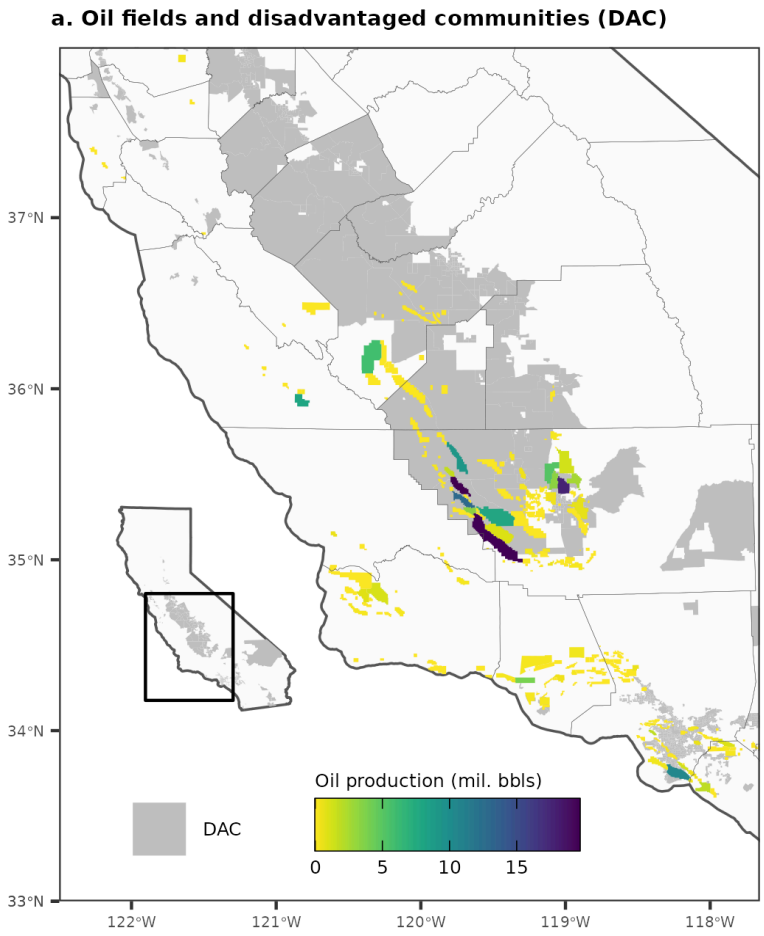

Figure 1: **Mapping the geographic distribution of active oil fields across census tracts in Central and Southern California**. This map visualizes the disproportionate burden borne by disadvantaged communities (DACs) from oil production activities in 2019. Oil fields are color-coded based on their production levels, with darker colors indicating higher production volumes (in millions of barrels). The gray areas represent census tracts classified as disadvantaged. This figure shows the concentration of oil production within or near these communities, emphasizing the need for targeted policies to address environmental justice issues and mitigate adverse health and environmental outcomes in these areas. The inset map provides a broader context of the area of interest within California.

The figure above confirms that the oil production data has been successfully loaded into the Bren Taylor server which the coding was completed on, which is crucial as this dataset is the primary driver for the entire workflow. Given the size of the datasets, some comprising millions of rows, this visual verification ensures that no issues arose during the data upload process onto the Taylor server. Oil production data from the Department of Conservation (DOC) involved uploading numerous extensive datasets into the server, making it essential to validate that the data has been read in accurately and is ready for processing. The visual match with the original study provides assurance that the data ingestion phase has been completed correctly, allowing the analysis to proceed.

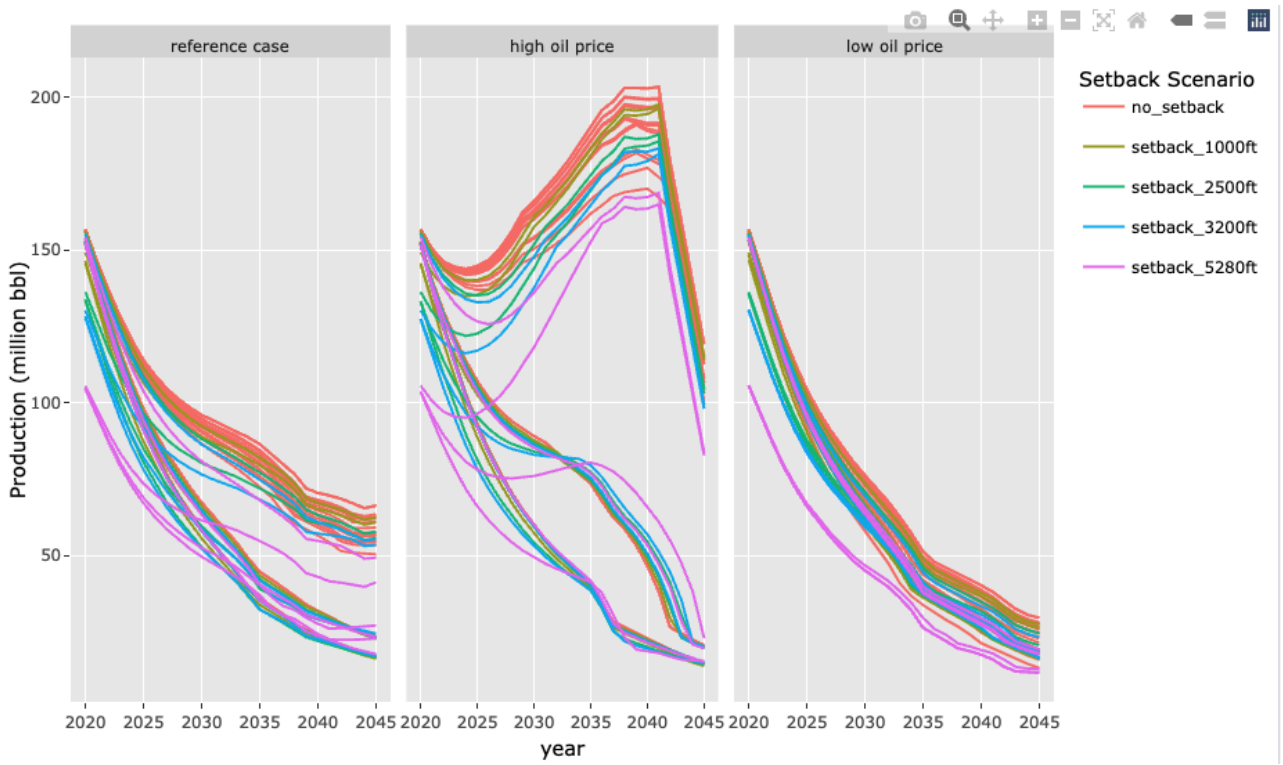

Figure 2: **Visualizing forecasts of oil production from 2020 to 2045 under low, reference, and high oil price scenarios using Poisson well entry and exit models**. Each line represents a different combination of scenarios, including innovation, carbon price, carbon capture, production quota, and excise tax. While these scenarios are not used directly in this capstone project, the values for excise and carbon tax are used in plots later on to communicate the benefits of the setbacks. The lines labeled for each setback scenario demonstrate that increasing the distance of the setback results in lower production levels for each of the oil price scenarios.

The image above confirms the validity of the new results as the values align with what the client generated in the original study. Also, the production values for the non-3,200 foot setbacks fall in line with what is expected based on its location in the setback distribution. The visual shows the impact that the setback has on future production: a larger setback decreases oil production over time.

To assess the validity of the results of the recreated workflow, figures from the Nature Energy study done by the clients are recreated to check the accuracy of the existing setback scenarios as well as the new 3,200 foot distance.

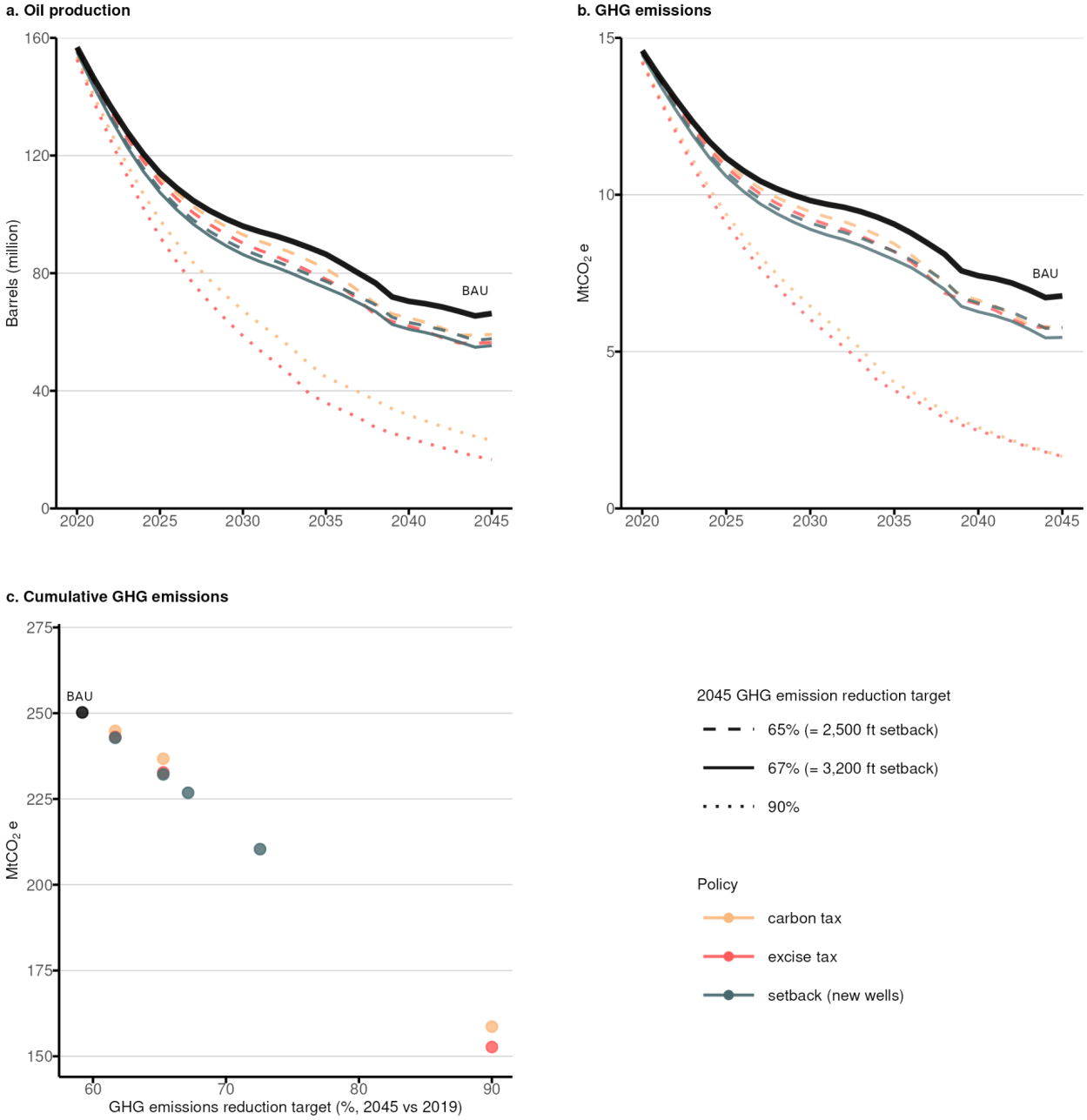

Figure 3: **Projected trajectories of California's oil production and related greenhouse gas emissions under various policy scenarios.** This figure compares the annual oil production and GHG emissions in California under a business-as-usual (BAU) scenario and three different policy approaches: setbacks for new wells (3,200 foot distance), an excise tax on oil production, and a carbon tax on emissions from oil extraction. Panel (a) shows the annual oil production, in millions of barrels) from 2020 to 2045, visualizing the effects of each policy compared to a BAU scenario. The solid green line represents the amount of oil production under the setback imposed on new wells. The dotted lines represent the associated production

under carbon and excise taxes to achieve a similar 65% GHG reduction as the 2,500 foot setback. These lines show that a setback policy that achieves the same GHG reduction produces the least oil out of the three policies. Panel (b) displays the corresponding annual greenhouse gas emissions, in million metric tons of CO_2 equivalent. Panel (c) presents the cumulative greenhouse gas emissions reduction targets for 2045, indicating the effectiveness of each policy in achieving various reduction percentages relative to 2019 levels. The dotted lines in panels (a) and (b) represent the projected oil production levels under policy scenarios where the greenhouse gas emission reductions achieved are equivalent to those expected from implementing a 2,500-foot setback regulation, but instead achieved through the use of carbon taxes or excise taxes on oil production.The orange and red dots in panel (c) correspond to the excise tax and carbon tax associated with the equivalent GHG emission reduction of 1,000 and 2,500 foot setbacks. The two points furthest right represent the DAC share of benefits under a 90% GHG reduction scenario under carbon and excise taxes.

The curves and points in the figure above align precisely with the results from the Nature Energy study, confirming the accuracy of the generated outputs. Notably, the new 3,200-foot setback scenario, illustrated by the second furthest grey dot from the right in panel (c), demonstrates a significant reduction in cumulative greenhouse gas emissions. This addition enriches the analysis by offering insights into the impact of increased setback distances compared to the existing scenarios, emphasizing its effectiveness in achieving substantial GHG reductions by 2045.

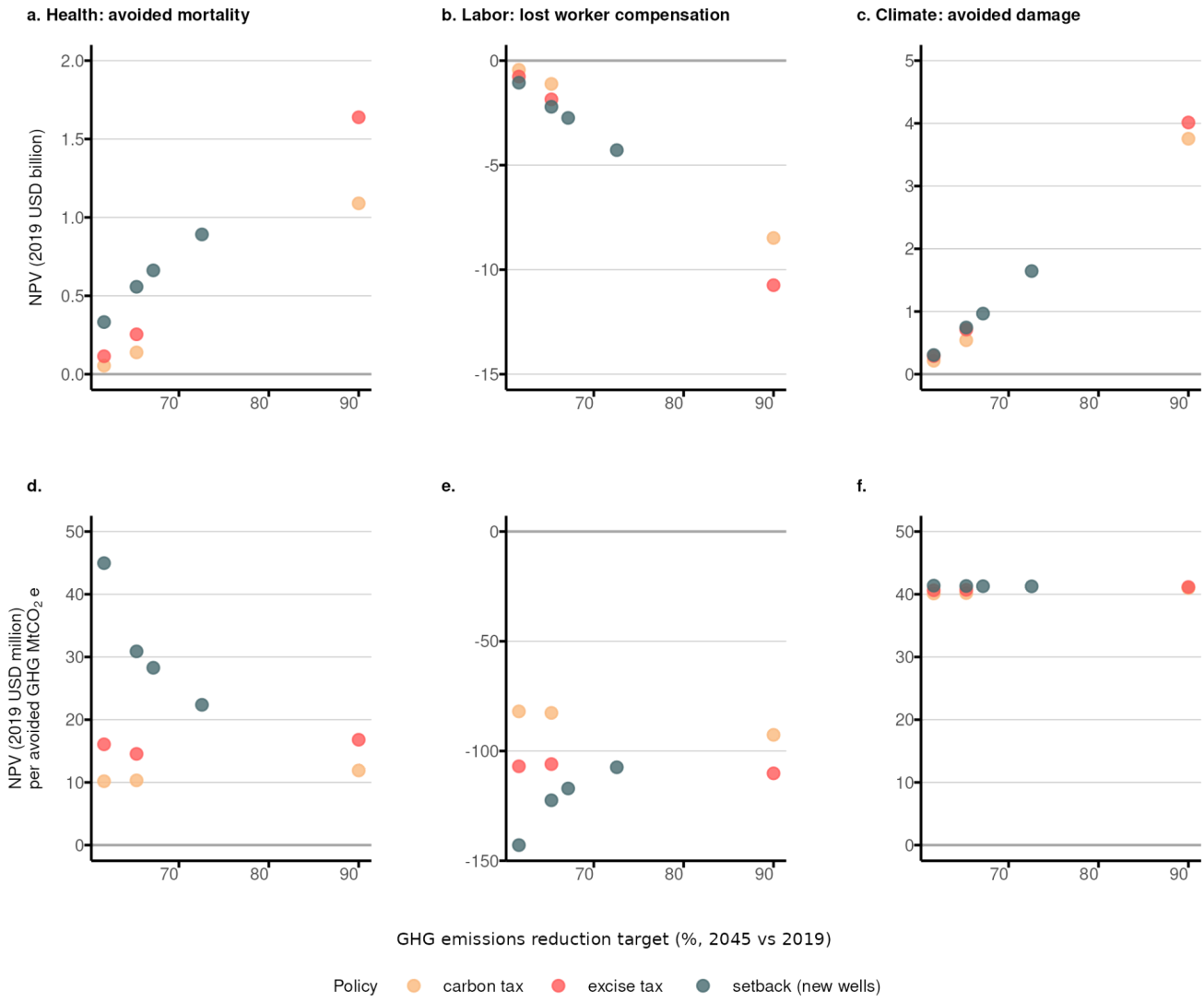

Figure 4: **Evaluating the effectiveness of California's oil-production policies in terms of health, economic, and environmental outcomes.** This figure compares the projected impacts of three different policy approaches – setbacks for new wells (with the 3,200 foot setback distance included), an excise tax on oil production, and a carbon tax on emissions from oil extraction – against a business-as-usual (BAU) scenario. The Poisson model for well entry and exit is used in this plot, as the new well estimates from 2020 to 2023 are more closely aligned with this model than the random forest estimates. The upper row (a-c) showcase the total benefits accrued between 2020 and 2045, including the prevention of premature death (a), the reduction in worker compensation (b), and the avoidance of climate-related damages measured using the social cost of carbon (c). The bottom row (d-f) shows the same benefits normalized by the total reduction in GHG emissions achieved from 2020 to 2045. The orange and red dots correspond to the excise tax and carbon tax associated with the equivalent GHG emission reduction of the 1,000 and 2,500 foot setback distances. The two points furthest right represent the DAC share of benefits under a 90% GHG reduction scenario under carbon and excise taxes. This comparison shows that the setback policy has a greater impact on avoided mortality and avoided climate damage than the tax-based policies. All monetary values are reported as net present values in 2019 US dollars, calculated using a discount rate of 3%.

The figure above is a key figure in regards to confirming the validity of the recreated results, and understanding the impact of the 3,200 foot setback scenario. The figures that have been recreated from the Nature Energy paper implement the setback starting in 2020 as this

was how the setback was implemented in the original project done by the clients. This figure above is created in the *figure3.R* script. The 3,200 foot setback points reflect the total impacts of the policy on new wells from 2020 to 2045. The *npv_dt* dataframe from this script calculates the total values for each of the categories in the plot: health, labor, and climate. **From 2020 to 2045, the model estimates that a 3,200 foot setback imposed on new wells would avoid $662,142,900 in mortality costs, avoid $966,669,540 in climate damage, and create $2,741,446,700 in forgone wages in total across California.** Note that all of these figures are measured in 2019 USD net present value. These values represent the points of 3,200 foot setback on new wells in the plots in the top row. The bottom row represents the net present value of each category per avoided megaton of carbon dioxide emitted, with the net present value being in 2019 value terms. For each megaton of CO2 emissions avoided, the 3,200 foot setback on new wells is estimated to provide $28,278,249 in avoided mortality costs and $41,284,858 in avoided climate damage, while resulting in $117,079,394 in forgone wages. These outputs are associated with a 67.15% reduction in greenhouse gas emissions in 2045 compared to 2019 levels.

The health benefits (subplot (a) in Figure 4) exhibit a significant increase with higher GHG reduction targets, particularly for the setback policy, which outperforms the carbon and excise taxes of equivalent GHG levels in regards to health benefits. The same level of GHG reduction achieved through excise and carbon taxes does not translate into equally substantial health benefits in terms of avoided mortality as benefits that arise from the setback policies. This discrepancy can be attributed to the distinct ways these policies target emission sources and their broader public health impacts. While excise and carbon taxes focus on reducing overall emissions, setback policies specifically limit the number of wells near residential areas. Consequently, this leads to reduced well production and lower PM2.5 levels in these critical regions, thereby yielding greater health benefit since production is lowered in places with her population density.

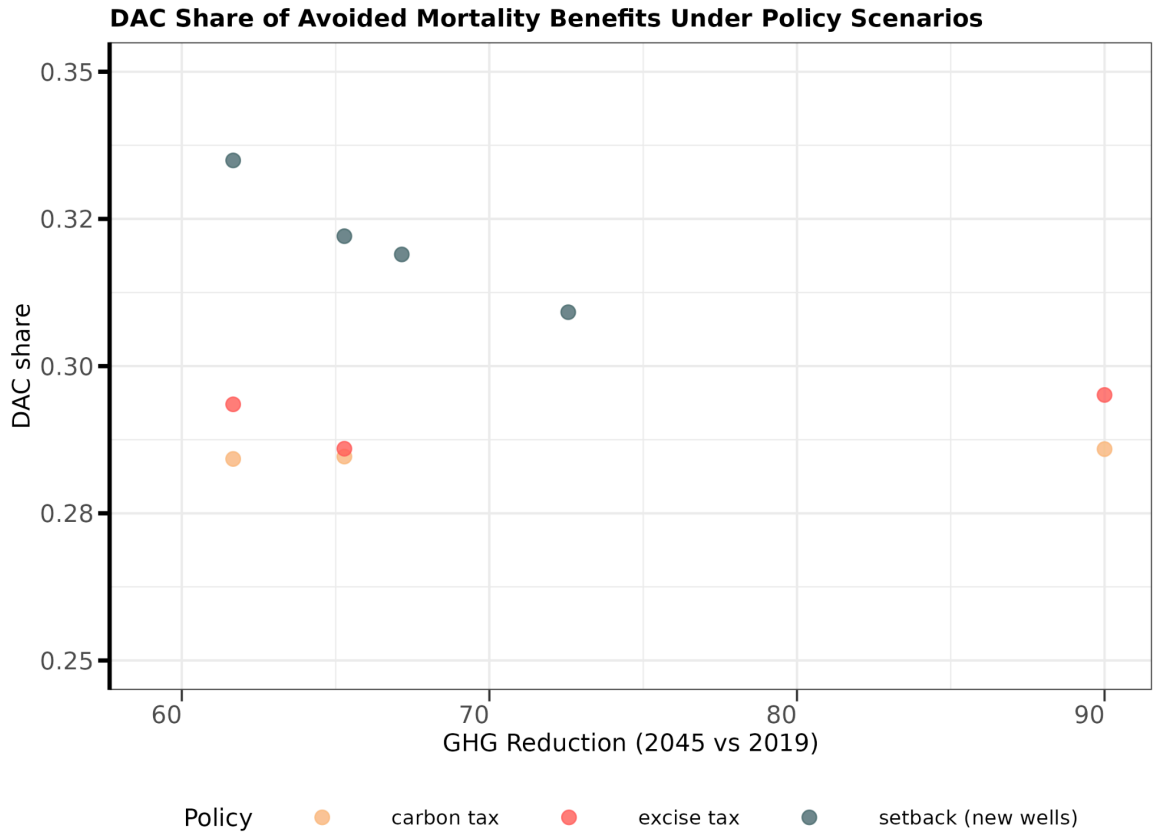

Figure 5: **Comparing the distribution of health benefits across disadvantaged communities under different oil production policies**. The scatter plot shows the relationship between the stringency of greenhouse gas reduction targets for 2045 and the share of avoided mortality benefits taken on by disadvantaged communities under three policy scenarios: carbon tax, excise tax, and a 3,200 foot setback on new wells. The setback policy consistently results in higher DAC share of avoided mortality compared to the other two policies. For example, the 3,200 foot setback scenario, represented by the third gray dot from the left, demonstrates a notably higher DAC share of avoided mortality compared to the tax policies at a similar emissions reduction level. As the setback distance increases and the GHG reduction target becomes more strict, the DAC share of avoided mortality under the setback policy gradually decreases, converging towards the shares observed under the tax policies. The orange and red dots correspond to the excise tax and carbon tax associated with the equivalent GHG emission reduction of 1,00 and 2,500 foot setback distances, showing that the disadvantaged community share of the benefits under equivalent tax scenarios are not as significant as under the setback scenarios. The two points furthest right represent the DAC share of benefits under a 90% GHG reduction scenario under carbon and excise taxes. This plot shows the potential for setback policies to disproportionately benefit DACs in regards to health outcomes, especially at moderate setback distances and emissions reduction targets.

The figure above illustrates the role of setback policies in providing health benefits to disadvantaged communities (DACs). It shows that the 3,200-foot setback on new wells consistently results in higher DAC shares of avoided mortality compared to carbon and excise tax policies that achieve equivalent GHG reductions. This result suggests that physical restrictions on oil well proximities can directly improve health outcomes for vulnerable populations by reducing exposure to harmful pollutants. While the relative advantage of setback policies diminishes as greenhouse gas reduction targets become more stringent, the plot highlights the importance of integrating setback policies with broader emissions reduction strategies to ensure that environmental regulations benefit the most vulnerable communities effectively. This balanced approach can promote both immediate health improvements and long-term environmental justice.

The differences between setback policies and tax policies are important to note in their approach and impact. Setback policies target specific regions near sensitive areas, forcing producers to avoid certain areas. This in turn protects communities living near active oil fields. Setback policies benefit DACs, which are often situated close to oil production sites, more significantly from reduced exposure to harmful pollutants. In contrast, tax policies, whether carbon or excise, do not impose geographical constraints. Instead, they allow producers to decide where to limit production, which may not necessarily reduce the disproportionate impacts on DACs. This flexibility can result in uneven benefits, as producers might prioritize reductions in less populated or less vulnerable areas to minimize their tax liabilities. Furthermore, the setback policy's targeted nature means it can effectively address the localized environmental justice issues that DACs face. Since a higher proportion of DACs are located near oil wells, implementing setbacks directly improves health outcomes by reducing pollutant exposure in these specific communities. Conversely, while tax policies can drive overall emissions reductions, their indirect approach may fail to address the concentrated health risks experienced by DACs. Thus, the setback policy is more adept at mitigating the disproportionate impacts imposed on these vulnerable communities, highlighting the need for combining both spatial and economic strategies in environmental regulation to achieve comprehensive and equitable health benefits.

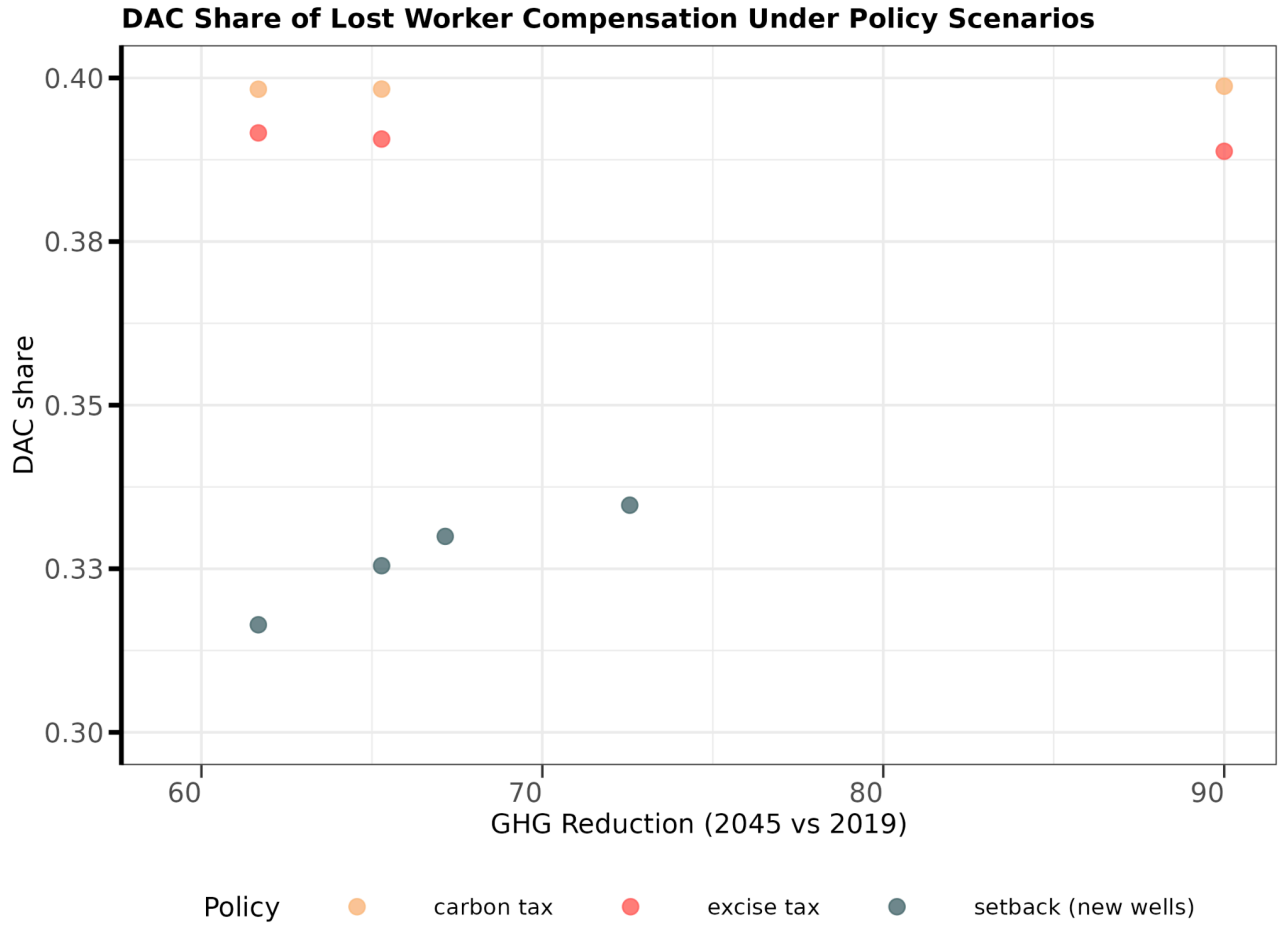

Figure 6: **Disadvantaged community share of lost worker compensation under different policy scenarios**. This plot shows the share of lost worker compensation experienced by disadvantaged communities (DACs) under different policy scenarios aimed at achieving 2045 greenhouse gas emission reduction targets. The y-axis represents the share of lost worker compensation for DACs, and the x-axis indicates the stringency of the 2045 GHG emissions targets. DACs experience a consistently lower share of lost worker compensation under setback policies compared to excise and carbon taxes. Notably, the third gray point from the left represents the new 3,200 foot setback distance, where the share of lost worker compensation for DACs is lower than under other policies. DACs benefit from a lower share of economic impacts under setback policies.

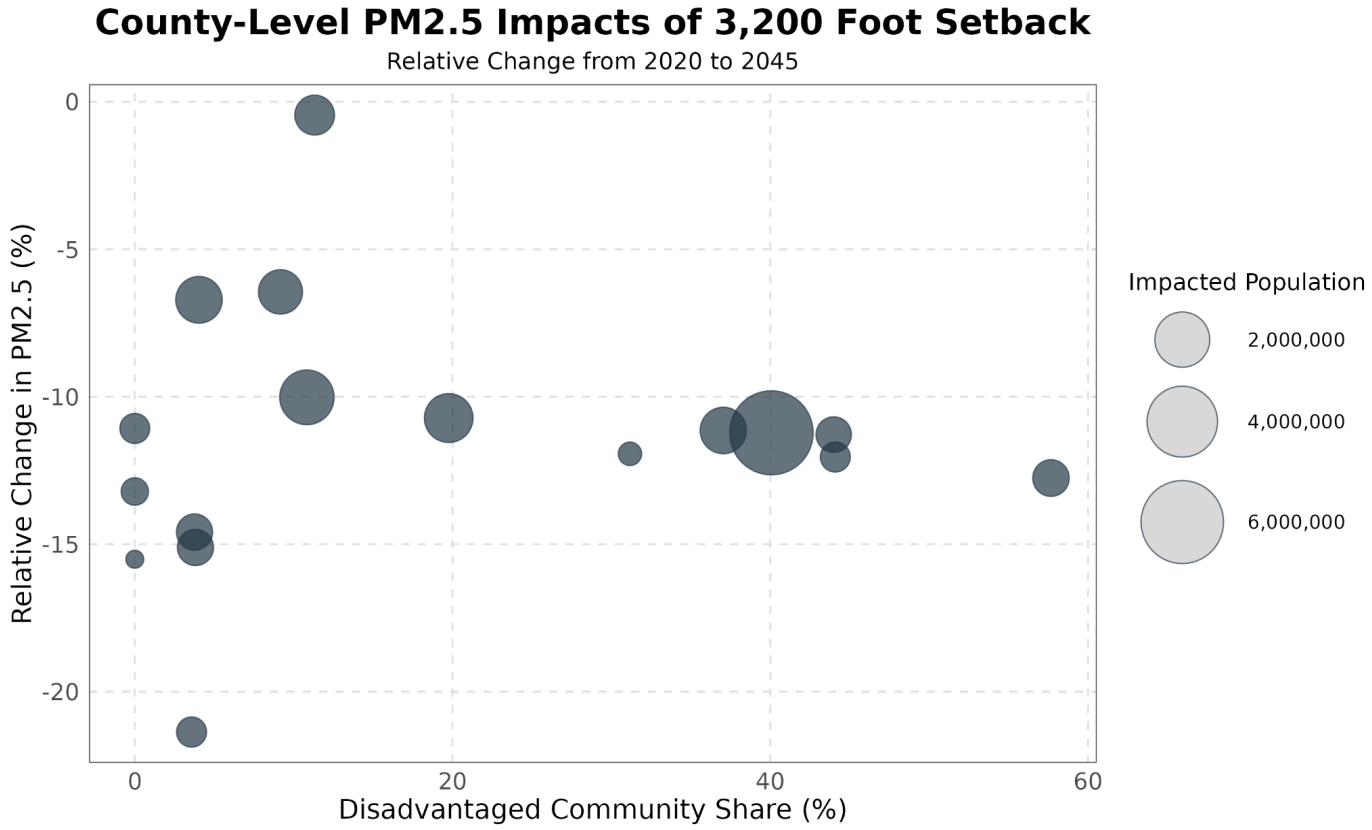

Figure 7: **Understanding the relative change in PM2.5 exposure across oil-producing counties in California from 2020 to 2045 under a 3,200 foot setback policy**. The scatter plot shows that counties with a higher proportion of disadvantaged communities tend to experience greater reductions in PM2.5 exposure, suggesting that the setback policy has the potential to provide significant air quality benefits to vulnerable populations. The bubbles represent the impacted population in each county, or the cumulative population of all census tracts in the county with oil production. The visualization shows the additional improvements in PM2.5 reduction under the 3,200 foot setback compared to a no setback scenario. For example, Los Angeles County, the largest point on the plot, would see an 11% greater reduction in PM2.5 exposure in 2045 under the 3,200 foot setback compared to no setback. Roughly 40% of Los Angeles county is classified as disadvantaged, and these communities would see significant reduction in PM2.5 exposure with the implementation of a 3,200 foot setback starting in 2020.

The visualizations in Figures 5, 6, and 7 underscore the potential for setback policies to address environmental justice concerns by disproportionately benefitting disadvantaged communities. Figure 5 demonstrates that setback policies consistently result in a higher share of avoided mortality benefits for DACs compared to carbon and excise taxes with equivalent GHG reduction outcomes. This suggests that the targeted spatial restrictions imposed by setbacks

are more effective at reducing health risks in vulnerable communities living near active drilling sites. Further, Figure 6 shows that DACs also experience a lower share of lost worker compensation under setback policies compared to tax-based approaches. This implies that the economic impacts of setbacks are less concentrated in disadvantaged communities, potentially mitigating concerns about job losses in already vulnerable communities. Figure 7 further reinforces the environmental justice benefits of setbacks by showing that counties with a higher proportion of DACs tend to experience greater reductions in PM2.5 exposure under the 3,200 foot setback scenario.

In summary, these results highlight the importance of considering the distributional impacts of environmental policies and the potential for setbacks to mitigate the disproportionate health burdens faced by disadvantaged populations, while also reducing economic disruption in these communities. By prioritizing the well-being of communities most affected by oil production, setback policies such as SB 1137 can contribute to a more equitable transition towards sustainable energy practices

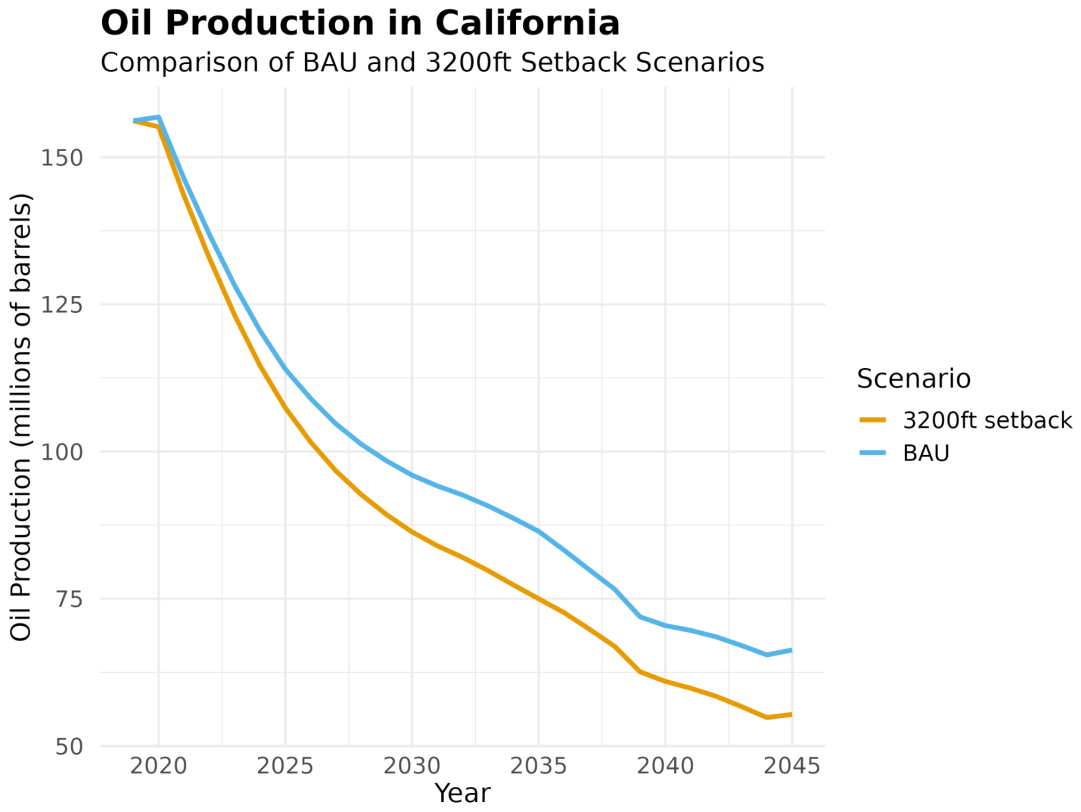

Figure 8: **Visualizing the reduction in oil production under a 3,200 foot setback policy.** The figure illustrates the projected impact of a 3,200 foot setback policy on California's oil production compared to a business-as-usual (BAU) scenario. The setback line demonstrates the policy's increasing effectiveness over time due to its cumulative effect on reducing the number of new wells drilled each year. As fewer new wells are added annually under the setback scenario, the overall oil production decreases more significantly with each passing year compared to the BAU scenario. This compounding effect highlights the long-term benefits of implementing a setback policy, as the reduction in new well drilling will lead to a decrease in oil production over the course of the forecasted period of 2020 to 2045. By visualizing these trajectories, policymakers and stakeholders can better understand the potential of setback policies to reduce oil production.

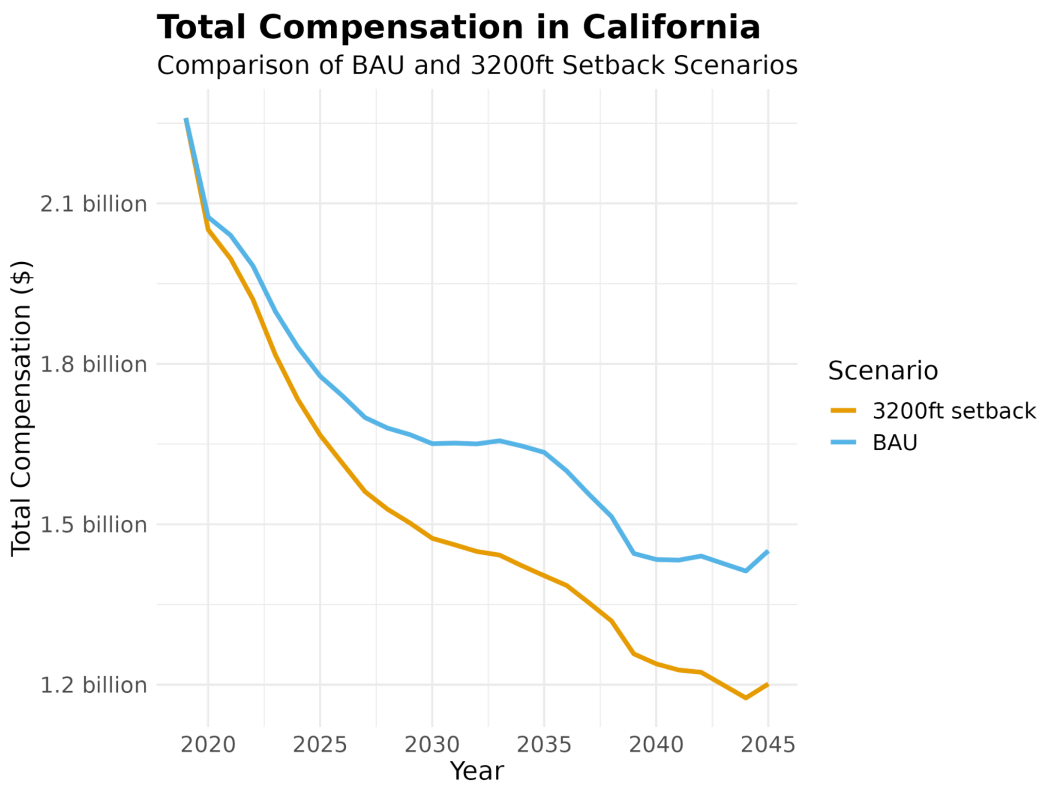

Figure 9: **Visualizing the economic impact of a 3,200 foot setback policy on total worker compensation**. This figure compares the projected total compensation under a 3,200 foot setback policy and a business-as-usual (BAU) scenario for the oil industry in California. The setback line shows the policy's growing influence on reducing total compensation over time, as the cumulative effect of fewer new wells drilled each year leads to a decrease in overall industry compensation compared to the BAU scenario. This visualization shows the long-term economic implications of implementing a setback policy on the oil industry workforce.

4.5: Machine Learning Implementation

The second phase of the project is to model well entry and exit projections using machine learning methods. In the original study, well entry and exit were estimated using Poisson models. Poisson models are a type of generalized linear model that are commonly used to model count data, where the response variable represents the number of occurrences of an event within a fixed interval of time or space. The Poisson distribution assumes that the mean and variance of the count data are equal, and the models relate the mean of the response variable to a linear combination of predictor variables through a logarithmic link function. Poisson models are well-suited for modeling rare events, such as the entry or exit of oil wells in a given time period, and can handle overdispersion in the data. However, Poisson models may be a weaker choice for modeling well entry and exit projections due to their simplifying assumptions. Poisson models assume that events occur independently of each other and at a constant rate over time, which may not hold true for oil well entries and exits. Additionally, Poisson models do not account for the potential presence of zero-inflation, where there are more zero counts than expected under the Poisson distribution. These limitations may lead to biased or inaccurate predictions of well entry and exit. The purpose of the updated models is to provide more detailed and powerful predictions of the total number of new wells entering and

exiting oil fields in the future. Brent price, capital expenditures, operational expenditures, and depletion rate by field are the features used in the Poisson models to estimate the number of total wells based on the change in entry or exit year over year. Existing literature in energy economics suggests that oil producers have several options at their disposal, including the option to delay investment and the option to abandon a producing field. The decision for producers to extract oil is ultimately decided by the profitability of the opportunity, which is influenced by factors such as oil price, capital expenditures, and operational expenditures (Abadie et al. 2017).

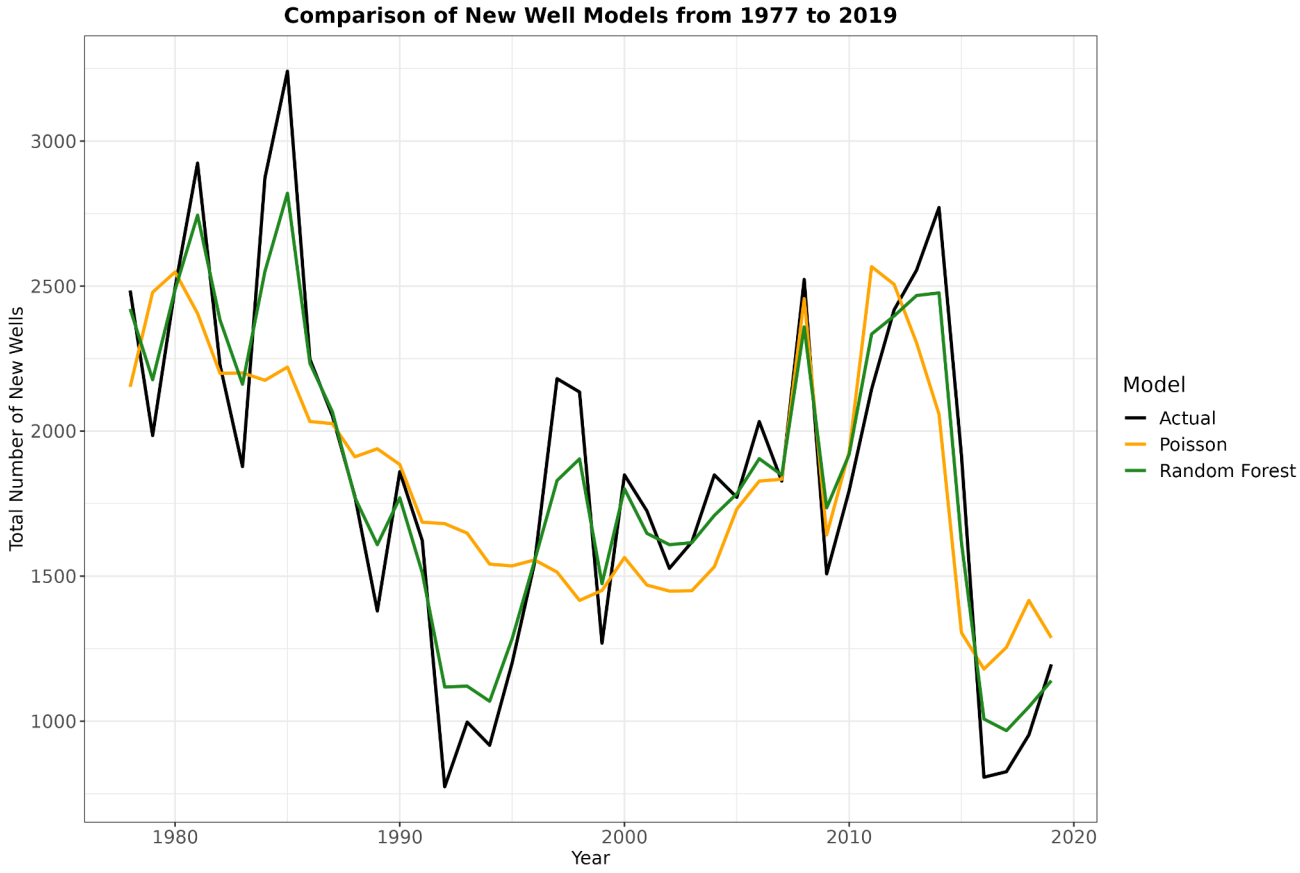

Figure 10: **Comparing the predictive power of each of the models versus the actual number of well entries**. This figure shows the predictions of total new wells by Random Forest and Poisson models compared to the actual historical data from 1977 to 2019. The Random Forest model, which was trained with 500 trees and 4 randomly selected features at each split (mtry = 4), captures larger shifts and trends than the Poisson model, which is more conservative and does not fluctuate as much as the Random Forest. The ensemble nature of the Random Forest model allows it to capture complex nonlinear relationships between the predictor variables and the response, making it a powerful tool for modeling well entries. The Poisson model provides more conservative estimates, reflecting its tendency to predict fewer new wells overall, and often underestimated the peaks seen in the true data. This comparison highlights the strengths and limitations of each modeling approach in capturing the dynamics of well entries over time.

Random Forest regression is a powerful machine learning technique that offers several advantages over traditional statistical models like Poisson regression in modeling well entry and exit projections. Tree-based models like Random Forest can capture complex, nonlinear relationships between input features and the target variable, allowing them to uncover

interactions in the data that might have been missed by the Poisson model. This is evident in Figure 10, where the Random Forest predictions capture larger shifts in the number of new wells from 1991 to 1995. Historical estimations of the number of new wells have proven to be more accurate than the original Poisson model when testing on historical well entry based on the Brent price, weighted means of capex and opex, depletion rate, and field code. Well exit models are trained on the same features, except capital expenditures is removed from the feature list as the amounts are not forecasted to change much, if at all, from 2020 to 2045. The doc_field_code, or the oil field code, is added to the Random Forest model to improve its predictive power. Note that the Brent price has been manually inserted for the training of these models, as the data given was in real dollar values instead of adjusted present value. This change to adjusting the historical Brent prices into present dollar terms allows for the model to not be negatively impacted by misunderstanding the impact of oil price on well entry. The architecture of the Random Forest model used in this study is intentionally kept simple to avoid overfitting and maintain flexibility. The purpose is not to create a model that is too sensitive to the training data, as the forecasted data are only estimates. A more complex model might overfit to the noise in the training data, leading to poor generalization performance on unseen data. By using a simpler model with 500 trees and 4 randomly selected features at each split, the Random Forest model strikes a balance between capturing important patterns in the data and avoiding overfitting. This approach results in a more flexible model that can adapt to new data without being overly influenced by the idiosyncrasies of the training set.

It's important to acknowledge the potential limitations of the Random Forest used for predicting well entry and exit projections. One significant concern is the reliability of the forecasted feature data, particularly the oil price and operational expenditures, from 2020 to 2045. These variables are highly susceptible to fluctuations caused by various external factors, such as geopolitical tensions, global macroeconomic conditions, and unforeseen events like natural disasters or pandemics. These kinds of unexpected changes in the input features can impact the models' performance and lead to inaccurate projections. Another potential limitation arises from the discrepancy between the training data and the out-of-bag data. The models are trained on historical data from 1977 to 2019, while the projections are made for the period from 2020 to 2045. An underlying assumption in the Random Forest model is that the relationships between the input features and the target variable remain relatively consistent over time. However, this assumption may not hold true, as the dynamics of the oil industry can evolve in unexpected or unforeseen ways. To enhance the robustness and reliability of the well entry and exit projection models, it is important to update the models as new data becomes available. Regular model validation and recalibration can help identify any deviations from the expected patterns and allow for timely adjustments.

In the workflow, the Random Forest models are injected into the *fun_extraction_model_targets.R* script to update the entry and exit predictions, particularly in the *func_yearly_production* function. The code has been updated so that the user now selects which model to use (either Random Forest or Poisson) in the *00_extraction_steps.R* script, allowing for flexibility in model choice based on the user's preferences or the specific requirements of the analysis. The predictions from the selected model for each year are then joined into the model workflow, and the amount of producing wells in each field is multiplied by the average production per well in the forecasted years to calculate the total production in future years. This approach is consistent with the existing framework, which forecasts field production by the average production per well multiplied by the number of active wells in the field.

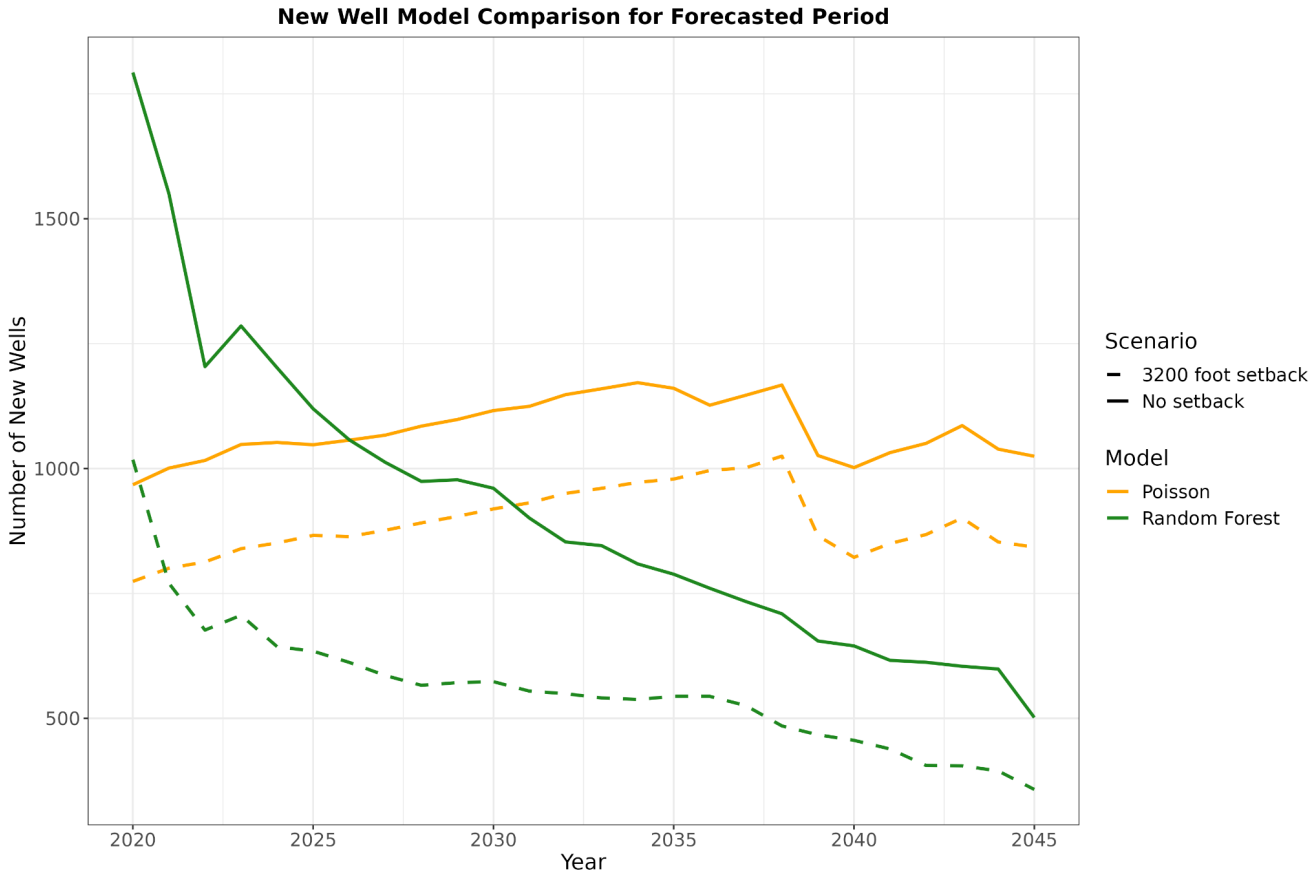

Figure 11: **New well model comparison for forecasted period.** This plot compares the predictive power of the Poisson and Random Forest models under two policy scenarios: no setback and a 3,200 foot setback on new wells. The Random Forest model predicts fewer new wells after 2026 compared to the Poisson model, suggesting that it may be more sensitive to changes in the input features over time. The impact of the 3,200 foot setback is slightly larger on the Random Forest model, as evidenced by the wider gap between the two scenarios compared to the Poisson model. The overall trends in both models show a decrease in new wells over time, with the setback scenario resulting in a lower number of new wells compared to the no setback scenario.

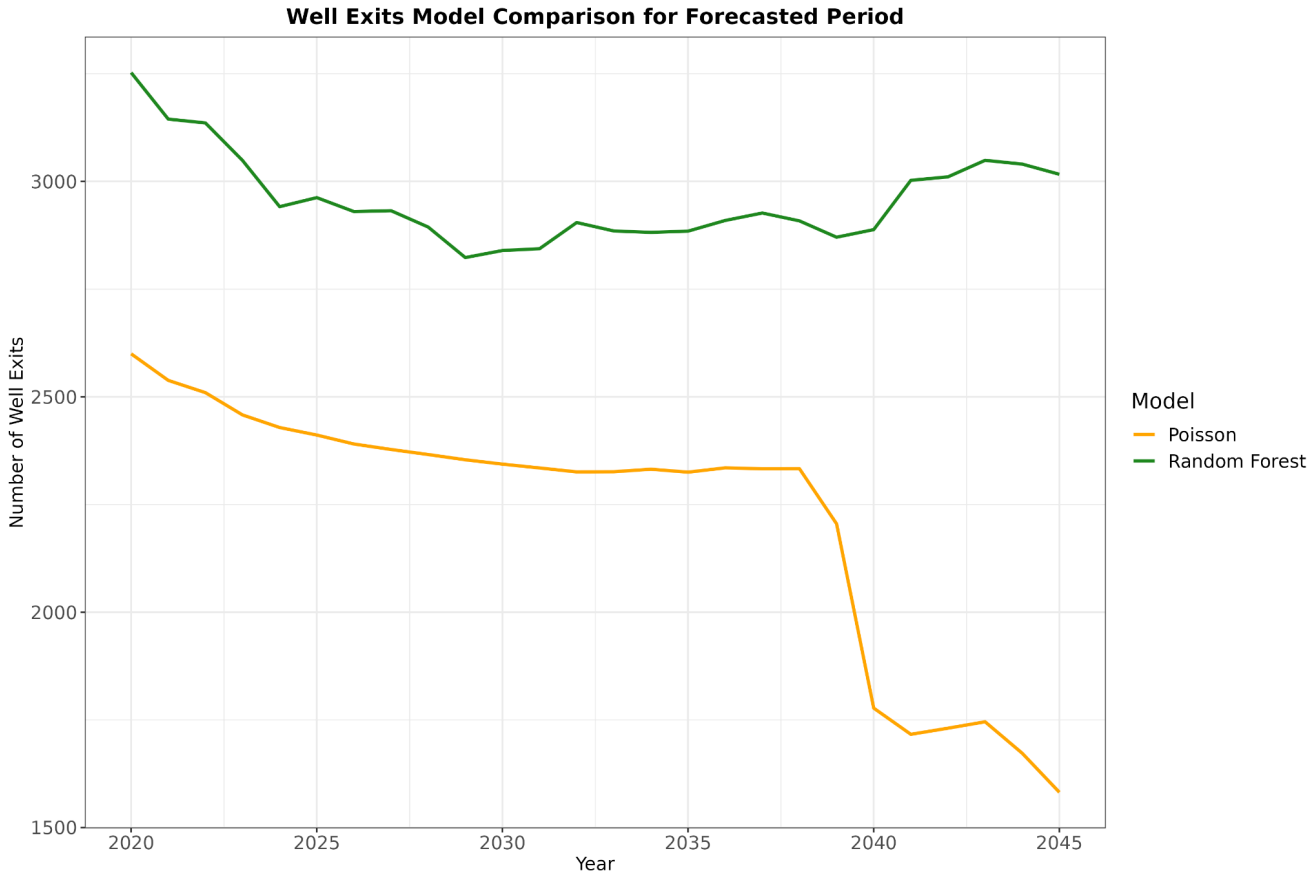

Figure 12: **Exit well model comparison for forecasted period.** This plot compares the predictive power of the Poisson and Random Forest models under two policy scenarios: no setback and a 3,200 foot setback on new wells. The Random Forest model predicts more exit wells compared to the Poisson model, suggesting that it may be more sensitive to changes in the input features over time.

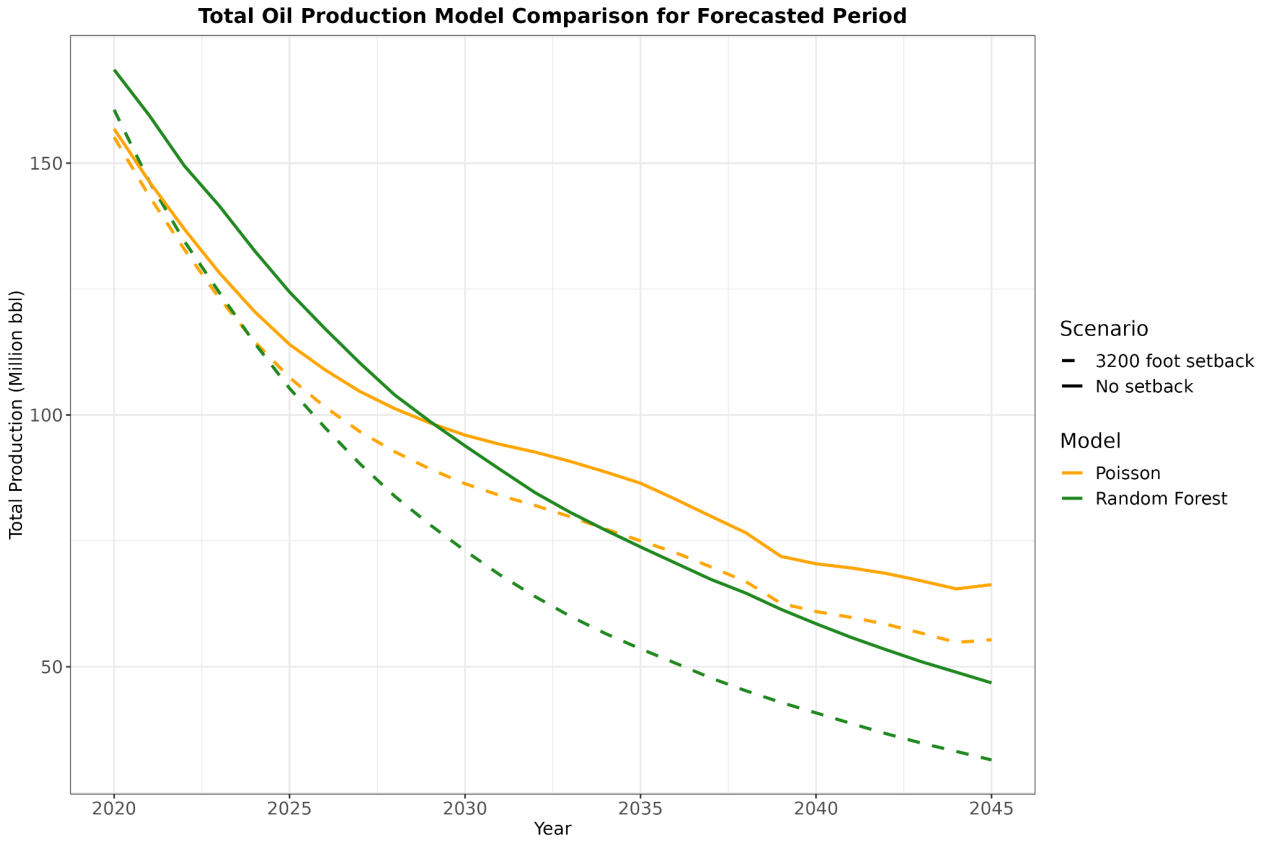

Figure 13: **Total production model comparison for forecasted period**. This plot compares the predictive power of the Poisson and Random Forest models under two policy scenarios: no setback and a 3,200 foot setback on new wells. The production for the Random Forest model is slightly lower after about 2028 because it predicts fewer new wells starting around this time, and also more well exits over the entire forecasted period. This suggests that the Random Forest model may be more sensitive to changes in the input features over time, capturing the combined effect of reduced new well development and increased well abandonment on total production. The Poisson model, on the other hand, shows a more gradual decline in production, likely due to its more conservative estimates of new well entries and exits. The impact of the 3,200 foot setback is more pronounced in the Random Forest model, as the gap between the two scenarios is larger compared to the Poisson model.

The comparison of the Poisson and Random Forest models for forecasting new well entries, well exits, and total production provides valuable insights into the potential impacts of the 3,200 foot setback policy on the oil industry. The Random Forest model predicts a larger impact of the setback policy, suggesting that it may be more sensitive to changes in the input features. In terms of new well entries, the Random Forest model predicts fewer new wells after 2026 compared to the Poisson model, indicating a more significant decline in well development. Similarly, the Random Forest model forecasts more well exits throughout the entire period, which, combined with the reduced new well entries, leads to a slightly lower total production after 2028. In contrast, the Poisson model provides more conservative estimates and shows a more gradual decline in new well entries, well exits, and total production. These differences highlight the importance of considering multiple modeling approaches to assess the potential impacts of the setback policy, as each model may capture different aspects of the complex dynamics at play in the oil industry.

The plots created in the other sections of this document rely on the Poisson models for well entry and exit. Further work on these models can focus on improving the understanding of

well entry and exit by integrating new data and incorporating government policies on well licenses into future models. Bayesian modeling presents a promising approach to achieve this goal. Unlike traditional machine learning methods, Bayesian models allow for the incorporation of prior knowledge and historical data into the modeling process. By using this information, Bayesian models can update thor predictions based on observed data, which results in more accurate estimates. In the context of well entry and exit, historical data on well licenses, production levels, and regulatory changes can be used to inform the priors of the Bayesian models. As new data becomes available, the models can adapt and refine their predictions, providing stakeholders with reliable insights for decision making. Bayesian models also offer the advantage of quantifying uncertainty in the predictions. By generating probability distributions over the model parameters and outputs, Bayesian models can capture the inherent variability and uncertainty associated with well entry and exit dynamics.

4.6: Delaying Setback to 2025

Since the data on oil production goes up to 2019, implementing the setback does not directly relate to understanding the implications of SB 1137 as the effects of the bill will be realized, if it is passed, starting January 1, 2025. A code chunk is added to *load_input_info_fc.R* to introduce setbacks starting for a specified intervention year, which is set to 2025 in this case, instead of applying the setbacks from the beginning of the simulation period (2020). First, the update generates a dataframe with all combinations of setback scenarios, field codes, and years from 2020 to 2045. The data is then separated into pre-intervention and post-intervention periods. For the pre-intervention period, the “no_setback” scenario is applied, while for the post-intervention period, the original setback scenarios are used. Finally, the pre-intervention and post-intervention data are combined to create an updated setback dataframe that reflects the introduction of setbacks starting from the specified intervention year. This allows for a more realistic representation of the Senate Bill and its impact on the projections of oil production. More information on this can be found in Section 7.3.

4.7: Interactive Dashboard

The final phase of the project involves leveraging the results of the first two phases to create a public-facing dashboard. This dashboard contains information including the locations of wells, buffer areas around sensitive receptors like schools, hospitals, and disadvantaged communities impacted by the significance of Senate Bill 1137. Additional contents of the dashboard will include a brief background on the effects of oil extraction on health, the difference between active and inactive wells, and the purpose behind the dashboard. There are four pages in the interactive dashboard named: Oil Well Explorer, About this App, Statewide Impacts of SB 1137, and Research Methods. Our client requested that information about active and non-active wells be the first page displaying hard hitting information. Although the *Wells_All.shp* contained five different type of well statuses, for the purpose of simplicity the wells classified as "Plugged","Idle","Canceled","Unknown","PluggedOnly", and "Abeyance" were grouped to be classified as Non-Active.

The generated a reactive Leaflet map that takes in two inputs: county and well type. After the user selects both inputs, the map on the right will update, displaying desired results. Well locations are aggregated when zoomed out, but users have the option to click on the cluster in order to view the disaggregated distribution. On this map, the dashboard uses Leaflet’s World Street Map so that users are able to easily identify locations wells can be in in order to make internal connections as to where these oil wells may reside in relation to the user. This map also includes a layer in cornflower blue displaying a generated 3,200 foot buffer

around sensitive areas so that users are able to assess the dangerous proximity that oil wells are to them, and so that users can conclude that if SB 1137 were to be implemented, no further oil wells could be drilled anywhere in the cornflower blue area. At the top, the Oil Well Explorer page hosts a brief description of SB 1137 so that users who may not be familiar with SB 1137 are able to attain an understanding of its implications.The results of the additional 3,200 foot setback scenario are incorporated in a pop-up message. When the user clicks on a desired county, a pop-up message 'County Facts' appears. It contains the name of the county the user clicked on, percent reduction in PM 2.5 relative to a business as usual reduction, percentage of disadvantaged census tracts, as well as population.The *county_health_results.csv* dataset is used to calculate the percentage of disadvantaged census tracts in each county, using a column called `dac_share`. The average of both percentage of disadvantaged census tracts and population were calculated using values from 2019 to 2023. To calculate the percentage of PM 2.5 reduction with SB 1137 implemented, the added column to our input data sets implemented the following formula :
$percreduction = SB1137_{pm2.5} - BAU_{pm2.5} / BAU_{pm2.5}$. The reason behind using the relative difference was to ensure the user could better interpret the results.

Content on the About This App page will contain further information about motivation behind the project and analysis, background information about environmental racism and a brief history on oil wells in California, and relevant information about SB 1137 so that users can gain a full understanding of the Bill and its importance, along with being able to visualize its importance.

Outputs from the updated setback model, including data and visualizations on health, emission, and labor implications of the Senate Bill, serve to inform people on the costs and benefits of the implementation of the bill. Senate Bill 1137 proposes a setback that is expected to provide health benefits, especially for disadvantaged census tracts. These disadvantaged areas currently experience particulate matter (PM) 2.5 levels that are two times higher compared to non-disadvantaged tracts. Moreover, it is critical that the findings associated with the impacts and implementation of SB 1137 are accessible to voting Californians amid voting season in November. The findings associated with the additional setback scenario will be found on the Statewide Impacts of SB 1137 page. Here, visualizations comparing projections of avoided mortality costs, labor impacts, and emissions will be readily available for the public eye. Along with the visualizations will include a short summary describing the graphs, as well as what key takeaways the user can form.

A key goal of the dashboard is to emphasize how SB 1137 contributes to reducing emissions in these disadvantaged communities. Users who have an interest in understanding the societal implication of this bill will be able to interact with the dashboard to obtain quick, important information based on the user's location on how they are likely to be impacted by the effects of the bill. The dashboard has been and will continue to be designed in a way that allows for people of all backgrounds to easily digest the impact of the Senate Bill on their community and state at large. There will be information on the Bill and an interactive map where users can investigate well locations and visualize the results of the new setback policy. 'The Statewide Impacts of SB 1137' page contains short and concise information on the top, summarizing the main takeaways from the analysis performed by the freshCAir team. There exists three key visuals on this page, two include the disadvantaged communities' share of avoided mortality, as well as the disadvantaged communities' share of lost worker compensation associated with the implementation of SB 1137. The first visual demonstrates that the setback policy brings in better health benefits than the excise tax and carbon tax. The key takeaway with this visual is that the setback policy (represented with a gray dot) can protect disadvantaged communities from further environmental harm. The second point graph shows the share of lost worker compensation experienced by disadvantaged communities under different policy scenarios:

carbon tax, excise tax, and setback. The x-axis represents the stringency of the 2045 greenhouse gas emissions targets. This graph shows that, with the 3,200 foot setback in place (gray dot), the share of lost worker compensation is lower for disadvantaged than under the other two policies. This means that disadvantaged communities benefit from facing a lower share of economic impacts under the setback policy. Lastly, the last visualization is a line graph showing the projection of oil production in California with SB 1137 in place and with no setback policy in place. Over time, it is predicted that the policy's effectiveness will increase over time due to the fact that the effect of reducing the number of new wells being drilled each year will become more apparent with time. The gap between the two lines becomes notable at 2025, and continues to widen over time. To understand the methods and selection that came into play, the user can shift over to the last page to learn more about the freshCAir team's approach to predicting these values.

The final page of the interactive dashboard, Research Methods, contains a breakdown of the different approaches made in assessing the impact of SB 1137. This page is added per request of our faculty advisor in order to maintain all technicalities in one area if the user is interested in gaining a deeper understanding of the decisions and approaches made. The Research Methods page contains a breakdown of the impacts associated with the regulatory policy that prohibits new well locations 3,200 feet away from sensitive receptors. It also includes a description of the types of data that were used to conduct analyses, as well as the reasoning behind using different predictive models. The Machine Learning Development subsection breaks down important predictor variables used for the model, the type of machine learning model used, and the tuning metrics used to improve the random forest model. In the Model Training subsection, there is also a brief description of what the target variable is, as well as the metrics used to assess the performance of the model. Three visuals representing new well predictions, oil well production prediction, and a comparison of new well models from 1977 to 2019. The first graph is a historical graph showing the number of new wells ranging from 1977 to 2019, along with two other line graphs representing the different type of predictive model used. The green line represents the random forest model, and the orange line represents Poisson model the client crafted. The main takeaway with this visualizaiton is that the random forest model more closely resembles the actual number of new wells throughout this historical period. The second visualization that shows the number of forecasted new wells from 2020 to 2045. There are a total of four lines on this graph, with two representing the Poisson model and the other two representing random forest. There are two types of lines on this graph, the dashed line representing the 3,200 foot setback. Considering the number of new wells impacted by the setback implementation, the dashed line being below the full line is consistent. The last graph, representing forecasted oil production, also shows two different models with two different setback policies: 3,200 feet and no setback. The y-axis is the total amount of oil produciton in millions of barrels. The gap between each set of lines in the random forest model is larger compared to the Poisson model, and this graph suggests that the random forest model may be sensitive to changes in the different types of inputs over time.

The repository of the Github repository will be publicly available as the data itself was included in the .gitignore file. In terms of reproducibility, this repository will facilitate that, with the exception of the well locations. Moreover, the freshCAir team will also release the repository for the interactive dashboard which will mark the specific point in time in which we submitted the repository for the Master of Environmental Data Science Capstone.

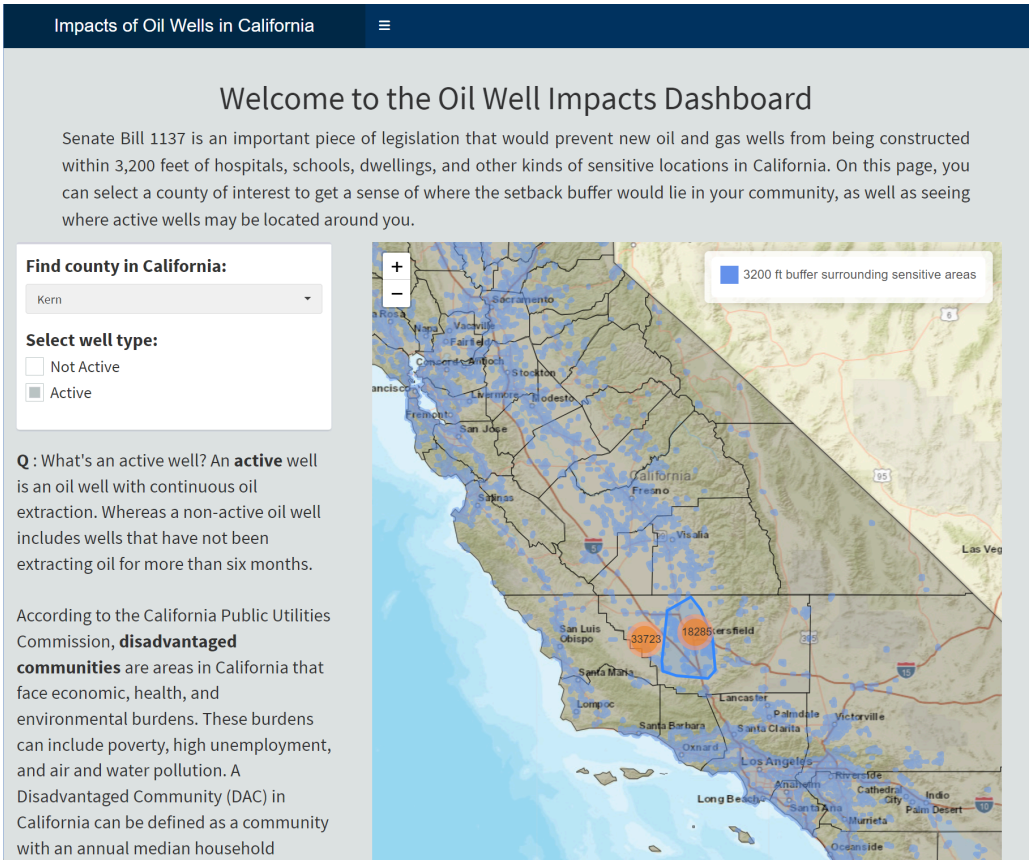

Figure 14: **Current layout of Shiny dashboard landing page selecting active wells in Kern County with clusterMarkers representing the aggregated number of active oil wells in Kern County.** This map also displays the distinction between what an active and not-active oil well is, as well as California Public Utilities Commission's definition of disadvantaged communities in California. The second cluster of active oil wells shows the area coverage that cluster has. The top right corner displays an explanation of the cornflower blue layer on the map so users can visualize the distribution of sensitive areas in California, on a statewide and county-wide basis.

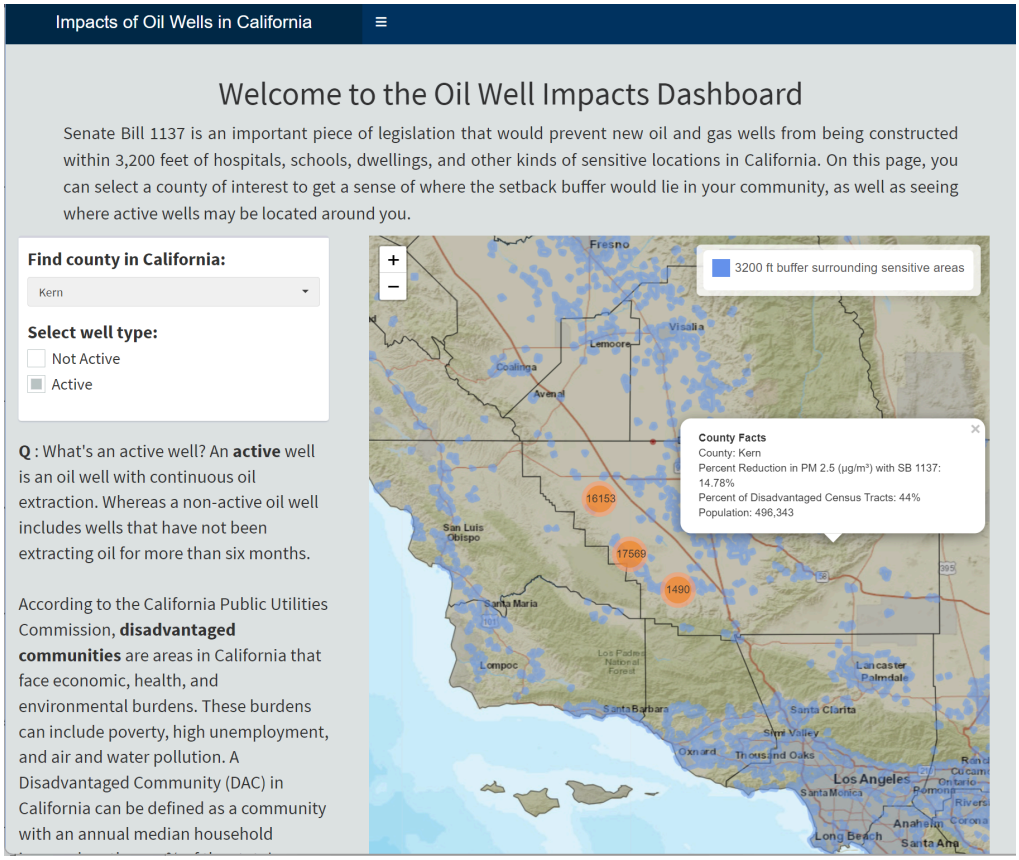

Figure 15: **Popup message displaying county-level and county-specific details including population and percentage of potential PM 2.5 reduction.** This feature does not require two pickerInputs to work, as the user can click on an area on the map and a pop-up message would show up on the user's screen. This feature works for all counties in California.

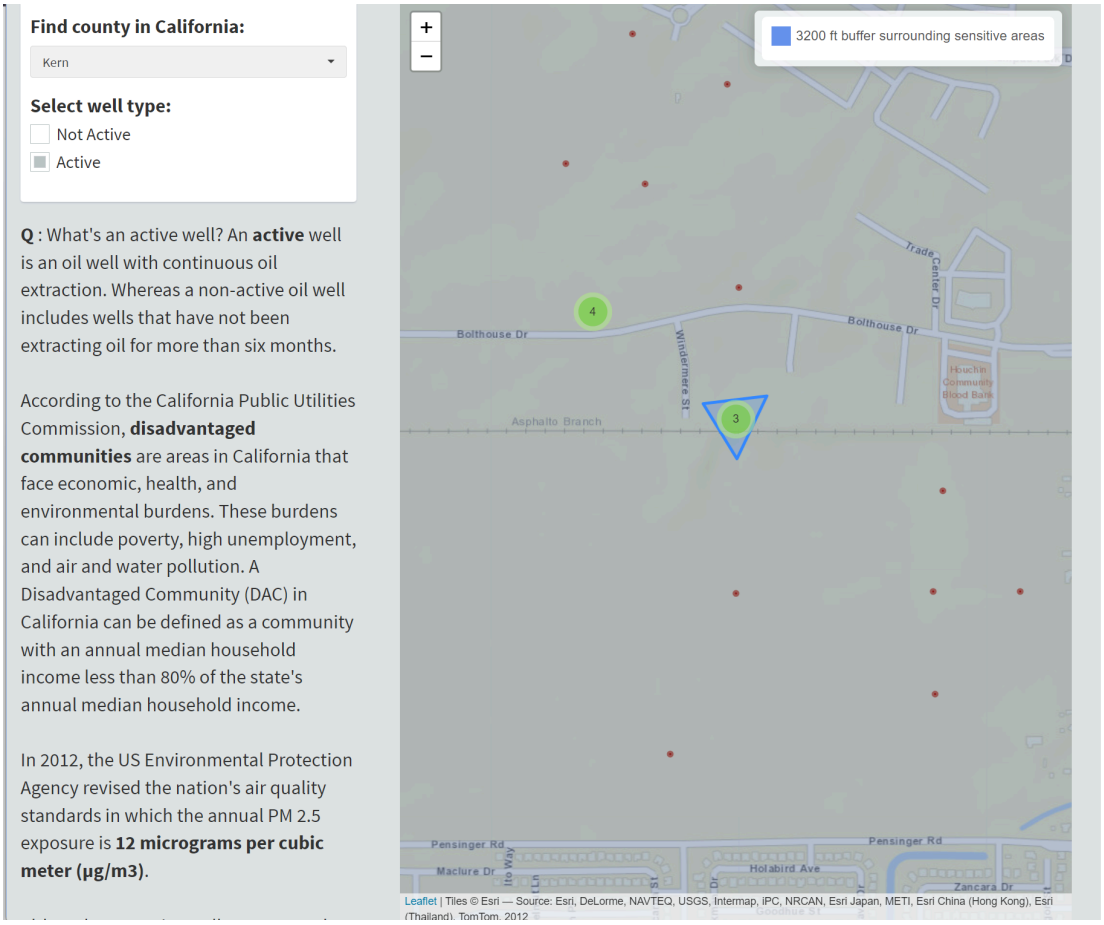

Figure 16: **Disaggregated cluster of active oil well locations in Kern County.** The highlighted cluster, represented by a triangle, shows the disaggregation of the cluserMarker on the Leaflet map. These clusters are eventually broken down into red points shown in the map above. On the left, additional information on PM 2.5 can be found so that users are able to better understand the metrics behind yearly PM 2.5 exposure.

5. Products and Deliverables

5.1: Interactive Dashboard

The interactive dashboard will be hosted on the client's website, in hopes that it will be publicly distributed for Californians to access. The two main goals that motivated the development of the dashboard were education and communication. With the Oil Well Explorer page, Californians will be able to personalize the information being displayed in hopes that they are now informed about the issue at hand. The statewide impacts page will also help guide the user through further implications of the Bill, with digestible information and visualizations at hand.

The repository for the interactive dashboard will be submitted to the client for future modification and edits.This is in effort to facilitate our clients' communication and outreach related to SB 1137. The Github organization for the interactive dashboard, freshcair-capstone, will host all materials used for building the dashboard. Instructions on upkeep and potential additions will be included in the repository. Contents of the dashboard will include background information on the effects of oil well activity as well as the purpose behind developing a public

facing dashboard. Moreover, each tab on the dashboard will include different findings pertaining to freshCAir's Capstone Project. Maintenance will be made possible as the freshCAir team will incorporate a reproducible workflow for other users to replicate the dashboard. The interactive dashboard is currently being developed, and per request of the client, the freshCAir team will also ensure that verbiage used in the Summary and/or About pages are representative of the Nature Energy article and Policy Brief. The landing page of the interactive dashboard will contain a well locator map, in which all maps, categorized as active and inactive, will be displayed on a map with county borders. This map will also include average yearly particulate matter concentration, percentage of disadvantaged census tracts, and population. Moreover, the landing page will have most of the visuals so that most relevant information to SB 1137 will be first seen.

5.2: Updated Repository

The forked repository README has been updated by adding more descriptions to the scripts for future users, including information about the updates made and metadata documentation. The new README provides a thorough overview of the project, including its purpose, objectives, methodology, and expected outcomes. It outlines the data sources utilized, the preprocessing steps undertaken, and the machine learning models employed. Additionally, it furnishes clear instructions for installation, setup, and execution of the codebase, facilitating reproducibility and collaboration. The link to the GitHub Organization housing the Shiny dashboard repository can be found here.

5.3: Data Structure

For this project, the capstone project directory is located on the Taylor server. By setting the working directory to their local directory, future developers can ensure that the scripts access the necessary data files relative to the project's root directory. This update simplifies running the data processing scripts for future users.

The data structure within the data-str folder has two main components: public and private. The public data, initially received from clients, follows the Zenodo archive's format. Private data, also received from clients, cannot be shared publicly due to confidentiality. Therefore, the private data structure is designed to protect sensitive information, organizing the folders based on the level of detail in the columns.

To accommodate both public and private data structures, the current scripts adjust file paths and update them to facilitate easy interpretation of the model for future users. Originally, most data paths referenced the Zenodo archive, likely serving as a centralized location for data storage and access in the original project. To improve the workflow and make the project more user-friendly, future users will set their working directory at the top of each script once the data folder has been uploaded to the archive. A visualization of the detailed data structure is included in the Appendix.

6. Testing

6.1: Overview

The main components of testing in this project entail making sure that the data generated in the scripts leading up to the final extraction model are correct. There are 21 data sets used in the final extraction model (00_extraction_steps.R), with most of the data being

loaded in from the load_input_info.R script. To ensure accurate results, the inputs into the model are compared to the intermediate subfolder from Zenodo since all of this data is publicly available. Data inputted into these scripts has been tested and outputs confirmed through inspection. The *comparedf* function from the *arsenal* package is implemented to compare the data generated in the new workflow with what is in Zenodo from the original run-through.

6.2: Setback Coverages

To ensure validity of results, the regenerated output data is compared to the data inputted into the final extraction model (intermediate data) from the original study. The business as usual scenarios are compared to determine if the outputs are identical. The production forecast results match exactly with those generated by the clients, given that the *comparedf()* function from the *arsenal* package displays that there are no differences across each of the datasets after thorough inspection.

The summary table below displays the square mileage of oil fields that are covered under each setback scenario.

	no_setback	setback_1000	setback_2500	setback_3200	setback_5280
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	0	276.	593.	708.	976

The setback buffer coverage shows that the amount of oil fields covered by the 3,200 foot setback is about 20% larger than the 2,500 foot setback, which seems reasonable considering that many of the sensitive receptor buffer areas overlap. As a result, not the entirety of the setback region is included in the total area covered for most of the receptors, since these receptors tend to be joined close together especially in urban areas such as Los Angeles.

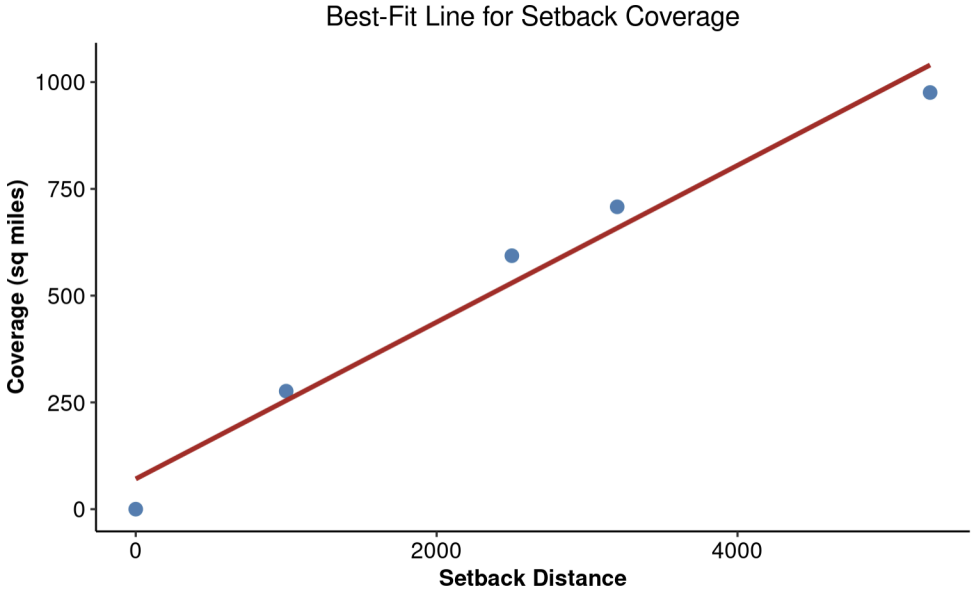

Figure 17: **Relative total square mileage of oil fields covered under each setback scenario**. This plot shows the relationship between setback distance and the amount of oil field covered in square miles. The plot shows a fairly strong positive linear relationship between setback distance and average coverage of oil fields. The linear trend implies that the relationship between setback distance and oil field coverage remains consistent, facilitating informed decision making and policy analysis.

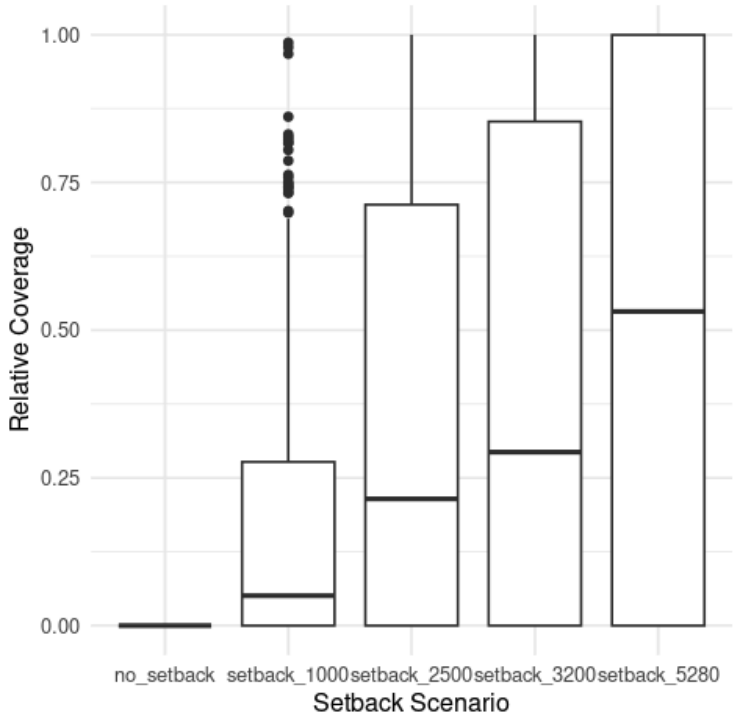

Figure 18: **Relative coverage of each oil well field under different setback scenarios**. This plot presents a comparison of the relative coverage of oil fields in California under different setback scenarios, with the y-axis representing the relative coverage and the x-axis displaying the different setback distances. The plot shows that as the setback distance increases, the relative coverage of oil fields also increases. The key takeaway from the plot is that the mean coverage for the 3,200 foot setback distance is slightly greater than the 2,500 foot setback and less than 5,280 foot setback, showing that the results for the new setback distance of 3,200 are in the appropriate range.

Note that the recreated figures also serve as tests of validity of the new setback scenario results.

6.3: Machine Learning Model Testing

To test the random forest predictions for the number of new wells by year, data is split into training and testing sets. The root mean squared error (RMSE) is a commonly used metric for evaluating model performance, including random forests. RMSE measures the average magnitude of the errors between the predicted and actual values, with lower values indicating better model performance. It is calculated by taking the square root of the mean of the squared differences between the predicted and actual values. RMSE is particularly useful because it penalizes larger errors more heavily than smaller ones, making it sensitive to outliers and providing a clear indication of the model's predictive accuracy. The RMSE for the Random Forest model is 13.956 and Poisson is 32.442. The improved accuracy of the predictions from the Random Forest model on the historical data can also be observed visually in Figure 10.

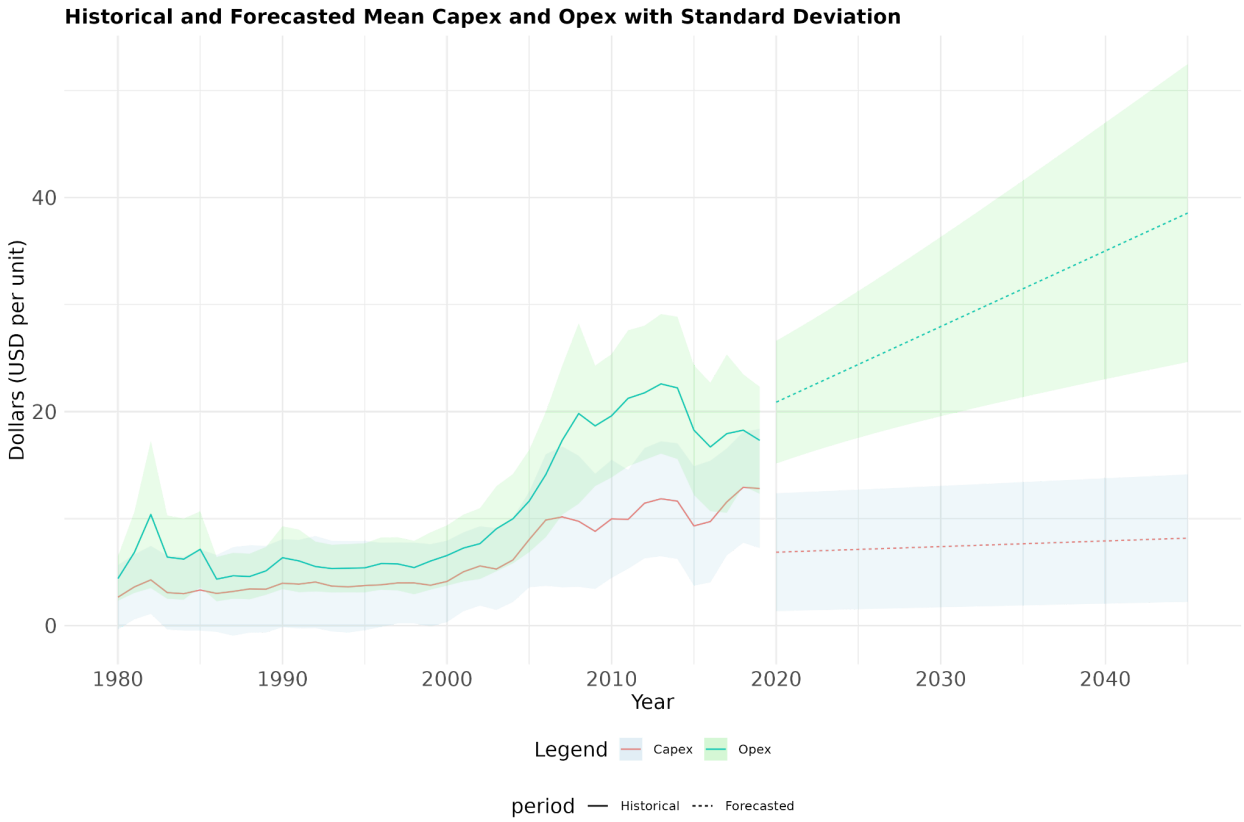

Figure 19: **Historical and forecasted mean capital expenditure and operational expenditure with standard deviation buffer**. This plot shows the historical and forecasted capital expenditures (Capex) and operational expenditures (Opex) per unit from 1980 to 2045. Historical data is represented by solid lines, while forecasted data is indicated by dashed lines, with shaded areas representing the standard deviation. Notably, Opex peaked around the 2008-2015 period, reflecting significant operational costs during these years. This plot provides insight into the random forest model's projections for higher numbers of new wells. The model may be projecting these higher numbers due to the historical correlation of high new well counts with periods of elevated costs, potentially influenced by multicollinearity among the variables. This suggests that the model may be capturing the complex relationships between expenditures and well counts in its forecasts. The random forest model is likely to disregard data prior to 2008 since the operational expenditures in the forecasted period have no overlap with the period from 1977-2008.

While the RMSE values indicate that the Random Forest and Gradient Boosted models perform better than the Poisson model on the historical data, it is important to consider the underlying factors influencing these estimates. The Random Forest model's strong accuracy may be attributed to its ability to capture nonlinear relationships between the features of the model. However, this complexity also introduces potential issues like multicollinearity, where the interdependence of variables may lead to overfitting. The forecasted feature data, or out-of-bag data, differs from the historical period due to higher operational expenditure, as seen in the figure above, and depletion rates, even with oil prices adjusted to present dollar values. This discrepancy highlights the need for further work to enhance the robustness and reliability of these machine learning models.

6.4: Interactive Web Dashboard Testing

Since the dashboard does not have user inputs, rather selections, there is no testing that needs to be conducted for the dashboard.

7. User Documentation

7.1 Overview of Repository Structure and Organization

The README file in the forked project repository contains important information on what each script does. The README, like it is set up now, will list the scripts in order of how they should be run to recreate the results from start to finish. Since all of the data leading up to the final model has been generated and stored in the *data-str* folder, individual scripts may be run in any desired order.

7.2 Computing Environment

To ensure reproducible practices and facilitate future interactions with the updated workflow, the computing environment has been documented, recording all packages and their respective versions. The session information has been saved in a *sessioninfo.txt* file, which details the exact R version and package configurations used when the project was completed. To activate the environment used for this project, first verify that you have the same R version installed as noted in *sessioninfo.txt*. Next, use the package details in *sessioninfo.txt* to install the required packages and their specific versions. This can be done manually or with the help of tools like *remotes* for precise version control. By following these steps, the computing environment used for the project can be accurately restored, ensuring that the workflow runs smoothly without issues caused by package updates. Code is included in the Appendix for how to run *sessioninfo.txt*.

To avoid permanently setting these package versions and to create a reproducible environment, users can set up a renv, or a virtual environment, in R Studio before setting the session info packages. Here's how to create and activate a renv for this project:

1. Install the renv package if not already installed:

```
install.packages("renv")
```

2. Load the renv package:

```
library(renv)
```

3. Initialize a new renv environment in your project directory:

```
renv::init()
```

4. Activate the renv environment:

```
renv::activate()
```

5. Install the packages specified in the sessioninfo.txt file:

```
renv::restore()
```

By following these steps, a new renv environment will be created specifically for this project, ensuring that the required packages and their versions are installed and activated within the virtual environment. This approach allows for reproducibility while keeping the project-specific dependencies separate from the global R environment. While this approach ensures reproducibility, it's important to note that packages might become outdated over time.

Future users may need to manually update or add packages to keep the environment current. The sessioninfo.txt file provides a reliable reference for the project's dependencies, making the environment portable and reproducible, thereby simplifying future updates and collaborations.

7.3 Important Code Updates

The most important update that users must make for the scripts to run is changing the working directory to the user's designated directory. The code for this project was completed on the Taylor server, and the working directory is set for each script using the *setwd()* function. Users must update the path inside this function to their new directory location. This update must be added for all scripts, as the Taylor server required the directory to be set for each script to access data used in the project. The tables below highlight the important scripts which users may investigate to digest the injections that have been made into the code to generate the 3,200 foot setback scenario.

There is an important update in the *load_input_info.R* script that users must be aware of. The user will select the well entry and exit model for the final extraction model, with the options of Poisson or Random Forest . The user may enter 1 or 2 depending on which model they would like to use for that run. Another key update is imposing a delay on when well setbacks begin. Users can set an *intervention_year* for which the setbacks will begin, and the code will pick up if a year has been set and institute the setback starting that year. The business-as-usual scenarios will be run for years leading up to the intervention year, if one is set. The user simply has to define the variable in the script before the if statement which starts on line 233 of *load_input_info.R*.

7.4 Future Work

The "extraction_2024-06-05_rf" subfolder contains the model run for the Random Forest approach, while the "extraction_2024-05-13" subfolder contains the Poisson run. These subfolders serve as a starting point for comparing the performance of the two modeling approaches and identifying areas for improvement. Since the figures in this project were created using the Poisson results, implementing the Random Forest results could potentially improve the estimates of future oil production. Note that the 5,280 foot setback GHG reduction equivalent for excise and carbon taxes was not generated in the project due to unforeseen issues with the scenarios generated and stored in the outputs of the final extraction model. Since this project did not entail calculating the 3,200 foot setback equivalent GHG reduction of excise and carbon taxes, this issue should be investigated when calculating the equivalent GHG reduction for the new setback distance.

As new data becomes available it can be integrated into the models to improve their predictive power. By incorporating this data, the models can adapt to the evolving industry landscape and provide more accurate estimates of well entry and exit. Future work can also focus on incorporating government policies on well licenses into the models. By considering the impact of regulatory changes and policy interventions, the models can provide more comprehensive insights into the factors influencing well entry and exit decisions.

Bayesian modeling presents a promising approach to leverage new data and incorporate prior knowledge into the modeling process. The historical data from 1977 to 2019, used to train the Random Forest models implemented in this project, can serve as the prior knowledge in the Bayesian model. As new data is collected, the Bayesian model can update its posterior distributions, resulting in more refined predictions. In practice, this can be implemented by specifying prior distributions for the model parameters based on the historical data and the current Random Forest model's performance. To implement Bayesian modeling in R, several packages can be utilized, such as *rstan* for Bayesian inference, *brms* for a high-level interface for Bayesian regression models, and *loo* for model comparison and selection. By utilizing the

strengths of Bayesian modeling, integrating new data oil production, and incorporating government policies, the well entry and exit models can be continuously updated and improved to give more accurate and actionable models and results.

7.3 Important Workflow Scripts

Figure 5 in the Appendix shows a diagram of the key scripts that are impacted by injecting the 3,200 foot setback scenario. The table below includes a description of these scripts that precede the final extraction model.

Table 1. Key Setback Scripts

Script	Description of Scripts	Importnt Outputs
well_setback_sp_prep.R	Processes spatial data from the FracTracker Setback dataset to analyze and visualize sensitive receptors (e.g., dwellings, playgrounds, healthcare facilities) around oil and gas extraction sites in California. It involves reading and transforming spatial layers from a Geographic Database (GDB), applying buffers to identify setback areas, simplifying complex geometries for efficiency, and ultimately creating and saving spatial buffers around sensitive sites, which are then visualized using various mapping libraries in R.	buffer_3200ft.shp: Create a new shapefile of 3200 foot buffer around sensitive receptors.
gen_well_setback_status.R	Processes well and field data to determine their proximity to sensitive receptors based on predefined setback distances (1000ft, 2500ft, 3200ft, and 5280ft) around oil and gas extraction sites in California. It involves reading spatial buffer data and then calculating which wells and fields fall within these buffers, ultimately generating attributes for each well and field regarding their inclusion within the setbacks, and visualizing these relationships through maps.	coverage_map.html: create a MapView image to display 3,200ft buffer around sensitive receptors
county-setback.R	Calculates and visualizes the percentage of each county in California covered by oil and gas setback zones of different distances (1,000ft, 2,500ft, 3,200ft, and 5,280ft) from oil and gas wells. It uses spatial data manipulation to intersect county and field boundaries with setback buffer zones, computes the area covered by each setback within counties, and saves the results for further analysis.	county_level_setback_coverage.csv: includes ratio of county area covered by 3,200 foot buffer
predict_existing_production.R	Predicts future oil production from existing wells that have not exited production up to the year 2045. It merges well production data with decline parameters and peak production information, adjusts for wells within setback areas, calculates production per well considering both active and non-setback wells, and finally aggregates and saves the adjusted production data for analysis, accounting for various scenarios including setbacks and plugged wells.	pred_prod_no_exit_2020-2045_field_start_year_revised.csv: provide the predictive number of wells in the 3,200 foot scenario.

7.4 Output Data and Figures Scripts

The following table below details important scripts used to generate output data and data used to create figures in the project. This table will be helpful for new future users of the project,

and especially for injecting the data generated using the Random Forest well entry and exit models into the figures.

Table 2. Figure Scripts

Script	Description of Scripts	Outputs
fig_outputs.R	Refines and extracts the scripts to generate outputs for creating figures for the manuscript. This includes various emission targets, social cost of carbon, carbon price scenarios, census tract data, oil extraction information, and disadvantaged communities.	dac_bau_health_labor_all_oil.csv, dac_health_labor_all_oil.csv, health_ct_results.csv: data relate to health labor_county_results.csv : data relate to labor state_levels_all_oil.csv, npv_x_metric_all_oil.csv: data relate to census and oil extraction
field_characteristics.R	Collects essential data to develop the plot and map, encompassing census tract information, disadvantaged communities, oil production, as well as setback scenarios at both field and county levels.	field_characteristics.csv , county_characteristics.csv
figure1.R	Processes and transforms various datasets related to oil production fields, census tracts, and county-level data. Generates a series of maps and plots to visually represent the distribution and impacts of oil production across different regions, focusing on aspects like disadvantaged communities, oil production volumes, PM2.5 pollution, and worker compensation, culminating in an assembled figure that integrates these visualizations for presentation.	fig1a.csv, fig1b.csv, fig1d.csv: necessary data to print out images. figure1a.png, figure1b.png: printed image files with new wells
figure2.R	Focuses on analyzing and visualizing data related to oil production and greenhouse gas (GHG) emissions under various policy interventions and scenarios in California. It first sets up the necessary R environment, loads data, and preprocesses it by filtering and adjusting based on specific criteria like policy interventions and oil price scenarios. Then, it creates a series of plots to visually represent the impact of different policies on oil production and GHG emissions over time, culminating in a combined figure that includes plots for reference, low, and high oil price scenarios along with their corresponding GHG emissions and cumulative effects, all formatted for clear and informative presentation.	fig2ab.csv, fig2c.csv: necessary data to print out images. figure2-high.png, figure2-low.png, figure2-ref-case.png: printed image files with new wells.
figure3.R	Based on the health, labor and climate impact data across the different setback distances, carbon tax and excise tax, this analysis filters out avoided mortality, total lost worker compensation and climate damaged value by the net present value (NPV). Six series of plots generated to compare the total value of different scenarios relative to BAU in order to achieve various 2045 GHG emissions targets(%2045 vs 2019), depending on the high, reference, and low oil price.	fig3a-f.csv: necessary data to print out images. Figure3-high.png, figure3-low.png, figure3-sb-all.png, figure3-sb-new.png: printed image files with new wells.

7.5 New Scripts

The table below details the new scripts that have been created in this project. While not essential for future users, they provide insight into some of the work done to generate plots and data from this document. The *testing.R* script is especially important, as the recreated data which was used in the final extraction model was validated in this script by comparing the recreated data with the data generated by the clients in the original project.

Table 3: *new-scripts* Folder Contents

Script	**Description**
eda.R	Performs exploratory analysis and visualization on oil production data. Processes and analyzes data on well activity, production volumes, and geographic distribution of wells across counties and census tracts. Examines the coverage of different setback scenarios and creates interactive maps to visualize the distribution of wells and their characteristics.
fr_viz.R	Used to create figures for the faculty review presentation.
ml-analysis.R	Trains random forest models to predict the number of new and exit wells based on oil price, capital expenditures, operational expenditures, and depletion rate. Generates visualizations to compare the performance of the random forest models with historical data and the Poisson model. Explores the historical and forecasted trends in capex, opex and oil prices
output-review.R	Generates plots and summary statistics to compare the effects of different setback scenarios. Wrangles census tract, county, and state-level data.
pred-dev.R	Code used in the development of the new and exit well predictive models. Note that the models are implemented in the *load_input_info_fc.R* script.
rel-coverage.R	Calculates the total area covered by each setback scenario, summarizes the relative coverage statistics, and creates plots to show the relationship between setback distance and coverage. Fits a linear model to the setback distance and coverage data, plotting the best-fit line and displaying the equation. These plots are used in the Testing section of this document.
testing.Rmd	This code performs data comparisons and checks across numerous datasets related to oil and gas production, emissions, policy scenarios, and environmental justice metrics. It uses the *comparedf* function from the *arsenal* package to verify consistency in dimensions, variable names, row counts, and attributes between different versions or sources of data frames to confirm the validity of the data being used for the final model. Data generated by the clients is compared to the new data to ensure consistency in the new outputs. The datasets being checked include crude oil production, greenhouse gas emissions from oil fields, carbon pricing and excise tax scenarios, emission reduction targets, county-level health incidence rates, industrial emissions, disadvantaged community shares, and projected impacts of policy interventions on production, emissions, and health outcomes.

7.6 Intermediate Data

Table 4 describes the data used in the final extraction model. This table can assist future users of the project by deepening their understanding of what kinds of information is used in the model.

Table 4: Intermediate Data Overview

Data	Column name	Purpose
scenario_id_list_targets.csv	scen_id: create scenario name by combining all policy scenarios BAU_scen: a binary whether the scenario is business-as-usual (no policies activated) setback_scenario: includes 1000ft, 2500ft, 3200 ft, 5280ft setback scenarios setback_existing: 1 if the setback is imposed on existing wells, 0 if not	Contains essential information on various energy scenarios for oil price, setback distance, carbon price, ccs technologies including various impact of these factors on emissions. Only the main columns are utilized.
setback_coverage_R.csv	NAME: county name, area_sq_mi, area_acre, orig_area_m2: overall area with different units setback_scenario: includes 1000ft, 2500ft, 3200 ft, 5280ft setback scenarios rel_coverage: the coverage of the oil field under the specific setback n_wells: number of well in this area	Contains information about oil and gas fields. This data was used to analyze the impact of different setback distances on the coverage and production of oil resources in each field.
coverage_map_files/	N/A	Contains spatial files of 1000, 2500, 3200, and 5280 foot setback coverages.
crude_prod_x_field_revised.csv	doc_field_code: specific code for extraction field doc_fieldname: name of the extraction field year: entire full year of oil extraction total_bbls: total extract barrel per oil for entire year	Contains information on crude oil production by field and year for historical trend analysis.
entry_df_final_revised.csv	doc_field_code, doc_fieldname, year doc_prod, capex, capex_bbl_rp, capex_per_bbl_reserves, capex_per_bbl_nom, opex, opex_bbl_rp, opex_per_bbl_nom, m_cumsum_div_my_prod, m_cumsum_div_max_res, capex_imputed, wm_capex_imputed, opex_imputed, wm_opex_imputed, wm_cumsum_div_my_prod, wm_cumsum_div_max_res, wm_cumsum_eer_prod_bbl, brent, new_prod, n_new_wells, top_field	Contains information on oil fields and is used for in-depth analysis of the economic performance and operational characteristics of oil fields over time. It is showing the main three categories here, including basic information, oil price, and new wells information.
field_capex_opex_forecast_revised.csv	doc_field_code, year, m_opex_imputed, m_capex_imputed, wm_opex_imputed, wm_capex_imputed	Used to project future costs associated with oil production operations, including field-level capital expenditures (CapEx) and field-level operational costs(OpEx)

field-year_peak-production_yearly.csv	doc_field_code, doc_fieldname, start_year, Peak_prod_year: the year of highest oil production for an oil field peak_tot_prod: peak annual total oil production by field no_wells: number of wells in each field peak_avg_well_prod: average of oil production per each wells and field peak_well_prod_rate: peak production rate by oil field	Contains information about the peak production year for each oil field. Used to analyze the performance and decline characteristics of oil fields based on their highest peak production levels.
forecasted_decline_parameters_2020_2045.csv	doc_field_code, doc_fieldname, year, q_i: peak production rates D: decline rates, by calculating the percent change in production rate within from the previous year's decline rate b: hyperbolic decline exponents, using a non-linear least squares (NLS) regression, int_year: the number of years since the start of production, d: exponential decline rates	Contains forecasted decline parameters for oil fields from 2020 to 2045, aiding in projecting future oil production.
ghg_emissions_x_field_2018-2045.csv	doc_field_code, doc_fieldname, year steam_field: steam injection in oil extraction process for binary indicator upstream_kgCO2e_bbl: upstream GHG emissions, including exploration, drilling, and crude production, intensity in kilograms of CO2 equivalent per barrel of oil produced	Contains information about greenhouse gas (GHG) emissions for oil fields from 2018 to 2045, facilitating analysis of the carbon footprint of oil production across different fields and to project future GHG emissions based on production forecasts.
pred_prod_no_exit_2020-2045_field_start_year_revised.csv	doc_field_code, doc_fieldname, start_year: starting year of production for each field, no_wells: the number of wells in the field, year: year of the production forecast adj_no_wells: adjusted number of wells based on the setback scenario, production_bbl: forecasted production volume in barrels	Contains predicted oil production volumes for fields from 2020 to 2045, considering different setback scenarios and assuming no field exits. This dataset is used to analyze the impact of different setback regulations on future oil production at the field level.
emission_reduction_90.csv	emission_reduction: 90 percent reduction scenario, ghg_emission_MtCO2e: GHG emissions in million metric tons of CO2 equivalent (MtCO2e)	Provides the corresponding GHG emissions in million metric tons of CO2 equivalent (MtCO2e) for 90% reduction scenario.
excise_tax_non_target_scens.csv	year: forecasting year from 2020 to 2058, tax_rate: a fraction of the oil price, excise_tax_scenario: showing either no tax or 5 percent, units: tax rate	Contains information about excise tax rates for non-target scenarios from 2020 to 2058. This dataset is used to analyze the impact of different excise tax scenarios on oil production and revenues.

inmap_processed_srm/srm_XX_fieldYY.csv	GEOID: column represents the unique identifier for each county, total chemical amount (NH3, NOX, PM2.5, SOX, VOC), average weighted chemical amount: "totalXX" and "totalXX_aw" columns represent the chemical concentrations and area-weighted chemical concentrations resulting from emissions related to the oil field's operations.	Contains information about the impact of 26 areas in California. Used to assess the spatial distribution of air quality impacts from the oil field across different counties in California.

7.6 Guidelines for Dashboard Users

Currently, the main mapping page includes two pickerInputs: county name and well type. The user will first select from a list of California counties and then choose which well type they would like to show on the map, either active or inactive. There are currently three layers to the map: point geometries of well locations throughout California, California county polygons, and a 3200 foot buffer around almost every select sensitive area. The goal is to have county-level information as a pop-up message, ideally for it to show up when the user's mouse is hovered over the county polygon.

The other pages, including the page that will display the statewide results of the implementation of the 3,200 foot setback on health, production, and labor outcomes, of the interactive dashboard are still underway. Definitions and useful information about PM 2.5, distinction between active and inactive oil wells, and what is considered a disadvantaged census tract are included on the dashboard.

8. Archive Access

The existing model, along with the revised code and file paths, will be saved on GitHub for anyone to access the scripts. Two data structure types exist for future developers: *data* and *data-str*. The data folder replicates the structure shared by clients, with the *processed* subfolder under the data folder serves as a catch-all for data generated throughout the project to regenerate the existing models. The *data-str* folder, with the full structure outlined in the Appendix, has been updated for easy injection by future developers.

There are two tracks for archiving data: one for Bren affiliates and another for clients. For Bren affiliates, both the intermediate and output public data from the *data-str* folder will be archived using Dryad. For clients, both the data and data-str folders, containing all necessary data, have been delivered.

9. References

Abadie LM, Chamorro JM. Valuation of Real Options in Crude Oil Production. Energies. 2017; 10(8):1218. https://doi.org/10.3390/en10081218

California SB1137 | 2021-2022 | Regular Session. LegiScan. https://legiscan.com/CA/text/SB1137/id/2606996

Czolowski, Eliza D et al. “Toward Consistent Methodology to Quantify Populations in Proximity to Oil and Gas Development: A National Spatial Analysis and Review.” *Environmental health perspectives* 125.8 (2017): 086004–086004. Web.

Deshmukh, R., Weber, P., Deschenes, O. *et al.* Equitable low-carbon transition pathways for California’s oil extraction. *Nat Energy* 8, 597–609 (2023). https://doi.org/10.1038/s41560-023-01259-y

Lewis C, Greiner LH, Brown DR (2018) Setback distances for unconventional oil and gas development: Delphi study results. PLOS ONE 13(8): e0202462. https://doi.org/10.1371/journal.pone.0202462

Referendum on SB 1137 | California Independent Petroleum Association. Accessed 4 June 2024. https://www.cipa.org/i4a/pages/index.cfm?pageid=1048

Zhang, Huanjia. “Californians Living within Miles of Oil and Gas Wells Have Toxic Air.” EHN, 21 Oct. 2021, www.ehn.org/oil-and-natural-gas-industry-air-pollution-2655333610.html#:~:text=People%20living%20within%202.5%20miles,Science%20of%20The%20Total%20Environment.

Appendix:

A. Tables and Figures

Table 5: Code Updates

Script	Additions
health_data.R	Called *dplyr* for select() function; updates to how missing data is handled to ensure consistency in results
source_receptor_matrix.R	Added checks if the CRS of the shapefiles and the data to be intersected are different and if so transforms CRS's to match, calling *purrr* for map() function
srm_extraction_population.R	Called *janitor* package to clean column names
ica_multiplier_process.R	Column names in the ICA files are updated to reflect the actual column names in the input file; data joined using a left join instead of an inner join; *na.rm = TRUE* argument is used in some summarize()and sum() functions to remove missing values; some changes in the arguments and column names in the pivot_wider()
stocks_flows.R	N/A
create_ccs_scenarios.R	N/A
social_cost_carbon.R	After melting the *scc_df* data.table, it is converted back to a data.table object
clean_doc_prod.R	Replacing *readtext* package with *dplyr*; call *dplyr* for select() statements
process-monthly-prod.R	N/A
process-monthly-inj.R	N/A
opgee-carb-results.R	dt_res data.table is converted using *dcast.data.table* instead of *dcast*; dt_res is explicitly set as a data.table object
rystad_processing.R	Updates to data cleaning based on column names; call *dplyr* in many cases to specify operation
zero_prod.R	Called *dplyr* for select() and filter() operations
income_data.R	Called *dplyr* for select() operations; addition of a code block to list available variables using listCensusMetadata()
ccs_parameterization.R	N/A
well_setback_sp_prep.R	sr_dwellings and sr_s objects are filtered for valid geometries using st_is_valid() and st_make_valid(); added code block for testing the conversion of data frames to spatial data frames which is commented out; created setback buffer for 3,200 foot distance scenario
gen_well_setback_status.R	Added and processed 3,200 foot setback distance scenario; field coverage calculations are performed using as_tibble() instead of as.tibble()

economically_recoverable_resources.R	N/A
create_entry_econ_variables.R	Called *dplyr* for select() operations
init_yr_prod.R	N/A
match_fields_assets.R	Called *raster* for unique() function in several cases; called *dplyr* for several select() functions
create_entry_input.R	Converted the month_year and start_date columns in init_yr_prod to Date format
crude_prod_x_field.R	N/A
field_county_producton.R	N/A
field_emission_factors_2015.R	N/A
county-setback.R	Added new code chunk to calculate the county coverage for the 3,200ft buffer scenario; updated the rbind() function to include the new county_coverage_df_3200 data frame when combining the setbacks
well_exits.R	N/A
prep_data_field_year.R	N/A
field-vintage-exit-rule.R	N/A
field-vintage-exit.R	N/A
historic-extraction-emissions.R	N/A
prep_data_field_vintage.R	N/A
decline_parameters_field_start_year.R	Updated doc_field_code to be numeric for res_all and peak_prod to ensure join compatibility
predict_existing_production.R	Added a section to handle doc_field_code as numeric for all data frames; updated the left_join of wells and setbacks to define the relationship as many-to-many; modified n_wells_area calculation to use length(unique(paste(api_ten_digit, start_year))) instead of n() due to package issues; updated group_by() and summarise() functions to use *dplyr* prefix; converted data frames to *data.table* objects in order to perform operations
analyze-parameters.R	N/A
extraction_fields.R	N/A
injection-type-by-field.R	N/A
forecast_ghg_emission_factors.R	Called *dplyr* to operate on several functions by adding it as a prefix; updated the file paths in the list.files() function to use the res_path variable and full.names = T

emissions-target-90.R	N/A
prep-excise-non-target.R	N/A
load_input_info.R	Added several libraries; used readxl or read_excel function instead of read.xslx; updated several column names and types when read in data files; converted some data frames to data.tables using setDT(); added lines to convert doc_field_code to numeric and then back to character in several places; removed the sprintf() function that pads field codes with leading zeros since it converted doc_field_code into a numeric when it is supposed to be a character; implemented Random Forest model
scenario-list-targets.R	N/A
00_extraction_steps.R	Added model toggle variable
fun_extraction_model_targets.R	Added lines for testing purposes; updated the logic for handling different target policies; added checks using if/else statements to ensure required columns exist in various data tables before performing operations; replaced some data manipulation operations with *dplyr* functions, like pivot_wider() instead of dcast(); added code to remove duplicates from prod_existing_vintage_z and handle missing fields in zero_prod_quota_old; replaced some *data.table* syntax with equivalent *dplyr* syntax, like rename() instead of setnames(); converted data frames into *data.table* objects when necessary, usually after melting or merging to uphold object type
review_target_out.R	N/A
compile_extraction_outputs_full.R	N/A
compile_subset_csvs.R	Added testing to ensure all scenarios are read in properly

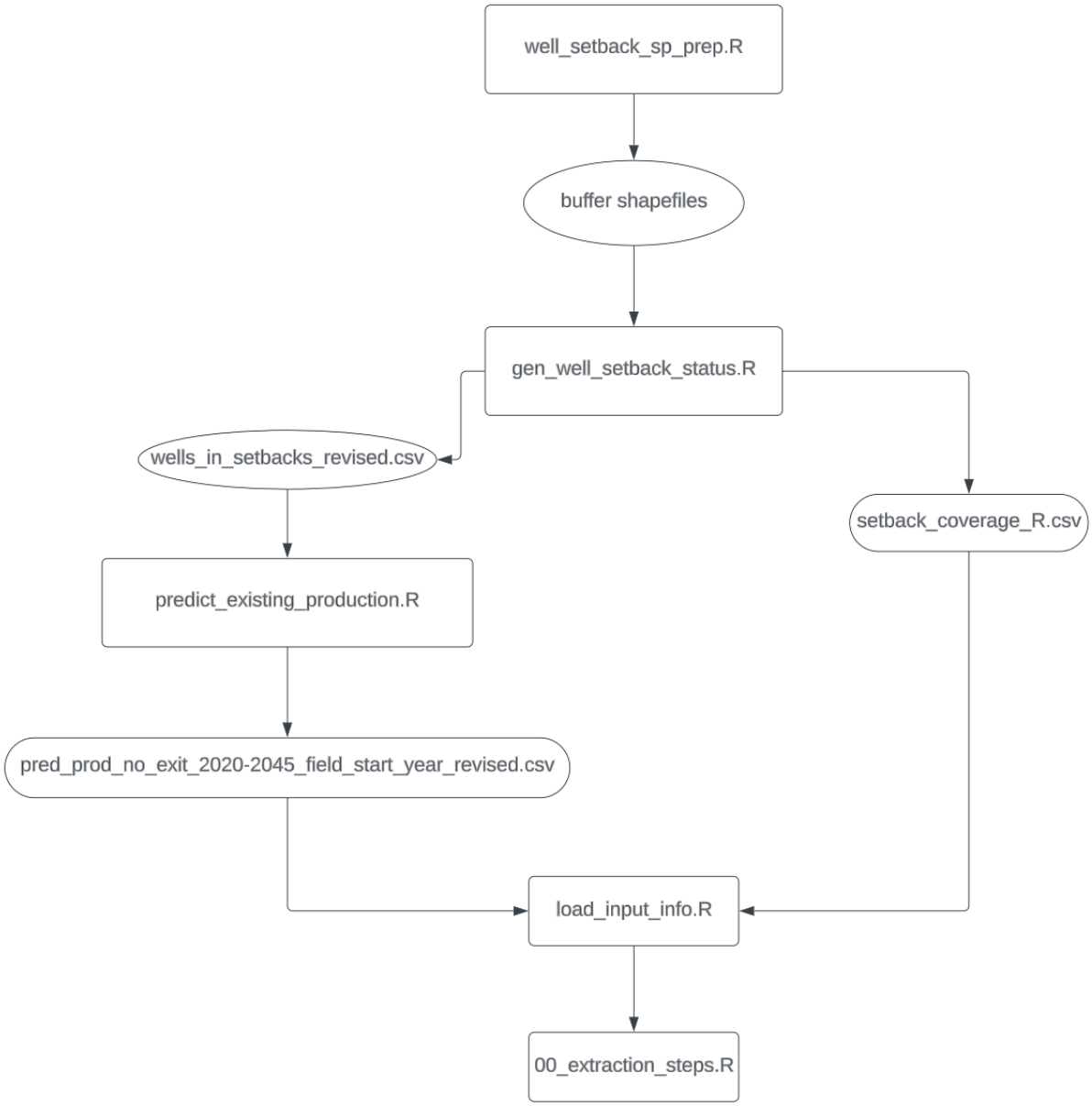

Figure 20: Visualizing the scripts that process the 3,200 foot setback scenario data.

B. Code

```
# Read the session information from the file
session_info <- readLines("sessioninfo.txt")

# Install the remotes package if not already installed
if (!require(remotes)) install.packages("remotes")

# Function to install a specific version of a package
install_specific_version <- function(package, version) {
  remotes::install_version(package, version = version)
}

# Extract package names and versions from sessioninfo.txt and install them
for (line in session_info) {
  if (startsWith(line, "package")) {
    package_info <- unlist(strsplit(line, " "))
    package_name <- gsub("'|'", "", package_info[2])
    package_version <- gsub("'|'", "", package_info[4])
    install_specific_version(package_name, package_version)
  }
}
```