Forecasting Renewable Energy Production Based on Historical Weather Data in California: Predictive Analysis using Regression Modeling

Problem

California often guesses wrong about how much solar and wind will show up **tomorrow**. Those misses force operators to curtail clean energy and keep fossil plants ready "just in case," costing **over $4.2 billion each year** *and adding avoidable CO_2.* The hard part is the weather: clouds and wind can change quickly, vary by location, and shift across seasons. My project turns reliable, widely available weather data into **unified day-ahead forecasts for both solar and wind**. The aim is practical: fewer curtailments, less fossil backup, lower costs, and steady reliability, using a method simple to retrain and update as conditions change.

Abstract

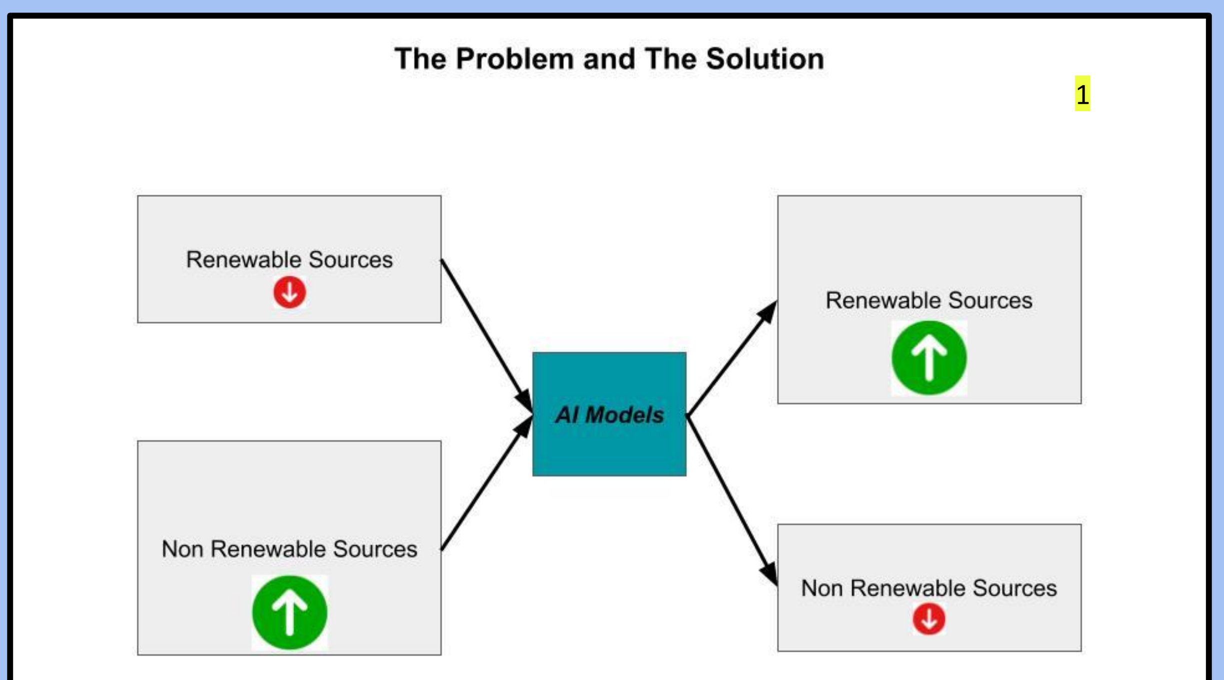
The California electricity sector wastes over $4.2 billion annually due to reliance on non-renewable energy. This project addresses the issue by developing regression models to forecast renewable production, focusing on solar and wind, using California weather and generation data from 2022–2023. Key predictors included temperature, cloud cover, solar radiation, wind speed, and day length.

The solar model achieved a mean absolute error (MAE) of ~10,000 MWh, under 39% of average daily output, while the wind model had an MAE of ~17,200 MWh (~53.6% of output), reflecting wind's volatility. A website was also built to display interactive forecasts, making the results accessible for planners and users. Attempts to model hydro were unsuccessful due to poor correlation with predictors.
https://www.weather2watts.com/

By improving forecast accuracy, the models can reduce unnecessary fossil fuel use and enable better grid management. Results suggest potential daily CO_2 reductions of up to 9,100 metric tons. While centered on California, this approach is adaptable to other regions, supporting cleaner, more resilient energy systems.

Research Question

How can renewable energy production, specifically solar and wind energy, be accurately predicted using weather data to improve energy grid management and reduce reliance on fossil fuels?



Procedure

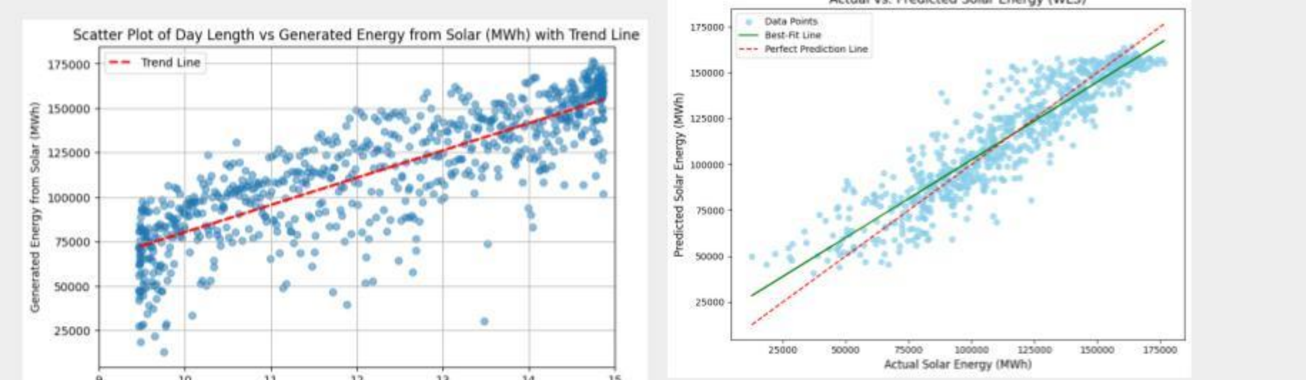
1. Data Collection: Daily renewable energy generation data (solar, wind, hydro) and weather records were collected for California from 2022–2023. Weather variables included temperature, cloud cover, solar radiation, wind speed, humidity, and day length.
2. Data Consolidation: Using Python, the datasets were merged into a single table where each row represented one day. This ensured that weather observations were directly matched with the corresponding day's energy generation.
3. Exploratory Data Analysis (EDA): Relationships between predictors and energy output were visualized using scatter plots. A global F-test confirmed the significance of predictors, while variance inflation factor (VIF) analysis and correlation matrices were used to detect and reduce multicollinearity. Added Variable Plots (AV plots) clarified the effect of individual predictors, and diagnostic plots (Residuals vs. Fitted and Normal Q-Q) were reviewed for model assumption violations.
4. Data Preprocessing: Outliers and anomalies were identified and either removed or adjusted to prevent skewing model performance.
5. Model Development: Initial models were created using Ordinary Least Squares (OLS). Weighted Least Squares (WLS) was later applied to improve robustness against heteroscedasticity. Model accuracy was assessed with R^2, adjusted R^2, residual standard error (RSE), mean absolute error (MAE), and ratio metrics.
6. Model Refinement: The models were iteratively improved by minimizing MAE and RSE while ensuring assumptions of linearity, independence, normality, and equal variance were satisfied.
7. Final Evaluation: The optimized WLS model was selected for its balance of accuracy, generalizability, and complexity. It consistently outperformed alternative approaches such as Random Forest and XGBoost given the dataset size.
8. **Website Development and Deployment:** Developed a web app to make forecasts accessible, supporting solar and wind predictions in two modes: historical data and custom weather inputs. Built with React (frontend) and FastAPI (backend), the system used WLS-trained models and was deployed on AWS (Elastic Beanstalk + Amplify with custom domain). This provides a practical, user-friendly tool for renewable energy forecasting. Website link:
https://www.weather2watts.com/

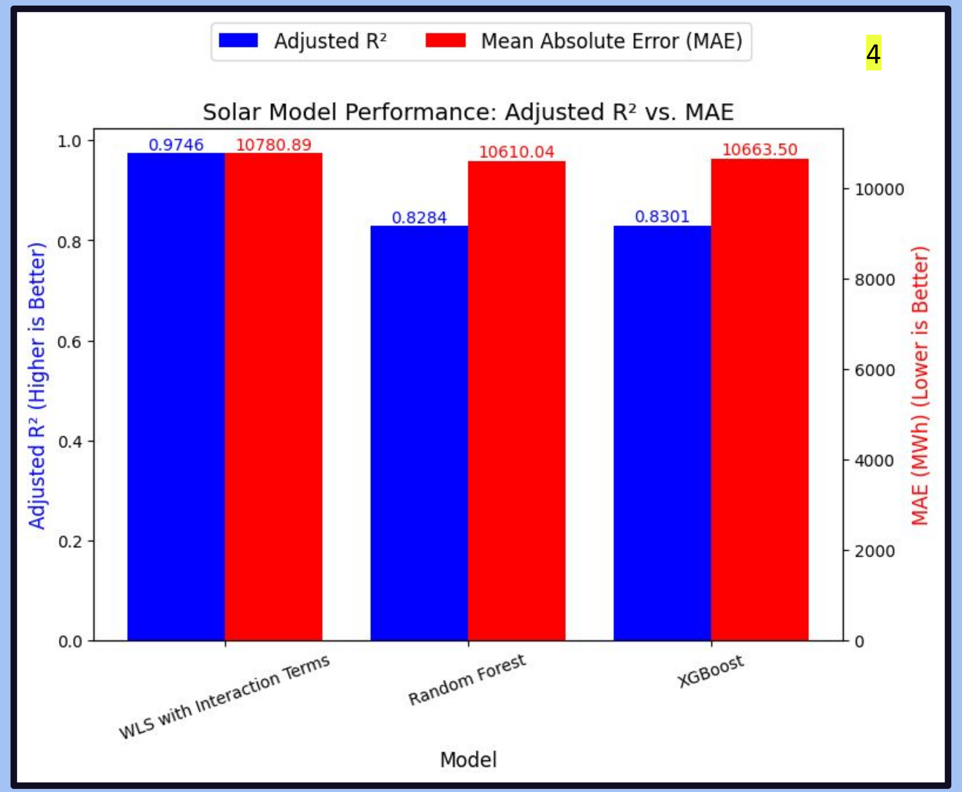
Findings - Solar

2

What did you discover and/or prove?

This project demonstrated that **solar energy** production can be reliably predicted using key weather factors such as temperature, humidity, cloud cover, solar radiation, and day length. The model proved highly effective, explaining 97% of the variation in solar energy production, meaning these weather variables account for nearly all changes in daily solar energy output. The model's predictions were also highly accurate, with an average error of ~10,780 MWh per day (Mean Absolute Error, MAE). This means that, on average, the model's predictions deviate by less than 39% from actual daily energy production. These findings highlight the strength of weather-driven predictions for solar energy forecasting and demonstrate that it is possible to estimate solar energy output with a high degree of precision.



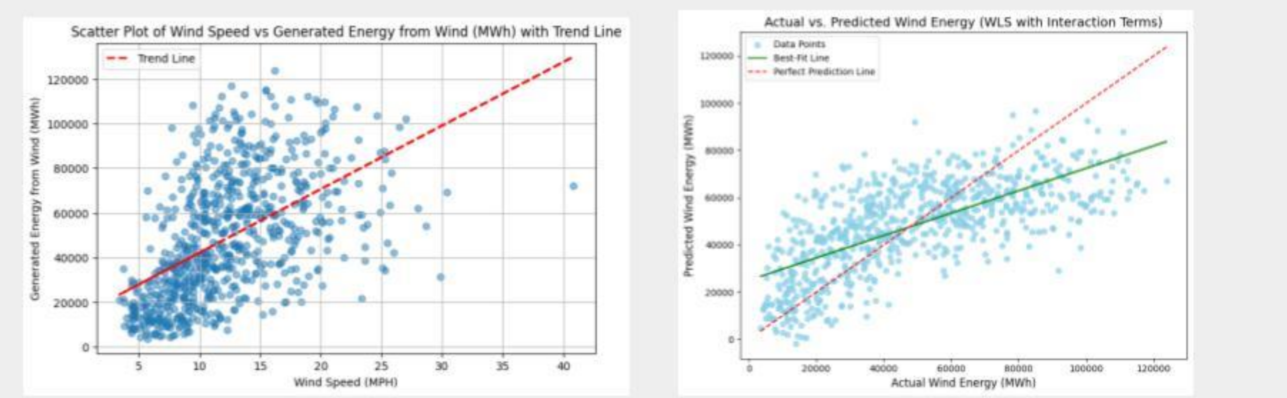


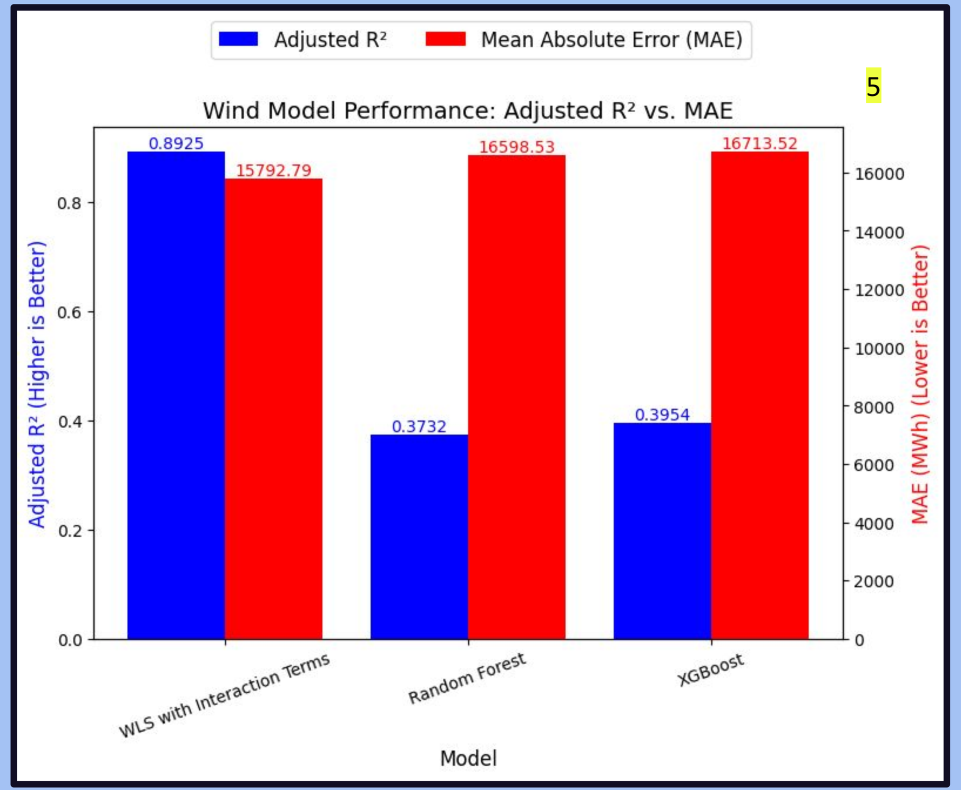
Findings - Wind

3

What did you discover and/or prove?

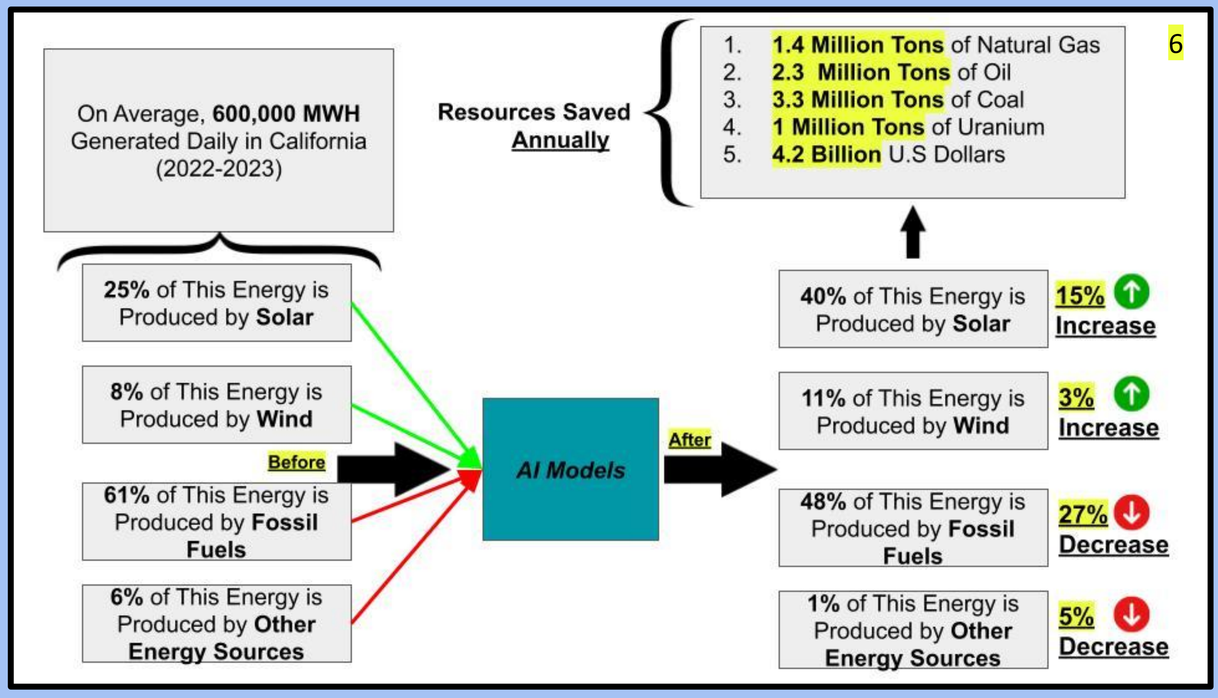
Predicting **wind energy** production was more challenging due to the unpredictable nature of wind speeds, other weather factors, and a lack of data. The model was able to capture 89% of the variation in wind energy production, meaning it successfully identified most patterns in the data. However, predictions were slightly precise, with an average error of about 15,792.79 MWh per day, accounting for 42% of daily wind energy production. This shows that while the model recognizes trends, wind energy is much harder to predict than solar energy due to its natural variability.

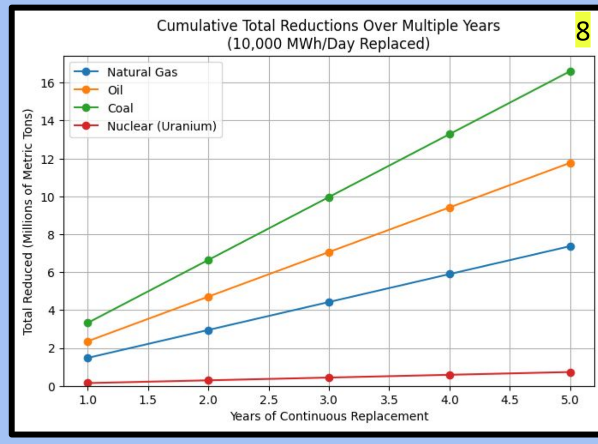




Conclusion

Overall, the predictive models for solar and wind energy demonstrated substantial promise, with the solar model achieving a mean absolute error (MAE) of about **10,000 MWh—below 39% of the average daily solar output—and the wind model at roughly 17,200 MWh (~53.6% of average daily wind production).** While wind's higher error margin reflects inherent volatility, both models offer valuable insights for daily planning by enabling grid operators to anticipate fluctuations more accurately than less sophisticated estimates. Beyond their technical performance, these linear regression models deliver tangible real-world benefits. By improving accuracy in renewable energy predictions, the project can avert up to 10,000 MWh of fossil-fueled generation per day—equivalent to about 15% of California's daily energy production—resulting in roughly **9,100 metric tons of CO_2** emissions avoided each day (or 2.35 million metric tons annually, if coal is replaced). Concretely, that's 5,000 fewer tons of coal burned daily or **33,333 barrels of oil saved per day**. In addition, more precise scheduling has the potential to reduce California's estimated **$4.2 billion in annual energy losses** tied to overproduction and fossil fuel reliance. In short, these findings highlight how robust forecasting **not only cuts emissions but also lessens costs for the state's energy sector.**





Sources

1. EIA. (n.d.). *U.S. Energy Information Administration (EIA)*. Retrieved March 28, 2025, from https://www.eia.gov/
2.Blunt, K. (2024, November 15). *Why Californians have some of the highest power bills in the U.S.* The Wall Street Journal. https://www.wsj.com/business/energy-oil/why-californians-have-some-of-the-highest-power-bills-in-the-u-s-a831b60e
3. Sample, I. (2024, December 4). *Google DeepMind predicts weather more accurately than leading system*. The Guardian. https://www.theguardian.com/science/2024/dec/04/google-deepmind-predicts-weather-more-accurately-than-leading-system
4. National Renewable Energy Laboratory. (n.d.). *State estimation and forecasting*. Retrieved March 28, 2025, from https://www.nrel.gov/grid/state-estimation-forecasting.html
5. Blumsack, S., & Fernandez, A. (2013). *Ready or not, here comes the smart grid!* Renewable and Sustainable Energy Reviews, 19, 528–543. https://doi.org/10.1016/j.rser.2012.11.032
6. Lawrence Berkeley National Laboratory. (n.d.). *Berkeley Lab estimates sustained electric efficiency*. Retrieved March 28, 2025, from https://emp.lbl.gov/news/berkeley-lab-estimates-sustained-electric
7. Lawrence Berkeley National Laboratory. (n.d.). *Energy markets & policy*. Retrieved March 28, 2025, from https://energyanalysis.lbl.gov/energy-markets-policy
8. Lawrence Berkeley National Laboratory. (n.d.). *Data for Rethinking Energy and Agriculture in Mali (DREAM)*. Retrieved March 28, 2025, from https://international.lbl.gov/dream
9. Lawrence Livermore National Laboratory. (n.d.). *Flowcharts*. Retrieved March 28, 2025, from https://flowcharts.llnl.gov/
10. National Renewable Energy Laboratory. (n.d.). *National Renewable Energy Laboratory (NREL)*. Retrieved March 28, 2025, from https://www.nrel.gov/
11. National Renewable Energy Laboratory. (n.d.). *100 percent clean electricity by 2035 study*. Retrieved March 28, 2025, from https://www.nrel.gov/analysis/100-percent-clean-electricity-by-2035-study.html
12 .ReStack. (n.d.). *Neural networks: Answer deep neural network – GeeksforGeeks Cat AI*. Retrieved March 28, 2025, from https://www.restack.io/p/neural-networks-answer-deep-neural-network-geeksforgeeks-cat-ai
13. Octopus Energy. (n.d.). *When will fossil fuels run out?* Retrieved March 28, 2025, from https://octopus.energy/blog/when-will-fossil-fuels-run-out/

Key: 1, 2, 3, and 6 were made in google slides. 4, 5, 7, and 8 were made in Google Colab through matplotlib.