Original Research Article

# Leveraging Machine Learning to Optimize Modern Renewable Energy Implementation

Nihal Kadamba[1], Mauricio Hernandez[2]

[1]*High School, Milpitas, California, United States;*
[2]*Nicholas School of the Environment, Duke University, Durham, North Carolina, United States*

## ABSTRACT

The increasing use of renewable energy sources such as solar and wind power has led to significant challenges with energy curtailment, causing substantial losses in energy efficiency. To address this, we investigated the use of machine learning models to predict solar energy output and mitigate curtailment across California, Nevada, Arizona, and New Mexico from 2018 to 2022. We hypothesized that time of the day, temperature, and irradiance would be the factors most predictive of energy curtailment across the aforementioned states. Specifically, we used XGBoost and Random Forest to predict solar curtailment and evaluate if such predictors can be used to accurately estimate curtailment in each state. Using historical weather, the characteristics of solar power plants, and energy data, we found that these models effectively predicted general energy patterns. Notably, we found that the hour of the day and the average temperature across all solar plants in each state were the main predictors in the models. Using a set of 16 predictors, the evaluated models reported mean absolute errors ranging from 1.8% to 4.8% of the historical solar energy curtailed. These results highlight the potential of machine learning to optimize renewable energy use, reduce curtailment, and improve grid reliability, offering a scalable solution to address the challenges of oversupply in renewable energy systems.

**Keywords:** renewable energy; solar curtailment; machine learning; Random Forest; XGBoost; grid reliability

## INTRODUCTION

The increasing need for renewable energy sources is driven by the rapid growth of global energy consumption. Due to policies, incentives, and cost reductions in technology, wind and solar energy have seen rapid growth in many regions around the world (1). For example, California, a major energy consumer, is projected to increase its energy demand by 76% by 2045 (2). Consequently, the state has set ambitious goals to transition to renewable sources for 60% of its energy needs (3).

While renewable energy sources can readily generate large amounts of power, the variability and unpredictability of renewable energy sources, such as high levels of wind and solar power, can pose challenges for integration into electric power systems. For instance, oversupply from these sources can lead to substantial energy curtailment.

**Corresponding author:** Nihal Kadamba, E-mail: nihalkadamba@gmail.com.

**Accepted** October 8, 2025
https://doi.org/10.70251/HYJR2348.35755763

In California, for instance, while summers are generally sunny, occasional cloudy days can lead to sudden drops in solar energy generation (4). Conversely, on days with unexpectedly high solar energy production, a significant amount of energy may be unused. This phenomenon is also observed in Texas, a state with substantial wind energy production (5). Oversupply poses a significant challenge in the realm of renewable energy, as it undermines the primary goal of decarbonizing our energy system. Curtailments cause large amounts of energy to be needlessly lost. According to the EIA, the solar and wind curtailment totaled 2.4 million MWh, a 63% increase to 2021 (6).

Western US states, particularly those on the West Coast, have a unique advantage in renewable energy due to abundant solar and wind resources. California, for example, has successfully integrated substantial renewable energy into its grid. However, managing oversupply and maintaining grid stability remain challenges (7). Similarly, Nevada (managed by NEVP) and Arizona (managed by AZPS) are expanding solar capacity and load management but also grapple with oversupply issues.

Machine learning (ML) can be used to forecast the energy output of photovoltaic and wind plants, thereby mitigating energy curtailment. ML models can minimize the uncertainty caused by unpredictable weather conditions by using region-specific data to predict wind patterns and solar output. Energy curtailment can be mitigated by identifying the efficiency of existing renewable energy sources and using ML models to predict their energy output, which can lead to improvements in the future.

The application of ML techniques in managing curtailment has garnered significant attention within the research community. Recent studies have explored the potential of ML algorithms to enhance the predictability and efficiency of renewable energy generation. For instance, Alassery et al. conducted a comparative study that investigated the efficacy of Artificial Neural Networks (ANNs), Support Vector Machines (SVMs), and Random Forests (RFs) in predicting solar radiation. Their findings suggested that ANNs exhibited superior performance compared to SVMs and RFs in accurately forecasting solar radiation patterns (9). Similarly, Shams et al. employed a range of ML models, including RFs, ANNs, Gradient Boosting Trees (GBTs), and Long Short-Term Memory (LSTM) networks, to predict wind and photovoltaic (PV) curtailment. Their research concluded that RF models yielded the most optimal results in forecasting curtailment events, highlighting the potential of RFs in optimizing renewable energy systems (8, 10).

These studies underscore the growing significance of ML as a valuable tool in addressing the challenges associated with renewable energy integration. However, solar curtailment prediction is complex due to the volatility of solar power generation and its dependence on power exchange, other units' power generation, and load demand. This paper presents a new methodology for predicting solar curtailment in California, Nevada, Arizona, and New Mexico. The methodology includes data preprocessing, normalization, ML method selection, proposed cross-validation (CV) approaches, and comparisons. The ML models implemented include Random Forest (RF), and Extreme Gradient Boosting Regression (XGBoost). These models are evaluated through 90–10% holdout and the best hyperparameters are selected for each model. The accuracies of all the approaches are compared to identifying the best solar curtailment prediction model.

We hypothesized that time of day, temperature, and irradiance would be the most predictive factors of solar energy curtailment across California, Nevada, Arizona, and New Mexico. This hypothesis was based on the understanding that these variables directly influence solar energy generation and grid balancing. By including them into our machine learning models, our goal was to evaluate whether these features could explain patterns of curtailment across different states and grid conditions. This approach builds on prior research and addresses the need for machine learning based tools to optimize renewable energy.

## METHODS AND MATERIALS

The methodology encompassed a multi-step approach (Figure 1), starting with the collection of historical weather data, encompassing solar irradiance, temperature, and wind speed, alongside comprehensive energy consumption and grid load data. This data collection was coupled with the mapping of existing PV power plant locations, incorporating their installed capacity and technological specifications.

Utilizing this gathered information, we calculated both historical PV generation, based on actual weather conditions and plant performance, and theoretical PV generation, based on optimal weather conditions and PV nameplate capacity. By comparing these two values, we estimated the amount of PV energy curtailment that

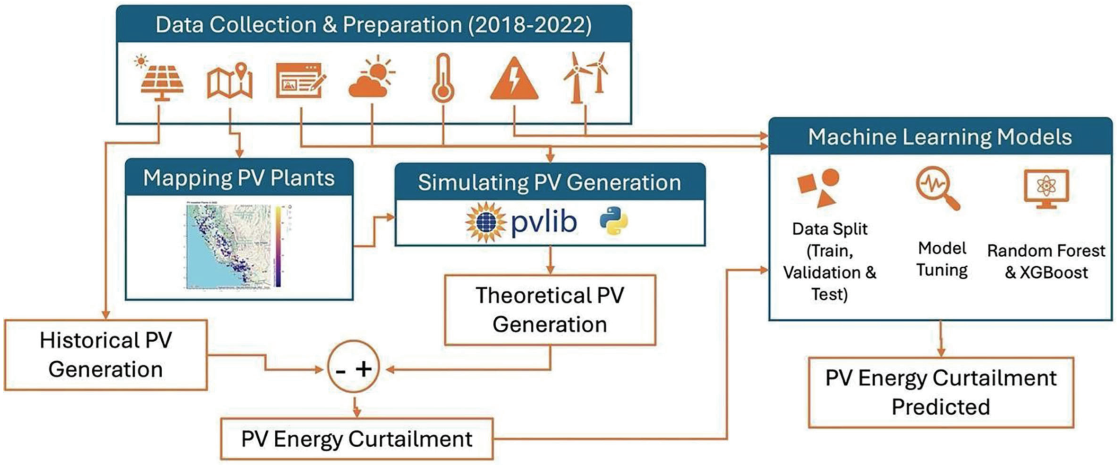


**Figure 1.** Methodology used for predicting PV energy curtailment. The flow diagram outlines data collection, feature engineering, model training, and evaluation phases for machine learning analysis.

occurred.

This rich dataset, encompassing weather, energy, and curtailment information, was then leveraged to train and fine-tune several ML models. These models were designed to predict PV energy curtailment in each of the four states included in our study.

**Data Collection**

We collected weather and energy systems data from California, Nevada, New Mexico, and Arizona which was mainly retrieved using Application Programming Interfaces (APIs). The dataset, spanning from July 2018 to December 2022, encompasses hourly intervals of PV load, actual load, forecasted load, generation capacity, and weather data. Additionally, it includes details about the location of power plants and the specific technologies employed for energy generation. Weather data was sourced from the National Solar Radiation Database (NSRDB) (11) which is a serially complete collection of satellite-derived measurements of solar radiation, global horizontal, direct normal, diffuse horizontal irradiance, and meteorological data. For the energy-related data, it was obtained from the Energy Information Administration (EIA) which collects and publishes data for every Balancing Authority in the US. These government datasets were selected for their accuracy and comprehensive coverage. Python v3.8 was used to develop code for data retrieval and management, employing the Physical Solar Model v3.2 (12), EIA API v2 (13), and Python libraries such as PVLIB (14), requests, and Pandas.

**Data Management**

In our data, the total number of hourly records we compiled was 39,420. The column types used for this dataset were wind and PV generation. For the data in 2018, we had missing values for the first part of the year, so we ignored that data and decided to only analyze data after 07/01/2018. For the rest of the year, and for the other years, we found some missing values that were replaced by interpolated values.

After obtaining the information of weather and the power grid from our four states, we prepared the data by checking for missing values, split the weather into each of the following categories: temperature, dew point, direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), global horizontal irradiance (GHI), and wind speed. Then, we merged the weather data and the power systems data. We also filtered all the PV plants from our inventory.

**Estimation of Theoretical PV Generation**

We collected the weather data using the aforementioned categories to construct our comprehensive data matrix. We identified missing values and addressed them using interpolation to maintain data integrity. A crucial step in our analysis involved obtaining the total PV generation for each state on an hourly basis. Subsequently, we used the location (latitude and longitude) of the PV plants to estimate their power generation for each hour in the 5-year period. This estimation was achieved by simulating the energy generation of each individual solar plant using the open-source software PVLIB.

Then we computed the PV energy curtailment, a critical variable in our analysis, by subtracting the historical PV generation from the theoretical PV generation. This calculation allowed us to quantify the discrepancy between expected and actual energy production.

The wealth of weather data, combined with the rest of the energy-related data collected from the Energy Information Administration (including actual power demand, forecasted power demand, power generation by technology, and total capacity of the system), served as predictors for our machine learning methods: Random Forest and XGBoost. For the model training and evaluation, the data was split into a 90/10 ratio, with the oldest data comprising the training set and the newest data comprising the testing set. Standard K-Fold cross-validation was not applied because it is not appropriate for time series data. Time series datasets have a natural temporal order where past values affect future data. K-Fold shuffles the data and disrupts this sequence, leading to data leakage by exposing future information during model training on past data.

### Machine Learning Models Applied for the Prediction of PV Energy Curtailment

Random Forest (RF), which is an ML algorithm that uses regression trees to reach a predicted output. RF creates decision trees to create an element of randomness, which is beneficial for prediction reliability because it averages outputs, which reduces overfitting. To use this, we first collect training data. Then, we adjusted its parameters, followed by a sampling of data. Next, each of the regression trees predict an output based on the testing datasets. The outputs are then averaged out to reach a final value.

XGBoost, which stands for "Extreme Gradient Boosting", is used for handling large datasets and is known for its high accuracy, making it helpful for identifying patterns in energy production and predicting curtailments. RF, which creates decision trees to create an element of randomness, is beneficial for prediction reliability because it averages output, which reduces overfitting.

### Error Metrics

Finally, we used two common error metrics used in these types of ML models, Mean Absolute Error (MAE) and the Root Mean Squared Error (RMSE), to evaluate the effectiveness of the machine learning models employed in this work.

Mean Absolute Error (MAE)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|Y_i - \hat{Y}_i\right|$$

where $n$ is the number of data points, $\hat{Y}_i$ is the predicted values, and $Y_i$ is the actual values. The MAE metric helps quantify how well a model's predictions match the actual outcomes. In our case, MAE is used to evaluate the prediction models' performance by assessing how far the forecasts deviate from observed values.

Root Mean Squared Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\Sigma^N \left(Y_i - \hat{x}_i\right)^2}{n}},\ i = 1$$

where n is the number of data points, $\hat{x}_i$ = predicted values. $x_i$ = actual values, RMSE measures the difference between predicted and observed values. This metric penalizes larger errors more than smaller ones (which is due to the squaring of the error), it is a good metric when you want to give more weight to large prediction errors.

$$R^2 = 1 - \frac{SSR}{SST}$$

where SSR (Sum of Squared Residuals) represents the total error between the observed and predicted values the sum of squared errors) and SST (Total Sum of Squares), which is the total variance in the observed data (the difference between the observed values and their mean). This formula is useful for finding a good fit of the model to the data.

## RESULTS AND DISCUSSION

Before benchmarking models, we summarized five-year meteorological variability to contextualize the main predictors. In Nevada (2018–2022), temperature and irradiance (GHI, DNI, DHI) exhibit pronounced diurnal and seasonal structure, while wind speed varies more modestly across seasons (Figure 2). These patterns align with curtailment clustering during high-irradiance hours and spring shoulder months, motivating our focus on time of day, temperature, and irradiance as primary features. Across all states and models, MAE ranged from 0.0175 to 0.0510, RMSE from 0.0317 to 0.1089, and $R^2$ from 0.0886 to 0.7705. To benchmark a linear baseline, we also evaluated Lasso; it consistently underperformed the non-linear models—for example, AZ: MAE = 0.195, $R^2 = -0.670$; NV: MAE = 0.055, $R^2 = -0.627$. This confirmed the need for RF/XGB on this problem (Table 2).

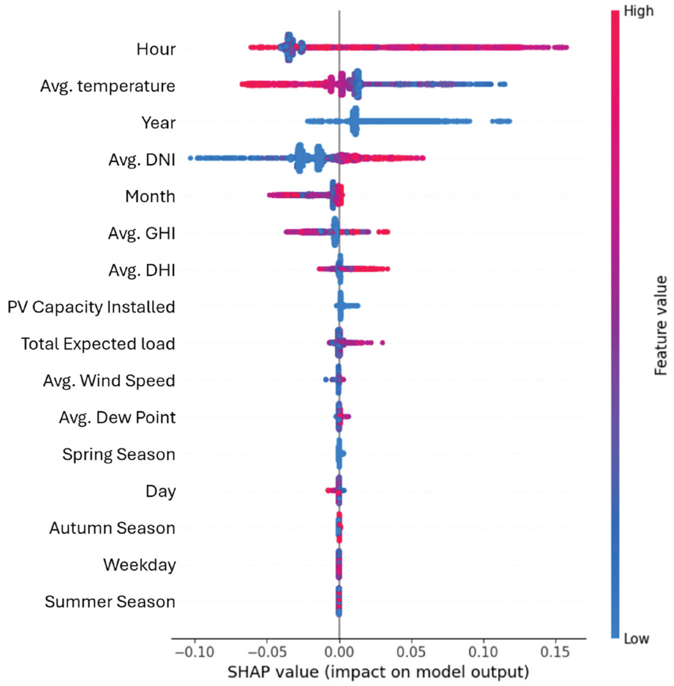


**Figure 2.** SHAP (SHapley Additive exPlanations) analysis for XGBoost energy curtailment predictions. Time, temperature, and global horizontal irradiance (GHI) were the most influential features, while seasonal variables had lower importance.

**Arizona**

XGB outperformed RF across all metrics. RMSE decreased from 0.1089 (RF) to 0.1002 (XGB), a 7.99% improvement. MAE fell from 0.0510 to 0.0485 (4.9% reduction). $R^2$ rose from 0.0886 to 0.2281 (+157.6% relative gain). While overall explanatory power remained modest, XGB better captured Arizona's sharp mid-day irradiance spikes and steep evening declines typical of desert conditions. Boosting's ability to model nonlinear transitions likely explains the advantage over RF's averaging tendency during rapid ramps.

**California**

XGB again led RF. RMSE improved from 0.0510 (RF) to 0.0431 (XGB) (15.49% reduction), MAE from 0.0234 to 0.0175 (25.21% reduction), and $R^2$ from 0.4059 to 0.5758 (+41.86%). These gains indicate better fit to California's pronounced "duck-curve" dynamics, where mid-day surplus and evening ramps drive curtailment. SHAP (Figure 2) shows hour of day and GHI dominating contributions, consistent with operational curtailment windows in CAISO.

**New Mexico**

XGB modestly outperformed RF. RMSE dropped from 0.0643 to 0.0603 (6.22%), MAE from 0.0312 to 0.0309 (0.96%), and $R^2$ increased from 0.6917 to 0.7294 (+5.45%). Both models performed well (high $R^2$), suggesting more stable mapping between weather/load features and curtailment in New Mexico. Where XGB helped most was in smoothing residual spikes around peak irradiance hours.

**Nevada**

Here RF was best. Relative to XGB, RF achieved 26.45% lower RMSE (0.0317 vs 0.0431), 25.71% lower MAE (0.0130 vs 0.0175), and $R^2$ improved from 0.5758 (XGB) to 0.7705 (+33.81%). Nevada's feature–target relationships appear more partitionable (e.g., regime-like dayparts or facility clusters), where RF's ensemble of decision regions excels. Figure 4's multi-year weather panel shows clear irradiance/temperature regimes, aligning with RF's strength.

Both models captured day-to-day trends but under-predicted extreme peaks (Figure 3) These residuals likely reflect unmodeled operational factors (e.g., transmission constraints, intertie schedules, non-solar generation dispatch, contractual curtailment rules) and sub-hourly volatility. Incorporating additional grid features (load ramp rate, import/export schedules, storage charge/discharge state) may reduce peak error. SHAP results confirm our hypothesis: time of day, temperature, and GHI were consistently top drivers. Seasonal dummies contributed less once direct irradiance features were present. This pattern held across states, with the magnitude of importance varying by climate (e.g., temperature higher in AZ/NV). The Nevada weather summary (Figure 4) visually reinforces why these features matter. To verify that model comparisons weren't driven by distribution drift, we checked train–test feature distributions; average temperature, dew point, GHI, and expected load were broadly similar with only minor tail differences (Figure 5), indicating limited covariate shift across the 90/10 split. After the figures, we report full metrics by state and model. Across all states and models, MAE ranged 0.0175–0.0510, RMSE 0.0317–0.1089, and $R^2$ 0.0886–0.7705 (Table 1). For a linear benchmark, Lasso consistently underperformed non-linear models (e.g., AZ: MAE = 0.195, $R^2 = -0.670$; NV: MAE = 0.055, $R^2 = -0.627$), reinforcing the need for RF/XGB on this task (Table 2).

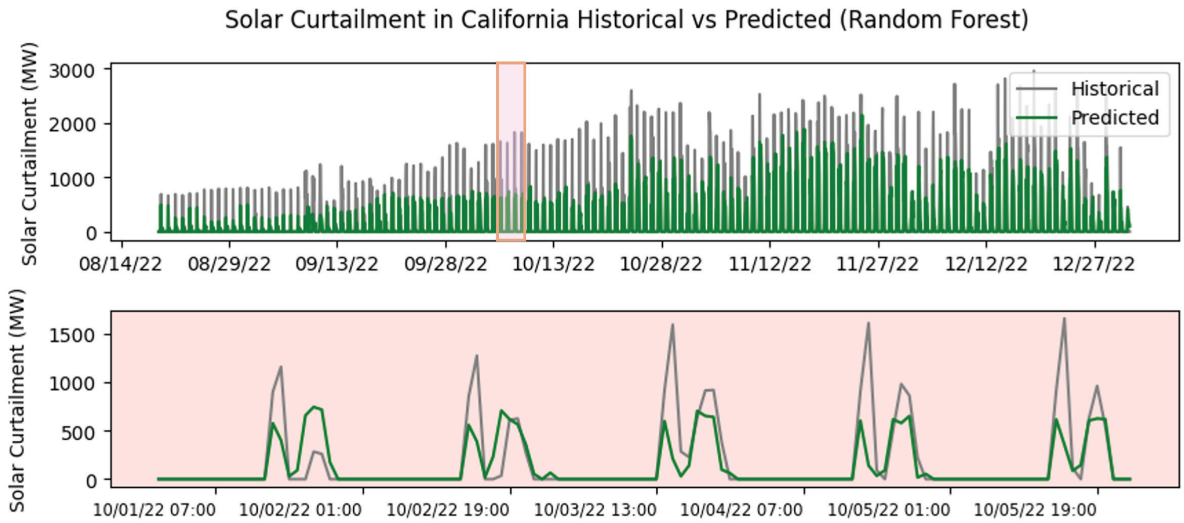


**Figure 3.** Actual and predicted solar curtailment across four states using Random Forest and XGBoost models. The graph shows model predictions compared to actual curtailment values. Error peaks indicate areas for model improvement. Models were evaluated using a 90–10 holdout validation.

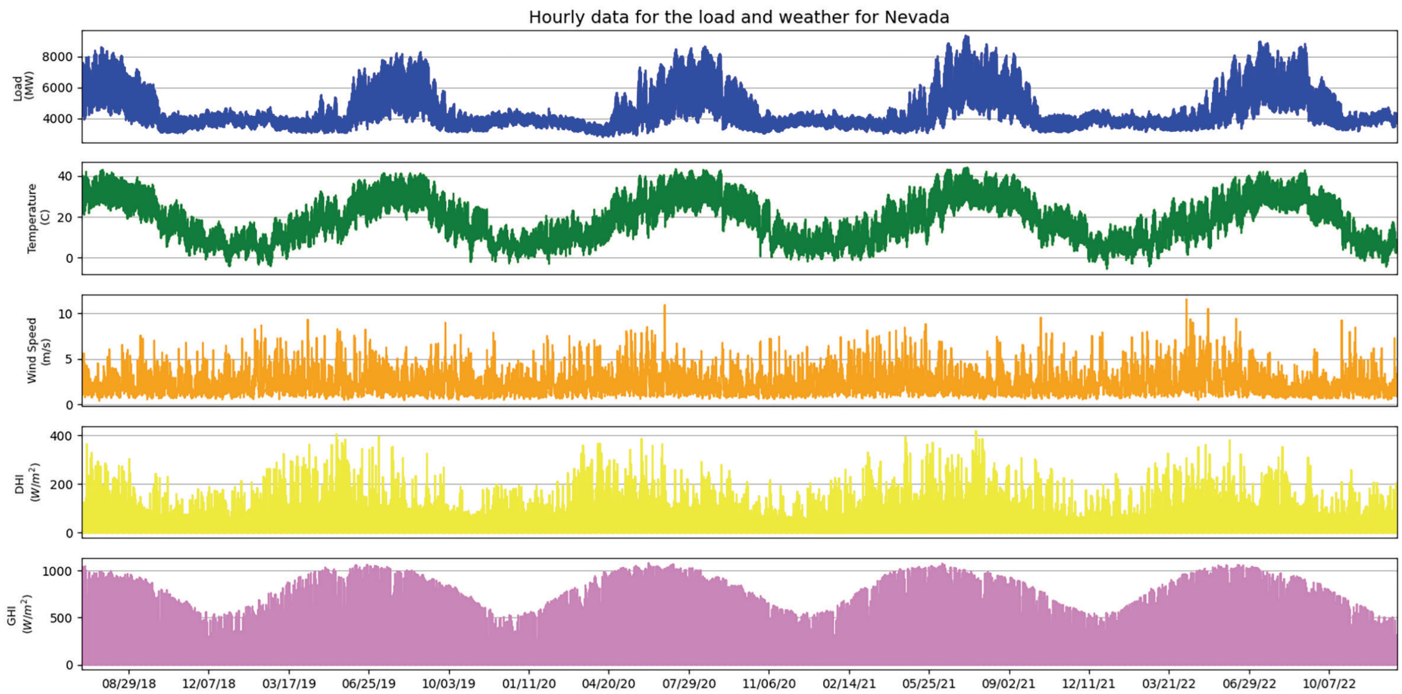


**Figure 4.** Weather data visualization for Nevada from 01/01/2018 to 12/31/2022. Includes trends in temperature, irradiance (DNI, DHI, GHI), and wind speed, used in the predictive modeling of solar curtailment.

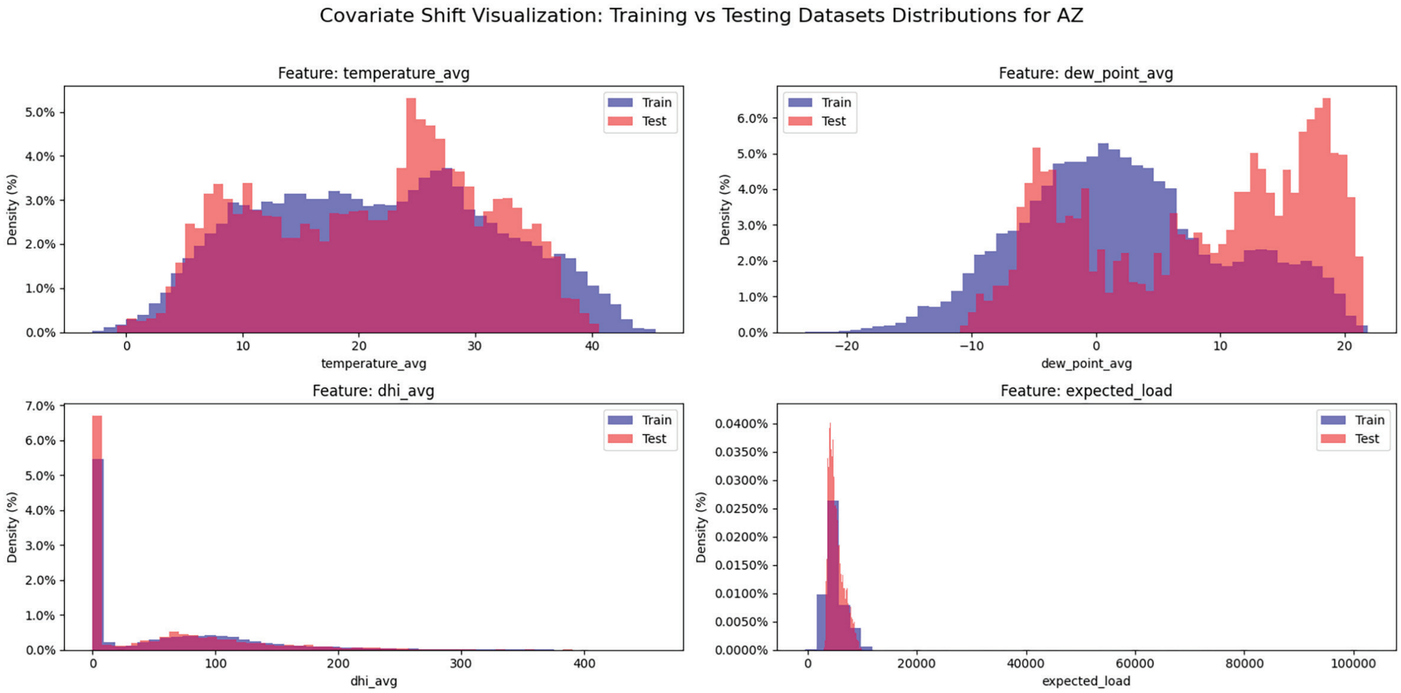


**Figure 5.** Covariate shift visualization (training vs test datasets distributions) for average temperature (top-left), dew point (top-right), direct horizontal irradiance (bottom-left), and expected electricity load (bottom-right).

**Table 1.** Machine Learning Models Performance Metrics

| State | Model | RMSE | MAE | $R^2$ |
|---|---|---|---|---|
| AZ | RF | 0.1089 | 0.051 | 0.0886 |
| AZ | XGBoost | 0.1002 | 0.0485 | 0.2281 |
| CA | RF | 0.051 | 0.0234 | 0.4059 |
| CA | XGBoost | 0.0431 | 0.0175 | 0.5758 |
| NM | RF | 0.0643 | 0.0312 | 0.6917 |
| NM | XGBoost | 0.0603 | 0.0309 | 0.7294 |
| NV | RF | 0.0317 | 0.013 | 0.7705 |
| NV | XGBoost | 0.0431 | 0.0175 | 0.5758 |

Performance metrics of machine learning models (Random Forest and XGBoost) in predicting solar energy curtailment in Arizona (AZ), California (CA), New Mexico (NM), and Nevada (NV). RMSE = Root Mean Squared Error, MAE = Mean Absolute Error, $R^2$ = Coefficient of Determination. XGBoost generally outperformed RF in CA and NM, while RF had the highest accuracy in NV.

**Table 2.** Model performance (Lasso, Random Forest, XGBoost) across Arizona, California, Nevada, and New Mexico, evaluated using Mean Absolute Error (MAE) and coefficient of determination ($R^2$)

| State | Arizona | | | California | | | Nevada | | | New Mexico | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | **Lasso** | **RF** | **XGB** | **Lasso** | **RF** | **XGB** | **Lasso** | **RF** | **XGB** | **Lasso** | **RF** | **XGB** |
| **MAE** | 0.195 | 0.051 | 0.049 | 0.042 | 0.023 | 0.018 | 0.055 | 0.013 | 0.018 | 0.043 | 0.031 | 0.031 |
| **R2** | -0.670 | 0.089 | 0.228 | -0.267 | 0.406 | 0.576 | -0.627 | 0.771 | 0.576 | 0.595 | 0.692 | 0.729 |

Lower MAE and higher $R^2$ values indicate better predictive accuracy.

## CONCLUSION

This research highlights the use of ML models, including Random Forest (RF) and XGBoost, in forecasting energy curtailments and enhancing the management of renewable energy resources in states such as California, Nevada, Arizona, and New Mexico. The models successfully identified patterns in solar and wind energy generation, with RF demonstrating particularly strong performance in Nevada, attributed to its unique climatic conditions. Evidently, our research shows the need for optimization of renewable energy grids, which is crucial for reducing oversupply and curtailment.

However, difficulties were encountered in accurately predicting peaks in energy production, showing the necessity for adjustments during periods of extreme weather. These prediction errors suggest the models may not fully capture the nonlinear dynamics or rapid shifts present during sudden weather or load changes. SHAP analysis further revealed that time, temperature, and global horizontal irradiance (GHI) were the most important features in predicting curtailment, supporting our hypothesis. Additional features such as cloud cover, humidity, load variability, or battery storage activity could enhance model accuracy in future studies.

While these findings are consistent with prior research, they imply that XGBoost may provide similar or even enhanced performance in specific contexts. Future research could improve predictive accuracy by incorporating additional data, such as detailed weather variables, as well as using more sophisticated ML methodologies. Deep learning models like Long Short-Term Memory (LSTM) networks or ensemble methods could further improve predictions by capturing temporal dependencies and edge-case behavior. Beyond curtailment prediction, these models could also be applied to real-time grid management, demand response systems, and renewable energy market forecasting. Nonetheless, the results show the potential of ML in renewable energy and in energy curtailment.

## FUNDING SOURCES

This research received no external funding.

## CONFLICT OF INTERESTS

The authors declare that there are no conflicts of interest related to this work.

## REFERENCES

1. Bird L, Milligan M, Lew D. Wind and solar energy curtailment: a review of international experience. *Renew Sustain Energy Rev.* 2016; 65: 577-86. https://doi.org/10.1016/j.rser.2016.06.082
2. California Air Resources Board. 2022 Scoping Plan for Achieving Carbon Neutrality. Sacramento (CA): California Air Resources Board; 2022. Available from: https://ww2.arb.ca.gov/sites/default/files/2023-04/2022-sp.pdf (accessed 2025-07-16).
3. U.S. Energy Information Administration. Solar and wind power curtailments are rising in California. Today Energy. 2023. Available from: https://www.eia.gov/todayinenergy/detail.php?id=60822 (accessed 2025-07-16).
4. Prol JL, Zilberman D. No alarms and no surprises: dynamics of renewable energy curtailment in California. *Energy Econ*. 2023; 120: 106974. https://doi.org/10.1016/j.eneco.2023.106974
5. Niaz H, Mahmud R, Shah A, Raza M. Can Texas mitigate wind and solar curtailments by leveraging Bitcoin mining? *J Clean Prod*. 2022; 369: 132700. https://doi.org/10.1016/j.jclepro.2022.132700
6. California Energy Commission. New data indicates California remains ahead of clean electricity goals. Calif Energy Comm. 2022. Available from: https://www.energy.ca.gov/news/2022-02/new-data-indicates-california-remains-ahead-clean-electricity-goals (accessed 2025-07-16).
7. California Independent System Operator. Integration of Renewable Resources. Folsom (CA): CAISO; 2021. Available from: https://www.caiso.com/Documents/Integration-RenewableResourcesReport.pdf (accessed 2025-07-16).
8. Shams MH, Zadeh M, Alavi M, Tahmasebi A, Ahmadi P. Artificial-intelligence-based prediction and analysis of the oversupply of wind and solar energy in power systems. *Energy Convers Manag*. 2021; 239: 114892. https://doi.org/10.1016/j.enconman.2021.114892
9. Alassery F, Khan MA, Kumar R, Alharthi M, et al. An artificial-intelligence-based solar-radiation prophesy model for green energy utilization in energy management systems. *Sustain Energy Technol Assess*. 2022; 53: 102060. https://doi.org/10.1016/j.seta.2022.102060
10. Shams MH, Zadeh M, Ghorbani N, Ahmadi P. Machine-learning-based utilization of renewable-power curtailments under uncertainty by planning of hydrogen systems and battery storages. *J Energy Storage.* 2021; 40: 103010. https://doi.org/10.1016/j.est.2021.103010
11. National Renewable Energy Laboratory. NSRDB:

National Solar Radiation Database. Golden (CO): National Renewable Energy Laboratory; 2023. Available from: https://nsrdb.nrel.gov/ (accessed 2025-07-16).

12. National Renewable Energy Laboratory. Physical Solar Model (PSM), version 3.2. Golden (CO): National Renewable Energy Laboratory; 2023. Available from: https://developer.nrel.gov/docs/solar/nsrdb/psm3-2-2-download/ (accessed 2025-07-16).
13. U.S. Energy Information Administration. EIA's Technical Documentation. Washington (DC): U.S. Energy Information Administration; 2023. Available from: https://www.eia.gov/opendata/documentation.php (accessed 2025-07-16).
14. Holmgren WF, Hansen CW, Mikofski MA. pvlib Python: a Python package for modeling solar-energy systems. *J Open Source Softw*. 2018; 3 (29): 884. https://doi.org/10.21105/joss.00884