Predicting Solar Curtailment Using XGBoost and Random Forests

Matthew Ng & Mauricio Hernandez

Received July 14, 2024
Accepted October 28, 2024
Electronic access November 15, 2024

Excess renewable energy is often wasted, or "curtailed," due to demand or transmission constraints. Predicting the timing and amount of curtailment is critical for demand response, which shifts load to periods of curtailment, maximizing renewable usage. Recent machine learning efforts have predicted when solar curtailment occurs in California, but have not yet achieved accuracy and broad applicability in predicting curtailment amounts. Here we developed models using XGBoost and Random Forests to predict solar curtailment amounts. We also estimate historical curtailment at a high spatial granularity, which can potentially enable model validation for U.S. ISOs that do not report historical curtailment. In Nevada, our XGBoost model achieved strong predictive accuracy, with an R-squared of 0.92, MAE of 22.8 MW, RMSE of 52.3 MW, and normalized RMSE of 0.045. Further testing is required to determine the models' generalizability across other regions.

Keywords: Machine learning, solar photovoltaic energy, renewable curtailment

Introduction

Renewable energy curtailment is the reduction of renewable energy due to excess production or transmission constraints. Regions with high variable renewable energy (VRE) penetration, like California and Texas, already experience significant curtailment. In 2023, the California Independent System Operator (CAISO) curtailed over 2.5 million megawatt hours (MWh) of solar energy [1]. As renewable energy penetration grows, curtailment levels are likely to increase.

Curtailment wastes energy and represents a missed opportunity to reduce CO2 emissions. Recognizing this, researchers have studied strategies to maximize the use of excess renewable generation through electricity load shifting. Many works propose synchronizing electric vehicle charging with periods of high renewable output [2,3]. Similarly, researchers have suggested allocating excess renewable energy to heat pumps [4], data centers [5], and batteries [6,7]. However, implementing these solutions requires accurate forecasts of renewable curtailment, which is challenging due to the variability of solar and wind production.

To address this problem, recent efforts have applied machine learning to forecast curtailments of renewable sources. In Shams et al., various regression models were employed to predict wind and solar power curtailment (WSPC) in CAISO, but these models yielded significant prediction errors [8]. Similarly, Hadian and Naderkhani applied an exhaustive set of regression models to predict WSPC in CAISO, with improved solar prediction accuracy [9]. Gorka and Roald introduced a gradient-boosted binary classification model to identify the presence or absence of solar curtailment in CAISO, illustrating its utility for load shifting applications [10].

While these studies are inspiring and provide valuable insights, they are not without limitations. We identify three main limitations that we aim to resolve in this work:

First, these models rely exclusively on data provided by CAISO, limiting their applicability beyond this region. They require a complete list of supply data including solar, wind, hydro, geothermal, biogas, coal, natural gas, nuclear, batteries, and imports. This extensive list of data is not easily accessible for other independent system operators (ISOs).

Furthermore, validation of these models relies on system-wide curtailment data provided by CAISO. As of 2024, the majority of US ISOs lack historical system-wide curtailment data [11].

Second, these models lack several crucial inputs that significantly affect curtailment. Specifically, since wind and solar power generation are driven by weather conditions, the absence of weather data in previous models significantly limits their forecasting capabilities [11]. Moreover, these models do not account for changes in renewable capacity across the multiple-year time frames of the studies. As renewable capacity grows without corresponding transmission system upgrades, curtailment is likely to increase [12].

Third, these works lack a method for accurately predicting the amount of solar curtailment that can be applied to other US regions. To the best of our knowledge, existing research on predicting solar curtailment focuses on California and is limited to CAISO. In Shams et al., the models had significant predictive error [8]. Although Hadian and Naderkhani [9] improved predictions using a similar methodology, they encountered the same limitations mentioned above. The models developed in Gorka and Roald [10] focused solely on identifying curtailment.

While our study aims to develop a methodology applicable beyond California, we acknowledge that further validation is necessary to establish the models' effectiveness across multiple regions.

To address these problems, this paper develops multiple regression models to predict solar curtailment amounts, using weather, irradiance, capacity, and other relevant inputs. We apply these models to the Nevada balancing authority (NEVP) as a case study. NEVP is chosen because it lacks system-wide curtailment data, similar to most ISOs, allowing us to demonstrate our approach's applicability to a U.S. region other than California. The models are trained on publicly available data on weather, capacity, and load in NEVP (39423 hours). To validate our models in the absence of ground truth curtailment data, we estimate curtailment using theoretical solar production and actual output. Since all datasets used are publicly available and cover all US independent system operators, the proposed methodology can be readily extended to other regions. We do not predict wind curtailment due to Nevada's limited onshore wind capacity, which consists of a single utility-scale wind plant generating 150 MW [13].

Nevada, with its high solar potential, already experiences renewable curtailment. As illustrated in Figure 1, NEVP has experienced a consistent expansion of total solar photovoltaic (PV) capacity for the last five years. As solar penetration in Nevada increases, solar curtailment is likely to increase.

In summary, the paper makes the following contributions:

- We develop regression models based on random forests and gradient-boosted learning to predict the amount of solar curtailment.
- The proposed approach uses publicly available datasets that collect nationwide data, with the potential goal of enabling replication across other US regions.
- It is shown that the models can accurately predict solar curtailment in the Nevada region without detailed curtailment data.
- We demonstrate the accurate performance of both models in a real world case study of the Nevada region using standard performance metrics.

Methodology

In this section, we detail the steps taken to develop models capable of accurately predicting the amount of solar curtailment in NEVP. We first discuss the sourcing and processing of input data and provide a concise overview of the photovoltaic models and ML models employed. We then discuss the performance metrics and hyper-parameter optimization.

The data collection, model training, and hyperparameter tuning were implemented using Python, pandas, and Scikit-learn version 1.2.2. The processed dataset used and an implementation of the proposed model are publicly available.

Data Sourcing and Processing

The complete list of input variables is presented as follows, categorized into three distinct groups:

- Weather: Direct normal irradiance, global horizontal irradiance, diffuse horizontal irradiance, dew point, wind speed, and temperature.
- Time: Month, day, hour, year, season
- System: Installed photovoltaic capacity, total load

Solar irradiance and meteorological data were collected from the NREL National Solar Radiation Database (NSRDB) using the Physical Solar Model (PSM) v3.2.2 [14]. For input to the machine learning models, we assigned weights to each region's weather data according to the nameplate capacity of its solar plants. We then calculated a weighted average of the weather data. Additionally, load and capacity data were obtained from the US Energy Information Administration (EIA) API v2 [15]. The total installed solar capacity, as used in the model and presented in Figure 1, was calculated by adding the individual capacities of each solar plant in NEVP. Using the Python requests library, we obtained the aforementioned raw data from the respective API services and filtered it to include only values specific to the Nevada balancing authority (NEVP).

The data were standardized to Coordinated Universal Time (UTC), averaged to one-hour intervals, and preprocessed using linear interpolation to replace empty values. We discarded the data for the first six months of 2018 because of missing information.

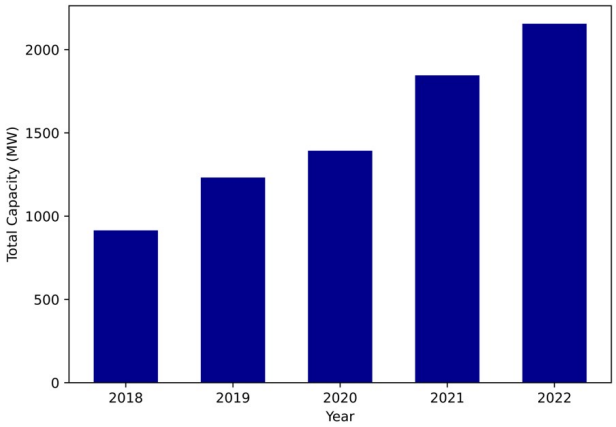

Fig. 1 Annual Solar Nameplate Capacity in Nevada (NEVP). Raw data sourced from EIA API v2.

Curtailment Estimation

The target variable, historical photovoltaic (PV) curtailment, was computed by subtracting the historical PV generation from the theoretical PV generation [16]:

$$E^c(t) = \sum_i \left(N_i(t) - E^h(t) \right) \quad (1)$$

where i denotes the index of the PV power plants operated by NEVP, $E^c(t)$ denotes the energy curtailment at time t, $E^h(t)$ denotes the amount of PV power transmitted to the grid at time t, and $N_i(t)$ denotes the estimated power output of PV plant i at time t.

We utilized photovoltaic modeling tools from the PVLIB Python library to compute theoretical PV generation for each solar plant in Nevada (NEVP). Weather data for PVLIB was sourced from the NSRDB PSM v3.2.2 at 2km spatial resolution [17]. Standard module parameters (such as efficiency and temperature coefficients) were taken from the 2009 Sandia Advent Solar Ventura module. For modeling temperature effects, parameters from the Sandia Array Performance Model [18] tailored for open rack mounted PV modules with dual glass (front and back) construction were employed.

Next, we subtracted the historical PV generation data, available from the EIA APIv2 [15], to create the historical curtailment dataset. This dataset was concatenated with the preprocessed data, resulting in a total of 39,423 timesteps between July 20th, 2018 to December 31st, 2022.

The data were split into training and testing sets in a 90/10 proportion, with the most recent data allocated to the test dataset as follows:

- Train: 7/01/18-7/20/2022 (35,481 samples)
- Test: 7/21/2022-12/31/2022 (3,942 samples)

Before training, we computed a Pearson correlation matrix to assess the relationships between all predictor variables and the target variable. Figure 2 visualizes the resulting matrix as a heatmap, with deeper shades of red indicating stronger positive correlations and deeper blues indicating stronger negative correlations. The irradiance features—direct normal irradiance (DNI), diffuse horizontal irradiance (DHI), and global horizontal irradiance (GHI)—show strong correlations due to their shared dependence on solar radiation. Installed PV capacity (PV Cap) demonstrates a strong positive correlation with the year, reflecting consistent growth in solar installations in Nevada, as illustrated in Figure 1.

For clarity, we have omitted the diagonal because each variable is correlated with itself (R=1). Likewise, we omitted the upper triangle of the matrix since it mirrors the lower triangle.

Model Selection and Tuning

We used two machine learning models: random forests (RFs), and extreme gradient boosting (XGBoost).

Random forests [19] are an ensemble learning method that constructs a multitude of decision trees during training and outputs the mean regression prediction of the individual trees. Each tree is trained on a random subset of the data and a random subset of the features. The final prediction is made by averaging the predictions of all the individual trees. For regression tasks, the predicted value is the average of predictions from all trees in the forest.

XGBoost [20] is an optimized gradient boosting algorithm designed for speed and performance. It sequentially builds trees in an additive manner, where each new tree attempts to correct the errors made by the previous trees. The predictions are made by summing the predictions of all the trees. XGBoost aims at optimizing the objective function, expressed as:

$$\Omega(\theta) = \sum_{i=1}^{n} l(y_i, y_i) + \sum_{k=1}^{K} \beta(f_k) \quad (2)$$

where l is the loss function, β is a regularization term, y_i are the actual values, $\bar{y}_i$ are the predicted values, n is the number of instances in the training set, and K is the number of trees.

Performance Metrics

We evaluate the accuracy of our machine learning algorithms using several error metrics: the coefficient of determination (R^2), root mean square error (RMSE), and mean absolute error (MAE).

These metrics are expressed in the following equations:

$$R^2 = 1 - \frac{\sum_t (E^c(t) - E^f(t))^2}{\sum_t (E^c(t) - \overline{E^c(t)})^2} \quad (3)$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} \left(E^c(t) - E^f(t)\right)^2} \quad (4)$$

$$\mathrm{MAE} = \frac{1}{n} \sum_{t=1}^{n} \left| E^c(t) - E^f(t) \right| \quad (5)$$

where E^c (t) represents the actual solar curtailment. Each metric provides different insights into the accuracy of the regression model. R^2 quantifies how much of the variance in actual curtailment is explained by the input variables in the regression model. RMSE measures the average error magnitude between forecasted and actual curtailment values, emphasizing larger errors. MAE represents the average error magnitude in forecasted curtailment values.

Since RMSE and MAE are measured in the same units as solar curtailment, they provide useful metrics to evaluate the regression models. However, since RMSE and MAE can be

misleading depending on the scale of curtailment error, we also include R^2 for a more comprehensive assessment. We do not include Mean Absolute Percent Error (MAPE) because it becomes arbitrarily high when $E^c(t)$ is 0, making it a non-useful metric.

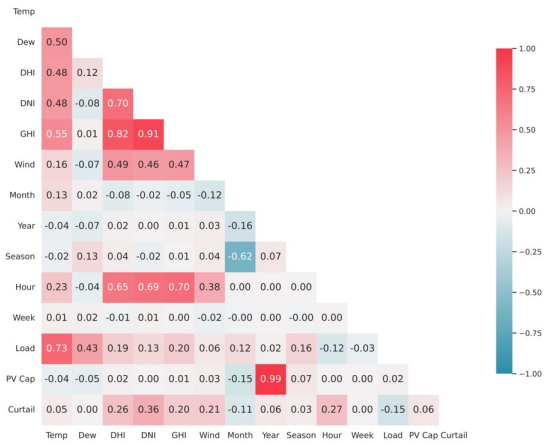

Fig. 2 Pearson correlation heatmap depicting the relationships among predictor variables and the target. Dew, DNI, DHI, GNI, Wind, and PV Cap denote average dew point, diffuse horizontal irradiance, direct normal irradiance, global horizontal irradiance, wind speed, and installed PV capacity, respectively.

Hyperparameter Optimization

We performed a grid search to tune the following hyperparameters for the XGBoost model: maximum tree depth (max_depth), minimum sum of instance weights required in a child node (min_child_weight), and minimum loss reduction required to partition a leaf node (gamma). The grid search algorithm trained the model on every combination of these hyperparameters and evaluated its accuracy using RMSE. The training set validation employed a 5-fold rolling cross-validation, maintaining temporal order to ensure the model did not access future data. Table 1 presents the specific search space and the optimal values found.

To reduce computation time, we preselected a learning rate (eta) of 0.1 and the number of estimators (n_estimators) to 100, both commonly used defaults. All other hyperparameters were set to their default value in version 2.0.3 of the XGBoost library. We used MAE as the scoring metric for hyperparameter selection.

Hyperparameter	Tested Values	Optimal
max_dept	2, 4, 6, 8	4
min_child_weight	1, 3, 5, 7	3
gamma	0, 0.1, 0.2, 0.3	0.1

Table 1 Hyperparameter Grid Search Results

Results

This section presents the performance of both models on various metrics, comparing them against each other and briefly to the state of the art.

ML Method	R^2	RMSE (MW)	MAE (MW)
Random Forest	0.89	60.28	24.71
XGBoost	0.92	52.35	22.83

Table 2 Performance comparison of Random Forest and XGBoost

We first evaluate the performance of both models on the test set, corresponding to a six month period. Table 2 presents the error metrics of the RF and XGBoost models implemented in this paper. We notice the R2 is high for both models and the RMSE and MAE are quite low, which suggests that both models are able to accurately predict solar curtailment. To interpret the RMSE relative to the scale of curtailment values, we normalized the RMSE as

$$\text{NRMSE} = \frac{\text{RMSE}}{y_{\max} - y_{\min}} \tag{6}$$

where $y_{\max} - y_{\min}$ represents the range of actual values in the dataset. Using this normalization, we found an NRMSE of 0.0.045 for XGB and 0.051 for RF. While the XGBoost and RF models show comparable accuracy, the XGBoost model demonstrates superior performance in all other evaluated metrics.

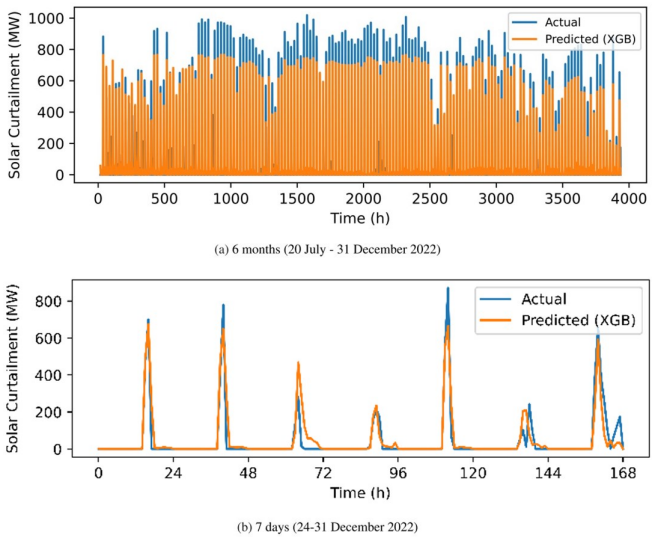

Fig. 3 Comparison of XGBoost (XGB) predicted versus actual solar curtailment.

We further evaluate the performance by presenting graphs to visualize the accuracy of the model. Figure 3 presents a comparison between the actual and predicted solar curtailment using the

XGBoost (XGB) model over the six month test period, as well as a detailed view over one week. Similarly, Figure 4 presents the actual and predicted solar curtailment using the Random Forest (RF) model over the same time period. In both graphs, the actual solar curtailment is depicted with a blue line, whereas the predicted is depicted with an orange line. We notice that both the RF and XGB model predictions are close to the actual solar curtailment values. In Figure 3(b) and Figure 4(b), we can see that the XGB and RF models often underpredict peak solar curtailment levels during this one week period. Nevertheless, these models significantly outperform those discussed in Shams et al.[8], which reported substantial inaccuracies in solar curtailment predictions. Overall, these results indicate that the proposed models accurately predict the amount of curtailment.

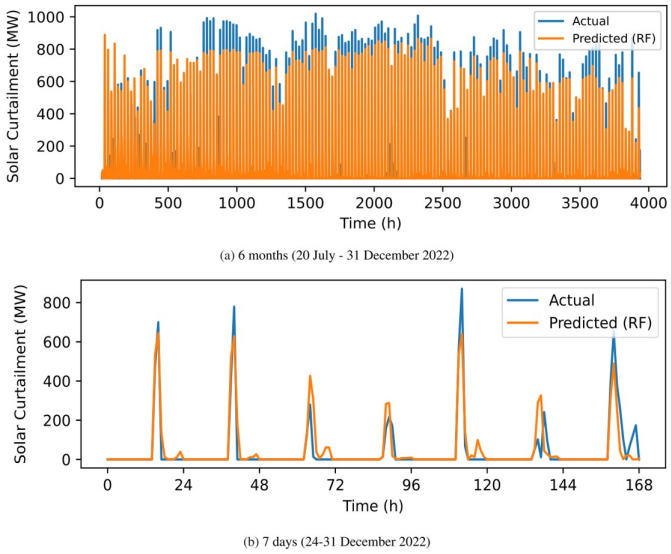

Fig. 4 Comparison of Random Forest (RF) predicted solar curtailment versus actual curtailment

Figure 5 presents the feature importance for the tuned XGBoost model, visualized as a beeswarm plot. The SHAP (SHapley Additive exPlanations)[21] values measure the contribution of each feature on the XGB output: positive values correlate with increased predictions, while negative values indicate the opposite. To investigate which features account for the majority of the model's predictions, we conducted a Pareto analysis. The results are shown in Figure 6. As expected, the hour of day significantly affects curtailment predictions due to the time-dependent nature of solar production. For instance, there are periods in the day, such as at night, for which solar curtailment cannot occur. Additionally, residential electricity demand lowers during the midday hours, as shown by the well-documented "duck curve." This temporal mismatch between demand and generation is a key predictor of curtailment. Furthermore, direct normal irradiance (DNI) and global horizontal irradiance (GHI) are shown to be the 2nd and 3rd important features. We can see that higher (red) values of average DNI in Figure 5 are associated with increased curtailment predictions, because high irradiance causes more solar generation. Without sufficient demand or storage capacity, this excess energy must be curtailed.

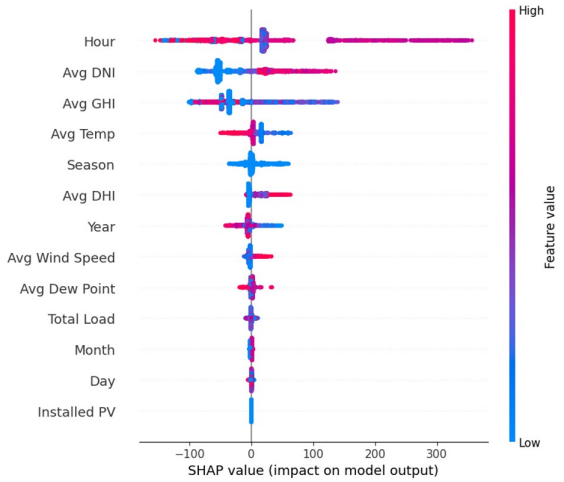

Fig. 5 Feature importance for solar curtailment in XGBoost model. The SHAP (SHapley Additive exPlanations) value measures the impact of each feature on model output. All weather inputs is a weighted average.

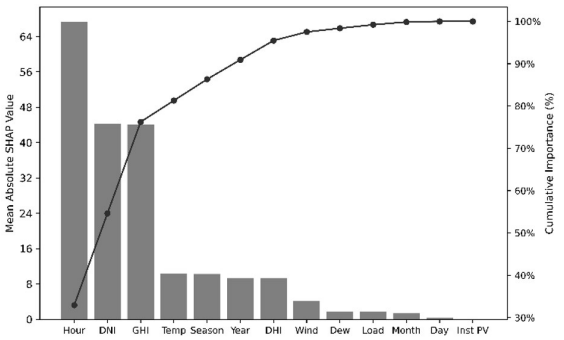

Fig. 6 Pareto chart displaying mean absolute SHAP values for features in the XGBoost model. Three features-hour of day, direct normal irradiance, and global horizontal irradiance-account for nearly 80% of the model's predictions.

From an economic standpoint, the importance of hour and irradiance as predictors suggests that strategies such as load-shifting, which align energy consumption with peak solar production, could yield significant cost savings. For instance, Hledik et. al estimate that deploying 1,000 MW of battery storage would be necessary to achieve a curtailment reduction of 50%, at a capital cost in the low hundreds of millions[22].By using the proposed machine learning approach to predict curtailment and optimize load shifting, these capital costs could be potentially reduced or avoided.

To analyze the percent of solar capacity wasted by curtailment,

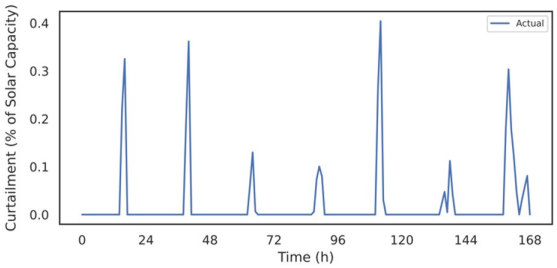

Fig. 7 Solar curtailment in Nevada as a percentage of installed solar capacity over a one-week period (24-31 December 2022)

we normalized curtailment values by dividing them by Nevada's installed solar capacity. Figure 7 presents the percentage of solar capacity wasted by curtailment over a one-week period. We observe that the percentage of NEVP's capacity that is wasted by curtailment is below 0.5%, indicating that solar curtailment is currently not a significant issue for Nevada. However, this scenario may change as solar penetration increases. We also see that peak curtailments consistently occur between 2 PM and 4 PM daily, when solar production is high and demand is relatively low.

Discussion

Due to the lack of curtailment forecast data, it is difficult to implement energy load shifting strategies that effectively optimize excess renewable generation. While recent efforts have used machine learning to forecast curtailment within CAISO, they lack the key capability of accurately predicting the magnitude of energy curtailed. Additionally, these works do not incorporate weather or PV capacity data, and their applicability to regions beyond CAISO is limited by data availability constraints.

In this work, we introduced two solar curtailment forecasting models using XGBoost and Random Forests to help fill these gaps. Our approach accurately predicted solar curtailment amounts in Nevada using publicly available datasets that cover the entire US. We also developed a method for simulating curtailment in the absence of ground truth data by implementing photovoltaic modelling tools alongside historical solar generation. . The proposed method shows potential promise for allowing curtailment models to be validated even when historical curtailment data is unavailable-a common limitation faced by most US ISOs.

While our models successfully demonstrated accurate predictions in the Nevada region, we acknowledge that the generalizability of these findings to other regions remains untested. Future work could evaluate these models with multiple other U.S. independent system operators to assess their broader applicability. Additionally, testing the model over a one year time period would enable a more comprehensive analysis of seasonality. Nonetheless, we point out that no previous models were capable of accurately forecasting the amount of solar curtailment in any US regions outside of California. The proposed models offer an immediate benefit by providing detailed curtailment information that is currently scarce in Nevada. Moreover, these models hold potential promise in facilitating strategies to optimize surplus solar energy, including electric load shifting, battery storage, and demand response initiatives. One limitation of these models lies in the use of photovoltaic modelling tools to estimate curtailment. To address the general unavailability of ground truth curtailment, we rely on estimated theoretical solar production to determine curtailment as shown in equation (1). While PVLIB [17] provides suitable accuracy for modelling theoretical solar production, future work could benefit from incorporating ground-truth curtailment data to fully validate models.

When deploying these models, there are several potential methods, such as forecasting day-ahead inputs and feeding them into the trained model. However, a primary challenge faced by the model is the effect of climate change on weather inputs. To address this issue, we propose periodically retraining the models to account for new climate trends. The implementation cost of deploying the proposed approach is unknown; however, it is likely more cost-effective than alternate methods of reducing curtailment.

The application of machine learning for forecasting renewable energy curtailment remains a promising approach to optimize renewable energy generation. We acknowledge that physics-based models are an alternate approach that can also yield successful results. The machine learning approach provides advantages such as reduced complexity and computational time when compared to physics-based models; however, a hybrid approach that combines these methodologies may yield improved results and better generalizability. Future studies could investigate the merits of this approach.

In future studies, we aim to enhance our models with real-time forecasting capabilities by leveraging real-time datasets, such as those published by CAISO [23]. We also plan to incorporate higher time resolution datasets, available from some ISOs, to further refine our models' temporal granularity beyond hourly intervals.

References

1 U. I. Administration, *Solar and wind power curtailments are rising in california*, `https://www.eia.gov/`, accessed: 2024-07-08.

2 J. Dixon, W. Bukhsh, C. Edmunds and K. Bell, *Scheduling electric vehicle charging to minimise carbon emissions and wind curtailment*, `https://doi.org/10.1016/j.renene.2020.07.017`, Online]. Available:.

3 H. Kikusato, Y. Fujimoto, S.-i. Hanada, D. Isogawa, S. Yoshizawa, H. Ohashi and Y. Hayashi, *Electric vehicle charging management using auction mechanism for reducing pv curtailment in distribution systems*.

4 M. Brunner, K. Rudion and S. Tenbohlen, *Pv curtailment reduction with smart homes and heat pumps*.

5 J. Zheng, A. Chien and S. Suh, *Mitigating curtailment and carbon emissions through load migration between data centers*.

6 C. Root, H. Presume, D. Proudfoot, L. Willis and R. Masiello, *Using battery energy storage to reduce renewable resource curtailment*.

7 T. Masuta, J. Silva Fonseca, H. Ootake and A. Murata, *Application of battery energy storage system to power system operation for reduction in pv curtailment based on few-hours-ahead pv forecast*.

8 M. Shams, H. Niaz, B. Hashemi, J. Liu, P. Siano and A. Anvari-Moghaddam, *Artificial intelligence-based prediction and analysis of the oversupply of wind and solar energy in power systems*, https://doi.org/10.1016/j.enconman.2021.114892, Online]. Available:.

9 H. Hadian and F. Naderkhani, *Deep Learning-Based Models for Wind and Solar Curtailment Forecasting*, https://, Online]. Available:.

10 J. Gorka and L. Roald, *Classification models for forecasting and real-time identification of solar curtailment in the california grid*.

11 B. Acun, B. Morgan, H. Richardson, N. Steinsultz and C.-J. Wu, *Unlocking the potential of renewable energy through curtailment prediction*, arXiv preprint arXiv:2405.18526,.

12 U. I. Administration, *A case study of transmission limits on renewables growth in texas*, https://www.eia.gov/electricity/markets/quarterly/archive/2023/transmission_limits_07_2023.pdf, Online]. Available:.

13 U. I. Administration, *Nevada state energy profile*, https://, accessed: July 8, 2024. [Online]. Available:.

14 M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin and J. Shelby, *The national solar radiation data base (nsrdb*.

15 U. I. Administration, *EIA api documentation*, https://www.eia., accessed: July 8, 2024. [Online]. Available:.

16 *California Independent System Operator, "Impacts of renewable energy on grid operations*, https://www.caiso.com/.

17 K. Anderson, C. Hansen, W. Holmgren, A. Jensen, M. Mikofski and A. Driesse, *pvlib python: 2023 project update*, https://doi.org/10.21105/joss.05994, Online]. Available:.

18 J. Kratochvil, W. Boyson and D. King, *Photovoltaic array performance model*, https://energy.sandia.gov/wp-content/gallery/, Online]. Available:.

19 L. Breiman, *Random forests*.

20 T. Chen and C. Guestrin, *Xgboost: A scalable tree boosting system*.

21 S. Lundberg and S.-I. Lee, *A unified approach to interpreting model predictions*, https://arxiv.org/abs/1705.07874, Online]. Available:.

22 R. Hledik, J. Chang, R. Lueken, J. Pfeifenberger, J. Pedtke and J. Vollen, *The Brattle Group, prepared for Public Utilities Comission of Nevada Governor's Office of Energy*, p. 201.

23 *California Independent System Operator, "Today's Outlook*, https://www.caiso.com/todays-outlook, accessed: July 8, 2024. [Online]. Available:.