Advancing Solar Irradiation Prediction in Extreme Climates: A LASSO Regression Analysis in Tomsk

David Akpuluma[1,3] and Aleksey Yurchenko[1,2]
[1]*National Research Tomsk Polytechnic University, Lenin Avenue 30, 634050 Tomsk, Russia*
[2]*National Research Tomsk State University, Lenin Avenue 36, 634050 Tomsk, Russia*
[3]*Centre for Nuclear Energy Studies, University of Port Harcourt, East-West Road, PMB 5323 Choba, Port Harcourt, Nigeria*
{aa06, reaper}@tpu.ru, david.akpuluma@uniport.edu.ng, akpoebi@gmail.com

Keywords: Solar Irradiation Forecasting, LASSO Regression Analysis, Climatic Variability in Energy, Modelling Renewable Energy Prediction in Siberia, Meteorological Data Analytics in PV Systems.

Abstract: This study presents a robust approach to predicting solar irradiation in the challenging climatic conditions of Tomsk using LASSO regression, with a particular emphasis on interpretability and climatic variability. Two distinct models were developed: Model 1, integrating specific humidity at 2 meters, and Model 2, excluding this variable to assess the impact of a wider range of meteorological factors. The comprehensive meteorological dataset from NASA's POWER database underpinned the analysis. The models' efficacy was demonstrated by impressive R-squared values: 0.843 for Model 1 and 0.813 for Model 2, indicating a substantial proportion of variance in solar irradiation was captured. Notably, Model 1's RMSE of 0.0353 and Model 2's RMSE of 0.0386 affirm the precision of the predictions. The study advances the predictive modeling of solar power output, offering valuable contributions to renewable energy forecasting literature and operational practices by providing a methodological framework that is both accurate and comprehensible, even amidst the complexities of extreme weather patterns.

1 INTRODUCTION

In the wake of pressing global challenges such as climate change and resource scarcity, the United Nations' Sustainable Development Goal 7 (SDG7) underscores the imperative of universal access to sustainable energy, emphasizing renewable sources and energy efficiency as pivotal elements in the achievement of this goal [1].

Concurrently, transition engineering emerges as a crucial discipline, orchestrating the shift towards sustainability in energy systems [2]. This field meticulously blends engineering principles with long-term strategic planning to ensure that future energy demands are met in harmony with the environment, thus facilitating a robust and sustainable societal infrastructure [3].

The intersection of SDG7 and transition engineering's targets encapsulates the essence of modern energy strategies that are environmentally sound, economically feasible, and socially inclusive [4], [5].

Against the current backdrop, the utilisation of advanced machine learning techniques, particularly LASSO (Least Absolute Shrinkage and Selection Operator), marks a notable advancement in the field of predictive analytics for photovoltaic (PV) power generation. This paper introduces a pioneering case study in Tomsk, leveraging these sophisticated methods to enhance the accuracy of PV power predictions. LASSO, with its proven effectiveness in feature selection and its ability to manage overfitting in complex datasets, provides a robust analytical framework. This approach adeptly addresses the challenges associated with the variability and complexity of PV power output, thereby facilitating more precise and reliable solar energy forecasting. This methodological choice is especially pertinent in contexts requiring clear and transparent model-driven decisions, aligning with the increasing need for comprehensible and accountable predictive models in various sectors.

The purpose of this study is to showcase the efficacy of LASSO in precisely predicting PV power generation [6]. The use of LASSO, an interpretable

model, is particularly significant in contexts where legal considerations demand transparency and justification in decision-making processes. By employing LASSO, this research not only adheres to the core principles of SDG7 but also aligns with the objectives of transition engineering. This alignment enhances the reliability and efficiency of renewable energy resources, ensuring that the methods used for forecasting and analysis are both legally compliant and transparent. The importance of such interpretable models is underscored in environments where the decisions and predictions of AI-driven models must be clear and justifiable, especially in light of increasing regulatory scrutiny in the use of complex algorithms in various sectors.

The future scope of this research is expansive. It will delve into the integration of real-time data feeds to enhance model responsiveness, the exploration Explainable Artificial Intelligence (XAI) tools for broader interpretability spectrum, and the scalability of the proposed framework to other regions and renewable energy forms. By doing so, it aims to contribute significantly to the literature on renewable energy forecasting and the operational optimization of PV systems, thereby supporting the global endeavour towards a sustainable and resilient energy future.

2 METHODOLOGY

2.1 General Methods

In the field of photovoltaic power prediction, various methods are commonly used, each, with its strengths and weaknesses. Statistical techniques like the integrated moving average (ARIMA) model are known for their ability to estimate solar power effectively by examining the linear relationships in past data [7]. However, these approaches may not fully grasp the complexities in solar energy data. To address this, machine and deep learning methods have been increasingly applied, with a focus on data processing, feature extraction, and uncertainty evaluation[12], [13].

Machine learning methods such as Random Forest (RF) and Support Vector Machines (SVM) are lauded for their capability to identify linear connections between multiple input variables and solar energy production [8]. These algorithms excel at handling datasets with numerous variables, offering a deeper understanding of solar energy systems [14].

Nevertheless, their performance relies heavily on the quality and quantity of training data available which can be a limitation. Artificial Neural Networks (ANNs) including variants, like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have emerged as tools in solar power forecasting [9]. Their ability to model patterns and adapt to forecasting tasks is unmatched.

However, the lack of transparency, in Artificial Neural Networks (ANNs) can make it difficult for practitioners to understand how the models make decisions. Physical models, which are based on photovoltaic principles and solar geometry use factors like irradiance, panel positioning and temperature to predict energy output [10].

While these models provide insights into power generation, they may not be able to account for all real-world variables accurately leading to potential prediction errors. Hybrid models aim to combine the strengths of modeling approaches by integrating machine learning with methods. This combination seeks to improve prediction accuracy by harnessing the capabilities of machine learning and the foundational principles of physical models [11].

Although hybrid models show promise in enhancing forecasting accuracy, their complexity and reliance on data sources can be challenging. Ultimately choosing a power forecasting method depends on balancing factors like accuracy, interpretability and computational requirements based on task needs. The ongoing advancements in techniques and data availability are driving the improvement of these methodologies, for more reliable and efficient solar power forecasting.

2.2 LASSO Framework

The approach followed in this work focuses on two models, Model 1 and Model 2, for predicting solar irradiation in Tomsk using LASSO regression. Model 1 examines the relationship between all-sky surface UVB irradiance (allsky_sfc_uvb) and various independent meteorological variables. It highlights the significant influence of specific humidity at 2 meters (qv2m) on the predictive model. The performance of Model 1 is evaluated using metrics like RMSE, MSE, and R-squared, indicating a strong predictive capability.

Model 2, in contrast, explores the same dependent variable (allsky_sfc_uvb) but excludes the qv2m variable from its analysis. This model assesses the impact of other meteorological variables, including wind direction at 10 meters (wd10m) and wind speed at 10 meters (ws10m), on solar irradiation

predictions. Similar to Model 1, Model 2's effectiveness is gauged using RMSE, MSE, and R-squared metrics, which provide insights into its predictive accuracy. Both models demonstrate the utility of LASSO regression in handling complex datasets for solar irradiation forecasting, emphasizing the importance of specific meteorological factors in predicting solar power output. The flow diagram of the process can be seen in Figure 1.

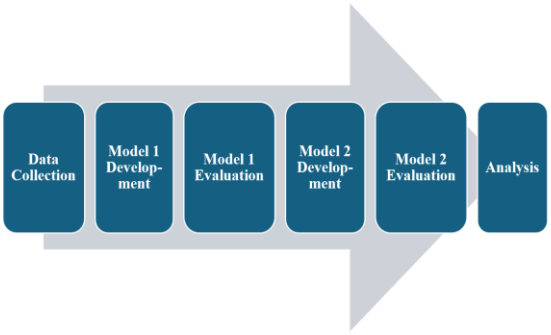

Figure 1: Flow diagram of the approach.

Loss function = OLS loss function

$$+\lambda * \sum_{i=1}^{n} \left| a_i \right|. \quad (1)$$

Equation (1) describes the loss function used in LASSO regression. It consists of two parts: the Ordinary Least Squares (OLS) loss function, which is the sum of the squared differences between the observed and predicted values, and a penalty term. The penalty is applied to the absolute values of the regression coefficients (represented by a_i), summed up across all coefficients (from i=1 to i=n). The term λ is a non-negative regularization parameter that controls the strength of the penalty. By increasing λ, the LASSO method can shrink less important coefficients to zero, effectively performing feature selection. This penalty encourages the model to maintain simplicity and prevent overfitting, leading to more interpretable models.

2.3 Data Source

Data for this work was accessed from NASA's POWER (Prediction of Worldwide Energy Resources) database. Monthly meteorological data for Tomsk from January 2001 to December 2020 was collected and prepared for input in the LASSO models. Tomsk, a city in Siberia, with the coordinates, approximately 56.5010 degrees latitude North and 84.9924 degrees longitude East was chosen as the case study location due to its distinctive climatic and geographical characteristics, which present a unique opportunity to study solar irradiation patterns in a region with significant seasonal variations. The choice of Tomsk adds to the diversity and comprehensiveness of solar irradiation research, particularly in areas with extreme climatic conditions.

2.4 Feature Selection Analysis

In our research we conducted a feature selection analysis to identify the variables that impact how well our predictive model works. We looked at the dataset sourced from NASA POWER, with 260 data points across 9 variables as outlined in Table 1. This thorough examination was crucial in capturing the essence of our dataset, ensuring that only important variables with significant predictive power were included in the model. Our goal with this feature selection process was to improve the accuracy and clarity of the model setting a groundwork, for further analysis.

3 RESULTS AND DISCUSSION

3.1 Model 1

In Model 1, we analysed allsky_sfc_uvb vs other independent variables. The model shows in Figures 2 and 3 that qv2m was the most influential variable, followed by y2m.

Figures 4 and 5 are plots for visualisation of the results. Normal Q-Q (Quantile-Quantile) plots are graphical tools used to assess if a dataset follows a particular distribution, usually a normal distribution. If the points in the plot fall approximately along a straight line, it suggests the data are normally distributed. Series residual plots are used in regression analysis to visualize the residuals (differences between observed and predicted values) across the data series. These plots help to identify any patterns in the residuals, suggesting issues with the model, such as non-linearity, heteroscedasticity, or outliers.

The metrics used to evaluate the performance of a LASSO regression model are usually the root mean squared error (RMSE), the mean squared error (MSE) and the r squared (r^2). Here's an interpretation of each metric:

Delving into the performance of our Lasso regression model, we observe that it boasts a promising precision in its predictions. An RMSE of 0.0353 suggests that our model's forecasts are, on

average, only a small fraction off from the actual figures, which is quite commendable.

Further affirming its accuracy, the MSE stands at a minimal 0.00125, reflecting minor average errors in the predictions squared. Moreover, an R-squared value of 0.843 is noteworthy, indicating that the model can explain over 84% of the variability in the dependent variable—pointing to a robust model that captures the essence of the data well. These indicators collectively point towards a model that performs reliably and can be trusted for its predictive insights even in extreme weather climates like Tomsk.

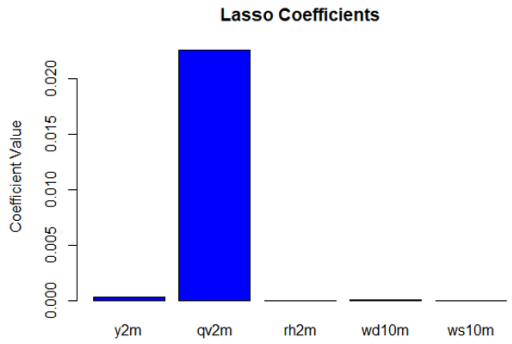

Figure 2: LASSO analysis from model 1.

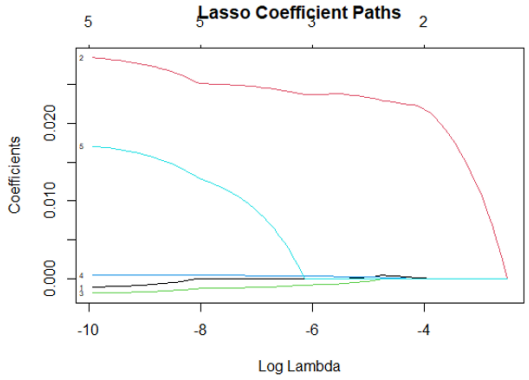

Figure 3: LASSO coefficient paths for model 1.

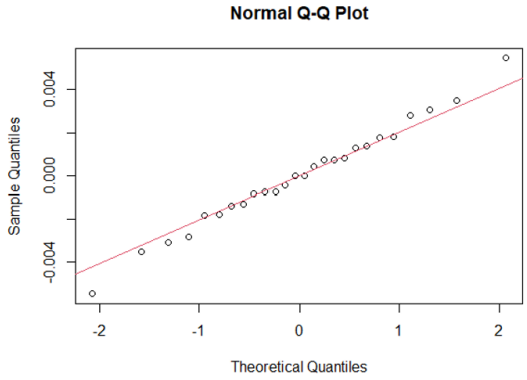

Figure 4: Normal Q-Q plot of model 1.

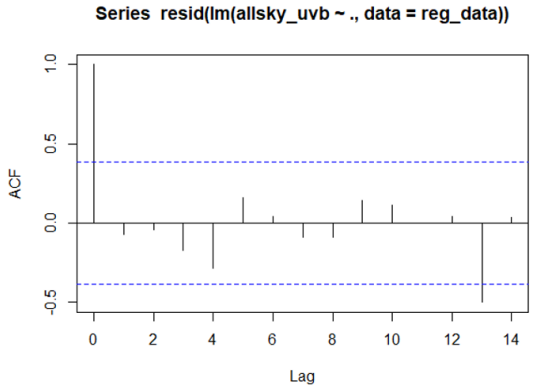

Figure 5: Series residual plot of model 1.

3.2 Model 2

In model 2 we excluded the specific humidity at 2 meters (qv2m) from the analysis and instead assessed the impact of other variables such as the temperature in degrees Celsius at 2 meters, wind direction at 10 meters (wd10m), and wind speed at 10 meters (ws10m) on solar irradiation forecasts. Computations were done for allsky_sfc_uvb vs other independent variables except qv2m. The model shows that y2m was the most influential variable, followed by ws10m and wd10m. The effectiveness of model 2 is evaluated using RMSE, MSE, and R-squared metrics, which provide insights into its predictive accuracy.

Table 1: Characteristics and description of data set.

Short name	Unit	Type
y2m	merra-2 temperature at 2 meters (c)	Independent
qv2m	merra-2 specific humidity at 2 meters (g/kg)	Independent
rh2m	merra-2 relative humidity at 2 meters (%)	Independent
wd10m	merra-2 wind direction at 10 meters (degrees)	Independent
ws10m	merra-2 wind speed at 10 meters (m/s)	Independent
allsky_sfc_uvb	ceres syn1deg all sky surface uvb irradiance (w/m^2)	Dependent
clrsky_sfc_par_tot	ceres syn1deg clear sky surface par total (w/m^2)	Dependent

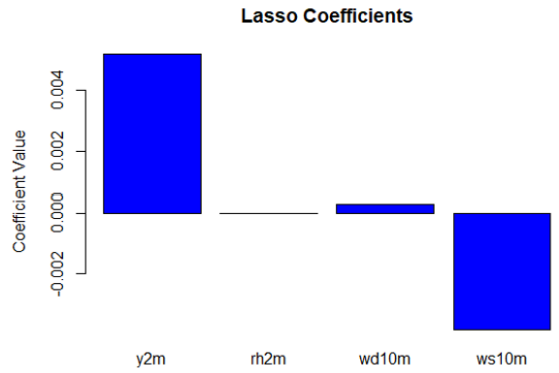

Figure 6: LASSO analysis from model 2.

In Figure 6, using LASSO regression we highlight the most important meteorological factors required for predicting power output. We can see the coefficients assigned to four variables: y2m, rh2m, wd10m and ws10m. The y axis displays the coefficient values selected by the LASSO model. This plot illustrates how each variable influences the prediction outcome. Specifically, the positive impact of temperature (y2m) on the prediction is evident. However, relative humidity at 2 meters (rh2m) does not strongly predict outcomes in this model. The small coefficient for wind direction at 10 meters (wd10m) suggests a minor relationship with the predicted value. Conversely a substantial negative coefficient for wind speed at 10 meters (ws10m) indicates an inverse relationship with the target variable.

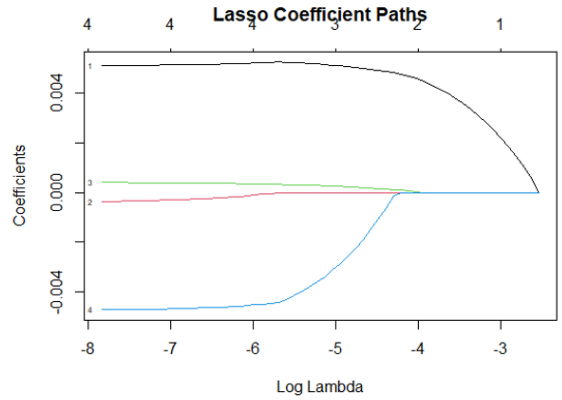

Figure 7: LASSO coefficient paths for model 2.

Figure 7 illustrates the LASSO coefficient paths for model 2, providing insight into the feature selection process as regularization is applied. Here, the x-axis, which represents the logarithm of lambda (the regularization parameter), shows the trajectory of each feature's coefficient as the model complexity is varied. LASSO simultaneously performs feature selection and regularization to enhance the prediction model's robustness, particularly under conditions of multicollinearity or when a parsimonious model is desired [15].

The plot shows several paths corresponding to different coefficients, with their values converging towards zero as the log lambda increases. This convergence is indicative of the LASSO method's ability to shrink less important coefficients down to zero, effectively eliminating them from the model. Notably, the coefficient paths can also inform us about the relative importance and stability of the features across different regularization strengths; features with paths that quickly converge to zero are less robust, whereas those that remain non-zero at higher lambda values are more influential.

For a renewable energy application, such as solar irradiation prediction, the interpretability of this model is crucial. Figure 7 suggests that some features have a more consistent influence on the model's predictions, maintaining a non-zero coefficient across a range of lambda values. In contrast, features whose coefficients drop to zero more quickly are deemed less relevant. The key advantage of this approach is that it reduces the model's complexity and potential overfitting, leading to a model that is both interpretable and generalizable.

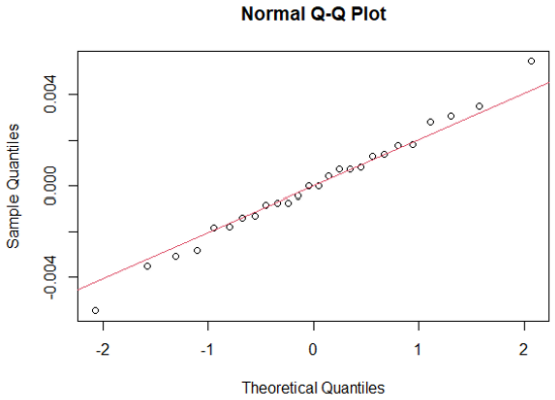

Figure 8: Normal Q-Q plot of model 2.

Figure 8 shows the Normal Q-Q plot of model 1 which, as mentioned earlier, is a tool used to check the normality of residuals, in regression analysis. This plot compares the quantiles to a standard normal distribution with the theoretical quantiles of the normal distribution on one axis and the sample quantiles of residuals on another. In a situation where

residuals follow a distribution the data points should align closely with a reference line. The alignment of points in Figure 8, along the line indicates that model 2's residuals adhere closely to a distribution suggesting that the model is well tuned, and its predictions are trustworthy.

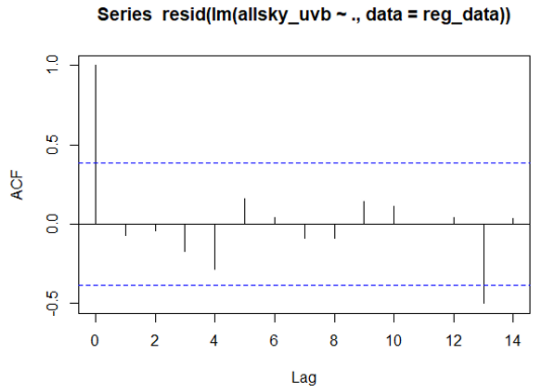

Figure 9: Series residual plot of model 2.

Figure 9 is the Autocorrelation Function (ACF) plot, for the residuals from model 2. It illustrates how the series correlates with itself at time lags. In a scenario where the model fits well, we would anticipate the autocorrelations at all lag points to fall within the confidence bounds (indicated by the dashed lines). This indicates no autocorrelation. It also shows that the model's residuals are random. The randomness signifies that the model has effectively captured all information from the data.

Our LASSO regression model's predictive accuracy is quantified through several statistical measures. The RMSE of 0.0386 highlights that our predictions deviate from the actual values by this small margin, indicating a tightly fitted model. The MSE, at 0.00149, reaffirms this, showing a minimal average of squared errors.

Impressively, the model accounts for approximately 81.3% of the variability in the target variable, as suggested by an R-squared of 0.813. Such a high R-squared value reflects the model's robustness in capturing the underlying data patterns.

To round up, these metrics underscore a strong predictive capability, suggesting the model is well-tuned to the nuances of our data.

4 CONCLUSIONS

The aim of the study was achieved through a systematic approach involving LASSO regression to forecast solar irradiation in Tomsk, which is a method particularly well-suited to handle the challenges posed by climatic variability. The study created two distinct models to interpret the complex relationships between meteorological factors and solar power output. Model 1 included specific humidity at 2 meters as an independent variable, while Model 2 excluded it, thus allowing for the analysis of other significant meteorological factors.

The effectiveness of these models was demonstrated by robust statistical metrics: Model 1 showed an R-squared value of 0.843, indicating that it could explain over 84% of the variability in the dependent variable, while Model 2 had an R-squared value of 0.813, accounting for approximately 81.3% of the variability in the target variable. These high R-squared values signify that both models have strong predictive capabilities and can reliably capture the underlying data patterns.

Furthermore, the use of the NASA POWER database provided a comprehensive set of meteorological variables, ensuring that the models had a solid data foundation to work from. The choice of Tomsk, with its distinctive climate, offered a unique case for examining solar irradiation patterns, contributing to the literature on renewable energy forecasting and the operational practices in the field.

In essence, the study succeeded in contributing to predictive modeling by developing interpretable models that effectively address climatic variability and demonstrate strong predictive performance, thus advancing the field of solar power output forecasting.

To further enhance the research presented, one could explore the integration of real-time data feeds to improve model responsiveness, use Explainable Artificial Intelligence (XAI) tools to broaden the interpretability spectrum, and test the scalability of the proposed framework across different regions and forms of renewable energy.

REFERENCES

[1] "Transforming our world: the 2030 Agenda for Sustainable Development," UN Doc. A/RES/70/1, Sept. 25, 2015.

[2] S. Krumdieck, Transition Engineering: Building a Sustainable Future. Taylor & Francis, 2020.

[3] J. Stephenson, B. Barton, G. Carrington, A. Doering, R. Ford, D. Hopkins et al., "The energy cultures framework: exploring the role of norms, practices and material culture in shaping energy behaviour in New Zealand," Energy Research & Social Science, vol. 7, pp. 117-123, 2015, [Online]. Available: https://doi.org/10.1016/j.erss.2015.03.005.

[4] B. Sovacool, M. Burke, L. Baker, C. Kotikalapudi, and H. Wlokas, "New frontiers and conceptual frameworks for energy justice," Energy Policy, vol. 105, pp. 677-691, 2017, [Online]. Available: https://doi.org/10.1016/j.enpol.2017.03.005.

[5] D. McCollum, W. Zhou, C. Bertram, H. Boer, V. Bosetti, S. Busch et al., "Energy investment needs for fulfilling the Paris agreement and achieving the sustainable development goals," Nature Energy, vol. 3, no. 7, pp. 589-599, 2018, [Online]. Available: https://doi.org/10.1038/s41560-018-0179-z.

[6] N. Tang, S. Mao, W. Yu, and R. Nelms, "Solar power generation forecasting with a lasso-based approach," IEEE Internet of Things Journal, vol. 5, no. 2, pp. 1090-1099, 2018, [Online]. Available: https://doi.org/10.1109/jiot.2018.2812155.

[7] S. Pasari and A. Shah, "Time series auto-regressive integrated moving average model for renewable energy forecasting," Sustainable Production, Life Cycle Engineering and Management, pp. 71-77, 2020, [Online]. Available: https://doi.org/10.1007/978-3-030-44248-4_7.

[8] C. Voyant, G. Notton, S. Kalogirou, M. Nivet, C. Paoli, F. Motte et al., "Machine learning methods for solar radiation forecasting: a review," Renewable Energy, vol. 105, pp. 569-582, 2017, [Online]. Available: https://doi.org/10.1016/j.renene.2016.12.095.

[9] F. Niu and Z. O'Neill, "Recurrent neural network based deep learning for solar radiation prediction," Building Simulation Conference Proceedings, 2017, [Online]. Available: https://doi.org/10.26868/25222708.2017.507.

[10] M. Sengupta, Y. Xie, A. Lopez, A. Habte, G. Maclaurin, and J. Shelby, "The national solar radiation data base (nsrdb)," Renewable and Sustainable Energy Reviews, vol. 89, pp. 51-60, 2018, [Online]. Available: https://doi.org/10.1016/j.rser.2018.03.003.

[11] V. Sharma and S. Chandel, "Performance and degradation analysis for long term reliability of solar photovoltaic systems: a review," Renewable and Sustainable Energy Reviews, vol. 27, pp. 753-767, 2013, [Online]. Available: https://doi.org/10.1016/j.rser.2013.07.046.

[12] S. Sobri, S. Koohi-Kamali, and N. Rahim, "Solar photovoltaic generation forecasting methods: a review," Energy Conversion and Management, vol. 156, pp. 459-497, 2018, [Online]. Available: https://doi.org/10.1016/j.enconman.2017.11.019.

[13] R. Ahmed, V. Sreeram, Y. Mishra, and M. Arif, "A review and evaluation of the state-of-the-art in PV solar power forecasting: techniques and optimization," Renewable and Sustainable Energy Reviews, vol. 124, p. 109792, 2020, [Online]. Available: https://doi.org/10.1016/j.rser.2020.109792.

[14] C. Yen, H. Hsieh, K. Su, M. Yu, and J. Leu, "Solar power prediction via support vector machine and random forest," E3S Web of Conferences, vol. 69, p. 01004, 2018, [Online]. Available: https://doi.org/10.1051/e3sconf/20186901004.

[15] Y. Kim and J. Kim, "Gradient lasso for feature selection," Twenty-First International Conference on Machine Learning - ICML '04, 2004, [Online]. Available: https://doi.org/10.1145/1015330.1015364.