Proceedings of the 6th ITB International Graduate School Conference
Multidisciplinary Collaboration for Sustainable Energy: Science, Technology, Policy, and Society

Prediction of Hydropower Plant Electricity Production Dependence on Weather Conditions Using Machine Learning Approach

Dennis Hasnan Zulfialda [1], Hakim Luthfi Malasan [2]

[1] Program Studi S2, Sains Komputasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung, PT PLN (Persero) Jakarta, Indonesia
[2] KK Astronomi dan Program Studi Sains Komputasi, Fakultas Matematika dan Ilmu Pengetahuan Alam, Institut Teknologi Bandung
Bandung, Indonesia
Email: 20923014@mahasiswa.itb.ac.id

Abstract. Optimizing the operation of hydropower plants within the PLN Sulawesi Generation Unit, this study proposes a data-driven approach to analyze electricity production by incorporating weather-related variables. Using historical data from January 2014 to December 2023, indicators relevant to PLTA electricity output were extracted using machine learning algorithms. The integration of electricity production data, dam-related variables, temperature, and rainfall allows for accurate forecasting of electricity generation as the model's output. The processed data were obtained from PLTA in Sulawesi, as well as weather data from websites of Accuweather and NOAA. The results demonstrate the predictive performance of the proposed approach through model validation and comparison with similar methods. The machine learning methods used in this study include SARIMAX, Random Forest Regressor, Support Vector Regression, and Extreme Gradient Boosting XGBoost. These models utilized a combination of electricity production records, dam data, meteorological information, and four ENSO indicators. The findings show that merging diverse data sources when significantly correlated with the target variable can improve prediction accuracy, with one algorithm emerging as the best performer. Every model was then applied to forecast electricity production on a new, unseen dataset. The results of this study indicate that machine learning is effective in predicting hydropower electricity output and can serve as a strategic consideration for PLN's management in planning and operating PLTA which was interconnected in a big electricity system. This structured approach aligns with organizational goals and supports informed decision-making in hydropower plant operations.

Keywords: *electricity production, hydropower plant, machine learning, temperature, weather, PLN*

1 Introduction

In this era of rapid technological advancements, the power sector is also required to keep up with the fast-paced developments. Moreover, the shift in electricity generation is now focusing on the effectiveness of operating Renewable Energy (RE) plants, with this effectiveness becoming a key determinant of organizational success. This is particularly true for large companies like PLN, where the operation of hydropower plants (PLTAs) becomes a primary focus. The adoption of advanced technologies such as machine learning is highly relevant in this context. The implementation of this technology is not limited to specific industries but has proven successful across various sectors in facilitating faster and more accurate decision-making.

Hydropower remains one of the most reliable sources of renewable energy, contributing significantly to global electricity generation. However, its performance is inherently sensitive to weather conditions, particularly precipitation and temperature. These environmental factors directly affect river discharge levels, which are critical for hydropower plant operations. A study conducted in Nepal highlighted that seasonal fluctuations in rainfall and temperature significantly impacted hydropower generation, underlining the sector's vulnerability to climatic variability [1].

Climate-related disruptions to hydropower are not confined to tropical regions. In Switzerland, for example, studies have indicated that hot and dry weather can significantly impact the performance of hydropower facilities—particularly those based on run-of-the-river systems with minimal storage capacity [2]. On a broader scale, global climate oscillations such as the El Niño–Southern Oscillation (ENSO) have been recognized as major factors contributing to fluctuations in hydropower generation worldwide [3]. These insights underline the necessity of developing forward-looking and adaptive planning methods to strengthen the resilience of hydropower infrastructure in the face of climate variability.

The Southern Oscillation Index (SOI) is one of the principal tools used to observe and analyze ENSO behavior. It is derived from comparing sea-level atmospheric pressure between two key locations: Tahiti and Darwin, Australia. This comparison reveals broad pressure variations across the tropical Pacific. When SOI values are negative, it generally signals an El Niño episode—characterized by lower-than-normal pressure in Tahiti and elevated pressure in Darwin. Conversely, positive SOI values typically point to La Niña conditions [4].

El Niño and La Niña episodes are further characterized by variations in sea surface temperatures (SSTs) in the equatorial Pacific, especially in the Niño 3.4

region. These events are defined by the Oceanic Niño Index (ONI), which considers a five-period moving average of SST anomalies. An anomaly exceeding +0.5°C over five consecutive 3-month intervals is categorized as El Niño, while a drop below -0.5°C over the same span marks a La Niña event [4].

The Niño 3.4 region is widely used to classify El Niño strength due to its strategic location along the equatorial Pacific, where fluctuations in SSTs strongly influence atmospheric convection patterns. Typically, a temperature increase of just +0.5°C is enough to trigger deep convection from March to June. However, during the rest of the year, larger anomalies are necessary—sometimes reaching +1.5°C during November to January—to maintain strong convection patterns [4].

While the Niño 3.4 region is standard for identifying La Niña events, some argue that the Niño 4 region may provide better accuracy since its baseline SSTs are usually at or above the deep convection threshold year-round. Thus, a negative anomaly of -0.5°C in Niño 4 can effectively disrupt convection, causing it to shift westward across the Pacific [4].

To complement these observations, Outgoing Longwave Radiation (OLR) data are also utilized to understand atmospheric convection and cloud dynamics. These OLR values are captured by NOAA's AVHRR instruments aboard orbiting satellites, with a focus on the equatorial zone between 160°E and 160°W. During El Niño, there is a notable drop in OLR values, indicating enhanced cloudiness and rainfall. On the other hand, higher OLR values reflect a decrease in cloud activity, which is commonly associated with La Niña patterns [4].

Despite the increasing integration of renewable energy into national grids, there remains a significant gap in accurately forecasting hydropower electricity generation, particularly in regions like Sulawesi where climate variability and hydrological conditions are complex. Existing studies have predominantly focused on global or national-scale hydropower prediction, often overlooking localized environmental and operational factors that influence power output. Furthermore, few studies have combined meteorological variables with dam-specific and ENSO-related indicators in a unified machine learning framework. This study addresses that gap by proposing a data-driven approach tailored to the operational context of PLTA in Sulawesi, Indonesia. The main objective is to evaluate and compare the performance of several machine learning algorithms—SARIMAX, SVR, Random Forest, and XGBoost—in forecasting electricity production using an integrated dataset. By doing so, the study contributes a practical framework for PLN and similar utilities to enhance short-term planning and optimize hydropower operations through intelligent forecasting.

2 Methodology

2.1 Overview of Machine Learning Approaches

This study utilizes supervised machine learning algorithms to predict electricity production in a hydropower plant. Supervised learning involves training models on labeled historical data, where the goal is to learn the mapping between input features (e.g., weather and dam data) and the output target (electricity production). This approach is suitable for regression tasks and is commonly used in predictive modeling across various domains. The algorithms applied in this study are: Seasonal AutoRegressive Integrated Moving Average with eXogenous variables (SARIMAX), Random Forest Regressor (RFR), Support Vector Regression (SVR), and Extreme Gradient Boosting (XGBoost).

2.2 SARIMAX

SARIMAX is a time series forecasting model that extends the ARIMA model by incorporating both seasonal components and exogenous variables. This enables the model to account for external influences like rainfall or temperature, which can significantly affect electricity generation. SARIMAX combines autoregressive (AR), differencing (I), and moving average (MA) elements with seasonal terms and additional regressors [6].

The relevance of SARIMAX for forecasting with exogenous features has been demonstrated in prior research. For instance, Farikh et al. (2024) achieved superior accuracy using SARIMAX for sales forecasting with weather variable compared to the Vector Auto Regression (VAR) model, as indicated by a lower RMSE (8.966 vs. 24.171) [6].

In this study, SARIMAX parameters were selected using grid search to optimize hyperparameters such as p, d, q, P, D, Q, and s.

2.3 Random Forest Regressor (RFR)

Random Forest Regressor is an ensemble learning algorithm that builds multiple decision trees using random subsets of the training data and features. It then aggregates their predictions to improve accuracy and prevent overfitting [7]. The model is known for its robustness, interpretability, and ability to rank feature importance. In this study, the RFR was tuned using grid search to optimize hyperparameters such as n_estimators, max_depth, and min_samples_split.

In previous work, Random Forest models have been successfully applied to predict energy demand and production. However, studies such as Gökçe et al. (2022) found that while RFR was effective, it was outperformed by XGBoost in terms of RMSE [8].

2.4 Support Vector Regression (SVR)

Support Vector Regression (SVR) extends the Support Vector Machine (SVM) framework to regression tasks. It attempts to fit a function within a specified margin while minimizing model complexity [11]. SVR is particularly effective when paired with kernel functions such as the Radial Basis Function (RBF), which allows it to model non-linear relationships in data.

In this study, SVR was used with an RBF kernel, and its performance was fine-tuned using C, gamma, and epsilon through grid search. Prior studies have highlighted the effectiveness of SVR in time series prediction. Abba et al. (2021) compared standalone SVR with hybrid SVR models using optimization algorithms such as Harris Hawks Optimization (HHO), with SVR-HHO achieving an R^2 of 0.9951 for electricity load demand prediction [12].

2.5 Extreme Gradient Boosting (XGBoost)

XGBoost is a high-performance implementation of gradient boosted trees designed for efficiency and accuracy. It builds additive regression models in a forward stage-wise manner and allows for regularization to reduce overfitting. The model has gained popularity for its consistent success in machine learning competitions [9].

In this study, XGBoost was applied to model electricity output, and hyperparameters such as learning_rate, max_depth, and n_estimators were optimized. The study by Gökçe et al. (2022) found that XGBoost outperformed Random Forest and linear models in forecasting electricity consumption, achieving the lowest RMSE [8]. Similarly, Li et al. (2024) demonstrated that XGBoost, especially when combined with optimization techniques like the Sparrow Search Algorithm (SSA), resulted in improved accuracy for energy consumption forecasting [10].

Here are the parameter used on each model based on grid search that has been implemented shown below on Table 1.

Table 1 Best parameter for each model

No	Model	Parameter
1	SARIMAX	p=1, d=0, q=1, P=1, D=1, Q=1, s=12
2	RFR	max depth= None, min samples split =5, n estimators= 150
3	SVR	C=10, epsilon=0,01, kernel= linear
4	XGBoost	learning rate= 0,1, max depth= 3, n estimators= 100

2.6 Research Methodology Steps

The historical hydrology data used in this study comes from a hydropower plant (PLTA) located in a city in South Sulawesi. Historical weather data was obtained from the Accuweather website, and ENSO data was sourced from the NOAA website. The overall process flow can be seen in Figure 1.

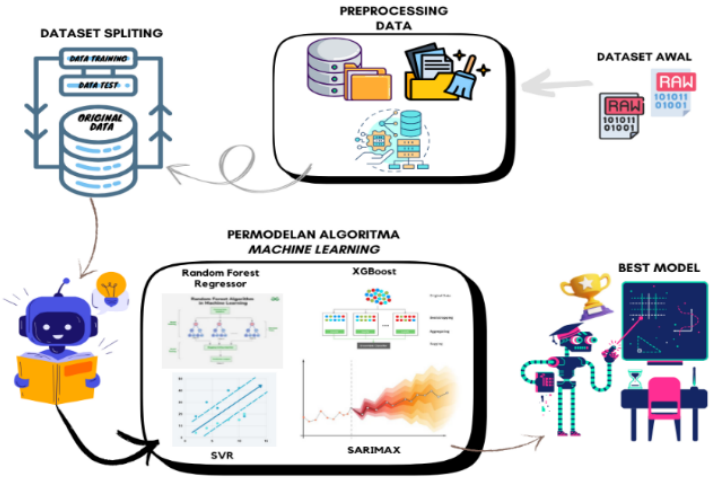

Figure 1 Research Process Flowchart

For the simulation of the four machine learning algorithms in this study, the Python programming language was used via the Jupyter Notebook application.

Based on Figure 1, a detailed explanation of each testing step is described as follows:

Preparing data from the hydropower plant (PLTA), which was originally in daily format and then averaged to obtain monthly values. The daily data spans from January 2014 to December 2023. Although the original dataset contained 3,652 daily observations, the data was aggregated into 120 monthly samples. This decision is justified by both methodological and practical considerations. Daily data often exhibit high variability due to short-term fluctuations such as extreme weather, sensor errors, or local anomalies. Aggregating to monthly values helps to reduce this noise, providing a more stable and representative signal for machine learning models. Moreover, granular data may lead to overfitting, particularly when used with traditional machine learning algorithms that are not specifically designed to handle high-frequency time series.

Monthly data, on the other hand, align better with the long-term analytical goals, such as identifying seasonal trends or evaluating the impact of climate variables on energy production. This format is also more suitable for strategic and policy-level decision-making due to its simplicity and interpretability. From a computational perspective, using fewer samples enables faster model training and reduces the complexity of parameter tuning. Thus, the choice to use 120 monthly samples is both analytically sound and practically efficient.

The selected variables include load, electricity production, elevation, discharge, spillway, and inflow. The data is shown in Table 2.

Table 2 Hydrology data overview

	Beban	Discharge	Inflow	Limpasan	Elevasi	Produksi Listrik
0	111.62	37.16	67.33	29.66	615.44	83724.80
1	78.25	27.10	28.31	1.73	615.31	52353.10
2	88.26	33.44	35.12	5.30	615.29	66592.80
3	110.17	35.99	36.85	9.60	615.19	79574.20
4	120.58	38.91	67.33	29.66	615.44	89737.70
...	...	...	...	...	...	...

Weather-related data were then prepared, including values for SLP Anomaly, OLR Anomaly, SST Region 3.4 Anomaly, and ONI Anomaly, as shown in Table 3.

Table 3 Weather data overview

Act. Low	Act. Avg	Norm High	Norm Low	Norm Avg.	Norm Dept.	Precip . Amt	Cool Deg Day	SLP Anom	OLR Anom	SST 3.4 Anom	ONI Anom
23.71	27.13	29.00	23.00	26.00	1.13	3.24	9.13	2.40	3.80	-0.49	-0.42
23.39	27.71	29.00	23.04	26.00	1.71	1.49	9.71	0.10	3.30	-0.85	-0.46
23.42	28.10	29.00	23.00	26.52	1.58	0.82	10.10	-1.50	-17.30	-0.34	-0.27
23.80	28.33	29.53	23.00	27.00	1.33	0.85	10.33	1.30	-15.00	0.18	0.04
23.97	28.87	30.03	23.00	27.00	1.87	0.70	10.87	0.90	-2.00	0.45	0.21
…	…	…	…	…	…	…	…	…	…	…	…

These two datasets were combined into a single dataset, ensuring that all data was numerical and there were no non-null categorical data. Correlation matrix was then calculated between variables to measure the strength and direction of linear relationships. For weather data, a cross-correlation was performed with a lag of 3 months to detect the delayed effect of weather on electricity production, which is the target variable of this study.As shown in Figure 2, the correlation matrix among variables, especially weather data, reveals a moderate to fairly strong positive correlation ranging from 0.26 to 0.69 with electricity production.

This indicates that weather conditions, even with a 3-month lag, are related to current electricity production. The highest correlation within the temperature category and electricity production was observed in the variable Act. Low (actual lowest temperature) with a value of 0.69. The correlation between rainfall and electricity production was 0.41, and for ENSO indicators, the correlation ranged from 0.39 to 0.40. Based on these results, weather data variables such as temperature, rainfall, and ENSO indicators can be used as input features in the four machine learning algorithms tested in this study.

The dataset was then preprocessed using the ‘Min-Max Scaler’ library, as the range of values among the variables in the dataset varied significantly. This step was intended to bring all features into the same scale range to enhance the performance of the four machine learning algorithms used. The next step was splitting the dataset into 70% training data and 30% testing data.

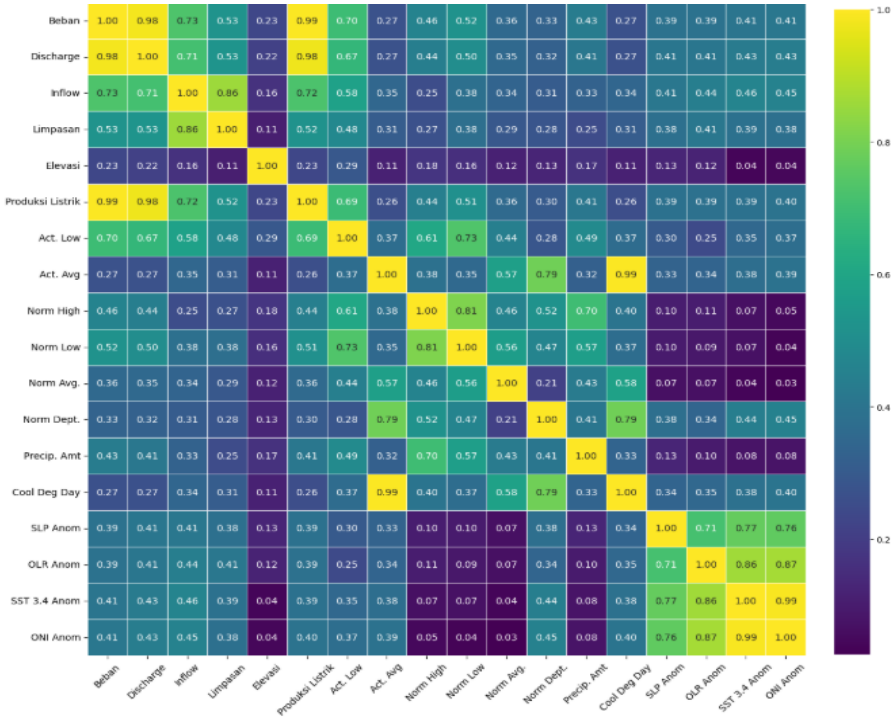

Figure 2 Matrix correlation of dataset

After that, 4 machine learning models were developed using grid search to tune parameters in order to achieve the best model performance, evaluated using accuracy metrics such as RMSE, MAE, MAPE, and R^2. The best-performing model was then tested by making predictions on a data outside the training dataset, which were then compared to actual data with the same variables for the first semester of 2024.

3 Result and Discussion

Based on the testing results, as shown in Table 4 below:

Table 4 Evaluation metric of 4 models

No	Algorithm	Model Performance			
		MSE	RMSE	R^2	MAPE
1	SARIMAX	0.0002	0.0153	0.9919	1.67%
2	RFR	0.0013	0.0367	0.9639	5.08%
3	SVR	0.0006	0.0253	0.9828	2,48%
4	XGBoost	0.0022	0.0472	0.9403	5.83%

From the table above, it can be concluded that the SARIMAX model with parameters p=1, d=0, q=1, P=1, D=1, Q=1, and s=12 is the best model for this dataset, as it shows the most optimal performance based on evaluation metrics including MSE, RMSE, R^2, and MAPE. The other models (RFR, SVR, and XGBoost) also demonstrated reasonably good performance, though not as strong as SARIMAX during the model building.

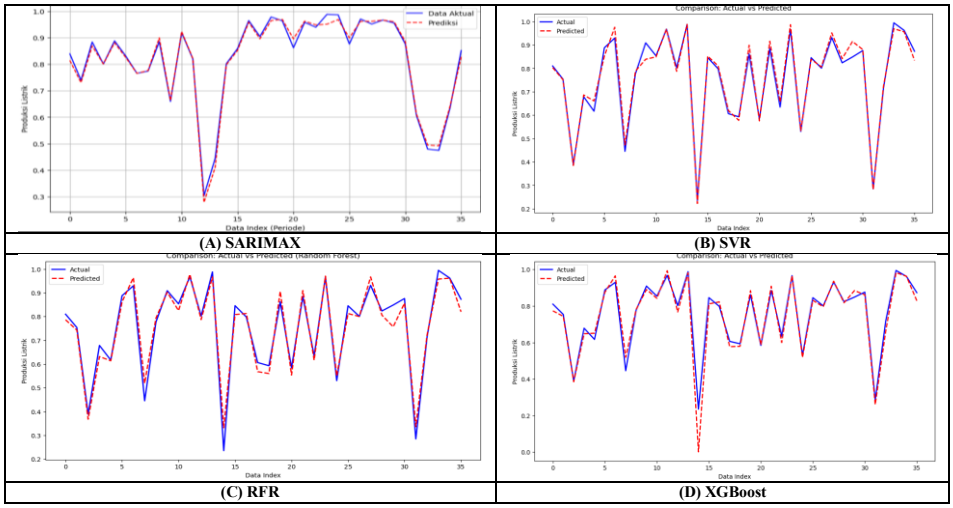

Figure 3 Prediction Model

However, when testing each model using their best-tuned parameters for forecasting first semester 2024 data, the RFR model achieved the overall best performance, with the lowest MAPE of 4.10%, and the highest R^2 value of 0.9686. This indicates a very strong and accurate predictive capability. XGBoost model competitively has a high R^2 of 0.9626 and MAPE of 4.29%. a little bit higher than RFR model.

On the other hand, SARIMAX failed to demonstrate good performance at both scales, as indicated by a negative R^2 value and a MAPE exceeding 28.14%, suggesting that the model was unable to capture the data patterns in the new dataset effectively, despite being the best-performing model during the initial model development phase, as shown in Table 5 and Figure 4. below.

Table 5 Evaluation metrics of 4 models in forecasting first semester 2024 data

No	Algorithm	Model Performance			
		MSE	RMSE	R^2	MAPE
1	SARIMAX	0.2015	0.4489	-0.664	28.14%
2	RFR	0.0017	0.0415	0.9686	4.10%
3	SVR	0.0216	0.1471	0.6057	21.12%
4	XGBoost	0.002	0.0451	0.9629	4.29%

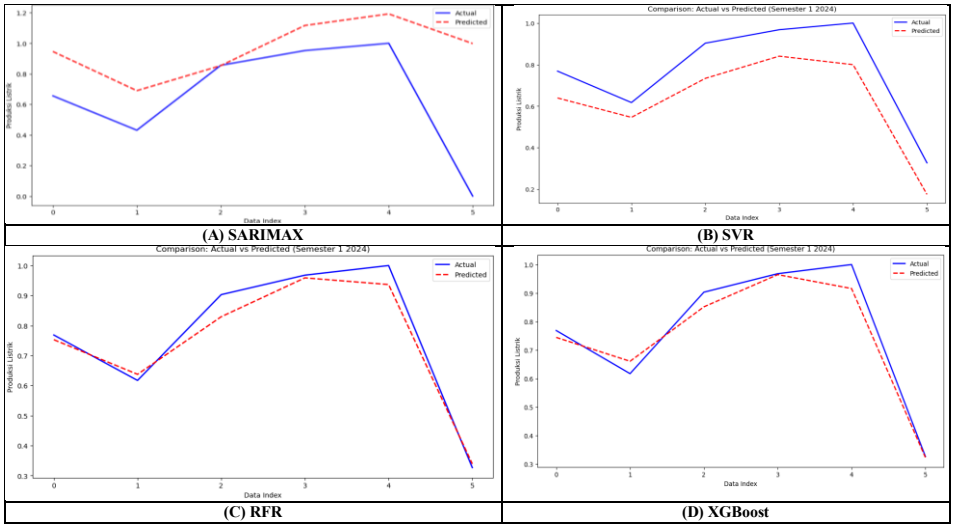

Figure 4 2024 First Semester Forecast

A drastic and suspicious drop in SARIMAX performance was observed, with the R^2 value declining from 0.9919 during internal testing to -0.664 on out-of-sample data. This significant degradation strongly suggests a critical issue in the model's ability to generalize beyond the training distribution. A negative R^2 indicates that the model performs worse than a naive mean predictor, typically resulting when the residual sum of squares exceeds the total sum of squares.

Several factors may explain this anomaly. First, SARIMAX is highly sensitive to data structure and may have overfitted to short-term seasonal or trend patterns specific to the training set. Second, the out-of-sample data likely exhibit different temporal dynamics or non-stationarities that the model was not exposed to during training. Third, SARIMAX assumes a degree of stationarity, and any structural shifts or noise in the test data can severely impact its accuracy.

Furthermore, the internal evaluation (Table 1) may have used a validation set that was too similar to the training data, thereby inflating performance metrics. Notably, while other models such as Random Forest, SVR, and XGBoost also experienced declines in performance on out-of-sample data, the magnitude was significantly smaller. Random Forest and XGBoost maintained high R^2 values (0.9686 and 0.9629, respectively), while SVR, although affected more strongly ($R^2 = 0.6057$, MAPE = 21.12%), still outperformed SARIMAX.

These results underscore the importance of validating time series models on truly unseen data and highlight the advantages of using more adaptive or hybrid machine learning approaches in forecasting tasks involving complex, evolving data.

4 Conclusion and Suggestion

While SARIMAX excelled during the fitting process, XGBoost is more recommended for future predictions, as it proved to be superior in forecasting unseen data (first semester 2024). This is a crucial aspect for model deployment in actual prediction systems.

Based on the analysis, it can be concluded that weather variables such as temperature and humidity, rainfall, and ENSO indicators play a significant role in influencing electricity production, particularly for renewable energy-based power plants like hydropower plant. These three groups of variables show correlated patterns, either directly or with time lags, indicating that climate conditions and global weather anomalies can have tangible impacts on energy availability and efficiency. Therefore, these variables are highly relevant and suitable as predictors in future electricity production forecasting models.

For future research, it is recommended to include additional features such as land surface temperature index, wind speed, and others to further improve the model's performance. It would also be beneficial to experiment with model combinations or hybrid approaches to enhance prediction accuracy even further.

References

[1] A. Devkota, A. R. Shrestha, and P. Koirala, "Assessment of climate change impact on hydropower generation: A case study of the Upper Tamakoshi Hydroelectric Project, Nepal," *Heliyon*, vol. 8, no. 12, p. e12529, 2022. [Online].Available:https://www.sciencedirect.com/science/article/pii/S2405844022035289

[2] S. Rössler, M. Zappa, and M. Weingartner, “Hot and dry weather impacts on Swiss hydropower production,” *Environmental Research Letters*, vol. 18, no. 7, 2023. [Online]. Available: https://iopscience.iop.org/article/10.1088/1748-9326/acd8d7

[3] Y. van Vliet *et al.*, “The influence of climate variability and extremes on global hydropower generation: A multi-model assessment,” *Hydrology and Earth System Sciences*, vol. 26, pp. 2431–2450, 2022. [Online]. Available: https://hess.copernicus.org/articles/26/2431/2022/

[4] National Centers for Environmental Information (NCEI), "El Niño Southern Oscillation (ENSO) Monitoring," NCEI Access Monitoring. [Online]. Available: https://www.ncei.noaa.gov/access/monitoring/enso/sst

[5] P. Santoso, H. Abijono, and N. L. Anggreini, “Algoritma Supervised Learning Dan Unsupervised Learning Dalam Pengolahan Data,” *Unira Malang*, vol. 4, no. 2, 2021.

[6] F. Alzami, A. Salam, I. Rizqa, C. Irawan, P. N. Andono, and D. Aqmala, "Demand prediction for food and beverage SMEs using SARIMAX and weather data," *IETIA*, vol. 29, no. 1, Feb. 2024. [Online]. Available: https://doi.org/10.18280/isi.290129

[7] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[8] M. Gökçe and E. Duman, “Performance Comparison of Simple Regression, Random Forest and XGBoost Algorithms for Forecasting Electricity Demand,” in *Proc. IISEC*, 2022, pp. 1–6, doi: 10.1109/IISEC56263.2022.9998213.

[9] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 785–794.

[10] X. Li, Z. Wang, C. Yang, and A. Bozkurt, “An advanced framework for net electricity consumption prediction: Incorporating novel machine learning models and optimization algorithms,” *Energy*, vol. 296, 2024, Art. no. 131259. [Online]. Available: https://doi.org/10.1016/j.energy.2024.131259

[11] I. R. Isnaeni, Sudarmin, and Z. Rais, “Analisis Support Vector Regression (SVR) dengan Kernel Radial Basis Function (RBF) untuk Memprediksi Laju Inflasi di Indonesia,” *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 4, no. 1, pp. 30–38, 2022, doi: 10.35580/variansiunm13.

[12] S. I. Abba *et al.*, “Emerging Harris Hawks Optimization based load demand forecasting and optimal sizing of stand-alone hybrid renewable energy systems – A case study of Kano and Abuja, Nigeria,” *Results in Engineering*, vol. 12, 2021, Art. no. 100260. [Online]. Available: https://doi.org/10.1016/j.rineng.2021.100260

Acknowledgement

We sincerely express our gratitude to PT PLN (Persero) for funding the Master's degree program at the Bandung Institute of Technology through the Distance Learning Program. This research is part of an evaluation of the operational system of one of the hydropower plants in South Sulawesi, which has consistently strived to mitigate the impacts of each El Niño phenomenon.