Bias correction of global climate model using machine learning algorithms to determine meteorological variables in different tropical climates of Indonesia

Juan Nathaniel
School of Information System
Singapore Management University
jnathaniel@smu.edu.sg

Campbell D. Watson
IBM Research
T. J. Watson Research Center, NY, USA
cwatson@us.ibm.com

Abstract

Accurate and localized forecasting of climate variables are important especially in the face of uncertainty imposed by climate change. However, the data used for prediction are either incomplete at the local level or inaccurate because the simulation models do not explicitly consider local contexts and extreme events. This paper, therefore, attempts to bridge this gap by applying tree-based machine learning algorithms to correct biases inherent in simulated, reanalysed climate model against local climate observations in differing tropical climate subsystems of Indonesia. The new observation datasets were compiled from various weather stations and agencies across the country. Our results show that regions of tropical savanna experience greatest bias corrections, followed by the tropical monsoon and tropical forest. Finally, to account for extreme events, we embed regional large-scale climate events into these models. In particular, we incorporate ENSO to account for the residual error of extreme rainfall observations, and have achieved an improved bias-correction of 36.67%.

1 Introduction

Across the tropical region such as Indonesia, an accurate prediction of climate variables is a prerequisite to robust irrigation planning [16] and forest fire decision-making [22], among many others [11][19]. However, Indonesia's heterogeneous climate systems, which can be broadly grouped into the tropical savanna (Aw), tropical monsoon (Am), and tropical forest (Af) [13], imposes additional constraints due to their unique underlying climate patterns. In addition, the archipelago which boasts approximately 1.9 million square kilometers of land area only has less than 200 reliable climate stations [1]. Therefore, prediction of different climate variables with greater localization and spatial coverage are urgently needed.

One solution would be to deploy Internet of Things (IoT) weather sensors across critical areas, including the agricultural regions and areas prone to forest fire or natural hazards [15]. This method, however, is difficult to implement given the sparsity of internet and reliable electrical coverage [9]. On the other end of the solution spectrum, global climate simulation models are often opted for. However, they tend to perform worse on tropical climates, like Indonesia, where meteorological observations are often fluctuating and changing abruptly by the days [21]. Thus, we are proposing a middle-ground approach where local climate context is incorporated into the global simulation output to account for or correct these reported biases. This paper will therefore map ERA5 reanalysed, simulated global output against our 169 stations, 10-year-worth of local climate observations through different tree-based machine learning ensembles, so that given any area within Indonesia of known tropical climate subsystem, we can predict its meteorological variables with increased confidence.

AI for Earth Sciences Workshop at NeurIPS 2020.

Finally, this paper will explore the potential application of regional large-scale climate events in making bias-correction of local extremes more robust. In particular, we will observe whether the incorporation of El Niño/Southern Oscillation (ENSO) measurements will improve the performance of bias-correcting models for extreme rainfall events.

The machine learning models we are considering in this paper are tree-based, including Random Forest (RF), Gradient Boosting Decision Tree (GB), and Extreme Gradient Boosting (XGB) as they have been widely reported to have performed well in predicting fluctuating climate variables as compared to other algorithms including the support vector machine (SVM) and shallow artificial neural network (ANN) [6][7][17].

2 Materials and Methods

2.1 Dataset and analysis

2.1.1 Indonesia's climate observation data

The meteorological data includes daily observations from 169 weather stations located across the Indonesian archipelago. The 10-year data spanning from January 2010 to December 2019 was manually collected, cross-checked, and cleaned from diverse agencies that are governed by the Meteorology, Climatology, and Geophysical Agency of Indonesia (BMKG)[1]. Each station captures a set of standardized climate variables, which can be grouped into four categories: (1) average, maximum, and minimum temperature (Tavg, Tmax, Tmin), (2) average relative humidity (RHavg), (3) average and maximum windspeed (Uavg, Umax), and (4) daily rainfall rate (RR).

In total there are three tropical climate subsystems in the country, namely tropical savanna (Aw), tropical monsoon (Am), and tropical forest (Af) [13], with a corresponding distribution of 17, 37, and 115 stations respectively. Figure 1 highlights the spatial distribution of the stations where our observation datasets are collected.

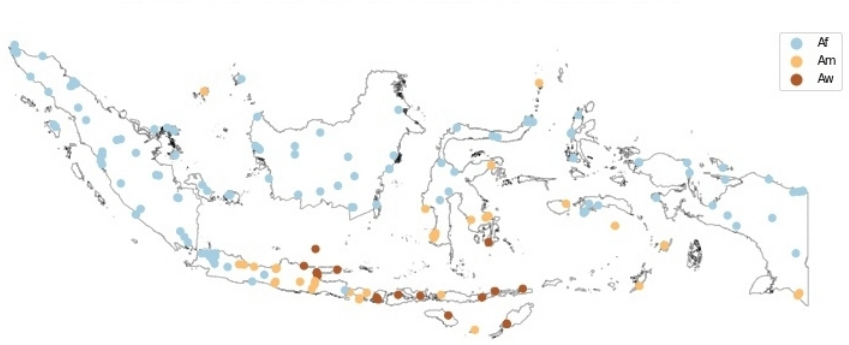

Figure 1: Spatial distribution of meteorological stations in Indonesia according to the tropical climate subtypes.

2.1.2 Copernicus ERA5 reanalysis data

ERA5 is the fifth generation European Center for Medium-Range Weather Forecast (ECMWF) reanalysis for the global climate and weather with data dating back to 1979. ERA5 is produced by adjusting output from simulation model, which in this case is the CY41R2 of ECMWF's Integrated Forecast System, with coarse & proxied meteorological observations. The ERA5 dataset has an hourly temporal resolution, is covering the earth on a 30km grid (0.25-degree resolution), and resolves the atmosphere using 137 pressure levels that spans up to a height of 80km above ground [5]. In order to match the spatio-temporal resolutions of our observation dataset, we aggregate the relevant ERA5 reanalysed data (RHavg is omitted due to its unavailability in ERA5) into daily variables and assign these values to stations that are spatially contained within the corresponding grids.

2.2 Machine learning algorithms for bias-corrected climate variables

Random Forest (RF): The RF model was conceptualized by Breiman (2001) [3] which uses an idea called bagging to aggregate a collection of decision trees with controlled variance.

Gradient Boosting Decision Tree (GB): The GB model was first proposed by Friedman (2002) [8] which uses a collection of regression trees as weak classifiers, similar to RF. One major difference, however, lies in GB's individual binary tree's attempt to incorporate residuals from previous trees, which results in the overall reduction of prediction's bias.

Extreme Gradient Boosting (XGB): The XGB algorithm was put forward by Chen and Guestrin (2016) [4] that combines all the predictions made by weak tree learners to develop a stronger learner through the additive training strategy.

2.3 Experimental setup

Dataset preparation: We are performing an input-output mapping between each of the ERA5 reanlysed variables (Tavg, Tmin, Tmax, Uavg, Umax, and RR) as the input feature and its corresponding observed daily climate data to perform bias correction. Thereafter, for each of the three climate subsystems, training and testing dataset will be randomly selected on station level with a 80:20 split, translating to a train:test split of 14:3 stations for Aw, 30:7 stations for Am, and 92:23 stations for Af.

Metrics: We are using root-mean-squared error (RMSE) to evaluate the performance of different ML algorithms in correcting the biases of simulated and reanalysed ERA5 outputs.

Parameter tuning and model selection: Selecting a fine-tuned ML model requires extensive search for optimal hyperparameters. This iterative process can be resource-intensive and time-consuming, and depending on the granularity of each hyperparameter, the search space can grow exponentially [14]. Hence, to accelerate the speed of parameter tuning and model selection, we will consider a two-step approach: (1) performing a random grid search with coarse variable granularity, followed by a (2) complete search with a finer variable-step [2]. Throughout this process, we will evaluate each of the proposed model's performance with a k-fold cross validation approach (where k=5), using RMSE as the error metrics to be minimized.

3 Results and Discussion

3.1 Baseline evaluation between observed and simulated, reanalysed datasets

The evaluation results between data collected across the 169 Indonesian meteorological stations and ERA5 output are presented in Table 1, aggregated at the three tropical climate subsystems.

Table 1: Baseline evaluation between climate observations and ERA5 simulated, reanlysed output

	RMSE					
Tropical subsystem	Tmin (°C)	Tmax (°C)	Tavg (°C)	Uavg (m/s)	Umax (m/s)	RR (mm/day)
Savanna (Aw)	1.7	3.2	1.2	4.2	6.2	12.39
Monsoon (Am)	2.4	3.9	1.7	4.3	5.6	14.15
Forest (Af)	2.4	3.5	2.2	2.5	3.9	20.15

Generally, there are greater bias and uncertainties surrounding rainfall and wind speed data, especially in regions of tropical forest.

3.2 Evaluation of machine learning models in correcting bias in simulated, reanalysed ERA5 output

The individual model's performance is evaluated against the corresponding observed weather variable across different climate subsystems in Indonesia. Table 2 summarizes the evaluation of the best performing model. For full result, please refer to the table at Appendix A.

Table 2: Statistical evaluation for the best performing algorithms between bias-corrected ERA5 simulated, analysed output and observed climate data

Climate variable	Savanna (Aw) RMSE (Best ML)	Savanna (Aw) Error reduction (%)	Monsoon (Am) RMSE (Best ML)	Monsoon (Am) Error reduction (%)	Forest (Af) RMSE (Best ML)	Forest (Af) Error reduction (%)
RR (mm/day)	10.94 (GB)	13.25	12.57 (RF)	11.17	18.79 (RF)	6.75
Tmax (°C)	1.35 (GB)	57.41	2.23 (RF)	42.67	1.69 (RF)	51.30
Tmin (°C)	1.43 (GB)	15.38	2.18 (RF)	8.02	1.83 (RF)	23.11
Tavg (°C)	0.86 (GB)	26.50	1.72 (RF)	0.58	1.63 (RF)	27.23
Umax (m/s)	2.12 (GB)	61.58	4.20 (RF)	24.87	2.00 (GB)	48.59
Uavg (m/s)	1.16 (RF)	63.29	1.11 (GB)	74.07	2.03 (RF)	17.81

Overall, RF and GB models perform better than XGB for every climate variables. The greatest improvements are observed in Aw where percentage error reductions can go as high as 63.29%, 61.58%, and 57.41% for Uavg, Umax, and Tmax respectively. Across the different climate subsystems, RR boasts the least improvement with error reduction ranging only between 6% in tropical forest and 13% in tropical savanna, primarily due to the extreme fluctuation of observations across months. Next, we will consider whether incorporating an index of a regional climate oscillation, ENSO, could account for these residual biases in extreme rainfall events.

3.3 Evaluation of bias-correction models in extreme events using localized climate knowledge

El Niño–Southern Oscillation (ENSO) is an irregular yet periodic fluctuation of sea surface temperature and winds over the tropical eastern Pacific Ocean, affecting the climate of much of the tropics, especially rainfall [21]. We evaluated the performance of our bias-correcting models when NINO3.4 [18], an index commonly used to explain ENSO, is incorporated as an additional feature to explain for RR. Table 3 summarizes bias-correction performance when compared to observed climate data.

Table 3: Statistical evaluation for the best performing algorithms between bias-corrected ERA5 RR output after embedding ENSO and its corresponding climate observation

Climate variable	Savanna (Aw) RMSE (Best ML)	Savanna (Aw) Error reduction (%)	Monsoon (Am) RMSE (Best ML)	Monsoon (Am) Error reduction (%)	Forest (Af) RMSE (Best ML)	Forest (Af) Error reduction (%)
RR (mm/day)	11.17 (GB)	9.85	11.89 (RF)	15.97	12.76 (RF)	36.67

As observed, Am and Af experience improved bias-correction when we consider ENSO phenomenon, with 15.97% and 36.67% error reductions respectively, as compared to earlier results of 11.17% and 6.75%. Thus, the inclusion of NINA3.4 as an additional feature improves the prediction of extreme rainfall days, reinforcing the notion that regional large-scale climate conditions have a local-scale influence on extreme events.

4 Conclusion

This paper presented a novel application of three tree-based machine learning algorithms, including RF, GB, and XGB, to correct biases in global simulated, reanalysed climate model namely ERA5, by calibrating it to local meteorological variables (Tmax, Tmin, Tavg, Umax, Uavg, RR), and taking into account the differing tropical climate subsystems in Indonesia. Overall, RF and GB produce better results as compared to XGB, especially when correcting for climate variables in regions of tropical savanna. Regions of tropical forest show more modest results primarily due to their greater fluctuations and abrupt daily changes, especially for RR. Nonetheless, by incorporating regional large-scale climate conditions, namely ENSO, we have improved bias-correction especially in extreme rainfall observations.

References

[1] BMKG. (2020, August 21). BMKG | Badan Meteorologi, Klimatologi, dan Geofisika. https://www.bmkg.go.id/

[2] Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The annals of statistics, 24(6), 2350-2383.

[3] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[4] Chen, T., Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In: In Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp. 785–794.

[5] Copernicus Climate Change Service (C3S) (2017): ERA5: Fifth generation of ECMWF atmospheric reanalyses of the global climate . Copernicus Climate Change Service Climate Data Store (CDS), date of access. https://cds.climate.copernicus.eu

[6] Fan, J., Wang, X., Wu, L., Zhou, H., Zhang, F., Yu, X., Lu, X., Xiang, Y. (2018). Comparison of Support Vector Machine and Extreme Gradient Boosting for predicting daily global solar radiation using temperature and precipitation in humid subtropical climates: a case study in China. Energy Convers. Manage. 164, 102–111.

[7] Feng, Y., Cui, N., Gong, D., Zhang, Q., Zhao, L. (2017). Evaluation of random forests and generalized regression neural networks for daily reference evapotranspiration mod- elling. Agric. Water Manage. 193, 163–173.

[8] Friedman, J. H. (2002). Stochastic gradient boosting. Computational statistics & data analysis, 38(4), 367-378.

[9] Handayani, K., Krozer, Y., & Filatova, T. (2017). Trade-offs between electrification and climate change mitigation: An analysis of the Java-Bali power system in Indonesia. Applied energy, 208, 1020-1037.

[10] Hargreaves, G. H., & Samani, Z. A. (1985). Reference crop evapotranspiration from temperature. Applied engineering in agriculture, 1(2), 96-99.

[11] Hidayat, R., Ando, K., Masumoto, Y., & Luo, J. J. (2016). Interannual variability of rainfall over Indonesia: Impacts of ENSO and IOD and their predictability. IOP Confrence Series: Earth and Env. Sci, 31(20136), 012043.

[12] Kazemitabar, J., Amini, A., Bloniarz, A., & Talwalkar, A. S. (2017). Variable importance using decision trees. In Advances in neural information processing systems (pp. 426-435).

[13] Köppen, Wladimir (1884). Translated by Volken, E.; Brönnimann, S. "Die Wärmezonen der Erde, nach der Dauer der heissen, gemässigten und kalten Zeit und nach der Wirkung der Wärme auf die organische Welt betrachtet" [The thermal zones of the earth according to the duration of hot, moderate and cold periods and to the impact of heat on the organic world)]. Meteorologische Zeitschrift (published 2011). 20 (3): 351–360

[14] Koch, P., Wujek, B., Golovidov, O., & Gardner, S. (2017). Automated hyperparameter tuning for effective machine learning. In Proceedings of the SAS Global Forum 2017 Conference (pp. 1-23). Cary, NC: SAS Institute Inc.

[15] Kraemer, F. A., Ammar, D., Braten, A. E., Tamkittikhun, N., & Palma, D. (2017, October). Solar energy prediction for constrained IoT nodes based on public weather forecasts. In Proceedings of the Seventh International Conference on the Internet of Things (pp. 1-8).

[16] Mardianto, M. F. F., Tjahjono, E., & Rifada, M. (2019). Statistical modelling for prediction of rice production in Indonesia using semiparametric regression based on three forms of fourier series estimator. ARPN J. Eng. Appl. Sci, 14, 2763-70.

[17] Men, H., Fu, S., Yang, J., Cheng, M., Shi, Y., & Liu, J. (2018). Comparison of SVM, RF and ELM on an Electronic Nose for the Intelligent Evaluation of Paraffin Samples. Sensors, 18(1), 285.

[18] PSL Web Team. (1948). Climate indices: Monthly atmospheric and ocean time series: NOAA physical sciences laboratory. Home: NOAA Physical Sciences Laboratory. https://www.psl.noaa.gov/data/climateindices/list/

[19] Supriyadi, B., Windarto, A. P., & Soemartono, T. (2018). Classification of natural disaster prone areas in Indonesia using K-means. International Journal of Grid and Distributed Computing, 11(8), 87-98.

[20] Tangang, F., Salimun, E., Aldrian, E., Sopaheluwakan, A., & Juneng, L. (2018). ENSO modulation of seasonal rainfall and extremes in Indonesia. Climate Dynamics, 51(7-8), 2559-2580.

[21] Vimont, D. J., Battisti, D. S., & Naylor, R. L. (2010). Downscaling Indonesian precipitation using large-scale meteorological fields. International journal of climatology, 30(11), 1706-1722.

[22] Wijayanto, A. K., Sani, O., Kartika, N. D., & Herdiyeni, Y. (2017, January). Classification model for forest fire hotspot occurrences prediction using ANFIS algorithm. In IOP Conference Series: Earth and Environmental Science (Vol. 54, No. 1, p. 012059). IOP Publishing.

Appendix A: Complete evaluation metrics for each individual ML model

	Tropical Savanna (Aw) RMSE					Tropical Monsoon (Am) RMSE					Tropical Forest (Af) RMSE				
	ERA5	ML			Best % error reduction	ERA5	ML			Best % Error Reduction	ERA5	ML			Best % Error Reduction
		RF	GBDT	XGB			RF	GBDT	XGB			RF	GBDT	XGB	
RR (mm/day)	12.39	10.95	10.94	11.70	13.25	14.15	12.57	12.60	15.82	11.17	20.15	18.79	18.81	21.10	6.75
Tx (°C)	3.17	1.41	1.35	1.46	57.41	3.89	2.25	2.23	2.25	42.67	3.47	1.69	1.71	1.83	51.30
Tn (°C)	1.69	1.45	1.43	1.69	15.38	2.37	2.18	2.21	2.15	8.02	2.38	1.83	1.86	2.10	23.11
Tavg (°C)	1.17	0.86	0.86	1.22	26.50	1.73	1.72	1.73	4.31	0.58	2.24	1.63	1.69	1.21	27.23
Ux (m/s)	6.22	2.19	2.12	2.39	61.58	5.59	4.20	4.20	4.37	24.87	3.89	2.01	2.00	2.40	48.59
Uavg (m/s)	3.16	1.16	1.37	1.39	63.29	4.28	1.12	1.11	1.37	74.07	2.47	2.03	3.03	3.16	17.81

Figure 2: Statistical evaluation for each of the machine learning models against observed weather variables, with the best performing algorithm (lowest RMSE) highlighted in yellow

Appendix B: Time-series of bias-corrected ERA5 values against observational climate data

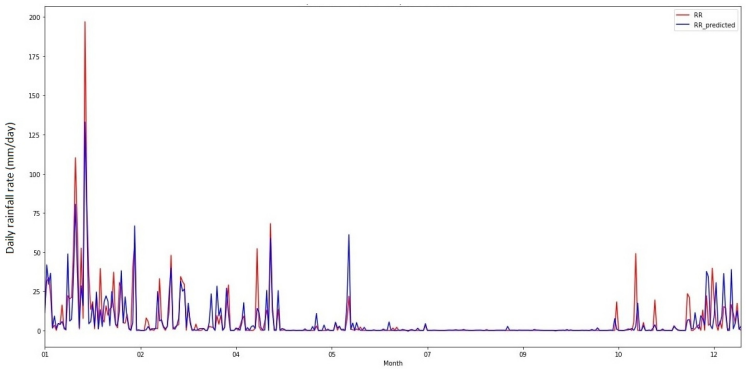

Figure 3: Time-series between bias-corrected ERA5 output and local climate variable for daily rainfall rate in one of the tropical forest's station in the year 2019

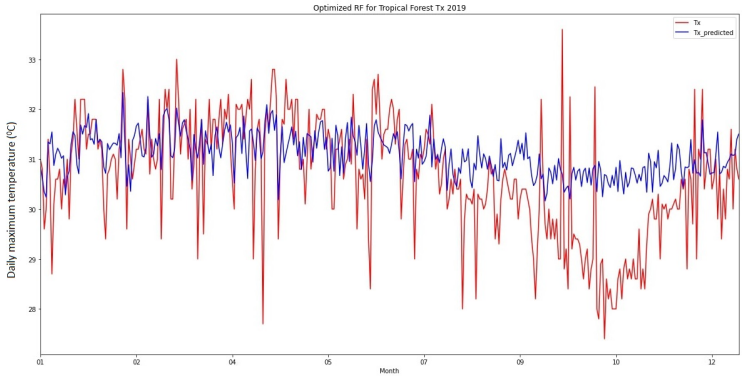

Figure 4: Time-series between bias-corrected ERA5 output and local climate variable for daily maximum temperature in one of the tropical forest's station in the year 2019

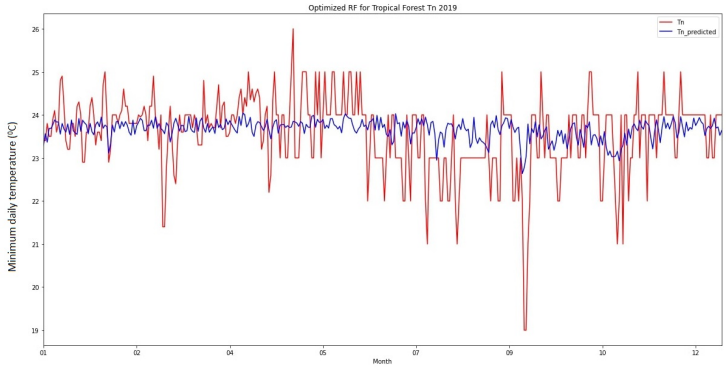

Figure 5: Time-series between bias-corrected ERA5 output and local climate variable for daily minimum temperature in one of the tropical forest's station in the year 2019

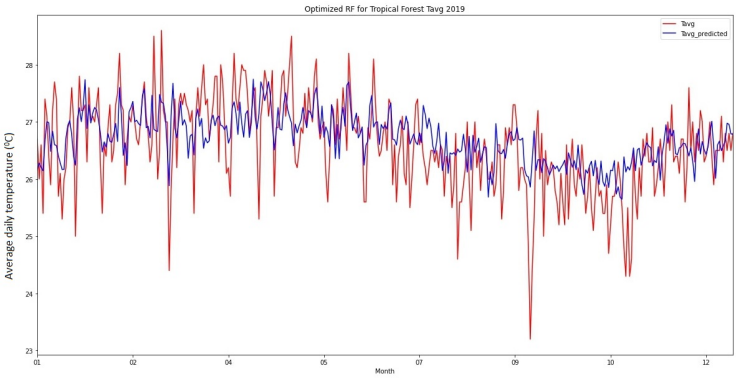

Figure 6: Time-series between bias-corrected ERA5 output and local climate variable for daily average temperature in one of the tropical forest's station in the year 2019

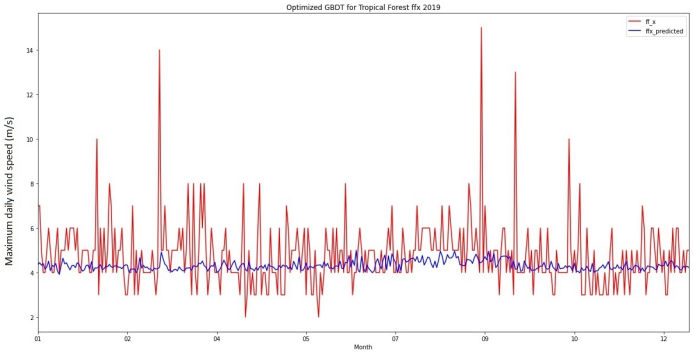

Figure 7: Time-series between bias-corrected ERA5 output and local climate variable for daily maximum windspeed in one of the tropical forest's station in the year 2019

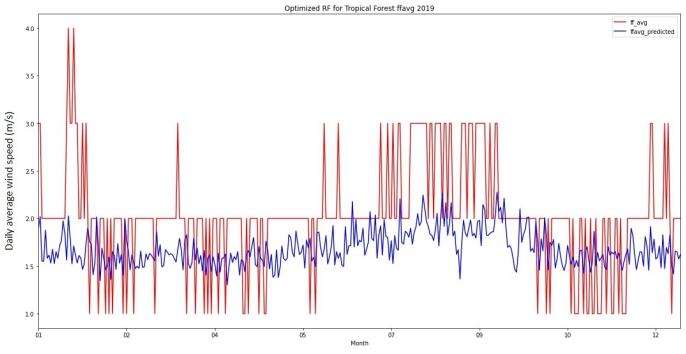

Figure 8: Time-series between bias-corrected ERA5 output and local climate variable for daily average windspeed in one of the tropical forest's station in the year 2019