Monthly daily-mean rainfall forecast over Indonesia using machine learning and artificial intelligence ensemble

H Harsa¹, M N Habibie, A S Praja, S P Rahayu, T D Hutapea, Y Swarinoto, R Kurniawan and S Noviati

Meteorology Department of Research and Development Center
Indonesia Meteorology Climatology and Geophysics Agency
Jl. Angkasa I No.2, Kemayoran, Jakarta Pusat, DKI Jakarta 10720.

¹hasatuadi@gmail.com

Abstract. A daily mean rainfall in a month forecast method is presented in this paper. The method provides spatial forecast over Indonesia and employs ensemble of Machine Learning and Artificial Intelligence algorithms as its forecast models. Each spatial grid in the forecast output is processed as an individual dataset. Therefore, each location in the forecast output has different stacked ensemble models as well as their model parameter settings. Furthermore, the best ensemble model is chosen for each spatial grid. The input dataset of the model consists of eight climate data (i.e., East and West Dipole Mode Index, Outgoing Longwave Radiation, Southern Oscillation Index, and Nino 1.2, 3, 4, 3.4) and monthly rainfall reanalysis data, ranging from January 1982 until December 2019. There are four assessment procedures performed on the models: daily mean rainfall establishment as a response function of climate patterns, and one-up to three-month lead forecast. The results show that, based on their performance, these non-Physical models are considerable to complement the existing forecast models.

1. Introduction

Atmospheric circulation at every location influences climatic condition everywhere on earth. This phenomenon is known as teleconnection and driven mainly by the results of fluid dynamic interaction between atmosphere and ocean [1] [2] [3] [4]. The behaviour of teleconnection in numerical atmospheric model has been investigated in some researches [5] [6] [7]. Given a set of initial states, numerical atmospheric models apply physics formula to forecast the future state of atmosphere variables. Since atmospheric teleconnection means that the condition of atmosphere in a particular location is triggered by events at distant locations, then the model aiming to capture the teleconnection process should be a global model. In global model, there are grids representing earth regardless they are evenly or unevenly distributed in distance. The representation of earth as grids is needed by the model’s equations to gather information propagated from a triggering event location through adjacent locations until the impact reaches distant locations. In the middle of the modelling process, the local atmosphere state also contributes to the amplification or diminishment of the triggering event.
This paper discusses a development of teleconnection-based models to establish a daily mean rainfall forecast in a month. Different from numerical models which rely on physics formulas as the main instrument, the models developed in this study used Machine Learning (ML) and Artificial Intelligence (AI) algorithms as the main instrument to build the model as well as data collection as the source of information on which the model was built. Conceptually, ML and AI were trained to learn the underlying pattern of how events at some places triggered other events at some distant locations. The learning phase of ML and AI was based on the data recorded both at the source and impacted locations. By using this technique, the need to provide continuous propagation media can be minimized.

2. Data and Methods
2.1. Data
There were two datasets provided for the model development, i.e. input and target. There were eight variables in the input dataset as learning materials of ML and AI. They were Nino1+2, Nino3, Nino4, Nino3.4, Dipole Mode West, Dipole Mode East, SOI, and OLR. All Nino are variable used to observe the Pacific Ocean. They are derived from the Sea Surface Temperature (SST) mean at each determined regions indicated by the number following the Nino variable names. Dipole variables are similar to Nino except that they are located at the Indian Ocean and represent the difference between western and eastern part of the Indian Ocean. There were two dipole indices, i.e. West and East to identify the
regions where the data were obtained. Southern Oscillation Index (SOI) is a standardized sea level pressure difference between Tahiti and Darwin. Outgoing Longwave Radiation (OLR) is a quantity of energy released by earth to space in the form of electromagnetic radiation. All data were obtained from National Oceanic and Atmospheric Administration (NOAA).

The target dataset given to ML and AI were monthly daily-mean rainfall dataset. The dataset was obtained from Japan Meteorological Agency (JMA). It consists of spatial reanalysis data spanning from 95°E to 145°E and -11.25°N to 7.5°N having a 1.25° of resolution. Hence, there were 41 and 16 grid divisions of longitude and latitude respectively. Both input and target datasets ranged from January 1982 to December 2019. The datasets used in this research is displayed by Figure 1. The input data were eight climate index variables having monthly time series interval. The target data were spatial datasets covering Indonesia area. The time series length of both input and target was 450 records.

2.2. Method

This research utilized ML and AI to build models. Since there were 41×16 (=656) grids in the target dataset, there must also be the same number of models provided by ML and AI. In this research, the models were also built for four lead records. There were total of 656×4 (=2624) models built. Furthermore, the output of each model was arranged spatially according to the location of the grid it represented. In other words, each model corresponded to the same input dataset but with different target. We chose XGBoost [8] for ML and Deep Learning (DL) [9] for AI as the algorithm to build models and H2O [10] as the programming interface. There were some model candidates provided for each algorithm at all grids. These model candidates had different parameter settings. After model candidates had been available, a stacked ensemble to all of them was constructed. The best model was then chosen as the final model for a particular grid. The best model for a grid was one having the smallest value of a loss function and might be a DL, XGBoost, or the ensemble of many models. Mean residual deviance was used as the loss function performance metric.

In the model building phase, the input and target datasets were randomly split to ten-fold cross validation for every grid and also had different fold compositions with other grids. To speed up model generalization, the input and target data were transformed. The transformation performed at the input data were scaling using each variable’s mean value and standard deviation as the center and scaling factor respectively, as shown by equation (1). This transformation changed the dispersion of all input variables to have the mean value of zero and the standard deviation of one. The target data were transformed using a base10-log transformation because rainfall value fluctuates from zero to infinitely large. This transformation reduced the great range of rain values so that the ML and AI can easily estimates data having great deviation. The target transformation is given by (2). In the prediction phase, the output of the model built by ML and AI must be inversed to the real rainfall value using the inverse transform as denoted by (3).

\[
X_t = \frac{X - \mu}{\sigma} \quad (1)
\]

\[
Y_t = \log_{10}(Y + 1) \quad (2)
\]

\[
Y = 10^{Y_t} - 1 \quad (3)
\]

where:

- \( \mu \) = mean
- \( \sigma \) = standard deviation
3. Results and Discussions

The increasing RMSE is also displayed by Figure 2 (b). The RMSE values are displayed a density function for each lead. There are two peaks at each lead in the figure. The two peaks indicate that there were two distinguishable categories of the model performance. These two categories of the model performance can be labelled subjectively as, for example, good and better. Using good and better label, the better category is denoted by the left peak and the good category is done by the right peak. Figure 2 (b) also tells us that as the lead increases, the better category is qualitatively stable but the good category is not. The good category slides slightly to right and has greater RMSE variance. Using the composition of RMSE at all leads, then the category labels can also be assigned to the model for assessing the qualitative spatial performance of the model. To assign the category label, a k-means clustering technique was utilized. The output of k-means resulted in an area separation as shown by Figure 3. The area separation can distinguish easily the spatial type of the model characteristics.
Figure 3. Area separation. The $k$-means clustering result of RMSE composition from all leads as displayed by Figure 2.

Figure 3 supports the hypothesis obtained by qualitatively compare the spatial RMSE in Figure 2 (a) that the Southern area were relatively more difficult to predict than the Northern area. The performance of the models was assessed using Root Mean Square Error (RMSE) metric. Since each model represented a spatial location, then the RMSE were also displayed spatially as shown by Figure 2 (a). The RMSE is presented for each monthly lead from 0 to 3, where each lead indicates the input dataset response of the model to the daily mean rainfall prediction. Lead 0 denoted the inline record series between the input and target datasets, whereas lead 1 to 3 did between the input and their consecutive target record indicated by the lead number. In Figure 2 (a), the Grater RMSE values were displayed as darker areas and vice versa. There are also triangle markers in each sub-figure to indicate the location where the model had the highest and lowest RMSE. It is also shown by Figure 2 (a) that the southern part of Indonesia has a greater RMSE value than that of the northern part and it increases at the subsequent leads. By observing the data pattern based on the RMSE, the locations having greater RMSE possessed higher value deviation. In terms of rainfall, the locations having greater RMSE might have easily-distinguishable dry and wet season. Conversely, the locations having smaller RMSE mostly have a lower deviation in their values. This also means that the seasonality is less distinguishable mostly in northern part of Indonesia.

Some examples of prediction obtained from the models are displayed by Figure 4. The prediction results were collected from the grids marked by triangles in Figure 2. Therefore, there are eight samples in Figure 4, i.e., the grid location where the highest and lowest RMSE detected for each lead. It can be observed from Figure 4 that the patterns of rainfall among locations having a low and high RMSE value were distinguishable. Locations having a high RMSE also have a higher variance than those with lower RMSE. Also note that there was almost no any bias between all models and their target indicated by the horizontal line as mean value of target and model data. The biases were only visible in a very small amount at “lead0” models for both highest and lowest RMSE locations. These biases were indicated by the presence of two horizontal lines although they were not separated far.

4. Conclusion
A data-driven climate model has been presented in this paper. The input data were climate indices and the target data were spatial daily rainfall reanalysis. The study area coverage of this paper was Indonesia. The ensembles of Machine Learning (ML) and Artificial Intelligence (AI) algorithms were utilized to build the daily rainfall estimates models for current month up to three months lead. There were as many models as the number of spatial reanalysis rainfall data grids. Generating The results show that there were two performance type from the RMSE of the models built by ML and AI ensembles. The northern part of Indonesia mostly had lower RMSE values than that of the Southern
The pattern of rainfall data contributed significantly to the performance of the models. There are some future works can be performed, for example, developing a system to facilitate the automation of prediction by the models and incorporating lagged values in the input dataset so that more information can be collected.

5. References

[1] Anteneh Z A, Assefa M M and Wossenu A 2019 Teleconnection of Regional Drought to ENSO, PDO, and AMO: Southern Florida and the Everglades Atmosphere 10 295
[2] Tao G, J Yu and H Paek 2016 Impacts of four northern-hemisphere teleconnection patterns on atmospheric circulations over Eurasia and the Pacific Theoretical and Applied Climatology, 129 815–831
[3] William R, Ian H, John B, Mark C and Ron C 2019 Understanding the potential of climate teleconnections to project future groundwater drought Hydrology and Earth System Sciences 23 3233–3245
[4] Yufei Z, Yuhang W, Zuowei X, Hailong W and Philip J R 2020 Atmospheric teleconnection processes linking winter air stagnation and haze extremes in China with regional Arctic sea ice decline Atmospheric Chemistry and Physics 20 4999–5017
[5] Sloan J C, Jason E S, Benjamin I C and Richard S 2013 Stationarity of the Tropical Pacific Teleconnection to North America in CMIP5/PMIP3 Model Simulations Geophysical Research Letters 40 1-6

[6] Erzsébet K, Zoltán B, Roland H, Judit B and Rita P 2020 Evaluation of Historical CMIP5 GCM Simulation Results Based on Detected Atmospheric Teleconnections Atmosphere 11 723

[7] Dörthe H and Klaus D 2012 How well do state-of-the-art atmosphere-ocean general circulation models reproduce atmospheric teleconnection patterns Tellus A: Dynamic Meteorology and Oceanography 64 19777

[8] Tianqi C and Carlos G 2016 XGBoost: A Scalable Tree Boosting System Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 785-794

[9] Konstantinos V, Magdalena K, Thomas J, Alexander P, Archimides A, Ioannis P S, Michael L, Anastasia V, and Elizabeth O J, Vassilis G, Antonia V, Paul T, Dimitris T, Jiri B and Vassilis G G 2016 Deep Learning and Association Rule Mining for Predicting Drug Response in Cancer. A Personalised Medicine Approach bioRxiv

[10] Arno C, Erin L, Viraj P and Anisha A 2016 Deep Learning with H2O Jessica L H2O.ai, Inc Leghorn St. Mountain View, CA