Monthly reservoir inflow prediction based on artificial neural network over the Saguling catchment area

A Rachmawati1,2, I P Santikayasa1, Impron1 and F Alfahmi2

1IPB University, St. Raya Dramaga, Bogor, West Java, Indonesia
2BMKG, St. Angkasa 1 No. 2, Central Jakarta, Indonesia

asrirachmarachmawati@apps.ipb.ac.id

Abstract. In line with the objectives of the Impact-Based Forecast, hydrological predictions are used to support water resources management, mitigation of natural disasters, and climate variability impacts. The Artificial Neural Network algorithm was applied to build the prediction model for the Saguling Reservoir monthly inflow by utilizing observation data; the monthly rainfall (P) and inflow (Q) as the predictors in different lag time; t (recent), t-1 (previous month), and t-11 (11 months ago). The predictors are simulated in three variations of hidden layer numbers (2, 6, and 10). The best model under the normal period is model nine with RMSE 31.23 and R 0.88. This model is also the best model under La Nina's condition with RMSE 30.01 and R 0.83. For the El Nino period, we found that model five is the best model with the highest accuracy at the level of generalization in both the training process and predict extreme conditions at the validation stage. Overall, this model has a good performance and high potential usage from a practical point of view and costs but needs further simulation to make it more reliable and robust in any climate conditions.

1. Introduction
Impact-based forecast (IBF) is being so popular since this program is very useful to reduce misunderstanding of the weather and climate information dissemination. The World Meteorological Organization (WMO) has suggested a potential solution is for national meteorological and hydrological services (NMHSs) to use impact-based warnings to bridge the gap between forecasts and possible impacts of impending hazards [1]. Service in weather forecasting and warning is the primary role of BMKG to support the sustainability of various sectors of life, one of which is closely related to hydrological prediction. For many years, hydrological prediction methods based on hydro-meteorological data were developed to improve the prediction performance and to support the efficient water resources management that is related to infrastructure and mitigate the impact of natural disasters and climate variability [2]. Reservoirs are used as a problem-solving of water-related demand problems like hydropower, urban water daily need, conservation, irrigation, and flood mitigation. Due to the management operation under control, the development of model prediction such as the reservoir inflow prediction is being improved continuously, on the other hand, the black-box model especially machine learning-based models are still very rarely used as an operational routines program. That for large-scale hydropower systems, choosing one of the mechanistic models is going to be hard due to the lack of data availability [3]. Although the machine learning-based models are the black-box model that the internal processes involved cannot be understood but from a practical point of view, it has so many advantages [4]. In the ANN, the learning objective is to recognize the best relationship or pattern between the output and input variables, so it must be trained in an appropriate training algorithm. Backpropagation is one
of the algorithms commonly used and proven reliable in several studies in various fields, and this algorithm is adopted to train the ANN models in this study.

Various types of literature prove that an Artificial Neural Network (ANN) is a powerful model in numerous problems, including the operational management of water resources [5]. Coulibaly et al. (2000) initiated the use of a stop-training approach to train networks on the multi-layer feed-forward neural networks (FNN) model in building a hydrological prediction model in the Chute-du-Diable hydro system in northern Quebec (Canada). The results show that the proposed method is an effective model for reservoir inflow in real-time forecasting with relatively affordable computation costs [6]. Kumari and Srinivasu (2019) proposed an ANN Reservoir computing approaches to predict the Srisilam Dam of daily inflow and compared this proposed approach with ARIMA and ANN, which proved that the ANN is better than the ARIMA, even farther the ANN reservoir computing get beyond the ANN performances [7]. Lee et al. (2020) developed a statistical inflow forecasting model for the Boryoung Dam based on the relationship between the monthly reservoir inflow and the climate index variable. The simulated models are multiple linear regression, supporting vector machines, and ANN. Hence, this study cannot achieve perfect model performance, the study was being a promising start of statistical model for the operation of water resources and provides recommendations for future research to further combine the climate index with other variables such as rainfall and flow, and previous monthly flow for better skills models [8].

The purpose of this study is to simulate the prediction of the Saguling monthly inflow on ANN models approach, with the precipitation (mm) over the Saguling Catchment area and the average monthly inflow (m3 / s) as the input. Various scenarios are simulated in the training, testing, and validation of models. Furthermore, this study can provide another perspective in supporting the role of BMKG in the Citarum Cascade Dam Management Coordination Team (TKPBKC) in hydrological prediction in water resources management.

2. Methodology

2.1. Study Area and Data

The focus area of this study is the Saguling Reservoir catchment area located in the upper Citarum river basin (UCRB) between 60° 54’45” S and 107° 21’ 58” E (fig. 1). It is located in West Java and one of the largest and most important in Indonesia. The Saguling Reservoir is upstream of two other reservoirs, namely Cirata Reservoir and Jatiluhur or Juanda Reservoir. The reservoir which was built in 1981 has a catchment area of 53,000 hectares with a water capacity of 609 million m3. The reservoirs are used for hydroelectric power to supply electricity to Java and Bali, which are managed by PT. Indonesia Power POMU and agricultural irrigation purposes, also.

![Figure 1. The Saguling catchment area](image)
The primary data are the observation of monthly rainfall obtained from the ombrometer station around the Saguling reservoir that collected by Bogor Climatological Station as the coordinator of West Java climate data, there are more than twenty-five rain gauge stations over the Saguling catchment area and mostly scattered in the Bandung area (figure 1), this fairly large distribution of rain posts makes observational rainfall data availability is better than other climate parameters. Another primary data is monthly Saguling reservoir inflow that is observed and managed by PT. Indonesia Power POMU and for the additional information is Southern Oscillation Index (SOI) from National Centres for Environmental Prediction (NCEP). The data periods are 30 years or between January 1990 - December 2019.

2.2. Preprocessing.
The data processing started with the pre-processing step to treat and prepare the raw (observation) data before building the model. The beginning of the process is data collecting and filtering under the following rules:
- the data series up to 30 years (1990-2019)
- having good continuity data histories or not much missing data
- and it has a good correlation with monthly Saguling Reservoir inflow.

The rainfall data from twenty-five rain gauge stations over the Saguling catchment area are collected and quality controlled. Figure 2 shows the coefficient correlation score between monthly rainfall data and the monthly inflow (orange bar). The blue and yellow bar shows the presentation of each data station's availability. The yellow bars are the selected station as an input data model. The condition of the observation data is often under the expectations, there are conditions that we have to deal with missing data either due to human error factors or technical constraints that are tools or natural disasters. On the other hand, we need the complete data to build the model, so we have to do the missing data filling step. One of the simple method data is using the normal or climatological value data of precipitation for at least 30 years.

![Figure 2](image)

**Figure 2.** The Rainfall data availability from 25 rain gauges are scattered over the Saguling catchment area (blue and yellow bar (selected data)) and the data correlation scores to monthly inflow data (orange bar).

Then, we use PCA to determine which lag time data are efficient to use as model input data. The PCA is a multi-variable analysis method that aims to reduce the dimensions of the original variables into the new variables (the main components) that still store most of the information [9]. Table 1 is the
analysis of eigenvalues analysis for each component. Since we need the univariate variable, so we have to choose PC1 only, with 76.2 % as the proportion of representing the monthly rainfall data from 9 rainfall station data.

Table 1. Eigen analysis of the covariance PCA

| Description | PC1     | PC2     | PC3     | PC4     | PC5     | PC6     | PC7     | PC8     | PC9     |
|-------------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| Eigenvalue  | 141010  | 14654   | 6922    | 6494    | 5251    | 3796    | 3525    | 1948    | 1416    |
| Proportion  | 0.762   | 0.079   | 0.037   | 0.035   | 0.028   | 0.021   | 0.019   | 0.011   | 0.008   |
| Cumulative  | 0.762   | 0.841   | 0.879   | 0.914   | 0.942   | 0.963   | 0.982   | 0.992   | 1       |

The data transform equation is:

\[ PC_1 = 0.372P_1 + 0.311P_2 + 0.527P_3 + 0.221P_4 + 0.332R_5 + 0.304P_6 + 0.284P_7 + 0.27P_8 + 0.288P_9 \] (1)

The correlation between the new variable rainfall (PC1) and the reservoir inflow data at the different lag times with the predicted reservoir inflow data \((Q_t + 1)\) is analyzed. This analysis will determine the proper input models and the best scenarios (Table 2). The selected predictors are highlighted in yellow.

Table 2. The predictor coefficient correlation score at the various lag time

| Input   | Correlation | P-value | Input   | Correlation | P-value |
|---------|-------------|---------|---------|-------------|---------|
| Q      | 0.634       | 0       | P       | 0.698       | 0       |
| Q_{t-1}| 0.332       | 0       | P_{t-1} | 0.479       | 0       |
| Q_{t-2}| 0.063       | 0.238   | P_{t-2} | 0.252       | 0       |
| Q_{t-3}| -0.165      | 0.002   | P_{t-3} | 0.033       | 0.534   |
| Q_{t-4}| -0.37       | 0       | P_{t-4} | -0.222      | 0       |
| Q_{t-5}| -0.464      | 0       | P_{t-5} | -0.448      | 0       |
| Q_{t-6}| -0.391      | 0       | P_{t-6} | -0.482      | 0       |
| Q_{t-7}| -0.22       | 0       | P_{t-7} | -0.381      | 0       |
| Q_{t-8}| -0.038      | 0.482   | P_{t-8} | -0.232      | 0       |
| Q_{t-9}| 0.168       | 0.002   | P_{t-9} | -0.017      | 0.745   |
| Q_{t-10}| 0.432      | 0       | P_{t-10}| 0.284       | 0       |
| Q_{t-11}| 0.567      | 0       | P_{t-11}| 0.532       | 0       |
| Q_{t-12}| 0.428      | 0       | P_{t-12}| 0.51        | 0       |

Notes: \(Q = \) monthly Reservoir Inflow (m³/s) \(P = \) monthly rainfall (mm) \(t = \) time (month)

The last step of preprocessing is separating data for the training, testing, and validating periods. Then, we take the hold-out cross-validation method to separate the whole data into 75% data are for the training period, 25%. This technique is common for any machine learning model for its easiness and efficiency [10]. For the validation stage, we use the last one-year data, Normal, El Nino, and La Nina period, the data validation filtering uses additional information namely 30 years of SOI from NCEP. this validation stage is expected to measure the model reliability and robustness in various real situations that may occur.

2.3. Prediction Model Based on ANN

As the input model, we have monthly rainfall data from nine selected rain gauge stations data (P) and monthly reservoir inflow (Q) in 3 different lag time specifically are recent time (Pt and Qt), previous month (Pt-1 and Qt-1), and the 11 previous months (Pt-11, Qt-11). The combinations are 10, 20, and 30
input layers. It is simulated with three different number of hidden layers (2, 6, and 10 hidden layers) so that there are nine model scenarios simulated in the ANN model to predict the inflow (table 3).

Table 3. Nine scenarios of monthly Saguling reservoir inflow prediction model

| Model | Num. of Input Layer | Data Input | Num. of Hidden Layer | Output |
|-------|---------------------|------------|----------------------|--------|
| One   | 10                  | $P_t, Q_t$ | 2                    |        |
| Two   | 10                  | $P_t, Q_t$ | 6                    |        |
| Three | 10                  | $P_t, Q_t$ | 10                   |        |
| Four  | 20                  | $P_t, Q_t, Q_{t-1}$ | 2 | $Q_{t+1}$ |
| Five  | 20                  | $P_t, Q_t, Q_{t-1}$ | 6 |        |
| Six   | 20                  | $P_t, P_{t-1}, Q_t, Q_{t-1}$ | 10 |        |
| Seven | 30                  | $P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_t$ | 2 |        |
| Eight | 30                  | $P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_t$ | 6 |        |
| Nine  | 30                  | $P_t, P_{t-1}, P_{t-2}, Q_{t-1}, Q_t, Q_{t-1}$ | 10 |        |

Notes : $t$ = recent data (month), $t-1$ = last month, $t-11$ = 11 months ago

ANN is an enormously distributed parallel information processing system that is very similar to the system of the human brain [11]. The fundamental rules of the ANN model development are:

- The Information is processed on many single elements of neurons (nodes/units/cells).
- Information signals propagate among the connected neurons.
- Each connected nets have a weight unit as the strength of the connection.
- the nonlinear transformation applied to each unit of the neuron as the activation function [4].

![Figure 3a](image-a). The basic algorithm of ANN

![Figure 3b](image-b). ANN BP architecture design.

The main components of the ANN model are input data, information processing among the neuron, and output that compare with the target (fig. 3a). This component is adjustable freely to get the best model performance. Information processing from input data occurs on many single elements (neurons), each node or neuron applies a nonlinear transformation called an activation function. Signals are passed between nodes via a net that is connected among the neuron a layer, and each connection in the net has an associated weight [3]. This process has a feed-forward propagation of error. In the ANN BP, the error unit or weights are re-propagation back to update the value of the weights to get the best output model and this process may occur in many repetitions based on the predetermined epoch or iteration (fig. 3b).
3. Result And Discussion
The whole procedure is systematically to get the inflow prediction, several scenarios are simulated to minimize the limitations of the model which are influenced by different types of architecture, and to see how the number of data input layers for the model affects the performance of this model. This study used two model performances indicator, the root means square error (RMSE) and the regression value ($R$) is applied. RMSE value shows the mean error or the difference value of the mean of the output model and the target or the prediction, while the $R$-value that has ranged from 0 to 1 measure how well the regression line is close between the output model and target [12].

Babei et al. (2019) simulate an ANN model to predict the monthly inflow using rainfall data with time lag variations as the input. The result shows that the number of input variables can increase the skill model. On the other side, the number of neurons requires the accuracy of predictor selection based on the characteristics of the existing data to obtain results of a good and efficient learning process [13].

Based on the results presented in table 4, obtained results are also similar. It is worth remembering that the rainfall characteristics play an important role as this is the basis of the pattern recognition process at this stage of training models.

Table 4. Skill model measurement recapitulation

| Model | Num. Hidden Layer | Predictor              | Training |              | Testing |              | Validating |              |
|-------|------------------|------------------------|----------|--------------|---------|--------------|------------|--------------|
|       |                  |                        | RMSE     | $R$          | RMSE    | $R$          | RMSE       | $R$          |
| one   | 2                | $P_t, Q_t$             | 36.55    | 0.83         | 45.87   | 0.73         | 32.72      | 0.90         |
| two   | 6                | $P_t, Q_t$             | 34.62    | 0.85         | 47.38   | 0.87         | 25.41      | 0.94         |
| three | 10               | $P_t, Q_t$             | 28.32    | 0.90         | 52.04   | 0.87         | 42.32      | 0.84         |
| four  | 2                | $P_{t-1}, P_t, Q_{t-1}, Q_t$ | 28.56 | 0.90         | 53.39   | 0.86         | 29.93      | 0.92         |
| five  | 6                | $P_{t-1}, P_t, Q_{t-1}, Q_t$ | 27.02 | 0.91         | 50.21   | 0.89         | 30.81      | 0.92         |
| six   | 10               | $P_{t+1}, P_t, Q_{t-1}, Q_t$ | 31.01 | 0.88         | 44.96   | 0.86         | 38.90      | 0.85         |
| seven | 2                | $P_{t+1}, P_{t+1}, P_t, Q_{t+1}, Q_{t+1}, Q_t$ | 24.62 | 0.92         | 30.52   | 0.85         | 47.28      | 0.77         |
| eight | 6                | $P_{t+1}, P_{t+1}, P_t, Q_{t+1}, Q_{t+1}, Q_t$ | 34.52 | 0.85         | 48.60   | 0.69         | 43.22      | 0.71         |
| nine  | 10               | $P_{t+1}, P_{t+1}, P_t, Q_{t+1}, Q_{t+1}, Q_t$ | 32.90 | 0.87         | 29.57   | 0.85         | 57.81      | 0.64         |

In table 4 we can see the performance or skills of all models at all stages. Almost all models performed well during the training stage, even being able to reach an $R$ value of 0.92 (model seven), this performance dropped quite drastically at the testing stage for model one and model eight. Then, we obtained the satisfaction model performances for the validating stages, the best performances show model four with RMSE is 25.41 with the $R$-value is 0.94. In general, the model that can achieve satisfactory results at every stage is model 5. This model has high accuracy at the level of generalization in the training process but can predict extreme conditions at the validation stage. For example, the inflow of 2019 April was an extreme event. The inflow is 279 m$^3$/s, which is 67% bigger than the climatological value. The detail of the validations is shown by figure 4.
The characteristic of rainfall in West Java is especially the monsoon type and is greatly influenced by extreme conditions such as La Nina and El Nino, that rainfall in the West Java region tends to fall during El Nino and vice versa, it relatively increases during the La Nina period [14]. So that, it is necessary to evaluate the model under those conditions. Based on table 5, we can see that some scenarios may suitable in normal conditions, but the others are better for extreme conditions, both El Nino and La Nina.

![Figure 4. Output model for January – December 2019](image)

### Table 5. Skill model for validation period recapitulation

| Model | Normal | El Nino | La Nina |
|-------|--------|---------|---------|
|       | RMSE   | R       | RMSE    | R       | RMSE    | R       |
| one   | 41.00  | 0.78    | 38.70   | 0.88    | 35.71   | 0.76    |
| two   | 40.90  | 0.78    | 36.47   | 0.89    | 34.61   | 0.79    |
| three | 37.60  | 0.82    | 32.62   | 0.91    | 35.27   | 0.77    |
| four  | 39.70  | 0.79    | 30.77   | 0.92    | 35.20   | 0.76    |
| five  | 35.63  | 0.84    | 31.09   | 0.92    | 34.92   | 0.76    |
| six   | 35.75  | 0.84    | 34.80   | 0.90    | 33.87   | 0.77    |
| seven | 30.67  | 0.88    | 34.89   | 0.90    | 32.27   | 0.80    |
| eight | 41.16  | 0.78    | 37.78   | 0.89    | 33.96   | 0.78    |
| nine  | 31.23  | 0.88    | 35.49   | 0.89    | 30.01   | 0.83    |

The best scenario for normal conditions is shown by models seven and nine (figure 5), while at El Nino period the best prediction model scenario is shown by model four or five (figure 6), while for conditions with more rainfall in the La Nina period the model nine performed better than other modeling scenarios (Figure 7).
Figure 5. Output model seven for the normal period

Figure 6. Output model four for the El Nino period

Figure 7. Output model nine for the La Nina period
4. Conclusion
This study applied an ANN-BP model by simulating various combinations of model input data into models one to nine models. The scenarios are simulated step by step, started with data pre-processing in the training, testing, and validating stage. The model has a good performance based on the RMSE and R values. Model five is the best model with the highest accuracy at the level of generalization in the training process and is capable to predict extreme conditions at the validation stage. The further validation in the normal and extreme periods (El Nino and La Nina), model 7 or 9 is best used in normal conditions with RMSE 30.67 and 31.23 and R both 0.88, model 9 is also the best model used in La Nina conditions with RMSE 30.01 and R- 0.83. Whereas in El Nino conditions models 4 and 5 have RMSE 30.77 and 31.09 with R 0.92. Overall, this model has a good performance and high potential usage from a practical point of view and costs to be applied in the operational. Hence, there are early indications of the possibility of overfitting and the less robustness of the scenario changes, the further simulations are needed for a more robust and more reliable architectural model design in any climate condition.

References
[1] WMO 2015 WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services
[2] Butts M, Jensen R, Larsen J, Mueller H, Richaud B, Larsen P, Draw B dan Alejandro E 2017 Seasonal Forecasting of Reservoir Inflows and Their Uncertainties 6865 4431–7
[3] Irvine K N dan Eberhardt A J 1992 MULTIPLICATIVE, SEASONAL ARIMA MODELS FOR LAKE ERIE AND LAKE ONTARIO WATER LEVELS JAWRA J. Am. Water Resour. Assoc.
[4] ASCE 2000 Task Committee on Application of Artificial Neural Networks in Hydrology, Artificial Neural Networks in Hydrology. II:Hydrologic Application J. Hydrol. Eng. 5 124–37
[5] Noori R, Khakpour A, Omidvar B dan Farokhnia A 2010 Comparison of ANN and principal component analysis-multivariate linear regression models for predicting the river flow based on developed discrepancy ratio statistic Expert Syst. Appl.
[6] Coulibaly P, Anctil F dan Bobée B 2000 Daily reservoir inflow forecasting using artificial neural networks with stopped training approach J. Hydrol.
[7] Pradeepakumari B dan Srinivasu K 2019 Dam Inflow Prediction by using Artificial Neural Network Reservoir Computing Int. J. Eng. Adv. Technol. 9 662–7
[8] Lee D, Kim H, Jung I dan Yoon J 2020 Monthly reservoir inflow forecasting for dry period using teleconnection indices: A statistical ensemble approach Appl. Sci. 10
[9] Jollife I T dan Cadima J 2016 Principal component analysis: A review and recent developments
[10] May R J, Maier H R dan Dandy G C 2010 Data splitting for artificial neural networks using SOM-based stratified sampling Neural Networks
[11] Haykin S dan Network N 2004 A comprehensive foundation
[12] Dawson C W, Abrahart R J, Shamseldin A Y dan Wilby R L 2006 Flood estimation at ungauged sites using artificial neural networks J. Hydrol.
[13] Babaei M, Moeini R dan Ehsanzadeh E 2019 Artificial Neural Network and Support Vector Machine Models for Inflow Prediction of Dam Reservoir (Case Study: Zayandehroud Dam Reservoir) Water Resour. Manag. 33 2203–18
[14] Nabilah F, Prasetyo Y dan Sukmono A 2017 Analisis Pengaruh Fenomena El Nino Dan La Nina Terhadap Curah Hujan Tahun 1998 - 2016 Menggunakan Indikator Oni (Oceanic Nino Index) (Studi Kasus : Provinsi Jawa Barat) J. Geod. Undip 6 402–12