Peformance improvement of Neural Networks in daily streamflow prediction using principal component analysis

M. R. Fallah Haghgoo Lialestani[1] , M. B. Sharifi[2], M. Mirabi[3]

1- MSc Graduate in Water Engineering, Sarvab Consulting Engineers Co., Mashhad, Iran

2- Associate Professor, Department of Civil Engineering, Ferdowsi University of Mashhad, Iran

3- Islamic Azad University, Torbat-e Heydariyeh Branch, Departemant of Civil Engineering, Torbat-e Heydariyeh, Iran

Moh_haghgoo@yahoo.com
Mbsharif@ferdowsi.um.ac.ir
Mirabi.mehrdad@gmail.com

Abstract

Successful river flow time series forecasting is a major goal and an essential procedure that is necessary in water resources planning and management. The objective of this study is to evaluate the effectiveness of Principal Component Analysis (PCA) in the improvement of the ANN model performance. In this work, we develop and test two artificial neural networks (MLP, GRNN) to forecast streamflow. MLP and GRNN models, including the original ANN model without data preprocessing, were set up and evaluated. Understanding the temporal relationships between climatic drivers and streamflow is fundamental to the model development. statistical approach depending on cross-, auto- and partial-autocorrelation of the observed data is used as a good alternative to the trial and error method in identifying model inputs. these analyses revealed temporal dependencies between the climate-flow datasets. we modeled the flow using the Principal Component Analysis (PCA) technique that reduces the number of input variables to include only the ones effective in ANN (PCA-ANN). The performance of ANN models in training and testing stages are compared with the observed streamflow values to identify the best fit forecasting model based upon a number of selected performance criteria. Under the overall consideration including the model performance and the complexity of modeling, the MLP-PC4 model was optimal.

Keywords: Streamflow, MLP, GRNN, Principal Component Analysis

1. Introduction

Stream flow discharge forecasting has been considered as an important challenge for the researchers in the two past decades. For the purpose of stream flow discharge modeling different approaches such as regression, conceptual and black box models are used [1].

Physicsbased rainfall-runoff models require considerable data and human effort to calibrate, validate, and test but are extremely useful in understanding the governing physics or processes. Because of the limited resources associated with developing and calibrating conceptual, metric, and physics models, data-driven hydrological methods have been widely adopted for forecasting streamflow. Multiple linear regression (MLR), variations of autoregressive moving average (ARMA) models and artificial neural networks (ANNs) are commonly used methods [2].

In data collection you encounter situations where there are large number of variables in the database. In such situations it is very likely that subsets of variables are highly correlated with each other. The accuracy and reliability of a classification or prediction model will suffer if you include highly correlated variables or variables that are unrelated to the outcome of interest. One of the key steps in data collection is finding ways to reduce dimensionality without sacrificing accuracy. Principal component analysis (PCA) is a classic technique in data analysis. It can be used for compressing higher dimensional data sets to lower dimensional ones for data analysis, visualization, feature extraction, or data compression.

The objective of principal component analysis is to reduce the dimensionality (number of variables) of the data set but retain most of the original variability in the data. The objective of this study is to evaluate the