EPiC Series in Engineering
Volume 3, 2018, Pages 1457–1464
HIC 2018. 13th International Conference on Hydroinformatics

Real-Time Quality Control and Infilling of Precipitation Data Using Neural Networks

Mohammadreza Moslemi[1], Darko Joksimovic[1]
[1] Ryerson University, 350 Victoria Street, Toronto, Ontario M5B 2K3, Canada
moslemi@ryerson.ca, darkoj@ryerson.ca

Abstract. Due to advancements in instrumentation and communication technologies, monitoring of water infrastructure is experiencing a significant growth worldwide and water managers are increasingly deploying monitoring equipment for decision-making purposes. Hydrological events and relevant datasets including rainfall data are of a complex nature and are potentially susceptible to errors from various sources. Hence, it is essential to develop efficient methods for the quality control of the acquired data. The present work introduces an artificial neural network-based approach for real-time quality control and infilling of rain gauge data. Available rainfall measurements from neighboring rain gauges are employed to train and develop the neural network model. Trained artificial neural network model was able to validate up to about 97% of the data using 95% confidence intervals. This finding suggests that artificial neural networks can be successfully implemented for erroneous data identification/correction and reconstruction of missing data points. Given its short processing time and reportedly superior performance to traditional quality control strategies, neural network methodology can be deployed as an efficient tool for the processing and control of large sets of timeseries with complex natures including precipitation data.

Keywords: Artificial Neural Networks, Quality Control, Rain Gauge, Rainfall

1 Introduction

Flow monitoring and rainfall data are widely deployed by municipalities and engineering firms for a variety of purposes including analyses, design and operation of urban drainage systems. Also, increasing attention is being paid to real-time monitoring and performance assessment of collection systems which enables collection systems' operation/maintenance optimization, infiltration/inflow (I/I)) tracking, and combined sewer overflow (CSO) assessments. As the cost of implementing monitoring networks is reduced due to technological advancement, the volume and availability of reliable real-time data will have an increasing and significant role in water resources management and engineering. However, flow and rainfall sensors are associated with measurement deficiencies and errors [1-3] which in many cases are inevitable. Generated datasets may include faulty or missing data points that can diminish overall data quality and influence subsequent analyses and decisions that rely upon them. Moreover, manual data processing approaches are time consuming, costly, reliant on analysts' experience, and also prone to human error and oversight. Insufficient existing conventional quality control (QC) methodologies along with the rapidly growing volume of generated data and real-time applications, intensifies the necessity to develop efficient automated QC procedures for sensor-generated timeseries and provide end users with accurate and reliable real-time hydrometric data. The importance of accessibility to reliable precipitation data becomes more prominent when it comes to short-term monitoring of small catchments as is the case for I/I studies. I/I which is the entry of water originating from rainfall or groundwater sources into sanitary sewer systems, can impair the performance of the collection systems. I/I studies in many cases rely on the accuracy of single precipitation monitoring stations within a monitored catchment.

2 Current Data Processing Approaches

Conventional approaches for quality control of timeseries data typically rely on manual inspections or semi-automated simple statistical methodologies, including checks for outliers through comparison with temporal and spatial observations at that particular station and at neighboring stations such as tests for threshold/maximum values, flat line, spike, rate of change, etc. Moreover, traditional data processing methods are usually based on certain assumptions that are not necessarily valid and applicable given the actual properties of the data [3-7].

A number of researchers have focused on rain gauge data QC utilizing statistical approaches through cross checking with contemporary weather radar data, however radar precipitation data can also be associated with various types of errors arising from radar calibration, variation of the vertical reflectivity profile, bright band errors, attenuation, ground clutter, anomalous propagation, and wind drift errors [8]. In addition, radars are often reported to underestimate rainfall data (up to 50% underestimation has been reported [2,9,10]). Therefore, rain gauge data are normally assumed as ground truth in most quality control schemes and are used for the calibration and validation of radar data [10]. As an alternative approach, researchers have also tried to integrate manual QC procedures into a systematic algorithm, the so-called Case Based Reasoning (CBR). Aside from performing traditional data QC and correction, the CBR approach also identifies and proposes the most relevant or frequently used data correction methodology on the basis of previous similar incidents and respective operators' judgment/decisions. Although such approaches are capable of adjusting on the basis of operators' inputs, they are quite labour intensive and are reliant on traditional statistical methodologies. Various CBR tools such as gapIT, WaterQuality CBR, CBR Shell, myCBR, Free CBR, jCOLIBRI, and CASPIAN with various fields of applications are reported elsewhere [11-15].

More recent data mining approaches have recognized machine learning using artificial neural networks (ANNs) as a robust and efficient methodology for data analytics in various fields of practice including water data analytics [4,16]. Despite its versatility and successful applications in various QC fields, machine learning approaches have seldom been used as the main scheme for precipitation data processing. This lack of implementation in hydrometeorological arena offers further opportunities to develop and enhance ANN based QC approaches within the hydrometric data processing field. Moreover, limited work with the aim of QC of rain gauge data through concurrent use of neighboring gauge readings and radar precipitation data has been reported [3,18]. It is believed that such data can be employed as a valuable source for the fine tuning and optimization of neural networks QC procedures. ANN quality control algorithms can also be combined with traditional statistical detection and flagging of outliers in order to enhance the accuracy and robustness of the QC procedures.

3 Objectives

Given the shortcomings of traditional QC schemes, the existing project aims to develop a real-time machine learning algorithm for the automated detection/correction of erroneous data and reconstruction of missing values in a precipitation data collection and processing system using contemporary measurements at nearby gauges. The outcome of this research can bring about significant benefits through automation of data quality assurance and control (QA/QC) processes, resulting in the reduction of the time and cost associated with manual data procedures, and the enhancement of the quality and reliability of the generated data. This will subsequently support municipalities and other stakeholders relying on large datasets being generated by water monitoring networks, resulting in better design and operational decisions for the benefit of urban populations and the receiving water environment.

4 Available Database

Available hydrometeorological database pertaining to the geographical area of study, the City of Toronto, Canada, was investigated and collected. Major available data sources include the following.
- National Oceanic and Atmospheric Administration (NOAA): Radar hourly precipitation data from the KBUF S-band dual-polarimetric radar station located in Buffalo, New York, US is available by volume scan (Level III format) over the period of February 1996 to September 2017.
- Environment Canada: Hourly precipitation data are available from various rain gauges in Toronto area. Hourly radar imagery from the King City C-band dual-polarization radar station located near King City, Ontario, Canada is also available for extraction.
- The City of Toronto: The City's rain gauge network consists of five 4-season and thirty nine 3-season gauges. Datasets from these gauges are available in 5-min intervals as of June 2015.
- Toronto and Region Conservation Authority (TRCA): Precipitation data in 5-min intervals from various rain gauges within the City of Toronto is available as early as 2000's up to 2017.

5 Methodology

In the present research, a rainfall estimation algorithm based on ANNs is developed that takes advantage of nearby rain gauge data. The following is a summary of the deployed methodology, which is also represented schematically in Figure 1.

- Reference rain gauge stations neighboring the rain gauge of interest (target station) are selected on the basis of geographic proximity (distance/altitude differences), and contemporary availability of rainfall data.
- ANN is calibrated (trained and validated) using pre-processed historical data available from reference rain gauges as model's input and corresponding rainfall measurements at the target station as model's target.
- Trained ANN is used to estimate rainfall at the target station using measurements available at the respective reference stations. The estimated values will be used to quality control (error detection/correction) of suspicious observations or to fill the missing data.
- Rainfall measurements at the target station are marked as validated if they fall within the confidence intervals obtained by means of neural networks for a desirable significance level α (e.g. 5%), implying that the prospective rainfall value is expected (with a probability of 1- α) to have fallen within the derived intervals. Average values estimated using the ANN model are used for data correction/infilling. In case owing to scattered/inaccurate/missing reference data, where an appropriate estimate to confirm the measurement at the target station cannot be proposed, target measurement are deemed suspicious, requiring further processing and human intervention. As such, depending on the outcomes of the quality control procedure, data is classified as "Valid", "Corrected", "Infilled", or "Suspicious".

The ANN was developed using MATLAB (MathWorks, Inc., MA, USA). Feed-forward neural networks with backpropagation comprising of three layers (one input, one hidden, and one output layer) consisting of ten neurons in the hidden layer were used for the training and validation of data (see Figure 2). Feed-forward, backpropagation networks are frequently used elsewhere for the analysis of precipitation data [4,16,19-21] and is reported to be the most common type of ANN models used in water resources applications [17]. The Levenberg–Marquardt algorithm was used to train the networks owing to its fast convergence capability and satisfactory implementation by various researchers in the water resources arena [16,22-24]. Sigmoid (Logsig) and linear (Purelin) transfer functions were used in the hidden and output layers, respectively. R (correlation coefficient), RMSE (root mean square error) and mean absolute error were used to evaluate the performance of the model, and a sigmoid function was employed as the transfer function.

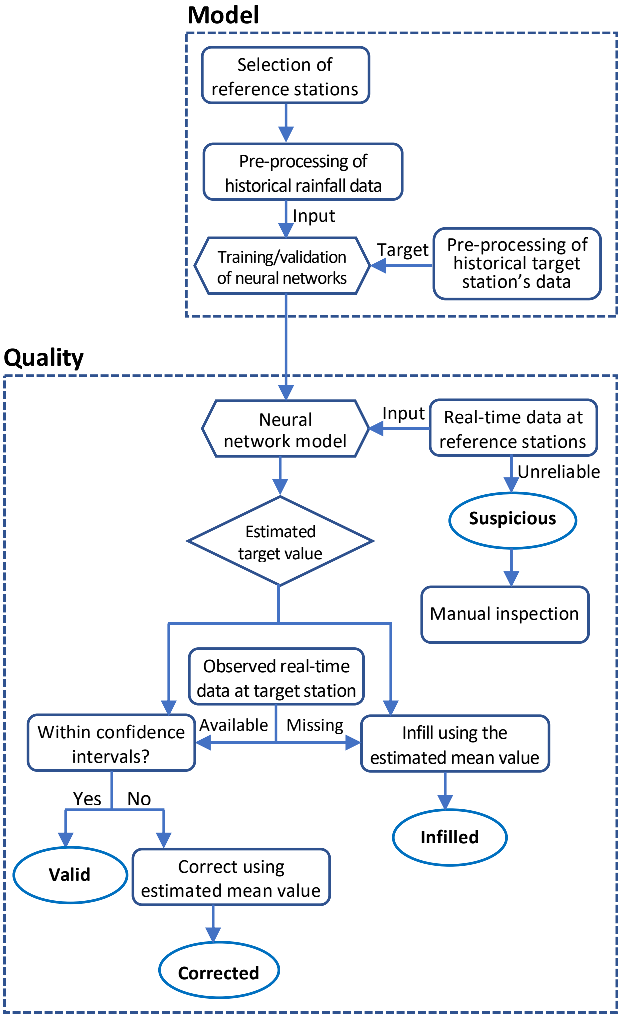
Model
Selection of reference stations
Pre-processing of historical rainfall data
Input
Training/validation of neural networks
Target
Pre-processing of historical target station's data
Quality
Neural network model
Input
Real-time data at reference stations
Unreliable
Suspicious
Manual inspection
Estimated target value
Observed real-time data at target station
Available
Missing
Within confidence intervals?
Infill using the estimated mean value
Yes
No
Valid
Correct using estimated mean value
Infilled
Corrected

Figure 1. Schematic of the automatic real-time quality control algorithm

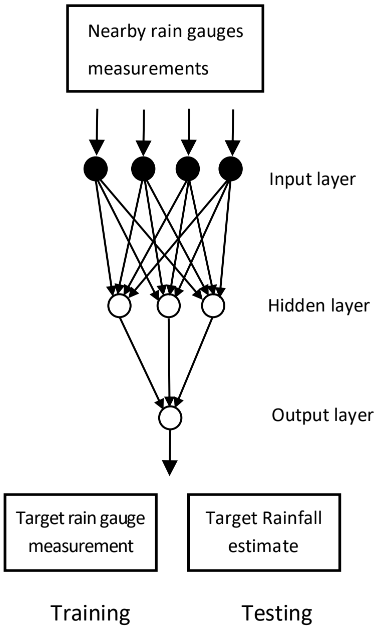

Figure 2. Topology of ANN model with three layers deployed for training/testing

Data pre-processing plays an important role in determining the performance and efficiency of the ANN models [16,18,24,25]. To ensure that model predictions are not influenced by the magnitude of data which are temporally and spatially distributed, rain gauge and data can be normalized into dimensionless values in the range of 0-1 using the following formula:

$$I_n = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \quad (1)$$

where, I, I_n, I_{min}, and I_{max} are, respectively, an input value, the respective normalized value, minimum value, and maximum value within a dataset. Suitable criteria (keeping the ratio of training samples to free variables greater than 30, stopping the training process as soon as the MSE is minimized, and the selection of appropriate values for rainfall threshold and inter-event periods) were also incorporated to prevent model overfitting, which can result in poor ANN model performance [19,26,27].

6 Case Study and Results

The QC algorithm developed in this research project was applied to a timeseries of precipitation depth measurements in 5-min intervals observed by selected rain gauges within the City of Toronto operated by TRCA. A number of available rain gauges within the City of Toronto are illustrated in Figure 3. Selected reference stations (HY003, HY016, HY027) were used for the QC of rainfall data at the target station (HY008). The specifications of gauges are listed in Table 1. Reference stations are selected on the bases of the availability of data, proximity, and elevation. Historical dataset ranging from April 2012 to December 2015 were used for modelling purposes. In order to eliminate redundant zero rainfall values in the input set, these data were consolidated by separating 5-min rainfall pulses using PCSWMM software (Computational Hydraulics International, Canada). As reported elsewhere [11,22], data aggregation can assist with reducing data size and noise. Therefore, in addition to including 5-min increments, in a separate set of modeling, data aggregated to hourly totals were employed. Datasets comprising of about 8,500 and 1,950 points were used in cases of 5-min and hourly ANN models, respectively. Seventy percent of data (in sequence) was used for ANNs training, whereas fifteen percent of data was used for validation phase and the remaining fifteen percent were used for testing purposes. Figure 4 illustrates the correlation between cumulative rainfall values derived using rainfall measurements in the input data set and respective model estimation at the target station (HY008). As it can be observed, model estimates closely correlate to actual measurements. Modeling results indicate that assuming 95% confidence intervals, 96.9% and 95.6% of data were automatically validated using 5-min and hourly ANN models, respectively. RMSE and mean absolute error were respectively calculated to be 0.38 mm and 0.19 mm for the 5-min data. These parameters in case of using hourly totals were measured at 0.92 mm and 0.40 mm. Figure 5 indicates ANN models performances for 5-min and hourly totals during modeling segments including training, validation, and testing. As it is evident the overall models R value were 0.78 and 0.93 for the 5-min and hourly data.

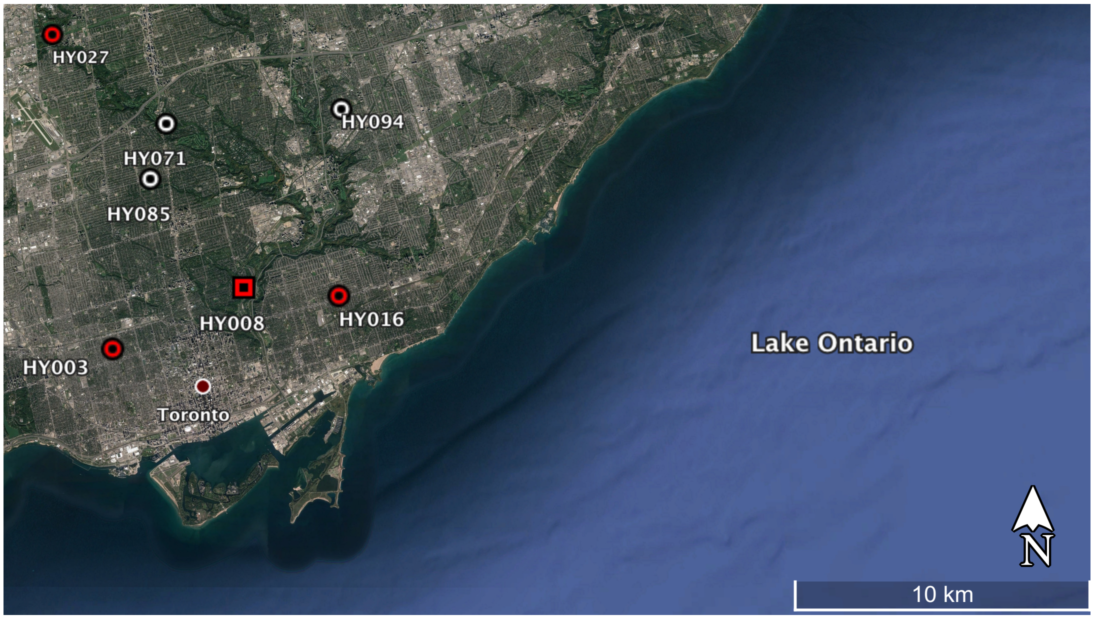

Figure 3. The position of the target rain gauges (square) and respective reference stations (circles)

Table 1. Specifications of rain gauge stations

Station ID	Operation period	Equipment type	Elevation (masl)
HY003	3-season	Tipping bucket	109
HY008	4-season	Ott Pluvio2	78
HY016	3-season	Tipping bucket	115
HY027	3-season	Tipping bucket	175
HY071	3-season	Tipping bucket	139
HY085	3-season	Tipping bucket	169
HY094	3-season	TB3	152

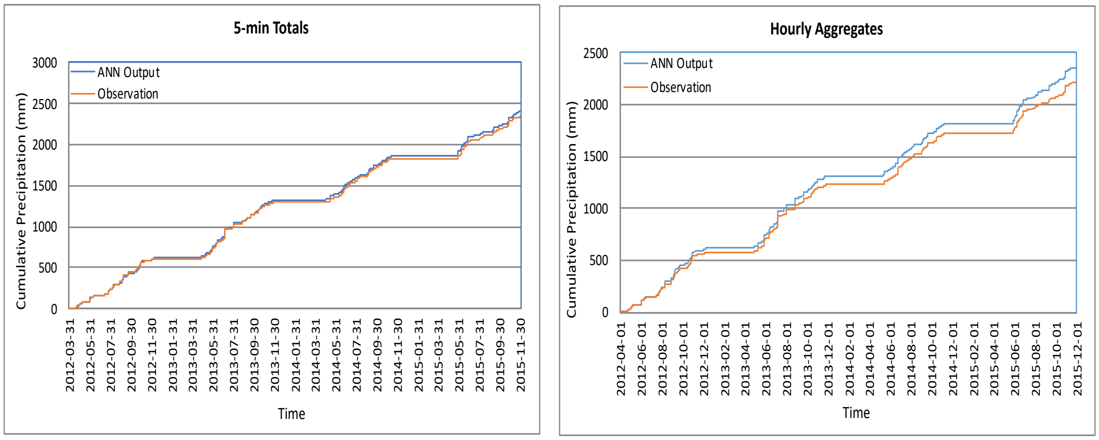

Figure 4. Correlation between observed and estimated data at the reference station, HY008. Left: 5-min totals, Right: Hourly aggregates

7 Conclusion

This study revealed that neural networks can be employed as an effective tool for the quality control and infilling of rainfall data. As discussed above, the proposed ANN algorithm was able to verify up to about 97% of data using 95% confidence intervals. It is believed that the developed methodology can provide a valuable tool for water practitioners engaged in small-scale rainfall/discharge monitoring studies, providing increased efficiency and enabling them to maximize the outcomes of such studies. It is believed that the performance of ANN model can be improved through integration of available radar data into rainfall QC procedure for the selection of reference stations (on the basis of storm direction) and also as an additional input parameter into the model and for the cross checking of estimated data. Moreover, conventional QC approaches such as tests for threshold/maximum values, flat line, spike, and rate of change can be incorporated into the ANN algorithm in order to enhance the performance and accuracy of the model.

8 Acknowledgement

The authors acknowledge the National Sciences and Engineering Research Council of Canada (NSERC) for financial support, Civica Infrastructure Inc. for project support and TRCA for providing precipitation data used in the study.

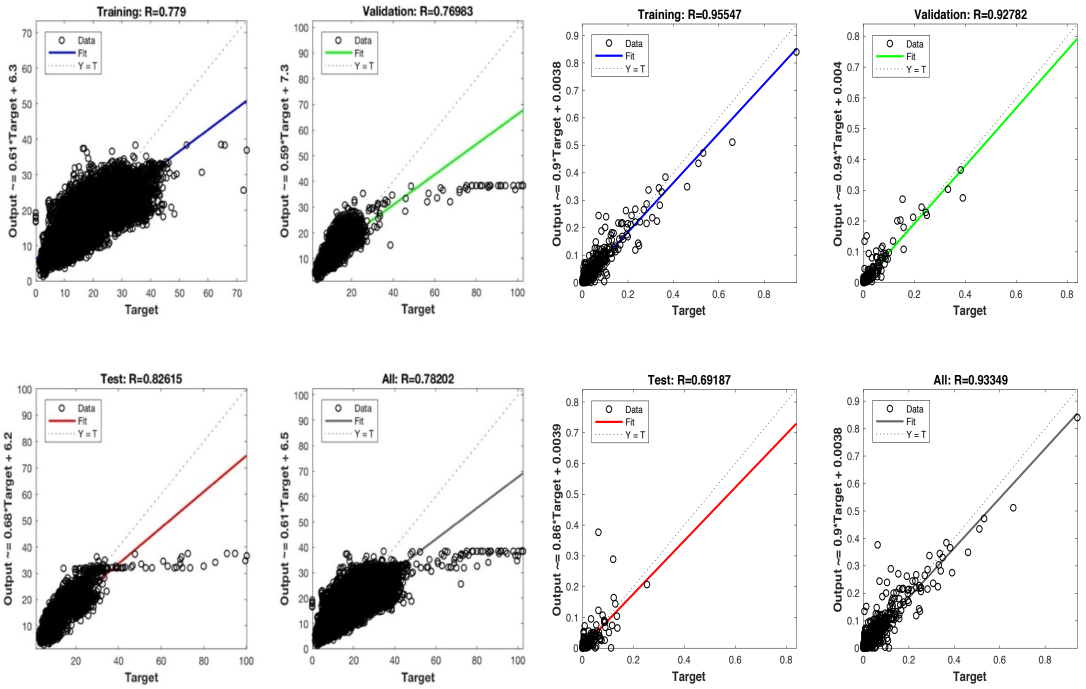

Figure 5. ANN models performance. Left: 5-min totals, Right: Hourly aggregates (using normalized data)

References

[1] Golz, C., Einfalta, T., Gabella, M., Germann, U., 2005. Quality control algorithms for rainfall measurements. Atmospheric Research, 77, 247-255.

[2] Stellman, K.M., Fuelberg, H.E., Garza, R., Mullusky, M., 2001. An Examination of Radar and Rain Gauge–Derived Mean Areal Precipitation over Georgia Watersheds. Weather and Forecasting, 16(1), 133-144.

[3] Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., Grams, H., Wang, Y., Cocks, S., Martinaitis, S., 2016. Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. Bull. Am. Meteorol. Soc., 97, 388–393.

[4] Sciuto, G., Bonaccorso, B., Cancelliere, A., Rossi, G., 2009. Quality control of daily rainfall data with neural networks. Journal of Hydrology, 364, 13– 22.

[5] U.S. Integrated Ocean Observing System Program, 2016. Manual for Real-Time Quality Control of Water Level Data: a Guide to Quality Control and Quality Assurance for Water Level Observations. Silver Spring, MD.

[6] Wang, X., Cheng, Y., Wu, S., Zhang, K., 2016. An effective toolkit for the interpolation and gross error detection of GPS time series. Survey Review, 48(348), 202-211.

[7] Kondragunta, C.R., Shrestha, K., 2006. Automated Real-Time Operational Rain Gauge Quality-Control Tools in NWS Hydrologic Operations. In: Preprints, 20th AMS conference on hydrology.

[8] Zhong, L., Zhang, Z., Chenc, L., Yang, J., Zou, F., 2016. Application of the Doppler weather radar in real-time quality control of hourly gauge precipitation in eastern China. Atmospheric Research, 172–173, 109–118.

[9] Stanzani, R., Alberoni, P.P., Nanni, S., Mulazzani, C., Pasquali, A., 2000. Raingauge and C-Band Radar Monthly Rainfall Comparison in the Po Plain Area. Phys. Chem. Earth, 25(10-12), 981-984.

[10] Lumpe, M., 2015. Tests of Quantitative Precipitation Estimates Using National Weather Service Dual- Polarization Radar in Missouri. Master's Thesis, University of Missouri, US.

[11] Giustarini, L., Parisot, O., Ghoniem, M., Hostache, R., Trebs, I., Otjacques, B., 2016. A user-driven case-based reasoning tool for infilling missing values in daily mean river flow records. Environmental Modelling & Software, 82, 308-320.

[12] Corchardo, J.M., Lees, B., Fyfe, C., Rees, N., Aiken, J., 1998. Neuro-Adaptation Method for a Case-Based Reasoning System. IEEE International Joint Conference on Neural Networks Proceedings, 713-718.

[13] ElKafrawy, P., Mohamed, R.A., 2015. Comparative Study of Case Based Reasoning Software. International Journal of Scientific Research and Management Studies, 1(6), 224-233.

[14] Atanassov, A., Antonov, L., 2012. CComparative Analysis of Case Based Reasoning Software Frameworks jCOLIBRI and myCBR. Journal of the University of Chemical Technology and Metallurgy, 47(1), 83-90.

[15] Li, L., Li-na, Z., Li-wu, Z., 2012. Forecast system of pear scab management based on case-based reasoning and fuzzy ISODATA clustering. 24th Chinese Control and Decision Conference, 1701-1706.

[16] Maier, H.R., Dandy, G.C., 2000. Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications. Environmental Modelling & Software, 15, 101–124.

[17] May, D.B., Sivakumar, M., 2009. Prediction of urban stormwater quality using artificial neural networks. Environmental Modelling & Software, 24, 296–302.

[18] Xiao, R., Chandrasekar, V., 1997. Development of a Neural Network Based Algorithm for Rainfall Estimation from Radar Observations. IEEE Transactions on Geoscience and Remote Sensing, 35(1), 160-171.

[19] Sahoo, G.B., Ray, C., 2006. Flow forecasting for a Hawaii stream using rating curves and neural networks. Journal of Hydrology, 317, 63–80.

[20] Kim, J.W., Pachepsky, Y.A., 2010. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology, 293(3-4), 305-314.

[21] Nkuna, T.R., Odiyo, J.O., 2011. Filling of missing rainfall data in Luvuvhu River Catchment using artificial neural networks. Physics and Chemistry of the Earth, 36, 830-835.

[22] Sahoo, G.B., Ray, C., De Carlo, E.H., 2006. Use of neural network to predict flash flood and attendant water qualities of a mountainous stream on Oahu, Hawaii. Journal of Hydrology, 327, 525–538.

[23] Dutta, D., Sharma, S., Sen, G.K., Kannan, B.A.M., Venketswarlu, S., Gairola, R.M., Das, J., Viswanathan, G., 2011. An Artificial Neural Network based approach for estimation of rain intensity from spectral moments of a Doppler Weather Radar. Advances in Space Research, 47(11), 1949-1957.

[24] Roman, U.C., Patel, P.L., Porey, P.D., 2012. Prediction of missing rainfall data using conventional and artificial neural network techniques. ISH Journal of Hydraulic Engineering, 18(3), 224–231.

[25] Lazri, M., Ameur, S., Mohia, Y., 2014. Instantaneous rainfall estimation using neural network from multispectral observations of SEVIRI radiometer and its application in estimation of daily and monthly rainfall. Advances in Space Research, 53, 138–155.

[26] Amari, S.I., Murata, N., Muller, K.R., Finke, M., Yang, H.H., 1997. Asymptotic statistical theory of overtraining and cross-validation. IEEE Transactions on Neural Networks, 8(5), 985–996.

[27] Abbot, J., Marohasy, J., 2014. Input selection and optimisation for monthly rainfall forecasting in Queensland, Australia, using artificial neural networks. Atmospheric Research, 138, 166–178.