ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 10 , October 2018

Bad Data Analysis on Streamflow Forecasting Using Nonlinear Echo State Network

E. P. Wigand, S. Bahrami*

*Department of Hydrological Science, College of Science, University of Nevada, Reno
Department of Geography, University of Nevada, Reno

ABSTRACT:A correct estimation of the stream flow is crucial to reduce the consequences of flash floods. Hydrologic prediction or simulation, especially in ungauged basins, is essential for responsible and sustainable water resource management. In the current study, we develop a framework on a study area including twelve gauged watersheds spanning across different climatic settings in the US. In this work we will propose a novel approach of Nonlinear Echo State Network using Multivariable Polynomial (NESN-MP) to forecast daily stream flow in ungauged basin with bad data. This work aims to demonstrate the ability of NESN-MP to solve a simulation task in comparison with ANFIS. Publicly available climate and US Geological Survey streamflow records are used to train and test the model. The model inputs include time-lagged records of precipitation, solar radiation, day length, vapor pressure and temperature. Furthermore, recurrent feedback loops allow ANN streamflow estimates to be used as model inputs. The successful of these flow prediction approach indicates that the NESN-MP can predict streamflow with bad data entry as accurately as good data set entry in the basins on which they were trained.

KEYWORDS: Forecasting, nonlinear echo state network using multivariable polynomial (NESN-MP), Streamflow, Bad data, Ungauged-basins.

I. INTRODUCTION

State estimation and forecasting have always been general concerns for engineers. State estimation is applied in all energy management systems to identify the present operating state of a system [1-2]. Forecasting is also an important and necessary aid to planning and planning is the backbone of effective operations. In hydrology, streamflow forecasting is vital for water resources engineers, reservoir operators and water managers who strive to balance a range of competing objectives to support their decisions about hydroelectric power programming, flood mitigation, agricultural and domestic water supplies, irrigation management as well as maintenance of environmental flows.Accurate streamflow prediction and developing an optimal streamflow forecasting model, as a stochastic property of environmental modelling, is one of the most important component of watershed planning and sustainable water resource management[3].The streamflow is under influence of various factors such as evapotranspiration, rainfall, atmospheric circulation and temperature which makes its generation process nonlinear and time-varying. The magnitude and locality of extreme streamflow events due to climate change and anthropogenic factors can end up to damaged infrastructure, degraded surface water quality, loss of agricultural lands, phosphorus diffusion, and sediment pollutants[4].Therefore, Accurate and timely predictions of high and low streamflow events at either gauged or ungauged watershed will provide required information to make strategic decisions as fallowing; (1)Ensure sustainable watershed planning; (2) Define the dilution potential of catchments; (3) Set ecological streamflow limits; (4) Allocate water resources.

Due to poor data availability greatly compounds with accurately forecasting daily streamflow, water managers must rely on the streamflow estimates from various prediction models [5]. There are four different streamflow forecasting models: conceptual, metric, physics based, and data-driven. The first three mentioned models assume that the relation between the input and output series is linear or even near linear. They thus ignore the nonlinear information hidden in the streamflow series. In contrast to these models, data-driven methods focus on using nonlinear relation between inputs and outputs. However, they have some disadvantages including high complexity along with high processing time and high dependence on parameter tuning and optimization [6-7].

To overcome these drawbacks, application of the NESN-MP(called NESN in this paper) forecasting engine in stream flow forecasting is presented in [8-9]. It has been shown that this model works well for the circumstance that there is

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 10 , October 2018

precise observed stream flow data. However, there are many streams all over the world which do not have accurate observed streamflow data, or the data could exist only in a form that is extremely difficult to access while some other data should be kept secret due to policy concern which will produce bad data. Poor decision being made due to poor data. Therefore, reasonable forecasting of any hydrological process is the call of the time and valuable to responsible and sustainable water resources management. In fact, the National Research Council has noted growing attention to minimize the impacts of bad data among stakeholders for uncertainty assessments of hydrologic prediction [10] which can be because of ungauged basins, potentially inaccurate measurement, incomplete data collection, uncertain estimate, "fat-fingered" data entry, policy concerns, mis-categorization, etc[11].

There are some methodological studies for predicting streamflow response in ungauged basins with bad data which utilized deterministic physically based models to calculate streamflow. They performed based on distributed hydrologic parameters, and statistical regionalization which uses regression models to transfer hydrologic information from gauged to ungauged basins. The distributed hydrologic parameters approach, focusses on dispersing errors into measurement, parameter and structural uncertainty, the produced uncertainties are then disseminated toward model output. The statistical regionalization, is a challenging task in hydrological science [12] due to poor streamflow data, which is normally calibrated [13]. Moreover, the obtained results have been usually examined on different basins, while every catchment characteristics is different from one case to another [14]. Subsequently, there is no universal method for regionalization. While this is a broadly accepted procedure, uniqueness of the watersheds and the obscurity of parameters bring major uncertainty in the ungauged basins'simulations.

As an inherent symptom of any modelling task, all hydrologic models will suffer from some degree of uncertainty [15]. Movement away from methods grounded in traditional statistics toward conceptual, process-based models has blurred our understanding of model uncertainty to the point that most models are considered as almost purely deterministic tools [16]. Qamar et.al. [3], use non-parametric distance-based method to assess streamflow duration curve in ungauged basins. Their work acquires a more robust model with better global performance even if the extension of the selected model to the whole workspace may be less optimal [17]. Given the hydrological process complexity, using an adaptation of globalized/ regionalized uncertainty is optimal [18].

Some international organizations such as the United Nations Development program (UNDP) and World Bank are concerning to generate a precise approach for development, and management of freshwater recourses. Artificial intelligence (AI) methods are recent developments in several hydrological areas due to their ability to incorporate a tried-and-true model with no need to prior knowledge of the existing functional or nonlinear relationship between input and output [19] One of the most common AI methods to predict stream-flows in ungauged catchments is to identify the train model with homogenous nearby basins to forecast the stream-flow with different climate input[20]. Shu and Ouarda (2008) [21]considered the homogeneous region characteristics to find similar hydrological sites forpredicting flood quantile at ungauged basins; their result showed that the ANFIS approach had more capability compared with the other techniques examined in general, however in sites under 1000 m3/s flood quantile, ANN yields better results. Chang Shian Chen et al. (2010) [22] tried to employ the available hydrological record of nearby catchments with similar homogenous characteristics to estimate ungauged catchments. They concluded the temporal distribution and spatial characteristics considered in the model, reflect most of the behaviour of rainfall–runoff in nature.

A method of random forest models and an ensemble of artificial neural networks, has been used to predict several components of streamflow [23]. Some researchers used regression trees and model tree ensembles to predict a complete flow-duration curve (FDC) for streams, [24]. Senent-Aparicio et al[25] Combined machine learning with Soil and Water Assessment Tool(SWAT) to estimate instantaneous peak flow (IPF) in areas where sub-daily observational data are scarce. The results of this study can contribute to superior ability of extreme learning machine (ELM) to estimate IPF, thereby reducing uncertainties associated with IPF estimations.All previous studies for ungauged estimation have applied the homogeneous nearby basin parameters which result in inaccurate results Since every catchment is unique in its characteristics hence a direct transfer of model parameter values from gauged to ungauged basins may not be appropriate. Therefore,there is yet a need to produce a more accurate estimation of daily streamflow at ungauged basins [26].

However, all These studies granted prized baseline application of machine learning to streamflow perdition, their model performance could not be compared due to one unique accurate data set used for every individual research. To

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 10 , October 2018

circumvent the above challenges, we focus this paper on developing a novel method on streamflow forecasting with bad data set input. the primary objective of this research is developing a model to yield valuable estimation, in (1) problems corrupted by noise (2), complex systems that, may not be dittoed, and (3) circumstances where input is incomplete or ambiguous by nature producing bad data.

In this proposed method, the model employs the concepts of ANN in an iterative procedure to produce bad data set. The generated bad data set derived from original accurate data set of the gauged basins, are then applied to develop a new input. Subsequently, this new input is used for evaluating the model applicability, which in turn is used for generating ensemble simulations in the ungauged basin. To test the generality of the method, twelve different watersheds across the United States are considered. While all the basins considered in this study were gauged with precise data input, the current study assumed some basins to be un-gauged or producing bad data to evaluate the effectiveness of the proposed methodology. This algorithm will always converge, with no need to stochastic training, and is also applicable to any ungauged basins. Recurrent feedback loops are added to this algorithm, allowing future predictions to be based on time-lagged predictions not time-lagged measurements. To evaluate the effectiveness of the proposed methodology in ungauged basin prediction, we compare our result with ANFIS. The process was repeated by considering representative basins from different climatic and land use scenarios as ungauged. The results of the study indicated that the ensemble simulations in the ungauged basins with NESN were closely matching with the observed streamflow and yield better result comparing to ANFIS. The remainder of this study is as follows. Section II provides an overview of the NESN. Simulation results and discussion are given in Section III, and conclusions are summarized in Section IV.

II. NONLINEAR ECHO STATE NETWORK

In the most of practical circumstances, where the main concern is generating accurate predictions with no insight on the internal structure of the process involved, the authors believe NESN approaches can provide appropriate and accurate solutions.As it has been pointed out in the literature, this novel method is easy, effective, with less computations [27].

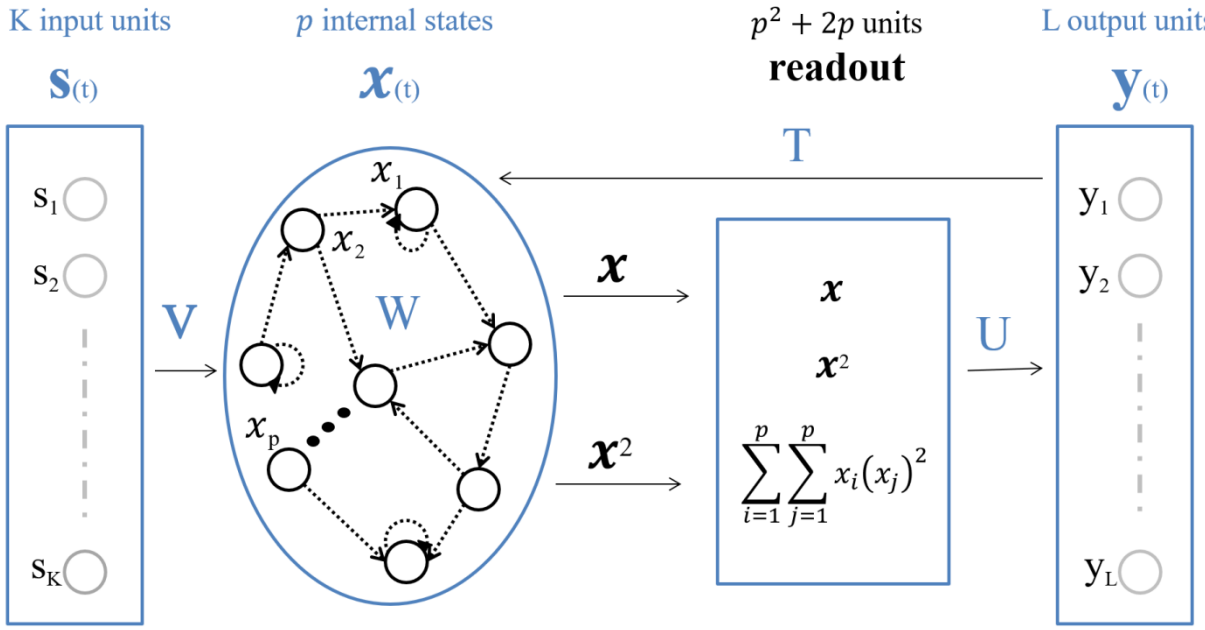

Fig. 1. Schematic of NESN.

NESN provides a total of $2p + p^2$ units;$p$internal states; p squares of the internal states; and p^2 units gained by multiplying the internal states and squares of the internal states. This process will minimize the order of weight

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 10 , October 2018

matrices radically. The weight matrices (W, T, and V) are then applied to calculate the internal states of the reservoir. The vector of internal states is updated using

$$\boldsymbol{x}_{(t+1)} = f(W.\boldsymbol{x}_{(t)} + V.\boldsymbol{s}_{(t+1)} + T.\boldsymbol{y}_{(t)}) \quad (1)$$

and the readout vector is

$$\bar{\boldsymbol{x}}_{(t+1)} = [\boldsymbol{x}_{(t+1)}, \boldsymbol{x}^2{}_{(t+1)}, \sum_{i_1=1}^{p}\sum_{i_2=1}^{p} \boldsymbol{x}_{i_1(t+1)}.\boldsymbol{x}^2{}_{i_2(t+1)}] \quad (2)$$

where $\boldsymbol{x}^2{}_{(t+1)} = \left[x^2_{1(t+1)}, x^2_{2(t+1)}, \dots, x^2_{p(t+1)}\right]$, p is the number of internal states $\left[\frac{N}{p+2}\right]$, $\boldsymbol{s} \in R^{K\times 1}$ is the input vector,$\boldsymbol{x} \in R^{p\times 1}$ is the internal state vector, $\bar{\boldsymbol{x}} \in R^{(p^2+2p)\times 1}$ is the readout vector, and$\boldsymbol{y} \in R^{L\times 1}$ denotes the output states.

The matrix $W \in R^{p\times p}$ defines the internal state interconnections within the reservoir. The values in W are fixed values generated randomly over a symmetric interval.

$$W = (w_{ij})_{p\times p} \; ; \; w_{ij} \in (-1,1)(i,j = 1,2,\dots,p) \quad (3)$$

Matrix $V \in R^{p\times K}$, containing randomly chosen fixed values, defines the connections of the input with the internal states of the reservoir.

$$V = (v_{ij})_{p\times k} \; ; v_{ij} \in (-1,1)(i = 1,2,\dots,p, j = 1,2,\dots,k) \quad (4)$$

The output feedback matrix,$T \in R^{p\times L}$ is

$$T = (t_{ij})_{p\times L} \; ; \; t_{ij} \in (-1,1)(i = 1,2,\dots,p, j = 1,2,\dots,L) \quad (5)$$

The output matrix,$U \in R^{L\times(p^2+2p)}$ is

$$U = (u_{ij})_{L\times(p^2+2p)} \; ; \; u_{ij} \in (-1,1)(i = 1,2,\dots,L, j = 1,2,\dots,2p+p^2) \quad (6)$$

where K is the number of inputs, p is the number of internal states, and L is the number of outputs.

III. SIMULATION RESULTS

The performance of the NESN in presence of bad datais tested using climate (day length, precipitation, solar radiation, maximum and minimum temperature per day, and vapor pressure)and US Geological Survey streamflow datawith a time interval of 24 hours used to train and test the models.

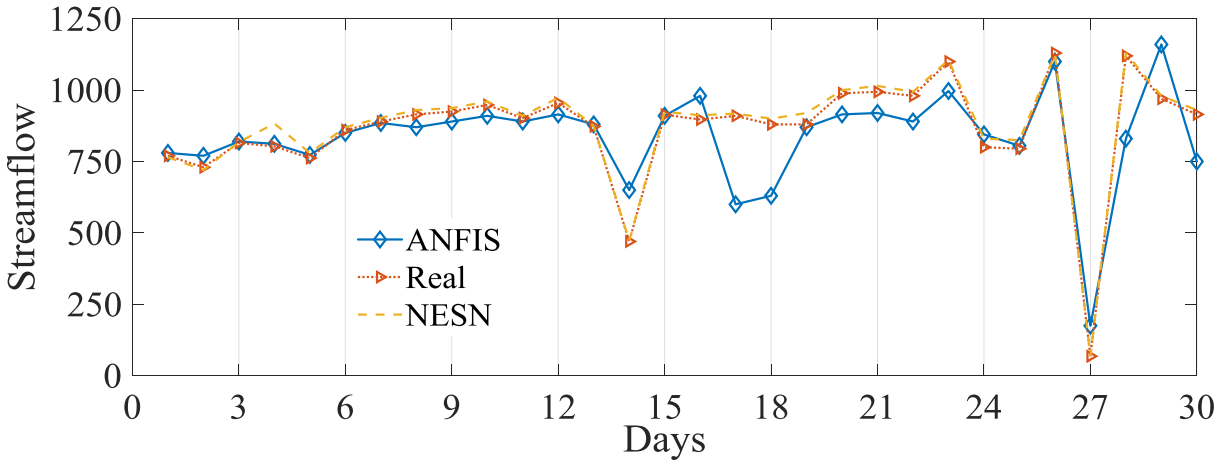

Fig. 2. Streamflow forecasting without bad data.

ISSN: 2350-0328
International Journal of AdvancedResearch in Science, Engineering and Technology
Vol. 5, Issue 10 , October 2018

For the purpose of testing and training, each data set has been divided into two separate parts with their lengths denoted as L_{train} and L_{test}, respectively. To evaluate the performance of the proposed methods the MSE, root mean squared error (RMSE), normalized root-mean-square error (NRMSE), normalized mean-absolute error (NMAE), and mean absolute error (MAE) have been compared. The streamflow forecasting is carried out for 67 days ahead. $l_{train} = 200$, $l_{test} = 30$ with no overlap and with the test data starting immediately after the training data. Fig. 2shows the prediction for 30 days ahead for NESN and ANFIS without bad data.

To validate the performance of the proposed method in presence of bad data, severe changes have been made in the input data. The changes vary between 10% and 100% of the initial values. Fig. 3 and Fig. 4 show the comparison between the streamflow forecasting with and without the bad data for NESN and ANFIS respectively. Table 1 also shows the error indices for different forecasting results shown in Fig. 3 and Fig. 4.

It is shown that the NESN provides the MAE of 15 and 20 with and without bad data which are80% and 76.7% below those for ANFIS, respectively. In case of RMSE, NESN gives the respective values of 21 and 26 which are well below the RMSE of 115 and 144 for ANFIS with and without bad data respectively. It is shown that the bad data provide uneven impact on the prediction. Therefore, the changes in the predicted results vary during different days.

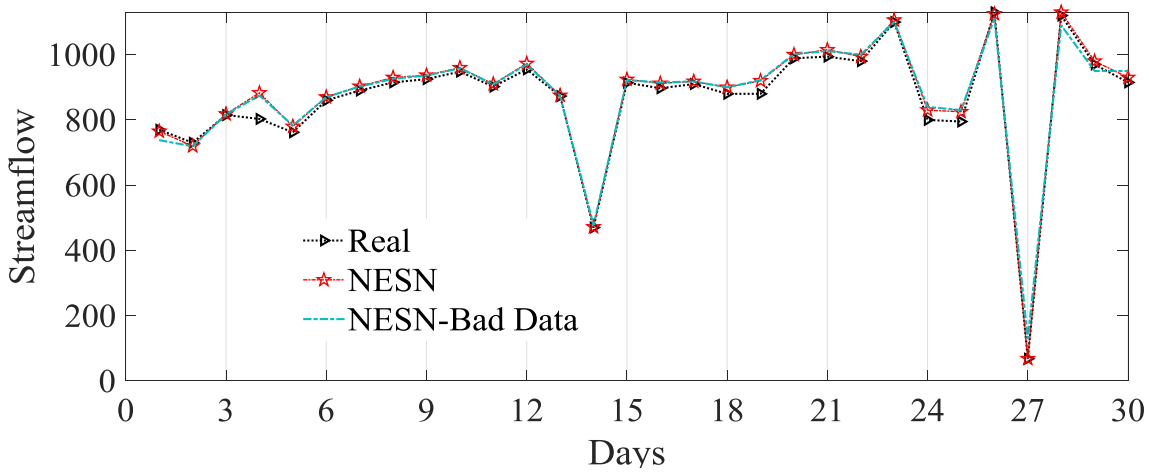

Fig 3. The performance of the NESN in presence of bad data in streamflow forecasting.

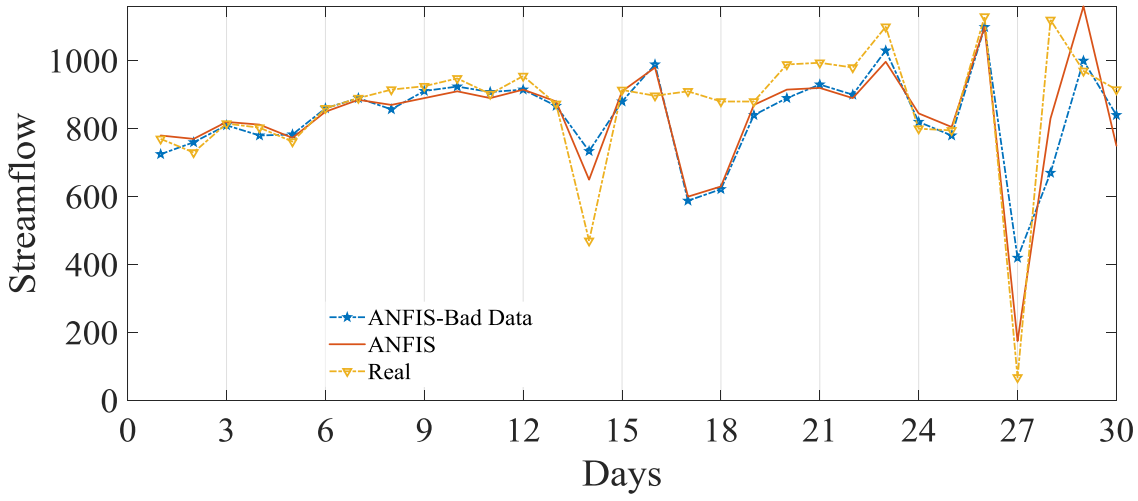

Fig 4. The performance of the ANFIS in presence of bad data in streamflow forecasting.

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 10 , October 2018

Table 1. Error indices for ANFIS and NESN with and without bad data.

	ANFIS	ANFIS-Bad Data	NESN	NESN-Bad Data
MSE	13270	20687	449	690
RMSE	115	144	21	26
NRMSE	0.595	0.743	0.110	0.136
MAE	76	86	15	20
NMAE	0.067	0.076	0.013	0.018

IV. CONCLUSION

This paper addressed the task of predicting daily stream flows with bad data input for water resource purposes, comparing NESN to ANFIS. The purpose of this study was twofold: (1) we aimed to find an effective ANN procedure able to predict mean daily streamflow with bad data input, and (2) we highlighted pros and cons of the two different modelling approaches. Results confirm that NESN can stand the comparison with a ANFIS procedure, producing good performances if correctly trained and appropriately supplied with a good amount of well-chosen information. NESN provides significantly lower values than those given by ANFIS for MAE, NMAE, MSE, RMSE, and NRMSE. Thus, it can be considered as a powerful tool for predicting stream flow events, even allowing for their lack of physical interpretability. In fact, as long as the user is interested in forecasting the streamflow (e.g. filling in missing data) NESN seem to be a useful option; however, if a physical interpretation of the process is needed, then the parsimonious conceptual/ANN models would be preferred, given also the less time necessary for calibration. Future work will run a sensitivity analysis to explore the most important affecting factor on streamflow forecasting in different circumstance of good data input and bad data input.

REFERENCES

[1] M. A. Chitsazan, M. S. Fadali, A. M Trzynadlowski, “State estimation of IEEE 14 bus with unified interphase power controller (UIPC) using WLS method”, Energy Conversion Congress and Exposition (ECCE), 2017 IEEE, pp. 2903-2908, Oct. 2017.

[2] M. A. Chitsazan, A. M Trzynadlowski, “State estimation of IEEE 14 bus with interphase power controller using WLS method”, Energy Conversion Congress and Exposition (ECCE), 2016 IEEE, pp. 1-5, Sep. 2016.

[3] M.U. Qamar, M.Azmat, M.J.M. Cheema, M.A. Shahid, R.A. Khushnood, S.Ahmad, “A comparative performance signature for the prediction of flow duration curves in ungauged basins”, 2016, J. Hydrol, pg no : 1030-1041

[4] N. Arnell, “impacts, adaptation and vulnerability: hydrology and water resources”, 2001, United Nations Environmental Program, Intergovernmental Panel on Climate Change.

[5] K.E. Kapo, K. McDonough, T. Federle, Scott. Dyer, R. Vamshi,“ Mixing zone and drinking water intake dilution factor and wastewater generation distributions to enable probabilistic assessment of down-the-drain consumer product chemicals in the U.S.”, 2015, journal of Science of the Total Environment, pg no : 302–309

[6] T. Peng, J. Zhou, C. Zhang, W.Fu, “Streamflow Forecasting Using Empirical Wavelet Transform and Artificial Neural Networks”, 2017, journal of Water.

[7] A. Jakeman, I. Littlewood, P. Whitehead, “Computation of the instantaneous unit hydrograph and identifiable component flows with application to two small upland catchments”, 1990,Journal of Hydrology, pg no : 275–300.

[8] M. A. Chitsazan, M. Sami Fadali, A. M Trzynadlowski, “Wind speed and wind direction forecasting using echo state network with nonlinear functions”, Renewable Energy, 2019, pg no :879-889.

[9] S. Bahrami, P.E. Wigand, “Daily Streamflow Forecasting Using Nonlinear Echo State Network of advance science in research, engineering and technology”, 2018, International Journal of Advanced Research in Sience, Engineerng and T echnology, pg 3619-3626.

[10] J. Doherty, J.M. Johnston,“ Methodologies for Calibration and Predictive Analysis of a Watershed Model.”, 2003, Journal of the American Water Resources Association, 39(2), pg no :251-265

[11] W.H. Farmer , Sara Levin, “ Characterizing Uncertainty in Daily Streamflow Estimates at Ungauged Locations for the Massachusetts Sustainable Yield Estimator”, 2018, jornal of hydrol, pg no :198-210

[12] J. Samuel, P. Coulibaly, R. A. Metcalfe,“Estimation of Continuous Streamflow in Ontario Ungauged Basins: Comparison of Regionalization Methods”, 2011, J. Hydrol, pg no : 447-459

[13] M. Sivapalanet, “IAHS decade on predictions in ungauged basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences”,2010,Journal of Hydrological Sciences – Journal Des Sciences Hydrologiques, pg no : 857–880.

[14] L. Oudin, V.Andreassian, C.Perrin, C.Michel,N.LE. Moine, “Spatial proximity , physical similarity, regression and ungauged catchment: A comprision of regionalization approaches based on 913 French catchment”, 2008, Journal of Water Resource Researc, W03413.

[15] W.H. Farmer, R.M. Vogel, “the Deterministic and Stochastic Use of Hydrologic Models”, 2016, Journal of Water Resources Research ,pg no :5619-5633.

[16] T.Wagener, H.S. Wheater,“Parameter Estimation and Regionalization of Continuous Rainfall-Runoff Models Including Uncertainty”,2005,Journal of Hydrology, pg no :132-154

[17] G. Cybenkot, “Approximation by Superpositions of a Sigmoidal Function”, 1989, Journal of Math. Control Signals Systems,pg no :303-314

ISSN: 2350-0328

International Journal of AdvancedResearch in Science, Engineering and Technology

Vol. 5, Issue 10 , October 2018

[18] Z.M.Yaseen, M. Fu, C.Wang, W.Hanna, M.Wan, R.C. Deo, A.El-shafie, "Application of the Hybrid Artificial Neural Network Coupled with Rolling Mechanism and Grey Model Algorithms for Streamflow Forecasting Over Multiple Time Horizons",2018, journal of Water Res, pg no : 1883–1899

[19] D.J.Booker,R.A.Woods, "Comparing and combining physically-based and empirically-based approaches for estimating the hydrology of ungauged catchments", 2014, Journal of Hydrology, pg no : 227-239

[20] N. Valizadeh, M.Mirzaei, M. Falah Allaw, " Artificial intelligence and geo-statistical models for stream-flow forecasting in ungauged stations: state of the art", 2017, Journal of Natural Hazards, Pg no: 1377-1392

[21] C. Shu, T. Ouarda, "Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system", 2008, J Hydrol, pg no :31–43

[22] C. S. Chen, F. N. F. Chou, B. P. T. Chen," Spatial information-based back-propagation neural network modeling for outflow estimation of ungauged catchment", 2010, journal of Water Resour Manag ,pg no :4175–4197

[23] S. C.Worland,W. H.Farmer,J. E.Kiangc, "Improving predictions of hydrological low-flow indices in ungaged basins using machine learning", 2018, Environmental Modelling & Software, pg no : 169-182.

[24] S. Schnier, X. Cai, "Prediction of regional streamflow frequency using model tree ensembles",2014, Journal of Hydrology, pg no: 298-309.

[25] J. Senent-Aparicio, P. Jimeno-Sáez, A.Bueno-Crespo, J. Pérez-Sánchez, D. Pulido-Velázquez, "Coupling machine-learning techniques with SWAT model for instantaneous peak flow prediction",2018, Journalof Biosystems Engineering,

[26] W.Buytaert, K. Beven , "Regionalization as a learning process", 2009, journal of Water Resource.

[27] M. A. Chitsazan, M. Sami Fadali, Amanda K. Nelson, A. M Trzynadlowski, "Wind speed forecasting using an echo state network with nonlinear output functions", American Control Conference (ACC), 2017 IEEE, pp. 5306-5311, May. 2017.