Journal of Hydrology 510 (2014) 103–109

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

# A new quality control procedure based on non-linear autoregressive neural network for validating raw river stage data

M. López-Lineros [a,*], J. Estévez [b], J.V. Giráldez [c,d], A. Madueño [e]

[a] Design Engineering Department, University of Sevilla, C/Virgen de África, 7, 41011 Sevilla, Spain
[b] Projects Engineering Area, University of Córdoba, Cra Madrid km 396, 14071 Córdoba, Spain
[c] Agronomy Department, University of Córdoba, Cra Madrid km 396, 14071 Córdoba, Spain
[d] Agronomy Department, IAS, CSIC, Alameda del Obispo, 14080 Córdoba, Spain
[e] Aerospatial Engineering and Fluids Mechanics Department, University of Sevilla, Ctra, de Utrera km 1, 41013 Sevilla, Spain

ARTICLE INFO

*Article history:*
Received 30 July 2013
Received in revised form 10 December 2013
Accepted 12 December 2013
Available online 20 December 2013
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Sheng Yue, Associate Editor

*Keywords:*
River stage data
Validation
Quality control
Non-linear autoregressive neural networks

SUMMARY

The main purpose of this work is the develop of a new quality control method based on non-linear autoregressive neural networks (NARNN) for validating hydrological information, more specifically of 10-min river stage data, for automatic detection of incorrect records. To assess the effectiveness of this new approach, a comparison with adapted conventional validation tests extensively used for hydro-meteorological data was carried out. Different parameters of NARNN and their stability were also analyzed in order to select the most appropriate configuration for obtaining the optimal performance. A set of errors of different magnitudes was artificially introduced into the dataset to evaluate detection efficiency. The NARNN method detected more than 90% of altered records, when the magnitude of error introduced was very high, while conventional tests detected only around 13%. In addition, the NARNN method maintained a similar efficiency at the intermediate and lower error ratios, while the conventional tests were not able to detect more than 6% of erroneous data.



## 1. Introduction

Quality control is a major prerequisite for using hydro-meteorological information. High quality databases are essential tools for scientists, engineers, and planners alike (Shafer et al., 2000). Validation of meteorological data ensures the quality of the information, identifying incorrect values and detecting problems that require immediate maintenance attention (Estévez et al., 2011). The application of quality assurance procedures is especially necessary to provide hydro-meteorological data in real time or near real time for different purposes (WMO, 2008).

The hydro-meteorological information acquired on monitoring networks frequently contains erroneous data (Sciuto et al., 2009), restraining its subsequent use (Madsen, 1989). Therefore, raw data must be revised by qualified personnel. Nevertheless reliable, objective, and fast methods are required for flagging potentially erroneous data (Abbott, 1986; Reek et al., 1992).Many procedures have been developed for quality assurance of hydro-meteorological variables such as precipitation or other input climate variables (temperature, relative humidity, solar radiation, wind speed) for a reference evapotranspiration equation (Durre et al., 2010; Feng et al., 2004; Gandin, 1988; Geiger et al., 2002; Hubbard et al., 2005; Kunkel et al., 2005; Shafer et al., 2000; You et al., 2007). However, literature related to river stage data quality control is scant.

In order to develop a data validation model, raw data must be sorted by measuring values of certain variables sequentially in time (Koskela et al., 2003). Data sets are usually incomplete and consist of mixed signal and noise. In addition to this, in most cases, the underlying hydrological process is assumed to be stochastic which makes signal identification harder. The goal in model building is to reveal the underlying process from the data. Models are estimated through statistical methods to find regularities and dependencies existing in the data.

Several computational techniques have been proposed to gain more insight into the processes and phenomena which contain temporal information. Statistical methods based on linear and nonlinear models have been effective in many applications (Gershenfeld and Weigend, 1993). Among these methods, linear regression autoregressive (AR) and autoregressive moving average models (ARMA) (Box et al., 1994) have been the ones most accepted since the original contribution reported by Yule (1927). Linear models are well known, and many algorithms for model building are available.

* Corresponding author. Tel.: +34 637215844.
*E-mail address:* mlopezlineros@us.es (M. López-Lineros).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.12.026

Nevertheless, most of the intervening processes are, to some extent, nonlinear, questioning the adequacy of linear models. Nonlinear methods became widely applicable in the 1980s with the growth of computer processing speed and data storage. Among the nonlinear methods, Artificial Neural Networks (hereafter ANN) soon became very popular.

The ANN model was inspired by how the human brain works. The success of ANN models lies in their ability to approximate Borel-measurable functions to any degree of accuracy, as indicated by Hornik et al. (1989). ANNs became very useful in temporal sequence prediction due to their learning capacity of nonlinear dependencies from a large volume of potentially noisy data. In most cases, neural network prediction models give a better performance than other similar models (Hen and Hwang, 2002).

Elman (1990) warned of the difficulties in finding a global optimal way for representing the effect of time. Consequently, a vast number of different neural network architectures have been proposed in an attempt to capture the time context in the data.

In time series prediction with neural networks, the main problems have usually been the selection of the length of the input vectors and the actual structure of the network. These problems are similar in all neural architectures. In addition to this, with the recurrent networks, the stability of the model and the learning algorithm must be considered.

The first step in the design of data quality control tools is an analysis of the main error sources. This step requires the establishment of some standards to assess the quality of the data including the successive manipulations from the raw data acquired by field sensors. In the next step, control levels must be set for a periodic revision of the data. Estévez et al. (2011) proposed a progressive application of conventional tests for controlling hydro-meteorological information. A different approach was based on statistical decisions (Hubbard et al., 2005), computing data thresholds for different time periods and based on: (i) rate of change, (ii) temporal consistency,(iii) seasonal persistence, and (iv) spatial consistency. Alternatively, the data can be arranged in different categories depending on the chosen quality control method (Hasu and Aaltone, 2011), or the introduction of data assimilation methods (McLaughlin, 2002), based on interpolation, smoothing, and filtering techniques.

Neural network methods are common hydrological tools either based on individual networks to estimate flow rates in ungauged basins, or coupled to other methods (Hong, 2012) such as hybrid neural networks with Kalman filters (MLP-EFFQ), within recursive algorithms for hydrological models, or in the analysis of time series with neural networks and fuzzy logic (Lohani et al., 2012).

The main objective of this work was the comparison of two methods to estimate the fraction of the total detected errors in a raw data series artificially modified with the introduction of errors generated with a random simulator. The two methods are:

(a) A conventional method commonly used in the hydro-meteorological field (Estévez et al., 2011).
(b) A recursive nonlinear autoregressive neural network (NARNN) algorithm trained on a previously validated data series (Shaw et al., 1997; Ruano, 2005).

## 2. Materials and methods

### 2.1. Source of data

In this study, river stage data from one gauge station (Villoldo), located in the basin of the River Carrión, an affluent of the Duero River in the northwest of the Iberian Peninsula (Fig. 1), have been used. This station belongs to the Spain National Gauging Network (ROEA), controlled by the Confederación Hidrográfica del Duero (CHD, 2013).

This station has been selected because it is one of the most reliable gauging points of Carrión system. Villoldo has been supplying hydrologic information since the early 20th century, given the importance of the Carrión River as a water supplier to the Canal of Castile, an artificial canal built between the 18th and 19th centuries to open a fluvial route to carry the wheat cropped in the area to the ports of the Bay of Biscay. The canal was soon superseded by the railway, and was relegated to being a water distribution system for irrigated farms. Currently, the Canal of Castile is a relevant natural park.

The gauge is a triangular profile V-flat weir, adequate for natural streams (e.g. WMO, 2010), converting the water depth into an electrical signal recorded in binary code, (BCD).This signal is sent by cable to a remote station. The remote station adds redundant codes, (CRC) for error correcting purposes before re-sending it via satellite to the CPC at 10 min intervals. The BCD decoder is able to process 20,000 steps with a maximum resolution of 5 mm on a 10 m range. At Villoldo gauging station readings are taken in fractions of 20 steps, which imply a real resolution of 10 mm.

The 10-min data base contained 602,928 readings from February 2, 1999 to July 20, 2010. The dataset used for quantifying the number of errors detected by quality control methods ranged between the records 1001 and 6000, selecting the records 1–1000 as the initial dataset (hereafter seed) for the NARNN training, checking and control. The dataset has already been revised by an expert technician, using the first 80 data (zero values because the sensor was still disconnected), for the purpose of checking the performance of the neural network stability. This test of NARNN stability was done with variable length seed ranging from 100, 80% error, to 1000, 8% error.

### 2.2. Errors introduction

In order to evaluate the percentage of erroneous records detected by each of the two methods, a set of known errors was introduced into the initial data series. The perturbations of the raw river stage values consisted of random errors, $x_{tRE}$ to be added the original data, $X_{tO}$, resulting in a new value $X_{tE}$

$$X_{tE} = X_{tO} + x_{tRE} \quad (1)$$

Error ratios defined as the random error divided by the standard deviation ($\sigma x$) of the river stage data over a specific time period (Meyer et al., 1989) were analyzed. The ratios were used on a monthly basis, following the simulation approach reported by Estévez et al. (2009) for the error introduction to several meteorological variables. Random error ratios, of $e = x_{tRE}/\sigma_x$, of 0.2, 0.4, 0.6, 0.8 and 1.0 were introduced (Camillo and Gurney, 1984; Ley et al., 1994). Random errors were generated from uniformly distributed pseudorandom numbers in the range [0–1] (e.g. Press et al., 2007, Chap. 7). 10% of records of the initial series for each error ratio were modified. Thus, five altered 10-min river stage - datasets were computed to assess the efficacy of the validation procedures in detecting them, and the relationship with their magnitude.

### 2.3. Conventional quality control procedures

Three conventional quality control procedures were applied to modified river level data sets from Villoldo station. These tests have been extensively used in many works, especially in hydrometeorological data analysis (Estévez et al., 2011; Feng et al., 2004; Meek and Hatfield, 1994; Shafer et al., 2000).The tests are usually based on a set of three computer-based rules proposed by O'Brien and Keefer (1985). These rules include computation of fixed or dynamic high/low bounds for each variable (range test), the use of fixed or dynamic rate of change limits for each variable