Journal of Hydrology 511 (2014) 530–545

Contents lists available at ScienceDirect

Journal of Hydrology

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

Inter-comparison of time series models of lake levels predicted by several modeling strategies

R. Khatibi [a], M.A. Ghorbani [b,*], L. Naghipour [c], V. Jothiprakash [d], T.A. Fathima [d], M.H. Fazelifard [b]

[a] Consultant Mathematical Modeller, Swindon, UK
[b] Department of Water Engineering, Tabriz University, Tabriz, Iran
[c] Department of Civil Engineering, Tabriz University, Tabriz, Iran
[d] Department of Civil Engineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India

ARTICLE INFO

Article history:
Received 18 July 2013
Received in revised form 20 December 2013
Accepted 6 January 2014
Available online 18 January 2014
This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Lake water level prediction
Chaos theory
ARIMA
Neural networks
Evolutionary methods
Pluralism in modeling

SUMMARY

Five modeling strategies are employed to analyze water level time series of six lakes with different physical characteristics such as shape, size, altitude and range of variations. The models comprise chaos theory, Auto-Regressive Integrated Moving Average (ARIMA) – treated for seasonality and hence SARIMA, Artificial Neural Networks (ANN), Gene Expression Programming (GEP) and Multiple Linear Regression (MLR). Each is formulated on a different premise with different underlying assumptions. Chaos theory is elaborated in a greater detail as it is customary to identify the existence of chaotic signals by a number of techniques (e.g. average mutual information and false nearest neighbors) and future values are predicted using the Nonlinear Local Prediction (NLP) technique. This paper takes a critical view of past inter-comparison studies seeking a superior performance, against which it is reported that (i) the performances of all five modeling strategies vary from good to poor, hampering the recommendation of a clear-cut predictive model; (ii) the performances of the datasets of two cases are consistently better with all five modeling strategies; (iii) in other cases, their performances are poor but the results can still be fit-for-purpose; (iv) the simultaneous good performances of NLP and SARIMA pull their underlying assumptions to different ends, which cannot be reconciled. A number of arguments are presented including the culture of pluralism, according to which the various modeling strategies facilitate an insight into the data from different vantages.

1. Introduction

Five different modeling strategies are used to investigate the variation in water levels of six lakes using their time series recorded monthly and spanning from 58 to 109 years of data record. If lakes are closed systems, one expects that their water levels would display distinct patterns and probably would be predictable by any reasonable model. This pragmatic expectation is challenged by a comparison of the performance of different modeling strategies using different datasets from different lakes with different shapes and sizes. A study of this nature is of a practical significance as noted among others by Sen et al. (2000) that water levels play a significant role in management of fresh water supply, designing and planning of lakeshore structures and the environment. Modeling is the key for the simulation of level variations and understanding their baseline and future states.

There are various reasons for the importance of lake water levels, e.g. Hayshi and Kamp (2007) in studying hydrological processes in water balance of lakes, who note that "Certain types of plants require relatively high water levels, while others cannot tolerate standing water. Therefore, water level change is considered a disturbance to many aquatic plants." The balance between inputs and outputs of water is controlled by the hydrological processes and this gives rise to dynamic changes in water level that can be explained by simple equations. Water level changes may also be driven by surface winds leading to the setup of seiches as standing waves. There are other processes taking place within lakes that are driven by thermal currents and mixing processes creating physical movements in the body of lake waters but the amount of change in water level is often small. These important processes are studied by using Navier–Stokes equations and provide a deeper insight into the ongoing processes, see Rodi (1984), Hodges et al. (2000) for a review. The implementation of these models is only feasible by using expensive data and long computational times but this is not the case with yet another category of models, known as time series analysis, which make use of simple records of water levels

* Corresponding author. Tel.: +98 411 3392786; fax: +98 411 3345332.
E-mail address: m_Ali_ghorbani@ymail.com (M.A. Ghorbani).

http://dx.doi.org/10.1016/j.jhydrol.2014.01.009
0022-1694/

comprising the time history of water levels. The focus of this paper is on five modeling strategies, popular with researchers, used for time series analysis, and the aim of this paper is an inter-comparison of these models using the data from six different lakes.

The five time series modeling strategies employed in this paper are chaos theory, Auto-Regressive Integrated Moving Average (ARIMA), Artificial Neural Networks (ANN), Gene Expression Programming (GEP) and the more traditional approach of Multiple Linear Regression (MLR). The latter still remains popular and serves as the baseline for model comparison. Some authors use just one technique and compare its performance with conventional approaches but often a number of techniques are used with the aim of their inter-comparison towards identifying the best performing modeling strategy. However, the cultural tendency is that the less successful models are rejected. One drawback of the search for the best model is that when a new modeling strategy is found as good as the conventional ones but not better, its launch can become less attractive, even though the evidence is not exhaustive but anecdotal. However, this is the point where this paper takes a more critical view and discusses the various possibilities.

The application of these techniques to the study of lake water levels or to other hydrological processes has been topical in research activities since the last decade or so but is often sporadic. The application of chaos techniques include: Vittori (1992) investigating nonlinear time series analysis using delay coordinate embedding on the tidal data from the Venice Lagoon from 1980 to 1984; Koçak (1997) applying nonlinear prediction to water level time series; Zaldivar et al. (1998) employing nonlinear time series analysis for the detection of high water levels in Venice (Italy); Frison et al. (1999) comparing the results of nonlinear dynamic analysis of coastal waters with other existing methods; Solomatine et al. (2000) suggested the possibility of accurate predictions of the surge water level in the North Sea with similar techniques; Rahmstorf (2003) used a semi-empirical approach to study sea level fluctuations based on earth temperature changes. Khatibi et al. (2011a) employed chaos theory to hourly water level at Hillarys Boat Harbour, Western Australia. This paper uses the Nonlinear Local Prediction model (NLP) based on chaos theory as a prediction model.

The ARIMA models are stochastic forecasting approaches formulated for prediction or forecasting purposes, with applications to study water level fluctuations in lakes using their recorded time series and the applications include: Altunkaynak (2007) compared the performance of the traditional Auto-Regressive Moving Average Exogenous input (ARMAX) models with ANN and other models in forecasting one month-ahead of lake levels; Yarar et al. (2009) used Seasonal Auto-Regressive Integrated Moving Average (SARIMA), ANN and other modeling technique for time series of Lake Beysehir. De Domenico et al. (2013) compared predictions of chaos theory with that of ARIMA for sea level modeling for daily, weekly, 10-day and monthly time scale at the Cocos (Keeling) islands and their comparative analyses show that the chaos theory is more suitable than ARIMA in forecasting, although both approaches are acceptable to a certain extent for this purpose.

The application of ANN has gained great popularity in time-series prediction because of its robustness and simplicity. Predictive ANN studies of lake water level time series include: (Altunkaynak, 2007; Ondimu and Murase, 2007); Talebizadeh and Moridnejad (2011) developed several ANN and other models to forecast the lake level fluctuations in Lake Urmu (Urmia), Azerbaijan, northwest Iran.

GEP provides another modeling strategy for predictive water level of lakes using evolutionary computing to automatically generate regression-type equations describing possible cause-and-effect relationships in the data. In recent studies, Ghorbani et al. (2010) applied GEP to forecast sea water-level variations in Hillarys Boat Harbor and compared the results with those of ANNs, which showed the feasibility of GEP in modeling time series; Kavehkar et al. (2011) used Genetic Programming (GP) to forecast daily water level variations at Lake Urmu and the results were compared with a corresponding outputs from ANN model. More recently, Kisi et al. (2012) investigated the abilities of GEP, ANN, ARMA and other modeling techniques to forecast daily lake levels at Iznik in Western Turkey.

Karimi et al. (2013) applied ANN and MLR, ARMA and other models to forecast hourly sea levels for Darwin Harbor, Australia. They also used the MLR technique for selecting the optimum input combinations of hourly sea level and predicting optimum ANN models compared with those of optimum ARMA models.

The above is indicative that investigations on time series analysis of lake water levels are topical to the extent that these research tools are likely to be at their "proof-of-concept" stage for applying to practical problems or already used by modeling practitioners as well but this paper aims to bring together the academic mindset of delivering models and tools at their proof-of-concept stages and practitioners' mindset in search of usable tools. This paper shows that there are issues yet to be addressed when different datasets fail to identify a clear-cut winner. The five modeling strategies used in this paper employ water level time series of six lakes from different regions of the world for this investigation. These are: (i) Lake Trafford, Collier County, Florida, USA, (ii) Lake Istokpoga, Highlands County, Florida, USA, (iii) Cypress Lake, Belknap County, New Hampshire, USA, (iv), Lake Winnipesaukee, New Hampshire, USA, (v) the Great Salt Lake, Salt Lake County, Utah, USA and (vi) Lake Van, Turkey. Each has more than 50 years of data comprising monthly records of water levels and this paper presents a review of their shape, size and other physical characteristics to match the performance of each model with their physical context. Both the Great Salt Lake and Lake Van are noted for being the largest of the cases within the chosen cases.

2. Mathematical basis of the modeling strategies used

2.1. Chaos theory

The implementation of chaos theory is concerned, in the first place, with the identification of the presence of deterministic chaotic signals in time series. It is customary to use several methods for the identification of these signals in a particular time series and this study employs the following nonlinear methods: Average Mutual Information (AMI), False Nearest Neighbors (FNN), Correlation Dimension (CD) and largest Lyapunov exponents. The theoretical basis of each of these methods is outlined below.

2.1.1. Phase space reconstruction

Phase space is a tool for characterizing dynamical systems, which is reconstructed from univariate or multivariate time series (Takens, 1981) generated by a deterministic chaos system with a degrees of freedom to be determined. The Takens theorem states that the underlying (unknown) dynamics can be fully recovered by building a m-dimensional space wherein the components of each state vector $\boldsymbol{Y}_t$ are defined through the delay coordinates.

$$\boldsymbol{Y}_t = \{X_t, X_{t-\tau}, X_{t-2\tau}, \ldots, X_{t-(m-1)\tau}\} \tag{1}$$

where $m > 2D_2$ is called *embedding dimension* and τ is referred to as *lag time*, D_2 referred to later as correlation exponent and X_t represents time series. If the dynamics of the system can be reduced to a set of deterministic laws, trajectories converge towards a subset of the phase space with fractional dimension, called attractor.

2.1.2. AMI and FNN

This paper employs the minimization of the AMI to identify τ and the FNN to that of m, as suggested by Cellucci et al. (2003). For a given time series sequence $\{x_0, x_1, x_2, \ldots, x_i, \ldots, x_n\}$, the mutual information indicates the amount of information about the