20. APHA, *Standard Methods for the Examination of Water and Wastewater*, American Public Health Association, American Water Works Association, Water Pollution Control Federation, Washington, DC, 2002, 22nd edn.

21. Haiping, L., Pei, X., Timberley, M., Roane, E., Jenkins, P. and Zhiyong, R., Microbial desalination cells for improved performance in wastewater treatment, electricity production and desalination. *Bioresour. Technol.*, 2012, **105**, 60–66.

22. Mehanna, M. *et al.*, Using microbial desalination cells to reduce water salinity prior to reverse osmosis. *Energy Environ. Sci.*, 2010, **3**, 1114–1120.

23. Franks, A. E., Nevin, K. P., Jia, H. F., Izallalen, M., Woodard, T. L. and Lovley, D. R., Novel strategy for three-dimensional real-time imaging of microbial fuel cell communities: Monitoring the inhibitory effects of proton accumulation within the anode biofilm. *Energy Environ. Sci.*, 2009, **2**, 113–119.

24. Lee, H. S., Parameswaran, P., Kato-Marcus, A., Torres, C. I. and Rittmann, B. E., Evaluation of energy-conversion efficiencies in microbial fuel cells (MFCs) utilizing fermentable and non-fermentable substrates. *Water. Res.*, 2008, **42**(6–7), 1501–1510.

25. Mercer, Microbial Fuel Cells: Generating Power from Waste, illumin.usc.edu April (vol. XV (II), Online), 2014, 26.

26. Katuri, K. P., Enright, A., Flaherty, V. O. and Leech, D., Microbial analysis of anodic biofilm in a microbial fuel cell using slaughterhouse wastewater. *Bioelectrochem*, 2012, **87**, 164–171.

27. Kokabian, B. and Gude, V. G., Sustainable photosynthetic biocathode in microbial desalination cells. *J. Chem. Eng.*, 2015, **262**, 958–965.

28. Rabaey, K., Clauwaert, P., Aelterman, P. and Verstraete, W., Tubular microbial fuel cells for efficient electricity generation. *Environ. Sci. Technol.*, 2005, **39**, 8077–8082.

29. Min, B. and Logan, B. E., Continuous electricity generation from domestic wastewater and organic substrates in a flat plate microbial fuel cell. *Environ. Sci. Technol.*, 2004, **38**, 5809–5814.

30. Sun, J. J., Zhao, H. Z., Yang, Q. Z., Song, J. and Xue, A., A novel layer-by-layer self assembled carbon nanotube-based anode, preparation, characterization, and application in microbial fuel cell. *Electrochim. Acta*, 2010, **55**, 3041–3047.

31. Rabaey, K., Angenent, L., Schroder, U. and Keller, J., *Bioelectrochemical Systems: From Extracellular Electron Transfer to Biotechnological Application*, IWA Publishing, London, 2009, 1st edn.

32. Aelterman, P., Versichele, M., Marzorati, M., Boon, N. and Verstraete, W., Loading rate and external resistance control the electricity generation of microbial fuel cells with different three-dimensional anodes. *Bioresour. Technol.*, 2008, **99**, 8895–8902.

33. Xie, X., Hu, L. B., Pasta, M., Wells, G. F., Kong, D. S., Criddle, C. S. and Cui, Y., Three-dimensional carbon nanotube-textile anode for high-performance microbial fuel cells. *Nano. Lett.*, 2011, **11**, 291–296.

# A time-series forecasting-based prediction model to estimate groundwater levels in India

## Debasish Sena* and Naresh Kumar Nagwani

National Institute of Technology, Raipur 492 010, India

**India is one of the fast developing countries in the world with a growth rate of 6.4%. Rapid industrialization is the main cause behind such growth. Although industrialization is of utmost importance for growth, sustainability of ecology is also a matter of concern. India has a vast coastline, but the saline water is not suitable for industrialization; so groundwater is the primary source for both industrialization and human consumption. Agriculture plays a major role in India's economy and irrigation is also dependent on groundwater to some extent. Hence the study of groundwater levels is the need of the hour. In this study, time-series techniques like fuzzy time-series analysis and ARIMA are utilized for forecasting monthly groundwater levels. Experiments are performed on the datasets collected from different regions of India. The experimental results demonstrate that fuzzy time series analysis yields more accurate forecast of groundwater levels compared to the ARIMA model. The results of this study can be utilized for planning a suitable policy for groundwater use and its proper regulation to avoid future crisis.**

GROUNDWATER is a major resource in our country. In fact, India tops the list of groundwater abstracting countries. Groundwater is essential for sustainability of ecosystem; it provides stream water during drought conditions. Considering the effects of climate change, land-use change and global environmental changes like change in the amount of precipitation, increase in temperature and increase in demand of groundwater because of population growth, it is important to assess them[1]. Water being a dynamic resource, its storage undergoes continuous change either by recharge from various sources or discharge due to extraction or natural basin outflow. Hence periodic monitoring of groundwater levels is imperative for planning systematic development and management of groundwater resources[2].

Groundwater level prediction in India is of utmost importance as our large population is heavily dependent on groundwater for daily consumption. Also groundwater is heavily used both for irrigation and industrialization in India. Due to faulty irrigation system, a lot of groundwater is wasted. Prediction of groundwater levels is the

*For correspondence. (e-mail: dsena.mtech2013.ee@nitrr.ac.in)

need of the hour to avoid future crisis. Earth science data like groundwater level are large and complex and often represent a time series, making them difficult to analyse[3]. A wide variety of data mining, machine learning and information theoretic approaches are applicable to groundwater-level data. Artificial neural network and model tree ensembles methods have generally been employed for future prediction of groundwater levels[3]. This communication focuses on methods like auto-regressive integrated moving average (ARIMA) using Box–Jenkins methodology and fuzzy time-series analysis for forecasting groundwater level.

In the past, various approaches have been suggested for predicting water level, including physical and statistical models. However, none of them is considered best because of the high degree of uncertainty and time-varying characteristics of the hydro-system. Principal component analysis (PCA) and neural network models have been designed for predicting water level of Hoek Van Holland during storm situations by van de Weg[4]. However, the PCA method has the disadvantage of difficulty in calculation of covariance matrix in an accurate manner[5]. Chang and Chang[6] have proposed an adaptive neuro-fuzzy inference system for forecast of water level in reservoirs. Not having any systematic method for designing the controller is the main issue with fuzzy logic while a lot of computational resource is needed to fully implement a standard neural network. Scitovski *et al.*[7] have utilized the periodicity of water-level behaviour using trigonometric regression for long-term forecasting and nearest neighbour method for short-term forecasting. Forecasting of groundwater level using conceptual physical models has also been proposed. But these are constrained with the limitation of too many dependent variables[8]. A hybrid model combining genetic algorithm and wavelet network model has been proposed by Wang and Zhao[8]. But the genetic algorithm does not assure of a global optimal solution.

The ARIMA model and fuzzy time-series model used in this study have been designed based on past data and some random noise component with mathematical manipulation and can predict the groundwater levels more accurately. As the fuzzy time-series analysis provides a better result in comparison to ARIMA, the latter model can be used in combination with various parametric and non-parametric methods of forecasting. Artificial neural network (ANN), **k**-nearest neighbour (k-NN), Markov chains, etc. can be used with the ARIMA model to reduce the forecasting error. The ARIMA forecast models are usually governed by three components: variables of the model, coefficients of the variables and some unobserved errors or random shock[9]. All the three components contribute towards the uncertainty of forecasting. In this communication, all the three components have undergone thorough analysis experimentally. The effect of temporal aggregation on ARIMA processes has been discussed by

Stram and Wei[10]. Fuzzy time-series analysis on groundwater level dataset is performed by following the work of Song and Chissom[11]. Trend analysis of pre- and post-monsoon groundwater levels has been performed by Gokhale and Sohoni[12]. The ARIMA model takes care of the seasonal variation of groundwater level in addition to the trend component.

For the ARIMA model monthly groundwater-level data have been collected from Groundwater Information System, Government of India (GoI), Ministry of Water Resources, Central Ground Water Board (CGWB)[13]. In this study the ARIMA model is used for future prediction of groundwater level utilizing Box–Jenkins methodology. As multiple ARIMA models can be proposed for a single dataset, a suitable model has been chosen by studying aspects such as mean square error (MSE) and mean magnitude of relative error (MMRE).

The Box–Jenkins methodology is useful for stationary time series, i.e. it must have a stable mean, variance and autocorrelation over the series[14]. For a stationary series, the correlogram dies down rapidly, or it lasts for four to five lags above the significant level. One way of removing non-stationarity from time series is by simply applying difference operation to the time series.

The first-order differencing is expressed as $X'_t = X_t - X_{t-1}$, where $X_t$ is the value of the time-series variable at time $t$, $X_{t-1}$ the value of the time-series variable at time $t-1$ and $X'_t$ is the first differenced time-series value. Likewise second-order differencing is expressed as $X''_t = X'_t - X'_{t-1}$. In most cases, up to second-order differencing is performed for a time series. The backshift operator $B$ is often used to represent the equation in a compact manner; the first-order difference operation is expressed as $X'_t = (1 - B)X_t$ and the second-order difference operation as $X''_t = (1 - B)^2 X_t$.

A time series generally consists of two parts: a deterministic part representing the time-series values and a white noise part induced implicitly. The ARIMA model includes both parts. The auto-regressive part represents the deterministic component and determines how the data values of a time series regress upon themselves. The moving average part corresponds to the memory of the time series for the preceding random noise components[23]. The integrated part represents the degree of differencing needed to convert a time series to a stationary one.

An auto-regressive model of order $p$ (AR($p$)) suggests how the current time series value is regressed upon $p$ number of past time-series values. So an AR($p$) model can be mathematically represented as

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + \cdots + \phi_p X_{t-p} + \varepsilon_t, \tag{1}$$

where $\{X_i\}$ are the time-series values at instance $i$, $\{\phi_i\}$ are the auto-regressive parameters and $\varepsilon_t$ is the white noise component at instance $t$. Equation (1) can be written in terms of backshift operator as

$$\phi_p(B)X_t = \varepsilon_t, \tag{2}$$

where $\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ is the AR characteristic polynomial calculated on $B$ (ref. 16).

A moving-average model of order $q$ (MA($q$)) indicates the current value of the time series as a linear regression of $q$ previous white noise values[15]. Mathematically, an MA($q$) model can be expressed as

$$X_t = \varepsilon_t - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}, \tag{3}$$

where $\{\theta_i\}$ are the moving-average parameters. Equation (3) can be written in terms of backshift operator as

$$X_t = \theta_q(B)\varepsilon_t, \tag{4}$$

where $\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$ is the MA characteristic polynomial calculated on $B$ (ref. 16).

A $d$ order differenced time series can be expressed in terms of backshift operator $B$ as $(1 - B)^d X_t$. So an ARIMA($p, d, q$) model is a combination of an auto-regressive model of order $p$ and a moving-average model of order $q$ applied over a $d$ times differenced time series. Mathematically, an ARIMA($p, d, q$) can be expressed as

$$\phi_p(B)(1 - B)^d X_t = \theta_q(B)\varepsilon_t. \tag{5}$$

The values of $p, d, q, \{\phi_i\}$ and $\{\theta_i\}$ can be calculated by building appropriate ARIMA models.

Generally the ARIMA model is expressed as ARIMA($p, d, q$). While applying ARIMA model to a time series, first the differencing is performed to convert the time series to a stationary one and then auto-regressive moving average (ARMA) model is applied to the differenced series. The ARIMA($p, d, q$) model can be represented as given in eq. (5). Sometimes the time series displays seasonality, i.e. dependency on past data seems prominent at multiples of some seasonal lag $s$. So the ARIMA model for such time series comprises a seasonal auto-regressive component and a seasonal moving-average component employed over a seasonally differenced time series. The model is referred as ARIMA($P, D, Q$)$_s$ and is expressed as

$$\Phi_P(B^s)(1 - B^s)^D X_t = \Theta_Q(B^s)\varepsilon_t, \tag{6}$$

where $\Phi_P(B^s) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$ and $\Theta_Q(B^s) = 1 - \Theta_1 B^s - \Theta_2 B^{2s} - \cdots \Theta_Q B^{Qs}$ are the seasonal AR and MA operators of orders $P$ and $Q$ respectively, with seasonal lag $s$ (ref. 27). In general the seasonal and non-seasonal operators can be aggregated into a multiplicative seasonal ARIMA, denoted by SARIMA($p, d, q$) ($P, D, Q$)$_s$ and expressed as

$$\phi_p(B)\Phi_P(B^s)(1 - B)^d(1 - B^s)^D X_t = \theta_q(B)\Theta_Q(B^s)\varepsilon_t. \tag{7}$$

For a given time series, first the order of difference $d$ is determined. Then, the order of auto-regression $p$ and order of moving average $q$ are determined. There can be multiple possible sets of $p, q, d, P, Q$ and $D$ for a particular time series. So to derive a suitable model, three steps[20] are followed according to the Box–Jenkins models: (a) Model structure identification; (b) Parameter estimation and calibration; (c) Validation or model testing.

In model structure identification, the order of auto-regression ($p$), order of seasonal auto-regression ($P$), moving average ($q$), order of seasonal moving average ($Q$), order of differencing ($d$) and seasonal order of differencing ($D$) are estimated. The order of difference is determined from the number of difference (time period changes) operations applied on the time series to make it a stationary one. The order of auto-regression is determined from the number of significant partial auto-correlation values and the order of moving average is the number of significant auto-correlation values with either exponentially decaying or in a dampened sine wave. After obtaining a stationary time-series model, it can be identified from the theory[20,21] given in Table 1.

The seasonal differencing ($D$) can be indicated by a correlogram decaying gradually at multiples of some seasons, but negligible between consecutive periods[17]. The seasonal auto-regression order ($P$) is the number of significant partial auto-correlation values occurring at some season and the seasonal moving average order ($Q$) is the number of significant auto-correlation values occurring at some season. The seasonality orders can be identified from Table 2.

In parameter estimation and calibration, after determining the order of auto-regression ($p$) and order of moving average ($q$), the next big task is to estimate the auto-regressive parameters $\{\phi_i\}$ and the moving average parameters $\{\theta_i\}$. The auto-regressive parameters can be determined by the Yule–Walker's equation[21,22] and the moving average parameters can be determined using the equation

$$\rho_k = \frac{(-\theta_k + \theta_1\theta_{k-1} + \theta_2\theta_{k-2} + \cdots + \theta_{q-k}\theta_q)}{(1 + \theta_1^2 + \theta_2^2 + \cdots + \theta_q^2)}, \tag{8}$$

for $k = 1, 2, ..., q$. There are various algorithms like Marquadt's algorithm, 'armax' toolbox in 'MATLAB' and 'arima' functions in $R$ available for parameter estimation[23]. Maximum likelihood estimation is also used for estimating parameters[24].

**Table 1.** Behaviour of ACF and PACF for ARMA models

| Model | ACF | PACF |
|---|---|---|
| AR($p$) | Dies down | Cuts off after $p$ lags |
| MA($q$) | Cuts off after $q$ lags | Dies down |
| ARMA($p, q$) | Dies down | Dies down |

**Table 2.** Behaviour of ACF and PACF for seasonal ARMA models

| Model | ACF* | PACF* |
|---|---|---|
| $AR(P)_s$ | Tails off at lags $k_s$, $k = 1, 2, \ldots$, | Cuts off after lag $P_S$ |
| $MA(Q)_s$ | Cuts off after lags $Q_s$ | Tails off at lags $k_s$, $k = 1, 2, \ldots$, |
| $ARMA(P, Q)_s$ | Tails off at lags $k_s$ | Tails off at lags $k_s$ |

*Values at non-seasonal lags $h \neq k_s$ for $k = 1, 2, \ldots$, are zero.

As there are many possible models, choosing the appropriate one is of utmost importance in time-series analysis. Model testing and validation are used to validate the proposed model. Among the various possible models, the one best suitable for the time series is determined either by maximum likelihood estimation (MLE)[22], MSE[23] or MMRE criteria. Also criteria like AIC, AICc and BIC are used to decide the suitable model for a particular time series[17]. More often, MLE method is used for long-term data generation, whereas MSE and MMRE methods are advisable for short-term forecasting of the time series. In this study, MMRE was calculated for each model and the one with minimum MMRE selected as the best model for forecasting. Residual analysis was also performed to check the fitness of the model.

Fuzzy time-series analysis is a recent technique of future forecasting. It is basically established on the fuzzy set theory. The drawback of conventional set theory is that in the real world, many concepts cannot be explained by their membership or non-membership within the set. So fuzzy set theory appears as the solution to the problems posed by conventional set theory.

Let $U$ be the universe of discourse divided into $n$ intervals as $U = \{u_1, u_2, \cdots, u_n\}$, where $u_i$ is an interval in the universe of discourse $U$. A fuzzy set $A_i$ of $U$ is defined as

$$A_i = f_{Ai}(u_1)/u_1 + f_{Ai}(u_2)/u_2 + \cdots + f_{Ai}(u_n)/u_n,$$

where $f_{Ai}$ is the membership function for fuzzy set $A_i$, $f_{Ai} : U \to [0,1]$. $u_k$ is the element of fuzzy set $A_i$ and $f_{Ai}(u_k)$ is the degree of membership of $u_k$ to $A_i$. $f_{Ai}(u_k) \in [0,1]$, where $1 \leq k \leq n$.

Let $Y(t) (t = \cdots, 0, 1, 2, \cdots)$ be a subset of $R$, the universe of discourse on which fuzzy sets $f_i(t) (i = 1, 2, \cdots)$ are defined, and let $F(t)$ be a collection of $f_i(t)$. Then, $F(t)$ is called a fuzzy time series on $Y(t) (t = \cdots, 0, 1, 2, \cdots)$. $F(t)$ can be called a linguistic variable[28] and $f_i(t)$ $(i = 1, 2, \cdots)$ can be viewed as possible linguistic values of $F(t)$ and are presented by fuzzy sets. As $F(t)$ is time-dependent and according to Song and Chissom[11], if $F(t)$ is caused by $F(t - 1)$ only, then the relationship can be represented by $F(t - 1) \to F(t)$. The above dependency can be represented by $F(t) = F(t-1) \times R(t-1, t)$, where $R(t - 1, t)$ represents the fuzzy relationship between $F(t)$ and $F(t - 1)$, and '$\times$' represents an operator (can be max–min[11], min–max[29] or arithmetic operator[30]). If $F(t - 1)$ can be represented by $A_{i-1}$ and $F(t)$ by $A_i$, then $F(t - 1) \to F(t)$ can be represented as $A_{i-1} \to A_i$.

A fuzzy logical relationship group can be constructed by grouping all right-hand-side fuzzy sets preceded by the same fuzzy set in the left-hand-side of fuzzy logical relationship[28]. If there are fuzzy logical relationships such that $A_i \to A_j, A_i \to A_k, A_i \to A_l, \cdots$, then they can be merged into a fuzzy logical relationship group as $A_i \to A_j, A_k, A_l, \cdots$.

Determining the length of the interval to divide the fuzzy time series into multiple fuzzy sets is an important task as different lengths of intervals may produce different forecasting results. An effective length of interval should not be too large or too small, as too large intervals lead to no fluctuation in the fuzzy time series and too small, intervals will diminish the mining of fuzzy time series[31]. A heuristic for determining the effective length of the interval is set in a manner so that at least half of the fluctuations in the time series should be reflected by the interval. Based on this concept, two approaches are proposed[31]. They are average-based length and distribution-based length. In this communication, the distribution-based approach is used for effective length determination.

The calculations of forecasting are carried out by the following procedure as given by Chen[28].

(a) If fuzzified value of time $i$ is $A_i$ and there exists a fuzzy logical relationship $A_i \to A_j$ and the maximum membership value of $A_j$ occurs in the interval $u_j$, then forecasted value of time $i + 1$ is $m_j$, where $m_j$ is the midpoint of $u_j$.

(b) If fuzzified value of time $i$ is $A_i$ and there exists fuzzy logical relationships $A_i \to A_{i1}, A_i \to A_{i2}, \cdots$, $A_i \to A_{jp}$ and the maximum membership values of $A_{i1}, A_{i2} \cdots, A_{jp}$ occur in the interval $u_1, u_2, \cdots, u_p$ respectively, then forecasted value for time $i + 1$ is $(m_1 + m_2 + \cdots + m_p)/p$, where $m_1, m_2, \cdots, m_p$ are the midpoints of intervals $u_1, u_2, \cdots, u_p$ respectively.

(c) If fuzzified value of time $i$ is $A_i$ and there does not exist any fuzzy logical relationship group whose current state of value is $A_i$ and the maximum membership value of $A_i$ occurs in the interval $u_i$ with a midpoint $m_i$, then the forecasted value for time $i + 1$ is $m_i$.

We now explain details of the experiment, performed on monthly groundwater level of Jainath region, Adilabad district, Andhra Pradesh, India, based on the methodology described above.

The dataset used here is taken from the Groundwater Information System, GoI. The datasets taken are monthly groundwater level data from 2005 to 2012.

The obtained dataset has some missing values and is filled in by linear interpolation. The dataset is also analysed for possible outliers and these have been replaced by the average value of the corresponding month.

The steps for designing the ARIMA model for the dataset are explained using the R software package[25]. Like R, there are several other software packages available for time series analysis. The time-series for the monthly groundwater level has been used for building the ARIMA models.

Figure 1 shows the dataset taken for the analysis. It shows the monthly groundwater level of Jainath region from 2005 to 2012. Figure 2 shows the time plot for the dataset. The time (in years) is represented on the X-axis and the monthly groundwater level (in metres) is represented on the Y-axis. The plot presents a stationary time series with a seasonality $s = 12$. Hence, no difference operation is needed for the dataset and the order of integration ($d$) for the time series is zero.

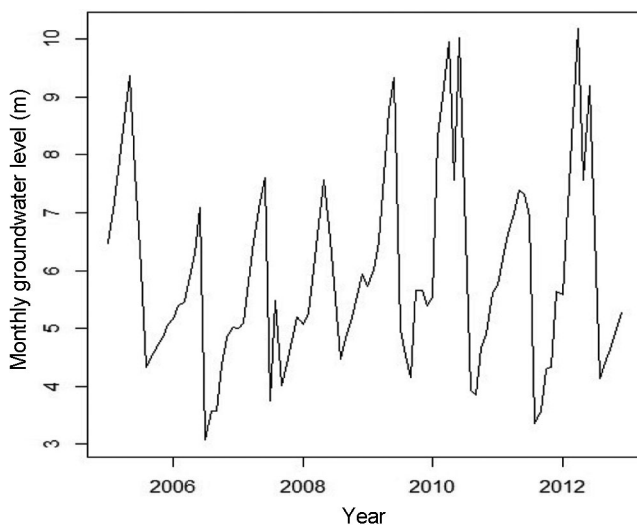Then the order of auto-regression and moving average is determined by drawing the ACF and PACF plot of the stationary series, as shown in Figures 3 and 4 respectively. The dotted line around the abscissa represents the 95% confidence interval, and the ACF and PACF values within the confidence interval are considered as insignificant. The ACF shows a damping sine wave with significant auto-correlation values at lag 1, lag 12 and lag 24. So the order of moving average ($q$) and seasonal moving average order ($Q$) are determined as 1 and 2 respectively. Significant partial auto-correlation values at lag 0 and lag 12 can be observed from the PACF plot. So the order of auto-regression ($p$) and seasonal auto-regression ($P$) can be determined as 1 and 1 respectively. Hence the model identified for the monthly groundwater level is ARIMA$(1, 0, 1)(1, 0, 2)_{12}$.

The auto-regressive parameters $\phi_1$ and $\Phi_1$ are to be 0.7735 and 0.9181 respectively and the moving average
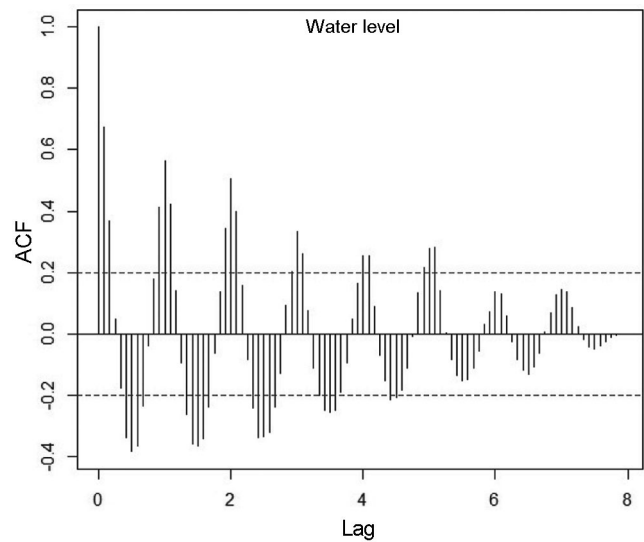
|   | Jan | Feb | Mar | Apr | May | Jun | Jul | Aug | Sep | Oct | Nov | Dec |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 2005 | 6.47 | 7.20 | 7.92 | 8.64 | 9.37 | 7.70 | 6.03 | 4.34 | 4.50 | 4.66 | 4.84 | 5.07 |
| 2006 | 5.17 | 5.40 | 5.47 | 5.88 | 6.30 | 7.09 | 3.07 | 3.57 | 3.57 | 4.37 | 4.85 | 5.02 |
| 2007 | 5.00 | 5.10 | 5.80 | 6.50 | 7.20 | 7.59 | 3.75 | 5.48 | 4.01 | 4.39 | 4.79 | 5.19 |
| 2008 | 5.08 | 5.26 | 6.03 | 6.80 | 7.57 | 6.54 | 5.51 | 4.47 | 4.83 | 5.19 | 5.57 | 5.93 |
| 2009 | 5.73 | 6.01 | 6.51 | 7.64 | 8.78 | 9.33 | 4.99 | 4.53 | 4.15 | 5.65 | 5.65 | 5.39 |
| 2010 | 5.54 | 8.28 | 9.19 | 9.95 | 7.56 | 10.01 | 7.27 | 3.92 | 3.85 | 4.67 | 4.89 | 5.62 |
| 2011 | 5.75 | 6.21 | 6.62 | 7.00 | 7.38 | 7.32 | 6.92 | 3.36 | 3.57 | 4.30 | 4.34 | 5.63 |
| 2012 | 5.58 | 7.11 | 8.64 | 10.17 | 7.56 | 9.18 | 6.65 | 4.13 | 4.41 | 4.70 | 4.99 | 5.27 |
| 2013 | 4.69 | 5.10 | 5.52 | 5.93 | 6.35 | 5.66 | 4.96 | 4.27 | 4.45 | 4.63 | 4.82 | 5.32 |

**Figure 1.** Dataset of monthly groundwater level.



**Figure 3.** ACF plot of monthly groundwater level.



**Figure 2.** Plot of monthly groundwater level.
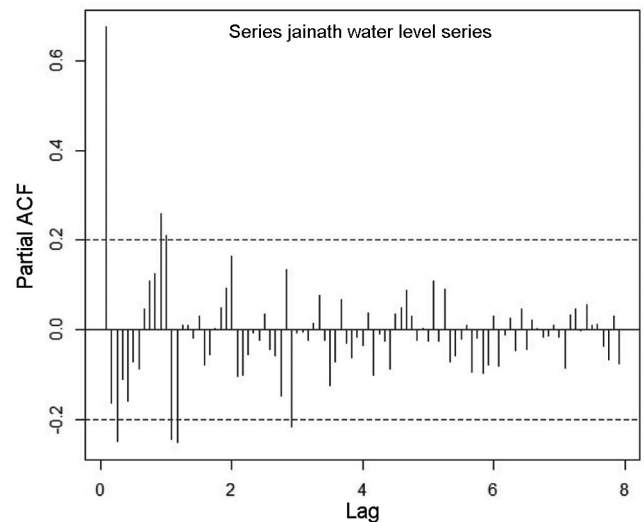


**Figure 4.** PACF plot of monthly groundwater level.

parameters $\theta_1$, $\Theta_1$ and $\Theta_2$ are $-0.3038$, $-0.6261$ and $0.1494$ respectively. There is an intercept of 6.0654 for the time series. So, the mathematical expression for the built model is as follows:

$$(1-0.9181B^{12})(1-0.7735B)X_t = 6.0654 +$$
$$(1-0.6261B^{12}+0.1494B^{24})(1+0.3038B)\varepsilon_t. \qquad (9)$$

Equation (9) can be simplified as given in eq. (10).

$$X_t = 0.7735X_{t-1} + 0.9181X_{t-12} - 0.7101X_{t-13}$$
$$+0.045\varepsilon_{t-25} + 0.1494\varepsilon_{t-24} - 0.1902\varepsilon_{t-13}$$
$$-0.6261\varepsilon_{t-12} + 0.3038\varepsilon_{t-1} + \varepsilon_t + 6.0654. \qquad (10)$$

Forecast of the monthly groundwater level for the year 2013 is performed using the model designed above and is shown in Figure 5. The plot indicates a seasonal groundwater-level fluctuation within a confidence interval of 80% and 95%.

In this experiment, groundwater-level data from 2005 to 2012 have been used for the design of the model and data of 2013 have been used for verifying the designed model. As the groundwater-level data of 2013 only contains information for the months of January, May, August and November, the forecast values of the corresponding months only have been used for the calculation of the MMRE. After forecasting, the MMRE for forecast values has been calculated followed by calculation of the percentage error, which is 9.39. So the predicted model is
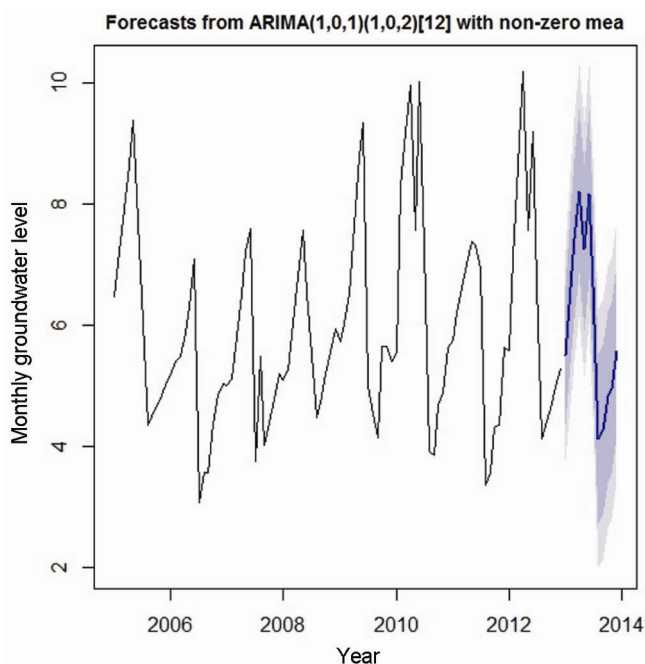
better for the prediction of earth science data like groundwater level. Again to test the goodness of the desired model, diagnosed checking has been performed. Residual analysis is used here as a method of diagnostic checking. The quantile–quantile (Q–Q) plot shown in Figure 6 is almost linear, which implies a normal distribution of residuals. Figure 7 shows a symmetric histogram with a normal curve. Figures 6 and 7 validate a good fitness of the model[26].

The dataset considered here for implementation purpose is the monthly groundwater-level data of Jainath region from 2005 to 2012. The groundwater level varies from 3.07 to 10.17 m. So the universe of discourse is
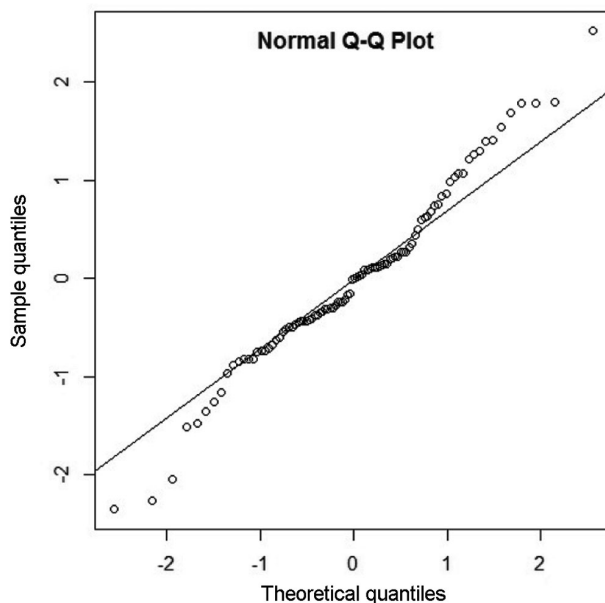


**Figure 6.** Quantile–quantile plot for monthly groundwater level.



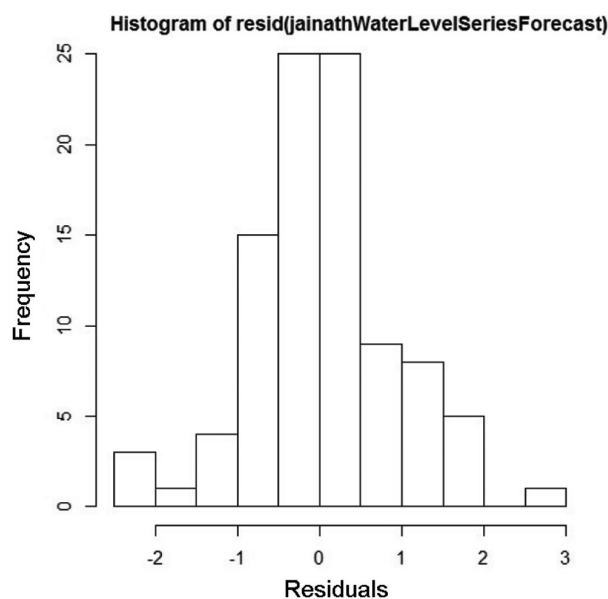**Figure 5.** Forecast plot of monthly groundwater level.



**Figure 7.** Histogram of the residuals of monthly groundwater level.

chosen from 3.00 to 10.20. To fuzzify the universe, division of the overall interval is performed using the distribution-based length approached as discussed by Huarng[31]. Initially the average value the absolute values of the first difference of the series is calculated; it is 0.85 m. Then the base for the length of the interval is calculated as 0.1 according to Table 3. The length of the interval is chosen as 0.4, which is the largest value less than at least half of the first differences.

After determining the effective length of the intervals, the universe of discourse is divided into 18 intervals as shown in Table 4.

While defining fuzzy sets on the universe, the linguistic variable is 'monthly water level' and the universe of discourse is divided into 18 fuzzy sets, $A_1, A_2, ..., A_{18}$ and each $A_i$ ($i = 1, 2, ..., 18$) is defined by intervals $u_1, u_2, ..., u_{18}$ as follows:

$$A_1 = \{1/u_1, 0.5/u_2\};$$

$$A_2 = \{0.5/u_1, 1/u_2, 0.5/u_3\};$$

$$A_3 = \{0.5/u_2, 1/u_3, 0.5/u_4\};$$

$$A_4 = \{0.5/u_3, 1/u_4, 0.5/u_5\};$$

$$A_5 = \{0.5/u_4, 1/u_5, 0.5/u_6\};$$

$$A_6 = \{0.5/u_5, 1/u_6, 0.5/u_7\};$$

$$A_7 = \{0.5/u_6, 1/u_7, 0.5/u_8\};$$

$$A_8 = \{0.5/u_7, 1/u_8, 0.5/u_9\};$$

$$A_9 = \{0.5/u_8, 1/u_9, 0.5/u_{10}\};$$

$$A_{10} = \{0.5/u_9, 1/u_{10}, 0.5/u_{11}\};$$

$$A_{11} = \{0.5/u_{10}, 1/u_{11}, 0.5/u_{12}\};$$

$$A_{12} = \{0.5/u_{11}, 1/u_{12}, 0.5/u_{13}\};$$

$$A_{13} = \{0.5/u_{12}, 1/u_{13}, 0.5/u_{14}\};$$

$$A_{14} = \{0.5/u_{13}, 1/u_{14}, 0.5/u_{15}\};$$

$$A_{15} = \{0.5/u_{14}, 1/u_{15}, 0.5/u_{16}\};$$

$$A_{16} = \{0.5/u_{15}, 1/u_{16}, 0.5/u_{17}\};$$

$$A_{17} = \{0.5/u_{16}, 1/u_{17}, 0.5/u_{18}\};$$

$$A_{18} = \{0.5/u_{17}, 1/u_{18}\}.$$

After defining the fuzzy sets, each value of the monthly groundwater level series is assigned with its corresponding fuzzy sets.

After fuzzifying the whole dataset, the fuzzy logical relationship group is obtained from each fuzzy logical rela-

tionship by following the theory as mentioned earlier in the text. The fuzzy logical relationship groups are as given in the Table 5.

Using the monthly groundwater level dataset of Jainath region, the groundwater level for the last 12 months of the dataset has been forecast and relative error calculated by comparing the forecasted groundwater level against their actual values for January, May, August and November. Table 6 shows details of the forecasting. MRE is obtained as 0.0687 and the percentage error is calculated to be 6.87.

**Table 3.** Base mapping table

| Range | Base |
|---|---|
| 0.1 to 1.0 | 0.1 |
| 1.1 to 10 | 1 |
| 11 to 100 | 10 |
| 101 to 1000 | 100 |

**Table 4.** Fuzzy set intervals

| | | |
|---|---|---|
| $u_1 = [3.0, 3.4]$ | $u_2 = [3.4, 3.8]$ | $u_3 = [3.8, 4.2]$ |
| $u_4 = [4.2, 4.6]$ | $u_5 = [4.6, 5.0]$ | $u_6 = [5.0, 5.4]$ |
| $u_7 = [5.4, 5.8]$ | $u_8 = [5.8, 6.2]$ | $u_9 = [6.2, 6.6]$ |
| $u_{10} = [6.6, 7.0]$ | $u_{11} = [7.0, 7.4]$ | $u_{12} = [7.4, 7.8]$ |
| $u_{13} = [7.8, 8.2]$ | $u_{14} = [8.2, 8.6]$ | $u_{15} = [8.6, 9.0]$ |
| $u_{16} = [9.0, 9.4]$ | $u_{17} = [9.4, 9.8]$ | $u_{18} = [9.8, 10.2]$ |

**Table 5.** Monthly groundwater level fuzzy logical relationship group

$A_9 \rightarrow A_{11}, A_7, A_{12}, A_{10}$
$A_{11} \rightarrow A_{13}, A_1, A_{12}, A_3, A_{11}, A_{10}, A_{15}$
$A_{13} \rightarrow A_{15}$
$A_{15} \rightarrow A_{16}, A_{18}$
$A_{16} \rightarrow A_{12}, A_5, A_{18}, A_{10}$
$A_{12} \rightarrow A_8, A_2, A_9, A_{15}, A_{18}, A_{16}$
$A_8 \rightarrow A_4, A_9, A_{10}, A_7$
$A_4 \rightarrow A_4, A_5, A_3, A_7$
$A_5 \rightarrow A_5, A_6, A_4, A_7$
$A_6 \rightarrow A_6, A_7, A_5, A_8$
$A_7 \rightarrow A_8, A_9, A_3, A_4, A_7, A_6, A_{14}, A_{11}, A_5$
$A_1 \rightarrow A_2$
$A_2 \rightarrow A_2, A_4, A_7$
$A_3 \rightarrow A_4, A_7, A_3, A_5$
$A_{10} \rightarrow A_{12}, A_{10}, A_{11}, A_1, A_3$
$A_{14} \rightarrow A_{16}$
$A_{18} \rightarrow A_{12}, A_{11}$

**Table 6.** Forecast details of monthly groundwater level

| Year | Month | AGL | FC | FGL | RE |
|---|---|---|---|---|---|
| 2013 | January | 4.69 | $A_5$ | 5 | 0.066 |
| 2013 | May | 6.35 | $A_9$ | 6.8 | 0.0708 |
| 2013 | August | 4.27 | $A_4$ | 4.7 | 0.1007 |
| 2013 | November | 4.82 | $A_5$ | 5 | 0.0373 |

In this study, different models for groundwater-level prediction have been proposed for 'Jainath' region using ARIMA and fuzzy time-series analysis. The models have been built using the past groundwater-level fluctuation patterns. The current predicted groundwater level is linearly related to its previous value because the ARIMA models are based on auto-correlations. The models are verified using the groundwater-level values of year 2013 of the dataset. Percentage error is calculated for both ARIMA and fuzzy time-series analysis as 9.35 and 6.87 respectively. This clearly indicates that the fuzzy time-series analysis is better than the ARIMA model for forecasting. The CGWB along with the state groundwater agencies can apply these models for the quinquennial periodic groundwater assessment (GWA) for estimating the dynamic groundwater resource. The groundwater estimation committee conducting the national GWA exercise can adopt these models for the estimation of groundwater level of individual GWA units. There is scope for further improvement of the present model by combining it with other parametric and non-parametric models.

Conflict of interest:   The authors certify that there is no conflict of interest regarding the publication of this paper.

1. Stoll, S., Hendricks Franssen, H. J., Barthel, R. and Kinzelbach, W., What can we learn from long-term groundwater data to improve climate change impact studies? *Hydrol. Earth Syst. Sci.*, 2011, **15**(12), 3861–3875.
2. Reddy, A. G. S., Water level variation in fractured, semi-confined aquifers of Anantpur district, southern India. *J. Geol. Soc. India*, 2012, **80**, 111–118.
3. Hoffman, F. M. *et al.*, Data mining in earth system science. In Proceeding of International Conference on Computational Science, Reykjavik, Iceland, 2011, pp. 1450–1455.
4. van de Weg, M. C., Prediction of water level during storm situations using neural networks, Department of Computer Science, Thesis, Leiden University, The Netherlands, 1997.
5. Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M. and Hooman, A., An overview of principal component analysis. *J. Signal Inf. Proc.*, 2006, **4**(1), 173–175.
6. Chang, F. J. and Chang, Y. T., Adaptive neuro-fuzzy inference system for prediction of water level in reservoir. *Adv. Water Resour.*, 2006, **29**, 1–10.
7. Scitovski, R., Maričić, S. and Scitovski, S., Short term and long term water level prediction at one river measurement location. *Croat. Oper. Res. Rev.*, 2012, **3**, 1–11.
8. Wang, L. and Zhao, W., Forecasting groundwater level based on WNM with GA. *J. Comput. Inf. Syst.*, 2011, **7**(1), 160–167.
9. Gavirangaswamy, V. B., Gupta, G., Gupta, A. and Agrawal, R., Assessment of ARIMA based prediction techniques for road traffic volume. In Fifth International Conference on Management of Emergent Digital EcoSystems, New York, USA, 2013, pp. 246–251.
10. Stram, D. O. and Wei, W. W. S., Temporal aggregation in the ARIMA process. *J. Time Series Anal.*, 1986, **7**(4), 279–292.
11. Song, Q. and Chissom, B. S., Forecasting enrolments with fuzzy time series part-I. *Fuzzy Sets Syst.*, 1993, **54**, 1–9.
12. Gokhale, R. and Sohoni, M., Detecting appropriate groundwater-level trends for safe groundwater development. *Curr. Sci.*, 2015, **108**(3), 395–404.
13. Groundwater Information System, Government of India (on-line); http://gis2.nic.in/cgwb/Gemsdata.aspx (accessed October 2014).
14. Goulão, M., Fonte, N., Wermelinger, M. and Abreu, F. B., Software evolution prediction using seasonal time analysis: a comparative study. In 16th European Conference on Software Maintenance and Reengineering, Szeged, Hungary, 2012, pp. 213–222.
15. Wu, W., Zhang, W., Yang, Y. and Wang, Q., Time series analysis for bug number prediction. In Second International Conference on Software Engineering and Data Mining, Chengdu, China, 2010, pp. 589–596.
16. Crye, J. D. and Chan, K. S., *Time Series Analysis with Applications in R*, Springer Verlag, New York, 2010, 2nd edn, ISBN: 978-0-387-75958-6.
17. Shumway, R. H. and Stoffer, D. S., *Time Series Analysis and its Application with R Examples*, Springer, New York, 2010, 3rd edn, ISBN 978-1-4419-7864-6.
18. Contreras, J., Espínola, R., Nogales, F. J. and Conejo, A. J., ARIMA models to predict next day electricity prices. *IEEE Trans. Power Syst.*, 2003, **18**(3), 1014–1020.
19. Ahmad, S. and Latif, H. A., Forecasting on the crude palm oil and kernel palm production: seasonal ARIMA approach. In IEEE Colloquium on Humanities, Science and Engineering Research, Penang, Malaysia, 2011, pp. 939–944.
20. Singh, L. L., Abbas, A. M., Ahmad, F. and Ramaswamy, S., Predicting software bugs using ARIMA model. In 48th Annual Southeast Regional Conference, New York, USA, 2010.
21. Nielsen, H. B., Univariate time series analysis; ARIMA models. *Econometrics*, 2005, **2**, 1–21.
22. Chatfield, C., *Time Series Forecasting*, Chapman & Hall/CRC, Boca Raton, Florida, USA, 2000, ISBN: 1-58488-063-5.
23. National Program on Technology Enhanced Learning, Government of India (on-line); http://nptel.ac.in/courses/105108079/ (accessed September 2014).
24. Hagan, M. T. and Behr, M., The time series approach to short term load forecasting. *IEEE Power Eng. Rev.*, 1987, **2**(8), 56–57.
25. Gentleman, R., Ihaka, R. and Bates, D., The R project for statistical computing (on-line); http://www.r-project.org
26. Yan, Z., Traj-ARIMA: a spatial-time series model for network-constrained trajectory. In Second International Workshop on Computational Transportation Science, New York, USA, 2010, pp. 11–16.
27. Engle, R. F. and White, H., *Co-integration, Causality and Forecasting*, EconPapers, Oxford University Press, Oxford, UK, 1999, pp. 1–44.
28. Chen, S. M., Forecasting enrollments based on fuzzy time series. *Fuzzy Sets Syst.*, 1996, **81**, 311–319.
29. Song, Q. and Chissom, B. S., Forecasting enrollments with fuzzy time series part-II. *Fuzzy Sets Syst.*, 1994, **62**, 1–8.
30. Chen, S. M. and Hwang, J. R., Temperature prediction using fuzzy time series. *IEEE Trans. Syst., Man Cybernatics – Part B*, 2000, **30**(2), 263–275.
31. Huarng, K., Effective length of intervals to improve forecasting in fuzzy time series. *Fuzzy Sets Syst.*, 2001, **123**, 387–394.