Stoch Environ Res Risk Assess (2005) 19: 326–339
DOI 10.1007/s00477-005-0238-4

ORIGINAL PAPER

A. K. Mishra · V. R. Desai

Drought forecasting using stochastic models

Published online: 21 June 2005
© Springer-Verlag 2005

Abstract Drought is a global phenomenon that occurs virtually in all landscapes causing significant damage both in natural environment and in human lives. Due to the random nature of contributing factors, occurrence and severity of droughts can be treated as stochastic in nature. Early indication of possible drought can help to set out drought mitigation strategies and measures in advance. Therefore drought forecasting plays an important role in the planning and management of water resource systems. In this study, linear stochastic models known as ARIMA and multiplicative Seasonal Autoregressive Integrated Moving Average (SARIMA) models were used to forecast droughts based on the procedure of model development. The models were applied to forecast droughts using standardized precipitation index (SPI) series in the Kansabati river basin in India, which lies in the Purulia district of West Bengal state in eastern India. The predicted results using the best models were compared with the observed data. The predicted results show reasonably good agreement with the actual data, 1–2 months ahead. The predicted value decreases with increase in lead-time. So the models can be used to forecast droughts up to 2 months of lead-time with reasonably accuracy.

Keywords Kansabati catchment · ARIMA model · SARIMA model · SPI · Forecasting

1 Introduction

Drought is considered by many to be the most complex but least understood of all the natural hazards affecting more people than any other hazard. Drought is a normal feature of climate and its occurrence appears inevitable. However, much confusion remains within the scientific and policy-making community about its characteristics. Research has shown that the lack of a precise and objective definition of drought in specific situations has been an obstacle in understanding drought. This has led to indecision and inaction on the part of managers, policy makers, and others (Wilhite and Glantz 1985, Wilhite et al. 1986). The global climate change in recent years is likely to enhance the frequency of droughts. While much of the weather that we experience is brief and short-lived, drought is a more gradual phenomenon, slowly affecting an area and tightening its grip with time. In severe cases, drought can last for many years, and can have devastating effects on agriculture and water supplies. It is very difficult to determine when a drought begins or ends. A drought can be short, lasting for just a few months, or it may persist for years before climatic conditions return to normal. Like many countries drought is common in India also and these drought areas are mainly confined to the southern and western parts of the country. In addition, there are few more drought-prone pockets in other parts of India. Out of 3.28 million km^2 of geographical area in India about 1.07 million km^2 of land is subjected to different degrees of water stress and drought conditions.

One of the basic deficiencies in mitigating the effects of drought is the inability to forecast drought conditions reasonably well in advance by either few months or seasons. Accurate drought forecasts would enable optimal operation of irrigation systems. Yevjevich (1967) was among the first at attempting a prediction of properties of droughts using the geometric probability distribution, defining a drought of k years as k consecutive years when there are not adequate water resources. Saldariaga and Yevjevich (1970) continued the development of run theory, incorporating concepts of time series analysis in formulations to predict drought occurrence. Sen (1976, 1977) continued this work in applying run theory to water resource predictions, evaluating run sums of annual flow series. Rao and

A. K. Mishra (✉) · V. R. Desai
Department of Civil Engineering,
Indian Institute of Technology, Kharagpur, 721302, India
E-mail: akmishra@civil.iitkgp.ernet.in
E-mail: venkapd@civil.iitkgp.ernet.in

Padmanabhan (1984) investigated the stochastic nature of yearly and monthly Palmer's drought index (PDI) and to characterize them using valid stochastic models to forecast and to simulate PDI series. Moye et al. (1988) developed a pertinent probability distribution based on difference equations to forecast drought of prespecified duration and average drought length of desired period. Sen (1990) derived exact probability distribution functions of critical droughts in stationary second order Markov chains for finite sample lengths on the basis of the enumeration technique and predicted the possible critical drought durations that may result from any hydrologic phenomenon. Kendel and Dracup (1992) proposed a drought event generator using alternating renewal–reward model. Loaiciga and Leipnik (1996) modeled the occurrence of drought events by the renewal processes. Lohani and Loganathan (1997) used PDSI in a non-homogenous Markov chain model to characterize the stochastic behavior of drought and based on these drought characterizations an early warning system is used for drought management. Chung and Salas (2000) used low-order discrete autoregressive moving average (DARMA) models for estimating the occurrence probabilities of drought events. Kim and Valdes (2003) used PDSI as drought parameter to forecast drought in the Conchos River basin in Mexico using conjunction of dyadic wavelet transforms and neural network.

There has been considerable research on modeling for various aspects of drought, such as the identification and prediction of its duration and severity. It is rather easy to sense that a drought has set in, particularly during a cropping season. There is a need to develop methods and techniques to forecast the initiation/termination point of droughts. The ARMA models, pattern recognition techniques, physically based models using Palmer drought severity index (PDSI), standardized precipitation index (SPI), a moisture adequacy index involving Markov chains, or the notion of conditional probability, seems to offer a potential to develop reliable and robust forecasts towards this goal (Panu and Sharma 2002). Such research efforts would be of considerable importance in mitigating the impacts of agricultural droughts and/or short-term hydrological droughts.

The stochastic models presented in this paper are based on SPI as drought index. The SPI is used in this study for the following advantages, which are discussed by Hayes et al. (1999).

- The primary reason is that SPI is based on rainfall alone, so that drought assessment is possible even if other hydro-meteorological measurements are not available.
- The SPI is also not adversely affected by topography.
- The SPI is defined over various timescales; this allows it to describe drought conditions over a range of meteorological, hydrological and agricultural applications.
- The fourth advantage of SPI comes from its standardization, which ensures that the frequencies of extreme events at any location and on any time scale are consistent.
- The SPI also detects moisture deficit more rapidly than PDSI, which has a response time scale of approximately 8–12 months. Hughes and Saunders (2002) have demonstrated that SPI-12 exhibits a close correspondence to the PDSI in studying drought climatology for Europe.

The main objective of present study is to calculate time series of SPI for multiple time scales and to develop valid stochastic models to forecast and simulate SPI series. Since the calculation of SPI is based on the moving sum of rainfall series as part of the procedure, linear stochastic models will be useful for forecasting the SPI series. The importance for considering the present study area is because of the following reasons. (a) The basin is situated in an underdeveloped part of India, so no study was conducted earlier for drought analysis, (b) The people in the region are very poor and they mostly depend on agriculture, so it is very important to analyse the drought in the basin, and (c) The basin was affected by severe droughts in the years 1965–1967 and around 1980s, which was for a longer duration. The severity of drought in 1990s was for short period. Since the basin is affected by short-term drought frequently (Mishra and Desai 2005) it was necessary for the researchers to investigate drought in the basin.

2 Background information on application of stochastic models

The stochastic models, which are often known as time series models have been used in scientific, economic and engineering applications for the analysis of time series. Time series modeling techniques have been shown to provide a systematic empirical method for simulating and forecasting the behavior of uncertain hydrologic systems and for quantifying the expected accuracy of the forecasts. Some of the literatures dealing with different types of time series where stochastic models are as good as ANN models can be found in literature (Brace et al. 1991; De Groot and Wurtz 1991; Caire et al. 1992; Foster et al. 1992; Gorr et al. 1994).

The ARIMA model approach has several advantages over others such as exponential smoothing and neural network in particular, its forecasting capability and its richer information on time-related changes. In most time series, there is a serial correlation among observations. This characteristic is effectively considered by ARIMA model. This model also provides systematic searching stage (identification, estimation and diagnostic check) for an appropriate model. Characteristic of many types of hydrologic time series has periodically varying components. Data of this type may be modeled using a linear stochastic model that is commonly referred to as

autoregressive integrated moving average (ARIMA) model (Lewis and Ray 2002). An inherent advantage of the SARIMA family of models is that few model parameters are required for describing time series, which exhibit non-stationarity both within and across the seasons. Some useful applications of these models in seasonal river flow forecasting are reported in McKerchar and Dellur (1974), Panu et al. (1978), Cline (1981), Govindaswamy (1991) and Yurekli et al. (2005). Hydrologists have also widely used stochastic analogy for the analyzing and modeling of hydrologic time series. It is observed from literature that the type of model fits to a particular time series is problem dependent. There are two classes of stochastic models, which are described below:

2.1 Nonseasonal models

Autoregressive (AR) models can be effectively coupled with moving average (MA) models to form a general and useful class of time series models called autoregressive moving average (ARMA) models. In ARMA model the current value of the time series is expressed as a linear aggregate of p previous values and a weighted sum of q previous deviations (original value minus fitted value of previous data) plus a random parameter, However, they can be used when the data are stationary. This class of models can be extended to non-stationary series by allowing differencing of data series. These are called autoregressive integrated moving average (ARIMA) models. Box and Jenkins (1976) popularized ARIMA models. The general non-seasonal ARIMA model is AR to order p and MA to order q and operates on dth difference of the time series z_t; thus a model of the ARIMA family is classified by three parameters (p, d, q) that can have zero or positive integral values.

The general non-seasonal ARIMA model may be written as

$$\phi(B)\nabla^d z_t = \theta(B)a_t \tag{1}$$

where ϕ (B) and θ (B) are polynomials of order p and q, respectively.

$$\phi(B) = (1 - \phi_1 B - \phi_2 B^2 - \cdots \phi_p B^p) \tag{2}$$

and

$$\theta(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots \theta_q B^q) \tag{3}$$

2.2 Seasonal models

Many time series contain cyclic features. Very often in hydrologic time series these features are of an annual cycle primarily due to the earth's rotation about the sun. Such series are cyclically non-stationary. Once the deterministic cyclic effects have been removed from a series, the ARIMA approach can be applied to obtain a linear model for the stochastic part of the series. Box et al. (1994) have generalized the ARIMA model to deal with seasonality, and define a general multiplicative seasonal ARIMA model, which are commonly known as SARIMA models. In short notation the SARIMA model described as ARIMA (p, d, q) (P, D, Q)$_s$, where (p, d, q) is the non-seasonal part of the model and (P, D, Q)$_s$ is the seasonal part of the model, which is mentioned below:

$$\phi_p(B)\Phi_P(B^s)\nabla^d\nabla_s^D z_t = \theta_q(B)\Theta_Q(B^s)a_t \tag{4}$$

where p is the order of non-seasonal autoregression, d the number of regular differencing, q the order of non-seasonal MA, P the order of seasonal autoregression, D the number of seasonal differencing, Q the order of seasonal MA, s is the length of season.

The time series model development consists of three stages, i.e. identification, estimation and diagnostic check (Box and Jenkins 1976), which are available in literatures and time series books.

3 Case study

The physical area considered in this study is the portion of the Kansabati river basin (Fig. 1) upstream from the Kangsabati dam, in the extreme western part of West Bengal state in eastern India. The region has an area of 4265 km^2. The elevation ranges from minimum of 110 m to a maximum of 600 m. The average elevation of the region is approximately 200 m. The basin experiences very hot summer and temperature in the region reaches up to 45°C in May and June. Generally the dry periods are accompanied with high temperatures, which lead higher evaporation affecting natural vegetation and the agriculture of the region along with larger water resource sectors. The mean annual precipitation in the

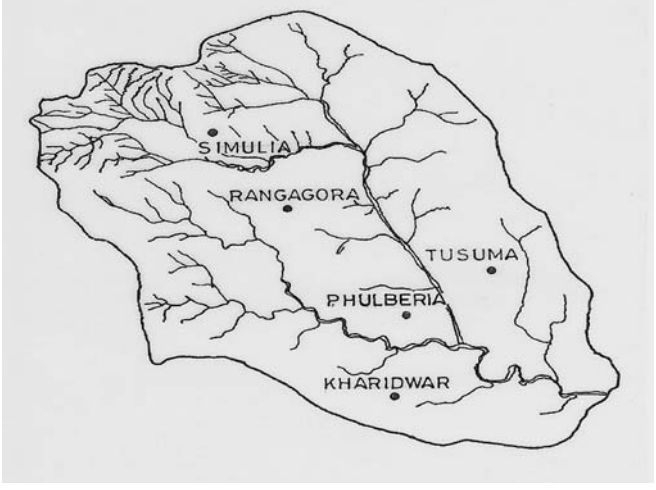

Fig. 1 Location of precipitation stations used in the study

basin is about 1268 mm. Mainly three rivers are contributing the flow in Kansabati catchment that is Kansai, Kumari, and Tongo. There is Kansabati dam constructed at the confluence of three rivers in Purulia district. The waters primarily used for irrigation. The major crops grown in the catchment are paddy, maize, pulses, and vegetables. It is considered a drought prone area with irregular rainfall and the soils are mostly laterite in nature having a low water holding capacity. About 50–60% of the study area is upland, which is managed by the poor farmers. Lands are mostly monocropped having limited surface irrigation facilities. The water demand due to the extensive cultivation leads to over-exploitation of groundwater resources. The over-exploitation of groundwater, especially in summer has led to degradation of water resources. Irrigated crops are not widespread because there is not always enough water for the purpose. For this study, five rain gauge stations were considered and monthly rainfall data was procured for the period from 1965 to 2001. The statistical properties of rainfall series along with their geographic location are shown in Table 1. The data for these backward areas are very difficult to get and that too not for long periods. Wei (1990) stated that the minimum number of 50 observations is needed to build reasonable ARIMA model, which is reasonably correct in the present study. The minimum rain gauge density for flat regions of temperate Mediterranean and Tropical zone is one station per 600–900 km^2 (according to World Meterological Organization), which seems reasonably correct for present study. In India the rain gauge density is about 1.7 gauges/1000 km^2 area. Understanding the difficulties about rain gauge density and importance of the basin, the present work is carried out based on the average rainfall over the basin.

4 Standardized precipitation index (SPI) for drought analysis

In the present study SPI is used as drought index, due to its several advantages as mentioned earlier. A deficit of precipitation impacts on soil moisture, stream flow, reservoir storage, and groundwater level, etc. at different time scales. McKee et al. (1993) developed the SPI to quantify precipitation deficits on multiple time scales. Shorter or longer time scales may reflect lags in the response of different water resources to precipitation anomalies. McKee et al. (1993) defined the criteria for a "drought event" for any time scales. A drought event occurs at the time when the value of SPI is continuously negative. The event ends when the SPI becomes positive. Weather classification based on SPI is shown in Table 2.

After the conceptualization of SPI, many researchers in drought studies have used it. Bussay et al. (1999) and Szalai and Szinell (2000) assessed the utility of SPI for describing droughts in Hungary. They concluded that SPI was suitable for quatifying most types of drought events. Stream flow was best described by SPIs with time scale of 2–6 months. Strong relationships between SPI and ground water level were found at time scales of 5–24 months. Agricultural drought (quantified by soil moisture content) was indicated by the SPI on a scale of 2–3 months. More recently Lana et al. (2001) have used the SPI to investigate patterns of rainfall over Catalonia, Spain while Hughes and Saunders (2002) have studied drought climatology for the entire Europe based on SPI values at time scales of 3, 6, 9, 12, 18, and 24 months for the period 1901–1999. Mishra and Desai (2005) studied the spatial and temporal variation of drought over Kansabati basin in India using SPI as the drought index.

4.1 Computation of SPI

The SPI is computed by fitting a probability density function to the frequency distribution of precipitation summed over the time scale of interest. This is performed separately for each month (or any other temporal basis of the raw precipitation time series) and for each location in space. Each probability density function is then transformed into a standardized normal distribution.

The gamma distribution is defined by its probability density function as

$$g(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} \quad \text{for } x > 0 \tag{5}$$

where $\alpha(>0)$ is a shape factor, $\beta(>0)$ is a scale factor, and $x > 0$ is the amount of precipitation. $\Gamma(\alpha)$ is the gamma function which is defined as

Table 1 Raingauge stations in the Kansabati river basin

Raingauge stations	Elevation (m) (a.m.s.l)	Geographic coordinates		Statistical properties of annual rainfall series (1965–2001)					
		Latitude	Longitude	Mean (mm)	Max (mm)	Min (mm)	Standard deviation	Skewness	Kurtosis
Simulia	220.97	23° 10′	86° 22′	1300.68	1840	828	260.32	0.174	−0.605
Rangagora	222.92	23° 4′	86° 24′	1152.57	1729	743	219.1	0.782	0.656
Tusuma	158.6	23° 08′	86° 43′	1268.3	1683	746	239.31	−0.221	−0.547
Kharidwar	135.96	23° 00′	86° 38′	1216.97	1814	827	248.2	0.637	−0.306
Phulberia	144.32	22° 55′	86° 37′	1345.7	2081	674	322.73	0.329	−0.006

Table 2 Weather classification based on SPI

SPI values	Class
> 2	Extremely wet
1.5–1.99	Very wet
1.0–1.49	Moderately wet
−0.99 to 0.99	Near normal
−1 to −1.49	Moderately dry
−1.5 to−1.99	Severely dry
< −2	Extremely dry

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy \tag{6}$$

Fitting the distribution to the data requires that α and β be estimated. For this Edwards and McKee (1997) suggested a method using the approximation of Thom (1958) for maximum likelihood as follows:

$$\hat{\alpha} = \frac{1}{4A}\left(1 + \sqrt{1 + \frac{4A}{3}}\right) \tag{7}$$

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} \tag{8}$$

where

$$A = \ln(\bar{x}) - \frac{\sum \ln(x)}{n} \quad \text{for } n \text{ observations} \tag{9}$$

The resulting parameters are then used to find the cumulative probability of an observed precipitation event for the given month or any other time scale.

$$G(x) = \int_0^x g(x)dx = \frac{1}{\hat{\beta}^{\hat{\alpha}}\,\Gamma(\hat{\alpha})} \int_0^x x^{\hat{\alpha}-1} e^{-x/\hat{\beta}} dx \tag{10}$$

Substituting t for $x/\hat{\beta}$ reduces Eq. 6 to incomplete gamma function:

$$G(x) = \frac{1}{\Gamma(\hat{\alpha})} \int_0^x t^{\hat{\alpha}-1} e^{-1} dt \tag{11}$$

Since the gamma function is undefined for $x = 0$ and a precipitation distribution may contain zeros, the cumulative probability becomes:

$$H(x) = u + (1-u)\, G(x) \tag{12}$$

where u is the probability of zero precipitation.

The cumulative probability, $H(x)$ is then transformed to the standard normal random variable Z with mean zero and variance one, which is the value of SPI. Following Edwards and McKee (1997), Hughes and Saunders (2002), an approximate conversion is used in this paper, as provided by Abramowitz and Stegun (1965) as an alternative:

$$Z = \text{SPI} = -\left(k - \frac{c_0 + c_1 k + c_2 k^2}{1 + d_1 k + d_2 k^2 + d_3 k^3}\right) \quad \text{for } 0 < H(x) \le 0.5 \tag{13}$$

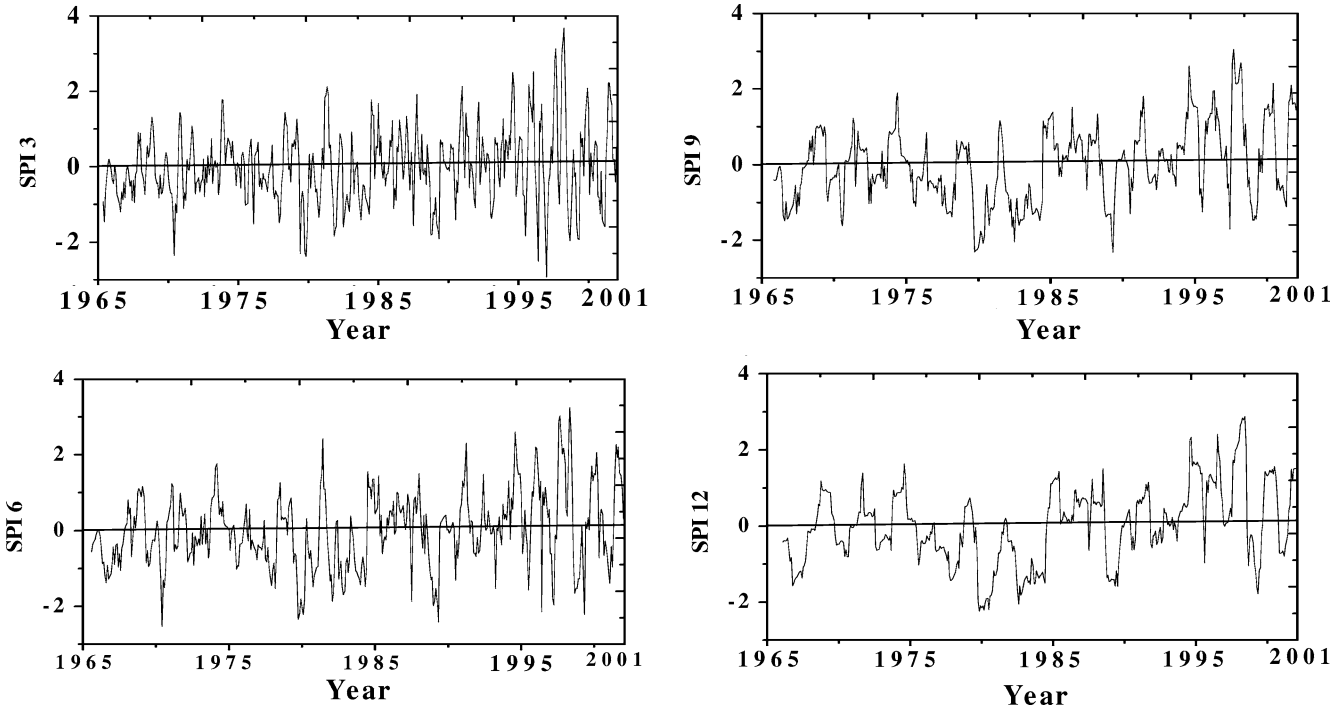

Fig. 2 SPI time series based on the average rainfall over the Kansabati basin

$$Z = \mathrm{SPI} = +\left(k - \frac{c_0 + c_1 k + c_2 k^2}{1 + d_1 k + d_2 k^2 + d_3 k^3}\right) \quad \text{for } 0.5 < H(x) < 1 \tag{14}$$

where

$$k = \sqrt{\ln\left[\frac{1}{(H(x))^2}\right]} \quad \text{for } 0 < H(x) \leq 0.5 \tag{15}$$

$$k = \sqrt{\ln\left[\frac{1}{(1 - H(x))^2}\right]} \quad \text{for } 0.5 < H(x) < 1 \tag{16}$$

and $c_0 = 2.515517$ $c_1 = 0.802853$ $c_2 = 0.010328$ $d_1 = 1.432788$ $d_2 = 0.189269$ $d_3 = 0.001308$.

The statistical tests, Kolgomorov–Smirnov (K–S) and Chi-square tests show that rainfall in the basin follows a gamma distribution. The regional time series of SPI value is calculated using the mean areal rainfall over the Kansabati basin. The time series of SPI 3, SPI 6, SPI 9, and SPI 12 are shown in Fig. 2.

5 Results and discussion

5.1 Model development

Time series model development consists of three stages identification, estimation, and diagnostic checking (Box and Jenkins 1976; Bras and Rodriguez-Iturbe 1985; Makridakis et al. 2003). The identification stage involves transforming the data (if necessary) to improve the normality and the stationarity of the time series and determining the general form of the model to be estimated. During the estimation stage the model parameters are calculated using the method of moments, least square methods, or maximum likelihood methods. Finally, diagnostic checks of the model are performed to reveal possible model inadequacies and to assist in selecting the best model. The data set from 1965 to 1994 is used for model development for SPI 3, SPI 6, and SPI 9 and SPI 12 series. The data set for SPI 24 is from 1965 to 1989 is used to have longer testing set (as these type of drought are rare). The models were developed for SPI 3, SPI 6, SPI 9, SPI 12, and SPI 24. For illustration, two examples were described briefly (for SPI 3 and SPI 12). The model identified for SPI 3 is a ARIMA model and for SPI 12 is a SARIMA model, so these two series were identified for the illustration purpose.

5.1.1 Identification

Identification of the general form of a univariate model involves two steps. First, the data series is analyzed for stationarity and normality. Appropriate differencing of the series is performed (if necessary) to achive stationarity and normality. Second, the temporal correlation structure of the transformed data are identified by examining its autocorrelation (ACF) and partial autocorrelation (PACF) functions (Box and Jenkins 1976). This information is then used to determine the general form of the univariate model to be fit.

The ACF and PACF are estimated for SPI-3, as shown in Fig. 3. The ACF and PACF show the series is stationary. The ACF is damping out in sine-wave manner with significant spikes at the first two lags. The first five values are significant in PACF, which indicates the process can be modeled as a combination of both AR and MA processes. Alternative ARIMA models were identified by considering the ACF and PACF graphs of the SPI series. This indicates a possible ARIMA (p, 0, q) model with p = 1–5 and q = 1–3. So all the combination were tried to determine the best model out of these candidate models. The model that gives the minimum Akaike Information Criterion (AIC) and Schwarz Bayesian Criterion (SBC) is selected as best fit model. Usually the model with the smallest AIC will have residuals, which resemble white noise (Makridakis et al. 2003). The mathematical formulation for the AIC (Akaike 1974) is defined as:

$$\mathrm{AIC} = -2\log L + 2m \tag{17}$$

where $m = (p + q + P + Q)$ is the number of terms estimated in the model and L denotes the likelihood function of the ARIMA models and it is a monotonically decreasing function of the sum of squared residuals. The mathematical formulation for the SBC (Schwarz 1978) is defined as:

$$\mathrm{SBC} = -2\log L + m\ln(n) \tag{18}$$

where n denotes the number of observations.

In the original time series of SPI 12, it is observed that ACF curve decays with mixture of sine and exponential curve and in PACF there is significant lag at 1, which suggests AR process. In the PACF, there are significant spikes present near lag 12, 24, and 36. So the series was seasonally differenced with 12 as period. The plot of ACF and PACF after seasonal difference is shown in Fig. 4. In the seasonal differenced series of SPI 12, it is observed that the ACF curve decay fast with a mixture of sine and exponential waves. In the PACF there is a significant spike at lag 1, which indicates an AR (1) as non-seasonal part of model. The significant spike at 12 and 24 in PACF indicates a SARIMA model. The best model out of different candidate models is identified using minimum AIC and SBC criteria. The identification of best model for different SPI series based on minimum AIC and SBC criteria is shown in Table 3.

5.1.2 Parameter estimation

After the identification of model using the AIC and SBC criteria, estimation of parameters is done. During the estimation stage, model estimates were calculated simultaneously for AR and MA parameters. Model

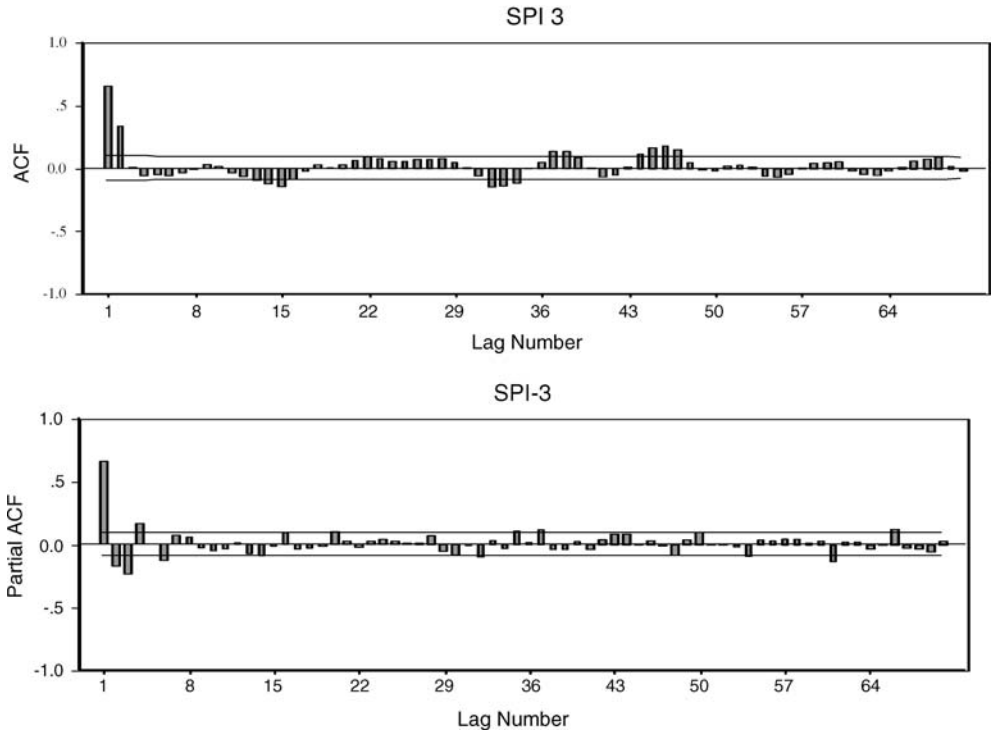

Fig. 3 ACF and PACF plots used for the selection of candidate models for SPI 3 series

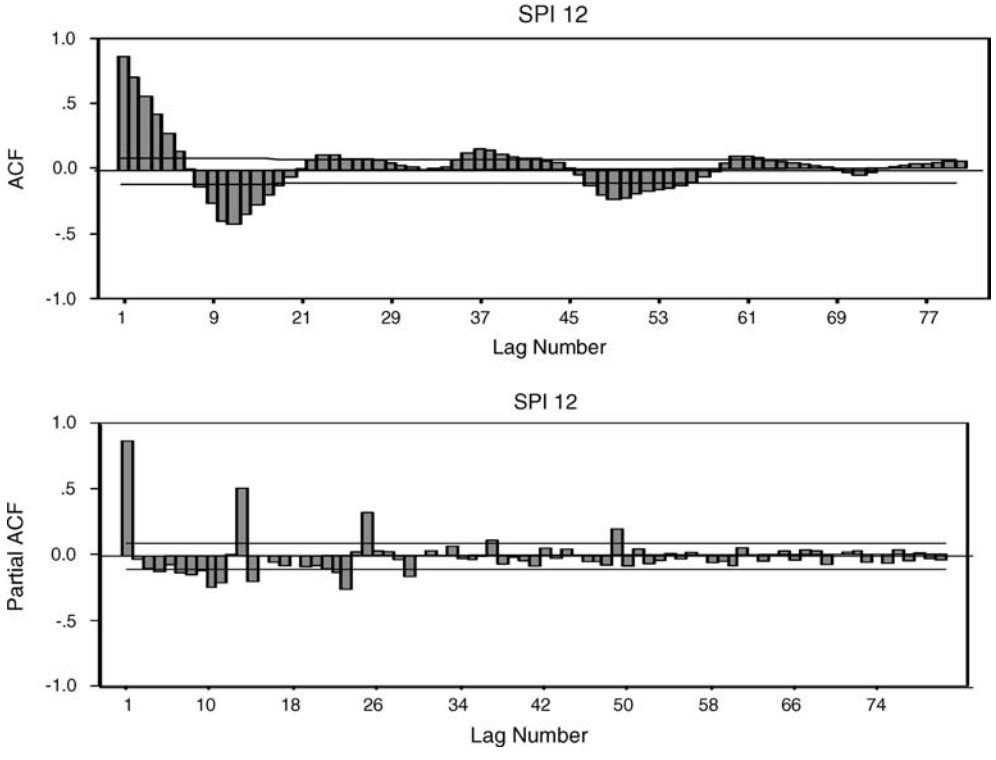

Fig. 4 ACF and PACF plots used for the selection of candidate models for SPI 12 series

estimates were made using the procedure outlined by Box and Jenkins (1976). Preliminary estimates of the parameters were computed from the ACF of the series developed in the identification stage. These preliminary estimates were then used as the starting values in an iterative Marquarardt optimization algorithm for non-linear least squares that minimize the residual sum of squares.

The value of the parameters, associated standard errors, *t*-ratios and *p*-values are listed in Table 4. The

Table 3 Comparison of AIC and SBC for selected candidate models

SPI series	MODEL	AIC	SBC
SPI-3	ARIMA (5,0,0)	393.8496	405.7677
	ARIMA (5,0,1)	393.2485	411.1502
	ARIMA (5,0,2)	392.9962	401.0815
	ARIMA (4,0,0)	394.0761	401.8105
SPI-6	ARIMA (1,0,0)	394.9166	404.1272
	ARIMA (1, 0, 0) $(1, 0, 0)_6$	399.3026	405.3239
	ARIMA (1, 0, 0) $(2, 0, 1)_6$	403.3026	415.3452
	ARIMA (1, 0, 0) $(3, 0, 1)_6$	403.3788	418.4320
	ARIMA (1, 0, 0) $(4, 0, 1)_6$	406.1496	424.2134
	ARIMA (1, 0, 0) $(1, 1, 1)_6$	394.74583	403.7777
	ARIMA (1, 0, 0) $(2, 1, 1)_6$	98.8442	410.8868
	ARIMA (1, 0, 0) $(3, 1, 1)_6$	400.8442	415.8974
SPI-9	ARIMA (1,0,0)	384.5656	387.5763
	ARIMA (1, 0, 0) $(1, 0, 0)_9$	386.9911	393.0124
	ARIMA (1, 0, 0) $(2, 0, 1)_9$	392.2528	404.2954
	ARIMA (1, 0, 0) $(2, 1, 1)_9$	405.0011	417.0436
	ARIMA (1, 0, 0) $(3, 1, 1)_9$	384.0637	487.0169
SPI-12	ARIMA (0, 0, 1)	379.4519	382.4625
	ARIMA (1, 0, 0) $(2, 1, 0)_{12}$	368.9238	377.9557
	ARIMA (1, 0, 0) $(3, 1, 0)_{12}$	371.1187	383.1613
	ARIMA (1,0,0)(0,1,1)12	369.0126	378.0339
	ARIMA (2, 0, 0) $(1, 1, 1)_{12}$	375.8339	387.8765
	ARIMA (2, 0, 0) $(2, 1, 1)_{12}$	372.1500	387.2032
SPI-24	ARIMA (1, 0, 0)	345.5579	348.5685
	ARIMA (1, 0, 0) $(1, 0, 0)_{12}$	347.5618	353.5831
	ARIMA (1, 0, 0) $(2, 0, 0)_{12}$	367.3457	376.3777
	ARIMA $(1,0,0)(1,1,0)_{24}$	373.3837	379.4050
	ARIMA $(1,0,0)(2,1,0)_{24}$	341.2924	350.3243
	ARIMA $(1,0,0)(3,1,0)_{24}$	352.4660	364.5086
	ARIMA $(1,0,0)(0,1,1)_{24}$	328.1283	334.1495
	ARIMA $(1,0,0)(1,1,1)_{24}$	353.0663	362.0982

Table 4 Statistical analysis of model parameters

SPIs series	Model parameters	Variables in the model			
		Value of parameters	Standard error	*t*-ratio	$P < 0.05$
SPI 3	ϕ_1	0.7219	0.0662	10.91	0.000
	ϕ_2	−0.794	0.0766	−10.37	0.000
	ϕ_3	0.2617	0.0775	3.38	0.001
	ϕ_4	0.1445	0.0669	2.16	0.031
	ϕ_5	−0.2098	0.0534	−3.93	0.000
	θ_1	−0.0404	0.0466	−0.87	0.387
	θ_2	−0.8932	0.0367	−24.35	0.000
SPI 6	ϕ_1	0.7833	0.0308	25.42	0.000
	$\Phi 1$	−0.3157	0.0477	−6.62	0.000
	$\Theta 1$	0.9675	0.0204	47.46	0.000
SPI 9	ϕ_1	0.9170	0.0199	46.03	0.000
	Φ_1	−0.5760	0.0529	−10.89	0.000
	Φ_2	−0.3176	0.0591	−5.37	0.000
	Φ_3	−0.1224	0.0532	−2.30	0.022
	Θ_1	0.9537	0.0243	39.30	0.000
SPI 12	ϕ_1	0.9684	0.0125	77.67	0.000
	Φ_1	−0.6083	0.0476	−12.78	0.000
	Φ_2	−0.2764	0.0484	−5.72	0.000
SPI 24	ϕ_1	0.9529	0.0156	61.19	0.000
	θ_1	0.8979	0.0413	21.75	0.000

standard errors calculated for the model parameters were generally small compared to the parameter values. Therefore most of the estimates of parameters are statistically significant and these parameters should be included in the models.

5.1.3 Diagnostic check

The model having been identified and the parameters estimated, diagnostic checks are then applied to the fitted model to verify that the model is adequate. The

residuals are studied to see if any pattern remains unaccounted for. For a good forecasting model, the residuals left over after fitting the model should be white noise. All validation tests are carried out on the residual series only. The tests are summarized briefly in the following paragraph.

5.1.3.1 ACF and PACF of residuals The residual ACF function (RACF) should be obtained to determine whether residuals are white noise. There are two useful applications related to RACF for the independence of residuals. The first is the correlogram drawn by plotting r_k against lag k, where r_k is the residual ACF function. If some of the RACFs are significantly different from zero, this may indicate that the present model is inadequate. The ACF and PACF of residuals for SPI 3 and SPI 12 are shown in Figs. 5a, b and 6a, b. Most of the values lies within confidence limits except very few individual correlations appear large compared with the confidence limits, which is expected among 65 lags. The figure indicates no significant correlation between residuals.

5.1.3.2 Histogram of residuals Histograms of residuals for SPI-3 and SPI-12 are shown in Figs. 5c and 6c. These histogram shows the residuals are normally distributed. This signifies residuals to be white noise.

5.1.3.3 Normal probability of residuals The graph of the cumulative distribution for the residual data normally appears as a straight line when plotted on normal probability paper (Chow et al. 1988). The plots of SPI 3

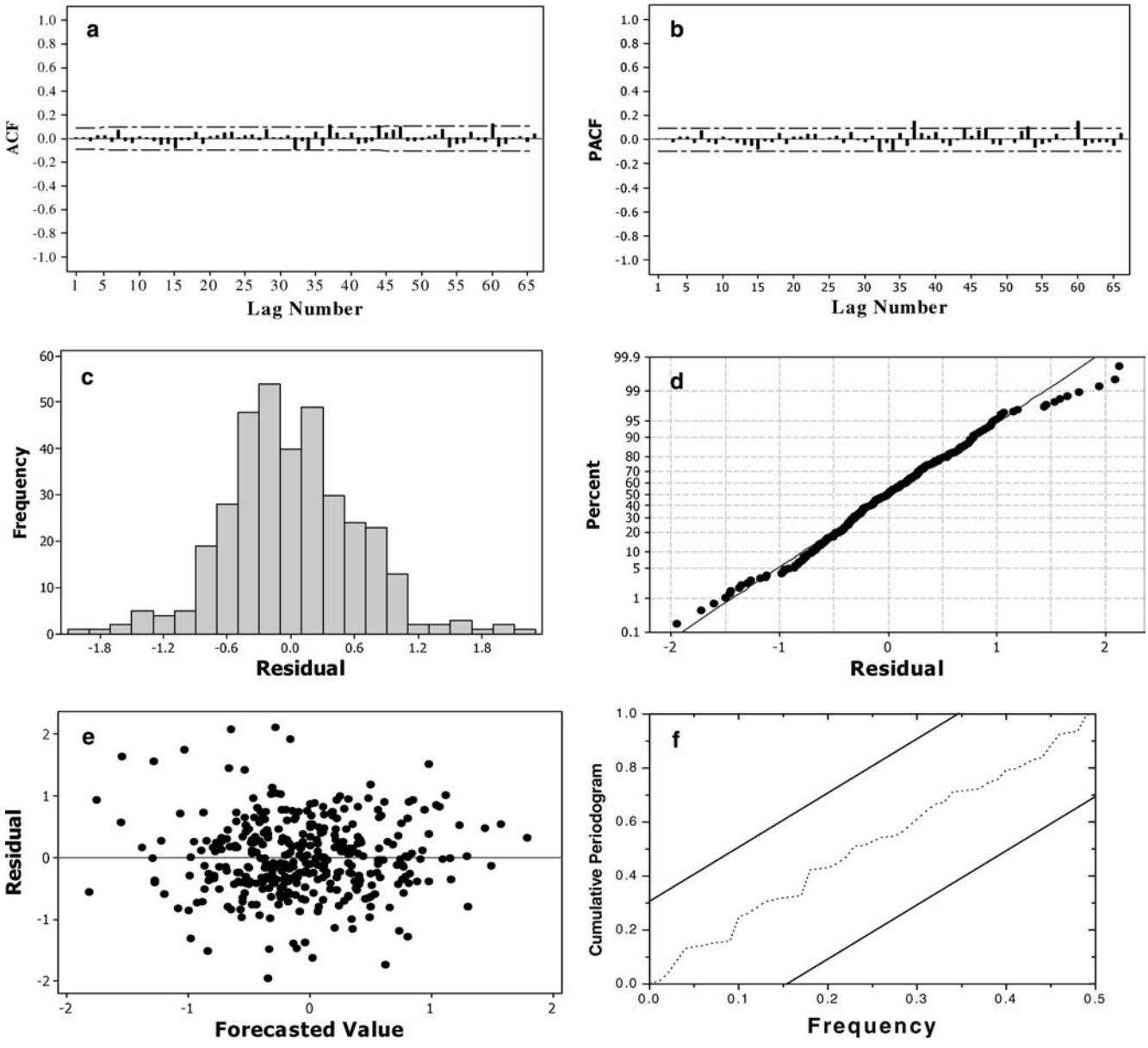

Fig. 5 Diagnostic check for best fitted model for SPI 3 series

and SPI 12 are shown in Figs. 5d and 6d. The figures show the normal probability plot of the residuals look fairly linear, the normality assumptions of the residuals hold (Durbin 1960).

5.1.3.4 Residual values versus forecast values Residuals are plotted against forecast values. The SPI-3 and SPI-12 are shown in Figs. 5e and 6e. These figures indicate residuals are evenly distributed around mean, which shows models are adequate (Govindaswamy 1991).

5.1.3.5 Periodogram check The significance of periodicities in the residual series can be tested using the cumulative periodogram test, also known as Bartletts test (Bartlett 1946). This test provides an effective means for the detection of periodic nonrandomness. If a significant periodicity is observed, the next significant periodicity will be detected by carrying out the test from which the first periodicity is removed, and so on. The test is briefly described below.

The periodogram of a time series a_t, where $t = 1, 2, 3,..., n$, is defined as

$$\gamma_i^2 = \frac{2}{n}\left[\left(\sum_{t=1}^{n} a_t \cos(2\pi f_i t)\right)^2 + \left(\sum_{t=1}^{n} a_t \sin(2\pi f_i t)\right)^2\right] \tag{19}$$

Where $f_i = i/n$ is the frequency $i = 1, 2,..., N/2$ here N the number of observations is even.

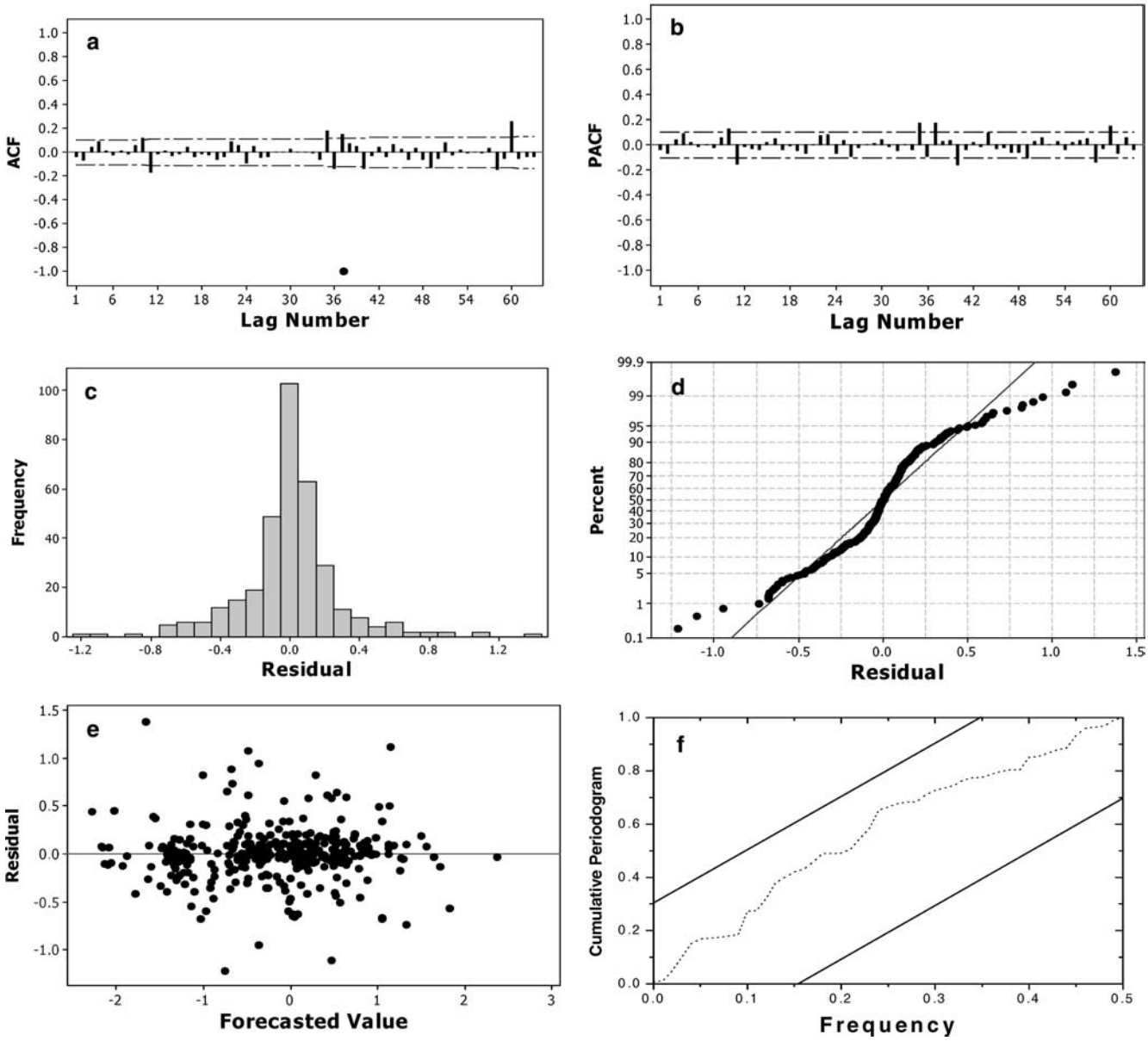

Fig. 6 Diagnostic Check for best-fitted model for SPI 12 series

Table 5 K-S test and $Q(r)_{stat}$ calculation of residuals for SPIs series

SPI	Model	K–S test		$Q(r)_{stat}$	Degrees of freedom	χ^2 distribution
		D_{tab}	D_{cal}			
SPI 3	ARIMA (5, 0, 2)	0.1461	0.0407	57.2939	43	59.282
SPI 6	ARIMA (1, 0, 0) $(1, 1, 1)_6$	0.1461	0.0768	54.3466	47	63.978
SPI 9	ARIMA (1, 0, 0) $(3, 1, 1)_9$	0.1461	0.1124	54.2845	45	61.63
SPI 12	ARIMA (1, 0, 0) $(2, 1, 0)_{12}$	0.1461	0.1268	62.8763	47	63.978
SPI 24	ARIMA$(1,0,0)(0,1,1)_{24}$	0.1461	0.1315	64.8691	48	65.152

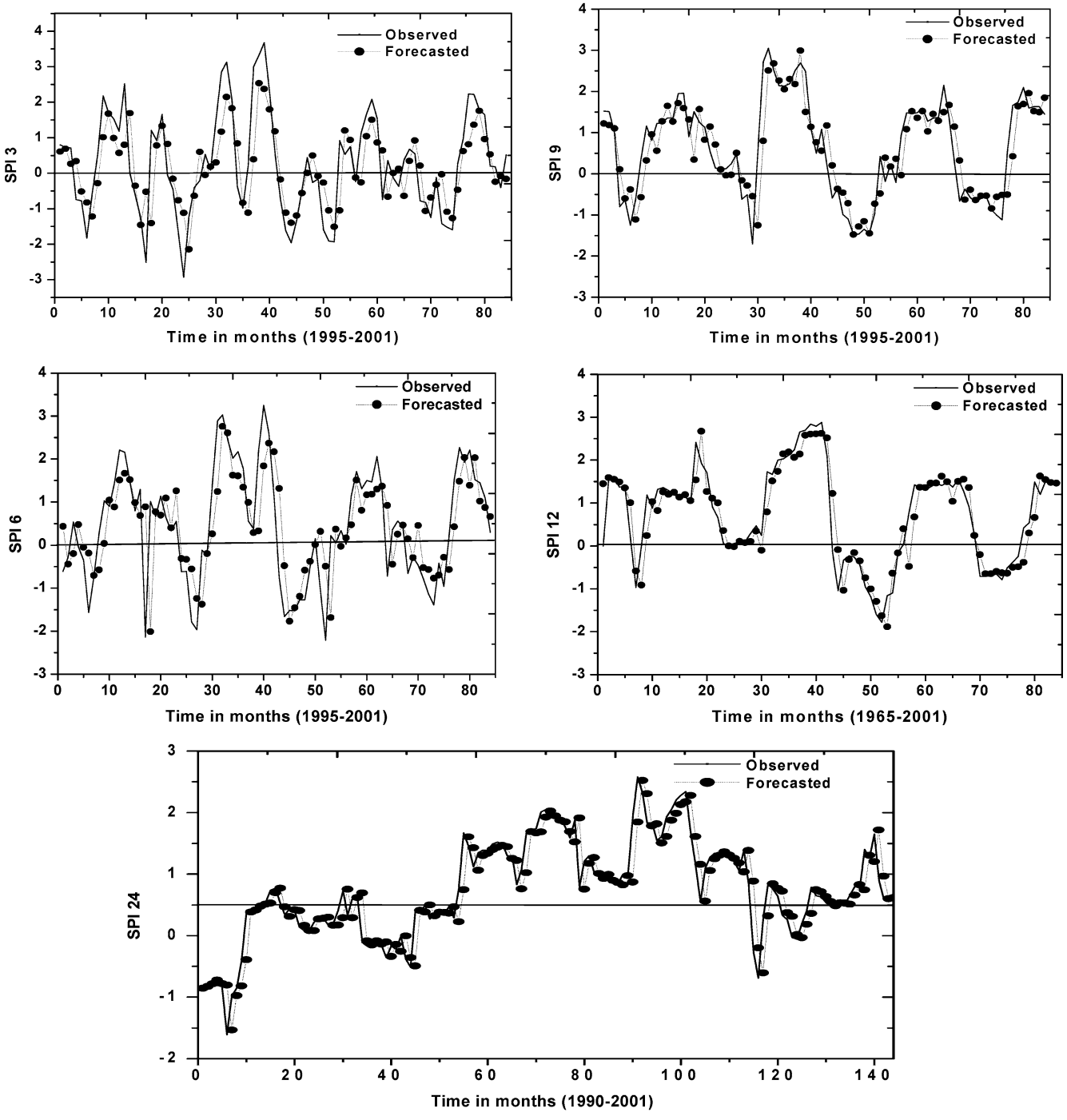

Fig. 7 Comparison of observed data with predicted data using best ARIMA models

Compute

$$g_i = \frac{\sum_{j=1}^{i} \gamma_j^2}{\sum_{i=1}^{N/2} \gamma_i^2} \quad (20)$$

It is observed that g_i lies between 0 and 1. The plot of $g_i$versus f_i is known as cumulative periodogram. If all the values of g_i lie within the significance band, then there is no significant periodicity present in the series.

The cumulative periodogram for the residuals of different series are shown in Figs. 5f and 6f. It is observed all the values of g_i lie within the significance band, which confirms that no significant periodicity is present in the residual series at 95% confidence level.

5.1.3.6 Portmantateau lack-of-fit test The modified Ljung–Box–Pierce statistics proposed by Ljung and Box (1978) is used in this study to test the adequacy of the model. The modified Ljung–Box–Pierce statistic, i.e, $Q(r)$ statistic is calculated as (Makridakis et al. 2003):

$$Q(r) = n(n+2)\sum_{k=1}^{k}(n-k)^{-1}r_k^2 \quad (21)$$

Here n is the number of observations in series. The first 50 ACF of the residuals from the model are taken for calculation of $Q\ (r)_{\text{stat}}$ shown in Table 5. These $Q\ (r)_{\text{stat}}$ values are compared with χ^2distribution with respective degree of freedom at a 5% significant level. It is observed that the calculated value is less than the actual χ^2value, which signifies that the present models are adequate on the available data.

5.1.3.7 Kolgomorov–Smirnov (K–S) tests This is a non-parametric test used to test the normality of residuals from different set of models of the fit of data (Haan 1977)

$$D_{\text{cal}} = \max/P_x(x) - S_n(x)/ \quad (22)$$

where D_{cal} is the maximum deviation, $P_x(x)$ the completely specified theoretical cumulative distribution function under the null hypothesis, $S_n(x)$ is the sample cumulative density function based on n observations.

For a chosen significance level α, the value of D_{cal} statistics is compared with table D_{tab} statistics. If D_{cal} is greater than the critical value D_{tab}, the null hypothesis related to normality is rejected for the chosen level of significance.

The K–S test is used to test the normality of residuals from different set of models. It is observed that for all models the D_{cal} is less than than D_{tab} at 5% significant level, shown in Table 5.This test satisfies that the residuals are normally distributed.

5.2 Drought forecasting from selected models

The forecast was done for 1-month lead-time using the best models from historical data. The data set from 1994 to 2001 is used for validating model for SPI 3, SPI 6, SPI 9, and SPI 12. The testing data set for SPI 24 is from 1990 to 2001. The different data set is taken for SPI 24 because these droughts are rare and to include the droughts of 1990s. The plot between observed data and predicted data using the selected best model for all SPI series is shown in Fig. 7. It is observed that the predicted data follows the observed data very closely. Basic statistical properties are compared between observed and forecasted data for one month lead time, using Z-test for the means and F-test for standard deviation (Haan 1977), shown in Table 6. Since Z_{cal} values related to means were between Z-critical table values (± 1.96 for two tailed at a 5% significance level), the data shows that there is no significant difference between the mean values of observed and predicted data. Similarly, the F_{cal} values of standard deviation were smaller than the F-critical values at a 5% signif-

Table 6 Comparison of statistic properties of the observed and predicted data

SPI series	Mean observed	Mean forecasted	Decision $\lvert z_{\text{cal}}\rvert < 1.96$	Variance observed	Variance forecasted	Decision $F_{\text{cal}} < F_{\text{tab}}$
SPI 3	0.2619	0.15	0.5966	2.06	1.02	0.4952 < 1.462
SPI 6	0.4498	0.4251	0.1347	1.7446	1.0672	0.6117 < 1.462
SPI 9	0.6279	0.5919	0.2039	1.4613	1.1523	0.7886 < 1.462
SPI 12	0.7307	0.7102	0.1198	1.2663	1.1979	0.946 < 1.462
SPI 24	0.7540	0.7366	0.1851	0.6419	0.6175	0.9619 < 1.00

Table 7 Coefficient of correlation between observed and predicted data for different lead-time

SPI series	1-month lead-time	2-month lead-time	3-month lead-time	4-month lead-time	5-month lead-time	6-month lead-time
SPI 3	0.801	0.66192	0.4459	0.3234	0.2782	0.2073
SPI 6	0.799	0.6176	0.411	0.352	0.283	0.219
SPI 9	0.877	0.7298	0.595	0.5078	0.444	0.388
SPI 12	0.925	0.828	0.73	0.648	0.56	0.476
SPI 24	0.9055	0.797	0.714	0.654	0.619	0.588

icance level. Thus, the results show that predicted data preserves the basic statistical properties of the observed series. The forecast is done with 1-month to 6-month lead-time. For example 1-month ahead forecast means that during April 2000 the forecast for May 2000 is done. Coefficient of correlation between observed and predicted data for different lead-time is shown in Table 7.It is observed that with a longer and longer lead-time the coefficient of correlation decreases between observed and predicted data. Therefore the selected best models from ARIMA building approach using a time series data of SPI series can be used for the drought forecasting.

6 Conclusion

Droughts being pernicious and creeping phenomena than other climatic events, it becomes difficult to predict drought in a basin. Like many other river basins forecasting drought is of utmost importance in the Kansabati river basin for planning and optimal operation of irrigation systems as agriculture is the primary activity in the basin. This study was focused on drought forecasting using SPI as a drought indicator. The SPI is used due to its lot of advantages over other drought indices and it is used to quantify most type of drought. The seasonal ARIMA model developed for different SPI series using the correlation methods of Box and Jenkins and the AIC and SBC structure selection criteria. In the present investigation it is observed that, when the moving window series is plotted for ACF and PACF then the seasonal effect will arise due to the moving window effect. So it is important to remove these seasonal effects using moving window length as seasonal period. Among the best models it is observed that the MA parameters in non-seasonal part vanishes which may be due to cancellation of MA part of original series with MA part of filter. The stochastic models developed to forecast drought found to give reasonably good results up to 2-month lead-time. The results seem to be better for higher SPI series (SPI 9, SPI 12, and SPI 24) and even can be used up to 3-month lead-time. These results may be due to increase in filter length which reduces the noise more effectively. So it is recommended that linear stochastic models can be used in this and other hydrometerologically similar basins for forecasting SPI series of multiple time scales to know the drought severity in future. These stochastic models developed in Kansabati basin can be used for the development of a drought preparedness plan in the region so as to ensure sustainable water resource planning within the basin.

Acknowledgments The authors would like to thank two anonymous reviewers for giving valuable suggestions for improving the quality of the paper. The authors would also like to acknowledge the editor G. Christakos for the timely handling the review processes of the paper.

References

Abramowitz M, Stegun A (1965) Handbook of mathematical formulas, graphs, and mathematical tables. Dover Publications Inc., New York

Akaike H (1974) A look at the statistical model identification. IEEE Trans Automatic Control AC 19:716–723

Bartlett MS (1946) On the theoretical specification of sampling properties of auto correlated time series. J R Stat Soc B8:27

Box GEP, Jenkins GM (1976) Time series analysis forecasting and control. Holden-Day, San Francisco

Box GEP, Jenkins GM, Reinsel GC (1994) Time series analysis, 'forecasting and control.' Prentice Hall, Englewood Cliffs, NJ

Brace MC, Schmidt J, Hadlin M (1991) Comparison of the forecasting accuracy of neural networks with other established techniques. In: Proceedings of the 1st forum on application of neural networks to power systems. Seattle, WA, pp 31–35

Bras RL, Rodriguez-Iturbe I (1985) Random functions and hydrology. Addison-Wesley, Reading MA, USA

Bussay A, Szinell C, Szentimery T (1999) Investigation and measurements of droughts in Hungary. Hungarian Meteorological Service, Budapest

Caire P, Hatabian G, Muller C (1992) Progress in forecasting by neural networks. In: Proceedings of the international joint conference on neural networks, vol. 2, pp540–545

Chow VT, Maidment DR, Mays LW (1988) Applied hydrology. Mcgraw-Hill Book Company, New York

Chung CH, Salas JD (2000) Drought occurrence probabilities and risks of dependent hydrological processes. J Hydrol Eng ASCE 5(3):259–268

Cline TB (1981) Selecting seasonal streamflow models. Water Resour Res 17(4):975–984, Company Inc., New York, p 478

De Groot C, Wurtz D (1991) Analysis of univariate time series with connectionist nets: a case study of two classical examples. Neurocomputing 3:177–192

Durbin J (1960) The fitting of time series models, review of the international institute of Statistics, vol. 28, pp 233–140

Edwards DC, McKee TB (1997) Characteristics of 20th century drought in the United States at multiple timescales. Colorado State University Fort Collins, Climatology Report No. 97-2

Foster WR, Collopy F, Ungar LH (1992) Neural network forecasting of short, noisy time series. Comput Chem Eng 16(4):293–297

Gorr WL, Nagin D, Szczypula J (1994) Comparative study of artificial neural network and statistical models for predicting student grade point averages. Int J Forecast 10:17–34

Govindaswamy R (1991) Univeriate Box-Jenkins forecasts of water discharge in Missouri river. Water Resour Dev 7(3):168–177

Haan CT (1977) Statistical methods in hydrology. Iowa State Press, Iowa

Hayes MJ, Svoboda MD, Wilhite DA, Vanyarkho OV (1999) Monitoring the 1996 drought using the standardized precipitation index. Bull Am Meterol Soc 80:429–438

Hughes BL, Saunders MA (2002) A drought climatology for Europe. Int J Climatol 22:1571–1592

Kendel DR, Dracup JA (1992) On the generation of drought events using an alternating renewal-reward model. Stochastic Hydro Hydr 6(1):55–68

Kim T, Valdes JB (2003) Nonlinear model for drought forecasting based on a conjunction of wavelet transforms and neural networks. J Hydrol Eng ASCE 8(6):319–328

Lana X, Serra C, Burgue∼no A (2001) Patterns of monthly rainfall shortage and excess in terms of the standardized precipitation index. Int J Climatol 21:1669–1691

Lewis PAW, Ray BK (2002) Nonlinear modeling of periodic threshold autoregressions using TSMARS. J Time Ser Anal 23(4):459–471

Ljung GM, Box GEP (1978) On a measure of lack of fit in time series models. Biometrica 65:297–303

Loaiciga HA, Leipnik RB (1996) Stochastic renewal model of low-flow stream sequences. Stochastic Hydro Hydr 10(1):65–85

Lohani VK, Loganathan GV (1997) An early warning system for drought management using the Palmer drought index. J Am Water Resour Assoc 33(6):1375–1386

Makridakis S, Wheelwright SC, Hyndman R (2003) Forecasting methods and applications. Wiley (ASIA) Pvt Ltd., Singapore

McKee TB, Doesken NJ, Kliest J (1993) The relationship of drought frequency and duration to time scales. In: Proceedings of the 8th conference on applied climatology 17–22 January, Anaheim, CA. American Meteorological Society, Boston, MA, pp 179–184

McKerchar AI, Dellur JW (1974) Application of seasonal parametric linear stochastic models to monthly flow data. Water Resour Res 10(2):246–255

Mishra AK, Desai VR (2005) Spatial and temporal drought analysis in the Kansabati river basin, India. Int J River Basin Manag IAHR 3(1):31–41

Moye AL, Kapadia AS, Cech IM, Hardy RJ (1988) The theory of run with application to drought prediction. J Hydrol 103:127–137

Panu US, Sharma TC (2002) Challenges in drought research: some perspectives and future directions. Hydrol Sci 47(S)

Panu US, Unny TE, Ragade RK (1978) A feature prediction model in synthetic hydrology based on concepts of pattern recognition. Water Resour Res 14(2):335–344

Rao AR, Padmanabhan G (1984) Analysis and modelling of Palmers drought index series. J Hydrol 68:211–229

Saldariaga J, Yevjevich V (1970) Application of run-lengths to hydrologic series Hydrol Paper. Colorado State University Publication, Colorado State University, Fort Collins, CO

Schwartz G (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Sen Z (1976) Wet and dry periods of annual flow series. J Hydrol Div ASCE 106(HY1):99–115

Sen Z (1977) Run sums of annual flow series. J Hydrol 35:311–324

Sen Z (1990) Critical drought analysis by second order Markov chain. J Hydrol 120:183–202

Szalai S, Szinell C (2000) Comparison of two drought indices for drought monitoring in Hungary—a case study. In: Vogt JV, Somma F (eds) Drought and drought mitigation in Europe. Kluwer, Dordrecht, pp 161–166

Thom HCS (1958) A note on gamma distribution. Monthly Weather Rev 86:117–122

Wei WWS (1990) Time series analysis. Addison-Wesley Publishing, Reading, MA

Wilhite DA, Glantz MH (1985) Understanding the drought phenomenon: the role of definations. Water Int 10:111–120

Wilhite DA, Rosenberg NJ, Glantz MH (1986) Improving federal response to drought. J Climate Appl Meteorol 25:332–342

Yevjevich V (1967) An objective approach to definitions and investigations of continental hydrologic droughts. Hydrol Papers Colorado State University, Fort Collins, CO

Yurekli K, Kurunc K, Ozturk F (2005) Application of linear stochastic models to monthly flow data of Kelkit Stream. Ecol Model 183:67–75