

1 **An improved ARIMA model for precipitation simulations**

2 H.R. Wang¹, C. Wang^{1*}, X. Lin², and J. Kang²

3 1. College of Water Sciences

4 Key Laboratory for Water and Sediment Sciences Ministry of Education,

5 Beijing Normal University,

6 19 Xijiekouwai Street, Beijing, 100875 China

7 2. College of Mathematic Sciences, Beijing Normal University

8 19 Xijiekouwai Street, Beijing, 100875 China

9 * Correspondence email: chengw@knights.ucf.edu

10

11

12 **Abstract**

13 Auto Regressive Integrated Moving Average (ARIMA) models have been widely used to
14 calculate monthly time series data formed by inter-annual variations of monthly data or inter-monthly
15 variation. However, the influence brought about by inter-monthly variations within each year is often
16 ignored. An improved ARIMA model is developed in this study accounting for both the inter-annual
17 and inter-monthly variation. In the present approach, clustering analysis is performed first to
18 hydrologic variable time series. The characteristics of each class are then extracted and the correlation
19 between the hydrologic variable quantity to be predicted and characteristic quantities constructed by
20 linear regression analysis. ARIMA models are built for predicting these characteristics of each class
21 and the hydrologic variable monthly values of year of interest are finally predicted using the modeled
22 values of corresponding characteristics from ARIMA model and the linear regression model. A case
23 study is conducted to predict the monthly precipitation in Lanzhou precipitation station, China, using
24 the model, and the results show that the accuracy of the improved model is significantly higher than
25 the seasonal model, with the mean residual achieving 9.41 mm and the forecast accuracy increasing
26 by 21%.

27 **Keywords** Hydrological Process, Seasonal ARIMA model, Clustering Regression, Precipitation
28 prediction

29

30 **1. Introduction**

31 Hydrological processes are complicated; they are influenced by not only deterministic, but also
32 stochastic factors (Wang et al. 2007). The deterministic change in a hydrological process is always
33 accompanied by the stochastic change. Generally speaking, determinism includes periodicity,
34 tendency, and abrupt change. A strict deterministic hydrological process is rare. Stationary time series
35 has been widely used in hydrological data assimilation and prediction to tackle the stochastic factors
36 in hydrological processes. From the point of view of stochastic processes, hydrological data series
37 usually comprises trend term and stationary term. The basic idea of Auto Regressive Integrated
38 Moving Average (ARIMA) model, one of the most commonly used time series model, is to remove
39 the trend term of series by difference elimination, so that a nonstationary series can be transformed
40 into a stationary one. Some researchers have used ARIMA model for the analysis of hydrological
41 process without considering the effects of seasonal factors (Jin et al. 1999; Niua et al. 1998; Toth et al.
42 1999). However, most studies (Ahmad et al. 2001; Lehmann et al. 2001; Qi et al. 2006) neglected
43 stationary test and the influence from inter-monthly variation within a year. In this paper, the seasonal
44 ARIMA model is improved by removing the effect of seasonal factors, and the improved model is
45 tested through a case study. The paper is organized as follows: the ARIMA model is introduced first,
46 followed by the introduction of the issues in the currently existing ARIMA model and our proposed
47 methods to improve it. A case study is conducted and discussion is addressed finally.

48 **2. ARIMA model**

49 A hydrological time series $\{y_t, t=1,2,\dots,n\}$ could be either stationary or nonstationary. Given
50 that there are essentially no strictly deterministic hydrological processes in nature, the analysis of
51 hydrological data by means of nonstationary time series is of importance, among which ARIMA
52 model is one of the available choices.

53 **2.1 ARIMA model**

54 For a stationary time series, ARMA (p, q) model is defined as follows:

$$55 \quad y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + u_t - \theta_1 u_{t-1} - \theta_2 u_{t-2} - \dots - \theta_q u_{t-q} \quad (1)$$

56 Where p denotes the autoregressive (AR) parameters, q represents the moving average (MA)
 57 parameters, the real parameters ϕ_1, ϕ_2, \dots , and ϕ_p are called autoregressive coefficients, the real
 58 parameters θ_j ($j = 1, 2, \dots, q$) are moving average coefficients, and u_t is an independent white
 59 noise sequence, i.e. $u_t \sim N(0, \sigma^2)$. Usually the mean of $\{y_t\}$ is zero; if not, $y'_t = y_t - \mu$ is used in
 60 the model.

61 Lag operator (B) is then introduced, thus

$$62 \quad \varphi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p \quad (2)$$

$$63 \quad \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q \quad (3)$$

64 where $\varphi(B)$ is the autoregressive operator and $\theta(B)$ is the moving-average operator.

65 Then the model can be simplified as

$$66 \quad \varphi(B)y_t = \theta(B)u_t \quad (4)$$

67 If $\{y_t\}$ are nonstationary, we can obtain the stationarized sequence z_t by means of difference, i.e.,

$$68 \quad z_t = (1 - B)^d y_t = \nabla^d y_t \quad (5)$$

69 where d is the number of regular differencing. Then the corresponding ARIMA(p, d, q) model for
 70 y_t can be built (Box et al. 1997), where d is the number of differencing passes by which the
 71 nonstationary time series might be described as a stationary ARMA process.

72 **2.2 Seasonal ARIMA(p, d, q) model**

73 Most hydrological time series have obviously seasonal (quasi-periodic) variation (Box et al.
 74 1967), representing recurring of hydrological processes over a relatively (but not strictly) fixed time
 75 interval. Monthly data series often shows a seasonal period of 12 months while quarterly data series
 76 always present a period of 4 quarters. Seasonality can be determined by examining whether the

77 autocorrelation function of the data series with a specified seasonal order is significantly different
 78 from zero. For instance, if the autocorrelation coefficient of a monthly data series with new data series
 79 formed by a lag of 12 months is not significantly different from 0, the monthly data series does not
 80 have a seasonality of 12 months; if the autocorrelation coefficient is significantly different from 0, it is
 81 very likely this monthly data series has a seasonality of 12 months. A seasonal ARIMA model can be
 82 built for a data series with seasonality.

83 For a time series $\{y_t\}$, its seasonality can be eliminated after D orders of differencing with a
 84 period of S . If a further d orders of regular differencing is still needed in order to make the data
 85 series stationary, a seasonal ARIMA can be built for the data series as follows,

$$86 \quad \phi_p(B)\Phi_p(B^S)(1-B)^d(1-B^S)^D y_t = \theta_q(B)\Theta_Q(B^S)u_t \quad (6)$$

87 where P is the number of seasonal autoregressive parameter, Q is the seasonal moving average order,
 88 S is the period length (in month in this work), and D denotes the number of differencing passes.

89 **2.3 Implementation of ARIMA model**

90 The procedure of estimating ARIMA model is given by the flowchart in **Fig. 1** which involves
 91 the following steps:

92 **(1) Stationary identification.** The input time series for an ARIMA model needs to be stationary,
 93 i.e., the time series should have a constant mean, variance, and autocorrelation through time.
 94 Therefore, the stationarity of the data series needs to be identified first. If not, the non-stationary time
 95 series is then required to be stationaried. Although the stationary test, such as unit root test and KPSS
 96 test are used to identify if a time series is stationary, plotting approaches based on scatter diagram,
 97 autocorrelation function diagram, and partial correlation function diagram are often used. The latter
 98 approach can usually provide not only the information whether the testing time series is stationary but
 99 indicate the order of the differencing which is needed to stationarize the time series. In this paper, we
 100 identify the stationarity of a time series from the autocorrelation function diagram, and partial
 101 correlation function diagram.

102 If a time series is identified nonstationary, differencing is usually made to stationarize the time
103 series. In the differencing method, the correct amount of differencing is normally the lowest order of
104 differencing that yields a time series which fluctuates around a well-defined mean value and whose
105 autocorrelation function (ACF) plot decays fairly rapidly to zero, either from above or below. The
106 time series is often transformed for stabilizing its variance through proper transformation, e.g.,
107 logarithmic transformation. Although logarithmic transformation is commonly used to stabilize the
108 variance of a time series rather than directly stationarize a time series, the reduction in the variance of
109 a time series is usually helpful to reduce the order of difference in order to make it stationary.

110 **(2) Identification of the order of ARIMA model.** After a time series has been stationarized,
111 the next step is to determine the order terms of its ARIMA model, i.e., the order of differencing, d
112 for nonstationary time series, the order of auto-regression, p , the order of moving average, q , and
113 the seasonal terms if the data series show seasonality. While one could just try some different
114 combinations of terms and see what works best strictly, the more systematic and common way is to
115 tentatively identify the orders of the ARIMA model by looking at the autocorrelation function (ACF)
116 and partial autocorrelation (PACF) plots of the stationarized time series. The ACF plot is merely a bar
117 chart of the coefficients of correlation between a time series and lags of itself and the PACF plot
118 present a plot of the partial correlation coefficients between the series and lags of itself. The detailed
119 guidelines for identifying ARIMA model parameters based on ACF and PACF, can be found
120 elsewhere, e.g. Pankratz (1983). It should be noted that, to be strict, the ARIMA model built in this
121 step is actually an ARMA model with if the time series is stationary, which is in fact a special case of
122 ARIMA model with $d = 0$.

123 **(3) Estimation of ARIMA model parameters.** While least square methods (linear or nonlinear)
124 are often used for the parameter estimation, we use the maximum likelihood method (McLeod, 1983;
125 Melard, 1984) in this paper. A t -test is also performed to test the statistical significance.

126 **(4) White noise test for residual sequence.** It is necessary to evaluate the established ARIMA
127 model with estimated parameters before using it to make forecasting. We use white noise test here. If
128 the residual sequence is not a white noise, some useful information has not been extracted and the

129 model needs to be further tuned. The method is illustrated as follows.

130 Null hypothesis: $H_0 : \text{corr}(e_t, e_{t-k}) = 0 \quad \forall k, t$

131 Alternative hypothesis: $H_1 : \text{corr}(e_{t_0}, e_{t_0-k_0}) \neq 0 \quad \exists k_0, t_0$

132 The autocorrelation of the data series is measured by the autocorrelation coefficient which is
133 defined as

$$134 \quad r_k = \frac{\sum_{t=k+1}^n e_t e_{t-k}}{\sum_{t=1}^n e_t^2} \quad (k = 1, 2, \dots, m) \quad (7)$$

135 where n is the number of cases of sample of series for white noise test, m is the maximum number of
136 lag. In practice, m uses the value of $\left\lceil \frac{n}{10} \right\rceil$ when n is very large and $\left\lceil \frac{n}{4} \right\rceil$ when n is small.

137 When $n \rightarrow \infty$, $\sqrt{nr_k} \sim N(0,1)$.

138 The test statistics is given by

$$139 \quad Q = n(n+2) \sum_{k=1}^m \frac{r_k^2}{n-k} \quad (8)$$

140 Given the degree of confidence of $1 - \alpha$, if

$$141 \quad Q < \chi_\alpha^2(m-p-q) \quad (9)$$

142 Then Q fits the χ^2 distribution at the significance of $1 - \alpha$ and the null hypothesis is accepted.

143 **(5) Hydrological forecasting.** The linear least squares method is usually applied for
144 rainfall-runoff prediction. In general, based on the n observation values, the values of future L
145 time steps can be estimated (Kohn et al. 1986).

146 **3. Improvement of conventional ARIMA model**

147 Seasonal ARIMA models apply for time series which arranges in order with a certain time

148 interval or step, e.g., a month. However, in this case, while the seasonal ARIMA model is capable of
149 dealing with the inter-annual variation of each monthly of a monthly data series, the information of
150 inter-monthly variation of the time series may be lost. For example, after an order of 12 of seasonal
151 differencing (term S in a general seasonal ARIMA model) of a monthly time series, the original
152 monthly series has been migrated to a new time series without seasonality. A nonseasonal ARIMA
153 model is then fitted to the new time series where the inter-monthly variation of original monthly time
154 series has also migrated to the inter-monthly variation of the new series after seasonal differencing.
155 The transformation of inter-monthly variation of original monthly data to the new inter-monthly
156 variation of seasonally differenced series may result in loss of accuracy of model performance. In this
157 study, twelve individual seasonal ARIMA models for precipitation prediction for each month are built
158 from each monthly data series, e.g., the January data series from 1951 to 2000, which are referred to
159 as ARIMA models of inter-annual variation ignoring the inter-monthly variation.

160 In order to prevent from losing the inter-monthly variation information, we propose in this study
161 the following improvement to the conventional seasonal ARIMA model, which simultaneously takes
162 into account both kinds of temporal variation (inter-annual variation and inter-monthly variation).
163 Clustering analysis is first applied to classify the monthly data series and extract characteristics of
164 each data series class (Sun et al. 2005). In this study, we use Euclidean distance as the distance
165 measurement in clustering analysis. The characteristics of each data series refer to the maximum,
166 minimum, and truncated mean of the series of this class. A linear regression model is then built with
167 hydrological variable to be predicted, e.g., monthly precipitation, as dependent variables and with
168 maximum, minimum, and truncated mean of each class as independent variables in the linear
169 regression model. For example, a monthly precipitation would be described as a linear regression
170 function of the maximum, minimum, and truncated mean of the data series of a class where this
171 month's precipitation has been clustered in the clustering analysis. A conventional seasonal ARIMA
172 model is built for the maximum, minimum, and truncated mean of each class, respectively, accounting
173 for the inter-monthly variation of each characteristic variable. By this way, we are trying to avoid
174 losing the inter-monthly variation information. The implementation of the improved ARIMA model
175 involves the following procedure, as illustrated in Fig. 2.

176 i). Perform clustering analysis on monthly data, and group the months with similar
177 hydrological variation.

178 ii). Find the maximum, minimum, and truncated mean of each cluster.

179 iii). Build linear regression models and determine the associated parameters for each monthly
180 data series. For example, for the precipitation in the i -th month,

$$181 \quad y_i = a_i y_{j,\max} + b_i y_{j,\min} + c_i y_{j,\text{avg}} + d_i \quad (10)$$

182 where a_i , b_i , c_i , and d_i are the coefficients in the model for the i -th month

183 hydrologic parameter, e.g., precipitation, which need to be estimated, and $y_{j,\max}$, $y_{j,\min}$,

184 and $y_{j,\text{avg}}$ are respectively the maximum, minimum, and truncated mean of the j -th

185 class where the time series of the i -th month is identified in cluster analysis.

186 iv). Build ARIMA models for the maximum, minimum, and truncated mean of each class and
187 predict the characteristics for the time year of interest using the established ARIMA models.

188 v). Substitute the predicted characteristics into the linear regression model built in Equation (10)
189 and obtain the monthly hydrologic variable, say precipitation.

190 **4. Case study**

191 In this section, we are presenting an application of the proposed improved ARIMA model to the
192 precipitation forecasting of Lanzhou precipitation station in Lanzhou, China. Lanzhou is located in the
193 upper basin of Yellow River. It has a continental climate of mid-temperate zone, with an average
194 precipitation of 360 mm and mean temperature of 10°C. In general, rainfall seasons are May through
195 September, while drought occurs in spring and winter. The Lanzhou precipitation station is located at
196 103.70°E, 35.90°N. The monthly precipitation data from 1951 to 2000 is used for parameter
197 estimation and the monthly precipitations of 2001 are then predicted using the proposed model and
198 compared with the observation values. In order to show the improvement of this present approach, we
199 first build a conventional seasonal ARIMA model and a set of 12 ARIMA models for each monthly

200 precipitation series which account for the seasonal variation. The improved ARIMA model
201 accounting for both inter-month and inter-annual variation of monthly precipitation time series is then
202 built using the presented approach and its prediction results are compared with the conventional
203 ARIMA model and seasonal ARIMA model, as well as auto-regressive models.

204 **4.1 Conventional seasonal ARMA modeling**

205 The precipitation at the Lanzhou precipitation station from 1951 through 2001 and from 1991
206 through 2001 are plotted as shown Fig. 3 (a) and (b) respectively. The two figures show less
207 precipitation in winter and spring and more in summer and autumn. Fluctuation occurs to the data
208 during high precipitation seasons. Using power transformation with an order of 1/3, fluctuations at
209 high values are removed and the data become stationary, as shown in Fig. 3(c). According to
210 autocorrelation and partial correlation functions, as shown in Fig. 4, seasonal term with a period of 12
211 exists. With the difference elimination method, the order of the model can be determined from, and
212 the following seasonal ARIMA model is obtained.

$$213 \quad (1 - B^{12})y_t = (1 - \theta_1 B)(1 - \theta_2 B^{12})u_t \quad (11)$$

214 The maximum-likelihood method is then used for parameter estimation and the results are listed
215 in Table 1. As shown in Table 1, parameter estimation is statistically significant. A white noise test is
216 performed for the residual sequence. If the test does not pass, the model needs to be improved. As
217 shown in Table 2, with a significance level of 5%, the test is passed, i.e., the useful information is
218 extracted and the model is acceptable.

219 **4.2 Individual ARIMA model for each month data series**

220 As discussed in Section 2.2, the data can be classified into 12 groups associated with each month
221 respectively. Stationary identification, stationary treatment, model identification, parameter estimation
222 and residual test are performed for the 12 groups of data. A total of 12 ARIMA models are built and
223 the estimated parameters are shown in Table 3.

224 4.3 The improved ARIMA model based on clustering and regression analysis

225 Box-Cox transformation is applied as a pretreatment of data for clustering analysis in order to
226 stable the variance of the monthly precipitation data series. Given that the precipitation has values of
227 zero resulting in negative infinity in the transformation, Box-Cox transformation (Thyer et al., 2002;
228 Meloun et al., 2005; Ip et al., 2004) is corrected as follows.

$$229 \text{ Data after transformation} = \begin{cases} \frac{(\text{original data} + 1)^\alpha - 1}{\alpha} & \alpha \neq 0 \\ \log(\text{original data}) & \alpha = 0 \end{cases}$$

230 After Box-Cox transformation, as shown in Fig. 6, the data are much more symmetric than the
231 original data series, which is helpful for the later clustering analysis. Moreover, it can be seen that
232 there are many zero precipitation values in the raw monthly precipitation data series and so does the
233 transferred data. This indicates that the samples of data sequence may not be from one individual
234 population but from multiple populations which further implies the necessarily of clustering analysis
235 for the data series. Clustering analysis with Euclidean distance is then applied which indicates that the
236 monthly precipitation sequences can be clustered into three classes, as shown in Fig. 7.

$$237 \begin{cases} \text{Class 1: Jan., Feb., Nov., and Dec.} \\ \text{Class 2: Mar., Apr., and Oct.} \\ \text{Class 3: May, Jun., Jul., Aug., and Sep.} \end{cases}$$

238 It is interesting that the clustering results are mostly coincides with the precipitation season. For
239 example, Class 1 looks like corresponding to the drought season while Class 3 corresponds to the
240 rainfall season. After the clustering analysis to the monthly precipitation time series, the
241 characteristics of each class, i.e., maximum, minimum, and truncated mean, are identified, as shown
242 in Fig. 8. Whereas fluctuations in the mean and minimum data series are relatively small, relatively
243 larger variation are shown in the maximum data series.

244 Linear regression models for each monthly precipitation are fitted using the characteristics of
245 each class where the monthly precipitation data series is located. The parameters corresponding to
246 each linear regression model are presented in Table 4 which pass the t -test at the significance of 0.05

247 indicating that those linear models fit their data series well respectively. Following the steps described
248 in Section 2.3, nine ARIMA modes are built for each of the characteristic variables of each class. The
249 estimated parameters are shown in Table 5. Auto-regressive models with orders of 24 and 36, or AR
250 (24) and AR (36), are also fitted to the monthly precipitation time series for comparative study with
251 the improved ARIMA model and conventional ARIMA model.

252 **5. Results and discussion**

253 The monthly precipitations of 2001 are predicted using the improved ARIMA model as well as
254 the conventional seasonal ARIMA model, the 12 seasonal ARIMA models for the precipitation of
255 each month, and AR(24) and AR(36) models, the prediction results shown in Table 6 and Fig. 9. The
256 absolute error of each method is 9.41, 11.49, 11.78, 17.05, and 17.82 mm for the improved ARIMA
257 model, conventional ARIMA model, individual ARIMA for each month data series, AR(24), and
258 AR(36), respectively, indicating that the improved ARIMA presented in this paper performs the best
259 with the smallest errors. Compared with the conventional ARIMA model, the improved ARIMA
260 model increases the prediction accuracy by 24%.

261 The conventional ARIMA model predicts accurately for March, June, August, and November but
262 mismatches the other months' precipitation. It predicts more accurately for October precipitation than
263 the improved ARIMA model. The 12 individual ARIMA models for each month data series performs
264 similarly to the conventional ARIMA model. The overall performance of AR(24) model does not
265 show difference from that of AR(36) model; neither models perform as good as the improved ARIMA
266 model or the conventional ARIMA model. However, the AR models give a better prediction for
267 September precipitation of 2001 than the other two models.

268 While the improved ARIMA model catches the correct trend overall and predicts the monthly
269 precipitation in most months with high accuracy, it predicts highly accurately for the dry seasons,
270 such as January, February, March, November, and December. However, it overestimates the
271 precipitation of July and October and underestimates the September precipitation significantly. After a
272 closer look at the data, we find that the mean precipitations of July and October are 63.8 and 23.48
273 mm over the period of 1951 through 2000, respectively, whereas the observation precipitations of

274 both months in 2001 are 39.5 and 5.2mm, respectively, much lower than the average precipitation of
275 the two month. Over the 51 years period of 1951 through 2001, the precipitations of July and October
276 in 2001 are 8th and 14th smallest, respectively. However, the precipitations of July and October in
277 2001 are the 2nd and 3rd smallest from 1991 to 2001, respectively and significantly smaller than the
278 precipitation of other months. This may be the reason that the improved and conventional model
279 underestimates for these two months. However, it is interesting that the AR models underestimates the
280 July precipitation but overestimates the October precipitation. This may be because of the much lower
281 precipitation in July, 2000 and much higher precipitation in October, 2000, relative to the July and
282 October in 2001, which, we believe, dominate the prediction of AR models. Similarly, the September
283 precipitation of 2000 is close to the precipitation of September in 2001, which results a better AR
284 prediction in that month. According to the performance of AR models, we expect an improvement if
285 we apply AR model to stationarized data series rather than the raw data series.

286 While the mean precipitation of September is 44.99 mm over the period of 1951 through 2000,
287 the observation of September in 2001 is 82mm, the 4th largest one from 1951-2001, and the largest on
288 in past 45 years. Furthermore, September, 2001 is the only one whose precipitation is larger than the
289 August's precipitation in the previous ten years. These facts clearly show that the precipitation of
290 September, 2001, is an extreme value, or outlier from statistical point of view. Therefore, it is fair to
291 conclude that the built ARIMA model needs to be further improved for extreme situations.

292 Given that both the inter-annual variation and inter-monthly variation of the hydrological data
293 effect the prediction of hydrological time series, it is better to account for both for better prediction.
294 Inter-monthly data may result from different populations as well as nonstationary factors, so the
295 conventional seasonal ARIMA model which usually neglect the inter-monthly variations is not
296 effective enough. An improved ARIMA model has been built in this paper taking account for both
297 inter-annual and inter-monthly variation of hydrological data. Based on clustering analysis and
298 regression, much more information is extracted from the data series. A case study is conducted for the
299 precipitation of Lanzhou precipitation station with the improved ARIMA model and the comparison
300 with the conventional ARIMA model indicates that the accuracy of the improved ARIMA model is

301 significantly higher than that of the conventional ARIMA model. This improved approach can be
302 applicable to other hydrological processes prediction with time series data, such as runoff, water level,
303 and water temperature.

304 Apparently, the present model could be further improved, especially for the prediction of
305 extreme phenomena. Given that the selection of clustering method does affect model performance,
306 different clustering methods, e.g., the definition of distance in the hierarchical clustering can be
307 applied (Wang et al. 2005) to obtain better fittings. Characteristics value should be constructed by the
308 features of hydrological time series, not limited to the extreme or mean values. A higher order of
309 regression model rather than the linear regression may be used for the hydrologic forecasting. Last but
310 not the least, artificial intelligence approaches, such as neural network or support vector machine, can
311 be used to further improve the proposed ARIMA model.

312 **References**

- 313 Ahmad, S., Khan, I.H., and Parida, B.P.: Performance of stochastic approaches for forecasting river
314 water quality. *Water Research*, 35(18): 4261-4266, 2001.
- 315 Box, G.E.P.: *Time Series Analysis Forecasting And Control*. China Statistics Press, Beijing, China,
316 1997.
- 317 Box, G.E.P.: Models for forecasting seasonal and nonseasonal time series, in: *Spectral Analysis of*
318 *Time Series*, Harris, B. ed., Wiley, New York, pp 271-311, 1967.
- 319 Toth, E.A. and Montanari, B.A.: Real-time flood forecasting via combined use of conceptual and
320 stochastic models, *Phys. Chem. Earth (B)*, 24(7): 793-798, 1999.
- 321 Ip, W.C., Wong, H., and Wang, S.G.: A GIC rule for assessing data transformation in regression.
322 *Statistics & Probability Letters*, 68(1): 105-110, 2004.
- 323 Jin, J.L., Ding, J., and Wei, Y.M.: Threshold autoregressive model based on genetic algorithm and its
324 application to forecasting the shallow groundwater level. *Hydraulic Engineering*, 27(6):51-55,
325 1999.
- 326 Kohn, R.: Estimation, prediction, and interpolation for ARIMA models with missing data. *J. Amer.*
327 *Statist. Assoc.*, 81: 751-761, 1986.
- 328 Lehmann, A. and Rode, M.: Long-term behaviour and cross-correlation water quality analysis of the
329 River Elbe, Germany. *Water Research*, 35(9):2153-2160, 2001.
- 330 Mcleod, A.I. and Sales, P.R.H. An algorithm for approximate likelihood calculation of ARMA and
331 seasonal ARMA models. *Applied Statistics*, 32:211-223. 1983

332 Melard, G. A fast algorithm for the exact likelihood of autoregressive-moving average models.
333 *Applied Statistics*, 33: 104-119. 1984.

334 Meloun, M., Sáňka, M., and Němec P.: The analysis of soil cores polluted with certain metals using
335 the Box–Cox transformation. *Environmental Pollution*, 137(2):273-280, 2005.

336 Pankratz, A.: *Forecasting with Univariate Box-Jenkins Models: Concepts and Cases*. Wiley, New
337 York, 1983.

338 Qi, W. and Zhen, M.P.: Winters and ARIMA model analysis of the lake level of salt Lake Zabuye,
339 Tibetan Plateau. *Journal of Lake Science*, 18(1):21-28, 2006.

340 Shumway, R. H. and Stoffer, D. S.: *Time Series Analysis and Its Applications* Springer-Verlag New
341 York, Inc., Secaucus, NJ, USA, 2005.

342 Sun, L.H., Zhao, Z.G., and Xu, L.: Study of summer rain pattern in monsoon region of east China and
343 its circulation cause. *Journal of Applied Meteorological Science*, 16(z1):57-62, 2005.

344 Thyer, M., Kuczera, G., and Wang, Q.J.: Quantifying parameter uncertainty in stochastic models
345 using the Box–Cox transformation. *Journal of Hydrology*, 265(1-4):246-257, 2002.

346 Wang, H.R., Ye, L.T., and Liu, C.M.: Problems in wavelet analysis of hydrologic series and some
347 suggestions on improvement. *Progress in Natural Science*, 17(1):80-86, 2007.

348 Wang, Y., Yin, L.Z., and Zhang, Y.: Judging model of fuzzy optimal dividing based on improved
349 objective function clustering method. *Mathematics Practice and Theory*, 35(11):142-147, 2005.

350 Niua, X.F., Edmiston, H.L., and Bailey, G.O.: Time series models for salinity and other
351 environmental factors in the Apalachicola estuarine system. *Estuarine, Coastal and Shelf Science*,
352 (46):549–563, 1998.

353

354

Table 1. Estimated parameters of the conventional seasonal ARMA model

Parameter	Estimated value	Standard deviation	<i>t</i> - test	Tail probability
θ_1	-0.16379	0.03959	-4.14	<.0001
θ_2	0.93434	0.02117	44.14	<.0001

355

356

Table 2. Autocorrelation of the residuals of the conventional seasonal ARIMA model

AR Order	χ^2 statistic	Degree of freedom	Tail probability	Autocorrelations of residue*					
6	0.770	4	0.943	0.000	-0.007	-0.018	0.021	-0.007	0.020
12	6.910	10	0.734	0.013	0.014	0.012	-0.043	0.086	-0.019
18	13.400	16	0.643	0.092	0.014	0.031	-0.004	0.021	0.020
24	16.810	22	0.774	0.042	0.007	-0.022	-0.026	-0.032	0.039
30	20.650	28	0.840	0.050	-0.031	-0.048	0.003	0.018	0.008
36	28.100	34	0.752	0.045	0.018	0.064	-0.044	0.036	0.044
42	30.900	40	0.849	0.057	-0.015	0.019	0.023	0.006	-0.001
48	52.940	46	0.224	-0.012	0.040	-0.022	0.032	-0.079	-0.156

357

*: Autocorrelations of residue for lag 1 through lag 48, 6 lags per row from Column 5 through 10.

358

359

360

Table 3. Seasonal ARIMA models for each month

Month	Model	ML parameter estimation
1	$(1 - \alpha B)y_t = (1 - \beta B)u_t$	$\alpha = -0.95, \beta = -0.97$
2	$(1 - \alpha B^2)y_t = u_t$	$\alpha = -0.49$
3	$y_t = (1 - \beta B)u_t$	$\beta = 0.38$
4	$y_t = (1 - \beta_1 B - \beta_2 B^2)u_t$	$\beta_1 = 0.27, \beta_2 = -0.22$
5	$y_t = (1 - \beta B^2)u_t$	$\beta = -0.30$
6	$y_t = (1 - \beta B)u_t$	$\beta = -0.32$
7	$y_t = (1 - \beta B^2)u_t$	$\beta = -0.3349$
8	$(1 - \alpha B)y_t = (1 - \beta B)u_t$	$\alpha = -0.182, \beta = -0.0528$
9	$(1 - \alpha B)y_t = (1 - \beta B)u_t$	$\alpha = 0.956, \beta = 0.469$
10	$y_t = (1 - \beta B)u_t$	$\beta = -0.32$
11	$(1 - \alpha B)y_t = (1 - \beta B)u_t$	$\alpha = 0.681, \beta = 0.741$
12	$(1 - \alpha B)y_t = (1 - \beta B)u_t$	$\alpha = 0.650, \beta = 0.766$

361

362

Table 4. Estimated parameters for linear regression models

Class	Month	d_i^*	a_i^*	c_i^*	b_i^*
1	1	0.16	0.09	0.39	0.23
	2	0.21	-0.12	1.21	-0.14
	11	-0.54	0.30	1.51	-0.62
	12	0.16	-0.27	0.89	0.53
2	3	1.92	-0.50	0.46	0.53
	4	-0.39	-0.57	2.33	-0.62
	10	-1.53	1.07	0.21	0.09
3	5	2.17	-0.41	0.22	0.98
	6	-0.19	-0.22	1.49	-0.35
	7	-0.22	0.27	1.05	-0.35
	8	-2.11	1.07	0.24	0.05
	9	0.35	-0.72	2.01	-0.33

*: See Eq. (10) for definition.

363

364

365

366

367

Table 5. Parameters of ARIMA models for characteristic variables of each class

Class	Characteristic variable	ARIMA model	ML parameter estimating		Standard deviation estimating		Value of P
1	maximum	$(1-B)(1-\alpha B)y_t = u_t$	-0.56		0.13		<0.0001
	mean	$(1-B)y_t = (1-\beta B)u_t$	0.92		0.07		<0.0001
	minimum	$(1-B)^2 y_t = (1-\beta B)^2 u_t$	0.84		0.09		<0.0001
2	maximum	$(1-B)y_t = (1-\beta B)^2 u_t$	-0.30		0.14		0.00311
	mean	$(1-\alpha B^2)(1-B)^2 y_t = u_t$	-0.52		0.12		<0.0001
	minimum	$(1-\alpha B^2)(1-B)^2 y_t = u_t$	-0.64		0.11		<0.001
3	maximum	$(1-\alpha B^2)(1-B)^2 y_t = u_t$	-0.45		0.13		0.0006
	mean	$(1-\alpha B)^2(1-B)^2 y_t = (1-\beta B^4)u_t$	-0.82	0.81	0.20	0.16	<0.0001
	minimum	$(1-\alpha B)^2(1-B)^2 y_t = (1-\beta B^4)u_t$	-0.81	0.80	0.12	0.17	<0.0001

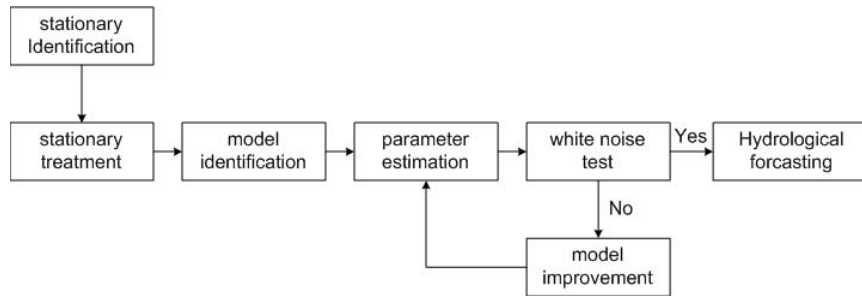
368

369

Table 6. Predicted monthly precipitation data for 2001

Month (2001)	Observation (mm)	Prediction by improved ARIMA model (mm)		Prediction by conventional ARMA model (mm)		Prediction by 12 seasonal ARIMA models (mm)		Prediction by AR(24) model (mm)		Prediction by AR(36) model (mm)	
		prediction	residual	prediction	residual	prediction	residual	prediction	residual	prediction	residual
1	2.8	2.54	-0.25	0	-2.8	1.14	-1.66	0.27	-2.53	0.57	-2.23
2	1.9	1.897	-0.003	0	-1.9	3.58	1.68	6.4	4.5	6.4	4.5
3	0	0.099	0.099	5.38	5.38	12.10	12.10	4.89	4.89	5.24	5.24
4	22.2	12.32	-9.871	11.99	-10.21	12.32	-9.88	5.81	-16.3	7.25	-14.9
5	11.1	12.61	1.515	31.26	20.16	33.17	22.07	6.49	-4.61	12.05	0.95
6	33	33.58	0.582	41.28	8.28	38.16	5.16	77.86	44.86	79.75	46.75
7	39.5	60.26	20.76	64.88	25.38	47.19	7.69	22.55	-16.9	20.09	-19.4
8	69.8	72.92	3.12	71.82	2.02	84.12	14.32	110.5	40.72	114.5	44.73
9	82	32.5	-49.5	37.98	-44.02	35.17	-46.83	65.89	-16.11	63.2	-18.8
10	5.2	32.03	26.83	20.15	14.95	24.37	19.17	55.45	50.25	58.78	53.58
11	1.9	1.532	-0.368	0	-1.9	2.68	0.78	3.9	2	3.79	1.89
12	0.9	0.898	-0.002	0	-0.9	0.94	0.04	0	-0.9	0	-0.9
Mean absolute error (mm)		9.41		11.49		11.78		17.05		17.82	

370

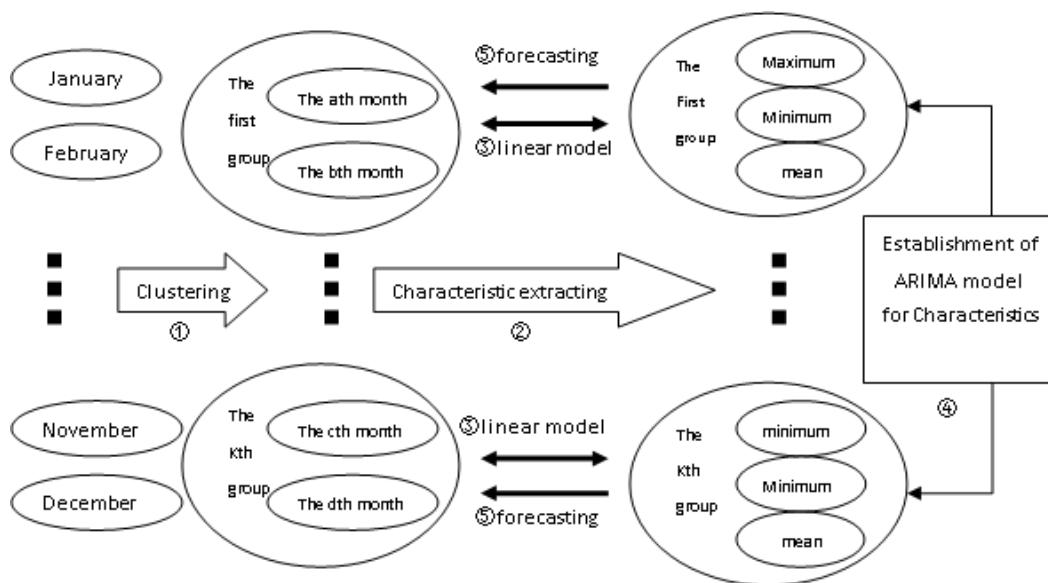


371

372

Fig. 1. Procedure of applying ARIMA model

373



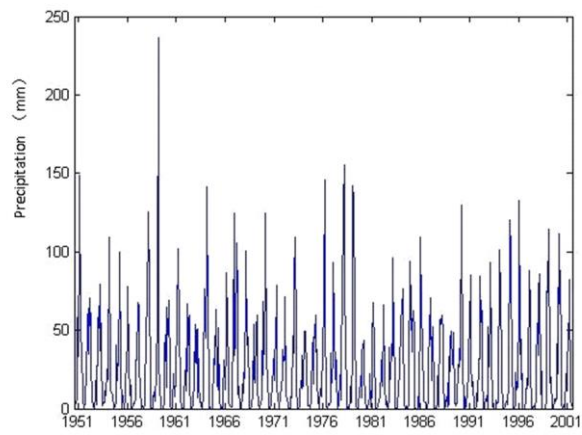
374

375

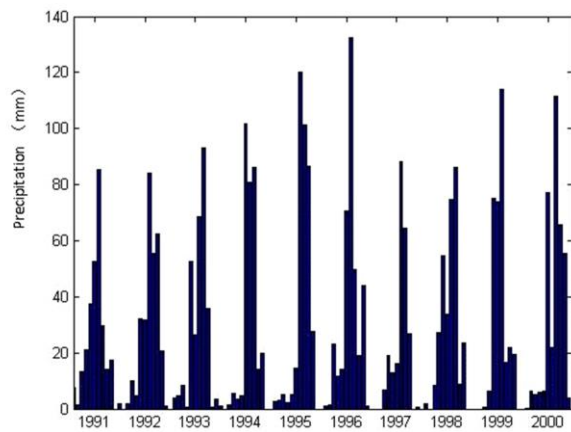
Fig. 2. Prediction steps of ARIMA model based on clustering and regressive analysis

376

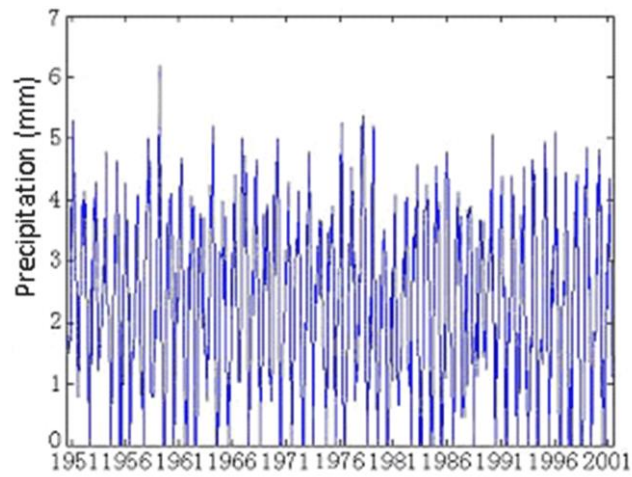
377



378



379



380

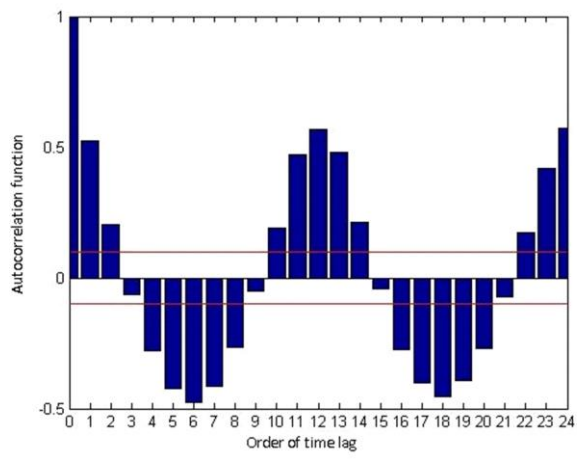
381

382

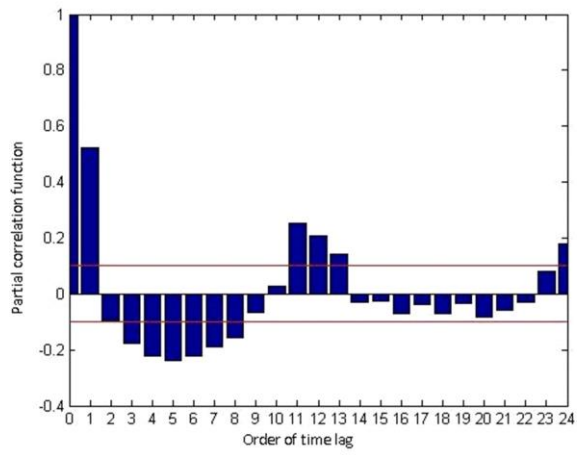
383

Fig. 3. Monthly precipitation in Lanzhou Precipitation Station.
Upper: Observation (1951-2001); Middle: Observation (1991-2000); Lower: After power transformation (1951-2001)

384



385



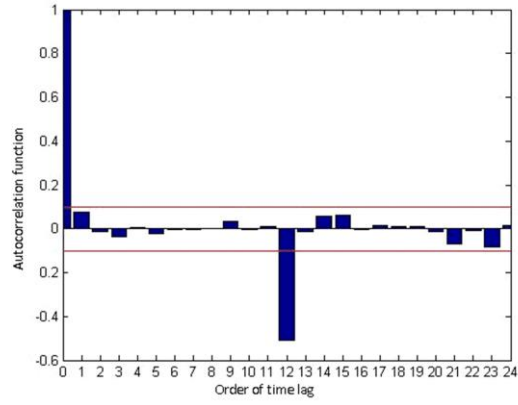
386

Fig. 4. Autocorrelation and Partial Correlation plots of data series

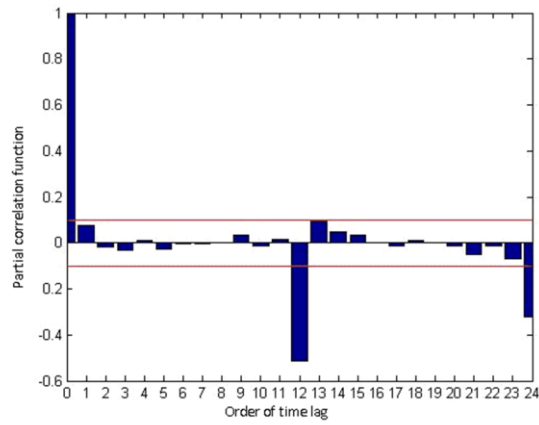
387

Upper: Autocorrelation; Lower: Partial correlation

388



389



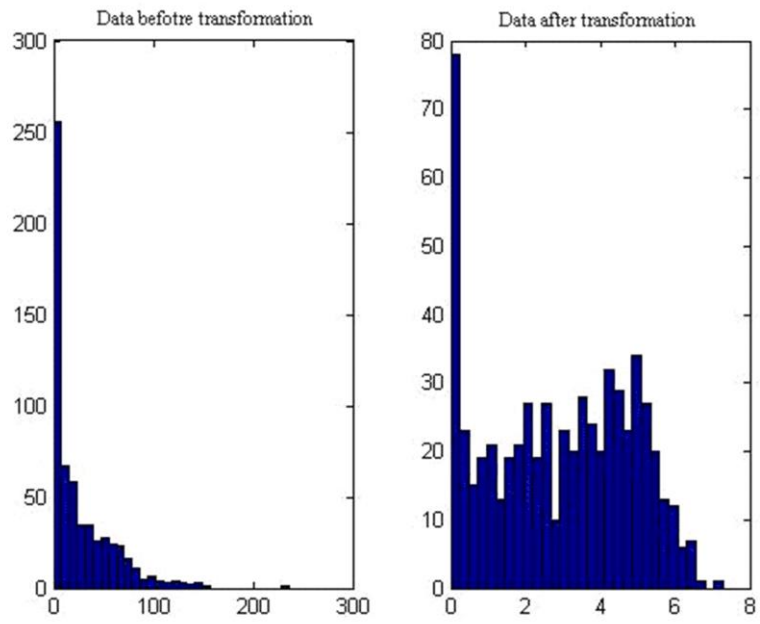
390

391

392

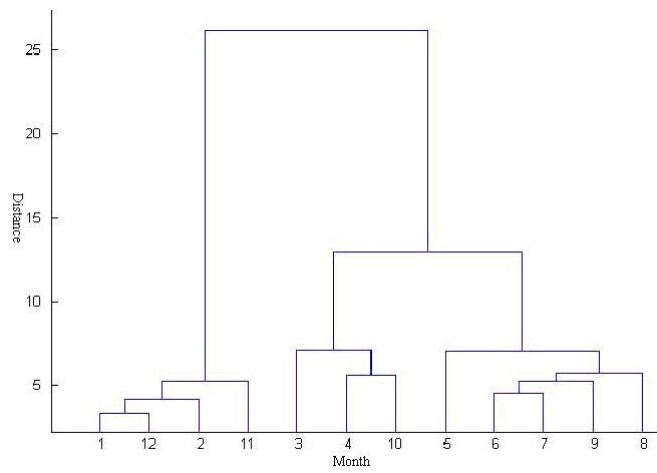
Fig. 5. Autocorrelation and Partial Correlation plots of data series after differencing
Upper: Autocorrelation; Lower: Partial correlation

393



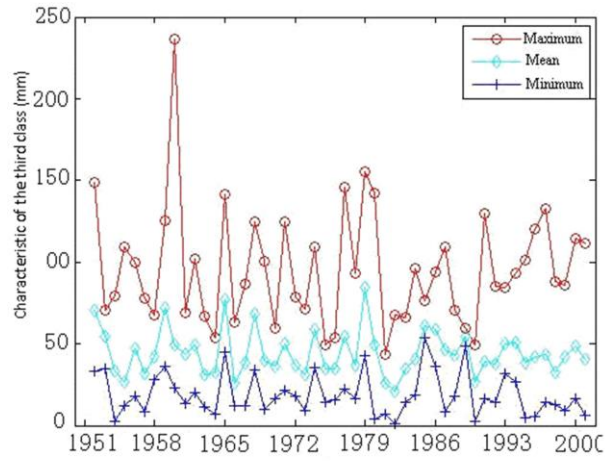
394
 395
 396

Fig. 6. Monthly precipitation series before and after Box-Cox transformation

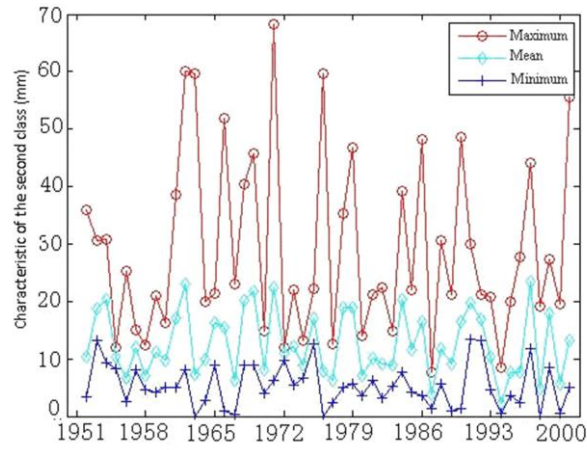


397
 398
 399

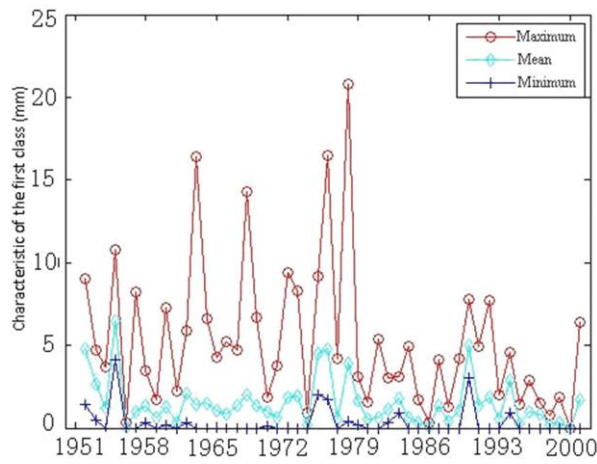
Fig. 7. Clusters of monthly precipitation time series



400



401



402

403

404

Fig. 8. Characteristics of each time series class.
Upper: first class; Middle: second class; Lower: third class

405

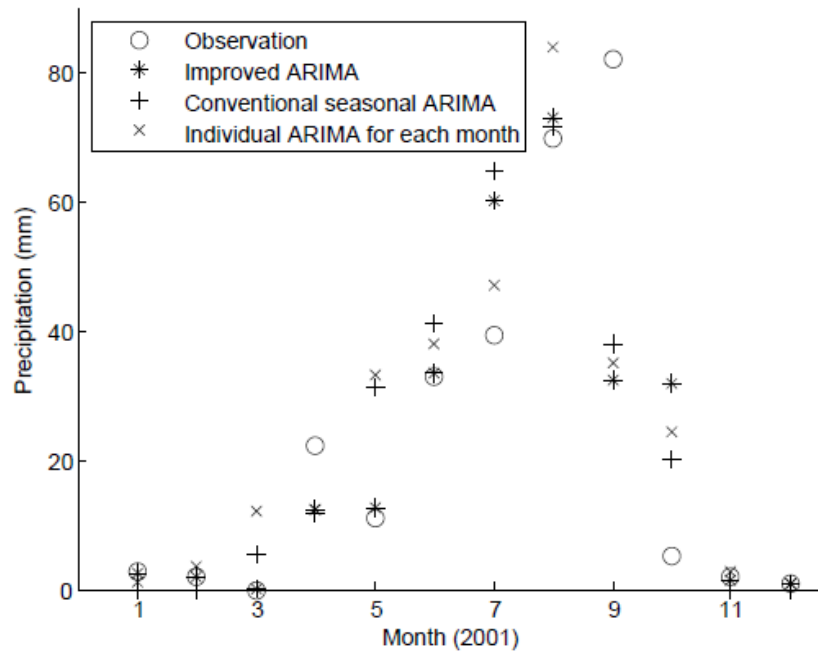


Fig. 9. Comparison between predictions and observation

406
407