6th International Conference on Hydroinformatics - Liong, Phoon & Babovic (eds)

PERIODIC AUTOREGRESSIVE MODEL APPLIED TO DAILY STREAMFLOW

WEN WANG[1, 2], PIETER H.A.J.M. VAN GELDER[1], J.K. VRIJLING[1]
1. TU Delft, Faculty of Civil Engineering & Geosciences, Hydraulic and Offshore Engineering Section. P.O.Box 5048, NL-2600 GA Delft, Netherlands
2. College of Water Resources and Environment, Hohai University Nanjing, 210098, China

Periodic models are ideal for modeling hydrological time series. But when applying periodic models to daily flow series, there is a problem of how to make the periodic model "parsimonious", because it is infeasible to fit a model for each day of the year. Therefore, an approach to periodic modeling is presented which groups the neighboring days based on cluster analysis, then fits an AR model to each group. A periodic model is fitted to the daily streamflow process of the upper Yellow River, and compared with an ARMA model that is fitted to the entire deseasonalized streamflow series.

INTRODUCTION

Streamflow processes usually have seasonal mean, variance, skewness and serial dependence structure. A usual procedure in modeling seasonal streamflow series is to deseasonalize the series by first subtracting the seasonal mean and then dividing by the seasonal standard deviation. However such a procedure can only remove the seasonality in the mean and variance, but the seasonality in serial dependence structure remains. To model appropriately such seasonality, periodic models can be employed. Two popular periodic models are the PAR (periodic autoregressive) and PARMA (periodic autoregressive and moving average) models, which are extensions of ARMA models that allow periodic (seasonal) parameters.

Literature on PAR or PARMA models has abounded since late 1960s' (e.g., [1-5]). However, PAR or PARMA usually applied to time series at the time scale of a month or at least a quarter-monthly. Even when applying to such large-time scale time series, there is also a problem of making the model parsimonious, namely, decreasing the number of model parameters required, and fitting a parsimonious PAR (PPAR) model. Salas et al. (1980) [6] proposed a Fourier series approach for reducing the number of parameters in PAR or PARMA models. The same approach is adopted for fitting so-called seasonally varying runoff coefficient (SVRC) model in which more independent exogenous inputs are included in the model (Kachroo and Liang, 1992) [7]. Thompstone et al. (1985) [8] proposed alternative approach for developing PPAR model by evaluating the compatibility of the AR equations for neighboring seasons based on residual variance analysis, then combining compatible individual AR models for adjacent seasons. However, when applying to daily flow series, with respect to the former approach, although the number of parameters could be reduced, the order of PAR model would be

the same for all seasons, which loses generality and may not be appropriate for catching seasonal dynamics of streamflow. With respect to the latter approach, separate AR models should be fitted to each day of the year, and the computation for calculating the compatibility between neighboring days are overwhelming which makes it infeasible.

The purpose of this paper is to present a method for fitting PAR model to daily flow series based on cluster analysis, make forecasting experiments to daily flow of the upper Yellow River at Tangnaihai, and compare the forecasting ability of PAR model with that of ARMA which is fitted directly to entire deseasonalized streamflow series.

DATA USED

Daily average streamflow at Tangnaihai has been recorded since January 1, 1956. The discharge gauging station Tangnaihai has a 133,650 km^2 contributing watershed in the upper Yellow River basin. Because the watershed is partly permanently snow-covered and sparsely populated, without any large-scale hydraulic works, it is fairly pristine. The plot of mean daily discharge and standard deviation of the streamflow at Tangnaihai based on 45 years' records is shown in Figure 1.

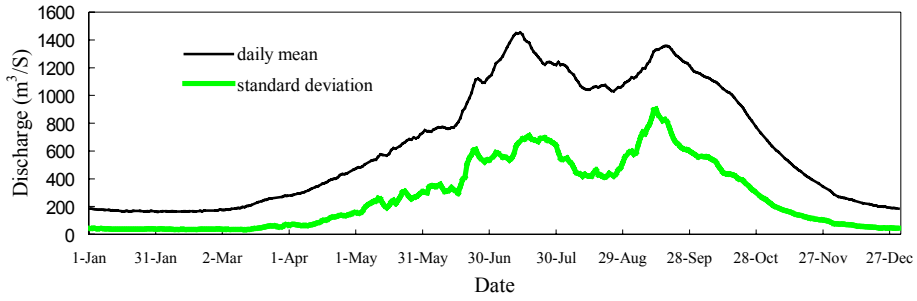

Figure 1 Variation in daily mean and standard deviation of the streamflow at Tangnaihai

Hydrologic gauging stations and weather gauging stations in the watershed are sparse. In the study, besides the streamflow series at Tangnaihai, data from some gauging stations in the upstream area are also used, including: streamflow at Mqau and Jimai, daily precipitation at Maqu and Dari, daily temperature at Dari. Because the starting date for measuring for these stations are different, to make the series have the same length when making cluster analysis, daily data from January 1, 1960 to December 31, 2000 are used for all the stations. But for fitting univariate time series models to daily streamflow at Tangnaihai, data from January 1, 1956 to December 31, 2000 are used.

The analysis procedures followed in this study is briefly illustrated in Figure 2.

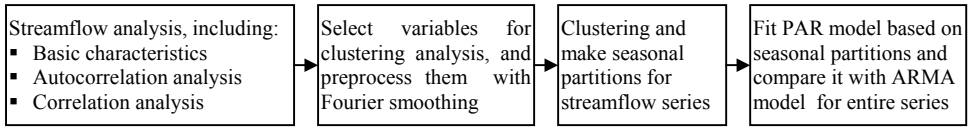

Figure 2 Data analysis procedures followed in this study

SEASONAL PARTITIONING FOR STREAMFLOW SERIES

Data preparation

Partitioning time series seasonally is performed through cluster analysis. Cluster analysis could be viewed as an instance of unsupervised learning to partition cases of a data set into a distinct number of groups. If data are represented as an $N \times m$ matrix

$$\boldsymbol{X} = (x_1, \ldots, x_i, \ldots, x_N)' \quad (1)$$

where $x_i=(x_{i1}, \ldots, x_{im})$, the goal of cluster analysis is to partition the rows of $\boldsymbol{X}$ into k distinct groups. Partitioning of daily average streamflow series is based on each day's mean discharge, standard deviation, autocorrelation at different lag time (1-10 days), mean residual after linear filtering, and correlation between each day's streamflow with previous day's temperature and precipitation. Specifically, in the clustering analysis for daily streamflow at Tangnaihai, in matrix $\boldsymbol{X}$, the number of rows N is 365, the number of columns m may change depending on the number of variables chosen, and x_i is represented generally by

$$x_i = (\mathrm{acf}_{i1}, \ldots, \mathrm{acf}_{i10}, \mathrm{mean}_i, \mathrm{sd}_i, \mathrm{res}_i, \mathrm{d\_jm}_i, \mathrm{d\_mq}_i, \mathrm{r\_mq}_i, \mathrm{r\_dr}_i, \mathrm{t\_dr}_i) \quad (2)$$

where the elements of x_i are variables calculated for day i of the year, including:

- $\mathrm{acf}_{i1}, \ldots, \mathrm{acf}_{i10}$ are the autocorrelation coefficients between the average discharge in day i and in day i +1, ..., i +10 calculated for deseasonalized daily streamflow series;
- mean_i is the mean value of day i of the year;
- sd_i is the standard deviation of day i of the year;
- res_i is the average residual for day i of a linear ARMA-type model fitted to the deseasonalized daily streamflow series;
- $\mathrm{d\_jm}_i$ is the correlation coefficient between the average discharge at Tangnaihai in day i with the streamflow at Jimai in day i-3;
- $\mathrm{d\_mq}_i$ is the correlation coefficient with the average discharge at Maqu in day i -1;
- $\mathrm{r\_mq}_i$ is the correlation coefficient with the precipitation at Maqu in day i -6;
- $\mathrm{r\_dr}_i$ is the correlation coefficient with the precipitation at Dari in day i -6;
- $\mathrm{t\_dr}_i$ is the correlation coefficient with the temperature at Dari in day i -7.

The reason for choosing these variables are as follows: 10 autocorrelation coefficients are used because serial dependence is of most importance when fitting univariate model to a time series; daily mean and standard deviation are basic statistics of flow series; residuals reflect the remaining information after linear filtering for different seasons; the correlation coefficients with upstream discharge gauging stations, weather gauging stations reflect the changes of relationship between streamflow at Tangnaihai and exogenous inputs with the change of season, and the lag time are chosen as the delay time when the largest correlation coefficients occur.

To make the scale of different variables the same, logarithmization first applied to mean streamflow of each day, and standardization is applied to all variables when making cluster analysis to deduce the influence of large-scale variables.

Cluster analysis of daily data set

There are a wide range of clustering algorithms available, which are usually divided into two broad classes called hierarchical and nonhierarchical clustering methods. A popular nonhierarchical clustering technique, *k*-means clustering (Hartigan and Wong, 1979) [9] is used here. It proceeds in the following steps:

(1) Select *k* initial clusters, where *k* is the specified number of clusters;

(2) Calculate the mean or centroids of the *k* clusters.

(3) For each case, calculate its distance to each centroid. If the case is closest to the centroid of its own cluster, leave it in that cluster, otherwise, reassign it to the cluster whose centroid is closest to it;

(4) Repeat step 2 and step 3 until no cases are reassigned.

When applying clustering techniques to time series, the temporal continuity must be kept, which differs from the cluster analysis for ordinary data sets. Because most of these factors fluctuate significantly because of short data length for each day and probable strong noise influence, and cluster analysis is quite sensitive to outliers, if no appropriate measures are taken, the clustering result for a time series could be very fragmented, which is useless for seasonal partitioning. To make the clusters of streamflow series temporally continuous, one approach could be to modify the aforementioned *k*-means clustering algorithm by constraining the reassignment according to the sequential relationship in step (3). We take another approach here by preprocessing the data series with Fourier smoothing.

The idea of this approach is simple. Smoothing each column of the data matrix X in equation (1) by taking several first Fourier harmonics of each column, then the variables of each case for clustering analysis will change smoothly by row, i.e., by day. With this preprocessing, the cases in each cluster will be temporally continuous, and the final clustering result can then easily be treated as seasonal partitions of a time series. In our study, first 10 harmonics are taken for each variable. Figure 3 and Figure 4 show some of the Fourier smoothed variables.

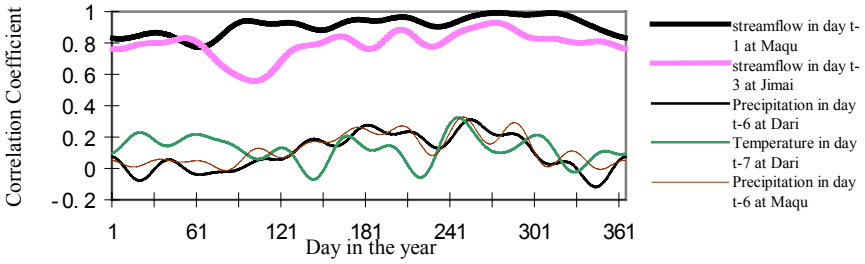

Figure 3 Fourier smoothed correlation coefficients between streamflow at Tangnaihai and exogenous variables

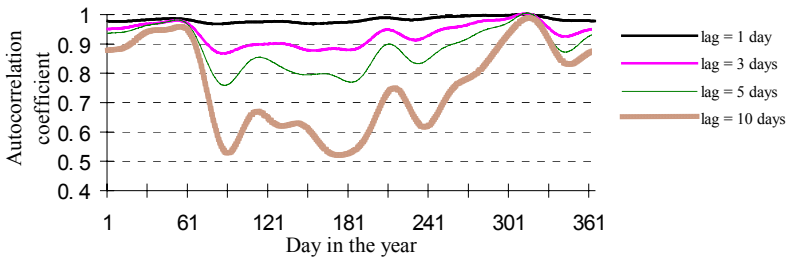

Figure 4 Fourier smoothed autocorrelation coefficients of streamflow at Tangnaihai

Clustering results and seasonal partitions

Cluster analysis is carried out for two data sets, one data set only taking the variables derived from streamflow series at Tangnaihai into account (we call it dataset A), and another data set taking the variables derived from both streamflow series at Tangnaihai and exogenous inputs into consideration (we call it dataset B).

Some indices can be calculated for determining the number of clusters (Timm, 2002)[12]. But in our study, the number of clusters is decided based on whether the results make intuitive sense, that is, the clustering should agree with the autocorrelation and correlation structure and streamflow statistics, and the total seasonal partitions of the streamflow should be not be too large. After trail and error, the number of clusters is chosen as 7 for both data sets. Clustering results for the two data sets are shown in Figure 5 and Figure 6.

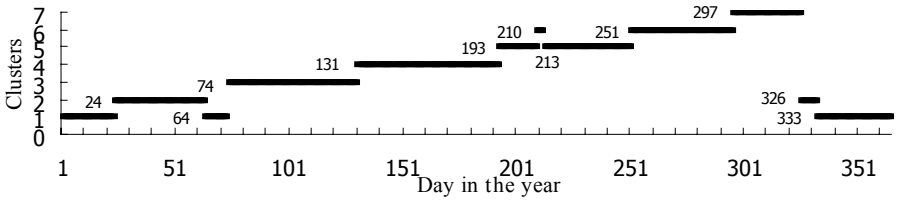

Figure 5 Clustering result based on dataset A (the numbers denote the starting day of each cluster)

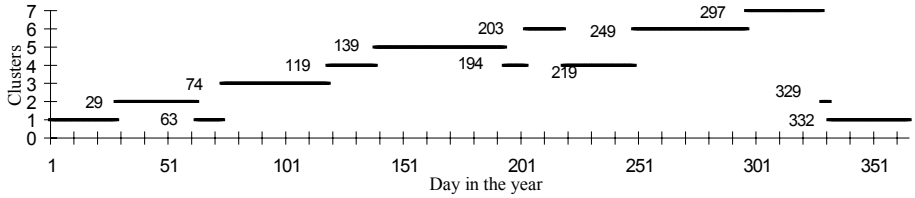

Figure 6 Clustering result based on dataset B (the numbers denote the starting day of each cluster)

Although generally the clustering results for the two data sets are close, we can find many differences. For fitting univariate periodic AR model to the streamflow, we make

seasonal partitioning in terms of the clustering result based on dataset A which comprises only variables derived from the streamflow series itself. In the clusters based on dataset A (Figure 5), one segment, from day 210 to 212, is too short, therefore we merge it with its two neighboring segments. The first segment and the last segment belong to the same cluster. Therefore, we made 9 seasonal partitions based on the cluster solution derived from dataset A, listed in Table 1. The cluster analysis result for dataset B will be used in the follow-up papers.

Table 1 Seasonal partitions of daily streamflow at Tangnaihai

Partition	1	2	3	4	5	6	7	8	9
Day span	1-23, 333-365/366	24-63	64-73	74-130	131-192	193-250	251-296	297-325	326-332

PERIODIC AUTOREGRESSIVE MODEL FITTING

Periodic autoregressive model

The periodic model is essentially a nonlinear model that treats a nonlinear process piecewise in different periods. And in each period, the model could be linear (e.g., autoregressive model or regression model with exogenous variables) or nonlinear (e.g., polynomial model or artificial neural network model). Periodic autoregressive model, which treats nonlinear process piecewise linearly is fitted here for daily streamflow at Tangnaihai based on seasonal partitions obtained in section 3.

When fitting a PAR model to a time series, we should select the order of the AR model for each season. Alkaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) are widely used criteria for determining the order of ARMA model. The chosen orders for each partition according to minimum AIC or BIC are shown in Table 2.

Table 2 The selected AR order according to minimum AIC or BIC

Partition	1	2	3	4	5	6	7	8	9
AIC	28	24	3	14	18	10	14	8	9
BIC	1	5	1	5	6	6	5	5	2

As having been noticed by many authors, BIC always tends to give smaller estimate of autoregressive orders. According to AIC, we fit AR(28), AR(24), AR(3), AR(14), AR(18), AR(10), AR(14), AR(8) and AR(9) to partition 1 to 9 respectively. Before fitting the models, the daily streamflow series is deseasonalized first. The parameter estimation is performed by ordinary least squares estimation procedure.

Comparison with ARMA model

For comparison, an ARMA(19,1) model is also fitted to the logarithmized and deseasonalized daily streamflow at Tangnaihai. The PAR model whose order of autoregressive is determined by AIC is compared with ARMA(19,1) model, and both models are applied to make 1-10 days ahead forecast for year 1996 - 2000. And the model fitting is carried out on a rolling ahead basis, that is, when forecasting daily flow in

1996, we use the model fitted to data from 1956 - 1995, and so forth. Forecasting accuracy is measured by coefficient of efficiency (*CE*), which is one of the most widely used forms of fitting criterion in hydrology community.

To partial out the influence of seasonality and compare the model performance for different seasons, forecasting accuracy of PAR and ARMA(19,1) is compared for winter (December to February), spring (March to May), summer (June to August) and autumn (September to November) separately. The plot of forecast accuracy versus lead time is shown in Figure 7.

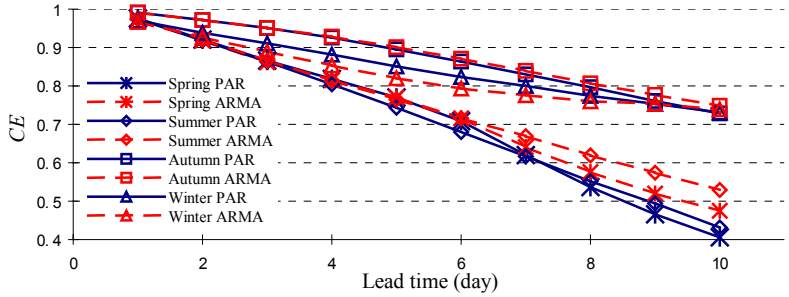

Figure 7 Comparison of the forecast accuracy of PAR model and ARMA model

It is shown that, neither models outperforms the other for all seasons. For spring, PAR performs slightly better for lead times 1-5, but gives poor forecast for longer lead-time. For summer, ARMA obviously outperform PAR for lead times longer than 4 days. For autumn, both models gives similar forecast accuracy. For winter, PAR outperforms ARMA for most lead times.

The performance of PAR is not as good as expected. Although the gain of forecast accuracy for winter is significant, at the same time, the drop of forecast accuracy for summer is also obvious. This is surprising because daily streamflow series is usually considered as nonlinear, because nonlinear type models like PAR which treat streamflow in different seasons separately should perform better than linear ARMA model. The possible reason may be that the partitioning of the streamflow is not good enough for catching the seasonal dynamics, therefore further improvement will focus on this aspect.

CONCLUSION

Periodic models are ideal for modeling seasonal hydrological time series, and many researches have been executed in the area of periodic modeling for monthly river flow. But when applying periodic models to daily flow series, there is a problem of how to make the periodic model “parsimonious”, because it is infeasible to fit a model for each day of the year. Therefore, an approach to periodic modeling is presented which groups the neighboring days based on cluster analysis, then fits an AR model to each group.

The variables considered in cluster analysis include each day’s mean discharge and standard deviation, each day’s autocorrelation with previous day's streamflow, the

relationship between each day's streamflow with upstream discharge, and temperature and precipitation in the watershed. To insure the temporal continuity of the cases in each cluster, Fourier smoothing is applied to clustering variables sorted by the order of the day in the year. Then, the Fourier smoothed variables are clustered, and the cases in each cluster will be mostly temporally continuous, therefore the clusters can then easily be treated as seasonal partitions of a time series.

After partitioning the streamflow series of the upper Yellow River at Tangnaihai, a periodic model is constructed which fits an AR model to each partition of the daily streamflow series. The PAR model is compared with an ARMA model, which is fitted to the entire deseasonalized streamflow series. Generally, PAR model gives better forecast for winter, performs similarly for other seasons for short lead time, but perform worse for long lead time, especially for summer.

Clustering technique employed here restricts that each point of the data set belongs to an exact class and omit the possibility that they belong to two or more classes simultaneously. Since observed hydrologic time series are inevitably disturbed by various kinds of noise and the climate uncertainty cannot be avoided, the starting date of a given season may vary rather than being fixed. Therefore, further extension this work could be made to apply temporal fuzzy clustering technique that may provide a solvent for better catching the dynamics of streamflow.

REFERENCES

[1] Jones, R.H., and Brelsford W.M., "Time series with periodic structure", *Biometrika*, Vol. 54, (1967), pp 403-408.

[2] Salas, J.D., Boes D.C., and Smith R.A., "Estimation of ARMA models with seasonal parameters", *Water Resources Research*, Vol. 18, (1982), pp 1006-1010.

[3] Noakes, D.J., Mcleod A.I., and Hipel K.W., "Forecasting monthly riverflow time series". *International Journal of Forecasting*, Vol. 1, (1986), pp 179-190.

[4] Salas, J.D., and Abdelmohsen M.W., "Initialization for generating single-site and multisite low-order periodic autoregressive and moving average processes", *Water Resources Research*, Vol.29, No.6, (1993), pp 1771-1776.

[5] McLeod, A.I., "Diagnostic Checking Periodic Autoregression Models With Application", *The Journal of Time Series Analysis*, Vol. 15, (1994), pp 221-233

[6] Salas, J.D., Delleur J.W., Yevjevich V. and Lane W.L., "*Applied Modelling of Hydrologic Time Series*", Colorado: Water Resources Publications, (1980)

[7] Kachroo, R.K., Liang G.C., "River flow forecasting Part 2. Algebraic development of linear modeling technique", *Journal of Hydrology*, Vol.133, (1992), pp 17-40.

[8] Thompstone, R.M., K.W. Hipel, and A.I. Mcleod. "Forecasting quarter-monthly riverflow", *Water Resources Bulletin*, Vol. 21, No. 5, (1985), pp 731-741.

[9] Hartigan, J.A., and Wong, M.A., "A K-means clustering algorithm", *Applied Statistics*, 28, (1979), pp 126-130.

[10] Timm, N. H. "*Applied Multivariate Analysis*", New York: Springer, (2002)