Application of a New Hybrid Model to Predicting Daily Runoffin a Week

Beilei Qin [1], Wensheng Wang [1], Jing Ding [1], Qiang Fu [2]

（1. Department of Hydrology and Water Resources, Hydraulic School of Sichuan University, Chengdu, Sichuan 610065, China, 1207qbl@sohu.com;
2. School of Water Conservancy & Civil Engineering, Northeast Agricultural University, Harbin, Heilongjiang 150030, China, fuqiang100@371.net）

Abstract: A new hybrid model is presented for predicting daily runoff in a week, which is combined by threshold regressive model and nearest neighbor bootstrapping disaggregation model. In this hybrid model, firstly, threshold regressive model is used to predict weekly runoff. Secondly, nearest neighbor bootstrapping disaggregation model is used to disaggregate this weekly runoff, and then daily runoff in the week is obtained. Analysis and examples indicate that this new hybrid model is practical, efficient and has prominent merit. [Nature and Science, 2004,2(1):48-52].

Key words: threshold regressive model; nearest neighbor bootstrapping disaggregation model; prediction; disaggregation

1 Introduction

Now, inflow prediction of next week must be obtained in order to constitute correct weekly dispatching plan of one hydroelectric plant. For the sake of satisfying hydroelectric plant's demand, predicting daily runoff in a week is tried in this paper. Predicting daily runoff in a week refers to that daily runoff in the next week is predicted according to the historical runoff (or rainfall) data of the river basin.

Traditional hydrology prediction method usually gives one mathematical model with certain prediction period. If traditional hydrology prediction method is used to predict daily runoff in a week, it means that seven mathematical models with different prediction period (from 1d to 7d) must be constituted, and prediction accuracy may descend remarkably with the increase of prediction period. In order to obtain relatively stable prediction accuracy, a new hybrid model is presented, which first predicts weekly runoff and then disaggregates weekly runoff into daily runoff.

In the first part of the hybrid model, threshold regressive model (TR) is used to predict weekly runoff, owing to that, not only can TR model take non-linear characteristic of weekly runoff which is a complex dynamic phenomenon into account, but also consider linear correlation of weekly runoff as its special example, and TR model can ensure good stability and prediction accuracy of the model by controlling threshold values. In the disaggregation part of the hybrid model, nearest neighbor bootstrapping disaggregation model (NNBD) is applied, because compared with traditional classical disaggregation model, NNBD model considers much more historical information.

The principle and algorithm of TR model and NNBD model, and the constitution and application of the hybrid model, are introduced respectively as following sections.

2 Threshold Regressive Model

2.1 The principle and algorithm of TR model

The principle of TR model is，that non-linear correlation between dependent variable and independent variable is described by several linear regressive models with different threshold values. Its general formula is as follows:

$$\hat{W}_t = b(j,0) + \sum_{i=1}^{P} b(j,i)Q_{t-i} + e_t(j)$$

$$As Q_{t-k} \in [r(j-1), r(j)] \quad (1)$$

where, $r(0) = -\infty$, $r(L) = +\infty$, $r(j)$ ($j = 1,2,\cdots,L-1$) is threshold value, L is the number of threshold intervals,

$b(j,i)$ is regressive coefficient of threshold interval number j, $\hat{W}_t$ is independent variable series (for example weekly runoff), Q_{t-i} is dependent variable series (for example daily runoff), Q_{t-k} is threshold variable, P is the number of dependent variables, and $e_t(j)$ is stochastic series, which is independent series. As the mean of $e_t(j)$ series is zero, $E\,[e_t(j)] = 0$, it can be neglected when predicting.

2.2 Parameter estimation of TR model

Constituting TR model substantially is that estimating its parameters, the number of threshold intervals L, threshold value $r(j)$ (j =1,2,⋯,L－1) and aggressive coefficient $b(j,i)$. Till now, there still isn't one specific method to confirm these parameters. So, choosing initial parameters, threshold variable, L and $r(j)$ (j =1,2,⋯,L－1), is usually based on correlation of independent variable and dependent variables. Then through trial method, threshold variable, L and $r(j)$ (j =1,2,⋯,L－1) are confirmed. And last, aggressive coefficient $b(j,i)$ is calculated through one optimization method.

3 Nearest Neighbor Bootstrapping Disaggregation Model

Traditional classic disaggregation is usually based on similarity principle and a typical module is chosen to disaggregate water volume. But this kind of method doesn't make use of the historical information fully. NNBD model is quoted in the hybrid model in order to improve the accuracy of disaggregation. NNBD model has clear concept and single structure, and avoids the assumption of the form of dependence and probability function. Taking the characteristic of components into account and several typical modules required to constitute an expected disaggregation module when confirming disaggregation module, are its prominent merits. Thus ensures full use of information and stable result.

3.1 The principle and algorithm of NNBD model

Let the observed daily runoff be represented by $\{Q_t\}_n$, where n is the number of daily runoff series. If the number of variables studied is m, let vector X_t＝（Q_t，Q_{t+1}，⋯，Q_{t+m-1}）, and we can suppose

$$W_t = \sum_{i=0}^{m-1} Q_{t+i}$$

(if m＝7, then W_t represents weekly runoff). Generally, there exists correlation between hydrology phenomena along time scale. So, to some extent, X_t depends on the historical daily runoff Q_{t-1}, Q_{t-2}, ⋯, Q_{t-P}. Make D_t＝(Q_{t-1}, Q_{t-2}, ⋯, Q_{t-P}) and name it as eigenvector of the daily runoff series. Then, X_t＝（Q_t，Q_{t+1}，⋯，Q_{t+m-1}）(t=P+1，P+2，⋯，n-m+1) can be defined as the succeeding value of D_t.

Among D_t（t= P+1，P+2，⋯，n） which are constituted by $\{Q_t\}_n$, there must be some eigenvectors nearest neighbor to current eigenvector D_i. Suppose the number of nearest neighbor eigenvectors is K, and represented by $D_{1(i)}$，$D_{2(i)}$，⋯，$D_{K(i)}$. $X_{1(i)}$，$X_{2(i)}$，⋯，$X_{K(i)}$ must be the succeeding values of each corresponding eigenvector. The nearest neighbor is judged by the difference between D_i and D_t, which is defined as:

$$r_{t(i)} = \Big(\sum_{j=1}^{P}(d_{ij} - d_{tj})^2\Big)^{\frac{1}{2}} \tag{2}$$

where, $r_{t(i)}$ represents the difference between D_i and D_t, d_{ij} and d_{tj} are number j variable of D$_i$ and D_t respectively, and P is the dimension of eigenvector. Then, $r_{j(i)}$ (j=1,2,…,K) is denoted as the difference between $D_{j(i)}$ and D_i, and it should be mentioned that $r_{1(i)} < r_{2(i)} < \cdots < r_{K(i)}$ (the number j is ordered according to the value of $r_{j(i)}$). The less $r_{j(i)}$ is, the nearer D_i and $D_{j(i)}$ will be and X_i is more similar to $X_{j(i)}$. Let $G_{j(i)}$ be the nearest neighbor bootstrapping weight of $X_{j(i)}$, which shows similarity between X_i and $X_{j(i)}$. Obviously, $G_{j(i)}$ is related to $r_{j(i)}$.

As discussed above, $X_{j(i)}(l) / W_{j(i)}$, the relative value of number l variable of number j nearest neighbor succeeding vector $X_{j(i)}$, is known. Number l variable of current succeeding vector X_i can be obtained through multiplying predicted weekly runoff $\hat{W}_i$ by weighted average of $X_{j(i)}(l) / W_{j(i)}$. So, the ultimate formula of NNBD model can be written as:

$$X_i(l) = \left(\sum_{j=1}^{K} G_{j(i)} \, X_{j(i)}(l) \big/ W_{j(i)}\right)\hat{W}_i,$$

$$(l=1,2,\cdots,m), \tag{3}$$

where, X_i (l) refers to number l variable of current succeeding vector X_i, $\hat{W}_i$ is the corresponding gross of X_i calculated according to equation (1), $X_{j(i)}(l)$ is number l variable of number j nearest neighbor succeeding vector $X_{j(i)}$, $W_{j(i)}$ is the corresponding gross of $X_{j(i)}$, and $G_{j(i)}$ is the nearest neighbor bootstrapping weight of $X_{j(i)}$.

3.2 Parameter estimation of NNBD model

NNBD model is confirmed, on the condition that the number of nearest neighbor K, the dimension of eigenvector P, and the nearest neighbor bootstrapping weight $G_{j(i)}$ are estimated.

Generally, $K = \mathrm{int}(\sqrt{n-P})$ is taken. On the condition of $P \geqslant 2$, the dimension of eigenvector P can be estimated by runoff auto-correlation graph or partial-correlation graph. But the ultimate value of K and P should be estimated suited to the instance through trial method.

There are a number of methods to estimate bootstrapping weight $G_{j(i)}$. When estimating, first of all, its restraint condition

$$\sum_{j=1}^{k} G_{j(i)} = 1.0$$

must be satisfied, and then bootstrapping weight $G_{j(i)}$ should be related to $r_{j(i)}$, and the bootstrapping weight function should be a simple one. As the number j is ordered according to the value of $r_{j(i)}$, in this paper, the following formula is adopted.

$$G_{j(i)} = \frac{1/j}{\sum_{l=1}^{K} 1/l}, \quad (j=1,2,\cdots,K) \tag{4}$$

When K is confirmed, we can only calculate Gj(i) once.

4 Case Study

4.1 Constitution of the hybrid model

5-year daily runoff of Jinsha River is used to demonstrate how to constitute the hybrid model and apply it to practice. Let Q_t be observed daily runoff, and weekly runoff can be denoted as

$$W_t = \sum_{i=0}^{6} Q_{t+i}$$

Figure 1 shows the hydrograph of observed daily runoff at Pingshan station, Jinsha river. Table 1 shows correlation coefficient between daily runoff and its succeeding weekly runoff.

According to correlation coefficient between weekly runoff and its nearest neighbor historical runoff showed above, it is proposed that P=3. So weekly runoff prediction TR model is constituted, considering Q_{t-1}、Q_{t-2} and Q_{t-3} as dependent variables (correlation coefficient between them and their succeeding weekly runoff is above 0.9) and their succeeding weekly runoff

$$W_t = \sum_{i=0}^{6} Q_{t+i}$$

as independent variable. And then, NNBD model is adopted, letting eigenvector $D_t = (Q_{t-1}, Q_{t-2}, \cdots, Q_{t-3})$. and succeeding value $X_t = (Q_t, Q_{t+1}, \cdots, Q_{t+6})$. According to equation (3), the predicted weekly runoff $\hat{W}_t$ is disaggregated into daily runoff in the week.

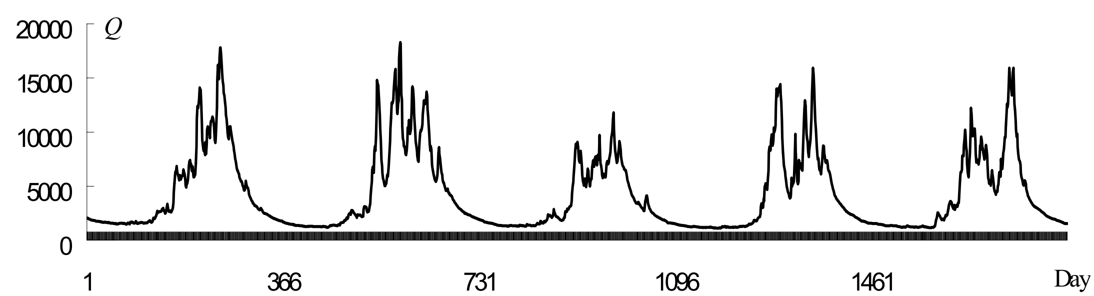

Figure 1 The hydrograph of observed daily runoff at Pingshan station, Jinsha river

Table 1 The correlation coefficient between daily runoff and its succeeding weekly runoff at Pingshan station, Jinsha river

$R(W_t, Q_{t-i})$	i=1	i=2	i=3	I=4	i=5	i=6
W_t	0.95	0.93	0.91	0.89	0.87	0.86

4.2 Weekly runoff prediction

The first four years runoff series is used for fitting of TR model, and the remaining one-year data is used for prediction or testing purpose. In this research, the number of samples during fitting phase and prediction phase should be 1443 and 365, respectively.

Since the correlation coefficient between Q_{t-1} and W_t is the most one, Q_{t-1} is chosen to be the threshold variable. The number of threshold interval L and threshold value $r(j)$ (j =1,2,⋯,L−1) can be confirmed based on correlation graph of W_t and Q_{t-1} and the optimization method chosen by TR model. Weekly runoff prediction TR model is obtained. The formula is showed below, and Tab.2 shows the accuracy of weekly runoff prediction:

$$\hat{W}_t = \begin{cases} 40.3 - 0.03Q_{t-3} - 2.27Q_{t-2} + 9.31Q_{t-1}, Q_{t-1} < 1630 \\ -1561.9 - 3.58Q_{t-3} - 3.25Q_{t-2} + 14.68Q_{t-1}, 1630 \le Q_{t-1} < 5540 \\ 18335.6 - 0.67Q_{t-3} - 4.50Q_{t-2} + 10.04Q_{t-1}, 5540 < Q_{t-1} \end{cases} \quad (5)$$

From the result of weekly runoff prediction, we can see that accuracy in fitting phase is accordant to that in prediction phase, owing to TR model's superiority.

At the same while, the rate of relative error less than 20% is over 80% both in fitting phase and prediction phase. The result is satisfied.

4.3 Daily runoff prediction

According to the main idea of NNBD model, K ＝10 and P＝3 are taken after calculation and comparison, then the first four years runoff series is used for constituting eigenvector D_t=(Q_{t-1}, Q_{t-2}, ⋯, Q_{t-3}) (t=1,2, ⋯ ,1443), just like weekly runoff prediction part, and last, the predicted weekly runoff $\hat{W}_t$ from equation (5) in prediction phase can be disaggregated into daily runoff in the week. Table 3 shows the accuracy of daily runoff predicted by hybrid model.

Table 2 The accuracy of weekly runoff prediction

Relative error	<10%	<20%	<30%	*MAE*	*MSE*
Fitting phase	64.73	86.35	95.36	563	1030
Prediction phase	69.32	85.48	93.97	517	899

Note: *MAE* is mean absolute error, and *MSE* is mean square error.

Table 3 The accuracy of daily runoff predicted by hybrid model

Relative error	<10%	<20%	<30%	*MAE*	*MSE*
1d	83.29	97.81	100	290	512
2d	75.89	92.05	98.08	389	714
3d	70.41	86.85	94.52	497	918
4d	64.93	81.92	92.05	584	1080
5d	60.82	79.45	88.77	670	1220
6d	58.08	76.16	86.03	755	1350
7d	53.42	72.88	83.56	834	1440

Note: *MAE* is mean absolute error, and *MSE* is mean square error.

From the result of daily runoff prediction in a week, it can be found that the hybrid model can give satisfied prediction accuracy for shorter prediction period. With the increase of prediction period, prediction accuracy decreases gradually. But the rate of decrease isn't prominent. Even when prediction period gets to 7d, the rate of relative error less than 20% can get to 76.16%. So, it can be concluded that prediction accuracy of this hybrid model is relatively more stable.

5 Summary

A new hybrid model that effectively blends the merits of TR model and NNBD model has been firstly presented for predicting daily runoff in a week. Analysis and examples indicate that this hybrid model can be well applied to predicting daily runoff in a week and it has following characteristics: (1) Taking the advantage of TR model, no matter the characteristic of runoff is linear or non-linear, the hybrid model can give finer runoff prediction. (2) Substituting NNBD model for traditional classical disaggregation model to choose distribution module, the hybrid model improves disaggregation accuracy effectively. (3) The hybrid model has strong adaptability. It not only can be applied to daily runoff prediction in a week, but also can be applied to daily runoff prediction in ten-day or in a month, etc.

Correspondence to:

Beilei Qin
Hydraulic School of Sichuan University
Chengdu, Sichuan 610065, China
Telephone: 01186-28-8541-0803
Cellular Phone: 01186-13036660654
E-mail: 1207qbl@sohu.com

References

[1] Srinivas VV, Srinivasan K. A hybrid stochastic model for multiseason streamflow simulation. Journal of Water Resources Research in India 2001;37(10):2537-2549.

[2] Wang WS, Xiang HL, Ding J. Prediction of hydrology and water resources with nearest neighbor bootstrapping regressive model. Journal of Hydroelectric Energy in China 2001;19(2):8-10,14.

[3] Jin JL, Yang XH, Jin BM, Ding J. Application of threshold regressive model to predicting annual runoff. Journal of Glaciology and Geocryology in China 2000;22(3):230-234.

[4] Ding J, Liu QS. Stochatis Hydrology, Chinese Water resource and Hydroelectricity Press, 1997.

[5] Wang HW. Partial least-squares regression-method and application. Chinese Natural Defense Industry Press, 1999.

[6] Jin JL, Wei YM, Ding J. Application of projection pursuit threshold regressive model for predicting annual runoff. Journal of Scientia Geographica Sinica in China 2002;22(2):171-175.

[7] Liu SY, Liao JP. The theory and method of non-linear application of runoff. Journal of Hubei Polytechnic University in China 2003;18(1):42-49.

[8] Jin JL, Wei YM, Ding J. Nearest neighbor bootstrapping model for predicting daily flow process. Journal of System Simulation in China 2002;14(11):1494-1496.