Hydrology Forecasting Model of Upper-lower Bound Estimation Based on Projection Pursuit Regression

Wei Li, Jianzhong Zhou*, Kuaile Feng, Chengwei He & Xinwei Deng
School of Hydropower and Information Engineering, Huazhong University of Science and Technology, Wuhan, Hubei, China
State Key Laboratory for Science and Technology of Digital Watershed, Huazhong University of Science and Technology, Wuhan, Hubei, China

ABSTRACT: Due to its simple structure and easy calculation, upper-lower bound estimation method has been applied in more and more researches related to probabilistic hydrology forecast. Upper and lower bounds are usually estimated by artificial neural network method, namely forecast intervals. This thesis proposes an upper-lower bound interval estimation forecasting method based on projection pursuit regression model. Coverage ratio, forecasting interval width, and symmetry shall be set as forecast criterions for upper-lower bound forecast of runoff. This thesis selected the flow data monitored in Yichang Station which is on the upstream of the Yangtze River as research object. It introduced wavelet analysis method to conduct denoising process on input data. Then, this thesis compared the results respectively forecasted by projection pursuit upper-lower bound estimation model after wavelet denoising process, projection pursuit upper-lower bound estimation model, and upper-lower estimation model of neural network model. The results showed that: projection pursuit interval forecasting model after wavelet denoising process has similar effect to BP neural network model and they both have better effect than interval forecasting model only with projection pursuit regression model.

Keywords: interval hydrology forecast; projection pursuit; wavelet analysis; upper-lower bound estimation method

1 INTRODUCTION

Probabilistic hydrological forecasting accomplishes quantitative interval forecasting in forms of flow or water level etc., thus can provide convenience for hydrological workers and provide guarantee of reservoir operation and water allocation for reservoir and drainage basin managers. In recent years, studies of drought forecasting, water resources estimation, and reservoir operation influence[2] have obtained significant progress.

Among those studies, the method based on Bayesian Theory is widely applied in hydrology[3]. The method has even been incorporated into a theoretical structure which can provide accurate theoretical guidance on probabilistic hydrological forecasting. It firstly assumes flow or water level distribution pattern. Then it applies normal quantile transform to transform likelihood function and posterior distribution, so as to calculate posterior distribution of flow or water level [4-7]. Sharma proposed a nonparametric rainfall probabilistic hydrological forecasting method which can be both seasonal and annual to reinforce management on water resources supply [8-9]. Sharma also tried to do rainfall runoff forecasting from perspective of Bayesian theory, aiming to find a method that can replace parametric estimation, model comparison, and hierarchical model development [10]. Besides, Biondi D et al. [11] and Zhang H-g et al. [12] proposed a real-time flood forecasting model based on Bayesian Theory. Kim Y-O et al. used Bayesian random planning method to do seasonal flow forecasting. The results showed this method can have higher forecasting precision [13]. Although hydrological experts have made plenty of exploration on Bayesian probabilistic forecasting, certain difficulties of hydrological forecasting timeliness have occurred due to the massive calculation and long time required in Bayesian forecasting.

*Corresponding author: jz.zhou@mail.hust.edu.cn
In addition, experts have accumulated rich experiences in ensemble hydrological forecasting [14] which has been widely used in some areas. Ensemble flow forecasting can provide data support to reservoir optimization operation. Asefa T used GLUE method to realize ensemble forecasting by combining the regular ratios of multi-group parameters of hydrological forecasting model [15]. Zhou R et al. applied GLUE (Generalized Likelihood Uncertainty Estimation) method to conduct multiple-scale uncertainty analysis and improved the sampling process [16]. Ye L et al. used multiple-target optimization algorithm to construct the forecasting intervals of ensemble hydrological forecasting model [17]. By using atmosphere model to do rainfall forecast and then applying rainfall runoff model to do runoff forecast can help achieve ensemble forecast. Furthermore, this method can consider the uncertainty of rainfall forecast in a more comprehensive and systematic way. For example, Chen J et al. combined random weather generator and ensemble weather report to do short-term flow forecast [18]. Yu P-S et al. used a random method based on seasonal weather forecast to conduct seasonal water shortage probability forecast [19].

LUBE (Lower and Upper Bound Estimation) method based on upper bound and lower bound uses neural network model to forecast upper and lower bounds of flow. Based on coverage ratio and relative width of forecast interval, it constructs a target function of penalty function to reflect forecasting precision of the model. In this way, calculation can be reduced to a large extent with good applicability. As more symmetric upper-lower bounds forecasted under condition of same coverage ratio and relative width appearing around actual measured value can bring better interval forecasting effect, Zhang introduced symmetry as an indicator. Compared with the methods mentioned above, the coverage ratios and relative widths of forecasted results were similar; however, the symmetry was higher [20].

The LUBE methods of upper and lower bounds use neural network model to forecast upper and lower bounds. Although neural network model has good forecasting effect, it is a black-box statistical model based on data driving that relies on data. In addition, internal neuron structure of neural network model cannot be presented in details. It is easy for neural structure model to enter overfitting status. PPR is a mathemathical model that does forecasting simulation through projection direction and reflects calculation process by mathematical formula. It is widely used in hydrological frequency analysis [21] and hydrological forecasting. This thesis used PPR model to forecast upper and lower bound. It used a combined indicator of coverage ratio, interval width, and symmetry as the target function to calculate. Furthermore, it applied real number encoded GA algorithm optimization projection vector and target parameter to realize interval probabilistic forecasting based on PPR.

As there’s systematic error in actual measured flow data, data is usually processed before forecasting. Besides some simple normalization processing for calculating convenience, wavelet is popular noise processing method. Sang Y-F used improved wavelet model structure in hydrological sequential simulation [22]. Wavelet analysis eliminates white noise from actual measured data by decomposition and reconstruction, so as to obtain stable and accurate hydrological forecasting. Before actual calculation, this thesis firstly used Shannon wavelet to get two-layer decomposition resulting in a time flow sequence of one low frequency and three high frequencies. After the model was calculated, this thesis reconstructed the forecasted data, and then eliminated the noise from wavelet of actual measured upstream and downstream flow data.

2 METHODS

2.1 Wavelet analysis

Wavelet analysis is a time-frequency multiresolution analytical method [23]. The key is to realize wavelet transform in signals. For finite energy signal or time sequence, its continuous wavelet transform is as follows:

$$W_f(a,b) = \int_{-\infty}^{+\infty} f(t) \psi^*\left(\frac{t - b}{a}\right) dt$$

(1)

In which $W_f(a,b)$ refers to wavelet transform coefficient; $a$ refers to contraction-expansion factor; $b$ refers to shift factor; $\psi(t)$ refers to mother wavelet; and $^*$ refers to complex conjugate.

There are many algorithms to realize wavelet transform. This thesis applied Mallat Algorithm. It set $Q$ as univariate hydrological time sequence. See (2) below for Mallat decomposition algorithm:

$$\begin{align*}
|e^{+1}(t) &= Hc_j, j = 0,1,\cdots,J \\
|d^{+1}(t) &= Gd_j
\end{align*}$$

(2)

See (3) below for Mallat reconstructing algorithm:

$$\begin{align*}
c_j &= \tilde{H}c_{j+1} + \tilde{G}d_{j+1}, \\
j &= J - 1, J - 2,\cdots, 0
\end{align*}$$

(3)

In which $c_j$ and $d_j$ respectively refer to approximate signal and detail signal under decomposition layer $j$; $H$ refers to decomposition low-pass filter; $G$ refers to decomposition high-pass filter; $\tilde{H}$ refers to reconstructing low-pass filter; and $\tilde{G}$ refers to reconstructing high-pass filter.

As Daubechies wavelet system is very sensitive to irregular signal while Shannon wavelet has the short-
est time window and better time resolution compared to other db wavelets [22], this thesis used Shannon wavelet as mother wavelet function.

2.2 Projection pursuit regression

Projection Pursuit Regression (PPR) [22-26] technology is a new multiple-factor modeling technology that combines Projection Pursuit (PP) and Regression Analysis (RA) methods. PPR does not require any human intervention for statistical data, such as assumption or transform. Instead, it uses computer to reduce dimensionality and optimize data. It can examine data structure objectively and obtain sufficient nonlinear and non-normal useful information. In addition, PPR also describes data structure by numerical function and use it for forecasting. PPR models \( y = f(x) \) and \( x = (x_1, x_2, x_p) \) are one-dimensional random variable and P-dimension variable respectively. In order to objectively reflect characteristics of high-dimensional nonlinear data structure, PPR uses the sum of a series of ridge functions \( G_m(Z) \) to approach regression function as shown below:

\[
f(x) - \sum_{m=1}^{M} \beta_m G_m(Z_m) = \sum_{m=1}^{M} \beta_m G_m(\sum_{j=1}^{M} \alpha_{mj} x_j) \quad (4)
\]

In Equation (4), \( M \) refers to number of ridge function; \( \beta_m \) refers to weight coefficient of the \( m \)-th ridge function; \( \alpha_{mj} x_j \) refers to independent variable of ridge function \( G_m \) which is the projection of \( p \)-dimensional vector on \( a \) direction; and \( \alpha_{mj} \) refers to the \( j \)-th component of \( a \). In order to confirm parameters in Equation (4), target function and optimization algorithm are usually used to confirm the detailed forecasting model of PPR.

2.3 Upper-lower bound estimation method

Upper-lower bound estimation method takes Coverage Width Symmetry-based Criterion (CWSC) as criterion. It applies forecasting method to directly forecast upper and lower bound of runoff; and to directly construct the probabilistic forecasting method of forecast interval. Both current upper and lower bound method uses neural network to forecast upper bound and lower bound. This thesis used projection pursuit method to forecast intervals. The CWSC criterion estimated by upper and lower bounds is penalty function based on coverage ratio (PICP), interval width (PINRW) and symmetry indicator (PIS). See the following equation for CWSC penalty function:

\[
\text{CWSC} = \gamma_{(\text{PIS})} e^{-\eta_1(\text{PIS}-\mu_1)} + \eta_2(\text{PIARW}) + \gamma_{(\text{PICP})} e^{-\eta_3(\text{PICP}-\mu_3)} \quad (5)
\]

In which PIARW refers to relative width and PICP refers to coverage ratio:

\[
\text{PIARW} = \frac{1}{n} \sum_{i=1}^{n} \frac{U_i - L_i}{y_i} \times 100\% \quad (6)
\]

\[
\text{PICP} = \left( \frac{1}{n} \sum_{i=1}^{n} c_i \right) \times 100\% \quad (7)
\]

In which, \( c_i \) = \begin{cases} 1 & L_i \leq y_i \leq U_i \\ 0 & \text{otherwise} \end{cases} \quad (8)

\[
\text{PIS} = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - (U_i + L_i)/2|}{U_i - L_i} \times 100\% \quad (9)
\]

In the formulae: \( U_i \) and \( L_i \) are separately the upper and lower bound forecasted at time \( t \). \( y_i \) is the measured data at \( t \). \( \mu_i, \mu_2, \eta_1, \eta_2, \eta_3 \) are the parameters of the function.

Coverage ratio is probability of actual measured value’s appearance within forecasted interval range which can reflect the simulation ability of forecasted interval in actual measured sequence. Forecasting interval width is the range of calculating upper bound and lower bound. PIARW is usually used for calculation. Smaller forecasting interval width can bring more accurate simulating degree. RIS is symmetry indicator which reflects the symmetric degree of forecasting interval in comparison to relative actual measured flow.

2.4 Model Structuring

Wavelet analysis is to decompose complex time sequence into approximate signals and detail signal sequences among which approximate signals belong to low-frequency part and detail signals belong to high-frequency part. There are differences between the functioning mechanisms and represented objects of these two. Each kind of frequency elements has its own constraints and laws of development. Set the various different frequency band sequences decomposed by moment wavelet as the input of PPAR model. Respectively output \( t \) of \( T \) sequence (\( T \) refers to forecast period). Then, start the reconstruction. The combined model established in the way mentioned above is called Projection Pursuit Auto Regression Ensemble Model (PPARWD). See Reference [27] for detailed introduction to this model. This thesis used one-dimensional projection direction. The highest order of orthogonal polynomials refers to the minimum CWSC value of second-order calculation. Thus, we can confirm the intervals of probabilistic forecasting. This thesis used genetic algorithm and set CWSC criterion as the target function to determine coefficient of Herimate orthogonal polynomials, projection direction, and parameter of LUBE. See Figure 1 for the flow chart of this model.
3 RESEARCH REGION

This thesis used upstream of Yangtze River basin as the research area. It took the actual measured flow data during flood seasons from 1953 to 2007 (Junes to September) of Pingshan Station, Gaochang Station, Lijiawan Station, Beibei Station, Wulong Station and Yichang Station on the stem stream of Yangtze River as the input, so as to forecast the flow of Yichang Station with forecast period of one day. Each station location and branch inflow of Yangtze River upstream can be found in Figure 2. 35 years (from 1953 to 1987) were selected as the regular ratio period while 20 years (from 1988 to 2007) were selected as the verification period for the model.

The total length of Yangtze River is around 6,300 km. Yangtze River is the third longest river in the world and the longest in Asia. It is the river with the most water yield in China with total water resources of 961.6 million m³, accounting for almost 36% of the total water resources in the whole country. Yangtze River basin locates between 90°33'~122°25' on eastern longitude and 24°30'~35°45' northern latitude with an area of 1.8 million km². It belongs to subtropical monsoon climate. The average rainfall can be around 1,100 mm in many years. During rainy season from April to October, its rainfall can account for 85% of the average annual rainfall. Upstream basin of Yangtze River starts from Gela Dandong of Qinghai-Tibet Plateau on the west and extends to Yichang of Hubei Province on the east with a total length of 4,504 km. The main branches include Ya-lung River, Min River, Jialing River and Wu River covering a control basin area of 1 million km². The upper reaches of Yangtze River are complex in terrain with many river rapids and speedy flow. The river fall and slope of river bed are big with abundant hydropower resources. Wu-dongde, Baihetan, Xiluodu, and Xiangjia Dam cascade reservoirs are established on Jinsha River, the upper of Yangtze River. Both Gezhouba and Three Gorges Dam, the largest dam in the world, are set in Yichang of Hubei Province. Therefore, precise hydrological forecasting contains wide practical significance in basin water resources management and reservoir operation.

4 RESULTS AND DISCUSSION

See the wavelet decomposition results from Yichang Station during the verification period from Junes to September of 1988 to 2007 in Figure 3. Shannon wavelet was used to divide the actual measured sequences into 1 low-frequency sequence and 3 high-frequency sequences. Projection pursuit regression model was applied to each sequence respectively for upper-lower bound forecasts. The last step was to combine all the data. From the figure, we can see the decomposed low-frequency wavelet had big flow but small fluctuation. The change frequency and amplitude of corresponding high-frequency wavelet made it pretty difficult to do forecasting.
During the projection pursuit regression simulation of analyzing each frequency interval by wavelet, coverage ratio of 0.9 was set as the target for low-frequency intervals. Due to its dramatic changes and high amplitude, high-frequency intervals were hard to obtain precise forecasts. During calculation, 0.3, 0.5 and 0.6 were set as the targets for the second frequency interval to do forecasting. As the flow in the third and the fourth frequency intervals were small and thus had little influence on the overall forecasting, 0.3 was set as the target for these two intervals to do forecasting. Forecasting results were given in comparison with the 0.3, 0.5 and 0.6 coverage ratios set in the second frequency interval.

Table 1. Result comparison of projection pursuit regression model and neural network model.

| Period          | Model | Index  | PICP | PIRAW | PIS  | RMSE    |
|-----------------|-------|--------|------|-------|------|---------|
| Regular rate    |       |        |      |       |      |         |
| basis period    | BP    | 0.934  | 0.339| 0.214 | 2432.82|
|                 | 0.3   | 0.899  | 0.358| 0.241 | 3041.75|
|                 | 0.5   | 0.922  | 0.388| 0.221 | 2983.46|
|                 | 0.6   | 0.934  | 0.408| 0.205 | 2945.02|
| Verification    | PPR   | 0.934  | 1.040| 0.239 | 6175.91|
| period          | BP    | 0.932  | 0.346| 0.210 | 2370.34|
|                 | 0.3   | 0.901  | 0.358| 0.234 | 2931.14|

Interval forecasting aims to include the most actual measured values with the minimum interval width and makes upper-lower boundaries arranged in symmetric distribution taking actual measured values as the benchmarks. In other words, it aims to make coverage ratio bigger while keep forecasting interval width as small as possible. At the same time, symmetry shall be close to 0 as much as possible. The contrariety of interval width and coverage ratio makes it only possible for interval forecasting to find the balance point of these two indicators. From Table 2, we can see that the overall coverage ratios after reconstruction were respectively 0.899, 0.922 and 0.934. With the increase of coverage ratio, the forecasting interval width also gradually grew bigger: 0.358, 0.388 and 0.408. In the meantime, the symmetric indicators gradually grew smaller: 0.241, 0.221 and 0.205. By comprehensively considering the forecasting effect of these three indicators, this thesis applied the second frequency interval with coverage ratio of 0.3 as the forecasting model. Through data verification during the verification period, this thesis found the coverage ratio was 0.901, the interval width was 0.358, the symmetry was 0.234, and the root-mean-square error was 2931.14. Figure 3(a) and Figure 3(b) respectively refer to the results of flow forecasted during flood seasons in 1954 and 1998 by applying wavelet analysis projection pursuit regression interval forecasting model.

Compare the interval forecasting results from BP neural network model structure and that of wavelet analysis-projection pursuit regression model, we can find during regular ratio period and verification period, wavelet analysis-projection pursuit regression forecasting model is slightly lower than neural network model in interval coverage ratio; is similar in interval width; is slightly higher in symmetry; and is slightly higher in root-mean-square error. It shows the interval forecasting effect of wavelet analysis-projection pursuit regression structure is similar to that of BP neural network model. However, wavelet analysis-projection pursuit regression model has more specific model structure. For model parameter and model structure improvement, wavelet analysis-projection pursuit regression model can avoid overfitting.

Figure 3. Wavelet breakdown diagram of Yichang Station during probative term from June to September since 1988 till 2007.

Figure 4. (a) and (b) respectively refer to the flow results during flood season in 1954 and 1998 estimated by wavelet analysis projection pursuit regression interval.
5 CONCLUSION

This thesis selected 55a flow data in total from 6 hydrological stations on the upstream of the Yangtze River as the input. It introduced wavelet analysis method to conduct denoising process on the input data and applied projection pursuit regression model to construct upper and lower interval bounds. CWSC penalty function combined with interval coverage ratio, interval width and symmetry as indicators is the target function in this thesis. The thesis also used genetic algorithm of real-part coding to do parameter optimization, so as to get the upper-lower bound interval forecasting of projection pursuit regression model. The forecasting results showed the effect of projection pursuit interval forecasting model processed by wavelet denoising process is almost the same with that of BP neural network model; and they both are better than that of interval forecasting model only with projection pursuit regression model.

ACKNOWLEDGEMENT

This work was supported by the State Key Program of National Natural Science of China for Major Research Projects (No. 91547208), and the Surface Project of National Natural Science of China (No. 51579107), the State Key Program of National Natural Science of China (No. 51239004).

REFERENCES

[1] Fan F.M., Schwanenberg D., Alvarado R., dos Reis A.A., Collischonn W, Naumann S. 2016. Performance of deterministic and probabilistic hydrological forecasts for the short-term optimization of a tropical hydropower reservoir. Water Resour. Manag., pp: 1-17.
[2] Araghinejad S. 2011. An approach for probabilistic hydrological drought forecasting. Water Resour. Manag., 25: 191-200. doi:10.1007/s11269-010-9694-9.
[3] Knill D.C., Pouget A. 2004. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci., 27: 712-719.
[4] Krzysztofowicz R. 1999. Bayesian theory of probabilistic forecasting via deterministic hydrologic model. Water Resour. Res. 35:2739-2750. doi:10.1029/1999wr900099.
[5] Krzysztofowicz R. 2014. Probabilistic flood forecast: Exact and approximate predictive distributions J Hydrol 517: 643-651 doi:10.1016/j.jhydrol.2014.04.050.
[6] Krzysztofowicz R., Kelly K.S. 2000. Hydrologic uncertainty processor for probabilistic river stage forecasting. Water Resour. Res., 36: 3265-3277 doi: 10.1029/2000wr 900108.
[7] Li, W., Zhou, J., Sun, H. et al. 2017. Water Resour. Manage. doi:10.1007/s11269-016-1557-6.
[8] Sharma A. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 1 - A strategy for system predictor identification. J. Hydrol., 239: 232-239. doi: http://dx.doi.org/10.1016/S0022-1694(00)00346-2.
[9] Sharma A. 2000. Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part 3 - A nonparametric probabilistic forecast model J. Hydrol., 239: 249-258. doi: http://dx.doi.org/10.1016/S0022-1694(00)00348-6.
[10] Sharma A., Marshall L, Nott D., Franks S.W., Sivapalan M., Takeuchi K., Tachikawa Y. A Bayesian view of rainfall-runoff modelling: alternatives for parameter estimation, model comparison and hierarchical model development. In: Predictions in Ungauged Basins, 2005. pp: 299-311.
[11] Biondi D., De Luca D.L. 2012. A Bayesian approach for real-time flood forecasting. Physics and Chemistry of the Earth, Parts A/B/C 42-44: 91-97. doi:10.1016/j.pce. 2011.04.004.
[12] Zhang, H.G., Guo, S.L., Liu, P., Peng, D.Z., Wang, C.J. 2005. Real-time flood updating model based on Bayesian method. Journal of Wuhan University of Hydraulic and Electric Engineering, 1.
[13] Kim Y.-O., Palmer R.N. 1997. Value of seasonal flow forecasts in Bayesian stochastic programming. Journal of Water Resources Planning and Management, 123: 327-335.
[14] Zhou R., Li Y., Lu D., Liu H., Zhou H. 2016. An optimization based sampling approach for multiple metrics uncertainty analysis using generalized likelihood uncertainty estimation. J. Hydrol., 540: 274-286., doi:10.1016 /j.jhydrol.2016.06.030.
[15] Asea T. 2009. Ensemble streamflow forecast: A GLUE - Based Neural Network Approach. J. Am. Water Resour. As. 45: 1155-1163.
[16] Ye L., Zhou J., Zeng X., Guo J., Zhang X. 2014. Multi-objective optimization for construction of prediction interval of hydrological models based on ensemble simulations. J. Hydrol., 519: 925-933. doi:10.1016/j.jhydrol. 2014.08.026.
[17] Chen J., Brissette F.P. 2015. Combining stochastic weather generation and ensemble weather forecasts for short-term streamflow prediction. Water Resour. Manag., 29: 3329-3342.
[18] Yu P.-S., Yang T.-C., Kuo C.-M., Wang Y.-T. 2014 A stochastic approach for seasonal water-shortage probability forecasting based on seasonal weather outlook. Water Resour. Manag., 28: 3905-3920.
[19] Zhang H., Zhou J., Ye L., Zeng X., Chen Y. 2015. Lower upper bound estimation method considering symmetry for construction of prediction intervals in flood forecasting. Water Resour. Manag., 29: 5505-5519. doi: 10.1007/s11269-015-1131-7.
[20] Duroucher M., Chebana F., Ouarda TBMJ. 2015. A non-linear approach to regional flood frequency analysis using projection pursuit regression. Journal of Hydrometeorology, 16: 1561-1574. doi:10.1175/jhm-d-14-0227.1.
[21] Sang Y.-F. 2013 Improved wavelet modeling framework for construction of prediction intervals in flood forecasting. Water Resour. Manag., 27: 2807-2821.
[22] Jones L.K. 1992. A simple lemma on greedy approximation for construction of system predictor identifications. J. of Water Resources Planning and Management: Part 1 - A strategy for system predictor identifications. J. Hydrol., 239: 232-239. doi: http://dx.doi.org/10.1016/S0022-1694(00)00346-2.
[25] Wang W.-C., Chau K.-W., Xu D.-M., Qiu L., Liu C.-C. 2016. The annual maximum flood peak discharge forecasting using Hermite projection pursuit regression with SSO and LS method. *Water Resour. Manag.*, pp: 1-17.

[26] Friedman J.H., Stuetzle W. 1981. Projection pursuit regression. *Journal of the American statistical Association*, 76: 817-823.

[27] Ji, C.M., Li, R.B., Zhang, J.K., Liu, D., Zhang, P. & Du, L. 2015. Combined projection pursuit self-regression model based on wavelet breakdown and its application in annual runoff forecasting. *Journal of Hydroelectric Engineering*, 34: 27-35.