DOI:10.13476/j.cnki.nsbdqk.2020.0089
胡义明,罗序义,梁忠民,等.基于分位点回归模型的洪水概率预报方法[J].南水北调与水利科技(中英文),2020,18(5):01-12. HU Y M,LUO X Y,LIANG Z M,et al. Flood probqbilistic forecasting based on quantile regression method[J]. South-to-North Water Transfers and Water Science & Technology,2020,18(5):01-12. (in Chinese)

基于分位点回归模型的洪水概率预报方法

胡义明,罗序义,梁忠民,黄一昕,蒋晓蕾

(河海大学 水文水资源学院,南京 210098)

摘要:采用分位点回归模型分析洪水预报的不确定性,提供洪水预报倾向值(预报概率分布的中位数)和 90%置信度的预报区间成果,实现了洪水概率预报。基于"精度-可靠性"联合评价指标对分位点回归模型计算的预报倾向值和预报区间成果进行了评估。在信江流域梅港站的应用结果表明:基于分位点回归模型提供的倾向值定值预报结果可进一步提升洪水预报的精度;同时该模型提供的 90%预报区间结果具有较高的覆盖率(约 90%)且离散度较小(小于 0.20),表明预报区间以较窄的宽度包含了绝大多数的实测值,预报可靠性较强。

关键词:洪水概率预报;分位点回归模型;预报倾向值;预报区间;精度-可靠性评价

中图分类号:TV213 **文献标志码**:A **开放科学(资源服务)标志码(OSID)**:

由于自然过程的复杂性和人类认识水平的局限性,采用水文模型进行洪水预报不可避免地会存在诸多不确定性,进而导致洪水预报成果的不确定性[1-3]。为了描述洪水预报的这种不确定性,众多洪水预报不确定性分析方法或概率预报方法被相继提出。但无论是哪一种方法,一般都是在分析预报不确定性基础上实现的,即通过将确定性水文模型与不确定性分析方法相耦合,获得未来任一时刻洪水要素的概率分布,从而实现洪水过程的概率预报。

纵观国内外现行的洪水概率预报方法,大体上可分为两类途径:一是全要素耦合途径;二是预报总误差分析途径[4]。在全要素耦合途径中,分别量化降雨-径流过程各个环节或主要要素的不确定性,如降雨输入不确定性[5-6]、模型参数不确定性[7-8]和模型结构不确定性[9-10]等,再对这些不确定性进行耦合,实现概率预报[11-14]。在预报总误差分析途径中,不直接处理输入、模型结构和参数的不确定性,代之以处理其综合误差,即从确定性预报结果入手,分析预报结果与实际洪水过程的总误差。通过采用数理统计方法构建确定性模型输出与实际洪水过程的数学描述方程,直接量化洪水预报的综合不确定性。在此基础上,推求以确定性预报值为条件的预报量的预报分布函数,实现概率预报。代表性方法包括贝叶斯预报系统的水文不确定性处理器[15-18]、模型条件处理器[19-21]、三维误差矩阵[22]和误差异分布模型[4]等。

分位点回归模型属于预报总误差分析类方法,其通过直接分析预报结果与实际洪水过程的差异进行洪水概率预报,在提供预报量的预报倾向值(中位数 Q50)的同时,也可提供某一置信度下的预报区间。本文以信江流域梅港水文站已有的实时洪水预报方案成果为基础,采用分位点回归模型,开展实时洪水概率预报方法研究。

1 方法原理

1.1 分位点回归模型

分位点回归模型是对以古典条件均值为基础的最小二乘算法的延伸,分位点回归可以估计一组回

收稿日期:2020-05-25 **修回日期**:2020-06-18 **网络出版时间**:2020-06-24
网络出版地址:https://kns.cnki.net/KCMS/detail/13.1430.TV.20200624.1046.002.html
基金项目:国家重点研发计划(2016YFC0402709;2016YFC0402707);国家自然科学基金(41730750;51709073)
作者简介:胡义明(1986—),男,江苏宿迁人,副教授,博士,主要从事水文水资源研究。E-mail:yiming.hu@hhu.edu.cn

归变量与被解释变量之间的线性关系，依据被解释变量的多种条件分位数对解释变量进行回归，可以更加精确地描述解释变量对于被解释变量的条件分布形状以及变化范围的影响，特征的分析与刻画将更加全面。利用分位点回归模型可获得任一分位点水平(如 5%和 95%等分位点)上的回归方程，进而可计算指定置信水平的洪水概率预报成果，如预报倾向值(中位数)或 90%置信度下的预报区间成果，在量化洪水预报不确定性的同时，可提供更为丰富的预报信息[23-24]。

为便于描述分位点回归模型的基本理论，现以变量 X、S 和 Y 分别表示待预报流量、确定性模型预报流量和前期实测流量系列，则分位点多元回归方程可表示为

$$X(\tau)=\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y+\varepsilon(\tau) \tag{1}$$

式中：τ 是选取的分位点($0<\tau<1$)，决定了因变量的回归水平，即在给定 S 及 Y 条件下，待预报变量对应于 τ 分位点水平的条件分位数为 $X(\tau)$，τ 越大，表明回归水平越高；$\beta_i(\tau)$，$i=0,1,2$ 是回归水平 τ 下的方程系数，其可采用加权最小一乘准则估计，即

$$\begin{aligned}&Q(\beta_0(\tau),\beta_1(\tau),\beta_2(\tau))=\\&\min\{\sum_{X_t\geqslant\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y}\tau|X(\tau|t)-\beta_0(\tau)-\\&\beta_1(\tau)S-\beta_2(\tau)Y|+\\&\sum_{X_t<\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y}(1-\tau)|X(\tau|t)-\beta_0(\tau)-\\&\beta_1(\tau)S-\beta_2(\tau)Y|\}\end{aligned} \tag{2}$$

式中：$t=1,2,3,\cdots,n$ 为观测系列样本个数。

利用上述公式可获得任一分位点水平(如 5%和 95%等分位点)上的回归方程，进而可计算指定置信水平的概率预报成果，如中位数(50%)预报值或 90%置信区间预报成果，进而提供更为丰富的预报信息。

1.2　概率预报评估指标

概率预报模型在提供类似于确定性模型的定值预报结果(分布函数的某一分位数，如期望值或中位数等)的同时，还提供某一置信度(如 90%)的区间预报结果来进行不确定性分析。针对概率预报模型提供的期望值或中位数定值预报结果，采用确定性系数和洪峰误差指标评估各模型的预报精度；针对概率预报模型提供的预报区间结果，采用覆盖率和离散度指标评估各模型的可靠度[25]。各指标计算公式如下。

(1)洪峰相对误差是用于描述预报洪峰相对于实测洪峰的偏差，其值越接近于 0 表明预报精度越高。其计算公式为

$$\Delta R=\frac{Q_f-Q_o}{Q_o}\times100\% \tag{3}$$

式中：Q_o、Q_f 分别为实测洪峰和预报洪峰，m^3/s。

(2)确定性系数用于描述预报系列和实测系列之间的吻合程度，其值越接近于 1 表明预报精度越高。其计算公式为

$$NSE=1-\frac{\sum_{i=1}^{N}(Q_{o,i}-Q_{f,i})^2}{\sum_{i=1}^{N}(Q_{o,i}-\bar{Q}_o)^2} \tag{4}$$

式中：$Q_{o,i}$、$Q_{f,i}$ 分别为第 i 时刻的实测流量、预报流量，m^3/s；$\bar{Q}_o$ 为实测流量序列均值，m^3/s；N 为系列时段总数。

(3)覆盖率是预报区间覆盖实测流量数据的比率，计算公式为

$$CR=\frac{\sum_{i=1}^{N}k_i}{N}\quad k_i=\begin{cases}1, q_i^d\leqslant o_i\leqslant q_i^u\\0, o_i<q_i^d \text{ 或 } o_i>q_i^u\end{cases} \tag{5}$$

式中：q_i^u、q_i^d 分别为第 i 时刻置信区间(如 90%)的上、下限，m^3/s；o_i 为第 i 时刻的实测流量，m^3/s；N 为系列时段总数。

(4)离散度是预报区间宽度与实测值之比，计算公式为

$$DI=\frac{1}{N}\sum_{i=1}^{N}\frac{q_i^u-q_i^d}{o_i} \tag{6}$$

某一指定置信度下，离散度越小表明置信区间越窄，洪水预报结果可能的变化幅度就越小，意味着预报结果稳定性高、不确定性越小，预报结果越实用；但置信区间越窄，其覆盖率可能就越低，表明置信区间不能覆盖大多数实际观测值或远离实测值，误差可能较大。所以，覆盖率与离散度两个指标一般情况下是矛盾的。从洪水预报角度，希望在保证有较高覆盖率的前提下，离散度尽可能小，反之，在离散度较小情况下，覆盖率尽可能高。

2　应用实例

采用信江流域的主要控制站梅港站 2012—2019 年共 10 场洪水资料对分位点回归模型进行率定及验证，其中 2012—2017 年的 8 场洪水用于分位点回归模型的率定，而 2019 年的 2 场洪水用于模型的验证。场次洪水对应的模拟预报值通过新安江模型计算获得。

在分析实测洪水过程时发现，t 时刻的流量 X_t 与($t-1$)时刻的流量 X_{t-1} 具有较强的相关性(相关系数达 0.96)，为此，选取预报时刻 t 的模型预报值 S_t 和上一时刻的实测流量 X_{t-1} 为自变量，t 时刻的流量 X_t 为因变量，构建了分位点多元回归分析模

型，推求在给定 t 时刻模型预报值 S_t 及上一时刻（$t-1$）的实测流量 X_{t-1} 条件下，t 时刻流量 X_t“实测值”的条件分布，并计算了 0.05、0.50 和 0.95 概率对应的预报值，以实现洪水概率预报。

表 1 给出了率定期 8 场洪水的分析结果，并与原始模型预报成果进行对比分析。表 1 中，Q50 表示分位点回归模型提供的预报倾向值（中位数预报）成果，而 90%预报区间的下限和上限分别为分位点回归模型提供的 5%和 95%分位点预报值。针对分位点回归模型提供的预报倾向值（Q50）定值预报结果，采用确定性系数和洪峰误差指标评估其预报精度；针对分位点回归模型提供的 90%预报区间结果，采用覆盖率和离散度指标评估其可靠度。就分位点回归模型提供的预报倾向值结果来看：率定期 8 场洪水的确定性系数值均高于原始定值预报的确定性系数值，8 场洪水确定性系数的均值为 0.99，而原始定值预报的确定性系数平均值为 0.92；基于 Q50 预报成果计算的 8 场洪水的洪峰预报相对误差均远小于原始定值预报的洪峰相对误差，8 场洪峰的相对误差绝对值的平均值为 0.44%，而原始定值预报的洪峰相对误差绝对值的平均值为 6.23%。这表明分位点回归模型在量化预报不确定性的同时也提高了定值预报精度。从 90%置信度下的预报区间来看：8 场洪水的区间覆盖率在 80%～96%，平均值达 90%；而其区间离散度指标值在 0.13～0.14，平均值达 0.14。这表明分位点回归模型提供的 90%预报区间，在具有较高覆盖率（达 90%）情况下，其离散度也较小（小于 0.20），预报区间较窄，可靠性较强。

表 1　基于分位点回归模型的预报评估结果（率定期）

洪号	确定性系数		洪峰相对误差/%		90%预报区间	90%预报区间
	原始预报	Q50	原始预报	Q50	覆盖率/%	离散度
20120610	0.90	0.99	4.90	−0.60	96	0.13
20130630	0.95	0.99	−5.20	−0.30	91	0.14
20140624	0.97	0.99	−3.40	−0.50	92	0.14
20150608	0.91	0.99	−7.30	−0.30	89	0.14
20150619	0.97	0.99	−0.80	−0.50	87	0.13
20160507	0.88	0.99	−7.90	−0.30	92	0.13
20160604	0.84	0.99	−10.60	−0.30	93	0.13
20170626	0.91	0.99	9.80	−0.70	80	0.15
平均值	0.92	0.99	6.23	0.44	90	0.14

表 2 给出了验证期 2 场洪水的分析结果，并与原始预报成果进行了对比分析。就分位点回归模型提供的预报倾向值（Q50）结果来看：验证期 2 场洪水的确定性系数的均值为 0.99，高于原始定值预报的确定性系数平均值 0.91；2 场洪水的洪峰预报相对误差绝对值的均值为 0.45%，远小于原始定值预报的洪峰相对误差绝对值的均值 7.15%。这表明分位点回归模型提供的预报倾向值（Q50）具有较高的精度。从 90%置信度下的预报区间来看，2 场洪水的区间覆盖率的平均值为 92%，而其区间离散度的平均值为 0.14。这同样表明分位点回归模型提供的 90%预报区间在具有较小离散度情况下保证了较高的区间覆盖率。

表 2　基于分位点回归模型的预报评估结果（验证期）

洪号	确定性系数		洪峰相对误差/%		90%预报区间	90%预报区间
	原始预报	Q50	原始预报	Q50	覆盖率/%	离散度
20190608	0.94	0.98	10.80	−0.60	91	0.14
20190710	0.87	0.99	3.50	−0.30	92	0.14
平均值	0.91	0.99	7.15	0.45	92	0.14

图 1、2 以率定期 20120610 号场次洪水的预报效果为例，直观地展现分位点回归模型的应用效果。图 1 给出了率定期 20120610 号场次洪水的实测系列及分位点回归模型 90%置信度下的预报区间成果（横坐标 1 表示起始时刻 2012 年 5 月 9 日 22:00）。从图中可以看出：实测系列基本都位于 90%预报区间的上下限内，尤其是该预报区间很好地覆盖了洪峰值，且具有较窄的区间宽度。

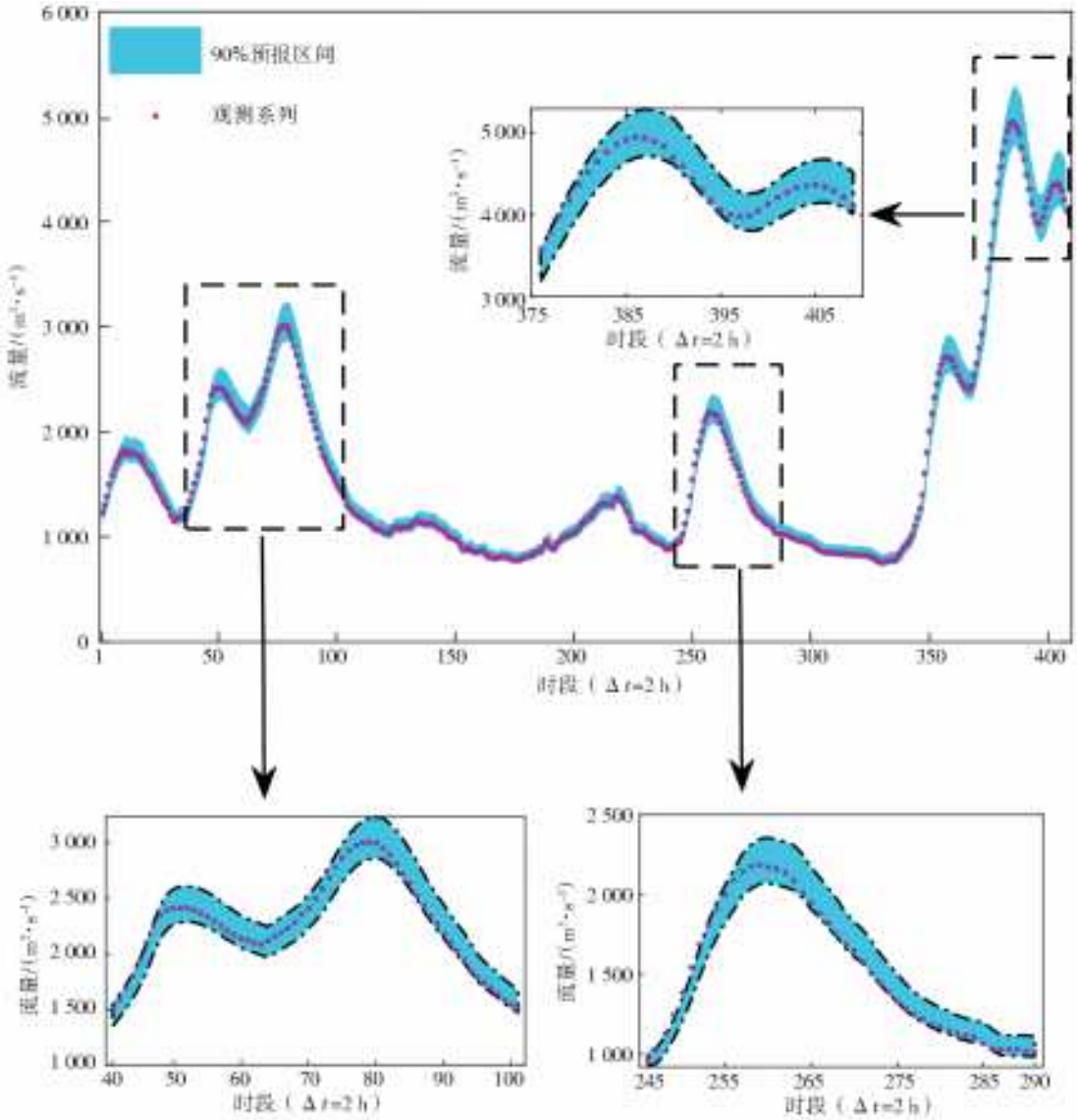

图 1　实测值系列及 90%置信度预报区间(20120610 号洪水)

图 2 给出了 20120610 号实测洪水系列、原始模型预报系列及分位点回归模型预报倾向值(Q50)系列成果。从图中可以看出:相比较于原始模型预报结果,分位点回归模型提供的预报倾向值结果与实际观测系列更接近。这表明分位点回归模型在量化预报不确定性的同时,也可对模型预报进行校正,以提供更为精确的定值预报成果。

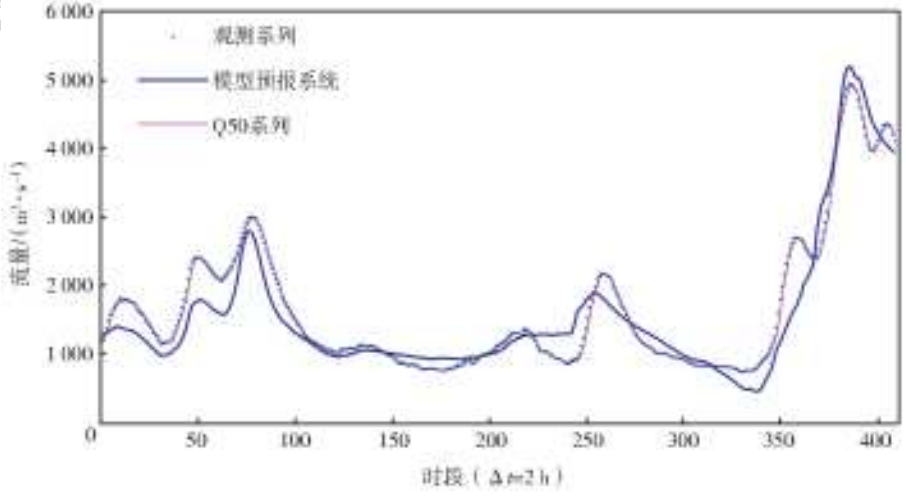

图 2　分位点回归模型 Q50 预报系列及原模型预报系列

3　结　论

基于分位点回归模型开展了梅港站实时洪水概率预报研究，在量化预报不确定性的同时，提供了洪水概率预报成果；并采用确定性系数和洪峰相对误差指标及区间离散度和覆盖率指标对分位点回归模型的精度及可靠性进行了评估。

(1)以分位点回归模型提供的预报倾向值(Q50)作为确定性预报，基于洪峰相对误差和确定性系数指标的评价结果表明，分位点回归模型可以进一步提升洪水定值预报精度。

(2)以分位点回归模型提供的90%置信度下的预报区间结果来看：实测系列基本都位于预报区间内，且区间宽度较窄。这表明基于分位点回归模型获得的预报区间，具有较高的覆盖率，且离散度也较小，预报区间的可靠性较强。

参考文献：

[1] 梁忠民，戴荣，李彬权. 基于贝叶斯理论的水文不确定性分析研究进展[J]. 水科学进展，2010，21(2)：274-281.

[2] 张洪刚，郭生练，何新林，等. 水文预报不确定性的研究进展与展望[J]. 石河子大学学报(自然科学版)，2006，24(1)：15-21.

[3] RAMOS M H，VAN ANDEL S J，PAPPENBERGER F. Do probabilistic forecasts lead to better decisions? [J]. Hydrology and Earth System Sciences，2013，17(6)：2219-2232.

[4] 梁忠民，蒋晓蕾，钱名开，等. 考虑误差异分布的洪水概率预报方法研究[J]. 水力发电学报，2017，36(4)：18-25.

[5] 梁忠民，蒋晓蕾，曹炎煦，等. 考虑降雨不确定性的洪水概率预报方法[J]. 河海大学学报(自然科学版)，2016，44(1)：8-12.

[6] KAVETSKI D，KUCZERA G，FRANKS S W. Bayesian analysis of input uncertainty in hydrological modeling：1. Theory[J]. Water Resources Research，2006，42(3)：W03407. DOI：10. 1029/2005WR004368.

[7] BEVEN K，BINLEY A. The future of distributed models：model calibration and uncertainty prediction[J]. Hydrological Processes，1992，3(6)：279-298.

[8] 赵盼盼，吕海深，朱永华，等. 基于GLUE和标准Bayesian方法对TOPMODEL模型的参数不确定性分析[J]. 南水北调与水利科技，2014，12(6)：44-48.

[9] CLARK M P，SLATER A G，RUPP D E，et al. Framework for Understanding Structural Errors (FUSE)：A modular framework to diagnose differences between hydrological models[J]. Water Resources Research，2008，44(12). DOI：10. 102912007WR006735.

[10] 李致家，黄鹏年，姚成，等. 灵活架构水文模型在不同产流区的应用[J]. 水科学进展，2014，25(1)：28-35.

[11] KRZYSZTOFOWICZ R. Bayesian theory of probabilistic forecasting via deterministic hydrologic model[J]. Water Resources Research，1999，35(9)：2739-2750.

[12] KUCZERA G，KAVETSKI D，FRANKS S，et al. Towards a Bayesian total error analysis of conceptual rainfall-runoff models：Characterising model error using stormdependent parameters[J]. Journal of Hydrology，2006，331(1-2)：161-177.

[13] AJAMI N K，DUAN Q，SOROOSHIAN S. An integrated hydrologic Bayesian multimodel combination framework：Confronting input，parameter，and model structural uncertainty in hydrologic prediction[J]. Water Resources Research，2007，43(1)：W1403.

[14] 李明亮，杨大文，陈劲松. 基于采样贝叶斯方法的洪水概率预报研究[J]. 水力发电学报，2011，30(3)：27-33.

[15] KRZYSZTOFOWICZ R，KELLY K S. Hydrologic uncertainty processor for probabilistic river stage forecasting[J]. Water Resources Research，2001，36(11)：3265-3277.

[16] 刘章君，郭生练，李天元，等. 贝叶斯概率洪水预报模型及其比较应用研究[J]. 水利学报，2014，45(9)：1019-1028.

[17] 邢贞相，芮孝芳，崔海燕，等. 基于AM-MCMC算法的贝叶斯概率洪水预报模型[J]. 水利学报，2007，38(12)：1500-1506.

[18] 蒋晓蕾，梁忠民，王春青，等. BFS-HUP模型在潼关站洪水概率预报中的应用[J]. 人民黄河，2015，37(7)：13-15.

[19] TODINI E. A model conditional processor to assess predictive uncertainty in flood forecasting[J]. International Journal of River Basin Management，2008. 6(2)：123-137.

[20] MONTANARI A，GROSSI G. Estimating the uncertainty of hydrological forecasts：A statistical approach[J]. Water Resources Research，2008，44(12)：W00B08. DOI：10. 1029/2008WR006897

[21] 王艳兰，梁忠民，王凯，等. 基于多模型MCP方法的洪水概率预报[J]. 南水北调与水利科技，2018，16(6)：39-45.

[22] VAN STEENBERGEN N，RONSYN J，WILLEMS P. A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication[J]. Environmental Modelling & Software，2012，33：92-105.

[23] WEERTS A H，WINSEMIUS H C，VERKADE J S. Estimation of predictive hydrological uncertainty using quantile regression：Examples from the National Flood Forecasting System (England and Wales)[J]. Hydrology and Earth System Sciences，2011，15(1)：255-265.

[24] 乔舰，李再兴. 分位数回归的理论再说明及实例分析[J]. 统计与决策，2012(19)：104-107.

[25] 蒋晓蕾，梁忠民，胡义明，等. 洪水概率预报评价指标研究[J]. 湖泊科学，2020，32(2)：539-552.

·译文(Translation)·

DOI:10.13476/j.cnki.nsbdqk.2020.0089

Flood probabilistic forecasting based on quantile regression method

HU Yiming, LUO Xuyi, LIANG Zhongmin, HUANG Yixin, JIANG Xiaolei

(College of Hydrology and Water Resources, Hohai Uuniversity, Nanjing 210098, China)

Abstract: The quantile regression model is used to analyze the uncertainty of flood forecasting. The results of preferred value (median of predicted probability distribution function) and 90% confidence interval of flood forecasting are provided to realize the forecast of flood probability. The performance of probabilistic forecasting obtained by the quantile regression model is evaluated using "accuracy-reliability" joint evaluation index. The application results of Meigang station in the Xinjiang River basin show that in term of prediction preferred value, the quantile regression model can further improve the accuracy of the flood forecasting. Simultaneously, the prediction interval results with 90% confidence level provided by the model have higher coverage (about 90%) and less dispersion (less than 0.20), which means that the narrow prediction interval contains most of the observation, and the reliability of the forecast interval is strong.

Key words: flood probabilistic forecasting; quantile regression model; prediction preferred value; prediction interval; accuracy-reliability assessment

Due to the complexity of natural process and the limitation of human understanding, the use of hydrological models for flood forecasting will inevitably have many uncertainties, which will lead to the uncertain results[1-3]. In order to describe the uncertainty of flood forecasting, many uncertainty analysis methods of flood forecasting or probabilistic forecasting methods have been proposed successively. However, no matter which method is used, it is generally realized on the basis of analyzing the uncertainty of forecasting. In other words, by coupling the deterministic hydrological model with the uncertainty analysis method, the probability distribution of flood elements at any time in the future is obtained to realize the probabilistic forecasting of flood process.

The current flood probabilistic forecasting methods can be generally divided into two types. One is the total-factor coupling approach, and the other is the total forecasting error analysis approach[4]. In the total-factor coupling approach, the uncertainties of each link or major factor in the rainfall-runoff process, such as rainfall input uncertainty[5-6], model parameter uncertainty[7-8], and model structure uncertainty[9-10], are quantified respectively. Then, these uncertainties are coupled to realize probabilistic forecasting[11-14]. In the total forecasting error analysis approach, the uncertain-

Received: 2020-05-25 **Revised**: 2020-06-18 **Online publishing**: 2020-06-24
Online publishing address: https://kns.cnki.net/KCMS/detail/13.1430.TV.20200624.1046.002.html
Fund: National Key Research and Development Program of China (2016YFC0402709; 2016YFC0402707); National Natural Science Foundation of China (41730750; 51709073)
Author brief: HU Yiming (1986-), male, PhD, associate professor, Suqian Jiangsu Province, mainly engaged in the research on hydrology and water resources. E-mail: yiming.hu@hhu.edu.cn

ties of input, model structure, and parameters are not directly dealt with. Instead, the comprehensive error is dealt with. Specifically, the total error between the forecasting results and the actual flood process is analyzed from the deterministic forecasting results. The mathematical statistics method is used to construct the mathematical description equation of the output of the deterministic model and the actual flood process to directly quantify the comprehensive uncertainty of flood forecasting. On this basis, the forecast distribution function of the forecast quantity in the condition of the deterministic forecast value is deduced to realize the probabilistic forecasting. The representative methods include the hydrologic uncertainty processor of Bayesian forecasting system[15-18], the model conditional processor[19-21], the three-dimensional error matrix[22], and the error heterogeneous distribution model[4].

Quantile regression model belongs to the total forecasting error analysis method. Through direct analysis of the difference between the forecasting results and the actual flood process, the flood probabilistic forecasting is carried out. It provides not only the prediction preferred value (median Q50) of the forecast quantity but also the prediction interval at a certain confidence level. In this paper, based on the existing results of real-time flood forecasting scheme of Meigang hydrological station in the Xinjiang River basin, the quantile regression model is used to study the real-time flood probabilistic forecasting method.

1 Method principle

1.1 Quantile regression model

The quantile regression model is an extension of the least squares algorithm based on the classical conditional mean. Quantile regression can estimate the linear relationship between a set of regression variables and explained variables. Regression of explanatory variables according to various conditional quantiles of explained variables can more accurately describe the influence of explanatory variables on the conditional distribution shape and variation range of explained variables, and the analysis and characterization of features will be more comprehensive. The regression equation at any quantile level (such as 5% and 95% quantiles) can be obtained using the quantile regression model, and then the flood probabilistic forecasting results at a specified confidence level can be calculated, such as the prediction preferred value (median) or the prediction interval results at 90% confidence level. Furthermore, the uncertainty of flood forecasting is quantified, and more abundant forecasting information can be provided at the same time[23-24].

In order to describe the basic theory of the quantile regression model, variables X, S, and Y are used to represent the flow to be predicted, the forecast flow of the deterministic model, and the early-stage observed flow series respectively. The quantile multiple regression equation can be expressed as follows

$$X(\tau)=\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y+\varepsilon(\tau) \tag{1}$$

where τ is the selected quantile ($0<\tau<1$), which determines the regression level of dependent variable. In other words, under the given conditions of S and Y, the conditional quantile of the variable to be predicted corresponding to the τ quantile level is $X(\tau)$. A higher τ leads to a higher regression level. $\beta_i(\tau)$, ($i=0,1,2$) is the equation coefficient at the regression level τ, which can be estimated by the weighted least-absolute criterion

$$\begin{aligned}&Q(\beta_0(\tau),\beta_1(\tau),\beta_2(\tau))=\\&\min\{\sum_{X_t\geqslant\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y}\tau\,|X(\tau|t)-\beta_0(\tau)-\\&\beta_1(\tau)S-\beta_2(\tau)Y|+\\&\sum_{X_t<\beta_0(\tau)+\beta_1(\tau)S+\beta_2(\tau)Y}(1-\tau)\,|X(\tau|t)-\beta_0(\tau)-\\&\beta_1(\tau)S-\beta_2(\tau)Y|\}\end{aligned} \tag{2}$$

where $t=1,2,3,\cdots,n$, which is the number of samples of the observed series.

The regression equation at any quantile level (such as 5% and 95% quantiles) can be obtained using the above equation, and then the probabilistic forecasting results at a specified confidence level can be calculated, such as the median (50%) forecast value or the prediction interval results at 90% confidence level, thus providing more abundant forecasting information.

1.2 Evaluation index of probabilistic forecasting

The probabilistic forecasting model offers not only deterministic forecasting results similar to those of the deterministic model (a quantile of the distribution function, such as the expected value or median) but also the prediction interval results with a certain confidence level (such as 90%) for uncertainty analysis. According to the deterministic forecasting results of the expected value or median provided by the probabilistic forecasting model, the certainty coefficient and the flood peak error are used to evaluate the forecast accuracy of each model. According to the prediction interval results provided by the probabilistic forecasting model, the coverage rate and dispersion are used to evaluate the reliability of each model[25]. The calculation equation of each index is as follows.

(1) The relative error of flood peak is used to describe the deviation of the forecasted flood peak from the observed flood peak. The value closer to 0 indicates the higher forecast accuracy. The calculation equation is as follows

$$\Delta R=\frac{Q_f-Q_o}{Q_o}\times 100\% \tag{3}$$

where Q_o and Q_f are respectively the observed and forecasted flood peaks (m^3/s).

(2) The certainty coefficient is used to describe the degree of coincidence between the forecast series and the observed series. The value closer to 1 indicates the higher forecast accuracy. The calculation equation is as follows

$$NSE=1-\frac{\sum_{i=1}^{N}(Q_{o,i}-Q_{f,i})^2}{\sum_{i=1}^{N}(Q_{o,i}-\bar{Q}_o)^2} \tag{4}$$

where $Q_{o,i}$ and $Q_{f,i}$ are respectively the observed flow and the forecast flow at time i (m^3/s); $\bar{Q}_o$ is the mean value of the observed flow series (m^3/s); N is the total number of periods in the series.

(3) The coverage rate is the percentage of the observed flow data covered by the prediction interval. The calculation equation is as follows

$$CR=\frac{\sum_{i=1}^{N}k_i}{N}\quad k_i=\begin{cases}1, q_i^d\leqslant o_i\leqslant q_i^u\\ 0, o_i<q_i^d \text{ or } o_i>q_i^u\end{cases} \tag{5}$$

where q_i^u and q_i^d are respectively the upper and lower limits of the confidence interval (such as 90%) at time i (m^3/s); o_i is the observed flow at time i (m^3/s); N is the total number of periods in the series.

(4) The dispersion is the ratio of the width of the prediction interval to the measured value. The calculation equation is as follows

$$DI=\frac{1}{N}\sum_{i=1}^{N}\frac{q_i^u-q_i^d}{o_i} \tag{6}$$

At a certain specified confidence level, when the dispersion is lower, the confidence interval is narrower and the possible variation range of the flood forecasting results is narrower. These signal higher stability of the forecasting results, smaller uncertainty, and more practical forecasting results. However, a narrower confidence interval may lead to a lower coverage rate, indicating that the confidence interval cannot cover most of the actual observed values or is far from the measured values, and the error may be large. Therefore, the coverage rate and dispersion generally show reverse trends. From the perspective of flood forecasting, it is hoped that the dispersion is as low as possible on the premise of ensuring a high coverage rate. Conversely, it is hoped that the coverage rate will be as high as possible in the case of high dispersion.

2 Application examples

The data of ten floods from 2012 to 2019 at Meigang Station, the main control station of the Xinjiang River basin, is used to calibrate and verify the quantile regression model. The eight floods from 2012 to 2017 are used for the calibration of the quantile regression model, and the two floods in 2019 are used for model verification. The simulated forecast value corresponding to the floods is calculated through the Xin'an River model.

In analysis of the observed flood process, it is found that the flow X_t at time t has a strong correlation with the flow X_{t-1} at time $(t-1)$ (the correlation coefficient is 0.96). Therefore, a quantile multiple regression analysis model is constructed

by selecting the model forecast value S_t at the forecast time t and the observed flow X_{t-1} at the previous time as the independent variables and the flow X_t at the time t as the dependent variable. The conditional distribution of the "measured value" of the flow X_t at time t under the given conditions of the model forecast value S_t at time t and the observed flow X_{t-1} at the previous time $(t-1)$ is deduced, and the forecast values corresponding to the probabilities of 0.05, 0.50, and 0.95 are calculated, so as to realize the flood probabilistic forecasting.

Tab. 1 shows the analysis results of eight floods in the calibration period and makes a comparative analysis with the forecasting results of the original model. In the table, Q50 represents the prediction preferred value (median forecast) results provided by the quantile regression model, and the lower and upper limits of the prediction interval with 90% confidence level are the 5% and 95% quantile forecast values provided by the quantile regression model, respectively. According to the deterministic forecasting results of the prediction preferred value (Q50) provided by the quantile regression model, the certainty coefficient and the flood peak error are used to evaluate the forecast accuracy. According to the prediction interval results with 90% confidence level provided by the quantile regression model, the coverage rate and dispersion are used to evaluate the reliability. From the results of the prediction preferred value provided by the quantile regression model, the certainty coefficient values of the eight floods in the calibration period are all higher than those of the original deterministic forecast. The mean value of the certainty coefficients of the eight floods is 0.99, while that of the original deterministic forecast is 0.92. The relative error of flood peak forecast of the eight floods calculated based on Q50 forecasting results is far less than that of original deterministic forecast. The average absolute value of relative errors of the eight flood peaks is 0.44%, while the average absolute value of relative errors of flood peak of the original deterministic forecast is 6.23%. The results show that the quantile regression model not only quantifies the uncertainty of forecasting, but also improves the accuracy of deterministic forecast. From the prediction interval at 90% confidence level, the interval coverage rate of the eight floods ranges from 80% to 96%, with an average of 90%. The interval dispersion is between 0.13 and 0.14, and the mean value is 0.14. The results show that the prediction interval with 90% confidence level provided by the quantile regression model has a low dispersion (less than 0.20) under the condition of high coverage rate (up to 90%), indicating that the prediction interval is narrow and the reliability is strong.

Tab. 1 Forecast evaluation results based on quantile regression model (calibration period)

Flood No.	Certainty coefficient		Relative error of flood peak/%		Coverage rate of prediction interval with 90% confidence level/%	Dispersion of prediction interval with 90% confidence level
	Original forecast	Q50	Original forecast	Q50		
20120610	0.90	0.99	4.90	−0.60	96	0.13
20130630	0.95	0.99	−5.20	−0.30	91	0.14
20140624	0.97	0.99	−3.40	−0.50	92	0.14
20150608	0.91	0.99	−7.30	−0.30	89	0.14
20150619	0.97	0.99	−0.80	−0.50	87	0.13
20160507	0.88	0.99	−7.90	−0.30	92	0.13
20160604	0.84	0.99	−10.60	−0.30	93	0.13
20170626	0.91	0.99	9.80	−0.70	80	0.15
Mean value	0.92	0.99	6.23	0.44	90	0.14

Tab. 2 shows the analysis results of the two floods in the verification period, and makes a comparative analysis with the original forecasting results. According to the prediction preferred value (Q50) results provided by the quantile regression model, the mean value of the certainty

coefficient of the two floods in the verification period is 0.99, which is higher than that of the original deterministic forecast (0.91). The average absolute value of relative errors of flood peak forecast of the two floods is 0.45%, which is far less than the average absolute value of relative errors of flood peak of the original deterministic forecast (7.15%). The results show that the prediction preferred value (Q50) provided by the quantile regression model has high accuracy. From the prediction interval with 90% confidence, the average interval coverage rate of the two floods is 92%, and the average interval dispersion is 0.14. The results also show that the prediction interval with 90% confidence level provided by the quantile regression model ensures a high coverage rate under the condition of small dispersion.

Tab. 2 Forecast evaluation results based on quantile regression model (validation period)

Flood No.	Certainty coefficient		Relative error of flood peak/%		Coverage rate of prediction interval with 90% confidence level	Dispersion of prediction interval with 90% confidence level
	Original forecast	Q50	Original forecast	Q50		
20190608	0.94	0.98	10.80	−0.60	91	0.14
20190710	0.87	0.99	3.50	−0.30	92	0.14
Mean value	0.91	0.99	7.15	0.45	92	0.14

Fig. 1 and Fig. 2 take the forecast effect of No. 20120610 flood in the calibration period as examples to visually show the application effect of the quantile regression model. Fig. 1 shows the observed series of No. 20120610 flood in the calibration period and the prediction interval results with 90% confidence level provided by the quantile regression model (the abscissa 1 represents the starting time of 22:00 on May 9, 2012). It can be seen from the figure that the observed series are basically within the upper and lower limits of the prediction interval with 90% confidence level. In particular, the prediction interval covers the flood peak well and has a narrow interval width.

Fig. 2 shows the results of the No. 20120610 observed flood series, the original model forecast series, and the prediction preferred value (Q50) series provided by the quantile regression model. It can be seen from the figure that compared with the forecasting results of the original model, the prediction preferred value results provided by the quantile regression model are closer to the actual observed series. This shows that the quantile regression model cannot only quantify the uncertainty of forecasting, but also correct the model forecast, so as to provide more accurate deterministic forecasting results.

3 Conclusions

Based on the quantile regression model, the real-time flood probabilistic forecasting of Meigang Station is studied, which not only quantifies the uncertainty of forecasting, but also provides the flood probabilistic forecasting results. In addition, the indexes of certainty coefficient and relative error of flood peak and the indexes of interval dispersion and coverage rate are used to evaluate the accuracy and reliability of the quantile regression model.

(1) The prediction preferred value (Q50) provided by the quantile regression model is used as the deterministic forecast. The evaluation results based on the relative error of flood peak and the certainty coefficient show that the quantile regression model can further improve the accuracy of flood deterministic forecasting.

(2) According to the prediction interval results at 90% confidence level provided by the quantile regression model, the observed series are basically within the prediction interval, and the interval width is relatively narrow. The results show that the prediction interval obtained based on the quantile regression model has a high coverage rate and a low dispersion, and the prediction interval has a strong reliability.

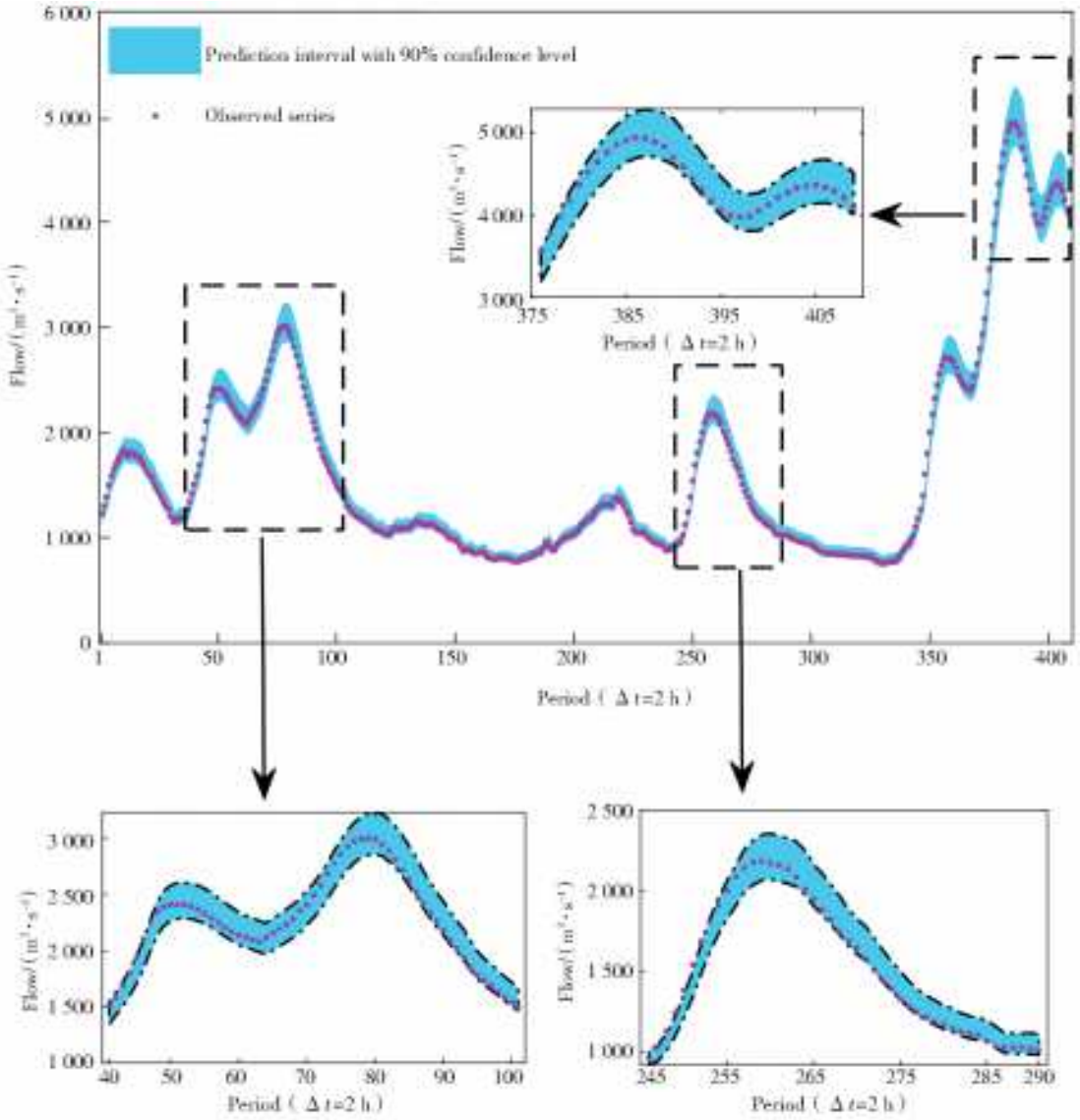

Fig. 1 Measured value series and the prediction interval with 90% confidence level (20120610)

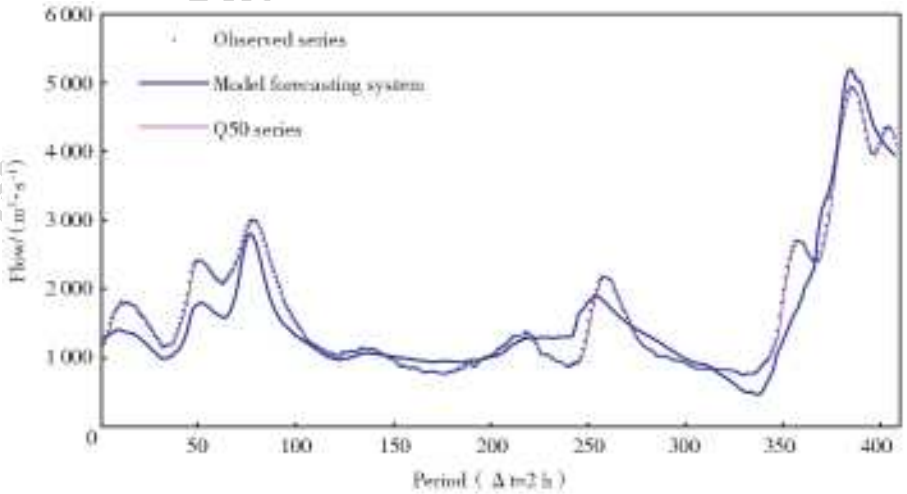

Fig. 2 Quantile regression model Q50 forecast series and original model forecast series

References:

[1] LIANG Z M, DAI R, LI B Q. Research progress of hydrological uncertainty analysis based on Bayesian theory[J]. Advances in Water Science, 2010, 21(2): 274-281. (in Chinese)

[2] ZHANG H G, GUO S L, HE X L, et al. Recent advancement and prospect of hydrological forecasting uncertainty study[J]. Journal of Shihezi University: Natural Science, 2006, 24(1): 15-21. (in Chinese)

[3] RAMOS M H, VAN ANDEL S J, PAPPENBERGER F. Do probabilistic forecasts lead to better decisions? [J]. Hydrology and Earth System Sciences. 2013, 17(6):2219-2232.

[4] LIANG Z M, JIANG X L, QIAN M K, et al. Probabilistic flood forecasting considering heterogeneity of error distributions[J]. Journal of Hydroelectric Engineering, 2017, 36(4):18-25. (in Chinese)

[5] LIANG Z M, JIANG X L, CAO Y X, et al. Probabilistic flood forecasting considering rainfall uncertainty [J]. Journal of Hohai University (Natural Sciences), 2016, 44(1):8-12. (in Chinese)

[6] KAVETSKI D, KUCZERA G, FRANKS S W. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory[J]. Water Resources Research, 2006, 42(3):W03407. Doi:10.1029/2005WR004368.

[7] BEVEN K, BINLEY A. The future of distributed models: model calibration and uncertainty prediction[J]. Hydrological Processes. 1992, 3(6):279-298.

[8] ZHAO P P, LYU H S, ZHU Y H, et al. Uncertainty analyses of parameters in TOPMODEL model based on GLUE and standard Bayesian method[J]. South-to-North Water Transfers and Water Science & Technology, 2014, 12(6):44-48. (in Chinese)

[9] CLARK M P, SLATER A G, RUPP D E, et al. Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models[J]. Water Resources Research. 2008, 44(12):

[10] LI Z J, HUANG P N, YAO C, et al. Application of flexible-structure hydrological models in different runoff generation regions[J]. Advances in WaTer Science. 2014, 25(1):28-35. (in Chinese)

[11] KRZYSZTOFOWICZ R. Bayesian theory of probabilistic forecasting via deterministic hydrologic model[J]. Water Resources Research, 1999, 35(9):2739-2750

[12] KUCZERA G, KAVETSKI D, FRANKS S, et al. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using stormdependent parameters[J]. Journal of Hydrology, 2006, 331(1-2):161-177.

[13] AJAMI N K, DUAN Q, SOROOSHIAN S. An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction[J]. Water Resources Research. 2007, 43(1):W1403.

[14] LI M L, YANG D W, CHEN J S. Probabilistic flood forecasting by a sampling-based Bayesian model[J]. Journal of Hydroelectric Engineering, 2011, 30(3):27-33. (in Chinese)

[15] KRZYSZTOFOWICZ R, KELLY K S. Hydrologic uncertainty processor for probabilistic river stage forecasting[J]. Water Resources Research, 2001, 36(11):3265-3277.

[16] LIU Z J, GUO S L, LI T Y, et al. Bayesian probability flood forecasting model and its comparative application study[J]. Shuili Xuebao, 2014, 45(9):1019-1028. (in Chinese)

[17] XING Z X, RUI X F, CUI H Y, et al. Bayesian probabilistic flood forecasting model based on AM-MCMC algorithm[J]. Journal of Hydraulic Engineering, 2007, 38(12):1500-1506. (in Chinese)

[18] JIANG X L, LIANG Z M, WANG C Q, et al. Application of BFS-HUP model in flood probability prediction of Tongguan station[J]. Yellow River, 2015, 37(7):13-15. (in Chinese)

[19] TODINI E. A model conditional processor to assess predictive uncertainty in flood forecasting[J]. International Journal of River Basin Management. 2008. 6(2):123-137.

[20] MONTANARI A, GROSSI G. Estimating the uncertainty of hydrological forecasts: A statistical approach [J]. Water Resources Research. 2008, 44(12):W00B08. DOI:10.1029/2008WR006897

[21] WANG Y L, LIANG Z M, WANG K, et al. Probabilistic flood forecasting based on multi-model MCP [J]. South-to-North Water Transfers and Water Science & Technology, 2018, 16(6):39-45. (in Chinese)

[22] VAN STEENBERGEN N, RONSYN J, WILLEMS P. A non-parametric data-based approach for probabilistic flood forecasting in support of uncertainty communication[J]. Environmental Modelling & Software. 2012, 33:92-105.

[23] WEERTS A H, WINSEMIUS H C, VERKADE J S, 2011. Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales)[J]. Hydrology and Earth System Sciences 15(1), 255-265.

[24] QIAO J, LI Z X. Quantile regression theory to explain and instance analysis[J]. Statistics and Decision, 2012, 19:104-107. (in Chinese)

[25] JIANG X L, LIANG Z M, HU Y M, et al. Research on assessment criteria in probabilistic flood forecasting[J]. Journal of Lake Sciences, 2020, 32(2):539-552. (in Chinese)