

基于数据分析技术的水文组合预报应用研究

张弛*, 周惠成, 李伟, 唐国磊

(大连理工大学 土木水利学院, 辽宁 大连 116024)

摘要: 水文组合预报方法是对多种预报模型的预报结果进行组合分析的一种预报方法. 针对历史洪水数据丰富和相对贫乏两种情况, 分别提出基于多目标模糊优选和基于贝叶斯分析的组合预报模型. 前者是选用多目标模糊优选模型根据各预报方案在不同流量级别下的预报精度确定方案优属度, 而后加权平均的组合预报模型; 后者则是以贝叶斯分析为基础, 同时结合专家经验、马尔可夫蒙特卡罗模拟、Gibbs抽样法, 并引入实时校正的组合预报模型. 以嫩江流域为实例, 分别对两种组合预报模型的精度进行了验证. 验证结果表明: 两种模型可行而且实用, 预报精度均明显高于单个模型的预报精度.

关键词: 水文组合预报; 模糊优选; 贝叶斯分析
中图分类号: TV 124 **文献标识码:** A

0 引言

水文预报是建立在尽可能掌握客观水文规律的基础上, 预报未来水文现象变化的应用科学技术, 及时准确的水文预报结果可以直接为防汛抢险、水资源合理利用与保护、水利工程建设和调度运用管理、工农业生产发展服务. 目前国内外主要集中于对三类水文预报模型的研究: 物理模型、黑箱模型、概念模型. 物理模型又叫做确定性模型或白箱模型, 依据水流的连续方程和动量方程来求解水流在流域的时间和空间变化规律, 比较典型的为 SHE、NWSH 和 SRFCN 等模型. 黑箱模型只研究水文系统与外界的信息交换情况, 而不研究水文系统内部的性质与结构, 即输入输出型模型, 比较典型的为人工神经网络模型. 统计模型. 概念模型是介于物理模型和黑箱模型之间的一种类型, 它含有较少的变量, 具有一定的物理意义但在某些方面又只能设置概念性参数, 比较典型的为新安江模型、大伙房模型等. 不同的模型有着不同的特点, 物理模型精度较高, 但需要进行大量的测量和对变化的连续物理过程有深入的了解; 黑箱模型简单灵活, 适用性广, 但对历史数据具有依赖性; 概念模型有一定的物理基础, 精度可

靠, 输入变量少于物理模型, 但该类模型许多环节上要借助于概念性元素的模拟或基于经验函数关系, 然而这些关系难以完全反映复杂的水文现象和物理作用机制.

各式各样的水文预报模型, 无论是基于流域的地形、气候、下垫面等实际特征为建模基础的传统水文预报模型, 还是基于长期水文历史数据挖掘为基础的数据挖掘水文预报模型, 都有自身的优缺点. 单个水文预报模型往往强化了水文预报的某些方面而忽视了另一些方面. 当前水文预报的发展趋势是多模型结合预报, 即选取能代表流域特征的多个模型, 将它们在预报过程中互相嵌入或将它们的预报结果进行综合, 这样可从多个角度模拟流域状况, 提高预报精度. Shan seldin^[1]将 5 个水文预报模型在 11 个流域进行预报应用, 采用 3 种不同的方法 (简单平均、加权平均和神经网络) 将各个预报结果进行综合, 通过比较得出用神经网络把 3 个方法各自结果综合后的预报结果明显好于利用单个水文模型得到的结果. Xiong 等^[2]在前者的基础上, 利用 Takagi-Sugeno 模糊系统对各个模型的预报结果进行模糊结合, 取得了更好的结果. See^[3]同样将几个水文预报模型 (神经网络、ARM、模糊推理

等)的结果输入神经网络得到数据融合后的预报结果,充分弥补单个模型对特定状况下预测不准的不足.在水文预报中,类似的根据实际情况,将多个预报模型结合起来做水文预报的应用还有很多^[4-8].

1 水文组合预报体系

目前国外研究水文组合预报时,往往使用几类智能算法进行组合预报,来判别各种组合预报方法的精度,很少针对流域的不同特点,尤其是流域历史数据的不同情况使用不同的组合预报方法进行预报.而中国是发展中国家,近几十年来区域经济发展不平衡导致各流域具体条件与情况不同,某些流域信息化建设时间长,积累了大量的水文数据;反之也存在着许多缺乏水文资料的流域,因此需要一个针对不同情况的合理的水文组合预报体系.

针对数据丰富和数据相对贫乏两种情况,本文分别应用基于多目标模糊优选^[9]的组合预报模型和基于贝叶斯分析的组合预报模型.基于多目标模糊优选的组合预报模型的原理是通过历史洪水来判断某种预报模型相对于某种类型洪水的

适合度,然后在实时预报中,判断洪水类别后根据不同的适合度利用各模型组合预报,这种模型需要较多的各种流量级别的历史洪水过程,一部分数据用于训练和检验基于数据挖掘的水文预报模型,另一部分数据用于模糊优选的训练.可以说,历史数据越丰富,数据范围涵盖的量级越大,基于模糊优选的组合预报模型精度就越高.但是在有些数据资源不是很丰富的流域,只能提供建立预报模型的水文数据,如果将这些数据划出一部分用于模糊模式识别,那么减少训练数据的预报模型精度必然会受到影响.针对这种情况,可采用基于贝叶斯分析的组合预报模型,该模型首先利用专家经验确定模型的分布类型,然后通过贝叶斯分析将已知各模型预报结果的后验分布转化成预报值的分布函数,利用马尔可夫蒙特卡罗模拟^[10](MCMC)方法抽样模拟得到组合预报的结果,在抽样模拟中,将基于时间序列的实时校正方法结合到MCMC中的Gibbs抽样中,得到了比单个模型预报结果更好的组合预报结果.由两种组合预报模型构建的水文组合预报体系如图1所示.

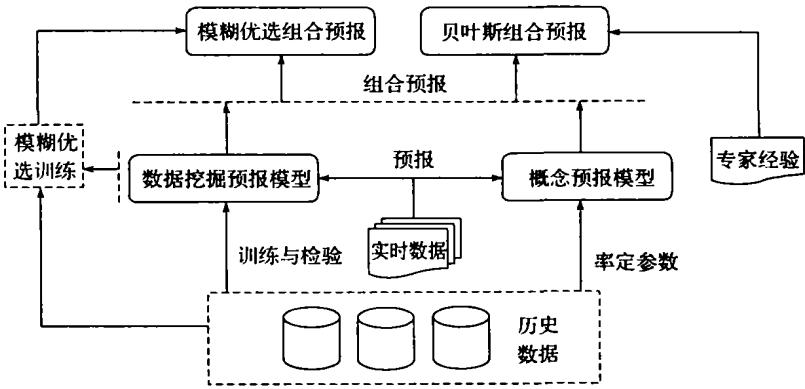


图 1 水文组合预报体系

Fig 1 The architecture of hydrological combined forecasting

2 水文组合预报模型

2.1 基于多目标模糊优选的组合预报模型

2.1.1 模型结构 模型结构如图2所示,用Fuzzy Mus Neural分别代表模糊推理法、马斯京干法和神经网络法,组合预报的步骤为

(1) 确定洪水流量级别的隶属度关系:利用模糊聚类分析所求控制站历史流量数据来确定该控制站的洪水流量级别隶属度关系.

(2) 单个模型预报:利用以上介绍的马斯京干法、神经网络模型和模糊推理模型分别对历史检验数据的数据集进行河道洪水预报,将预报结果放入结果集.

(3) 多目标模糊优选:在各个流量级别下,利用多目标模糊优选模型计算各预报模型的优属度.

(4) 预测流量级别:在预见期较长的情况下,预报精确的流量数值很难,但预报流量级别却有

很多方法.

(5) 针对不同级别的流量过程进行预报: 针

对不同的类别, 利用各预报模型相应的优属度组合进行预报.

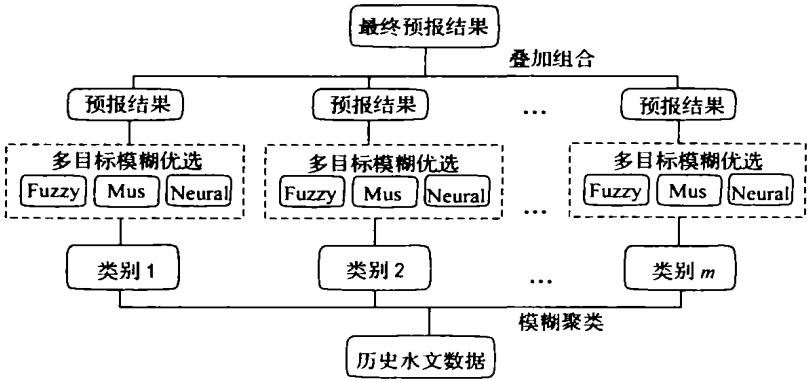


图 2 组合预报模型结构图

Fig 2 The structure of combined forecast model

2.1.2 预报方案多目标模糊优选 对所求控制站的历史汛期流量数据进行聚类分析, 根据聚类得到的结果确定该控制站的洪水流量级别隶属度. 流量级别聚为 M 类 $C_m \in \{C_1, \dots, C_M\}$. 针对 N 场历史洪水, 分别利用马斯京干、神经网络和模糊推理水文预报模型进行预报, 得到每场洪水预报的洪峰、峰现时间和流量过程的相对误差, 用矩阵表示

$$E^k = \begin{pmatrix} e_{11}^k & e_{12}^k & e_{13}^k \\ e_{21}^k & e_{22}^k & e_{23}^k \\ e_{31}^k & e_{32}^k & e_{33}^k \end{pmatrix} = (e_{ij}^k), \quad k = 1, \dots, N \quad (1)$$

其中 k 为洪号, $i = 1, 2, 3$ 分别代表预报洪峰、峰现时间和流量过程 3 个指标, $j = 1, 2, 3$ 分别代表模糊推理法、马斯京干法和神经网络法. 利用所求控制站的洪水流量级别隶属度分别求出这 N 场洪水流量级别隶属度

$$\{\{\mathbb{T}_1, \mathbb{T}_2, \dots, \mathbb{T}_M\}, \{\mathbb{T}_1, \mathbb{T}_2, \dots, \mathbb{T}_M\}, \dots, \{\mathbb{T}_1, \mathbb{T}_2, \dots, \mathbb{T}_M\}\} \quad (2)$$

其中 $\mathbb{T}_m$ 为第 k 场洪水属于第 m 个洪水流量级别的隶属度. 利用重心法的原则, 采用下式

$$r_{ij}^m = \frac{\sum_{k=1}^N \mathbb{T}_m^k e_{ij}^k}{\sum_{k=1}^N \mathbb{T}_m^k} \quad (3)$$

求出 3 种预报模型针对各个洪水流量级别的多指标预报相对误差矩阵

$$R^m = \begin{pmatrix} r_{11}^m & r_{12}^m & r_{13}^m \\ r_{21}^m & r_{22}^m & r_{23}^m \\ r_{31}^m & r_{32}^m & r_{33}^m \end{pmatrix} = (r_{ij}^m), \quad m = 1, \dots, M \quad (4)$$

利用得到的结果和陈守煜的多目标模糊优选模型^[9]求出在各流量级别下的各预报方案的优属度, 从而在实时预报中利用各预报方案的优属度组合多模型预报结果得到预报值.

2.2 基于贝叶斯分析的组合预报模型

2.2.1 模型简介 图 3 为贝叶斯组合预报的模型结构图. 假定利用 3 个模型进行水文预报, 得到 3 个结果 $\{X_1, X_2, X_3\}$, 而 X_0 为水文真实值. 利用贝叶斯分析, 问题就变为将 3 个确定水文预报模型的预报结果作为条件, 预报在该条件下水文值的问题, 基本公式如下:

$$P(X_0 | \{X_1, X_2, X_3\}) = \frac{P(\{X_1, X_2, X_3\} | X_0) \cdot P(X_0)}{P(\{X_1, X_2, X_3\})} \quad (5)$$

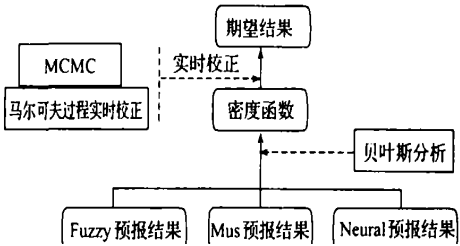


图 3 基于贝叶斯分析的组合预报模型结构图
Fig 3 The structure of Bayesian theory-based combined forecast methods

其中由于 3 个预报模型相互独立, 则有

$$P(\{X_1, X_2, X_3\} | X_0) = \prod_{i=1}^3 P(X_i | X_0) \quad (6)$$

根据经验, 一般预报值符合以真实值为期望的正态分布:

$$P(X_i | X_0) \sim N(X_0, \sigma_i^2) \quad (7)$$

对于参数 X_0, X_i 的先验分布, 如果没有任何信息, 一般假设服从均匀分布:

$$X_0 \sim U(a, b), X_i \sim U(e, f)$$

这里使 $\lambda = 1/e^2$ 表示模型的精度, 一般情况下, 正态分布的方差可选用伽马分布: $\lambda \sim \Gamma(c, d)$. 则这时 X_0 的条件密度概率函数为

$$f(X_0 | \{X_1, X_2, X_3\}) = \prod_{i=1}^3 [g(X_i | X_0)] \cdot h(X_0) \Big/ \left[\prod_{i=1}^3 q(X_i) \right] \quad (8)$$

其中

$$g(X_i | X_0) = \frac{\lambda_i}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}(X_i - X_0)^2 \lambda_i\right\} \cdot \frac{d^c}{\Gamma(c)} \lambda_i^{c-1} \exp\{-d\lambda_i\} \quad (9)$$

$$h(X_0) = \frac{1}{b-a} \quad (10)$$

$$q(X_i) = \frac{1}{f-e} \quad (11)$$

将式 (9)、(10)、(11) 代入式 (8) 中, 得到

$$f(X_0 | \{X_1, X_2, X_3\}) = C \prod_{i=1}^3 \left[\lambda_i^{c-1/2} \cdot \exp\left\{-\frac{1}{2}(X_i - X_0)^2 \lambda_i - d\lambda_i\right\} \right] \quad (12)$$

其中 C 为多个常数表达式形成的常数, 式 (12) 为所求值的后验分布形式.

2.2.2 模型流程

(1) 各模型预报

首先利用 3 个预报模型分别进行预报, 求出各个模型的预报结果.

(2) 贝叶斯确定后验密度函数

将式 (12) 中表达的有关 X_0 的后验分布转化为有关 $X_1, X_2, X_3, \lambda_1, \lambda_2, \lambda_3, X_0$ 等参数的联合概率密度分布, 由于过于复杂难以直接求解, 可以在下一步通过 MCMC 法抽样形成稳定的分布求解.

(3) 马尔可夫蒙特卡罗 (MCMC) 模拟分布

简单地说, 马尔可夫蒙特卡罗 (MCMC) 先对每个密度函数中的变量赋初值, 然后对每个变量随机抽样, 用抽取的样本计算并更新分布, 如此依次循环, 直至满足要求. 这里首先根据式 (12) 求出各个参数 $\lambda_1, \lambda_2, \lambda_3, X_0$ 的后验分布, 然后在下一步中通过 Gibbs 抽样器抽样模拟, 在抽样过程中, 利用马尔可夫过程进行实时校正, 详见文献 [11].

(4) Gibbs 抽样

对 $\lambda_1, \lambda_2, \lambda_3, X_0$ 反复抽样迭代 m 次 (例如

50 000), 其间每隔一段距离计算一次参数的遍历均值, 当均值稳定后, 认为它收敛了. 若假定在第 b 次以后 Gibbs 抽样收敛, 则可以通过 $m - b$ 个样本得到 $\lambda_1, \lambda_2, \lambda_3, X_0$ 的分布与期望等参数, 其中 X_0 的期望值即为预报值.

3 应用实例

以嫩江流域的诺敏河上的古城子、讷漠尔河上的德都、嫩江干流上游的阿彦浅 3 个河道控制站为对象 (如图 4 所示), 研究嫩江干流下游同盟的河道洪水预报.

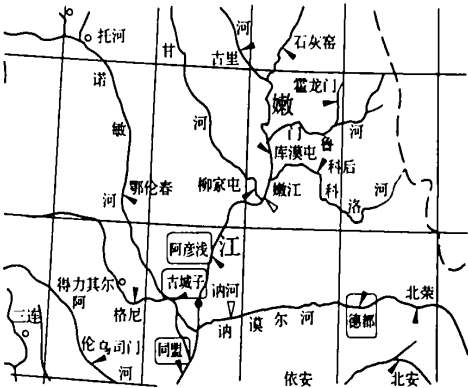


图 4 嫩江流域古城子、阿彦浅、德都、同盟四站位置关系

Fig. 4 Spatial distribution of Guchengzi, Ayanqian, Dedu and Tongmeng in Nenjiang river basin

从图 4 可以看出, 同盟站来水主要由上游的干流和支流洪水传播产生, 因此选取干支流上具有代表性质的三站进行分析. 和诺敏河、讷漠尔河、嫩江上游流域比较, 从古城子、阿彦浅、德都到同盟站之间的流域面积比较小, 因此忽略了此区间的降雨对同盟站的影响. 选取该流域历史上相对丰水年 1973、1978、1979、1980、1984、1988 年的汛期四站径流流量数据为训练数据, 建立基于关联规则分析的模糊推理预报模型、神经网络预报模型, 同时建立概念预报模型——马斯京干法.

3.1 基于多目标模糊优选的组合预报

利用这 3 种预报模型分别对具有洪水级别代表性的 1969、1976、1977、1981、1982、1983 年的汛期同盟站的径流过程进行预报, 预报结果详见文献 [11], 不同的流量级别下, 不同的预报模型精度往往不同.

选择洪水过程的峰值、峰现时间、流量过程三

者的预报误差分别为指标 1 2 3 计算出针对各个洪水流量级别的多指标预报相对误差矩阵, 如

表 1 所示, 模糊推理、马斯京干、神经网络分别简记为方法 1 方法 2 方法 3

表 1 低、中、高流量级别的多指标预报相对误差
Tah 1 The relative errors of multi-objective forecast for low, medium and high flow ranks

指标	低流量			中流量			高流量		
	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3	方法 1	方法 2	方法 3
1	0.065 7	0.048 05	0.048 7	0.086 3	0.025 27	0.075 2	0.036 1	0.054 74	0.138 2
2	0.009 5	0.028 56	0.010 5	0.014 4	0.044 74	0.025 8	0.010 7	0.039 41	0.024 3
3	0.113 2	0.130 92	0.072 2	0.081 9	0.092 28	0.056 2	0.061 8	0.082 59	0.047 9

利用陈守煜提出的多目标模糊优选模型, 对各种情况下预报模型的优属度进行计算, 结果如表 2 所示.

表 2 不同流量级别下的预报方案优属度
Tah 2 The memberships of forecast scheme at different flow ranks

流量级别	预报方案优属度		
	方法 1	方法 2	方法 3
低流量	0.180	0.520	0.999
中流量	0.175	0.520	0.481
高流量	0.989	0.338	0.263

根据表 2 结果, 针对不同量级采用加权平均的预报方法, 计算 1989 年径流预报, 结果如图 5 所示, 图中所示的是基于多目标模糊优选的组合预报与真实值在 3 种预报模型所围成的预报范围中的位置. 基于多目标模糊优选的组合预报的结果要好于单个模型预报结果.

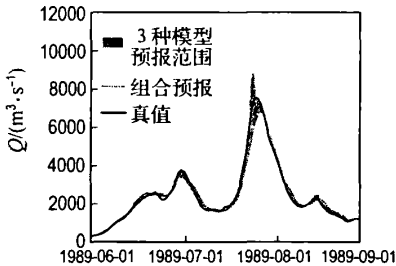


图 5 基于多目标模糊优选的组合预报结果对比图
Fig. 5 Comparison of combined forecast results based on fuzzy optimization of multi-objective

3.2 基于贝叶斯分析的组合预报

这里直接利用基于贝叶斯分析的组合预报模型对 1976 1977 1981 1982 1983 1989 年汛期径流进行预报, 6 年汛期 20 场洪水预报结果的峰值、峰现时间在不同保证率下的合格率见表 3

表 3 汛期 20 场洪水多模型预报结果对比
Tah 3 The forecasting results of 20 floods using different models

预报方法	峰值相对误差合格率 /%			峰现时间预报合格率 /%		确定性系数 (平均) /%
	≤ 2%	≤ 5%	≤ 10%	= 0 时段	≤ 1 时段	
贝叶斯组合预报	35	85	95	55	100	90.37
模糊推理	5	45	80	50	100	83.94
马斯京干	35	80	90	40	90	78.83
神经网络	10	55	80	50	100	85.54

从表 3 可以看出, 贝叶斯组合预报模型的预报结果也明显好于单个模型的预报结果.

4 结 论

针对数据丰富和数据相对贫乏两种情况, 本文分别提出基于多目标模糊优选的组合预报模型和基于贝叶斯分析的组合预报模型.

报结果中的峰值误差、峰现时间误差、流量过程误差为指标, 利用模糊优选计算出各预报模型针对不同类别洪水的优属度, 然后对各模型的预报结果进行组合, 得到了比单个模型预报结果更好的组合预报结果. 但由于要对各模型优属度进行训练, 此组合方法需要较长时间的流域径流资料. 基于贝叶斯分析的组合预报模型, 利用贝叶斯分析对已知各模型预报结果为后验分布的函数进行

转换, 然后利用马尔可夫蒙特卡罗 (MCMC) 将难以积分的分布函数抽样模拟, 在每次进行下一时段预报时, 利用马尔可夫时间序列性质将前一时段预报精度通过 Gibbs 抽样的迭代函数影响到后一时段的预报中, 即在迭代过程中考虑了时间序列数据 (近期的历史数据), 最终得到较好的预报结果。与基于多目标模糊优选的组合预报模型相比, 该模型可以在历史数据较少的前提下保证一定的精度, 但需要根据专家经验确定各个分布函数。并且当预见期比较长时, 根据时间序列的实时校正的效果相对下降, 此时的预报精度也会随之下降。因此, 两种组合预报方法各有特点与适用范围, 可根据资料等情况选用之。

参考文献:

- [1] SHAM SELDIN A Y, O'CONNOR K M, LIANG G C, *et al* Methods for combining the outputs of different rainfall-runoff models [J]. **J of Hydrol** 1997, **197**: 203-229
- [2] XIONG Li-hua, SHAM SELDIN A Y, O'CONNOR K M. A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system [J]. **J of Hydrol** 2001, **245**: 196-217
- [3] SEE L, ABRAHART R J Multi-model data fusion for hydrological forecasting [J]. **Comput & Geosci** 2001, **27**: 987-994
- [4] SUDHEER K P, NAYAK P C, RAMASASTRIK S Improving peak flow estimates in artificial neural network river flow models [J]. **Hydrol Processes** 2003, **17**: 677-686
- [5] JAIN A, SUDHEER K P, SRINIVASULU S. Identification of physical processes inherent in artificial neural network rainfall runoff models [J]. **Hydrol Processes** 2004, **18**(3): 571-581
- [6] TINGSANCHALIT, GAUTAM M R. Application of tank, NAM, ARMA and neural network models to flood forecasting [J]. **Hydrol Processes** 2000, **14**: 2473-2487
- [7] ABRAHART R J, SEE L. Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments [J]. **Hydrol Processes** 2000, **14**: 2157-2172
- [8] ÖZELKAN E C, DUCKSTEIN L. Fuzzy conceptual rainfall-runoff models [J]. **J of Hydrol** 2001, **253**: 41-68
- [9] 陈守煜. 工程水文水资源系统模糊集分析理论与实践 [M]. 大连: 大连理工大学出版社, 1998
- [10] GILKS W S. **Markov Chain Monte Carlo in Practice** [M]. London: Chapman & Hall, 1996
- [11] 张 弛. 数据挖掘技术在水文预报与水库调度中的应用研究 [D]. 大连: 大连理工大学, 2005

Research on applications of data analysis technology to hydrological combined forecasting

ZHANG Chi, ZHOU Hui-cheng LI Wei TANG Guo-lei

(School of Civil and Hydraulic Eng., Dalian Univ. of Technol., Dalian 116024 China)

Abstract Hydrological combined forecasting is a method giving summary and analysis to different forecasting results produced by different predication models. Aiming at the two situations—abundance or lack of history flood data, the combined forecasting models separately based on multi-objective fuzzy optimization theory and Bayesian analysis theory are proposed correspondingly. The former model introduces multi-objective fuzzy optimization theory to find out optimal relative membership degree of each projection on some precision at different discharges, and then by means of weighted average to confirm the optimal forecasting result; the latter model is based on Bayesian theory, combined with experts' experiences, MCMC simulation, Gibbs sampling and real-time auto-tuning technology. Taking the drainage area of Nenjiang for instance, the precision of the two integrated models was tested, and the result indicates that the established models are available and practical with higher precision than that of any single model.

Key words hydrological combined forecasting; fuzzy optimization; Bayesian analysis