DOI:10.20040/j.cnki.1000-7709.2023.20221477

基于时空水文数据相似性度量的洪水预报模型

谢 敏[1],朱松挺[1],傅 韬[2],欧阳志勇[1]

(1. 江西省防汛信息中心,江西 南昌 330009; 2. 江西省水投江河信息技术有限公司,江西 南昌 330096)

摘要:在水文数据只包括雨量站数据和河道监测断面水位实测数据的情况下,提高中小型流域洪水预报精度一直是一个重大挑战。根据数据特征实现洪水的自动化编码,利用决策树模型对历史洪水分类,得到候选洪水组和淘汰洪水组,提高计算效率。在此基础上,建立基于时空水文数据相似性度量的洪水预报模型,以江西省大汾水流域水文数据为例开展水文数据分析,并随机选取四场洪水进行验证,结果表明洪峰水位合格率为100%,峰现时间合格率为75%,精度较高,在水文数据研究中具有重要理论价值和现实意义。

关键词:洪水识别;相似性搜索;决策树;洪峰预测

中图分类号:TV122 **文献标志码**:A **文章编号**:1000-7709(2023)08-0077-04

1 概况

大汾水流域位于江西省西南部、湖南省东南部,是遂川江上游右岸的一级支流。流域面积275 km^2,涉及湖南省桂东县和江西省遂川县,流域东毗左溪,南依桥头水,西、北邻遂川江。干流流经湖南省桂东县清泉镇庄川村和江西省遂川县大汾镇、西溪乡、堆子前镇,主河道长46 km。流域多年平均年降水量1 590 mm、年水面蒸发量652 mm、年径流量3.2×10^8 m^3。流域呈树叶形,西南高东北低,上游为山区,中下游以丘陵为主。属山岳丘陵区,主河道纵比降7.35‰。流域地处华南地层区赣中南褶隆,岩性为早古生代寒武纪砂岩和页岩。近年来长江流域夏季的短时期降雨量突破历史极值,各大湖泊河流均超过保证水位或达到历史最高水位,给受灾地区造成了重大经济损失[1],因此迫切需要对流域水文特性做出准确预报。但影响水位的因素众多,包括各支流流量、流域内降雨量等,使得水位等因素的规律难以把握,精确预报洪水面临着重大挑战[2]。目前,洪水预报方法多采用物理水文模型,为了增加水文模型预报的精度,已对各类水文模型进行不断改进[3]。然而,现有水文模型如新安江模型、HEC-HMS模型等均存在明显的不足[4,5]。因此,本文运用统计学理论,分析大汾水流域水文数据,实现洪水的自动化识别和编码,建立了基于时空雨量相似性度量的洪水预报模型,旨在提高洪峰预报精度,减轻洪涝灾害造成的经济损失。

2 洪水自动化识别和编码

2.1 洪水自动化识别

大汾水流域目前共有20个监测站,分别为中村、洛阳、大汾、石花、西溪、廖坊、仙人井、文溪、茶洞、五指峰、集龙洞、堆子前、遂川中村、遂川廖坊、遂川寨南、杨芳、遂川文坳、遂川鹿坑、遂川秋平、井坑水库等站点。部分测站见图1。根据大汾水流域水情特性,使用降雨量和水位涨幅来量化洪水。其量化规则为:①获得年最高水位,判断其24 h内涨率是否达到最小洪水的涨率要求,若满

图1 大汾水部分测站点

Fig. 1 Some survey stations in Dafenshui

收稿日期:2022-07-18,**修回日期**:2022-10-01

基金项目:江西省水利厅一般科技项目(202123YBKT14)

作者简介:谢敏(1985-),男,硕士、高级工程师,研究方向为水利信息化,E-mail:xm@jxsl.gov.cn

通讯作者:朱松挺(1982-),男,硕士、高级工程师,研究方向为水利信息化,E-mail:zhust@jxsl.gov.cn

足涨幅要求，则说明该时段可能发生了洪水，否则该时段未发生洪水。②若该时段发生洪水，则记录洪峰时间。起涨点判断依据为从洪峰时间开始向前找到最低点和次高点，判断次高点涨幅是否满足起涨的涨幅要求，若不满足则最低点为起涨点，若满足则继续往前找最低点和次高点，直到确定起涨点。③退水点判断依据为从洪峰时间开始往后找到最低点和次高点，判断次高点跌幅是否满足退水的跌幅要求；若不满足则最低点为退水点，若满足则继续往前找最低点和次高点，直到确定退水点。④删除降雨总量较小的情况。计算起涨点到退水点的平均雨量和时长，若平均雨量小于最小平均雨量或时长小于最小时长，则该场降雨未达到洪水要求，若满足要求则该时刻发生了洪水。

2.2 洪水编码

在对洪水自动化识别过程中，可以识别出每场洪水的主要特征，分别为时段、时长、平均雨量、雨量方差、洪峰等特征。为方便人工和机器快速了解洪水特征，将洪水的一些特征组成一个特征码。特征码使用“时段＋时长＋平均雨量＋雨量方差＋峰值＋预留位”共 32 位字符。

3 基于时空雨量相似性度量的洪水预报模型

3.1 基于决策树的相似筛选

决策树作为机器学习的十大经典算法之一，对处理分类和回归问题具有良好的性能[6]。因此，采用决策树的分类思想，筛选出满足要求的候选洪水。对于决策树，需要确定每个样本点的输入特征和输出变量。在初筛过程中，每次洪水都有对应的特征码，对特征码拆分，将得到样本点的多个输入特征，根据多个输入特征将其分为候选洪水组和淘汰洪水组。其主要思想为：①获取特征码后，将其分解为多个输入特征。②判断每个输入特征是否满足初筛要求，若其中一个特征不满足要求，则将该洪水分为淘汰洪水组；若所有的特征都满足要求后，则将该洪水分为候选洪水组。

3.2 基于时空雨量的洪水预报模型

采用相似性搜索的基本原理，依据定义的相似策略，在已建立的洪水特征库中找出最接近当前雨型（洪水）的数据序列。在实际的降雨过程中，降雨分布具有时空特性。同一时刻各个监测站点收集的雨量信息会呈现出一定的差异性，称为空间差异性。同样，在同一站点的不同时刻收集到雨量信息也具有差异性，称为时间差异性。因此，考虑将时间影响因素和空间影响因素融合到模型中，提出了采用矩阵形式的时空相似搜索。设当前洪水的时空雨量分布矩阵为 $\boldsymbol{A} \in S \times T$ 。

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1T} \\ a_{21} & a_{22} & \cdots & a_{2T} \\ \vdots & \vdots & \vdots & \vdots \\ a_{S1} & a_{S2} & \cdots & a_{ST} \end{bmatrix}$$

式中，S 为监测站的个数；T 为当前洪水的时长；a_{ST} 为第 S 个监测站 T 时刻的雨量。

为获得当前洪水 T 小时后的水位分布，计算过程如下。

步骤 1 从洪水特征库中获得历史第 j 场洪水的时空雨量分布矩阵，即 $S \times H$ 维矩阵 $\boldsymbol{B}$，其中 $T<H$，

$$\boldsymbol{B}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1H} \\ b_{21} & b_{22} & \cdots & b_{2H} \\ \vdots & \vdots & \vdots & \vdots \\ b_{S1} & b_{S2} & \cdots & b_{SH} \end{bmatrix}$$

步骤 2 基于曼哈顿距离的矩阵相似搜索。求解矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$ 相似度的方法是先求两矩阵对应列向量的曼哈顿距离，取平均值，得到两矩阵的距离度量，距离越小，矩阵越相似。其中曼哈顿距离 d 计算公式[7]为：

$$d(X,Y)=\sum_{i=1}^{n} | x_i - y_i | \tag{1}$$

其计算过程为：

$$C_j = | \boldsymbol{A} - B_j | = \begin{bmatrix} | a_{11}-b_{11} | & | a_{12}-b_{12} | & \cdots & | a_{1T}-b_{1T} | \\ | a_{21}-b_{21} | & | a_{22}-b_{22} | & \cdots & | a_{2T}-b_{2T} | \\ \vdots & \vdots & \vdots & \vdots \\ | a_{S1}-b_{S1} | & | a_{S2}-b_{S2} | & \cdots & | a_{ST}-b_{ST} | \end{bmatrix} \tag{2}$$

$$D_j = \left[\frac{1}{S}\sum_{s=1}^{S} | a_{s1}-b_{s1} | \ \ \frac{1}{S}\sum_{s=1}^{S} | a_{s2}-b_{s2} | \ \cdots \ \frac{1}{S}\sum_{s=1}^{S} | a_{sT}-b_{sT} | \right] \tag{3}$$

使 $c_{jt}=\dfrac{1}{S}\sum_{s=1}^{S} | a_{st}-b_{st} |$，$t=1,2,\cdots,T$，则矩阵 $\boldsymbol{A}$、$\boldsymbol{B}$ 的距离 d_j 为：

$$d_j=\frac{1}{T}\sum_{t=1}^{T} c_{jt} \tag{4}$$

步骤 3 循环遍历该小流域的候选组结果，重复步骤 1、2 得到与矩阵 $\boldsymbol{A}$ 最相似的历史洪水。相似洪水对应的水位曲线即为预测当前洪水 T 小时后的水位曲线图。

4　大汾站应用结果分析

选取研究流域 2008～2021 年的水文数据(包括降雨量和水位),对大汾站进行洪水峰值和峰现预报。

4.1　洪水自动化识别及编码

在洪水自动化编码前,需要提前设定 3 个参数,即涨率、最小平均雨量和最小时长。通过试验发现,最小平均雨量和最小时长对洪水识别的结果影响不大,其作用只是排除小部分降雨量很小的洪水。涨率是识别洪水最重要的参数,通过试验分析,将涨率设定为 18 cm/h,最小平均雨量设定为 0.1 mm,最小时长设定为 16 h。将历史水文数据输入到事先设定的洪水定义算法中,智能算出所有历史洪水场次共 96 场,其中 9 场洪水列表见表 1。

表 1　洪水特征库中 9 场洪水列表

Tab. 1　List of nine floods in flood characteristic reservoir

序号	洪水特征码	起涨点	退水点
a	080501003601.02016.30297.0900000	2008-05-01 20:00:00	2008-05-03 08:00:00
b	080624004406.04098.87298.3500000	2008-06-24 12:00:00	2008-06-26 08:00:00
c	090519009604.68078.89297.1200000	2009-05-19 16:00:00	2009-05-23 16:00:00
d	100617008505.70104.26297.3800000	2010-06-17 22:00:00	2010-06-21 11:00:00
e	110805005001.50024.01297.1000000	2011-08-05 16:00:00	2011-08-07 18:00:00
f	120303009204.83098.92296.9300000	2012-03-03 23:00:00	2012-03-07 19:00:00
g	140521005405.83094.72297.5200000	2014-05-21 13:00:00	2014-05-23 19:00:00
h	140811007607.52135.88297.8100000	2014-08-11 14:00:00	2014-08-14 18:00:00
i	160415008904.16067.52296.9300000	2016-04-15 06:00:00	2016-04-18 23:00:00

每个特征码都有 32 位,其中 1～6 位代表洪水时段,7～10 位代表洪水时长,11～15 位代表平均雨量,16～21 位代表雨量方差,22～27 位代表洪水水位的峰值,28～32 位代表预留位,表示以后可加入的其他要素。如查询第一场洪水的特征码,就可以知道该洪水的开始时间为 2008 年 5 月 1 日,时长 36 h,平均雨量 1.02 mm,洪峰 297.09 m 等。对应 9 场洪水的过程曲线见图 2。由图 2 可知,自动化识别的每次洪水均有明显涨幅。

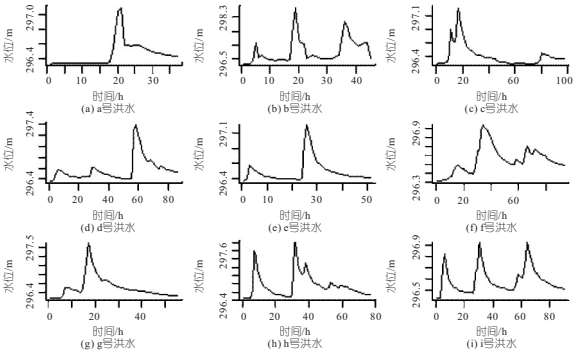

图 2　洪水过程曲线

Fig. 2　Flood hydrographs

4.2　洪水预报结果与分析

根据《水文情报预报规范》[8],将预测洪峰水位小于真实洪峰水位变幅的 20%,预测洪峰出现时间小于真实峰现变幅 3 h 视为合格。

4.2.1　洪水预报初步筛选

为了提高运行速度,需要在寻找相似洪水前进行初步筛选,去掉明显不匹配的洪水。洪水的主要特征有时段、时长、平均雨量和水位峰值等,根据这些特征,运用决策树模型初步筛选出与待测洪水明显不匹配的洪水。

为了验证初步筛选能加快运行效率,使用未初步筛选的时空雨量预测方法和采用了初步筛选的时空雨量预测方法进行对比。为了确保试验的真实性,对两组试验分别运行了 10 次,两组试验运行 10 次所需时间见表 2。由表 2 可看出,初步筛选的运行时间变少。在测试第 10 场洪水时,未初步筛选的算法需要从 95 场洪水中匹配最相似的洪水,而初步筛选后的算法只需从 19 场洪水中匹配最相似的洪水。

表 2　10 场洪水预报所需时间

Tab. 2　Time required for ten flood forecasts　s

场数	未初步筛选的时空雨量预测法	初步筛选的时空雨量预测法
第 1 场	1.720 000 000 001 16	1.389 999 999 999 42
第 2 场	1.729 999 999 995 93	1.349 999 999 998 54
第 3 场	1.739 999 999 997 96	1.400 000 000 001 46
第 4 场	1.760 000 000 002 04	1.379 999 999 997 38
第 5 场	1.760 000 000 002 04	1.340 000 000 003 78
第 6 场	1.800 000 000 002 91	1.520 000 000 004 07
第 7 场	1.959 999 999 999 13	1.360 000 000 000 58
第 8 场	1.870 000 000 002 62	1.330 000 000 001 75
第 9 场	1.809 999 999 997 67	1.349 999 999 998 54
第 10 场	1.830 000 000 001 75	1.419 999 999 998 25

4.2.2　洪水预报结果

为验证所提预报模型的精度,从 96 场洪水中随机取出 4 场洪水(#1～#4)作为验证集,其中只取每场洪水前 1/4 时段的水位和雨量时间序列数据作为输入特征,从剩余 92 场洪水中找到最相似的洪水,进而对 1/4 时段后的洪峰水位和峰现时间进行预报。表 3 描述了大汾站洪水预报的结果,图 3 展示了待测洪水的实际曲线和预测曲线。

由表 3 可看出,4 场洪水的洪峰相对误差都在 20%以内,峰现误差时间有 3 场在 1 h 之内,表明洪峰预报合格率 100%,峰现误差时间合格率 75%。虽然有 1 场洪水峰现时差 8 h,但从图 3 右下角的水位曲线可以发现,预测洪水的次洪峰时间刚好处于实测峰值时间点。从试验结果可发现,所提出的基于时空雨量相似性度量的水文预测模型具有较高的精度。

表 3 大汾站洪水预报结果

Tab. 3 Flood forecast results of dafen station

洪号	实测洪峰水位/m	预报洪峰水位/m	洪峰相对误差/%	峰现时间误差/h
#1	297.15	297.23	0.027	1
#2	297.23	297.12	0.037	2
#3	296.84	296.91	0.024	0
#4	296.82	296.74	0.030	8

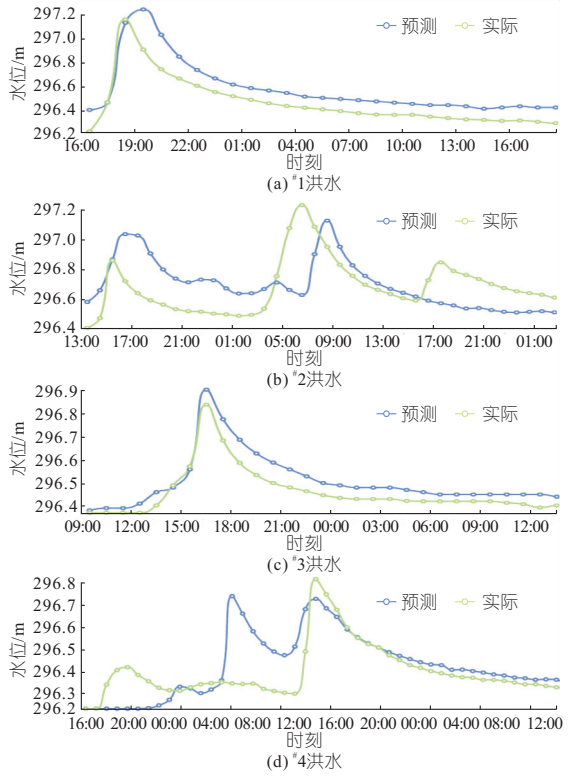

图 3 大汾站 4 场洪水预报

Fig. 3 Four flood forecasts at dafen station

5 结论

a. 构建了一种适合数字水文的中小型流域洪水预报模型，通过分析雨量和水位数据，自动化识别洪水并对洪水的洪峰和峰现时间进行预测，实例应用表明，洪水的自动化识别技术能较好地识别出洪水，并反映洪水的基本特征；基于决策树模型的初步筛选技术能够有效地降低模型运行时间，快速获得预测结果。所构建的基于相似性度量的洪水预报模型只需要雨量和水位数据作为输入数据，无其他额外的参数，相比物理模型如新安江模型等，不仅参数少，耗时少，而且预报精度高。

b. 针对现有相似性搜索的水文模型精度不高的问题，提出了一种基于时空雨量相似性度量的洪水预报模型。从时间和空间两个维度考虑雨量的相似性，进而改进基于曼哈顿的最小距离公式。试验结果表明考虑时空相似性后，精度更高。

参考文献：

[1] 徐帅帅. 基于分布式水文模型的洪水预报及水库削峰的案例研究[D]. 济南：山东建筑大学，2019.

[2] 杨冬. 水文数据分析中应用数据挖掘技术的若干研究[J]. 黑龙江水利科技，2017，45(9)：35-37.

[3] 李振亚，黄国新，肖凤林，等. 基于 TOPMODEL 的分布式水文模型在中小流域的应用研究[J]. 江西水利科技，2020，46(5)：374-381.

[4] 孙小洪，赵兵，孙逸群，等. 机器学习技术在曹娥江流域洪水预报中的应用[J]. 浙江水利科技，2022，50(2)：83-87.

[5] 周聂，侯精明，陈光照，等. 基于机器学习的山洪灾害快速预报方法[J]. 水资源保护，2022，38(2)：32-40，111.

[6] 李会，胡笑梅. 决策树中 ID3 算法与 C4.5 算法分析与比较[J]. 水电能源科学，2008(2)：129-132，163.

[7] 李伟华. 水库健康监测大数据清洗方法研究[D]. 泰山：山东农业大学，2019.

[8] 郭磊，习雪飞，李军，等. 基于水文数据挖掘技术的洪峰预测分析研究[J]. 水电能源科学，2021，39(12)：80-83.

Flood Forecasting Model Based on Similarity Measurement of Spatio-temporal Hydrological Data

XIE Min[1]，ZHU Song-ting[1]，FU Tao[2]，OUYANG Zhi-yong[1]

(1. Jiangxi Flood Control Information Center，Nanchang 330009，China；

2. Jiangxi Shuitou Jianghe Information Technology Co.，Ltd.，Nanchang 330096，China)

Abstract：In the case that hydrological data only include rainfall station data and measured water level data of river monitoring section，it is always a major challenge to improve the accuracy of flood forecasting in small and medium-sized basins. In this paper，flood automatic coding was realized according to data characteristics，and historical floods were classified by using decision tree model，thus the candidate flood groups and eliminated flood groups were obtained and the calculation efficiency was improved. On this basis，a flood forecasting model was established based on the similarity measurement of spatio-temporal hydrological data. Taking the hydrological data of Dafenshui Basin in Jiangxi Province as an example，a case analysis of hydrological data was carried out，and four floods were randomly selected for verification. The results show that the qualification rate of flood peak water level is 100%，and the qualification rate of flood peak time is 75%，the accuracy is high，which has important theoretical value and practical significance in hydrological data research.

Key words：flood identification；similarity search；decision tree；flood peak prediction