DOI:10.20040/j.cnki.1000-7709.2023.20230125

面向长江上游面雨量长期预报的聚类模型构建和改进

顾 丽,陈瑜彬,李春龙

(长江水利委员会水文局,湖北 武汉 430010)

摘要:聚类预报模型是一种基于长系列数据分析,采用数理统计方法进行量级预报的长期预报模型,并应用于国内多个流域长期预报中。经过多年实践,模型在准确性和建模效率上还有提升空间。基于该模型原理,选取长江上游为预报示范区,6~8月逐月面雨量为预报对象,预报与均值相比偏多或偏少和量级作为该模型的检验方法,阐述聚类模型构建的流程和改进。最后利用改进前后的模型分别预测2020~2022年长江上游6~8月逐月面雨量,预测结果表明改进后的模型在距平符号一致率和距平误差百分率上有较大提升,且建模过程简单,操作性强,有一定参考价值。

关键词:预报;聚类;模型;长江上游;面雨量

中图分类号:[TV124] **文献标志码**:A **文章编号**:1000-7709(2023)06-0005-04

1 引言

由于水文气象要素长期变化的复杂性,影响其未来变化的因素是多样的。在现有长期水文气象预报方法中,经验分析、数理统计、聚类分析等较为常见。这些方法对长期预报能力的提升有指导性意义,但也存在建模过程复杂、主观性强等问题。聚类预报模型将现有数理统计模型与聚类分析模型相结合,先根据数据特征将数据进行聚类预处理,再利用数理统计方法对数据进一步加工和优化,目前该方法已在国内多个流域长期预报中加以应用[1-3]。多年的实践应用表明,该模型在预测效果和建模效率上还有一定提升空间。本文针对预报对象计算、建模因子筛选规则和建模效率等方面进行进一步优化和改进,建模效率和预测效果有一定提高,对开展水文气象长期预报业务具有一定推广和应用参考价值。

2 数据与方法

以长江上游为示范区,选择示范区内78个国家气象站,采用动态泰森多边形法计算流域面雨量,统计1971~2019年6~8月逐月面雨量序列,并将此数据序列作为建模的预报对象输入,选取北半球100 hPa高度场(以下简称100 hPa),北半球500 hPa高度场(以下简称500 hPa),北半球场面气压场(以下简称SLP),太平洋海温场(以下简称SST),全国18区降水、18区温度及130项环流指数7类气候因子资料(数据来自国家气候中心)作为建模的预报因子输入,构建聚类预报模型,再对2020~2022年长江上游面雨量进行模型检验[4]。

本文预报对象为长江上游面雨量,由于金沙江上游站点分布较稀疏,岷沱江下游、乌江上游站点分布较密集,故采用动态泰森多边形法计算流域面雨量以增强预报对象值的代表性。动态指当站点存在缺测数据或站点搬迁,计算缺测时间段面雨量时排除该站点,但其他时间该站点参与计算,这使得可利用的站点范围大大增加;同时采用泰森多边形法,每个站点的权重大小为其疏密程度的反映,考虑了站点的分布情况。

长期面雨量预报主要关注预报与均值相比偏多或偏少和量级,距平符号可表示偏多还是偏少,距平值与均值的百分率可表示量级,符号一致率越高且误差百分率越低说明模型的准确性越高[5]。因此,本文采用距平符号一致率及距平误差百分率衡量模型改进前后的准确性,作为聚类

收稿日期:2023-02-03,**修回日期**:2023-03-21

基金项目:国家自然科学基金青年基金项目(52009012);国家重点研发计划(2021YFC3200301)

作者简介:顾丽(1990-),女,工程师,研究方向为水文信息化,E-mail:1393116725@qq.com

通讯作者:陈瑜彬(1981-),男,教授级高级工程师,研究方向为预报调度一体化,E-mail:chenyb@cjh.com.cn

预报模型的检验方法。

3 聚类模型的构建

3.1 模型构建流程

聚类预报模型建模过程见图 1。其中前 10 步为模型构建，后 2 步为模型检验。

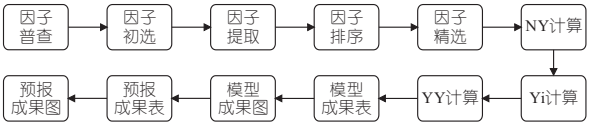

图 1 聚类预报模型流程图

Fig. 1 Flow chart of the clustering model

步骤 1 因子普查即计算预报对象与预报因子的相关性。

步骤 2 因子初选即将因子普查中达标的格点(分区、项)挑选出来作为初选因子。

步骤 3 因子提取指提取初选因子及预报对象的数据资料并将相关系数等计算展示。

步骤 4 因子排序指将初选因子历年观测值根据相关系数排序。

步骤 5 因子精选即在因子排序的基础上根据精选规则进一步筛选预报因子。

步骤 6 NY 计算指预报因子值正贡献指数与预报对象值进行一元回归分析得到一个 NY 方程。

步骤 7 Yi 计算指将各类预报因子值与预报对象值进行多元一次回归，有几类因子就得到几个 Yi 方程，不同因子组合后为一类因子。

步骤 8 YY 计算指将各类预报因子值代入多元一次回归方程得到的 y 值当作 x 值，预报对象值当做 y 值进行二次回归，得到一个 YY 方程。

步骤 9 模型成果表即将最终因子序号 NY/Yi/YY 方程系数的计算过程汇总到一个表，此表为预报的基础。

步骤 10 模型成果图指以历年 NY 值为 x 轴，YY 值为 y 轴，建立二维坐标系，展示预报对象实际距平值及对应年份信息，此图为模型好坏判断依据。

步骤 11 预报成果表指将预报时间对应最终因子值代入 NY 方程得到 NY 值，代入 Yi 方程得到的值当做 x 代入 YY 方程，得到 YY 值，NY 与 YY 的平均值即为预报结果。

步骤 12 预报成果图是在模型成果图的基础上标出预报年份对应的 NY/YY 值及对应距平值和年份信息，至此预报结束。

本文重点针对预报准确性和建模效率两个方面进行模型优化。

3.2 改进因子初选规则

因子初选时，在改进前模型对 3 类非场面因子和 4 类场面因子计算相关系数，查看预报因子与预报对象是否线性相关并以此挑选预报因子。相关系数 r 的计算公式为：

$$r=\frac{\sum_{t=1}^{n}(y_t-\bar{y})(x_t-\bar{x})}{\sqrt{\sum_{t=1}^{n}(y_t-\bar{y})^2\sum_{t=1}^{n}(x_t-\bar{x})^2}} \tag{1}$$

式中，x_t 为预报因子单个格点(分区或项)各年观测值；y_t 为预报对象各年观测值；$\bar{x}$、$\bar{y}$ 分别为 x_t、y_t 的均值。

改进后，模型除了对 3 类非场面因子计算相关系数外，增加了相关概率计算，只有因子的相关系数和相关概率同时大于阈值才能被挑选为初选因子。相关概率 p 的计算公式为：

$$p=\sum_{t=1}^{n}\frac{f(x)}{n} \tag{2}$$

其中 $f(x)=\begin{cases}1 & (y_t-\bar{y})(x_t-\bar{x})>0\\ 0 & (y_t-\bar{y})(x_t-\bar{x})\leqslant 0\end{cases}$

p 的范围在 0～1 之间，当 p 接近 0 时，说明预报因子距平与预报对象距平同号可能性较小，当 p 接近 1 时，说明预报因子距平与预报对象距平同号可能性较大。

3.3 改进因子精选规则

因子精选时，被挑中的因子需满足 3 个条件：①符号分离，排序后的因子中最佳界值以上预报对象正距平较多，最佳界值以下负距平较多；②大小分离，预报对象正距平中较大值大多在最佳界值以上，负距平中较小值大多在最佳界值以下；③时间分离，最近几年的预报对象距平值，若为正大多在最佳界值上，若为负大多在最佳界值下。若预报因子不能同时满足 3 个条件无法挑选最佳界值，则为假相关因子，因子精选时应去掉。计算流程见图 2。图 2 中，p_{t1} 为排序后 t 行以上预报对象正距平个数与 $t-1$ 的比值；p_{t2} 为 t 行以下负距平个数与 $T-t$ 的比值；T 为某个因子的总行数，即参与建模的年份数；n_{t1} 为 t 行以上预报对象极大正距平个数；n_{t2} 为 t 行以下极小负距平个数；m_{t1} 为 t 行以上最近年份预报对象正距平个数；m_{t2} 为 t 行以下最近年份预报对象负距平个数；当 P、N、M 值固定时，得到的因子精选结果惟一，且规则的满足程度最优。

3.4 改进建模效率

采用数据包括预报对象和预报因子值，预报

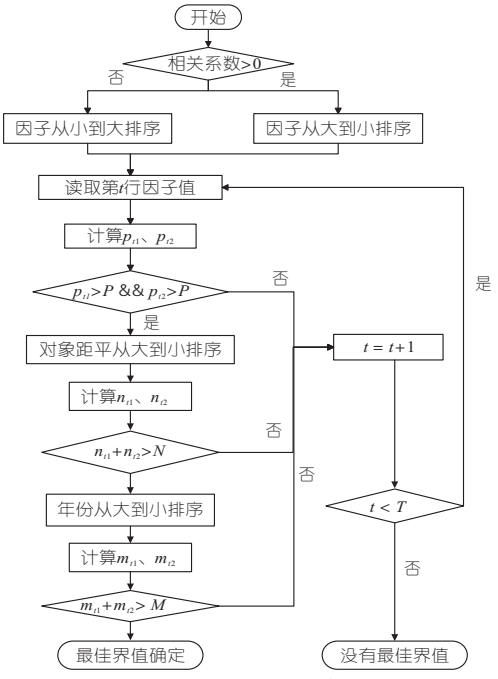

图 2　因子精选规则计算流程图

Fig. 2　Calculation process of factor selection rule

对象为长江上游面雨量,预报因子包括 4 类场面因子和 3 类非场面因子。本文建立了一系列数据库表,用于存储预报对象、预报因子值的源数据及场面预报因子中格点的地理位置,新建表的含义见表 1。

表 1　模型新建表及含义

Tab. 1　Model creation table and its meaning

表名	含义
LT_ST_RAIN	长江上游站点逐日雨量
LT_ST_WEIGHT	长江上游站点逐日权重
LT_SUB_RAIN	小分区逐日面雨量值
LT_SUB_WEIGHT	小分区面积权重
TQ_SUB135_G	全国站点分区对应关系
TQ_Rain160_G	全国站点逐日雨量
TQ_Temper160_G	全国站点逐日温度
TQ_Circu_G	100、500 hPa 格点逐日值
TQ_GRID_100PA	100 hPa 所有格点地理位置
TQ_GRID_500PA	500 hPa 所有格点地理位置
TQ_SLP_G	SLP 格点逐月值
TQ_GRID_SLP	SLP 所有格点地理位置
TQ_SST_G	SST 格点逐月值
TQ_GRID_SST	SST 所有格点地理位置
TQ_CIR130_G	130 项逐月值
TQ_GRID_GIR130	130 项每项位置

本文将参与计算的源数据入库,建模时直接计算并供模型调用,在确保一致性的同时也提高了建模效率。场面预报因子中格点的地理位置是因子初选的一个重要因素,需要 3 个及以上格点达标且成片才能被挑选为初选因子。改进前这项操作由人工筛选,将因子的位置信息入库后模型可自动筛选达标因子,将繁琐的工作简化,提高了建模过程效率。

3.5　聚类预报模型方程率定

分别以 1971～2019 年长江上游 6、7、8 月面雨量为预报对象,七类气象要素作为预报因子建模。建模得到的 YY 方程依次为:

$$y=-56.211+0.480x_1+0.351x_2+0.411x_3+0.154x_4 \quad (3)$$

$$y=-65.256+0.385x_1+0.351x_2-0.253x_3+0.276x_4+0.638x_5 \quad (4)$$

$$y=-127.947+0.121x_1+0.133x_2+0.658x_3+0.562x_4+0.434x_5 \quad (5)$$

式(3)将 18 区雨量与 SLP 合并共同作为 x_1,SST 与 500 hPa 合并共同作为 x_3,18 区温度没有被选中的因子;式(4)将 100 hPa 与 SLP 合并共同作为 x_2,18 区温度同样没有被选中的因子;式(5)将 18 区温度与 18 区雨量合并共同作为 x_1,SST 与 100 hPa 合并共同作为 x_3。

将被挑选因子 1971～2019 年对应月份的值依次代入 NY、Yi、YY 方程,得到一系列模型计算值;再将被挑选因子 2020～2022 年对应月份的值计算并代入方程,得到 2020～2022 年主汛期 3 个月的长江上游面雨量预报值。

4　模型应用分析讨论

以长江上游主汛期 6、7、8 月的建模及预报情况说明模型改进前后的改变。距平符号一致率 P' 和距平误差百分率 S 的计算公式分别为:

$$P'=\sum_{t=1}^{n}\frac{f(x)}{n} \quad (6)$$

$$S=\frac{1}{n}\sum_{t=1}^{n}\left[\mathrm{abs}\left(\frac{(b_t-a_t)}{\bar{a}}\right)\times 100\right] \quad (7)$$

其中　$f(x)=\begin{cases}1 & (a_t-\bar{a})(b_t-\bar{a})>0\\ 0 & (a_t-\bar{a})(b_t-\bar{a})\leqslant 0\end{cases}$

式中,a_t 为实际值;b_t 为模型值;$\bar{a}$ 为实际值均值。

计算后改进前后的准确性对比见表 2。由表 2 可知,模型经过改进后,在距平符号一致率上有所提升,在距平误差百分率上有所下降,这表明准确性有所提升。

表 2　改进前后准确性对比

Tab. 2　Comparison of accuracy before and after improvement

月份	距平符号一致率		距平误差百分率/%	
	改进前	改进后	改进前	改进后
6	0.79	0.87	9.06	4.60
7	0.69	0.94	12.03	5.24
8	0.79	0.90	13.75	11.88
平均	0.76	0.90	11.51	7.24

6、7、8 月模型计算值与实际值对比见图 3。由图 3 可知，除 2020 年 6 月外，其他模型计算值的趋势与实际值基本一致，说明模型对未来变化趋势预测的准确性较高；当实际值与均值较接近时(2020 年 8 月、2021 年 6～8 月，2022 年 6 月)，预报值与实际值不仅距平符号一致，且在量级上也较接近；当实际值极大或极小时(2020 年 7 月，2022 年 7、8 月)，预报值与实际值虽然距平符号一致，但在量级上差异较大，说明模型对极端气候预测的准确性还需提升。

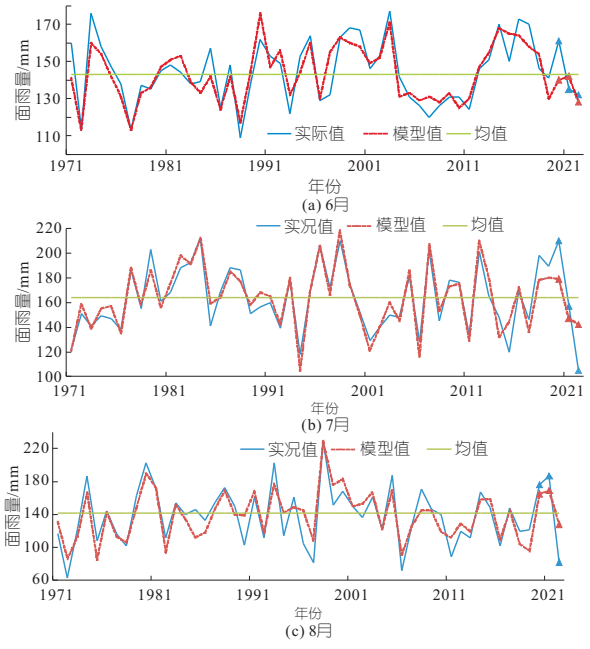

图 3 长江上游 6～8 月模型计算与实际值对比

Fig. 3 Comparison of model value and actual value in the upper reaches of Yangtze River in June, July and August

5 结论

a. 预报对象和预报因子由手工计算升级为模型自动计算，因子筛选由人工筛选改为自动筛选，提高了建模效率。

b. 改进后的模型在距平符号一致率和距平误差百分率上均有所提高，模型的准确性增强。

c. 当实际值极大或极小时，模型预报值与实际值的量级相差较大；由于方程的建立以局部规律的独立性假定为基础，当作为长期预测时，验证期应用结果总体较建模率定期差，这是模型仍需改进的地方。

参考文献：

[1] 李福威，包爱美，疏杏胜，等. 基于水文—气象因子的综合多模型长期径流预报研究[J]. 中国农村水利水电，2022(11)：6-12.

[2] 程忠良，刘勇，高成，等. 基于马氏距离判别的丹江口水库长期径流分级预报[J]. 中国农村水利水电，2018(7)：1-4.

[3] 雷晓辉，王浩，廖卫红，等. 变化环境下气象水文预报研究进展[J]. 水利学报，2018，49(1)：9-18.

[4] HU Y J，ZHONG Z，ZHU Y M，et al. A statistical forecast model using the time-scale decomposition technique to predict rainfall during flood period over the middle and lower reaches of the Yangtze River Valley [J]. Theoretical and applied climatology，2018，132：479-489.

[5] 吴旭树，王兆礼，陈柯兵，等. 基于大气环流和海温场的降水组合预报模型[J]. 水资源保护，2022，38(6)：81-87.

Construction and Improvement of Clustering Model for Long-term Areal Rainfall Forecasting in the Upper Reaches of the Yangtze River

GU Li，CHEN Yu-bin，LI Chun-long

(Bureau of Hydrology，Changjiang Water Resources Commission of the Ministry of Water Resources，Wuhan 430010，China)

Abstract：The cluster forecasting model is a long-term forecasting physical model based on mathematical statistics. This method has been applied in the long-term forecast of the upper reaches of the Yangtze River. After years of practice，there is space for improvement in accuracy and efficiency of the model. This paper selected the upper reaches of the Yangtze River as the demonstration area for forecast，the monthly areal rainfall from June to August as the forecast object，and used the coincidence rate of anomaly symbol and the percentage of anomaly error as the test methods of the model. And then the construction process and improvement of the model were described. Finally，the improved model was used to forecast the monthly area rainfall in the upper reaches of the Yangtze River from June to August during 2020 to 2022. The prediction results show that the improved model has a great improvement in the anomaly sign agreement rate and the percentage of anomaly error，and it has been implemented easily，which has certain reference value.

Key words：forecast；cluster；model；the upper reaches of the Yangtze River；areal rainfall