

# 基于SOM的流域分类和无资料区径流模拟

伊璇<sup>1</sup>, 周丰<sup>2</sup>, 王心宇<sup>1</sup>, 杨永辉<sup>1</sup>, 郭怀成<sup>1</sup>

(1. 北京大学 环境科学与工程学院, 水沙科学教育部重点实验室, 北京 100871;

2. 北京大学 城市与环境学院, 地表过程分析与模拟教育部重点实验室, 北京 100871)

**摘要:**无资料区的径流模拟问题是国内外水文研究的难点之一。基于相似流域的参数移植法是常用的解决方法之一,但如何判断相似流域是制约此类方法发展的难点。本文以滇池流域为例,采用自组织映射神经网络(SOM)和层次聚类分析(HCA)联合模式,选取16个流域物理特征为指标进行子流域分类,以确定相似流域。运用无分层的K-means分类的SOM法将整个滇池流域划分为7类具有水文属性的子流域组,分类情景与HCA基本一致,两者实现相互验证。采用HBV水文模型模拟子流域径流过程,并选择部分子流域进行组内参数移植交叉检验。结果显示,HBV模型可较好的模拟滇池流域径流过程;此外,子流域交叉检验结果优良,表明同组内参数可以相互移植。本文不仅为解决滇池流域无资料问题提供了可靠手段,而且由于SOM实现了高维流域特征可视化展示,有助于管理者全面、深入的把握滇池流域水文属性的空间分布特征,为进行水资源管理提供指导。

**关键词:**无资料地区;径流模拟;流域分类;自组织映射神经网络;HBV模型;滇池流域

doi: 10.11820/dlkxjz.2014.08.011

中图分类号:TV 212.4

文献标识码:A

## 1 引言

无资料或稀缺资料地区的径流模拟问题,一直是国内外水文学界的难点之一(Sivapalan et al, 2003)。中国拥有3400多个国家级水文站点,但相对于960万km<sup>2</sup>的国土面积而言,站点数量偏少,且站点分布不均(谈戈等, 2004)。世界气象组织的数据显示,全球每个水文站点的控制面积约为15~10000 km<sup>2</sup>(柴晓玲, 2005),难以满足研究需要,加上气候变化和人类活动的干扰,使得许多历史数据无法适用于现况。针对这种情况,国际水文科学协会(International Association of Hydrological Sciences, IHAS)在2003-2012年开展了无测站流域水文预测(Prediction in Ungauged Basins, PUB)计划(Sivapalan et al, 2003),旨在用10年时间加强对无资料地区水文预报问题的研究。

观测资料的不完整性导致无法采用常规方法率定水文模型参数。为解决该问题,科学家们作

了很多尝试。学者们应用过的方法主要有比拟法(靳晓莉等, 2008)、插值法(Guo et al, 2001; Merz, 2004)、回归法(Xu, 2003; Lamb et al, 2004; Wagener et al, 2006)和聚类法(McIntyre et al, 2005; He et al, 2011; Kileshye Onema et al, 2012)等。刘苏峡等(2010)又将比拟法和插值法归纳为移植法、将回归法归为替代法的一种。其中,比拟法和插值法需借用邻近流域流量数据,对流域的地理位置和空间变化幅度有一定限制(Blöschl, 2005)。回归法试图建立流域属性特征值与模型参数回归关系,在实际应用上并不可取。通过流域物理特征判断相似流域,实现参数区域化的聚类法逐渐成为解决无资料问题的热点之一。其原理是流域的水文行为由地形和气候特征决定(Oudin et al, 2010),因此,当流域在物理属性上相似时,水文模型参数是可以相互传递的。

但是,过去的研究多集中于一条河流干支流或上下游河段的资料缺失问题上,针对具有复杂河网

收稿日期:2014-01;修订日期:2014-04。

基金项目:国家水体污染控制与治理科技重大专项项目(2013ZX07102-006)。

作者简介:伊璇(1989-),女,山东泰安人,博士生,主要研究方向为水资源管理与水污染控制,E-mail: xuanyi.eva@gmail.com。

通讯作者:郭怀成(1953-),男,北京人,教授,博导,主要从事环境规划、流域管理等方面研究,E-mail: hcguo@pku.edu.cn。

结构的湖泊流域的无资料径流模拟研究还较少,且缺乏普适性。究其原因,一是大尺度的湖泊流域水文过程及影响因素复杂,数据解译及模式识别困难;二是流域地貌特征迥异,典型指标及结论普适性差。滇池是云南省最大的高原湖泊,大量氮磷营养盐通过流域内河流携带流入湖体,致使湖泊富营养化严重。径流模拟对该流域的水资源和水环境的有效管理非常重要,但流量数据的严重缺乏限制了模拟研究。

鉴于此,本文以滇池流域为例,采用自组织映射神经网络(Self-organizing Maps, SOM)和层次聚类法(Hierarchical Cluster Analysis, HCA)进行流域相似性分类,以实现参数区域化,尔后,采用HBV水文模型进行参数移植检验,尝试解决流域内的无资料问题。基于K-means聚类的SOM法不仅是一种分类算法,还可以将高维数据进行可视化展示(Yang et al, 2012),实现各属性的空间特异性分析。研究结果一方面有助于管理者全面、深入地把脉滇池流域水文属性的空间分布特征;另一方面,为解决当地无资料问题提供了有效手段,有助于提高流域水文过程、营养盐负荷模拟结果的有效性,为控制滇池富营养化和开展流域管理提供科学支撑。

2 研究区域与方法

2.1 研究区域

滇池流域(24°28'~25°28'N, 102°29'~103°01'E)位于中国西南云贵高原中部,海拔为1880.6~2837.8 m。流域总面积为2920 km<sup>2</sup>,湖体面积达309 km<sup>2</sup>。流域内主要地形有山地、丘陵和平原,其中山地和丘陵占整个流域的70%左右;主要土地利用类型为林地、建设用地、草地、耕地和水域等,其中建设用地集中在滇池北部—昆明市一带,林地集中在海拔较高的山区;主要土壤类型为粘土,占整个区域的70.51%,其次是沙土和壤土。

滇池流域位于亚热带湿润地区,气候温和,多年平均气温为14.7℃,年平均降雨量为1035 mm,主要分布在5-10月。年均相对湿度为73%~75%,年潜在蒸发量为1205.6 mm。流域内共有30条河流流入,年均入流量为6.6亿m<sup>3</sup>,1条河流流出湖体,年均出流量为2.5亿m<sup>3</sup>。

本文选取滇池流域的43个子流域作为研究对象(图1),子流域控制面积从几平方千米到几百平方千米不等,共占整个流域的61%。由于湖滨带地区没有形成汇流,因此不列入本文研究范围。

2.2 研究方法

本文的研究步骤可分为三步:一是通过ArcMap等软件提取43个子流域气候和地形属性特征值,并对提取的数据进行标准化,消除指标间由于单位和度量差别造成的影响;二是采用SOM和HCA法,将43个子流域基于水文相似性进行分类;三是HBV水文模型交叉验证,选取相似流域进行组内参数组移植,检验区域化结果的可靠性(图2)。

2.2.1 流域物理特征提取

一个流域的产汇流过程取决于流域内的气象和地形特征。在Yadav等(2007)研究的基础上,选取以下流域特征指标,即地形(面积、流域长度、流域坡度、河网密度、形状系数)、地势(平均海拔)、气候(多年平均降雨量)、土地利用类型(耕地、林地、草地、水域、建筑用地、未利用地)和土壤类型(沙土、粘土、壤土)进行研究(表1)。其中,子流域中面积、长度、坡度等地形特征相差较大,林地、农业用地和建筑用地是该区域最常见的土地利用类型。此外,流域主要土壤类型为粘土。

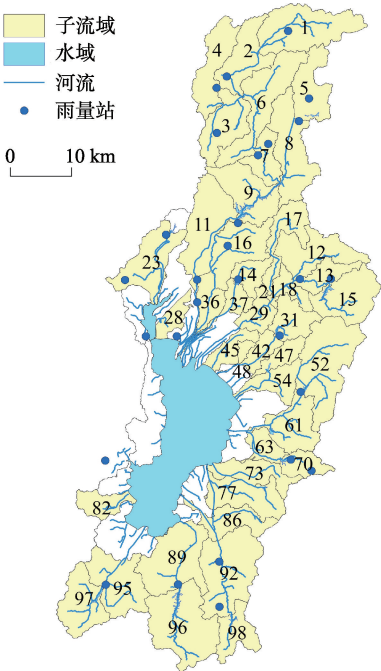


图1 滇池流域河网、观测站点及研究区域分布  
Fig.1 River network, rainfall stations and research area in Lake Dianchi Basin

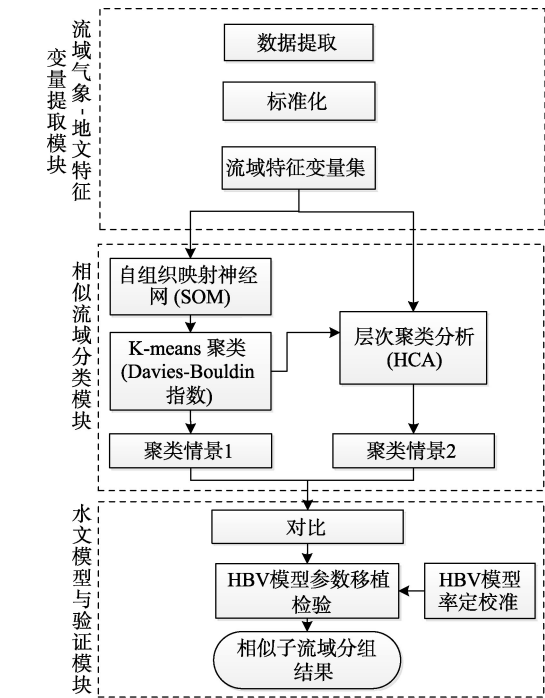


图2 研究流程图

Fig.2 Flow-progress diagram of research

2.2.2 相似流域分类情景判断

基于流域特征变量,采用SOM法对流域进行分类。将分类结果与HCA的结果进行对比,以达到相互验证的目的。

SOM是一种聚类和高维可视化的无监督学习算法,为芬兰学者Kohonen从人脑对信号的处理特

点上获得启示而提出的。该算法将输入的高维空间数据经过非线性映射,转换到低维平面阵列(常用六边形网格表示),具有“无监督”和概率保持等特征(Kohonen, 1982)。其中,Kohonen六边形网格的个数由公式 $N=5\sqrt{M}$  ( $M$ 为样本个数)确定(Davies et al, 1979; Astel et al, 2007)。本文选用基于无分层的K-means分类的SOM法进行子流域分类,Davis-Bouldin指数(Davies et al, 1979)用于确定最优分类组数。计算平台为MATLAB 2011 SOM Toolbox(Ver.2.0)工具箱(<http://www.cis.hut.fi/som-toolbox>),计算程序重复运行1000次,以获得稳定的分类结果。

2.2.3 HBV水文模型验证

为了检验上述分类方法在解决无资料地区水文模拟问题中的有效性,选用HBV模型进行径流模拟以验证流域分类结果的准确性。HBV模型(Bergström et al, 1995)是由瑞典气象和水文研究所(SMHI)开发的一种概念流域水文模型,其输入变量为日降雨量、气温和月潜在蒸发量,输出数据为日径流量。模型包含4个模块:融雪和积雪模块、土壤湿度和有效降雨模块、蒸发散模块及径流响应模块。本文采用的模型版本为HBV light version 1.3 (Seibert, 1998)。

采用Nash-Sutcliffe效率系数(NSE)和决定系数( $R^2$ )来评价模型拟合结果:

表1 43个子流域的流域特征数据

Tab.1 Physical properties of 43 sub-basins

类型	特征变量	来源	平均值	标准差	最大值	最小值
地形	面积/km <sup>2</sup>	50 m×50 m DEM, ArcGIS 9.3 处理	41.42	31.93	115.13	1.66
	流域长度/m		9731.34	5850.99	20993.73	120.71
	流域坡度/(m/km)		10.42	5.08	20.07	0.31
	河网密度/(m/km <sup>2</sup> )		818.64	1566.33	7545.00	2.18
	形状系数		2.27	0.35	3.46	1.81
地势	平均海拔/m	50 m×50 m DEM, ArcGIS 9.3 处理	2074.16	124.88	2409.56	1885.16
气候	多年平均降雨量/mm	22个降雨站(1999-2000年日值数据)	921.11	123.35	1581.69	786.29
土地类型	农业用地/%	土地利用.shp 文件(1:50000)	20.86	10.56	45.02	4.47
	林地/%		50.39	23.62	84.78	1.07
	草地/%		2.54	2.56	11.30	0.13
	水域/%		0.72	0.92	4.69	0.00
	建筑用地/%		21.40	22.59	92.68	1.12
土壤类型	未利用地/%	土壤分布.shp 文件(1:500000)	4.08	2.78	11.73	1.20
	沙土/%		26.27	22.80	89.32	0.00
	粘土/%		72.02	23.65	100.00	10.09
	壤土/%		1.71	5.38	24.61	0.00



$$NSE = 1 - \frac{\sum_{i=1}^n (Q_{obs,i} - Q_{sim,i})^2}{\sum_{i=1}^n (Q_{obs,i} - \bar{Q}_{obs})^2} \tag{1}$$

$$R^2 = \frac{\sum_{i=1}^n (Q_{obs,i} - \bar{Q}_{obs})(Q_{sim,i} - \bar{Q}_{sim})}{\sqrt{\sum_{i=1}^n (Q_{obs,i} - \bar{Q}_{obs})^2} \sqrt{\sum_{i=1}^n (Q_{sim,i} - \bar{Q}_{sim})^2}} \tag{2}$$

式中： $Q_{obs,i}$  和  $Q_{sim,i}$  分别为第*i*个时刻的观测和模拟径流； $\bar{Q}_{obs}$  和  $\bar{Q}_{sim}$  分别为时段内的平均值；*n* 为时间步长。其中，*NSE* 和 *R*<sup>2</sup> 取值范围为 0~1，越接近 1 则说明拟合结果越好。

3 结果与分析

3.1 相似流域分类情景分析

基于 43 个子流域的气候和地形属性特征变量，判断相似流域。HCA 法将 43 个子流域最明显地分为两类(图 3)。其中 A 组的子流域集中在流域下游平缓地区，人口相对密集，具有城市型河流的特征，即河道两侧多为硬质护坡，河流分支较少，受到人为干扰较大；B 组子流域群则位于滇池流域上游山地丘陵地区，坡度较大，植被较多，受人为干扰较少。

使用 SOM 法将 43×16 个高维特征变量投影于 33 个六边形组成的拓扑网格中，基于 Davis-Bouldin

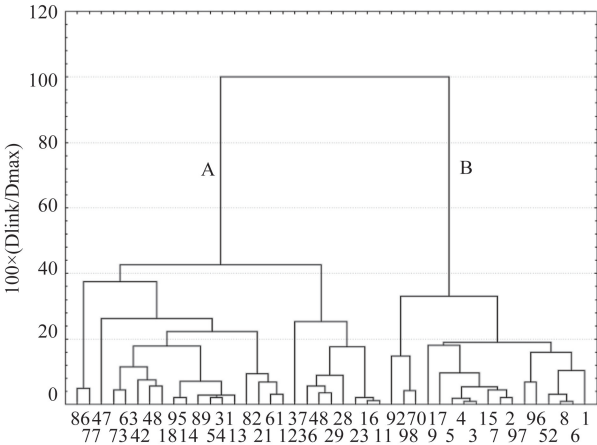


图3 HCA 聚类情景  
Fig.3 Result of Hierarchical Cluster Analysis (HCA)

指数对 33 个网格进行进一步分类。通过 1000 次重复计算分析，分 7 组的情况与其他分组数相比，发生次数最高，因此，7 组是最适合的分组数目(图 4a)，因此，将 43 个子流域分为 7 组(I-VII)。图 4b 为分组结果，其中，六边形中数字代表子流域的编号，编号在图形中距离越近，则认为流域越具有水文相似性；用不同颜色区分组别，罗马数字代表相应分组编号。其中，组 I、IV 和 VII 在整个流域中所占比例最大。

从图 3 和图 4 上看，HCA 和 SOM 两种分类结果基本一致。仅子流域 21 和 42 号，在 SOM 分类归为

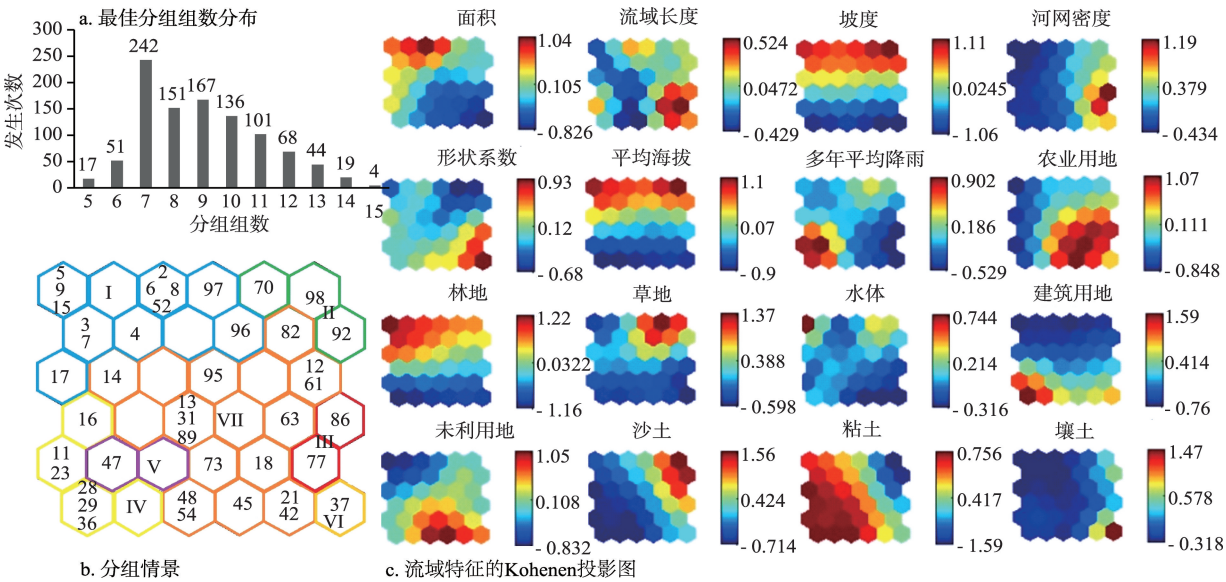


图4 SOM 结果  
Fig.4 Results of self-organizing maps (SOM)

第Ⅶ组,HCA归为第Ⅶ组。这是由于这两个子流域位于组Ⅵ和Ⅶ的交界处,容易被不同的方法分到不同的组中。

通过分析 Kohenen 图(图 4c),可以发现 7 组子流域是具有特定水文属性的(表 2)。其中,组Ⅰ位于海拔较高的林地地区,河网密度较低,形状指数偏低,表明该处降水容易迅速汇集,加之土壤为粘土,透水性弱,易发生洪水,在流域管理上应该重视防洪。组Ⅳ、Ⅴ特征差异较小,都处于海拔较低,坡度较小的地区,从土地利用类型来看,主要为建筑用地。该地河道受人为干扰较大,河道两岸常为直立堤坝。组Ⅶ处于海拔较低的地区,在滇池流域中分布较广。该区域形状指数较高,因而径流变化缓和,土地利用类型以农业用地为主,径流过程受灌溉的影响较大,为滇池面源污染的主要源头。图 5 为 7 组子流域在整个滇池流域的分布情况。

3.2 日径流模拟及参数移植检验

为了进一步检验该分组情景确实具有水文相似性特征,可以解决无资料问题,选取组Ⅰ、Ⅳ和Ⅶ中 6 个流量站(图 5)进行 HBV 模型模拟和参数移植交叉验证。HBV 模型在模拟流量时, $R^2$ 和  $NSE$ 结果可参见表 3。其中拟合最差的为大清河站( $R^2=0.718$ ,  $NSE=0.495$ ),最好的为干海子站( $R^2=0.840$ ,  $NSE=0.638$ )。以干海子站为例,绘制 HBV 模型模拟和实测的流量过程曲线(图 6)。从图 6 可以看出,模拟过程线与实测过程线拟合程度较好,说明 HBV 模型可以较好地捕捉到流量波动过程及峰值。

参数移植交叉检验结果见表 3。 $R^2$ 和  $NSE$ 代表贡献组参数移植到目标组模型的拟合度检验结果,表中数据为模型率定结果。虽然有些组别的  $R^2$ 和  $NSE$ 偏小,但也满足了拟合度要求。如最小的组Ⅲ(昆明站移植到大清河站)的  $R^2$ 和  $NSE$ 分别为 0.654

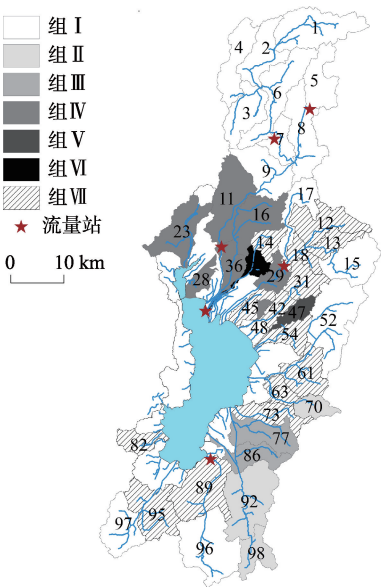


图 5 子流域聚类情景  
Fig.5 Sub-basins classification scenario

和 0.472,拟合度检验结果也达到了要求。该组检验结果不够理想的原因是大清河流量较小,波动较大,HBV 模型直接进行模拟时结果受到影响,因此参数移植的结果也相应较差。移植效果最好的为组Ⅶ(干海子站向柴河站移植), $R^2$ 为 0.839, $NSE$ 为 0.636。从图 7 可以看出,虽然这两个测站分别位于不同河流,距离相隔较远,但是两者参数移植效果较佳,表明聚类方法对于解决无资料问题具有可行性。

可见,虽然滇池流域径流量不大、且波动较大,但 HBV 模型仍可以得到较好的模拟结果,有效地捕捉到径流丰水期和枯水期的波动变化。交叉检验的结果证明 SOM 法分类方式可以有效地判断水文特征相似的流域,同组子流域间不仅具有相似的流域特征,同时具有相似的水文行为。因此,无

表 2 7 组子流域水文特征分布  
Tab.2 Characteristics of the seven sub-basins

分组编号	面积	流域长	坡度	河网密度	形状指数	海拔	降雨	土地利用	土壤
I	高	偏低	偏大	低	偏低	高	偏低	林地	粘土
II	中	中	大	中	低	高	中	草地	沙土
III	低	高	偏小	高	高	偏低	低	农业	粘土
IV	偏高	偏低	小	低	偏低	低	高	建筑用地	粘土
V	偏低	低	小	低	中	低	偏高	建筑用地	粘土
VI	低	高	小	偏高	高	低	低	未利用地	壤土
VII	偏低	偏低	中	偏低	偏高	偏低	偏低	农业	沙土

注:其中“高”,“偏高”,“中”,“偏低”和“低”代表该组变量值在所有流域中的分布情况;土地利用和土壤为该组流域主要类型。

表3 模型率定和组内交叉检验

Tab.3 Calibration and cross-validation within groups

目标组\贡献组		中和站	白邑站	昆明站	大清河站	干海子站	柴河站
中和站	$R^2$	0.823	0.779	—	—	—	—
	$NSE$	0.616	0.556	—	—	—	—
白邑站	$R^2$	0.821	0.835	—	—	—	—
	$NSE$	0.532	0.611	—	—	—	—
昆明站	$R^2$	—	—	0.734	0.726	—	—
	$NSE$	—	—	0.524	0.424	—	—
大清河站	$R^2$	—	—	0.654	0.718	—	—
	$NSE$	—	—	0.472	0.495	—	—
干海子站	$R^2$	—	—	—	—	0.840	0.809
	$NSE$	—	—	—	—	0.638	0.561
柴河站	$R^2$	—	—	—	—	0.839	0.812
	$NSE$	—	—	—	—	0.636	0.631

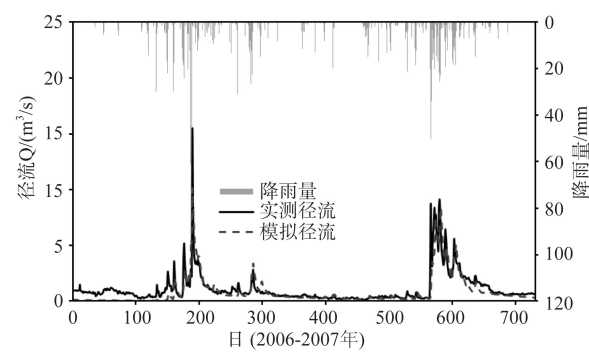


图6 干海子站HBV模型率定与验证结果  
Fig.6 HBV model calibration and validation result in Ganhaizi Station

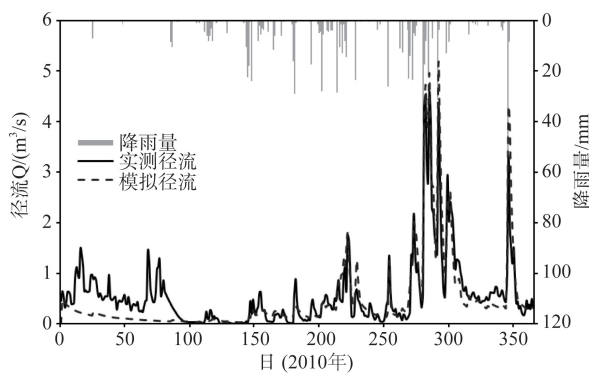


图7 柴河站HBV模型验证结果  
Fig.7 HBV model validation result in Chaihe Station

资料流域可以移植有资料流域的参数,这为解决滇池流域径流监测数据缺失问题提供了有效方法。

4 结语

本文选取地形、地势、气候、土地利用类型和土壤类型等属性特征,基于SOM和HCA两种聚类方法,将滇池流域43个子流域分成7个相似流域组,并利用HBV模型验证该区域化方法的可靠性。两种方法均将滇池流域分为7个相似子流域组,各组间具有不同的水文属性,而组内的参数可以相互传递。本文将SOM法用于具有复杂河网结构的湖泊流域,一方面解决了滇池流域监测资料严重缺失的问题,另一方面,流域的水文分类(分区)为今后进行水资源和水质管理提供了指导。此外,确定相似组进行参数区域化的研究,为今后快速模拟流域内各子流域径流提供了技术支持。

参考文献(References)

柴晓玲. 2005. 无资料地区水文分析与计算研究[D]. 武汉: 武汉大学. [Chai X L. 2005. Hydrologic estimation methods and analysis in un-gauged basins[D]. Wuhan, China: Wuhan University.]  
靳晓莉, 张奇, 许崇育. 2008. 一个概念性水文模型的参数区域化研究: 以东江流域为例. 湖泊科学, 20(6): 723-732.

- [Jin X L, Zhang Q, Xu C Y. 2008. Regionalization study of a conceptual hydrological model in the Donjiang Basin. *Journal of Lake Sciences*, 20(6): 723-732.]
- 刘苏峡, 刘昌明, 赵卫民. 2010. 无测站流域水文预测(PUB)的研究方法. *地理科学进展*, 29(11): 1333-1339. [Liu S X, Liu C M, Zhao W M. 2010. Towards the methodology for predictions in ungauged basins. *Process in Geography*, 29(11): 1333-1339.]
- 谈戈, 夏军, 李新. 2004. 无资料地区水文预报研究的方法与出路. *冰川冻土*, 26(2): 192-196. [Tan G, Xia J, Li X. 2004. Hydrological prediction in ungauged basins. *Journal of Glaciology and Geocryology*, 26(2): 192-196.]
- Astel A, Tsakovski S, Barbieri P, et al. 2007. Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. *Water Research*, 41(19): 4566-4578.
- Bergström S. 1995. The HBV model//Singh V P. *Computer models of watershed hydrology*. Highlands Ranch, CO: Water Resources Publications: 443-476.
- Blöschl G. 2005. Rainfall-runoff modeling of ungauged catchments//Anderson M G. *Encyclopedia of hydrological sciences*. Chichester, UK: John Wiley & Sons: 2061-2080.
- Davies D L, Bouldin D W. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2): 224-227.
- Guo S, Wang J, Yang J. 2001. A semi-distributed hydrological model and its application in a macroscale basin in China. Gothenburg, Sweden: IAHS Publication: 167-174.
- He Y, Bárdossy A, Zehe E. 2011. A catchment classification scheme using local variance reduction method. *Journal of Hydrology*, 411(1-2): 140-154.
- Killeshye Onema J M, Taigbenu A E, Ndiritu J. 2012. Classification and flow prediction in a data-scarce watershed of the equatorial Nile region. *Hydrology and Earth System Sciences*, 16(5): 1435-1443.
- Kohonen T. 1982. Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43(1): 59-69.
- Lamb R, Kay A L. 2004. Confidence intervals for a spatially generalized, continuous simulation flood frequency model for Great Britain. *Water Resources Research*, 40(7): doi: 10.1029/2003WR002428.
- McIntyre N, Lee H, Wheater H, et al. 2005. Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, 41(12): doi: 10.1029/2005WR004289.
- Merz R, Blöschl G. 2004. Regionalisation of catchment model parameters. *Journal of Hydrology*, 287(1): 95-123.
- Oudin L, Kay A, Andreassian V, et al. 2010. Are seemingly physically similar catchments truly hydrologically similar? *Water Resources Research*, 46(11): doi: 10.1029/2009WR008887.
- Seibert J. 1998. HBV light version 1.3, user's guide. Uppsala, Finland: Uppsala University, Dept of Earth Science.
- Sivapalan M, Takeuchi K, Franks S W, et al. 2003. IAHS decade on predictions in ungauged basins(PUB), 2003-2012: shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, 48(6): 857-880.
- Wagener T, Wheater H S. 2006. Parameter estimation and regionalization for continuous rainfall-runoff models including uncertainty. *Journal of Hydrology*, 320(1): 132-154.
- Xu C Y. 2003. Testing the transferability of regression equations derived from small sub-catchments to a large area in central Sweden. *Hydrology and Earth System Sciences*, 7(3): 317-324.
- Yadav M, Wagener T, Gupta H. 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. *Advances in Water Resources*, 30(8): 1756-1774.
- Yang Y H, Wang C Y, Guo H C, et al. 2012. An integrated SOM-based multivariate approach for spatio-temporal patterns identification and source apportionment of pollution in complex river network. *Environmental Pollution*, 168: 71-79.



## Classification and runoff simulation of data-scarce basins based on self-organizing maps

YI Xuan<sup>1</sup>, ZHOU Feng<sup>2</sup>, WANG Xinyu<sup>1</sup>, YANG Yonghui<sup>1</sup>, GUO Huaicheng<sup>1</sup>

(1. Laboratory of Water and Sediment Sciences, Ministry of Education, College of Environmental Sciences and Engineering, Peking University, Beijing 100871, China; 2. Laboratory for Earth Surface Process, Ministry of Education, College of Urban and Environmental Sciences, Peking University, Beijing 100871, China)

**Abstract:** Runoff prediction in ungauged basins (PUB) is one of the difficult research areas in hydrological studies. Parameter replacement using data from similar basins is one of the common methods in dealing with the PUB problem. When basins are similar in physical properties, their hydrological behaviors are assumed to be also similar and thus the hydrological model parameters can be transferred from the donor basin to the target basin. However, it is hard to determine whether a donor basin is indeed similar to a target basin and therefore it is not always clear whether the parameters can be transferred between the basins. Existing research often focus on river basin PUB problem, with inadequate attention on lake basins that contain a number of river streams. This study addresses the PUB question using Lake Dianchi Basin as an example. Lake Dianchi Basin has a complicated river network as well as serious PUB problems. Self-organizing maps (SOM) and hierarchical clustering analysis (HCA) were jointly used to identify analogy basins based on 16 physical attributes, including area, length, slope, drainage density, Ke, mean elevation, average precipitation, six land use types and three soil types. SOM method with K-means cluster was applied to classify similar sub-basins into distinct groups and Davis-Bouldin index was used to determine the optimal group numbers. After 1000 iterations the 43 sub-basins were classified into seven groups (I-VII). This SOM-based classification result is the same as the result of HCA except for two sub-basins. Among the seven groups, group I, IV, and VII contains most of the sub-basins and the other four groups contain no more than three sub-basins each. Different groups have different characteristics and the classification result provides a guidance for local management of the lake basin. For instance, group I is located in high elevation area where the density of streams and infiltration rate of the soil are both low therefore the area is flood-prone, thus the local government should pay more attention on flood control in such area. HBV model was used to simulate the runoff process and for sub-basins where the simulation went well, their parameters were used in the cross-basin test. The cross-basins test was applied to test whether or not hydrological model parameters could be transferred between two sub-basins in the same group. Six stations in three groups were selected as examples and sub-basins in each two sub-basin pair are from the same group. The result shows that the HBV model performs well in the runoff simulation of Lake Dianchi Basin ( $R_2 \geq 0.718$ ,  $NSE \geq 0.495$ ). The cross-basin test result is also very promising ( $R_2 \geq 0.654$  and  $NSE \geq 0.472$ ) — it proves that ungauged sub-basins could borrow the model parameters of gauged basins in the same group. Thus, this research provides a solution for solving the PUB problem in the Lake Dianchi Basin. This research provides a basis for solving the problem of lack of data for runoff modeling for the basin. Meanwhile, SOM visualizes multi-dimensional properties of the basin, which is useful for practitioners in water resource management to comprehensively understand the spatial distribution of hydrological characteristics of Lake Dianchi Basin.

**Key words:** ungauged basin; runoff simulation; catchments classification; self-organizing map; HBV model; Lake Dianchi Basin