

基于 UMAP 与 HDBSCAN 的北京市极端暴雨 时空动态分布规律研究^{*}

徐淑高¹⁾ 王纤阳¹⁾ 蒋卫威²⁾ 鱼京善^{1)†} 刘 源³⁾ 周桂欢¹⁾

(1)北京师范大学水科学研究院, 数字流域实验室, 100875, 北京;

2)中国科学院地球化学研究所, 环境地球化学国家重点实验室, 550081, 贵州贵阳;

3)中国水利水电科学研究院, 100038, 北京)

摘要 极端暴雨具有历时短、雨强大、破坏性强等特点, 是引发城市内涝的主要原因之一, 探究其时空动态分布规律, 有助于提高城市内涝风险精细化管理水平。本文以北京市 2004—2016 年 308 个气象站点的降雨资料为研究样本, 利用 UMAP 降维算法和 HDBSCAN 聚类算法, 构建了各类极端暴雨事件的时空动态分布模型。首次提取了北京市全域 4 类极端暴雨模式: 模式 1, 暴雨集中在主城区, 并围绕主城区缓慢移动; 模式 2, 暴雨从西南山区途经主城区, 向东北方向移动; 模式 3, 暴雨自西部山区向主城区西北部扩散, 最后向主城区北部方向移动; 模式 4, 暴雨从西部山区途经主城区向东移动。研究表明, 北京市极端暴雨的主要类型为短历时降雨过程, 暴雨中心在空间上均存在从西向东移动的趋势。其中, 主城区、东南山区以及南部平原地区面临相对更高的极端暴雨风险。各类极端暴雨的模式特征物理机制明确, 其重构特征可以充分表征实际暴雨事件特性。研究成果可为北京市降雨设计、城市内涝风险管理等工作提供一定的参考。

关键词 极端暴雨; UMAP; HDBSCAN; 动态时空分布; 北京

中图分类号 P458.1+21.1

DOI: 10.12202/j.0476-0301.2022303

0 引言

进入 21 世纪以来, 我国城镇化规模快速扩张, 城镇人口比例从 2000 年的 36% 增长到 2020 年的 64%。高速城镇化在推动社会经济快速发展的同时, 对资源、环境产生了巨大的压力^[1-2]。加之全球气候持续变暖, 极端天气事件发生的频率明显增加^[3]。在此背景下, 我国城市面临的洪涝灾害问题日益严峻, 2006 年以来, 每年受淹城市多达 100~250 座, 大多为局地暴雨内涝^[4], 其中北京“7·21”、郑州“7·20”等极端暴雨袭击特大城市所引发的严重内涝灾情, 再次引起人们对于暴雨内涝的思考^[5]。

城市内涝风险精细化管理水平受限于降雨预报的精度。目前, 降雨预报以数值模拟方法为主, 对暴雨的时空分布进行精准预报仍然十分困难。在时间上, 数值模拟方法对初始条件、同化算法等因素敏感, 导致其在 1 h 内预报能力下降过快, 3 h 后的预报准确性较差; 在空间上, 该方法对大气运动物理的过程细节缺乏考虑, 预报结果的空间分辨率往往偏低^[6], 支撑应急防灾措施具有一定的局限性。为及时启动应

急响应行动, 提高极端暴雨的可预见性, 应更换视角, 从历史观测数据中挖掘极端暴雨事件的时空动态分布规律, 对可能形成极端暴雨的天气形势进行分类识别, 预估降雨时空分布特征与动态移动路径, 为应急响应赢取宝贵的时间, 增强防灾的预见性和有效性, 从而为内涝风险管理提供更多科学依据。

目前, 极端暴雨时空分布的相关研究大致可分为 2 类: 一类是针对极端暴雨的静态要素, 如单站点雨量以及整个降雨事件总雨量的空间分布展开的^[7]; 另一类是针对雨强随时间的变化特征, 如整个降雨事件中雨强的时程分布展开的^[8-10]。这些研究所发现的规律都具有一定的片面性: 前者往往忽略了降雨的时间变异性; 后者则大多只集中在单个站点, 忽略了降雨的空间异质性。在实际的降雨过程中, 降雨的时间、空间特性是由雨团本身物理性质和移动特征共同决定的, 需要对降雨的时空动态变化过程展开研究才能兼顾以上 2 类特性。此外, 也有研究表明, 在雨强相同的情况下, 分别以流域上游和下游区域为起点发展的降雨将产生截然不同的水文效应^[11-12]。在实际应用层面, 对降雨时空性质的精准把握, 也有助于对后续的

^{*} 国家自然科学基金重点资助项目(52239003); 济南市水务科技资助项目(JNSWKJ202001)

[†] 通信作者: 鱼京善(1965—), 男, 教授, 博士。研究方向: 水文水资源。E-mail: jingshan@bnu.edu.cn

收稿日期: 2022-10-13

水文响应进行更加精确的预估,便于防洪减灾等工作的展开.因此,开展极端暴雨的时空动态分布规律研究,具有重要理论意义和实用价值.

随着数据采集和传输技术的进步和完善,降雨数据得到了极大的丰富,应用机器学习算法分析降雨时空特征成为可能.近年来,胡睿等^[13]利用 Ward 聚类分析法对城市暴雨雨型开展研究,将暴雨雨型划分为 5 种类型,并分析了雨型对城市洪涝的影响;刘媛媛等^[14]基于 *K-means* 算法建立了短历时降雨时空特征分析模型,成功提取降雨时空分布规律.然而,在 Ward 与 *K-means* 聚类算法中,聚类中心的选择很大程度地影响聚类效果,难以保证分类结果与实际降雨特征的精准匹配.同时,聚类算法在处理数据量大、高维度的降雨数据时,数据预处理工作必不可少. Liu 等^[15]与 Liu 等^[16]分别利用 PCA(主成分分析)和 LLE(局部线性嵌入)对高维降雨数据进行预处理,并结合聚类算法建立降雨时空分布模型,揭示了长时间序列的降雨时空分布规律.然而极端暴雨数据具有非高斯性,不符合 PCA 假定样本数据服从高斯分布的前提, LLE 非线性降维算法在处理大型高维降雨数据集时,不能很好保持原有高维数据的全局结构.

针对上述问题,本文提出了一种均匀流形逼近与投影(uniform manifold approximation and projection, UMAP)降维算法和 HDBSCAN(hierarchical density-based spatial clustering of applications with noise)聚类算法相结合的极端暴雨时空动态特征分析方法.其中:UMAP 是一种基于流形学习的降维算法,与 PCA、LLE 等传统降维算法相比,在非高斯大容量样本上表现良好,能够较好地保持原有高维数据集的全局结构. HDBSCAN 是一种基于层次的密度聚类算法,与传统方法相比,具有以下优势:可以对任意形状和任意密度的集群进行聚类;不依赖于数据集的数量以及聚类中心;对用户设定参数敏感性较低;可以对每个集群中的每个样本进行隶属度评价.

本文以北京市为例,将 UMAP 与 HDBSCAN 算法引入极端暴雨的时空动态分布格局分析中,应用 308 个气象站点的场次降雨数据为研究对象,挖掘研究区极端暴雨时空动态分布特征,为提高城市内涝风险精细化管理水平提供科学依据.

1 研究数据及预处理

1.1 研究区概况 北京地处华北平原,北依燕山,西靠太行山,东部距离海洋仅几百 km,整体地形复杂,加之城市快速扩张而导致“热岛效应”和“雨岛效应”,形成了复杂多变的降水特征. 2004 年以来,强降

雨对北京市区的夏季总降水量贡献比不断增加^[17],内涝事件时有发生.北京市的汛期主要集中在每年的 6—8 月份,而其降水主要来自 6 h 短历时暴雨事件^[18],也是引发城市内涝灾害的主导因素之一.如 2012 年“7·21”历史特大暴雨事件,引发了严重内涝,受灾人口高达 190 万人,79 人死亡,经济损失接近百亿元^[19].

1.2 数据来源及处理 本研究的降雨数据来源于北京市水务局,共 308 个气象站点,选取 2004—2016 年各站点汛期(6—8 月)时段内 5 min 时间分辨率的降雨资料作为样本进行分析,气象站点分布如图 1 所示.数据分析之前进行预处理和质量控制^[16].具体如下:1)如果单站 5 min 降雨量 > 10 mm,且是孤立的数据点,则认为该记录不合理;2)距某站 5 km × 5 km 范围内雨量站监测的降雨量为 0,但该站 5 min 降雨量 > 10 mm 的,认为该站记录的降雨量不合理.核对后,仅有 80 条记录标注不合理,约占 400 万条记录总数的 2×10^{-5} .对于单站不合理记录,查阅同期雷达图并进行对比.如果确认数据记录是不合理的异常值,则删除数据记录.此外,在距离该站 5 km × 5 km 范围内的其他雨量站收集的降雨数据中插值的 1 个值,用来代替不合理的记录.采用反距离平方加权法进行插值,并根据降雨事件对质量控制的数据进行分类.

此外,选取暴雨资料作为样本,并基于暴雨资料筛选极端暴雨事件,具体步骤为:1)暴雨筛选,根据中国气象中心规定,24 h 降水量达 50 mm、1 h 降雨量达 20 mm 的降雨定义为一次“暴雨”,记为一次暴雨事件;2)为凸显暴雨事件的极端性与灾害性特征,本研究选取每次暴雨事件最大 6 h 降雨作为一次极端暴雨事件,且暴雨区内至少有 1 个自动观测站的降水强度 $\geq 40 \text{ mm} \cdot \text{h}^{-1}$.

基于上述标准和处理过程,筛选出 45 场极端暴雨事件,将其划分为 12 个时段,用于构建降雨时空矩阵数据库.

2 研究方法

2.1 降雨时空矩阵构建 降雨移动路径、雨强变化、降雨范围等,均属于极端暴雨事件的时空分布特征,为充分描述这些特征,对筛选出的极端暴雨事件从时间维度和空间维度构建降雨时空矩阵.一场极端暴雨被定义为一个矩阵,若历史暴雨事件样本集中存在 N 个暴雨事件,则存在构建 N 个高维矩阵.利用该方法建立了一个降雨事件样本集 Ω ,时空矩阵为

$$\Omega = \{x_1, x_2, \dots, x_N\}, \quad (1)$$

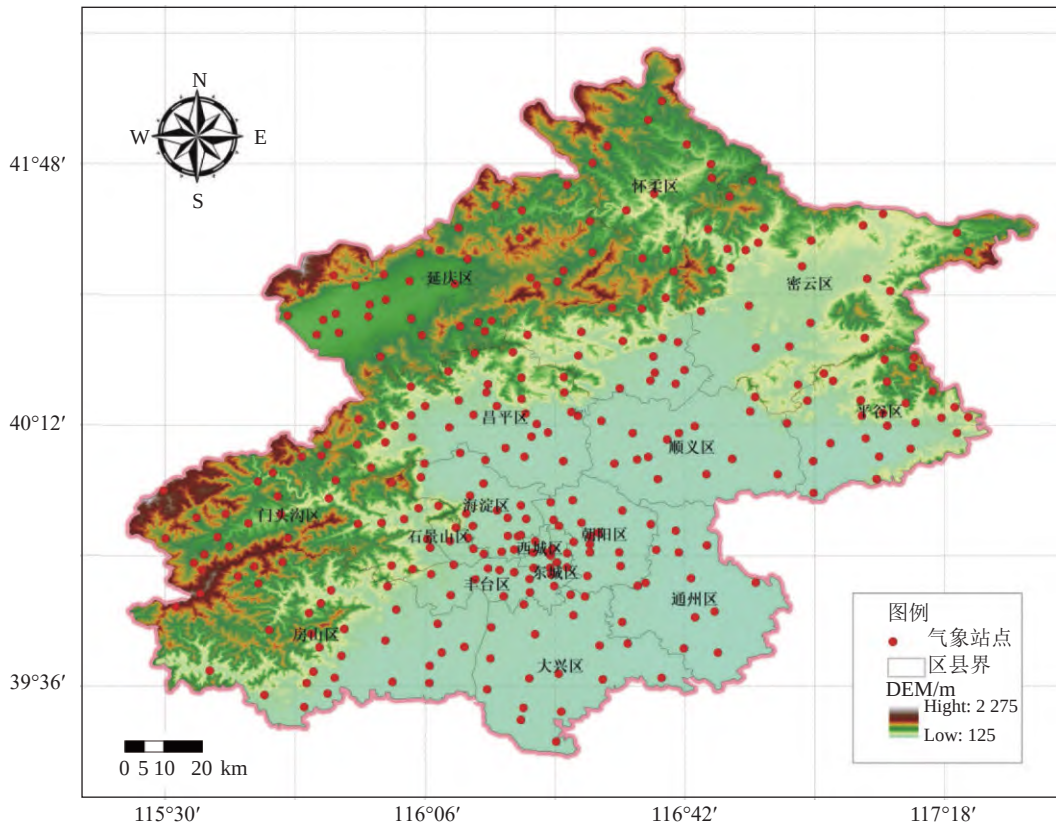


图1 北京气象站点分布

$$\mathbf{x}_j = \begin{bmatrix} H_{11,j} & H_{21,j} & \cdots & H_{s1,j} \\ H_{12,j} & H_{22,j} & \cdots & H_{s2,j} \\ \vdots & \vdots & & \vdots \\ H_{1m,j} & H_{2m,j} & \cdots & H_{sm,j} \end{bmatrix}, \quad (2)$$

式中: Ω 是由 N 次暴雨组成的历史极端降雨样本数据集; $\mathbf{x}_j$ 为第 j 次降雨时间的比例矩阵; $H_{in,j}$ 为第 j 次降雨事件中第 i 个雨量站的降雨量占 n 时刻所有雨量站总降雨量的百分比, $i=\{1,2,\cdots,s\}$, $n=\{1,2,\cdots,m\}$. 其中, $\Omega=45$, $s=308$, $m=12$.

2.2 UMAP 极端暴雨样本数据集是一个非线性、高维度的数据空间, 为提高数据分析效率及分析结果的准确性, 采用非线性数据降维算法 UMAP 对其进行降维分析. UMAP 是一种基于黎曼几何和代数拓扑理论而构建的数据降维算法, 依据高维空间映射到低维空间的相似度理论对高维数据进行降维^[20]. 算法计算步骤如下.

1) 设 $X=\{x_1,\cdots,x_n\}$ 为输入数据集, 给定输入超参数 k , 对于每个 x_i , 使用最近邻下降算法计算集合 $\{x_i,\cdots,x_k\}$. 对于每一个 x_i , 定义

$$\rho_i = \min\{d(x_i, x_j) \mid 1 \leq j \leq k, d(x_i, x_j) > 0\}, \quad (3)$$

$$\sum_{j=1}^k \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right) = \log_2 k, \quad (4)$$

式中: ρ_i 要确保选择至少有 1 个边缘权值为 1 的数据点与 x_i 相连; σ_i 为长度尺度参数.

2) 定义有向加权图 $\bar{G}=(V,E,w)$, 利用 $\bar{G}$ 的对称性质定义无向加权图 G , G 的顶点 V 是集合 X , 由此形成有向边集合 $E=\{(x_i, x_j) \mid 1 \leq j \leq k, 1 \leq i \leq N\}$. 定义权重函数

$$\omega(x_i, x_j) = \exp\left(\frac{-\max(0, d(x_i, x_j) - \rho_i)}{\sigma_i}\right). \quad (5)$$

定义 A 为 $\bar{G}$ 的加权邻接矩阵, 可以得到一个对称矩阵, 得到无向加权图 G 的邻接矩阵 B :

$$B = A + A^T - A \otimes A^T, \quad (6)$$

式中 $\otimes$ 为 Hadamard 乘积.

3) UMAP 沿边界和顶点分别施加引力和斥力演化出一个包含点集 $\{y_i\}, i=1,\cdots,N$ 的等价加权图 H , 2 顶点 i 和 j 在坐标 y_i 与 y_j 处的引力分别为

$$\frac{-2ab\|y_i - y_j\|_2^{2(b-1)}}{1 + \|y_i - y_j\|_2^2} \omega(x_i, x_j)(y_i - y_j), \quad (7)$$

$$\frac{2b(1 - \omega(x_i, x_j))(y_i - y_j)}{(\varepsilon + \|y_i - y_j\|_2^2)(1 + a\|y_i - y_j\|_2^{2b})}, \quad (8)$$

式中: ε 是一个很小的值, 防止分母为 0; a 和 b 为超参数.

UMAP 降维分析主要受超参数 k (降维后的数据

维度)、 a 与 b 的影响,其中超参数 a 和 b 的取值由 n_n (集群最小数量)与 $\min(d)$ (样本间最小距离)的大小决定. UMAP 可将数据降维至任意维度,算法认为高维空间通过映射变幻得到低维空间后,其相关特性和数据结构仍可与原高维空间一致或相似. 因此,在本文的实际分析中,为了对样本间的拓扑结构及局部特征进行可视化分析,设置 $k=2$,同时对 n_n 取值 2~4, $\min(d)$ 取值 0.1~0.3 进行了检验. 最终发现,设置 $n_n=2$, $\min(d)=0.1$ 是最佳选择,具体分析过程在此不作赘述,详见 3.1 节.

2.3 HDBSCAN 为提取不同类别极端暴雨事件的时空分布特征,采用 HDBSCAN 密度聚类算法对降维样本进行分类,使得同类中样本近似,而各类之间样本不同. HDBSCAN 是由 Campello 等^[21]提出的一种基于密度聚类与层次聚类相结合的算法,算法引入相互可达距离(图 2)的度量方法对不同族群的树形层次结构进行搭建,以实现不同密度族群的聚类. HDBSCAN 可为每个样本进行隶属度评价,隶属度取值范围为 $[0, 1.0]$. 若隶属度取值为 0,表明样本点为噪声点,不属于任何簇;若隶属度取值为 1.0,表明样本点为中心点,簇中心点属性可代表该簇的典型特征. 算法计算步骤如下.

1) 数据点的空间变换. 核心距离 $c_k(x) = d(x, N^k(x))$ 指数据点与其第 k 个最邻近数据点的距离. 2 点之间的互达距离计算式为

$$d_k(a, b) = \max\{c_k(a), c_k(b), d(a, b)\}, \quad (9)$$

式中, $d(a, b)$ 指的是 a 、 b 2 点之间的直线距离.

2) 最小生成树的构建. HDBSCAN 算法内部采

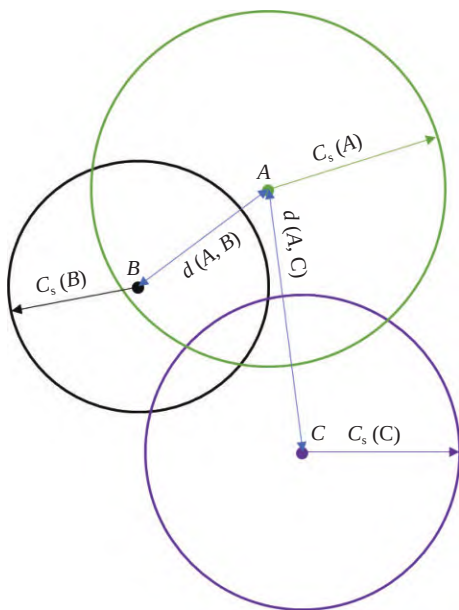


图 2 互达距离示意

用 Prim(普里姆算法),将样本点作为原始点、其他点与原始点的互达距离作为权重,构建最小生成树.

3) 集群层次结构的建立. 算法以互达距离为权重,对最小生成树的边进行遍历、重新排序,将每条边归类到一个新的簇.

4) 聚类树的压缩. 算法在对粗层次结构进行压缩分割时,会对新分割出的簇样本数与最小簇的样本数大小进行对比,把数量较小的一方进行剔除.

5) 聚类簇的提取. 算法以 $\lambda = 1/d$ 为标准对聚类结果进行度量,并引入 λ_{birth} 、 λ_{death} 、 λ_p 3 个量,分别用来代表聚类簇形成时、聚类簇分裂为 2 个子簇时以及从父簇分离出的值. 用稳定度 S_{cluster} 来评估聚类簇的稳定性. S_{cluster} 的计算式为

$$S_{\text{cluster}} = \sum_{p \in C_{\text{cluster}}} (\lambda_p - \lambda_{\text{death}}), \quad (10)$$

式中 C_{cluster} 为聚类簇集合.

HDBSCAN 聚类分析中用到的主要参数为 C_{cluster} ($\min_cluster_size$, 最小聚类簇数量),本研究对 3~8 的不同 C_{cluster} 取值进行了检验,发现当类别数为 4 时,2 分类的特征差异明显,重复消失. 由此可知,从降雨的时空动态分布来看,极端暴雨有 4 种类型,详见 3.2 节.

2.4 降雨质心 为直观对比不同降雨模式的时空分布规律,基于物理中质心的概念,采用降雨质心来表示研究区中某个时段平均降雨中心的位置及雨量大小. 根据降雨质心的移动和变化,可以更加直观地对比 4 类降雨模式的时空变化规律. 第 i 个时段的降雨质心的坐标以及雨量^[22]计算式为:

$$P(x) = \frac{\sum_{j=1}^n P_j W(j, i)}{\sum_{j=1}^n W(j, i)}, \quad (11)$$

$$C_i(x) = \frac{\sum_{j=1}^n X_j W(j, i)}{\sum_{j=1}^n W(j, i)}, C_i(y) = \frac{\sum_{j=1}^n Y_j W(j, i)}{\sum_{j=1}^n W(j, i)}, \quad (12)$$

式中: $P(x)$ 为质心在第 i 个时段的降雨量; P_j 表示第 j 个站点的降雨量; n 表示站点的个数; $W(j, i)$ 表示第 j 个站点在第 i 个时段所有站点中的降雨比例; $C_i(x)$ 和 $C_i(y)$ 为第 i 个时段降雨质心的经纬度坐标; X_j 和 Y_j 分别为第 j 个站点的经纬度坐标.

3 结果与讨论

3.1 UMAP 模型参数选取 在确定降维维度 k 后,UMAP

降维效果主要受 n_n (集群最小数量) 与 $\min(d)$ (样本间最小距离) 的影响. 图3为应用UMAP模型对极端暴雨高维矩阵降维时, 不同模型参数组合对降维结果的影响. 当 n_n 和 $\min(d)$ 越小时, 聚类簇之间的距离越大, 同一簇样本间的距离越小; 当 n_n 和 $\min(d)$ 的值变

大时, 全部的样本趋向一个整体结构, 无法保留样本间的簇内拓扑关系, 该结果表明降雨时空矩阵的高维结构在合适的参数选择下, 才能通过聚类簇呈现出来. 因此, 结合实际降雨过程, 选择 $n_n = 2$ 、 $\min(d) = 0.1$ 作为模型最优参数组.

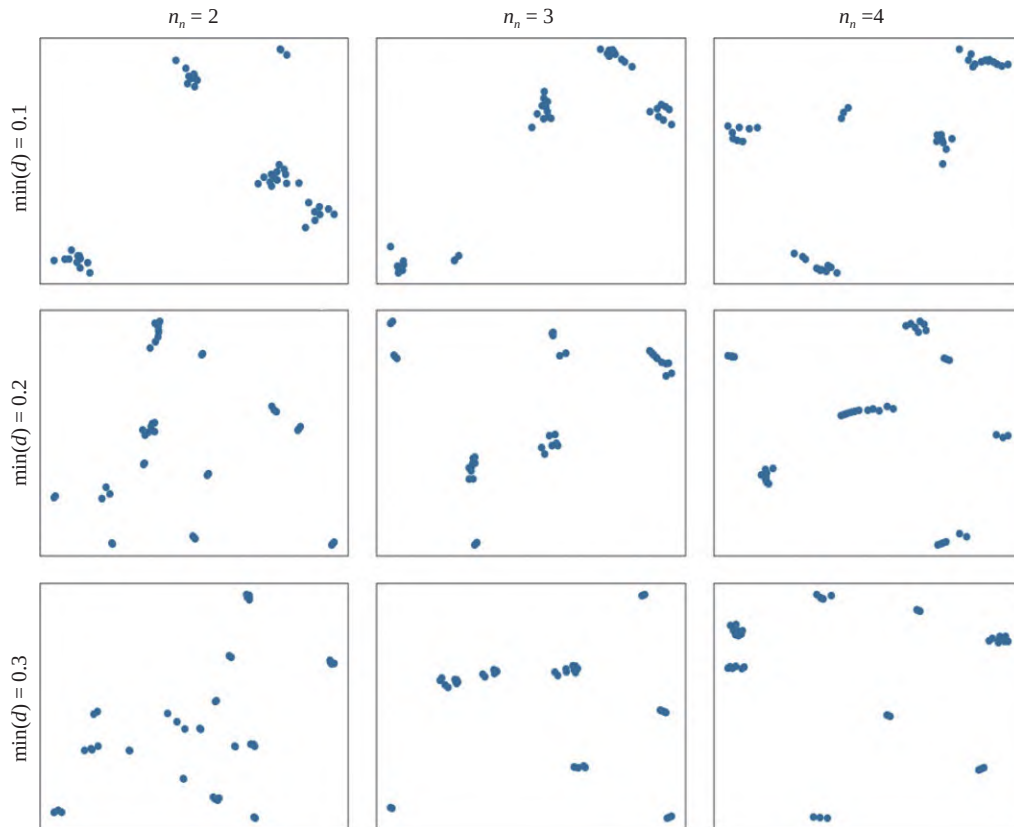


图3 参数选择对UMAP降维结果的影响

3.2 HDBSCAN 聚类分析 $\min(c)$ (最小聚类簇数量) 是决定 HDBSCAN 聚类效果的主要参数. 经过反复测试, 当 $\min(c)$ 取值为 4 时, 分类特征差异明显, 不同类别样本间无重复情况, 表明根据暴雨的时空分布特征可将其分为 4 种模式. 图4为 HDBSCAN 聚类结果, 45 个暴雨样本被划分为 4 簇, 其中 3 个暴雨样本隶属度取值为 0, 被标识为噪声点, 经验证, 噪声点的时空特征与 4 类模式特征差异较大, 且无相似特征. 该结果表明 HDBSCAN 可以较好地捕捉“异常”暴雨样本, 并对存在“共性”的暴雨事件进行归类.

3.3 极端暴雨时空动态分布模式识别 利用 HDBSCAN 聚类算法得到 4 个分类簇是降维后的低维特征空间. UMAP 算法是基于高维空间映射到低维空间的相似度理论, 即低维空间同一子集的样本在高维空间中也是相似的, 因此低维空间中的同一子集的样本在高维空间中也归入同一子集. 选取高维空间中的隶属度为 1 的样本作为对应类别的聚类中心, 构成最终的

特征空间, 用于描述该类样本的时空动态分布格局. 结果表明, 北京市极端暴雨事件按时空分布格局可分为 4 种模式: 模式 1, 暴雨中心主要集中在主城区, 并围绕主城区缓慢移动; 模式 2, 暴雨中心从西南山区穿过主城区, 向东北方向移动; 模式 3, 暴雨中心自西

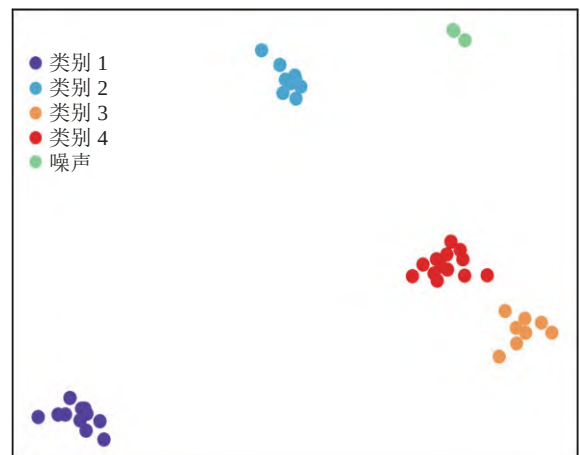


图4 HDBSCAN 聚类结果

部山区向城区西北部扩散,最后向城区北部移动;模式 4,暴雨中心从西部山区途经城区向东移动。

图 5 展示了 4 类降雨模式的质心运动轨迹和相应的雨强变化趋势。本研究将 360 min 降雨过程划分为 72 个 5 min 时段,每个点代表特定时间段整个城市的降雨中心。点的颜色的深浅与大小分别代表不同降雨时段和不同雨强大小。在 360 min 内,暴雨中心在主城区上空盘旋,移动距离较短,雨强整体变化较小,呈递减趋势,定义为第 1 类暴雨模式;第 2 类暴雨

模式中,暴雨质心从西南山部经过主城区向东北方向移动,移动距离较长,雨强呈递减趋势;第 3 类暴雨模式中,降雨质心从西部山区移向东南部,紧接着向主城区北部移动,移动距离较短,雨强呈递减趋势;第 4 类降雨模式中,暴雨质心从中部山区,横跨主城区北部向东移动,移动速度快,雨强变化快。总体来看,4 类暴雨模式均有从西向东移动的趋势,暴雨中心的散落位置主要集中在主城区,雨强随质心移动整体呈减小趋势。

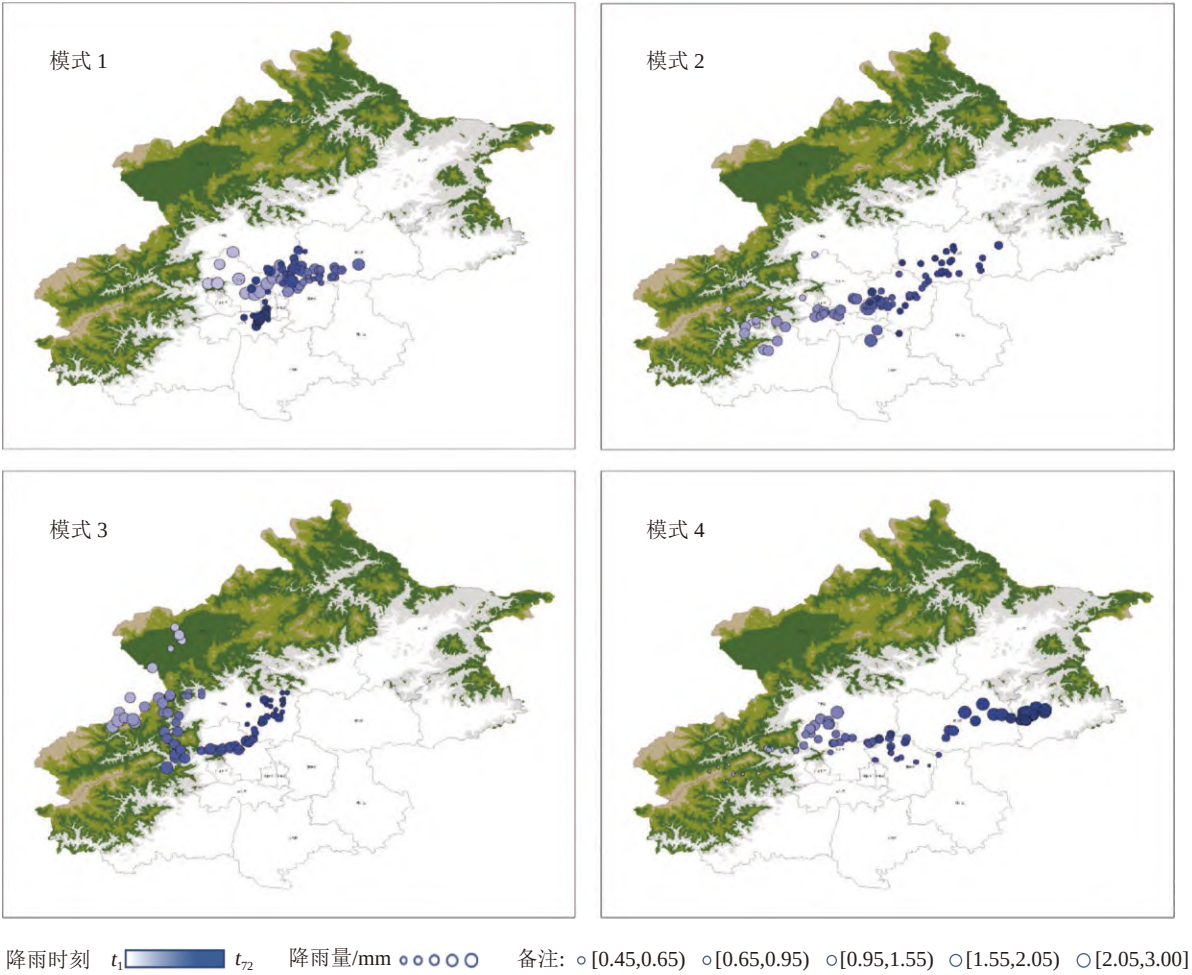


图 5 不同极端暴雨模式质心时空分布

3.4 极端暴雨时空格局重构 根据时空分布格局重构每一类特征暴雨过程,通过克里金插值方法得到 30 min 累积降雨全图,用于表示该模式暴雨的“典型”状态。与质心时空分布图相比,全图侧重于描述不同时段降雨累积量的空间分布。通过 2 种类型图间的对比,可以全面了解暴雨过程的动态发展。

图 6 所示为模式 1 极端暴雨“典型”过程,在东亚中高维“两脊一槽”环流和副热带高压复合作用下形成,为大面积降水。模式 1 类型降雨样本占比 24.4%,为北京城区极端降雨主要类型之一。2016 年“7·20”

降雨过程为模式 1 的一次典型极端降雨事件,受中纬度槽与副热带高压相互作用,以低涡系统性降水为主^[23],“城市雨岛”效应明显,雨强大,覆盖范围广,全市平均降雨 203 mm,主城区平均降雨 291 mm,降雨中心主要集中在中心城区,基本不发生移动,整个降雨过程中雨强明显减小,与模式 1 所描述过程的降雨状态基本吻合。

图 7 展示了第 2 类模式的极端暴雨过程,为暖区降水,由北京西南山区槽前的低空强对流或暖湿气流造成,因冷暖气流交汇处不稳定性较大^[14],整体雨强

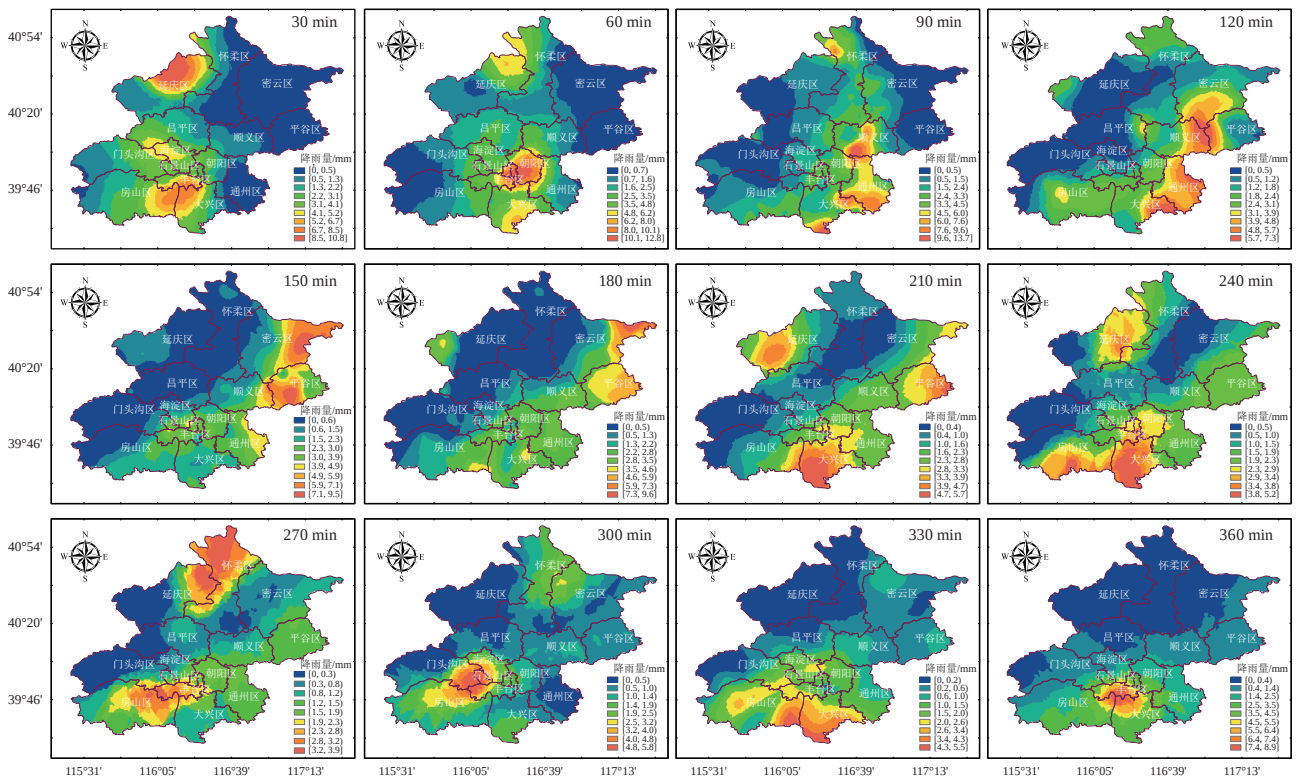


图6 模式1暴雨时空分布

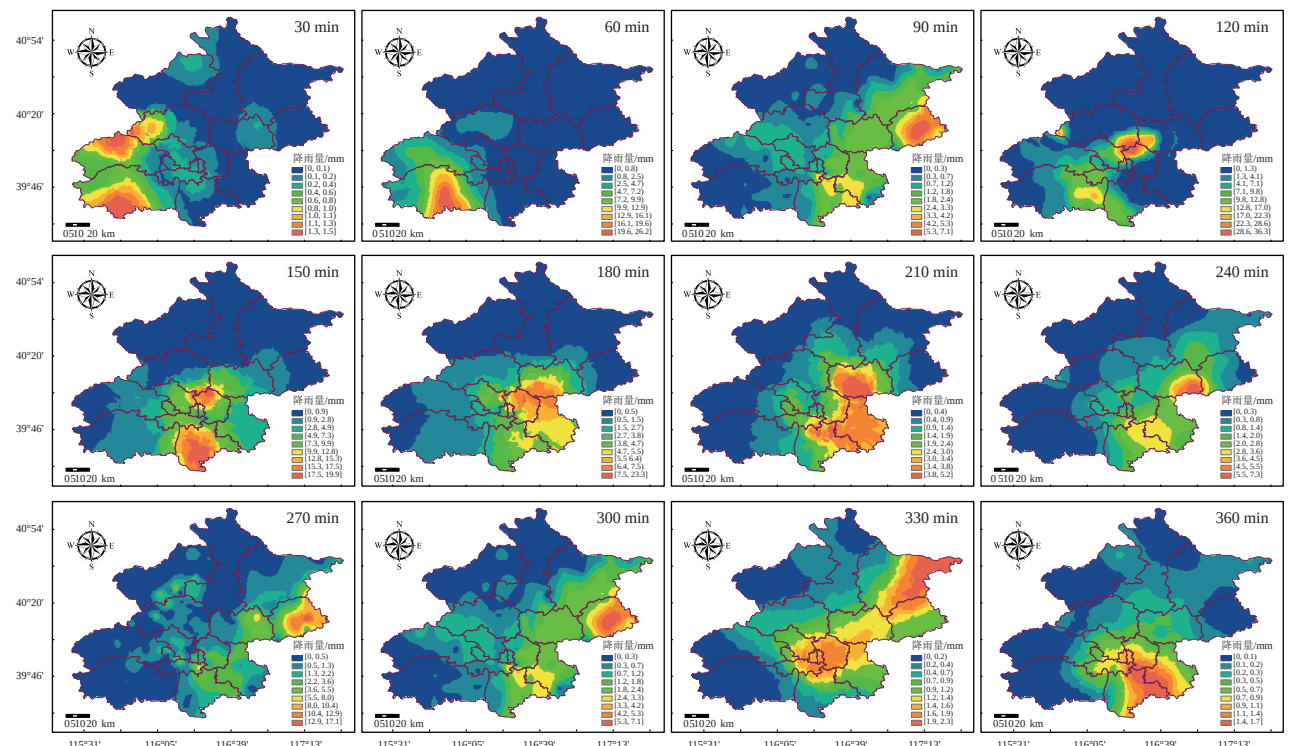


图7 模式2暴雨时空分布

大。这一类暴雨模式中,降雨从西南开始,向城市中西部扩散,雨带还有向城市南部延伸的趋势,整个过程中,雨强不断减小。该模式降雨类型占总样本 22.2%,与模式 1 相比,雨带移动过程相对较快。2013 年 7 月 1 日降雨事件为模式 2 极端暴雨类型的实际案例,此

次暴雨移动过程基本符合模式 2 所描述的规律,雨强先增后减,整体符合雨强减小的趋势。

图 8 所示为第 3 类模式的极端暴雨过程,由东部山前低涡东移过程中遇到充足水汽所产生的强降水^[18]。降雨从西北部山区逐步向中心城区扩散,然后

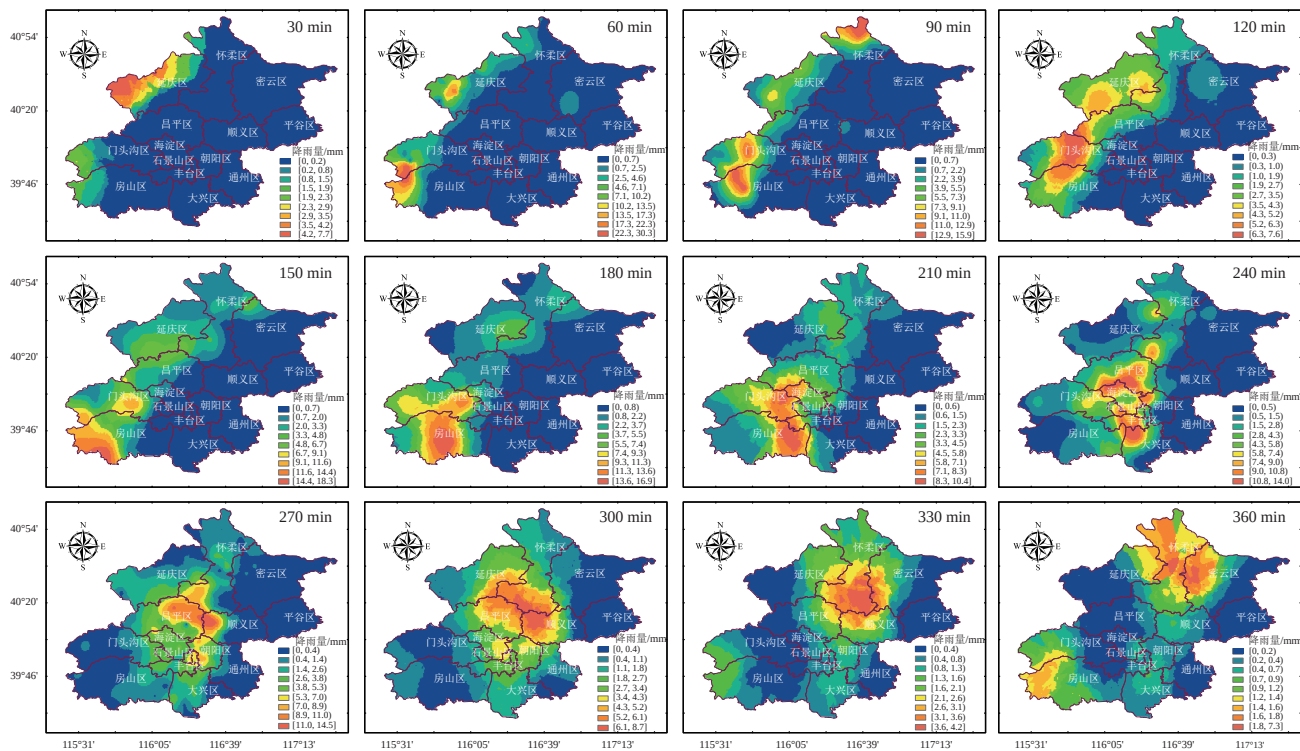


图 8 模式 3 暴雨时空分布

向北移动。整个降雨过程中,雨强偏小且呈减小趋势,降雨覆盖范围小,呈较强的局地性。该模式降雨事件约占总样本的 18%。2006 年 6 月 27 日发生的降雨事件,同模式 3 极端暴雨具有相似的时空动态变化特征。具体来说,暴雨从西北部向主城区扩散,再向城市北部移动,整体降雨范围较大,雨强呈先增后减的趋势。

图 9 为第 4 类模式极端降雨过程,由多单体合并或多单体列车效应引起^[7, 18]。降雨从城市西部移动至城市东部,雨强无明显变化趋势,降雨范围呈条带分布。模式 4 降雨事件占样本总数 28.9%,为最主要的极端降雨模式。著名的“7·21”极端暴雨事件^[23],时空动态分布特征与模式 4 所描述规律高度吻合,降雨条带自西向东快速移动,雨强较大,降雨分布范围广,前期呈多个暴雨中心,雨强呈增强趋势。

综上所述,从极端暴雨的形成机制考虑,聚类模型所识别的 4 类降雨事件之间具有明显差异。聚类模型的训练过程从极端暴雨的特征出发,不直接考虑极端暴雨形成的物理机制,但其结果可以根据不同类型极端暴雨的形成机制差异进行进一步挖掘和有效解释。

3.5 讨论 对比其他相关研究,在雨强随时间变化的趋势方面,本研究的 4 类极端暴雨模式均具有在 6 h 内雨强衰减的特征,表明北京极端暴雨事件的主要类型为 6 h 短历时降水过程,这一特征与孙继松等^[18]的

研究结果相吻合;在极端暴雨的空间发展趋势方面,本研究的 4 类极端暴雨过程均有从西向东的移动趋势,主城区、东南山区以及南部平原地区面临更高的极端降雨风险,与 Song 等^[7]的研究结果基本相似;同时考虑北京极端暴雨时间、空间分布特征并对降雨进行分类的研究不多,仅 Liu 等^[15]的研究可用来进行横向对比与参考。该研究基于北京主城区 14 个气象站点的降雨资料,总结了 3 类极端暴雨模式,其中“降雨从西北部山区移动到中心城区”“降雨集中在城区基本不发生移动”2 类降雨模式与本研究中的第 1、3 类降雨模式基本对应;而本研究中的第 2、4 类降雨模式,则被 Liu 等^[15]统一划分为“降雨自城区西南部向城中心与城北部扩散”一类,根据本研究结果可知,2 类降雨模式仅存在时空局部交叉,但其物理发生机制、移动路径、降雨范围均存在明显区别,属于 2 类不同的极端暴雨模式。该研究与本研究的结论之间的差异可能是数据在空间层面的详细程度导致的。由于该研究的雨量站点数量相对较少,空间精度受到一定限制,可能难以有效把握暴雨的局部特征,单一站点数据的不确定性也被放大。本研究采用的站点数据相对更翔实,空间分布更密集,在一定程度上降低了输入数据的不确定性,能够较好地地区分时空分布存在相似性的不同降雨模式,从而更全面地揭示了北京全域降雨时空动态分布规律。

在研究方法方面,相比于传统的 PCA 与 LLE 等

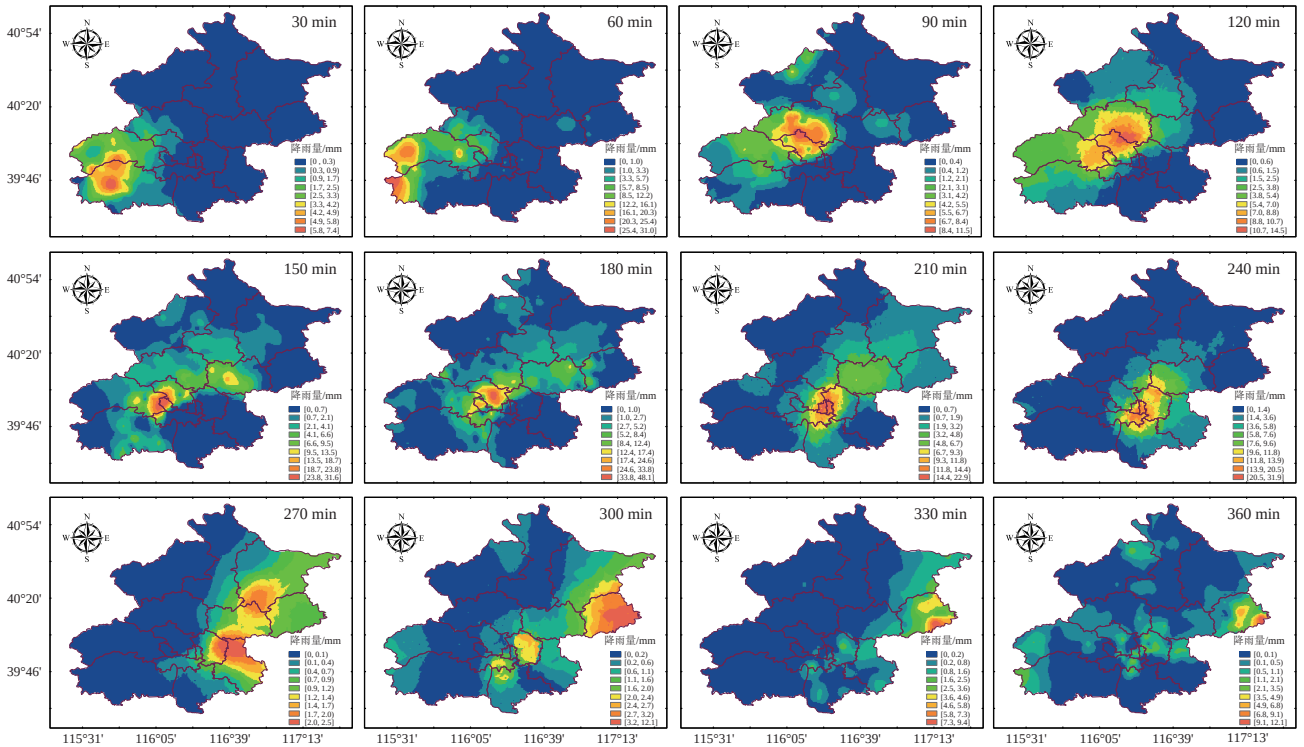


图9 模式4暴雨时空分布

降维方法^[15-16],本研究采用的UMAP方法在处理北京地区的高维降雨数据集时,可以在降低数据冗余度的基础上有效保留数据集原有的局部特征(图3),表明这一方法在处理高维数据集时具有良好的适用性。刘媛媛等^[14]应用动态聚类算法对城市短历时暴雨模式进行分析,由于算法对降雨样本中的异常点不敏感,聚类结果易受初始聚类中心选择的影响,难以得到理想的聚类效果。而本研究采用的HDBSCAN算法,能够有效识别样本数据间的局部结构,捕捉样本数据中的异常点(图4),实现对任意形状的样本数据簇进行聚类,得到良好的分类效果。本研究的结果显示,相比其他已有方法体系,UMAP与HDBSCAN方法可以更有力地支撑基于高维数据集的极端暴雨时空分布格局分析,能够有效地捕捉不同极端暴雨类型之间的细微差异。

4 结论

本文将UMAP降维算法与HDBSCAN密度聚类算法引入降雨时空分布特征研究中,成功提取北京全域4类极端暴雨模式:模式1,暴雨集中在主城区,并围绕主城区缓慢移动;模式2,暴雨从西南山区途经主城区,向东北方向移动;模式3,暴雨自西部山区向主城区西北部扩散,最后向主城区北部方向移动;模式4,暴雨从西部山区途经主城区向东移动。研究结果表明,北京市极端暴雨的主要类型为短历时降雨

过程,暴雨中心在空间上均存在从西向东移动的趋势。其中,主城区、东南山区以及南部平原地区面临相对更高的极端暴雨风险。本研究提出的方法具有良好的适用性。UMAP结合HDBSCAN的机器学习方法可以对暴雨事件进行分类,定量描述各类降雨事件的发展和动态时空降雨分布规律。有关极端暴雨的动态时空分布规律,可以促进城市洪涝风险管理。需要指出的是,虽然北京短历时极端暴雨事件的频率高于长历时暴雨,但长历时暴雨带来的负面影响同样不可忽视。为更加完整地描述北京极端暴雨模式及其特征,未来可收集更长时间序列的降雨数据,开展长历时暴雨的动态时空分布特征分析。

5 参考文献

- [1] 程晓陶,刘昌军,李昌志,等.变化环境下洪涝风险演变特征与城市韧性提升策略[J].水利学报,2022,53(7):757
- [2] 陈文龙,徐宗学,张印,等.高密度城市暴雨洪涝治理理论框架及其应用研究[J].水利学报,2022,53(7):769
- [3] 舒章康,李文鑫,张建云,等.中国极端降水和高温历史变化及未来趋势[J].中国工程科学,2022,24(5):116
- [4] 张炜,李思敏,时真男.我国城市暴雨内涝的成因及其应对策略[J].自然灾害学报,2012,21(5):180
- [5] 栾清华,秦志宇,王东,等.城市暴雨道路积水监测技术及其应用进展[J].水资源保护,2022,38(1):106
- [6] 邸苏闯,李卓蔓,刘玉,等.基于气象雷达反演和云图外推法的临近期降雨预报方法研究[J].水利水电技术(中英

- 文), 2022, 53(5): 13
- [7] SONG X M, ZHANG J Y, ZOU X, et al. Changes in precipitation extremes in the Beijing metropolitan area during 1960-2012[J]. *Atmospheric Research*, 2019, 222: 134
- [8] PILGRIM D H, CORDERY I. Rainfall temporal patterns for design floods[J]. *Journal of the Hydraulics Division*, 1975, 101(1): 81
- [9] KEIFER C J, CHU H H. Synthetic storm pattern for drainage design[J]. *Journal of the Hydraulics Division*, 1957, 83(4): 1332
- [10] WILLEMS P, ARNBORG NIELSEN K, OLSSON J, et al. Climate change impact assessment on urban rainfall extremes and urban drainage: methods and shortcomings[J]. *Atmospheric Research*, 2012, 103: 106
- [11] 夏军, 石卫. 变化环境下中国水安全问题研究与展望[J]. *水利学报*, 2016, 47(3): 292
- [12] RAFIEEINASAB A, NOROUZI A, KIM S, et al. Toward high-resolution flash flood prediction in large urban areas - analysis of sensitivity to spatiotemporal resolution of rainfall input and hydrologic modeling[J]. *Journal of Hydrology*, 2015, 531: 370
- [13] 胡睿, 王双银, 王鹏飞. 基于聚类分析的短历时暴雨雨型研究[J]. *水电能源科学*, 2021, 39(4): 8
- [14] 刘媛媛, 刘洪伟, 霍凤霖, 等. 基于机器学习短历时暴雨时空分布规律研究[J]. *水利学报*, 2019, 50(6): 773
- [15] LIU W, CHEN S, TIAN F. Spatiotemporal modes of short time rainstorms based on high-dimensional data: a case study of the urban area of Beijing, China[J]. *Water*, 2021, 13(24): 3597
- [16] LIU Y Y, LI L, LIU Y S, et al. Dynamic spatial-temporal precipitation distribution models for short-duration rainstorms in Shenzhen, China based on machine learning[J]. *Atmospheric Research*, 2020, 237: 104861
- [17] YUAN Y F, ZHAI P M, LI J, et al. Changes in classified precipitation in the urban, suburban, and mountain areas of Beijing[J]. *Advances in Climate Change Research*, 2017, 8(4): 279
- [18] 孙继松, 雷蕾, 于波, 等. 近10年北京地区极端暴雨事件的基本特征[J]. *气象学报*, 2015, 73(4): 609
- [19] ZHOU Y P, LAU W K M, LIU C T. Rain characteristics and large-scale environments of precipitation objects with extreme rain volumes from TRMM observations[J]. *Journal of Geophysical Research: Atmospheres*, 2013, 118(17): 9673
- [20] MCINNES L, HEALY J, MELVILLE J. UMAP: uniform manifold approximation and projection for dimension reduction[EB/OL]. 2018[2022-10-08]. <https://arxiv.org/abs/1802.03426>
- [21] CAMPello R J G B, MOULAVI D, SANDER J. Density-based clustering based on hierarchical density estimates[C]//Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17. Springer Berlin Heidelberg, 2013: 160-172
- [22] LIU Y S, YANG Z S, HUANG Y H, et al. Spatiotemporal evolution and driving factors of China's flash flood disasters since 1949[J]. *Science China Earth Sciences*, 2018, 61(12): 1804
- [23] 陆婷婷, 崔晓鹏. 北京两次特大暴雨过程观测对比[J]. *大气科学*, 2022, 46(1): 111

Spatiotemporal dynamics distribution patterns of extreme rainstorms in Beijing revealed by UMAP and HDBSCAN

XU Shugao¹⁾ WANG Xianyang¹⁾ JIANG Weiwei²⁾ YU Jingshan¹⁾
LIU Yuan³⁾ ZHOU Guihuan¹⁾

(1) Digital Watershed Laboratory, College of Water Science, Beijing Normal University, 100875, Beijing, China;

2) State Key Laboratory of Environment Geochemistry, Institute of Geochemistry, Chinese Academy of Sciences, 550081, Guiyang, Guizhou, China;

3) China Institute of Water Resource and Hydropower Research, 100038, Beijing, China)

Abstract Extreme rainstorm is characterized by short duration, strong downpours and widespread destruction—all major causes of flash flood and waterlogging disasters. Exploring spatiotemporal dynamics of extreme rainstorm is helpful to refined management level of urban waterlogging risk. A method of clustering combined with dimensionality reduction algorithm was proposed in this study, to extract spatiotemporal dynamics of extreme rainstorm events. Precipitation data from 308 meteorological stations in Beijing from 2004 to 2016 were used, to construct spatiotemporal dynamics models of various extreme rainstorm events with the HDBSCAN clustering method. This

was combined with UMAP dimensionality reduction algorithm to extract 4 types of extreme rainstorm models in Beijing. In mode 1, rainstorm was found concentrated in the main urban area, moving slowly around. In mode 2, rainstorm was found to move from the southwest mountainous area through the main urban area to the northeast. In mode 3, the rainstorm was found to spread from the western mountainous area to the northwest of the city, and to finally move to the north of the city. In mode 4, torrential rain was found to move westward from western mountain area to urban area. These results show that the main type of extreme rainstorm in Beijing is short-duration precipitation, and the rainstorm center has the tendency to move from west to east. Downtown area, southeastern mountainous area and southern plain area all face relatively higher extreme precipitation risks. The physical mechanism of model characteristics of all kinds of extreme rainfall is clear, and the reconstruction characteristics of rainfall events could fully represent actual rainfall events. This study provides important basis for flood forecasting and scientific management in Beijing.

Keywords extreme rainstorm; UMAP; HDBSCAN; dynamic spatiotemporal distribution; Beijing

【责任编辑: 武 佳】