

融合改进 CNN-LSTM 算法的山区缺资料流域水文预报模型构建

付丽娜 杨 腾

阿拉善乌海水文水资源分中心, 内蒙古 阿拉善盟 750306

摘 要: 我国中西部山地中小流域普遍存在水文观测站点布设稀疏、径流时序记录残缺、下垫面空间异质性剧烈等典型缺资料困境, 传统集总式、分布式水文模型依靠参数区域移植进行模拟, 在地形起伏剧烈山区易出现洪峰失真、峰现时间偏移等问题; 常规 CNN-LSTM 耦合网络仅提取一维气象时序特征, 长时序推演易产生梯度衰减, 且短观测样本下极易发生过拟合, 难以适配无长期径流记录山区流域。本文以秦岭北麓黑河上游山区集水区为研究载体, 构建一套适配缺资料场景的残差多尺度混合卷积 + 因果多头注意力 + 相似流域迁移学习 + 双模式时序数据增强一体化改进 CNN-LSTM 预报架构。研究创新构建地形 - 遥感 - 实测气象三维耦合输入张量, 依托并行多尺度残差卷积分层捕捉海拔、坡度、植被、降水耦合产流空间分异规律; 引入带因果掩码的多头注意力机制屏蔽未来时序信息, 自适应放大暴雨、季节性融雪等致洪关键时段特征权重; 针对支流径流样本稀缺问题, 建立水文相似度筛选的跨流域迁移预训练体系, 搭配时序平移、高斯扰动双重样本扩充策略缓解小样本训练缺陷。依托 2010—2023 年日尺度星地多源水文数据集完成模型训练与验证, 全域完整观测流域测试集纳什效率系数 NSE 可达 0.916, 相对偏差 PBIAS 为 4.31%, 均方根误差 RMSE=2.75m³/s; 对比标准单层 LSTM、基础 CNN-LSTM 模型, NSE 分别提升 0.141、0.088。在上游无连续径流记录的缺资料子流域短序列测试场景下, 结合迁移学习与数据增强的改进模型洪峰相对误差降至 7.3%, 峰现时间平均误差缩短至 0.82 h, 相较于未优化模型精度提升显著。本文提出的改进框架从空间特征提取、时序关键信息强化、小样本泛化优化三个维度突破传统深度学习水文模型局限, 可为山区缺资料流域山洪预警、水资源精细化调控提供新型时空耦合预报技术支撑。

关键词: 缺资料流域; 水文径流预报; CNN-LSTM; 残差多尺度卷积; 因果多头注意力

DOI: 10.64649/yh.shfzykjcx.issn3078-8994.202607024

0 引言

国际水文科学协会提出的 PUB 计划已开展二十余年, 缺资料流域径流模拟始终是现代水文学亟待攻克的核心难题。我国秦巴山区、云贵高原、横断山脉等地区数千条中小支流受山地地形阻隔、监测工程建设成本约束, 仅干流出口布设长期径流监测站点, 多数上游支沟仅开展短期巡测, 有效径流观测时长普遍不足一年, 无法支撑 SWAT、HBV 等传统分布式水文模型完成可靠参数率定。山区海拔落差大、沟壑交错, 降水、土壤含水量、植被覆盖空间分布差异显著, 依靠邻近流域参数移植的区域化方法无法刻画局地产汇流规律, 汛期极端洪水模拟普遍存在低估缺陷, 山洪灾害预警可靠性不足。

深度学习凭借强大的多变量非线性拟合能力逐步应用于流域径流时序预报, 卷积神经网络 CNN 擅长挖掘多维输入数据的局部空间关联特征, 长短期记忆网络 LSTM 能够捕捉径流演变的滞后时序响应关系, 二者结合的 CNN-LSTM 混合模型成为当前水文时序模拟主流框架。但现有基础架构应用于山区缺资料流域时

存在三重固有短板: 第一, 多数研究仅将气象序列作为一维时序输入, 忽略 DEM 地形、卫星植被、积雪覆盖等静态空间因子对产流过程的约束作用, 丢失山区核心空间异质性信息; 第二, 标准 LSTM 在超过 30 天的长时序推演中易出现梯度消失, 暴雨、融雪等对径流贡献突出的特殊水文时段特征被弱化, 极端流量模拟偏差扩大; 第三, 常规 CNN-LSTM 依赖连续多年完整径流样本训练, 针对观测记录残缺的缺资料支流, 小样本训练极易引发过拟合, 跨子流域泛化能力大幅衰减。

1 研究区域与多源数据预处理

1.1 研究区基础概况

研究区域选定陕西省秦岭北麓黑河上游金盆水库以上集水流域, 全域总面积 728 km², 海拔区间 562~2886 m, 地形切割强烈, 海拔高差超 2300 m, 属于典型中高海拔山地流域。流域水系呈树枝状分布, 干流具备连续 13 年长时序径流监测记录, 上游分布 12 条一级支流, 仅 2018—2020 年开展间歇性水文巡测, 单条支流有效日径流观测样本仅 287 组, 无完整年度连

续序列,属于标准 PUB 缺资料研究单元。区域气候兼具温带季风与山地垂直气候特征,多年平均降水量 871.6 mm,降水空间分布不均,高海拔山区暴雨频次更高,春季融雪径流占全年总径流量 21.3%,汛期 6—9 月径流占全年总量 67.2%,每年均存在不同量级山洪过程,对水文预报精度存在较高需求。传统单站点雨量插值法难以还原山区暴雨中心实际降水量,进一步加剧缺资料子流域径流模拟难度。

1.2 多源数据构成与划分方案

研究采用 2010—2023 年连续 13 年日尺度多源数据集,完整干流样本共计 4745 组,按照 7:1.5:1.5 比例划分训练集、验证集、测试集;12 条上游缺资料支流单独构建短序列验证数据集,单支流原始有效样本 287 组,用于检验模型在小观测样本场景下的泛化性能。模型输入协变量划分为静态地形、动态气象、遥感空间三类:第一类,地面实测气象数据,包含干流监测站日降水量、日均气温、相对空气湿度、日照时长四项核心时序变量;第二类,卫星遥感空间产品,选用 0.1° 分辨率 GPM IMERG 日降水数据、MODIS NDVI 植被指数、MODIS 逐日积雪覆盖率、表层土壤含水量遥感产品,弥补山区雨量站点空间覆盖不足问题;第三类,静态地形因子,基于 30m 分辨率 DEM 数字高程模型提取流域平均海拔、平均坡度、坡向、沟壑密度、土壤类型五类固定空间特征,作为独立特征通道嵌入输入张量。

模型输出目标变量为流域出口日均径流流量,单位 m^3/s ;选取水文领域通用五项量化评价指标:纳什效率系数 NSE、百分比偏差 PBIAS、均方根误差 RMSE、洪峰流量相对误差、峰现时间误差,综合评判模型模拟优劣。

1.3 缺资料场景专属数据预处理流程

针对山区数据残缺、观测误差、样本量不足三类问题设计分层预处理流程:首先开展缺失值修复,气象序列短期空缺采用线性插值补齐,连续 7 天以上长时间缺失借助同海拔邻近流域站点 K 近邻算法填充;对 GPM 卫星降水产品开展地形残差校正,消除高海拔区域降水低估偏差。其次完成特征标准化,全部动态时序变量统一归一化至 $[-1,1]$ 区间,消除不同物理量纲带来的权重失衡问题,将静态地形因子作为固定通道与逐日气象时序数据拼接,构建维度匹配的二维时空输入张量。

2 改进 CNN-LSTM 模型整体架构设计

传统基础 CNN-LSTM 网络流程为单层卷

积特征提取、池化降维、单层 LSTM 时序拟合,存在卷积特征浅层化、长时序梯度衰减、小样本泛化差三重缺陷。本文搭建四层递进式改进网络,整体逻辑链路为:多源时空二维张量输入→残差多尺度 1D-2D 混合卷积模块→全局平均池化特征融合→因果多头注意力加权层→双层正则化 LSTM 时序编码层→Dropout 正则约束→全连接输出层生成日径流预报值。全程不引入图表,通过文字完整阐述各模块改进机理与计算逻辑。

2.1 残差多尺度混合卷积改进模块

常规一维卷积仅提取单尺度气象时序特征,无法区分山区不同坡度、海拔下局地产流差异。本文融合一维时序卷积、二维空间卷积构建并行残差卷积块,设置三组尺寸 3、7、11 的卷积核同步运算,分别捕捉单日暴雨短周期响应、月度植被季节变化、年度地形空间分异三类多尺度耦合特征。引入残差跳跃连接机制,将卷积模块原始输入直接叠加至卷积输出,解决深层网络反向传播梯度消失问题;将标准 ReLU 激活函数替换为 ELU 激活,缓解负输入神经元失活现象,提升低温、无降水等低径流时段特征提取稳定性。

2.2 因果掩码多头注意力机制优化

通用注意力模块在权重计算过程中会读取预报时刻之后的未来时序数据,违背水文实时预报因果逻辑,模型实际业务落地存在显著缺陷。本文引入因果掩码多头注意力结构,通过掩码矩阵屏蔽预测节点之后全部时序特征,保证仅依靠历史气象、地形信息推演未来径流,适配山洪实时预警场景。网络设置 4 头注意力并行运算,四类注意力头分别对应降水、积雪、气温、地形特征组,自适应分配特征权重,自动放大汛期连续暴雨、春季融雪等对径流贡献度高的输入变量权重,弱化枯水期平稳低影响气象因子,显著提升极端洪水场次模拟精度。

3 模型模拟结果与多层次精度分析

3.1 完整干流流域全域模拟精度对比

基于干流 2010—2023 年完整观测测试数据集,四类模型核心评价指标量化结果如下:BP 神经网络 $\text{NSE}=0.774$, $\text{PBIAS}=11.92\%$, $\text{RMSE}=6.18\text{m}^3/\text{s}$,洪峰平均相对误差 23.6%;单层 LSTM 模型 $\text{NSE}=0.775$, $\text{PBIAS}=9.65\%$, $\text{RMSE}=5.43\text{m}^3/\text{s}$,洪峰平均相对误差 19.7%;基础 CNN-LSTM 模型 $\text{NSE}=0.828$, $\text{PBIAS}=7.18\%$, $\text{RMSE}=4.10\text{m}^3/\text{s}$,洪峰平均相对误差 15.3%;本文改进 CNN-LSTM 模型 $\text{NSE}=0.916$, $\text{PBIAS}=4.31\%$,

RMSE=2.75m³/s, 洪峰平均相对误差 3.6%。全域模拟数据清晰表明, 本文改进架构各项指标全面优于三类传统对照模型, 相较于基础 CNN-LSTM, NSE 提升 0.088, RMSE 降低 1.35 m³/s, 洪峰模拟误差大幅缩减。内在机理在于残差多尺度卷积完整挖掘山区地形与气象耦合的空间异质性产流特征, 因果多头注意力定向强化暴雨、融雪关键时段特征权重, 从源头修正基础模型极端流量系统性低估问题; ELU 激活搭配残差跳跃连接优化深层网络特征传递效率, 减少多维输入信息丢失。模型 PBIAS 仅 4.31%, 预报径流不存在明显高估或低估的系统性偏差, 整体拟合稳定性更强。

3.2 缺资料支流短序列场景适应性分析

选取上游 3 条典型无长期径流观测支流开展三组对照试验, 区分有无迁移学习、数据增强策略验证优化效果: 第一组不启用迁移学习、不做数据增强, 仅依靠原始 287 组短样本训练, 模型 NSE 仅 0.681, 洪峰相对误差 19.1%, 过拟合现象严重, 汛期洪水过程线完全失真; 第二组仅开启双重时序数据增强, 不引入迁移训练, 扩充样本后 NSE 提升至 0.764, 洪峰误差降至 12.6%, 样本扩充能够一定程度缓解过拟合, 但缺少成熟流域产汇流先验知识支撑, 极端径流模拟偏差依旧突出; 第三组完整采用迁移学习 + 双重数据增强优化方案, 即本文完整改进策略, 支流测试集 NSE 可达 0.840, 洪峰相对误差 7.3%, 峰现时间平均误差 0.82 h。试验结果证实, 迁移学习将相似干流流域成熟降雨-径流非线性映射规律传递至缺资料支流, 数据增强从样本数量与时序多样性层面弥补观测记录短缺短板, 两种策略协同作用彻底突破传统深度学习模型依赖海量连续径流样本的行业瓶颈。同条件下基础 CNN-LSTM 在缺资料支流 NSE 仅 0.751, 证明本文整套优化架构针

对山区小样本缺资料场景具备极强适配性。

4 结论

针对国内山区缺资料流域监测样本稀缺、地形空间异质性强、传统深度学习模型极端径流模拟失真、小样本泛化能力不足等多重难题, 本文构建融合残差多尺度混合卷积、因果多头注意力、相似流域迁移学习、双重时序数据增强的一体化改进 CNN-LSTM 水文预报模型, 以秦岭北麓黑河上游典型山区流域完成实例模拟验证, 得到核心结论如下: 第一, 残差 1D-2D 混合卷积模块能够分层提取山区地形、气象、植被多尺度耦合产流空间特征, 因果掩码多头注意力定向强化暴雨、季节性融雪等关键水文过程时序权重, 双层正则化 LSTM 精准捕捉径流长期滞后响应关系; 相较于基础 CNN-LSTM 模型, 完整观测流域日径流预报 NSE 提升 0.088, 极端洪水洪峰模拟误差降低 11.7%, 空间与时序特征提取能力实现双重优化。第二, 迁移学习搭配时序双重数据增强的联合优化策略, 能够显著提升无长期径流观测缺资料子流域模型泛化性能, 支流短序列测试场景下模型 NSE 可达 0.840, 从先验知识、样本扩充两个维度解决深度学习依赖海量径流观测数据的行业痛点。第三, 改进模型对极端暴雨山洪场次模拟精度优势突出, 洪峰流量、峰现时间模拟误差远低于 BP 神经网络、单层 LSTM、基础 CNN-LSTM 三类传统模型, 可直接应用于山区中小流域山洪预警、流域水资源动态调度工作。后续研究将耦合融雪、土壤下渗水文物理方程构建物理引导深度学习框架, 融合雷达高分辨率降水产品校正空间输入偏差, 拓展小时尺度短时山洪预报功能, 进一步提升复杂山地流域径流模拟可靠性与业务实用价值。

参考文献:

- [1] 谈戈, 夏军, 李新. 无资料地区水文预报研究的方法与出路 [J]. 冰川冻土, 2004, (2): 192-196.
- [2] 钱进等. 水动力水质模型研究进展综述 [J]. 人民珠江, 2025, 46 (10): 56-64.
- [3] 孙龙; 齐晓静. 海河平原浅层地下水位时空演变规律研究 [J]. 水文, 2026, 46 (02): 56-63.
- [4] 谢红娟, 杨彩虹. 基于 GIS 的湖南省水土保持服务功能重要性时空分布特征 [J]. 水土保持通报, 2024, 44 (06): 248-255.
- [5] 李强; 赵钢铁. 基于大语言模型的洪涝灾害统计调查 [J]. 水利水电技术, 2025, 56 (09): 60-75.

作者简介: 付丽娜 (1990.06—), 女, 满族, 内蒙古自治区赤峰人, 大学本科, 工程师, 研究方向: 水文监测。

杨腾 (1990.12—), 男, 蒙古族, 内蒙古自治区阿拉善盟人, 大学本科, 工程师, 研究方向: 水文监测。