

文献调研：山洪灾害监控与风险预测

苏锦华 2021103740 chasesu@ruc.edu.cn

January 4, 2022

Abstract

山洪灾害是我国最为严重的自然灾害之一，山洪灾害空间分布频度变化大的特性，地面监测站点仅能监控有限流域，难以满足山洪灾害全空间区域的监测。随着机器学习等人工智能技术的广泛应用，基于人工智能技术和大数据驱动新一代水文模型成为水文学研究的热点。笔者总体收集了 500 余篇 AI 前沿文章，主要覆盖应用地质学、水文分析、遥感视觉 AI、物理驱动 AI 的模型与技巧，涉及从山洪预警到损失估计、从多源数据到数据缺失的全方位解决方案。本文梳理了山洪监控与预警、山洪风险评估与损失估计、小流域场景下多源数据协同的山洪监控相关方向的研究现状和发展趋势，并针对小流域数据缺失场景下提出了一种构思方案：结合基于强化学习 + 变分推断的山洪模拟的数据增强，在少量标注数据的情况下即可对无资料流域进行稳健预测；原理是采用迁移学习和半监督学习的方法，来借助全空间区域检测的遥感图像，以攫取有资料流域（Gauged Basins）的气象数据和水位信息。最后本文讨论了尚未解决的关键问题、技术难点和物理 AI 对该方向研究的未来影响。

1 背景与问题

1.1 问题背景

山洪灾害是我国最为严重的自然灾害之一，严重地威胁着人民生命财产安全，给国民经济造成重大损失。传统的山洪灾害监测方法分为地面观测和遥感监测两种，由于山洪灾害空间分布具有多发、少发和不发等频度变化大的特性，所以设立的有限地面监测站点只能代表局部地点的信息，缺乏宏观性和代表性，难以满足山洪灾害全空间区域的监测。随着机器学习等人工智能技术的广泛应用，基于人工智能技术和大数据驱动新一代水文模型成为水文学研究的热点。将人工智能方法与气象降水和水文监测产品以及遥感产品相结合应用于暴雨洪水灾害监测与损失评估成为了山洪灾害防御的核心技术难点。因此，建立山洪灾害识别与损失评估系统在山洪灾害监测具有重大意义。

1.2 问题与目标

对山洪灾害的损失进行评估。依托水文模型确定山洪灾害致灾临界降雨量，利用逐日的河流水位和流量、逐小时降雨数据对水文模型进行参数率定，预报子流域（或格网点）的径流过程、洪峰等洪水特征，根据流域河流防洪标准，发布洪水预警，从而推算出河流达到警戒水位、保证水位的流域临界面雨量，建立小流域山洪风险评估模型。依据气象预报相关产品进行流域面雨量预报，对山洪灾害的可能发生的强度和影响范围进行监测和损失评估。

2 重点概念

2.0.1 无资料流域 (Ungauged Basins)

山洪全方位监控的核心难点在于监测无资料流域。无资料流域 (Ungauged Basins) 指无地面监测站的小型流域,但其本身并不是孤立的,而是与上下游共同满足流量方程,满足水文模型的约束。‘如何利用有资料流域 (Gauged Basins) 的气象数据和水位信息,借助全空间区域检测的遥感图像,来对无资料流域进行监控并实现对山洪级别的重大灾害进行预警,损失估计’是核心问题。

2.0.2 山洪预警系统

山洪 [?] 是最危险的自然灾害之一,因为它们固液混合且急流流动。泥石流发生速度快、固液混合的特点冲击力巨大,监测主要依靠于传感器网络,主要分为事先预警和事件监控。事先预警系统会利用气象数据在泥石流发生前对此做出预警,事件监控在泥石流发生的极短时间内做出判断 [?]. Li 因此指出 [?] 山洪灾害现场建设的预警作用与影响,预警系统仍是减少山洪损失最重要的设备系统。

2.0.3 物理驱动人工智能

深度学习依赖于数据驱动方法,所有与物理模拟背景下的任务,特别是水文问题,变量之间满足物理条件的约束。研究物理损失约束、更紧密耦合的可微模拟学习算法,以及强化学习和不确定性建模,是物理驱动人工智能的核心。

2.0.4 强化学习水文模拟

基于强化学习 (RL) 的实时控制方法,已经成为一种最先进的自主控制方法。在人工智能领域。使用深度神经网络控制的系统交互来学习控制策略-有效地尝试各种控制策略,直到收敛到那些达到期望的目标。这在水文领域中模拟复杂的水动力模型具有优势,雨水测绘站点实时反馈的数据用于训练 RL 算法,可以非常有效地有效控制和指导分布式站点对于洪水和洪峰流量的模拟预测。

3 国内外研究发展趋势

笔者总体收集了 500 余篇应用水文研究与 AI 技术的前沿文章,主要覆盖应用地质学、水文分析、遥感视觉 AI、物理驱动 AI 的模型与技巧,涉及从山洪预警到损失估计、从多源数据到数据缺失的全方位解决方案。这一章节主要概况性地介绍研究热点和研究趋势,涉及的模型方法将在下一章节重点介绍。

3.1 山洪监控与预警

对于预警系统来说,降水控制指数 (rainfall control index) 是关键的监控指标,地形地质以及流域变化可以共同构筑综合预警指数,[?] 针对山东山洪数据给出了具体方案。Zhao[?] 以山东省临朐县为例,分析了山洪发生与实际降水,结合水位给出了一种预警雨量的设置原则。

对于事先预警系统来说,多源数据指导预测是需要的,降水量、水文监控、滑坡位移监控组成的传感器网络构成了物联网 [?],从物联网优化的角度提升事先预警系统的预测效果与预测延迟 [?]. Jiang.[?] 介绍了云南山洪防范总体布局,同时讨论了现有房屋结构建设在防范山洪中重要性。

当然,事后监控也是有价值的,泥石流发生后的传感器数据用于山洪受灾分析,总结洪水灾后受损评估的统计方法 [?]. 长时间降水、地质地形以及人类活动是造成山洪的前三大影响因素,近年人类活动在加剧山洪的破坏性 [?].

预警系统本来是复杂功能,除了关键事件的识别,稳定运行、软件架构与效率、能源与升级、数据采集源等多个侧面均有相关研究与探讨。Zhang.[?] 在分析了全中国山洪发生的分布特征后,认为在抵御山洪中,非结构建设措施应占主导地位,房屋结构建设为辅。Guan.[?] 认为利用遥感和信息系统可以全局的监控山洪受灾情况,为山洪风险评估与决策提供有力武器。Ji.[?] 利用 webGIS 开发一套公布山洪发生风险指数的 GIS 应用,重点探讨了系统架构、数据库设计等技术细节,同时对预测效果进行了相应评估。Jamil.[?] 则重点设计了实时预警系统,重点关注降水自动监控、水位自动监控以及预测预警的联动系统。Wenchuan.[?] 讨论了山洪技术的发展脉络,突出成因分析、预测、预警和灾情分析。Jiang.[?] 对中国 1987-2013 年减灾应急相关工作做了总结,其中减灾救急经验让中国减灾救援能力提升了一个台阶。Guo.[?] 则重点谈及中国山洪防控设备的升级换代,能稳定监控降水与水位,防止能源切断的情况发生,为山洪救灾提供可靠支撑。Mizuyama.[?] 利用无人机检测水位监测缺失区域水位和检测情况。

人工智能在水文过程中扮演着日益重要的角色 [?],其中,深度学习和机器学习在水文过程的建模和预测方面有更好的效果,可以实现更高的精度、鲁棒性、效率、计算成本和整体模型性能 [?]. 传统的水文分析模型包括物理模型和统计模型,比如全球水文模型 (GHMs)、时间序列模型等对洪水模拟的准确性还不足以满足大多数应用,而基于机器学习算法的数据驱动技术在水文建模中越来越流行,特别是在预测方面 [?]. 于是很多学者基于机器学习算法对模型进行改进。Schmidt.[?] 采用人工神经网络和随机森林两种流行的机器学习方法来分析德国的洪水事件,发现机器学习模型比线性回归有更高的预测精度,Noymanee.[?], Ko.[?] 同样使用机器学习技术更准确的预测水文影响;遥感可以更精确的估计洪水面积的时间变化及其与降水和蒸散的季节性相关的水文模式,也有学者将遥感与机器学习模型结合, Dona.[?] 利用遥感和机器学习模型监测临时湖泊的水文模式。

目前的水文建模方法通常依赖于基于物理或数据科学的方法。而只基于物理的模型倾向于刚性结构,在某些情况下会产生不切实际的参数值,只基于 ML 算法会忽略物理过程的约束建立输入-输出关系。虽然有一种观点认为,物理模型能够更好地理解过程,ML 算法表现出更好的预测能力,但不增加预测能力的科学知识可能具有欺骗性。因此,需要一种混合建模方法,以耦合 ML 算法和基于物理的模型。于是 Zaherpour[?] 构造了全球水文模型 (GHM) 模拟的优化多模型组合 (MMC),并使用来自全球 40 个大型集水区的 5 个 GHM 的径流模拟来说明 MMC 方法有更好的表现。Bhasme.[?] 也构建了一个物理信息机器学习 (PIML) 模型,该模型将概念水文模型的过程理解与机器学习模型的预测能力相结合,结果表明,PIML 模型的性能优于纯概念模型和 ML 算法,因此将概念模型结构与 ML 算法相结合的混合方法可用于提高关键水文过程的预测精度,这对洪水风险评估是至关重要的。Yaseen.[?] 将经典的极限学习机 (ELM) 训练算法和 Salp 群算法 (SSA) 混合,对月河流流量进行预测,SSA-ELM 的河流流量预测精度优于经典 ELM 和其他 AI 模型。Papacharalampous.[?] 使用 GR4J 水文模型获得点水文预测,并将其用作分位数回归设置中的预测变量。

3.2 山洪风险评估与损失估计

将人工智能方法与气象降水和水文监测产品以及遥感产品相结合应用于暴雨洪水风险评估和损失估计成为了山洪灾害防御的核心技术难点。损失的估计是支持山洪发生后实施救援顺序的决策依据,没有山洪就没有对应的损失,数据的缺少导致山洪发生的规律不够清晰。保险是降低山洪等突发灾害造成损失的良好手段,保险精算厘定费率依赖于对区域洪水发生概率的预测与估计。损失估计与风险评估相辅相成,目标分别是利用模型算法尽可能逼近实际损失的观测值和损失的数学期望。

1995 年起至今,中国学者开始借鉴国外测绘经验,对山洪风险进行评估。Li[?] 对北京山区的山洪发生概率进行了风险预测与划分,开展了山区山洪分类及其危险区测绘技术研究,制定了适合北京的急流分类和危险区测绘技术的细则和规定。Tang.[?] 是在考虑山洪受灾风险时将经济信息考虑进损失估计,展示了区域化的山洪风险。Zheng[?] 曾对山洪灾害风险进行了度量,通过拓扑结构对大小流域关系进行建模,并通过灰色系统理论构建风险评价分,通过文登盆地介绍了文登盆地调查的结果和结论。Xiu[?] 使用随机森林模型,对中国江西省山洪发生风险进行预测,江西省的高风险区域主要是分布在降雨量高的山区附近,暴雨大,农田生产潜力大。

山洪损失评估是水利工程防洪评估、洪水风险测绘和洪水保险索赔的重要组成部分。如何科学、准确地评估洪灾损失一直是洪灾研究的重点和难点。Shen.[?] 从灾害分类、洪灾损失评估指标体系、洪灾损失评估方法三个方面阐述了洪灾损失评估的研究现状,并对目前评估规模、评估指标体系、评估方法等方面存在的问题进行了探讨。[?] 回顾与总结了山洪灾害中建筑物的物理脆弱性评估方法,将评估方法分为三类:脆弱性矩阵、脆弱性曲线和脆弱性指数。Duan.[?] 统计了 1950 年以来中国山洪对社会经济造成的损失,揭示了洪水相关灾害的趋势和中国的社会经济损失的关联关系。Totschnig.[?] 通过风险要素脆弱性评估,对山洪风险中的单元损失程度和空间分异规律进行分析。不同于大多数研究使用过程强度来粗略估计脆弱性值,Fuchs.[?] 设计了量化评估建筑物损毁的函数方法。Papathoma.[?] 也设计了一套针对山洪、滑坡的房屋损毁评估工具。Xiao.[?] 山洪灾害给山带来更为严重的水土流失,对损失估计需要纠偏。Ze.[?] 利用统计资料对凤山县包括山洪在内地质灾害进行了分析。

3.3 小流域场景下多源数据协同的山洪监控

流域监控站点可以收集水域数据,依赖于传感器的发展和无线网络技术的发展。Guesmi.[?] 讨论了传感器的无线物联网在山洪监测中的应用与早期监测。Jun.[?] 讨论了山洪的地理空间分布规律,定量分析山洪灾害与各因子之间的数量和概率相关关系。基于 ArcGIS 将广东省划分为 179 801 个网格,按网格分别统计每个因子中山洪灾害的数量和发生概率。结果表明:相对高差、坡度、水系密度与山洪灾害发生概率之间的相关关系可以用 $y=a_1e^{b_1x}+a_2e^{b_2x}$ 模拟;地层岩性、多年平均 24h 降雨和距水系距离与山洪灾害发生概率之间的相关关系可以用 $y=ae^{bx}$ 模拟。Ou.[?] 讨论了在山区和复杂的地形和地质条件下,山洪的发生过程。Shu.[?] 建立建全降雨监控网对山洪预警具有重要意义。Li.[?] 由于水位和降水的密切绑定关系,中国已经在利用降雨数据对水位缺失区域进行山洪监测预警。

4 基准技术方法

上一章节对山洪识别与风险预测的相关概念进行了讨论,以研究者的关注热点和发展趋势为主,本章节将影响力较大和效果较大的技术方案进行介绍,以供后续研究者选取合适的基准方法

来进行对比。

4.1 监控与预警山洪

山洪预测的最简单清晰是多点时序预测,即预测多个监控站点水位的变化序列,关键技术是时序预测和水文拓扑结构的利用,主流方法基本引入了循环神经网络和水文物理规则。Atiquzzaman.[?] 将极端学习机 (ELM) 应用于水文流量序列建模,并将其性能与 GP 和基于进化计算的支持向量机 (EC-SVM) 进行比较。Kratzert.[?] 提出了一种 LSTM 的变式 (EA-LSTM),允许深度学习模型将流域相似性作为特征层进行学习。Ji.[?] 使用 LSTM 模拟中亚天山两个数据稀疏冰川流域的日流量,效果取得显著提升。Yang.[?] 试图解决全球水文模型 (GHMs) 中洪水模拟的准确性不足的问题,采用 LSTM 来改进基于 GHMs 的洪水模拟。将经典洪水模拟和机器学习技术相结合,得到一种更加稳健和自信的洪水风险评估方法。Yaseen.[?] 引入现代非调谐机器学习数据驱动方法 (极限学习机 ELM) 来解决流量预测问题,与人工神经网络 (ANN) 模型进行比较发现 ELM 方法在大部分区域预测精度上优于 ANN 模型,耗时也更低。Kan.[?] 将 ANN 与 K-最近邻法相结合,提出了一种用于洪水预报的新型混合机器学习 (HML) 水文模型。Tewelde.[?] 耦合机器学习和可接受极限方法在分布式水文模型的参数识别中,取得了较好应用效果。Yang.[?] 使用变形测试的方法比较不同水文机器学习模型的可靠性。Zhou.[?] 首次将单调复合分位数回归神经网络 (MCQRNN) 用于洪水过程的概率密度预测,同时提出了一个综合模型 (XAJ-MCQRNN),结合了 XAJ 和 MCQRNN,以克服多步超前洪水概率密度预测中遇到的误差传播和累积现象,同时有助于减轻与极端洪水和降雨事件的不确定性相关的影响,并提高洪水预报和预警的准确性和可靠性。

4.2 损失与风险

随机过程、时间序列模型等统计学模型一直被应用与风险和损失的评估,针对该任务的具体情形设计专门化的模型仍然具有竞争力。Yang.[?] 利用一个多准则决策过程,通过结合概率统计和模糊集理论,使得云模型 (Cloud Model) 可以帮助定性概念和定量数据之间进行必要的转换。在多准则决策过程中,采用层次分析法 (AHP) 和熵权相结合的混合权重法确定指标权重。结果表明,基于云模型的方法不仅能够表征山洪灾害的风险水平,而且能够表征相同风险水平下的相对风险概率。建议的研究为未来流域洪水风险管理提供了指导,并扩展了现有风险评估方法的范围。

现近流行的机器学习树结构模型近三年已经被广泛地用于风险预测、损失评估,本研究也将以树结构模型作为基准线。Xiu.[?] 选择中国江西省作为案例研究,该省经常遭受山洪灾害,选取 14 个洪水风险评估指标和随机森林算法构建山洪灾害风险评估模型。Ma.[?] 极端梯度增强 (XGBoost) 是一种优秀的集成学习算法,在许多领域取得了显著的成果。它不仅优化了算法,而且自动应用 CPU 的多线程进行并行计算,从而大大提高了模型的训练速度和预测精度。

深度神经网络、迁移学习、半监督学习等 AI 前沿技术已经在山洪损失分析中得到了初步应用,相关研究收到关注,相关技术瓶颈和上限尚有待进一步挖掘。Zhang.[?] 从分析山洪灾害机理和属性特征入手,提出了基于粗糙集的 RBF 神经网络集成洪灾损失评估模型,并将基于神经网络集成的洪灾损失评估模型应用于洪灾损失预测,效果出众。Kim.[?] 基于混合机器学习 (无监督聚类技术和监督分类技术) 的水文评估工具经证实,性能评级的一致性率为 98%。Gong.[?] 以雅安雨城为研究对象,充分利用多源遥感信息来反映雅安的气象和降水状况,通过选取降雨量、径流深度、降雨强度、降雨强度等指标,构建了评价指标体系,从致灾因素和孕灾环境两个方面分

析了流域的库容、高程、坡度和水系。利用 BP 神经网络对历史洪水发生地点进行训练, 得到各研究指标的权重, 对各指标进行叠加分析, 得出雅安市洪涝灾害分布图。

4.3 小流域场景下数据协同

小流域场景下数据缺失是主要问题, 但水文信息属于整体性系统, 学者主要从水文物理模型来利用周围水文信息实现对数据缺失区域的监控, 但近年也涌现出利用神经网络等模型的泛化性和相关迁移学习技巧来实现协同监控的技术方案。Potdar.[?] 探讨了洪峰流量、降雨变异性和流域地形之间的一阶相关性, 并使用多维统计建模方法揭示了这些复杂的相互作用。在不同的机器学习技术中, XGBoost 用于通过变量重要性分析确定影响峰值流量的重要地形和降雨特征。创建了一个具有低偏差和方差的简约模型, 可用于未来的山洪爆发预测。结果证实, 尽管流域内降雨的空间组织对流域响应有重大影响, 但流域地形是洪峰流量的主要驱动因素。Zhu.[?] 采用了八个大气环流模型 (GCM) 和两种基于人工智能的方法 (支持向量回归、SVR 和极限学习机 (ELM)) 的集合, 以建立对气候变化的历史径流响应, 发现基于人工智能的 SVR 和 ELM 方法在预测未来水文响应方面表现出非常好的性能。结果还表明, 在类似山区流域的局部区域模型性能更优。Ragetti.[?] 使用决策树学习来探索参数集在整个流域描述符空间中的可转移性, 建立了中国十个省份 35 个集水区的半分布式降雨径流模型。在一系列遗漏测试中, 评估了模型在假定无资料集水区检测洪水事件的能力。Zhang.[?] 利用径向基神经网络来对村庄、小镇等小流域区域进行损失预测。Ouali.[?] 使用统计和机器学习方法研究水文变量 RFA 中与非线性相关的方面, 使用人工神经网络 (ANN) 和广义加性模型 (GAM) 与基于非线性 ANN 的典型相关分析 (NLCCA) 程序相结合, 以确保对所涉及的复杂过程进行适当的非线性建模。结果表明, 完全非线性模型是最合适的。

5 方案提出与讨论

由上述文献研究可以发现, 物理驱动的研究逐渐成为热点, 但与小流域场景的结合还不够, 笔者试图结合最新的相关山洪 AI 研究, 给出小流域场景下的解决方案。

5.1 基于物理 AI 驱动的小流域预警模型

笔者基于 Google 团队提出的 Hydronets[?], 发展出基于中国水文灾情数据的模型:

将水文区域 \mathcal{R} 定义为有向图 $\mathcal{R} = (\mathcal{B}, \mathcal{E})$, 其中节点集中的每个节点 $\mathcal{B} = \{b_1, \dots, b_n\}$ 表示一个流域和有向边, $b_i \rightarrow b_j \in \mathcal{E}$ 表示 b_i 是 b_j 的直接子流域。边缘方向表示从子流域到其包含流域的水流, 并且每当 $b_i \rightarrow b_j$ 存在, 就代表着 b_i 是 b_j 的来源 (可以有多个来源), b_j 是 b_i 的下游节点。向外度为零的流域称为区域排水沟。没有来源 (度数为零) 的流域称为区域来源。对于每个流域 $b \in \mathcal{B}$, 用 $S(b) \subseteq \mathcal{B}$ 表示 b 的来源集合。对于每个流域 b_i , 考虑它的时间特征的序列, $\mathcal{X}_i^{1:t} \mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}, \dots, \mathbf{x}_i^{(t)}$, 其中 $\mathbf{x}_i^{(t)}$ 是时间 t 的特征向量, 它可以包括特征信息, 例如降水、温度等。特定盆地 i 还包括一个静态特征向量 z_i , 可包括土壤类型、高程等信息。对于每个盆地 b_i , 将 $\mathbf{y}_i^{1:t}$ 作为其目标标签序列。通常, 目标标签为水位或流量 (即水的体积流量)。给定所需的预测层位 h (例如, 两天), 要为区域 \mathcal{R} 创建模型 $F_{\mathcal{R}}$, 该模型可根据过去长度为 T 的输入窗口准确预测层位 h 处所有流域的目标标签, $F \left(\mathcal{X}_1^{(t-T:t)}, \dots, \mathcal{X}_n^{(t-T:t)}, \mathbf{z}_1, \dots, \mathbf{z}_n \right) \rightarrow (y_1^{t+h}, \dots, y_n^{t+h})$ 在水文预测中, 传统上使用纳什-萨特克利夫效率 (NSE) 来衡量预测质量, 这相当于经典统计的 R^2 (“解释方差”)。在这项工作中, 还将使用 R^2 -persist 衡量预测质量。

构建了新的水文预报结构族。该族中的模型称为水网，利用河流结构提供的先验信息，给定一个水文区域 $\mathcal{R} = (\mathcal{B}, \mathcal{E})$ ，水文网跨越一个区域 \mathcal{R} 的计算图，使得每个流域 $b_i \in \mathcal{B}$ 在河流图中，网络 $H(\mathcal{R}) (H_1, \dots, H_n)$ 包含一个子网络（也称为节点） H_i 。 H_i 对应到 H_j 当且仅当 $(b_i \rightarrow b_j) \in \mathcal{E}$ 。每个节点 H_i 由三个子模型组成，其中两个子模型是特定流域，第三个子模型存在于在所有流域。在每个节点 b_i 中，共享模型输出一个时间嵌入向量，对该流域信息和上游信息编码。此外，利用该嵌入，流域特定模型在流域 b_i 处输出目标标签（如水位）。除区域排水流域外的所有节点都将其时间嵌入信息传递给其下游节点。定义 $K \left| E_i^{(t)} \right|$ 为每个嵌入向量的大小

$$\begin{aligned} F_i^{\text{CMB}} \left(\left\{ E_j^{(t-T:t)} \mid j \in S(b_i) \right\}, \mathbf{z}_i \right) &\rightarrow C_i^{(t-T:t)} \\ F^{\text{SHA}} \left(\mathcal{X}_i^{(t-T:t)}, \mathbf{z}_i, C_i^{(t-T:t)} \right) &\rightarrow E_i^{(t-T:t)} \\ F_i^{\text{PRD}} \left(E_i^{(t-T:t)} \right) &\rightarrow l_i^{(t+h)} \end{aligned}$$

5.2 基于物理 AI 驱动的小流域损失评估模型

充分利用多源遥感信息来反映雅安的气象和降水状况，通过选取降雨量、径流深度、降雨强度、降雨强度等指标，构建了损失评价指标体系，从致灾因素和孕灾环境两个方面分析了流域的库容、高程、坡度和水系所带来的风险大小。其中结合遥感和水位，充分利用深度学习方法具有重大挑战性，基于物理 AI 以及强化学习模拟的水文模型 [?] 提出了如下的模型框架：

物理分布水文模型在大型流域水文模拟中是有效的，但水文特征的复杂特性限制了其应用。为了在实践中有效地管理水资源，需要一个易于使用和高效的水文模型。基于机器学习 (ML) 的模型有可能提供气象预报和水文响应之间的快速映射路径，而无需详细描述相应的物理过程。然而，广泛的数据需求、对空间变异性的忽视以及极端流量的低性能限制了 ML 模型的应用。本研究试图通过将基于物理的分布式水文模型与人工神经网络 (ANN)、计算机视觉 (CV) 和分类方法 (CA) 相结合，开发基于 ML 的水文模型。为了解决训练不足的问题，使用物理分布式水文模型 (GBHM) 和随机降雨发生器生成大量的合成数据 (GBHM-ANN)。为了改进极端流模拟，在 GBHM-ANN (GBHM-ANN-CA) 中加入了分类方法。为了捕获预测因子的空间变异性，还使用基于局部二元模式的计算机视觉方法来形成 GBHM-ANN-CA-CV 模型。综合案例研究表明了这三种建模方法的有效性。最后，使用泰国湄南河上游盆地的真实数据评估 GBHM-ANN-CA-CV。结果表明，在数据有限的流域，新的数据驱动模型的预测精度大大提高。具体而言，CV 提取的空间信息可以提高数据驱动水文模型的鲁棒性，CA 可以极大地改进高流量模拟。对于长期的日径流模拟，组合模型具有令人满意的精度。这项研究证明了基于 ML 的水文模型在水资源管理中的潜力，特别是在不断变化的环境中。

本方案试图构建一个物理过程和机器学习混合水文模型 (GBHM-ANN-CA-CV) 进行水文模拟，建模框架如图 2 所示。该模型主要由四个主要步骤组成。第一步是数据预处理，包括输入变量选择、训练数据生成和空间特征提取。由于该模型由 GBHM 生成的合成数据驱动，因此使用相同的输入。模型输入由三类变量组成，包括气象变量、植被参数和集水区地貌信息。气象输入包括通过 Penman-Monteith 方程计算的日降水量、日平均气温、日相对湿度和日潜在蒸散量。植被输入包括被线性插值为日分辨率的 LAI。集水区地貌输入包括从 DEM 导出的坡度、坡度长度和高程信息。关于模型输入的详细信息见表 2。确定模型输入后，使用 SRG 和 GBHM 生成用于模型开发的基于物理的合成数据。利用合成数据集，通过 LBP-CV 提取输入变量的空间特征，并将其串联形成 ML 的特征空间。为了避免模型过度拟合，采用主成分分析 (PCA) 进一步降低特征空间的维数。

5.3 技术难点与讨论

本方案可以理解为国内特色的 HydroNet，其核心更是使用了在 AI 领域一统 NLP 和 CV 的 Transformer。结合基于强化学习 + 变分推断的山洪模拟的数据增强，在少量标注数据的情况下即可对无资料流域进行稳健预测，原理是采用迁移学习和半监督学习的方法，来借助全空间区域检测的遥感图像，以攫取有资料流域（Gauged Basins）的气象数据和水位信息。不仅仅对于无资料流域的预测有质的提升，其变分推断涉及的隐变量在深度学习的框架下引入统计学概率图模型不仅具有难得的解释力，对于任何形式的数据缺失都能够较好支撑。

1. 所有解决方案均同时考虑深度结合水文模型和深度学习方法，保证鲁棒性、极端情况依旧稳定的条件下追求更高的性能。并未一味地追求模型的参数量，在深度学习的基础上引入变分推断的概率图，与传统统计模型、物理模型的原有工作有衔接、有发展。

2. 对于无资料流域的数据，采用迁移学习和半监督学习的方法，来借助全空间区域检测的遥感图像，以攫取有资料流域（Gauged Basins）的气象数据和水位信息，方法上借鉴大量 AI 前沿方法。

3. 对于遥感级别的数据处理使用像素级别的语义分割方法，使得损失评估能够达到建筑物级别的精细度，得益于预训练和迁移学习，遥感模型在小规模图像上也具备很好的信息提取能力。