论文审稿意见： 论文标题：“Threshold Effects and Generalization Bias in AI-based Urban Pluvial Flood Prediction: Insights from a Dataset Design Perspective.” **作者：** Hao Hu 等

总体而言，引言部分撰写较好，提出的问题逻辑清晰且具有相关性。使用合成数据库确实存在优势：除了解决真实数据不足的问题外，它还能帮助我们克服导致过拟合的偏差-方差困境。

然而，在文章后续部分，我们发现方法论的部署存在明显不足：

1. 生成水文响应的物理基模型并未完整呈现或经过充分验证。这一点非常重要，因为如果物理模型无法充分表征案例研究的复杂性，那么整个研究的意义将大打折扣。

   回复：修订稿 2.1 节大幅补充了模型验证内容，包括 InfoWorks ICM 一维-二维耦合模型的率定过程、图 1-2（监测管道流量模拟与实测对比）、以及新增的验证表（3 场实测降雨事件的 NSE，峰值误差、径流体积误差均给出数值分析）。原版 2.2 节也保留了 NSE>0.5 的明确表述。模型参数率定（糙率、土壤条件）和验证指标均已完整呈现。

   Response: Section 2.1 of the revised manuscript has been substantially expanded to include detailed model validation. This includes the calibration process of the InfoWorks ICM 1D–2D coupled model, Figures 1–2 (comparison between simulated and observed pipe flow), and a newly added validation table presenting quantitative results for three observed rainfall events (including NSE, peak flow error, and runoff volume error). The original Section 2.2 retains the explicit statement that NSE > 0.5. The model parameter calibration (e.g., roughness and soil conditions) and validation metrics are now fully presented.

2. 模拟降雨极不现实，模拟降雨强度（100 mm/h）远高于物理模型所观测和应用的实际强度（1 mm/h，见图 7）。这些关键信息不应仅通过阅读图表才能获知。必须考虑其后果：一个用 1 mm/h 降雨校准的模型，能否可靠地模拟 100 mm/h 降雨的响应？后果如何？此外，所有模拟降雨事件持续时间完全相同，这是为什么？人们可以想象，更长的降雨事件由于总雨量更大，对模型的影响应更显著。因此，模拟降雨缺乏评估系统所有可能状态所需的多样性。

   回复：修订稿 2.1 节明确说明“采用芝加哥雨型，结合小区降雨历史统计特征，生成 10 场重现期典型降雨过程……固定降雨时长 30 小时，雨峰系数取 0.4”，并在图 3 中展示 10 场雨叠加曲线。同时新增说明：“当前设置更偏向可控制的比较……目的是隔离数据长度和雨强分布对模型性能的影响，避免历时变化引入额外混杂因素”。

   Response: Section 2.1 of the revised manuscript now clearly states that the Chicago rainfall pattern was adopted. Based on the historical rainfall characteristics of the study area, ten typical rainfall events with different return periods were generated. The rainfall duration was fixed at 30 hours, and the peak coefficient was set to 0.4. Figure 3 presents the superimposed curves of these ten rainfall events.

In addition, a clarification has been added: “The current setup is designed to favor controlled comparisons. The objective is to isolate the effects of dataset length and rainfall intensity distribution on model performance, while avoiding additional confounding factors introduced by variations in rainfall duration.”

3. 关于 LSTM 模型的选择，我们不应仅仅因为它“流行”就选用它，而应因为它真正适合当前问题。而且 LSTM 并非当前最常用的模型；对于高强度/快速洪水，感知机（perceptron）似乎更为常用，后者复杂度低得多，能耗也低得多，更有利于构建可持续的世界。 此外：
   回复：修订稿摘要明确：“采用长短期记忆网络（LSTM）作为统一的基准序列预测模型……因其在水文时序预测中广泛采用且性能得到验证”。原版 Highlights 和 2.4 节也强调“chosen for its widespread adoption and proven performance in hydrology”，并说明洞见对其他序列模型具有参考意义。
   修改思路：若需对比感知机，可在未来版本的讨论中补充一句：“与更轻量的 MLP 相比，LSTM 在捕捉长时依赖方面更具优势，本研究优先选择 LSTM 作为基准以保持与主流水文文献的一致性。”
   Response: The revised manuscript clearly states in the Abstract that a Long Short-Term Memory (LSTM) network is adopted as the unified baseline sequence prediction model, due to its widespread use and well-validated performance in hydrological time series prediction. The original Highlights and Section 2.4 also emphasize that it was “chosen for its widespread adoption and proven performance in hydrology,” and further note that the insights derived are informative for other sequence models as well.
   Revision approach: If a comparison with a multilayer perceptron (MLP) is required, an additional sentence can be incorporated in the discussion of a future version: “Compared with lighter MLP models, LSTM has a stronger capability to capture long-term dependencies. Therefore, this study prioritizes LSTM as the baseline to maintain consistency with mainstream hydrological literature.”

4. 关于 LSTM 模型的验证，其结果会因验证事件的选择而发生剧烈变化，因此强烈建议采用交叉验证，而本文并未进行。

回复：修订稿 2.3 节（2 3-2.5）完整引入“5 折交叉验证”策略，对每种数据集配置（长度、特征组合、雨量分级）独立进行随机分层抽样，报告 mean±std。
Response: Section 2.3 of the revised manuscript (Sections 2.3–2.5) fully introduces a 5-fold cross-validation strategy. For each dataset configuration (including data length, feature combinations, and rainfall categories), stratified random sampling is performed independently. The results are reported in the form of mean ± standard deviation.

以上所有要素表明，所采用的方法论并未经过充分思考。我们对研究结果的意义

表示怀疑。
在呈现质量方面，必须指出本文写作水平较差。多处内容重复（例如对降雨的描述），而另一些关键信息却缺失（如何评估泛化能力？）。文章中的符号表示前后不一致。即使是模型输出也未统一定义：有时是流量（L127），有时是淹没面积（L146），有时又是水深（L230）。这怎么可能？唯一合理的解释是，部分内容可能由 AI 生成，且未经过适当校正。
回复：已统一模型输出为“积水面积 Y”，并规范符号体系与指标定义；关于内容重复问题（降雨描述等），已对降雨生成及实验设置部分进行合并与精简，删除重复描述，并统一在方法部分集中说明降雨构建流程；关键信息缺失（泛化能力评估），已在实验设计中补充不同雨量分级训练–测试对比实验，并基于 NRMSE 与 R^2指标明确泛化能力评估方式。关于符号不统一问题，已统一符号体系，逐行检查并修正相关位置的符号错误与表述不一致问题，确保公式与文本表达一致，并在全文进行一致性修正。
Response: The model output has been standardized as “inundation area (Y),” and the notation system as well as the definitions of evaluation metrics have been unified. Regarding content redundancy (e.g., rainfall descriptions), the sections on rainfall generation and experimental setup have been consolidated and streamlined, with duplicate descriptions removed and the rainfall construction process consistently presented in the Methods section.
To address missing key information (i.e., generalization assessment), additional experiments have been incorporated to compare training–testing performance across different rainfall intensity categories. The evaluation of generalization ability is explicitly defined based on NRMSE and R^2 metrics.
Regarding inconsistent notation, the symbol system has been fully standardized. All related expressions have been carefully reviewed line by line, with errors corrected and inconsistencies resolved to ensure alignment between equations and textual descriptions, and overall consistency throughout the manuscript.

结论部分提供的新信息很少：是的，模型质量受数据库规模影响——我们早已知道；是的，训练模型必须包含不同类型和强度的事件——我们也早已知道，训练数据库必须覆盖整个状态空间；计算时间随训练数据库长度增加而增加，这也是显而易见的。另一方面，不对：从合成数据（即无噪声、无不确定性）中学习的模型不可能出现过拟合。因此，关于这一点的结论是错误的，必须从其他地方寻找解释。
遗憾的是，降雨事件的简化性质无法让我们探索其他途径或量化阈值，例如需要多少事件才能达到平台期。
尽管课题本身看似有趣，但鉴于上述所有问题，我很难推荐该论文发表。它需要彻底重新构思和重写。
回复：针对 1.数据集构建合理性不足 / 实验设计解释不清的问题；我们已在修订稿方法部分系统补充数据集构建流程，明确数据长度、特征组合及降雨分级的设计逻辑，并强调控制变量原则。2. 不同数据集配置之间缺乏清晰对比逻辑问题；
我们已通过多因素实验设计（数据长度 × 特征组合 × 降雨等级）构建完整对比体系，并在结果部分进行系统分析。3. 实验结果解释偏经验化；引入多因素

方差分析（MANOVA），对不同因素的显著性及贡献率进行定量评估，增强结果解释的统计支撑。4. 模型性能差异原因分析不足；已结合数据规模、特征复杂度及降雨分布差异，对 NRMSE 与 R^2变化趋势进行机制性分析。

Response:

1. **Insufficient justification of dataset construction / unclear experimental design:** The revised manuscript systematically supplements the dataset construction process in the Methods section. The design logic of data length, feature combinations, and rainfall categorization is clearly explained, with emphasis on the controlled variable principle.
2. **Lack of a clear comparison framework among different dataset configurations:** A comprehensive comparison framework has been established through a multifactor experimental design (data length × feature combinations × rainfall categories), and the results are analyzed systematically in the Results section.
3. **Overly empirical interpretation of experimental results:** A multivariate analysis of variance (MANOVA) has been introduced to quantitatively evaluate the significance and contribution of different factors, thereby strengthening the statistical support of the result interpretation.
4. **Insufficient analysis of the causes of model performance differences:** A mechanism-oriented analysis has been conducted by integrating data scale, feature complexity, and rainfall distribution differences to explain the variation trends in NRMSE and R^2.

具体意见

- 最大降雨强度的变异性无法代表真实情况。90 mm/h 至 140 mm/h 的范围本应更广泛地分布，例如从 10 mm/h 开始。

  回复：在修订稿中补充说明降雨强度范围基于设计暴雨（IDF 曲线）构建，旨在覆盖中高强度降雨情景，以突出城市内涝响应特征，而非刻画完整降雨频率分布。关于低强度降雨情景（如<20 mm/h），本研究未予以纳入，主要原因是在研究区域内该类降雨通常不会引发显著的城市内涝响应，因此对研究目标影响较小。

  Response: The revised manuscript clarifies that the rainfall intensity range is constructed based on design storms (IDF curves), with the aim of covering medium- to high-intensity rainfall scenarios to highlight urban inundation response characteristics, rather than to represent the full rainfall frequency distribution.

  Low-intensity rainfall scenarios (e.g., < 20 mm/h) are not included in this study. This is primarily because such rainfall events typically do not trigger significant urban inundation responses in the study area, and therefore have limited relevance to the research objectives.

- L128-135：降雨-径流模型的信息严重不足，特别是其响应时间是多少？我们希望看到该模型输出的水文过程线。Nash 效率系数为 0.5 可能掩盖了模拟结果：它既可能反映出能够再现水文系统动态的模拟，也可能是峰值

流量/水深极低的平庸模拟，无法代表实际观测情况。我们希望看到对强烈和极强烈降雨事件的模型输出进行更详细的分析。是否存在最强事件下的饱和现象？ 这很成问题，因为如果降雨-径流模型高估了洪水或过度简化了真实情况（水文系统的运行机制），那么问题就被大大简化了……这甚至可能导致对结论可靠性的质疑。

回复：已在修订稿中补充水动力模型的率定与验证说明，并增加多场降雨事件的模拟结果分析，以提高模型可靠性说明；同时对模型输出变量（如积水面积/流量）的来源进行了明确说明。Nash 效率系数（NSE）为 0.5 表明模型与参考结果之间具有中等一致性，在大尺度城市水文模拟研究中通常被认为是可接受的性能水平。针对高强度降雨事件，进一步分析了系统响应特征，以识别潜在的饱和现象及非线性水文过程。尽管模型存在一定简化，但其能够反映降雨–径流过程的主要动态特征，满足本研究中数据驱动模型构建与评估的需求。

Response: The revised manuscript includes additional descriptions of the calibration and validation of the hydrodynamic model, along with analyses of simulation results from multiple rainfall events, in order to improve the demonstration of model reliability. The sources of model output variables (e.g., inundation area and flow) have also been clearly specified.

A Nash–Sutcliffe Efficiency (NSE) value of 0.5 indicates a moderate level of agreement between the model and reference data, which is generally considered an acceptable performance level in large-scale urban hydrological simulations. For high-intensity rainfall events, further analyses of system response characteristics have been conducted to identify potential saturation effects and nonlinear hydrological processes.

Although certain simplifications exist, the model is capable of capturing the key dynamics of the rainfall–runoff process, and is therefore adequate for supporting the development and evaluation of the data-driven models in this study.

- L171 中序列长度与 LSTM 的 2 个时间步长（30 分钟）存在混淆。

  回复： 已在修订稿中区分“数据集序列长度（样本规模）”与“LSTM 时间步长（时间窗口）”两个概念，并分别进行定义说明。

  Response: The revised manuscript clearly distinguishes between “dataset sequence length (sample size)” and “LSTM time steps (time window),” and provides separate definitions and explanations for each concept.

- L137：q 被定义为降雨强度，这是不恰当的，因为 q 通常代表流量。

  回复：此处的符号 q 为芝加哥设计暴雨公式中的表达，并非流量。

  Response: The symbol q here refers to the expression in the Chicago design storm formula, rather than representing flow.

- L139：应明确指出公式对应的图号，而非“本公式”。

  回复：已在修订稿中将公式引用统一修改为明确编号形式，并对应具体图号或公式编号进行说明。

Response: In the revised manuscript, all formula references have been standardized to explicit numbering, with clear correspondence to specific figure numbers or equation numbers.

- L148：需明确“总降水量”指什么——是事件开始以来的累积量吗？
  回复：已在修订稿中明确“总降水量”为降雨事件全过程的累积降雨量，并在变量定义部分进行补充说明。
  Response: The revised manuscript clearly defines “total precipitation” as the cumulative rainfall over the entire duration of a rainfall event, and this has been explicitly clarified in the variable definition section.
- 方程 2 的问题在于时间并未出现。如果系统是真正动态且时间依赖的，必须在方程中体现，并更好地定义“总降水量”的含义。
  回复：在修订稿中对方程相关描述进行补充说明，强调模型输入为时间序列数据，从而体现系统的时间依赖性。
  Response: The revised manuscript supplements the descriptions related to the equations, emphasizing that the model inputs are time series data, thereby reflecting the temporal dependency of the system.

- L149：水文模型（称为“降雨-径流”）的输出实际上是淹没面积。这是一种与现实相差甚远的近似，如果不加以解释和论证，则不可接受。再次说明，模型描述过于简略。
  回复：已在方法部分补充说明：水动力模型输出的各要素通过空间统计转换得到地面积水面积，用作深度学习模型预测目标。
  Response: The Methods section has been supplemented to clarify that the outputs of the hydrodynamic model are transformed into surface inundation area through spatial statistical processing, which is then used as the prediction target for the deep learning model.

- L156-166：我们希望看到按实验类别划分的 Nash 效率系数。
  回复：
- L172-177：请明确以分钟为单位的持续时间以及洪水事件数量。
  回复：已明确数据时间步长（如分钟级）及每场降雨事件的持续时长，并补充不同雨型对应的事件数量说明。
  Response: The revised manuscript clearly specifies the data time step (e.g., minute-level resolution) and the duration of each rainfall event, and further supplements the number of events corresponding to different rainfall patterns.
- L178：有验证集是好的，但数据库是如何划分为训练、测试和验证集的？
  回复：全文增加样本随机性抽取和 k 折交叉实验验证后删除了测试集，按照统一的 8：2 比例（训练集：测试集）设计。
  Response: After introducing random sampling and k-fold cross-validation across the entire dataset, the independent test set has been removed. Model performance is now evaluated based on cross-validation results, ensuring robustness and reducing sampling bias, rather than relying on a fixed 8:2 train–test split.

- L172 至 L177 依次代表 5 场洪水、6 场洪水、7 场洪水……直至 10 场洪水。这是个细节，但在 90 mm/h 强度下，在城市环境中很难将这种降雨视为“轻度”。
  回复：已在修订稿中对降雨事件数量与分级方式进行了说明，明确不同洪水事件数量用于构建数据集规模变化，而非直接对应实际城市降雨频率；同时补充说明降雨强度基于芝加哥雨型设计暴雨（IDF 曲线）生成，用于覆盖不同降雨情景范围。
  Response: The revised manuscript clarifies the number and classification of rainfall events, explicitly stating that different numbers of flood events are used to construct variations in dataset size, rather than to directly represent actual urban rainfall frequency. In addition, it is specified that rainfall intensity is generated based on the Chicago design storm (IDF curves), in order to cover a range of rainfall scenarios.
- L202-203：我不理解两个时间步长的序列如何能帮助近似长期记忆。LSTM 在这里显然被误用，或者说，对于一个响应时间仅为 2 个时间步长的合成系统（前述信息缺失），LSTM 并不推荐。
  回复：已在摘要部分补充 LSTM 适用于时间序列建模的原因，并说明其在捕捉降雨–径流滞后关系中的优势。
  Response: The Abstract has been revised to include the rationale for adopting LSTM in time series modeling, highlighting its advantage in capturing the lagged relationship between rainfall and runoff.
- L209：请明确什么有助于避免过拟合？是批量大小、学习率还是训练轮次？过拟合通常通过正则化方法避免。但在合成系统上，先验不应存在过拟合，因为没有噪声和不确定性。
  回复：已在修订稿中对模型训练过程进行了补充说明，明确给出了训练轮次（epochs）、批量大小（batch size）及学习率等关键参数设置，并说明模型选择基于验证集性能，从而在一定程度上控制过拟合风险。同时，修订稿已弱化原文中对“过拟合”的直接表述，将性能下降更多归因于数据规模与特征复杂度之间的不匹配关系。
- Response: The revised manuscript provides additional details on the model training process, explicitly specifying key parameters such as the number of epochs, batch size, and learning rate. It also clarifies that model selection is based on validation performance, which helps mitigate the risk of overfitting to some extent.
- In addition, the revised version softens the original direct attribution to “overfitting,” instead explaining performance degradation more in terms of the mismatch between data scale and feature complexity.

- 表 1：请更精确地说明时间步长、序列长度和批量大小的单位。
  回复：已精确说明。

2.5.1 L220：为什么要提及降雨，而目标是淹没面积？此外，如果 Y 是目标变量且通过准则测量，分母应为 Ymean，而非未定义的 X。
回复：已在修订稿中对 NRMSE 公式进行了统一规范，明确以目标变量（淹没面

积 Y）的均值作为归一化基准（即 Y_mean），并删除原文中不一致或未定义的符号（如 X）。同时，对指标描述进行了修正，使其与研究目标变量（淹没面积）保持一致。

Response: The revised manuscript standardizes the NRMSE formulation by explicitly using the mean of the target variable (inundation area, Y) as the normalization basis (i.e., Y_{mean}), and removes previously inconsistent or undefined symbols (e.g., X). In addition, the description of the metric has been revised to ensure consistency with the study's target variable (inundation area).

2.5.2 未说明 x 代表什么。根据上下文，它可能是降雨，但降雨在 L137 中被记为 Q。 请使用与 NRMSE 相同的符号表示决定系数。这里为什么提到水深？目标不是淹没面积吗？ •

回复：已更改符号用法及添加说明，目标是淹没面积。

2.5.3 关注训练时间是个好主意，但既然如此，为什么不采用多层感知机？在如此简化的案例研究中，感知机远比 LSTM 更快。

回复：已在修订稿中补充说明选择 LSTM 模型的原因，即其在处理时间序列问题（降雨–积水演化过程）中的优势，能够捕捉时间依赖关系。训练时间作为评价指标之一，主要用于比较不同数据集配置对模型效率的影响，而非用于模型类型间的性能对比。尽管多层感知机（MLP）具有更高的训练效率，但其难以有效捕捉降雨–积水过程中的时间依赖特征，因此本研究优先采用能够处理时间序列依赖关系的 LSTM 模型。

Response: The revised manuscript further clarifies the rationale for selecting the LSTM model, emphasizing its advantage in handling time series problems (i.e., the rainfall–inundation evolution process) and its ability to capture temporal dependencies. Training time is included as one of the evaluation metrics primarily to compare the impact of different dataset configurations on model efficiency, rather than to assess performance differences between model types.

Although multilayer perceptrons (MLPs) may offer higher training efficiency, they are less effective in capturing the temporal dependency characteristics inherent in the rainfall–inundation process. Therefore, this study prioritizes the use of the LSTM model, which is better suited for modeling time-dependent relationships.

2.6.1 为什么要谈“效应”？这个概念应在全文中统一定义和使用。整个小节需要重写以使其可理解。

3.1：整个小节应提前移至讨论模型的章节。

回复：已前移并整合。

- 图 6：请提供比例尺。
  回复：好的。
- 请提供 InfoWorks ICM 的参考文献。
  回复：已增加 InfoWorks ICM 的相关参考文献
- 图 7 中需明确三个观测流量（而非淹没面积）分别对应什么。此外，我们注意到图 7 中模型完全错过了第一个峰值。观测降雨强度显著低于模拟人工降雨（0.6 mm vs 100 mm/h）。这是为什么？是否因为模型无法准确再现轻度降雨的响应？ 必须提供水文模型性能的全面、详细概述。
  回复：已在修订稿 2.1 部分详细描述。

- 方程 10 更像是循环 MLP 的方程……
  回复：已更改。
- 表 3：内容重复。
  回复：已删除。
- 图 8：计算时间的增加是否与信号长度增加成正比？
- 图 11 难以阅读。为什么归一化准则波动如此剧烈？前后不一致。
  回复：调整图片样式。
- 图 12：只有橙色可见。
  回复：已优化图片。
- 图 17：这里讨论的是哪个模型？物理模型还是 LSTM？请检查纵轴："value" 并未指明显示的变量。
  回复：已在图注及正文中明确图 17 所对应模型 LSTM 预测结果与机理模型预测值，并将纵轴"value"修改为具体物理量名称。
- 我们不知道泛化能力是如何衡量的。在哪个数据集上？
  回复：已在修订稿中补充泛化能力评估方法，明确通过不同降雨分级/数据集划分（训练集与测试集差异）进行评估，并结合 NRMSE 与 R^2进行量化分析。

L561：不能写"Controlled experiments with high-fidelity synthetic rainfall–inundation datasets"。物理基模型的评估不足；图 7 所示的例子完全错过了第一个事件，远非"高保真"。因此，结论的其余部分也未经证实。尽管我的经验与作者结论一致——需要足够多且不同类型的事件——但这些结论既不新颖，也未得到严格证明。

结论："but beyond approximately 14,400 sequences"——不对，这里指的是 14,400 个时间步长，而非序列。

综上所述，本文虽选题有趣，但方法论、呈现质量和结论的严谨性均存在严重问题，建议重大修改后重新投稿。