arXiv:2603.02418v1 [stat.AP] 2 Mar 2026

研究论文预印本

地理定位气象数据和建筑物相关数据对保险洪水风险评估的贡献

Mulah Moriah,*†‡ Franck Vermet,†‡ Pierre Ailliot,†‡ Philippe Naveau,¶ 和 Juliette Legrand†‡

†大西洋布列塔尼数学实验室，布雷斯特西布列塔尼大学，布雷斯特，法国
‡让·迪厄多内欧洲精算学院（EURIA），布雷斯特西布列塔尼大学，布雷斯特，法国
¶气候与环境科学实验室，巴黎文理研究大学和巴黎-萨克雷大学，吉夫叙尔伊韦特，法国 *通讯作者。电子邮箱：mulah.moriah@etudiant.univ-brest.fr

摘要
洪水是造成损失最惨重的自然灾害之一，2013年至2023年间导致超过1,000亿美元的保险损失。在法国，自然灾害保险制度持续出现赤字，凸显了开展精准建筑物尺度洪水风险评估的必要性。尽管保险公司通常依托灾害地图和区域气候指标，采用频率—严重度模型，但以往研究表明，此类大尺度变量只能解释个体洪水损失变异性的一小部分。本研究评估了多个地理参考数据层对大型法国住宅保险组合中洪水理赔发生和严重程度建模的边际贡献。研究从使用标准核保信息的基准模型出发，依次引入气候专家变量、极端降雨指标，以及精细尺度的地理定位建筑物和环境属性。本分析针对一种实际应用场景：由于预算、许可或运营方面的限制，保险公司无法部署完整的水文或水力巨灾模型。结果显示，基于降雨的指标，尤其是一项新构建的、用于捕捉局地强降水的指标，显著改善了理赔建模；建筑物和环境变量则进一步提升了发生概率预测的精度。总体而言，研究结果表明，高分辨率地理定位数据如何提升暴露和脆弱性评估、补充官方洪水地图，并为保险公司提供一套可操作的框架，以完善洪水风险评估和定价。

关键词： 洪水风险、降雨数据、地理定位的建筑物及周边环境数据、财产保险

1. 引言

1.1 洪水风险与保险的作用

保险是限制气候相关灾害造成的经济和社会成本的重要工具。通过汇集风险并在保单持有人之间分摊损失，保险公司帮助家庭和公共机构从破坏性事件中恢复。在法国，标准住宅保险由《自然灾害保险》（CatNat）计划补充，该计划于1982年创立（Grislain-Letrémy 和 Peinturier，2010），通过强制附加费和公私合作再保险安排，为全国范围内的重大自然灾害提供保障。近年来，这一框架面临的压力不断加大。气候变化正使强降雨的发生频率和强度双双上升，而城市扩张和土地利用变化则加剧了暴露和脆弱性。因此，CatNat计划已连续八年出现承保亏损，凸显出改进洪水风险量化、预测和定价方法的必要性（参见 France Assureurs，2023年）。洪水是法国造成损失最严重的自然灾害，自1982年以来，CatNat赔付累计超过€ 288亿欧元（France Assureurs，2021年）。洪水动力学差异很大，从快速径流和山洪到较缓慢的河流洪水，且深受局地地形、土壤条件和建成环境影响，因此建筑物层面的损失估算尤其困难。

1.2 现有方法及其局限性

洪水建模传统上依赖水文和水力学方法，用于模拟径流过程、河流流动动态、水深、流速和淹没范围。这些基于物理的模型构成洪水灾害制图的科学基础，并广泛应用于巨灾建模框架。它们通常需要高分辨率地形数据、水文网络和土地利用信息，以及专业的建模技能。随后，研究人员会将模型输出与损失函数相结合，以估算潜在经济损失；这通常需要使用来自保险组合或建筑专业知识的汇总数据。洪水建模和损失估算的大规模应用包括法国公共再保险机构 CCR（Caisse Centrale de Réassurance）以及法国保险联合会 France Assureurs 开展的研究。（例如，参见 France Assureurs，2021年）。

保险理赔数据，包括洪水发生记录和核保信息，日益被视为评估灾害影响的宝贵资源。此类数据已被用于验证灾害地图和降雨指标，尤其是在市镇或地区等聚合空间尺度上（例如 Spekkers等人，2014年；Zhou等人，2013年）。尽管这些方法对大规模风险评估有效，但它们不太适合解释在保险组合或单体建筑层面观察到的强烈损失异质性。

在实践中，大多数保险公司不具备为常规定价或保险组合风险管理部署详细水文或水力模型所需的财务、技术或运营能力。这一局限因保险组合在空间上呈集中而非均匀分布而进一步加剧，使得局部暴露效应尤为重要。因此，保险公司必须在个体保单尺度上评估洪水风险，同时考虑建筑物的特征以及高度局部化的损失驱动因素。因而，洪水风险定价通常依赖于基于核保数据构建的统计频率—严重程度模型，有时辅以灾害地图或宽泛的气候指标。然而，实证证据表明，此类大尺度变量仅能解释建筑物层面损失变异性的有限部分（Merz等人，2013年；Spekkers等人，2011年；Zhou等人，2013年）。地形坡度、土壤不透水性、与水道的邻近程度、周边土地利用以及强降雨的空间集中等局部因素在洪水损失中起主导作用，但市镇或地区指标难以充分刻画这些因素。与此一致，Merz等人，2004 表明，仅水文相关变量只能解释观测损害变异性的一小部分，凸显了对更丰富的精细尺度背景信息的需求。

在政府间气候变化专门委员会（IPCC）和联合国减少灾害风险办公室（UNDRR）所采用的洪水风险框架中，风险源于灾害、暴露和脆弱性之间的相互作用。在保险情境下，暴露通常可以通过地理编码保单位置得到较好的近似。脆弱性部分由核保变量体现，但这些变量往往缺乏捕捉建筑物层面特征和周边环境条件所需的空间与结构细节。在这三个组成部分中，降雨强度和地表径流的局部动态通常是标准保险模型中表现最不足的部分。

本研究着重于在传统保险定价框架下对洪水索赔发生和严重程度进行建模，旨在实现双重目标：估计用于费率厘定的洪水风险，并评估保险组合中已持有保险合同的风险。分析聚焦于一个实际应用场景：由于预算、许可或运营方面的限制，保险公司无法部署完整的水文或水力巨灾模型。我们力求通过系统纳入地理定位数据和气象情境数据，在保持经典频率—严重度统计模型可解释性及日常实用性的同时，改进传统的保险公司洪水风险评估。据我们所知，此前尚无研究在统计方法中同时整合保单级保险人记录、详细的建筑属性与环境特征，以及基于降雨的变量，用于预测洪水相关建筑损失。现有研究通常使用汇总索赔数据，或仅掌握有关单个建筑物的有限信息。例如，Torgersen等人，

2017年 纳入了有关建筑周边环境的详细数据，但未纳入气象变量和具体的建筑属性。通过填补这一空白，本研究为建筑级洪水风险提供了更细致的视角，并为希望加强洪水风险建模和减灾工作的保险公司及研究人员提供了可操作的见解。

1.3 议程与目标

本研究旨在评估，逐步纳入一致的地理定位背景信息，如何提升保险公司使用的传统建筑级洪水风险模型。我们从保险公司的视角出发：在无法常规应用全面的物理水文或水力模型的环境中，保险公司依赖频率—严重度统计模型进行定价、监控保险组合和划分风险。为此，我们构建了一系列嵌套模型。第一个模型作为基准模型，仅使用保险组合中的传统核保变量和保单级数据。

第二步，我们通过添加保险实践中广泛使用的水文与气候指标来增强基准模型，例如灾害地图、分区信息和汇总地理变量。这一建模层反映了业界将洪水灾害数据纳入统计定价模型的主流做法。第三个建模层加入了源自网格化气象数据集的基于降雨的变量。这些变量旨在更准确地捕捉局部灾害动态，并反映静态灾害地图无法呈现的降水事件在空间和时间上的变异性。最后，我们纳入建筑特征及其周边区域的详细地理定位信息。通过纳入局部结构和环境属性，这一层旨在解释预测洪水发生及相关成本中更大比例的变异。

在每一步中，我们都会评估新增数据层在多大程度上改善了洪水理赔发生及严重程度的建模，并使用适用于保险情境的性能指标。分层流程使我们能够衡量各类信息在实际操作中的贡献。本文其余部分结构如下。第2 节介绍保险组合，并详细说明气象、建筑物层面及环境变量的构建。第3节概述建模框架，考察各阶段取得的增量性能改进及其影响。第4 节讨论研究结果对保险实践的更广泛意义，以及本研究的局限性和未来研究方向。

2. 数据

本研究结合多个数据来源，在建筑尺度上对洪水理赔的发生及其严重程度进行建模。数据包括：（i）保险公司的保险组合和理赔记录；（ii）公开可获取的水文和气候数据集；以及（iii）描述承保建筑物及其周边环境的地理参考信息。图1 总结了这些数据层，并详细说明了其来源和空间分辨率。本研究使用的大多数外部数据来源均可公开获取。我们以这些原始开放数据集为基础，构建并设计了一系列气象、环境和建筑层面变量，使其适用于标准的保险公司建模框架。从数据角度来看，我们的主要贡献在于确定哪些变量可以从这些开放数据源中推导出来，说明如何构建这些变量，并评估它们对于建筑物层面洪水风险评估各自具有的增值作用。尽管 GASPAR 等国家灾害数据库以及法国生态转型部的2023a 数据对洪水事件具有广泛的保险保障范围，但它们仅记录法国自然灾害制度下由政府认定的自然灾害。由于认定取决于市镇提出正式申请，规模较小、高度局部化或未被报告的洪水事件往往会被遗漏。此外，GASPAR 的空间分辨率仅达到市镇尺度，对于建筑物层面的建模而言过于粗略。因此，我们依靠保险公司业务组合数据来评估完整的建模框架，并

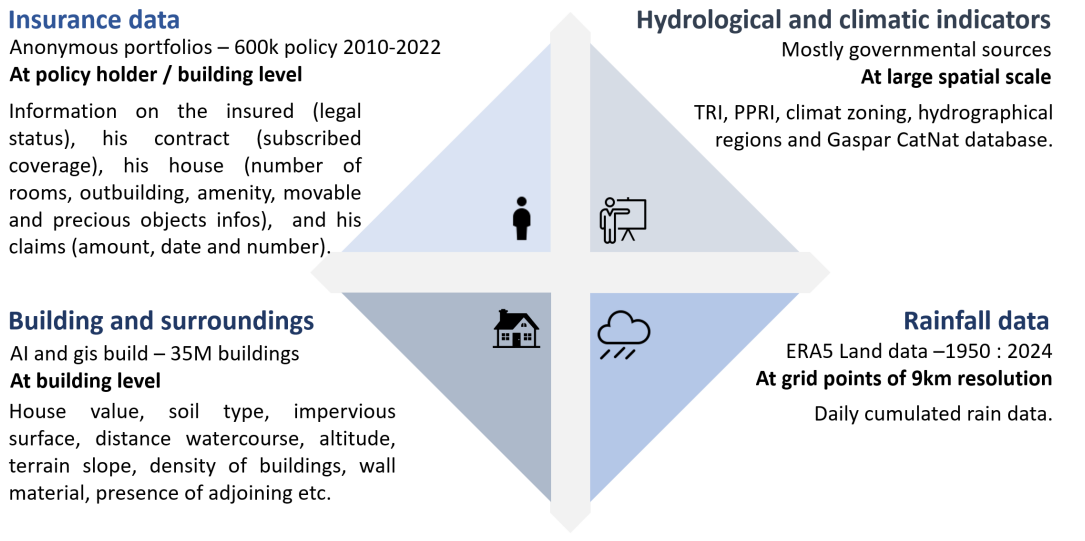


图 1. 数据来源概 述。本研究使用的数据来源。建筑物、水文和气候数据均采用各数据源最新可用的 e
版本。 各变量的数据生产年份不尽相同。

仅在考察降雨相关变量的大尺度行为和内部一致性时，才采用汇总后的全国数据集。

如图 2 所示，大多数结果地图集中展示法国的三个地区：奥克西塔尼、上法兰西以及大东部的一部分。这些地区呈现出各异的洪水风险特征。奥克西塔尼主要受地中海气候影响，降雨历时短但强度高，地形复杂，这些因素共同促使快速径流和山洪的发生，塞文诺尔型事件期间尤为如此。相比之下，上法兰西的气候属于温带海洋性气候，地势相对平坦，其洪水机制更多由缓慢的河流漫溢以及低洼地区的排水系统饱和所主导。大东部所选省份的气候、水文地理和地形条件则处于中间状态。除可视化用途外，本文报告的所有分析和性能指标均基于完整的全国数据集计算得出。

本节其余部分将介绍各数据来源以及解释变量的推导过程。第2.1 节呈现保险组合和理赔数据集。第2.2 节讨论保险应用中通常使用的水文与气候指标。第2.3 节介绍由降雨数据推导出的变量，包括MILRE指标的构建。第2.4节详细说明基于地理位置的建筑物相关变量和周边变量。有关数据来源和变量构建的更多细节见附录1。

2.1 保险数据

本研究使用了从法国大型家庭保险组合中提取的数据，涵盖2014-2022年。研究期间，该保险组合约包含968,000名不同的保单持有人，其中许多保险单跨越多年，并包括10,800条洪水相关理赔。该数据集包括面临洪水风险的住宅保险保单及其相关理赔记录。每份保险单均精确定位至建筑物地址，从而能够整合外部地理空间信息。

对于每栋承保建筑物，可用的核保信息包括保险定价中常用的标准保险单特征，例如占用类型、被保险价值、保险保障

选项和合同期限。这些变量构成了保险公司使用统计频率—严重度方法对洪水风险进行建模时传统上可用的核心信息集。理赔数据包括洪水相关理赔的发生情况及相应的赔偿金额。理赔成本以欧元计，并调整为不变价格，以反映研究期间的通货膨胀。

该建模框架区分两个互补的目标：洪水理赔发生，定义为保险单在给定暴露期内至少发生一次洪水相关理赔的概率；洪水理赔严重程度，定义为在发生理赔的条件下，洪水理赔的总成本。这种区分符合标准精算实践，并允许不同因素分别影响损失发生的概率和损失的规模。

保险合同异质性是利用理赔数据时众所周知的问题（参见 André 等人，2013年）。为确保数据集同质性，我们将分析范围限制在保险保障和保险责任相近的保单子群体。此筛选依据合同相关信息进行，并与保险公司数据专家合作完成。此筛选步骤所用的合同变量不会纳入建模阶段，从而使分析聚焦于洪水风险的驱动因素，而非合同效应。

2.2 水文与气候指标

为利用保险公司通常可获取的信息刻画洪水灾害，我们采用一组基于公开数据集构建的水文与气候指标。这些指标与法国公共机构编制的参考洪水风险评估保持一致。

洪水灾害分区主要通过重大洪水风险区域（TRI）和洪水风险预防计划（PPRI）来体现。TRI依据欧盟指令 2007/60/EC（欧洲洪水风险中心（CEPRI），2022年）设立，划定面临重大洪水风险的区域，其依据是针对选定的主要河流和海岸地带开展的详细水力学和地形研究。这些区域以高空间分辨率绘制，旨在识别洪水可能对人口和经济活动造成重大影响的位置。然而，由于其针对性范围有限且制定成本高昂，TRI分析仅覆盖我们保险组合中5%的建筑物。相比之下，PPRI（环境、规划与交通区域及跨省部门局，2023年）是以市镇尺度制定的监管规划工具。它们利用历史洪水事件、简化的水力模拟以及专家评估所得信息，将市镇区域划分为较大范围的灾害分区。尽管PPRI覆盖了领土的很大一部分（约占被保险建筑物的30%），但其空间粒度相对较粗，区域划定往往受行政和土地利用标准驱动，而非基于对洪水过程的详细呈现。

除 TRI 和 PPRI 外，我们还纳入了与水文和气候相关的其他地理描述指标，例如流域边界和气候区。这些特征反映了长期环境背景，并捕捉了法国各地降水型的整体差异，但它们要么是静态的，要么随时间缓慢变化。

这些水文与气候变量共同构成了保险公司通常采用的标准开源基准灾害信息。它们为洪水风险评估提供了机构参考，但在表征单体建筑尺度上降雨诱发洪水的精细空间和时间变异性方面，能力存在根本限制。在本研究中，这些变量既用作解释变量，也用作参考，以评估基于降雨的指标和建筑层级属性所带来的增量价值。

2.3 降雨相关变量

洪水损失通常由强烈或持续的降雨事件引发，但降水与由此造成的损害之间存在间接且高度非线性的关系。洪水可能通过

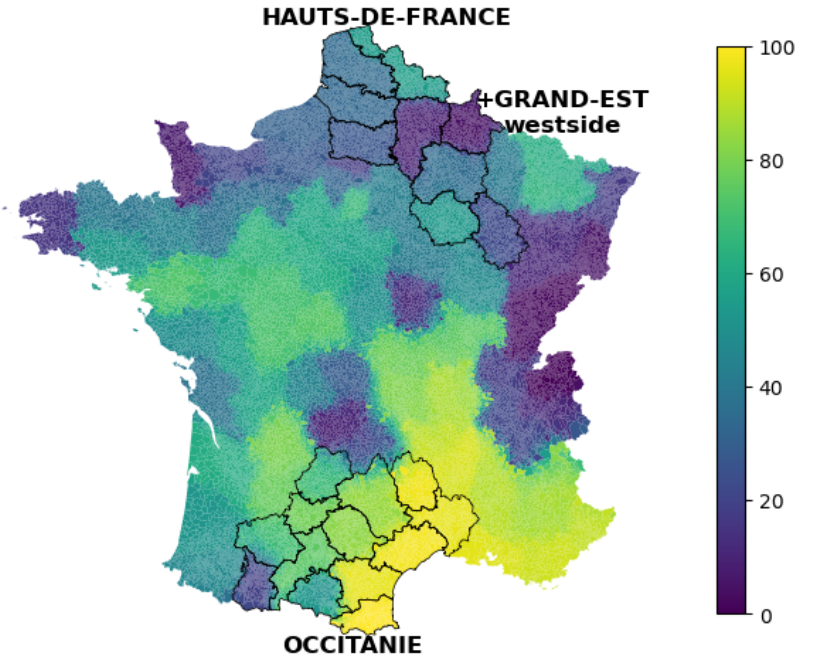


图2. 基于广义帕累托分布形状参数估计的尾部权重，已重新缩放至0 与100之间，其中100 对应最厚重的尾部。用于建模结果可视化的选定关注区域也以黑色标出。

多种机制发生，包括地表径流、河流溢流、土壤饱和或排水系统故障；损害往往是由降雨的水平再分布造成的，而非由某一时刻单点的降水强度造成。因此，原始降雨测量数据很少适合作为预测建筑物层面洪水发生或损失成本的直接预测变量。

为了在保险建模框架内更好地刻画这些动态，我们构建了降雨相关变量，以同时考虑极端降雨特征的空间异质性以及降雨事件的时间结构。所有降雨变量均源自ERA5-Land再分析数据集，该数据集在覆盖法国本土的规则网格上提供每日降水估算值，约包含6200个网格点，观测数据时间跨度为1950年至2024年（Muñoz Sabater，2019）。

降雨极端事件的空间特征：极端降雨分布在法国全境表现出强烈的空间变异性，反映出对比鲜明的气候区制。为考虑这种异质性，我们运用极值理论估计日降雨量分布的尾部特征。具体而言，我们对高阈值以上的超阈值量拟合广义帕累托分布，并将估计得到的形状参数作为尾部厚重程度的代理指标（Coles 等人，2001）。随后，将尾部特征相似的网格点用聚类方法归为一组，从而构建一个记为尾部_权重_聚类的分类变量。图2 展示了聚类之前的原始尾部权重，并重新缩放至0至100之间，其中100对应最厚重的尾部。所得格局与已知气候差异一致：南部地区的降雨分布尾部更厚，北部地区的尾部则较轻。

基于事件 d 降雨聚合：除了空间异质性之外，洪水发生

还强烈依赖于降雨的时间结构。有些事件由短时强降水引发，而另一些则是数日累积降雨的结果。为捕捉这些机制，我们计算多个时间窗口内的降雨聚合量，并针对每个暴露期，提取各窗口中的最大累积量。这些构造基于水文考量，旨在近似洪水形成的关键驱动因素，例如土壤饱和度和径流潜势，同时不依赖显式的水力建模。

我们引入一种合成降雨指标，称为最强局地相对事件（MILRE），旨在同时处理空间异质性和多尺度时间累积问题。在洪水严重程度建模中，每笔理赔都对应一个已知的洪水日期。从该日期向前回溯，计算事件发生前多个时间窗内的累积降雨深度，即 $nd \in \{1, 3, 5, 7, 10, 30\}$ 天。对于每栋承保建筑物，我们从 ERA5-Land 网格中选取距离最近的四个点获取降雨值，并保留这些点中的最大值，以代表局地降水状况。该值记为i_{nd}。

对于每个累积时间窗 nd，将洪水日期当天观测到的降雨值转换为经验累积概率，该概率根据该位置相应累积降雨序列的历史分布估算：

$$\widehat{F}_{nd}(i^*_{nd}) = \frac{1}{M_{nd}} \sum_{k=1}^{M_{nd}} \mathbb{1}_{i^k_{nd} \leqslant i^*_{nd}},$$

其中 M_{nd} 是向量中的天数，i^*_{nd} 是与洪水事件相关的观测值。我们将以毫米为单位的绝对降雨深度值转换为经验概率，以便用相对值表示，从而便于比较不同时间窗以及降雨特征各异的区域。图 3 示意了针对给定时间窗和洪水日期，计算与建筑物相关的 $\widehat{F}_{nd}(i^*_{nd})$ 所需的过程。

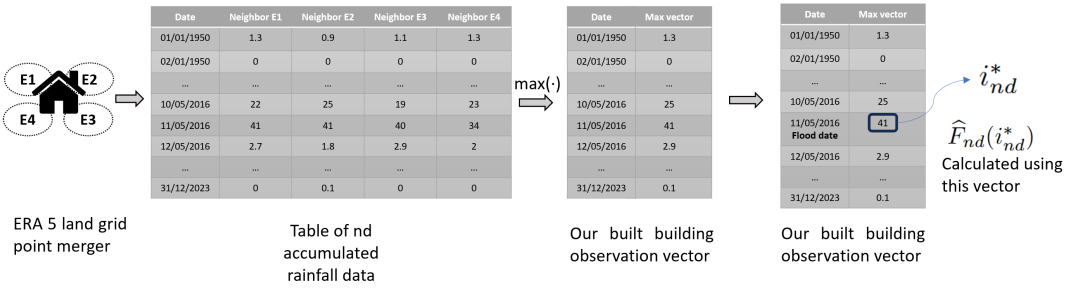
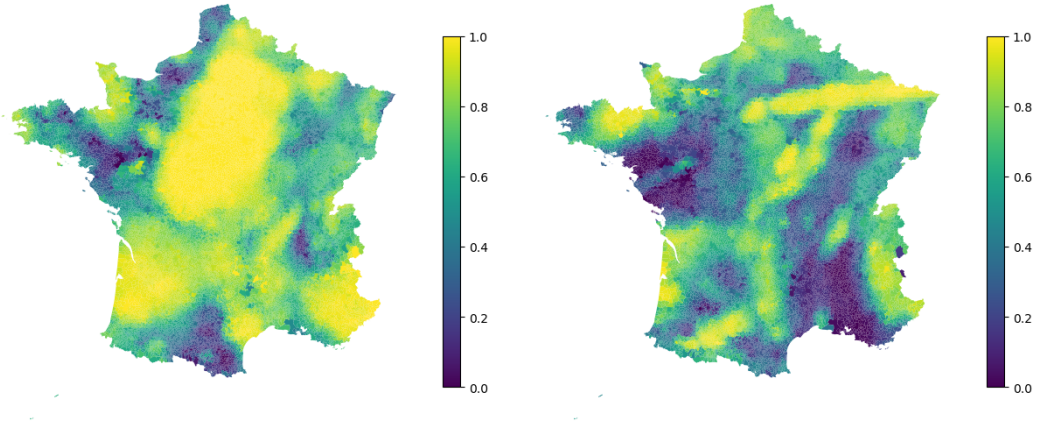


图3. 示意图说明了如何获取某一建筑物在给定时间窗和洪水日期下对应的经验概率。

随后，将所有累积时间窗中观测到的最大经验概率定义为 MILRE,

$$MILRE = \widehat{F}_{max} = \max_{nd \in \{1,3,5,7,10,30\}} \widehat{F}_{nd}(i^*_{nd}). \tag{2.1}$$

根据其构造，MILRE（最强局地相对事件）的取值介于0和1之间，数值越高，表示降雨事件**相对**于当地历史条件越极端。使用多个累积时间窗可确保我们同时从短历时降雨和长期累积效应中获取**最强事件**。



(a) 2016年的 ann_MILRE 值。

(b) 2017年的 ann_MILRE 值。

图4。2016年和2017年的 ann_MILRE（年度最强局地相对事件）空间分布。数值越高，表示相应地区经历的极端事件相对于其历史降雨水平越突出。

用于发生建模的年度MILRE：对于频率数据集而言，使用洪水日期并无意义。房产在一年中的部分时段获得承保，并且我们知道提交了多少洪水相关理赔。

因此，我们定义了一个类似的年度指标，记为 `ann_MILRE`：针对每一年和每个累积时间窗，选取该年记录到的最大降雨量，将此数值转换为经验概率，然后在所有时间窗中取最大值。更具体地说，对于给定的累积时间窗nd （回顾：1天、3天、5天等），我们首先确定1950年到2024年间相应累积降雨量的年度最大值。由此得到nd 个向量，每个向量包含年度最大值。随后，我们按照计算i^*_{nd} 和$F_{nd}(i^*_{nd})$时采用的相同方法，计算每个向量对应的经验累积分布。从概念上说，这为每一年提供nd 个值，用以表示年度最大值的经验累积概率。因此，我们得到：

$$\widehat{F}_{nd}(y^*_{nd}) = \frac{1}{75}\sum_{k=1950}^{2024} \mathbb{1}_{y^k_{nd}\leqslant y^*_{nd}},$$

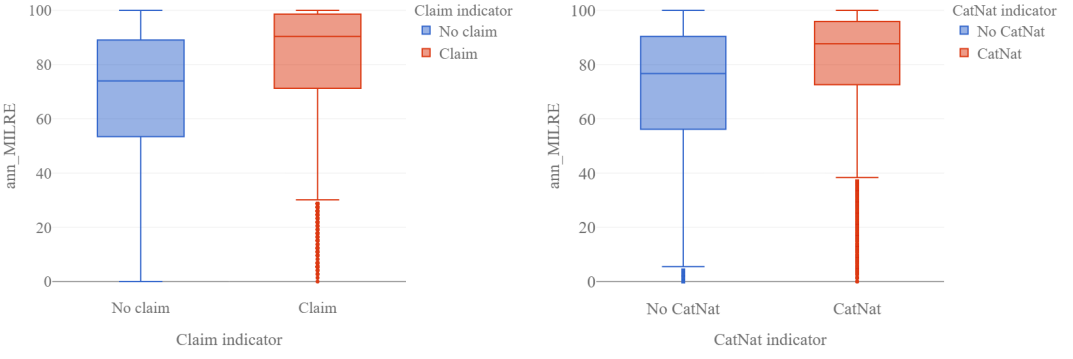
其中，y^*_{nd} 表示特定考察年份中，在该累积时间窗nd 内观测到的年度最大值。

最后，对于每一年，我们取nd 概率值中的最大值，将这些向量合并为一个向量，如2.1所示：

$$ann_MILRE = \max_{nd\in\{1,3,5,7,10,30\}} \widehat{F}_{nd}(y^*_{nd}). \tag{2.2}$$

该变量表示某一年是否发生了极端降雨事件；如果发生，则标示其位置。其基本作用是帮助排除未经历显著降雨相关极端事件的地区和年份。在图 4中，该图展示了_2016年和2017年的 MILRE 地图，高值分布的对比尤为明显。2017年的地图显示，高强度事件远少于2016年。2016年因法国发生的严重洪水事件而广为人知（法国生态转型部，2023b），而2017年则以

存在更多具有极低值的区域为特征，这表明缺乏极端降雨事件。尽管2016年也有部分地区未经历显著降雨，但这些区域的空间分布范围小于2017年。为量化2016年和2017年洪水发生次数之间的差异，我们参考了 Gaspar CatNat 事件计数。根据本研究所使用的 GASPAR 数据提取结果（于2022年下载），按正式宣布受洪水影响的市镇数量计算，2016年排名第8位，而2017年仅排名第41位。



(a) 按理赔发生筛选的保险组合数据集中的 ann_MILRE 箱线图。 (b) 按至少颁布一项获批 CatNat法令筛选的 GASPAR数据集中的 ann_MILRE 箱线图。

图 5. GASPAR 和理赔数据集按洪水发生情况划分的 ann_MILRE 箱线图，展示了有事件期间与无事件期间的分布差异。

在图 5中，我们分析了ann_MILRE 值的分布如何随是否提交理赔以及是否颁布 CatNat法令而变化。子图5a 比较了某一年报告理赔的保险合同（建筑物）与未报告理赔的保险合同（建筑物）的ann_MILRE值。子图5b 展示了按是否颁布 CatNat法令分组的箱线图，使用的是GASPAR数据集中的开放数据。该数据集按年份提供颁布 CatNat法令的市镇名单及相关事件日期。对于每一年，我们都会将未出现在数据集中的市镇补充到名单中，以确保每年都涵盖所有法国市镇。随后，我们绘制各类别对应ann_MILRE值的箱线图。

在两个子图中，橙色箱线图的数值都集中在100%左右，表明发生理赔或颁布法令时，我们的变量通常高于未发生洪水事件时的值。大多数数值超过35%，而在没有发生事件时则并非如此。尽管两者的分布部分重叠，但仍明显不同。在GASPAR数据集上得出类似结论，是一项值得关注的成果。该数据集的时间和空间覆盖范围比我们的保险组合中的理赔记录更广（涵盖1983年至2023年的所有法国市镇）。

2.4 建筑物及其周边环境

洪水损失在很大程度上受局部脆弱性条件影响，而这些条件无法仅通过灾害地图或降雨指标反映。这些条件既涉及建筑物本身的固有属性，也涉及其周边环境的特征；两者共同决定洪水如何到达承保建筑物，以及损害如何发展。与冰雹等灾害不同，冰雹造成损害的因素会直接撞击建筑物，而洪水属于间接灾害：降雨会引发一系列过程，涉及地形、土地利用和建成环境，最终导致损害。

为表征这些影响，我们汇集了一组描述承保建筑及其局部环境的地理定位变量。这些变量源自开源地理数据

并在单体建筑尺度上与承保位置进行空间关联。它们旨在补充保险公司的核保数据，同时与典型定价做法及既有洪水风险评估程序保持一致。

建筑特征：建筑物层面变量描述了会影响洪水易损性和潜在损失规模的结构属性。这些变量包括建筑占地面积、建筑类型、楼层数和主要墙体材料。其中一些特征已用于保险定价和核保。在保险公司系统中可获取这些信息时，它们还可通过支持地理编码验证、检测报告错误或遗漏，以及提高风险暴露信息的整体可靠性，促进数据一致性的改善。在本研究中，建筑特征用于捕捉被保险资产之间洪水易损性的相对差异，而不是用于表示详细的结构或工程属性。其作用是提供建筑尺度上的额外背景信息，以补充保险公司的核保数据和灾害指标。

周边环境（建成环境和自然环境）：局部环境在洪水形成、传播以及局部尺度上的汇聚过程中发挥着关键作用。因此，我们纳入了描述承保建筑物周边土地利用、土壤封闭、地形和水文环境的变量。这些指标包括地表不透水性、植被覆盖、局部坡度、相对高程以及与水道的距离等度量。总体而言，这些变量决定了地表径流的强度、汇流范围，以及降雨能否下渗，还是会被迅速排走。我们在每栋建筑物周围预先设定的空间邻域内计算这些变量，从而简洁且实用地概括局部条件，而无需明确模拟水流路径。尽管蒸散和防洪基础设施等因素也会影响洪水动力学，但目前在国家层面无法获得关于这些方面的一致且空间覆盖全面的数据。因此，本研究所使用的环境变量集合未明确纳入这些因素。

在建模框架中，建筑物层面和环境变量的贡献将通过递进方式进行评估，既单独评估，也评估其与降雨相关变量之间的交互作用，以判断局地背景如何调节极端降水对洪水发生及其严重程度的影响。

我们希望通过指出以下事实来结束本节：France Assureurs 预计，到2021年2050年，洪水相关损失将增加81%。其中，34%的增幅与气候影响有关，包括13.77%可直接归因于气候变化。例如，Luu等人2018年的研究报告称，在法国南部，超过百年一遇洪水幅度的概率增加了2.5(±0.8)个百分点。Blöschl等人2019以及Philip等人2018的归因研究提供了进一步证据。总之，这些结果凸显了将气象和水文指标纳入洪水风险模型的必要性，既要体现风险的主要驱动因素，也要跟踪变暖气候下洪水灾害的变化。尤其是，France Assureurs 使用日降雨量的90%和99%分位数来解释和预测洪水损失，这表明，在保险分析中更充分地利用此类变量，将有助于提升我们对洪水损失的理解和预测能力。

France Assureurs 还将预期成本增幅的65%归因于“财富效应”，这反映了城市化程度的提高以及资产价值的上升。这凸显了在气候风险评估中考察家庭行为与聚居模式的重要性（Bézy，2023年）。尽管这一维度超出了本研究的范围，但它进一步促使我们将理赔记录与详细的建筑属性及当地环境条件联系起来（Torgersen 等人，2017年）。表1 概述了本研究建模所考虑的变量。

表1. 变量组、空间分辨率和建模作用概览。详细的变量定义和描述性统计见附录A。

	变量	空间分辨率	备注
保险数据			
	理赔发生、索赔金额	建筑物	建模目标
	房间数量、动产、贵重物品，外-建筑规模	建筑物	核保信息
水文及气候数据			
	PPRI、气候区，CATNAT法令数量	市镇	法规与历史专家分区
	TRI、集水区	精细化多边形	法规与历史专家分区
降雨数据			
	极端尾部权重聚类	市镇	降雨极端事件的空间特征极端降雨
	MILRE , ann_MILRE	重新计算的 ERA5 网格在建筑物层面	基于事件的降雨强度指标
建筑及周边数据			
	居住面积、房屋价值、建造时期，楼层数、墙体材料、建筑外表面，墙体总长度、相邻建筑的存在	建筑物	结构与脆弱性描述指标
	地形最大坡度、建筑密度、主要土壤类型、不透水表面比例	建筑周边建筑物	局部环境背景
	至最近水体的距离与海拔差	建筑物	与水文网络的距离
	河道		work

3. 洪水索赔成本与概率建模框架

3.1 建模方法

对于每位被保险人 $i \in 1, \ldots, n$，令 $\mathbf{Y_i}$ 表示目标变量，在意外保险中，该变量通常代表索赔频率、发生概率、严重程度或纯保费。令 N_i 为索赔数量，$C_{i,j}$ 为第 j 笔索赔的成本。在实践中，保险公司使用不同的数据集分别对发生情况与严重程度进行建模（参见 Goldburd 等人，2016年）。在每位被保险人每年至多观察到一次索赔的情况下（如我们的数据集），预期总成本（或纯保费）可建模为：

$$\mathbb{E}(Y_i) = \underbrace{\mathbb{E}(\mathbb{1}_{N_i>0})}_{\text{probability module}} \times \underbrace{\mathbb{E}(C_{i,1})}_{\text{severity module}} .$$

自 Brockman 和 Wright 于 1972年提出后，传统定价方法便一直基于广义线性模型（GLM）；Nelder 和 Wedderburn 于 1972年也对此进行了阐述。我们有意作出这一选择：我们的目标不是提出新的建模技术，而是评估额外的背景变量和地理定位变量如何改善保险公司运营环境中的洪水风险建模。在我们的设置中，我们使用逻辑回归（采用伯努利分布和 Logit 链接函数的广义线性模型）来建模索赔概率，并使用对数链接的伽马广义线性模型来建模索赔成本。在这两种情况下，线性预测项的形式为：

$$g(\mathbb{E}(Y_i)) = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_p x_{i,p}, \tag{3.1}$$

其中，$g(\cdot)$ 表示链接函数，$(x_{i,1}, \ldots, x_{i,p})$ 为解释变量，$(\beta_0, \ldots, \beta_p)$ 为对应的系数。

第一个建模层仅采用标准核保信息，作为基线。后续层逐步纳入其他数据来源，包括专家提供的气候与水文洪水分区变量、降雨指标以及建筑物及周边特征。正文中展示的所有模型均采用相同的广义线性模型（GLM）结构，并使用相同的训练与评估方案进行估计。因此，预测性能的提升可直接归因于新增数据层，而非建模技术的变化。

空间异质性在洪水风险中发挥着核心作用，保险公司通常依靠基于行政区划或专家来源的分区变量来考虑暴露的地域差异。多项研究提出了两步法：首先估计非空间模型，然后通过潜在分区变量或空间平滑技术捕捉残差空间结构，再将其重新纳入定价模型（参见 Assunção 等，2014年；Boskov 和 Verrall，1994；Scheel 等，2013年；Taylor，1989）。本研究主要通过显式协变量引入空间信息。此选择旨在直接将洪水的发生及严重程度与建筑物层面可解释的物理与情境驱动因素联系起来。

3.2 拟合细节与模型评估

为评估模型性能与有效性，我们采用了多种评估指标，并针对具体的建模挑战分别进行了调整。成本模型采用以下指标进行评估：

- 均方根误差 (RMSE) $\left[{}^{1}{}_{n}\sum_{mi=1}(\hat{}\gamma_i - \gamma_i)^2\right]^{1}/2$ 会对较大的预测误差施加更大惩罚，并提供预测精度的绝对衡量。数值越低，表明模型性能越好。
- 基尼指数用于评估模型按风险重要性从低到高排序的能力（参见 Denuit et al., 2019；Frees et al., 2014）。预测结果按 $\hat{}y(n) \geq \cdots \geq \hat{}y(1)$ 的顺序排列，并根据这一排序计算累计暴露比例 $P(i) = i/n$ 和累计观测总成本 $R(i) = \sum_{ij=1} yj/\sum_{nj=1} yj$。该指数根据相应的洛伦兹曲线，采用标准公式计算（Dorfman, 1979）。指数值越高，排序能力越强。
- 偏差 $2\sum_{mi=1}\left[(\gamma_i - \hat{}\gamma_i)/\hat{}\gamma_i - \log(\gamma_i/\hat{}\gamma_i)\right]$ 用于比较拟合模型的对数似然与饱和模型的对数似然；饱和模型能够完美预测观测数据。值越低，模型越好。

索赔发生建模的特点是极端的类别不平衡，正结果约占0.22%。在此类情形下，标准的训练与评估程序可能导致有偏模型，从而低估稀有事件He和Garcia，2009。为解决这一问题，我们在广义线性模型估计过程中采用类别重加权策略Wang等，2017年；Zhu等，2018。观测权重定义如下：

$$w_0 = \frac{n}{2 * \sum_{i=1}^{n} \mathbb{1}_{\gamma_i=0}} \text{ and } w_1 = \frac{n}{2 * \sum_{i=1}^{n} \mathbb{1}_{\gamma_i=1}},$$

$w0$与$w1$分别是多数类与少数类的权重。该选择在保持总有效样本量的同时，对稀有事件的误分类施加更重的惩罚。它可视为代价敏感学习的一种简化变体，因为仅改变了一种错误惩罚。

在不平衡分类问题中，传统指标（例如准确率）往往会产生误导（参见 Guo 和 Viktor，2004；Japkowicz，2013年）。因此，我们采用：

- 对数损失（-1/2n 偏差）：$-\frac{1}{n}\sum_{Ni=1}\left[(1-\gamma i)\log(1-\hat{}\gamma i)+\gamma i\log(\hat{}\gamma i)\right]$，用于评估概率校准。在逻辑回归情况下，对数损失与偏差成正比。结果表中报告的是 -对数损失；数值越小，表现越好。
- 基尼指数用于评估排序能力；在类别严重不平衡的情况下，它比基于 ROC 的指标更具信息量（Jeni 等，2013年）。
- 临界成功指数：真正例/(真正例+假正例+假负例)，由 Schaefer 于1990 年提出，用于衡量分类器在极端类别不平衡情况下的技能。有关该指标的最新应用和讨论，参见 Legrand 等人，2021年。与概率指标或排序指标不同，该指标基于预测类别标签而非概率定义。为确保模型间公平比较，我们确定每个模型产生最大 CSI 值的阈值，并将该最大值报告为指标。指标值越高，表示性能越好。利用该阈值和相应的预测类别，我们还计算了受洪水影响个体的分类精确率和召回率。

所有模型均采用五折交叉验证进行评估。除绝对性能指标外，结果还以仅使用核保数据的基准保险公司模型为参照进行报告。此外，还纳入一个不含协变量的哑模型作为最简基准模型。通过报告估计参数的数量来评估模型的简约性。

为评估各个协变量的贡献，我们通过将完整模型与每次排除一个变量的受限模型进行比较，执行似然比检验。在被排除变量不会改善模型这一零假设下，检验统计量服从卡方分布。结果以对数刻度报告，显著性阈值设为 5%。对于选定的重点变量，我们提供观测值与预测值图，以说明其对风险估计的边际影响。我们还报告模型参数的数量，以评估模型的简约性。

以下建模结果部分按层次展开：

1. GLM ins：仅使用核保变量进行基线建模。2. GLM ins+c：在基准模型基础上，加入基于专家知识的水文与气候变量。3. GLM ins+r：进一步扩展模型，在水文与气候变量之外加入降雨衍生指标。4. GLM all：完整模型，纳入降雨、气候专家知识、建筑物及周边变量。

3.3 成本建模结果与讨论

表2 汇报了各模型的评估指标。表中显示的模型名称在地图和图表中保持一致，以便于解读。图6展示了所有模型的皮尔逊残差。残差图通过在市镇尺度上按组汇总数值并取中位数生成，而性能指标则在单体建筑层级上计算。尽管这种汇总不可避免地会减少部分细节，但它是在清晰呈现结果与保留足够细致程度以考察模型空间模式之间的最佳折衷。

表**1** 中仅使用保险数据的保险公司模型“GLM ins”

首先值得注意的是，我们的保险公司模型虽然略优于虚拟基准，但整体表现仍然较弱。该模型在误差度量方面仅有小幅改善，风险区分能力也十分有限，表2中的基尼指数为 0.0562，印证了这一点。这一结果并不意外，因为保险数据主要反映被保险资产价值，这与传统保险产品结构相对应。

此外，由于洪水相关损失主要由水量、流速和水深等水文驱动因素决定，成本与房屋价值之间的关联较弱。保险公司模型

表2. 使用五折交叉验证计算的成本模型指标汇总，以及相对于仅使用保险公司数据的模型的变化情况。指标改善时，变化以蓝色显示；否则以红色显示。

模型	RMSE	Gini	偏差	参数数量
哑模型 无变量	9165.4983 ↑0.21%	0 ↓100%	13340.42 ↑1, 21%	**0** ↓100%
GLM ins 仅保险公司数据	9146.6254	0.0562	13180.49	15
GLM ins+c 保险公司数据 + 气候	9054.3267 ↓1.01%	0.1474 ↑162.28%	12586.02 ↓4.51%	19 ↑26.67%
GLM ins+r 保险公司数据 + 降雨	9000.7446 ↓1.59%	0.1661 ↑195.55%	12424.03 ↓5.74%	19 ↑26.67%
GLM all 所有数据	**8781.6942** ↓3.99%	**0.2439** ↑333.99%	**11619.99** ↓11.84%	33 ↑120%

未纳入强度相关预测变量，也缺乏用于判断房产是否位于高风险洪水区域的指标。因此，该模型与理赔成本之间并未建立起强关联。对模型贡献最大的变量是可移动财产的价值，以及与建筑物的规模和价值相关的变量，例如房间数量。这些变量通过调整估算的损失金额，有助于解释保险损失的部分差异。然而，该模型并未形成有意义的空间分层，尽管这对于气候敏感型风险至关重要。

先前关于保险定价的研究表明，仅凭核保数据无法实现足够精确的风险分层。因此，保险公司经常利用外部来源增强其内部数据，包括地理与环境信息（参见 Denuit 和 Lang，2004）。与此一致，我们通过添加一个整合气候专业变量的额外层来扩展模型。

保险公司 + 气候模型，“GLM ins+c”，见表1 保险 + 水文与气候变量这一建模层在基准设定的基础上，加入了由专家定义的气候与洪水分区变量，例如 TRI、PPRI、气候区、水文结构和历史自然灾害记录。这些变量直接用作协变量，代表了当无法采用更复杂的建模方法时，保险公司通常用于考虑暴露空间差异的信息。在这些基于专家判断的风险变量中，TRI类别和气候区类别对保险 + 气候成本模型的影响最大。TRI划定于这样一些区域：这些区域兼具过往暴露、特定水文条件以及较高的潜在经济损失，因此它们与观测到的损失密切相关。气候区则通过根据降雨模式和主要洪水形成机制区分区域，提供互补的大尺度信息，从而捕捉洪水损失模式中持续存在的空间差异，并凸显出本身更易发生洪水的气候类型。

加入气候专家变量后，均方根误差降低了 1%，基尼指数提高了三倍，表明风险分层效果有所改善。如图6所示，空间模式开始显现，尤其是在地中海沿岸以及奥克西塔尼东部的加龙河沿岸。在上法兰西和大东部，气候模型提高了对这些地区南部风险的识别能力，而北部地区的分层效果仍然有限。地图上残差的减少体现了这一改进。

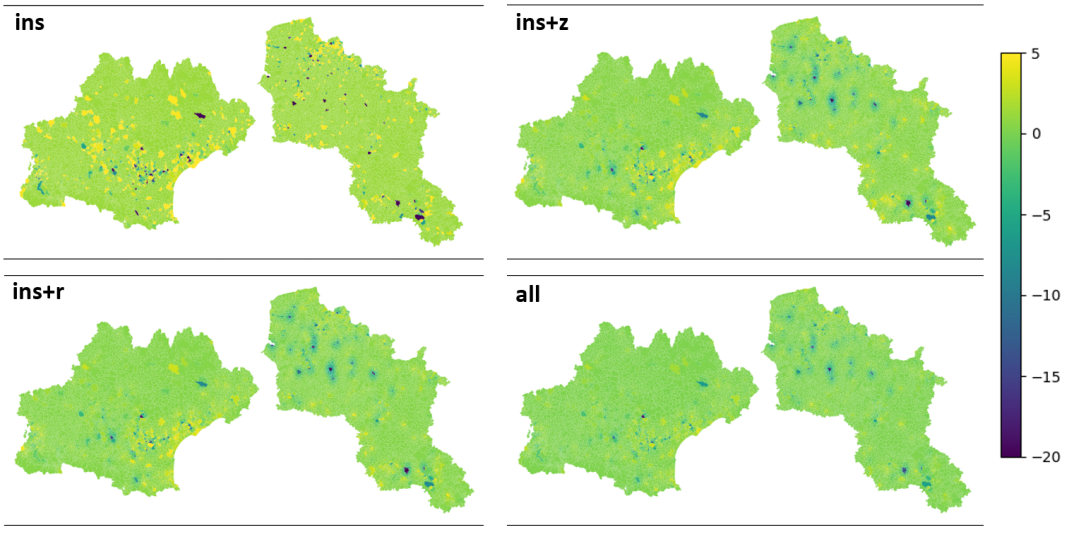


图6. 对于表 2 和图 2中定义的两个区域，展示了市镇层面上表中定义的四种广义线性模型成本模型的皮尔逊残差。残差明显偏离零，表明观测值与拟合值之间存在拟合偏差。

保险公司 +降雨模型，“GLM ins+r”，见表 1 保险 + 气候 + 降雨变量
现在，我们引入基于降雨的指标，并通过 MILRE 变量聚焦极端事件。该变量记录了与每栋受洪水影响的建筑物相关的最强局地降雨事件。MILRE 被划分为基于分位数的类别，以兼顾灵活性和稳健性，但在模型中作为连续预测变量使用。该变量提升了模型解释理赔严重程度的能力。其逻辑很简单：较弱的降雨事件往往造成较少的保险赔付款，而极端事件则更可能导致重大损失。表 2 中的实证结果支持这一观点：纳入极端降雨信息后，均方根误差降低了 1.6%，基尼指数升至 16.6%，优于仅气候模型。这表明，降雨强度提供了补充专家分区的信息，描述的是单次事件的严重程度，而非长期暴露。

与保险 + 气候（ins+c）模型相比，尽管两者的大尺度残差结构相似，ins+r 模型给出的预测结果更为细致、差异更明显（图6）。例如，在北部省和奥克西塔尼海岸线上，该模型产生了更平缓的残差梯度，而不是ins+c模型中观察到的普遍较高残差。这表明，纳入极端降雨信息能够提高风险估计的空间分辨率。

全变量模型，“GLM all”，见表 1 保险 + 降雨 + 气候 + 建筑变量最终，我们构建了一个完全整合模型，使用所有可用的数据来源，即保险记录、气象变量、气候相关指标以及建筑及周边环境的特征。残差误差分析显示，主要高误差区域显著改善，尤其是奥克西塔尼海岸沿线（图 6）。我们还观察到所有性能指标均有显著提升，而参数数量仅适度增加（从19个增至33个）。均方根误差从9000降至8781，基尼指数上升（从0.16升至0.24），反映出风险排序有所改善。

纳入建筑物和当地环境变量后，若干交互项变得显著。尤其是降雨强度（MILRE）与建筑特征之间的交互作用显著提升了模型拟合度，证实洪水损失由事件

严重程度和当地暴露条件共同驱动。在核保变量中，可移动资产的价值和建筑规模仍然很重要，但可移动资产与 MILRE 之间的交互作用成为最显著的预测变量。为更好地捕捉暴露机制并减轻共线性，我们引入了一个交互变量，将 TRI 分类与水道相关特征（至最近河流的距离和相对高差）结合起来。这一复合指标称为*WCTRII*，将 TRI 分区划分为九种暴露类别，并成为严重程度模型和发生模型中的核心预测变量。虽然将降雨累积量直接纳入该指标并未提升模型表现，但允许其与降雨变量显式交互则提供了更多信息。

图 7a 使用似然比检验对主要预测变量进行排序。总体而言，WCTRII和MILRE在模型中占主导地位，其次是由保险公司提供、并经我们的数据集丰富和补充的建筑变量。图7c 展示了新增变量如何使我们捕捉到ins+c中缺失的信息。尽管与建筑物相关的因素仍能提供有用信号，但其贡献次于WCTRII和MILRE，后两者直接刻画洪水暴露和事件强度。WCTRII至关重要，因为它量化了暴露：高WCTRII得分对应于靠近水道的低洼地区房产，从而使其处于洪水风险的最前沿。MILRE则刻画降雨强度，有助于区分常规事件与极端的灾难性事件。

3.4 概率建模结果与讨论

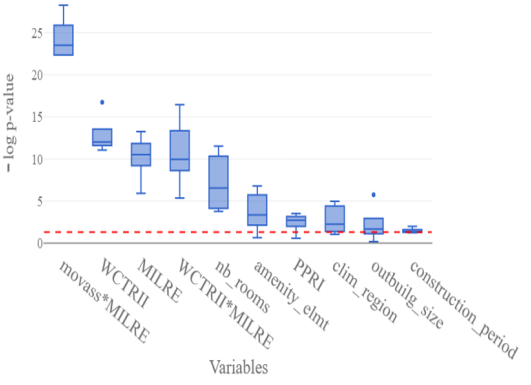
表3. 概率模型指标汇总，这些指标通过五折交叉验证计算，并给出其相对于仅使用保险公司数据的模型的相对变化。当指标改善时，变化以蓝色显示；否则以红色显示。召回率和精确率针对受洪水侵袭的建筑计算。

模型	-对数损失	Gini	CSI	召回率	精确率	参数数量
GLM ins 仅保险公司数据	0.6631	0.2339	0.0049	**0.1577**	0.0051	**17**
GLM ins+c 保险公司数据 + 气候	0.6245 ↓5.82%	0.3766 ↑61.01%	0.0075 ↑59.06%	0.0902 ↓42.80%	0.0083 ↑62.75%	21 ↑23.53%
GLM ins+r 保险公司数据 + 降雨	0.6082 ↓8.28%	0.4770 ↑103.93%	0.0159 ↑224.49%	0.0562 ↓64.36%	0.0233 ↑356.86%	21 ↑23.53%
GLM all 所有数据	**0.5614** ↓15.34%	**0.5462** ↑133.52%	**0.0261** ↑432.65%	0.1280 ↓18.83%	**0.0342** ↑570.59%	40 ↑135.29%

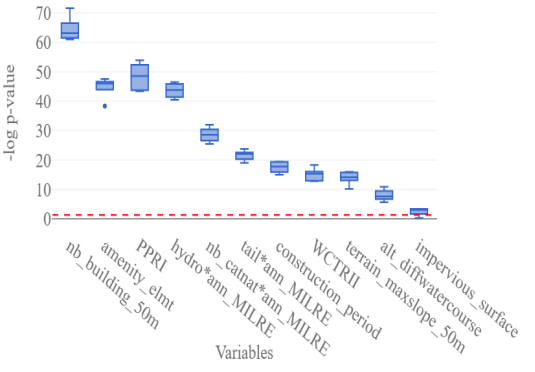
表 3 总结了概率模型的性能指标。鉴于我们的数据集中类别严重不平衡，其中 99.78% 的建筑物未被洪水淹没，只有 0.22% 的建筑物被洪水淹没，因此将结果与几乎只预测多数类的哑模型进行比较并不相关。

保险公司模型“GLM ins”，表 **1** 仅使用保险数据
与成本模型一样，仅保险人模型呈现出的空间变异较小，这一点可从残差图中看出（图 8）。在该模型下，我们正确识别出15.77%的受洪水侵袭的建筑，但这需要将全部观测数据中的6%标记为受洪水影响，因此CSI和精确率都很低。此时，_便利设施（即水池/户外设施的存在）成为影响最大的预测变量。这反映了脆弱性差异：户外要素通常在损害发生时具有较低的阈值，因此即使相对轻微的地表径流事件也会造成损失，而这类事件可能不会到达或影响主体建筑结构。

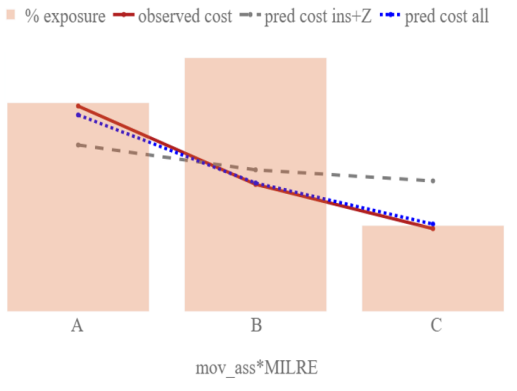
保险公司 +气候模型，“GLM ins+c”，表 **1** 保险 +水文与气候变量接下来，我们纳入源自水文洪水区的气候相关专业知识变量以及

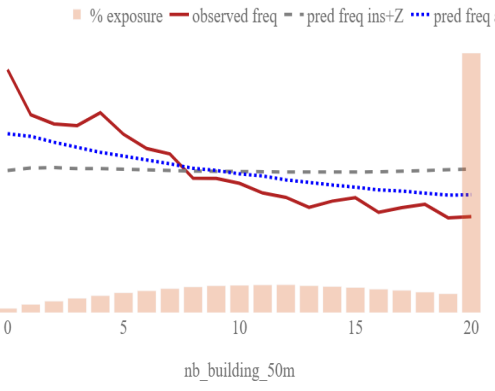


(a) 某成本模型的变量重要性。



(b) 某概率模型的变量重要性。





(c) 关于 mov_ass*MILRE 的平均观测洪水损失和预测洪水损失。(d) 关于 nb_building_50 的平均观测洪水频率和预测洪水频率。

图7。上方图表展示似然比检验的结果（对数尺度上的p值），用于评估变量重要性。红色虚线表示显著性阈值。对于下方图表，数据集按相应变量进行“分组”，柱形表示总暴露量，圆点表示各组的平均观测值和预测值。为保密起见，纵轴数值已屏蔽。

历史分类。该扩展模型的大多数性能指标均显著提升，尤其是基尼系数和CSI指数（表 3），这表明概率排序和校准效果更好。按照这一设定，实际遭受洪水侵袭的建筑物中约有9%被正确识别，而整个保险组合中仅有2%被标记为受洪水影响，表明精确率明显提高。空间区分度大幅改善，但部分地区仍被高估，这与基于专家判断的分区粒度相对粗略相符（图 8）。奥克西塔尼南部就是一个典型例子。由于该地区曾发生严重的总损失，且主要由强烈的地中海降雨造成，分区变量将整个地区一律标记为“高风险”（见图 11）。因此，模型在该区域内处处赋予较高概率，却未能反映实际物理风险的高度异质性；这种风险主要集中在特定水道及径流路径沿线。

保险公司 +降雨模型，“GLM ins+r”，见表1 保险 + 降雨变量

接下来，我们通过ann_MILRE变量纳入降雨强度信息，该变量衡量每栋建筑物所在位置对极端降雨事件的年度暴露量。`ins+r` 模型实现了性能提升，预测结果更为精确，空间区分度也有所改善。在此

设定下，5%的受洪水侵袭的建筑被正确检出，而样本中所有建筑物的0.5%被标记为受洪水影响。总体而言，残差变得更加结构化，尤其是在奥克西塔尼沿海地区以及北部地区中部等高误差区域。这些改进在图8的残差模式中清晰可见，同时也体现在更高的基尼系数和精确率得分上。

从概念上讲，ann_MILRE充当一个动态时空滤波器，能够有效区分遭受强降水的年份–位置组合与仅面临有限气象风险的组合。在数据中，当ann_MILRE保持在中等水平以下时，洪水几乎从不发生。这表明极强降雨是发生洪水的必要条件，但并非充分条件。因此，我们将该变量理解为引导模型将其学习“集中”于已发生潜在洪水触发事件、因而更难分类的区域。

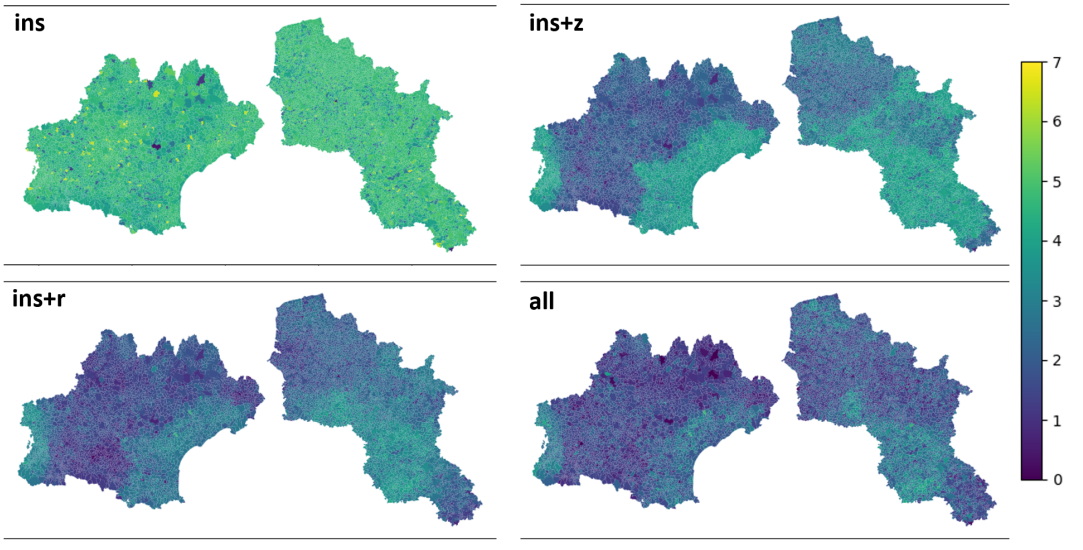


图8. 对于表3 中定义的四个广义线性模型概率模型以及图2中定义的两个地区，皮尔逊残差在市镇层面显示。与零的显著偏离表明观测值与拟合值之间存在拟合偏差。

全变量模型，“GLM all”，表1 保险 + 降雨 + 气候 + 建筑变量 最后，我们构建了一个综合模型，整合所有可用的数据来源。在该设定中，我们还纳入了WCTRII变量；该变量在成本模型中具有显著性，并且对概率估计仍然具有相关性。该模型在概率校准方面取得了显著提升，其幅度与首次引入空间信息时所实现的初始改进相当（见指标，表3）。该模型正确识别了12.80%的受洪水侵袭的建筑，同时仅将保险组合中0.7%的部分标记为受洪水侵袭，从而进一步限制了假阳性。针对受洪水侵袭的建筑的预测结果精确率大幅提高，同时与前一模型不同，仍保持了较高的召回率。残差图颜色更深，反映出所有地区的预测性能均有所改善。此外，此前因开放数据中的大尺度模式而被高估的高风险区域如今得到了更准确的划定，凸显了高分辨率地理定位信息的关键优势之一。与粗略的分区分类相比，此类数据支持精细得多的风险分层，降低了保险组合中大段被错误标记为高风险或被忽视为低风险的可能性。

图7b 显示，建筑属性和环境背景在概率模型中的影响强于在成本模型中的影响。最重要的预测变量是 nb_建筑物_50m，

紧随其后的是 amenity_elmt。变量 nb_建筑物_50m 用于量化半径为 50 米的范围内建筑物的数量，在概率模型中尤其具有参考价值，因为它同时反映了城市化模式以及附近水体造成的空间限制。洪水风险高的区域通常建筑密度较低，这是分区法规和过去洪水遗留影响共同塑造的结果。因此，这些区域内的建筑物往往 nb_建筑物_50m 值较低，使这一指标成为反映洪水暴露的有用空间代理变量。当建筑物位于河流或其他水体附近时，50 米缓冲区的一部分会与水体重叠，从而减少统计到的周边建筑物数量。这种自然截断有效反映了邻水建筑物面临的更高洪水风险，因为它们更容易受到溢流影响。此外，地形和防洪措施也会影响建筑密度，这可能进一步强化 nb_建筑物_50m 值较低与洪水风险之间的联系。

变量 ann_MILRE 出现在三个排名靠前的交互项中，证实其在捕捉风险年度变异方面发挥着关键作用（图7b）。变量 nb_catnat 标记了曾发生过洪水的区域，无论洪水发生频繁、偶尔还是罕见。尽管这类信息很有参考价值，但它是静态的：在没有极端事件的年份，即使是历史上的易发洪水区域，实际风险也可能很低；而在发生严重洪水的年份，过去洪水较少的区域也可能遭受重创。这正是与 ann_MILRE 的交互作用发挥价值之处：它们为历史信息提供背景，从而实现动态风险评估。

3.5 讨论

官方洪水分区是法国洪水风险管理的重要组成部分。在现有数据集中，重要洪水风险区（TRI）对易发洪水区域的划定最为详细、权威，尽管其空间覆盖范围有限（主要受实施成本影响），且其设计初衷主要是用于监管规划，而非保险定价。因此，在下文中，我们将 TRI 同时用作解释性基准和参考真实值。

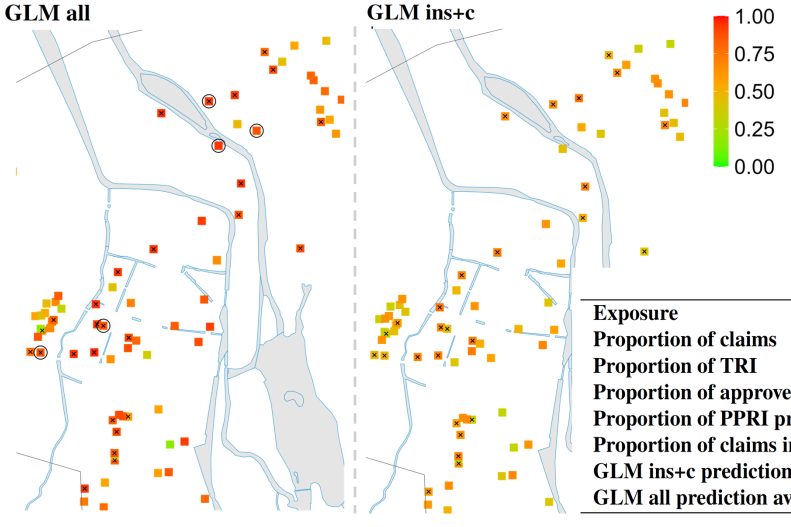
成本模型和概率模型的估算结果表明，在有官方分区数据的地区，官方分区具有很强的解释力。在我们的数据中，只有5%的建筑物位于TRI洪水区内，但这些区域却涵盖了约11%的记录在案的洪水事件。在TRI高风险区，这种差异更加显著：其中的建筑物不足1%，却约占洪水发生次数的4%，且平均成本是TRI区域外建筑物的1.6倍。这种集中分布说明，基于专家判断绘制的洪水地图有助于识别结构性暴露区域。在TRI区域内，纳入分区数据的模型与完全整合的设定几乎表现一致，无论是预测概率还是损失水平皆如此。这种相似性表明，我们框架中的额外变量与监管洪水地图所体现的专家判断相契合，而非与之相悖。该框架的主要增值体现在TRI划定区域以外的地区，这些地区约占保险组合的95%。在这些地区，官方洪水地图几乎无法提供指导，因此保险公司必须借助其他信息来源来区分风险水平。

在成本模型中，将降雨强度（MILRE）与详细的环境指标和水道相关变量结合起来，可以识别出平均损失显著较高的非 TRI 建筑群体，这一区别对定价至关重要。在 TRI 区域以外的这些高风险群体中，保险公司的基准模型加上气候变量只能解释观测总成本的 57%，而完全整合模型则可解释 97%。然而，在整个非 TRI 区域内，两种模型对总损失的解释比例相近。这表明，新增变量并未提高整体风险估计，而是更准确地刻画了未分区区域内损失的分布，从而减少高暴露风险特征的定价过低，并抑制定价风险较低对象时的定价过高。

概率模型也呈现出类似趋势。在 TRI 区域之外，`ins+c`模型会系统性地高估未受洪水侵袭的建筑的受灾概率，其赋值比`all` 模型高约 15%。相反，对于曾遭洪水侵袭的建筑，`all` 模型

给出的概率平均高出约 10%，表明其校准效果更好，也能更准确地识别易受灾结构。尤其是，`all` 模型还标示出官方气候风险清单（TRI、PPRI 等）中不存在、但具有明显加剧因素的高风险区域，例如陡坡、自然滞蓄区、大面积不透水表面以及重尾降雨。我们认为，这表明官方洪水地图可能不完整、已过时，或未能设置为捕捉某些洪水过程。该模型还识别出一些鲜为人知的水道，尽管它们未被划为主要易发洪水区域，却仍可能构成灾害。

巴黎以南的 Le Perray-en-Yvelines 市镇及其周边地区，是一个很有代表性的案例。近年来，上游城市化显著扩大了不透水表面的范围，改变了强降雨期间当地的水文行为。尽管当地报告提及过径流引发的洪水，但该地区位于 TRI 和 PPRI 边界之外，而且在我们的校准期内，GASPAR 数据库中没有记录任何洪水事件。尽管缺乏官方分区，`all` 模型仍将该市镇的多栋建筑物识别为高风险，与之相反，气候调整保险模型（`ins+c`）则将这些建筑物评为接近组合平均值。这种差异源于`all` 模型考虑了不透水程度、地形坡度和极端降雨。在本研究期间，即2024年11月，柯克风暴给 Perray-en-Yvelines 带来了洪水。随后颁布的两项部长令正式认定当地发生过洪水事件，这也是该市镇首次获此认定。虽然这只是一个个案，但这一事件表明，纳入精细尺度暴露度和降雨信息的基于保险的模型，能够在洪水风险出现在官方数据集之前揭示其新兴趋势。



Area statistics	
Exposure	219
Proportion of claims	0.114
Proportion of TRI	0
Proportion of approved PPRI	0
Proportion of PPRI proposal	0.367
Proportion of claims in TRI/PPRI	0.06
GLM ins+c prediction average	0.61 (0.11)
GLM all prediction average	0.685 (0.19)

图9. Montargis 市区一段区域中 ins+c 模型与 all 模型的预测洪水发生概率对比。每个方格代表一栋建筑物；十字表示保险组合中至少有一次已观察到的洪水理赔的建筑物。圆圈标出的建筑物对应于模型预测结果差异最大的地点。表格汇总了所示区域的关键统计数据。

Montargis 市（图 9）展示了一种洪水风险广为人知、但仅得到部分正式认定的情况。自 1980 年以来，该市镇多次遭遇洪水，GASPAR 数据库中列有 12 次部长级洪水认定。然而，该市尚未被指定为 TRI，自 2018 年以来也仅有一份未经批准的 PPRI，这可能反映出该镇人口较少（15,000 人），经济利益有限。在这种情况下，气候增强保险模型（`ins+c`）在大部分城区给出的概率估计大致均一，主要依据粗尺度的气候和分区变量。相比之下，全 模型加入了更精细的

空间细节，包括至河流距离、相对高程和局部建筑布局等指标。如图 9所示，这使得河流廊道沿线和低洼地区的预测概率较高，这些地方的理赔记录也更常见；而市镇其他区域的风险则更接近基线风险。圈出的建筑物标示了两个模型差异最大的地点，揭示了值得进一步考察的暴露模式。

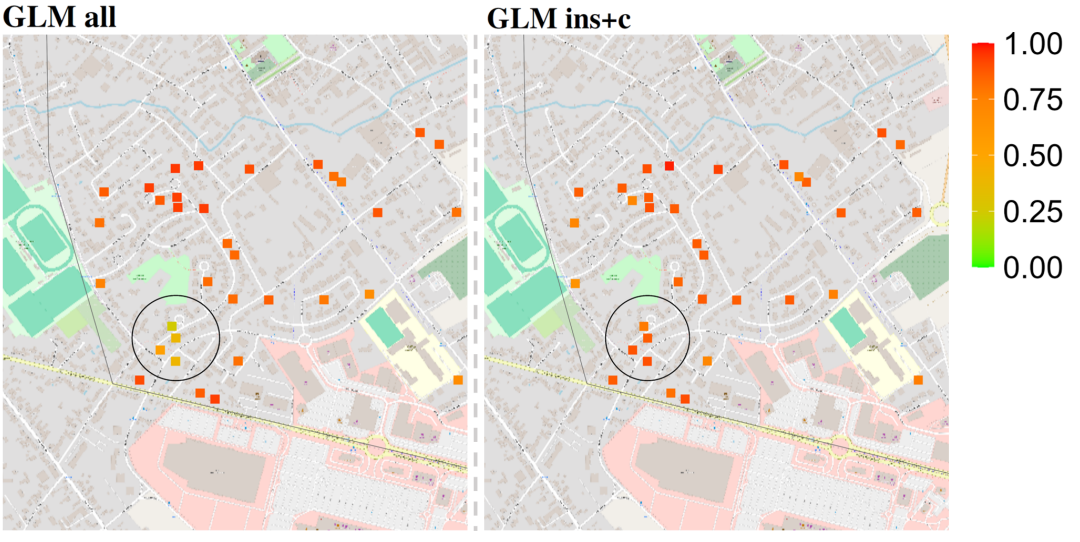


图 10. 对比 Saint-Memmie 市一处区域的 ins+c 模型与 all 模型预测的洪水发生概率。每个方块代表一栋建筑物。圆圈标出的区域显示了两个模型给出的概率估计差异最大的位置。

图 10 展示了 Saint-Memmie 的另一种情况：当地既有获批的 PPRI，也有官方确认的洪水事件。在这种情况下，`ins+c`模型已能捕捉洪水风险的主要空间分布模式，在受管制洪泛区内给出的预测结果高于平均值，但相当均一。`all`模型的增益在于细化局部风险估计。图 10 中圆圈标出的区域内，建筑物地势高于河流，坡面驱动径流有限，积水倾向较低。对于这些建筑物，`all` 模型给出的概率低于`ins+c`模型，表明尽管它们位于官方绘制的大尺度洪水区内，其暴露程度仍较低。PPRI 的保险保障范围很广，但地图相对粗略。这个案例说明，细粒度数据可以补充现有分区：在已有监管地图的情况下，高分辨率环境与建筑属性有助于避免过度平滑，揭示大致划定的风险区域内部微尺度的暴露差异。

4. 结论

本研究采用传统的广义线性模型（GLM）框架，旨在剥离数据而非建模技术的影响，考察整合详细的地理定位信息如何提升保险领域的洪水风险评估。仅依赖核保数据的模型，其解释能力和空间区分能力较弱；而在能够获取相关信息的情况下，纳入官方分区以及专家气候与水文变量可显著提升模型表现。然而，最大的改进来自在建筑物层面纳入高分辨率环境与降雨数据，这些数据能够更好地捕捉局地暴露机制和事件强度。这种精细化建模在官方洪水区之外尤为有价值，既能实现更细致的风险分配，也能更准确地为高暴露房产定价，并减少定价过高

低风险房产的定价过高现象，同时不挑战现有监管分类。

解释变量的作用因建模目标而异。严重程度模型主要由暴露指标和事件强度（例如与水道的距离、相对高程、极端降雨指标）驱动，并辅以被保险金额，在发生洪水后对损失规模进行调整。因此，其他新增建筑变量所起的作用有限。相比之下，发生模型更依赖建筑密度、土地覆盖和局地地形等周边环境描述变量，这些变量决定洪水是否发生。年度汇总降雨指标也有助于提升模型表现。

本研究中的降雨变量并非用于短期或实时预测，这超出了本研究的范围。相反，在定价和投资组合风险评估中，可根据不同强度水平对预期损失进行条件化，并在其经验分布上积分，从而得出符合精算原则的无条件风险估计。除了直接提升预测能力外，这些变量还能通过捕捉事件严重程度、减少遗漏变量偏差，以及提升其他协变量预测效应的稳健性和可解释性，来稳定模型。汇总方式选择反映了保险应用场景，旨在捕捉多个时间跨度内最强的降雨事件。在模型开发过程中，我们测试了不同的汇总时间窗口，结果一致，表明关键在于对比短期和长期强度测量，而非采用何种确切的时间定义。尽管该降雨指标是基于法国数据校准的，但其构建依据局部极端降水暴露的一般原则，因此有可能推广至降雨信息具有可比性的其他地区。

除了定价之外，这些结果也对预防和适应具有重要意义。通过凸显当地小尺度特征的影响，基于保险的模型可以补充公共灾害地图，并支持更有针对性的风险意识提升和减灾工作。从这个意义上说，高分辨率数据不应被视为排斥他人的工具，而应成为通过更充分知情的决策增强集体韧性的手段。在以国家团结为基础的法国自然灾害赔偿制度中，改善风险区分并非旨在限制保险保障。相反，更精准的建模有助于实现更公平的定价、改进投资组合监测，并制定更有效的预防策略，同时遵循现有监管框架。

总体而言，本研究表明，结合地理定位环境数据的精算模型能够通过补充空间细节以及与保险相关的暴露视角，有效补充官方洪水风险评估。更广泛而言，通过数据整合加强洪水风险评估，是实现长期气候韧性和备灾的关键组成部分。

致谢 我们感谢 ADDACTIS France 提供本研究所用保险数据的访问权限。

利益冲突 无。

数据可用性声明 本研究中用于构建解释变量的所有数据均为开源数据。法国灾害数据库（GASPAR）同样为开源数据。出于保密原因，本研究使用的保险组合不予提供。

资助声明 本研究未获得任何资助机构（无论商业机构还是非营利机构）的专项资助。

注释

参考文献

André, C., Monfort, D., Bouzit, M., & Vinchon, C. (2013). 保险数据对住宅建筑沿海洪水损失成本评估的贡献：从 johanna (2008) 和 xynthia (2010) 风暴事件中获得的见解。《自然灾害与地球系统科学》，13(8), 2003–2012。Assunção, R., Costa, M. A., Prates, M. O., & Silva e Silva, L. S. G. (2014). 空间分析。载于 A. Charpentier (编)，R语言计算精算科学（第 207–256 页）。Chapman & Hall。Bézy, T. (2023). 洪水风险的影响。Blöschl, G., Hall, J., Viglione, A., Perdigão, R. A., Parajka, J., Merz, B., Lun, D., Arheimer, B., Aronica, G. T., Bilibashi, A., 等 (2019). 气候变化既增加也减少了欧洲河流洪水。《自然》，573(7772), 108–111。Boskov, M., & Verrall, R. (1994). 基于空间模型的地理区域保费定价。《ASTIN公报：国际精算协会期刊》，24(1), 131–143。Brockman, M. J., & Wright, T. (1992). 机动车统计定价：有效利用数据。《精算师协会期刊》，119(3), 457–543。欧洲洪水风险中心（CEPRI）。(2022). 法国实施欧洲洪水指令 [在线；访问日期：2024-12-16]。CES OSO – THEIA。(2023). 土地覆盖图。Chatelain, P. (2023). 个人住宅保险按地址定价 [博士论文，克洛德·贝纳尔里昂第一大学]。Coles, S., Bawa, J., Trenner, L., & Dorazio, P. (2001). 极值统计建模导论（第208卷）。施普林格。法国国家地理与森林信息研究所 (I. N.)。(2019). 地形数据库。Denuit, M., & Lang, S. (2004). 基于贝叶斯广义加性模型的非寿险费率厘定。《保险：数学与经济学》，35(3), 627–647。Denuit, M., Sznajder, D., & Trufin, J. (2019). 基于洛伦兹曲线、集中曲线、基尼指数和凸序的模型选择。《保险：数学与经济学》，89, 128–139。法国公共财政总局 (DGFiP)。(2019). 土地价值查询（dvf）：房地产销售查询。环境、规划与交通区域及跨省部门局。(2023). 洪水风险预防计划。《2000/60/EC号指令》。(2000). 欧洲议会和理事会于2000年10月23日颁布，旨在建立共同体在水政策领域采取行动的框架 [存档]。eur-lex.europa.eu。Dorfman, R. (1979). 基尼系数公式。《经济学与统计学评论》，146–149。Fortin, G., Poirier, C., Duhamel, F., & Germain, D. (2020). 洪水风险与脆弱性：以加拿大新不伦瑞克省肯尼贝卡西斯河流域为例。《IdeAs：美洲思想》，(15)。France Assureurs。(2021). 气候变化对2050年展望下保险业的影响。FA研究报告。France Assureurs。(2023). 2022年自然灾害保险，财产与责任保险。FA研究报告。Frees, E. W., Meyers, G., & Cummings, A. D. (2014). 保险费率厘定与基尼指数。《风险与保险期刊》，81(2), 335–366。Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2016). 保险定价的广义线性模型。意外险精算师协会，CAS专著系列，5, 77。Grislain-Letrémy, C., & Peinturier, C. (2010). 法国自然灾害保险制度。Guo, H., & Viktor, H. L. (2004). 通过提升法和数据生成从不平衡数据集中学习：DataBoost-IM方法。《ACM SIGKDD探索通讯》，6(1), 30–39。He, H., & Garcia, E. A. (2009). 从不平衡数据中学习。《IEEE知识与数据工程汇刊》，21(9), 1263–1284。法国国家地理与森林信息研究所。(2022). RGE ALTI高程数据。Japkowicz, N. (2013). 不平衡学习的评估指标。《不平衡学习：基础、算法与应用》，187–206。Jeni, L. A., Cohn, J. F., & De La Torre, F. (2013). 应对不平衡数据：性能指标使用建议。2013年人类情感计算与智能交互协会会议，245–251。Joly, D., Brossard, T., Cardot, H., Cavailhes, J., Hilal, M., & Wavresky, P. (2010). 法国本土的气候类型：一种空间构建。《Cybergeo：欧洲地理学杂志》。Legrand, J., Naveau, P., & Oesting, M. (2021). 渐近相关与渐近独立极值的二元分类器评估。arXiv预印本 arXiv:2112.13738。Luu, L. N., Vautard, R., Yiou, P., van Oldenborgh, G. J., & Lenderink, G. (2018). 利用 EURO-CORDEX 模拟归因法国南部极端降雨事件。《地球物理研究快报》，45(12), 6242–6250。Merz, B., Kreibich, H., Thieken, A., & Schmidtke, R. (2004). 建筑物洪水直接经济损失估算的不确定性。《自然灾害与地球系统科学》，4(1), 153–163。https://doi.org/10.5194/nhess-4-153-2004 Merz, B., Kreibich, H., & Lall, U. (2013). 多变量洪水损失评估：一种基于树的数据挖掘方法。《自然灾害与地球系统科学》，13(1), 53–64。法国气象局。(2024). 气候概览：一些地理知识 [访问日期：2024-03-29]。法国生态转型部。(2023a). 风险相关行政程序辅助管理国家数据库（GASPAR）。法国生态转型部。(2023b). 法国洪水风险概述。

Muñoz Sabater, J. (2019). ERA5-Land 1950年至今逐小时数据［访问日期：DD-MMM-YYYY]. https://doi.org/10.24381/cds.e2161bac Nelder, J. A., & Wedderburn, R. W. (1972). 广义线性模型. 英国皇家统计学会杂志A辑：社会统计, 135(3), 370–384. Philip, S., Kew, S. F., van Oldenborgh, G. J., Aalbers, E., Vautard, R., Otto, F., Haustein, K., Habets, F., & Singh, R. (2018). 验证气候变化对2016年5月至6月法国致洪降水的快速归因研究. 《水文气象学杂志》, 19(11), 1881–1898. Schaefer, J. T. (1990). 临界成功指数作为预警技巧指标. 《天气与预报》, 5(4), 570–575. Scheel, I., Ferkingstad, E., Frigessi, A., Haug, O., Hinnerichsen, M., & Meze-Hausken, E. (2013). 一种纳入空间变量选择的贝叶斯层次模型：天气对保险索赔的影响. 英国皇家统计学会杂志C辑：应用统计, 62(1), 85–100. Spekkers, M., Ten Veldhuis, M.-C., Kok, M., & Clemens, F. (2011). 基于荷兰保险公司数据的暴雨洪涝损失分析. 载于《国际城市洪水风险管理研讨会》论文集. Spekkers, M., Kok, M., Clemens, F., & Ten Veldhuis, J. (2014). 影响降雨导致的建筑结构及内部财产损失的因素之决策树分析. 《自然灾害与地球系统科学》, 14(9), 2531–2547. Taylor, G. C. (1989). 利用样条函数按地理区域厘定保费. 《ASTIN公报：国际精算协会期刊》, 19(1), 91–122. Torgersen, G., Rød, J. K., Kvaal, K., Bjerkholt, J. T., & Lindholm, O. G. (2017). 运用多变量建模技术评估城市区域房产的洪水暴露. Water, 9(5), 318. Wahl, J. C., Aanes, F. L., Aas, K., Froyn, S., & Piacek, D. (2022). 水损保险风险保费的空间建模. 《斯堪的纳维亚精算杂志》, 2022(3), 216–233. Wang, Y.-X., Ramanan, D., & Hebert, M. (2017). 学习对尾部进行建模. 《神经信息处理系统进展》, 30. Zhou, Q., Panduro, T., Thorsen, B. J., & Arnbjerg-Nielsen, K. (2013).基于保险数据的洪水损失建模验证. 《水科学与技术》, 68(2), 425–432. Zhu, M., Xia, J., Jin, X., Yan, M., Cai, G., Yan, J., & Ning, G. (2018). 用于处理类别不平衡医疗数据的类别权重随机森林算法. 《IEEE Access》, 6, 4641–4652.

附录1. 数据来源和变量构建的补充细节附录*1.1* 水文与气候指标的补充信息

- **TRI（洪水风险地区）：** TRI 区域依据欧盟洪水指令（2007/60/EC）设立，旨在确定面临重大洪水风险的区域，并指导国家和欧洲层面的洪水风险管理措施优先排序（欧洲洪水风险中心（CEPRI），2022年）。其目的是识别洪水灾害程度高且人口稠密、资产价值高或经济活动密集的领土，而非绘制覆盖整个领土的完整洪水地图。

  TRI 区域的界定依赖多种数据来源，包括有记录的过往洪水事件、具有代表性的洪水情景的水文和水力建模，以及人口密度和经济价值等暴露度指标。这些要素会结合分析，以评估领土层面的潜在洪水影响。因此，TRI 边界体现了对洪水风险的综合评估，同时纳入了灾害和暴露度维度。在 TRI 边界内，洪水灾害分为高、中、低三个等级，通常对应逐渐降低的超越概率，常被解释为约 10 年、100 年和 1,000 年的重现期。这些类别描述了参考情景下预期的洪水强度和空间范围。它们并未明确考虑建筑物层面的脆弱性、地方防护措施（某些情况下）或洪水发生的年际变化。需要强调的是，TRI 的构建方法受到严格规范，但各地区之间仍可能有所不同。高质量水力建模、先进的数字高程模型及其他数据的可用性，都会影响特定 TRI 的整体质量。尽管存在这些局限，TRI 目前仍代表法国公共机构所开展的最先进洪水评估水平。

  按照设计，TRI覆盖范围仅限于被认定为洪水风险具有国家重要性的区域。因此，法国领土中有相当大一部分未被分类。在我们的数据集中，约95%的建筑物位于TRI边界之外。因此，我们明确引入“无”类别，以区分TRI覆盖范围之外的区域与TRI 区域内被归类为低

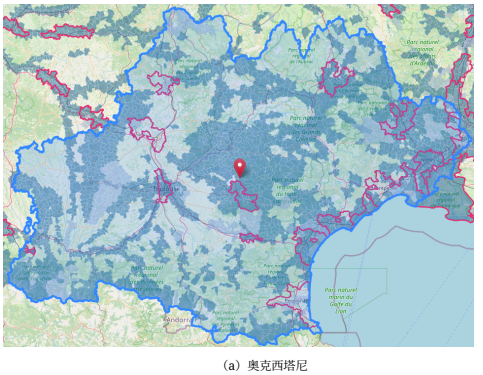
灾害的区域。

- **洪水自然灾害预防计划（PPRI）：**这些地图是地方监管工具，在市镇或跨市镇尺度上制定，旨在指导土地利用规划并限制洪水暴露（Direction régionale et interdépartementale de l’environnement, de l’aménagement et des transports，2023年）。与旨在支持国家层面优先排序的 TRI 不同，PPRI 服务于地方决策和分区法规。

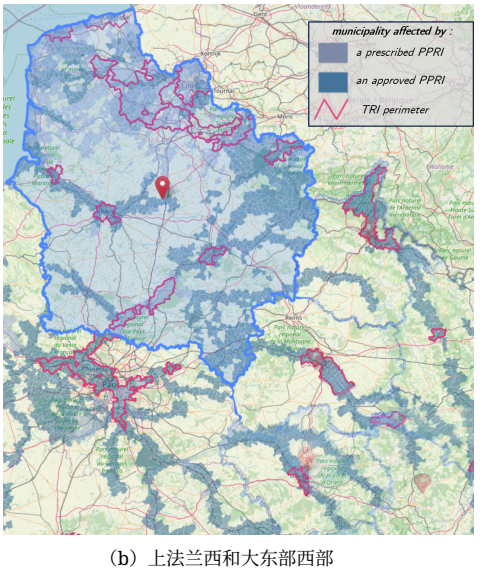
  PPRI 通常依据参考洪水的水力模拟制定，参考洪水最常以百年一遇重现期定义；当历史极端事件超过参考情景时，也可将其纳入补充依据。模拟会使用河流几何形态、数字高程模型提供的地形数据，以及简化的水流表示，但通常不会考虑未来的土地利用变化、气候变化或建筑尺度的减灾措施。最终生成的地图将区域划分为从极低到极高的灾害等级，并据此转化为土地利用方面的监管限制。在我们的数据集中，详细的市镇内部洪水自然灾害预防计划占保险组合的1.4%，其中约20%与 TRI 区域重叠。

  PPRI 的保险保障分布不均，因为其编制和审批需要投入大量技术与行政资源。因此，许多市镇没有详细的市镇内部地图，只能依靠针对整个市镇潜在洪水风险的评估；这类评估可能已获批准、已过时，或仍在审核中（尚未获批）。关于 TRI，我们增加了“无”这一类别，以明确标示缺少 PPRI 信息的情况。从建模角度看，详细的 PPRI 与 TRI 功能相似；而市镇层级的 PPRI 扩大了空间覆盖范围并提升了模型性能，但代价是空间精度降低，且对潜在暴露区域采用了较粗略的分类。

  图 11 概述了我们研究地区的易发洪水区域，这些区域根据 TRI 和 PPRI 界定，并突出显示已知存在暴露风险的区域。



（a）奥克西塔尼



（b）上法兰西和大东部西部

图11. 用于洪水风险评估的法国专家参考地图。PPRI（洪水风险预防计划）标示了可能受到百年一遇参考事件影响的区域，为土地利用规划和风险减缓提供支持。PPRI 由省长批准。TRI（洪水风险区）针对洪水风险高且涉及重大利益的区域编制。采用三种情景（低、中和高重现期）来确定暴露面积及相关风险。

- **水文分　　区：**法国水文系统按层级划分为四个

嵌套层级：水文区域、水文分区、水文子分区和区域，分别对应《水框架指令》（《2000/60/EC号指令》，2000）规定的主要河流流域逐级划分的结果。这些流域被界定为这样一些区域：所有地表径流都经由连通的河网汇聚至同一出口，因此它们是河流洪水过程的天然空间参照。

在本研究中，我们采用水文分区层级，将法国本土划分为24个单元，各单元具有不同的河流系统、流域结构和水文响应（de l'Information Géographique et Forestière，2019年）。该变量并非旨在对上游—下游累积或动态集水区过程进行显式建模，而是作为区域分割因子，用于捕捉水文行为方面的广泛差异，例如径流生成、河流密度和主要洪水机制。文献中常采用类似的水文分区来控制具有空间连贯性的水文异质性（Fortin 等人，2020年）。

- **气候区：**气候分区反映了降水、温度和大气环流的长期空间模式，这些因素塑造了洪水形成过程。多种气候分类体系划分出法国本土的不同气候型（Joly 等人，2010年；法国气象局，2024年）。一种常用的分类方法区分出五种主要气候类型：海洋性、受影响的海洋性、大陆性、地中海性和山地气候，各自具有特定的降雨季节性和极端降水特征。

  法国气象局进一步将这一框架细分为29个气候亚区，以捕捉降雨强度、风暴类型和水文气象动力过程更细致的空间变异性。我们采用这一精细分区作为背景变量，以考虑区域间降水型和洪水暴露的差异；这些差异未在事件尺度上明确建模。与水文分区一样，气候区发挥的是结构性控制作用，而非单次洪水事件的预测变量。

- **GASPAR 数据库：**该数据库（风险相关行政程序辅助管理国家数据库）是一个国家级行政登记系统，用于汇总法国 CatNat 保险制度下对自然灾害的官方认定（法国生态转型部，2023a）。对于洪水，该数据库记录受影响的市镇、日期，以及与每项部级法令相关的灾害类型。

  该数据集为重大洪水事件的时空分布提供了宝贵的历史参考。不过，纳入记录须由市镇正式提出申请，并经部际委员会批准，这意味着 GASPAR 主要收录在行政层级上被认定为重大事件的洪水。因此，局地、低强度或未报告的洪水可能未得到充分记录。在本研究中，GASPAR 因此被用作历史洪水发生情况的背景指标，而非全面的洪水事件目录。

附录 *1.2* 从地址到变量

通常而言，保险数据集不会提供承保建筑的精确位置，而通常只包含市镇代码。尽管这一详细程度足以满足大多数空间分析和基于分区的风险评估（Wahl et al., 2022年），准确的地理定位能够纳入精细尺度的环境和结构信息。为充分利用这一潜力，我们设计了一个多步骤地理编码工作流程，将投保人地址与具体建筑物关联起来。首先，我们清理并规范化地址记录，使其符合法国国家地址数据库（BAN），即法国官方地址数据库的格式。这种统一处理提高了数据的一致性和匹配效果。随后，将每个规范化后的地址与一栋或多栋建筑物关联起来；这些建筑物来自一个综合建筑物数据库，该数据库整合了多个开放数据源，涵盖研究区域内所有已知的地理定位建筑物。在 25% 的情况下，一个地址对应多栋建筑物；对于这些情况，我们仅保留主要建筑物，因为其余建筑物通常是附属建筑或次要结构。为减少建筑物匹配中的误分类，我们

依据业务规则和保险公司提供的信息进行完整性检查，确保信息与投保人记录一致。在确定承保建筑物及其精确 GPS 坐标后，可以结合运用机器学习、计算机视觉、自然语言处理和地理信息系统技术，自动从地图、影像、文本文件及相关来源中提取信息。这使我们能够构建详细变量，用于描述建筑物及其周边环境。有关将原始数据转化为结构化风险变量所采用方法的更详细说明以及示例，请参见 Chatelain，2023年。该参考文献探讨了建筑物及周边环境数据集的开发，并展示了其在保险风险建模中的应用。

附录 *1.3* 建筑物及周边数据构建方法

根据底层数据来源的结构和限制，采用不同的处理策略来构建解释变量，相关内容将在以下各节中介绍。示例结果以市镇层面呈现，并使用中位数概括研究区域内各变量的情况。

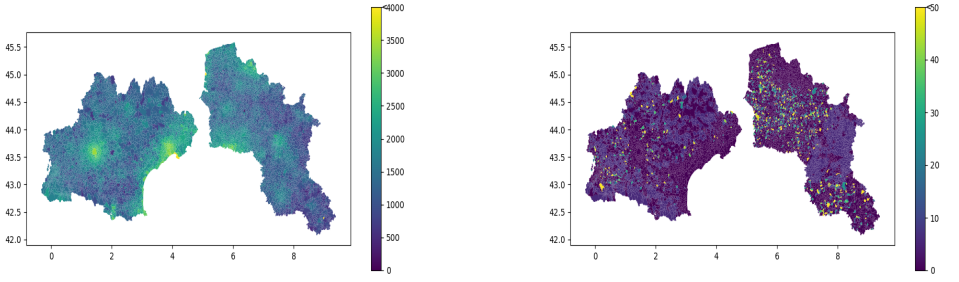
集成数据 是指结构化的开放数据集，经过有限转换即可直接关联到保险组合。根据其空间属性，集成工作可通过行政标识符（例如市镇代码）或几何对象之间的空间连接来完成。图 13a展示了每栋建筑物周围 200 米缓冲区内不透水表面的比例，该数据源自国家土地覆盖数据集“Occupation des Sols”（CES OSO – THEIA，2023年）。该栅格产品的分辨率为 10 米，通过与建筑物邻域相交来量化局部土壤封闭程度。在主要城市区域及其周边可观察到较高的数值，各地区之间呈现出明显的空间梯度。

与补全数据、创建数据和计算变量不同，集成数据不依赖推断或建模。后续的数据类别体现了更具创新性的特征，这些特征要么源自非结构化数据源，要么需要经过大量处理才能在建筑尺度上加以利用，要么结合了多个无法在标准数据集中同时获取的背景输入数据。

补全数据来源于覆盖范围全面、但在建筑物层面存在缺漏的开放数据源或行政数据源，因此无法直接用于保险建模。缺失值通过监督式机器学习模型进行推断；这些模型基于记录完善的观测数据进行训练，并结合内部保险数据、建筑特征和空间背景信息加以丰富。房屋价值就是一个典型例子。法国所有房地产交易的记录均可通过 DVF 数据库获取（法国公共财政总局（DGFiP），2019），但近年来仅有一小部分建筑物进行过交易。为获得完整的建筑物层面变量，研究人员利用观测价格训练预测模型，所用解释特征包括建筑物占地面积、楼层数、位置及周边特征，再用模型填补缺失值。图12a 展示了由此得出的房屋平均价值，较高值集中在主要城市区域。

创建数据 是通过先进的数据处理技术从原始、非结构化来源中提取的，包括对航空影像或行政文本等数据应用计算机视觉和自然语言处理。这些变量捕捉了传统数据集中无法直接获取的信息，例如附属建筑的存在情况和规模。为确保可靠性，生成的变量会与人工标注的参考样本进行系统验证，将预测结果与像素级图像标注或整理后的文本摘录进行比较。图 12b 展示了附属建筑的空间分布，表明其在高密度城市区域中较少见。该变量是创建数据的一个范例，因为它是通过对卫星影像应用图像识别模型生成的，而不是通过补充已有的结构化数据集得到的。

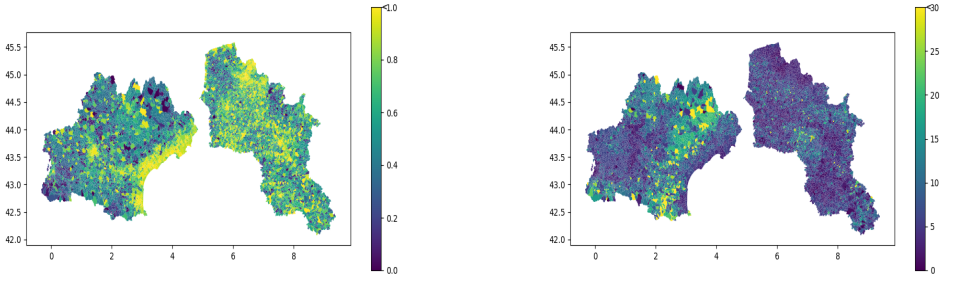
计算数据 计算变量是通过空间分析或简化的物理指标组合多个数据集得出的。通常需要进行几何运算、邻域聚合，或将连续空间场转换为建筑级指标。图 13b展示了一个示例，表示每栋建筑与其周边100米半径范围内地形之间的最大高程差。该变量利用国家数字高程模型（法国国家地理与森林信息研究所，2022年），通过将高程数据与建筑足迹进行空间相交，并汇总局部海拔梯度来计算。它反映了影响地表径流汇集和流向的局部地形差异。正如预期，在奥克西塔尼等山区地区观测到的数值较高。



(a) 房产价值，补全数据示例。

(b) 附属建筑存在情况，创建数据示例。

图 12. 按市镇层面汇总的变量中位数。



(a) 建筑物周围 200米缓冲区内不透水表面的百分比，集成数据示例。

(b) 给定建筑物之间以米为单位的最大海拔差以及100米缓冲区内的环境数据，计算数据示例。

图13。变量　　按市镇层面的中位数进行汇总，展示数据构建的四种不同方法　　.

表 4. 本研究所考虑变量概览。ML 指机器学习。

变量名称	取值	全局统计	受淹统计	详细信息
保险		针对所有 建筑物生成	针对受洪水侵袭的 建筑物生成	统计值：平均值（标准差） 或比例 %
索赔金额 单笔索赔费用	[5, 69k]	-	5884.13（9093.98）	特定合同的索赔金额 和洪水发生
索赔数量 索赔数量	{0,1}	0.221% (0.0473)	-	-
房间数量 房间数量	[2, 6]	3.68 (0.8)	3.86 (0.79)	-
可移动资产价值 可移动资产价值	[0, 456k]	[5千,10千]：15.77% [10k，20k]：14.06%	[30k,45k]：16.16% [10k,20k]：14.82%	离散化形式中使用的连续变量 保密目的
贵重物品 贵重物品价值	[0, 25k]	[0, 5k]: 91.34% [5k, +∞]：8.65%	[0, 5k]：86.91% [5k, +∞]：13.09%	以离散化形式使用的连续变量 保密目的
便利设施要素 便利设施	存在, 不存在	不存在: 93.01% 存在: 6.99	不存在：84.18% 存在：15.82	主要为水池的存在
附属建筑_尺寸 最终附属建筑尺寸	存在、不存在	不存在：98.56% 存在：1.44	不存在：98.25% 存在：1.75	离散化为布尔值 用于质量问题
专家气候				
tri 河流溢流	低、中，高，无	无：94.94% 中：2.44%	无：90.23% 中：4.56%	-
tri 径流	低，中，高，无	无：99.98% 中：0.01%	无：99.95% 中：0.03%	-
ppri 除浸没外的所有类别	7个类别	无：98.08% 低：0.75%	无：93.11% 高：2.62%	类别从极 高到无
水文区 集水区	24 区域	-	-	-
气候区 气候区	10 个区域	平原海洋性：31% 变异海洋性：18%	平原海洋性：36% 地中海性：14%	oc：海洋性 medit：地中海性
自然灾害数量 已确认的自然灾害数量	[0, 40]	5.6 (4.95)	6.65 (4.47)	用于定义历史上 易受影响的区域
降雨				
尾部权重聚类 按尾部形状划分的区域		53.08 (27.79)	69.09 (27.58)	基于 ERA5-Land 数据构建
MILRE 最强局地降雨事件	[14.24, 100]	95.53 (9.65)	-	成本数据版本
ann_MILRE 最强局地降雨事件	[0, 100]	68.66 (25.37)	82.64 (20.22)	年度发生版本
建筑及周边环境				
居住 g_su表面 m住宅表面	[2.41,2443.52]	112.54 (56.77)	118.13（59.26）	使用机器学习补全的数据
房屋 -价值 房屋价值 2	[200, 28544]	2306.66（1237.34）	2593.38 (1439.02)	使用机器学习补全的数据
建造时期 建造时期	[0, 2022]	[0, 1915]: 26.28% [1949, 1968]：11.61%	[0, 1915]：26.26% [1949, 1968]：11.88%	基于文本处理构建
楼层数 楼层数	[0, 4]	0.94 (0.69)	0.90 (0.63)	使用机器学习补全的数据
附属建筑面积 最终附属建筑面积	[0, 1518.03]	9.87 (29.14)	10.06 (23.69)	利用计算机视觉构建
墙体材料 主要墙体材料	集料、砖，混凝土等。	集料：19.53% 砖：18.84%	砖：19.16% 集料：16.87%	利用计算机视觉构建
50米范围地形最大坡度 建筑物周围50米缓冲区	[0.09,307.68]	7.76 (7.64)	8.92 (9.06)	最大坡度 建筑物与周边环境之间
50米范围内建筑数量 50米范围内的建筑物数量	[0, 164]	19.73 (15.24)	16.83 (13.38)	利用计算机视觉和地理信息系统构建
相邻建筑存在情况 相邻建筑的存在	存在，不存在	不存在：60.54% 存在：39.46	不存在：67.43% 存在：32.57	-
距水道距离 距最近水道距离	[0.021909.2]	464.83 (632.67)	338.88 (453.32)	米，欧几里得距离和 所有纳入考虑的水道
水道海拔差 海拔差	[−2196.2,525.8]	−12.92 (36.29)	−9.78 (29.31)	考虑河床 用于最近点海拔
土壤类型 主要土壤类型	砂、黏土、泥灰岩，石灰岩等。	泥灰岩：25.23% 砂：24.92%	泥灰岩：31.53% 砂：26.96%	取值的唯一值或组合值
im透水 _su表面 建筑物周围的百分比	[0, 1]	0.72 (0.29)	0.66 (0.31)	建筑物周围的200米网格内
共用墙长度 墙体总长度	[0, 661.51]	5.14 (8.5)	3.98 (7.42)	使用所有相邻建筑计算