融合卫星与雨量计测量的降水数据：采用 LightGBM，重点关注极端分位数

Hristos Tyralis[1], Georgia Papacharalampous[2], Nikolaos Doulamis[3], Anastasios Doulamis[41] 雅典国立技术大学农村测绘与地理信息工程学院地形学系，Iroon Polytechniou 5，157 80 佐格拉富，希腊（montchrister@gmail.com，hristos@itia.ntua.gr，https://orcid.org/0000-0002-8932-4997）

[2] 雅典国立技术大学农村测绘与地理信息工程学院地形学系，Iroon Polytechniou 5，157 80 佐格拉富，希腊（papacharalampous.georgia@gmail.com，https://orcid.org/0000-0001-5446-954X）[3] 雅典国立技术大学农村测绘与地理信息工程学院地形学系，Iroon Polytechniou 5，157 80 佐格拉富，希腊（ndoulam@cs.ntua.gr，https://orcid.org/0000-0002-4064-8990）

[4] 地形学系，农村测绘与地理信息工程学院，雅典国立技术大学，Iroon Polytechniou 5, 157 80 佐格拉富，希腊（adoulam@cs.ntua.gr，https://orcid.org/0000-0002-0612-5889）

摘要：在水文建模应用中，了解降水在时空上的真实情况至关重要，但受经济条件限制，雨量站的空间覆盖范围有限。网格化卫星降水数据集覆盖范围广且均匀，为估算实际降水提供了另一种选择，尽管相关估算并不准确。为提高降水估算精度，人们应用机器学习，将雨量计测量值与网格化卫星降水产品融合。在此背景下，观测降水作为因变量，卫星数据则作为预测变量。随机森林是相关应用中占主导地位的机器学习算法。在这类空间预测场景中，会给出因变量的点预测（通常是条件分布的均值或中位数）。本文旨在解决空间插值场景中降水概率预测的问题，重点关注极端分位数。为此，我们提出使用轻量级梯度提升机（LightGBM）对降水进行概率空间预测。LightGBM 是一种提升算法，在预测和预报竞赛中屡获殊荣的参赛方案中表现突出

。为评估 LightGBM，我们开展了一项大规模应用，将美国本土连续地区的日降水量测量值与 PERSIANN 和 GPM-IMERG 卫星降水数据融合。我们关注因变量概率分布的极端分位数；在极端分位数方面，LightGBM 的分位数评分优于分位数回归森林（QRF，一种随机森林变体）。我们的研究有助于理解在空间场景中使用机器学习进行概率预测。

关键词：IMERG；轻量级梯度提升机；机器学习；PERSIANN；分位数回归；分位数回归森林；随机森林；遥感；卫星降水校正；空间插值

这是已被接受并发表于*IEEE*应用地球观测与遥感精选主题期刊的一篇文章的接受

1. 引言

经济条件限制了设有雨量站区域的覆盖范围及其空间覆盖密度。因此，在水文应用中，网格化卫星数据集被用作观测降水的替代数据。然而，网格化卫星数据集对实际降水的估计并不准确，因此需要将网格数据集与雨量计降雨测量值融合进行后处理；相关综述见 Abdollahipour 等（2022）以及 Hu 等（2019）。

将网格化卫星数据集与雨量计测量相结合的一种常见方法，是在回归场景中应用机器学习算法。在此背景下，卫星降水数据充当预测变量，而观测降水则充当因变量。这类场景的主要假设是，雨量计测量能够代表实际降水，尽管一些研究对这些测量的准确度提出了质疑（Sun 等 2018）。这类回归场景中最先进的算法是 Breiman 的（2001年）随机森林（Baez-Villanueva 等 2020、Chen 等 2021年、Fernandez-Palomino 等 2022、He 等 2016、Hengl 等 2018、Lin 等 2022、Militino 等 2023、Nguyen 等 2021年、Tyralis 等 2019b、Zhang 等 2021年）。

其他机器学习算法也已得到应用（例如 Meyer 等 2016、Lei 等 2022、Papacharalampous 等人 2023a、2023b、Shen 和 Yong 2021年），但应用频率较低。为了更好地了解这些算法的表现，重要的是使用覆盖大范围区域且雨量站网络密集的大型数据集，对多种算法进行比较（Lei 等 2022、Papacharalampous 等人 2023a、2023b）。

大多数融合卫星数据与站点观测的研究具有一个共同特征，即发布空间点预测，并使用平方误差评分函数、绝对误差评分函数或相关技巧评分（例如纳什–萨特克利夫效率（NSE）和克林–古普塔效率（KGE））对其进行评估。对于概率分布的均值泛函而言，平方误差评分函数是一致的（关于一致性的定义，读者可参阅第 3.4 节）（Gneiting 和 Raftery 2011），也就是说，通过使用平方误差评分函数训练机器算法，可以发布回归算法响应变量条件概率的均值预测。对于绝对误差评分函数（其对于中位数泛函是一致的）以及相关技巧评分，也适用类似的论证（Gneiting 和 Raftery 2011）。然而，当预测以概率分布的形式呈现时，其信息量更大（Gneiting 和 Raftery 2007）；Papacharalampous 等人（2019）以及 Papacharalampous 和 Tyralis（2022a）曾评论过 Hydrology 中对概率预测的需求，并指出在总体不确定性估计的背景下，这已成为 Hydrology 中的一个重要问题（Blöschl 等人 2019）。

在密集的分位数水平网格上预测条件概率分布的分位数，可近似完整的概率分布（Tyralis 和 Papacharalampous 2021b，2022a）。我们关注条件概率分布的极端分位数（例如，参见 Curceac 等人 2020，Tyralis 和 Papacharalampous 2023），但也评估了响应变量条件概率分布中心处的泛函预测值。鉴于极端分位数在开展洪水研究方面的重要性，它们备受关注。我们的应用包括将 PERSIANN 和 GPM-IMERG 日降水数据集与覆盖美国本土连续地区（美国本土）的基于雨量计的日降水测量数据进行融合。

本文旨在解决空间插值情境下的降水概率预测问题，并重点关注极端分位数。为此，我们拟采用轻量级梯度提升机（LightGBM）算法

（Ke 2017）通过分位数评分函数进行训练（Koenker 和 Bassett Jr 1978）。分位数评分函数及其相关平均得分已应用于水文建模和预测研究，例如 Dogulu 等人 2015、Papacharalampous 等人 2019、Tyralis 和 Papacharalampous 2021b。我们将 LightGBM 与分位数回归森林（QRF，Meinshausen 2006）进行比较，后者是随机森林的一种变体。对 LightGBM 的评估基于其相对于分位数回归森林的表现；鉴于随机森林在空间插值情境下因预测性能高且使用方便而备受青睐（Hengl 等 2018），这一比较尤为重要。QRF 在融合卫星与雨量计降水测量方面的既有应用包括 Bhuiyan 等人（2018）、Zhang 等（2022）；而其他能够在空间插值情境下生成概率预测的机器学习算法应用则较少（例如 Glawion 等人的深度学习应用，2023）。如前所述，我们重点关注极端分位数，并以包含分位数评分函数作为组成部分的技巧评分为比较依据。尽管评估侧重于极端分位数，我们也会进一步讨论中心分位数水平上的分位数；这些分位数的结果在很大程度上受降水间歇性特征的影响。

本文其余部分的结构安排如下：第2节介绍 LightGBM 和 QRF，并重点讨论与分位数回归相关的主题。第3节介绍研究所用的数据、问题表述以及用于评估算法的指标。结果见第4节，随后在第 5 节中讨论这些结果。本文最后以第 6 节作结。

2. 方法

我们将轻量级梯度提升机（LightGBM）应用于一个包含网格化卫星降水产品（见 第3.1.2节）和雨量计测量（见 第3.1.1节）的数据集。由于我们的目标是提供降水的概率预测（尤其是高分位数），因此使用分位数评分函数训练 LightGBM（见 第2.3节）。我们使用留出样本对这些算法与 QRF 进行了比较（见 第3.2节）。本节介绍 LightGBM 和 QRF，软件实现则见 附录A。由于 LightGBM 和 QRF 分别是提升算法和随机森林的变体，我们重点介绍使它们有别于相应基础算法的某些特性。更详细的介绍

关于提升法和随机森林的内容可参见教科书（Hastie 等，2009；James 等，2013；Efron 和 Hastie，2016），而遥感领域的科学家和技术人员对此也很熟悉。

2.1 分位数回归森林

分位数回归森林（Meinshausen 2006）是随机森林（Breiman2001）的一种变体，用于生成概率预测。随机森林是一种用于空间插值的先进算法（Hengl 等 2018），并已广泛应用于水文领域（Tyralis 等 2019b）。

用于回归的随机森林算法会构建一个决策树集成，而算法的预测值等于各棵树预测值的均值。通过将装袋法与随机节点优化相结合来构建树森林。装袋法（自助聚合）是指对训练集进行有放回重采样，并使用该样本训练单棵树的过程。此外，随机森林会在每个候选分裂处随机选取一组特征。

分位数回归森林对响应变量的条件分布进行近似，而不是对各棵树的预测值求平均（后者近似于响应变量的条件均值）。随机森林的特性同样适用于分位数回归森林。这些特性总结见 Tyralis 等（2019b），其中包括与当前问题相关的高预测性能、速度、大规模应用的可行性、抗过拟合能力、高效处理高度相关变量的能力以及稳定性。

这里我们采用了 Wright 提出的分位数回归森林的 `R` 语言实现（Wright 和 Ziegler，2022；Wright 和 Ziegler，2017），使用 100 棵决策树和默认超参数值，因为默认实现效率很高（Tyralis 等，2019b）。随机森林的这一软件实现计算速度较快；然而，对于大型数据集，其速度远慢于 LightGBM，因此对本研究的大样本进行超参数优化变得难以实现。由于分位数回归森林是一种回归算法，其训练和预测方式与常规方法相同，即训练样本包含一组观测到的预测变量以及一组观测到的因变量。具体应用将在后文的 第3.2节中介绍。

2.2 轻量级梯度提升机（LightGBM）

梯度提升决策树是一种集成学习算法，其中决策树依次加入集成中。在每次迭代中，都会针对算法迄今为止产生的误差训练一棵新的决策树。基于梯度下降的公式化方法使提升法的概念得以形式化（Friedman，2001；Natekin 和 Knoll，2013；Mayr 等，2014）。梯度提升决策树可以通过不同的评分函数进行优化，因此可以生成符合用户需求的预测。Tyralis 和 Papacharalampous（2021a）总结了提升算法的若干特性。尽管提升算法与随机森林具有一些相似特性，但在多种场景下，其表现往往更好，不过仍需进行超参数调优。

LightGBM（Ke 等，2017）是一种提升算法，与常见的梯度提升算法相比具有一些优势。具体而言，它特别适用于特征维度高且规模较大的数据集。它采用基于梯度的单边采样（GOSS），排除梯度较小的数据实例（相比之下，常见的提升算法会扫描所有数据实例，以估计所有可能分裂点的信息增益），从而减少训练时间。此外，它还采用互斥特征捆绑（EFB）将相互排斥的特征进行捆绑，以减少特征数量。另外，它使用基于直方图的算法来寻找最佳分裂点，这与早期取得成功的提升算法变体（例如 XGBoost；Chen 和 Guestrin，2016）类似。

LightGBM 包含多个参数，通过调优这些参数可以提升其预测性能。优化过程详见第3.3节。此外，LightGBM 使用分位数评分函数进行训练（见第2.3节和第3.4节），以生成概率预测。由于 LightGBM 是一种回归算法，因此它按照通常的方式进行训练和预测，即训练样本包括一组观测到的预测变量和一组观测到的因变量。具体应用将在后文第3.2节中介绍。

2.3 分位数回归

分位数回归算法用于预测回归场景中的条件分位数。下面我们将解释分位数回归在实践中的工作原理。下文中，观测值将用小写字母表示，而随机变量将用带下划线的小写字母表示。设 y 为一个随机变量，其累积分布函数为 $F_{\underline{y}}$ 定义如下：

$$F_{\underline{y}}(y) := P(\underline{y} \le y) \tag{1}$$

那么，y 的 τ^{th} 分位数 $Q_{\underline{y}}(\tau)$ 定义如下：

$$Q_{\underline{y}}(\tau) := \inf\{y: F_{\underline{y}}(y) \ge \tau\}, \tau \in (0, 1) \tag{2}$$

其中，$\inf\{\cdot\}$ 表示实数集的下确界。

令 $F_{\underline{y}|\boldsymbol{x}}$ 表示随机变量 y 在给定 p 维向量 $\boldsymbol{x}$ 条件下的分布：

$$F_{\underline{y}|\boldsymbol{x}}(y|\boldsymbol{x}) := P(\underline{y} \le y|\boldsymbol{x}) \tag{3}$$

那么，给定 $\boldsymbol{x}$ 时 y 的 τ^{th} 分位数 $Q_{\underline{y}|\boldsymbol{x}}(\tau|\boldsymbol{x})$ 定义为：

$$Q_{\underline{y}|\boldsymbol{x}}(\tau|\boldsymbol{x}) := \inf\{y: F_{\underline{y}}(y|\boldsymbol{x}) \ge \tau\}, \tau \in (0, 1) \tag{4}$$

分位数损失函数 $\rho_\tau(u)$ 定义如下，见 式 (5)：

$$\rho_\tau(u) := u\,(\mathbb{I}(u \ge 0) - \tau), u \in \mathbb{R} \tag{5}$$

这里 τ 是所关注的分位数水平，且 $\mathbb{I}(A)$ 表示示性函数：当事件 A发生时其值为 1，否则为 0。由式 (5) 定义的分位数损失函数为正且方向为负，即目标是将其最小化；当 $u = 0$ 时，其值为 0。

令 $\boldsymbol{\theta}$ 为回归模型的参数（例如第 2.2 节中的 LightGBM）。令 $\mathrm{y}(\mathrm{x}, \boldsymbol{\theta}(\tau))$ 表示当预测变量取值为 x、模型参数取值为 $\boldsymbol{\theta}(\tau)$ 时，回归模型在分位数水平 τ 下的预测。为了估计 $\boldsymbol{\theta}(\tau)$，其中 $\tau \in \in(0, 1)$，应最小化平均分位数得分 $(1/\mathrm{n})\,\Sigma_{i=1}^{n}\,\rho_\tau(\mathrm{y}(\mathrm{x}_i, \boldsymbol{\theta}(\tau)) - \mathrm{y}_i)$；这正是 Koenker 和 Bassett Jr（1978年）阐述的线性参数分位数回归的核心思想。参数为 $\boldsymbol{\theta}(\tau)$ 的回归模型可预测条件分位数 $\mathrm{Q}_{\underline{y}|\boldsymbol{x}}(\tau|\mathrm{x})$。

3. 数据与应用

3.1 数据

我们使用全球历史气候网络日值数据集（GHCNd）中提取的地球观测日降水来评估这些算法，具体如第3.1.1节所述；此外，还使用了当前业务化 PERSIANN（基于人工神经网络的遥感信息降水估算）系统提供的网格化卫星降水，以及 GPM IMERG（综合多卫星反演）晚期降水数据集（如第3.1.2节所述），并使用从亚马逊云科技（AWS）地形瓦片应用中提取的高程数据（如第3.1.3节所述）。雨量计的位置为

见图1。

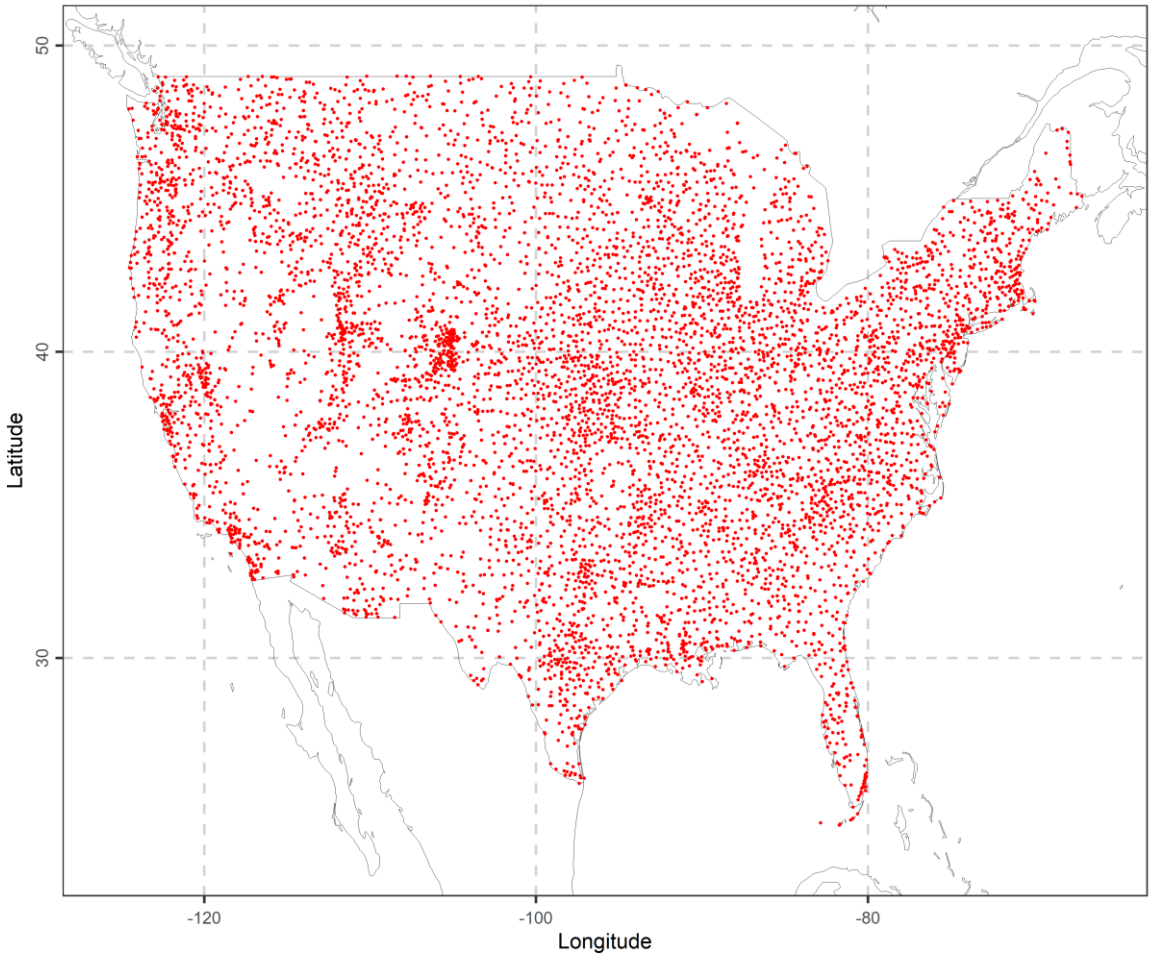


图1。提供本研究所用数据的地面站点地理位置（红点）地图。此处，纬度以赤道以北的距离计量，而经度以英格兰格林尼治子午线以西的距离（单位：度（°））计量。

3.1.1 地球观测降水数据

我们使用从全球历史气候网络日值数据集（GHCNd）中提取的每日总降水数据（Durre 等，2008、2010；Menne 等，2012；美国国家海洋和大气管理局国家气候数据中心（https://www1.ncdc.noaa.gov/pub/data/ghcn/daily）；访问日期：2022-02-27）。具体而言，我们提取了美国本土（见图1）范围内 7,261 个站点的数据。数据涵盖两年时间段 2014–2015。

3.1.2 卫星降水数据

我们使用了当前业务化 PERSIANN 系统的网格化卫星日降水数据（Hsu 等，1997，Nguyen 等2018，2019）。该系统由加利福尼亚大学欧文分校（UCI）的水文气象与遥感中心（CHRS）开发。PERSIANN 数据取自水文气象与遥感中心网站

水文气象与遥感中心（CHRS）（https://chrsdata.eng.uci.edu；访问日期：2022-03-07）。

此外，我们使用了 NASA 戈达德地球科学（GES）数据与信息服务中心（DISC）开发的 GPM IMERG 晚期降水 L3 1 天、0.1 度 × 0.1 度 V06 数据集（Huffman 等，2019）。GPM IMERG 数据通过双线性插值映射到 CMORPH0.25 网格上的 0.25 度 × 0.25 度网格。虽然还存在其他插值方法，但本文的结论不受插值类型影响，而取决于所采用的算法类型（详见第 5 节）。IMERG 数据取自美国国家航空航天局（NASA）地球数据（https://doi.org/10.5067/GPM/IMERGDL/DAY/06网站；访问日期：2022-12-10）。

提取的数据覆盖美国本土，时间跨度为两年2014–2015。本文使用的 GPM IMERG 版本卫星降水产品未使用地面降水数据进行偏差订正。PERSIANN 数据集经过地面数据订正；因此，对这些数据应用回归算法，实际上是一种进一步改进卫星数据集的后处理框架。这是该领域的常见做法，例如可参见 Baez-Villanueva 等（2020）和 Bhuiyan 等（2018）等。

可以使用不同或更多的卫星降水数据集。后一种情况下，可用信息和预测变量的数量将增加，从而提高订正后的降水产品的准确度。不过，虽然构建高准确度的新数据集时，采用多个卫星降水数据集是常见做法，但这超出了本研究的范围。本研究旨在了解发布概率预测时 LightGBM 的特性。此外，结论（见 第 6 节）不会受到影响，因为结果源于两种已实现算法的理论特性，后文 第 5 节将对此进行讨论。尽管如此，同时纳入两个数据集（其中一个已经订正，另一个则未经订正）还有一项额外优势。具体而言，正如 Papacharalampous 等人（2023b）已经指出的，未订正数据集包含的信息比订正后的数据集更多（事实上，后者对最终产品准确度的提升较不显著）。因此，产品的多样性或有助于更好地了解订正算法的特性。

3.1.3 高程数据

在合并基于雨量站观测的降水数据与卫星降水数据时，高程是一个有用的预测变量（Xiong 等人 2022）。因此，我们使用亚马逊云科技（AWS）的地形瓦片（https://registry.opendata.aws/terrain-tiles；访问日期：2022-09-25）应用程序，计算了 第3.1.1节中各站点的高程。

3.2 问题表述与算法评估

问题设置的表述方式与 Papacharalampous 等人（2023a，b）提出的程序类似。站点日总降水量是回归问题中的因变量。预测变量是距离站点最近的网格点的日总降水量。具体而言，PERSIANN 数据集对应四个预测变量，GMP IMERG 数据集也对应另外四个预测变量。此外，我们计算了站点与每个卫星数据集最近网格点之间的距离，因此又得到了八个预测变量。站点海拔及其经度和纬度也作为预测变量。正如前文在第 2.1 节 和 2.2 节中分别所述，预测变量之间可能存在的交互作用不会影响随机森林（及其变体）和提升法（及其变体）的性能。

为了解我们如何将站点降水量与网格化卫星降水关联起来，我们设计了图 2。对于单个网格（例如 PERSIANN），我们确定了与每个降水站最近的四个网格点，计算了这些网格点的距离 d_i、i =1、2、3、4，并按递增顺序排列为 $d_1 < d_2 < d_3 < d_4$（见图 2）。当我们提及 PERSIANN 数据集时，距离 d_i、i = 1、2、3、4 被称为“PERSIANN 距离 1–4”；而当我们提及 IMERG 数据集时，这些距离被称为“IMERG 距离 1–4”。网格点1–4 处相应的日降水值称为“PERSIANN 值 1–4”或“IMERG 值 1–4”。

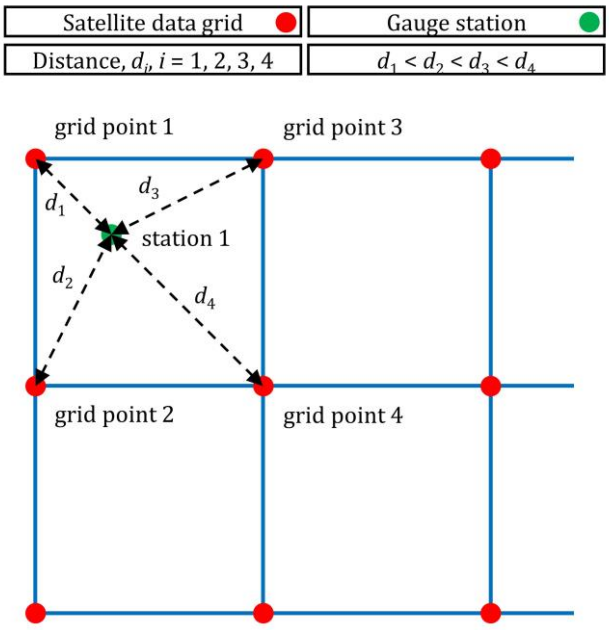


图 2。回归问题的设置。请注意，术语“网格点”用于描述具有卫星数据的地理位置，而术语“站点”用于描述具有地面测量的地理位置。还请注意，在本文中，距离d_i、i = 1、2、3、4也分别称为“PERSIANN 距离 1–4”或“IMERG 距离 1–4”（取决于我们指的是 PERSIANN 网格还是 IMERG 网格），网格点处的日降水值1–4 称为“PERSIANN 值 1–4”或“IMERG 值1–4”（取决于我们指的是 PERSIANN 网格还是 IMERG 网格）。

表 1 给出了该问题回归设置中的预测变量。具体而言，预测变量包括PERSIANN 值1–4、IMERG 值1–4、PERSIANN 距离 1–4、IMERG 距离 1–4 以及站点的经度、纬度和高程。最终数据集包含 4 833 007 个样本。我们将数据集随机划分为三个大小相等的折。随机森林在前两个折的并集中进行训练，并在第三个折中进行测试（其中包含 1 611 002 个样本）。LightGBM 在第一个折中训练，并在第二个折中验证。我们实施该程序以估计其超参数（见第3节第3.3节）。在估计 LightGBM 的超参数后，我们使用前两个折的并集和最终参数重新训练该算法。随后在第三个折中测试该算法。LightGBM 低于 0 的预测被转换为 0（众所周知，降水是正量），而 QRF 未产生负的降水预测。

Tab表 1本研究考察的预测变量集中的预测变量 k.

预测变量	预测变量集
PERSIANN 值 1	✓
PERSIANN 值 2	✓
PERSIANN 值 3	✓
PERSIANN 值 4	✓
IMERG 值 1	✓
IMERG 值 2	✓
IMERG 值 3	✓
IMERG 值 4	✓
PERSIANN 距离 1	✓
PERSIANN 距离 2	✓
PERSIANN 距离 3	✓
PERSIANN 距离 4	✓
IMERG 距离 1	✓
IMERG 距离 2	✓
IMERG 距离 3	✓
IMERG 距离 4	✓
经度	✓
纬度	✓
站点海拔	✓

3.3 LightGBM 超参数优化

LightGBM 有多个可调参数。我们选择优化其中一些参数，其余参数则保留R 软件实现中的默认值（Shi 等，2022）。待调优超参数的选择依据是算法文档以及实际应用经验，因为该算法曾用于国际预测竞赛中的获奖方案。选定用于调优的参数及其说明依据软件文档

(https://lightgbm.readthedocs.io/en/v3.3.2/Parameters.html) 见下文表2。

表2。LightGBM 参数。

参数	描述	取值
最大深度_	树模型的最大深度。最大值 _可用于显式限制树深度。	6、8、10
叶节点中的最小_数据_量_	这是防止按叶生长的树出现过拟合的一个非常重要的参数。其最优值取决于训练样本的数量以及 num_叶子节点。将其设为较大的值可以避免树生长得过深，但可能导致欠拟合。实践中，将其设为数百或数千对于大型数据集来说已经足够	20、100、200、500、1 000
学习率_	收缩率。一般来说，如果减少 num_迭代次数 ，那么就应增加学习 _rate	0.02、0.05、0.1
num_迭代次数	迭代次数。num _迭代次数 该参数控制提升轮数将执行的轮数。由于 LightGBM 使用决策树作为学习器，因此也可以将其视为“树的数量”	400
num_叶子数	单棵树中的最大叶子数。这是控制树模型复杂度的主要参数	20, 40, 60, 80, 100, 200、500

参数空间包含一个网格，其中列出参数所有可能取值的组合，但不包括 num_leaves > 2^{max_depth} 的参数组合。此外，我们采用了带有提前停止功能的算法，并将参数 early_stopping_round 设为 20。在这种情况下，如果连续 20 次迭代得分都没有提升，算法就会在达到指定迭代次数之前停止。提前停止有助于缩短训练时间。

3.4 性能指标与评估

我们使用以下定义的分位数评分函数对分位数回归森林和 LightGBM 进行了比较：

$$S_\tau(x, y) := \rho_\tau(x - y) \tag{6}$$

这里，y是空间过程的实现值（观测值），而x是在τ水平下的预测分位数。水文预测应当具有概率性，其形式应为概率分布（例如参见 Gneiting 和 Raftery 2007、Papacharalampous 和 Tyralis 2022a）。在多个水平上预测概率分布的分位数，是对完整概率分布的一种良好替代。对于预测分布的分位数泛函而言，分位数评分函数具有严格一致性，即如果接收到预测某个分位数的指令，那么遵循该指令时，期望分位数评分将达到最小（Gneiting 2011）。因此，当接收到

预测分位数泛函的指令时，使用分位数评分函数训练模型是很自然的做法，正如第2.3节中所述。

分位数水平为τ时，机器学习算法的性能准则形式如下：

$$\bar{S}_\tau := (1/n)\sum_{i=1}^{n} S_\tau(x_i, y_i) \tag{7}$$

其中，$\{x_i、y_i\}$、$i = 1、\ldots、n$分别表示第i^{th}个样本的预测值和观测值，n表示测试折的样本数。我们在多个分位数水平τ下，对两种算法均计算了$\bar{S}_\tau$。我们将分位数回归森林给出的预测视为参考预测，并计算了参考算法在指定分位数水平下的技巧评分，定义如下：

$$S_{\tau,\text{skill}} := 1 - \bar{S}_{\tau,\text{LightGBM}}/\bar{S}_{\tau,\text{QRF}} \tag{8}$$

一般而言，$S_{\tau,\text{skill}} \le 1$；而对于分位数水平为$\tau$的优异预报，我们有$\bar{S}_{\tau,\text{LightGBM}} = 0$且$S_{\tau,\text{skill}} = 1$。当$S_{\tau,\text{skill}} > 0$时，LightGBM优于分位数回归森林；而$S_{\tau,\text{skill}}$越高，LightGBM的表现越好。我们没有使用其他评分函数（例如平方误差评分函数，或相关的技巧评分，如Nash-Sutcliffe效率）比较这些算法，因为此类函数不适用于分位数泛函（Gneiting 2011）。

此外，我们为每种算法计算了一个评分，用于衡量各算法给出的预测是否符合名义频率。具体而言，相应评分定义为

$$\overline{\text{FR}}_\tau := \left|(1/n)\sum_{i=1}^{n} \mathbb{I}(y_i \le x_i) - \tau\right| \tag{9}$$

为了更好地理解该评分，令$\tau = 0.95$。假设给出的预测完全准确，则$(1/n)\sum_{i=1}^{n}\mathbb{I}(y_i \le x_i)$应等于0.95（即95%的观测值应小于或等于相应的预测），且$\overline{\text{FR}}_\tau$应等于0。同样，我们计算了相应的技巧评分，并以分位数回归森林作为参考算法：

$$\text{FR}_{\tau,\text{skill}} := 1 - \overline{\text{FR}}_{\tau,\text{LightGBM}}/\overline{\text{FR}}_{\tau,\text{QRF}} \tag{10}$$

总体而言，$\text{FR}_{\tau,\text{skill}} \le 1$；而对于分位数水平为$\tau$的优秀预报，我们有$\overline{\text{FR}}_{\tau,\text{LightGBM}} = 0$且$\text{FR}_{\tau,\text{skill}} = 1$。当$\text{FR}_{\tau,\text{skill}} > 0$时，LightGBM优于分位数回归森林；而$\text{FR}_{\tau,\text{skill}}$越高，LightGBM的表现越好。

4. 结果

此处展示算法的应用结果，这些结果将在下一节中详细讨论，并随后给出相应解释。关于算法在测试集中的表现，我们在以下分位数水平进行了测试：$\tau \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99, 0.999\}$。请注意，技巧评分高于 0 表示 LightGBM 优于 QRF。

关于频率技巧评分的情况（请回顾公式（τ, skill 10）之后对频率技巧评分 FR 数值的解释），见图3a。对于 $\tau \leq 0.8$，两种算法的表现几乎相同；而对于 $\tau \in (0.8, 0.95)$，评分在 0 附近波动，且在较高分位数水平下 LightGBM 优于 QRF。回顾第 3.4 节，与频率相关的评分函数对于感兴趣的泛函并不一致，因此可以使用分位数评分函数对算法进行严格评估。分位数评分函数的技巧评分值（请回顾公式（δ_{τ}, skill 之后对分位数技巧评分值的解释）见图3b。从图中可见，当 $\tau \geq 0.97$ 时，LightGBM 的表现似乎优于 QRF；而在较低分位数水平下，两种算法的表现大致相同。在两种技巧评分的情况下，当 τ 增加时，评分也会随之增加，$\tau \geq 0.97$。

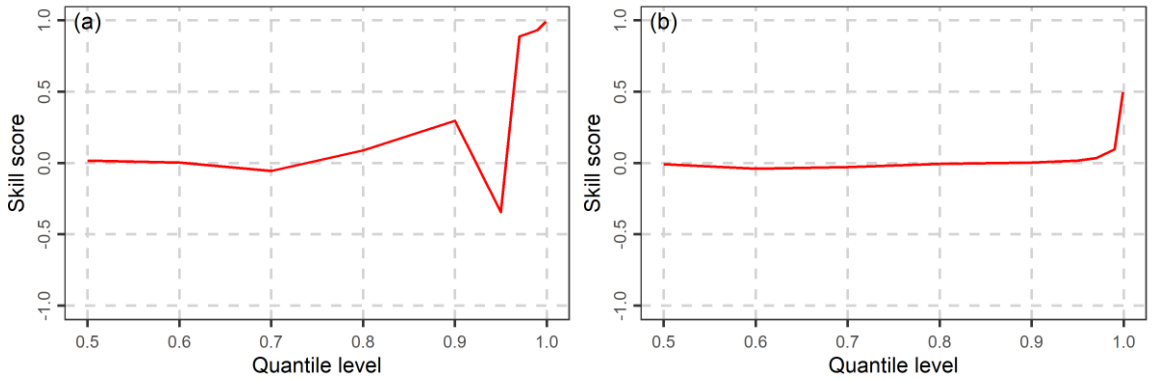


图3。不同分位数水平下（a）频率和（b）分位数损失的技巧评分τ（测试集中的完整数据）。

了解当测试集中的观测值等于 0 时各算法的表现很有意义。零降水约占日观测值总数的 72%，而时间和空间上的间歇性是降水的主要特征。图4展示了频率和分位数评分函数的相关技巧评分。图4。在频率方面，两种算法的表现相同（图4a）。这是意料之中的，因为算法始终会给出大于或等于 0 的预测（对于

LightGBM，预测经过调整后如此；见第3.2节），另请参见下一节的讨论。然而，就分位数评分函数而言，QRF 在 $\tau \geq 0.97$ 时似乎优于 LightGBM；见图4b。

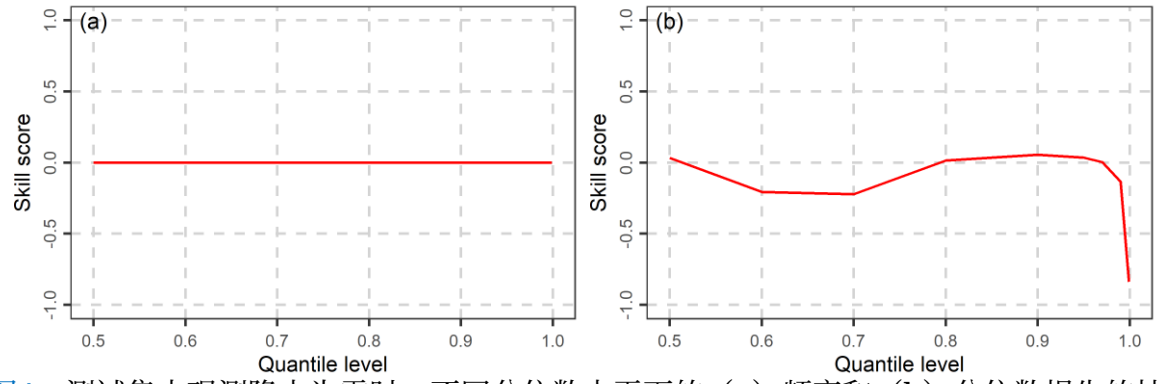


图4。测试集中观测降水为零时，不同分位数水平下的（a）频率和（b）分位数损失的技巧评分。

当在观测降水大于 0 的情况下测试算法时，整体情况有所变化；见 图5。在这里，对于 $\tau \geq 0.95$，LightGBM 在频率和基于分位数评分函数的技巧评分方面似乎都优于 QRF。对于这两种技巧评分，随着 τ 增加，当 $\tau \geq 0.97$ 时，评分均有所提高。

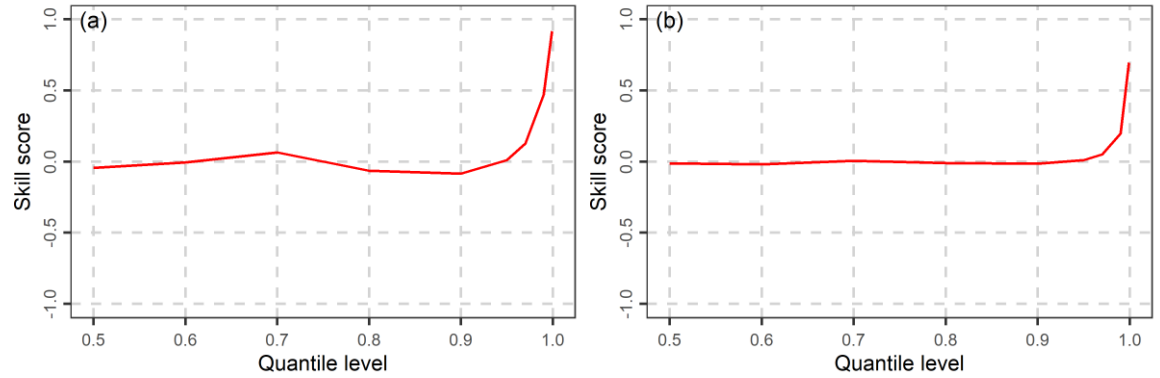


图5。测试集中观测降水大于零时，不同分位数水平下的（a）频率和（b）分位数损失的技巧评分。

了解各算法在每个站点上的单独表现也很有意义；参见图6。这里，我们考察基于分位数评分函数的技巧评分；请注意，分位数评分函数与分位数泛函相一致。技巧评分低于–1 的站点已从图6中移除。原因是，有些技巧评分值低至–10 或更低，这会在结果的呈现中造成一些伪影。移除这些站点不会影响结论，因为技巧评分呈偏态分布，这是由于其值不可能超过 1，

尽管其值可以等于−∞。此外，我们移除了两个算法的平均得分均为 0 的站点（在这种情况下，技巧评分未定义）。在图6中，我们观察到，随着分位数水平$\tau \to 1$的变化，技巧评分会提高。此外，我们观察到，在相同分位数水平下，不同站点的技巧评分有所差异，不过这种差异相对较小。当分位数水平为$\tau \geq 0.97$ 时，技巧评分明显偏离 0。

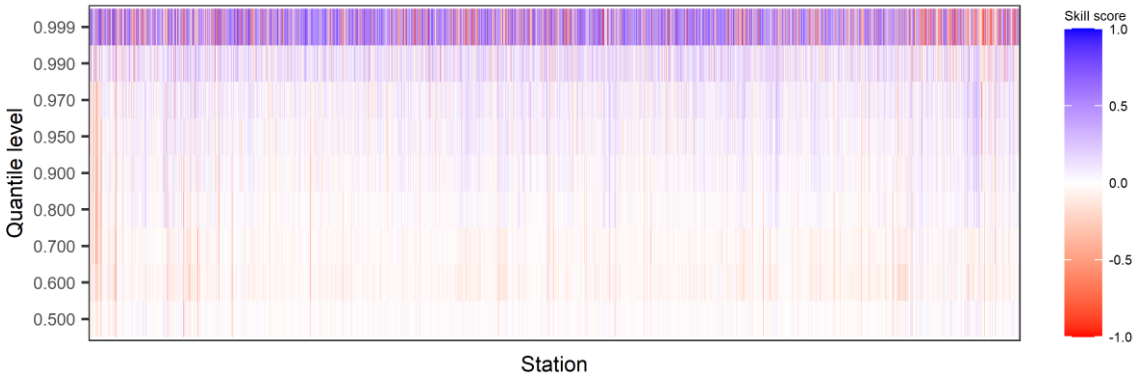


图6。不同分位数水平及各站点分位数损失的技巧评分热图。

5. 讨论

关于两种方法的总体表现，LightGBM 是一种特别适用于维度数量较多的大型数据集的算法。此外，鉴于它属于高度参数化的提升算法家族，具有很高的灵活性，因此其平均表现优于随机森林并不令人意外。在完整数据的情况下，这一点尤为明显：使用分位数评分函数评估时，LightGBM 总体上表现更好。然而，LightGBM 并非在所有分位数水平上都优于 QRF。在较低的分位数水平下，这两种算法的表现似乎相近；而在较高的分位数水平下，LightGBM 明显优于 QRF。

低分位数水平下出现这种行为的一种可能解释，是数据集中零值占比较高。具体而言，QRF 是一种基于自助抽样的算法，因此可以对零值进行重采样。另一方面，LightGBM 基于最小化分位数评分函数；当数据集具有高度间歇性时，这种做法可能不是最优的。例如，降水条件分布的中位数应为零（因为数据集中零值的数量超过 50%）。在水平为 $\tau = 0.5$ 时，分位数评分函数

等同于绝对误差评分函数；而当需要预测在零值处有质量点的概率分布的中位数值时，这种函数可能并不适用。尽管如此，LightGBM 和 QRF 在 0.5 分位数水平上的总体表现仍然相近。这可能是因为，虽然 QRF 对零值的预测更准确，但在给出非零预测时，其表现会明显较差。

在较高的分位数水平下，LightGBM 在所有技巧评分方面均明显优于 QRF，而且随着τ的增加，这种差异也随之扩大；当$\tau \to 1$时，技巧评分趋于 1。一种可能的解释是，QRF 无法预测超出训练集取值范围的值（Hengl 等 2018）。在较高的分位数水平下，这一弱点更为明显：训练集中高值变得更加稀少，算法也倾向于转向较低值。另一方面，LightGBM 通过将基学习器添加到先前的误差上来构建模型；尽管基学习器是决策树，它似乎仍能更好地外推训练集的取值范围。此外，较高分位数水平下的降水条件概率分布似乎符合正则条件，在这些条件下，分位数评分函数的性质似乎成立。

当观测降水为零时，QRF 在较高分位数水平下的分位数技巧评分优于 LightGBM；而当观测降水大于零时，情况则相反。在完整测试集上，LightGBM 的表现优于 QRF，因为观测降水为零时的分位数评分绝对值通常低于观测降水非零时的评分，因此平均评分的主要部分来自非零值。

考察单个算法对于理解其自身特性至关重要。然而，算法的组合（机器学习文献中称为集成学习，Sagi 和 Rokach 2018）可能带来更大的提升。实践中，简单组合（例如 Papacharalampous 和 Tyralis 2020）或堆叠（Tyralis 等 2021年）已被证明能够有效预测条件概率分布的均值泛函。用于概率预测的算法组合（例如堆叠，Wolpert 1992，Lichtendahl 等人 2013，Yao 等人 2018），或许还包括空间特征（例如 Papacharalampous 和 Tyralis 2022b），是值得研究的课题，并且已在其他水文应用中证明了其有效性（Papacharalampous 等人 2019，Tyralis 等 2019a），也已应用于合并基于雨量站观测的降水数据集与卫星降水数据集（关于均值泛函的情形，见 Zandi 等人 2022）。基于极值理论的极端分位数预测也值得研究，但仍需评估

在实践中，空间场景下的条件分布是否具有厚尾特性。尽管如此，极端分位数回归已应用于时间域中的后处理应用（Tyralis 和 Papacharalampous 2023）。考察其他算法，例如深度学习算法（LeCun 等人 2015，Schmidhuber 2015），在降水概率预测中的表现，也可能是值得研究的课题。

6. 结论

我们提出，在融合雨量计测量和卫星降水产品的空间场景中，使用 LightGBM 发布日降水的概率预测。在预测响应变量的条件概率分布的极端分位数时，LightGBM 优于此类场景下最先进的分位数回归森林（一种随机森林变体）；而在预测概率分布中心的分位数时，两种算法表现相近。随着所比较的分位数水平升高并趋近于 1，方法性能的差异逐渐扩大，且有利于 LightGBM。通过在包含美国本土连续地区观测降水的大型数据集上比较这些算法，增强了对结果的信心。

我们还对结果提供了直观解释：由于随机森林无法外推至训练集范围之外，再加上 LightGBM 凭借其结构提升了发布准确预测的能力，因此 LightGBM 能够更好地预测极端分位数。另一方面，在预测条件概率分布中心的分位数时，分位数回归森林与 LightGBM 表现相当，这是由于降水具有高度间歇性的特征，加之分位数回归森林采用了基于自助法的结构；与基于分位数评分函数的算法结构相比，这种结构在此情形下似乎更为适用。

利益冲突： 作者声明不存在利益冲突。

作者贡献： HT 和 GP 在 ND 和 AD 的意见基础上构思并设计了这项工作。HT 和 GP 完成了分析和可视化，并撰写了初稿；ND 和 AD 对初稿进行了评论，并补充了新的文字、解读和讨论内容。

资助： 本研究是在 BETTER RAIN 项目（利用机器学习算法和概念改进卫星降雨产品）框架下开展的。该研究项目由希腊基金会

研究与创新（H.F.R.I.）通过“3rd 面向博士后研究人员的 H.F.R.I. 研究项目资助征集”（项目编号：7368）提供支持。

附录A　　统计软件

我们使用 R 编程语言（R Core Team 2022）来实现算法，并报告和可视化结果。

在数据处理和可视化方面，我们使用了由社区贡献的 R 软件包：data.table（Dowle and Srinivasan 2022）、elevatr（Hollister 2022）、ncdf4（Pierce 2021）、rgdal（Bivand et al. 2022）、sf（Pebesma 2018、2022）、spdep（Bivand 2022、Bivand and Wong 2018、Bivand et al. 2013）、tidyverse（Wickham et al. 2019、Wickham 2022）。

这些算法通过使用由社区贡献的 R 软件包 ranger（Wright 2022、Wright and Ziegler 2017）、lightgbm（Shi et al. 2022）实现。

性能指标通过实现由社区贡献的 R 软件包 scoringfunctions（Tyralis and Papacharalampous 2022a、2022b）计算。

报告使用以下贡献软件包生成：R 软件包 devtools（Wickham 等，2022）、knitr（Xie，2014、2015、2022）、rmarkdown（Allaire 等，2022、Xie 等，2018、2020）。

参考文献

[1] Abdollahipour A, Ahmadi H, Aminnejad B (2022) 卫星降水估算降尺度方法综述。地球科学信息学 15(1)。https://doi.org/10.1007/s12145-021-00669-4。[2] Allaire JJ, Xie Y, McPherson J, Luraschi J, Ushey K, Atkins A, Wickham H, Cheng J, Chang W, I annone R (2022) rmarkdown：R 的动态文档。R 软件包版本 2.18。https://CRAN.R-project.org/package=rmarkdown。[3] Baez-Villanueva OM, Zambrano-Bigiarini M, Beck HE, McNamara I, Ribbe L, Nauditt A, Birkel C, Verbist K, Giraldo-Osorio JD, Xuan Thinh N (2020) RF-MEP：一种用于合并网格化降水产品和地面测量的新型随机森林方法。环境遥感 239:111606。https://doi.org/10.1016/j.rse.2019.111606。[4] Bhuiyan MAE, Nikolopoulos EI, Anagnostou EN, Quintana-Seguí P, Barella-Ortiz A (2018) 一种用于合并全球降水数据集的非参数统计技术：在伊比利亚半岛的开发与水文评估。Hydrology and Earth System Sciences 22：1371–1389。https://doi.org/10.5194/hess-22-1371-2018。[5] Bivand RS (2022) spdep：空间依赖：加权方案、统计量。R 软件包版本 1.2-7。https://CRAN.R-project.org/package=spdep。

[6] Bivand RS, Wong DWS (2018) 全球和局部空间关联指标实现方式的比较。TEST 27(3)：716–748。https://doi.org/10.1007/s11749-018-0599-x。

[7] Bivand RS、Pebesma E、Gómez-Rubio V（2013）《R空间数据分析实用指南》：第二版。施普林格，纽约，纽约州。https://doi.org/10.1007/978-1-4614-7618-4。

[8] Bivand RS、Keitt T、Rowlingson B（2022）`rgdal`：‘地理空间’数据抽象库的绑定。R 软件包版本 1.6-2。https://CRAN.R- project.org/package=rgdal。

[9] Blöschl G、Bierkens MFP、Chambel A、Cudennec C、Destouni G、Fiori A、KirchnerJW、McDonnell JJ、Savenije HHG、Sivapalan M 等（2019）水文学未解难题（UPH）中的二十三个问题–群体视角。《水文科学杂志》64(10)：1141–1158。https://doi.org/10.1080/02626667.2019.1620507。

[10] Breiman L（2001年）随机森林。《机器学习》45(1):5–32。https://doi.org/10.1023/A:1010933404324。

[11] Chen T, Guestrin C（2016）XGBoost：一种可扩展的树提升系统。第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集，785–794。https://doi.org/10.1145/2939672.2939785。

[12] Chen C, Hu B, Li Y（2021年）一种易于使用的基于空间随机森林的降尺度校准方法，用于生成高分辨率、高准确度的降水数据。《水文学与地球系统科学》25(11):5667–5682。https://doi.org/10.5194/hess-25-5667-2021。

[13] Curceac S, Atkinson PM, Milne A, Wu L, Harris P（2020）校正水文极端事件过程模型模拟中的条件偏差：一项采用North Wyke农场平台开展的实验。《人工智能前沿》3:82。 https://doi.org/10.3389/frai.2020.565859。

[14] Dogulu N, López López P, Solomatine DP, Weerts AH, Shrestha DL (2015) 使用分位数回归和UNEEC方法估算水文预测不确定性，并在不同集水区对这些方法进行比较。《水文与 地球系统科学》19(7)：3181–3201。https://doi.org/10.5194/hess-19- 3181-2015。

[15] Dowle M, Srinivasan A (2022) `data.table`：'data.frame' 的扩展。R 软件包版本 1.14.6。 https://CRAN.R-project.org/package=data.table。

[16] Durre I, Menne MJ, Vose RS (2008) 评估质量保证程序的策略。《应用气象学与气候学杂志》47(6)：1785–1791。 https://doi.org/10.1175/2007JAMC1706.1。

[17] Durre I, Menne MJ, Gleason BE, Houston TG, Vose RS (2010) 对每日地面观测进行全面的自动化质量保证。《应用气象学与气候学杂志》49(8)：1615–1633。 https://doi.org/10.1175/2010JAMC2375.1。

[18] Efron B, Hastie T（2016）《计算机时代的统计推断》。剑桥大学出版社，纽约。https://doi.org/10.1017/CBO9781316576533。

[19] Fernandez-Palomino CA, Hattermann FF, Krysanova V, Lobanova A, Vega-Jácome F, Lavado W, Santini W, Aybar C, Bronstert A（2022年）秘鲁和厄瓜多尔流域高分辨率网格化降水数据集：开发与水文评估。《水文气象学杂志》23(3):309–336。https://doi.org/10.1175/JHM-D-20-0285.1。

[20] Friedman JH（2001年）贪心函数逼近：一种梯度提升机。《统计年鉴》29(5):1189–1232。https://doi.org/10.1214/aos/1013203451。

[21] Glawion L, Polz J, Kunstmann HG, Fersch B, Chwala C（2023）spateGAN：采用条件生成对抗网络方法对降雨场进行时空降尺度。ESS开放档案。https://doi.org/10.22541/essoar.167690003.33629126。

[22] Gneiting T（2011）制作和评估点预报。《美国统计学会杂志》106(494)：746–762。https://doi.org/10.1198/jasa.2011.r10138。

[23] Gneiting T, Raftery AE（2007）严格适当的评分规则、预测与估算。《美国统计学会杂志》102(477)：359–378。https://doi.org/10.1198/016214506000001437。

[24] Hastie T, Tibshirani R, Friedman J（2009）《统计学习要素》。施普林格，纽约。https://doi.org/10.1007/978-0-387-84858-7。

[25] He X, Chaney NW, Schleiss M, Sheffield J (2016) 基于自适应随机森林的降水空间降尺度。《Water Resources Research》52(10):8217–8237. https://doi.org/10.1002/2016WR019034。

[26] Hengl T, Nussbaum M, Wright MN, Heuvelink GBM, Gräler B (2018) 随机森林作为空间和时空变量预测建模的通用框架。PeerJ 6(8):e5518. https://doi.org/10.7717/peerj.5518。

[27] Hollister JW (2022) `elevatr`：从各种应用程序接口获取高程数据。R 软件包版本 0.4.2. https://CRAN.R-project.org/package=elevatr。

[28] Hsu K-L, Gao X, Sorooshian S, Gupta HV (1997) 基于人工神经网络的遥感信息降水估算。《应用气象学杂志》36(9):1176–1190. https://doi.org/10.1175/1520- 0450(1997)036<1176:PEFRSI>2.0.CO;2。

[29] Hu Q, Li Z, Wang L, Huang Y, Wang Y, Li L (2019) 降雨空间估算：从空间插值到多源数据融合的综述。Water 11(3):579. https://doi.org/10.3390/w11030579。

[30] Huffman GJ, Stocker EF, Bolvin DT, Nelkin EJ, Tan J (2019) GPM IMERG晚期降水 L3 1 day 0.1 degree x 0.1 degree V06，编辑者 Andrey Savtchenko，格林贝尔特，MD，戈达德地球科学数据与信息服务中心 (GES DISC)，访问日期：[2022-10-12], https://doi.org/10.5067/GPM/IMERGDL/DAY/06。

[31] James G, Witten D, Hastie T, Tibshirani R (2013)《统计学习导论》。施普林格，纽约。https://doi.org/10.1007/978-1-4614-7138-7。

[32] Ke G, Meng Q, Finkey T, Wang T, Chen W, Ma W, Ye Q, Liu TY (2017) LightGBM：一种高效的梯度提升决策树。《神经信息处理系统进展》30:3146–3154。

[33] Koenker RW, Bassett Jr G（1978年）回归分位数。《计量经济学》46(1):33–50。https://doi.org/10.2307/1913643。

[34] LeCun Y, Bengio Y, Hinton G（2015）深度学习。《自然》521(7553):436–444。https://doi.org/10.1038/nature14539

[35] Lei H, Zhao H, Ao T（2022）利用机器学习分类与回归，在中国采用两步合并策略整合多源降水产品和雨量站观测。《水文学与地球系统科学》26（11）：2969–2995。https://doi.org/10.5194/hess-26-2969-2022。

[36] Lichtendahl Jr. KC、Grushka-Cockayne Y、Winkler RL（2013）对概率取平均还是对分位数取平均更好？《管理科学》59（7）：1594–1611。https://doi.org/10.1287/mnsc.1120.1667。[37] Lin Q、Peng T、Wu Z、Guo J、Chang W、Xu Z（2022）三峡水库区 GPM IMERG 和 TRMM 3B42 产品的性能评估、误差分解及基于树的机器学习误差校正。《Atmospheric Research》268：105988。https://doi.org/10.1016/j.atmosres.2021.105988。[38] Mayr A、Binder H、Gefeller O、Schmid M（2014）提升算法的演进：从机器学习到统计建模。《医学信息方法》53（6）：419–427。https://doi.org/10.3414/ME13-01-0122。[39] Meinshausen N（2006）分位数回归森林。《机器学习研究杂志》7：983–999。[40] Menne MJ、Durre I、Vose RS、Gleason BE、Houston TG（2012）全球历史气候网络日值数据库概述。《大气与海洋技术杂志》29（7）：897–910。https://doi.org/10.1175/JTECH-D-11-00103.1。[41] Meyer H、Kühnlein M、Appelhans T、Nauss T（2016）比较四种机器学习算法在基于卫星的光学降雨反演中的适用性。《Atmospheric Research》169：424–433。https://doi.org/10.1016/j.atmosres.2015.09.021。[42] Militino AF、Ugarte MD、Pérez-Goya U（2023）纳瓦拉（西班牙）降雨日尺度插值的机器学习方法。《系统、决策与控制研究》445：399–413。https://doi.org/10.1007/978-3-031-04137-2_34。[43] Natekin A、Knoll A（2013）梯度提升机教程。《神经机器人学前沿》7：21。https://doi.org/10.3389/fnbot.2013.00021。[44] Nguyen P、Ombadi M、Soroos hian S、Hsu K、AghaKouchak A、Braithwaite D、Ashouri H、Rose Thorstensen A（2018）PERSIANN 全球卫星降水数据系列：产品综述与评估。《水文学与地球系统科学》22（11）：5801–5816。https://doi.org/10.5194/hess-22-5801-2018。[45] Nguyen P、Shearer EJ、Tran H、Ombadi M、Hayatbini N、Palacios T、Huynh P、Braithwaite D、Updegraff G、Hsu K、Kuligowski B、Logan WS、Sorooshian S（2019）CHRS 数据门户：易于访问的 PERSIANN 全球卫星降水数据公共存储库。《科学数据》6：180296。https://doi.org/10.1038/sdata.2018.296。[46] Nguyen GV、Le X-H、Van LN、Jung S、Yeon M、Lee G（2021）随机森林算法在韩国多种卫星降水产品合并中的应用。《Remote Sensing》13（20）：4033。https://doi.org/10.3390/rs13204033。[47] Papacharalampous G、Tyralis H（2020）使用简单组合进行水文时间序列预测：大数据测试及一年期河流流量可预测性研究。《Journal of Hydrology》590：125205。https://doi.org/10.1016/j.jhydrol.2020.125205。[48] Papacharalampous G、Tyralis H（2022a）用于应对概率水文后处理与预报挑战的机器学习概念和方法综述。《水领域前沿》4：961954。https://doi.org/10.3389/frwa.2022.961954。

[49] Papacharalampous G, Tyralis H（2022b）利用大型数据集中的时间序列特征，支持无测站地点的水文气象探索与预测。Water 14(10):1657. https://doi.org/10.3390/w14101657。

[50] Papacharalampous G, Tyralis H, Langousis A, Jayawardena AW, Sivakumar B, Mamassis N, Montanari A, Koutsoyiannis D（2019）大规模概率水文后处理：为何以及如何应用机器学习分位数回归算法。Water 11(10):2126. https://doi.org/10.3390/w11102126。

[51] Papacharalampous G, Tyralis H, Doulamis A, Doulamis N（2023a）用于融合网格化卫星降水数据与地球观测降水数据的机器学习算法比较。Water 15(4):634. https://doi.org/10.3390/w15040634。

[52] Papacharalampous G, Tyralis H, Doulamis A, Doulamis N（2023b）基于树的集成算法在日时间尺度上融合卫星和地球观测降水数据的比较。Hydrology 10(2):50. https://doi.org/10.3390/hydrology10020050。

[53] Pebesma E（2018）用于 R 的简单特征：空间向量 数据的标准化支持。The R Journal 10 (1):439–446。 https://doi.org/10.32614/RJ-2018-009。

[54] Pebesma E（2022）sf：用于 R 的简单特征。R 软件包版本 1.0-9。https://CRAN.R-project.org/package=sf。

[55] Pierce D（2021）ncdf4：用于 Unidata netCDF（版本 4 或更早版本）格式数据文件的接口。R 软件包版本 1.20。 https://CRAN.R-project.org/package=ncdf4。

[56] R Core Team（2022）R：用于统计计算的语言和环境。R Foundation for Statistical Computing，奥地利维也纳。 https://www.R- project.org。

[57] Sagi O，Rokach L（2018）集成学习：综述。Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 8(4):e1249。 https://doi.org/10.1002/widm.1249。

[58] Schmidhuber J（2015）神经网络中的深度学习：概述。Neural Networks 61：85–117。https://doi.org/10.1016/j.neunet.2014.09.003

[59] Shen Z，Yong B（2021）使用梯度提升决策树方法对中国大陆基于 GPM 的卫星降水反演进行降尺度。Journal of Hydrology 602:126803。https://doi.org/10.1016/j.jhydrol.2021.126803。

[60] Shi Y，Ke G，Soukhavong D，Lamb J，Meng Q，Finley T，Wang T，Chen W，Ma W，Ye Q，Liu T-Y，Titov N（2022）lightgbm：轻量级梯度提升机。R 软件包版本 3.3.4。https://CRAN.R-project.org/package=lightgbm。

[61] Sun Q, Miao C, Duan Q, Ashouri H, Sorooshian S, Hsu K-L（2018）全球降水数据集综述：数据源、估算与相互比较。《地球物理学评论》56(1):79–107。https://doi.org/10.1002/2017RG000574。

[62] Tyralis H, Papacharalampous G（2021a）能源研究中的提升算法：系统综述。《神经计算与应用》33(21):14101-14117。https://doi.org/10.1007/s00521-021-05995-8。

[63] Tyralis H, Papacharalampous GA（2021b）基于分位数的水文建模。Water 13(23):3420。https://doi.org/10.3390/w13233420。

[64] Tyralis H, Papacharalampous G（2022a）机器学习概率预报与预测综述。https://arxiv.org/abs/2209.08307。

[65] Tyralis H，Papacharalampous G（2022b）scoringfunctions：用于评估点预报的评分函数集合。https://CRAN.R-project.org/package=scoringfunctions。

[66] Tyralis H, Papacharalampous GA（2023）用于预测极端分位数的水文后处理。Journal of Hydrology 617（C部分）：129082。https://doi.org/10.1016/j.jhydrol.2023.129082。

[67] Tyralis H、Papacharalampous G、Burnetas A、Langousis A（2019a）使用分位数回归算法堆叠泛化进行水文后处理：在美国本土的大规模应用。Journal of Hydrology 577:123957。https://doi.org/10.1016/j.jhydrol.2019.123957。

[68] Tyralis H、Papacharalampous G、Langousis A（2019b）面向水文学科学家和从业者的随机森林简要综述及其在水资源领域的近期发展。Water 11(5):910。https://doi.org/10.3390/w11050910。

[69] Tyralis H、Papacharalampous G、Langousis A（2021）用于日径流预测的超级集成学习：大规模演示及与多种机器学习算法的比较。《神经计算与应用》33(8):3053–3068。https://doi.org/10.1007/s00521-020-05172-3。

[70] Wickham H (2022) `tidyverse`：轻松安装并加载“Tidyverse”。R 软件包版本 1.3.2。https://CRAN.R-project.org/package=tidyverse。

[71] Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Paige Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019) 欢迎使用 tidyverse。《开放源代码软件杂志》4(43):1686。https://doi.org/10.21105/joss.01686。

[72] Wickham H, Hester J, Chang W, Bryan J (2022) `devtools`：让开发 R 软件包更轻松的工具。R 软件包版本 2.4.5。https://CRAN.R- project.org/package=devtools。

[73] Wolpert DH (1992) 堆叠泛化。Neural Networks 5(2):241–259. https://doi.org/10.1016/S0893-6080(05)80023-1。

[74] Wright MN (2022) `ranger`：随机森林的快速实现。R 软件包版本 0.14.1。https://CRAN.R-project.org/package=ranger。

[75] Wright MN, Ziegler A (2017) ranger：在 C 中处理高维数据++ 和 R 的随机森林快速实现。统计软件杂志 77(1):1–17。https://doi.org/10.18637/jss.v077.i01。

[76] Xie Y (2014) knitr：在 R 中实现可复现研究的综合工具。载于：Stodden V、Leisch F、Peng RD（编）《实现可复现计算研究》。Chapman and Hall/CRC。

[77] Xie Y (2015)《使用 R 和 knitr 编写动态文档》，第 2 版。Chapman and Hall/CRC。

[78] Xie Y（2022）`knitr`：用于 R 动态报告生成的通用软件包。R 软件包版本 1.41。https://CRAN.R-project.org/package=knitr。

[79] Xie Y、Allaire JJ、Grolemund G（2018）《R Markdown：权威指南》。Chapman and Hall/CRC。ISBN 9781138359338。https://bookdown.org/yihui/rmarkdown。

[80] Xie Y、Dervieux C、Riederer E（2020）《R Markdown 烹饪手册》。Chapman and Hall/CRC。ISBN 9780367563837。https://bookdown.org/yihui/rmarkdown- cookbook。

[81] Xiong L、Li S、Tang G、Strobl J（2022）地貌测量学与地形分析：数据、方法、平台和应用。《地球科学评论》233:104191。https://doi.org/10.1016/j.earscirev.2022.104191。

[82] Yao Y, Vehtari A, Simpson D, Gelman A（2018）使用堆叠对贝叶斯预测分布进行平均。《贝叶斯分析》13(3):917–1003. https://doi.org/10.1214/17-BA1091。

[83] Zandi O, Zahraie B, Nasseri M, Behrangi A（2022）在地形复杂区域生成高分辨率月降水量：堆叠机器学习模型与局部加权线性模型的比较。Atmospheric Research 272:106159。 https://doi.org/10.1016/j.atmosres.2022.106159。 [84] Zhang L, Li X, Zheng D, Zhang K, Ma Q, Zhao Y, Ge Y（2021）采用一种新型双重机器学习方法融合多种卫星降水产品与雨量站观测。Journal of Hydrology 594:125969。 https://doi.org/10.1016/j.jhydrol.2021.125969。 [85] Zhang Y, Ye A, Nguyen P, Analui B, Sorooshian S, Hsu K（2022）QRF4P-NRT：使用分位数回归森林对近实时卫星降水估计进行概率后处理。Water Resources Research 58(5):e2022WR032117。 https://doi.org/10.1029/2022WR032117。