arXiv:2303.06501v2 [stat.ME] 4 Apr 2024

从有限时间数据中学习：应用于地球科学的动态稀疏历史函数线性模型

Joseph Janssen *不列颠哥伦比亚大学地球、海洋与大气科学系和 Shizhe Meng不列颠哥伦比亚大学数学系和 Asad Haris不列颠哥伦比亚大学地球、海洋与大气科学系和 Stefan Schrunner挪威生命科学大学数据科学系和 Jiguo Cao西蒙菲莎大学统计与精算科学系和 William J. Welch不列颠哥伦比亚大学统计系和 Nadja Kunz不列颠哥伦比亚大学Norman B Keevil矿业工程研究所和 Ali A. Ameli不列颠哥伦比亚大学地球、海洋与大气科学系2024年4月5日

*本研究由加拿大统计科学研究所（CANSSI）授予 Ali A. Ameli、William J. Welch 和 Jiguo Cao 的协作研究团队项目资助。Joseph Janssen 曾

摘要

科学家和统计学家常常希望了解两个随时间变化的变量之间复杂的关联关系。近期关于稀疏函数历史线性模型的研究证实，这类模型很有望用于此目的，但仍存在若干明显局限。最重要的是，以往研究对历史系数函数施加了稀疏性约束，却未允许稀疏性（因而也包括滞后）随时间变化。我们通过采用矩形系数结构并结合惠特克平滑，简化了稀疏函数历史线性模型的框架；随后，通过从分层系数结构中估计动态时间滞后，减少了以往框架所需的假设。我们的研究旨在从水文数据中提取隐藏的物理降雨径流过程。我们通过八项模拟研究展示了该方法的潜力与准确性，并进一步利用两组真实水文数据验证了这一点。

关键词： 函数型数据分析、水文学、降雨径流关系、河川流量、非平稳时间滞后、单位线

由加拿大自然科学与工程研究委员会（NSERC）博士奖学金资助。Asad Haris 获得了 UBC 数据科学研究所提供的博士后经费支持。Jiguo Cao 和 William J. Welch 的研究分别部分获得了 NSERC探索基金（RGPIN-2018-06008 和 RGPIN-2019-05019）的资助。Stefan Schrunner 衷心感谢挪威生命科学大学为其在加拿大不列颠哥伦比亚大学进行国际访问提供的经费支持（项目编号 1211130114）。

1 引言

在许多复杂系统中，因果变量会经过复杂且未知的非平稳过程滤波，转化为某种感兴趣的响应。在众多学科领域中，因果影响如何随时间传播并发生变化，是科学家和政策制定者十分关注的问题。如果能够以数据驱动的方式学习这种滤波函数，人们便能从各种时间序列数据源中提取有关生成变异性的机制的有用知识。

具体而言，本研究旨在了解径流时间序列背后的数据生成过程。水文学中称为集水区的观测单位会将降水转化为河川流量（图1）。这个过程听起来可能很简单；然而，略有差异的地质、气候或地形特征，都可能产生截然不同的滤波函数。了解特定集水区的滤波函数至关重要，原因有很多。例如，若知道某个集水区会将大部分降雨缓慢地通过地下和深层径流路径滤出，就意味着该河道容易受到采矿等人类活动造成的土壤和深层岩石污染（Kunz，2020年）。此外，推断导致过程变化的地质或气候因素，有助于改进未测量河川流量地区的预测。事实上，准确估计将降水转化为河川流量的滤波函数，有助于水文学家理解复杂时间水文数据背后的物理过程。

正如 Botter 等人（2010年）所指出的，对于给定的集水区，水输入（x）与河川流量（y）可通过时变单位线模型（方程（1））中的历史滤波函数（h）建立因果关系，其中单位线定义为单位降雨脉冲之后出现的河川流量响应。

$$y(t) = \int_0^\infty h(s, \mathbf{P}(\mathbf{s}), \mathbf{T}(\mathbf{s}), \mathbf{E}(\mathbf{s}), \mathbf{L}(\mathbf{s}))x(t-s)\,ds \tag{1}$$

尽管方程（1）被表述为水输入（x）与滤波函数（h）之间的线性卷积，但由于滤波函数会随前期水输入（**P**）、蒸散（**E**）、温度（**T**）及景观变量（**L**）的变化而随时间改变，因此该方程实际上是非线性的。请注意，函数 h 取决于其输入的所有历史值，因此加粗的 **s** 由 $t-(s+1)$ $t-(s+2)$...0 组成。蒸发或蒸腾会改变滤波函数的效应大小，因为蒸发量较大的集水区中，只有少部分水可供滤入河川流量。另一方面，水流动或储存在地表下方的深度决定了降水事件与河川流量上升之间的时间间隔（Jasechko 等人，2016年；Somers 和 McKenzie，2020年）。地表径流（OF）是降水可经过的最快、最浅路径；经滤波进入浅层地下径流（SF）的降水会渗入土壤并以较慢的速度流向河道；而深层地下水流（GF）则深入地下，是响应最慢的径流路径（图1）。

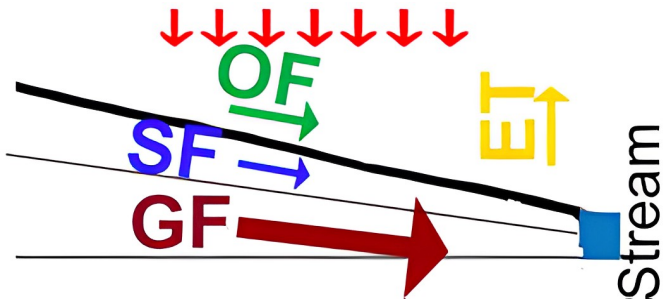


图1：流路示意图展示了降水（红色箭头）如何分配为蒸散（ET）、地表径流（OF）、浅层地下径流（SF）或深层地下水流（GF）；箭头宽度表示相对影响大小的示意值。

科克西拉河（不列颠哥伦比亚省）和 Withlacoochee 河（佛罗里达州）的河川流量与降雨时间序列表明，降雨与河川流量之间的关系确实复杂且高度非平稳（图2）。在科克西拉河，冬季降雨几乎不会蒸发（第1—50天和第320—365天），而是迅速

渗入集水区，形成河川流量。相反，在夏季和初秋（第100—310天），即使有降雨，也不会显著增强河川流量。在 Withlacoochee 河，与科克西拉河相比，少雨或无雨期间河川流量下降得没那么快（例如第1—50天）。该集水区的蒸发量可能较高，因此只有在发生强降雨时，河川流量才可能增加，如第50天所示。

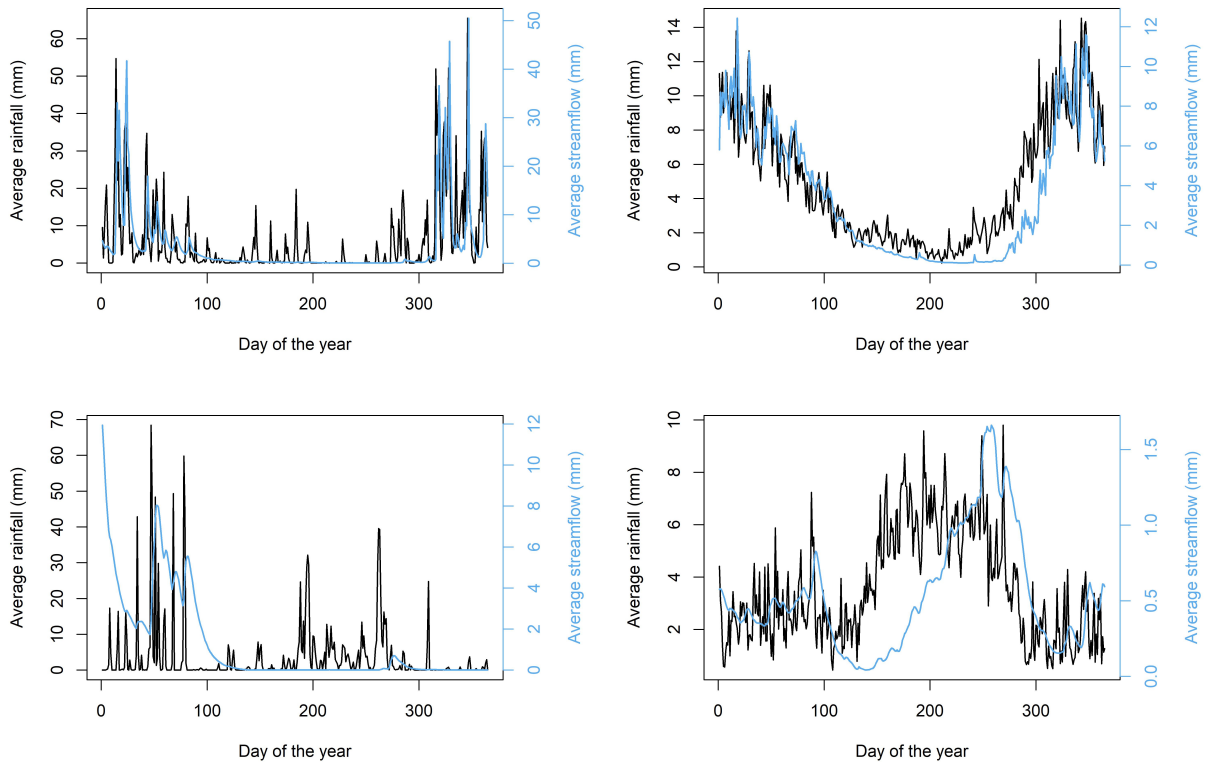


图2：左：1998年科克西拉河（上）和 Withlacoochee 河（下）的每日河川流量（蓝色）和降雨（黑色）。右：1979—2018年科克西拉河（上）和 Withlacoochee 河（下）的日均河川流量（蓝色）和降雨（黑色）。

最贴近方程（1）所述因果过滤关系的统计模型来自 Xun 等人（2022），他们研究了历史函数线性模型：

$$y_i(t) = \int_{t-\delta}^{t} \beta(s,t) x_i(s)\, ds + \epsilon_i(t). \tag{2}$$

这里，$y_i(t)$ 是重复序列 i 在时间 t 的关注结果变量，$x_i(s)$ 是同一重复样本在过去时间 s 的时间解释变量，而 $\epsilon_i(t)$ 是残差项。双变量系数函数 $\beta(s,t)$ 表示过去时间 s 的解释变量对当前时间 t 的结果变量的影响，其中 $t-\delta \leq s \leq t$。为估计 $\beta(s,t)$，Xun 等人（2022）最小化了一个包含三项的目标函数：（1）最小二乘损失；（2）嵌套组桥惩罚项，用于确定常数滞后 δ，超过该滞后后，$\beta(s,t)$ 为零；（3）平滑度惩罚项。这一方法建立在许多既有方法之上，包括 Malfait 和 Ramsay（2003）最初提出的函数历史线性模型，以及 Harezlak 等（2007）施加于系数的离散差分惩罚，该惩罚与 P 样条框架（Eilers 等，2015）类似。Malfait 和 Ramsay（2003）及 Harezlak 等（2007）主要关注预测能力，而 Xun 等人（2022）则更关注准确的参数估计。参数估计也是本文的关注重点。

Xun 等人（2022）提出的方法颇具吸引力，但也存在若干明显缺点。首先，该方法过于复杂。其原始形式是非凸的，因此需要反复执行等价于 Lasso 的优化步骤，直至收敛。用于参数化 $\beta(s,\ t)$ 的帐篷状基函数会增加计算成本，因此基函数的数量通常受到限制（以往研究中使用的基函数约为 230 个）。其次，三角形有限元框架人为地对最大时间滞后施加了限制（即，公式（2）中的积分域不能为负）。例如，若要预测图2中 Withlacoochee 河第一天的河川流量，则只能使用一个降雨值（0毫米），导致对1998年河川流量峰值的预测不准确。第三，以往的框架不允许时间滞后参数 $\delta(t)$ 随时间变化。尽管通常会假设一个任意固定的时间窗口，认为降水会在该窗口内影响河川流量，但这是以往研究的一项重大局限（Janssen 等人，2021），因此找出真实的潜在 $\delta(t)$ 对水文

领域而言十分重要（Tennant 等人，2020；Chiu 和 Bittler，1969）。分布滞后文献中还可以找到其他类似方法，它们也存在类似问题（Almon，1965年；Pesando，1972年；Rushworth 等人，2013；Asencio 等人，2014；Gupta，1968年；Liao 等人，2023）。

本文有三项主要贡献。首先，据我们所知，这是首次尝试在历史函数线性模型中估计时变时间滞后参数。其次，我们以张量积惠特克基取代帐篷状基函数，从而简化了以往的历史函数线性模型方法并提高了其灵活性，使得能够恢复分辨率更高的系数和时间滞后函数。最后，我们成功证明，我们的方法能够通过准确估计真实滤波函数 $\beta(s,\ t)$ 和时变时间滞后参数 $\delta(t)$，为水文学开辟新的研究方向。

2 具有动态稀疏性的历史函数线性模型（HFLM-DS）

令 $y_i(t)$ 表示重复样本 $i \in 1...n$ 在时间 $t \in 1 \dots T$ 的观测响应，并令 $x_i(t)$ 表示相应的协变量。在我们的应用中，$y_i(t)$ 表示第 i 年的每日河川流量，$x_i(t)$ 表示每日降雨量。我们首先考虑方程（1）中描述的精确河川流量因果关系，其中 $\mathbf{V}$ 是所有重要前期变量的集合{**P**、**T**、**E**、**L**}。随后，我们将真实滤波函数 $h(s,\mathbf{V})$ 分解为一个可识别分量和一个不可识别分量：前者因前期条件的季节性变化保持一致而在各年份间稳定（$\beta(s,t)$），后者则逐年变化（$h'(s,\mathbf{V})$）。

$$
\begin{aligned}
y_i(t) &= \int_0^\infty h(s,\mathbf{V}) x_i(t-s)\ ds && \text{Exact filtering function model} \\
&= \int_0^\infty \left[\beta(s,t) + h'(s,\mathbf{V})\right] x_i(t-s)\ ds && \text{Expand } h \text{ into subcomponents} \\
&= \int_0^\infty \beta(s,t) x_i(t-s)\ ds + \int_0^\infty h'(s,\mathbf{V}) x_i(t-s)\ ds && \text{Distribute and separate terms} \\
&= \int_0^\infty \beta(s,t) x_i(t-s)\ ds + \epsilon_i(t) && \text{Unparamaterized term as error}
\end{aligned}
$$

分离各项后，将不可识别分量视为自相关误差项 $\epsilon_i(t)$，并令 D 表示所考虑滞后的总数，则我们的函数型历史线性模型为

$$
y_i(t) = \int_0^{D-1} \beta(s,t) x_i(t-s)\ ds + \epsilon_i(t), \tag{3}
$$

系统部分如图3所示。

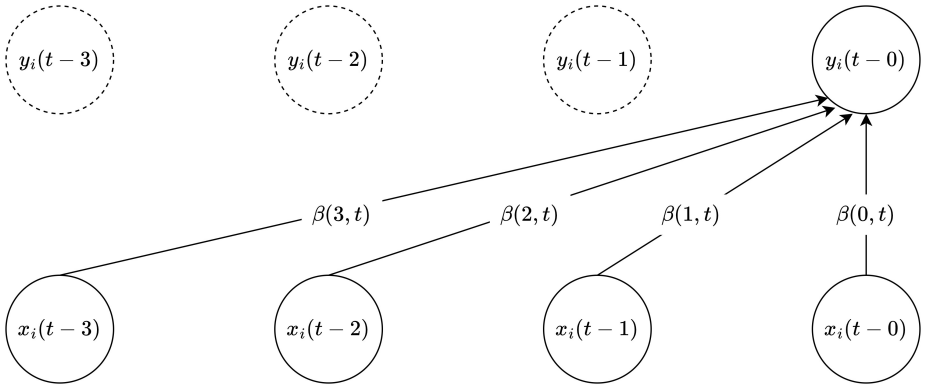


图3：展示了响应 y 与原因 x 在 D=4 下的关系。

与 Xun 等人（2022）一样，我们不失一般性地假设，响应和协变量中的季节性信号已被去除，因此它们在每个年内日序上的均值均为零。使用降雨和径流异常不仅消除了函数型截距项的必要性，也消除了两个信号中可能存在的任何周期性混杂效应（ Moges 等人，2022年）。我们注意到 $t-s$ 可以为负值。出现这种情况时，我们从前一个重复样本 $i-1$的末尾取值。

设 $\phi_1(s,t)...,\phi_K(s,t)$ 表示一个由 K 个已知基函数组成的序列。给定这些基函数及其对应的基系数 $\mathbf{b} = [b_1,...,b_K]^T$，系数函数 $\beta(s,t)$ 可表示为

$$\beta(s,t) = \sum_{k=1}^{K} b_k \phi_k(s,t).$$

利用这一基展开，我们可以将公式（3）重新表述为

$$\begin{aligned} y_i(t) &= \int_0^{D-1} \sum_{k=1}^{K} b_k \phi_k(s,t) x_i(t-s)\ ds + \epsilon_i(t) \\ &= \sum_{k=1}^{K} b_k \int_0^{D-1} \phi_k(s,t) x_i(t-s)\ ds + \epsilon_i(t) \\ &= \sum_{k=1}^{K} b_k z_{ik}(t) + \epsilon_i(t), \end{aligned}$$

其中

$$z_{ik}(t) = \int_0^{D-1} \phi_k(s,t) x_i(t-s)\ ds.$$

如引言中所述，我们的首要目标之一是推断动态时间滞后 $\delta(t)$，在此之后，原因对响应不再有可辨识的影响，即

$$\delta(t) = \max\{s\} \qquad \text{subject to } \beta(s,t) \neq 0. \tag{4}$$

确定 $\delta(t)$ 的准确度直接取决于用于参数化 $\beta(s,\ t)$ 的基函数分辨率；但如果继续使用以往研究采用的帐篷状基函数（Xun 等，2022），增加基函数数量可能会带来难以承受的计算开销。在这种情况下，使用惠特克基是最合适的解决方案（Eilers，2003；Whittaker，1922）。在本研究中，我们用惠特克基函数的张量积 $\phi_k(s,\ t) = w_m(s) \otimes w_\ell(t)$ 对系数函数 $\beta(s,\ t)$ 进行参数化，适用于$m \in \{0,...,D-1\}$ 和 $\ell \in \{1,...,T\}$，其中 $m = k-1-(\lceil k/D \rceil - 1) \times D$ 和 $\ell = \lceil k/D \rceil$。这里，$w_m(s)$ 和 $w_\ell(t)$ 表示惠特克基函数，它们等价于节点间距为 1 的零次 B 样条基函数（Eilers，2003）：

$$w_p(s) = \begin{cases} 1 & p \le s < p+1 \\ \\ 0 & \texttt{otherwise} \end{cases}.$$

假设对于所有 t，$x_i(t)$ 与 $y_i(t)$ 之间合理的最大滞后影响出现在滞后 $s = D-1$，那么我们必须估计 $K = DT$ 个参数 $\{b_k : k = 1, \ldots\ldots, K\}$。一般而言，我们预期 $nT = N \ll K$，因此该系统通常是欠定的。若不对系数函数 $\beta(s, t)$ 的结构，或等价地对系数 b_k，作出某些假设，就无法解决这一问题。因此，我们提出三项主要假设。第一，对于相同的滞后，相邻时间点的系数值应当相近（例如，$\beta(s, t) \approx \beta(s, t+1)$）。第二，对于相同的时间点，相邻滞后的系数值应当相近（例如，$\beta(s, t) \approx \beta(s+1, t)$）。最后，我们预期 $\beta(s, t)$ 具有相当程度的稀疏性。在拟合模型之前，稀疏程度尚不确定，不过我们预期稀疏区域会出现在较大的滞后处。我们尽量减少假设的数量，以使方法具有灵活性并广泛适用（Clark 等人，2011）；不过，针对本研究，我们也可以进一步假设系数函数应为非负，因为额外的降水不可能减少河川流量。我们没有加入这一约束，而是让模型保持灵活，以便进一步评估该方法在科学上的有效性。

基于这三项假设，我们现在对 $\beta(s, t)$施加正则化和稀疏性约束。根据第一项假设，我们对系数函数在相邻时间点上的平方差进行正则化。这可以通过水平一阶差分惩罚矩阵（D_H）实现；该矩阵最早由 Harezlak 等（2007）应用于函数型历史线性模型，之后又被 Xun 等人（2022）使用。我们还以类似 Garcia（2010）的方式为 D_H 添加周期边界条件，使得 $\beta(s, T) \approx \beta(s, 1)$。同样，根据第二项假设，我们通过引入垂直一阶差分惩罚矩阵（D_H），对同一时间点但滞后值相邻的系数之间的平方差施加惩罚。

进一步地，我们引入垂直一阶差分惩罚矩阵（D_V）。此外，我们考虑在垂直惩罚项上施加顶部零边界条件，因为假设超过 $D-1$ 的滞后不会产生影响。图4展示了系数和惩罚项的结构。此外，若取 $D=2$、 $T=3$ 和 $K=6$，则 D_H、 D_V 和 $\beta(s,\ t)$ 可表示为：

$$D_H = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \\ 1 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & 0 & -1 \end{bmatrix} \quad D_V = \begin{bmatrix} -1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \beta(s,t) = \begin{bmatrix} \beta(0,1) \\ \beta(1,1) \\ \beta(0,2) \\ \beta(1,2) \\ \beta(0,3) \\ \beta(1,3) \end{bmatrix}.$$

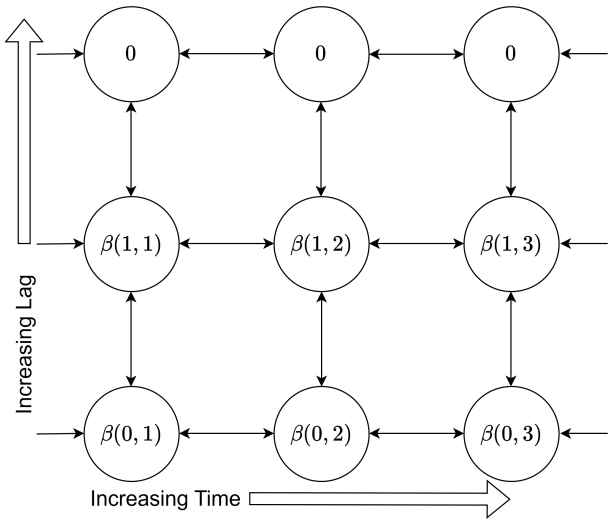


图4：对于 $D=2$ 和 $T=3$，$\beta(s,\ t)$ 的一般结构如图所示，其中 s 表示滞后，t 表示时间。节点之间的双向连线表示节点应具有相似的值。单向连线表示存在周期为 $T+1$ 的周期性行为（即，$\beta(0,\ 3)$ 和 $\beta(0,\ 1)$ 相连）。底部节点行表示将当前时刻的原因与当前时刻的响应联系起来的系数，而其上各行表示连续的过去滞后项对应的系数。顶部行显示零边界条件。

以往大多数提出历史函数线性模型的研究都在模型中施加了稀疏性（Harezlak 等人，2007；Malfait 和 Ramsay，2003年；Xun 等，2022），然而，它们都假设解释特征对响应不再产生影响的滞后在时间上是静态的。对于许多应用而言，尤其是在地球科学领域，这一假设确实过于严格。因此，我们通过嵌套组二范数阈值化，以更灵活的方式施加稀疏性，从而能够可靠地估计动态时间滞后 $\delta(t)$。嵌套组框架使我们能够根据层级结构，将每个系数置于一系列组中（Zhao 等人，2009年）。设 $A_{0,\ t}$，$A_{1,\ t}$，……，$A_{D-1,\ t}$ 是为每个时间点 $t \in 1$，……，T 定义的一系列组。组 $A_{s,\ t}$ 包含对应于 $\beta(s,\ t)$ ……，$\beta(D-1,\ t)$ 的系数索引。请注意，这些组是嵌套的，即 $A_{D-1,\ t} \subset A_{D-2,\ t} \subset \ldots \subset A_{0,\ t}$。在第 2.2 节详细介绍我们的方法和算法时，将进一步讨论这些组。

2.1 自相关误差

正如我们在本节开头的推导中所示，由于用年内日序替代前期温度等动态成分，我们的模型在一定程度上存在设定错误 t (Granger 和 Newbold，1974年)。虽然 t 可作为这些动态前期条件的良好代理变量，但由于不同年份之间存在气候变化，它并不完美。这种不足，加上河川流量固有的日间自相关以及可能存在的系统性测量误差（Horner 等人，2018年；Sorooshian 和 Dracup，1980年；Schoups 和 Vrugt，2010年；Kim 等人，2023），可能导致误差具有很强的自相关性（Sun 等人，2021年；Thursby，1987年）。例如，假设某个集水区在大多数年份的五月都较为干燥，但某一特定年份的五月却异常湿润。为了拟合大多数年份的数据，我们的模型很可能会将五月的 $\beta(s,\ t)$ 设得相当低；然而，对于这个特别湿润的五月，由于河川流量的强度具有持续性，模型会系统性地低估河川流ˆ量（即产生自相关误差），而我们的 $\beta(s,\ t)$

在该特定年份的五月过低。若忽略误差的这一特性，系数估计值仍将保持无偏，但估计效率会较低、方差较大，可能导致无法很好地还原潜在的因果过程（Hibbs，1973年）。自然，我们可以考虑将响应的前几个滞后项纳入模型，以提升预测能力并降低自相关，这也是经典时间序列框架 ARIMAX（Box 等人，2015）所采用的做法；然而，优先追求预测准确度会带来若干问题。首先，过去的河川流量并非未来河川流量的原因，因此将其纳入模型会使模型进一步偏离真实的数据生成过程，进而导致因果效应估计值错误（Dafoe，2018年；Barnett 等人，2017年）。此外，尽管降水数据在全球范围内均可获得且准确，河川流量或蒸发量等其他数据通常并不准确或无法获取。由于水文学家通常需要一种易于推广至所有无测站或数据有限区域的模型（Janssen 和 Ameli，2021年），模型必须排除对过去河川流量值的明确使用。

在不显式将滞后响应值纳入模型的情况下，处理自相关误差最常见的方法是 Cochrane-Orcutt 或 Prais-Winsten 校正方法（Cochrane and Orcutt，1949年；Prais and Winsten，1954年）。遗憾的是，在我们最初的模拟研究中，无论是 Cochrane-Orcutt 还是 Prais-Winsten 校正，在参数估计方面都未能胜过不进行自相关校正的方法。目前尚不清楚这些方法失败的确切原因，但我们有几种推测。首先，我们发现，人为降低验证集中得到的平滑度惩罚项能够改善结果，但我们无法确定一种可用于获得最优结果的、一致适用的降低幅度。其次，我们注意到，如果只考虑少数几个滞后（即 $D < 5$），校正方法的表现远胜于默认回归；然而，如果 $D > 100$，默认回归的表现则远胜于校正方法。探究这些问题及其他问题的成因，或许是未来研究的一个有趣方向（Dagenais，1994年；Sims，1972年；McGuirk and Spanos，2009年；Mizon，1995年；Dafoe，2018年；Thursby，1987年）。

相反，我们发现，使用自回归分布滞后模型后进行反卷积，在存在自相关误差时能取得更好的结果。Kirchner（2022年）对这一方法作了全面阐述，但许多更早的研究中也曾提及该方法（Baltagi and Baltagi，2011年；Pagano and Hartley，1981年；Tsay，1985年；Young，2002年）。简而言之，方程

$$y_t = \alpha_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + b_2 x_{t-2} + \epsilon_t$$

$$y_t = \beta_0 x_t + \beta_1 x_{t-1} + u_t$$

等价，因为我们可以用第二个方程减去其自身乘以α_1 的滞后项，从而得到

$$y_t = \alpha_1 y_{t-1} + \beta_0 x_t + (\beta_1 - \alpha_1 \beta_0) x_{t-1} - \alpha_1 \beta_1 x_{t-2} + \epsilon_t,$$

其中 $u(t)$ 是由 $u_t = \alpha_1 u_{t-1} + \epsilon_t$ 定义的一阶自回归过程。显然，我们可以直接求解第一个方程，无需担心自相关误差，然后通过反卷积，从修正后的 b 系数中得到我们感兴趣的β 系数（更多详情请参见 Kirchner（2022年））。上述简单示例很容易推广到我们更复杂的方法，并在下面的算法中实现。值得庆幸的是，与 Prais-Winsten 程序相比，这也降低了计算成本，因为只需拟合一个模型，通过诊断检验确认不存在自相关，随后进行简单的反卷积即可，而无需拟合两个模型。请注意，所有自相关误差校正方法都隐含地假设共同因子约束成立，这意味着响应不会导致滞后解释变量发生变化（McGuirk and Spanos，2009年）；对于我们的应用而言，这是一个安全的假设。

2.2 算法

令响应$\mathbf{Y}$定义为长度为 N –($D-1$) 的列向量。然后，我们定义稀疏矩阵$\mathbf{Z}$ $\in \mathbb{R}^{N-(D-1)\times K}$，其每一行中有 D 个非零元素，这些元素对应于给定当前时间 $y_i(t)$ 时 $\beta(s,t)$ 的有效条目。这里，$\mathbf{b}$ 是向量，它的】【。

b_k $\forall k \in 1$，.....，K。例如，假设 $n = 2$、$D = 2$、$T = 3$ 和 $K = 6$，那么

$$\mathbf{Y} = \begin{bmatrix} y_1(2) \\ y_1(3) \\ y_2(1) \\ y_2(2) \\ y_2(3) \end{bmatrix} \quad \mathbf{Z} = \begin{bmatrix} \cancel{z_{1,1}(2)}^{0} & \cancel{z_{1,2}(2)}^{0} & z_{1,3}(2) & z_{1,4}(2) & \cancel{z_{1,5}(2)}^{0} & \cancel{z_{1,6}(2)}^{0} \\ \cancel{z_{1,1}(3)}^{0} & \cancel{z_{1,2}(3)}^{0} & \cancel{z_{1,3}(3)}^{0} & \cancel{z_{1,4}(3)}^{0} & z_{1,5}(3) & z_{1,6}(3) \\ z_{2,1}(1) & z_{2,2}(1) & \cancel{z_{2,3}(1)}^{0} & \cancel{z_{2,4}(1)}^{0} & \cancel{z_{2,5}(1)}^{0} & \cancel{z_{2,6}(1)}^{0} \\ \cancel{z_{2,1}(2)}^{0} & \cancel{z_{2,2}(2)}^{0} & z_{2,3}(2) & z_{2,4}(2) & \cancel{z_{2,5}(2)}^{0} & \cancel{z_{2,6}(2)}^{0} \\ \cancel{z_{2,1}(3)}^{0} & \cancel{z_{2,2}(3)}^{0} & \cancel{z_{2,3}(3)}^{0} & \cancel{z_{2,4}(3)}^{0} & z_{2,5}(3) & z_{2,6}(3) \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \\ b_5 \\ b_6 \end{bmatrix}.$$

我们的方法首先通过以下方式估计一个平滑系数函数：

$$\min_{\mathbf{b}} ||\mathbf{Y} - \mathbf{Z}\mathbf{b}||_2^2 + w_h||D_H\mathbf{b}||_2^2 + w_v||D_V\mathbf{b}||_2^2, \tag{5}$$

并使用这个初始系数函数估计值计算组 2-范数（见第 2 节末尾）。随后，使用对应组范数大于一系列 q 值的系数来拟合一系列模型，其中

$$\min_{\mathbf{b}} \ ||\mathbf{Y} - \mathbf{Z}\mathbf{b}||_2^2 + w_h||D_H\mathbf{b}||_2^2 + w_v||D_V\mathbf{b}||_2^2 \text{满足条件 } \mathbf{b}^{A_{s,t}} = 0 \text{ 当 } ||\mathbf{b}^{A_{s,t}}||_2^2 < q.$$

选定最优的 q 后（通过用户选择的方法进行选择），将先前的响应值作为协变量加入，并计算修改后的系数。最后，对系数进行反卷积，使其恢复为预期形式。算法的详细说明见算法 1。

2.3 置信区间

为了近似估计所得 $\beta(s,\ t)$ 和 $\delta(t)$ 关注参数的不确定性，我们可以构造自助法置信区间。然而，通常使用自助法构造置信区间，严重依赖于独立观测的假设（Kunsch，1989）。为解决这一问题，我们建议使用块

算法 1：要求：一个响应时间序列 $Y = [y_1, ..., y_N]$ 和一个解释变量时间序列 $X = [x_1, ..., x_N]$。1：在验证集上寻找使某个目标函数（即 R^2）最大化的最优权重 w_h, w_v。2：使用最优权重 w_h 和 w_v 以及方程 (5) 中的最小二乘准则，求得最优 b。3：计算所有 (s, t) 对应的组范数 $||\mathbf{b}_{A_{s,t}}||_2^2$。4：寻找最优稀疏性阈值 q，以获得 $\delta(t)$。5：给定 q，重新计算最优平滑权重 w_h 和 w_v，并使用全部数据和作为额外协变量的 $y_{t-1}...y_{t-c}$ 重新拟合模型。6：对优化后的系数进行反卷积，得到 $\beta(s, t)$。7：返回 $\beta(s, t)$ 和 $\delta(t)$

如Asencio等人（2014）或Paparoditis和Politis（2003）所述的残差自助法。其基本思路是，在拟合模型时不对自相关误差进行校正，并估计残差序列；随后，我们可以将残差序列划分为多个块，每个块的长度为一年。最后，通过有放回地抽取这些块，并将其添加到模型预测值序列中，即可构造自助样本。我们认为置信区间对本应用并非直接重要，因此在实验中没有采用这种方法；不过，残差自助法和区块自助法已在以往研究中得到广泛检验（Chatterjee和Lahiri，2011年；Lahiri，1999；Paparoditis和Politis，2003年）。

3 水文学数据应用

我们将该方法应用于两个差异显著的集水区数据，以展示其在水文学中的通用性。在真实数据实验中，我们对实际降雨与

河川流量之间的关系进行建模。模拟研究同样使用真实降雨数据，以模拟时间序列中的实际模式；但河川流量则根据已知的系数函数和滞后结构进行模拟，因此我们能够评估估计性能。

我们首先探讨的集水区是科克西拉河集水区，位于加拿大不列颠哥伦比亚省的 Cowichan。该集水区面积为 236km^2 ，平均海拔为 461 米。约 11% 的降水以雪的形式降落；这是一个极其湿润的集水区，干旱指数（潜在蒸散与降水的比值）仅为 0.37。我们探讨的第二个集水区是 Withlachoochee河集水区，位于美国佛罗里达州的戴德城。该集水区面积为 650km^{2,} ，地势低平，平均海拔仅为海平面以上 42 米。这里不会降雪，且集水区相当干燥，干旱指数为 1.07。为获取这两个区域的降雨数据，我们使用 EMDNA（Tang 等人，2021），这是一个高质量气候数据集，包含 1979—2018年北美洲范围内、以 10km 网格单元为单位的完整逐日降水和温度值。基于这些数据，我们通过对科克西拉河和 Withlacoochee 河集水区内的降水和温度值取平均，生成两地的逐日降雨值；然后使用 $0°C$ 的温度阈值计算逐日降雨（图2）。移除闰日后，我们得到 $365*40 = 14,$ 600 个观测值。我们认为最大合理滞后应为 $D = 150$ 天。事实上，无论集水区位于何处，雨水主导型集水区中的所有降水都应在 150 天（5 个月）内排出、被深层地下水系统截留，或从集水区蒸发（Jasechko 等人，2016年；Tennant 等人，2020），因此总共需要估计 $365 \times 150 = 54,$ 750 个参数，以构建 $\beta(s,\ t)$（公式 5）。最大滞后值的选择也会改变可用观测值的总数，因为要预测响应，我们需要此前 150 天的解释变量数据。因此，最终观测值数量为 14,451。

超参数 w_h 和 w_v 使用 *GPfit* 通过贝叶斯优化进行优化

R 软件包（MacDonald 等，2015）。我们首先采用 30 个均匀随机生成的初始点，随后通过最大化验证集R^2 的预期增益，继续进行 35 次迭代以选择下一个点。在多个集水区进行实验后，我们注意到，最优的 w_h 通常远大于最优的 w_v，因此我们将 w_h 的搜索范围设为 $e^{[8,24]}$，并将 w_v的搜索范围设为 $e^{[-5,15]}$ 。

选择最优阈值 q 通常是算法 1 中最困难的一步。在方法部分中，我们没有具体说明应如何进行选择，而是将其留给用户根据自身目标决定，不过我们会结合水文应用提供一些指导。如果目标是特征选择，那么与只选择确定重要的特征相比，剔除少数重要特征并非那么关键，因此可能需要更高的稀疏性。反之，如果只有在确定某个滞后项不重要时才应将其剔除（例如确定性独立筛选（Fan 和 Lv，2008）），那么未来的用户可能会希望稀疏性更低。尽量减小验证集上的误差似乎能得到较高的测试集 R^2 得分，但最终的系数函数表明稀疏性不足。Rushworth 等人（2013）也发现了这一现象：优化赤池信息准则会导致平滑度不足，因此他们将赤池信息准则最优的平滑度参数设为下限，并人为提高该值。在下文的所有实验中，我们按照 Satopaa 等人（2011）所述的方法寻找“拐点”来选择 q，因为该方法始终能得到较好的结果。

我们在所有实验中都令 $c = 2$，这意味着在算法 1 的第 5 步中，我们将响应变量的前两个值作为额外协变量加入自回归分布滞后模型（即误差按二阶自回归过程建模）。这是一个保守的选择，因为我们也能够准确地对一阶自回归误差建模（Schmidt, 1971）。所有实验均使用 R 4.2.1 版运行。

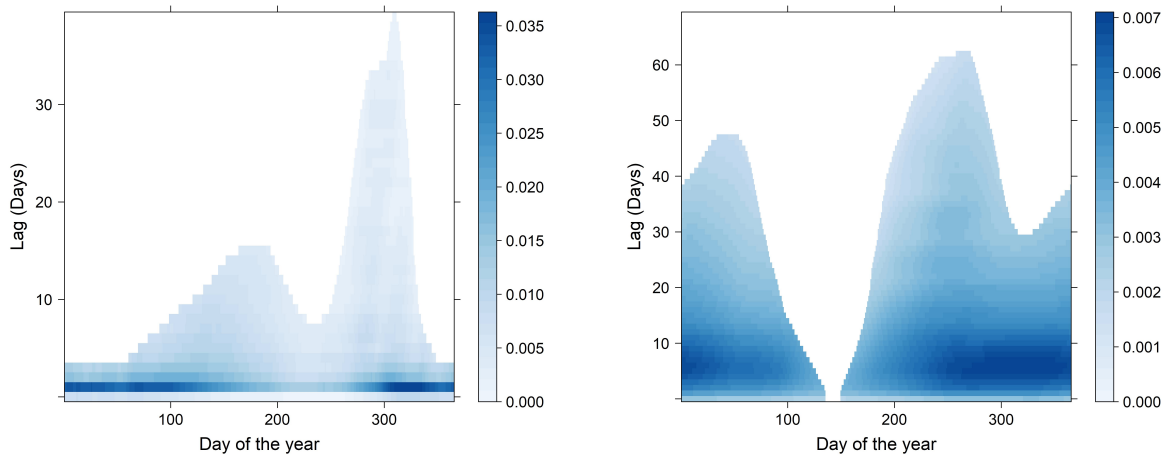
3.1 实际河川流量的应用

科克西拉河和 Withlachoochee河的每日径流数据分别取自加拿大环境部 HYDAT数据库和 CAMELS（Addor 等，2017年）。图2展示了两条河流的日均河川流量。为提高预测性能，并消除残差图中的模式和非高斯性，我们通过 $y_{new} = \log(y + 1)$ 对河川流量 y 进行变换。将数据拆分为 80% 用于训练、20% 用于验证后，我们获得超参数 w_h 和 w_v 的估计值。随后，使用所有可用数据进行训练，获得 $\beta(s,\ t)$、$\delta(t)$ 以及整个数据集 R^2 的最终估计值。我们还使用测试集评估模型的预测性能。为计算测试集 R^2 值，我们将数据拆分为 60% 用于训练、20% 用于优化超参数 w_h 和 w_v，以及 20% 用于测试。

在整个数据集上完成训练后，科克西拉河的估计值 $\beta(s,\ t)$ 如图5a所示。在最终模型中，通过二阶自回归过程对自相关误差进行建模，其拟合系数为（0.704，0.122），表明存在显著的自相关。冬季，降雨频率和强度在经历了从九月到十月再到十一月（第260—320天）的快速上升后达到峰值（图2）；因此，到十二月（第335天）时，集水区内的土壤已完全饱和，地下水也已完全连通。在这种情况下，任何额外的降雨要么立即汇入流向河道的洪水，要么迅速推动土壤中已有的水分流向河道。此外，这一行为也可从图2中看出：十二月、一月和二月（第330—70天）的河川流量与前一天的降雨量呈强相关。因此，从经验上观察到第350—60天期间约3天的 $\delta(t)$，证实并支持了现有的专家知识。图5a中，夏季和秋季分别出现了两个明显的最大滞后峰值。夏季约第190天的峰值，其

$\delta(t)$ 约为18天，显示出该集水区长期的退水行为。从三月开始，该集水区似乎进入干化阶段，此时蒸散和径流超过水输入（图2）。在干化阶段，由于前期干旱程度是河川流量形成的重要驱动因素，较早之前发生的降雨开始对当前河川流量产生相对更大的影响。八月（约第230天）时，土壤完全干燥且温度较高，因此大部分降雨都用于增加土壤湿度和蒸发，而不是形成河川流量。从九月中旬（第260天）开始，随着降雨迅速变得更加频繁，科克西拉河集水区进入快速湿润阶段。此前约40天内发生的降雨开始对河川流量产生显著影响，因为在此期间，如果前几天处于干燥状态，任何额外的降水都会增加土壤湿度；但如果前几天处于湿润状态，额外的降水则会形成河川流量。湿润阶段于十二月（约第350天）结束，此时无论最湿润月份十一月的降水存在何种年际变化，土壤都会完全饱和。在测试集上，$R^2 = 0.82$。

对于第二个集水区——佛罗里达州的 Withlachoochee河，测试数据集上的 R^2 为 0.55。在最终模型中，使用二阶自回归过程对自相关误差进行建模，拟合系数为（ 1.389，-0.408），显示出极强的自相关。尽管图5b中 Withlachoochee河的结果与科克西拉河的结果（图5a）看起来大不相同，但可以用类似的方式来解读。首先需要指出的是，该集水区可能不会出现地表径流，即降雨绕过蓄水过程并在短时间内显著影响河川流量的现象。这是合理的，因为这是一个干燥集水区，干旱指数大于1，意味着潜在蒸发量大于总降水量。从五月开始（第130—175天），集水区进入干燥阶段。每年这个时候，集水区已经经历了约6个月相对干燥的天气，气温也开始升高，



(a) 科克西拉河　　(b) Withlachoochee河

ˆ 图 5：整个数据时期（1979—2018年）内，不列颠哥伦比亚省的科克西拉河（左）和佛罗里达州的 Withlachoochee河（右）在各滞后 s 和时间 t 下的估计函数 $\beta(s,t)$。

因此，五月的大部分降雨会蒸发或补充土壤水分。到了六月中下旬（约第180天），通常会出现持续且稳定的降雨，降雨量超过蒸发量（图2），于是流域土壤变得足够湿润，能够产生河川流量。在集水区完全饱和之前，到了九月（第260天），降雨量低于蒸发量和河川流量，导致较长的滞后开始变得显著，并从十月持续至次年进入干化阶段。

两项真实世界研究均得出了有趣且有意义的结果，并且在测试集上取得了不错的 R^2 分数。两种情况下，都能区分出湿润、湿润状态、干燥和干旱期。此外，尽管我们的模型并未将系数函数为正作为约束条件，但在两种情况下，模型都正确识别出降雨与河川流量在所有时段和所有滞后下均存在很强的正相关关系。这进一步表明，我们的方法可靠，能够正确识别主导过程。两个集水区在测试集上的 R^2 结果与神经网络的结果相似；神经网络更复杂，也较难解释（Hoedt 等人，2021年；Anderson 和 Radi´c,

2022年；Kratzert 等人，2018年）。

4 模拟研究

我们的模拟研究旨在量化并了解我们在各种噪声环境下估计 $\beta(s,\ t)$ 和 $\delta(t)$ 的能力。

4.1 评估标准

Xun 等人（2022）通过计算 δ 的均方根误差、 δ 的百分比偏差，以及 β（$s,\ t$）的平均积分平方误差来比较各种方法。当 δ 随时间变化时，δ 的百分比偏差无法充分反映 δ 估计值的准确性；与 R^2 相比，积分平方根误差等其他指标也较难解释。因此，我们为模拟研究引入三项评估标准。第一项标准 β-R^2 旨在以易于解释的方式评估对真实系数函数的估计准确性：

$$R^2(\beta, \hat{\beta}) = 1 - \frac{\int\int\{\beta(s,t) - \hat{\beta}(s,t)\}^2 \ ds \ dt}{\int\int\{\beta(s,t) - \bar{\beta}\}^2 \ ds \ dt} \tag{6}$$

第二个准则， $\delta(t)$-偏差，用于评估我们对 $\delta(t)$ 的估计中出现的偏差。我们通过计算 $\delta(t)$ 随时间的平均值，再减去 $\delta(t)$ 随时间的平均值，来得到该准则。第三个准则 $\delta(t)$- 相关性，用于评估我们估计 $\delta(t)$ 随时间变化的期际变异性的能力。该指标的计算方法很简单，即计算真实值 $\delta(t)$ 与估计值 $\delta(t)$ 之间的相关性。偏差可能会随时间相互抵消，而较强的相关性并不意味着 $\delta(t)$ 的尺度正确；但这两项评估指标相辅相成，弥补了彼此的不足，同时分别提供了关于我们估计 $\delta(t)$ 尺度和时间动态准确性的相关信息。理想情况下， β-R^2 和 $\delta(t)$-相关性应接近 1，而 $\delta(t)$-偏差应接近 0。

4.2 模拟情景

本节开展八项模拟研究，旨在量化我们的方法在噪声环境下估计 $\beta(s,\ t)$ 和 $\delta(t)$ 的能力。我们使用上一节中相同的真实降雨数据构建协变量矩阵 **Z**，随后设置八种情景，分别采用不同的 $\beta(s,\ t)$、$\delta(t)$、噪声水平和噪声自相关系数，以便模拟响应 $y_i(t)$的已知真实值。河川流量众所周知难以预测。根据集水区的位置及其动态，即使使用长短期记忆（LSTM）神经网络（Hochreiter 和 Schmidhuber，1997）等强大的黑箱方法，并加入更多预测变量，测试集 R^2 的表现范围也可能从 0 到 1 不等（Ayzel 和 Heistermann，2021年；Hoedt 等人，2021年；Kratzert 等人，2018年）。我们假设，在未来可能应用本方法的情景中，基于具有 $R^2 < 0.4$ 的模型进行推断将十分困难。此外，根据以往研究，我们知道在水文学领域，具有 $R^2 > 0.8$ 的可解释模型将十分罕见。因此，我们在响应向量中加入额外噪声，使得如果根据真实的 $\beta(s,\ t)$ 计算真实响应 $y_{\texttt{true}}(t)$，将得到$R^2(y,\ y_{\texttt{true}}) = 0.8$ 或 $R^2(y,\ y_{\texttt{true}}) = 0.4$，其中

$$R^2(y, y_{\texttt{true}}) = 1 - \frac{\int \{y(t) - y_{\texttt{true}}(t)\}^2 \, dt}{\int \{y(t) - \bar{y}\}^2 \, dt}, \tag{7}$$

$y(t)$表示在 $y_{\texttt{true}}(t)$中加入噪声后得到的模拟数据，而 $\bar{y}$ 是 $y(t)$在t上的平均值。此外，我们为模拟噪声设定了两种可能的自相关水平。根据上一节的结果，我们知道噪声可能具有很强的自相关性，因此设定了中等自相关情景 AR(0.6,0.1) 和高自相关情景 AR(1.5,-0.52)。

这八种情景围绕第3.1节真实水文学研究中获得的结果展开。我们采用图5a和图5b中的结果，并将其设为真实情景。将这些 $\beta(s,\ t)$ 值与我们的降雨值矩阵结合，即可计算真实响应值。在两种不同的噪声水平、两种不同的

集水区以及两个自相关水平的组合下，共有八种模拟情景。对于每种情景，我们重复模拟100次，以获得稳定的结果和不确定性水平。每次模拟迭代在配备32 GB内存的 Intel i9-9980HK 2.4 GHz 处理器上耗时不到10分钟，因此我们的方法在任何现代硬件配置上都具有计算可行性。

4.3 模拟结果

表1显示，所提出的方法很有前景且稳健。我们始终观察到，无论位置如何，噪声较少的数据都能对 $\beta(s,\ t)$和 $\delta(t)$进行更准确、稳定的推断。我们还始终观察到，较高的自相关会导致较差的 $\delta(t)$-相关结果。表1所示的平均结果以及各次模拟运行结果均始终呈现负的 $\delta(t)$偏差，揭示了系统性误差。这表明，平均而言， $\delta(t) > \delta(t)$，也就是说，估计的系数函数往往过于稀疏，因此所选的 q 应当降低，这表明我们选择阈值 q 的方法并非最优。

对于科克西拉河，高噪声情景下的 $\beta - R^2$ 始终高于0.9，低噪声情景下则高于0.96。高噪声情景下对 $\delta(t)$的估计也始终较差，而较高的自相关也会产生影响。

对于 Withlachoochee河， $\beta(s,\ t)$ 的估计始终准确：高噪声情景下，R^2 的范围为 0.888 至 0.927；低噪声情景下， R^2 的范围为 0.934 至 0.968。在两种噪声情景下，高相关性情景的偏差约为 -4 天，而低相关性情景的偏差约为 -2.5 天。考虑到 $\delta(t)$ 的范围可能从 2 到 60 多（图5b），这些偏差结果相当令人鼓舞。此外，我们的方法在两种低噪声情景下都能以高于 0.98 的相关性恢复 $\delta(t)$。在高噪声、低自相关情景下， $\delta(t)$ 的相关性略低（0.943）；而在高噪声、高自相关

表1：模拟研究结果汇总。每项第4.1节中定义的评估标准均显示100次重复试验的平均值（标准差）。

位置	$R^2(y,\ y_{\texttt{true}})$	AR(2)系数	β-R^2	$\delta(t)$-偏差	$\delta(t)$-相关系数
科克西拉河	0.4	0.6, 0.1	0.926 (0.016)	-2.699 (1.806)	0.772 (0.126)
科克西拉河	0.8	0.6, 0.1	0.971 (0.005)	-1.773 (0.769)	0.925 (0.043)
科克西拉河	0.4	1.5, -0.52	0.904 (0.030)	-4.133 (2.60)	0.534 (0.280)
科克西拉河	0.8	1.5, -0.52	0.961 (0.009)	-2.42 (0.944)	0.850 (0.145)
Withlacoochee 河	0.4	0.6, 0.1	0.927 (0.014)	-2.773 (1.684)	0.943 (0.039)
Withlacoochee 河	0.8	0.6, 0.1	0.968 (0.005)	-2.391 (0.571)	0.992 (0.003)
Withlacoochee 河	0.4	1.5, -0.52	0.888 (0.037)	-4.912 (4.155)	0.848 (0.097)
Withlacoochee 河	0.8	1.5, -0.52	0.934 (0.015)	-3.736 (1.045)	0.982 (0.009)

情景（0.848）。

5 讨论与结论

在这项工作中，我们以 Malfait 和 Ramsay（2003）、Harezlak 等（2007）、Rushworth 等人（2013）以及 Xun 等人（2022）提出的函数型历史线性模型迭代版本为基础开展研究。我们大幅简化了这些模型的形式，同时借助惠特克基实现动态稀疏性，并对系数函数进行高分辨率估计。在第 2 节中，我们通过若干图示说明算法，并介绍为将高度过参数化的问题简化为可处理问题而作出的三项较小假设。我们仅假设水平平滑性、垂直平滑性，以及稀疏性集中在较大的滞后处，以保持方法的灵活性，使其适用于主要水文研究目标以外的各种应用。在全文中，我们

其主要驱动力是准确估计降雨径流关系这一目标。从温哥华岛和佛罗里达州收集降雨和河川流量数据后，我们估计了两个位置上将降雨转化为河川流量的系数函数，并将所得结论与专家知识进行了比较。我们发现，这些函数具有明确且在水文意义上重要的解释：集水区会经历四个不同阶段：（1）湿润；（2）湿润状态；（3）干燥；（4）干燥状态。由于温哥华岛的科克西拉河流域温度和降雨的季节变化显著，这四个阶段可以清晰地区分。另一方面，我们发现，佛罗里达州的 Withlachoochee河流域极其干燥，因此从未进入湿润状态，但可以推断出其他所有阶段。事实上，若没有针对动态稀疏性的这些方法改进，就无法得出上述任何推断。最后，在一项广泛的模拟研究中，我们发现，简单的方法能够以很高的准确度恢复真实系数函数 $\beta(s,\ t)$ 以及动态时间滞后 $\delta(t)$。虽然系数函数和时间滞后的恢复准确度显然取决于噪声水平和自相关程度，但我们发现，即使当 R^2 小于或等于 0.4 且误差具有极强的自相关性时，仍能根据推断得出准确结论。

我们的方法和实验存在若干不足之处。首先，尽管我们的算法可能纳入多个特征，但这些额外特征未必符合我们的稀疏性和平滑性假设。例如，如果我们将降雪作为多雪集水区河川流量的预测变量，就可能预期降雪需要数天、数周甚至数月才能融化并开始对河川流量产生贡献。这与我们的稀疏性假设相矛盾，因为我们会预期滞后极低和极高时的系数为零。其次，在整个研究中，我们假设集水区的行为不会随年份变化。尽管我们的方法无法捕捉年际关系变异，但可用于提取多个

系数函数，以比较某个已知变化发生前后的情况。例如，可以估计土地利用发生重大变化前后的降雨径流关系，或气候发生严重变化前后的降雨径流关系。现在聚焦于实验的局限性，我们注意到实验仅涉及水文学中的一项特定应用，不过我们推测这些方法在其他领域也能发挥良好作用。此外，我们的模拟研究存在局限，因为我们提供的真实值是已知算法能够恢复的。这个局限又回到了第二项方法论局限：我们的方法可能只能恢复某些类型的系数函数。如果我们改为使用非光滑真实值开展模拟研究，模拟结果很可能会更差。

未来的研究可以解决上述不足之处中的一项或多项。或许也可以通过类似自适应套索的方式引入自适应平滑性，以提高我们方法的灵活性，因为不同日期或滞后可能具有不同的平滑性（Zou，2006；Centofanti 等人，2022；Yang 和 Hong，2017；Martinez 和 Carroll，2010；Ballout 等人，2023）。在自回归分布滞后模型中加入时变 AR 项，或考虑异方差性，也可能改善滤波函数的提取效果（Kim 等人，2023；Rao，2004）。此外，将这些方法拓展至空间分析领域也可能颇有意义。通过平滑的空间归因，可以了解某个原因的特定网格单元对响应的影响程度。此外，借助稀疏性，可以划定对响应有影响的空间区域，并将其与对响应没有显著影响的区域区分开来。我们目前的研究采用惩罚最小二乘法和稀疏性阈值化的视角，但另一种有趣的视角或许可以来自因果推断和条件独立性框架，并由此发展出不同的方法（Laumann 等人，2023）。例如，或许可以利用偏相关或传递熵找到正确的稀疏性（Murari 等人，2018；Moges 等人，2022），尽管高维、小样本量的环境使这一任务颇具挑战。还可以探索更好的超参数选择方法（Bach，2008；Hall 等人，2009；Chatterjee和Lahiri，2011）。我们希望我们的方法能够成为可靠的工具，帮助科学多个领域从时间序列数据中学习，或为进一步的方法学发展提供启发。这一研究领域无疑蕴藏着丰富的科学发现潜力。

我们希望我们的方法能够成为可靠的工具，帮助科学多个领域从时间序列数据中学习，或为进一步的方法学发展提供启发。这一研究领域无疑蕴藏着丰富的科学发现潜力。

补充材料

R代码及实验运行方法： R代码文件 jasa code Submitted.R 包含运行实验所需的全部代码。此外，四个 .rds 文件包含所有河川流量和降雨数据。所有文件请见 `https://github.com/HydroML/HFLM-DS` 。

参考文献

Addor, N., Newman, A. J., Mizukami, N., 和 Clark, M. P. (2017年)。CAMELS数据集：用于大样本研究的流域属性和气象学数据。《水文与地球系统科学》，21(10):5293–5313。

Almon, S. (1965年)。资本拨款与支出之间的分布滞后。《计量经济学会计量经济学》，第178–196页。

Anderson, S. 和 Radi´c, V. (2022年)。卷积长短期记忆网络在区域水文建模中的评估与解释。《水文与地球系统科学》，26(3):795–825。

Asencio, M., Hooker, G., and Gao, H. O.（2014年）。函数卷积模型。《统计建模》，14(4):315–335。

Ayzel, G. and Heistermann, M.（2021年）。校准数据长度对概念性水文模型与长短期记忆网络和门控循环单元性能的影响：基于 CAMELS 数据集中的六个流域的案例研究。《计算机与地球科学》，149:104708。

Bach, F. R.（2008年）。Bolasso：通过自助法实现模型一致的最小绝对收缩和选择算子估计。载于第*25*届国际机器学习大会论文集，第33–40页。

Ballout, N.、Etievant, L. 和 Viallon, V.（2023）。关于使用交叉验证校准自适应套索。《生物计量学杂志》，65(5)：2200047。

Baltagi, B. H. 和 Baltagi, B. H.（2011年）。分布滞后与动态模型。计量经济学，第131–147页。

Barnett, A. G.、Stephen, D.、Huang, C. 和 Wolkewitz, M.（2017年）。环境暴露的时间序列模型：预测准确，还是理解透彻。《环境研究》，154：222–225。

Botter, G.、Bertuzzo, E. 和 Rinaldo, A.（2010年）。水文响应中的输运：旅行时间分布、土壤水分动态与老水悖论。《水资源研究》，46(3)。

Box, G. E., Jenkins, G. M., Reinsel, G. C. 和 Ljung, G. M.（2015）。时间序列分析：预测与控制。约翰·威立父子出版公司。

Centofanti, F., Lepore, A., Menafoglio, A., Palumbo, B. 和 Vantini, S.（2022年）。函数对函数线性回归模型的自适应平滑样条估计量。计算统计学，第 1–26 页。

Chatterjee, A. 和 Lahiri, S. N.（2011年）。LASSO估计量的自助法。《美国统计学会会刊》，106(494):608–625。

Chiu, C.-L. 和 Bittler, R. P.（1969年）。降雨径流关系的线性时变模型。《水资源研究》，5(2):426–437。

Clark, M. P., Kavetski, D. 和 Fenicia, F.（2011年）。在水文建模中践行多重工作假设的方法。《水资源研究》，47(9)。

Cochrane, D. 和 Orcutt, G. H.（1949年）。将最小二乘回归应用于包含自相关误差项的关系。《美国统计学会会刊》，第 32–61 页。

Dafoe, A.（2018年）。时间依赖下因果效应的非参数识别。《社会学方法与研究》，47(2)：136–168。

Dagenais, M. G.（1994年）。含变量误差和自相关扰动项的回归模型中的参数估计。《计量经济学杂志》，64(1-2)：145–163。

Eilers, P. H.（2003年）。完美的平滑器。《分析化学》，75(14)：3631–3636。

Eilers, P. H., Marx, B. D. 和 Durb´an, M.（2015）。P样条二十年。《*SORT*：统计与运筹学汇刊》，39(2):0149–186。

Fan, J. 和 Lv, J.（2008年）。针对超高维特征空间的确定性独立筛选。《英国皇家统计学会杂志：*B*辑（统计方法学）》，70(5):849–911。

Garcia, D.（2010年）。对一维及高维含缺失值网格数据进行稳健平滑。《计算统计与数据分析》，54(4):1167–1178。

Granger, C. W. 和 Newbold, P.（1974年）。计量经济学中的伪回归。《计量经济学杂志》，2(2):111–120。

Gupta, Y.（1968年）。一种估计分布滞后模型的有效方法。技术报告。

Hall, P.、Lee, E. R. 和 Park, B. U.（2009年）。基于自助法的最小绝对收缩和选择算子惩罚项选择，实现预言机性能。《中华统计学志》，第449–471页。

Harezlak, J., Coull, B. A., Laird, N. M., Magari, S. R., 和 Christiani, D. C.（2007）。函数回归问题的惩罚解。《计算统计与数据分析》，51(10):4911–4925。

Hibbs, D. A.（1973）。时间序列回归模型中的统计估计与因果推断问题。《社会学方法论》，5:252–308。

Hochreiter, S. 和 Schmidhuber, J.（1997）。长短期记忆。《神经计算》，9(8):1735–1780。

Hoedt, P.-J., Kratzert, F., Klotz, D., Halmich, C., Holzleitner, M., Nearing, G. S., Hochre- iter, S. 和 Klambauer, G.（2021）。质量守恒长短期记忆网络：一种质量守恒长短期记忆网络。在国际机器学习会议，第4275–4286页。机器学习研究论文集。

Horner, I., Renard, B., LeCoz, J., Branger, F., McMillan, H., and Pierrefeu, G.（2018年）。水位测量误差对径流不确定性的影响。《水资源研究》，54(3):1952–1976。

Janssen, J. 和 Ameli, A.A.（2021年）。一种水文功能方法，可提升水文监测稀疏地区的大样本水文学表现。《水资源研究》，57(9):e2021WR030263。

Janssen, J., Radi´c, V., 和 Ameli, A.（2021年）。评估北美洲各地未来季节性市政供水短缺的风险。《地球科学前沿》，9:730631。

Jasechko, S., Kirchner, J. W., Welker, J. M., 和 McDonnell, J. J. (2016年). 全球径流中不到三个月的水占相当大的比例。《自然·地球科学》, 9(2):126– 129.

Kim, M., Bauser, H. H., Beven, K., 和 Troch, P. A. (2023). 流量衰退动力学的时间变异性：机器学习的应用与从机器中学习。《水资源研究》, 59(5):e2022WR032690.

Kirchner, J. W.（2022年）。通过对含噪时间序列进行反卷积和解混，估计非线性、非平稳和异质系统的脉冲响应函数。《传感器》，22(9):3291。

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., and Herrnegger, M.（2018年）。使用长短期记忆网络（LSTM）进行降雨–径流建模。《水文与地球系统科学》，22(11):6005–6022。

Kunsch, H. R.（1989年）。针对一般平稳观测值的刀切法和自助法。《统计年刊》，第1217–1241页。

Kunz, N. C.（2020年）。拓展矿区水安全视角。《水安全》，11:100079。

Lahiri, S. N.（1999年）。区块自助法的理论比较。《统计年刊》，第386–404页。

Laumann, F., Von K¨ugelgen, J., Park, J., Sch¨olkopf, B., andBarahona, M.(2023). 用于函数型数据因果结构学习的核独立性检验。《熵》，25(12):1597。

Liao, Z., Qian, M., Kronish, I. M., and Cheung, Y. K. (2023). 使用具有自相关误差的贝叶斯分布滞后模型分析单一个体试验。《医学统计学》。

MacDonald, B., Ranjan, P., and Chipman, H.（2015）。Gpfit：用于拟合确定性模拟器输出的高斯过程模型的 R 软件包。《统计软件杂志》，64：1–23。

Malfait, N. and Ramsay, J. O.（2003年）。历史函数线性模型。加拿大统计学杂志，31(2)：115–128。

Martinez, J. G. 和 Carroll, R. J.（2010年）。Oracle 并非最优：调整自适应套索。生物统计学，1(1):1–27。

McGuirk, A. 和 Spanos, A.（2009年）。重新审视误差自相关校正：公共因子约束与格兰杰非因果性。《牛津经济学与统计学通报》，71(2):273–294。

Mizon, G. E.（1995年）。给自相关校正方法的一则简单忠告：不要使用。《计量经济学杂志》，69(1):267–288。

Moges, E.、Ruddell, B. L.、Zhang, L.、Driscoll, J. M. 和 Larsen, L. G.（2022年）。降水对美国本土连续区域河川流量的影响强度与记忆效应。《水资源研究》，58(3):e2021WR030186。

Murari, A., Lungaroni, M., Peluso, E., Gaudio, P., Lerche, E., Garzotti, L., Gelfusa, M. 和 Contributors, J.（2018年）。论利用传递熵研究信号间因果影响的时间范围。《熵》，20(9):627。

Pagano, M. 和 Hartley, M. J.（1981年）。关于拟合受多项式约束的分布滞后模型。《计量经济学杂志》，16(2):171–198。

Paparoditis, E. 和 Politis, D. N.（2003年）。用于单位根检验的基于残差的区块自助法。《计量经济学》，71(3):813–855。

Pesando, J. E.（1972年）。分布滞后模型中的季节变化。《美国统计学会会刊》，67(338):311–312。

Prais, S. J. 和 Winsten, C. B.（1954年）。趋势估计量与序列相关。技术报告，考尔斯委员会讨论论文，芝加哥。

Rao, S. S.（2004年）。关于具有非平稳相关误差的多元回归模型。*Biometrika*, 91(3):645–659。

Rushworth, A. M., Bowman, A. W., Brewer, M. J., 和 Langan, S. J.（2013）。水文数据的分布滞后模型。*Biometrics*，69(2):537–544。

Satopaa, V., Albrecht, J., Irwin, D., 和 Raghavan, B.（2011）。在干草堆中寻找"膝点"：检测系统行为中的膝点。载于*2011 31st international confer-ence on distributed computing systems workshops*，第 166–171 页。IEEE。

Schmidt, P.（1971年）。带有二阶自回归扰动的分布滞后模型估计：一项蒙特卡洛实验。*International Economic Review*，第 372–380 页。

Schoups, G. 和 Vrugt, J. A.（2010年）。用于对具有相关、异方差和非高斯误差的水文模型进行参数推断和预测推断的形式化似然函数。《水资源研究》，46(10)。

Sims, C. A.（1972年）。近似先验约束在分布滞后估计中的作用。《美国统计学会会刊》，第169–175页。

Somers, L. D. 和 McKenzie, J. M.（2020年）。高山环境中的地下水综述。*Wiley*跨学科评论：水，7(6):e1475。

Sorooshian, S. 和 Dracup, J. A.（1980年）。用于水文降雨径流模型的随机参数估计程序：相关误差和异方差误差情形。《水资源研究》，16(2):430–442。

Sun, F.-K., Lang, C.，和 Boning, D.（2021年）。针对时间序列神经网络中的自相关误差进行调整。神经信息处理系统进展，34:29806– 29819。

Tang, G.、Clark, M. P.、Papalexiou, S. M.、Newman, A. J.、Wood, A. W.、Brunet, D. 和 Whitfield, P. H.（2021年）。Emdna：一个面向北美洲的集合气象数据集。地球系统科学数据，13(7):3337–3362。

Tennant, C., Larsen, L., Bellugi, D., Moges, E., Zhang, L. 和 Ma, H.（2020年）。信息流在制定流量预测模型中的效用：来自干旱雪主导流域的案例研究。《水资源研究》，56(8):e2019WR024908。

Thursby, J. G.（1987年）。存在设定误差时采用普通最小二乘法还是广义最小二乘法？：一种期望损失方法。《计量经济学杂志》，35(2-3):359–374。

Tsay, R. S.（1985年）。动态回归（分布滞后）模型中的模型识别。《商业与经济统计杂志》，第228–237页。

Whittaker, E. T.（1922年）。一种新的毕业方法。爱丁堡数学学会会刊，41:63–75。

Xun, X., Guan, T. 和 Cao, J.（2022年）。采用嵌套组桥接方法对历史函数线性模型进行稀疏估计。加拿大统计学杂志，50(4):1254– 1269。

Yang, L. 和 Hong, Y.（2017年）。数据平滑的自适应惩罚样条。计算统计与数据分析，108：70–83。

Young, P. C.（2002年）。实时洪水预报的进展。伦敦皇家学会哲学汇刊。*A*辑：数学、物理与工程科学，360（1796）：1433–1450。

Zhao, P., Rocha, G. 和 Yu, B.（2009年）。用于分组和分层变量选择的复合绝对惩罚族。《统计年刊》，37(6A):3468–3497。

Zou, H.（2006）。自适应套索及其预言机性质。《美国统计学会会刊》，101(476):1418–1429。