arXiv:2603.02059v1 [stat.ML] 2 Mar 2026

TRAKNN：面向罕见气象轨迹检测的高效轨迹感知时空 kNN

Guillaume Coulaud[1] Davide Faranda[2,3,4]

[1] 蒙彼利埃大学、Inria、法国国家科学研究中心、LIRMM，蒙彼利埃，法国 guillaume.coulaud@inria.fr[2] 气候与环境科学实验室，UMR 8212 CEA-CNRS-UVSQ，巴黎萨克雷大学与 IPSL，CEA 萨克雷 l'Orme des Merisiers，91191 吉夫叙尔伊韦特，法国[3] 伦敦数学实验室，8 Margravine Gardens，伦敦 W6 8RH，英国[4] LMD/IPSL，巴黎高等师范学院，PSL大学，巴黎综合理工学院，巴黎理工学院，索邦大学，法国国家科学研究中心，巴黎，法国 davide.faranda@cea.fr

摘要。 极端天气事件，如风暴和热浪，由持续数天演变的大气环流型驱动。传统基于环流的研究往往聚焦于瞬时大气状态，而捕捉这些空间场的时间演变（即轨迹）对于刻画罕见且可能具有重大影响的大气行为至关重要。然而，在数十年、大陆尺度的格点数据集上进行穷尽式相似性搜索会带来显著的计算与内存挑战。本文提出 TRAKNN（轨迹感知 KNN），一种完全无监督且数据无关的框架，采用精确 kNN 方法检测时空数据中几何上罕见的短轨迹。TRAKNN 利用基于递归的算法，将计算复杂度与轨迹长度解耦，并通过高效的批处理操作最大化计算强度。这些优化使得在标准工作站上（无论 CPU 还是 GPU）即可完成穷尽分析。

我们在 75 年的欧洲逐日海平面气压数据上评估了我们的方法。结果表明，TRAKNN 识别出的稀有轨迹对应于物理上一致的大气异常，并且与独立的极端事件数据库相吻合。

关键词： 时空数据 · K近邻 · 环境科学 · 气候 · 类比

1 引言

风暴、热浪或寒潮等极端天气事件对生态系统 [31] 和社会 [17]有着重要影响。在人类活动导致的

气候变化背景下，这些事件的频率和强度正在发生变化 [30]，其社会经济成本在过去数十年间不断上升 [12], ，促使人们持续努力，以更好地理解其背后的物理驱动因素 [19] ，并确定气候变化所发挥的作用 [26]。在欧洲，许多极端事件与持续数天的大尺度大气环流型有关 [14]。许多基于环流的现有研究 [18,20] 聚焦于瞬时大气状态（即单个时间步的空间场），然而，空间场的时间演变在极端事件的形成、增强和持续过程中发挥着核心作用。因此，要刻画罕见大气行为，捕捉短时空轨迹，而非孤立的瞬时快照，至关重要。检测稀有轨迹有助于系统地识别异常的动力演变、检索历史类比，并为将大气动力学与观测到的极端影响联系起来提供合理依据。采用完全由数据驱动的无监督方法论，还能确保该方法不依赖于特定问题或数据，从而无需依靠预定义的天气型、物理阈值或特定领域的先验知识，即可开展系统性研究。

从数据挖掘的角度来看，这种情形自然会引出高维时空数据中的异常检测问题。基于距离的方法，例如基于最近邻的方法 [3,27] ，在这一背景下尤其具有吸引力，因为它们具有非参数特性、易于解释，且建模假设极少。然而，将这类方法系统地应用于跨越数十年的大陆尺度数据集，会带来巨大的计算挑战。对轨迹进行穷举式精确比较，需要随着时间步数量呈二次缩放，并需要大量内存来加载数据；对于涵盖数十年每日观测的数据集而言，在常规硬件上很快就会变得难以承受。

在实践中，这些限制往往导致采用子采样、近似相似性搜索，或依赖专用的高性能计算基础设施。相比之下，我们有意将设计目标设定为另一种方向：在不需要难以承受的内存开销的前提下，使基于轨迹的分析能够在常规硬件上高效、系统地进行穷举计算。要实现这一目标，必须谨慎设计算法，在保持精确性的同时，消除重叠时间窗口所导致的冗余计算。

现有异常检测研究涉及了这一问题的若干方面，但尚未直接解决在实际硬件限制下，如何在大陆尺度上兼顾高空间维度、数十年时间跨度以及基于轨迹的比较。为此，我们开发了一种针对网格化时空数据的高效且节省内存的算法。

本文的贡献如下：

– 我们提出一种通用的完全无监督框架，用于检测网格化时空数据中几何上稀有的短轨迹，且不依赖底层物理变量。– 我们推导出一种精确的基于递归的算法来计算轨迹距离，使计算成本不受轨迹长度影响。

– 我们提出一种高效的 TRAKNN 算法，可在配备 CPU 的标准工作站上进行跨数十年的穷举分析，并提供 CPU+GPU 版本。– 我们以 75 年的欧洲每日海平面气压数据为例展示该方法，结果表明，稀有轨迹呈现连贯的大气异常，且与独立的极端事件数据库相符。

这些结果共同表明，几何意义上精确的稀有轨迹对应着有意义的天气异常；在大陆尺度和跨数十年的时间尺度上估算稀有程度在计算上是可行的，并能为极端大气行为提供可解释的洞见。

本文其余部分安排如下。第2节介绍相关工作。随后在第3节描述问题及我们的算法。第4节评估算法的可扩展性。第5节通过一个真实案例研究展示该方法。第6节讨论更广泛的应用，并在第7节作出总结。

2 相关工作

时空数据集中的稀有模式检测处于气候科学、机器学习和异常检测的交汇处。在大气科学中，极端事件通常通过基于影响的阈值或基于环流的诊断方法来识别。许多研究采用流场类比方法 [20,28,7], ，其中类比是指历史上在欧几里得距离下大尺度场与目标日最为相似的日期。这类方法主要关注瞬时大气状态，未明确考虑环流模式随时间的演变。在现有类比与相似性搜索研究的基础上，我们探讨基于距离的方法在检索几何意义上稀有轨迹方面的潜力。

从数据挖掘的角度来看，稀有轨迹检测可表述为高维时间序列数据中的无监督异常检测。经典的基于距离的方法，例如 k 最近邻（kNN） [27,3],、局部离群因子 [5], 和孤立森林 [23], 通过特征空间中的几何隔离来定义异常。这些方法所需的建模假设极少且具有可解释性，因此颇具吸引力；但由于穷举相似性搜索的成本高昂，将其应用于跨越数十年的时空数据集在计算上颇具挑战。

现代相似性搜索库（如 FAISS [21,9] ）支持大规模近邻检索，通常会使用近似索引结构或专用计算硬件。然而，它们未能利用滑动窗口轨迹构建所产生的强时间重叠，因此直接应用于连续子序列时会导致重复计算。

时间序列数据挖掘已应对与子序列相似性搜索和模式发现相关的挑战。矩阵剖面（MP）框架 [34] 通过基于递推的距离更新，为单变量时间序列引入了精确且可扩展的全对相似性搜索。后续研究将这一思路扩展到

多维和流式场景。然而，这些方法并未直接解决时空数据带来的挑战。

在高维空间中应用基于距离的方法时，一个关键问题是距离集中现象 [4,1]。然而，理论分析表明，当数据位于低维结构上或具有强依赖性时，最近邻方法仍然具有意义 [10]。气候再分析场具有显著的空间相关性和较低的内在维数，经验正交函数分析也证实了这一点 [15]。更一般地说，内在维数估计方法 [22]提供了一种工具，可用于量化高维观测数据背后的有效自由度。

深度学习的最新进展也已应用于时空异常检测 [33,6], ，包括卷积自编码器、变分自编码器和循环神经网络。尽管这些方法能够捕捉复杂的非线性结构，但通常需要大量训练数据、模型调优和计算资源。已有多项研究探讨了利用深度学习方法检测极端天气事件 [24,25,32]。总而言之，现有方法要么 (i) 关注瞬时状态，要么 (ii) 依赖近似相似性搜索而未考虑时间重叠，要么 (iii) 需要复杂的参数化建模。相比之下，我们的工作旨在针对网格化时空数据集，开展精确、具备轨迹感知能力且计算高效的 k基于 NN 的罕见性检测，无需强大的硬件支持。

此外，我们的研究目标并非直接检测极端天气事件，而是检测稀有轨迹，并评估：1）这些轨迹是否对应有意义的物理模式；2）它们是否与极端事件相关联。

3 方法论

3.1 问题概述

我们要解决的核心问题是，如何在大规模网格化气候数据中系统地识别异常的大气演变。设想一组涵盖数十年、记录欧洲每日海平面气压图的序列。诸如风暴或热浪这样的极端天气事件，很少只是某个单一时刻的快照，而通常是持续数天的演变过程。例如，温带气旋通常会经历三至七天的生成、发展和消散过程。在此过程中，气压的空间模式会以连贯但可能稀有的方式演变。

我们的目标是在不依赖预先定义的阈值、标注事件或特定领域物理指标的情况下，检测此类罕见时间演变。我们采用纯粹由数据驱动的几何视角：将形状为 $h \times w$ 的连续逐日地图序列视为高维空间中的一条轨迹 。如果一条轨迹位于该空间中远离所有其他轨迹的区域，即在几何上孤立，则认为它是罕见的。我们通过它与 k 个最近邻之间的平均距离来量化这种孤立程度：平均最近邻距离较大的轨迹被视为孤立，因而也被视为罕见。

这一表述直观易懂：如果某种大气演变不同寻常，那么它在历史记录中就应该鲜有相近的对应情况。难点在于，如何让这一思路在包含数万个时间步和数千个空间网格点的数据集上具备计算可行性。本节余下部分将对问题进行形式化，并介绍一种高效算法，使得在标准硬件上进行穷尽轨迹分析成为现实。

3.2 问题定义

我们致力于在高维时空数据集中识别罕见时间演变。设$\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$为一个由空间场构成的离散时间序列，其中每个空间场 $X_t \in \mathbb{R}^{h\times w}$都是一个物理或抽象变量的矩阵（例如气压、温度或像素强度），在时间 t观测得到。

我们将持续时间为 d 、起始时间为 t 的**时空轨迹** $T_t^{(d)}$ 定义为由 d 个连续空间场组成的有序序列：

$$T_t^{(d)} = (X_t, X_{t+1}, \ldots, X_{t+d-1}) \in \mathbb{R}^{d\times h\times w} \tag{1}$$

轨迹是一个三阶张量，表示系统在时间区间 $[t, t+d-1]$内于高维构型空间中的路径。

为了量化两条轨迹 $T_i^{(d)}$ 和 $T_j^{(d)}$ 之间的相似性，我们采用平方欧氏距离 $\mathcal{D}$，该距离由两条轨迹之差的弗罗贝尼乌斯范数定义。它刻画了空间与时间维度上的累积逐点方差：

$$\mathcal{D}(T_i^{(d)}, T_j^{(d)}) = \|T_i^{(d)} - T_j^{(d)}\|_F^2 = \sum_{m=0}^{d-1} \|X_{i+m} - X_{j+m}\|_F^2 \tag{2}$$

在许多现实系统中，时间持续性会导致一种平凡匹配效应：时间上重叠的轨迹看起来高度相似。为确保发现非平凡孤立性，我们定义一个**排除区** e。只有当 $|i-j| > e$时，轨迹 T_i和 T_j才被视为候选邻居。

目标是计算**稀有度得分** s_t，其定义为 $T_t^{(d)}$ 在所有非重叠轨迹集合中的 k 个最近邻（kNN）的平均距离。

$$s_t = \frac{1}{k} \sum_{j\in\mathcal{N}_t} \mathcal{D}(T_t^{(d)}, T_j^{(d)}), \tag{3}$$

其中， $\mathcal{N}_t$ 是表示 $T_t^{(d)}$ 的 K 近邻的索引集合。

3.3 轨迹感知 KNN

我们提出了**轨迹感知 KNN （TRAKNN）**算法，如算法1所述。该算法分为两部分：高效计算所有空间场之间的成对相似度，以及通过基于递推的优化计算所有轨迹之间的距离，从而将计算复杂度与轨迹持续时间 d 解耦。

算法1： TRAKNN 算法

输入： 场序列 $\mathbf{X} \in \mathbb{R}^{h\times w\times n}$，持续时间 d，排除区 e，
邻居 k，批大小 b
输出： 稀有度分数 $\mathbf{s} \in \mathbb{R}^{n-d+1}$
// 1. 对称分批通用矩阵乘法空间距离计算
为 $i = 1 \dots n$预计算平方范数$N_i = \|X_i\|_F^2$
对批索引 $i \leftarrow 1$、$b+1$、……、n **执行**
  对批索引 $j \leftarrow i$、$i+b$、……、n **执行**
    $G \leftarrow B_i^\top B_j$ // 仅计算上三角
    对局部索引 u ，v 分批**执行**
      $S_{u,v} \leftarrow N_u + N_v - 2G_{u,v}$
      $S_{v,u} \leftarrow S_{u,v}$ // 利用对称性
// 2. 初始化与递推
对 $j \leftarrow 1$ **到** $n-d+1$ **执行**
  $D_{1,j} \leftarrow \sum_{m=0}^{d-1} S_{1+m,j+m}$
对 $t \leftarrow 2$ **到** $n-d+1$ **执行**
  // 向量化更新
  $D_{t,\cdot} \leftarrow D_{t-1,\text{prev}} - S_{t-1,\text{prev}} + S_{t+d-1,\text{curr}}$
  // 3. k近邻稀有度估计
  $\mathcal{N}_t \leftarrow$ 中 D_t 里 $|t-j| > e$最小的 k 个值的索引
  $s_t \leftarrow \frac{1}{k}\sum_{j\in\mathcal{N}_t} D_{t,j}$

优化空间距离计算 TRAKNN 算法首先预计算空间平方距离矩阵 $S \in \mathbb{R}^{n\times n}$。对高分辨率空间场之间的成对空间距离进行重复计算，构成了主要的计算瓶颈。我们将两条轨迹之间的平方距离计算重新表述为批量通用矩阵-矩阵乘法（GeMM）操作，以最大限度地提高计算强度。

设 $X_i, X_j \in \mathbb{R}^{h\times w}$ 为两个空间场。它们之间的平方欧氏距离计算如下：

$$\|X_i - X_j\|_F^2 = \|X_i\|_F^2 + \|X_j\|_F^2 - 2\langle X_i, X_j\rangle_F, \tag{4}$$

其中

$$\langle X_i, X_j\rangle_F = \sum_{p=1}^{h}\sum_{q=1}^{w}(X_i)_{p,q}(X_j)_{p,q} \quad \text{and} \quad \|X\|_F^2 = \langle X, X\rangle$$

空间场 $\|X_i\|_F^2$ 的范数会预先计算并缓存，从而将后续距离评估简化为内积计算和向量加法。通过将每个空间场 X_i 重塑为带有 $m = h\cdot w$的向量 $x_i \in \mathbb{R}^m$，我们将多个空间场汇总为批量稠密矩阵 B_1，$B_2 \in \mathbb{R}^{m\times b}$。随后通过一次三级BLAS操作$G = B_1^\top B_2$即可获得点积批次。由于空间平方距离矩阵是对称的（$S_{i,\ j} = S_{j,\ i}$），

我们只计算上三角批次的 G，从而将所需的 GeMM 运算数量有效减半，并最大限度地提高现代硬件上的浮点运算利用率。此步骤对应算法1的第1至7行。

常数时间轨迹距离递推 计算出 S 后，TRAKNN 会计算轨迹距离。由于时间上连续的两条轨迹 $T_i^{(d)}$ 和 $T_{i-1}^{(d)}$ 共享 $d-1$ 个空间场，因此任意一对（i, j）的平方距离都可在常数时间内 $\mathcal{O}(1)$ 更新：

$$\mathcal{D}(T_i^{(d)}, T_j^{(d)}) = \mathcal{D}(T_{i-1}^{(d)}, T_{j-1}^{(d)}) - S_{i-1,j-1} + S_{i+d-1,j+d-1} \tag{5}$$

必须在第 8-9 行显式初始化此递推，然后利用第 11 行的向量化运算，按行增量计算。

复杂度分析 TRAKNN 的计算复杂度主要由初始空间距离矩阵的计算决定，其规模为 $\mathcal{O}(hw \cdot n^2)$。使用公式 5（第 6 行）计算所有轨迹距离的开销为 $\mathcal{O}(n^2)$。初始化递推的开销为 $\mathcal{O}(d \cdot n)$。第 11 行检索前 k 个最近邻的开销为 $\mathcal{O}(n \log k)$。因此，TRAKNN 的计算复杂度为 $\mathcal{O}(hw \cdot n^2)$。在内存方面，该算法需要存储 $n \times n$空间距离矩阵 S、一个长度为 n 的预计算范数向量，以及两个长度为 $(n-d+1)$ 的向量，分别用于存储轨迹距离和分数。由于算法还会存储形状为 $h \times w \times n$ 的时空数据，因此内存占用为 $\mathcal{O}(\max(hw,\ n) \cdot n)$。

4 性能评估

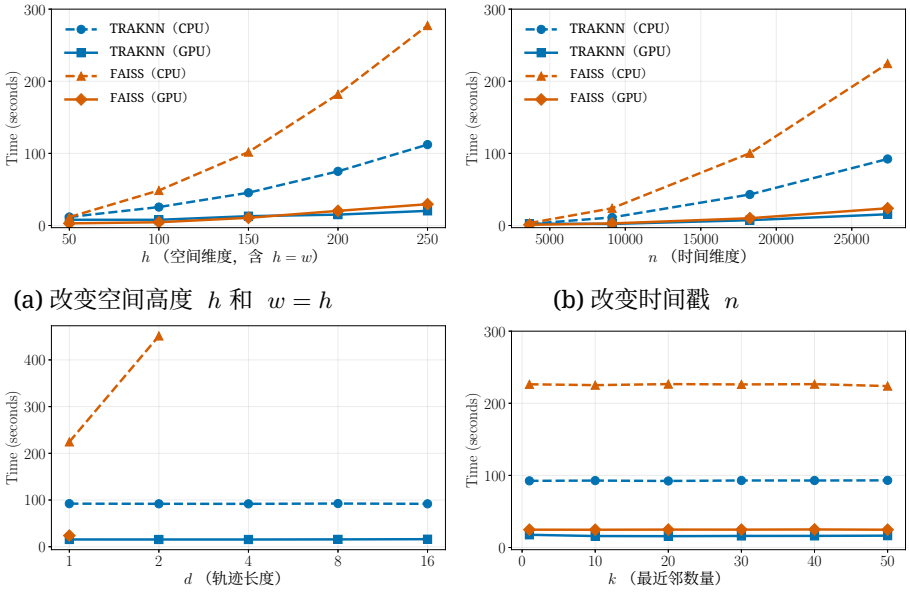
实验设置 代码使用 Python 3.11 和 PyTorch 2.5 编写，使算法能够在中央处理器和图形处理器上无缝执行。对于 GPU 版本，仅空间矩阵的分批计算在 GPU 上完成，矩阵则存储在 RAM 中。算法代码及全部实验可在以下 GitHub代码仓库[5]获取。实验在一台配备英特尔中央处理器（i9-13900H@2.6GHz）、32 GB RAM 以及英伟达图形处理器（RTX 2000 Ada，8GB）的笔记本电脑上进行。

在合成随机时空数据上评估该算法的执行时间和内存使用量。作为基线，我们将计算时间与 FAISS算法 [9,21], 进行比较，后者是一个支持在中央处理器和图形处理器上进行相似性搜索的先进库。

执行时间图1突出展示了TRAKNN在空间维度 $h \times w$（图1a）、时间步数量 n（图1b）、轨迹持续时间 d （图1c）以及邻居数量 k（图1d）方面的强可扩展性。对于长度为1的轨迹，TRAKNN在GPU上的性能与FAISS相当，而在CPU上则略优于FAISS。更重要的是，

[5] https://github.com/GuillaumeCld/轨迹k近邻

TRAKNN在轨迹长度增加时几乎保持恒定的计算时间，展现了其在大规模轨迹稀有性分析中的效率。相比之下，FAISS需要构建并存储数据库，这会带来显著的内存开销，并使其无法扩展到更长的轨迹。



(a) 改变空间高度 h 和 $w = h$

(b) 改变时间戳 n

(c) 改变轨迹长度 d

(d) 改变邻居数量 k

图 1：TRAKNN 与 FAISS 在 CPU 和 GPU 上的执行时间（秒），分别随 (a) 空间维度长度、(b) 时间步数量、(c) 轨迹长度和 (d) 邻居数量变化。默认参数为 $n = 27{,}000$, $h = 180, w = 280$, $d = 1$ 和 $k = 1$。

内存使用量 图 2 展示了所有算法在具有 75 年每日观测的180×280 网格上的峰值 RAM 使用量，对应于 $n \approx 27,000$、$k = 10$和 $d \in \{1, 5\}$。结果表明，TRAKNN 的内存使用量为 6 GB，与 d无关；而 FAISS 需要更多内存，$d = 1$ 需要 10 GB，并且内存使用量随 d呈线性增长。对于 $d = 2$ FAISS，需要超过 20 GB 的内存，无法在 GPU 上使用。

5 案例研究

我们以1950年至2025年欧洲（30°N $-$75°N，30°W $-$40°E）的平均海平面气压（SLP）数据为例，展示该方法的适用性。这些数据对应分辨率为 0.25° 的网格地图，对应的图像为

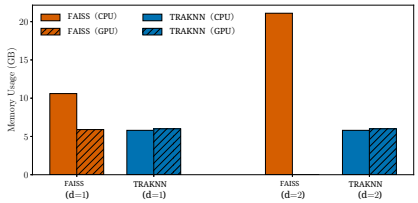


图 2：各算法的 RAM 峰值内存使用量

分辨率为 181×281，即约50,000个空间点；75年来的每日观测对应约27，000个时间点。总计，我们处理了超过十亿个数据点。数据可直接从C3S气候数据存储库 [16]下载。

我们的实验旨在回答以下三个问题：1）欧几里得距离对于高维轨迹是否仍有意义？2）几何意义上的稀有轨迹是否对应有意义的大气异常？3）稀有轨迹是否可能对应极端天气事件?

5.1 与降维方法的比较

内在维数研究 为更好地理解高维大气轨迹的几何结构，我们估计轨迹空间的内在维数（ID）随轨迹持续时间 d变化的情况，即表示数据所需的最少变量数。尽管每条长度为 d 的轨迹都位于维数为 $d \times h \times w$的环境空间中，但由于存在强烈的时空相关性，其有效自由度可能低得多。因此，我们使用最大似然 k NN估计器 [22], 计算局部 ID 的平均值，该估计器能够在局部量化 ID。鉴于此估计器取决于参数 k ，且 TRAKNN 能够高效检索 k 个邻居，我们使用 $k \in \{20, 30, 40\}$应用该估计器。为降低 k选择所带来的影响，我们报告局部 ID 的宏平均值和标准差。本分析有两个目的：（i）评估尽管环境维数较高，欧几里得最近邻距离是否仍有意义；（ii）刻画随着时间嵌入长度增加，动力学复杂度如何变化。表 1 所示结果表明，ID 比环境维数低两个数量级，且呈次线性增长并明显趋于饱和：从单日场的8 ± 7 增加到持续时间为五天或更长的轨迹的约 18 ± 5 。值得注意的是，超过五天后，额外的时间背景仅使 ID 略有增加，这表明多日大气演变仍局限于结构化的低维流形中。

PCA 对检测的影响 我们现在研究，在空间维度降低后的数据上应用 TRAKNN，是否会影响对最罕见轨迹的检测。为降低空间维度，我们使用 PCA 低秩近似

表 1：平均海平面气压下，环境空间维数（$d \times h \times w$）和平均局部内在维数（ID）随轨迹持续时间的变化对比

d	环境空间维数	ID
1	50,000	8 ± 7
3	150,000	16 ± 6
5	250,000	18 ± 5
7	350,000	19 ± 4
15	750,000	21 ± 3

该方法来自 PyTorch [13], ，它通过随机奇异值分解来近似矩阵的较低秩q 。我们设置 $q = 100$，然后保留能够解释 99% 方差的成分。这样可将维度从约 50,000 降至 33。我们使用 $d = 5$ 和 $k = 10$应用 TRAKNN。在降维前后检测到的得分最高的前 100 个日期中，有 99 个相同。共同日期之间的斯皮尔曼秩相关系数为 0.99，表明其秩排序几乎相同。保留 95% 的解释方差会将维度降至 13，得到 93 个相同日期，秩相关系数为 0.96。

这一结果与此前关于内在维数的结果相呼应，表明尽管轨迹维数很高，使用欧几里得距离估算轨迹之间的距离仍然是恰当的。

5.2 合成分析

为评估检测到的轨迹在物理上的一致性，我们对使用 $k = 10$、长度为 $d = 5$ 天的轨迹中识别出的最罕见的 100 条大气轨迹进行合成分析。我们旨在考察高稀有度轨迹是否对应结构清晰且可解释的大尺度环流模式，而非孤立的噪声驱动构型。为进行合成分析，我们采用大气分析中的标准流程，并在下一段中加以说明。

对于每条轨迹，我们考察相对于长期气候态的异常场，并将每个空间场向量化为维度为 $h \times w$的状态向量，该维度对应于空间域中的网格点数量。为考虑网格单元的实际面积向两极递减这一情况，我们根据纬度的余弦值对每个网格点进行加权。这种面积加权可确保后续分析反映具有物理意义的空间结构，不会受到高纬度地区网格密度较高的影响。随后，对加权异常向量进行标准化：在每个网格点上减去均值并除以标准差。这一步确保所有空间位置对分析的贡献大致相当，不受其固有变率的影响。然后，我们使用低秩主成分降低维度，并保留 95% 的解释方差。接着，我们在降维后的主成分分析

空间中使用平方欧氏距离的 k-均值算法，对相似日期进行重新分组。该过程将稀有轨迹归为少数几个连贯的动力型。对于每个聚类，通过对分配到该聚类的所有日期对应的原始异常场取平均值，计算合成异常图。

图3所示的合成结果揭示了持续存在的大尺度结构，其特征为强气压梯度和空间上有组织的异常型。特别是，我们发现三个相关聚类，聚类规模为 s。第一个聚类（$s = 25$）在北欧显示出较高的正气压异常，而在区域南部则显示出较低的负异常。第二个聚类（$s = 41$）与第一个相反，北部为较高的负异常，下部为较低的正异常。第三个聚类（$s = 34$）对应于覆盖几乎整个欧洲的大范围低压型（负异常）。

本分析证实，排名最高的罕见轨迹对应的是动态连贯且具有物理意义的环流演变，而非随机离群点。

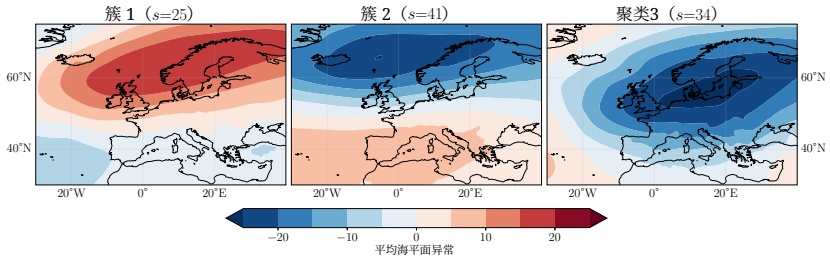


图 3：对平均海平面气压异常的三个主要簇进行合成分析

5.3 与极端事件匹配

现在，我们将探究排名前 100 的轨迹是否对应已知的极端天气事件，尤其是风暴。

数据库说明 我们使用三个公开且相互独立的数据库：XWS [29], CLIMK-WINDS [11],和EM-DAT [8]。

XWS数据库由1979年至2013年间50场最极端的欧洲冬季风暴组成，这些风暴由基于事件物理特征的气象指数确定。分析在以下区域进行： $35°\text{N} - 70°\text{N}$、 $15°\text{W} - 25°\text{E}$；在检测到的5730场风暴中，最极端的50场被保留在目录中。检测到的风暴分为两类：保险风暴，由23场造成高额保险损失的风暴组成；以及大型风暴，包括选出的27场风暴

即根据其在风暴严重程度指数中的排名位列前列的“非保险”风暴。

CLIMK–WINDS 数据库由 1995 年至 2015 年间，在以下空间域27°N $-$ 72°N、 22°W $-$ 45°E 内检测到的 50 场最严重的欧洲冬季风暴组成，这些风暴是通过气象指数检测得出的。

EM-DAT 数据库是一个收录自然与技术灾害的全球数据库；具体而言，其中包含风暴、热浪和寒潮等极端气象事件。在这项研究中，我们仅收集欧洲极端事件，并在 1950 年至 2025 年的完整空间域上计算稀有度得分。

匹配方法论 为进行公平比较，我们研究各目录在相同空间域和相同日期范围内的罕见海平面气压轨迹。对于每项分析，我们保留稀有度得分最高的前 100 个日期；若某个事件属于一条罕见轨迹，即发生在该日期至该日期加上 d 天的时间段内，则视为检测到该事件，这段时间对应轨迹持续时间。实验中我们始终设置 $k = 10$ 。

结果 表2总结了在三个数据库：XWS、$CLIMK–WINDS$和$EM\text{-}DAT$中，不同轨迹长度对应的风暴日期数量。结果突显出，检测结果会随轨迹持续时间以及数据库性质（气象类与影响类）而变化。

随着轨迹长度 d 增加，识别出的天数总体呈增加趋势，这一趋势在气象目录中尤为明显。

对于 CLIMK–WINDS，覆盖风暴的前100个日期中，被检测到的日期数量从6个（1天轨迹）稳步增加到12个（7天轨迹）。同样，对于XWS大型风暴，该数量从3个（1天）增加到7个（7天）。这种单调或近似单调的增长表明，较长的轨迹能够更好地捕捉严重冬季大风风暴的时间范围。由于此类系统通常会持续演变数天，延长 d 会增加至少一个罕见海平面气压配置与目录中记录的事件发生重叠的概率。然而，增加d 会减少与保险风暴对应的天数。XWS中保险风暴与大型 风暴之间的差异，可能反映了基于影响的严重程度排名与纯气象严重程度排名之间的区别。保险风暴可能具有更局地性的特征，或与特定的高损失日相关，这解释了它们在较短持续时间下更高的代表性。

相比之下，一些类别呈现非单调变化。例如，EM-DAT 一般风暴的检测数量从 $d = 1$ 时的 8 次降至 $d = 3$ 时的 5 次，随后在持续时间更长时又再次增加。这表明，增加轨迹持续时间并不能在所有数据集中都一致地提升检测效果。

作为对比，随机选取 100 个日期，运行 100 次后平均可发现 0.7 个 CLIMK-WINDS 风暴和 0.4 个 XWS 风暴（大型 $+$ 保险）。

此外，与 EM-DAT 报告的极端温度相比，瞬时维数（$d = 1$）显示，在 100 个日期中，有 9 个与严寒冬季条件重合，17 个与寒潮重合。随着 d 增大，

表2：TRAKNN按数据集和风暴亚型统计的与风暴相关的已识别日期数量，该数量取决于轨迹持续时间 d （以天为单位），并与平均随机估计器进行比较。

数据集	风暴类型	d	随机
		1 3 5 7	
CLIMK-WINDS –		6 9 11 12	0.7
XWS	保险风暴	6 5 2 4	0.2
	大型风暴	3 4 5 7	0.2
EM-DAT	风暴（一般）	8 5 7 4	0.9
	风暴（热带）	0 0 1 1	0.1
	风暴（温带） 3 2 3 3		0.4

与极端温度的对应关系显著减弱。例如，在 $d = 5$ 时，100 个日期中没有一个对应严寒冬季条件，只有 7 个对应寒潮。

启示 总体而言，结果表明：(i) 较长的轨迹有助于提升对大尺度冬季大风风暴的检测效果，尤其是在气象目录中；(ii) 与 EM-DAT 等多灾种灾害数据库相比，基于气象严重程度指数构建的目录（CLIMK–WINDS 和 XWS）中的检测结果更加一致；(iii) 对于极端温度事件，空间地图快照更具相关性

6 讨论

实践见解 TRAKNN 可在几分钟内，利用标准工作站对跨数十年的大陆尺度数据集进行穷尽式分析。这种可扩展性使从业者能够快速开展“假设分析”，例如改变轨迹长度或测试不同变量；由于精确相似性搜索的计算复杂度呈二次增长，以往这类分析难以实现。此外，该方法能够直接处理任意轨迹长度的格点数据，无需降维或复杂的特征工程，为历史相似个例检索和极端事件归因提供了一种透明且易于使用的工具。

基于 kNN 的算法中，一个关键考量是如何选择 k。这种方法的一个优势是， k 的取值对计算成本影响极小。此外，较大的 k 值所得到的前 k 个结果，自然包含任何较小 k 值所对应的前 k 个结果。因此，只需少量额外的后处理开销，即可针对多个 k 值评估稀有度得分，无需从头重新计算整个过程。这使用户能够以类似集成分析的方式，高效评估不同 k 值对得分以及识别最稀有轨迹的影响。

尽管内存占用看起来可能相当大，但在实际的类比大气研究中，可用记录通常至多涵盖 ≈ 75 年的逐日数据，而且空间域受到所研究现象物理尺度的限制。因此，我们案例研究中考虑的问题规模反映了现实应用中合理的上限，而非对方法的人为限制。

与矩阵剖面的关系所提出的基于递推的优化，从本质上说，是将 MP 形式体系扩展至时空数据。通过利用式 5 中的递推关系，我们采用了类似于 MP高效算法 [34,2], 的更新逻辑，但将其适用于任意维度的序列，而非标量值。因此，在 $k = 1$这一特殊情况下，我们的稀有度得分 s_t 即为一种时空矩阵轮廓，将全对相似性搜索理念扩展至时空场，并为识别复杂物理系统中的非平凡模式和异常提供了稳健的基础。

7 结论

我们提出了 TRAKNN，旨在弥补气候科学中的一项空白，使分析重点能够从瞬时大气快照转向任意持续时间的空间场的时间演变。未来，如果实际应用有所需要，可以研究外存计算，以应对大型空间或时间域：这些数据域可能无法装入内存，并且可能造成严重的数据加载瓶颈。此外，未来还将研究在流场类比分析中纳入基于轨迹的类比。

致谢

我们感谢奥克西塔尼大区的资助，以及欧盟“地平线2020”研究与创新计划（资助协议号101003469，XAIDA）、ANR项目TASE PowDev（未来电网战略发展，ANR-22-PETA-0016）、ANR项目Templex（ANR-23-CE56-0002）和ANR项目TRACCS“为气候服务变革气候建模”（ANR-22-EXTR-0005，Extending）的资助。

参考文献

1. Charu C. Aggarwal、Alexander Hinneburg 和 Daniel A. Keim。《高维空间中距离度量的意外表现》。载于 Jan Van den Bussche 和 Victor Vianu 编，《数据库理论 — ICDT 2001 》，第 420–434 页。Springer。2. Reza Akbarinia 和 Bertrand Cloez。使用不同距离函数高效计算矩阵剖面。arXiv 预印本 arXiv:1901.05708，2019。

3. Fabrizio Angiulli 和 Clara Pizzuti。高维空间中的快速离群点检测。载于《欧洲数据挖掘与知识发现原理会议》，第 15–27 页。Springer，2002。4. Kevin Beyer、Jonathan Goldstein、Raghu Ramakrishnan 和 Uri Shaft。“最近邻”何时有意义？载于 Catriel Beeri 和 Peter Buneman 编，《数据库理论 — ICDT’99》，第 217–235 页。Springer。5. Markus M. Breunig、Hans-Peter Kriegel、Raymond T. Ng 和 Jörg Sander。LOF：识别基于密度的局部离群点。载于《ACM SIGMOD 数据管理国际会议论文集》，第 93–104 页。ACM，2000。6. Raghavendra Chalapathy 和 Sanjay Chawla。用于异常检测的深度学习：综述。arXiv 预印本 arXiv:1901.03407，2019。7. Soledad Collazo、Solange Suli、Pablo G Zaninelli、Ricardo García-Herrera、David Barriopedro 和 José M Garrido-Perez。大尺度环流和局地反馈对阿根廷 2022/23 年极端夏季高温的影响。《Communications Earth & Environment》，5(1):231，2024。8. Damien Delforge、Valentin Wathelet、Regina Below、Cinzia Lanfredi Sofia、Margo Tonnelier、Joris A. F. van Loenhout 和 Niko Speybroek。EM-DAT 紧急事件数据库档案。9. Matthijs Douze、Alexandr Guzhva、Chengqi Deng、Jeff Johnson、Gergely Szilvasy、Pierre-Emmanuel Mazaré、Maria Lomeli、Lucas Hosseini 和 Hervé Jégou。FAISS 库。第 1–17 页。10. Robert J. Durrant 和 Ata Kabán。“最近邻”何时有意义：一条逆定理及其启示。25(4):385–397。11. Clare M. Flynn、Julia Moemken、Joaquim G. Pinto、Michael K. Schutte 和 Gabriele Messori。CLIMK–WINDS：欧洲极端冬季风暴的新数据库。17(9):4431–4453。12. Giovanni Forzieri、Alessandra Bianchi、Filipe Batista e Silva、Mario A Marin Herrera、Antoine Leblois、Carlo Lavalle、Jeroen CJH Aerts 和 Luc Feyen。气候极端事件对欧洲关键基础设施的影响日益加剧。《全球环境变化》，48:97–107，2018。13. N. Halko、P. G. Martinsson 和 J. A. Tropp。利用随机性发现结构：用于构造近似矩阵分解的概率算法。53(2):217–288。14. John Hanley 和 Rodrigo Caballero。大尺度大气流和罗斯贝波破碎在欧洲极端风暴演变中的作用。39(21)。15. Abdel Hannachi、Ian T Jolliffe、David B Stephenson 等。大气科学中的经验正交函数及相关技术：综述。《International Journal of Climatology》，27(9):1119–1152，2007。16. Hans Hersbach、Bill Bell、Paul Berrisford、Gionata Biavati、András Horányi、Joaquín Muñoz Sabater、Julien Nicolas、Carole Peubey、Raluca Radu、Iryna Rozum、Dinand Schepers、Adrian Simmons、Cornel Soci、Dick Dee 和 Jean-Noël Thépaut。ERA5 自 1940 年至今的单层逐小时数据。哥白尼气候变化服务（C3S）气候数据存储库（CDS），2023。doi: 10.24381/cds.adbb2d47（访问日期：2024-06-10）。17. Peter Hoeppe。天气相关灾害的趋势——对保险公司和社会的影响。11:70–79。18. Daniel E Horton、Nathaniel C Johnson、Deepti Singh、Daniel L Swain、Bala Rajaratnam 和 Noah S Diffenbaugh。大气环流模式变化对极端气温趋势的贡献。《Nature》，522(7557):465–469，2015。

19. Radley M. Horton、Justin S. Mankin、Corey Lesk、Ethan Coffel 和 Colin Raymond。《极端高温事件研究的最新进展综述》。2(4):242–259。20. Aglaé Jézéquel、Pascal Yiou 和 Sabine Radanovics。利用流场类比研究环流在欧洲热浪中的作用。《气候动力学》，50(3):1145–1159，2018。21. Jeff Johnson、Matthijs Douze 和 Hervé Jégou。利用图形处理器进行十亿级相似性搜索。7(3):535–547。22. Elizaveta Levina 和 Peter Bickel。内在维数的最大似然估计。载于神经信息处理系统进展，第17卷。麻省理工学院出版社。23. Fei Tony Liu、Kai Ming Ting 和周志华。孤立森林。载于*IEEE*国际数据挖掘会议（*ICDM*）论文集，第413–422页。IEEE，2008。24. Yunjie Liu、Evan Racah、Joaquin Correa、Amir Khosrowshahi、David Lavers、Kenneth Kunkel、Michael Wehner、William Collins等。将深度卷积神经网络应用于气候数据集中的极端天气检测。*arXiv*预印本 *arXiv:1605.01156*，2016。25. Mayur Mudigonda、Prabhat Ram、Karthik Kashinath、Evan Racah、Ankur Mahesh、Yunjie Liu、Christopher Beckham、Jim Biard、Thorsten Kurth、Sookyung Kim等。利用深度学习检测极端天气模式。《地球科学深度学习：遥感、气候科学与地球科学综合方法》，第161–185页，2021。26. 美国国家科学院、医学部、地球与生命研究部、大气科学委员会、极端天气事件委员会和气候变化归因委员会。《气候变化背景下的极端天气事件归因》。美国国家科学院出版社，2016。27. Sridhar Ramaswamy、Rajeev Rastogi 和 Kyuseok Shim。从大型数据集中挖掘离群点的高效算法。载于*2000*年*ACM SIGMOD*数据管理国际会议论文集，第427–438页，2000。28. Liwen Ren、Tianjun Zhou 和 Wenxia Zhang。2018年夏季东北亚破纪录高温事件的归因：环流的作用。《环境研究快报》，15(5):054018，2020。29. J. F. Roberts、A. J. Champion、L. C. Dawkins、K. I. Hodges、L. C. Shaffrey、D. B. Stephenson、M. A. Stringer、H. E. Thornton 和 B. D. Youngman。1979年至2012年欧洲极端风暴XWS开放获取目录。14(9):2487–2501。30. Peter Stott。气候变化如何影响极端天气事件。352(6293):1517–1518。31. Caroline C. Ummenhofer 和 Gerald A. Meehl。具有生态相关性的极端天气和气候事件：综述。372(1723):20160135。32. Shikha Verma、Kuldeep Srivastava、Akhilesh Tiwari 和 Shekhar Verma。极端天气事件中的深度学习技术：综述。*arXiv*预印本 *arXiv:2308.10995*，2023。33. Senzhang Wang、Jiannong Cao 和 S Yu Philip。用于时空数据挖掘的深度学习：综述。*IEEE*知识与数据工程汇刊，34(8):3681–3700，2020。34. Yan Zhu、Zachary Zimmerman、Nader Shakibay Senobari、Chin-Chia Michael Yeh、Gareth Funning、Abdullah Mueen、Philip Brisk 和 Eamonn Keogh。矩阵轮廓II：利用新算法和图形处理器，突破时间序列模式与连接的一亿规模壁垒。载于*2016*年*IEEE*第*16*届国际数据挖掘会议（*ICDM*），第739–748页。IEEE，2016。