互评作业 3：数据挖掘应用系统方案
——基于气象数据的极端天气预测系统

周明轩 3120240977

1. 场景描述

一个气象研究机构希望利用气象数据（如温度、降水量、风速等）开发一个极端天气预测系统，以便及时预警。随着全球气候变化的加剧，极端天气事件（如高温热浪、强降雨、寒潮、台风等）发生频率显著上升，给农业生产、城市运行、能源保障与人民生命财产安全带来了严峻挑战。及时、准确地预测极端天气事件已成为现代气象服务体系中的核心任务之一。传统气象预测方法主要依赖物理数值模型进行计算，但这些模型在应对突发、区域性强的极端天气时存在响应滞后、空间分辨率不足等问题。近年来，随着大数据与人工智能技术的发展，基于数据驱动的预测方法逐渐显示出在识别复杂气象模式、提前捕捉极端事件信号方面的优势。

本报告旨在提出一套完整的基于气象数据的数据挖掘应用系统方案，构建一个能够自动分析气象数据并进行极端天气预警的智能预测系统，服务于政府应急管理、农业防灾减灾、交通安全等多个领域。

2 数据收集

极端天气预测系统的基础是全面、准确、高时空分辨率的气象数据。系统的预测精度在很大程度上依赖于所采集数据的类型、质量与时效性。因此，本系统需从多个维度收集和整合气象数据，包括地面观测、遥感监测、历史再分析数据等。以下将详细说明所需收集的气

象数据类型及其在极端天气预测中的作用。

2.1 地面观测类数据

来自全国自动气象站、区域站网的地面观测数据具有高时效、高精度的特点，是极端天气识别和建模的核心数据源。主要包括：

- 气温：用于识别高温热浪、寒潮、昼夜温差剧变等事件。突变的温度升降是热浪和冷空气活动的重要先兆指标。
- 降水量：用于暴雨、强对流、短时强降水等事件的判断。连续高强度降水还可用于滑坡、山洪等次生灾害风险评估。
- 风速与风向：对台风、龙卷风、大风等极端天气事件尤为关键。风场突变和风切变特征是突发性强风的预测基础。
- 气压：气压骤降往往伴随气旋系统、台风、冷锋等极端系统活动，是分析天气演变的重要参数。
- 相对湿度：高湿环境下气温升高更易导致热应激，对湿热型热浪的识别极为关键。
- 能见度：对沙尘暴、大雾等能见度灾害有直接表征意义，适用于交通和航空预警场景。

2.2 遥感与雷达数据

为了弥补地面站点覆盖的空间不足，需要融合遥感卫星和雷达数据，以实现对高空、海洋和偏远区域的全天候监测：

- 气象卫星数据：提供云图、地表温度、云顶高度等信息，用于识别热对流发展、飓风/台风结构监测、雷暴活动识别等。
- 天气雷达数据：用于捕捉短时强对流天气的动态过程，包括

雷暴、冰雹、龙卷风等高影响事件的早期探测。

- 雷电定位系统数据：可实时记录雷暴发展过程，有助于识别强对流区的活跃程度及其潜在风险。

2.3 气候历史数据

历史数据的积累不仅用于模型训练，也可帮助识别极端事件的长期演化趋势和异常周期性：

- 再分析数据：提供多层次气象场变量（温度、风场、湿度、气压等）的长时间序列数据，适用于极端天气的回溯分析和模型初始化。
- 长期极端事件记录（极端温度、极端降水日数等）：用于构建事件样本库、训练分类预测模型，并辅助提取极端天气先兆特征。

3 数据预处理

气象数据具有时序性强、维度多、来源复杂、质量差异显著等特点，直接使用原始数据进行极端天气预测可能会导致模型性能下降甚至预测失败。因此，构建一个高性能的预测系统必须以高质量、结构化的数据为基础。本系统的数据预处理流程主要包括缺失值处理、异常检测与校正、数据平滑、时间序列对齐和标准化处理，以下将逐一说明关键步骤与方法。

3.1 缺失值处理

气象观测数据常因设备故障、通信中断、恶劣环境干扰等原因产生缺测，尤其在远程站点和卫星数据中更为普遍。对于时间序列建模

而言，缺失值若处理不当将显著影响模型学习连续性和时序特征。

常用缺失值处理策略包括：

- 线性插值：对于短时、连续的缺测区间，采用线性插值进行平滑填补，适用于气温、湿度等连续性强的变量。
- 时间窗口均值法：利用缺失点前后固定时间窗口内的均值填充，适合短期波动不大的时间序列特征（如风速、气压）。
- 空间插值：对于区域性缺测数据（如雷达图像缺块），根据临近站点或格点进行空间插值，恢复观测场的连续性。
- 机器学习填补：在多变量数据中，使用其他相关变量构建预测模型进行缺失填补，提升插补精度，适用于卫星遥感等高维数据场景。

缺失值比例较高的数据（例如连续缺测超过 30%）将被标记，并根据建模需求进行降权或剔除，以确保输入数据质量。

3.2 异常值检测与校正

气象观测中常出现物理不合理的观测值或“跳变点”，如气温瞬间下降 20℃、风速超过理论极值等。这些异常点若不加处理将极大干扰模型训练。

处理流程如下：

- **规则阈值法**：根据国家或行业气象数据质量标准设定物理上下限，超出范围的数据自动标记为异常。
- **滑动窗口 Z-score 检测**：在每个时间滑动窗口内计算均值和标准差，识别超过 3σ 的离群点。

- **趋势断点检测**：检测序列中趋势突变处，用于定位设备故障或突发异常。

异常值在确认后可根据插值或邻近时段回归替代，或直接剔除用于短期预测模型的训练。

3.3 数据平滑

气象数据中常存在由观测误差、局地扰动或测量噪声引起的高频抖动，这种波动不具备预警价值，反而会干扰极端趋势的提取。因此，在保持长期趋势与突变性的基础上，需对数据进行适度平滑。

常用方法包括：

- **移动平均**：对序列进行窗口平均，有效去除高频噪声，适合处理温度、风速、湿度等稳定性变量。
- **指数加权平滑**：对近期数据赋予更高权重，保留部分突变特征，适合对突发天气具有一定敏感性的变量。
- **小波变换/EMD 分解**：将序列分解为趋势项和噪声项，只保留低频信号输入模型，适用于高噪声的雷达/遥感数据。

需要注意的是，平滑处理需权衡信息保留与噪声去除之间的关系，避免过度平滑导致极端信号被掩盖。

3.4 时间序列对齐

由于数据来源多样，不同气象数据的采样频率、时间戳格式可能存在差异，导致模型无法直接进行多变量联合建模。因此，统一时间轴对齐是数据预处理的关键环节。

主要对齐策略如下：

- **时间标准化处理**：统一时间戳格式为 ISO 8601，消除时区、夏令时等问题。
- **重采样**：将高频数据（如分钟级风速）聚合到统一粒度（如小时或 3 小时）以便与其他数据对齐，可使用均值、最大值、累计值等策略。
- **插值对齐**：对于采样频率较低的变量（如卫星云图每小时一帧），可进行时间插值或帧间重建，提高与高频地面观测数据的对齐精度。
- **时间窗口滑动拼接**：以固定滑动窗口方式（如过去 6 小时→当前时刻）生成样本序列，满足时序建模输入需求。

该阶段的输出为统一时间轴、等长等频的多变量时间序列样本，为后续特征工程和建模打下基础。

4 特征工程

特征工程是极端天气预测系统中至关重要的环节，直接影响模型的表达能力和泛化性能。得益于高质量的预处理数据，本系统进一步通过合理构造时间序列特征、统计特征、衍生物理特征和空间相关特征，以提升模型对极端天气事件的敏感性和预测精度。

在本系统中，特征工程的目标不仅是压缩和规整高维气象数据，更重要的是提取对极端事件（如暴雨、寒潮、热浪、大风等）具备判别力的关键因素，从而增强模型对异常模式的学习能力。

4.1 时间序列特征提取

极端天气事件往往具有**时序演化特性**，如强降水前常伴有气压突

降和湿度骤升。因此，充分挖掘时间序列中的动态变化规律，对于提升系统的预警敏感性至关重要。

本系统提取的时序特征包括：

- **滞后特征**：如过去 1 小时、3 小时、6 小时的温度、风速、降水量等值。滞后特征帮助模型学习天气演变的惯性趋势，适用于 LSTM、Transformer 等时序模型。
- **变化率特征**：如温度变化速率（ΔT）、气压下降速率（ΔP），有助于识别突变信号，是暴风、雷暴等灾害前兆的重要指标。
- **滑动统计特征**：包括滑动平均、最大值、标准差、偏度、峰度等，提取局部波动模式，如风力波动强烈可能预示龙卷风来临。
- **周期性特征**：考虑日周期（昼夜变化）、年周期（季节变化）等信息，通过正余弦函数编码时间戳，用于建模气象变量的长期趋势。

4.2 多源气象变量构造特征

基于原始气象观测数据（如温度、湿度、气压、风速、风向、降水量、太阳辐射量等），我们构建出更具物理含义的组合特征与指数指标，用于提升模型对极端气象条件的识别能力。

典型构造特征包括：体感温度：综合温度、湿度和风速等因素计算，更能反映人体感知到的热冷程度，对热浪、寒潮等事件更敏感；气压梯度：在邻近区域或不同时间点间计算气压变化率，有助于捕捉锋面活动、台风路径等重大天气系统；湿球温度：结合温度与湿度，是判断中暑、高温风险的重要指标，常用于极端高温事件预测；风切

变强度：评估不同高度风速/风向变化程度，对预测雷暴、大风灾害尤为重要；雷暴指数：利用多气象变量计算的大气不稳定性指标，用于预估雷暴、冰雹等对流性天气。

4.3 特征选择与降维策略

为避免冗余特征带来噪声干扰，同时提升计算效率与模型泛化能力，系统采用以下特征筛选方法：

- 相关性分析：计算皮尔逊系数、互信息量等，剔除冗余或弱相关变量。
- 模型嵌入式选择：利用树模型（如随机森林、XGBoost）或注意力机制提取关键特征权重进行排序。
- 主成分分析与自编码器降维：用于处理高维遥感数据和格点特征，保留最主要的空间变化信息。

5 算法选择与模型训练

在极端天气预测任务中，算法的选择对模型性能和实用性起着决定性作用。由于气象数据天然具有多变量、多尺度、强时序性与空间相关性等复杂特征，传统的单变量预测方法在面对突发性强、演变快的极端天气事件时往往存在响应滞后和预测不稳定等问题。因此，本系统在算法选择上不仅需考虑预测准确性，还要关注多变量建模能力、非线性建模能力与长依赖捕捉能力。

本节将对常用的时间序列预测模型进行系统比较，结合实际气象数据特点选择适用模型，并制定对应的训练策略。

5.1 模型对比分析

模型类型	核心特点	优点	局限性
ARIMA	基于线性自回归+滑动平均+差分	模型结构清晰，适合短期平稳序列预测	仅适用于单变量、线性关系强的数据，对非平稳和非线性特征建模能力差
Prophet	面向商业时间序列，内置季节性、节假日影响建模	自动化强，易于部署，解释性好	对极端值建模能力弱，难以处理高频气象数据和异常波动
LSTM	长短期记忆神经网络，擅长建模长依赖关系	能处理非线性、多变量序列，适应复杂时间动态模式	训练时间长，对输入格式依赖较强，不擅长捕捉局部突变
Temporal CNN	一维卷积神经网络，建模局部时间依赖	参数更少，训练更快，适合高频数据预测	对长期依赖学习能力有限
Transformer	自注意力机制建模全局依赖	可并行训练，适合处理长序列，支持多变量输入	对小样本不敏感，训练资源开销大

5.2 模型选择与应用场景匹配

根据本项目对极端天气的预警需求，我们重点考虑以下三类场景：单站点、短时预测任务（如未来 1~6 小时降雨量预测）；区域性极端事件建模（如暴雨、热浪、大风的区域演变趋势）；长时序、高维输入（如融合遥感图像+气象多变量+历史灾害记录）预测大尺度极端天气事件。综合考虑上述情景，选择使用多变量 LSTM 模型 + 注意力机制增强。这一模型选择兼顾非线性建模能力与时间依赖学习能力；适配滑动窗口输入结构，适合逐小时预测；加入注意力机制后，可自适应聚焦对极端天气影响最大的特征变量或时间节点；易于结合格点数

据、遥感统计特征，实现区域预测扩展。

6 模型验证

模型验证是极端天气预测系统开发过程中的关键环节，旨在全面评估所选模型在不同时间尺度、空间区域及气象条件下的预测能力、稳健性和实用性。在极端天气预警这一高敏感度应用中，仅依赖单一指标可能掩盖模型在极端情况下的失效风险。因此，本系统采用多种评估指标相结合的方式，从误差精度、事件检测能力和极端值捕捉能力等多维度对模型性能进行全面验证。

6.1 常用回归评估指标

针对气象变量（如温度、降水量、风速等）预测值与真实观测值之间的误差评估，系统采用以下指标：

指标名称	计算公式
MAE（平均绝对误差）	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i-\hat{y}_i\rvert$
MSE（均方误差）	$\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2$
RMSE（均方根误差）	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}$
MRE（平均相对误差）	$\frac{1}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i-\hat{y}_i}{y_i}\right\rvert$

6.2 极端事件识别能力评估指标

考虑到本系统目标是“极端天气预警”，仅关注连续变量的误差不足以评估模型实际应用效果。我们引入事件识别类指标来专门评估模

型对极端事件（如强降雨、强风、极端高温）的判断能力：

指标名称	指标意义
命中率	在所有真实发生的极端事件中，有多少被模型成功识别。衡量系统对实际事件的敏感性。
精确率	模型预测为极端事件的时刻中，有多少是真正的极端事件。反映系统误报的频率。
F1 分数	精确率与召回率的调和平均，综合衡量检测准确性与完整性，适合极端事件稀疏场景。

7 模型调优

模型调优是极端天气预测系统建模流程中的核心环节，直接决定模型在实际应用中能否实现高精度预测与稳定泛化的能力。由于气象数据具有高噪声性、强非线性、季节性强烈等特点，调优过程不仅需要优化传统的超参数组合，还需引入提升模型泛化与鲁棒性的策略，从而增强其对极端事件的感知与学习能力。

7.1 超参数优化方法

手动调参

- 初期模型探索阶段，结合业务理解与模型反馈，人工尝试调节学习率、隐藏层大小、时间步长等关键参数；
- 优点是直观、便于快速验证假设，但效率低、主观性强。

网格搜索

- 在预定义的参数组合空间内穷举搜索最佳超参数组合；
- 适合参数空间不大、模型训练成本可接受的情形；
- 可用于传统模型（如 ARIMA）中对滞后阶数（p,d,q）或季节性

周期的组合优化。

随机搜索

- 在定义的参数范围内随机采样，效率更高，适合高维参数空间；
- 对 LSTM、GRU 等深度模型中大量参数组合（如隐藏单元数、dropout 率、batch size）具有良好效果。

贝叶斯优化

- 基于先前的试验结果构建代理函数，智能地引导搜索过程；
- 能在有限计算资源下迅速收敛到高性能区域；
- 推荐使用开源工具如 Optuna、Hyperopt、Ray Tune 实现深度模型调参。

自动机器学习（AutoML）框架

- 如 Google AutoML、AutoKeras、AutoGluon，可自动完成特征选择、模型搜索与参数调优；
- 在气象预测中适合进行大规模模型基线比较与自动化部署探索。

7.2 增强学习能力的训练策略

数据增强与扰动训练

- 通过对历史气象数据添加微弱噪声或进行平滑扰动，增强模型对小幅波动和测量误差的鲁棒性；
- 对极端天气样本可通过加权复制（oversampling）方式增强训练频次，避免模型在样本不平衡情况下的偏差。

正则化技术

- 引入 L1/L2 正则化、Dropout 等方式防止过拟合；

- 在 LSTM 等循环神经网络中结合 Recurrent Dropout 技术，防止时间步长增加导致的冗余记忆问题。

多任务学习（Multi-task Learning）

- 通过同时预测多个相关变量（如温度、降雨概率、风速等级），实现特征共享，提升模型对气象模式的整体理解能力；
- 对提升极端天气检测与常规天气预报的协调性具有重要价值。

自注意力机制与时间动态建模

- 在深度模型结构中引入自注意力模块（如 Transformer、Informer），以捕捉长时间依赖关系，增强模型对渐进式极端事件（如热浪、寒潮）的提前预判能力。

在线学习与迁移学习

- 将历史训练模型参数作为初始化，对不同区域/季节/年份进行微调迁移，提升新环境下的学习效率；
- 在线学习机制可实现模型在实时反馈下的快速迭代更新，应对天气演变趋势的结构性变化。

8 部署与监控

为了实现极端天气预测系统在实际气象业务中的落地应用，必须将模型部署在稳定可靠、具备高可用性与实时响应能力的系统环境中，并建立完善的模型监控与更新机制。本部分从模型部署架构设计以及监控与自适应更新机制等方面，详细阐述如何构建一个具备业务实用价值的极端天气预警平台。

8.1 系统集成与部署架构

系统应采用模块化微服务架构，包括：数据接入服务、模型推理服务、结果发布服务、预警触发机制、模型更新调度器等；使用容器化技术（如 Docker）封装模型服务，结合 Kubernetes 实现高并发、高可用部署与弹性伸缩；接入 GPU 推理服务（如 NVIDIA Triton）提升深度模型（如 LSTM、Transformer）推理效率。

与气象数据平台集成,接入国家或地方气象局提供的实时数据 API，支持多源数据同步接入（地面站、卫星、雷达）；实时拉取并存储数据至分布式数据仓库（如 Apache Hive、ClickHouse）或消息队列（如 Kafka）用于后续推理处理。

将预测模型封装为标准化接口（如 RESTful API 或 gRPC 服务），部署于推理服务器；支持批量预测和滑动窗口预测模式，满足不同时间尺度（如小时级、日级）预报需求；输出包括未来温度、降水概率、风速变化趋势及极端事件风险等级等指标。

8.2 系统集成与部署架构

在极端天气预测系统中，环境模式变化频繁，需定期或自适应地更新模型，以保持其预测准确性与适应性。

定期模型再训练，每隔固定周期（如每周或每月）基于最新实测数据重新训练或微调模型；使用持续集成（CI）和持续部署（CD）流程自动完成训练、验证、部署闭环。

在线学习机制，部分模型可采用在线学习方式，对新数据进行快速适配；支持部分权重冻结+局部微调，避免灾难性遗忘。

多模型版本管理，利用模型版本管理工具（如 MLflow、DVC）记录不同训练周期下的模型版本与性能指标；支持回滚历史模型、A/B 测试等方式进行多模型对比验证。

9 效果评估与反馈迭代

极端天气预测系统的实际应用效果不仅取决于模型在历史数据上的训练精度，更关键的是其在真实环境中的预警有效性与用户接受度。因此，必须构建完善的效果评估与反馈闭环机制，将真实天气结果与用户反馈纳入系统，持续优化模型性能与预警策略，提高系统在实战中的可靠性与适用性。

9.1 反馈数据收集

为了全面评估模型在真实场景下的表现，需持续收集并分析以下类型的实测气象数据：接入国家气象局、地方气象站提供的实况观测数据（如气温、降水、风速、湿度）；包括分钟级、小时级和日级的历史实况，确保覆盖短时强降雨、雷暴大风、高温等极端天气类型；用于与模型预测值进行逐时间点比对，量化模型在不同场景下的误差。利用雷达回波图像、卫星云图等遥感资料辅助分析预测准确性，特别用于降水空间分布、强对流区域定位等高精度需求场景；可与模型空间分布预测结果进行匹配分析，评估模型空间推理能力。建立极端事件记录库，包括历史台风路径、暴雨积水报告、高温预警触发时间等；用于统计模型在极端天气预测上的召回率、准确率等业务关键指标。

在气象预警系统的应用中，用户感知和响应行为对系统优化同样具有重要价值：通过问卷调查、系统内嵌反馈功能收集气象专家或业

务人员对预测结果的主观评价；记录他们对预测时效、空间分辨率、置信度输出、预警触发规则等方面的改进建议；有针对性地优化模型设计和预警逻辑。面向公众用户的移动端/网页端提供简单反馈入口；收集用户对预警准确度、发布时间、信息表达清晰度等感知层面的评价；可借助自然语言处理分析意见建议文本，提炼共性问题。

9.2 动态性能评估与迭代优化机制

基于上述数据反馈，系统可建立以下自动评估与模型优化流程：

滚动验证与动态评分机制

- 使用滑动时间窗口对模型预测效果进行持续验证，
- 通过定期评分，识别性能退化趋势。

模型版本管理与对比测试

- 使用模型管理平台（如 MLflow）跟踪每次模型更新的训练参数、验证精度、部署版本；
- 支持多版本 A/B 测试，评估不同算法、不同训练样本、不同特征配置下模型的实际表现。

基于反馈的模型重训练

- 采集反馈数据并融合到训练集中，尤其是极端事件样本与误判样本；
- 采用增量训练或迁移学习机制更新模型，使其更好适应当前气候背景与区域特征；
- 可引入主动学习机制，在预测置信度较低或用户频繁反馈的区域优先补充标注样本。