基于有监督机器学习的珠三角气温数据统计与预测研究

摘要

近年来，全国范围内极端高温事件频发。气温的持续升高加剧了台风、暴雨、高温热浪等极端天气事件，对珠三角地区的稳定发展构成了严峻挑战。珠三角，作为华南地区的经济枢纽，是中国城市群高质量发展的代表。本研究聚焦于珠三角地区，采用有监督机器学习技术对珠三角地区的气温相关数据进行统计分析与预测，为区域气候适应和减灾工作提供科学的决策支持。本团队结合自动化网络爬虫技术和气象信息源下载，整理出近年来珠三角地区的详细气象数据，并进一步对数据进行严格的清洗与预处理。本研究建立了岭回归、多层感知机（MLP）和卷积-长短期记忆（CNN-LSTM）三种预测模型，并通过对比模型在测试集上相关系数、均方误差（MSE）和平均绝对误差（MAE）等评价指标，综合评估了模型的预测性能。研究结果表明，CNN-LSTM 模型在测试集上表现出色，相关系数高达 0.99997，显示出其在捕捉气象数据复杂关系方面的卓越能力。

鉴于 CNN-LSTM 模型的卓越预测能力，本研究进一步对模型灵敏度进行分析，并讨论其优缺点，为未来预测模型的改进和应用提供了新方向。本研究不仅为珠三角地区的气温预测提供了新的视角和方法，也为其他地区的气温预测和气候变化研究提供了参考。

关键词：气温预测；机器学习；珠三角；CNN-LSTM

目录

表格与插图清单

基于有监督机器学习的珠三角气温数据统计与预测研究

一、引言

（一）研究背景

随着全球变暖的日益加剧，气温变化早已成为研究热点问题。气温作为全球气候系统中的一个关键指标，对于理解气候模式和预测未来气候变化至关重要。为贯彻落实第七十五届联合国大会一般性辩论讲话中领导人提出的“双碳”目标，我国将应对气候变化作为国家战略，纳入生态文明建设整体布局和经济社会发展全局。珠三角，位于广东省中南部珠江口两侧，其气候特征主要表现为典型的南亚热带海洋季风气候，具有显著的季节性变化，决定着当地农业生产模式、城市规划、能源消耗、及居民的生活质量。珠三角作为中国南部的重要经济区域，在保持高质量快速发展的同时，也面临着气温升高和城市热岛效应对城市规划和布局提出的新要求，珠三角城市群气温数据统计与预测研究愈发重要。近百年来，大气学科群发展迅速，探索出一系列高效准确的气温预测模型。然而传统的统计模型气温预测方法仍存在一定的局限性，如自回归移动平均（ARMA）模型，虽在短期内具有一定的预测能力，但捕捉气候变化中的非线性动态和长期趋势方面表现不足。传统模型往往忽略了气候系统中的复杂相互作用，如大气、海洋和陆地之间的耦合效应，且容易忽略由人类活动引起的温室气体排放增加等因素。近年来，随着人工智能和大数据技术的发展，有监督机器学习技术因其在处理复杂数据和预测建模方面的优势，逐渐成为大气科学领域研究的新趋势。

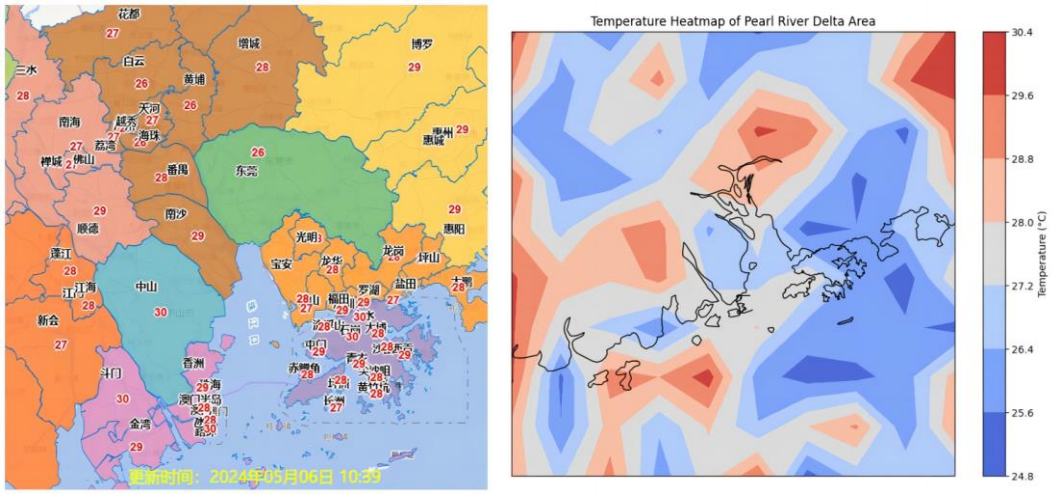


图 1：珠三角实时气温图（图源：https://www.gbaweather.net/sc/）与珠三角热图

（二）研究意义

在全球气候变暖的背景下,气温预测的准确性对于适应气候变化、保障社会经济稳定和推动可持续发展具有至关重要的作用。气温变化不仅影响生态系统的结构和功能，也与传染病的传播、热浪引起的健康问题以及过敏反应等密切相关，对生态系统和人类健康带来挑战。同时，气温的波动直接影响到旅游业、户外工作和零售等行业，进而影响经济的发展脉络。在农业领域，气温对作物的生长周期、产量和病虫害的发生具有决定性的影响，关系到农业规划和粮食安全。精准的气温预测可以帮助珠三角各界制定有效的规划和决策，降低因气温波动而带来的经济损失，促进珠三角地区多行业的高质量可持续发展。

鉴于气温变化对珠三角地区社会经济发展的重要性，本团队展开珠三角气温数据统计与预测研究，并期望为区域气候适应和减灾工作提供科学的决策支持。本研究旨在探索有监督机器学习技术在珠三角地区气温数据统计与预测中的应用，引入多种先进的有监督机器学习技术，并根据多个评价指标比较气温预测模型，探索一种高效的气温数据统计与预测方法。本研究深入地理解了气温变化的内在机制，不仅为珠三角地区的气温预测提供了新的视角和方法，也为其他地区的气温预测和气候变化研究提供了参考。

（三）研究现状

1. 传统气温预测模型

传统气温预测主要依赖于统计模型，如自回归移动平均（ARMA）和自回归积分滑动平均（ARIMA）模型等。程研等人基于全球平均温度的动态数据，对全球平均气温情况进行了统计分析，采用时间序列分析方法，进行数据预处理并建立 ARMA 模型，对全球气温变化进行预测[1]。刘闯基于时间序列分解模型将气温时间序列信息分解为长期趋势、循环趋势与季节变动,并在此基础上较好地实现了实测气温在时间序列上的拟合。对于预测而言,作者进一步搭建 ARIMA 模型与循环神经网络模型,结果表明,ARIMA 模型拟合的序列所产生残差较为平稳。而在循环神经网络的测试集中,生成的序列时间分辨率较高,对短期气温的预报而言模型具有一定的可信度[2]。这类模型在捕捉短期气温波动方面表现良好，但面对长期趋势和非线性特征时，其预测能力受限。伍红雨等人分析华南冬季气温的年际和年代际变化的多尺度气候特征，研究不同时间尺度上华南冬季气温与大气

环流以及海温的关系,建立了不同时间尺度上华南冬季气温预测模型并进行检验,为华南冬季气温预测提供参考依据[3]。

此外，传统气温预测模型往往不能充分考虑气候系统中的多因素相互作用，如大气、海洋和陆地的耦合效应。Cheng, Li 系统性探讨了海洋热含量变化的科学意义、定量评估了自上世纪 50 年代以来的热含量变化观测事实、提供了全球和区域热含量变化的未来预估[8]。考虑到耦合效应，为了更全面地理解气候系统，气象学家们发展了气候系统耦合模型，这些模型能模拟大气、海洋、陆地和海冰等多个圈层的相互作用。S Zhang, Z Liu 等人回顾了过去 20 年来海气耦合资料同化领域的主要进展及其对天气气候可预报性研究的影响，总体上按耦合资料同化领域由弱耦合同化向强耦合同化的发展，用耦合模式平衡协调地融合地球系统观测信息[9]。气候系统耦合模型虽然在模拟气候系统的整体行为方面取得了进展，但在计算资源和时间上的需求较高，限制了其在实时预测中的应用,需要进一步研究和开发高效的算法和计算技术，以提高模型的运行效率。

2.机器学习气温预测

近年来，有监督机器学习技术，因其处理复杂非线性关系和大规模数据集的优势，已广泛应用于气温预测。线性回归、多层感知机（MLP）和长短期记忆递归神经网络（LSTM）等算法已被证明在气温预测中具有较高的准确性和泛化能力。田淼选取重庆沙坪坝区 57516 站台的气象数据，应用基于一维卷积-双向 LSTM 神经网络的温度预测方法，用平均绝对误差(MAE)评估模型，结果表明不管是沙坪坝区的日最高气温预测,还是日最低气温预测,CNN-LSTM 模型的预测效果都为最优、预测方法有效[4]。刘剑南基于国家气候中心发布的中国 160 站月平均数据集以及 130 项环流指数,探索了机器学习在中国月、季节尺度气温及降水预测的应用,提出一个动态建模的机器学习(TD-ML)方案,对中国季节尺度气温及降水进行预测，结果表面所提出的预测模型,在具有挑战性的气候异常空间分布预测技巧方面表现出了较大的优势[5]。刘丹秀以机器学习中长短期记忆神经网络和随机森林模型两种多因素预测模型为研究方法,然后计算各影响因素与气温特征的相关系数,并结合散点图筛选出与气温特征强相关的多个因素，结果表明相较于改进的多因素随机森林预测模型和单因素 SARIMA 预测模型,改进的 LSTM 模型有着良好的预测效果,预测的平均绝对百分误差最小,低于其他模型约 3%[6]。

向德萍，张普等人提出了一种基于 Transformer 的多模态气象预测方法，能够同时考虑多种类型的气象数据，并有效地建模气象数据的特征分布和特征重要性。该方法在多种气象变量的预测任务中具有较好的性能，可以实现准确和稳定的气象预测[7]。在交通流预测领域，图神经网络（GNN）结合循环神经网络（RNN）的方法取得了显著的成功。此方法通过在静态切片空间中对图结构进行抽象，并在动态时间上对嵌入进行累积，有效地解决了时空预测场景中图的动态性问题，与交通流预测类似，气象预测也可以抽象为一个时空预测问题。H Lin,Z Gao 等人采用全球天气预报基准数据集 WeatherBench，提出了因地制宜的图卷积核（location-characterized kernel）来改良 DCRNN（一种用于交通流预测的图神经网络）有效地提升了气象预测的精度，并缩短预测时间[10]。

（四）研究内容

本研究聚焦于珠三角地区，采用三种有监督机器学习技术对珠三角地区的气温相关数据进行统计分析与预测，为区域气候适应和减灾工作提供科学的决策支持。团队成员利用自动化网络爬虫技术爬取 2020 年-2024 年 5 月以来“天气后报”官方网站公开的珠三角城市群天气数据，包括日期、天气和气温等数据，并从“rp5.ru”官方网站的权威数据库中下载更多有关相对湿度、风速、风向和平均气压等详细气象数据，整合爬取的天气数据及下载的详细气象数据，将其作为本研究的原始数据集，以完成后续研究内容。

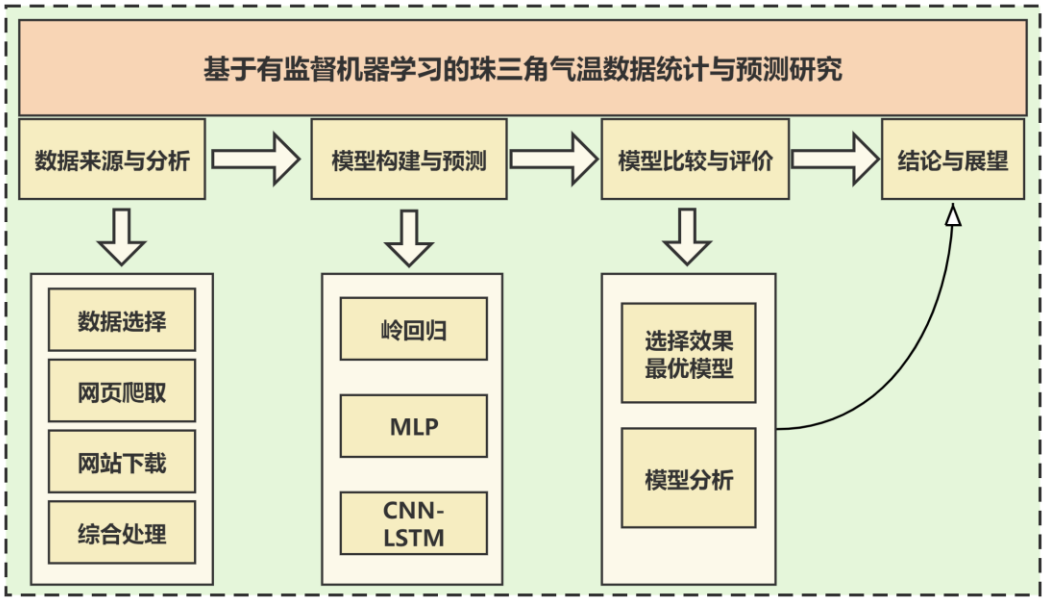


图 2：本研究整体框架图

本研究进一步对原始数据集据进行清洗和预处理，包括处理缺失值、异常值等，并将数据集进行可视化分析。然后基于我们的数据集，训练与构建三种有监督机器学习模型，包括岭回归模型、多层感知机模型（MLP）和卷积-LSTM 神经网络（CNN-LSTM）。随后，我们利用相关系数、均方误差(MSE)和平均绝对误差(MAE)等一系列通用指标对模型的预测性能进行进一步评估和比较，研究结果表明，CNN-LSTM 模型在测试集上表现出色，相关系数高达 0.99997，显示出其在捕捉气象数据复杂关系方面的卓越能力。

基于最优模型 CNN-LSTM，本研究对模型的预测结果进行深入分析，并针对模型的灵敏性、优缺点等方面进行了深入讨论。根据本研究所有工作，我们总结出本研究的主要成果并提出未来研究的方向，包括模型性能与改进、特征影响分析、预测准确性优化和更广泛的应用等。

二、数据来源与分析

（一）数据来源

本研究数据收集工作基于公开的气象信息源，主要选自“气象后报” 和“rp5.ru”网站。我们采用自动化网络爬虫技术，从“气象后报”网站采集近年来珠三角地区的天气数据，并从 rp5.ru 网站的权威数据库中下载更多相关气象数据。为确保研究的准确性和时效性,我们特别选取了 2020 年 5 月至 2024 年 5 月期间，珠三角地区气温，气压，相对湿度和风速等关键气象指标作为研究样本数据。此类数据不仅覆盖了四季变换对气象条件的影响,也充分考虑了地域性气候特征的差异,为后续的模型构建与分析提供了坚实的数据支撑。

表 1：研究所用数据集

数据来源	数据内容	样本数量
天气后报网站爬取	日期、天气状况、气温	1361 天
rp5.ru 网站下载	气温、湿度、气压、风速、降水、风向等	123675 项

（二）数据分析

本研究深入分析了以广州、深圳、珠海三市为代表的珠三角地区近五年的天气数据,研究关注于气温、相对湿度、风速和平均气压四个关键气象指标。通过

对这些数据进行综合分析与可视化，揭示了珠三角地区天气模式的相似性和变化趋势，为理解该地区的气候类型、地理条件以及可能的气候极端事件提供了重要参考，并为进一步统计分析和预测模型构建提供可靠依据。

以下就本研究的数据分析与可视化进行详细介绍：

1. 气象指标分析

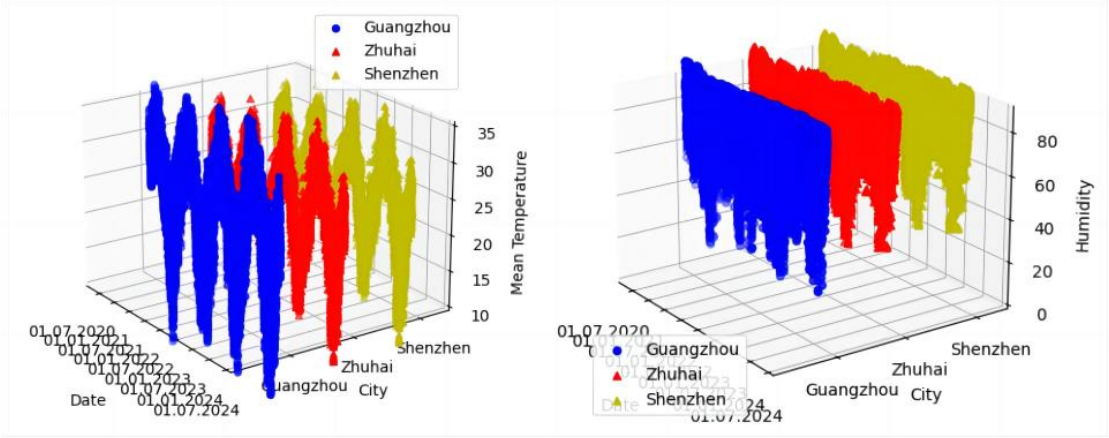


图 3：三市气温和湿度三维图

从图 3 的三市气温和湿度三维图可以看出，三市的气象数据变化曲线呈现出相似的趋势，且波动范围相近，表明三个地区在天气变化上具有较高的一致性，反映出珠三角城市群相似的气候类型和地理条件。基于此相似性，本节重点选择广州市详细气象数据分析并进行相应可视化，以反映整个珠三角气象数据分析。

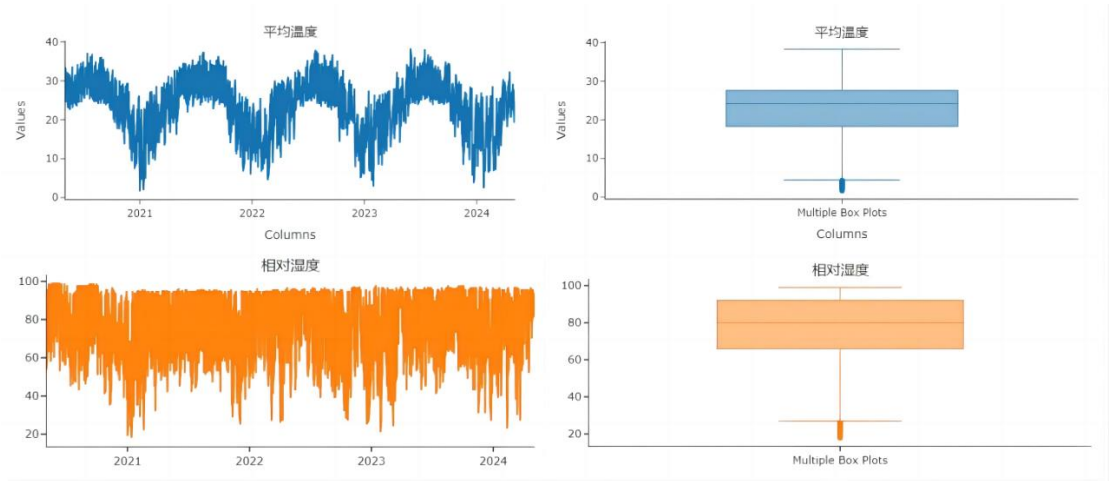


图 4：广州平均气温和相对湿度时间序列图及箱线图

图 4 反映了 2020 年至 2024 年 5 月广州市平均气温和相对湿度时间序列图及箱线图。由时间序列图可以看出，平均气温和相对湿度在年尺度上表现出平稳的波动，在月尺度上波动较大，表面广州市年气温和湿度变化较为稳定并呈现出明

显的季节性变化。由箱线图可看出温湿度范围的波动相对稳定，数据较为集中。

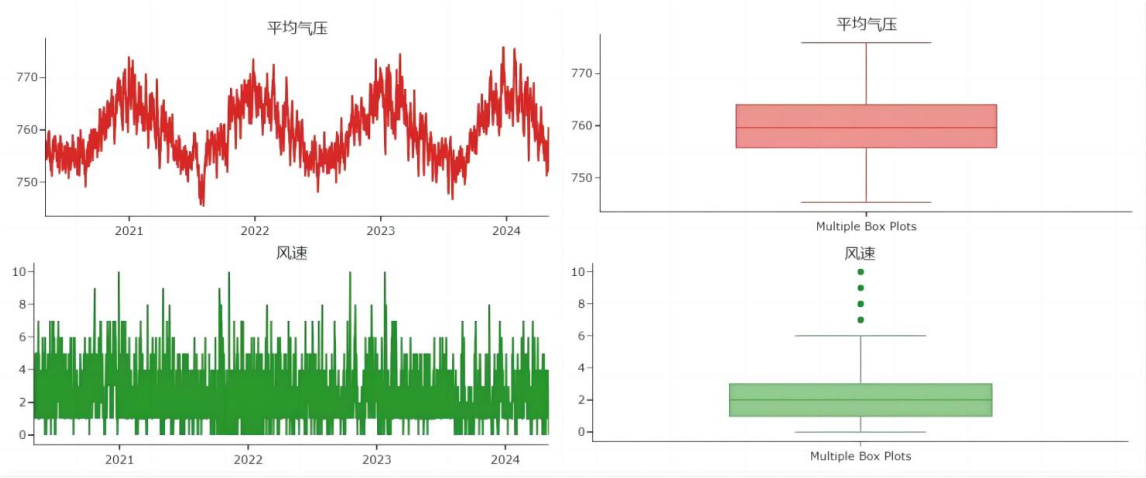


图 5：广州平均气压和风速时间序列图及箱线图

图 5 反映了 2020 年至 2024 年 5 月广州平均气压和风速时间序列图及箱线图。由时间序列图可以看出，平均气压在年尺度上表现出平稳的波动，在月尺度上波动较大，与上述平均气温和湿度的变化趋势一致，符合与季节性气温变化，如夏季高温导致热低压，冬季低温导致冷高压。而风速时间序列图波动较为明显，反映出广州市天气系统的多样性和集中的夏季极端天气事件（如台风等）。同时，由箱线图可以看出平均气压和风速范围的波动相对稳定，数据较为集中，而风速较多异常值的存在明显指示着台风类极端气候事件。

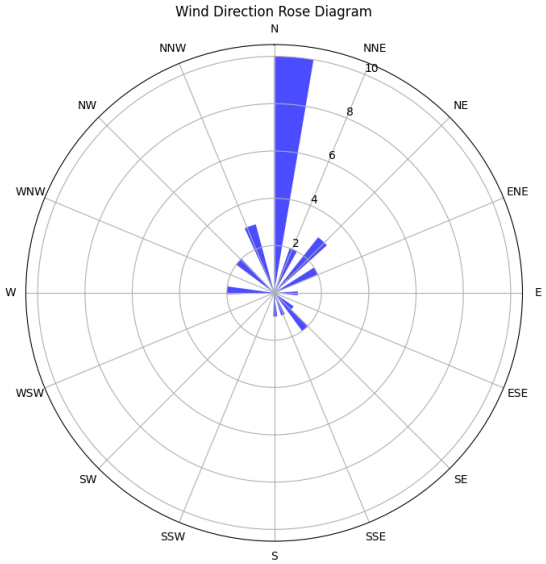


图 6：广州风向玫瑰图

图 6 以风向玫瑰图直观地展示了广州市风向频率。风向玫瑰图通常用来表示风向的频率或者主导风向。图中每个方向的扇区大小可表示该方向风的频率或者

常见程度。从图中可以看出 N 区偏 NNE 扇区面积较大，且标记的数值较高，表示该方向是主导风向，即风最常从北方吹来。

2. 相关性分析

由图 7 的关系矩阵热图我们可以分析湿度、气压、风速和降水这四个变量与气温之间的相关性，并考虑它们对未来气温拟合与预测模型的潜在影响。气温与气压之间存在强烈的相关性，气压是预测气温变化中一个非常重要的预测因子。因此，在建立预测模型时，应充分利用气压这一变量，以提高预测精度。

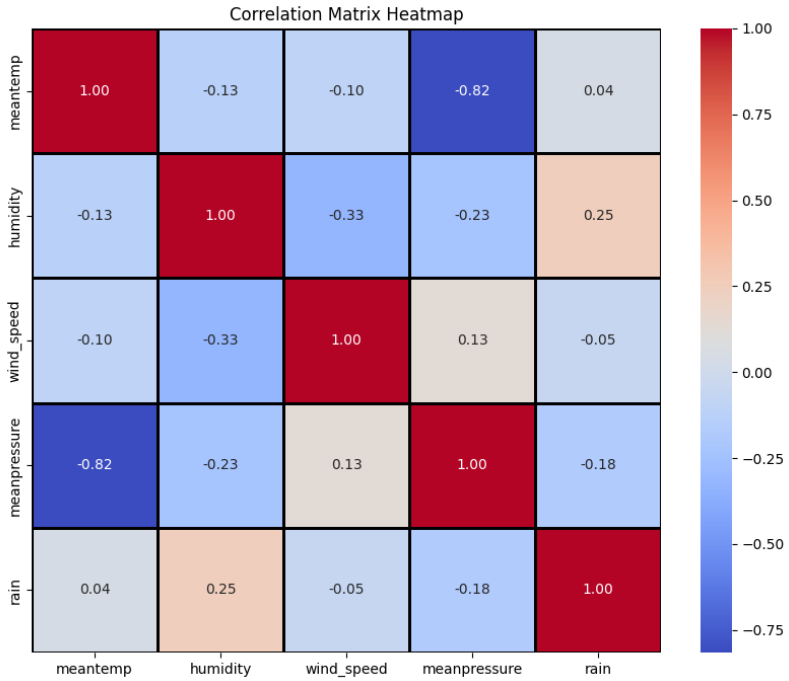


图 7：关系矩阵热图

降水与气温的相关性相对较弱，但在某些季节性或地区性气候模型中，考虑降水对气温的潜在影响仍是必要的，例如在考虑到降水可能引起的局部气温变化时。此外，尽管湿度和风速与气温的直接相关性较弱，但仍可能在特定条件下或与其他变量相互作用时影响气温。例如，在模拟温室效应或城市热岛效应时，湿度和风速的变化则变得重要。因此，在构建气温预测模型时，我们需综合考虑变量与场景的影响，尤其是在分析特定气候事件或地理区域时。

3. 加法模型分解

加法模型分解是一种常用的时间序列分析，用于将时间序列数据分解为趋势、季节性和残差三个成分

$$y(t) = T(t) + S(t) + R(t) \qquad \text{公式(1)}$$

其中，$y（t）$是时间点 t 的观测值，$T（t）$是时间点 t 的趋势成分，$S（t）$是在时间点 t 的季节性成分，$R（t）$是在时间 t 的残差。

首先，通过滤波器或移动平均方法平滑原始时间序列，得到趋势成分。这个过程的目的是捕捉时间序列中的长期变化趋势。我们采用了移动平均法用于消除时间序列中的随机波动：

$$T(t) = \frac{1}{2m+1} \sum_{i=-m}^{m} y(t-i) \qquad \text{公式（2）}$$

其中 $T(t)$ 表示在时间点 t 的趋势成分，m 是移动平均的窗口大小，决定了平滑的程度。较大的 m 会产生更平滑的趋势，但可能会导致滞后$y(t-i)$是在时间点 $t-i$的原始时间序列值。

然后对原始时间序列减去趋势成分，得到去趋势后的序列，再进行周期性分析，提取出季节性成分。这里使用傅里叶变换提取周期性变化：

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-j2\pi f t} dt \qquad \text{公式(3)}$$

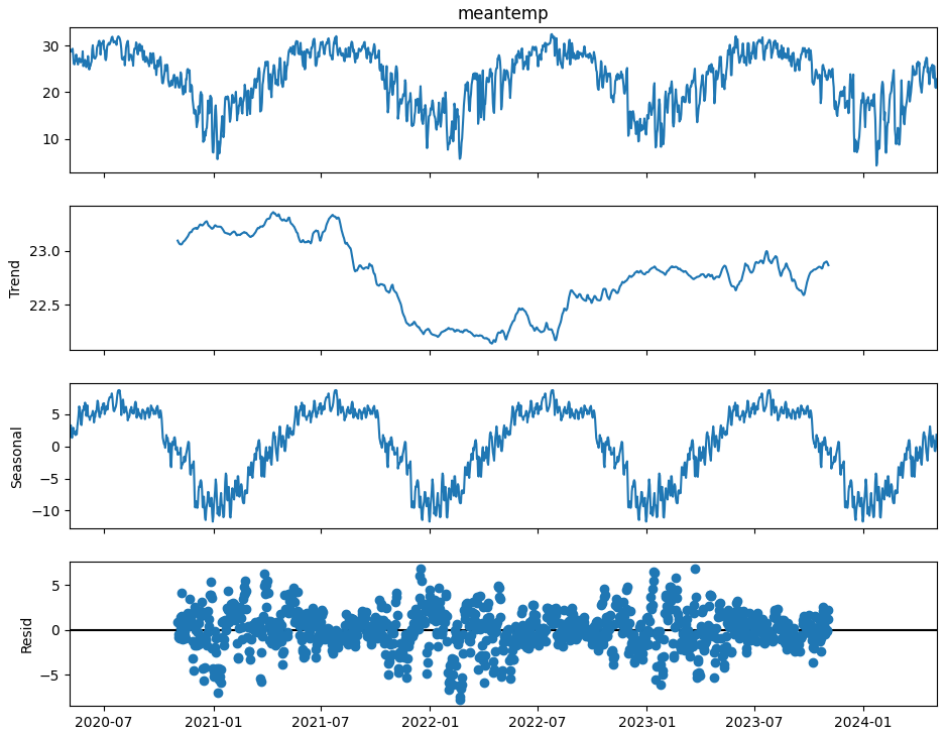


图 8：加法模型分解可视化

由图 8 可以看出，最上方部分展示了原始的气温数据。可观察到数据在一定范围内波动，显示出明显的季节性变化，即在特定周期内气温上升和下降；

趋势分量显示了气温数据随时间变化的平滑长期趋势，可看出气温趋势在 2021 年中期达顶峰后略有下降再趋于稳定，反映了气候变化等长期因素的影响。

季节性分量可明显看出，每年都有规律的气温升高和降低，对应于季节变化。

残差分量可看出，大部分残差聚集在零附近，表明模型已经相当好地捕捉了数据中的主要模式。残差中没有明显的模式或结构，表明分解是成功的。

（三）数据处理

1. 缺失值处理：

处理缺失值是数据预处理中的重要步骤。缺失值的存在会严重影响模型的性能和预测的准确性。面对大规模数据集时，选择合适的填充方法对于保持数据的统计性质和提高分析的准确性至关重要。对缺失值可视化处理，如图 9 所示。

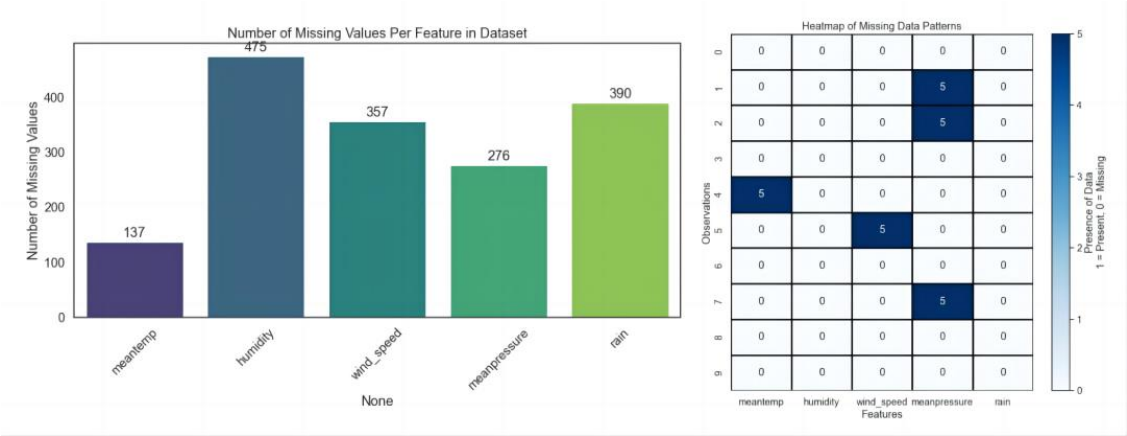


图 9：缺失值可视化处理

本研究使用前向填充处理缺失值。前向填充是处理时间序列数据中缺失值的常用方法，特别适用于连续记录但偶尔出现数据缺失的情况。前向填充假定一个缺失的观测值可以合理地被前一个非缺失值所代替，基于的假设是相邻的观测值之间有很高的相关性。

2. 异常值检测

自编码器是一种神经网络，用于学习数据的高效编码。在异常值检测中，自

编码器可用来识别重建误差异常大的数据点，这类点可能是异常值。对于气温数据，自编码器可有效地表示正常气温变化（日变化和季节性波动）以区分不符合常规模式的异常气温读数。将气温 x 压缩成一个较低维度的隐藏表示z，而解码器则尝试从这个隐藏表示重构原始输入$\hat{x}$。在输入层接受具有 2 个特征的输入（时间与气温），使用 4 个神经元和 ReLU 激活函数进行数据编码。编码器公式：

$$z = f(W_e x + b_c) \qquad \text{公式(4)}$$

然后在解码层尝试使用线性激活函数从编码表示中重构输入，解码器公式如下：

$$\hat{x} = g(W_d z + b_d) \qquad \text{公式(5)}$$

之后整个网络通过最小化均方误差损失函数训练，计算重构误差来检测异常值：

$$e = \left|\left|x - \hat{x}\right|\right|^2 \qquad \text{公式(6)}$$

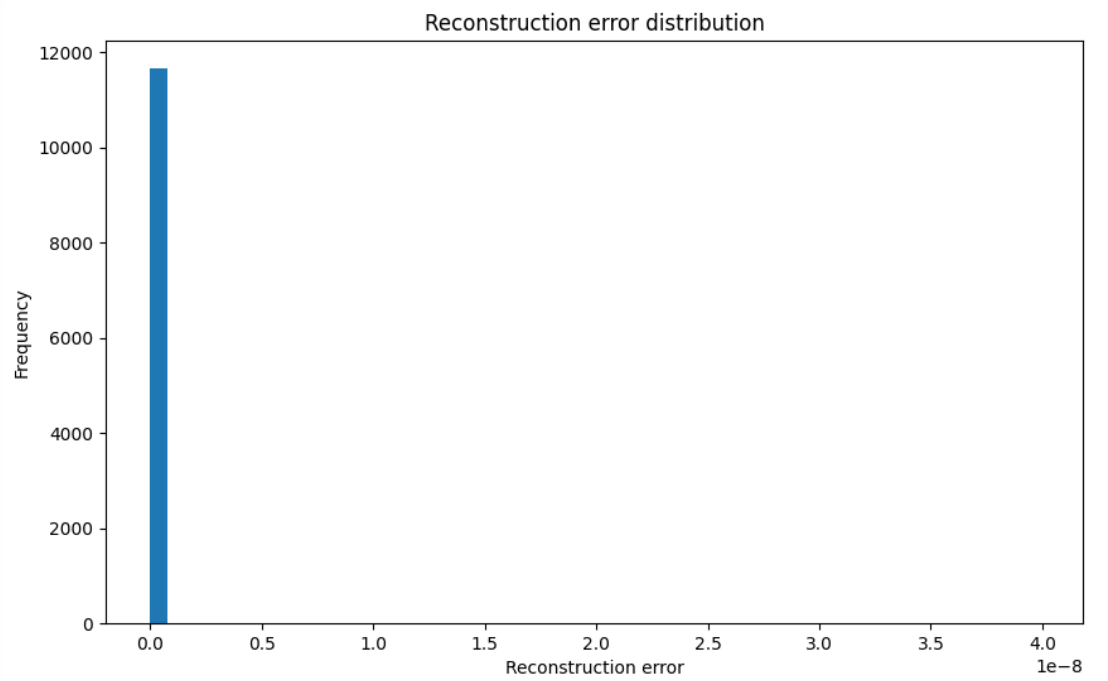


图 10：误差可视化

极少数误差值稍大，但仍然非常接近 0（小于 1e-8），这表明对所有数据点的重建误差都很小，没有显著的大误差点。

三、模型构建与预测

（一）岭回归模型

岭回归是一种线性回归模型的改进方法，用于解决普通线性回归中可能存在的过拟合问题。在岭回归中，通过添加一个正则化项可以有效地控制模型的复杂度，从而提高模型的泛化能力。这个正则化项基于模型的权重向量的平方和，通过调节正则化参数来平衡拟合数据和控制模型复杂度之间的权衡。岭回归的数学表达式如下：

$$\hat{\beta}^{ridge} = \arg\min_{\beta}\left\{\sum_{i=1}^{n}(y_i - X_i\beta)^2 + \alpha\sum_{j=1}^{p}\beta_j^2\right\} \qquad \text{公式(7)}$$

其中，$\hat{\beta}^{ridge}$是岭回归的系数向量，y_i是观察到的目标值，X_i是与观察值y_i相对应的特征向量，β是模型权重参数，p 是特征数量，α 是岭回归的正则化参数。

岭回归通过最小化目标函数来拟合数据，其中第一项是普通的最小二乘损失，第二项是正则化项，用于惩罚模型的复杂度。调节参数 α 可以控制正则化项的影响程度，从而平衡模型的拟合效果和复杂度。

我们首先构建了一个全面的特征矩阵 X，它综合了多种气象指标，包括气温、湿度、风速、气压、风向、天气状况、雨量和能见度。这些特征为我们的模型提供了丰富的信息，以便更准确地捕捉和预测气象变化。在选择回归模型的多项式次数时，我们通过多次实验和验证，最终确定了一次多项式作为回归拟合的模型。一次多项式因其简单性和有效性，在多个测试中均展现出了良好的预测性能。

为了进一步提高模型的泛化能力，防止过拟合现象，我们在模型中引入了一个正则化项。正则化是一种常用的技术，通过在损失函数中添加一个额外的惩罚项来限制模型的复杂度。在本研究中，我们选择了 L2 正则化，其参数（正则化系数）经过调整后设定为 0.1，以实现对模型复杂度的有效控制。在模型训练完成后，我们在独立的测试集上进行了预测，并对模型的性能进行了评估。测试结果表明，岭回归（Ridge Regression）在测试数据集上的表现达到了相关系数 0.77，这一结果表明模型具有较好的预测效果。

表 2：岭回归模型性能

测试集表现	结果值
测试集损失	0.8297648916096542
均方误差	8.012017281352431
均方根误差	2.1544674070887733
相关系数	0.7760400821179322

为了进一步验证模型的预测性能，我们提供了一系列的可视化图表进行分析。

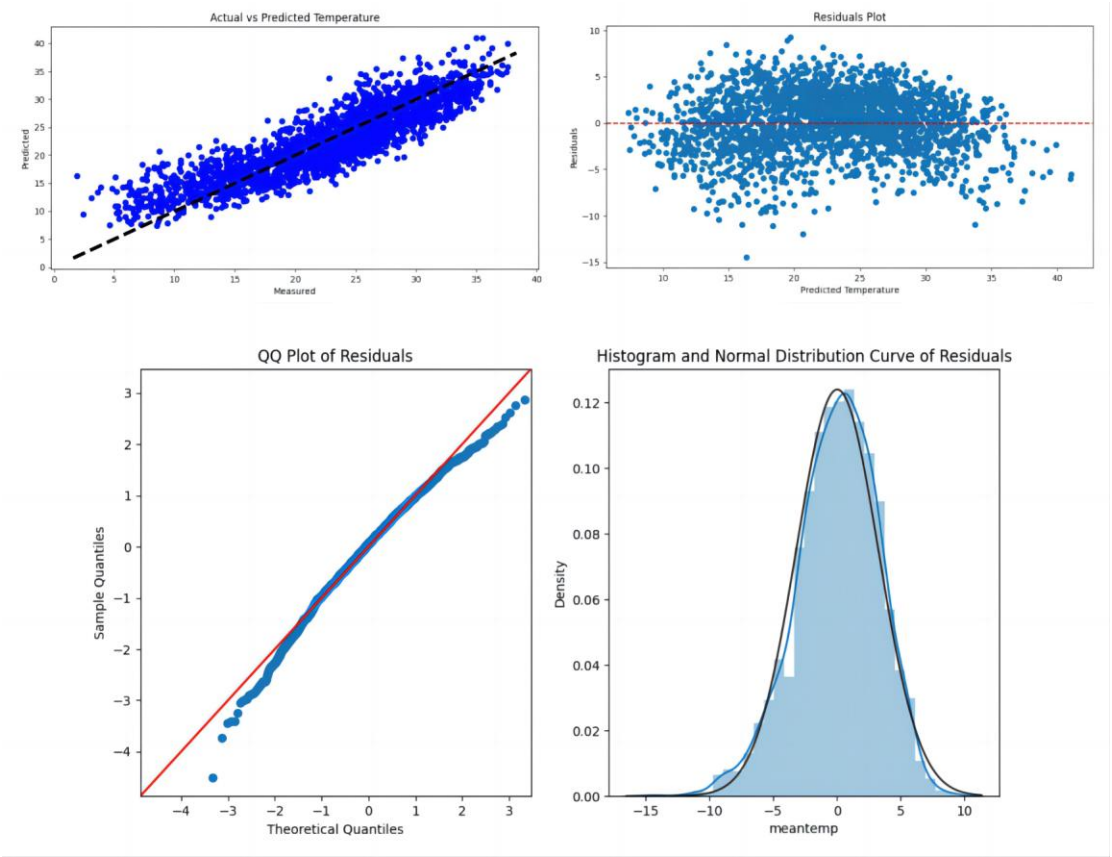


图 11：岭回归性能可视化

首先，实际与预测气温的对比图显示了模型预测值与实际观测值之间的一致性，该图直观地展示了每个观测点的预测值与实际值之间的关系。图中，如果预测值紧密地围绕着实际值，这表明模型能够准确地捕捉到气温变化的趋势，从而验证了模型的预测准确性。接着，残差图表明了预测误差的随机性，没有明显的

模式，这表明模型没有系统性偏差。此外，残差的 Q-Q 图近似地呈现出一条直线，这暗示了残差近似正态分布，是模型拟合良好的一个迹象。最后，残差的直方图与正态分布曲线的对比进一步证实了残差分布的正态性，表明模型的预测误差符合正态分布的假设。

通过上述可视化分析，我们得出结论：岭回归模型在气温预测任务上表现良好。实际与预测气温的对比图显示了模型的预测准确性，残差图证明了模型没有系统性偏差，而 Q-Q 图和残差直方图的正态性分析则表明了模型预测误差的统计特性。这些结果共同支持了模型的有效性和可靠性。

（二） 多层感知机模型

多层感知机（MLP）是一种广泛使用的前馈人工神经网络模型，它通过模拟人脑神经元的连接方式来处理和解决复杂问题。MLP 模型由三个主要部分组成：输入层、隐藏层以及输出层，每一层由多个节点（神经元）组成。输入层层负责接收外部输入信号，并将这些信号传递到网络的下一层。隐藏层是 MLP 的核心，可以有多个。每个隐藏层由一定数量的神经元组成，这些神经元通过加权和的输入信号，并应用一个非线性激活函数来产生输出。隐藏层的存在使得网络能够学习和执行复杂的函数映射。最终的输出由输出层产生，它根据网络的特定任务（如分类、回归等）来决定输出的形式。

MLP 模型的每个连接都有对应的权重，这些权重在训练过程中通过反向传播算法进行调整。训练过程中，网络通过比较实际输出和期望输出，计算损失函数，然后通过梯度下降等优化算法更新权重，以最小化损失函数。网络结构图如下：

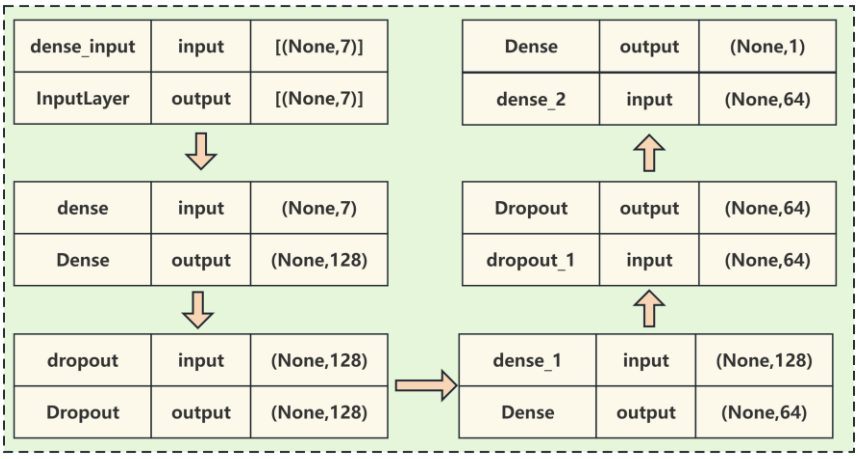


图 12：多层感知机网络结构图

我们使用风速，降雨量，湿度，时间日期等指标来对降水量进行预测，使用了 ReLU 激活函数，Adam 优化器，损失函数使用了均方误差。且经过多次尝试，我们设置了两个隐藏层，神经元数量分别为 128 个和 64 个。

下表 3 展示了训练得到的 MLP 模型在测试集上的表现，相关系数达到了 0.84，实现了比较好的效果，这一统计指标表明模型能够很好地捕捉数据之间的关系，实现了较高的预测准确性。此外，模型在测试集上的损失函数值也相对较小，这进一步验证了模型的预测效果。

表 3: 多层感知机模型性能

测试集表现	结果值
测试集损失	7.061423778533935
均方误差	7.061425430708115
均方根误差	2.6573342715413344
相关系数	0.8499625648106409

为了更直观地展示模型的预测性能，我们提供了真实结果与预测结果的对比图，以及相应的残差图和 Q-Q 图。

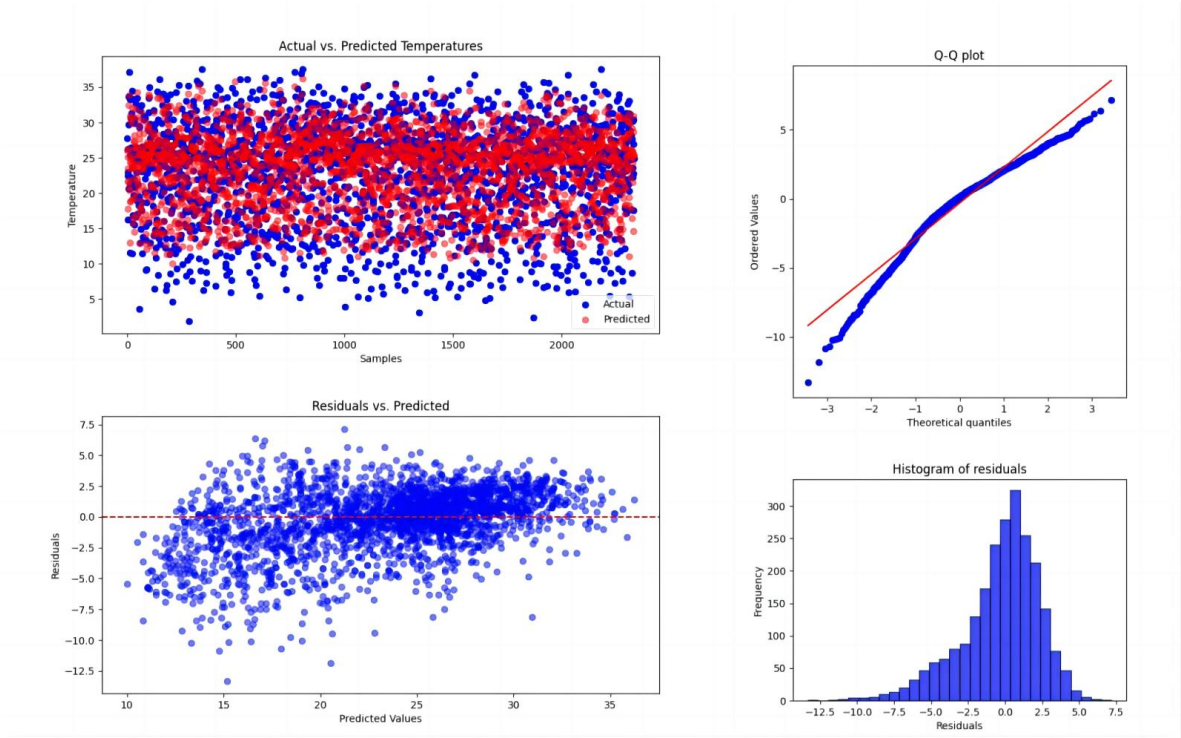


图 13：多层感知机性能可视化

从残差图中可以看出，绝大多数的残差值（即真实值与预测值之差）集中在0附近，这表明模型的预测值与真实值之间的偏差很小。尽管存在少数几个较大的残差值，但这些离群点并不影响模型整体的预测能力。

Q-Q 图进一步揭示了真实值与预测值的分布特性。通过比较样本数据的分位数与正态分布的分位数，Q-Q 图可以直观地展示两者的一致性。在我们的分析中，真实值和预测值的分布大致相似，这为模型的正态性假设提供了一定的支持。

综上，MLP 模型在测试集上展现出良好的预测性能和较高的数据拟合度。

（三） CNN-LSTM

卷积神经网络（CNN）是一种深度学习架构，在处理具有网格状拓扑结构的数据领域表现出色，尤其是在图像和视频分析任务中。CNN 的核心优势在于其能够通过卷积操作有效地提取输入数据的局部特征，并通过网络的多个层次逐步构建出更加复杂和抽象的特征表示。

在 CNN 中，卷积层扮演着至关重要的角色。每个卷积层都由一系列的可学习的滤波器（或称为卷积核）组成，这些滤波器在输入数据上滑动以识别局部模式。通过这种方式，卷积层能够捕捉到数据中的局部依赖关系，为后续的网络层次提供丰富的特征表示。尽管 CNN 最初是为处理图像数据设计的，但其强大的特征提取能力也适用于时间序列数据分析。在时间序列数据中，卷积层可用于捕捉时间窗口内的趋势和周期性模式。将时间序列数据视为一维的“图像”，并应用卷积操作，CNN 能识别出随时间变化的模式，这对预测和异常检测等任务至关重要。

长短期记忆网络（LSTM）是一种特殊类型的循环神经网络（RNN），专门用于处理序列数据，如文本、语音、时间序列等。相比于传统的 RNN 结构，LSTM 在处理长序列数据时更有效，因为它能够有效地捕捉和利用序列数据中的长期依赖关系。LSTM 的主要优点在于它具有记忆单元和门控机制，这使得它能够有效地控制信息的流动和遗忘，从而更好地处理长序列数据。

结合 CNN 和 LSTM，可以充分发挥它们各自的优势。首先，CNN 层能够从输入的多维时间序列数据（例如时间、气温、湿度等）中提取有用的局部特征。这些特征提取器通过卷积操作，有效地捕获了数据中的空间相关性，使得模型能够识别出不同时间点的重要模式和趋势。接着，这些提取的特征被传递到 LSTM 层，用于捕捉时间依赖性。LSTM 网络的记忆单元和门控机制允许模型有效地处理长

序列数据，并学习到序列中的长期依赖关系。这样，模型能够更好地理解数据的时间动态，并据此进行准确的预测。

我们设计了一个综合卷积神经网络（CNN）和长短期记忆网络（LSTM）的混合模型，以提高降水量预测的准确性。与多层感知机（MLP）类似，我们的模型输入包括风速、降雨量、湿度以及时间日期等多个气象指标。这些输入指标经过精心选择，旨在为模型提供全面的数据视图，从而增强预测能力。

在数据预处理阶段，我们执行了以下步骤以适配网络结构：

(1)数据标准化：通过将数据缩放到零均值和单位方差，我们减少了不同量级指标之间的差异，这有助于模型训练的稳定性和收敛速度。

(2)数据维度重塑：为了符合 CNN 的输入要求，我们将时间序列数据重塑为合适的维度，这允许模型利用其卷积层有效地捕捉时间序列数据中的局部模式。

表 4：CNN-LSTM 模型性能

测试集表现	结果值
测试集损失	8.984350077851995e-07
均方误差	0.0009478581158513121
均方根误差	0.0006990742062658162
相关系数	0.9999734371487724

我们对 CNN-LSTM 网络在测试集上的表现进行了综合评估，如表 4。模型的相关系数接近 100%，这表明模型能够非常精确地捕捉输入指标与降水量之间的复杂关系。此外，模型的损失函数值极低，这进一步证实了模型的预测性能。在误差分析方面，我们观察到所有类型的误差都保持在较低水平，这表明模型对不同情况下的降水量预测都具有较高的准确度。这些结果为我们提供了信心，表明所设计的 CNN-LSTM 模型能够为气象预测任务提供强大的支持。

下图 14 分别从不同维度对预测结果和真实结果进行了可视化，所设计的网络模型在气温预测任务上表现良好。模型的预测误差较小，残差近似正态分布，且没有明显的系统性偏差。这些结果表明，模型能够准确地捕捉到气温变化的趋势，为气温预测提供了可靠的支持。

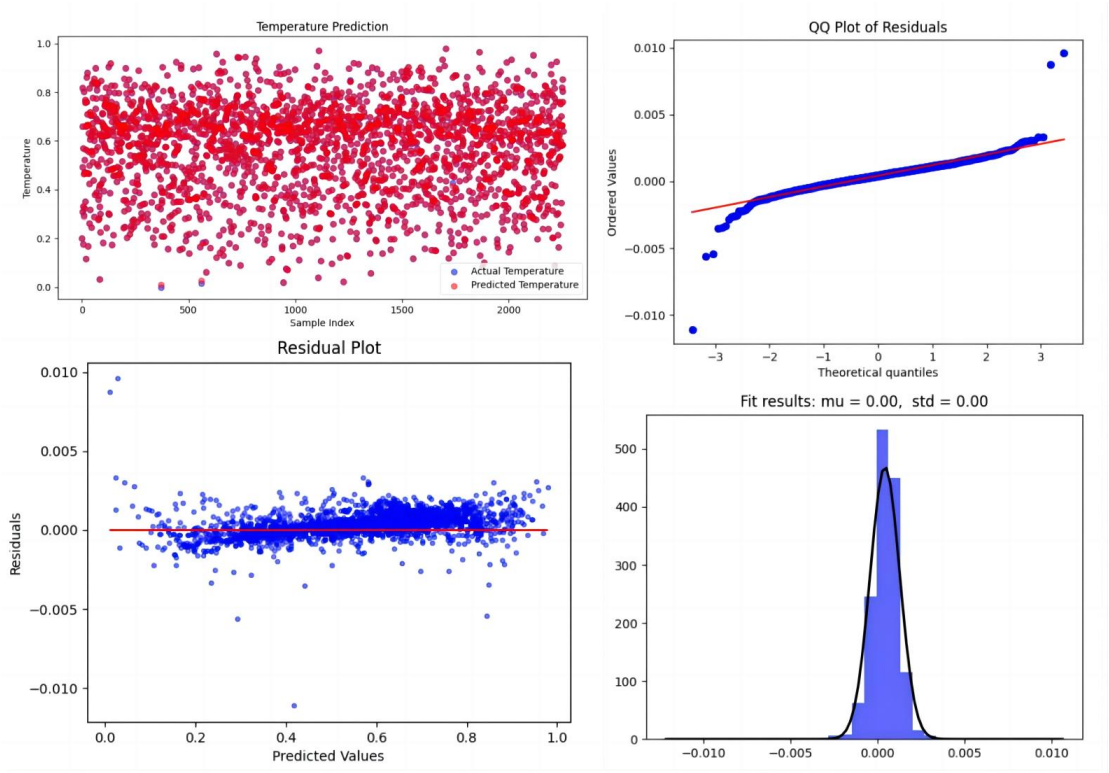


图 14：CNN-LSTM 性能可视化

四、模型比较与评价

（一）模型比较

在比较不同模型的性能时，我们选择了三个关键的指标：均方误差（MSE）、均方根误差（RMSE）和相关系数。这些指标在评估回归模型性能方面具有的独特优势和互补性。MSE 衡量的是模型预测值与实际观测值之差的平方的平均值。它对大的预测误差给予更大的惩罚，因此能够更敏感地捕捉模型的预测偏差。RMSE 是 MSE 的平方根，它与原始数据具有相同的量纲，使我们能够直观地理解预测误差的大小。相关系数衡量的是模型预测值与实际观测值之间的线性关系强度。一个高的相关系数表明模型能够很好地捕捉数据之间的关系。

基于上述指标，我们对三种不同的模型进行了比较：岭回归模型、多层感知机（MLP）模型和长短期记忆网络（LSTM）模型,结果如表 5 所示。

岭回归模型模型在均方误差和均方根误差上的表现分别为 8.01 和 2.15，相关系数为 0.83。表明岭回归模型能够以中等的准确度进行预测，但由于其简单性，可能无法捕捉到数据的所有复杂特征

MLP 模型的均方误差和均方根误差分别为 7.06 和 2.66，相关系数略高于岭

回归模型，达到 0.85。MLP 作为一种深度学习模型，能够通过非线性变换捕捉更复杂的关系，但其误差仍高于 LSTM 模型

CNN-LSTM 模型在所有指标上均展现出卓越的性能，均方误差和均方根误差极低，分别为 0.0009 和 0.0007，相关系数接近 1，达到 0.99997。CNN 和 LSTM 模型结合，特别适合处理时间序列数据，其门控机制能有效地捕捉长期依赖关系。

表 5：三种模型比较

模型	均方误差	均方根误差	相关系数
岭回归模型	8.012017281352	2.1544674	0.82976
MLP 模型	7.0614254307	2.6573342715	0.8499625648
CNN-LSTM	**0.000947858**	**0.000699074**	**0.99997**

综合考虑，CNN-LSTM 模型表现最为出色，其极低的误差和极高的相关系数表明了其在气温预测任务上的优越性能。

（二）模型评价

1.灵敏性分析:

对 CNN-LSTM 模型进行灵敏性，检测其鲁棒性效果，分别从两个方面进行改变：训练数据集的大小以及不同自变量的组合。

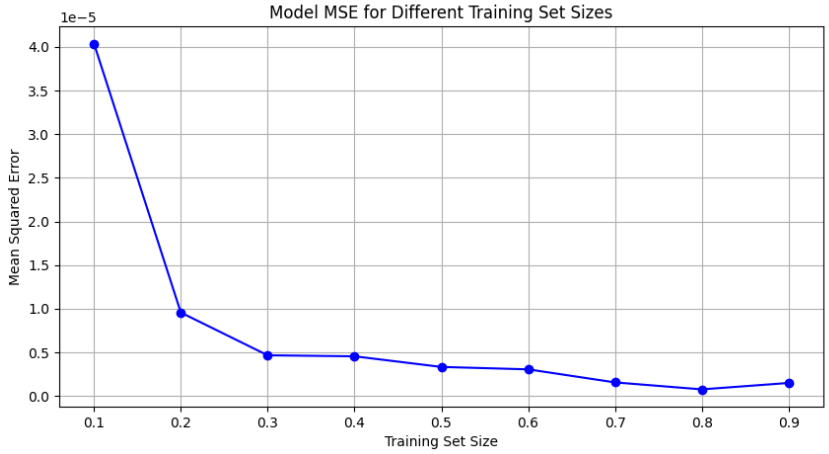


图 15：训练集大小与 MSE 关系

折线图 15 展示了不同规模的训练集大小对于模型 MSE 值变化的影响，随着训练集大小的增加，模型的均方误差（MSE）显著下降，尤其是在训练集大小从 10%增加到 30%的阶段。表明增加训练数据的数量可显著提高模型的预测准确性。

如下是组合不同的自变量，观察不同自变量组合对于模型 MSE 值变化的影响，

一般而言，组合变量的 MSE 值比单个变量的值要高一些。

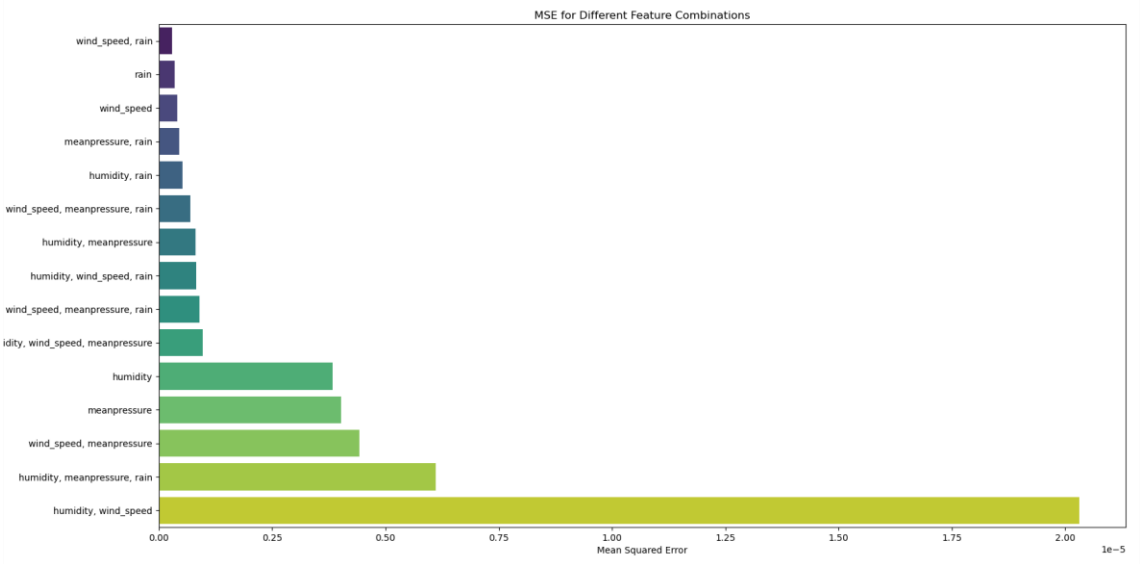


图 16：不同自变量的组合与 MSE 关系

包含“湿度”、“平均气压”、“降水”这三个特征的组合产生了最低的 MSE，表明这组特征在预测气温方面非常有效。单独的“降水”特征以及与“风速”结合的特征组合表现出较高的 MSE，这表明这些特征单独或者在没有“平均气压”或“湿度”的情况下对模型的预测能力贡献较小。

上述结果表明不同的训练集大小以及自变量组合会对模型的结果产生比较大的影响。符合 CNN-LSTM 模型对训练数据量和输入特征敏感的特性，通过适当调整训练集大小和精心选择特征组合，可以显著提升模型的预测性能和效率。

2. 模型优缺点：

模型优点：（1）考虑序列的时空特性：CNN-LSTM 结合了卷积神经网络（CNN）和长短期记忆网络（LSTM），能有效地处理时间序列数据中的时空特征，而气温具有明显的时序性和空间相关性；（2）自动特征提取：CNN-LSTM 能自动从输入数据中学习到特征，而不需手动提取特征，可减轻特征工程的负担；（3）适用于多维输入：CNN-LSTM 能够处理多维输入数据，例如本研究中气象数据包含的多个气象特征。

模型缺点：（1）计算量大：CNN-LSTM 模型因其结构复杂，具有较高的计算复杂度，在处理大规模数据集时往往需更多的时间开销和计算资源，这一特限制了它在资源受限环境下的应用。但这种计算强度也是其能够捕获复杂数据模式和

提高预测准确性的基础，对于需高精度预测的应用场景，如天气预测，这种深度学习模型的使用是合理的选择；（2）序列长度敏感：对于很长的时间序列数据，CNN-LSTM 模型可能会面临梯度消失或梯度爆炸的问题，需进一步采用各种策略，如引入梯度裁剪、使用更高级的优化算法，或改进模型架构来缓解这一问题。此外，这也推动了技术的进步和新算法的开发，如 LSTM 变体和注意力机制，都有助于更好地处理长序列且能提高模型的整体性能。

五、结论与展望

本研究通过综合应用自动化网络爬虫技术和有监督机器学习模型，对珠三角地区 2020 年至 2024 年 5 月的气象数据进行了深入的统计分析与预测。研究发现，相较于岭回归和多层感知机（MLP）模型，卷积-长短期记忆（CNN-LSTM）模型在气温预测任务上展现出了显著的优越性，在测试集上的相关系数高达 0.999973，验证模型在处理复杂时空数据方面的高效性和准确性。

本研究进一步对 CNN-LSTM 模型的灵敏度进行了细致的分析，探讨了不同气象指标和训练集大小对模型性能的影响，为未来模型的优化提供了方向。同时我们也注意到，尽管 CNN-LSTM 模型在本研究中表现出色，但其计算成本较高，对长序列数据的处理仍需进一步的技术改进。

本研究计划在以下方面进行深入探索：（1）进一步优化 CNN-LSTM 模型的结构，以降低计算成本并提高模型的运行效率；（2）探索更多气象指标和外部数据源，以增强模型的泛化能力和预测精度；（3）考虑到气候变化的不确定性，可引入概率预测方法，为决策者提供更全面的风险评估；（4）将本研究的方法和模型应用于其他地区的气温预测，以验证其普适性和适应性。

本研究不仅为珠三角地区的气候适应和减灾工作提供了科学依据，也为全球其他地区的气候预测和气候变化研究提供了新的视角。随着技术的不断发展和数据资源的日益丰富，人工智能技术将在气象预测研究中发挥越来越重要的作用。

参考文献

[1] 程研,华志强,黄玉洁,侯云艳. ARMA模型在预测全球平均温度情况中的应用[J]. 内蒙古民族大学学报(自然科学版),2023,38(02):103-108.

[2] 刘闯. 南极Dome A时序气温建模与预测初探[D].武汉大学,2023.

[3] 伍红雨, 华南冬季气温的多尺度特征及预测研究. 广东省,广东省气候中心,2016-05-27.

[4] 田淼. 基于LSTM模型的重庆气温预测研究[D].重庆大学,2022.

[5] 刘剑南. 基于机器学习方法的中国短期气候预测研究[D].南京信息工程大学,2024.

[6] 刘丹秀. 基于随机森林和长短期记忆神经网络的气温预测研究[D].安徽建筑大学,2023.

[7] 向德萍, et al. "基于 Transformer 的多模态气象预测." Journal of Computer Engineering & Applications 59.10 (2023).

[8] Cheng, Lijing, et al. "Past and future ocean warming." Nature Reviews Earth & Environment 3.11 (2022): 776-794.

[9] Zhang, Shaoqing, et al. "Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: a review." Climate Dynamics 54 (2020): 5127-5144.

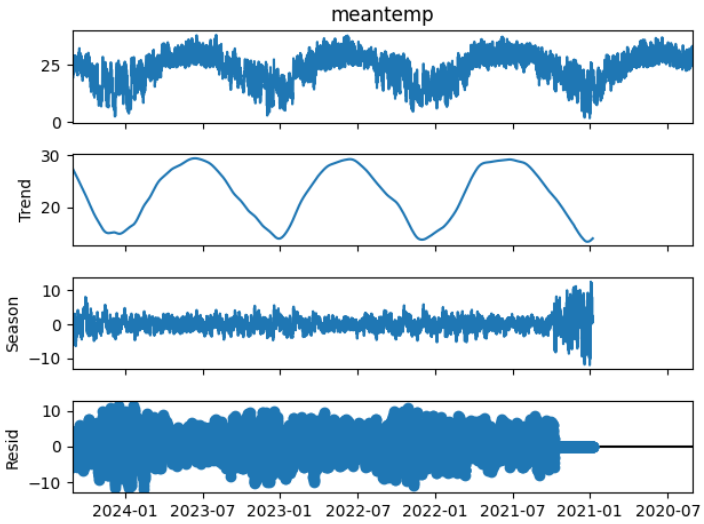
[10] Lin, Haitao, et al. "Conditional local convolution for spatio-temporal meteorological forecasting." Proceedings of the AAAI conference on artificial intelligence. Vol. 36. No. 7. 2022.

附录

附录一
介绍：数据分析方法补充

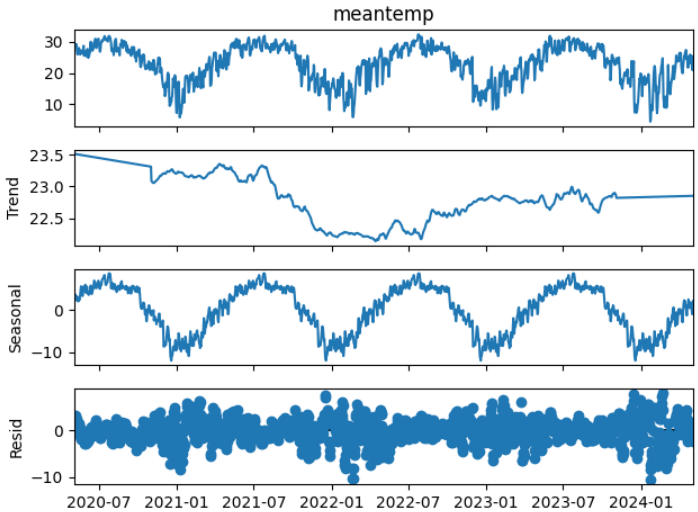
（1）STL 模型分解：

STL 分解是非常灵活的季节性分解方法，允许季节性成分随时间变化，且趋势和季节性的平滑参数可以独立调整。STL 的优势在于对异常值具有鲁棒性，并且可处理任意季节性周期长度。



反映出数据的长期变化趋势，趋势线相对平滑，呈从开始到结束的温度逐渐下降的趋势。

（2）动态调和回归分解：



趋势线更动态，季节性变化比 STL 分解的幅度大，体现出气温的高峰和低谷。

附录二

介绍：本文提及的简称与全称对应表

附表：简称与全称对应表

全称	简称
珠江三角洲，包括广州、深圳、珠海等九个城市	珠三角
多层感知机	MLP
卷积-长短期记忆神经网络	CNN-LSTM
均方误差	MSE
平均绝对误差	MAE