

本科生毕业设计

|  |
| --- |
| 基于数据分析的人工降雨效果评估 |

|  |  |
| --- | --- |
| 院 系 | 计算机科学与技术 |
| 专业班级 | 计科1703 |
| 姓 名 | 马恒奔 |
| 学 号 | U201714601 |
| 指导教师 | 秦磊华 |

2021年05月30日

**学位论文原创性声明**

本人郑重声明：所呈交的论文是本人在导师的指导下独立进行研究所取得的研究成果。除了文中特别加以标注引用的内容外，本论文不包括任何其他个人或集体已经发表或撰写的成果作品。本人完全意识到本声明的法律后果由本人承担。

作者签名： 年 月 日

**学位论文版权使用授权书**

本学位论文作者完全了解学校有关保障、使用学位论文的规定，同意学校保留并向有关学位论文管理部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。本人授权省级优秀学士论文评选机构将本学位论文的全部或部分内容编入有关数据进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本学位论文属于 1、保密囗，在 年解密后适用本授权书

2、不保密囗 。

（请在以上相应方框内打“√”）

作者签名： 年 月 日

导师签名： 年 月 日

摘 要

“天水计划”是以带电粒子催化降雨为核心原理的人工降雨项目。为了对带电粒子催化降雨的效果进行评估，需要对“天水计划”外场试验期间的天气数据进行分析，对外场试验设计一套可靠的评估方案，并得出可靠的评估结果，作为之后开展外场试验的重要依据，以及为人工影响天气的其他研究提供参考。

参考了国内外的人工影响天气试验的效果评估，根据收集到的试验数据的特点，设计了一套科学的评估方案。按照侧重点不同，分为宏观效果评估和微观效果评估两部分。宏观上，主要采用机器学习方法，以试验区整体为研究对象，使用了神经网络模型对试验区自然降雨量进行了预测，近似求出实验期间假设未进行作业时的自然降雨量，通过与实际降雨量对比，说明外场试验的宏观效果。微观上，主要采用统计检验方法，对试验区内部的播种差异、风向差异进行研究，说明试验区内部降雨差异与外场试验之间的关系，从微观层面说明外场试验的实际效果。

**关键词**：人工增雨；机器学习；降雨预测；统计检验

Abstract

"tianshuijihua" is an artificial rainfall project based on the core principle of catalytic rainfall with charged particles. In order to evaluate the effect of charged particle-catalyzed rainfall, it is necessary to analyze the weather data during the field test of "tainshuijihua", to design a reliable evaluation scheme for the field test, and to produce reliable evaluation results as an important basis for the subsequent field test and for other research on artificially influenced weather. The results will be used as a reference for other research on artificial weather.

A scientific evaluation scheme was designed based on the characteristics of the collected experimental data, with reference to the effect evaluation of artificial weather experiments at home and abroad. According to the different focus, it is divided into two parts: macro-effects assessment and micro-effects assessment. Macroscopically, machine learning methods were mainly used to predict the natural rainfall in the test area as a whole, and a neural network model was used to approximate the natural rainfall during the experiment period assuming no operation was conducted, and to illustrate the macroscopic effect of the outfield experiment by comparing it with the actual rainfall. At the micro level, statistical tests were mainly used to study the seeding differences and wind direction differences within the experimental area, to illustrate the relationship between rainfall differences within the experimental area and the outfield experiment, and to illustrate the actual effect of the outfield experiment at the micro level.

**Keywords:** Artificial rainfall enhancement, machine learning, rainfall prediction, statistical test

目 录

[摘 要 I](#_Toc72601408)

[Abstract II](#_Toc72601409)

[1 绪 论 5](#_Toc72601410)

[1.1 课题背景 5](#_Toc72601411)

[1.2 国内外研究现状 7](#_Toc72601412)

[1.3 论文结构 9](#_Toc72601413)

[1.4 课题来源 10](#_Toc72601414)

[2 方案论证 11](#_Toc72601415)

[2.1 需求分析 11](#_Toc72601416)

[2.2 可行性分析 11](#_Toc72601417)

[2.3 开发工具分析及选择 12](#_Toc72601418)

[2.4 关键技术 13](#_Toc72601419)

[2.5 基本方案的制定 15](#_Toc72601420)

[2.6 设计中考虑的制约因素 16](#_Toc72601421)

[2.7 成本估算 17](#_Toc72601422)

[2.8 本章小结 17](#_Toc72601423)

[3 基于全连接神经网络的自然降雨预测模型 18](#_Toc72601424)

[3.1 六盘山自然降雨预测模型说明 18](#_Toc72601425)

[3.2 数据处理与特征选取 18](#_Toc72601426)

[3.3 预测模型的建立 21](#_Toc72601427)

[3.4 模型应用 28](#_Toc72601428)

[3.5 本章小结 31](#_Toc72601429)

[4 基于统计分析的人工增雨效果评估模型 32](#_Toc72601430)

[4.1 效果的统计检验 32](#_Toc72601431)

[4.2 随机对比试验 32](#_Toc72601432)

[4.3 风向影响的降雨差异分析 40](#_Toc72601433)

[4.4 本章小结 50](#_Toc72601434)

[5 总结与展望 51](#_Toc72601435)

[致 谢 52](#_Toc72601436)

[参考文献 53](#_Toc72601437)

# 绪 论

本章介绍了“天水计划”项目的由来、主要内容和意义，然后指出了外场试验效果评估的重要性，也就是本文工作的意义。接着说明了国内外人工降雨效果评估方面的常见做法，提出了引入机器学习对传统评估进行改进的想法。

## 课题背景

### 人工降雨与“天水计划”

水资源短缺是一个全球性的难题。与世界主要发达国家相比，我国的水资源短缺状况尤其严峻，人均水资源量仅为世界平均水平的四分之一[1]。气象统计表明，大气中只有16%-18%的水汽能形成降雨。如果能够采用技术手段，将大气水资源的利用率提高若干个百分点，这将显著提升我国有效水资源总量，经济效益潜力巨大。

基于干冰、碘化银等催化剂的传统人工增雨技术，是目前大气水资源开发的一种主要技术手段。但是近年来，我们对水资源的需求量越来越大，供需矛盾日益突出，传统人工增雨技术的一些固有缺点逐渐凸显，如作业效率低，温湿度窗口过小，火箭、高炮、飞机等播撒方式管控严格等。这些缺点大大限制了大气水资源开发的成效和规模。

所以，本世纪以来，越来越多的西方国家开始进行人工降雨新技术的研究，其中最具有代表性的技术之一就是带电粒子催化降雨。该技术至今已在墨西哥、阿联酋、阿曼等多个国家的干旱地区进行了成功的外场试验。[2-4]鉴于此技术对于充分缓解水资源短缺问题意义重大，国外相关机构都对此严加保密。

在此背景下，我们国家在2016年推出了“天水计划”。该项目以基于带电粒子催化降雨雪的新技术为研究对象，围绕大气水资源开发与调控技术体系的构建，从理论、关键技术及装备到应用示范开展全链条研究。该项目的主要内容包括，开展带电粒子催化降雨雪的科学原理的研究，突破关键技术与研发试验装备，进行人工降雨雪过程的模拟与空陆一体化区域水资源适应性调控，以及进行带电粒子催化降雨雪系统集成与应用示范研究[5]。“天水计划”的基本原理是在通过试验装置，向大气中播撒带电粒子，从而使空气中的部分气溶胶带电。这些带电气溶胶粒子的静电场对其他中性水分子簇团存在极化效应，产生带电气溶胶粒子对被极化的水分子簇团的非接触的电场凝聚力，促使其凝结速率增加，促进降雨的形成。

“天水计划”的外场试验区选择了六盘山地区和乌鞘岭地区，在六盘山地区建立了三个单电极带电粒子发生器装置和一个双电极带电粒子发生器装置，在乌鞘岭地区建立了三个单电极带电粒子发生器装置。通过在外场试验区开展试验，检验粒子发生装置的实际效果。

“天水计划”意在自行研发带电粒子催化降雨雪的核心技术，通过对外场试验中带电粒子发生装置的实际效果的分析，对带电粒子催化降雨雪试验的方案进行评估，进而更加充分利用大气水资源以缓解国内水资源匮乏的现状。

### 降雨效果评估

“天水计划”从2016年开始实施，先后完成了带电粒子催化降雨雪的科学规律和关键技术的研究，完成了应用示范具体选址，完成了气象监测网络的设计以及基站的集成控制。该项目在2020年确定了外场试验方案，并于2020年7月到2020年11月，在乌鞘岭和六盘山两个试验区成功开展了外场试验。

为了对外场试验的结果进行分析，以及对带电粒子人工降雨方案的效果进行评估，从而作为进一步开展实验的重要参考，需要进行带电粒子人工降雨外场实验效果评估。人工降雨效果检验是人工降雨试验中重要的一部分，它的基本要求为科学、客观和定量。[6]

为了进行科学有效的效果评估，一个重要工作是估计出实验区假设未播种情况下的自然降雨量，从而区分出实际降雨中的自然降雨部分和催化增加的部分。一般情况下，自然降雨量的求取采用线性回归的方法，通过线性回归的方法建立目标区域和对比区域的历史回归方程。但是预测精度不高，也不能充分利用采集到的气象数据。同时，“天水计划”所采取带电粒子催化降雨雪技术在国内并无先例可循，装置的运行特性、影响范围、作用时间都尚未可知，需要对多种气象要素进行分析。

而随着计算机技术的不断发展，机器学习已经在计算机视觉[7]、自然语言处理[8]、数据挖掘[9]、数据分析[10]等多个领域取得了很好的效果。机器学习通过选取合适的算法，从大量数据中自动地归纳数据内在的逻辑和规则，再根据这个归纳的结果进行新数据的预测。相比于传统的基于统计学的算法，利用机器学习可以更好地解决非线性的降雨问题[11]。将机器学习地方法和传统统计方法相结合，可以进一步提高效果检验地精确度和可信度。本文使用了神经网络对自然降雨量进行了预测，具有较高的可靠度，通过分析预测降雨量和实际降雨量，完成了部分评估工作。

神经网络的预测模型也有自身的限制，它本身需要大量的数据进行训练来提高模型的精确度。但是“天水计划”外场试验周期较短，只有部分历史数据丰富的站点可以使用神经网络模型来预测自然降雨。“天水计划”项目的自建站点建成时间较晚，采集数据较少，这些数据无法构建神经网络模型，需要结合传统的统计检验方法和显著性检验方法。为此设计了多种分析手段，来从多方面的对多种气象数据进行分析评估。

### 面临的问题和挑战

降雨评估手段没有规范化的定量方法。目前还没有令人信服的科学依据表明各种形式的播云试验是有效或可行的增加降雨量的手段。而且，带电粒子人工降雨在国内也是首例，对于降雨效果的评估设计是困难的而没有依据可循，需要设计科学有效的评估方法。也要进行多方面的评估，从多角度来对结果进行评估。

数据清理、整合和分析。效果评估的数据多来源于自动气象站，该类数据的数据量较大，采样密度大，缺测数据也较多，且该类数据的观测项目较多，包含了很多干扰项目。不同地区，不同气象站的数据结构不同，没有统一标准。需要对这些数据进行辨别，甄别出有用数据与无用数据，以及对数据要进行清理和整合，最后这些数据才能进行分析，也要设计多种分析方法，来找到降水的变化以及其他气象要素和降雨之间的联系。

## 国内外研究现状

人工降水是目前大气水资源开发的主要技术手段。世界气象组织在近期发布的报告中指出，各国应该将人工降水作为水资源综合管理战略的一部分，并建议在各国建立开展云、雾和降水气候学分析，开展跨国外场试验和独立专家评估等，为人工影响天气的研究提供参考和依据。

人工降雨效果检验主要采用物理效果检验和统计效果检验[24]。其中物理效果检验是对降雨技术有效性的定性检验，而统计效果检验是在一定显著性水平下的定量的效果分析。统计效果检验从最初的简单对比，逐步发展出了随机试验的各种统计效果检验方法，科学性和可靠性都有了明显的提升。

以色列在1961-1967年进行的冬季人工降水试验中，使用了随机交叉试验的方案，分别利用秩和检验法和双样本检验法对增雨效果进行了评估，实验结果发现增雨效果可以达到21.6%[12]。在美国怀俄明州开展的WWMPP人工增雨试验中，也采用了区域回归随机试验的方法，并提出了更加精确的效果评估指标[13]。我国于1975至1978年，在福建古田水库开展的小火箭人工降雨试验中，也采用了区域回归随机试验的方法，通过分别拟合试验区和对比区之间的回归关系直线，采用了多个事件检验法与双样本分析，并根据催化云的类型以及天气的不同，进一步进行了分层检验，最终得出在显著性水平0.05的情况下，对锋面天气以及小雨天气的积状云催化效果最佳的结论[14]。北京地区在2001到2004年开展的火箭、高炮催化作业中，分别采用序列试验法、回归分析法、双比分析法与区域回归数值模拟对效果进行了评估，分别在一定显著性水平上给出了增雨效果[15]。随着自动气象站普及和互联网技术的进步，我国在2011年实现了基于.NET的华南区域人工增雨效果统计检验共享平台，该平台采用区域回归的方法进行分析，效果检验的精度得到了大幅度提高[16]。

在目前超过56个国家开展的人工影响天气项目中，不乏有取得可观成果的[17]，如CLIMAX I&II[18]，ISRAEL I &II[19]等。但是，绝大多数项目没有经过科学的定量评估与效果检验，从而不能在较高置信度水平上给出科学的人工影响天气试验的定量效果评估。此外，2003年美国国家研究委员会(NRC)[20]与2007年世界气象组织(WMO)的报告[21]中都指出，目前还没有令人信服的科学依据表明各种形式的播云试验是有效或可行的增加降雨量的手段。即便是目前有记录的结果积极的试验，在经过审查和质疑后，其科学性与准确性依然有待商榷。

我国在开展试验的早期，由于缺乏必要的观测设备和观测手段，以及试验设计缺乏周密性和合理性，尽管开展了数次人工降雨作业，但几乎不能提供任何科学上有力的证据证明降雨作业的作用效果[22]。随着试验设计越来越科学完善，观测设备与观测手段进一步增强，我国对人工降雨试验的评估也逐渐走上正轨。就目前而言，国内外对人工降雨评估的手段和方法都比较单一，都是基于传统的统计方法对降雨效果进行评估[14]。

对于降雨预测问题，使用机器学习与统计手段相结合进行效果评估的方法，在学术界几乎是空白的。只有王伟健等人利用随机森林算法，对单次降雨作业进行了探索性的研究[23]。相比于传统的基于统计学的分析方法，使用机器学习可以更好地解决非线性的降雨问题[11]，提高降雨量的预测精度。使用机器学习还可以使用气象站测得的多种气象数据进行辅助分析预测，从而进一步的提高精度，这在传统效果检验中做不到的。通过引入温度、压强、湿度、能见度等多种气象要素，将降雨这一过程中的物理变化也纳入了考虑的范畴，这与传统的线性回归模型相比更加科学，结果也更加精确。

## 论文结构

本文的主要内容如下：

第一章，首先介绍了天水计划的背景与研究内容，然后介绍了人工增雨的效果评估方法，对效果评估过程中会遇到的问题和挑战进行了分析，对国内外的人工降雨的研究情况进行了介绍，并对效果评估的意义和主要内容进行了说明。

第二章，通过分析收集的试验数据，确定了效果评估的方向，对开发工具进行了选择，介绍了分析过程涉及到的一些关键技术，制定了比较详细的分析方案，并介绍了设计中考虑到的制约因素和成本估算。

第三章，搭建了神经网络对六盘山试验区试验期间假设未开装置的自然降雨量进行了预测，先通过历史数据对模型进行训练和评估，达到一定的精确度，再对试验期间降雨结果进行预测，并分析了实验结果。

第四章，对试验区内部的雨量差异进行了讨论和探究，通过随机对比实验和风向差异研究，通过热图和玫瑰风向图等多种手段，将降雨差异进行了可视化的表现，从时间、地理位置两个方面探究了带电粒子生成装置带来的影响。

第五章，对论文的研究内容进行了总结，指出了在研究过程中存在的不足，指出了下一步的研究方向。

## 课题来源

本课题来源于“带电粒子催化人工降雨雪新原理新技术及应用示范（天水计划）”（项目编号：2016YFC0401000），该项目是国家重点研发计划“水资源高效开发利用”专项在非常规水资源开发领域首批启动项目。

# 机器学习与统计分析结合的评估方案

本章对如何开展效果评估工作进行了讨论，参考了评估工作中使用到的试验数据，分析了评估工作中使用到的技术手段和具体的评估方法，对评估工作的整体方案进行了设计。

## 需求分析

“天水计划”于2020年7月至11月，在乌鞘岭和六盘山地区分别开展了带电粒子催化人工降雨雪外场试验。该项目在两个试验区自建的雨量站点和综合站点，在试验期间收集到了大量的气象数据，此外试验区的国家自动气象站记录了包括试验期间在内的长期历史数据。

为了对外场试验结果进行评估，用来作为带电粒子发生装置有效性的外场试验证据，现需要对已经收集到的数据进行分析评估，针对实验特性和数据特点设计一套科学、有效、定量的外场试验评估方法。

## 可行性分析

“天水计划”于2019年分别在六盘山和乌鞘岭试验区建立了雨量站和综合站，这些站点记录了雨量、风向、风速、气温、相对湿度、空气中离子数浓度等数据。同时，六盘山和乌鞘岭国家气象站提供了十余年的多种气象数据，作为评估的数据支持。通过数据处理和数据分析，为外场试验结果设计一套评估方法是可行的。

人工催化增雨试验的实验结果中最重要的部分就是实验区域的降雨量，我们把实验期间该地的降雨量分为自然降雨量和催化增雨量。我们要求出催化增雨在降雨中所占的比例，所以重点在于得到自然降雨量。传统的气象评估通常使用线性的区域回归方法进行降雨量的预测，但是线性回归的方法只使用了雨量这一统计量，对于气象站统计的其他多种气象要素无法进行分析。且线性回归模型通常研究的是月雨量、季雨量，考虑到外尝试验的时间跨度，这一预测时间精确度不够。另外，自然降雨预测是一个非线性的问题，直接使用线性回归预测的方法效果会不够理想。而气象站有长达十余年的气象数据，比较充足。

基于上述分析，选择引入机器学习的方法来对自然降雨量进行预测。通过构建神经网络模型对六盘山试验区的试验期间自然降雨量进行预测，来得到精度较高的预测结果。这是对外场试验开展效果的宏观检验，通过将模型的预测结果与实际结果进行对比，清晰地说明在试验区降雨量增加与否，增加量多少。

使用可靠性高的神经网络模型对目标研究量进行预测是一个直观、可信度高的方法。但是，神经网络的训练需要时间维度较长的数据。对于“天水计划”自行建立的雨量站点和综合站点收集到的降雨数据，以及其他气象要素，由于数据的时间范围相对较小，导致实验数据比较少，不适宜采用神经网络的方法进行预测。对于这些数据，需要灵活变通，设计相应的科学的评估方法，来检验实验的效果。这些数据相对比较微观，讨论的是一个实验区内部的降雨变化。

## 开发工具分析及选择

开展效果评估首先要进行数据处理，由于收集到的数据比较多，且各种数据格式不同，数据均是来自自动测量仪的原始格式，所以需要进行大量的数据处理和数据清洗工作。另外，需要针对外场试验设计神经网络模型预测自然降雨量，所以选择在数据处理和机器学习学习领域广泛应用的python语言作为开发语言。

对于大量的数据处理和清洗工作，使用基于NumPy（提供高性能的矩阵运算）的Pandas分析工具。Pandas是python提供的，用于数据处理和分析的软件库，它提供用于数据表格和时间序列的数据结构和操作方法。它是一个强大的分析结构化数据的工具集，提供数据清洗、数据挖掘和数据分析等功能。在进行数据处理分析时，也用到了Scikit-learn(sklearn)，它是一个机器学习库，包含有简单有效的数据分析工具。

搭建神经网络来预测自然降雨量时，选择使用TensorFlow框架。TensorFlow是一个开源软件库，是基于数据流编程的符号数字系统，在机器学习领域得到了广泛应用，是一种主流框架。TensorFlow功能强大且非常灵活，同时易于上手。

## 关键技术

### 局部异常因子算法

本次设计处理的所有数据，包括气象站的历史数据和自建站点在试验期间的采集数据均为自动站原始数据，数据量较大，且有较多的异常和缺测情况。在数据送入模型训练之前，需要对收集到的数据进行清洗，剔除其中的异常数据。采用人工清洗数据的方法不能满足评估模型对数据质量的要求，所以引入机器学习的异常检测算法，来帮助得到质量更高的数据。

局部异常因子算法（Local Outlier Factor，LOF）是一种有效的异常检测算法[25]。LOF通过计算一个数值来表征一个样本的离群程度，具体为用一个样本点周围样本的平均密度比上该样本密度，这个比值越大，该点是异常点的可能性也越大。下面介绍LOF算法的相关定义。

1. d(p,o)：p点和o点之间的距离
2. k-distance（k-邻近距离）：在距离p的最近的几个点中，第k近点与p点之间的距离记为k-distance(p)
3. k-distance neighborhood of p（第k距离邻域）：指点p的第k距离内所有点。
4. reach-distance（可达距离）：p点到o点的可达距离，为o点的k-临近距离和p点与o点之间直接距离的较大值。计算见式2-1。

(2‑1)

1. local reachability density（局部可达密度）：点p的局部可达密度的计算方法见式2-2。该值表示p的第k邻域内的点到p的平均可达距离的倒数。

(2‑2)

1. local outlier factor：局部离群因子，计算见式2-3。

(2-3)

局部离散因子表示p点的邻域点的局部可达密度与p点的局部可达密度之比的平均数。若这个值接近1，说明p与邻域点密度差不多；若小于1，说明p点处在一个相对密集的区域，是异常点的可能小比较小；若远大于1，说明p点距离其他点较远，该点为异常点的可能性很大。

### 全连接神经网络

本设计的一个重要问题是设计神经网络，对试验期间目标区的自然降雨量进行预测。由于降雨是一个随机性相对较强的问题，且降雨量的多少与其他气象要素之间的关系不是明确的。所以为了提升预测精度，我们选择非线性拟合能力较强的全连接神经网络进行预测。

全连接神经网络的结构相对比较简单，包括输入层、输出层和隐藏层，其中输入层和输出层都是单层结构。输入层的神经元个数是由数据的特征数决定的，输出层的神经元数量是网络预测的标签数量决定的。层与层之间神经元相互连接，而层内神经元不连接，下一层神经元会连接上一层所有的神经元，这就是全连接的由来。一个仅有一个隐含层的神经网络就能拟合任何一个函数，所以他的非线性拟合能力很强，它的缺点是网络中需要用到较多的神经元，资源消耗比较大。每个神经元的模型如图2‑1所示。



图2‑1 神经元模型

激活函数f本质上是一个非线性函数，它的作用是将线性的输出转换为非线性，来解决非线性的问题。若不存在激活函数，那么神经网络就只能处理线性问题，而实际问题大多是非线性的。神经元中最重要的部分是神经元之间的连接，如图2‑1所示，神经元之间的每个连接都有一个权重w与之对应。权重w的值不是一成不变的，在对神经网络的训练过程中，权重的值一直在变化。从某种意义上说，神经网络的训练就是对权重值的不断调整，直到到达一个最佳状态，使得预测结果最好。调整权重是通过反向传播(Back propagation，简称BP)算法实现的，它通常与一种最优化方法结合使用，如梯度下降法。它的基本原理是，将神经网络模型训练的过程中产生的模型输出与期望输出之间的损失loss，向前传播反馈到隐含层和输入层，各层根据反馈信号，以及输入输出权重和该神经元的阈值，使用梯度下降法计算出参数调整的方法，调整后进行新一轮的训练。不断调整权重，改善模型性能。

## 评估方案的设计

为了使降雨评估更加完善和全面，将工作分为宏观和微观两个方面进行研究。既要完成宏观层面的试验区的降雨预测，也要把握一个试验区内部的微观的降雨差异。所以基于得到的数据的实际情况，通过分析各站点数据的特征要素以及各种数据的时间维度，将主要的工作划分为两个部分，主要内容为：

(1) 建立基于神经网络的降雨量预测模型

对六盘山试验区求出实验期间的自然降雨量的预测值，利用气象学界普遍认为与降雨关联性强的气象要素，根据获取数据的情况及其特征数量选择合适的神经网络或某种特定的机器学习算法建立降雨量预测模型，通过对模型的训练和评估，获得相对准确的目标区域降雨量预测模型。将得到的模型用于预测实验期间的自然降水，通过比较自然降水和实际降水来评估人工增雨效果。具体流程如图2‑2所示。



图2‑2 神经网络预测降雨

(2)建立基于统计分析的人工增雨效果评估模型

对于试验区内部，由地理、时间、风向等因素带来的降雨差异，需要分别进行分析。由于这些数据来源于“天水计划”建立的测量装置，工作时间较短，所以历史数据不足，无法使用神经网络分析。就要建立基于统计分析的效果评估模型，对各天气要素分别设计与之相应的分析方法，力求从微观上说明降雨的差异，完成对人工降雨的效果检验。具体流程如图2‑3所示。



图2‑3 基于统计分析的效果评估

## 设计中考虑的制约因素

在进行效果检验方案设计时，首先要考虑的是能拿到什么样的数据，没有试验数据的支持，再好的检验方案也不会有效果。在收集到所有可得数据后发现，自动气象站有长期数据，适合建立神经网络预测降雨量。而外场试验自建站点收集到的数据时间范围较短，不适合神经网络模型，要进行基于统计的分析。

在进行实际的评估时，考虑到要处理大量的数据，进行数据分析，以及可视化等相关工作，编程语言选择python是一个不错的选择。虽然之前没有学习过python，但是考虑到其他语言完成该类工作很困难，且python相对易学，所以选择先对语言进行学习再完成实验任务。

机器学习任务要选择一个合适的框架，由于之前没有学习过机器学习相关知识，所以选择了官方文档和其他教学资源比较丰富的Tensorflow。且该项目在2020年进行前序试验设计时，使用了Tensoeflow作为部分机器学习内容的研究工具，所以使用相同的框架对于我理解之前的工作比较方便。

本此课程设计属于对试验结果的评估，目的是做到客观、可靠、定量的试验评估，方便后续试验的开展，以及对相关研究提供参考，所以对社会、健康、安全、法律、文化及环境等方面不会带来不利影响，但要尽可能得提高评估的可靠性，为后续研究提供参考。

## 成本估算

本次设计的难点在于面对的是一个全新的问题，不论是人工降雨效果评估还是神经网络、机器学习算法的实现都是之前本人从未涉及的领域。所以，了解项目、学习一般人工降雨的试验方法和效果评估，学习python和tensorflow，学习数据处理方法、可视化方法都花了很大的时间和经历，有很高的学习成本。此外，由于没有现有的完备科学的评估方法，所以要对评估方案进行研究设计，尽可能地做到多方面的、定量的、可信的评估。最后，在进行实际工作时，神经网络的搭建、评估以及改善等等都是开发过程中花费的代价。

使用软件开发成本估算模型Putnam，计算得到的工作量K为0.3374人年，估算开发时间td为0.3300年。

## 本章小结

本章从需求分析入手，分析项目收集到的实验数据，对实现分析的编程语言、框架等进行了选择和确定。针对不同类型和时间维度的数据指定了不同的分析方法对实验结果进行评估，确定了基于统计分析和机器学习相结合的总体评估思路，设计了具体的研究流程。

# 基于全连接神经网络的自然降雨预测

本章分析处理了六盘山气象站十余年的气象数据，搭建了全连接神经网络进行自然降雨量的预测，通过评估验证了模型的可靠性，并对实际问题进行了应用。

## 六盘山自然降雨预测模型说明

六盘山试验区内的六盘山自动气象站具有丰富的实验数据和历史数据，有利于使用神经网络，进行实验期间降雨量的预测估计。

首先对神经网络的类型进行选择。在深度学习中，循环神经网络(RNN)和卷积神经网络(CNN)都是十分流行且强大的神经网络结构。循环神经网络是专门处理序列数据的一种神经网络，目前主要运用在自然语言处理、语音识别等方面。而卷积神经网络主要运用于图像处理，它考虑了输入值（像素）的空间分布，对图像有更好的识别效果。所以，虽然两者是十分热门的网络结构，但是对于我们预测降雨来说都不合适。

而全连接神经网络(DNN)结构虽然比较简单，但是非线性拟合能力比较强，几乎可以拟合任何一个函数。当然它也有缺点，如难以处理高维度数据，所用神经元比较多等。但是考虑到气象站采集到的天气要素种类以及数据规模相对来说比较小，且不包含高维数据，所以可以忽略全连接模型的这些缺点。综合分析，使用全连接神经网络进行降雨量的预测是一个不错的选择。

使用历史数据作为神经网络模型的输入，对网络进行训练，来预测实验期间，如果不加人工干扰的情况下的自然降水量。通过与实验期间的实际降雨量进行对比，分析带电粒子装置开启后带来的实际效果。整体上看，我们的神经网络模型是一个回归模型。

## 数据处理与特征选取

在建立神经网络预测模型之前，首先要完成历史数据的筛选和清洗，以及进行特征的选取。由于气象站数据为自动站自动记录的，存在缺测、异常等多种干扰，所以首先对拿到的原始数据使用LOF算法进行异常检测，使用scikit-learn包的neighbors模块的LocalOutlierFactor，作为LOF算法的具体实现。主要通过调节参数n\_neighbors，也就是一个点p周围的邻居数量，以及用来判断是否异常的LOF阈值threshold，来进行异常检测。当参数调节到结果合适时，剔除掉异常数据。

### 数据整合

六盘山气象站拥有的气象数据时间维度为2008年1月1日至2020年11月30日，它的采集频率为一小时一统计，采集要素包括气压、温度、最高气温、最低气温、露点温度、相对湿度、最小相对湿度出现的时间、水汽压、过去1小时降雨量、两分钟平均风向、两分钟平均风速、最大风速的风向等多种气象数据。

由于降雨过程一般情况下是持续数个小时甚至几天的，所以每小时的降雨数据对我们来说意义不大。同时，对每小时的降雨预测难度非常大，所以需要对数据进行整合。按照一天为单位，对数据进行重采样，降低数据的时间分辨率。

为了方便对数据进行操作，使用了pandas中的一种重要的数据结构dataframe来进行数据处理。dataframe是一种表格型数据结构，它与数据库的二维表格比较相似，dateframe既有行索引(index)也有列索引(column)。通过读取数据，转化为dataframe格式，我们可以很方便地对数据进行增删改查，且有很多功能强大的内置函数，可以方便地对数据进行操作。dataframe的结构示意图如图3‑1所示。



图3‑1 dataframe结构示意图

来自自动气象站的数据按照一天为单位保存为不同的excel文件。使用pandas读取文件，转换为dataframe数据结构，就可以使用append方法方便得对不同文件的数据进行连接。而且，利用dataframe的重采样方法resample，就可以对数据进行整合，但是resample方法对所有的数据列均采用相同的采样模式，如全部求平均值或全部求和，这一点不满足我们的要求。

整合中要对不同的数据类型采用不同的方式。对于气温、气压、湿度、可见度等连续型标量，为了保持其物理意义，取它们24小时统计量的均值作为日均值。对于降雨量这种离散型标量，由于每小时统计的是过去1小时收集到的数值，所以将这些数据进行相加处理。对于风向等矢量数据，直接对角度求代数平均或者进行矢量运算都是不符合物理意义的，所以对于每天的24个风向信息，要找出其中的主导风向。

基于以上要求，在使用dataframe.resample方法时，对所有信息进行求均值的处理，对雨量再乘24代表日雨量。忽略风向风速的生成结果，对原始风向风速数据进行单独处理。将风向的360度分为16个区域，将每天的24个风向数据整除16，得到风向的区域分布，保存为列表。通过寻找列表中的众数，来找到该日的主导风向区域，计算区域内的几组风速的平均值，来得到主导风速。用单独计算的风向风速数据替换resample方法计算得到的风向风速结果。

### 数据标准化处理

由于各个数据的量纲和数值范围不同，如果将原始特征直接放入网络中训练，数值较小的特征对网络的影响也会较小，实际上各个特征对降雨的影响程度是不清楚的。而且，神经网络中的激活函数是有非线性范围的，数值较大的特征进行训练会影响网络的初始化，甚至会影响网络的收敛性。所以需要在进行训练之前，要对数据进行标准化处理，如式3-1所示进行计算。

(3-1)

式中是标准化处理后的数据矩阵，*D*是未经处理的原始数据矩阵，是训练集数据矩阵各个特征的均值，是训练集数据矩阵各特征的标准差。对于神经网络划分的训练集、验证集和测试集，进行数据标准化时，采用的均值和标准差参数均是训练集数据的，因为我们假设认为训练集的均值和标准差代表了这项气象特征在该地的分布规律。

### 特征选取

自动气象站记录的数据中包含的特征对我们来说并不都是有用的，其中包含了相关特征和冗余特征。相关特征对于我们的学习任务是有帮助的，它和降雨量存在内在联系。但是冗余特征不仅对我们求解问题没有帮助，而且会浪费资源，延长收敛时间。我们在选取特征时，要选择与降雨关系密切的气象要素，如气温、湿度等，对一些明显是冗余信息的不进行选取，如该地的区站号、站点等级、该地海拔等。

选取特征的数量不宜过多也不宜过少，特征过少的话，神经网络难以从已有的特征中获取降雨规律的信息；特征选择过量的话，会人为地给神经网络引入噪音，使非相关因素影响了网络的效果。我们在进行特征选取时，先剔除了一些明显的冗余特征，然后将其他特征提取出来进行训练。再依次剔除一个特征进行训练，如果预测结果没有变差，就说明该项特征对学习内容没有帮助，除去后能够节省资源。因为记录的气象要素数量不是很多，所以这种方法不会尝试很多次就能得到结果。最终决定采用的特征包括时间、气压、温度、湿度、风向、风速。

## 预测模型的建立

### 模型数据说明

为了得到一个有效的模型，可靠地预测出2020年7月到11月的降雨数据。我们需要对设计的神经网络模型可靠性进行检验。因此，我们先不预测2020年实验期间的降雨量，而是使用历史数据来对未开展试验的一个年份进行自然降雨预测，通过比较预测数据与实际降雨数据之间的差异，来评价该模型是否有效，再使用该模型预测2020年7月到11月的自然降雨量。

所以我们选择将2008年到2016年的数据作为训练集， 2017年和2018年的数据作为验证集，2019年的数据作为测试集。选择全连接神经网络作为网络结构，选取的特征为月份、气压、温度、每小时降雨量、风速、风向。

### 神经网络的实现

首先我们构建神经网络的主体结构，建模时采用序列化建模，具体操作为先实例化一个tensorflow.keras.Sequential类，来对model进行初始化。接着使用model.add()方法为网络添加层，由于我们的数据维度较低，所以只隐藏层的网络结构只需要设计一层。我们的神经网络选择使用全连接神经网络，层结构为全连接层，需要调用tensorflow.layers.dense()方法来创建全连接层。模型的训练特征为6个，所以输入层的神经元个数设置为6个；根据输入神经元的个数，相应的设置隐藏层神经元为25个；模型输出为预测的降雨量，所以只有1个神经元作为输出层。设置激活函数为sigmoid，它将变量映射到0和1之间。

在设计完网络主题结构后，还需要对模型的学习过程进行配置，这一步骤通过model.compile()方法进行实现，它有三个参数，分别为优化器optimizer，损失函数loss和评估标准metrics。优化器optimizer选择使用广泛使用的Adam优化算法，来加速模型的学习速度和学习效果。损失函数loss是网络模型试图最小化的目标函数，使用广泛使用的均方误差mse是一个稳妥的选择。metrics是用于监控训练的评估标准，我们选择使用均方误差mse和平均决定误差mae作为评估标准。

搭建完模型后还要对数据进行梳理，清洗整合过的数据不能直接的用于训练，先将数据分为训练集、验证集和测试集。由于我们之前的处理都是基于dataframe数据结构的，而模型的数据输入格式应该为dataset格式，所以要先处理dataframe数据来生成dataset。

首先要对三个数据集进行数据标准化处理。再设计函数dataframe\_to\_dataset来对dataframe进行转换，该函数的具体工作为：先将数据集中的降雨量数据列取出作为模型的label，再将剩余的数据传入dataset.from\_tensor\_slices()函数，按照设置好的batch\_size（一次训练所选取的样本数）来对传入的datafram数据进行切分，生成dataset格式数据。为了使训练时每批次训练用到的数据集不同，需要将数据进行打乱，这就用到了dataset.shuffle()函数，对训练集数据进行打乱处理。为了增大数据量，使用dataset.repeat()函数将数据集进行复制扩充。至此，函数功能完成，生成了dataset数据集，用于模型输入。

数据处理完之后，开心进行模型训练，使用model.fit()函数进行模型训练。该函数需要配置较多参数：设置训练轮数epochs，来控制训练模型迭代次数。但是迭代次数不是越多越好，当迭代次数过多时，会出现过拟合现象，也就是模型过度地“迎合”训练集的数据特点，导致模型在训练集上表现很好，但是在验证集上的效果越来越差。为了尽量避免过拟合的出现，设置参数callbacks来查看每个epoch开始和结束时的状态，具体为调用callbacks.EarlyStopping()方法作为callbacks的具体实参。设置EarlyStopping对验证集的loss值进行监控，通过设置patience的值来规定在多少个epoch内验证集的loss值没有提升就停止训练。还要设置参数batch\_size，来作为每次梯度更新的样本数。再使用训练集样本数和验证集样本数分别除以batch\_size，来对steps\_per\_epoch和validation\_steps进行赋值。还要传入处理好的训练集和验证集的dataset格式数据。至此，模型训练已经完成。

### 模型的评估

因为2019年没有进行人工降雨，所以理论上来说，好的模型的预测结果应该与实际降雨量基本吻合，为了定量地说明模型的预测结果的可靠程度，引入决定系数R2对预测结果进行评估，它的定义见式3-2。

(3-2)

决定系数R2反应了，因变量的变化波动中有多大的比例能够被自变量的波动所解释。由公式可以看出，预测值和实际值的差异越小，R2的值越接近1。按照这一评估手段对模型预测结果进行检验。

### 模型初步结果与改进

按照上述的模型搭建方法，搭建一个全连接神经网络，根据划分好的数据集和选取的特征数据进行训练。图3‑2给出了神经网络迭代过程的损失曲线。

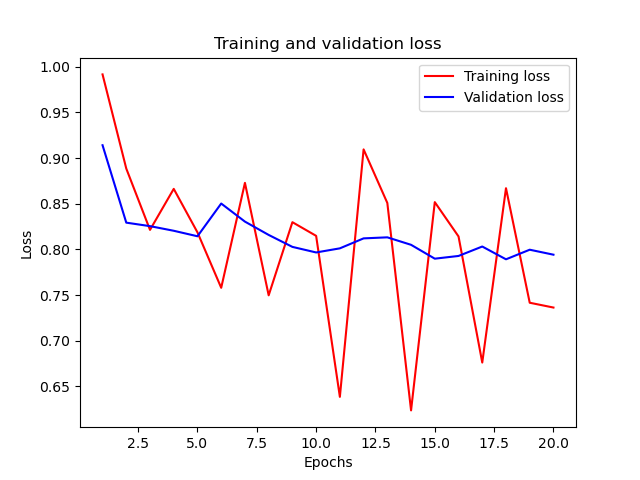


图3‑2 训练过程中的loss曲线

可以看到不论是训练集还是验证集，对应的loss曲线都没有随着训练轮数的增加而不断下降，训练集的loss值在上下振荡，而验证集的loss值虽有下降，但是下降幅度较小，一直维持在较高水平，这两者说明模型不仅在验证集中表现不好，甚至在训练集中都没有达到较好的效果，我们的模型准确度不够。我们使用R2对测试结果进行定量分析，将2019年的预测降雨量和实际降雨量代入公式进行计算，得到计算结果为0.21。这说明我们的预测量与实际量相差很大，当前模型效果很差。我们分析对2019年降雨量的预测结果，尝试从中寻找原因，雨量的预测值和实际值如图3‑3所示。

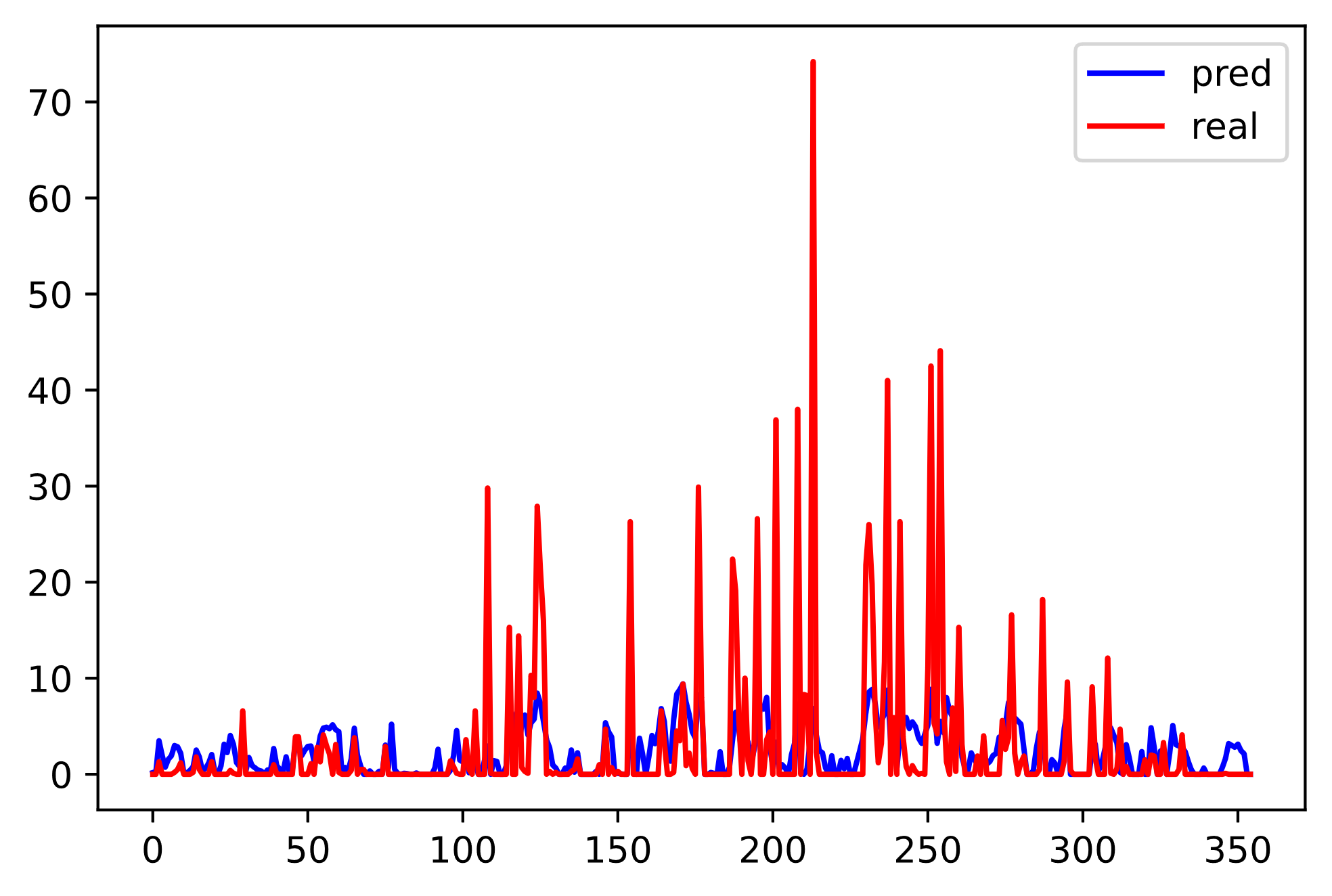


图3‑3 预测2019年日降雨

红色曲线代表了2019年的实际日降雨量随日期的变化曲线，蓝色曲线代表了模型预测的日降雨量随日期的变化曲线。x轴表示2019年的第几天，y轴表示该日的降雨量。

由于我们在六盘山地区开展试验的时间为2020年的8月至11月，所以我们要着重观察2019年同期的降雨预测效果。发现2019年8月到11月的真实降水量为508.20mm，使用模型预测的降水量为383.06mm，预测值与真实值相比，误差达到了32.67%。再观察图3‑3展示的降水量对比图像，我们可以发现，每年降水的时间分布是不均匀的，春冬季节降水量较小，大概在每年的第一百天左右开始，降水量突然增大，在200天左右达到峰值，每年的第100天到第250天之间一直维持在较高水平。但是模型预测值随时间的变化并不明显，预测值一直在较低水平。对于夏季雨水的增加，模型不能及时地做出响应，导致不能预测到降雨量大的周期的到来。

因为降水存在季节差异，在雨量少的季节中，日雨量中有较多零值，这样网络可能会偏向于使用零值来作为网络的输出。为了凸显出降雨的季节差异，以及减少大量零值对神经网络的影响，尝试将日雨量作为预测值的标准进一步修改，改为预测月雨量。与之前每小时雨量整合为日雨量类似，将日雨量按照月份累加，得到月雨量。将气压、湿度等气象特征求平均值作为每月的特征值。将数据重新处理后，对网络模型中的batch\_size、size\_train,、size\_val等参数进行了相应的修改。使用新的数据，重新对修改后的模型进行训练，预测出来的月降雨量的统计结果如表3‑1所示。

表3‑1 2019年预测结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际月平均降雨量 | 预测月平均降雨量 | 误差 |
| 2019年1-7月 | 71.31mm | 54.77mm | -30.20% |
| 2019年8-11月 | 124.97mm | 82.67mm | -51.17% |

从表中可以看到，月平均降雨量的预测结果与实际值相差较大。同时，预测值和真实值之间的决定系数R2的值为0.55，这也验证了两者之间存在较大差异。

分析月降雨量的预测结果，发现预测值除了在2月左右有稍许偏高外，其他时间都低于实际降雨量。而且我们发现，七八月份实际降雨量会出现徒增，但是模型并没有发现这一规律。考虑到模型用到的特征值，如气温、气压这些，不会像降雨量一样发生明显的突变。我们猜想8月左右的强降雨天气不能被该模型识别，也就是说当前的特征的不足以让模型学会何时会发生强降雨。因此我们决定加入其他特征来进行训练，希望这些特征的加入，可以让模型“嗅”到强降雨到来前的味道。

因为我们对降雨的机制不了解，我们已经将与降雨有关联的天气数据都加入进行了训练，所以没有哪个天气的特征值加入会让模型大幅度提升。实验之前进行方案设计时，设计了区域回归实验，也就是使用其他没做实验的站点作为参考，来预测实验地区的自然降水量。而我们当前的工作是用实验站点的历史数据来进行预测的，我们可以将区域回归实验融入进来，将其他站点的降雨量作为特征值。要求这些地方不但没有受到人工降雨的影响，而且与实验地区相距不远，且气候条件非常相似。将这些加入作为特征值进行训练，可以让模型知道8月份左右会有强降雨发生，从而改善模型的效果。经过筛选，我们选择加入隆德气象站和泾源气象站同时期的降雨数据作为模型的特征值，重新训练模型。得到预测的降雨数据如图3‑4所示，训练过程中的loss曲线如图3‑5所示。

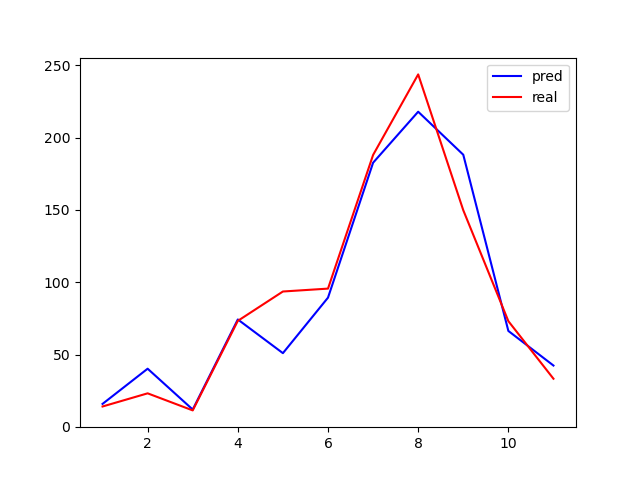


图3‑4 预测月降雨曲线

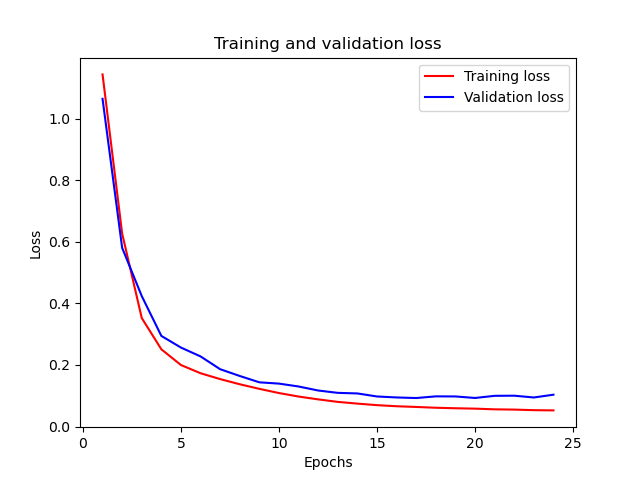


图3‑5 训练过程中loss曲线

从雨量图中可以看到降雨量吻合的效果比较好，而且图像的loss曲线平滑下降。将预测的结果进行统计，如表3‑2所示。

表3‑2 2019年月雨量预测结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际月平均降雨量 | 预测月平均降雨量 | 误差 |
| 2019年1-7月 | 71.31mm | 66.47mm | -6.79% |
| 2019年8-11月 | 124.97mm | 128.70mm | +2.98% |
| 2019年1-11月 | 90.83mm | 89.11mm | -1.89% |

可以看到误差值都非常小，模型效果比较理想。注意到2019年8月到11月的预测值是大于实际值的，这说明我们的模型不会得到较低的预测值来使实际值看起来虚高，来迎合我们想要的实验结果。如果我们将模型应用于2020年，发现实际降雨量比预测值更大，这是比偏高的预测值还要大，更有力地说明了我们实验的有效性。

在改善模型的过程中，我们发现，固原、泾源气象站降雨量的加入，会让模型得到大幅度的提升，但是再添加更多其他气象站的数据不会使预测精度有较大的提升，猜想这是由于其他站点距离六盘山气象站较远，且降雨量差距较大造成的。现在，我们重新着眼于日雨量，希望通过加入两个气象站降水量数据的模型可以成功地预测出日雨量，来对模型的预测结果的时间精度进行提升。对模型的训练数据进行相应的调整，将月雨量修改为日雨量，重新调整参数并进行训练，首先预测2019年的自然降雨量。结果如表3‑3：

表3‑3 2019年日雨量预测结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际日平均降雨量 | 预测日平均降雨量 | 误差 |
| 2019年1-7月 | 2.35mm | 2.29mm | -2.55% |
| 2019年8-11月 | 3.50mm | 3.71mm | +6% |
| 2019年1-11月 | 2.81mm | 2.86mm | +1.78% |

训练过程中的loss曲线如图3‑6所示，可以看到训练集和测试集的loss曲线均下降明显，并最终保持在一个较低值。进一步地，2019年的日降雨量的预测值和实际值的R2值为0.91，比较接近1。这说明该模型对日雨量的预测是比较精确的。

对预测结果和实际降雨量进行具体分析，发现2019年全年日平均降水量的预测值和实际值相差1.78%，理想情况下该值应该接近0，但是这个数值不大，说明该模型还是比较合适的。上半年的预测值比实际值低2.55%，这个数值较低，说明上半年的预测比较准确。下半年的预测值比实际值大了6%，这个值相对来说比较大，但是考虑到预测降雨量偏大，这说明该模型不会输出较低的预测值来使实际值看起来虚高。如果对2020年进行预测，发现实际降雨量比预测值更大，这是比偏高的预测值还要大，更有力地说明了我们外场试验的有效性。

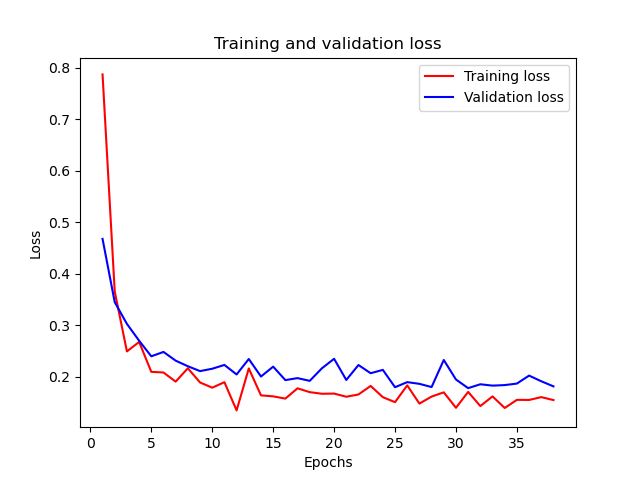


图3‑6 训练过程中loss曲线

## 模型应用

使用改进后的模型，对六盘山气象站2020年的月降雨量和日降雨量分别进行预测。训练集为2008年到2017年的数据，2018年和2019年的数据作为验证集，2020年的数据作为测试集，来对2020年降水量进行预测。

### 预测月雨量

模型输入的雨量格式为月雨量，对模型进行训练，训练过程中的loss曲线如图3‑7所示。

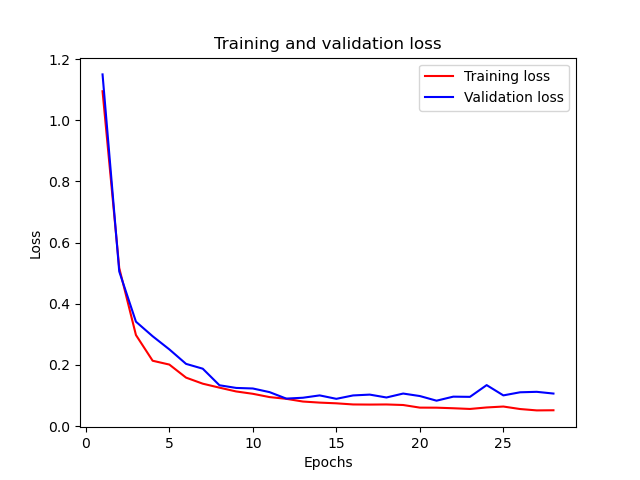


图3‑7 训练过程中的loss曲线

可以看到训练集和验证集的loss曲线都不断下降，说明模型效果较好，对2018和2019年的预测结果是比较满意的。预测结果如表3‑4所示。

表3‑4 2020年月雨量预测结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际月平均降雨量 | 预测月平均降雨量 | 误差 |
| 2020年1-7月 | 64.96mm | 64.82mm | -0.23% |
| 2020年8-11月 | 122.30mm | 103.99mm | -15.79% |
| 2020年1-11月 | 85.81mm | 79.03mm | -7.9% |

可以看到全年的降水量实际值是比预测值大的，在保证预测值是可靠的前提下，这个结果是我们希望看到的。

观察降雨增量在上下半年的分布，戴兴安1月至7月预测值和真实值仅相差0.23%，因为这期间没有进行实验，所以预测值和真实值理论上来说应该是相同的，这一数据再次证明了模型的可靠性。下半年预测值比实际值小了15.79%，这说明下半年的降雨受到了除自然降雨以外的影响，也就是受到试验影响。结合之前预测2019年下半年降雨的情况，该模型会对下半年降雨量的预测值偏高，预测值比实际值高6%左右。但是在2020年，预测值与实际值低了15.79%，所以考虑到模型的误差，这个实际值增加量比15.79%更大，为21%左右。

### 预测日雨量

模型输入的雨量格式为日雨量，对模型进行训练，训练过程中的loss曲线如图3‑8所示。

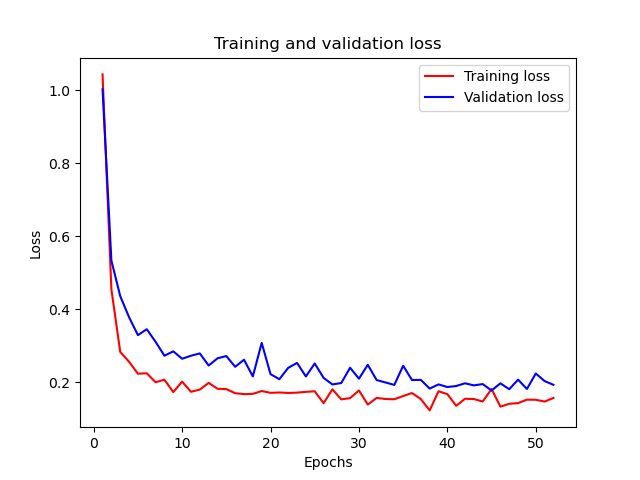


图3‑8 训练过程中loss曲线

可以看到训练集和验证集的loss曲线均不断下降，保持在了较低水平。模型计算结果如表3‑5所示。

表3‑5 2020年日雨量预测结果

|  |  |  |  |
| --- | --- | --- | --- |
|  | 实际日平均降雨量 | 预测日平均降雨量 | 误差 |
| 2020年1-7月 | 2.14mm | 2.25mm | +6.54% |
| 2020年8-11月 | 4.01mm | 3.51mm | -12.47% |
| 2020年1-11月 | 2.82mm | 2.71mm | -3.9% |

结合2019年的结果（见表3‑3），可以看到2019年的上半年预测值小了2.55%；但是在2020年，预测值大了6.54%。若只看实际降雨量会发现，2019上半年的实际日降水量平均值为2.35mm，2020年为2.13mm。这两组数值都表明，2020年上半年的雨量比2019年是有所降低的，实际降雨量大致降低9.36%。也就是说，在只有自然降雨的影响下，2020年的上半年比2019年的雨水偏少，这样导致了2020年的预测值相对来说会有些偏大。

从表3‑5中我们可以看到，2020年下半年的预测值低了12.47%。同时，我们注意到预测2019年时，下半年预测值大了6%（见表3‑3）。也就是说，该模型得到的下半年的预测值会偏高6%左右。假设模型在2020年下半年也有该误差，则扣除误差，增量大概为12.47%+6%=18.47%左右。

另一方面，只参照实际降雨量来分析，2020年下半年的实际日降雨量为4.01mm，而2019下半年的实际日降雨量为3.50mm，2020年相比增加了12.47%。而根据上半年显示2020年雨水相对较少，上半年减少了9.36%，假设下半年也会减少9.36%。但是2020年下半年实际降雨量（月均值为4.01mm）相比2019年（月均值为3.50mm）却增加14.57%，这表示降雨效果达到了9.36%+14.57%=23.93%。两个方面的证据都表示降雨的增加会达到20%。

我们再来分析一下2020年的降水量的残差图，如图3‑9所示，该图显示的数据是用真实值减去预测值得来的。如下所示。可以看到，上半年的残差变化比较小，维持在0值附近，下半年的起伏比较大，而多为正值，这说明下半年的真实降雨量是普遍大于预测值的。这是实验带来的结果，这一图非常直观。

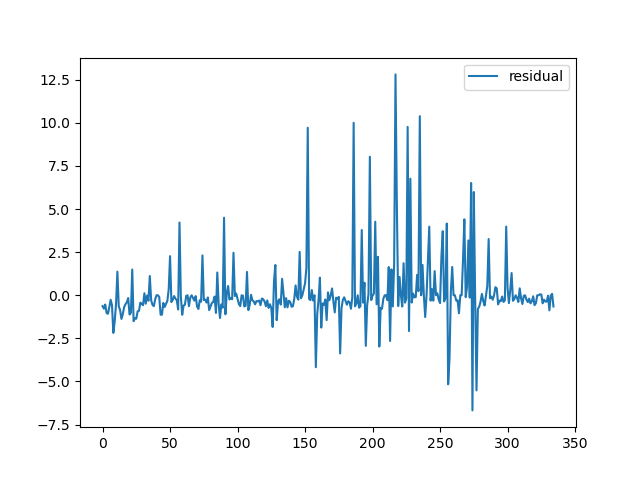


图3‑9 2020年日降雨量残差图

## 本章小结

本章搭建了一个全连接神经网络模型，先通过对六盘山地区历史数据降雨量进行预测来不断改善模型精度，后对试验期间的自然降雨量进行了预测，与实际降雨量进行对比发现，实验期间降水增加量达到了20%左右，符合外场试验的预期效果。另外，在改善模型预测精度的过程中发现，如果要预测出一个气象站的降雨量，其他气象站的降雨数据非常重要。仅仅从气压、温度、湿度、风等当地的自然气象条件很难预测出降雨的强弱。

# 基于统计分析的人工增雨效果评估

本章是针对“天水计划”自建站点收集的数据所作的评估分析，这些数据由于时间周期较短，不符合神经网络对数据量的要求，所以采用基于统计分析的方法，对随机对比试验的结果进行分析，并对实验区内部各雨量站的差异与风向建立联系，采用可视化手段将实验结果清晰地展现出来。

## 效果的统计检验

统计检验的主要内容是显著性检验。对我们的评估而言，若实际降雨量与估计的降雨量之间存在差异，那么就要通过显著性检验，指出由于自然原因造成这个差值的可能性是多少。如果可能性很大，我们就不能认为人工降雨试验带来的的影响显著改变了降雨，也就是说我们的试验效果是不显著的；如果可能性很小，那么我们就有较大的把握认为人工降雨试验是有效的，也就是说试验效果显著。上面提到的可能性大小就表示了显著性水平，用α表示。

显著性检验又称为假设检验，人工降水效果评估中常用的检验方法是t-检验法。它也称为Student t检验，适用于样本较小，样本标准差未知的正态分布。使用t检验法时，要求我们人工降雨试验的影响只会对总体平均值进行改变，而不对总体方差有影响。满足这两个条件时，t-检验的依据才成立。

## 随机对比试验

### 随机对比试验说明

“天水计划”于2020年在乌鞘岭地区和六盘山地区开展了随机对比实验。将两个试验区的带电粒子发生装置按照随机规则进行开关，每次开关周期为一天，装置开关次数比例为1:1，乌鞘岭地区从2020年7月开始实验，六盘山地区从2020年8月开始实验，两个试验区都在2020年11月结束实验。装置开启的日期作为随机对照试验的实验组，关闭的日期作为随机对照试验的对照组。“天水计划”在两个试验区分别安装了8个雨量筒，用来记录实验期间该地的雨量情况，通过清洗雨量筒的数据，得到该地实验期间的降雨。通过统计实验组和对照组的降雨量，来计算出实验组相比于对照组的增加量，同时还要对两组数据进行显著性检验，来推论差异发生的概率。

### 随机对比试验统计分析

对乌鞘岭和六盘山地区分别进行随机对比试验的统计分析，具体操作为先处理数据，分别得到试验区的雨量站数据，再对实验组和对照组的降雨结果进行对比，计算出实验组的降雨增加量，并进行t检验。

#### 数据处理

我们得到的原始数据为各个站点每分钟的信息，按照站点和日期以excel格式保存。与之前处理历史数据相似，我们使用pandas对各个站读取数据，转化为dataframe格式进行操作，使用append方法将一个站点的所有数据整合到一个dataframe表中。因为我们的随机对比试验的时间单位是天，所以使用resample方法对每分钟的降雨数据进行重采样，得到每天的降雨数据。对各个站点分别进行处理后，按照日期为关键字，进行dataframe的合并，就得到了各个站点的每日的降雨数据，对每天中各站点雨量进行相加，得到了该实验区该日的降雨量。

#### 乌鞘岭统计结果分析

完全按照随机对照试验的原则，乌鞘岭试验区实际实验组数据为14天，对照组数据为13天。将这些数据清洗并统计之后，得到实验组合计14天的日降雨总和为176.74mm，日均降水量12.62mm，而对照组合计13天的日降雨总和为129.12mm，日均降水量为9.93mm，则由公式4-1，如下

(4-1)

可计算出实验期间的带电粒子催化降雨的效率为27.1%，达到了项目要求的20%的增雨效率。式中，A为增雨效率，R0为平均自然降雨（即对照组平均雨量），R1为平均催化降雨（即实验组平均雨量）。

我们实验组和对照组的差别理论上只有装置开启与否，所以我们可以把对照组作为自然降雨量，而实验组就是催化后的降雨量。为了排除自然环境的影响，指出自然影响做成这个差异的可能性有多大，针对这两组数据进行t检验。因为实验组的均值偏大，所以选择单尾检验，选择双样本等方差假设作为t检验的类型，得到的t检验结果为0.3743。这个结果表示，实验组和对照组之间的差异是在自然条件下发生的概率有37.43%，也就是说两者的降雨量不属于同一分布的概率为62.57%。

进一步地，由于随机实验采取以天为周期随机开展实验的方法，尽管其置信度较高，但其实验数据偏少且不连续，无法体现出装置长时间开启的增雨效果，故而继续对长期实验进行统计。选取研究时间为，将2020年8月24日到2020年8月30日，其间9月15日至9月22日期间由于装置闪络损坏未能正常开启，将这个时间段作为对照组，其余作为实验组进行相同的统计与分析。

结果显示实验组共30天的日降雨总和为707.23mm，日均降水量23.57mm，而对照期间合计8天的日降雨总和为83.19mm，日均降水量为10.40mm。计算出实验期间的带电粒子催化降雨的效率为126.63%，远远超出了项目要求的20%的增雨效率。对这一数据同样进行t检验，得到结果为0.2293，也就是大概有22.93%的概率说明实验组和对照组的均值相同，不属于同一分布的概率为77.07%。

将上述所有的实验组数据合并在一起，所有对照组数据合并在一起，得到实验组共44天，对照组共21天。实验组的总降雨量为883.97mm，日平均降雨量为20.09mm。对照组的总降雨量为212.3mm，日平均降雨量为10.11mm。相对于对照组，实验组的日均降雨量高了99%，远高于要求的20%。t检验结果为0.1521，也就是大概有15.21%的概率说明实验组和对照组的均值相同，也就是说均值不同，不属于同一分布的概率为84.79%，这一值相对还是很高的。

#### 六盘山统计结果分析

与乌鞘岭单电极随机对比试验相似，需要选取实验组和对照组，通过比较雨量的增量，以及进行显著性检验来分析降雨效果。但是与乌鞘岭单电极不同的是，六盘山地区考虑到试验周期，对试验方案进行了调整，在实验期间，所有装置功率全开，这样导致了随机对比试验没有了对照组，所以选择了没开装置的7月份的降雨数据作为对照组。

由于装置始终开启，能获取到的没有装置开启的时间只有几天。再考虑12小时的缓冲时间，则8、9月没有可以用来做同等时间尺度下对比的数据。故而选用7月的雨量数据与8、9月实验期间的进行对比。

经过数据的清洗和整理，发现7月雨量缺测比较严重，1号与8号雨量筒基本没有采集到数据，其余雨量筒只有前半月有数据。所以不再考虑1号和8号雨量筒，将2号到7号雨量筒的数据进行加和，作为六盘山地区的降雨数据。除去缺测的情况，在8月和9月中清洗出了54天作为催化降雨的实验组，选择7月1日到7月13日共13天的数据作为对比区。

根据统计结果并进行对比可知，实验期间合计54天的日降雨总和为1100.89mm，日均降水量20.39mm，而对照期间合计8天的日降雨总和为234.24mm，日均降水量为18.02mm。可计算出实验期间的带电粒子催化降雨的效率为13.15%，相比项目要求的20%的增雨效率尚有一定距离。对两组数据进行t检验，结果为0.4184，也就是大概有41.84%的概率说明实验组和对照组的均值相同，也就说两个均值不同，不属于同一分布的概率为58.16%。

### 可视化分析处理

4.2.2节中，将实验组和对照组两组降雨数据进行对比，并进行t检验，说明实验组和对照组之间的差异。这一方法可以清晰地看到带电粒子发生装置开启与否对整个乌鞘岭实验区带来的影响。但是对我们实验来说，试验区雨量是由8个雨量站的相加得到的，这8个雨量站的位置以及各自的降雨量存在差异，我们也要分析实验区内部的降雨差异。为了凸显出各站点在试验期间的变化，也就是试验区内不同位置的降雨增加差异，使用热图(heatmap)将试验区内的降雨增加情况可视化地表达出来，通过颜色对比可以直观地看出增加的降雨量的地区差异。

#### 热图和插值

热图(heatmap)是一种常见的可视化手段，它的色彩变化丰富明显，表达信息生动形象，所以它在数据分析领域得到了广泛的应用。热图的生成可以使用python的seaborn包来完成，seaborn是基于matplotlib的图形可视化python包，能够高度兼容numpy与pandas数据结构，通过它可以地绘制heatmap。

绘制heatmap的一个主要方法是seaborn.heatmap()，该方法要求的输入数据为矩阵数据集，如numpy的数据，或是pandas的dataframe数据。而我们的数据只是8个雨量点的数据，为了得到矩阵数据，我们将8个雨量站按照坐标位置标注在地图上，按照经纬度的划分，将试验区域划分为网格，将雨量站的雨量值赋值给所在的网格。这还有一个问题就是，雨量站在地图上分布相对比较稀疏，划分完网格之后，有较多网格没有数据。因为网格的数据表示该地的降雨量，所以将其近似假设为0是不合适的，这就要使用插值法对网格进行赋值。

插值法是根据已知的离散数据，拟合出一个连续函数，来使该函数经过所有的已知的离散点，并用函数值来近似求出缺失的离散值。python的scipy包中有插值模块interpolate，其中有多种插值方法。我们的目标是基于网格数据插值，是一个二维插值问题，选用了griddata()方法进行插值，该方法有强大的处理多维散列取样点并进行插值运算的能力。

griddata()方法有5个主要参数， points参数表示所有已知点的坐标，我们使用numpy.array()方法，将所有雨量站点的实际地理坐标传入。values代表各个点的函数值，将雨量站点的实际雨量传入。xi表示插值的范围，也就是要对什么区域的数据进行插值处理，它也是一个二维表结构，我们可以使用numpy.mgrid()方法，通过设置横纵坐标的上下限以及步长来生成网格用于数据的填充。method表示插值方法，有nearest、linear和cubic三种，分别表示阶梯插值、线性插值以及二阶、三阶B样条曲线，经过尝试，发现对于我们的问题cubic的插值效果更好。fill\_value表示在无数据时网格中填充的值，为了在后续生成的heatmap图像更加清晰，填充值为None。

#### 乌鞘岭可视化分析

按照上述分析，将乌鞘岭各个雨量站点按照实际经纬度，标注起来，如图4‑1所示。

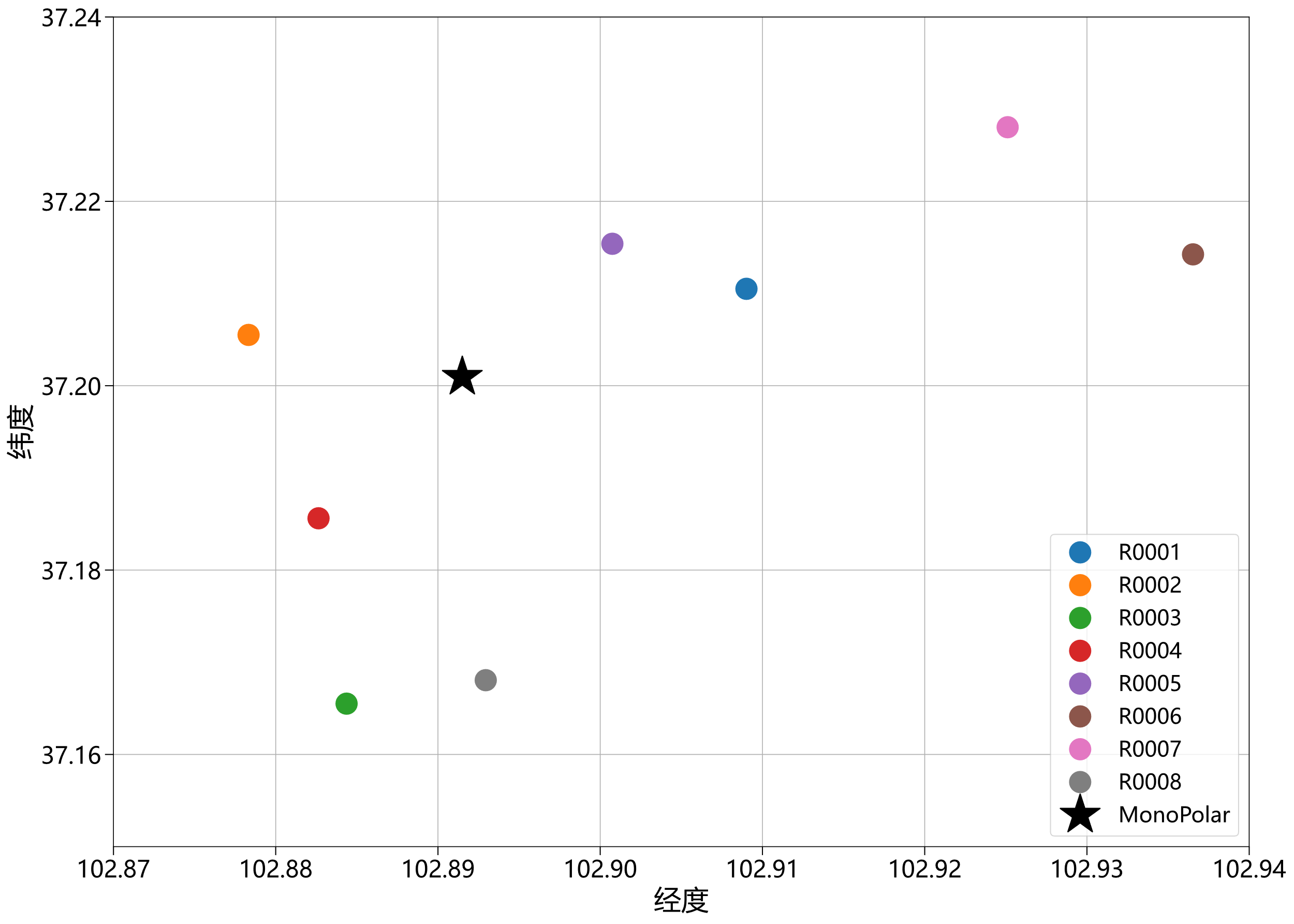


图4‑1 乌鞘岭各个雨量站位置示意图

我们将东经102.87°~ 102.94°，北纬37.15° ~ 37.24°的范围作为伍绍陵地区的试验区，按照纬度和经度将该区域划分为网格，将雨量站数据按照实际经纬度写入到对应网格内。对于没有雨量站的网格，要进行二维插值，如4.2.3.1节中介绍的，使用griddata()方法，对网格进行二维插值，也就是根据已知的雨量站数据估算出其他点。分别对实验组和对照组进行插值操作，插值完成后，使用plt.imshow()方法，将网格数据按照图片的形式画出，不同的数值对应了不同的颜色，如图4‑2所示。

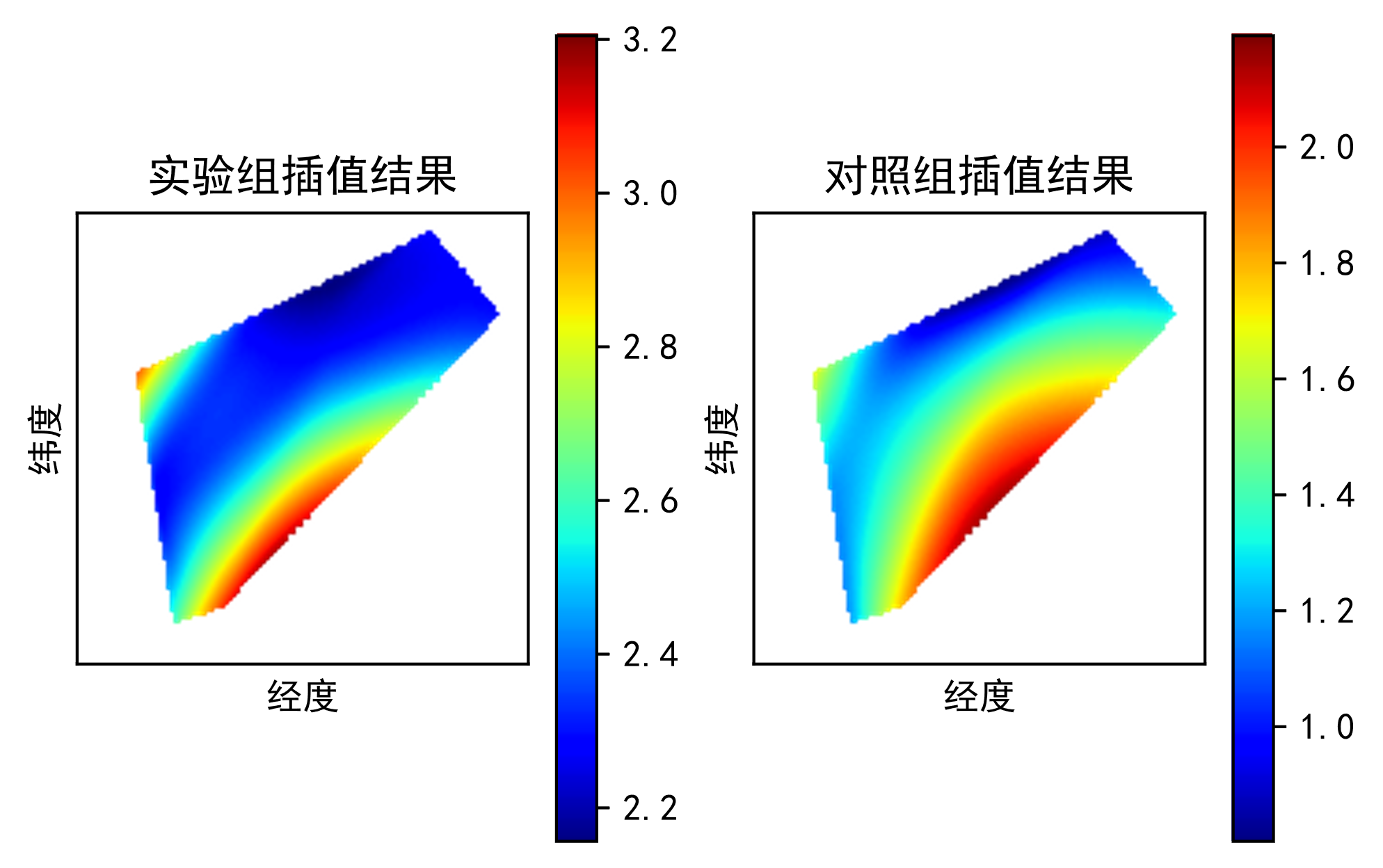


图4‑2 乌鞘岭试验区的降雨插值结果

在使用scipy包的griddata()方法进行插值时，我们对于无插值结果的网格填充的是空值None，所以插值结果中不是所有地方都有颜色。结合图4‑1可以看到，有插值数据的范围的轮廓与雨量站集合轮廓基本吻合，这一部分是实验的核心区域，也是我们重点探讨的范围。我们近似地将该地区的插值结果看作该地区的降雨分布。虽然从颜色上看，实验组的插值结果和对照组没有明显的区别，但是观察到两张图的颜色栏colorbar是不同的，实验组的数值相对较大。所以，可以看到实验组的插值结果是较大于对照组的。

随机对比试验结果分析的一个核心原则是对比实验组相较对照组的变化量，结合到本试验，就是要突出实验组降雨相比于对照组的增加量。为了可以更加直观地观察实验组相对对照组的降雨增加量，我们将实验组的插值结果减去对对照组的插值结果，也就是用两个插值得到的网格结构进行相减，得到降雨增长分布的网格结构，再把该结果画为heatmap图像，如图4‑3所示。

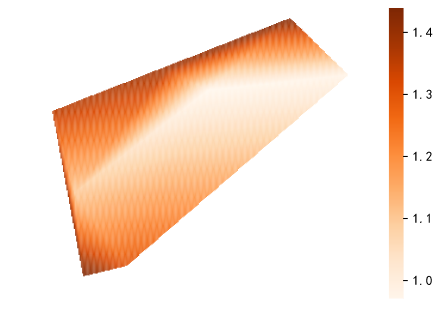


图4‑3 试验区降雨增长的heatmap图像

观察上heatmap图中的颜色栏数值，可以看到最小值是1.0，为正数。所以从图可以直观地看出，在实验期间，乌鞘岭实验组各地降雨量都有不同程度的增加。试验区地西北区域增加相对较大，中心区域增加量相对较小。

#### 六盘山可视化分析

与乌鞘岭地区不同地是，该实验区的8个雨量站中只有6个雨量站在试验期间记录下了有效数据，所以六盘山只有6个站点的数据可以被研究使用。先将各个雨量站点按照实际经纬度，在地图上标注起来，如图4‑4所示。

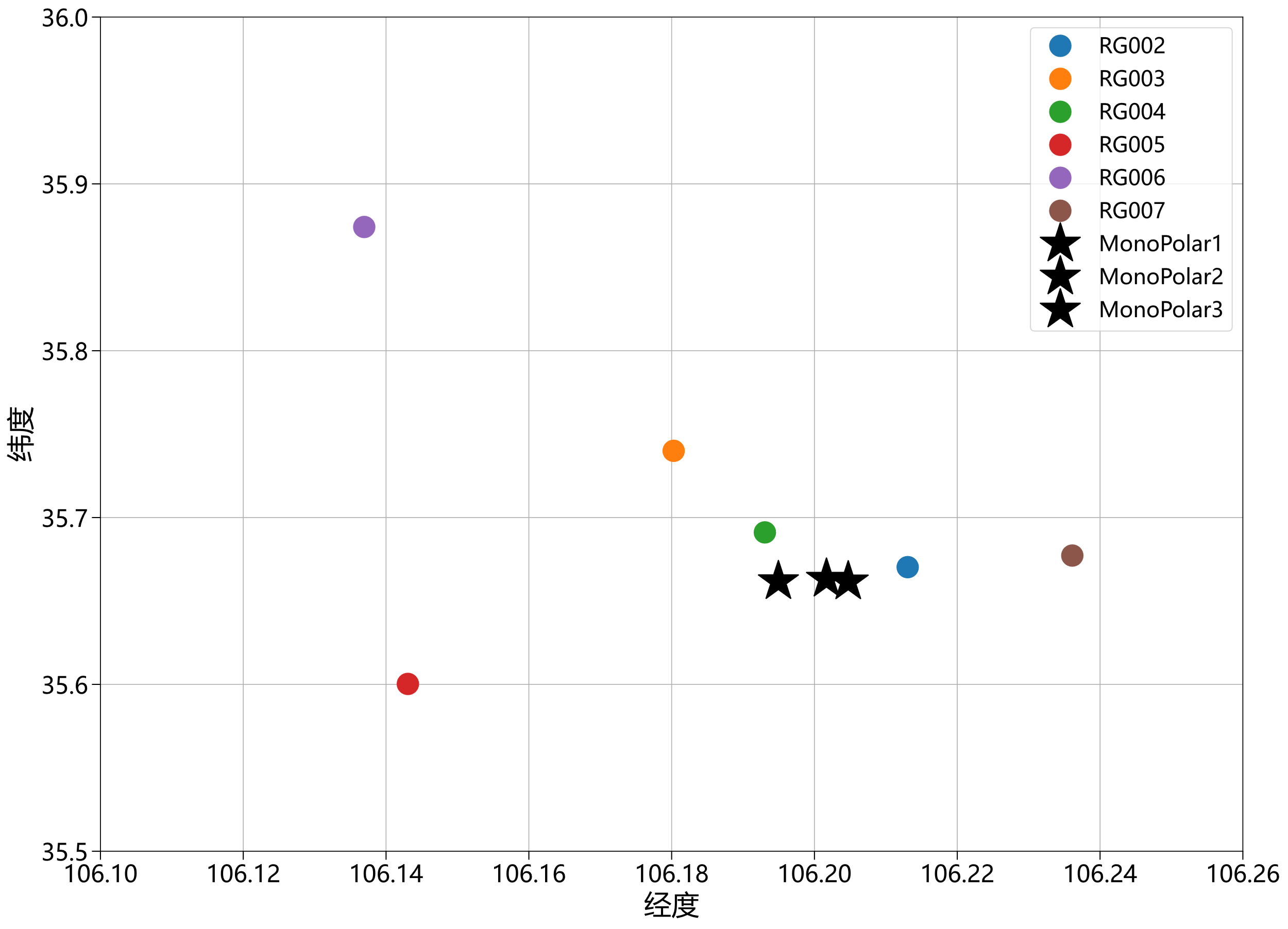


图4‑4 六盘山各个雨量站位置示意图

我们将东经106.10°~ 106.26°，北纬35.5° ~ 36.0°的范围作为六盘山地区的实验区，按照纬度和经度将试验区划分为网格，将雨量站的数据按照各自实际的纬度和经度写入到网格内。对于没有雨量站的网格，仍使用Scipy中的griddata方法，对未知点进行二维插值。分别对实验组和对照组进行插值操作，将插值后的数据绘制成heatmap图像，如图4‑5所示。

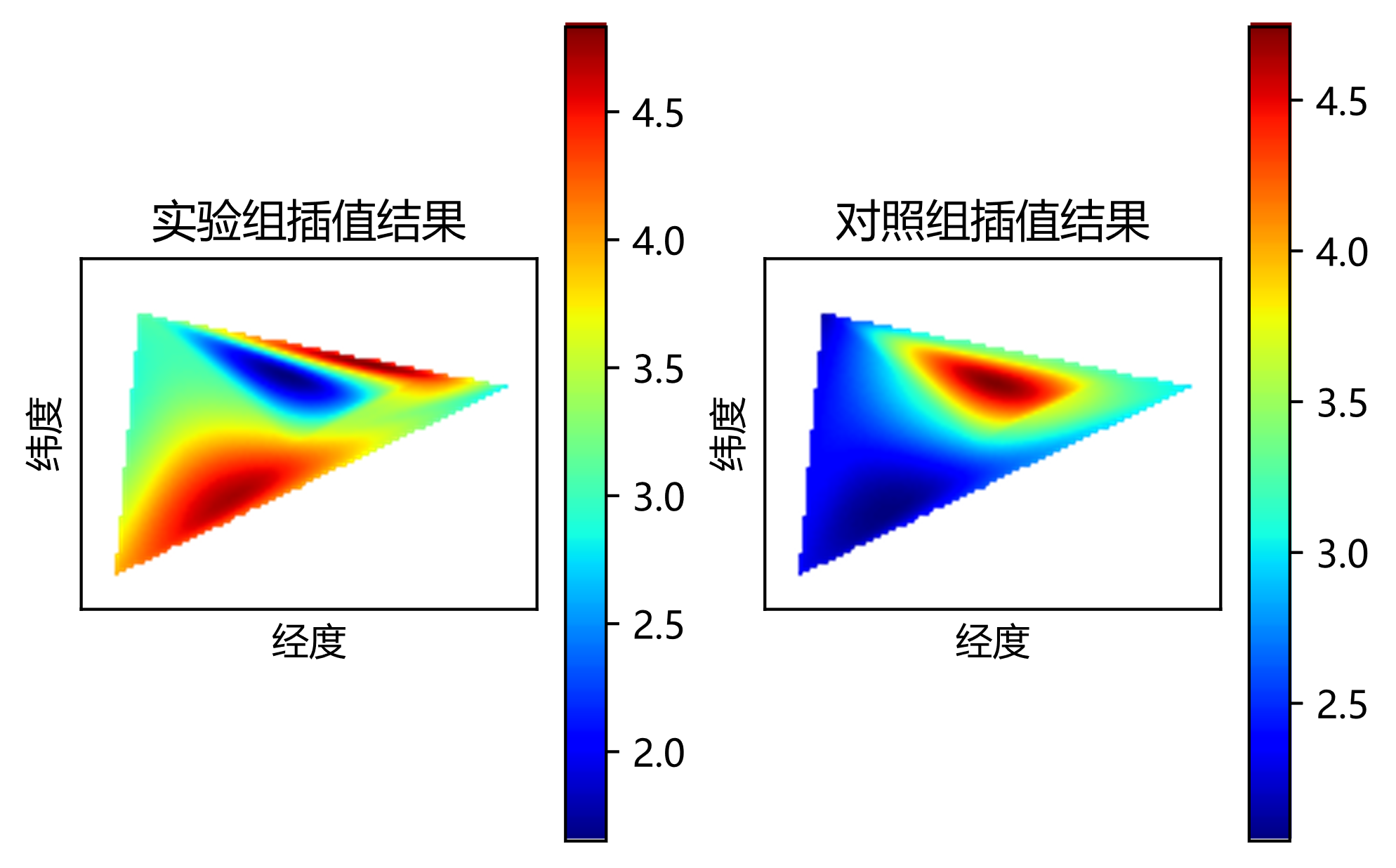


图4‑5 六盘山试验区的降雨插值结果

与乌鞘岭地区相似，图中有效数据范围的轮廓与雨量站集合轮廓基本吻合。通过观察上图，可以发现，除图示三角区域的中心靠北位置以外，实验组的图像颜色相比对照组的相同位置处颜色更深。这说明实验组的降雨大范围内是多于对照组的，但是有区域除外。

为了分析上述情况出现的原因，我们列举出了六盘山的6个有效雨量站的实验组和对照组的日均降雨量数据，如表4‑1所示。

表4‑1 六盘山有效站点雨量说明

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | R0002 | R0003 | R0004 | R0005 | R0006 | R0007 |
| 实验组 | 4.386 | 4.045 | 2.352 | 3.108 | 3.956 | 2.542 |
| 对照组 | 3.336 | 2.835 | 4.491 | 2.042 | 2.330 | 2.935 |

从表中可以看到，除了R0004和R0007站点以外，其他站点的实验组降雨量均比对照组降雨量多。参照图4‑4可知，这两个站点在试验核心区域的位置正是在中心偏右上角，与图像表现出来的信息相符。这是外场试验实际存在的现象，但是考虑到随机试验持续时间不长，降雨结果具有偶然性，两个站点的异常属于可以接受的范围。为了更加直观地观察实验组相比对照组的增加量，我们将实验组的插值结果减去对对照组的插值结果，得到降雨增长分布的网格数据，将他画为heatmap图像，如图4‑6所示。

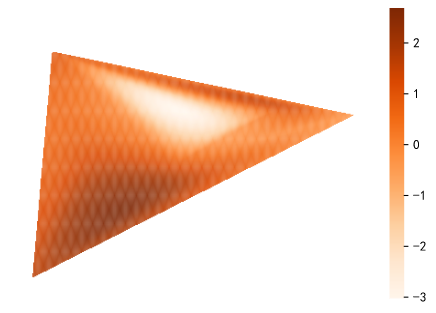


图4‑6 试验区降雨增长的heatmap图像

从图中我们可以直观的看出，除中心偏右上角区域外，核心区域内降雨其他地点的降雨量都有所增加，此结果与上述分析相符合。

## 风向影响的降雨差异分析

### 风向影响降雨的总体说明

带电粒子生成装置在工作过程中，风的影响是很大的。生成的带电粒子需要在风的作用下扩散出去，然后来催化该地区的降雨。如果没有风，带电粒子的扩散速度很慢，就会导致装置的影响范围很小，只能影响距离装置很近的地方。如果有风，理论上，下风处的降雨催化效果要比上风处的催化效果要好，因为下风处获得了更多的带电粒子。以上的推断都建立在带电粒子生成器正常工作且有效的前提下。

为了分析外场试验期间，风对试验区内不同位置降雨量的影响，我们分析了乌鞘岭地区和六盘山地区的自建雨量站点的降雨数据，以及综合站的记录下的风向数据，来探究风向与降雨催化效果之间的关系。

乌鞘岭地区和六盘山地区各自建有8个雨量站和若干个综合测试站，综合测试站记录了该地每分钟的风向风速信息。雨量站的位置不同，收集到的降雨数据不同，以降雨数据为核心，通过分析风向的变化与降雨变化之间的关系，找到上风处和下风处的雨量站降雨差异，根据各个雨量站与带电粒子发生器之间的位置关系，找出风向影响降雨是否与各自相对位置之间的联系。

### 数据处理

首先对拿到的原始数据使用LOF算法进行异常检测，具体为调用scikit-learn包的neighbors模块的LocalOutlierFactor()方法，剔除掉异常数据。

风向影响降雨实验的数据处理主要分为降雨数据和风向风速数据。雨量站和综合站点采集的数据是按照每分钟采样的，但是降雨过程往往会持续数个小时，且负离子传播催化是需要时间的，所以我们选择以天为单位进行研究。

降雨数据仍然使用pandas进行处理，转化为dataframe格式之后调用resample()方法，按照天为单位进行重采样。需要注意的是，现在我们研究的是不同站点在同一时间内的降雨差异，所以要尽可能地消除不同站点由于自身原因带来的降雨差异，凸显出风向对不同站点降雨数据的影响。最简单的方法就是根据未进行实验时的降雨数据，分别求出各个站的降雨均值，然后用实验期间的降雨数据减去这一均值，就得到了催化的效果，公式如下：

(4-2)

其中是实际降雨量，是根据未进行实验时的降雨数据计算出的平均将欲练，是催化的雨量。

综合站点收集的风向数据是按照角度为单位记录的。我们认为风向代表了催化的方向，为了可以清晰的展现出一天中的风向风速信息，我们使用了风向玫瑰图进行辅助分析。

### 风向玫瑰图

风向玫瑰图又叫风向频率玫瑰图，是用来简明扼要地描述风速和风向在特定位置的典型分布的图形工具。在python中有一个windrose库，用来管理风数据，并绘制风向玫瑰图，它使用matplotlib包作为后端，可以使用numpy数组和pandas的dataframe结构进行数据的传递。

我们将综合气象站的数据读取为dataframe格式，再从中读取风向风速数据列通过dataframe.values.tolist()方法转化为列表结构，使用WindroseAxes来绘制风向玫瑰图。

### 乌鞘岭试验区的详细分析

乌鞘岭试验区的带电粒子发生装置附近有8个自建的雨量站和4个综合站，这些站点分别记录了该地每小时的降雨数据和风向风速数据。经过数据处理、分析，发现综合站中只有W0003综合站的风向数据比较完整，只能用它作为该地区的风向数据。但是需要注意的是该站只有7、8、9三个月的数据，也就是只能对2020年7-9月的实验结果进行分析。

为了消除8个雨量站点本身的降雨差异，先要求出各雨量站的自然降雨量。因为雨量站在2019年5月已经建好并开始记录数据，所以对所有雨量站5月之后的数据使用pandas进行处理，同样地，要使用到dateframe格式方便进行重采样操作，得到2019年5月到2020年6月的各个站点的日降雨量数据。而且，因为在乌鞘岭地区进行了随机对照试验，所以在2020年7月之后仍然有几天未开装置；并且在2020年9月15日至9月22日装置出现故障，也未进行试验。将这些日期的降雨数据统计在一起，计算出每个站点的平均日降雨量数据如表4‑2所示。

表4‑2 雨量站的自然降雨数据日均值

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 雨量站 | R0001 | R0002 | R0003 | R0004 | R0005 | R0006 | R0007 | R0008 |
| 自然降雨量 | 3.213 | 3.825 | 3.378 | 2.976 | 2.832 | 3.390 | 3.039 | 3.657 |

将试验期间的各个站点的实际降雨数据，减去各自的日均自然降雨量，就得到了催化效果降雨数据。实验期间的催化增雨效果时将各个站点均无降雨的日期删除，因为实际降雨都是0时，再减去自然降雨量之后，各站的数据均为负值且都不相等，容易做出错误的判断。

求风向时，因为乌鞘岭地区只有W0003气象站的风向信息能用，所以将W0003的风向信息作为单电极的风向信息，找出主导方向后，需要以单电极装置为界限确定上风处和下风处，计算出上风处的站点的平均增雨量和下风处的站点的平均增雨量，再计算出下风处增量相比上风处增量的增长值。发现在很多日期中，下风处的雨量增加量都比上风处的雨量增加量要大，如表4‑3所示。

表4‑3 上下风雨量增量

| 日期 | 上风处站点雨量平均增量 | 下风处站点雨量平均增量 | 下风处增量相比上风处的增幅 |
| --- | --- | --- | --- |
| 2020/7/1 | 2.888151 | 5.495293 | 90% |
| 2020/7/11 | 0.406718 | 1.272235 | 213% |
| 2020/7/16 | -6.21792 | -4.40449 | 29% |
| 2020/7/17 | 7.790935 | 12.77013 | 64% |
| 2020/7/18 | 11.34078 | 23.34237 | 106% |
| 2020/7/23 | 28.07284 | 35.03279 | 25% |
| 2020/7/24 | -11.0821 | -7.57941 | 32% |
| 2020/8/3 | -1.88452 | -0.07358 | 96% |
| 2020/8/7 | -8.20118 | -4.61615 | 44% |
| 2020/8/13 | -12.6059 | -10.9839 | 13% |
| 2020/8/14 | 48.93238 | 51.39373 | 5% |
| 2020/8/17 | -0.75605 | 3.574104 | 573% |
| 2020/8/19 | -12.4133 | -11.381 | 8% |
| 2020/8/22 | 8.685797 | 13.9958 | 61% |
| 2020/8/23 | 0.334511 | 5.389129 | 1511% |
| 2020/8/29 | 99.64446 | 105.7174 | 6% |
| 2020/8/30 | -6.61201 | -3.06868 | 54% |
| 2020/9/5 | -5.51891 | -3.45382 | 37% |
| 2020/9/8 | 17.3028 | 18.71562 | 8% |
| 2020/9/13 | 21.13172 | 29.57258 | 40% |
| 2020/9/14 | 16.93538 | 30.81197 | 82% |
| 2020/9/17 | -13.1992 | -11.8777 | 10% |
| 2020/9/21 | 14.32866 | 33.30321 | 132% |
| 2020/7/9 | -0.7750 | -2.8075 | -262% |
| 2020/7/10 | -3.0225 | -3.4250 | -13% |
| 2020/7/13 | 1.1600 | 0.5920 | -49% |
| 2020/7/22 | -1.3300 | -2.8750 | -116% |
| 2020/7/25 | -3.0675 | -3.4050 | -11% |
| 2020/7/27 | -2.9675 | -3.1025 | -5% |
| 2020/8/1 | -3.0425 | -3.4050 | -12% |
| 2020/8/4 | 2.7675 | 2.6325 | -5% |
| 2020/8/5 | 9.1600 | 7.3650 | -20% |
| 2020/8/6 | -3.0175 | -3.3525 | -11% |
| 2020/8/8 | -3.1175 | -3.4025 | -9% |
| 2020/8/9 | -2.4100 | -2.4900 | -3% |
| 2020/8/11 | -3.1175 | -3.3050 | -6% |
| 2020/8/12 | 1.7500 | 1.4850 | -15% |
| 2020/8/16 | -2.8200 | -2.9300 | -4% |
| 2020/8/21 | 6.3700 | 6.3500 | 0% |
| 2020/8/31 | 0.0575 | -1.3275 | -2409% |
| 2020/9/1 | -2.5700 | -1.1200 | 56% |
| 2020/9/6 | -3.1500 | -3.1460 | 0% |
| 2020/9/9 | 1.1775 | 1.1750 | 0% |
| 2020/9/11 | -3.0675 | -3.4525 | -13% |
| 2020/9/15 | -3.1175 | -3.4275 | -10% |
| 平均值 | 4.0256 | 6.3461 | 58% |

将上表中所有日期的降雨求平均值，上风处雨量站的平均增加雨量为4.0256mm，下风处雨量站的平均增加雨量为6.3461mm，下风处的增长量比上风处的增长量大58%，也就是说下风处的催化效果比上风处强58%。这说明了催化效果与风向关系很大，也证明了我们一开始的假设是正确的。同时，不管上风处、下风处，雨量增加量均是正值，说明均受到了催化的正向作用，这一点也证明了装置的有效性。

图4‑7和图4‑8是选取下风处催化效果更好的几个日期的数据绘制的风向雨量图。图中的散点代表了8个雨量站的具体地理位置，数值为该日降雨量减去该站的平均降雨量，也就是催化的雨量值。黑色的五角星代表了乌鞘岭单电极装置，红色的五角星是综合气象站W0003，由该站的数据来计算风向。右下角是风向玫瑰图。图中的带箭头的射线代表了主导风，箭头的方向是风向，线段长度代表风速的大小。需要注意的是，风向玫瑰图上的风向指的是从外面吹向地区中心的方向，图中线条突出的程度代表了吹该风的频率。

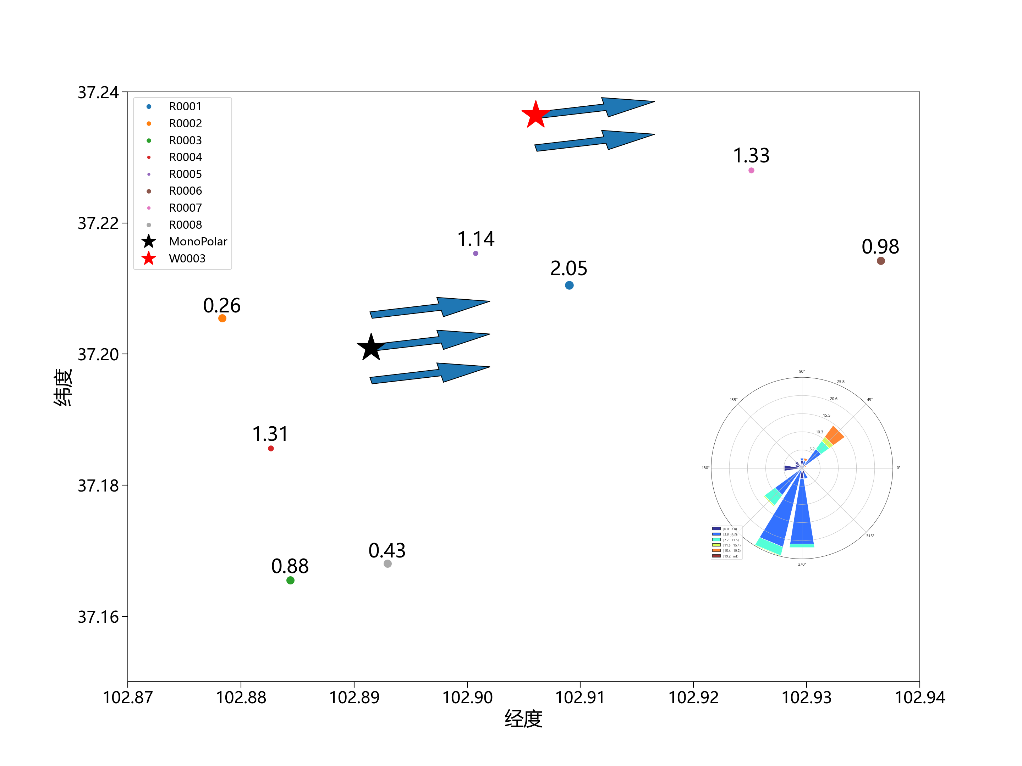


图4‑7 7月1日风向雨量图

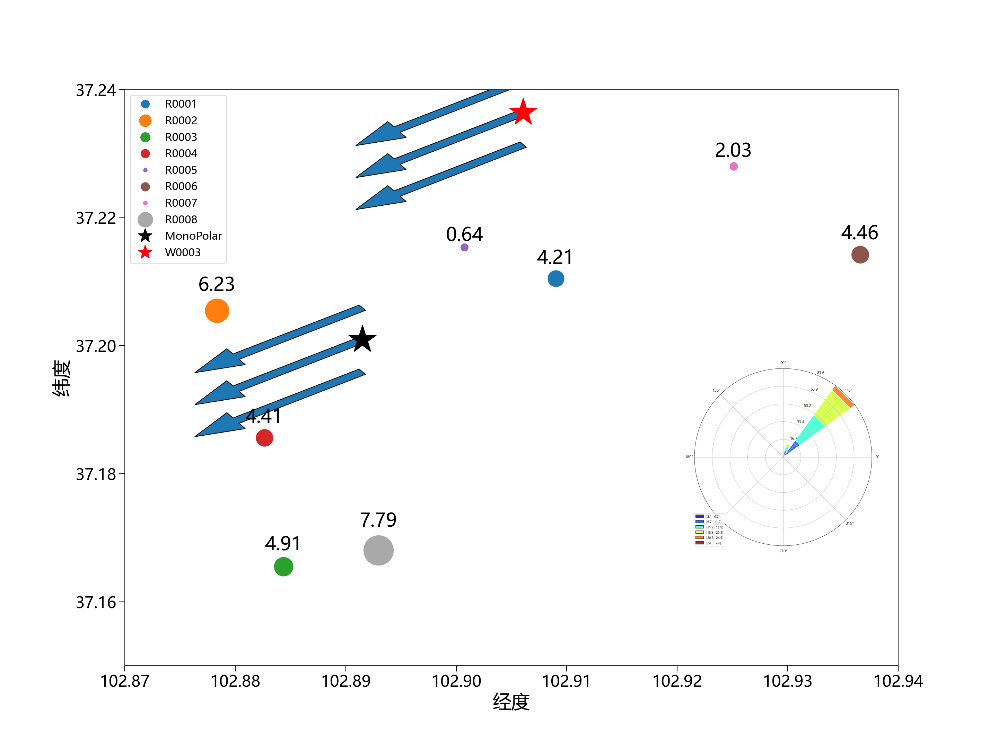


图4‑8 7月18日风向雨量图

为了使我们的结果更具有说服力，我们将研究周期从一天改为了一个降雨过程。把一个雨量站测得的日降雨量不为0，定义为一个降雨过程的开始；把8个雨量站的日降雨全为0，定义为一个降雨过程的结束。将三个月的数据进行分析，总结出了以下几个降雨过程，如表4‑4所示。

表4‑4 上下风雨量增量

|  |  |  |  |
| --- | --- | --- | --- |
| 降雨过程 | 上风处站点雨量平均增量 | 下风处站点雨量平均增量 | 下风处增量相比上风处的增幅 |
| 0709-0713 | -1.89 | -4.01 | -112% |
| 0716-0718 | 3.228449 | 7.927001 | 146% |
| 0722-0727 | -3.34571 | -2.51216 | 25% |
| 0803-0809 | 1.1975 | -0.7725 | -165% |
| 0811-0817 | 5.7275 | 5.23 | -9% |
| 0819-0823 | 5.778832 | 8.092871 | 40% |
| 0829-0901 | 19.35542 | 22.88791 | 18% |
| 0905-0909 | 1.383653 | 1.437085 | 4% |
| 0911-0917 | 0.360255 | 4.923628 | 1267% |
| 平均值 | 3.5329 | 4.8004 | 36% |

可以看到在9个降雨过程中，有6个降雨过程中，下风处的降雨增加量要大于上风处。从平均值来看，在9个降雨过程中，上风处的降雨平均增加量为

3.5329mm，下风处的降雨平均增加量为4.8004mm，下风处增量相比上风处增量增加了36%，也就是说下风处的催化效果比上风处强了36%左右。同时，下风处和上风处的降雨增加量均为正值，证明了装置的有效性。图4‑9是一个降雨过程。

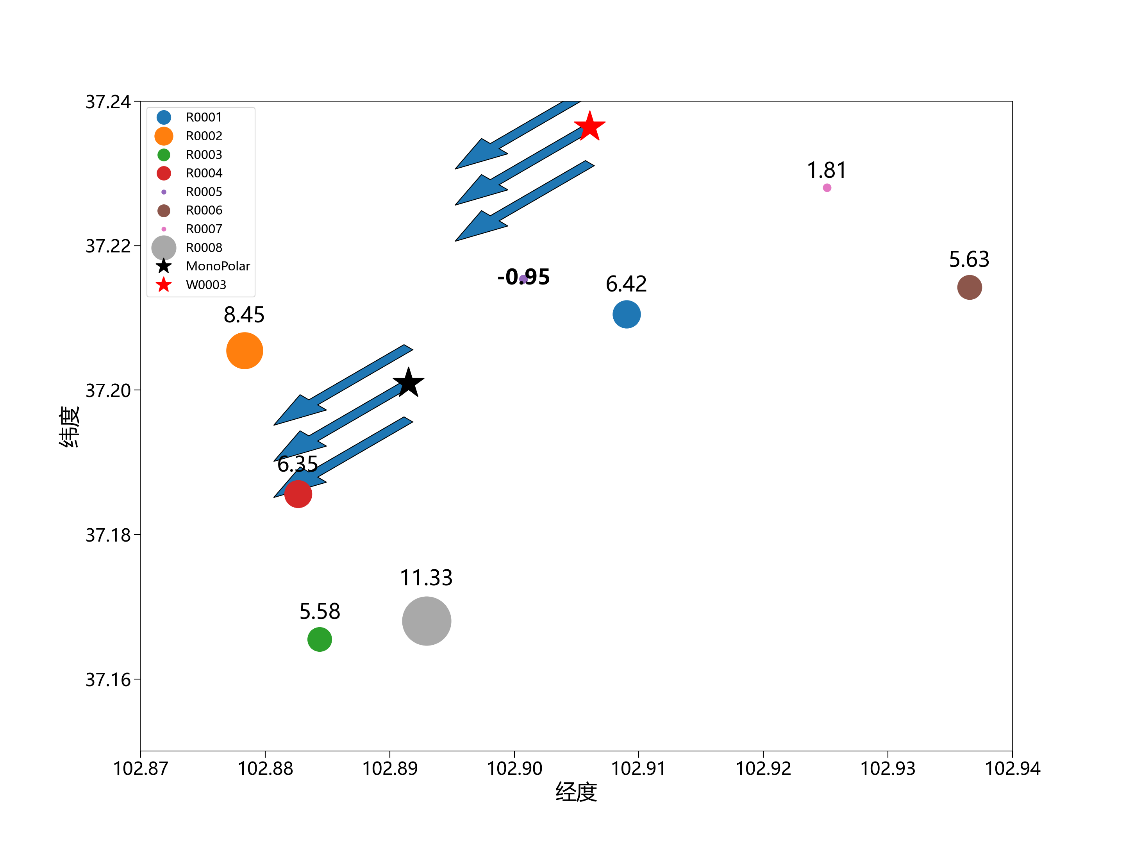


图4‑9 7月16日到18日风向雨量图

### 六盘山试验区的详细分析

六盘山试验区的单电极装置也有8个自建的雨量计站点，但是外场试验开始后，RG001雨量站实际没有降雨数据记录下来，所以只能选择RG002~RG008共7个雨量计站点的数据。六盘山的3个综合站点 WG005、WG006、WG007在外场试验期间均记录了风向风速数据。其中WG006建在单电极装置处，所以以该站的风向数据为主要风向参考数据，将WG005和WG007的数据作为辅助风向参考。

同样的，为了消除8个雨量站点本身的降雨差异，先要求出各雨量站的自然降雨量。观察到雨量站在2020年5月开始记录数据，六盘山试验区从8月开始进行外场试验。所以统计计算出2020年5月到2020年7月之间的降雨信息，将这些日期的日降雨量数据统计在一起，得到7个站点各自的自然条件下的平均降雨数据，如表4‑5所示。

表4‑5 雨量站的自然降雨数据日均值

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 雨量站 | RG002 | RG003 | RG004 | RG005 | RG006 | RG007 | RG008 |
| 自然降雨量 | 3.718 | 7.584 | 4.929 | 3.544 | 2.760 | 3.701 | 2.321 |

将试验期间的各个站点的实际降雨数据，减去各自的日均自然降雨量，就得到了催化效果降雨数据。观察实验期间的催化增雨效果时将各个站点均无降雨的日期删除，因为实际降雨都是0时，再减去自然降雨量之后，各站的数据均为负值且都不相等，容易做出错误的判断。

此处与乌鞘岭地区还有个不同点，由于乌鞘岭地区的雨量站在未做试验前有13个月的降雨数据，所以求出的平均值比较可靠。但是六盘山地区未做试验的数据只有3个月的，且在5~7月份雨水较多，所以求得的各个的平均值较高。我们的实际降雨量数据减去平均值后会有较多的负值。为了减小带来的影响，探究催化效果时，不仅删去了没有降雨的日期，而且删去了只有一个站有少量降水的日期，降低了减去大的平均值带来的对实验结果的误判。

求风向时，因为WG006建在单电极装置处，所以以该站的风向数据为主要风向参考数据，将WG005和WG007的数据作为辅助风向参考。找出主导方向后，需要以单电极装置为界限确定上风处和下风处，计算出上风处的站点的平均增雨量和下风处的站点的平均增雨量，再计算出下风处增量相比上风处增量的增长值。由于数据较多，每天的具体数据不再列出。计算的得到下风处站点雨量平均增量为0.893mm，上风处站点雨量平均增量为0.751mm，下风处相比上风处增加了19%。也就是说下风处的催化效果比上风处强19%，这说明了催化效果与风向关系比较大，也证明了我们一开始的假设是正确的。同时，不管上风处还是下风处，雨量增加量均是正值，证明了装置的有效性。

下面的几张图（图4‑10，图4‑11）是从选取下风处催化效果更好的几个日期的数据绘制的。图中的圆形散点代表了7个雨量站的具体地理位置；数值代表了该日降雨量减去该站的平均降雨量，也就是催化的降雨增加量。三个黑色的五角星代表了六盘山三个单电极装置。红色的三角形代表了综合气象站WG006，由该站点的风向作为该试验区的主要参考风向，将WG005和WG007的风向作为辅助参考风向。因为有三个独立的气象站风向数据，所以在图中没有绘制风向玫瑰图。图中的带箭头的射线代表了风，箭头的方向是风向，线段长度代表风速的大小。

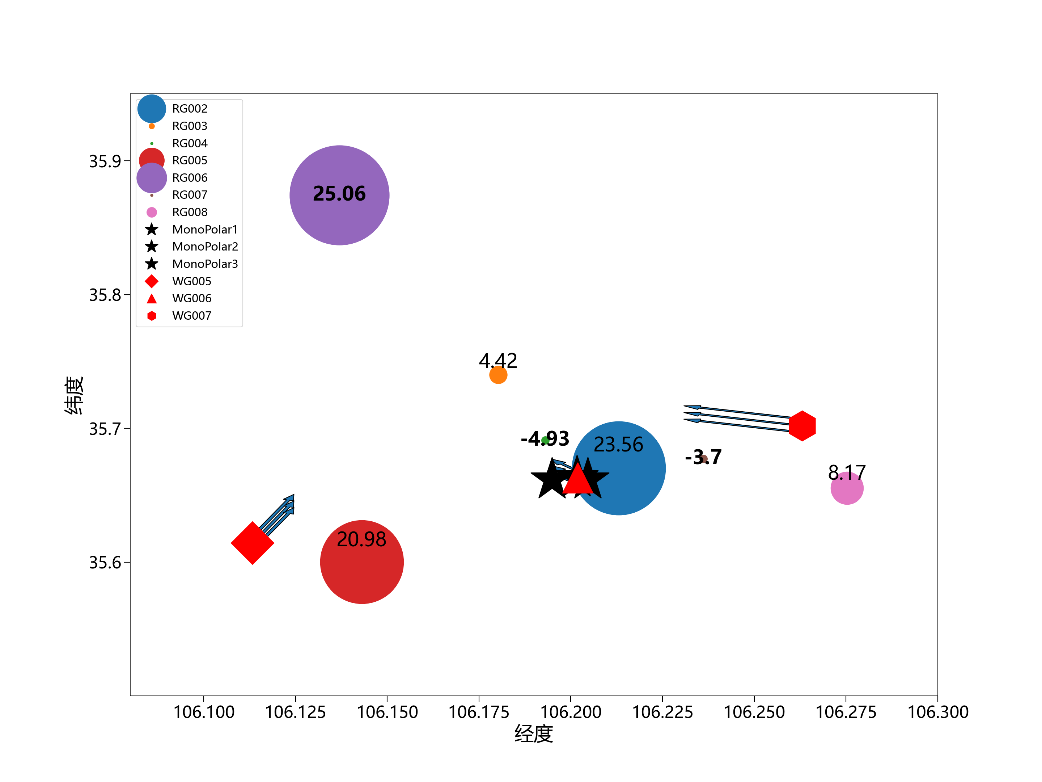


图4‑10 8月6日风向雨量图

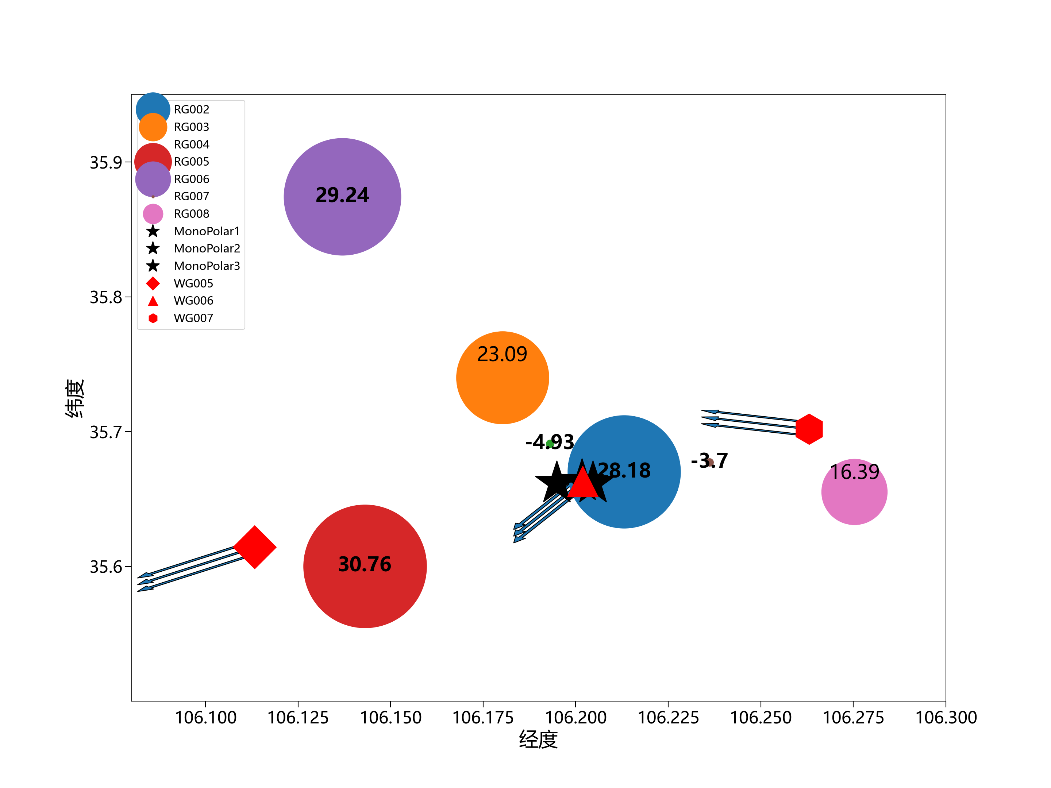


图4‑11 8月16日风向雨量图

同样的，我们对试验划分出了降雨过程，如表4‑6所示。

表4‑6 上下风雨量增量

|  |  |  |  |
| --- | --- | --- | --- |
| 降雨过程 | 下风处站点雨量平均增量 | 上风处站点雨量平均增量 | 下风处增量相比上风处的增幅 |
| 0804-0808 | 2.8175 | 0.896667 | 214% |
| 0810-0817 | 10.518 | 3.66 | 187% |
| 0819-0905 | -0.15167 | 0.08 | -290% |
| 0909-0917 | -1.156 | -0.53 | -118% |
| 0920-1002 | -0.57333 | -1.445 | 60% |
| 1012-1015 | -2.7 | -3.1875 | 15% |
| 1027-1031 | -1.816 | -1.38 | -32% |
| 1104-1106 | -4.52 | -0.52 | -769% |
| 1117-1123 | -5.92 | -2.98667 | -98% |
| 平均值 | -0.38906 | -0.60139 | 35% |

从平均值来看，上风处的降雨平均增加量为-0.60139mm，下风处的降雨平均增加量为-0.38906mm，下风处增量相比上风处增量增加了35%，也就是说下风处比上风处的催化效果强了35%左右，这一数值和乌鞘岭地区的36%非常相近。

与乌鞘岭地区不同的时，乌鞘岭地区的下风处和上风处的降雨增加量均为正值，但时六盘山地区都是负值，探究原因应该是前文提到的，由于求得的平均值相比自然水平引起的。图4‑12是一个降雨过程。

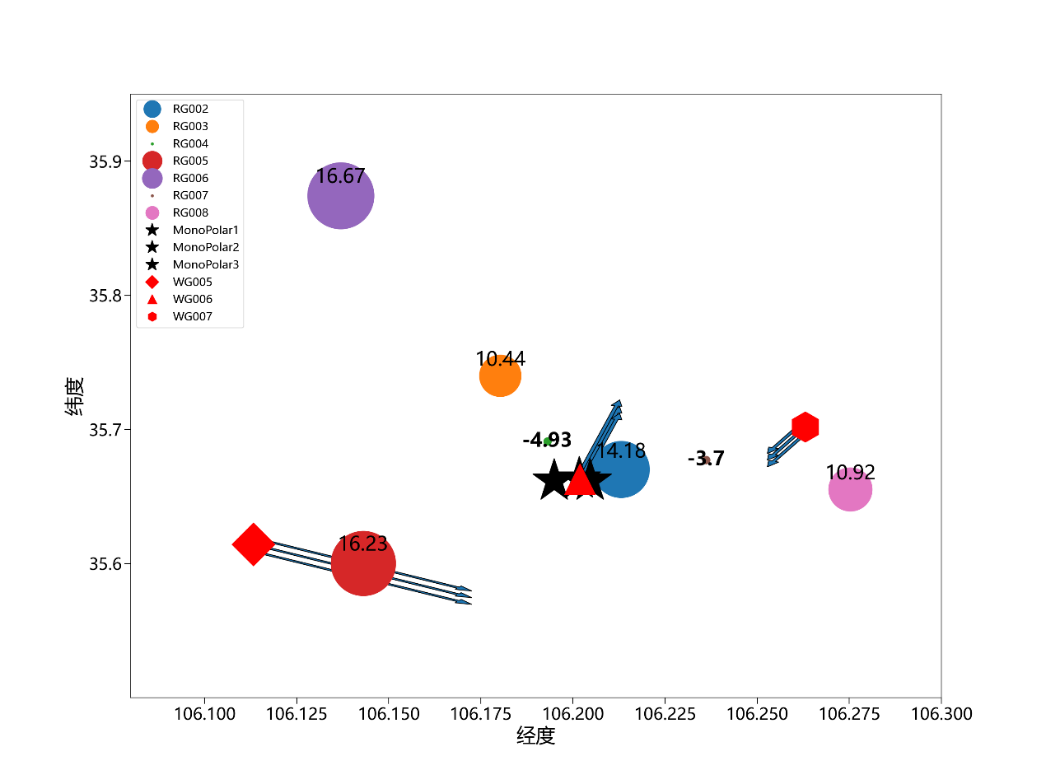


图4‑12 8月10日到17日风向雨量图

## 本章小结

本章按照统计统计分析的方法，采用了显著性检验，对“天水计划”设计的随机对比试验的实验结果进行了分析，并采用了热图和插值等手段突出了随机对比试验的实验效果。本章还探究了风向对六盘山和乌鞘岭实验区内部的雨量分布带来的影响，通过定量的分析说明了带电粒子发生装置的有效性，将风向和雨量的关系通过图像清晰地表现了出来。

# 总结与展望

论文对2020年7月到11月开展的“天水计划”外场试验进行了效果评估的方案设计，并对外场实验结果进行了评估。主要工作如下：

1. 提出了结合机器学习和传统统计分析的效果评估方法。根据数据的不同，采用不同的分析策略。
2. 对六盘山试验区搭建了神经网络模型，通过一步步的改进和评估，模型具有一定的精度，对试验期间的自然降雨量进行了预测，对于预测结果和实际结果进行了评估。
3. 对六盘山和乌鞘岭地区各自内部的随机对比试验结果进行了分析处理和显著性检验，通过可视化手段展现了试验效果。同时，针对试验区内部各雨量站的实验位置，研究了风向的影响。

本文为人工影响天气项目的效果评估做出了有益的探索，但是由于个人能力、数据获取等方面的限制，仍然存在一定的问题，包括：

1. 搭建神经网络模型时，考虑不够全面，导致走了很多弯路，前期模型效果很差，且模型的精度还有进一步提升的空间。
2. 乌鞘岭试验区周围气象站的数据由于种种原因，迟迟没有拿到完备数据，时间所限，无法对乌鞘岭地区搭建神经网络模型。
3. 在完成基于统计分析的评估时，用到的数据处理手段、以及可视化手段不够丰富。

根据上述问题，结合本文已有的研究成果，未来工作的侧重点如下：

1. 进一步改善神经网络模型，提高模型的预测精度。
2. 尽快拿到乌鞘岭的气象数据，尝试使用新的模型。
3. 对更多的气象要素进行分析，采用更多的分析方法和可视化手段。

致 谢

转眼间，四年过去了。四年前踏入校园，清晰地就像发生在昨天，而不久后，我们都将挥手作别，开启下一段旅程。在四年的学习生活中，我成长了好多，一路上得到了很多人的陪伴、支持和帮助。在论文完成之际，我衷心地向我的老师、同学以及亲人们致以感谢。

首先要感谢我的导师秦磊华教授。秦老师十分认真，每次我向他寻求帮助总能得到他耐心的讲解和针对性很强的建议。在我的论文写作期间，秦老师对我提出了很多宝贵的意见，帮助我顺利地完成了毕业设计。

还要感谢我的师兄薛凤鸣对我毕业设计的帮助。不论效果评估还是机器学习，在本次设计之前我都是门外汉，多亏师兄的耐心指导和无私帮助，我才能一步步地克服困难完成论文。也要感谢实验室的其他师兄师姐，在这段实验室的学习时光中，他们都给予了我很多关心和帮助。

感谢我的室友，四年来大家朝夕相伴，互相进步，互帮互助，一起度过了美好的大学时光。感谢我的朋友，大家的开心和烦恼总是会一起交流，和他们在一起的时间总是很快乐。感谢我的老师们，四年来我们遇到了很多老师，可能他们不能记住每一个人的名字，但是他们严谨的治学态度、丰富的知识，以及对我们诲人不倦的态度总是让我们如沐春风。

最应该感谢的是我的父母。不论何时，他们都在我身后默默的支持我，给予我物质的帮助和精神上的慰藉，没有他们的付出，我也不可能顺利的完成学业。

最后，向四年中遇到的所有美好的人致以感谢。

参考文献

[1] 陈灿平. 我国水资源问题的成因分析和对策研究[J]. 西南民族大学学报(人文社会科学版),2014, 35(06): 141-144.

[2] Khain A, Arkhipov V, Pinsky M,et al. Rain enhancement and fog elimination by seeding with charged droplets. Part I: Theory and numerical simulations[J]. Journal of applied meteorology, 2004, 43(10):1513-1529.

[3] Rohwetter P, Kasparian J, Stelmaszczyk K,et al. Laser-induced water condensation in air[J]. Nature Photonics, 2010, 4(7):451-456.

[4] Ju J, Liu J, Wang C,et al. Laser-filamentation-induced condensation and snow formation in a cloud chamber[J]. OPTICS LETTERS, 2012, 37(7):1214-1216.

[5] 潘垣, 于克训. 带电粒子催化人工降雨雪新原理新技术及应用示范(天水计划)[J]. 中国环境管理,2017, 9(03): 115-116.

[6] 房彬, 肖辉, 王振会, 等. 聚类分析在人工增雨效果检验中的应用[J]. 南京气象学院学报,2005,(06): 739-745.

[7] Liang M, Hu X. Recurrent convolutional neural network for object recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2015:3367-3375.

[8] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling[C]. Thirteenth annual conference of the international speech communication association, 2012.

[9] Lu H, Setiono R, Liu H. Effective data mining using neural networks[J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 1996, 8(6):957-961.

[10] Yu L, Wang S, Lai K K. An integrated data preparation scheme for neural network data analysis[J]. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2005, 18(2):217-230.

[11] Alpaydin E. Introduction to machine learning[M]. MIT press, 2014.

[12] Gabriel R K, Baras M. The Israeli rainmaking experiment 1961-67: Final statistical tables and evaluation[M]. Hebrew University, 1970.

[13] Breed D, Rasmussen R, Weeks C,et al. Evaluating winter orographic cloud seeding: design of the Wyoming Weather Modification Pilot Project (WWMPP)[J]. Journal of Applied Meteorology and Climatology, 2014, 53(2):282-299.

[14] 叶家东, 范蓓芬. 人工影响天气的统计数学方法[M]. 科学出版社, 1982.

[15] 李书严, 李伟, 赵习方. 北京市人工增雨效果评估方法分析[J]. 气象科技,2006,(03): 296-300.

[16] 黄彦彬, 姚展予, 游积平, 等. 基于“.NET”的华南区域人工增雨效果统计检验共享平台的设计与实现[J]. 气象,2011, 37(06): 727-734.

[17] Silverman B A. A Critical Assessment of Glaciogenic Seeding of Convective Clouds for Rainfall Enhancement[J]. BULLETIN OF THE AMERICAN METEOROLOGICAL SOCIETY, 2001, 82(5):903-924.

[18] Mielke P W, Berry K J, Medina J G. Climax I and II: Distortion Resistant Residual Analyses[J]. Journal of Applied Meteorology, 1982, 21(6):788-792.

[19] Levi Y, Rosenfeld D. Ice Nuclei, Rainwater Chemical Composition, and Static Cloud Seeding Effects in Israel[J]. Journal of Applied Meteorology, 1996, 35(9):1494-1501.

[20] Council N R. Critical issues in weather modification research[M]. National Academies Press, 2004.

[21] WMO Statement on Weather Modification. The Commission for Atmospheric Sciences Management Group, Second Session, Oslo, Norway, 24-26 September 2007[EB/OL]. https://www.wmo.int. 2007

[22] 叶家东. 人工降水的试验设计和效果检验[J]. 气象,1979,(02): 26-29.

[23] 王伟健, 姚展予, 贾烁, 等. 随机森林算法在人工增雨效果统计检验中的应用研究[J]. 气象与环境科学,2018, 41(02): 111-117.

[24] 刘晴. 人工增雨效果统计检验方案优选及个例分析[D]. 中国气象科学研究院, 2013.

[25] Sander J. LOF: Identifying Density-Based Local Outliers.[J]. 2000, 29(2):93-104.