Summary

## 1.介绍（Introduction）

### 1.1问题背景（Background）

近两年来，我国河南、陕西、湖北等地遭遇了极为罕见的强降雨。与此同时，一些北方城市遭遇了历史罕见的暴风雪。这些暴雨和降雪对当地人民的生命、安全和财产构成严重威胁。以郑州为例。在今年，7月18日18时至21日0时，郑州市出现暴雨和极端暴雨。累计平均降水量449毫米，20日16时至17时郑州站降雨量达到201.9毫米，超过了我国陆地上每小时降雨量的极值。7月17日开始断断续续下大雨，到了7月20日上午，雨突然开始加大。截至当日下午，郑州多处社区和道路被雨水淹没。郑州市气象局发布信息，郑州年平均降雨量为640.8毫米，这三天降雨量接近甚至超过往年。截至7月23日12时，经过初步统计，郑州市共紧急转移安置395989人，农作物受灾面积44209.73公顷，直接经济损失为6550亿元，暴雨引发的洪涝及次生灾害已造成数百人死亡。

据相关研究人员介绍，在全球变暖的背景下，我国未来降水的数量、强度、频率和类型都将受到直接影响。预计本世纪末降水量将增加10%左右，极端降水概率将显著增加。由于我国国土面积大，加上各种类型的地形地貌等因素的共同影响，不同城市的降水特征呈现出不同的特点。因此，有必要建立不同潜在极端降水事件城市的预测模型及其损失的定量分析模型。

### 1.2团队任务（Team Mission）

在本文中，我们将我们的工作分为以下几个部分：

（1）数据预处理。该部分包括了数据集的获取，对数据集数据的初步处理，例如删除确实值较多的记录等等。

（2）问题分析、构建模型。该部分包括对各个问题的分析，并尝试构建合理的数学模型去解决实际问题。

（3）模型评估。该部分包括了对模型的检验，以及整体优点和不足的分析。

我们团队要解决的问题包括：

（1）对郑州地区降水特征的年变化特征进行相关分析，筛选出降水量较高的年份。同时，对2021年郑州市淹没事件进行具体定量分析。

（2）收集整理我国更多城市多年的降水数据，分析这些城市的降水趋势。

（3）建立普适的模型，利用收集到的城市天气数据进行预测和分析。

## 2.问题分析（Problem Analysis）

## 2.1 数据分析与预处理（Data Analysis And Pre-processing）

本文数据集来自附件1中提供的郑州市3个气象站近70年的日降水观测量资料与NOAA所提供的开放气象资料。根据NOAA网站所提供的的参考文档，数据集中各个指标如下表：

|  |  |
| --- | --- |
| 气象指标对照表 | |
| STATION | 观测站编号 |
| DATE | 日期 |
| DEWP | 露点Mean dew point (.1 Fahrenheit) |
| FRSHTT | 标志当天是否发生了（1表示发生了，0表示没有发生）  总共有六位数分别为:雾/雨/雪/冰雹/雷/台风  Indicator for occurrence of:  Fog/Rain or Drizzle/  Snow or Ice Pellets/Hail  Thunder  Tornado/Funnel Cloud |
| GUST | 阵风  Maximum wind gust (.1 knots) |
| MAX | 最高温度  Maximum temperature (.1 Fahrenheit) |
| MIN | 最低温度  Minimum temperature (.1 Fahrenheit) |
| MXSPEED | 最大风速  Maximum sustained wind speed (.1 knots) |
| PRCP | 降水量  Precipitation amount (.01 inches) |
| SLP | 海平面气压  Mean sea level pressure (.1 mb) |
| SNDP | 雪深  Snow depth (.1 inches) |
| STP | 本站气压  Mean station pressure (.1 mb) |
| TEMP | 温度  Mean temperature (.1 Fahrenheit) |
| VISIB | 能见度  Mean visibility (.1 miles) |
| WDSP | 风向风速  Mean wind speed (.1 knots) |

上述指标提供了温度、风速、气压等方面的指标，这些因素都将从一定程度上影响降水。这些在后续的相关分析中都将作为主要指标。

在气象站实际的观测过程中，有时只会观测并记录部分信息，没有记录的信息往往会被标记特定的值，表明它们没有被观测。如果某个指标的值全部都是9，例如在降水量值为99.99英寸，或者是阵风为999.9节，就意味着气象站此时没有观测降水量与阵风大小。当然，这些值将不能作为有效数据，应当舍去。

从时间跨度上来看，观测站1记录了从1957年12月3日起至2021年11月5日的气象数据，观测站2记录了1983年7月1日起至2021年11月6日的气象数据，观测站3记录了1961年10月22日起至2021年11月6日的气象数据。这些观测站都没有记录1965年至1972年的数据，因此在之后的绘图与建模过程中，将略去这部分年度的信息。

同时，为了最大化的利用三个观测站的观测信息，在时间跨度上有重合的年份，对观测站收集的一些定量数据（例如降水量、温度等指标）取平均值。

## 2.2 问题分析（Question Analysis）

问题一中要对郑州地区降水的年变化特征进行相关分析，并筛选出降水量较高的年份。由于提供的数据集含有多个指标，包括了温度、风速、气压等等，这里可以利用多元分析的方法进行建模。

常见的多元分析模型包括：

（1）多元方差分析、多元回归分析和协方差分析，称为线性模型方法，用以研究确定的自变量与因变量之间的关系；

（2）判别函数分析和聚类分析,用以研究对事物的分类；

（3）主成分分析、典型相关和因素分析，研究如何用较少的综合因素代替为数较多的原始变量。

本文使用的方法是主成分分析与多元回归分析方法。以年降水量为因变量，以其他各个指标为自变量，建立多元分析模型。由于数据集中存在较多的指标，所以可以使用主成分分析来进行降维，经过特征选择后，余下的指标就是影响城市降水量的重要特征指标。

问题二中要求收集我国更多城市多年的降水数据，并分析这些城市的降水趋势。本文数据来源为美国国家海洋和大气管理局（NOAA）的开放气象数据。通过查阅城市的气象台编号，我们从该网站收集了北京、济南、广州、银川、大连五个城市的近70年的气象信息。通过作出这些城市近70年的降水量变化图，构造差分整合移动平均自回归模型，可以直观看出这些城市的降水趋势。

问题三中要求用收集到的城市天气数据进行预测和分析。考虑到天气数据为时间序列，所以可以使用构造人工神经网络的方法来进行城市天气数据预测。本文构造的模型是长短期记忆循环神经网络（LSTM）模型，来对城市的降水量和温度进行预测。

## 3. 模型构建（Models And Answers）

### 3.1 Model 1: 利用主成分分析进行相关分析

#### 3.1.1 模型假设

（1）这里假设各个观测站所观测到的数据都是准确可靠的

（2）假设忽略GUST，SNDP，STP数据列对主成分分析没有影响。因为提供的数据集中，GUST，SNDP，STP中大部分数据都为缺失值（Missing），被标记为999.99，因此在进行主成分分析时，忽略了上述指标。

#### 3.1.2 模型分析与符号表示

主成分分析(PCA)是一种线性变换技术，被广泛应用于各种不同的领域，特别是特征提取与降维。PCA帮助我们根据特征之间的相关性来识别数据中的模式，这与我们的所研究的问题是相符的。简单的说，PCA的目的是寻找高维数据中存在最大方差的方向，并将数据投影到维数小于或等于原始数据的新子空间。

如果用PCA降维，我们可以构造d\*k维的变换矩阵W，它能把训练样本的特征向量x映射到新的k维特征子空间，该空间的维数比原来的d维特征空间要少。例如，我们有一个特征向量x：

x = [x1,x2,…,xd] ,x ∈Rd

接着通过一个变换矩阵W∈Rd\*k进行变换：

xW = z

结果以向量方式表达如下：

z = [z1, z2 ,z3, z4,…,zk] z∈Rk

#### 3.1.3 模型构建

（1）数据标准化

（2）提取主成分

#### 3.1.4 降水特征分析

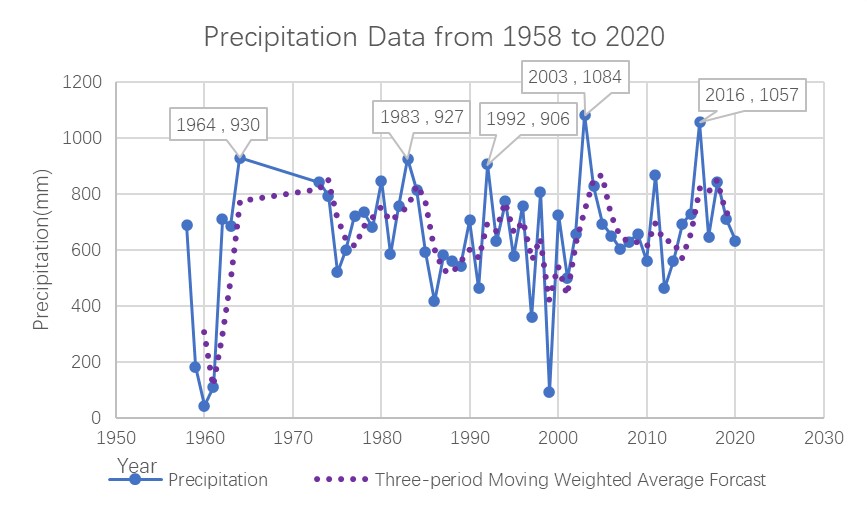


图. 1958~2020年郑州市降水量数据图

自1958年至2020年，郑州市年平均降水量为654mm。通过线性回归的方法，可以求得降水的变化趋势线。同时，回归系数可以反映气候变量的倾向率，即平均每一年气候变量的变化量。

结合数据集，容易求得趋势线方程为：

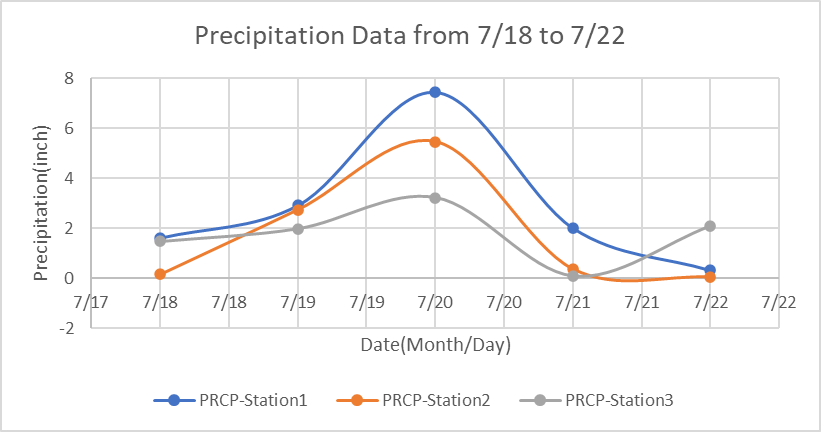
Y = 2.6857x – 4696.4

由此可见，郑州市近60年的年平均降水量有上升趋势。通过上述折线图，可以发现降水量较高的年份为1964年，1983年，1992年，2003年，2016年，其年降水量分别为930mm，927mm，906mm，1084mm，1057mm，均超过了900mm。

除此之外，由上述折线图可以直观观察出多雨期主要为1964年前后，1982~1984年，与2002年前后；少雨期主要为1965年~1981年，1985~2001年。同时，从折线图中，可以直接观察到郑州市降水具有一定的周期性，且不断震荡，同时存在突变的情形。通过三周期移动加权平均的方法，可以作出郑州市近70年降水量的趋势线。

在2021年郑州市“7·20”特大暴雨中，郑州市在短短几天内的降水量就已经超过了其年平均值。根据所提供的三个观测站的数据，本次强降水从7月18日开始，一直至7月22日，持续约4天。我们团队查阅了中国气象台网的有关数据，了解到本次强降雨的最强时段在19至20日，这与提供的数据是一致的。在2021年7月20日，三个观测站记录的降水量分别为7.43英寸（即约189mm），5.46英寸（即约139mm），3.22英寸（即约82mm）。同时，郑州在7月20日16时至17时，其最大小时降水量达到了201.9mm，突破了中国大陆小时降雨量历史极值。郑州局部3小时最大降雨量达到了333毫米。由此可见，本次强降水过程具有雨量大、时间长、短期强降水强、降水极端性突出的特点。

截止至8月2日下午，根据河南省人民政府新闻办公室公布的消息，郑州市本次淹没事件已经造成了本市292人遇难，47人失踪。全省倒塌房屋30106户，89001间，农作物成灾面积872万亩，绝收面积380万亩，直接经济损失1142.69亿元。



### 3.2 Model 2: 利用差分整合移动平均自回归模型分析城市降水趋势

#### 3.2.1 数据来源

本模型的数据来源为美国国家海洋和大气管理局的开放气象数据。我们团队通过访问该网站Global Surface Summary of the Day – GSOD数据库，查询并下载了北京、广州、银川、大连和济南5个城市1956年至2021年近70年的气象资料。这些气象资料包括的各个气象指标与问题一中郑州市三个观测站的各个指标是一致的，如下表：

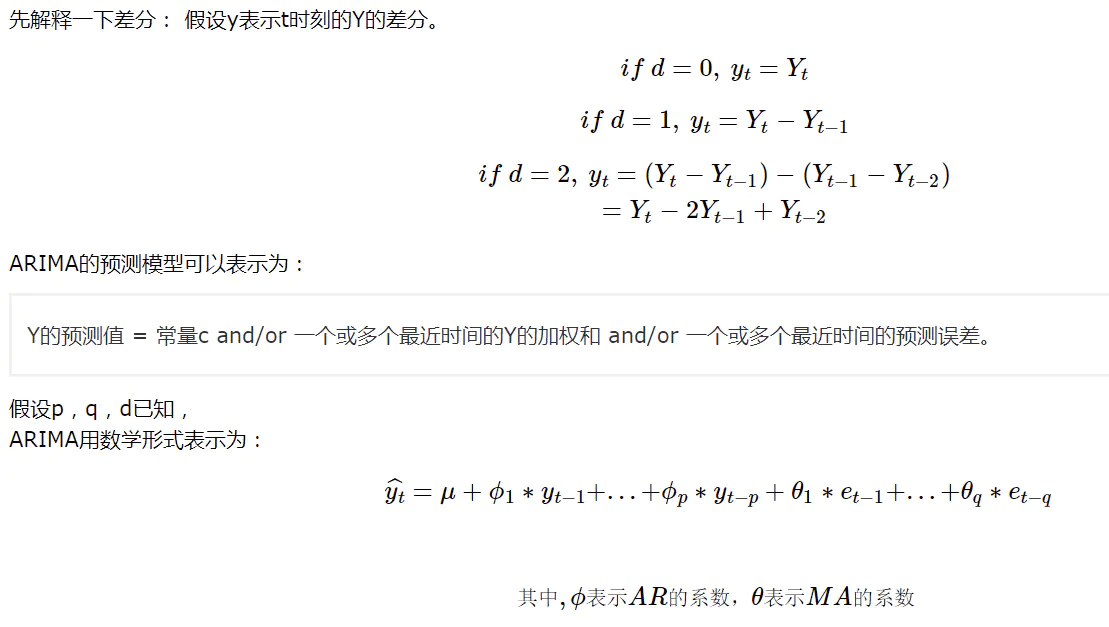
|  |  |
| --- | --- |
| 气象指标对照表 | |
| STATION | 观测站编号 |
| DATE | 日期 |
| DEWP | 露点Mean dew point (.1 Fahrenheit) |
| FRSHTT | 标志当天是否发生了（1表示发生了，0表示没有发生）  总共有六位数分别为:雾/雨/雪/冰雹/雷/台风  Indicator for occurrence of:  Fog/Rain or Drizzle/  Snow or Ice Pellets/Hail  Thunder  Tornado/Funnel Cloud |
| GUST | 阵风  Maximum wind gust (.1 knots) |
| MAX | 最高温度  Maximum temperature (.1 Fahrenheit) |
| MIN | 最低温度  Minimum temperature (.1 Fahrenheit) |
| MXSPEED | 最大风速  Maximum sustained wind speed (.1 knots) |
| PRCP | 降水量  Precipitation amount (.01 inches) |
| SLP | 海平面气压  Mean sea level pressure (.1 mb) |
| SNDP | 雪深  Snow depth (.1 inches) |
| STP | 本站气压  Mean station pressure (.1 mb) |
| TEMP | 温度  Mean temperature (.1 Fahrenheit) |
| VISIB | 能见度  Mean visibility (.1 miles) |
| WDSP | 风向风速  Mean wind speed (.1 knots) |

由于要分析降水趋势，因此自变量是时间，因变量是降水量。通过作出5个城市降水量随时间的变化关系图像，可以大体分析出这5个城市的降水趋势。同时，差分整合移动平均自回归模型在长时间序列预测具有优良的性能，其也能够完成对降水趋势的分析。

#### 3.2.2 模型分析与符号表示

构造ARIMA模型最重要的地方在于时序数据的平稳性。平稳性是要求经由样本时间序列得到的拟合曲线在未来的短时间内能够顺着现有的形状惯性地延续下去，即数据的均值、方差理论上不应该有过大的变化。因此，其第一步是观测数据是否为平稳时间序列。对于非平稳时间序列，应先进行d阶差分运算，化为平稳时间序列：

在得到平稳时间序列之后，要对平稳时间序列分别计算其自相关系数ACF和偏自相关系数PACF，通过对自相关图和偏自相关图的分析，得到最佳的阶层p和阶数q,最终得到下述的ARIMA模型：

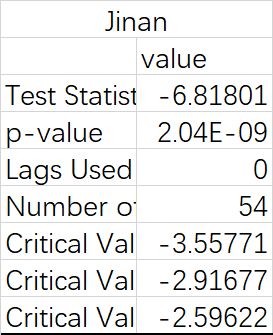
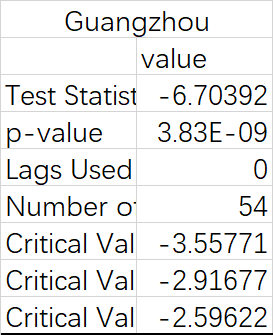
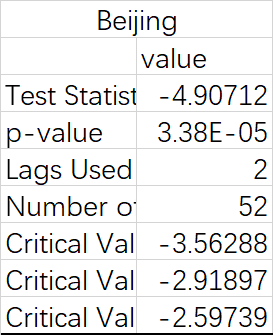


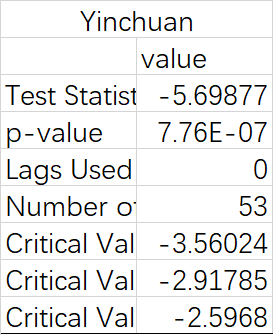
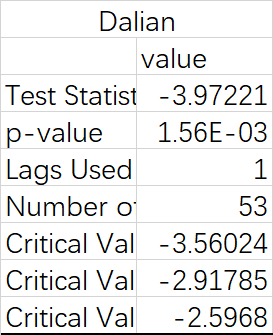
本文中，用来描述各个城市的ARIMA模型。

#### 3.2.3 模型构建

（1）对给出的时间序列进行平稳性检验：

通过折线图，可以大体看出这些数据都属于平稳时间序列。接下来通过ADF检验来证明这些时间序列都是平稳的。





由于ADF检验数值均小于-3.56024,同时注意到所有的ADF检验中p值很小，因此能够拒绝原先的非平稳性假设，能够认为这些时间序列都是平稳的。

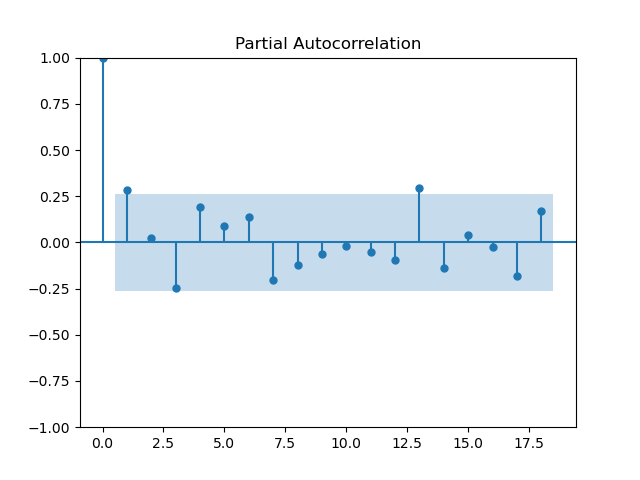
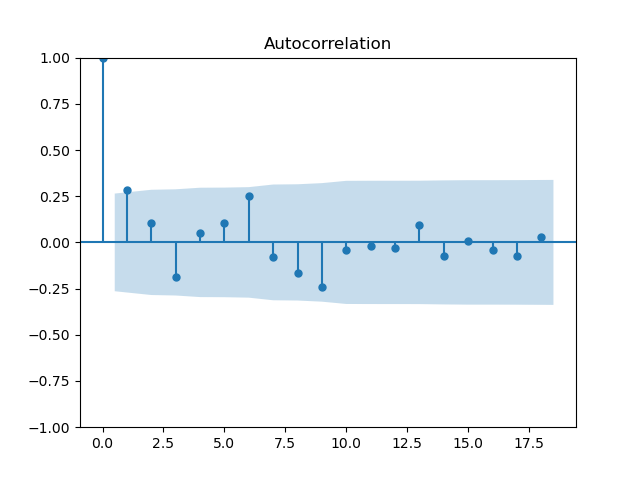
因此，能够确定在5个趋势分析模型中，d=0

（2）确定参数p, q

为了确定参数p和q，需要作出ACF图（自相关图）与PACF图（偏自相关图）。其中，p为自回归模型阶数，q为移动平均模型阶数。由ACF图可以确定参数p，由PACF图可以确定参数q。

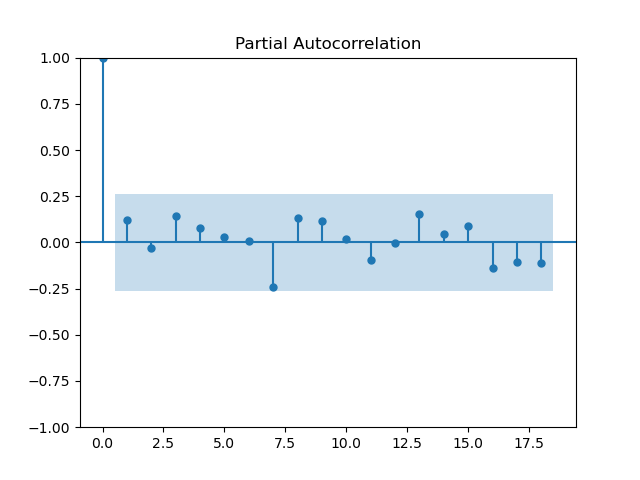
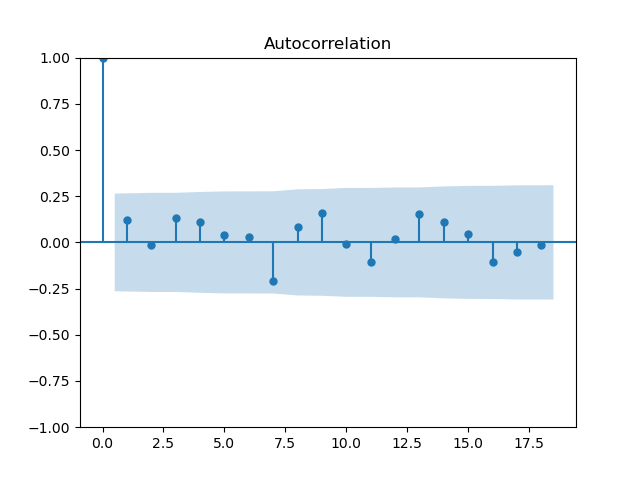


下图为北京市降水量系列数据的ACF图与PACF图：



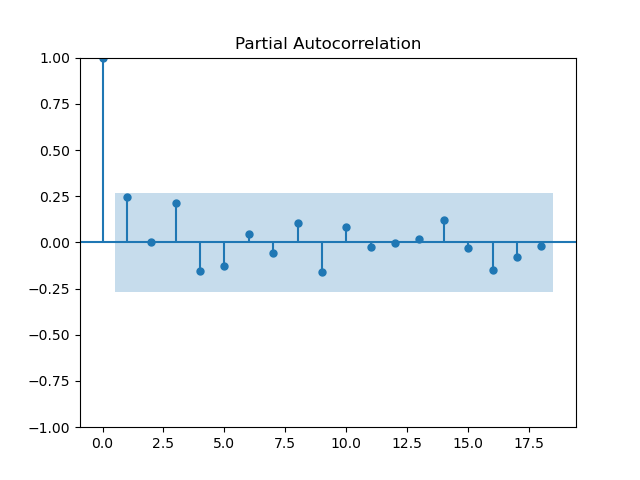
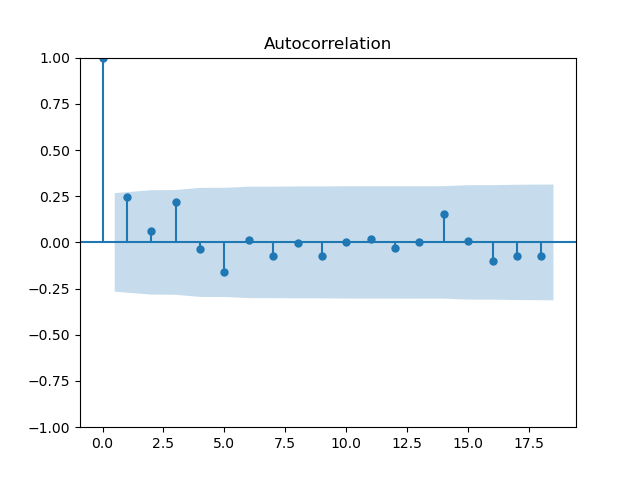
所以能够确定，对于北京市降水量系列数据而言，ARIMA模型中参数p应取0，q可以取1

类似的，下图为广州市降水量系列数据的ACF图与PACF图：



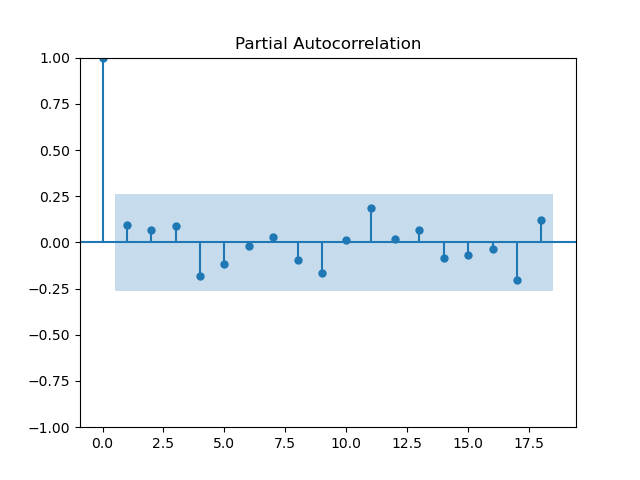
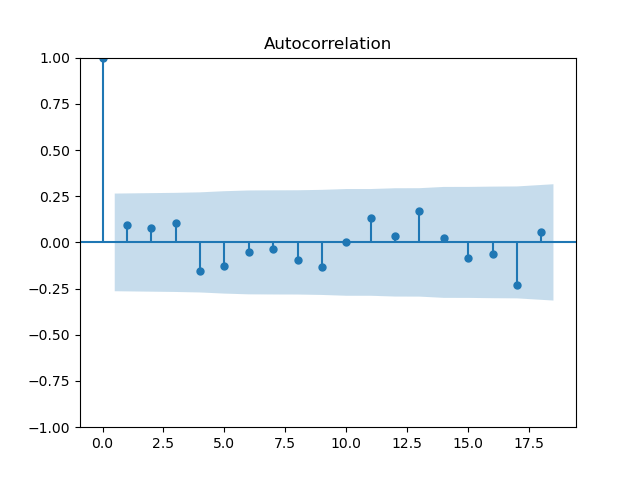
能够确定，对于广州市降水量系列数据而言，ARIMA模型中参数p应取1，q可以取1

下图为银川市降水量系列数据的ACF图与PACF图：



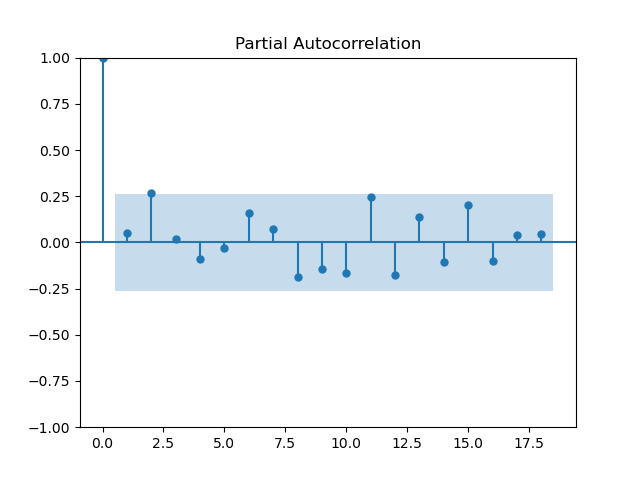
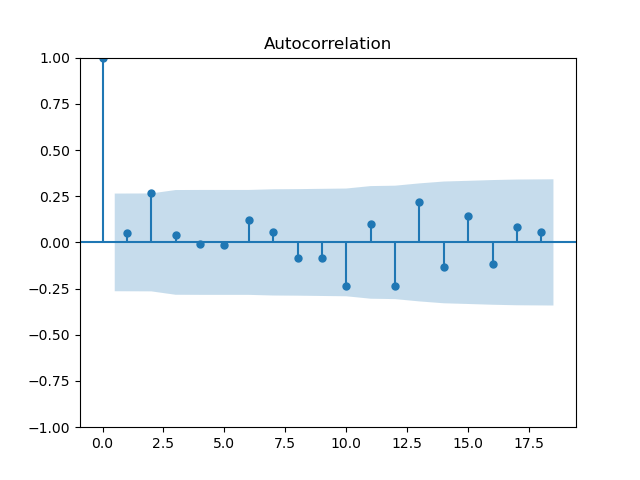
对于银川市降水量系列数据而言，ARIMA模型中参数p应取1，q可以取1

下图为济南市降水量系列数据的ACF图与PACF图：



对于济南市降水量系列数据而言，ARIMA模型中参数p应取2，q应取0

下图为大连市降水量系列的ACF图与PACF图：



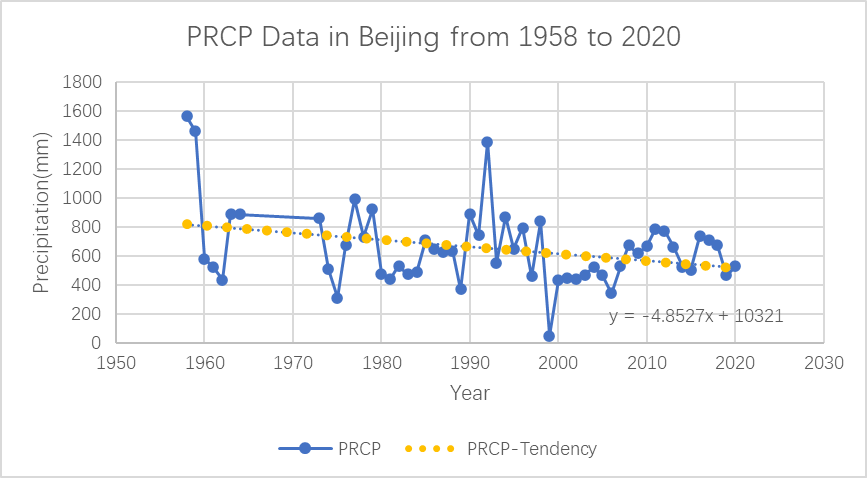
对于大连市降水系列数据而言，ARIMA模型中参数p可以取1，q可以取1

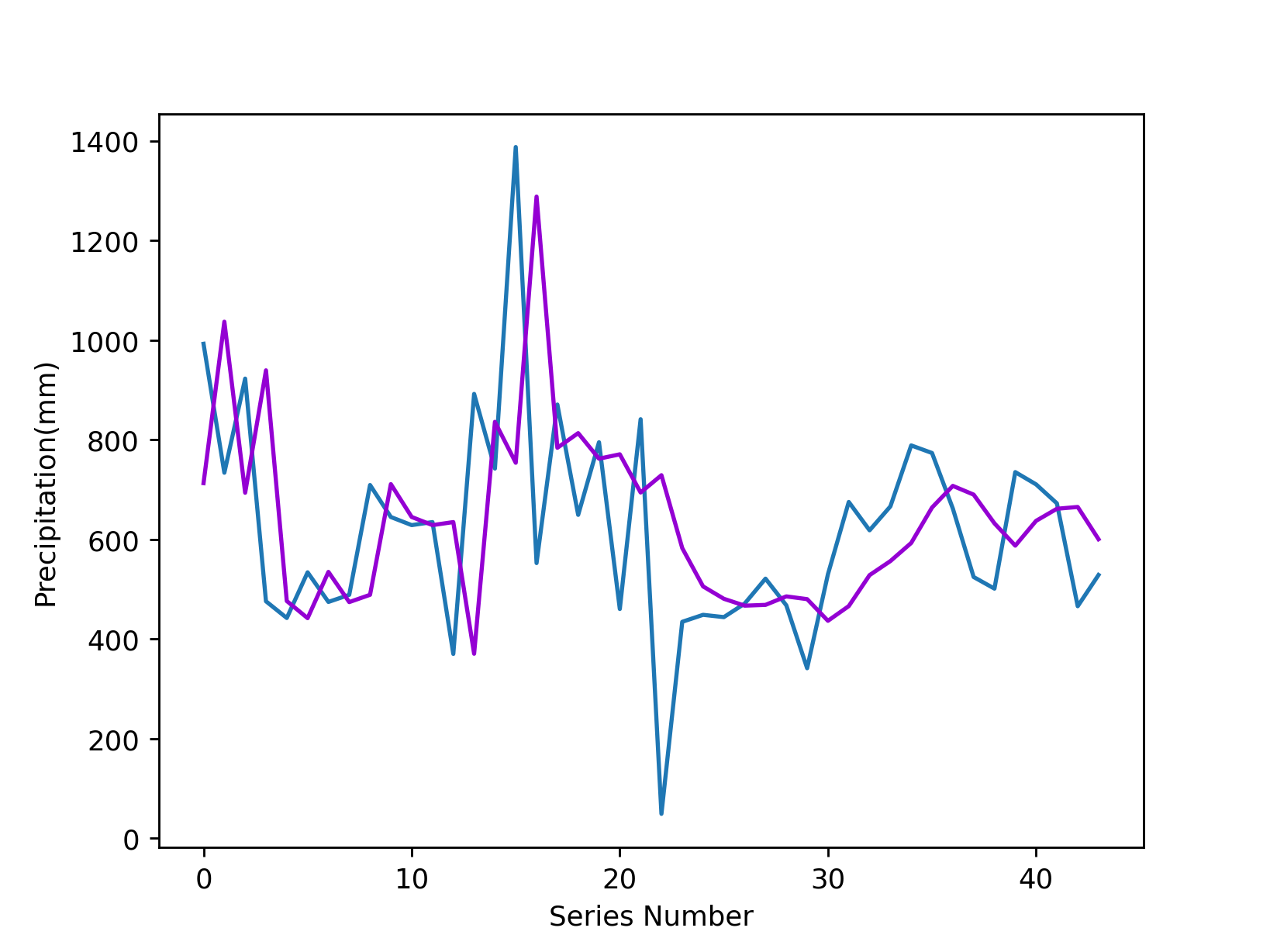
（3）构造模型

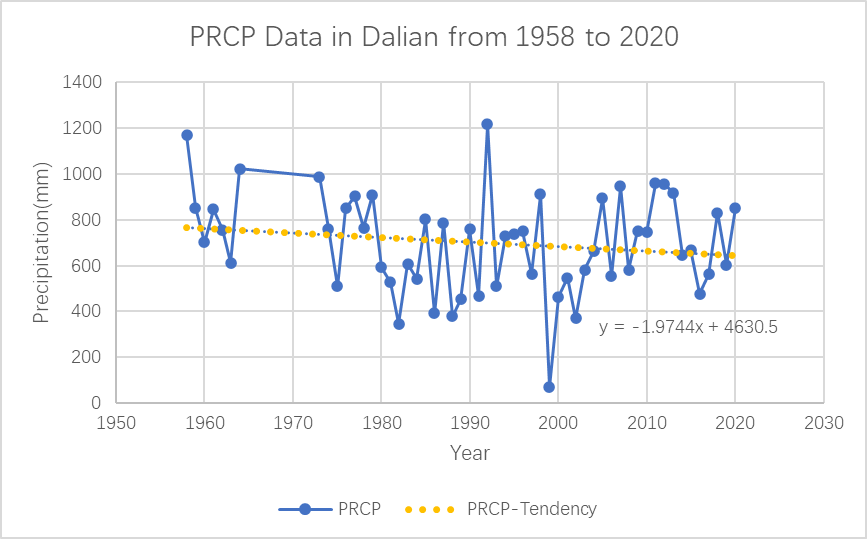
根据之前求得的参数，可以构造5个ARIMA模型，即：

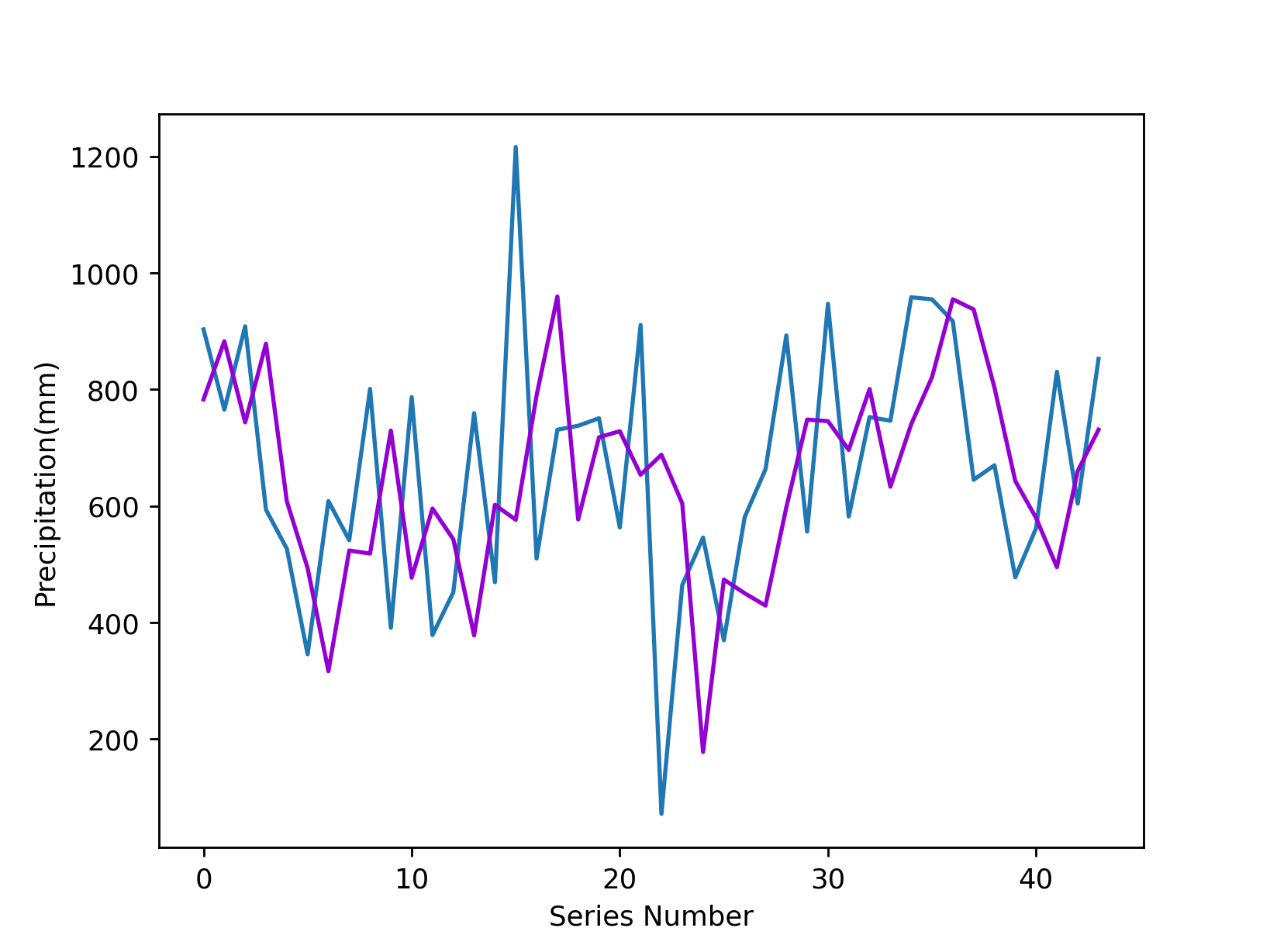
#### 3.2.4 降水趋势分析

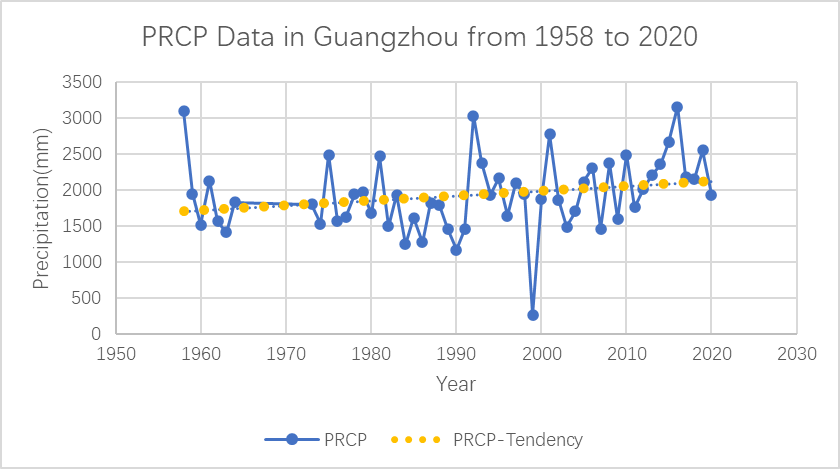
在构造ARIMA模型后，可以分别作出5个城市的降水量变化图。从图中可以直接观察出降水量变化趋势。

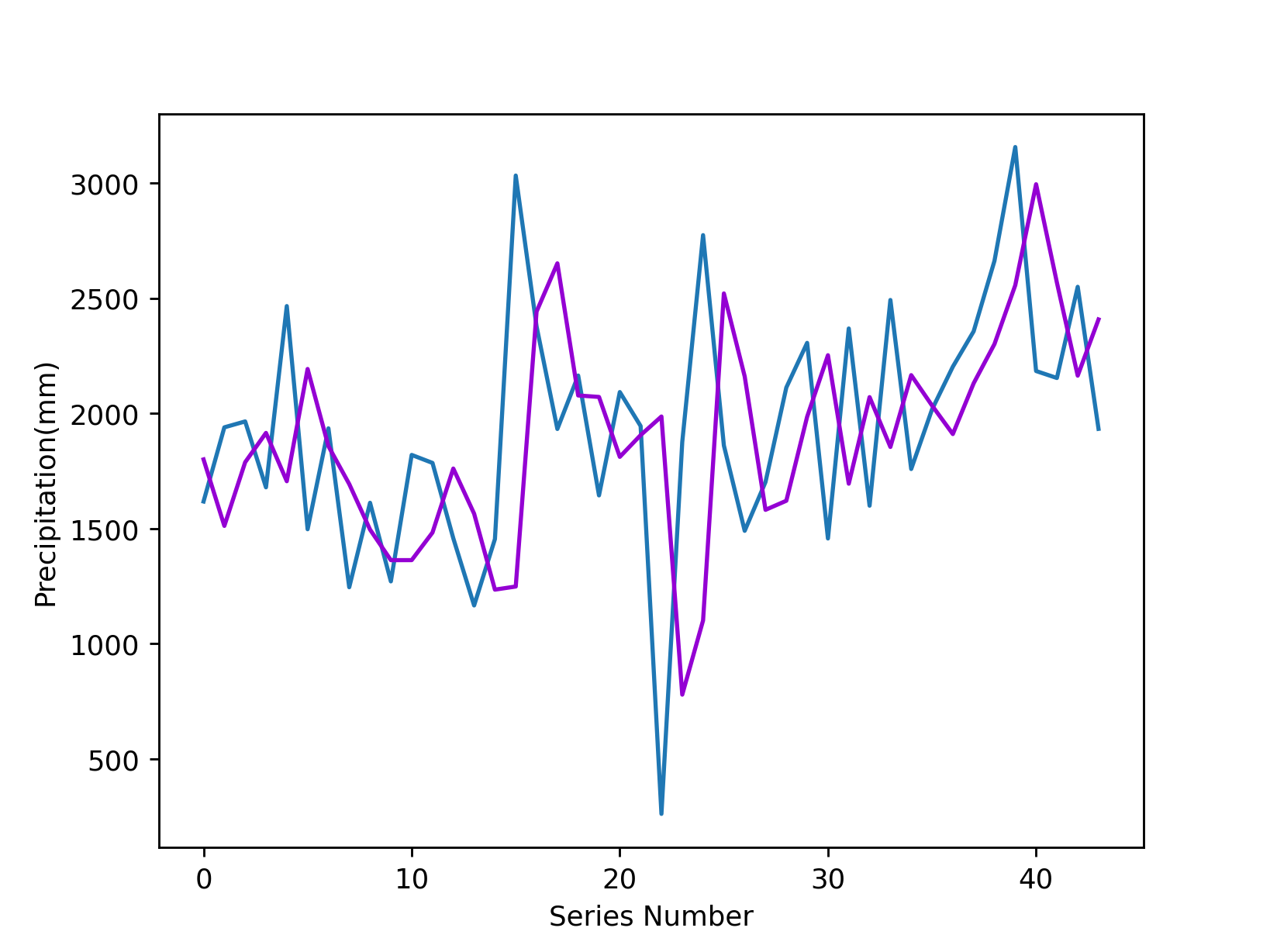


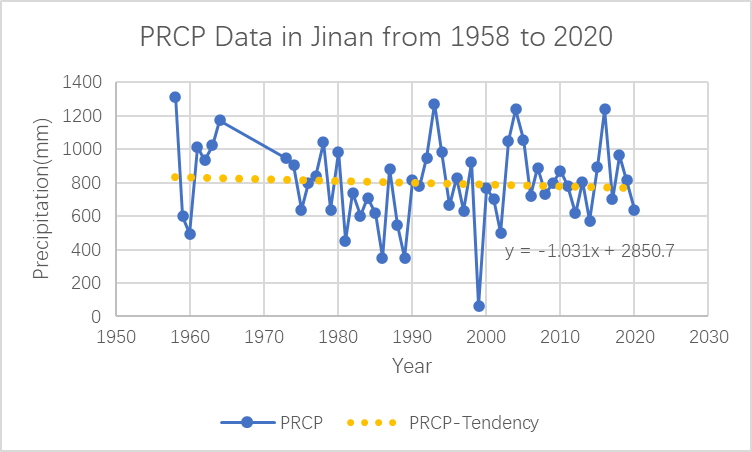


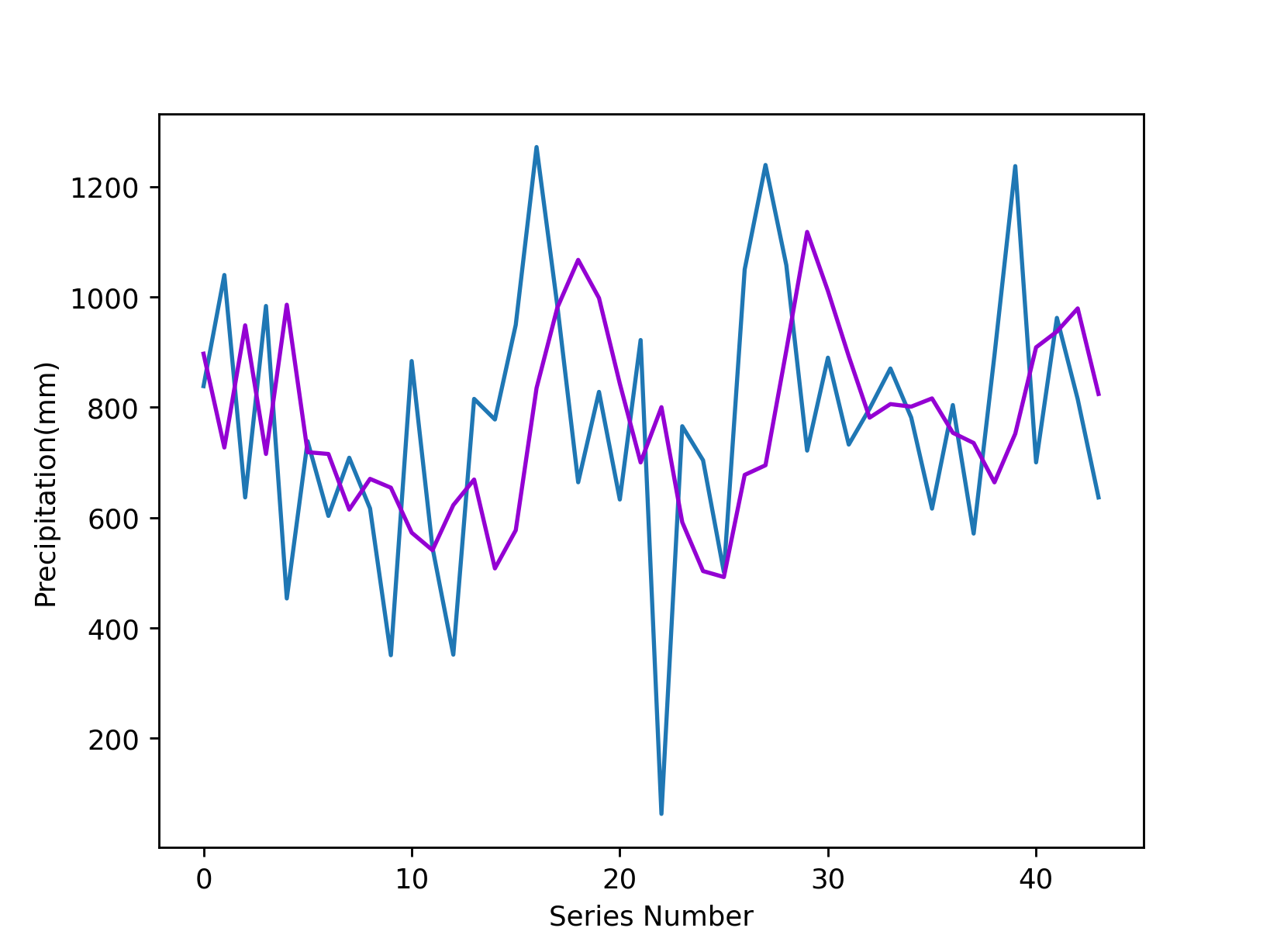


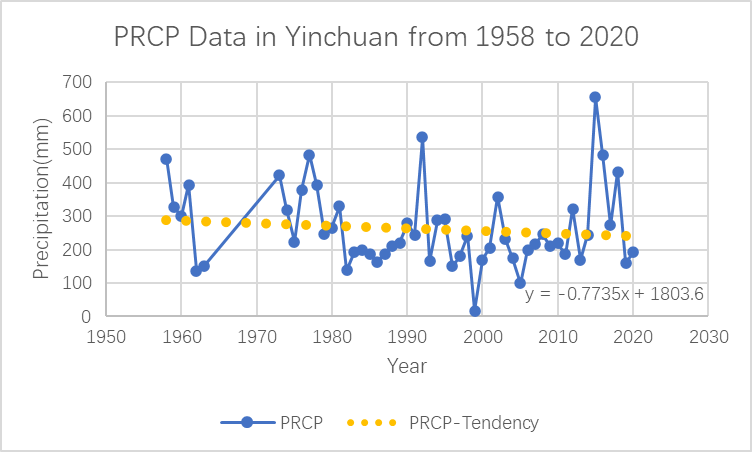


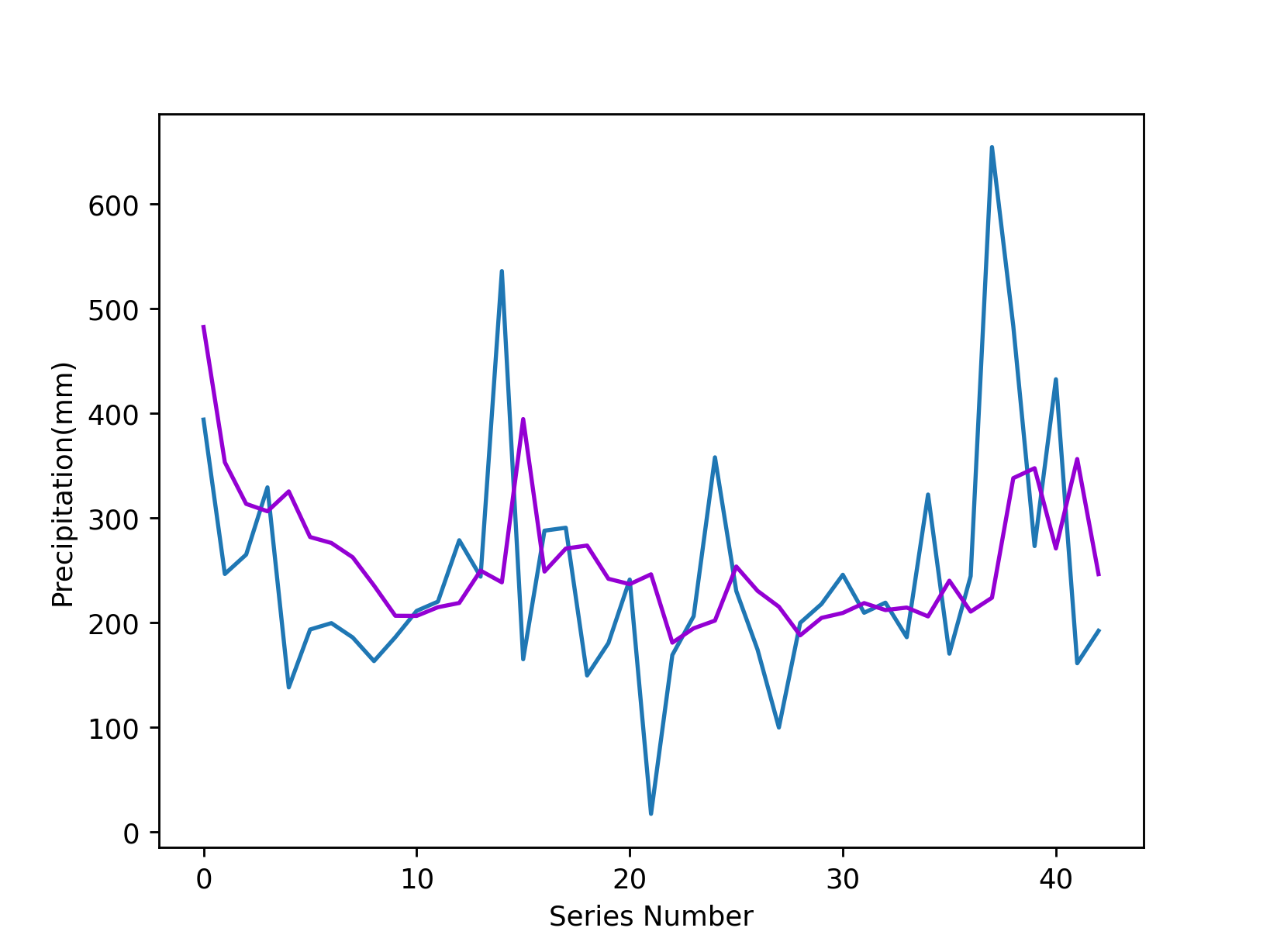












### 3.3 Model 3: 利用长短期记忆循环神经网络进行天气预测

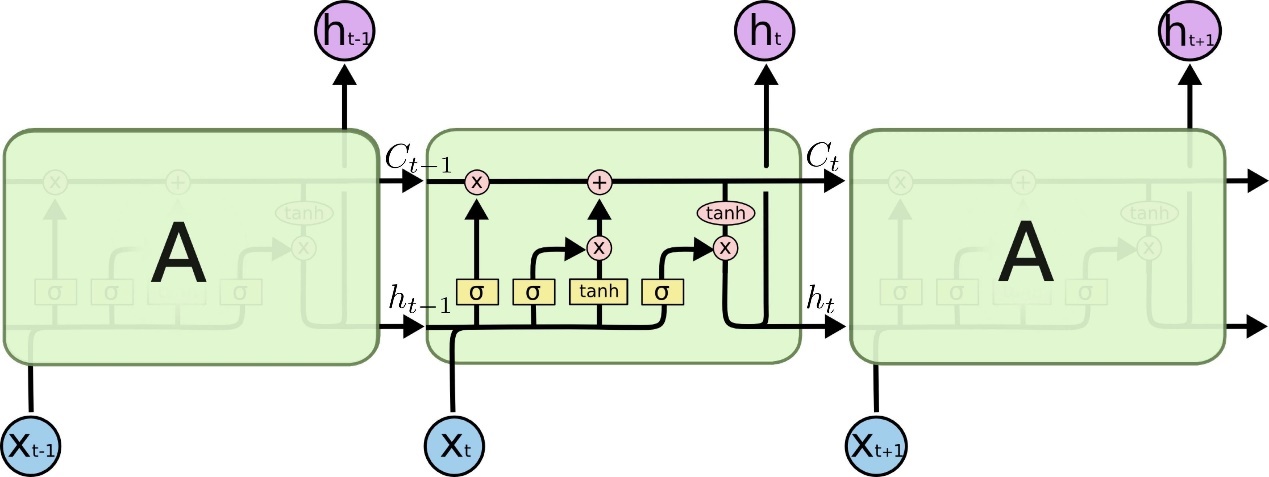
#### 3.3.1 数据来源

本模型的数据来源于北京市气象台所提供的从2020年1月1日起至2021年11月9日近2两年的记录。这些记录包含的指标有DEWP MAX MIN MXSPD PRCP TEMP VISIB WDSP，这些指标的含义与模型一中的定义相同。

#### 3.3.2 模型分析与符号表示

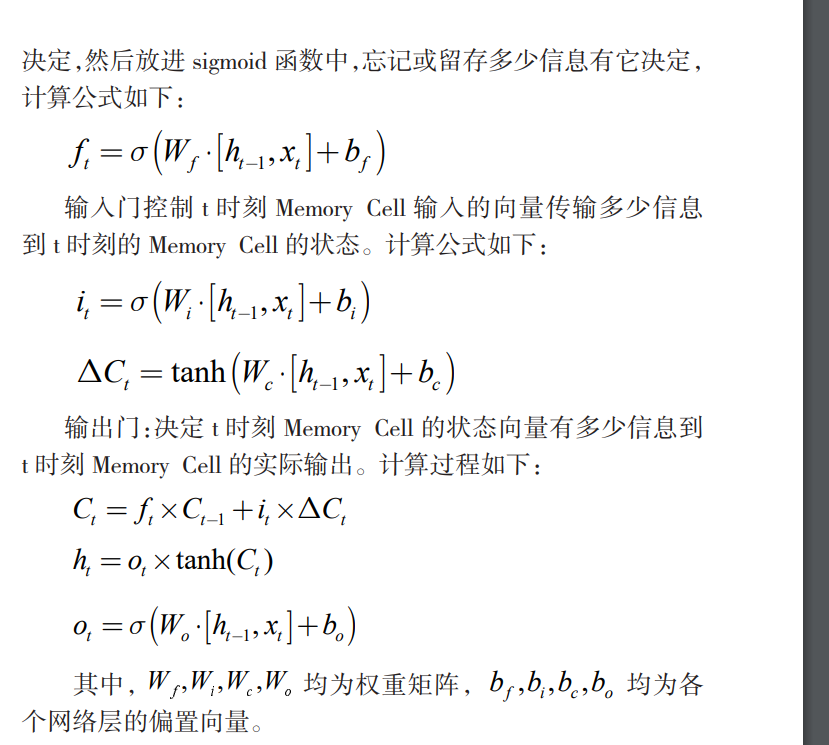
长短期记忆循环神经网络是基于RNN模型衍生出的特殊形式，其提出的目的是为了解决循环神经网络中存在的梯度消失、梯度爆炸等问题而提出来的一种优化算法。它能够通过门结构与细胞状态更新来更好的保存长期记忆，并通过存储单元保留与时间有关的信息来解决梯度消失和梯度爆炸问题。

LTSM神经网络隐层内部的链状结构如下图所示：



其每一个记忆细胞(Memory Cell)内部结构都有三个门结构，分别是遗忘门（forget gate），输入门（input gate），输出门（output gate）。





#### 3.3.3 模型构建

#### 3.3.4 天气预测和分析

## 4. 模型评估与优化（Model Evaluation And Optimization）

## 5. 优点与不足（Strengths and Weakness）

## 参考资料（References）

## 附录（Appendix）