分类号 TP399 学号 18020045

UDC 004.8 密级 公 开

工学硕士学位论文

基于机器学习的数值天气预报降水产品偏差订正方法研究

硕士生姓名 张永顺

学 科 专 业 计算机科学与技术

研 究 方 向 海洋信息工程

指 导 教 师 张卫民 研究员

国防科技大学研究生院

二〇二〇年十月

论文书脊

（此页只是书脊样式，学位论文不需要印刷本页。）

基于机器学习的数值天气预报降水产品偏差订正方法研究

国防科技大学研究生院

Research on the Method of Correcting Precipitation Product Deviation in Numerical Weather Prediction Based on Machine Learning

Candidate：Zhang Yongshun

Supervisor：Zhang Weimin

A dissertation

Submitted in partial fulfillment of the requirements

for the degree of Master of Engineering

in Computer Science and Technology

Graduate School of National University of Defense Technology

Changsha，Hunan，P.R.China

October，2020

独 创 性 声 明

本人声明所呈交的学位论文是我本人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表和撰写过的研究成果，也不包含为获得国防科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文题目：基于机器学习的数值天气预报降水产品偏差订正方法研究

学位论文作者签名：张永顺　　日期：2020 年 10 月 26 日

学位论文版权使用授权书

本人完全了解国防科技大学有关保留、使用学位论文的规定。本人授权国防科技大学可以保留并向国家有关部门或机构送交论文的复印件和电子文档，允许论文被查阅和借阅；可以将学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密学位论文在解密后适用本授权书。）

学位论文题目：基于机器学习的数值天气预报降水产品偏差订正方法研究

学位论文作者签名：张永顺　　日期：2020 年 10 月 26 日

作者指导教师签名：　　日期：2020 年 10 月 26 日

目 录

表 目 录

图 目 录

摘 要

近年来，降水引发的洪涝灾害频发，定量、定时、定点的降水预报显得尤为重要。数值预报是我国中短期降水预报的主要方法，但是数值预报模式输出的降水量往往与实际情况存在偏差，需要对数值预报产品进行偏差订正。然而传统方法通常都是基于线性相关分析挑选预报因子，再在线性相关的基础上建立预报模型，对于降水这种高度非线性的问题往往难以得到令人满意的预报效果。

本论文主要研究机器学习对数值预报模式降水产品的偏差订正。首先基于银河全球谱模式预报数据构造再预报数据集；然后利用随机森林对降水预报因子进行重要性评估，挑选出最合适的预报因子；其次通过机器学习方法对大量训练数据进行学习，建立预报因子与逐 3h 降水量之间的晴雨分类模型、降水等级分类模型以及降水量回归模型；最后对逐 3h 降水量预报进行偏差订正。主要工作如下：

1、构造了训练数据集并研究了基于随机森林的降水预报因子选取方法，利用随机森林对预报因子进行重要性评估，挑选出最合适的预报因子。模式数据选择观测站附近的四个网格点预报数据，不再将其插值到观测站点上，这样既避免了站点预报场插值计算的不准确，同时也考虑了站点附近天气过程的影响。

2、研究了基于机器学习的降水预报产品偏差订正方法，利用随机森林、支持向量回归和长短期记忆网络方法通过对大量训练数据的学习，建立预报因子与逐 3h 降水量之间的晴雨分类模型、降水等级分类模型以及降水量回归模型，从而得到更加准确的逐 3h 降水量预报。本文选择了安徽省黄山市的屯溪站进行实验，以 2017-2019 年的逐 3h 降水观测作为标签，银河全球谱模式相应的格点预报数据作为预报因子。实验结果表明机器学习方法的逐 3h 降水量预报结果较模式输出降水量和多元线性回归方法的预报结果更好，尤其是对于强降水的预报，在均方根误差和降水预报业务评价指标上均表现良好。

3、研究了低分辨率偏差订正模型对高分辨率数据的偏差订正适用性。再预报数据集的构造需要非常大的计算资源和存储开销，为了减小再预报数据集构造的计算资源和存储开销，同时为了使低分辨率偏差订正模型适用于高分辨率数据的偏差订正，在构造数据集时，需要对低分辨率数据进行基于机器学习的空间插值提升其分辨率，进而改进偏差订正的效果。本文设计了基于高斯过程回归的空间插值算法并以海表面温度进行实验验证，实验结果表明该算法的均方根误差比最近邻、双三次和双线性插值法更低，更能有效地提升数据的空间分辨率。

关键词：机器学习；降水预报；预报因子；偏差订正

ABSTRACT

In recent years, flood disasters caused by precipitation have occurred frequently. Quantitative, timing and fixed-point precipitation forecasts have become particularly important. Numerical weather prediction is the main method of medium- and short-term precipitation forecasting in China, but the precipitation output of numerical model often deviate from the actual situation, and further deviation corrections for numerical forecasting products are needed. However, traditional methods usually select predictors based on linear correlation analysis, and then establish a forecast model based on linear correlation. It is often difficult to obtain satisfactory forecast results for highly nonlinear problems such as precipitation.

This thesis mainly studied the deviation correction of the short-term precipitation products of the numerical weather prediction by the machine learning method. Firstly, construct a re-forecast dataset based on the forecast data of the Yinhe Global Spectrum Model, and then use the random forest method to evaluate the importance of the precipitation predictors and select the most suitable predictors. Secondly, use machine learning methods to learn from a large amount of training data to establish the clear and rain classification model, the precipitation grade classification model, and the precipitation regression model between the predictor and the 3h precipitation, and finally the deviation correction of the 3h precipitation forecast. Main tasks as follows:

1. Constructed a training data set and studied the selection method of precipitation predictors based on random forests. Random forests were used to evaluate the importance of predictors and select the most suitable predictors. The model data selected the forecast data of four grid points near the observation station and no longer interpolated them to the observation station. This avoided the inaccuracy of the interpolation calculation of the station forecast field and also considered the influence of the weather process near the station.

2. Research on the deviation correction method of precipitation forecast products based on machine learning, using random forest, support vector regression and long short-term memory network methods to establish the clear and rain classification model, the precipitation grade classification model, and the precipitation regression model between the predictor and the 3h precipitation through the learning of a large amount of training data, so as to get more accurate 3h precipitation forecast. In this thesis, the Tunxi Station in Huangshan City, Anhui Province is selected for the experiment, with the 3h precipitation observations from 2018 to 2019 as the label, and the corresponding grid point forecast data of the Yinhe Global Spectrum Model as the predictor. The experimental results showed that the 3h precipitation forecast result of the machine learning method was better than the forecast result of the model output precipitation and

the multiple linear regression method, especially for the forecast of heavy precipitation, both in the root mean square error and the precipitation forecast business evaluation index good performance.

3. The applicability of low-resolution deviation correction model to high-resolution data was studied. The construction of the re-forecast data set requires very large computing resources and storage overhead. In order to reduce the computing resources and storage overhead of the re-forecast data set construction, and to make the low-resolution deviation correction model suitable for high-resolution data deviation correction, When constructing a data set, it is necessary to perform spatial interpolation based on machine learning on low-resolution data to improve its resolution, thereby improving the effect of deviation correction. This thesis designed a spatial interpolation algorithm based on Gaussian process regression and experimentally verified it with sea surface temperature. According to the experimental results, the interpolation results of this algorithm had lower root mean square errors than nearest neighbor interpolation, bicubic interpolation and bilinear interpolation methods, could effectively improve the spatial resolution of the data.

Key Words：Machine learning, Precipitation forecast, Predictor, Deviation correction

第一章　绪论

1.1 研究背景和意义

近年来，我国多地洪涝灾害频繁发生，对人民群众的生命安全以及财产安全造成了极大的危害。2016 年，我国江淮、西南东部以及长江中下游地区等地发生严重的洪涝灾害，造成大量人员伤亡和农作物受灾。2020 年入汛以来，我国南方 20 多个省份出现了严重的洪涝灾害，受灾人次超过 3000 万，经济损失超过 800 亿元。降水尤其是强降水是导致洪涝灾害发生的主要原因，及时、准确的降水预报可以有效地减少洪涝带来的危害，对于防灾减灾工作意义重大。

然而，由于降水产生的物理过程和化学过程的复杂性，降水预报是天气预报中的重点和难点之一[1]。目前，降水预报主要依靠数值天气预报方法[2]（Numerical Weather Prediction，NWP）。数值天气预报的基本概念如图 1.1 所示，首先应用数学和物理学的知识建立大气方程组，即数值模式，根据实际的大气状态，在一定的初值条件以及边界条件下，对所建立的大气方程组进行时间积分，然后通过数值计算求解方程组，最终得到未来一段时间的各种气象要素的预报场，即预测未来的大气状态。

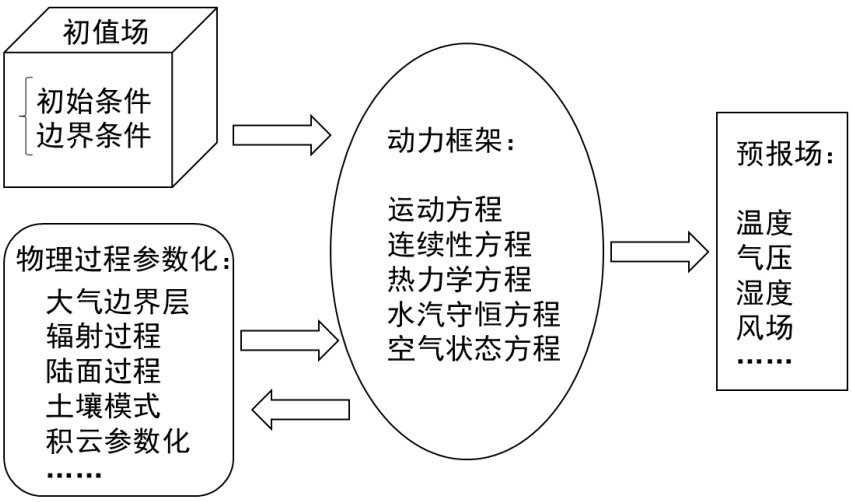

图 1.1　数值天气预报概念图

随着计算机技术以及探测技术的不断发展，数值天气预报产品的预报时效进一步延长，网格格距也不断地缩小。不仅如此，模式中的物理过程及参数化方案也不断地细化和完善，天气预报业务正逐渐向以数值预报为主的方向上发展。同

时，依靠数值模式的气象要素预报业务的预报准确率也不断提升，数值天气预报产品的准确率已经成为衡量现代的气象科学发展水平的几项重要指标之一。业务预报中的气象要素预报工作建立在数值天气预报产品的基础上，目前在气象领域的研究中已经有了广泛的应用，为了提高客观气象要素预报的预报性能，最重要的是要提高数值天气预报产品的准确率。

数值天气预报假设大气系统的运动变化遵循一定的自然规律，大气系统未来某一时刻的状态由其初始状态及其所遵循的自然规律所决定，因此可以将大气运动在时间和空间上的变化研究转化为给定初始条件以及边界条件下的大气方程组的求解问题。数值天气预报的核心是数值模式，它是一组极为复杂的方程组，一般来说由动量守恒方程、质量守恒方程、水汽守恒方程、能量守恒方程以及空气状态方程组成。然而，人类目前对大气运动的认识仍然是有限的，在这种情况下，数值模式不可能将所有的大气运动规律用一个方程组来完美表达，也就是说数值模式无法描述大气运动中的所有物理规律。此外，数值模式对于初值极其敏感。模式在启动时初始场与强迫场的不匹配导致模式需要一段时间来适应强迫场[3]，这个过程称为 spin-up。模式的 spin-up 阶段会造成一些预报偏差，这严重限制了模式的短临预报，尤其是强降水预报。由于观测误差的存在以及资料同化所带来的误差，根据目前的技术仍然难以得到绝对准确的初始场。以上事实说明，数值天气预报并不是完全精确的，数值预报的结果是存在一定误差的。

在实际的预报工作中，考虑到数值预报产品与观测实况之间存在的误差，基层的气象台站往往会增加一个天气会商过程[4]。目前主要以人工天气会商为主，依靠预报员自身的预报经验和实况数据对数值预报产品进一步解释应用，从而得到最终的预报结果。然而随着预报格点化、精细化的不断发展，人工天气会商已经逐渐无法适应天气预报发展的需要，为了更进一步地提升预报效率及预报精度，利用机器学习算法模拟天气会商这一过程，成为了数值预报产品应用研究以及天气预报智能化的重点之一。

数值预报产品包括各种气象要素如温度、风场以及降水等。其中，数值预报模式输出的降水量场相较于其它的物理量场的精度更低[5]，如果直接将其用于降水预报，则会产生较大的误差，尤其是对于强降水的预报。为了获得更加准确的降水预报结果，需要对数值天气预报的降水产品进行偏差订正。数值预报产品解释应用就是对数值预报产品这一综合性的结果，利用统计学、动力学以及人工智能等方法，结合观测实况数据，建立偏差订正预报模型，进一步提高数值预报产品的预报精度[6]。

以降水为例，假设通过数值预报产品解释应用方法对某一站点的降水量进行预报。则该站点的降水量即为预报对象，选取与降水量密切相关的一些气象要素

作为预报因子，再利用同一时期的预报因子和站点降水量实况资料，采用多元线性回归方法建立站点降水量的预报模型。进行预报时，只要将相应时刻的数值预报产品作为输入，就可以得到最终的降水预报结果。

然而，降水预报一直是数值预报产品解释应用中的难点，其释用预报的效果不如温度、风场等气象要素[7]。这是因为当前对于数值预报产品的解释应用基本上都是应用相关分析以及线性回归分析，这些方法都是建立在线性相关的基础上的统计学 ，然而对于降水预报这种非线性问题存在一定局限性[8]。降水量是一段时间的累积量，一般为 1 小时、3 小时、6 小时以及 12 小时等。以 3 小时为例，逐 3 小时降水量并不是一个在时间上连续的变量，也并非某个时刻上的观测量，这就决定了它的非线性。并且，与温度等正态分布的气象要素不同的是，降水量是偏态分布的，一般来说，无雨和小雨的情况较多，大雨和暴雨的情况则较少，如果利用传统的统计学方法直接对降水量进行偏差订正是存在问题的。

近年来，人工智能以及机器学习在各个领域的应用越来越广泛。气象领域存在的大量数据十分适合机器学习的应用，国内外也有一些研究使用人工智能的方法进行天气预报，并取得了一定成果[9]。传统的天气预报方法旨在研究大气运动的规律和原理，而人工智能将其视为一种模糊的“黑箱”问题，通过大量数据的训练和学习，建立机器学习的预报模型，从而解决气象领域中存在的无法直接通过数学物理模型解决的预报问题[10]。

基于数值预报产品解释应用的气象要素预报问题有两个关键环节，一是选取合适的预报因子，二是建立相应的预报模型[11]。传统的解释应用方法普遍采用基于统计学的相关分析和线性回归分析来处理这两个问题。本文将机器学习算法应用到数值天气预报降水产品的解释应用中，使用机器学习算法分别对筛选预报因子和建立预报模型这两个关键环节进行改进，并以地面气象观测站的实际降水量为例，证明其能够对数值预报产品进行有效的偏差订正，从而提供更为准确的降水量预报结果，进而提高数值天气预报系统的精细定量预报水平和智能化水平。此外，目前业务上通常对 T1279L137 分辨率的数值预报产品进行解释应用，而本文使用的数值预报产品的分辨率为 T639L137，这是因为目前仍然没有成熟的数据集可以使用，而重新构造数据集的难度极大。低分辨率的数值预报产品可能无法满足解释应用的需求，于是对数值预报产品进行空间插值进而提高分辨率就变得十分必要。许多机器学习算法都可以应用于空间插值，为了更好地进行数值预报产品的解释应用，本文还研究了一种用于空间插值的机器学习算法，并设计了基于海表面温度（SST）的插值实验进行验证。

1.2 国内外研究现状

1.2.1 降水预报

降水的产生过程十分复杂，它是由一系列不同尺度的物理过程同时作用的综合结果[12]。目前，国内外预报短期降水的方法主要有三种，分别是天气学方法、雷达回波外推预报方法以及数值天气预报方法[13]。

天气学方法是早期发布降水预报的方法，它以天气图为基础，根据预报员的经验和预报知识，主要对影响降水的一些典型天气系统进行预报[14]。其缺点是预报结果并没有经过客观的定量计算，因此不同预报员依据自身的经验知识所做出的降水量预报可能存在较大差异。此外，中小尺度暴雨在天气图上很难分辨，因此天气学方法的降水预报存在着一定的局限性。

随着高时空分辨率的遥感技术的发展，基于多普勒雷达资料的雷达回波外推技术已经广泛应用于短时降水预报。多普勒雷达回波外推图具有 6min 的高时间分辨率和 1km 的高空间分辨率，包含着丰富的时序信息和空间信息[15]。对于 2 小时以内的定量降水预报，雷达回波外推预报方法较其它方法拥有更高的时空分辨率和预报准确性。但是对于更长时间的定量降水预报而言，数值天气预报相比于雷达回波外推技术则具有明显的优势。

数值天气预报是集大气探测、天气学、动力气象学以及计算机通信技术为一体的综合性科学，它通过数值计算求解一定初值条件以及边界条件下的模式方程组，利用高性能计算机来预测未来的大气状态[16]。随着数值天气预报以及相关技术的快速发展，数值天气预报不再是简单的形势预报，已经逐渐扩展为各种气象要素在内的要素预报，并且其预报时效从 1-2 天延长至 3-10 天，预报的地理空间范围也扩展到平流层。然而，由于受到模式初始场、边界条件、地形以及次网格过程等各种因素的限制，数值天气预报的降水量场不可能绝对准确，与真实的降水量实况数据相比一定会存在着一些偏差[17]。

1.2.2 数值预报产品释用

数值天气预报产品是现代天气预报业务的基础之一，为了获取更为精确的气象要素以及天气现象的预果，仅仅使用数值预报的结果仍然不够[18]。定时、定点、定量的气象要素预报是当前基层气象台站天气预报业务的主要任务，要求预报准确率要高，对于气象要素的预报产品要有针对性。然而，目前的数值天气预报的要素预报能力仍然难以满足实际应用的需求，在季节转换时数值预报对天气形势预报还很不稳定，对一些灾害性、转折性天气的预报能力仍然不足，比如强对流、台风和暴雨等[19]。

同时，不同数值预报模式有不同的特性，数值预报的结果在一定程度上存在系统性误差，可以对数值天气预报的系统性误差做出订正，进而提升数值天气预报的预报效果[20]。尽管大多数数值预报产品已经具有较高的质量，但是考虑到其仍然不可避免的误差，我们不可能将任何数值预报产品直接用于预报。在修正这些预报误差的过程中，需要通过对数值预报产品进行偏差订正，不断挖掘数值天气预报产品的预报潜能，使其预报精度得到进一步的提高。

近几十年来，国内外的数值预报产品释用技术发展迅速。早在上世纪 50 年代，美国国家气象局就开展了完全预报（Perfect Prediction，PP）方法的试验，60 年代实现了 PP 方法的业务运行，70 年代建立了基于模式输出统计法（Model Output Statistic，MOS）的预报系统[21]，逐渐实现了各种基础的气象要素的预报。随后，日本、加拿大、英国等也相继建立了基于 PP、MOS、卡尔曼滤波、神经网络等方法的气象要素预报系统，并取得了较好的预报效果。Zbynk 利用 MOS 方法预报了欧洲地区 7 条江河流域的逐日降水量，实验结果表明 MOS 方法预报的逐日降水量结果比数值天气预报方法有较大的提升[22]；Kretzschmar 等使用神经网络方法预报了 24 小时局地风速，并得到了较为准确的预报结果[23]；Andtew 等利用神经网络方法预报了机场云雾的消散，并与线性回归方法的预报结果进行对比，实验表明神经网络方法拥有更好的预报效果[24]；Laurence 等对比分析了加拿大的业务预报中常用的 UMOS（Updateable Model Output Statistics）系统、神经网络以及卡尔曼滤波等方法的预报性能，结果显示不同方法都有优点和缺点，在具体应用时需要考虑实际情况进行选择[25]。

我国的数值预报产品解释应用技术相对来说开展较晚。1982 年，国家气象中心 B 模式投入业务使用，随后在此基础上开展了一些数值天气预报产品释用工作，主要以模式直接输出和统计学释用方法为主。经过多年的努力，我国的数值预报产品释用技术取得了很多研究成果，并在实际的业务预报中取得了较好的效果[26]。孙永刚等使用我国的高分辨率有限域预报系统（HLAFS）的格点预报资料，建立了内蒙古地区降水的 MOS 预报方程，证明了 MOS 方法对降水具有一定的预报能力 [27]。陈力强等利用我国 MM5 模式预报产品和气象站点的降水实况观测数据，使用 MOS 方法建立了辽宁省县级站的晴雨以及雨量等级预报模型，实验结果表明降水的 MOS 预报模型具有一定的预报能力，并将其应用到业务中[28]。

1.3 研究内容

相对于温度、湿度等气象要素预报来说，降水预报一直是数值预报产品解释应用中的难点问题，预报效果一直不是很理想，尤其是对于强降水的预报。目前使用较为普遍的方法是降水概率预报（Precipitation Probability Forecasts, PPF），

将降水预报问题转化成二分类问题或者多分类问题，预报结果以百分率表示降水是否发生或者发生各级降水的可能性。本文将降水预报视作一种回归问题，研究对象为 3h 累积降水量，使用机器学习算法建立数值预报模式输出的各种物理量场与降水量之间的回归关系，从而实现数值预报降水产品的偏差订正。

本文工作主要包括以下三个部分：

1.使用随机森林方法评估预报因子的重要性，从而避免了线性相关分析的局限性，挑选出更加合适的预报因子。具体来说，影响降水的气象要素众多且它们与降水量之间的关系是非线性的，传统的线性相关分析难以挑选出对降水量影响最显著的那些预报因子。而随机森林是多棵决策树集成学习的机器学习算法，可以对降水预报的预报因子进行重要性评估，从而挑选出更合适的预报因子。

2.将多种机器学习算法应用到降水预报产品的解释应用中，进而得到更加准确的降水量预报结果，从而提高数值预报产品解释应用的智能化水平。随机森林模型可以建立降水晴雨分类和降水等级分类模型，对相应预报时效内的模式预报降水进行订正。此外，针对降水量的预报是典型的非线性回归问题，机器学习中随机森林和支持向量机算法可以很好地解决这类问题。此外，降水预报是一种典型的时间序列预测类问题，即通过某种现象一段时间的状态来判断其未来一段时间的状态，而深度学习中的长短时记忆网络（LSTM）在时间序列预测类问题上表现良好。相比于传统的线性回归分析，这些机器学习算法可以通过对某一气象观测站点大量历史数据的训练和学习，建立起此站点的降水量与预报因子之间的非线性回归模型，从而得到更加精确的定时、定点、定量的降水量预报结果。

3.研究了一种基于高斯过程回归的数值预报产品的空间插值方法。受到模式分辨率的限制，许多数值预报产品的空间分辨率有时无法满足解释应用的需求，于是对数值预报产品进行空间插值进而提高分辨率就变得十分必要。传统的插值方法只考虑了与待插值点邻近的区域，再利用数学方法计算得到插值点处的结果，但是由于气象、海洋这些地理数据往往具有较高的空间变异性和复杂的非线性，传统方法的插值结果仍然不够精确。高斯过程回归作为一种核方法，通过引入核函数可以有效地处理这类非线性问题并进行空间插值。为了更好地进行数值预报产品的解释应用，本文研究了一种基于高斯过程回归的数值预报产品的空间插值方法，并针对海表面温度设计了插值实验进行验证。

1.4 论文结构

本文共分为六个章节，图 1.2 为本文的总体框架。

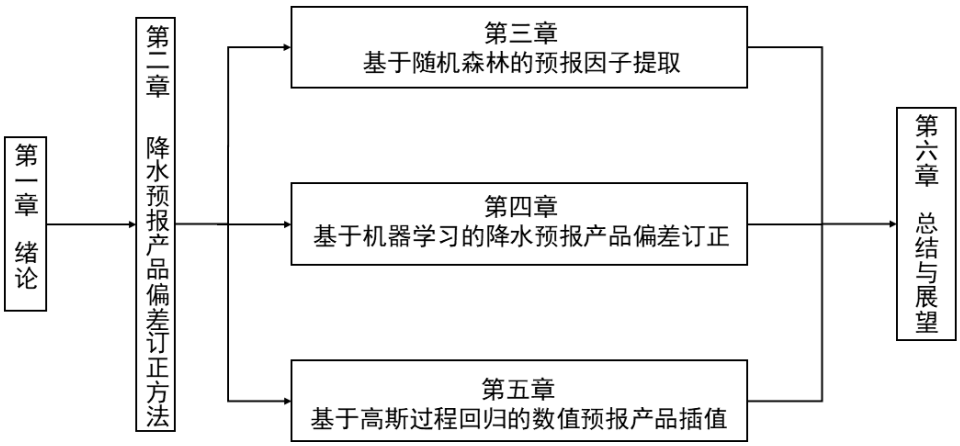

图 1.2　论文的总体框架

具体内容组织如下：

第一章，绪论。首先论述了降水预报研究的重要性，指出了目前数值预报模式输出的降水产品的精度较低、降水产品的统计学释用方法存在的局限性和不足之处，以及使用机器学习方法解决这些问题的可行性。然后对降水预报和数值预报产品释用的国内外研究现状进行了论述，最后明确了研究内容以及论文结构。

第二章，降水预报产品偏差订正方法。首先介绍了目前普遍采用的数值预报产品的统计学释用方法，重点阐述基于 MOS 法的降水预报产品偏差订正方法的基本流程，并分析其存在的不足之处和局限性。然后介绍了可以应用于降水预报产品偏差订正的相关机器学习算法，分别对随机森林、支持向量回归、长短期记忆神经网络以及高斯过程回归方法的理论依据与算法特点进行了描述。

第三章，基于随机森林的预报因子提取。首先介绍了课题研究所使用的实验数据，分析模式预报的降水量存在的不足之处，并指出预报因子筛选的重要性。然后介绍传统的基于线性相关的预报因子选取方法并选取预报因子。最后利用随机森林重要性评估方法对预报因子进行筛选后与传统方法的结果对比，并根据降水的影响因素分析对预报因子进行解释。

第四章，基于机器学习的降水预报产品偏差订正。在随机森林筛选预报因子的基础上，使用随机森林建立预报因子与逐 3h 降水之间的晴雨分类模型和降水等级分类模型，然后分别使用随机森林、支持向量回归和长短期记忆神经网络建立预报因子与降水量之间的回归模型，对模式预报的降水量进行偏差订正，最后根

据选定的评价指标与模式输出的降水量和统计学释用的预报结果对比，进而验证机器学习算法的有效性。

第五章，基于高斯过程回归的数值预报产品插值。通过构造一个组合的核函数，以影响待插值变量的一些物理因素作为特征输入，使用高斯过程回归方法建立插值模型。最后设计了一个针对海表面温度的插值实验，与传统插值方法进行对比并分析实验结果。

第六章，总结与展望。首先对本文的研究进行总结并概括了本文的创新点，然后分析了研究中仍然存在的几点局限性和不足，并对未来的研究方向进行了展望。

第二章　降水预报产品偏差订正方法

2.1 统计学释用方法

2.1.1 数值预报产品的统计学释用

数值预报产品的解释应用方法较多，主要可以归纳为四种[6]：1.以经验预报为主的天气学释用方法，这种方法只是定性的方法，并不能给出定量的预报结果；2.模式直接输出方法（Direct Model Output，DMO），将模式的格点预报值直接插值到观测站点上，从而得到该站点的预报结果；3.统计学释用方法，包括 PP 法、MOS 法、卡尔曼滤波方法等[29]；4.人工智能方法，包括神经网络方法（Neural Networks, NN）等[30]。目前，国内外普遍用于业务的数值天气预报产品释用方法多为统计学释用方法，其中以 PP 方法和 MOS 方法为代表。

PP 方法是 1959 年由美国气象学家 Klein 提出的，其基本思路是将某一时刻的实际气象参数（如温度、湿度、风场、气压等）作为预报因子，与预报对象（例如降水量）之间建立预报模型，预报时再将模式预报值作为预报因子代入预报模型中，即可求得相应时刻的预报结果。MOS 方法是 Glathn 和 Lowry 在上个世纪 70 年代提出的，其基本思路是将数值预报产品当作预报因子，与相应时刻的预报对象的实况数据（例如降水量实况数据）之间建立预报方程，最后将预报因子（即模式预报值）代入到预报方程中得到相应的预报结果[21]。

PP 方法和 MOS 方法的预报模型的数学基础均为统计回归方程，预报对象为某一时刻的天气现象或者气象要素。它们的不同点是方程中预报因子的选择（即回归方程的自变量），PP 方法以实况观测值作为自变量，而 MOS 方法以模式预报结果为自变量。PP 方法在建立方程时没有将模式预报的系统误差考虑进去，而 MOS 方法则在方程中修正了模式预报的误差以及不确定性，它的预报效果要优于 PP 方法，因此逐渐成为数值天气预报产品释用的主要方法。PP 法和 MOS 法的技术流程分别如图 2.1 和图 2.2 所示。

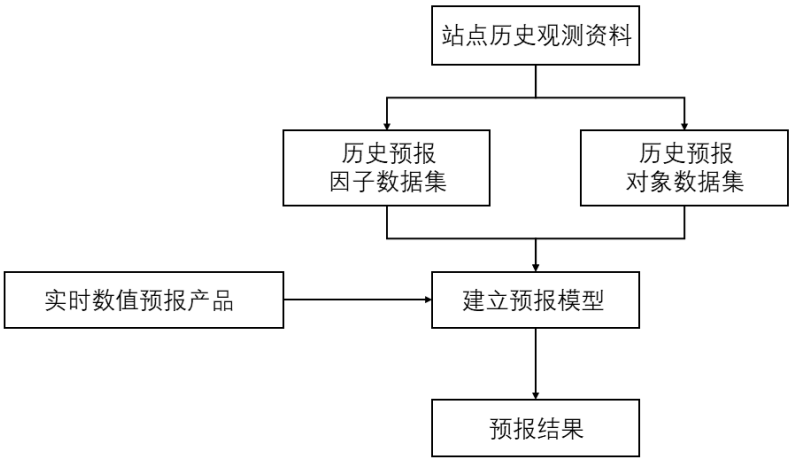

图 2.1　PP 方法技术流程

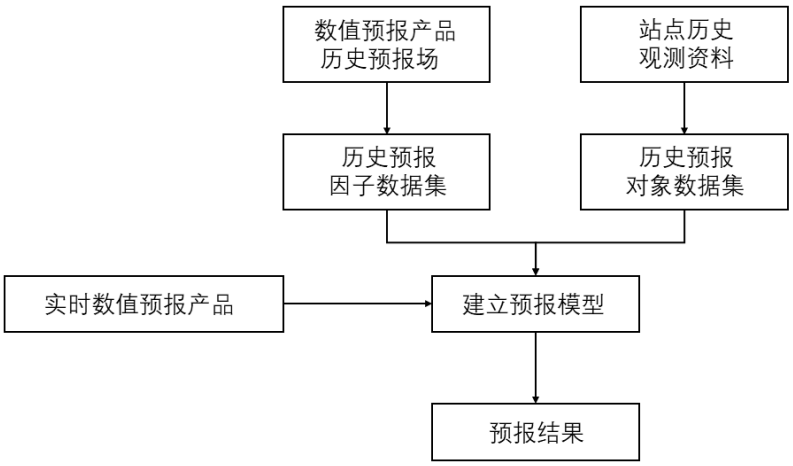

图 2.2　MOS 方法技术流程

2.1.2 基于 MOS 方法的降水预报

MOS 方法建立在多元线性回归模型的基础上[21]。相比于传统的一元线性回归，如果因变量同时受到两个以及两个以上的自变量的影响，就称为多元线性回归[31]。

对于降水预报问题而言，就是研究降水量 Y 与多个预报因子（x_1，x_2，$x_3 \cdots x_n$，n 为正整数）之间的回归方程。

$$Y = x_1 * a_1 + x_2 * a_2 + x_3 * a_3 + \cdots + x_n * a_n + b \tag{2.1}$$

其中，x_1，x_2，$x_3 \cdots x_n$为选取的预报因子，即数值预报产品；而a_1，a_2，$a_3 \cdots a_n$和b为待求解的回归系数。回归方程建立之后，需要对回归方程进行显著性检验，一般采用 F 检验[31]，通过显著性检验得到最终的预报方程。

此外，有时也使用逐步回归方法建立回归方程，该方法将筛选预报因子和建立预报方程作为一个整体。与一般的线性回归方法不同的是，逐步回归将预报因子逐个引入到方程中。每一个预报因子在引入之后都要对回归方程进行 F 显著性检验，并对方程中的每一个预报因子进行 t 检验[31]。当预报因子因为之后的因子的引入而无法通过 t 检验时，则从回归方程中删去该预报因子。这样可以确保每次引入新的因子之后方程中只包含显著性预报因子。这是一个反复的过程，直到回归方程中既没有显著的因子选入也没有不显著的因子剔除为止，以保证最终所得到的预报因子是最优的。

MOS 方法已经广泛应用于基层气象台站的要素预报中，且取得了较好的预报效果。使用 MOS 方法进行降水预报主要包括以下五个部分：资料来源与处理、预报因子预选、预报因子和预报对象之间的相关性分析、建立 MOS 预报方程和预报效果检验及分析[32]。

资料来源与处理。研究所用资料通常分为两个部分，一是数值预报模式输出的各要素场，二是研究地区气象站点的逐小时降水实况观测资料，一般来说要求至少 1-2 年的数据。预报对象为气象站点的降水量，有时也对降水量进行分级处理；预报因子即为数值预报模式输出的物理量场，一般是将格点上的预报结果内插到气象观测站点上，可以选择站点附近的 4 个格点利用双线性插值法进行插值，或者定义一个权重函数衡量附近更多格点对该站点的影响，以引入更多格点值进行插值。

预报因子预选。由于数值预报产品种类繁多、数量巨大，在建立预报方程前，往往需要先预选一些预报因子。考虑研究区域地形、气候等各种影响因素，一般从以下 3 个方面选取物理意义明确的因子：1）水汽因子：降水产生的基础是充足的水汽，能够描述水汽的气象因子有比湿、相对湿度和总柱水汽量等。2）热力不稳定因子：包括位势高度、温度、2 米露点温度等。3）动力因子：主要描述降水产生及发展的动力条件，包括散度、水平纬向风、水平经向风、垂直速度、地面气压、对流有效位能等。

预报因子和预报对象之间的相关性分析。首先计算每一个预选因子与降水量之间的相关系数，并对它们进行 t 检验，将能通过 0.05 显著性 t 检验的因子按相关系数大小排序，一般最终选择 10 个预报因子左右。

建立 MOS 预报方程。为了提高预报精度，可以对数据按照不同季节划分不同的时间段，例如以 3-5 月份、6-8 月份、9-11 月份、12 月至次年 2 月份划分春夏秋冬。通常，一年之中降水最多是 6-8 月份。根据不同时段的预报因子和对应的降水实况资料，一般使用多元线性回归建立降水量预报方程并进行 F 检验。

预报效果检验及分析。TS 评分（threat score，TS）是业务上经常使用的降水预报评估方法，其值域为 0~1，评分越大说明预报效果越好。除此之外，如果进行降水量的偏差订正，还可以使用均方根误差（RMSE，Root Mean Squared Error）进行评估，RMSE 越小代表预报效果越好。

基于 MOS 方法的降水预报的基本流程如图 2.3 所示，

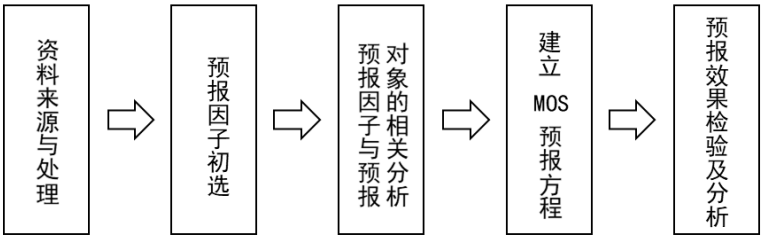

图 2.3　基于 MOS 方法的降水预报的基本流程

MOS 方法中的预报因子以单相关系数反映预报因子对预报对象影响的重要性程度[32]。但是相关系数只能用来描述两个变量之间线性函数关系，并不能真实反映使用这些预报因子进行预报的预报效果。虽然相关分析和回归的方法在处理温度等强线性相关的气象要素时表现优异[31]，但是对于较复杂的非线性问题表现则存在一定局限性。显然，降水由于其复杂的形成机制以及各种不断变化的气象要素影响，是一个典型的非线性问题。此外，降水还受到各个地区地形的影响，仅仅利用相关系数挑选预报因子并不一定能找到合适的预报因子。

MOS 方法建立预报方程的数学基础是多元线性回归，而多元线性回归除了假设预报对象和预报因子之间线性相关之外，还需要预报对象应该满足正态分布，如果不符合这个条件，那么预报效果就会变得不稳定，预报模型也就难以具备参考性。而对于降水来说，由于降水量不是连续变量，而是一段时间的累积值，降

水量是严重偏态分布的变量，无雨和小雨的情况较多，大雨和暴雨的情况较少。在这种情况下，对降水使用MOS方法建立预报模型可能难以得到较好的预报效果。

2.2 相关机器学习方法

为了提升降水预报的预报准确率，就需要更为合适的建模方法。尽管线性相关分析和多元线性回归的方法在处理温度这类强线性相关数据时预报效果较好，但是对于非线性问题例如降水预报时的预报效果则较差。机器学习中有许多方法适用于处理这类非线性问题[33-35]。孙全德等分别使用 LASSO（Least absolute shrinkage and selection operator）回归、随机森林和深度学习算法对欧洲中期天气预报中心预报的 10m 风速进行订正，订正结果显示优于 MOS 方法[36]。孙俊奎等使用概率神经网络、支持向量机和逻辑回归算法对数值预报产品中的逐 3h 降水量进行偏差订正，其中支持向量机模型的 TS 评分达到 40%以上，有效地提高了降水预报的准确率[37]。众多机器学习算法中，随机森林、支持向量回归和长短时记忆网络在处理数值天气预报产品的偏差订正中效果较好[38]。

2.2.1 随机森林

随机森林（Random forests, RF）是一种基于集成学习的机器学习算法，能够将多个决策树集成在一起，被广泛应用于处理各种分类问题和回归问题。RF 方法由 Leo Breiman 提出[39]，其基本思路是首先使用自助法（bootstrap）重采样[39]，对训练数据集 D 中的样本进行采样，生成新的数据集。新数据集中的多个决策树就组成了一个随机森林模型，由每个决策树的分类或回归结果决定整个随机森林模型的最终预测结果。

集成学习的优势在于，单个决策树的分类或回归能力也许较小，但是通过集成许多个决策树之后，随机森林可以利用每个决策树的预测结果的投票或者平均值决定最终的分类或者回归结果，因此集成的机器学习算法一般比单个的算法的学习能力要更强。

此外，随机森林算法还可以在训练数据的过程中自动地进行特征重要性评估[36]，可以将其应用在降水预报问题中，计算预报因子的重要性并筛选对降水影响最大的一些预报因子。其基本思路是计算每个预报因子在随机森林的每个决策树上的贡献，再对其求平均得到每个预报因子的贡献，即该预报因子的重要性。

随机森林的构建过程如图 2.4 所示[39]：

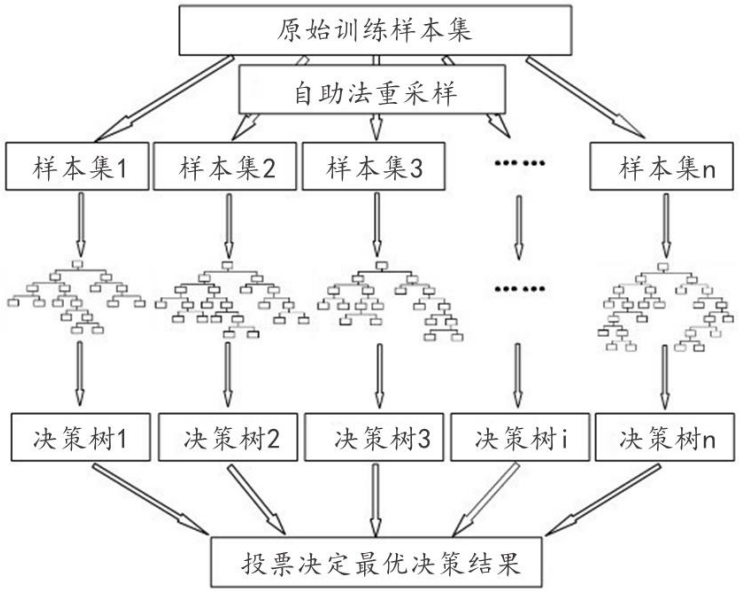

图 2.4　随机森林构建流程图

首先，使用自助法从原始训练样本中进行 n 次重采样，最终生成 n 个样本训练集，再根据每个样本进行训练得到相应的决策树；对于每一个决策树，都各自得到相应的分类或者回归结果。如果随机森林处理的是一个分类问题，那么决策树即为分类树，根据投票就可以决定最优的分类结果；如果面对的是一个回归问题，那么决策树即为回归树，取所有回归结果的平均值，即为随机森林的回归结果。

2.2.2 支持向量回归

1963 年，Vanpik 提出了支持向量机（Support vector machines, SVM）方法。上世纪 90 年代，SVM 方法又经过了不断地完善和发展，能够有效地处理非线性问题，并在很多相关的学科领域里得到了广泛的应用[40]。

以分类为例，支持向量机通常用于处理线性可分的问题。对于线性不可分的问题，支持向量机利用核函数完成一个非线性映射的过程，这一过程使得原本在低维特征空间中线性不可分的数据，能够映射至线性可分的高维空间中，再利用解决一般线性可分问题的方法进行推导和计算，从而能有效地处理非线性问题[40]。

还有一点需要注意的是，支持向量机中的“支持向量”是数据中的小部分特殊样本，它们分布在最优超平面的附近。因此，使用支持向量机算法进行实验时，其计算量的大小取决于支持向量的个数，而不是整个数据样本。支持向量机算法的这一特性使得其更能充分地利用关键样本，因此算法的“鲁棒”性较好[40]。

一般来说，支持向量机算法常用来处理分类问题，但实际上支持向量也能够

应用于回归问题[41]，此时算法称为支持向量回归（Support Vector Regression，SVR）。SVR 和 SVM 算法相似的是都需要求得一个最优超平面，但 SVM 的超平面是为了将数据分成两个类别，因此要求样本点离超平面越远越好；而 SVR 的超平面要求样本点离超平面越近越好。如图 2.5 所示[41]，给定训练样本 $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$，$y_i \in R$，SVR 的最终目标是找到一条红色的实线（超平面），要求这个超平面 f(x)距离 y（样本的标签）越近越好，f(x)中的参数 ω 和 b 需要通过训练样本的学习来确定。

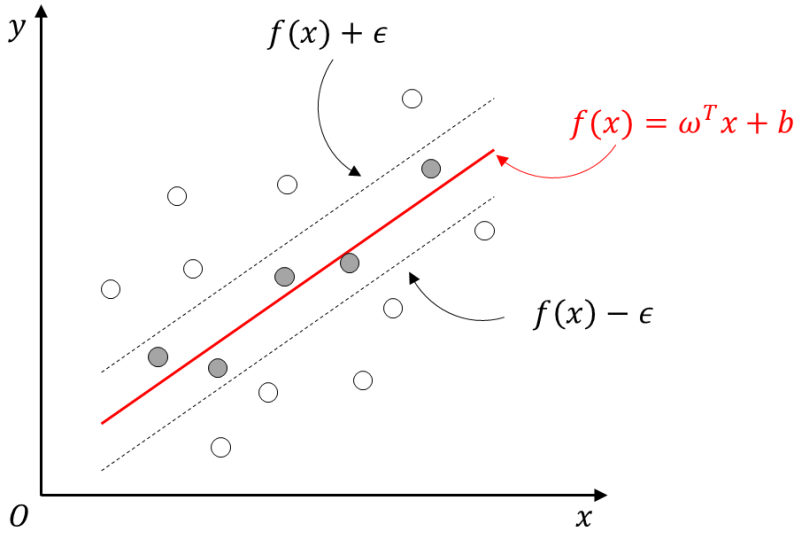

图 2.5　SVR 算法示意图

与一般回归方法不同的是，SVR 允许 f (x)与 y 之间不超过 ϵ 的偏差存在，即当且仅当 $f(x) - y| > \epsilon$ 时，才会将该样本作为损失进行计算，于是可以将 SVR 问题转化为式(2.2)，其中 ω 是最优超平面的法向量。

$$\min_{\omega,b} \frac{1}{2} \| \omega \|^2 + C \sum_{i=1}^{m} l_{\epsilon}(f(x_i), y_i) \tag{2.2}$$

公式(2.2)中 C 是正则化常数，而函数 l_{ϵ} 是 ϵ -不敏感损失函数(ϵ -insensitive loss function)，其表达式如(2.3)所示：

$$l_{\epsilon}(z) = \begin{cases} 0, & if |z| \leq \epsilon; \\ |z| - \epsilon, & otherwise. \end{cases} \tag{2.3}$$

引入松弛变量 ξ_i 和 $\hat{\xi}_i$，可将(2.2)式重写为:

$$\begin{aligned}&\min_{w,b,\xi_i,\hat{\xi}_i}\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}(\xi_i,\ \hat{\xi}_i)\\&s.t.\ \ f(x_i)-y_i\le\epsilon+\ \xi_i,\\&\qquad y_i-f(x_i)\le\epsilon+\ \hat{\xi}_i,\\&\qquad \xi_i\ge 0,\ \hat{\xi}_i\ge 0,\ i=1,2,\cdots m.\end{aligned}\tag{2.4}$$

引入拉格朗日乘子 $\mu_i\ge 0$，$\hat{\mu}_i\ge 0$，$\alpha_i\ge 0$，$\hat{\alpha}_i\ge 0$ 可以得到式(2.4)的拉格朗日函数：

$$\begin{aligned}&L(\omega,b,\alpha,\hat{\alpha},\xi,\hat{\xi}_i,\mu,\hat{\mu})\\&=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}(\xi_i+\hat{\xi}_i)-\sum_{i=1}^{m}\xi_i\mu_i-\sum_{i=1}^{m}\hat{\xi}_i\hat{\mu}_i\\&+\sum_{i=1}^{m}\alpha_i(f(x_i)-y_i-\epsilon-\xi_i)+\sum_{i=1}^{m}\hat{\alpha}_i(y_i-f(x_i)-\epsilon-\hat{\xi}_i)\end{aligned}\tag{2.5}$$

再令 $L(\omega,b,\alpha,\hat{\alpha},\xi,\hat{\xi}_i,\mu,\hat{\mu})$ 对 ω，b，ξ，$\hat{\xi}_i$ 的偏导为零可得：

$$\omega=\sum_{i=1}^{m}(\hat{\alpha}_i+\alpha_i)x_i\tag{2.6}$$

最终可以求得 SVR 的解形如：

$$f(x)=\omega^T x+b=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)x_i^T x+b\tag{2.7}$$

式中，参数 α，$\hat{\alpha}$ 和 b 可以通过约束条件求得。最终结果表明，通过作为支持向量的样本点即能够完全确定待求的最优超平面。最后引入核函数，可得到 SVR 方法最终确定的非线性回归函数如式(2.8)所示。

$$f(x)=\sum_{i=1}^{m}(\hat{\alpha}_i-\alpha_i)k(x,x_i)+b\tag{2.8}$$

2.2.3 长短期记忆网络

长短期记忆网络（Long short-term memory, LSTM）是一种深度学习方法，它是循环神经网络（Recurrent Neural Network , RNN）中极为特殊的一种。长短期记忆网络能够在训练的过程中让误差保持在一个较为恒定的水平，使得循环神经网络可以在误差的允许范围内进行多个时间步长的学习，进而建立起长期的因果联系的通道[42]。此外，LSTM 可以解决由于循环神经网络太复杂时权重更新不稳定所造成的梯度消失问题，因此更适合用来处理长时间序列的问题。

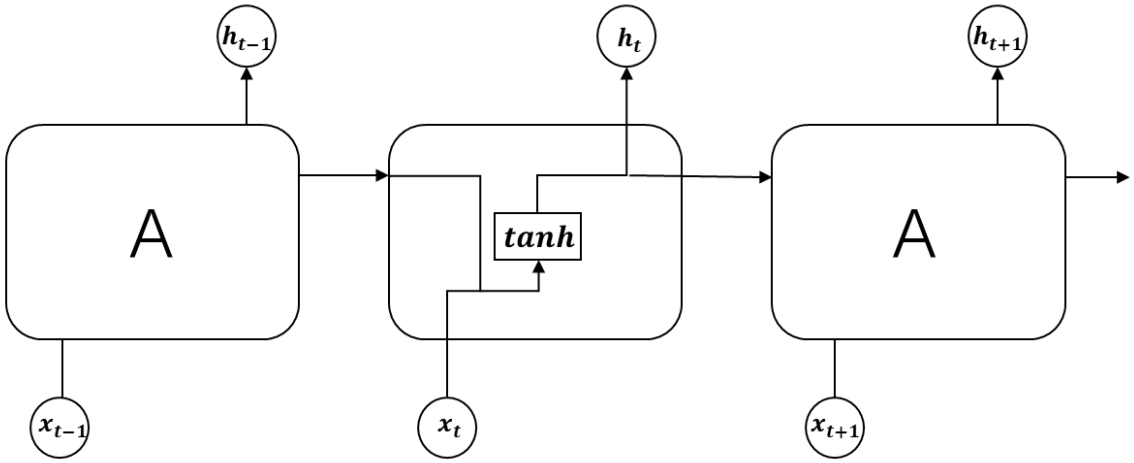

图 2.6　RNN 算法示意图

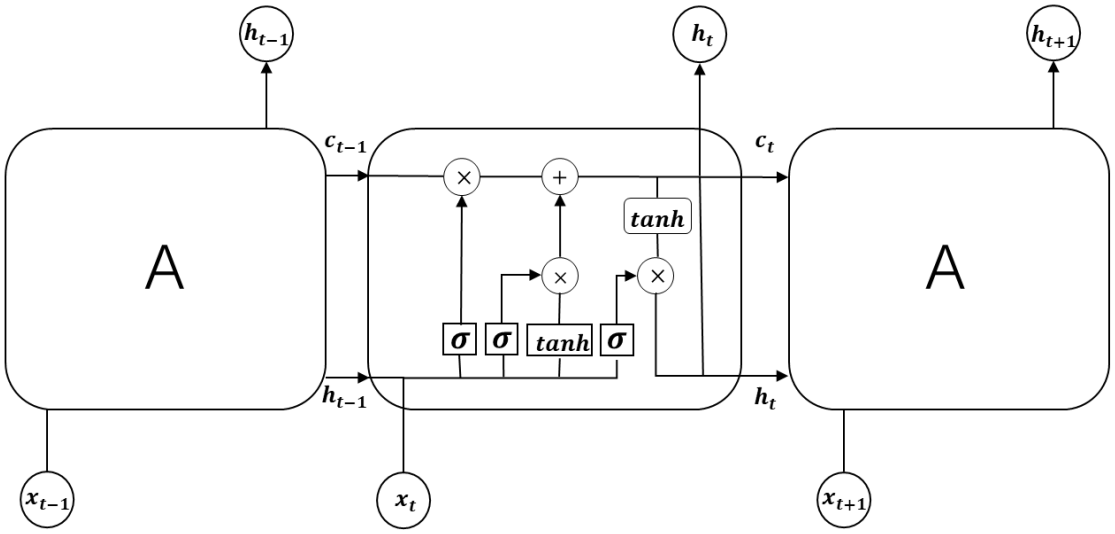

图 2.7　LSTM 算法示意图

LSTM 和 RNN 的区别在于，RNN 在隐藏层的传输中只是一个简单的 tanh 函数（如图 2.6 所示，图中左右两侧的 A 隐含了中间的复杂细节），而 LSTM 在算法中增加了一个独特的信息处理器，该处理器被称为“细胞”（cell）。如图 2.7 所示，一个“细胞”中被设置了三种门结构：输入门、遗忘门和输出门。信息被传入到 LSTM 中之后，这三种门结构就会按照自身的数学规则来进行判断。符合条件的信息将被算法所记忆，而不符合条件的信息将使用遗忘门进行舍弃。通过不同的门结构之间的协同作用，LSTM 可以捕捉到长期依赖。三种门结构的数学形式如下：

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + b_i\right) \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + b_o\right) \end{aligned} \tag{2.9}$$

输入门 i_t 主要用于更新细胞的状态；遗忘门 f_t 决定应记忆或遗忘哪些信息；输出门 o_t 用来确定下一步隐藏状态的值。如公式(2.9)中所示，三种门结构的数学形式相似，都是将上一步的隐藏状态信息 h_{t-1} 以及当前输入信息 x_t 传递至函数 σ（σ 为 sigmoid 函数的简称）中，式中 W_{xi}、W_{hi}、W_{xf}、W_{hf}、W_{xo} 和 W_{ho} 是不同的权值矩阵，b_i、b_f 和 b_o 代表相应的偏置向量。

LSTM 利用这三个门结构以及上一步的隐藏状态信息 h_{t-1}、细胞状态 c_{t-1} 和当前输入的信息 x_t 来确定当前的隐藏状态信息 h_t 和细胞状态 c_t，再输出到下一个 cell 中。具体计算过程如式(2.10)所示。

$$\begin{aligned} c_t &= f_t \odot c_{t-1} + i_t \odot tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right) \\ h_t &= o_t \odot tanh\left(c_t\right) \end{aligned} \tag{2.10}$$

当前的细胞状态 c_t 主要包括两个部分之和，一部分实遗忘门 f_t 乘以上一步的细胞状态 c_{t-1}（$\odot$ 表示向量在元素间的点积），另一部分是输入门 i_t 乘以当前记忆的某些信息（通过 tanh 函数激活）。可以发现，当前的细胞状态 c_t 由遗忘门和输入门共同决定。

当前的隐藏状态 h_t 根据当前的细胞状态 c_t 计算所得，由于 c_t 以线性方式进行更新，因此首先让其通过非线性的 tanh 函数进行处理，之后再使用输出门 o_t 对其进行过滤从而得到当前的隐藏状态 h_t。当前的隐藏状态 h_t 将会成为细胞的输出，并

与当前的细胞状态c_t一同传递给下一个时间步长。

2.2.4 高斯过程回归

高斯过程是这样一个集合，在这个集合中任意有限个随机变量均具有联合高斯分布[43]。一个高斯过程的全部属性都可以用它的均值函数以及协方差函数来表示，即：

$$\begin{aligned} m(x) &= E(x) \\ k(x,\ x^*) &= E\Big[\big(f(x)-m(x)\big)\big(f(x^*)-m(x^*)\big)\Big] \end{aligned} \tag{2.11}$$

其中$x,\ x^* \in R^d$为任意随机变量，可以将高斯过程定义为：

$$f(x) \sim GP\big(m(x),\ k(x,\ x^*)\big) \tag{2.12}$$

通常，为了简化符号，我们取均值函数为0。

假设有训练集$\mathrm{D}=\{(x_i, y_i)\,|\,i=1,2,\cdots,n\}=(X,\ y)$.其中$x \in R^d$代表 d 维输入矢量，$\mathrm{X}=\{x_1, x_2, \cdots, x_n\}$表示 d×n 维输入矩阵，$y_i \in R$表示输出标量，y 表示输出矢量。

假设训练集是有噪声的，那么考虑如下模型：

$$y = f(x) + \varepsilon \tag{2.13}$$

假设$\varepsilon \sim N(0,\ \sigma_n^2)$，可得到 y 的先验分布为：

$$y \sim (0,\ K(X,\ X) + \sigma_n^2 I_n) \tag{2.14}$$

假设y_*是测试点x_*对应的预测值，则可以得到y和y_*的联合高斯分布为：

$$\begin{bmatrix} y \\ y_* \end{bmatrix} \sim N\left(0,\ \begin{bmatrix} K(X,\ X)+\sigma_n^2 I_n & K(X,\ x_*) \\ K(x_*,\ X) & K(x_*,\ x_*) \end{bmatrix}\right) \tag{2.15}$$

如果有n个训练点和n_*个测试点，那么$K(X,\ x_*)$代表度量x和x_*之间相关性的

$n\times n_*$阶协方差矩阵；类似地，对于$K(X,\ X)$，$K(x_*,\ x_*)$，$K(x_*,\ X)$也是这样。其中I_n为 n 维单位矩阵。

通过贝叶斯公式计算y_*的后验分布为：

$$y_*\mid X,y,x_*\ \sim\ N(\overline{y_*},\ cov(y_*)) \tag{2.16}$$

其中，

$$\begin{aligned}\overline{y_*}&=K(x_*,\ X)\left[K(X,\ X)+\sigma_n^2 I_n\right]^{-1}y\\ cov(y_*)&=K(x_*,\ x_*)-K(x_*,\ \mathrm{X})\left[K(X,\ X)+\sigma_n^2 I_n\right]^{-1}K(X,\ x_*)\end{aligned} \tag{2.17}$$

即为x_*对应的预测值y_*的均值和方差。

高斯过程回归的关键是核函数，联合高斯分布的协方差矩阵 K 就是核函数，并且 K 必须是对称半正定的。而核函数矩阵 K 就是对称半正定的。理论上机器学习中所应用的核函数都可以作为这个协方差矩阵。

平方指数核函数（Squared Exponential, SE）是核机器学习中较为常用的核函数，它是无限可微的，即这个协方差函数的高斯过程具有所有阶的均方导数，这就使得它非常平滑，不适用于建模物理过程。平方指数核函数的函数表达式如下：

$$k_{SE}(r)=\exp\left(-\frac{r^2}{2l^2}\right) \tag{2.18}$$

其中 r=x-x'，l 是长尺度参数。

不同于 SE 核函数的平滑特性，Matern 类核函数比较粗糙，它的一般表达式如下：

$$k_{Matern}(r)=\frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}r}{l}\right)^{v}K_v\left(\frac{\sqrt{2v}r}{l}\right) \tag{2.19}$$

其中 l 是长尺度超参数，K_v是一个修正的贝塞尔函数，Γ 表示伽马函数，v=p+1⁄2（p 是非负整数），v 的值越小，则函数越粗糙，理论上更适合处理气象和海洋过程。常用的，v=1⁄2，v=3⁄2，或者 v=5⁄2。

当 v=1⁄2 时，核函数被简化为：

$$k_{\nu=\frac{1}{2}}(r)=\exp\left(-\frac{r}{l}\right) \tag{2.20}$$

有理二次核函数（Rational Quadratic, RQ）不同于之前两类核函数，它本身就可以处理不同长度尺度的特征。对于沿岸和岛屿附近各种突变的 SST 来说，它可以很细微地进行模拟。它的函数表达式如下：

$$k_{RQ}(r)=\left(1+\frac{r^2}{2\alpha l^2}\right)^{-\alpha} \tag{2.21}$$

当超参数 l 和 α 大于 0 时，可以将它看成是具有不同特征长度尺度的 SE 核函数的无限总和（scale mixture），这意味着它本身就是许多个核函数的和。

已经证明，现有核函数通过加法、乘法和比例缩放得到的仍然是核函数[44]。具体规则如下：

$$\begin{aligned} k(x,\ \mathrm{x}^*)&=k_1(x,\ \mathrm{x}^*)\ +\ k_2(x,\ \mathrm{x}^*)\\ k(x,\ \mathrm{x}^*)&=k_1(x,\ \mathrm{x}^*)\ *\ k_2(x,\ \mathrm{x}^*)\\ k(x,\ \mathrm{x}^*)&=\alpha k_1(x,\ \mathrm{x}^*) \end{aligned} \tag{2.22}$$

其中 α 为实数。通过组合核函数得到的新的核函数能够提取多尺度的特征信息，根据组合公式选择合适的核函数进行组合，可以有针对性地提取不同物理过程中的特征信息，而长尺度超参数 l 则隐含了不同特征的影响半径。

2.3 本章小结

本章主要介绍了降水预报产品偏差订正的两类方法，一种是目前常用的以 MOS 法为代表的统计学释用方法，另一种是可以应用于降水预报产品偏差订正的相关机器学习算法。首先介绍了目前普遍采用的统计学释用方法，描述了 PP 方法和 MOS 方法的技术流程，重点阐述基于 MOS 法的降水预报产品偏差订正方法的基本流程，并分析其存在的不足之处和局限性。主要有以下两点：

1. MOS 方法基于线性相关的假设挑选降水预报因子，对于降水这类复杂的非线性问题存在一定局限性，无法找到最合适的预报因子，进而影响降水的预报效果。

2. MOS 方法利用多元线性回归建立降水预报模型，理论上需要预报对象满足

正态分布或近似正态分布的假设，然而降水量是严重偏态分布的变量，使用 MOS 方法建立降水预报模型可能难以得到较好的预报效果。

针对 MOS 方法的这两个不足之处，介绍了可以应用于降水预报产品偏差订正的相关机器学习算法，分别对随机森林、支持向量回归、长短期记忆神经网络以及高斯过程回归方法的理论依据与算法特点进行了描述。随机森林的重要性评估可以改进基于线性相关分析的预报因子筛选，随机森林、支持向量回归和长短期记忆神经网络方法可以建立降水量的预报模型，而高斯过程回归方法可以应用于空间插值进而提升数据的分辨率。

第三章　基于随机森林的预报因子提取

3.1 实验数据及预处理

本文使用银河全球谱模式（Yinhe Global Spectral Model，YHGSM）数据以及地面气象观测站的降水历史观测资料。模式数据为 72 小时预报时效的逐 3 小时预报数据（起报时刻为 00GMT，即北京时间 8:00），观测数据为历史逐小时降水量观测实况数据，使用过去三天的模式预报产品和历史观测资料对数值预报产品进行解释应用。

地面气象观测站为安徽省黄山市的屯溪站，屯溪站的区站号为 58531，该站为基准站，具有一定代表性。屯溪站的经纬度为北纬 29.43°、东经 118.17°，屯溪站的观测场海拔高度为 142.7 米，其气压传感器海拔高度为 143.9 米。如图 3.1 所示，本文使用的模式数据为观测站附近的四个网格点预报数据（仅为示意图，不代表绝对位置），不再将其插值到观测站点上，即站点预报场使用覆盖该站点的网格点的预报场替代，其潜在好处是：（1）避免了站点预报场插值计算的不准确；（2）同时考虑了站点附近气象因素的影响。

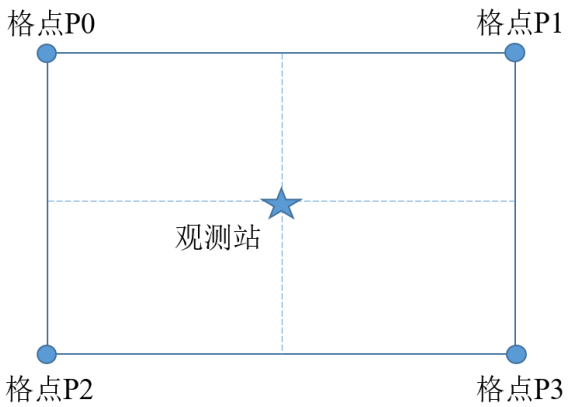

图 3.1　观测站与预报格点相对位置示意图

银河全球谱模式的模式方程组基于浅薄大气与静力平衡近似，垂直方向上采用基于三次样条的有限元离散以及基于气压的地形跟踪混合垂直坐标，水平方向上采用基于三角截断的球谐谱离散以及线性高斯精简网格，时间离散方案使用两时间层半隐半拉格朗日方案[44]。物理过程参数化包括积云对流、湍流扩散、次网格地形拖曳、陆面过程、辐射以及大尺度降水过程等[45]。YHGSM 数值预报产品的要素包括模式预报地面场和气压层上的高空形势场，气压层包括 10-1000hpa。模式预报地面场和气压层上的高空形势场的主要参数如表 3.1 和 3.2 所示。

表 3.1 YHGSM 预报地面场主要参数

变量中文名	Grib 编码	变量英文名
对流有效位能	59	Convect avail potential energy
地面气压	134	Surface pressure
总柱水汽量	136	Total column water
表面潜热通量	147	Surface latent heat flux
平均海平面气压	151	Mean sea level pressure
10 米 u 风	165	10 meter u wind
10 米 v 风	166	10 meter v wind
2 米温度	167	2 meter temperature
2 米露点温度	168	2 meter dew point temperature
地表太阳辐射	176	Surface solar radiation
地面热辐射	177	Surface thermal radiation
总降水量	228	Total precipitation
对流降雪水当量	228220	Convective snowfall rate water equivalent
大尺度降雪水当量	228221	Large scale snowfall rate water equivalent

表 3.2 YHGSM 预报高空形势场主要参数

变量中文名	Grib 编码	变量英文名
位势高度	129	Geopotential height
温度	130	Temperature
U 风	130	U component of wind
V 风	132	V component of wind
比湿	133	Specific humidity
垂直速度	135	Vertical velocity
涡度	138	Vorticity
散度	155	Divergence
相对湿度	157	Relative humidity
比云液态水含量	246	Specific cloud liquid water content
比云冰水含量	247	Specific cloud ice water content
云覆盖率	248	Fraction of cloud cover

实验数据的预处理主要是将模式预报地面场和气压层上的高空形势场的主要

参数按照不同气压层和位置（指站点附近的四个格点）处理为不同的因子存放成 csv 格式，其命名规则采用“因子英文名首字母_气压层_格点位置”格式，例如“r_500_P1”表示“P1 点处 500hpa 的相对湿度”。模式预报数据时间间隔均为 3 小时，因此还需要将 1 小时降水量观测实况进行逐 3 小时的累加处理，最终形成与模式数据起报时刻相同的逐 3 小时数据。

与大多数物理量场相比，数值预报模式输出的降水量场的精度更低，如果直接将其应用于降水预报，则会产生较大的误差，尤其是对于强降水的预报。为了获得更加准确的降水预报结果，需要对数值预报产品进行订正。

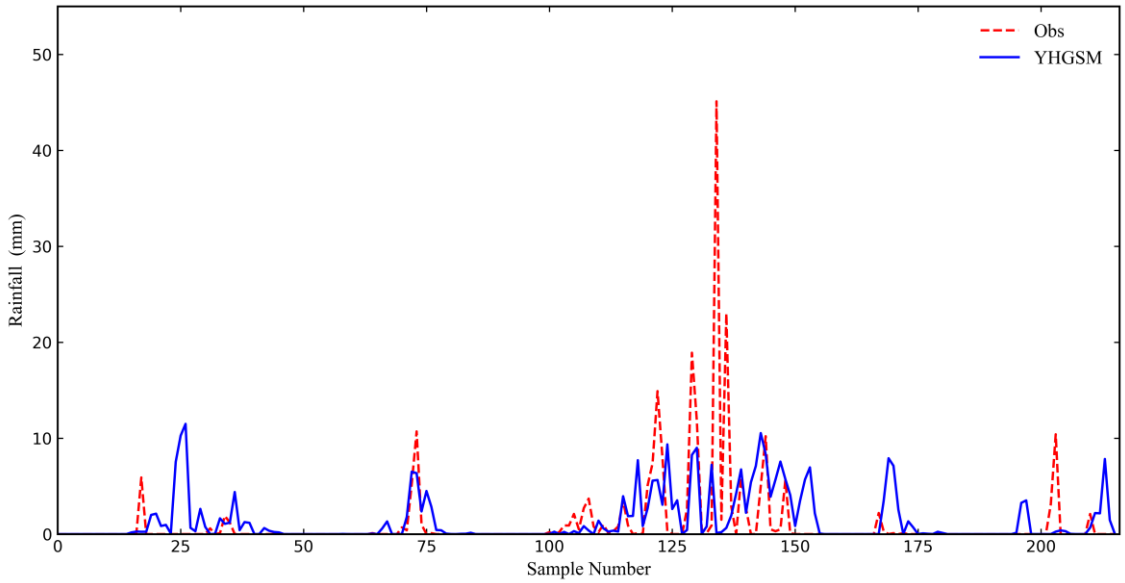

图 3.2　模式预报降水与降水实况对比

如图 3.2 所示为屯溪站 2019 年 6 月份的逐 3 小时降水实况和相应的 YHGSM 逐 3 小时降水预报，红色虚线为降水实况，蓝色实线为 YHGSM 输出的降水量，起报时刻均为北京时间 8:00。6 月份共 30 天，逐 3 小时数据每天有 8 个，由于预报因子需要构造前 3 天的数据，因此共有 27×8=216 个样本。可以直观地看出，虽然 YHGSM 预报的降水可以很好地进行晴雨预报，即能有效地区分有无降水，但是在降水量较大时的预报效果较差，预报降水量偏小，即不能较好地预报强降水。因此，需要对模式输出的降水量进行偏差订正。

3.2 预报因子筛选

无论是 PP 方法还是 MOS 方法，最终都是要建立预报对象和预报因子之间的统计关系。对于某一个确定的气象要素的预报问题，其预报对象是唯一的，于是

预报因子的选取对预报结果的影响极大。张诚忠利用 T106 数值预报产品，在使用相同的数学回归模型的基础上，研究不同的预报因子选取与处理方法对降水预报准确率的影响，并对广西 89 个站点进行了降水预报实验，结果表明预报因子的选取与处理对 MOS 预报方程的预报准确率影响极大[46]。

选取合理的预报因子对降水预报至关重要，一般先尽可能多地将与降水关系密切的影响因子全部挑选出来建立预报因子库。模式预报数据包括各个高度层的相对湿度、温度、水平纬向风、水平经向风、比湿、垂直速度以及位势高度等。

通常在选取最终预报因子前会先计算大量相关因子作为预选，需要注意以下几点：1）选择具有明确物理意义的因子。2）选择预报精度相对预报对象来说较高的因子。3）选取的物理因子要包括影响预报对象的各种物理因素。4）除了单站上的因子，也可以使用附近站点上的影响因子。

构造预报因子通常先根据以上方面预选一些因子，然后计算这些因子以及预报对象之间的相关系数，选取相关系数较大的因子建立因子库，再对其进行 0.05 显著性的 t 检验，无法通过显著性检验的因子将被舍弃，最终入选的预报因子将和降水量实况数据之间建立回归方程。

相关系数能够衡量变量之间的线性相关程度，一般用 r 表示，对任意变量 X 和 Y，它们的相关系数为：

$$r(X,\ Y)=\frac{Cov(X,\ Y)}{\sqrt{Var[X]Var[Y]}} \tag{3.1}$$

式(3.1)中，$Cov(X,\ Y)$ 为 X 与 Y 的协方差，$Var[X]$ 为 X 的方差，$Var[Y]$ 为 Y 的方差。

具体计算时，对气象观测站点同一时刻的预报因子和预报对象按照公式（3.1）计算相关系数。相关系数并不是用来挑选预报因子的唯一依据，因为并不见得每个自变量都对因变量有显著的解释作用。不显著有很多原因造成，可能是自变量本身与被解释变量没有相关关系，也可能是自变量过多，由多重共线性引起。所以还应对每个自变量的系数进行显著性检验，也就是 t 检验。

3.2.1 基于线性相关的预报因子筛选

由于数值预报产品数量巨大，在建立预报方程前，往往先预选一些预报因子。考虑研究区域地形、气候背景等各项影响因素，一般从以下 3 个方面选取物理意义明确的因子：1）水汽因子：产生降水的基础是充足的水汽，描述水汽条件的因子有比湿、相对湿度以及总柱水汽量等。2）热力不稳定因子：包括位势高度、温

度、2 米露点温度等。3）动力因子：用来描述降水产生及发展的动力条件，包括水平纬向风、水平经向风、垂直速度、散度、地面气压、对流有效位能等。以上因子的气压层包括高度层 500、600、700、850、925、1000hpa。

预选一些预报因子之后，需要对预报因子与预报对象进行相关分析，计算各预选因子与降水量之间的相关系数，并进行 t 检验，将能通过 0.05 显著性 t 检验的因子按相关系数大小排序，根据经验，一般选择 10 个预报因子左右[27]。基于线性相关选取的因子及其相关系数排序如表 3.3 所示（其中 P0、P1、P2、P3 分别代指观测站点附近的四个网格点，具体位置如图 3.1 所示）。

表 3.3 基于线性相关筛选的预报因子及其相关系数

预报因子	具体解释	相关系数
cp_P3	P3 点处对流降水	0.220474
cp_P2	P2 点处对流降水	0.209067
tcw_P3	P3 点处总柱水汽量	0.179413
r_600_P3	P3 点处 600hpa 相对湿度	0.172017
q_500_P0	P0 点处 500hpa 比湿	0.166833
mode_rain_P1	P1 点处模式预报降水量	0.161200
d_600_P3	P3 点处 600hpa 散度	0.155833
r_500_P0	P0 点处 500hpa 相对湿度	0.155260
r_925_P1	P1 点处 925hpa 相对湿度	0.148621
q_850_P0	P0 点处 850hpa 比湿	0.138283

3.3 基于随机森林的预报因子重要性评估

MOS 方法中的预报因子以单相关系数反映预报因子对预报对象影响的重要性程度。但是相关系数只能用来描述两个变量之间线性函数关系，并不能真实反映综合使用多个预报因子进行预报的性能。通过随机森林进行特征重要性评估，建立预报因子提取器，可以挑选出对降水量影响最大的一些预报因子。此外，MOS 方法一般只将同一时刻的数值预报产品作为预报因子[6]，本节将在基于随机森林重要性评估选择同一时刻的数值预报产品作为预报因子的基础上，增加了预报时效之前 72 小时的预报因子进行预报，在使用 LSTM 方法进行预报时又加入了降水量历史观测实况资料作为预报因子。

预报因子的筛选在机器学习中可以看作是特征工程的一部分，即从原始数据中寻找机器学习模型的训练数据。特征工程在机器学习中极为重要，因为训练数

据在某种程度上决定了机器学习模型性能的上限，而采用不同的算法只是不断地接近这个上限。类似地，没有合适的预报因子，即使采用最好的算法，也很难得到较好的预报效果。

数据标准化使用离差标准化方法，其标准化的计算过程如公式（3.1）所示，其中 $max(x)$ 是 x 中的最大值。该方法和最大最小值标准化方法类似，但使用该方法标准化处理后的数据将会落在区间[-1,1]中，相对于最大最小值标准化方法来说，数据的原始结构仍然可以保留，因此更适用于降水量这一类偏态分布的数据。

$$x^* = \frac{x}{|max(x)|} \tag{3.2}$$

随机森林的特征重要性评估的基本原理是以不同特征在所有决策树上的贡献的平均值的大小为依据，计算每个特征的变量重要性评分（Variable Importance Measures，VIM），进而评估不同特征的重要性。一般使用袋外数据（OOB）错误率或者基尼指数（Gini index，GI）计算贡献值。

假设数据中有特征 X1，X2，X3，…，Xm，重要性评估的目标是计算每个特征 Xj 的基尼指数评分 $VIM_j^{(Gini)}$，也就是第 j 个特征在每个决策树中节点分裂不纯度的改变量的平均值。

首先，基尼指数的计算公式为：

$$GI_m = \sum_{k=1}^{|K|}\sum_{k^* \neq k} PmkPmk^* = 1 - \sum_{k=1}^{|K|} P^2mk \tag{3.3}$$

其中，K 代表类别数量，Pmk 指类别 k 在节点 m 中的比例。

分枝前后节点 m 的基尼指数变化量，也就是节点 m 上特征 Xj 的重要性，如公式(3.4)所示：

$$VIM_{jm}^{(Gini)} = GI_m - GI_l - GI_r \tag{3.4}$$

公式(3.4)中，GI_l 和 GI_r 即为分枝之后新节点的基尼指数。如果在集合 M 中出现了特征 Xj 在决策树 i 中的节点，那么 Xj 在第 i 个决策树上的重要性的计算公式如下：

$$VIM_{ij}^{(Gini)} = \sum_{m \in M} VIM_{jm}^{(Gini)} \tag{3.5}$$

假设 RF 中共有 n 颗树，那么：

$$VIM_j^{(Gini)} = \sum_{i=1}^{n} VIM_{ij}^{(Gini)} \tag{3.6}$$

最后，将重要性评分进行归一化处理即为最终的重要性评分：

$$VIM_j = \frac{VIM_j}{\sum_{i=1}^{c} VIM_i} \tag{3.7}$$

在计算特征重要性评分的基础上，特征选择的步骤如下：

1）计算每个特征的重要性，并将它们按照降序的顺序进行排序；

2）确定要舍弃的比例，按照特征重要性的标准舍弃一定比例的特征，最终得到一个新的特征集；

3）用新的特征集重复过程（1）（2），不断地舍弃不必要的特征，最后保留 m 个特征（m 为根据经验设置的值，本文中 m 的值为 10）；

根据随机森林的特征重要性评估结果，选取了最为重要的 10 个预报因子，相较于原始数据剔除了一些对降水无关紧要或者影响不大的因子，这使得计算资源也大大减少。选取的因子及其重要性评分如表 3.4 所示。

表 3.4 基于随机森林筛选的预报因子及其重要性评分

预报因子	具体解释	重要性评分
d_500_P2	P2 点处 500hpa 散度	0.054917
d_925_P0	P0 点处 925hpa 散度	0.039240
cape_P1	P1 点处对流有效位能	0.026925
r_1000_P2	P2 点处 1000hpa 相对湿度	0.022184
w_850_P1	P1 点处 850hpa 垂直速度	0.021170
mode_rain_P2	P2 点处模式预报降水量	0.018602
w_700_P0	P0 点处 700hpa 垂直速度	0.017667
r_850_P2	P2 点处 850hpa 相对湿度	0.016847
tcw_P0	P0 点处总柱水汽量	0.016554
q_925_P2	P2 点处 925hpa 比湿	0.014393

基于线性相关分析得到的预报因子包括不同格点和气压层的对流降水、总柱

水汽量、相对湿度、比湿、模式预报降水量以及散度。而通过随机森林重要性评估筛选的高影响预报因子包括不同格点和气压层的散度、对流有效位能、相对湿度、垂直速度、模式预报降水量、总柱水汽量和比湿。其中，散度、比湿、相对湿度、总柱水汽量和模式预报降水量是两种方法共同筛选出的预报因子，从影响降水的物理因素来看，这些因子也的确是与降水量密切相关的物理量。此外，随机森林重要性评估方法还筛选出了对流有效位能和垂直速度这两个预报因子。对流有效位能是评估大气不稳定度的指标，并对强对流天气的预报具有重要作用；而垂直速度可以反映雨滴下落的速度特征，对于降水过程的雨量强度变化有着明显的指示作用。

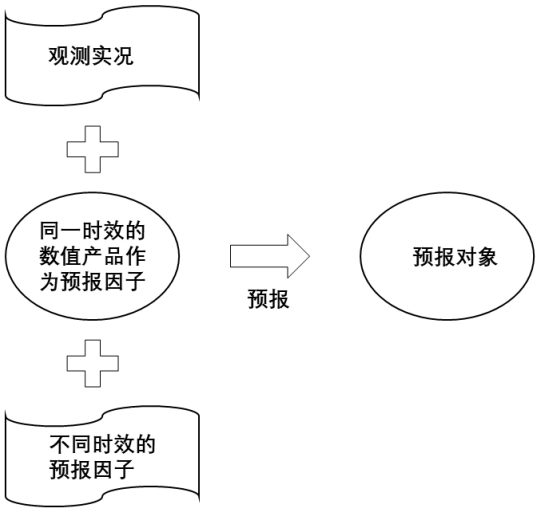

图 3.3　预报因子构造示意图

本文在挑选预报因子时，除了使用随机森林重要性评估改进基于线性相关的预报因子筛选之外，还在预报因子中增加了观测实况资料，同时采用了不同预报时效的预报因子，如图 3.3 所示。在使用随机森林和支持向量回归方法进行预报时，采用了预报时效之前 72 小时的预报因子进行预报。而在使用 LSTM 方法进行预报时又加入了降水量历史观测实况资料作为预报因子，因为 LSTM 神经网络可以从降水量本身的时序结构中寻找因果关系。

3.4 本章小结

本章首先介绍了课题研究所使用的实验数据，观测数据为安徽省黄山市屯溪站的逐小时降水量实况数据，模式预报数据为银河全球谱模式数据的预报地面场和气压层上的高空形势场，模式数据为观测站附近的四个网格点预报数据，其潜

在好处是：1.避免了站点预报场插值计算的不准确；2.同时考虑了站点附近气象因素的影响。

然后分析了模式预报的降水量存在的不足之处，虽然 YHGSM 预报的降水可以很好地进行晴雨预报，即能有效地区分有无降水，但是在降水量较大时的预报效果较差，预报降水量偏小，即不能较好地预报强降水。

本章还指出了预报因子筛选的重要性，对于某一个确定的气象要素的预报问题，其预报对象是唯一的，于是选取预报因子对于预报效果影响极大。接着介绍了传统的基于线性相关的预报因子筛选方法和筛选结果。最后利用随机森林的重要性评估方法对预报因子进行筛选后与传统方法的结果进行对比，并根据降水的影响因素分析对预报因子进行解释。此外，本章还对传统的预报因子筛选从两个方面做出了改进：1.除了模式预报数据之外，还加入了观测数据作为预报因子；2.除了相应预报时效的模式数据之外，还增加了其它预报时效的模式数据作为预报因子。

第四章　基于机器学习的降水预报产品偏差订正

4.1 思路分析

数值预报产品的统计学释用方法主要是根据统计计算来实现，即在数值天气预报产品基础上，结合大量的历史观测实况数据，利用统计学的技术及方法，最终建立气象要素的预报模型，获取更加精确的预报结果。目前业务上用得最多、效果较好的统计学释用方法以 PP 法和 MOS 法为代表，其数学模型大多采用多元线性回归或逐步线性回归方法。这些方法尽管计算过程简便，但是在理论上和实际应用中仍有不足，尤其是对于降水预报这种非线性问题，如果预报因子选取不合适或者预报模型建立不准确时，预报效果较差。

为了进一步提高降水预报产品偏差订正的精度，除了选取合适的预报因子之外，还需要使用合适的建模方法。基于相关分析和线性回归的方法在处理温度等强线性相关性的数据时表现优异，但是在研究非线性问题如降水时的预报效果则较差。而机器学习中有许多适合研究非线性问题的分类或者回归方法，比如 RF、SVR 和 LSTM 方法等，本章将使用这些机器学习方法建立降水预报产品的偏差订正模型。

研究课题的技术路线如图 4.1 所示：

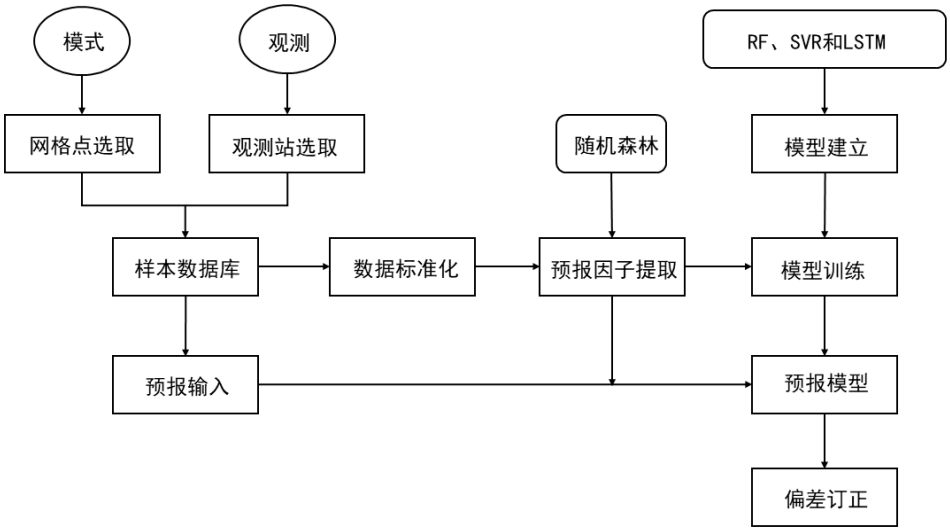

图 4.1　基于机器学习的降水预报产品偏差订正的技术路线

基于机器学习的降水预报产品偏差订正可分为三个部分，首先是处理模式数

据和实况数据构造样本数据集，然后使用随机森林的重要性评估模型筛选预报因子，其次分别使用随机森林、支持向量回归和长短期记忆神经网络搭建和训练降水产品的偏差订正模型，最后输入模式产品进行偏差订正。具体的实施方案如图 4 所示：

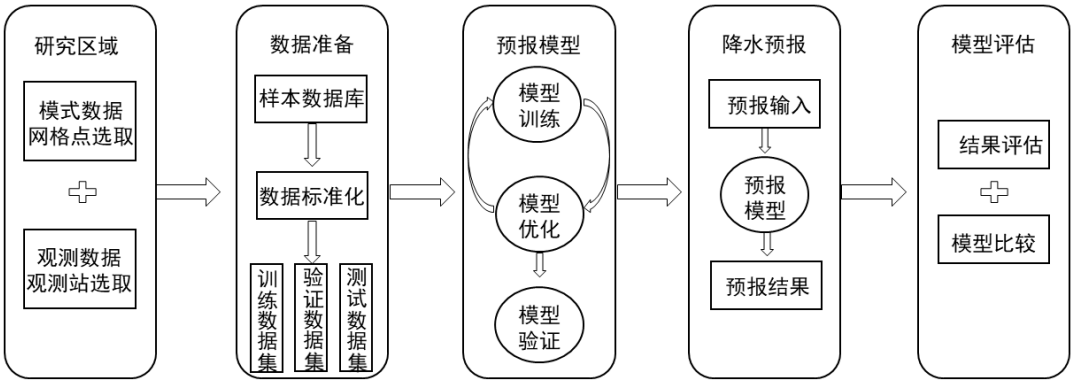

图 4.2　基于机器学习的降水预报产品偏差订正实施方案

基于机器学习的降水预报产品偏差订正实施步骤如下：

1）确定研究区域。将研究区域确定为安徽省黄山市，数据选取包括模式网格点选取以及观测站点选取。通常预报站点不在模式网格点上，因此需要选取预报站点附近的格点进行插值得到预报站点的气象要素值。本章选取的是预报站点附近的 4 个网格点。同时为了避免插值带来的误差，4 个网格点的要素值也作为预报因子进入模型训练。本章观测站点为黄山市的屯溪站。网格点选取和观测站点选取之后生成样本数据库。

2）数据准备及处理。主要包括偏差标准化处理、特征选取以及数据集划分。对于每一个样本，分别存储每个气象因子预报时效前三天对应的数据。模式数据和观测数据的时间间隔均为 3 小时，因此对于单个样本的单个气象因子特征数为 3×24/3=24 个。利用随机森林算法选取重要性较大的一些特征作为预报因子，此部分工作主要在第 3 章中完成。最后将数据集划分训练集、测试集和验证集。

3）生成偏差订正模型。选择随机森林、支持向量回归和 LSTM 方法建立偏差订正模型。模型建立后利用训练数据集、验证数据集和测试数据集对模型进行训练和优化，生成最终的偏差订正模型。

4）利用偏差订正模型对降水预报产品进行偏差订正。输入为第三章中随机森林重要性评估所筛选的预报因子（模型输入内容与训练样本类似，但是剔除了观

测值），然后使用偏差订正模型进行降水偏差订正。

5）模型评估及对比。控制实验为基于多元线性回归的 MOS 方法，同时三种机器学习方法之间互为对比实验，最后利用 TS（threat score）评分、漏报率、空报率、预报准确率这些预报业务上常用的评估方法以及均方根误差（Root Mean Square Error，RMSE）对这些模型进行评估，并分析其优缺点。另外需要注意的是，建立模型时主要考虑站点的本地天气及气候因素，所以针对某一站点的降水预报产品的偏差订正模型一般来说只针对此区域，而模型是否具有区域的通用性可能还需要进一步的实验验证。

4.2 算法介绍

4.2.1 基于随机森林的降水预报偏差订正算法

随机森林是基于决策树的集成学习算法，可以应用于分类问题和回归问题。使用随机森林方法对降水预报产品进行偏差订正属于回归问题，即以逐 3 小时累积降水量为预测对象 Y，所选择的 10 个预报因子即为预测对象的特征向量 X，回归的任务就是要建立预测对象 Y 与特征向量 X 之间的函数关系，这与建立预报方程的思路是一致的。

由于多个决策树的集成组合，随机森林可以较好地处理非线性问题，比大多数单个算法的精度要高。并且，样本随机和特征随机的特点让随机森林不会轻易地在训练过程中出现过拟合。

随机选择样本：随机森林的单个决策树的训练数据需要通过自助采样法进行选择，具体过程如下：假设训练集 D 含有 n 个训练样本，那么重新构建一个新的空集 D_1，每次都从 D 中有放回地随机抽取一个样本 x 放入到 D_1，重复这个过程 n 次，最终得到一个新的含有 n 个样本的训练集 D_1。虽然 D 和 D_1 的样本数量相同，但是 D_1 中的 n 个样本可能存在重复的样本。可以确定的是，D_1 的样本都存在于 D 中，而 D 中的一些样本可能不存在于 D_1 中。

随机提取特征：随机森林的单棵决策树在切分节点时，并不会将所有的特征遍历，而是随机抽取部分特征作为特征子集，再从这个特征子集中寻找最优特征。这种方法可以增大随机森林中每棵决策树之间的区分度，从而整体上提升随机森林模型的泛化能力。

随机森林的预测性能主要取决于两个方面的参数：一是建立的决策树的数量，即随机森林的结果最终由多少个决策树决定，理论上决策树的数量增大可以提升模型的预测性能，但是同样会消耗更多的时间和计算资源；二是随机森林中每个决策树在进行决策时使用的特征（预报因子）的最大数量（max features），通常

有三种选择：特征总数的 20%、特征总数的平方根个或者直接选取所有特征。一般来说，增大 max features 可以提升模型性能，这样每棵决策树就可以利用更多的特征进行回归，但是同样也会降低每棵决策树的多样性以及增加模型的运行时间，因此需要综合考虑各方面来确定数量。通过逐步筛选的方法可以确定随机森林中的最佳决策树数量，首先设置一个数量范围 100-1000，以 50 为间隔步长设置不同的决策树数量从而构建多个随机森林模型进行预测，以均方根误差为模型评估方法，当模型的均方根误差最小时即可大致确定决策树的最佳数量，本章使用逐步筛选方法确定的决策树数量为 600。通过使用类似的方法，确定了 max features 的最佳数量为特征总数的平方根个。

4.2.2 基于 SVR 的降水预报偏差订正算法

支持向量机的功能十分强大，可以处理分类问题和回归问题，处理分类问题时为 SVM 模型，处理回归问题时为 SVR 模型。支持向量机也可以处理线性问题和非线性问题，处理线性问题时和其它的线性学习器一样进行一般地推导处理，而在处理降水预报这类非线性问题时，支持向量机采用引入核函数的方法将其拓展成非线性学习器[47]。

给定训练样本 $D=\{(x_1, y_1), (x_2, y_2),\cdots,(x_n, y_n)\}$，其中 x 是特征输入，y 是逐三小时累积降水量，通过对 SVR 模型的训练求解回归模型中的最优参数从而得到降水量的预报模型。基于随机森林重要性评估得到的 10 个预报因子为特征输入，逐 3 小时累积降水量作为预报标签，使用划分好的训练集对 SVR 模型进行训练。

对于支持向量回归来说，在处理非线性问题时，一般来说使用径向基函数（Radial Basis Function, RBF）作为模型的核函数，径向基函数的数学形式如公式(4.1)所示，其中 r 为核函数的超参数[47]。

$$K(x, x_i)=exp(-r\,||\,x-x_i\,||^2) \tag{4.1}$$

径向基函数中有一个 gamma 参数[47]，gamma 描述了核函数将原始数据映射至高维特征空间之后的具体分布。如果 gamma 参数越大，那么支持向量的个数就会减少，而如果 gamma 参数的值越小，支持向量的个数就会增加。该参数通过影响支持向量的数量进而影响到 SVR 模型的运行速度及时间消耗。

此外，模型中还有一个惩罚项 C 十分关键[47]。增大 C 的值会导致模型在训练集上的学习变得更加严格，随之而来的结果就是模型在训练集上的拟合能力很强但是在新的测试集上的拟合能力降低，也就是会降低模型的泛化能力。类似地，如果惩罚项 C 的值减小，就会在一定程度上提高模型的泛化能力。

通过逐步筛选的方法可以确定 SVR 模型中的 gamma 参数和惩罚项 C 的较优数值，首先设置一个数值范围，以相应间隔步长设置参数构建多个 SVR 模型进行预测，以均方根误差为模型评估方法，当模型的均方根误差最小时即可大致确定参数的最佳数量，本章使用逐步筛选方法确定的 gamma 参数和惩罚项 C 的值分别为 0.01 和 8。

4.2.3 基于 LSTM 的降水预报偏差订正算法

降水预报是一种典型的时间序列预测类问题，即通过某种现象一段时间的状态来判断其未来某一段时间的状态。LSTM 神经网络基于其遗忘门的结构，在处理时序预测类问题效果较好。此外，LSTM 方法作为一种深度学习方法[48]，随着神经网络层的加深以及训练数据的增加，可以很好地拟合任何非线性函数。

与其它机器学习方法不同的是，使用 LSTM 模型进行预测时，第一步需要将原始数据集处理为适用于监督学习问题，从而让 LSTM 模型能够实现通过前几个时刻的预报因子和降水实况去预测下一个时刻的降水量。不同于其它的回归方法只是简单地使用同一时刻的预报因子去预测降水量，LSTM 方法有两个方面在建立模型时做出了改进：（1）将降水作为时间序列类问题进行预测，将前三天的预报因子都作为输入变量，与预报对象建立回归模型，考虑了过去天气系统的演变。（2）与 MOS 方法只使用模式预报数据作为预报因子不同的是，LSTM 方法加入了前三天的降水实况数据作为预报因子，利用 LSTM 模型中门结构的特点，学习降水量自身在时间序列上的因果关系。

LSTM 作为一种深度学习方法，其网络结构的构建十分重要。深度学习实质上是深层的神经网络。模型越复杂，模型的参数越多，那么模型就拥有更强大的预测性能。提高神经网络的复杂度有两种办法，即增加隐藏层的数量和增加隐藏层的神经元的数量。此外，激活函数的选择也十分关键。

本章使用的 LSTM 算法采用双隐藏层结构，每一个隐层都含有 50 个神经元，由于处理的是回归问题，模型的输出层只有 1 个神经元。模型的输入变量为 24 个时间步长（过去 72 小时内的逐 3 小时数据）的特征，损失函数使用均方误差（Mean Absolute Error，MAE)，优化算法为自适应的 Adam 优化器。模型的激活函数选择 relu 函数，对于呈偏态分布的降水来说，它要优于常用的 tanh 和 sigmoid 函数[48]。

过拟合问题是深度学习中不可避免的问题之一，常常导致模型在训练集上的误差很低，而模型对于其它数据集的预测能力则降低，一般来说，过拟合会降低模型的泛化能力。有一些方法可以缓解过拟合问题，最简单的方法是增加数据集的容量，这样可以让模型在数据集上学习得更加全面。除此之外，还可以对模型使用正则化方法，对模型的参数进行正则化约束，进而提升深度学习模型的泛化

能力。正则化包括 L_1 正则化和 L_2 正则化，分别通过对权重的 L_1 范数和 L_2 范数进行约束，从而减小深度学习模型在训练数据集上产生的过拟合问题。本文使用 L_2 正则化对所建立的 LSTM 模型进行约束。

4.3 屯溪站逐 3 小时降水预报实验

4.3.1 实验设置

本章使用的是 2017-2019 年屯溪站的逐 3 小时降水量实况数据以及屯溪站附近 4 个格点相应时间的模式预报数据。其中逐 3 小时降水量实况数据为预报对象，即分类模型和回归模型所预测的标签。模式的格点预报数据为第三章中使用随机森林重要性评估方法所选取的 10 个预报因子，将其前三天的逐 3 小时数据作为模型的特征输入。

实验内容包括分类和回归两类实验，分类实验首先使用随机森林方法建立 0-1 二分类模型对降水进行晴雨分类，具体做法是将逐 3 小时降水量实况数据按 0.1mm 的阈值划分为无雨（标签为 0）和有雨（标签为 1），使用模型训练后进行晴雨分类，分类结果与 YHGSM 的模式预报降水进行对比（模式预报降水按同等标准划分为 0-1 标签）；然后，剔除无雨的数据，在有雨的条件下进行降水等级的分类，以 2.9mm 为阈值划分为小雨（标签为 1）和中到大雨及以上（标签为 2），使用模型训练后进行降水等级分类，并将结果与 YHGSM 的模式预报降水进行对比（模式预报降水按同等标准划分标签）。将 2017-2019 年的数据打乱顺序之后分别选取 80%作为训练集、10%作为验证集、10%作为测试集。

回归实验分别使用随机森林、支持向量回归和 LSTM 方法（以下统一以 RF、SVR 和 LSTM 代称）建立回归模型从而对模式预报的逐 3 小时降水量进行偏差订正。2017-2018 年的数据作为训练集对模型进行训练，2019 年 3-4 月份的数据作为验证集对模型进行验证，最后将 2019 年 6 月份的数据作为测试集对模型进行测试。对照实验为基于多元线性回归的 MOS 方法（以下以 Linear 代称），预报对象与机器学习方法相同，预报因子为第三章中使用线性相关分析所选取的 10 个预报因子，训练集、验证集和测试集的设置均与机器学习方法一致。

实验结果的评估方法分为两类，一种是基于回归的准确性度量的方法，用 RMSE 来衡量；另一种是业务上针对降水预报的评分方法，有 TS 评分、空报率、漏报率和预报准确率，可以对分类结果和回归结果进行评估。RMSE 是常用的回归方法的评价指标，其公式定义如下：

$$RMSE=\sqrt{\frac{1}{n}\sum_{1}^{n}(original_i - predicted_i)^2} \tag{4.2}$$

式中，$original_i$表示第 i 个实况值，$predicted_i$表示第 i 个预测值，n 表示样本数量。RMSE 表示预测值与实况值之间的平均偏差，RMSE 的值越小表示回归的精度越高，在本实验中则意味着预报越准确。

表 4.1 降水预报检验表

实况\预报	有	无
有	NA	NC
无	NB	ND

业务上针对降水预报的评分方法有 TS 评分、空报率和漏报率。对于单站的降水预报而言，如表 4.1 所示，分别定义 NA、NB、NC、ND 为预报正确次数、空报次数、漏报次数以及预报和实况都无降水的次数，则 TS 评分的定义如公式(4.3)所示，其取值范围为 0~1，最佳评分为 1，最差评分为 0。从公式(4.3)中可以看出，TS 评分与 ND 没有关系，因此即使是同样的预报模型，TS 评分对于降水频率较高时的评分较高，而对于降水频率较低时的评分较低。

$$TS = \frac{NA}{NA+NB+NC} \tag{4.3}$$

空报率 NH 的定义如公式(4.4)所示，主要用于描述降水预报中空报的比率，其取值范围为 0~1，最佳评分为 0，最差评分为 1。

$$NH = \frac{NB}{NA+NB} \tag{4.4}$$

漏报率 PO 的定义如公式(4.5)所示，主要用于描述降水预报中漏报的比率，与空报率类似，其取值范围为 0~1，最佳评分也为 0，最差评分为 1。

$$PO = \frac{NC}{NA+NC} \tag{4.5}$$

预报准确率 EH 的定义如公式(4.6)所示，与 TS 评分不同的是，它将降水的准确预报与不发生降水的准确预报都考虑进去，其取值范围为 0~1，最佳评分为 1，

最差评分为 0。

$$EH = \frac{NA+ND}{NA+NB+NC+ND} \tag{4.6}$$

4.3.2 降水晴雨分类实验结果

降水晴雨分类实验使用随机森林分类模型，无雨和有雨的标签分别设置为 0 和 1 作为模型的输出，模型的输入为之前选择的预报因子的前 72 小时内的逐 3h 数据。实验数据的时间范围为 2017-2019 年的逐 3h 的降水实况和模式预报数据。样本总数为 8702 个，其中 6961 个样本（80%）作为训练集，871 个样本（10%）作为验证集，870 个样本（10%）作为测试集。

晴雨分类实验的预报时效为 24h，即使用随机森林分别建立 0-3h、3-6h、6-9h、9-12h、12-15h、15-18h、18-21h 以及 21-24h 共 8 个 3 小时累积降水的晴雨分类模型，将训练集的顺序随机打乱后对模型进行训练，利用验证集优化模型之后，使用对应预报时效的随机森林分类模型对测试集进行测试，并与 YHGSM 的模式预报降水结果进行对比。随机森林模型以及 YHGSM 的 24h 内逐 3h 晴雨分类的空报率、漏报率、TS 评分以及预报准确率分别如图 4.3、图 4.4、图 4.5 和图 4.6 所示。

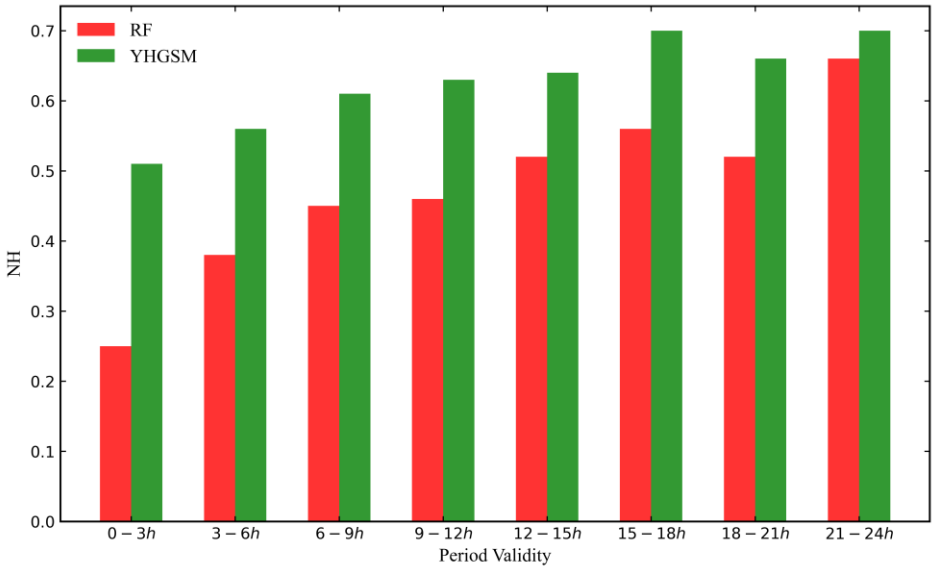

图 4.3　24h 内逐 3h 晴雨分类的空报率对比图

如图 4.3 所示，红色柱形表示随机森林晴雨分类的空报率，绿色柱形为 YHGSM 模式预报降水晴雨分类的空报率，横坐标表示预报时效。0-3h 的晴雨分类的空报

率结果显示，RF（0.25）比 YHGSM（0.51）降低了 26 个百分点，通过随机森林模型的订正之后，有效地减少了模式预报降水的空报现象。随着预报时效的延长，RF 模型的空报率基本呈现逐渐上升的趋势，并且 RF 的空报率始终低于 YHGSM。

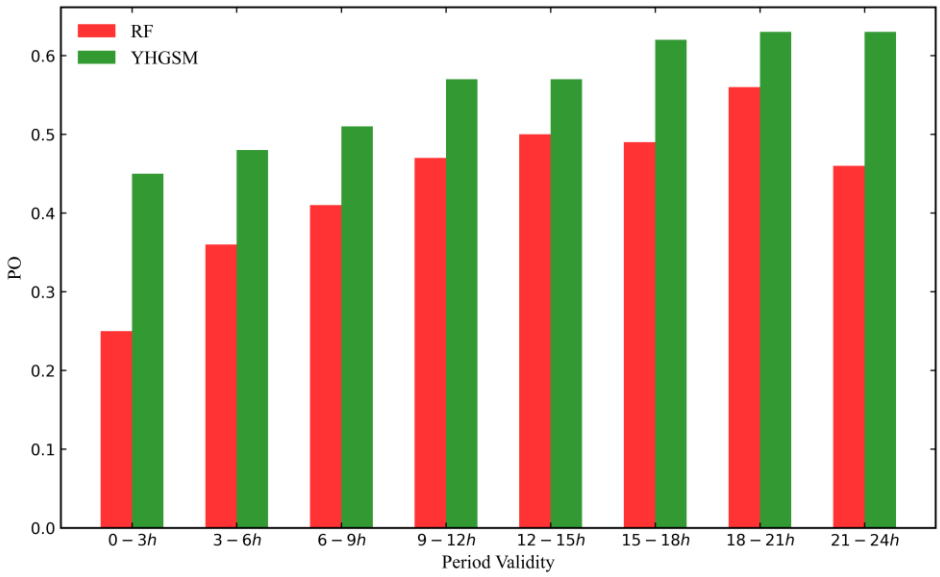

图 4.4　24h 内逐 3h 晴雨分类的漏报率对比图

如图 4.4 所示，红色柱形表示随机森林晴雨分类的漏报率，绿色柱形为 YHGSM 模式预报降水晴雨分类的漏报率，横坐标表示预报时效。0-3h 的晴雨分类的漏报率结果显示，RF（0.25）比 YHGSM（0.45）降低了 20 个百分点，通过随机森林模型的订正之后，有效地减少了模式预报降水的漏报现象。随着预报时效的延长，RF 模型的漏报率基本呈现逐渐上升的趋势，并且 RF 的漏报率始终低于 YHGSM。

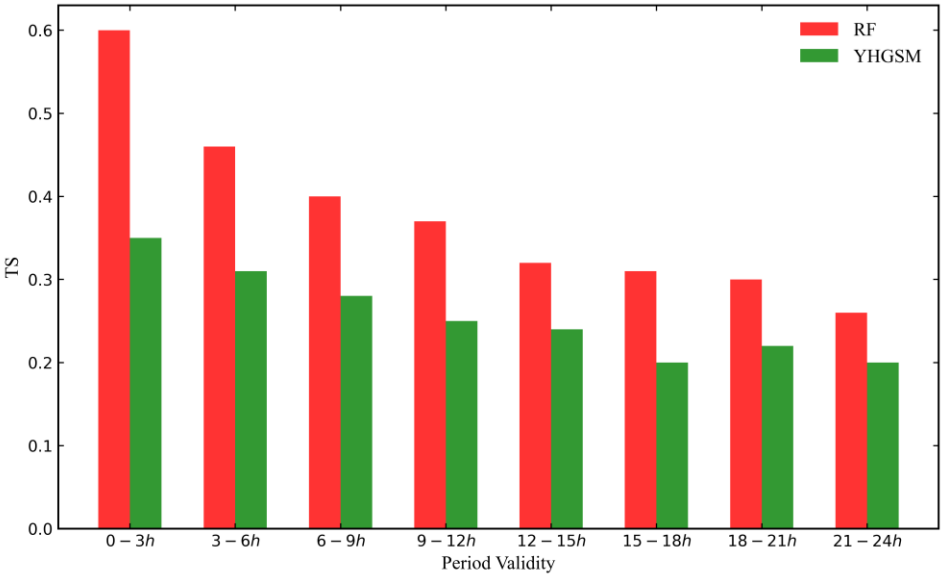

图 4.5　24h 内逐 3h 晴雨分类的 TS 评分对比图

如图 4.5 所示，红色柱形表示随机森林晴雨分类的 TS 评分，绿色柱形为 YHGSM 模式预报降水晴雨分类的 TS 评分，横坐标表示预报时效。0-3h 的晴雨分类的 TS 评分显示，RF（0.60）比 YHGSM（0.35）提高了 25 个百分点，通过随机森林模型的订正之后，有效地提升了模式预报降水的晴雨预报能力。随着预报时效的延长，RF 模型的 TS 评分呈现逐渐下降的趋势，但是 RF 的 TS 评分始终高于 YHGSM。

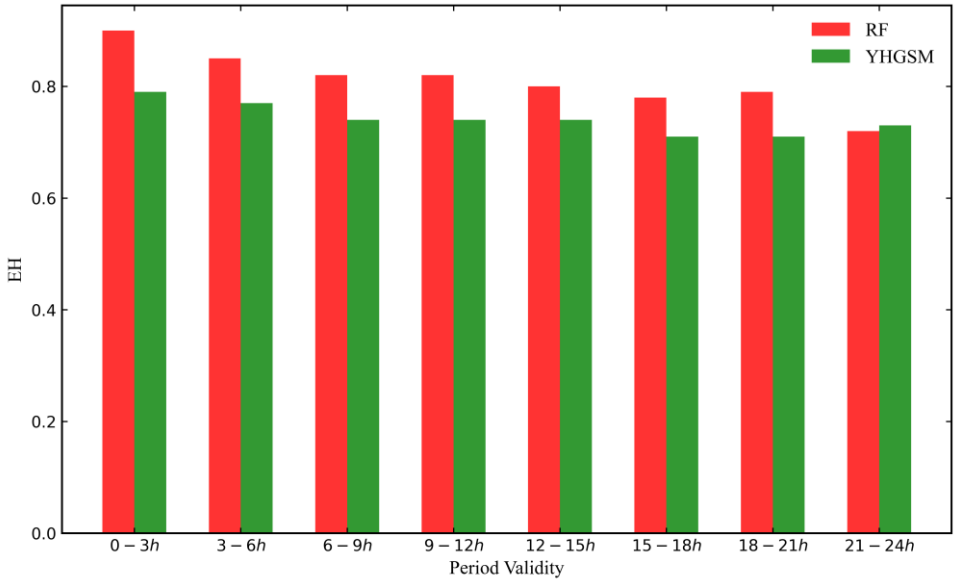

图 4.6　24h 内逐 3h 晴雨分类的预报准确率对比图

如图 4.6 所示，红色柱形表示随机森林晴雨分类的预报准确率，绿色柱形为 YHGSM 模式预报降水晴雨分类的预报准确率，横坐标表示预报时效。预报准确率不仅能衡量模型成功预报降水的能力，也能衡量模型成功预报无降水的能力，0-3h 的晴雨分类的预报准确率达到了 0.90。随着预报时效的延长，RF 模型的预报准确率基本呈现逐渐下降的趋势，RF 的预报准确率在 21h 以内一直高于 YHGSM，21-24h 略低于 YHGSM。

降水晴雨分类实验结果表明，使用随机森林模型对模式预报降水进行偏差订正之后，可以有效提升 24h 以内的晴雨预报的能力，同时也能有效地减少空报率和漏报率。

4.3.3 降水等级分类实验结果

降水等级分类实验使用随机森林分类模型，剔除无降水的数据之后，2017-2019 年还剩下 1425 个有雨的样本。对有雨样本进行降水等级分类，以 2.9mm 为阈值划

分为小雨（标签为 1）和中到大雨及以上（标签为 2）两种类别，模型的输入为之前选择的预报因子的前 72 小时内的逐 3h 数据。其中 1140 个样本（80%）作为训练集，143 个样本（10%）作为验证集，142 个样本（10%）作为测试集。

考虑到有雨的降水样本数量较少，降水等级分类实验的预报时效设置为 9h，即使用随机森林分别建立 0-3h、3-6h 以及 6-9h 共 3 个 3 小时累积降水的等级分类模型，将训练集的顺序随机打乱后对模型进行训练，利用验证集优化模型之后，使用对应预报时效的随机森林分类模型对测试集进行测试，并与 YHGSM 的模式预报降水结果进行对比。表 4.2、表 4.3 和表 4.4 分别表示随机森林模型和 YHGSM 的 0-3h、3-6h 以及 6-9h 降水等级分类的空报率、漏报率、TS 评分以及预报准确率。

表 4.2 0-3h 降水等级分类的空报率、漏报率、TS 评分以及预报准确率

预报方法	空报率	漏报率	TS 评分	预报准确率
YHGSM	0.67	0.74	0.17	0.64
RF	0.52	0.45	0.35	0.68

如表 4.2 所示，0-3h 的预报结果显示，RF 的 TS 评分比 YHGSM 提高了 18 个百分点。另外，预报准确率也有提升，空报率和漏报率有明显下降。这说明使用随机森林对模式预报降水进行偏差订正后能够提升 0-3h 的中到大雨及以上的降水预报的预报性能。

表 4.3 3-6h 降水等级分类的空报率、漏报率、TS 评分以及预报准确率

预报方法	空报率	漏报率	TS 评分	预报准确率
YHGSM	0.76	0.76	0.14	0.65
RF	0.70	0.76	0.16	0.68

如表 4.3 所示，3-6h 的预报结果显示，RF 在各项指标上与 YHGSM 的差异很小，总体上优于 YHGSM，而漏报率已经持平。

表 4.4 6-9h 降水等级分类的空报率、漏报率、TS 评分以及预报准确率

预报方法	空报率	漏报率	TS 评分	预报准确率
YHGSM	0.73	0.81	0.12	0.65
RF	0.73	0.85	0.11	0.63

如表 4.4 所示，6-9h 的预报结果显示，RF 的 TS 评分和预报准确率已经低于 YHGSM。这说明在预报时效超过 6h 之后，RF 已经无法提升 YHGSM 的预报性能。

图 4.7、图 4.8、图 4.9 和图 4.10 分别表示随机森林模型以及 YHGSM 的 9h 内逐 3h 降水等级分类的空报率、漏报率、TS 评分以及预报准确率。如图 4.7 和图 4.8 所示，0-3h 时 RF 的空报率和漏报率仍然低于 YHGSM，但 3-6h 时 RF 的漏报率已经和 YHGSM 持平，而 6-9h 时 RF 的漏报率已经低于 YHGSM。

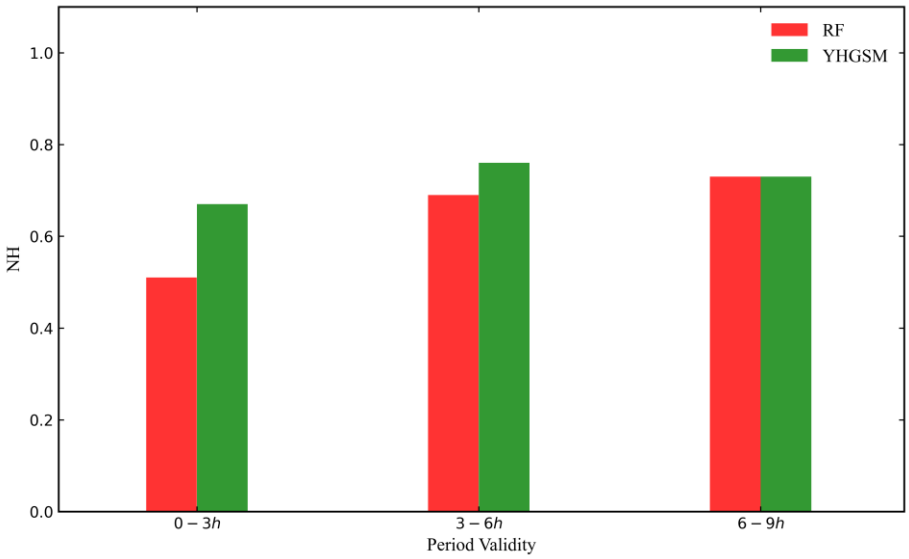

图 4.7　9h 内逐 3h 降水等级分类的空报率对比图

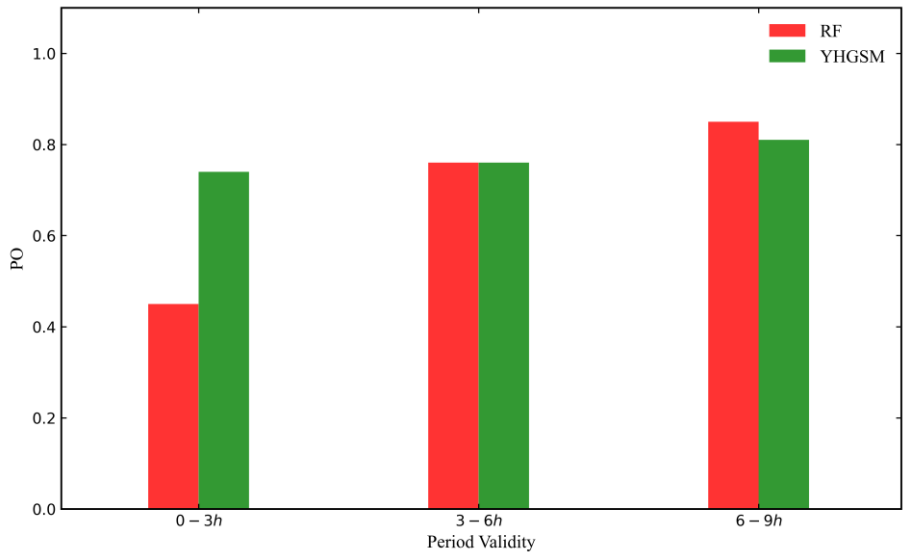

图 4.8　9h 内逐 3h 降水等级分类的漏报率对比图

如图 4.9 和图 4.10 所示，随着预报时效的延长，RF 的 TS 评分和预报准确率呈现下降趋势，并且在预报时效 6 小时以内 RF 的 TS 评分和预报准确率均高于 YHGSM，而 6-9h 时 RF 的 TS 评分和预报准确率均低于 YHGSM。

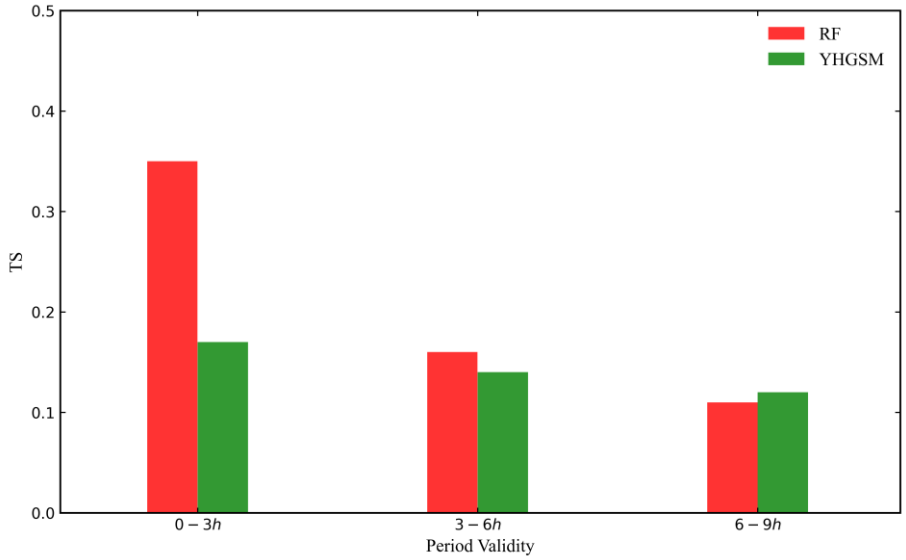

图 4.9　9h 内逐 3h 降水等级分类的 TS 评分对比图

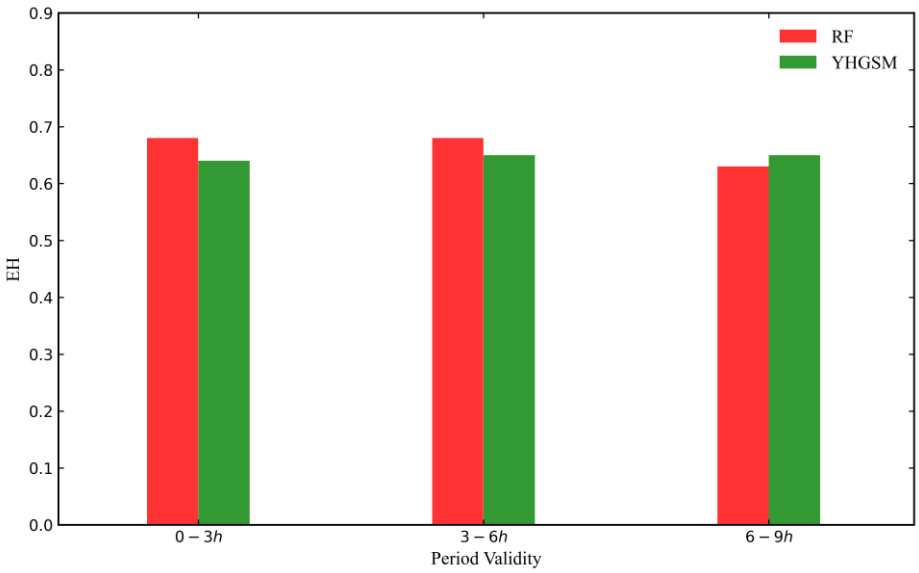

图 4.10　9h 内逐 3h 降水等级分类的预报准确率对比图

降水晴雨分类实验结果表明，使用随机森林模型对模式预报降水进行偏差订正之后，可以提升 6h 以内对中到大雨及以上降水的预报能力，预报时效越短预报效果越好，当预报时效超过 6h 之后随机森林模型对 YHGSM 模式预报降水的偏差订正没有什么效果。

4.3.4 降水量回归实验结果

降水量回归实验需要得到定量的降水量结果，因此其难度更大，本章仅对 0-3h

的 YHGSM 模式预报降水量进行偏差订正。利用验证集优化模型之后，分别使用随机森林、支持向量回归、长短期记忆网络和 MOS 方法对测试集进行测试，测试集为屯溪站 2019 年 6 月份的数据。6 月份共 30 天，逐 3 小时数据每天有 8 个，由于预报因子需要构造前 3 天的数据，因此共有 27×8=216 个样本。因为每年 5-9 月份是黄山地区的汛期，所以选择 6 月份降水较多的时候对模型进行测试可以得到更有效的结论。另外需要说明的是，业务上对于逐 3 小时降水量超过 0.1mm 的情况均视为有降水，而小于 0.1mm 时视为无降水。

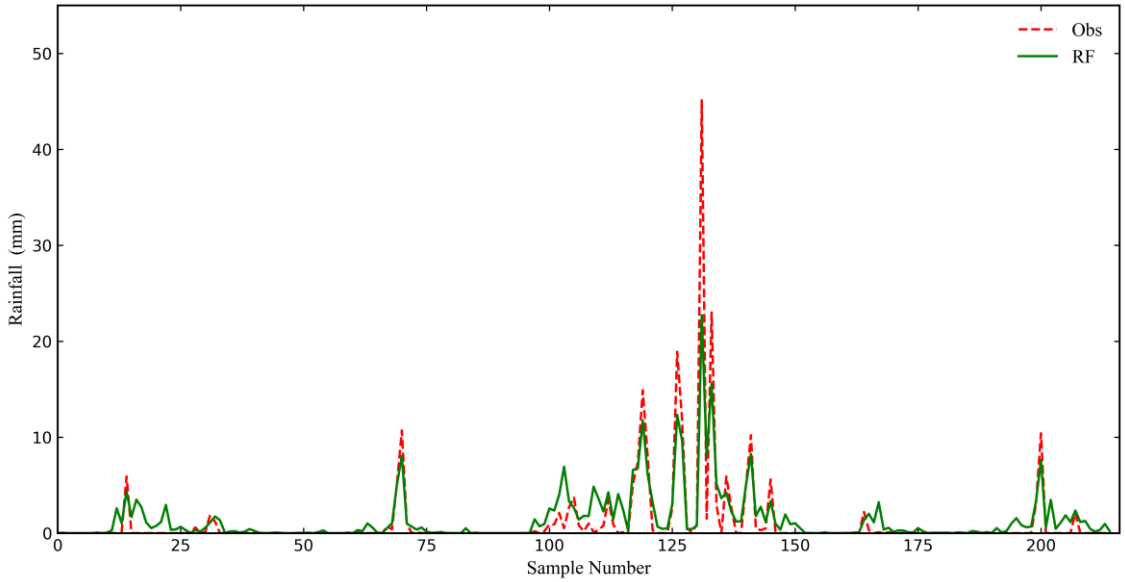

图 4.11　随机森林方法预报的降水量

如图 4.11 所示，红色虚线为 6 月份的降水量实况数据，绿色实线为随机森林方法的预报结果。从图中可以直观地看出，随机森林可以较准确地预报有无降水的情况，但是在面对强降水时难以预报出降水量的大小。

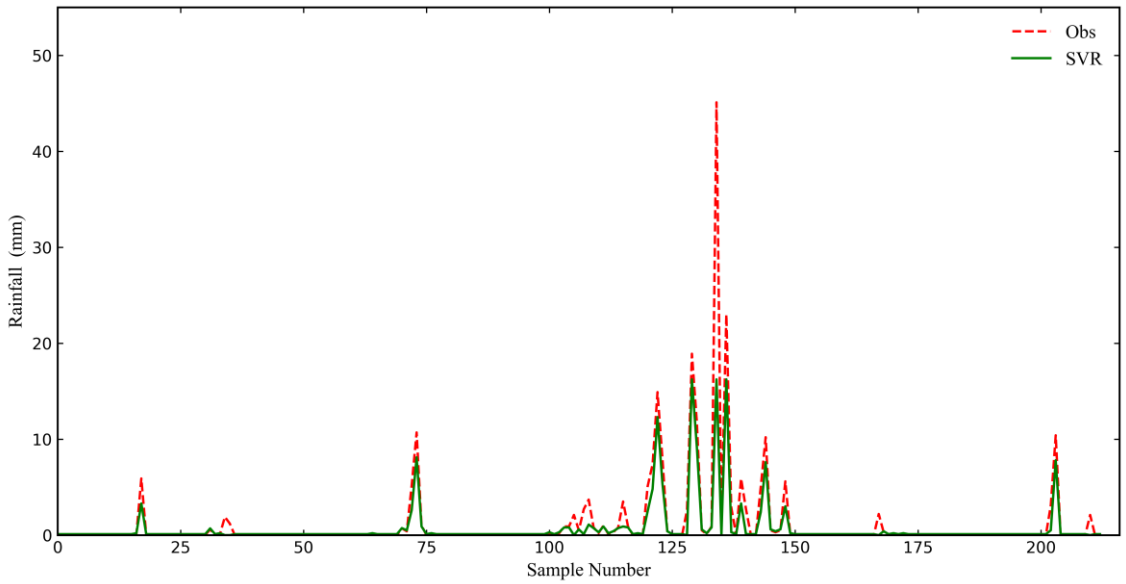

图 4.12　支持向量回归方法预报的降水量

如图 4.12 所示，红色虚线为 6 月份的降水量实况数据，绿色实线为支持向量回归方法的预报降水量。从图中可以看出，支持向量回归方法预报的降水量在曲线上显得相对平滑，预报降水较实况数据总体上来说偏小，但仍然能反映出 6 月份整体降水的轮廓。

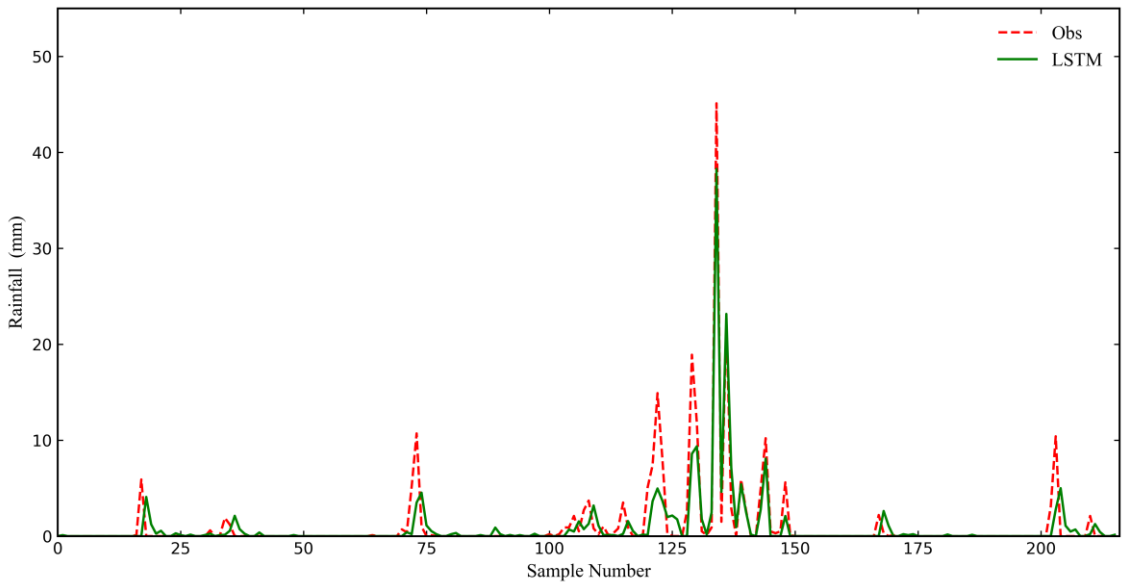

图 4.13　LSTM 方法预报的降水量

如图 4.13 所示，红色虚线为 6 月份的降水量实况数据，绿色实线为长短期记忆神经网络方法的预报降水量。LSTM 方法可以有效地区分有无降水的情况，并且

能够准确地预报出 6 月中旬的强降水的量级（20.0~49.9mm 为暴雨），整体的回归精度较高。

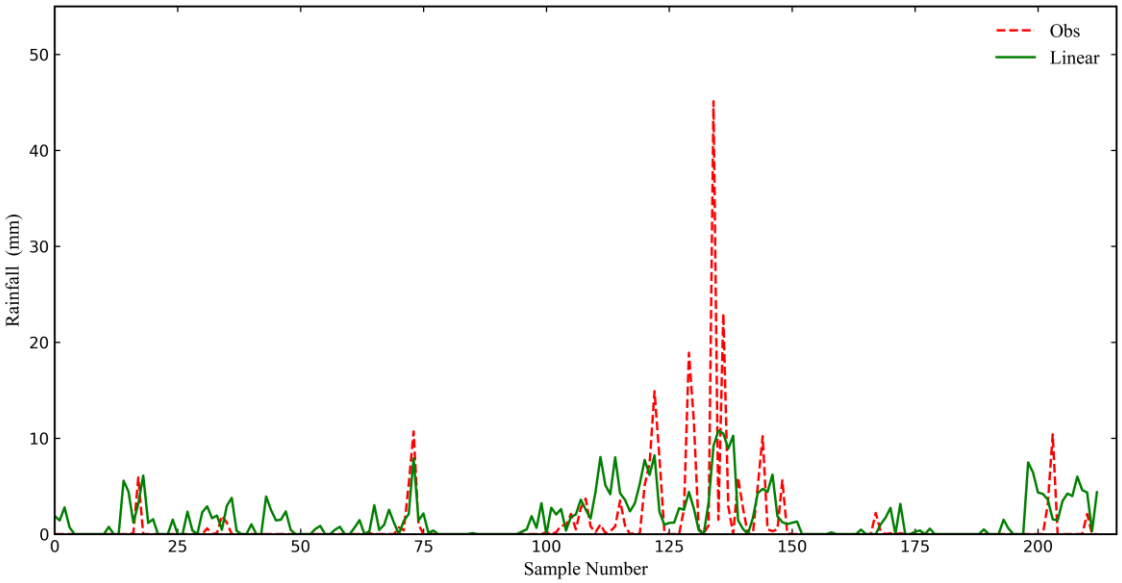

图 4.14　多元线性回归方法预报的降水量

如图 4.14 所示，红色虚线为 6 月份的降水量实况数据，绿色实线为多元线性回归方法的预报降水量。可以看出，由于 6 月份整体降水偏多且降水量级较大（达到大雨、暴雨级），这种情况对于多元线性回归方法来说很难预报出较为准确的降水量。

以上是各种预报方法在图像上的直观结果，并不能充分反映出各种方法的预报性能。下面通过 RMSE、空报率 NH、漏报率 PO、TS 评分和预报准确率 EH 这些评价指标对不同预报方法的预报效率进行对比，具体结果如表 4.5 所示。

表 4.5 不同预报方法之间的 RMSE、空报率、漏报率、TS 评分以及预报准确率

预报方法	RMSE	空报率	漏报率	TS 评分	预报准确率
YHGSM	3.958	0.20	0.53	0.42	0.63
RF	2.107	0.08	0.52	0.47	0.67
SVR	2.214	0.15	0.50	0.46	0.72
LSTM	1.680	0.31	0.39	0.48	0.79
Linear	3.876	0.07	0.56	0.42	0.64

首先可以看出 LSTM 方法的 RMSE 最小，RMSE 为 1.680(mm)，比 YHGSM 的模式输出降水量降低了 57.6%，整体上其它所有方法都比 YHGSM 的 RMSE 要低，从回归的角度来说，针对模式预报的降水量的偏差订正都是有效的，且机器

学习方法比多元线性回归方法的回归精度高很多，这也印证了对于降水这类非线性问题更适合用机器学习方法进行处理。

空报率和漏报率分别描述了针对降水预报的空报和漏报的情况，LSTM 方法的空报率高于其余所有方法，而漏报率低于其它方法，这说明 LSTM 方法预报的降水量普遍较其它方法偏高，导致在没有降水的时候仍然预报了降水，这可能是 LSTM 方法有效拟合了 6 月中旬暴雨的一个副作用。而 RF、SVR 和 Linear 方法的空报率都低于 YHGSM，且 RF 和 SVR 方法的漏报率也要低于 YHGSM 和 Linear 方法。这说明机器学习方法在一定程度上可以降低模式预报降水的空报情况和漏报情况。

最后，LSTM 方法的 TS 评分和预报准确率都是所有预报方法中最高的，另外两种机器学习方法 RF、SVR 的 TS 评分和预报准确率也明显高于 YHGSM，而多元线性回归方法在这两项指标上与 YHGSM 基本持平。综合以上所有结果，说明基于机器学习方法的降水量偏差订正是十分有效的，并且要优于基于多元线性回归方法的降水量偏差订正。

4.4 本章小结

本章首先论述了基于多元线性回归方法的降水预报偏差订正的缺点，引出机器学习方法的优势，并介绍了随机森林、支持向量回归和长短期记忆神经网络方法的算法特点和参数选择。

然后在随机森林方法筛选预报因子的基础上设置了三个实验，首先使用随机森林建立降水晴雨分类模型对 YHGSM 预报降水进行订正，然后使用随机森林建立降水等级分类模型对 YHGSM 预报降水进行订正，最后分别使用随机森林、支持向量回归和长短期记忆神经网络方法建立了预报因子与逐 3 小时降水量之间的回归模型，对 YHGSM 预报的降水量进行偏差订正，并与基于多元线性回归方法的预报结果进行对比，有以下结论：

1.随机森林建立的降水晴雨分类模型对 YHGSM 模式预报降水的偏差订正在 24h 内都是有效的，而降水等级分类模型对 YHGSM 模式预报降水的偏差订正在 6h 内有效，预报时效超过 6h 之后则无效。

2.相对于 YHGSM 和多元线性回归方法，机器学习方法在预报强降水时效果更加明显，尤其是 LSTM 方法。

3. 机器学习方法在一定程度上可以降低模式预报降水的空报情况和漏报情况。基于 TS 评分和预报准确率的评估结果表明，机器学习方法的降水量偏差订正是十分有效的，并且要优于基于多元线性回归方法的降水量偏差订正。

第五章　基于高斯过程回归的数值预报产品插值

5.1 思路分析

目前业务上通常对 T1279 的数值预报产品进行解释应用。受制于模式分辨率的影响，YHGSM 的预报产品的分辨率及其精度仍然难以满足释用的要求，考虑到数值模式和数据同化方法的发展水平以及巨大的计算资源、存储资源，想要快速获取高分辨率的数值预报产品不是一件那么简单的事情，针对数值预报产品进行插值提升其空间分辨率是一个很好的思路。

然而常规的插值方法往往考虑与插值点邻近的区域，再利用数学方法计算得到插值点处的结果，比如最近邻插值、双三次插值以及双线性插值等。这些方法的优点是时间代价较少，但是因为只考虑了局部信息，难免会有许多局限性。尤其是气象、海洋领域的地理数据有较高的空间变异性和复杂的非线性，这导致插值结果往往不够精确[49]。

机器学习方法越来越多地应用于地理数据的插值。Jia 将机器学习方法应用于地震数据的插值且取得了良好的结果[50]。Antonić 很早就开始利用神经网络进行气候数据的插值[51]。神经网络可以准确模拟复杂的非线性函数[52]，正适合用来处理地理数据中的一些复杂非线性关系。Li 使用随机森林、支持向量机等 23 种方法对澳大利亚西南边缘的泥浆含量样本进行插值并作出对比，其中机器学习方法的结果更加准确[53]。

众多机器学习方法中，核方法通过引入核函数使模型能够处理非线性问题[54]。常见的有支持向量机方法以及高斯过程回归方法等。核方法也已应用于气象领域，且取得了较好的表现。Wang 使用加权最小二乘支持向量机算法对风速进行估计，能够准确地预测风速的变化趋势，且具有较高的预测精度[55]。Paniagua-Tineo 使用支持向量回归方法，通过引入温度、降水、相对湿度和气压等预测变量，能够准确预测 24 小时之后的最高温度[56]。

高斯过程回归（Gaussian process regression，GPR）是一种常用的核机器学习方法。与支持向量机和神经网络方法相比，该算法更易实现，并且模型中的超参数可以通过自适应的方法获取，最重要的是，高斯过程回归的输出结果具有一定的概率意义[57]。高斯过程回归已成功应用于图像超分辨领域[58]。而通过插值提高预报产品的空间分辨率其实可以类比于图像领域的超分辨技术，两者都是从低分辨率数据中获取更高分辨率的数据，况且许多数值预报产品也很容易以图像的方式展示。

针对气象要素的空间插值一般来说不需要考虑插值边界处的缺失值，而对于

海洋要素来说，由于岛屿和岸边处的值存在缺失，插值点处的邻近区域也就会存在缺失，常规的插值方法就极容易产生较差的插值结果。而机器学习方法使用回归的思想，综合考虑所有训练样本中的插值点，建立待插值要素及其影响因素之间的回归方程，可以提升插值的综合效果，尤其是对于插值点邻近区域存在缺失的情况，插值精度提升显得更为明显。考虑到这一点，本章针对数值预报产品中的海洋要素设计插值算法并进行实验验证，后续将逐渐应用到降水数据中。

海表面温度（Sea Surface Temperature，SST）是数值预报产品中的一个重要要素，对很多实际的海洋过程分析及海气相互作用分析影响极大[59]。多种因素都对会对 SST 造成影响[60]，比如海表风应力、海表热通量、海洋流速等。由于气象海洋过程的即时性，这些物理因素的影响半径和强度各不相同。单个核函数无法捕捉这些多尺度信息。而通过构造一个组合而成的核函数，引入多变量输入特征，就可以提取这些物理因素不同的影响半径和强度。

考虑到以上这些情况，本章设计了一种基于高斯过程回归的 SST 插值算法。通过构造一个组合的核函数，以经纬度、海表风应力、海表热通量、海洋流速作为其特征输入（由于实验数据的限制，本章仅考虑了这些因素作为特征输入），来提取多尺度的特征信息。

5.2 插值算法

海表面温度是海气相互作用、海洋热力及动力过程等多种因素共同作用下的结果， 由于其影响因素的复杂性和随机性，海表面温度在经纬度上的分布是不均匀的。此外，由于受到海面风及海表热通量的影响，SST 可能在有些地区会产生局部的（异常）现象。另外，沿岸区域及岛屿附近的海域，SST 还可能受到入海径流的影响。当然，SST 也并不是完全没有规律可循，总体来说，SST 的变化有着显著的年周期并受到季节性影响[61]。

考虑到空间位置上相近的 SST 有着相似的分布，首先将地理位置（经度、纬度）作为输入特征。其次，考虑 SST 的物理成因及影响因素，另选 taux（纬向海表风应力）、tauy（经向海表风应力）、net_heating（海表热通量）、u（纬向海流速度）、v（径向海流速度）作为输入特征。

SST 这种粗糙的分布非常适合使用 Matern 类核函数去提取特征[40]，并且函数越粗糙越能提取到更多细节的变化。而有理二次核函数可以看成是无数个具有不同长度尺度特征的 SE 核函数的总和[40]，它可以同时拟合海洋表面大范围的均匀温度分布和局部的温度突变过程，也包括沿岸和岛屿附近的海域。每个特征的长尺度超参数刻画了该特征的影响半径。

考虑到 SST 的各种影响因素，构造了一个组合而成的核函数 k_s 来描述 SST 的分布特点：

$$k_s(\cdot\,,\,\cdot) = k_m(\cdot\,,\,\cdot) + k_r(\cdot\,,\,\cdot) \tag{5.1}$$

其中，$k_m(\cdot\,,\,\cdot)$ 表示 v = 1⁄2 时的 Matern 类核函数，而 $k_r(\cdot\,,\,\cdot)$ 表示有理二次核函数。Matern 类核函数描述了影响 SST 的气象海洋过程的（细节特征）粗糙度，而有理二次核函数则刻画了从近岸到远海的 SST 的各种不同尺度的分布特征。

具体来说，

$$k_m(x,\ x^*) = exp(-r) \tag{5.2}$$

$$k_r(x,\ x^*) = \left(1 + \frac{r^2}{2\alpha}\right)^{-\alpha} \tag{5.3}$$

其中，

$$r = \sqrt{(x - x^*)P^{-1}(x - x^*)^T} \tag{5.4}$$

而 $P = diag(l)^2$。核函数中的超参数 α 和 l 都是未知的，可以先置为 0，再由边际似然的对数推导出来。

算法的训练和预测的过程如图 5.1 所示。

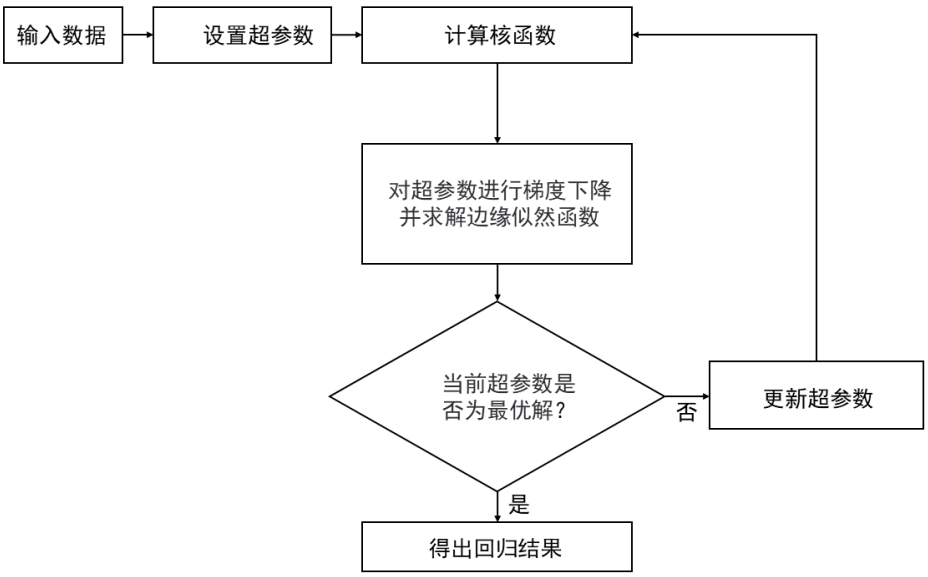

图 5.1　GPR 插值算法流程

5.3 基于 SST 的插值实验

本章使用的是马里兰大学以及德州农工大学共同开发的全球海洋再分析产品 SODA 数据集。SODA 数据集的空间分辨率为 0.5°×0.5°，其经度范围是 0.25°～359.75° E，纬度范围是 74.75° S～89.75°N，并且数据在垂直方向上的各层间距是不等的，共有 50 层（本文选取深度为 5 米的海温作为 SST）。

本章使用了 2014-2015 年的逐月数据（该数据集为月平均数据），研究区域为 0°～66°N， 100°～180°E。训练集的分辨率为 1°×1°，分别从 2014 年 12 个月份的原始数据集中采样而来，剩余数据作为验证集，利用回归方法进行插值时所用的是训练集中的所有网格点。这样选取的原因是为了从单个月份的数据中训练获取超参数，探索海表面温度的月度变化和季节变化对插值效果的影响。通过这 12 组训练集和验证集的实验，选出效果最好的一组超参数。选择 2015 年 5 月份南海地区的数据作为测试集。实验过程如图 5.2 所示。

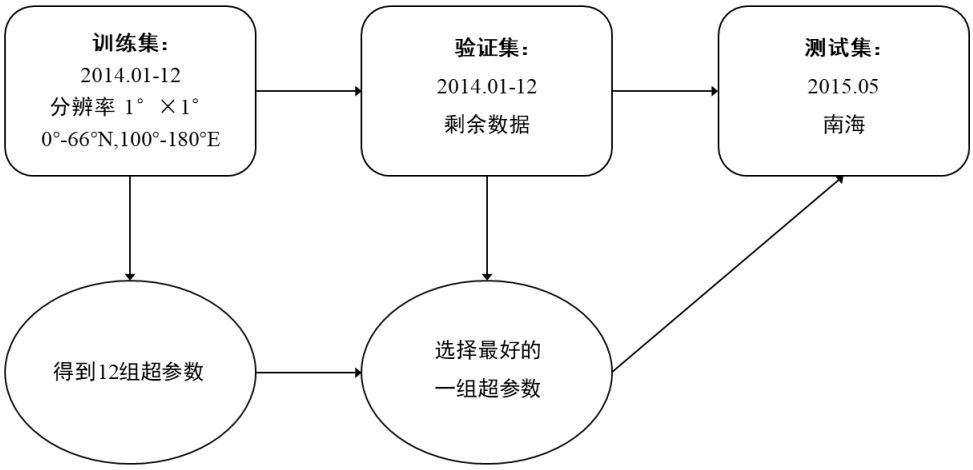

图 5.2　基于 SST 的插值实验过程

另外需要注意的是，由于标签值为海表面温度，需要剔除陆地区域的无效值。输入的特征有经度、纬度、5 米处纬向海流速度、5 米处径向海流速度、纬向海表风应力、经向海表风应力以及海表热通量。对照实验有最近邻插值、双三次插值、双线性插值、支持向量回归和主成分回归(以下简称 Bilinear、Cubic、Nearest、SVR 和 PCR)。

利用 RMSE 来评估插值的精度。此外，SST 的插值结果很容易以图像形式描述。结构相似性（Structural Similarity，SSIM）是衡量两个图像相似度的常用指标，

所以也可以利用 SSIM 来对插值结果进行评估。假设输入的图像分别为 x 和 y，那么它们之间的 SSIM 可以表示为：

$$SSIM(x, y) = [l(x, y)]^{\alpha}[c(x, y)]^{\beta}[s(x, y)]^{\gamma} \tag{5.5}$$

公式(5.5)要求 $\alpha > 0,\ \beta > 0,\ \gamma > 0$，其中 $l(x, y)$ 是亮度比较，$c(x, y)$ 是对比度比较，$s(x, y)$ 是结构比较，其具体定义如下。

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2 + \mu_y^2 + c_1} \tag{5.6}$$

$$c(x, y) = \frac{2\sigma_{xy} + c_2}{\sigma_x^2 + \sigma_y^2 + c_2} \tag{5.7}$$

$$s(x, y) = \frac{\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \tag{5.8}$$

其中，μ_x 和 μ_y 分别表示 x 和 y 的平均值，σ_x 和 σ_y 表示 x 和 y 的标准差，σ_{xy} 表示 x 与 y 之间的协方差。通常 c_1=6.5025，c_2=58.5225，$c_3 = c_2/2$，$\alpha = \beta = \gamma = 1$，均为常数项[62]。

SSIM 的值域为 0~1。特别地，当两张图像完全相同时，SSIM 的值为 1。

5.3.1 单个测试集的情况

通过验证集的实验发现，由 2014 年 9 月份的数据训练产生的超参数为最优模型，利用该模型对测试集进行预测。去除测试集（南海地区，以下简称“区域一”）中的陆地区域，有效插值点数有 1397 个。

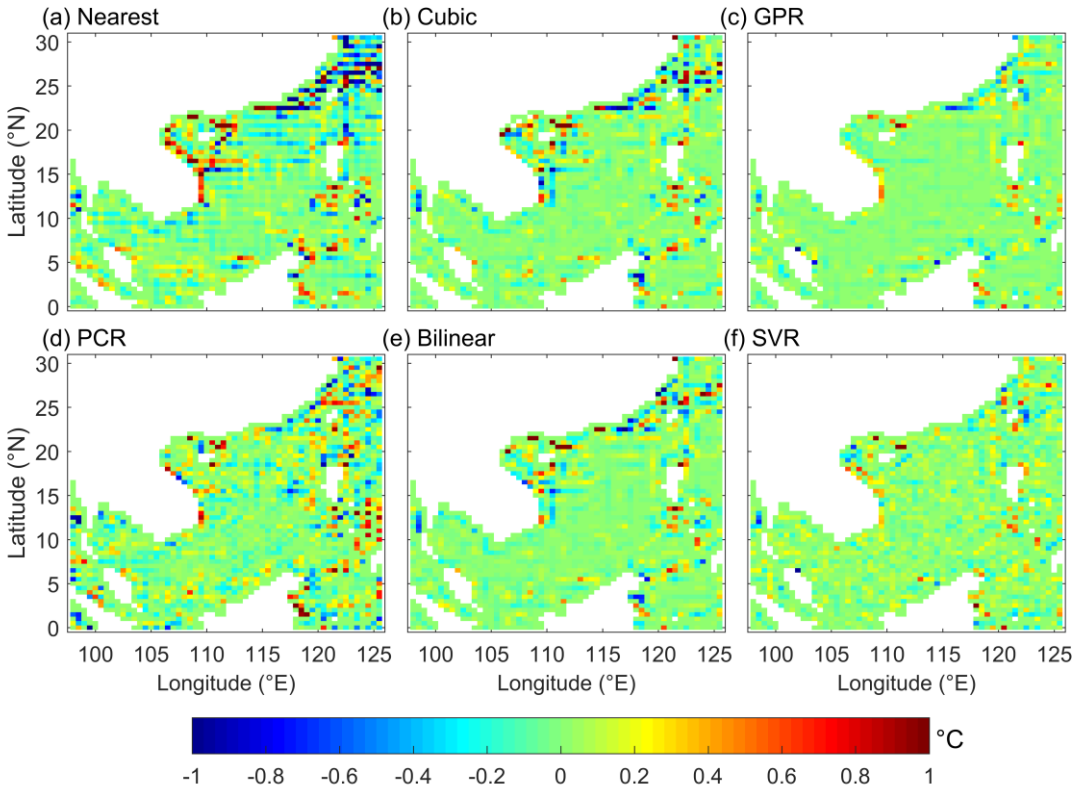

图 5.3　区域一的 SST 插值结果与原始的 SST 之差

如图 5.3 所示，(a)、(b)、(c)、(d)、(e)和(f)分别是最近邻插值、双三次插值、GPR 插值、PCR 插值、双线性插值和 SVR 插值法的插值结果和原始 SST 之间的差值，绿色区域近似为 0，表明插值结果与原始值非常接近。可以直观地看出，与 Cubic 和 Nearest 插值法对比，使用 GPR 进行插值的大部分区域插值效果都较好，尤其是陆地边界处插值效果有明显改善。表 5.1 中 RMSE 的对比结果显示，GPR 插值得到的 RMSE 比 Nearest 插值降低了 62.4%，比 Cubic 降低了 43.7%，比 Bilinear 降低了 38.9%。

此外，假如插值的结果与真值完全一致，那么可以画出一张差值全部为 0（即全部为淡绿色）的图.若以此图作为参照，分别计算以上三种方法产生的差值图与参照图之间的 SSIM 值，可以衡量三种插值方法所得结果绘成图像之后与原始图像的差异，如表 5.1 中所示，发现依然是 GPR 插值法的值最接近 1，这说明从图像的角度来衡量超分辨的精度，依然是 GPR 更胜一筹。

表 5.1 不同插值方法在区域一的 RMSE 和 SSIM

	Nearest	Cubic	GPR	PCR	Bilinear	SVR
RMSE	0.3908	0.2607	0.1468	0.2892	0.2403	0.1717
SSIM	0.8926	0.9300	0.9587	0.8939	0.9354	0.9393

5.3.2 插值算法在时间上和空间上的泛化能力

泛化能力是指通过机器学习的方法学得的模型在新的数据集上的预测能力。特别地，对于 SST 这种在时空上连续的地理数据的插值算法而言，需要分别讨论一下该算法在时间和空间上的泛化能力。

由于模型的超参数是由单个月份的数据训练生成的，所以很有必要讨论一下插值算法在时间上的泛化能力。由图 5.4 可知，Nearest 插值法的结果在 12 个月份中都是最差的，与另外五种方法相比，GPR 插值法在 3-10 月份的插值效果明显要好，其中 9 月和 10 月插值效果最好，原因可能是因为验证集所选最佳模型的超参数是 2014 年 9 月份，而 SST 在年际变化中是有规可循的。当然，总体来说 GPR 在各个月份上的插值效果均要优于另外五种方法，说明算法在时间上的泛化能力是可靠的。

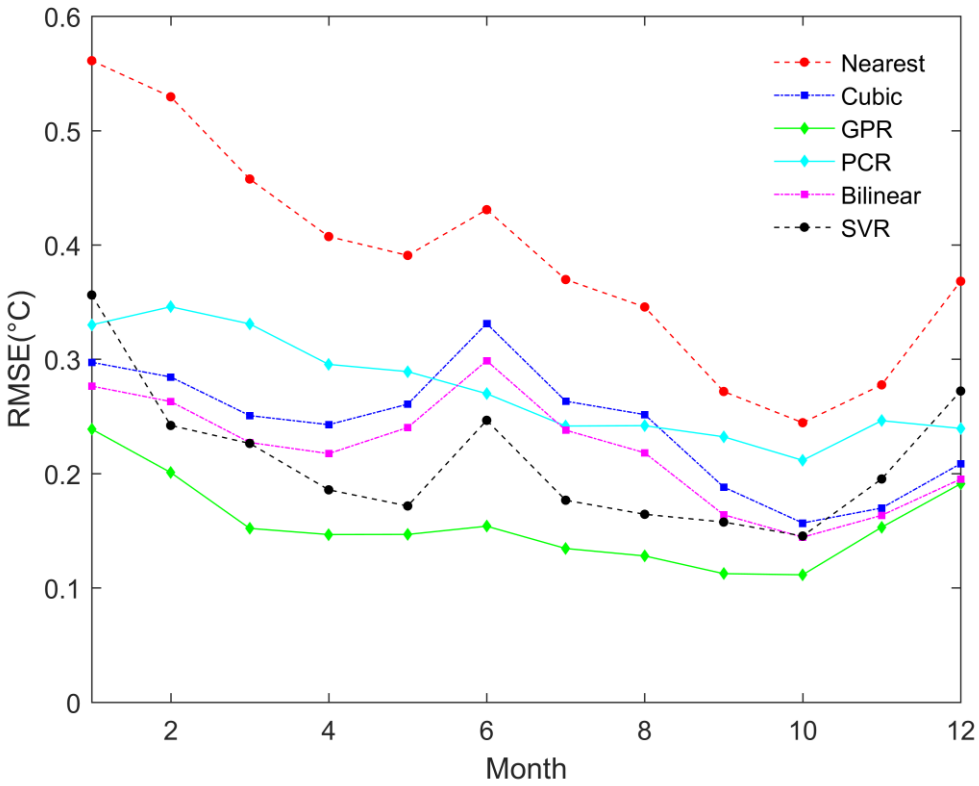

图 5.4　不同插值方法对于不同月份的 SST 插值结果

以上实验均选择了在区域一进行，下面讨论插值算法在空间上的泛化能力。测试时间统一为 2015 年 5 月份，另选两个区域进行插值结果的对比。测试区域分别为 0°～30°N， 125°～150°E 和 30°～65°N， 115°～150°E（下面统称为区域二

和区域三）。区域二为远离陆地的海洋区域，区域三为陆地周围海区以及岛屿附近海区。

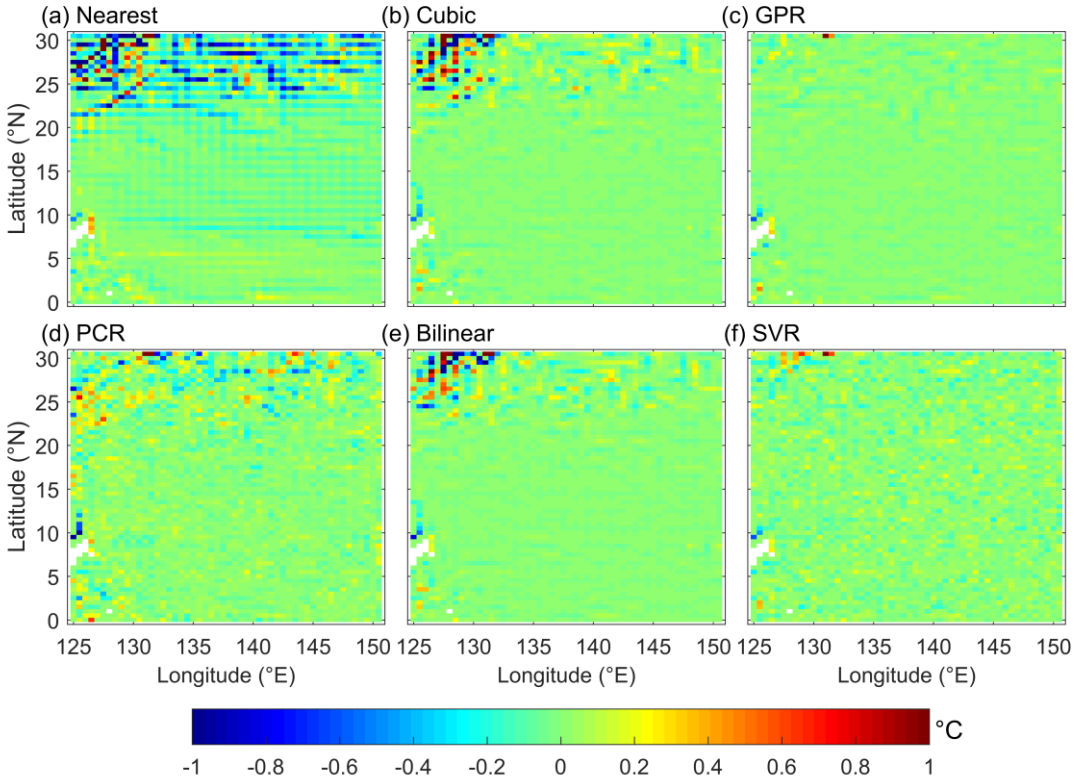

图 5.5　区域二的 SST 插值结果与原始的 SST 之差

如图 5.5 所示，(a)、(b)、(c)、(d)、(e)和(f)分别是最近邻插值、双三次插值、GPR 插值、PCR 插值、双线性插值和 SVR 插值法在区域二上的插值结果和原始 SST 之间的差值。区域二的有效插值点有 2444 个。

表 5.2 不同插值方法在区域二的 RMSE 和 SSIM

	Nearest	Cubic	GPR	PCR	Bilinear	SVR
RMSE	0.3364	0.2524	0.0606	0.1492	0.2058	0.0896
SSIM	0.8699	0.9554	0.9844	0.9309	0.9609	0.9522

综合图 5.5 和表 5.2 的结果，可以发现，对于岛屿附近和沿岸的区域，GPR 的插值效果很好，RMSE 比 Bilinear 插值降低近 70.6%，比 Cubic 和 Nearest 插值分别降低 76.0%和 81.9%。从图 5.5 中还可以看出，在区域 2 的西北部分，GPR 方法明显优于传统方法。这是因为该区域位于岛屿的南部，在岛屿附近使用 GPR 插值效果更好。同样，机器学习方法 SVR 和 PCR 插值方法在这方面也比传统方法有更

好的插值效果。

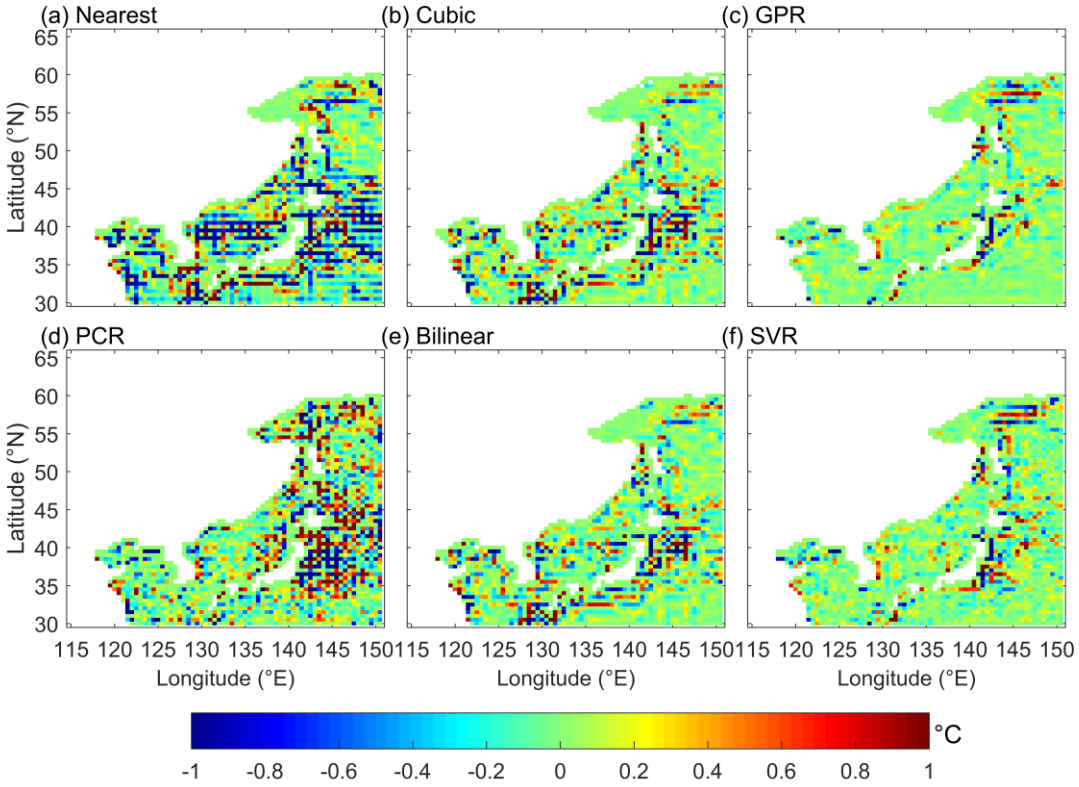

图 5.6　区域三的 SST 插值结果与原始的 SST 之差

如图 5.6 所示，(a)、(b)、(c)、(d)、(e)和(f)分别是最近邻插值、双三次插值、GPR 插值、PCR 插值、双线性插值和 SVR 插值法在区域三上的插值结果和原始 SST 之间的差值。区域三的有效插值点有 1496 个。

表 5.3 不同插值方法在区域三的 RMSE 和 SSIM

	Nearest	Cubic	GPR	PCR	Bilinear	SVR
RMSE	0.9329	0.6050	0.3484	0.8326	0.5311	0.3888
SSIM	0.8560	0.8989	0.9375	0.8656	0.8991	0.9172

综合图 5.6 和表 5.3 的结果可以看出，对于岛屿附近的海域，GPR 的插值效果仍然较好，RMSE 比 Bilinear 插值、Cubic 插值和 Nearest 插值分别降低了 34.4%、42.4%和 62.7%。同时也能看出，除了两种核机器学习方法，其它方法在岛屿附近的插值效果都较差。

综上所述，可以得出以下结论：

1.无论使用什么插值方法，陆地和岛屿附近的 SST 插值效果都不如远离陆地和岛屿区域的。

2.GPR 的插值结果整体上优于传统的插值方法，无论是在远离陆地和岛屿的海洋区域，还是在陆地和岛屿边界处，GPR 的插值结果都优于其它方法。

3.GPR 插值算法在时间上和空间上的泛化能力都是可靠的。

5.3.3 季节变化对算法的影响

由于验证集所挑选的最好的一组超参数及其对应核函数是由 2014 年 9 月份的数据训练生成的，而在讨论插值算法在时间上的泛化能力时，发现测试集为 2015 年 9 月份和 10 月份时插值效果最好。考虑到 SST 大约有 12 个月左右的显著变化周期，尝试探索由相近月份训练生成核函数。实验设置如下，分别选择 2014 年 2 月、5 月、8 月、11 月训练生成的 GPR 算法，去测试 2015 年 1-3 月、4-6 月、7-9 月、10-12 月的数据，测试区域仍然选择南海地区。测试结果与原插值算法、Bilinear 插值和 Cubic 插值结果进行对比。

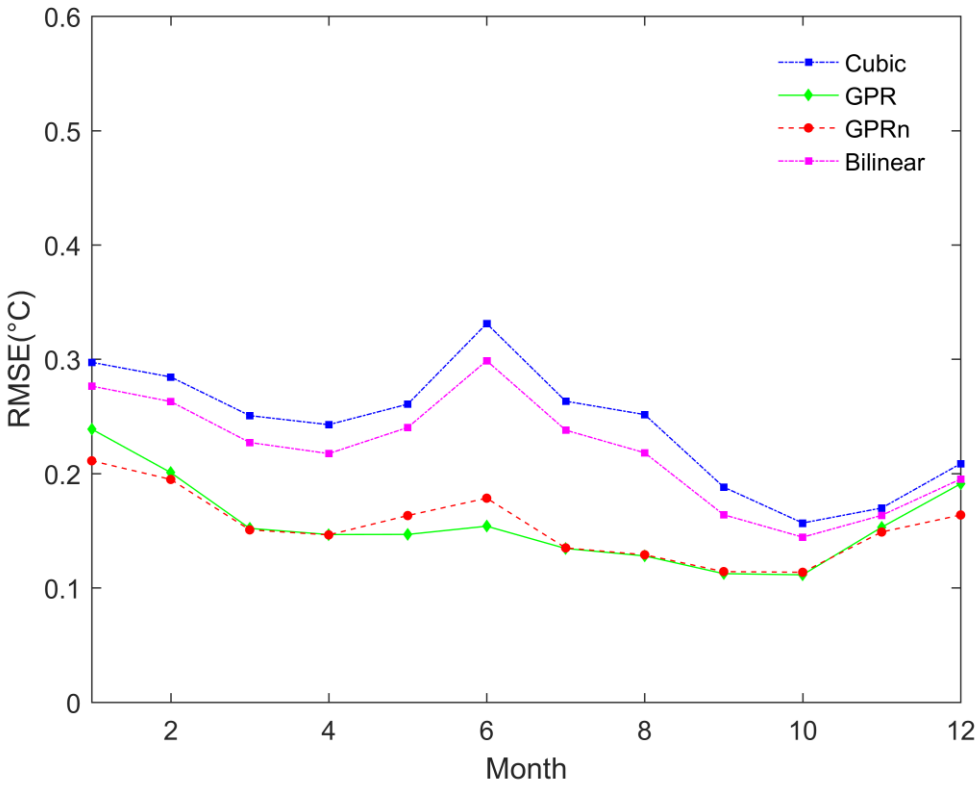

图 5.7　相近月份 SST 训练生成的 GPR 插值结果

如图 5.7 所示，分别利用原 GPR 算法和相近月份生成核函数的 GPR 插值方法对区域一的 2015 年 12 个月份的 SST 进行插值，GPRn 表示相近月份生成核函数的 GPR 插值方法。从图中过得结果可以看出，新算法在 3-10 月的表现不如原插值算法，而另外四个月的插值结果有些许改进。考虑选择原算法插值 3-10 月的结果，其余四个月选择新算法进行插值。总得来说，这两种插值算法都要优于双线性插值和双三次插值的结果。

5.3.4 算法运行时间的讨论

插值实验所用的软件是 Matlab R2016a。由于算法每次运行时间不同，采用五次运行时间的平均值代替。图 5.8 所示是不同插值方法在区域一上对 2015 年的 12 个月份的 SST 进行插值的时间消耗。双线性插值、双三次插值和最近邻插值法消耗的时间几乎相同，当有效插值点为 1397 个时，时间约为 20-21 秒，而 GPR、PCR 和 SVR 插值法用时约为 22-23 秒。因此，GPR 插值方法的额外时间成本约为 2 秒 (10%)。考虑到其对于传统方法改进的插值精度而言，这一代价是完全可以接受的。

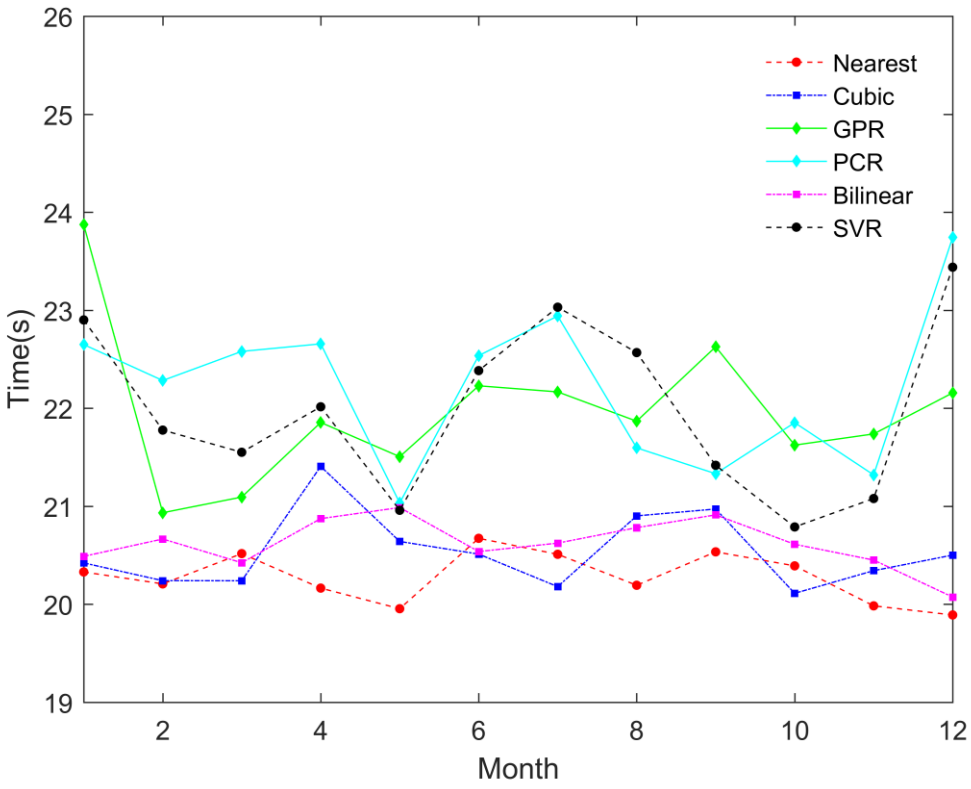

图 5.8　区域一上不同插值方法的时间消耗

5.4 本章小结

本章设计了一种基于高斯过程回归的 SST 插值算法，通过构造一个组合的核函数，以影响 SST 的经纬度、海表风应力、海表热通量、海洋流速作为特征输入，建立插值模型，进而提高数值预报产品的空间分辨率。并设计了插值实验，与最近邻插值、双三次插值、双线性插值、主成分回归和支持向量回归的插值结果进行对比，实验结果表明：

1.无论使用什么插值方法，陆地和岛屿附近的 SST 插值效果都不如远离陆地和岛屿区域的。

2.GPR 的插值结果整体上优于传统的插值方法，无论是在远离陆地和岛屿的海洋区域，还是在陆地和岛屿边界处，GPR 的插值结果都优于其它方法。

3.GPR 插值算法在时间上和空间上的泛化能力都是可靠的。

然后分析了季节变化对插值算法的影响，结果表明季节变化对 GPR 插值算法的影响并不明显。最后讨论了算法的时间消耗，GPR 插值方法相比于传统插值方法的额外时间成本约为 2 秒(10%)，考虑到其对于传统方法改进的插值精度而言，算法的时间消耗是完全可以接受的。

第六章　总结与展望

6.1 论文总结

数值预报模式输出的降水量与实际情况往往存在偏差，需要进一步的偏差订正。然而传统的基于线性相关分析挑选的预报因子，以及在线性相关的基础上建立的预报模型，对于降水这种高度非线性的问题存在一定局限性。本论文针对这一具体问题，研究了机器学习方法对数值天气预报模式短期降水产品的偏差订正方法，同时探索了低分辨率偏差订正模型对高分辨率数据的偏差订正适用性。本论文的创新点具体如下：

1.针对线性相关分析挑选降水预报因子存在的局限性，提出了基于随机森林的降水预报因子选取方法。利用随机森林对预报因子进行重要性评估，挑选最重要的降水预报因子。同时，模式数据选择观测站附近的四个网格点预报数据而不再将其插值到站点上，这样既可以避免站点预报场插值计算的不准确，也能考虑到站点附近天气过程的影响。此外，本论文还对传统的预报因子筛选从两个方面做出了改进：a.在使用 LSTM 方法进行预报时，预报因子中除了模式预报数据之外，还加入了观测数据；b.除了相应预报时效的模式数据之外，还增加了其它预报时效的模式数据作为预报因子。

2.针对模式预报降水存在的误差，研究了基于机器学习的降水预报产品偏差订正方法。通过对大量训练数据的学习，首先使用随机森林建立降水晴雨分类模型和降水等级分类模型对模式预报降水进行订正，然后分别利用随机森林、支持向量回归和长短期记忆网络方法建立预报因子与降水量之间的回归方程，从而得到更加准确的降水量预报。屯溪站逐 3h 降水量预报实验的结果表明，机器学习方法的降水量偏差订正是十分有效的，其在各项指标上均优于多元线性回归方法的偏差订正。尤其是 LSTM 神经网络方法，预报的降水量的 RMSE 比 YHGSM 的模式输出的降水量降低了 57.6%，同时可以对强降水的降水量进行准确的预报。

3.针对当前低分辨率的数值预报产品无法满足解释应用的需求，研究了一种基于高斯过程回归的数值预报产品插值的方法，用于提升数值预报产品的空间分辨率，探索低分辨率偏差订正模型对高分辨率数据的偏差订正适用性。设计了基于高斯过程回归的空间插值算法并以海表面温度进行实验验证，实验结果表明该算法的插值结果相比传统插值方法的均方根误差更低，更能有效地提升数据分辨率。插值算法在时间上和空间上的泛化能力都是可靠的，并拥有相对较少的时间消耗。

6.2 研究展望

虽然本论文的研究取得了一定的成果，但是仍然有几个方面的问题需要在未来的研究工作中进一步完善，具体如下：

1.在使用机器学习方法建立降水预报偏差订正模型时，建立了过去三天的预报因子与未来 3h 降水量之间的回归方程，下一步将延长预报的时效，研究未来 24 小时内逐 3h 降水量的偏差订正预报。

2.本论文研究的基于机器学习的降水预报产品偏差订正方法仅针对单个站点进行预报，下一步将探索偏差订正方法的区域通用性，对一个区域的多个站点甚至全国范围内的站点的降水预报进行偏差订正预报。

3.在使用高斯过程回归进行空间插值时，目前只验证了该插值方法在 SST 插值上的有效性，下一步将使用高斯过程回归对 YHGSM 中与降水预报相关的所有气象因子进行插值提升其空间分辨率，并在此基础上使用机器学习方法对降水预报产品进行偏差订正，提高其实际应用价值。

致　谢

时间仿佛拥有了加速度。刚入学时还在埋怨日子过得真慢，可转眼间，两年半的硕士生涯马上就要结束了。坦白地说，我的性格是不大适合搞研究的，好动、不喜静，坐不得冷板凳，这是我的缺点，这两年里也在不断地鞭策自己。当初保研时又跨了专业，于是整个读研期间是比较吃力的。如果说现在的我也取得了一定的成果，那么首先要感谢的，是我的导师张卫民老师。

感谢张老师对我一直不抛弃不放弃，在我硕士期间的每一个关键节点上，从课程的选修到实验的设计，从论文的选题到论文的撰写……总能给我及时的鼓励和专业的指导，让我在灰心丧气时重整旗鼓，最终度过一个个难关。千言万语浓缩成三个字，谢谢您！

此外，还要感谢银福康老师对我的毕业论文的指导和帮助，从数据的获取、实验的进行，到毕业论文的撰写，银老师给我提供了极大的帮助。不仅如此，银老师积极向上的生活态度也一直感染着我，每次在朋友圈看到银老师的长跑打卡记录，总是会给我带来一种莫名的力量。感谢王辉赞老师在小论文撰写上给与的专业指导。

感谢冯淼师姐在实验上和编程技术上给予的耐心指导，让我从一个科研小白，逐渐成长为也能发表 SCI 论文的合格研究生。感谢师兄师姐们在生活上以及学习上提供的无私帮助，有你们在实验室的时候，会有一种莫名的安宁。

感谢和我同年级的蒋涛、范茂廷、戴俊、张琪、朱勋江、卢竞择以及桑浩同学，硕士生涯有你们，真好！

感谢我的父母，疫情期间在家躺了几个月，给您们添了太多麻烦！感谢陪伴我四年的女朋友，虽然异地见面不多，但是这期间有太多的心事只能与你诉说。

最后，以一首五言小诗结束本文：

汝阴游学子，偶入长沙城，修身治学苦，待得学成归。

参考文献

[1] Wu J, Long J, Liu M. Evolving RBF neural networks for rainfall prediction using hybrid particle swarm optimization and genetic algorithm[J]. Neurocomputing, 2015, 148:136-142.

[2] Lorenc A C. Analysis methods for numerical weather prediction[J]. Quarterly Journal of the Royal Meteorological Society, 2010, 112(474).

[3] Huth R, Richard M, Metelka L, et al. On the integrability of limited-area numerical weather prediction model ALADIN over extended time periods[J]. Studia Geophysica Et Geodaetica, 2003, 47(4):863-873.

[4] 余兴明, 卢绍宗. Windows media 编码器实现单收站全国天气会商网络视频广播[J]. 广西气象, 2006, 27(S1):122-123.

[5] Ke F, Huijun W, Young-Jean C. A physically-based statistical forecast model for the middle-lower reaches of the Yangtze River Valley summer rainfall[J]. Chinese Science Bulletin, 2008, 53(4):602-609.

[6] 潘晓滨,何宏让, 王春明. 数值天气预报产品解释应用[M]. 北京:气象出版社, 2016:1~3.

[7] Olson D A, Junker N W, Korty B. Evaluation of 33 Years of Quantitative Precipitation Forecasting at the NMC[J]. Weather & Forecasting, 1936, 10(3):498-511.

[8] Li W. Short-range forecast of heavy area rainfall combined PP method with MOS method in the upper reaches of Changjiang river[J]. Meteorological Monthly, 2003.

[9] Lai L L, Braun H , Zhang Q P , et al. Intelligent weather forecast[C]// International Conference on Machine Learning & Cybernetics. IEEE, 2004.

[10] Russell, Stuart J. Artificial Intelligence: A Modern Approach[M]. 北京:人民邮电出版社, 2002:25~28.

[11] 严明良, 曾明剑, 濮梅娟. 数值预报产品释用方法探讨及其业务系统的建立[J]. 气象科学, 2006, 26(1):90-96.

[12] Pruppacher H R, Klett J D, Wang P K. Microphysics of clouds and precipitation[J]. Aerosol Science & Technology, 1980, 28(4):381-382.

[13] Lin C, Vasi S , Kilambi A, et al. Precipitation forecast skill of numerical weather prediction models and radar nowcasts[J]. Geophysical Research Letters, 2005, 32(14):86~88.

[14] 寿绍文. 天气学分析[M]. 北京:气象出版社, 2002:18~20.

[15] Novak P, Brezkova L, Frolik P. Quantitative precipitation forecast using radar echo extrapolation[J]. Atmospheric Research, 2009, 93(1-3):328-334.

[16] Lorenc A C. Analysis methods for numerical weather prediction[J]. Quarterly

Journal of the Royal Meteorological Society, 2010, 112(474).

[17] Rodwell M J, Richardson D S, Hewson T D, et al. A new equitable score suitable for verifying precipitation in numerical weather prediction[J]. Quarterly Journal of the Royal Meteorological Society, 2010, 136(650):1344-1363.

[18] 矫梅燕. 天气业务的现代化发展[J]. 气象, 2010, 36(7):1-4.

[19] Peng X, Che Y, Chang J. A novel approach to improve numerical weather prediction skills by using anomaly integration and historical data[J]. Journal of Geophysical Research Atmospheres, 2013, 118(16):8814-8826.

[20] Huchao L I, Aimei S, Dengxin H E, et al. Application of Back-Propagation Neural Network in Predicting Non-Systematic Error in Numerical Prediction Model[J]. Plateau Meteorology, 2015, 42(6):1198-1201.

[21] Glahn H R, Lowry D A. The use of model output statistics in objective weather forecasting[J]. Journal of Applied Meteorology, 1972, 11:1203-1211.

[22] Sokol Zbynk. MOS-Based Precipitation Forecasts for River Basins[J]. Weather & Forecasting, 2003, 18(5):769-781.

[23] Kretzschmar R, Eckert P, Cattani D, et al. Neural Network Classifiers for Local Wind Prediction[J]. Journal of Applied Meteorology, 2004, 43(5):727-738.

[24] Dean A R, Fiedler B H. Forecasting Warm-Season Burnoff of Low Clouds at the San Francisco International Airport Using Linear Regression and a Neural Network[J]. Journal of Applied Meteorology, 2002, 41(6):629-639.

[25] Wilson L J, et al. The Canadian Updateable Model Output Statistics (UMOS) System: Validation against Perfect Prog[J]. Weather & Forecasting, 2003, 18(2):288-302.

[26] 丁士晟. 中国 MOS 预报的进展[J]. 气象学报, 1985(03):78-84.

[27] 孙永刚, 李彰俊, 孟雪峰等. 天气动力学组合因子在 MOS 降水预报中的应用[J]. 气象, 1998(02):27-30.

[28] 陈力强, 韩秀君, 张立祥. 基于 MM5 模式的站点降水预报释用方法研究[J]. 气象科技, 2003(05):13-17.

[29] 窝春蓉, 冯汉中. 利用卡尔曼滤波方法释用数值预报产品[J]. 高原山地气象研究, 2004, 024(002):16-19.

[30] Moustris K P, Larissi I K, Nastos P T, et al. Precipitation Forecast Using Artificial Neural Networks in Specific Regions of Greece[J]. Water Resources Management, 2011, 25(8):1979-1993.

[31] 王振友, 陈莉娥. 多元线性回归统计预测模型的应用[J]. 统计与决策, 2008, 000(005):46-47.

[32] 王达文, 王述舜, 班显秀等. MOS 预报业务系统[J]. 气象, 1987(11):31-32.

[33] Silverman D, Dracup J A. Artificial Neural Networks and Long-Range

Precipitation Prediction in California[J]. Journal of Applied Meteorology, 2000, 39(1):57-66.

[34] Azimi S, Moghaddam M A. Modeling Short Term Rainfall Forecast Using Neural Networks, and Gaussian Process Classification Based on the SPI Drought Index[J]. Water Resources Management, 2020, 34(4):1369-1405.

[35] Mark N, Witold F, Robert R. Cuykendall. Rainfall forecasting in space and time using a neural network[J]. J Hydro, 1992, 137(1-4):1-31.

[36] 孙全德, 焦瑞莉, 夏江江等. 基于机器学习的数值天气预报风速订正研究[J]. 气象, 2019, 45(03):132-142.

[37] 孙俊奎, 王占良, 张颖. 3 种修正的机器学习算法在逐 3h 降水量预报中的比较应用[J]. 甘肃科学学报, 2020.

[38] Cho D, Yoo C, Im J, et al. Comparative Assessment of Various Machine Learning-Based Bias Correction Methods for Numerical Weather Prediction Model Forecasts of Extreme Air Temperatures in Urban Areas[J]. Earth and Space Ence, 2020, 7(4).

[39] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.

[40] Hearst M A, et al. Support vector machines[J]. Intelligent Systems, 1998, 13(4):18-28.

[41] Smola A J, Schlkopf B. A tutorial on support vector regression[J]. Stats and Computing, 2004, 14(3):199-222.

[42] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[43] Rasmussen C E, Williams C K I. Gaussian Processes for Machine Learning[M]. MIT Press, 2005:65~66.

[44] Andreas M, Kopera M A, Marras S, et al. Strong Scaling for Numerical Weather Prediction at Petascale with the Atmospheric Model NUMA[J]. International Journal of High Performance Computing Applications, 2016, 100.

[45] Thuburn J, Wood N, Staniforth A. Normal modes of deep atmospheres. I: Spherical geometry[J]. Quarterly Journal of the Royal Meteorological Society, 2010, 128(584).

[46] 张诚忠. 不同因子处理方法对广西 MOS 方程降水预报准确率影响的试验[J]. 气象研究与应用, 2001, 22(3):24-28.

[47] Cortes C, VaPnik V. Support vector networks[J]. Machine Learning, 1995, 20:273-295.

[48] Schmidhuber, Jurgen. Deep Learning in Neural Networks: An Overview[J]. Neural Netw, 2015, 61:85-117.

[49] Li J. A Review of Spatial Interpolation Methods for Environmental Scientists[M]. Canberra:Geoscience Australia, 2008:11~12.

[50] Jia Y, Ma J. What can machine learning do for seismic data processing? An interpolation application[J]. Geophysics, 2017, 82(3): 163-177.

[51] Antonić O, Križan J, Marki A, Bukovec D. Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks[J]. Ecological Modelling, 2001, 138(1-3): 0-263.

[52] Bryan B A, Adams J M. Three-Dimensional Neurointerpolation of Annual Mean Precipitation and Temperature Surfaces for China[J]. Geographical Analysis, 2002, 34(2): 93-111.

[53] Li J, Heap A D, Potter A, Daniell J J. Application of machine learning methods to spatial interpolation of environmental variables[J]. Environmental Modelling & Software, 2011, 26(12): 1647-1659.

[54] Hofmann T, Schölkopf B, Smola A J. Kernel methods in machine learning[J]. The Annals of Statistics, 2008, 36(3): 1171-1220.

[55] Wang Q J, Zhang X F. Effective wind speed estimation for variable speed wind turbines based on WLS-SVM[J]. Journal of System Simulation, 2005, 17(7): 1590-1593.

[56] Paniagua-Tineo A, Salcedo-Sanz S, Casanova-Mateo C, Ortiz-García E G, Cony M A, Hernández-Martín E. Prediction of daily maximum temperature using a support vector regression algorithm[J]. Renewable Energy, 2011, 36(11): 3054-3060.

[57] He Z K, Liu G B, Zhao X J, Yang J. Temperature Model for FOG Zero-Bias Using Gaussian Process Regression[J]. Advances in Intelligent Systems and Computing, 2013, 37-45.

[58] He H, Siu W C. Single image super-resolution using Gaussian process regression[C]//CVPR. IEEE, 2011, 449-456.

[59] Thompson B, Tkalich P, Malanotte-Rizzoli P. Regime shift of the South China Sea SST in the late 1990s[J]. Climate Dynamics, 2017, 48(5-6): 1873-1882.

[60] Katsaros K B, Soloviev A V, Weisberg R H, Luther M E. Reduced Horizontal Sea Surface Temperature Gradients Under Conditions of Clear Skies and Weak Winds[J]. Boundary-Layer Meteorology, 2005, 116(2): 175-185.

[61] Du Y, Wang D, Xie Q, Church J. Harmonic analysis of sea surface temperature and wind stress in the vicinity of the maritime continent[J]. Journal of Meteorological Research, 2003, 17(S1): 226-237.

[62] Wang Z, Bovik A C, Sheikh H R, Simoncelli E P. Image Quality Assessment: From Error Visibility to Structural Similarity[J]. IEEE Transactions on Image Processing, 2004, 13(4).

作者在学期间取得的学术成果

[1] Zhang Y S, Feng M, Zhang W M, Wang H Z, Wang P Q. A Gaussian Process Regression-based Sea Surface Temperature Interpolation Algorithm[J]. Journal of Oceanology and Limnology. 2020.