



博士学位论文

基于机器学习的时间序列分析：以太阳黑子、泉流量和地震预报为例

作者姓名: 程 术

指导教师: 石耀霖 教授 中国科学院大学

学位类别: 理学博士

学科专业: 固体地球物理学

培养单位: 中国科学院大学 地球与行星科学学院

2021 年 12 月

Time series analysis based on machine learning: The cases of
sunspot, spring discharge and earthquake prediction

A dissertation submitted to
University of Chinese Academy of Sciences
in partial fulfillment of the requirement
for the degree of
Doctor of Natural Science
in Solid Earth Geophysics

By

Cheng Shu

Dissertation Supervisor: Professor Shi Yaolin

**College of Earth and Planetary Sciences, University of Chinese
Academy of Sciences**

December, 2021

中国科学院大学 学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师的指导下独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明或致谢。

作者签名：

日 期：

中国科学院大学 学位论文授权使用声明

本人完全了解并同意遵守中国科学院有关保存和使用学位论文的规定，即中国科学院有权保留送交学位论文的副本，允许该论文被查阅，可以按照学术研究公开原则和保护知识产权的原则公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存、汇编本学位论文。

涉密及延迟公开的学位论文在解密或延期后适用本声明。

作者签名：

导师签名：

日 期：

日 期：

摘要

时间序列分析是指从时间序列数据中挖掘有用信息的统计技术。随着机器学习的发展，尤其是具备长短期记忆单元的神经网络（Long Short-Term Memory Recurrent Neural Network，简称 LSTM-RNN）以及一维卷积神经网络（One Dimensional Convolutional Neural Networks，简称 1DCNN），基于机器学习的方法处理时间序列数据取得了一些成果。鉴于部分时间序列数据具有非平稳性、非高斯性和非线性等特征，基于机器学习的分析方法从这种类型的时间序列数据中挖掘有价值的信息仍然具有挑战性。本论文基于机器学习方法，根据数据集的复杂度选择了几种具备代表性的时间序列数据，试图构建不同输入和输出时间窗口长度探索时间序列数据的特征。本论文研究工作的内容和结论如下：

(1) 利用神经网络预测太阳黑子活动。太阳黑子是太阳局部强磁场活动在太阳光球表面上产生的黑色斑块，是太阳活动强弱的重要指标。太阳黑子活动影响着地球气候、人类空间活动等。针对第 25 太阳周的峰值与历史太阳周相比增强还是减弱，目前已有的研究并未得出统一的结论。传统意义上基于物理机制的模型被人为简化，因此这类模型难以准确捕获太阳黑子时间序列趋势。这里基于月均太阳黑子数量和面积，构建不同的输入时间窗口长度（72 个月、132 个月和 264 个月）和输出时间窗口长度（1 个月和 72 个月），利用 LSTM-RNN、1DCNN、具备 LSTM 神经元和一维卷积层的神经网络（Long Short-Term Memory and One Dimensional Convolutional Neural Networks，简称 LSTM-1DCNN）捕获太阳黑子时间序列趋势。最终发现，三种神经网络算法均能捕获到太阳黑子活动，且 LSTM-1DCNN 的预测性能略优于 LSTM-RNN 和 1DCNN。基于 LSTM-1DCNN 的预测结果发现，第 25 太阳周太阳黑子数峰值为 132.86（出现时间 2024 年 12 月），太阳黑子占比面积峰值为 1469.01（出现时间 2025 年 3 月）。两者结果显示，第 25 太阳周的峰值跟第 24 太阳周基本持平。

(2) 利用机器学习预测龙子祠泉流量变化。地下水过度开采导致井水干枯、淡水区的水量减少、地面沉降等现象。泉流量是地下水到地表水过渡的重要指标，反映了含水层中水位的动态变化。我们这里选取龙子祠泉作为研究对象，收集了该泉域的泉流量和周围九个区域的降水量，在设置不同输入和输出时间窗

口长度（分别设置为 1 至 4 个月）下，基于机器学习的分析方法预测泉流量。这些方法包括 LSTM-RNN、1DCNN、LSTM-1DCNN、支持向量机（Support Vector Regression，简称 SVR）、线性回归（Linear Regression，简称 LR）、随机森林（Random Forest，简称 RF）、决策树（Decision Tree，简称 DT）、K 近邻（k-Nearest Neighbor，简称 KNN）。总体来看，这些方法都适合预测未来泉流量变化；增加输入时间窗口长度，算法的预测能力会逐渐降低；增加输出时间窗口长度，算法的性能会出现一定幅度的下降。进一步研究发现，仅仅利用历史 1 个月的泉流量就能精确预测未来 1 个月龙子祠的泉流量。

(3) 利用机器学习对南加州地区进行中期地震预报。有效地预报强震能够减小人员伤亡和经济损失。地震预报包括时间、地点、震级、可能发生的概率这几个要素。根据预报时间的长短，地震预报可分为长期、中期和短期预报。短期预报不可控因素太多，这里不予以考虑；长期预报需要较长时间的观测资料，目前数据时长仍旧不够。因此，地震中期预报成为关注的重点。这里数据来源于美国南加州地区地震目录，该目录记录时间较长（~90 年），完备性较好，且南加州地区的地震活动较为频繁。从地震目录中计算了 16 种不同的地震因子，采用 LSTM-RNN、SVR、LR、RF、DT、KNN、梯度提升回归（Gradient Boosting Regression Tree，简称 GBRT）、极端随机森林回归（Extra Trees Regressor，简称 ETR）这几种方法，对区域内可能发生的最大震级进行了中期预报。从机器学习算法的表现来看，多数算法均出现了过拟合问题，即训练集的误差很小，而测试集的误差很大。对于 DT、KNN 和 ETR，训练集的误差甚至达到了 0，而且所有的训练集都处于置信区间内，即算法能够拟合训练集中的所有特征（包括噪声）。整体来看，本研究中机器学习对南加州地区的地震中期预报是过拟合的，算法预测的结果极易受到数据集扰动的影响，可能是因为输入数据中遗漏了未知的重要变量或需要更长时间的数据集。

综合以上三个方向，发现利用机器学习探索时间序列数据时，输入和输出时间窗口长度会影响算法的性能：适当的时间窗口长度有助于提高算法性能；增加输出时间窗口长度会降低算法性能。机器学习分析时间序列数据时，需要考虑到数据集的完备性，还要找到合适的输入特征。

关键词：机器学习，时间序列，泉流量，太阳黑子，中期地震预报

Abstract

Time series analysis is a statistical technique for mining useful information from time series data. With the progress of machine learning (ML), especially the long short term memory recurrent neural network (LSTM-RNN) and one dimensional convolutional neural networks (1DCNN), there are achieving some results dealing with time series data based on ML. However, time series data owns characteristics of non-stationary, non-Gaussian and nonlinear, thus it is difficult to extract information from time series data by ML. Based on ML, we select several time series data, and try to construct different input and output time window length to explore time series data. The contents and conclusions of the research works are as follows:

(1) **Prediction of sunspot activity by neural network.** Sunspots are black patches produced on the surface of the solar photosphere by the strong local magnetic activity of the Sun, and an important indicator of the strength of solar activity. Sunspot activity will affect earth's climate, human space activities and so on. There are still no unified conclusion on whether the peak value of the cycle 25 shows an upward or downward trend compared with the historical solar cycle. The traditional physical models are simplified, which makes it difficult for accurately capturing the sunspot time series trend. Here, we use the monthly average sunspot number and area, construct different input time window length (72 months, 132 months or 264 months) and output time window length (one month or 72 months), and adopt LSTM-RNN, 1DCNN and LSTM-1DCNN (Long Short-Term Memory and One Dimensional Convolutional Neural Networks). Finally, we find that LSTM-RNN, 1DCNN and LSTM-1DCNN can capture the sunspot activity and the performance of LSTM-1DCNN is slightly better than LSTM-RNN and 1DCNN. Based on the prediction of LSTM-1DCNN, the maximum sunspots' number and area in the cycle 25 predicted by LSTM-1DCNN is 132.86 which occurs in December 2024 and 1469.01 which occurs in March 2025. Both results show that the peak values of cycle 25 are basically the same as cycle 24.

(2) **Predicting spring discharge in Longzici spring by machine learning.** Over

exploitation of groundwater results in dry well water, reduction of water volume in fresh water areas, land subsidence and so on. Spring discharge is an important indicator of the transition from groundwater to surface water, which reflects the dynamic change of aquifer and the normal operation of water flow system in spring area. We choose the Longzici spring. The dataset include the spring discharge and the precipitation at nine surrounding areas, and we set the input and output time window length from 1 to 4 months respectively, and use LSTM-RNN, 1DCNN, long short term memory and one dimensional convolutional neural networks (LSTM-1DCNN), support vector regression (SVR), linear regression (LR), random forest (RF), decision tree (DT), and k-nearest neighbor (KNN) to simulate the trend of spring discharge. Overall, these methods are suitable for predicting the various of spring discharge. Increasing the input time window length will not improve the performance of the models; Increasing the output time window length will slightly reduce the performance of the model. Otherwise, we find that historical spring discharge for one month can accurately predict the future spring discharge.

(3) **Medium term earthquake prediction in Southern California by machine learning.** Effective forecasting of powerful earthquakes can reduce casualties and economic damage. Earthquake prediction involves time, location, magnitude and probability of occurrence. According to time span, earthquake prediction can be divided into long-term, medium-term and short-term prediction. For short-term prediction, there are too many uncontrollable factors; For long-term prediction, the observation dataset are required at least hundreds of years. Our observations duration are still insufficient. Therefore, medium-term earthquake prediction becomes the focus of our study. The seismic dataset comes from the earthquake catalog of Southern California, USA. This catalog is an ideal database because it has been recoding for a long time (~ 90 years) with good completeness, and seismic activity is frequent in the Southern California region. From the earthquake catalog, we calculate 16 different earthquake factors and use several ML methods, namely LSTM-RNN, SVR, LR, RF, DT, KNN, gradient boosting regression tree (GBRT) and extra trees regression (ETR), to predict the maximum earthquake magnitude that may occur in the region within the medium term. From the

performance of all models, they have a large degree of over fitting. That is, the error of the training set is very small, while the error of the testing set is very large. For DT, KNN and ETR models, the error of the training set even reaches 0 and the scope of the training data set is in the confidence interval. That is, the model can fit all the features of the training set, including noises. It can be seen from these results that ML methods appears over fitting to explore the medium-term earthquake prediction in Southern California in this study. The possible reasons are followings: We miss some important factors; We need to collect more earthquake catalog.

Combining these three field, we find that the input and output time window length will affect the performance of the model when exploring time series data by machine learning. Choosing the right input time window length helps to improve the model performance; Increasing the output time window decreases the model performance. Machine learning needs to consider the completeness of the dataset and finding the appropriate input features when exploring time series.

Keywords: Machine Learning, Time Series, Spring Discharge, Sunspot, Earthquake Prediction

目 录

第 1 章 引言	1
1.1 问题的提出	1
1.2 研究现状、挑战与机遇	3
1.3 论文结构安排	4
第 2 章 机器学习的基础理论	6
2.1 引言	6
2.2 机器学习算法	7
2.2.1 线性回归 (LR)	7
2.2.2 k 近邻 (KNN)	8
2.2.3 决策树 (DT)	9
2.2.4 随机森林 (RF)	10
2.2.5 极端随机森林回归 (ETR)	11
2.2.6 梯度提升回归树 (GBRT)	11
2.2.7 支持向量回归 (SVR)	11
2.2.8 卷积神经网络 (CNN)	13
2.2.9 长短期记忆循环神经网络 (LSTM-RNN)	14
2.3 训练前的必要准备	16
2.3.1 滑动窗口法	17
2.3.2 数据归一化	17
2.3.3 优化器	18
2.3.4 优化目标函数	19
2.3.5 性能度量	20
2.3.6 欠拟合和过拟合	21
2.3.7 计算工具与平台	24
2.4 小结	24
第 3 章 基于神经网络预测太阳黑子活动	25
3.1 研究背景	25
3.2 数据与方法	27
3.2.1 数据简介	27
3.2.2 方法描述	30
3.3 试验结果与分析	31

3.3.1 预测未来 1 个月太阳黑子活动	32
3.3.2 预测未来 72 个月太阳黑子活动	33
3.3.3 与其他研究的比较	42
3.4 讨论与小结	42
第 4 章 基于机器学习预测龙子祠泉流量	44
4.1 研究背景	44
4.2 数据与方法	46
4.2.1 研究区域	46
4.2.2 数据描述	46
4.2.3 方法描述	50
4.3 试验结果与分析	51
4.3.1 预测未来 1 个月泉流量	53
4.3.2 预测未来 2 个月泉流量	53
4.3.3 预测未来 3 个月泉流量	55
4.3.4 预测未来 4 个月泉流量	57
4.3.5 利用历史泉流量预测未来泉流量	59
4.4 讨论与小结	62
第 5 章 基于机器学习对南加州地区的地震中期预报	64
5.1 研究背景	64
5.2 数据与方法	66
5.2.1 研究区域	66
5.2.2 地震目录	66
5.2.3 地震因子	68
5.3 结果分析	70
5.3.1 基于 6 个区块预测未来 1 年的最大震级	70
5.3.2 基于整个区块预测未来 1 年的最大震级	74
5.3.3 基于整个区块预测未来 10 年的最大震级	76
5.4 讨论与小结	82
第 6 章 总结与展望	84
附录 A 附录	90
A.1 基于机器学习对南加州地区的地震中期预报	90
A.1.1 表	90
A.1.2 图	91
参考文献	97

致谢	105
作者简历及攻读学位期间发表的学术论文与研究成果	106

图形列表

1.1 人工智能、机器学习和深度学习的逻辑关系	2
2.1 两层 LSTM-RNN 的示例	15
2.2 LSTM 神经元示意图	15
2.3 滑动窗口法	17
2.4 AIC 信息准则	23
3.1 1949 年至 2021 年期间月均太阳黑子数	27
3.2 1949 年至 2021 年期间太阳黑子面积随时间和纬度变化的蝴蝶图	29
3.3 1874 年至 2021 年期间月均太阳黑子数和面积	30
3.4 最佳模型预测未来 1 个月的太阳黑子数	34
3.5 最佳模型预测未来 1 个月的太阳黑子面积	35
3.6 最佳模型预测未来 72 个月的太阳黑子数	38
3.7 最佳模型预测未来 72 个月的太阳黑子面积	41
4.1 龙子祠泉地理条件	47
4.2 龙子祠泉流量变化趋势	47
4.3 龙子祠泉周围 9 个区域降水量随时间变化的趋势图	48
4.4 1987 年至 2018 年期间龙子祠泉年域降水量与泉流量分布	49
4.5 龙子祠泉多年月均泉流量与降水量变化趋势	49
4.6 不同输入时间窗口长度下最佳模型预测未来 1 个月泉流量	54
4.7 不同输入时间窗口长度下最佳模型预测未来 2 个月泉流量	56
4.8 不同输入时间窗口长度下最佳模型预测未来 3 个月泉流量	58
4.9 不同输入时间窗口长度下最佳模型预测未来 4 个月泉流量	60
4.10 最佳模型利用历史 1 个月泉流量预测未来 1 个月泉流量	62
5.1 南加州地区区域构造	66
5.2 震级与频度的关系	67
5.3 确定 b 值的两种方法	68
5.4 b_{mle} 与 b_{lstsq}	70
5.5 基于 6 个区块预测未来 1 年的最大震级	72
5.6 不同模型基于区块 1 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	73

5.7 不同模型基于整个区块预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	75
5.8 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	78
5.9 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.85: 0.15）	79
5.10 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.9: 0.1）	81
A.1 不同模型基于区块 2 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	92
A.2 不同模型基于区块 3 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	93
A.3 不同模型基于区块 4 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	94
A.4 不同模型基于区块 5 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	95
A.5 不同模型基于区块 6 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8: 0.2）	96

表格列表

2.1 过拟合与欠拟合的判断标准	22
3.1 13 个月平滑的月均太阳黑子数量、峰值和谷值出现的时间	28
3.2 不同模型在不同的输入时间窗口长度下预测未来 1 个月太阳黑子数的拟合指标效果	32
3.3 最佳模型预测的太阳黑子活动	32
3.4 不同模型和输入时间窗口长度下预测未来 1 个月太阳黑子面积的拟合指标效果	36
3.5 不同输入时间窗口长度和层数下 LSTM-1DCNN 预测未来 72 个月太阳黑子数的拟合指标效果	37
3.6 最佳模型预测未来 72 个月太阳黑子活动的最大值	39
3.7 不同输入时间窗口长度和层数的 LSTM-1DCNN 预测未来 72 个月太阳黑子面积的拟合指标效果	40
3.8 不同研究给出的预测第 25 太阳周的太阳黑子活动相比于第 24 太阳周的强弱	42
4.1 不同模型在不同的输入和输出时间窗口长度下预测泉流量的拟合指标效果	52
4.2 最佳模型预测 2019 年 1 月泉流量	53
4.3 最佳模型预测 2019 年 1 月和 2 月泉流量	55
4.4 最佳模型预测 2019 年 1 月至 3 月泉流量	55
4.5 最佳模型预测 2019 年 1 月至 4 月泉流量	59
4.6 不同模型利用历史 1 个月泉流量预测未来 1 个月泉流量的拟合指标效果	61
4.7 不同模型利用历史 1 个月泉流量预测 2019 年 1 月泉流量	61
5.1 1932 年至 2021 年期间研究区域内 7 级以上的地震	66
5.2 基于地震目录的地震因子	69
5.3 每个时空窗口至少有 30 个地震数的时空窗口信息	71
5.4 不同模型基于区块 1 预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）	74
5.5 不同模型基于整个区块预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）	76
5.6 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）	77

5.7 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数 据集划分比例为 0.85: 0.15）	77
5.8 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数 据集划分比例为 0.9: 0.1）	80
A.1 不同模型基于区块 2 至区块 6 预测未来 1 年最大震级的拟合指标效 果（数据集划分比例为 0.8: 0.2）	90
A.1 不同模型基于区块 2 至区块 6 预测未来 1 年最大震级的拟合指标效 果（数据集划分比例为 0.8: 0.2）（续）	91

符号列表

缩写

ANNs	Artificial Neural Networks
ML	Machine Learning
CNN	Convolutional Neural Network
1DCNN	One Dimensional Convolutional Neural Network
2DCNN	Two Dimensional Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
LSTM-RNN	Long Short-Term Memory Recurrent Neural Network
LSTM-1DCNN	Long Short-Term Memory and One Dimensional Convolutional Neural Network
SVC	Support Vector Classification
SVR	Support Vector Regression
KNN	k -Nearest Neighbor
DT	Decision Tree
RF	Random Forest
GBRT	Gradient Boosting Regression Tree
GBCT	Gradient Boosting Classification Tree
ETR	Extra Trees Regression
ETC	Extra Trees Classification
LR	Linear Regression
LC	Linear Classification
LS	Least Square
MLE	Maximum Likelihood Estimation
MSE	Mean Squared Error

RMSE	Root Mean Squared Error
MSSN	Monthly Mean Total Sunspot Number
MSSA	Monthly Mean Total Sunspot Area

第1章 引言

1.1 问题的提出

近几十年来数据采集和存储技术突飞猛进，科学界可用数据总量和种类显著增加，挖掘数据中隐藏的信息非常具有挑战性，这是因为人类在给定时间内处理和分析大量数据的能力非常有限 (Bougher, 2016)。很多自然现象是由多种因素所致，基于这些因素所构建的模型复杂度较高 (Reitsma, 2010)，因此人类很容易丢失一些从数据中捕获有效信息的机会，进而对复杂现象内在机制的理解存在一定的局限 (Feyyad, 1996)。

人类一直致力于理解自然现象的本质，期望在不同的条件下做出最优决策。过去几十年里，人类创造和收集的数据远超出了从数据中提取的信息量，很多模型的预测能力难以与数据增长的速度同步，因此数据的价值并未完全被开发出来。为了最大化利用爆炸式增长的数据，人类需要解决几个重要问题：(1) 能够从大量数据中提取信息；(2) 从数据中推导出的模型比传统的经验模型能获取到更多的信息；(3) 这些信息需要遵循自然界普适性的规律。

人工智能 (Artificial Intelligence, 简称 AI) 是一个大领域，与任何智力工作相关。图灵测试指出，询问者在提出一些书面问题后，如果人类不能区分书面回答来自人类还是计算机，那么这台计算机就通过了测试 (Turing, 1950)。图1.1展示了 AI、机器学习 (Machine Learning, 简称 ML) 和深度学习的逻辑关系。ML 是 AI 的主流算法，深度学习又是 ML 近些年来发展最迅速的分支之一。本论文选择 ML 作为重点关注的技术，它是机器从数据中通过算法学习规律进而预测新数据的方法。

根据目标任务的差异，ML 可用于分类、回归、聚类和降维等问题；根据数据结构的差异，ML 可以分为空间结构数据学习、序列结构数据学习和时空结构数据学习；根据数据是否带有标签以及标签的数量，ML 又可分为监督学习、半监督学习和无监督学习等。监督学习处理的数据全部带有标记，常被用来处理分类和回归问题。ML 中多数算法都属于监督学习，如线性回归 (Linear Regression, 简称 LR)、支持向量回归 (Support Vector Regression, 简称 SVR)、决策树 (Decision Tree, 简称 DT)、随机森林 (Random Forest, 简称 RF)、梯度提

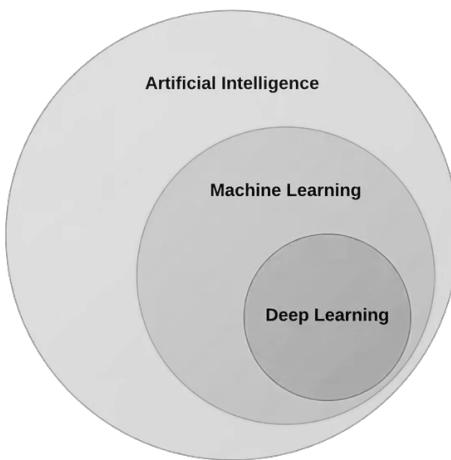


图 1.1 人工智能、机器学习和深度学习的逻辑关系。

Figure 1.1 The relationship between artificial intelligence, machine learning and deep learning.

升回归树（Gradient Boosting Regression Tree，简称 GBRT）、极端随机森林回归（Extra Trees Regression，简称 ETR）、 k 近邻（ k -Nearest Neighbor，简称 KNN）、卷积神经网络（Convolutional Neural Network，简称 CNN）、长短期记忆神经网络（Long Short-Term Memory Recurrent Neural Network，简称 LSTM-RNN）等。

很多自然现象可看作是一个复杂系统，例如气候变暖、地震发生机制等 (Fan 等, 2021)。现有的可操作且简化的经验模型通过降低复杂度（研究个别现象）或加强假设/近似进行推论，但对于大尺度系统的运行机制难以精准描述。ML 模型排除了传统经验模型中人为干扰的因素，直接从数据中学习，进而为探究复杂现象提供了一条新的研究路径。

目前，机器学习在复杂问题中的应用领域较为广泛，这里重点关注地球科学领域。地球科学中很多数据集属于时空结构数据，例如大气模拟、海洋运输、火势蔓延、土壤运动、植被碳循环过程等，这些都是时空动力学领域内重要的研究问题 (Mathieu 等, 2015; Oh 等, 2015)；卫星图像的地形和纹理特征属于空间结构数据，类似于计算机视觉，研究对象通常具有一些描述边缘、纹理、形状和颜色等特征，利用这些特征构建机器学习算法，从而实现对象的定位、分类和检测等目标 (Lee 等, 1990)；地震波动态时间序列类似于自然语言与信号识别，属于时间序列数据 (Rouet-Leduc 等, 2017; Perol 等, 2018; DeVries 等, 2018)。

根据数据出现的时间顺序排列的数据称为时间序列。时间序列预测基于时间结构数据，可利用机器学习分析时间序列数据。自然界中很多预测问题都涉及到时间因素。可通过时间序列推断过去数据点序列发生的事件，预测未来将会发

生的事件。本论文根据数据复杂度的不同选取了几种时间序列数据，研究太阳黑子活动、龙子祠泉流量的动态趋势以及南加州地区中期地震预报。

1.2 研究现状、挑战与机遇

深度学习兴起前，关于计算机视觉和语音识别等目标任务多数是基于人工设计的特征和典型的机器学习方法。在计算机视觉领域，常用的人工设计的特征包括 SIFT、SURF、HOG、HOF、MBH 等。人工设计的特性增强了对解释性驱动因素的控制，但这个过程繁琐，迁移能力差，结果可能是非最优的。与使用人工设计特征的算法相比，使用由数据驱动的机器学习有时非常有效。通常，机器学习算法在处理复杂任务方面优于物理模型，尤其是在物理机制不完善而观测数据较多时。机器学习可预测某时间段内相对静止的特征，也可用来研究动态变化过程。与半经验性的物理建模方法相比，机器学习少了人工的干预，使用起来更加灵活。

尽管机器学习常见的应用领域（计算机视觉和语音识别）与地球科学中的实际问题存在很多的相似之处，但还是有不同的方面。例如，计算机视觉处理的照片有三个通道（红绿蓝），但高光谱卫星图像会扩展到数百个光谱通道，远超出可见光范围，这与自然图像的统计特性不同；而且高光谱卫星图像的光谱通道存在空间上的依赖性，违反了数据独立同分布的重要假设；另外，由于地球科学研究中使用不同种类的传感器采集不同类型的数据，这些数据存在不同的成像几何、时空分辨率等特征，很难集成这些数据，而且多种类型的传感器收集到的观测数据存在噪声来源不同、数据丢失和系统偏差等问题。

一般地，机器学习需要具备以下几个要素：

- **解释能力。**在解决实际问题时，不仅要考虑提高预测的精度，更要重点关注预测结果的可解释性。目前来说，可解释性是机器学习普遍存在的缺陷 (Montavon 等, 2018)。机器学习从观测数据中得到的是统计相关关系，而不是因果关系 (Runge 等, 2015; Reichstein 等, 2019)。为了提高机器学习的可解释性，我们需要挑选合适的输入特征；再者，基于神经网络的算法能够给出每个隐藏层的特征值，可视化这些特征值在一定程度上也能观测到算法学到的规则。

- **泛化能力。**机器学习通常在训练时性能表现良好，但在外推时结果可能会出现很大的偏差，这可能是由过拟合或观测数据中存在的偏差所致 (Friedlingstein

等, 2014)。为达到预期的任务目标, 可减小算法的规模或提供更丰富的数据。

• **表达能力。**表达能力本质上是函数逼近。万能逼近定理指出, 一层的神经网络几乎可以近似所有的连续函数。但一层的神经网络在实际应用时很难拟合。可通过增加网络层数和节点数获取更强表达能力的神经网络。

一般地, 物理建模和机器学习被看作是两类学科。前者由理论驱动建模, 后者由数据驱动建模。在原理方面, 物理方法可以解释; 在适应数据方面, 机器学习具备高度的灵活性。从机器学习发展趋势来看, 物理建模和机器学习协同作用能够增加模型的可信度 (Karpatne 等, 2017a,b; Camps-Valls 等, 2018)。

1.3 论文结构安排

本论文基于不同的时间序列数据, 利用机器学习方法进行处理和分析, 发现和理解新知识, 推动机器学习在时间序列问题中的应用。第2章为机器学习理论部分, 简要介绍论文涉及到的机器学习的技术原理和方法, 然后引出这些算法训练前必要的准备条件。按照数据集的复杂程度, 此后几个章节分别研究太阳黑子活动、泉流量变化趋势和地震中期数值预报。

第3章基于一种输入特征, 利用神经网络探测太阳黑子活动。太阳黑子是太阳局部强磁场活动在太阳光球层上产生的黑色斑块, 是太阳活动强弱的重要指标。太阳黑子活跃时会影响地球环境和人类经济发展。若人类能够提前预知太阳黑子活动并做出相应的防御措施, 可以减少太阳活动带来的损失。从长期记录来看, 近几十年来太阳黑子峰值呈现持续下降的趋势。为了继续跟踪太阳黑子活动, 我们采用神经网络预测未来 1 个月、72 个月太阳黑子活动。本论文中将未来 72 个月太阳黑子活动最剧烈时视为第 25 太阳周的峰值。这里重点关注第 25 太阳周太阳黑子活动的峰值。

第4章基于两种输入特征, 利用机器学习预测龙子祠泉流量变化。近些年来, 人类过度开采地下水, 使部分泉水面临干涸的风险, 因此合理管理地下水的用量就显得尤为重要。如果能够准确预知未来 1 个月甚至几个月地下水动态变化趋势, 就能够最大化地管理地下水, 从而实现地下水的可持续供应。龙子祠泉具备喀斯特地貌特征, 地下水流路径错综复杂, 基于半经验性的物理模型难以准确预测地下水变化趋势。机器学习擅长处理复杂问题, 为预测龙子祠泉的地下水位提供了研究方法。

第5章基于多种输入特征，利用机器学习对南加州地区进行地震中期数值预报。原始数据集为1932年至今的南加州地区地震目录。这里选择了中期预报，是因为长期预报需要长时间的数据资料积累（南加州地区地震目录记录最长年限为~90年），而短期预报机制更加复杂。基于地震目录我们得到16个地震因子，其中7个地震因子与空间地震分布密切相关。基于这些地震因子，我们尝试探索未来可能发生的最大震级。

第6章分别对太阳黑子活动、泉流量动态变化和地震中期数值预报这几个目标任务进行了总结与展望。需要指出的是，本论文中很多关于统计学、机器学习中的专业术语并未加以详细区分，比如算法/模型/方法/映射、机器学习/神经网络/深度学习、训练/优化/拟合/模拟/学习等。论文中这些含义近似的术语将会被无差别使用。

第2章 机器学习的基础理论

本章首先介绍了机器学习的基本概念（第2.1节）；接着介绍论文中涉及的机器学习算法的基础理论，包括LR（第2.2.1节）、KNN（第2.2.2节）、DT（第2.2.3节）、RF（第2.2.4节）、ETR（第2.2.5节）、GBRT（第2.2.6节）、SVR（第2.2.7节）、CNN（第2.2.8节）、LSTM-RNN（第2.2.9节）；然后引出这些算法训练前的必要准备，包括滑动窗口法（第2.3.1节）、数据归一化（第2.3.2节）、优化器（第2.3.3节）、优化目标函数（第2.3.4节）、性能度量（第2.3.5节）、欠拟合和过拟合（第2.3.6节）、计算工具与平台（第2.3.7节）。

2.1 引言

本质上，各类算法设计的初衷是为解决某一特定任务。若想将某一种算法迁移到其他任务中，可能需要再次耗费大量的精力重新修缮算法。机器学习算法从大量样本中学习，若想将训练好的机器学习算法迁移到其他类似的任务上，只需要在原算法的基础上稍微进行调整训练。因此，即使面临的任务有所差异，学到的算法仍具有一定的通用性 (Goodfellow 等, 2016)。

机器学习从数据中提取知识。或者说，机器学习是一种从数据中学习经验的算法。Mitchell 等 (1997) 将机器学习定义为：“假设性能 P 用来评估计算机计算的某类任务 T ，若该计算利用经验 E 在任务 T 上性能 P 有所提升，则对于 T 和 P 而言，计算机程序从 E 进行了有效的学习。”

传统的机器学习从训练样本出发，试图通过数据本身而不是原理分析获得规律，实现对数据行为或趋势的准确预测。机器学习通过平衡学习结果的有效性与学习算法的可解释性，为解决有限样本的学习任务提供了研究范式。本论文的研究方法将重点集中在监督学习算法上。被标记的数据可来自实际观测，也可来自数值模拟。基于这些被标记的数据集，使用监督学习算法挖掘数据中隐藏的信息，从而利用评估良好的算法预测未来趋势。

2.2 机器学习算法

假设数据集为 $D = \{(\mathbf{x}_1, \mathbf{y}_1), (\mathbf{x}_2, \mathbf{y}_2), \dots, (\mathbf{x}_m, \mathbf{y}_m)\}$, 其中 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, \mathbf{X} 为输入向量, $\mathbf{x}_i \in \mathbf{X}$ 为输入特征空间, $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m\}$, \mathbf{Y} 为输出向量, $\mathbf{y}_i \in \mathbf{Y}$ 为输出特征空间。抽取数据集 D 中部分数据作为训练集, 学习从输入空间 \mathbf{X} 到输出空间 \mathbf{Y} 的映射 $f : \mathbf{X} \rightarrow \mathbf{Y}$ 。通过该映射, 使用未经训练的数据(即测试集)评估算法。测试过程中要选择性能度量的指标。需要注意的是, 机器学习重点关注的是算法能够很好地适应测试集, 而不仅仅在训练集中表现良好, 因此要求算法具备较强的泛化(Generalization)性能。

2.2.1 线性回归 (LR)

线性回归是机器学习中很多算法的基石, 是一种经典的、简单的回归算法, 经常用来预测回归问题。机器学习中许多非线性算法都是在线性回归算法的基础上通过引入高维映射或层级结构演化而来。线性回归通过学习一个函数 $f(\mathbf{X})$ (式2.1), 使函数值 $f(\mathbf{X})$ 与目标值 \mathbf{Y} 尽可能地接近。

$$f(\mathbf{X}) = \mathbf{W}^T \mathbf{X} + \mathbf{b}. \quad \dots (2.1)$$

其中, \mathbf{W} 和 \mathbf{b} 都是待定的参数。确定 \mathbf{W} 和 \mathbf{b} 的关键在于衡量 $f(\mathbf{X})$ 与 \mathbf{Y} 之间的距离, 均方差(Mean Square Error, 简称 MSE)是最常用的距离度量指标。这里试图让 MSE 最小化。将 \mathbf{W} 和 \mathbf{b} 合并成向量 $\hat{\mathbf{W}} = (\mathbf{W}; \mathbf{b})$ 。根据均方差的定义, 可将式2.1写成

$$J(\hat{\mathbf{W}}) = \min_{\hat{\mathbf{W}}} \sum_{i=1}^m (f(\mathbf{x}_i) - \mathbf{y}_i)^2. \quad \dots (2.2)$$

当样本特征较多而样本数较少时, 线性回归算法很容易陷入过拟合。在线性回归的基础上对式2.2加上正则化项, 可有效消除多重共线性特征的影响, 在一定程度上缓解了过拟合问题。若使用 L_2 范数正则化, 则有

$$J(\hat{\mathbf{W}}) = \min_{\hat{\mathbf{W}}} \sum_{i=1}^m (f(\mathbf{x}_i) - \mathbf{y}_i)^2 + \lambda \|\hat{\mathbf{W}}\|_2^2. \quad \dots (2.3)$$

其中, 正则化参数 $\lambda > 0$ 。式2.3亦称岭回归(Ridge Regression)。若使用 L_1 范数正则化, 则有

$$J(\hat{\mathbf{W}}) = \min_{\hat{\mathbf{W}}} \sum_{i=1}^m (f(\mathbf{x}_i) - \mathbf{y}_i)^2 + \lambda \|\hat{\mathbf{W}}\|_1. \quad \dots (2.4)$$

其中，正则化参数 $\lambda > 0$ 。式2.4亦称 LASSO (Least Absolute Shrinkage and Selection Operator)。

利用最小二乘法对 \mathbf{W} 和 \mathbf{b} 进行估计，则有

$$\hat{\mathbf{W}}^* = \arg \min_{\hat{\mathbf{W}}} (\mathbf{Y} - (\mathbf{X}; \mathbf{I})\hat{\mathbf{W}})^T (\mathbf{Y} - (\mathbf{X}; \mathbf{I})\hat{\mathbf{W}}). \quad \dots (2.5)$$

令 $E_{\hat{\mathbf{W}}} = \hat{\mathbf{W}}(\mathbf{Y} - (\mathbf{X}; \mathbf{I})\hat{\mathbf{W}})^T(\mathbf{Y} - (\mathbf{X}; \mathbf{I})\hat{\mathbf{W}})$ ，对 $\hat{\mathbf{W}}$ 求导得到

$$\frac{\partial E_{\hat{\mathbf{W}}}}{\partial \hat{\mathbf{W}}} = 2(\mathbf{X}; \mathbf{I})^T((\mathbf{X}; \mathbf{I})\hat{\mathbf{W}} - \mathbf{Y}). \quad \dots (2.6)$$

令式2.6为零，可得 $\hat{\mathbf{W}}$ 最优解的封闭式：

$$\hat{\mathbf{W}} = [(\mathbf{X}; \mathbf{I})^T(\mathbf{X}; \mathbf{I})](\mathbf{X}; \mathbf{I})^T \mathbf{Y}. \quad \dots (2.7)$$

2.2.2 k 近邻 (KNN)

KNN 是机器学习中常用的监督学习方法。工作机制可描述为：给定测试样本，基于某种度量找出训练集中与其最接近点的 k 个训练样本，再基于这 k 个样本的信息进行预测。在分类问题中，使用投票法，即利用 k 个样本中出现最多的样本类别作为最终的预测结果；在回归问题中，使用平均值法，即利用 k 个样本的实际值的平均作为最终的预测结果。这两类问题也可以基于距离远近进行加权平均或加权投票。KNN 的关键在于选择合适的 k 值、距离度量和决策规则。KNN 在训练时会将整个数据集一次性输入，因此需要大量的内存空间来存储数据。在高维情况下，KNN 容易出现维数灾难，即样本稀疏、距离计算困难等问题。

以分类问题举例。在数据和标签已知的情况下，输入测试样本，将测试样本的特征与训练集中对应的特征相互比较，找到训练集中与之最为相似的前 k 个样本，则该测试样本对应的类别就是 k 个数据中出现次数最多的那个分类。KNN 算法可描述为：

- (1) 计算测试样本与各个训练样本之间的距离；
- (2) 按照距离的递增关系进行排序；
- (3) 选取距离最小的 k 个样本；
- (4) 确定前 k 个样本所在类别的出现频率；
- (5) 返回前 k 个样本中出现频率最高的类别作为测试样本的预测分类。

2.2.3 决策树 (DT)

根据目标任务类型的不同，决策树可分为分类树和回归树。分类树用于处理离散型数据，回归树用于处理连续型数据。决策树由节点和有向边组成。一棵决策树包含一个根节点、若干个内部节点和若干个叶节点。根节点包含整个样本集 \mathbf{X} ，内部节点表示特征，叶节点对应于决策结果。

在进行分类或回归任务时，从根节点开始，对样本的某一特征进行判定测试，根据测试结果划分到子节点；这时每个子节点对应着该特征的一个取值，这个过程不断循环，直至目标到达叶节点。本质上，决策树是将空间超平面进行划分的一种方法。每分割一次，都会将当前的空间根据特征的取值进行划分，使每个叶节点在空间中不会相交。

算法 1 决策树学习基本算法

Input: 数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$ ； 特征集 $A = \{a_1, a_2, \dots, a_d\}$.

```

1: procedure TREE( $D, A$ )
2:   生成节点 Node;
3:   if 数据集  $D$  中的所有样本属于同一类别  $C$ , 即  $D \subset C$  then
4:     将 Node 标记为  $C$  类叶节点; return
5:   end if
6:   if  $A = \emptyset$  OR  $D$  中样本在  $A$  上取值相同 then
7:     将 Node 标记为叶节点, 其类别标记为  $D$  中样本数量最多的类; return
8:   end if
9:   从  $A$  中选择最优划分特征  $a^* = \{a_1^*, a_2^*, \dots, a_v^*\}$ ;
10:  for  $a_i^*(i \in \{1, 2, \dots, v\})$  do
11:    在 Node 下再生成一个分支;  $D_v$  表示  $D$  在  $a^*$  上取值为  $a_v^*$  的样本子集;
12:    if  $D_v = \emptyset$  then
13:      将分支节点标记为叶节点, 其类别标记为  $D$  中样本最多的类; return
14:    else 以 Tree( $D_v, A, \{a^*\}$ ) 为分支节点
15:    end if
16:  end for
17: end procedure

```

Output: 输出最优决策树 Tree(D, A).

决策树的基本流程遵循分而治之 (Divide and Conquer) 的策略，见算法1。决策树的生成过程是递归的。决策树在以下三种情况时会出现递归：

- 当前节点包含的样本全部属于同一类别，这种情况不需要划分（见算法1中第4行）；
- 当前特征集为空，或所有样本在所有特征上取值均相等，无法划分（见算法1中第7行）；
- 当前节点包含的样本集合为空，不能划分（见算法1中第13行）。

2.2.4 随机森林 (RF)

RF 是由多个决策树组成的集成算法。RF 中不同决策树之间没有关联性。通过组合多个决策树，最终结果投票取均值，使算法的结果具有较高的精度和泛化性能。RF 有两个关键特性，“随机”和“森林”。“随机”使 RF 具备高抗过拟合能力。“森林”将多个决策树组合在一起，这是 RF 拟合能力强大的根本原因。

算法 2 RF 基本算法

Input: 数据集 $D = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m)\}$; 特征集 $A = \{a_1, a_2, \dots, a_d\}$.

```

1: procedure RF( $D, A$ )
2:   随机抽样，训练决策树；
3:   while 节点可以分裂 do
4:     随机选取特征，做节点分裂特征；return
5:   end while
6:   建立大量决策树，形成森林。
7: end procedure
```

Output: 输出最优随机森林 $\text{RF}(D, A)$.

构造 RF 需要四步（见算法2），详细过程可描述为：

- (1) 样本容量为 N ，从中有放回地抽取 N 次，每次抽取 1 个样本，最终形成了 N 个样本。将选择的 N 个样本作为决策树根节点处的样本，用来训练一棵决策树。这个过程也被称作自助采样法（Bootstrap）；
- (2) 每个样本有 M 个特征，在节点需要分裂时，随机从这 M 个特征中选取出 m 个特征，满足条件 $m << M$ 。然后从这 m 个特征中采用某种策略（比如说信息增益）来选择 1 个特征作为该节点的分裂特征；
- (3) 决策树形成过程中每个节点都要按照步骤 2 进行分裂。如果下一次该节点选出来的特征是其父节点分裂时用过的特征，则该节点就成了叶节点，无须继续分裂。此时形成了决策树；

(4) 按照步骤 1-3 建立决策树，通过投票或取均值取得最终结果，这样就构建了 RF。

2.2.5 极端随机森林回归（ETR）

ETR 同样是一种由多棵决策树集成的学习器。对比 RF，主要有两点不同：

- 对于每棵决策树的训练集，RF 采用 Bootstrap 采样来选择样本集，将其作为每棵决策树的训练集。而 ETR 中每棵决策树采用原始训练集；
- 在选定了划分特征后，RF 中的决策树会基于信息增益、基尼系数、均方差等原则，选择一个最优的特征值划分点，这和传统的决策树相同。但 ETR 会随机选择一个特征值来划分决策树。

从第二点可以看出，由于随机选择了特征值的划分点，而不是最优点，因此 ETR 中决策树的规模一般大于 RF。也就是说，ETR 的方差相对于 RF 进一步减小，但是偏差相对于 RF 进一步增大。一般情况下，ETR 在精度方面要优于 RF。

2.2.6 梯度提升回归树（GBRT）

GBRT 是另一种决策树集成方法。与 RF 方法不同，GBRT 采用连续的方式构造树，每棵树都试图纠正前一棵树的错误。默认情况下，GBRT 中处理特征值时没有使用随机化，而是用到了强预剪枝。GBRT 使用深度很小（1 到 5 之间）的树，这样算法占用的内存更少，预测速度也更快。

GBRT 背后的主要思想是合并许多简单的算法，比如深度较小的树。每棵树只能对部分数据做出更好的预测，因此可以通过添加不同的树来提高 GBRT 的性能。除了预剪枝与决策树的数量之外，GBRT 的另一个重要参数是学习率，它用于控制每棵树纠正前一棵树错误的强度。较高的学习率意味着每棵树都可以做出较强的修正。通过向 GBRT 中添加许多不同的树，可以增加算法的复杂度，让算法有更多的机会纠正错误。

2.2.7 支持向量回归（SVR）

SVR 是机器学习中经典的监督学习算法之一。同线性回归的目标一样，SVR 学习的目标是使 $f(\mathbf{X})$ 与目标值 \mathbf{Y} 之间的偏差尽可能地小。传统回归算法计算损失通常直接计算算法输出 $f(\mathbf{X})$ 与目标值 \mathbf{Y} 之间的距离，当且仅当 $f(\mathbf{X}) = \mathbf{Y}$ 时，误差才为 0。在很多现实情况中，一定范围内的误差可以被容忍。而 SVR 正是基于这一点：假设 $f(\mathbf{X}) - \mathbf{Y} \leq \epsilon$ ，若 $f(\mathbf{X})$ 与 \mathbf{Y} 之间的距离大于 ϵ 时，才会计

算损失。SVR 问题可形式化为

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{b}} \frac{1}{2} \|\mathbf{W}\|^2 + C \sum_{i=1}^m \ell_\epsilon(f(\mathbf{x}_i) - \mathbf{y}_i), \\ s.t. |f(\mathbf{x}) - (\mathbf{W}^T \mathbf{x} + \mathbf{b})| \leq \epsilon. \end{aligned} \quad \dots (2.8)$$

$$\ell_\epsilon(z) = \begin{cases} 0, & \text{if } |z| \leq \epsilon; \\ |z| - \epsilon, & \text{otherwise.} \end{cases} \quad \dots (2.9)$$

其中 C 为正则化常数， ℓ_ϵ 为 ϵ -不敏感损失函数。

这里引入两个松弛变量 ε_i 和 $\hat{\varepsilon}_i$ ，式2.8可改写为

$$\begin{aligned} \min_{\mathbf{W}, \mathbf{b}, \varepsilon_i, \hat{\varepsilon}_i} \frac{1}{2} \|\mathbf{W}\|^2 + C \sum_{i=1}^m (\varepsilon_i + \hat{\varepsilon}_i), \quad s.t. \begin{cases} |f(\mathbf{x}_i) - (\mathbf{W}_i^T \mathbf{x}_i + b_i)| \leq \varepsilon_i + \hat{\varepsilon}_i, \\ \varepsilon_i \geq 0, \hat{\varepsilon}_i \geq 0, i = 1, 2, \dots, m. \end{cases} \end{aligned} \quad \dots (2.10)$$

再引入拉格朗日乘子 $\mu_i, \hat{\mu}_i, \alpha_i, \hat{\alpha}_i \geq 0$ ，从而得到拉格朗日函数

$$\begin{aligned} L(\mathbf{W}, \mathbf{b}, \epsilon, \hat{\epsilon}, \mu, \hat{\mu}, \alpha, \hat{\alpha}) = & \frac{1}{2} \|\mathbf{W}\|^2 + C \sum_{i=1}^m (\varepsilon_i + \hat{\varepsilon}_i) - \sum_{i=1}^m (\mu_i \varepsilon_i + \hat{\mu}_i \hat{\varepsilon}_i) \\ & + \sum_{i=1}^m \alpha_i (f(\mathbf{x}_i) - \mathbf{y}_i - (\varepsilon_i + \hat{\varepsilon}_i)) + \sum_{i=1}^m \hat{\alpha}_i (f(\mathbf{x}_i) - \mathbf{y}_i - (\varepsilon_i + \hat{\varepsilon}_i)). \end{aligned} \quad \dots (2.11)$$

对式2.11分别求 $\mathbf{W}, \mathbf{b}, \varepsilon_i, \hat{\varepsilon}_i$ 的偏导，并将结果整合到式2.11中，得到 SVR 的对偶问题：

$$\begin{aligned} \max_{\alpha, \hat{\alpha}} \sum_{i=1}^m [\mathbf{y}_i (\hat{\alpha}_i - \alpha_i) - \epsilon (\hat{\alpha}_i + \alpha_i)] - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) \mathbf{x}_i^T \mathbf{x}_j, \\ s.t. \sum_{j=1}^m (\hat{\alpha}_j - \alpha_j) = 0, \quad \alpha_i, \hat{\alpha}_i \in [0, C]. \end{aligned} \quad \dots (2.12)$$

以上过程需要满足 Karush-Kuhn-Tucker (KKT) 条件，即

$$\begin{cases} \alpha_i (f(\mathbf{x}_i) - \mathbf{y}_i - \epsilon_i - \hat{\varepsilon}_i) = 0, \\ \hat{\alpha}_i (f(\mathbf{x}_i) - \mathbf{y}_i - \epsilon_i - \hat{\varepsilon}_i) = 0, \\ \alpha_i \hat{\alpha}_i = 0, \quad \epsilon_i \hat{\varepsilon}_i = 0, \\ (C - \alpha_i) \hat{\varepsilon}_i = 0, \quad (C - \alpha_i) \hat{\alpha}_i = 0. \end{cases} \quad \dots (2.13)$$

可以看出,当且仅当 $f(\mathbf{x}_i - \mathbf{y}_i - \epsilon_i - \hat{\epsilon}_i) = 0$ 时, $\alpha_i = 0$; 当且仅当 $f(\mathbf{x}_i - \mathbf{y}_i - \epsilon_i - \hat{\epsilon}_i) = 0$ 时, $\hat{\alpha}_i = 0$ 。也就是说, 在样本 $(\mathbf{x}_y, \mathbf{y}_i)$ 不在 ϵ 间隔带中, $\alpha_i = 0$ 和 $\hat{\alpha}_i = 0$ 才能取非零值。此外, $f(\mathbf{x}_i - \mathbf{y}_i - \epsilon_i - \hat{\epsilon}_i) = 0$ 和 $f(\mathbf{x}_i - \mathbf{y}_i - \epsilon_i - \hat{\epsilon}_i) = 0$ 两者只能有一个成立, 因此 α_i 和 $\hat{\alpha}_i$ 至少有一个为零。

为了增强 SVR 的非线性特征, 可将 \mathbf{X} 使用核函数映射到高维空间。

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j). \quad \dots (2.14)$$

式2.12可转化为

$$\begin{aligned} \max_{\alpha, \hat{\alpha}} \sum_{i=1}^m & [\mathbf{y}_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i)] - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m (\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j) \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j), \\ \text{s.t. } & \sum_{j=1}^m (\hat{\alpha}_i - \alpha_i) = 0, \quad \alpha_i, \hat{\alpha}_i \in [0, C]. \end{aligned} \quad \dots (2.15)$$

在非线性情况下, 最优问题在特征空间(而不是输入空间)中函数需要满足可微条件。最终 SVR 可表示为

$$f(\mathbf{X}) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) + \mathbf{b}. \quad \dots (2.16)$$

本论文选择了线性核函数, 其数学表达式如下:

$$\mathbf{k}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j. \quad \dots (2.17)$$

2.2.8 卷积神经网络 (CNN)

CNN 是机器学习理论中发展最迅速的一个子领域。目前在图像识别、自动驾驶等领域都使用了二维卷积神经网络 (Two Dimensional Convolutional Neural Network, 简称 2DCNN)。而一维卷积神经网络 (One Dimensional Convolutional Neural Network, 简称 1DCNN) 擅长于自然语言处理、信号处理等时间序列数据。一般来讲, CNN 有两大特点:

- 能有效地将较大的数据压缩为较小的形状;
- 在压缩形状的同时, 还能有效地保留图片特征。

经典的 CNN 由三部分组成:

- **卷积层**。卷积层负责提取图像中的局部特征。卷积层的运算过程可理解为: 使用卷积核来过滤原始数据中的部分小区域, 从而得到这些小区域的特征值。在

具体应用中，卷积核可以有多个，每个卷积核代表了一种特征模式。如果某个区域与此卷积核的加权和较大，则该区块与该卷积核十分接近。可以说，卷积层通过卷积核提取小区域中的特征。

- **池化层。**池化层，也称作下采样层，用来大幅度降低参数量级，即降维。常见的池化运算包括最大池化和平均池化。加入池化层，是因为当仅仅加入卷积层时，所得到的特征图的维度可能依旧很大。与卷积层相比，池化层能够更加有效地降低数据维度，大大减小运算量，还可以缓解过拟合。

- **全连接层。**全连接层是传统神经网络的层，用来输出最终结果。CNN 的输出层多数情况下为全连接层。经过卷积层和池化层处理过的数据输出到全连接层，可以得到最终的结果。如果输入数据仅仅使用全连接层，而没有使用卷积层和池化层，全连接层会面临着学习参数太多、计算成本高、计算效率低下等问题。

CNN 通过在输入层和输出层之间使用隐藏层，可找到数据的中间表征。CNN 中的卷积层、池化层、dropout 等方法可有效减小数据维度。使用 CNN 的前提是具备较多的数据、较多需要估算的参数、强大的计算能力等。实际应用过程中，尤其是在小样本情况下，深度学习的性能不一定优于传统意义上的机器学习。

2.2.9 长短期记忆循环神经网络（LSTM-RNN）

第2.2.8节提到，1DCNN 能处理时序问题。除了 1DCNN，ANNs 中还有另一种处理时间序列的神经网络，即循环神经网络（Recurrent Neural Network，简称 RNN）。RNN 将历史时间步神经元的输出引入到当前时间步神经元的输入中，而当前时间步神经元的输入影响着未来时间步神经元的输出。传统的 RNN 很容易出现梯度消失现象，这是因为梯度连乘导致参数更新非常缓慢。

LSTM-RNN 是一种特殊的 RNN，由Hochreiter 等 (1997) 提出。LSTM-RNN 通过引入历史时间步的神经元状态并更新当前时间步神经元隐藏层的输出，将不需要的信息去掉，解决了梯度消失和梯度爆炸问题。RNN 只更新历史一个时间步的状态，而 LSTM-RNN 可以学习到什么时候以及多长时间遗忘和保存某些信息。LSTM-RNN 被广泛应用于文本生成、语音识别、机器翻译、图像描述和视频标记等。图2.1绘制了两层的 LSTM-RNN。

为了展示 LSTM-RNN 如何工作，将图2.1中的 LSTM 神经元展开，从而得到 LSTM 神经元结构图（见图2.2）。图2.2中， x_t 为在时间步为 t 时的输入信

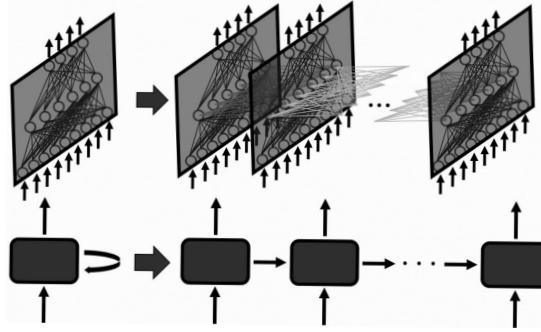


图 2.1 两层 LSTM-RNN 的示例。

Figure 2.1 The example of two-layer LSTM-RNN.

号, \mathbf{h}_t 为时间步为 t 时的隐藏层状态, \mathbf{h}_{t-1} 为时间步为 $t-1$ 时的隐藏层状态。
 $\mathbf{g}_t = \tanh(\mathbf{U}_g \mathbf{x}_t + \mathbf{V}_g \mathbf{h}_{t-1} + \mathbf{b}_g)$ 。 \tanh 为双曲正切函数, i_t 为输入门, f_t 为遗忘门,
 \mathbf{o}_t 为输出门。 \mathbf{c}_t 表示时间步为 t 时单元状态, \mathbf{c}_{t-1} 表示时间步为 $t-1$ 时单元状态。 \odot 表示元素相乘。黑色方块代表循环阶段。这些神经元具有各种组件, 分别
为具有记忆单元的输入门、自循环连接的神经元、遗忘门和输出门。记忆单元类
似于一个累加器, 可以学习序列中的长期依赖关系, 从而使优化变得更加容易。
同时, 每个单元格由三个乘法单元控制, 即输入门、输出门和遗忘门, 这些决定
了是忘记过去单元状态还是将输出传递到之后的状态, 从而使 LSTM 神经元能
够长期存储和访问信息。

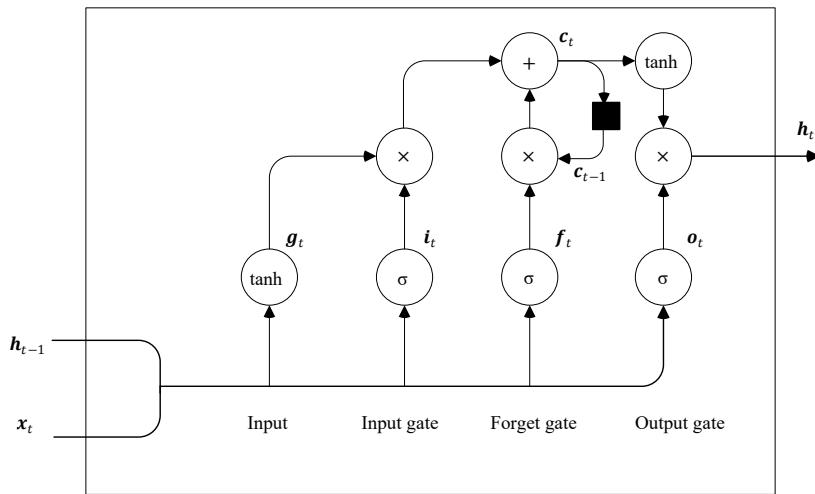


图 2.2 LSTM 神经元示意图。

Figure 2.2 An LSTM cell diagram.

LSTM 神经元具体工作机制如下:

(1) 首先将序列值 \mathbf{x}_t 与神经元 \mathbf{h}_{t-1} 作为先前输出，该组合输入的第一步会通过 $\tanh(\cdot)$ 层进行压缩，即 $\mathbf{g}_t = \tanh(\mathbf{U}_g \mathbf{x}_t + \mathbf{V}_g \mathbf{h}_{t-1} + \mathbf{b}_g)$ ，其中 \mathbf{U}_g 和 \mathbf{V}_g 分别代表输入和前一个神经元的输出的权重， \mathbf{b}_g 表示输入的偏置。 $\mathbf{g}_t \in (-1, 1)$ 。第二步此组合输入通过输入门。输入门是一层 $\sigma(\cdot)$ 激活节点，其输出乘以压缩的输入。这些输入门 $\sigma(\cdot)$ 可“杀死”输入向量中不需要的任何元素。 $\sigma(\cdot)$ 函数的输出值 $\mathbf{i}_t = \sigma(\mathbf{U}_i \mathbf{x}_t + \mathbf{V}_i \mathbf{h}_{t-1} + \mathbf{b}_i)$ ， $\mathbf{i}_t \in (0, 1)$ 。其中 \mathbf{U}_i 、 \mathbf{V}_i 和 \mathbf{b}_i 均是需要学习的参数， $\mathbf{i}_t \in (0, 1)$ 。将输入连接到这些节点的权重以输出接近于零的值，从而“关闭”某些输入值；或者相反，输出接近于 1 以“记住”其他值。LSTM 神经元输入部分的输出可以由 $\mathbf{g}_t \odot \mathbf{i}_t$ 得到，其中 \odot 表示元素相乘。

(2) 通过这个单元的数据流的下一步是遗忘循环门，可表示为 $\mathbf{f}_t = \sigma(\mathbf{U}_f \mathbf{x}_t + \mathbf{V}_f \mathbf{h}_{t-1} + \mathbf{b}_f)$ 。其中 $\mathbf{f}_t \in (0, 1)$ ， \mathbf{U}_f 、 \mathbf{V}_f 和 \mathbf{b}_f 分别为两个的权重参数和一个偏置参数。LSTM 神经元有一个内部状态变量 \mathbf{c}_t 。这个变量滞后一个时间步，即 \mathbf{c}_{t-1} 被添加到输入数据中以创建一个有效的循环层，即 $\mathbf{c}_{t-1} \odot \mathbf{f}_t$ 。 \mathbf{c}_t 可表达为 $\mathbf{c}_{t-1} \odot \mathbf{f}_t + \mathbf{g}_t \odot \mathbf{i}_t$ ，这种加法运算而不是乘法运算有助于降低梯度消失的风险。然而，这个循环是由遗忘门控制的，它的工作原理与输入门相同，有助于网络学习哪些状态变量应该被“记住”或“忘记”。

(3) 最后，获得一个输出层，由 $\tanh(\cdot)$ 压缩函数控制。这个门决定了哪些值被允许作为神经元 \mathbf{h}_t 的输出 \mathbf{o}_t 。 $\mathbf{o}_t = \sigma(\mathbf{U}_o \mathbf{x}_t + \mathbf{V}_o \mathbf{h}_{t-1} + \mathbf{b}_o)$ ， $\mathbf{o}_t \in (0, 1)$ 。 \mathbf{U}_o 、 \mathbf{V}_o 和 \mathbf{b}_o 为输出门中一系列可学习的参数。新的隐藏层 \mathbf{h}_t 可通过 $\tanh(\mathbf{c}_t) \mathbf{o}_t$ 计算得到。最后一层的输出 \mathbf{h}_n 会连接到一个传统的全连接神经网络，表达为 $\mathbf{y} = \mathbf{W}_d \mathbf{h}_n + \mathbf{b}_d$ 。其中 \mathbf{y} 作为神经网络的输出， \mathbf{W}_d 和 \mathbf{b}_d 分别为权重和偏置。

总体来看，整个 LSTM-RNN 的运行过程可描述如下。首先，将一系列的时间序列数据 $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\}$ 作为输入，其中 \mathbf{x}_t 是时间步 t 的输入数据。在堆叠的 LSTM 层中，下一层接收上一层的输出 $\mathbf{h} = \{\mathbf{h}_1, \mathbf{h}_2, \dots, \mathbf{h}_n\}$ ， \mathbf{h} 作为下一层的输入。最后一层使用全连接层。

2.3 训练前的必要准备

开展机器学习，首先需要将收集的资料进行整理分析，得到满足目标任务的特征因子。利用滑动窗口法将这些特征因子转化为监督学习数据集。然后，对监督学习数据集进行划分和归一化处理。有时，数据集划分和归一化处理不分前后

次序。在优化算法时，需要选择合适的性能度量指标、优化目标函数、优化器等。对算法输出结果进行分析时，还会涉及到算法是否出现过拟合或欠拟合的问题。

2.3.1 滑动窗口法

本论文中选择的数据集（太阳黑子活动、泉流量、地震数值预测）都是时间序列数据。历史时间序列数据 $\{x_{t-M+1}, x_{t-M+2}, \dots, x_t\}$ 作为预测下 N 个时间步 $\{x_{t+1}, x_{t+2}, \dots, x_{t+N}\}$ 的输入。需要将原始数据集转化为监督学习数据集。监督学习数据集的生成可采用滑动窗口法。图2.3绘制了滑动窗口法的过程。其中，将时间序列中输入时间窗口 M 个观测值作为输入，输出时间窗口 N 个观测值作为输出。将窗口一次滑动一个时间步，对整个原始数据集重复此过程，最终可获得监督学习数据集。

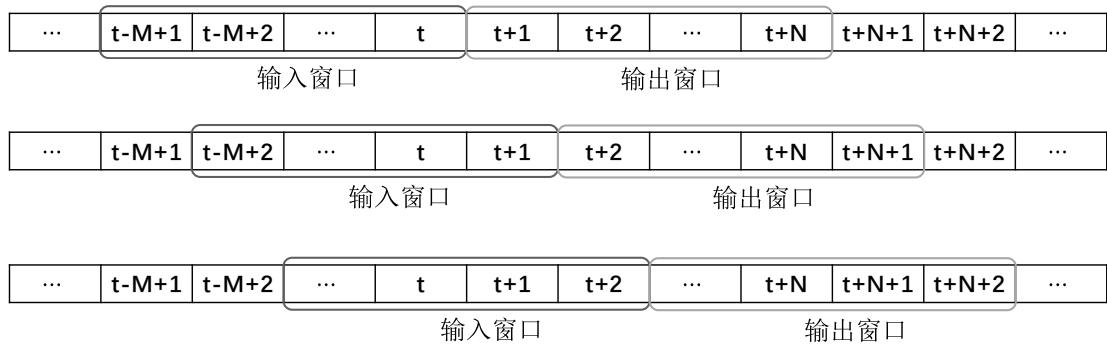


图 2.3 滑动窗口法。每一行是时间序列数据集。方框中每个数字代表一个时间步。输入窗口和输出窗口在整个可用的数据集中按月/年滑动。

Figure 2.3 The sliding window method. Each row is time series dataset. Each number in the box represents a time step. The input and output windows monthly/yearly slide based on entire available dataset.

2.3.2 数据归一化

将数据喂给神经网络之前，需要对其进行归一化（Normalization）处理。目的是使预处理的数据被限定在一定范围内，提高算法精度，并加速算法的收敛速度。目前有 2 种最常用的归一化方法，分别为：

- **线性函数归一化 (Min-Max Scaling)**。该方法是一种线性转换的方法，归一化后的数据值被映射到 $[0, 1]$ 之间。转化公式如式下：

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \quad \dots (2.18)$$

其中， \mathbf{X}_{norm} 是归一化后的数据， \mathbf{X} 是原始观测数据， \mathbf{X}_{min} 是观测数据中的最小值， \mathbf{X}_{max} 是观测数据中的最大值。

- **0 均值归一化** (Z-Score Standardization)。0 均值归一化方法将原始观测数据集归一化为均值为 0、方差 1 的数据集，归一化公式如下：

$$\mathbf{X}_{\text{norm}} = \frac{\mathbf{X} - \mu}{\sigma}. \quad \dots (2.19)$$

其中， μ 、 σ 分别为原始数据集的均值和方差。该种归一化方式要求原始数据的分布可以被近似为高斯分布。

因本论文所涉及的几种数据集都不满足高斯分布特征，因此统一采用线性函数归一化法。

2.3.3 优化器

在寻找最优参数（权重和偏置）的过程中，要使目标函数尽可能地小。这个过程需要计算参数的导数（即梯度），以梯度为指引方向逐步更新参数的值。常见的优化算法如下：

- **随机梯度下降法** (Stochastic Gradient Descent, 简称 SGD)。SGD 计算过程中基于单个样本，即每读一个数据，则立刻计算代价函数的梯度。

- **批量梯度下降** (Batch Gradient Descent, 简称 BGD)。BGD 在计算过程中基于整个数据集，即读取所有数据后才会计算代价函数的梯度。

- **小批量梯度下降** (Mini-Batch Gradient Descent, 简称 Mini-BGD)。Mini-BGD 选择小批量数据更新梯度。SGD、BGD 和 Mini-BGD 三类方法都存在一个问题，即更新方向完全依赖于梯度，容易陷入局部最优。

- **动量** (Momentum)。Momentum 引入了动量的物理概念，使得未来梯度方向与历史关联起来，从而跳出了局部最优。

- **自适应梯度** (Adaptive Gradient, 简称 Adagrad)。Adagrad 会累加历史所有梯度的平方。

- **均方根反向传播** (Root Mean Square Propagation, 简称 RMSprop)。MSProp 计算所有梯度的平均值。Adagrad 和 RMSprop 都属于自适应学习率方法，训练过程中的步长（即学习率）时刻在调整。

- **Adam** (Adaptive Moment Estimation)。Adam 融合了 RMSProp 和 Adagrad，通过组合两者的优点，实现参数空间的高效搜索。

本论文在训练神经网络时使用了 Adam 优化算法，它由 Kingma 等 (2014) 提出。Adam 简单高效，不需要消耗过多内存，尤其适用于具有大量的数据和很多参数的问题中。Adam 第一分量 (2.20) 和第二分量 (2.21) 分别表示为：

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t. \quad \dots (2.20)$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2. \quad \dots (2.21)$$

其中， m_t 和 v_t 分别是第一分量和第二分量的估计值。 m_t 和 v_t 均被初始化为 0。当衰减值很小时，梯度会出现偏差。利用式2.22和式2.23，通过修正偏置来解决偏差问题。

$$\hat{m}_t = \frac{m_t}{1 - \beta_1^t}. \quad \dots (2.22)$$

$$\hat{v}_t = \frac{v_t}{1 - \beta_2^t}. \quad \dots (2.23)$$

其中， β_1 默认值为 0.9， β_2 默认值为 0.999。在修正了偏置后，权重也得以修正（见式2.24）。

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t. \quad \dots (2.24)$$

其中， ϵ 默认值为 10^{-8} 。

2.3.4 优化目标函数

在学习过程中，算法会通过某个指标寻找最优权重参数，这个指标为代价函数 (Cost Function)。代价函数能够很直观地呈现数据集的拟合程度。训练算法的过程，也就是最小化预测值与目标值差距的过程，或者说最小化代价函数的过程。这里需要仔细区分损失函数 (Loss Function)、代价函数和目标函数：

(a) **损失函数**。损失函数针对单个样本，计算的是单个样本的误差，损失函数可表述为 $L(\mathbf{y}_i - f(\mathbf{x}_i))$ 。通常损失函数有以下几种：

- 绝对损失函数：

$$L(\mathbf{y}_i, f(\mathbf{x}_i)) = |\mathbf{y}_i - f(\mathbf{x}_i)|. \quad \dots (2.25)$$

- 平方损失函数：

$$L(\mathbf{y}_i, f(\mathbf{x}_i)) = (\mathbf{y}_i - f(\mathbf{x}_i))^2. \quad \dots (2.26)$$

◦ 0-1 损失函数：

$$L(\mathbf{y}_i, f(\mathbf{x}_i)) = \begin{cases} 0, & \mathbf{y}_i = f(\mathbf{x}_i) \\ 1, & \mathbf{y}_i \neq f(\mathbf{x}_i) \end{cases}. \quad \dots (2.27)$$

◦ 交叉熵损失函数：

$$L(\mathbf{y}_i, f(\mathbf{x}_i)) = -\mathbf{y}_i \log f(\mathbf{x}_i). \quad \dots (2.28)$$

通常，式2.25和式2.26被用在回归问题上，式2.25被用在二分类问题上，式2.28被用在多分类问题上。

(b) 代价函数。与损失函数不同的是，代价函数针对的是整个训练样本，是所有样本误差的平均值，也就是损失函数的平均值，一般被称作经验风险最小化函数，表达式为

$$J = \frac{1}{N} \sum_{i=1}^N L(\mathbf{y}_i, f(\mathbf{x}_i)). \quad \dots (2.29)$$

但经验风险过小在多数情况下会出现过拟合，而过拟合的原因是 $f(\mathbf{X})$ 太过复杂。

(c) 目标函数。对于小样本，用经验风险估计期望经常不太理想，需要对其进行矫正。结构风险最小化是为了抗过拟合而提出的策略。结构风险在经验风险上加入了表示算法复杂度的正则化项 $\lambda J(f)$ ，因此结构方程为

$$J_\lambda = \frac{1}{N} \sum_{i=1}^N L(\mathbf{y}_i, f(\mathbf{x}_i)) + \lambda J(f). \quad \dots (2.30)$$

最终的优化函数为 $\min J_\lambda$ ，此函数又被称作目标函数。

2.3.5 性能度量

机器学习中，通常将数据集划分为训练集和测试集，还假设这两种数据集同分布。一般在训练集上构建算法，寻找最优算法，然后使用测试集验证算法的实际预测能力。如果一个算法能够对未见样本（即测试集）做出准确预测，则认为该算法能够从训练集泛化到测试集。泛化能力指预测测试集的能力，获得较强的泛化能力是机器学习的最终目标。

如果算法只能正确识别已有的训练集，而预测测试集时出现很大的偏差，那有可能只学习到了训练集中的无关紧要的细节（如噪声）。需要注意的是，仅仅

用训练集去学习和评价参数，常常会导致算法只可以用来处理某种数据集，对于其他数据集难以正确处理。

对算法的泛化性能进行评估，需要衡量算法泛化能力的评价标准，即性能度量（Performance Measure）。通常，使用不同的性能度量指标，评判结果会有差异。学习的最终目标是，预测结果与真实值尽可能地接近。性能度量指标有：

- 均方误差（Mean Squared Error，简称 MSE）。数学公式为：

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (f(\mathbf{x}_i) - \mathbf{y}_i)^2. \quad \dots (2.31)$$

- 绝对误差（Mean Absolute Error，简称 MAE）。数学公式为：

$$\text{MAE} = \frac{1}{m} \sum_{i=1}^m |f(\mathbf{x}_i) - \mathbf{y}_i|. \quad \dots (2.32)$$

- 均方根误差（Root Mean Absolute Error，简称 RMSE）。数学公式为：

$$\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m (f(\mathbf{x}_i) - \mathbf{y}_i)^2}. \quad \dots (2.33)$$

• 平均绝对百分比误差（Mean Absolute Percentage Error，简称 MAPE）。数学公式为

$$\text{MAPE} = \frac{100\%}{m} \sum_{i=1}^m \left| \frac{f(\mathbf{x})_i - \mathbf{y}_i}{\mathbf{y}_i} \right|. \quad \dots (2.34)$$

- 交叉熵误差（Cross Entropy Error，简称 CEE）。数学公式为：

$$\text{CEE} = - \sum_{i=1}^m (\mathbf{y}_i \log f(\mathbf{x}_i)). \quad \dots (2.35)$$

- 准确度（Accuracy）。数学公式为：

$$\text{Accuracy} = \frac{1}{m} \sum_{i=1}^m (\mathbf{y}_i == f(\mathbf{x}_i)). \quad \dots (2.36)$$

在处理回归问题时，本论文采用 MSE 作为损失函数，RMSE 作为预测的误差。

2.3.6 欠拟合和过拟合

欠拟合（Under-Fitting）也称欠学习，直观表现是训练得到的算法在训练集上表现差，未学到数据的规律。引起欠拟合的原因可能有：

- 算法本身过于简单。例如，数据本身是非线性的，但使用了线性算法；

- 特征数太少，无法正确的建立统计关系。

过拟合（Over-Fitting）也称过学习，它的直观表现是算法在训练集上表现良好，但在测试集上表现不好，泛化性能差。过拟合是在算法参数拟合过程中由于训练数据包含抽样误差，在训练时复杂的算法将抽样误差也进行了拟合。所谓抽样误差，是指抽样得到的样本集和整体数据集之间的偏差。直观来看，引起过拟合可能的原因有：

- 算法本身过于复杂，以至于拟合了训练样本集中的噪声。需要选用更简单的算法，或者对算法进行简化；
- 训练样本太少或缺乏代表性。需要增加样本数或增加样本的多样性；
- 算法拟合了噪声。需要剔除噪声或者改用对噪声不敏感的算法。

过拟合是有监督机器学习算法长期以来需要面临的一个问题。表2.1给出了实际应用时判断过拟合与欠拟合的准则。

表 2.1 过拟合与欠拟合的判断标准。

Table 2.1 The criterion for under-fitting and over-fitting.

训练集上的表现	测试集上的表现	判断
欠佳	欠佳	欠拟合
良好	欠佳	过拟合
良好	良好	适度拟合

还可以从偏差和方差的角度来看待欠拟合和过拟合。算法的泛化误差来源于三部分，即偏差（Bias）、方差（Variance）和噪声（Noise）。偏差度量了学习算法的期望预测值与真实值的偏离程度，即刻画了学习算法本身的拟合能力。假设输入特征向量为 \mathbf{X} ，输出为 \mathbf{Y} ，实际输出为 \mathbf{Y}' ，要拟合的目标函数为 $f(\mathbf{X})$ ，训练的函数为 $\hat{f}(\mathbf{X})$ ，则偏差为：

$$\text{Bais}^2(\mathbf{X}) = (\hat{f}(\mathbf{X}) - \mathbf{Y})^2. \quad \dots (2.37)$$

高偏差意味着算法本身的输出值与期望值差距大，导致欠拟合。

与偏差不同，方差度量了训练集的变动影响算法性能的程度，即刻画了数据扰动所造成的影响。它是对训练样本的小波动敏感而导致的误差。方差可以理解为算法的波动程度。根据概率论中方差的定义：

$$\text{Variance}(\mathbf{X}) = E[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2]. \quad \dots (2.38)$$

高方差意味着算法本身与期望的算法差距大，从而导致过拟合。

除了偏差与方差，噪声也会影响到算法性能的好坏。噪声表达了在当前任务上任何学习算法所能达到的期望泛化误差的下界，即刻画了学习问题本身的难度。期望预测值与真实值之间的误差称为噪声，即

$$\text{Noise}(\mathbf{X}) = E[(Y' - Y)^2]. \quad \dots (2.39)$$

当存在噪声时，复杂的算法会尽量覆盖噪声，即产生了过拟合。这样，即使训练误差很小，由于没有描绘真实的数据趋势，测试误差反而更大。还有一种情况，如果数据是由未知的非常复杂的系统产生的，实际上有限的数据很难去“代表”这个复杂系统。采用不恰当的数据集拟合，算法性能会很差，因为部分数据在不恰当的复杂假设下就像是“噪声”，从而产生过拟合。

综上，泛化误差可分解为偏差、方差和噪声，即

$$\text{Error}(\mathbf{X}) = \text{Bias}^2(\mathbf{X}) + \text{Variance}(\mathbf{X}) + \text{Noise}(\mathbf{X}). \quad \dots (2.40)$$

以上分解说明，泛化性能是由学习算法的能力、数据的充分性以及学习任务本身的难度共同决定。一般来讲，噪声难以避免，更难以被剔除。为了获得更好的泛化性能，需要同时减小偏差（能充分拟合数据）和方差（数据扰动产生的影响小）。

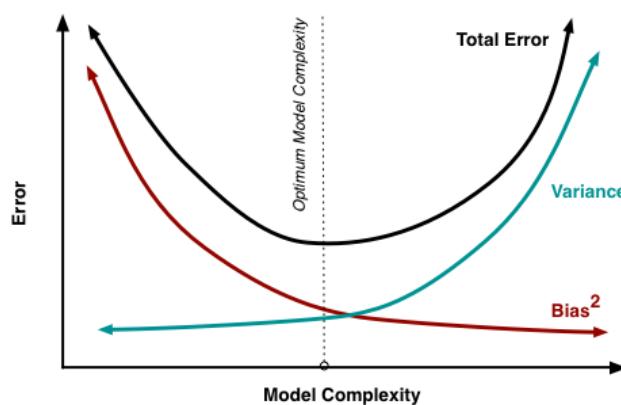


图 2.4 AIC 信息准则。

Figure 2.4 Akaike information criterion.

但是偏差和方差是有冲突的。图2.4给出了AIC信息准则(Akaike Information Criterion)，又称赤池信息准则，表达了泛化误差与偏差、方差的关系。它建立

在熵的概念基础上，可以权衡所估计算法的复杂度和此算法拟合数据的优良性。当算法训练不足时，训练数据的特征没有被很好地拟合，会产生较大偏差和较小方差。当算法得到进一步训练时，偏差会逐渐减小，而方差会逐渐增大。当算法被充分训练之后，学习器的拟合能力非常强，会出现较小的偏差和较大的方差，算法会将训练数据中的所有特征（包括噪声）都学习到，此时就会发生过拟合。

学习率和批量处理大小（Batch-Size）都是影响算法优化效率和泛化能力的超参数。在实际训练过程中，会观察算法每一次更新过程中测试集误差的变化趋势。当测试集误差在指定的循环次数范围内未能有进一步改善的空间，算法会被人为终止，这种策略也被称作“早停”（Early Stopping）。这种方法简单有效，是深度学习中最常见的正则化策略之一。

2.3.7 计算工具与平台

论文中涉及的数据分析全部基于 Python 中的开源库编码实现。开源库主要包括 Tensorflow2.0、Scikit-learn、Numpy、Pandas、Matplotlib、Julian 等，其中 Tensorflow2.0 用来构建神经网络算法，Scikit-learn 用来构建常用的机器学习算法，Numpy 和 Pandas 主要针对数据进行预处理，Matplotlib 则是在可视化结果时使用，Julian 将数据中的序列数据转化为日期。论文计算任务在本地服务器（1×GPU1080Ti）上完成。

2.4 小结

本章详细介绍了机器学习的理论基础以及相关优化的概念和方法。第2.2节介绍了不同的机器学习算法的基础理论，包括 LR、KNN、DT、RF、SVR、GBRT、ETR、CNN、LSTM-RNN。第2.3节说明了算法需要的前期准备，包括基于滑动窗口法获取监督学习数据集、归一化、优化器和优化函数的选择、性能度量标准、欠拟合和过拟合等。

第3章 基于神经网络预测太阳黑子活动

要进行机器学习，先要有数据集。数据集由很多样本组成。在研究太阳黑子活动时，只考虑了历史太阳黑子活动与未来太阳黑子活动的关联性。因此模型的输入和输出都来自同一种特征。本论文从一种特征（太阳黑子活动）出发，利用机器学习探索在不同的输入和输出时间窗口长度下算法的性能。

3.1 研究背景

太阳活动是太阳大气中局部区域各种不同活动现象的总称，包括太阳黑子、日珥、光斑、谱斑、耀斑、太阳风和日冕质量抛射等。处于活动剧烈期的太阳会辐射出大量紫外线、X射线、粒子流和强射电波，导致地球上出现极光、磁暴和电离层扰动等现象。太阳活动对地球的影响有以下几个方面 (Jie 等, 2012):

- **地球大气层。** 地球大气层在太阳的辐射下形成电离层，无线电波依靠电离层的反射向远距离传播。当太阳活动剧烈时，在向阳的半球电离层容易衰减甚至中断；
- **地球气候。** 太阳活动周期与地球气候的关系也比较明显，其作用机制并不清楚。太阳活动达到高峰时，地球上太平洋热带及亚热带地区气温升高、海水加速蒸发、西太平洋热带海域的降雨增多，东太平洋热带海域气温降低；
- **地球磁场。** 太阳活动剧烈时，地球磁场会产生磁暴现象；
- **航天活动。** 大耀斑出现时，会射出的高能量质子，对航天飞行、卫星设备、太阳能电池等有极大的破坏性。

太阳黑子是太阳局部强磁场活动在太阳光球层上产生的黑色斑块。Noyes (2013) 详细阐述了太阳黑子的产生机制。太阳黑子位于太阳表面的强磁场区，是太阳表面的炽热气体形成的巨大漩涡，温度高达 3000°C - 4500°C 。太阳黑子由太阳大气中的电磁过程引起，时烈时弱，周期为 ~ 11 年。通常太阳黑子活动越剧烈，太阳活动越频繁。太阳黑子可以成群出现，也可以是孤立的本影，周围有半影。

传统意义上，基于半经验性的物理机制模型已经被广泛应用于太阳黑子时间序列预测。然而，这些模型都有一个前提假设，时间序列是从线性过程产生的，

难以有效地抓住非高斯性、非稳态性和非线性的时间序列关系 (Jiang 等, 2011; Arlt 等, 2014)。

近些年来，机器学习中的神经网络在太阳黑子活动预测中的应用非常普遍 (Pala 等, 2019)。使用的神经网络模型愈加复杂，容错能力也逐渐增加。例如，Zhao 等 (2008) 使用径向基神经网络预测未来 4 个月平滑月均太阳黑子数，发现相对误差控制在 38% 以内，而且误差会随着预测时间的延长而逐渐增加；Ding 等 (2012) 基于反向传播神经网络 (Backpropagation Neural Network, 简称 BPNN) 预测平滑月均太阳黑子面积，发现相对误差不超过 5%；Li 等 (2018) 在预测太阳黑子数时，使用了 BPNN 及其变种的神经网络，发现组合的神经网络强于 BPNN，RMSE 为 1.7117，MAPE 为 0.0435。

尽管很多学者利用不同的模型预测第 25 太阳周的峰值，却没有得出统一的结论。对于第 23 太阳周而言，峰值出现时间为 2001 年 9 月，太阳黑子数为 238.2，太阳黑子面积为 2171.7；对于第 24 太阳周而言，峰值出现时间为 2014 年 2 月，太阳黑子数为 146.1，太阳黑子面积为 1439.8。对于预测的第 25 太阳周太阳黑子活动的峰值，可分为以下几种结论：

(1) 与第 24 太阳周相比，第 25 太阳周太阳黑子活动更剧烈。例如，McIntosh 等 (2020) 基于磁活动周期的物理模型，预测到第 25 太阳周是有史以来观察到的最剧烈的太阳黑子活动，峰值为 ~ 233 ；Pesnell (2018) 采用 naïve 方法，得到第 25 太阳周太阳黑子数的峰值为 $\sim 180 \pm 60$ 。

(2) 与第 24 太阳周相比，第 25 太阳周太阳黑子活动基本与之持平。例如，Hiremath (2008) 预测未来 ~ 170 年的太阳黑子活动，发现第 24 和 25 太阳周的太阳黑子数均为 ~ 110 ；Bhowmik 等 (2018) 得到第 25 太阳周太阳黑子数的峰值为 ~ 118 ，出现在 ~ 2024 年；Singh 等 (2019) 得出第 25 太阳周太阳黑子数的峰值为 124 ± 11 ，出现在 2024 年 2 月；Bisoi 等 (2020) 得出第 25 太阳周太阳黑子数峰值为 $\sim 131\text{--}134$ ，出现在 2025 年 7 月。

(3) 与第 24 太阳周相比，第 25 太阳周太阳黑子更弱。例如，Kitiashvili (2020) 发现第 25 太阳周最大太阳黑子数为 ~ 50 （发生在 2024 年或 2025 年）。该数值小于 Dalton 最低点期间的太阳黑子数，可能与 Maunder 最低点相当。

总之，无论最终模型的表现效果如何，目前很难对第 25 太阳周太阳黑子活动的非线性动态特征（太阳黑子峰值和太阳周持续时间）进行精准预测。究其原

因，太阳黑子时间序列具有非平稳性、非高斯性、非线性特征。目前，有关太阳黑子的记录长达 400 多年，太阳黑子时间序列数据显示出周期性震荡。通过机器学习探索太阳黑子活动，有助于理解太阳活动的机制。鉴于预测未来第 25 太阳周太阳黑子活动难以取得一致性的结果，本章尝试预测未来 1 个月和 72 个月的太阳黑子活动，将未来 72 个月太阳黑子活动最剧烈时的数量/面积定义为第 25 太阳周的峰值，探索在什么情况下能够较为精准地预测太阳黑子活动。

本章结构安排如下。第3.2节描述了太阳黑子时间序列和预处理过程。第3.3节描述太阳黑子活动的试验过程，并将试验结果可视化。第3.4节对预测太阳黑子动态变化进行了总结与展望。

3.2 数据与方法

本节首先介绍了太阳黑子数量和面积的时间序列以及两者长期的变化趋势。接着介绍原始数据集进行预处理的过程，包括生成监督学习数据集、数据集划分、归一化处理等。最后选取几种不同结构的神经网络训练处理后的数据。

3.2.1 数据简介

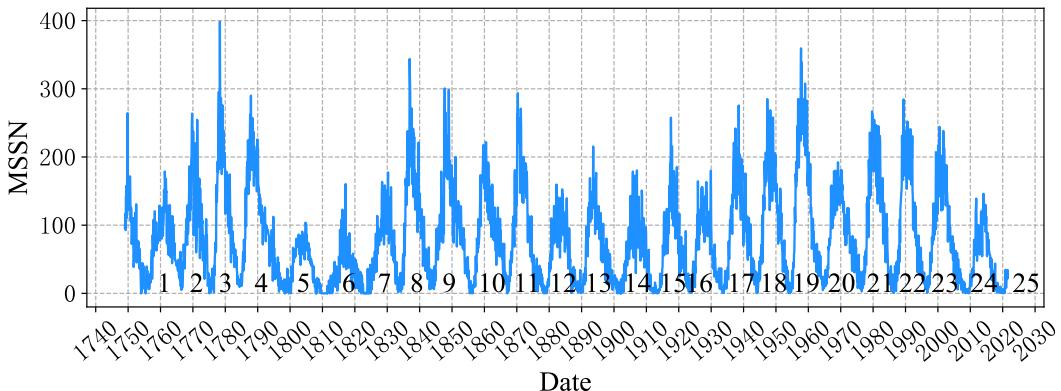


图 3.1 1949 年至 2021 年期间月均太阳黑子数。

Figure 3.1 Monthly mean total sunspot number between 1749 and 2021.

本章采用两种时间序列数据集。第一类为月均太阳黑子数（Monthly Mean Total Sunspot Number，简称 MSSN），来自 SILSO（Sunspot Index and Long-Term Solar Observation）网站¹发布的太阳黑子 2.0 版本。MSSN 数据集跨越了 ~273 年（从 1749 年 1 月至 2021 年 8 月），包含 3272 条记录。截至目前，MSSN 数据集

¹ 数据来源：WDC-SILSO，Royal Observatory of Belgium，Brussels，<http://sidc.be/silso/datafiles>

表 3.1 13 个月平滑的月均太阳黑子数量、峰值和谷值出现的时间。

Table 3.1 Solar minimum and maximum timings and amplitudes of 13-month smoothed monthly mean sunspot numbers.

太阳周	谷值		峰值		时间 (年)		
	日期	黑子数	日期	黑子数	上升	下降	最小值到最小值
1	1755.204	14.0	1761.455	144.1	6.251	5.000	11.251
2	1766.455	18.6	1769.707	193.0	3.252	5.748	9.000
3	1775.455	12.0	1778.371	264.3	2.916	6.337	9.253
4	1784.708	15.9	1788.124	235.3	3.416	10.164	13.580
5	1798.288	5.3	1805.123	82.0	6.835	5.835	12.670
6	1810.958	0.0	1816.373	81.2	5.415	6.998	12.413
7	1823.371	0.2	1829.874	119.2	6.503	4.000	10.503
8	1833.874	12.2	1837.204	244.9	3.330	6.334	9.664
9	1843.538	17.6	1848.124	219.9	4.586	7.834	12.420
10	1855.958	6.0	1860.124	186.2	4.166	7.080	11.246
11	1867.204	9.9	1870.623	234.0	3.419	8.335	11.754
12	1878.958	3.7	1883.958	124.4	5.000	6.246	11.246
13	1890.204	8.3	1894.042	146.5	3.838	8.000	11.838
14	1902.042	4.5	1906.123	107.1	4.081	7.500	11.581
15	1913.623	2.5	1917.623	175.7	4.000	6.000	10.000
16	1923.623	9.4	1928.290	130.2	4.667	5.417	10.084
17	1933.707	5.8	1937.288	198.6	3.581	6.836	10.417
18	1944.124	12.9	1947.371	218.7	3.247	6.917	10.164
19	1954.288	5.1	1958.204	285.0	3.916	6.587	10.503
20	1964.791	14.3	1968.874	156.6	4.083	7.332	11.415
21	1976.206	17.8	1979.958	232.9	3.752	6.749	10.501
22	1986.707	13.5	1989.874	212.5	3.167	6.750	9.917
23	1996.624	11.2	2001.874	180.3	5.250	7.084	12.334
24	2008.958	2.2	2014.288	116.4	5.330	—	—
平均值	9.29		178.7	4.33	6.74	11.03	
标准差	5.70		57.76	1.11	1.24	1.18	
中值	9.65		183.3	4.08	6.75	11.25	

包含了 24 个完整的太阳周。表3.1展示了这 24 个太阳周太阳黑子数量、峰值和谷值出现时间。第 1 太阳周为 1755 年 2 月至 1766 年 5 月，目前正处于第 25 太阳周的起始阶段²。图3.1基于 1749 年 1 月至 2021 年 8 月期间月均太阳黑子数时间序列数据绘制了太阳黑子数的变化趋势。图3.1显示，MSSN 数据集的分布是右偏的。[Panigrahi 等 \(2021\)](#) 指出，MSSN 数据集的峰度大于 3。鉴于 MSSN 数据集呈现出复杂的非高斯型分布，因此较难预测出太阳黑子数的变化趋势。

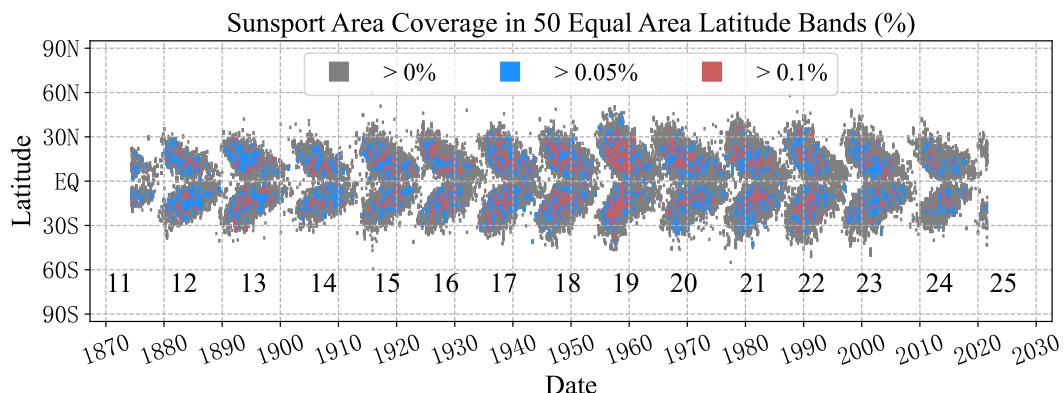


图 3.2 1949 年至 2021 年期间太阳黑子面积随时间和纬度变化的蝴蝶图。

Figure 3.2 Butterfly diagram of sunspot area which marks the latitude of sunspot locations as a function of time from 1749 to 2021.

第二类数据集为太阳黑子面积（Sunspot Area，简称 SSA）。SSA 数据集来自 Lisa Upton 和 David Hathaway 提供的网站³，该数据集中的面积是指太阳黑子面积占可见太阳半球的比例。SSA 数据集可用来预测太阳黑子面积和太阳黑子蝴蝶图。[Hathaway \(2015\)](#) 指出，太阳黑子面积也是太阳磁活动指标，同样也可以应用在预测太阳黑子活动方面。SSA 数据集跨越了 ~147 年（从 1874 年 5 月 1 日至 2021 年 8 月 7 日），包含 247,693 条记录。因此 SSA 数据集包含了 13 个完整的太阳周。同 MSSN 数据集相比，SSA 数据集还有太阳黑子发生的位置信息，因此 SSA 数据集在理解长期太阳磁活动和变化时非常重要。图3.2根据 1874 年至 2021 年期间太阳黑子面积随着时间的变化绘制出蝴蝶图。在每个太阳周早期，太阳黑子出现在高纬度地区，然后向赤道方向移动。

为保持太阳黑子数和面积在时间上的一致性，将 SSA 数据集中的太阳黑子面积按月均，得到月均太阳黑子面积（Monthly Mean Sunspot Areas，简称 MSSA）数据集。将 MSSN 和 MSSA 两种时间序列数据集绘制在同一张图中，可显示太

²参考：[www://sidc.be/silso/cyclesmm](http://sidc.be/silso/cyclesmm)

³参考：<http://solarcyclescience.com/activeregions.html>

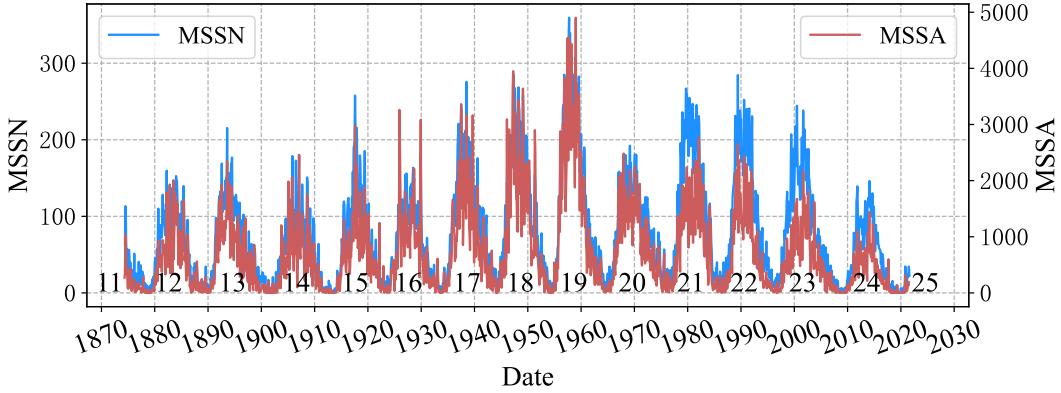


图 3.3 1874 年至 2021 年期间月均太阳黑子数和面积。

Figure 3.3 Monthly mean sunspot numbers and monthly mean sunspot areas from 1874 to 2021.

阳黑子数和太阳黑子面积之间的动态变化关系。图3.3截取了 MSSA 在 1874 年至 2021 年的数据集，纵轴按照不同尺度标准绘制太阳黑子数和太阳黑子面积的变化趋势。在重叠的时间范围内，太阳黑子数和面积在太阳黑子活动和发生时间上存在高度的关联性。

3.2.2 方法描述

本节围绕预处理方法和机器学习模型展开。为了更精准地捕获太阳黑子活动趋势，本章使用了近几百年来观测到的太阳黑子数量/面积，使用不同种类的神经网络，试图从历史观测数据中提取关键信息，预测太阳黑子的动态变化特征。太阳黑子活动可以从长期历史记录中得到，依赖的历史时间长度从几个月到几个太阳周不等。将历史观测数据集转化为监督学习时间序列数据集。太阳黑子活动趋势可以被近似表达为：

$$S(t + \Delta, t + 2\Delta, \dots, t + f\Delta) = F[S(t, t - \Delta, \dots, t - (k - 1)\Delta)]. \quad \dots (3.1)$$

其中， Δ 代表样本的采样间隔（1 个月）， $S(t + \Delta, t + 2\Delta, \dots, t + f\Delta)$ 代表未来 f 个月的太阳黑子活动， $S(t, t - \Delta, \dots, t - (k - 1)\Delta)$ 代表历史 k 个月的太阳黑子活动。

通过第2.3.1节的滑动窗口法，可以将原始太阳黑子数量数据集转换为监督学习数据集。紧接将数据集分割为训练集和测试集，分割比例为 0.8: 0.2。在训练集和测试集被使用之前，需要对两者分别进行归一化处理（第2.3.2节）。MSSA 数据集也进行了同样的操作。式3.1明确指出，模型的输入是历史记录的太阳黑

子数量/面积，输出为未来的太阳黑子数量/面积。

3.3 试验结果与分析

太阳黑子时间序列呈现出非线性、非平稳性和非高斯性等特征。第2.2.8节和第2.2.9节提到，LSTM-RNN 和 1DCNN 均适用于时间序列数据，而且原始记录持续了几百年，因此这里选择了 LSTM-RNN 和 1DCNN。考虑到 LSTM 神经元、卷积运算在处理时间序列数据方面具有不同的优势，综合这两种优势，可以组成 LSTM-1DCNN。

太阳黑子数据集记录的时间相对较长。MSSN 数据集跨越了~273 年，MSSA 数据集跨越了~147 年。引言 3.1 提到，太阳黑子活动存在周期性，平均为~11 年。这里考虑输入时间窗口长度为 72、132、264，评估不同输入时间窗口长度对模型的影响。每个输入响应可表示如下：

$$I_{-k} = [S(t), S(t-1), \dots, S(t-k+1)]. \quad \dots (3.2)$$

其中， $k = 72, 132, 264$ 。

本研究重点关注未来一个太阳周内的太阳黑子活动。为了预测第 25 太阳周太阳黑子的峰值，可以将输出时间窗口长度设置为 72。将未来 72 个月太阳黑子最大值定义为第 25 太阳周的峰值。每个输出响应可以表示如下：

$$O_{+f} = [S(t+1), S(t+2), \dots, S(t+f)]. \quad \dots (3.3)$$

其中， $f = 1, 72$ 。

这里简要介绍不同模型的超参数配置。所有模型的输入节点数由输入时间窗口长度 k 决定，输出节点数由输出时间窗口长度 f 决定。以下是本章所有神经网络通用的超参数：训练经过 1000 回合；网络每层的激活函数均使用线性整流函数（Rectified Linear Unit，简称 ReLU）；批量训练的数据量设为 132；Adam 方法作为优化器；学习率初始值设为 10^{-3} ，同时训练次数每增加 10 次，学习率会减小 $1 - 10^{-6}$ 倍，这样逐步减小学习率能够防止模型陷入局部极小值；所有试验均使用“早停”策略，即 100 次迭代后，若测试集的 MSE 值没有下降，则停止训练。

3.3.1 预测未来 1 个月太阳黑子活动

本节采用了三层 LSTM-RNN、两层 1DCNN 和三层 LSTM-1DCNN。因各种网络独特的性质，在设置以下超参数时会有所差异。针对 LSTM-RNN，隐藏层含 LSTM 神经元，隐藏层的节点数分别设为 64 和 32，最后一层为全连接层。针对 1DCNN，隐藏层为一维卷积层，过滤器个数、过滤器大小、步长分别为 32、3、2；卷积层后连接了最大池化层，池化大小为 2，步长为 1；最后一层为全连接层。针对 LSTM-1DCNN，第一层含有 LSTM 神经元，神经元个数分别为 64；第二层为一维卷积层，过滤器个数、过滤器大小、步长分别为 32、3、1；最后一层为全连接层。

表 3.2 不同模型和输入时间窗口长度下预测未来 1 个月太阳黑子数的拟合指标效果

Table 3.2 The indicators for predicting the next monthly sunspot numbers by different models with different input window length.

模型	I_{-k} -[节点或过滤器数]- O_{+f}	测试集	
		MSE	RMSE
LSTM-RNN	72-[64-32]-1	0.0030	0.0548
LSTM-RNN	132-[64-32]-1	0.0032	0.0569
LSTM-RNN	264-[64-32]-1	0.0033	0.0578
1DCNN	72-[32]-1	0.0041	0.0640
1DCNN	132-[32]-1	0.0042	0.0647
1DCNN	264-[32]-1	0.0043	0.0658
LSTM-1DCNN	72-[64-32]-1	0.0029	0.0543
LSTM-1DCNN	132-[64-32]-1	0.0032	0.0567
LSTM-1DCNN	264-[64-32]-1	0.0032	0.0566

表 3.3 最佳模型预测的太阳黑子活动。

Table 3.3 Prediction the sunspot activity by the best models.

日期	太阳黑子活动	输入时间窗口长度	LSTM-RNN	1DCNN	LSTM-1DCNN
2021 年 9 月	数量	72 个月	38.87	31.66	40.97
2021 年 8 月	面积	72/132 个月	274.14	359.78	288.61

讨论输出时间窗口长度为 1 个月时太阳黑子数。表 3.2 展示不同模型（LSTM-RNN、1DCNN 和 LSTM-1DCNN）在不同的输入时间窗口长度（72 个月、132 个月、

264个月)下预测未来1个月太阳黑子数的拟合指标效果。表3.3第一行展示了预测2021年9月太阳黑子数。就网络性能而言,LSTM-1DCNN性能略优于LSTM-RNN和1DCNN。就输入时间窗口长度而言,历史72个月的太阳黑子数作为时间窗口长度所得到的模型是最优的。当输入时间窗口长度为72个月时,LSTM-RNN的拟合指标较小($MSE=0.0030$ 和 $RMSE=0.0548$),预测2021年9月的太阳黑子数为38.87;1DCNN的拟合指标较小($MSE=0.0041$ 和 $RMSE=0.0640$),预测2021年9月的太阳黑子数为31.66;LSTM-1DCNN的拟合指标较小($MSE=0.0029$ 和 $RMSE=0.0543$),预测2021年9月的太阳黑子数为40.97。

讨论输出时间窗口长度为1个月时太阳黑子面积。表3.4展示在不同模型(LSTM-RNN、1DCNN和LSTM-1DCNN)和输入时间窗口长度(72、132、264)下预测未来1个月太阳黑子面积的拟合指标效果。表3.3第二行展示了预测2021年8月太阳黑子面积。就网络性能而言,LSTM-1DCNN的性能同样略优于LSTM-RNN和1DCNN。当输入时间窗口长度为72个月时,LSTM-RNN的拟合指标较小($MSE=0.0020$ 和 $RMSE=0.0446$),预测2021年8月太阳黑子面积为291.30;当输入时间窗口长度为132个月时,1DCNN的拟合指标较小($MSE=0.0028$ 和 $RMSE=0.0531$),预测2021年8月太阳黑子面积为359.78;当输入时间窗口长度为132个月时,LSTM-1DCNN的拟合指标较小($MSE=0.0021$ 和 $RMSE=0.0453$),预测2021年8月太阳黑子面积为288.61。

图3.4绘制了输入时间窗口长度为72的最佳模型预测未来1个月的太阳黑子数量,图3.5绘制了输入时间窗口长度为132的最佳模型预测未来1个月的太阳黑子面积。从图3.4和图3.5可以看出,对于预测值和测试集,LSTM-1DCNN具备良好的拟合能力。所有的结果均显示,预测的峰值比观测值略低,预测的谷值比观测值略高,这是因为太阳黑子活动较弱或者较强时样本量偏少,模型难以准确地学习到这些特征。

3.3.2 预测未来72个月太阳黑子活动

第3.3.1节预测未来1个月太阳黑子数量/面积时,得出LSTM-1DCNN性能略优于LSTM-RNN和1DCNN。同时考虑到LSTM-1DCNN结合了LSTM-RNN和1DCNN两者的优点,以下试验不再考虑LSTM-RNN和1DCNN两种模型。本节讨论输出时间窗口长度为72个月的太阳黑子活动。表3.5展示了不同架构的LSTM-1DCNN模型在不同输入时间窗口长度(72、132、264)下预测未来72个

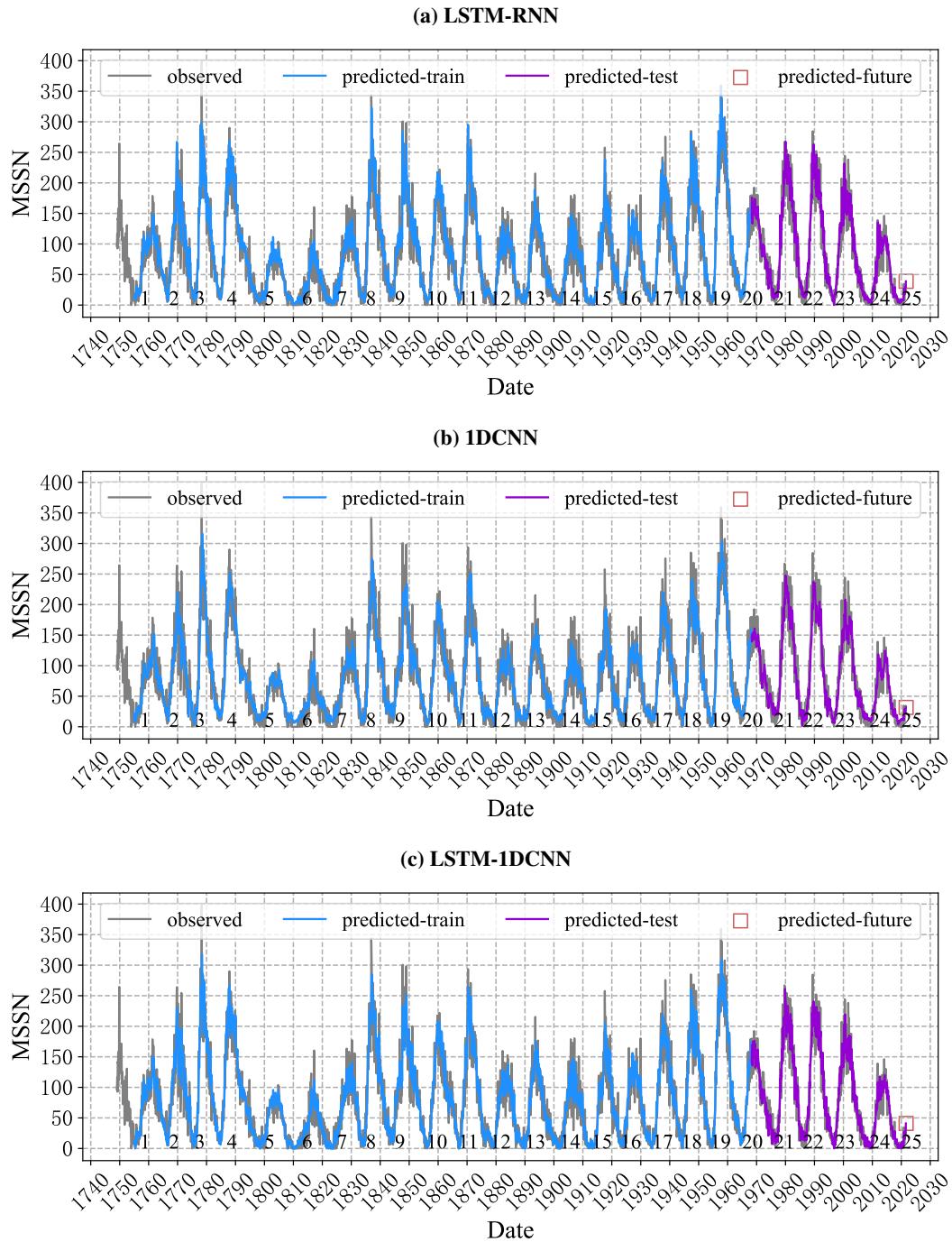


图 3.4 最佳模型预测未来 1 个月的太阳黑子数。

Figure 3.4 Predicting the next monthly sunspot numbers by the best models.

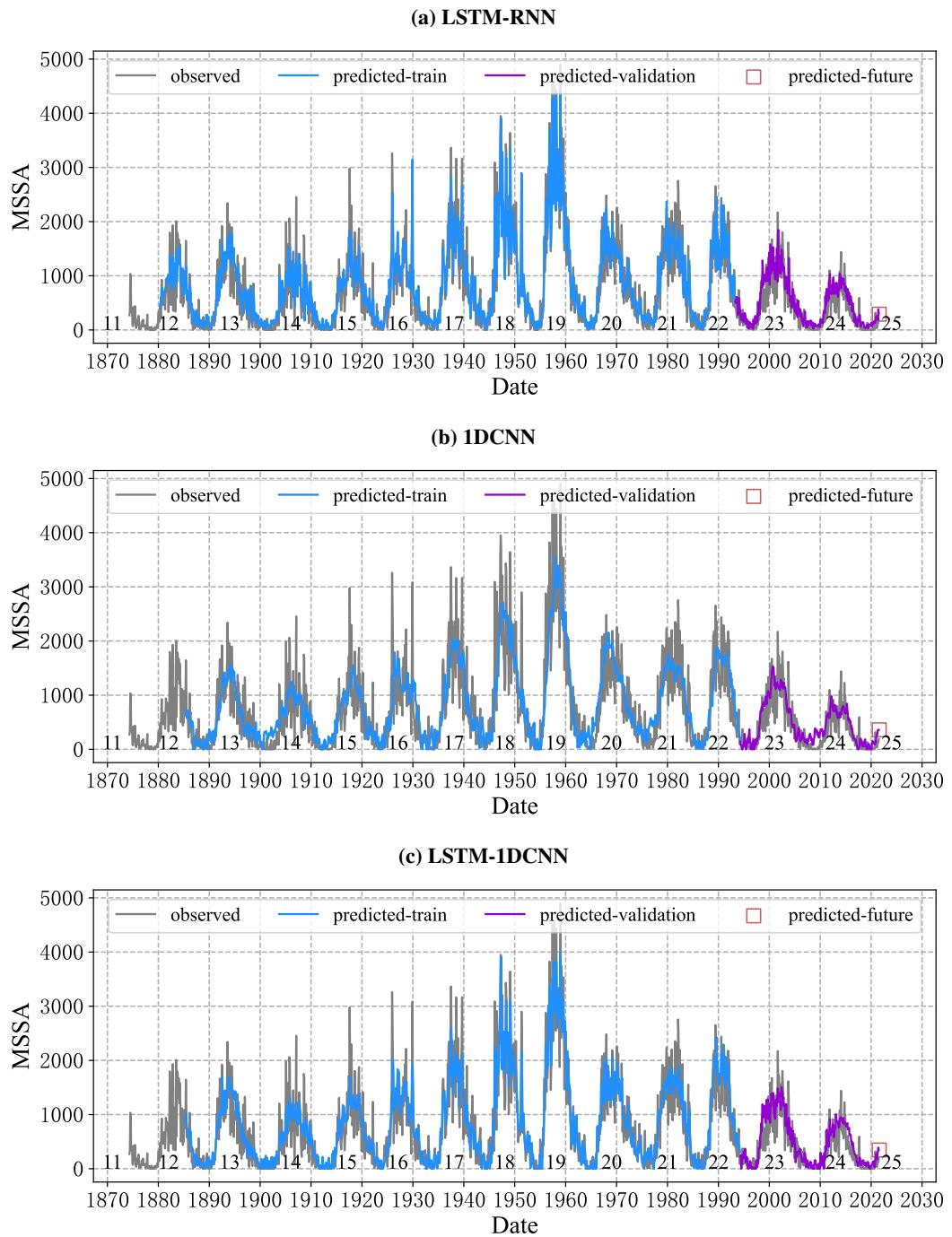


图 3.5 最佳模型预测未来 1 个月的太阳黑子面积。

Figure 3.5 Predicting the next monthly sunspot areas by the best models.

表 3.4 不同模型和输入时间窗口长度下预测未来 1 个月太阳黑子面积的拟合指标效果。

Table 3.4 The indicators for predicting the next monthly sunspot areas by different models with different input window length.

模型	I_{-k} -[节点或过滤器数]- O_{+f}	测试集	
		MSE	RMSE
LSTM	72-[64-32]-1	0.0021	0.0454
LSTM	132-[64-32]-1	0.0023	0.0479
LSTM	264-[64-32]-1	0.0024	0.0492
1DCNN	72-[32]-1	0.0035	0.0588
1DCNN	132-[32]-1	0.0028	0.0531
1DCNN	264-[32]-1	0.0030	0.0551
LSTM-1DCNN	72-[64-32]-1	0.0021	0.0454
LSTM-1DCNN	132-[64-32]-1	0.0021	0.0453
LSTM-1DCNN	264-[64-32]-1	0.0023	0.0479

月太阳黑子数的拟合指标效果。与表3.2相比，表3.5中 MSE 值和 RMSE 值出现了一定程度的增加。也就是说，随着输出时间窗口长度的增加，模型性能呈现下降趋势。在相同的输入时间窗口长度下，隐藏层层数和节点数越多，模型性能先上升后下降。就输入时间窗口长度而言，历史 264 个月的太阳黑子数作为输入时间窗口长度所得到的模型是最优的。

表3.6第一行展示了基于最佳模型预测未来 72 个月最大的太阳黑子数和出现时间。当输入时间窗口长度为 264 个月时，3 层 LSTM-1DCNN 的拟合指标较小 ($MSE=0.0085$ 和 $RMSE=0.0920$)，预测未来 72 个月的太阳黑子数的最大值为 151.55，出现在 2023 年 9 月；4 层 LSTM-1DCNN 的拟合指标较小 ($MSE=0.0082$ 和 $RMSE=0.0905$)，预测未来 72 个月的太阳黑子数最大值为 174.71，出现在 2025 年 1 月；5 层 LSTM-1DCNN 的拟合指标较小 ($MSE=0.0072$ 和 $RMSE=0.0849$)，预测未来 72 个月的太阳黑子数的最大值为 132.86，出现在 2024 年 12 月。对黑子数而言，第 23 太阳周最大 MSSN 出现在 2001 年 9 月，其值为 238.2；第 24 太阳周最大 MSSN 出现在 2014 年 2 月，其值为 146.1。研究结果显示，第 25 太阳周的峰值跟第 24 太阳周基本持平。

图3.6绘制了不同结构的 LSTM-1DCNN 的最佳模型预测未来 72 个月的太阳黑子数。为了更清晰地展示结果，图3.6在绘制训练集和测试集的结果时，只绘

表 3.5 不同输入时间窗口长度和层数下 LSTM-1DCNN 预测未来 72 个月太阳黑子数的拟合指标效果。

Table 3.5 The indicators for predicting the next 72 monthly sunspot numbers by LSTM-1DCNN with different input window length and layers.

层数	I_{-k} -[节点或过滤器数]- O_{+f}	测试集	
		MSE	RMSE
3	72-[32(LSTM)-16(Conv)]-72	0.0189	0.1373
	72-[64(LSTM)-32(Conv)]-72	0.0165	0.1283
	72-[128(LSTM)-64(Conv)]-72	0.0140	0.1182
	132-[32(LSTM)-16(Conv)]-72	0.0195	0.1397
	132-[64(LSTM)-32(Conv)]-72	0.0124	0.1115
	132-[128(LSTM)-64(Conv)]-72	0.0133	0.1151
	264-[32(LSTM)-16(Conv)]-72	0.0155	0.1245
	264-[64(LSTM)-32(Conv)]-72	0.0122	0.1103
	264-[128(LSTM)-64(Conv)]-72	0.0085	0.0920
4	72-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0179	0.1339
	72-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0135	0.1162
	72-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0146	0.1206
	132-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0170	0.1305
	132-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0132	0.1148
	132-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0149	0.1220
	264-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0089	0.0944
	264-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0071	0.0841
	264-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0093	0.0965
5	72-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0173	0.1315
	72-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0134	0.1157
	72-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0158	0.1258
	132-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0150	0.01226
	132-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0163	0.1276
	132-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0161	0.1268
	264-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0087	0.0935
	264-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0072	0.0849
	264-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0074	0.0859

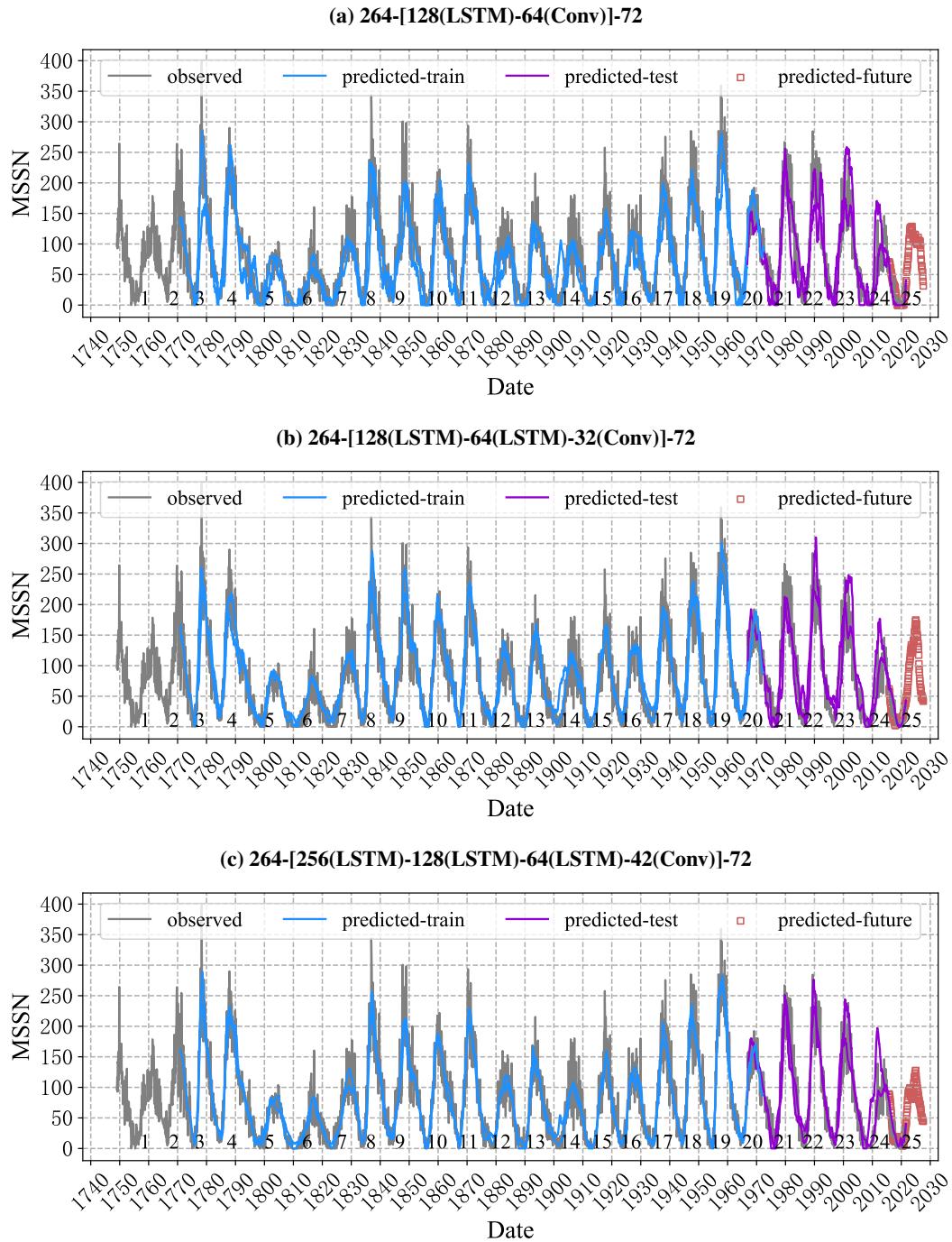


图 3.6 最佳模型预测未来 72 个月的太阳黑子数。

Figure 3.6 Predicting the next 72 monthly sunspot numbers by the best models.

表 3.6 最佳模型预测未来 72 个月太阳黑子活动的最大值。**Table 3.6 Prediction the maximum of sunspot activity by the best models.**

	3 层		4 层		5 层		
	输入时间窗口长度	最大值	发生时间	最大值	发生时间	最大值	发生时间
数量	264	151.55	2023 年 9 月	174.71	2025 年 1 月	132.86	2024 年 12 月
面积	132/264/72	1016.32	2024 年 8 月	1469.01	2025 年 3 月	1397.77	2024 年 4 月

制了每个输入样本的第一个和最后一个预测值。图3.6显示，未来第 72 个月太阳黑子数的最大值与第 24 太阳周的峰值基本持平，而且在第 25 太阳周出现了双峰性质。相比于图3.6b和图3.6c，图3.6a中的第一个峰值比第二个大，导致预测的峰值出现的时间提前。

表3.6第二行展示了基于最佳模型预测未来 72 个月最大的太阳黑子面积和出现时间。3 层 LSTM-1DCNN 模型的输入时间窗口长度为 132 个月时，拟合指标较小 ($MSE=0.0078$ 和 $RMSE=0.0884$)，预测未来 72 个月太阳黑子面积最大值为 1016.32，出现在 2024 年 8 月；4 层 LSTM-1DCNN 的拟合指标较小 ($MSE=0.0082$ 和 $RMSE=0.0905$)，预测未来 72 个月太阳黑子面积最大值为 1469.01，出现在 2025 年 3 月；5 层 LSTM-1DCNN 的拟合指标较小 ($MSE=0.0072$ 和 $RMSE=0.0849$)，预测未来 72 个月太阳黑子面积最大值为 1397.77，出现在 2024 年 4 月。对于太阳黑子面积，第 23 太阳周 MSSA 的峰值为 2171.7，出现在 2001 年 9 月；第 24 太阳周 MSSA 的峰值为 1439.82，出现在 2014 年 2 月。研究结果显示，第 25 太阳周的峰值跟第 24 太阳周基本持平。

为了更清晰地展示结果，在展示训练集和测试集的结果时，只绘制了每个输入样本的第一个和最后一个预测值。图3.7绘制了不同结构的 LSTM-1DCNN 的最佳模型预测未来 72 个月的太阳黑子面积。图3.7显示，未来第 72 个月超过了在第 25 太阳周峰值的到达时间。图3.7a和图3.7c均在第 25 太阳周均出现了双峰性质。

图3.6和图3.7分别绘制了不同层数下最佳模型预测太阳黑子数量和面积。从图3.6和图3.7可以看出，对于预测值和测试集，LSTM-1DCNN 具备良好的拟合能力。所有的结果均显示，预测的峰值比观测值略低，预测的谷值比观测值略高，因为太阳黑子活动较弱或较强的样本量偏少，模型难以准确学习到这些特征。

表 3.7 不同输入时间窗口长度和层数的 LSTM-1DCNN 预测未来 72 个月太阳黑子面积的拟合指标效果。

Table 3.7 The indicators for predicting the next 72 monthly sunspot areas by LSTM-1DCNN with different input window length and layers.

模型	I_{-k} -[节点或过滤器数]- O_{+f}	测试集	
		MSE	RMSE
3	72-[32(LSTM)-16(Conv)]-72	0.0105	0.1025
	72-[64(LSTM)-32(Conv)]-72	0.0085	0.0923
	72-[128(LSTM)-64(Conv)]-72	0.0097	0.0987
	132-[32(LSTM)-16(Conv)]-72	0.0093	0.0965
	132-[64(LSTM)-32(Conv)]-72	0.0091	0.0952
	132-[128(LSTM)-64(Conv)]-72	0.0078	0.0884
	264-[32(LSTM)-16(Conv)]-72	0.0091	0.0956
	264-[64(LSTM)-32(Conv)]-72	0.0090	0.0950
	264-[128(LSTM)-64(Conv)]-72	0.0084	0.0919
4	72-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0107	0.1036
	72-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0097	0.0986
	72-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0099	0.0993
	132-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0100	0.0998
	132-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0091	0.0953
	132-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0091	0.0953
	264-[64(LSTM)-32(LSTM)-16(Conv)]-72	0.0096	0.0978
	264-[128(LSTM)-64(LSTM)-32(Conv)]-72	0.0075	0.0865
	264-[256(LSTM)-128(LSTM)-64(Conv)]-72	0.0054	0.0733
5	72-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0105	0.1023
	72-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0071	0.0841
	72-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0061	0.0781
	132-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0089	0.0941
	132-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0063	0.0791
	132-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0067	0.0817
	264-[128(LSTM)-64(LSTM)-32(LSTM)-21(Conv)]-72	0.0101	0.1006
	264-[256(LSTM)-128(LSTM)-64(LSTM)-42(Conv)]-72	0.0075	0.0867
	264-[512(LSTM)-256(LSTM)-128(LSTM)-85(Conv)]-72	0.0096	0.0980

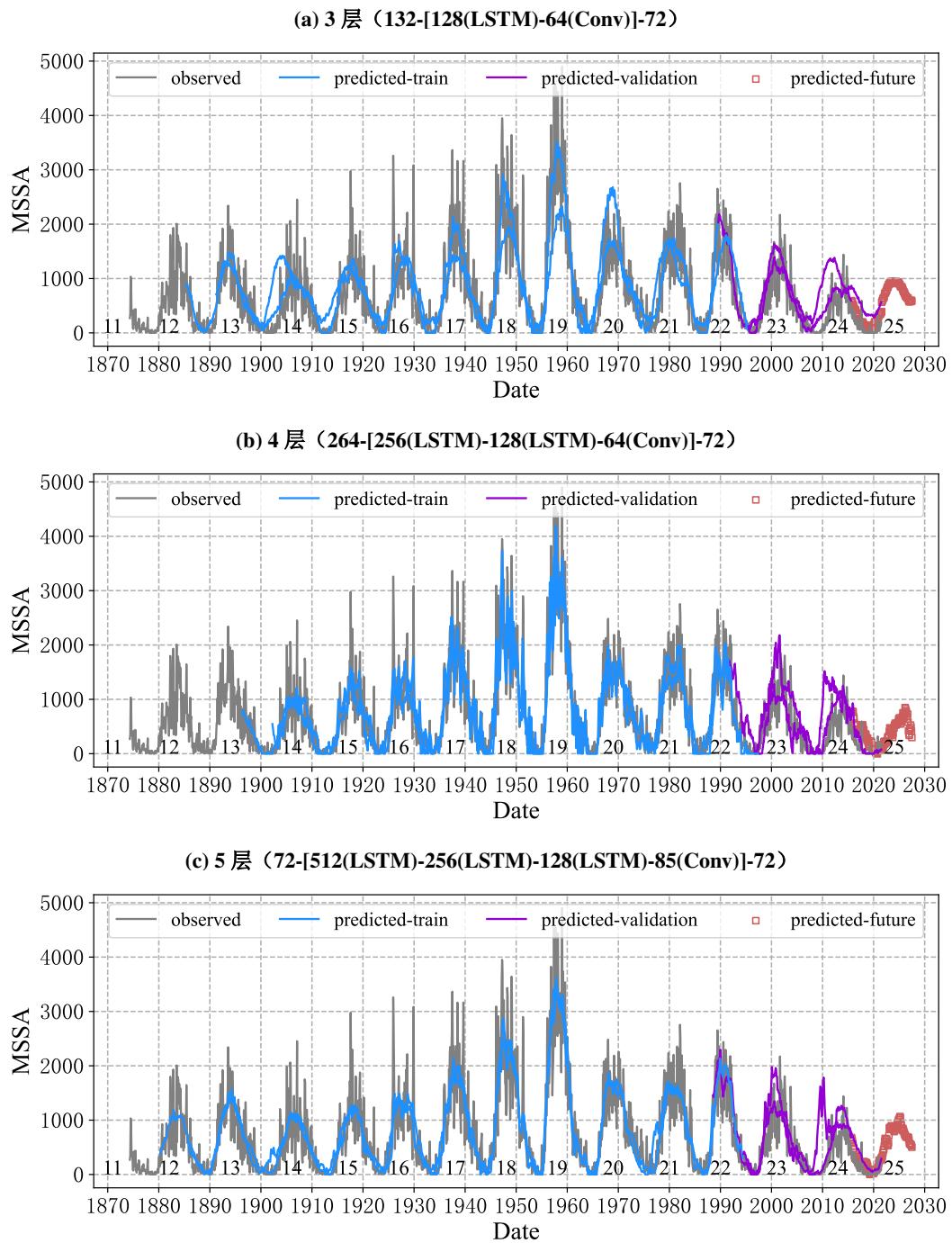


图 3.7 最佳模型预测未来 72 个月的太阳黑子面积。

Figure 3.7 Predicting the next 72 monthly sunspot areas by the best models.

3.3.3 与其他研究的比较

表 3.8 不同研究给出的预测第 25 太阳周的太阳黑子活动相比于第 24 太阳周的强弱

Table 3.8 The different studies for predicting sunspot amplitude of cycle 25 comparing with cycle 24.

方法	相比第 24 太阳周	峰值达到时间	参考文献
LSTM-1DCNN	基本相等	2024 年 12 月–2025 年 3 月	本研究
多维混合神经网络	基本相等	2025 年 1 月 (± 6 个月)	Okoh 等 (2018)
光谱分量外推法	29% 更强	2022 年至 2023 年之间	Kane (2007)
光谱小波分解树	17% 更强	2023 年 4 月	Rigozo 等 (2011)
前向传播神经网络	更弱	2022 年至 2023 年左右	Covas 等 (2019)
相似周期方法	30% 更强	2024 年 10 月	Du (2020)
LSTM-RNN	55% 更强	2022 年 10 月	Li 等 (2021)

已有研究中采取不同的方法预测已经到来的第 25 太阳周的太阳黑子活动，得到结果并不统一。可能的原因是太阳黑子时间序列具备极其复杂的易变性、数据集长度、预处理以及选择模型的差异等。表 3.8 列出部分关于第 25 太阳周的研究。例如，Covas 等 (2019) 使用具备时空特性的神经网络，预测第 25 太阳周可能是有史以来太阳黑子数最低的时期，预测峰值出现的时间为 2022 年至 2023 年。Okoh 等 (2018) 利用多维混合模型估计 MSSN，得到第 25 太阳周的峰值为 122.1 (± 18.2)，达到时间为 2025 年 1 月 (± 6 个月)。

3.4 讨论与小结

Solanki 等 (2011) 指出不同的关于长期预测太阳黑子活动得出的结论差异较大，这主要是因为太阳黑子活动具有非常复杂的非线性特征。MSSN 和 MSSA 时间序列数据是非稳态的，数据中所包含的随机的分量使太阳黑子在长时间预测尺度上不太可靠。由于数据中含有非线性效应的随机波动，难以对未来几个周期进行准确预测 (Charbonneau, 2010; Petrovay, 2010)。Charbonneau (2010) 认为，传统物理的物理模型没有建立在观测数据集上，难以反应真实情况。Mendoza 等 (2011) 指出，太阳黑子活动起源于一定范围内的周期过程，而不是随机或间歇性过程。

截至目前，太阳黑子预测最大的挑战是预测下一个太阳周的最大振幅和持

续时间 (Petrovay, 2010)。Gleissberg (1939) 曾指出太阳黑子周期的波动不够规则，在获取响应的表达函数时相当困难。Herrera 等 (2015) 假设太阳黑子每个周期存在一个固定的能量和给定频率（即太阳黑子活动周期在 8 年至 14 年之间），太阳黑子估计模型的准确度会受到不确定性原则的限制。因此，准确预测太阳黑子的任何参数（相位、振幅和周期）都十分困难。不仅未来太阳周期的特性难以预测，而且准确预知太阳黑子活动也几乎不可能。太阳黑子周期预测的问题可被视作一个概率问题，因此太阳周期预测中经常会出现太阳黑子周期的预测概率。Camporeale (2019) 建议在空间天文研究中需要改变预测模式，从点预测到具有可靠不确定性的概率方法。太阳周期的振幅取决于其在震荡阶段的位置变化，理论上可以定性预测第 25 太阳周的最大振幅。然而，不同的物理模型和数值模型预测未来太阳周最大振幅时还存在较大差异。

本研究使用了 LSTM-1DCNN 预测未来太阳黑子活动，采用了 MSSN 和 MSSA 数据集，发现第 25 太阳周的最大太阳黑子数为 132.86（出现在 2024 年 12 月），最大太阳黑子面积为 1469.01（出现在 2025 年 3 月）。该预测结果与 Covas 等 (2019) 的研究结果不仅在黑子活动上类似，而且在出现时间上也相近。

同时研究还发现，模型性能会随着输出时间窗口长度的增加而略微下降，随着输入时间窗口长度的增加而先上升后下降。如果要提供非常准确的预报，可以将输出时长设置为 1 年或 12 年，每隔一段时间更新模型，重新预测的太阳黑子活动。输入时间窗口长度和输出时间窗口长度直接决定了模型的构建，而且输出时间窗口长度对预测结果有较大的影响，这使得合适的输出序列长度成为构建良好模型的一项前提条件。

第4章 基于机器学习预测龙子祠泉流量

第3章基于一种特征（太阳黑子强度）研究时间序列数据。在本章的研究中，输入数据的特征（历史泉流量和降水量）有所增加，利用机器学习探索龙子祠泉流量动态变化趋势。本章可以检验增加特征种类对机器学习探索时间序列数据的影响。

4.1 研究背景

近几十年来，全球人口持续增长，生活用水和农业用水的需求也随之增长，淡水短缺现象日益严重 (Portmann 等, 2010; Iglesias 等, 2015)。地下含水层中存储着大量的未冻结淡水，这些淡水有助于人类生存的可持续发展。然而，全球地下水每年消耗 $\sim 283 \text{ km}^3/\text{year}$ ，出现了一定程度的过度开采 (Wada 等, 2010)。地下水过度开采的后果包括井水干涸、溪流和湖泊水量减少、水质退化、抽水成本增加、地面沉降、油井产量下降等 (Nayak 等, 2006)。Dalin 等 (2017) 和 Jr. 等 (2018) 研究表明，灌溉是地下水损耗的主要原因。

有效管理水资源是一项艰巨的任务，需要考虑到不同的时间尺度 (Galelli 等, 2010)。Kresic 等 (2009) 在水资源管理中引入了可持续发展的概念，对泉流量进行了深入研究。Coppola Jr 等 (2003) 指出，地下水位预测对于实施最佳水管理政策和跨地缘政治保护水资源政策至关重要。地下水是一种可再生资源，在干旱时期地下水具备良好的抵御能力。通常地下水资源按半季节至季节进行规划，优化水的利用效率，可保持农地的土壤含水容量以及维持水系统平衡。

过去的几十年里，由数据驱动的模型在水文领域的应用发展迅速，综述见 (Abrahart 等, 2012; Raghavendra. N 等, 2014)。应用领域包括降水—径流模型 (Dibike 等, 2001; Solomatine 等, 2003)、土壤湿度 (Ahmad 等, 2010)、干旱预测 (Le 等, 2016)、波浪冲刷深度 (Etemad-Shahidi 等, 2011)、河流产沙量 (Goyal, 2014)、桥墩周围的最大冲刷深度 (Najafzadeh 等, 2016)、下游水闸的局部冲刷深度 (Najafzadeh 等, 2017b)、评估泥沙输移 (Najafzadeh 等, 2017a)、主河道和漫滩的流量 (Zahiri 等, 2018) 等。

已有研究表明，利用机器学习预测地下水水位具备可行性。例如，Coppola Jr

等 (2003) 利用 ANNs 预测地下水位; Sun (2013) 通过空间插值技术耦合数据驱动模型预测地下水水位变化; Tapoglou 等 (2014) 基于混合 ANN-Kriging 模型, 模拟德国巴伐利亚州 Isar 河流域的每日地下水水位变化; Sun 等 (2015) 基于 ANNs 预测新加坡沼泽森林地下水水位变化; Yadav 等 (2017) 比较了极端学习机和 SVM 在预测加拿大两个不同油井的月度地下水水位方面的性能, 发现极端学习机的表现优于 SVM; Sahoo 等 (2017) 使用微分机器学习算法预测高平原含水层和密苏里河流域的水位变化, 利用光谱分析确定合成输入, 得出 ANNs 优于混合线性和非线性回归模型; Guzman 等 (2017) 采用非线性自回归神经网络预测密西西比河含水层中井水水位; Wunsch 等 (2018) 也使用非线性自回归神经网络对德国西南部的几个油井预测每月地下水水位, 表明了非线性自回归神经网络预测地下水水位具备优势; Amaranto 等 (2018) 比较了五种不同的数据驱动模型在不同水文状况下季节性地下水水位的预测性能, 发现所有由数据驱动的模型均优于基线模型, 且在缺水条件下误差有所增加; Rakhshandehroo 等 (2018) 基于小波神经网络预测佛罗里达州浅井和阿肯色州深井的地下水水位, 得出浅井中嘈杂的地下水波动导致误差更高; Amaranto 等 (2019) 利用两步数据驱动建模方法, 基于气候变量、地表水可用量、地下水位变化和人类水管理预测未来地下水可用量。

泉流量在地下水研究中非常重要 (Toth, 1971; Tóth, 1999)。泉流量代表了地下水到地表水的过渡指标, 反映了含水层的动态变化以及整个水流系统的运转情况。泉流量是一系列动态变化的结果, 主要受该地区降水量的影响。此外, 泉流量本身还会影响着泉域排入和排出的水量。水平衡条件表明, 补给含水层中储存水的变化率与水流入和流出的速度基本平衡。一般来说, 对水平衡条件进行定量分析通常需要考虑以下条件: 历史泉流量、降水量、地下水开采量、入渗、地表径流、蒸散、地下水补给、土壤水分、侧向水流至蓄水层、地表含水层和地下含水层之间的渗漏、蓄水层中蓄水量的变化等。多数情况下, 水平衡条件的评估非常复杂。

基于水平衡条件的模型来模拟泉流量的动态变化, 这很难实用化。出于实际应用, 研究者们经常采用更简化的方法。例如, Zhang 等 (1996) 使用块状参数模型和最小二乘法模拟爱荷华州石灰岩含水层泉流量随时间的变化; Lambrakis 等 (2000) 将非线性时间序列分析和 ANNs 应用于岩溶泉流量的动态变化和短期预测; Hu 等 (2008) 基于 ANNs 模拟中国娘子关泉流量; Fiorillo 等 (2010) 基于

互相关分析，研究了意大利南部两处岩溶泉的降水量和泉流量之间的关系；Fan 等 (2013) 提出了组合极值统计模型，用于研究极端气候变化和高度地下水开发条件下的泉流量消耗过程；Diodato 等 (2014) 提出了用于泉流量估算集总气候模型；Cheng 等 (2021) 利用 LSTM-RNN 和 SVR 预测了龙子祠泉未来 1 个月泉流量。

本章结构安排如下。第4.2节描述了数据与方法。第4.3节描述了试验过程，并将试验结果可视化。第4.4节对试验结果进行了总结，并对下一步研究进行了展望。

4.2 数据与方法

本节首先介绍研究区域龙子祠泉的地理条件。接着可视化分析该泉域的降水量和泉流量，找出数据的基本特征。在数据正式被训练之前，需要将原始数据集进行预处理，包括生成监督学习数据集、数据集划分、归一化处理等。最后选取几种不同的机器学习方法训练处理后的数据。

4.2.1 研究区域

龙子祠泉域坐落于山西省临汾市，属于汾河水系。从地图上看，经度范围为 $[110^{\circ}45', 111^{\circ}30']$ ，纬度范围为 $[35^{\circ}40', 36^{\circ}40']$ ，流域面积为 $\sim 2250 \text{ km}^2$ 。图4.1绘制了龙子祠泉的地理条件。该泉域位于构造剥蚀、溶蚀中低山区，属于非全排型泉域。泉水出露于西山与临汾盆地交界处的坡积物中，可划分为东池、南池和北池 3 个泉。泉群露出面积为 $\sim 0.12 \text{ km}^2$ ，泉水大多以散流的形式溢出地表。

4.2.2 数据描述

针对龙子祠泉域，本研究收集了 1987 年 1 月至 2018 年 12 月（共 32 年）的泉流量月观测资料和泉域内化乐、克城、山头、一平坦、台头、光华、河底、双凤渰、关王庙九个气象观测站的月降水资料。图4.2绘制了从 1987 年至 2018 年泉流量的长时间序列。由图4.2a看出，这 32 年期间 1987 年 7 月观测到泉流量最大值 $5.79 \text{ m}^3/\text{s}$ （排除了 1992 年 11 月出现的异常值），2011 年 7 月观测到泉流量最小值 $2.54 \text{ m}^3/\text{s}$ 。除了突发性的泉流量增加事件外，1987 年至 2005 年期间泉流量呈下降趋势，而 2006 年以后泉流量趋于稳定。出现这种现象的原因是，泉流量主要由降水量和开采量共同决定。但该地区泉水开采量数据难以收集，很难详

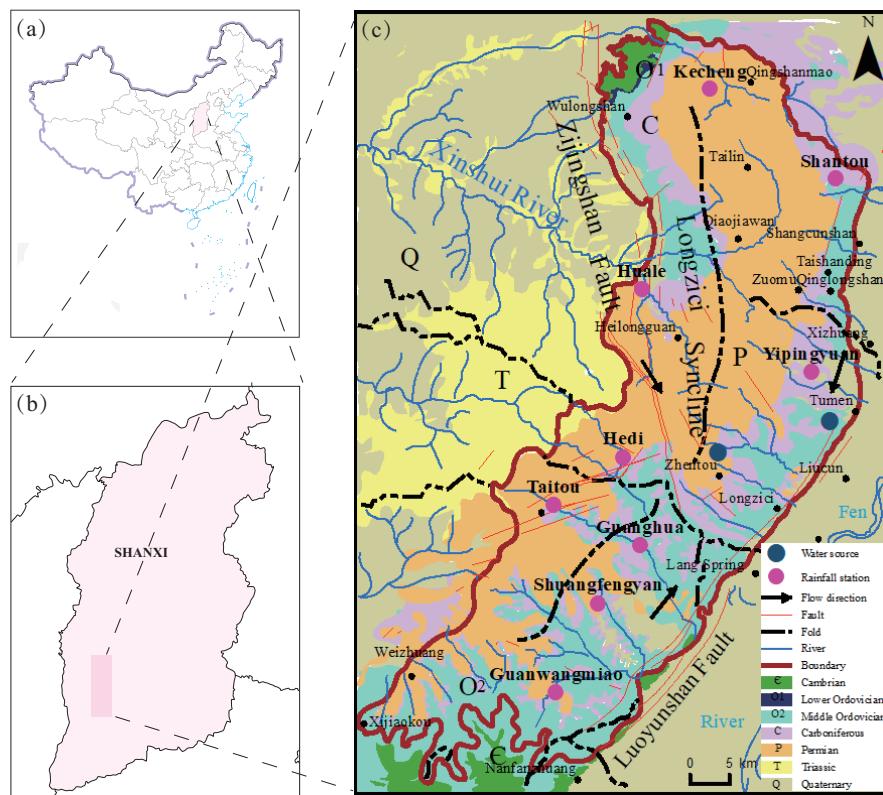


图 4.1 龙子祠泉地理条件 (Cheng 等, 2021)。

Figure 4.1 The hydrogeological conditions in Longzici area (Cheng 等, 2021).

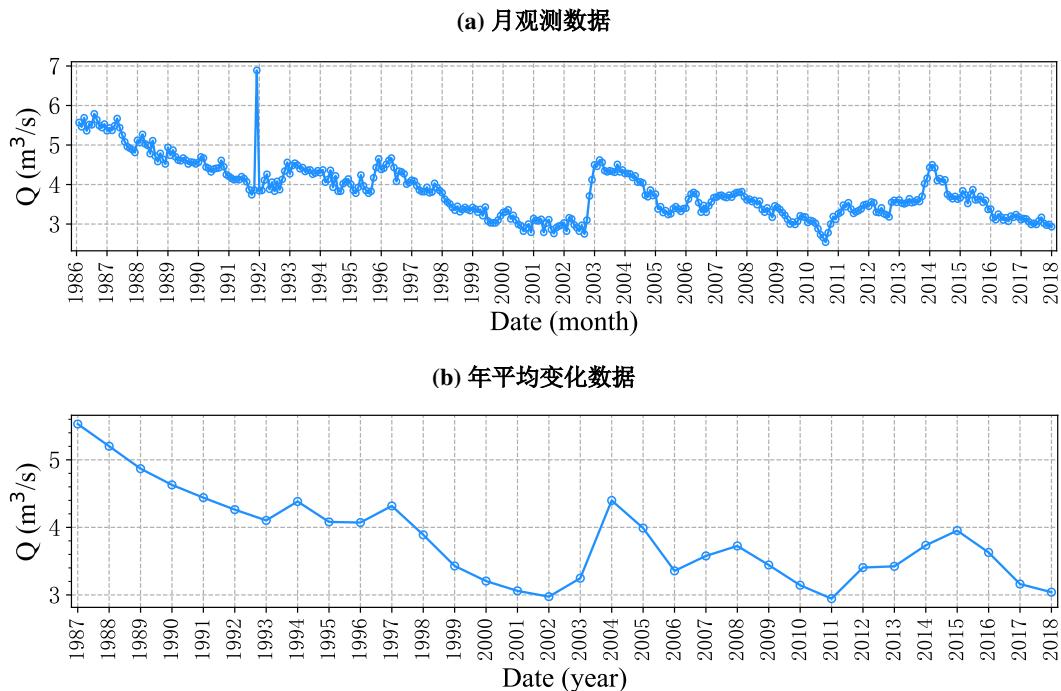


图 4.2 龙子祠泉流量变化趋势。

Figure 4.2 The discharge variation trend of Longzici spring.

细分析开采量对泉流量的影响。根据粗略统计的勘探数据，2003 年前该地区泉水开采量保持在较高的水平 ($\sim 3.16 \times 10^6 \text{m}^3/\text{year}$)，这很可能是泉流量持续下降的原因。2003 年该地区出现了极端降水天气，使泉流量显著增加。在随后的几年里，泉流量波动相对规律，且略有下降趋势。由于开采量数据难以收集，本研究只关注降水量对泉流量的影响。

图4.2b描述了 1987 年至 2018 年年均泉流量变化趋势图。这 32 年来，1987 年年均泉流量的观测值最大 ($5.53 \text{ m}^3/\text{s}$)，2011 年年均泉流量的观测值最小 ($2.95 \text{ m}^3/\text{s}$)。年均泉流量的谷值分别出现在 1993 年、2002 年、2011 年，间隔为 ~ 10 年、这种波动可能与自然环境有关（如泉水的自我调节功能）。

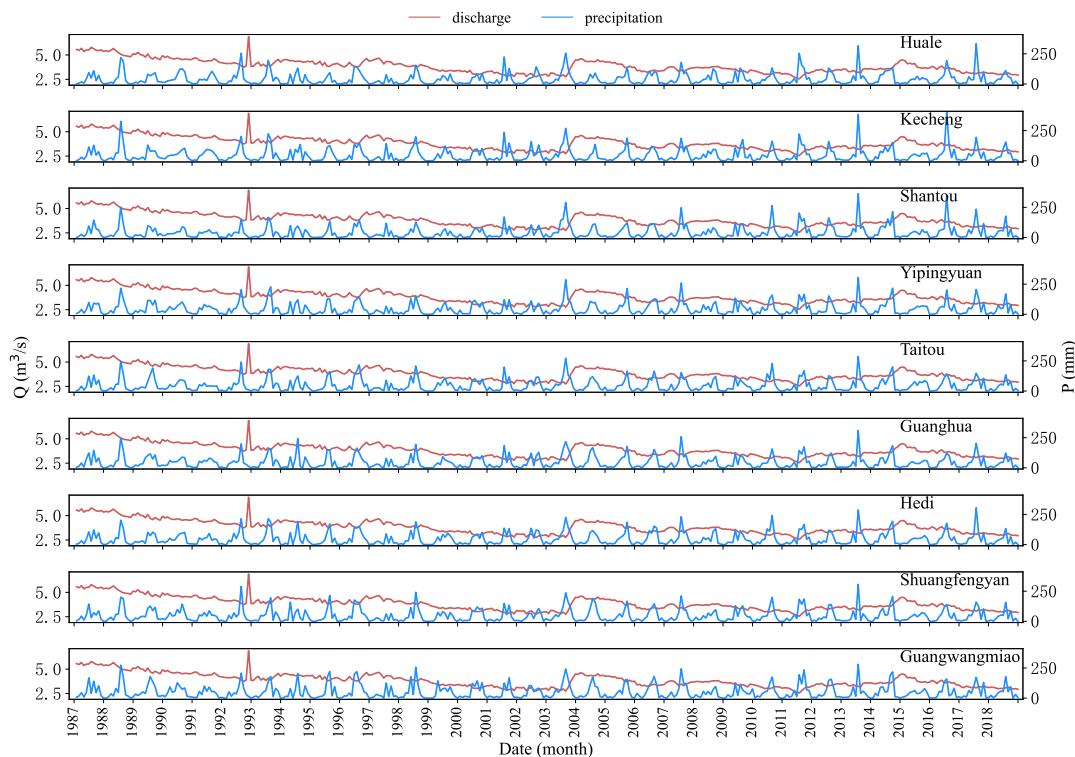


图 4.3 龙子祠泉周围 9 个区域（即化乐、克城、山头、一平垣、台头、光华、河底、双风渰、关王庙）降水量随时间变化的趋势图。

Figure 4.3 The precipitation variation trend of the nine city (i.e., Huale, Kecheng, Shantou, Yipingyuan, Taitou, Guanghua, Hedi, Shuangfengyan and Guangwangmiao) around Longzici spring.

图4.3描述了 1987 年 1 月至 2018 年 12 月期间龙子祠泉周围九个区域月均降水量随时间变化的趋势图。由图4.3可知，泉域降水量在空间上的分布存在一致性波动的原则，即九个地区的降水时间序列显示出降水波动规律较为同步。另

外，每年降水量也有一定的规律，夏季降水量明显上升，而冬季降水量明显下降，春季和秋季的降水量则相对平稳。

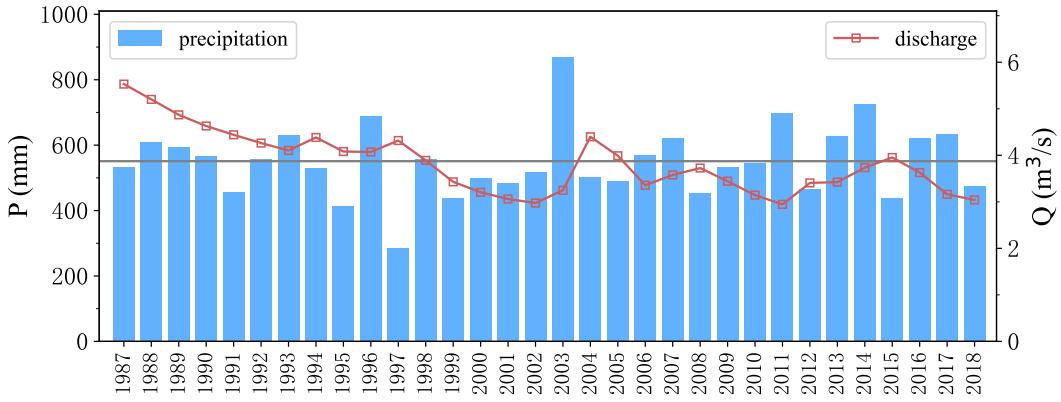


图 4.4 1987 年至 2018 年期间龙子祠泉年域降水量与泉流量分布。

Figure 4.4 The annual precipitation and discharge distribution for Longzici spring from 1987 to 2018.

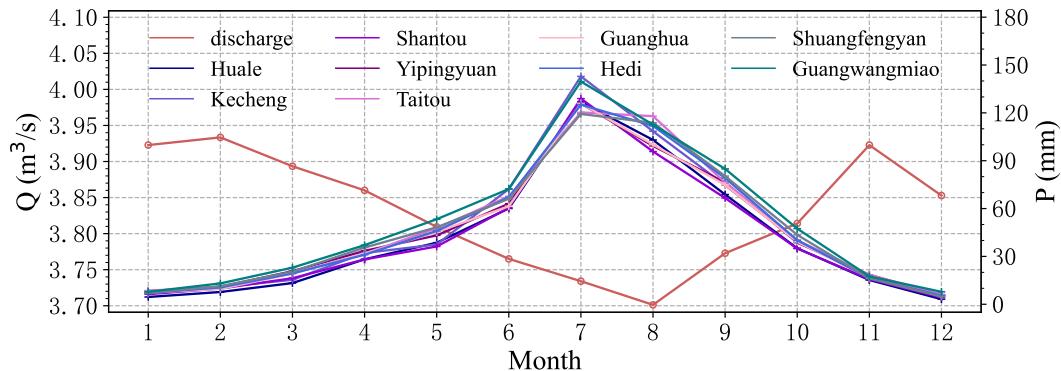


图 4.5 龙子祠泉多年月均泉流量与降水量变化趋势。

Figure 4.5 Monthly mean discharge and precipitation distribution for Longzici spring and its surrounding area.

图4.4描述了泉域年际降水量分布。1987年至2018年期间，降水量变化范围为[283.79 mm, 870.60 mm]，年均降水量为550.97 mm。若以年均降水量为丰枯标准，则枯水年（小于550.97 mm）与丰水年（大于550.97 mm）各占一半。2003年年降水量为870.60 mm，远超出其他年份，结合泉流量波动情况，得知强降水给泉流量带来了重要的水源供给，这也是泉流量在2004年迅速回升的重要原因。图4.5中绘制了泉域九个地区多年月均泉流量与降水量的关系。由图4.5可知，降水量是泉流量变化的重要影响因素，直接给龙子祠泉补充水源，而且降水量对泉

流量的补给响应存在~4个月的滞后。

4.2.3 方法描述

本节围绕预处理方法和机器学习方法展开。未来泉流量走势主要依赖于历史泉流量和降水量，这里需要将数据集转化为监督学习时间序列数据。图4.5表明龙子祠泉流量与降水量之间存在时间上的滞后。未来泉流量不仅依赖于历史1个月泉流量和降水量，还可能依赖历史几个月甚至几年的泉流量和降水量。泉流量的变化趋势可被近似表达为：

$$\begin{aligned} Q(t + \Delta, t + 2\Delta, \dots, t + f\Delta) = & F[P_1(t), P_1(t - \Delta), \dots, P_1(t - (n - 1)\Delta); \dots; \\ & P_9(t), P_9(t - \Delta), \dots, P_9(t - (n - 1)\Delta); \\ & Q(t), Q(t - \Delta), \dots, Q(t - (m - 1)\Delta)]. \end{aligned} \quad \dots (4.1)$$

其中， Δ 代表采样间隔（1个月）。 $m = n$ ，即假设输入的历史降水量和泉流量时间同步。 $Q(t + \Delta, t + 2\Delta, \dots, t + f\Delta)$ 代表未来 f 个月泉流量。 $P_e(t - i\Delta)(i \in \{0, 1, \dots, n - 1\}, e \in \{1, 2, \dots, 9\})$ 代表第 e 区域历史第 i 个月的降水量， $Q(t - j\Delta)(j \in \{0, 1, \dots, m - 1\})$ 代表历史第 j 个月泉流量。式4.1表明，输入是泉域内历史降水量与泉流量，输出为未来泉流量。

龙子祠泉地理条件复杂，泉流量与降水量时间序列均呈现出复杂的非线性特征，而且可用泉流量的记录有限。在小样本情况下，可以适当简化机器学习模型。本章不仅选择了三种神经网络（包括 LSTM-RNN、1DCNN、LSTM-1DCNN），还选择了其他几种简单的机器学习方法（包括 SVR、LR、RF、DT、KNN）。

由图4.5可知，泉流量与降水量之间存在~4个月的滞后，因此这里考虑输入的历史泉流量和降水量的时间窗口长度为1至4个月，评估不同输入时间窗口长度对模型的影响。每个输入响应可以表示如下：

$$\begin{aligned} I_{-k} = & [Q(t), Q(t - 1), \dots, Q(t - k + 1); P_1(t), P_2(t), \dots, P_9(t); \dots; \\ & P_1(t - k + 1), P_2(t - k + 1), \dots, P_9(t - k + 1)]. \end{aligned} \quad \dots (4.2)$$

其中， $k = 1, 2, 3, 4$ 。

通常地下水资源按半季节至季节进行规划，因此这里将输出时间窗口长度设置为1至4个月。每个输出响应可以表示如下：

$$O_{+f} = [Q(t + 1), Q(t + 2), \dots, Q(t + f)]. \quad \dots (4.3)$$

其中， $f = 1, 2, 3, 4$ 。

通过第2.3.1节的滑动窗口法，可以将原始数据集转换为监督学习数据集。紧接着，将数据集分割为训练集和测试集，分割比例为 0.8: 0.2。在训练集和测试集被使用之前，需要对两者分别进行归一化处理（第2.3.2节）。

4.3 试验结果与分析

在利用机器学习之前，我们首先定义一个基准模型——未来 1 个月泉流量 O_{+1} 等于历史 1 个月泉流量 I_{-1} ($O_{+1} = I_{-1}$)。得到拟合指标 $MSE=0.0301\text{ m}^3/\text{s}$, $RMSE=0.1736\text{ m}^3/\text{s}$ 。可将该基准模型的拟合结果与机器学习模型进行对比，从而了解机器学习模型拟合的优劣。

这里简要介绍各种机器学习的超参数配置。所有模型的输入节点数由输入时间窗口长度决定，输出节点数由输出时间窗口长度决定。本研究采用了三层的 LSTM-RNN 和 LSTM-1DCNN、四层的 LSTM-1DCNN，这里采取较少的层数是因为数据集只有几百对样本，三到四层的网络足以拟合数据集。因不同网络的独特性质，在设置以下超参数时会有所差异。针对 LSTM-RNN，前两层含 LSTM 神经元，最后一层为全连接层。隐藏层的节点数分别设为 32 和 16，输出层为全连接层。针对 1DCNN，第一层和第二层均为一维卷积层，过滤器个数、过滤器大小、步长分别为 32/16、3、1。卷积层后连接了最大池化层，池化大小为 2，步长为 1。最后一层为全连接层。针对 LSTM-1DCNN，第一层为一维卷积层，过滤器个数、过滤器大小、步长分别为 128、3、1；卷积层后连接了最大池化层，池化大小为 2，步长为 1；第二、三层均含有 LSTM 神经元，神经元个数分别为 64 和 32。最后一层为全连接层。

以下是本章所有神经网络通用的超参数：训练经过 1000 回合；网络每层均会使用 ReLU 激活函数；批量训练的数据量设为 120；Adam 方法作为优化器；学习率初始值设为 10^{-3} ，同时训练次数每增加 10 次，学习率会减小 $1 - 10^{-6}$ 倍；所有试验均使用“早停”方案。

表 4.1 不同模型在不同的输入和输出时间窗口长度下预测泉流量的拟合指标效果。

Table 4.1 The metrics for predicting spring discharge by using different models, input and output window length.

k	模型	O_{+1}		O_{+2}		O_{+3}		O_{+4}	
		MSE	RMSE	MSE	RMSE	MSE	RMSE	MSE	RMSE
1	LSTM	0.0542	0.2328	0.0013	0.0365	0.0017	0.0412	0.0022	0.0466
	1DCNN	0.0542	0.2328	0.0279	0.1669	0.0197	0.1404	0.0020	0.0451
	LSTM-1DCNN	0.0009	0.0297	0.0012	0.0352	0.0015	0.0389	0.0022	0.0474
	SVR	0.0010	0.0310	0.0013	0.0361	0.0016	0.0400	0.0019	0.0433
	LR	0.0011	0.0336	0.0015	0.0390	0.0019	0.0440	0.0023	0.0476
	RF	0.0011	0.0338	0.0030	0.0548	0.0028	0.0528	0.0028	0.0533
	DT	0.0075	0.0868	0.0133	0.1152	0.0079	0.0892	0.0072	0.0846
	KNN	0.0015	0.0389	0.0025	0.0503	0.0032	0.0568	0.0035	0.0593
2	LSTM	0.0542	0.2328	0.0013	0.0362	0.0019	0.0434	0.0019	0.0438
	1DCNN	0.0008	0.0288	0.0276	0.1660	0.0019	0.0436	0.0023	0.0484
	LSTM-1DCNN	0.0013	0.0364	0.0015	0.0391	0.0021	0.0463	0.0026	0.0510
	SVR	0.0010	0.0314	0.0013	0.0360	0.0016	0.0400	0.0018	0.0424
	LR	0.0011	0.0326	0.0016	0.0396	0.0020	0.0452	0.0024	0.0493
	RF	0.0043	0.0658	0.0022	0.0469	0.0028	0.0525	0.0029	0.0541
	DT	0.0138	0.1173	0.0101	0.1007	0.0101	0.1007	0.0077	0.0879
	KNN	0.0039	0.0623	0.0043	0.0654	0.0048	0.0696	0.0051	0.0716
3	LSTM	0.0015	0.0393	0.0017	0.0415	0.0015	0.0390	0.0022	0.0466
	1DCNN	0.0015	0.0389	0.0017	0.0415	0.0024	0.0487	0.0024	0.0487
	LSTM-1DCNN	0.0015	0.0382	0.0020	0.0481	0.0020	0.0450	0.0031	0.0554
	SVR	0.0012	0.0345	0.0015	0.0393	0.0018	0.0422	0.0020	0.0448
	LR	0.0016	0.0402	0.0021	0.0462	0.0025	0.0504	0.0029	0.0534
	RF	0.0010	0.0314	0.0013	0.0360	0.0017	0.0409	0.0021	0.0455
	DT	0.0012	0.0349	0.0031	0.0556	0.0041	0.0642	0.0044	0.0666
	KNN	0.0047	0.0689	0.0053	0.0731	0.0057	0.0755	0.0057	0.0756
4	LSTM	0.0544	0.2334	0.0020	0.0451	0.0022	0.0471	0.0023	0.0475
	1DCNN	0.0020	0.0444	0.0023	0.0475	0.0025	0.0504	0.0035	0.0596
	LSTM-1DCNN	0.0024	0.0492	0.0026	0.0508	0.0028	0.0533	0.0037	0.0610
	SVR	0.0013	0.0365	0.0015	0.0392	0.0019	0.0437	0.0022	0.0474
	LR	0.0024	0.0487	0.0027	0.0519	0.0031	0.0554	0.0034	0.0585
	RF	0.0010	0.0319	0.0014	0.0368	0.0018	0.0420	0.0022	0.0468
	DT	0.0020	0.0442	0.0022	0.0474	0.0036	0.0597	0.0070	0.0838
	KNN	0.0065	0.0808	0.0068	0.0822	0.0068	0.0824	0.0070	0.0835

4.3.1 预测未来1个月泉流量

表4.2 最佳模型预测2019年1月泉流量。

Table 4.2 Predicting the spring discharge in January 2019.

输入时间窗口长度	1	2	3	4
2019年1月泉流量 (m ³ /s)	2.97	2.92	2.99	2.99

本节讨论输出时间窗口长度为1个月的情况。表4.1中第三至四列展示了不同模型、输入和输出时间窗口长度下预测未来1个月泉流量的拟合指标效果。表4.2展示了不同输入时间窗口长度下，利用最佳模型预测未来1个月泉流量变化趋势。当输入时间窗口长度为1个月时，LSTM-1DCNN拟合指标相较于基准模型偏小（MSE=0.0009 m³/s和RMSE=0.0297 m³/s），预测2019年1月泉流量为2.97 m³/s；当输入时间窗口长度为2个月时，1DCNN拟合指标相较于基准模型偏小（MSE=0.0008 m³/s和RMSE=0.0288 m³/s），预测2019年1月泉流量为2.92 m³/s；当输入时间窗口长度为3个月时，RF拟合指标相较于基准模型偏小（MSE=0.0010 m³/s和RMSE=0.0313 m³/s），预测2019年1月泉流量为2.99 m³/s；当输入时间窗口长度为4个月时，RF拟合指标相较于基准模型偏小（MSE=0.0010 m³/s和RMSE=0.0319 m³/s），预测2019年1月泉流量为2.99 m³/s。所有模型在预测2019年1月泉流量时，最大差距为0.06 m³/s。

图4.6绘制了不同输入时间窗口长度下最佳模型预测未来1个月泉流量 O_{+1} 。由图4.6可知，RF具备良好的拟合能力，且训练集中预测值和观测值时间上不存在偏差，测试集中预测值和观测值在时间上存在1个月的偏差；而LSTM-1DCNN和1DCNN训练集和测试集中的预测值和观测值在时间上均存在1个月的滞后。绝大多数的预测值和观测值的绝对误差在0.1 m³/s以内。利用这些最佳模型预测未来1个月（即2019年1月）龙子祠泉流量，均可得到较为可靠的结果。

4.3.2 预测未来2个月泉流量

本节讨论输出时间窗口长度为2个月的情况。表4.1中第五、六列展示了不同模型在输入时间窗口长度下预测未来2个月泉流量的拟合指标效果。表4.3展示了不同输入时间窗口长度下，利用最佳模型预测未来2个月泉流量变化趋势。当输入时间窗口长度为1个月时，1DCNN拟合指标相较于基准模型偏小

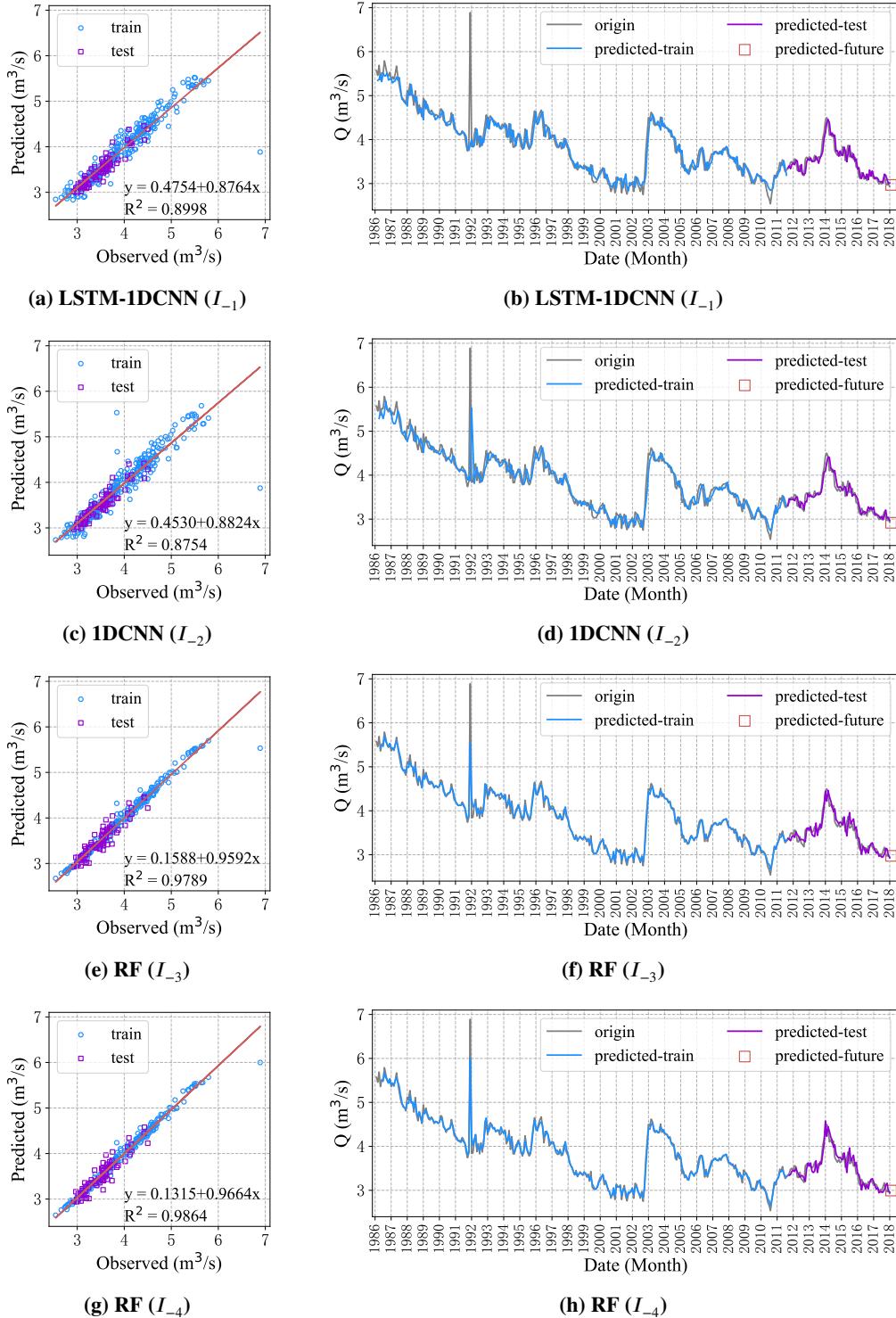


图 4.6 不同输入时间窗口长度下最佳模型预测未来 1 个月泉流量。

Figure 4.6 Predicting the next monthly spring discharge by the best models with different input window length.

表 4.3 最佳模型预测 2019 年 1 月和 2 月泉流量。

Table 4.3 Predicting the spring discharge in January 2019 and February 2019.

输入时间窗口长度	1	2	3	4
2019 年 1 月泉流量 (m^3/s)	2.93	2.98	2.99	3.05
2019 年 2 月泉流量 (m^3/s)	2.94	2.91	2.93	3.07

($\text{MSE}=0.0012 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0352 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $2.93 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $2.94 \text{ m}^3/\text{s}$; 当输入时间窗口长度为 2 个月时, SVR 拟合指标相较于基准模型偏小 ($\text{MSE}=0.0013 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0360 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $2.98 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $2.01 \text{ m}^3/\text{s}$; 当输入时间窗口长度为 3 个月时, RF 拟合指标相较于基准模型偏小 ($\text{MSE}=0.0013 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0363 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $2.99 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $2.93 \text{ m}^3/\text{s}$; 当输入时间窗口长度为 4 个月时, RF 拟合指标相较于基准模型偏小 ($\text{MSE}=0.0014 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0373 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $3.05 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $3.07 \text{ m}^3/\text{s}$ 。

图4.7绘制了不同输入时间窗口长度下最佳模型预测未来 2 个月泉流量 O_{+2} , 这里只展示了每个输入样本的最后一个预测值。由图4.7可知, 对于预测值和测试集, RF 具备良好的拟合能力, 且训练集中预测值和观测值时间上不存在偏差, 测试集中预测值和观测值时间上存在 2 个月的偏差; LSTM-1DCNN 和 SVR 所有的预测值和观测值在时间上存在 2 个月的偏差。绝大多数的预测值和观测值的绝对误差基本在 $0.2 \text{ m}^3/\text{s}$ 以内。利用最佳模型预测未来 2 个月 (即 2019 年 1 月和 2 月) 龙子祠泉流量, 均可得到较为理想的结果。

4.3.3 预测未来 3 个月泉流量

表 4.4 最佳模型预测 2019 年 1 月至 3 月泉流量。

Table 4.4 Predicting the spring discharge from February 2019 to March 2019.

输入时间窗口长度	1	2	3	4
2019 年 1 月泉流量 (m^3/s)	2.98	2.98	3.03	3.06
2019 年 2 月泉流量 (m^3/s)	2.98	2.91	3.04	3.07
2019 年 3 月泉流量 (m^3/s)	2.89	2.90	2.98	2.99

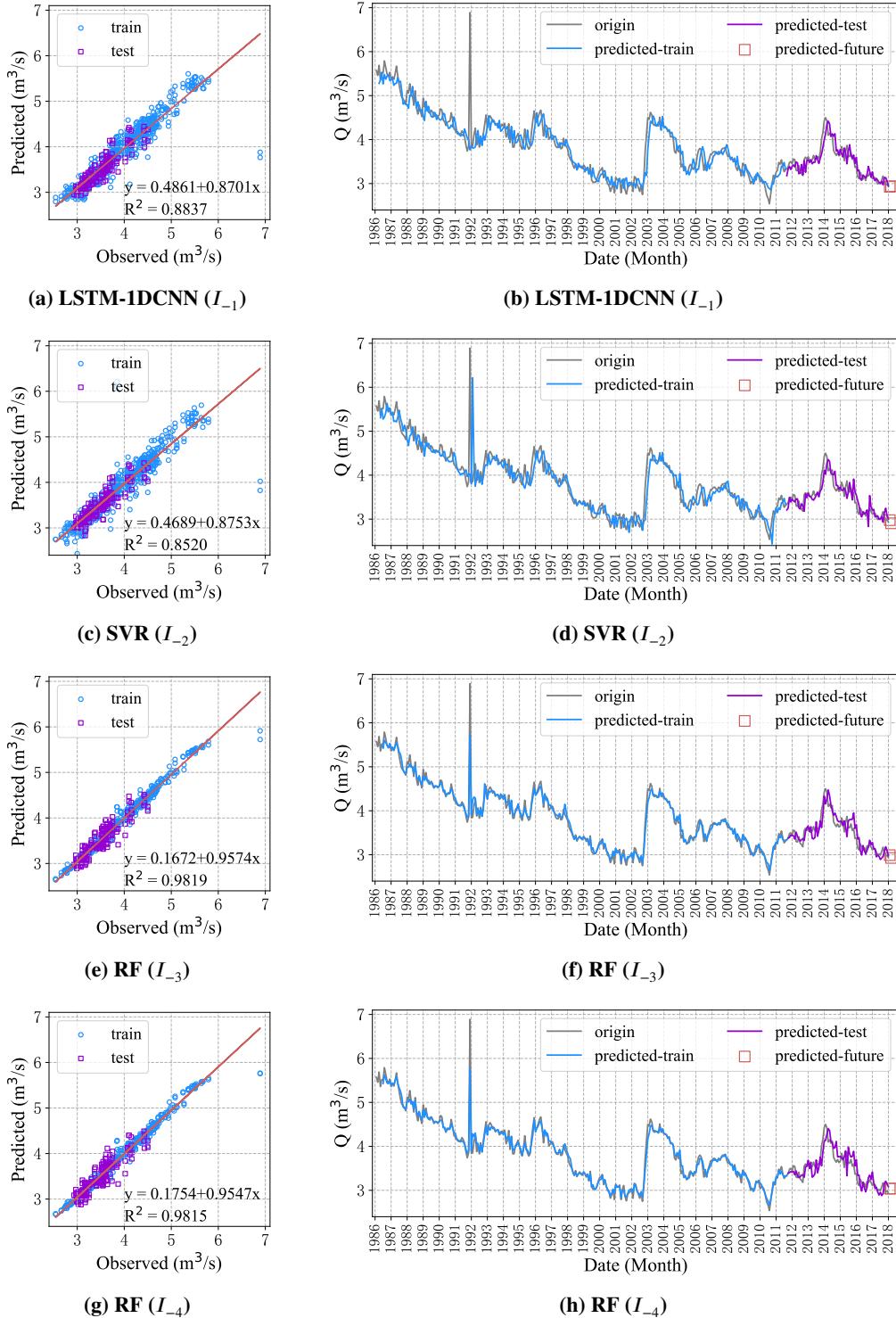


图 4.7 不同输入时间窗口长度下最佳模型预测未来 2 个月泉流量。

Figure 4.7 Predicting the next two monthly spring discharge by the best models with different input window length.

本节讨论输出时间窗口长度为3个月的情况。表4.1中第七、八列展示了不同模型在不同输入时间窗口长度下预测未来3个月泉流量的拟合指标效果。表4.4展示了不同输入时间窗口长度下，利用最佳模型预测未来3个月泉流量变化。当输入时间窗口长度为1个月时，LSTM-1DCNN 拟合指标相较于基准模型偏小（ $MSE=0.0015\text{ m}^3/\text{s}$ 和 $RMSE=0.0389\text{ m}^3/\text{s}$ ），预测2019年1月泉流量为 $2.98\text{ m}^3/\text{s}$ ，2019年2月泉流量为 $2.98\text{ m}^3/\text{s}$ ，2019年3月泉流量为 $2.89\text{ m}^3/\text{s}$ ；当输入时间窗口长度为2个月时，SVR 拟合指标相较于基准模型偏小（ $MSE=0.0013\text{ m}^3/\text{s}$ 和 $RMSE=0.0360\text{ m}^3/\text{s}$ ），预测2019年1月泉流量为 $2.98\text{ m}^3/\text{s}$ ，2019年2月泉流量为 $2.91\text{ m}^3/\text{s}$ ，2019年3月泉流量为 $2.90\text{ m}^3/\text{s}$ ；当输入时间窗口长度为3个月时，LSTM-RNN 拟合指标相较于基准模型偏小（ $MSE=0.0015\text{ m}^3/\text{s}$ 和 $RMSE=0.0390\text{ m}^3/\text{s}$ ），预测2019年1月泉流量为 $2.30\text{ m}^3/\text{s}$ ，2019年2月泉流量为 $2.94\text{ m}^3/\text{s}$ ，2019年3月泉流量为 $2.86\text{ m}^3/\text{s}$ ；当输入时间窗口长度为4个月时，RF 拟合指标相较于基准模型偏小（ $MSE=0.0017\text{ m}^3/\text{s}$ 和 $RMSE=0.0414\text{ m}^3/\text{s}$ ），预测2019年1月泉流量为 $3.06\text{ m}^3/\text{s}$ ，2019年2月泉流量为 $3.07\text{ m}^3/\text{s}$ ，2019年3月泉流量为 $2.99\text{ m}^3/\text{s}$ 。

图4.8绘制了不同输入时间窗口长度下最佳模型预测未来3个月泉流量 O_{+3} ，这里只展示了每个输入样本的最后一个预测值。由图4.8可知，对于预测值和测试集，RF具备良好的拟合能力，且训练集中预测值和观测值时间上不存在偏差，测试集中预测值和观测值时间上存在3个月的偏差；LSTM-1DCNN、SVR 和 LSTM 的预测值和观测值在时间上存在3个月的偏差。大多数的预测值和观测值的绝对误差基本在 $0.3\text{ m}^3/\text{s}$ 以内。利用这些最佳模型预测未来3个月（即2019年1月、2月和3月）龙子祠泉流量，均可得到较为可靠的结果。

4.3.4 预测未来4个月泉流量

本节讨论输出时间窗口长度为4个月的情况。表4.1中第九、十列展示了不同模型在不同输入时间窗口长度下预测未来4个月泉流量的拟合指标效果。表4.5展示了不同输入时间窗口长度下，利用最佳模型预测未来4个月泉流量变化趋势。从表4.5可知，2019年1月至2019年4月泉流量略微呈下降趋势，这是因为冬季降水量很少，无法给泉提供水源供给。

当输入时间窗口长度为1个月时，SVR 拟合指标相较于基准模型偏小（ $MSE=$

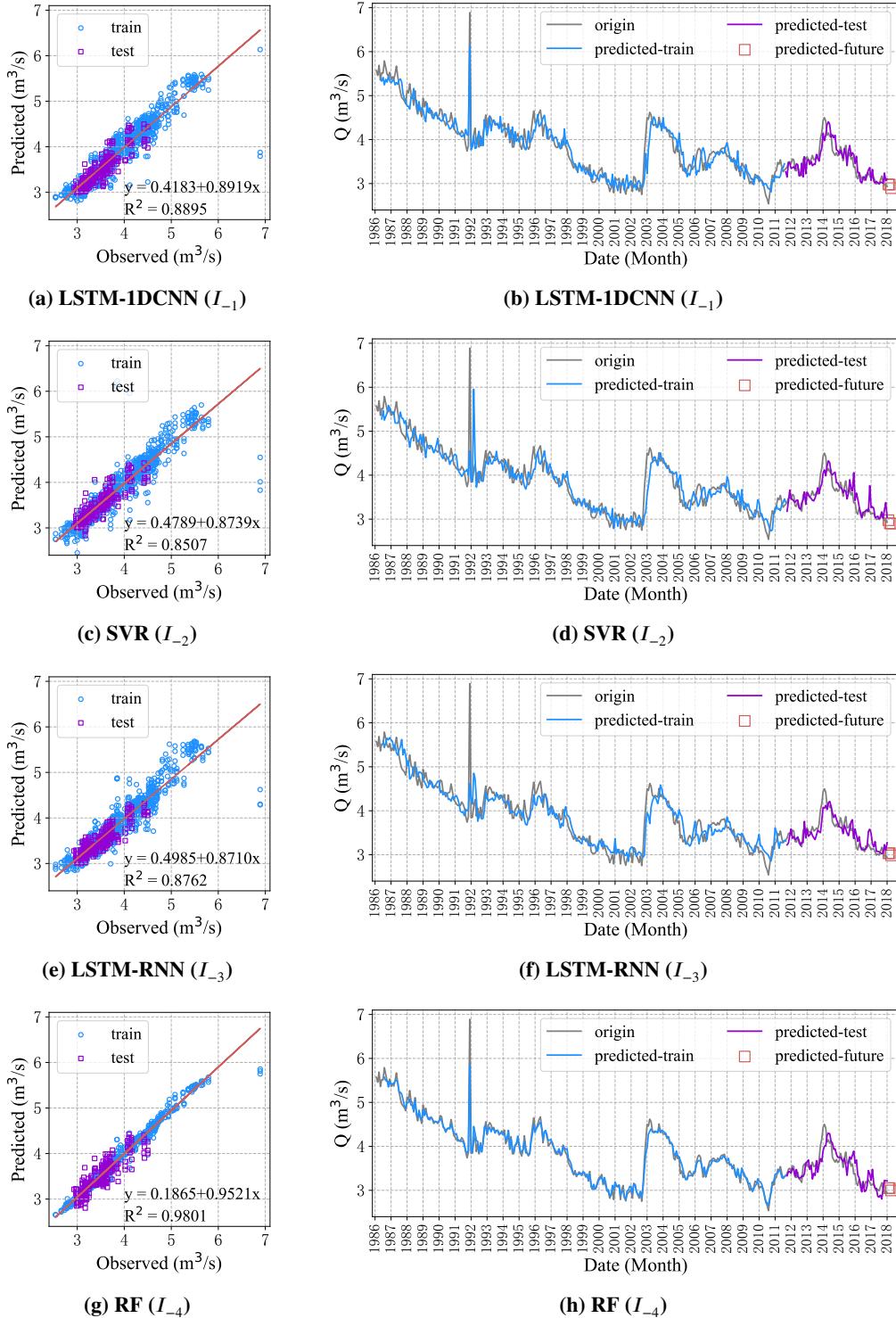


图 4.8 不同输入时间窗口长度下最佳模型预测未来 3 个月泉流量。

Figure 4.8 Predicting the next three monthly spring discharge by the best models with different input window length.

表 4.5 最佳模型预测 2019 年 1 月至 4 月泉流量。

Table 4.5 Predicting the spring discharge from January 2019 to April 2019.

输入时间窗口长度	1	2	3	4
2019 年 1 月泉流量 (m ³ /s)	2.97	2.98	2.30	3.07
2019 年 2 月泉流量 (m ³ /s)	2.91	2.91	2.91	3.01
2019 年 3 月泉流量 (m ³ /s)	2.91	2.90	2.83	2.97
2019 年 4 月泉流量 (m ³ /s)	2.90	2.93	2.78	2.93

0.0019 m³/s 和 RMSE=0.0433 m³/s)，预测 2019 年 1 月泉流量为 2.97 m³/s，2019 年 2 月泉流量为 2.91 m³/s，2019 年 3 月泉流量为 2.91 m³/s，2019 年 4 月泉流量为 2.90 m³/s；当输入时间窗口长度为 2 个月时，SVR 拟合指标相较于基准模型偏小 (MSE=0.0018 m³/s 和 RMSE=0.0424 m³/s)，预测 2019 年 1 月泉流量为 2.98 m³/s，2019 年 2 月泉流量为 2.91 m³/s，2019 年 3 月泉流量为 2.90 m³/s，2019 年 4 月泉流量为 2.93 m³/s；当输入时间窗口长度为 3 个月时，SVR 拟合指标相较于基准模型偏小 (MSE=0.0020 m³/s 和 RMSE=0.0448 m³/s)，预测 2019 年 1 月泉流量为 2.30 m³/s，2019 年 2 月泉流量为 2.91 m³/s，2019 年 3 月泉流量为 2.83 m³/s，2019 年 4 月泉流量为 2.77 m³/s；当输入时间窗口长度为 4 个月时，RF 拟合指标相较于基准模型偏小 (MSE=0.0022 m³/s 和 RMSE=0.0468 m³/s)，预测 2019 年 1 月泉流量为 3.07 m³/s，2019 年 2 月泉流量为 3.01 m³/s，2019 年 3 月泉流量为 2.97 m³/s，2019 年 4 月泉流量为 2.93 m³/s。

图4.9绘制了不同输入时间窗口长度下最佳模型预测未来 4 个月泉流量 O_{+4} ，这里只展示了每个输入样本的最后一个预测值。由图4.9可知，对于预测值和测试集，RF 具备良好的拟合能力，且训练集中预测值和观测值时间上不存在偏差，测试集中预测值和观测值时间上存在 4 个月的偏差；SVR 中所有的预测值和观测值都存在 4 个月的偏差。大多数的预测值和观测值的绝对误差基本在 0.5 m³/s 以内。利用这些最佳模型预测未来 4 个月（即 2019 年 1 月、2 月、3 月和 4 月）龙子祠泉流量，得到泉流量具有一定的可靠性。

4.3.5 利用历史泉流量预测未来泉流量

本节考虑输入中仅有历史泉流量时预测未来泉流量走势情况。当输入和输出时间窗口长度均为 1 个月时，未来 1 个月泉流量为 $Q(t+1) = f(Q(t))$ 。具备一

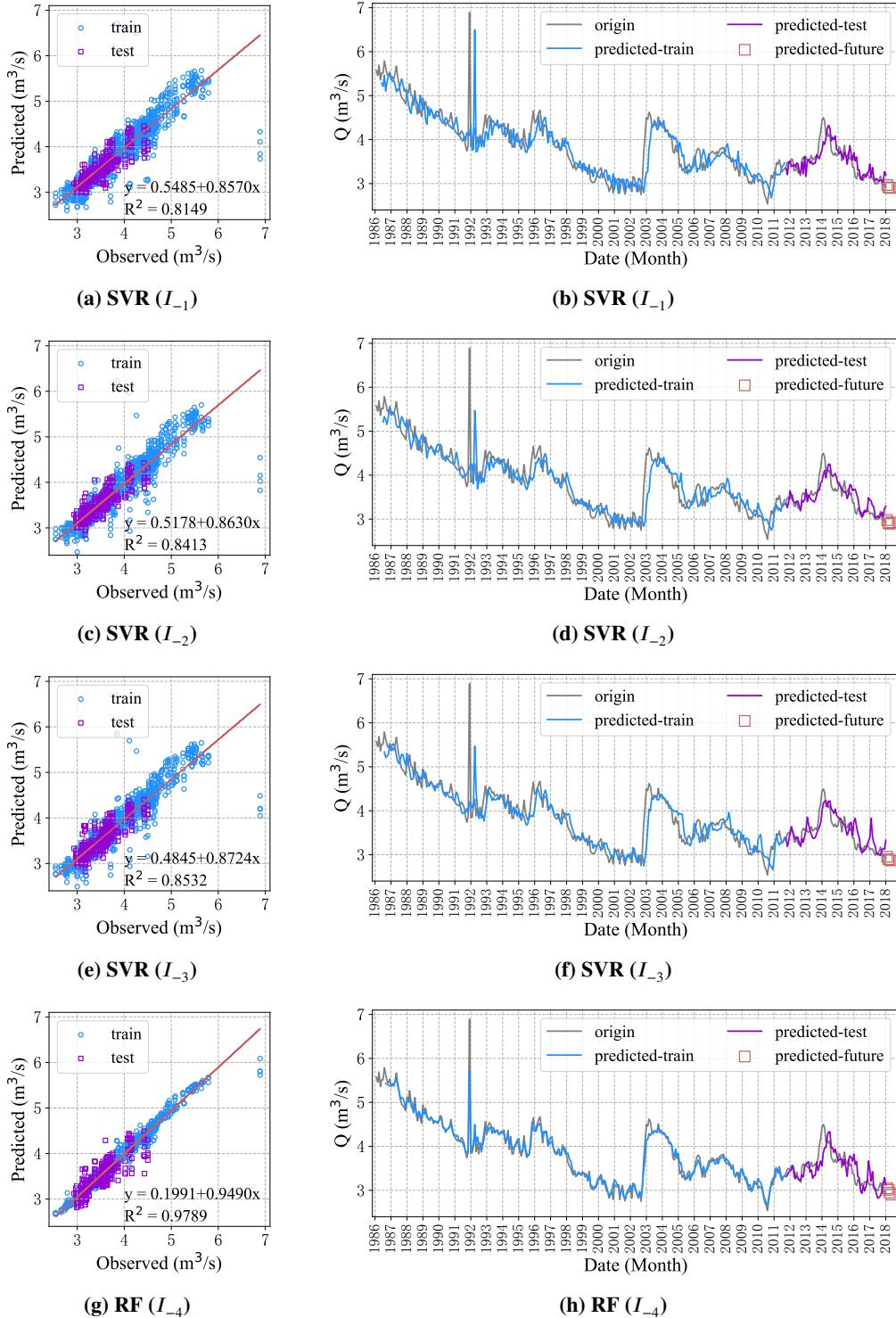


图 4.9 不同输入时间窗口长度下最佳模型预测未来 4 个月泉流量。

Figure 4.9 Predicting the next four monthly spring discharge by the best models with different input window length.

维卷积的神经网络其输入需要具备高维特征，因此本节只使用 LSTM-RNN。经过不断测试，发现 3 层的 LSTM-RNN 最优，且两个隐藏层的 LSTM 单元数分别为 64 和 32。表4.6展示了基于不同模型在历史 1 个月泉流量下预测未来 1 个月泉流量的拟合指标效果，其中 LSTM-RNN 和 SVR 的性能最佳。表4.7展示了不同模型利用历史 1 个月泉流量预测 2019 年 1 月泉流量。

表 4.6 不同模型利用历史 1 个月泉流量预测未来 1 个月泉流量的拟合指标效果。

Table 4.6 The indicators for predicting the next monthly spring discharge by different models with one input window length.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM-RNN	0.0029	0.0540	0.0009	0.0297
SVR	0.0043	0.0659	0.0009	0.0297
LR	0.0043	0.0655	0.0009	0.0306
RF	0.0018	0.0421	0.0014	0.0374
DT	0.0016	0.0400	0.0017	0.0409
KNN	0.0026	0.0506	0.0015	0.0388

表 4.7 不同模型利用历史 1 个月泉流量预测 2019 年 1 月泉流量。

Table 4.7 Predicting the monthly spring discharge in January 2019 by different models with one input window length.

模型	LSTM-RNN	SVR	LR	RF	DT	KNN
2019 年 1 月泉流量 (m^3/s)	2.98	2.97	3.01	2.94	2.97	2.91

讨论输入时间窗口长度为 1 个月且输入中仅含有历史泉流量而不含泉域的降水量的情况。LSTM-RNN 拟合指标相较于基准模型偏小（ $\text{MSE}=0.0009 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0297 \text{ m}^3/\text{s}$ ），预测 2019 年 1 月泉流量为 $2.98 \text{ m}^3/\text{s}$ ；SVR 拟合指标相较于基准模型偏小（ $\text{MSE}=0.0009 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0297 \text{ m}^3/\text{s}$ ），预测 2019 年 1 月泉流量为 $2.97 \text{ m}^3/\text{s}$ ；LR 拟合指标相较于基准模型偏小（ $\text{MSE}=0.0009 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0306 \text{ m}^3/\text{s}$ ），预测 2019 年 1 月泉流量为 $3.01 \text{ m}^3/\text{s}$ ；RF 拟合指标相较于基准模型偏小（ $\text{MSE}=0.0014 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0374 \text{ m}^3/\text{s}$ ），预测 2019 年 1 月泉流量为 $2.94 \text{ m}^3/\text{s}$ ；DT 拟合指标相较于基准模型偏小（ $\text{MSE}=0.0017 \text{ m}^3/\text{s}$ 和

$\text{RMSE}=0.0409 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $2.97 \text{ m}^3/\text{s}$; KNN 拟合指标相较于基准模型偏小 ($\text{MSE}=0.0015 \text{ m}^3/\text{s}$ 和 $\text{RMSE}=0.0388 \text{ m}^3/\text{s}$), 预测 2019 年 1 月泉流量为 $2.91 \text{ m}^3/\text{s}$ 。

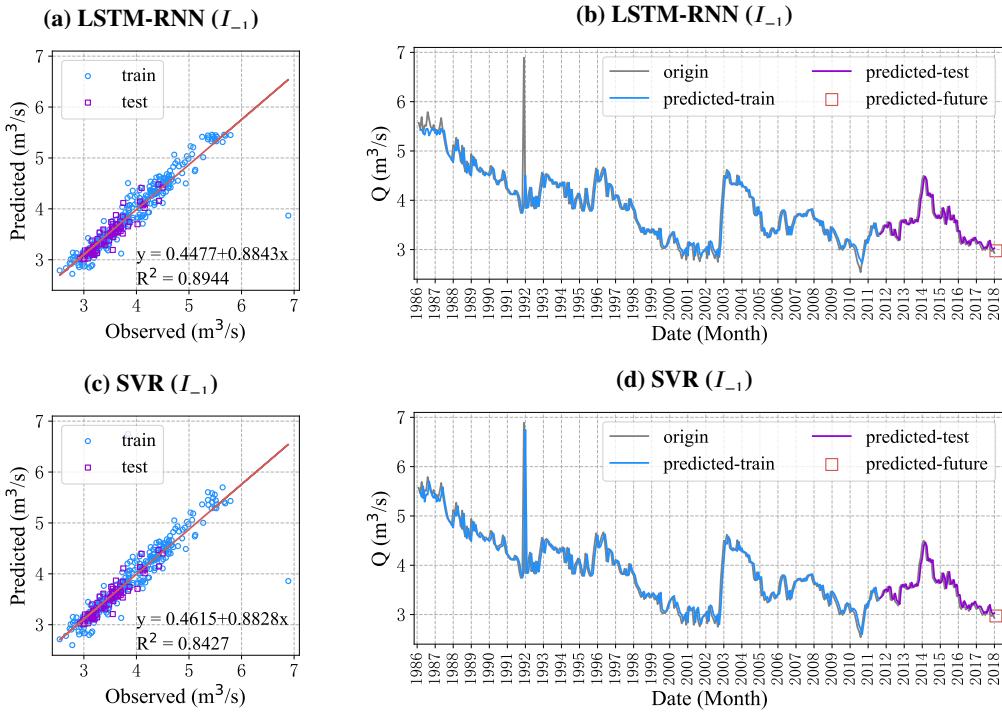


图 4.10 最佳模型利用历史 1 个月泉流量预测未来 1 个月泉流量。

Figure 4.10 Predicting the next monthly spring discharge by the different models with one input window length.

图4.10绘制了最佳模型基于历史 1 个月泉流量预测未来 1 个月泉流量。与图4.6对比, 图4.10中 LSTM-RNN 和 SVR 也能够很好地拟合训练集和测试集, 但训练集中预测值和观测值时间上都存在 1 个月的偏差。因此, 仅仅利用历史 1 个月泉流量就能精确预测未来 1 个月龙子祠的泉流量。这里发现输入中只需要利用泉流量而不需要降水量信息就能够较为精确地估计未来泉流量走势, 可能是降水量随季节变化, 而模型学到了泉流量按季节变化的特征。

4.4 讨论与小结

捕获泉流量动态变化对于管理和规划水资源至关重要。许多情况下, 根据历史观测的资料预测未来泉流量非常复杂。机器学习模型擅长于处理这类问题。本章使用了八种不同的机器学习算法, 分别 LSTM-RNN、1DCNN、LSTM-1DCNN、SVR、LR、RF、DT、KNN, 选择了山西省具备喀斯特地貌特征的龙子祠泉作为

研究案例。结果显示，在时间序列数据集非常有限的情况下，以上所有模型都能为预测提供较为满意的结果。从表4.1得出以下结论：

- (1) 机器学习模型适用于捕获泉流量的动态变化过程。
 - (2) 随着输入时间窗口长度的增加，模型性能不会大幅度提升。该结论与第3.3章研究太阳黑子强度基本类似，即选择合适的输入时间窗口长度很关键。本研究表明输入时间窗口长度为1时，输入数据中含有足够的信息捕捉泉流量的动态变化。若输入时间窗口长度大于某个最佳值（这里为1个月），输入数据中会产生冗余信息，这些信息在模型学习过程中会被自动忽视。极端情况下，冗余信息会使模型产生更多的参数，让模型更难训练，甚至会产生更大的误差。
 - (3) 随着输出时间窗口长度的增加，模型的性能会出现一定幅度的下降。也就是说，预测未来1个月泉流量是最可靠的。预测时间跨度越长，模型的性能会逐渐下降。这说明短期泉流量更多地由历史泉流量和降水量控制，而长期泉流量可能会更多地受到外界其他因素的干扰。在第4.1节提到，泉流量不仅受到历史降水量和泉流量的影响，还会受到其他因素的干扰，比如地下水开采量、入渗、地表径流、蒸散、地下水补给、土壤水分、侧向水流至蓄水层、地表含水层和地下含水层之间的渗漏、蓄水层中蓄水量的变化等。这些因素均没有被考虑到模型中。
 - (4) 输入中仅考虑历史泉流量也能精确预测出未来泉流量。
 - (5) 输出时间窗口长度大小在很大程度上决定了预测值（相对于观测值）的滞后时间。
- 根据以上研究结果，未来机器学习应用于泉流量可能的研究方向如下：
- (1) 研究对象为龙子祠泉，具有喀斯特地貌特征。可以尝试其他研究区域，进一步验证模型是否具备普适性。
 - (2) 时间采样间隔为1个月，如果减小采样间隔时间，检验能否进一步提高模型的性能。
 - (3) 这里尝试了8种不同的机器学习算法，并未涵盖所有机器学习算法，可尝试其它机器学习模型或组合这些模型，查看拟合效果能否得到进一步提升。

第5章 基于机器学习对南加州地区的地震中期预报

本研究围绕着利用机器学习探索时间序列数据展开，第3章基于一种特征（太阳黑子强度），第4章基于两种特征（泉流量和降水量），本章继续增加特征种类，利用机器学习对南加州地区的地震进行中期预报。

5.1 研究背景

过去的几十年里，地球上发生了许多起大震。例如，2008年川滇地区7.9级地震，2011年日本东北9.0级地震，2015年尼泊尔7.8级地震，2017年墨西哥南部海岸8.2级地震，美国南加州7.1级地震等。倘若大震发生在人群居住密集区，可能会导致严重的人员伤亡和大量的财产损失。若能提前预报大震事件，就能够有效地拯救生命和减少财产损失。

地震预报一般涉及到几个要素，包括发生时间、经纬度、震级和发生的概率(Allen, 1976)。地震预报是一个极具挑战性任务(Geller等, 1997)。一个多世纪以来地震学家仍然无法精准预报地震，很大一部分归咎于地震的自组织临界性。根据预报的时间长度，地震预报分为长期、中期和短期预报。本章将重点放在地震中期数值预报。

地震数值预报离不开数据的支撑。数据来自地震目录、地震前兆（比如土壤中氡含量）、三分量地震仪测得的地震波、卫星数据等(Banna等, 2020)。本章利用地震目录衍生出能够反映地震特征的地震因子。随着先进技术的发展，地震目录实现了实时更新，先进的方法和计算能力的提升为地震数值预报提供了新的机遇。

目前，预报地震已发展了很多方法，这些方法可分为由物理驱动的模型和由数据驱动的模型。由物理驱动的模型包括加速动量释放(Ben-Zion等, 2002)、RTL(Region-Time-Length)算法(Sobolev, 2007)、ETAS(Epidemic-Type Aftershock Sequence)(Ogata, 1986)等。由物理驱动的模型会因经验的局限性，难以拟合较为复杂的非线性特征，阻碍了对地震发生机制的进一步了解。

由数据驱动的方法因其强大的提取数据信息的能力，被越来越多地被应用到地震学中(Alves, 2006; Panakkat等, 2007; Madahizadeh等, 2009; Sunkara等,

2009)。Banna等(2020)系统性地总结了84篇科学论文,这些论文基于不同的数据和机器学习方法预报地震。Banna等(2020)指出,被应用在地震学中的机器学习方法可被分为基于规则的方法、浅层机器学习和深层机器学习。基于规则的方法包括模糊逻辑(Zamani等,2013;Mirrashid,2014)和模糊神经网络(López等,2019)。浅层机器学习包括SVR(Asencio-Cortés等,2017)、RF(Asim等,2017)、DT(Asencio-Cortés等,2017)、KNN(Panakkat等,2007;Asencio-Cortés等,2017)、BPNN(Panakkat等,2007;Narayananakumar等,2016)、径向基神经网络(Alexandridis等,2014)、概率神经网络(Adeli等,2009)、聚类(Shodiq等,2018)等。深层神经网络包括RNN(Panakkat等,2009;Asim等,2017)、LSTM-RNN(Wang等,2020)、多层的神经网络(Huang等,2018)等。

地震因子是指与地震发生有关的地球物理信息的特征。已有一些研究提出使用地震因子进行地震预报(Panakkat等,2007;Martínez-Álvarez等,2013;Reyes等,2013;Morales-Esteban等,2013;Asencio-Cortés等,2016)。例如,Reyes等(2013)提出了基于一些地震因子作为地震预报的输入;Martínez-Álvarez等(2013)结合了Panakkat等(2007)中的地震因子,使用特征选择方法分析了地震因子与二元分类(地震发生与否以及是否发生大震)的相关性;Asencio-Cortés等(2016)基于几种不同的监督学习模型,同时利用计算的**b**值调整地震因子,进而提高地震预报的精度;Asim等(2018a)利用集成模型基于地震因子预报未来15天内是否会出现5.0级以上地震。

这项工作的重点是利用几个有监督的机器学习模型,基于各种地震因子作为模型的输入,以提高地震预报的准确性,这些研究关注的是中小震。为了实现预报大震任务,本研究采取了LSTM-RNN(Wang等,2020)和其他几种机器学习算法(SVR、LR、RF、DT、KNN、GBRT、ETR)对地震进行中期预报。本章选择美国南加州地区作为研究区域。该地区属于地震孕育区(地震孕育区是基于地震震源和浅层地质构造背景的空间分布)。南加州地震目录记录的时间长达~90年,是一个较为理想的地震研究区域。

本章结构安排如下。第5.2节介绍了数据与方法,具体包括研究区域介绍、地震目录特征描述、地震因子的计算、时空窗口法、数据的预处理过程等。第5.3节描述地震预报的试验过程,并将试验结果可视化。第5.4节对本章地震预报进行了小结。

5.2 数据与方法

5.2.1 研究区域

加利福尼亚州位于北美和太平洋板块之间，地壳运动频繁，是地震多发地带。南加州地区的地震目录记录时间较长（~90年），且地震记录相对较为完备，故本研究选择南加州地区作为研究区域，地理范围在[32°N–37°N]和[114°W–122°W]。图5.1绘制了南加州地区区域构造和地震目录中记录地震的震级和发生位置。

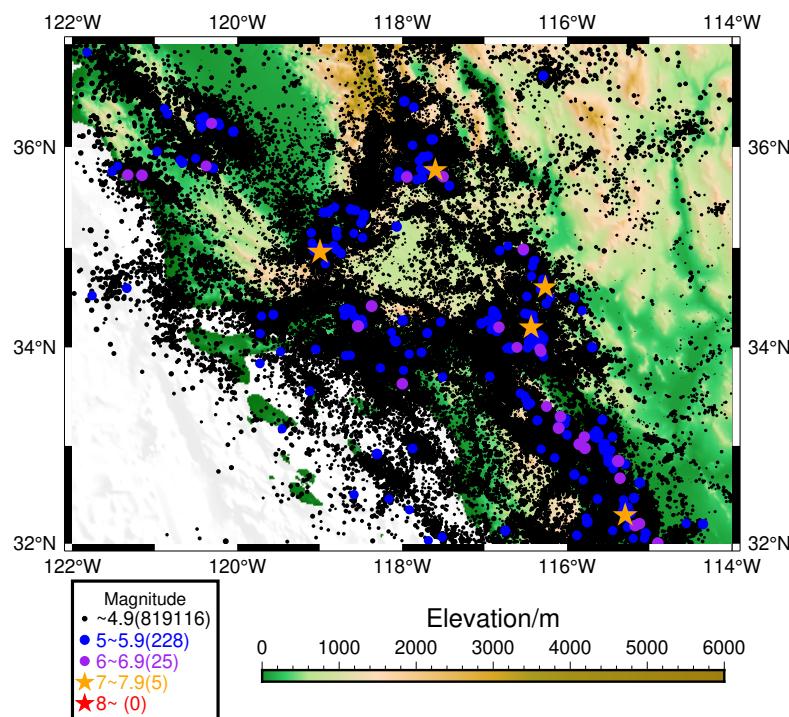


图 5.1 南加州地区区域构造。

Figure 5.1 Tectonics of the Southern California area.

5.2.2 地震目录

表 5.1 1932 年至 2021 年期间研究区域内 7 级以上的地震。

Table 5.1 Earthquake with magnitude no less than 7 in the study area from 1932 to 2021.

发生时间	纬度 (N)	经度 (W)	震级
1952-07-21 11:52:14	34.958	118.998	7.5
1992-06-28 11:57:34	34.200	116.437	7.3
1999-10-16 09:46:44	34.603	116.265	7.1
2010-04-04 22:40:42	32.286	115.295	7.2
2019-07-06 03:19:53	35.770	117.599	7.1

地震目录来自南加州地震数据中心 (Southern California Earthquake Data Center, 简称 SCEDC)¹。地震目录从 1932 年 1 月 1 日开始记录, 2021 年 9 月 19 日截至, 位于研究区域内一共有 831,906 条数据。每条记录包含以下信息:

- (1) 发生时间 (年, 月, 日, 时, 分, 秒);
- (2) 发生地点, 分别为纬度 (N)、经度 (E)、深度 (向下);
- (3) 震级。

在地震预报中, 重点关注大震。表5.1列出研究区域内 7 级以上的地震。由表5.1可知, 近三四十年来, 每隔 ~10 年该地区会发生 7 级以上的地震。

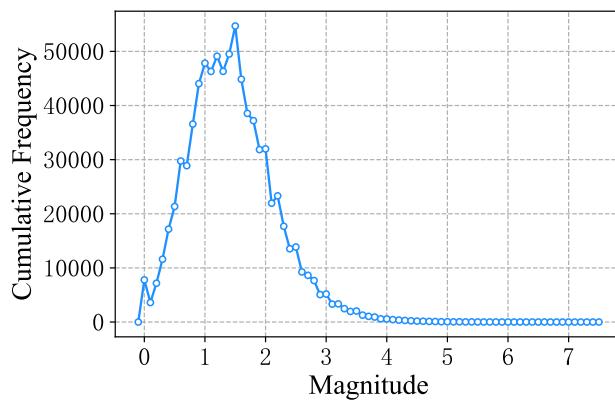


图 5.2 震级与频度的关系。

Figure 5.2 Relationships between magnitude and cumulative frequency.

Gutenberg-Richter 法则是指, 随着震级逐渐增大, 地震频度出现指数级下降 (Asim 等, 2018b)。在给定的研究区域内, 频度和震级需要服从 Gutenberg-Richter 法则 (Gutenberg 等, 1994)。该法则的数学表达形式如下:

$$\log N = a - bM. \quad \dots (5.1)$$

其中, $N > 0$, N 指震级不小于阈震级 M_c 的地震事件的累积数量。 b 值为斜率, 由震级和频度决定。图5.2绘制了南加州地区震级与频度的关系, 发现小震发生的次数远超出大震, 但震级在小于 2 时, 图5.2并不满足 Gutenberg-Richter 法则, 小震存在缺失的情况。

公式5.1可知, b 值反映了震级频度关系, b 值越小表示较大震级的地震发生频次越高。Schorlemmer 等 (2005) 研究发现, 断层类型与 b 值存在显著的关系,

¹ 数据来源: <https://service.scedc.caltech.edu/ftp/catalogs/>

因此 b 值能够体现研究区域的应力水平。 b 值一般可以通过两种方法获取，分别为最小二乘（Least Square，简称 LS）法和最大似然估计（Maximum Likelihood Estimation，简称 MLE）法。图5.3a绘制了 1932 年至 2021 年期间不同起算震级下利用 MLE 法得到的 b 值。图5.3b则绘制了 Gutenberg-Richter 法则的关系图。图5.3a和图5.3b拐角处的震级可看作是阈震级。综合两图的效果，发现从 1932 年至 2021 年期间，南加州地震目录中 3 级以上的地震基本完备，原始地震目录减少到 24,865 条。

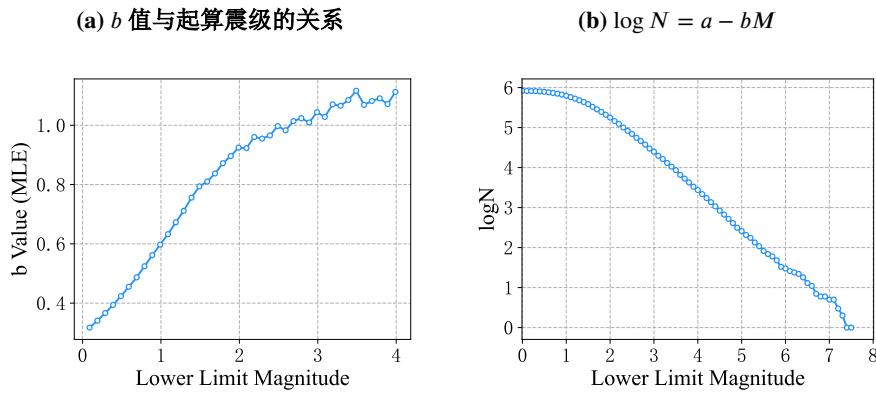


图 5.3 确定 b 值的两种方法。(a) 1932 年至 2021 年期间不同起算震级下利用 MLE 计算的 b 值；(b) 1932 年至 2021 年期间 Gutenberg-Richter 关系式 $\log N = a - bM$ 。

Figure 5.3 The methods of the calculating b value. (a) The b value calculated by MLE between 1932 and 2021; (b) The Gutenberg-Richter relation $\log N = a - bM$ between 1932 and 2021.

5.2.3 地震因子

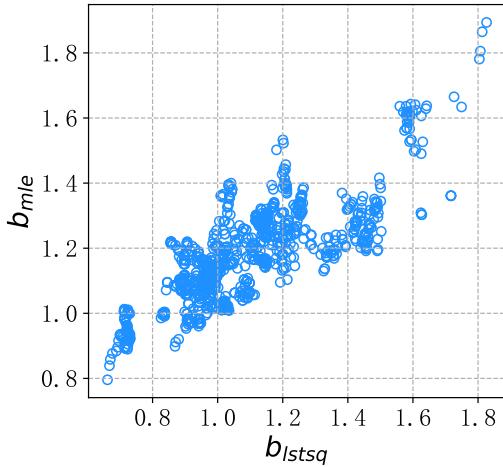
本章目标是预报研究区域内某时间段的最大震级，采取“以震报震”的策略（基于历史地震目录预报未来最大震级）。但仅依靠地震目录预报研究区域内的震级还不够，需要找出地震目录中隐藏的能够反应地震活动特征的信息，这些信息被称作地震因子。地震因子在预测未来震级时至关重要。这些地震因子基于地震目录计算得来。

上节已经提到， b 值的重要性和计算方法。图5.4绘制了 b_{mle} 与 b_{lstsq} 关系图，发现两者的斜率接近于 1。在获取 b 值后，也可得到其他几个与 b 值相关的地震因子，例如 a 值、最小二乘法的均方根误差 η 、最大震级欠缺 ΔM 等。表5.2描述了 17 个不同的地震因子，利用前 16 个因子作为输入预测最后 1 个因子。地震因子不仅包括 Panakkat 等 (2007) 使用的最大震级 M_{max} 、频度 frequency、Gutenberg-

表 5.2 基于地震目录的地震因子。

Table 5.2 Indicators related with earthquake catalog.

因子	公式	描述
1 M_{\max}		输入时间窗口内的最大震级
2 frequency		地震发生次数
3 M_{mean}	$M_{\text{mean}} = \frac{\sum_i M_i}{\text{frequency}}$	输入时间窗口内的平均震级
4 b_{lssq}	$b_{\text{lssq}} = \frac{\log_e(\sum_i M_i \log N_i) - \sum_i M_i \sum_i \log N_i}{n \sum_i (M_i \log N_i)^2 - n \sum_i M_i^2}$	由最小二乘法计算得到 Gutenberg-Richter 法则中的 b 值
5 b_{mle}	$b_{\text{mle}} = \frac{\log_e(\sum_i M_i) - M_c}{\sum_i (\log N_i + b M_i)}$	由最大似然估计法计算得到 Gutenberg-Richter 法则中的 b 值
6 a_{lssq}	$a_{\text{lssq}} = \frac{n}{\Delta M} = \frac{a_{\text{lssq}}}{\Delta M}$	由最小二乘法计算得到 Gutenberg-Richter 法则中的 a 值
7 ΔM	$\Delta M = M_{\max, \text{observed}} - \frac{a_{\text{lssq}}}{b_{\text{lssq}}}$	最大震级欠缺
8 η	$\eta = \sqrt{\frac{\sum_i (\log N_i - (a_{\text{lssq}} - b_{\text{lssq}}) M_i)^2}{n}}$	基于最小二乘法产生的均方根误差
9 \sqrt{E}	$\sqrt{E} = \sum_i \sqrt{E_i}, E_i = 10^{11.8+1.5M_i}$	地震能量平方根 (Asim 等, 2017)
10 Lat_{mean}	$\text{Lat}_{\text{mean}} = \frac{\sum_i \text{Lat}_i}{\text{frequency}}$	平均纬度
11 RMSE_{Lat}	$\text{RMSE}_{\text{Lat}} = \sqrt{\frac{\sum_i (\text{Lat}_i - \text{Lat}_{\text{mean}})^2}{\text{frequency}}}$	纬度的均方根误差
12 Lon_{mean}	$\text{Lon}_{\text{mean}} = \frac{\sum_i \text{Lon}_i}{\text{frequency}}$	平均经度
13 RMSE_{Lon}	$\text{RMSE}_{\text{Lon}} = \sqrt{\frac{\sum_i (\text{Lon}_i - \text{Lon}_{\text{mean}})^2}{\text{frequency}}}$	经度的均方根误差
14 $\text{Lat}_{\sqrt{E}}$	$\text{Lat}_{\sqrt{E}} = \frac{\sum_i \text{Lat}_i \sqrt{E_i}}{\sum_i \sqrt{E_i}}$	按能量加权计算的震中平均纬度
15 $\text{Lon}_{\sqrt{E}}$	$\text{Lon}_{\sqrt{E}} = \frac{\sum_i \text{Lon}_i \sqrt{E_i}}{\sum_i \sqrt{E_i}}$	按能量加权计算的震中平均经度
16 k	$\text{Lat} = k \text{Lon} + b$	由最小二乘法计算得到经纬度之间的斜率
17 M_{future}		输出时间窗口内的最大震级

图 5.4 b_{mle} 与 b_{lstsq} Figure 5.4 b_{mle} and b_{lstsq}

Richter 法则中基于最小二乘法计算得到的 a_{lstsq} 值、 b_{lstsq} 值、均方根误差 η 、最大震级欠缺 ΔM 、最大似然估计法得到的 b_{mle} 值、地震能量平方根 \sqrt{E} 和平均震级 M_{mean} ，还包括平均纬度 Lat_{mean} 、纬度的均方根误差 RMSE_{Lat} 、平均经度 Lon_{mean} 、经度的均方根误差 RMSE_{Lon} 、按能量加权计算的震中平均纬度 $\text{Lat}_{\sqrt{E}}$ 、按能量加权计算的震中平均经度 $\text{Lon}_{\sqrt{E}}$ 、用最小二乘法计算的经纬度之间的斜率 k ，这 7 个地震因子能够体现地震发生的空间位置、地震成团或成带分布等。尽管部分因子在一定程度上高度相关，比如 M_{mean} 和 b_{mle} ， b_{mle} 和 b_{lstsq} ， M_{max} 和 \sqrt{E} ，但研究时仍旧保留了这些冗余信息，因为机器学习不一定能够很好地捕获这些冗余信息。

5.3 结果分析

5.3.1 基于 6 个区块预测未来 1 年的最大震级

第2.3节已经介绍了窗口滑动法。在探索太阳黑子和泉流量时均采取了窗口滑动法，将原始观测数据集转化为监督学习数据集。与窗口滑动法有所差异，这里不仅在时间上对数据集进行滑动划分，在空间上也进行了同样的操作。这样做的目的是缩小预测震级的区域范围。因为研究时不仅要关注震级，还要关注该震级发生的空间范围，而且该范围越小越好。

需要注意的是，部分地震因子的计算是基于统计学基础上的，因此每个时空窗口内地震数目需要达到某个值 N 。如果达不到该值，部分地震因子将会出现

表5.3 每个时空窗口至少有30个地震数的时空窗口信息。

Table 5.3 The time and space windows satisfy with earthquake events at least 30.

输入时间窗口长度	区块序号	空间窗口	
		纬度范围	经度范围
6年	1	[35.0, 37.0N]	[114.0W, 118.0W]
	2	[35.0N, 37.0N]	[118.0W, 122.0W]
	3	[33.5N, 35.5]	[114.0W, 118.0W]
	4	[33.5N, 35.5N]	[117.5W, 121.5W]
	5	[32.0N, 34.0N]	[114.0W, 118.0W]
	6	[32.0N, 34.0N]	[117.0W, 121.0W]

极端异常值。为了避免这种情况的出现，将每个时空窗口的地震数限制在至少30个。通常，7级以上的大震前兆会伴有半年至几年不等的前兆信息。因此，为了满足每个时空窗口的地震数至少为30，将输入时间窗口长度设为6年，空间窗口最小可设为纬度 $2^\circ \times$ 经度 4° 。表5.3展示了每个时空窗口至少有30个地震数的时空窗口信息。

时间窗口按月滑动。空间窗口按照表5.3中区块序号进行滑动，即从研究区域的西北角开始向东滑动 4° 。到达区域边缘后，再向南滑动 1.5° 。向南滑动开始时离左边缘 0.5° 向东滑动 3.5° 。到达区域边缘后，再次向南滑动 1.5° 。向南滑动开始时离左边缘 1° 向东滑动 3° 。这样重复几次后，一共得到6个空间区块。每次开始向南滑动时，西边的区块会右移。这是因为研究区域西南角为海域，地震目录中缺失了这部分的数据。根据时空窗口滑动，前16个地震因子和未来最大震级可被近似表达为：

$$\begin{aligned}
M_{\text{future}}^{\text{loc}}(t+T) = & F[M_{\text{max}}^{\text{loc}}(t-\Delta), \text{frequency}^{\text{loc}}(t-\Delta), M_{\text{mean}}^{\text{loc}}(t-\Delta), \\
& b_{\text{lstsq}}^{\text{loc}}(t-\Delta), b_{\text{mle}}^{\text{loc}}(t-\Delta), a_{\text{lstsq}}^{\text{loc}}(t-\Delta), \Delta^{\text{loc}} M(t-\Delta), \\
& \sqrt{E}^{\text{loc}}(t-\Delta), \eta^{\text{loc}}(t-\Delta), \text{Lat}_{\text{mean}}^{\text{loc}}(t-\Delta), \dots (5.2) \\
& \text{RMSE}_{\text{Lat}}^{\text{loc}}(t-\Delta), \text{Lon}_{\text{mean}}^{\text{loc}}(t-\Delta), \text{RMSE}_{\text{Lon}}^{\text{loc}}(t-\Delta), \\
& \text{Lat}_{\sqrt{E}}^{\text{loc}}(t-\Delta), \text{Lon}_{\sqrt{E}}^{\text{loc}}(t-\Delta), k^{\text{loc}}(t-\Delta), M_{\text{future}}^{\text{loc}}(t-\Delta)]
\end{aligned}$$

其中， t 代表当前时刻， Δ 代表历史时间窗口长度（即输入时间窗口长度）， T 代表未来时间窗口长度（即输出时间窗口长度）， loc 代表区块序号， $\text{loc} \in \{1, 2, \dots, 6\}$ 。

$M_{\text{future}}^{\text{loc}}(t+T)$ 代表未来时间窗口长度 T 内第 loc 个区块的最大震级。 $M_{\text{max}}^{\text{loc}}(t-\Delta)$ 代表历史时间窗口长度 Δ 内第 loc 个区块的最大震级。其他变量的含义以此类推。

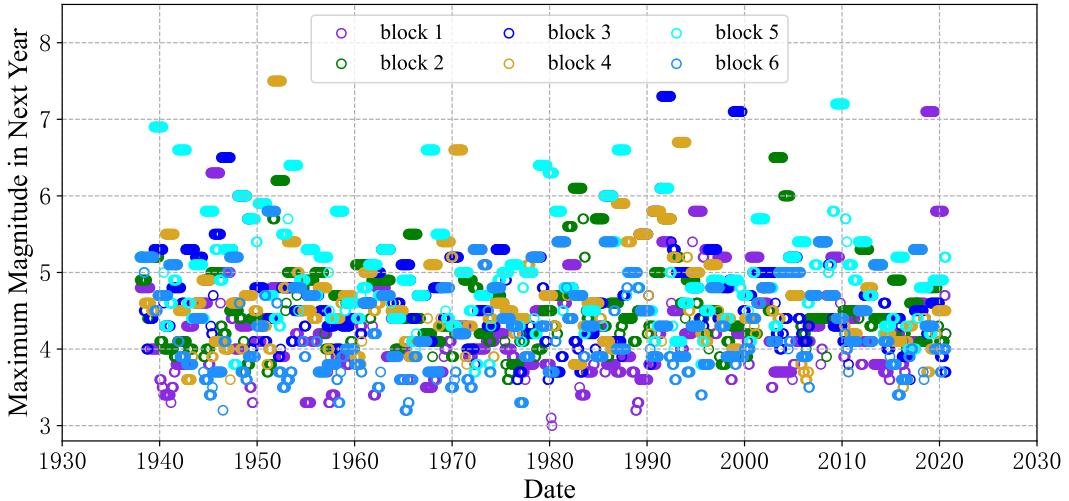


图 5.5 基于 6 个区块预测未来 1 年的最大震级。

Figure 5.5 The maximum magnitude in next year based on six blocks.

式5.2明确指出，模型的输入是历史 16 个地震因子在 6 个不同区块的统计值，输出为未来时间窗口内的最大震级。图5.5绘制了不同区块未来 1 年最大震级。从图5.5得知，该地区未来 1 年最大震级多数位于 3.5 至 5.5 之间。

这里采取了 8 种不同类型的模型，分别为 LSTM-RNN、SVR、LR、RF、GBRT、DT、KNN、ETR。模型输入是 6 个区块在某个特定时刻前 6 年时间窗口长度内的 16 个地震因子，即输入为 $6 \times 16 = 96$ 个，输出是某个区块下一年的最大震级。数据集划分比例为 0.8: 0.2。针对两层的 LSTM-RNN，第一层含 32 个 LSTM 神经元，最后一层为全连接层。网络每层使用 ReLU 激活函数。训练经过 1000 回合。批量训练的数据量设为 128。Adam 方法作为优化器。学习率初始值设为 10^{-3} ，同时训练次数每增加 10 次，学习率会减小 $1 - 10^{-6}$ 倍。所有试验均使用“早停”方案。

将区块分成 6 个子任务进行试验。区块 2-6 的试验结果被放在了附录中（见第A.1.1节和见第A.1.2节）。表5.4和表A.1中展示了 6 个不同区块在不同模型下预报未来 1 年最大震级的拟合指标效果。图5.6、图A.1至图A.5展示了 6 个不同区块拟合未来 1 年最大震级的时间序列，横轴时间 t 代表着未来 $[t, t + 1]$ 年时间段

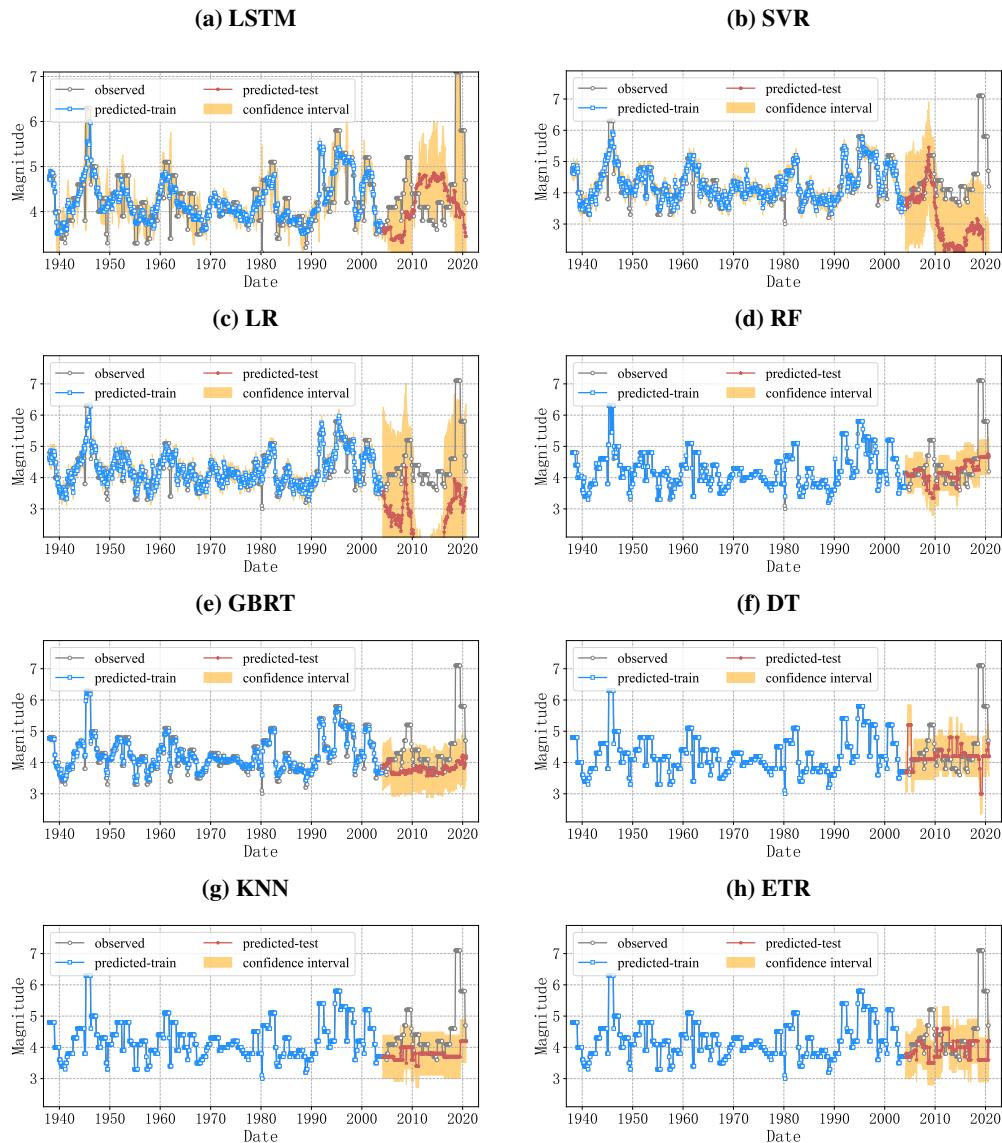


图 5.6 不同模型基于区块 1 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure 5.6 The time series of predicting the maximum magnitute of block 1 with the split ratio 0.8:0.2 in next year by different models.

表 5.4 不同模型基于区块 1 预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8:0.2）。

Table 5.4 The metrics for predicting the maximum magnitude of block 1 with the split ratio 0.8:0.2 in next year by different models.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM	0.0068	0.0826	0.0734	0.2710
SVR	0.0069	0.0832	0.2116	0.4600
LR	0.0054	0.0736	0.5800	0.7616
RF	0.0014	0.0376	0.0452	0.2126
GBRT	0.0012	0.0339	0.0663	0.2574
DT	0.0000	0.0000	0.0558	0.2361
KNN	0.0000	0.0000	0.0673	0.2595
ETR	0.0000	0.0000	0.0778	0.2789

内的最大震级。由表5.4和表A.1、图5.6至图A.5可知，两层的 LSTM-RNN、SVR、LR、RF、GBRT、DT、KNN、ETR 的拟合结果均出现了很大程度的过拟合问题。对于 DT、KNN、ETR 这三种机器学习模型，甚至能够拟合全部的训练集，而对测试集却无能为力。机器学习相对于本数据集而言，过于复杂。

5.3.2 基于整个区块预测未来 1 年的最大震级

鉴于第5.3.1节的结果，这里试图不对区域进行划分滑动，采用时间滑动窗口法获取模型的输入和输出。同样采取了 8 种不同类型的模型。模型输入是整个区块在某个特定时刻前 6 年时间窗口长度内 16 个地震因子，即输入为 16 个，输出是整个区块下一年的最大震级。数据集划分比例为 0.8: 0.2。LSTM-RNN 网络和超参数的设置同上。

表5.5展示了不同模型基于整个区块预报未来 1 年最大震级的拟合指标效果。图5.7展示了不同模型基于整个区块拟合未来 1 年最大震级的时间序列，横轴时间 t 代表着未来 $[t, t + 1]$ 年时间段内的最大震级。由表5.5和图5.7可知，几种模型仍均出现了很大程度的过拟合问题，该结果同第5.3.2节一样。

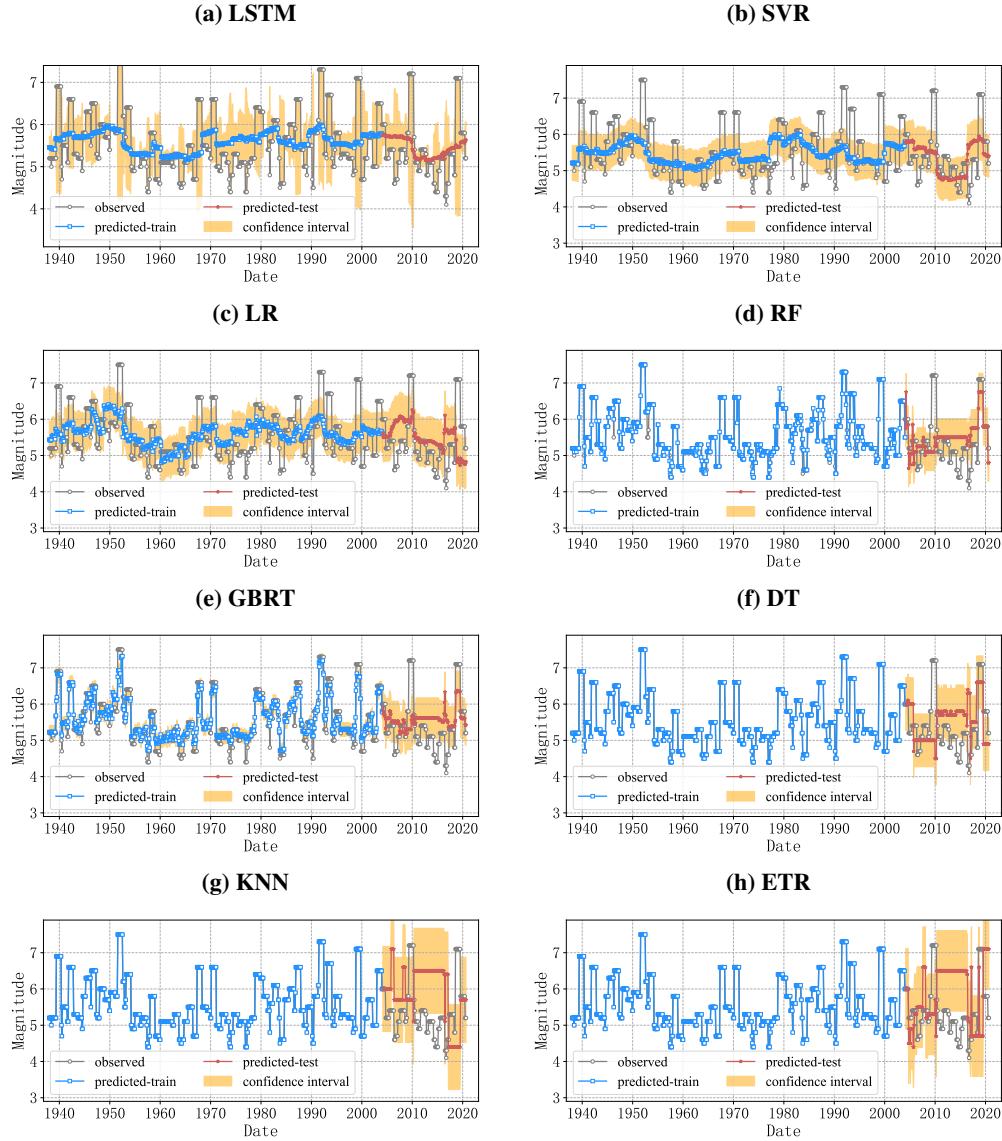


图 5.7 不同模型基于整个区块预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure 5.7 The time series of predicting the maximum magnitude with the split ratio 0.8:0.2 in next year by different models.

表 5.5 不同模型基于整个区块预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）。

Table 5.5 The metrics for predicting the maximum magnitute with the split ratio 0.8:0.2 in next year by different models.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM	0.0367	0.1916	0.0444	0.2108
SVR	0.0396	0.1989	0.0453	0.2127
LR	0.0349	0.1868	0.0636	0.2522
RF	0.0026	0.0511	0.0433	0.2082
GBRT	0.0053	0.0728	0.0451	0.2125
DT	0.0000	0.0000	0.0767	0.2770
KNN	0.0000	0.0000	0.1603	0.4004
ETR	0.0000	0.0000	0.1457	0.3817

5.3.3 基于整个区块预测未来 10 年的最大震级

5.3.3.1 训练集：测试集 =0.8: 0.2

鉴于第5.3.2节的结果，这里试图延长输出时间窗口长度，即预测未来 10 年的最大震级。同样采取了 8 种不同类型的模型。模型输入是整个区块在某个特定时刻前 10 年时间窗口长度内的 16 个预报因子，即输入为 16 个，输出是整个区块未来 10 年的最大震级。数据集划分比例为 0.8: 0.2。LSTM-RNN 网络和超参数的设置同上。

表5.6展示了不同模型基于整个区块预报未来 10 年最大震级的拟合指标效果。图5.8展示了不同模型基于整个区块预报未来 10 年最大震级的时间序列。图5.8显示出未来 10 年最大震级时间序列较为规律，横轴时间 t 代表着未来 $[t - 10, t]$ 年时间段内的最大震级。由表5.6和图5.8可知，几种模型还是出现了很大程度的过拟合问题，该结果同第5.3.2节和第5.3.1节的结论一样。

5.3.3.2 训练集：测试集 =0.85: 0.15

这里数据集划分比例为 0.85: 0.15。表5.7展示了不同模型基于整个区块预报未来 10 年最大震级的拟合指标效果。图5.9展示了不同模型基于整个区块预报未来 10 年最大震级的时间序列。由表5.7和图5.9可知，几种机器学习模型仍均出现

表 5.6 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）。

Table 5.6 The metrics for predicting the maximum magnitute with the split ratio 0.8:0.2 in next ten years by different models.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM	0.0127	0.1125	0.0409	0.2021
SVR	0.0207	0.1439	0.0164	0.1280
LR	0.0114	0.1067	0.0320	0.1789
RF	0.0004	0.0201	0.1288	0.3589
GBRT	0.0001	0.0099	0.2016	0.4490
DT	0.0000	0.0000	0.2668	0.5165
KNN	0.0000	0.0000	0.0265	0.1627
ETR	0.0000	0.0000	0.2285	0.4781

表 5.7 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数据集划分比例为 0.85: 0.15）。

Table 5.7 The metrics for predicting the maximum magnitute with the split ratio 0.85:0.15 in next ten years by different models.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM	0.0136	0.1168	0.0469	0.2167
SVR	0.0212	0.1455	0.0052	0.0723
LR	0.0121	0.1100	0.0250	0.1581
RF	0.0008	0.0287	0.1339	0.3660
GBRT	0.0001	0.0110	0.1269	0.3562
DT	0.0000	0.0000	0.0287	0.1693
KNN	0.0000	0.0000	0.0265	0.1627
ETR	0.0000	0.0000	0.1021	0.3196

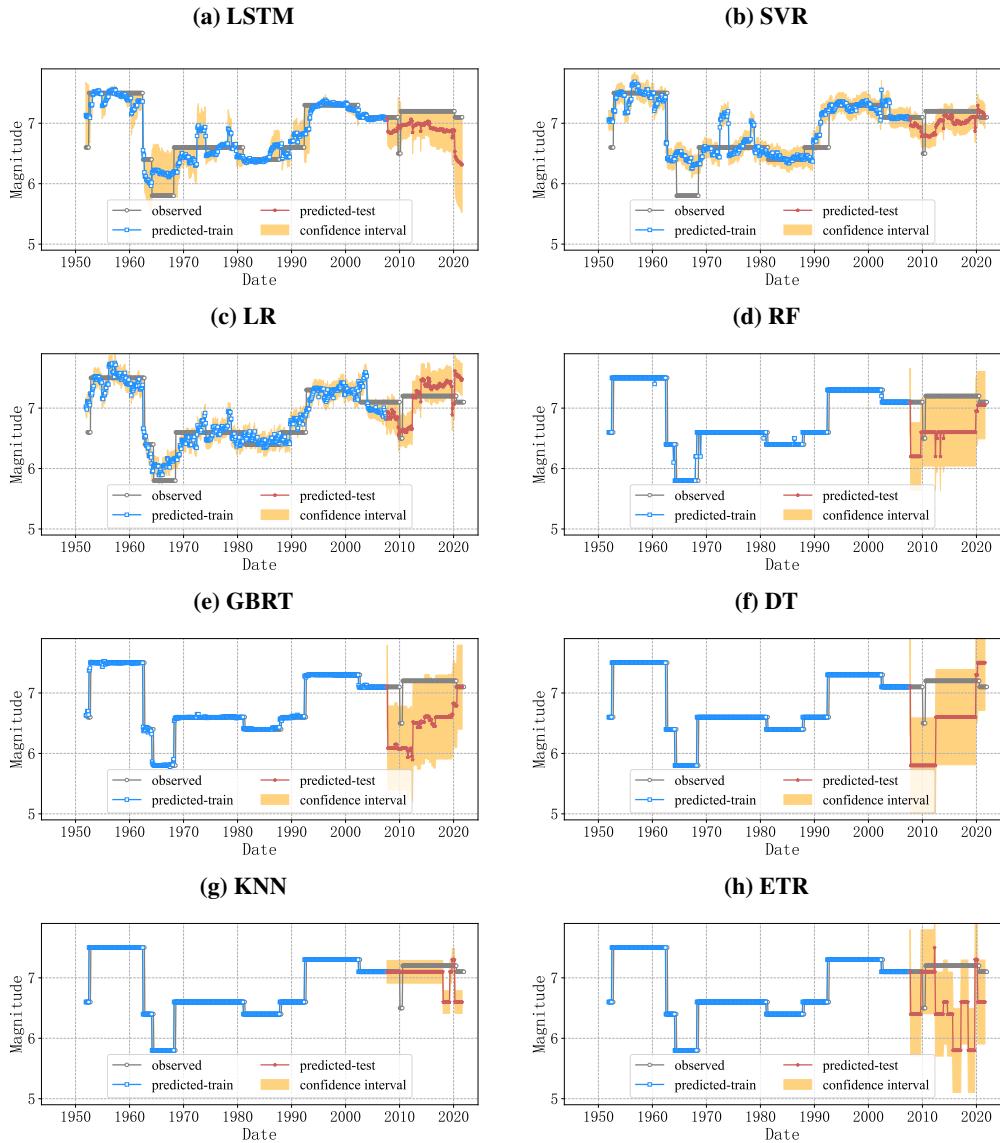


图 5.8 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure 5.8 The time series of predicting the maximum magnitude with the split ratio 0.8:0.2 in next ten years by different models.

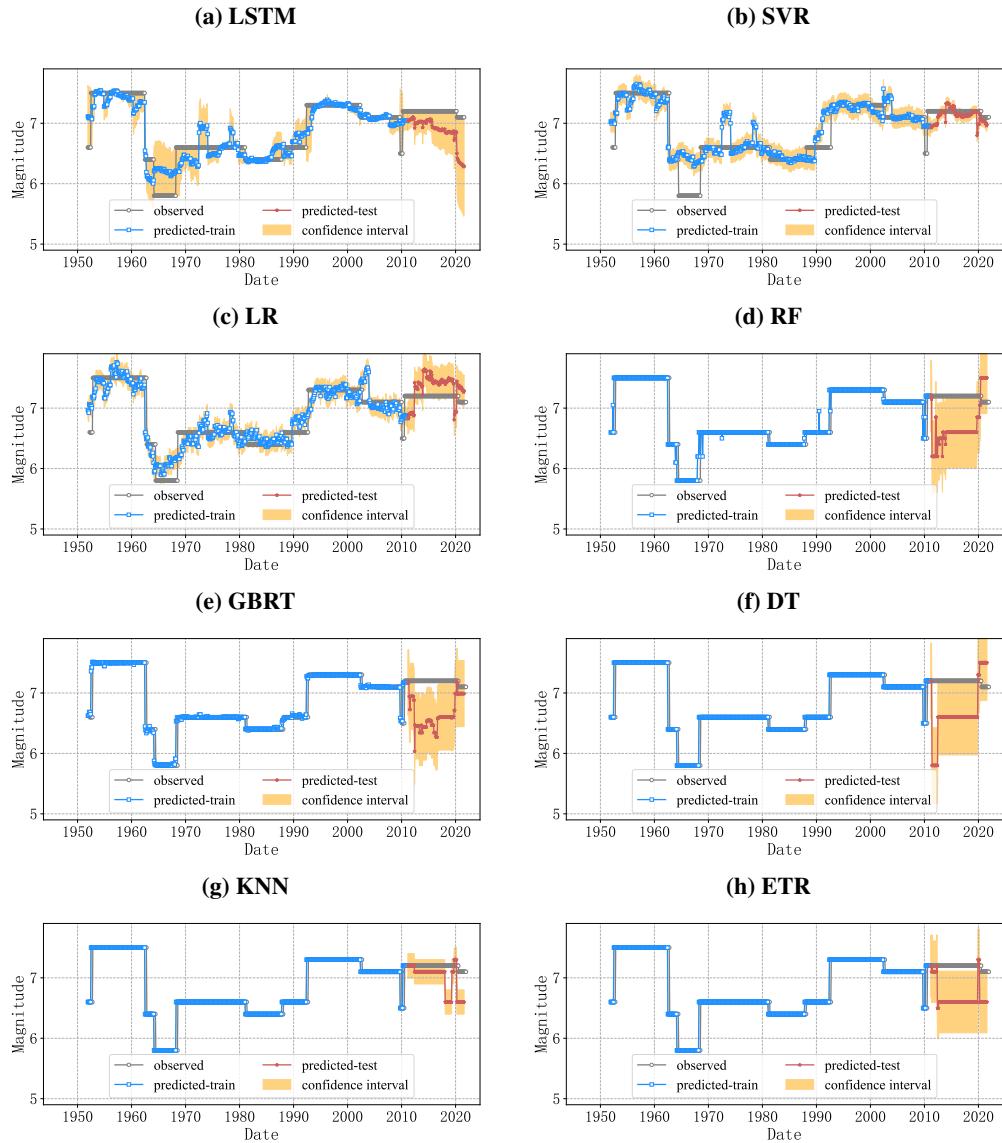


图 5.9 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.85: 0.15）。

Figure 5.9 The time series of predicting the maximum magnitude with the split ratio 0.85:0.15 in next ten years by different models.

了很大程度的过拟合问题。相对而言，KNN 和 SVR 能够较好地拟合测试集数据，但容易受到数据扰动的影响。

5.3.3.3 训练集：测试集 =0.9: 0.1

表 5.8 不同模型基于整个区块预测未来 10 年最大震级的拟合指标效果（数据集划分比例为 0.9: 0.1）。

Table 5.8 The metrics for predicting the maximum magnitute with the split ratio 0.9:0.1 in next ten years by different models.

模型	训练集		测试集	
	MSE	RMSE	MSE	RMSE
LSTM	0.0210	0.1450	0.0053	0.0728
SVR	0.0212	0.1455	0.0052	0.0723
LR	0.0123	0.1108	0.0074	0.0860
RF	0.0005	0.0225	0.0179	0.1337
GBRT	0.0001	0.0116	0.0068	0.0827
DT	0.0000	0.0000	0.0096	0.0980
KNN	0.0000	0.0000	0.0178	0.1332
ETR	0.0000	0.0000	0.0169	0.1300

这里数据集划分比例为 0.9: 0.1。表5.8展示了不同模型基于整个区块预报未来 10 年最大震级的拟合指标效果。图5.10展示了不同模型基于整个区块预报未来 10 年最大震级的时间序列。由表5.8可知，出现过拟合的模型有 RF、GBRT、DT、KNN、ETR，出现欠拟合的模型有 LSTM-RNN、SVR、LR。从图5.10可以看出，除了 LSTM-RNN，其他几种机器学习模型均能拟合出测试集中 7.1 级的地震。尽管如此，这几种机器学习方法稳健性较差，数据的略微变动就会影响模型性能。

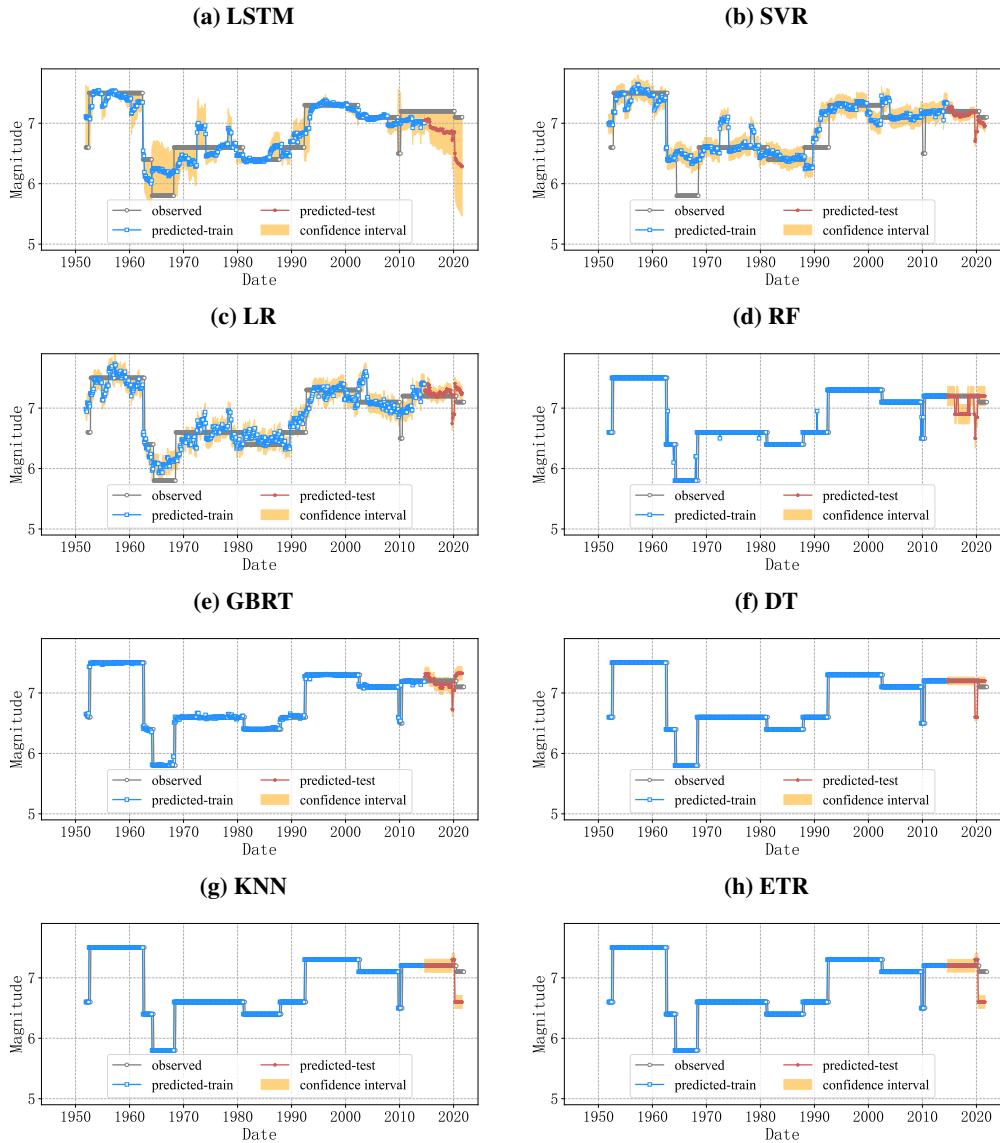


图 5.10 不同模型基于整个区块预测未来 10 年最大震级的时间序列图（数据集划分比例为 0.9: 0.1）。

Figure 5.10 The time series of predicting the maximum magnitude with the split ratio 0.9:0.1 in next ten years by different models.

5.4 讨论与小结

第2.3.6节详细介绍了过拟合产生的原因。模型的泛化误差来源于偏差、方差和噪声。假设输入地震特征因子为 \mathbf{X} , 输出未来时间窗口内的最大震级为 $M_{\text{predicted}}$, 实际发生震级为 M_{observed} , 要拟合的目标函数为 $f(\mathbf{X})$, 训练函数为 $\hat{f}(\mathbf{X})$, 则偏差为:

$$\text{Bais}^2(\mathbf{X}) = E[(\hat{f}(\mathbf{X}) - M_{\text{observed}})^2]. \quad \dots (5.3)$$

高偏差意味着训练集中预测的未来最大震级 $M_{\text{predicted}}$ 与实际发生的最大震级 M_{observed} 差距大, 因此会导致欠拟合的问题。根据第5.3节的结果, 训练集中的 $M_{\text{predicted}}$ 与 M_{observed} 差距较小, 不存在欠拟合问题。偏差越小, 模型越能充分拟合训练数据集。

与偏差不同, 方差度量了训练集的变动导致学习性能的变化, 即刻画了数据扰动所造成的影响。它是由训练样本的小波动敏感而导致的误差。方差可以理解为模型预报震级的变化范围, 可表达为:

$$\text{Variance}(\mathbf{X}) = E[(\hat{f}(\mathbf{X}) - f(\mathbf{X}))^2]. \quad \dots (5.4)$$

高方差意味着训练的模型 $\hat{f}(\mathbf{X})$ 与期望模型 $f(\mathbf{X})$ 差距大, 因此会导致过拟合问题。根据第5.3节的结果, 测试集中的 $M_{\text{predicted}}$ 与 M_{observed} 差距较大, 即方差较大, 模型存在过拟合问题。方差越大, 数据扰动产生的影响越大。可能的原因有: 模型过于复杂, 以至于拟合了训练集中的噪声; 训练样本过少; 输入特征缺乏代表性等。

除了偏差与方差, 噪声也会影响到模型性能的好坏。噪声表达了在预测震级时算法所能达到的期望泛化误差的下界, 刻画了学习问题本身的难度。期望预报 $f(\mathbf{X})$ 与真实值 M_{observed} 之间的误差称为噪声, 即

$$\text{Noise}(\mathbf{X}) = E[(f(\mathbf{X}) - M_{\text{observed}})^2]. \quad \dots (5.5)$$

当存在噪声时, 复杂的模型会尽量覆盖噪声点。即使训练误差很小, 但没有描绘真实的数据趋势, 测试误差反而会更大。如果数据是由未知的某个非常复杂的模型产生的, 实际上有限的数据集也很难去“代表”这个复杂模型。采用不恰当的数据集拟合, 模型性能很难得到提升。一般来讲, 噪声难以避免, 更难以被剔除。训练样本存在噪声的干扰, 导致模型拟合了这些噪声。

从第5.3节来看，无论是哪种机器学习模型，即便采取正则化、Dropout、早停等防止过拟合的策略，过拟合的现象并未好转。因此，模型复杂度不是过拟合产生的根本原因。经过观察发现，部分模型在小震时预报偏大，大震时预报偏小。小震预报偏大可能是因为小震缺失，大震预报偏小可能是统计的时长不够或需要更多的大震数据。数据长度问题暂时难以解决，只能通过时间的积累才能获取更长时间的地震记录。

地震目录中基本不存在噪声。几十年前的地震仪对小震反应可能不灵敏，小震容易被忽视。在地震目录还是由人工记录时，可能会存在误记或漏记。当然，小震不是关注的重点。还有一方面的噪声来自人工地震的影响。但是噪声对本章研究的影响实际上微乎其微。还需要重点考虑训练样本过少、输入特征缺乏代表性等方面。

第6章 总结与展望

本论文关注的是利用机器学习探索时间序列数据。因为机器学习能够较为精确地从大量数据中提取关键信息，因此受到了国内外学术界和工业界越来越多的关注。鉴于机器学习从数据中提取关键信息的优异性能，本论文基于太阳黑子（一类特征）、泉流量（两类特征）、地震数值模拟（多类特征）等时间序列问题展开研究。针对“太阳黑子”，本论文基于神经网络探测未来不同时间窗口长度的太阳黑子强度，这里重点关注第 25 太阳周的太阳黑子峰值；针对“泉流量”，本论文基于多种机器学习模型预测未来不同时期龙子祠泉流量变化趋势；针对“地震数值模拟”，本论文选取了南加州地区作为研究对象，预测未来最大震级。具体内容为：

(1) 基于神经网络预测太阳黑子活动。太阳黑子是太阳活动过程中很容易观测到的特征。太阳内部的强磁场导致太阳黑子的产生，太阳黑子能够反映太阳活动的剧烈程度。目前有关太阳黑子记录长达 400 余年，长期的记录为理解太阳活动机制奠定了基础。太阳黑子的时间序列数据具有非平稳性、非高斯性和非线性等特征，目前预测太阳黑子活动始终具有挑战性。已有研究利用年均或月均太阳黑子数量或面积预测未来太阳黑子活动，但针对未来第 25 太阳周太阳黑子的峰值与过去太阳周的峰值相比是上升还是下降，目前的研究并未给出一致的结论。传统意义上的基于物理机制的模型在拟合太阳黑子活动时会刻意简化模型，从而难以准确捕获太阳黑子时间序列关系。这里采用月均太阳黑子数量和面积（这里的面积是指太阳黑子占太阳可见半球的面积），采纳了不同的输入时间窗口长度和输出时间窗口长度，输入时间窗口长度选择了 72 个月（6 年）、132 个月（11 年）和 264 个月（22 年），输出时间窗口长度选择了 1 个月和 72 个月，基于不同结构的神经网络（LSTM-RNN、1DCNN 和 LSTM-1DCNN），分别进行预测和对照分析。

(a) 讨论输出时间窗口长度为 1 个月的太阳黑子强度。

• 太阳黑子数。历史 72 个月作为输入时间窗口长度所得到的模型均是最优的。LSTM-RNN 的拟合指标 $MSE=0.0030$, $RMSE=0.0548$, 预测 2021 年 9 月太阳黑子数为 38.87; 1DCNN 的拟合指标 $MSE=0.0041$, $RMSE=0.0640$, 预测 2021 年 9

月太阳黑子数为 31.66; LSTM-1DCNN 的拟合指标 $MSE=0.0029$, $RMSE=0.0543$, 预测 2021 年 9 月太阳黑子数为 40.97。

• **太阳黑子面积。**当输入时间窗口长度为 72 个月, LSTM-RNN 的拟合指标 $MSE=0.0021$, $RMSE=0.0454$, 预测 2021 年 9 月太阳黑子面积为 274.14; 当输入时间窗口长度为 132 个月, 1DCNN 的拟合指标 $MSE=0.0028$, $RMSE=0.0531$, 预测 2021 年 9 月太阳黑子面积为 359.78; 当输入时间窗口长度为 132 个月, LSTM-1DCNN 的拟合指标 $MSE=0.0021$, $RMSE=0.0453$, 预测 2021 年 9 月太阳黑子面积为 288.61。

(b) 讨论输出时间窗口长度为 72 个月的最大太阳黑子强度。

• **太阳黑子数。**这里将预测的未来 72 个月最大太阳黑子强度作为第 25 太阳周太阳黑子的峰值。历史 264 个月作为输入时间窗口长度所得到的模型均是最优的。3 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0085$, $RMSE=0.0920$, 预测未来 72 个月太阳黑子数最大值为 151.55, 发生在 2023 年 9 月; 4 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0082$, $RMSE=0.0905$, 预测未来 72 个月太阳黑子数最大值为 174.71, 发生在 2025 年 1 月; 5 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0072$, $RMSE=0.0849$, 预测未来 72 个月太阳黑子数最大值为 132.86, 发生在 2024 年 12 月。对黑子数而言, 第 23 太阳周最大太阳黑子数出现在 2001 年 9 月, 其值为 238.2; 第 24 太阳周最大太阳黑子数出现在 2014 年 2 月, 其值为 146.1。研究结果显示, 第 25 太阳周的峰值跟第 24 太阳周基本持平。

• **太阳黑子面积。**历史 132 个月作为输入时间窗口长度所得到的模型均是最优的。3 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0078$, $RMSE=0.0884$, 预测未来 72 个月太阳黑子面积最大值为 1016.32, 发生在 2024 年 8 月; 4 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0082$, $RMSE=0.0905$, 预测未来 72 个月太阳黑子面积最大值为 1469.01, 发生在 2025 年 3 月; 5 层最佳的 LSTM-1DCNN 的拟合指标 $MSE=0.0072$, $RMSE=0.0849$, 预测未来 72 个月太阳黑子面积最大值为 1397.77, 发生在 2024 年 4 月。对于太阳黑子面积, 第 23 太阳周 MSSA 的峰值为 2171.7, 出现在 2001 年 9 月; 第 24 太阳周 MSSA 的峰值为 1439.82, 出现在 2014 年 2 月。研究结果显示, 第 25 太阳周的峰值跟第 24 太阳周基本持平。

经过多次试验后, 发现模型的性能并不会随着输入时间窗口长度的增加而

逐渐提高。也就是说，输入时间窗口长度存在一个范围，超出该范围时模型性能会有所下降。另外，本研究还发现输出时间窗口长度增加会降低模型的性能，可能的原因是长期来看，时间序列很容易受到外界其他因素的干扰。

(2) 利用机器学习预测龙子祠泉流量变化。预测太阳黑子活动时只考虑了一种特征，探测泉流量变化时我们试图增加输入特征。目前，人类经济活动致使全球气候变得愈加不稳定，人口的增加又迫使生活用水和农业用水需求的激增。全球大部分未冻结的淡水都存储在地下含水层中，地下水的可持续利用有助于人类经济和社会的稳定发展。然而，近些年来，地下水过度开采，井水干枯、溪流和湖泊等淡水区的水量减少、抽水成本日益增加、地面沉降、供水不足等现象时有发生。泉流量是地下水到地表水过渡的指标之一，反映了含水层的动态变化。准确预知泉流量的变化，优化泉水的利用效率，保持合理的农业用水，有助于整个水源的可持续管理。泉流量是一系列动态变化的结果，整个泉域可看作是一个非线性的复杂系统，因为地下水流途径错综复杂。一般来讲，对于泉流量的定量分析需要考虑水文地质特征、地貌特征、土地利用、土地覆盖、人为开采和气候条件等，但这些特征的数据难以收集。

这里研究对象为龙子祠泉，数据涵盖了龙子祠泉流量、其周围九个区域（即化乐、克城、山头、一平垣、台头、光华、河底、双风溝、关王庙）的降水量。龙子祠泉是具有喀斯特地貌的泉域。喀斯特地貌是指地下水和地表水对可溶性岩石腐蚀与沉淀、侵蚀与沉积以及重力崩塌、坍塌、堆积等作用形成的地貌。在溶洞中，地下管道多，水流会形成一个不断扩大的循环网。理论上，基于平衡方程模型可估计未来泉流量。但在实际应用过程中，会经常简化分析。

本论文基于不同的机器学习模型在预测时间窗口长度不超过 4 个月的情况下预测泉流量并进行比较分析。经过分析后，将输入和输出时间窗口长度分别设置为 1 至 4 个月。将不同时间窗口长度的输入数据和输出数据喂给不同的机器学习模型，这些模型包括 LSTM-RNN、1DCNN、LSTM-1DCNN、SVR、LR、RF、DT 和 KNN。

。输出窗口为 1 个月。当输入时间窗口长度为 1 个月时，LSTM-1DCNN 的拟合指标 $MSE=0.0009\text{ m}^3/\text{s}$, $RMSE=0.0297\text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.97\text{ m}^3/\text{s}$; 当输入时间窗口长度为 2 个月时，1DCNN 的拟合指标 $MSE=0.0008\text{ m}^3/\text{s}$, $RMSE=0.0288\text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为

2.92 m³/s；当输入时间窗口长度为3个月时，RF的拟合指标MSE=0.0010 m³/s, RMSE=0.0313 m³/s, 预测2019年1月泉流量为2.99 m³/s；当输入时间窗口长度为4个月时，RF的拟合指标MSE=0.0010 m³/s, RMSE=0.0319 m³/s, 预测2019年1月泉流量为2.99 m³/s。另外，多数模型在预测值和观测值上存在1个月的滞后。

• **输出窗口为2个月。**当输入时间窗口长度为1个月时，1DCNN的拟合指标MSE=0.0012 m³/s, RMSE=0.0352 m³/s, 预测2019年1月泉流量为2.93 m³/s, 2019年2月泉流量为2.94 m³/s；当输入时间窗口长度为2个月时，SVR的拟合指标MSE=0.0013 m³/s, RMSE=0.0360 m³/s, 预测2019年1月泉流量为2.98 m³/s, 预测2019年2月泉流量为2.91 m³/s；当输入时间窗口长度为3个月时，RF的拟合指标MSE=0.0013 m³/s, RMSE=0.0363 m³/s, 预测2019年1月泉流量为2.99 m³/s, 2019年2月泉流量为2.93 m³/s；当输入时间窗口长度为4个月时，RF的拟合指标MSE=0.0014 m³/s, RMSE=0.0373 m³/s, 预测2019年1月泉流量为3.05 m³/s, 预测2019年2月泉流量为3.07 m³/s。另外，多数模型在预测值和观测值上存在2个月的滞后。

• **输出窗口为3个月。**当输入时间窗口长度为1个月时，LSTM-1DCNN的拟合指标MSE=0.0015 m³/s, RMSE=0.0389 m³/s, 预测2019年1月泉流量为2.98 m³/s, 2019年2月泉流量为2.98 m³/s, 2019年3月泉流量为2.89 m³/s；当输入时间窗口长度为2个月时，SVR的拟合指标MSE=0.0013 m³/s, RMSE=0.0360 m³/s, 预测2019年1月泉流量为2.98 m³/s, 2019年2月泉流量为2.91 m³/s, 2019年3月泉流量为2.90 m³/s；当输入时间窗口长度为3个月时，LSTM-RNN的拟合指标MSE=0.0015 m³/s, RMSE=0.0390 m³/s, 预测2019年1月泉流量为3.03 m³/s, 2019年2月泉流量为3.04 m³/s, 2019年3月泉流量为2.98 m³/s；当输入时间窗口长度为4个月时，RF的拟合指标MSE=0.0017 m³/s, RMSE=0.0414 m³/s, 预测2019年1月泉流量为3.06 m³/s, 2019年2月泉流量为3.07 m³/s, 2019年3月泉流量为2.99 m³/s。另外，多数模型在预测值和观测值上存在3个月的滞后。

• **输出窗口为4个月。**当输入时间窗口长度为1个月时，SVR的拟合指标MSE=0.0019 m³/s, RMSE=0.0433 m³/s, 预测2019年1月泉流量为2.97 m³/s, 2019年2月泉流量为2.91 m³/s, 2019年3月泉流量为2.91 m³/s, 2019年4

月泉流量为 $2.90 \text{ m}^3/\text{s}$ ；当输入时间窗口长度为 2 个月时，SVR 的拟合指标 $\text{MSE}=0.0018 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0424 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.98 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $2.91 \text{ m}^3/\text{s}$, 2019 年 3 月泉流量为 $2.90 \text{ m}^3/\text{s}$, 2019 年 4 月泉流量为 $2.93 \text{ m}^3/\text{s}$ ；当输入时间窗口长度为 3 个月时，SVR 的拟合指标 $\text{MSE}=0.0020 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0448 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.30 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $2.91 \text{ m}^3/\text{s}$, 2019 年 3 月泉流量为 $2.83 \text{ m}^3/\text{s}$, 2019 年 4 月泉流量为 $2.78 \text{ m}^3/\text{s}$ ；当输入时间窗口长度为 4 个月时，RF 的拟合指标 $\text{MSE}=0.0022 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0468 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $3.07 \text{ m}^3/\text{s}$, 2019 年 2 月泉流量为 $3.01 \text{ m}^3/\text{s}$, 2019 年 3 月泉流量为 $2.97 \text{ m}^3/\text{s}$, 2019 年 4 月泉流量为 $2.93 \text{ m}^3/\text{s}$ 。另外，多数模型在测试集和其观测值上存在 4 个月的滞后。

- 讨论输入时间窗口长度为 1 个月且输入中仅含有历史泉流量。LSTM-RNN 的拟合指标 $\text{MSE}=0.0009 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0297 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.98 \text{ m}^3/\text{s}$; SVR 的拟合指标 $\text{MSE}=0.0009 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0297 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.97 \text{ m}^3/\text{s}$; LR 的拟合指标 $\text{MSE}=0.0009 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0306 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $3.01 \text{ m}^3/\text{s}$; RF 的拟合指标 $\text{MSE}=0.0014 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0374 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.94 \text{ m}^3/\text{s}$; DT 的拟合指标 $\text{MSE}=0.0017 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0409 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.97 \text{ m}^3/\text{s}$; KNN 的拟合指标 $\text{MSE}=0.0015 \text{ m}^3/\text{s}$, $\text{RMSE}=0.0388 \text{ m}^3/\text{s}$, 预测 2019 年 1 月泉流量为 $2.91 \text{ m}^3/\text{s}$ 。

总体来看，八类不同模型的性能评价指标（ MSE 和 RMSE ）都偏小，因此几种方法都适合预测未来泉流量变化。输入时间窗口长度从 1 个月增加到 4 个月会逐渐降低模型的预测能力，这在某种程度上说明了输入数据存在一定程度上的冗余，这些冗余信息会在拟合过程中被忽视。随着输出时间窗口长度的增加，模型的性能会出现一定幅度的下降。这里输入和输出时间窗口长度对预测泉流量的模型性能的影响同研究太阳黑子类似，即需要找到合适的输入时间窗口长度，而输出时间窗口长度则在目标时间内越短越好。另外，进一步研究发现，仅仅利用历史 1 个月泉流量就能精确预测未来 1 个月龙子祠泉流量，可能是因为降水量随季节变化，而模型学到了泉流量随季节变化的特征。

泉流量不仅受到历史泉流量的影响，还会受到其他因素的干扰，比如地下水开采量、入渗、地表径流、蒸散、地下水补给、土壤水分、侧向水流至蓄水层、

地表含水层和地下含水层之间的渗漏、蓄水层中蓄水量的变化等。这些因素均没有被考虑到模型中，导致所学到的模型总会出现一定程度的偏差。未来机器学习应用于泉流量可能的研究方向为：(1) 研究对象为龙子祠泉，可以尝试其他研究区域，进一步验证机器学习是否具备普适性；(2) 减小采样间隔时间，查看能否进一步提高模型的性能。

(3) **基于机器学习对南加州地区的地震进行中期预报。**因地震发生机制极其复杂，这里我们试图使用多种输入特征预报地震。大震发生很可能会导致严重的人员伤亡和经济损失，影响社会经济的持续发展。为了减小这些损失，预报强震就显得尤为重要。地震预报一般会涉及到时间、地点、震级、可能发生的概率这几个要素。根据预报的时间长度，地震预报分为长期、中期和短期预报。其中，短期预报不可控因素太多，这里不予以考虑。长期预报需要的观测资料至少上百年，目前数据时长不够。因此，地震中期预报成为关注的重点。地震数据来源于美国南加州地区地震目录。从地震目录中计算了 16 种不同的地震因子，采用不同机器学习方法（LSTM-RNN、SVR、LR、RF、DT、KN、GBRT、ETR）对区域内可能发生的最大震级进行了以下几种试验：

- (a) 基于时空窗口滑动法将研究区域划分为 6 个不同区块，数据集划分比例为 0.8: 0.2，预报未来一年的最大震级；
- (b) 基于整个区块，数据集划分比例为 0.8: 0.2，预报未来一年的最大震级；
- (c) 基于整个区块，数据集划分比例分别为 0.8: 0.2、0.85: 0.15 和 0.9: 0.1，预报未来十年的最大震级。

从多数模型的表现来看，它们均出现了很大程度的过拟合问题。当基于整个区块且划分比例为 0.9: 0.1 时，LSTM-RNN、SVR、LR 均出现了欠拟合状态，RF、GBRT、DT、KNN、ETR 均出现了过拟合状态。尽管在这种情况下试验是最优的，但模型极易受到数据集的影响，即数据集的微小变化对模型的性能会产生巨大的影响，模型产生了高方差。这种情况出现的原因可能是选取的数据集长度不够或忽略了某些重要的输入特征。另外，本研究未曾尝试对多个地区、多种类型、多种形式的地震目录，未来可进一步探索以多种前兆观测资料与地震目录相结合的方法，更深入地研究中期地震预报问题。

附录 A 附录

A.1 基于机器学习对南加州地区的地震中期预报

A.1.1 表

表 A.1 不同模型基于区块 2 至区块 6 预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）。

Table A.1 The metrics for predicting the maximum magnitute from block 2 to block 6 with the split ratio 0.8:0.2 in next year by different models.

区块号	模型	训练集		测试集	
		MSE	RMSE	MSE	RMSE
2	LSTM	0.0129	0.1138	0.0290	0.1703
	SVR	0.0162	0.1274	0.0511	0.2261
	LR	0.0118	0.1085	0.2937	0.5419
	RF	0.0045	0.0668	0.0654	0.2557
	GBRT	0.0023	0.0478	0.0666	0.2580
	DT	0.0000	0.0000	0.2584	0.5083
	KNN	0.0000	0.0000	0.0525	0.2290
	ETR	0.0000	0.0000	0.0602	0.2453
3	LSTM	0.0085	0.0920	0.2715	0.5210
	SVR	0.0104	0.1019	0.5286	0.7271
	LR	0.0072	0.0848	0.3738	0.6114
	RF	0.0019	0.0438	0.0546	0.2337
	GBRT	0.0017	0.0412	0.0399	0.1999
	DT	0.0000	0.0000	0.0787	0.2805
	KNN	0.0000	0.0000	0.0843	0.2903
	ETR	0.0000	0.0000	0.0621	0.2491
4	LSTM	0.0101	0.1006	0.0263	0.1623
	SVR	0.0131	0.1143	0.0397	0.1991
	LR	0.0075	0.0863	0.2995	0.5473
	RF	0.0020	0.0447	0.0478	0.2187
	GBRT	0.0018	0.0420	0.0188	0.1371
	DT	0.0000	0.0000	0.0344	0.1856
	KNN	0.0000	0.0000	0.0226	0.1502
	ETR	0.0000	0.0000	0.0418	0.2045

表 A.1 不同模型基于区块 2 至区块 6 预测未来 1 年最大震级的拟合指标效果（数据集划分比例为 0.8: 0.2）(续)。

Table A.1 The metrics for predicting the maximum magnitute from block 2 to block 6 with the split ratio 0.8:0.2 in next year by different models (continued).

区块号	模型	训练集		测试集	
		MSE	RMSE	MSE	RMSE
5	LSTM	0.0173	0.1316	0.3035	0.5509
	SVR	0.0187	0.1367	0.9922	0.9961
	LR	0.0114	0.1068	0.3835	0.6192
	RF	0.0030	0.0549	0.0375	0.1937
	GBRT	0.0032	0.0565	0.0593	0.2435
	DT	0.0000	0.0000	0.0489	0.2212
	KNN	0.0000	0.0000	0.1025	0.3201
	ETR	0.0000	0.0000	0.1060	0.3256
6	LSTM	0.0201	0.1419	0.1112	0.3334
	SVR	0.0208	0.1443	0.2319	0.4816
	LR	0.0140	0.1182	2.9787	1.7259
	RF	0.0043	0.0657	0.0660	0.2569
	GBRT	0.0029	0.0538	0.0438	0.2094
	DT	0.0000	0.0000	0.1217	0.3489
	KNN	0.0000	0.0000	0.0972	0.3118
	ETR	0.0000	0.0000	0.1145	0.3384

A.1.2 图

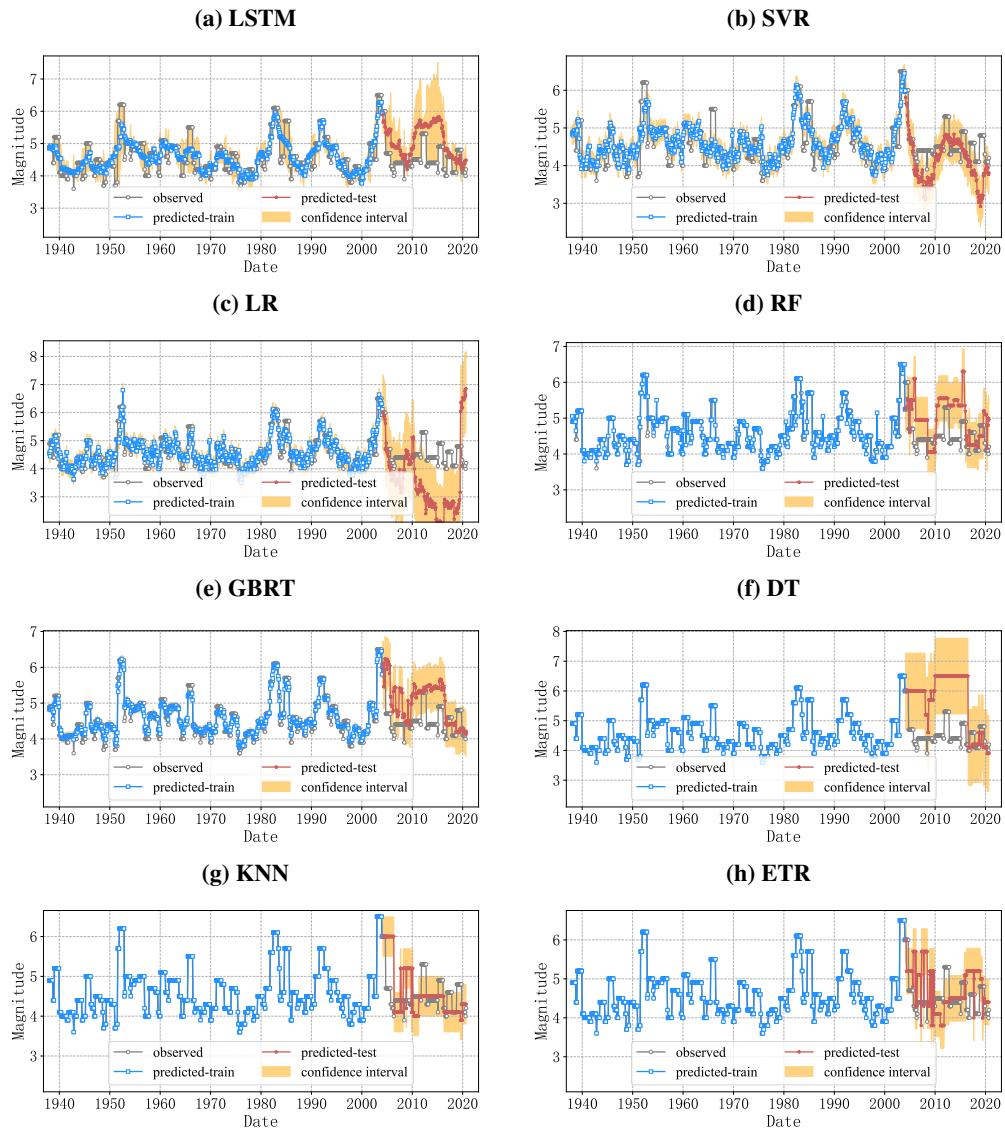


图 A.1 不同模型基于区块 2 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure A.1 The time series of predicting the maximum magnitude of block 2 with the split ratio 0.8:0.2 in next year by different models.

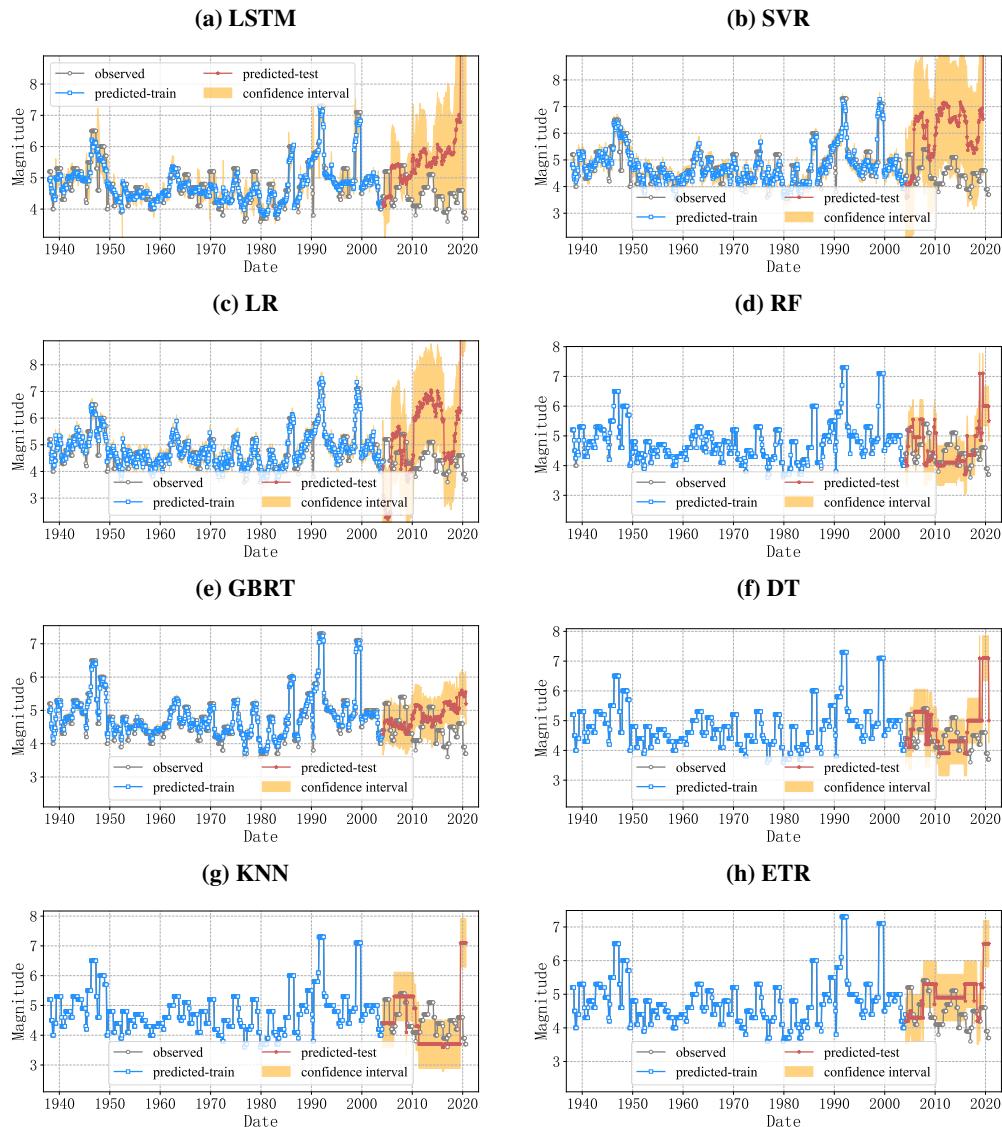


图 A.2 不同模型基于区块 3 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure A.2 The time series of predicting the maximum magnitude of block 3 with the split ratio 0.8:0.2 in next year by different models.

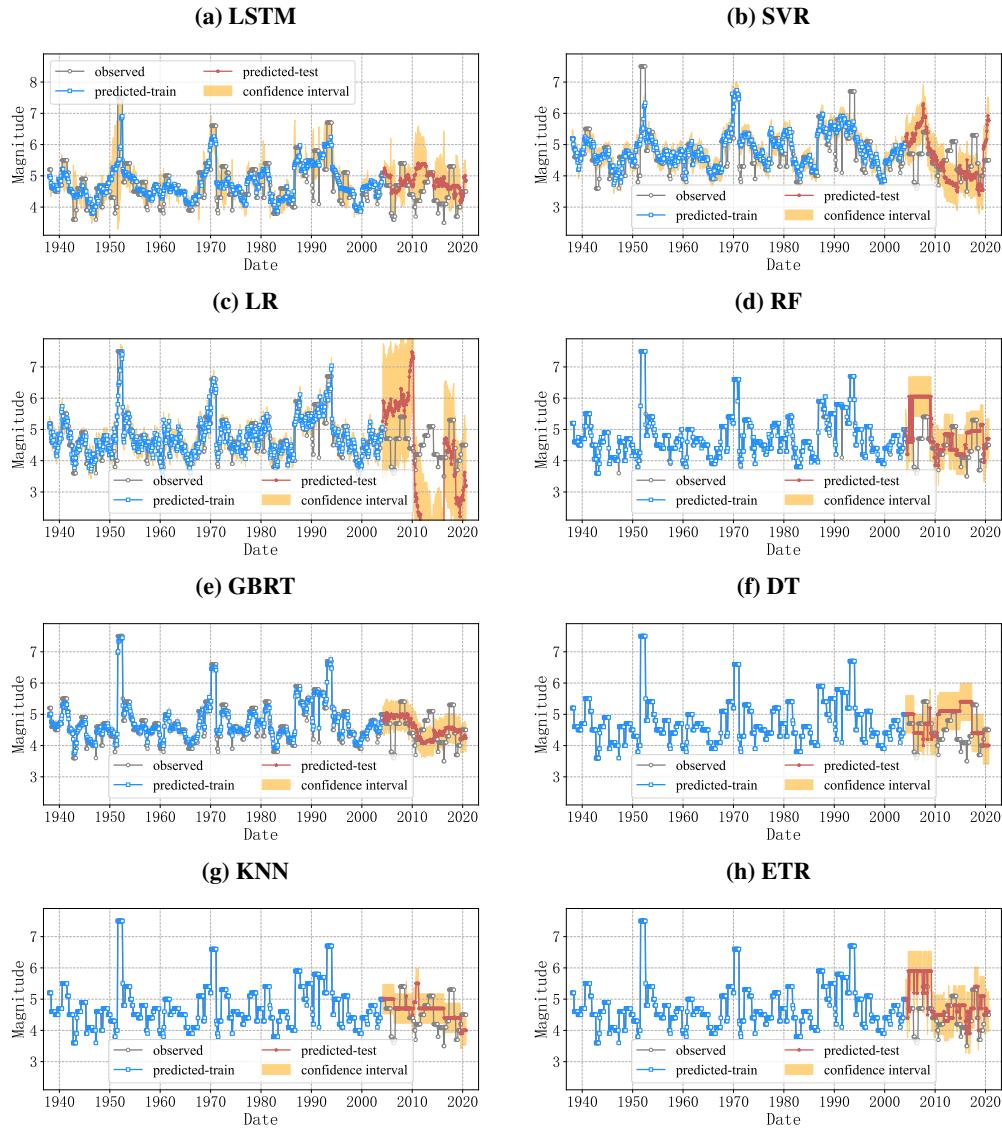


图 A.3 不同模型基于区块 4 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure A.3 The time series of predicting the maximum magnitude of block 4 with the split ratio 0.8:0.2 in next year by different models.

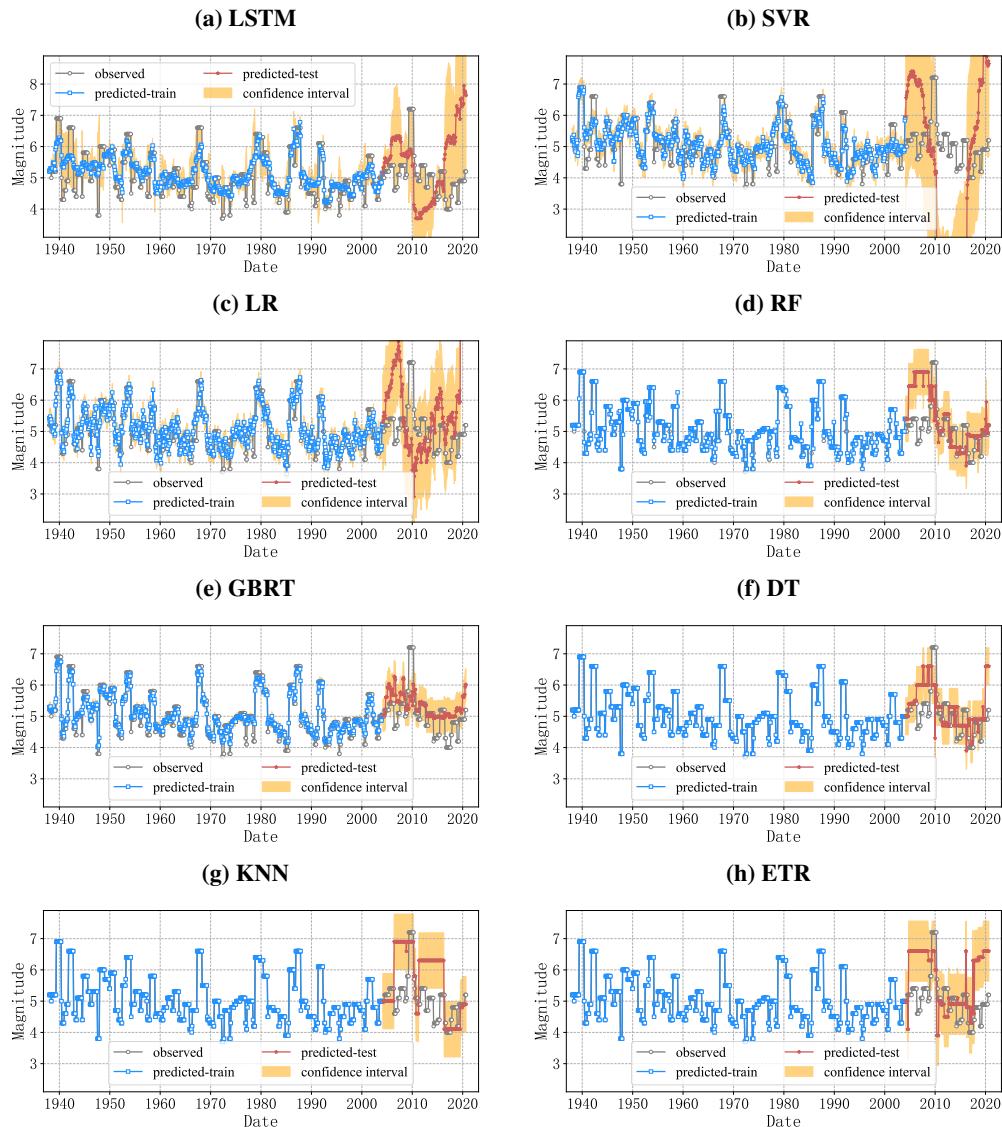


图 A.4 不同模型基于区块 5 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure A.4 The time series of predicting the maximum magnitude of block 5 with the split ratio 0.8:0.2 in next year by different models.

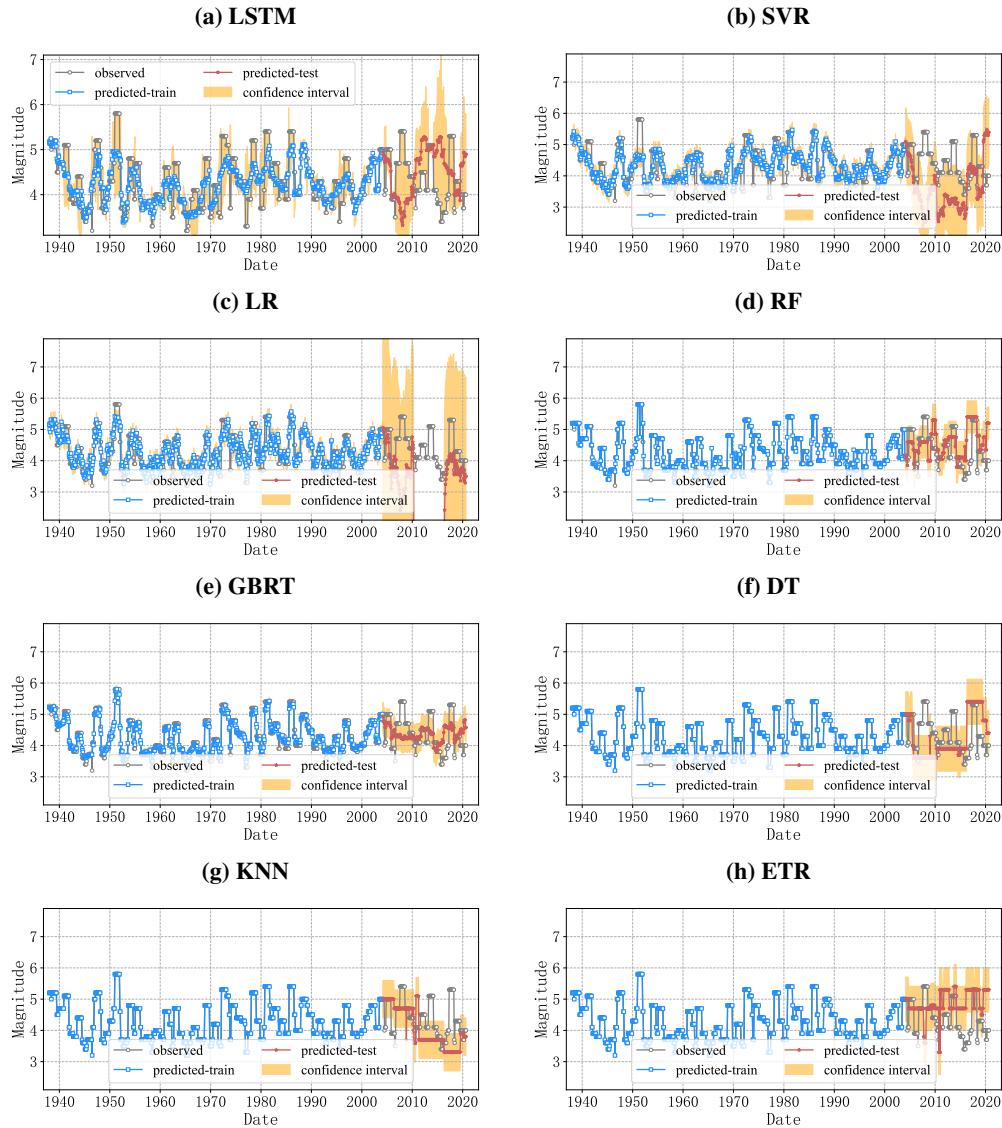


图 A.5 不同模型基于区块 6 预测未来 1 年最大震级的时间序列图（数据集划分比例为 0.8:0.2）。

Figure A.5 The time series of predicting the maximum magnitude of block 6 with the split ratio 0.8:0.2 in next year by different models.

参考文献

- Abrahart R J, Anctil F, Coulibaly P, et al. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting [J]. *Progress in Physical Geography*, 2012, 36(4): 480-513.
- Adeli H, Panakkat A. A probabilistic neural network for earthquake magnitude prediction [J]. *Neural Networks*, 2009, 22(7): 1018-1024.
- Ahmad S, Kalra A, Stephen H. Estimating soil moisture using remote sensing data: A machine learning approach [J]. *Advances in Water Resources*, 2010, 33(1): 69-80.
- Alexandridis A, Chondrodima E, Efthimiou E, et al. Large earthquake occurrence estimation based on radial basis function neural networks [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2014, 52(9): 5443-5453.
- Allen C R. Responsibilities in earthquake prediction: To the seismological society of America, delivered in Edmonton, Alberta, May 12, 1976 [J]. *Bulletin of the Seismological Society of America*, 1976, 66(6): 2069-2074.
- Alves E. Earthquake forecasting using neural networks: Results and future work [J]. *Nonlinear Dynamics*, 2006, 44: 341-349.
- Amaranto A, Munoz-Arriola F, Corzo G, et al. Semi-seasonal groundwater forecast using multiple data-driven models in an irrigated cropland [J]. *Journal of Hydroinformatics*, 2018, 20(6): 1227-1246.
- Amaranto A, Munoz-Arriola F, Solomatine D, et al. A spatially enhanced data-driven multimodel to improve semiseasonal groundwater forecasts in the High Plains aquifer, USA [J]. *Water Resources Research*, 2019, 55(7): 5941-5961.
- Arlt R, Weiss N. Solar activity in the past and the chaotic behaviour of the dynamo [J]. *Space Science Reviews*, 2014, 186: 525-533.
- Asencio-Cortés G, Martínez-Álvarez F, Morales-Esteban A, et al. A sensitivity study of seismicity indicators in supervised learning to improve earthquake prediction [J]. *Knowledge-Based Systems*, 2016, 101: 15-30.
- Asencio-Cortés G, Scitovski S, Scitovski R, et al. Temporal analysis of Croatian seismogenic zones to improve earthquake magnitude prediction [J]. *Earth Science Informatics*, 2017, 10(3): 303-320.
- Asim K M, Martínez-Álvarez F, Basit A, et al. Earthquake magnitude prediction in Hindu Kush region using machine learning techniques [J]. *Natural Hazards*, 2017, 85(1): 471-486.
- Asim K M, Idris A, Iqbal T, et al. Seismic indicators based earthquake predictor system using genetic

- programming and AdaBoost classification [J]. *Soil Dynamics and Earthquake Engineering*, 2018, 111: 1-7.
- Asim K M, Idris A, Iqbal T, et al. Earthquake prediction model using support vector regressor and hybrid neural networks [J]. *PLoS ONE*, 2018, 13(7): e0199004.
- Banna M H A, Taher K A, Kaiser M S, et al. Application of artificial intelligence in predicting earthquakes: State-of-the-art and future challenges [J]. *IEEE Access*, 2020, 8: 192880-192923.
- Ben-Zion Y, Lyakhovsky V. Accelerated seismic release and related aspects of seismicity patterns on earthquake faults [J]. *Pure and Applied Geophysics*, 2002, 159: 2385-2412.
- Bhowmik P, Nandy D. Prediction of the strength and timing of sunspot cycle 25 reveal decadal-scale space environmental conditions [J]. *Nature Communications*, 2018, 9(1): 1-10.
- Bisoi S K, Janardhan P, Ananthakrishnan S. Another mini solar maximum in the offing: A prediction for the amplitude of solar cycle 25 [J]. *Journal of Geophysical Research: Space Physics*, 2020, 125(7): e2019JA027508.
- Bouger B B. Machine learning applications to geophysical data analysis [D]. University of British Columbia, 2016.
- Camporeale E. The challenge of machine learning in space weather: Nowcasting and forecasting [J]. *Space Weather*, 2019, 17(8): 1166-1207.
- Camps-Valls G, Martino L, Svendsen D H, et al. Physics-aware Gaussian processes in remote sensing [J]. *Applied Soft Computing*, 2018, 68: 69-82.
- Charbonneau P. Dynamo models of the solar cycle [J]. *Living Reviews in Solar Physics*, 2010, 7 (1): 1-91.
- Cheng S, Qiao X, Shi Y, et al. Machine learning for predicting discharge fluctuation of a karst spring in North China [J]. *Acta Geophysica*, 2021, 69(1): 257-270.
- Coppola Jr E, Szidarovszky F, Poulton M, et al. Artificial neural network approach for predicting transient water levels in a multilayered groundwater system under variable state, pumping, and climate conditions [J]. *Journal of Hydrologic Engineering*, 2003, 8(6): 348-360.
- Covas E, Peixinho N, Fernandes J. Neural network forecast of the sunspot butterfly diagram [J]. *Solar Physics*, 2019, 294(3): 1-15.
- Dalin C, Wada Y, Kastner T, et al. Groundwater depletion embedded in international food trade [J]. *Nature*, 2017, 543(7647): 700-704.
- DeVries P M, Viégas F, Wattenberg M, et al. Deep learning of aftershock patterns following large earthquakes [J]. *Nature*, 2018, 560(7720): 632-634.
- Dibike Y B, Velickov S, Solomatine D, et al. Model induction with support vector machines: Introduction and applications [J]. *Journal of Computing in Civil Engineering*, 2001, 15(3): 208-216.

- Ding L, Lan R, Jiang Y, et al. Prediction of the smoothed monthly mean sunspot area based on neural network [J]. *Transactions of Atmospheric Sciences*, 2012, 35(4): 508-512.
- Diodato N, Guerriero L, Fiorillo F, et al. Predicting monthly spring discharges using a simple statistical model [J]. *Water Resources Management*, 2014, 28(4): 969-978.
- Du Z. The solar cycle: Predicting the peak of solar cycle 25 [J]. *Astrophysics and Space Science*, 2020, 365(6): 1-5.
- Etemad-Shahidi A, Ghaemi N. Model tree approach for prediction of pile groups scour due to waves [J]. *Ocean Engineering*, 2011, 38(13): 1522-1527.
- Fan J, Meng J, Ludescher J, et al. Statistical physics approaches to the complex earth system [J]. *Physics Reports*, 2021, 896: 1-84.
- Fan Y, Huo X, Hao Y, et al. An assembled extreme value statistical model of karst spring discharge [J]. *Journal of Hydrology*, 2013, 504: 57-68.
- Feyyad U. Data mining and knowledge discovery: Making sense out of data [J]. *IEEE Expert*, 1996, 11(5): 20-25.
- Fiorillo F, Doglioni A. The relation between karst spring discharge and rainfall by cross-correlation analysis (Campania, Southern Italy) [J]. *Hydrogeology Journal*, 2010, 18(8): 1881-1895.
- Friedlingstein P, Meinshausen M, Arora V K, et al. Uncertainties in CMIP5 climate projections due to carbon cycle feedbacks [J]. *Journal of Climate*, 2014, 27(2): 511-526.
- Galelli S, Gandolfi C, Soncini-Sessa R, et al. Building a metamodel of an irrigation district distributed-parameter model [J]. *Agricultural Water Management*, 2010, 97(2): 187-200.
- Geller R J, Jackson D D, Kagan Y Y, et al. Earthquakes cannot be predicted [J]. *Science*, 1997, 275: 1616-1616.
- Gleissberg W. A long-periodic fluctuation of the sun-spot numbers [J]. *The Observatory*, 1939, 62: 158-159.
- Goodfellow I, Bengio Y, Courville A. Deep learning [M]. The MIT Press, 2016.
- Goyal M K. Modeling of sediment yield prediction using M5 model tree algorithm and wavelet regression [J]. *Water Resources Management*, 2014, 28(7): 1991-2003.
- Gutenberg B, Richter C. Frequency of earthquakes in California [J]. *Bulletin of the Seismological Society of America*, 1994, 34: 185-188.
- Guzman S M, Paz J O, Tagert M L M. The use of NARX neural networks to forecast daily groundwater levels [J]. *Water Resources Management*, 2017, 31(5): 1591-1603.
- Hathaway D H. The solar cycle [J]. *Living Reviews in Solar Physics*, 2015, 12(4).
- Herrera V V, Mendoza B, Herrera G V. Reconstruction and prediction of the total solar irradiance: From the medieval warm period to the 21st century [J]. *New Astronomy*, 2015, 34: 221-233.

- Hiremath K. Prediction of solar cycle 24 and beyond [J]. *Astrophysics and Space Science*, 2008, 314(1): 45-49.
- Hochreiter S, Schmidhuber J. Long short-term memory [J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- Hu C, Hao Y, Yeh T C J, et al. Simulation of spring flows from a karst aquifer with an artificial neural network [J]. *Hydrological Processes: An International Journal*, 2008, 22(5): 596-604.
- Huang J, Wang X, Zhao Y, et al. Large earthquake magnitude prediction in Taiwan based on deep learning neural network [J]. *Neural Network World*, 2018, 28(2): 149-160.
- Iglesias A, Garrote L. Adaptation strategies for agricultural water management under climate change in Europe [J]. *Agricultural Water Management*, 2015, 155: 113-124.
- Jiang C, Song F. Sunspot forecasting by using chaotic time-series analysis and NARX network [J]. *Journal of Computers*, 2011, 6(7): 1424-1429.
- Jie T, Xiong Z. Prediction of smoothed monthly mean sunspot number based on chaos theory [J]. *Acta Physica Sinica*, 2012, 61(16).
- Jr. J J B, Whittemore D O, Wilson B B, et al. Sustainability of aquifers supporting irrigated agriculture: A case study of the High Plains aquifer in Kansas [J]. *Water International*, 2018, 43(6): 815-828.
- Kane R. Solar cycle predictions based on extrapolation of spectral components: An update [J]. *Solar Physics*, 2007, 246(2): 487-493.
- Karpatne A, Atluri G, Faghmous J H, et al. Theory-guided data science: A new paradigm for scientific discovery from data [J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(10): 2318-2331.
- Karpatne A, Watkins W, Read J, et al. Physics-guided neural networks (PGNN): An application in lake temperature modeling [J]. arXiv: 1710.11431, 2017.
- Kingma D P, Ba J. Adam: A method for stochastic optimization [J]. arXiv: 1412.6980, 2014.
- Kitiashvili I N. Application of synoptic magnetograms to global solar activity forecast [J]. *The Astrophysical Journal*, 2020, 890(1): 36.
- Kresic N, Stevanovic Z. Groundwater hydrology of springs: Engineering, theory, management and sustainability [M]. Butterworth-heinemann, 2009.
- Lambrakis N, Andreou A S, Polydoropoulos P, et al. Nonlinear analysis and forecasting of a brackish karstic spring [J]. *Water Resources Research*, 2000, 36(4): 875-884.
- Le M H, Perez G C, Solomatine D, et al. Meteorological drought forecasting based on climate signals using artificial neural network — A case study in Khanhhoa Province Vietnam [J]. *Procedia Engineering*, 2016, 154: 1169-1175.

- Lee J, Weger R C, Sengupta S K, et al. A neural network approach to cloud classification [J]. *IEEE Transactions on Geoscience and Remote Sensing*, 1990, 28(5): 846-855.
- Li G, Ma X, Yang H. A hybrid model for forecasting sunspots time series based on variational mode decomposition and backpropagation neural network improved by firefly algorithm [J]. *Computational Intelligence and Neuroscience*, 2018, 2018: 3713410.
- Li Q, Wan M, Zeng S G, et al. Predicting the 25th solar cycle using deep learning methods based on sunspot area data [J]. *Research in Astronomy and Astrophysics*, 2021, 21(7): 184.
- López S, Márquez A, Hernández F A. Evolutionary design of linguistic fuzzy regression systems with adaptive defuzzification in big data environments [J]. *Cognitive Computation*, 2019, 11: 1-12.
- Madahizadeh R, Allamehzadeh M. Prediction of aftershocks distribution using artificial neural networks and its application on the May 12, 2008 Sichuan earthquake [J]. *Journal of Seismology and Earthquake Engineering*, 2009, 11.
- Martínez-Álvarez F, Reyes J, Morales-Esteban A, et al. Determining the best set of seismicity indicators to predict earthquakes. Two case studies: Chile and the Iberian Peninsula [J]. *Knowledge-Based Systems*, 2013, 50: 198-210.
- Mathieu M, Couprie C, LeCun Y. Deep multi-scale video prediction beyond mean square error [J]. arXiv: 1511.05440, 2015.
- McIntosh S W, Chapman S, Leamon R J, et al. Overlapping magnetic activity cycles and the sunspot number: Forecasting sunspot cycle 25 amplitude [J]. *Solar Physics*, 2020, 295(12): 1-14.
- Mendoza B, Velasco-Herrera V M. On mid-term periodicities in sunspot groups and flare index [J]. *Solar Physics*, 2011, 271(1): 169-182.
- Mirrashid M. Earthquake magnitude prediction by adaptive neuro-fuzzy inference system (ANFIS) based on fuzzy C-means algorithm [J]. *Natural hazards*, 2014, 74(3): 1577-1593.
- Mitchell T M, et al. Machine learning [J]. Burr Ridge, IL: McGraw Hill, 1997, 45(37): 870-877.
- Montavon G, Samek W, Müller K R. Methods for interpreting and understanding deep neural networks [J]. *Digital Signal Processing*, 2018, 73: 1-15.
- Morales-Esteban A, Martínez-Álvarez F, Reyes J. Earthquake prediction in seismogenic areas of the Iberian Peninsula based on computational intelligence [J]. *Tectonophysics*, 2013, 593: 121-134.
- Najafzadeh M, Rezaie Balf M, Rashedi E. Prediction of maximum scour depth around piers with debris accumulation using EPR, MT, and GEP models [J]. *Journal of Hydroinformatics*, 2016, 18(5): 867-884.
- Najafzadeh M, Laucelli D B, Zahiri A. Application of model tree and evolutionary polynomial regression for evaluation of sediment transport in pipes [J]. *KSCE Journal of Civil Engineering*, 2017, 21(5): 1956-1963.

- Najafzadeh M, Tafarognoruz A, Lim S Y. Prediction of local scour depth downstream of sluice gates using data-driven models [J]. ISH Journal of Hydraulic Engineering, 2017, 23(2): 195-202.
- Narayananakumar S, Raja K. A BP artificial neural network model for earthquake magnitude prediction in Himalayas, India [J]. Circuits and Systems, 2016, 7: 3456-3468.
- Nayak P C, Rao Y S, Sudheer K. Groundwater level forecasting in a shallow aquifer using artificial neural network approach [J]. Water Resources Management, 2006, 20(1): 77-90.
- Noyes R W. The sun, our star [M]. Harvard University Press, 2013.
- Ogata Y. Statistical models for earthquake occurrences and residual analysis for point processes [J]. Mathematical Seismology, 1986, 1: 228-281.
- Oh J, Guo X, Lee H, et al. Action-conditional video prediction using deep networks in Atari games [J]. arXiv: 1507.08750, 2015.
- Okoh D, Seemala G, Rabiu A, et al. A hybrid regression-neural network (HR-NN) method for forecasting the solar activity [J]. Space Weather, 2018, 16(9): 1424-1436.
- Pala Z, Atici R. Forecasting sunspot time series using deep learning methods [J]. Solar Physics, 2019, 294(5): 1-14.
- Panakkat A, Adeli H. Neural network models for earthquake magnitude prediction using multiple seismicity indicators [J]. International journal of neural systems, 2007, 17(1): 13-33.
- Panakkat A, Adeli H. Recurrent neural network for approximate earthquake time and location prediction using multiple seismicity indicators [J]. Computer-Aided Civil and Infrastructure Engineering, 2009, 24: 280-292.
- Panigrahi S, Pattanayak R M, Sethy P K, et al. Forecasting of sunspot time series using a hybridization of ARIMA, ETS and SVM methods [J]. Solar Physics, 2021, 296(1): 1-19.
- Perol T, Gharbi M, Denolle M. Convolutional neural network for earthquake detection and location [J]. Science Advances, 2018, 4(2): e1700578.
- Pesnell W D. Effects of version 2 of the international sunspot number on naïve predictions of solar cycle 25 [J]. Space Weather, 2018, 16(12): 1997-2003.
- Petrovay K. Solar cycle prediction [J]. Living Reviews in Solar Physics, 2010, 7(1): 1-59.
- Portmann F T, Siebert S, Döll P. MIRCA200-Global monthly irrigated and rainfed crop areas around the year 2000: A new high-resolution data set for agricultural and hydrological modeling [J]. Global Biogeochemical Cycles, 2010, 24(1).
- Raghavendra. N S, Deka P C. Support vector machine applications in the field of hydrology: A review [J]. Applied Soft Computing, 2014, 19: 372-386.
- Rakhshandehroo G R, Akbari A, Igder M, et al. Long term groundwater level forecasting in shallow and deep wells using wavelet neural networks trained by an improved harmony search algorithm [J]. Journal of Hydrologic Engineering, 2018, 23: 04017058.

- Reichstein M, Camps-Valls G, Stevens B, et al. Deep learning and process understanding for data-driven earth system science [J]. *Nature*, 2019, 566(7743): 195-204.
- Reitsma F. Geoscience explanations: Identifying what is needed for generating scientific narratives from data models [J]. *Environmental Modelling & Software*, 2010, 25(1): 93-99.
- Reyes J, Morales-Esteban A, Martínez-Álvarez F. Neural networks to predict earthquakes in Chile [J]. *Applied Soft Computing*, 2013, 13(2): 1314-1328.
- Rigozo N, Echer M S, Evangelista H, et al. Prediction of sunspot number amplitude and solar cycle length for cycles 24 and 25 [J]. *Journal of Atmospheric and Solar-Terrestrial Physics*, 2011, 73 (11-12): 1294-1299.
- Rouet-Leduc B, Hulbert C, Lubbers N, et al. Machine learning predicts laboratory earthquakes [J]. *Geophysical Research Letters*, 2017, 44(18): 9276-9282.
- Runge J, Petoukhov V, Donges J F, et al. Identifying causal gateways and mediators in complex spatio-temporal systems [J]. *Nature Communications*, 2015, 6(1): 1-10.
- Sahoo S, Russo T, Elliott J, et al. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the US [J]. *Water Resources Research*, 2017, 53(5): 3878-3895.
- Schorlemmer D, Wiemer S, Wyss M. Variations in earthquake-size distribution across different stress regimes [J]. *Nature*, 2005, 437(7058): 539-542.
- Shodiq M N, Kusuma D H, Rifqi M G, et al. Neural network for earthquake prediction based on automatic clustering in Indonesia [J]. *JOIV: International Journal on Informatics Visualization*, 2018, 2(1): 37-43.
- Singh A, Bhargawa A. Prediction of declining solar activity trends during solar cycles 25 and 26 and indication of other solar minimum [J]. *Astrophysics and Space Science*, 2019, 364(1): 1-7.
- Sobolev G. On applicability of the RTL prognostic algorithms and energy estimation to Sakhalin seismicity [J]. *Journal of Volcanology and Seismology*, 2007, 1: 198-211.
- Solanki S K, Krivova N A. Analyzing solar cycles [J]. *Science*, 2011, 334(6058): 916-917.
- Solomatine D P, Dusal K N. Model trees as an alternative to neural networks in rainfall-runoff modelling [J]. *Hydrological Sciences Journal*, 2003, 48(3): 399-411.
- Sun A Y. Predicting groundwater level changes using GRACE data [J]. *Water Resources Research*, 2013, 49(9): 5900-5912.
- Sun Y, Wendi D, Kim D E, et al. Technical note: Application of artificial neural networks in groundwater table forecasting — A case study in Singapore swamp forest [J]. *Hydrology and Earth System Sciences Discussions*, 2015, 12: 9317-9336.
- Sunkara S, Tiwari R. Model dissection from earthquake time series: A comparative analysis us-

- ing modern non-linear forecasting and artificial neural network approaches [J]. *Computers & Geosciences*, 2009, 35: 191-204.
- Tapoglou E, Karatzas G P, Trichakis I C, et al. A spatio-temporal hybrid neural network-Kriging model for groundwater level simulation [J]. *Journal of Hydrology*, 2014, 519: 3193-3203.
- Toth J. Groundwater discharge: A common generator of diverse geologic and morphologic phenomena [J]. *Hydrological Sciences Journal*, 1971, 16(1): 7-24.
- Tóth J. Groundwater as a geologic agent: An overview of the causes, processes, and manifestations [J]. *Hydrogeology Journal*, 1999, 7(1): 1-14.
- Turing A M. Computing machinery and intelligence [M]//Mind. 1950: 433-460.
- Wada Y, Van Beek L P, Van Kempen C M, et al. Global depletion of groundwater resources [J]. *Geophysical Research Letters*, 2010, 37(20): L20402.
- Wang Q, Guo Y, Yu L, et al. Earthquake prediction based on spatio-temporal data mining: An LSTM network approach [J]. *IEEE Transactions on Emerging Topics in Computing*, 2020, 8(1): 148-158.
- Wunsch A, Liesch T, Broda S. Forecasting groundwater levels using nonlinear autoregressive networks with exogenous input (NARX) [J]. *Journal of Hydrology*, 2018, 567: 743-758.
- Yadav B, Ch S, Mathur S, et al. Assessing the suitability of extreme learning machines (ELM) for groundwater level prediction [J]. *Journal of Water and Land Development*, 2017(32): 103-112.
- Zahiri A, Najafzadeh M. Optimized expressions to evaluate the flow discharge in main channels and floodplains using evolutionary computing and model classification [J]. *International Journal of River Basin Management*, 2018, 16(1): 123-132.
- Zamani A, Sorbi M R, Safavi A A. Application of neural network and ANFIS model for earthquake occurrence in Iran [J]. *Earth Science Informatics*, 2013, 6(2): 71-85.
- Zhang Y K, Bai E W, Libra R, et al. Simulation of spring discharge from a limestone aquifer in Iowa, USA [J]. *Hydrogeology Journal*, 1996, 4(4): 41-54.
- Zhao H J, Wang J L, Zong W G, et al. Prediction of the smoothed monthly mean sunspot numbers by means of RBF (radial basic function) neural networks [J]. *Chinese Journal of Geophysics*, 2008, 51(1): 20-24.

致 谢

时光如白驹过隙，恍惚间博士生涯已进入尾声阶段，感慨万千。3年前，自己怀着对学术生涯的期待与中国科学院大学结下了缘分。这段路程饱含辛酸与苦楚，但我终不后悔当初的选择。这段博士生涯已深深烙印在我的记忆中，成为我人生中宝贵的财富。回顾这段求学时光，我品尝过无数次的试验失败带来的挫折感。但在自我怀疑后，我仍旧选择振奋精神重新来过。

首先，我想感谢我的导师石耀霖院士。教诲如春风，师恩似海深。三年多的时光，受到了石老师的悉心教导。石老师在科研上给予我最大的支持与关照。在与石老师沟通互动过程中，尤其是在 PPT 中全面详细的修改与耐心指导，让我深刻体会到了石老师对学术的严谨与执着，对学生的热忱与极力支持，其终身学习的态度令我生畏并让我努力向石老师的精神世界看齐。博士阶段，对于科学的研究方法和过程，如何成为一名合格的研究者，石老师都为我倾注了大量心血。

同时，我想感谢的是我的第二指导老师张怀教授。无论在科研还是生活上，张老师是我的良师益友。他始终激励着我，让我从容自信地面对每个挑战，特别是在科研的瓶颈期和焦灼期。

另外，我还想感谢实验室的其他老师、同学和行政人员。感谢周元泽教授对科研工作中的照顾与指导。感谢乔小娟教授跟我开展的关于泉流量研究的合作。同时感激程惠红教授对关于汪荣江老师程序的指导和 GMT 软件的使用。感谢金一民师兄在程序上提供的交流，王然师兄在 \LaTeX 使用上提供的帮助，郭一村师兄积极交流讨论科研的思路与论文写作方法。感谢实验室其他兄弟姐妹的陪伴。感谢实验室为科研提供的工作环境和人文关怀，让我体验到温馨的家园感。

当然，家人的支持与理解给了我精神上的援助。感谢我的父亲支撑起了整个家庭，母亲为这个家庭提供了温暖的港湾。有了这些积淀，我才有机会迈入博士队伍。同时还要感激我的姐姐和弟弟。还要感谢我的男友小灰灰，有了你的鼓励与情感共鸣，为我的人生开启了新的一扇窗。

作者简历及攻读学位期间发表的学术论文与研究成果

作者简历：

2011年9月—2015年7月，在青岛理工大学计算机与工程学院获得学士学位。

2015年9月—2018年7月，在云南大学发展研究院获得硕士学位。

2018年9月—2021年12月，在中国科学院大学地球与行星科学学院攻读博士学位。

已发表（或正式接受）的学术论文：

1. **Cheng Shu**, Qiao Xiaojuan, Shi Yaolin, et al. Machine learning for predicting discharge fluctuation of a karst spring in North China [J]. *Acta Geophysica*, 2021, 69(1): 257-270.
2. 程术, 石耀霖, & 张怀. 基于神经网络预测太阳黑子变化. *中国科学院大学学报* [J]. 2021.

参加的研究项目及获奖情况：

参加的研究项目：

编号	题目	类别
U1839207	大数据与地震数值预测探索	国家自然科学联合基金
41725017	计算地球动力学	国家自然科学基金
U2039207	基于数值模拟的确定性—概率地震危险性分析方法研究	国家自然科学基金

获奖情况：荣获2021学年“三好学生”荣誉称号。