分类号： TP39　　单位代码： 10431
密　级： 公开　　学　　号： 10431200532

齊魯工業大學

硕士学位论文

（专业学位）

基于机器学习算法的海洋观测数据质量控制技术研究

作　者　姓　名　王龙飞
领　　域　电子信息
所　在　学　部　海洋技术科学学部
指导教师姓名
专业技术职务　宋苗苗　副研究员

2023 年　6 月 6 日

A Thesis Submitted for the Application of

the Master's Degree of Engineering

Research on quality control technology of marine observing data based on machine learning algorithms

Candidate: Wang Longfei

Specialty: Electronic information

Supervisor: Professor Song Miaomiao

Qilu University of Technology, Jinan, China

June, 2023

目 录

摘　要

海洋观测数据的正确与否直接影响着海洋基本特征描述、规律分析和管理决策的准确性，失真或错误的深海大洋环境资料若被放进长时间序列海洋数据库中，用于分析、研究物理海洋现象的分布特征及其变化规律，或者直接用于海洋和天气、气候科学领域的业务化预测预报中，终将严重影响预测预报的可靠性。

为了进一步实现数据质量控制的智能化和准确性，本课题结合机器学习算法、统计学等方法对海洋监测数据质量控制技术进行研究，保证海洋观测数据准确有效。

数据的质量控制主要是指从原始数据中找出异常数据并对异常数据进行标记和校正。本文选用多种获取自浮标和观测台站检测的海洋观测数据进行实验和验证，主要包括浮标观测波浪数据、温盐数据和小麦岛观测台站温度数据等，并针对不同海洋观测数据提出了多种以机器学习算法为核心的数据质量控制方法，此外，考虑到统计学方法也具有一定的适用性，本文将部分结合统计学方法进行研究。本文的研究内容主要包括以下几个部分：

（1）海洋观测数据的异常检测方法研究。将海洋观测数据集划分为单变量海洋观测数据和多变量海洋观测数据。对于单变量的浮标观测有效波高数据，使用统计学方法结合局地检验法进行异常检测。对于多变量的浮标观测温盐等海洋观测数据，使用具有数据重构功能的自编码器进行异常检测。

（2）海洋观测数据的异常值校正方法研究。海洋观测数据是一系列时间序列数据，本文分别以长短周期记忆神经网络（Long Short Term Memory: LSTM）算法为核心和以 ARIMA 算法为核心建立海洋观测数据预测模型，对检测到的异常数据进行预测，并使用预测数据替换异常数据，实现数据的异常值校正。

（3）海洋观测数据质量智能控制软件界面系统设计。结合上述两部内容，使用 Python 和 PyQt5 设计数据质量控制软件系统，将数据质量控制的各个过程，方法选择以及结果显示集成到可视化应用程序，支持人机交互，提供海洋观测数据质量控制的全流程化操作功能。

关键词：海洋观测；数据质量控制；统计学方法；机器学习；异常检测；异常值校正

ABSTRACT

The correctness of marine observing data directly affects the accuracy of the description of the basic characteristics of the ocean, the analysis of the laws and the management decisions. If the distorted or incorrect deep-sea ocean environmental data are put into the long-term time series ocean database, they can be used to analyze and study the distribution characteristics and change rules of physical ocean phenomena, or directly used in the operational prediction and forecast in the fields of ocean, weather and climate science, Finally, the reliability of prediction will be seriously affected.

In order to further realize the intellectualization and accuracy of data quality control, this thesis combines machine learning algorithm, statistics and other methods to study the quality control technology of marine observing data, to ensure the accuracy and effectiveness of marine observing data.

Data quality control mainly refers to finding abnormal data from the original data and marking and modifying the abnormal data. This thesis selects various marine observing data obtained from buoys and observation stations for experiments and validation, mainly including wave data, temperature and salinity data observed by buoys, and temperature data measured by XiaoMaiDao observation stations. Multiple data quality control methods centered on machine learning algorithms are proposed for different ocean observation data. In addition, considering that statistical methods also have certain applicability, this thesis will research in part with statistical methods. The research content of this thesis mainly includes the following parts:

(1) Research on anomaly detection model of marine observing data. The marine observing data set is divided into univariate marine observing data and multivariate marine observing data. For univariate buoy observation significant wave height data, the statistical method combined with local test method is used for anomaly detection. For multivariable marine observing data such as temperature and salinity observed by buoys, the self-encoder with data reconstruction function is used for anomaly detection.

(2) Research on outlier correction method of marine observing data. Because the marine observing data is a group of time series data, this thesis establishes the marine observing data prediction model with the Long Short Term Memory network Model: LSTM algorithm as the core and the ARIMA algorithm as the core, predicts the detected abnormal data, and replaces the abnormal data with the predicted data to achieve the

correction of the abnormal value of the data.

(3) The design of intelligent control software interface system for marine observing data quality. Combining the above two parts, a software system is designed using Python and PyQt5 to integrate the various processes, method selection and result display of data quality control into the visual interface, support human-computer interaction, and provide the full-process operation function of marine observing data quality control.

Keywords: Ocean observation, Data quality control, Statistical methods, Machine learning, Abnormal detection, Correction of outlier

第 1 章 绪论

1.1 研究背景和意义

海洋观测是海洋科学与技术的重要组成部分，人们认识海洋，发展海洋和保护海洋都离不开海洋观测技术的应用[1]。各项海洋工作和各种海洋生产活动的开展，如预防海洋灾害，维护海洋权益，防治海洋污染，保护海洋生态环境等，必须要做到对海洋空间的认知、对海洋现象的了解、对反映机理的寻求、对演变过程的探讨、对海洋秘密的发现和海洋规律的总结等，由此建立的先进的海洋观测网和预报服务系统，为海洋工作提供了重要的保障服务[2]。海洋观测数据的正确与否直接影响着海洋基本特征描述、规律分析和管理决策的准确性。缺失或异常的海洋观测数据若被放进长时间序列海洋观测数据库中，用于分析、研究物理海洋现象的分布特征及其变化规律，或者直接用于海洋观测和气象、气候等科学领域的预测预报中，最终会严重影响预测预报的准确性和可靠性。另外，如果缺少对数海洋观测数据的质量控制，直接将数据应用于海洋调查活动及各类海洋观测网，同样会使结果大打折扣，甚至会产生“数字垃圾”风险。

随着海洋观测技术的不断发展和进步，海洋资料数据已经得到了极大的积累，但是，由于海洋观测存在客观的外部影响，观测设备所得到的海洋观测数据不可避免地会存在缺失和异常，如何从海量的数据中检测出异常并有效地填补数据，提高海洋观测数据的可用性，成为现代海洋研究的一项重要课题。

质量控制（quality controlling, QC）是指采用一定方法、建立数学模型和调整参数来判断数据质量可靠性，并进行质量标识和数据修补的处理过程[3]。质量控制的目的是提高数据的准确性和可用性。影响海洋观测数据可靠性的关键因素是数据错误。错误数据的产生有各种原因，如设备故障、网络问题或通信问题等，这些都可能对数据分析结果产生重大影响[4]。

机器学习是当前计算机行业最为火热的话题之一。机器学习可以概括为利用已有的海量数据，分析推论出某种模型，预测未来的一种方法。随着机器学习的发展和机器学习在其他科学领域的应用，形成了模式识别、统计学、机器视觉、语音识别、自然语言处理等交叉学科；机器学习在数据挖掘领域的应用，得出了良好的结果并运用在各个行业。其中利用机器学习结合统计学、数据挖掘等进行海洋观测数据质量控制是获得准确可靠的海洋观测数据的一种重要途径。

海洋观测数据质量控制对保障海洋观测数据的可靠性具有重要意义，但海洋观测数据质量控制系统的设计受到很多方面制约，包括：

（1） 异常天气导致的正常的异常数据；

（2） 不同观测要素之间的影响；

（3） 不同方法对于不同要素的适用性；

综上所述，随着海洋观测技术的进步和海洋观测数据的不断积累，海洋观测数据质量问题急需解决，在计算机技术和机器学习算法的不断发展下，海洋观测数据质量控制技术有着十分广阔的应用场景，研究并设计出适用于海洋观测数据的数据质量控制方法具有重要且实际的意义，对于海洋观测行业的发展具有巨大的推动作用。

1.2 国内外研究进展

1993 年，联合国政府间海洋学委员会牵头组织和实施了全球海洋观测系统计划，使得海洋观测能力开始稳步增强[2]。随着多年海洋观测技术的发展，相关单位已经积累了大量的海洋观测数据，并在各个领域加以应用，海洋观测数据质量控制研究也得到了更多的重视。美国国家海洋和大气管理局国家环境信息中心（NOAA/NCEI）、英国气象局哈德莱中心（Met Office Hadley Centre）、全球温盐剖面计划（GTSPP）、Argo 全球海洋观测阵列计划、世界大洋环流实验室（WOCE）等研发机构系统的开展了海洋观测数据质量控制的研究。在中国，杭州全球海洋 Argo 系统野外科学观测研究站（中国 Argo 实时资料中心）、国家海洋信息中心、中国科学院海洋研究所、山东省科学院海洋仪器仪表研究所等科研机构也对海洋观测数据质量控制的研究愈发重视[4]。

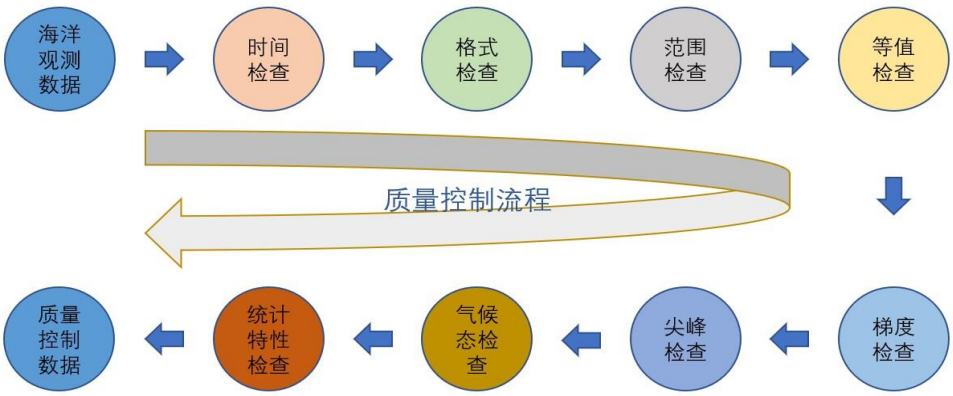


图1.1 数据质量控制的流程示意图

各个研究机构对海洋观测数据质量控制系统的设计子模块大致相同，但子模块检查顺序不尽相同。图 1.1 是上述机构数据质量控制的流程示意图，表 1.1 是对上述机构系统子模块的质量控制方法的整理和归纳。

目前，基于统计学与计算机技术的海洋观测数据质量控制技术不断取得进展，基于机器学习算法的海洋监测数据智能质量控制技术仍处在不断研究和实验的阶

段。基于统计学与机器学习的海洋观测数据质量控制技术包含两部分内容，即基于统计学与机器学习的海洋观测数据异常检测和基于机器学习的海洋观测数据异常值校正。

表 1.1 各机构或数据集数据质量控制方法的整理归纳

子模块类型		WOD	Argo	GTSPP	EN4	WAGHC	国家海洋信息中心		
							大面	定点	移动
格式转换		0	x	0	x	x	0	0	0
预检查（基础检查）	测站识别码检查	x	1	1	x	x	2	2	2
	日期时间检查	1	1	1	x	1	3	3	3
	经纬度检查	1	1	1	x	1	4	4	4
	陆地位置检查	1	1	1	5	4	5	5	5
	流速/速度检查	2	1	1	3	x	x	x	6
	声速检查	x	x	1	x	x	x	x	x
	深度递增检查	3	3	4	6	2	8	8	9
范围检查	全球范围检查	x	2	2	1	3	6	6	7
	区域参数范围检查	5	2	3	x	x	7	7	8
	全球深度-温/盐范围检查	5	x	5	x	3	x	x	x
	冰点检查	x	11	7	x	x	14	14	15
剖面检查	尖峰检查	x	4	8	2	8	9	9	10
	垂直梯度检查	6	x	9	x	9	x	x	x
	距离中值检查	x	5	x	x	x	10	10	11
	密度翻转检查	7	8	10	4	x	13	13	14
	数据翻转检查	x	6	x	x	x	11	11	12
	等值检查	x	7	6	x	6	12	12	13
	极值检查	x	x	x	x	7	x	x	x
	相邻廓线检查	x	x	14	7	x	x	x	x
	逆温检查	x	x	12	x	x	x	x	x
仪器类型检查	Argo 英文名检查	x	9	x	x	x	x	x	16
	仪器最大深度检查	x	x	x	x	5	x	x	1
	温盐漂移检查	x	10	x	x	x	x	x	17
统计特性检查	莱茵达准则检查	8	x	x	x	x	x	15	x
	拟合优度检查	x	x	x	x	x	x	x	x
	局地气候态检查	x	x	13	x	9	x	x	x
	温度-盐度图检验	x	x	x	x	x	15	x	18
	相邻数据检查	x	x	x	9	x	x	16	19
局地最大深度检查		x	0	11	5	4	1	1	x
航线检查		x	x	15	3	x	16	x	20
背景场检查		x	x	x	8	x	x	x	x
高精度廓线配对检查		4	x	x	x	x	x	x	x

注：表中数字表示此种检查方法在对应机构中的数据质量控制程序中的先后顺序，表中

值 x 表示对应的机构中没有使用该检查方法。表 1 引用自文献[5] 。

1.2.1 国外研究现状

在科学研究和生产生活的不同领域，异常具有各不相同的定义，但本质上均是基于 1980 年 Hawkins[6]的定义，即异常表示的是由于产生原理不同而与其他序列段存在较大差异的数据。而随着异常检测的不断研究，学者们对异常的定义进行了补充，即认为异常是指与数据集中其他数据表现不一致的观测值[7]。在海洋观测领域中，定义异常为：由海洋环境突变引起或者由海洋观测设备故障引起的海洋观测要素序列不符合正常模式的特殊观测数据[8, 9]。近年来，基于数据驱动的方法逐渐应用于海洋观测数据的异常检测领域[10, 11]，该方法可以分为两个大类：统计类和机器学习类。统计类方法是通过构建统计模型，然后采用假设检验的方式进行判定，数据出现的概率越小表示异常的概率越大[12]。统计类方法主要有参数和非参数方法。机器学习类异常检测方法主要有带有标签的异常检测和无标签的异常检测。有标签的机器学习异常检测是基于已标记的异常和非异常数据构建分类模型，然后分类模型将新的数据分类为异常类或非异常类，主要包括 K 近邻法[13]、决策树[14]、支持向量机[15]和人工神经网络[16]等。海洋观测数据大多是没有经过标记的无标签数据，无法使用监督学习模型进行异常检测，对于无标签数据，Kmeans 算法是一种应用广泛、研究使用较多的无监督学习算法，为了改进算法、提高算法的性能，研究人员在此基础上提出了 Min Max Kmeans 算法[17]、KMOR 算法[18]、Seeded-KMeans[19]和 IWO-KMeans 算法[20]等多种能够适用不同情境的算法，这些改进算法相较于传统 Kmeans 算法具有更好的聚类效果，使用这些方法的数据异常检测也具有更高的效率。海洋观测数据是一种时间序列数据，通常需要使用能够解决数据时间依赖性的方法来处理海洋观测数据，如 ARIMA[21]。此外，基于其他机器学习算法的海洋观测数据预测也不断得到验证和应用。

加拿大海洋网络（ONC）开发并实施了一个全面的面向过程的质量保证模型和面向产品的数据质量控制模型，主要包括自动的传感器范围检查、传感器关系测试、尖峰检测和梯度检查等以及定期手动检查[22, 23]。美国海洋观测计划(OOI)采用了系统级和人在回路数据质量控制方法[24]。在系统层面，采用了全局范围、局部范围、固定值、梯度、趋势和峰值测试六种自动化算法。在自动算法测试之后，执行人工回路质量控制方法[25, 26]。此外，Argo 项目在世界海洋上部署了 3000 多个浮标，对电导率、温度、深度等海洋要素采用实时自动质控和延迟质控两种数据质量控制程序。机器学习等应用算法的不断进步，使得质量控制系统的建立有了新的突破方向。Rahman 等人采用监督分类法，并使用多分类器框架对海洋传感器网络进行数据质量评估，其中包括平衡，以解决分类中的不公平性[27, 28]。Timms 等人提出了一种新的自动化数据质量评估框架，该框架使用模糊逻辑提供数据质

量的连续尺度，然后使用连续质量标尺计算数据上的误差条，他们研究的重点是量化数据的不确定性，并为数据的适用性提供更有意义的测量方法[29]。Smith 等人提出了一个动态贝叶斯网络（DBN）框架，其旨在改进复制专家生成的误差条方面，用于产生概率质量评估，并表示相关传感器读数的不确定性[30]。Karakuş 等人[31]建立了多项式自回归模型来预测风速数据，他们使用过去每小时平均风速数据进行后一天的风速预测，证明了在一定条件下，使用多项式自回归模型进行风速数据的预测能够取得较高精度的结果。Arumugam 和 Saranya 等[32]提出了一种基于 ARIMA 的月降雨量数据异常检测技术并进行了实验验证。Wani 等人[33]提出了一种基于聚类的方法来对气象数据进行预测，并利用气温、大气压、风向、相对湿度等多种要素进行实验验证，证明了该方法能够较好的预测风速数据。Demolli 等人[34]通过建立随机森林（RF）、支持向量机回归（SVR）、极端梯度提升算法（XGBoost）三种机器学习预测模型并对比三个模型在预测长期每日总功率数据上的模型性能，结果表明，使用随机森林算法建模的预测模型具有更好的预测能力。Tomin 等人[35]提出了一种基于引入 Hilbert 谱分析的经验模态分解方法结合机器学习算法的自适应风速预测方法，对于非平稳风速时间序列，使用该方法建立的预测模型，能够预测得到较高精度的未来数据。Shi[36]等人将小波技术和神经网络技术(wavelet-artificial neural networks, WANN)应用于水质异常检测，首先通过小波重构技术去除高频噪声，然后采用人工神经网络(artificial neural networks, ANN)预测水质的变化来进行异常检测。Pushe 等人[37]利用传感器序列间的皮尔逊相关性变化来进行异常检测，当序列间皮尔逊相关系数超出阈值时判定为异常，实验证明具有较好的检测效果。Li 等人[38] 提出了基于主成分分析(Principal Component Analysis, PCA)方法的数据异常检测，将原始多维数据通过 PCA 映射到低维空间中表示，通过训练数据计算残差阈值来判定异常。Kim 等人也针对风速数据提出了使用多层感知器、支持向量回归的质量控制方法，实现了对要素数据的质量控制[4]。Khodayar 等人[39]提出了一种基于自编码器的短期数据预测方法，并通过超短期和短期风速数据进行建模验证，证明了该方法对短期风速数据预测的精度性和有效性。

1.2.2 国内研究现状

我国沿海海域通过岸上的数据中心和海上布放的海洋观测浮标，建立了能够稳定可靠、安全运行的实现采集、传输、接收和处理海洋观测数据功能的浮标观测网，成为支撑起我国海洋科技发展和海洋生产活动的重要数据来源[40]。浮标观测是海洋观测数据获取的一种重要方式，对浮标和各种观测设备获取的海洋观测数据进行数据质量控制一直是海洋观测数据处理的重要方面，对此，国内外多位学者专家进行了海洋浮标观测数据质量控制研究。目前我国海洋观测数据质量控

制大多还是使用数据的时间检查、范围检查、梯度检查、气候特性检查等方法进行质量控制[40]，此外，统计学方法也是进行数据质量控制的重要研究方法。针对海洋波高观测数据，刘首华等人使用格拉布斯准则、局地异常值检验方法和波高观测误差控制方法建立了一种异常值检测模型，并通过实验验证了所建模型在波高异常值检测上的有效性[40]。机器学习算法也逐渐应用于海洋观测数据领域，进行数据的预测。郭颜萍等人使用小波变换和 LS-SVM 的方法进行了船面风速风向的估算研究[41]。周禹生人使用滑动窗口改进 ARIMA 模型，并将其应用于海底机器人观测数据质量控制[42]。王国松等人使用长短期记忆神经网络（LSTM）进行了沿海风速预报的研究[43]，贺琪等人也使用 LSTM 算法结合 STL 分解算法进行了海表面温度预测算法的研究[44]。而结合范围检查、时间检查及统计学等方法的使用机器学习算法进行数据评估和插值的数据质量控制技术还鲜有得到使用，同时，随着对海洋监测数据应用领域的增加，海洋监测数据质量的要求也越来越高，使用新的方法如结合机器学习算法进行质量控制对保证海洋监测数据质量具有重要意义。

1.3 论文结构和内容安排

本论文的结构和内容安排如下：

第 1 章：本论文的绪论部分。主要介绍了海洋观测异常数据对海洋研究和海洋行业的影响，以及海洋观测数据质量控制的重要性，介绍了海洋观测数据质量控制的数据特性。针对上述问题，介绍了近年来国内外海洋观测数据异常检测和预测模型，并对文章总体结构进行概述。

第 2 章：海洋观测数据质量控制关键技术。主要介绍了海洋观测数据质量控制总体技术路线，阐述了海洋观测数据的时间特性和分类，分析了应用于海洋观测时间序列的分解算法。针对海洋观测数据质量控制的异常检测部分剖析了几种传统的统计学方法异常检测和几种结合机器学习算法的异常检测方法，针对海洋观测数据质量控制的异常值校正部分重点阐述了几种以机器学习算法为基础的海洋观测数据预测方法。

第 3 章：海洋观测数据异常检测模型。在第二章单变量时间序列和多变量时间序列数据分析理论的基础上，对于单变量的有效波高观测数据，设计并开发了一种用于单变量海洋观测数据异常检测的基于统计学方法、局地异常检测和误差控制的异常检测方法；对于多变量或者单变量海洋观测数据，建立并实现了基于自编码器的海洋观测数据异常检测模型。通过对比实验，分析模型的性能并检测数据异常和做质量标记。

第 4 章：海洋观测数据异常值校正模型。对于第三章检测到的异常数据，需

要通过一定的方法将数据修改或填补，通过预测的方法产生更加接近真实值的数据来进行异常数据的插值是更加有效的。本章主要建立了基于 LSTM 的海洋观测数据预测模型和基于 ARIMA 的海洋观测数据预测模型，并通过对比分析实验验证模型的有效性。

第 5 章：海洋观测数据质量智能控制软件系统设计。基于上述模型和实验，将实验结果以及实验方法等设计到一个统一的软件界面上进行显示和操作。

第 6 章：总结与展望。首先对上文各章海洋观测数据质量控制过程进行进一步的总结，然后提出本文研究可以进一步改进和提升的思路。

第 2 章 海洋观测数据质量控制关键技术

海洋观测数据由一组或多组时间序列数据组成，具有时间序列数据所有的特性。海洋观测设备可能会因为传感器故障或老化、外部环境突变等得到异常的数据或者数据缺失，数据的异常和缺失会影响观测数据在海洋研究和生产中的使用，所以对海洋观测数据质量控制方法的研究深受重视。

本章主要介绍海洋观测数据质量控制的总体技术路线以及时序数据分解算法、数据异常值检测方法、数据预测方法等海洋观测数据质量控制关键技术。首先，设计了海洋观测数据质量控制的总体技术路线；其次，阐述了海洋观测时间序列数据的定义、STL 时序数据分解算法以及小波分解重构算法；然后，详细阐述了基于统计学方法和机器学习算法的数据异常值检测方法；之后，介绍了基于 ARIMA 和 LSTM 的海洋观测数据预测方法，最后是本章小结。

2.1 海洋观测数据质量控制的总体技术路线

海洋观测设备在收集海洋要素数据时，由于其本身的电气损耗和周围环境的影响，可能会导致观测数据的失真与缺失。为保证获取的海洋观测数据能够有效可靠的应用于海洋行业的各个方面，本论文针对不同海洋观测要素（海温和波浪有效波高），分析了多种单变量和多变量的异常检测方法，将统计学方法和机器学习算法结合应用于单变量海洋观测数据的异常检测，将自编码器结合神经网络、注意力机制、数据分解算法等应用于多变量海洋观测数据的异常检测，并运用 ARIMA、STL、小波分解、LSTM 等算法建立数据预测模型对异常数据进行异常值校正，设计开发海洋观测数据质量控制软件系统提高人机交互界面对数据质量控制结果进行显示和输出。

本文所设计的海洋观测数据质量控制总体技术路线如图 2.1 所示。针对海洋观测设备获取的原始数据存在的异常数据问题，首先，采用统计学方法和机器学习算法建立多种数据质量控制方法模型对异常数据进行检测和校正；然后，通过对比实验分析各模型的性能，并分析其在不同海洋观测要素下的性能；最后将数据质量控制模型集成到数据质量控制软件系统中。

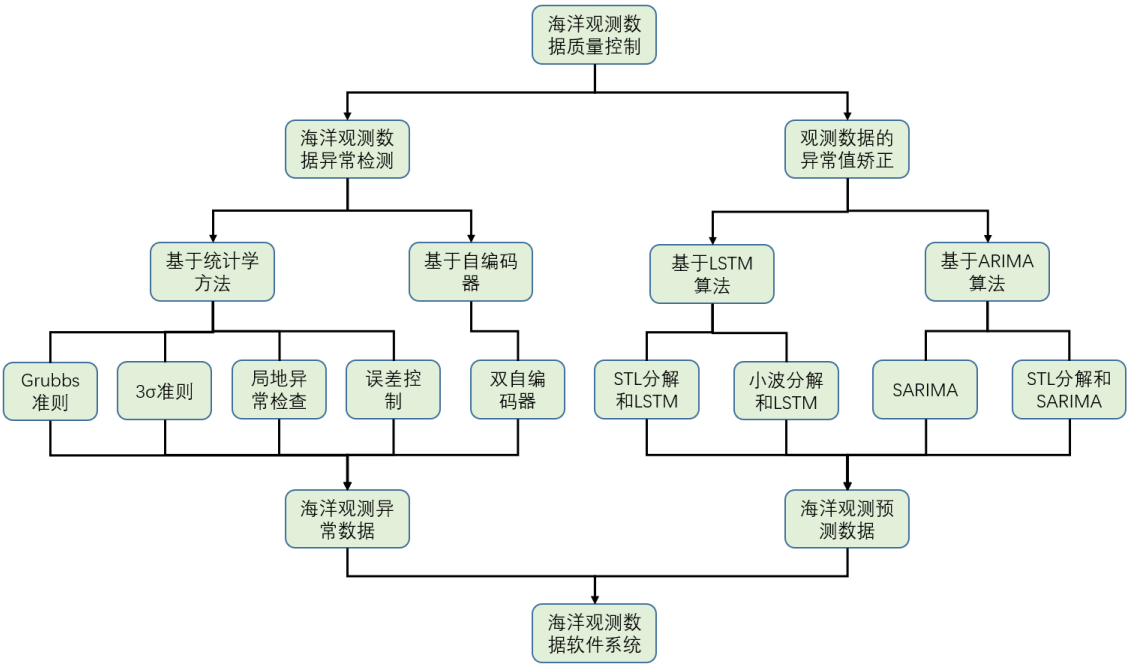


图2.1　海洋观测数据质量控制总体技术路线

针对海洋观测设备观测到的海温和波高数据，分析其准确性、完整性等特性，基于机器学习算法等建立数据质量智能控制模型对数据进行质量控制。数据质量控制主要包含异常检测和异常值校正两部分。统计学方法主要适用于单变量的数据分析，对于得到单一海洋要素的原始观测数据，通过统计学方法结合监督学习分类的方法进行单一要素的异常检查。对于具有较强相关性的多海洋要素的数据，即多维数据，使用孤立森林和自编码器等方法进行异常检测是主要的研究方法。对于检测标记的异常值，通过 LSTM、时间序列预测模型等得到的预测数据进行替换插值。海洋观测数据质量控制技术包括以下三个方面的内容：第一部分为单变量海洋观测数据的异常检测，通过时序数据分析得到的原始海温和波高数据特性，采用统计学方法和机器学习算法检测不同海洋观测要素的数据异常，研究无监督学习算法对单变量数据异常检测的适用性，分析比较其异常检测的性能，建立单变量海洋观测数据异常检测模型；第二部分多变量海洋观测数据的异常检测，基于自编码器机器学习算法建立数据异常检测模型，然后结合 LSTM 等机器学习算法优化其性能，建立多变量海洋观测数据的异常检测模型并分析其性能；第三部分运用 LSTM、ARIMA 等机器学习预测算法对异常检测得到的异常数据进行校正，通过分析海洋观测数据的特性采用不同的分解算法优化 LSTM 和 ARIAM 预测模型的性能，然后建立海洋观测数据预测模型，实现海洋观测数据的异常值校正；第四部分是设计完整的海洋观测数据质量智能控制软件系统，直观的将各项数据显示在软件系统上。

2.2 时间序列数据分解算法

2.2.1 海洋观测时间序列数据

时间序列 T 是一组随时间先后顺序不断变化的数字序列，其 n 个数值具有某些变化规律并统一在一个统计指标下：

$$\mathrm{T} = (t_1, t_2, ..., t_i, ...t_n), t_i \in R \tag{2.1}$$

海洋观测时间序列数据是对随时间变化而变化的海洋观测数据的统称，它用于描述海洋观测要素随时间变化的情况。通常来说，海洋观测时间序列数据是通过对海洋观测要素使用给定采样频率及时间间隔进行数据采集的结果。当观测要素为单个时，称时间序列为单变量时间序列，当观测要素不止一个时，称时间序列为多变量时间序列。

2.2.2 STL 时间序列分解算法

海洋观测数据是一组或多组的时间序列，对于时间序列数据，其数据变化主要指数据随时间的趋势变化和数据在一段时间内的周期性变化，所以，找到时间序列的趋势变化和变化周期，对于时间序列的分析、处理和预测等都有重要作用。本节分析本文选择使用的时间序列分解算法。

基于局部加权回归的周期趋势分解方法(seasonal-trend decomposition procedure based on loess, STL)[45]是一种可以将时间序列数据 Y_t 分解为周期项（季节分量，S_t）、趋势项（趋势分量，T_t）和冗余项（剩余分量，Rt）的时间序列分解方法。STL 分解算法可以用来探索历史数据的规律，还可以用于处理适用于任何周期性的数据，具有很好的鲁棒性，表达式为：

$$Y_t = S_t + T_t + R_t \tag{2.2}$$

STL 分解算法包含两个部分：内循环和外循环[46]。外循环的设计是为了通过分配鲁棒性权重来减少数据中离群值的影响。内循环的作用是更新分解得到的趋势项和季节项，过程如下[47]：

（1）对时间序列数据去趋势。去除趋分量后的结果为 Y_t-$T_t^{(k)}$。k 是内循环的循环数。

（2）循环子序列平滑得到$\tilde{S}_t^{(k+1)}$。

（3）子序列的低通滤波。得到平滑子序列趋势分量$\tilde{T}_t^{(k+1)}$。

（4）平滑后的循环子序列去趋势。(k+1)次循环中的季节分量为：$S_t^{(k+1)}=\tilde{S}_t^{(k+1)} - \tilde{T}_t^{(k+1)}$。

（5）去季节性 $Y_t - S_t^{(k+1)}$ 。

（6）趋势平滑。对去除季节性的序列做 Loess 回归，得到趋势分量 $T_t^{(k+1)}$。

内部循环不断运行，直到符合要求退出内部循环，开始外部循环。在外部循环中，计算序列的余项[48]R_t，计算公式为：

$$R_t^{(k+1)} = Y_t - T_t^{(k+1)} - S_t^{(k+1)} \tag{2.3}$$

外循环主要为了调节内循环步骤（2）和步骤（6）Loess 回归中的鲁棒权重。鲁棒权重 h 被定义用于评价 R_t 的鲁棒性，ρ_t 是时刻t的鲁棒权值[49]，其公式表示为：

$$h = 6median(|R_t|) \tag{2.4}$$

$$\rho_t = \frac{B(|R_t|)}{h} \tag{2.5}$$

权重函数 B 的计算公式为：

$$B(u) = \begin{cases} (1-u^2)^2 & 0 \le u < 1 \\ 0 & u > 1 \end{cases} \tag{2.6}$$

STL 分解算法不仅仅是处理月度和季度数据，还可以处理任何类型的季节性数据， STL 分解的季节分量可以根据实验需求进行时间变化速度的控制，STL 分解算法对异常值具有鲁棒性，少量的异常值不会影响趋势项和季节项的估计。基于上述优点，STL 分解算法在时间序列数据的处理中得到了广泛的应用。

2.2.3 小波分解重构算法

信号的主要特征主要在于其概貌信号的特征，一组数据的有一定的变化趋势和变化过程中的细微波动，如果能够分解出一组数据的变化趋势和造成波动的细节信息，对于处理数据、分析数据和预测数据都会有积极的影响。本节主要是对小波分解与重构算法的分析和研究。

1998 年，法国学者 S.Mallat 提出了小波分解与重构的快速算法 Mallat 算法，它将信号分解为代表低频成分的概貌信号和代表高频成分的细节信号，概貌信号反映信号变化的基本趋势，细节信号反映原始信号变化的随机波动。设 $c_0(n)$为待分解离散信号，根据 Mallat 算法分解式子可表示为[50]：

$$\begin{cases} a_{j+1}(k) = \sum_{m \in z} h(m-2k) a_j(m) \\ d_{j+1}(k) = \sum_{m \in z} g(m-2k) d_j(m) \end{cases} \tag{2.7}$$

其中，a_{j+1} 为 j+1 层的低频系数，d_{j+1} 为 j+1 层的高频系数，h，g 分别为低频和高频分解滤波器。小波重构是小波分解的逆过程，重构表达式为：

$$a_j(m)=\sum_{k\in z}h(m-2k)a_{j+1}(k)+\sum_{k\in z}g(m-2k)d_{j+1}(k) \tag{2.8}$$

其中，$\tilde{h}$，$\tilde{g}$分别为低频和高频重构滤波器。

2.3 海洋观测数据异常检测方法

海洋观测数据的质量控制，首先要做的就是检测出已有数据的异常和缺失，这是数据质量控制研究的基础。在各行各业，基于统计学方法的异常检测被广泛使用，而随着机器学习的发展和深入，各种应用机器学习算法的数据异常检测技术更是层出不穷，将这些方法应用于海洋观测数据领域，是本文研究的重要内容。本节介绍几种常用的异常检测方法。

数据异常检测使用最多的方法就是统计学方法，常用的异常值统计判别准则有 3σ 准则、奈尔准则、格拉布斯（Grubbs）准则、狄克逊（Dixon）准则等，在不同的情形下需要采用合适的准则进行异常检测。

格拉布斯准则和 3σ 准则都适用于单变量数据，而海洋观测数据的每一个要素都可以视为一组单变量数据，所以，对于海洋观测数据，使用 Grubbs 准则和 3σ 准则做异常检测是本文研究的一个重要方法。

孤立森林算法和自编码器模型是两种应用于没有数据标签的多变量数据异常检测方法。

本节主要是对两种统计学方法在数据异常检测领域计算过程的分析和总结，通过介绍和分析孤立森林算法和自编码器模型，最终确定本文使用的异常检测方法。

2.3.1 基于 Grubbs 准则的数据异常检测方法

Grubbs 准则作为统计学方法中使用最多的异常值检测方法之一，其基本原理是从一组数据中找出最远离数据平均值的观测数据并通过观测数据序列的标准差来判断数据脱离此观测序列的程度。由于格拉布斯准则主要适用于少量数据，而且根据格拉布斯表可以选择不同的数据量以及对应的临界值参数，对比其他统计学判别准则，格拉布斯准则在少量数据上进行的数据异常检测更具合理性。

Grubbs 准则的计算步骤：

步骤 1：准备数据。

步骤 2：数据排序，将待检测的观测数据按从小到大或从大到小的方式进行排

序，其可疑值只存在于最大值或最小值中。

步骤 3：计算平均值 x 和标准差 σ。

步骤 4：计算最小值与平均值的差 p1、最大值与平均值的差 p2，得到两个偏移值。

步骤 5：可疑值确定：p1 和 p2 更大的那个对应的最小值或最大值。

步骤 6：计算 G_i 值：$G_i=(x_i-x)/\sigma$。查询格拉布斯表得到临界值 $G_P(n)$，通过比较 G_i 和 $G_P(n)$大小判断数据是否可疑，如果 G_i 值不大于 $G_P(n)$，则判断该观测数据是正常值，否则判定该观测数据为异常值。

步骤 7：剔除被判定为异常值的数据，剩余数据重复上述计算，知道数据中没有异常值。

2.3.2 基于 3σ 准则的数据异常检测方法

拉依达准则是数据分析领域进行偏离点剔除的常用方法之一。对于一组观测数据，首先假定其只含有随机误差，并计算数据的标准偏差，然后根据数据的标准偏差和设置的概率确定一个区间，观察误差的区间分布，如果误差超过这个区间，则认为误差为粗大误差而不是符合假设的随机误差，此时，该误差对应的数据认为是异常值。

在 3σ 原则下，设置误差区间为 3 倍的标准差，只要超过这个范围区间，数据就被视为异常值。在使用 3σ 原则进行异常检测时，通常认为数据的取值有 99.7%的概率集中在(μ-3σ,μ+3σ)区间内（μ 为平均值，σ 为标准差），只有 0.3%的数据取值会超出这个范围，属于极个别的小概率事件，因此将超出 3 倍标准差范围的值认为是异常值。3σ 原则要求数据服从正态或近似正态分布，且样本数据不能太小。对于数据不服从正态分布的情况，则可以根据不同业务场景使用远离平均值的 k 倍标准差来控制误差范围，k 值就可以认为是阈值。

2.3.3 基于孤立森林的数据异常检测方法

孤立森林（Isolation Forest, iForest）[51]于 2008 年由 Fei Tony Liu, Kai Ming Ting 和周志华教授在提出。iForest 算法的主要应用领域是异常数据挖掘和离群点检测，即从所得数据中找出与总体数据规律不太符合的数据。其基本思想是对数据空间和切分而成的数据子空间使用一个随机超平面不断进行切分，每次切分的结果是产生两个下级子空间，切分结束的信号是每个子空间中只包含一个数据点。

孤立森林的基本原理是数据中的异常值可以通过较少次数的随机特征分割而从正常值中孤立出来。孤立森林的构建包含 t 棵孤立树(iTree)的构建和 t 棵孤立树构成孤立森林两个过程。孤立树结构如图 2.2 所示，包含作为孤立树起始点的根结点、含有子结点的内部结点和不可再分出子结点的叶结点。

孤立树的构建过程如下：

步骤 1：从训练数据中随机选取 n 个数据样本作为子集放入一棵树的根结点；

步骤 2：随机选取某个特征，并从其中随机选取一个阈值 p；

步骤 3：根据阈值 p 将数据划分到 2 个子空间；小于 p 的点放在左子空间，大于 p 的放在右子空间；

步骤 4：各内部节点重复步骤 2 和步骤 3，直到叶结点上只有一个数据或树已经长到了所设定的高度。

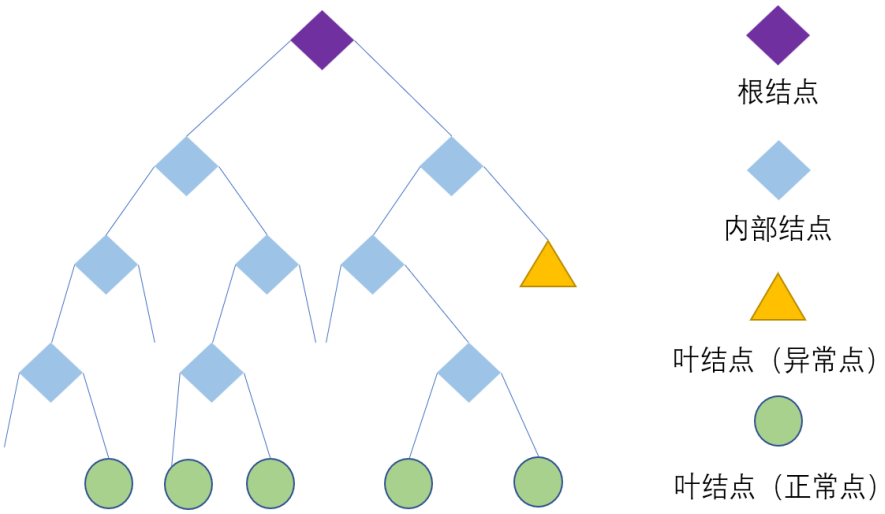


图2.2　孤立树结构图

在生成 t 棵孤立树后，需要对样本进行异常程度评估，即针对每个样本 x，使用如下公式估计异常得分 S：

$$S(x,m)=2^{\frac{-E[\mathrm{f}(x)]}{\mathrm{g}(m)}} \tag{2.9}$$

$$\mathrm{g}(m)=2H(m-1)-2(\frac{m-1}{m\sqrt{b^2-4ac}}) \tag{2.10}$$

在得到样本的异常得分 S 后，对样本进行异常判定，如果异常得分 S 越接近于 1，说明样本异常的可能性越大；如果异常得分越接近于 0，表示异常的可能性越小。

孤立森林是一种通过划分次数进行异常判断的异常检测方法，使用孤立森林进行异常检测的一个关键是设计合适的数据的划分规则。孤立森林异常检测有两个重要的前提：

1. 要求样本中异常值是少量的；

2. 异常值与正确值之间具有明显的差异，样本中越容易被切分出来的点越有

可能被判定为异常值。

基于孤立森林的海洋观测异常检测模型如图 2.3 所示：

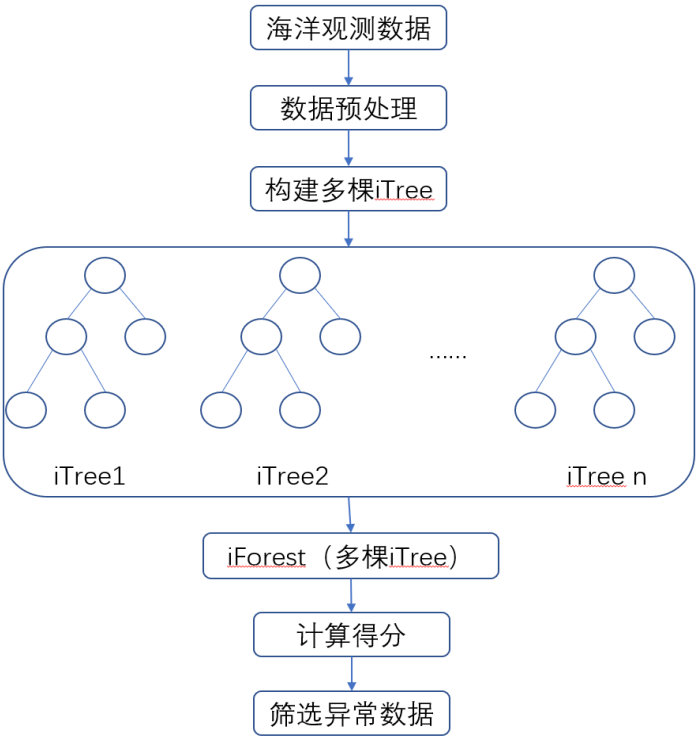


图2.3　基于孤立森林的海洋观测异常检测模型图

2.3.4 基于自编码器的数据异常检测方法

对于没有异常标签的数据的异常检测，自编码器是一种有效的检测方法，本文多变量数据的异常检测将以自编码器为核心建立模型。下面分析几种常见的自编码器模型。

自编码器(Auto Encoder, AE)是由 Rumelhart 等人[52]提出的一种模型，AE 具有结构简单，可根据具体情况进行多层堆叠的特点，AE 可以实现数据的降维、重构表示等功能，是无监督式学习模型中重要的一种，被广泛应用于数据异常检测、数据生成和图像处理等领域。

自编码器由两部分构成，分别是进行数据降维的编码器部分和实现数据重构的解码器部分。编码器的具体实现是将高维输入 W 编码成低维的隐变量 Z，得到可以表示高维输入 W 的最有信息量的数据特征；解码器则是用来将隐藏层的隐变量 Z 重构为与输入维度相同的 $Y=\{y_1, y_2, \ldots, y_k\}$，即 AE(W)。上述编码、解码过程可以描述为：

$$Z = E(w) = \varphi(a_E W + b_E) \tag{2.11}$$

$$AE(W) = D(Z) = \varphi(a_D Z + b_D) \tag{2.12}$$

其中，a_E、a_D 分别为自编码器各部分的权重，b_E、b_D 为偏置，ϕ 指非线性激活函数。

损失函数的作用是衡量因数据降维而造成的信息损失程度，保证 AE 能够有效地提取特征并较为准确地重构输入。AE 最理想的状态就是解码器的输出能够完美地重构出编码器的输入。自编码器常用的一种损失函数为：

$$\min_{AE} \| AE(W) - W \|_2 \tag{2.13}$$

其中，$\|.\|_2$ 表示 L_2 范数。

在时间序列异常检测场景下，异常数据对于正常数据来说只是少数，所以，如果使用自编码器重构出来的输出 AE(W)跟原始输入的差异超出一定阈值（threshold），原始时间序列即存在了异常。使用自编码器进行数据异常检测，在使用训练数据进行模型训练时，要保证训练数据中没有或只有少量异常数据。自编码器的基本结构如图 2.4 所示：

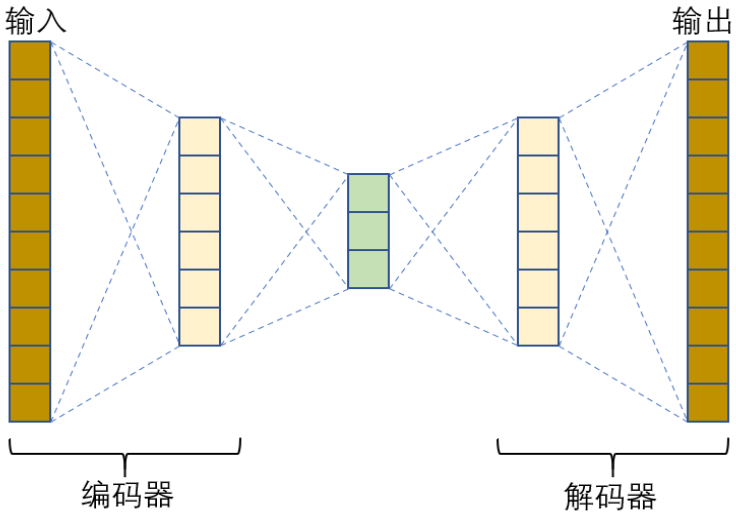


图2.4　自编码器基本结构图

随着神经网络的不断发展和成熟，自编码器的相关变型相继被提出，且对于不同场景和数据表现出了各自的优势。

深度自编码器是通过增加 AE 的网络层数使编码器能够提取出更加重要的数据特征并更多的减小干扰信息的影响，进而获得更加有效的隐变量低维特征表示。

AE 在进行数据的降维和重构时易受到噪声的干扰，影响重构数据的结果，造成异常检测的失准。深度自编码器是一种通过增加自编码器网络层数来减小噪声干扰的方法。此外，为了更好地缓解 AE 易受噪声干扰的问题，研究人员设计了降

噪自编码器(Denoising AutoEncoder, DAE) [53]。DAE 的目的是对有噪声的数据进行编码、重构并取得较小的重构误差，提取出更加能够代表正常数据的数据特征，增强网络的鲁棒性。DAE 的关键是加强网络从带有噪声的数据中提取能够表示正常数据特征的能力。DAE 网络架构如图 2.5 所示。DAE 的功能实现最重要的过程是对输入数据进行加噪处理，即对输入数据的某些信息进行破坏，然后将带有噪声的数据输入到编码器进行特征提取和解码器进行数据重构，其过程与自编码器相同。

DAE 的训练目标函数为：

$$\min_{DAE} \| DAE(W) - W \|_2 \tag{2.14}$$

DAE 相较于 AE 具有更有效地提取原始输入数据低维特征表示的能力，并重构出还原度较高的输入数据，具有更好的鲁棒性。

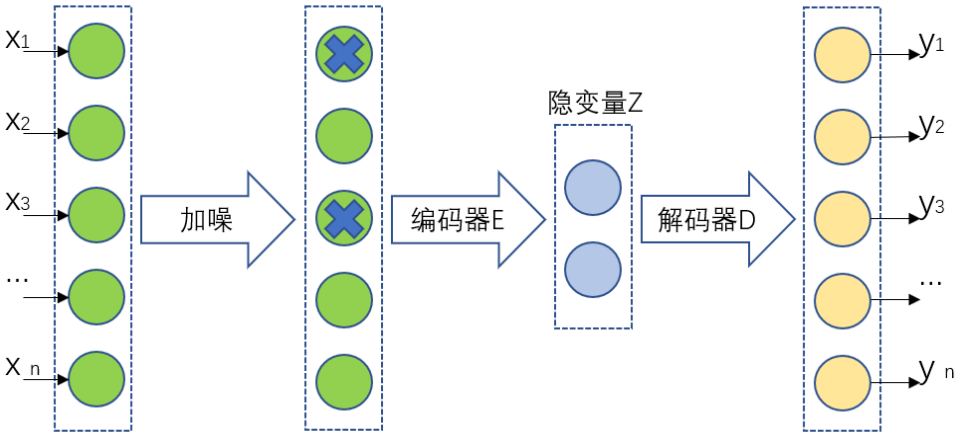


图2.5　降噪自编码器网络架构图

稀疏自编码器(Sparse AutoEncoder, SAE)[54]是一种在隐藏层神经元施加稀疏性约束的改进自编码器。其基本思想是通过在损失函数中加入适当的函数项达到减少编码器中活动神经元的数量的效果。SAE 的实现过程如图 2.6 所示。对于施加稀疏性约束 KL 的 SAE，其优化目标有两个，一是极小化输出与输入的差异，二是控制$\bar{\rho}$与 ρ 的 KL 散度最小：

$$\min_{SAE} \| SAE(W) - W \|_2 + \beta \sum_{j=1}^{h} KL(\rho \| \bar{\rho}_j) \tag{2.15}$$

其中，$\bar{\rho}_j$ 指平均激活值，h 为神经元总个数，β 为 KL 散度惩罚项在优化目标中的比重。

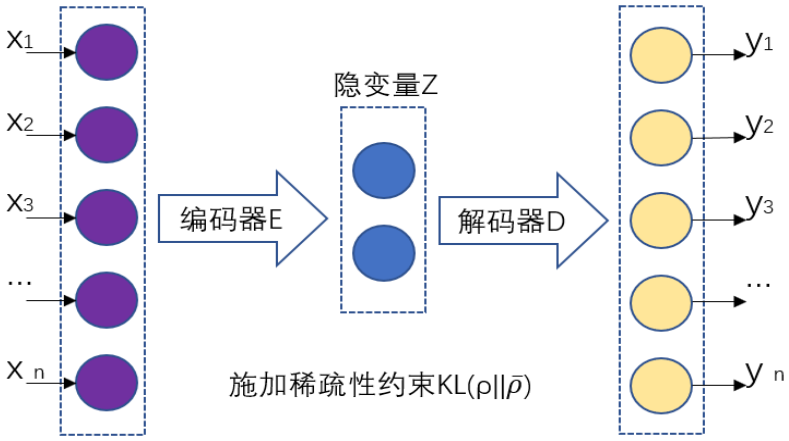


图2.6　稀疏自编码器基本架构图

变分自编码器(VAE)[错误!未找到引用源。]是变分推断和神经网络结合而成的一种方法，被广泛应用于数据分析等领域。它不同于上述各种自编码器，其在常规的自编码器的基础上中间是一个分布，VAE 从分布中采样然后输入到解码器中，重构出与输入数据相似但不完全相同的数据，使得解码器输出结果能够对噪声更有鲁棒性。VAE 的实现过程如图 2.7 所示。VAE 编码器的作用是计算出每个数据点的均值和方差。VAE 解码器的作用是对隐变量进行解码，通过计算每个数据点的概率分布来重构输入数据。VAE 生成的隐变量的各维特征之间的相关性不强，因此在用作特征编码时相比普通 AE 更能提取有效信息。

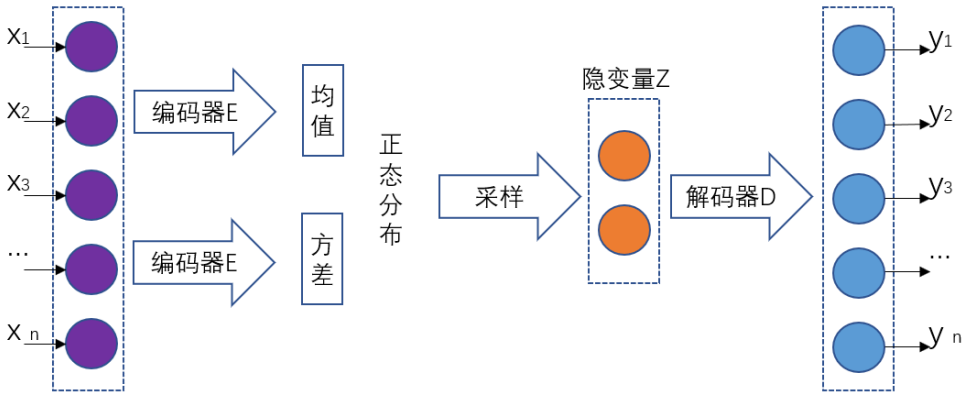


图2.7　变分自编码器基本架构图

通过上述分析的几种自编码器模型，考虑到所针对的数据和问题，降噪自编码器是更合适的一种自编码器。

使用各种类型的自编码器进行数据异常检测的过程基本上是相同的，其主要步骤如下：

步骤 1：根据原始数据选择合适的自编码器类型。

步骤 2：训练数据经过预处理之后输入到自编码器的编码器，不同类型的自编码器处理过程不同，将编码器生成的隐变量输入到自编码器的解码器进行数据的

解码重构，得到与原始数据结构一致的重构数据，并根据训练数据和重构数据之间的关系计算异常得分。

步骤 3：根据训练数据在模型中的表现和异常得分设置阈值。

步骤 4：将测试数据输入到模型，计算测试数据的异常得分并与阈值进行比较，将异常得分大于阈值的数据判定为异常数据。

2.4 海洋观测数据预测方法

海洋观测数据质量控制的另一个重要方面是异常值的校正处理，使用机器学习算法进行数据预测进而实现数据的异常值校正是一种重要方式。考虑到海洋观测数据作为时间序列数据具有时间特性，本节选择和分析了适用于时间序列数据预测的 ARIMA 算法和 LSTM 算法，并设计了两种机器学习算法基本的数据预测过程步骤。

2.4.1 基于 ARIMA 的数据预测方法

差分自回归滑动平均(ARIMA)模型由 Box 和 Jenkins 二人提出[错误!未找到引用源。]。ARIMA 被广泛应用于时间序列数据的预测，其模型结构简单，不需要借助其他变量，只需要自身变量的处理。ARIMA 模型的思想是对时间序列数据建立数学模型，然后对过去的数据和现在的数据行近似描述，并使用这个模型预测未来数据。ARIMA 模型的使用有一定的限制，其要求原始输入数据必须为平稳序列，对于非平稳数据，需要对数据进行差分处理，在得到平稳序列后才能作为模型的输入进行未来预测。此外，ARIMA 模型要求输入数据必须是单变量序列。

ARIMA(p, d, q)有三个重要的建模参数，分别为：自回归模型(AR)中的自回归项数 p，滑动平均模型(MA)中的滑动平均项数 q 和将非平稳时序转换为平稳时序的差分阶数 d。

AR 模型认为当前时刻和前面的 p 个时刻有关，其表达式为：

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + u_t \tag{2.16}$$

其中 ϕ 为自回归系数，u_t 表示白噪声，是时序中的随机波动成分。

MA 模型关注的是 AR 模型中的误差项的累加，滑动平均法能有效地消除预测中的随机波动。若时序中的白噪声序列为$\{u_1,u_2,\ldots,u_n\}$，则 q 阶滑动平均模型的表达式为：

$$x_t = \sum_{i=1}^{q} \theta_i u_{t-i} + u_t \tag{2.17}$$

其中，Θ 为滑动平均系数，u_t 是不同时期的白噪声。

自回归滑动平均模型(ARMA)是 MA 模型于 AR 模型的结合，能够同时解决当前数据与后期数据之间的关系和随机波动。ARMA 模型的表达式为：

$$x_t = \sum_{i=1}^{p} \varphi_i x_{t-i} + u_t + \sum_{i=1}^{q} \theta_i u_{t-i} \tag{2.18}$$

ARMA 模型只能处理平稳时间序列，对于非平稳时间序列则需要引入差分项 d。引入差分项 d 的 ARMA 模型则成为 ARIMA 模型。ARIMA 模型的表达式为：

$$(1-\sum_{i=1}^{p} \varphi_i L^i)(1-L)^d X_t = (1+\sum_{i=1}^{q} \theta_i L^i)\varepsilon_t \tag{2.19}$$

其中，L 为滞后算子，d 为正整数。

差分算子的表达式为：

$$\nabla^d x_t = (1-L)^d x_t \tag{2.20}$$

令：

$$w_t = \Delta^d x_t = (1-L)^d x_t \tag{2.21}$$

则 ARIMA 模型的表达式为：

$$w_t = \sum_{i=1}^{p} \varphi_i w_{t-1} + \sigma + u_t + \sum_{i=1}^{q} \theta_i u_{t-i} \tag{2.22}$$

基于 ARIMA 的海洋观测数据预测方法的主要步骤如下：

步骤 1：导入数据，划分训练集和测试集并进行平稳性检验，如果序非平稳，则进行 d 阶差分处理。

步骤 2：遍历搜索 AIC 或者 BIC 最小的参数组合，确定 ARIMA 的 p、q、d 参数，得到 ARIMA 模型并建模预测。

2.4.2 基于 LSTM 的数据预测方法

长短期记忆神经网络(LSTM)于 1997 年由 Hocjreiter 和 Schmidhuber 提出[错误!未找到引用源。]。LSTM 是循环神经网络[错误!未找到引用源。]（RNN）的一种，LSTM 对 RNN 神经网络的隐藏层进行改进形成 LSTM 隐藏层神经元，使 LSTM 在保留 RNN 记忆功能的基础上能选择性地遗忘数据信息并更新记忆单元，克服 RNN 模型的长期依赖问题，进而在一定程度上解决 RNN 存在的梯度弥散[59]。LSTM 模型结构如图 2.8 所示。

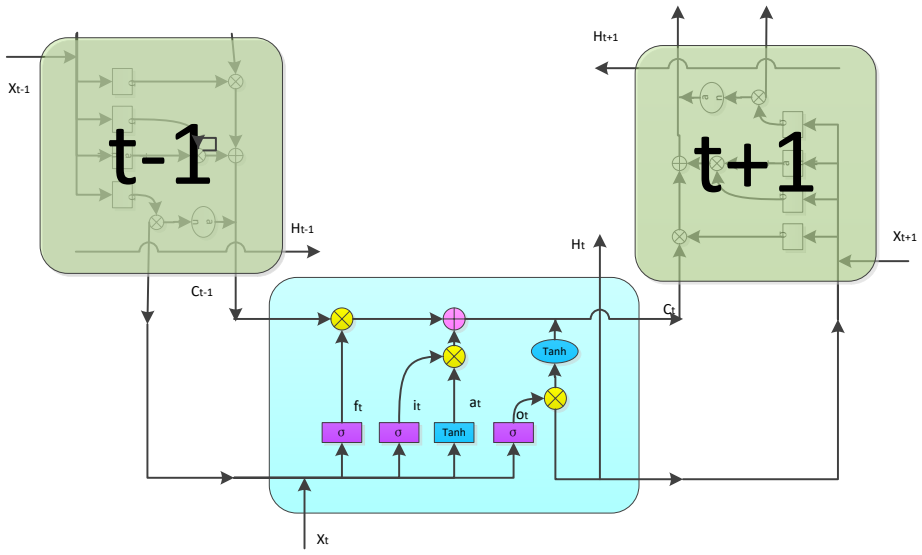


图2.8 LSTM 模型结构图

LSTM 中主要包含单元状态 C_t，细胞状态更新以及三个门结构，LSTM 门结构的作用是控制单元状态的信息获取，主要包括对信息的添加和删除，三个门结构分别是：输入门，遗忘门和输出门。

遗忘门的作用是对信息的限制和删除，它是以一定的概率遗忘上一层的隐藏单元状态，输出 f_t 就代表了这个概率，其数学表达式为：

$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f) \tag{2.23}$$

输入门负责处理当前序列位置的输入，并对单元状态进行更新，输出 i_t 和 a_t 的数学表达式为：

$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i) \tag{2.24}$$

$$a_t = \tanh(W_a h_{t-1} + U_a x_t + b_a) \tag{2.25}$$

遗忘门和输入门的结果都会作用于的单元状态 C_t，其数学表达式为：

$$C_t = C_{t-1} f_t + i_t a_t \tag{2.26}$$

输出门的作用是决定输出单元状态的哪些部分，输出 o_t 和 H_t 的数学表达式为：

$$o_t = \tanh(W_o h_{t-1} + U_o x_t + b_o) \tag{2.27}$$

$$H_t = o_t \tanh(C_t) \tag{2.28}$$

其中，H_{t-1} 是上一个单元的输出；x_t 是 t 时刻的输入。

基于 LSTM 的海洋观测数据预测方法的主要步骤如下：

步骤 1：数据预处理，根据实验数据设置滑动窗口的大小，并将原始数据划分为训练集和测试集；

步骤 2：建立 LSTM 模型，根据训练集数据确定 LSTM 的层数，节点数等参数值；

步骤 3：将当测试数据导入模型，预测未来数据，进行模型验证。

2.5 本章小结

本章主要是对海洋观测数据质量控制的总体技术路线和关键技术的分析，首先介绍了时间序列的基本概念，说明海洋观测数据作为一种时间序列在本文应用中可分为单变量海洋观测时间序列和多变量海洋观测时间序列。然后对海洋观测数据质量控制的总体技术路线进行了分析和设计，对本文的研究进行了内容上的划分。针对海洋观测数据存在的时间特性，分析和介绍了适用于时间序列分解的 STL 分解算法，针对数据的特征维度和数据特征所包含的数据信息，介绍了小波分解和重构算法的基本概念。另外，本章介绍了几种常用的数据异常检测方法的原理及其计算或检测过程，为后文海洋观测数据异常检测模型的设计提供理论依据，最后本章介绍了几种时间序列数据预测算法的基本原理、结构以及模型建立过程。

第 3 章 海洋观测数据异常检测方法

对于海洋观测设备采集到的观测数据，首先需要对其进行格式检查，确保数据格式的正确性，保证数据的可用性。在得到格式无误的数据后对于海洋观测数据的质量控制，第一步要做的就是海洋观测数据异常值检测，将缺失和异常的数值检测出来并进行标记。本章介绍了海洋观测数据异常检测的相关方法，并在第 2 章所述异常检测基本算法的基础上，针对单变量海洋观测数据设计了基于统计学方法和局地异常检测以及误差控制的海洋观测数据异常检测模型，针对多变量海洋观测数据设计了基于自编码器的海洋观测数据异常检测模型，并进行了模型验证、实验测试和结果分析，最后是本章小结。

3.1 海洋观测数据异常检测的常规方法

海洋观测领域对海洋观测数据进行异常检测的常规流程是通过日期时间检查、观测位置检查、观测要素范围检查、连续性检查、统计特性检查、相关性检查等方法将缺失和不符合海洋观测要素规律的数据检测出来，然后对异常的观测数据做质量标记。主要的海洋观测数据异常检测方法及方法介绍如表 3.1 所示。

表 3.1 海洋观测数据异常检测的常规方法

主要检测方法	方法介绍
日期检查	海洋观测日期时间的取值应位于合理范围内。
位置检查	海洋观测数据的观测设备所在位置应在合理取值范围内。
观测要素范围检查	简称范围检查，主要是对已有的国内外海洋观测数据进行统计分析，根据观测要素自身的特点，定义要素的取值变化范围，对数据进行检验。如果超出设定的正常范围，则认为该数据为异常数据。
统计特性检验	海洋观测数据在理论上往往服从于一定的概率统计特性，数据对应的随机变量和随机过程是相互独立并且服从一定的分布，时间序列资料对应的随机过程也是平稳的或周期性的。根据数据的这些特性，建立分布拟合函数，进行卡方拟合有度检验，最后采用轮次检验方法检验数据是否是独立的，独立的数据往往都是异常值。
气候特性检验	根据海域海洋环境气候要素季节性变化和日变化的特点，检验观测数据是否满足其季节性变化和日变化特性。

续表 3.1 海洋观测数据异常检测的常规方法

主要检测方法	方法介绍
相关性检验	根据海洋观测资料数据间的相互关系进行检验，即通过要素间的相互关系检验数据的异常。
连续性检验	海洋观测要素在一定时空范围内具有连续性，时间接近或者位置临近的观测要素差值在一定范围内。
极值检验	一般情况下，定点定时要素观测值的取值应在该地该要素的多年极值范围内。
梯度检验	对观测要素进行梯度检验，如果梯度较大，则判定其为异常值，需进一步分析。
内部一致性检验	同一时间观测的气象要素之间的关系符合一定物理联系的检验。
时间一致性检验	气象记录在一定时间范围内的变化是否具有特定规律的检验。
空间一致性检验	气象记录在一定空间范围内的变化是否符合其空间规律的检验。

3.2 基于统计学方法的海洋观测数据异常值检测方法

统计学方法异常检测主要适用于单变量的序列数据。基于统计理论的数据异常值检测方法认为正常的数据对象遵守统计模型的规律，而异常点明显不遵守该统计模型，因此对于大量数据中的单个异常数据，统计学方法能够较为准确地检测并进行异常判定。但是异常值在数据中的表现方式十分多样且复杂，通常以单个数值突变的方式、以连续多个数据脱离数据整体分布的方式或以大量数据块异常等方式出现，数据块异常在局部时间内异常值的数量可能会出现超过正常数据数量的现象，导致数据块异常成为统计学方法最难以判别异常的一种情况。这时，使用基于统计理论的异常值检测方法就难以做出有效的异常判定甚至产生大量的误判。此外，由于观测仪器本身存在观测误差，使用统计学方法判定为异常值的数据也可能存在错判的情况。本节中以格拉布斯准则（Grubbs 准则）和 3σ 准则为基础，考虑到使用 Grubbs 准则和 3σ 准则异常检测可能存在的漏判和误判问题，结合局地异常值检查和误差控制方法，提高基于统计学方法异常检测的可靠性。

3.2.1 基于 Grubbs 准则和 3σ 准则结合局地异常检测和误差控制的异常检测模型设计

基于统计理论的异常数据检测方法在满足一定数据量的前提下具有稳定性和准确性。具有比较广泛的应用范围。但是对于异常分布复杂的数据，仅仅使用统计学方法进行异常检测并不能取得良好的效果，因此在使用统计学方法进行检测后加入局地异常值检测是对基于统计学理论异常检测模型的补充与保障。局地异常检测是通过一定的数据模型对几个相邻数据的比较，对局部数据进行较为准确

的异常判定。局地异常值检测方法根据如下公式来设计：

$$|x_n-(x_{n-1}+x_{n+1})/2|-|(x_{n+1}-x_{n-1})/2|\geq\beta \quad (3.1)$$

β 是根据不同的观测要素、不同的应用领域进行设置的判定系数。使用局地异常检测公式，首先要假定 x_{n-1} 和 x_{n+1} 均为正常数据，然后通过上述公式比较认定异常数据。但是对于 x_{n-1} 和 x_{n+1} 中存在异常数据的情况，使用上述公式进行异常判定，x_n 就可能会出现误判，这是局地检测方法的缺点。

此外，由于观测本身存在的观测误差，也可能导致观测数据在使用统计学方法做异常检测时出现误判，所以需要加入观测设备的误差控制。

误差控制是对上述过程判定为异常值的数据进行的检测，被判定为异常值的数据与相邻正常数据的差值只有不小于设备的观测误差，才认定该数据为异常值，否则认为该异常值为错判，将其重新判定为正常值。不同观测要素的误差控制需要根据其特性设计。

本节异常检测只针对单变量的海洋观测数据，对于获得海洋观测数据，在经过简单的数据预处理之后，得到能够被模型识别的有效数据。首先，构建滑动窗口将数据切分为多个子数据集，滑动窗口的大小和滑动步长基于 Grubbs 准则的要求和数据的特性进行选择，然后根据 Grubbs 准则的计算过程建立基于 Grubbs 准则的异常检测模型，将检测到的异常值做质量标记，再根据 3σ 准则的要求建立基于 3σ 准则的异常检测模型并做质量标记。将两个模型结合，把两个模型检测到的异常值都视为异常值并做质量标记。之后根据本节介绍的局地异常检测方法公式建立局地异常检测模型，主要检测经过上述模型检测出的异常值是否存在误判。最后根据观测要素的数据特性设置误差控制范围，构成基于 Grubbs 准则、3σ 准则、局地异常检测和误差控制的海洋观测数据质量控制模型，其流程如图 3.1 所示：

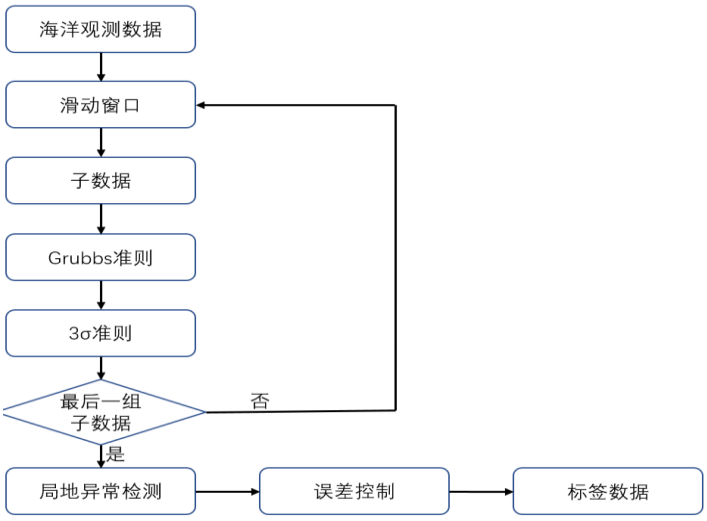


图3.1 基于统计学方法、局地异常检测和误差控制的海洋观测数据质量控制模型流程图

3.2.2 实验数据

本节方法可以用于海水温度、波浪等海洋观测数据的异常检测。本节实验使用 W70X 浮标波浪数据，选取 W70X 浮标波浪数据中的有效波高数据作为本节实验的单变量海洋观测数据进行异常检测，验证方法的有效性。本节从 W70X 浮标波浪数据中截选 2019-01-24 08:38:00 开始到 2021-04-25 10:38:00 为止以 1 小时为时间采样的 17127 个数据，其数据如表 3.2 所示。首先将选取实验的数据添加噪声，噪声的添加选择高斯噪声，并通过调整高斯噪声的均值和方差，使之产生 161 个异常值并整数化异常值，其数据散点图如图 3.2 所示，图 3.2 中黑色表示原始有效波高正常数据，红色表示添加噪声产生的异常数据。

表 3.2 W70X 浮标有效波高数据

数据采集时间	有效波高（分米）
2021/4/25 10:38	8
2021/4/25 9:38	7
2021/4/25 8:38	7
2021/4/25 7:38	9
2021/4/25 6:38	8
……	……
2019/1/24 12:38	8
2019/1/24 11:38	7
2019/1/24 10:38	7
2019/1/24 9:38	7
2019/1/24 8:38	8

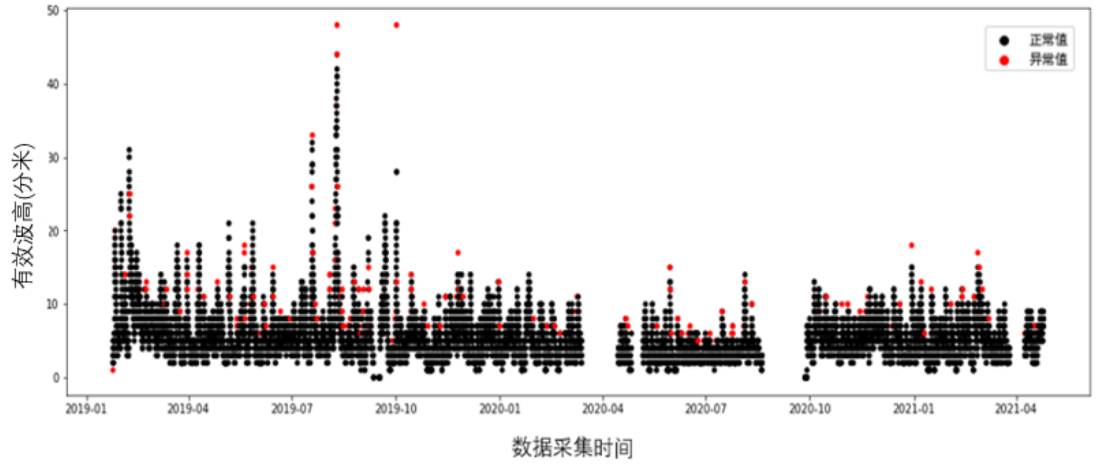


图3.2　加入噪声产生异常数据的有效波高数据散点图

将上述有效波高数据划分为训练数据和测试数据，以前 13700 个数据作为训

练集数据用来训练模型，确定基于统计学方法、局地异常检测和误差控制的异常检测方法的各个参数，将剩余数据作为测试集验证模型的有效性。其中，训练集数据包含 141 个异常值，测试集数据中包含 20 个异常值。

3.2.3 模型参数选择

基于 Grubbs 准则、3σ 准则、局地异常检测和误差控制的海洋观测数据异常检测模型的参数主要有滑动窗口的大小 k，滑动步长 m，置信概率参数 p，Grubbs 测量次数 n，局地异常检测的临界值系数 β，误差控制波动参数 α 和误差控制范围参数 b。在本节模型中，测量次数 n 就是滑动窗口步长 k。

对于本节实验使用的波浪数据，需要通过经验和不断地修改参数进行实验来确定合适的参数数值，最终本节模型选定滑动窗口大小 k=100，测量次数 n=k=100，滑动步长 m=20，置信概率参数 p=5，局地异常检测临界值系数 β=2.1，误差控制波动参数 α=0.2 和误差控制范围参数 b=0.1，这时，模型的异常检测准确率较高。

3.2.4 实验结果与分析

本节实验通过异常检测后的数据与加入噪声的有效波高数据进行对比，对比使用 Grubbs 准则的异常检测方法、Grubbs 准则结合 3σ 准则的异常检测方法、Grubbs 准则与 3σ 准则结合局地异常检测的异常检测方法、Grubbs 准则与 3σ 准则结合局地异常检测和误差控制的异常检测方法，得出每一种方法的异常检测结果，判断其正确检测和误判漏判现象，分析异常检测方法的有效性和准确性。

图 3.3 为在训练集数据上使用 Grubbs 准则进行异常检测的结果图，由图可知，使用 Grubbs 准则的异常检测模型检测出 64 个异常数据，数据存在漏判且存在错判，所以加入 3σ 准则继续对数据进行异常检测，此时，使用 3σ 准则的异常检测模型检测出 221 个异常数据，如图 3.4 所示，数据产生了大量的错判且存在漏判，结合两种方法的异常检测模型共检测出 221 个异常数据，如图 3.5 所示，结果说明使用 3σ 准则的异常检测方法检测出的异常数据包含使用 Grubbs 准则检测方法检测出的异常数据，但数据仍存在漏判和更多的错判，模型效果不佳。但是，在两种方法结合的基础上使用局地异常检测方法，由图 3.6 可见，数据漏判现象明显减少，更多的是统计学方法产生的数据错判，局地异常检测方法在统计学方法之后又检测出 34 个异常数据。误差控制方法用来控制异常检测的错判现象，如图 3.7 所示，判为异常的数据个数为 137 个，模型消除了大量的错判现象，说明本节使用的海洋有效波高数据在使用统计学方法进行异常检测时大量在误差控制范围内的正常数据被判为了异常。此时，经过本节方法的异常数据与加入噪声产生的异常数据个数相差 4 个且为漏判。结果表明，结合上述方法的有效波高数据异常检测方法在训练集上具有较高的准确性。

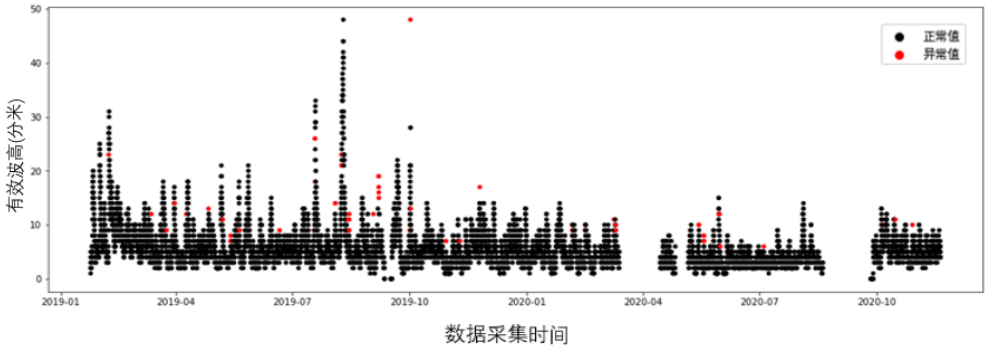


图3.3　训练集上使用 Grubbs 准则的有效波高数据异常检测结果图

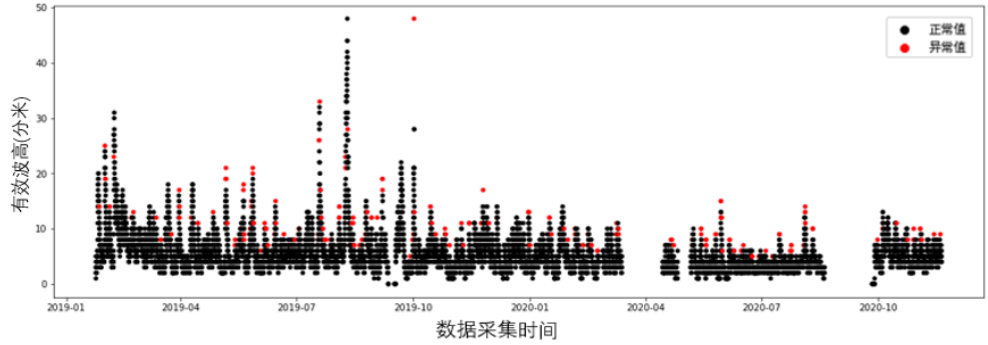


图3.4　训练集上使用 3σ 准则的有效波高数据异常检测结果图

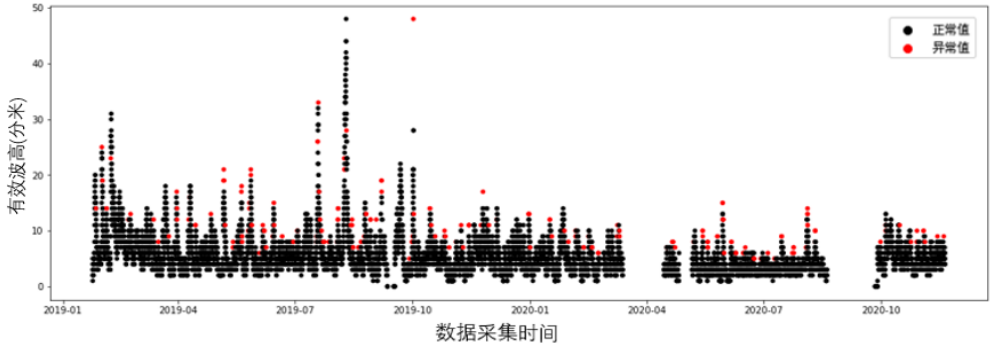


图3.5　训练集上结合 Grubbs 准则和 3σ 准则的有效波高数据异常检测结果图

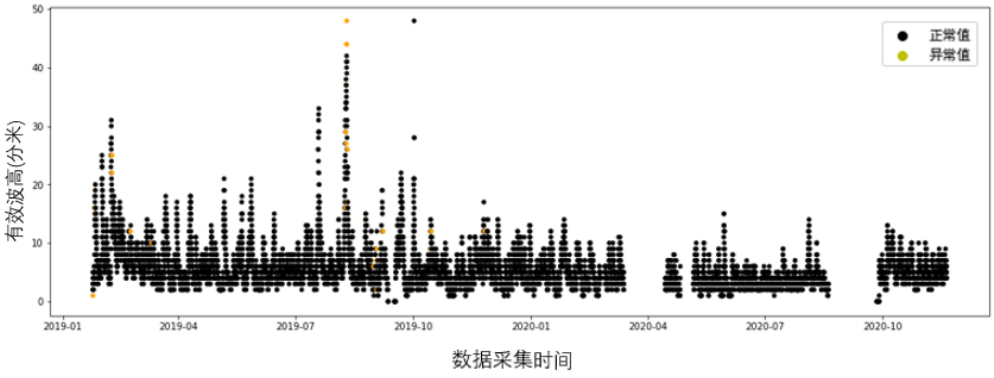


图3.6　训练集上使用局地异常检测方法的有效波高数据异常检测结果图

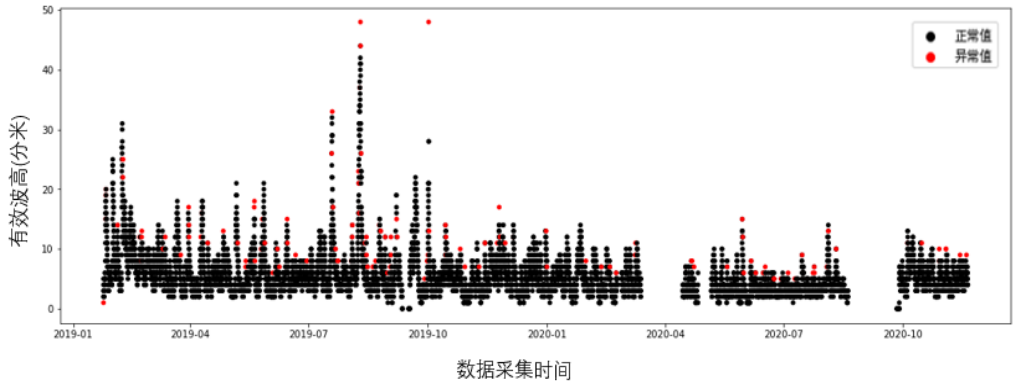


图3.7　训练集上经过误差控制的有效波高数据异常检测结果图

将测试集数据导入模型，经过上述方法建立的模型在测试集数据上共检测出19 个异常数据，如图 3.8 所示，异常检测结果只存在 1 个漏判数据，说明模型在测试集数据上仍具有较高的准确性和有效性。基于格拉布斯准则、3σ 准则结合局地异常检测和误差控制的整个过程的检测结果如表 3.3 所示：

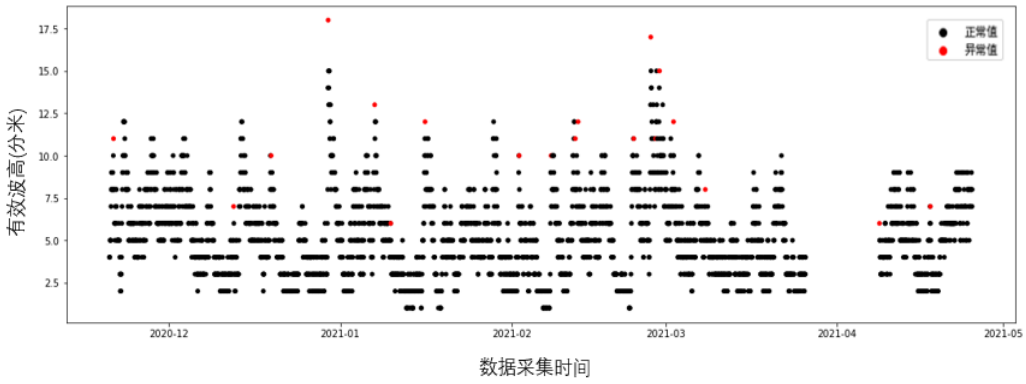


图3.8　测试集上基于统计学方法、局地异常检测和误差控制的有效波高数据异常检测结果图

表 3.3　基于格拉布斯准则、3σ 准则结合局地异常检测和误差控制的检测结果

数据 / 检测方法	数据数量					
	训练集数据	测试集数据	训练集异常数据	测试集异常数据	训练集上检测到的异常数据	测试集上检测到的异常数据
格拉布斯准则	13700	3427	141	20	64	
3σ 准则	13700	3427	141	20	221	
Grubbs 准则结合 3σ 准则	13700	3427	141	20	221	
局地异常检测	13700	3427	141	20	34	
误差控制方法	13700	3427	141	20	137	19

注：测试集上检测到的异常数据数量下的斜线表示模型无须分别显示各个过程异常检测结果，只须统计测试集数据通过异常检测模型得到的最终异常检测结果。

3.3 基于自编码器的海洋观测数据异常值检测方法

对于单变量的海洋观测数据，传统的一些异常检测模型和基于统计学方法的异常检测模型就已经实现了较好的检测性能。然而，统计学方法主要适用于单变量数据，且对于数据维度多，数据量较大的海洋观测时间序列数据，传统的异常检测方法不能很好地描述数据分布，检测效率也比较低。对于多变量的海洋观测数据，借助机器学习算法较强的特征提取和信息表示能力，通过选择合适的机器学习算法建立异常检测模型进行数据的异常检测是有效的缓解上述问题的重要方式。本节提出了一种基于自编码器的异常检测方法，实现多变量海洋观测数据的异常检测。

3.3.1 基于自编码器的海洋观测数据异常检测模型设计

基于重构误差的异常检测作为深度学习的一种重要应用被广泛使用。利用自编码器进行数据的编码和重构来实现数据的异常检测是一种重要的方法。海洋观测数据作为一种复杂的时间序列数据，同样可以使用自编码器进行异常检测，但是时间序列数据首先要解决的问题就是数据的时间依赖性，否则会造成模型准确的降低。本节通过设计两个串行连接的自编码器来简化模型，并通过 LSTM 神经网络提取数据的时间特性，实现海洋观测数据的异常检测。

基于自编码器的海洋观测数据异常检测模型是根据待检测数据的重构误差来进行异常检测的。为了能够更好的提取海洋观测时间序列的时间特性，本节模型的两个自编码器（AE_1，AE_2）的网络结构都由 LSTM 设计构成，将两个自编码器串行连接，AE_1 的解码器输出作为 AE_2 编码器的输入，数据通过自编码器模型，正常数据的重构误差较小，而异常数据的重构误差被放大，其数值表现出与正常数据重构误差明显的差异，基于异常数据重构误差的放大，自编码器能够实现更加准确有效地异常数据判别。本节模型中两个自编码器 AE_1 和 AE_2 的目标函数表达式分别为：

$$\min_{AE_1} \| AE_1(W) - W \|_2 \tag{3.2}$$

$$\min_{AE_2} \| AE_2(AE_1(W)) - W \|_2 \tag{3.3}$$

在建立了串行连接的自编码器模型后，基于此模型构建新的自编码器实现数据的异常检测。本节基于自编码器方法的海洋观测数据异常检测模型结构如图 3.9 所示：

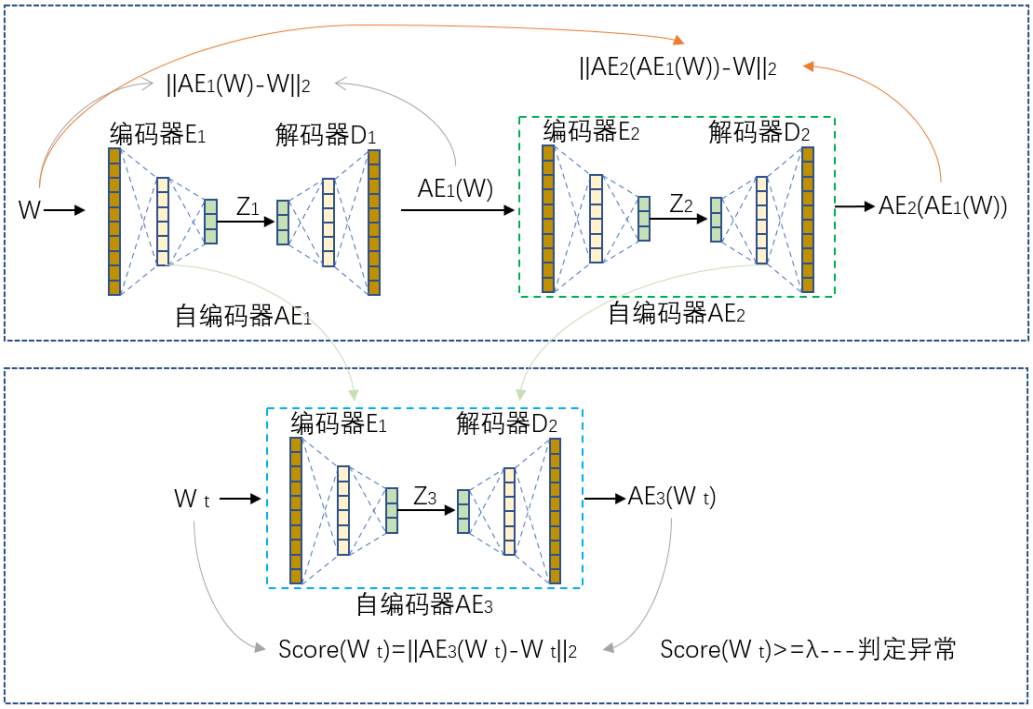


图3.9　基于自编码器的海洋观测数据异常检测模型

其中，W 表示用于训练的海洋观测数据通过滑动窗口得到的子数据集，Z_1、Z_2、Z_3 分别为自编码器的隐变量，W_t 表示待检测的海洋观测数据通过滑动窗口得到的子数据集。

本节基于自编码器的海洋观测数据异常检测模型设计分为两个阶段。阶段一是使用训练数据集训练，通过两个串行连接的自编码器（AE_1，AE_2）进行海洋观测数据的降维特征提取和重构并优化模型的参数。降维的目的是抑制数据中的干扰信息，提取更加有效的低维特征表示，重构是利用隐变量的低维特征升维得到与原始数据结构相同的重构数据。阶段二是使用测试数据集进行异常值的检测，将 AE_1 的编码器 E_1 和 AE_2 的解码器 D_2 结合组成新的自编码器 AE_3，通过 AE_3 计算数据的异常得分，进而通过异常得分与设置的阈值大小比较判定数据是否异常。本节模型将 AE_1 的编码器 E_1 和 AE_2 的解码器 D_2 结合组成新的自编码器 AE_3 是因为仅含编码器 E_1 可防止抑制过多的干扰信息，同时解码器 D2 对正常数据的解码能力更强，能够在放大异常序列重构误差的同时减小正常序列的重构误差。

本节模型数据的异常得分表达式为：

$$score(W_t) = \| AE_3(W_t) - W_t \|_2 \tag{3.4}$$

本节方法的具体实现步骤如下：

步骤 1：数据预处理，将数据中格式错误、时间地点等错误的数据调整为适合

输入到异常检测模型的数据结构。

步骤 2：将经过预处理的海洋观测数据划分训练集 W 和测试集 W_t。

步骤 3：构建两个自编码器 AE_1 和 AE_2，并将它们串行连接，如图所示，将 W 作为 AE_1 的输入，然后将其输出 $AE_1(W)$作为 AE_2 的输入，得到重构输出 $AE_2(AE_1(W))$。

步骤 4：使用 AE_1 和 AE_2 的目标函数公式优化 AE_1 和 AE_2 的参数，得到较为理想的模型参数。

步骤 5：利用步骤 3 和步骤 4 建立的串行连接的自编码器 AE_1 和 AE_2，将 AE_1 的编码器 E_1 和 AE_2 的解码器 D_2 重新组合成为新的自编码器 AE_3，将 W_t 输入到自编码器 AE_3 实现对 W_t 的编码和重构，得到重构结果 $AE_3(W)$。

步骤 6：计算异常得分 score(W_t)

步骤 7：将计算得到的 score(W_t)与设置的阈值 λ 进行比较，若 score(W_t) >λ, 则将其判定为异常，反之则为正常数据。

3.3.2 实验数据

本节方法主要针对多变量的海洋观测数据，可以用于海水温度、波浪等海洋观测数据的异常检测。本节实验使用浮标观测数据，从中选择温度、盐度、深度、氧含量、浊度和 YLS 六种观测要素作为本节的实验数据，验证基于自编码器的多变量海洋观测数据异常检测方法的有效性。本节截选从 2022-04-11 18:00:00 开始到 2022-05-23 09:00:00 为止以 1 小时为时间采样的 1000 个数据，其数据格式如表 3. 4 所示。

表 3.4 实验所用浮标观测海洋观测数据格式

数据采集时间	温度(℃)	盐度(S)	深度(m)	氧含量(ppt)	浊度(ftu)	YLS(spad)
2022/5/23 9:00	15.71	30.87	38.93	7.52	100	1.3
2022/5/23 8:00	25.50	10.87	18.74	7.63	110.3	2.1
2022/5/23 7:00	15.35	30.87	38.63	7.59	211.5	1.8
2022/5/23 6:00	15.26	30.89	38.56	7.62	159.3	2.3
2022/5/23 5:00	15.35	30.89	38.63	7.23	167.2	1.6
……	……	……	……	……	……	……
2022/4/11 22:00	10.16	29.98	33.22	4.98	93.5	2.4
2022/4/11 21:00	10.12	29.98	33.19	4.83	48.3	2.9
2022/4/11 20:00	10.24	29.98	33.29	4.86	25.5	1.7
2022/4/11 19:00	11.40	29.96	34.24	4.48	35.8	2.2
2022/4/11 18:00	10.48	29.96	33.48	0	144.5	1.9

为满足海洋观测数据异常检测的要求，在原始数据中加入少量噪声，构成实验所要检测出的异常数据。使用自编码器进行数据的异常检测需要训练集数据尽可能少的异常，所以添加噪声的方式选择随机地在海洋观测数据的某个要素上添

加噪声。多变量的海洋观测数据的异常表现不同于单变量海洋观测数据，多变量时间序列数据的异常不仅表现在数据的时间特性上，而且表现在不同变量的相关性上，比如某一时刻数据的所有要素全都表现出与前后时刻不同的变化，此时数据不一定为异常值，但是，如果数据某一时刻只有一个要素数据突变，此时，这组时间序列数据更可能是异常数据。根据上述分析，我们在原始数据中加入了 15 个随机噪声，作为实验的异常数据。本节实验将原始数据集划分为训练集和测试集，其中，训练集为数据的前 769 个数据，训练集为剩余数据，加入的随机噪声在训练集中含有 8 个，在测试集中包含 7 个。

3.3.3 模型参数选择

对于本节实验使用的自编码器 AE_3 的构建基于两个串行连接的自编码器 AE_1 和 AE_2，我们采用自编码器 AE_1 的编码器 E_1 和自编码器 AE_2 的解码器 D_2 组成，其中，AE_3 的编码器部分为两个 LSTM 层构建的网络，解码器部分为与编码器结构对称的两个 LSTM 网络，采用 LSTM 构成自编码器的目的主要是 LSTM 神经网络能够捕捉原始海洋观测数据的时间依赖性。编码器两层 LSTM 网络分别确定其参数为 16 个细胞单元和 4 个细胞单元，激活函数选择使用 relu；解码器的两层 LSTM 网络的参数为 4 个细胞单元和 16 个细胞单元。通过观察在训练集上的数据异常得分分布图，如图 3.10 所示，训练集上的异常得分大多小于 0.2，然后根据经验和实验确定异常检测的阈值为 0.22。

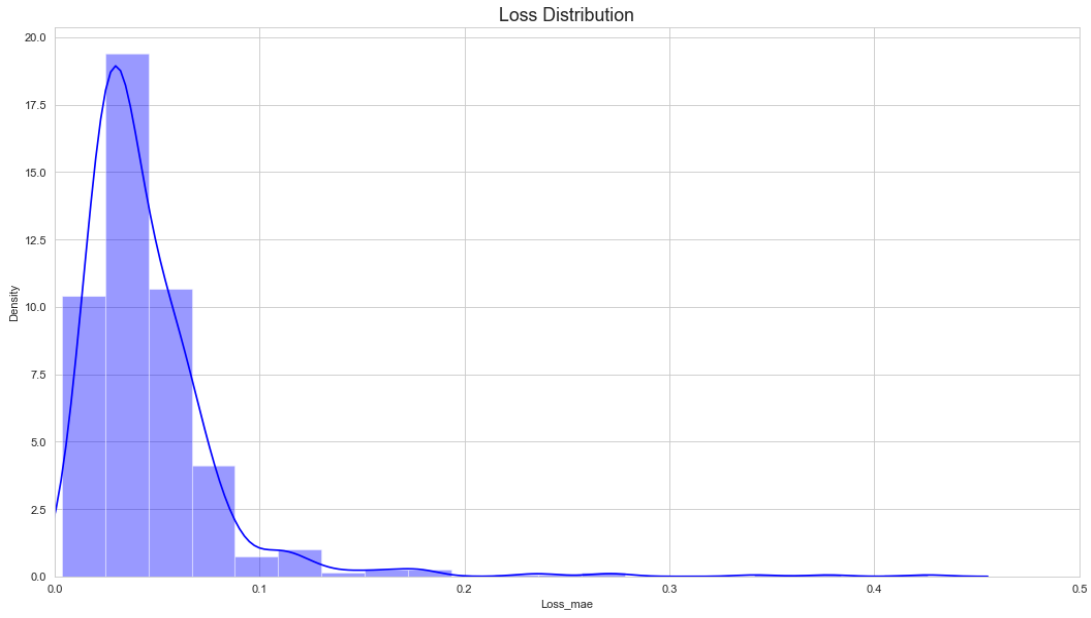


图3.10　训练集上的异常得分分布图

在训练模型之前，需要对模型的学习过程进行配置，即确定模型的的优化器和损失函数。本节使用 Python 进行实验，其中定义了十多种优化器种类，各种优

化器的信息如表 3.5 所示。Adam 优化器具有参数的更新不受梯度的伸缩变化影响、更新步长和梯度大小无关、对目标函数没有平稳要求、能较好的处理噪声样本等优点，本节模型最终选用 Adam 优化器；损失函数是指用于计算标签值和预测值之间差异的函数，对于不同的任务，使用的损失函数也不相同，本节模型最终确定 mae 作为模型的损失函数。

表 3.5 python 定义的各类优化器名称及参数信息

优化器	简介	主要调节参数
Adadelta	自适应学习率方法	params(iterable)：要优化的参数；rho(float, optional)：用于计算平方梯度运行平均值的系数；lr(float, optional)：学习率；weight_decay(float, optional)：权重衰减（L2 惩罚）
Adagrad	在线学习和随机优化的自适应次梯度方法	params(iterable):要优化的参数；lr(float, optional):学习率；lr_decay(float, optional):学习率衰减指数；weight_decay(float, optional):权重衰减（L2 惩罚）
Adam	随机优化方法	params(iterable):要优化的参数；lr(float, optional):学习率；betas(Tuple[float, float], optional):用于计算梯度及其平方的运行平均值的系数；weight_decay(float, optional):权重衰减（L2 惩罚）
AdamW	解耦权重衰减正则化的 Adam 变体	params(iterable):要优化的参数；lr(float, optional):学习率；betas(Tuple[float, float], optional):用于计算梯度及其平方的运行平均值的系数；weight_decay (float, optional):权重衰减（L2 惩罚）
SparseAdam	适用于稀疏张量的 Adam 算法	params (iterable)：要优化的参数；lr (float, optional)：学习率；betas (Tuple[float, float], optional)：用于计算梯度及其平方的运行平均值的系数
Adamax	基于无穷范数的 Adam 变体	params (iterable)：要优化的参数；lr (float, optional)：学习率；betas (Tuple[float, float], optional)：用于计算梯度及其平方的运行平均值的系数；weight_decay (float, optional)：权重衰减（L2 惩罚）
ASGD	实现平均随机梯度下降	params (iterable)：要优化的参数；lr (float, optional)：学习率；lambd (float, optional)：衰减项；alpha (float, optional)：平滑指数；t0 (float, optional)：开始求平均值的点；weight_decay (float, optional)：权重衰减（L2 惩罚）
LBFGS	BFGS 针对计算复杂度很大、目标函数非凸、有可能会收敛到鞍点的情况，LBFGS 在 BFGS 的基础上针对的是所需存储巨大的情况。	lr (float)：学习率 (default: 1)；max_iter (int)：每个优化步骤的最大迭代次数；；max_eval (int)：每个优化步骤的最大函数计算次数；tolerance_grad (float)：一阶最优终止公差；tolerance_change (float)：函数值/参数变化的终止容差；history_size (int)：更新历史大小

续表 3.5 python 定义的各类优化器名称及参数信息

优化器	简介	主要调节参数
RMSprop	在添加 epsilon 之前取梯度平均值的平方根	params (iterable)：要优化的参数；lr (float, optional)：学习率；momentum (float, optional):动量系数；alpha (float, optional):平滑常数；centered(bool, optional):如果为真，计算为中心的 RMSProp，梯度将通过其方差的估计进行归一化；weight_decay (float, optional)：权重衰减（L2 惩罚）
Rprop	实现弹性反向传播算法	params (iterable):要优化的参数；lr (float, optional):学习率；etas (Tuple[float, float], optional):一对(etaminus, etaplis)，它们是乘增和减因子；step_sizes (Tuple[float, float], optional):允许的最小步长和最大步长对
SGD	实现随机梯度下降（可选带有动量）	params (iterable):要优化的参数；lr (float):学习率；momentum (float, optional):动量系数；weight_decay (float, optional):权重衰减(L2 惩罚)

3.3.4 实验结果与分析

本节实验结果通过与加入噪声的测试集上的数据进行对比，计算得到每组数据的异常得分，通过与设置的阈值进行比较，判定数据是否异常。实验设置阈值为 0.22，此时，再测试集上数据的异常得分情况如图 3.11 所示，在整个数据集上数据的异常得分情况如图 3.12 所示。由图 3.11 和图 3.12 可知，测试集上超出阈值的数据有 7 个，与我们加入噪声后的测试集数据的异常值个数相同。此外，分别选择阈值为 0.16 和 0.22 时，本节方法数据异常检测的结果如表 3.6 所示。阈值设置为 0.16 时，可以完全将训练集中加入的异常信息检测出来，但同时也会将一些正常数据判定为异常，造成误判，所以阈值选择为 0.22 时异常检测效果更好，我们将判定为异常的每组数据的时间索引和异常得分表示出来，如表 3.7 所示，明显可知，判定为异常数据的时间索引与我们加入噪声的异常数据的索引相同，表明本节自编码器模型有效的检测出了异常数据，证明了模型的有效性。此外，我们在整个数据集上加入了 15 个噪声数据，但是实验在训练集上被判定为异常的数据仅为 4 个，比加入噪声少了 4 个，这说明对于本节建立的模型，将加入噪声的训练数据输入模型进行训练，然后将测试集输入模型，模型仍能很好的检测出异常数据，说明自编码器能够很好的从噪声数据中提取出有效信息并进行数据重构。

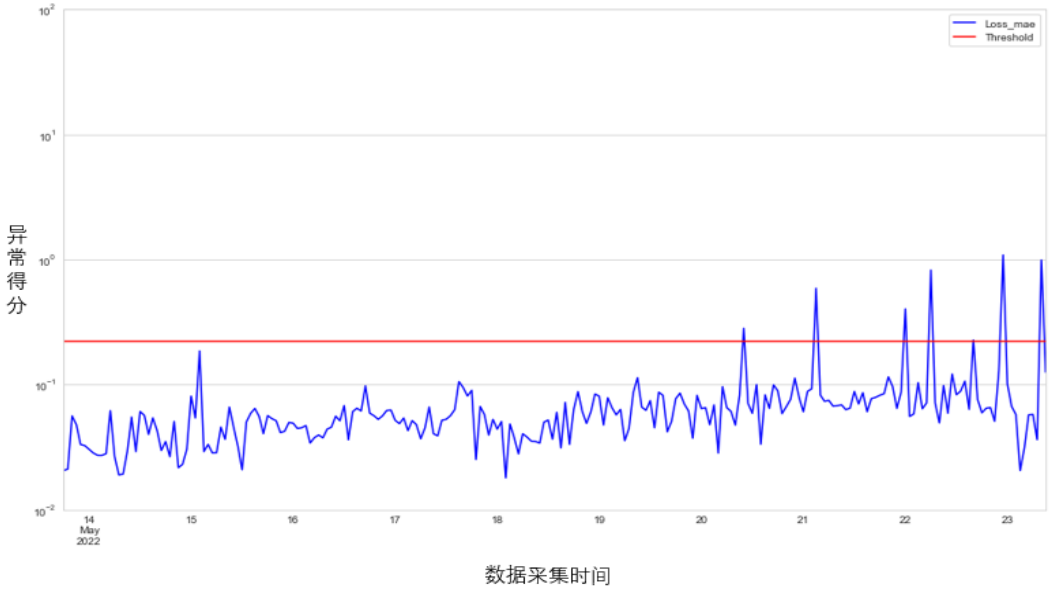


图3.11　在测试集上的异常得分（蓝色）与阈值（红色）

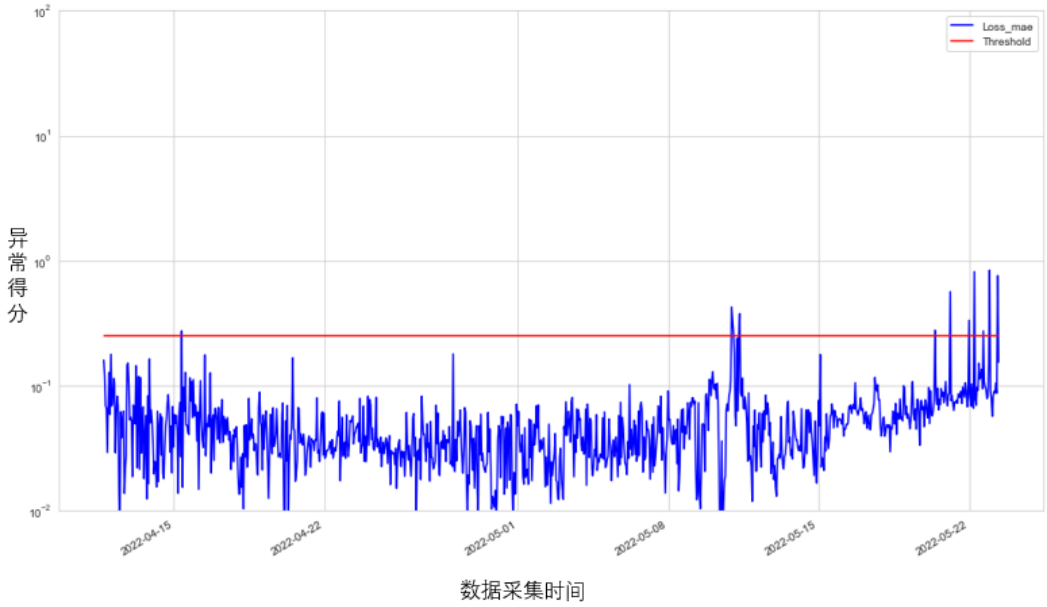


图3.12　数据异常得分曲线（蓝色）与阈值（红色）

表 3.6 不同阈值下自编码器数据异常检测的结果

数据 / 阈值	数据数量					
	训练集数据	测试集数据	训练集异常数据	测试集异常数据	训练集上检测到的异常数据	测试集上检测到的异常数据
0.16	769	231	8	7	8	10
0.22	769	231	8	7	4	7

表 3.7 整个数据集上被判定为异常值的数据时间索引及其得分

数据采集时间	Loss_mae	Threshold	Anomaly
2022-04-15 09:00:00	0.262239	0.22	True
2022-05-10 23:00:00	0.299522	0.22	True
2022-05-11 00:00:00	0.249695	0.22	True
2022-05-11 08:00:00	0.256324	0.22	True
2022-05-20 10:00:00	0.281588	0.22	True
2022-05-21 03:00:00	0.589284	0.22	True
2022-05-22 00:00:00	0.402810	0.22	True
2022-05-22 06:00:00	0.822883	0.22	True
2022-05-22 16:00:00	0.226880	0.22	True
2022-05-22 23:00:00	1.087134	0.22	True
2022-05-23 08:00:00	0.990720	0.22	True

3.4 本章小结

本章设计了两种海洋观测数据异常检测的方法。主要针对不同格式的海洋观测时间序列数据，对于单变量的海洋观测数据，基于统计学方法、局地异常检测和误差控制的海洋观测数据异常检测方法能够取得不错的检测效果，且不需要考虑数据的时间特性，也不需要考虑数据缺失对检测方法的影响。由于本章方法使用的数据量较小，本章没有使用诸如准确率、召回率等评估方法，仅比较了检测结果和初始标记的异常值之间的差量，本章基于 Grubbs 准则、3σ 准则、局地异常检测和误差控制的海洋观测异常检测方法能检测出大多数的异常数据，表明了所提方法的有效性。

对于多变量的海洋观测数据，建立基于自编码器的异常检测模型能够取得不错的效果。本章自编码器异常检测模型的建立过程基于两个串行连接的自编码器，通过重构两个自编码器的编码器和解码器，使得本章建立的异常检测模型结构相对简单，检测性能也更加优良。由于自编码器的编码器具有很好的提取有效数据特征的能力，所以使用带有少量异常信息的数据训练模型，并不影响模型的异常检测性能。由于海洋观测数据是一组时间序列数据，所以使用 LSTM 层提取数据的事件依赖性并构建自编码器，使得本章自编码器异常检测模型取得了较好的效果，验证了所提方法的有效性。

第 4 章 海洋观测数据异常值校正方法

海洋观测数据质量控制不仅仅是检测出数据中存在的异常和缺失，还要对异常和缺失的数据用更符合海洋要素时空特性的数据进行填补插值。在第 3 章中，几种海洋观测观测数据的异常和缺失已经被检测标记，如何将这些数据进行插值是本章要解决的问题。

目前，海洋观测数据质量控制相关的研究主要集中在数据异常值检测方面，数据插值方法也仅限于线性插值、二次插值、均值插值等，这些方法大多无法适用于复杂的海洋观测数据。机器学习算法的不断发展，为海洋观测数据的预测提供了更多的选择，而使用预测的数据进行插值，也成为一种新型的插值方法。

本章提出了以 LSTM 模型为核心的海洋观测数据异常值校正模型，并针对不同海洋要素数据分别建立了基于 STL 分解算法和 LSTM 神经网络的异常值校正模型与基于小波分解重构和 LSTM 神经网络的海洋观测数据预测模型。提出了以 ARIMA 模型为核心的海洋观测数据异常值校正模型，并针对不同海洋观测要素建立了基于 STL 分解算法的 SARIMA 海洋观测数据预测模型。本章将重点介绍模型的结构，对网络训练过程中使用的数据进行讲解，对预测结果进行分析，最后对模型进行相关评估。

4.1 基于 STL 分解算法和 LSTM 神经网络的海洋观测数据预测算法

本节提出了一种基于 STL 分解算法和 LSTM 神经网络的海洋观测数据预测方法，运用 STL 分解算法将海洋观测各要素数据分解为季节分量、趋势分量和余量，再将这些信号作为 LSTM 神经网络模型的输入，来训练模型预测未来的海洋观测数据，最后采用 RMSE、R^2 和 MAPE 等指标来评价模型预测性能的好坏。

4.1.1 基于 STL 分解算法和 LSTM 的预测模型设计

海洋观测设备采集的海洋要素序列均为时间序列，本节模型设计针对海洋观测数据的非平稳性，对海洋观测数据的每个要素数据都进行 STL 分解得到要素的趋势项、季节项和余项，分析出某些确定性因素影响下的序列分布规律，从而挖掘出其中潜在的关键信息。然后利用 LSTM 模型进行海洋观测数据的预测。

对于海洋观测数据，其各要素数据多是具有一定的趋势性和季节性，也具有较高的随机性，使用 STL 分解算法可以将其分解为趋势项、季节项和余项。趋势项，主要反映数据在时间范围的大致趋势，季节项，主要反映了数据在一段时间内的波动情况，与趋势项反映在不同的时间尺度上，从随机项看，海洋观测要素数据在整个时间序列中都具有较高的随机性。

STL 分解算法主要有四个主要的输入参数，参数的选择对于 STL 分解得到的数据是十分重要的，而这些参数主要是根据所给数据的特性以及实际经验进行选择。针对不同海洋观测要素数据，在选择了合适的参数之后，使用 STL 分解算法分解得到了各海洋观测要素的趋势项、季节项和余项，所得到的每一项的序列长度和原始序列相同，在将它们进行归一化处理之后，方便之后作为 LSTM 的输入供模型使用。

经过 STL 分解算法获得了合适的趋势分量、季节分量和余量，下一步工作是将这些分量作为 LSTM 预测模型的输入，实现对未来一段时间的海洋观测数据预测。这一过程的关键在于 LSTM 模型的搭建。LSTM 神经网络的搭建首先需要确定滑动窗口的大小，输入网络的滑动窗口大小主要由预测步长决定；其次是确定 LSTM 层输出空间维度，这取决于 LSTM 层输入空间维度；然后进行隐藏层的设计，即需要确定是否增加 LSTM 层和 Dense 层的层数，隐藏层的设计主要通过控制变量的方法不断地进行验证。LSTM 神经网络的搭建的每一步都需要不断地进行 MAE 或者 MSE 的计算来进行评估，最终确定 LSTM 神经网络的结构。结合 STL 分解算法和 LSTM 神经网络的海洋观测数据预测模型是以海洋观测数据集中的训练集为基础构建而成的，并最终用于测试集数据的预测。

综上，本节的基于 STL 分解算法和 LSTM 神经网络的海洋观测数据预测模型设计如图 4.1 所示：

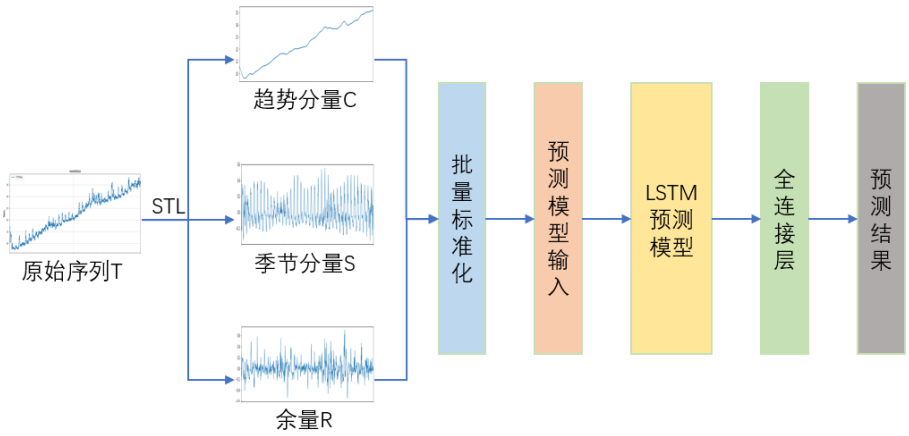


图4.1　基于 STL 分解算法和 LSTM 神经网络的海洋观测数据预测模型

本节预测模型的设计主要分为两个部分：第一部分主要是利用 STL 分解算法对原始海洋观测数据进行分解，得到合适的能够作为 LSTM 模型输入信号的趋势项、季节项和余项数据；第二部分是搭建 LSTM 预测模型，将得到的趋势项、季节项和余项一起作为模型的输入，通过 LSTM 预测模型得到未来一段时间的预测数据。

4.1.2 模型实现和参数选择

本节算法使用 python 实现，采用 python 集成库 statsmodels 下的 STL，采用的神经网络训练框架为 Python 下的 Keras，后端采用的是 TensorFlow。首先将原始海洋观测数据进行一系列的格式、时间、位置检查等得到能够被模型处理的海洋观测数据，然后将数据进行 STL 分解得到趋势分量、季节分量和余量，将经过 STL 分解的各分量进行归一化处理，加快模型的训练速度，防止过拟合，然后形成具有 3 个特征和对应一个结果的一组数据，其数据格式如表 4.1 所示，经过归一化处理的和 STL 分解的海洋观测数据各分量如图 4.2 所示，再以滑动窗口取连续 24h 的数据，其中第 24 个数据的结果为数据标签，构成新的数据集。将新构建的数据集前 80%作为训练集训练模型，后 20%作为验证集进行模型验证。

表 4.1 STL 分解得到的趋势项、余项和季节项数据

Timeindex	Trend	Season	resid
2020-11-01 00:01:00	1.000000	0.482017	0.543948
2020-11-01 00:01:00	0.995806	0.517545	0.538866
2020-11-01 00:01:00	0.991312	0.514184	0.541640
2020-11-01 00:01:00	0.987416	0.625958	0.510189
2020-11-01 00:01:00	0.983219	0.631843	0.512770
……	……	……	……
2020-11-01 00:01:00	0.719047	0.607745	0.769153
2020-11-01 00:01:00	0.719045	0.840450	0.617468
2020-11-01 00:01:00	0.719038	0.875754	0.582800
2020-11-01 00:01:00	0.719025	0.824425	0.546292
2020-11-01 00:01:00	0.719004	0.841919	0.523257

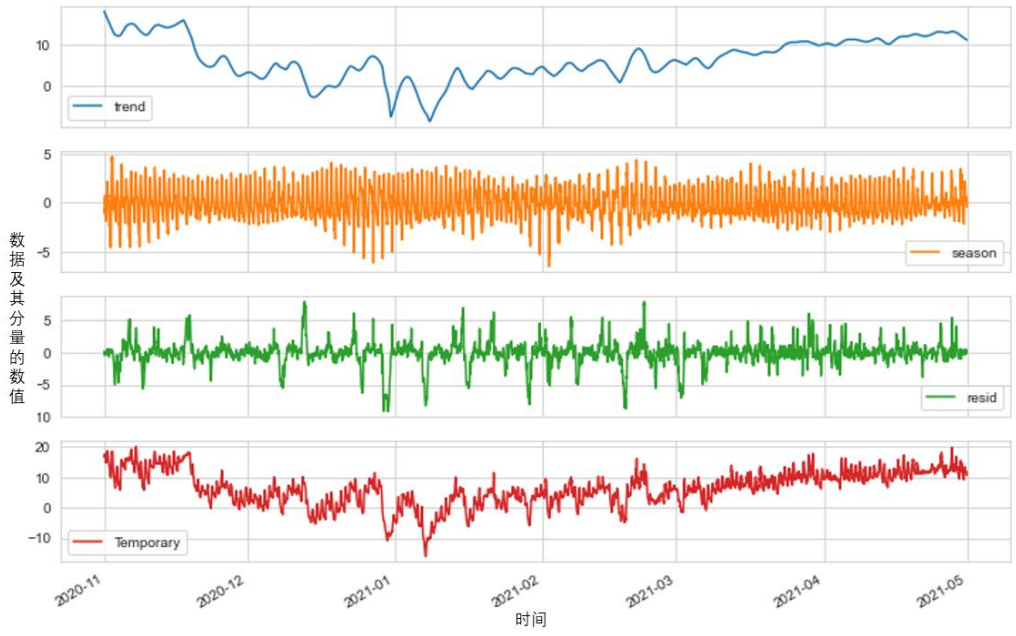


图4.2　STL 分解的海洋观测数据各分量

将海洋观测数据经过预处理，STL 分解和归一化处理之后，将它们作为 LSTM 的输入进行海洋观测数据的预测。这时，需要对 LSTM 模型的各个参数进行设计和不断优化。本节实验选用的实验数据较为平稳且相对较少，所以本节实验设计的 LSTM 预测模型结构相对简单，包括输入层，一个 LSTM 隐藏层和全链接层，LSTM 层主要有两个参数，即 LSTM 层的细胞单元数量 unit 用于指定输入的 shape=(TIME_STEPS, INPUT_SIZE)，本节模型最终确定 unit=16，shape=(24, 3)，其他参数设置采用默认值。此外，模型的编译需要设置优化器和损失函数，本节模型最终选用 Adam 优化器，确定 mae 作为模型的损失函数。

4.2 基于小波分解重构和 LSTM 神经网络的海洋观测数据预测算法

本节提出了一种基于小波分解单支重构和 LSTM 神经网络的海洋观测数据预测方法，运用小波分解将海洋观测各要素数据分解为概貌信号和细节信号并单支重构，再将重构的这些信号作为 LSTM 神经网络模型的输入，来训练模型预测未来的海洋观测数据，最后采用 RMSE、R^2 和 MAPE 等指标来评价模型预测性能的好坏。

4.2.1 基于小波分解重构和 LSTM 的预测模型设计

海洋观测数据的每一个要素都可以被视为一个一维信号，本节模型设计就是对海洋观测数据的每个要素数据都进行小波分解得到分解系数，然后对它们进行信号的单支重构，将高频信号和低频信号分离开，达到提取特征的目的，然后利用LSTM模型进行海洋观测数据的预测。小波分解和单支重构的过程如图4.3所示。

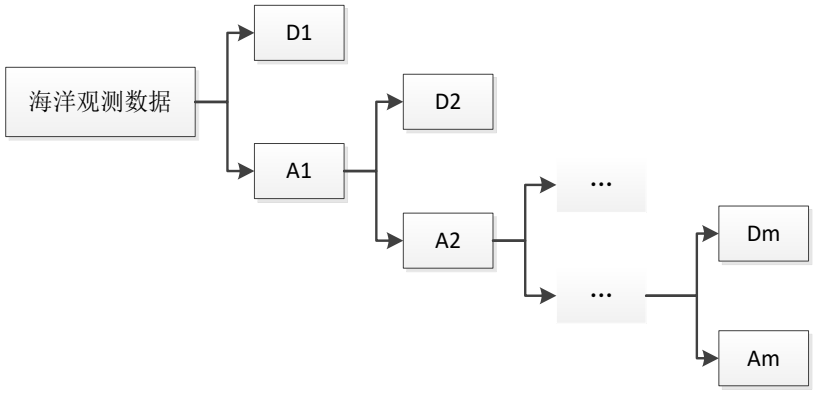


图4.3 小波分解过程

小波重构是小波分解的逆过程，但是，本节使用的小波单支重构是为了得到与原始信号相同结构的分量信号，所以利用全为零的信号结合 Mallat 算法进行单支重构得到需要的各个分量，过程（以 d8 为例）如图 4.4 所示。

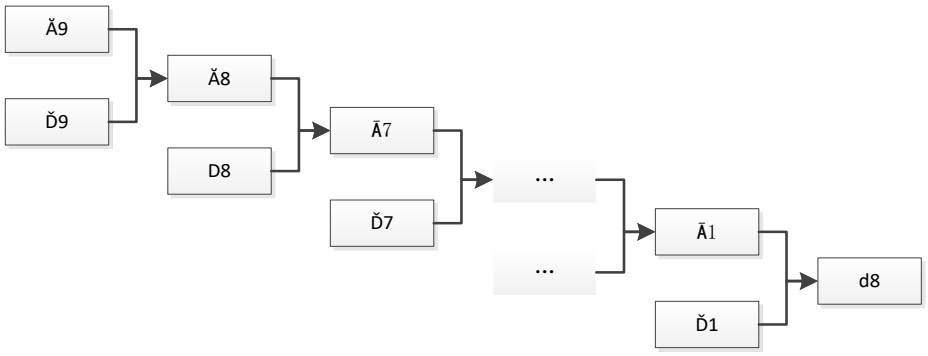


图4.4　小波单支重构过程（d8 为例）

其中，D8 为小波分解得到的 D8 细节信号，Ăi 是与 Ai 相同结构的全为零的一组信号，Ďi 是与 Di 相同结构的全为零的一组信号，Ā 是与 Ai 相同结构的新生成的一组信号。

在本节的海洋观测数据预测模型中，小波基函数的选择关系到海洋观测数据分解和重构的正确性。在信号处理过程中，选择不同的小波基函数会得到不同结果的分解重构信号，这些信号应用于后续模型中，效果会相差巨大。因此，在本节海洋观测数据预测中，小波基函数的选择是十分关键的一步，它需要在小波基特性的基础上结合实际经验进行选择，同样，对于小波分层数的确定，需要在小波基函数确定的基础上结合实际经验进行选择。

小波基的种类很多，常见的小波主要包括 coifN 小波、Hear 小波、symN 小波、dbN 小波、biorNr.Nd 小波等。本节主要 db3 小波基函数进行简单介绍。

Daubechies 小波作为十分常见的小波基是学者 Ingrid Daubechies 所提出的。db3 小波属于 dbN 小波的一种，具有较好的正则性。也正因为该小波具有较好的正则性，所以 db3 小波在信号重构中可以获得较为光滑的信号。此外，该小波并不具备对称性，通过图中的小波函数图像也可以看出这一点。db3 小波函数如图 4.5.所示。

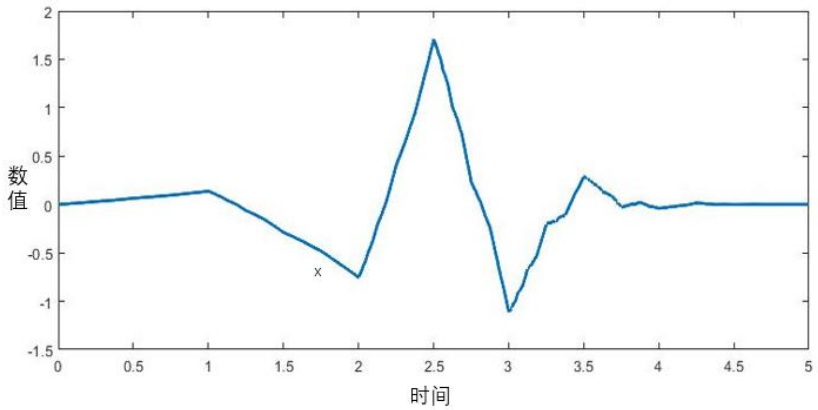


图4.5　db3 小波函数图

针对不同海洋观测要素数据，在选择了合适的小波基函数和小波分层数之后，小波分解得到各个分解的子信号并对它们单支重构，得到 n+1 个有效子序列，每个子序列长度和原始序列相同，方便之后作为 LSTM 的输入供模型使用，以温度数据为例，分解重构后的子序列如图 4.6 所示：

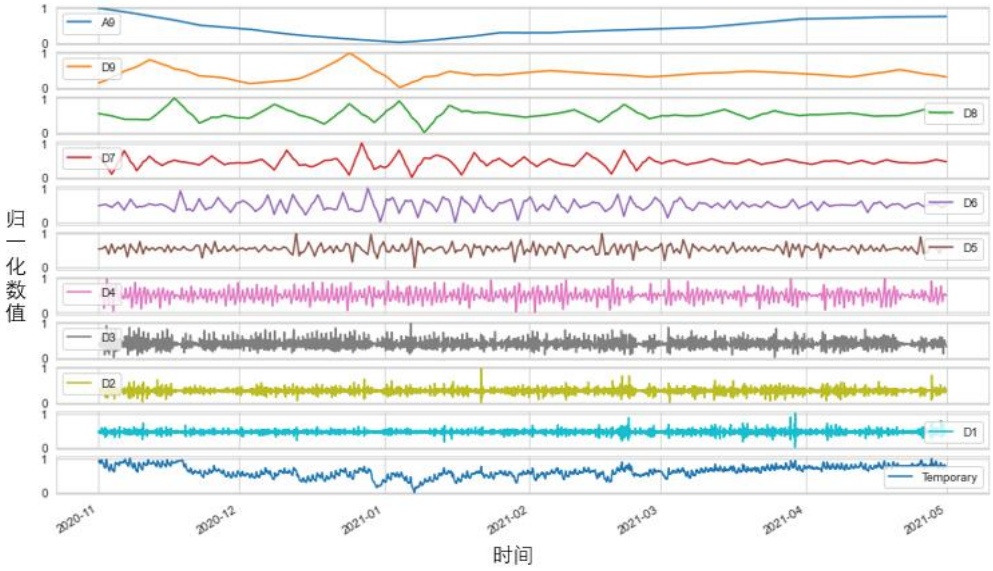


图4.6　小波分解重构子序列图

经过小波分解单支重构获得了合适的概貌信号和细节信号，下一步工作是将这些概貌信号和细节信号一起作为 LSTM 预测模型的输入，实现对未来一段时间的海洋观测数据预测。本节 LSTM 模型的搭建过程与 4.2.1 介绍的 LSTM 模型搭建过程相同。

综上，本节的基于小波分解重构和 LSTM 神经网络的海洋观测数据预测模型设计如图 4.7 所示：

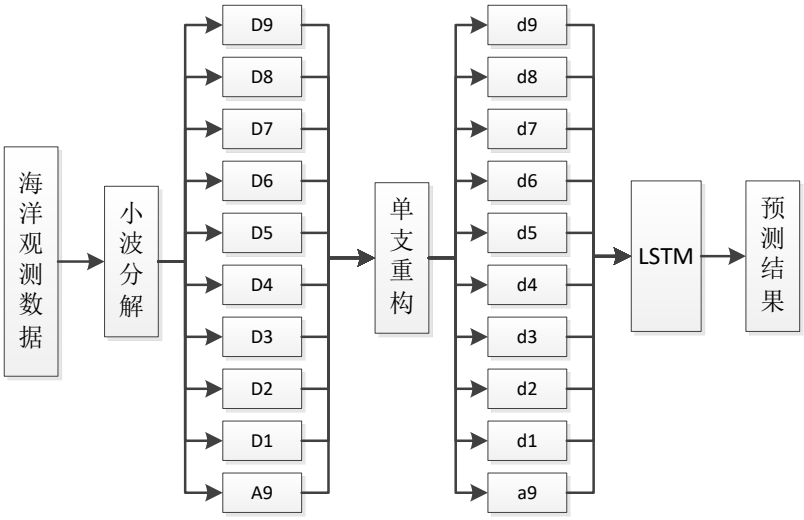


图4.7　基于小波分解重构和 LSTM 神经网络的海洋观测数据预测模型

本节模型设计主要分为两个部分：第一部分主要是利用小波分解与重构算法对原始海洋观测数据进行分解和单支重构，得到合适的能够作为 LSTM 模型输入信号的概貌信号和细节信号；第二部分是搭建 LSTM 预测模型，将得到的概貌信号和细节信号一起作为模型的输入，通过 LSTM 预测模型得到未来一段时间的预测数据。

4.2.2 模型实现和参数选择

本节算法使用 python 实现，采用的神经网络训练框架为 Python 下的 Keras，后端采用的是 TensorFlow。首先将原始海洋观测数据进行一系列的格式、时间、位置检查等得到能够被模型处理的海洋观测数据，然后将数据进行小波分解重构得到概貌信号和细节信号，本节选用 db3 小波基函数，将海洋观测数据分解为 9 层结构，经过小波分解重构后的各信号如图 4.8 所示，再将经过小波分解重构的概貌信号和细节信号进行归一化处理，加快模型的训练速度，防止过拟合，分解形成的各信号分量数据格式如图 4.9 所示，然后形成具有 10 个特征和对应一个结果的一组数据，再以滑动窗口取连续 24h 的数据，其中第 24 个数据的结果为数据标签，构成新的数据集。将新构建的数据集前 80%作为训练集训练模型，后 20%作为验证集进行模型验证。

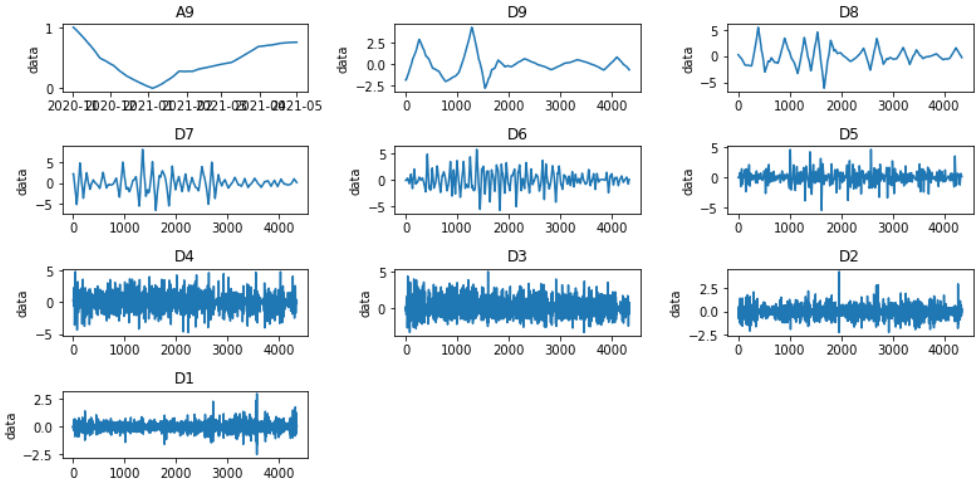


图4.8　海洋台站温度数据小波分解重构各信号图

timeindex	A9	D9	D8	D7	D6	D5	D4	D3	D2	D1
2020-11-01 00:01:00	1.000000	0.136621	0.549466	0.596229	0.490603	0.536701	0.522988	0.363098	0.333075	0.452847
2020-11-01 01:01:00	0.999883	0.134810	0.548487	0.598112	0.489974	0.536016	0.529337	0.386452	0.304554	0.479115
2020-11-01 02:01:00	0.999764	0.133006	0.547472	0.600021	0.489307	0.535353	0.536994	0.403653	0.308173	0.420161
2020-11-01 03:01:00	0.999646	0.131197	0.546452	0.601937	0.488629	0.534653	0.545178	0.427244	0.320162	0.468041
2020-11-01 04:01:00	0.999160	0.131790	0.545547	0.598371	0.489176	0.535009	0.537100	0.418264	0.359975	0.459179

图4.9　洋台站温度数据小波分解重构数据格式

海洋观测数据在预处理，小波分解重构和归一化处理后，得到用于输入到LSTM 预测模型的多维特征数据。这时，需要对 LSTM 的各个参数进行设计和不断优化。本节实验选用的实验数据同 4.1 节一样，所以本节实验设计的 LSTM 预测模型结构也同 4.1 节相似，即包括输入层，一个 LSTM 隐藏层和全链接层，LSTM 层的两个参数最终确定为细胞单元数量 unit=16，用于指定输入的 shape=(TIME_STEPS, INPUT_SIZE) =(24,10)。在训练模型之前，对于模型学习过程的优化器和损失函数的选择，本节也同样选用 adam 优化器和 mae 损失函数。

4.3 基于 ARIMA 的海洋观测数据预测模型设计

本节使用基于 STL 分解算法和 ARIMA 算法的数据预测方法来预测海洋观测数据，运用 STL 分解算法将海洋观测要素数据分解为季节分量、趋势分量和余量，再运用各分量数据分别建立 SARIMA 预测模型预测未来数据，海洋观测预测数据就是各分量预测数据的和，最后通过采用 RMSE、MAE 和 MAPE 等指标来评价本节模型预测性能的好坏。

4.3.1 基于 STL 分解算法和 SARIMA 的海洋观测数据预测模型设计

SARIMA 模型是 ARIMA 模型的扩展，具有处理季节趋势时间序列数据的特点。SARIMA(p, d, q)(P, D, Q，S）具有多个需要设置的参数，将其分为两个部分，非季节模型参数 p, d, q 与季节性模型参数 P, D, Q, S。其中，p 是趋势的自回归系数，d 是趋势差分阶数，q 代表趋势的移动平均阶数，P 为季节性自回归阶数，D 为季节性差分阶数，Q 代表季节性移动平均阶数，S 是指单个季节性周期的时间步长。

在海洋观测时间序列数据中，存在明显的周期性变化。STL 分解算法分解得到的趋势项反映数据的变化趋势，数据不会出现连续的巨大波动，使用 SARIMA 模型能较准确的预测未来数据，此外，使用 STL 分解算法将海洋观测数据分解为趋势项、余项和季节项并分别用于建立 SARIMA 模型，能更好的对每一分量数据对应的模型进行参数调整，提高每一部分的预测精度。本节使用 STL 分解算法的目的是提取海洋观测要素趋势分量、季节分量和余量分别建立 SARIMA 模型进行数据预测，最终海洋观测数据的预测结果就是趋势分量预测结果、季节分量预测结果和余量预测结果之和。

综上，本节的基于 STL 分解算法和 SARIMA 算法的海洋观测数据预测模型设计如图 4.10 所示：

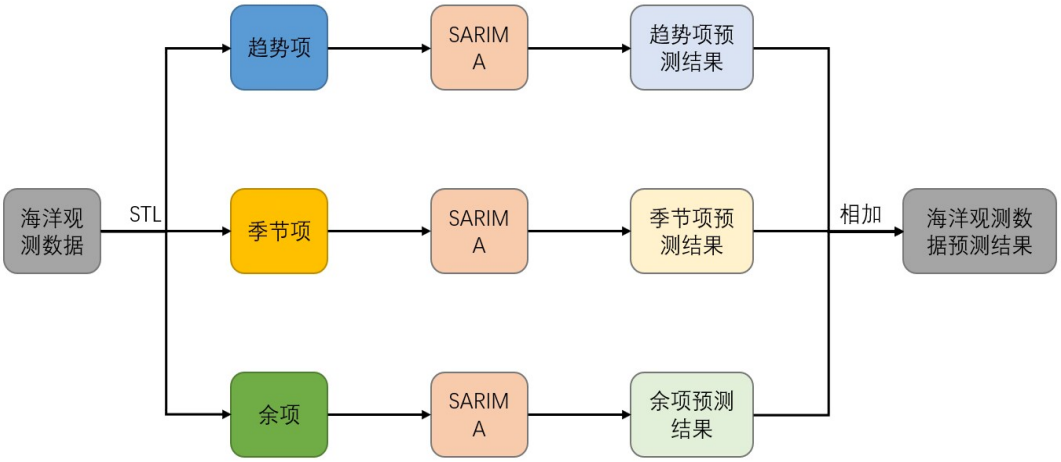


图4.10　基于 STL 分解算法和 SARIMA 的海洋观测数据预测模型图

本节模型设计主要分为两个部分：第一部分主要是使用 STL 分解算法对原始海洋观测数据进行分解，获得能够表示数据变化趋势、周期性以及冗余的趋势项、季节项和余项；第二部分主要是针对 STL 分解的各项分别建立 SARIMA 模型进行未来一段时间的数据预测，并将各项预测结果相加得到海洋观测数据的预测结果。

4.3.2 模型实现和参数选择

本节算法使用 python 实现，采用 python 集成库 statsmodels 下的 STL 以及库 pmdarima。首先将原始海洋观测数据进行一系列的格式、时间、位置检查等得到能够被模型处理的海洋观测数据，然后将数据进行 STL 分解得到趋势项、季节项和余项，如图 4.11、4.12、4.13 所示，将上述分解项进行数据的归一化处理之后，把各项数据的前 85%作为训练集训练模型，后 15%作为验证集进行模型验证。

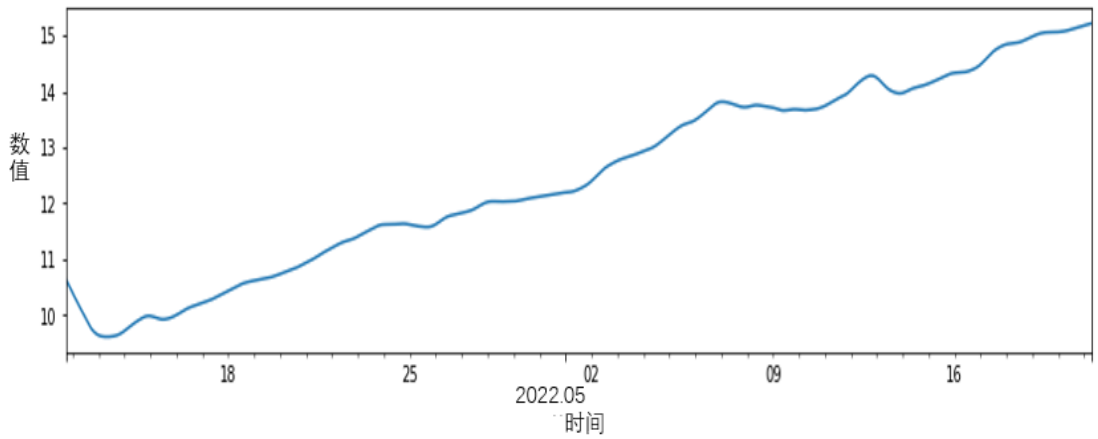


图4.11　浮标观测温度数据 STL 趋势项曲线图

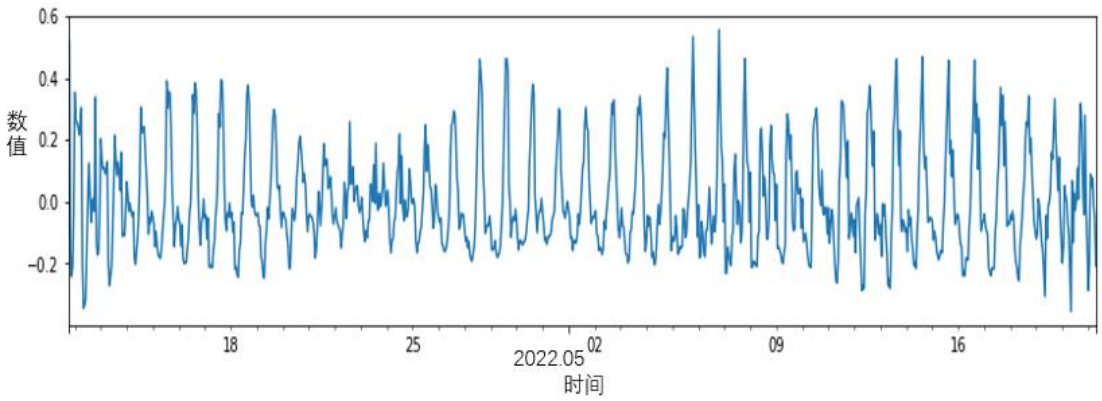


图4.12　浮标观测温度数据 STL 季节项曲线图

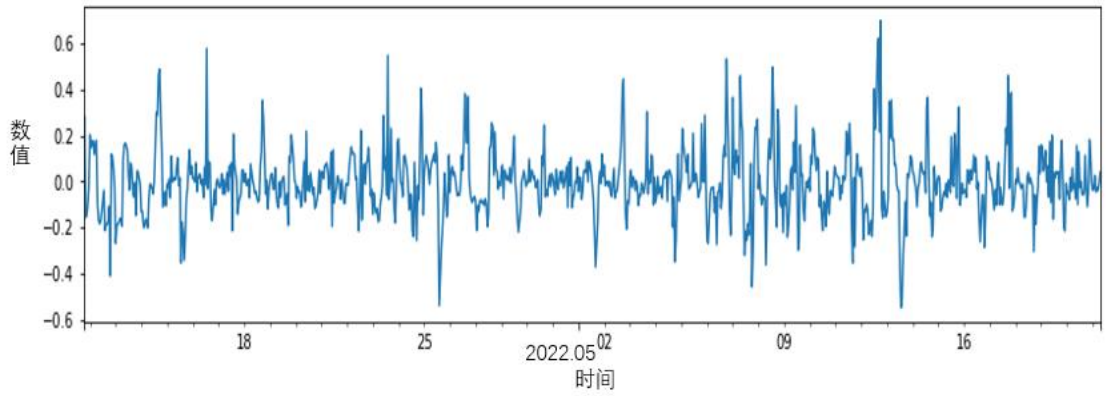


图4.13　浮标观测温度数据 STL 余项曲线图

在海洋观测数据经过 STL 分解之后，本节模型最关键的是 SARIMA 模型的建立和参数选择。模型的参数选择主要有两种方法，一种是采用图解法进行 SARIMA 模型的参数选择，但是这种方法计算并不容易，且耗时长，主观性大。另一种 SARIMA 模型的参数计算方法是使用网格搜索，网格搜索可以遍历不同的参数组合，可以根据合适的模型评价准则来选择合适的模型参数，所以本节选用网格搜索的方法来确定模型参数，使用 AIC 准则来评价选取模型使，AIC 函数达到最小的模型被认为是最优模型。本节模型设计使用 python 来实现，采用 python 库 pmdarima 中的 auto_arima()函数来确定各个分量 SARIMA 预测模型的各个参数，并采用 ADF 检验确认差分参数，最终，模型参数结果如图 4.14、图 4.15、图 4.16 所示。其中，趋势项的 SARIMA 模型参数为 ARIMA(2,0,2)(0,1,0)[12], 季节项的 SARIMA 模型参数为 ARIMA(1,0,0)(2,1,2)[12]，余项的 SARIMA 模型参数为 ARIMA(1,0,0)(2,1,0)[12]。此外，为了验证本节模型的有效性，本节还设计了不经过 STL 分解进行 SARIMA 的海洋观测数据预测，其模型参数结果如图 4.17 所示，其模型参数为趋势项的 SARIMA 模型参数结果为 ARIMA(2,0,0)(2,1,0)[12]。

```
Best model:  ARIMA(2,0,2)(0,1,0)[12] intercept
Total fit time: 578.833 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  807
Model:             SARIMAX(2, 0, 2)x(0, 1, [], 12)   Log Likelihood                4501.556
Date:                             Tue, 28 Feb 2023   AIC                           -8991.111
Time:                                     20:59:00   BIC                           -8963.041
Sample:                                          0   HQIC                          -8980.325
                                               - 807
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0003   8.12e-05      3.858      0.000       0.000       0.000
ar.L1          1.9599      0.002   1089.982      0.000       1.956       1.963
ar.L2         -0.9660      0.002   -548.678      0.000      -0.969      -0.963
ma.L1          0.5003      0.009     56.181      0.000       0.483       0.518
ma.L2          0.5308      0.010     55.802      0.000       0.512       0.549
sigma2      6.985e-07   7.43e-09     94.054      0.000    6.84e-07    7.13e-07
===================================================================================
Ljung-Box (L1) (Q):                   27.72   Jarque-Bera (JB):            128151.37
Prob(Q):                              0.00   Prob(JB):                         0.00
Heteroskedasticity (H):               0.27   Skew:                            -1.67
Prob(H) (two-sided):                  0.00   Kurtosis:                        65.11
===================================================================================
```

图4.14　STL-SARIMA 趋势项参数结果图

```
Best model:  ARIMA(1,0,0)(2,1,2)[12] intercept
Total fit time: 522.640 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                  y   No. Observations:                  807
Model:             SARIMAX(1, 0, 0)x(2, 1, [1, 2], 12)   Log Likelihood                1553.076
Date:                             Tue, 28 Feb 2023   AIC                           -3092.152
Time:                                     21:59:04   BIC                           -3059.403
Sample:                                          0   HQIC                          -3079.568
                                               - 807
Covariance Type:                               opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -0.0003      0.002     -0.161      0.872      -0.004       0.004
ar.L1          0.6099      0.025     24.890      0.000       0.562       0.658
ar.S.L12      -1.0704      0.046    -23.439      0.000      -1.160      -0.981
ar.S.L24      -0.0980      0.046     -2.146      0.032      -0.187      -0.009
ma.S.L12       0.0622      0.040      1.555      0.120      -0.016       0.141
ma.S.L24       0.7561      0.024     31.656      0.000       0.709       0.803
sigma2         0.0011   3.87e-05     27.799      0.000       0.001       0.001
===================================================================================
Ljung-Box (L1) (Q):                   0.02   Jarque-Bera (JB):               244.94
Prob(Q):                              0.88   Prob(JB):                         0.00
Heteroskedasticity (H):               1.14   Skew:                            -0.21
Prob(H) (two-sided):                  0.28   Kurtosis:                         5.69
===================================================================================
```

图4.15　STL-SARIMA 季节项参数结果图

```
Best model:  ARIMA(1,0,0)(2,1,0)[12] intercept
Total fit time: 452.873 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  807
Model:             SARIMAX(1, 0, 0)x(2, 1, 0, 12)   Log Likelihood                 430.776
Date:                            Tue, 28 Feb 2023   AIC                           -851.553
Time:                                    21:19:11   BIC                           -828.161
Sample:                                         0   HQIC                          -842.564
                                              - 807
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0002      0.005      0.033      0.974      -0.010       0.010
ar.L1          0.5418      0.023     23.461      0.000       0.497       0.587
ar.S.L12      -0.4383      0.026    -16.847      0.000      -0.489      -0.387
ar.S.L24      -0.5652      0.027    -20.878      0.000      -0.618      -0.512
sigma2         0.0195      0.001     25.251      0.000       0.018       0.021
===================================================================================
Ljung-Box (L1) (Q):                   0.33   Jarque-Bera (JB):               116.18
Prob(Q):                              0.57   Prob(JB):                         0.00
Heteroskedasticity (H):               1.73   Skew:                             0.43
Prob(H) (two-sided):                  0.00   Kurtosis:                         4.67
===================================================================================
```

图4.16　STL-SARIMA 余项参数结果图

```
Best model:  ARIMA(2,0,0)(2,1,0)[12] intercept
Total fit time: 405.021 seconds
                                SARIMAX Results
==========================================================================================
Dep. Variable:                                   y   No. Observations:                  807
Model:             SARIMAX(2, 0, 0)x(2, 1, 0, 12)   Log Likelihood                 160.574
Date:                            Tue, 28 Feb 2023   AIC                           -309.148
Time:                                    15:54:57   BIC                           -281.078
Sample:                                         0   HQIC                          -298.362
                                              - 807
Covariance Type:                              opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0220      0.008      2.642      0.008       0.006       0.038
ar.L1          0.7009      0.027     26.296      0.000       0.649       0.753
ar.L2          0.1020      0.029      3.573      0.000       0.046       0.158
ar.S.L12      -0.6873      0.027    -25.189      0.000      -0.741      -0.634
ar.S.L24      -0.2640      0.034     -7.868      0.000      -0.330      -0.198
sigma2         0.0388      0.001     32.465      0.000       0.036       0.041
===================================================================================
Ljung-Box (L1) (Q):                   0.01   Jarque-Bera (JB):               821.51
Prob(Q):                              0.92   Prob(JB):                         0.00
Heteroskedasticity (H):               1.62   Skew:                             0.13
Prob(H) (two-sided):                  0.00   Kurtosis:                         7.97
===================================================================================
```

图4.17　浮标观测温度数据 SARIMA 预测模型参数结果图

在参数选择之后，对模型的诊断是非常重要的一步，所以需要通过分析 SARIMA 模型的模型诊断图来确保模型所作的假设没有被违反。SARIMA 模型的残差应为一个均值为 0，方差为常数的正态分布的白噪声，此时，表示模型具有良好的效果，可以认为模型能够充分提取序列的信息，否则，需要对模型参数进行调整。本节各个模型的模型诊断图如图 4.18、图 4.19、图 4.20、图 4.21 所示。

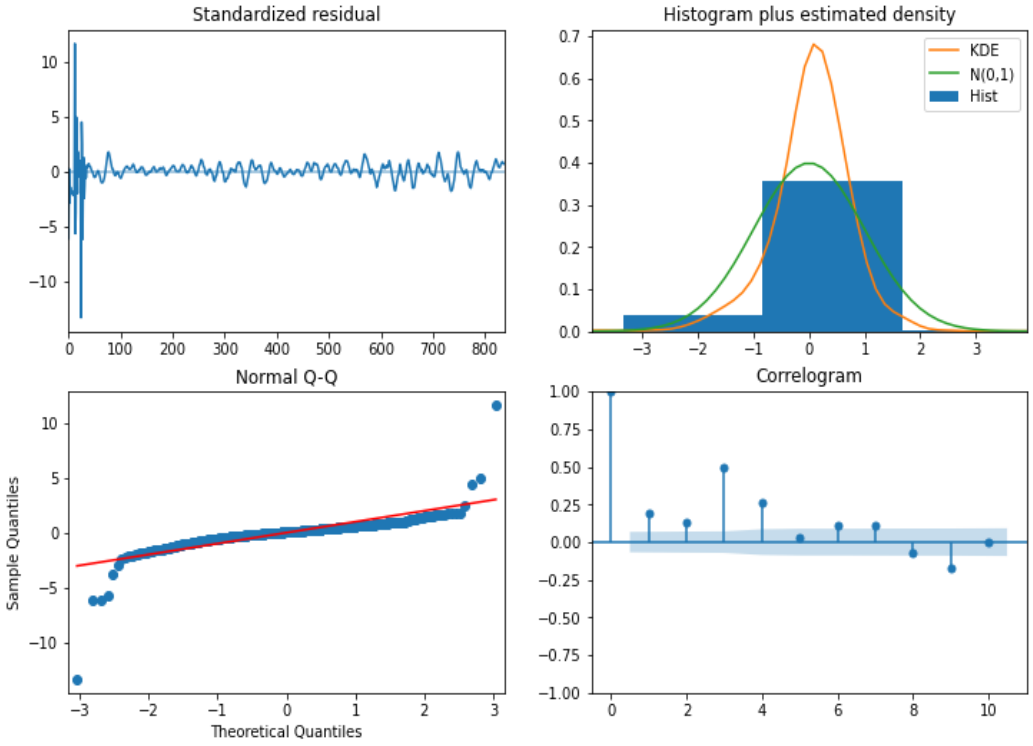


图4.18　STL-SARIMA 趋势项模型诊断图

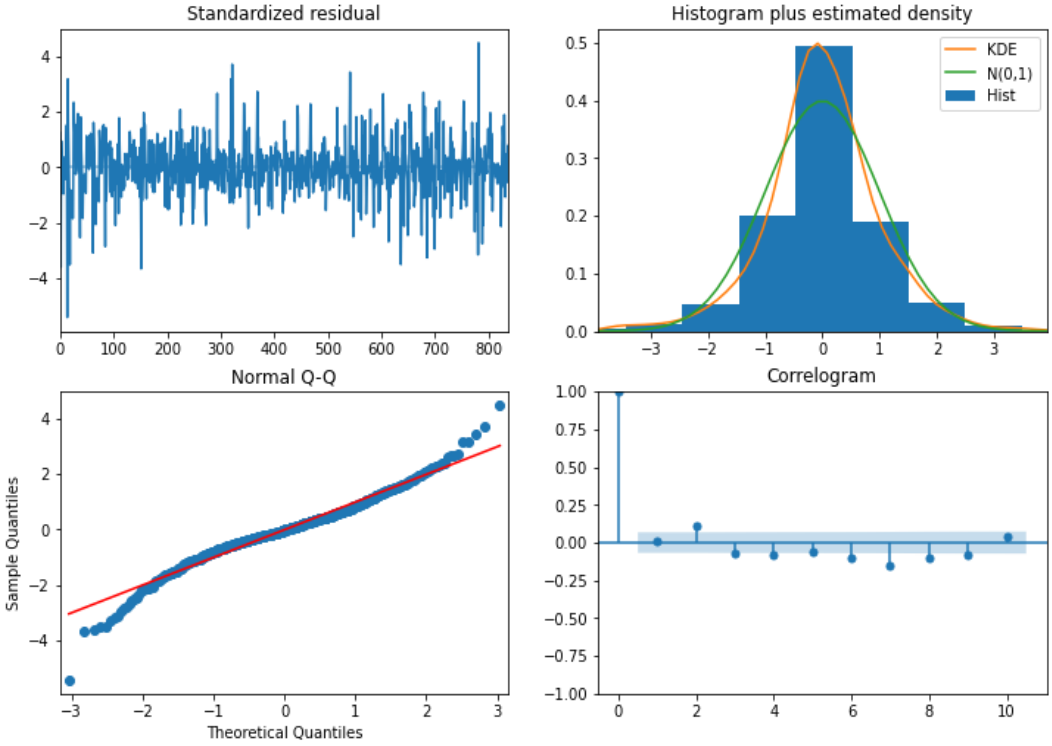


图4.19　STL-SARIMA 季节项模型诊断图

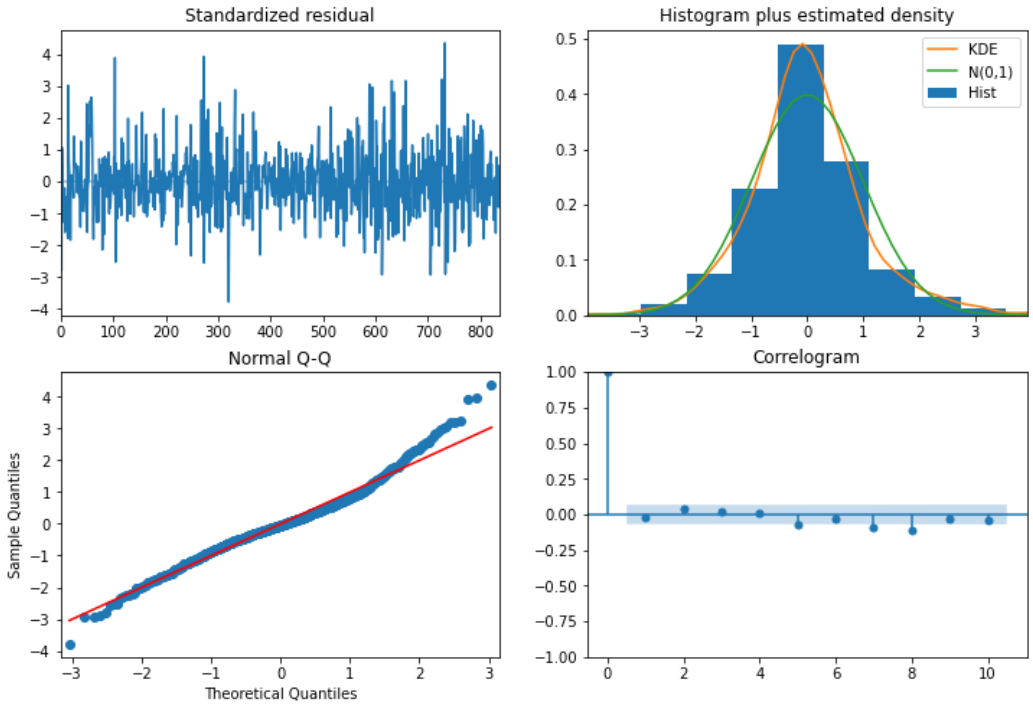


图4.20　STL-SARIMA 余项模型诊断图

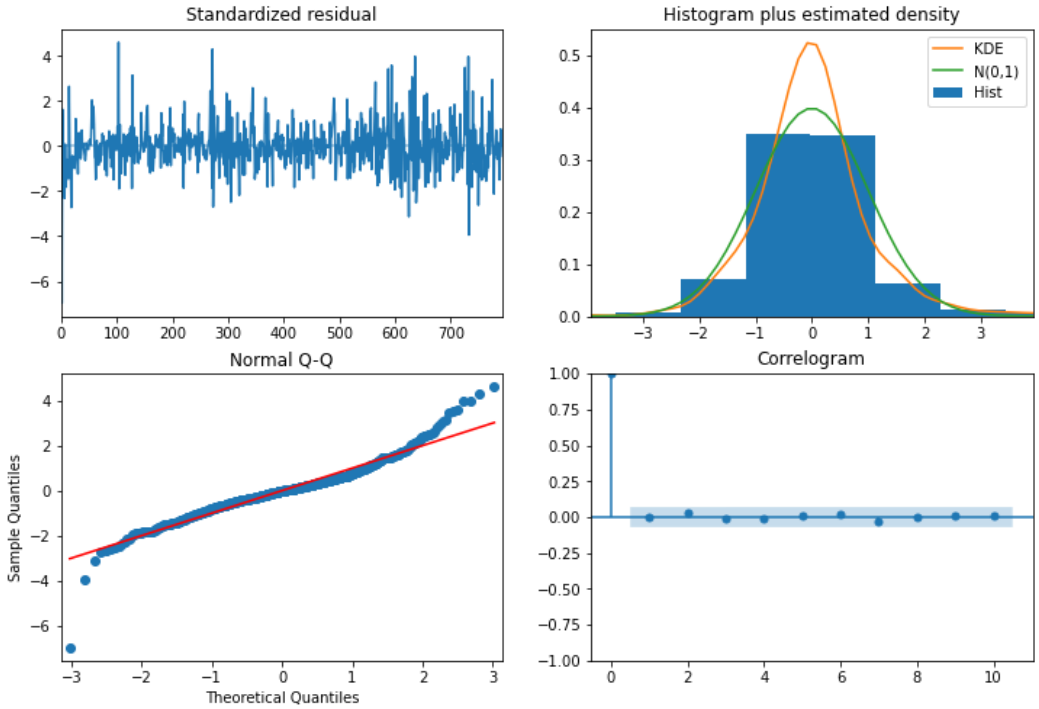


图4.21　浮标观测温度数据 SARIMA 模型诊断图

根据 SARIMA 模型诊断图可以清晰的看出，残差的时序图基本稳定，残差随着时间的变化波动不大，右上方的正态分布图也很好地说明了残差是服从正态分布的，左下方的正太 Q—Q 图也说明了残差服从正态分布。右下方的残差自相关

图显示残差不存在自相关，说明残差序列是白噪声序列。综上，可以得出结论，残差近似为均值为 0，方差为常数的正态分布的白噪声，模型参数确定为上述参数能够取得良好的预测效果。

4.4 模型测试及结果分析

4.4.1 模型评价指标

对于海洋观测数据的预测结果，本文采用平均绝对误差（Mean Absolute Error: MAE）、均方根误差（root mean square error: RMSE）、平均绝对百分比误差（mean absolute percentage error: MAPE）和 R^2 决定系数作为模型的评价指标，对比不同预测模型分别预测海洋观测数据预测性能的好坏。MAE、MAPE、RMSE 和 R^2 决定系数的公式如下表示：

$$\mathrm{RMSE}=\sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i-\tilde{y}_i)^2} \tag{4.1}$$

$$\mathrm{MAE}=\frac{1}{m}\sum_{i=1}^{m}|\tilde{y}_i-y_i| \tag{4.2}$$

$$\mathrm{MAPE}=\frac{100\%}{m}\sum_{i=1}^{m}\frac{|y_i-\tilde{y}_i|}{y_i} \tag{4.3}$$

$$R^2=1-\frac{\sum_{i=0}^{m}(y_i-\tilde{y}_i)^2}{\sum_{i=0}^{m}(y_i-\bar{y})^2} \tag{4.4}$$

其中，y_i 代表真实值，$\tilde{y}_i$ 代表预测值，$\bar{y}$ 代表均值，m 代表预测数量。对于 MAE、RMSE 和 MAPE，其数值越小表示模型预测效果越好，对于 R^2，其数值越接近于 1 表示模型预测效果越好。

4.4.2 实验数据集介绍

基于 LSTM 的预测模型使用的实验数据是国家海洋科学数据中心完全共享的中国台站观测数据中的 XiaoMaiDao 海洋台站温度数据，实验数据选择的时间跨度为 2020-11-01 00:01:00 至 2021-4-30 23:01:00，时间分辨率为 1h，共 4344 个温度要素观测数据。XiaoMaiDao 海洋台站温度原始序列数据及数据格式如表 4.2 和图 4.22 所示：

表 4.2 海洋观测温度原始数据格式

数据采集时间	温度（℃）
2020-11-01 00:01:00	16.9
2020-11-01 01:01:00	17.1
2020-11-01 02:01:00	17
2020-11-01 03:01:00	17.6
2020-11-01 04:01:00	17.6
2020-11-01 05:01:00	17.7
2020-11-01 06:01:00	17
2020-11-01 07:01:00	16.7
2020-11-01 08:01:00	15.8
2020-11-01 09:01:00	14.8
2020-11-01 10:01:00	15.1

图4.22　海洋台站温度原始序列

基于 ARIMA 预测模型的实验所用数据为海洋观测浮标观测到的温度数据，使用的温度数据为从 2022-04-11 18:00:00 至 2022-05-21 07:00:00，时间分辨率为 1h 的 950 个数据。浮标观测温度数据及数据格式如表 4.3 和图 4.23 所示：

表 4.3 浮标观测的温度数据 （部分）

数据采集时间	温度（℃）
2022/4/11 18:00	10.48
2022/4/11 19:00	11.40
2022/4/11 20:00	10.24
2022/4/11 21:00	10.12
2022/4/11 22:00	10.16
……	……
2022/5/21 03:00	15.20
2022/5/21 04:00	15.25
2022/5/21 05:00	15.16
2022/5/21 06:00	15.15
2022/5/21 07:00	15.05

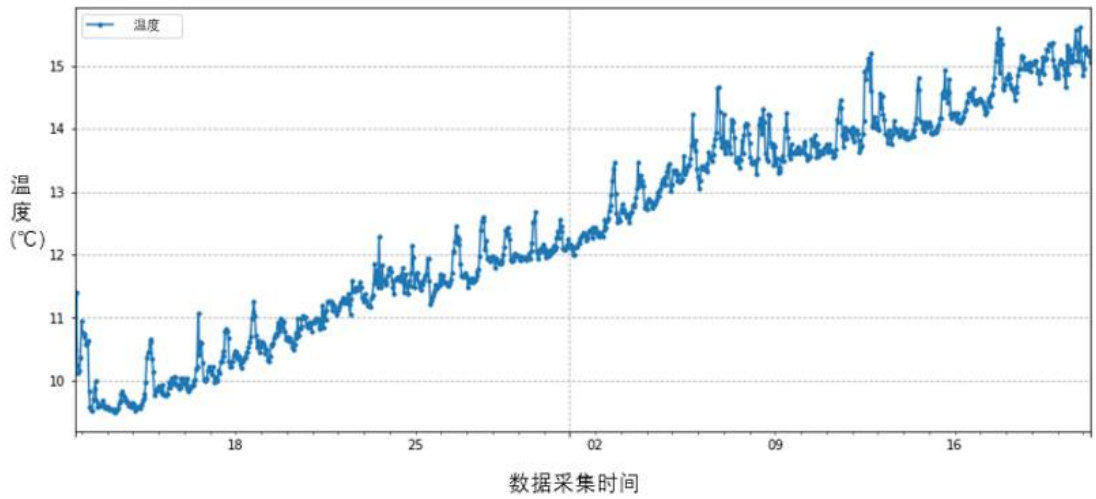


图4.23　浮标观测温度数据

4.4.3 实验结果分析

本章设计了只使用 LSTM 算法的海洋观测数据预测模型，在本章使用的温度数据集上的预测结果如图 4.24 所示。本章设计的基于 STL 分解算法结合 LSTM 算法的海洋观测数据预测模型在验证集上的结果如图 4.25 所示。通过小波分解重构结合 LSTM 预测模型，得到在验证集上的预测结果，如图 4.26 所示。为了验证模型的有效性，本章分别对比只使用 LSTM 算法构建的海洋观测数据预测模型和结合 STL 分解算法的 LSTM 预测模型以及结合小波分解重构的 LSTM 预测模型，通过比较其 RMSE 指标、MAPE 指标和 R^2 值来评价模型的性能，表 4.4 是三种预测模型的 RMSE 指标、MAPE 指标和 R^2 值。

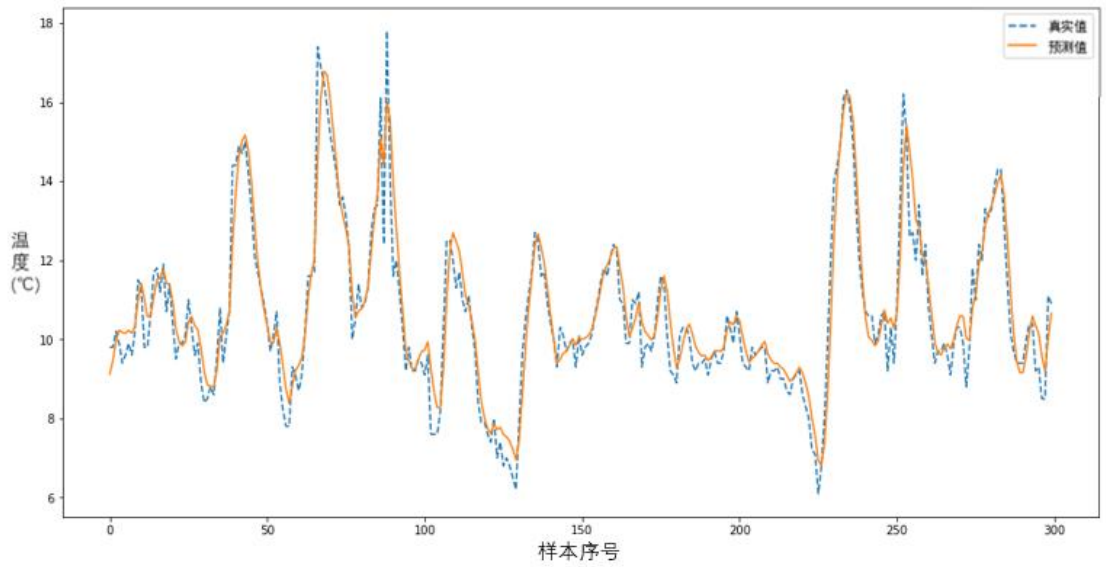


图4.24　LSTM 验证集上的预测结果

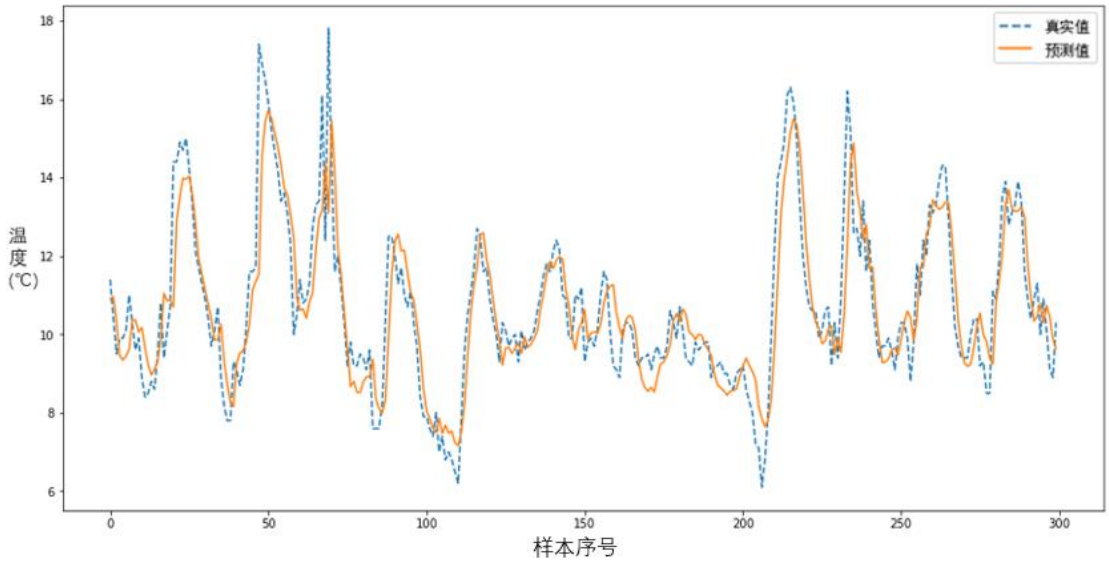


图4.25　STL-LSTM 验证集上的预测效果

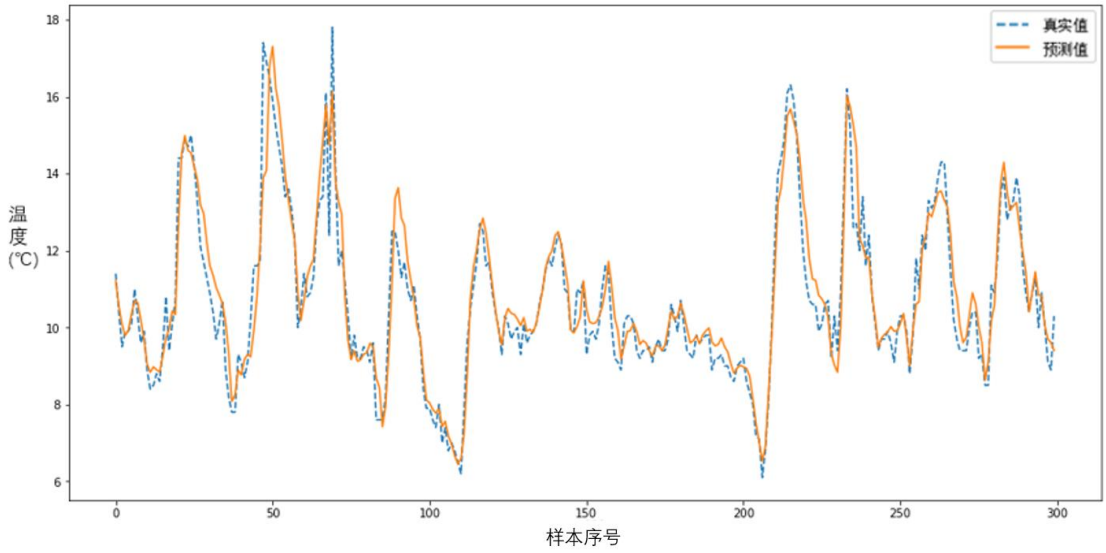


图4.26　Wavelet-LSTM 验证集上的预测效果

表 4.4 模型评价指标（RMSE、MAPE、R2）

预测模型	评价指标		
	RMSE	MAPE（%）	R^2
LSTM	0.0295	2.8660	0.7488
STL-LSTM	0.0239	2.1850	0.8373
Wavelet-LSTM	0.0225	1.5239	0.9207

由表 4.4 可知，在 XMD 海洋台站温度的数据预测上，Wavelet-LSTM 预测模型的 RMSE 指标和 MAPE 指标分别为 0.0225 和 1.5239，相比 LSTM 模型和

STL-LSTM 模型数值更小，这表示数据的波动性会对预测结果产生影响，而小波变换可以更加有效的处理这种特性，通过小波分解与重构时间序列数据的特征趋势，再将其作为 LSTM 模型的输入进行预测可以提高预测精度；此外，STL-LSTM 预测模型的 RMSE 指标和 MAPE 指标分别为 0.0239 和 2.1850，相比 LSTM 模型数值更小，说明 STL-LSTM 比 LSTM 模型预测效果更好，对比三种模型的 R^2 值，Wavelet-LSTM 模型的 R^2 最接近于 1，STL-LSTM 模型相较 LSTM 模型更接近于 1，说明 Wavelet-LSTM 模型是那个更好的预测模型，STL-LSTM 模型比 LSTM 模型更好，其主要原因是 Wavelet-LSTM 模型和 STL-LSTM 模型充分利用了 LSTM 算法的优点，使用小波分解重构获得了更加细节的信号，获得了更多的特征，使得误差减小，使用 STL 分解也得到了相似的效果。综上所述，无论在 RMSE、MAPE 指标下，还是在 R^2 数值比较中，Wavelet-LSTM 模型都优于其他模型，STL-LSTM 模型也取得了较好的效果，验证了本文方法的有效性。另外，在使用其他不同海洋观测数据的验证中，基于 STL 分解算法和小波分解重构的两种模型表现出不同的优劣，所以在不同的海洋观测数据的预测上，应选用多种算法进行比较，选择更好的模型进行数据预测。

对于第 3 章异常检测使用的 W70X 波浪数据中的有效波高数据，本节使用 STL-LSTM 预测模型对其测试集数据进行预测，并使用上述模型评价指标对模型进行评估，其中模型 R^2 值为 0.8518，接近且大于模型在温度数据集上的 R^2 值，RMSE 指标和 MAPE 指标分别为 0.0186 和 1.5836，均接近且小于模型在温度数据集上的评估数值，说明 STL-LSTM 预测模型在 W70X 有效波高数据上的预测具有良好的效果。其预测结果和原始数据曲线如图 4.27 所示：

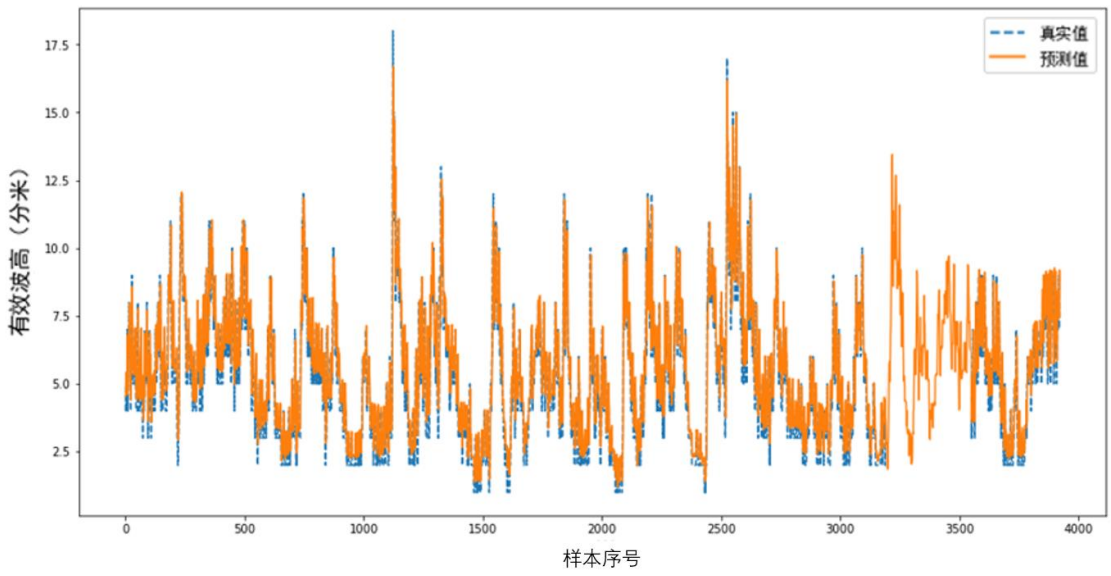


图4.27　W70X 浮标有效波高数据使用 STL-LSTM 预测模型的预测效果

注：图中真实值曲线具有部分缺失数据，可使用模型的预测值对缺失数据进行填补插值。

对于基于 ARIMA 的海洋观测数据预测，本章设计的基于 STL 分解算法的 SARIMA 预测模型在其各分量上的预测结果如图 4.28、图 4.29、图 4.30 所示，在浮标温度数据上获得的预测结果如图 4.31 所示。此外，本章设计了只使用 SARIMA 算法的预测模型作为基于 STL 分解算法的 SARIMA 预测模型的对比，其在浮标温度数据上的预测结果如图 4.32 所示，最后，使用 MAE，MAPE，RMSE 三种指标对上述两种模型进行评估，其评估结果如表 4.5 所示：

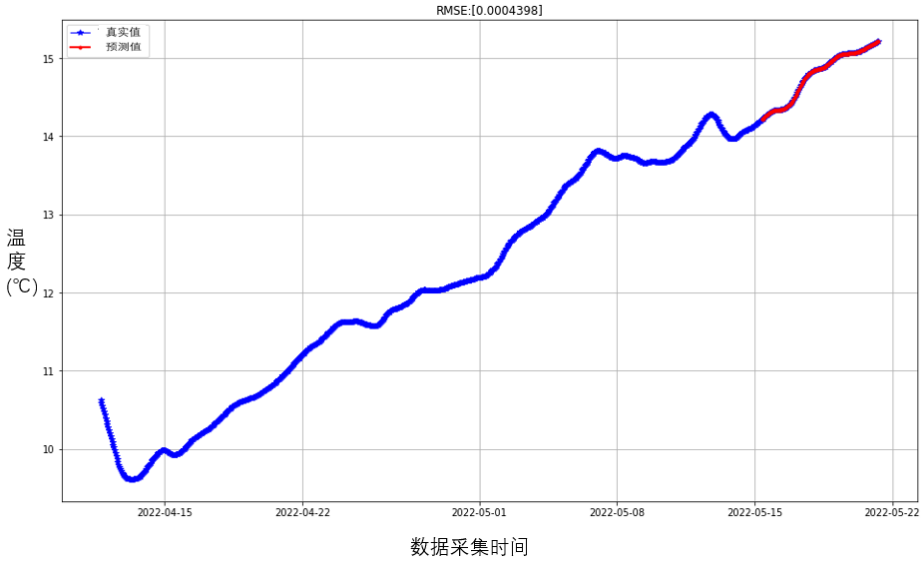


图4.28　趋势项 SARIMA 模型预测结果

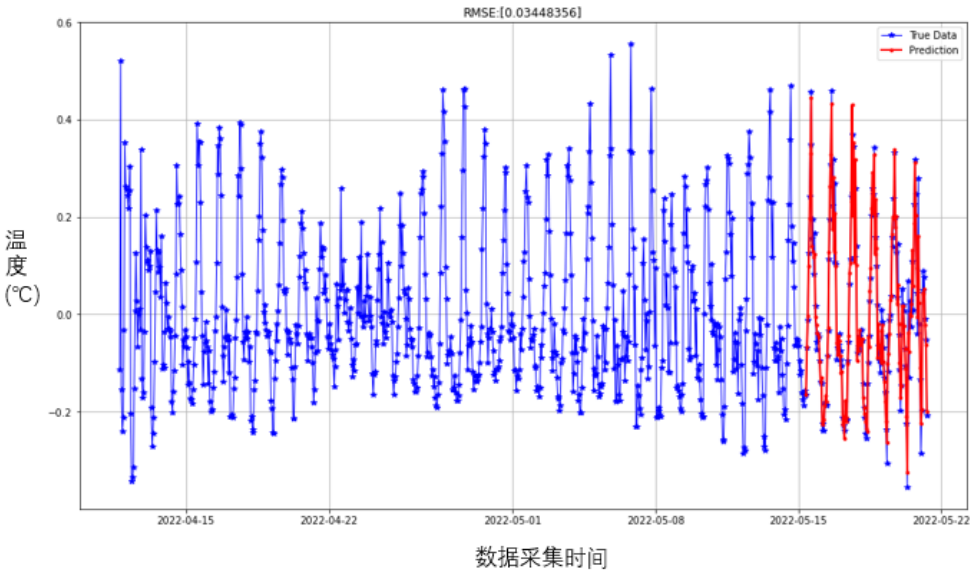


图4.29　季节项 SARIMA 模型预测结果

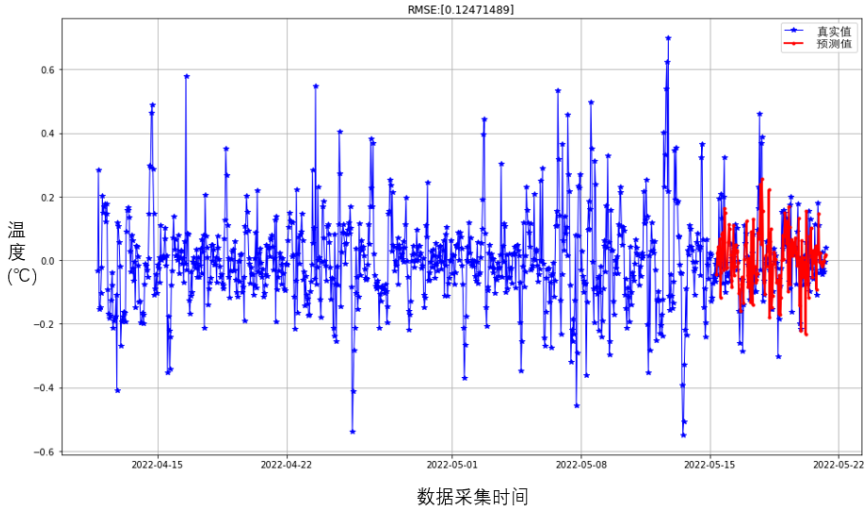


图4.30　余项 SARIMA 模型预测结果

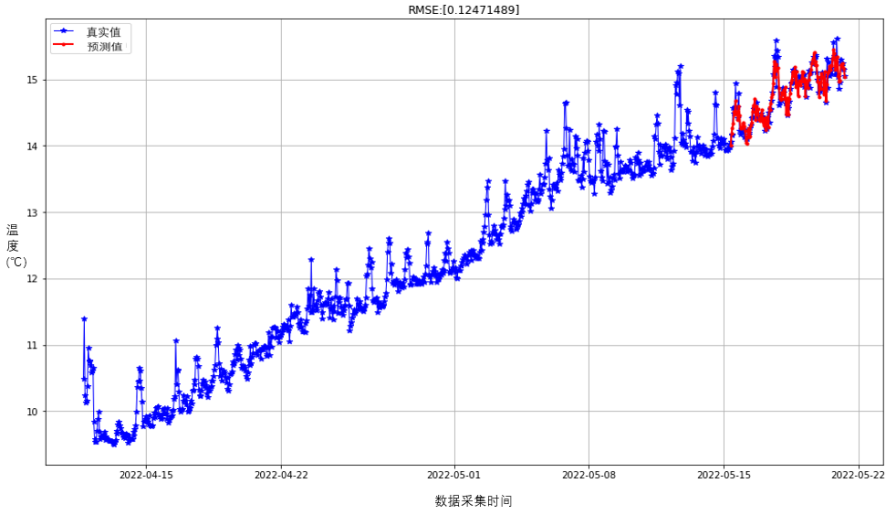


图4.31　海洋观测数据 STL-SARIMA 模型预测结果图

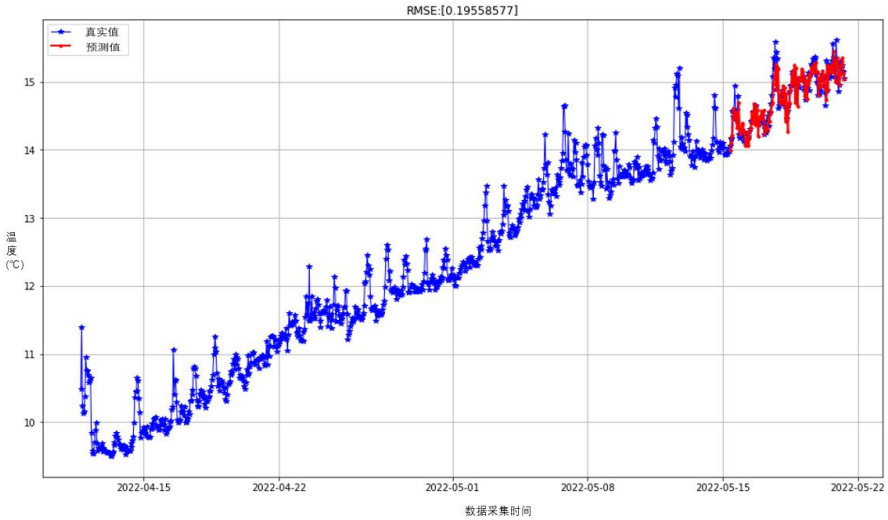


图4.32　SARIMA 预测结果图

表 4.5 SARIMA 与 STL-SARIMA 预测模型评价指标

预测模型	评价指标		
	MAPE	MAE	RMSE
SARIMA	0.09982	0.14825	0.19559
STL-SARIMA	0.06659	0.09865	0.12471

观察经过 STL 分解的趋势项的预测效果和分解得到的趋势项、季节项、余项预测和结果，发现在趋势项上模型预测精度很高，说明 STL-SARIMA 的预测误差主要集中在余项和季节项上，所以在浮标观测温度数据上，使用 STL-SARIMA 模型来预测未来温度的变化趋势具有良好的效果。

由表 4.5 可知，在浮标观测温度数据上的预测，STL-SARIMA 模型的 MAPE、MAE、RMSE 的数值分别为 0.06659，0.09865，0.12471，都小于 SARIMA 模型的各项指标，这表明 STL-SARIMA 模型在浮标观测温度数据上的取得的预测效果更好。

通过使用相同海域不同浮标检测到的温度数据验证本章基于 STL 分解算法和 SARIMA 的预测模型的有效性。实验选用与浮标观测数据检测时间相同、采样频率相同的一组海洋观测数据，并将数据输入到本节设计的预测模型，得到如图 4.33 所示的预测结果。然后通过上述评价指标进行评估，其中 MAPE 的值为 0.00711，MAE 的值为 0.01015，RMSE 的值为 0.01299，其评估结果均与模型在测试集上的评估结果相近，说明本节模型在同海域同时间段内的预测是有效的，验证了 STL-SARIMA 模型的有效性。

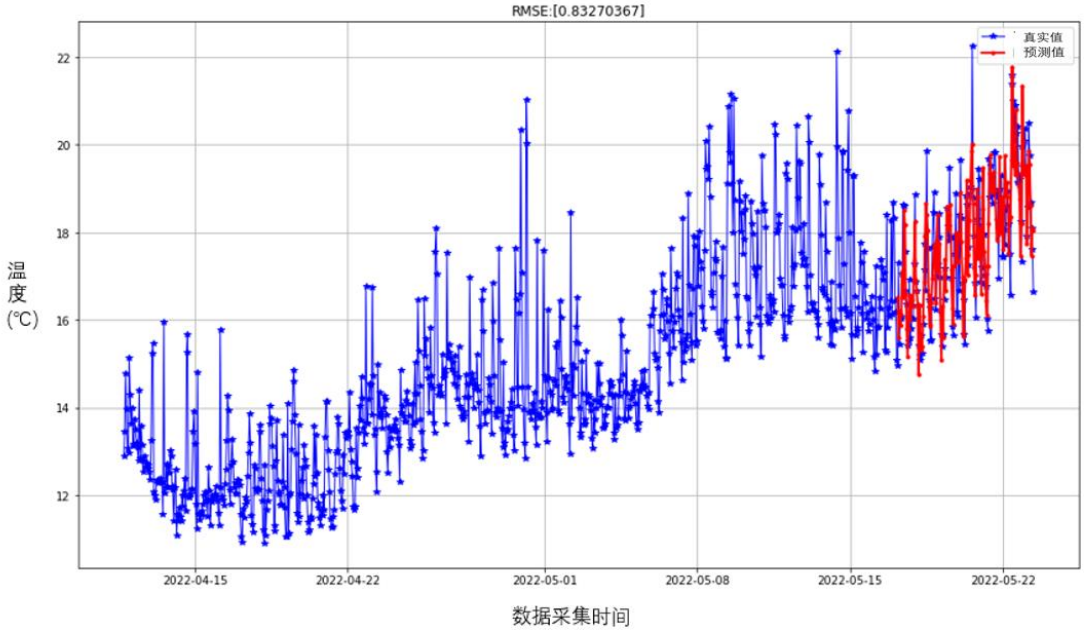


图4.33　同海域同时间不同浮标观测到的数据预测结果

4.5 海洋观测数据异常值校正

完成对原始海洋观测数据的异常检测后，将异常数据做质量标记，然后根据异常数据所在的时间节点确定模型预测的数据集，再将数据输入到海洋观测数据预测模型进行未来一段时间的预测，在预测数据中找出与异常数据相同时间点的数据作为插值的数据替换掉异常数据，最后，将原始海洋观测数据与插值修正的数据存储起来。此外，对于缺失值也可以采用异常值校正结果进行填补。

对于 3.1 在测试集中检测到的 19 个异常有效波高数据，本节采用 4.1 节的基于 STL 分解算法和 LSTM 神将网络的预测方法进行异常数据的插值修正。对 19 个异常有效波高数据的异常值校正信息如表 4.6 所示：

表 4.6 有效波高异常数据的异常值校正结果

数据采集时间	有效波高数据（分米）	加噪后的数据（分米）	异常值校正数据（分米）
2020-11-20 22:38:00	9	11	8.4
2020-12-12 16:38:00	6	7	5.1
2020-12-19 11:38:00	9	10	8.9
2020-12-29 19:38:00	15	18	16.4
2021-01-07 05:38:00	9	13	8.2
2021-01-10 03:38:00	8	6	5.1
2021-01-16 08:38:00	11	13	11.7
2021-02-02 09:38:00	8	10	8.8
2021-02-08 04:38:00	8	10	8.4
2021-02-12 13:38:00	8	11	8.1
2021-02-13 01:38:00	6	12	6.2
2021-02-23 03:38:00	9	11	10.1
2021-02-26 05:38:00	14	17	15.2
2021-02-27 02:38:00	10	11	9.4
2021-02-27 19:38:00	12	15	11.9
2021-03-02 08:38:00	9	12	9.0
2021-03-08 02:38:00	7	8	7.0
2021-04-08 15:38:00	7	6	7.2
2021-04-17 20:38:00	5	7	5.8

由表 4.6 可知，对于有效波高要素的异常值校正数据，其数据结果相对原始正常有效波高数据数值相差较小，取异常值校正数据为近似整数值，其结果同样接近于正常数据，说明可以通过有效波高预测数据对异常数据进行插值修正，证明了上述方法的海洋观测数据质量控制方法对本实验使用的有效波高数据是有效的。此外，对于有效波高数据集中的缺失值，也可由 4.1 设计的模型进行异常值校正，如上图 4.27 所示。

对于 3.2 中测试集数据检测到的 7 个异常数据中的温度数据，本节采用 4.3 节

基于 STL 分解算法和 SARIMA 的预测方法实现对温度异常数据的填补插值。对温度异常数据的异常值校正信息如表 4.7 所示：

表 4.7 温度异常数据的异常值校正结果

数据采集时间	温度原始数据（℃）	加噪后的数据（℃）	异常值校正数据（℃）
2022-05-20 10:00:00	14.87	16.477	15.024
2022-05-21 03:00:00	15.199	9.209	14.493
2022-05-22 00:00:00	15.287	19.297	15.736
2022-05-22 06:00:00	15.045	7.059	13.952
2022-05-22 16:00:00	15.517	11.52	14.966
2022-05-22 23:00:00	16.102	26.11	16.159
2022-05-23 08:00:00	15.494	25.49	16.103

由表 4.7 可知，对于温度要素的异常值校正数据，其数值与没有加入噪声的原始数据十分接近，说明可以通过预测数据对异常数据进行插值修正，证明了上述方法的海洋观测数据质量控制方法是有效且准确的。

4.6 本章小结

本章设计了基于小波分解结合 LSTM 神经网络的海洋观测数据预测模型和基于 STL 分解算法结合 LSTM 神经网络的海洋观测数据预测模型，设计了单一使用 LSTM 的海洋观测数据预测模型，并利用 RMSE、MAPE 指标和 R2 决定系数评价了模型的预测性能，对比基于小波分解重构的 LSTM 海洋观测数据预测模型、基于 STL 分解算法的 LSTM 海洋观测数预测模型的预测性能与单一的 LSTM 预测模型，结果表明本章提出的两种预测模型比只使用 LSTM 预测模型进行预测具有更高的预测精度，从而验证了本章方法的有效性，此外，对于本章使用的数据，基于小波分解重构的 LSTM 预测模型取得了更好的预测效果，但是，对于不同海洋观测数据，基于小波分解重构的 LSTM 预测模型和基于 STL 分解算法的 LSTM 预测模型需要相互验证，选取效果更好的模型进行海洋观测数据的预测。

本章 LSTM 预测模型的创新点在于小波分解单支重构信号和 STL 分解算法分解信号得到了更多的特征作为 LSTM 模型的输入，能更准确的预测未来海洋观测数据；STL 分解算法能很好的将各种周期性的时间序列进行分解；小波分解单支重构将每一支分解信号行进重构获得细节信号和概貌信号，将全部重构信号作为 LSTM 模型的输入，相较于小波分解重构是将分解信号分别预测再重构的方法，小波分解单支重构能更好的利用 LSTM 模型，获得更加准确的预测数据。

本章还设计了基于 STL 分解算法的 SARIMA 海洋观测数据预测模型，对比了不经过 STL 分解的 SARIMA 预测模型，基于 STL 分解的 SARIMA 预测模型具有

更好的预测效果，尤其是对于分解得到的趋势项可以进行高精度的预测，因此可以作为预测海洋观测数据变化趋势的重要方法。

本章最后对海洋观测数据做了异常值校正，通过使用预测数据替换异常数据的方法，实现了海洋温度要素数据的质量控制。

第 5 章 海洋观测数据质量控制软件系统的设计与实现

本章基于第 3 章的海洋观测数据异常检测模型和第 4 章的海洋观测数据异常值校正模型设计了海洋观测数据质量控制软件界面系统。本章首先对软件系统进行需求分析，然后针对不同需求进行功能划分，最后对各子功能模块进行设计，实现软件显示界面。

5.1 软件架构设计

海洋观测技术的发展促使我们获得了大量的海洋观测数据，这些数据在采集之后按照一定的数据格式被存储起来，本文研究及系统设计都是基于已经获取的延时数据。本章设计的海洋观测数据质量控制系统是为了集数据的导入、处理、异常检测、做质量标记、插值、显示和存储为一体，在保证全面性的基础上实现软件的简洁化、智能化。

海洋观测数据质量控制系统分为五个部分，数据导入模块、数据异常检测模块、数据异常值校正模块、显示模块和数据存储模块。系统的五个模块的具体功能如下：

（1） 数据导入模块

数据导入模块是用于将获取的要进行数据质量控制的海洋观测数据文件导入软件系统，导入的数据文件包括 txt 文件、csv 文件、xls 文件等。导入的数据包括时间、风速、海表面温度、波浪等数据。数据导入还包括了数据的预处理，如：数据的时间顺序处理、数据的格式处理和缺失值处理等。功能是为海洋观测数据质量控制系统提供方便系统处理的观测数据。

（2） 数据异常检测模块

数据异常检测模块主要是对导入的数据进行异常检测并做质量标识。数据异常检测模块包括导入数据的归一化、标准化处理、质控要素的选择、异常检测的方法选择、异常检测处理过程和质量标识过程。数据异常检测基于导入的数据和第三章建立的数据异常检测模型进行检测，并将结果显示在数据信息显示模块。

（3） 数据异常值校正模块

数据异常值校正模块的核心是数据预测。其主要是依据有异常标识的数据对局部数据进行建模预测，并将预测结果插值到带有异常标识的数据的位置。数据异常值校正模块包括读取异常检测模型检测的数据、根据异常标识建立预测模型和生成质控数据。数据异常值校正基于第三章得到的带有质量标识的检测数据和第四章建立的数据预测模型进行针对异常数据的预测，并将结果显示在数据信息

显示模块。

（4） 显示模块

显示模块将第三章和第四章的信息和数据显示在界面上。主要包括显示系统简介、导入数据、异常检测要素的数据曲线、质控信息、质控数据图示等。

（5） 数据存储模块

数据存储模块将得到的质控信息和质控数据存储在存储器中，以便之后的研究和使用。

系统用例图如图 5.1 所示：

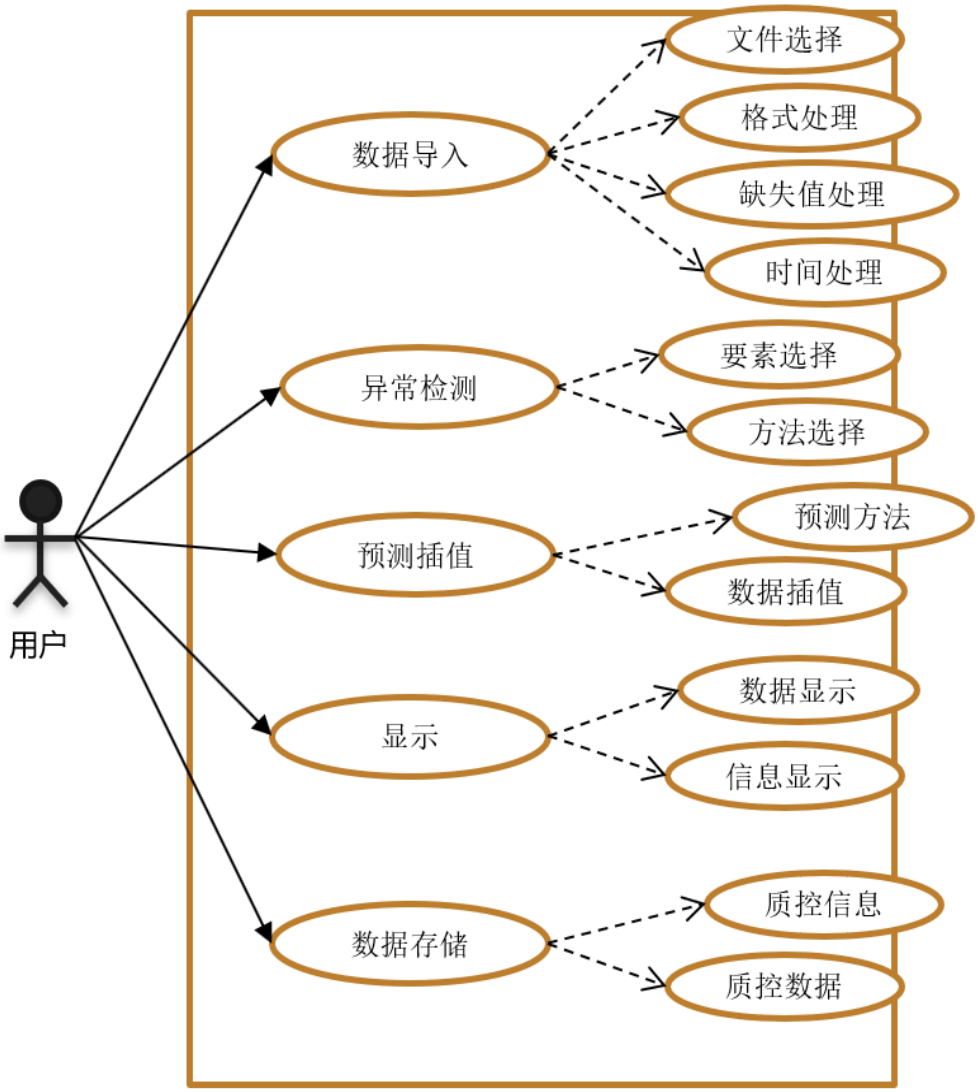


图5.1　系统用例图

5.2 软件功能划分与实现

5.2.1 系统功能模块设计

根据系统的需求分析将海洋观测数据质量控制系统分为五个部分，分别是数据导入模块、数据异常检测模块、数据异常值校正模块、显示模块和数据存储模块。其中，数据导入模块是将原始数据文件经过简单处理后导入系统，系统读取文件中的数据并进行格式处理、时间处理、缺失值处理处理之后交给异常检测模型；数据异常检测模块在选择了异常检测要素和异常检测方法后建立异常检测模型，将系统读取、处理后的数据作为异常检测模型的输入，进行异常检测得到异常检测结果；数据异常值校正模块在确定预测方法之后，读取数据异常检测结果，根据异常检测的质量标识进行数据预测，并将预测结果保存到带有异常标识的数据的位置；显示模块显示系统简介、数据信息和数据质量控制信息等，保证用户能简单快捷的看到数据的相关信息；数据存储模块将数据质量控制信息和数据质量控制数据存储到存储器中，并保留原始数据，便于之后的研究和应用使用。海洋观测数据质量控制系统功能模块如图 5.2 所示：

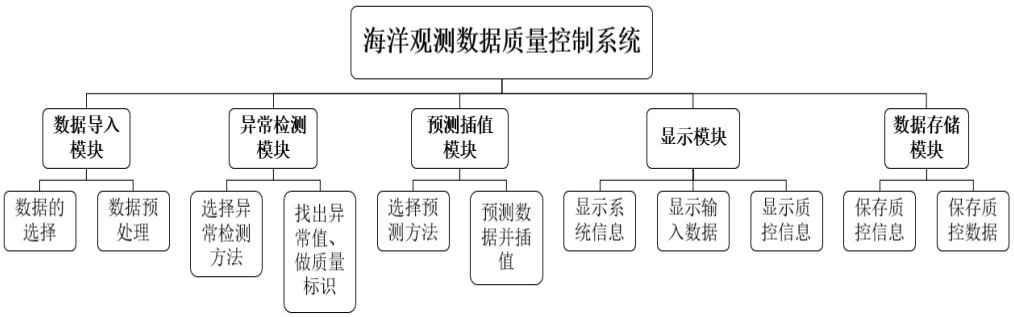


图5.2 海洋观测数据质量控制系统功能模块

5.2.2 数据导入模块的设计与实现

各个观测机构的海洋观测数据存储的文件格式不尽相同，主流的海洋观测数据存储文件有 txt 文件、csv 文件、xls 文件等。数据导入模块的设计要保证大多数种类的文件可以导入和读取，本文系统设计实现了 txt 文件、csv 文件、xls 文件和.xlsx 文件的有效读取，所以在导入数据文件时要确定文件格式。此外，数据导入模块还实现了数据的预处理，在读取到原始数据之后，数据导入模块将数据按照时间正序将数据进行排序，保证时间序列的时序特性，将数据中存在的缺失值进行简单的线性插值，减小缺失值对数据质量控制模型的影响等多种处理。数据导入模块流程图如图 5.3 所示

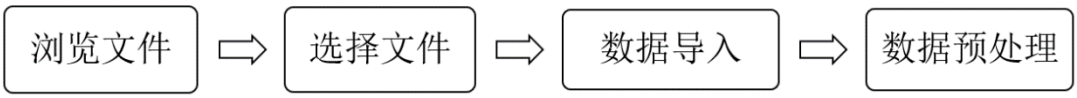


图5.3　数据导入模块流程图

5.2.3 数据异常检测模块的设计与实现

不同海洋观测数据，不同海洋观测要素根据其特性可以采用不同的异常检测方法，本章系统将第三章建立的多种海洋观测数据异常检测模型链接到数据质量控制系统，供用户自己选择异常检测方式。异常值检测模块设计了异常检测方法选择、海洋观测数据要素选择，异常信息显示等功能，其实现流程如图 5.4 所示：

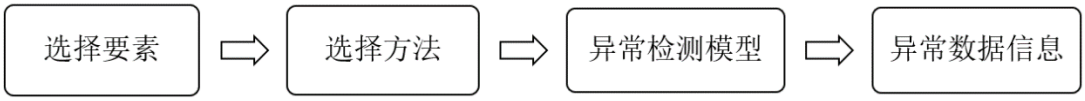


图5.4　海洋观测数据异常检测模块流程图

5.2.4 数据异常值校正模块的设计与实现

海洋观测数据异常值校正模型的核心是数据预测，本章系统将第四章建立的数据预测模型连接到数据质量控制系统，供用户自己选择预测方法。海洋观测数据异常值校正模型设计了预测方法选择、增加数据插值列功能，其实现流程如图 5.5 所示：

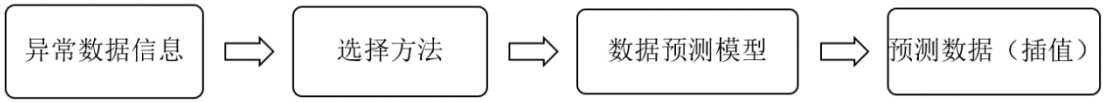


图5.5　海洋观测数据异常值校正模块流程图

5.2.5 显示模块

信息的显示对于海洋观测数据质量控制系统是必不可少的。根据数据导入模块、数据异常检测模块和数据异常值校正模块的处理，系统得到了输入数据、异常数据信息、质控数据和信息等许多用户期望直观观察到的信息，显示模块设计了原始数据表格显示窗口、目标质控要素曲线显示窗口、异常数据信息显示窗口、异常数据/异常值校正数据曲线显示窗口等，此外，显示模块也设计了显示系统简介的功能。

5.2.6 数据存储模块

完成数据质量控制的海洋观测数据需要生成新的文件并存储起来，为后续的研究和应用提供数据质量保障。数据存储模块设计的功能只有一个，就是对生成

的质控数据和异常信息进行存储。

5.3 软件界面展示

本文设计的内容均由 Python 编程实现，所以本章系统的界面设计可以选择一款适用于 python 的 GUI 设计工具。PyQt 跨平台工具包被应用于创建 GUI 应用程序，它将 Python 与 Qt 库结合，允许使用 Python 语言直接调用 Qt 库中的 API，所以本章使用 PyQt5 进行界面设计，相比直接使用 Qt 会大大提高开发效率。本章设计的海洋观测数据质量控制系统完整界面如图 5.6 所示。其中包含系统简介，数据导入与显示，异常检测信息，预测数据波形显示，质控信息显示，作者信息等。

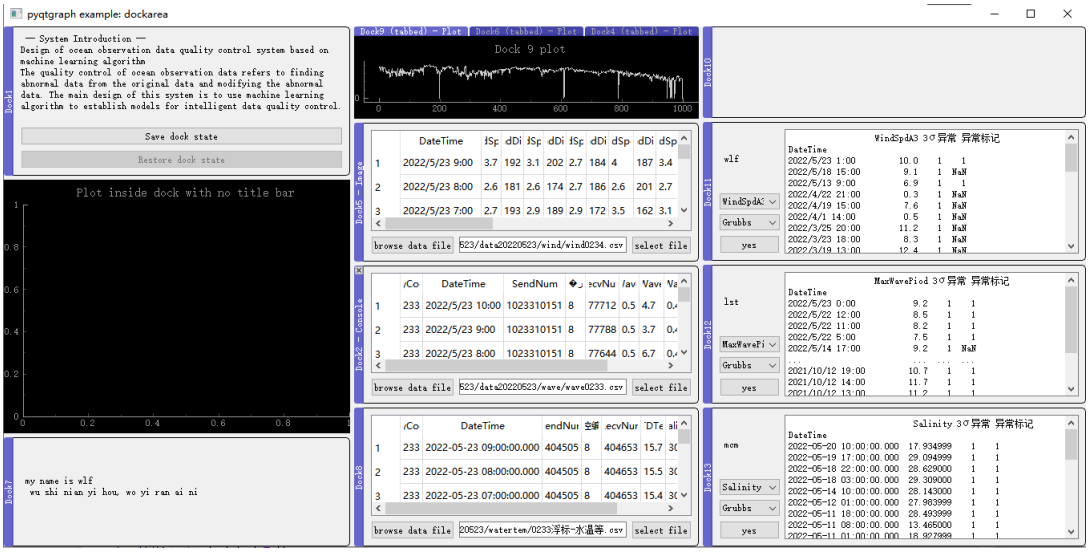


图5.6 软件完整界面

本章海洋观测数据质量控制系统界面每一部分都可以单独进行查看或者操作，数据导入模块和导入数据表格显示界面如图 5.7、图 5.8 所示：

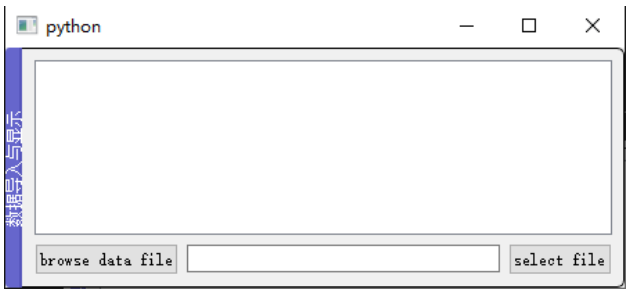


图5.7 数据导入模块

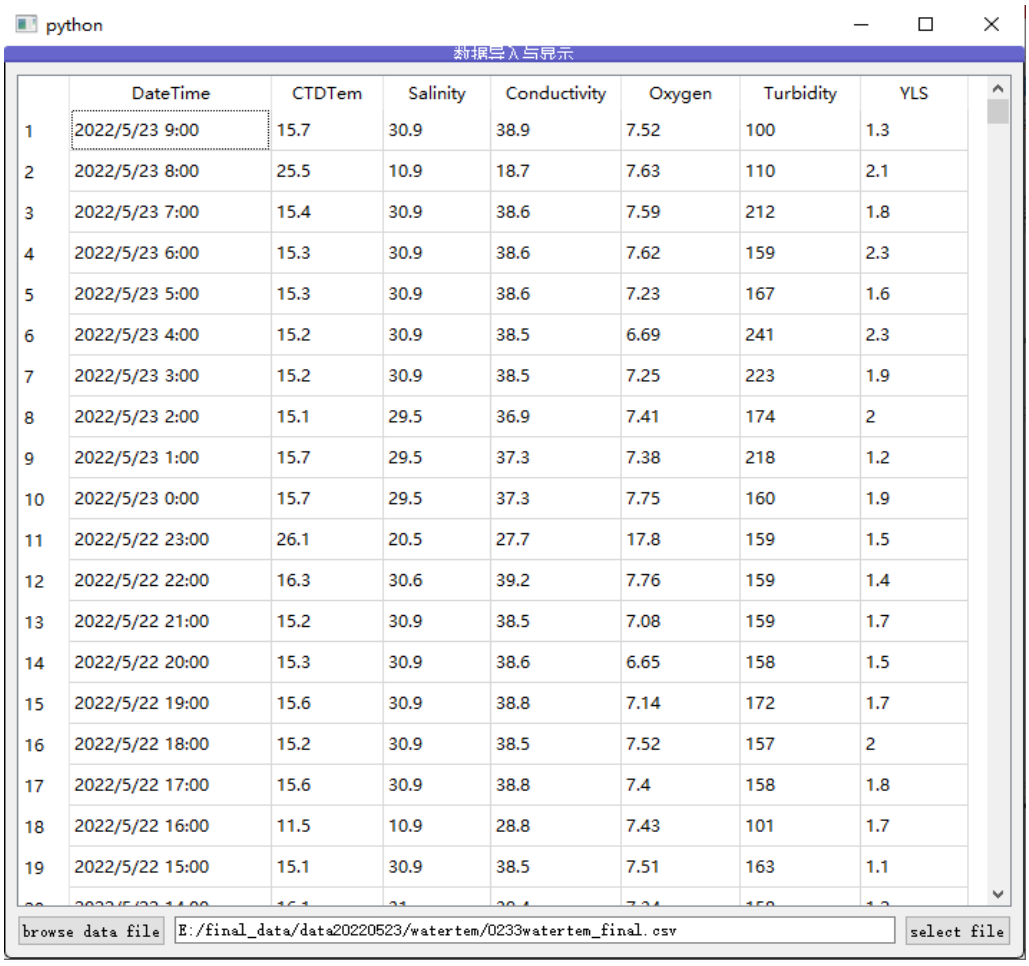


图5.8　导入数据表格显示界面

所选进行海洋观测数据质量控制的要素曲线显示界面如图 5.9 所示：

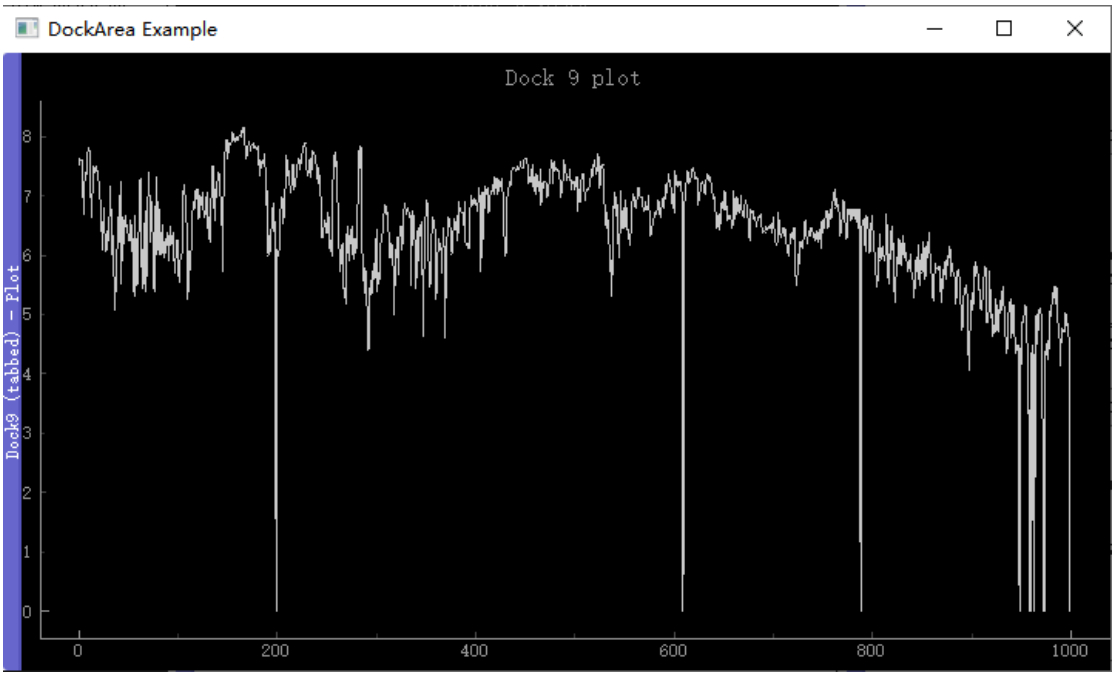


图5.9　质控要素曲线显示界面

异常检测信息显示界面如图 5.10 所示，显示的主要内容包括异常数据的时间索引、数值、异常标记等。

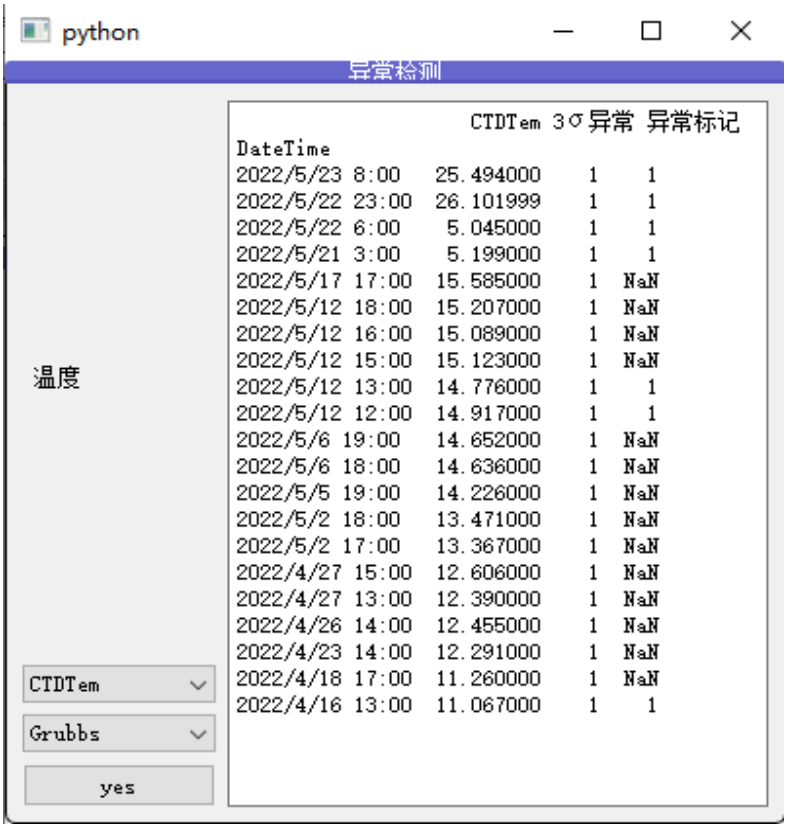


图5.10　异常检测信息显示界面

海洋观测数据预测数据曲线显示界面如图 5.11 所示：

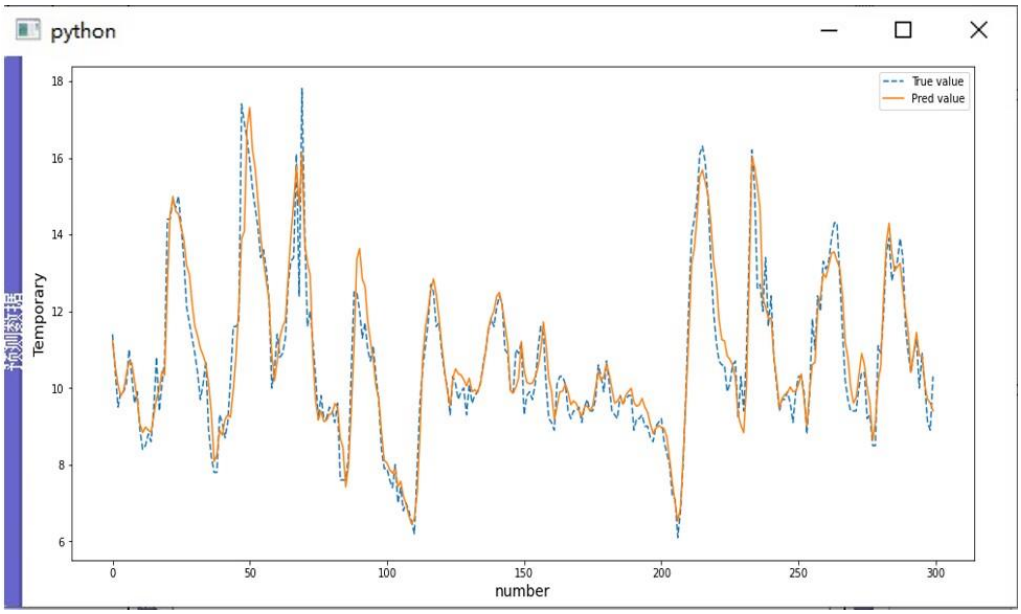


图5.11　异常数据/预测数据曲线显示界面

系统简介显示界面如图 5.12 所示：

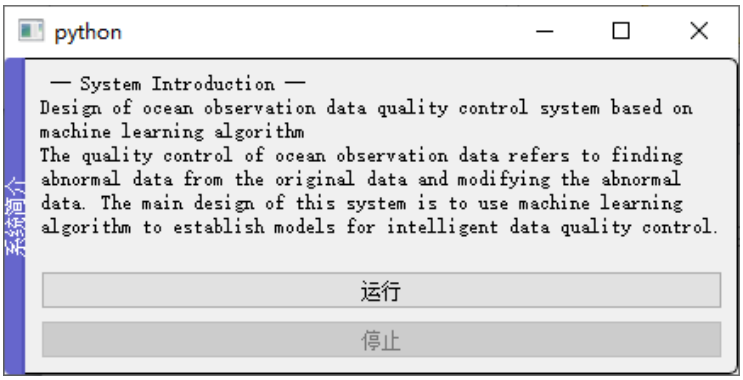
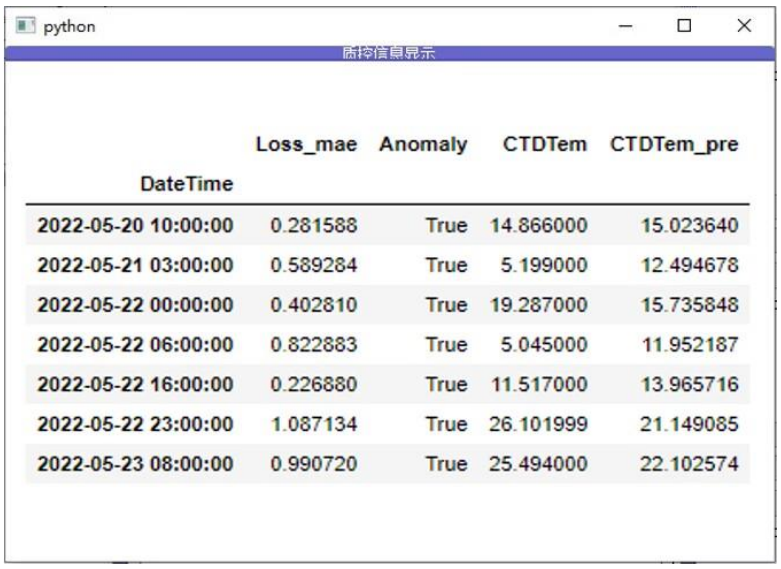


图5.12　系统简介信息

质控数据显示界面如图 5.13 所示，主要用来显示异常数据的时间索引、异常得分、异常标记、异常值和异常值校正数据。



DateTime	Loss_mae	Anomaly	CTDTem	CTDTem_pre
2022-05-20 10:00:00	0.281588	True	14.866000	15.023640
2022-05-21 03:00:00	0.589284	True	5.199000	12.494678
2022-05-22 00:00:00	0.402810	True	19.287000	15.735848
2022-05-22 06:00:00	0.822883	True	5.045000	11.952187
2022-05-22 16:00:00	0.226880	True	11.517000	13.965716
2022-05-22 23:00:00	1.087134	True	26.101999	21.149085
2022-05-23 08:00:00	0.990720	True	25.494000	22.102574

图5.13　质控数据显示界面

异常信息存储与质控数据存储操作界面如图 5.14 所示

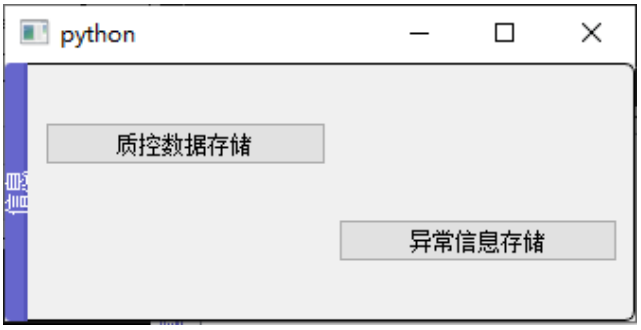


图5.14　异常信息存储与质控数据存储操作界面

5.4 本章小结

本章完成了海洋观测数据质量控制系统界面设计与实现。本章先对海洋观测数据质量控制系统进行了需求分析，然后根据需求对系统功能进行了划分和设计。本章划分的系统功能包括五个主要模块：数据导入模块、数据异常检测模块、数据异常值校正模块、显示模块和数据存储模块。再之后分别对各个功能模块进行了实现，最后使用 PyQt5 设计了系统界面并展示了相关模块的信息。

第 6 章 总结与展望

6.1 总结

海洋观测数据的正确性影响着海洋科学研究、海洋生态保护、海洋经济发展等诸多领域。由海洋浮标、海洋台站等海洋观测设备观测得到的海洋观测数据在外部因素或者数据传输过程失误的影响下，多多少少会出现数据缺失和数据异常的情况，这些影响正确性和有效性的数据可能带有大量的信息，在海洋观测数据的应用领域造成巨大的干扰。数据质量控制在许多领域被广泛研究和应用。在海洋观测领域，在海洋技术不断发展的前提下积累了大量可以用于海洋研究的观测数据，这些数据可能会存在异常且不存在异常标签，针对上述问题，本文提出了用于单变量异常检测的以统计学方法为核心的异常检测方法和用于多变量异常检测的以自编码器为核心的异常检测方法。海洋观测数据要应用于相关领域之中，不仅仅要求检测出异常和缺失，而且需要将更加准确可靠的数据对异常和缺失进行插值填补，针对这个问题，本文提出了以 LSTM 为核心的数据预测模型和以 ARIMA 为核心的数据预测模型。结合上述两个部分，就构成了海洋观测数据质量控制模型，能够以较高的标准完成质量控制任务。主要工作如下：

（1）对于单变量的海洋观测数据，以 Grubbs 准则和 3σ 准则为基础，考虑到统计学方法会带来的数据漏判和误判，提出了基于 Grubbs 准则和 3σ 准则结合局地异常检测和误差控制的海洋观测异常检测方法，并经过实验验证，说明所用方法的有效性。

（2）对于多变量的海洋观测数据，考虑到所得海洋观测数据没有数据标签，提出了以自编码器为核心的海洋观测数据异常检测方法。通过串行连接两个自编码器的方法得到了效果良好，结构简单的自编码器异常检测模型。

（3）对于经过异常检测模型检测得到的异常数据，通过建立基于 STL 分解和 LSTM 的预测模型，基于小波分解重构和 LSTM 的预测模型以及基于 STL 分解和 SARIMA 的预测模型对数据进行预测，验证了预测模型的有效性并用预测数据进行异常值校正。

（4）对整个海洋观测数据异常检测和异常值校正方法的实现设计软件界面显示系统，实现了数据导入、数据显示、方法选择、要素选择、数据存储的人机交互与智能显示。

6.2 展望

海洋观测数据质量控制技术有着广泛的应用前景，特别是在当下海洋观测技术快速发展的环境下，准确、快速、有效地对所得数据进行质量控制具有重要意义。本文针对海洋观测数据质量控制的两个方面，分别在异常检测和异常值校正两个过程上提出了一些方法。但这两部分所用方法也存在一些局限性，需要继续研究和不断尝试，而且随着机器学习算法的不断研究和发展，能够应用于海洋观测数据质量控制的方法也层出不穷，需要在后续研究中不断试验。总体来讲，本文所做的工作还可以在一下几个方面进行研究：

（1）在海洋观测数据异常检测方面，可以对由极端天气变化造成的数据波动进行分析，减少对特殊情形下造成的数据波动的误判。

（2）考虑结合一片海域内多个观测点位的观测数据进行海洋观测数据质量控制。

（3）在海洋观测数据预测方面，由于 transformer、informer 等算法的兴起，可以尝试将这些方法应用于海洋观测数据预测领域。

（4）对于海洋观测数据的应用研究，进一步结合海洋科学的相关理论进行分析研究是本文工作之后需要的重要研究方向。

（5）本文研究的是海洋观测延时数据质量控制，可以考虑将基于机器学习算法的海洋观测数据质量控制应用到实时海洋观测数据中，提高实时海洋观测数据质量控制的有效性，形成实时——延时海洋观测数据质量控制系统。

参考文献

[1] 卢勇夺,王朝阳,王豹,等. 我国海洋锚系浮标数据异常值检测方法研究——以QF110 和 QF306 为例[J]. 海洋预报,2019,36(6):37-43.

[2] National Data Buoy Center. Handbook of Automated Data Quality Control Checks and Procedures[M]. 刘愉强，郭少琼，朱鹏利译. 自动化数据质量控制检查和程序手册指南. 北京：海洋出版社, 2018.1:1-2.

[3] GB/T 14914.6-2021,海洋观测规范第 6 部分：数据处理与质量控制[S]. 中国标准出版社, 2021.

[4] Kim H J , Park S M , Choi B J , et al. Spatiotemporal Approaches for Quality Control and Error Correction of Atmospheric Data through Machine Learning[J]. Computational Intelligence and Neuroscience, 2020, 2020(4):1-12.

[5] 谭哲韬,张斌,吴晓芬,等. 海洋观测数据质量控制技术研究现状及展望[J]. 中国科学（地球科学）,2022,52(3):418-437.

[6] Hawkins D M. Identification of Outliers[M]. London: Chapman and Hall, 1980:30-100.

[7] 赵曼. 基于数据相关性的异常检测算法研究[D]. 北京：北京交通大学，2017.

[8] Beltrami G M. Automatic, real-time detection and characterization of tsunamis in deep-sea level measurements[J]. Ocean Engineering, 2011, 38(14-15):1677-1685.

[9] Hassani V, Pascoal, António M, S Rensen A J. Detection of mooring line failures using Dynamic Hypothesis Testing[J]. Ocean Engineering, 2018, 159(1):496-503.

[10] Baldacci L, Golfarelli M, Lombardi D, et al. Natural gas consumption forecasting for anomaly detection[J]. Expert Systems with Applications, 2016, 62(15):190-201.

[11] Ahamed A H R, Fariza N, Abdullah G, et al. Real-time big data processing for anomaly detection: A Survey[J]. International Journal of Information Management, 2018, 45:289-307.

[12] Wu J, Zeng W, Yan F. Hierarchical Temporal Memory method for time-series-based anomaly detection[J]. Neurocomputing, 2018, 273:535-546.

[13] Tran T.M., Le X.M.T., Nguyen, H.T., et al. A novel non-parametric method for time series classification based on k-Nearest Neighbors and Dynamic Time Warping Barycenter Averaging[J]. Engineering Applications of Artificial Intelligence, 2019, 78 (2):173–185.

[14] Benkercha R, Moulahoum S. Fault detection and diagnosis based on C4.5 decision tree algorithm for grid connected PV system[J]. Solar Energy, 2018, 173:610-634.

[15] Tian Y, Mirzabagheri M, Bamakan S M H, et al. Ramp loss one-class support vector machine; A robust and effective approach to anomaly detection problems[J]. Neurocomputing, 2018, 310(8):223-235.

[16] Mekki H, Mellit A, Salhi H. Artificial neural network-based modelling and fault detection of partial shaded photovoltaic modules[J]. Simulation Modelling Practice & Theory, 2016, 67:1-13.

[17] Grigorios, Tzortzis, Aristidis, et al. The MinMax k-Means clustering algorithm[J]. Pattern Recognition, 2014, 47(7): 2505-2516.

[18] Gan G , Ng K P . k -means clustering with outlier removal[J]. Pattern Recognition Letters, 2017, 90(15):8-14.

[19] Roul R K , Sahay S K . Semi-supervised clustering using seeded-kMeans in the feature space of ELM[C]// India Conference. IEEE, 2017.

[20] Fan C , Zhang T , Yang Z , et al. A Text Clustering Algorithm Hybriding Invasive Weed Optimization with K-Means[C]// Ubiquitous Intelligence & Computing & IEEE Intl Conf on Autonomic & Trusted Computing & IEEE Intl Conf on Scalable Computing & Communications & Its Associated Workshops. IEEE, 2016.

[21] Kadri F, Harrou F, Chaabane S, et al. Seasonal ARMA-based SPC charts for anomaly detection: Application to emergency department systems[J]. Neurocomputing, 2016, 173(15): 2102-2114.

[22] MARTIN, HEESEMANN, TANIA, et al. Ocean Networks Canada: From Geohazards Research Laboratories to Smart Ocean Systems[J]. Oceanography, 2014, 27(2):151-153.

[23] Abeysirigunawardena D , Jeffries M , Morley M G , et al. Data Quality Control and Quality Assurance Practices for Ocean Networks Canada Observatories: Challenges and Opportunities[C]// OCEANS 2015 - MTS/IEEE Washington DC. IEEE, 2015.

[24] Cowles T , Delaney J , Orcutt J , et al. The Ocean Observatories Initiative: Sustained Ocean Observing Across a Range of Spatial Scales[J]. Marine Technology Society journal, 2010(6):44.

[25] M. Smith , L. Belabbassi , L. Garzio , et al., Automated quality control procedures for real-time ocean observatories initiative datasets[C]. OCEANS 2017 - Anchorage, Anchorage, AK, USA, 2017, pp. 1-4.

[26] M. F. Vardaro et al., OOI data quality procedures and tools building on the first year of operations[C]. OCEANS 2017 - Anchorage, Anchorage, AK, USA, 2017, pp. 1-5.

[27] Rahman A , Smith D V , Timms G . Multiple classifier system for automated quality assessment of marine sensor data[C]// Intelligent Sensors, Sensor Networks and Information Processing, 2013 IEEE Eighth International Conference on. IEEE, 2013.

[28] Rahman A , Smith D V , Timms G . A Novel Machine Learning Approach Toward Quality Assessment of Sensor Data[J]. IEEE Sensors Journal, 2014, 14(4):1035-1047.

[29] Timms G P , Souza P , Reznik L , et al. Automated Data Quality Assessment of Marine Sensors[J]. Sensors (Basel, Switzerland), 2011, 11(10):9589-9602.

[30] Smith D , Timms G , Souza P D , et al. A Bayesian Framework for the Automated Online Assessment of Sensor Data Quality[J]. Sensors (Basel, Switzerland), 2012, 12(7):9476-9501.

[31] Karakuş O, Kuruoğlu E E, Altınkaya M A. One-day ahead wind speed/power prediction based on polynomial autoregressive model[J]. IET Renewable Power Generation, 2017, 11(11): 1430-1439.

[32] Arumugam P, Saranya R. Outlier Detection and Missing Value in Seasonal ARIMA Model Using Rainfall Data [J]. Materials Today: Proceedings, 2018, 5(1): 1791–1799.

[33] Wani M R, Wani M A, Riyaz R. Cluster based approach for mining patterns to predict wind speed[C]. 2016 IEEE International Conference on Renewable Energy Research and Applica-tions (ICRERA). Birmingham, UK: IEEE, 2016: 1046-1050.

[34] Demolli H, Dokuz A S, Ecemis A, et al. Wind power forecasting based on daily wind speed data using machine learning algorithms[J]. Energy Conversion and Management, 2019, 198: 111823.

[35] Tomin N, Sidorov D, Kurbatsky V, et al. A hybrid wind speed forecasting strategy based on Hilbert-Huang transform and machine learning algorithms[C]. 2014 International Conference on Power System Technology. Chengdu, China: IEEE, 2014: 2980-2986.

[36] Shi B, Wang P, Jiang J, et al. Applying high-frequency surrogate measurements and a wavelet-ANN model to provide early warnings of rapid surface water quality anomalies[J]. Science of The Total Environment, 2018, 610:1390–1399.

[37] Zhao Pushe, Masaru Kurihara, Tanaka Junichi, et al. Advanced Correlation-Based Anomaly Detection Method for Predictive Maintenance[C]. IEEE International Conference on Prognostics and Health Management (ICPHM). IEEE, 2017:78-83.

[38] Li S, Wen J. A model-based fault detection and diagnostic methodology based on PCA method and wavelet transform[J]. Energy & Buildings, 2014, 68:63-71.

[39] Khodayar M, Teshnehlab M. Robust deep neural network for wind speed prediction[C]. 2015 4th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS). IEEE, Zahedan, Iran: 2015: 1-5.

[40] 刘首华,陈满春,董明媚,高志刚,张建立,武双全,林峰竹. 一种实用海洋浮标数据异常值质控方法[J]. 海洋通报, 2016, 35(3):7.

[41] 郭颜萍，胡桐，漆随平. 基于小波变换和 LS-SVM 的船面风速风向估算方法[J]. 海洋技术学报, 2016, 35(2):66-70

[42] Zhou Y , Qin R , Xu H , et al. A Data Quality Control Method for Seafloor Observatories: The Application of Observed Time Series Data in the East China Sea[J]. Sensors, 2018, 18(8): DOI:10.3390/s18082628.

[43] 王国松,王喜冬,侯敏,等. 基于观测和再分析数据的 LSTM 深度神经网络沿海风速预报应用研究[J]. 海洋学报（中文版）,2020,42(1):67-77.

[44] 贺琪，查铖，宋巍，戚福明，郝增周，黄冬梅. 基于 STL 的海表面温度预测算法[J]. 海洋环境科学, 2020, 39(6):8.

[45] Robert C, William C, Irma T. STL: A seasonal-trend decomposition procedure based on lo-ess[J]. Journal of Official Statistics, 1990, 6(1): 3-73.

[46] Yin H, Jin D, Gu Y H, et al. STL-ATTLSTM: vegetable price forecasting using STL and attention mechanism-based LSTM[J]. Agriculture, 2020, 10(12): 612.

[47] Zhao Y, Ma Z, Yang Y, et al. Short-term passenger flow prediction with decomposition in urban railway systems[J]. IEEE Access, 2020, 8: 107876-107886.

[48] Li Y, Bao T, Gong J, et al. The prediction of dam displacement time series using STL, extra-trees, and stacked LSTM neural network[J]. IEEE Access, 2020, 8: 94440-94452.

[49] Chen D, Zhang J, Jiang S. Forecasting the short-term metro ridership with seasonal and trend decomposition using loess and LSTM neural networks[J]. IEEE Access, 2020, 8: 91181-91187.

[50] 周松林，茆美琴，苏建徽.基于小波分析与支持向量机的风速预测[J].太阳能学报,2012,33(3):452-456.

[51] Antonella M ,M.J. D T ,Manuele B . Detecting outliers from pairwise proximities: Proximity isolation forests[J]. Pattern Recognition,2023,138.

[52] Yikun Y ,Zengyi S ,Fengxia L , et al. Short-term Wind Speed Prediction Based on Improved Auto Encoder[J]. Journal of Physics: Conference Series,2023,2418(1).

[53] Sopheap K ,ChangSung K ,KwangJae S , et al. Fast Detection of Current Transformer Saturation Using Stacked Denoising Autoencoders[J]. Energies,2023,16(3).

[54] Xuegui L ,Shuo F ,Nan H , et al. Surface microseismic data denoising based on sparse autoencoder and Kalman filter[J]. Systems Science & Control Engineering,2022,10(1).

[55] Yue L ,Cong W ,Yanan T , et al. Parameter-shared variational auto-encoding adversarial network for desert seismic data denoising in Northwest China[J]. Journal of Applied Geophysics,2021(prepublish).

[56] Zhang J ,Huang Y ,Huang C , et al. Research on ARIMA Based Quantitative Investment Model[J]. Academic Journal of Business & Management,2022,4(17).

[57] Jianing W ,Hongqiu Z ,Yingjie Z , et al. A novel prediction model for wind power based on improved long short-term memory neural network[J]. Energy,2023,265.

[58] Bansal A C ,Sain V D . An adaptive RNN algorithm to detect shilling attacks for online products in hybrid recommender system[J]. Journal of Intelligent Systems,2022,31(1).

[59] 高梦琦, 昌锡铭, 王欢. 基于小波分解和长短时记忆网络的地铁进站量短时预测[J]. 山东科学, 2019, 32(4):8.