项目级别：＿＿校级＿

项目编号：＿20211469

中山大学大学生创新训练计划

项目总结报告

项目名称：基于机器学习的珠江口东侧水域低氧长期变化规律与驱动因素研究

项目完成人：杨涵，陈至雪，潘文轩

指导老师：胡嘉镗

所在院系：环境科学与工程学院

研究起止时间：2021.4至2021.12

教务部

二○二〇年十二月 制

**基于机器学习的珠江口东侧水域低氧长期变化规律与驱动因素研究**

**摘要：**近年来，由于周边城市人口的快速增长和社会经济发展，珠江接收了大量的营养物质和各种污染物，导致珠江口夏季频繁出现缺氧现象。珠江口的缺氧问题已得到广泛关注，以往研究大多关注低氧形成的潜在过程和控制因素，却未能有效量化。使用复杂的水质模型是预测水质的主要方法，但是，受大量复杂的计算，严格的边界条件等限制，这些模型不容易采用，需要极高的计算成本。本研究基于多元回归、随机森林等数据驱动的机器学习模型，定量地探究珠江口东侧的香港水域长期缺氧的规律及其驱动因素。研究结果表明，叶绿素a和风速使香港水域低氧长期变化的主要因素，其中，风速的影响更为显著。同时，风速也是促使低氧长期趋势形成的最主要原因。

**关键词：**珠江口，低氧长期变化，机器学习，驱动因素

**Research on long-term changes and driving factors of hypoxia in the waters east of the Pearl River Estuary based on machine learning**

**Abstract:** In recent years, due to the rapid population growth and socio-economic development in the surrounding cities, the Pearl River has received a large amount of nutrients and various pollutants, resulting in frequent hypoxia in the Pearl River Estuary in summer. The problem of hypoxia in the Pearl River Estuary has received widespread attention. Previous studies have mostly focused on the underlying processes and control factors of hypoxia, but they have not been effectively quantified. The use of complex water quality models is the main method to predict water quality. However, due to a large number of complex calculations and strict boundary conditions, these models are not easy to adopt and require extremely high calculation costs. Based on data-driven machine learning models such as multiple regression and random forest, this research quantitatively explores the law of long-term hypoxia and its driving factors in Hong Kong waters on the east side of the Pearl River Estuary. The research results show that chlorophylla and wind speed are the main factors for the long-term changes in low oxygen in Hong Kong waters. Among them, the influence of wind speed is more significant. At the same time, the wind speed in summer is also the main reason for the formation of a long-term trend of hypoxia.

**Keywords:** Pearl River Estuary, long-term changes in hypoxia, machine learning, driving factors

# 一、研究背景

由于人类活动与气候变化，全球海域频繁爆发水体缺氧现象，对其沿岸生态系统产生破坏性的影响，河口及其邻近海域的低氧/缺氧问题由此也成为全球关注的重点。

海域水体缺氧现象背后的机理极为复杂，涉及到水质，水生态，水动力过程及生物化学过程等多方面的影响及水体的响应机制。国内外学者对此进行了大量的研究。如李等人采用化学生物标志物及微化石等对长江口富营养化及低氧水域环境的形成原因进行了探究；1 Ni等采用耦合的物理-生物化学地球模型对切萨皮克湾长期低氧现象进行了模拟与探究。2

珠江是中国第二大河流，径流量在全球排第13位，约为330×109m3/a。自20世纪80年代以来，珠江中的缺氧问题持续收到关注。如图1所示为1976-2017年夏季珠江口底部的溶解氧分布，图中蓝色区域代表缺氧区域，该图直观反映出珠江口缺氧现象在过去几十年长期持续的扩展趋势。 3由于过去几十年工业、农业活动和珠三角城市群的快速发展，珠江口接纳了大量的人为碳物质和营养盐负荷。自而随着珠三角经济社会的进一步发展，珠江口的入海营养盐负荷将不断升高。4沿海低氧系统中水体养分含量与低氧环境之间的强相关关系已有诸多文献记载与说明。1由此，大量营养元素的输入为珠江河口水域低氧环境形成的重要因素。

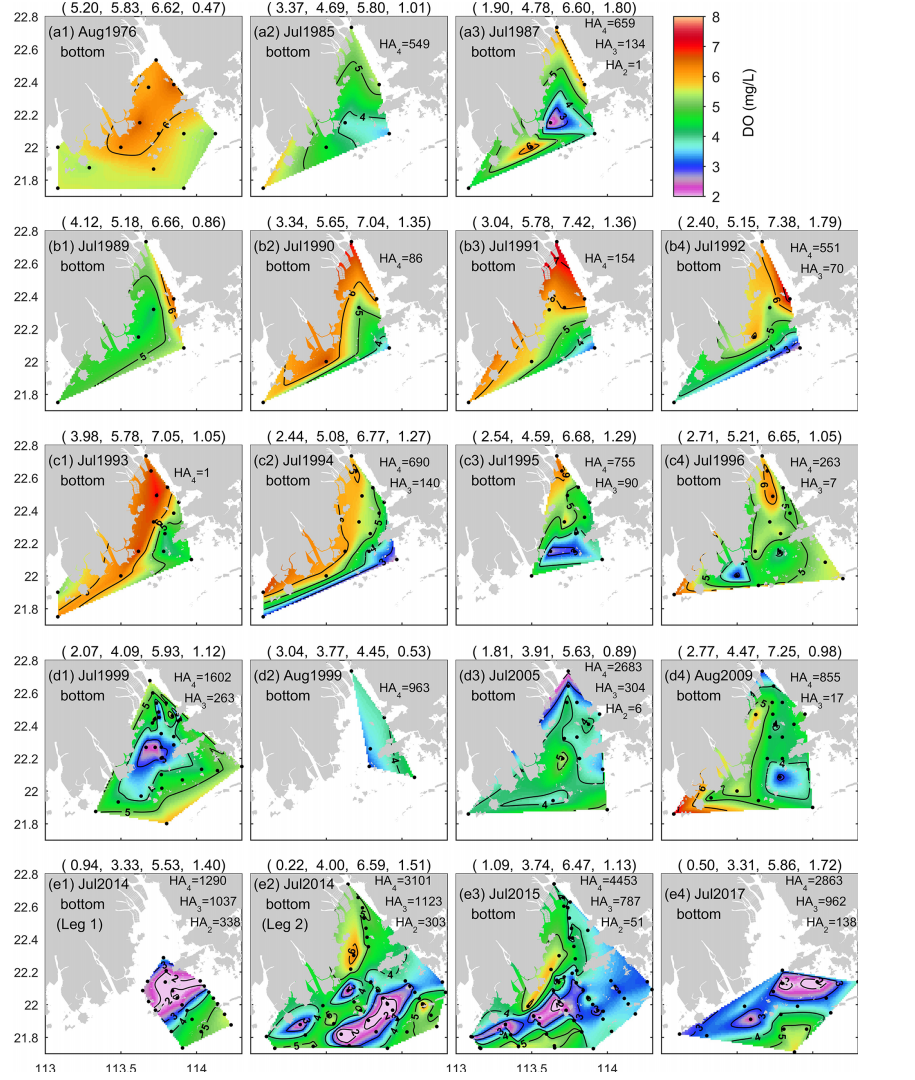


图 1 1976-2017年夏季PRE底部的DO分布。请注意，HA4、HA3和 HA2分别表示基于观测的面范围（km2）估计值**，分别为DO< 4、DO < 3 和 DO < 2 mg L−1。**3

Fig.1 DO distributions at the bottom of the PRE for the summer months of 1976–2017. Note that HA4, HA3, and HA2 represent the observation-based estimates of the areal extent (km2) with DO < 4, DO < 3, and DO < 2 mg L−1,respectively.

此外，近年来越来越多的研究表明气候变化助长了沿海海洋缺氧的蔓延。气候变化可以改变海洋环流或垂直分层，从而改变氧源和汇之间的平衡5除此外，珠江口海域低氧形成还与其水体化过程，有机碳颗粒(POC)等相关。6

前人对珠江水域的时空分布特征与低氧现象形成的影响因素已有所探究，总的来说，以前的研究表明，引起珠江口水域缺氧既有自然过程和人为过程，包括富营养化，水体分层，气候变化等等。 7但其采用的传统的统计学习方法，无法对珠江低氧与其影响因子非线性、时滞性、时空异质性的复杂响应关系进行准确地探究与预测。8

利用复杂的水质模型是预测水质的主要方法，但是，由于诸如因素如数据限制，大量的计算以及严格的边界条件的限制，这些模型不容易采用，需要极高的计算成本。尽管水质模型如EFDC模型非常有能力模拟水质的复杂动态，但是水质预测仍然难以进行,因其需要大量的时间与数据构建模型，且模型的准确率极度依靠边界条件的确定。9作为替代方案，使用数据驱动的方法可以有效的建立水质变量之间的关系，并不需要对边界条件进行划分，从而节省了计算

机器学习模型是一类新兴的数据分析技术，该类方法直接从数据中挖掘特征，而无需依赖于预定的方程，其主要优势在于在足够的数据训练后，它能对复杂非线性关系进行准确地拟合。如被誉为最好的机器学习算法之一的随机森林算法(Random forest, RF)，它具有如极高的准确率，极强的数据分析挖掘能力，分析复杂相互作用分类特征，以及能给出变量重要性评估等诸多优点。10

理论上说，如随机森林此类机器学习算法为分析及量化水体富营养化的影响因素及机制提供了新的思路与方法，但目前使用机器学习模型并将其应用于海域或湖泊低氧水体环境形成机理的探究仍鲜有报道。

近些年深度学习理论的突破，不断有更新更有效的神经网络模型被开发及应用。其中长短期记忆网络(Long short-term memory, LSTM)常用于时间序列变量的模拟，将尤其适用于对水质的模拟与预测，不过其目前在水质预测的应用极少。11尽管某些经典机器学习方法，例如人工神经网络（Artificial Neural Network，ANN），也可用于时间序列预测，12但人为创造的特征以解释时间依赖性，它们的所利用的时序信息是有限的。13而且，它们的性能总是比 LSTM 差。14

综上所述，前人对珠江口长期缺氧的现象已有观察，但对长期缺氧的影响因素未能有效量化。本研究旨在探索珠江口东侧香港水域底层溶解氧的长期变化规律及其驱动因素，并基于机器学习模型将其量化。先是依据溶解氧与其响应变量之间的响应关系构建了溶解氧的预测模型，基于所构建的精确模型进一步探讨各影响变量的贡献程度，即探究引起珠江口东侧水域长期缺氧的驱动因素，并比较了所构建不同机器学习模型的性能。

# 二、研究材料与方法：

## 2.1 研究区域概述

本研究以1984-2018年香港环保署水质监测站外围的十个站点测定的溶解氧含量等水质指标数据为研究对象。具体监测站位详见图2。香港环保署自1986 年以来实施了一项全面的海洋水质时空监测计划，计划对海域中的76个水质监测站点水柱表、中、底层进行每两月或每月一航次的巡航采样与监测。该计划为掌握香港水域富营养化的变化及其影响提供了一个有价值的、超过30年的长期数据集。考虑到维多利亚港和后海湾受内源排放的影响太大，低氧现象难以解释，故选取香港监测外围一圈站点，即NM5、NM6、NM8、SM17~SM20、MM8、MM13、MM14共计十个站点，代表珠江口东侧香港水体情况。此外还收集了由万山岛气象站观察的气象数据。

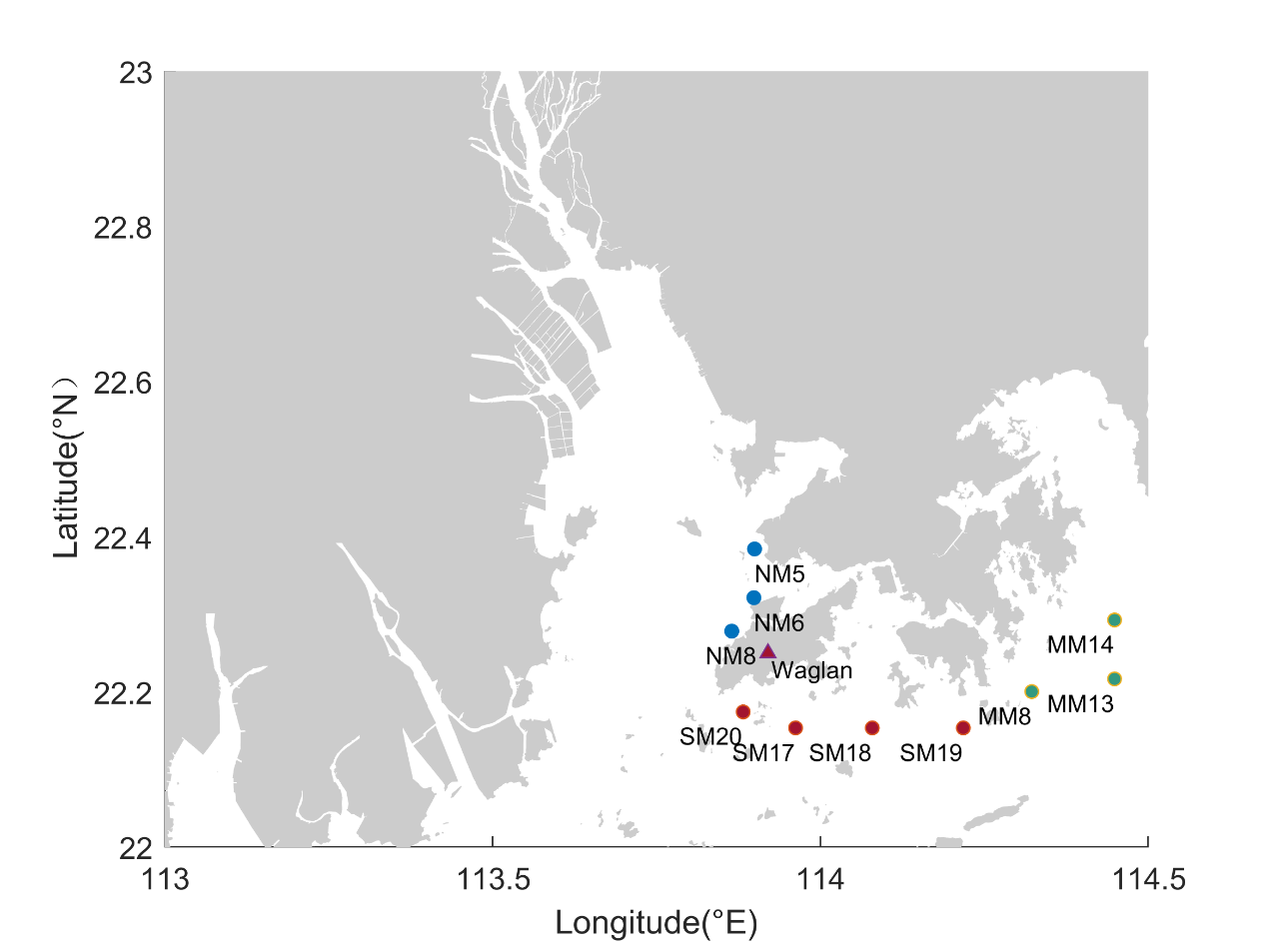
## 2.2 数据处理

**2.2.1数据同化**

根据前人研究，表层叶绿素含量，底层水温，表层营养盐，风速，径流流量等与珠江口东侧水体缺氧情况密切相关8，因此将其作为模型的输入变量。据观测数据，表层叶绿素含量，底层水温，表层营养盐，密度差在十个站点各不相同，此外，不同站点的时间段和数据缺口长度存在差异，为使每个站点用于研究的水文数据相对均衡，对已有的水文数据进行了数据同化。具体为将水质指标存在的缺失值使用对应月份平均值代替，同一年某个月份中存在多个数据的对其取平均值，使得在所研究时段1984-2018年25年内每一个月份都有数值，最终使每个站点的数据均成为25年逐月的共300个样本的连续时序数据。随后，为减少模型变量计量单位差异造成的误差，对输入变量和输出变量进行了归一化处理。通过减去变量平均值，然后除以其标准差，对所有输入输出变量进行了标准转换实现归一化。如式(1)所示：

（1）

式中，与分别为标准化处理前后的数据，为数据均值，为数据标准差。



**图2** **选取珠江口水文站点与气象站点分布**

Fig.2 Distribution of hydrological stations and weather station in the Pearl River Estuary

**2.2.2 主成分分析**

主成分分析（Principal Component Analysis，PCA）是一种十分有效的数据处理技术，可降低大型数据集的维度，提高模型可解释性，同时最大限度地减少信息丢失。将相互关联的变量合成创建新的变量作为主要组件，从而最大化数据集的方差。 15

为了量化缺氧严重程度，我们应用了 PCA将外围十个站点的溶解氧均值，溶解氧最小值，低氧频率，低氧深度，低氧面积等降维为一个反映珠江口外围面域缺氧情况的综合缺氧指标（Hypoxic Index, HI）。 16得到的HI与溶解氧均值，溶解氧最小值成反比关系，而与低氧频率，低氧深度，低氧面积等成正比关系。HI越大，说明水体缺氧现象越严重。

主成分分析得到的HI在1984~2018共25年内每个月份都有数据，对各个月份的HI数值进行了统计，发现在夏季HI的数值最高，在非夏季HI的数值均偏小，如图一所示。这与前人研究中发现珠江口东侧水域缺氧主要发生在夏季的结果一致，本研究所探究的主要对象即发生在夏季的珠江口东侧水域长期缺氧，因此此处取各年夏季6~8月份的HI数值平均得到夏季平均HI作为该年的缺氧指标。下文中所有HI均为夏季平均HI。

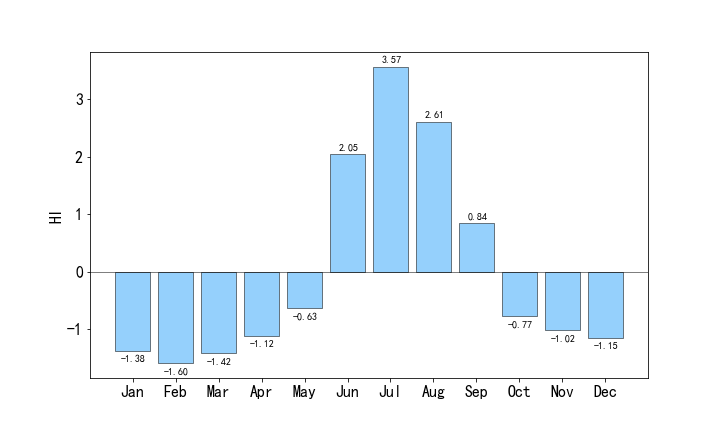
****

图3 1994~2018年HI月平均分布直方图

Fig.3 Histogram of monthly average distribution of HI from 1994 to 2018

## 2.3研究模型

**2.3.1 模型介绍**

本研究使用机器学习模型包括长短期记忆神经网络LSTM，随机森林RF，多元回归。多元回归模型是最简单也是最常使用的多元统计模型。随机森林目前在水质预测中已有一定应用，且成效较为显著。LSTM为新兴的机器学习模型，在预测时序数据上具有优势，但在水质预测中的应用较少。本质上说，从效果上看，都是通过预测变量与响应变量的响应关系建立回归模型。与多元回归最大的区别在于随机森林和LSTM能有效捕捉变量间的非线性作用。而LSTM与随机森林之间不同在于，LSTM因其模型原理输入需为连续发生的时序数据，而随机森林不受此限制。此外，LSTM通常需要更大的数据集才能逐渐体现其优势。

**2.3.2 建模过程**

机器学习模型的构建分为训练过程与测试过程，将一组数据作为模型的输入训练得到模型方程，再将另一组未曾在训练过程中出现的数据作为模型输入，将模型输出与实际数值比较，计算评价函数用于评测所构建模型的性能。我们将原始数据进行归一化后，按照一定的比例将数据集合随机划分为训练集与测试集。并通过十折交叉验证配合网格搜索等方法不断优化模型的超参数，减少模型过拟合，以生成最通用的预测统计模型。计算纳什系数，均方误差，决定系数等作为评判预测模型精度的指标。为减少训练测试集选择的不确定性造成的影响，共进行1000组训练与测试，其均值用以评测最终模型性能。本研究中所有统计模型均使用此种模式构建。

## 2.4模型比较与评估

**2.4.1 评价函数**

本研究中衡量评价模型的性能采用均方根误差(Root mean square error, RMSE)、皮尔逊相关系数(Pearson correlation coefficient, Pcorr)决定系数R2和纳什效率系数(Nash efficiency coefficient, NSE)。计算所得的RMSE的值越小, Pcorr，NSE，R2值越接近1,则模型预测的精度越高，可信度也越高，具体公式如下：

|  |  |  |
| --- | --- | --- |
|  |  | (2) |
|  |  | (3) |
|  |  | (4) |
|  |  | (5) |

对预测的数据我们选择了纳什系数用于评价模型的质量，均方根误差用于测量预测的精密度，决定系数用于衡量模型对数据的可解释程度，以及皮尔森相关系数衡量预测值与真实值之间的相关性。评价模型的整体表现时，以NSE作为主要的评判指标。通常认为介于 0 和 1 之间的 NSE 值是可以接受的，而 NSE < 0 表示模型预测比平均观测值更差的不可接受的性能。在这里，NSE值≥0.4被认为是令人满意的，NSE值≥0.7和0.4≤NSE<0.7分别被认为是优异和良好的表现。

**2.4.2 排列重要性**

分析输入变量的相对重要性对于提高黑箱模型的可解释性十分有效。我们采用了置换排列重要性(variable importance,VI)，这是一种与模型无关的解释方法，用于分析输入变量的相对重要性。VI 通过对变量的值进行随机打乱，根据模型误差的增加来评估变量的重要性。对重要变量进行置换会导致模型误差明显增加，而模型误差在响应于不显著变量的置换时保持稳定。VI的算法如下：17

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

其中 X 表示输入变量的矩阵，y 是输出变量的向量， 表示所使用的机器学习模型， 表示评价函数， 是置换模型误差矩阵。与第个变量的相对应的模型误差可以按如下方式计算：

|  |  |  |
| --- | --- | --- |
|  |  | (7) |

其中 是通过置换第 个输入变量得到的矩阵,而 表示对应于 的置换模型误差矩阵。第 个变量的排列重要性VI可以通过计算 和 之间的差值来衡量。如下所示。

|  |  |  |
| --- | --- | --- |
|  |  | (8) |
|  |  |  |

**2.4.3 趋势探究**

Theil-Sen Median方法又被称为 Sen 斜率估计，是一种稳健的非参数统计的趋势计算方法。该方法计算效率高，对于测量误差和离群数据不敏感，常被用于长时间序列数据的趋势分析中。 18 Mann-Kendall（MK）是一种非参数统计检验方法，最初由Mann在1945年提出，后由Kendall和Sneyers进一步完善，其优点是不需要测量值服从正态分布，也不要求趋势是线性的，并且不受缺失值和异常值的影响，在长时间序列数据的趋势显著检验中得到了十分广泛的应用。 19 为了计算时间序列的线性趋势，我们使用非参数的Sen斜率估计器计算Sen趋势值。随后将MK趋势检验应用于时间序列，以统计评估变量随时间变化是否存在单调的上升或下降趋势。

随后，设置情景案例，将输入变量分别去趋势化，探讨不同响应变量对珠江口东侧水域溶解氧长期趋势的影响，如式所示：

|  |  |  |
| --- | --- | --- |
|  |  | (9) |

其中，表示去趋势化后的第j个响应变量，表示未去趋势的原第j个响应变量，表示原第j个响应变量的Sen斜率，表示响应变量向量第个元素。

在每个情景案例中，将去趋势化后的变量取代原有的变量作为输入变量，送入模型训练后，可得到一组模拟的HI输出。

# 三、结果与讨论

## 3.1低氧长期变化模型拟合

**3.1.1 初步预测**

利用多元回归，随机森林，LSTM构建夏季平均HI预测模型，所得模型训练与预测结果如表1所示。

从表一中我们可以看出，训练集精度远优于测试集，说明预测存在过拟合现象。此外，对该时间尺度下所构建的不同机器学习模型在测试集上的表现进行比较，发现随机森林与多元回归相差不大，且均远优于LSTM模型。NSE＜0的模型基本是不可取的 20，说明LSTM不适用于该时间尺度下的数据集。对LSTM而言，通常需要较大的数据量用于训练才能得到较好的结果。 21

综上，我们对溶解氧的分布规律与模型的预测的性能进行了初步的测试，发现LSTM模型不适用于该数据集，因此，在后续讨论中均不再使用。随机森林与多元回归性能相近，但存在过拟合现象，接下来将对其进行进一步优化。

表1 三种机器学习模型初步预测结果

Table 1 Preliminary prediction results of three machine learning models

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NSE | | RMSE | |
| 训练集 | 测试集 | 训练集 | 测试集 |
| 随机森林 | 0.89±0.03 | 0.22±0.33 | 0.86±0.13 | 2.30±0.49 |
| 多元回归 | 0.64±0.09 | 0.24±0.44 | 1.39±0.21 | 2.42±0.61 |
| LSTM | 0.91 | -0.07 | 1.42 | 2.38 |

**3.1.2模型优化**

过拟合现象是指模型对数据进行了过度的训练，使得模型在训练集中表现良好，而在测试集中表现严重不佳。常用解决过拟合的方法包括，增加样本量，正则化，降低模型复杂度等方法。 22此处即以减少模型过拟合为目标，对现有模型进行优化。

（1）模型超参数调整

机器学习中的超参数指的是机器学习模型中使用者预设置的、无法在训练过程中学习的模型参数，模型的性能与超参数直接相关。机器学习中的超参数优化旨在寻找使得机器学习算法在验证数据集上表现性能最佳的超参数，这个过程一般依赖使用者的经验或进行大量参数组合的尝试进行。由于多元回归模型中不存在超参数，因此我们对随机森林模型进行超参数调整。

随机森林模型中所需调整的参数包括，树的数量，每棵树最大的特征数，树的最大生长深度，叶子的最小样本数量，分支节点的最小样本数量等等。 23在本研究中，主要对前两项超参数进行调整，因其余参数主要与样本数相关，而该研究的样本数较小，调整这些参数对结果的影响不大。此外，训练集与测试集的划分的比例也可能影响到预测的精度，因此我们将此比例也作为一项“超参数”进行优化。

基于数据集大小与模型使用经验，我们为随机森林的超参数调整设置了如表所示参数范围。在该参数空间使用网格搜索算法(Grid Search Algorithm)搜索最佳参数组合。遍历所有参数组合后，得到两组超参数组合(8:2, 64, 1)，(7:3, 32, 1)使测试集NSE达到最高，其平均NSE分别为0.3±0.35和0.3±0.23。在上述两组超参数组合中，后者标准差更小，说明训练结果更稳定，因此选择(7:3, 32, 1)作为随机森林模型的最优超参数。

表2 随机森林超参数设置

Table 2 Hyperparameter settings of random forest

|  |  |  |
| --- | --- | --- |
| 超参数 | 范围 | 最优超参数 |
| 训练集测试集比例 | [5:5, 6:4, 7:3, 8:2] | 7:3 |
| 树的数目 | [2, 4, 8, 16, 32, 64, 128] | 32 |
| 树最大特征数 | [1, 2, 3, 4, 5, 6] | 1 |

（2）删除潜在异常值

为进一步提升测试集性能，对测试集中常出现较大偏差的年份进行了统计。如图4所示，对于随机森林和多元回归模型，对2004，2016，2017的HI数值进行模拟时出现较高的偏差。将这些年份的数据删除后，再进行训练与测试，所得模型表现见表3。观察到多元回归与随机森林性能均得到了大幅的提升，多元回归测试集性能由初步预测时NSE = 0.24±0.44提升至0.69±0.18，而随机森林测试集性能由NSE = 0.3±0.23提升至NSE = 0.61±0.22。图5为1000组模型中预测精度最高的一组模型，可更直观展现模型的性能，这组结果也将作为后续探讨驱动变量时的模型基准。如图5所示，多元回归与随机森林于测试集性能均十分优秀，NSE分别为0.95与0.94，其余评估函数数值也令人十分满意。观察到虽然在某些突变点模型仍无法捕捉，但对HI的变化的整体趋势拟合较好。

表3 删除潜在异常值后模型表现

Table 3 The performance of the model with potential outliers deleted

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | NSE | | RMSE | |
| 训练集 | 测试集 | 训练集 | 测试集 |
| 随机森林 | 0.85±0.04 | 0.61±0.22 | 0.58±0.08 | 1.53±0.37 |
| 多元回归 | 0.61±0.22 | 0.69±0.18 | 0.89±0.04 | 1.46±0.27 |

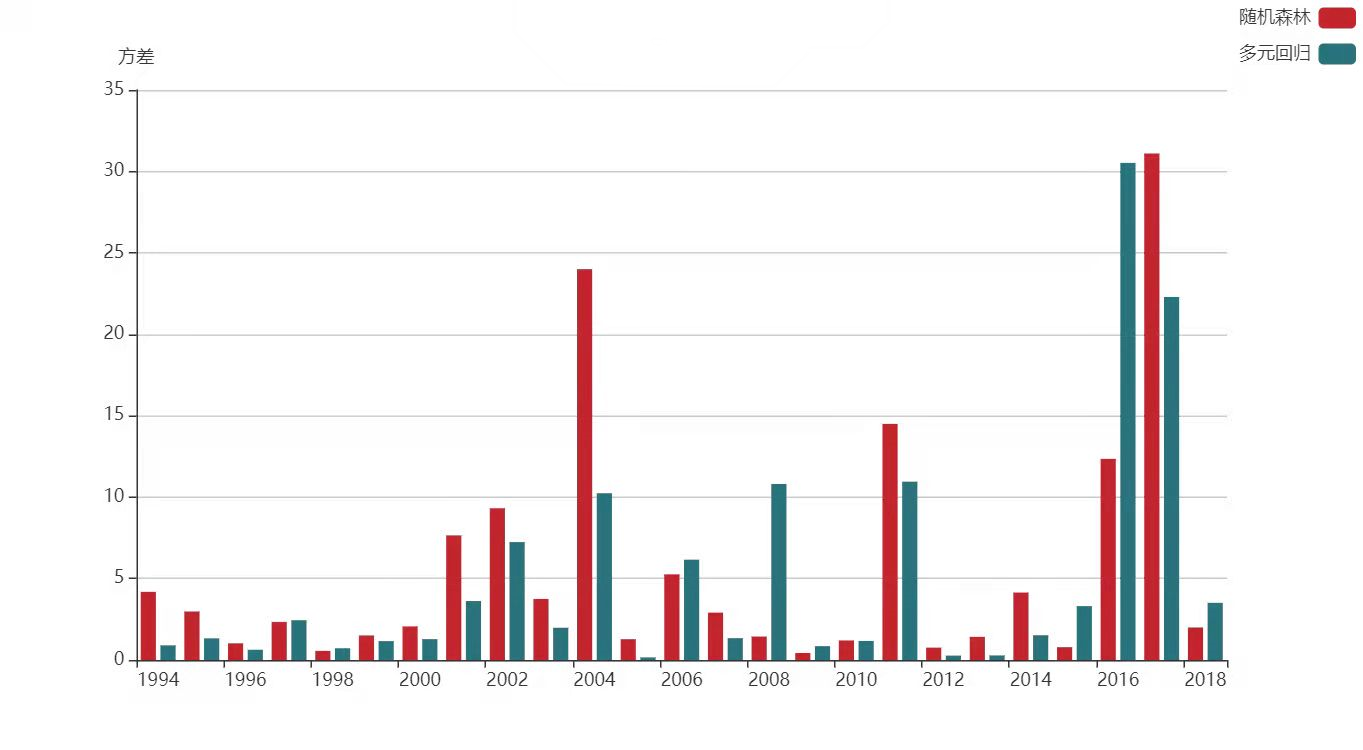


图4 各年份测试集误差统计

Fig.4 Error statistics of the test set in various years

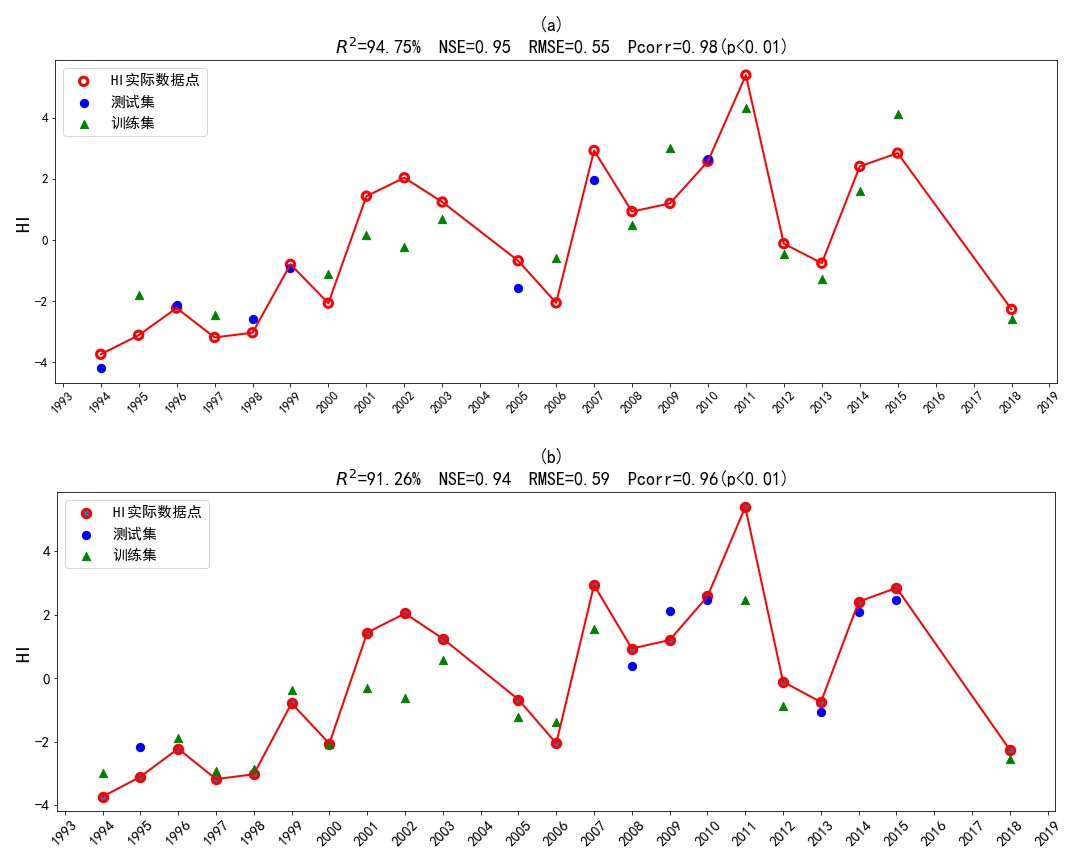


图5 1000组训练模型中测试集精度最高的一组。红线为原始HI的数据分布，散点图为模型预测的结果。（a.多元回归模拟结果 b.随机森林模拟结果）标题为测试集上各评估函数值。

Fig.5 One of the 1000 training models with the highest accuracy on the test set. The red line is the data distribution of the original HI, and the scatter plot is the result of the model prediction. (A. Multivariate regression simulation result b. Random forest simulation result) The title is the evaluation function value on the test set.

## 3.2驱动因素探究

缺氧事件对河口附件的水生生态构成严重威胁，确定引起造成低氧事件的主要原因对水质管理人员至关重要。尽管很重要，但通常被认为是黑箱的机器学习模型并没有揭示输入和输出变量之间的因果机制。相反，基于过程的模型的结果可以根据模型结构中的潜在因果机制直接解释。然而，基于过程的模型在准确预测溶解氧浓度方面受到限制，特别是当未能满足数据要求时。在本研究中，将所构建的机器学习模型与变量排列重要性分析，去趋势化情景模拟等相结合，有助于珠江口东侧水域溶解氧与其响应变量间的因果关系，解决机器学习方法的局限性。其中基于变量排列重要性分析得到的是响应变量对HI长期变化的影响，该变化既包括对年际波动的影响，也包括长期趋势的影响。去趋势化情景模拟则可直接反映各响应变量对HI长期趋势的影响。

**3.2.1 排列重要性分析探究响应变量对低氧长期变化的贡献**

由3.1节的讨论，我们最终选择了删除潜在异常值的夏季平均HI作为模型输入探究驱动变量，随机森林模型已选择最优超参数组合。探究驱动变量的主要方法是基于排列重要性VI。具体实现方法是随机打乱某一个输入变量的值，看它对结果的改变程度。改变大的说明该因素影响大。同样地，为减少选取不确定性带来的偏差，共进行了1000次训练与测试。考虑到当模型的预测精度足够高时，结论才能有效。因此仅选取所有结果中测试集达到满意水平的部分（NSE>0.4）。所得结果如图6所示。

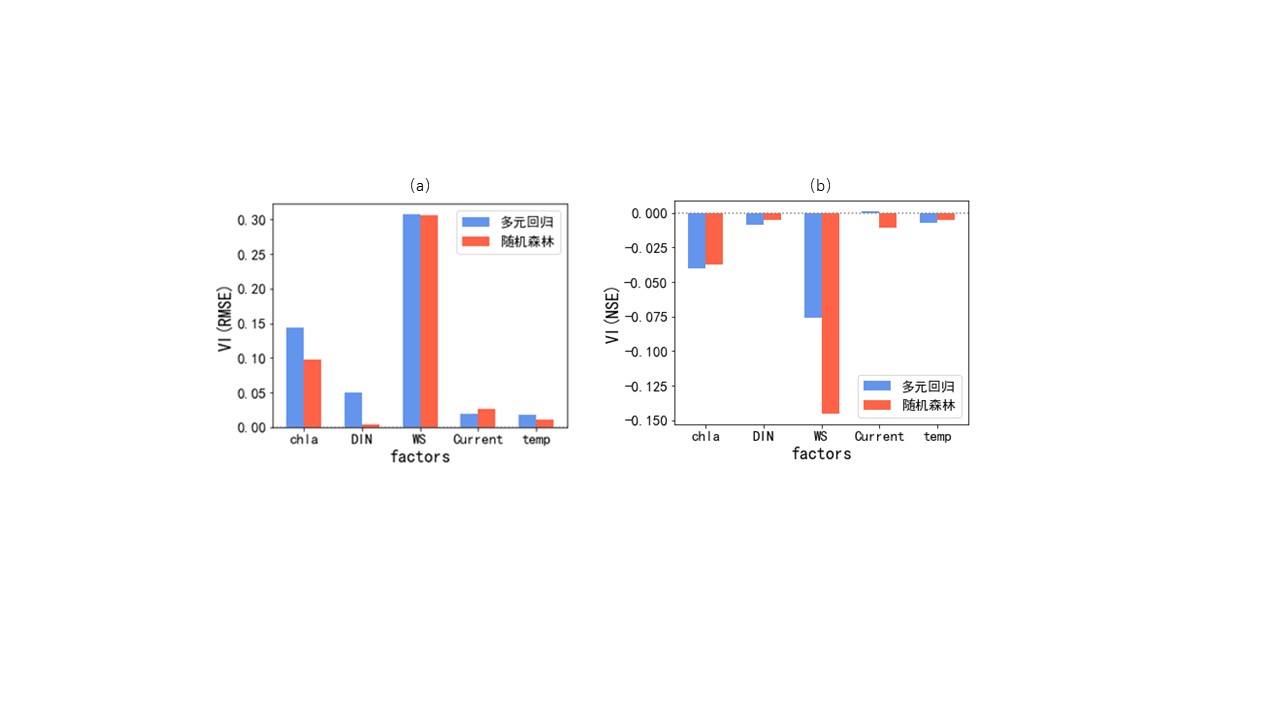


图6 各响应变量排列重要性VI

Fig.6 VI of each response variable

由图可知，无论从NSE还是RMSE出发，对多元回归及随机森林而言，其排列重要性最大的均为风速与表层叶绿素含量，其中WS又大于chla。由此可知，引起珠江口东侧水域溶解氧长期变化的主要因素是WS和chla。基于VI的驱动因素分析是基于数值而非基于过程，机器学习模型的因果解释仍然有限，因此，结果应与既定知识相比加以解释，以避免混淆因素的潜在影响。叶绿素a含量反应水体的富营养化程度，说明近年珠江口附近城市排放的大量污水向沿海海洋提供了过量的营养物质输入刺激了那里的初级生产力，随后底水中有机物的分解消耗了大量的DO，导致缺氧。

如王等人的研究，可以极大地影响珠江对河流输入扰动下缺氧的空间迁移。8而水体的再曝气过程又与夏季西南季风有密切的关联，风速小则使得水体搅拌作用不明显，水体出现分层现象，当底层溶解氧大量消耗时，得不到上层空气的补充，由此导致底层缺氧。该研究结果对于设计有效的管理策略以缓解珠江口水域中的缺氧现象十分有利。

**3.2.2 去趋势化情景模拟探究响应变量对低氧长期趋势的影响**

使用Theil-Sen估计器来计算各输入变量与输出变量的线性趋势。MK检验表明，风速，径流，HI均显示出统计学显著的长期趋势，具有较小的p值。而水温，无机盐含量，叶绿素a未显示显著的长期趋势。如图7所示。为了探讨不同响应变量对珠江口东侧水域溶解氧长期趋势的影响，我们设置了Dtr\_chla，Dtr\_DIN，Dtr\_WS，Dtr\_Current，Dtr\_temp的五个情景模型，其中消除了各变量长期趋势变化的影响。随后，取3.1中预测精度最高的一组随机森林与多元回归模型预测所得HI作为比较基准。比较情景模拟的HI输出与基准HI即得到响应变量各自的长期趋势对溶解氧的影响。以上情景案例同样使用这组模型模拟得到HI输出。如图所示，观察到仅当消除风速长期趋势后能有效缓解HI长期升高的趋势，而消除其他变量的长期趋势对缓解HI长期趋势的影响不大。由此得出结论，对HI在1984~2018年变化的长期趋势而言，风速为推动珠江口夏季长期缺氧的主要因素。

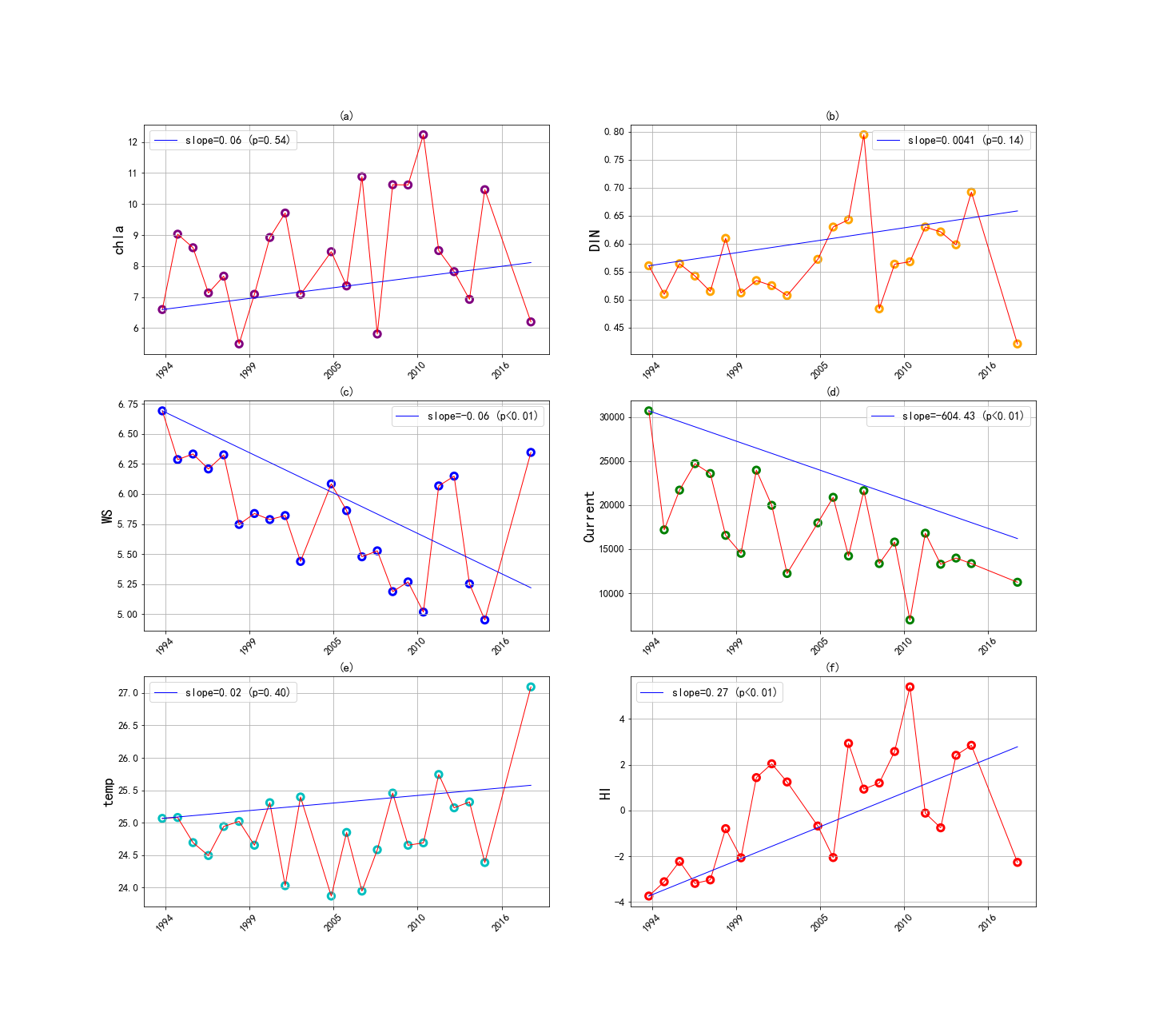


图7 响应变量与解释变量的长期趋势（a.叶绿素a，b.无机盐含量，c.风速，d.径流流量，e.水温，f.原始HI）蓝线表示对数据的线性拟合。

Fig.7 The long-term trends of the response variables and explanatory variables (a. Chlorophyll a, b. Inorganic salt content, c. Wind speed, d. Runoff flow, e. Water temperature, f. Original HI) The blue line represents the linear fit to the data.

## 3.3 不同机器学习性能比较

本研究使用的模型包括随机森林，多元回归，LSTM。在本研究所采用数据上，整体表现随机森林与多元回归相近，多元回归略优，而LSTM预测效果远低于多元回归与随机森林，在后续的诸多讨论中未将其纳入。LSTM是新兴机器学习模型，对时序数据预测具有很大潜力，但其本质也是神经网络模型，且较神经网络模型更为复杂，构建神经网络模型往往需要大量数据进行训练，否则极易预测过拟合。在初步预测的结果中已有所体现。此外，LSTM模型包含极多超参数，包括神经网络层数，各层神经元个数，学习率，训练批次等等，较随机森林与多元回归，模型的训练与计算成本更大。

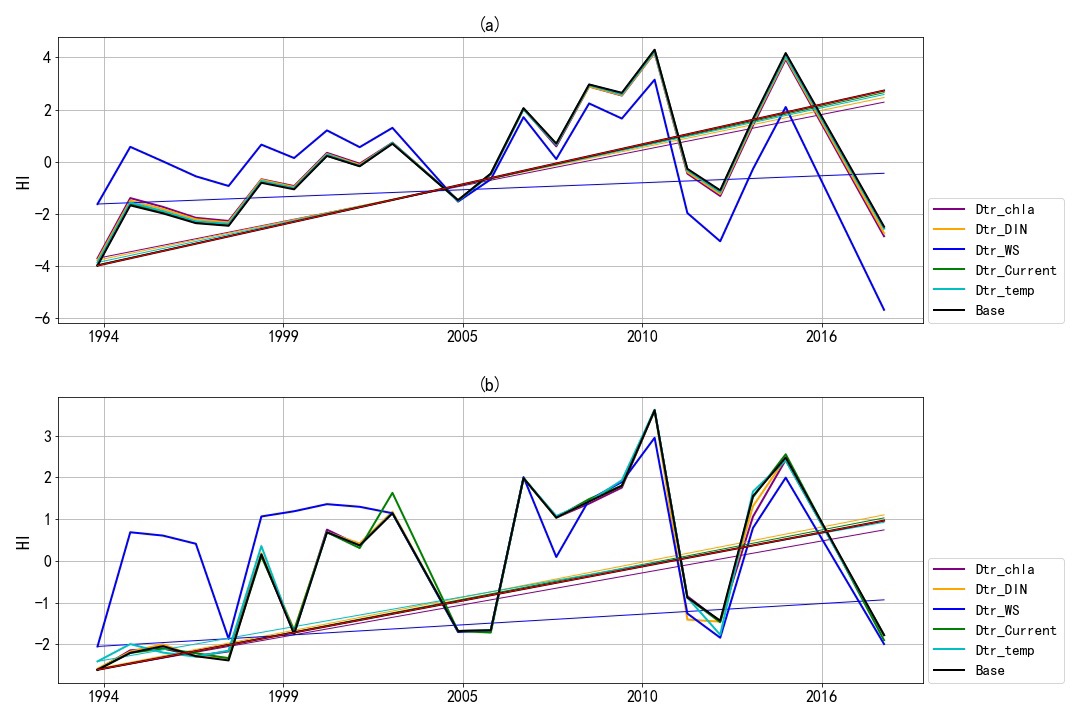


图8. 去趋势化情景模拟（a.多元回归模拟结果 b.随机森林模拟结果）直线为对HI线性拟合

Fig.8 De-trend scenario simulation (a. Multiple linear regression simulation results b. Random forest simulation results) The straight line is a linear fit to HI

随机森林与多元回归的性能相近，但多元回归略优于随机森林，这可能是因为随机森林仍存在过拟合现象。在3.1节讨论中，已进行讨论如何解决预测出现过拟合现象，且基本到达该数据下模型在测试集上性能所达到的上限。多元回归过拟合现象基本解决，而随机森林过拟合现象仍然存在。对机器学习模型而言，出现过拟合往往是因为模型过度的训练，使得模型训练出一些数据本身所不具备的特征。解决这种问题主要有两种思路，一是从数据出发，通过降噪，增加样本数丰富样本信息，使模型更复杂。二是从模型出发，对超参数进行优化，如对随机森林通过减少预测树的深度和每层树的特征数量，使模型更简单，或者干脆直接使用更简单的模型。在3.1讨论中基本已涉及，但随机森林仍出现过拟合，或许是对所预测的25个样本点数据集而言，随机森林仍是有些复杂。目前的机器学习模型主要是面向大数据，尤以各类神经网络为代表的深度学习模型往往十分复杂，对样本源的数量有极大的需求。该研究从另一方面验证了机器学习模型中“没有免费的午餐”定理，对不同的数据源往往要使用最恰当的模型才能得到最精确的结果，对只包含少量数据的数据集而言，使用更为简单的模型往往能得到更好的结果。

# 四、总结与展望

本研究基于香港环保署所观测的水文数据，使用多元回归和随机森林等机器学习模型近似模拟了珠江口东侧水域的溶解氧长期变化趋势，并根据所建立的模型定量地探究其变化的驱动因素。初步建立模型时，LSTM模型因其精度过低被淘汰，多元回归与随机森林则出现较严重过拟合现象。通过调参，删除潜在异常值后，对模型进行一千次训练与测试，在所构建的各1000组模型中，多元回归于测试集上的NSE平均可达0.68，最高可达0.95，随机森林于测试集上的NSE平均可达0.61，最高可达0.94。由此进一步借助统计方法排列重要性计算得到各变量的相对贡献率。研究结果表明，风速，表层叶绿素含量是造成珠江口东侧水域溶解氧长期变化的主要因素，而水温，无机氮含量，径流流量的贡献相对较小。其中，风速越小，表层叶绿素含量越大，低氧现象越严重。总的来说，研究表明，自然和人为过程都是导致珠江口缺氧的重要原因。随后，通过情景案例探究不同变量对HI长期趋势的影响时，得到风速是影响珠江口东侧长期缺氧趋势的主要因素。

在数据源充足的情况下，未来可将机器学习与水质模型结合，一方面可在时间领域作拓展，结合小波分析等手段探究引起珠江口东侧水域年际波动规律及其影响因素，评估其对生态系统的影响并针对性提出解决对策。另一方面，本研究是基于外围水域从面域角度进行探讨，进一步地可就某一区域内的站点水质情况进行研究。此外，有待探究其他环境因素如水体营养盐总量、有机质含量等的影响，进一步优化机器学习预测模型。

**参考文献：**

(1) Li, X.; Bianchi, T. S.; Yang, Z.; Osterman, L. E.; Allison, M. A.; DiMarco, S. F.; Yang, G. Historical trends of hypoxia in Changjiang River estuary: Applications of chemical biomarkers and microfossils. *Journal of Marine Systems* **2011**, *86*, 57-68.

(2) Ni, W.; Li, M.; Testa, J. M. Discerning effects of warming, sea level rise and nutrient management on long-term hypoxia trends in Chesapeake Bay. *Sci Total Environ* **2020**, *737*, 139717.

(3) Hu, J.; Zhang, Z.; Wang, B.; Huang, J. Long-term spatiotemporal variations in and expansion of low-oxygen conditions in the Pearl River estuary: a study synthesizing observations during 1976–2017. *Biogeosciences* **2021**, *18*, 5247-5264.

(4) Strokal, M.; Kroeze, C.; Li, L.; Luan, S.; Wang, H.; Yang, S.; Zhang, Y. Increasing dissolved nitrogen and phosphorus export by the Pearl River (Zhujiang): a modeling approach at the sub-basin scale to assess effective nutrient management. *Biogeochemistry* **2015**, *125*, 221-242.

(5) Levin, L. A.; Ekau, W.; Gooday, A. J.; Jorissen, F.; Middelburg, J. J.; Naqvi, W.; Neira, C.; Rabalais, N. N.; Zhang, J.

(6) Li, Y.; Song, G.; Massicotte, P.; Yang, F.; Li, R.; Xie, H. Distribution, seasonality, and fluxes of dissolved organic matter in the Pearl River (Zhujiang) estuary, China. *Biogeosciences* **2019**, *16*, 2751-2770.

(7) Xu, C.; Xu, Y. J.; Hu, J. T.; Li, S. Y.; Wang, B. A numerical analysis of the summertime Pearl River plume from 1999 to 2010: Dispersal patterns and intraseasonal variability. *Journal of Marine Systems* **2019**, *192*, 15-27.

(8) Wang, B.; Hu, J.; Li, S.; Yu, L.; Huang, J. Impacts of anthropogenic inputs on hypoxia and oxygen dynamics in the Pearl River estuary. *Biogeosciences* **2018**, *15*, 6105-6125.

(9) Jiang, L.; Li, Y.; Zhao, X.; Tillotson, M. R.; Wang, W.; Zhang, S.; Sarpong, L.; Asmaa, Q.; Pan, B. Parameter uncertainty and sensitivity analysis of water quality model in Lake Taihu, China. *Ecological Modelling* **2018**, *375*, 1-12.

(10) Iverson, L. R.; Prasad, A. M.; Matthews, S. N.; Peters, M. Estimating potential habitat for 134 eastern US tree species under six climate scenarios. *Forest Ecology and Management* **2008**, *254*, 390-406.

(11) Liang, Z.; Zou, R.; Chen, X.; Ren, T.; Su, H.; Liu, Y. Simulate the forecast capacity of a complicated water quality model using the long short-term memory approach. *Journal of Hydrology* **2020**, *581*.

(12) Palani, S.; Liong, S.-Y.; Tkalich, P. An ANN application for water quality forecasting. *Marine Pollution Bulletin* **2008**, *56*, 1586-1597.

(13) Reichstein, M.; Camps-Valls, G.; Stevens, B.; Jung, M.; Denzler, J.; Carvalhais, N.; Prabhat. Deep learning and process understanding for data-driven Earth system science. *Nature* **2019**, *566*, 195-204.

(14) Zhang, J.; Zhu, Y.; Zhang, X.; Ye, M.; Yang, J. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas. *Journal of Hydrology* **2018**, *561*, 918-929.

(15) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2016**, *374*.

(16) Wang, K.; Cai, W.-J.; Chen, J.; Kirchman, D.; Wang, B.; Fan, W.; Huang, D. Climate and Human-Driven Variability of Summer Hypoxia on a Large River-Dominated Shelf as Revealed by a Hypoxia Index. *Frontiers in Marine Science* **2021**, *8*.

(17) Fisher, A.; Rudin, C.; Dominici, F. All Models are Wrong, but Many are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. *J Mach Learn Res* **2019**, *20*.

(18) Sen, P. K. Estimates of the Regression Coefficient Based on Kendall's Tau. *Journal of the American Statistical Association* **1968**, *63*, 1379-1389.

(19) Gilbert, R. O.: *Statistical Methods for Environmental Pollution Monitoring*: United States, 1987.

(20) Zhi, W.; Feng, D.; Tsai, W. P.; Sterle, G.; Harpold, A.; Shen, C.; Li, L. From Hydrometeorology to River Water Quality: Can a Deep Learning Model Predict Dissolved Oxygen at the Continental Scale? *Environ Sci Technol* **2021**, *55*, 2357-2368.

(21) Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput* **1997**, *9*, 1735-1780.

(22) Srivastava, N.; Hinton, G.; Krizhevsky, A.; Sutskever, I.; Salakhutdinov, R. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research* **2014**, *15*, 1929-1958.

(23) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5-32.