DOI: 10.13758/j.cnki.tr.2023.04.001
高志炜, 吴电明, 陈曦, 等. 机器学习在氮循环领域的应用研究进展. 土壤, 2023, 55(4): 689–698.

机器学习在氮循环领域的应用研究进展①

高志炜[1,2,3], 吴电明[1,2,3,4*], 陈 曦[1,2,3], 潘月鹏[4]

(1 华东师范大学地理科学学院, 地理信息科学教育部重点实验室, 上海 200241; 2 崇明生态研究院, 上海 202162; 3 自然资源部超大城市自然资源时空大数据分析应用重点实验室, 上海 200241; 4 中国科学院大气物理研究所大气边界层物理和大气化学国家重点实验室, 北京 100029)

摘 要: 氮循环是地球圈层中水–土–气–生多介质、多界面的复杂过程, 与土壤健康、粮食安全、全球变暖、空气污染、水体质量等环境问题密切相关。近年来, 得益于计算机技术的快速发展和海量、多源数据的产生, 机器学习迅速成为研究氮素循环强有力的工具。本文系统梳理了机器学习的功能性概念, 包括典型开发流程和学习应用场景等; 总结了机器学习的典型应用算法, 包括经典机器学习(如随机森林、支持向量机等)和深度学习(如卷积神经网络、长短期记忆网络等); 并综述了机器学习在氮循环研究领域的应用研究进展, 包括大气、水体、土壤和植物/作物等介质的氮素代谢机制、模拟氮素循环过程及管理氮素流动等。未来基于大数据和机器学习技术的特征工程和模型融合的研究, 将会给氮循环领域的数据分析与建模带来巨大变革。同时, 将机器学习与基于物理过程的模型相结合解决氮循环过程中的复杂问题, 可为服务国家“双碳”战略以及控制全球变暖、空气污染等环境问题提供重要支撑。

关键词: 机器学习; 深度学习; 氮循环; 硝化; 反硝化; 氧化亚氮

中图分类号: S154.1; TP181 **文献标志码:** A

Machine Learning in Nitrogen Cycle Research: A review

GAO Zhiwei[1,2,3], WU Dianming[1,2,3,4*], CHEN Xi[1,2,3], PAN Yuepeng[4]

(1 *School of Geographical Sciences*, *East China Normal University*, *Key Laboratory of Geographic Information Sciences*, *Ministry of Education*, *Shanghai* 200241, *China*; 2 *Institute of Eco-Chongming* (*IEC*), *Shanghai* 202162, *China*; 3 *Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities*, *Ministry of Natural Resources*, *Shanghai* 200241, *China*; 4 *State Key Laboratory of Atmospheric Boundary Physics and Atmospheric Chemistry*, *Institute of Atmospheric Physics*, *Chinese Academy of Sciences*, *Beijing* 100029, *China*)

Abstract: Nitrogen cycle is a complex process of multi-media and multi-interface between water-soil-atmosphere-biology in the Earth's sphere, which is closely related to environmental problems such as soil health, food security, global warming, air pollution and water quality. With the rapid development of computer technology and the generation of massive and multi-source data in recent years, machine learning (ML) has rapidly become a powerful tool to study nitrogen cycle. This paper first introduces the functional concepts of ML, including typical development process and learning application scenarios. Then typical application algorithms of ML are summarized, including classical ML (such as random forest, support vector machine, etc.) and deep learning (such as convolutional neural network, long-term and short-term memory network, etc.). In addition, the application research progress of ML in the field of nitrogepn cycle research are reviewed, including nitrogen metabolism mechanism, simulating nitrogen cycle process and managing nitrogen flow in atmosphere, water, soil and plant/crop. In the future, the research of feature engineering and model fusion based on big data and ML technology will bring great changes to data analysis and modeling in the field of nitrogen cycle. Meanwhile, combine ML with process-based models to solve complex problems in the nitrogen cycle, which will provide important support for serving the national “double carbon” strategy and controlling global warming, air pollution and other environmental issues.

Key words: Machine learning (ML); Deep learning; Nitrogen cycle; Nitrification; Denitrification; Nitrous oxide

①基金项目: 上海市 2022 年度科技创新行动计划长三角科技创新共同体领域项目(22002400300), LAPC 国家重点实验室开放课题(LAPC-KF-2022-09)和中央引导地方科技发展资金项目(2021ZY0002)资助。

* 通讯作者(dmwu@geo.ecnu.edu.cn)

作者简介: 高志炜(1999—), 女, 山东德州人, 硕士研究生, 主要从事城市环境氮循环研究。E-mail: 51213901022@stu.ecnu.edu.cn

氮(N)是生命代谢必需的营养元素，参与蛋白质合成、信号调节等基础生理功能[1]。空气中 78% 的气体是氮气(N_2)，经由闪电作用、生物固氮、人工合成氨等途径形成活性氮，从而进入陆地、海洋、大气等圈层，参与氮素循环。自然生态系统一般处于“氮限制”的状态，少量的氮沉降、施肥等氮素输入可以促进生态系统生产力的提高[2]。但是，由于人口数量的不断增加和对粮食产量的需求，大量的氮肥被用于提高作物、森林树木和草场等产量，导致过量的氮素进入生态系统，超过了地球系统的行星边界(planetary boundary layer)，成为继生物多样性之后的又一全球性问题[3]。据估算，农田生态系统作物的氮素利用率只有 20% ~ 50%[4]，其余的氮素一部分被保留在土壤中，一部分通过硝态氮淋失、气体排放等进入水体和大气，引发了一系列的环境问题，包括土壤酸化、面源污染、大气污染、臭氧层空洞、生物多样性降低等[5-6]。

为了研究多介质氮素迁移转化过程及其环境效应，国内外学者已经发展了多种模型，主要包括自下而上(bottom-up)和自上而下(top-down)两种类型。前者包括排放清单估算、基于物理过程的模型预测等，决定其预测能力的关键因素在于数据量的大小、氮素循环机理的研究等，如土壤数据库的建立、氮循环的关键功能基因和驱动因素的解析等[7-9]；后者包括遥感定量反演等，其模型的准确度更依赖于仪器精度、大气条件和后期数据分析等因素[10]。虽然这些方法能够精确地定量评估氮素循环过程、驱动机制和环境影响等，对解决氮循环复合型问题起到了决定性的作用，但也存在一些缺陷，如自下而上的方法非常考验研究者的先验知识，自上而下的方法面对存在缺失值的时间序列的建模能力仍然有限[11]，且两者在拟合非线性关系的精度上也有待提高以及运行成本高昂等[12]。

机器学习(Machine Learning，ML)是近年来迅速发展起来的人工智能中的一个新兴领域，为科学家研究氮素代谢、循环和利用等提供了全新的视角。机器学习已被广泛地应用于土壤学[13]、大气科学[14]、环境科学[15]、水文学[16]和生物信息学[17]等多学科交叉研究领域。氮素循环作为生物地球化学循环的重要组成部分，也已经涌现出大量机器学习的应用性文章。如图 1 所示，关于机器学习和氮素交叉研究的论文数量呈现逐年增长的趋势，2010 年以来更是飞速发展。

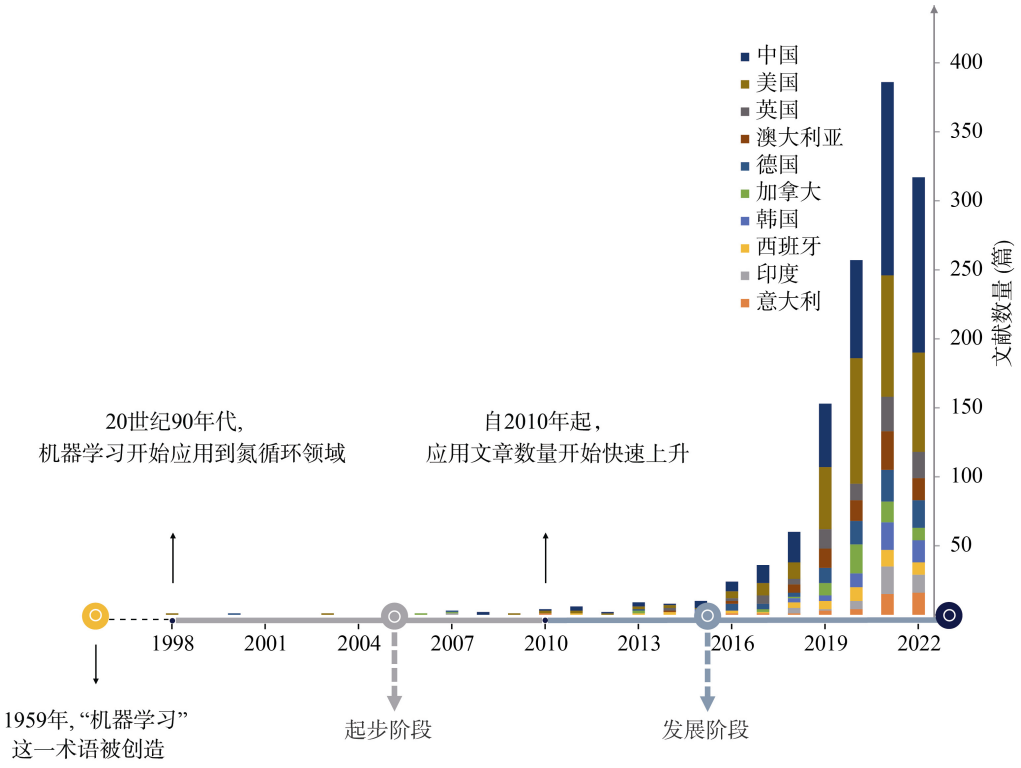


图 1　基于 Web of Science 以“机器学习”和“氮”为关键词搜索得到的世界各国已发表论文的数量(访问日期 2022 年 7 月 15 日)

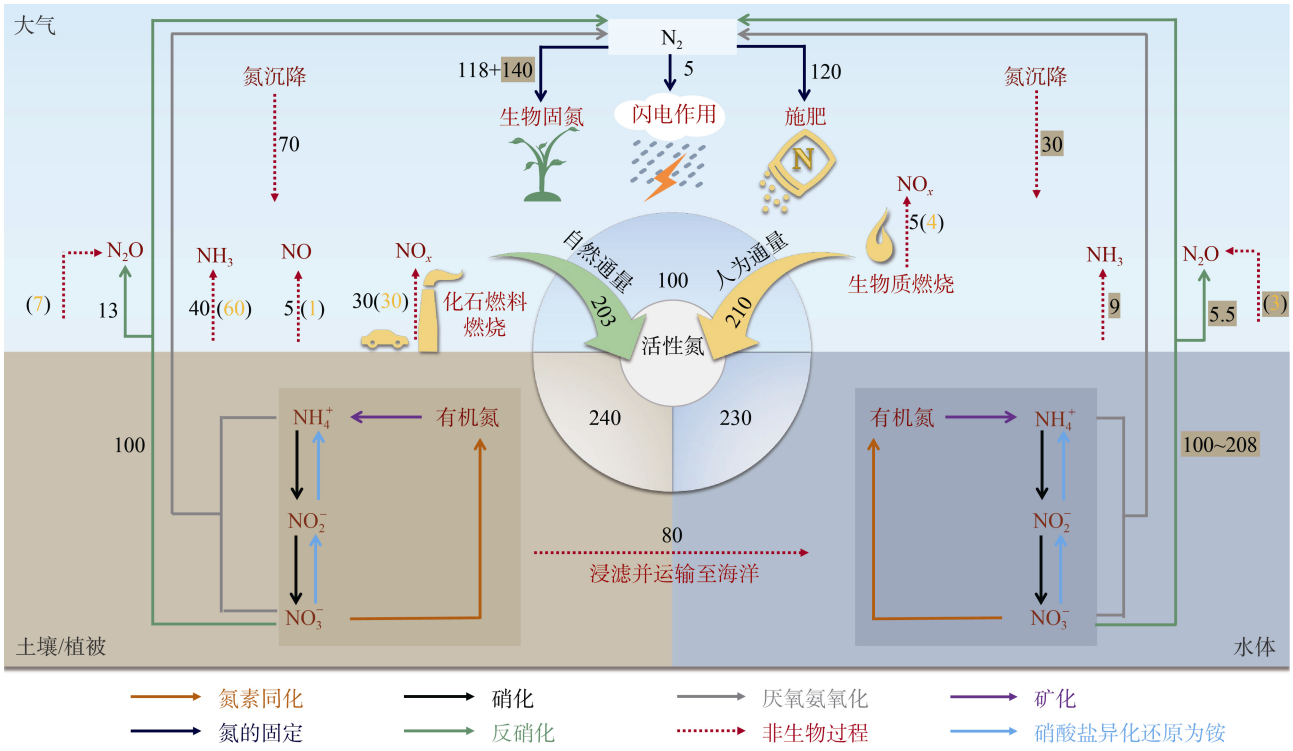
Fig. 1　Numbers of papers published from various countries with keyword “machine learning” and “nitrogen” based on Web of Science (access date 7/15/2022)

单一的、集成的或与基于物理过程模型混合建模的机器学习算法与氮素经典的研究手段结合被应用于生态系统氮素循环各个时空尺度的研究[18-19]。机器学习凭借更灵活的模型结构和更高的计算效率，能够定量构建社会、经济、环境要素到氮素浓度变化的动态响应关系中，从不断增长的地理空间数据流中提取模式和见解，提高季节性预测的预测能力，进行跨多个时间尺度的远程空间联系建模，从而获得对氮素科学问题的进一步理解。混合建模的方法也使得机器学习和物理过程模型各自在已经获得较好预测结果的情况下更进一步赋予和增添了彼此的竞争优势[20]。机器学习还在降低研究成本，宏观、全面、快速预测土壤中氮的流入、流出和转化过程，了解大尺度全球氮素通量及其空间分布，大大降低全球氮素收支的不确定性等方面发挥重要作用，迅速成为研究氮素循环强有力的工具[15]。

本文综述了机器学习在氮循环领域的研究进展和应用情况，比较了其与经典研究方法的优劣，提出了未来应该关注的研究方向，以期为推动氮素循环研究、解决氮素相关的环境问题等提供科学支撑，也为政府部门决策、联合国政府间气候变化专门委员会(IPCC)气候变化评估、实现联合国可持续发展目标等提供参考和建议。

1　氮循环

氮素生物地球化学循环主要由微生物参与的氧化还原反应驱动。固氮作用、好氧硝化、厌氧反硝化、厌氧氨氧化等多种过程驱使地球上不同价态或相态的氮素保持动态平衡[21]。空气中的惰性 N_2 是可自由获取的氮的最大库存，但由于其三键结构的高度化学稳定性，很难被生物直接利用，需要经过一系列氮转化过程，形成如铵盐(NH_4^+)和硝酸盐(NO_3^-)才能为生物所吸收[22]。通过生物固氮和闪电作用每年约 203 Tg 的 N_2 转化为活性氮，进入陆地和海洋生态系统[23](图 2)。大部分 N_2 被还原为铵化合物，随后在硝化作用下，NH_4^+被逐步氧化成 NO_3^-，并通过土壤、沉积物、淡水和海水的微生物反硝化、化学作用等以 N_2 的形式重新返回大气，构成氮的循环过程。同时，厌氧氨氧化微生物以亚硝酸盐(NO_2^-)为电子受体，将 NH_4^+氧化为 N_2，也起到脱氮作用[24]，因此，该过程经常和反硝化作用一起被应用到废水处理厂的脱氮工艺中。而硝酸盐异化还原为铵(DNRA)会与反硝化微生物竞



(图中圆环上的数字代表了三大系统的活性氮分配；各个细线箭头旁的数字表示氮素迁移转化的通量(N，Tg/a)，其中，黑色数字代表自然通量，全部自然通量 203 Tg/a 被汇入生态系统活性氮库；黄色数字代表人为排放通量，全部人为排放通量 210 Tg/a；加底纹的数字代表参与水体生态系统氮素循环过程的活性氮通量。数据来源于 Fowler 等[23]的文献)

图 2　生态系统中氮循环关键过程示意图

Fig. 2　Key processes of nitrogen cycle in ecosystems

争 NO_3^-和有机物，将 NO_3^- 还原为 NO_2^- 和 NH_4^+，再次将固定的氮回收利用[25]。土壤中超过 90% 的氮素以有机态形式存在，难以被植物利用[26]。矿化作用将土壤中有机态氮在微生物的作用下转化为易被植物吸收的无机氮(如 NH_4^+、NO_3^-)[27]，再经过氮素同化最终合成氨基酸和蛋白质，因此，该过程与作物产量和氮素利用效率等密切相关[28]。

在氮素转化过程中，大量的活性氮被释放到环境中，直接影响气候变化、空气污染和水体质量等。例如，硝化和反硝化作用产生的氧化亚氮(N_2O)是一种重要的温室气体，在地球的辐射平衡和平流层臭氧(O_3)循环中起着关键作用[29]。而通过氨挥发产生的氨气(NH_3)，以及硝化和反硝化过程排放的氮氧化物(NO_x)和气态亚硝酸(HONO)等是典型的空气污染物[30]，参与近地面 O_3 和氢氧自由基(·OH)的生产和消耗、挥发性有机化合物(VOCs)的循环等过程，在自由基光化学和大气氧化能力等方面起着关键作用[23]。这些短寿命活性氮气体(NH_3、NO_x 和 HONO)可以转化为 NO_3^- 或 NH_4^+，是形成气溶胶的重要前体物，影响着大气细颗粒物浓度($PM_{2.5}$)和空气质量[31]。大气干湿沉降可以移除空气中的活性氮，连同氮肥(主要是铵态氮和硝态氮)的输入，再次进入陆地或海洋生态系统的氮循环(图 2)。

2 机器学习

机器学习是实现人工智能的一种方法，是一门跨学科的学科，通过结合概率论、统计学等数学方法，从已知数据中模拟或实现，从已有数据中挖掘规则，从而实现对未知数据的“预测”[32]。机器学习发展到今日，已经积累了大量的算法。一般根据学习方式分为监督学习、非监督学习和强化学习。监督学习需要对每一个数据样本有明确标注，常应用于分类和回归问题，常见的算法有贝叶斯分类器、逻辑回归、随机森林、支持向量机、卷积神经网络等[33]。非监督学习算法的主要任务是在不对数据做任何标注情况下发现数据的分布规律，常应用于关联规则的学习以及聚类，常见的算法包括 K-Means 聚类、层次聚类、自组织映射等[33]。强化学习通过与外部环境交互获得的反馈中学习，常见的应用场景包括动态系统以及机器人控制等，常见的算法包括 Q-Learning 等[34]。

机器学习模型的开发遵循收集数据、处理数据、建立模型、训练和验证模型以及测试模型性能的系统步骤[15](图 3)。数据处理包括①数据清洗，识别“脏数据”：对缺失数据、异常数据和重复数据进行删除、填充和纠正等；②数理统计分析：对数据进行标准化或正态化处理；③数据挖掘：针对高维数据进行降维，或为了避免多重共线性进行特征提取。开发模型也是一项复杂的任务，首先将处理后的数据进行分组，分为训练集、验证集和测试集。训练集用于模型拟合；验证集用于调整模型的超参数，初步评估模型的能力；测试集用于评估最终模型的泛化能力和性能表现。模型的准确性通常根据不同的任务选择不同的准则。分类任务通常采用极大似然准则，回归任务通常采用均方误差准则。预测问题通常属于回归任务，常用的指标有决定系数(R^2)、卡方(χ^2)、平均偏差误差(MBE)、均方误差(MSE)、均方根误差(RMSE)、平均百分比误差(MPE)等[35]。

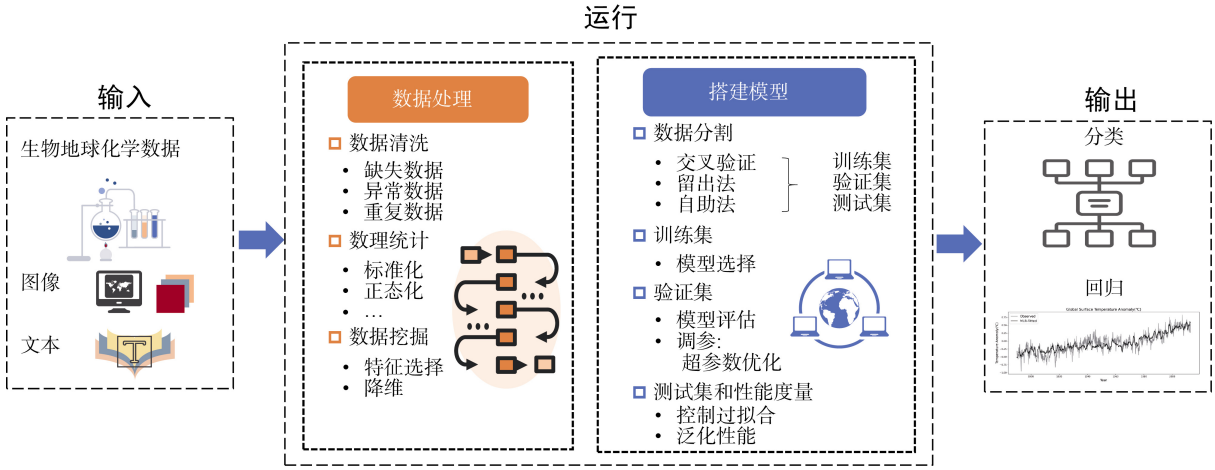


图 3 机器学习模型开发的典型工作流程

Fig. 3 Typical workflow for developing machine learning models

MATLAB 机器学习工具箱、R 的“程序包”、Python 的 scikit-learn 以及开源的算法等为非机器学习领域的从业人员搭建了能在其专业领域应用的桥梁。氮循环领域的输入变量主要来自于高光谱图像数

据、生物地球化学实验室模拟、外场测量数据和文本数据等[36]，非常规数据源还有智能手机等[37]。当输入变量很少时，通过统计学方法或研究人员的先验知识，可以筛选出变量的最佳集合，确保模型的准确性，并使模型具有可解释性。随着研究区的扩大，研究内容的复杂化，将会产生更高维数据集，评估所有变量的重要性将变得难以实现。虽然模型输入变量的增多能提供更高的准确性，但同时会降低模型的可解释性，并导致多重共线性[38]。因此，机器学习提供了特征选择技术以消除输入变量的多重共线性，包括粒子群优化、遗传算法(GA)、混合 GA–人工神经网络、平行 GA、人工蜂群算法等[39]。为了探索特征选择的数据集是否稳健，可利用重复的敏感性分析观察在不同输入下输出的波动范围，从而对输入进行取舍增减，进一步保证模型输入数据集的优质性[40]。一旦成功地构建模型，就能将其用于特定问题的预测，但此时它们仅适用于开发它们的数据范围或特定问题，想要实现模型的外推仍然需要新数据集的重新训练。迁移学习提供了模型的可移植性解决方案，它可以有机地利用源域中的知识对目标域更好地建模[41]。

3　机器学习典型应用算法

机器学习方法(例如随机森林、支持向量机和神经网络)应用非常广泛，其性能和适用性普遍优于更简单的方法，例如主成分回归、偏最小二乘回归、多元线性回归和 K 最近邻算法等[13, 38]。本文主要介绍代表性的经典机器学习方法以及深度学习进阶算法(图 4)。

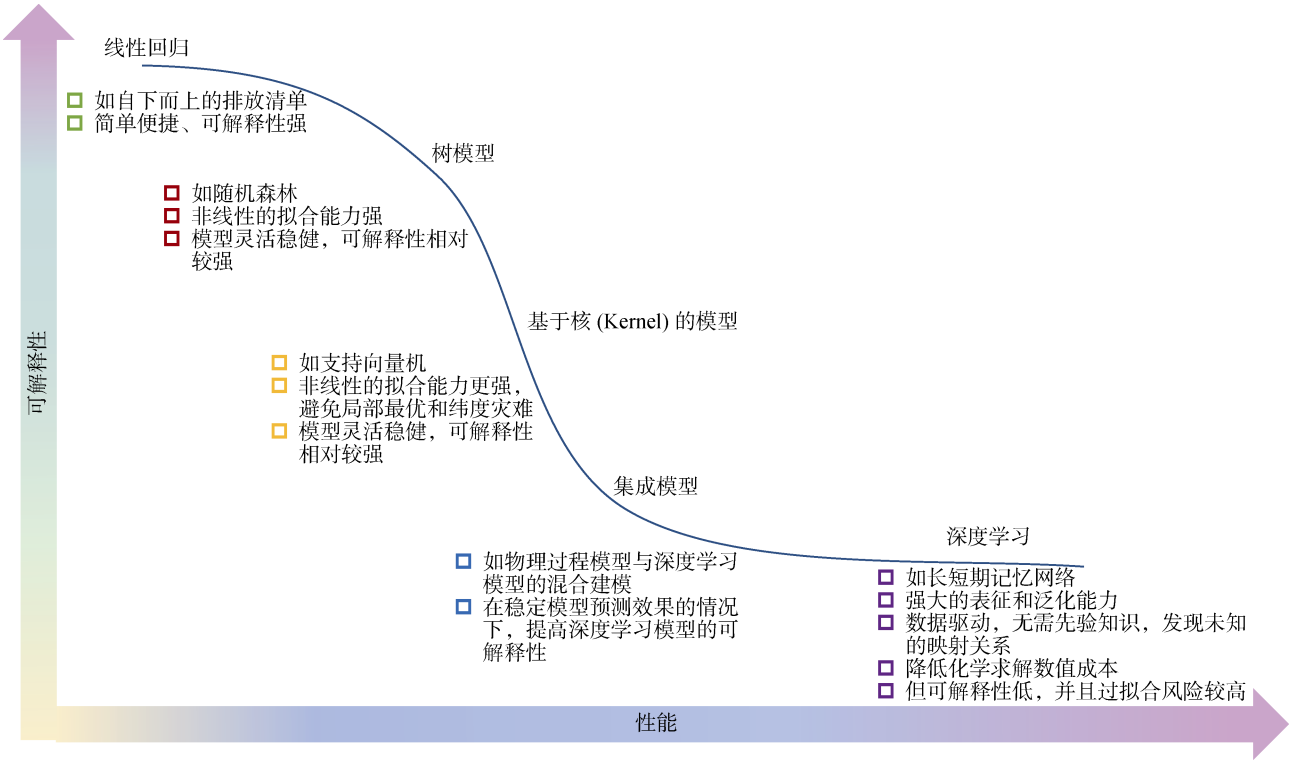


图 4　机器学习各类算法性能和可解释性之间的权衡以及各自优势[38]

Fig. 4　Trade-off between performance and interpretability for machine learning algorithms and their respective advantages

3.1　经典机器学习

3.1.1　随机森林　　随机森林(Random Forest，RF)是 Breiman 开发的一种基于若干决策树的集成学习算法。“随机”是它的精髓，主要体现为训练集随机抽样以及特征子集随机生成。正是因为这两个“随机”的引入，才较好地提升了它的泛化和抗噪能力，使其不易陷入单一决策树引发的过拟合问题。迄今为止，RF 在全球生态系统中氮素转化通量预测[42]、活性氮物种时空建模[43]、土壤/水体氮素浓度卫星反演评估[44-45]等方面都表现出了强大的应用潜力。

3.1.2　支持向量机　　支持向量机(Support Vector Machine，SVM)是一种用于二元分类的广义分类器，使用核技巧以及定义软间隔最大化，旨在维度空间中找到一个正确分类的最优决策面[46]。相比于容易过度拟合训练样本和基于贪心学习的策略来搜索假设空间的人工神经网络，SVM 以结构风险最小化(SRM)原理代替经验风险最小化(ERM)原理，大大降低了过拟合风险，并以凸优化的本质和核函数的使用有效避

免局部最优和“维度灾难”[47]。SVM 有 4 种核函数类型：线性、多项式、S 型(Sigmoid)和径向基函数(RBF)。核函数及其参数的选择影响 SVM 模型分析结果的准确性。RBF 核，有的时候也被称为高斯核(Gaussian Kernel)，广泛应用于土壤测绘数字制图[48]、水质监测和废水处理[49]、生态化学计量[50]等方面。

3.2 深度学习

深度学习或深度神经网络是指具有多层的人工神经网络(Artificial Neural Network，ANN)。ANN 最早由 Warren McCulloch 和 Walter Pitts 提出，很多理论分析和前瞻性成果在 20 世纪五六十年代相继出现。比如 1958 年心理学家 Rosenblatt 创造的感知机(Perceptron)，但由于感知机存在无法完成多种模式的训练识别的缺陷使得研究陷入冰冻期[51]。经过多层神经网络和反向传播算法的提出及应用，神经网络于 20 世纪 80 年代迎来了第二次研究热潮[52]。随着网络层数的递增，反向传播算法容易产生梯度消失或者爆炸的问题；另一方面，其他诸如 SVM 等新兴算法又可以在更少的硬件资源条件下达到很好的效果，因此神经网络复归沉寂[53]。但 21 世纪以来，随着计算能力和训练数据规模的大幅度提升，特别是云计算、高性能 GPU 硬件设备的广泛应用，使得神经网络研究再次复苏，迎来第三次繁荣。

深度学习也是机器学习的最蓬勃发展的分支[54]，并得益于计算机技术的快速发展以及海量数据的不断积累，其在医学、物理学、化学等领域大放异彩并启发了地学的发展。它可以在不依赖于先验知识的情况下完全由数据驱动，不断从增长的地理空间数据流中提取模式和见解，从而成为地理建模的新方法[20]。在地球系统大数据背景下，深度学习算法(例如卷积神经网络和长短期记忆网络)结合地理信息系统(GIS)和遥感(RS)或利用谷歌地球引擎(GEE)等云计算平台链接，通过编译大量数据进行环境资源监测、土地覆盖测绘和信息建设与预测，辅助决策者进行氮素管理。

3.2.1 卷积神经网络　卷积神经网络(Convolutional Neural Network，CNN)以生物视觉感知机制为灵感，是一种著名的深度学习架构，在计算机视觉领域取得了令人瞩目的成就。1990 年，Le Cun 等[55]发表了建立 CNN 现代框架的开创性论文。自 2012 年以来，CNN 逐渐成为图像分类、对象检测、语义分割等视觉识别任务的主流算法[56]。

3.2.2 长短期记忆网络　长短期记忆网络(Long Short-Term Memory，LSTM)属于一种特殊的循环神经网络(RNN)，擅长处理非线性时间序列数据。最初版本是由 Sepp Hochreiter 和 Jürgen Schmidhuber 在 1997 年提出，用于克服 RNN 在学习长期依赖项时通常出现的梯度爆炸/消失问题[57]。在地球系统科学背景下，通过建立气候和遥感协变量与目标变量(陆地、海洋和大气)相关联的 LSTM 模型，可推断大陆或全球估计值[58]。

4 机器学习在氮循环领域的应用研究进展

机器学习的应用场景非常广泛，包括文本处理、图像识别、数据挖掘等[35]。近年来，在地球科学领域涌现出了各类跨学科和应用型研究。例如，Reichstein 等[20]给出了机器学习的典型地学研究应用场景，包括分类问题、融合问题、预测问题、时间序列建模问题等。在氮素循环领域，机器学习已经承接室内培养试验、田间野外试验、大气外场观测、遥感观测和分子生物学手段产生的高通量数据流，在水–土–气–生多介质、多界面上，进行了各个研究尺度上的模型模拟，包括从单细胞、微生物纯菌等微观尺度，到培养瓶、盆栽等小尺度，以及到小区、田块、流域等中等尺度，再到国家、洲际和全球等大尺度，建立了包含海洋固氮[59]、预测硝化速率及 N_2O 的排放[60]等多个全球尺度时空模型，氮肥的输入和氮沉降[61]相关的多个农业模型，改善水体富营养化的反硝化过程和厌氧氨氧化模型[62]，以及识别固氮基因的分子生物学模型[63]等。

从 20 世纪 90 年代开始，随着 SVM 和 RNN 的流行，机器学习转变为数据驱动的方法。相比于传统模型，机器学习具有以下优势。一方面，机器学习无需太多先验知识。例如，在预测硝化作用过程中，随机梯度提升(SGB)相比于基于物理过程的 WNMM(水氮管理模型)、APSIM(农业生产系统模型)等模型表现出更佳的性能[60]。在评估空气质量的确定性方面，随机森林(RF)方法表现出比化学和物理传输模型 WRF-CMAQ 更高的准确性[64]。另一方面，机器学习方法已被证明比以往的机械或半经验建模方法更强大和灵活。例如，具有一个隐藏层的人工神经网络能够过滤去除噪声，预测 CO_2 通量的昼夜和季节性变化[65]。Cui 等[66]提出了一个由随机森林、梯度提升和反向传播神经网络组成的集成机器学习模型，实现了对未被以往氮循环模型量化的 HONO 估算。2010 年以来，随着深度学习的兴起，数据驱动的优势更加得到加强，传统模型高成本化学求解的束缚被进一步挣脱。同时，对于理论和经验知识还未达到或还未成熟

的情况，机器学习还可以提供一种依赖于数据来弥补未知的映射关系的解决方法[67]。然而，尽管机器学习算法具有强大的泛化能力和非线性学习能力，但大多数算法的黑箱特点导致其不可解释或模型的可解释性不足，且随着隐藏层层数的增多，可解释性越差[68](图 4)，这也催生了模型解释领域的发展[69]。例如，Hou 等[18]采用 RF 辅以 Shapley 加性解释算法和 post hoc 解释技术揭示了大气霾污染的驱动因素。另外，将机器学习算法集成到物理过程模型框架中也可弥补可解释性较差的短板，联合数据同化算法融合时空上离散分布的不同来源和分辨率的直接或间接观测信息来自动调整模型轨迹，以减少动态模型中的偏差[70]。Zhan 等[71]开发的新型混合模型随机森林时空克里格法(RF-STK)，填补了每日 NO_2 统计建模的空白，成为人体健康风险评估和解决空气污染问题的关键步骤。

机器学习还为实现氮素智能管理、提高作物产量、保障土壤健康和粮食安全等提供了新的研究途径和策略，成为精准农业系统科学决策的支持工具[72]。特别是机器学习的分支深度学习和强化学习，具有更强的表征能力或环境交互能力，与氮循环中存在的反馈控制循环相结合，使得环境氮素调节更加“智能”，并通过多系统耦合和动态调整策略找到符合目标的氮素最优配置[73]。基于机器学习的视觉传感技术可以有效识别叶片/冠层或土壤氮含量[44, 74]、诊断作物营养状况[75]、自动监测缺氮胁迫[75]、确定当前季节的植物氮需求[76]、开发控释尿素[77]、预测作物产量[78]等。氮素的有效分配将最大限度地提高作物生产力，这不仅节省了人力、物力和经济成本，还减少氮素流失所造成的一系列生态问题[79]。同样地，自动化和 DNA 测序技术的最新进展大大降低了分析微生物群落组成的成本，机器学习的回归和分类模型则可以利用从农田土壤中收集的 16S rRNA 基因数据对土壤健康进行综合评估[80]。在生物学上，Higdon 等[63]用 RF 训练分类模型识别具有生物固氮特征的基因，与泛基因组关联研究(Pan-GWAS)识别的基因进行比对和协同建模，鉴定出玉米分离株中乳球菌泛基因组与生物固氮相关的基因子集。

机器学习还在与氮循环相关的河流生态学、流域面源污染控制、溪流湖泊恢复生态学等领域具有潜在的应用前景。如，基于高分辨率卫星遥感产品，利用挺水植物对氮去除或水净化的光谱响应，引入 4 种机器学习方法来估算水体总氮浓度[44]，可能是一种新的水质参数光学估算方法[81]；基于 LSTM 架构，可提前几个小时预测污水处理厂氨氮和硝氮的排放浓度[82]；而 Xu 等[62]使用 ANN 模型进一步揭示了不同抗生素抑制下厌氧氨氧化脱氮过程的响应效应及潜在机制，并融合动力学建模方法对最大脱氮率进行了预测。

总体而言，各种经典稳健的机器学习方法和进阶深度学习算法已经应用于地球系统科学的主要子领域，并且越来越多地被整合、用于补充和增强现有的物理过程模型，在生态系统氮素循环的多个过程中成为支持科学决策的依据，也为理解生物地球化学氮素代谢、循环和利用等提供了新的视角。

5　总结与展望

大数据正在成为 21 世纪的关键资源之一，以数据驱动发现的模型也成为生物地球化学领域的热点议题。随着计算机性能的突破，深度学习和强化学习的持续发展，易于使用的机器学习工具箱的出现，预示着未来 10 年机器学习算法针对地球科学领域的预测研究将继续呈现持续性的增长[33]。从历史上看，机器学习已被证明具有强大的表征和泛化能力，可以进一步认识多源、多尺度、多介质、复杂高维的时空关系，研究者可以通过训练模型获取、筛选、分析和可视化生物地球化学数据，模拟氮循环重要生物或非生物转化过程，探索发现潜在转化机制，解决氮素失衡导致的土壤(如土壤酸化)、大气(如臭氧层空洞)和水体(如富营养化)等生态安全问题。通过将强化学习和深度学习结合，还能实现与环境交互，制定完整解决方案，自动改进算法，建立动态自动化系统。但在实际应用方面，未来还需要考虑模型的复杂性和可解释性，对此建议根据从地球系统物理模型派生的合成数据测试机器学习方法的性能，在遵守物理定律的框架下，同时在理论薄弱的地方发挥数据驱动和经验驱动的协同作用[20]。未来基于大数据和机器学习技术的特征工程和模型融合的研究，将会给氮循环领域的数据分析与建模带来巨大变革，为服务国家“双碳”战略以及控制全球变暖、空气污染等环境问题提供更多途径。

参考文献：

[1] Maathuis F J. Physiological functions of mineral macronutrients[J]. Current Opinion in Plant Biology, 2009, 12(3): 250–258.

[2] Melillo E D. The first green revolution: Debt peonage and the making of the nitrogen fertilizer trade, 1840-1930[J]. The American Historical Review, 2012, 117(4): 1028–1060.

[3] Rockström J, Steffen W, Noone K, et al. A safe operating space for humanity[J]. Nature, 2009, 461(7263): 472–475.
[4] Li S T, He P, Jin J Y. Nitrogen use efficiency in grain production and the estimated nitrogen input/output balance in China agriculture[J]. Journal of the Science of Food and Agriculture, 2013, 93(5): 1191–1197.
[5] Galloway J N, Townsend A R, Erisman J W, et al. Transformation of the nitrogen cycle: Recent trends, questions, and potential solutions[J]. Science, 2008, 320(5878): 889–892.
[6] Houlton B Z, Almaraz M, Aneja V, et al. A world of cobenefits: Solving the global nitrogen challenge[J]. Earth's Future, 2019, 7(8): 865–872.
[7] Wu D M, Zhang J W, Wang M D, et al. Global and regional patterns of soil nitrous acid emissions and their acceleration of rural photochemical reactions[J]. Journal of Geophysical Research: Atmospheres, 2022, 127(6): e2021JD036379.
[8] Tian H Q, Yang Q C, Najjar R G, et al. Anthropogenic and climatic influences on carbon fluxes from eastern North America to the Atlantic Ocean: A process-based modeling study[J]. Journal of Geophysical Research: Biogeosciences, 2015, 120(4): 757–772.
[9] Giltrap D L, Li C S, Saggar S. DNDC: A process-based model of greenhouse gas fluxes from agricultural soils[J]. Agriculture, Ecosystems & Environment, 2010, 136(3/4): 292–300.
[10] Overpeck J T, Meehl G A, Bony S, et al. Climate data challenges in the 21st century[J]. Science, 2011, 331(6018): 700–702.
[11] Das M, Ghosh S K. A deep-learning-based forecasting ensemble to predict missing data for remote sensing analysis[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2017, 10(12): 5228–5236.
[12] Lee H, Wang J F, Leblon B. Using linear regression, random forests, and support vector machine with unmanned aerial vehicle multispectral images to predict canopy nitrogen weight in corn[J]. Remote Sensing, 2020, 12(13): 2071.
[13] Padarian J, Minasny B, McBratney A B. Machine learning and soil sciences: A review aided by machine learning tools[J]. Soil, 2020, 6(1): 35–52.
[14] Zheng L M, Lin R, Wang X M, et al. The development and application of machine learning in atmospheric environment studies[J]. Remote Sensing, 2021, 13(23): 4839.
[15] Zhong S F, Zhang K, Bagheri M, et al. Machine learning: New ideas and tools in environmental science and engineering[J]. Environmental Science & Technology, 2021, 55(19): 12741–12754.
[16] Sit M, Demiray B Z, Xiang Z R, et al. A comprehensive review of deep learning applications in hydrology and water resources[J]. Water Science and Technology, 2020, 82(12): 2635–2670.
[17] Jin S T, Zeng X X, Xia F, et al. Application of deep learning methods in biological networks[J]. Briefings in Bioinformatics, 2021, 22(2): 1902–1917.
[18] Hou L L, Dai Q L, Song C B, et al. Revealing drivers of haze pollution by explainable machine learning[J]. Environmental Science & Technology Letters, 2022, 9(2): 112–119.
[19] Keller C A, Evans M J. Application of random forest regression to the calculation of gas-phase chemistry within the GEOS-Chem chemistry model v10[J]. Geoscientific Model Development, 2019, 12(3): 1209–1225.
[20] Reichstein M, Camps-Valls G, Stevens B, et al. Deep learning and process understanding for data-driven Earth system science[J]. Nature, 2019, 566(7743): 195–204.
[21] Canfield D E, Glazer A N, Falkowski P G. The evolution and future of Earth's nitrogen cycle[J]. Science, 2010, 330(6001): 192–196.
[22] Kuypers M M M, Marchant H K, Kartal B. The microbial nitrogen-cycling network[J]. Nature Reviews Microbiology, 2018, 16(5): 263–276.
[23] Fowler D, Coyle M, Skiba U, et al. The global nitrogen cycle in the twenty-first century[J]. Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 2013, 368(1621): 20130164.
[24] Broda E. Two kinds of lithotrophs missing in nature[J]. Zeitschrift Für Allgemeine Mikrobiologie, 1977, 17(6): 491–493.
[25] Woods D D. The reduction of nitrate to ammonia by *Clostridium* welchii[J]. The Biochemical Journal, 1938, 32(11): 2000–2012.
[26] Matsumoto S, Ae N. Characteristics of extractable soil organic nitrogen determine using various chemical solutions and its significance for nitrogen uptake by crops[J]. Soil Science and Plant Nutrition, 2004, 50(1): 1–9.
[27] Schimel J P, Bennett J. Nitrogen mineralization: Challenges of a changing paradigm[J]. Ecology, 2004, 85(3): 591–602.
[28] Xu G H, Fan X R, Miller A J. Plant nitrogen assimilation and use efficiency[J]. Annual Review of Plant Biology, 2012, 63: 153–182.
[29] Thompson R L, Lassaletta L, Patra P K, et al. Acceleration of global N_2O emissions seen from two decades of atmospheric inversion[J]. Nature Climate Change, 2019, 9(12): 993–998.
[30] 宋雅琦, 吴电明, 俞元春. 土壤活性氮气体排放研究进展[J]. 科技导报, 2022, 40(3): 130–144.
[31] Zhang X N, Ward B B, Sigman D M. Global nitrogen cycle: Critical enzymes, organisms, and processes for nitrogen budgets and dynamics[J]. Chemical Reviews, 2020, 120(12): 5308–5351.
[32] Xu Y J, Liu X, Cao X, et al. Artificial intelligence: A powerful paradigm for scientific research[J]. The Innovation, 2021, 2(4): 100179.
[33] Bergen K J, Johnson P A, de Hoop M V, et al. Machine learning for data-driven discovery in solid Earth

geoscience[J]. Science, 2019, 363(6433): eaau0323.
[34] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237–285.
[35] Zhou Z H. Machine Learning[M]. Singapore: Springer Singapore, 2021.
[36] Salcedo-Sanz S, Ghamisi P, Piles M, et al. Machine learning information fusion in Earth observation: A comprehensive review of methods, applications and data sources[J]. Information Fusion, 2020, 63: 256–272.
[37] Yang M D, Hsu Y C, Tseng W C, et al. Assessment of grain harvest moisture content using machine learning on smartphone images for optimal harvest timing[J]. Sensors, 2021, 21(17): 5875.
[38] Liu X, Lu D W, Zhang A Q, et al. Data-driven machine learning in environmental pollution: Gains and problems[J]. Environmental Science & Technology, 2022, 56(4): 2124–2133.
[39] 周慧颖, 汪廷华, 张代俐. 多标签特征选择研究进展[J]. 计算机工程与应用, 2022, 58(15): 52–67.
[40] 何坤龙, 赵伟, 刘晓辉, 等. 云雾覆盖下地表温度重建机器学习模型的训练集敏感性分析[J]. 遥感学报, 2021, 25(8): 1722–1734.
[41] 王惠. 迁移学习研究综述[J]. 电脑知识与技术, 2017, 13(32): 203–205.
[42] Glenn A J, Moulin A P, Roy A K, et al. Soil nitrous oxide emissions from no-till canola production under variable rate nitrogen fertilizer management[J]. Geoderma, 2021, 385: 114857.
[43] Li R, Cui L L, Zhao Y L, et al. Long-term trends of ambient nitrate (NO_3^-) concentrations across China based on ensemble machine-learning models[J]. Earth System Science Data, 2021, 13(5): 2147–2163.
[44] Wang J Z, Shi T Z, Yu D L, et al. Ensemble machine-learning-based framework for estimating total nitrogen concentration in water using drone-borne hyperspectral imagery of emergent plants: A case study in an arid oasis, NW China[J]. Environmental Pollution, 2020, 266(Pt 2): 115412.
[45] Mashaba-Munghemezulu Z, Chirima G J, Munghemezulu C. Modeling the spatial distribution of soil nitrogen content at smallholder maize farms using machine learning regression and sentinel-2 data[J]. Sustainability, 2021, 13(21): 11591.
[46] Noble W S. What is a support vector machine?[J]. Nature Biotechnology, 2006, 24(12): 1565–1567.
[47] 奉国和. SVM 分类核函数及参数选择比较[J]. 计算机工程与应用. 2011, 47(3): 123–124.
[48] Zhou T, Geng Y J, Chen J, et al. High-resolution digital mapping of soil organic carbon and soil total nitrogen using DEM derivatives, Sentinel-1 and Sentinel-2 data based on machine learning algorithms[J]. Science of the Total Environment, 2020, 729: 138244.
[49] Kim Y, Oh S. Machine-learning insights into nitrate-reducing communities in a full-scale municipal wastewater treatment plant[J]. Journal of Environmental Management, 2021, 300: 113795.
[50] Qiu Z C, Ma F, Li Z W, et al. Estimation of nitrogen nutrition index in rice from UAV RGB images coupled with machine learning algorithms[J]. Computers and Electronics in Agriculture, 2021, 189: 106421.
[51] 张驰, 郭媛, 黎明. 人工神经网络模型发展及应用综述[J]. 计算机工程与应用. 2021, 57(11): 57–69.
[52] Werbos P J. The roots of backpropagation: From ordered derivatives to neural networks and political forecasting[M]. New York: John Wiley & Sons, 1994.
[53] Cortes C, Vapnik V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273–297.
[54] Zhang Q C, Yang L T, Chen Z K, et al. A survey on deep learning for big data[J]. Information Fusion, 2018, 42: 146–157.
[55] Le Cun Y, Boser B, Denker J S, et al. Handwritten digit recognition with a back-propagation network[J]. Advances in Neural Information Processing Systems, 1990: 396–404.
[56] Chen L Y, Li S B, Bai Q A, et al. Review of image classification algorithms based on convolutional neural networks[J]. Remote Sensing, 2021, 13(22): 4712.
[57] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena. 2020, 404: 132306.
[58] Li K L, Duan H R, Liu L F, et al. An integrated first principal and deep learning approach for modeling nitrous oxide emissions from wastewater treatment plants[J]. Environmental Science & Technology, 2022, 56(4): 2816–2826.
[59] Tang W Y, Li Z C, Cassar N. Machine learning estimates of global marine nitrogen fixation[J]. Journal of Geophysical Research: Biogeosciences, 2019, 124(3): 717–730.
[60] Pan B B, Lam S K, Wang E L, et al. New approach for predicting nitrification and its fraction of N_2O emissions in global terrestrial ecosystems[J]. Environmental Research Letters, 2021, 16(3): 034053.
[61] Lu X C, Yuan D H, Chen Y A, et al. Estimations of long-term nss-SO_4^{2-} and NO_3^- wet depositions over East Asia by use of ensemble machine-learning method[J]. Environmental Science & Technology, 2020, 54(18): 11118–11126.
[62] Xu X X, Liu S, Zeng M, et al. Deciphering response effect and underlying mechanism of anammox-based nitrogen removal process under exposures to different antibiotics via big data analysis[J]. Bioresource Technology, 2022, 347: 126674.
[63] Higdon S M, Huang B C, Bennett A B, et al. Identification of nitrogen fixation genes in *Lactococcus* isolated from maize using population genomics and machine learning[J]. Microorganisms, 2020, 8(12): 2043.
[64] Vu T V, Shi Z B, Cheng J, et al. Assessing the impact of clean air action on air quality trends in Beijing using a machine learning technique[J]. Atmospheric Chemistry and

Physics, 2019, 19(17): 11303–11314.
[65] Papale D, Valentini R. A new assessment of European forests carbon exchanges by eddy fluxes and artificial neural network spatialization[J]. Global Change Biology, 2003, 9(4): 525–535.
[66] Cui L L, Wang S X. Mapping the daily nitrous acid (HONO) concentrations across China during 2006-2017 through ensemble machine-learning algorithm[J]. Science of the Total Environment, 2021, 785: 147325.
[67] Taki R, Wagner-Riddle C, Parkin G, et al. Comparison of two gap-filling techniques for nitrous oxide fluxes from agricultural soil[J]. Canadian Journal of Soil Science, 2019, 99(1): 12–24.
[68] Zdeborová L. Understanding deep learning is also a job for physicists[J]. Nature Physics, 2020, 16(6): 602–604.
[69] Toms B A, Barnes E A, Ebert-Uphoff I. Physically interpretable neural networks for the geosciences: Applications to earth system variability[J]. Journal of Advances in Modeling Earth Systems, 2020, 12(9): e2002M-e2019M.
[70] Ivatt P D, Evans M J. Improving the prediction of an atmospheric chemistry transport model using gradient-boosted regression trees[J]. Atmospheric Chemistry and Physics, 2020, 20(13): 8063–8082.
[71] Zhan Y, Luo Y Z, Deng X F, et al. Satellite-based estimates of daily NO_2 exposure in China using hybrid random forest and spatiotemporal kriging model[J]. Environmental Science & Technology, 2018, 52(7): 4180–4189.
[72] Ghahramani Z. Probabilistic machine learning and artificial intelligence[J]. Nature, 2015, 521(7553): 452–459.
[73] Irrgang C, Boers N, Sonnewald M, et al. Towards neural Earth system modelling by integrating artificial intelligence in Earth system science[J]. Nature Machine Intelligence, 2021, 3(8): 667–674.
[74] Patel A K, Ghosh J K, Pande S, et al. Deep-learning-based approach for estimation of fractional abundance of nitrogen in soil from hyperspectral data[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 6495–6511.
[75] Barbedo J G A. Detection of nutrition deficiencies in plants using proximal images and machine learning: A review[J]. Computers and Electronics in Agriculture, 2019, 162: 482–492.
[76] Qin Z S, Myers D B, Ransom C J, et al. Application of machine learning methodologies for predicting corn economic optimal nitrogen rate[J]. Agronomy Journal, 2018, 110(6): 2596–2607.
[77] Jiang Z W, Yang S H, Chen X, et al. Controlled release urea improves rice production and reduces environmental pollution: A research based on meta-analysis and machine learning[J]. Environmental Science and Pollution Research International, 2022, 29(3): 3587–3599.
[78] Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review[J]. Computers and Electronics in Agriculture, 2018, 151: 61–69.
[79] Yang Y, Shang X, Chen Z, et al. A support vector regression model to predict nitrate-nitrogen isotopic composition using hydro-chemical variables[J]. Journal of Environmental Management, 2021, 290: 112674.
[80] Wilhelm R C, van Es H M, Buckley D H. Predicting measures of soil health using the microbiome and supervised machine learning[J]. Soil Biology and Biochemistry, 2022, 164: 108472.
[81] Niu C, Tan K, Jia X P, et al. Deep learning based regression for optically inactive inland water quality parameter estimation using airborne hyperspectral imagery[J]. Environmental Pollution, 2021, 286: 117534.
[82] Farhi N, Kohen E, Mamane H, et al. Prediction of wastewater treatment quality using LSTM neural network[J]. Environmental Technology & Innovation, 2021, 23: 101632.