Climate Change Research Letters 气候变化研究快报, 2026, 15(3), 674-687
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ccrl
https://doi.org/10.12677/ccrl.2026.153072

Hans汉斯

机器学习在黄土高原沟壑区插补二氧化碳浓度的潜力评估

邓　炜[1,2]，刘登峰[1*]，李明亮[3]，郭凤年[1]，孟静静[1]，黄　强[1]

[1]西安理工大学水利水电学院，旱区水工程生态环境全国重点实验室，陕西　西安
[2]赣江下游水文水资源监测中心新余水文监测大队，江西　新余
[3]水利部水利水电规划设计总院，北京

收稿日期：2026年4月23日；录用日期：2026年5月22日；发布日期：2026年5月28日

摘　要

以陕西的淳化生态水文实验基地的沟底和塬上的实测二氧化碳浓度数据和空气温度数据，用沟底观测数据评价了多层感知机(MLP)、双向长短期记忆神经网络(Bi-LSTM)、随机森林(RF)、多元线性回归(MLR)等机器学习方法模拟空气二氧化碳浓度的潜力，遴选性能较好的机器学习方法对塬上二氧化碳浓度进行了模拟、插补，并评估插补值的质量。结果表明：在沟底植被覆盖度不同的区域，仅以实测二氧化碳浓度为输入，机器学习方法即可达到可接受的模拟精度；增加空气温度作为输入后，MLP和RF模型能进一步提升模拟精度。利用沟底二氧化碳浓度数据插补塬上浓度时，MLP和RF仍表现出较高的可行性，能有效还原变化趋势，虽然在部分时段对高值存在轻微低估，但其插补结果与实测值的均值、极大值、极小值的最大偏差仅为5%。MLP和RF可以用于进行区域二氧化碳浓度模拟与数据插补。

关键词

二氧化碳浓度，机器学习，生态水文，数据插补，黄土高原

Evaluation of the Potential of Using Machine Learning to Interpolate the CO_2 Concentration in the Gully Region of the Loess Plateau

Wei Deng[1,2], Dengfeng Liu[1*], Mingliang Li[3], Fengnian Guo[1], Jingjing Meng[1], Qiang Huang[1]

[1]State Key Laboratory of Water Engineering Ecology and Environment in Arid Area, Xi'an University of Technology, School of Water Resources and Hydropower, Xi'an Shaanxi
[2]Xinyu Hydrological Monitoring Brigade, Hydrology and Water Resources Monitoring Center of Lower Ganjiang River, Xinyu Jiangxi

[*]通讯作者。

文章引用: 邓炜, 刘登峰, 李明亮, 郭凤年, 孟静静, 黄强. 机器学习在黄土高原沟壑区插补二氧化碳浓度的潜力评估[J]. 气候变化研究快报, 2026, 15(3): 674-687. DOI: 10.12677/ccrl.2026.153072

[3]General Institute of Water Resources and Hydropower Planning and Design, Ministry of Water Resources, Beijing

Received: April 23, 2026; accepted: May 22, 2026; published: May 28, 2026

Abstract

Based on the measured data of carbon dioxide concentration and air temperature at the bottom of the gully and on the plateau of the Chunhua Ecological Hydrological Experimental Base in Shaanxi Province, the potential of machine learning methods such as Multilayer Perceptrons (MLP), Bidirectional Long Short-term Memory (Bi-LSTM), Random Forest (RF) and Multiple Linear Regression (MLR) in simulating air carbon dioxide concentration was evaluated using the observation data from the bottom of the gully. The machine learning method with better performance was selected to simulate and interpolate the carbon dioxide concentration on the tableland, and the quality of the interpolation value was evaluated. The results show that in areas with different vegetation coverage at the bottom of the ditch, when only the measured carbon dioxide concentration is used as the input, the machine learning method can achieve an acceptable simulation accuracy; when air temperature is added as an input, the MLP and RF models can further improve the simulation accuracy. When using the ditch bottom carbon dioxide concentration data to interpolate the concentration on the plateau, MLP and RF still show high feasibility and can effectively restore the changing trend, although there is a slight underestimation of high values at some times, the maximum deviation between the interpolation results and the measured values of mean, maximum and minimum values is only 5%. MLP and RF can be used for regional carbon dioxide concentration simulation and data interpolation.

Keywords

Carbon Dioxide Concentration, Machine Learning, Ecohydrology, Data Interpolation, Loess Plateau

1. 引言

黄土高原位于干旱半干旱地区，生态环境脆弱十分脆弱，且对气候变化的响应极其敏感[1] [2]。近百年来，大气中温室气体含量不断增加，全球气候变化逐渐加剧，人类社会和生态环境都受到严重影响[3] [4]。在人类活动和气候变化的共同作用下，许多生态环境脆弱区都在变成一个显著的碳汇[5]。黄土高原作为我国水土流失较严重的区域，我国于 20 世纪 50 年代设立水土保持科学实验研究机构系统开展黄土高原沟壑区水土流失科学研究工作，并不断完善监测站点布置，了解区域碳过程，科学开展区域生态治理工作[6]。

二氧化碳浓度是区域碳汇评估的重要变量。目前涡度相关系统和红外气体分析仪等设备监测二氧化碳浓度精度高、分辨率高，但设备监测普遍存在价格昂贵、铺设具有一定的局限性以及设备定期标定期间无法获取数据等问题[7]，严重影响了在大尺度、长期连续监测中的应用。随着机器学习方法在水文模拟上的广泛运用[8] [9]，已经有学者将机器学习方法用于二氧化碳浓度的模拟，以更低的成本获得高质量的二氧化碳浓度数据，更好地探究二氧化碳浓度变化过程。缪云飞[10]采用前馈神经网络与量化共轭梯度算法，构建了短波红外通道卫星 CO_2 反演模型，并利用 GOSAT 卫星观测光谱数据反演了 CO_2 浓度。赵嘉宁[11]等采

用随机森林等机器学习方法对利用秦岭北麓关中站实测二氧化碳浓度进行校正，明显提高了数据质量。邓炜[12]等采用多种机器学习学习方法，以空气温度等环境因子对沟壑区底部近地表二氧化碳浓度进行了模拟，以空气温度和土壤温度等环境因子作为输入，模拟了研究时段内二氧化碳浓度变化过程。

空气温度直接影响植被呼吸作用和光合作用，也有学者开始采用空气温度等环境因子用于模拟二氧化碳浓度，但目前大多学者仅聚焦于单一观测系统，环境变异性相对有限[12]。沟壑区作为黄土高原的典型生态脆弱区[13]，沟壑区局地海拔落差较大，沟壑区沟底和塬上环境变异性较大，设备布设和维护成本相对较高。因此，本文在黄土高原沟壑区沟底遴选模拟效果输入组合和机器学习方法；采用模拟效果较好的机器学习方法，以沟底二氧化碳浓度和塬上空气温度作为输入组合，对塬上二氧化碳浓度数据进行插值补全，进一步提升数据质量。这是对于跨区域二氧化碳数据处理的进一步尝试，有助于更好地了解黄土高原沟壑区沟底和塬上地面空气二氧化碳浓度的变化过程，优化站点设备布设，解决由于设备异常、设备标定导致的数据缺失问题，并对黄土高原沟壑区固碳增汇技术提供参考，为黄土高原乃至黄河流域在碳中和背景下快速发展提供科技支撑。

2. 研究区与数据

2.1. 研究区概况

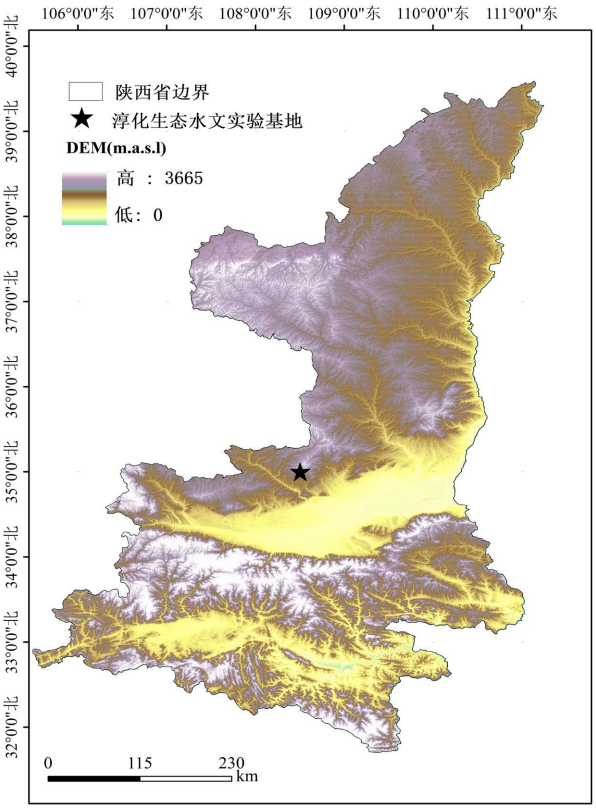

Figure 1. The location of Chunhua eco-hydrological experimental base
图 1. 淳化生态水文实验基地位置图

本研究的站点位于陕西省咸阳市淳化县的和家山小流域的淳化生态水文实验基地。实验基地位于渭河北岸的黄土高原沟壑区，如图 1 所示。气候属温带大陆性季风气候，盛行东风，降雨集中在夏季和秋季[14]。基地共两套观测系统可连续观测二氧化碳浓度，两处观测系统海拔相差约 100 米，一套二氧化碳

浓度观测系统位于沟道底部，另一套水碳通量观测系统位于塬上，其中沟道底部的植被覆盖度高，主要有灰绿藜和碱蒿等植被[15]。

2.2. 沟底的二氧化碳浓度观测系统

沟道底部的二氧化碳浓度观测系统主要观测设备包括六要素气象传感器(WXA100-06，中铭电气，China)和二氧化碳传感器(GMP252, Vaisala, Finland) [16]，可以进行二氧化碳浓度和常规气象要素进行连续观测。观测系统的现场布设如图 2 所示，共设置 3 个二氧化碳传感器(安装高度为 0.3 m)用于观测地面空气二氧化碳浓度，自北向南分别为 1 号传感器、2 号传感器和 3 号传感器，观测的二氧化碳浓度分别记为 C_1、C_2、C_3；在 2 号传感器西侧安装有六要素气象传感器(安装高度为 3 m)，可观测空气温度、空气相对湿度等多种环境因子。

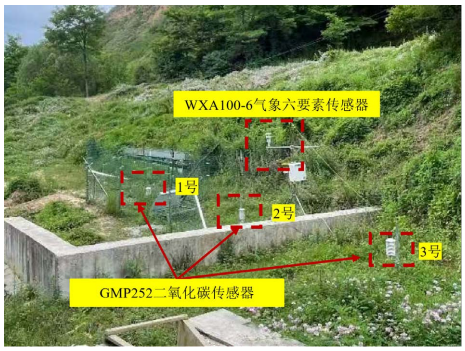

Figure 2. The layout scheme of carbon dioxide concentration observation system at the bottom of gully region
图 2. 沟底的二氧化碳浓度观测系统布设方案

2.3. 塬上通量塔的水碳通量观测系统

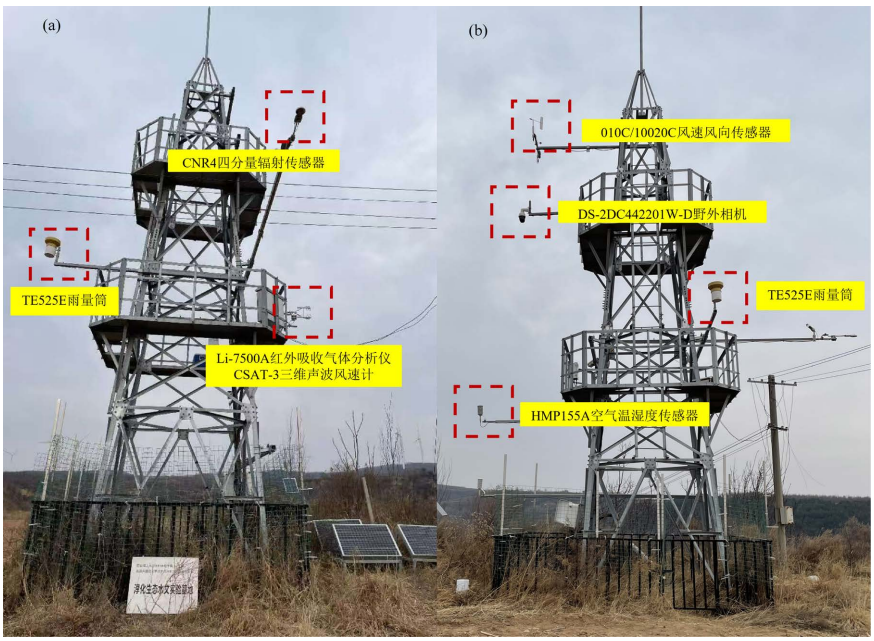

Figure 3. The layout scheme of water and carbon flux observation system (a) To observe the flux tower northward; (b) To observe the flux tower eastward
图 3. 水碳通量观测系统布设方案(a) 向北观察通量塔；(b) 向东观察通量塔

通量塔的水碳通量观测系统位于沟壑区的塬上，主要观测设备包括红外吸收气体分析仪(Li-7500A, LI-COR, USA)、三维声波风(CSAT-3, Campbell Scientific., USA)、空气温湿度传感器(HMP155A, Vaisala, Finland)和四分量辐射传感器(CNR4, Kipp&Zonen, Netherlands)等组成[16]，可以进行二氧化碳浓度和常规气象要素等进行连续观测。现场布设如图 3 所示。涡度相关系统(安装高度为 5 m)由红外吸收气体分析仪由红外吸收气体分析仪和三维声波风组成，可用于观测二氧化碳浓度，记为 C_4；在涡度相关系统的北侧安装有空气温湿度传感器(安装高度为 2 m)，可观测空气温度和空气相对湿度；在涡度相关系统的南侧安装有四分量辐射传感器(安装高度为 5 m)，可观测长波辐射和短波辐射。

2.4. 数据及预处理

本研究所采用的数据包括沟底 2 号传感器处的空气二氧化碳浓度 C_2、沟底 3 号传感器处的空气二氧化碳浓度 C_3、沟底空气温度 T_{air1}、塬上空气温度 T_{air2}、塬上水碳通量观测系统的二氧化碳浓度 C_4，均为淳化生态水文试验基地的实测数据。所用数据数据为 2021 年 2 月 1 日至 2021 年 12 月 3 日的日尺度数据，数据长度为 300 个。采用线性插补补全和 SG 滤波预处理后的数据如图 4 和图 5 所示。在整个研究时段内，塬上、沟底的空气二氧化碳浓度和空气温度的日变化都十分剧烈。2 月和 11 月份空气温度较低，二氧化碳浓度较高，6 月至 10 月空气温度较高，二氧化碳浓度较低[12]。对比不同传感器的数据以及实验基地设备维护记录可知，塬上水碳通量观测系统的二氧化碳浓度 C_4 于 2021 年 9 月初由于探头积尘出现数据异常，初步清洗后于 9 月中旬将设备带离实验基地进行标定，导致 9 月份至 12 月份部分数据异常或缺失。该时段的二氧化碳浓度数据通过线性插补补全，仅具有增长的趋势，不具有较剧烈的日尺度变化。

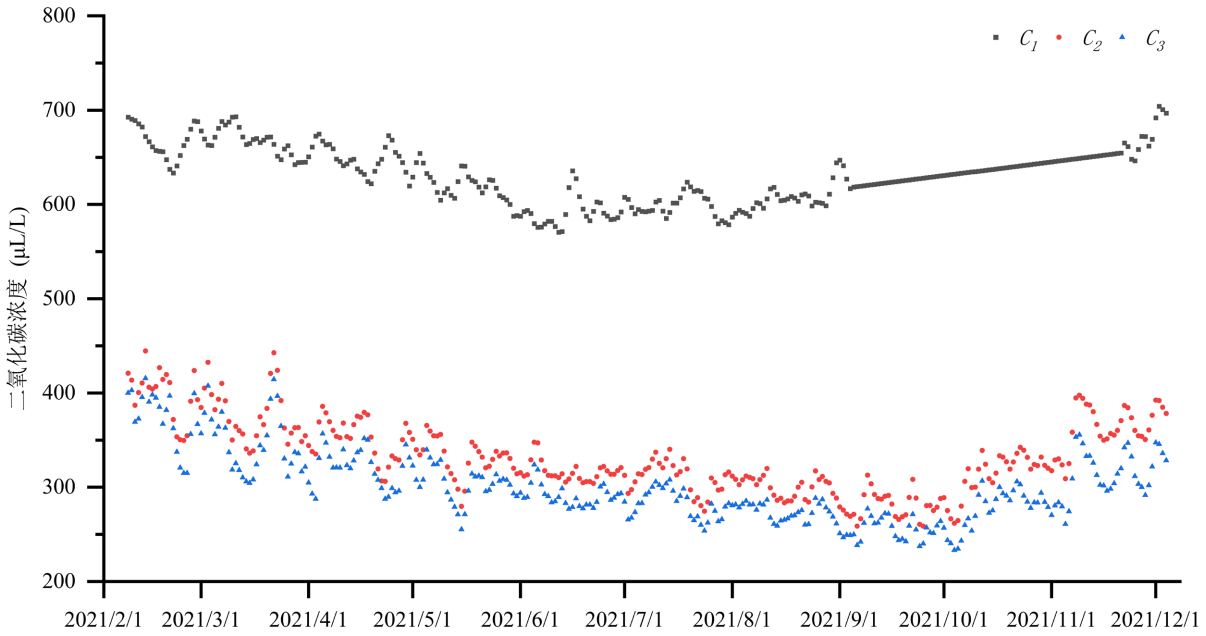

Figure 4. The data of air carbon dioxide concentration
图 4. 空气二氧化碳浓度的数据

为了更好地分析二氧化碳浓度(C_2, C_3, C_4)，空气温度(T_{air1}, T_{air2})的统计信息，本研究计算了研究时段序列的多个统计参数，计算结果见表 1。由计算结果可知，对比不同海拔的二氧化碳浓度来看，塬上气温的最大值为 25.7℃，最小值为−2.4℃，平均值为 13.3℃，标准差为 6.9℃；沟底气温的最大值为 23.8℃，最小值为−1.8℃，平均值为 12.8℃，标准差为 6.4℃。一般来说气温随着海拔的升高会逐渐降低，且不同

区域的气温垂直递减率也有所差异[17]。在本研究区，塬上的海拔要比沟底高 100 m，但研究时段内，塬上的气温平均值却略高于沟底。出现这种逆温现象的主要原因是沟壑区地形条件特殊。塬上接受太阳辐射的面积较大，而沟底西侧则存在较高坡地的遮挡，且存在树木遮挡现象，沟底接受太阳辐射的面积和时间要少于塬上。塬上气温和沟底气温变差系数、偏度和峰度最大仅相差 0.05 左右，塬上和沟底气温具有相似的统计规律。对比塬上和沟底的实测空气二氧化碳浓度来看，沟底的二氧化碳浓度平均值、最大值、最小值都要远低于塬上二氧化碳浓度，但其标准差高于塬上，沟底二氧化碳浓度变化更为剧烈。此外，对比沟底两个传感器处的空气二氧化碳浓度可知，随着植被覆盖率的增加，二氧化碳浓度的变差系数、偏度都有所增加，二氧化碳浓度的变化越剧烈[12]。

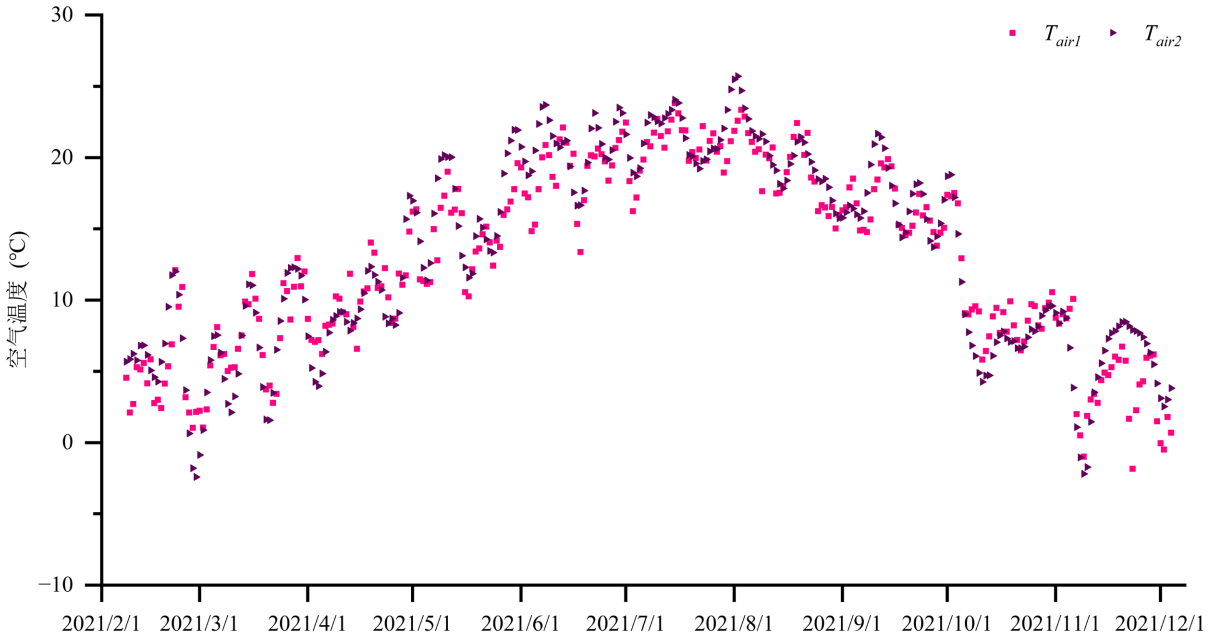

Figure 5. The data of air temperature
图 5. 空气温度数据

Table 1. The brief statistical parameters of carbon dioxide concentration and air temperature
表 1. 二氧化碳浓度和空气温度的简要统计参数

数据	x_{mean}	x_{max}	x_{min}	x_{std}	C_v	C_s	C_k
C_2/(μL/L)	332.5	444.6	258.7	39.3	0.12	0.50	−0.24
C_3/(μL/L)	301.6	415.6	233.2	37.9	0.13	0.82	0.41
C_4/(μL/L)	631.2	704.1	570.6	29.8	0.05	0.10	−0.68
T_{air1}/˚C	12.8	23.8	−1.8	6.4	0.50	−0.19	−1.12
T_{air2}/˚C	13.3	25.7	−2.4	6.9	0.52	−0.14	−1.17

二氧化碳浓度和温度的数值相差数十倍。为了提高模拟精度，采用空气温度和二氧化碳浓度时间序列用于模拟时，本研究按式(1)对使用的所有变量序列都进行归一化处理。

$$x_{normal} = \frac{x - x_{mean}}{x_{std}} \tag{1}$$

式中，x_{normal} 为经过归一化处理后的变量序列，x 为原始观测的变量序列，x_{mean} 为相应变量序列的平均值，

x_{std} 为相应变量序列的标准差。

3. 研究方法

3.1. 多元线性回归

多元线性回归(Multiple Linear Regression, MLR)是一种应用十分广泛的机器学习方法之一，具有计算效率高等优点，且不需要预设参数[16]。MLR (两个输入变量)通过下式表示多个输入变量和目标变量的关系：

$$y = \alpha_1 x_1 + \alpha_2 x_2 + \beta \tag{2}$$

式中：y 是模拟值；x_1、x_2 是输入变量；α_1 是输入变量 x_1 的斜率系数，α_2 是输入变量 x_2 的斜率系数，β 是截距。

3.2. 多层感知机

多层感知机(Multilayer Perceptrons, MLP)是最经典的神经网络模型之一，主要由输入层、隐藏层和输出层构成，其中输入层和输出层分别只有一个，隐层可以是一层或多层拓扑结构[18]。MLP 因具有较好的非线性全局作用，常用于解决回归问题[19]。MLP (一层隐藏层)的基本结构如图 6 所示，可用下式表示：

$$y = J_1\left(b^{(2)} + W^{(2)}\left(J_2\left(b^{(1)} + W^{(1)}x\right)\right)\right) \tag{3}$$

式中：y 是模拟值；J_1 和 J_2 是激活函数；$W^{(1)}$和 $W^{(2)}$是隐藏层和输出层的权值系数；$b^{(1)}$和 $b^{(2)}$是隐藏层和输出层偏置项。

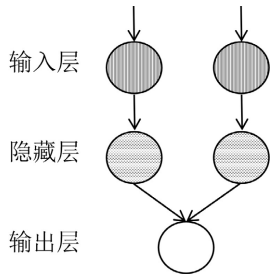

Figure 6. The structure of MLP
图 6. 多层感知机的结构

3.3. 双向长短期记忆神经网络

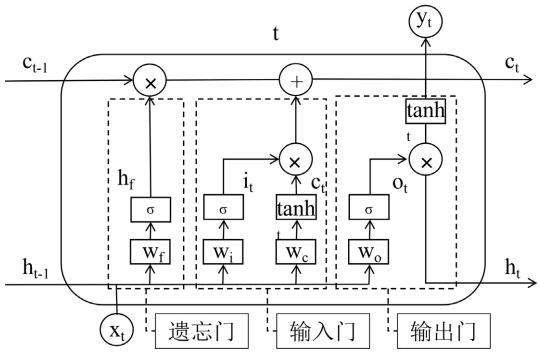

Figure 7. The structure of LSTM
图 7. 长短期记忆神经网络的单元结构

双向长短期记忆神经网络(Bidirectional Long Short-term Memory, Bi-LSTM)是基于长短期记忆神经网络(Long Short Term Memory, LSTM)产生的[20]。LSTM 由遗忘门、输入门和输出门构成[21]。Bi-LSTM 由前向 LSTM 和后向 LSTM 组合而成，其中前向 LSTM 获取输入序列的过去信息，后向 LSTM 获取输入序列的未来信息，预测结果由两个 LSTM 共同决定[22]。LSTM 的结构如图 7 所示，Bi-LSTM 的结构如图 8 所示。

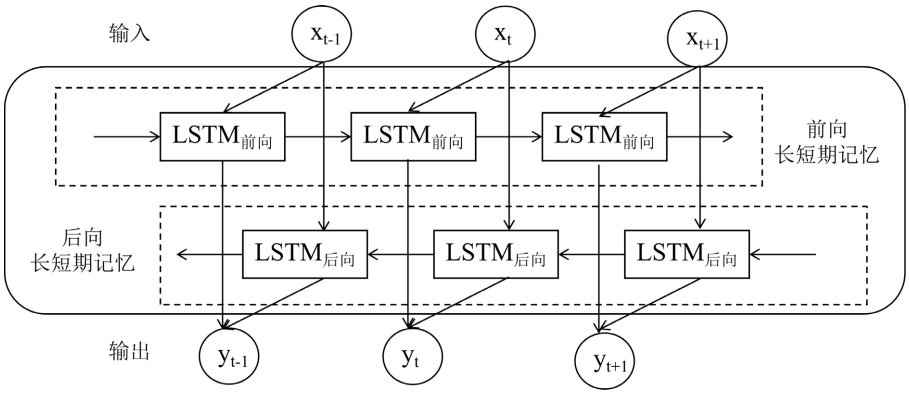

Figure 8. The structure of Bi-LSTM
图 8. 双向长短期记忆神经网络的结构

3.4. 随机森林

随机森林(Random Forest, RF)是一种基于决策树的集成学习算法，其核心在于通过构建多棵决策树并集成其结果，以提升模型的预测性能和泛化能力[23] [24]。RF 的构造过程如图 9 所示，训练集中随机抽取子数据集构成决策树，多个决策树构成随机森林，其中的每个决策树都根据数据子集和解释变量子集进行训练，预测值由各决策树共同决定[25] [26]。RF 是一种能够有效减少过拟合的非线性建模工具，对解决多变量的预测具有很好的效果，目前已广泛用于解决分类和回归问题[27]。

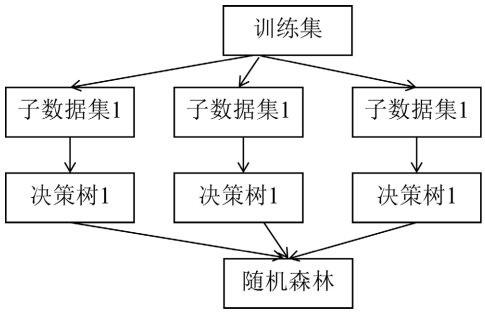

Figure 9. The construction process of RF
图 9. 随机森林的构造过程

3.5. 机器学习模型的参数设置

本研究所用超参数设置与前期研究保持一致，仍采用 Python 中 Sklearn 工具库的随机搜索确定 MLP、Bi-LSTM 和 RF 的超参数。MLP 的超参数设置包括隐藏层数量，神经元数量，激活函数等；Bi-LSTM 的超参数设置包括神经元数量，激活函数等；RF 的超参数设置包括树的数量，节点可分得的最小样本数等[12]。由于所用数据长度仅为 300 个，为减少数据量对模拟精度的影响，采用十折交叉验证。塬上的二氧

化碳浓度的从 9 月初开始出现异常，2 月初至 8 月底部分数据为正常数据，约占整个研究时段的 68%。综合考虑，本研究中 3 种机器学习中训练集所占比例均设为 60%。

3.6. 模型评价指标的选择

本研究选用平均绝对误差(Mean absolute error, MAE)，均方根误差(Root mean square error, RMSE)，以及 KGE 系数(Kling-Gupta efficiency coefficient, KGE)评价不同机器学习方法的性能差异。三个评价指标的计算公式如下：

$$\begin{aligned} \mathrm{MAE} &= \frac{\sum_1^n \left| y_i - \hat{y}_i \right|}{n} \\ \mathrm{RMSE} &= \sqrt{\frac{\sum_1^n \left(y_i - \hat{y}_i \right)^2}{n}} \\ \mathrm{KGE} &= 1 - \sqrt{(r-1)^2 + \left(\frac{\mu_s}{\mu_0} - 1 \right)^2 + \left(\frac{\sigma_s / \mu_s}{\sigma_0 / \mu_0} - 1 \right)^2} \end{aligned} \tag{4}$$

式中，n 是样本数据总数；y_i 是实测值，$\hat{y}_i$ 是模拟值，$\bar{y}$ 是实测值系列的均值；r 是模拟值和实测值系列之间的线性相关系数，μ_s 是模拟值系列的均值，μ_0 实测值系列的均值，σ_s 是模拟值系列的均方差，σ_0 实测值系列的均方差。三个评价指标中 MAE 和 RMSE 的值越小模拟效果越好，KGE 的值越接近 1 模拟效果越好。

4. 结构与分析

4.1. 沟底二氧化碳浓度模拟潜力评价

本文选用二氧化碳浓度和空气温度作用输入变量，其中空气温度和二氧化碳浓度又相互影响。一般情况下，输入变量与输出变量的相关性越高，模拟的准确性也越高。因此本研究首先对变量间相关关系进行分析，计算沟底两处不同植被覆盖度下的二氧化碳浓度和空气温度的线性相关系数，相关系数的计算结果见表 2。计算结果表明，虽然沟底两个传感器处的植被覆盖度有所不同，但是两处的二氧化碳浓度具有相同的变化规律，线性相关系数为 0.97。3 号传感器处的二氧化碳浓度 C_3 和空气温度 T_{air1} 的线性相关系数为−0.75，二氧化碳浓度和空气温度呈明显的负相关。共确定 2 种用于 C_3 模拟的输入组合，输入组合 1 (input1)由 2 号传感器处的二氧化碳浓度 C_2 单独作为输入；输入组合 2 (input2)由 2 号传感器处的二氧化碳浓度 C_2 和空气温度 T_{air1} 共同作为输入。

Table 2. The line correlation coefficient between data at the bottom of the gully
表 2. 沟底所用数据间的线性相关系数

数据	C_3/(μL/L)
C_2/μL/L	0.97
T_{air1}/˚C	−0.75

采用 MLR、MLP、Bi-LSTM 和 RF 对 C_3 进行模拟，不同机器学习方法、不同输入组合的评价指标计算结果见表 3。对比不同机器学习在测试集的评价指标，无论是以单一变量还是组合变量作为输入，MLP、Bi-LSTM 和 RF 的 KGE 均大于 0.8，MLP 的 KGE 大于 0.7，模拟效果都是可接受的，其

中 MLP 的模拟效果最好，Bi-LSTM 模拟效果最差。由于不同位置二氧化碳浓度线性相关系数高达0.97，采用 MLR 进行模拟时，其 MAE 和 RMSE 较低。采用 Bi-LSTM 进行模拟时，滑动窗格的设置直接影响模型的精度和运行速度。由于地面空气二氧化碳浓度和空气温度的日变化都十分剧烈，经过前期的多次手动调参，综合考虑运行效率和模拟精度，将 Bi-LSTM 在模拟中选用的滑动窗格大小设置为 3。即使滑动窗格设置得很小，Bi-LSTM 仍因其对数据的充分利用导致了模拟精度偏低。空气温度直接影响植被生长，植被生长直接影响二氧化碳浓度。对比不同输入组合来看，数据的有效性是有条件的，并非增加输入的数据量就可以增加模拟精度[28] [29]。四种机器学习方法中，MLR、MLP 和 RF 可以较好地捕捉空气温度和二氧化碳浓度之间的关系，在增加空气温度作为输入后，模拟精度有所提高。从评价指标计算结果来看，以二氧化碳浓度和空气温度(input2)共同作为输入，采用 MLP 的模拟精度最高，其在测试集的相对误差仅 3.1%，平均绝对误差仅 8.985 μL/L，均方根误差仅 11.992 μL/L。

Table 3. The evaluation indexes of ground air carbon dioxide concentration at sensor 3 (C_3) simulation using different machine learning models
表 3. 采用不同机器学习模型对 3 号传感器处地面空气二氧化碳浓度(C_3)模拟的评价指标

模型	输入	训练集			测试集		
		MAE (μL/L)	RMSE (μL/L)	KGE	MAE (μL/L)	RMSE (μL/L)	KGE
MLR	input1	4.71	6.66	0.98	10.19	13.37	0.76
	input2	4.60	6.56	0.98	9.48	12.22	0.75
MLP	input1	4.99	6.99	0.80	10.11	13.51	0.80
	input2	4.90	6.95	0.94	8.99	11.99	0.83
Bi-LSTM	input1	11.75	15.99	0.80	15.56	18.40	0.81
	input2	11.76	15.43	0.81	23.38	26.74	0.83
RF	input1	4.14	5.73	0.98	12.00	14.76	0.90
	input2	3.72	5.26	0.99	11.48	14.19	0.92

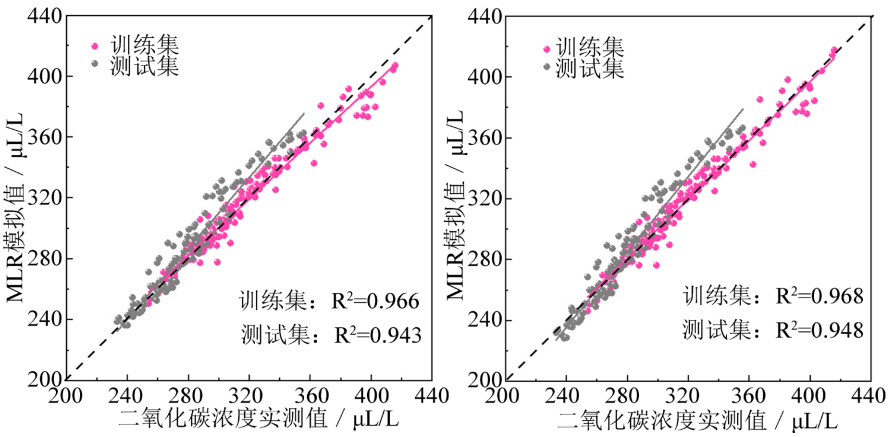

Figure 10. The scatter plot of C_3 measured value and simulated values of different machine learning models (input2)
图 10. C_3 实测值与不同机器学习方法模拟值(input2)散点图

四种机器学习方法中，MLP 和 MLR 的模拟效果相对较好，为进一步分析以 input2 作为输入时的整个模拟过程，模拟值与实测值的散点图如图 10 所示，对比图如图 11 所示。从散点图来看，MLP 和 MLR

在训练集和测试集的散点 R^2 都大于 0.9，模拟值与实测值的线性相关性很高。对比模拟值和实测值来看，MLP 和 MLR 可以较好地模拟出空气二氧化碳浓度在研究时段内的整体变化过程[12]。两种机器学习方法都可以较好地模拟出该时段的二氧化碳浓度的较低值和变化过程，但对于二氧化碳浓度较大值的模拟，都出现了一定程度的高估情况，值得提出的是 MLR 的低估和高估情况更为突出。

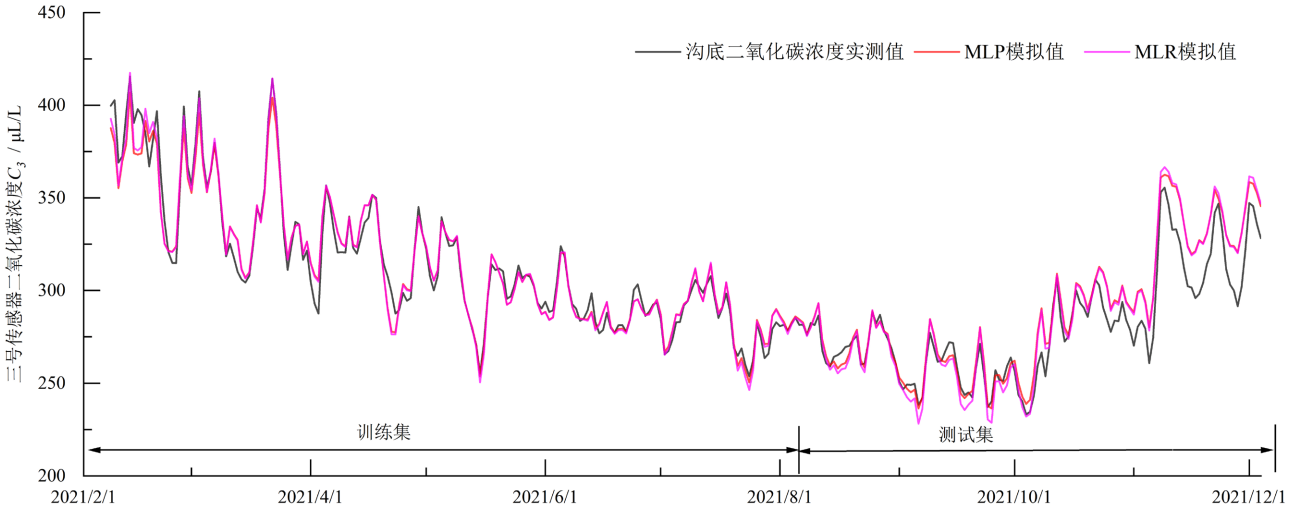

Figure 11. The comparison of measured values of C_3 and simulated values of different machine learning models (input2)
图 11. C_3 实测值与不同机器学习模型模拟值(input2)对比图

4.2. 塬上二氧化碳浓度模拟潜力评价

沟底和塬上存在海拔差，植被覆盖率也存在明显差异。气温作为影响植被生长的关键环境因子之一，直接影响植被覆盖率。气温随海拔变化也会发生变化。基于在沟底的前期模拟结果，MLP 和 MLR 的模拟精度相对较高。因此以沟底二氧化碳浓度(C_2)，结合塬上气温(T_{air2})，采用 MLP 和 MLR 对塬上二氧化碳浓度进行模拟，模拟的评价指标计算结果见表 4。从训练集上来看，MLP 在训练集的 R^2 大于 0.8，训练效果较好。从测试集来看，MLP 的三个评价指标都要优于 MLR，且 MLP 的 KGE 大于 0.7。值得提出的是，测试集内大部分数据均通过线性插补获得，评价指标的计算仅具有参考性。为了更直观地分析模拟值与实测值，散点图如图 12 所示，对比图如图 13 所示。在训练集上，机器学习方法可以较好地模拟出二氧化碳浓度的变化过程，但在测试集 MLP 和 MLR 仍然对某些时段内二氧化碳浓度的较高值存在低估。对于不同海拔二氧化碳浓度的模拟，在添加塬上气温数据作为输入，MLP 和 MLR 都可以较好的捕捉数据缺失的时段内气温变化和二氧化碳浓度变化特征，能够模拟出在确实数据段二氧化碳浓度的变化过程。综合来看，采用 MLP 用于插补数据效果最好。

Table 4. The evaluation indexes of ground air carbon dioxide concentration on the tableland (C_4) simulation using different machine learning models
表 4. 采用不同机器学习模型对塬上地面空气二氧化碳浓度(C_4)模拟的评价指标

模型	输入	训练集			测试集		
		MAE (μL/L)	RMSE (μL/L)	KGE	MAE (μL/L)	RMSE (μL/L)	KGE
MLP	Input2	10.35	12.83	0.80	11.78	14.63	0.76
MLR	Input2	10.36	12.29	0.90	13.92	17.90	0.61

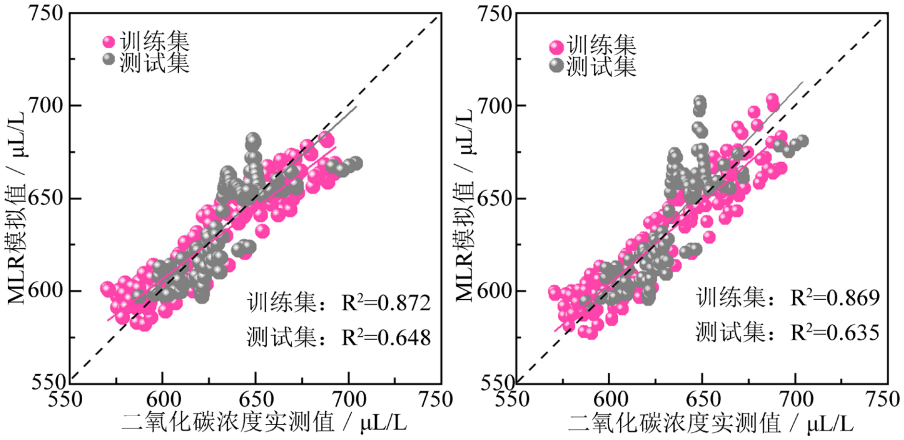

Figure 12. The scatter plot of C_4 measured value and simulated values of different machine learning models
图 12. C_4 实测值与不同机器学习方法模拟值散点图

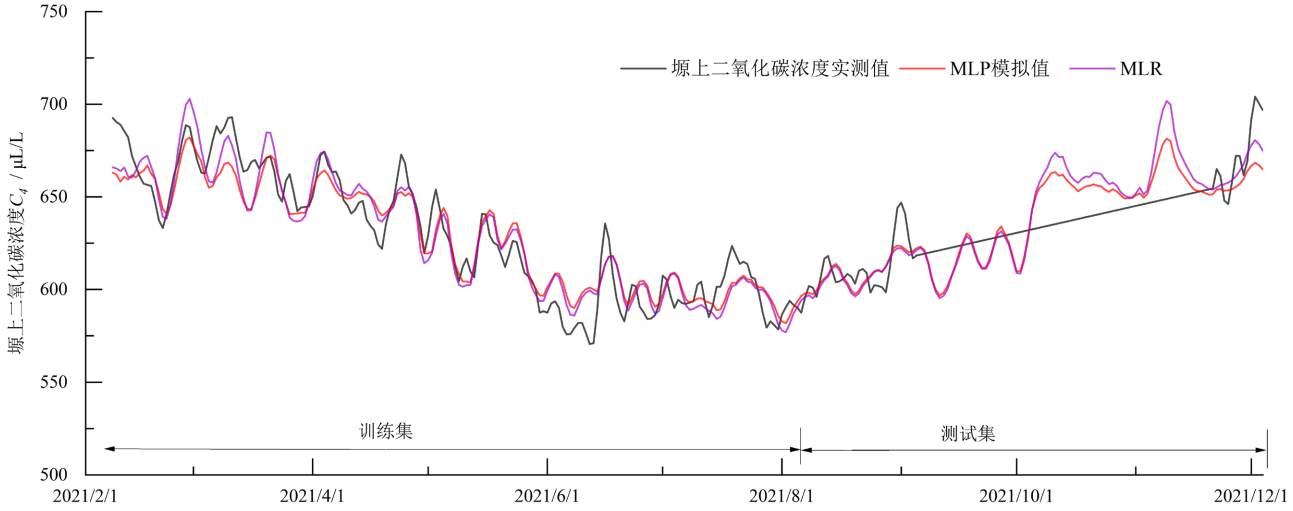

Figure 13. The comparison of measured values of C_4 and simulated values of different machine learning models
图 13. C_4 实测值与不同机器学习模型模拟值对比图

4.3. 定量评价采用机器学习方法插补数据

Table 5. The brief statistical parameters of different data
表 5. 不同数据的简要统计参数

数据	x_{mean}	x_{max}	x_{min}	x_{std}	C_v	C_S	C_k
C_{MLP}	642.1	681.5	596.8	21.8	0.0	−0.4	−0.9
C_{2022}	664.1	696.5	622.1	20.6	0.0	−0.4	−1.1
C_{2024}	653.1	710.1	610.2	26.8	0.0	0.3	−0.9

为进一步分析数据缺失时段(2021 年 9 月 6 日至 11 月 20 日)的数据质量，计算了 2021 年塬上二氧化碳浓度数据缺失段模拟值，计算了 2022 年同期实测二氧化碳浓度和 2024 年同期实测二氧化碳浓度的统计参数，评价 MLP 和 RF 插补数据的质量，确定最优的插补数据。这些数据中，2021 年缺失时段的 MLP 模拟值命名为 C_{MLP}，2022 年同期观测值数据命名为 C_{2022}，2024 年同期观测值数据命名为 C_{2024}。这些数

据的散点图见图 11，统计参数计算结果见表 5。散点图显示，C_{MLP} 与实测值的相关系数最小为 0.73。从统计参数的计算结果来看，这些数据的标准差、变差系数、偏度、峰度相差不大，主要差异出现在均值、最大值、最小值的计算结果。其中，MLP 模拟值和实测序列的均值、最大值的偏小 3%，最小值约偏小 5%。综合来看，插补后的数据与 2022 年、2024 年同期实测数据较为接近，数据质量较高。

5. 结论

本文采用陕西省淳化生态水文基地的塬上和沟底不同观测系统实测的多处二氧化碳浓度和空气温度数据，在沟底以不同输入组合驱动多种机器学习方法对二氧化碳浓度进行模拟，并遴选出模拟效果较好的机器学习模型；尝试以沟底二氧化碳浓度与塬上气温共同作为输入，对塬上二氧化碳浓度的缺失数据进行模拟和插补，并评估插补结果的效果。主要结论如下：

1、对于距离较近，海拔相近，植被覆盖度不同区域的空气二氧化碳浓度，仅以沟底不同位置的实测二氧化碳浓度作为输入，机器学习方法的模拟精度是可接受的。在增加空气温度作为输入后，三种机器学习方法中，Bi-LSTM 的模拟精度降低，MLP、RF 和 MLR 则可以较好地捕捉空气温度和二氧化碳浓度之间的关系，模拟精度有所提高。

2、对于不同海拔、植被覆盖度不同区域，以沟底实测二氧化碳浓度和塬上实测空气温度作为输入，采用 MLP 和 MLR 模拟塬上二氧化碳浓度是可行的。机器学习方法能够模拟出缺失数据段二氧化碳浓度的变化，但对某些时段内二氧化碳浓度的较高值存在低估。

3、以沟底实测二氧化碳浓度数据作为输入，采用机器学习方法插补塬上空气二氧化碳浓度数据是可行的。值得注意的是，MLP 对某些时段内二氧化碳浓度的均值、极大值和极小值存在低估，但误差最大仅 5%，插补数据质量较高。

在淳化生态水文基地出现了逆温现象，且塬上温度变化比沟底更为剧烈。2021 年 11 月下旬塬上气温出现了一次短暂降温过程，温度降低，植物光合作用减弱，二氧化碳浓度会短暂升高，这直接导致 MLP 在该时段内对二氧化碳浓度出现了明显的低估。本文采用了陕西省淳化生态水文基地的通量塔塬上和沟底的数据，是研究区内植被覆盖度和海拔差异最大的观测值，未来可以采用距离更远的站点的数据进行模拟，进一步探索采用机器学习方法模拟跨区域二氧化碳浓度的潜力，更经济且科学了解区域碳变化过程。

基金项目

国家自然科学基金(52279025)；国家重点研发计划项目(2022YFF1302200)。

参考文献

[1] 傅伯杰, 刘彦随, 曹智, 等. 黄土高原生态保护和高质量发展现状、问题与建议[J]. 中国科学院院刊, 2023, 38(8): 1110-1117.

[2] 张西宁, 郭文遥, 曹丹, 等. 水土保持与县域善治: 黄土高塬沟壑区的治理范式与创新实践[J]. 中国水土保持, 2026(1): 22-24.

[3] Ding, M., Flaig, R.W., Jiang, H. and Yaghi, O.M. (2019) Carbon Capture and Conversion Using Metal-Organic Frameworks and MOF-Based Materials. *Chemical Society Reviews*, **48**, 2783-2828. https://doi.org/10.1039/c8cs00829a

[4] 李恩, 侯锐, 侯方玲, 等. 高用水工业节水减碳分析: 以江苏省为例[J]. 水利水电技术(中英文), 2025, 56(8): 49-60.

[5] Noor, S., Jiang, X., Wang, X., Yang, J., Newman, S., Li, K., *et al*. (2026) Human-Induced Biospheric Carbon Sink: Impact from the Taklamakan Afforestation Project. *Proceedings of the National Academy of Sciences*, **123**, e2523388123. https://doi.org/10.1073/pnas.2523388123

[6] 郜国明, 喻权刚, 徐佳. 黄土高原水土保持碳汇研究及其价值实现路径[J]. 中国水利, 2024(23): 12-19.

[7] Solanki, H., Vegad, U., Kushwaha, A. and Mishra, V. (2025) Improving Streamflow Prediction Using Multiple Hydrological Models and Machine Learning Methods. *Water Resources Research*, **61**, e2024WR038192. https://doi.org/10.1029/2024wr038192

[8] 何志远, 钟九生, 代仁丽. 基于机器学习的综合干旱监测建模及在西南地区应用[J]. 水利水电技术(中英文), 2022, 53(2): 43-51.

[9] 陈思宇, 李肖男, 花续, 等. Kolmogorov-Arnold 网络在长江中下游水位预报中的应用[J]. 水力发电学报, 2025, 44(4): 97-107.

[10] 缪云飞. 基于机器学习的大气二氧化碳卫星反演研究[D]: [硕士学位论文]. 合肥: 安徽大学, 2023.

[11] 赵嘉宁, 牛振川, 梁单, 等. 秦岭北麓大气 CO_2 浓度的分层观测与通量估算研究[J]. 环境科学学报, 2026, 46(5): 11-21.

[12] 邓炜, 刘登峰, 李明亮, 等. 机器学习方法模拟黄土高原沟壑区二氧化碳浓度的潜力评估[J]. 生态学报, 2025, 45(13): 6559-6575.

[13] 郝旺林, 夏彬, 许明祥. 黄土丘陵区典型沟道侵蚀诱发的 CO_2 通量估算[J]. 水土保持学报, 2022, 36(6): 179-188.

[14] 张奎月, 刘登峰, 刘慧, 等. 黄土高原灌丛生态系统土壤呼吸特征及其影响因素[J]. 人民珠江, 2022, 43(4): 83-94.

[15] 郭凤年. 黄土高原典型灌丛系统蒸散发过程模拟与变化规律分析[D]: [硕士学位论文]. 西安: 西安理工大学, 2023.

[16] 邓炜. 基于机器学习的黄土高原沟壑区小流域水热碳过程模拟[D]: [硕士学位论文]. 西安: 西安理工大学, 2024.

[17] 徐月月, 何清, 毛东雷, 等. 2022-2023 年中昆仑山北坡不同海拔气象要素梯度对比分析[J]. 高原气象, 2025, 44(1): 224-239.

[18] 刘华玲, 马俊, 张国祥. 基于深度学习的内容推荐算法研究综述[J]. 计算机工程, 2021, 47(7): 1-12.

[19] 陈镜元. 基于机器学习模型的汉江上游径流预测及水文干旱评价[D]: [硕士学位论文]. 咸阳: 西北农林科技大学, 2025.

[20] 王超, 张耀飞, 张社荣, 等. 数字孪生水利监测感知网多参数时序预测模型[J]. 水力发电学报, 2025, 44(9): 73-88.

[21] 彭海波, 戴善进, 李泽华, 等. 基于 CNN-LSTM-Attention 融合模型的东江流域中长期水文预报方法及应用[J]. 中国水利, 2025(8): 61-66.

[22] 牟含笑, 郑建飞, 胡昌华, 等. 基于 CDBN 与 BiLSTM 的多元退化设备剩余寿命预测[J]. 航空学报, 2022, 43(7): 301-312.

[23] 赵安周, 徐瑞皓, 卢彦琦. 基于随机森林的长江中下游流域综合干旱指数构建与应用[J]. 农业现代化研究, 2026, 47(1): 213-224.

[24] 邢贞相, 侯泉滢, 王轶男, 等. 基于群组特征与随机森林算法的大庆地区湖泊健康评价[J/OL]. 水资源保护, 1-16. https://link.cnki.net/urlid/32.1356.TV.20260120.0821.002, 2026-05-26.

[25] 赖成光, 陈晓宏, 赵仕威, 等. 基于随机森林的洪灾风险评价模型及其应用[J]. 水利学报, 2015, 46(1): 58-66.

[26] 谭龙金, 冯震宇, 周家文, 等. 堰塞坝数据库插补及稳定性评价[J]. 水力发电学报, 2025, 44(5): 33-43.

[27] 黄青东智, 孙颖, 杨明新, 等. 基于异速生长方程和随机森林模型的青藏高原不同乔木林类型碳储量估测[J]. 生态学报, 2025, 45(15): 7191-7201.

[28] Belkin, M., Hsu, D., Ma, S. and Mandal, S. (2019) Reconciling Modern Machine-Learning Practice and the Classical Bias-Variance Trade-Off. *Proceedings of the National Academy of Sciences*, **116**, 15849-15854. https://doi.org/10.1073/pnas.1903070116

[29] Bartram, L., Correll, M. and Tory, M. (2022) Untidy Data: The Unreasonable Effectiveness of Tables. *IEEE Transactions on Visualization and Computer Graphics*, **28**, 686-696. https://doi.org/10.1109/tvcg.2021.3114830