DOI: 10.20103/j.stxb.202312312894

张江蕾,陈少辉.基于机器学习与再分析数据集的黄河水源涵养区蒸散发研究.生态学报,2024,44(18):8314-8325. Zhang J L, Chen S H.Evapotranspiration in the water source conservation area of the Yellow River Basin based on machine learning and reanalysis dataset. Acta Ecologica Sinica,2024,44(18):8314-8325.

基于机器学习与再分析数据集的黄河水源涵养区蒸散 发研究

张江蕾^{1,2},陈少辉^{1,*}

1 中国科学院地理科学与资源研究所,陆地水循环及地表过程重点实验室,北京 1001012 中国科学院大学 资源与环境学院,北京 100049

摘要:蒸散发是水循环的关键要素,分析其变化特征有助于理解区域水资源的时空分布格局。黄河水源涵养区是黄河流域重要的生态功能区,对该区域的蒸散发变化特征进行研究并归因分析,有助于缓解黄河流域的水资源供需矛盾。基于机器学习与ERA5-land 再分析数据集,探究黄河水源涵养区 2000—2022 年蒸散发时空变化特征及影响因素,利用驱动要素去趋势方法分析不同影响因素的作用区域。结果表明:(1)黄河水源涵养区蒸散发多年平均值分布区间为 256.49—841.45 mm,空间分布特征为自东向西递减,整体呈增加趋势;(2)黄河水源涵养区蒸散发的主要影响因素是地表净太阳辐射、总降水量、相对湿度,不同子流域内的主导影响因素不同,主导影响因素与区域内的水热条件及下垫面状况有关;(3)ERA5-land 再分析数据集有着较好的模拟精度,可以作为大空间尺度和长时间区间研究的数据来源,但是由于下垫面的复杂性,仍需要在研究区内开展适应性评估。 关键词:黄河流域;蒸散发;机器学习;ERA5-land 再分析数据集;影响因素

Evapotranspiration in the water source conservation area of the Yellow River Basin based on machine learning and reanalysis dataset

ZHANG Jianglei^{1, 2}, CHEN Shaohui^{1, *}

Key Laboratory of Water Cycle and Related Land Surface Processes, Institute of Geographic Sciences and Natural Resources Research, Beijing 100101, China
 College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract: Evapotranspiration is a key element of the water cycle, and analyzing its variations helps understand the spatiotemporal distribution patterns of regional water resources. The water source conservation area of the Yellow River Basin is an important ecological function area in the Yellow River Basin. Studying the characteristics of evapotranspiration changes in this area and conducting attribution analysis can help alleviate the water supply-demand contradictions in the Yellow River Basin. Based on machine learning and the ERA5-land reanalysis dataset, this study explored the spatiotemporal variations and influencing factors of evapotranspiration in the Yellow River water source conservation area from 2000 to 2022. The driving factor regression analysis method is used to analyze the influence of different factors in different regions. The results show that: (1) The multi-year average distribution range of evapotranspiration in the Yellow River water source conservation area is 256.49—841.45 mm, with a spatial distribution characteristic of decreasing from east to west and an overall increasing trend. (2) The main influencing factors of evapotranspiration, and relative humidity. The dominant influencing factors vary in different sub-basins and are related to the hydrothermal conditions and underlying surface conditions in the region.

收稿日期:2023-12-31; 网络出版日期:2024-07-12

基金项目:国家重点研发计划(2021YFC3201102);第二次青藏高原综合科学考察研究(2019QZKK1003);国家自然科学基金(U2003105)

^{*} 通讯作者 Corresponding author.E-mail: chensh@igsnrr.ac.cn

(3) The ERA5-land reanalysis dataset has good simulation accuracy and can serve as a data source for large spatial scale and long-time interval studies. However, due to the complexity of the underlying surface, adaptive assessment within the study area is still needed.

Key Words: Yellow River Basin; evapotranspiration; machine learning; ERA5-land reanalysis dataset; influencing factors

蒸散发(Evapotranspiration, ET)主要包括土壤蒸发、植被蒸腾及植被截留水蒸发,是地表水循环的核心, 联系了陆地生态系统与水文系统,在能量交换中起到重要作用^[1]。准确估算区域蒸散发对于水土保持研究、 区域水资源合理分配及生态安全风险评估有重要意义^[2]。

在探究蒸散发影响因素的方法中,机器学习可以在无先验条件下自主完成要素间隐藏关系的建立,对于 复杂系统有更高的模拟精度^[10]。王秀英等^[11]将机器学习方法运用到蒸散发研究当中,结果表明机器学习可 以处理大量混合类数据,对噪声有着很强的耐受性。钱云平等^[12]对比了机器学习方法和地表能量平衡模型 方法,结果表明,机器学习模型的蒸散发模拟误差仅为物理模型的 1/4。Zhang W B 等^[13]以我国西北干旱、半 干旱区草地下垫面的蒸散发为研究对象,基于机器学习中的随机森林方法构建估算模型,并得到良好的模拟 效果。

黄河水源涵养区是指包括黄河源区、渭河南山支流区、伊洛河流域的地区,这一概念由王国庆学者^[14]于2022年首次提出。黄河水源涵养区位于整个黄河流域的上游和中游,对于当地及下游地区的生产、生活有着举足轻重的作用。由于其概念较新,目前对于该区域的蒸散发研究较为匮乏。因此,本文基于机器学习与ERA5-land 再分析数据集,探究黄河水源涵养区2000—2022年蒸散发时空变化特征及影响因素,利用驱动要素去趋势方法分析不同影响因素的作用区域。研究结果有助于黄河水源涵养区的水资源保护及开发、水资源配置及优化。

1 数据资料与方法

1.1 研究区概况

黄河水源涵养区主要由三部分组成(图1)。其中,黄河源区地处青藏高原东北部,平均海拔3500m以上,面积为22.25万km²,是黄河流域内最大的冰雪融水和降雨补水区^[12],是中国西部地区可持续发展的生态 屏障^[14]。渭河南山支流流域总面积6.33万km²,渭河流域是黄河泥沙的重要来源地,也是我国水土的重点保 护区^[15]。伊洛河是黄河十大支流之一,由伊河和洛河组成,流域面积为1.86万km²。黄河水源涵养区横跨青



海、四川、甘肃、宁夏回族自治区、陕西和河南省,其地势西高东低,流域落差大,下垫面情况较为复杂[3]。

Fig.1 Water conservation area of the Yellow River

1.2 数据来源

研究数据采用欧洲中期天气预报中心(European Centre for Medium-Range Weather Forecasts, ECMWF) 2000—2022 年 ERA5-land 数据,空间分辨率为 0.1°,时间分辨率为月。ERA5 是欧洲中期天气预报中心全球 气候天气再分析的第五代产品,使用了 ECMWF 集合预报系统(Integrated Forecast System, IFS)中 CY41R2 的 模型预报结果以及四维变分数据同化技术(4D-Var data assimilation)^[16]。ERA5-land 则是 ERA5 陆地部分的 重演,由考虑水文过程的 ERA5 的气象强迫场驱动陆面模式计算得来,其拥有更高空间分辨率的同时考虑了 网格间高程的差异^[17]。

已有大量研究将 ERA5-land 再分析数据集与观测资料进行对比,发现 ERA5-land 数据精度高^[18-19],可以 提供更精确及更高时空分辨率的数据^{[16][20]}。由于下垫面的非均匀性,ERA5-land 再分析数据集在应用前需 要进行精度评估^[21]。为探究 ERA5-land 数据在研究区的适用状况,在 2022 年 7 月对研究区的 789 个观测点 进行实地调查(图 2),记录每个观测点的经纬度、高程、地表覆盖类型、温度、大气条件等数据,计算观测点的 参考作物蒸散发。计算方式采用世界粮农组织(FAO)修正后(1998 年) Penman-Monteith 公式,FAO Penman-Monteith 公式如下^[22]:

$$ET_{0} = \frac{0.480(R_{n}-G) + \gamma \frac{900}{T+273}u_{2}(e_{s}-e_{a})}{\Delta + \gamma(1+0.34 u_{2})}$$
(1)

式中: ET_0 为参考作物蒸散量,mm/d; R_a 为净辐射,MJm⁻² d⁻¹;G为土壤热通量,MJ m⁻² d⁻¹,日尺度记为0,忽略 不计; e_s 为饱和水汽压,kPa; e_a 为实际水汽压,kPa; Δ 为饱和水汽压—温度曲线斜率,kPa/ \mathbb{C} ; γ 为温度计常数, kPa/ \mathbb{C} ; u_2 为2 m 处的风速,m/s。T为日平均温度, \mathbb{C} 。除以上实地测量数据外,FAO Penman-Monteith 公式 中其他输入变量涉及多个公式及中间变量的计算,由以上基础气象数据计算得到。

在得到实地观测点的参考作物蒸散发后,根据国际粮农组织作物系数报表,结合实际地表覆盖类型调查结果,计算作物实际蒸散发。计算公式如下:

$$ET_c = K_c \times ET_0 \tag{2}$$

式中:ETc为作物实际蒸散发,ETo为参考作物蒸散发,Kc为作物系数,是一个无量纲的衰减系数,它是各种阻

抗的综合影响系数,包括表面阻抗、气孔阻抗和扩散阻抗。作物系数依据 FAO 推荐作物种类的作物系数,结 合实地调查点的地表覆盖类型确定,不同地表覆盖类型的作物系数见下表(表 1)。

Table 1 Crop coefficients for unrecent surface cover types at observation sites				
作物种类 Crop types	作物系数 Crop coefficients	作物种类 Crop types	作物系数 Crop coefficients	
草地 Grassland	0.6	阔叶林	0.65	
红薯 Sweet potato	0.4	竹	1	
小麦 Wheat	0.7	无作物	0.1	
辣椒 Pepper	0.7			
玉米 Corn	0.8			
青稞 Highland barley	0.7			
灌木丛 Shrubs	0.5			
针叶林 Coniferous forest	1			

数字高程模型(DEM)数据来源于中国科学院资源 环境科学与数据中心(http://www.resdc.cn),空间分辨 率为 250 m。

1.3 研究方法

1.3.1 多元线性回归算法

多元线性回归(MLR)算法是机器学习中的一种算法,目的是在已有数据集上通过2个或2个以上自变量构建线性模型来拟合该数据集特征向量的各个分量之间的关系,找到一条直线或者一个超平面来实现预测值



Fig.2 Distribution of observation points

和真实值之间的误差最小化^[11]。线性模型的目标函数(损失函数)表示为:

$$J(\vec{\theta}) = \frac{1}{2} \sum_{i=1}^{m} (h_{\vec{\theta}}(x^{(i)}) - y^{(i)})^2$$
(3)

式中: θ 为参数向量, m 为样本数, y⁽ⁱ⁾ 为第 i 个训练样本的真实值, h_θ(x⁽ⁱ⁾) 为第 i 个训练样本特征值组合预测 函数。线性回归主要应用于回归类问题, 实现比较简单, 结果有很好的可解释性。

1.3.2 决策回归树算法

决策树(DT)算法是机器学习中的一种算法,属于监督学习的一种,是一种树型结构。其结构由节点和有向边组成,具体可分成根节点、内部节点和叶子节点,典型算法包括: ID3 算法、CART 算法、SLIQ 算法、CLS 算法等,CART 决策树通过构建树、修剪树、评估树来构建一个二叉树。当终点节点是连续变量时,该树为回归树;反之为离散变量时,为分类树^[11]。

1.3.3 随机森林算法

随机森林(RF)算法是机器学习中的一种算法,是利用集成学习的思想将多颗独立决策树整合。其基本 单元是决策树,这种融合结构可以解决决策树泛化能力弱的问题。随机森林的本质是从样本中采样选出 n 个 样本,从所有属性中随机选择 k 个属性,选择最佳分割属性作为节点建立 CART 决策树,重复以上步骤 m 次, 建立 m 棵 CART 决策树,进而形成随机森林,通过投票表决结果,决定最终预测结果^[22–25]。

随机森林算法适合处理多维数据,既可以处理离散型数据,也可以处理连续性数据,训练速度快,抗噪声能力良好,准确度较高。缺点是当数据的噪声较大时会出现过拟合,其模型的参数也比较复杂^[26]。

1.3.4 支持向量回归算法

支持向量回归算法(SVR)是机器学习中的一种方法,主要是通过升维后在高维空间中构造线性决策函数

来实现线性回归。支持向量机算法可分为线性可分支持向量机、线性支持向量机和非线性支持向量机。给定 训练数据集,通过间隔最大化得到分离超平面,则得到相应的决策函数。支持向量算法复杂度是由支持向量 个数决定,并非由数据维度决定,所以算法本身相比其它型不太容易过拟合^[11]。

1.3.5 多层感知机算法

多层感知机(MLP)也称为人工神经网络,是机器学习中的一种方法,除了输入输出层,它中间可以有多 个隐层,最简单的 MLP 只含一个隐层,即三层的结构。MLP 所有的参数就是各个层之间的连接权重以及偏 置,因此模型训练目的就是求解最佳参数,首先随机初始化所有参数,然后迭代地训练,不断地计算梯度和更 新参数,直到满足某个条件时(比如误差足够小、迭代次数足够多时)为最优^[11]。

1.3.6 Theil-Sen Median 趋势分析

使用稳健的非参数化趋势度(sen)分析方法,并通过 Mann-Kendall 统计检验法对蒸散发及影响因素长期 变化的显著趋势进行解释分析^[27-29]。Sen 趋势度(*p*)的计算公式为:

$$p = \operatorname{median}\left(\frac{x_{i} - x_{i}}{j - i}\right) + \frac{1}{i < j < n}$$

$$\tag{4}$$

式中:x_j,x_i为蒸散发时间序列,当p为负时,表示时间序列蒸散发呈下降趋势,当p为正时,时间序列蒸散发表 示时间序列呈上升趋势。

由于 Mann-Kendall 统计检验法具有样本数据不需要遵从一定的分布,适用于非正态分布的数据,能够剔除少数异常值等优点。因此,Mann-Kendall 统计检验法是气象学、水文学、气候学中比较常用的时间序列趋势检验方法^[30-32]。Mann-Kendall 检验的公式为:

$$Q = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \operatorname{sign}(x_j - x_i)$$
(5)

$$\operatorname{sign}(x_{j}-x_{i}) = \begin{cases} 1 & (x_{j} - x_{i} - 0) \\ 0 & (x_{j}-x_{i} = 0) \\ -1 & (x_{j}-x_{i} < 0) \end{cases}$$
(6)

$$Z = \begin{cases} \frac{Q-1}{\sqrt{\operatorname{var}(Q)}} & (Q>0) \\ 0 & (Q=0) \\ \frac{Q+1}{\sqrt{\operatorname{var}(Q)}} & (Q<0) \end{cases}$$
(7)

式中:Q为检验统计量;Z为标准化后的检验统计量;x_j,x_i为蒸散发时间序列数据;n为样本数,当 n>8 时,Q 近似为正态分布,其均值和方差计算公式为:

$$E(Q) = 0 \tag{8}$$

$$\operatorname{var}(Q) = \frac{n(n-1)(2n-5)}{18} \tag{9}$$

式中:标准化后 Z 为标准正态分布,双边的趋势检验中,在给定 α 置信水平上,如果 | Z | ≥ Z_{1-a/2}则原假设是不可接受的。即 α 置信水平上, | Z | 大于等于 1.645,1.96,2.576 时,分别表示通过了可信度 90%,95%和 99%的 显著性检验。对于统计量 Z 大于 0 时是上升趋势, Z 小于 0 时是下降趋势。本次研究的 α 置信水平取 0.05, 当 Z 大于 0 且 | Z | 大于等于 1.96 时,认为目标要素明显增加、当 Z 大于 0 且 | Z | 小于 1.96 时,认为目标要素轻微增加;当 Z 小于 0 且 | Z | 大于等于 1.96 时,认为目标要素明显减少、当 Z 小于 0 且 | Z | 小于 1.96 时,认为目标要素轻微减少。基于 Theil-Sen 趋势分析和 Mann-Kendall 检验分析原理,对蒸散发和主要影响因素进行逐像元计算分析。

1.3.7 驱动要素去趋势分析

为探究单一影响因素对蒸散发的影响,使用去趋势分析法对机器学习筛选出的影响因素进行去趋势处

理。再分别将原始和去趋势的驱动要素逐一输入到多影响因素拟合的回归方程中。通过比较原始和去趋势 输入状态下蒸散发的相对变化,来量化各个要素对蒸散发的贡献。经过去趋势处理的驱动要素年趋势为零, 但保留了要素的年际变化和季节性波动^[33]。假设其中一种影响因素为*Y*,那么其去趋势操作步骤如下。

步骤一:在栅格尺度上,将逐月的 Y 数据累加到年尺度,基于 Mann-Kendall 非参数趋势检验获得 Y 的年趋势。

步骤二:基于下列公式去除 Y 年序列趋势:

$$Y(t) = Y_0(t) + k(t_0 - t)$$
(10)

式中: $Y_d(t)$ 和 $Y_0(t)$ 分别为 Y 在第 t 年的去趋势值和原始值, k 为原始 Y 年序列的趋势, t_0 为初始年份(2000年)。

步骤三:分别将原始和去趋势的 Y数据输入到回归方程中,其他驱动要素不要做处理(原始驱动),得到 两种输入状态下蒸散发年趋势的差异(ΔT_y):

$$\Delta T_Y = T_{0Y} - T_{dY} \tag{11}$$

式中:T_{0Y}为输入原始 Y 得到的蒸散发趋势,T_{dY}为输入去趋势 Y 得到的蒸散发趋势。

基于同样的步骤,可以得到其他驱动要素对蒸散发趋势的影响,那么影响因素 Y 对蒸散发趋势的贡献度 计算如下:

$$C_{Y} = \frac{\Delta T_{Y}}{\Delta T_{Y} + \Delta T_{X} + \Delta T_{Z} + \dots + \Delta T_{N}} \times 100\%$$
(12)

式中: ΔT_{Y} 、 ΔT_{X} 、 ΔT_{Z} 、 ΔT_{N} 分别代表影响因素 Y、X、Z、N 引起的蒸散发趋势变化,基于同样的步骤,可以得到其他要素变化对蒸散发趋势的贡献。

2 结果分析

2.1 ERA5-land 数据模拟效果验证

计算 789 个观测点的实际蒸散发值,提取观测点上 对应时间的 ERA5-land 实际蒸散发模拟值,将两种数据 进行线性拟合(图 3)。结果表明两种数据的拟合效果 较好,*R*² 为 0.755,其回归方程为:

 $ET_{ERA5} = 0.7076 \times ET_0 + 1.0579$ (13) 式中, ET_0 为实际蒸散发观测值, ET_{ERA5} 为 ERA5-land 实际蒸散发模拟值。

计算 ERA5-land 实际蒸散发模拟值与实际蒸散发观测值的相关系数,结果为 0.813,有着较强的相关性,

这说明 ERA5-land 数据的实际蒸散发值在研究区有着较好的适用性,可以作为评估气候和植被要素变化对蒸散发影响的数据基础。

2.2 蒸散发时空变化特点

ERA5-land 的实际蒸散发数据表明,黄河水源涵养区 2000—2022 年多年平均实际蒸散发介于 256.49— 841.45 mm(图4),呈现自东向西递减的趋势。实际蒸散发最高值出现在伊洛河流域,最低值出现在黄河源 区,这种显著的分异特征与研究区内地势特点、植被分布、植被类型及复杂气候有直接的联系。研究区东部为 伊洛河流域,降水丰富,地势平坦,植被密集、植被类型丰富,蒸散发强烈;研究区中部为渭河南山支流,该区域 北部为黄土高原,植被覆盖少,降雨量少,因此蒸散发比东部地区少;研究区西部为黄河源区,黄河源区地处高 原,多为基岩裸露山地,植被覆盖少,同时该区域属于水分限制区,降水量少,蒸散发较弱。

从蒸散发变化趋势(图4)上来看,黄河水源涵养区蒸散发呈现增加趋势,蒸散发增加的地区面积大于蒸





散发减少的地区面积。蒸散发增加的区域集中在黄河源区西部、北部及南部,渭河南山支流的中部区域及伊 洛河的西部区域,其余地区蒸散发减少。



图 4 研究区蒸散发多年平均值及变化趋势 Fig.4 Annual average ET and trend of ET in the study area

2.3 蒸散发影响因素筛选

为了探究研究区内影响实际蒸散发的可能因素,将研究区内的 ERA5-land 实际蒸散发多年平均值作为真 实值,将利用机器学习多种算法得到的实际蒸散发值作为模拟值,通过对比不同机器学习算法的模拟效果,确 定模拟效果最好的机器学习算法,利用算法模型中的重要性评估方法,可以得到影响蒸散发的主要因素。

ERA5-land 数据可提供多种变量数据,进行机器学习模拟需要进行变量筛选,根据 Penman-Monteith 公式 筛选出可能对蒸散发造成影响的变量,再根据 10 m 风速纬向 u 分量(u10)和 10 m 风速经向 v 分量(v10)得到 风速(WS),根据露点温度(d2m)和空气温度(t2m)得到相对湿度(RH),最终得到输入变量共计 15 种(表 2)。

Table 2 Variable category table						
变量类别		变量类别		变量类别		
Variable category		Variable category		Variable category		
总降水量	tn	风速	WS	土壤温度 (100—289 cm)	st]4	
Total precipitation	P	Wind speed		Soil temperature (100—289 cm)	541	
地表净太阳辐射	cer	相对湿度	RH	土壤含水量(0-7 cm)	ewyl1	
Surface net solar radiation	551	Relative humidity	iui	Volumetric soil water(0-7 cm)	50011	
空气温度	t2m	土壤温度(0-7 cm)	atl1	土壤含水量(7-28cm)		
2m temperature	ιΔm	Soil temperature(0-7 cm)	sui	Volumetric soil water(7 - 28cm)	SWVIZ	
地表温度	ماله	土壤温度 (7 —28cm)		土壤含水量 (28—100cm)	a	
Skin temperature	SKI	Soil temperature(7 — 28cm)	SU2	Volumetric soil water (28-100cm)	SWVIJ	
叶面积指数	TAT	土壤温度 (28—100cm)	-112	土壤含水量 (100—289cm)	14	
Leaf area index	LAI	Soil temperature (28—100cm)	sub	Volumetric soil water(100-289cm)	SWV14	

表 2 变量类别表

2.4 多种机器学习算法模拟精度

以筛选出的 15 种变量的多年平均值为输入变量,进行多种机器学习算法模拟,计算模拟结果的决定系数 (*R*²),多种算法的模拟结果如下(表 3)。结果表明随机森林(RF)算法的模拟效果最好,*R*²为 0.855,多层感 知机(MLP)的模拟效果最差,*R*²为 0.716。因此选用随机森林算法进行影响因子重要性评估,在此之前,为了 提高随机森林回归模型的精度,需要对随机森林算法参数进行调整进一步提高其模拟精度。

Table 3 Simulation accuracy of multiple machine learning algorithms				
机器学习算法名称 Machine learning algorithm name	R^2	机器学习算法名称 Machine learning algorithm name	R^2	
多元线性回归 Multiple Linear Regression (MLR)	0.843	支持向量回归 Support Vector Regression (SVR)	0.772	
决策树 Classification And Regression Tree (CART)	0.832	多层感知机 Multilayer Perceptrons (MLP)	0.716	
随机森林 Random Forest (RF)	0.855			

表 3 多种机器学习算法模拟料	情度
-----------------	----

2.5 随机森林算法模型参数调优

影响随机森林模拟精度的参数主要有 3 种,分别为最大迭代次数(n_estimators)、最大深度(max_depth)、 最大特征个数(max_features)。其中,最大迭代次数(n_estimators)这一参数决定了模型的学习能力,值越大, 模型的拟合效果越好。但是任何模型都有决策边界,当该参数达到一定大小后,模型的精确性不再上升而是 出现波动,简单来说,该参数过小时会出现欠拟合,而过大时则会出现过拟合。此外该参数越大所需要的计算 量和内存就越大,训练时间也会变长,因此该参数的确定需要在训练难度和模拟效果之间取得平衡。最大深 度(max_depth)是决策树的深度,代表了决策树的复杂程度,值越高模型就越复杂,在模型样本量多且特征也 多的情况下,通常会限制最大深度。最大特征个数(max_features)是决策树每次生成时,所使用的最大特征数 量,超过限制个数的特征都会被舍弃。该参数是一种既能让模型更简单,也能让模型更复杂参数。本文中随 机森林的初始参数设置为,最大迭代次数(n_estimators)为 200、最大深度(max_depth)为 10、最大特征个数 (max_features)为10。

采用交叉验证法(GridSearchCV)寻找最优参数,该方法缩减了超参数的值的空间,只搜索特定超参数和 有限的固定值,同时结合交叉验证方法确定最优超参数。基于此方法确定的各参数值分别为:最大迭代次数 (*n_estimators*)为500、最大深度(max_depth)为15、最大特征个数(max_features)为10。在优化模型参数后,计 算模型新的决定系数(*R*²),*R*² 由 0.855 提升为 0.932。随机选取相同的测试点进行误差检测,检测结果如下 (图 5、图 6),结果表明优化参数后,模拟误差变小,模型的模拟效果更好。



2.6 影响因素重要性程度评估

利用随机森林算法模型的重要性计算方法,对参与模拟的 15 种输入变量进行重要性程度计算,计算结果如下(表 6)。根据计算结果,tp(总降水量)、skt(地表温度)、t2m(空气温度)、ssr(地表净太阳辐射)、RH(相对湿度)、LAI(叶面积指数)、WS(风速)7 种因素对蒸散发有较高的重要性程度且重要性程度远大于其它 8 种因素。

Table 6 Sort the importance level of input variables					
影响因素 Influence factor	重要性程度 Importance level	影响因素 Influence factor	重要性程度 Importance level	影响因素 Influence factor	重要性程度 Importance level
tp	0.141950	LAI	0.106085	swvl4	0.026554
skt	0.117350	WS	0.104986	stl3	0.025387
t2m	0.108645	stl2	0.029953	swvl2	0.023067
ssr	0.108470	stl4	0.029840	swvl1	0.021294
RH	0.107626	stl1	0.027535	swvl3	0.021258

表 6 输入变量重要性程度排序

以实际蒸散发为因变量,以7种主要影响因素为自变量,进行多元线性回归模拟,得到回归方程以进行下 一步分析,回归方程为:

 $e = 59.114 - 1.143 \times tp - 0.145 \times ssr + 0.744 \times WS - 0.189 \times RH - 0.097 \times LAI - 0.772 \times t2m + 0.514 \times skt$ (14)

式中:e为实际蒸散发, tp为总降水量, ssr为地表净太阳辐射, WS为风速, RH为相对湿度, LAI为叶面积指数, t2m为空气温度, skt为地表温度。

2.7 不同区域的蒸散发主导影响因素分析

利用去趋势分析方法计算上述影响因素对蒸散发的影响程度,结果表明 ssr(地表净太阳辐射)、RH(相对湿度)、tp(总降水量)是研究区蒸散发变化的主导因素,它们的贡献程度分别是 40.73%、30%、26.97%。其余影响因素对蒸散发的影响有限,它们的贡献程度分别为 skt(地表温度)1.02%、WS(风速)0.25%、t2m(空气温度)0.88%、LAI(叶面积指数)0.14%。

在空间上,蒸散发变化的主导因素有着明显的区域分异规律(图7)。伊洛河流域的主要影响因素是降水量,从总降水量的变化趋势上来看(图7),2000—2022年伊洛河流域的总降水量呈现减少的趋势,输入水循环的总水量减少,由此导致了蒸散发减少。



图 7 蒸散发主要影响因素及其变化趋势 Fig.7 Main influencing factors and their changing trends of ET

渭河流域南山支流蒸散发的主导影响因素是总降水量、地表净太阳辐射及相对湿度。结合总降水量 (图7)、地表净太阳辐射(图7)、相对湿度(图7)的变化趋势来看,在3种影响因素对蒸散发变化起绝对影响

http://www.ecologica.cn

作用的区域,蒸散发的变化趋势与总降水量、地表净太阳辐射的变化趋势基本相同,与相对湿度的变化趋势相反。这是因为在一定温度条件下,相对湿度越高,蒸汽压达到饱和的程度就越高,蒸汽压饱和差小,蒸散发就 越慢。

黄河源区多种影响因素均有体现,这是由于黄河源区下垫面状况及水热条件较为复杂。黄河源区主导影 响因素是总降水量、地表净太阳辐射及相对湿度。蒸散发的变化趋势与总降水量、地表净太阳辐射的变化趋 势基本相同,与相对湿度的变化趋势相反。

综合来看,总降水量对蒸散发的贡献度自东向西逐渐减少,地表净太阳辐射和相对湿度的贡献度在伊洛 河流域没有体现,自渭河流域南山支流向西逐渐增高,并且在黄河源区成为影响蒸散发的主要因素。

3 讨论

根据已有研究,20世纪80年代以来,中国的气温持续上升,植被大范围变绿^[34],这与本文的研究结论一致(图7)。在气候和下垫面要素均变化的影响下,已有研究证明中国陆地蒸散发呈现显著增加的趋势^[33],该结果与本文的研究结果存在差异,这是因为两者研究的时间尺度不同,所依据的数据源也有所不同。但是该研究表明,黄河水源涵养区的主要影响因素是降水量、水汽压差(表征温度和相对湿度造成的影响)、叶面积指数,这与本文的研究结果一致。

在机器学习对蒸散发的研究上,已有研究表明 MLR、DT、RF、SVR、MLP 模型均能较好的估算蒸散发,其中基于集成学习理念的 RF 模型有着更好的模拟效果^[35],这与本文的研究结果一致。在对影响蒸散发的主要影响因子捕捉上,受限于时间尺度及研究区域的不同,已有研究对于首要影响因子的结论不同,但均表明净辐射、降水、气温、相对湿度是影响蒸散发的主要影响因子^{[11][35]},这与本文的研究结果一致。

对于蒸散发,黄河水源涵养区的不同子区域有着不同的主导因素。伊洛河流域的水热条件较好,但是总 降水量减少导致输入水循环的水量减少,成为制约蒸散发的主要影响因素。在渭河流域南山支流,总降水量 增加的区域,净太阳辐射减少,由此导致地表温度与空气温度下降,大气蒸发减弱,温度成为制约蒸散发的主 要影响因素。黄河源区的主导因素较多,该区域地势复杂,不同区块的水热条件差异较大,降水和温度均有成 为主导因素的区域。需要注意的是,相对湿度在黄河源区的较大区域成为了主导因素,这是因为黄河源区地 表植被经历了显著的变绿过程。植被蒸腾作用对于相对湿度较为敏感,植被变绿导致更多的太阳辐射被植被 截获,土壤蒸发量小于植被蒸腾量。此外,相对湿度的变化对于土壤蒸发本身也有较为明显的影响,因此,相 对湿度在较大区域内成为了主导因素。

本文的研究结果存在以下不确定性来源:第一,所使用的实际蒸散发数据来源于 ERA5-land 再分析数据 集,虽然经过了实地测量数据的验证,但由于蒸散发机理本身的复杂性,不可避免的存在误差。第二,确定的 研究区蒸散发主要影响因素(总降水量、地表净太阳辐射、相对湿度)并不是完全独立的,单一影响因素对蒸 散发的影响都存在不同程度的其他影响因素的贡献。第三,未考虑人类活动对蒸散发的影响,人类活动如退 耕还林、农业灌溉、水利工程等均会对下垫面产生影响,因此也会对蒸散发产生影响。

4 结论

(1)黄河水源涵养区蒸散发多年平均值分布区间为256.49—841.45 mm,空间分布特征为自东向西递减。 从变化趋势上来看,黄河水源涵养区的蒸散发总体增强,蒸散发增加的区域分布在黄河源区西部、北部及南 部、渭河南山支流的中部及伊洛河流域的西部,其余地区的蒸散发减少。

(2)黄河水源涵养区影响蒸散发的主要因素是地表净太阳辐射、总降水量、相对湿度、地表温度、风速、空 气温度、叶面积指数。不同子流域内的主导影响因素不同,黄河源区的主导因素是总降水量、地表净太阳辐射 及相对湿度;渭河南山支流的主导因素是总降水量、地表净太阳辐射及相对湿度;伊洛河流域的主导影响因素 是总降水量。流域内的主导影响因素取决于该流域的水热条件及下垫面状况。 (3) ERA5-land 再分析数据集有着较好的模拟精度,可以作为大空间尺度和长时间区间研究的数据来源。 但是由于下垫面的非均匀性,ERA5-land 再分析数据集在不同区域的适应性差异较大,应用于不同研究区域 前需要进行精度评估。

(4)随机森林算法易于训练,可以处理大量混合类型数据,模拟精度较高,还可以提供输入因素的重要性 得分。但是其模拟精度受到参数的调控,因此应用于不同研究前需要将其调整为最佳参数。

参考文献(References):

- [1] Yu D Y, Li X Y, Cao Q, Hao R F, Qiao J M. Impacts of climate variability and landscape pattern change on evapotranspiration in a grassland landscape mosaic. Hydrological Processes, 2020, 34(4): 1035-1051.
- [2] Wang F J, Liang W, Fu B J, Jin Z, Yan J W, Zhang W B, Fu S Y, Yan N N. Changes of cropland evapotranspiration and its driving factors on the Loess Plateau of China. The Science of the Total Environment, 2020, 728: 138582.
- [3] 鞠琴, 刘小妮, 刘娣, 申同庆, 谷黄河, 王国庆, 余钟波. 基于遥感和通量观测的实际蒸散发时空特征——以黄河流域水源涵养区为例. 水科学进展, 2023: 1-10. (2023-11-24). https://kns.cnki.net/kcms/detail/32.1309.P.20231123.1146.004.html.
- [4] 韩松俊, 王旭, 刘亚平, 田富强. 青藏高原潜在蒸发时空变化的南北分异特征. 水科学进展, 2023, 34(4): 490-498.
- [5] 张志胤,李瑞平,王秀青,王燕鑫,贾博中.河套灌区蒸散发时空变化特征及驱动力分析.生态科学,2023,42(4):9-17.
- [6] 翁升恒,张方敏,卢燕宇,段春锋,倪婷.淮河流域蒸散发时空变化与归因分析.生态学报,2022,42(16):6718-6730.
- [7] 涂晨雨, 贾绍凤, 朱文彬, 吕爱锋, 官云兰. 柴达木盆地蒸散发遥感估算与耗水有效性评价. 生态学报, 2022, 42(13): 5404-5415.
- [8] 詹云军,章文,严岩,王辰星,荣月静,朱捷缘,卢慧婷,郑天晨.长江流域实际蒸散发演变趋势及影响因素.生态学报,2021,41(17): 6924-6935.
- [9] 杨扬, 王丽娟, 黄小燕, 齐月, 谢蕊. 基于 ERA5-Land 产品的黄河流域蒸散时空变化特征. 干旱气象, 2023, 41(3): 390-402.
- [10] 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.
- [11] 王秀英,陈奇,杜华礼,张睿,马红璐. 基于机器学习的青藏高原高寒沼泽湿地蒸散发插补研究. 植物生态学报, 2023: 1-10. (2023-06-20). https://kns.cnki.net/kcms/detail/11.3397.Q.20230619.1615.004.html.
- [12] 钱云平,林银平,金双彦,宋瑞鹏,蒋秀华.黄河河源区水资源变化分析.水利水电技术, 2004, (5): 8-10.
- [13] Zhang W B, Zha X C, Li J X, Liang W, Ma Y G, Fan D M, Li S. Spatiotemporal change of blue water and green water resources in the headwater of Yellow River Basin, China. Water Resources Management, 2014, 28(13): 4715-4732.
- [14] 王国庆. 黄河流域水源涵养区界定. 水文, 2022, 42(2): 65.
- [15] 杨涛, 阎晓娟, 赵寒森, 王鹏, 朱涛, 蔡浩杰, 左旭刚, 奚仁刚, 张雨莲, 王立社, 吴硕. 渭河流域土地利用类型转换及其对生态空间格局的影响. 中国地质, 2022, 11(22): 1-16.
- [16] Hersbach H, Bell B, Berrisford P, Hirahara S, Horónyi A, Muñoz-Sabater J, Nicolas J, Peubey C, Radu R, Schepers D, Simmons A, Soci C, Abdalla S, Abellan X, Balsamo G, Bechtold P, Biavati G, Bidlot J, Bonavita M, Chiara G D, Dahlgren P, Dee D, Diamantakis M, Dragani R, Flemming J, Forbes R, Fuentes M, Geer A, Haimberger L, Healy S, Hogan R J, Hólm E, Janisková M, Keeley S, Laloyaux P, Lopez P, Lupu C, Radnoti G, Rosnay P D, Rozum I, Vamborg F, Villaume S, Thépaut J N. The ERA5 global reanalysis. Quarterly Journal of the Royal Meteorological Society, 2020, 146(730): 1999-2049.
- [17] Muñoz Sabater J, Dutra E, Agustí-Panareda A, Albergel C, Arduini G, Balsamo G, Boussetta S, Choulga M, Harrigan S, Hersbach H, Martens B, Miralles D G, Piles M, Rodríguez-Fernández N J, Zsoter E, Buontempo C, Thépaut J N. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth system science data, 2021, 13(9): 4349-4383.
- [18] Martens B, Schumacher D L, Wouters H, Muñoz-Sabater J, Verhoest N E C, Miralles D G. Evaluating the land-surface energy partitioning in ERA5. Geoscientific Model Development, 2020, 13(9): 4159-4181.
- [19] Wang C X, Graham R M, Wang K G, Gerland S, Granskog M A. Comparison of ERA5 and ERA-Interim near-surface air temperature, snowfall and precipitation over Arctic Sea ice: effects on sea ice thermodynamics and evolution. The Cryosphere, 2019, 13(6): 1661-1679.
- [20] Wu Z Y, Feng H H, He H, Zhou J H, Zhang Y L. Evaluation of soil moisture climatology and anomaly components derived from ERA5-land and GLDAS-2.1 in China. Water Resources Management, 2021, 35(2): 629-643.
- [21] 叶林媛, 鲁汉, 秦淑静, 张橹, 熊立华, 刘攀, 夏军, 程磊. 长江流域 1960—2019 年蒸发皿蒸发和实际蒸散发演变规律. 水科学进展, 2022, 33(5): 718-729.

- [22] Sheykhmousa M, Mahdianpari M, Ghanbari H, Mohammadimanesh F, Ghamisi P, Homayouni S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 6308-6325.
- [23] 张雷,王琳琳,张旭东,刘世荣,孙鹏森,王同立.随机森林算法基本思想及其在生态学中的应用——以云南松分布模拟为例.生态学报,2014,34(3):650-659.
- [24] 孟欣宁, 焦瑞莉, 刘念夏, 江江, 严中伟, 于爽, 娄晓, 李昊辰, 王立志, 陈亮, 郑子彦, 赵娜. 基于随机森林插值的中亚夏季极端高温变化特征. 干旱区研究, 2020, 37(4): 966-973. DOI:10.13866/j.azr.2020.04.17.
- [25] 杨北萍, 陈圣波, 于海洋, 安秦. 基于随机森林回归方法的水稻产量遥感估算. 中国农业大学学报, 2020, 25(6): 26-34.
- [26] 刘堃,何祺胜,荆琛琳,李金阳,陈丽.基于机器学习的蒸散量插补方法.河海大学学报(自然科学版),2020,48(2):109-115.
- [27] 马建琴,陈阳,郝秀平,李丹丹. 2001—2019 年河南省地表蒸散发时空变化及其影响因素.水土保持研究, 2021, 28(5): 134-141, 151.
- [28] 刘春艳,张科,刘吉平. 1976—2013年三江平原景观生态风险变化及驱动力. 生态学报, 2018, 38(11): 3729-3740.
- [29] 刘世梁,刘琦,张兆苓,邓丽,董世魁.云南省红河流域景观生态风险及驱动力分析.生态学报,2014,34(13):3728-3734.
- [30] Yue S, Pilon P, Cavadias G. Power of the Mann-Kendall and Spearman's rho tests for detecting monotonic trends in hydrological series. Journal of Hydrology, 2002, 259(1/2/3/4): 254-271.
- [31] 石玉琼, 王宁练, 李团胜, 汪晗. 榆林市景观生态风险及其时空分异. 干旱区研究, 2019, 36(2): 494-504.
- [32] 卿凤婷, 彭羽. 基于景观结构的北京市顺义区生态风险时空特征. 应用生态学报, 2016, 27(5): 1585-1593.
- [33] 白鹏, 蔡常鑫. 1982—2019年中国陆地蒸散发变化的归因分析. 地理学报, 2023, 78(11): 2750-2762.
- [34] Zhu Z C, Piao S L, Myneni R B, Huang M T, Zeng Z Z, Canadell J G, Ciais P, Sitch S, Friedlingstein P, Arneth A, Cao C X, Cheng L, Kato E, Koven C, Li Y, Lian X, Liu Y W, Liu R G, Mao J F, Pan Y Z, Peng S S, Peñuelas J, Poulter B, Pugh T A M, Stocker B D, Viovy N, Wang X H, Wang Y P, Xiao Z Q, Yang H, Zaehle S, Zeng N. Greening of the Earth and its drivers. Nature Climate Change, 2016, 6: 791-795.
- [35] 季鹏, 袁星. 基于多种机器学习模型的西北地区蒸散发模拟与趋势分析. 大气科学学报, 2023, 46(1): 69-81.