DOI:10.13878/j.cnki.jnuist.2021.06.008

韩慧敏[1]　沈润平[1]　黄安奇[1]　狄文丽[1]

# 基于集成学习方法的 CLDAS 土壤湿度降尺度研究

**摘要**

土壤湿度为陆地生态系统水循环和能量收支的关键参数,陆面数据同化系统可获得时空连续的土壤湿度数据,但由于空间分辨率较低限制了进一步应用.以华北地区作为研究区,基于单一模型(梯度提升机、深度前馈神经网络和随机森林)以及 Stacking 集成学习方法,对中国气象局陆面数据同化系统(CLDAS-V2.0)0~10 cm 土层土壤湿度产品开展了降尺度研究.2019 年 4—10 月降尺度模型的估算结果表明,0~10 cm 土层土壤湿度空间分辨率从 6 km 降尺度至 1 km,4 种降尺度方法结果均能反映出我国土壤湿度时空变化规律,一定程度上改进 CLDAS 产品高估现象,且空间分布细节更加丰富,精度得到提高,其中以 Stacking 集成学习的降尺度方法最好,降尺度的土壤湿度估算值与站点观测数据的相关性最高($R=0.756\ 8$),并且具有最小的误差(均方根误差为 0.050 5 $m^3/m^3$,偏差为 -0.005 2 $m^3/m^3$).时间上,Stacking 集成学习的降尺度结果同样与实测值的动态变化具有更高的相关性,均方根误差和偏差以 Stacking 集成学习方法最小,其次为随机森林和深度前馈神经网络.

**关键词**

土壤湿度;集成学习;降尺度;CLDAS

**中图分类号** S152.71

**文献标志码** A

**收稿日期** 2021-09-22

**资助项目** 国家重点研发计划(2018YFC1506602);国家自然科学基金(91437220)

**作者简介**

韩慧敏,女,硕士生,主要从事环境遥感方面的研究.1158864532@qq.com

沈润平(通信作者),男,博士,教授,主要从事陆面过程遥感研究.rpshen@nuist.edu.cn

1　南京信息工程大学　地理科学学院,南京,210044

## 0　引言

土壤水分参与地-气水分和能量交换,从而对作物生长、流域水文过程和气候变化产生重要影响[1-3],准确获取土壤湿度时空变化分布信息具有重要意义.近年来,多套基于陆面模式发展起来的土壤湿度陆面同化系统,包括全球陆面数据同化系统(GLDAS,0.25°×0.25°)[4]、北美陆面数据同化系统(NLDAS,0.125°×0.125°)[5]和中国气象局(CMA)陆面数据同化系统(CLDAS,0.062 5°×0.062 5°)[6].土壤湿度产品已被广泛用于陆面分析和气象业务,但是空间分辨率较低,极大地限制了其进一步应用,特别是在需要更高空间分辨率的精确农业管理和干旱监测领域.

为获得更高空间分辨率数据,许多学者对土壤湿度的降尺度方法开展了研究,目前主要有基于卫星遥感数据融合的方法、基于地理信息数据的方法以及基于模型的方法[7].基于卫星遥感数据融合的方法主要使用遥感数据,而不依赖站点资料,适用于区域大尺度的土壤湿度降尺度研究,但受卫星观测时间限制和云层覆盖的影响[8-9].基于地理信息数据的方法主要是利用地形、土壤质地和植被覆盖等参数,建立地理信息与土壤湿度之间的关系,获得高分辨率土壤湿度,但需要大量的实地数据来构建地统计或分形插值模型,从而限制了在较大尺度区域内的应用[10].基于模型的降尺度方法主要包括数理统计模型(如基于地统计学、多重分形或小波)和陆面模型,该类方法应用时需考虑模型关系的时空普适性,以及大量站点数据的输入[11].随着计算机性能的提高和人工智能技术的发展,机器学习方法被引入土壤湿度降尺度研究中,它能在缺乏连续数据的情况下,建立土壤湿度与陆表参数之间的关系,且具有较强的非线性问题学习能力和整合多源数据的灵活性,成为提高土壤湿度空间分辨率的有效技术.但单一机器学习方法易表现出对非线性及表征空间大的数据性能的不足,且易产生过拟合[12],难以全面考虑土壤湿度变化特征,导致估算精度不高、模型鲁棒性低等问题.而集成学习方法能结合多种学习器的优势,具有更高的模型准确性、鲁棒性和整体归纳能力,目前在土壤湿度降尺度研究中还鲜有应用.

CLDAS 是国家气象信息中心研发的中国气象局陆面数据同化系统(CMA Land Data Assimilation System,CLDAS),其土壤湿度产品是利用我国多种资料融合和同化获得的大气强迫数据,驱动多种陆面

模式模拟得到[13].该产品时空连续、不受天气影响,在中国区域的表现优于国际同类产品[14],目前空间分辨率只有 6 km.为此,本文利用梯度提升机、深度前馈神经网络、随机森林以及 Stacking 集成学习方法,以华北地区为例,开展了对 CLDAS 土壤湿度产品进行降尺度研究,将其空间分辨率降尺度至 1 km,以获得高时空分辨率连续的土壤湿度估算.

## 1 研究区域与数据

### 1.1 研究区概况

研究区位于我国华北地区,空间范围为 110°21′~122°43′E,31°23′~41°36′N,包括北京、天津、河北、河南和山东 5 个省(市),占地约 54 万 $km^2$,是我国最主要的粮食产区,耕地面积大(面积占比 71%)(图 1),历史上多次遭受重大干旱,土壤湿度能够直接表征地表水分状态,是关键地表参数.

### 1.2 研究数据

#### 1.2.1 CLDAS 土壤湿度数据

中国气象局陆面数据同化系统(CLDAS-V2.0)土壤湿度资料是基于数据融合和同化技术,利用气温、气压、湿度、风速、降水辐射数据和初始场信息,驱动 CLM 和 Noah-MP 陆面模式集合模拟而获得[15].CLDAS-V2.0 土壤湿度数据垂直分为 5 层,分别为 0~5、0~10、10~40、40~100、100~200 cm,单位为 $m^3/m^3$,产品覆盖东亚区域(60°~160°E,0°~65°N),空间分辨率为 0.062 5°,时间分辨率为逐小时.将每 24 h 的土壤湿度值求平均,获得逐日土壤湿度,研究利用 2019 年 4 月 1 日至 2019 年 10 月 31 日 0~10 cm 土层土壤湿度数据(中国气象数据服务中心:http://data.cma.cn)开展研究.

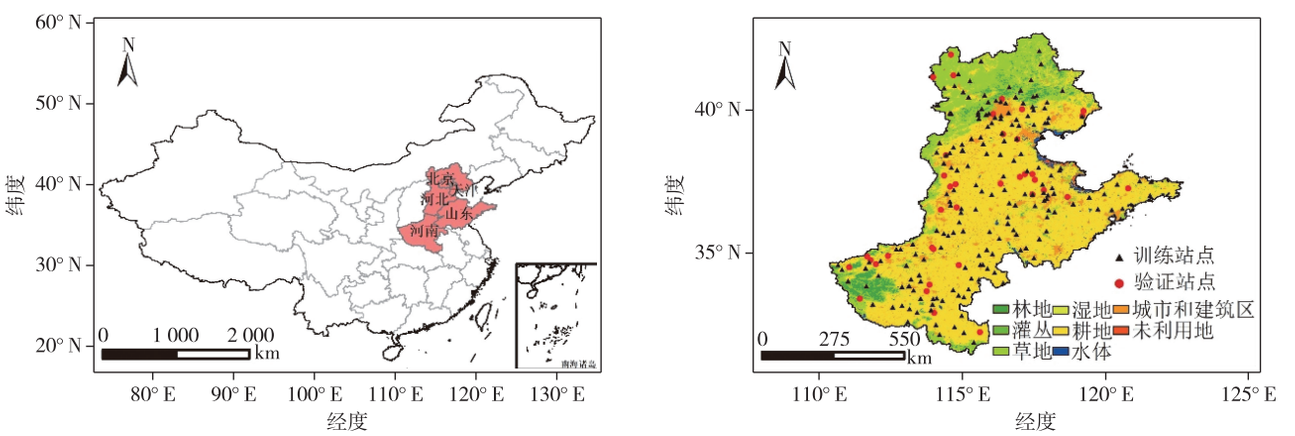
#### 1.2.2 遥感数据

采用 Terra 与 Aqua 卫星的 MODIS 数据(来源于美国国家航空航天局(http://ladsweb.modaps.eosdis.nasa.gov))获得的高分辨率陆面地表参量.时间范围为 2019 年 4—10 月.

1)地表温度数据:来自 MODIS 的 MOD11A1 和 MYD11A1 V6 产品,空间分辨率为 1 km,时间分辨率为 1 d.该数据采用广义分裂窗算法反演得到,同时它能够有效地消除大气的影响[16].为保持土壤湿度数据与地表温度数据的时间一致性,从数据集中提取白天地表温度数据和夜晚地表温度数据,经过投影、系数转换,并对同日次的白天和夜晚地表温度数据进行平均合成,得到逐日平均地表温度数据.

2)地表反照率数据:来自 MODIS BRDF/ALBEDO 系列 MCD43A3 产品,空间分辨率为 500 m,时间分辨率为 1 d,该数据是经过双向反射函数(BRDF)模型计算修正的反照率产品[17].本数据集包括 7 个窄波段和 3 个宽波段的黑空和白空反照率.白空反照率是漫射-半球反照率,反映了阴天条件下的地表反射状况;黑空反照率能够较为准确地反映正午时刻地球表面对太阳直射光线的反射情况[18].由于白空反照率和黑空反照率的平均值差异较小,且高度相关,实际地表反照率可选择白空、黑空反照率的平均值计算获得[19].

3)地表反射率数据:来自 MODIS 的 MOD09A1 和 MYD09A1 产品,空间分辨率为 500 m,时间分辨率为 8 d,共包含 7 个波段,其对应的波长范围如表 1




a. 华北地区区位　　b. 2019年华北地区土地利用及土壤湿度观测站点分布

图 1 研究区概况

Fig. 1 Overview of the study area　a.location map of North China;b.distribution map of land use and soil moisture observation sites in North China in 2019

所示.该数据是低观测角度条件下,受云、云阴影及气溶胶等影响最小的 8 d 日数据合成产品[20].归一化差异水体指数(NDWI)计算公式[21]如下:

$$NDWI = \frac{\{\rho(858\ nm) - \rho(1\ 240\ nm)\}}{\{\rho(858\ nm) + \rho(1\ 240\ nm)\}}, \quad (1)$$

其中,$\rho(858\ nm)$ 和 $\rho(1\ 240\ nm)$ 分别对应反射率数据的 2、5 波段.

**表 1 反射率数据的波长范围**

Table 1 Wavelength range of the reflectivity data

| 波段 | 波长范围/nm |
|---|---|
| 1 | 620~670 |
| 2 | 841~876 |
| 3 | 459~479 |
| 4 | 545~565 |
| 5 | 1 230~1 250 |
| 6 | 1 628~1 652 |
| 7 | 2 105~2 155 |

1.2.3 土壤质地数据

土壤质地数据来源于中国科学院资源环境科学数据中心(http://www.resdc.cn),投影为 Albers 正轴等面积双标准纬线圆锥投影,空间分辨率为 1 km.该数据是根据 1∶100 万土壤类型图和第二次土壤普查获取到的土壤剖面数据编辑制作而成的,依据砂粒、粉粒、黏粒含量进行土壤质地划分,将数据分为 Sand(砂土)、Silt(粉砂土)与 Clay(黏土)3 大类,每一类数据均通过百分比来反映不同质地颗粒的含量.

1.2.4 地形数据

数字高程模型(Digital Elevation Model,DEM)数据选用 SRTM DEM 数据,来源于中国科学院资源环境科学数据中心(http://www.resdc.cn).该数据是基于雷达测图技术通过美国“奋进”号航天飞机获得的,涵盖了 60°N~56°S 间陆地地表 80%面积范围,经 NASA 喷气推进实验室处理完成[22].研究采用的数据是基于版本 4.1 的 DEM 数据,经重采样生成 1 km 全国数据.

将以上遥感数据、土壤质地数据以及地形数据投影统一转换至经纬度投影,利用华北地区行政边界裁剪等预处理,经双线性插值方法,对数据重采样,分别生成分辨率 6 km 和 1 km 的两种数据,6 km 数据和 CLDAS 6 km 土壤湿度数据相匹配,用来训练降尺度模型,1 km 数据作为降尺度模型输入数据,用于估算高分辨土壤湿度.

1.2.5 站点观测数据

站点数据来源于国家气象信息中心资料服务室的 2019 年逐小时观测资料[23].该站点土壤湿度观测在垂直方向上分为 8 层:0~10 cm、10~20 cm、20~30 cm、30~40 cm、40~50 cm、50~60 cm、60~80 cm、80~100 cm.利用频域反射技术(Frequency Domain Reflectometry,FDR)来测定土壤体积含水量[24].研究参考韩帅[25]提出的方法对土壤湿度观测数据进行质量控制,得到有效站点 223 个,采用随机方法,抽出 185 个作为训练站点数据用于建模,38 个站点数据用于验证(图 1).

## 2 研究方法

### 2.1 降尺度因子的选择

华北地区的土壤湿度在时间尺度上具有明显的季节性,受到夏季风的影响,夏季土壤湿度高,冬季土壤湿度低.在地域上中部区域地势较平坦,土地利用类型以农业用地为主,土壤湿度高;北部与西南部地形复杂,土地利用类型以草地和灌木丛为主,涵养水源的作用较差,土壤湿度低;南部和沿海地区粉砂粒和黏粒含量高、砂粒含量低,土壤湿度高.总体上除降水外,本地区土壤湿度变化受到温度、地形、植被、土壤等因素的共同影响[26-27].地表温度为监测和降尺度土壤湿度中最重要的变量,土壤表层温度发生变化,其内在因素是土壤热惯量,且随着土壤湿度的增大,土壤热惯量增大,因此,地表温度与土壤湿度有密切的相关性[28].高程和坡度是影响土壤湿度空间分布的关键因素[26],华北地区高程为-23~2 539 m,高程变化大,坡度为 0°~22°,其中,94%的区域坡度为 0°~5°,坡度变化小,所以引入高程作为降尺度因子之一.在可见光和近红外区域,土壤湿度与植被光谱响应之间存在显著关系,研究表明,短波红外区域对土壤湿度的监测效果更好[29-31].同时,短波红外波段是植物叶片吸收水分的区域,植被反射率与叶片含水量呈负相关[32].因此,选择基于短波红外波段的归一化差异水体指数,作为降尺度因子之一.土壤异质性通过土壤质地的变化,包括土壤颗粒和孔隙分布的变化,影响土壤湿度的分布.此外,地表反照率受到土壤颜色的影响,从而影响植被稀疏土壤的蒸发效率,进而影响土壤湿度[33].因此,参照前人研究[34-38],选用更高分辨率地表温度、高程、归一化差异水体指数、土壤质地和地表反照率等对土壤水分较为敏感的因子指标,作为降尺度因子来开

展研究.

### 2.2 梯度提升机(GBM)

梯度提升机(Gradient Boosting Machine,GBM)是一种可以解决分类、回归和重要性排序问题的机器学习模型,是 Boosting 算法的典型代表.它遵循了集成学习的一种思想,即分多个阶段迭代训练一系列可叠加的基学习器模型,在迭代的推进过程中不断进行优化和提升,使每一次新的迭代都是为了减少上一次迭代的残差,使模型沿着残差减少最快的方向进行,由此产生一系列弱分类器,每个弱分类器都是一棵二叉树,最终将这些弱分类器组合形成能使损失函数达到极小的模型[39].GBM 对异常值和不平衡数据具有鲁棒性,确保了高效的性能.其算法过程[40]如下:

1)输入训练集数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$,损失函数 $L(y,f(x))$.

2)初始化模型 $f_0(x)=\arg\min\sum L(y_k,c)$,$c$ 为常量,是用来估计损失函数最小化的常数值.

3)迭代 $n=1,2,\cdots,N$($N$ 为样本数),$k=1,2,3,\cdots,K$($K$ 为基学习器的个数),计算$r_{kn}$:

$$r_{kn}=-\left[\frac{\partial L(y_n,f(x_n))}{\partial f(x_n)}\right]f(x)=f_{k-1}(x),\qquad(2)$$

式中,$r_{kn}$ 为损失函数负梯度在当前模型的值,若损失函数已达到最小值,则进行步骤 5),否则进行下一步.

4)对 $r_{kn}$ 建立基学习器模型$T_k(x)$,对梯度提升进行更新:

$$f_m(x)=f_{m-1}(x)+T_k(x).\qquad(3)$$

5)得到强学习器:

$$\hat{f}(x)=f_k(x)=\sum_{k=1}^{K}T_k(x).\qquad(4)$$

梯度提升机方法基于陆地表面变量和土壤湿度之间的统计关系,降尺度过程主要涉及两个阶段:

1)训练.基于 CLDAS 土壤湿度数据,与土壤湿度数据空间分辨率保持一致的陆表变量数据和站点数据建立梯度提升机回归模型.

2)预测.将高分辨率陆表变量数据输入第 1 阶段建立的回归模型,以生成高分辨率土壤湿度数据(图 2).

### 2.3 深度前馈神经网络(DFNN)

本研究采用深度前馈神经网络(Deep Feedforward Neural Network,DFNN)作为深度学习算法.在回归任务中,该模型可以从大量变量中提取高级特征,以实现较高的预测精度[41].每一层的神经元可以接收前一层神经元的信号,并产生信号输出到下一层.第 0 层叫做输入层,最后一层叫做输出层,其他中间层叫做隐藏层(图 3).这种网络模型在各层之间具有全连接的神经元结构,其中隐藏层根据模型的复杂程度可以设计成任意数量的多层,各层之间的连接表示特征的权重,其中信息没有反馈的从左向右传输.连接输入和输出的每层神经元结构具有以下映射关系[42]:

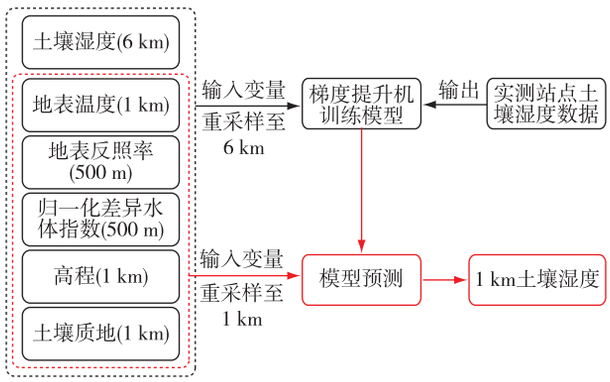



图 2 梯度提升机土壤湿度降尺度结构示意

Fig. 2 Downscaling of soil moisture based on gradient boosting machine

$$y=f(x,\theta),\qquad(5)$$

其中 $x$ 和 $y$ 分别代表输入和输出,$\theta$ 表示已知输入和期望输出值之间映射的最优参数解.为了避免深层网络反向传播可能带来的梯度消失和梯度爆炸的问题,各层的激活函数采用 ReLU 函数.输出层概率分布计算采用 softmax 函数.假设输入样本表示为:$X=(X_1,X_2,X_3,X_4,X_5)$,两个隐藏层 $h_1,h_2$ 的维度分别为 $H_1,H_2$,则两个隐藏层的输出分别如式(6)、式(7)所示,输出层的输出如式(8)所示.$\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3$ 为各连接层的连接权重矩阵,$\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3$ 为各层偏移量.模型训练优化的参数集合为 $\theta=\{\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{W}_3,\boldsymbol{b}_1,\boldsymbol{b}_2,\boldsymbol{b}_3\}$.

$$y_{h1}=\mathrm{ReLU}(\boldsymbol{W}_1X+\boldsymbol{b}_1),\qquad(6)$$

$$y_{h2}=\mathrm{ReLU}(\boldsymbol{W}_2\,y_{h1}+\boldsymbol{b}_2),\qquad(7)$$

$$y=\mathrm{softmax}(\boldsymbol{W}_3\,y_{h2}+\boldsymbol{b}_3).\qquad(8)$$

模型训练及预测过程如下:

1)变量因子标准化及输入.本研究使用标准差标准化方法,如式(9)所示,将处理好的标准化 6 km 空间分辨率自变量与因变量因子输入 DFNN 模型.

2)模型调参.需要通过调整模型参数以达到最

好的训练效果,主要参数包括隐藏层层数 $L$、隐藏层每层神经元数量 $N$ 和训练迭代次数 epochs.

3)土壤湿度预测.得到最优训练模型后,将 1 km 空间分辨率的自变量因子输入训练好的模型,得到 1 km 空间分辨率预测土壤湿度(图 3).

$$\text{标准化} = \frac{x_i - \text{mean}(x)}{\text{std}(x)}, \tag{9}$$

式中,$x_i$ 为一列自变量中的第 $i$ 个值,mean($x$)和 std($x$)分别是 $x$ 所在列自变量的均值和标准差.

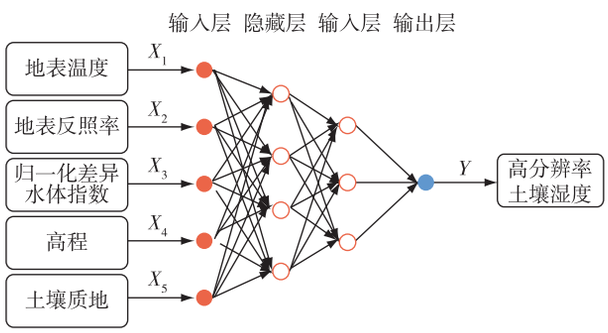



图 3 深度前馈神经网络土壤湿度降尺度结构示意

Fig. 3 Downscaling of soil moisture based on deep feedforward neural network

## 2.4 随机森林(RF)

随机森林是 Bagging 算法的典型代表,作为一种增强型决策树模型,可用于分类、回归等任务[43].此外,模型对异常值不敏感,在样本和变量的随机化训练阶段表现出良好的性能.与其他机器学习方法相比,随机森林模型被广泛应用于微波土壤湿度产品降尺度[44].随机森林模型的主要思想是基于回归树建立输入变量与输出土壤湿度之间的非线性函数[45]:

$$\text{SSMO} = f_{\text{RF}}(\boldsymbol{C}) + \varepsilon, \tag{10}$$

$$\boldsymbol{C} = (\text{LST}, \text{Albedo}, \text{DEM}, \text{soiltexture}, \text{NDWI}, \text{CLDAS-SSM}), \tag{11}$$

其中:SSMO 表示训练阶段的实测土壤湿度值;$\boldsymbol{C}$ 为输入向量,表示输入变量,包括地表温度(LST)、地表反照率(Albedo)、高程(DEM)、土壤质地(soil texture)、归一化差异水体指数(NDWI)和 CLDAS 土壤湿度(CLDAS-SSM);$f_{\text{RF}}$ 是一个非线性函数,在输入变量和输出 SSMO 之间建立关系.

在本研究的回归任务中,首先在训练期间内建立若干棵决策树,每棵决策树由 bootstrap 样本建立,其中训练输入数据约占总样本的 2/3,其余(1/3)的样本用于验证每棵树.为了进一步提高随机森林模型的泛化能力,通过对许多独立回归树的结果求算术平均来生成最终模型预测值,模型最终结果表示为

$$p(\text{SSMO} \mid \boldsymbol{C}) = \frac{1}{m}\sum_{i=1}^{m} P_i(\text{SSMO} \mid \boldsymbol{C}), \tag{12}$$

式中,$p(\text{SSMO} \mid \boldsymbol{C})$ 为最终预测结果,$m$ 是回归树的数量,$P_i(\text{SSMO} \mid \boldsymbol{C})$ 表示第 $i$ 棵树的预测结果.

## 2.5 Stacking 集成学习方法

集成学习的优势在于将多个基学习器的结果进行优化组合输出,以获得比任意基学习器更好的结果[46].集成学习主要包括并行化集成的 Bagging、序列化集成的 Boosting 以及堆叠式集成的 Stacking 等.以 Bagging 和 Boosting 为代表的集成方法可以对训练效果差的样本赋以较高的权重进行二次学习,提高组合预测的泛化能力.然而该类方法只能集成同类决策树模型,难以融合其他模型的优势特性,不同算法间数据观测的差异性难以体现[47].Stacking 是一种分层模型集成框架(图 4),首先调用不同类型的学习器对数据集进行训练学习,将各学习器得到的训练结果组成一个新的训练样例,作为元学习器的输入,最终第 2 层模型中元学习器综合多个基学习器的输出特征,作出最后的决策[48].因此,Stacking 增加了模型的准确性、鲁棒性和整体归纳能力.

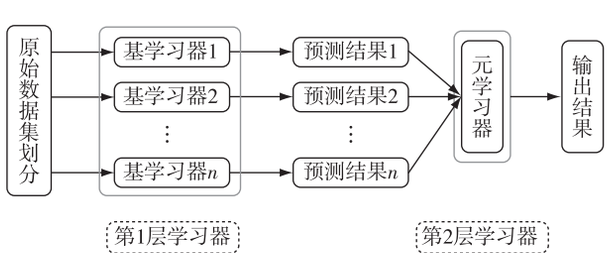



图 4 Stacking 集成学习结构示意

Fig. 4 Structure of Stacking ensemble learning

广义线性模型(Generalized Linear Models,GLM)作为一般线性模型的扩展,基本思想是通过概率分布函数模拟非线性过程,它具有清晰的变量权重结构,能够模拟非线性的响应关系,模型不会出现明显的过拟合,对每个基学习器的效应产生清晰的认识[49].GLM 模型主要通过连接函数,建立响应变量的数学期望值,及其与代表线性组合的预测变量间的关系.一个广义线性模型包括随机成分、系统成分和连接函数三部分:

$$f(y_i;\theta_i;\phi_i) = \exp\left\{\frac{y_i\,\theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i,\phi_i)\right\}, \tag{13}$$

$$n_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}\ , \tag{14}$$

$$E(Y_i) = \mu_i = g^{-1}(n_i)\ . \tag{15}$$

随机成分即因变量的概率分布,其中 $a(\phi_i)$, $b(\theta_i)$ ,$c(y_i,\phi_i)$ 为已知的函数;系统成分即自变量的线性组合.连接函数建立了随机成分与系统成分之间的特定关系.

本文基于 Stacking 集成学习的土壤湿度降尺度模型框架(图 5),算法步骤如下:

1)输入原始数据集 $T$,即包括 CLDAS 土壤湿度数据、与土壤湿度数据空间分辨率保持一致的陆表变量数据和站点数据,并按照 3∶1的比例随机划分训练集 $T_1$ 和测试集 $T_2$,$T = T_1 \cup T_2$,$T_1 \cap T_2 = \varnothing$.

2)学习并生成新的数据集.第 1 层包含 3 种基学习器:梯度提升机、深度前馈神经网络和随机森林,采用 $K$ 折交叉验证来训练第 1 层模型,3 种模型扩展之后生成第 2 层训练集 $T'_1$.在基学习器模型进行 $K$ 折交叉验证过程中,对测试集 $K$ 次计算结果求平均,3 种模型扩展后生成第 2 层测试集 $T'_2$.$T'_1$ 和 $T'_2$ 构成新的数据集 $T'$.

3)将得到的 $T'_1$ 用于训练第 2 层元学习器广义线性模型,并用 $T'_2$ 验证模型性能.训练得到最终土壤湿度降尺度模型.

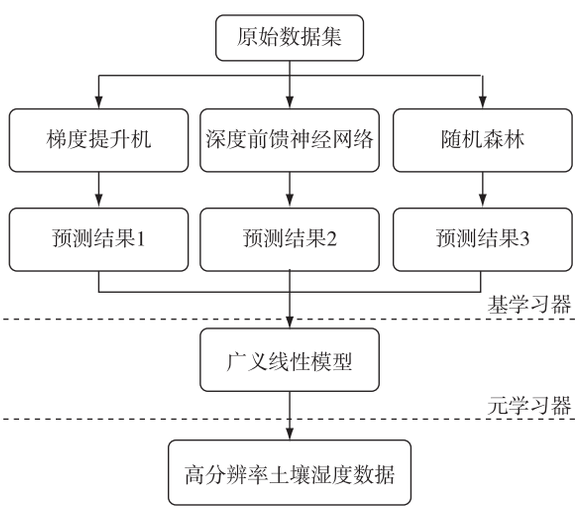



图 5　Stacking 集成学习方法土壤湿度降尺度结构示意

Fig. 5　Downscaling of soil moisture based on Stacking ensemble learning

### 2.6　评价指标

本研究采用相关系数($R$)、偏差(Bias,其量值记为 $B$)和均方根误差(RMSE)3 个指标定量地分析原始 CLDAS 土壤湿度和降尺度土壤湿度,计算公式[50]如下:

$$R = \frac{\sum (P_i - \bar{P})(O_i - \bar{O})}{\sqrt{\sum (P_i - \bar{P})^2 (O_i - \bar{O})^2}}\ , \tag{16}$$

$$B = \frac{1}{m}\sum_{i=1}^{m}(P_i - O_i)\ , \tag{17}$$

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{m}(P_i - O_i)^2}{m}}\ , \tag{18}$$

其中,$m$ 是样本数量,$O_i$ 和 $P_i$ 分别是第 $i$ 个实测值和预测的土壤湿度值,$\bar{O}$ 和 $\bar{P}$ 分别是实测值和预测结果的平均值.

精度评估过程中,采用双线性内插法,将 CLDAS 土壤湿度和不同方法降尺度结果内插到站点,然后与站点观测数据进行分析,计算相关系数、均方根误差和偏差.

采用决定系数($R^2$)和均方根误差(RMSE)对深度学习模型预测结果进行精度判定.均方根误差如式(18)所示,决定系数计算公式[51]如下:

$$R^2 = 1 - \frac{\sum (O_i - P_i)^2}{\sum (O_i - \bar{O})^2}\ , \tag{19}$$

$O_i$ 和 $P_i$ 分别是第 $i$ 个实测值和预测的土壤湿度值,$\bar{O}$ 是实测值平均值.

## 3　结果与分析

### 3.1　深度前馈神经网络参数调优

研究使用决定系数 $R^2$ 和均方根误差 RMSE 来表征深度前馈神经网络的拟合效果.通过调整隐藏层层数 $L$、神经元数量 $N$ 和迭代次数 epochs,得到模型训练集和测试集的决定系数 $R^2$ 和均方根误差 RMSE.当模型包含过多参数时,为了避免模型结果过拟合,提高模型的泛化能力,有必要同时考虑训练集和测试集具有最高的决定系数和最低的均方根误差.结果表明(表 2),固定神经元数量 $N = 400$,逐渐增加隐藏层层数 $L$ 或迭代次数 epochs 时,总体上,训练集和测试集的决定系数逐渐增大,均方根误差逐渐降低.当隐藏层层数 $L = 5$,迭代次数 epochs = 600 时,训练集和测试集的决定系数分别为 0.936 和 0.830,均方根误差分别为 0.018 和 0.030 1 $m^3/m^3$,此时再逐渐增加隐藏层层数 $L$ 或迭代次数 epochs,训练集和测试集的决定系数逐渐降低,均方根误差逐渐增大.因此,在本研究中最终选择以下数值作为模型的初始输入参数:隐藏层层数 $L = 5$,神经元数量 $N = 400$,迭代次数 epochs = 600.

**表 2　模型参数调整结果**

Table 2　Model parameter adjustment results

| $L$ | $N$ | epochs | 训练集 | | 测试集 | |
|---|---|---|---|---|---|---|
| | | | $R^2$ | RMSE/($m^3/m^3$) | $R^2$ | RMSE/($m^3/m^3$) |
| 4 | 400 | 400 | 0.865 | 0.024 | 0.804 | 0.032 3 |
| 4 | 400 | 500 | 0.905 | 0.023 | 0.814 | 0.031 4 |
| 4 | 400 | 600 | 0.912 | 0.022 | 0.815 | 0.031 4 |
| 5 | 400 | 400 | 0.919 | 0.021 | 0.810 | 0.031 9 |
| 5 | 400 | 500 | 0.927 | 0.019 | 0.819 | 0.031 1 |
| 5 | 400 | 600 | 0.936 | 0.018 | 0.830 | 0.030 1 |
| 6 | 400 | 400 | 0.915 | 0.021 | 0.822 | 0.030 9 |
| 6 | 400 | 500 | 0.899 | 0.023 | 0.803 | 0.032 5 |
| 6 | 400 | 600 | 0.888 | 0.024 | 0.804 | 0.032 4 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |

### 3.2　不同方法降尺度结果空间分布

为了比较不同方法降尺度效果,研究比较了梯度提升机(GBM)、深度前馈神经网络(DFNN)、随机森林(RF)和 Stacking 集成学习等 4 种方法,因土壤温度小于 0 ℃时,观测仪器异常导致缺少可靠的观测数据,研究仅分析 2019 年 4—10 月的 CLDAS 土壤湿度产品降尺度效果.从降尺度结果和原土壤湿度日均值分布(图 6)可以看出,总体上,华北地区的南部和沿海区域土壤湿度较高,中部和北部土壤湿度较低,平均土壤湿度均达到 0.2 $m^3/m^3$ 以上,降尺度前后产品均较好地反映出此变化规律.但降尺度前土壤湿度日均值的平均值为 0.226 $m^3/m^3$,降尺度后土壤湿度有所降低,日均值的平均值分别为:GBM(0.208 $m^3/m^3$)>DFNN(0.207 $m^3/m^3$)>RF(0.207 $m^3/m^3$)>Stacking(0.206 $m^3/m^3$).特别是华北地区的南部和沿海区域降尺度后降低明显,降尺度前土壤湿度日均值大于 0.25 $m^3/m^3$,降尺度后介于 0.2~0.3 $m^3/m^3$ 之间,并且降尺度后土壤湿度的空间分布细节更加丰富,这与降尺度过程中融合了对土壤湿度影响较大的高分辨率地表温度、地表反照率、地形等地表变量数据有关.

### 3.3　不同方法降尺度效果精度分析

利用站点观测数据逐日土壤湿度评估表明(图 7 和图 8):不同降尺度方法预测的降尺度结果与站点观测土壤湿度之间存在显著的相关性,相关系数介于 0.699 6~0.756 8,高于原土壤湿度的相关系数 0.651 1;降尺度后均方根误差介于 0.050 5~0.055 3 $m^3/m^3$ 之间,低于原土壤湿度均方根误差 0.062 3 $m^3/m^3$;降尺度后偏差介于-0.008 1~-0.005 2 $m^3/m^3$ 之间,比原土壤湿度偏差 0.023 9 $m^3/m^3$ 更接近于 0,说明降尺度后土壤湿度精度得到提高,相比于原土壤湿度,降尺度结果更接近于实测值.

对比 4 种不同的降尺度方法结果的相关系数表明(图 8),Stacking 集成学习土壤湿度和 RF 土壤湿度相关系数较高,分别为 0.756 8 和 0.740 2,其次为 DFNN 和 GBM 土壤湿度,分别为 0.7155 和 0.699 6,且 Stacking 集成学习土壤湿度和 RF 土壤湿度均方根误差较低,分别为 0.050 5 $m^3/m^3$ 和 0.050 9 $m^3/m^3$,其次为 GBM 土壤湿度和 DFNN 土壤湿度,分别为 0.055 3 $m^3/m^3$ 和 0.055 4 $m^3/m^3$.4 种不同降尺度方法结果偏差均在 0 以下,存在一定程度上的低估,但绝对偏差小于 CLDAS 产品,一定程度上改善了其高估现象,以 Stacking 集成学习方法低估程度最小,土壤湿度绝对偏差为 0.005 2 $m^3/m^3$.因此,相对来说,Stacking 集成学习方法优于其他 3 种方法,其相关系数最高,均方根误差和偏差相对较小,这与 Stacking 集成学习方法能够更好地挖掘出输入变量和土壤湿度的相关性,提升模型拟合效果有关.

### 3.4　不同方法降尺度结果时间序列及误差分析

从不同方法降尺度结果 0~10 cm 土层土壤湿度时间序列来看(图 9),4 种降尺度方法的降尺度结果和原土壤湿度的变化趋势与观测值总体上相似,均能反映出土壤湿度随时间变化的规律,但大多数日次原土壤湿度存在高估现象,在整个时间段内,比观测值高 0.013 7 $m^3/m^3$.4 种降尺度结果则存在一定程度上的低估,尤其在 91~98、180~190、200~210 和 240~270 日次,但整体上降尺度结果和观测值的曲线更为接近.按其与观测值偏离程度大小依次为:GBM>DFNN>RF>Stacking 集成学习方法,其中,Stacking 集成学习土壤湿度比观测值低大约 0.005 2 $m^3/m^3$,整体上,

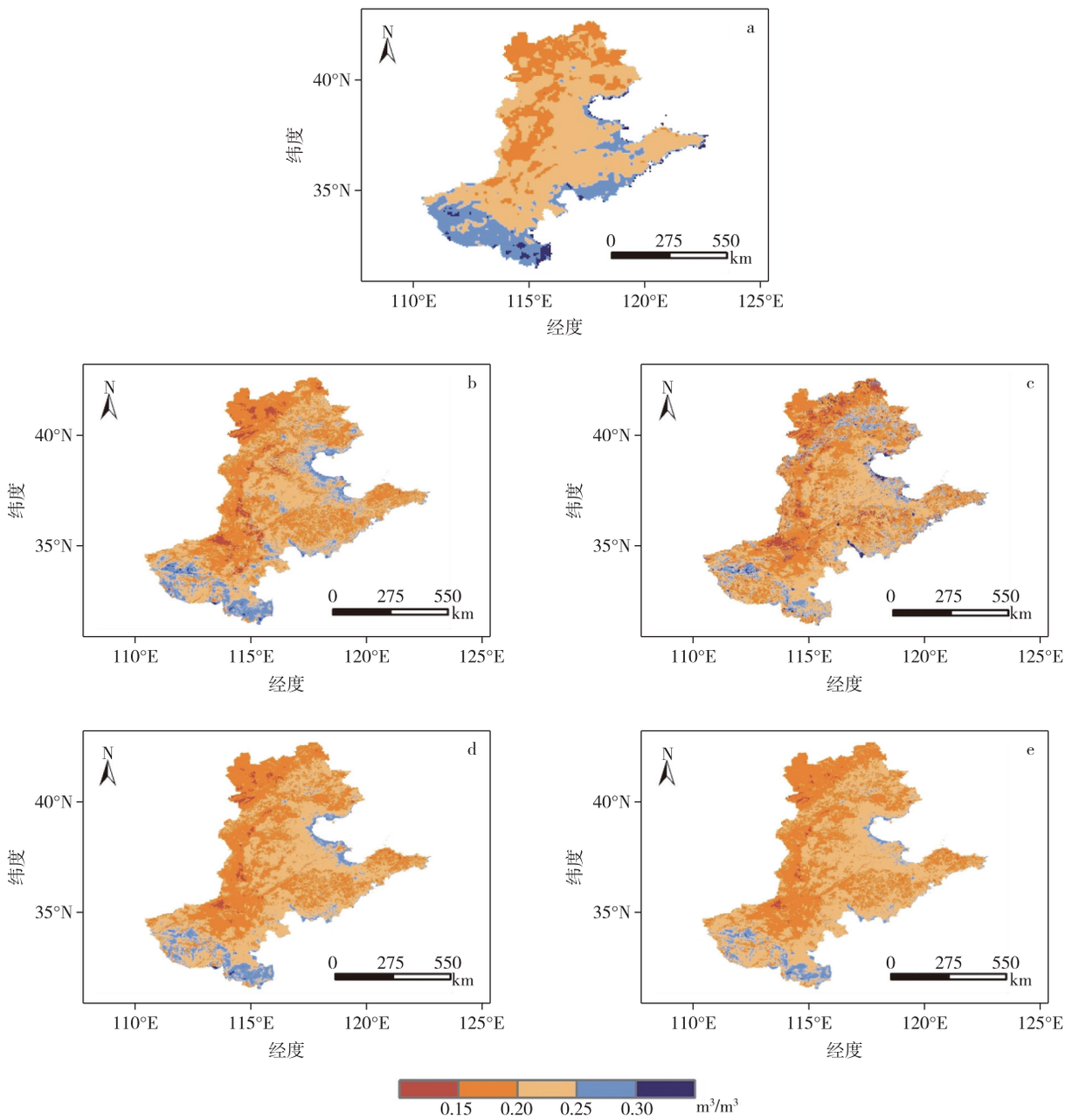



图 6　不同方法结果土壤湿度日均值空间分布　　a.CLDAS;b.GBM;c.DFNN;d.RF;e.Stacking 集成学习

Fig. 6　Spatial distribution of original and downscaled daily average soil moisture

a.CLDAS;b.GBM;c.DFNN;d.RF;e.Stacking ensemble learning

Stacking 集成学习降尺度土壤湿度与观测值曲线趋势变化更相吻合,更接近站点观测数据.

从土壤湿度与观测值的相关系数和误差分析(图 10)来看,与原土壤湿度相比,4 种降尺度方法降尺度结果相关系数均有所提升,其中,Stacking 集成学习土壤湿度相关系数提高较大,平均提高 0.13,其次为 RF 土壤湿度,平均提高 0.12.与原土壤湿度相比,4 种不同降尺度方法降尺度结果的均方根误差均有明显的降低,其中,Stacking 集成学习土壤湿度和 RF 土壤湿度比原土壤湿度的均方根误差,分别平均降低了 0.012 1 和 0.011 7 $m^3/m^3$,更接近于观测值.4 种方法降尺度结果的偏差绝大多数日次在 0 以下,即降尺度结果存在一定程度的低估现像.原土壤湿度偏差均值为 0.023 9 $m^3/m^3$、GBM 为-0.008 1 $m^3/m^3$、DFNN 为-0.005 8 $m^3/m^3$、RF 为-0.005 6 $m^3/m^3$、Stacking 集成学习方法为-0.005 2 $m^3/m^3$,4 种降尺度结果均改善了原土壤湿度的高估问题,其中,以 Stacking 集成学习方法最优.

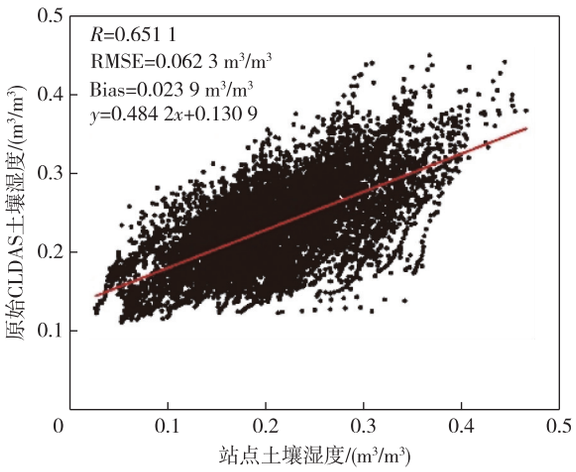



图 7　CLDAS 逐日土壤湿度精度站点评估散点图

Fig. 7　Scatter plot of CLDAS daily soil moisture and observations

表 3 为 4 种不同降尺度方法的降尺度结果与观测值在月尺度上的相关系数和误差.4 种方法降尺度结果与观测值的相关系数达到 0.45 以上,其中,4 月、6 月、7 月和 8 月相关系数较大,均达到 0.7 以上,9 月较小,介于 0.5~0.6 之间,这可能与夏季降水频繁,土壤水分具有较强的空间异质性,增加了土壤湿度估算的不确定性有关.整体来看,平均相关系数均大于 0.65,其中,Stacking 集成学习方法最高.除 9 月外,4 种方法降尺度结果与观测值在月尺度上的绝对偏差均小于 0.01 $m^3/m^3$,其中,4 月绝对偏差最小.在整个时间段内,除 RF 土壤湿度在 4 月偏差为正值,其余各月偏差均为负值,各月土壤湿度估算值均低于观测值.绝对偏差均值大小依次为:GBM>DFNN>RF>Stacking 集成学习.4 种方法降尺度结果与观测值在月尺度上的均方根误差介于 0.043~0.067 $m^3/m^3$之间,其中,9 月均方根误差较大,各方法降尺度结果的均方根误差均大于 0.059 $m^3/m^3$,4 月均方根误差较小,各方法降尺度结果的均方根误

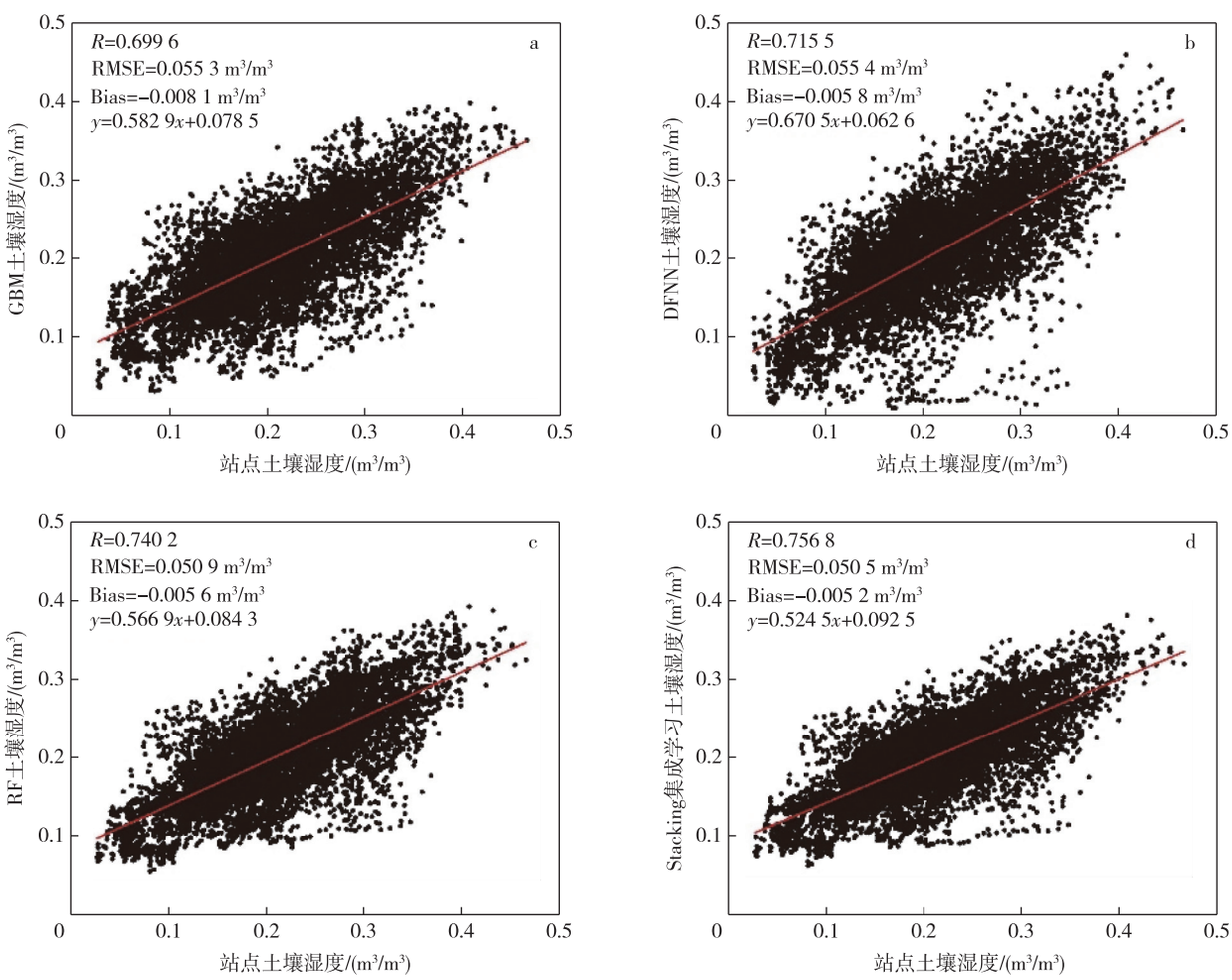



图 8　不同方法降尺度结果站点精度评估散点图　　a.GBM;b.DFNN;c.RF;d.Stacking 集成学习

Fig. 8　Scatter plots of downscaled soil moistures and observations

a.GBM;b.DFNN;c.RF;d.Stacking ensemble learning

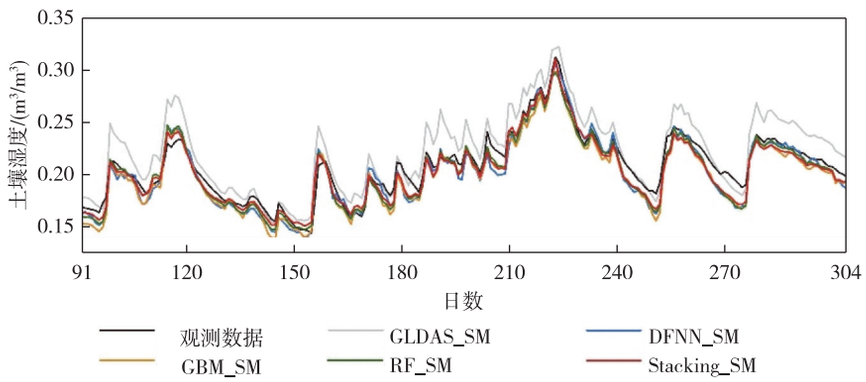



图 9　不同方法降尺度结果 0~10 cm 土层土壤湿度时间序列

Fig. 9　Time series of 0-10 cm soil moisture downscaled by different methods

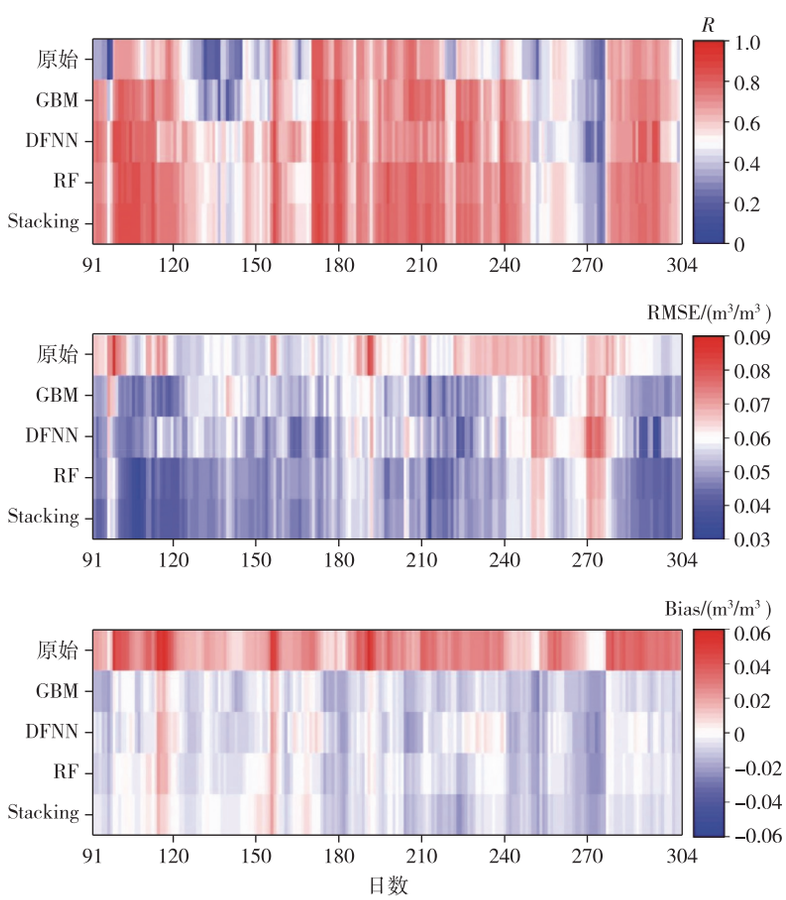



图 10　不同方法降尺度日均土壤湿度与观测值的相关系数与误差

Fig. 10　Correlation coefficients (up),RMSEs (middle),and biases (down) of downscaled daily average soil moistures compared with observations

差介于 0.043~0.050 $m^3/m^3$ 之间,以 Stacking 集成学习方法最小.因此,Stacking 集成学习方法在月尺度上优于其他方法.

## 4　结论与讨论

研究以华北地区为例,以地表温度、地表反照率、土壤质地、高程、归一化差异水体指数以及站点数据作为建模数据,基于 3 种单一模型(梯度提升机、深度前馈神经网络和随机森林)以及多模型 Stacking 集成学习的方法,开展了中国气象局陆面数据同化系统(CLDAS-V2.0) 0~10 cm 土层土壤湿度数据降尺度研究,使其空间分辨率从 6 km 降尺度至

表 3 不同方法降尺度月均土壤湿度与观测值的相关系数与误差

Table 3 Correlation coefficients and errors of different downscaling methods for monthly average soil moisture

| 月份 | 相关系数 | | | | 偏差/($m^3/m^3$) | | | | 均方根误差/($m^3/m^3$) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | GBM | DFNN | RF | Stacking | GBM | DFNN | RF | Stacking | GBM | DFNN | RF | Stacking |
| 4 | 0.742 4 | 0.773 4 | 0.778 3 | 0.790 9 | −0.004 7 | −0.002 3 | 0.000 3 | −0.000 3 | 0.049 6 | 0.0503 | 0.045 1 | 0.043 8 |
| 5 | 0.479 2 | 0.600 5 | 0.603 6 | 0.606 8 | −0.006 8 | −0.006 8 | −0.005 2 | −0.002 6 | 0.056 5 | 0.053 8 | 0.047 8 | 0.047 1 |
| 6 | 0.704 2 | 0.759 2 | 0.738 0 | 0.752 7 | −0.005 2 | −0.001 3 | −0.003 6 | −0.001 2 | 0.054 4 | 0.049 9 | 0.050 4 | 0.049 2 |
| 7 | 0.704 8 | 0.695 5 | 0.738 7 | 0.747 3 | −0.007 8 | −0.009 3 | −0.005 9 | −0.007 5 | 0.056 0 | 0.058 1 | 0.053 0 | 0.053 2 |
| 8 | 0.741 9 | 0.739 1 | 0.766 8 | 0.779 8 | −0.009 3 | −0.002 1 | −0.006 3 | −0.009 0 | 0.051 5 | 0.052 2 | 0.049 1 | 0.049 5 |
| 9 | 0.529 4 | 0.518 5 | 0.569 3 | 0.571 0 | −0.014 2 | −0.013 3 | −0.012 6 | −0.012 8 | 0.064 0 | 0.066 2 | 0.059 9 | 0.059 5 |
| 10 | 0.660 4 | 0.648 2 | 0.695 4 | 0.695 6 | −0.008 3 | −0.005 2 | −0.005 4 | −0.005 0 | 0.053 5 | 0.055 2 | 0.049 2 | 0.049 1 |
| 平均 | 0.651 8 | 0.676 3 | 0.698 6 | 0.706 3 | −0.008 1 | −0.005 8 | −0.005 6 | −0.005 2 | 0.055 3 | 0.055 1 | 0.050 6 | 0.050 2 |

1 km,并以站点观测数据对降尺度结果进行精度分析,得到以下结论:

1)4 种不同降尺度方法的降尺度结果和原土壤湿度在华北地区的空间分布具有相似规律,南部和沿海区域土壤湿度较高,中部和北部土壤湿度较低,平均土壤湿度均达到 0.2 $m^3/m^3$以上.降尺度后土壤湿度日均值有所降低,特别是在华北地区的南部和沿海区域.

2)4 种不同降尺度方法均有效提高了 CLDAS 土壤湿度产品的空间分辨率和精度,4 种方法绝对偏差均小于 CLDAS 产品,一定程度上改善了高估现象.原土壤湿度与站点观测数据的相关系数、均方根误差和偏差分别为 0.651 1、0.062 3 $m^3/m^3$和 0.023 9 $m^3/m^3$,降尺度土壤湿度的精度高低依次是 Stacking 集成学习方法、随机森林、深度前馈神经网络、梯度提升机.Stacking 集成学习降尺度方法估算精度最高,相关系数、均方根误差和偏差分别为 0.756 8、0.050 5 $m^3/m^3$和−0.005 2 $m^3/m^3$,梯度提升机估算效果较差,相关系数、均方根误差和偏差分别为 0.699 6、0.055 3 $m^3/m^3$和−0.008 1 $m^3/m^3$.

3)原土壤湿度和 4 种不同降尺度方法的降尺度结果均能较好地体现土壤湿度的日变化特征,但大多数日次原土壤湿度存在高估现象,4 种降尺度结果存在一定程度上的低估.整体上,降尺度结果和观测值的曲线更为接近,其中,Stacking 集成学习方法最优,与原土壤湿度相比,相关系数平均提高 0.13,均方根误差平均降低 0.012 1 $m^3/m^3$,偏差均值为−0.005 2 $m^3/m^3$.月尺度结果中,4 种不同降尺度方法的降尺度结果在 9 月相关系数均较小,均方根误差和偏差较大,整体来看,与其他 3 种方法相比,Stacking 集成学习土壤湿度在月尺度上相关系数较大,均方根误差和偏差较低,估算的土壤湿度精度较高.

本文利用梯度提升机、深度前馈神经网络、随机森林和 Stacking 集成学习方法,对 CLDAS 0~10 cm 土层土壤湿度产品开展了降尺度研究,使其空间分辨率从 6 km 降尺度至 1 km,且精度有所提高.但本研究选取的土壤湿度数据时间范围为 2019 年 4—10 月,即春夏秋三季,未考虑冬季降尺度模型的土壤湿度估算表现,主要是因为当前土壤湿度测量仪器在土壤含有冰水混合物时,测量结果存在误差与不确定性,冬季的降尺度结果有待进一步研究.另外,与土壤湿度相关的降尺度因子数据的质量对建模效果有较大影响,高质量的数据会提高降尺度结果的精度.

## 参考文献

References

[1] Clevers J, Leeuwen H. Combined use of optical and microwave remote sensing data for crop growth monitoring [J]. Remote Sensing of Environment, 1996, 56(1): 42-51

[2] Western A W, Zhou S L, Grayson R B, et al. Spatial correlation of soil moisture in small catchments and its relationship to dominant spatial hydrological processes [J]. Journal of Hydrology, 2004, 286(1): 113-134

[3] Seneviratne S I, Corti T, Davin E L, et al. Investigating soil moisture-climate interactions in a changing climate: a review [J]. Earth Science Reviews, 2010, 99(3/4): 125-161

[4] Jaroslaw Z, Mateusz K. Soil moisture variability over Odra watershed: comparison between SMOS and GLDAS data [J]. International Journal of Applied Earth Observations and Geoinformation, 2016, 45: 110-124

[5] Xia Y, Sheffield J, Ek M B, et al. Evaluation of multi-model simulated soil moisture in NLDAS-2 [J]. Journal of Hydrology, 2014, 512: 107-125

[6] Shi C X, Xie Z H, Qian H, et al. China land soil moisture EnKF data assimilation based on satellite remote sensing data [J]. Science China Earth Sciences, 2011, 54(9):

1430-1440
[7] Peng J,Loew A,Merlin O,et al.A review of spatial downscaling of satellite remotely sensed soil moisture[J].Reviews of Geophysics,2017,55(2):341-366
[8] Njoku E G,Wilson W J,Yueh S H,et al.Observations of soil moisture using a passive and active low-frequency microwave airborne sensor during SGP99 [J]. IEEE Transactions on Geoscience & Remote Sensing,2002,40(12):2659-2673
[9] Chauhan N S,Miller S,Ardanuy P.Spaceborne soil moisture estimation at high resolution:a microwave-optical/IR synergistic approach[J].International Journal of Remote Sensing,2003,24(22):4599-4622
[10] Werbylo K L,Niemann J D.Evaluation of sampling techniques to characterize topographically-dependent variability for soil moisture downscaling[J].Journal of Hydrology,2014,516:304-316
[11] Loew A,Mauser W.On the disaggregation of passive microwave soil moisture data using a priori knowledge of temporally persistent soil moisture fields[J].IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(3):819-834
[12] 张宏鸣,陈丽君,刘雯,等.基于 Stacking 集成学习的夏玉米覆盖度估测模型研究[J].农业机械学报,2021,52(7):195-202
ZHANG Hongming,CHEN Lijun,LIU Wen,et al.Estimation of summer corn fractional vegetation coverage based on Stacking ensemble[J].Transactions of the Chinese Society for Agricultural Machinery,2021,52(7):195-202
[13] 刘佩佩,宋海清,鲍炜炜,等.CLDAS 和 GLDAS 土壤温度数据在陕西省的适用性评估[J].气象科技,2021,49(4):604-611
LIU Peipei,SONG Haiqing,BAO Weiwei,et al.Applicability evaluation of CLDAS and GLDAS soil temperature data in Shaanxi province[J].Meteorological Science and Technology,2021,49(4):604-611
[14] 师春香,姜立鹏,朱智,等.基于 CLDAS2.0 驱动数据的中国区域土壤湿度模拟与评估[J].江苏农业科学,2018,46(4):231-236
SHI Chunxiang, JIANG Lipeng, ZHU Zhi, et al. Simulation and assessment of soil moisture in China based on CLDAS2.0 driven data[J].Jiangsu Agricultural Sciences,2018,46(4):231-236
[15] 卢晨媛,冯文兰,王永前,等.不同深度土壤水分同化产品在川西高原的应用[J].水土保持通报,2021,41(1):173-181
LU Chenyuan, FENG Wenlan, WANG Yongqian, et al. Application of soil moisture assimilation products at different depths in western Sichuan Plateau[J].Bulletin of Soil and Water Conservation,2021,41(1):173-181
[16] 孙丽君,刘晓,朱燕煌.基于 MODIS 数据的河南省干旱程度反演研究[J].测绘与空间地理信息,2021,44(3):140-142,145
SUN Lijun,LIU Xiao,ZHU Yanhuang,et al.Study on the inversion of drought degree in Henan province based on MODIS data[J].Geomatics & Spatial Information Technology,2021,44(3):140-142,145
[17] Crystal B S,Feng G,Alan H S,et al.First operational BRDF, albedo nadir reflectance products from MODIS [J]. Remote Sensing of Environment, 2002, 83(1):135-148
[18] 管延龙,王让会,姚建,等.气候变化背景下天山区域地表反照率特征分析[J].干旱区地理,2015,38(2):351-358
GUAN Yanlong, WANG Ranghui, YAO Jian, et al. Features of surface albedo of Tianshan Mountains area under the background of climate change[J].Arid Land Geography,2015,38(2):351-358
[19] Wang Z S,Schaaf C B,Sun Q S,et al.Capturing rapid land surface dynamics with Collection V006 MODIS BRDF/NBAR/Albedo (MCD43) products[J].Remote Sensing of Environment: an Interdisciplinary Journal, 2018,207:50-64
[20] 张浩彬,李俊生,向南平,等.基于 MODIS 地表反射率数据的水体自动提取研究[J].遥感技术与应用,2015,30(6):1160-1167
ZHANG Haobin,LI Junsheng,XIANG Nanping,et al.A study of extracting water bodies automatically based on the MODIS surface reflectance data[J].Remote Sensing Technology and Application,2015,30(6):1160-1167
[21] Wei Z S,Meng Y Z,Zhang W,et al.Downscaling SMAP soil moisture estimation with gradient boosting decision tree regression over the Tibetan Plateau [J]. Remote Sensing of Environment,2019,225:30-44
[22] 张朝忙,刘庆生,刘高焕,等.SRTM 3 与 ASTER GDEM 数据处理及应用进展[J].地理与地理信息科学,2012,28(5):29-34
ZHANG Chaomang,LIU Qingsheng,LIU Gaohuan,et al.Data processing and application progress of SRTM 3 and ASTER GDEM[J].Geography and Geo-Information Science,2012,28(5):29-34
[23] 孙帅,师春香,梁晓,等.不同陆面模式对我国地表温度模拟的适用性评估[J].应用气象学报,2017,28(6):737-749
SUN Shuai, SHI Chunxiang, LIANG Xiao, et al. Assessment of ground temperature simulation in China by different land surface models based on station obvervations[J].Journal of Applied Meteorological Science,2017,28(6):737-749
[24] 黄飞龙,李昕娣,黄宏智,等.基于 FDR 的土壤水分探测系统与应用[J].气象,2012,38(6):764-768
HUANG Feilong,LI Xindi,HUANG Hongzhi,et al.Soil moisture detection system based on FDR and its application[J].Meteorological Monthly,2012,38(6):764-768
[25] 韩帅.基于 CLDAS 驱动数据的 CLM3.5 和 SSIB2 陆面模式模拟评估及干旱监测应用[D].南京:南京信息工程大学,2015
HAN Shuai.The simulation and evaluation using CLM3.5 and SSIB2 land surface model based CLDAS forcing data with drought monitoring[D].Nanjing:Nanjing University of Information Science & Technology,2015
[26] Crow W T,Berg A A,Cosh M H,et al.Upscaling sparse ground-based soil moisture observations for the validation of coarse-resolution satellite soil moisture products[J]. Reviews of Geophysics,2012,50:1-20
[27] Gao B C.NDWI:a normalized difference water index for

remote sensing of vegetation liquid water from space[J]. Remote Sensing of Environment,1996,58(3):257-266

[28] 马春锋,王维真,吴月茹,等.基于MODIS数据的黑河流域土壤热惯量反演研究[J].遥感技术与应用,2012,27(2):197-207
MA Chunfeng, WANG Weizhen, WU Yueru, et al. Research on soil thermal intertia retrieval in Heihe river basin based on MODIS data[J]. Remote Sensing Technology and Application,2012,27(2):197-207

[29] Finn M P,Lewis M,Bosch D D,et al.Remote sensing of soil moisture using airborne hyperspectral data[J].GIScience & Remote Sensing,2011,48(4):522-540

[30] Lobell D B, Asner G P. Moisture effects on soil reflectance[J].Soil Science Society of America Journal, 2002,66(3):722-727

[31] Whiting M L,Lin L,Ustin S L.Predicting water content using Gaussian model on soil spectra[J].Remote Sensing of Environment,2004,89(4):535-552

[32] Fensholt R,Sandholt I.Derivation of a shortwave infrared water stress index from MODIS near and shortwave infrared data in a semiarid environment[J].Remote Sensing of Environment,2003,87(1):111-121

[33] Grote K,Anger C,Kelly B,et al.Characterization of soil water content variability and soil texture using GPR groundwave techniques[J]. Journal of Environment & Engineering Geophysics,2010,15(3):93-110

[34] Long D,Bai L L,Yan L,et al. Generation of spatially complete and daily continuous surface soil moisture of high spatial resolution[J]. Remote Sensing of Environment,2019,233:111364

[35] Hu F M,Wei Z S,Zhang W,et al.A spatial downscaling method for SMAP soil moisture through visible and shortwave-infrared remote sensing data[J].Journal of Hydrology,2020,590:125360

[36] Peilin S,Huang J F,Lamin R,et al.An improved surface soil moisture downscaling approach over cloudy areas based on geographically weighted regression[J].Agricultural and Forest Meteorology,2019,275:146-158

[37] Zhao W,Sánchez N,Lu H,et al.A spatial downscaling approach for the SMAP passive surface soil moisture product using random forest regression[J].Journal of Hydrology,2018,563:1009-1024

[38] Li J H,Wang S S,Grant G,et al.A model for downscaling SMOS soil moisture using Sentinel-1 SAR data[J].International Journal of Applied Earth Observations and Geoinformation,2018,72:109-121

[39] 徐继伟 ,杨云.集成学习方法:研究综述[J].云南大学学报(自然科学版),2018,40(6):1082-1092
XU Jiwei,YANG Yun.A survey of ensemble learning approaches[J].Journal of Yunnan University (Natural Sciences Edition),2018,40(6):1082-1092

[40] 张若愚.基于梯度提升机算法的变压器油中溶解气体含量短期预测方法[D].北京:华北电力大学(北京),2020
ZHANG Ruoyu. Short-term prediction method for dissolved gas in transformer oil based on gradient boosting machine [D]. Beijing: North China Electric Power University (Beijing),2020

[41] Zhan D Y,Zhang W,Huang W,et al.Upscaling of surface soil moisture using a deep learning model with VIIRS RDR [J]. ISPRS International Journal of Geo-Information,2017,6(5):1-20

[42] 刘斌,李立欣,李静.一种改进的基于深度前馈神经网络的极化码BP译码算法[J].移动通信,2019,43(4):8-14
LIU Bin, LI Lixin, LI Jing. An improved polar BP decoding algorithm based on deep feedforward neural network[J].Mobile Communications,2019,43(4):8-14

[43] Breiman L.Random forests machine learning[J].Journal of Clinical Microbiology,2001,2:199-228

[44] Liu-Yang X Y,Yang Y P,Jing W L,et al.Comparison of different machine learning approaches for monthly satellite-based soil moisture downscaling over Northeast China[J].Remote Sensing,2017,10(2):31

[45] Abbaszadeh P, Moradkhani H, Zhan X. Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method[J].Water resources research, 2019,55(1):324-344

[46] 阿里木·赛买提.基于集成学习的全极化SAR图像分类研究[D].南京:南京大学,2015
ALIM Samat. Ensemble learning based full polarimetric SAR image classification [D]. Nanjing: Nanjing University,2015

[47] 李阳,黄伟,席建忠.基于Stacking算法集成模型的电厂$NO_x$排放预测[J].热能动力工程,2021,36(5):73-81
LI Yang,HUANG Wei,XI Jianzhong.$NO_x$ emission forecasting based on stacking ensemble model[J].Journal of Engineering for Thermal Energy and Power, 2021, 36(5):73-81

[48] 宋相法.基于稀疏表示和集成学习的若干分类问题研究[D].西安:西安电子科技大学,2013
SONG Xiangfa.Study of classification problems based on sparse representation and ensemble learning[D].Xi'an: Xidian University,2013

[49] 蔡毅,朱秀芳,孙章丽,等.半监督集成学习综述[J].计算机科学,2017,44(增刊1):7-13
CAI Yi, ZHU Xiufang, SUN Zhangli, et al. Semi-supervised and ensemble learning:a review[J].Computer Science,2017,44(sup1):7-13

[50] 李磊,沈润平,黄安奇,等.土壤质地改变对CLDAS/Noah-MP土壤湿度模拟的影响研究[J].高原气象,2021,40(3):621-631
LI Lei, SHEN Runping, HUANG Anqi, et al. Impact of soil texture on the simulation of CLDAS/Noah-MP on simulating soil moisture[J].Plateau Meteorology,2021, 40(3):621-631

[51] 杜方洲,石玉立,盛夏.基于深度学习的TRMM降水产品降尺度研究:以中国东北地区为例[J].国土资源遥感,2020,32(4):145-153
DU Fangzhou,SHI Yuli,SHENG Xia.Research on downscaling of TRMM precipitation products based on deep learning: exemplified by Northeast China [J]. Remote Sensing for Land & Resources,2020,32(4):145-153

# Downscaling of CLDAS soil moisture based on ensemble learning method

HAN Huimin[1] SHEN Runping[1] HUANG Anqi[1] DI Wenli[1]
1 School of Geographical Sciences,Nanjing University of Information Science & Technology,Nanjing 210044

**Abstract** Soil moisture is a key parameter of the water cycle and energy budget in terrestrial ecosystems.Land data assimilation system can provide spatio-temporally continuous soil moisture data,however,its low spatial resolution limits the further application.Here,the soil moisture output in 0-10 cm soil layer from China Meteorological Administration Land Data Assimilation System (CLDAS-V2.0) was downscaled from 6 km to 1 km in North China by three single models (gradient boosting machine,deep feedforward neural network and random forest) and a Stacking ensemble learning method.The downscaled results for period of April to October in 2019 show that the four downscaling methods can reflect the temporal and spatial variation of soil moisture in North China and somehow alleviate the overestimation of CLDAS products.Both the spatial distribution details and accuracies are improved compared with original CLDAS soil moisture data.Furthermore,the Stacking ensemble learning method outperforms the other three in downscaling performance,including its highest correlation coefficient with observed data ($R$=0.756 8) and lowest error (RMSE=0.050 5 $m^3/m^3$,Bias=-0.005 2 $m^3/m^3$).Meanwhile,the downscaled results by Stacking ensemble learning are also highly correlated with the dynamic changes of soil moisture,with lowest RMSE and bias compared with station observations,followed by random forest and deep feedforward neural network.

**Key words** soil moisture;ensemble learning;downscaling;CLDAS