融合多源数据的高分辨率土壤水分模拟模型构建及应用

付平凡[1,2]，杨晓静[1,2※]，姜 波[3]，苏志诚[1,2]，孙东亚[1,2]

（1. 中国水利水电科学研究院，北京 100038；2. 水利部防洪抗旱减灾工程技术研究中心，北京 100038；
3. 吉林省墒情监测中心，长春 130033）

摘 要：实时动态高分辨率土壤水分产品可为区域农业生产安全保障提供重要支撑，目前常用的土壤水分遥感产品存在空间分辨率较低及时间序列不连续等问题。为了生成时空连续的高分辨率土壤水分结果，该研究引入集成学习中的随机森林（random forest, RF）和梯度提升机（grandient boosting machine, GBM）算法，构建了融合多源数据的高分辨率土壤水分模拟（high-resolution soil moisture simulation, HRSMS）模型。2017—2022 年 SMAP 微波土壤水分、植被指数、地表温度等遥感数据和墒情站点实测数据为模型输入和输出，利用 Savitzky-Golay 滤波方法和多元回归方法填补缺失的植被指数和地表温度数据，基于 RF 和 GBM 算法实现 SMAP 表层（0～5 cm）土壤水分数据分辨率提升（从 9 km 提高至 1 km）。以吉林省为例验证模型可行性，结果表明：1）HRSMS 模型相较于常用的多项式回归拟合法精度显著提升。均方根误差（root mean square error, RMSE）、平均绝对误差（mean absolute error, MAE）较多项式回归拟合法精度降低了 22.2%、43.9%，决定系数（R^2）提高了 0.270，西北部粮食主产区的误差减少了 33.2%；2）HRSMS 模型中，RF 与 GBM 算法计算效能相近，在吉林省开展相关研究时可结合数据条件任选其一进行模型构建。HRSMS 模型有效提升了土壤水分遥感数据产品的分辨率和精度，对进一步提升土壤水分精准监测能力具有重要意义。

关键词：土壤水分；随机森林；梯度提升机；SMAP SSM；降尺度；点面数据融合

doi：10.11975/j.issn.1002-6819.202407009

中图分类号：TP181；S152.7 **文献标志码**：A **文章编号**：1002-6819(2025)-05-0096-11

付平凡，杨晓静，姜波，等. 融合多源数据的高分辨率土壤水分模拟模型构建及应用[J]. **农业工程学报**，2025，41(5)：96-106. doi: 10.11975/j.issn.1002-6819.202407009 http://www.tcsae.org

FU Pingfan, YANG Xiaojing, JIANG Bo, et al. Construction and application of a high-resolution soil moisture simulation model integrating multi-source data[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2025, 41(5): 96-106. (in Chinese with English abstract) doi: 10.11975/j.issn.1002-6819.202407009 http://www.tcsae.org

0 引 言

土壤水分是表征区域水循环过程、农业灌溉调控及气候变化响应的关键参数，同时在水文过程模拟、气象预测及农业生产实践中具有重要的研究价值[1]。土壤含水量作为地表植被吸收水分的主要来源，对作物的生长发育至关重要[2]。准确监测土壤含水量对作物增产和粮食安全具有重要意义。

目前土壤水分数据获取方式主要为地面站点实测和遥感卫星反演，地面实测数据具备时间序列长、精度高等优点，但难以获取大范围、连续的土壤水分数据。遥感卫星数据具有空间覆盖范围大、动态周期性强、产品性价比高等优点，现已广泛应用于农业干旱监测等领域[3-5]。当前可从不同传感器中获取土壤含水量，如 AMSR-E（advanced microwave scanning radiometer-earth observing system）[6]、AMSR-2（advanced microwave scanning radiometer 2）[7]、Sentinel-2[8-9]、SMOS（soil moisture and ocean salinity mission）[10-11]、SMAP（soil moisture active passive mission）[12]、ESA-CCI（the european space agency's climate change initiative）[13]等。但上述遥感产品的空间分辨率大多为 10～40 km，无法满足现有区域农业旱情精细化监测的需求[14]。如何融合站点实测数据及遥感产品优势，提升土壤水分产品时空分辨率是当前土壤水分模拟预测研究中亟待解决的关键技术问题之一。

将低分辨率土壤水分产品转换为高分辨率产品的研究越来越多，目前常见的方法主要分为 4 类：1）基于卫星产品融合的方法，包括主被动数据融合[15-16]和光学和微波数据融合[17-19]；2）基于地理信息的方法，包括地统计模型[20]和分形插值模型[21]；3）基于数理模型的方法，包括统计模型[22]、陆面模型[23]和数据同化[24]；4）基于机器学习的算法，包括集成学习[25-26]和深度学习[27-28]。基于卫星产品融合的方法相对直接，已广泛应用于土壤水分降尺度研究中，该方法需要借助植被覆盖和地表温度（land surface temperature, LST）信息，仅适用于晴空条件[29]，其中典型的方法是多项式回归拟合法[30]。植被和地表温度数据对于土壤水分模拟至关重要，其数据的完整性将会严重影响土壤水分降尺度结果的连续性与准

收稿日期：2024-07-01 修订日期：2024-11-07
基金项目：国家重点研发计划项目（2023YFC3206001）；江西省重点研发计划项目（20232BBG70029）；江西省“科技+水利”联合计划项目（2022KSG01002）
作者简介：付平凡，博士生，研究方向为干旱监测预警等。
Email：fupf123456@163.com
※通信作者：杨晓静，高级工程师，研究方向为水灾害与水安全、旱情旱灾监测预报预警信息化等。 Email：yangxj@iwhr.com

确性[19]。当前植被指数的时间分辨率为 5、7、10、16 和 30 d，为获得连续的植被指数时间序列，插值、滤波、函数拟合等方法被用于植被数据重构[31]，其中 Savitzky-Golay（S-G）滤波方法广泛用于植被数据重构[32-33]。已有研究表明，全球约 65% 表面被云层覆盖[34]，导致云层覆盖区域内数据大面积缺失。处理地表温度信息缺失的过程中，部分学者利用 ATC 模型[35] 以及多元回归模型[28] 填补了缺失地表温度数据。综上所述，植被和地表温度数据对生成高分辨率土壤水分结果十分重要，相关数据重构方法也相对成熟，但将重构后数据应用于土壤水分模拟预测研究还相对较少。

近年来集成学习凭借其强大的泛化能力和从高维数据发现隐藏规律的优势，已广泛用于土壤盐分反演[36]、油菜花期预测[37]、番茄根长表型提取[38]、植被覆盖[39-40] 和地表温度[41-42] 重构研究，但将集成学习算法应用于高分辨率土壤水分生成全过程（包括辅助数据重构和土壤水分降尺度）的研究还相对较少。已有研究分析了不同机器学习算法间的性能差异[43]，但对比不同降尺度方法精度差异的研究还亟待进一步深入。

针对目前土壤水分模拟研究中存在输入数据时空不连续、不同降尺度方法精度差异不明确等问题，本文优选集成学习 Bagging 和 Boosting 中典型的随机森林（random forest, RF）梯度提升机（grandient boosting machine, GBM）算法，构建融合多源遥感数据和地面实测站点的高分辨率土壤水分模拟模型（high-resolution soil moisture simulation, HRSMS 模型），研究拟通过构建 HRSMS 模型解决光学/热红外遥感产品时空不连续问题，对比不同降尺度方法及 RF 和 GBM 算法在生成高分辨率土壤水分结果方面的性能，融合多源数据的优势实现 SMAP 表层（0～5 cm）微波土壤水分（soil moisture active passive surface soil moisture, SMAP SSM）遥感产品空间分辨率和精度提升，以期更好地服务粮食主产区旱情动态精准监测，为农业干旱预报、预警提供关键技术产品支撑。

1　材料和方法

1.1　研究区概况

吉林省（121°38′E～131°19′E、40°50′N～46°19′N）地处中国东北地区腹地，全省面积 18.74 万 km^2，总人口 2 347.69 万人。全省地貌差异显著，呈东南高、西北低的特征。该区域是著名的“黑土地之乡”，是中国重要的粮食主产区和商品粮基地。东部山地主要为长白山和低山丘陵区，中西部平原主要为松嫩平原和辽河平原。根据 2023 年中国统计年鉴，吉林省粮食产量位于全国第五（4 080.8 万 t），主要粮食作物玉米的产量位于全国第二（3 257.9 万 t），占全国玉米产量的 11.8%。

1.2　数据收集与处理

1.2.1　遥感数据

研究采用 6 种类型遥感数据：土壤水分、降水、植被指数、地表温度、土壤质地和数字高程数据，数据源信息详见表 1。

表 1　遥感产品数据集

Table 1　Remote sensing product datasets

数据集 Dataset	数据描述 Data details	数据来源 Data source	空间分辨率 Spatial resolution	时间分辨率 Temporal resolution	数据地址 Data address
Precipitation	CHIRPS v2.0 data	加州大学圣巴巴拉分校气候灾害中心（Climate Hazards Center，UC Santa Barbara）	0.05°	1 d	https://www.chc.ucsb.edu/data/chirps
SMAP Level 4 Soil Moisture	Version 4: Vv6032; Surface (0～5 cm)	美国国家冰雪数据中心（NSIDC）	9 km	3 h (09:00)	https://nsidc.org/data/smap/data
ERA5-LAND Vegetation	LAI	欧盟中期天气预报中心（ECMWF）	9 km	1 h (08:00)	https://cds.climate.copernicus.eu/cdsapp
ERA5-LAND LST	Temperature	欧盟中期天气预报中心（ECMWF）	9 km	1 h (08:00)	https://cds.climate.copernicus.eu/cdsapp
Vegetation	MODIS MOD13A2 v006 – NDVI, EVI	美国航天局地球观测系统数据和信息系统（NASA EOSDIS LPDAAC）	1 km	16 d	https://ladsweb.modaps.eosdis.nasa.gov/
LST	MODIS MOD11A1 v006–LST Day and Night Times	美国航天局地球观测系统数据和信息系统（NASA EOSDIS LPDAAC）	1 km	1 d	https://ladsweb.modaps.eosdis.nasa.gov/
Elevation	DEM、Slope	中国科学院资源环境数据云平台	1 km	静态数据	https://www.resdc.cn/
Soil Texture	Clay、Silt、Sand	中国科学院资源环境数据云平台	1 km	静态数据	https://www.resdc.cn/

土壤水分数据为 SMAP 数据，时间分辨率为 3 h，时间范围从 2015 年 3 月 31 日至今，其数据可覆盖到全球±45°纬度内的陆地区域[44]。降水（Precipitation）数据 CHIRPS（climate hazards group infrared precipitation with station data）空间分辨率为 0.05°（南纬 50°～北纬 50°和所有经度）、时间序列为 1981 年至今的网格降水数据。植被指数和地表温度数据来源分别为 MODIS（moderate resolution imaging spectroradiometer）和 ERA5-Land 第五代再分析数据。ERA5-Land 数据时空分辨率分别为 1 h 和 0.1°，时间跨度为 1950 年至今。数字高程模型数据（digital elevation model, DEM）和土壤质地数据为静态数据，土壤质地（soil texture）分为砂土（sand）、粉砂土（silt）、与黏土（clay）三大类，坡度（slope）数据基于 1 km 分辨率 DEM 数据生成。以上所有数据均选取 2017—2022 年间日值，其中 SMAP和 ERA5-Land 数据分别选择每日 09:00 和 08:00 时作为当日数据。

1.2.2　站点数据

地面墒情站点共有 305 个，其中 262 个自动站和 43 个人工站。自动站利用埋设在地下（10、20 和 40 cm）的传感器每隔 1 h 获取数据，人工站每隔 10 d 获取一次

数据。数据时间范围为 2011 年至今，由于 2017 年以前墒情站点数据连续性较差，故选择 2017—2022 年墒情数据作为集成学习的输出部分，经过数据序列完整性筛选后，共 303 个站点符合研究要求，从其中随机选 3 个站点作为验证站点，其余站点作为训练站点（图 1）。

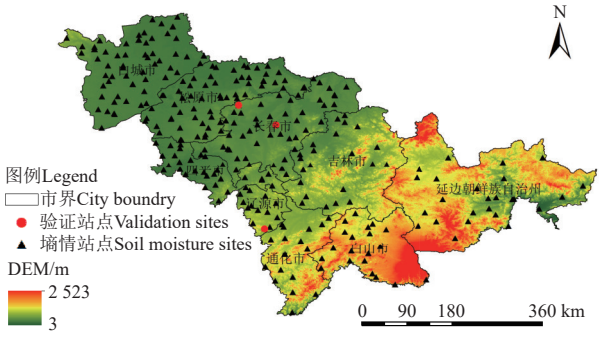


图 1 实测墒情站点空间分布

Fig.1 Spatial distribution of measured soil moisture stations

基于 2017—2022 年 300 个墒情站点数据共 273 392 条数据，按照 8:2 的划分规则将数据集划分为训练集和测试集，数据量分别为 154 416 和 38 605，土壤含水率统计结果如表 2 所示。

表 2 数据集统计分析结果

Table 2 Results of statistical analysis of data sets

数据集 Data set	最大值 Maximum/ ($m^3 \cdot m^{-3}$)	最小值 Minimum/ ($m^3 \cdot m^{-3}$)	均值 Mean/ ($m^3 \cdot m^{-3}$)	标准差 Standard deviation/ ($m^3 \cdot m^{-3}$)
总体数据集 Overall data sets	0.969	0.000	0.159	0.086
训练集 Training set	0.969	0.000	0.159	0.086
测试集 Test set	0.770	0.000	0.159	0.086
验证集 Validation set	0.038	0.043	0.181	0.052

1.3 融合多源数据的高分辨率土壤水分模拟模型（HRSMS 模型）构建方法

HRSMS 模型结构包括：1）多源数据预处理模块。将收集到的多源遥感数据（SMAP 地表土壤水分数据、植被数据、地表温度数据、降水数据、地形数据和土壤质地数据）进行投影、裁剪、重采样处理。地面实测墒情站点经过质量控制后，筛选出 303 个时间序列完整、无异常值的地面站点。2）时空不连续数据重构模块。针对 MODIS 植被指数时间分辨率（16 d）过长，地表温度数据由于云雨遮盖导致数据缺失过多等问题，构建 HRSMS 模型前，利用 S-G 滤波方法和多元回归方法（基于 RF 算法）将植被指数和地表温度数据进行重构，保证输入数据的连续性和完整性。3）高分辨率土壤水分生成模块。分为 2 个部分：①模型预训练阶段。将上述所有重构后时空连续的遥感数据提取至墒情站点处，遥感格点数据作为输入，实测墒情站点值作为输出，并对 RF 和 GBM 算法进行参数调优和模型验证；②高分辨率土壤水分结果生成。将连续的每日遥感格点数据输入训练后的模型中，生成高时空分辨率（1 d，1 km）的土壤水分结果。模型技术路线如图 2 所示。

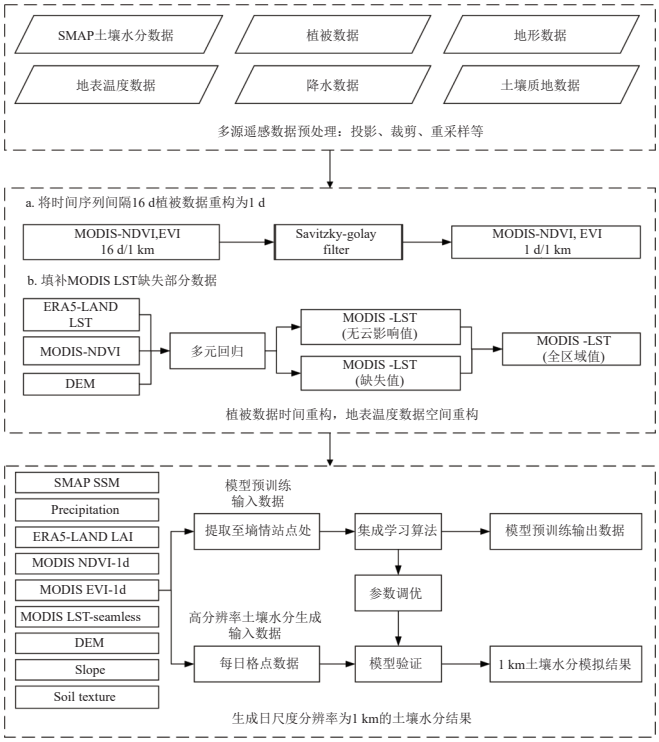


图 2 HRSMS 模型流程图

Fig.2 HRSMS model flowchart

1.3.1 多源数据预处理

由于 SMAP 地表土壤水分数据和 ERA5-Land 地表温度、植被指数数据分辨率为 9 km，为保证输入数据及重构数据分辨率的统一，将该数据（共 2 130 d）按照 MODIS 网格系统重采样为 1 km 并批量转为 WGS84 坐标系，统一裁剪至吉林省边界范围。

1.3.2 时空不连续数据重构

1）植被数据重构

由于植被指数时间分辨率为 16 d，无法满足生成每日高分辨率土壤水分结果的需求，因此需对数据进行重构处理。本研究使用 S-G 滤波方法[45]，通过调用 Python 中 savgol_filter 函数将 16 d 的 MODIS NDVI/EVI 时间序列重构为 1 d。savgol_filter 函数包含 2 个参数，分别为 window_length 和 K 值，window_length 必须为正奇数，值越小，曲线越贴近真实曲线；值越大，平滑效果越好。K 值必须小于 window_length 的长度，K 值越大，曲线越贴近真实曲线；K 值越小，曲线平滑越好。

2）地表温度重构

MODIS 提供的 LST 产品时间分辨率为 1 d，空间分辨率为 1 km。由于雨云的遮盖，MODIS LST 产品通常有很大比例的缺失区域，ERA5-Land LST 的产品未有缺失部分，其空间分辨率较粗（9 km），无法满足输入数据分辨率要求。因此，本研究基于 LST 与其相关控制因素（例如 NDVI，DEM），提出了一种基于 LST 相关变量的多元回归模型和时空融合模型相结合的 LST 填补方法，该方法结合二者优势生成 1 km 无缝每日地表温度。计算式为

$$LST_{value} = a \cdot LST_{ERA5\_value} + b \cdot NDVI_{1d\_value} + c \cdot DEM_{value} + d \quad (1)$$

$$LST_{non\_value}=a\cdot LST_{ERA5\_non\_value}+b\cdot NDVI_{1d\_non\_value}+c\cdot DEM_{non\_value}+d \quad (2)$$

式中 LST_{value} 和 LST_{non_value} 分别为 MODIS LST 中有值和空值区域；LST_{ERA5_value} 和 $LST_{ERA5\_non\_value}$ 分别为 MODIS LST 对应区域的 ERA5-LAND LST 有值和空值数据；$NDVI_{1d\_value}$ 和 $NDVI_{1d\_non\_value}$ 分别为 MODIS LST 对应区域的 1 d NDVI 有值和空值数据；DEM_{value} 和 DEM_{non_value} 分别为 MODIS LST 对应区域的 DEM 有值和空值数据；a，b，c，d 为模型的参数，最终将 LST_{value} 和 LST_{non_value} 进行拼接生成每日地表温度结果（$LST_{seamless}$）。

1.3.3　高分辨率土壤水分模拟模型

本研究通过优选 Bagging、Boosting 中 RF[46] 和 GBM[47-48] 算法构建表层高分辨率土壤水分模拟模型，模型具体构建过程如下：

1）模型预训练。SMAP L4 全球土壤水分数据集包含表层（0～5 cm）和根层（0～100 cm）土壤水分数据，由于不同深度土壤水分差距较大，因此选择跟 10 cm 深度较为接近的表层（0～5 cm）数据作为 HRSMS 模型的输入。具体操作为：将所有时空连续的多源遥感数据与地面实测墒情站点进行空间经纬度匹配，获取地面实测墒情站点所在点位的遥感像元值，将上述匹配后的遥感像元值作为输入数据，将预处理后的 10 cm 深度实测墒情站点数据作为输出数据，构建模型训练数据集，并进行模型参数调优和验证。

2）高分辨率土壤水分结果生成。利用 2017—2022 年的数据，在 300 个墒情站点所在 1 km 格点处进行训练。为获得区域内所有 1 km 格点的土壤水分结果，需将所有时空连续且重采样为 1 km 的每日多源遥感数据输入至上述训练后的模型中，最终生成表层日尺度空间分辨率为 1 km 的土壤水分结果（$SSM_{downscaled}$）。模型具体可表示为

$$SSM_{downscaled}=f\left(\begin{matrix}SMAP_{SSM},ERA5_{LAI},LST_{seamless},NDVI_{1d},\\EVI_{1d},Precipitation,Soil\ texture,DEM,Slope\end{matrix}\right) \quad (3)$$

式中 f 为 RF 和 GBM 算法，$SMAP_{SSM}$ 为 SMAP-L4 地表 0～5 cm 深度土壤水分数据，$ERA5_{LAI}$ 为 ERA5-Land 叶面积指数数据，$LST_{seamless}$ 为 ERA5-Land LST 和 MODIS LST 合成的 1 km 无缝日地表温度，其余变量同上文。

1.4　评价指标

选用平均绝对误差（mean absolute error，MAE）、决定系数（R^2）和均方根误差（root mean square error，RMSE）3 种指标分别对梯度提升机、随机森林算法的预测效果进行评估。各指标计算方法见文献 [26]。MAE 是绝对误差的平均值，它能够反映预测值误差的实际情况。RMSE 是含水量估计值与真值之差的平方的期望值，可以评价数据的变化程度。R^2 可以消除维数对评价测度的影响，MAE 和 RMSE 越小表明预测结果越好，R^2 越大表明预测结果越好。

2　结果与分析

2.1　指数重构结果

2.1.1　植被指数重构

将 2017—2022 年 16 d 的原始 NDVI/EVI 序列利用 S-G 滤波方法进行拟合，通过调用 Python 中 savgol_filter 函数实现。由于 NDVI 时间间隔为 16 d，为了兼顾曲线拟合和平滑效果，选择 window_length 长度为 7（该值必须为正奇数），此长度刚好能将 NDVI 时间序列进行 2 次拟合，并且能够保持曲线平滑。已有研究表明[32]，与 NDVI 变化趋势不一致的突然下降异常点应视为受云或其他原因影响的噪声点。为消除噪声点的影响，于是将 K 值从 6 至 3 逐步进行 S-G 滤波，结果表明 K 值为 6 时拟合了噪声点，K 值为 5 和 4 时既能消除噪声点的影响，也能达到拟合 NDVI 高值的效果，K 值为 3 时未能拟合 NDVI 高值结果。考虑到消除噪声点的稳定性，最终确定 K 值为 4 进行滤波处理。确定 window_length 和 K 值后，将滤波后的结果进行插值，最终获得 1 d 的 NDVI/EVI 产品（图 3）。7 月是玉米关键生育期，以 2017 年 7 月 12 日—2017 年 7 月 24 日每隔 4 d 的日尺度 NDVI 结果为例展示植被指数重构结果，重构后的结果表明 NDVI 值变化幅度较小，符合 NDVI 短时间内变化较小的特性[31]。

2.1.2　地表温度重构

2017—2022 年 MODIS LST 每日缺失率结果表明，该数据集在研究区每日的最大缺失比例为 99.1%，最小缺失率为 0.1%，多年平均缺失率为 55.7%，每日缺失 50.0% 以上数据占比为 58.8%。统计原始 MODIS LST 数据发现，每日最大缺失数据量为 213 220，而最小的 5 日非空值数据量大于此值。由于需要重构的数据量过大，在 2017—2022 年期间，将每 5 d 有值数据作为训练集，分批生成这 5 d 内每天缺失的 MODIS LST，然后将每日生成的缺失结果与原始 MODIS LST 非空值区域进行拼接，生成每日空间连续的 LST 结果。基于评价指标评估了 2017—2022 年重构后地表温度与原始 MODIS LST 的精度差异，结果表明 RMSE、MAE 和 R^2 均值分别为 0.526 K、0.338 K 和 0.986（图 4），均达到较好的模拟结果。

图 5 显示了填充缺失 LST 的效果，原始 MODIS LST 数据有大面积的缺失值（图中蓝色部分），LST 重建后空白区域被很好地填充。填补的 LST 空间分布与 ERA5-Land LST 和原始 MODIS LST 具有较高的一致性，且相对 ERA5-Land LST 产品分辨率也有所提高。

2.2　模型参数调优

在进行预测结果评估前，需对 2 种集成学习算法进行超参数调优。对于 RF 模型，需要进行调节的超参数分别为决策树的数量（n_estimators）、决策数的深度（max_depth）、建立决策树时选择的最大特征数目（max_features）和分割所需的最小样本数（min_sample_split）；GBM 模型需要调节的参数为学习率（learning_rate）、损失函数（loss）、决策树的数量（n_estimators）、决策数的深度（max_depth）、建立决策树时选择的最大特征数目（max_features）和叶节点所需最小样本数（min_samples_leaf）。

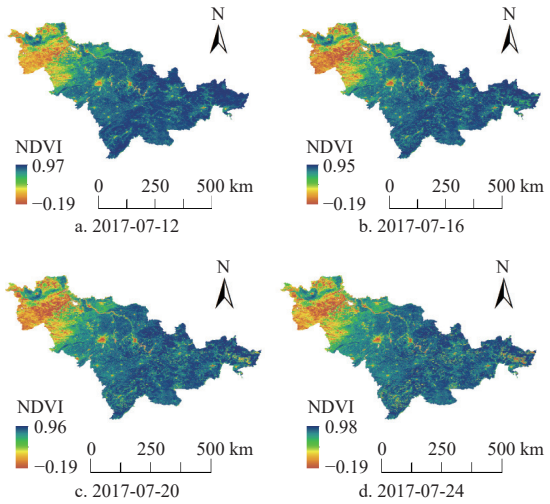
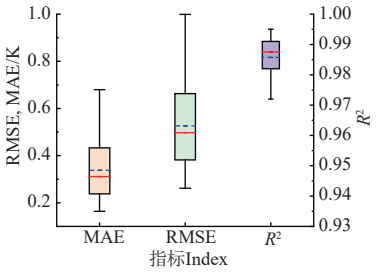


图 3 S-G 滤波后归一化植被指数每日空间结果

Fig.3 Normalized difference vegetation index (NDVI) daily spatial results after S-G filtering



注：虚线为均值，实线为中位数。

Note: Dashed line is the mean value, and solid line is the median.

图 4 2017—2022 年重构地表温度与原始地表温度

Fig.4 Reconstructed land surface temperature (LST) and original surface temperature for 2017-2022

研究中利用 GridSearch 方法[49]对 2 种模型进行 20 次超参数调节，GBM 模型在第 16 次误差最小，损失函数从'ls'，'huber'，'quantile'函数中优选，最终确定最优损失函数为'huber'函数；RF 模型在第 13 次误差最小，RF 模型和 GBM 模型的最优超参数结果如表 3 所示。

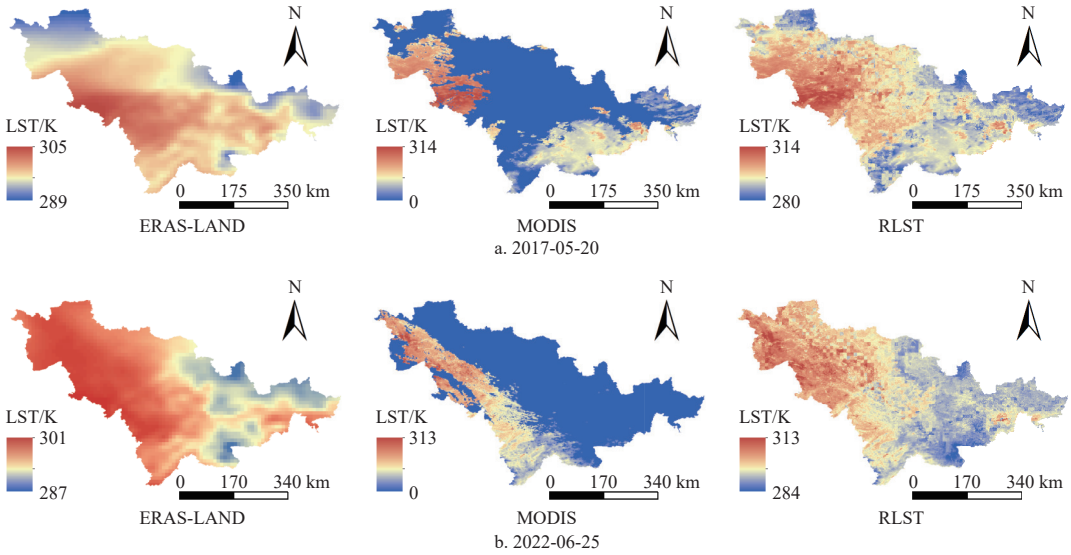


图 5 不同时间 ERA5-Land LST、MODIS LST 和重建 LST 的空间分布

Fig.5 Spatial distribution of ERA5-Land LST, MODIS LST, and reconstructed LST (RLST) at different time

表 3 超参数最优结果

Table 3 Hyperparameter optimization results

算法 Algorithms	超参数 Hyperparameter	开始 Start	结束 End	步长 Step	最优参数 Optimal parameter
随机森林 Random forest (RF)	n_estimators	10	10^3	100	610
	max_depth	10	10^3	100	110
	max_features	1	12	1	6
	min_sample_split	0	1	0.1	10^{-4}
梯度提升机 Grandient boosting machine (GBM)	learning_rate	0	1	0.05	0.1
	n_estimators	10	10^3	100	410
	max_depth	10	10^3	100	20
	max_features	1	12	1	2
	min_samples_leaf	0	10	1	5

2.3 模型验证

为了评估 HRSMS 模型的准确性，分别从时间和空间 2 个尺度进行结果验证。其中时间尺度直接与实测站点的长序列值进行对比，空间尺度则是利用评价指标的空间分布来进行验证。

1）时间尺度验证

随机选择 3 个验证站点，对比分析 10 cm 深度实测墒情站点和模型生成结果（图 6）。降水量主要集中在 5—9 月，SMAP SSM 资料的 SSM 峰值与降水事件符合。RF 和 GBM 算法均能较好拟合出地面观测结果，其中 GBM 算法模拟结果能够更好地反映 SSM 的时间变化特征，其与实测站点的 R^2 均大于 0.991，MAE 和 RMSE 均小于 0.100 m^3/m^3，优于 RF 算法结果。

2）空间尺度验证

为对比不同算法在空间上的准确性，从测试集的 RMSE、MAE 和 R^2 分布结果进行空间精度验证。测试集的验证结果表明，GBM 算法在各个评价指标上的结果均优于 RF 算法，且在西北地区（白城、松原和长春市）的表现结果更好。将评价指标结果划分为 5 个区间，RMSE 和 MAE 最小的区间，R^2 最大的区间为最优区间。

GBM 算法 RMSE 最优区间（0.110～0.410）、MAE 最优区间（0.050～0.160）和 R^2 大于 0.972 的占比分别为 38.0%、99.7% 和 95.7%，RF 算法分别为 0.6%、29.5% 和 74.1%，如图 7 所示。

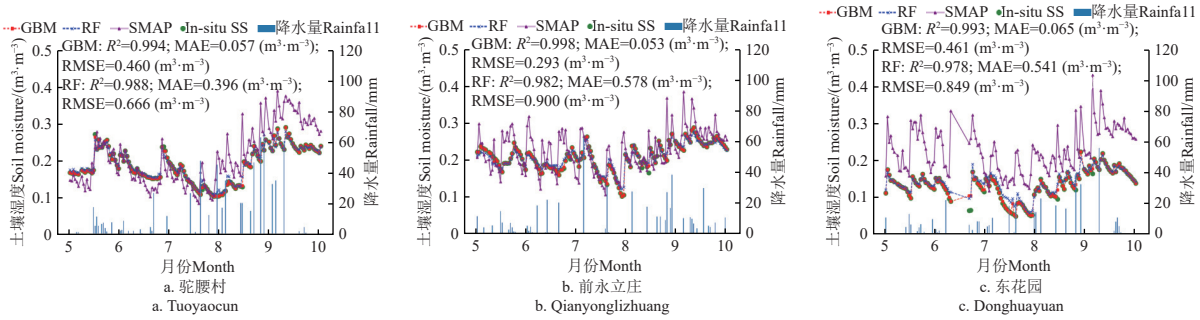


图 6　HRSMS 模型不同算法降尺度结果与 SMAP 产品、实测站点对比结果

Fig.6　Comparison of downscaling results of different algorithms of HRSMS model with SMAP products and measured sites

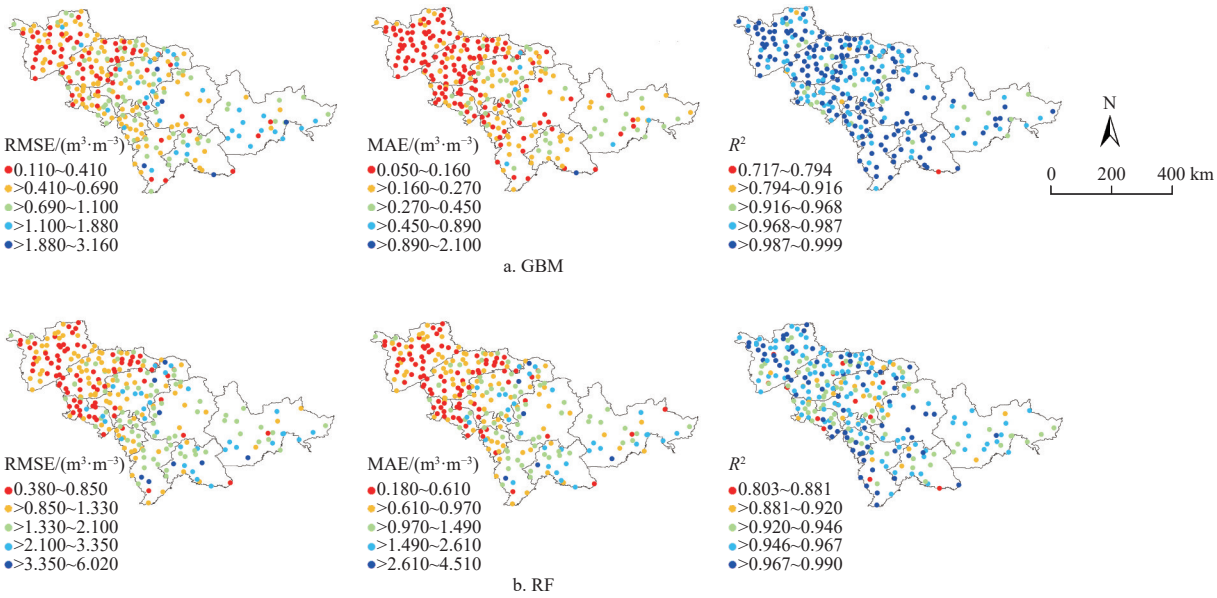


图 7　HRSMS 模型不同算法测试集与实测站点评价指标结果空间分布

Fig.7　Spatial distribution of evaluation metrics results for different algorithms of the HRSMS model for the test set and real sites

2.4　模型应用

2.4.1　高分辨率土壤水分结果生成

吉林省春玉米全生育期为 4—9 月，6—8 月是玉米的关键生育期（拔节-抽穗、抽穗-乳熟）[50-51]。因此随机选取 2018—2020 年玉米生育期 3 d（2018 年 6 月 16 日、2019 年 8 月 15 日和 2020 年 7 月 11 日）数据进行高分辨土壤水分结果生成。碍于篇幅限制，展示 2 d 降尺度模型结果（图 8）。结果表明，2 种算法均能很好地拟合实测站点结果，RF 算法 3 d MAE、RMSE 和 R^2 均值分别为 0.033 m^3/m^3、0.049 m^3/m^3 和 0.574；GBM 算法 3 d MAE、RMSE 和 R^2 均值 0.033 m^3/m^3、0.050 m^3/m^3 和 0.556。

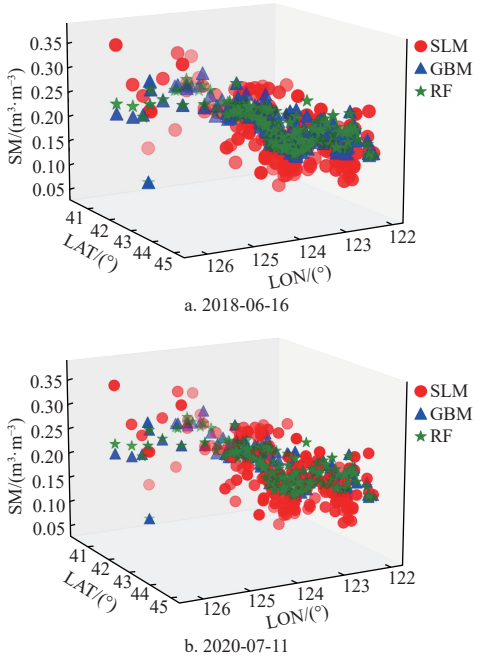
在土壤含水量较低的西北地区（LON 介于 122°～124°，LAT 介于 43°～45°），RF 和 GBM 算法的均值和中位数与实测站点值也较为接近（图 9），RF 和 GBM 算法与实测站点 3 d 平均误差分别为 2.7% 和 3.3%。2018 年 6 月 16 日 RF、GBM 算法和实测站点西北地区土壤水分均值结果分别为 0.120、0.119 和 0.122 m^3/m^3；2020 年 7 月 11 日 RF、GBM 算法和实测站点西北地区土壤水分均值分别为 0.106、0.106 和 0.107 m^3/m^3。

2.4.2　与多项式回归方法对比

为了验证本文构建的 HRSMS 模型较目前主流方法精度有何提升，选取土壤水分降尺度研究中常用的多项式回归拟合方法进行对比分析。多项式回归拟合方法通过将低分辨率土壤水分产品和光学/热传感器获得的植被指数以及地表温度等表面参数建立多项式关系，将该多项式应用于高分辨率遥感产品，进而获得高分辨率土壤水分结果，方法详见文献 [52]。多项式回归拟合方法与 HRSMS 模型生成高分辨率土壤水分结果的区别在于，多项式回归拟合方法是直接利用遥感产品的格点数据建立多项式关系；HRSMS 模型则是将实测墒情站点历史数据训练结果迁移推广至其余未有墒情站点格点处，生成高分辨率土壤水分结果。

为对比不同降尺度方法间的结果差异，利用多项式回归拟合方法建立多项式关系时，同样采用 RF 和 GBM 算法，生成了相同日期的结果（图 10），结果表明多项

式回归拟合方法高估西北部粮食主产区的土壤水分，低估东南部丘陵区的土壤水分。剖析吉林省西北部地区模拟结果发现（图 11），2018 年 6 月 16 日 RF 和 GBM 算法分别比实测站点值高 0.057 m^3/m^3 和 0.056 m^3/m^3，平均高出实际站点值 46.3%。2020 年 7 月 11 日比实测站点值高 0.028 m^3/m^3，平均高出实际站点值 26.2%。西北地区 RF 和 GBM 算法与实测站点 3 d 平均误差分别比多项式回归拟合法误差分别降低了 33.7% 和 32.7%。吉林省西北部地区的粮食产量占全省 63.0%，利用多项式回归拟合方法生成的高分辨率土壤水分结果对该区域进行农业干旱监测时，将会导致对农业旱情的严重误判。



注：LON 为经度；LAT 为纬度；SLM 为实测站点土壤水分。
Note: LON is longitude; LAT is latitude; SLM is measured site soil moisture.

图 8 HRSMS 模型生成与实测 0～10 cm 土层土壤湿度结果三维对比

Fig.8 Three-dimensional comparison of HRSMS model generation with measured soil moisture (SM) results in 0-10 cm soil layer

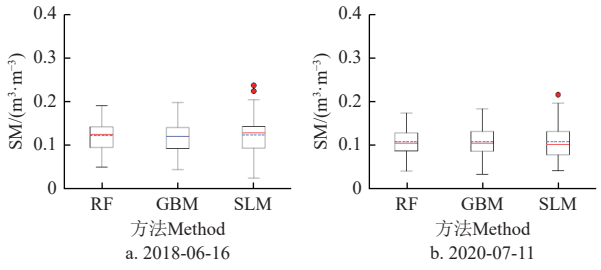


图 9 HRSMS 模型在吉林省西北部生成结果与站点实测土壤湿度值对比箱图

Fig.9 Boxplot of HRSMS model generated soil moisture in northwestern Jilin Province compared with measured values at site

多项式回归拟合方法生成的结果与实测站点值相差较多的原因可能有以下 2 点：1）原始遥感产品时间序列与实测站点序列的相关性不高。虽然 4—10 月的遥感产品和实测站点的时间序列有较高的相关性（平均相关系数为 0.63），但其余月份相关性较差，导致整个数据集与实测站点的相关性较低；2）文中遥感产品仅与有限的站点进行偏差校正，校正效果与墒情站点的数量和分布密切相关，因此导致生成的高分辨率土壤水分结果与实测站点误差较大。

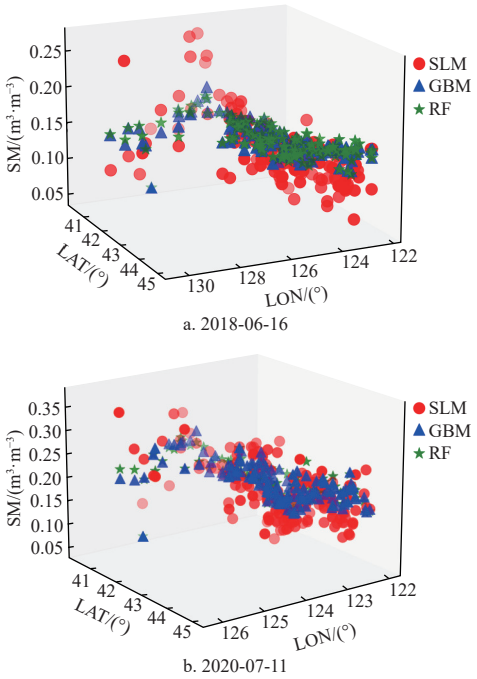


图 10 多项式回归拟合方法生成结果与实测站点结果三维对比

Fig.10 Three-dimensional comparison of the results generated by the polynomial regression fitting method with the results of the measured site

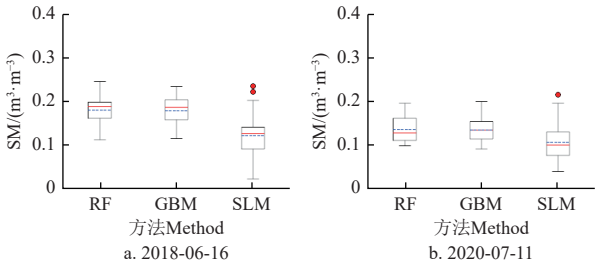


图 11 多项式回归拟合方法在吉林省西北部生成结果与站点实测值对比箱图

Fig.11 Boxplot of the polynomial regression fitting method in northwestern Jilin Province generating results versus measured values at site

HRSMS 模型中的 RF 和 GBM 算法能弥补多项式回归拟合方法因实测站点较少导致的结果偏差，提高了高分辨率土壤水分生成结果精度。其中 RF 算法 3 d RMSE、MAE 均值分别降低了 20.4%、45.5%，R^2 均值提高了 0.253；GBM 算法分别降低了 24.0%、42.4%，R^2 均值提高了 0.286，具体结果如表 4 所示。

HRSMS 模型相较于多项式回归拟合法至少有以下 2 个优点：1）简化了数据处理的流程，保证了输入数据

的时空完整性。HRSMS 模型是通过直接学习实测站点值，不需要校正遥感产品与实测墒情站点间的偏差，大大简化了前期数据质量控制等步骤。HRSMS 模型在模型训练前，将时空不连续的数据进行重构，保证了输入数据的连续性；2）显著提升了降尺度结果的精度。在同样使用 RF 和 GBM 算法的情况下，HRSMS 模型 3 d RMSE、MAE 均值较多项式回归拟合法降低了 22.2%、43.9%，R^2 提高了 0.270。

表 4　不同降尺度方法生成土壤含水量结果对比评估

Table 4　Comparative assessment of soil moisture results generated by different downscaling methods

方法 Methodologies	日期 Date	RF			GBM		
		MAE/ ($m^3 \cdot m^{-3}$)	RMSE/ ($m^3 \cdot m^{-3}$)	R^2	MAE/ ($m^3 \cdot m^{-3}$)	RMSE/ ($m^3 \cdot m^{-3}$)	R^2
HRSMS 模型 HRSMS model	2018-06-16	0.035	0.048	0.582	0.033	0.046	0.624
	2019-08-15	0.034	0.062	0.588	0.035	0.064	0.563
	2020-07-11	0.029	0.037	0.552	0.031	0.040	0.481
	均值 Mean	0.033	0.049	0.574	0.033	0.050	0.556
多项式回归拟合方法 Polynomial regression fitting method	2018-06-16	0.040	0.052	0.285	0.041	0.053	0.237
	2019-08-15	0.057	0.078	0.413	0.062	0.084	0.324
	2020-07-11	0.039	0.048	0.265	0.038	0.048	0.250
	均值 Mean	0.048	0.059	0.321	0.047	0.062	0.270

3　讨 论

基于随机森林和梯度提升机算法对比了 2 种降尺度方法生成的每日高分辨率土壤水分结果（1 km）的精度，结果表明本研究构建的模型相比多项式回归算法有以下优点：1）数据精度有所提高。HRSMS 模型总体 R^2 均值提高了 0.270，西北部粮食主产区的误差减少了 33.2%；2）数据预处理效率提升。模型省去遥感数据与实测站点偏差校正过程，有效缩短数据预处理时长。

表 5 列举了已有土壤水分降尺度研究，结果表明在使用同样算法、同样数据（SMAP 产品）的条件下，输入连续辅助数据相比未输入连续辅助数据生成的土壤水分结果更稳定。

表 5　对比已有表层土壤水分降尺度研究

Table 5　Comparison of recently published studies in the SSM downscaling

方法 Methods	数据 Data	时间分辨率 Spatial resolution	空间分辨率 Temporal resolution	R^2	来源 Source
RF	MODIS, SMAP, precipitation, topography	clear-sky (2015)	1 km	0.20～0.72	[53]
GBM	SMAP, Soil moisture indices	clear-sky (2015–2017)	1 km	0.17/0.09/0.28	[54]
XGBoost	MODIS, SMAP, precipitation, topography	Daily	1 km	0.56～0.97	[55]
RF/GBM	MODIS, SMAP, precipitation, topography	Daily	1 km	0.48～0.62	本文方法

ABBASZADEH 等[53-54] 分别采用了随机森林和梯度提升机算法来提高 SMAP 土壤水分产品的结果，但由于输入 MODIS LST 在云覆盖情况下有很大比例缺失，降尺度后的土壤水分无法实现空间连续性的目标，且精度也相对较差。文献 [53] 的最高 R2 与本文结果相差 0.100，但其最低 R^2 仅为 0.200。KARTHIKEYAN等[55] 输入连续的辅助数据，利用极端梯度提升算法生成了 5、10、20、50 和 100 cm 深度（5 层）的高分辨率土壤水分结果，R^2 相比未输入连续辅助数据的结果显著提高。

4　结　论

本研究构建的融合多源数据的高分辨率土壤水分模拟（high-resolution soil moisture simulation, HRSMS）模型，填补植被指数和地表温度缺失数据，并利用随机森林和梯度提升机算法实现将 9 km 分辨率的被动微波土壤水分产品（soil moisture active passive, SMAP）表层土壤水分数据提高至 1 km，并与多项式回归拟合方法进行了综合对比。主要的结论如下：

1）HRSMS 模型相比于多项式回归拟合法，显著的提升了土壤含水量模拟的精度，且解决了多项式回归拟合法高估吉林省西北部土壤含水量的问题。HRSMS 模型中的 RF 算法相比于多项式回归拟合法，RF 算法 3 d RMSE、MAE 均值分别降低了 20.4%、45.5%，R^2 均值提高了 0.253；GBM 算法分别降低了 24.0%、42.4%，R^2 均值提高了 0.286。西北地区 RF 和 GBM 算法与实测站点 3 d 平均误差分别比多项式回归拟合法误差分别降低了 33.7% 和 32.7%。

2）HRSMS 模型将时空不连续的植被指数和地表温度数据进行重构，相比采用同样算法但输入数据不连续的研究，生成的土壤水分结果更稳定。

3）RF 和 GBM 算法性能接近，在吉林省开展相关研究时，可根据研究需求应用 RF 和 GBM 算法。

[参　考　文　献]

[1] ZHANG D J, ZHOU G Q. Estimation of soil moisture from optical and thermal remote sensing: A review[J]. Sensors, 2016, 16(8): 1308.

[2] 程谅，焦雄，邸涵悦，等. 不同整地措施坡面土壤水分时空分布特征[J]. 土壤学报，2021，58(6)：1423-1435.
CHEN Liang, JIAO Xiong, DI Hanyue, et al. Spatio-temporal distribution of soil moisture on slopes relative to land preparation measure[J]. Acta Pedologica Sinica, 2021, 58(6): 1423-1435. (in Chinese with English abstract)

[3] SOUZA A G, RIBEIRO N A, SOUZA L L. Soil moisture-based index for agricultural drought assessment: SMADI application in Pernambuco State-Brazil[J]. Remote Sensing of Environment, 2021, 252: 112124.

[4] ZHOU K K, LI J Z, ZHANG T, et al. The use of combined soil moisture data to characterize agricultural drought conditions and the relationship among different drought types in China[J]. Agricultural Water Management, 2021, 243: 106479.

[5] 宋廷强，鲁雪丽，卢梦瑶，等. 基于作物缺水指数的农业干旱监测模型构建[J]. 农业工程学报，2021，37(24)：65-72.
SONG Tingqiang, LU Xueli, LU Mengyao, et al. Construction of agricultural drought monitoring model based on crop water stress index[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2021,

37(24): 65-72. (in Chinese with English abstract)
[6] NJOKU E G, JACKSON T J, LAKSHMI V, et al. Soil moisture retrieval from AMSR-E[J]. IEEE Transactions on Geoscience and Remote Sensing, 2003, 41(2): 215-229.
[7] PARINUSSA R M, HOLMES T R, WANDERS N, et al. A preliminary study toward consistent soil moisture from AMSR2[J]. Journal of Hydrometeorology, 2015, 16(2): 932-947.
[8] 贺玉洁，张智韬，巴亚岚，等. 基于 Sentinel-2 卫星数据的灌区农田土壤水盐协同反演[J]. 农业工程学报，2023，39(19)：111-121.
HE Yujie, ZHANG Zhitao, BA Yalan, et al. Synergistic inversion of water and salt in irrigated agricultural soils based on Sentinel-2[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2023, 39(19): 111-121. (in Chinese with English abstract)
[9] 陈芳芳，宋姿睿，张景涵，等. 融合多尺度特征的冬小麦空间分布提取方法[J]. 农业工程学报，2022，38(24)：268-274.
CHEN Fangfang, SONG Zirui, ZHANG Jinghan, et al. Extraction method for the spatial distribution of winter wheat using multi-scale features[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2022, 38(24): 268-274. (in Chinese with English abstract)
[10] KERR Y H, Al-Yaari A, Rodriguez-Fernandez N, et al. Overview of SMOS performance in terms of global soil moisture monitoring after six years in operation[J]. Remote Sensing of Environment, 2016, 180: 40-63.
[11] 姚晓磊，鱼京善，孙文超. 基于累积分布函数匹配的多源遥感土壤水分数据连续融合算法[J]. 农业工程学报，2019，35(1)：131-137.
YAO Xiaolei, YU Jingshan, SUN Wenchao. Continuous fusion algorithm analysis for multi-source remote sensing soil moisture data based on cumulative distribution fusion[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2019, 35(1): 131-137. (in Chinese with English abstract)
[12] ENTEKHABI D, NJOKU E G, O'Neill P E, et al. The soil moisture active passive (SMAP) mission[J]. Proceedings of the IEEE, 2010, 98(5): 704-716.
[13] DORIGO W, WAGNER W, ALBERGEL C, et al. ESA CCI soil moisture for improved earth system understanding: State-of-the art and future directions[J]. Remote Sensing of Environment, 2017, 203: 185-215.
[14] 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.
[15] 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 and Remote Sensing, 2002, 40(12): 2659-2673.
[16] DAS N N, ENTEKHABI D, NJOKU E G. An algorithm for merging SMAP radiometer and radar data for high-resolution soil-moisture retrieval[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 49(5): 1504-1512.
[17] 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.
[18] KIM J, HOGUE T S. Improving spatial soil moisture representation through integration of AMSR-E and MODIS Products[J]. IEEE Transactions on Geoscience and Remote Sensing, 2012, 50(2): 446-460.
[19] MOLERO B, MERLIN O, MALBETEAU Y, et al. SMOS disaggregated soil moisture product at 1 km resolution: Processor overview and first validation results[J]. Remote Sensing of Environment, 2016, 180: 361-376.
[20] KAHEIL Y H, GILL M K, MCKEE M, et al. Downscaling and assimilation of surface soil moisture using ground truth measurements[J]. IEEE Transactions on Geoscience and Remote Sensing, 2008, 46(5): 1375-1384.
[21] KIM G, BARROS A P. Downscaling of remotely sensed soil moisture with a modified fractal interpolation method using contraction mapping and ancillary data[J]. Remote Sensing of Environment, 2002, 83(3): 400-413.
[22] 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.
[23] INES V M, MOHANTY B P, SHIN Y. An unmixing algorithm for remotely sensed soil moisture[J]. Water Resources Research, 2013, 49(1): 408-425.
[24] REICHLE R H, ENTEKHABI D, MCLAUGHLIN D B. Downscaling of radio brightness measurements for soil moisture estimation: A four-dimensional variational data assimilation approach[J]. Water Resources Research, 2001, 37(9): 2353-2364.
[25] 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.
[26] ABOWARDA A S, BAI L L, ZHANG C J, et al. Generating surface soil moisture at 30m spatial resolution using both data fusion and machine learning toward better water resources management at the field scale[J]. Remote Sensing of Environment, 2021, 255: 112301.
[27] ZHANG D W, LU L J, LI X, et al. Spatial downscaling of ESA CCI soil moisture data based on deep learning with an attention mechanism[J]. Remote Sensing, 2024, 16(8): 1394.
[28] HUANG S Z, ZHANG X, CHEN N C, et al. Generating high-accuracy and cloud-free surface soil moisture at 1 km resolution by point-surface data fusion over the Southwestern U. S[J]. Agricultural and Forest Meteorology, 2022, 321: 108985.
[29] DJAMAI N, MAGAGI R, GOITA K, et al. A combination of DISPATCH downscaling algorithm with CLASS land surface scheme for soil moisture estimation at fine scale during cloudy days[J]. Remote Sensing of Environment, 2016, 184: 1-14.
[30] ZHAO W, LI A N, JIN H A, et al. Performance evaluation of the triangle-based empirical soil moisture relationship models based on Landsat-5 TM data and in situ measurements [J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(5): 2632-2645.
[31] LI S, XU L, JING Y H, et al. High-quality vegetation index product generation: A review of NDVI time series

reconstruction techniques[J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 105: 102640.

[32] BIAN J H, LI A N, SONG M Q, et al. Reconstruction of NDVI time-series datasets of MODIS based on Savitzky-Golay filter[J]. Journal of Remote Sensing, 2010, 14(4): 725-741.

[33] CHEN J, JONSSON P, TAMURA M, et al. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter[J]. Remote Sensing of Environment, 2004, 91(3/4): 332-344.

[34] MAO K B, YUAN Z J, ZUO Z Y, et al. Changes in global cloud cover based on remote sensing data from 2003 to 2012[J]. Chinese Geographical Science, 2019, 29(2): 306-315.

[35] ZHAO W, WEN F P, WANG Q M, et al. Seamless downscaling of the ESA CCI soil moisture data at the daily scale with MODIS land products[J]. Journal of Hydrology, 2021, 603: 126930.

[36] 巴亚岚，张智韬，谢坪良，等. 集成 Sentinel-1/2 和环境变量的新疆农田土壤含盐量反演[J]. 农业工程学报，2024，40(16)：171-179.
BA Yalan, ZHANG Zhitao, XIE Pingliang, et al. Inverting soil salinity of farmland in Xinjiang by integrating Sentinel-1/2 and environmental variables[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(16): 171-179. (in Chinese with English abstract)

[37] 谢乾伟，薛丰昌，陈剑飞. 结合虚拟样本生成的油菜花期集成学习预测模型[J]. 农业工程学报，2024，40(19)：159-167.
XIE Qianwei, XUE Fengchang, CHEN Jianfei. Ensemble learning prediction model for rapeseed flowering periods incorporating virtual sample generation[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(19): 159-167. (in Chinese with English abstract)

[38] 罗慧，刘星语，韦骁，等. 基于 THz 成像和集成学习的番茄根长表型提取及预测[J]. 农业工程学报，2024，40(18)：176-183.
LUO Hui, LIU Xingyu, WEI Xiao, et al. Extracting and predicting tomato root length phenotype using THz imaging and ensemble learning[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2024, 40(18): 176-183. (in Chinese with English abstract)

[39] TAHSIN S, MEDEIROS S C, HOOSHYAR M, et al. Optical cloud pixel recovery via machine learning[J]. Remote Sensing, 2017, 9(6): 527.

[40] LI X Q, PENG Q Y, ZHENG Y, et al. Incorporating environmental variables into spatiotemporal fusion model to reconstruct high-quality vegetation index data[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-12.

[41] LI Y L, ZHU S Y, LUO Y M, et al. Reconstruction of land surface temperature derived from FY-4A AGRI data based on two-point machine learning Method[J]. Remote Sensing, 2023, 15(21): 5179.

[42] YAO R, WANG L C, HUANG X, et al. Global seamless and high-resolution temperature dataset (GSHTD), 2001-2020[J]. Remote Sensing of Environment, 2023, 286: 113422.

[43] SENANAYAKE I P, YEO I Y, WALKER J P, et al. Estimating catchment scale soil moisture at a high spatial resolution: Integrating remote sensing and machine learning[J]. Science of the Total Environment, 2021, 776: 145924.

[44] 张云，张丹丹，孟婉婷，等. CYGNSS/SMAP 数据融合半经验模型的土壤湿度反演研究[J]. 北京航空航天大学学报，2022，49(11)：1-15.
ZHANG Yun, ZHANG Dandan, MENG Wanting, et al. Soil moisture retrieval using CYGNSS/SMAP data fusion semi-empirical model[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 49(11): 1-15. (in Chinese with English abstract)

[45] 于雷，朱亚星，洪永胜，等. 高光谱技术结合 CARS 算法预测土壤水分含量[J]. 农业工程学报，2016，32(22)：138-145.
YU Lei, ZHU Yaxing, HONG Yongsheng, et al. Determination of soil moisture content by hyperspectral technology with CARS algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(22): 138-145. (in Chinese with English abstract)

[46] BREIMAN L. Random forests[J]. Journal of Clinical Microbiology, 2001, 2: 199-228.

[47] FRIEDMAN J H. Greedy function approximation: A gradient boosting machine [J]. Annals of Statistics, 2001: 1189-1232.

[48] 万伦军. 基于梯度提升模型的负相关学习算法的研究与应用[D]. 合肥：中国科学技术大学，2014.
WAN Lunjun. Research and Application of Negative Correlation Learning Algorithm Based on Gradient Lifting Model. [D]. Hefei: University of Science and Technology of China, 2014. (in Chinese with English abstract)

[49] BELETE D M, HUCHAUAH M D. Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results[J]. International Journal of Computers and Applications, 2022, 44(9): 875-886.

[50] 任宗悦，刘晓静，刘家福，等. 近 60 年东北地区春玉米旱涝趋势演变研究[J]. 中国生态农业学报（中英文），2020，28(2)：179-190.
REN Zongyue, LIU Xiaojing, LIU Jiafu, et al. Evolution of drought and flood trend in the growth period of spring maize in Northeast China in the past 60 years[J]. Chinese Journal of Eco-Agriculture, 2020, 28(2): 179-190. (in Chinese with English abstract)

[51] 王蕊，张继权，郭恩亮，等. 近 55 a 吉林中西部玉米生长季旱涝时空特征分析[J]. 自然灾害学报，2018，27(1)：186-197.
WANG Rui, ZHANG Jiquan, GUO Enliang, et al. Spatiotemporal characteristics of drought and waterlogging during maize growing season in midwestern Jilin province for recent 55 years[J]. Journal of Natural Disasters, 2018, 27(1): 186-197. (in Chinese with English abstract)

[52] 宋每慧，辛景峰，黄诗峰，等. 基于地理加权回归的吉林省微波土壤水分降尺度研究[J]. 水电能源科学，2024，42(2)：23-29.
SONG Meihui, XIN Jingfeng, HUANG Shifeng, et al. Microwave soil moisture downscaling study of Jilin Province based on geographically weighted regression[J]. Water Resources and Power, 2024, 42(2): 23-29. (in Chinese with English abstract)

[53] ABBASZADEH P, MORADKHANI H, ZHAN X W. Downscaling SMAP radiometer soil moisture over the CONUS using an ensemble learning method[J]. Water Resources Research, 2019, 55(1): 324-344.

[54] WEI Zushuai, MENG Yizhuo, ZHANG Wen, 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.

[55] KARTHIKEYAN L, MISHRA A K. Multi-layer high-resolution soil moisture estimation using machine learning over the United States[J]. Remote Sensing of Environment, 2021, 266: 112706.

Construction and application of a high-resolution soil moisture simulation model integrating multi-source data

FU Pingfan[1,2] , YANG Xiaojing[1,2※] , JIANG Bo[3] , SU Zhicheng[1,2] , SUN Dongya[1,2]

(1. *China Institute of Water Resources and Hydropower Research*, *Beijing* 100038, *China*; 2. *Research Center of Flood and Drought Disaster Reduction of the Ministry of Water Resources*, *Beijing* 100038, *China*; 3. *Soil Moisture Monitoring Center of Jilin Province*, *Changchun* 130033, *China*)

Abstract: Soil moisture is one of the most critical hydrologic indicators in the land-atmosphere heat exchange and global climate dynamics. The high-resolution products of soil moisture are greatly contributed to the precise monitoring of agricultural droughts. However, the existing datasets of soil moisture are limited to the coarse spatial resolution (typically >9 km) and temporal discontinuity. In this study, a high-resolution soil moisture simulation (hrsms) framework was developed to incorporate an ensemble learning approach, particularly for multisource data fusion. Spatially continuous estimates of soil moisture were then captured at 1 km resolution with temporal consistency. The accuracy of estimation was improved significantly, compared with the conventional approaches. Three computational procedures are included in the framework. Firstly, the high-resolution ancillary datasets (e.g., vegetation indices and land surface temperature) were spatiotemporally reconstructed using Savitzky-Golay filtering with multivariate regression. Data gaps were also determined to preserve the temporal dynamics. Secondly, the spatial downscaling was performed on the soil moisture active passive (smap) observations (2017-2022, 0-5 cm depth) from 9 km to 1 km resolution. A systematic investigation was also made to clarify the synergistic relationships among vegetation indices, land surface temperature, soil properties, and topographic parameters. In situ measurements were then implemented using ensemble machine learning, including random forest (rf) and gradient boosting machine (gbm). Thirdly, the multi-scale assessments were selected to compare with the original moderate resolution imaging spectroradiometer land surface temperature (modis lst) products. The point-scale evaluation of in-situ networks was also carried out in Jilin Province, China. A systematic quantification was then performed on the computational efficiency and accuracy metrics, including the root mean square error (rmse), mean absolute error (mae), and coefficient of determination (R^2). Finally, the polynomial regression fitting (prf) was utilized to validate the hrsms model on three critical maize growth days (16 June 2018, 15 August 2019, and 11 July 2020). The results showed that: 1) The high performance was achieved in reconstructing the land surface temperature, with the rmse, mae, and R^2 values of 0.526 K, 0.338 K, and 0.986, respectively, compared with the original modis lst. Three sites were randomly selected to evaluate the performance of the hrsms model in both temporal and spatial dimensions. The gbm algorithm marginally outperformed the rf. 2) The rf algorithm was achieved in the mae, rmse, and R2 values of 0.033 m^3/m^3, 0.049 m^3/m^3, and 0.574, respectively, over three days. The gbm algorithm also yielded comparable metrics (MAE: 0.033 m^3/m^3; RMSE: 0.050 m^3/m^3; and R^2: 0.556). 3) The hrsms model significantly improved the accuracy of soil moisture simulation, compared with the prf. The improved model was realized to solve the prf overestimation of soil moisture in northwest Jilin Province. 4) The rf and gbm demonstrated similar efficacy, with the rf marginally outperforming gbm. As such, both improved models were equivalently deployed to implement the regional-scale simulation with operational flexibility. The hrsms framework successfully enhanced the spatial resolution and accuracy of soil moisture products, particularly with the temporal continuity. Multisource data and ensemble learning were integrated to solve the overestimation in the traditional models, suitable for the agriculturally vital regions. The operational adaptability of rf and gbm algorithms can be expected to tailor the applications to diverse data environments. The improved model also shared the significant potential for regional scalability, particularly in the necessitating areas for the high-resolution monitoring of soil moisture. The robustness and generalizability can be enhanced to validate the diverse geographical regions and climatic conditions. The complementary environmental variables (e.g., evapotranspiration) can also be integrated into future research. The findings can substantially contribute to the precision agriculture practices and climate resilience.

Keywords: soil moisture; random forest; gradient boosting machine; SMAP SSM; downscaling; point-surface data fusion