光学 精密工程
OPTICS AND PRECISION ENGINEERING
核心期刊
主管：中国科学院
主办：中国科学院长春光学精密机械与物理研究所
中国仪器仪表学会
主编：曹健林

极限梯度提升和长短期记忆网络相融合的土壤温度预测

李清亮，蔡凯旋，耿庆田，刘光洁，孙明玉，张崙，于繁华

引用本文:
李清亮, 蔡凯旋, 耿庆田, 等. 极限梯度提升和长短期记忆网络相融合的土壤温度预测[J]. 光学 精密工程, 2020, 28(10): 2337-2348.
LI Qing-liang, CAI Kai-xuan, GENG Qing-tian, et al. Estimation of soil temperature based on XGBoost and LSTM methods[J]. *Optics and Precision Engineering*, 2020, 28(10): 2337-2348.

在线阅读 View online: https://doi.org/10.37188/OPE.20202810.2337

文章编号　1004-924X(2020)10-2337-12

极限梯度提升和长短期记忆网络相融合的土壤温度预测

李清亮*，蔡凯旋，耿庆田，刘光洁，孙明玉，张　崙，于繁华*
（长春师范大学 计算机科学技术学院，吉林 长春 130032）

摘要：土壤温度是地球科学多个领域的重要变量。其时空变化受多种环境因素影响，这对土壤温度的准确预测带来巨大挑战。以机器学习为核心的数据驱动方法，在土壤温度预测中是重要研究领域，为基于物理过程模型提供重要补充。然而目前针对土壤温度影响因素量性研究较少，因此本文提出 XGBoost-LSTM 的数据驱动方法。基于极限梯度提升算法(XGBoost)分析土壤温度影响因素的重要性，然后根据影响因素重要性依次组合，并输入至长短期记忆网络(LSTM)，得到最优预测模型并实现土壤温度预测。最后在长白山和海北两个气象站完成实验，本文方法的最优均方根误差为 2.234、平均绝对误差为 1.716、纳什效率系数为 0.932、LMI 系数为 0.729 和威尔莫特一致性指数为 0.983。结果表明本文提出的 XGBoost-LSTM 预测模型与目前土壤温度中常用的数据驱动模型相比，均表现出更高的精确度。

关　键　词：土壤温度预测；长短期记忆网络；极限梯度提升；特征重要性；数据驱动方法
中图分类号：P401；P343.8　　**文献标识码**：A　　**doi**：10.37188/OPE.20202810.2337

Estimation of soil temperature based on XGBoost and LSTM methods

LI Qing-liang*, CAI Kai-xuan, GENG Qing-tian, LIU Guang-jie,
SUN Ming-yu, ZHANG Yu, YU Fan-hua*

(*College of Computer Science and Technology, Changchun Normal University, Changchun* 130032, *China*)
* *Corresponding author, E-mail: liqingliang@ccsfu.edu.cn; yufanhua@*163.*com*

Abstract: Soil temperature is an important variable in Earth sciences. The temporal and spatial variations in soil temperature are affected by numerous factors, resulting in various challenges in soil temperature prediction. For soil temperature prediction, the data-driven machine learning method is valuable and can be an important complement to physics-based process models. However, no extensive studies have been carried out on the importance of environmental factors on soil temperature. In this study, a data-driven XGBoost-LSTM method is proposed. The weights of the meteorological inputs

收稿日期：2020-03-16；**修订日期**：2020-05-24.
基金项目：国家自然科学基金资助项目(No. 61604019)；吉林省省级科技创新项目资助(No. 20190302026GX)；吉林省发改委产业技术研究与开发项目资助(No. 2020C019-3；No. 2019C054-8)；吉林省科技发展计划资助项目(No. 20180201086SF)；吉林省高等教育学会高教科研项目资助(No. JGJX2018D10)；辽宁省科技厅联合开放基金机器人学国家重点实验室开放基金资助项目(No. 2020-KF-22-08)；长春师范大学青年教师培育计划项目(长师大自科合字〔2019〕第 006 号)

are computed based on XGBoost, and then, the combination of meteorological inputs based on their weights is applied to obtain an optimal model by the LSTM method. An experiment is carried out at two stations in China (Changbai Mountain and Haibei). The most accurate performance for soil temperature estimation is attained, with highest values of NS = 0.932, WI = 0.983, and LMI = 0.729 and lowest values of RMSE and MAE of 2.234 and 1.716, respectively. These results show that the proposed model is generally superior to other state-of-the-art predictive models.

Key words: soil temperature estimation; long short-term memory; extreme gradient; data-driven method; weight of feature

1 引　言

土壤温度作为大气和陆地表面水热循环共同作用的结果，是地球科学多个领域的重要变量，如气象学[1-2]，农业[3-4]和环境科学[5]等。目前土壤温度的测量过程过于复杂[6-7]，无法在大部分区域提供监测点。因此，准确地土壤温度预测具有显著应用价值。然而土壤温度的时空变化受多种环境因素影响[8]，给土壤温度的准确预测带来巨大挑战，同时也使其成为具有深远意义的科学问题。目前基于机器学习的数据驱动经验模型是土壤温度预测一个重要的研究领域[2]。

近年来，随着机器学习技术迅速发展，已在多种领域成功应用[9-11]。其中基于机器学习的数据驱动模型在土壤温度预测中也取得了一定成果，最常见的有：人工智能神经网络(Artificial Neural Network, ANN)[12-14]、支持向量回归(Support Vector Regression, SVR)[15-16]、极限学习机(Extreme Learning Machine, ELM)[1-2]。ANN 由于可以模拟人脑组织结构进行推理的特点，已被广泛应用于多种预测领域。在土壤温度预测中，Mihalakakou 等人[12]基于 ANN 模型取得了较好的预测性能。Kisi 等人[13]指出土壤浅层预测时径向基神经网络(Radial Basis Neural Network, RBNN)相比广义回归神经网络(Generalized Regression Neural Network, GRNN)、多层感知机(Multilayer Perceptron, MLP)和线性回归(Linear Regression, LR)预测模型展示了更优的预测性能。支持向量回归(SVR)预测模型是由支持向量分类器(Support Vector Machines, SVM)发展而来，在土壤温度预测中发挥了重要作用。Delbari 等人[16]在伊朗地区的测试中，证明了 SVR 预测模型可以成功应用在土壤温度的预测。Feng 等人[1]在中国陕西省的测试中，通过比较 GRNN、反向传播神经网络(Back Propagation Neural Network, BPNN)、随机森林(Random Forests, RF) 和 ELM 的预测模型，发现 ELM 预测模型的预测能力最高。

然而，在前人研究主要基于 ANN，SVR 和 ELM 等方法实现土壤温度预测。缺少其他机器学习算法如极限梯度提升(XGBoost)和长短期记忆网络(Long Short-Term Memory, LSTM)等研究。XGBoost 是基于梯度下降方法，将多个弱分类器融合成一种强分类器，这种加强学习能力的模式，使 XGBoost 在多种领域的预测研究得到广泛应用[17-19]。同时，具有记忆细胞的 LSTM 模型既可以学习其短期行为，也可以学习其长期行为。这种模式使 LSTM 在模拟自然系统时非常有用[20-22]。同时土壤温度的影响因素很多，如太阳辐射、大气温度等，同时基于机器学习针对影响因素重要性的量化评价并结合预测模型的方式并未展开研究。事实上预测模型中不同影响因素的组合对预测能力有很大影响。

本文将在中国长白山和海北区域进行测试，首先基于 XGBoost 计算土壤温度预测中每个影响因素的重要性，然后将影响因素依据重要性大小依次进行组合并作为预测模型的输入，其中本文尝试基于 LSTM 作为预测模型。最后根据不同组合的预测性能选择最优的影响因素。

2 数据与方法

2.1 研究区域概况

本文在中国东北部吉林省长白山地区(41°41′N，127°42′E 至 42°51′N，128°16′E)和中国西北部青海省海北地区(31°36′N，86°35′E 至 39°19′N，103°104′E)进行了测试。在 FLUXNET(ht-

tps://FLUXNET.fluxdata.org/)上下载了半小时尺度的三年数据(2003 年 1 月 1 日至 2005 年 12 月 31 日)，其影响因素分别为大气温度、相对湿度、太阳辐射、蒸汽压、风速、降水量和风速，其中土壤温度是在土壤深度为 5 cm 时观测得到。长白山和海北都是我国重要的生态自然保护区。长白山地区以是种植业为主的农业地域类型，发展商品谷物农业，主要种植大豆、玉米、冬小麦及谷子(小米)等。海北属高原大陆性气候，太阳辐射强度大，日照时间长，日夜温差大，有利于油菜、青稞、马铃薯及蔬菜等种植业的发展。并且土壤温度对这些资源的生长有很大影响，因此，土壤温度预测对这些生态资源的保护和管理具有重要意义。

2.2　土壤温度预测框架

本研究主要考虑两个问题：(1)对于土壤温度预测时，不同影响因素起到不同作用，这些影响因素或其组合与土壤温度间的关系如何？(2)土壤温度预测是一种时间序列预测问题，哪一种机器学习算法最符合土壤温度时间序列的机制？为了解决上述问题，本文提出 XGBoost-LSTM 融合模型，如图 1 所示。对于多个气象数据来说，不同因素与土壤温度之间的相关性不同。首先基于 XGBoost 对各个影响因素进行排序。对排序后的影响因素，本文按其重要性依次进行组合作为 LSTM 模型的输入，找到效果最佳的输入组合，进而基于 LSTM 模型挖掘土壤温度与其影响因素间的关联，并实现预测。

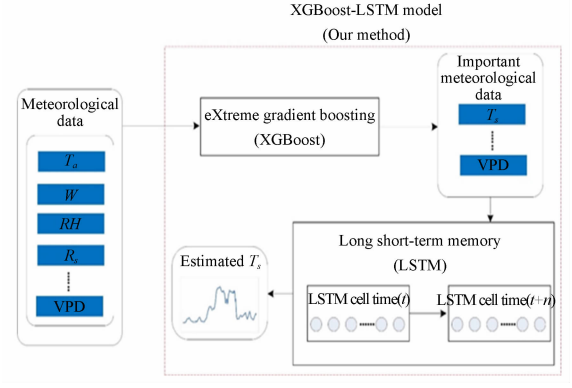

图 1　XGBoost-LSTM 模型的土壤温度预测流程图

Fig. 1　Framework of XGBoost-LSTM

2.3　XGBoost 模型

XGBoost 是一种新的梯度提升机器 Boosting 算法。Boosting 分类器作为集成学习模型，具有将弱分类器转化为强分类器的能力。其核心是将多个准确率较低的决策树模型组合成一个准确率较高的模型，通过不断迭代生成决策树，拟合残差优化目标函数，从而优化预测效果。当 XGBoost 模型完成训练，本文针对影响因素在该模型的贡献程度计算其重要性。具体方法如下：先计算特征 j 在单棵树中的重要程度，即找到与特征 j 有关的结点，计算该结点分裂前后的平方损失值，损失值越大说明该特征越重要。单个树中特征 j 的计算方法如下：

$$\hat{J}_j^2(L) = \sum_{t=1}^{T-1} i_t^2(v_t = j). \tag{1}$$

由于该算法构建的都为二叉树，T 为树的叶子结点数量，则 $T-1$ 即为树的非叶子结点数量，v_t 是和结点 t 相关联的特征，i_t^2 是结点 t 分裂之后平方损失减少值。特征 j 在整个模型中的重要程度通过对所有含特征 j 的决策树的重要程度求和并取平均值来衡量，公式如下：

$$\hat{J}_j^2(L) = \frac{1}{K}\sum_{k=1}^{K}(\hat{J}_j^2 L_K). \tag{2}$$

通过式(1)和式(2)量化每个特征的重要程度后，即可实现特征排序。

2.4　LSTM 模型

LSTM 是一种重要的递归神经网络模型，具有记忆细胞的 LSTM 既可以学习土壤温度与其环境因子之间的短期行为，也可以学习其长期行为。

图 2 说明了 LSTM 的结构，其工作流程如下式所述：

$$i(t) = \sigma(W_{ih}h(t-1) + W_{ix}x(t) + W_{ic}c(t-1) + b_i), \tag{3}$$

$$f(t) = \sigma(W_{fh}h(t-1) + W_{fx}x(t) + W_{fc}c(t-1) + b_f), \tag{4}$$

$$c(t) = f(t) \otimes c(t-1) + i(t) \otimes \tanh(W_{ch}h(t-1) + W_{cx}x(t) + b_c), \tag{5}$$

$$o(t) = \sigma(W_{oh}h(t-1) + W_{ox}x(t) + W_{oc}c(t) + b_o), \tag{6}$$

$$h(t) = o(t) \otimes \tanh(c(t)), \tag{7}$$

$$\hat{y}(t) = W_{yh}h(t) + b_o, \tag{8}$$

其中：$x(t)$和 $\hat{y}(t)$分别是 t 时刻的 LSTM 输入和输出，$i(t)$和 $f(t)$分别是 t 时刻的输入门和遗忘门，W 和 b 是模型参数，$c(t)$是 t 时刻的存储单元状态，它由遗忘门和输入门改变。$o(t)$是 t 时刻的输出门，它控制从状态到模型输出的信息流。$h(t)$是 t 时刻的递归输入；$\otimes$表示两个向量的元素相乘；$\sigma(\cdot)$表

示 sigmoid 激活函数；tanh(·)是双曲正切函数。

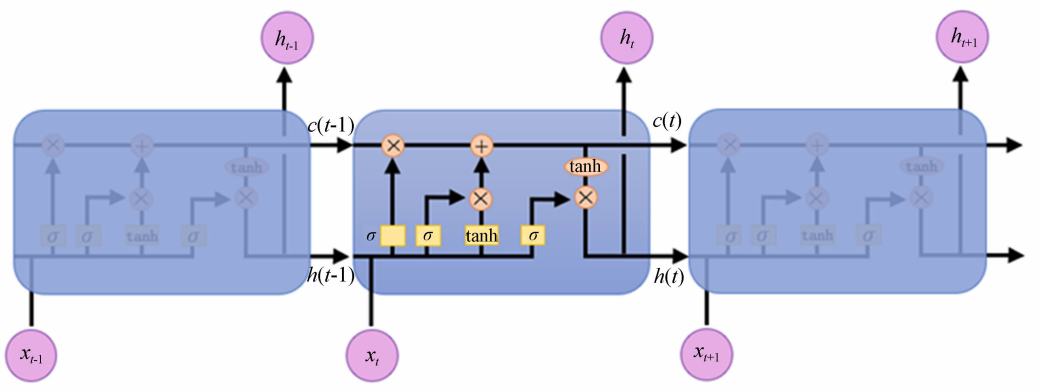

图 2　LSTM 流程图

Fig. 2　Structure of LSTM

2.5　实验设置

本文在 Intel Core(TM)i7-5820K，3.30 GHz CPU 和 64 Gb 内存的服务器上进行实验。在 Pycharm 2018.2.8，tensorflow 作为工具实现所有预测模型。整个数据集的前四分之三部分(长白山和海北区域，2003 年 1 月 1 日 00:00 至 2005 年 12 月 31 日 23:30 阶段的数据，共计 52 608 个)用于训练(2003 年 1 月 1 日 00:00 至 2005 年 4 月 1 日 12:30 阶段的数据，共计 39 456 个)，其余四分之一部分用于测试(2005 年 4 月 1 日 12:30至 2005 年 12 月 31 日 23:30 阶段的数据，总共 13 152 个)。为了验证本文方法的有效性，通过与五种数据驱动预测模型(LR，SVR，XGBoost，BPNN，ELM)作对比，并采用不同的统计评价标准评估模型性能，如均方根误差(RMSE)、平均绝对误差(MAE)、纳什效率系数(NS)、LMI 系数和威尔莫特一致性指数(WI)，定义如下：

$$RMSE = \sqrt{\frac{\sum_{n=1}^{N}(y_i-\hat{y}_i)}{N}}, \tag{9}$$

$$MAE = \frac{\sum_{n=1}^{N}|y_i-\hat{y}_i|}{N}, \tag{10}$$

$$NS = 1-\frac{\sum_{n=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{n=1}^{N}(y_i-\bar{y}_i)^2}, \tag{11}$$

$$WI = 1-\frac{\sum_{n=1}^{N}(y_i-\hat{y}_i)^2}{\sum_{n=1}^{N}(|\hat{y}_i-\bar{y}|+|y_i-\bar{y}|)^2}, \tag{12}$$

$$LMI = 1-\frac{\sum_{n=1}^{N}|y_i-\hat{y}_i|}{\sum_{n=1}^{N}|y_i-\bar{y}_i|}, \tag{13}$$

其中：N 表示测试数据总数，$\hat{y}_i$ 和 y_i 分别为土壤温度的预测值和观测值。$\bar{y}_i$ 是土壤温度观测值的平均值。其中 RMSE，MAE 的值越低，预测模型的性能越好。相反，NS，WI 和 LMI 值越大，预测模型的性能越好。

3　分析与讨论

本章节主要对比了 LR，SVR，XGBoost，BPNN，ELM 5 种预测模型，并基于 scikit-learn 工具实现所有方法。XGBoost 预测模型模型中，通过 scikit-learn 工具中提供的 selection.GridSearchCV 函数调整参数，并优化 XGBoost 模型。其以 CART 回归树作为 XGBoost 模型模型的二叉树进行构建、学习率为 0.005，叶子上的最小样本数为 1，树的个数为 800，最大深度为 5。BPNN 预测模型模型中，定义三层结构(输入层、隐藏层、输出层)，以平方误差作为损失函数，采用 Adam 进行优化，批处理数据量为 200，迭代次数为 300，学习率为 1×10^{-5}。ELM 预测模型模型中，利用 hpelm 工具提供的 elm 函数对 ELM 进行建模，初始化三层神经元网络(输入层、隐藏层、输出层)，定义隐藏层的 sigmoid 激活函数，设置节点数为 128。通过实验对比，发现上述设置预测结果最好。

3.1　XGBoost-LSTM 模型中 LSTM 结构超参数的测试

本文方法中，影响预测性能的主要超参数有：LSTM 神经单元个数、批量数据大小、迭代次数、

学习率、影响因素的选取。因此，下面本文将通过预测性能定义最优的超参数。本文方法中，影响预测性能的主要超参数有：LSTM 神经单元个数、批量数据大小、迭代次数、学习率、影响因素的选取。

在测试 LSTM 神经单元个数时，首先定义迭代次数为 200 次，批量数据大小为 200 次，学习率为 1×10^{-4}，选取大气温度、相对湿度和蒸气压。在(30，50，100 和 150)中分别测试 LSTM 神经单元个数；同时选择最优的 LSTM 神经单元个数后，再从(50，100，150，250，300，350，400，450，500)中选择批量数据大小；其次从(100，200，300，400，500，600，800，1 000，1 100，1 200)中选择迭代次数；再从(1×10^{-3}，1×10^{-4}，1×10^{-5})中选择学习率。最后等上述参数均选取最优之后，本文 LSTM 预测模型的最优结构随之确定。进而针对影响因素的选取进行测试，首先基于 XGBoost 算法计算所有影响因素(大气温度、相对湿度、太阳辐射、蒸汽压、风速、降水量和风速)的特征重要性，并依据重要程度从高到低进行组合，依次输入 LSTM 预测模型中，并测试哪些组合对土壤温度的预测更有帮助。其中 LSTM 预测模型输入特征中的输入维度(*input_size*)为影响因素的个数；历史数据序列的长度(*time_step*)定义为 7。本文首先以长白山区域进行实验，不同 LSTM 神经单元个数(*LSTM_size*)、批量数据大小(*batch_size*)、迭代次数(*train_epoch*)和学习率(*e*)的预测结果分别如表 1～表 5 所示(表中加粗数值为最优结果)。从表 1 可知，参数取值不同，预测模型的预测能力也不一样。首先 LSTM 神经单元个数过多，预测模型将过拟合，降低预测模型学习气象数据的泛化能力，同时个数太少，预测模型将遗失重要信息并无法有效挖掘土壤温度与其影响因素间的关联，因此 *LSTM_size* =50 时，预测性能最好(RMSE=2.413，MAE=1.893，NS=0.921，*WI*=0.980，LMI=0.740)；从表 2 可知，若迭代次数太小，预测模型无法收敛于最优值，而迭代次数太大，将明显出现过拟合现象，预测性能将不会提升，因此 *train_epoch* =1 100 时，预测性能最好(RMSE=2.234，MAE=1.756，NS=0.932，WI=0.983，LMI=0.759)；同理，从表 3 可知批量数据主要控制着训练过程一次处理样本的信息，过小将导致预测模型难于

表 1　不同 LSTM 神经单元个数在长白山气象站的预测结果

Tab. 1　Estimation results with different *LSTM _size* at Changbai Mountain meteorological station

LSTM _size	RMSE	MAE	NS	WI	LMI
30	2.493	1.901	0.915	0.979	0.739
50	**2.413**	**1.893**	**0.921**	**0.980**	**0.740**
100	2.558	2.089	0.911	**0.980**	0.714
150	2.785	2.318	0.894	0.976	0.682

表 2　不同批量数据大小在长白山气象站的预测结果

Tab. 2　Estimation results with different *batch _size* at Changbai Mountain meteorological station

batch_size	RMSE	MAE	NS	WI	LMI
50	2.696	2.105	0.901	0.978	0.711
100	2.890	2.381	0.886	0.975	0.674
150	2.686	2.217	0.902	0.978	0.696
250	2.452	1.980	0.918	0.980	0.729
300	2.406	1.918	0.921	**0.981**	0.737
350	**2.385**	1.895	**0.922**	**0.981**	0.740
400	2.392	**1.875**	**0.922**	**0.981**	**0.743**
450	2.408	1.889	0.921	0.980	0.741
500	2.453	1.934	0.918	0.980	0.735

表 3　不同迭代次数在长白山气象站的预测结果

Tab. 3　Estimation results with different *train_epoch* at Changbai Mountain meteorological station

train_epoch	RMSE	MAE	NS	WI	LMI
100	2.750	2.195	0.897	0.975	0.700
200	2.392	1.875	0.922	0.981	0.743
300	2.323	1.850	0.926	0.982	0.746
400	2.306	1.840	0.927	0.982	0.748
500	2.296	1.831	0.928	0.982	0.749
600	2.287	1.822	0.929	0.982	0.750
800	2.238	1.762	**0.932**	**0.983**	0.758
1000	2.235	1.758	**0.932**	**0.983**	**0.759**
1100	**2.234**	**1.756**	**0.932**	**0.983**	**0.759**
1200	2.235	1.759	**0.932**	**0.983**	**0.759**

表 4　不同学习率在长白山气象站的预测结果

Tab. 4　Estimation results with different e at Changbai Mountain meteorological station

e	RMSE	MAE	NS	WI	LMI
1×10^{-3}	2.641	2.176	0.905	0.978	0.702
1×10^{-4}	**2.392**	**1.875**	**0.922**	**0.981**	**0.743**
1×10^{-5}	2.755	2.156	0.897	0.973	0.704

收敛，过大将导致预测模型性能无法达到最优，因此 $batch\_size$ =400 时，预测性能最好(RMSE=2.392，MAE=1.875，NS=0.922，WI=0.981，LMI=0.743)；最后从表 4 可知学习率是控制训练过程中每一次逼近最优值的步长，过小容易选入局部最优，而过大将无法逼近最优值并在最优值附近不断徘徊。因此当 $LSTM\_size$ =50，$batch\_size$ =400，$train\_epoch$ =1 100，$e=1\times10^{-4}$时性能最佳，进而通过确定上述参数，本文得到了最优的 LSTM 预测模型结构。

3.2　XGBoost-LSTM 选取影响因素

得到最优 LSTM 预测模型结构后，然后分析影响因素的组合对预测能力的影响，XGBoost-LSTM 预测模型的整体的数据处理流程如图 3 所示。首先根据 2.3 节 XGBoost 部分计算所有影响因素的重要性，如图 4 所示。同时在 XGBoost 模型计算影响因素重要性的过程中，分别对 XGBoost 模型训练中学习率、树的个数为 800 和树的最大深度所对应的平方损失减少值进行分析，以选取最优参数，对比结果如表 5 所示。当学习率为 0.005，树的个数为 800，最大深度为 5 时，平方损失减少值最优(−0.078 481)。

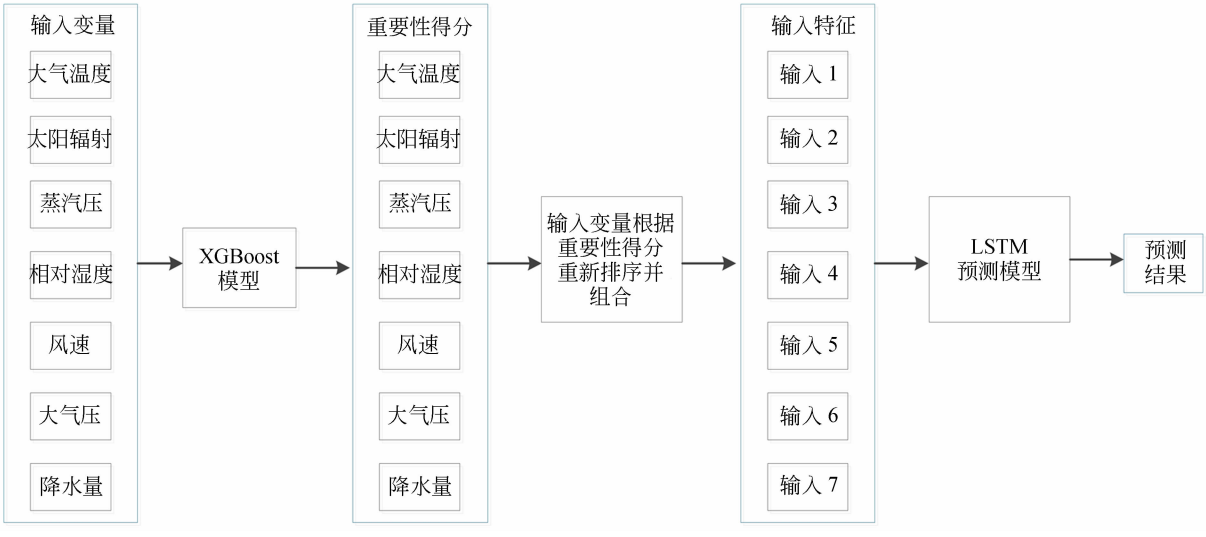

图 3　XGBoost-LSTM 整体流程

Fig. 3　Framework of XGBoost-LSTM model

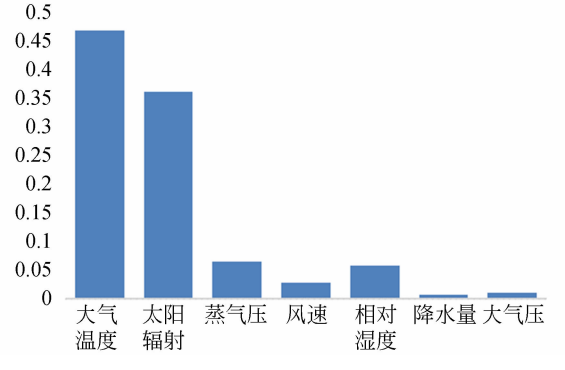

图 4　影响因素重要性得分

Fig. 4　Weight of the meteorological inputs

最终通过实验可知，影响因素重要性由高低排序如下：大气温度、蒸汽压、太阳辐射、相对湿度、风速、大气压和降水量。然后根据影响因素重要性，通过逐步组合生成 7 种预测模型输入：大气温度(输入 1)，大气温度＋太阳辐射(输入 2)，大气温度＋太阳辐射＋蒸汽压(输入 3)，大气温度＋太阳辐射＋蒸汽压＋相对湿度(输入 4)，大气温度＋太阳辐射＋蒸汽压＋相对湿度＋风速(输入 5)，大气温度＋太阳辐射＋蒸汽压＋相对湿度＋风速＋大气压(输入 6)，大气温度＋太阳辐射＋蒸汽压＋相对湿度＋风速＋大气压＋降水量(输入 7)。

表 5　不同参数的 XGBoost 模型性能对比

Tab. 5　XGBoost model performance of different parameters

学习率	树的个数	最大深度	方损失减少值	学习率	树的个数	最大深度	方损失减少值
0.01	4	700	−0.090	0.005	4	700	−0.084
0.01	4	800	−0.087	0.005	4	800	−0.084
0.01	4	900	−0.086	0.005	4	900	−0.086
0.01	5	700	−0.079	0.005	5	700	−0.078
0.01	5	800	−0.078	**0.005**	**5**	**800**	**−0.077**
0.01	5	900	−0.078	0.005	5	900	−0.078
0.01	6	700	−0.080	0.005	6	700	−0.080
0.01	6	800	−0.080	0.005	6	800	−0.081
0.01	6	900	−0.080	0.005	6	900	−0.081

最后基于 LSTM 预测模型对对上述 7 组合进行测试，以找到最优的影响因素组合，如表 6 所示。根据实验可知，以大气温度、太阳辐射和蒸汽压 3 种组合作为输入，可以得到更好的预测结果。

表 6　不同输入组合的预测结果

Tab. 6　Estimation results of the LSTM model by different input combinations

输入	RMSE	MAE	NS	WI	LMI
1	5.224	3.817	0.628	0.891	0.477
2	4.105	3.142	0.770	0.936	0.936
3	**2.234**	**1.756**	**0.932**	**0.983**	**0.759**
4	3.861	2.833	0.7967	0.947	0.612
5	3.969	2.964	0.785	0.942	0.594
6	4.064	3.141	0.785	0.939	0.569
7	4.100	3.137	0.771	0.937	0.570

3.3　不同数据驱动模型的对比

图 5 展示了在长白山区域，本文方法与目前常用的 5 种数据驱动模型的散点图对比。从图 5 可知，本文方法得到的土壤温度预测值（y）和观测值（x）之间的线性关系（$y=0.980\ 6x+0.210\ 7$）更接近理想线（$y=x$），同时相比其它数据驱动模型也得到了最高的 R^2（$R^2=0.934\ 2$）。需要注意的是 LR 模型虽然具有最高 R^2，但其线性关系相较于其他数据驱动模型具有较大差距。因此表明本文方法在长白山区域预测效果最好。图 6 展示了海北区域散点图对比结果。其中，ELM 预测模型预测结果相比本文模型更接近于线性关系，但由于其在长白山的表现一般，表现出其不稳定性。综合两区域散点图的线性关系可知，本文方法效果最佳。

表 7 和表 8 指出了不同数据驱动模型在长白山和海北区域性能数据统计对比。可知，本文方法在长白山站点得到的 RMSE，MAE，RAME 都比其他数据驱动模型的值小，同时 WI，LMI 的值最高。在海北站区域中，ELM 预测模型的 RMSE 值优于本文方法，但 NS，LMI 出现异常结果。其原因可能是 ELM 是一种随机选择隐层权值的神经网络结构，这种随机性使预测模型训练时产生非最优解，从而影响预测结构[1]。综上所述，本文方法在不同站点的土壤温度预测中都优于其他数据驱动模型。

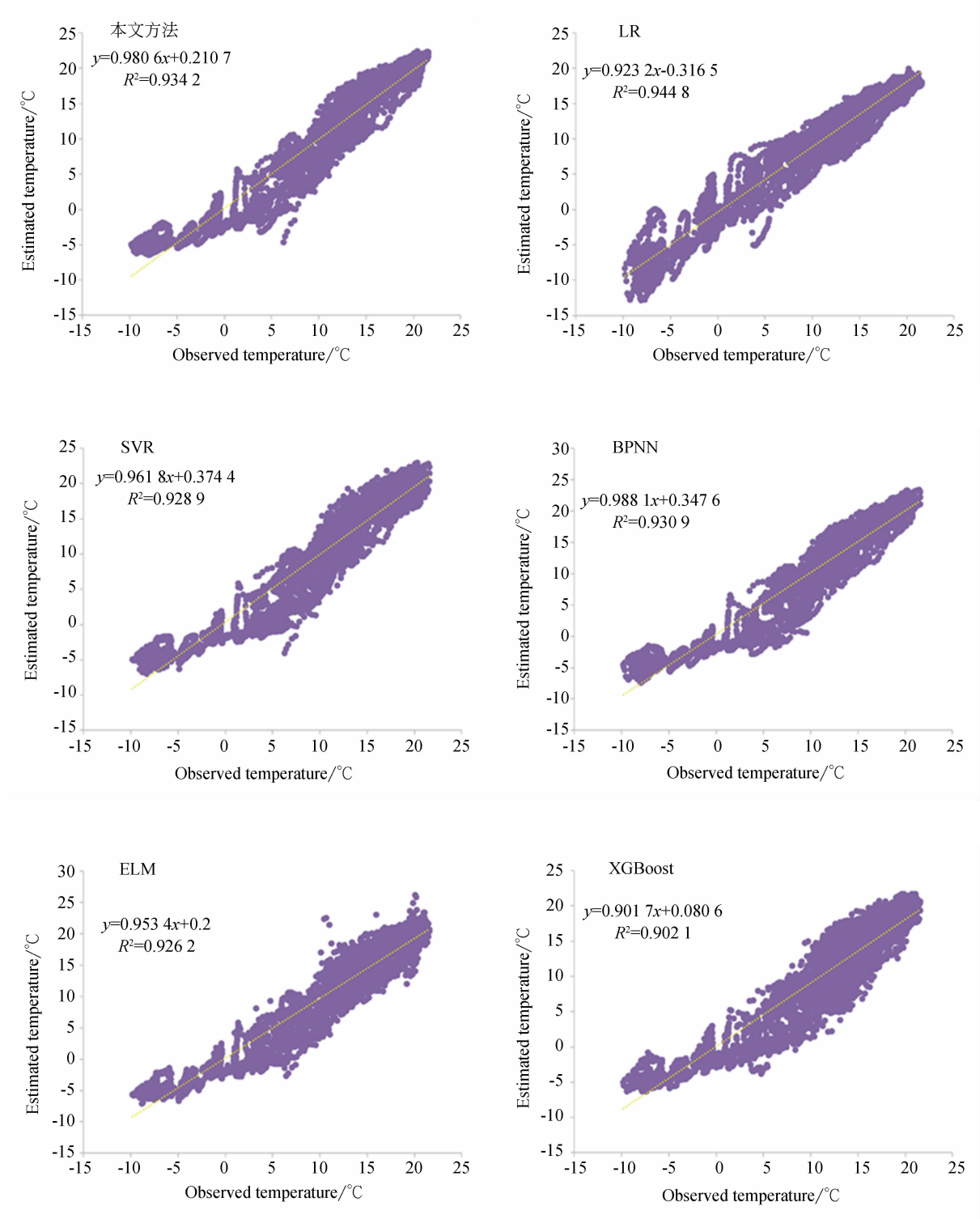

图 5　长白山区域温度预测值与真实值的数据驱动散点图

Fig. 5　Scatterplots of the estimated and observed values of temperature(℃) using the data-driven models for Changbai Mountain metrological station

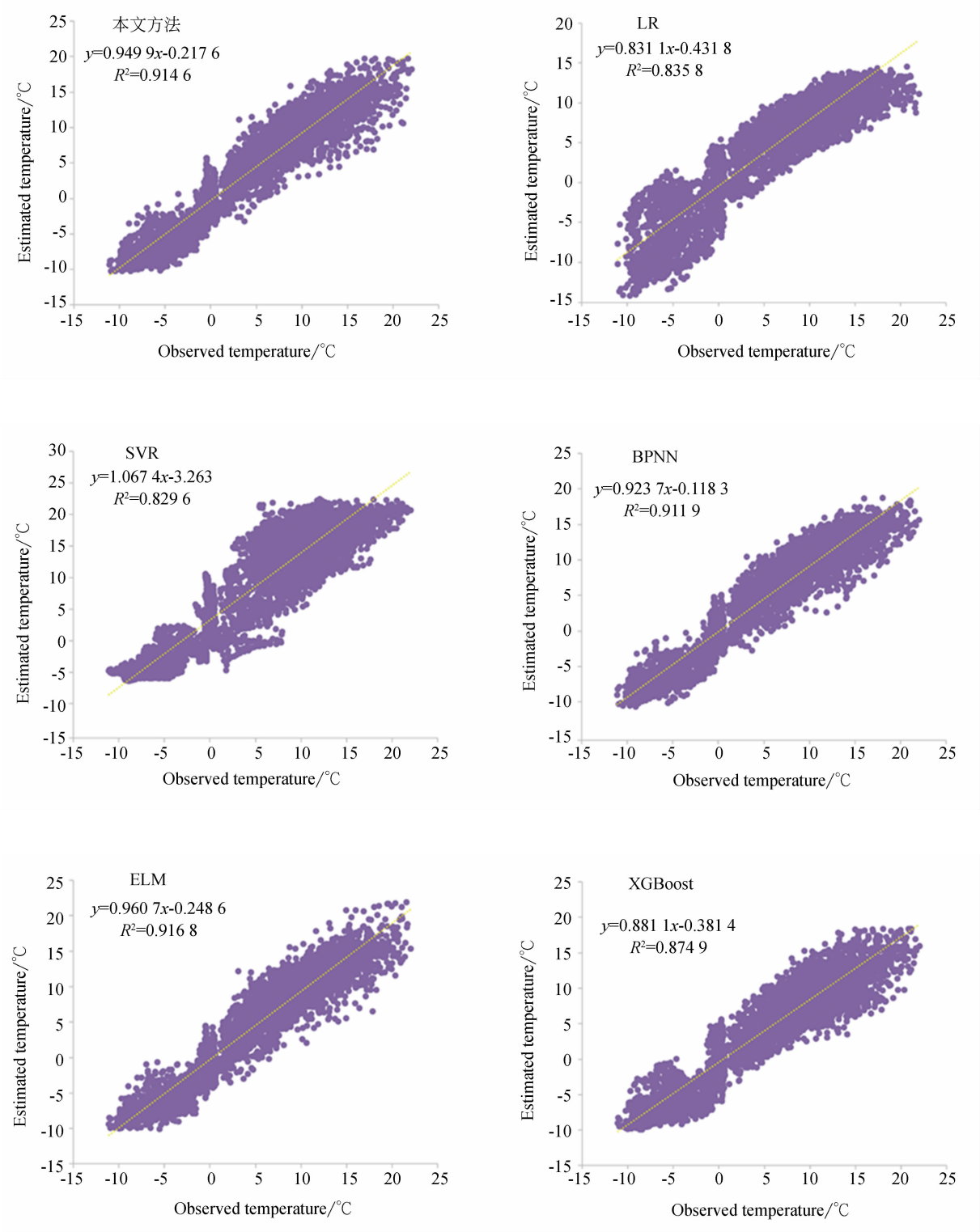

图 6　海北区域温度预测值与真实值的数据驱动散点图

Fig. 6　Scatterplots of the estimated and observed values of temperature (℃) using the data-driven models for Haibei Mountain metrological station

表 7　长白山地区不同数据驱动模型的性能对

Tab. 7　Testing phase results of the different models at Changbai Mountain

方法	RMSE	MAE	NS	WI	LMI
LR	2.266	1.836	−10 081.42	0.407	−6 720.342
SVR	2.303	1.794	0.928	0.982	0.754
XGBoost	2.804	2.213	0.893	0.971	0.697
BPNN	2.320	1.800	0.927	0.982	0.753
ELM	2.350	1.862	−10 427.05	0.429	−6 923.51
本文方法	**2.234**	**1.756**	**0.932**	**0.983**	**0.759**

表 8　海北地区不同数据驱动模型的性能对

Tab. 8　Testing phase results of the different models at Haibei

方法	RMSE	MAE	NS	WI	LMI
LR	3.307	2.598	−9 791.016	0.403	−6 717.998
SVR	3.941	3.027	0.718	0.918	0.522
XGBoost	2.825	2.197	0.855	0.961	0.653
BPNN	2.270	1.754	0.906	0.975	0.723
ELM	**2.216**	1.724	−10 577.625	0.422	−7 016.792
本文方法	2.240	**1.716**	**0.909**	**0.977**	**0.729**

4　结　论

然而，在前人研究主要基于 ANN、SVR 和 ELM 等方法实现土壤温度预测。缺少其它机器学习算法如极限梯度提升(XGBoost)和长短期记忆网络(LSTM)等研究。XGBoost 是基于梯度下降方法，融合多个弱分类器成为一种强分类器，这种加强学习能力的模式，使 XGBoost 在多种领域的预测研究得到广泛应用。同时，具有记忆细胞的 LSTM 模型既可以学习其短期行为，也可以学习其长期行为。这种模式使 LSTM 在模拟自然系统时非常有用。同时土壤温度的影响因素很多，如太阳辐射、大气温度等。但基于机器学习针对影响因素重要性的量化评价，并结合预测模型的方式并未展开研究。事实上，预测模型中不同影响因素的组合对预测能力有很大影响。

本文将在中国长白山和海北区域进行测试，首先基于 XGBoost 计算土壤温度预测中每个影响因素的重要性，然后将影响因素依据重要性大小依次进行组合并作为预测模型的输入，其中本文将尝试基于 LSTM 作为预测模型。最后根据不同组合的预测性能选择最优的影响因素。通过实验结果表明本文提出的预测模型与目前土壤温度中常用的数据驱动模型，在预测结果散点图和准确度统计结果的对比中，均表现出更高的精确度。

然而由于 FLUXNE 采集数据限制，本文主要针对白山和海北区域中 5 cm 土壤深度进行实验。不同的土壤深度和更多的观测站点对预测模型性能的影响需要进一步研究。同时在实验过程中发现，不同损失函数的使用导致预测结果不同，因此针对损失函数改进对预测能力的影响是必要的。

参考文献：

[1] FENG Y, CUI N, HAO W, *et al*.. Estimation of soil temperature from meteorological data using different machine learning models[J]. *Geoderma*, 2019, 338: 67-77.

[2] SANIKHANIA H, DEOB R C, YASEENC Z M, *et al*.. Non-tuned data intelligent model for soil temperature estimation: A new approach[J]. *Geoderma*, 2018, 330: 52-64.

[3] CORNU J Y, DENAIX L, LACOSTE J, *et al*.. Impact of temperature on the dynamics of organic matter and on the soil-to-plant transfer of Cd, Zn and Pb in a contaminated agricultural soil[J]. *Environmental Science and Pollution Research*, 2016, 23: 2997-3007.

[4] KIM Y, C STILL J, HANSON C V, *et al*.. Canopy skin temperature variations in relation to cli-

mate, soil temperature, and carbon flux at a ponderosa pine forest in central Oregon[J]. *Agricultural and Forest Meteorology*, 2016, 226: 161-173.

[5] YANGA J, BUSEN H, SCHERB H, *et al*.. Modeling of radon exhalation from soil influenced by environmental parameters[J]. *Science of The Total Environment*, 2019, 656: 1304-1311.

[6] BHADANI P, VASHISHT V, SOIL MOISTURE. Temperature and humidity measurement using arduino [C]. *The 9th International Conference on Cloud Computing, Data Science & Engineering, Noida, India*, 2019:567-571.

[7] HU G, LIN Z, WU X, *et al*.. An analytical model for estimating soil temperature profiles on the Qinghai-Tibet plateau of China[J]. *Journal of Arid Land*, 2016, 8 (2): 232-240.

[8] LIANG L L, RIVEROS-IREGUI D A, EMANUEL R E, *et al*.. A simple framework to estimate distributed soil temperature from discrete air temperature measurements in data-scarce regions[J]. *Geophys. Res*, 2017, 119: 407-417

[9] 徐英，谷雨，彭冬亮，等. 基于 DRGAN 和支持向量机的合成孔径雷达图像目标识别[J]. 光学 精密工程, 2020, 28(3):727-735.
XU Y, GU Y, PENG D L, *et al*.. SAR ATR based on disentangled representation learning generative adversarial networks and support vector machine[J]. *Opt. Precision Eng.*, 2020, 28(3):727-735. (in Chinese)

[10] 闫敬文，陈宏达，刘蕾. 高光谱图像分类的研究进展[J]. 光学 精密工程,2019,27(3):680-693.
YAN J W, CHEN H D, LIU L. Overview of hyperspectral image classification[J]. *Opt. Precision Eng.*, 2019, 27(3):680-693. (in Chinese)

[11] 魏彤，周银鹤. 基于机器学习识别与标记分水岭分割的盲道图像定位[J]. 光学 精密工程, 2019, 27 (1):201-210.
WEI T, ZHOU Y H. Blind sidewalk image location based on machine learning recognition and marked watershed segmentation[J]. *Opt. Precision Eng.*, 2019, 27(1):201-210. (in Chinese)

[12] MIHALAKAKOU G. On estimating soil surface temperature profiles[J]. *Energy and Buildings*, 2002, 34(3):251-259.

[13] KISI O, SANIKHANI H. Modellinglong-termmonthlytemperaturesby several data-driven methods using geographical inputs[J]. *Int. J. Climatol*, 2015, 35 (13): 3834-3846.

[14] HUR S O, KIM W T, JUNG KH, *et al*.. Estimation of Soil Surface Temperature by Heat Flux in Soil[J]. *Korean Journal of Soil Science & Fertilizer*, 2004, 37(3):131-135.

[15] MOAZENZADEH R, MOHAMMADI B. Assessment of bio-inspired metaheuristic optimisation algorithms for estimating soil temperature[J]. *Geoderma*, 2019, 353 (11): 152-171

[16] DELBARI M, SHARIFAZARI S, MOHAMMADI E. Modeling daily soil temperature over diverse climate conditions in Iran-a comparison of multiple linear regression and support vector regression techniques[J]. *Theoretical and Applied Climatology*, 2019, 135 (3-4): 991-1001.

[17] LI G, LI W, TIAN X, *et al*.. Short-term electricity load forecasting based on the xgboost algorithm [J]. *Smart Grid*, 2017, 7 (4): 274-285.

[18] JI S, WANG X, ZHAO W, *et al*.. An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise[J]. *Mathematical Problems in Engineering*, 2019, 2019 (Article ID 8503252): 1-15.

[19] ALBERTOTORRES-BARRÁNA, ALONSO Á, DORRONSOROAB J R. Regression tree ensembles for wind energy and solar radiation prediction [J]. *Neurocomputing*, 2017, 326-327 (1): 151-160.

[20] ZHANG J, ZHU Y, ZHANG X, *et al*.. Developing a Long Short-Term Memory (LSTM) based model for predicting water table depth in agricultural areas[J]. *Journal of Hydrology*, 2018, 561 (6): 918-929.

[21] QING X, NIU Y. Hourly day-ahead solar irradiance prediction using weather forecasts by LSTM [J]. *Energy*, 2018, 148 (4): 461-468.

[22] LI Q L, HAO H B, ZHAO Y, *et al*.. GANs-LSTM model for soil temperature estimation from meteorological: a new approach[J]. *IEEE ACCESS*, 2020, 8: 59427-59443.

作者简介：

李清亮(1988－)，男，博士，副教授，硕士生导师，中国电子学会计算机视觉专业技术人员，2016 年于吉林大学获得博士学位，主要从事大气物理 AI 预报，机器学习和图像处理等方面的研究。E-mail: liqingliang @ mail. ccsfu. edu. cn

于繁华(1970－)，男，博士，教授，硕士生导师，吉林省高校新世纪人才，长春市第五批有突出贡献专家，2008 年于吉林大学交通学院获得博士学位，主要从事大气物理 AI 预报，机器学习和智能优化等方面的研究。E-mail: yufanhua@163. com

耿庆田(1972－)，男，博士，教授，硕士生导师，2016 年于吉林大学获得博士学位，主要从事机器学习和汽车电子等方面的研究。E-mail: qtgeng @ mail. ccsfu. edu. cn