

DOI:10.7522/j.issn.1000-0240.2021.0079

LIU Hongchao, MA Junjie, LI Ren. Simulation of the water-thermal features within the surface soil in Tanggula region, Qinghai-Tibet Plateau, by using KNN model[J]. Journal of Glaciology and Geocryology, 2021, 43(4): 1243-1252. [刘宏超, 马俊杰, 李韧. 基于KNN机器学习方法对青藏高原唐古拉地区表层土壤水热状况的模拟[J]. 冰川冻土, 2021, 43(4): 1243-1252.]

# 基于KNN机器学习方法对青藏高原唐古拉地区 表层土壤水热状况的模拟

刘宏超<sup>1</sup>, 马俊杰<sup>2,3</sup>, 李 韧<sup>2</sup>

(1. 兰州大学 大气科学学院 半干旱气候变化教育部重点实验室, 甘肃 兰州 730000; 2. 中国科学院 西北生态环境资源研究院 冰冻圈科学国家重点实验室/青藏高原冰冻圈观测研究站, 甘肃 兰州 730000; 3. 中国科学院大学, 北京 100049)

**摘 要:** 利用唐古拉站2004—2012年气象观测资料, 基于KNN算法, 结合机器学习思想, 建立了一个气象回归模型, 模拟了2005年唐古拉地区表层土壤水热变化趋势, 结合实测数据, 将模拟值与观测值进行对比, 并对模型模拟效果进行了评估。结果表明: KNN模型能够较好地模拟活动层表层土壤水热状况, 各层土壤温度的模拟值与观测值的相关系数均在0.99以上, 均方根误差在1.25℃以内; 不同深度土壤水分的模拟值与观测值的相关系数均在0.95以上, 均方根误差在0.02 m<sup>3</sup>·m<sup>-3</sup>以内。总体上, KNN模型能够对青藏高原多年冻土区唐古拉地区表层土壤水热状况进行较为精确地模拟, 该模型对于青藏高原其他地区的适用性有待进一步研究验证。

**关键词:** KNN; 机器学习; 青藏高原; 多年冻土; 土壤水热过程

**中图分类号:** P642.14 **文献标志码:** A **文章编号:** 1000-0240(2021)04-1243-10

## 0 引言

青藏高原平均海拔高度4 000 m以上, 在中国境内部分西起帕米尔高原, 东至横断山脉, 横跨31个经度; 南至喜马拉雅山脉南缘, 北迄昆仑山-祁连山北侧, 纵贯约13个纬度<sup>[1]</sup>。青藏高原以其高大的地形隆升于周围大气之中, 其强大的热力和动力作用直接作用于对流层中层, 由于其独特的环流特征, 青藏高原对亚洲季风的形成和演化具有十分重要的作用, 进而对我国乃至全球的气候系统都有着重要的影响<sup>[2-4]</sup>。较高的海拔和独特的地形地貌特征使得青藏高原成为我国冰冻圈的主要分布区<sup>[5]</sup>, 其多年冻土总面积约为106万平方公里<sup>[6]</sup>。多年冻土作为气候的产物, 气候变化对多年冻土活动层冻融过程有着很大的影响; 而与此同时, 多年冻土也是气候变化的指示器, 土壤温湿度的变化通过改变地表和大气间的能量和水分交换来显著影响当地

和东亚的气候系统<sup>[7-8]</sup>。而活动层作为大气与多年冻土之间的纽带, 其变化不仅影响自身的水热性质, 而且对水循环<sup>[9]</sup>、能量交换<sup>[10]</sup>、碳循环<sup>[11]</sup>、生态系统<sup>[12]</sup>及气候系统<sup>[13]</sup>等都产生重要的影响。

自20世纪以来, 以气候变暖为主要特征的全球变化在青藏高原及周边地区表现尤为凸出<sup>[14]</sup>, 青藏高原地区升温速率是同期全球升温的两倍以上<sup>[15]</sup>, 且降水量不断增加<sup>[16]</sup>。青藏高原气候的不断变暖引发了多年冻土的退化, 进而对活动层水热过程产生了影响<sup>[17-18]</sup>。水热过程的改变对于陆气间能量和水分交换、地表感热、潜热输送、水循环过程及生态环境等都会产生重要影响<sup>[19]</sup>。可见, 活动层水热变化过程对陆面过程的研究意义较大<sup>[20]</sup>。目前, 许多学者针对青藏高原活动层水热变化开展了大量的研究。总的来说, 这些研究可以分为两个方面: 一类是针对观测站点实测资料的分析<sup>[21-22]</sup>; 另一类是使用模型模拟来分析活动层水热过程的变化<sup>[23-24]</sup>。

收稿日期: 2020-01-08; 修订日期: 2020-05-15

基金项目: 自然科学基金项目(42071093; 41671070; 41941015; 41690142); 冰冻圈科学国家重点实验室创新群体项目(41721091)资助

作者简介: 刘宏超, 硕士研究生, 主要从事边界层与陆面过程研究. E-mail: liuhch18@lzu.edu.cn

通信作者: 李韧, 研究员, 主要从事寒区陆面过程、气候变化方面的研究. E-mail: liren@lzb.ac.cn

然而青藏高原地区地形复杂,下垫面种类复杂多样,基于实测资料的研究分析虽然能够准确获取活动层水热分布特征,但却因监测站点数目有限,无法开展大范围的研究。因此,模型模拟成为了研究人员较为常用的手段。但是,由于冻融过程中水热过程极为复杂,水分在相变过程中会释放或吸收大量热量,从而影响土壤温度的变化,二者之间相互影响,使得相关研究十分困难<sup>[25-26]</sup>。目前陆面模式中还没有较为合适的土壤水分迁移方案,因此许多模式中关于土壤水分的模拟结果偏差较大,并且大多数模式中还需要输入一些较难获取的土壤质地及下垫面参数信息,这些都给青藏高原地区活动层水热过程的研究带来了困难。

KNN(K-Nearest Neighbor)算法作为一种信息处理手段,在分类与回归算法领域占有重要的位置,被广泛的应用在机器学习<sup>[27]</sup>、数据挖掘<sup>[28]</sup>、文本分类<sup>[29]</sup>等领域。作为一种用于分类和回归的机器学习算法之一,KNN算法非常有效和易于实现。本研究着眼于青藏高原多年冻土区,基于KNN算法,建立了一个气象数据的回归模型,对青藏高原多年冻土区活动层表层土壤水热过程进行模拟,并对该模型的模拟效果进行了评估。希望能够对青藏高原多年冻土区水热过程研究提供一种新的研究思路,并为相关研究提供一定的参考。

## 1 研究区概况和数据来源

本文研究区域选择在青藏高原多年冻土区唐古拉综合观测场(33°04' N, 91°56' E),海拔高度5 100 m,位于唐古拉山垭口西南部的一个平缓的坡地上,属于典型连续多年冻土区,下垫面类型为高寒草甸,成团簇状,高度较低矮,一年中最高约为10 cm,植被覆盖度约为20%~30%左右<sup>[30]</sup>。其年平均气温为4.9℃,极端最高温为17.6℃,极端最低温为-29.6℃,年平均气压为538 hPa,年均降水量为436.7 mm,降水主要集中在5月至9月,此期间降水量约占全年降水总量的92%,降水强度以小型降水为主<sup>[31]</sup>。

本次模拟研究所选用的数据资料来自于中国科学院青藏高原冰冻圈观测研究站唐古拉综合观测场的2004年至2012年气象观测资料。气象观测资料所包含的观测要素有雪深、降水、气温、辐射量、空气湿度、风速、土壤热通量、土壤温度、土壤湿度等。观测仪器信息及安装高度等如表1所示。

表1 观测项目及观测仪器信息

Table 1 The observation items and instruments

观测项目	观测仪器信息	安装高度或深度
雪深	SR50-L, Camball	2 m
降水量	T-200B, Geonor	1.5 m
空气温湿度	HMP45C-L, Vaisala	2, 5, 10 m
辐射	CNR-1, Kipp & Zonen	2 m
风速	05103-L, R. M. Young	2, 5, 10 m
土壤热通量	HFP01, Campell	5, 10, 20 cm
表层土壤温度	105T, Campell	5, 10, 20, 40 cm
表层土壤含水量	CS616-L, Campell	5, 10, 20 cm

## 2 研究方法

### 2.1 KNN机器学习算法

K邻近法,或者说K最近邻(KNN)算法,是一种分类和回归方法,在机器学习中,它是懒惰学习(lazy learning)的代表。本文将KNN算法应用于对气象数据的回归,以期对青藏高原多年冻土区表层土壤的水热状况进行模拟预测。其基本思路是采用平均法,在样本集合空间中,找出与预测样本A的曼哈顿距离小于S的最近的K个邻居,对这K个邻居的属性进行平均得到的结果,作为预测样本A的属性值。S值是设定的一个阈值,其目的是防止出现较大的偏差。

本文中的具体算法流程如图1所示:

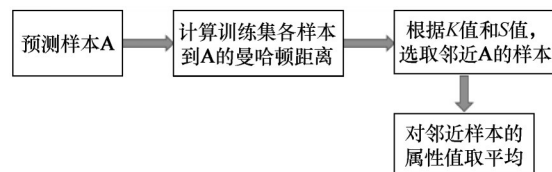


图1 KNN模型算法示意图

Fig. 1 Diagram of the KNN model structural

在将KNN算法应用于青藏高原多年冻土区表层土壤的气象要素的回归中,以5 cm深度土壤温度为例,具体的做法是将预测样本A输入到模型中,计算多维情况下训练集各个样本到预测样本A的曼哈顿距离,将曼哈顿距离小于设定的S值的K个最近样本视为预测样本A的邻近样本,计算这些邻近样本5 cm深度土壤温度,以此作为模型对样本A的预测值。

如果以X代表训练样本集,则

$$\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \dots, \mathbf{x}_n\} \quad (1)$$

对X中的每一个样本 $\mathbf{x}_i$ ,有

$$\mathbf{x}_i = [x_{(i,1)}, x_{(i,2)}, x_{(i,3)}, \dots, x_{(i,32)}] \quad (2)$$

对于预测样本  $\mathbf{A}$ , 则有:

$$\mathbf{A}=[a_1, a_2, a_3, \dots, a_{32}] \quad (3)$$

若以样本的第一个要素表示 5 cm 深度土壤温度, 即  $a_1$  代表预测样本  $\mathbf{A}$  的 5 cm 深度土壤温度, 则求取  $\mathbf{X}$  中每一样本到  $\mathbf{A}$  的曼哈顿距离得出集合  $\mathbf{L}$ :

$$\mathbf{L}=\{\sum_{j=2}^{32}(x_{(1,j)}-a_j), \sum_{j=2}^{32}(x_{(2,j)}-a_j), \dots, \sum_{j=2}^{32}(x_{(n,j)}-a_j)\} \quad (4)$$

对  $\mathbf{L}$  进行排序得到  $\mathbf{L}'$ :

$$\mathbf{L}'=\{\sum_{j=2}^{32}(x'_{(1,j)}-a_j), \sum_{j=2}^{32}(x'_{(2,j)}-a_j), \dots, \sum_{j=2}^{32}(x'_{(n,j)}-a_j)\} \quad (5)$$

模型的模拟值  $a'_1$  为:

$$a'_1=\frac{1}{K} \cdot \sum_{k=1}^K x'_{(k,1)} \quad (6)$$

式中:  $K$  值为设定选取的最近邻样本数量。

需要指出的是, 在模型搭建过程中,  $S$  值的取值很关键, 过大或者过小都有可能造成结果的偏差甚至模型搭建的失败。在本文中, 当  $S$  值取一固定值时, 会出现个别预测样本没有邻近样本的情况, 此时若增大  $S$  值会导致其余样本预测偏差增大, 由于需要对连续时间序列的要素进行预测, 故而本文采用了可变  $S$  值, 这是对传统 KNN 方法的一种改进。

## 2.2 样本集构建

如图 2 所示, 本文首先对中国科学院青藏高原冰冻圈观测研究站唐古拉综合观测场的 2004 年至 2012 年气象观测资料进行了预处理。具体做法是: 将每一日的各要素观测数据作为一个样本, 样本中任何一个要素观测值缺测即将该样本视为缺测样本, 剔除了缺测日数后共得到 3 009 个有效样本。为了去除 2005 年数据对模型模拟结果的影响, 从预处理数据中剔除了 2005 年的 365 个样本后, 剩余 2 644 个样本作为模型验证样本集。从模型验证样本集中随机抽取 80% 的样本构成了训练集, 剩余的 20% 构成了测试集。

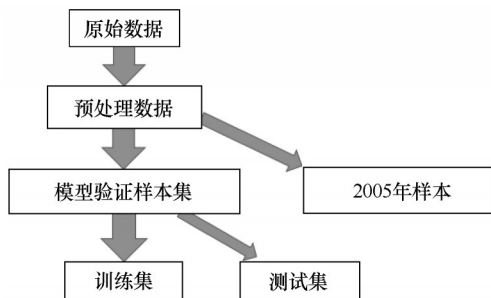


图 2 样本构建示意图

Fig. 2 Diagram of the sample construction

## 2.3 模型测试

利用模型对测试集的样本进行模拟, 模拟的 5 cm 深度的土壤温度与观测数值的偏差分布如图 3

所示。模型模拟 5 cm 深度的土壤温度的最大正偏差为 4.17 °C, 最大负偏差为 -3.77 °C, 平均偏差为 0.015 °C, 方差为 1.12, 模型模拟的 5 cm 深度的土壤温度与观测数值的相关系数为 0.98。从图 2 可以看出, 模型模拟偏差的绝对值主要集中在 1 °C 以内, 约占总数的 86.4%; 偏差绝对值在 2 °C 以上的仅有约 1.7%。如果将模型模拟结果偏差 2 °C 以内的结果视为正确的话, 那么该模型的模拟准确率在 98.3%, 可以认为该模型准确有效。

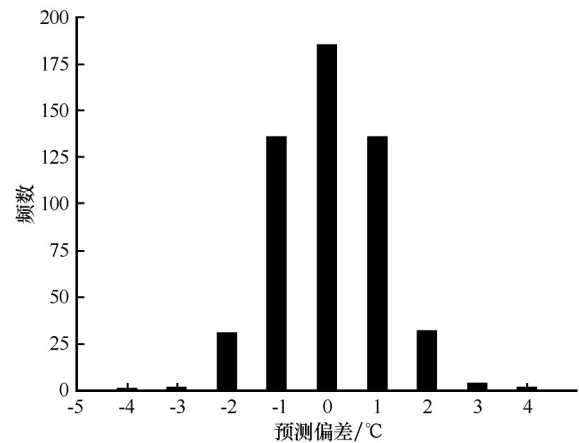


图 3 模型模拟偏差

Fig. 3 Model simulation bias

## 2.4 模型评估方法

### 2.4.1 统计参数

模型模拟的结果好坏可以根据一些统计参数进行评估, 相关系数  $CC$  (correlation coefficient) 和均方根误差  $RMSD$  (root-mean-square deviation) 是最常用的方法<sup>[32]</sup>。平均偏差  $MBE$  (mean bias error) 反映的是模拟值与观测值之间的偏离程度, 而标准差  $SD$  (standard deviation) 则代表了一个数据集数据分布的离散程度。这些统计参数的表达式为:

$$CC = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{N \sqrt{\frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\frac{1}{N} \sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (7)$$

$$RMSD = \sqrt{\frac{1}{N} \sum_{i=1}^n (X_i - Y_i)^2} \quad (8)$$

$$SD = \sqrt{\frac{1}{N} \sum_{i=1}^n (X_i - \bar{X})^2} \quad (9)$$

$$MBE = \frac{\sum_{i=1}^n (X_i - Y_i)}{n} \quad (10)$$

式中:  $X_i$ 、 $Y_i$  分别为模型预测值与观测值;  $N$  为样本容量;  $\bar{X}$ 、 $\bar{Y}$  分别为模型预测值与观测值的期望值, 在这里以其平均值代替。当模型效果理想时, 应有  $CC=1$ ,  $RMSD=0$ ,  $SD(X)=SD(Y)$ 。根据已有研究,



在进行模型效果检验时,当相关系数CC达到0.6时,可以认为模型的预测是成功的<sup>[33]</sup>。

#### 2.4.2 泰勒图

泰勒图能够将模型模拟结果与实测数据的标准差、均方根误差和相关系数在一张图上显示出来<sup>[34]</sup>。泰勒图可以比较直观地反映出模型模拟的效果的好坏,在泰勒图上,模拟结果至原点的距离代表了其标准差,方位角的余弦值代表了与观测值的相关系数,模拟结果至观测值之间的距离则代表了两者的均方根误差。

### 3 结果分析

#### 3.1 土壤温度模拟

土壤温度模拟是陆面过程模拟的重要部分,其模拟的准确性与精度对大气与土壤间的物质和能量交换过程的理解至关重要。图4给出了基于KNN机器学习模拟唐古拉活动层表层土壤(5~40 cm)温度与实测值之间的对比图。由图中可以看出,在不同深度,KNN模型模拟的土壤温度模拟值与实测值的跟随性都较好,成功地模拟出了表层土壤温度的年周期变化趋势。另外,不同深度的土壤温度在月尺度与旬尺度上模型模拟的效果较好,能较为精确地刻画出短周期内土壤温度的波动,这表明土壤温度的模拟是成功的。

然而,值得注意的是,12月初至2月底和6月中旬至8月中旬这两个时间段的模型模拟效果不是很理想。以5 cm深处的土壤温度为例,12月初至2月底时间段内,模型的模拟值与观测值相比明显偏大,该时段平均误差为0.93℃,均方根误差为1.45℃,而全年的平均误差仅为0.39℃,均方根误

差为1.19℃;6月中旬至8月中旬时间段内,月尺度与旬尺度的土壤温度波动未能很好地模拟,模拟值比实测值偏低,该时段平均误差为-0.20℃,均方根误差为0.79℃。对其原因进行分析,12月初至2月底和6月中旬至8月中旬这两个时间段,5 cm深度土壤温度的模拟值较年平均值分别偏高0.54℃与偏低0.59℃,可能与KNN算法模型的设计有关。这两个时期是全年5 cm深度土壤温度的最低(最高)时间段,K临近算法在筛选出给定样本的邻近样本后,在对目标要素(5 cm深度土壤温度)进行回归模拟时采用的方法是对邻近样本的属性进行平均,这会不可避免地造成极低(极高)值向平均值偏移。所以该模型在极值模拟的方面会存在一些缺陷,这也是下一步将要优化的内容。与模型模拟的全年5 cm深度土壤温度的均方根误差1.19℃相比,12月初至2月底的均方根误差偏大,而6月中旬至8月中旬的均方根误差偏小,这可能是后一时期的K邻近样本较多而前一时期的K近邻样本较少造成的。12月初至2月底和6月中旬至8月中旬这两个时间段5 cm深度实测土壤温度的平均值分别为-9.84℃、7.98℃,统计整个数据集中5 cm深度土壤温度在(-9.84±1.00)℃与(7.98±1.00)℃范围内的数据,其有效样本分别为222与125。由此可以看出样本集的容量对于KNN模型模拟结果的好坏有十分重要的影响,更大容量的样本集能降低KNN模型模拟结果的均方根误差,较小容量的样本集在进行KNN模型模拟时,则会因为K邻近样本的数量较少而产生比较大的误差。

总体可以认为,本研究中基于KNN机器学习所

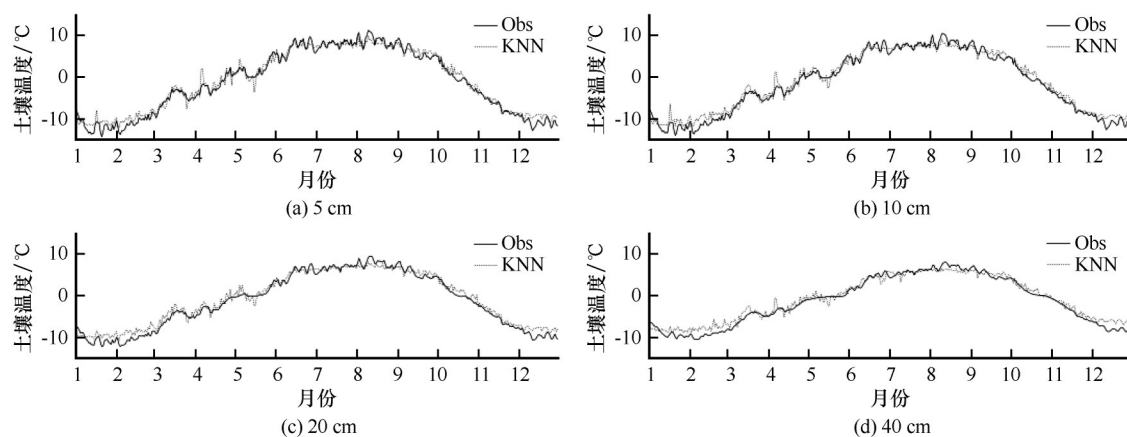


图4 不同深度土壤温度实测值(Obs)与KNN模拟值(KNN)对比

Fig. 4 Comparison of observed soil temperature (Obs) and KNN simulated value (KNN) at different depths

建立的模型可以非常好地对唐古拉地区活动层表层土壤温度进行模拟。

### 3.2 土壤水分模拟

多年冻土区活动层土壤水分是陆面过程的重要参量,它对地气热量交换、感热、潜热等都有重要影响,同时也对土壤温度有较大影响,因此,土壤水分的模拟也是陆面过程的重要研究内容<sup>[35]</sup>。图5为不同深度土壤体积含水量模拟值与实测值的对比图,由图中实测土壤水分曲线可以看出,冬春季不同深度土壤水分较低,而夏秋季节土壤水分较高。由春季至夏季,由于温度升高,导致表层土壤逐渐由冻结状态融化成未冻状态产生液态水,冻土的融化与雨季的到来共同造成了表层土壤含水量的逐渐升高<sup>[36]</sup>。夏秋季节,由于降水频繁,导致土壤水分较高,波动较大,这与唐古拉地区雨季区间相符,可见降水对表层土壤水分有较为显著的影响。

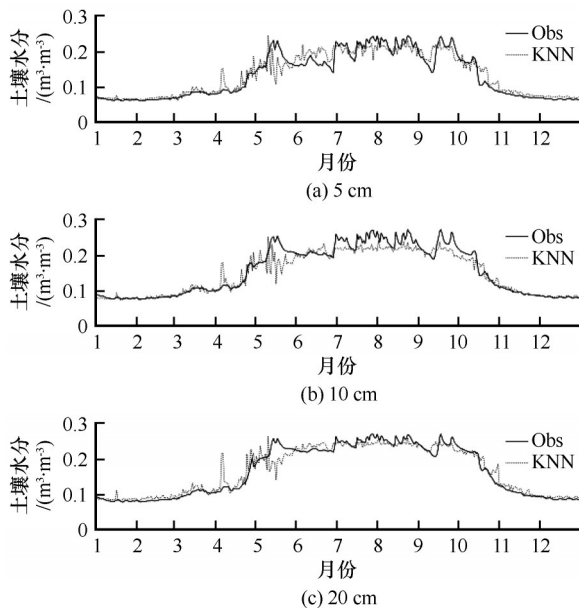


图5 不同深度土壤水分实测值(Obs)与KNN模拟值(KNN)对比

Fig. 5 Comparison of observed soil moisture (Obs) and KNN simulated value (KNN) at different depths

KNN模型不同深度土壤湿度的模拟值与观测值趋势较为一致,在冬春季模拟效果较好,数值上也较为吻合,两条曲线基本重合。但在5月至9月,5 cm深度土壤水分模拟与观测的平均值分别是 $0.190 \text{ m}^3 \cdot \text{m}^{-3}$ 与 $0.193 \text{ m}^3 \cdot \text{m}^{-3}$ ,模拟值的平均偏差 $-0.0026 \text{ m}^3 \cdot \text{m}^{-3}$ 与全年的平均偏差 $0.0029 \text{ m}^3 \cdot \text{m}^{-3}$ 相比有所偏低,造成该现象的原因与本文中土壤温度

模拟偏差的原因一致。从均方根误差来看,KNN模型对5 cm深度土壤水分模拟结果的均方根误差是 $0.029 \text{ m}^3 \cdot \text{m}^{-3}$ 与全年均方根误差 $0.021 \text{ m}^3 \cdot \text{m}^{-3}$ 相比略有偏大,表明5月至9月的模拟效果确实略有降低。这可能是由于表层土壤水分受外界降水影响较大,而该方法没有较好的捕捉到这一信息,因此导致模拟效果不如冻结期理想。然而考虑到土壤水分和温度相互影响,水热耦合机理认识尚不明确,土壤水分迁移机理复杂,所以对土壤水分的模拟较为困难。对10 cm、20 cm深度土壤水分的模拟结果与5 cm深度的情形相似,需要说明的是,随着深度的增加模拟的效果(平均偏差、均方根误差)并没有逐渐变差的趋势。

目前许多陆面模式并不能很好地模拟土壤水分的年变化特征<sup>[31,37]</sup>,KNN模型对表层土壤水分的模拟效果相较于大多陆面模式已有相当程度的提高,今后改善KNN模型对土壤水分模拟效果的工作或许可以从引入额外相关变量、配置适当权重、提高样本量等方面着手。

### 3.3 模拟结果评估

#### 3.3.1 土壤温度

图6~7给出了不同深度土壤温度模拟值与观测值之间的误差分析情况。由图6不同深度土壤温度模拟值与观测值散点图比较中可以看出,不同深度土壤温度的模拟值与观测值十分吻合,均较为集中地分布在1:1线附近,模拟效果较为理想。从图7中各统计指标的泰勒图中可以看出,不同深度土壤温度的模拟值与观测值的相关系数都非常高,均处于0.99以上水平,均方根误差也都在 $1^\circ\text{C}$ 附近。另外,不同深度土壤温度的模拟值与观测值的各统计指标均波动不大,说明模拟效果十分稳定。从图7(a)~7(d)中还可以看到,随着土壤深度的增加,无论观测值还是模拟值,其自身的标准差都在减少,也即数据本身的离散程度在减少,这可能是由于随着土壤深度的增加,土壤温度对外界环境变化(如气温、辐射等)的响应程度在减少。这点从图6中不同深度土壤的模拟值与观测值的散点图中也可以看出,随着土壤深度的增加,数据的分布更为集中,土壤温度的波动范围在减少。

#### 3.3.2 土壤水分

图8~9给出了不同深度土壤水分模拟值与观测值的误差分析情况,由图8中不同深度土壤水分模拟值与观测值散点图可以看出,KNN机器学习方法

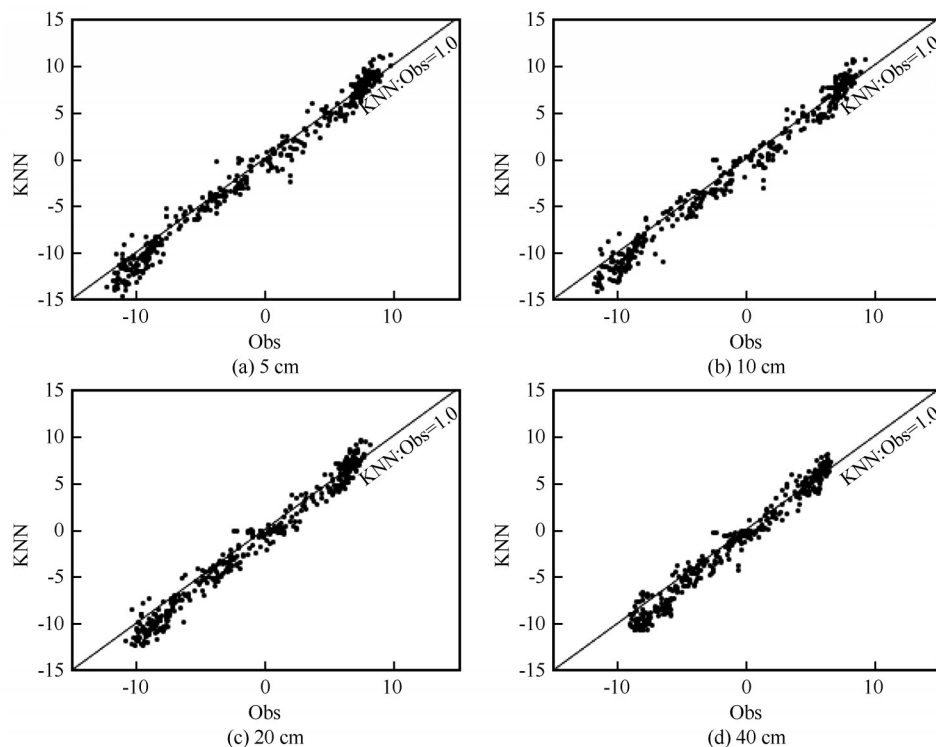


图6 不同深度土壤温度模拟值与观测值散点图

Fig. 6 Scatter plots of simulated and observed soil temperature at different depths

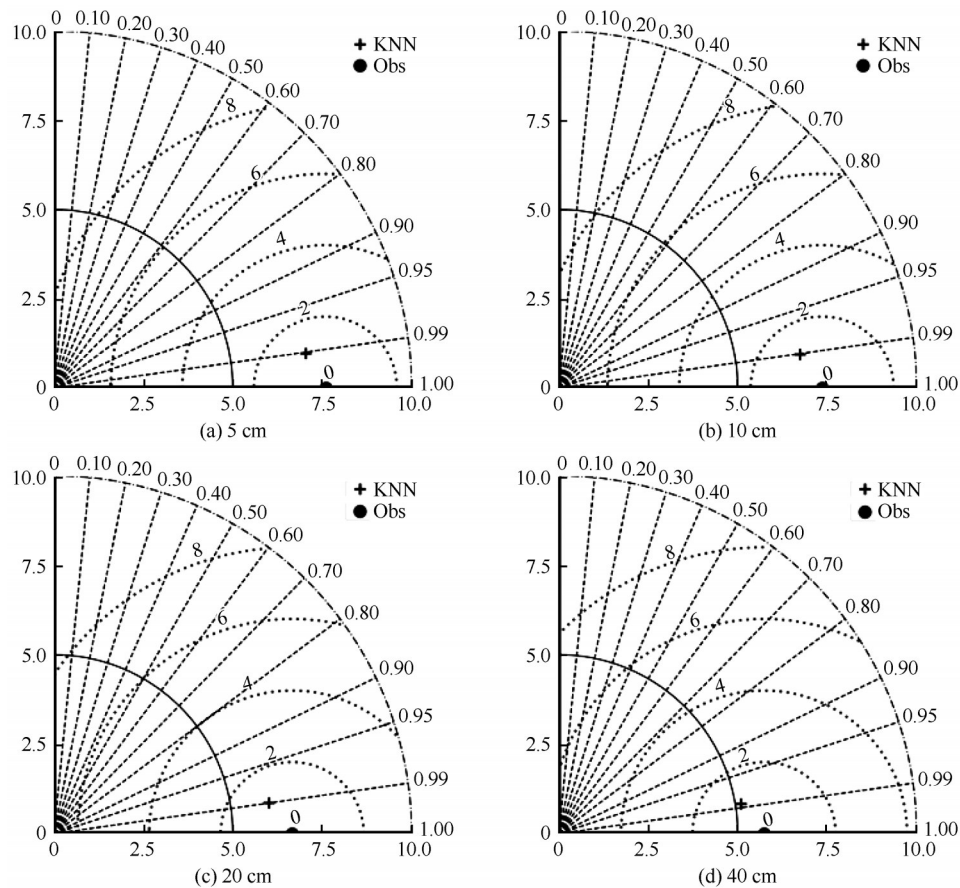


图7 不同深度土壤温度模拟值与观测值的统计指标泰勒图

Fig. 7 Taylor plots for statistical indicators of simulated and observed soil temperature at different depths



对土壤水分的模拟是成功的,都较为集中的分布在 1:1 线两侧,但是在土壤体积含水量处于  $0.15 \text{ m}^3 \cdot \text{m}^{-3}$  附近时,分布较为散乱,误差相对较大。从图 9 的统计指标的泰勒图中可以看出,不同深度土壤水分的模拟值与观测值的相关系数都在 0.95 以上,均方根误差都在  $0.02 \text{ m}^3 \cdot \text{m}^{-3}$  左右,与土壤温度不同的

是,随着土壤深度的增加,模拟值与观测值的标准差都在增加,这可能是由于表层土壤水分受下层土壤水分的影响较大的原因。另外,从泰勒图中可以看到,不同深度土壤水分的模拟值与观测值的统计指标都相差不大,这也反映了模型模拟的效果较为稳定。

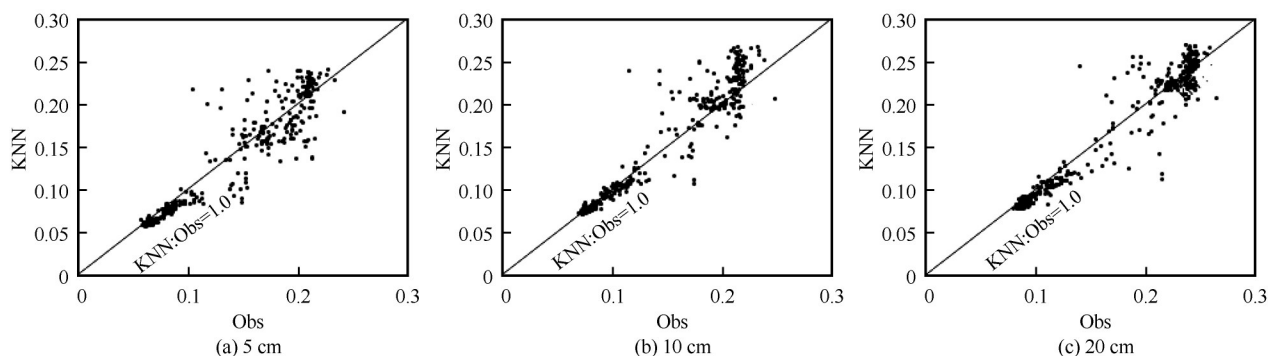


图 8 不同深度土壤湿度模拟值与观测值散点图

Fig. 8 Scatter plots of simulated and observed soil moisture at different depths

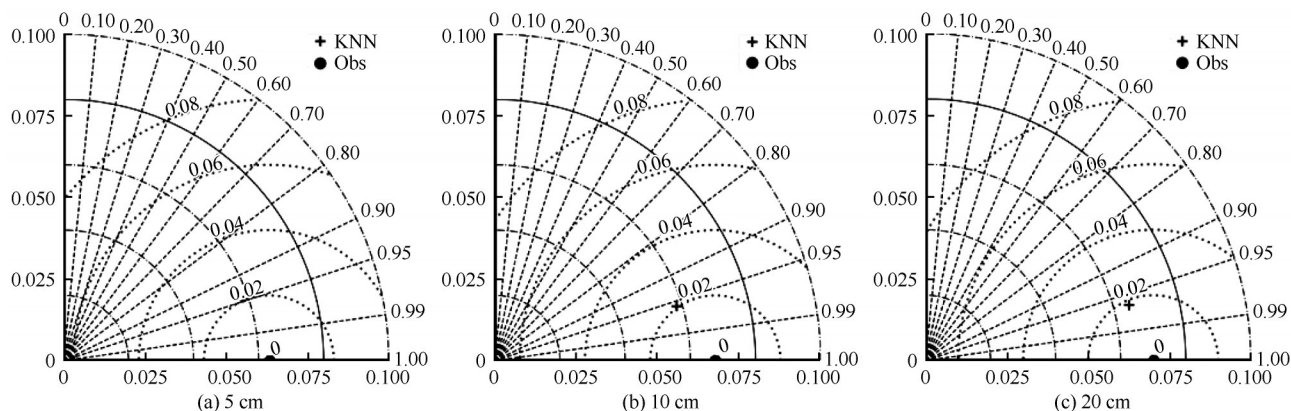


图 9 不同深度土壤水分模拟值与观测值的统计指标泰勒图

Fig. 9 Taylor plots for statistical indicators of simulated and observed soil moisture at different depths

#### 4 讨论

土壤温湿度的模拟是陆面过程研究的热点,已有多位研究者采用了不同的模型来对青藏高原多年冻土区土壤温湿度进行模拟<sup>[24,38]</sup>。许多学者利用 SHAW 模型来模拟土壤温度变化特征,然而在 SHAW 模型中,植被信息、地表反照率等输入变量的偏差都会对输出结果有很大影响<sup>[31]</sup>。GIPL2 模型由于对土壤温度模拟效果较好,同样得到了许多学者的青睐,然而该模型同样需要输入植被类型信息,土壤热导率参数如导热率和体积热容等<sup>[39]</sup>。本研究基于 KNN 算法,建立了一个气象数据的回归模型,对青藏高原多年冻土区水热过程进行了模拟,

并对该模型的模拟效果进行了评估。统计整理有关唐古拉地区水热过程其他陆面模式的模拟结果并同本文的模拟结果进行了比较(表 2),可以看出本文对土壤温湿度的模拟效果还是较为理想的。另外,值得注意的是,作为一种新的研究思路,本模型的搭建相比于其他陆面模式来说更为快捷高效,且只需输入较易获得的气象参数,不需要输入研究站点的植被及土壤质地等信息,操作方法简便,运行速度快。这在下垫面类型复杂且缺乏土壤质地等信息的青藏高原地区不失为一种较为理想的方法。

然而, KNN 模型还存在一定的缺陷,主要体现在以下几个方面:首先,该方法需要有一个较长时

表2 本文模型与其他模型模拟相关系数比较

Table 2 Comparison of correlation coefficients between the model in this paper and other models

深度	土壤温度相关系数			土壤水分相关系数		
	本研究	Coupmodel <sup>[38]</sup>	SHAW <sup>[31]</sup>	本研究	Coupmodel <sup>[38]</sup>	SHAW <sup>[31]</sup>
5 cm	0.99	0.99	0.96	0.95	0.88	0.86
10 cm	0.99	0.99	0.97	0.95	0.91	0.79
20 cm	0.99	0.99	0.97	0.95	0.93	0.73
40 cm	0.99	0.99	0.97			

间段的数据来作为样本进行训练和测试;其次,由于该模型对历史数据有较强的依赖性,无法较好地对极值进行模拟;最后,本文只针对唐古拉地区进行了模型的适用性分析,在其他站点是否能够很好的模拟还有待于进一步的分析验证。

5 结论

本文的研究中,利用唐古拉站 2004—2012 年气象站观测资料,基于 KNN 算法,建立了一个气象数据的回归模型,对 2005 年表层土壤温湿度进行了模拟试验,并对模型模拟结果进行了评估分析,将模拟值与观测值对比,得出以下结论:

(1)KNN 机器学习模型能够很好地模拟唐古拉地区表层土壤温度,不同深度土壤温度的模拟值与观测值的相关系数均达到了 0.99 以上,均方根误差都在 1.25℃以内,对其年内变化趋势及数值上均能很好的模拟再现。这在缺乏某一时间段观测资料的情况下,KNN 模型模拟结果能够作为一种较为理想的替代资料。

(2)在土壤水分的模拟试验中,KNN 机器学习模型能够较好地模拟表层土壤水分的年变化特征,不同深度土壤湿度的模拟值与实测值的相关系数均达到了 0.95 以上,均方根误差均在 0.02 m<sup>3</sup>·m<sup>-3</sup>以内。但在融化期的模拟效果不如冻结期,这可能是由于该模型没有很好地捕捉到外界降水对土壤水分的影响,下一步可以尝试通过优化降水量对土壤水分影响的权重来提高模型模拟精度。

总之,KNN 算法模型能够很好地应用于气象数据的回归模拟,其模拟的趋势与精度都较为理想,模拟结果具有很高的参考价值。同时,该模型还具有一定的优化空间,在最优 K 值设定、权重优化等方面还需要进一步的探索。下一步工作将主要放在模型的改进及模型在其他地区的适用性验证上。

参考文献(References):

[1] Zhang Yili, Li Bingyuan, Zheng Du. A discussion on the

boundary and area of the Tibetan Plateau in China[J]. Geographical Research, 2000, 22(3): 193-198. [张懿锂, 李炳元, 郑度. 论青藏高原范围与面积[J]. 地理研究, 2000, 22(3): 193-198. ]

[2] Qiao Yu, Zhou Shunwu, Ma Yue, et al. Dynamic effect of Tibetan Plateau and its impact on weather and climate in China[J]. Meteorological Science and Technology, 2014, 42(6): 1039-1046. [乔钰, 周顺武, 马悦, 等. 青藏高原的动力作用及其对中国天气气候的影响[J]. 气象科技, 2014, 42(6): 1039-1046. ]

[3] Yang Meixue, Wang Xuejia, Pang Guojin, et al. The Tibetan Plateau cryosphere: Observations and model simulations for current status and recent changes[J]. Earth- Science Reviews, 2019, 190: 353-369.

[4] Wu Guoxiong, Liu Yimin, He Bian, et al. Thermal controls on the Asian summer monsoon[J]. Scientific Reports, 2012, 2(1):1-7.

[5] Yao Tandong, Thompson L, Yang Wei, et al. Different glacier status with atmospheric circulations in Tibetan Plateau and surroundings[J]. Nature Climate Change, 2012, 1580: 1-5.

[6] Zou Defu, Zhao Lin, Sheng Yu, et al. A new map of permafrost distribution on the Tibetan Plateau[J]. The Cryosphere, 2017, 11: 2527-2542.

[7] Yang Meixue, Yao Tandong, Gou Xiaohua, et al. Water recycling between the land surface and atmosphere on the Northern Tibetan Plateau: a case study at flat observation sites[J]. Arctic Antarctic & Alpine Research, 2007, 39(4): 694-698.

[8] 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: 125-161.

[9] Wang Xuejia, Pang Guojin, Yang Meixue, et al. Effects of modified soil water-heat physics on RegCM4 simulations of climate over the Tibetan Plateau[J]. Journal of Geophysical Research, 2016, 121(12): 6692-6712.

[10] Li Ren, Yang Wen, Ji Guoliang, et al. The 40 a variational characteristics of surface heating field over Wudaoliang in the northern Tibetan Plateau[J]. Acta Energy Solaris Sinica, 2006, 26(6): 868-873. [李韧, 杨文, 季国良, 等. 40年来藏北高原五道梁地区地表加热场的变化特征[J]. 太阳能学报, 2006, 26(6): 868-873. ]

[11] Zhu Dan, Ciais P, Krinner G, et al. Controls of soil organic matter on soil thermal dynamics in the northern high latitudes[J]. Nature Communications, 2019, 10(1):1-9.

[12] Jorgenson M T, Osterkamp T E. Response of boreal ecosystems to varying modes of permafrost degradation[J]. Canadian Journal of Forest Research, 2005, 35(9): 2100-2111.

[13] Poutou E, Krinner G, Genthon C, et al. Role of soil freezing in future boreal climate change[J]. Climate Dynamics, 2004, 23(6): 621-639.

[14] Pang Guojin, Wang Xuejia, Yang Meixue. Using the NDVI to



- identify variations in, and responses of, vegetation to climate change on the Tibetan Plateau from 1982 to 2012[J]. *Quaternary International*, 2017, 444: 87-96.
- [15] Chen Deliang, Xu Baiqing, Yao Tandong, et al. Assessment of past, present and future environmental changes on the Tibetan Plateau[J]. *Chin. Sci. Bull.* 2015, 60(32): 3025-3035. [陈德亮, 徐柏青, 姚檀栋, 等. 青藏高原环境变化科学评估: 过去、现在与未来[J]. *科学通报*, 2015, 60(32): 3025-3035.]
- [16] Wang Xuejia, Pang Guojin, Yang Meixue. Precipitation over the Tibetan Plateau during recent decades: a review based on observations and simulations[J]. *International Journal of Climatology*, 2018, 38: 1116-1131.
- [17] Hu Guojie, Zhao Lin, Liren, et al. Variations in soil temperature from 1980 to 2015 in permafrost regions on the Qinghai-Tibetan Plateau based on observed and reanalysis products[J]. *Geoderma*, 2019, 337: 893-905.
- [18] Cuo L, Zhang Y, Bohn T J, et al. Frozen soil degradation and its effects on surface hydrology in the northern Tibetan Plateau[J]. *Journal of Geophysical Research: Atmospheres*, 2015, 120(16): 8276-8298.
- [19] Yang Yong, Chen Rensheng. Research review on hydrology in the permafrost and seasonal frozen regions[J]. *Advances in Earth Science*, 2011, 26(7): 711-723. [阳勇, 陈仁升. 冻土水文研究进展[J]. *地球科学进展*, 2011, 26(7): 711-723.]
- [20] Li Ren, Zhao Lin, Ding Yongjian, et al. Impact of surface energy variation on thawing process within active layer of permafrost[J]. *Journal of Glaciology and Geocryology*, 2011, 33(6): 1235-1242. [李韧, 赵林, 丁永建, 等. 地表能量变化对多年冻土活动层融化过程的影响[J]. *冰川冻土*, 2011, 33(6): 1235-1242.]
- [21] Hu Guojie, Zhao Lin, Li Ren, et al. Characteristics of hydro-thermal transfer during freezing and thawing period in permafrost regions[J]. *Soils*, 2014, 46(2): 355-360. [胡国杰, 赵林, 李韧, 等. 青藏高原多年冻土区土壤冻融期间水热运移特征分析[J]. *土壤*, 2014, 46(2): 355-360.]
- [22] Jiao Yongliang, Li Ren, Zhao Lin, et al. Processes of soil thawing-freezing and features of soil moisture migration in the permafrost active layer[J]. *Journal of Glaciology and Geocryology*, 2014, 36(2): 237-247. [焦永亮, 李韧, 赵林, 等. 多年冻土区活动层冻融状况及土壤水分运移特征[J]. *冰川冻土*, 2014, 36(2): 237-247.]
- [23] Zhang Wei, Wang Genxu, Zhou Jian, et al. Simulating the water-heat processes in the permafrost regions in the Tibetan Plateau based on CoupModel[J]. *Journal of Glaciology and Geocryology*, 2012, 34(5): 1099-1109. [张伟, 王根绪, 周剑, 等. 基于 CoupModel 的青藏高原多年冻土区土壤水热过程模拟[J]. *冰川冻土*, 2012, 34(5): 1099-1109.]
- [24] Yang Yong, Chen Rensheng, Ye Baisheng, et al. Heat and water transfer processes on the typical underlying surfaces of frozen soil in cold regions(I): model comparison[J]. *Journal of Glaciology and Geocryology*, 2013, 35(6): 1545-1554. [阳勇, 陈仁升, 叶柏生, 等. 寒区典型下垫面冻土水热过程对比研究(I): 模型对比[J]. *冰川冻土*, 2013, 35(6): 1545-1554.]
- [25] Gao Yanhong, Cheng Guodong, Shang Lunyu, et al. Application of atmospheric model coupled with frozen soil parameterization to simulating spring soil condition in the Qilian Mountains[J]. *Journal of Glaciology and Geocryology*, 2007, 29(1): 82-90. [高艳红, 程国栋, 尚伦宇, 等. 耦合冻土方案的大气模式对祁连山区春季土壤状况的模拟[J]. *冰川冻土*, 2007, 29(1): 82-90.]
- [26] Hinzman L D, Kane D L, Gieck R E, et al. Hydrologic and thermal properties of the active layer in the Alaskan Arctic[J]. *Cold Regions Science and Technology*, 1991, 19(2): 95-110.
- [27] Zhang Ning, Jia Ziyang, Shi Zhongzhi. Text categorization with KNN algorithm[J]. *Computer Engineering*, 2005, 31(8): 171-172. [张宁, 贾自艳, 史忠植. 使用 KNN 算法的文本分类[J]. *计算机工程*, 2005, 31(8): 171-172.]
- [28] Xiong Yajun, Liao Xiaonong, Li Ziming, et al. Application of KNN data mining algorithm to haze grade forecasting in Beijing[J]. *Meteorological Monthly*, 2015, 41(1): 98-104. [熊亚军, 廖晓农, 李梓铭, 等. KNN 数据挖掘算法在北京地区霾等级预报中的应用[J]. *气象*, 2015, 41(1): 98-104.]
- [29] Zhang Haoyu, Yao Gang, Yin Zhizhu, et al. Fault state perception method for six-phase PMSM based on wavelet neural network and KNN machine learning algorithm[J]. *Electrical Measurement & Instrumentation*, 2017, 56(2): 1-9. [张昊宇, 姚钢, 殷志柱, 等. 基于小波神经网络与 KNN 机器学习算法的六相永磁同步电机故障态势感知方法[J]. *电测与仪表*, 2017, 56(2): 1-9.]
- [30] Yao Jimin, Zhao Lin, Ding Yongjian, et al. The surface energy budget and evapotranspiration in the Tanggula region on the Tibetan Plateau[J]. *Cold Regions Science and Technology*, 2008, 52: 326-360.
- [31] Liu Yang, Zhao Lin, Li Ren. Simulation of the soil water-thermal features within the active layer in Tanggula region, Tibetan Plateau, by using SHAW model[J]. *Journal of Glaciology and Geocryology*, 2013, 35(2): 280-290. [刘杨, 赵林, 李韧. 基于 SHAW 模型的青藏高原唐古拉地区活动层土壤水热特征模拟[J]. *冰川冻土*, 2013, 35(2): 280-290.]
- [32] Jolliffe I T, Stephenson D B. Forecast verification: a practitioner's guide in atmospheric science[M]. New York: John Wiley and Son, 2003.
- [33] Wang Yuan, Qian Hao, Song Jinjie, et al. Verification of the T213 global spectral model of China National Meteorology Center over the East-Asia area[J]. *Journal of Geophysical Research*, 2008, 113(D10110): 1-7.
- [34] Taylor K E. Summarizing multiple aspects of model performance in a single diagram[J]. *Journal of Geophysical Research*, 2001, 106(D7): 7183-7192.
- [35] Diro G T, Sushama L, Huziy O. Snow-atmosphere coupling and its impact on temperature variability and extremes over North America[J]. *Climatic Dynamic*, 2017. DOI: org/10. 1007/s00382-017-3788-5.
- [36] Jiao Yongliang, Li Ren, Zhao Lin, et al. Processes of soil thawing-freezing and features of soil moisture migration in the permafrost active layer[J]. *Journal of Glaciology and Geocryology*, 2014, 36(2): 237-247. [焦永亮, 李韧, 赵林, 等. 多年冻土区活动层冻融状况及土壤水分运移特征[J]. *冰川冻土*, 2014, 36(2): 237-247.]
- [37] Yang Yong, Chen Rensheng, Ji Xibin, et al. Heat and water transfer processes on alpine meadow frozen grounds[J]. *Advances in Water Science*, 2010, 21(1): 30-35. [阳勇, 陈仁升, 吉喜斌, 等. 黑河高山草甸冻土带水热传输过程[J]. *水科学进展*, 2010, 21(1): 30-35.]
- [38] Hu Guojie, Zhao Lin, Li Ren, et al. The Water-thermal characteristics of frozen soil under freeze-thaw based on CoupModel[J]. *Scientia Geographica Sinica*, 2013, 33(3): 356-362. [胡国杰, 赵林, 李韧, 等. 基于 CoupModel 模型的冻融土壤水热耦合模拟研究[J]. *地理科学*, 2013, 33(3): 356-362.]
- [39] Qin Yanhui, Wu Tonghua, Li Ren, et al. Thermal condition of the active layer on the Qinghai-Tibet Plateau simulated by using

the model of GIPL2[J]. *Journal of Glaciology and Geocryology*, 2018, 40(6): 1153-1166. [秦艳慧, 吴通华, 李韧, 等.

基于GIPL2模型的青藏高原活动层土壤热状况模拟研究[J]. *冰川冻土*, 2018, 40(6): 1153-1166. ]

## Simulation of the water-thermal features within the surface soil in Tanggula region, Qinghai-Tibet Plateau, by using KNN model

LIU Hongchao<sup>1</sup>, MA Junjie<sup>2,3</sup>, LI Ren<sup>2</sup>

(1. Key Laboratory for Semi-Arid Climate Change of the Ministry of Education, College of Atmospheric Sciences, Lanzhou University, Lanzhou 730000, China; 2. Cryosphere Research Station on the Qinghai-Tibet Plateau, State Key Laboratory of Cryosphere Sciences, Northwest Institute of Eco-Environment and Resources, Chinese Academy of Sciences, Lanzhou 730000, China; 3. University of Chinese Academy of Sciences, Beijing 100049, China)

**Abstract:** Based on the meteorological observation data from 2004 to 2012 at Tanggula Station, using the KNN algorithm, combined with machine learning ideas, a meteorological regression model was established to simulate the surface soil water-thermal change trend in Tanggula region in 2005, and the measured values were combined with the observed values. The value comparison is made to evaluate the simulation effect of the model. The results show that the KNN model can well simulate the hydrothermal conditions of the soil in the active layer. The results show that the KNN model can well simulate the soil water and heat conditions in the active layer, and the correlation coefficients between the simulated values of the simulated soil temperature and the observed values are above 0.99, and the root mean square error is within 1.25 °C; The correlation coefficients between the simulated and observed values of soil moisture are above 0.95, and the root mean square error is within 0.02 m<sup>3</sup>·m<sup>-3</sup>. In general, the KNN model can simulate the hydrothermal process of Tanggula station in the permafrost region of the Qinghai-Tibet Plateau, and its applicability to other parts of the Qinghai-Tibet Plateau needs further research and verification.

**Key words:** KNN; machine learning; Qinghai-Tibet Plateau; permafrost; hydrothermal process

(责任编辑: 温智; 编辑: 庞瑜, 周成林)