文章编号：1672-6987(2023)02-0116-11；DOI：10.16351/j.1672-6987.2023.02.016

基于高斯混合聚类和 LightGBM 算法的印度洋次表层温度反演研究

汤贵艳，朱善良*，周伟峰，杨树国

（青岛科技大学 数理学院；数学与交叉科学研究院；青岛市人工智能海洋技术创新中心，山东 青岛 266061）

摘　要：海洋次表层的热力结构对于海洋环流和全球气候变化具有重要的意义。提出一种新的融合高斯混合模型(gaussion mixture model，GMM)和轻量级梯度提升机(light gradient boosting machine，LightGBM)算法的海洋次表层温度(ocean subsurface temperature，OST)反演模型，利用海表温度(sea surface temperature，SST)、海表盐度(sea surface salinity，SSS)、海表高度(sea surface height，SSH)、海表风场(sea surface wind，SSW)的水平分量(USSW)和垂直分量(VSSW)等多源海表参数对印度洋海域的次表层热力结构进行反演，并采用均方根误差和决定系数对模型进行验证。结果表明：所提出的模型可以准确反演印度洋海域的 OST 分布特征和季节变化规律。在此基础上，设计了不同海表参数输入组合的 3 种对比实验来定量分析不同海表参数对 LightGBM 模型的影响。结果表明：所有海表参数对模型都有积极作用，但 5 个输入参数(SST、SSS、SSH、USSW 和 VSSW)的 LightGBM 模型反演效果最好，3 个输入参数(SST、SSS 和 SSH)和 2 个输入参数(SST 和 SSH)的 LightGBM 模型次之。另外，与已有的极限梯度增强(extreme gradient boosting，XGBoost)反演模型相比，5 个输入参数的 LightGBM 模型具有更好的模拟能力。

关键词：高斯混合模型；轻量级梯度提升机；机器学习；海洋次表层温度

中图分类号：P 732　　**文献标志码：**A

引用格式：汤贵艳，朱善良，周伟峰，等.基于高斯混合聚类和 LightGBM 算法的印度洋次表层温度反演研究[J]. 青岛科技大学学报(自然科学版)，2023，44(2)：116-126.

TANG Guiyan，ZHU Shanliang，ZHOU Weifeng，et al. Estimation of Indian Ocean subsurface thermal structure based on Gaussion mixture clustering and LightGBM algorithm[J]. Journal of Qingdao University of Science and Technology(Natural Science Edition)，2023，44(2)：116-126.

Estimation of Indian Ocean Subsurface Thermal Structure Based on Gaussion Mixture Clustering and LightGBM Algorithm

TANG Guiyan，ZHU Shanliang，ZHOU Weifeng，YANG Shuguo

(College of Mathematics and Physics；Research Institute for Mathematics and Interdisciplinary Sciences；Qingdao Innovation Center of Artifical Intelligence Ocean Technology，Qingdao University of Science and Technology，Qingdao 266061，China)

Abstract：The thermal structure of the ocean subsurface is of great significance to ocean circulation and global climate change. In this paper，a new ocean subsurface temperature

收稿日期：2022-05-11
基金项目：中国科学院海洋环流与波动重点实验室开放基金项目(KLOCW2003).
作者简介：汤贵艳(1995—)，女，硕士研究生.　*通信联系人.

(OST) estimation model combining the Gaussian mixture model (GMM) and light gradient boosting machine (LightGBM) algorithm. The model uses multisource sea surface parameters including sea surface temperature (SST), sea surface salinity (SSS), sea surface height (SSH), northward and eastward components of sea surface wind (USSW and VSSW) to retrieve the OST of the Indian Ocean. Moreover, the root mean square error and coefficients of determination are employed to assess the performance of the model. The results show that the model proposed in this paper can accurately reflect the distribution characteristics and seasonal variation of the OST in the Indian Ocean. On this basis, three comparative experiments with different input combinations of sea surface parameters were designed to quantitatively analyze the influence of different input variables on the LightGBM model. The experimental results show that all sea surface parameters have a positive effect on the model. Moreover, the LightGBM model with five input parameters (SST, SSS, SSH, USSW, VSSW) has the best estimation effect, followed by the LightGBM model with three input parameters (SST, SSS, SSH) and two input parameters (SST, SSH). In addition, compared with the existing eXtreme gradient boosting (XGBoost) estimation model, the LightGBM model with five input parameters has better simulation capabilities.

Key words: Gaussian mixture model; light gradient boosting machine; machine learning; ocean subsurface thermal

海洋热含量是大尺度海洋环流结构和全球气候变化分析中的重要指标之一。大量的研究表明,近十余年来 300～2 000 m 的中深层海洋暖化现象影响着全球变暖进程[1-4]。因此,在全球变暖的背景下,海洋次表层热力结构的研究受到广泛关注[5-7]。而印度洋变暖在调节区域和全球范围内气候变化方面发挥着重要作用[8],例如,印度洋热含量的年际变化在 Walker 环流和南极涛动中起到重要影响[9];印度洋变暖不仅是对 El Nina 现象的被动反应,而且影响到印度-西太平洋地区夏季气候变化[10];热带印度洋变暖加剧有利于西太平洋通过大气形成更强的信风,可能促成类似 La Nina 现象的出现,其对太平洋气候波动造成影响[11]。由此可见,开展印度洋次表层热力结构的重构研究具有重要的理论意义和应用价值。然而,由于海洋次表层温度现场观测受到诸多因素制约而导致数据稀疏或缺失,所以如何基于海洋表层相关参数信息估算次表层热力结构成为当前物理海洋研究领域亟需解决的重要问题[6]。

近年来,随着卫星遥感数据、Argo 观测数据以及再分析数据的不断丰富,越来越多的学者基于经验统计模型开展海洋次表层温盐结构的研究[12-15]。然而,这些经验统计模型普遍存在输入参数单一,数据信息利用不足,模型精度不高等问题。随着人工智能技术的发展,人工智能在海洋学领域的应用受到越来越多的重视[16],人工神经网络(artificial neural network, ANN)、卷积神经网络(convolutional neural network, CNN)和自组织神经映射(self-organizing map, SOM)神经网络等人工智能技术已被用于估算海洋次表层热结构[8, 17-25]。例如,SU 等[19]采用随机森林(random forest, RF)算法反演了全球海洋次表层温度异常(subsurface temperature anomaly, STA),结果表明,RF 算法具有较好的反演能力。随后,SU 等[20]提出了地理加权回归模型反演印度洋的 STA,采用极限梯度增强(extreme gradient boosting, XGBoost)模型研究了 2015 年不同季节全球海洋的温盐异常[21];ZHANG 等[22]利用多层卷积长短期记忆神经网络算法,反演了全球海洋次表层三维温度结构,结果表明,中上层的估算精度高于中深层,而中深层估算精度还有待提高。LU 等[25]将多源海表遥感数据和 Argo 实测数据作为聚类变量,利用 K-means 聚类将全球海洋划分为不同的区域,在此基础上,运用 ANN 反演全球 STA,结果表明预聚类的模型优于未聚类的同类方法。

虽然学者们针对海洋次表层热力结构已做了一些研究,并取得了一些有意义的成果。但仍存在一些不足[16-17],例如,动力学模型具有昂贵的计算代价,统计回归模型存在输入参数少,时空分辨率低以及估算精度有待提高等问题。已有研究表明,融合

多种算法的混合模型不仅可以弥补单一模型训练时间过长的缺点,而且可以提高模型估算精度[25-26]。例如,与K-means聚类算法]相比,GMM聚类算法更适合大规模数据集,面向非等球型簇的聚类效果更好[27]。海洋数据具有多源、多类及多维度等特征,但融合GMM聚类算法的混合反演模型还不多见。另一方面,目前,针对反演海洋次表层热力结构的研究区域大多为全球海域,对局部海域的研究成果不足。因此,在局部海域上开展海洋次表层热力结构的混合反演模型研究,进一步提高模型估算精度仍是非常有意义的工作。

本工作针对印度洋海域,基于多源海表遥感数据和Argo实测数据,提出了一种融合GMM预聚类方法和LightGBM算法的海洋次表层温度(ocean subsurface temperature, OST)反演模型,该模型首先通过GMM聚类算法对印度洋海域聚类分析,进而利用LightGBM模型反演重构印度洋的次表层温度结构,并采用均方根误差(root mean square error, RMSE)和决定系数(coefficient of determination, R^2)等指标来评估反演模型的性能。并通过构建不同海表参数输入组合的对比实验来定量分析不同输入变量对OST估算的影响,与已有的XGBoost反演模型作对比来验证本模型的优越性。在此基础上,研究了印度洋海域OST的季节性变化特征和规律。

1 算法基础

1.1 高斯混合聚类算法

GMM聚类是一种常见的聚类算法[27],其假设数据是从多个高斯分布中生成的,并通过一定的权重将多个高斯模型融合成一个高斯混合模型。该聚类算法适合应用于海量数据,对于非等球型数据集的聚类效果优于层次聚类、K-Medoids和SOM聚类[27]。GMM聚类算法的数学原理如下:

首先,对GMM的各个变量进行初始化,由K个高斯分布组成的GMM的概率密度函数为

$$p(x)=\sum_{k=1}^{K}\pi_k N_k(x\mid\mu_k,\Sigma_k)\ 。 \tag{1}$$

其中,K为高斯分布的个数,每个高斯分布在GMM中被称为一个组件,$N_k(x\mid\mu_k,\Sigma_k)$为高斯概率密度函数,π_k、μ_k和Σ_k分别为GMM中第k个组件的权重系数、期望和协方差矩阵,且$\sum_{k=1}^{K}\pi_k=1$。

其次,针对所有的高斯模型,采用极大似然估计的方法来确定各个参数值。得到所有样本的似然函数$\Pi_{i=1}^{N}p(x_i)$,对其取对数后得到函数

$$\sum_{i=1}^{N}\lg\left\{\sum_{k=1}^{K}\pi_k N_k(x_i\mid\mu_k,\Sigma_k)\right\}\ 。 \tag{2}$$

用期望最大化算法(expectation-maximization algorithm, EM)估计GMM模型的各个参数,由期望步(E-step)得到每个数据x_i由第k个组件生成的概率

$$\Upsilon(i,k)=\frac{\pi_k N(x_i\mid\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\pi_j N(x_i\mid\mu_j,\Sigma_j)}\ 。 \tag{3}$$

由极大化步(M-step)得GMM的权重系数、期望和协方差矩阵的迭代公式

$$\pi_k=\frac{N_k}{N}\ , \tag{4}$$

$$\mu_k=\frac{1}{N_k}\sum_{i=1}^{N}\Upsilon(i,k)x_i\ , \tag{5}$$

$$\Sigma_k=\frac{1}{N_k}\sum_{i=1}^{N}\Upsilon(i,k)(x_i-\mu_k)(x_i-\mu_k)^{\mathrm{T}}\ 。 \tag{6}$$

根据以上迭代公式,不断迭代更新模型参数,直到似然函数的值收敛则停止,否则继续迭代,得到合适的参数值。

1.2 LightGBM算法

LightGBM算法是微软团队提出的一种基于梯度提升决策树(gradient boosting decision tree, GBDT)的Boosting算法[28]。其基本思想是将弱学习器提升为强学习器[23],它是以决策树作为弱学习器,每一次迭代,将之前学习器损失函数的负梯度作为残差去拟合建立新的学习器,通过不断迭代更新参数来训练模型。

首先,对模型进行初始化,设y_i为真实值,$\hat{y}_i$为预测值,构造损失函数,目标函数由损失函数和正则化项组成,可以表示为

$$Ob_j=\sum_{i=1}^{n}l(\hat{y}_i,y_i)+\sum_{k=1}^{K}\Omega(f_k)\ 。 \tag{7}$$

其中,f_k为预测函数,$\sum_{k=1}^{K}\Omega(f_k)$为正则化项,且$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda|\omega|^2=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}\omega_j^2$,$\gamma$为模型的惩罚系数,$T$和$\omega$分别为第$k$棵树叶子的数量和叶子的权重,正则化项可以控制模型的复杂度,防止过拟合现象的发生。

其次,在迭代优化目标函数时,对式(7)中的损

失函数进行二阶泰勒展开,第 t 步的迭代公式为

$$Ob_j^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) = \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + \text{cons}\, a\tan t = \sum_{i=1}^{n} \left[l(y_i, y_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) + \text{cons}\, a\tan t。 \tag{8}$$

其中,$g_i = \partial_{y^{(t-1)}} l(y_i, y_i^{(t-1)})$ 为损失函数中的一阶导数,$h_i = \partial_{y^{(t-1)}}{}^2 l(y_i, y_i^{(t-1)})$ 为损失函数中的二阶导数。

目标函数式(8)可简化为

$$Ob_j^{(t)} \approx \sum_{j=1}^{T} \left[\left(\sum_{i \in I_j} g_i \right) \omega_j + \frac{1}{2} \left(\sum_{i \in I_j} h_i + \lambda \right) \omega_j^2 \right] + \gamma T。 \tag{9}$$

通过优化策略求解目标函数的最小值,该值即为最优结构的决策树。当信息增益小于阈值 γ 时或者树达到最大深度时则停止建立新的决策树。

经过不断的迭代更新,每轮产生的弱学习器通过线性相加的方式得到最终的强学习器

$$F(x) = \partial_0 f_0(x) + \partial_1 f_1(x) + \cdots + \partial_m f_m(x)。 \tag{10}$$

其中,$F(x)$ 为强学习器,$f_m(x)$ 为弱学习器,∂_m 为第 m 个弱学习器的权重系数。

与 GBDT 算法和 XGBoost 算法相比,LightGBM 算法的改进优化策略主要包括直方图算法、带深度限制的 leaf-wise 的决策树生长策略等。LightGBM 算法采用的 leaf-wise 生长策略是在所有叶子节点中选择每次分裂增益最大的叶子节点进行分裂,这样可降低误差,提高效率并避免过拟合现象的发生。而且,LightGBM 算法还创新性地引入梯度单边采样算法,通过保留较大梯度样本和随即抽取较小梯度样本,减少了数据量,引入独立特征合并算法,将互斥特征捆绑以降低特征维数。通过数据采样和特征合并,大大加快了算法的训练速度,提高了运行效率。

2　改进的 OST 反演模型

本工作聚焦到印度洋海域,利用该海域的多源海表参数数据进行聚类分析,在此基础上构建新的融合 GMM 聚类和 LightGBM 算法的 OST 反演模型。

2.1　数据来源

本工作选取印度洋作为研究区域(30°E～120°E,50°S～30°N)。多源海表参数数据说明如表 1 所示,SST 和 SSS 来源于国际太平洋研究中心提供的 2005 年 1 月—2018 年 12 月的 Argo 网格化数据,Argo 数据的空间分辨率均为 1°×1°,时间分辨率为逐月,深度为 0～2 000 m(http://apdrc.soest.hawaii.edu/projects/Argo/data/gridded/On_standard_levels/index-1.html)。SSH 是由 AVISO 卫星高度计提供的逐月网格化数据集,空间分辨率为 0.25°×0.25°(https://www.aviso.altimetry.fr/en/data/products/sea-surface-height-products/global.html)。SSW(包括水平分量 USSW 和垂直分量 VSSW)来源于具有较高精度和适用性的交叉校准多平台(cross calibrated multi-platform, CCMP)v02.0 版本风场分析数据,CCMP 涵盖 1987 年 1 月～2019 年 4 月的数据,同时集成了多个微波辐射计和散射计观测数据,空间分辨率为 0.25°×0.25°(https://data.remss.com/ccmp/v02.0/)。

表 1　数据说明

Table 1　Interpretation of data

数据	来源	时间	空间分辨率
SSH	AVISO	200501—201812	0.25°×0.25°
SSW	CCMP	200501—201812	0.25°×0.25°
SST、SSS	Argo	200501—201812	1°×1°

2.2　改进的 OST 反演模型

为了提高模型的估算精度,本研究提出了一种改进的融合 GMM 聚类和 LightGBM 算法的 OST 混合反演模型。图 1 给出了反演模型估算 OST 的技术流程图。

1)对数据进行预处理。数据集包含 SST、SSS、SSH、USSW 和 VSSW 五个海表参量,时间跨度为 2005 年 1 月—2018 年 12 月。本工作数据的时间分辨率为月,通过插值方法空间分辨率统一为 0.5°×0.5°,为了在空间上匹配所有变量,如果任何变量在某一位置均为空,则删除该节点。最后,为了消除异常数据的影响,对数据进行归一化处理

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}}。 \tag{11}$$

其中,x^* 为样本数据归一化后的数据,x 为样本数据,$x_{\max}$ 为样本数据的最大值,$x_{\min}$ 为样本数据的最小值。另外,本工作将 2005 年 1 月—2018 年 12 月共 168 个月的数据划分为训练集和测试集,其中 2005 年 1 月—2017 年 12 月共 156 个月的数据作为训练集,2018 年 1 月—2018 年 12 月共 12 个月的数

据作为测试集。利用 Argo 观测的 OST 数据作为训练和测试标记。

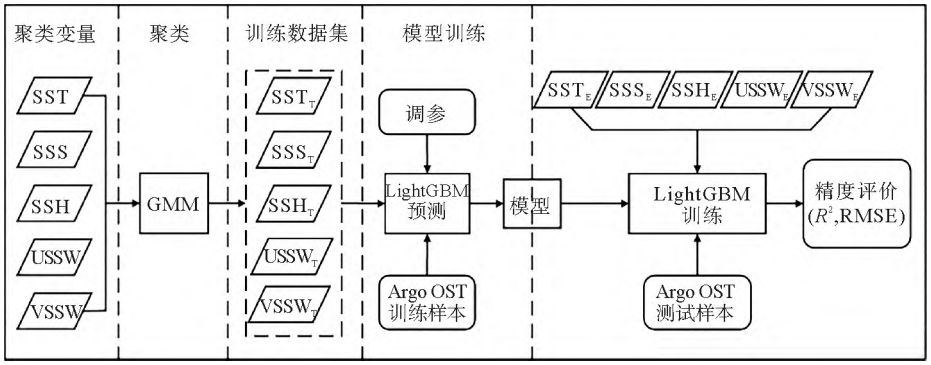

图1 融合 GMM 聚类和 LightGBM 算法的 OST 反演模型流程图

Fig.1 Flowchart of the OST estimation model combining the GMM clustering and LightGBM algorithm

2)基于 GMM 聚类算法对训练集和测试集的印度洋海表数据进行聚类分析。首先建立 GMM 的概率密度函数 $p(x)$,并初始化模型参数 π_k、μ_k 和 Σ_k,然后利用 EM 算法确定模型的参数,并基于贝叶斯信息准则确定最佳聚类数 K,算法详细流程如图 2 所示。

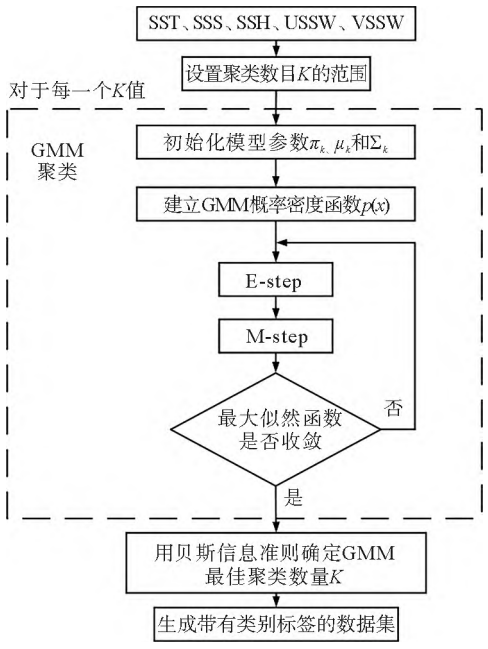

图2 GMM 聚类算法流程图

Fig.2 Flowchart of GMM clustering algorithm

步骤 1 给定聚类数目 K,随机选择一组模型,初始化模型的参数 π_k、μ_k 和 Σ_k,对这组 GMM 的概率密度函数 $p(x)$ 取对数,得最大似然函数式(2);

步骤 2 E-step:计算后验概率式(3),在计算后验概率时,参数 π_k、μ_k 和 Σ_k 取初始值或者上一轮迭代所得到的值;

步骤 3 M-step:根据 E-step 计算的后验概率 $\Upsilon(i,k)$ 再重新计算参数 π_k、μ_k 和 Σ_k;

步骤 4 重复迭代 E-step 和 M-step 直至似然函数收敛,获得最优参数;

步骤 5 用贝叶斯信息准则确定 GMM 的最佳聚类数 K,最终将印度洋海域划分属性不同的 4 种海域,生成带有类别标记的数据集。

3)将 GMM 聚类后带有类别标记的数据集作为 OST 反演模型的训练集和测试集。训练集用于训练 OST 反演模型,测试集用于评估该模型的准确率。将 SST、SSS、SSH、USSW 和 VSSW 等海表参数作为 LightGBM 算法的输入变量,将 Argo 观测的 OST 数据作为训练和测试标签。对 LightGBM 算法的关键参数调优,首先确定学习率,本工作测试了 0.01~0.1 范围内的数,结果表明,learning_rate=0.01 时,不仅收敛速度较快且效果较好;继而调整决策树的其他基本参数,对于控制树模型复杂度的 num_leaves,本工作测试了 10~50 范围内的数,结果表明,num_leaves=30 时效果最好;对于迭代次数 n_estimators,本工作测试了 1 000~5 000 之间的数,结果表明迭代次数为 2 000 时效果最好;对于控制每次迭代特征选择比例的 feature_fraction 以及采样比例 bagging_fraction,本工作测试了 0.6~1 之间的数,结果表明当其分别为 0.9 和 0.8 时效果最好。通过调整这些参数不断提高模型的准确率,当效果没有进一步提升时停止。最终该算法的参数值见表 2。

表 2　LightGBM 算法的主要参数

Table 2　The main parameters of LightGBM algorithm

参数	含义	数值
num_leaves	叶子数量	30
n_estimators	迭代次数	2 000
learning_rate	学习率	0.01
feature_fraction	建树的特征选择比例	0.9
bagging_fraction	建树的样本采样比例	0.8
bagging_freq	采样的频率	30
lambda_l1、lambda_l2	L1 和 L2 正则化项的权重系数	1×10^{-5}

4)利用本工作构建的基于 GMM 聚类和 LightGBM 算法的 OST 反演模型估算印度洋海域不同深度上的 OST,并采用 RMSE 和 R^2 两个指标来评价反演模型的估算精度和有效性,R^2 的计算公式为

$$R^2=1-\frac{\sum_{i=1}^{n}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n}(\bar{y}-y_i)^2}。\tag{12}$$

RMSE 的计算公式为

$$e_{\mathrm{RMSE}}=\sqrt{\frac{1}{N}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}。\tag{13}$$

其中,$\sum_{i=1}^{n}(\hat{y}_1-y_i)^2$ 为预测值与真实值的平方差之和,$\sum_{i=1}^{n}(\bar{y}-y_i)^2$ 为均值与真实值的平方差之和。

5)为了进一步评价改进的 OST 反演模型的估算性能,本文将该模型与 XGBoost 反演模型进行对比分析,以验证所构建的模型性能。

3　结果与分析

3.1　模型反演结果的验证

为检验融合 GMM 聚类和 LightGBM 算法的 OST 反演模型的有效性,利用该模型分别对印度洋次表层 26 个深度上的温度进行估算。图 3 分别给出了该模型估算的 OST 和 Argo 实测的 OST 在 2018 年 11 月 50、500 和 1 000 m 处的空间分布图。由图 3 所示,50、500 和 1 000 m 处反演模型估算的 OST 空间分布与 Argo 实测的 OST 空间分布都较吻合。

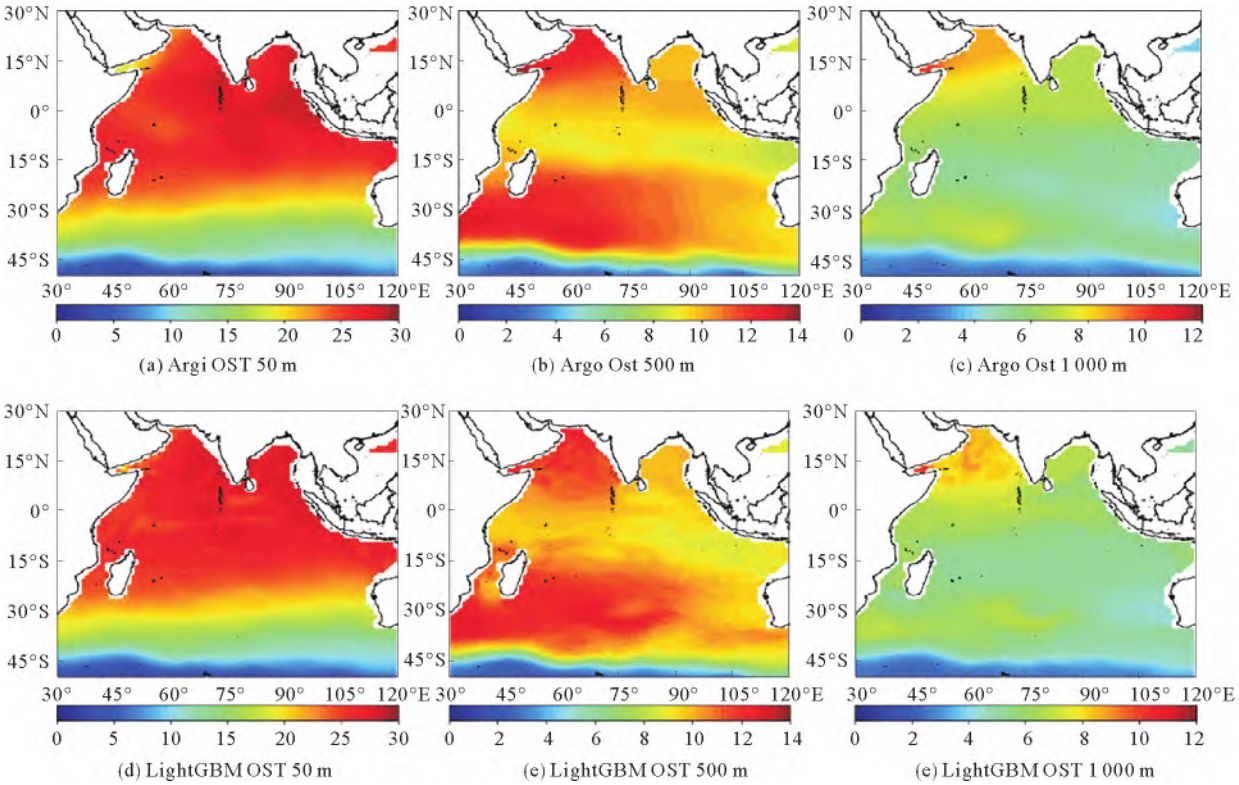

图 3　50、500 和 1 000 m 处 Argo 观测和反演模型估算的 OST 分布

Fig.3　Spatial distribution of the OST from Argo-observed and estimation model-estimated at depths of 50, 500 and 1 000 m

图 4 给出了模型估算与 Agro 观测的 OST 的残差分布,如图 4 所示,模型估算的 OST 残差大部分都较小,在深度为 500 m 处,残差控制在 2.5 ℃以内,在深度为 1 000 m 处,残差控制在 2 ℃以内。

这表明本工作模型能够准确反演印度洋的OST,较好地揭示了OST的空间分布特征。垂向上,OST的变化范围逐渐缩小,如50 m处不同海域的OST空间分布大致为0～30 ℃、500 m处0～14 ℃、1 000 m处0～12 ℃,这说明内部海洋温度随深度增加逐渐趋于稳定且空间异质性不显著。

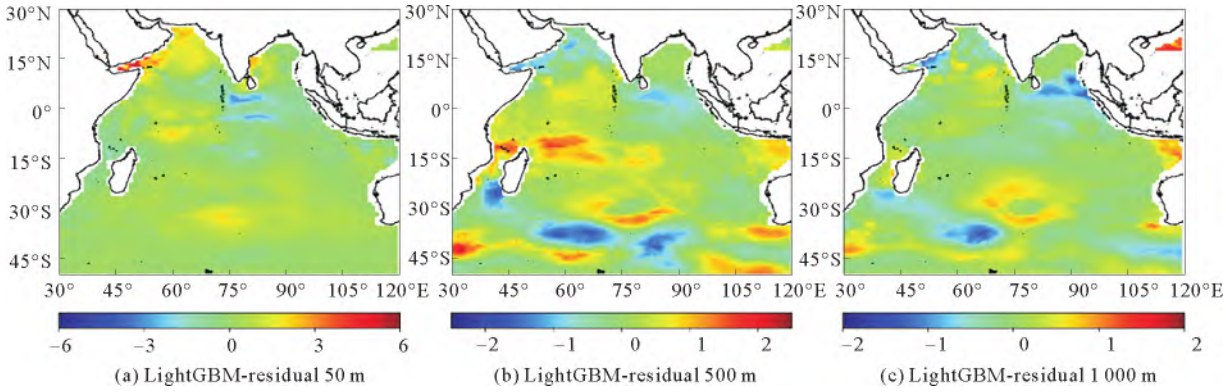

图4 在50、500和1 000 m深度反演模型估算和Argo观测的OST的残差分布

Fig.4 Spatial distribution of the difference between the estimation model-estimated and Argo-observed OST at depths of 50, 500 and 1 000 m

为定量评估OST反演模型的估算精度,本工作又计算了2018年11月在垂向上的RMSE和R^2值,表3展示了10～1 000 m不同深度上的计算结果。结果表明,RMSE的最小值和最大值分别为0.007 4 ℃和0.338 9 ℃,垂向平均的RMSE值为0.197 0 ℃;R^2的最大值和最小值分别为0.999 9和0.878 0,垂向平均的R^2值为0.953 0。上述结果表明该模型估算精度较高,反演效果好。然而该模型在不同的深度下性能有所差异,随着深度增加,模型的RMSE和R^2存在波动,大致上RMSE值越来越大,R^2值越来越小,说明模型反演效果有所下降,这可能是由于印度洋表层与内部存在复杂的非线性动力过程,该模型难以准确捕捉到这些变化规律。

表3 10～1 000 m深度下反演模型的RMSE和R^2值

Table 3 RMSE and R^2 values of estimation model at depths from 10 m to 1 000 m

深度/m	e_{RMSE}/℃	R^2	深度/m	e_{RMSE}/℃	R^2
10	0.011 0	0.999 9	250	0.227 0	0.949 2
20	0.027 8	0.999 2	300	0.231 9	0.946 1
30	0.042 9	0.998 2	400	0.235 9	0.943 0
50	0.100 2	0.990 2	500	0.247 8	0.936 5
75	0.169 1	0.997 2	600	0.275 4	0.921 5
100	0.196 0	0.963 2	700	0.272 3	0.923 9
125	0.216 2	0.955 8	800	0.281 3	0.919 3
150	0.189 4	0.965 8	900	0.274 8	0.924 4
200	0.213 8	0.956 0	1000	0.257 5	0.931 7

3.2 模型对比分析

为验证本工作所构建模型的可靠性,定量分析不同海表参数输入组合对该模型估算OST的影响,并与已有的XGBoost反演模型[21]作对比,本工作设计了4种实验,2个输入参数(SST和SSH)的LightGBM-2模型、3个输入参数(SST、SSS和SSH)的LightGBM-3模型、5个输入参数(SST、SSS、SSH、USSW和VSSW)的LightGBM-5和XGBoost-5模型,如表4所示。

表4 实验设计

Table 4 Experimental design

实验名称	输入参数	反演模型
LightGBM-2	SST、SSH	LightGBM
LightGBM-3	SST、SSS、SSH	LightGBM
LightGBM-5	SST、SSS、SSH、USSW、VSSW	LightGBM
XGBoost-5	SST、SSS、SSH、USSW、VSSW	XGBoost

图5展示了2018年11月4个实验条件下反演模型的RMSE和R^2垂向分布。如图5所示,不同深度的OST估算精度有所差异,在深度5～1 500 m之间,各个模型的RMSE值总体呈现出先增大后缓慢减小的趋势,R^2值呈现出先减小后缓慢增大的趋势,在深度1 500 m以下海域,RMSE值又变大,R^2值变小。结果表明,模型的估算精度随深度降低,各个模型反演性能都有减弱趋势,这说明仅用海表参数信息来反演海洋深层的温度是不够准确的,

还需要考虑其他海洋现象的影响。但随着深度的增加,LightGBM-5 模型的优势逐渐显现出来,其估算精度最高,XGBoost-5 模型的估算精度次之。从所有 26 个深度的平均 RMSE 和 R^2 来看,LightGBM-2 模型的平均 RMSE 值为 0.387 2 ℃,R^2 值为 0.800 3;LightGBM-3 模型的平均 RMSE 值为 0.231 1 ℃,R^2 值为 0.934 3;LightGBM-5 模型的平均 RMSE 值为 0.197 0 ℃,R^2 值为 0.953 1;XGBoost 模型的平均 RMSE 值为 0.228 1 ℃,R^2 值为 0.936 5。可以发现,输入变量的数量影响模型估算精度,所选择的海表参数对模型都有积极影响,LightGBM-5 的估算结果较可靠,估算精度也较高,与 XGBoost-5 模型相比较,该模型更适合估算印度洋的 OST。

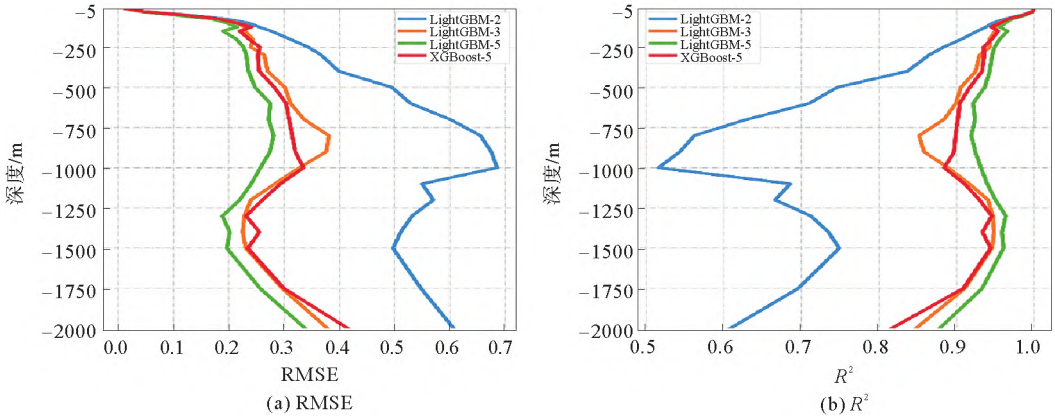

图 5　不同反演模型的 RMSE 和 R^2 垂向剖面

Fig.5　The vertical profile of RMSE and R^2 for different estimation models

为进一步对比 5 个输入参数的 LightGBM 模型与已有 XGBoost 模型的反演效果,图 6 给出了 2018 年 11 月 LightGBM 模型和 XGBoost 模型在 30、400 和 700 m 上模拟的 OST 和 Argo 观测的 OST 对比。结果表明,在深度为 30 m 处,Argo 观测的 OST 整体呈现西冷东暖趋势,LightGBM 模型能够模拟出这种空间分布特征,反演效果较好,而 XGBoost 模型估算的 OST 分布较为均匀,无法准确揭示这种分布特征;在深度为 400 和 700 m 处,两种模型估算的 OST 基本一致,都能够估算出 OST 的空间分布特征,相比较而言,LightGBM 模型的反演效果更好。

图 7 又给出了 2018 年 11 月 50、500 和 1 000 m 深度的 XGBoost 模型估算 OST 与 Argo 观测 OST 的残差分布,对比图 4 和图 7 可以看出,LightGBM 模型和 XGBoost 模型两种模型估算 OST 的残差分布有差异,LightGBM 模型估算残差整体上小于 XGBoost 模型估算残差。因此,LightGBM 模型能更好地估算印度洋的 OST,可以准确捕捉到深层海洋的温度分布特点。此外,在印度洋边界和南极环流区域附近的南印度洋的估算残差要大于印度洋的其他区域,这可能是由于边界流和南极环流的影响。

3.3　估算结果的季节性分析

为研究模型估算精度在不同季节上的表现,本工作利用改进的 OST 反演模型估算了 2018 年冬季(1 月)、春季(4 月)、夏季(7 月)、秋季(10 月)4 个季节的 OST,图 8 展示了 26 个深度上的 RMSE 和 R^2 的变化规律。

如图 8 所示,模型在四个季节的 RMSE 值大约在 5 ～ 800 m 深度不断增大,R^2 不断减小;而 RMSE 值在 800～2 000 m 深度先减小再逐渐增大,R^2 值先增大又逐渐减小,这些变化特征可能是由于印度洋上层海域存在不稳定的海洋现象,而下层海域复杂的动力过程难以被海表面参数准确表达[29],造成模型估算精度下降。模型在不同季节反演性能也存在一些差异,总体而言,模型在夏季(7 月)的预测结果准确性最高,平均 RMSE 值为 0.216 0 ℃,平均 R^2 值为 0.941 0;春季(4 月)和秋季(10 月)的预测结果准确性次之,平均 RMSE 值分别为 0.223 0 和 0.232 0 ℃,平均 R^2 值分别为 0.940 0 和 0.935 0;冬季(1 月)的预测结果准确性最低,平均 RMSE 值为 0.237 0 ℃,平均 R^2 值为 0.934 0。但 4 个季节的估算结果整体上趋势较为一致,估算精度也较高,这说明该模型可较好的捕捉到季节变化对 OST 的影响。

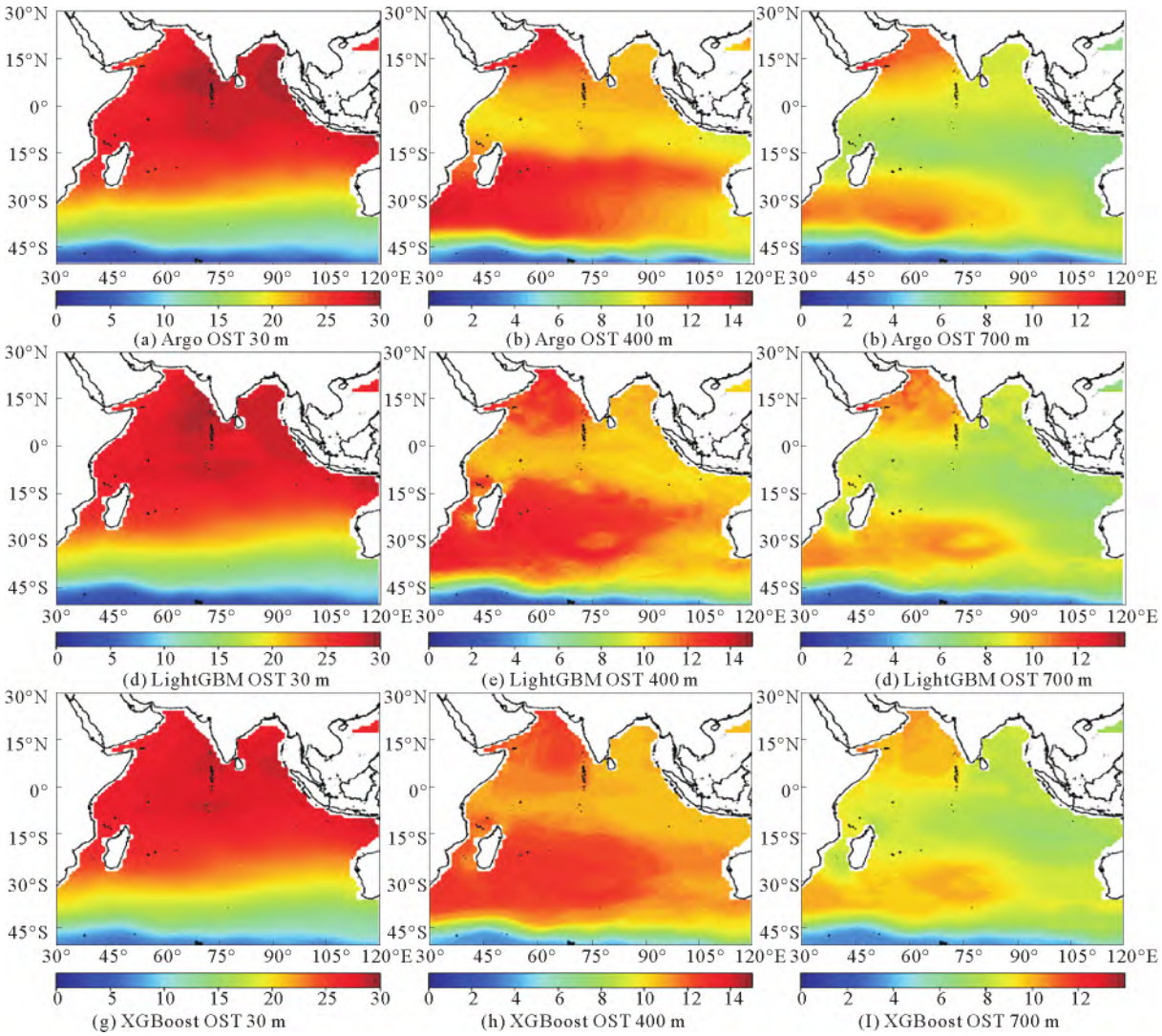

图6 30、400和700 m深度Argo观测的OST和2个模型估算的OST分布

Fig.6 Spatial distributions of the OST from Argo-observed and two models-estimated at depths of 30, 400 and 700 m

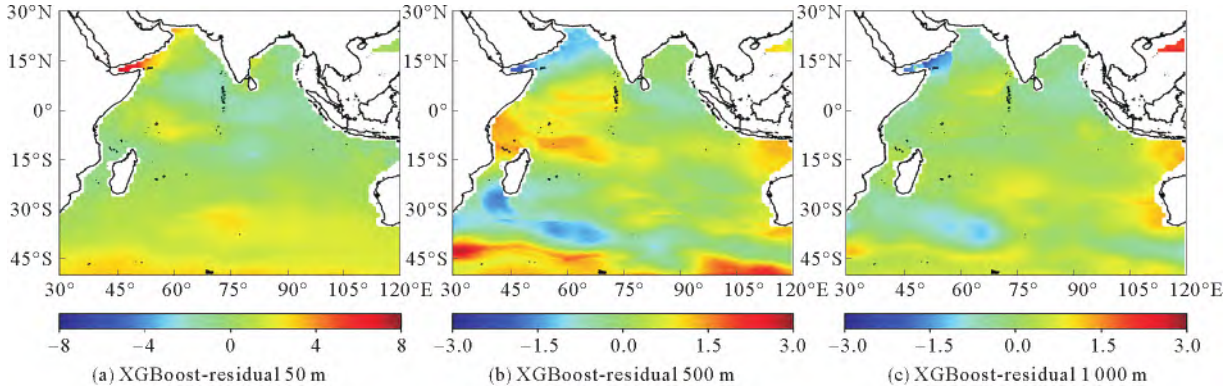

图7 在50、500和1 000 m深度XGBoost模型估算和Argo观测的OST的残差分布

Fig.7 Spatial distributions of the difference between XGBoost model-estimated and Argo-observed OST at depths of 50, 500 and 1 000 m

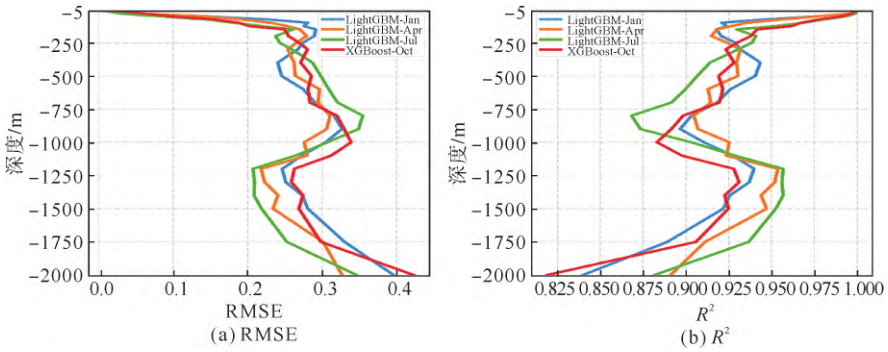

图 8　不同季节的 26 个深度下 LightGBM 模型的 RMSE 和 R^2 值

Fig.8　RMSE and R^2 values for the LightGBM model at 26 depths in different seasons

4　结　语

本研究面向印度洋海域，结合多源海表数据（SST、SSH、SSS、USSW 和 VSSW），提出了一种新的融合 GMM 聚类和 LightGBM 算法的 OST 反演模型，该模型采用 GMM 聚类算法将印度洋海域聚类成属性不同的 4 种海域，针对不同聚类海域，构建基于 LightGBM 算法的 OST 反演模型，估算出印度洋海域次表层（5～2 000 m）的温度分布，并利用 Argo 观测数据，通过 RMSE 和 R^2 来评估模型反演效果。结果表明，本模型能够较好反演出印度洋海域 OST 的空间分布。

为定量分析不同输入参数组合对反演模型估算 OST 的影响，设计了 3 种不同海表参数输入组合的对比实验。结果表明，当输入海表参数分别为 5 个（SST、SSS、SSH、USSW 和 VSSW），3 个（SST、SSS 和 SSH）和 2 个（SST 和 SSH）时，该模型 26 个深度上的平均 RMSE 值分别为 0.387 2、0.231 1 和 0.197 ℃，平均 R^2 值分别为 0.800 3、0.934 3 和 0.953 1，即 5 个输入参数（SST、SSS、SSH、USSW 和 VSSW）的 LightGBM 模型反演效果最好，3 个输入参数（SST、SSS 和 SSH）和 2 个输入参数（SST 和 SSH）的 LightGBM 模型次之。结果表明，所选择的海表参数对该模型都有积极影响，5 个输入参数的 LightGBM 模型有助于提高 OST 的估算精度。通过与 XGBoost 模型的对比发现，LightGBM 模型具有更好的反演性能，该模型可以更准确捕捉到印度洋次深层的温度分布特征。

最后，本工作基于构建的反演模型研究了季节因素对印度洋 OST 的影响，结果表明，该模型可以较好的捕捉到季节变化对 OST 的影响，反演效果较好，4 个季节的 RMSE 和 R^2 变化趋势较为一致，这种趋势可能与印度洋次表层的海洋动力环境变化特征有关。相对而言，该模型的估算精度在不同季节有所差异，夏季（7 月）OST 的估算精度最高，春季（4 月）和秋季（10 月）次之，冬季（1 月）最低，这说明该模型估算 OST 的能力一定程度上受季节因素的影响。

综上所述，本研究所提出的基于人工智能技术的 OST 反演模型，能够准确估算出印度洋海域的 OST 分布。该模型丰富了人工智能技术在深海遥感技术方面的应用，可为研究局部海域或全球海洋的内部动力环境、重构海洋次表层热结构以及研究全球气候变化等科学问题提供技术支持。然而，本研究模型仅考虑了 SST、SSH、SSS 和 SSW 等海表参数，输入参数有限，同时机理分析不足，估算精度还有待提高，未来可在这些方面开展进一步的探讨。

参　考　文　献

[1] MEEHL G A, ARBLASTER J M, FASULLO J T, et al. Model-based evidence of deep-ocean heat uptake during surface-temperature hiatus periods[J]. Nature Climate Change, 2011, 1(7): 360-364.

[2] BALMASEDA M A, TRENBERTH K E, KÄLLÉN E. Distinctive climate signals in reanalysis of global ocean heat content[J]. Geophysical Research Letters, 2013, 40(9): 1754-1759.

[3] 丁一汇. 2014—2016 年超强 El Niño 事件的发生发展过程与机理分析[J]. 大气科学学报, 2016, 39(6): 722-734.
DING Yihui. Analysis of the process and mechanisms of genesis and development for 2014—2016 mega El Niño event[J]. Transactions of Atmospheric Sciences, 2016, 39(6): 722-734.

[4] CHENG L J, ZHU J, ABRAHAM J, et al. 2018 Continues record global ocean warming[J]. Advances in Atmospheric Sciences, 2019, 36(3): 249-252.

[5] ALI M M, SWAIN D, WELLER R A. Estimation of ocean subsurface thermal structure from surface parameters: A neural

network approach[J]. Geophysical Research Letters, 2004, 31(20): L20308.

[6] KLEMAS V. Subsurface and deeper ocean remote sensing from satellites: An overview and new results[J]. Progress in Oceanography, 2014, 122: 1-9.

[7] 闫恒乾, 洪梅, 张韧, 等. 海洋表层-次表层反演与重构方法概述[J]. 海洋信息, 2016(3):1-8.
YAN Hengqian, HONG Mei, ZHANG Ren, et al. Overview of ocean surface subsurface inversion and reconstruction methods [J]. Marine Information, 2016(3): 1-8.

[8] SU H, WU X, YAN X H, et al. Estimation of subsurface temperature anomaly in the Indian Ocean during recent global surface warming hiatus from satellite measurements: A support vector machine approach[J]. Remote Sensing of Environment, 2015, 160: 63-71.

[9] SCHOTT F A, XIE S P, MCCREARY J P Jr. Indian Ocean circulation and climate variability[J]. Reviews of Geophysics, 2009, 47(1):1002.

[10] YANG J, LIU Q, XIE, S P, et al. Impact of the Indian Ocean SST basin mode on the Asian summer monsoon[J], Geophysical Research Letters, 2007, 34(2): L02708.

[11] LUO J J, SASAKI W, MASUMOTO Y. Indian Ocean warming modulates Pacific climate change[J]. Proceedings of the National Academy of Sciences of the United States of America, 2012, 109(46): 18701-18706.

[12] KHEDOURI E, SZCZECHOWSKI C, CHENEY R E. Potential oceanographic applications of satellite altimetry for inferring subsurface thermal structure[C]//Proceedings OCEANS'83, San Francisco, CA, USA: IEEE, 1983: 274-280.

[13] FISCHER M. Multivariate projection of ocean surface data onto subsurface sections[J]. Geophysical Research Letters, 2000, 27(6): 755-758.

[14] MAES C, BEHRINGER D, REYNOLDS R W, et al. Retrospective analysis of the salinity variability in the western tropical Pacific Ocean using an indirect minimization approach[J]. Journal of Atmospheric and Oceanic Technology, 2000, 17(4): 512-524.

[15] JEONG Y, HWANG J, PARK J, et al. Reconstructed 3-D ocean temperature derived from remotely sensed sea surface measurements for mixed layer depth analysis[J]. Remote Sensing, 2019, 11(24): 3018.

[16] LI X F, LIU B, ZHENG G, et al. Deep learning-based information mining from ocean remote sensing imagery[J]. National Science Review, 2020, 7(10): 1584-1605.

[17] WANG H, SONG T, ZHU S, et al. Subsurface temperature estimation from sea surface data using neural network models in the western Pacific Ocean[J]. Mathematics, 2021, 9(8): 852.

[18] WU X, YAN X H, JO Y H, et al. Estimation of subsurface temperature anomaly in the north Atlantic using a self-organizing map neural network[J]. Journal of Atmospheric and Oceanic Technology, 2012, 29(11): 1675-1688.

[19] SU H, LI W, YAN X. Retrieving temperature anomaly in the global subsurface and deeper ocean from satellite observations [J]. Journal of Geophysical Research: Oceans, 2018, 123(1): 399-410.

[20] SU H, HUANG L, LI W, et al. Retrieving ocean subsurface temperature using a satellite-based geographically weighted regression model[J]. Journal of Geophysical Research: Oceans, 2018, 123: 5180-5193.

[21] SU H, YANG X, LU W, et al. Estimating subsurface thermohaline structure of the global ocean using surface remote sensing observations[J]. Remote Sensing, 2019, 11(13): 1598.

[22] ZHANG K, GENG X, YAN X. Prediction of 3-D Ocean Temperature by Multilayer Convolutional LSTM[J]. IEEE Geoscience and Remote Sensing Letters, 2020, 17(8):1303-1307.

[23] SU H, WANG A, ZHANG T, et al. Super-resolution of subsurface temperature field from remote sensing observations based on machine learning[J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 102: 102440.

[24] SU H, ZHANG T, LIN M, et al. Predicting subsurface thermohaline structure from remote sensing data based on long short-term memory neural networks[J]. Remote Sensing of Environment, 2021, 260(8): 112465.

[25] LU W, SU H, YANG X, et al. Subsurface temperature estimation from remote sensing data using a clustering-neural network method[J]. Remote Sensing of Environment, 2019, 229: 213-222.

[26] 李大虎, 江全元, 曹一家. 基于聚类的支持向量回归模型在电力系统暂态稳定预测中的应用[J]. 电工技术学报, 2006, 21(7): 75-80.
LI Dahu, JIANG Quanyuan, CAO Yijia. Clustering based on support vector regression model and its application in power system transient stability prediction[J]. Transactions of China Electrotechnical Society, 2006, 21(7): 75-80.

[27] LI K, MA Z, ROBINSON D, et al. Identification of typical building daily electricity usage profiles using Gaussian mixture model-based clustering and hierarchical clustering[J]. Applied Energy, 2018(231): 331-342.

[28] KE G, QI M, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS'17), Curran Associates Inc, Red Hook, NY, USA: Neural Information Processing Systems Foundation, 2017: 3149-3157.

[29] 黎文娥,苏华,汪小钦, 等. 多源卫星观测的全球海洋次表层温度异常信息提取[J]. 遥感学报, 2017, 21(6): 881-891.
LI Wene, SU Hua, WANG Xiaoqin, et al. Estimation of global subsurface temperature anomaly based on multisource satellite observations[J]. Journal of Remote Sensing, 2017, 21(6): 881-891.

(责任编辑 姜丰辉)