DOI:10.11737/j.issn.1003-0239.2023.01.005

基于Stacking机器学习模型的南海北部海温预报

孙昭[1,2]，李云[1]，江毓武[2]，王兆毅[1]

（1.国家海洋环境预报中心 自然资源部海洋灾害预报技术重点实验室，北京 100081；2.厦门大学海洋与地球学院，福建 厦门 361102）

摘 要： 基于Stacking（ET-ET）的机器学习算法，利用美国国家环境预报中心再分析数据和MGDSST海温融合数据，建立了一套高效的海温长期预报方法，并在南海北部海域开展了1 a的表层海温长期预报实验。结果表明：基于Stacking（ET-ET）机器学习模型的表层海温长期预报的均方根误差降至0.52 ℃，平均绝对百分比误差降至1.58%，明显优于基于支持向量机、人工神经网络和长短期记忆模型的预报结果。

关键词： 机器学习；Stacking；南海北部；海温预报

中图分类号：P731.31 **文献标识码：**A **文章编号：**1003-0239（2023）01-0039-07

1 引言

海水温度是海洋环境的主要影响因子之一，开展海温预报对于海洋防灾减灾、气候变化研究、滨海旅游、海水养殖和渔业捕捞、海洋资源开发和保护以及我国国防建设都具有重要意义[1-2]。海温预报方法主要包括经验外推方法、统计方法、数值预报方法和大数据预报方法等，目前数值预报方法的应用最为普遍，但其也存在计算量大、预报时效短、预报结果高度依赖于初始条件和边界条件等问题。前人运用数值模式、观测和资料同化等手段，对各种海洋要素和海洋现象进行分析和预测[3-6]。近年来，将机器学习用于海洋预报的方法逐渐体现出其独特的优势。ZHANG等[7]采用长短期记忆神经网络模型预测了海表温度，该模型在中国沿海的海温数据集中应用效果良好。郝日栩等[8]提出基于改进经验正交函数和非线性自回归（Empirical Orthogonal Function-Nonlinear Auto Regressive，EOF-NAR）的神经网络混合模型并进行中长期海温时空预报。2019年，国家海洋环境预报中心开发的海温智能网格预报产品开始业务化运行，产品可以提供未来7 d西北太平洋10 km分辨率的逐小时海表面温度（Sea Surface Temperature，SST）预报。2020年国家海洋环境预报中心利用人工智能订正方法对有观测数据的站点开展释用订正，提高了县级单元海温预报的准确率[9]。将人工神经网络用于海浪数值预报和潮汐数值预报中，可以明显改进数值预报的精度[10-12]。以历史数据为核心的大数据智能预报方法用于推断未来海温，可以克服数据计算量大的缺点，能够更高效地进行海温预测。Stacking是机器学习回归问题的有效方法之一，它的主要思想是使用上一层机器学习模型的预测作为下一层模型的输入变量，用以提高模型的预测能力[13]。本文对极端随机树（Extra Tree，ET）模型进行Stacking算法堆叠，得到Stacking（ET-ET）机器学习算法。利用南海北部的MGDSST（Merged Satellite and In-situ Data Global Daily Sea Surface Temperature）海温融合数据和美国环境预报中心（National Centers for Environmental Prediction, NCEP）的NCEP / DOE AMIP-II（Reanalysis-2）再分析气象数据，通过分析数据的相关性，建立一套高效的南海北部表层海温长期预报方法。

收稿日期： 2021-09-22；**修回日期：** 2021-12-09。
基金项目： 国家重点研发计划（2022YFC3105102）。
作者简介： 孙昭（1997-），男，博士在读，主要从事深度学习在海洋预报中的应用工作。E-mail：22320191151067@stu.xmu.edu.cn

2 研究区域与数据

2.1 研究区域

南海位于中国大陆南部，天然海域面积350万平方公里，是中国三大边缘海之一。南海的石油、天然气、矿产和港口资源丰富，季风、台风等海气相互作用强烈，海洋灾害频发，因此开展长时效的南海海温变化预测对科学研究、气候预测、海上活动安全保障等具有重要意义。本文的研究区域为南海北部海域，空间范围为17°～21°N，113°～120°E（见图1蓝色方框）。本文对该区域进行基于多种机器学习方法的表层海温预报研究和对比实验，以期获取最优参数和算法。

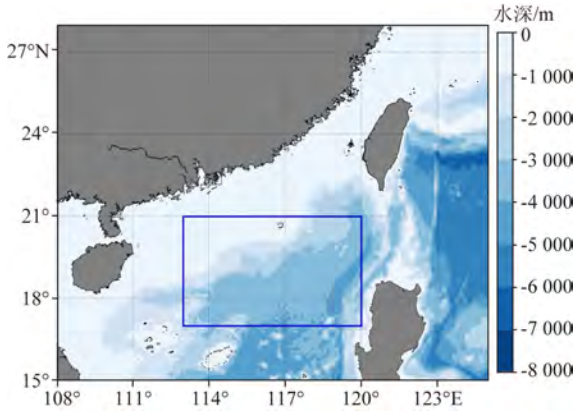

图1 研究区域（蓝色方框）

Fig.1 Study domain (blue box)

2.2 气象数据

本研究的海温预报模型使用Reanalysis-2气象数据。它是NCEP将来源于地面、船舶、无线电探空、探空气球、飞机和卫星等气象观测资料进行同化处理后研制的全球气象资料数据库。数据库空间分辨率为2.5°，时间为1979年1月—2021年8月[14]。本文选取1990—2011年每日4次预报的2 m处气温（air）、向下长波辐射通量（dlwrf）、表面向下平均短波辐射通量（dswrf）、表层平均潜热净通量（lhtfl）、表层气压（pres）、表面平均感热净通量（shtfl）、2 m处比湿（shum）、表面向上平均长波辐射量（ulwrf）、表面向上平均短波辐射量（uswrf）、10 m U方向风速（uwnd）和10 m V方向风速（vwnd）。

2.3 海温数据

本文使用日本气象厅（Japan Meteorological Agency，JMA）的MGDSST数据，数据空间分辨率为1/4°，时间范围为1982年至今。MGDSST数据根据卫星红外传感器（NOAA / AVHRR、MetOp / AVHRR）、微波传感器（Coriolis /WINDSAT、GCOM-W1/AMSR-2）和原位SST（来自浮标和船舶）进行分析[15]。本文将其作为训练中的标签值。

3 研究方法与数据处理

3.1 研究方法

ET模型最早是由GEURTS等[16]构造的随机森林算法（Random Forest，RF）的扩展版本。ET模型与RF非常相似，是利用集成学习思想将多棵决策树集合成一种算法。RF使用随机选择的子集进行训练，而ET则训练使用所有的样本，但是特征因子是随机选择的。RF可在一个随机子集中得到最好的分裂效果，而ET模型完全随机分裂，因此在一定程度上避免了过拟合。

Stacking模型的概念最早由WOLPERT[17]提出。它是一种将不同预测因子形成线性组合以提高预测精度的方法，其思想是利用交叉验证数据和非负性约束下的最小二乘来确定组合中的系数。Stacking模型是一种集合方法，它将多个机器学习算法提供的预测值作为新的训练集，再使用一个新的模型进行最终预测[17-18]。Stacking模型的具体运行结构见图2。在第一层模型中两次使用ET模型，第二层模型中选择线性模型进行最终回归和预测，即为本文提出的Stacking（ET-ET）机器学习算法。

3.2 数据处理

本文对气象数据进行日平均处理，并采用双线性插值法将研究区域内的SST数据和气象数据插值到分辨率为1°×1°的网格中，以保证两种数据具有相同的分辨率，最终得到40个网格的数据。对于机器学习算法来说，学习样本的质量是训练模型的关键之一。本文采用Pearson相关分析来确定各个气象特征（air、dlwrf、dswrf、lhtfl、pres、shtfl、shum、ulwrf、uswrf、uwnd和vwnd）对SST变化的敏感性。

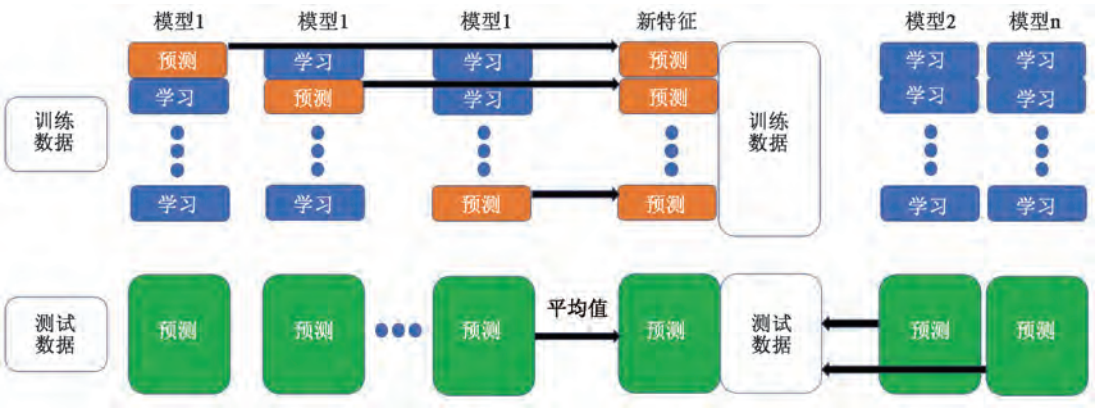

图2　Stacking模型结构

Fig.2　Stacking model structure

Pearson相关系数(r)的定义如下:

$$r=\frac{\sum_{i=1}^{n}(y_i-\bar{y}_i)(y_i'-\bar{y}_i')}{\sqrt{\sum_{i=1}^{n}(y_i-\bar{y}_i)^2}\sqrt{\sum_{i=1}^{n}(y_i'-\bar{y}_i')^2}} \tag{1}$$

式中:n是样本总数;y_i代表第一种特征的值,y_i'代表第二种特征的值,$\bar{y}_i$和$\bar{y}_i'$分别代表第一种特征和第二种特征的平均值。r越高,代表气象特征与SST的相关性越显著。使用1990—2011年的数据进行Pearson相关性分析,结果见图3。

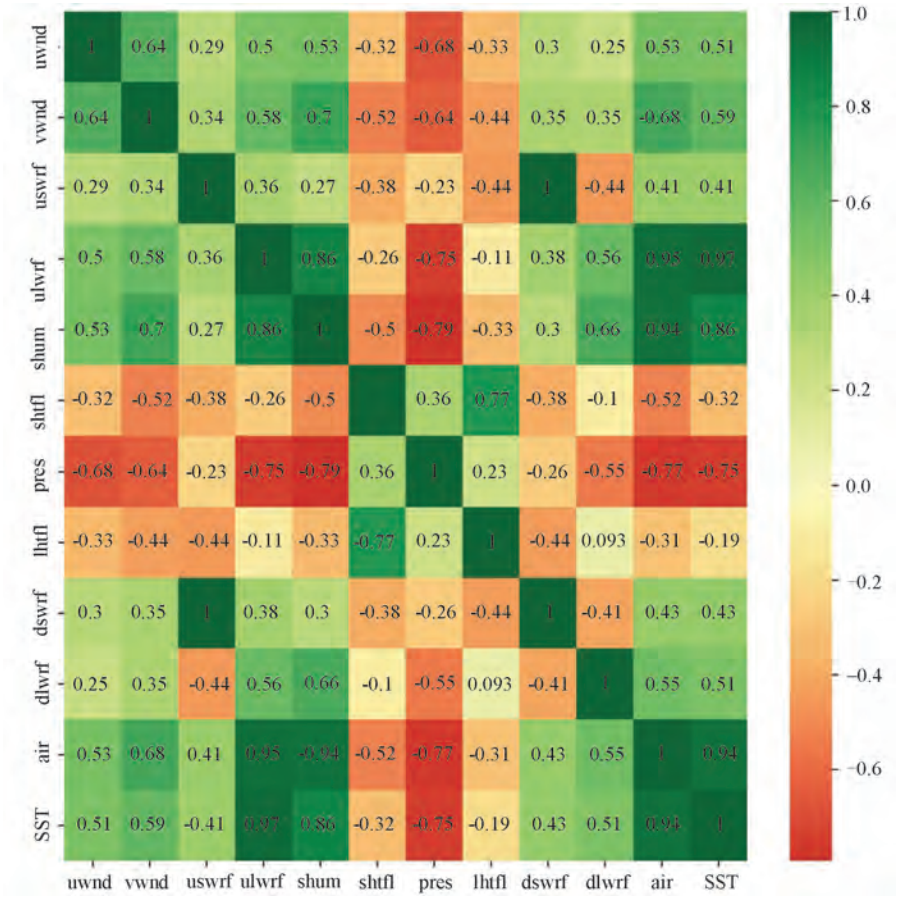

图3　SST与其他气象特征的相关性分析

Fig.3　Correlation analysis between SST and other meteorological characteristics

从图3可以看出，shtfl、pres和lhtfl与SST呈负相关，其他特征与SST呈正相关。另外，uwnd、vwnd、ulwrf、shum、pres、air和dlwrf 7个特征与SST有较强的相关性，它们与SST的r的绝对值均超过0.50，因此，本文将这7个特征作为SST预测的输入变量。

4 模型建立

对特征数据进行预处理后，需要对模型进行超参数优化。机器学习模型一般都有很多超参数，原则上选取最重要的超参数进行组合调整。调整的超参数为决策树个数、最大特征数量和最大深度（见表1）。在定义的参数范围内使用网格搜索方法对超参数进行优化。将17°N，113°E空间点的气象数据和SST数据分为两组，以数据集的90%作为训练集、10%作为测试集进行参数调整。将训练数据输入模型进行训练，然后用测试集对模型进行评估。为了计算不同模型预测的SST值与标签值的误差与拟合度，本文从统计学角度利用均方根误差（Root Mean Square Error，RMSE）、平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）和拟合优度（R^2）对各个模型的预测性能进行评价。

如果将效果较差的模型加入Stacking模型进行算法堆叠，可能会降低模型的预测效果。因此，将ET与RF、梯度提升迭代决策树（Gradient Boosting Decision Tree，GBDT）、AdaBoost（Adaptive Boosting）算法、XGBoost（Extreme Gradient Boosting）算法和LightGBM（Light Gradient Boosting Machine）算法这5种常见的机器学习模型进行比较。以1990—2010年的数据作为训练集，2011年的数据作为测试集，比较6种机器学习算法的预测效果以及每种算法的运行时间和效率。

表2为6种基于决策树模型的机器学习算法在研究区域所有网格的评价指标的平均值，可以用来评价算法在整个研究区域的预测效果。从表中可以看出，ET模型的RMSE和MAPE最小，分别为0.526 9和1.596 9%，且R^2最大，为0.936 2，这说明ET算法的预测误差最小，拟合度最高，对整个研究区域的预测效果最好。RF模型和LightGBM模型在预测效果上与ET模型相差不大，LightGBM模型由于自身的特点，在6种模型中的训练时间最短。结果表明ET模型、RF模型和LightGBM模型是相对有效的3种算法。将5种Stacking模型进行比较（见表3），可以看出，单个模型的预测效果较差，会降低Stacking模型的整体预测效果。将ET模型使用两次，第二层模型选择简单的线性回归，得到的预测效果最好，训练时间最短。如果将强学习模型RF和ET作为第二层模型，会因数据过拟合导致预测效果变差。因此本文使用预测效果最好的Stacking（ET-ET）模型与其他模型进行比较。

表1 极端随机树模型的超参数调整

Tab.1 Hyper-parameters adjustment of Extra Tree model

超参数	调整范围	调整结果	RMSE	MAPE	R^2
决策树个数	[50，100，200，300]	[200]			
最大特征数量	[2，3，4，5]	[3]	0.446 8	1.246 6	0.956 8
最大深度	[5，10，15，20]	[15]			

表2 6种方法在研究区域中所有网格的RMSE、MAPE和R^2的平均值

Tab.2 The average of RMSE, MAPE and R^2 of all grids in the study domain by the six methods

模型	RMSE	MAPE/%	R^2	训练时间/s
AdaBoost	0.564 1	1.704 3	0.928 8	95
ET	0.526 9	1.596 9	0.936 2	38
GBDT	0.531 7	1.612 6	0.935 1	57
LightGBM	0.530 9	1.612 0	0.935 1	2
RF	0.528 5	1.602 4	0.935 8	74
XGBoost	0.541 2	1.642 4	0.931 7	8

表3　5种Stacking模型在所有空间点上预测的平均值

Tab.3　Average value forecasted by five stacking models at all spatial points

模型	第一层模型	第二层模型	RMSE	MAPE/%	R^2	训练时间/s
1	ET-RF-LightGBM	线性回归	0.537 3	1.622 1	0.934 5	171
2	ET-RF	线性回归	0.526 1	1.591 5	0.936 3	137
3	ET-ET	线性回归	0.522 7	1.581 7	0.937 2	81
4	ET-ET	ET	0.588 8	1.801 0	0.906 5	161
5	ET-ET	RF	0.612 3	1.911 5	0.907 7	181

5　实验结果

图4为Stacking(ET-ET)模型与支持向量机(Support Vector Machine,SVR)模型、人工神经网络(Artificial Neural Network,ANN)模型和长短期记忆(Long Short-Term Memory, LSTM)模型的RMSE、

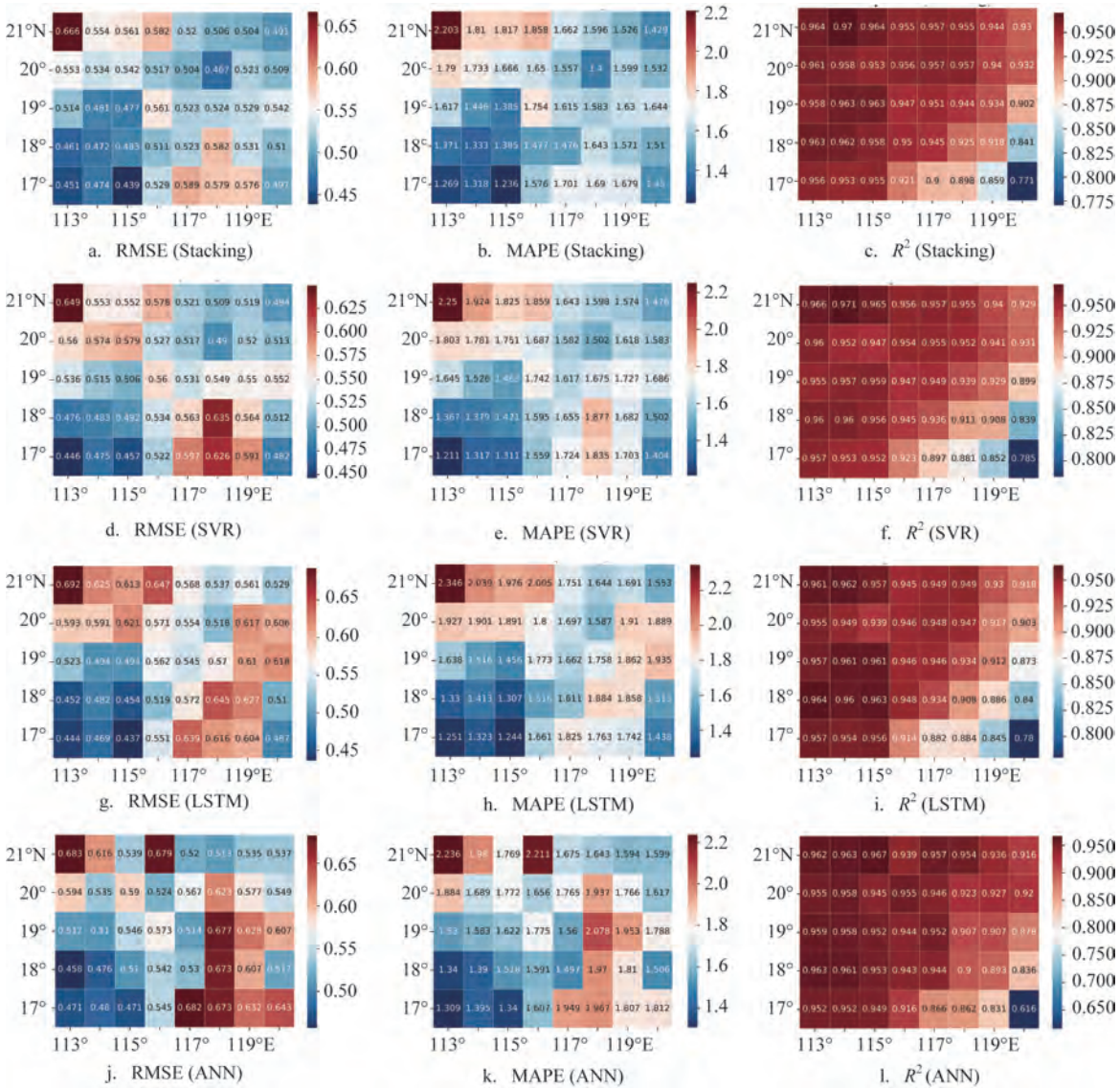

图4　利用Stacking(ET-ET)、SVR、ANN和LSTM方法计算研究区域RMSE、MAPE(单位:%)和R^2的空间分布。

Fig.4　The spatial distribution of RMSE, MAPE (unit:%)and R^2 in the study domain use Stacking(ET-ET), SVR, ANN and LSTM methods

MAPE和R^2的空间分布。从图中可以看出,Stacking(ET-ET)模型的空间分布与其他3种模型的RMSE、MAPE和R^2大致相似。17°N,120°E是拟合效果最差的空间点,21°N,113°E是预测误差最大的空间点。ANN和LSTM模型的误差分布与Stacking(ET-ET)模型结果略有不同。在研究区域东南部,Stacking(ET-ET)模型的RMSE和MAPE明显小于其他3种模型结果。Stacking(ET-ET)模型在大部分位置的RMSE和MAPE分别小于0.6和2%。21°N,113°E空间点误差较大可能与其位于近海有关,SST会受到更多环境因素的影响。

如表4所示,由研究区域中所有空间点的误差和拟合程度的对比分析可知,在4种机器学习算法中,Stacking(ET-ET)模型的RMSE和MAPE最小,其他3种模型的RMSE均大于0.53。Stacking(ET-ET)模型的拟合程度最好,并且训练时间也相对较小。由此可以看出,在长期预测中,Stacking(ET-ET)模型有较大的应用前景。

表4 Stacking(ET-ET)、SVR、ANN和LSTM模型预测结果对比

Tab.4 Comparison of forecast results between Stacking (ET-ET) model and SVR, ANN and LSTM models

模型	RMSE	MAPE/%	R^2	训练时间/s
Stacking(ET-ET)	0.522 7	1.581 7	0.937 2	81
SVR	0.535 3	1.626 9	0.934 4	103
ANN	0.566 4	1.712 2	0.922 9	111
LSTM	0.559 5	1.697 2	0.928 2	866

6 结论与讨论

本文基于南海北部1990—2011年的NCEP/DOE AMIP-II Reanalysis再分析气象数据以及MGDSST全球每日数据,利用Stacking(ET-ET)机器学习算法,实现了对南海北部1 a的长期预测。结论如下:

(1)本文构建的Stacking(ET-ET)机器学习模型在南海北部海温长期预报应用上具有较好的效果,1 a预报的RMSE和MAPE分别约为0.52 ℃和1.58%。相较于SVR、ANN和LSTM 3种常见的机器学习算法,Stacking(ET-ET)方法在控制误差和训练时间方面都有明显的优势,具有较好的应用前景。

(2)在利用气象特征对SST进行回归预测时,大部分研究区域的预测效果较好,这主要是因为机器学习算法输入的气象要素与SST具有较好的相关性。但距离陆地较近时预测误差较大,这可能是因为越靠近陆地,影响因素越多,SST的变化越复杂,预测越难。因此,在以后的研究中会尝试加入更多的特征因子,以期改进预测效果。

参考文献:

[1] 张建华. 海温预报知识讲座 第一讲 海水温度预报概况[J]. 海洋预报, 2003, 20(4): 81-85.
ZHANG H J. Lecture on knowledge of sea temperature prediction, Lecture 1, Overview of sea water temperature prediction[J]. Marine Forecasts, 2003, 20(4): 81-85.

[2] 吴磊, 王彬, 潘锡山, 等. 融合海表温度产品在渤黄东海的对比分析及初步验证[J]. 海洋通报, 2020, 39(6): 657-668.
WU L, WANG B, PAN X S, et al. Intercomparison analysis of merged sea surface temperature products for the Bohai, Yellow and East China Seas[J]. Marine Science Bulletin, 2020, 39(6): 657-668.

[3] 吴新荣, 王喜冬, 李威, 等. 海洋数据同化与数据融合技术应用综述[J]. 海洋技术学报, 2015, 34(3): 97-103.
WU X R, WANG X D, LI W, et al. Review of the application of ocean data assimilation and data fusion techniques[J]. Journal of Ocean Technology, 2015, 34(3): 97-103.

[4] 王辉, 万莉颖, 秦英豪, 等. 中国全球业务化海洋学预报系统的发展和应用[J]. 地球科学进展, 2016, 31(10): 1090-1104.
WANG H, WAN L Y, QIN Y H, et al. Review of the application of ocean data assimilation and data fusion techniques[J]. Advances in Earth Science, 2016, 31(10): 1090-1104.

[5] 刘娜, 王辉, 凌铁军, 等. 全球业务化海洋预报进展与展望[J]. 地球科学进展, 2018, 33(2): 131-140.
LIU N, WANG H, LING T J, et al. Review and prospect of global operational ocean forecasting[J]. Advances in Earth Science, 2018, 33(2): 131-140.

[6] 张培军, 周水华, 梁昌霞. 基于卫星遥感海温数据的南海SST预报误差订正[J]. 热带海洋学报, 2020, 39(6): 57-65.
ZHANG P J, ZHOU S H, LIANG C X. Study on the correction of SST prediction in South China Sea using remotely sensed SST[J]. Journal of Tropical Oceanography, 2020, 39(6): 57-65.

[7] ZHANG Q, WANG H, DONG J Y, et al. Prediction of sea surface temperature using long short-term memory[J]. IEEE Geoscience and Remote Sensing Letters, 2017, 14(10): 1745-1749.

[8] 郝日栩, 赵玉新, 何忠杰, 等. 基于EOF-NAR神经网络混合模型的海温预报方法研究[C]//中国海洋学会2019海洋学术(国际)双年会. 三亚, 2019: 31-45.
HAO R X, ZHAO Y X, HE Z J, et al. Study on sea surface temperature prediction method based on EOF-NAR neural network hybrid model[C]//2019 Marine Academic (International) Biennial

Meeting of China Oceanographic Society. Sanya, 2019: 31-45.

[9] 王兆毅, 李云, 王旭. 中国近岸海域基础预报单元海温预报指导产品研制[J]. 海洋预报, 2020, 37(4): 59-65.
WANG Z Y, LI Y, WANG X. Development of forecast guidance product for sea temperature of basic forecast units in the Chinese coastal waters[J]. Marine Forecasts, 2020, 37(4): 59-65.

[10] 陈希, 沙文钰, 李妍, 等. 人工神经网络技术在海浪预报中的应用[J]. 海洋通报, 2002, 21(2): 11-15.
CHEN X, SHA W Y, LI Y, et al. Application of the artificial neural network in the sea wave forecast[J]. Marine Science Bulletin, 2002, 21(2): 11-15.

[11] 齐义泉, 张志旭, 李志伟, 等. 人工神经网络在海浪数值预报中的应用[J]. 水科学进展, 2005, 16(1): 32-35.
QI Y Q, ZHANG Z X, LI Z W, et al. Application of artificial neural network to numerical wave prediction[J]. Advances in Water Science, 2005, 16(1): 32-35.

[12] 王建华, 于红兵, 宋运法. 人工神经网络在潮汐数值预报中的应用[J]. 海洋预报, 2007, 24(2): 47-51.
WANG J H, YU H B, SONG Y F. Application of artificial neural network to numerical tidal prediction[J]. Marine Forecasts, 2007, 24(2): 47-51.

[13] PAVLYSHENKO B. Using stacking approaches for machine learning models[C]//2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP). Lviv, Ukraine: IEEE, 2018: 255-258.

[14] KANAMITSU M, EBISUZAKI W, WOOLLEN J, et al. NCEP-DOE AMIP-II reanalysis (R-2) [J]. Bulletin of the American Meteorological Society, 2002, 83(11): 1631-1644.

[15] SAKURAI T, YUKIO K, KURAGANO T. Merged satellite and in-situ data global daily SST[C]//Proceedings.2005 IEEE International Geoscience and Remote Sensing Symposium. Seoul, Korea (South): IEEE, 2005: 2606-2608.

[16] GEURTS P, ERNST D, WEHENKEL L. Extremely randomized trees[J]. Machine Learning, 2006, 63(1): 3-42.

[17] WOLPERT D H. Stacked generalization[J]. Neural Networks, 1992, 5(2): 241-259.

[18] BREIMAN L. Stacked regressions[J]. Machine Learning, 1996, 24 (1): 49-64.

Sea temperature forecast in the northern South China Sea base on Stacking machine learning model

SUN Zhao[1,2], LI Yun[1], JIANG Yuwu[2], WANG Zhaoyi[1]

(1. Key Laboratory of Marine Hazards Forecasting, National Marine Environmental Forecasting Center, Ministry of Natural Resources, Beijing 100081, China; 2.College of Ocean and Earth Sciences,Xiamen University, Xiamen 361102, China)

Abstract: In this paper, an efficient long-term SST forecast method is established based on Stacking (ET-ET) machine learning algorithm using reanalysis data of National Centers for Environmental Prediction and Mergid satellite and in situ data Global Daily sea surface temperature (SST) fusion data, and long-term SST forecast experiment is carried out in the northern South China Sea for one year. The results show that the root mean square error of long-term SST forecast based on Stacking (ET-ET) machine learning model is reduced to 0.52 ℃, and the mean absolute percentage error is reduced to 1.58%, which is significantly better than the forecast results based on the support vector machine, artificial neural network and long short-term memory model.

Key words: machine learning; Stacking: northern South China Sea; sea temperature forecast