张巍, 杜超凡, 郭安博宇, 等. 一种机器学习海面风场快速融合的方法[J]. 海洋学报, 2022, 44(11): 144–158, doi:10.12284/hyxb2022137
Zhang Wei, Du Chaofan, Guo Anboyu, et al. Sea surface wind field smart fusion base on machine learning method[J]. Haiyang Xuebao, 2022, 44(11): 144–158, doi:10.12284/hyxb2022137

一种机器学习海面风场快速融合的方法

张巍[1,2]，杜超凡[2]，郭安博宇[1*]，宋晓姜[1]，沈世莹[2]

(1. 国家海洋环境预报中心, 北京 100081; 2. 中国海洋大学 计算机科学与技术学院, 山东 青岛 266100)

摘要: 基于多源资料进行海面风场的同化融合或插值融合, 目前受到计算能力的较大制约。本文提出在多源卫星数据和 ERA-5 再分析数据重叠区域, 训练基于 XGBoost 的机器学习 ERA-5 数据修正融合模型。然后基于该模型快速修正 ERA-5 数据（机器学习推理）。由于机器学习推理的快速性, ERA-5 全区域修正融合的时间仅需 2 s 左右, 可以较小计算代价构建整个海面融合风场。本文以 10 m 风速、10 m 风向、*U*10 分量和 *V*10 分量等典型风场变量展开, 考虑了海陆分布差异使用陆地掩膜消除陆地区域, 分别构建 D_S_A_XGBoost、D_S_O_XGBoost、U_V_A_XGBoost、U_V_O_XGBoost 4 个 ERA-5 修正模型, 并最终生成海面融合风场。通过修正前后的 ERA-5 再分析数据与卫星数据进行比较, 上述 4 个模型均减小了 ERA-5 再分析数据与卫星数据的差距。特别是在风速方面, 不论是均方根误差（RMSE）还是绝对误差（MAE）都得到有效降低。在风向方面上, $RMSE_d$ 以及 MAE_d 也呈现降低趋势。在利用热带大气海洋观测计划（Tropical Atmosphere Ocean Array, TAO）浮标数据对 4 种 XGBoost 模型进行评价发现, U_V_O_XGBoost 模型对于 ERA-5 数据的修正结果最好, 其相关性达到 0.893, 提高了约 0.011, 结果表明本文在保证风场精度的情况下较大地提高了融合速度。

关键词: XGBoost; HY-2B; CFOSAT; MetOp-B; ERA-5; 海面风场

中图分类号: P717; P732　　**文献标志码:** A　　**文章编号:** 0253−4193(2022)11−0144−15

1　引言

作为海洋学最重要的物理参数之一, 海面风场是海洋上层运动的主要动力来源, 几乎所有的海水运动都与之直接相关[1−7]。与此同时, 海面风场对于海洋渔业、海上交通及工程活动、风能开发等都有着直接的影响[8−9]。对于海面风场的测量, 其中常规的测量手段包括船舶、浮标以及沿岸站等。相对于全球海洋来说, 常规测量手段获取到的风场数据资料非常缺乏, 很难满足人类的生产或研究的需求。此时, 卫星遥感技术的出现很好地解决了常规测量手段所存在的问题。卫星遥感技术有着覆盖范围广, 空间分辨率高, 能够实时或准实时获取数据的优势[10−11]。但是单一卫星提供的海面风场产品在覆盖率等方面存在着不可避免的缺陷, 因此研究如何将多源卫星海面风场等产品进行融合, 以此提高海面风场数据的覆盖范围和精度, 从而满足当前数值预报研究以及海洋中小尺度系统研究的需求变得尤为重要。

当前有许多数据融合算法被研究者提出并利用。海面风场作为数据融合的应用领域, 目前主要的融合方法有插值类融合算法和同化变分类融合算法。其中插值算法有 Cressman 插值、Kriging 插值和时空加权分析方法等, 同化变分算法包括最优插值法、三维变分法等[12]。凌征等[13] 通过 Cressman 插值融

收稿日期: 2021−10−11; **修订日期:** 2022−05−15。
基金项目: 国家重点研发计划（2018YFC1407001）。
作者简介: 张巍（1975—）, 男, 北京市人, 副教授, 研究方向为海洋大气智能预报预警。E-mail: weizhang@ouc.edu.cn
*** 通信作者:** 郭安博宇, 工程师, 研究方向为海洋气象。E-mail: guoanboyu@nmefc.cn

合了我国近海的卫星风场和沿岸气象站风场资料。Zhang 等[14−15] 对包括 SSM/I、TMI、QuikSCAT、AMSR-E 等在内的多颗卫星海面风速数据进行了时空权重插值融合, 产生了全球范围 1987−2006 年的时间分辨率为 12 h、每天、每月的 0.25°网格的风速。齐亚琳和林明森[16] 对海洋二号卫星海面风场和 NCEP 数值风场资料进行融合, 融合算法中同样采用时空权重插值。Yan 等[17] 对多源散射计和辐射计风场与模式在分析风场进行了融合研究, 利用最优插值法建立了时间分辨率为 6 h, 空间分辨率为 0.25°的 2000−2015 年的全球风场产品。Chao 等[18] 基于二维变分分析的方法融合了卫星散射计海面风场与区域中尺度大气模式风场。

综上所述, 不论是插值类融合算法, 还是同化变分类融合算法, 它们都可以基本解决海面风场融合的问题。但是在实际应用中, 受到当前计算能力的制约[19]。这些算法由于计算过程复杂, 往往需要使用计算机集群, 且较难实现实时化融合。

为了以较低的计算代价实现实时化海面风场融合, 本文提出在多源卫星数据和 ERA-5 再分析数据重叠区域, 训练基于 XGBoost 的机器学习 ERA-5 数据修正模型。然后利用该模型在无卫星数据区域快速修正(机器学习推理)ERA-5 数据, 使得修正后得到的融合风场数据更加贴近卫星观测值, 最终得到时间分辨率为 12 h、每天的 0.25°的网格融合风场数据, 实现无缝网格风场[20]。其中最核心的修正过程是利用已经训练好的模型进行快速推理, 而由于机器学习推理的快速性, 可以减小计算代价, 构建整个海面融合风场。

2　海面风场相关数据集

本文使用的卫星有海洋二号 B(HY-2B)卫星、中法海洋卫星(CFOSAT)以及欧洲气象卫星 B(MetOp-B)卫星。3 颗卫星均可提供 2020 年 12 月以及 2021 年 1 月的海面风场资料。

HY-2B 卫星散射计 L2B 级数据存储经过风场反演和模糊去除处理后得到轨道各个风元的中心位置、风速、风向、观测时间及其他相关数据。HY-2B 卫星散射计每天约有 16 轨数据, 可覆盖全球 90% 的海域[21]。陈克海等[21] 使用 ECMWF 再分析风场数据、热带大气海洋观测计划(TAO)浮标和 NDBC 浮标实测数据对 HY-2B 风场进行了总体质量分析。分析发现, 在 4～24 m/s 风速区间内, HY-2B 卫星风速、风向均方根误差(RMSE)分别优于 2 m/s 和 20°, 能较好满足 HY-2B 卫星散射计业务化应用的精度要求。本文使用 2020 年 12 月以及 2021 年 1 月数据来进行实验, 选取的 HY-2B 卫星散射计 L2B 级数据的时间跨度为 12 h, 空间分辨率为 25 km×25 km, 且空间分布在 0°～45°N, 100°E～180°。

中法海洋卫星采用成熟的 CAST2000 小卫星平台, 设计寿命为 3 年, 运行于轨道高度为 521 km、降交点地方时 07:00 的太阳同步轨道, 探测数据分别传输至中法两国地面站, 由两国地面应用系统接收并进行处理。该卫星在海洋动力环境业务监测、海洋灾害监测和预报预警、海洋科学研究中发挥重要作用。本文同样使用 2020 年 12 月以及 2021 年 1 月数据来进行实验, 选取的 CFOSAT 卫星 L2B 级数据时间跨度为 12 h, 空间分辨率为 12.5 km×12.5 km, 且空间分布在 0°～45°N, 100°E～180°, 其风速精度为 1.5 m/s, 风向精度为 20°[22]。

2013 年 4 月 24 日, 欧洲航天局和欧洲气象卫星开发组织联合发射的 MetOp-B 代替 MetOp-A 作为主要的业务观测卫星, 其提供的海面风场数据产品风速精度为 2 m/s, 风速范围为 0～50 m/s。本文选取的 MetOp-B 风场数据空间分辨率为 12.5 km×12.5 km, 且空间分布在 0°～45°N, 100°E～180°。

ERA-5 是欧洲中期天气预报中心对过去 40～70 年全球气候和天气的第 5 代再分析数据。目前的数据是从 1950 年开始的, 分为 1950−1978 年的气候数据存储条目和 1979 年以后的。ERA-5 提供了大量大气、海浪和陆地表面数量的每小时估计数。本文选用的 ERA-5 再分析风场时间区间为 2020 年 12 月以及 2021 年 1 月, 其空间分辨率为 0.25°×0.25°, 其空间分布在 0°～45°N, 100°E～180°。

浮标数据选自离岸 50 km 以上, 具有连续风矢量观测能力的 TAO 浮标数据。该浮标具有较高的观测频率, 每 10 min 观测一次风速、风向。由于选定的 TAO 浮标上的测风计距离海面 4 m, 而散射计测量的是高度 10 m 处的风速, 因此需要将浮标观测风速转换到 10 m 高度上的风速, 转换公式为

$$s_{10} = 8.874\,03 \times s_z / \ln(z/0.001\,6)^{[21]}, \quad (1)$$

式中, z 表示距离海面的高度; s_{10} 和 s_z 分别表示 10 m 高度处的风速和在 z 高度上的风速。

3　融合方法

对于融合风场的生成, 研究共分为两部分进行,

即修正融合风场模型的训练及其机器推理。文中首先以卫星数据作为实况数据，通过 XGBoost 模型方法对 ERA-5 数据进行修正训练，得到修正融合风场模型，使得修正后的 ERA-5 数据更加接近于卫星数据分布，然后利用训练完毕的模型生成海面融合风场。

文中将混合的卫星数据统一处理成为 0.25°×0.25° 的标准网格数据。在插值处理过程中，由于卫星数据之间分辨率的不同，即 12.5 km×12.5 km 和 25 km×25 km 不等，为了方便统一插值，本文在空间上采用反距离加权插值算法，时间上采用最近邻方法对混合卫星数据进行插值，插值完成后的卫星数据与 ERA-5 数据共同完成修正融合风场模型的训练，并最终得到全区域的时间分辨率为 12 h 的 0.25°×0.25°的标准网格数据，具体融合流程如图 1 所示。

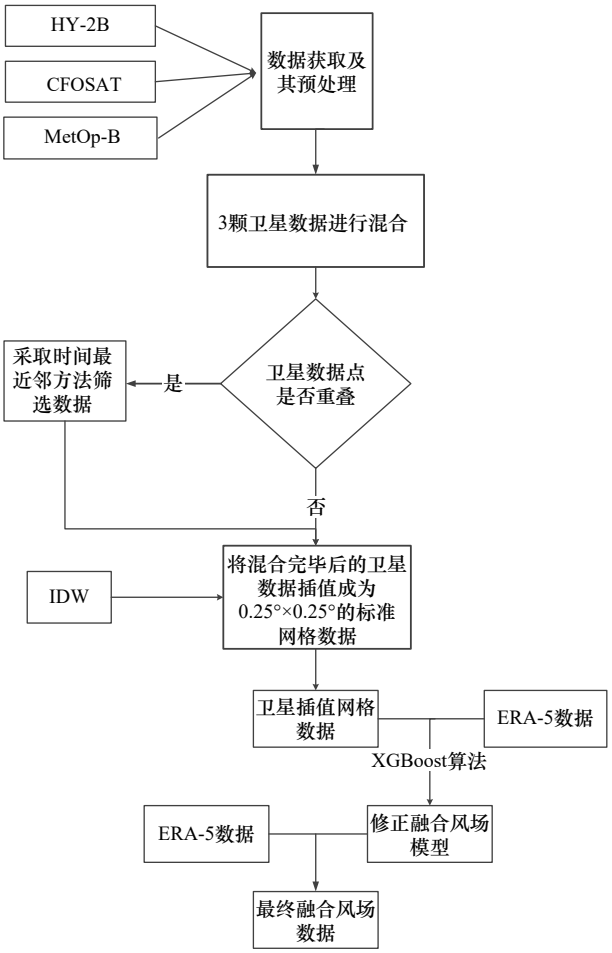

图 1 海面融合风场生成流程

Fig. 1 The process of obtaining the sea surface fusion wind

3.1 插值方法

3.1.1 反距离加权插值算法

当前对气象要素等进行插值的算法有很多[23−28]，本文选取的插值算法为反距离权重法（IDW）。IDW 插值是一种经常使用的空间插值方法，在 1972 年被美国国家气象局首次提出[29−31]。它的逻辑来源于地理学第一定律——相近相似原理。IDW是通过插值点与样本点之间距离的倒数为权重进行加权平均，与插值点越靠近的样本点计算时所被赋予的权重值越大，权重值一般与距离成反比关系，所以称之为“反距离”加权。其计算公式可以表示为

$$Z(X_0)=\frac{\sum_{i=1}^{n}Z(X_i)\times W_i}{\sum_{i=1}^{n}W_i}, \tag{2}$$

$$W_i=\frac{1}{\left[\sqrt{(x_0-x_i)^2+(y_0-y_i)^2}\right]^p}, \tag{3}$$

式中，$Z(X_0)$是待估计的X_0的属性值；$Z(X_i)$为X_0周围区域内的第i个点的属性值；W_i表示的是反距离权重；p表示的是权重的幂，默认选择p=2；(x_0, y_0)表示的是待估计点的坐标位置；(x_i, y_i)为待估计点周围第i点的坐标位置。

3.2 生成修正融合风场模型

正如引言所说，本文使用卫星数据对 ERA-5 数据进行修正融合，使得修正融合后的风场数据更加贴近真实值。研究流程如图 1 所示，首先对混合后的卫星数据进行插值操作，空间上使用反距离加权插值算法（IDW），时间上采用最近邻方法将其插值成为 0.25°×0.25°的标准网格数据。然后利用卫星插值数据和 ERA-5 数据获取训练样本后进行训练，最终得到所需的 XGBoost 模型，即修正融合风场模型。

3.2.1 修正融合方法

ERA-5 数据是全区域数据，其风场数据既涵盖了海洋区域，也包括了陆地区域。由于陆地风场和海洋风场的差异较大，详细分析请见 4.1 节。因此为了研究的科学性及其可靠性，本文采用 4 种方法来对 ERA-5 数据进行修正，具体方法如下：

方法 1：风速、风向修正（全区域）即 D_S_A_XGBoost 模型。在 XGBoost 训练的过程中，不区分海洋和陆地风场数据，全部用来进行模型的训练。

方法 2：U、V 修正（全区域）即 U_V_A_XGBoost 模型。与方法 1 相同，训练过程中不区分海洋和陆地风场数据。区别在于方法 1 中使用的训练数据为风速和风向，而方法 2 中使用的训练数据为 U10 和 V10，训练结束后再合成风速和风向。

方法 3：风速、风向修正（陆地掩码）即 D_S_O_XG-

Boost 模型。训练过程中区分海洋和陆地风场数据，即使用陆地掩码将陆地风场数据剔除，不参与模型的训练。

方法 4: U、V 修正(陆地掩码)即 U_V_O_XGBoost 模型。与方法 3 相同，训练过程中区分海洋和陆地风场数据。不同点在于方法 3 中使用的训练数据为风速和风向，方法 4 中使用的训练数据为 $U10$ 和 $V10$，训练结束后再合成风速和风向进行修正。

3.2.2　样本生成

噪声与偏差、方差共同构成机器学习的泛化误差[32]。噪声普遍存在，具有随机性和不可控性，例如数据采集仪器等带来的随机性偏差就是噪声的一种，本文中海陆交界处的无效数据可视为卫星观测的噪声。机器学习训练允许且需要数据中噪声的存在，由含噪声数据训练得到的模型通常更具有鲁棒性，能够更好地在未知分布数据上推理。

本文采用局部训练，全局推理的方式进行研究。即使用所能获取到的区域内样本数据进行训练，训练得到的模型可以应用于整片区域。本文对于训练样本的获取过程如图 2 所示。经过插值处理后的卫星数据与 ERA-5 数据均为 0.25°×0.25°的网格数据，本文使用卫星插值风场数据作为学习目标，选取卫星插值格点及其周围(5×5 窗口)的 ERA-5 值作为训练特征，进行训练。图 2a 绿色点表示的是卫星插值数据，周围 5×5 格点为 ERA-5 数据，当 ERA-5 数据在 5×5 空间格点中全部存在时，那么就会得到如图 2b 的训练样本，若 ERA-5 数据存在缺失，那么在该点就无法获取到训练样本。本文使用的训练样本为 2020 年 12 月 21 日至 2021 年 1 月 21 日数据，测试数据为 2021 年 1 月 31 日卫星初始数据以及修正前后的 ERA-5 数据。在研究过程中，本文针对 0 时和 12 时数据分别训练模型，即 0 时刻修正模型以及 12 时刻修正模型，其中训练过程中使用的训练集约 400 000，验证集约 40 000，测试集约 60 000。

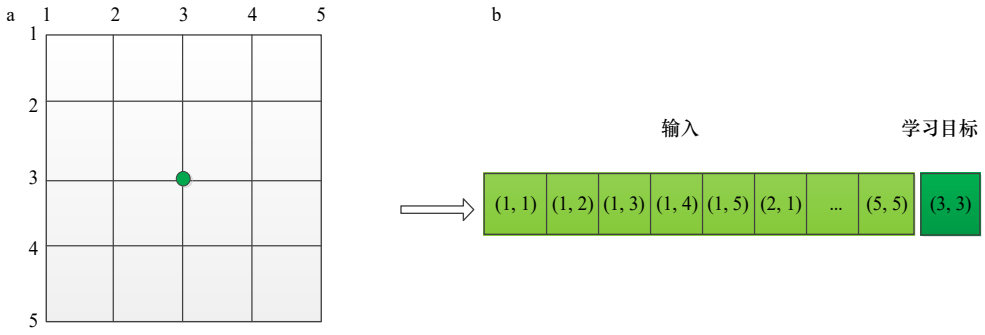

图 2　训练样本生成

Fig. 2　Generation of training samples

3.2.3　XGBoost 算法

集成学习通过构建并结合多个学习器来完成学习任务，比单一学习器获得显著优越的泛化性能。XGBoost 是在梯度下降树(Gradient Boosting Decision Tree, GBDT)的基础上对 boosting 算法进行的改进，由多棵决策树迭代组成[33–36]。

XGBoost 算法的核心思想是每次构建一棵新树来学习上次预测得到的残差，即首先初始构建一棵树来预测一个值，得到预测值与实际值的残差，然后构建下一棵树来学习残差，直至构建 K 棵树，并在训练中对构建树不断优化，算法的整体思路如图 3 所示。XGBoost 算法将训练得到的各个决策树预测值相加，得到模型最终的预测值。如公式所示:

$$\hat{y}_i = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F, \tag{4}$$

式中，$\hat{y}_i$为模型对于第i个样本的预测值；x_i为第i个样本的标签；K为分类回归树的数量；f_k为第k棵树模型函数。

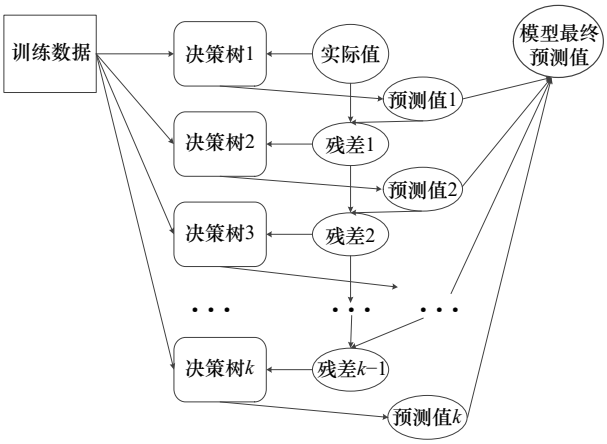

图 3　XGBoost 模型训练流程

Fig. 3　Training flow of the XGBoost model

4 模型评价及其分析

本文采用均方根误差(Root Mean Square Error, RMSE)、绝对误差(Mean Absolute Error, MAE)、相关系数(R)、标准差(σ)以及中心均方根误差(E')5 种误差统计方法来对风速模型性能进行评估[21, 37]。

$$R=\frac{Cov(a,b)}{\sigma_a\sigma_b}, \tag{5}$$

$$\sigma_s=\sqrt{\frac{\sum_{n=1}^{N}(s-\bar{s})^2}{N}}, \tag{6}$$

$$\mathrm{MAE}=\frac{\sum_{i=1}^{N}|Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i|}{N}, \tag{7}$$

$$RMSE=\sqrt{\frac{\sum_{i=1}^{N}(Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i)^2}{N}}, \tag{8}$$

$$E'=\sqrt{\sigma_a{}^2+\sigma_b{}^2-2\sigma_a\sigma_bR}. \tag{9}$$

对于风向来说,使用常规的 RMSE 以及 MAE 并不能够很好地衡量研究结果,因此本文采用 $\mathrm{RMSE_d}$ 以及 $\mathrm{MAE_d}$[21] 进行评价。

$$\mathrm{RMSE_d}=\sqrt{\frac{\sum_{i=1}^{N}(Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i)^2}{N}-\mathrm{ME}^2}, \tag{10}$$

$$\mathrm{MAE_d}=\frac{\sum_{i=1}^{N}|E_i|}{N}, \tag{11}$$

其中,

$$E_i=\begin{cases}Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i, & -180^\circ\leqslant Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i\leqslant 180^\circ,\\ Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i+360, & Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i<-180^\circ,\\ Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i-360, & Y_{\mathrm{mod}}^i-Y_{\mathrm{obs}}^i>180^\circ,\end{cases} \tag{12}$$

$$\mathrm{ME}=\frac{\sum_{i=1}^{N}E_i}{N}. \tag{13}$$

4.1 研究区域风场分析

在数据获取的过程中对 ERA-5 风场数据中陆地部分和海洋部分进行分析,如图 4 所示,陆地风场的风速分布和海洋风场的风速分布存在着很大的不同。陆地风场整体风速较小,其分布峰值约为 2.5 m/s,大部分风力等级在 5 级风以下,而在海洋风场中,风速分布峰值在 6～8 m/s 之间,整体分布在 0 m/s 至 20.0 m/s,并且 6 级以上大风发生频率较大。本文分析海洋中由于海面宽阔,没有遮挡物,对空气移动的

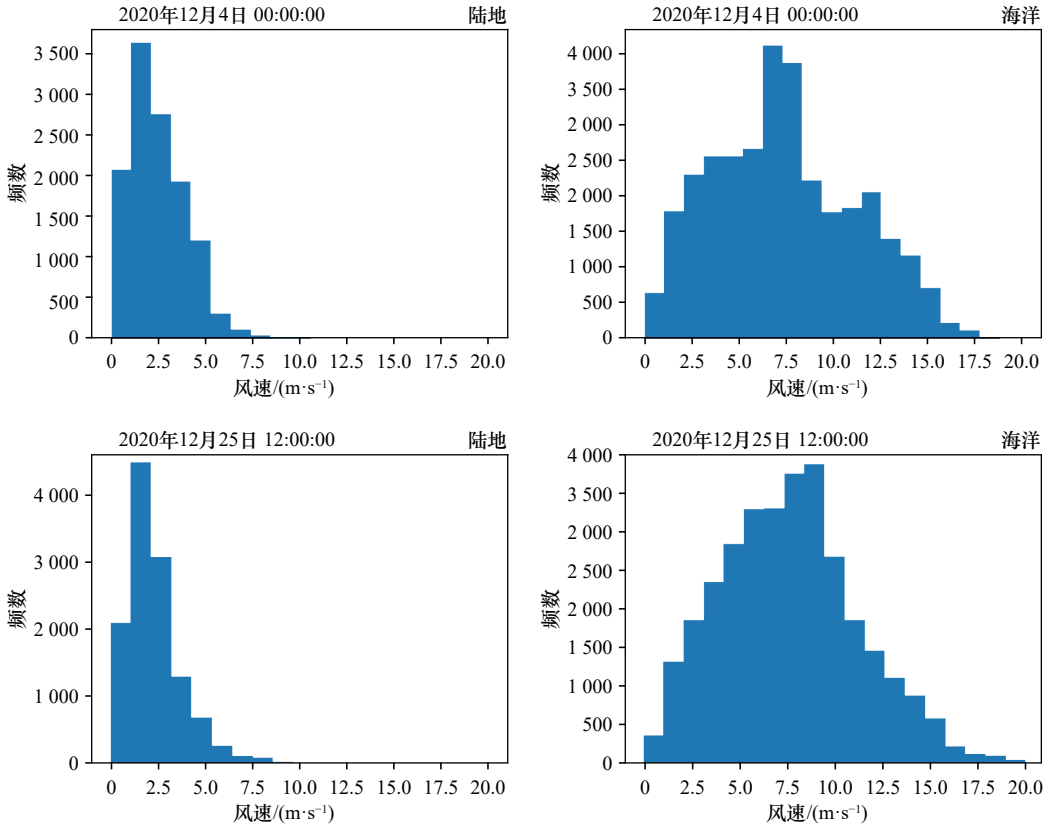

图 4 海陆风场风速统计直方图

Fig. 4 Histogram of wind speed statistics for sea and land wind fields

摩擦力小, 从而风速较大, 陆地上由于地面粗糙, 地形起伏, 有植被及建筑物阻碍等对空气移动的摩擦较大, 导致风速较小。介于陆地风场和海洋风场分布的不同, 本研究采用 4 种修正方法来对风场进行修正, 即 3.2 节中提出的修正方法。

4.2　ERA-5 修正融合结果分析

融合风场模型在(0°～45°N, 100°E～180°)研究区域内进行机器推理, 其中全区域内共计 58 101 个点, 在推理过程中由于 5×5 窗口的存在, 模型最终对 56 109 个点进行修正。推理过程中模型输入为 ERA-5 数据, 且在 CPU 上进行, 当前实验使用的 CPU 型号为 Intel(R) Xeon(R) CPU E5-2 690 v4 @ 2.60GHz, 单时刻推理平均用时约为 2.1 s。

4.2.1　原始卫星数据评价

在机器学习领域中, 以未参与训练的真值数据检验模型(模型测试与评价)必不可少。本文以 ERA-5 风场数据作为输入, 以卫星插值数据作为学习目标训练融合模型, 该模型期望从 ERA-5 风场推理得出卫星风场(本文称为融合风场便于和卫星真值相区分)。若推理得出的融合风场相较于 ERA-5 风场更加接近卫星原始数据, 即说明融合风场模型有效。所以本文以未参与模型训练的卫星原始数据进行测试评价。

一般机器学习的评价所用真值数据和模型推理数据处于同样的网格点。融合模型推理得到数据处于 ERA-5 的网格点, 与卫星原始数据位置并不一样, 而作为评价的卫星原始数据是不能做任何插值处理的。本文是将融合风场数据再插值回到卫星原始数据点进行比较。由于模型本身的学习目标是插值后卫星数据, 而检验和评价却使用卫星原始数据, 这其实是超出一般机器学习检验的更高和更严的要求。如能在这一更高要求下, 融合模型也能得到很好的结果, 则说明该融合方法是有效的。

实验中测试数据为 2021 年 1 月 31 日 00 时和 12 时数据, 共计约 130 000 个。实验使用训练完毕的 XGBoost 模型对 ERA-5 数据进行修正, 得到修正后的 ERA-5 数据分别插值到对应时间点的卫星数据上, 即将卫星数据作为真值, 计算 RMSE 等, 最终实验结果如下所示。

(1)表 1 和表 2 展示了对于 U_V_A_XGBoost 模型和 D_S_A_XGBoost 模型的评价信息及结果。

(2)表 3 和表 4 展示了对于 U_V_O_XGBoost 模

表 1　卫星评价数据信息(全区域)

Table 1　Satellite data information used in the test (whole region)

卫星	时间	数据数量
HY-2B	2021年1月31日 00:00:00	7 026
	2021年1月31日 12:00:00	6 233
CFOSAT	2021年1月31日 00:00:00	6 419
	2021年1月31日 12:00:00	12 928
MetOp-B	2021年1月31日 00:00:00	54 982
	2021年1月31日 12:00:00	55 586

表 2　全区域训练模型评价结果

Table 2　Evaluation results of the whole regional training model

卫星	时间	模型	风向/(°)		风速/(m·s^{-1})	
			$RMSE_d$	MAE_d	RMSE	MAE
HY-2B	2021年1月31日 00:00:00	原始	42.941	12.480	1.313	0.917
		U_V_A_XGBoost	42.133	**11.614**	1.166	0.803
		D_S_A_XGBoost	**39.497**	12.490	**1.083**	**0.774**
	2021年1月31日 12:00:00	原始	50.959	11.917	1.238	0.946
		U_V_A_XGBoost	48.910	**10.757**	**1.118**	0.853
		D_S_A_XGBoost	**43.517**	10.989	1.133	**0.845**
CFOSAT	2021年1月31日 00:00:00	原始	37.334	7.685	1.814	1.461
		U_V_A_XGBoost	**35.012**	**7.150**	1.465	1.150
		D_S_A_XGBoost	35.529	8.060	**1.434**	**1.107**
	2021年1月31日 12:00:00	原始	79.938	14.234	1.340	1.030
		U_V_A_XGBoost	78.729	**13.814**	**1.194**	**0.903**
		D_S_A_XGBoost	**76.858**	15.671	1.269	0.952
MetOp-B	2021年1月31日 00:00:00	原始	25.232	**9.860**	1.270	0.940
		U_V_A_XGBoost	24.408	10.122	1.118	0.806
		D_S_A_XGBoost	**24.391**	10.063	**1.053**	**0.771**
	2021年1月31日 12:00:00	原始	32.589	**8.530**	1.190	0.883
		U_V_A_XGBoost	**31.433**	8.534	1.076	0.786
		D_S_A_XGBoost	33.728	9.318	**1.011**	**0.735**

注: 加粗数字表示最优结果。

表 3 卫星评价数据信息（陆地掩码）

Table 3 Satellite data information used in the test (land mask)

卫星	时间	数据数量
HY-2B	2021年1月31日 00:00:00	7 026
	2021年1月31日 12:00:00	6 233
CFOSAT	2021年1月31日 00:00:00	6 419
	2021年1月31日 12:00:00	12 928
MetOp-B	2021年1月31日 00:00:00	54 977
	2021年1月31日 12:00:00	55 575

表 4 陆地掩码训练模型评价结果

Table 4 Evaluation results of land mask training model

卫星	时间	模型	风向/(°)		风速/(m·s^{-1})	
			$RMSE_d$	MAE_d	RMSE	MAE
HY-2B	2021年1月31日 00:00:00	原始	42.941	12.480	1.313	0.917
		U_V_O_XGBoost	41.342	**11.508**	1.182	0.814
		D_S_O_XGBoost	**38.860**	12.409	**1.076**	**0.767**
	2021年1月31日 12:00:00	原始	50.959	11.917	1.238	0.946
		U_V_O_XGBoost	49.686	**10.843**	**1.120**	0.851
		D_S_O_XGBoost	**42.023**	10.990	**1.120**	**0.834**
CFOSAT	2021年1月31日 00:00:00	原始	37.334	7.685	1.814	1.461
		U_V_O_XGBoost	**34.275**	**7.214**	1.461	1.146
		D_S_O_XGBoost	36.178	8.031	**1.431**	**1.103**
	2021年1月31日 12:00:00	原始	79.938	14.234	1.340	1.030
		U_V_O_XGBoost	79.149	**13.827**	**1.219**	**0.914**
		D_S_O_XGBoost	**76.327**	15.197	1.241	0.937
MetOp-B	2021年1月31日 00:00:00	原始	25.213	**9.860**	1.265	0.937
		U_V_O_XGBoost	25.183	10.093	1.140	0.815
		D_S_O_XGBoost	**24.996**	10.079	**1.093**	**0.787**
	2021年1月31日 12:00:00	原始	32.586	**8.525**	1.182	0.880
		U_V_O_XGBoost	**31.501**	8.627	1.111	0.806
		D_S_O_XGBoost	33.056	9.357	**1.058**	**0.763**

注：加粗数字表示最优结果。

型和 D_S_O_XGBoost 模型的评价信息及结果。

对比表 1 数据信息和表 3 数据信息可以发现，表 3 中的 MetOp-B 卫星测试数据比表 1 中 MetOp-B 卫星测试数据少，这是因为模型 U_V_O_XGBoost 和 U_V_O_XGBoost 是基于陆地掩码的模型，所以在测试的时候贴近陆地的卫星数据可能无法进行评估，从而导致了测试数据减少。

从表 2 中分析，对于风向来说，U_V_A_XGBoost 模型在 MAE_d 方面表现最好，除了在 MetOp-B 卫星上有所上升，在 HY-2B 和 CFOSAT 卫星上均下降，在 $RMSE_d$ 方面，D_S_A_XGBoost 模型的表现较好，但在 2021 年 1 月 31 日 12 时的测试样例中，在 MetOp-B 评价结果中出现了上升的情况，而 U_V_A_XGBoost 模型表现稳定，全部呈现下降趋势。对于风速来说，不论是 D_S_A_XGBoost 模型还是 U_V_A_XGBoost 模型，在 RMSE 以及 MAE 方面结果均下降。整体来说，U_V_A_XGBoost 模型的表现较稳定。

从表 4 进行分析，对于风向来说，U_V_O_XGBoost 模型在 MAE_d 方面表现最好，与表 2 中 U_V_A_XGBoost 模型的表现类似，除了在 MetOp-B 卫星上有所上升，在 HY-2B 和 CFOSAT 卫星结果中均下降，在 $RMSE_d$ 方面，D_S_O_XGBoost 模型表现整体要好于 U_V_O_XGBoost 模型，但是同样在 2021 年 1 月 31 日 12 时的测试样例中，出现了上升的情况，而 U_V_O_XGBoost 模型一直保持下降。对于风速来说，U_V_O_XGBoost 模型和 D_S_O_XGBoost 模型均表现良好，不论是在 RMSE 还是在 MAE 方面，测试结果均下降。整体来说，U_V_O_XGBoost 模型的稳定性较好。

综上所述，所有的模型在 HY-2B 卫星和 CFOSAT 卫星上的测试结果表现良好，但是在 MetOp-B 卫星的风向修正方面表现不理想，研究分析认为，导致该现象的原因可能有两点，一是 ERA-5 再分析数据的制作过程中使用了 MetOp-B 卫星数据，所以修正后的 ERA-5 数据可能会与 MetOp-B 卫星数据偏差增大；二是 HY-2B 卫星和 CFOSAT 卫星都是中国参与研制并运行的卫星，而 MetOp-B 卫星是欧洲卫星，卫星数据之间可能存在差异，模型的训练过程中可能更加偏向了 HY-2B 卫星和 CFOSAT 卫星，所以导致 MetOp-B 卫星的模型结果不佳。根据表 2 和表 4 的模型结果发现，使用 U、V 分量修正风速风向的研究方法在稳定性上要好于使用直接风速风向进行修正的研究方法。

图 5 表示的是 U_V_O_XGBoost 模型（模型选择的具体原因参见 4.2.2 节）修正结果在 HY-2B 卫星、

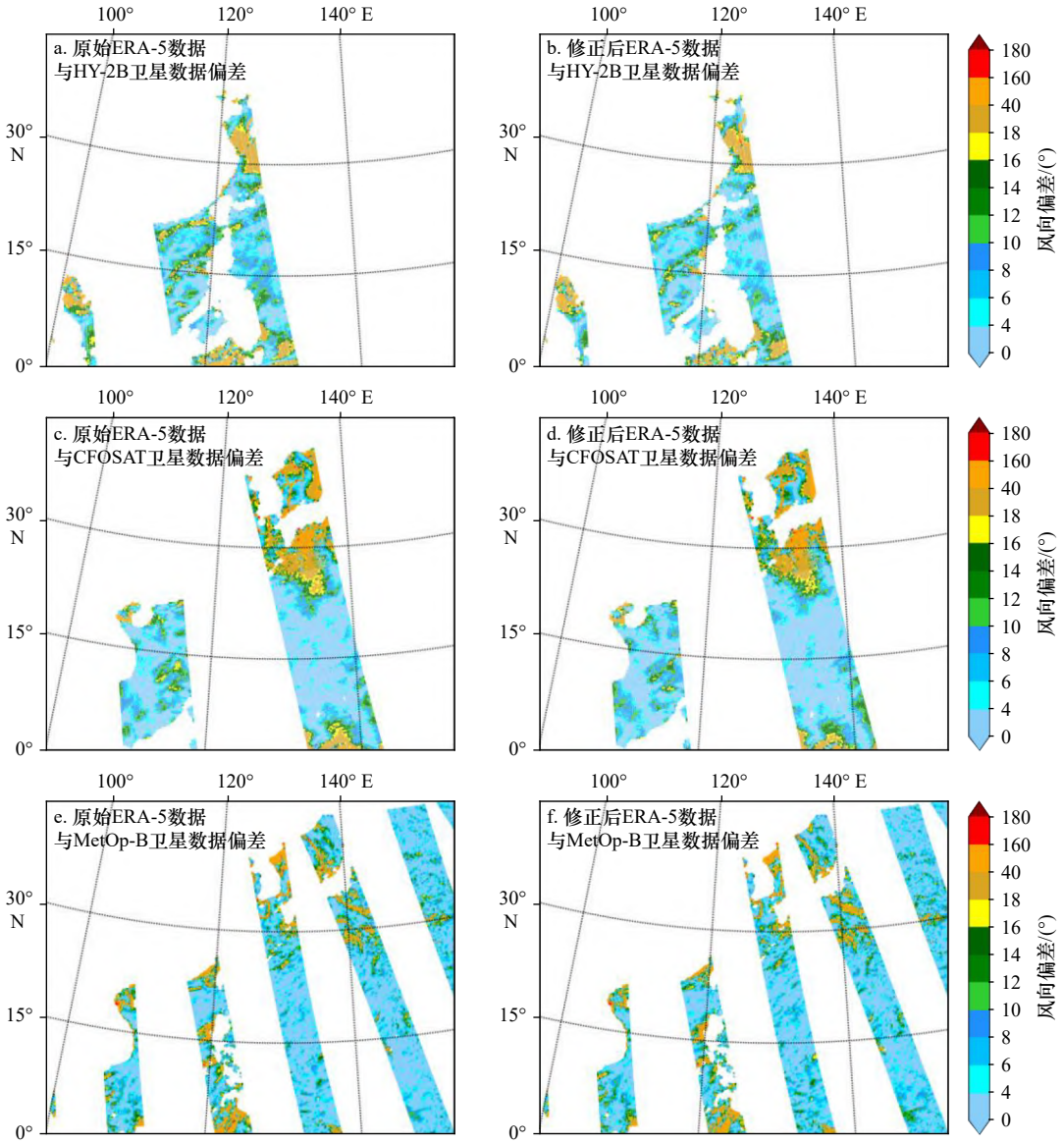

图 5　ERA-5 修正的风向效果图

Fig. 5　Wind direction effect diagram of ERA-5 correction experiment

CFOSAT 卫星以及 MetOp-B 卫星上的关于风向的展示, 挑选的时间为 2021 年 1 月 31 日 12 时。其中图中描述的是 ERA-5数据的插值结果与该点上卫星数据偏差的绝对值即 MAE_d, 图 5a 和图 5b 表示 HY-2B 卫星效果图, 图 5c 和图 5d 表示 CFOSAT 卫星效果图, 图 5e 和图 5f 表示 MetOp-B 卫星效果图。左侧图表示的是原始 ERA-5数据与卫星数据之间的偏差, 右侧图表示的是修正后的 ERA-5 数据与卫星数据之间的偏差。图 6 为 U_V_O_XGBoost 模型修正结果在 HY-2B、CFOSAT 以及 MetOp-B 卫星上的关于风速的展示, 所选时间为 2021 年 1 月 31 日 12 时。其中图中描述的是 ERA-5 数据的插值结果与该点上卫星数据偏差的绝对值即 MAE。图 6a 和图 6b 表示 HY-2B 卫星效果图, 图 6c 和图 6d 表示 CFOSAT 卫星效果图, 图 6e 和图 6f 表示 MetOp-B 卫星效果图。

4.2.2　浮标评价

本文使用浮标数据对 ERA-5 数据的修正方法进行评价, 选取的浮标为经纬度在 8°N, 165°E, 距离海面 4 m 高的 TAO 浮标, 选取的时间范围为 2020 年 12 月至 2021 年 1 月共计两个月的数据。本文剔除与浮标风速相差 3 倍标准差的数据, 并剔除与浮标风向相差大于 90°的数据[21, 38], 原因在于本文认定该点数据可能存在较为明显的误差, 该点数据可能会对整体的数据评价造成较大的影响。最终 ERA-5 数据与浮标数据匹配后得到 123 个测试样例, 计算相关系数等评价指标, 结果如图 7 所示。

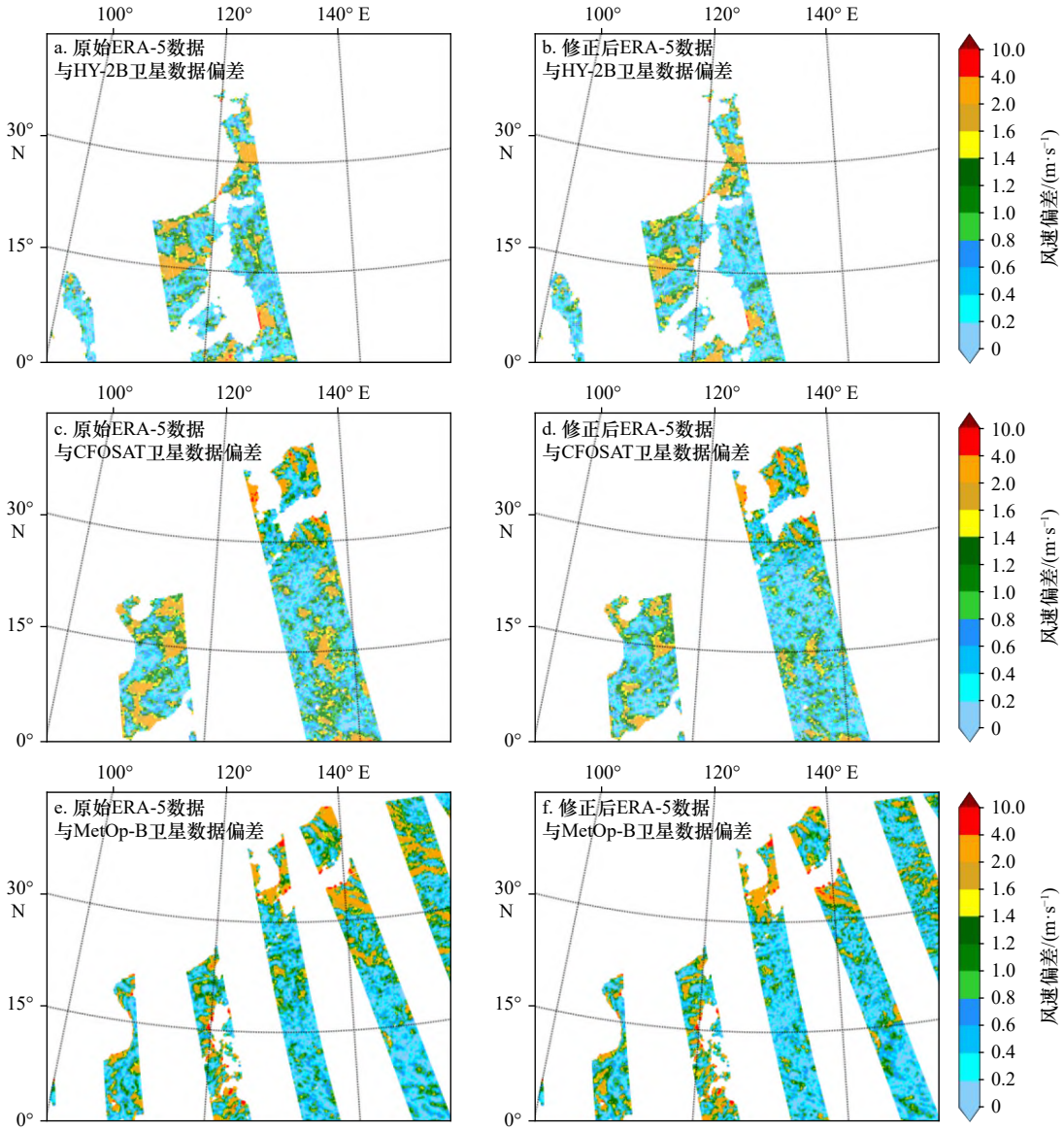

图 6 ERA-5 修正的风速效果图

Fig. 6 Wind speed effect diagram of ERA-5 correction experiment

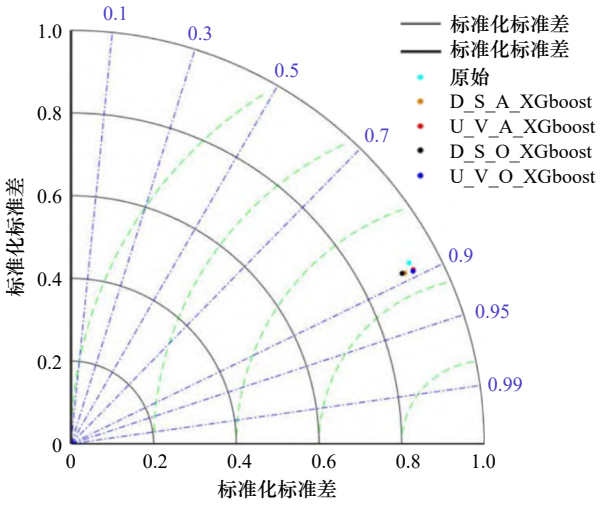

图 7 风速泰勒图

Fig. 7 Tyler diagram of wind speed

图 7 中展示的是 4 个模型方法的结果和原始 ERA-5 风速数据与浮标风速数据之间的差异，从图中可以看出，使用 U_V_O_XGBoost 模型修正的 ERA-5 数据与浮标数据的相关系数最高，中心均方根误差最小，整体结果要好于原始 ERA-5 数据的结果，意味着生成的融合风场数据更加接近浮标数据。

图 8 表示的是浮标数据与 ERA-5 原始数据以及 U_V_O_XGBoost 模型修正后的融合风场数据的匹配情况。从图中可以看出风速在不同时刻差异明显，例如风速可以从 5 m/s 迅速增到 9 m/s，同样可以从约 13 m/s 迅速减小到 5 m/s，风速前后相差较大。通过观察发现，在图中黑框区域，修正融合风场数据与浮标数据的差距明显减小，表明修正融合风场数据更加接

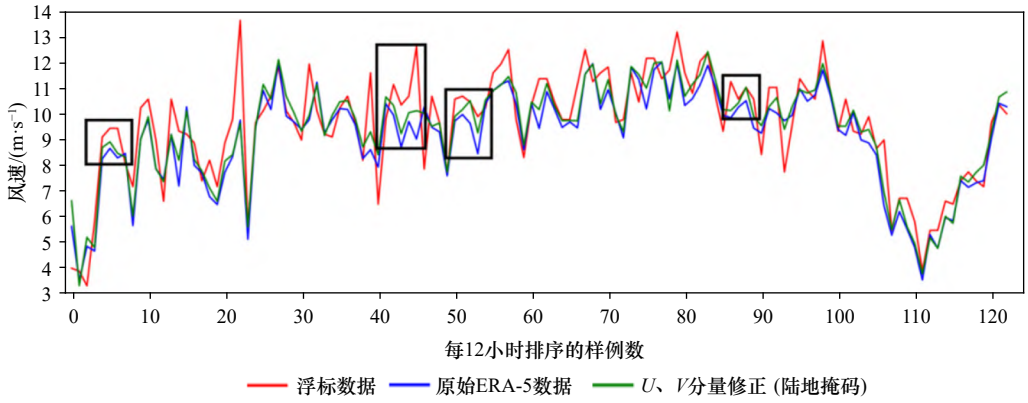

图 8　浮标数据与 ERA-5 数据的风速对比

Fig. 8　Comparison of wind speed between buoy data and ERA-5 data

近浮标数据。图 9 中分别表示的是 ERA-5 数据与浮标数据的风速相关性以及融合风场与浮标数据的风速相关性。从图 9 中可以看出融合风场风速相较于 ERA-5 数据来说相关系数有所提高。

本文采用 Adaboost 以及 Random Forest 算法进行风场融合研究，与 XGBoost 方法进行比较，结果如表 5 所示，其中相关系数、均方根误差以及标准差的计算公式在第 4 章进行了说明。从表中可以看出，Adaboost、Random Forest 以及 XGBoost 等算法生成的融合风场数据相比 ERA-5 数据来说与浮标的相关系数均有所提高，即更加接近于浮标数据，且 XGBoost 算法相对来说效果最好。

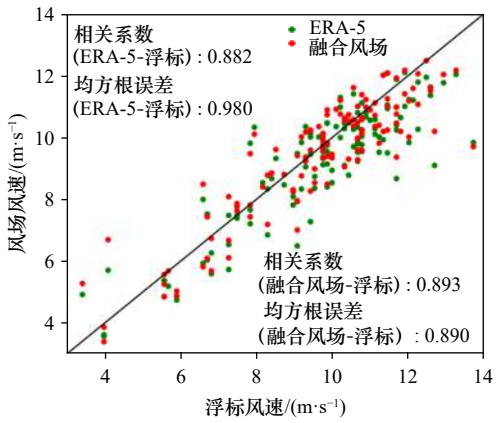

图 9　风速散点图

Fig. 9　Scatter plot about wind speed

表 5　不同机器学习算法风场融合结果

Table 5　Wind field fusion results of different machine learning algorithms

	相关系数	均方根误差	标准差
ERA-5	0.882	0.980	1.938
XGBoost	0.893	0.890	1.936
Random Forest	0.890	0.915	1.955
Adaboost	0.892	0.906	1.978

4.2.3　融合时间对比

本文目的在于降低风场融合的硬件要求，提高融合速度，且保证融合风场的质量。因此本文对融合时间进行统计对比，数据结果如表 6 所示。表中 XGBoost 表示的是本文采用 XGBoost 模型针对单一风场要素进行海面风场融合的方法，插值方法表示的是采用传统的 IDW 方法针对单一风场要素进行海面风场融合。本文在 0°～45°N，0°～180°区域共计 58 101 个网格点进行海面风场融合，针对 1 个月数据共计 60 次融合时间进行统计分析，结果如表 6 所示。XGBoost 模型方法融合时间明显优于传统插值方法。

表 6　融合时间对比

Table 6　Comparison of fusion time

融合方法	平均推理时间/s
XGBoost模型	2.063
插值方法(IDW)	226.616

4.3　融合风场展示

本文以 ERA-5 数据作为模型输入，以卫星插值数据作为学习目标进行模型训练，得到海面风场修正融合模型，最终采用训练完毕的海面风场修正融合模型进行推理，得到融合风场。图 10 中表示的是 2021 年 1 月 30 日 00 时的融合情况，从图 10 中可以看出卫星数据与 ERA-5 数据以及融合风场数据均具有大致相同的数据分布。从上述 3 处风场分布来看，融合风场数据更加贴近卫星数据，即风速达到 12.5 m/s 以上的区域中融合风场更加接近卫星数据分布情况。图 11 展示的是融合风场中风速在 2021 年 1 月 27 日 12 时

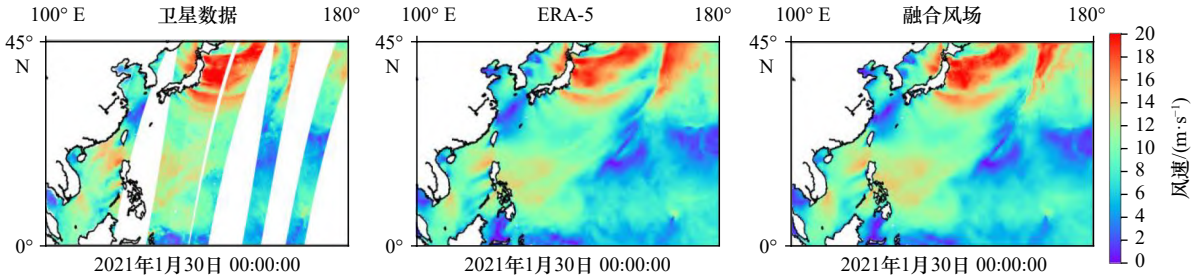

图 10 风速对比图

Fig. 10 Comparison chart of wind speed data

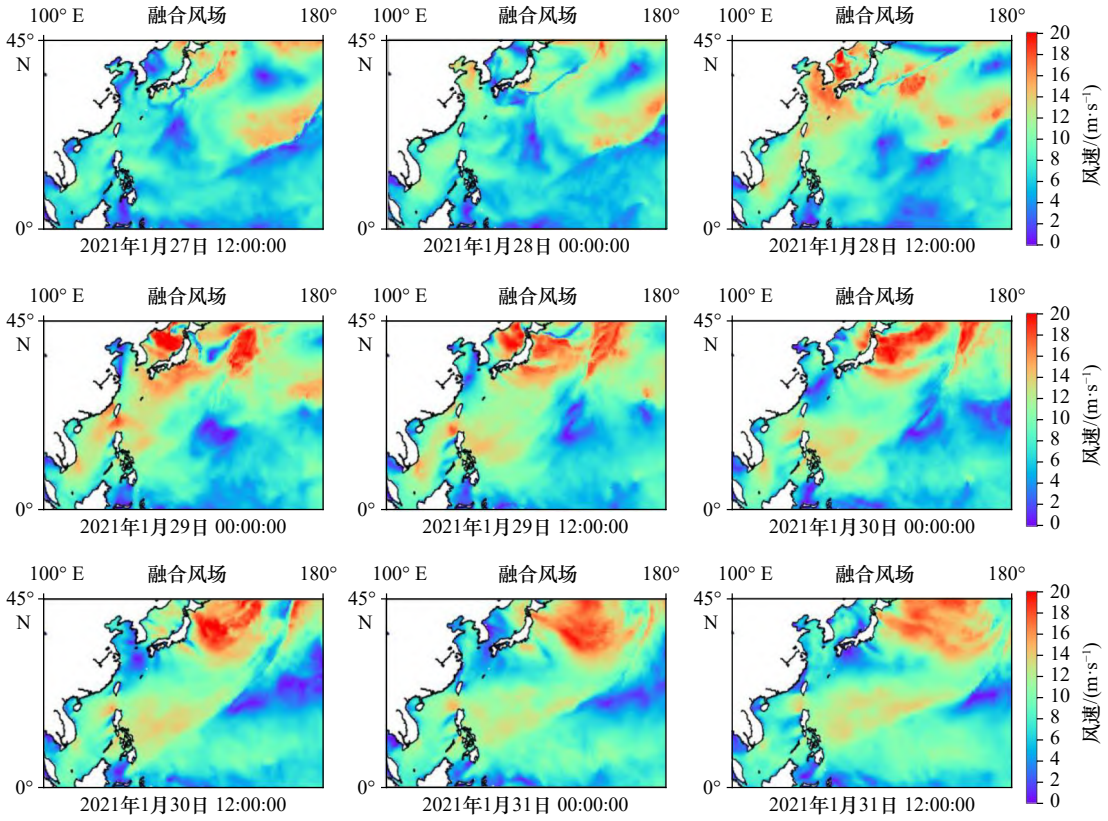

图 11 融合风速效果图

Fig. 11 The effect of wind speed after fusion

至 2021 年 1 月 31 日 12 时的连续时空分布情况，其时间分辨率为 12 h，图 12 展示的是该时间段融合风场整体分布情况，由图可以看出该时段西北太平洋区域风场多为东北风或西北风。

5 总结

本文使用 CFOSAT 卫星、HY-2B 卫星、MetOp-B 卫星数据以及 ERA-5 再分析数据，利用传统机器学习 XGBoost 在研究区域（0°～45°N，100°E～180°）内进行生成融合风场的研究。研究首先以卫星数据作为学习目标，将 ERA-5 数据作为模型输入训练得到修正融合风场生成模型，然后利用融合风场生成模型进行机器推理最终得到全区域空间分辨率为 0.25°×0.25°，时间分辨率为 12 h 的融合风场。其中，在机器推理过程中，生成单时刻全区域融合风场的时间仅需要约 2 s，相比较传统融合方法来说，该模型方法更加快速高效。文中共提出 4 种模型进行融合风场的研究，结论如下：

（1）使用 U、V 分量修正风速风向的研究方法比直接修正风速风向的研究方法在结果上更加稳定。

（2）U_V_O_XGBoost 模型得到的融合风场数据在风速方面最为接近浮标数据，同时风场修正结果稳定。

（3）研究中出现了修正融合结果在 MetOp-B 卫星

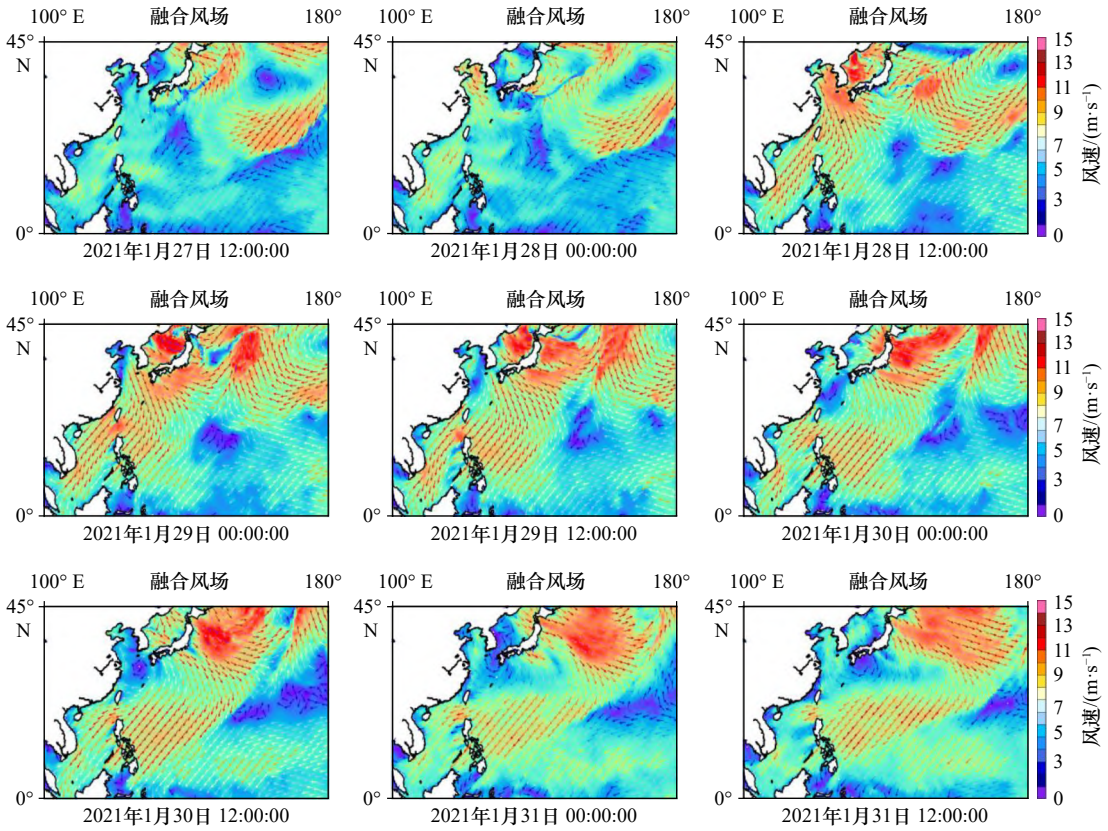

图 12　融合风场效果图

Fig. 12　Rendering of the fusion wind field

风向方面上升, 在 HY-2B 卫星和 CFOSAT 卫星表现良好的情况, 分析得到 HY-2B 卫星和 CFOSAT 卫星均为中国参与研制并运行, 而 MetOp-B 卫星为欧洲气象卫星, 两者存在差异, 该差异导致了模型在学习过程中出现偏向。

总而言之, 传统机器学习方法在对 ERA-5 再分析数据修正融合的过程中, 能够有效地学习到卫星数据的分布特征, 使得修正融合后的风场数据更加贴近研究区域内卫星数据分布, 从而提高生成的融合风场的质量。对于目前, 深度学习取得了重大进展, 深度学习擅长抽取高维数据的复杂结构, 通过足够多的数据和组合, 学习到非常复杂的函数关系[39]。因此本文下一步准备将深度学习方法应用到融合风场的研究中, 提高融合风场精度。

参考文献:

[1] 旷芳芳, 张友权, 张俊鹏, 等. 3种海面风场资料在台湾海峡的比较和评估[J]. 海洋学报, 2015, 37(5): 44−53.
Kuang Fangfang, Zhang Youquan, Zhang Junpeng, et al. Comparison and evaluation of three sea surface wind products in Taiwan Strait[J]. Haiyang Xuebao, 2015, 37(5): 44−53.

[2] 廖菲, 邓华, 曾琳, 等. 南海北部海面风速概率分布特征[J]. 海洋学报, 2018, 40(5): 37−47.
Liao Fei, Deng Hua, Zeng Lin, et al. The probability distribution of sea surface wind speeds over the northern South China Sea[J]. Haiyang Xuebao, 2018, 40(5): 37−47.

[3] 韩玉康, 周林, 赵艳玲, 等. 3种海面风场资料在吕宋海峡的比较与评估[J]. 海洋预报, 2019, 36(6): 44−52.
Han Yukang, Zhou Lin, Zhao Yanling, et al. Evaluation of three sea surface wind data sets in Luzon Strait[J]. Marine Forecasts, 2019, 36(6): 44−52.

[4] 张毅, 蒋兴伟, 林明森, 等. 星载微波散射计的研究现状及发展趋势[J]. 遥感信息, 2009(6): 87−94.
Zhang Yi, Jiang Xingwei, Lin Mingsen, et al. The present research status and development trend of spacebonre microwave scatterometer[J]. Remote Sensing Information, 2009(6): 87−94.

[5] 解学通, 郁文贤, 郭丽青, 等. 基于遗传算法的微波散射计海面风矢量反演研究[J]. 海洋通报, 2008, 27(4): 1−11.
Xie Xuetong, Yu Wenxian, Guo Liqing, et al. Research on genetic algorithm based ocean surface wind vector retrieval for microwave scatterometer[J]. Marine Science Bulletin, 2008, 27(4): 1−11.

[6] 林溢园, 邹巨洪, 何原荣, 等. 我国海洋二号卫星微波散射计数据处理软件设计[J]. 海洋通报, 2016, 35(4): 443−448.

Lin Yiyuan, Zou Juhong, He Yuanrong, et al. Design of data processing software for HY-2 satellite microwave scatterometer[J]. Marine Science Bulletin, 2016, 35(4): 443−448.

[7] 陈心一, 郝增周, 潘德炉, 等. 中国近海海面风场的时空特征分析[J]. 海洋学研究, 2014, 32(1): 1−10.
Chen Xinyi, Hao Zengzhou, Pan Delu, et al. Analysis of temporal and spatial feature of sea surface wind field in China offshore[J]. Journal of Marine Sciences, 2014, 32(1): 1−10.

[8] Hung S C, Chang W Y, Tsai W F, et al. Development of high-precision wind, wave and current forecast system for offshore wind energy industry in Taiwan: a two-stage method of numerical simulation and AI correction[J]. Journal of the Chinese Institute of Engineers, 2021, 44(6): 532−543.

[9] 刘付前, 骆永军, 王超. 基于遥感资料南海月平均风场分析[C]. 2009 航海技术理论研究论文集, [出版地不详: 出版者不详], 2009.
Liu Fuqian, Luo Yongjun, Wang Chao. Analysis of the monthly average wind field in the South China Sea based on remote sensing data [J]. 2009 Research Papers on Navigation Technology Theory, [S.l.: s.n.], 2009.

[10] 唐焕丽, 姚琴, 吕晓莹, 等. 多源卫星融合的广东海域海面风场特征[J]. 遥感信息, 2020, 35(1): 117−122.
Tang Huanli, Yao Qin, Lü Xiaoying, et al. Characteristics of sea surface wind field in Guangdong sea area with multi-source satellite fusion[J]. Remote Sensing Information, 2020, 35(1): 117−122.

[11] 冯倩. 多传感器卫星海面风场遥感研究[D]. 青岛: 中国海洋大学, 2004.
Feng Qian. Study of sea surface wind remote sensing by satellite multi-sensor data[D]. Qingdao: Ocean University of China, 2004.

[12] 柳婧. 基于最优插值方法的中国近海海面风场资料融合研究[D]. 北京: 国家海洋环境预报中心, 2018.
Liu Jing. Research on data fusion of sea surface wind in China's offshore based on optimal interpolation method[D]. Beijing: National Marine Environmental Forecasting Center, 2018.

[13] 凌征, 王桂华, 陈大可, 等. 中国近海风场融合[C]// 首届中国"数字海洋"论坛. 天津: 国家海洋信息中心, 2008, 90−94.
Ling Zheng, Wang Guihua, Chen Dake, et al. Integration of offshore wind fields in China [C]// The First China "Digital Ocean" Forum. Tianjin: National Maritime Information Centres, 2008, 90−94

[14] Zhang H M, Reynolds R W, Smith T M. Adequacy of the *in situ* observing system in the satellite era for climate SST[J]. Journal of Atmospheric and Oceanic Technology, 2006, 23(1): 107−120.

[15] Zhang H M, Reynolds R W, Bates J J. P2. 23 blended and gridded high resolution global sea surface wind speed and climatology from multiple satellites: 1987-present[C]//Proceedings of the 14th Conference on Satellite Meteorology and Oceanography. Atlanta, GA: American Meteorological Society 2006 Annual Meeting, 2006, 2.

[16] 齐亚琳, 林明森. 数据融合技术在海洋二号卫星数据中的应用[J]. 航天器工程, 2012, 21(3): 117−123.
Qi Yalin, Lin Mingsen. Application of the data fusion technique in the HY-2 satellite data[J]. Spacecraft Engineering, 2012, 21(3): 117−123.

[17] Yan Q S, Zhang J, Meng J M, et al. Use of an optimum interpolation method to construct a high-resolution global ocean surface vector wind dataset from active scatterometers and passive radiometers[J]. International Journal of Remote Sensing, 2017, 38(20): 5569−5591.

[18] Chao Y, Li Z J, Kindle J C, et al. A high-resolution surface vector wind product for coastal oceans: Blending satellite scatterometer measurements with regional mesoscale atmospheric model simulations[J]. Geophysical Research Letters, 2003, 30(1): 13−1−13−4.

[19] 张东翔. 多源卫星海面风场产品检验及融合研究[D]. 长沙: 国防科技大学, 2018.
Zhang Dongxiang. Research of multi-source satellite sea surface wind validation and data fusion[D]. Changsha: National University of Defense Technology, 2018.

[20] 金荣花, 代刊, 赵瑞霞, 等. 我国无缝隙精细化网格天气预报技术进展与挑战[J]. 气象, 2019, 45(4): 445−457.
Jin Ronghua, Dai Kan, Zhao Ruixia, et al. Progress and challenge of seamless fine gridded weather forecasting technology in China[J]. Meteorological Monthly, 2019, 45(4): 445−457.

[21] 陈克海, 解学通, 张金兰, 等. HY-2B卫星散射计海面风场产品质量分析[J]. 热带海洋学报, 2020, 39(6): 30−40.
Chen Kehai, Xie Xuetong, Zhang Jinlan, et al. Accuracy analysis of the retrieved wind from HY-2B scatterometer[J]. Journal of Tropical Oceanography, 2020, 39(6): 30−40.

[22] 黄耀辉, 赵晓磊, 阎诚, 等. 中法海洋卫星及典型应用[J]. 卫星应用, 2020(5): 32−37.
Huang Yaohui, Zhao Xiaolei, Yan Cheng, et al. CFOSAT and typical applications[J]. Satellite Application, 2020(5): 32−37.

[23] Shen S S P, Dzikowski P, Li G L, et al. Interpolation of 1961−97 daily temperature and precipitation data onto Alberta polygons of ecodistrict and soil landscapes of Canada[J]. Journal of Applied Meteorology and Climatology, 2001, 40(12): 2162−2177.

[24] Hofstra N, Haylock M, New M, et al. Comparison of six methods for the interpolation of daily, European climate data[J]. Journal of Geophysical Research: Atmospheres, 2008, 113(D21): D21110.

[25] 潘留杰, 薛春芳, 王建鹏, 等. 一个简单的格点温度预报订正方法[J]. 气象, 2017, 43(12): 1584−1593.
Pan Liujie, Xue Chunfang, Wang Jianpeng, et al. A simple grid temperature forecast correction method[J]. Meteorological Monthly, 2017, 43(12): 1584−1593.

[26] Jones P D, Raper S C B, Bradley R S, et al. Northern hemisphere surface air temperature variations: 1851−1984[J]. Journal of Applied Meteorology and Climatology, 1986, 25(2): 161−179.

[27] 陈小燕, 杨劲松, 黄韦艮, 等. 多源卫星高度计有效波高数据融合方法研究[J]. 海洋学报, 2009, 31(4): 51−57.

Chen Xiaoyan, Yang Jinsong, Huang Weigen, et al. Research on the fusion methods of significant wave height data from multisatellite altimeters[J]. Haiyang Xuebao, 2009, 31(4): 51−57.
[28] 李彦, 王丽娜, 蒋镇. 一种针对气象要素的空间插值算法[J]. 重庆理工大学学报(自然科学), 2014, 28(6): 94−98, 116.
Li Yan, Wang Li'na, Jiang Zhen. One kind of spatial interpolation algorithm for meteorological elements[J]. Journal of Chongqing University of Technology (Natural Science), 2014, 28(6): 94−98, 116.
[29] 饶莉娟, 王健林, 张星. 不同插值方法对精细化预报产品在青岛地区的检验比较[J]. 中国农学通报, 2020, 36(32): 100−108.
Rao Lijuan, Wang Jianlin, Zhang Xing. Different interpolation methods: comparison for refined forecast products in Qingdao area[J]. Chinese Agricultural Science Bulletin, 2020, 36(32): 100−108.
[30] 肇毓锋, 吴奇. 多时间尺度下Kriging与IDW空间插值方法的适用性研究[J]. 黑龙江水利科技, 2020, 48(11): 9−14.
Zhao Yufeng, Wu Qi. Applicability of Kriging and IDW spatial interpolation methods on multiple time scales[J]. Heilongjiang Hydraulic Science and Technology, 2020, 48(11): 9−14.
[31] 蒋伟达, 孙永福, 刘绍文, 等. 基于IDW的埕岛海域水下三角洲地形演变[J]. 海洋科学进展, 2020, 38(4): 697−707.
Jiang Weida, Sun Yongfu, Liu Shaowen, et al. Terrain evolution of subaqueous delta in Chengdao Sea area based on IDW[J]. Advances in Marine Science, 2020, 38(4): 697−707.
[32] 周志华. 机器学习[M]. 北京: 清华大学出版社, 2016: 23−47.
Zhou Zhihua. Machine Learning[M]. Beijing: Tsinghua University Press, 2016: 23−47.
[33] 马良玉, 於世磊, 赵尚羽, 等. 基于随机搜索算法优化XGBoost的过热汽温预测模型[J]. 华北电力大学学报(自然科学版), 2021, 48(4): 99−105.
Ma Liangyu, Yu Shilei, Zhao Shangyu, et al. Superheated steam temperature prediction models based on XGBoost optimized with random search algorithm[J]. Journal of North China Electric Power University (Natural Science Edition), 2021, 48(4): 99−105.
[34] 潘进, 丁强, 江爱朋, 等. 基于XGBoost的冷水机组不平衡数据故障诊断[J]. 机械强度, 2021, 43(1): 27−33.
Pan Jin, Ding Qiang, Jiang Aipeng, et al. Fault diagnosis of unbalanced data of chillers based on XGBoost[J]. Journal of Mechanical Strength, 2021, 43(1): 27−33.
[35] 孙晓黎, 马超群, 朱才华. 基于XGBoost的轨道交通短时客流预测精度分析[J]. 交通科技与经济, 2021, 23(1): 54−58.
Sun Xiaoli, Ma Chaoqun, Zhu Caihua. XGBoost-based analysis of prediction accuracy for short-term passenger flow in rail transit[J]. Technology & Economy in Areas of Communications, 2021, 23(1): 54−58.
[36] Chen T Q, Guestrin C. XGBoost: A scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco: ACM, 2016: 785−794.
[37] 曾晓青, 薛峰, 姚莉, 等. 针对模式风场的格点预报订正方案对比[J]. 应用气象学报, 2019, 30(1): 49−60.
Zeng Xiaoqing, Xue Feng, Yao Li, et al. Comparative study of different error correction methods on model output wind field[J]. Journal of Applied Meteorological Science, 2019, 30(1): 49−60.
[38] Freilich M H, Dunbar R S. The accuracy of the NSCAT 1 vector winds: comparisons with national data buoy center buoys[J]. Journal of Geophysical Research: Oceans, 1999, 104(C5): 11231−11246.
[39] 王国松, 王喜冬, 侯敏, 等. 基于观测和再分析数据的LSTM深度神经网络沿海风速预报应用研究[J]. 海洋学报, 2020, 42(1): 67−77.
Wang Guosong, Wang Xidong, Hou Min, et al. Research on application of LSTM deep neural network on historical observation data and reanalysis data for sea surface wind speed forecasting[J]. Haiyang Xuebao, 2020, 42(1): 67−77.

Sea surface wind field smart fusion base on machine learning method

Zhang Wei[1,2], Du Chaofan[2], Guo Anboyu[1], Song Xiaojiang[1], Shen Shiying[2]

(1. *National Marine Environmental Forecasting Center*, *Beijing* 100081, *China*; 2. *School of Computer Science and Technology*, *Ocean University of China*, *Qingdao* 266100, *China*)

Abstract: The assimilation fusion or interpolation fusion of the sea surface wind field based on multi-source data is currently restricted by computing power. This paper proposes to train the XGBoost-based machine learning ERA-5 data correction fusion model in the overlapping area of the multi-source satellite data and the ERA-5 reanalysis data, and then use the model to quickly correct (machine learning inference) ERA-5 data, of which the ERA-5 whole area correction fusion it only takes about 2 seconds. Due to the rapidity of machine learning inference, the entire sea surface fusion wind field can be constructed at a lower computational cost. This paper expands on typical wind field variables such as 10 m wind speed, 10 m wind direction, *U*10 component and *V*10 component, taking in-

to account the difference in sea and land distribution, using land masks to eliminate land areas, and constructing D_S_A_XGBoost, D_S_O_XGBoost, U_V_A_XGBoost, U_V_O_XGBoost corrections model, and finally generate sea surface fusion wind field. By comparing the ERA-5 reanalysis data before and after the correction with the satellite data, the above four models all reduce the gap between the ERA-5 reanalysis data and the satellite data. Especially in terms of wind speed, both root mean square error (RMSE) and mean absolute error (MAE) are effectively reduced. In terms of wind direction, $RMSE_d$ and MAE_d also show a decreasing trend. Using Tropical Atmosphere Ocean Array (TAO) buoy data to evaluate the four XGBoost models, it is found that the U_V_O_XGBoost model has the best correction results for ERA-5 data, and its correlation reaches 0.893, an increase of about 0.011, and the results show that the fusion speed is greatly improved under the condition of ensuring the accuracy of wind field.

Key words: XGBoost; HY-2B; CFOSAT; MetOp-B; ERA-5; sea surface wind field