DOI:10.13878/j.cnki.jnuist.2022.06.004

郭佳宾[1]　程丽丹[1]　金双根[2,3]

利用机器学习方法改进风云 3C 星载 GNSS 掩星温度廓线

摘要

本文使用 BP 神经网络、随机森林回归算法,对 2017 年全年风云三号 C 星(FY-3C)GNSS 掩星温度廓线数据进行修正和评估.结果表明:在全球范围内,两种方法均可以修正 GNSS 掩星温度数据,随机森林回归算法的修正效果优于神经网络方法,随机森林回归算法和神经网络方法修正后的结果与再分析数据的平均绝对误差分别为 0.03 K 与 0.32 K,均方误差分别为 0.09 K^2 与 1.02 K^2.将全球按照 10°×10°划分为 324 个网格后,随机森林回归算法对平均绝对误差与均方误差修正的正向收益分别为 97.53%与 92.9%,神经网络方法对平均绝对误差与均方误差修正的正向收益分别为 75.61%与 67.9%.

关键词

GNSS 掩星;温度廓线;随机森林;FY-3C;神经网络

中图分类号 P228.4

文献标志码 A

收稿日期 2022-07-31

资助项目 河南省农业气象保障与应用技术重点实验室应用技术研究基金(KM202224)

作者简介

郭佳宾,男,研究方向 GNSS 气象学.1292196066@qq.com

程丽丹(通信作者),女,高级工程师,研究方向大气遥感.158107010@qq.com

1 河南省气象灾害防御技术中心,郑州,450003

2 南京信息工程大学 遥感与测绘工程学院,南京,210044

3 河南理工大学 测绘与国土信息工程学院,焦作,454000

0 引言

全球卫星导航系统(GNSS)无线电掩星技术利用导航卫星与低轨卫星之间的信号延迟来反演全球高精度大气参数,在大气探测和气象预报中具有重要的应用前景[1].1995 年,美国成功进行了 GPS/MET 探测计划,首次证明了大气掩星探测的可行性[2].2001 年,德国发射了 CHAMP 卫星,该卫星搭载的掩星载荷更为先进,在掩星资料的数量以及资料精度上都有了较大改进[3-4].2006 年,中国台湾和美国联合研制的 COSMIC 卫星成功发射,该星座共有在轨卫星 6 颗[5].2012 年 9 月,欧洲气象卫星组织正式发射了 METOP-B 星[6].2013 年 8 月,韩国发射了 KOMPSAT-5 卫星[7].2018 年 11 月,欧洲气象卫星组织又再次发射了 METOP-C 星.在 COSMIC 取得巨大成功后,美国与中国台湾再次合作,开展了 COSMIC-2 计划,并于 2019 年 6 月下旬发射[8].2013 年 9 月,我国发射了 FY-3C 卫星.FY-3C 星上新增的 GNOS 载荷是国内第一个星上 GNSS 无线电掩星探测仪,该载荷可以同时接收北斗与 GPS 信号,从而大大提升了探测能力[9].

GNSS 掩星探测技术拥有全天时、高精度、高分辨率等优势,但搭载低轨卫星数量少,数据空间分辨率低于传统再分析资料,且在较低高度上,由于水汽以及折射、超折射现象的存在,导致掩星数据质量较差.廖蜜等[10]研究证明了 FY-3C 的中性大气折射率产品的精度基本能够达到预定目标;徐晓华等[11]将 FY-3C 掩星数据与 IGRA2 探空资料进行比较,证明了两种资料的一致性,但存在一定的差异;魏晋德[12]通过对 FY-3C 的掩星产品质量进行研究,证明了产品的可靠性,并使用相关产品对对流层顶特征进行了相关研究.上述文献均指出了 FY-3C 掩星数据的质量问题,但并未提出对数据质量进行改进的方法.GNSS 掩星数据量大,对其精度进行分析时,通常是一个统计平均的结果.因此可以使用机器学习方法对掩星廓线数据进行修正.本文将 FY-3C 的温度廓线数据与 ERA5 再分析数据作为输入值,分别使用神经网络方法和随机森林回归算法对其进行修正,并对修正结果做出评价.

1 观测数据与方法

1.1 观测数据

1.1.1 GNSS 掩星数据

本文所采用的数据是由风云数据网提供的 2017 年 1 月 1 日—12

月 31 日 FY-3C 的 L2 温度廓线数据,其中 6 月 1—31 日没有数据.图 1 展示了 2017 年 3 月 1—7 日的掩星事件在中国区域的分布状况.

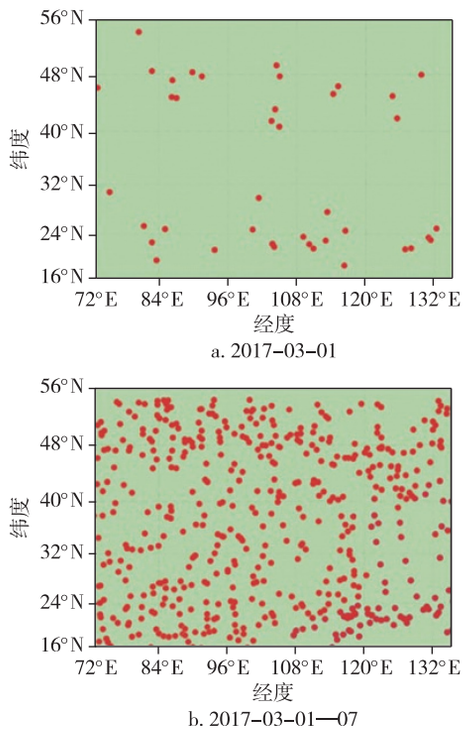


图 1　2017 年 3 月 1 日与 3 月 1—7 日掩星事件在中国区域分布状况

Fig. 1　Distribution ofradio occultation events in China on March 1,2017 and during March 1-7,2017

1.1.2　ERA5 再分析数据

ERA5 再分析数据的前身是 ERA-Interim[13-14],是由欧盟提供资助、ECMWF(欧洲中期天气预报中心)进行运营的新一代再分析资料[15].在此之前,再分析资料已经历经了 FGGE、ERA-15、ERA-40 等产品[16].ERA5 再分析数据水平分辨率为 0.25°×0.25°,垂直分辨率为 37 层,时间分辨率为 1 h.本文使用的是 150 hPa 的 ERA5 数据,其高度在 10 km 左右.

1.2　机器学习方法

1.2.1　神经网络方法

BP 神经网络方法可以学习与存储较多的输入-输出模式的映射关系,且无需事先知道这种映射关系的数学方程.BP 神经网络的拓扑结构中包括输入层、隐层以及输出层.首先在输入层输入学习样本,然后使用反向传播方法,不断地计算每个节点的权值与偏差,并进行调整,使输出层的值与预期值尽可能靠拢.当输出值与预期值满足设定条件时,保存整个网络的权值与偏差[17].本文的输入层、隐层以及输出层关系如图 2 所示.

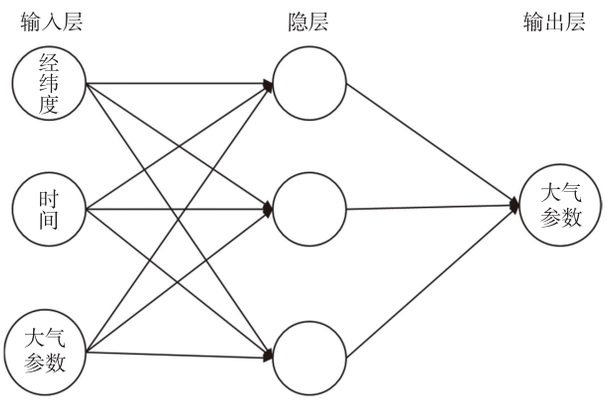


图 2　神经网络方法原理

Fig. 2　Principle of neural network algorithm

1.2.2　随机森林

随机森林是指利用多棵树对样本进行训练,并预测的一种分类器.随机森林回归算法对于多种资料,可以产生高准确度的分类器,可以处理大量的输入变数.在存在 N 个数据的样本集中,每个样本的输入特征向量都有 k 个特征,通过依次有放回的抽样得到它们的子样本集,将子样本集带入决策树中,这样每棵决策回归树会随机选取特征,进而通过训练得到一系列回归结果,再对这些回归结果取平均得到最终的回归结果[18],以此来降低回归方差.随机森林回归算法结构如图 3 所示.

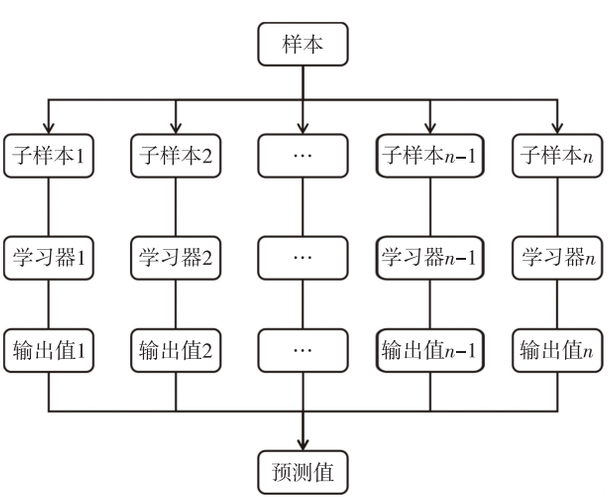


图 3　随机森林回归算法结构示意

Fig. 3　Structure of random forest regression algorithm

1.3　数据处理

1.3.1　GNSS 掩星数据与再分析资料处理

使用机器学习算法对掩星数据修正前,要对掩

星数据和再分析数据的时空特征进行匹配,生成若干组数据对.具体匹配规则为:时间间隔 1 h;空间上选择距离掩星点最近点的 ERA5 温度数据.将经纬度、时间等数据进行归一化处理[19],处理规则如下:

$$I_{\mathrm{lat}}=\frac{I_{\mathrm{lat,ro}}}{90}, \tag{1}$$

$$I_{\mathrm{lon}}=\frac{I_{\mathrm{lon,ro}}}{90}, \tag{2}$$

$$I_{\mathrm{time}}=\frac{I_{\mathrm{time,ro}}}{86\ 400}, \tag{3}$$

其中:$I_{\mathrm{lat,ro}}$为掩星事件的纬度信息;I_{lat}为归一化的掩星事件的纬度信息;$I_{\mathrm{lon,ro}}$为掩星事件的经度信息;I_{lon}为归一化的掩星事件的经度信息;$I_{\mathrm{time,ro}}$为掩星事件的时间信息;I_{time}为归一化的掩星事件的时间信息.

1.3.2 机器学习参数设置

在经过数据时空特征匹配后,随机选取 80%的数据对组成训练集,剩下的 20%数据对组成测试集.从图 4 可以看到,训练集与测试集具有相似的纬度分布特征.

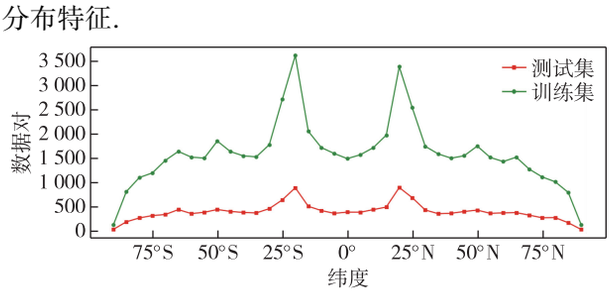


图 4 掩星数据训练集与测试集纬度特征分布

Fig. 4 Latitude distribution of training set and test set of radio occultation data

神经网络模型选择 5 层全连接的神经网络,每个隐藏层设置 10 个神经元,损失函数设置为 mse,参数更新采用 Adam 方法.随机森林回归模型中设置了 100 棵树,且不限制每棵决策树的树最大深度和最大叶节点数目,将决策树放入随机森林避免过拟合.

将全球化分为 18×18 个网格,即 10°(lat)×10°(lon).计算每一个网格的平均绝对误差与均方误差.

$$T_{\mathrm{mae}}=\frac{1}{N}\sum_{t=1}^{N}\left|\left(T_{\mathrm{ro},t}-T_{\mathrm{rea5}.t}\right)\right|, \tag{4}$$

$$T_{\mathrm{mse}}=\frac{1}{N}\sum_{t=1}^{N}\left(T_{\mathrm{ro},t}-T_{\mathrm{rea5}.t}\right)^{2}, \tag{5}$$

式中:T_{mae}是该网格的温度平均绝对误差;T_{mse}是该网格的温度的均方误差;$T_{\mathrm{ro},t}$是网格内任一掩星廓线的温度值;$T_{\mathrm{rea5},t}$为对应的再分析资料的温度值;N 为该网格内数据对的个数.

$$\Delta T_{\mathrm{mae}}=T_{\mathrm{mae,ro-ec}}-T_{\mathrm{mae,pre-ec}}, \tag{6}$$

$$\Delta T_{\mathrm{mse}}=T_{\mathrm{mse,ro-ec}}-T_{\mathrm{mse,pre-ec}}, \tag{7}$$

式中:$T_{\mathrm{mae,ro-ec}}$为网格内未修正前的掩星数据与再分析数据的平均绝对误差;$T_{\mathrm{mae,pre-ec}}$为使用相应方法修正后的掩星数据与再分析数据的平均绝对误差;$T_{\mathrm{mse,ro-ec}}$为网格内未修正前的掩星数据与再分析数据的均方误差;$T_{\mathrm{mse,pre-ec}}$为使用相应方法修正后的掩星数据与再分析数据的均方误差;ΔT_{mse}为修正前后均方误差的差值,该值越大表明修正效果越好,反之则修正效果越差;ΔT_{mae}为修正前后平均绝对误差的差值,该值越大表明修正效果越好,反之则修正效果越差.

2 结果与分析

图 5 为不同纬度带上神经网络方法与随机森林回归算法对 FY-3C 掩星数据的修正结果.可以看到,在全球范围内,两种方法都可以对掩星数据进行修

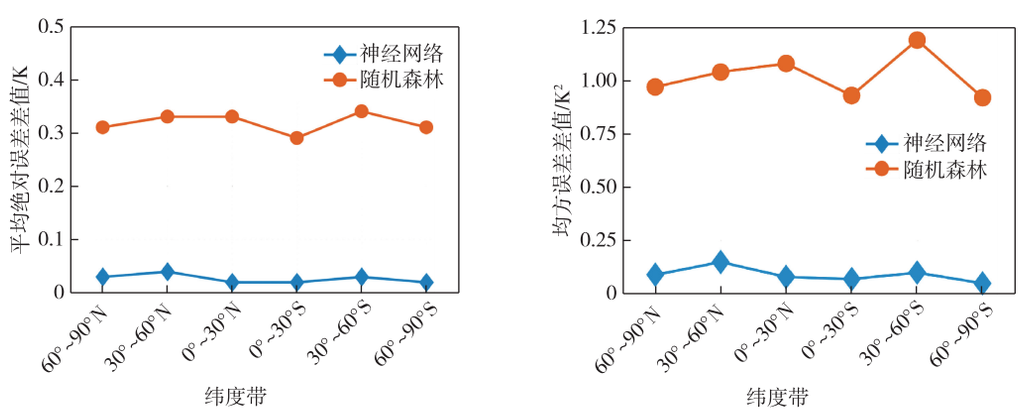


图 5 不同纬度带平均绝对误差差值与均方误差差值

Fig. 5 Differences of MAE and MSE at different latitudes

正,且随机森林算法的修正效果远胜神经网络方法.

两种方法在中纬度地区的修正效果要优于其他两个纬度带.北半球的修正效果略优于南半球的修正效果,这是 FY-3C 星自身原因造成的:北半球的廓线数据略多于南半球,更多的数据意味着更多的样本与特征,能让模型对经纬度参数更加敏感.

2.1 高纬度地区

从表 1 可以看出,在高纬度地区,使用神经网络方法修正后的温度数据均方误差与平均绝对误差,北半球的正向收益均大于南半球.随机森林回归算法的南北半球修正结果较为一致.

表 1 高纬度地区两种方法修正结果

Table 1 Statistics of correction results of two methods in high latitudes

方法	ΔT_{mae}/K		ΔT_{mse}/K^2	
	南半球	北半球	南半球	北半球
神经网络	0.02	0.03	0.05	0.09
随机森林	0.31	0.31	0.92	0.97

从图 6、7 得知,在高纬度地区的 108 个网格中,经过神经网络与随机森林修正后的掩星温度数据大部分具有正向收益,且随机森林回归算法的修正效果远高于神经网络方法.神经网络与随机森林回归算法对平均绝对误差的正向修正率分别为 74.07%与 96.3%,对均方误差的正向修正率分别为 66.67%与 90.74%.

2.2 中纬度地区

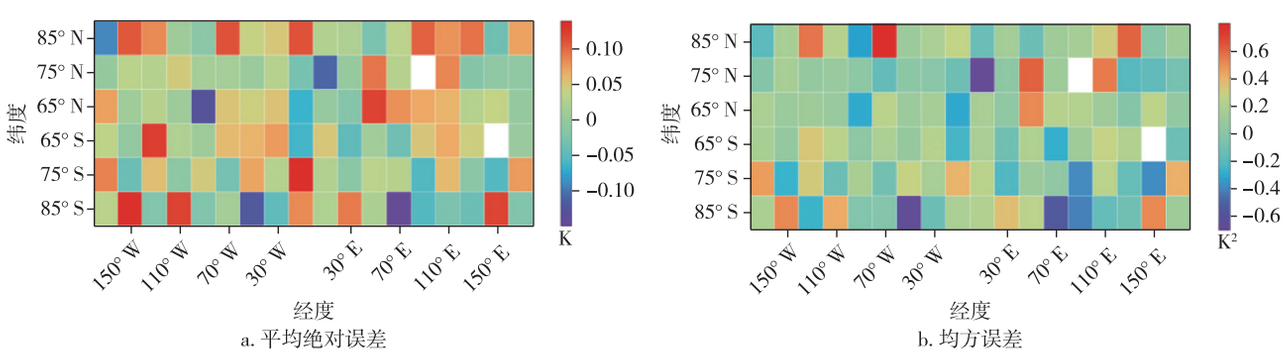
从表 2 可以看出,在中纬度地区,两种方法的修正结果都具有正向收益.在每项修正指标中,随机森林回归算法的修正效果约为神经网络方法的 10 倍.

表 2 中纬度地区两种方法修正结果

Table 2 Statistics of correction results of two methods in middle latitudes

方法	ΔT_{mae}/K		ΔT_{mse}/K^2	
	南半球	北半球	南半球	北半球
神经网络	0.03	0.04	0.10	0.15
随机森林	0.34	0.33	1.19	1.04

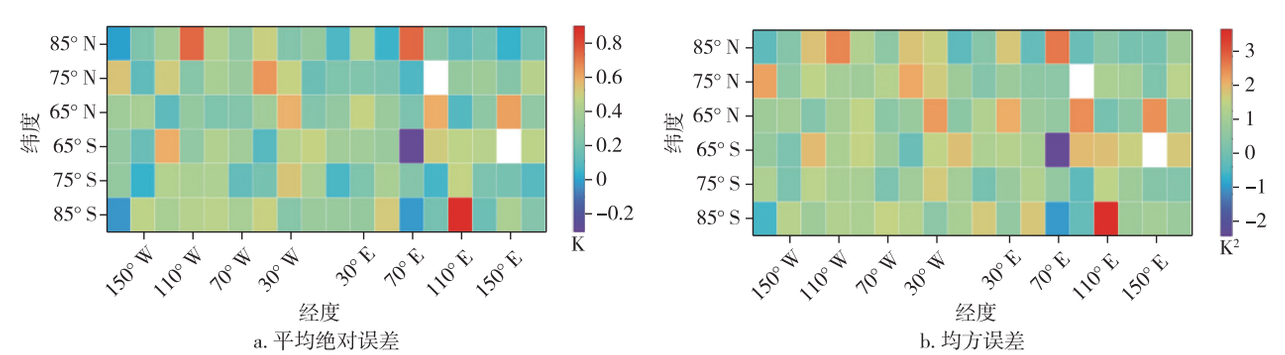
从图 8 可以看到,均方误差和平均绝对误差的差值范围集中在-0.4~0.6 K^2 与-0.1~0.15 K 之间,相比于修正前的结果提升不大.对均方误差与平均绝对误差的修正率分别为 70.37%与 80.55%.



a. 平均绝对误差　　b. 均方误差

图 6 高纬度地区神经网络方法对平均绝对误差与均方误差的修正结果

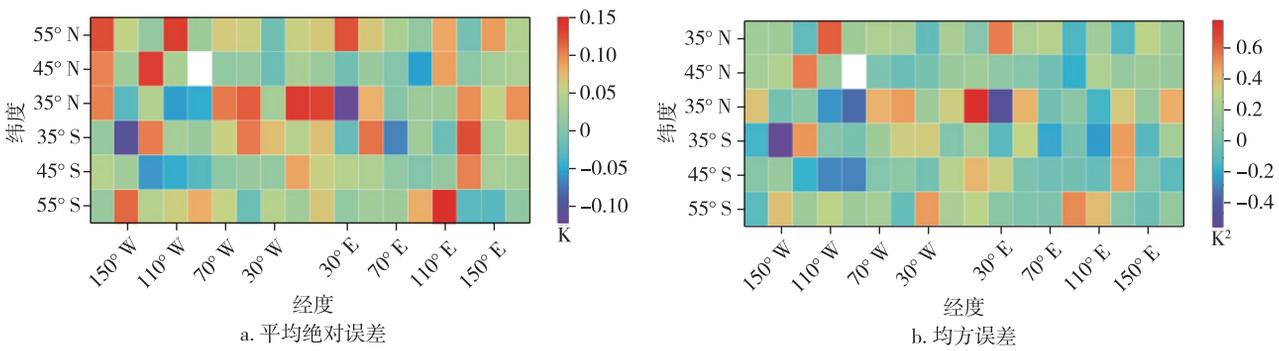
Fig. 6 Correction of MAE and MSE by neural network in high latitudes



a. 平均绝对误差　　b. 均方误差

图 7 高纬度地区随机森林回归算法对平均绝对误差与均方误差的修正结果

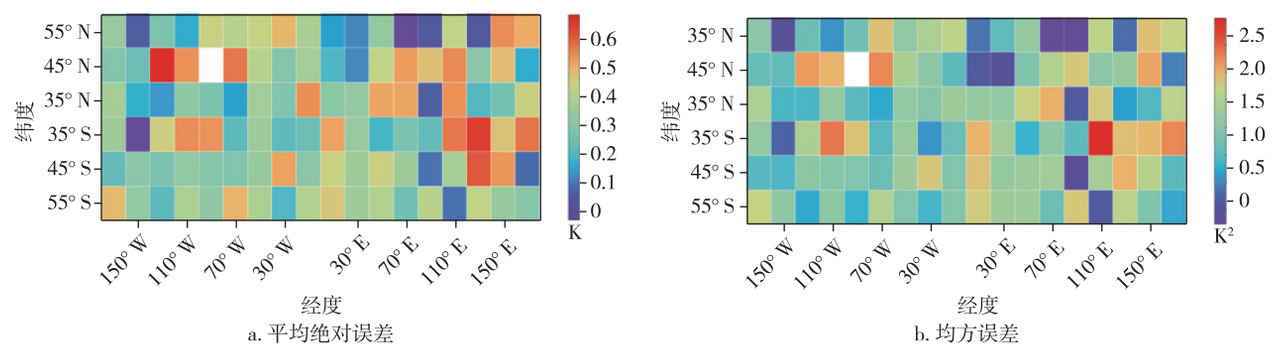
Fig. 7 Correction of MAE and MSE by random forest regression in high latitudes



a. 平均绝对误差　　b. 均方误差

图 8　中纬度地区神经网络方法对平均绝对误差与均方误差的修正结果

Fig. 8　Correction of MAE and MSE by neural network in middle latitudes



a. 平均绝对误差　　b. 均方误差

图 9　中纬度地区随机森林回归算法对平均绝对误差与均方误差的修正结果

Fig. 9　Correction of MAE and MSE by random forest regression in middle latitudes

从图 9 可以看到,使用随机森林回归算法后,均方误差与平均绝对误差的差值范围分别集中于 0~2.5 K^2 与 0~0.6 K.对均方误差与平均绝对误差的修正率分别为 92.59%与 98.15%.

2.3　低纬度地区

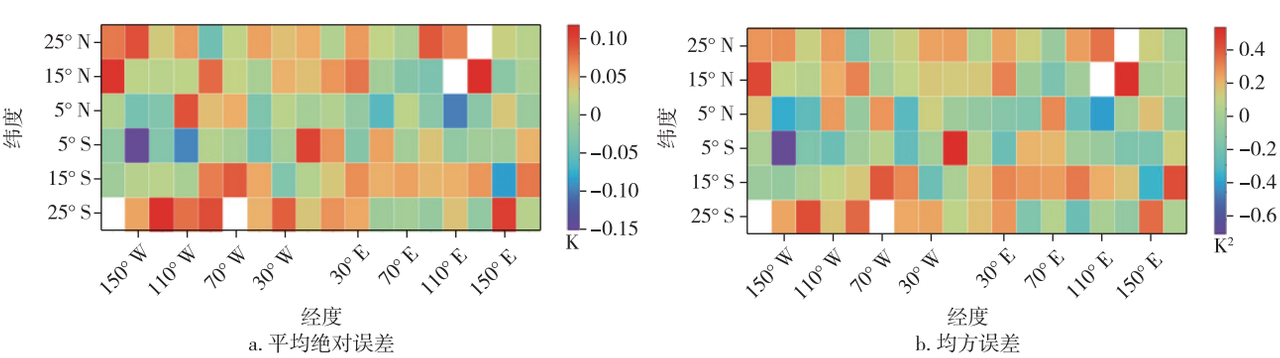
从图 10 可以看到,对均方误差与平均绝对误差的修正率分别为 66.67%与 72.22%,且在某一块区域整体呈现为正向收益与负向收益.如 5°S~5°N 处大部分表现为负收益,25°S 与 25°N 处表现为正收益.低纬度地区两种方法修正结果如表 3 所示.

从图 11 可以看到,在低纬度地区,随机森林回

表 3　低纬度地区两种方法修正结果

Table 3　Statistics of correction results of two methods in low latitudes

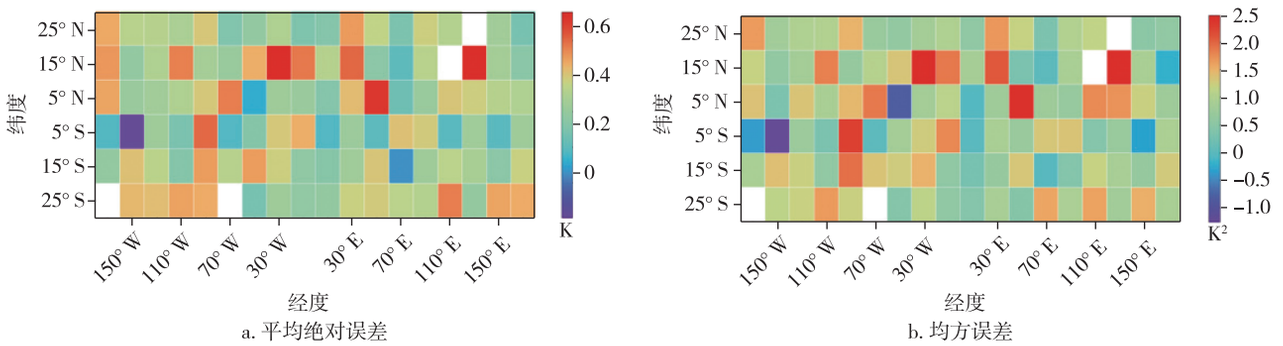
方法	ΔT_{mae}/K		ΔT_{mse}/K^2	
	南半球	北半球	南半球	北半球
神经网络	0.02	0.02	0.07	0.08
随机森林	0.29	0.33	0.93	1.08



a. 平均绝对误差　　b. 均方误差

图 10　低纬度地区神经网络方法对平均绝对误差与均方误差的修正结果

Fig. 10　Correction results of MAE and MSE by neural network in low latitudes



a. 平均绝对误差　　b. 均方误差

图 11　低纬度地区随机森林回归算法对平均绝对误差与均方误差的修正结果

Fig. 11　Correction of MAE and MSE by random forest regression in low latitudes

归算法对均方误差与平均绝对误差的修正率分别为 95.37%与 98.15%,且随机森林回归算法的正向收益与负向收益的分布没有明显的分布规律.

3　结论

本文采用神经网络方法和随机森林回归算法对 2017 年 FY-3C 掩星廓线的温度数据进行修正和评估,按照 10°×10°将全球划分为 324 个网格计算有效修正率,对两种修正效果的空间分布特征进行研究,得到如下结论:

1)神经网络方法与随机森林回归算法均可以对 FY-3C 掩星温度数据进行修正,其中随机森林回归算法对平均绝对误差与均方误差的正向修正率超过 90%,神经网络方法对平均绝对误差与均方误差的正向修正率超过 66.67%.

2)将修正结果按照高中低三个纬度划分,随机森林回归算法对三个纬度带的平均绝对误差的正向修正率分别为 96.3%、98.15%和 98.15%;均方误差的正向修正率分别为 90.74%、92.59%和 95.37%.神经网络方法对三个纬度带的平均绝对误差的正向修正率分别为 74.07%、80.55%和 72.22%;均方误差的正向修正率分别为 66.67%、70.37%和 66.67%.

3)神经网络方法和随机森林回归算法在北半球 GNSS 掩星温度剖面修正效果略优于南半球.

参考文献

References

[1] Gobiet A, Kirchengast G. Advancements of global navigation satellite system radio occultation retrieval in the upper stratosphere for optimal climate monitoring utility [J]. Journal of Geophysical Research: Atmospheres, 2004, 109(D24): D24110

[2] Kursinski E R, Hajj G A, Bertiger W I, et al. Initial results of radio occultation observations of earth's atmosphere using the global positioning system[J]. Science, 1996, 271(5252): 1107-1110

[3] 赵齐乐,刘经南,葛茂荣,等.CHAMP 卫星 cm 级精密定轨[J].武汉大学学报(信息科学版),2006,31(10):879-882
ZHAO Qile, LIU Jingnan, GE Maorong, et al. Precision orbit determination of CHAMP satellite with cm-level accuracy[J]. Geomatics and Information Science of Wuhan University, 2006, 31(10): 879-882

[4] Wickert J, Reigber C, Beyerle G, et al. Atmosphere sounding by GPS radio occultation: first results from CHAMP [J]. Geophysical Research Letters, 2001, 28(17): 3263-3266

[5] 青盛.地基 GPS 水汽反演的研究[D].成都:西南交通大学,2009
QING Sheng. Research on the computation of atmospheric water vapour base on ground-GPS [D]. Chengdu: Southwest Jiaotong University, 2009

[6] Hao N, Koukouli M E, Inness A, et al. GOME-2 total ozone columns from MetOp-A/MetOp-B and assimilation in the MACC system [J]. Atmospheric Measurement Techniques, 2014, 7(9): 2937-2951

[7] Hwang Y, Lee B S, Kim Y R, et al. GPS-based orbit determination for KOMPSAT-5 satellite[J]. ETRI Journal, 2011, 33(4): 487-496

[8] Lin C Y, Lin C C H, Liu J Y, et al. The early results and validation of FORMOSAT-7/COSMIC-2 space weather products: global ionospheric specification and Neaided Abel electron density profile[J]. Journal of Geophysical Research: Space Physics, 2020, 125(10): e2020JA028028

[9] 郭佳宾,金双根.利用 FY-3C 卫星 GNSS 掩星数据分析中国区域对流层顶参数变化[J].大地测量与地球动力学,2021,41(1):21-26
GUO Jiabin, JIN Shuanggen. Variations of tropopause parameters over China from FY-3C GNSS radio occultation observations[J]. Journal of Geodesy and Geodynamics, 2021, 41(1): 21-26

[10] 廖蜜,张鹏,毕研盟,等.风云三号气象卫星掩星大气产品精度的初步检验[J].气象学报,2015,73(6):1131-1140

LIAO Mi, ZHANG Peng, BI Yanmeng, et al. A preliminary estimation of the radio occultation products accuracy from the Fengyun-3C meteorological satellite [J].Acta Meteorologica Sinica,2015,73(6):1131-1140

[11] 徐晓华,朱洲宗,罗佳.利用 IGRA2 探空数据和 COSMIC 掩星资料对 FY-3C 掩星中性大气产品进行质量分析[J].武汉大学学报(信息科学版),2020,45(3):384-393
XU Xiaohua,ZHU Zhouzong,LUO Jia.Quality analysis of the neutral atmospheric products from FY-3C radio occultation based on IGRA2 radiosonde data and COSMIC radio occulation products[J].Geomatics and Information Science of Wuhan University,2020,45(3):384-393

[12] 魏晋德.风云三号 C 星大气掩星数据质量分析及对流层顶特征研究[D].徐州:中国矿业大学,2021
WEI Jinde.An evaluation of FY-3C radio occultation data quality and study of tropopause characteristics [D]. Xuzhou: China University of Mining and Technology,2021

[13] Uppala S,Dee D,Kobayashi S,et al.Towards a climate data assimilation system:status update of ERA-Interim [J].ECMWF Newsletter,2008,115(7):12-18

[14] Dee D P,Uppala S M,Simmons A J,et al.The ERA-Interim reanalysis:configuration and performance of the data assimilation system[J]. Quarterly Journal of the Royal Meteorological Society,2011,137(656):553-597

[15] 孟宪贵,郭俊建,韩永清.ERA5 再分析数据适用性初步评估[J].海洋气象学报,2018,38(1):91-99
MENG Xiangui, GUO Junjian, HAN Yongqing. Preliminarily assessment of ERA5 reanalysis data[J]. Journal of Marine Meteorology,2018,38(1):91-99

[16] Uppala S M, KÅllberg P W, Simmons A J, et al. The ERA-40 re-analysis[J]. Quarterly Journal of the Royal Meteorological Society,2005,131(612):2961-3012

[17] Haykin S.Neural networks and learning machines[M]. 3rd ed.Upper Saddle River,NJ:Prentice Hall,2008

[18] Segal M R.Machine learning benchmarks and random forest regression[J].Center for Bioinformatics & Molecular Biostatistics,2004

[19] Jayalakshmi T,Santhakumaran A.Statistical normalization and back propagation for classification[J].International Journal of Computer Theory and Engineering, 2011, 3(1):1793-8201

Improving temperature profile of FY-3C GNSS radio occultation by machine learning methods

GUO Jiabin[1] CHENG Lidan[1] JIN Shuanggen[2,3]

1 Henan Meteorological Disaster Prevention Technology Center,Zhengzhou 450003

2 School of Remote Sensing & Geomatics Engineering,Nanjing University of Information Science & Technology,Nanjing 210044

3 School of Surveying and Land Information Engineering,Henan Polytechnic University,Jiaozuo 454000

Abstract In this paper,BP neural network and random forest regression algorithm are used to correct the temperature profile data of FY-3C GNSS radio occultation in 2017.The results show that both methods can correct FY-3C radio occultation temperature data,but the performance of the random forest regression is better than that of the neural network.For the random forest regression algorithm and the neural network,the mean absolute errors between the corrected results and the reanalysis data are 0.03 K and 0.32 K,respectively,and the mean square errors are 0.09 K^2 and 1.02 K^2,respectively.When the globe is divided into 324 grids of 10°×10°,the random forest regression algorithm yields positive returns of 97.53% and 92.9% for average absolute error and mean square error corrections,respectively,and neural network produces positive returns of 75.61% for average absolute error correction and 67.9% for mean square error correction.

Key words GNSS radio occultation;temperature profile;random forest;FY-3C;neural network