

文章编号: 1672-8785(2025)09-0041-08

基于卫星红外资料的 PM_{2.5} 信号识别与可解释反演

王 根^{1*} 袁 松² 叶 松¹ 范传宇¹

(1. 巢湖学院电子工程学院, 安徽 合肥 238024;

2. 安徽省气象台, 安徽 合肥 230031)

摘 要: 直径小于 2.5 μm 的地面细颗粒物 (Particulate Matter, PM)——PM_{2.5} 对人类健康和社会经济产生了负面影响。大多数方法都是从卫星反演的气溶胶光学厚度间接产品或白天大气顶部反射率中获取 PM_{2.5} 数据。本文旨在直接使用风云四号 B 星多通道扫描成像辐射计 (Advanced Geosynchronous Radiation Imager, AGRI) 红外资料和人工智能模型, 近实时全时段 (包括白天和夜晚) 对江淮地区进行 4 km 空间分辨率和 15 min 时间分辨率的 PM_{2.5} 反演。首先分季节探讨了 AGRI 亮温对不同量级 PM_{2.5} 的信号响应; 其次, 基于随机森林方法, 分季节开展了 AGRI 亮温反演 PM_{2.5} 研究。试验结果表明, 四个季节反演的 PM_{2.5} 相关系数均超过 0.87。最后, 基于 SHapley 加性预测实现了模型可解释 (地理信息对 PM_{2.5} 的贡献率较大), 并进一步探讨了文中产品的应用。

关键词: 风云四号 B 星; PM_{2.5} 反演; 人工智能; SHAP 可解释; 江淮地区

中图分类号: P407 **文献标志码:** A **DOI:** 10.3969/j.issn.1672-8785.2025.09.006

PM_{2.5} Signal Identification and Interpretable Retrieval Based on Satellite Infrared Data

WANG Gen^{1*}, YUAN Song², YE Song¹, FAN Chuan-yu¹

(1. School of Electronic Engineering, Chaohu University, Hefei 238024, China;

2. Anhui Meteorological Observatory, Hefei 230031, China)

Abstract: Ground-level fine particulate matter with a diameter less than 2.5 μm (PM_{2.5}) has negative impacts on human health and social economy. Most methods obtain PM_{2.5} data from indirect products of aerosol optical depth retrieved from satellites or daytime top-of-atmosphere reflectivity. This paper aims to directly utilize infrared data from the advanced geosynchronous radiation imager (AGRI) on the Fengyun-4B satellite and arti-

收稿日期: 2025-04-08

基金项目: 安徽省高校杰出青年科研项目 (2022AH020093); 安徽省自然科学基金项目 (2408085MD102); 巢湖学院学科建设质量提升工程项目 (XLZ202404); 安徽省重点研究与开发计划项目 (2022h11020002); 安徽省教育厅科学研究项目 (KJ2021A1027); 巢湖学院科研基金资助项目 (KYQD-202211)

作者简介: 王根 (1983-), 男, 江苏姜堰人, 高级工程师, 主要从事卫星遥感定量化应用、智能优化与控制、大数据与人工智能可解释性等方面的研究。

***通讯作者:** E-mail: 203wanggen@163.com

cial intelligence models to perform near-real-time $PM_{2.5}$ retrievals over the Yangtze-Huaihe region throughout the entire day (including day and night), with a spatial resolution of 4 km and a temporal resolution of 15 minutes. Firstly, the seasonal signal response of AGRI brightness temperature to different $PM_{2.5}$ levels is explored. Secondly, a study on AGRI brightness temperature retrieval of $PM_{2.5}$ is conducted based on the random forest method in different seasons. The experimental results show that the correlation coefficients for $PM_{2.5}$ retrievals in all four seasons exceed 0.87. Finally, the model is interpretable using SHapley additive prediction (given the significant contribution of geographic information to $PM_{2.5}$), and the application of the proposed product is further explored.

Key words: Fengyun-4B satellite; $PM_{2.5}$ retrieval; artificial intelligence; SHAP interpretable; Yangtze-Huaihe region

0 引言

长期以来,空气污染一直是影响人类健康和生态系统的世界性问题之一^[1]。直径小于 $2.5\ \mu\text{m}$ 的地面细颗粒物($PM_{2.5}$)是典型的空气污染物,已被世界卫生组织和许多国家列入空气质量监测系统^[2]。目前, $PM_{2.5}$ 资料主要来源于地面观测和卫星遥感数据反演。环境质量监测站是分析 $PM_{2.5}$ 的最直接站点^[3]。与地面监测数据相比,卫星遥感资料具有空间分辨率高、监测范围广、全天候实时监测等特点,结合人工智能机器学习方法,可以克服因地面监测数据缺乏或时空分布不均而出现的不足,为 $PM_{2.5}$ 相关研究提供新思路^[4-5]。

基于卫星遥感的气溶胶光学厚度(Aerosol Optical Depth, AOD)已被广泛用于 $PM_{2.5}$ 反演^[6]。学者们提出不同方法来描述 $PM_{2.5}$ 与 AOD 之间的关联,包括物理模型^[7]、化学模型^[8]和统计模型^[9]。例如, Li T W 等^[10]提出了知识信息神经网络模型,使用卫星观测数据、再分析数据和地面站测量数据,联合反演了 $PM_{2.5}$ 和臭氧。Xue Y 等^[11]提出了改进的地理和时间加权回归模型,使用 Himawari-8/AOD 产品生成了华中和华东地区的 $PM_{2.5}$ 数据。

随机森林、梯度提升树和极端梯度提升树等机器学习方法已被广泛用于估计局部和全球 $PM_{2.5}$ ^[12-13]。Yu X Y 等^[3]结合地面环境监测站 $PM_{2.5}$ 数据、卫星 AOD 产品、气象数据等,利用时空加权法改进了 LightGBM 模型,得到了

较高的 $PM_{2.5}$ 反演精度。但其存在的问题是, AOD 反演过程中的误差和不确定性将不可避免地引入 $PM_{2.5}$ 反演中。Shen H F 等^[14]提出,直接利用从卫星可见光波段获得的大气顶部反射率(Top-of-Atmosphere Reflectance, TOAR)资料反演 $PM_{2.5}$,而不使用中间 AOD 产品,但此方法仅适用于白天。

综上所述,为了避免中间 AOD 产品引入的误差,并使卫星资料能够近实时全时段(白天和夜晚)反演 $PM_{2.5}$ 资料,本文提出了一种基于静止卫星红外资料的 $PM_{2.5}$ 反演方法,并对其进行了可解释性分析。

1 模型和方法

1.1 卫星资料反演 $PM_{2.5}$ 的基本框架

基于机器学习的卫星资料反演 $PM_{2.5}$ 的本质是在输入量(自变量、特征变量)与输出量(因变量)之间建立传递函数。其关系定义为

$$O_{PM_{2.5}} = f(X_S) + error \quad (1)$$

式中, $O_{PM_{2.5}}$ 为模型输出量; $f(\bullet)$ 为采用机器学习回归模型建立的传递函数; X_S 为模型输入量(自变量),本文中 X_S 表示 FY-4B/AGRI 红外通道亮温、亮温衍生变量和地理信息(经度和纬度)。

1.2 随机森林

随机森林是基于分类和回归树的算法集合^[15]。作为目前卫星资料反演大气廓线和降水中常用的数据挖掘方法之一^[16],随机森林中的每个独立树都是由训练样本和输入变量的随机选择子集创建的。对于回归问题,将多个

独立树的结果求平均值, 以产生随机森林的输出量。该算法有两个需要调整的基本模型超参数: 树的数量 ($n_estimators$) 和树的最大深度 (max_depth)。文中使用默认值设置除这两个参数以外的其他参数。为了得到两个超参数最优或次优组合, 本文基于均方误差优化此类参数。

1.3 模型 SHAP 可解释

SHapley 加法解释 (SHapley Additive explanations, SHAP) 是一种事后解释方法^[17]。SHAP 将所有特征视为“贡献者”, 并通过计算特征对模型输出的边际贡献来构建加法模型。SHAP 可以解释机器学习能够预测模型的原因, 甚至能够揭示一些“黑匣子”模型内部的信息。基于博弈论, 通过计算模型预测中的变化来评估每个特征的影响或贡献程度。SHAP 的定义为

$$P_i = \sum_{S \subseteq F(i)} \frac{|S|!(|F|-|S|-1)!}{|F|!} \times [F_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_s(x_s)] \quad (2)$$

式中, P_i 表示特征 i 的 SHAP 值; F 表示全部特征集合; $S \subseteq F(i)$ 表示 F 中剔除索引 i 特征后的一个子集 S ; $|S|$ 表示子集 S 中的特征数量; x_s 、 f_s 表示仅包含 S 中变量的模型输入值。

1.4 模型精度评估方法

使用皮尔逊相关系数 (Correlation Coefficient, CC) 和均方根误差 (Root Mean Square Error, RMSE) 作为 $PM_{2.5}$ 反演精度评估的标准。

CC 定义为

$$CC = \frac{\sum_{k=1}^m (S_k - \bar{S})(R_k - \bar{R})}{\sqrt{\sum_{k=1}^m (S_k - \bar{S})^2} \sqrt{\sum_{k=1}^m (R_k - \bar{R})^2}} \quad (3)$$

RMSE 定义为

$$RMSE = \sqrt{\frac{1}{m} \sum_{k=1}^m (S_k - R_k)^2} \quad (4)$$

式中, m 是匹配的样本总数量; S_k 是通过 FY-4B/AGRI 等资料反演的 $PM_{2.5}$; R_k 表示地面观测或目标 $PM_{2.5}$; $\bar{S}$ 和 $\bar{R}$ 分别表示其平均值。

2 模型特征构建

2.1 研究区域

江淮地区在我国社会和经济发展中起到较为重要的作用^[18]。本文聚焦江淮地区 $PM_{2.5}$ 反演研究。如图 1 所示, 该地区的经度范围为 $110^\circ E \sim 123^\circ E$, 纬度范围为 $28^\circ N \sim 34^\circ N$ 。图 1 中的背景为风云四号 B 星 (FY-4B)/先进静止轨道辐射成像仪 (Advanced Geostationary Radiation Imager, AGRI) 在 2022 年 6 月 19 日 01 时获得的红外窗区亮温分布。

2.2 数据来源

FY-4B/AGRI 有 15 个通道。本文的目的是基于 FY-4B/AGRI 数据近实时全天候反演 $PM_{2.5}$, 故仅考虑红外通道 8~15 的亮温。通道中心波长分别为 $3.75 \mu m$ 、 $6.25 \mu m$ 、 $6.95 \mu m$ 、 $7.42 \mu m$ 、 $8.55 \mu m$ 、 $10.8 \mu m$ 、 $12.0 \mu m$ 和 $13.3 \mu m$ 。FY-4B/AGRI 数据的时间分辨

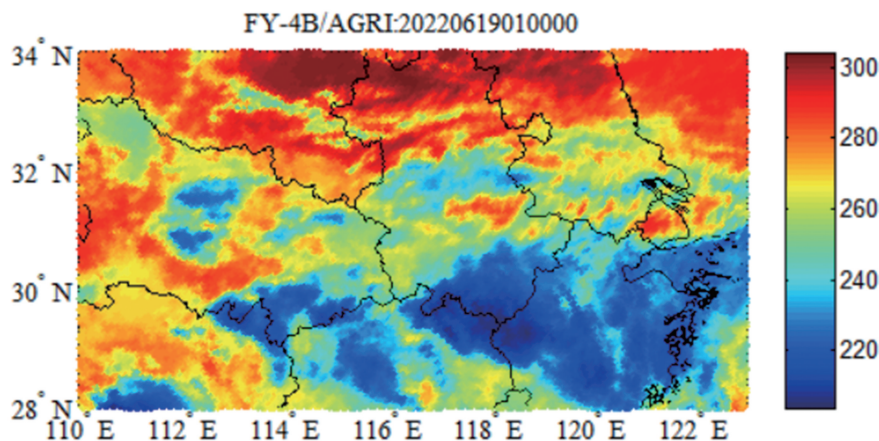


图 1 文中 $PM_{2.5}$ 反演研究区域

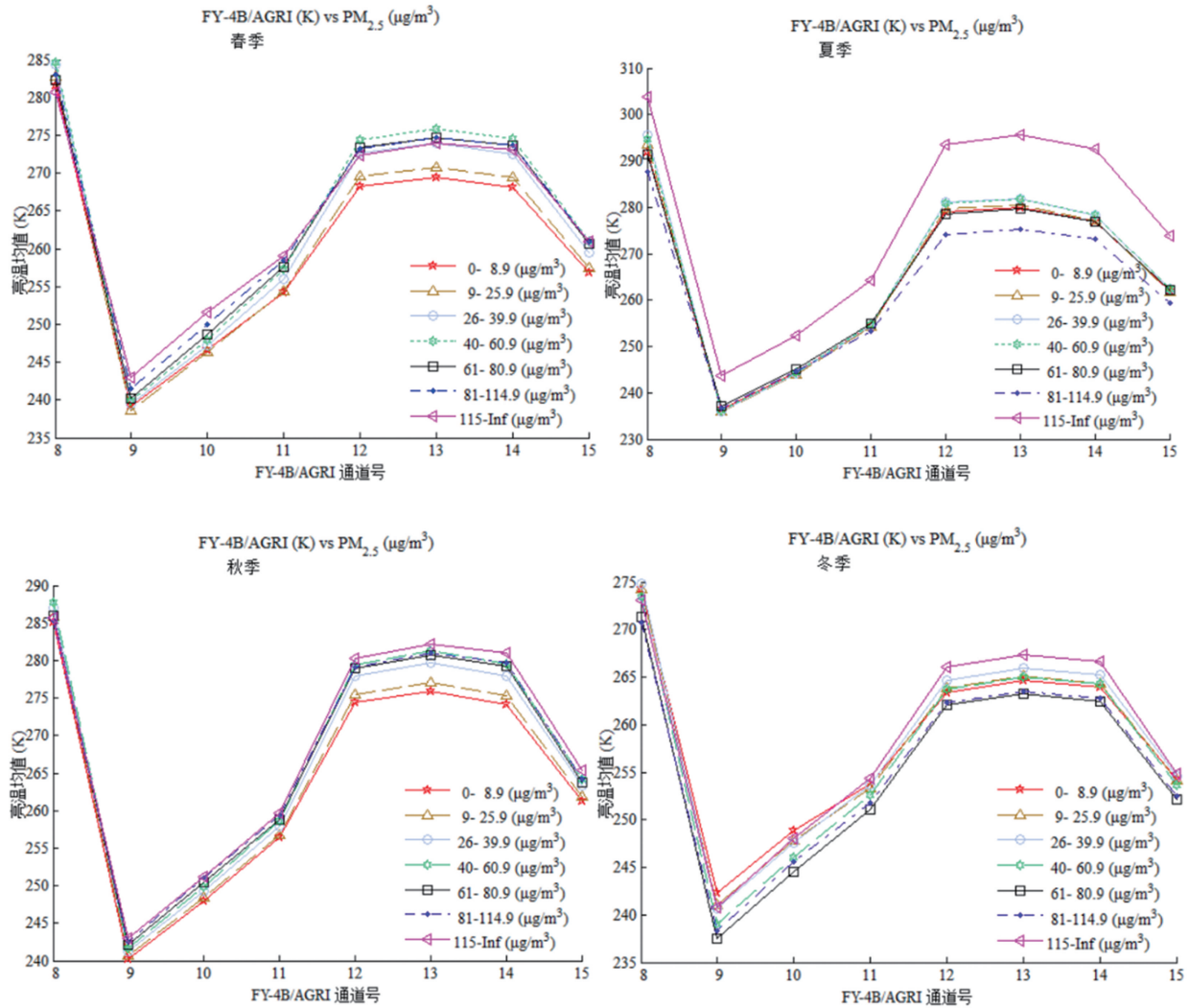


图 2 不同 $PM_{2.5}$ 等级下的 FY-4B/AGRI 红外通道亮温均值分布

率为 15 min, 空间分辨率为 4 km。资料来源网址为 <http://satellite.nsmc.org.cn/portalsite/default.aspx?currentculture=en-US>。

本文将“真实性”较高的地面 $PM_{2.5}$ 观测作为因变量进行相关研究。

2.3 模型特征构建

对于数据的时间分辨率, 选取 FY-4B/AGRI 和地面 $PM_{2.5}$ 观测的小时数据构建机器学习样本, 例如 00:00 和 01:00 (协调世界时) 的数据。对于数据的空间分辨率, 采用最近邻方法将 FY-4B/AGRI 红外资料及相关模型特征输入量插值到 $PM_{2.5}$ 地面站, 并将其作为机器学习样本。

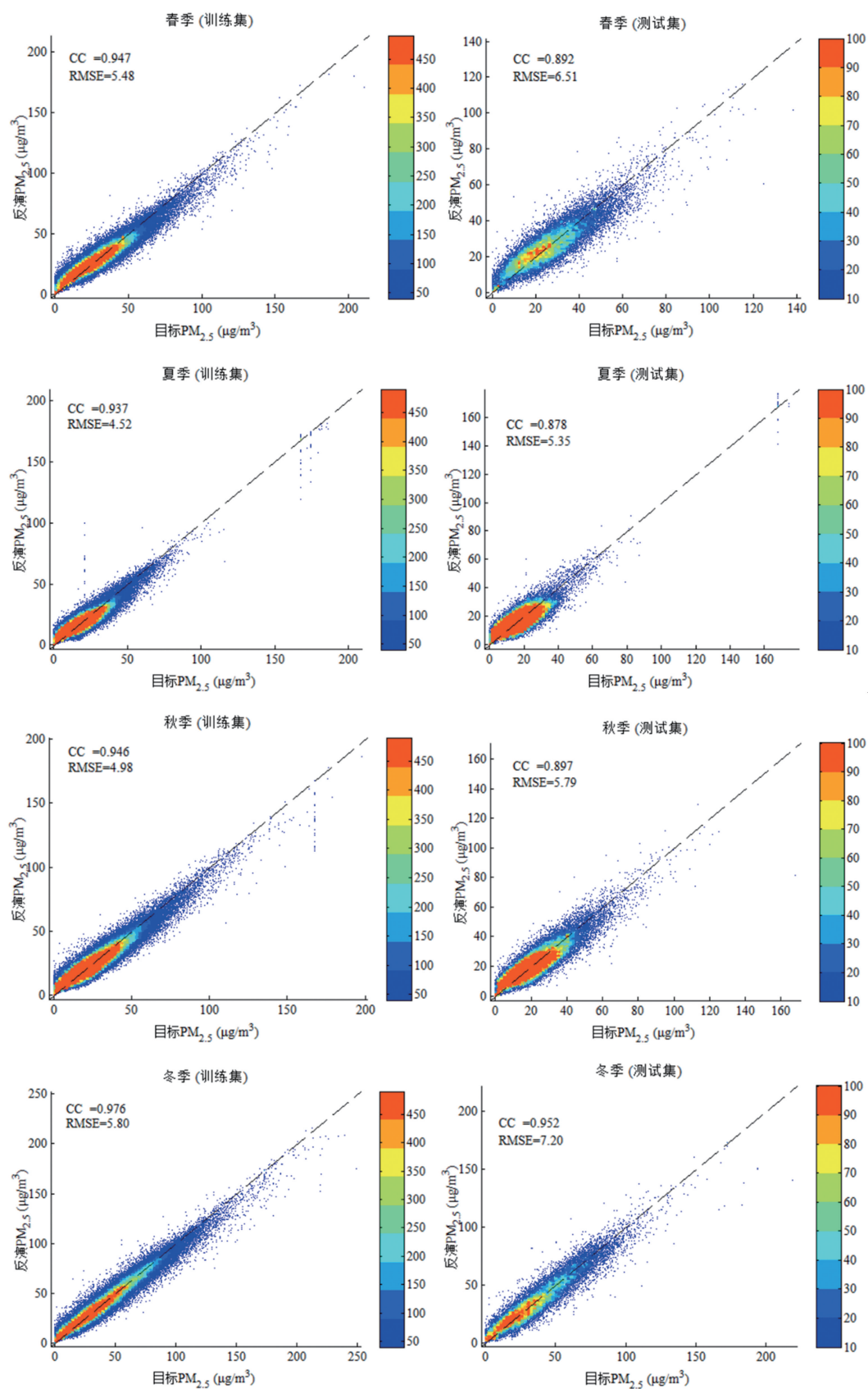
文中模型的输入量为 FY-4B/AGRI 通道 8~15 亮温、通道 8~15 亮温的衍生变量以及

地理信息(经度和纬度)。通道 8~15 亮温的衍生变量定义如下: 通道 8 亮温与通道 9 亮温至通道 15 亮温之差、通道 9 亮温与通道 10 亮温至通道 15 亮温之差; 依此类推到通道 14 亮温与通道 15 亮温之差, 共计 28 个通道亮温衍生变量。

3 红外亮温 $PM_{2.5}$ 信号识别与可解释反演试验分析

3.1 基于 FY-4B/AGRI 红外通道亮温的 $PM_{2.5}$ 信号识别

本文选取 2022 年整年的 FY-4B/AGRI 数据用于识别 $PM_{2.5}$ 信号分析。其中, 3~5 月为春季, 6~8 月为夏季, 9~11 月为秋季, 12~2 月为冬季^[5]。

图 3 FY-4B/AGRI 红外亮温反演 PM_{2.5} 的结果图

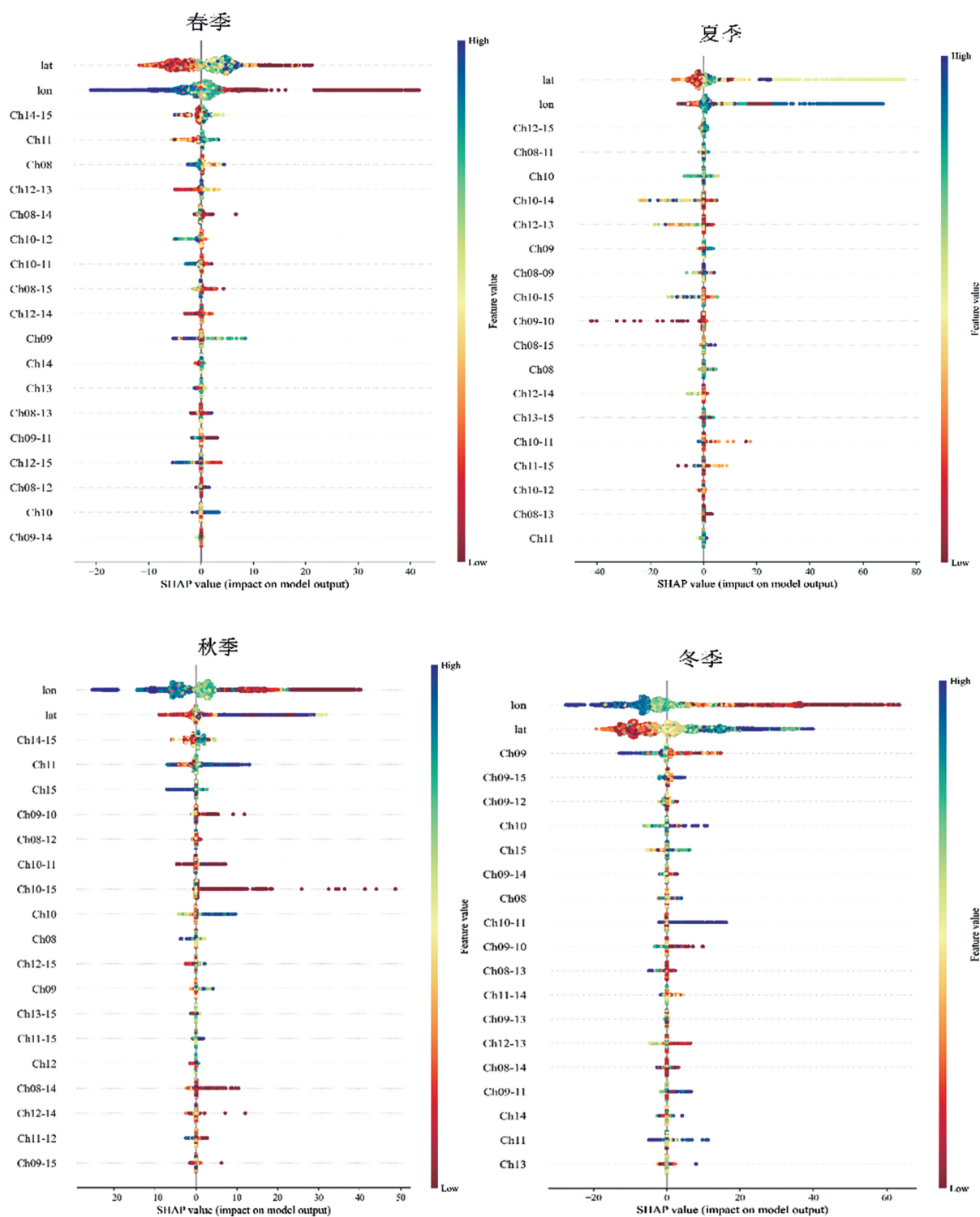


图 4 不同季节反演 $PM_{2.5}$ 的 SHAP 可解释性分析

为分析 FY-4B/AGRI 通道 8~15 亮温对不同 $PM_{2.5}$ 量级的响应，文中将 $PM_{2.5}$ 分为 7 个等级进行统计。等级分别为 0~8.9、9~25.9、26~39.9、40~60.9、61~80.9、81~114.9、115~ $+\infty$ (记为 115-Inf)。不同等级下 $PM_{2.5}$

的 FY-4B/AGRI 红外通道亮温均值分布如图 2 所示。

从图 2 中可以看出，FY-4B/AGRI 通道 8~15 在不同季节、不同 $PM_{2.5}$ 量级下均有所响应，尤其是通道 10~15。这说明采用红外通

道亮温及其衍生变量组合进行 $\text{PM}_{2.5}$ 反演是可行的。

3.2 FY-4B/AGRI 亮温反演 $\text{PM}_{2.5}$ 试验分析

本文进一步基于春季、夏季、秋季和冬季不同季节采用随机森林算法反演 $\text{PM}_{2.5}$ (结果见图 3)。由图 3 可知, 使用 FY-4B/AGRI 通道 8~15 亮温反演不同季节 $\text{PM}_{2.5}$ 均得到了较好的结果。从相关系数分析, 冬季 $\text{PM}_{2.5}$ 反演效果最好。从均方根误差分析, 夏季 $\text{PM}_{2.5}$ 反演效果最好。这可能与气温等气象要素有关, 后期将深入研究。

在文中构建模型和训练参数的基础上, 可以根据 FY-4B/AGRI 红外资料的数据实现江淮地区 4 km 空间分辨率、15 min 时间分辨率的近实时 $\text{PM}_{2.5}$ 反演, 从而为环境监测提供应用支持。

3.3 基于 SHAP 的 $\text{PM}_{2.5}$ 反演模型可解释分析

本文进一步基于 SHAP 开展了不同特征输入量对不同季节 $\text{PM}_{2.5}$ 反演的贡献率分析 (结果见图 4)。其中, “lon” 和 “lat” 分别表示经度和纬度; “Ch08” 表示通道 8 的亮温, “Ch08-09” 表示通道 8 与通道 9 的亮温之差, 其他以此类推; 水平轴是 SHAP 值, 每行代表一个特征; 蓝色和红色分别表示高特征值和低特征值。

由图 4 可知, 对于不同季节, 不同特征输入量对 $\text{PM}_{2.5}$ 反演的贡献率不同, 但地理信息 (经度和纬度) 对 $\text{PM}_{2.5}$ 反演的影响较大。例如, 对于秋季, 较低的特征值 “lon” 会推高 $\text{PM}_{2.5}$ 的反演值, 而特征值 “Ch11-15” 的散射在等于 0 的 SHAP 值附近振荡, 并且没有向两侧扩散, 表明这些特征与反演值无关。其他季节可作类似分析。通过这种方法可能可以更好地定位 $\text{PM}_{2.5}$ 的排放源, 进行 “源汇” 分析。

4 总结与展望

本文在分季节探讨 FY-4B/AGRI 红外亮温对不同量级 $\text{PM}_{2.5}$ 的信号响应或敏感性的基础上, 开展了 AGRI 亮温反演 $\text{PM}_{2.5}$ 研究, 并利用 SHAP 方法对反演模型进行了解释性分

析。主要结论如下:

(1) 分季节讨论 AGRI 红外亮温对 $\text{PM}_{2.5}$ 的敏感性。对于 FY-4B/AGRI 通道 8~15, 在不同季节、不同 $\text{PM}_{2.5}$ 量级下均有所响应, 尤其是通道 10~15。此结论或规律为卫星红外亮温反演 $\text{PM}_{2.5}$ 奠定了基础。

(2) AGRI 亮温反演 $\text{PM}_{2.5}$ 。区别于大多数方法依赖卫星获取的气溶胶光学厚度或白天可见光大气顶部反射率值反演 $\text{PM}_{2.5}$, 本文结合随机森林方法直接从红外资料近实时全时段 (包括白天和夜晚) 反演 $\text{PM}_{2.5}$ 。四个季节的 $\text{PM}_{2.5}$ 反演相关系数均超过 0.87, RMSE 值均低于 $7.21 \mu\text{g}/\text{m}^3$ 。

(3) 基于 SHAP 分析 $\text{PM}_{2.5}$ 反演的特征值贡献率。对于不同季节, 不同特征输入量对 $\text{PM}_{2.5}$ 反演的贡献率不同, 但地理信息 (经度和纬度) 对 $\text{PM}_{2.5}$ 反演的影响较大。

虽然本文的初步反演结果尚可, 但仍存在一些不足。本文仅从 FY-4B/AGRI 亮温及其衍生变量组合反演 $\text{PM}_{2.5}$, 未考虑 $\text{PM}_{2.5}$ 与气象变量之间的关系, 导致文中的分析不够深入。未来将考虑耦合气象变量的物理因素, 结合卫星遥感资料开展多模型集成 $\text{PM}_{2.5}$ 反演, 并将其反演产品应用于公共环境服务, 以实现 $\text{PM}_{2.5}$ 近实时监测。

参考文献

- [1] Duarte A L, Schneider I L, Artaxo P, et al. Spatiotemporal assessment of particulate matter (PM_{10} and $\text{PM}_{2.5}$) and ozone in a Caribbean urban coastal city [J]. *Geoscience Frontiers*, 2022, **13** (1): 101168.
- [2] Krzyzanowski M, Cohen A. Update of WHO air quality guidelines [J]. *Air Quality Atmosphere and Health*, 2008, **1**: 7-13.
- [3] Yu X Y, Xi M Z, Wu L Y, et al. Spatiotemporal weighted for improving the satellite-based high-resolution ground $\text{PM}_{2.5}$ estimation using the light gradient boosting machine [J]. *Remote Sensing*, 2023, **15**(16): 4104.

- [4] Chen G B, Li S S, Knibbs L D, et al. A machine learning method to estimate $PM_{2.5}$ concentrations across China with remote sensing, meteorological and land use information [J]. *Science of the Total Environment*, 2018, **636**(15): 52–60.
- [5] He Q Q, Huang B. Satellite-based mapping of daily high-resolution ground $PM_{2.5}$ in China via space-time regression modeling [J]. *Remote Sensing of Environment*, 2018, **206**: 72–83.
- [6] Sayer A M, Munchak L A, Hsu N, C et al. MODIS Collection 6 aerosol products: Comparison between Aqua's e-Deep Blue, Dark Target, and 'merged' datasets, and usage recommendations [J]. *Journal of Geophysical Research: Atmospheres*, 2014, **119**(24): 13965–13989.
- [7] Zhang Y, Li Z Q. Remote sensing of atmospheric fine particulate matter ($PM_{2.5}$) mass concentration near the ground from satellite observation [J]. *Remote Sensing of Environment*, 2015, **160**: 252–262.
- [8] Van D A, Martin R V, Brauer M, et al. Global estimates of ambient fine particulate matter concentrations from satellite-based aerosol optical depth: development and application [J]. *Environmental Health Perspectives*, 2010, **118**(6): 847–855.
- [9] Wei J, Li Z Q, Cribb M, et al. Improved 1km resolution $PM_{2.5}$ estimates across China using enhanced space-time extremely randomized trees [J]. *Atmospheric Chemistry and Physics*, 2020, **20**(6): 3273–3289.
- [10] Li T W, Yang Q Q, Wang Y, et al. Joint estimation of $PM_{2.5}$ and O_3 over China using a knowledge-informed neural network [J]. *Geoscience Frontiers*, 2023, **14**(2): 101499.
- [11] Xue Y, Li Y, Guang J, et al. Hourly $PM_{2.5}$ estimation over central and eastern China based on Himawari-8 data [J]. *Remote Sensing*, 2020, **12**(5): 855.
- [12] Zhang T, He W, Zheng H, et al. Satellite-based ground $PM_{2.5}$ estimation using a gradient boosting decision tree [J]. *Chemosphere*, 2021, **268**: 128801.
- [13] Wei J, Huang W, Li Z Q, et al. Estimating 1-km-resolution $PM_{2.5}$ concentrations across China using the space-time random forest approach [J]. *Remote Sensing of Environment*, 2019, **231**(15): 111221.
- [14] Shen H F, Li T W, Yuan Q Q, et al. Estimating regional ground-level $PM_{2.5}$ directly from satellite top-of-atmosphere reflectance using deep belief networks [J]. *Journal of Geophysical Research: Atmospheres*, 2018, **123**(24): 13875–13886.
- [15] Breiman L. Random forests [J]. *Machine Learning*, 2001, **45**(1): 5–32.
- [16] Wang G, Han W, Yuan S, et al. Retrieval of high-frequency temperature profiles by FY-4A/GIIRS based on generalized ensemble learning [J]. *Journal of the Meteorological Society of Japan*, 2024, **102**(2): 241–264.
- [17] Huang H, Liu Y L, Liu Y F, et al. The spatial distribution and driving mechanism of soil organic matter in hilly basin areas based on genetic algorithm variable combination optimization and Shapley Additive Explanations Interpretation [J]. *Remote Sensing*, 2025, **17**(7): 1186.
- [18] Zhu L H, Kang W Z, Li W, et al. The optimal bias correction for daily extreme precipitation indices over the Yangtze-Huaihe River Basin, insight from BCC-CSM1.1-m [J]. *Atmospheric Research*, 2022, **271**: 106101.