基于小波分析优化 $PM_{2.5}$ 浓度预测模型

许艺馨[1],任杰[2],冯磊[1],梁莹露[3],刘怡明[3]
(1. 广西物流职业技术学院,广西　贵港　537100;2. 南京理工大学环境科学与工程学院,
江苏　南京　210044;3. 贵港市气象局,广西　贵港　537100)

摘　要:采用多元线性回归方法(MLR)和 BP 神经网络方法(BPNN),按 1 h、3 h、6 h、12 h、24 h、48 h 预测时长对贵港市 2015—2018 年 $PM_{2.5}$ 浓度建模并检验对比模型准确率。结果表明,基于 MLR 与 BPNN 都能对 $PM_{2.5}$ 浓度作预测,预测效果随着预测时长的增加而下降,MLR、BPNN 模型预测结果平均绝对误差(MAE)分别为 4.01 μg/m³ ~ 15.48 μg/m³、3.89 μg/m³ ~ 15.63 μg/m³。采用小波分析方法对污染物数据优化并再次建模,结果表明,小波 - 多元线性回归(W-MLR)模型与小波 - 神经网络(W-BPNN)模型均得到优化,3 h ~ 24 h 预测时长优化效果尤为显著,W-MLR、W-BPNN 模型预测结果分别使 MAE 降低 1.6% ~ 13.5%、0.8% ~ 9.8%,且后者预测效果优于前者。

关键词: $PM_{2.5}$;气象要素;多元线性回归;BP 神经网络;小波分析
中图分类号:X513　　**文献标志码**:B　　**文章编号**:1006 - 2009(2021)02 - 0024 - 05

Wavelet Analysis for Optimizating $PM_{2.5}$ Concentration Prediction Models

XU Yi-xin[1], REN Jie[2], FENG Lei[1], LIANG Ying-lu[3], LIU Yi-ming[3]
(1. *Guangxi Logistics Vocational and Technical College*, *Guigang*, *Guangxi* 537100, *China*;
2. *College of Environmental Science and Engineering*, *Nanjing University of Science and Technology*, *Nanjing*, *Jiangsu* 210044, *China*; 3. *Guigang Meteorologic Bureau*, *Guigang*, *Guangxi* 537100, *China*)

Abstract: Using multiple linear regression method (MLR) and BP neural network method (BPNN), $PM_{2.5}$ concentration prediction models were built by taking 1 h, 3 h, 6 h, 12 h, 24 h, 48 h as prediction time from 2015 to 2018 in Guigang. The accuracy of the models were tested and compared. The results showed that both MLR and BPNN could be used in $PM_{2.5}$ concentration prediction. The prediction accuracy declined as the prediction time increased. The mean absolute error (MAE) of the prediction models by MLR and BPNN were 4.01 μg/m³ ~ 15.48 μg/m³ and 3.89 μg/m³ ~ 15.63 μg/m³, respectively. Using wavelet data analysis method for optimizing the contaminant data and making the model again, both W-MLR and W-BPNN were optimized. The optimization was significant when the prediction time was from 3 h to 24 h. By W-MLR and W-BPNN models, the MAE reduced 1.6% ~ 13.5% and 0.8% ~ 9.8%, respectively, and W-BPNN model was superior to W-MLR model in prediction.

Key words: $PM_{2.5}$; Meteorological elements; Multiple linear regression; BP neural network; Wavelet analysis

$PM_{2.5}$ 浓度变化影响着人类健康和生态环境[1-3],相比于一次污染物而言,$PM_{2.5}$ 生成机制和空间输送机理更为复杂,使得对其浓度预测难度较大,探索实现高精度的 $PM_{2.5}$ 浓度预测模型显得尤为重要。目前,国内外学者常用来预报空气污染物浓度的经典方法主要有数值预报方法[4]、统计预报方法[5]等,随着计算机技术的日趋成熟,涌现了

收稿日期:2020 - 03 - 24;**修订日期**:2021 - 01 - 23
基金项目:广西气象科研基金资助项目(桂气科 2019M22);贵港市科技局基金资助项目(贵科攻 1505004)
作者简介:许艺馨(1989—),女,广西贵港人,工程师,硕士,研究方向为大气污染控制技术。

人工神经网络模型[6-7]（Artificial Neural Network，ANN）、支持向量回归模型（Support Vector Machine，SVM）等机器学习预报方法。小波分析方法在数字信号分解和重构方面有着很好的效果[8-9]，与传统 ANN、SVM 相比，添加小波后的模型预测性能得到了提高[10-11]。然而，现阶段预测模型的研究大多采用日均值数据进行建模，采用随后几天或几个月的数据来验证模型，虽然一定程度上可拟合 $PM_{2.5}$ 的日变化趋势，但无法预测短时变化和突发变化，预测精度无法满足精细化空气质量预报的需求。此外，仅用几天或几个月的数据验证模型，难以说明模型的稳定性和重复性是否具备一定的科学性。今以主要大气污染物 $PM_{2.5}$ 作为预报指标，应用长时间的污染物浓度和气象要素小时序列，提出基于小波结合多元线性回归方法和 BP 神经网络方法建立不同预测时长的模型，为大气污染精细化预测与评价提供新的思路和方法。

1　材料与方法

1.1　数据来源

环境空气质量小时监测数据来自贵港市环境监测中心，包括 SO_2、NO_2、CO、O_3、$PM_{2.5}$、PM_{10} 等 6 项主要污染物和空气质量指数（AQI）共 7 项指标，空气质量分指数（IAQI）和污染物项目浓度限值依据《环境空气质量标准》（GB 3095—2012）中的相关规定。目前贵港市设有 4 个环境空气质量国控监测点，分别是贵城子站、江南子站、荷城子站和德智子站，其中德智子站作为环境空气质量对照点，数据不参与全市小时平均浓度、日均浓度统计。4 个监测点中贵城子站、荷城子站和德智子站位于港北区，江南子站位于港南区，见图 1。

气象要素小时观测数据来自贵港市国家气象观测站，包括过去 1 h 雨量（R）、气温（T）、气压（P）、相对湿度（RH）、平均风速（V）、日照时数（S）等 6 个要素。

1.2　数据处理

收集的 2015—2018 年污染物小时浓度存在较多缺测和无效数据，为了提高数据的使用性及有效性，依据王振波等[12]的数据质量控制方法，将原始数据中的缺测数据和经系统质控后确定为无效的数据剔除，再采用线性插值方法对污染物时间序列进行补充，计算公式为：

$$X_i = X_n + (i-n)(X_m - X_n)/(m-n)\ ,\ n < i < m \qquad (1)$$

式中：X_i 为缺测值和无效数据；X_m 和 X_n 分别为 X_i 前后的浓度值。若当天各个小时的缺测值和无效数据较多，则用相邻日期的小时数据替换。

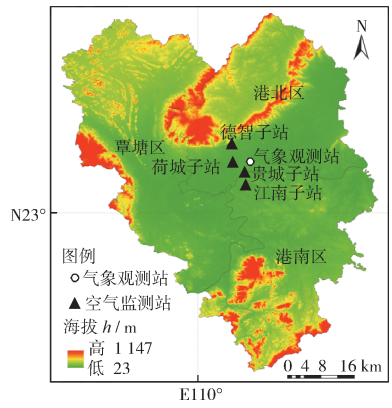

图 1　贵港空气自动监测站和气象观测站分布

Fig. 1　Distribution of air automatic monitoring stations and a meteorological station in Guigang

2　结果与讨论

2.1　相关性分析

为能定性定量探究对 $PM_{2.5}$ 浓度影响显著的因素，减少构建预测模型的输入量，采用 SPSS 22.0 中 Pearson 相关分析方法对 2015—2017 年各监测站 $PM_{2.5}$ 小时浓度与其他污染物浓度、气象要素进行相关性分析（见表 1），共 26 303 组样本。相关系数 R 反映两组样本的相似程度。由表 1 可知，各监测子站的相关性表现得较为一致，$PM_{2.5}$ 浓度与其他污染物浓度、气象要素的相关性大多通过 95% 显著性检验，只有 S 未通过 95% 显著性检验。当前小时 $PM_{2.5}$ 浓度与 SO_2、NO_2、PM_{10}、P 呈正相关，与 T、RH、R、V 呈负相关，与 O_3、S 呈弱正相关或负相关。由于贵城子站与气象观测站距离最近，故该子站 $PM_{2.5}$ 与各因子综合相关性最高，全市与 $PM_{2.5}$ 呈正相关因子排序为 $PM_{10} > NO_2 > P > SO_2 > S$，负相关因子按绝对值排序 $V > T > RH > R > O_3$。

表 1 中当前 $PM_{2.5}$ 与其他污染物的相关性表明，$PM_{2.5}$ 与 PM_{10} 有极显著的相关性。从粒径范围来看，PM_{10} 中包含了部分 $PM_{2.5}$，且 $PM_{2.5}$ 可以随时转化为 PM_{10}，故两者浓度保持高度正相关关系。

近来有学者发现[13]，随着大气氧化性的增强，将会显著促进二次气溶胶的生成，O_3 对大气中二次硫酸盐、二次硝酸盐等二次污染物的生成起到不

表 1　2015—2017 年各监测子站 $PM_{2.5}$小时质量浓度与各因子的相关系数

Table 1　Correlation coefficient between $PM_{2.5}$ hourly mass concentration and each factor in each monitoring substation from 2015 to 2017

各因子	德智子站	江南子站	贵城子站	荷城子站	全市
SO_2 ρ/(μg·m^{-3})	0.355	0.187	0.249	0.194	0.285
NO_2 ρ/(μg·m^{-3})	0.367	0.565	0.555	0.475	0.589
O_3 ρ/(μg·m^{-3})	0.188	−0.054	−0.028	0.028	−0.013①
PM_{10} ρ/(μg·m^{-3})	0.855	0.889	0.913	0.858	0.926
P p/(hPa)	0.273	0.311	0.326	0.315	0.329
T θ/℃	−0.189	−0.249	−0.257	−0.231	−0.255
RH/%	−0.167	−0.132	−0.144	−0.175	−0.155
R h/mm	−0.077	−0.072	−0.075	−0.065	−0.073
V v/(m·s^{-1})	−0.190	−0.277	−0.274	−0.217	−0.266
S t/h	0.031	0.008②	−0.001②	0.023	0.010②

①代表未通过 99% 显著性检验；②代表未通过 95% 显著性检验。

可忽视的作用[14]。德智子站 $PM_{2.5}$与 O_3 的正相关性最高，德智子站作为清洁对照点，能够真实反映未受大环境污染源影响下 $PM_{2.5}$的变化特征。

2.2　多元线性回归(MLR)模型

为评估其他污染物浓度与气象要素对 $PM_{2.5}$逐时变化的影响，利用相关性分析结果(见表 1)，选取 2015—2017 年已通过 95% 显著性检验的污染物浓度和气象要素作为自变量，建立未来 1 h ~ 48 h 的 $PM_{2.5}$浓度 MLR 预测模型，见表 2。

根据已建立的 2015—2017 年的 MLR 预测模型(见表 2)，1 h ~ 48 h 模型训练精度 R^2 在 0.404 ~ 0.953 之间，模型中各因子均通过显著性水平为 95% 的假设检验，并满足独立性假设。

利用 2018 年的污染物浓度、气象要素数据对模型进行检验，见图 2(a)(b)。由图 2 可见，污染物浓度实测值和预测值的变化趋势较为一致，验证精度 R^2 在 0.365 ~ 0.947 之间，总体能够反映实际 $PM_{2.5}$浓度的变化趋势，可随着预测时长的增加，模型对于突变高值存在一定低估，往往突变值即为污染物超标值。

表 2　2015—2017 年 $PM_{2.5}$小时平均质量浓度多元线性回归(MLR)模型

Table 2　Multiple linear regression (MLR) model of $PM_{2.5}$ hourly average concentration from 2015 to 2017

预测时长	多元线性回归方程
预测 1 h	$y_{1\,h}=0.979PM_{2.5}+0.115NO_2+0.023O_3-0.041PM_{10}-0.111P-0.183T-0.048RH-0.23R-0.982V+120.308$
预测 3 h	$y_{3\,h}=0.844PM_{2.5}-0.037SO_2+0.121NO_2+0.076O_3-0.018PM_{10}-0.351P-0.622T-0.153RH-0.394R-2.699V+384.359$
预测 6 h	$y_{6\,h}=0.607PM_{2.5}-0.072SO_2+0.126O_3+0.097PM_{10}-0.335P-0.824T-0.276RH-0.263R-3.363V+386.498$
预测 12 h	$y_{12\,h}=0.441PM_{2.5}-0.055SO_2+0.083O_3+0.133PM_{10}+0.369P-0.323T-0.378RH-2.568V-320.55$
预测 24 h	$y_{24\,h}=0.402PM_{2.5}-0.126SO_2+0.173NO_2+0.096PM_{10}+0.096P-0.876T-0.449RH-2.623V-22.421$
预测 48 h	$y_{48\,h}=0.293PM_{2.5}-0.183SO_2+0.209NO_2+0.035PM_{10}+0.278P-1.098T-0.52RH-2.288V-188.151$

2.3　BP 神经网络(BPNN)模型

根据表 1 中相关性分析结果，采用 Matlab R2016b 软件进行建模。将 2015—2017 年对 $PM_{2.5}$影响显著的污染物浓度和气象要素作为输入层；中间为一个隐含层，网络参数的节点数为 10；将 1 h ~ 48 h 后的 $PM_{2.5}$小时浓度作为输出层[15]。

模型的训练精度 R^2 为 0.494 ~ 0.956，运用 2018 年污染物浓度和气象要素来验证模型，见图 3(a)(b)。验证精度 R^2 为 0.327 ~ 0.949，BPNN 模型虽然训练精度较高，但验证精度低，说明模型泛化能力不够理想。由图 3 可见，随着预报时长由 3 h 增加到 24 h，预报精度会显著下降，在 $PM_{2.5}$小时浓度出现高、低值时，预测值出现不稳定的现象，说明 BPNN 模型对高低值的预测皆不够灵敏。

2.4　小波方法优化预测模型

由于 $PM_{2.5}$原始数据具有噪声干扰，经过 daubechies (db)小波预处理大气污染物浓度序列可滤除因仪器异常等原因导致的较大或较小数据，还原数据的真实序列。根据测试不同的 db 小波消失矩 N，最终确定利用 db5 小波变换方法对 2015—2018 年大气污染物浓度小时序列进行 5 层分解重构，得到低频系数 $A5$ 和高频系数 $D1$ ~ $D5$(见图 4)。其

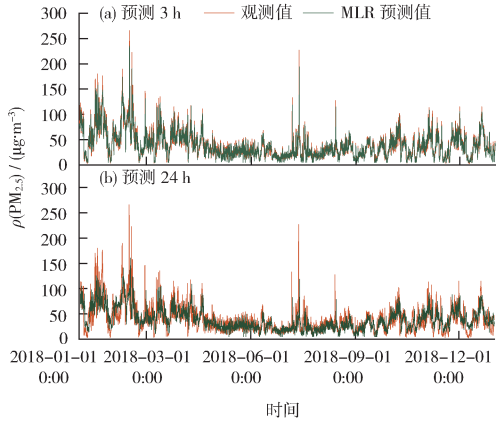

图 2　MLR 模型 3 h、24 h 预测值与观测值对比

Fig. 2　Comparison of the predicted values at 3 h and 24 h by MLR model with the observed values

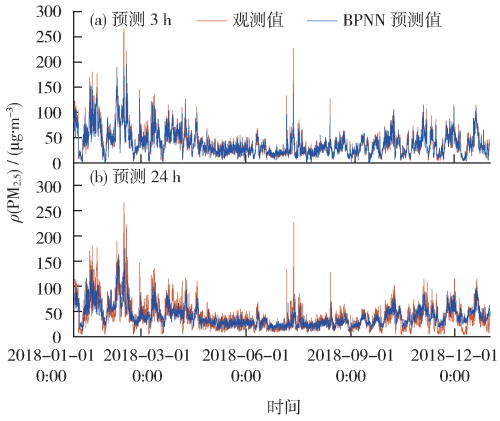

图 3　BPNN 模型 3 h、24 h 预测值与观测值对比

Fig. 3　Comparison of the predicted values at 3 h and 24 h by BPNN model with the observed values

中，由整体信息构成低频系数，反映了 $PM_{2.5}$ 演变的总体趋势和周期[16]；各种异常突变、干扰噪声构成高频系数，反映了 $PM_{2.5}$ 的突变和细小波动[17-18]。将 2015—2017 年大气污染物浓度滤波值及气象要素实测值作为 MLR 模型和 BPNN 模型的输入层，1 h～48 h 后的 $PM_{2.5}$ 浓度滤波值作为输出层，构建小波－多元线性回归(W-MLR)模型和小波－神经网络(W-BPNN)模型。

运用 2018 年大气污染物浓度和气象要素对预测模型准确度进行验证，见图 5(a)—(d)。由图 5 可见，W-MLR 模型和 W-BPNN 模型的验证效果较好，模型很好地捕捉了 $PM_{2.5}$ 浓度峰值、谷值。相比而言，在 2018 年 2 月 13 日—16 日及 7 月 12 日等重污染过程的 24 h 预测模型中，W-BPNN 模型比

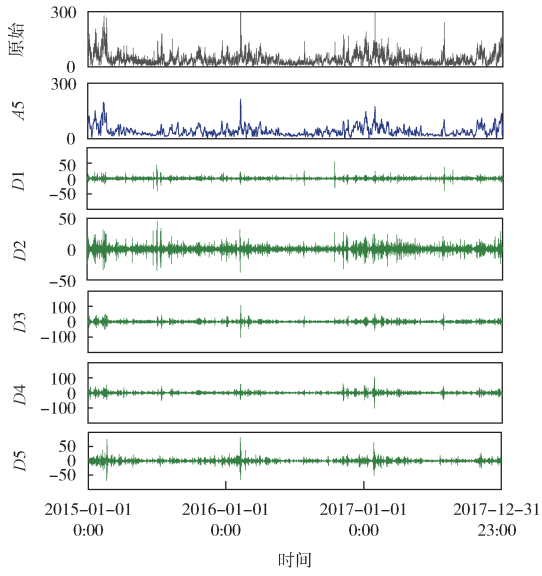

图 4　$PM_{2.5}$ 经 db5 小波分解结果

Fig. 4　Results of $PM_{2.5}$ concentration by db5 wavelet decomposition

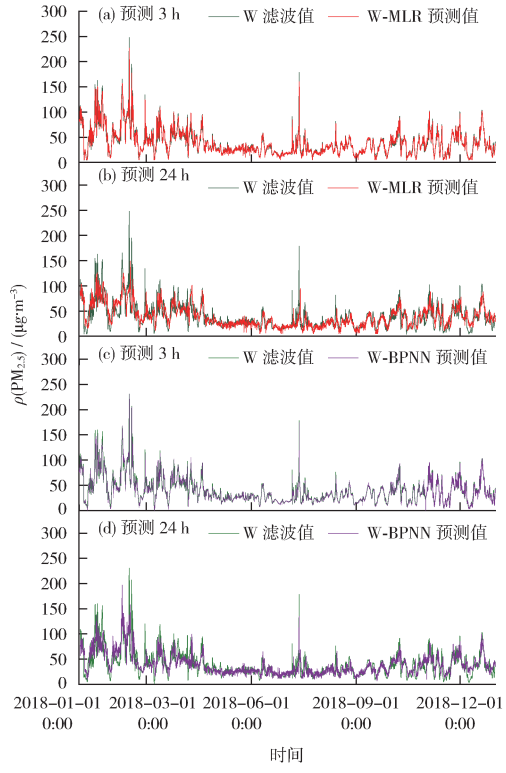

图 5　db5 滤波值与 W-MLR、W-BPNN 模型的 3 h、24 h 预测值对比

Fig. 5　Comparison of db5 filtering values with the prediction values by W-MLR and W-BPNN models at 3 h and 24 h

W-MLR 显示出更好的优势。

2.5　各模型预测性能参数对比

为能够定量描述模型在预测 $PM_{2.5}$浓度方面的性能,用 $PM_{2.5}$实测值与预测值的平均绝对误差(MAE)和均方根误差(RMSE)作为衡量模型预测精度的指标,指标越接近 0 表示预测结果越准确。

计算得出 MLR、BPNN、W-MLR、W-BPNN 模型预测性能参数,见表 3。

由表 3 可知,随着预测时效的增长,4 种模型预测效果均逐渐降低。在 3 h ~ 24 h 的预测时长中,W-MLR 模型预测结果与 MLR 比较,使 MAE 降低 1.6% ~ 13.5%,RMSE 降低 0.1% ~ 13.1%;W-BPNN 模型预测结果与 BPNN 比较,使 MAE 降低 0.8% ~ 9.8%,RMSE 降低 1.1% ~ 9.1%;而 W-BPNN 模型的验证指标又比 W-MLR 模型稍有提高,W-BPNN 模型(最优模型)与 MLR 比较,能使 MAE 下降 3.4% ~ 14.2%,RMSE 下降 2.0% ~ 13.6%。

表 3　4 种模型预测性能参数对比

Table 3　Comparison of predictive performance parameters of the four models

模型		训练 R_1^2	验证 R_2^2	验证 MAE/($\mu g \cdot m^{-3}$)	验证 RMSE/($\mu g \cdot m^{-3}$)	模型		训练 R_1^2	验证 R_2^2	验证 MAE/($\mu g \cdot m^{-3}$)	验证 RMSE/($\mu g \cdot m^{-3}$)
预测 1 h	MLR	0.953	0.947	4.007	6.267	预测 12 h	MLR	0.578	0.526	12.936	18.741
	BPNN	0.956	0.949	3.891	6.143		BPNN	0.651	0.533	12.587	18.517
	W-MLR	0.988	0.933	5.021	7.095		W-MLR	0.697	0.578	11.944	17.674
	W-BPNN	0.990	0.929	5.210	7.246		W-BPNN	0.778	0.591	11.658	17.260
预测 3 h	MLR	0.808	0.801	8.040	12.135	预测 24 h	MLR	0.533	0.457	13.781	20.026
	BPNN	0.854	0.819	7.647	11.533		BPNN	0.581	0.472	13.416	19.839
	W-MLR	0.922	0.846	6.953	10.546		W-MLR	0.608	0.480	13.565	19.554
	W-BPNN	0.939	0.850	6.898	10.489		W-BPNN	0.656	0.475	13.311	19.619
预测 6 h	MLR	0.681	0.661	10.774	15.834	预测 48 h	MLR	0.404	0.365	15.478	21.457
	BPNN	0.752	0.697	9.889	14.869		BPNN	0.494	0.327	15.628	22.117
	W-MLR	0.815	0.717	9.511	14.409		W-MLR	0.441	0.361	15.548	21.470
	W-BPNN	0.869	0.738	8.977	13.819		W-BPNN	0.545	0.349	15.727	21.871

4 种 $PM_{2.5}$预测模型在技术上互有优劣,从检验效果来看,皆达到了统计模型预测的技术要求。其中 BPNN、W-BPNN 模型能较好地捕捉 $PM_{2.5}$小时浓度与模型输入层因子之间的非线性影响规律,能够对未来 3 h ~ 24 h $PM_{2.5}$作较好的预测,可随着预测时效的增加,当应变量与自变量不再满足于某种内在规律性时,预测效果开始下降。

刘杰等[19]利用北京地区的大气污染物和气象要素观测资料,建立的多元线性回归方程预测 $PM_{2.5}$质量浓度的 MAE 为 31.7 $\mu g/m^3$,RMSE 为 42.9 $\mu g/m^3$,利用神经网络建模预测 $PM_{2.5}$质量浓度的 MAE 为 27.6 $\mu g/m^3$,RMSE 为 39.9 $\mu g/m^3$。李祥等[20]利用 BPNN 模型对天津地区 $PM_{2.5}$日均值监测数据建模,预测模型的 MAE 为 44 $\mu g/m^3$,RMSE 为 59 $\mu g/m^3$。文中选取的资料时间序列较长,考虑的模型输入因子较全面,得出模型验证精度较高,W-BPNN 模型的 6 个预测时长的平均 MAE 为 10.29 $\mu g/m^3$,平均 RMSE 为 15.05 $\mu g/m^3$。

3　结论

(1) 利用相关分析方法,得出与 $PM_{2.5}$小时浓度呈高度相关的污染物为 PM_{10},气压则是影响 $PM_{2.5}$小时浓度最显著的气象要素。

(2) MLR 模型虽然可定性预测 $PM_{2.5}$浓度变化趋势,定量预测 $PM_{2.5}$小时浓度,但需要筛选确定影响 $PM_{2.5}$浓度的显著性因子,并根据不同的预测时效构建不同的方程,过程较为烦琐。

(3) BPNN 方法虽然能较好地解决关系复杂的非线性问题,但其对 $PM_{2.5}$高、低值预报效果不够理想,容易产生过度拟合、局部最小化、泛化能力低等缺点。

(4) 相比而言,在 3 h ~ 24 h 的预测时长中,小波方法能优化前面两种传统的预测模型,能更好地预测 $PM_{2.5}$浓度突变情况,有效提高重污染天气下 $PM_{2.5}$浓度拟合效果。其中 W-BPNN 模型优于 W-MLR 模型,可作为一定步长 $PM_{2.5}$浓度预测的优选方法,为实现其他大气污染物浓度精细化预报提供新的思路和方法。

(下转第 34 页)