Supplement of
Winter brown carbon over six China’s megacities: Light absorption, molecular characterization, and improved source apportionment revealed by multilayer perceptron neural network
Diwei Wang¹, Zhenxing Shen¹*, Qian Zhang², Yali Lei³, Tian Zhang¹, Shasha Huang¹, Jian Sun¹, Hongmei Xu¹, Junji Cao⁴
¹Department of Environmental Science and Engineering, Xi’an Jiaotong University, Xi’an 710049, China
²Key Laboratory of Northwest Resource, Environment and Ecology, MOE, Xi’an University of Architecture and Technology, Xi’an 710055, China
³Key Lab of Geographic Information Science of the Ministry of Education, School of Geographic Sciences, East China Normal University, Shanghai 200241, China
⁴Key Lab of Aerosol Chemistry & Physics, SKLLQG, Institute of Earth Environment, Chinese Academy of Sciences, Xi’an, China
Correspondence to: Zhenxing Shen (zxshen@mail.xjtu.edu.cn)

Contents:
Text S1. Optical properties of brown carbon (BrC) calculation
Text S2. The ANN-MLP model construction
Table S1. Information of sampling sites
Table S2. Concentrations of PM2.5 and carbonaceous components in six Chinese cities
Figure S1. The MLP neural network architecture used in this study
Figure S2. The relationship between the abundance of POC & SOC and BrC \( b_{abs365} \) in six cities.
Figure S3. The relationship between the abundance of K⁺ and BrC \( b_{abs365} \) in BJ, HrB, XA and WH.
Figure S4. Source profiles (bars and left y-axis) and percentage contributions (dots and y-axis) of each chemical component resolved from PMF model analysis in six megacities.
Figure S5. Average source contribution to aerosol mass concentration that is estimated by PMF source factor.
Figure S6. The correlation of the observed and modelled BrC \( b_{abs365} \) for the six cities using MLP analysis.
Optical properties of brown carbon (BrC) calculation

The light absorption coefficient ($b_{abs}$, Mm$^{-1}$) of the BrC was calculated using the following Eq. (S1):

\[ b_{abs} = \left( A_\lambda - A_{700} \right) \times \left( V_{ext} \times \text{Portions} \right) \times \ln(10) \div \left( V_{aero} \times L \right) \]  

(S1)

where $A_\lambda$ and $A_{700}$ represent the measured absorbance at a specified $\lambda$ value and at 700 nm, respectively. In this study, $\lambda$ was set to 365 nm. Furthermore, $V_{ext}$ represents the volume of the solvent extract (5 mL) in which different portions of the filter were used to extract and estimate the absorption signal for the full filter. Finally, $V_{aero}$ represents the sampling volume, and $L$ represents the path length of the cell (100 cm).

In this study, ambient OC was used to replace methanol-soluble OC (MSOC) because several studies have indicated that most OC (~95%) can be extracted using methanol (Cheng et al., 2016; Huang et al., 2018). The mass absorption efficiency (MAE, m$^2$ g$^{-1}$) of the filter extracts at $\lambda$ was calculated using the following Eq. (S2):

\[ \text{MAE}_\lambda = \frac{b_{abs}}{OC} \]  

(S2)

The wavelength dependence of light absorption by BrC in the solvent extracts was derived using the following Eq. (S3):

\[ b_{abs} = K \times \lambda^{-\text{AAE}} \]  

(S3)

where $K$ denotes a constant, $\lambda$ denotes the wavelength of BrC, and AAE denotes the absorption Ångström exponent. In this study, to avoid interference from inorganic species, AAE was calculated through the linear regression fitting of log $b_{abs}$ versus log $\lambda$ in the 330–550-nm wavelength range.

The ANN-MLP model construction

As shown in Figure S1, the ANN-MLP model includes three main layers: input layer, hidden layer, and output layer. The two adjacent layers are fully connected (i.e., any neuron in the layer has connections to all the neurons in the layer below). The input layer receives the daily contributions of the PM$_{2.5}$ sources obtained from the PMF, and the BrC $b_{abs365}$ of six cities is set as the response variables in the output layer gives. In this study, we limited only a one hidden layer was used to design MLP models to explore the applicability of non-linear models. The neurons in the hidden layer computes the input data, realizes the nonlinear mapping of the input information, and passes the information to the output layer. The number of neurons in the hidden layer was determined automatically by the estimation algorithm (Borlaza et al., 2021). The important parameters of neural network are the connection weights, bias and activation functions between different layers. The weight represents the connection strength between neurons, and
the bias ensures that the output value calculated through the input cannot be activated randomly. The activation function plays a role of nonlinear mapping, which can limit the output amplitude of neurons within a certain range. The process of finding optimal parameters is the training process of neural network. For instance, the transformation of the data from the input layer in the hidden layer can be expressed by Eq. (S4):

$$\forall j \in \{1, \ldots, l\}, a_j = H\left(\sum_{i=1}^{l_i} w_{i,j}^G x_i + w_{0,j}^G\right)$$

(S4)

with $w_{i,j}^G$ the weight of the neuron between the input and hidden layer and $w_{0,j}^G$ an activation constant for neuron j. The activation function $H$ is often non-linear (Borlaza et al., 2021).

The feedforward ANN-MLP model was trained with a back-propagation process (Chang et al., 2019). For training the ANN and obtaining the optimal model, the following treatments were developed:

(a) the dataset was standardized by subtracting the mean of the observed values and dividing by the standard deviation, then the standardized values were saved as variable;

(b) 70% of the dataset was used as training set to train the model, and 30% of the dataset was used as test set to monitor errors during training process;

(c) The nonlinear functions (activation functions) of sigmoid and hyperbolic tangent (TanH) were introduced to perform nonlinear transformation on hidden variables, and then serve as the input of the next fully connected layer;

(d) initialized randomly the weights of adjacent layer nodes, and then the scaled conjugate and stochastic gradient descent optimization algorithms were used for iterative training to find the optimal weights between nodes of each layer;

(e) the MLP training stops when the model output error reaches the set error standard.
Figure S1. The MLP neural network architecture used in this study, where \( n \) is the number of PM\(_{2.5}\) sources from PMF, \( G \) is the daily standardized contribution of sources, and \( b_{abs} \) is the light absorption coefficient of BrC.

### Table S1. Information of sampling sites

<table>
<thead>
<tr>
<th>Observation megacity</th>
<th>Location</th>
<th>Geographical China</th>
<th>Site description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>39.97° N, 116.36° E</td>
<td>China</td>
<td>~8 m above ground level, in the north part of Beijing, which is close to several major roads including a highway and is surrounded by residences and restaurants.</td>
</tr>
<tr>
<td>Harbin</td>
<td>45.74° N, 126.73° E</td>
<td>North China</td>
<td>~18 m above ground level, in the east of Harbin, surrounded by campus, roads, residential commercial emission sources</td>
</tr>
<tr>
<td>Xi’an</td>
<td>34.23° N, 108.88° E</td>
<td>China</td>
<td>~15 m above ground level, in the southeast of downtown Xi’an, surrounded by two lane roads, residential commercial districts.</td>
</tr>
<tr>
<td>Chengdu</td>
<td>30.70° N, 104.06° E</td>
<td>China</td>
<td>~18 m above ground level, on the rooftop of a building of Southwest Jiaotong University, surrounded by commercial and residential areas and close to a train station</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>23.12° N, 113.35° E</td>
<td>South China</td>
<td>~30 m above ground level, in the central of Guangzhou, there is no obvious industrial pollution source near the monitoring station.</td>
</tr>
<tr>
<td>Wuhan</td>
<td>30.53° N, 114.39° E</td>
<td>China</td>
<td>~18 m above ground level, in the southeast of Wuhan city, surrounded by roads, residential commercial districts, this is a typical urban site with no industrial emission sources nearby.</td>
</tr>
</tbody>
</table>

### Table S2. Concentrations of PM\(_{2.5}\) and carbonaceous components in six Chinese cities

<table>
<thead>
<tr>
<th>sites</th>
<th>PM(_{2.5})((\mu\text{g}\cdot\text{m}^{-3}))</th>
<th>OC((\mu\text{g}\cdot\text{m}^{-3}))</th>
<th>EC((\mu\text{g}\cdot\text{m}^{-3}))</th>
<th>SOC((\mu\text{g}\cdot\text{m}^{-3}))</th>
<th>POC((\mu\text{g}\cdot\text{m}^{-3}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>55.5 ± 41.5</td>
<td>12.5 ± 5.9</td>
<td>2.1 ± 1.7</td>
<td>5.0 ± 1.2</td>
<td>6.3 ± 2.9</td>
</tr>
<tr>
<td>Harbin</td>
<td>85.5 ± 43.9</td>
<td>19.4 ± 8.5</td>
<td>7.5 ± 5.6</td>
<td>9.2 ± 3.9</td>
<td>10.2 ± 5.5</td>
</tr>
<tr>
<td>Xi’an</td>
<td>80.7 ± 49.8</td>
<td>15.5 ± 7.9</td>
<td>3.6 ± 2.9</td>
<td>6.9 ± 3.8</td>
<td>7.8 ± 2.8</td>
</tr>
</tbody>
</table>
Table S.1. BrC $b_{abs365}$ (Mm$^{-1}$) in six cities.

<table>
<thead>
<tr>
<th>City</th>
<th>$b_{abs365}$</th>
<th>SOC (μg/m$^3$)</th>
<th>POC (μg/m$^3$)</th>
<th>SOC &amp; POC (μg/m$^3$)</th>
<th>POC &amp; POC (μg/m$^3$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chengdu</td>
<td>71.8 ± 28.2</td>
<td>5.6 ± 2.7</td>
<td>2.3 ± 1.0</td>
<td>1.0 ± 0.4</td>
<td>4.6 ± 2.2</td>
</tr>
<tr>
<td>Guangzhou</td>
<td>42.5 ± 17.2</td>
<td>10.9 ± 3.7</td>
<td>2.8 ± 2.0</td>
<td>6.9 ± 1.4</td>
<td>4.0 ± 2.7</td>
</tr>
<tr>
<td>Wuhan</td>
<td>63.9 ± 26.1</td>
<td>11.7 ± 4.8</td>
<td>4.2 ± 2.0</td>
<td>3.1 ± 1.6</td>
<td>8.2 ± 3.5</td>
</tr>
</tbody>
</table>

Figure S2. The relationship between the abundance of POC & SOC and BrC $b_{abs365}$ in six cities.
Figure S3. The relationship between the abundance of K⁺ and BrC $b_{abs365}$ in BJ, HrB, XA and WH.
Figure S4. Source profiles (bars and left y-axis) and percentage contributions (dots and y-axis) of each chemical component resolved from PMF model analysis in six megacities.
Figure S5. Average source contribution to aerosol mass concentration that is estimated by PMF source factor.

Figure S6. The correlation of the observed and modelled BrC $b_{\text{abs}365}$ for the six cities using MLP analysis.
Reference

Trébluchon, C., and Uzu, G.: Disparities in particulate matter (PM$_{10}$) origins and oxidative potential at a
city scale (Grenoble, France) – Part 2: Sources of PM$_{10}$ oxidative potential using multiple linear
regression analysis and the predictive applicability of multilayer perceptron neural network analysis,

Chang, T., Lu, J., Shen, Z., Huang, Y., Lu, D., Wang, X., Cao, J., and Morent, R.: Simulation and
optimization of the post plasma-catalytic system for toluene degradation by a hybrid ANN and NSGA-
II method, Applied Catalysis B: Environmental, 244, 107-119,

characteristics of brown carbon aerosol during winter in Beijing, Atmos. Environ., 127, 355-364,

Huang, R. J., Yang, L., Cao, J., Chen, Y., Chen, Q., Li, Y., Duan, J., Zhu, C., Dai, W., Wang, K., Lin, C.,
Carbon Aerosol in Urban Xi’an, Northwest China: The composition and light absorption properties,