Hourly ozone and PM$_{2.5}$ prediction using meteorological data –
alternatives for cities with limited pollutant information

Felipe Cifuentes$^1$*, Angel Gálvez$^1$, Carlos M. González$^1$, Mauricio Orozco-Alzate$^2$, Beatriz H. Aristizábal$^{1**}$

$^1$ Hydraulic Engineering and Environmental Research Group, Universidad Nacional de Colombia Sede Manizales, Cra 27 64-60 Bloque H Palogrande. Manizales, Colombia.

$^2$ Department of Informatics and Computing, Universidad Nacional de Colombia Sede Manizales, Km 7 via al Magdalena Bloque S La Nubia. Manizales, Colombia

Supplementary material

\[
\begin{align*}
R &= \frac{1}{n-1} \sum_{i=1}^{N} \left( \frac{O_i - \bar{P}}{\sigma_o} \right) \left( \frac{O_i - \bar{O}}{\sigma_o} \right) \\
MB &= \frac{1}{N} \sum_{i=1}^{N} P_i - O_i \\
RMSE &= \left( \frac{1}{N} \sum_{i=1}^{N} (P_i - O_i)^2 \right)^{1/2} \\
NRMSE &= \frac{RMSE}{\sigma_o} \\
MAE &= \frac{1}{N} \sum_{i=1}^{N} |P_i - O_i|
\end{align*}
\]

(S1) (S2) (S3) (S4) (S5)

Where $P_i$ represent the predicted values, $O_i$ represent the observed values, $\sigma$ represents the standard deviation and $N$ is the total number of observations.

* Corresponding author. 576 8879300 ext 55337; Fax: 576 8879300 ext 55129

E-mail address: *fcifuentesc@unal.edu.co **bharistizabalz@unal.edu.co
**Table S1.** Periods of analysis used in each step of the study.

<table>
<thead>
<tr>
<th>Date range</th>
<th>Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>October 1 to December 31, 2019</td>
<td>PCA, Spearman, training, and validation of all the models</td>
</tr>
<tr>
<td>January 1 to January 30, 2020</td>
<td>Testing of all the models</td>
</tr>
<tr>
<td>May 1 to May 31, 2020</td>
<td>Additional testing for the best performing models for O3 and PM2.5 forecasting</td>
</tr>
</tbody>
</table>
Table S2. Statistics summary of variables on hourly time scale from October 01, 2019 to January 31, 2020 in Manizales.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>O₃ (µg m⁻³)</th>
<th>PM₂.₅ (µg m⁻³)</th>
<th>SO₂ (µg m⁻³)</th>
<th>CO (µg m⁻³)</th>
<th>T (°C)</th>
<th>RH (%)</th>
<th>SRa (W m⁻²)</th>
<th>Pptb (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>15.8</td>
<td>11.3</td>
<td>4.1</td>
<td>790.0</td>
<td>17.6</td>
<td>82.6</td>
<td>326.6</td>
<td>0.2 (545.8)</td>
</tr>
<tr>
<td>Std. Dev.</td>
<td>108</td>
<td>4.5</td>
<td>1.9</td>
<td>319.7</td>
<td>2.6</td>
<td>6.9</td>
<td>249.9</td>
<td>2.7</td>
</tr>
<tr>
<td>Max.</td>
<td>45.1</td>
<td>24.3</td>
<td>9.8</td>
<td>1752.8</td>
<td>24.8</td>
<td>95.0</td>
<td>994.2</td>
<td>134.2</td>
</tr>
<tr>
<td>Min.</td>
<td>2.1</td>
<td>0.6</td>
<td>0.5</td>
<td>152.3</td>
<td>12.3</td>
<td>56.5</td>
<td>66</td>
<td>0</td>
</tr>
<tr>
<td>Count</td>
<td>1942</td>
<td>1942</td>
<td>1942</td>
<td>1942</td>
<td>2841</td>
<td>2841</td>
<td>2841</td>
<td>2841</td>
</tr>
<tr>
<td>Daily average (Std. Dev.)</td>
<td>15.2 (3.3)</td>
<td>12.6 (3.5)</td>
<td>4.3 (0.9)</td>
<td>839.3 (149.1)</td>
<td>17.8 (1.2)</td>
<td>82.3 (4.9)</td>
<td>290.5 (105)</td>
<td>4.7 (545.8)</td>
</tr>
</tbody>
</table>

a Solar radiation average values correspond to the periods between 6 am and 6 pm.
b Rainfall corresponds to the sum of hourly values registered at the station and the average value between parentheses represents the sum during the whole period of analysis.
Table S3. Best RMSE of regression techniques using a 10-fold cross validation for O₃ and PM₂.₅.

<table>
<thead>
<tr>
<th>Set of predictor variables</th>
<th>O₃ training set</th>
<th>O₃ validation set</th>
<th>PM₂.₅ training set</th>
<th>PM₂.₅ validation set</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLR</td>
<td>SVR</td>
<td>ANN</td>
<td>MLR</td>
</tr>
<tr>
<td>Spearman coefficients</td>
<td>3.87</td>
<td>2.76</td>
<td>3.10</td>
<td>3.45</td>
</tr>
<tr>
<td>PCA</td>
<td>4.22</td>
<td>3.15</td>
<td>3.37</td>
<td>3.89</td>
</tr>
<tr>
<td>Meteorology</td>
<td>5.32</td>
<td>3.96</td>
<td>4.30</td>
<td>4.74</td>
</tr>
<tr>
<td>All available variables</td>
<td>3.87</td>
<td>2.54</td>
<td>2.93</td>
<td>3.52</td>
</tr>
<tr>
<td>Number of observations</td>
<td>1077</td>
<td></td>
<td></td>
<td>120</td>
</tr>
</tbody>
</table>
Figure Captions

**Fig. S1.** Characteristics of the study area of Manizales, Caldas - Colombia. Image adapted from (CORPOCALDAS and UNAL, 2020; Cuesta et al., 2020)

**Fig. S2.** Histogram of hourly concentrations and Log₁₀ concentrations for O₃, PM₂.₅, SO₂ and CO.

**Fig. S3.** O₃ model performance on the test set for each predictor variable method: Spearman coefficients (SO₂, CO, T, ws, HR, RS, wd_i, wd_j). PCA (SO₂, T, HR, RD, wd_i). Meteorological (T, ws, HR, RS, Ppt, wd_i, wd_j). All variables (SO₂, CO, PM₂.₅, T, ws, HR, Ppt, RS, wd_i, wd_j).

**Fig. S4.** PM₂.₅ model performance on the test set for each predictor variable method: Spearman coefficients (SO₂, CO). PCA (SO₂ CO, T, HR). Meteorological (T, ws, HR, RS, Ppt, wd_i, wd_j). All variables (SO₂, CO, PM₂.₅, T, ws, HR, Ppt, RS, wd_i, wd_j).
Fig. S1
Fig. S2

- **O<sub>3</sub>**
  - Skewness: 0.9
  - Kurtosis: 2.9

- **PM<sub>2.5</sub>**
  - Skewness: 2.6
  - Kurtosis: 16.2

- **SO<sub>2</sub>**
  - Skewness: 0.8
  - Kurtosis: 3.6

- **CO**
  - Skewness: 1.8
  - Kurtosis: 4.7

- **O<sub>3</sub>**
  - Skewness: -0.1
  - Kurtosis: 2.0

- **PM<sub>2.5</sub>**
  - Skewness: -1.1
  - Kurtosis: 7.1

- **SO<sub>2</sub>**
  - Skewness: -0.3
  - Kurtosis: 2.7

- **CO**
  - Skewness: 0.3
  - Kurtosis: 2.7
Fig. S3
Fig. S4