Adriana Szulecka*, Marian Mazur*

Application of the Statistical Error and Quantitative Performance Measures in the Evaluation Process of Short-Term Air Quality Forecasts for Krakow (Poland)**

1. Introduction

Currently, the results of mathematical air quality modelling serve many purposes (e.g. global and local pollution prevention strategies, industrial emissions reduction, actions of local governments). Additionally, in the past few years, due to continuous digital technology development, prognostic models have been utilised for drawing up publicly available short-term forecasts of the upcoming air quality. This information is of particular importance in the areas of densely populated cities, which are frequently struggling with poor air quality. Issuing alerts about possible exceedances of the air quality standards [8] helps to mitigate adverse effects on the health of residents [6].

Krakow, located in the south-eastern Poland and with more than 750 000 inhabitants [3], has been recently qualified as one of the most polluted European cities due to excessive concentrations of PM2.5 [11]. In 2011, on behalf of the Marshal Office of Malopolska Voivodeship, an air quality forecasting system based on the results of global multiscale chemical weather modelling system GEM-AQ was developed and fully operational since then [23, 24]. The air quality forecasts for the area of Malopolska Voivodeship and major cities within are prepared on the basis of a modelling system created by the EkoForecast foundation, which utilises a global multiscale chemical weather model – GEM-AQ [9].

In order to ensure the correct functioning of the short-term air pollution forecasting system, an appropriate assessment at every stage of this application is essential. Basically, the validation process of a mathematical model consists of scientific, operational and statistical evaluation. Scientific evaluation requires a thorough knowledge of the model’s basis. It examines the accuracy of described physical and chemical processes, as well as other assumptions in the model. Operational

---

* AGH University of Science and Technology, Faculty of Mining Surveying and Environmental Engineering, Department of Environmental Management and Protection, Krakow, Poland

** The paper has been prepared within the scope of the AGH UST statutory research no. 11.11.150.008
evaluation considers user-related issues connected with, among others, the user interface, error checking of data and internal model diagnostics. Statistical assessment focuses mainly on the comparison between forecasted and measured observations. This method is rather intuitive and may not provide precise reasons for the divergence between observations, but it gives information about the nature and severity of possible error [2, 7, 14]. Due to its simplicity, this assessment should be carried out regularly to ensure reliable air quality forecasts and to improve general performance of the modelling system.

Basic statistical analysis may be performed with the use of the typical error measurements and correlation coefficients (Pearson, Spearman) for assessing model accuracy, which are applicable in many fields (economics, weather and air quality prediction). Those can be extended with a set of quantitative performance measures suggested by the U.S. Environmental Protection Agency as a basis for air quality model evaluation. To facilitate the interpretation of the values obtained one can begin with plotting the data in different ways (scatter, quantile-quantile, residual or conditional scatter plots) [4, 5].

2. Description of the Short-Term Air Quality Forecasting System for Krakow

Short-term air quality forecasts drawn up for the area of Krakow are based on the deterministic system modelling dynamics and atmospheric chemistry – GEM-AQ (Global Environmental Multiscale – Air Quality). It was developed as a result of extending the operational weather prediction model GEM by implementing air quality chemistry processes (i.e. transport, deposition, emission, limited wet chemistry). The GEM model was originally created by the Meteorological Services of Canada (MSC) and is presently used for weather prediction over Canada. Current mechanism in the GEM-AQ model contains 50 gas-phase compounds, 116 chemical and 19 photolysis reactions. The CAM module (Canadian Aerosol Model) implemented in this model describes 5 aerosol types (sulphate, sea-salt, organic carbon, black carbon and soil dust) and their physiochemical reactions. Gas-phase chemistry is based on the modified ADOM model (Acid Deposition and Oxidants Model). The chemical module is operating “on-line” – advection of the chemical compounds is performed at each timestep. Advection and vertical diffusion processes are computed with Semi-Lagrangian scheme deriving from the GEM model. The variable resolution capability makes it possible to perform high resolution local simulations based on the global runs [9, 12, 15].

Calculations in the short-term air quality forecast system for Małopolska Voivodeship are performed in a two-stage run. Results of the global simulation over Central Europe with variable resolution of 0.135 (3-D meteorological fields, chemical
composition of the atmosphere) are used as initial and boundary conditions for the nested run. Nested simulation covers the territory of Poland with resolution of 0.05° (Fig. 1). The forecast horizon is 75 hours (from 21:00 UTC on the previous day) [24]. The emission input data have been prepared with the EMEP inventories with 0.5° resolution for 2011 and 2012 [19, 20]. The emission rates of PM\textsubscript{coarse}, PM2.5, nitrogen oxides, sulphur dioxide, carbon oxide and NMVOCs have been used in the calculations.

Fig. 1. Visualization of the global (a) and nested (b) run in the GEM-AQ model for air quality forecasting system in Małopolska Voivodeship

Source: [19, 20]

The results of the analysed air quality forecasting system are accessible via the web service Wrota Małopolski of the Marshal Office [24] and available for selected cities within the area of Małopolska Voivodeship. Forecasts are presented in the form of daily averaged concentrations of the following compounds: PM10, PM2.5, nitrogen dioxide (NO\textsubscript{2}), sulphur dioxide (SO\textsubscript{2}), carbon oxide (CO) and ozone (O\textsubscript{3}) for the next three days. An additional statistical correction of forecast data is performed to obtain better results [19, 20, 24].

In addition, maps of the distribution of the Common Air Quality Index (CAQI) over Małopolska Voivodeship are presented at this website. This indicator was created in the CITEAIR project and has been used at the Air Quality in Europe web service [1] for assessing air quality among more than 100 European cities from 2006. Its value is computed based on the three main pollutants in Europe: PM10, nitrogen dioxide (NO\textsubscript{2}) and ozone (O\textsubscript{3}), which can be extended by additional substances: PM2.5, sulphur dioxide (SO\textsubscript{2}) and carbon oxide (CO) [1, 10]. Therefore, an accurate and reliable forecasting of these concentrations is notably significant, as they are forming the overall air quality in the region.
3. Statistical Methods of Model Evaluation

3.1. Description of the *ex post* Forecast Error Measurements

The parameters applied in this study indicate differences between the actual and forecasted values and have been selected from a larger set of variables. Those are described below [17–20, 22, 25, 26]:

– Mean Bias Error:

\[
MBE = \frac{1}{m} \sum_{t=1}^{m} (C_p - C_o)
\]  

(1)

– Mean Percentage Error:

\[
MPE = \frac{\sum_{t=1}^{m} C_p - C_o}{m} \times 100 \% 
\]  

(2)

– Mean Absolute Bias Error:

\[
MABE = \frac{1}{m} \sum_{t=1}^{m} |C_p - C_o|
\]  

(3)

– Mean Absolute Percentage Error:

\[
MAPE = \frac{1}{m} \sum_{t=1}^{m} \left|\frac{C_p - C_o}{C_o}\right| \times 100 \% 
\]  

(4)

– Root Mean Square Error:

\[
RMSE = \sqrt{\frac{1}{m} \sum_{t=1}^{m} (C_p - C_o)^2}
\]  

(5)

where \(C_p\) and \(C_o\) denotes a predicted and observed concentration, respectively, and is a total amount of observations during the time series (\(t = 1, 2, ..., m\)). Ideally, the value of MBE should be equal or close to zero. A positive or negative value indicates overprediction or underprediction of prognostic model, respectively. A significant difference between RMSE and MABE point to presence of high values of discrepancies between observed and predicted concentrations [2, 26].
3.2. Description of Quantitative Performance Measures (US EPA)

In order to assess the general performance of the analysed prognostic model, a set of quantitative measures developed in 1993 by the National Environmental Research Institute (NERI) of Denmark was applied in this study. They have been recommended by the US Environmental Protection Agency for air quality model evaluation tool and later implemented in a software package BOOT (Statistical Model Evaluation Software Package, Version 2.0). Those measures include [4, 5, 16, 21]:

a) Fractional Bias:

\[
FB = \frac{\overline{C_O} - \overline{C_p}}{0.5 \cdot (\overline{C_O} + \overline{C_p})}
\]

b) Geometric Mean Bias:

\[
MG = \exp\left(\ln \overline{C_O} - \ln \overline{C_p}\right)
\]

c) Normalized Mean Square Error:

\[
NMSE = \frac{(\overline{C_O} - \overline{C_p})^2}{\overline{C_O} \cdot \overline{C_p}}
\]

d) Geometric Variance:

\[
VG = \exp\left[\ln \overline{C_O} - \ln \overline{C_p}\right]^2
\]

e) Fraction within a factor of two:

\[
FAC2 = \text{fraction of data that satisfy } 0.5 \leq \frac{C_p}{C_O} \leq 2.0
\]

where \(C_p\) and \(C_O\) denotes a predicted and observed concentration, respectively, and \(m\) represents a total amount of observations during the \(\tau\) time series. An ideal model would have \(MG\), \(VG\) and \(FAC2 = 1.0\), and \(FB\) and \(NMSE = 0.0\). Based on the evaluation of many models with many field data sets, model acceptance criteria for those measures were developed [4, 21]:

- fraction within a factor of two should be equal or greater than 50\% (FAC2 > 0.5),
- \(|FB| < 0.3\) or \(0.7 < MG < 1.3\),
- values of \(NMSE\) and \(VG\) should be less than 1.5 and 4, respectively.
4. Statistical Evaluation of Short-Term Forecasts for Krakow

In this study a brief comparison of prognostic and observed concentrations of selected air pollutants in Krakow was carried out. Prognostic observations regarding the upcoming daily average concentrations of PM10, PM2.5, SO₂, NO₂ and O₃ were systematically collected over a period of April 2014 – March 2015 from the Wrota Małopolski web service of the Marshal Office of Małopolska Voivodeship [24]. Those values were subsequently confronted with the corresponding measurements recorded at the urban background measuring station (Bujaka St., Krakow) at the same time. This station belongs to the surveillance grid of the Voivodeship Environmental Protection Inspectorate in Krakow and it has been operational since 2010. Continuous measures at this station are performed for the following substances: PM10, PM2.5, nitrogen oxides (NOₓ), nitrogen dioxide (NO₂), nitrous oxide (NO), sulphur dioxide (SO₂) and ozone (O₃). It is located in the southern part of Krakow (geogr. 50.010575; 19.949189) in a residential area. There are no significant emission sources in the area surrounding the station [13].

Below the time series of daily average modelled and measured concentrations of the analysed air pollutants are presented (Figs 2–6). They are supplemented by the values of Spearman correlation coefficients (Tab. 1) and the results of statistical evaluation of the examined modelling system (Tabs 2, 3). Due to the variability of the observed concentrations during the year, the error statistics have been calculated with additional division of the analysed time period to non-heating season (April – September) and heating season (October – March).

![PM10 concentration vs time](image)

**Fig. 2.** Time series (April 2014 – March 2015) of forecast and measured daily average concentration of PM10 [μg/m³]

Predicted PM10 concentrations are properly modelled by the evaluated forecasting system as evidenced by the high correlation coefficient (0.79) and the value of FAC2 = 0.801. However, over the whole analysed period of time the model tends
to slightly overestimate the observed concentrations of PM10 (MBE = 12.70 μg/m³). Higher rates of errors have been obtained during the heating season with a maximum of 163 μg/m³ (on December 7th 2014).

In the case of PM2.5 predictions the overestimation tendency of the model is similar to those observed for PM10 forecasts. It is yet consistently lower in value (MBE = 6.97 μg/m³) which provides a slightly higher correlation rate of observations ($r = 0.82$). It is important to note that the percentage absolute percentage errors are comparable ($\text{MAPE}_{\text{PM10}} = 61.17\%$; $\text{MAPE}_{\text{PM2.5}} = 63.38\%$) and better representation of PM2.5 observations may be superficial.

**Fig. 3.** Time series (April 2014 – March 2015) of forecast and measured daily average concentration of PM2.5 [μg/m³]

**Fig. 4.** Time series (April 2014 – March 2015) of forecast and measured daily average concentration of nitrogen dioxide (NO₂) [μg/m³]
Observed variation of the forecasted nitrogen dioxide concentrations during the non-heating season are not coherent with the measured values which affects the lower value of Spearman correlation coefficient ($r = 0.49$). For most of the observed days the evaluated prognostic model is overestimating NO$_2$ concentrations. The values of error statistics are yet not significantly high (MAPE = 54.80%) which allows meeting the model acceptance criteria (FB = −0.254; MG = 0.8; FAC2 = 0.793).

Daily average sulphur dioxide concentrations measured at Bujaka St. station only rarely exceed 20 μg/m$^3$ whereas model predictions indicates the possible occurrence of significantly higher values during the analysed year of observations. Thus, error statistics are great in value with MBE = 14.13 μg/m$^3$ and MAPE = 382.74%. Accordingly, the quantitative performance measures have deficient values (FB = −1.148; MG = 0.250; FAC2 = 0.180) in reference to the assumed model quality criteria.

![SO2 concentration vs Time](image1)

**Fig. 5.** Time series (April 2014 – March 2015) of forecast and measured daily average concentration of sulphur dioxide (SO$_2$) [μg/m$^3$]

![O3 concentration vs Time](image2)

**Fig. 6.** Time series (April 2014 – March 2015) of forecast and measured daily average concentration of ozone (O$_3$) [μg/m$^3$]
Table 1. Spearman correlation coefficients for GEM-AQ model evaluation (April 2014 – March 2015 observations)

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Correlation coefficient (r)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM10</td>
<td>0.79</td>
</tr>
<tr>
<td>PM2.5</td>
<td>0.82</td>
</tr>
<tr>
<td>NO₂</td>
<td>0.49</td>
</tr>
<tr>
<td>SO₂</td>
<td>0.42</td>
</tr>
<tr>
<td>O₃</td>
<td>0.52</td>
</tr>
</tbody>
</table>

High discrepancies between modelled and measured ozone concentrations are particularly evident during the non-heating season. Over warm months of the year GEM-AQ modelling system produces highly overestimated predictions for this air pollutant (MBEₙ₉₅ = 63.95 μg/m³), yet the mean percentage error value remains similar for the heating season (MPEₙ₉₅ = 191.55% and MPEₕ₉₅ = 194.88%). As a result the correlation coefficient denotes a moderate relationship between observations (r = 0.52). Although the model performs better during the heating season (FAC²ₕ₉₅ = 0.524), the overall error statistics indicate insufficient reliability of the short-term forecasts of ozone in the EkoForecast system.

Table 2. Selected error statistics for GEM-AQ model evaluation – daily average concentrations [μg/m³]

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Seasons</th>
<th>MBE</th>
<th>MABE</th>
<th>MPE</th>
<th>MAPE</th>
<th>RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM10</td>
<td>non-heating</td>
<td>9.43</td>
<td>11.34</td>
<td>52.48</td>
<td>56.72</td>
<td>13.28</td>
</tr>
<tr>
<td></td>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>16.14</td>
<td>27.59</td>
<td>56.62</td>
<td>65.83</td>
<td>33.95</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>12.70</td>
<td>19.25</td>
<td>54.50</td>
<td>61.17</td>
<td>25.53</td>
</tr>
<tr>
<td>PM2.5</td>
<td>non-heating</td>
<td>4.42</td>
<td>6.37</td>
<td>45.64</td>
<td>52.63</td>
<td>8.02</td>
</tr>
<tr>
<td></td>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>9.64</td>
<td>20.96</td>
<td>62.82</td>
<td>74.70</td>
<td>26.93</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>6.97</td>
<td>13.50</td>
<td>54.03</td>
<td>63.38</td>
<td>19.68</td>
</tr>
<tr>
<td>NO₂</td>
<td>non-heating</td>
<td>7.08</td>
<td>12.09</td>
<td>31.62</td>
<td>52.56</td>
<td>16.67</td>
</tr>
<tr>
<td></td>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>8.72</td>
<td>14.12</td>
<td>45.95</td>
<td>58.08</td>
<td>18.45</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>7.89</td>
<td>13.09</td>
<td>38.71</td>
<td>54.80</td>
<td>17.58</td>
</tr>
<tr>
<td>SO₂</td>
<td>non-heating</td>
<td>10.99</td>
<td>10.99</td>
<td>392.57</td>
<td>392.57</td>
<td>12.94</td>
</tr>
<tr>
<td></td>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>17.37</td>
<td>17.85</td>
<td>374.44</td>
<td>372.56</td>
<td>23.05</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>14.13</td>
<td>14.34</td>
<td>383.72</td>
<td>382.74</td>
<td>18.58</td>
</tr>
<tr>
<td>O₃</td>
<td>non-heating</td>
<td>63.95</td>
<td>64.26</td>
<td>191.55</td>
<td>192.07</td>
<td>68.76</td>
</tr>
<tr>
<td></td>
<td>season</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>20.23</td>
<td>23.50</td>
<td>194.88</td>
<td>202.77</td>
<td>31.07</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>41.96</td>
<td>43.76</td>
<td>193.23</td>
<td>197.45</td>
<td>53.25</td>
</tr>
</tbody>
</table>
Table 3. Selected quantitative performance measures (US EPA) for GEM-AQ model evaluation – daily average concentrations [μg/m³]

<table>
<thead>
<tr>
<th>Pollutant</th>
<th>Year seasons</th>
<th>Quantitative performance measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>PM10</td>
<td></td>
<td>FB</td>
</tr>
<tr>
<td></td>
<td>non-heating season</td>
<td>−0.309</td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>−0.257</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>−0.274</td>
</tr>
<tr>
<td>PM2.5</td>
<td></td>
<td>FB</td>
</tr>
<tr>
<td></td>
<td>non-heating season</td>
<td>−0.254</td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>−0.212</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>−0.223</td>
</tr>
<tr>
<td>NO₂</td>
<td></td>
<td>FB</td>
</tr>
<tr>
<td></td>
<td>non-heating season</td>
<td>−0.258</td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>−0.251</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>−0.254</td>
</tr>
<tr>
<td>SO₂</td>
<td></td>
<td>FB</td>
</tr>
<tr>
<td></td>
<td>non-heating season</td>
<td>−1.257</td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>−1.085</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>−1.148</td>
</tr>
<tr>
<td>O₃</td>
<td></td>
<td>FB</td>
</tr>
<tr>
<td></td>
<td>non-heating season</td>
<td>−0.857</td>
</tr>
<tr>
<td></td>
<td>heating season</td>
<td>−0.595</td>
</tr>
<tr>
<td></td>
<td>year</td>
<td>−0.775</td>
</tr>
</tbody>
</table>

5. Summary and Conclusions

The main objective of the carried out analysis was to perform a statistical evaluation of the results of a short-term air quality forecasting system for the area of Krakow. Available forecasted data derived from the Wrota Małopolski web service [24] were compared to the urban background measurements at the station located in Krakow, Bujaka St., over the period of April 2014 – March 2015 with heating (October – March) and non-heating (April – September) seasons taken into account.

In general, the analysed observations show a tendency of overestimation in reference to the actual concentrations recorded at the urban background station, which is particularly noticeable during the non-heating season (in the case of sulphur dioxide and ozone). The results obtained indicate a good reliability of PM10 and PM2.5 forecasted daily average concentrations as evidenced by the high Spearman correlation coefficient values (0.79 and 0.82, respectively) as they are also meeting the model acceptance criteria [4]. Despite the rather accurate realization of the short-term forecasts of nitrogen dioxide in relation to the error statistics values, the correlation coefficient indicates moderate correlation rate of the modelled and measured...
observations \((r = 0.49)\). This might be due to the higher variability of forecasted values compared to the actual observations during the analysed time series. In the case of sulphur dioxide and ozone forecasts the values of the statistical measures indicate high discrepancies in reference to the measurement data, which qualifies the assessed model as insufficiently accurate in predicting the upcoming concentrations of these compounds.

It is important to emphasize that a general evaluation of the analysed modelling system has been published annually for the area of Małopolska Voivodeship since 2010. Those reports [19–20] point to the on-going process of improving the model configuration using more recent input data. The extended statistical evaluation presented in this paper is consistent with those reports in relation to most of the concerned pollutants, showing generally good reliability of predicted values with slight overestimation noticeable during the non-heating season (especially for PM10 and PM2.5). Discrepancies regarding the results of the assessment for ozone predictions may be related to different averaging time of data for the evaluation or with the availability of more accurate modelling data for the location of Kurdwanów urban background monitoring station in Krakow.

The conducted analysis highlights the importance of publicly available air quality forecasting data system and its limitations associated with spatial accuracy and data quality. Furthermore, discrepancies in the forecasted air pollutants concentrations are likely to affect the verifiability of the Common Air Quality Index [1, 10] which is calculated on the basis of these values and published at the Wrota Małopolski website for general information about current air quality. An additional examination of the model characteristics is recommended for adequate identification of the causes and nature of the noticed errors in order to improve the air quality management system in Krakow [22].

References


