Ground Based Validation of Copernicus Atmosphere Monitoring Service Data for Kyiv

Andrii Shelestov  
Space Research Institute National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»  
Kyiv, Ukraine  
andrii.shelestov@gmail.com

Hanna Yailymova  
Space Research Institute National Academy of Sciences of Ukraine and State Space Agency of Ukraine  
Kyiv, Ukraine  
anneyailymova@gmail.com

Oleg Samoilenko  
Kyiv Academic University, Department of Mathematics  
Kyiv, Ukraine  
oilsamoilenko@gmail.com

Leonid Shumilo  
Space Research Institute National Technical University «Igor Sikorsky Kyiv Polytechnic Institute»  
Kyiv, Ukraine  
shumilo.leonid@gmail.com

Bohdan Yailymov  
Space Research Institute National Academy of Sciences of Ukraine and State Space Agency of Ukraine  
Kyiv, Ukraine  
yailymov@gmail.com

Abstract— Due to the insufficient number of ground stations for air quality indicators measurement in Ukrainian cities there is a need to use satellite data that provide similar air quality indicators, but cover the entire area of interest. This paper describes the validation of the Copernicus Atmosphere Monitoring Service (CAMS) based on the publicly available ground measurement of the air quality indicator PM2.5 for Kyiv for correct using of satellite information in the future.

Keywords—CAMS, air quality, satellite data, validation, neighbouring countries, PM2.5

I. INTRODUCTION

Currently, there are a large number of projects in the world that develop and use information technology to aggregate, analyze and visualize information on air pollution in cities to timely inform humanity about possible dangers [1]. The sources of information for such monitoring systems are both ground air quality measurement stations and satellite information (including Copernicus Sentinel-5p, Sentinel-3, CAMS model data). Due to coarse resolution of CAMS data in-situ information can be used for upsampling of satellite-derived air products and for calibration data from different sources.

The issue of validation of global products is still relevant and is being addressed by many countries. In particular, Chinese universities, analyzing and comparing the CAMS PM2.5 Forecast’s data with ground data for China and the United States, concluded that the accuracy is not satisfactory due to the spatial resolution of 40 km [2]. According to the official report of the Copernicus Atmosphere Monitoring Service, validation of CAMS data was officially conducted for such areas as Scandinavia, Mid-Europe, Mediterranean, Iberian Peninsula, France, East Europe, British Isles, and the Alps. For some of them, the correlation of ground data with CAMS data was high (0.6-0.8), and for some very low (0.2-0.3) [3]. That is why the issue of correct validation remains open for Ukraine.

Ukraine has already taken the first steps to create air monitoring systems in large cities - significantly increased the number of stations for measuring air pollution [4]. However, the greatest activity still occurs in large cities, as they make the largest contribution to air pollution. Therefore, the question of the absence or small number of ground posts for air quality measurement indicators in less populated cities or villages remains open. It is recommended to use satellite or simulated open data to fill the gaps where there are no ground measurements. However, satellite and model data that measure and model air quality indicators are still not validated and verified for the territory of Ukraine. Besides that, in-situ data often contain from so-called outliers and may have a lot of errors in measurements. In this regards, all the available data should be used together. Therefore, this paper considers the open issue of validation of satellite and model data based on available ground-based measurements for the territory of Kyiv with the possibility of further expansion to the entire territory of Ukraine.

II. INPUT DATA

Data collected from ground air quality monitoring stations for the city of Kyiv from the open resource SaveEcoBot will be used for data analysis and comparison [4]. The choice of such a pilot area is due to the fact that Kyiv is one of the largest cities in Ukraine, which has an anthropogenic impact on air quality and a sufficient number of ground stations for qualitative statistical comparison with satellite information. CAMS model data are selected as global data for the territory of Ukraine, which will be validated in this paper. In this work, the data for 2019 and 2020 were used for analysis.

The pollutant under study is PM2.5. According to research conducted in [5], [6], PM2.5 is the most significant pollutant used to monitor overall air quality in international projects, as well as to monitor sustainable development goals, in particular the SGD 11.6.2 indicator.

A. Ground Data

The experiment used ground-based measurement data provided by different providers and aggregated on a single SaveEcoBot platform [4]. The main providers of air quality information are Eco City, Lun Air and Save Dnipro. The Eco City public monitoring project is developing public monitoring of pollution levels not only of fine dust particles PM10 and PM2.5, but also of toxic pollutants – nitrogen dioxide, carbon monoxide, formaldehyde, ammonia, ground-level ozone and radiation. LUN's social project uses PM1.0, PM2.5 and PM10 sensors to monitor air quality, and SaveDnipro
provides information on the following air parameters: PM$_{2.5}$, PM$_{10}$, temperature, relative humidity, and atmospheric pressure.

The total number of air quality measurement posts for Kyiv is 213, of which 52 did not provide any measurements during the period of its existence, 137 measure PM$_{2.5}$ (Fig. 1), of which 89 are active (as of November 2020). The frequency of measurements depends on the station and can vary from every second to hourly information.

B. Copernicus Atmosphere Monitoring Service Data

Copernicus Atmosphere Monitoring Service (CAMS) [7] has been developed for sharing data and processed information about the Earth’s surface, aerosols, ozone and other reactive gases, aiming to support policymakers, business and citizens with enhanced atmospheric environmental information. In particular, the service provides daily hourly data PM$_{2.5}$ and PM$_{10}$ in the form of close to real-time analysis for Europe and Ukraine with a spatial resolution of 11 km. The Fig. 1-left presents the coverage of the territory of Ukraine by CAMS data for April 18, 2020 (19:00). Active fires, which were determined according to own methodology according to satellite data of high spatial resolution Sentinel-2 of the European program Copernicus, are marked in red [8]. The developed methodology is the basis of a project to determine fire safety indices, which is also performed by the authors within the NRFU Competition 2020.01/0268 «Information technology for fire danger assessment and fire monitoring in natural ecosystems based on satellite data».

As can be seen from Fig. 1-right, one pixel of CAMS data gets a large number of ground air quality measurement posts. That is why there is an important task of aggregation of ground and satellite data, satellite data validation, as well as, in the future, increase the spatial distinction of satellite data using deep learning algorithms.

III. METHODOLOGY

Before starting the validation of satellite data, we need to have an understanding of the reliability of the data that will be analyzed. In particular, the first step is to identify and eliminate the outliers observed in the data. These outliers can be caused by several factors - the failure of the ground station (error in determining the indicators of air pollution), as well as the high sensitivity of the sensor to sudden changes in the environment (sharp deterioration of air quality with subsequent rapid recovery). Such cases (outliers) were detected and eliminated from the data set using a boxplot or box-and-whisker diagram [9].

We also note that we validate rather rough global model data (with spatial resolution of 11 kilometers) with ground data, indicating that the major peaks in ground data physically cannot be identified on global data.

The next step in data analysis is to compare the outliers-cleaned data with the CAMS model data. To do this, the ground data were aggregated so that they could be easily compared with the CAMS data, namely, calculated the average hourly values of the PM$_{2.5}$ pollutant for each station.

Two commonly used statistical metrics, including correlation coefficient (R) and root mean squared error (RMSE) were hereby calculated between spatially and temporally co-located ground PM$_{2.5}$ measurements and CAMS PM$_{2.5}$ to quantitatively evaluate the accuracy and uncertainty of the latter. Mathematically, these metrics can be derived from the following equations:

\[
R = \frac{\sum_{i=1}^{n} (o_i - \bar{o})(p_i - \bar{p})}{\sqrt{\sum_{i=1}^{n} (o_i - \bar{o})^2} \sqrt{\sum_{i=1}^{n} (p_i - \bar{p})^2}}
\]

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (p_i - o_i)^2}
\]

where $o_i$ denotes ground-based PM$_{2.5}$ measurements and $p_i$ represents the CAMS PM$_{2.5}$ data, respectively. $o$ and $p$ are arithmetic means of the observed and forecasted PM$_{2.5}$ concentrations, respectively, while $n$ denotes the number of data pairs.
Since we will be interested in how to predict or model ground observations with CAMS data, we will also be interested in the coefficient of determination $R^2$ between ground data and CAMS data, which is determined by the following formula (in our case, the forecast data will be ground data):

$$R^2 = \frac{\sum_{i=1}^{n}(i - \hat{i})^2}{\sum_{i=1}^{n}(i - \bar{i})^2},$$

where $\hat{i}$ - simulated value of air quality according to CAMS.

**IV. RESULTS**

After identifying the outliers in the ground datasets, the statistical indicators ($R$, $R^2$ and RMSE) were calculated, which are presented in Fig. 2. For 114 stations, the correlation between ground data and CAMS is more than 0.5, which indicates a sufficiently high quality of CAMS model data. For 124 stations the RMSE indicator does not exceed 15 mg/m$^3$, and for 33 - 10 mg/m$^3$.

The Fig. 3 shows a comparison of daily data from all ground stations for 2019-2020. The R-squared ratio is not very high (0.36 - 0.38), but is stable for both years. The correlation coefficient shows the best values (0.5-0.6), which allows us to conclude that CAMS data can be used for Kyiv in analytical problems. Excluding from the survey those stations that are statistically anomaly (their values are very different from neighboring stations), all statistics increase for both years.

Also, from the dependence graph it is seen that the value of PM$_{2.5}$ from ground stations mostly exceeds the data from CAMS. There is a logical explanation for this - it is a spatial distinction of CAMS data. It does not allow to increase the variability of the values of the contaminant and thus increase the accuracy of CAMS products. A great advantage of satellite and model data for air quality monitoring is the continuous coverage of the study area with data, which cannot be guaranteed in the presence of ground stations only.

In Fig. 4 shows graphs comparing different ground stations within a single cell of CAMS data. Red shows CAMS data, green shows data from a specific ground station that is geographically within the investigated cell, and blue shows the average PM$_{2.5}$ values for all ground stations within this cell.

The characteristic jump for 100-110 days of the year (April 2020) is associated with a sharp deterioration in air quality in Kyiv due to fires in the Zhytomyr region and the Chornobyl zone. Stations with a red outline are defined as anomaly and removed from analysis because the values in them are much larger than the average value for all stations within a given cell.
V. CONCLUSIONS AND FURTHER STEPS

The work carried out has shown the effective joint use of model data and data from ground stations. In particular, this paper describes the results of a comparative analysis of PM2.5 air quality based on ground data and CAMS model data. This work will be the basis for further use of CAMS data for the whole territory of Ukraine and construction of air quality monitoring models.

As a result of the analysis, it is possible to draw the following conclusions. CAMS data are reliable global data that can be used in scientific and practical research in Ukraine. To improve the quality and increase compliance with ground data, scientists are faced with the task of developing mathematical models and methods to improve the spatial resolution of satellite information that have air quality products and CAMS data. A network of ground stations can be used as reference points, but it should be used in mind that they have a certain property to fail after a certain time of operation. Therefore, for reliability, within the project, which is implemented under a grant from the National Research Fund of Ukraine, will be used own sensors for measuring air quality, both portable and stationary.

It is also planned to conduct the deep analysis of air quality data in Kyiv. In particular, validation not only of PM2.5, but also of other air quality indicators, as well as improvement of validation methodology by using ground cover maps, which are obtained by own algorithms according to satellite data of high spatial resolution [10], [11]. It is also planned to increase the scope of validation to the entire territory of Ukraine.

ACKNOWLEDGMENT

The authors acknowledge the funding received by the National Research Foundation of Ukraine from the state budget 2020.02/0284 “Geospacial models and information technologies of satellite monitoring of smart city problems” (NRFU Competition “Leading and Young Scientists Research Support”).

REFERENCES

[1] SMart URBan Solutions for air quality, disasters and city growths (SMURBS).