Improving performance of oxidant prediction using machine learning by optimizing input data based on atmospheric science knowledge

*Tomohiro Sato[1], Peijiang Zhao[1], Koji Zettsu[1]

1. National Institute of Information and Communications Technology

[Background] Air pollution was the cause of 3.7 million premature deaths globally in 2012 (WHO, 2014). In Japan, the amount of oxidant (Ox) keeps high level or has been still increasing, although other air pollutants, such as NO_2 and $PM_{2.5}$. has been decreasing. High level of Ox affects human health and crop yields (e.g., Zheng et al., 2018). There are many factors for Ox generation, such as precursors especially NO_2 and VOC, sunlight and wind. They are non-linearly connected, thus, modeling of physical and chemical processes for Ox amount prediction is difficult and requires large calculation costs. There are many studies for Ox prediction using machine learning (ML) as an alternative of traditional modeling. Keller et al., (2017) showed that the random forest demonstrated the GEOS-Chem model calculation of ozone (O_3), major component of Ox, within an error less than 10% during the first 10 days with a 1/100 calculation cost. Other ML methods have been tested for O_3 prediction, Multi Linear Regression, Artificial Neural Network (Pires and Martins, 2011), Gene Expression Programming (Samadianfard et al., 2013).

[Aim] Previous studies of air pollution prediction by ML have well developed the model algorithm. Guennec et al., (2016) reported that improvement of input data, such as data augmentation, also increased the prediction performance. In this study, we approached to improve the Ox prediction using ML technique by optimizing the input data.

[Dataset and Method] Ox, especially O_3, is formed via photo-chemical reaction. The O_3 amount increases after the sunrise (e.g., $NO_2 + hn \rightarrow NO + O$, $O + O_2 + M \rightarrow O_3 + M$), takes a maximum during the day, and decreases during the night due to collision with surface and other species. This O_3 trend is observed throughout a year with a variation depending on whether, season and human activity. In this study, we compared the predicted results of Ox amount using two types of input data; in case 1, the Ox amount data itself (V_{obs}) and case 2, difference ($DeltaV = V_{obs} - V_{diurnal}$) from the reference diurnal variation ($V_{diurnal}$). We used data of amounts of major air pollutants (such as Ox, NO_2, SPM, $PM_{2.5}$, SO_2, CO), temperature, humidity, and wind at 16 points in Fukuoka city in 2014 to 2016 (Atmospheric Environmental Regional Observation System). The reference diurnal variation $V_{diurnal}$ was mean in each local hour and year. The ML model that we employed was a Convolutional Recurrent Neural Network (CRNN) based on a Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) (Zhao and Zettsu, 2018). The value of measurement at each point was translated into 2-D map based on the longitude and latitude in the CNN part and the spatial features was extracted. A short-term prediction was performed by the LSTM part. We used the data for the last 24 hours as an input. The output was mean value of Ox amount in all points after six hours.

[Results and Discussion] We demonstrated the Ox prediction using the CRNN model with two types of input data (V_{obs} for Case 1 and DeltaV for Case 2). Root mean squared error and accuracy for the two predictions were compared; Case 1: 10.3 ppb/75.4%, Case 2: 6.7ppb/84.7% (in 2014), Case 1: 11.0 ppb/72.7%, Case 2: 8.3ppb/79.6% (in 2015), Case 1: 9.4 ppb/76.7%, Case 2: 8.6ppb/79.4% (in 2016). Using DeltaV, the accuracy was increased at most 10%. We also compared the predictions in terms of time in a day. In case 1, the model overestimated the Ox amount during the night and underestimated during the day. This bias was significantly improved in case 2 for all years.

©2019. Japan Geoscience Union. All Right Reserved.

[Conclusion and remarks] We demonstrated an optimization of input data for Ox prediction using the CRNN model. Our approach showed the accuracy of the Ox prediction was improved at most 10% using a difference from the reference diurnal variation instead of using a raw data. This study indicates an importance to optimize input data based on atmospheric science knowledge for air pollution prediction by ML.

Keywords: Air pollution, Oxidant, Machine learning, Optimization, Short-term prediction

©2019. Japan Geoscience Union. All Right Reserved.