Contribution ID: **12** Type: **Oral presentation**

Integration of open-source remote sensing data for the investigation of subgrid scale drivers of pollution inferred from model-based inference and machine learning

Tuesday, 7 March 2023 10:00 (20 minutes)

Air pollution is the largest environmental cause of disease and premature death, resulting in more than 9 million premature deaths in 2015 - several times more than from AIDS, tuberculosis, and malaria combined [1-2]. While significant progress has been made in reproducing regional and global ozone fields and their attributions using chemical transport models and data assimilation techniques, including JPL's Multi-mOdel Multi-cOnstituent Chemical data assimilation (MOMO-Chem) framework [3], there is still a challenge to reproduce surface ozone, especially at urban scales relevant for human health impact assessments. The NASA Jet Propulsion Laboratory's Scientific Understanding from Data Science (SUDS) strategic initiative is designed to form interconnected teams between the data science and physical science communities to leverage data science techniques for improving scientific research through revealing new connections in data. The work presented here encapsulates the subgrid scale drivers of pollution SUDS project, mainly focusing on surface ozone and its precursors, inferred from model-based inference and machine learning (ML). Combining machine learning and data science techniques with the domain knowledge and science expertise from the SUDS team, the objectives of this work are to improve our understanding and prediction of global air quality with machine learning.

Using approximately 80 physical and chemical parameters from MOMO-Chem and TOAR-2 observations as the input feature space, maximum daily average 8 h (MDA8) ozone global bias is predicted using a Random Forest Regressor pipeline. This framework has been used to investigate global surface ozone variations and their drivers using explainable ML techniques (Miyazaki et al., in prep). Recent progress in the project has integrated open-source remote sensing data from Google Earth Engine (GEE), including MODIS land cover and Population Density data [4-5]. A generalizable data processing tool has been built for the ML pipeline to automatically extract and process data matching the MOMO-Chem global grid from Google Earth Engine to generate features to improve ML model performance. Early results of bias prediction across the global TOAR-2 ground station network with the added GEE features show an RMSE and R-squared performance improvement, with a 4% and 15.5% respective improvement in January and 2% and 8% performance respective improvement in July experiments. Especially, the high-resolution MODIS data provided additional constraints to improve the representation of high ozone events. Both the MODIS and Population Density features are ranked in the top 15 of permutation importance by the ML model, showing promise for the use of the added datasets in ozone bias prediction with Random Forest, and the potential of leveraging open-source data to improve our understanding of the global drivers of air quality. Future work will include leveraging the GEE tool utilize further high resolution data, such as the MODIS Burned Pixel Area Product and the VIIRS Nighttime Composites [6-7] imagery, to better understand the subgrid-scale drivers of global ozone, which would contribute to the broader TOAR-2 community.

References

[1] R. Fuller, P. Landrigan, K. Balakrishnan, G. Bathan, S. Bose-O'Reilly, and M. Brauer, "Pollution and health: a progress update,"The Lancet Planetary Health, vol. 6, no. 6, pp. E535–E547, 2022.

[2] World Bank, "Topics - Pollution", Online. Available: https://www.worldbank.org/en/topic/pollution

[3] Miyazaki, K., Bowman, K. W., Yumimoto, K., Walker, T., and Sudo, K.: Evaluation of a multi-model, multi-constituentassimilation framework for tropospheric chemical reanalysis, Atmos. Chem. Phys., 20, 931–967, https://doi.org/10.5194/acp-20-931-2020, 2020

[4] Friedl, M., Sulla-Menashe, D. (2019). MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006 [Data set]. NASA EOSDIS Land Processes DAAC. Accessed 2023-01-23 from https://doi.org/10.5067/MODIS/MCD12Q
[5] Center for International Earth Science Information Network - CIESIN - Columbia University. 2018. Gridded Population ofthe World, Version 4 (GPWv4): Population Count, Revision 11. Palisades, New York: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4JW8BX5.
[6] NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), MCD64A1.061 MODIS Burned Area Monthly Global 500m [Data set]. Accessed 2023-01-23 from https://developers.google.com/earth-engine/datasets/catalog/MODIS_061_MCD64A1
[7] C. D. Elvidge, K. E. Baugh, M. Zhizhin, and F.-C. Hsu, "Why VIIRS data are superior to DMSP for mapping nighttime lights,"Asia-Pacific Advanced Network 35, vol. 35, p. 62, 2013.

ML method

Random forest

Main air pollutant of interest

Primary author: DOERKSEN, Kelsey (University of Oxford)

Co-authors: Dr KALAITZIS, Freddie (University of Oxford); Mr MONTGOMERY, James (NASA Jet Propulsion Laboratory); Dr MIYAZAKI, Kazu (NASA Jet Propulsion Laboratory); Dr BOWMAN, Kevin (NASA Jet Propulsion Laboratory); Mr LU, Steven (NASA Jet Propulsion Laboratory); Dr GAL, Yarin (University of Oxford); Dr MARCHETTI, Yuliya (NASA Jet Propulsion Laboratory)

Presenter: DOERKSEN, Kelsey (University of Oxford)

Track Classification: Machine learning applications