Leveraging Deep Learning for Physical Model Bias of Global Air Quality Estimates

Kelsey Doerksen*
OATML
University of Oxford
Jet Propulsion Laboratory
California Institute of Technology

Yuliya Marchetti
Jet Propulsion Laboratory
California Institute of Technology

Kazuyuki Miyazaki
Jet Propulsion Laboratory
California Institute of Technology

Kevin Bowman
Jet Propulsion Laboratory
California Institute of Technology

Steven Lu
Jet Propulsion Laboratory
California Institute of Technology

James Montgomery
Jet Propulsion Laboratory
California Institute of Technology

Yarin Gal
OATML
University of Oxford

Freddie Kalaitzis
OATML
University of Oxford

Abstract

Air pollution is the world's largest environmental risk factor for human disease and premature death, resulting in more than 6 million premature deaths in 2019 [1]. It is challenging to accurately simulate one of the most important air pollutants, ozone (O3), particularly at scales relevant for human health impacts. Meanwhile, the observing network coverage is largely limited. Therefore, the drivers of regional and global ozone trends at these scales remain largely unknown. In this study, we employ a 2-D Convolutional Neural Network (CNN)-based U-Net architecture that predicts surface ozone residuals (i.e., model bias) in Jet Propulsion Laboratory's (JPL) atmospheric composition modeling system. We demonstrate the potential of this technique in North America and Europe, highlighting its ability to better capture physical model residuals compared to a traditional machine learning method. We also assess the impact of incorporating land use information from high-resolution satellite imagery to improve model estimates. Importantly, we discuss how our results can be the first steps to improving our scientific understanding of the factors impacting ozone bias that can be used to guide environmental policy.

1 Introduction

The NASA's 2017-2027 Decadal Survey for Earth Science and Applications from Space stated its priority to define "What processes determine the spatio-temporal structure of important air pollutants and their concomitant adverse impact on human health, agriculture, and ecosystems?" [2]. However, modeling air pollution to date has been challenging due to the complex influences

*Correspondence to kelsey.doerksen@cs.ox.ac.uk

Machine Learning and the Physical Sciences Workshop, NeurIPS 2023.

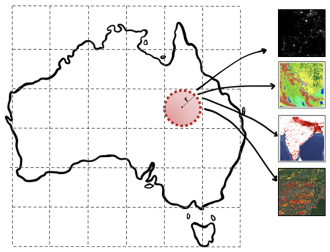


Figure 1: airPy GEE extraction over Australia. Top to bottom datasets: VIIRS nightlight, MODIS landcover, World Population and MODIS Burnt Pixel data. r represents the buffer radius extent selected by the user to match the desired grid size.

from atmospheric chemistry, emissions, planetary boundary layer dynamics and unknown non-linear processes. Recently, the Multi-mOdel Multi-cOnstituent Chemical data assimilation (MOMO-Chem) for tropospheric chemical reanalysis was developed at JPL to integrate various observational information with a chemical transport model to simulate atmospheric composition distributions, including near-surface air pollutants [3]. MOMO-Chem has made significant progress in reproducing large-scale ozone estimates. Nevertheless, there is a gap in finer-scale ozone analysis and drivers of physical model bias relevant for human health assessments.

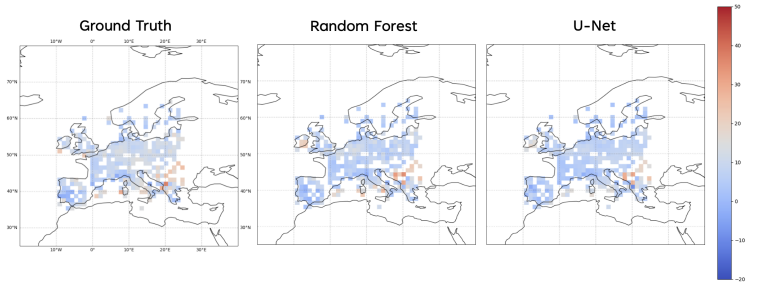
In this work, we investigate the integration of high-resolution satellite data products with the MOMO-Chem physical model parameters to train 2-D U-Net and Random Forest (RF) models to predict daily 8-hour surface ozone bias across North America and Europe, and showcase the performance improvement of deep learning over RF in this context. **Physical model bias** is defined as the difference, or systematic error, between a physical model's output and ground truth observations for a given target variable i.e. surface ozone. This work provides a first application of Deep Learning for predicting and diagnosing MOMO-Chem physical model residuals in an effort to identify the main drivers of air pollution model simulation bias for the globe. The study serves as a first step to help improve Earth system models and to assess and predict the effects of air pollution on human health.

2 Background and Related Work

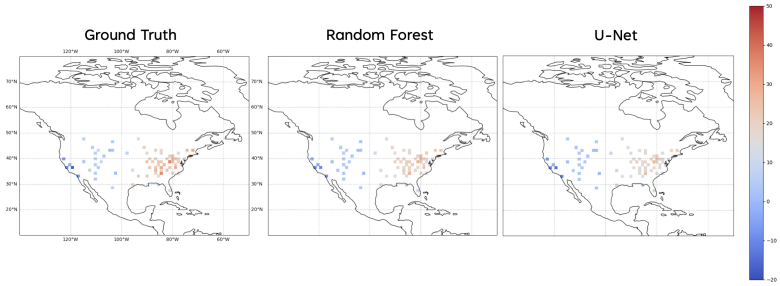
At the Earth's surface, ozone is an air pollutant formed through chemical reactions in the atmosphere when ultraviolet radiationfrom the sun interacts with its precursors, such as nitrogen oxides and volatile organic compounds [4]. The MOMO-Chem model is a state-of-the-art data assimilation framework used to estimate various atmospheric composition, including surface-level ozone. However, it suffers from large systematic estimation errors, i.e. biases, over polluted areas associated with complex chemical and physical processes. This leads to a limited understanding of air quality and its health impacts.

2.1 Machine Learning for Air Quality Physical Model Bias

Deep learning can be leveraged to identify mechanisms driving near-surface pollution and correct for their impact on air quality predictions, thereby improving physical models. Recent work has been developed to capture physical model bias for climate, weather and Earth system models with deep learning in [5,6,7], and traditional Random Forest (RF) ML techniques have been applied to model ozone concentration bias of the GEOS-Chem Chemical Transport Model in China in 2018 [8]. Our work proposes that a U-Net-based architecture is better suited to capture ozone bias than a RF, based on its ability to capture spatial relationships between neighbouring pixels, and through a combination analysis of RF and U-Net results, a clearer picture of the drivers of ozone bias can begin to be uncovered.



(a) Europe



(b) North America

Figure 2: Ground Truth vs RF baseline vs U-Net model average predictions over Europe and North America for 2016 test set. Locations are shown where ground truth data is available from the TOAR network.

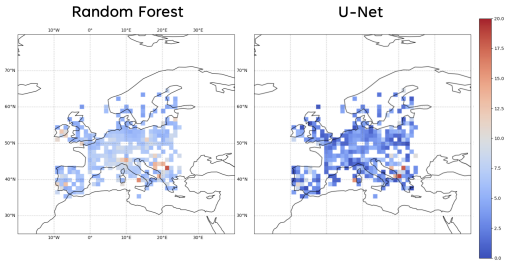
Table 1: Model Input Features

Data Source	**Variables**
MODIS	maxmium, mean, minimum, average percent coverage per land class type
GPWv411	maximum, mean, minimum, variance
MOMO-Chem	ammonia, dimethyl sulfide, nitric acid, carbon monoxide, bromine nitrate, temperature, nitrogen dioxide, peroxyacetyl nitrate, chemical productions of hydrogen oxide radicals, surface pressure, hydrogen superoxide, 1-Pentyne, sulfur dioxide, hydroxide clear-sky longwave radiation flux at surface and clear string outgoing longwave radiation to space

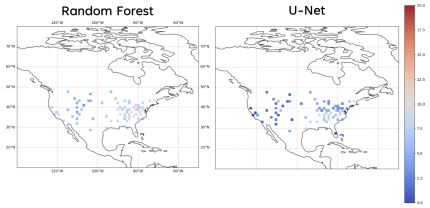
3 Dataset

Input The 2-hourly outputs of MOMO-Chem's 126 meteorological and chemical parameters obtained at 1.125 degree resolution (160x120 latitudinal and longitudinal grid points for the globe) were first down-selected to an emulator version of 16 of the top-ranked variables by importance from RF experiments and domain expert insight and used as input features. Land use data was extracted and processed into an ML-ready format using the airPy package (see 4.2) and includes variables from the MODIS Land Cover Yearly and GPWv411 Population Density products, encompassing 23 features [9,10]. It is hypothesized that the addition of land use information in our models can help to better predict MOMO-Chem bias, as we know that land use changes can impact air quality [11]. The full list of features is detailed in Table 1.

Ground Truth Ground truth surface ozone data is provided by the Tropospheric Ozone Assessment Report (TOAR) database which contains the world's largest collections of near-surface ozone measurements [12]. The majority of TOAR stations are located across North America and Europe, with some coverage over Asia, and virtually no coverage over remote areas and across the Global South. The dense TOAR coverage over North America and Europe allows us to focus our study on these regions as a first analysis, with future work to extend to regions with very sparse ground truth. Bias in the context of this work is the difference between the MOMO-Chem daytime 8-hour average



(a) Europe



(b) North America

Figure 3: RF vs U-Net Average RMSE over Europe and North America

surface ozone output and the ground-based TOAR observation of daytime 8-hour average surface ozone.

4 Methodology

4.1 Experimental Setup and Model Architecture

Individual models are trained for North America and Europe for the Summer season (June-August) respectively. We use a CNN-based model inspired by the U-Net architecture that accounts for spatial context in the data during training. The deep learning model is compared to a RF baseline. The North American extent experiments train with multi-channel images (arrays) of size 31x49 covering latitude, longitude ranges of (20, 55) and (-125, -70) respectively, and the European extent experiments train with multi-channel images of size 27x31 that cover latitude, longitude ranges of (35, 65), (-10, 25) respectively, matched to the MOMO-Chem grid resolution. Models are trained on two feature-space configurations (number of channels) to compare the performance with and without land use information; Experiment 1 uses 16 MOMO-Chem features and Experiment 2 uses 16 MOMO-Chem features plus 23 GEE MODIS and population data. We employ a channel-wise z-score normalization of the input to improve training.

We adapted the U-Net architecture to include dropout after each of the 2D convolutions in the Double Convolution module which was found to improve performance over batch normalization regularization for this application [13]. We train with Adam optimization and a weight decay of 1e-3, constant learning rate of 1e-2, training for 200 epochs and dropout rate of 0.1. Training and validation was done using samples from years 2005-2015 with a 90-10 train/validation split and testing on samples from 2016. To combat the ground-truth sparsity impacting U-Net performance, NaN locations were masked before calculating the loss and back propagating during training.

4.2 Integrating Land Use Information from Satellite Data with airPy

The airPy python package was developed to extract high-resolution surface information from Google Earth Engine (GEE) and compute relevant metrics for air quality studies for any location on the Earth for use in ML models and other statistical analysis. For a given latitude, longitude point and specified area of interest buffer extent, airPy extracts land surface data for the specified GEE product and calculates relevant statistical features that can match any grid resolution (Figure 1). To support open science, airPy is open-sourced and is available at: airPy.

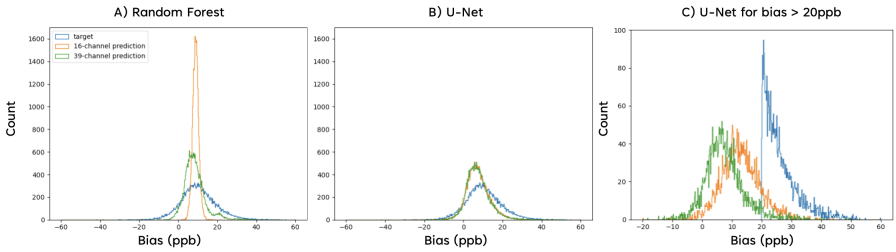


Figure 4: From left to right: Bias prediction histograms over Europe for Random Forest, U-Net, and U-Net predictions compared to bias ground truth greater than 20 parts per billion (ppb)

5 Results

5.1 Random Forest vs U-Net

Testing model performance during 2016 Summer months, the U-Net outperforms the RF baseline in North America with an average rmse of 9.6 vs 10.1 respectively, and in Europe with an average rmse of 8.5 vs 9.1, respectively. Figure 3a-3b showcases the higher performance of the U-Net model, on average, of capturing MOMO-Chem bias across Europe and North America over the RF baseline between June-August 2016. Intuitively this makes sense, as the CNN-based model includes spatial context during training, and supports our hypothesis that including this information is valuable to better capture ozone.

5.2 Incorporating Land Use Information

Figure 4 supports our hypothesis that including information derived from high-resolution satellite imagery about land use improves model performance for the RF in capturing bias, and in particular bias extremes, over Europe. Interestingly however, this is not the case with the U-Net, where the 39-channel feature space predicts closer to the mean of the bias distribution, in particular for cases of high bias greater than 20 ppb. These results are consistent with the North America experiments. Further investigation into this phenomena is ongoing and future work will focus on imbalanced regression and extreme value prediction problems particular to sparse data to support our investigation into driving factors of bias.

6 Conclusion

In this work, we have shown for the first time the capability of deep learning to estimate physical model bias of surface ozone for the MOMO-Chem framework. The U-Net-based model outperforms the RF baseline for both Europe and North America experiments. Land use information extracted from high-resolution satellite data improves the RF model in capturing bias, but does not show improvement for the U-Net in capturing bias extremes, with further investigations ongoing in this direction. To provide additional value to the scientific community, future work will integrate uncertainty quantification methodology into our model to provide pixel-wise uncertainty estimates for predictions and explore deep learning explainability metrics including SHAP [14].

7 Broader Impact

This work serves as a first step to leveraging deep learning to estimate ozone bias with the objective to improve the MOMO-Chem framework as well as integrating additional land use information into the understanding of the bias. Improved predictions of ozone bias are integral to correct for their impact on air quality estimates to make informed decisions to reduce global air pollution and its adverse health impacts. The development of airPy will reduce computational barriers for the science community in leveraging Earth data that can extend beyond air quality studies to other Earth Science applications.

8 Acknowledgements

KD acknowledges funding from EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines and Systems (Grant No. EP/S024050/1) and co-funding from the Oxford- Singapore Human-Machine Collaboration Programme, supported by a gift from Amazon Web Services. Main part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration (80NM0018D0004).

References

[1] Health Effects Institute. 2019. State of Global Air 2019, www.stateofglobalair.org.

[2] E. National Academies of Sciences and Medicine, "Thriving on our changing planet: A decadal strategy for earth observation from space.," Washington, DC: The National Academies Press., 2018.

[3] K. Miyazaki, K. W. Bowman, K. Yumimoto, T. Walker, and K. Sudo, "Evaluation of a multi-model, multi-constituent assimilation frame- work for tropospheric chemical reanalysis," Atmos. Chem. Phys, vol. 20, no. 931-967, 2020.

[4] Duncan, Bryan. "Surface-Level Ozone." Air Quality Observations from Space, 21 Apr. 2023, airquality.gsfc.nasa.gov/.

[5] Hess, Philipp, Stefan Lange, and Niklas Boers. "Deep Learning for bias-correcting comprehensive high-resolution Earth system models." arXiv preprint arXiv:2301.01253 (2022).

[6] Wang, F., Tian, D., and Carroll, M.: Customized deep learning for precipitation bias correction and downscaling, Geosci. Model Dev., 16, 535–556, https://doi.org/10.5194/gmd-16-535-2023, 2023.

[7] Laloyaux, P., Kurth, T., Dueben, P. D. and Hall, D. (2022). Deep learning to estimate model biases in an operational NWP assimilation system. Journal of Advances in Modeling Earth Systems, 14, e2022MS003016. https://doi.org/10.1029/2022MS003016

[8] Diagnosing the Model Bias in Simulating Daily Surface Ozone Variability Using a Machine Learning Method: The Effects of Dry Deposition and Cloud Optical Depth Xingpei Ye, Xiaolin Wang, and Lin Zhang. Environmental Science & Technology 2022 56 (23), 16665-16675 DOI: 10.1021/acs.est.2c05712

[9] D. Sulla-Menashe and M. A. Friedl, "Mcd12q1 modis/terra+aqua land cover type yearly l3 global 500m sin grid v006," NASA EOSDIS Land Processes DAAC.

[10] C. for International Earth Science Information Network CIESIN Columbia University, "Gridded population of the world, version 4 (gpwv4): Population density," New York: NASA Socioeconomic Data and Applications Center (SEDAC), 2018

[11] Massad, R. S., Lathière, J., Strada, S., Perrin, M., Personne, E., Stéfanon, M., Stella, P., Szopa, S., and de Noblet-Ducoudré, N.: Reviews and syntheses: influences of landscape structure and land uses on local to regional climate and air quality, Biogeosciences, 16, 2369–2408, https://doi.org/10.5194/bg-16-2369-2019, 2019.

[12] Schultz, M. G., et al.: Tropospheric Ozone Assessment Report: Database and Metrics Data of Global Surface Ozone Observations, Elem. Sci. Anth., 5, 58, https://doi.org/10.1525/elementa.244, 2017.

[13] Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. "U-net: Convolutional networks for biomedical image segmentation." Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015.

[14] Lundberg, S. and Lee, S.-I., "A Unified Approach to Interpreting Model Predictions", arXiv e-prints, 2017. doi:10.48550/arXiv.1705.07874.