– Supplementary Material: Appendix –
Environmental Justice and COVID-19 Outcomes: Uncovering Hidden Patterns with Geometric Deep Learning and New NASA Satellite Data

Ignacio Segovia-Dominguez[1, 2*], **Huikyo Lee**[1], **Zhiwei Zhen**[2], **Meichen Huang**[3], **Yuzhou Chen**[4], **Michael Garay**[1], **Daniel Crichton**[1], **Yulia Gel**[2, 5]

[1] Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA
[2] Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080, USA
[3] Department of Neurology, Baylor College of Medicine, Houston, TX 77030, USA
[4] Department of Computer & Information Sciences, Temple University, Philadelphia, PA 19122, USA
[5] Division of Mathematical Sciences, National Science Foundation, Alexandria, VA 22314, USA
ignacio.segoviadominguez@utdallas.edu, huikyo.lee@jpl.nasa.gov, zhiwei.zhen@utdallas.edu, meichen.huang@bcm.edu,
yuzhou.chen@temple.edu, michael.j.garay@jpl.nasa.gov, daniel.j.crichton@jpl.nasa.gov, ygl@utdallas.edu

Appendix A: Further Details of the collected NASAdat Dataset

Data Preprocessing and Format AOD is a measure of the amount of light that atmospheric aerosols scatter and absorb and a monotonic function of air quality related to particulate matter near the ground. We generated daily climatology of AOD using the 19-year observations between January 1st, 2001 and December 31st, 2019 (Figure 1 (a)) and used the climatological AOD in the team's previous studies (Segovia-Dominguez et al. 2021a,b). To calculate a climatological mean for each day of the year, we average 19 observations between January 1st, 2001 and December 31st, 2019. For example, the climatological AOD on January 1st is an average of the 19 New Year's days from 2003 through 2019.

We also provide data on daily climatology of surface air temperature and RH from the Atmospheric InfraRed Sounder (Aumann et al. 2003) as shown in Figures 1 (b) and (c). To fully take advantage of its high spatial resolution, we use surface air temperature and relative humidity from AIRS and CrIs in years 2020 and 2021. For example, GDL models can use topological summaries of the Community Long-term Infrared Microwave Coupled Atmospheric Product System (CLIMCAPS) products as input. The underlying hypothesis to be tested over the next three years is that surface air temperature and RH may affect COVID-19 hospitalization and death indirectly.

The collected dataset include a unique identifier for each county and is saved in the netCDF format. `NASAdat` can be accessed via:

Temperature
DOI: 10.48577/jpl.z31y-2r10
https://doi.org/10.48577/jpl.z31y-2r10
Metadata (url)
https://commons.datacite.org/doi.org/10.48577/jpl.z31y-2r10

Relative Humidity
DOI: 10.48577/jpl.ws86-1q81
https://doi.org/10.48577/jpl.ws86-1q81
Metadata (url)
https://commons.datacite.org/doi.org/10.48577/jpl.ws86-1q81

AOD
DOI: 10.48577/jpl.k37v-y751
https://doi.org/10.48577/jpl.k37v-y751
Metadata (url)
https://commons.datacite.org/doi.org/10.48577/jpl.k37v-y751

By including the Federal Information Processing Standard code (FIPS) of each county, now NASA's atmospheric data in `NASAdat` is easily matched with county level datasets from other public and private entities.

Uniqueness The collected `NASAdat` dataset is unique in multiple aspects. First, long-term AOD observations from a single instrument over the entire CONUS, such as our `NASAdat`, is only available from satellites. While AOD observations are also available from NASA's remote sensing Aerosol Robotic Network (AERONET) stations, AERONET coverage is noticeably sparser. In turn, many previous studies which compare AOD observations from MODIS with those from AERONET report reasonable agreement between the two, which also can serve as an additional measure of data quality control. Second, while NOAA through NCEI provides data on such weather variables as temperature, precipitation, drew point, visibility, etc. Almost all of NOAA's records rely on ground-based stations. As a result, in contrast to `NASAdat`, the NOAA data are limited to the resolution on covered areas across U.S., and many counties are far away from land-based stations which further increases uncertainty in applications requiring better resolution, such as biosurveillance. Third, in comparison to all other existing data, our daily climatologies of temperature and relative humidity provide annual cycles in these variables for each county with the Federal Information Processing Standard Publication 6-4 (FIPS 6-4) code, thereby making it easier to match `NASAdat` with various key biosurveillance, socio-economic and socio-demographic information of the best available granularity (i.e., at a county level) such as COVID-19 hospitalizations, cancer rates, and number of houses with solar panels. Fourth, temperature and relative humidity data for the entire globe including those over ocean are another benefit of using satellite observations

[*] Correspondence Author

when running ML models for different spatial domains other than the US. Fifth, the climatology datasets such as NASAdat can be used to study the impacts of the nation's climate change on various sectors, from digital agriculture to resilience of critical infrastructures to adverse climate events. Moreover, given multiple types of ground truth instances associated with these data, e.g., dust storms and teleconnection patterns, the presented benchmark NASAdat can serve as a test bed for a very broad range of ML tasks such as spatio-temporal forecasting with graph neural networks, transfer learning of climatic scenarios, dynamic clustering, anomaly detection, and multi-resolution pattern matching.

Quality of the dataset NASAdat undergoes standard data quality control checks under NASA guidelines. The original datasets were generated by averaging quality-controlled observations. As a part of retrieval algorithms, a quality flag is automatically assigned to each retrieved value of temperature, relative humidity, and AOD. The algorithms assign a quality flag of each pixel by comparing the observed values with predefined ranges of valid observations. A quality flag is a kind of automated annotation by a machine that is already considered in the original datasets. As such, we were confident about the quality of our newly generated datasets. Due to low-quality retrievals, there exists a small fraction of missing values in the original datasets. As per the standard statistical practice, these missing values are stripped when calculating a spatial and temporal average for each county.
Both MODIS and AIRS missions provide more detailed information on the quality flag.

MODIS
https://atmosphere-imager.gsfc.nasa.gov/sites/default/files/ModAtmo/documents/QA_Plan_C61_Master_2021_09_22.pdf

AIRS
https://docserver.gesdisc.eosdis.nasa.gov//public/project/AIRS/V7_L2_Quality_Control_and_Error_Estimation.pdf

Both AIRS and MODIS datasets cover the entire globe. The total sizes of 6209 AIRS and and 6939 MODIS files are about 2.5 and 4.2 gigabytes respectively. In our processed data, each file for temperature, relative humidity, or AOD has a size of 95 MB.

Maintenance Plan Our previous work (Lee et al. 2018) indicates that even 19 years (2001-2019) may not be long enough to define statistically stable AOD climatology. Also, we recognize that continuous updates are the key for these data utilities, especially for biosurveillance and other time sensitive applications. JPL NASA/-Caltech will update our datasets 2 times per year and also whenever new versions of the NASA products are released through NASA's Distributed Active Archive Centers (DAACs).

In our maintenance plan we are taking advantage from the fact that these benchmark data are one of the first projects within the most recent broader NASA's JPL initiative on hosting datasets, such as these and assigning DOIs so there is persistence for papers, and also capturing the raw and any derived results. As such, JPL will continue updating and maintaining these benchmark data under this broader NASA's initiative, with external access to a hub under the subdomain of jpl.nasa.gov. Our team will keep producing daily temperature, relative humidity, and AOD datasets from AIRS/CrIS and MODIS/VIIRS in a NetCDF format which can serve as input for multiple projects across the ML and atmospheric sciences communities. To take full advantage of the highest spatial resolution, we plan to expand and use level 2 surface air temperature and relative humidity from AIRS and CrIs of next years. With the combination of using NASA front-end servers, NVIDIA DGX clusters at the NASA Center for Climate Simulation, and parallel processing capabilities and elastic scalability of the Advanced Data Analytics Platform (ADAPT) science cloud, we expect to have no issue maintaining our data for years to come as these services will provide us all the resources necessary with no cost to NASAdat end-users.

Appendix B: Further Details on the Experimental Setup

Benchmarking neural network models We benchmark two broad classes of neural networks (i) Recurrent Neural Networks (RNNs): Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) can forecast univariate time series with LSTM hidden units; (ii) Spatio-Temporal Graph Convolutional Networks: spatio-temporal model with the framework of graph convolutional network (GCN) exploit GCN and temporal convolution to capture dynamic spatial and temporal patterns and correlations; we report performances for eight types of state-of-the-arts methods on our benchmark datasets including (1) Diffusion Convolutional Recurrent Neural Network (DCRNN) (Li et al. 2018): diffusion convolution recurrent neural network that captures both spatial and temporal dependencies through random walks on graph and encoder-decoder architecture for multi-step forecasting, (2) Long Short-Term Memory R-GCN (LRGCN) (Li et al. 2019): time-evolving neural network which integrates relational GCN (R-GCN) into the LSTM to fully investigate both intra-time and inter-time relations, (3) Attention Temporal Graph Convolutional Network (A3T-GCN) (Bai et al. 2021): an attention temporal GCN that combines GCNs and GRUs with attention mechanism which can capture both spatio-temporal dependencies and global variation trends; (4) Message Passing Neural Networks with LSTM (MPNN+LSTM) (Panagopoulos et al. 2021): a time-series version of message passing neural networks consists of a series of neighborhood aggregation layers to model in detail the dynamics of the spreading process; (5) Evolving Graph Convolutional Networks (EvolveGCNO and EvolveGCNH) (Pareja et al. 2019): evolving graph convolutional network that utilizes the recurrent model to update the trainable parameters of GCN for understanding and forecasting graph structure dynamics; (6) Graph Convolutional Recurrent Network (GconvLSTM) (Seo et al. 2018): graph convolutional recurrent network model which replaces convolution by graph convolution to extract the spatial-temporal information; (7) Gated Graph Neural Networks for Dynamic Graphs (DyGrEncoder) (Taheri et al. 2019): gated graph neural networks for dynamic graphs which uses a gated graph neural network equipped with standard LSTM for dynamic graph classification.

Why RMSE? In our experiments we use the RMSE metric rather than R^2 since RMSE is the standard metric for validation of predictive models in space-time forecasting (Brockwell et al. 1991). Despite statistical criticism, R^2 is still used in epidemiology. As such, we present a summary of results for R^2. While we find that R^2 for actual observations and hospitalization forecasts with/without AOD are generally similar in CA, in TX and PA R^2 for GCNs *with* AOD tends to be from 0.05 to 0.25 higher than R^2 for the same GCN but *without* AOD, with ranges from 0.6 to 0.88 in PA and from 0.71 to 0.93 in TX. These findings echo our conclusions on contributions of AOD to COVID-19 clinical severity, based on predictive RMSE.

Why Not Regression Models? Furthermore, we do not consider simpler models, such as regression, ARIMA and other Box-Jenkins class of models, because such tools focus only on linear relationships between variables and, as a result, cannot capture nonlinear nonseparable spatio-temporal dependencies of COVID-19 dynamics (and many other infectious diseases with high virulence). In turn, our analysis includes a broad range of DL architectures that allow us to address such nonlinear dependencies. Furthermore, the model consensus analysis presented in our paper enables us to address such pressing question as whether a relative risk to be affected by COVID-19 is higher for some areas due to their higher exposure to poor air quality.

Appendix C: Additional Literature on Environmental (In)justice, COVID-19 and Air Pollution

The COVID-19 pandemic provides a unique opportunity to study environmental (in)justice from both global and local perspectives. A thoroughly literature review suggests that novel dimensions of social problems arise as new data and information comes out. Below, we summarizes some of the newer studies in this trend.

- People of color and poor communities are dis-proportionally impacted by pollution since they have historically been dumped on when it comes to elements that other people do not want (e.g., land usage, facilities, etc) (Wilson et al. 2020).
- Less privileged resident are at greater risk of COVID-19 infection because they live in crowded and inadequate conditions. Gentrification pushes out socially marginalized residents due to the influx of capital in communities which, in turn, transforms neighborhoods economically, socially and demographically (Cole et al. 2021).
- COVID-19 pandemic have produced a social catharsis by showing the racial disparities in health care access and health outcomes, and teaching us that environmental justice goes beyond local hazards and exposure to pollutants (Cooper and Nagel 2021).
- Environmental injustice affects most the lives of children from lower socioeconomic backgrounds and, as a group, children are critical of the intergenerational environmental injustice in society (Rios, Neilson, and Menezes 2021).
- Sustainability discourses keep prioritizing the economy over social issues, thus, keeping environment, in all its forms, as the last pillar to take care of. Social consequences of environmental injustice can potentially be tackle in the future by increasing the emphasis on justice-centered education which, in turn, will overcome future sociopolitical constraints when dealing with environmental issues and their diversified layers of (in)justice (Rodrigues and Lowan-Trudeau 2021).
- During COVID-19 pandemic, the government and industry declared meatpacking as critical infrastructure, thus deepening the already precarious conditions of the meatpacking labor market. As such, stabilization of meat supply chain produced destabilization in health and security of a workforce mostly comprised of people of color; with similar implications for other front-line critical workers (Carrillo and Ipsen 2021).
- Eco-pandemic injustice explains interrelationships between global infectious diseases and socioecological systems, and demonstrate how the current pandemic exposes the structural inequalities contributing to higher mortality in people of color communities. Exploitation of the COVID-19 crisis by governments come in the form of austerity measures, conservative politics and deregulation (Powers et al. 2021).

Acknowledgements

The project has beted by NASA grant 20-RRNES20-0021 under the Rapid Response and Novel Research in Earth Science, the UTSystem-CONACYT ConTex program, and NSF ATD grant DMS 1925346.

References

Aumann, H.; Chahine, M.; Gautier, C.; Goldberg, M.; Kalnay, E.; McMillin, L.; Revercomb, H.; Rosenkranz, P.; Smith, W.; Staelin, D.; Strow, L.; and Susskind, J. 2003. AIRS/AMSU/HSB on the Aqua mission: design, science objectives, data products, and processing systems. *IEEE-TGRS*, 41(2): 253–264.

Bai, J.; Zhu, J.; Song, Y.; Zhao, L.; Hou, Z.; Du, R.; and Li, H. 2021. A3t-gcn: Attention temporal graph convolutional network for traffic forecasting. *ISPRS International Journal of Geo-Information*, 10(7): 485.

Brockwell, P.; Davis, R.; Fienberg, S.; Berger, J.; Gani, J.; Krickeberg, K.; Olkin, I.; and Singer, B. 1991. *Time Series: Theory and Methods: Theory and Methods*. Springer Series in Statistics. Springer New York. ISBN 9780387974293.

Carrillo, I. R.; and Ipsen, A. 2021. Worksites as Sacrifice Zones: Structural Precarity and COVID-19 in U.S. Meatpacking. *Sociological Perspectives*, 64(5): 726–746.

Cole, H. V. S.; Anguelovski, I.; Baró, F.; García-Lamarca, M.; Kotsila, P.; del Pulgar, C. P.; Shokry, G.; and Triguero-Mas, M. 2021. The COVID-19 pandemic: power and privilege, gentrification, and urban environmental justice in the global north. *Cities & Health*, 5(sup1): S71–S75.

Cooper, D.; and Nagel, J. 2021. Lessons from the Pandemic: Climate Change and COVID-19. *International Journal of Sociology and Social Policy*, ahead-of-print.

Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*, 9(8): 1735–1780.

Lee, H.; Garay, M.; Kalashnikova, O.; Yu, Y.; and Gibson, P. 2018. How Long should the MISR Record Be when Evaluating Aerosol Optical Depth Climatology in Climate Models? *Remote Sensing*.

Li, J.; Han, Z.; Cheng, H.; Su, J.; Wang, P.; Zhang, J.; and Pan, L. 2019. Predicting path failure in time-evolving graphs. In *KDD*, 1279–1289.

Li, Y.; Yu, R.; Shahabi, C.; and Liu, Y. 2018. Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting. In *ICLR*.

Panagopoulos, G.; Nikolentzos, G.; Vazirgiannis, M.; None; and None. 2021. Transfer Graph Neural Networks for Pandemic Forecasting. In *AAAI*, volume 35, 4838–4845.

Pareja, A.; Domeniconi, G.; Chen, J.; Ma, T.; Suzumura, T.; Kanezashi, H.; Kaler, T.; Schardl, T. B.; and Leiserson, C. E. 2019. EvolveGCN: Evolving Graph Convolutional Networks for Dynamic Graphs. arXiv:1902.10191.

Powers, M.; Brown, P.; Poudrier, G.; Ohayon, J. L.; Cordner, A.; Alder, C.; and Atlas, M. G. 2021. COVID-19 as Eco-Pandemic Injustice: Opportunities for Collective and Antiracist Approaches to Environmental Health. *Journal of Health and Social Behavior*, 62(2): 222–229. PMID: 33843313.

Rios, C.; Neilson, A. L.; and Menezes, I. 2021. COVID-19 and the desire of children to return to nature: Emotions in the face of environmental and intergenerational injustices. *The Journal of Environmental Education*, 52(5): 335–346.

Rodrigues, C.; and Lowan-Trudeau, G. 2021. Global politics of the COVID-19 pandemic, and other current issues of environmental justice. *The Journal of Environmental Education*, 52(5): 293–302.

Segovia-Dominguez, I.; Lee, H.; Chen, Y.; Garay, M.; Gorski, K. M.; and Gel, Y. R. 2021a. Does Air Quality Really Impact COVID-19 Clinical Severity: Coupling NASA Satellite Datasets with Geometric Deep Learning. In *KDD*.

Segovia-Dominguez, I.; Zhen, Z.; Wagh, R.; Lee, H.; and Gel, Y. R. 2021b. TLife-LSTM: Forecasting Future COVID-19 Progression with Topological Signatures of Atmospheric Conditions. In *PAKDD (1)*, 201–212. Springer.

Seo, Y.; Defferrard, M.; Vandergheynst, P.; and Bresson, X. 2018. Structured sequence modeling with graph convolutional recurrent networks. In *ICONIP*, 362–373. Springer.

Taheri, A.; Gimpel, K.; Berger-Wolf, T.; None; and None. 2019. Learning to represent the evolution of dynamic graphs with recurrent models. In *WWW*, 301–307.

Wilson, S. M.; Bullard, R.; Patterson, J.; and Thomas, S. B. 2020. Roundtable on the Pandemics of Racism, Environmental Injustice, and COVID-19 in America. *Environmental Justice*, 13(3): 56–64.