Does Air Quality Really Impact COVID-19 Clinical Severity: Coupling NASA Satellite Datasets with Geometric Deep Learning

Ignacio Segovia Dominguez
The University of Texas at Dallas and
Jet Propulsion Lab, Caltech
Pasadena, CA91109, USA
Ignacio.SegoviaDominguez@UTDallas.edu

Huikyo Lee
Jet Propulsion Lab, Caltech
Pasadena, CA 91109, USA
huikyo.lee@jpl.nasa.gov

Yuzhou Chen
Southern Methodist University and
Lawrence Berkeley National Lab
Berkeley, CA 94720, USA
yuzhouc@smu.edu

Michael Garay
Jet Propulsion Lab, Caltech
Pasadena, CA 91109, USA
michael.j.garay@jpl.nasa.gov

Krzysztof M. Gorski
Jet Propulsion Lab, Caltech
Pasadena, CA91109, USA
krzysztof.m.gorski@jpl.nasa.gov

Yulia R. Gel
The University of Texas at Dallas
Richardson, USA
ygl@utdallas.edu

ABSTRACT

Given that persons with a prior history of respiratory diseases tend to demonstrate more severe illness from COVID-19 and, hence, are at higher risk of serious symptoms, ambient air quality data from NASA's satellite observations might provide a critical insight into which geographical areas may exhibit higher numbers of hospitalizations due to COVID-19, how the expected severity of COVID-19 and associated survival rates may vary across space in the future, and most importantly how given this information, health professionals can distribute vaccines in a more efficient, timely, and fair manner.

Despite the utmost urgency of this problem, there yet exists no systematic analysis on linkages among COVID-19 clinical severity, air quality, and other atmospheric conditions, beyond relatively simplistic regression-based models.

The goal of this project is to glean a deeper insight into sophisticated spatio-temporal dependencies among air quality, atmospheric conditions, and COVID-19 clinical severity using the machinery of Geometric Deep Learning (GDL), while providing quantitative uncertainty estimates. Our results based on the GDL model on a county level in three US states, California, Pennsylvania and Texas, indicate that AOD attributes to COVID-19 clinical severity in 39, 30, and 132 counties out of 58, 67, and 254 total counties, respectively. In turn, relative humidity is another important factor for understanding dynamics of clinical course and mortality risks due COVID-19, but predictive utility of temperature is noticeably lower. Our findings do not only contribute to understanding of latent factors behind COVID-19 progression but open new perspectives for innovative use of NASA's datasets for biosurveillance and social good.

KDD '21, August 14–18, 2021, Virtual Event, Singapore.

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467207

CCS CONCEPTS

• **Applied computing** → **Health informatics**.

KEYWORDS

climate informatics, health informatics, COVID-19, biosurveillance

ACM Reference Format:
Ignacio Segovia Dominguez, Huikyo Lee, Yuzhou Chen, Michael Garay, Krzysztof M. Gorski, and Yulia R. Gel. 2021. Does Air Quality Really Impact COVID-19 Clinical Severity: Coupling NASA Satellite Datasets with Geometric Deep Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3447548.3467207

1 INTRODUCTION

Since mortality due to COVID-19 tends to be closely linked to a prior medical history of lung and other respiratory diseases [25, 29, 32], ambient air quality and other atmospheric conditions might shed an important light on assessing and predicting the severity of COVID-19 and associated survival rates. Furthermore, with proliferating studies on COVID-19 vaccine acceptance, it becomes much more important to better understand what the optimal strategies for vaccine allocation shall be, while accounting for various latent factors associated with COVID-19 dynamics and, in particular, the increasingly more evidenced impact of polluted air and higher risks of hospitalization due to COVID-19. However, most current studies on the linkage of air quality and severity of COVID-19 manifestation tend to suffer from the following shortcomings. First, the majority of existing results are based on linear models such various forms of multiple linear regression. Second, the reported results do not provide a systematic analysis of the associated uncertainties.

In turn, better understanding the impact of atmospheric factors and air quality on COVID-19 progression and associated mortality is both urgent and critical, not only in terms of efficient responses to the ongoing pandemic (e.g., deploying an adequate health care workforce in areas with expected higher clinical coronavirus severity), but also in terms of forecasting impending hot spots and developing more efficient, timely and fair strategies for vaccine allocation.

Recognizing the need, the current project aims to meet the decision makers' need to allocate scarce medical resources including vaccines for COVID-19 hot spots by addressing the following questions:

(1) Is regional air quality represented by climatological aerosol optical depth (AOD) related to normalized COVID-19 mortality and clinical severity?
(2) Does information extracted from satellite observations of temperature, humidity, and AOD exhibit any predictive utility for forecasting COVID-19 progression and clinical severity? If yes, what are the best ways to extract the most relevant information and what can be said regarding the resulting prediction under uncertainty?
(3) Can we use climatological annual cycles in temperature, humidity, and AOD to perform short- and medium-term forecasting of impending COVID-19 hot spots that can be used by disaster managers to develop more effective responses? If yes, what are the most useful predictors and how can we quantify the uncertainty of the resulting forecasts?

The answers to the three overarching questions above are of critical importance for enhancing our understanding of the hidden mechanisms behind COVID-19 progression and clinical severity, that is, which factors do or do not contribute to COVID-19 clinical course and mortality risks.

In turn, tracking the spatio-temporal dynamics of the spread of COVID-19 may be viewed as semi-supervised classification on two-dimensional spaces (i.e., development of maps of a future COVID spread as a function of temperature, humidity, and AOD maps). This is why we employed the machinery of geometric deep learning (GDL) [7, 24] (i.e., deep learning models that are developed specifically for non-Euclidean objects such as graphs and manifolds for image and shape analysis) with a goal of classifying the intensity of mapped COVID-19 hospitalizations using observations from NASA's satellites.

Our findings based on the GDL delivered forecasts for COVID-19 related hospitalizations in three US states: California, Pennsylvania, and Texas on a county level basis, suggest that contribution of AOD, relative humidity and temperature to COVID-19 hospitalizations may vary drastically from county to county, but overall, AOD tends to be the most important factor of COVID-19 clinical course. Our results open up new possibilities for predictive platforms based on the most relevant NASA's satellite data, ideally suited for forecasting the subseasonal COVID-19 clinical severity at a county scale in the United States.

2 RELATED WORK

Artificial Intelligence (AI) and Machine Learning (ML) for COVID-19 Biosurveillance Over the last few months, we witness a tremendous spike of interest in exploring utility of machine learning (ML) and artificial intelligence (AI) approaches for COVID-19 biosurveillance and forecasts [8, 36]. Arguably, the largest class of such approaches concern integration of models from mathematical biology (i.e., various forms of Susceptible-Exposed-Infected-Recovered (SEIR) and other compartmental models) with machine learning tools, aiming to improve interpretability and yield explanatory insights on COVID-19 dynamics [2, 5].

Several recent studies use long short-term memory (LSTM) type models [1, 3, 6, 12, 33] which show promising results in forecasting COVID-19 progression at the country level and prediction horizons up to 14 days. There are studies using Recurrent Neural Networks (RNN) architectures, often in combination with LSTM [31, 35]. Nevertheless, besides the very few most recent studies [26–28], the analysis of complex relationships between atmospheric conditions, air quality, and COVID-19 dynamics, using AI and ML algorithms remain largely unexplored.

COVID-19 and Atmospheric Conditions From the onset of COVID-19 pandemics, there has appeared a number of contradictory studies on the linkage between atmospheric variables and COVID19. For instance, Islam et al. [20] report that calm, cold, dry and overcast conditions are favorable to the transmission of COVID-19. In contrast, Martins et al. [23] suggest that higher precipitation may result in higher levels of COVID-19 spread. Cai et al. [9] find that there exist no correlation between the growth rate of the epidemics and daily mean temperature. In turn, studies of Chen et al. [11] indicate that temperature of $13 - 19^{\circ}$ C and humidity in the range of 50% -8% are favorable to the transmission of COVID-19. Finally, as suggested by Daneshvar et al. [14], Gupta et al. [18], risky precipitation ranges vary among countries and prediction horizons. More detailed discussion and comparison of recent findings on whether atmospheric conditions affect (or do not affect) COVID-19 dynamics can be found, for example, in [25, 29, 32]. One of the major unifying problem for these contradictory reports is that these studies predominantly focus only on *relatively simple linearized* relationships among atmospheric variables and COVID-19 records and tend to lack uncertainty quantification analysis. The goal of this project is to address those unrealistic linearized assumptions, using the GDL machinery.

3 PROPOSED FRAMEWORK

3.1 Forecasting Problem

Spatio-temporal Graph Neural Networks aim to discover patterns in data by learning from temporal and spatial dependencies simultaneously. In this paper, we represent the connection between adjacency counties as a weighted directed graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}, W\}$, where $\mathcal{V}$ is the node set, $|\mathcal{V}| = N$, $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is an edge set and W is the adjacency matrix with entries $\{\omega_{ij} > 0\}_{1 \le i,j \le N}$ for any $e_{ij} \in \mathcal{E}$ and $\omega_{ij} = 0$, otherwise. Let $P \in \mathbb{Z}_{>0}$ be the number of different node features associated with each node $v \in \mathcal{V}$. Then, a $N \times P$ feature matrix $\boldsymbol{X}_t$ serves as graph signal observed at time t, e.g. number of hospitalizations and atmospheric variables. Let τ be the windows size of past graph signals and h be the time ahead horizon. The Spatio-temporal Graph Neural Network aims to learn a mapping function $\mathcal{F}(\cdot)$ that maps the historical data $\{\boldsymbol{X}_{t-\tau}, \ldots, \boldsymbol{X}_{t-1}\}$ to future data $\{\boldsymbol{X}_t, \ldots, \boldsymbol{X}_{t+h}\}$, given a graph $\mathcal{G}$.

3.2 Predictive Geometric Deep Learning methodology

To model heterogeneous spatial-temporal graph structures as a homogeneous process of diffusion, here we focus on diffusion convolutional recurrent neural network (DCRNN) [22], a state-of-the-art geometric deep learning model for COVID-19 clinical severity forecasting. Based on daily COVID-19 datasets, to capture the spatial

and temporal dependency, we first incorporate the information in spatial dimension to the diffusion process. Here, we infer diffusion using random walk on the graph $\mathcal{G}$, i.e., a Markov process with transition matrix $\boldsymbol{R} = \boldsymbol{W}\boldsymbol{D}^{-1}$, where $\boldsymbol{W} \in \mathbb{R}^{N\times N}$ is the adjacency matrix of $\mathcal{G}$ and $\boldsymbol{D}_{i,i} = \sum_j \boldsymbol{W}_{i,j} \in \mathbb{R}^{N\times N}$ is the diagonal matrix of node degrees. Thus, the stationary geometric scattering can be constructed efficiently based on the transition matrix $\boldsymbol{R}$:

$$\boldsymbol{P} = \sum_{k=0}^{\infty} c(1-c)^k \boldsymbol{R}^k, \tag{1}$$

where $c \in [0, 1]$ denotes the restart probability and k is the diffusion step. The K-steps diffusion convolution between node feature matrix $\boldsymbol{X} \in \mathbb{R}^{N\times P}$ and the filter g_θ can be described as:

$$\boldsymbol{X} \star g_\theta = \sum_{k=0}^{K-1} \left(\theta_{k,1} (\boldsymbol{R})^k + \theta_{k,2} \left(\boldsymbol{R}^\top\right)^k\right), \tag{2}$$

where $\theta \in \mathbb{R}^{K\times 2}$ is the parameter of filter g_θ. Based on the diffusion convolution operator defined in Eq. 2, a diffusion convolutional layer of neural network can be built as follows:

$$\boldsymbol{H}_{:,q} = \sigma\left(\sum_{p=1}^{P} \boldsymbol{X}_{:,p} \star g_{\Theta_{q,p,:,:}}\right) \quad \forall q \in \{1, \ldots, Q\}, \tag{3}$$

where $\Theta \in \mathbb{R}^{Q\times P\times K\times 2}$ is the trainable parameter tensor, $\sigma(\cdot)$ is the activation function, e.g., ReLU. Except for the spatial domain, COVID-19 progression forecasting also involves temporal correlations (i.e., temporal dependency) in temporal domain, we utilize Gated Recurrent Unit (GRU) to learn temporal dynamics. Formally:

$$\begin{aligned}
\boldsymbol{z}_t &= \sigma\left(\boldsymbol{W}_z \left[\boldsymbol{H}_{t-1}, X_t\right] + \boldsymbol{b}_z\right) \\
\boldsymbol{r}_t &= \sigma\left(\boldsymbol{W}_r \left[\boldsymbol{H}_{t-1}, X_t\right] + \boldsymbol{b}_r\right) \\
\hat{\boldsymbol{H}}_t &= \tanh\left(\boldsymbol{W}_{\hat{h}} \left[\boldsymbol{r}_t \odot \boldsymbol{H}_{t-1}, X_t\right] + \boldsymbol{b}_{\hat{h}}\right) \\
\boldsymbol{H}_t &= \boldsymbol{z}_t \odot \boldsymbol{H}_{t-1} + (1 - \boldsymbol{z}_t) \odot \hat{\boldsymbol{H}}_t,
\end{aligned} \tag{4}$$

where $\odot$ is the elementwise product; $\boldsymbol{z}_i$ and $\boldsymbol{r}_i$ are update gate and reset gate, respectively; $\boldsymbol{b}_z$, $\boldsymbol{b}_r$, $\boldsymbol{b}_o$, $\boldsymbol{W}_z$, $\boldsymbol{W}_r$, and $\boldsymbol{W}_{\hat{h}}$ are trainable parameters; $[\boldsymbol{H}_{t-1}, X_t]$ and $\boldsymbol{H}_t$ are the input and output of GRU model, respectively.

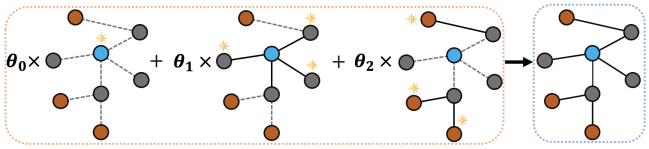


Figure 1: Example K-steps diffusion convolution (where $K = 3$). The blue node represents the starting node, gray nodes represent the 1-hop neighborhood of the starting node, brown nodes represent the 2-hop neighborhood of the starting node, and the star $\star$ denotes the node visited via diffusion procedure. $\theta_1, \theta_2, \theta_3$ are the parameters of the filter.

4 EXPERIMENTS AND RESULTS

4.1 NASA Satellite Datasets

As key predictors, we used daily climatology of surface air temperature and relative humidity (RH) from the Atmospheric InfraRed Sounder (AIRS; [4]) gridded with spatial resolution of $1^\circ \times 1^\circ$ (latitude $\times$ longitude). The underlying hypothesis is that surface air temperature and RH may affect the airborne survival of coronaviruses, and AIRS have provided nearly continuous global coverage of these two key variables since 2002. To calculate climatological mean for each day of the year, we averaged 17 observations between January 1st, 2003 and December 31st, 2019. For example, the climagological temperature on January 1st is an average of the 17 New Year's days from 2003 through 2019. Our hypothesis here is that the COVID-19 severity may be related to the annual cycles of temperature and humidity. As such, using observations for 2020 would have not made any considerable difference in our result. In addition, satellite observations always include missing values. By using the 17-year averages, we could fill most of missing values. The atmospheric variables representing each county are spatially averaged using all of the grid points within the county, so these values are not on regularly spaced grids any more.

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. To investigate whether NASA's AOD observations exhibit prediction skills for the COVID-19 clinical severity and, as such, can be used for detection of impending COVID-19 hotspots, we assess information in AOD distributions from the MODerate resolution Imaging Spectroradiometer (MODIS; [21]) onboard NASA's Terra satellite. Numerous epidemiological studies (e.g., [13, 19, 30, 34]) have associated local hotspots of relatively high AOD (representing poor air quality) with increased risk of morbidity and mortality. Along the same vein, to evaluate relationship of AOD and COVID-19 (if any), we generate daily climatology of AOD using the 19-year observations between January 1st, 2001 and December 31st, 2019. The AOD climatology is then gridded with the same spatial resolution as temperature and RH. Figure 2 presents maps of the mean AOD, temperature, and RH from the MODIS and AIRS instruments over the contiguous United States. Not surprisingly, AOD is relative high in heavily polluted metropolitan areas in coastal counties.

4.2 COVID-19 Datasets

Our experiments have been carried out using collected data in the three states: California, Pennsylvania and Texas. Particularly, our methodology produces daily COVID-19 progression and hospitalization forecasts at county-level resolution. Daily records on COVID-19 cases, deaths and hospitalizations are taken from the CovidActNow project[1] and Johns Hopkins University[2], see [15]. These data sources also include curated time series from official state and county dashboards, the U.S. Department of Health and Human Services, Centers for Medicaid and Medicare Services, New York Times, Covid Tracking Project; and aggregated data sources from the World Health Organization, European Center for Disease

[1] Available at https://covidactnow.org/
[2] Available at https://github.com/CSSEGISandData/COVID-19

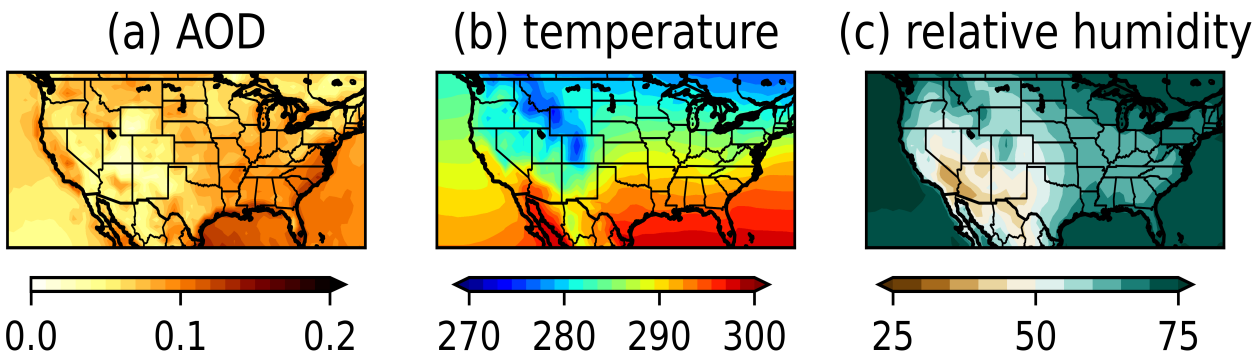


Figure 2: (a) Aerosol optical depth (AOD) in the MODIS averaged for the 19 years between 2001 and 2019. (b) Surface air temperature [K] and (c) relative humidity [%] in the AIRS averaged for the 17 years between 2003 and 2019.

Prevention and Control. For information about the COVID-19 disease progression, and additional modeling resources, we use the MIDAS online portal for COVID-19 modeling research[3]. We build a county connection network, for each state in our study, based on the official County Adjacency File Record Layout[4] provided by the United State Census Bureau.

4.3 Experimental Setting

In our experiments, we use daily data of eleven months of 2020, from February 1 to December 31, and split the graph signals into training set, first 80% of days (268), and test set, last 20% of days (67). We train the DCRNN architecture with lagged daily reported counts, i.e. 5 lags, to produce a 15 days ahead forecasting. The setting for the Recurrent Graph Neural Network methodology (i.e., DCRNN) is as described in Section 3, including a rectified linear activation function. In practice, we use RMSE as metric to assess the predictive capability of the forecasts.

We have special interest in modelling the severity of COVID-19 and evaluate its relationship with atmospheric variables. Here, we consider that the time series of hospitalizations at county level describes the seriousness of the infectious disease, thus we select such time series as our target variable. Since hospitalizations exhibit complex spatio-temporal dependencies, including nonseparability of the covariance structure, we do not normalize time series of hospitalizations at the county level. To verify the impact of adding NASA Satellite data in the Recurrent Graph Neural Network model, we use an experimental setting with next key elements:

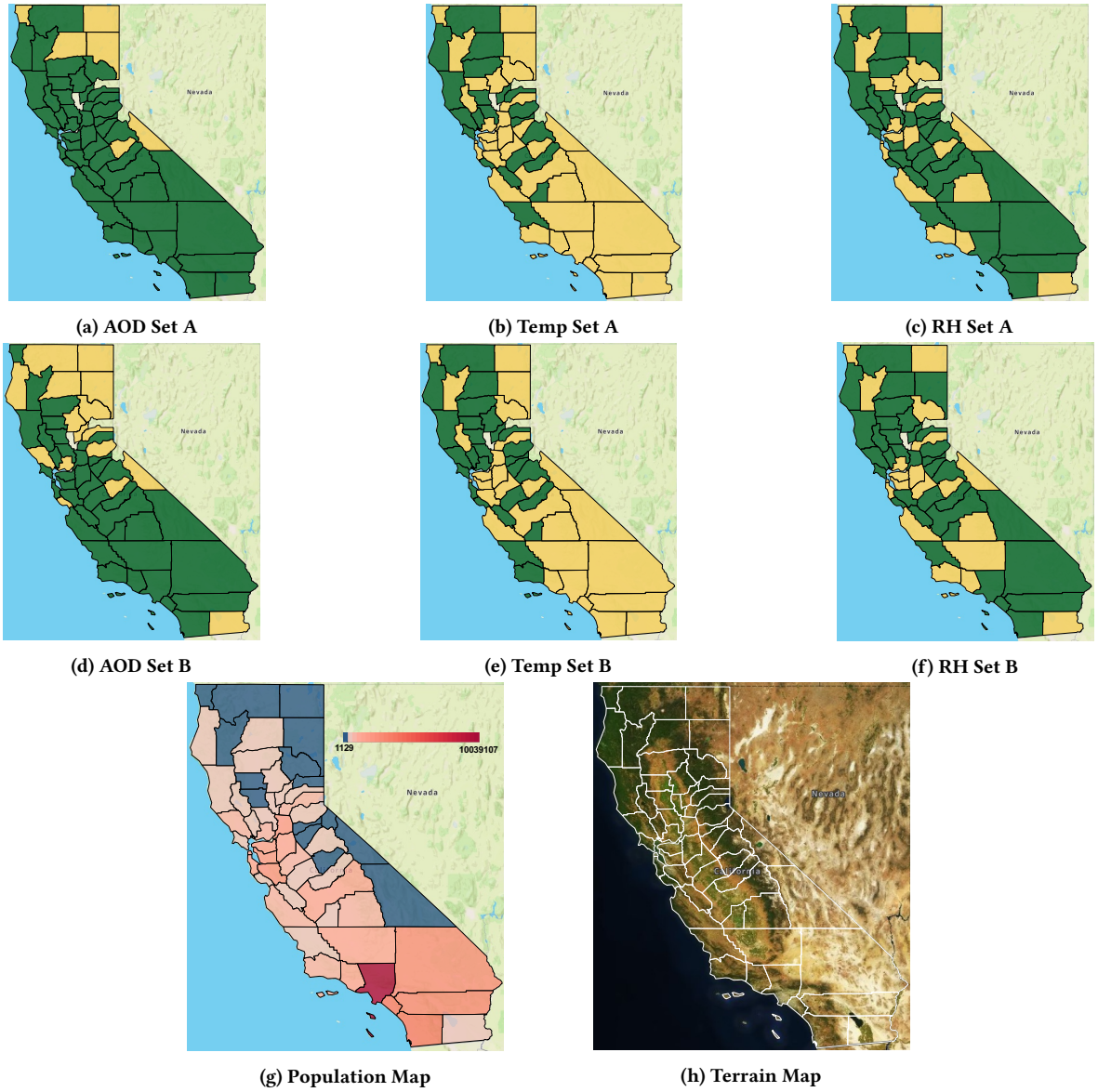
- As primary input variables we use daily historical values of two set of variables, 1) Set A: only number of hospitalizations at county level, and 2) Set B: number of hospitalizations and number of deaths at the county level.
- To contrast the overall performance on each experiment we measure RMSE of global model, average on 10 runs, and perform statistical hypothesis testing on equality of RMSEs.
- We geographically locate the counties in which notice an improvement in forecasting results when adding an atmospheric variable.

Note that Set A is based on historical COVID-19 related hospitalization records, while Set B includes additional historical COVID-19 related death records. We use both sets A and B to evaluate whether any additional historical variables on COVID-19 clinical severity (here in a form of COVID-19 related death records) contain valuable predictive information, also shared by AOD, temperature and RH. That is, in statistical terms, we assess whether AOD, temperature and RH bring any predictive information compared to a broad range of the already existing knowledge on COVID-19 severity. Note that while certainly the COVID-19 related death records and hospitalizations are dependent variables, their dependence is also non-trivial both over space and time. For instance, correlations among death and hospitalization records range from 1% in Trinity County in CA to 96% in Placer County in CA (similar dynamics occurs in TX and PA). This phenomenon is due to multiple factors, ranging from highly varying time periods after initial hospitalization to death to socio-economic disparities in healthcare access [16, 17]. In turn, analyzing Set B allows us to account for additional latent factors not available in Set A, while investigating the impact of AOD, temperature and RH on COVID-19 clinical severity.

Furthermore, there are a few missing counties in our experiment since these counties do not provide hospitalization data on a systematic basis (no imputation is performed here). Similarly, there exist multiple missing at random daily AOD measurements in various counties. We apply a quantile-based imputation method to address AOD missing values.

Finally, we emphasize that in this project we focus on assessing *predictive utility* of AOD, temperature and RH for COVID-19 clinical severity, rather than on forecasting COVID-19 clinical severity per se. As such, following the conventional time series methodology (see, e.g., [10]), we condition our analysis on the predictive model, that is, in our case DCRNN. As mentioned before this GDL model allows us to address nonlinearities in relationships of atmospheric and COVID-19 variables which are inaccessible with simpler linearized approaches such as autoregression. Hence, the considered

[3] Available at https://midasnetwork.us/covid-19/
[4] Available at link: County Adjacency File



(a) AOD Set A (b) Temp Set A (c) RH Set A

(d) AOD Set B (e) Temp Set B (f) RH Set B

(g) Population Map (h) Terrain Map

Figure 3: California's counties which show an improvement (green) in forecasting by adding AOD, Temp and RH during testing phase, using variables in sets A (top) and B (medium) as baseline (see Table 2), along with population and terrain maps.

DCRNN model forms an adequate basis for evaluating contributions of various informational sources to forecasting performance. In turn, comparison of DCRNN with other competing predictive approaches falls under the theme of forecasting COVID-19 clinical severity per se, and we leave it for future extensions of this project.

Model parameters We use the adjacency matrix as weight matrix, where each node represents a county and edges represent border connections between counties. DCRNN layer includes learning of additive bias, has 256 output channels, and set the filter size to 1.0. To reduce overfitting, we apply the dropout regularization method with probability set to 0.2. By executing 10 runs in each experiment, we ensure having enough data to analyze the weight initialization influence on our model performance.

All source codes and datasets are available online[5].

4.4 Averaged relative Contribution of Atmospheric Conditions to COVID-19 Related Hospitalizations

Table 1 summarizes the root mean squared errors (RMSE) for 15-day ahead forecasts of COVID-19 related hospitalizations in California, Pennsylvania, and Texas, averaged over all counties in the

[5]Source code available at Github repository (click here)

corresponding state. We observe that in California RH leads to the highest decrease of average RMSE for 15-day ahead forecasts of COVID-19 related hospitalizations, based both on the previous history of hospitalizations and mortality (Sets A and B). While AOD tends to contribute less than RH, it is also found to be a highly statistically significant predictor for both Sets A and B. In turn, temperature has no utility in explaining COVID-19 related hospitalizations in California. In general, magnitude of forecast errors in California is much higher than in Texas and Pennsylvania, which is likely to be attributed to substantially higher population, more diverse terrain and more heterogeneous land use classes (see panels (g) and (h) in Figures 3, 4 and 5.)

In Pennsylvania, AOD exhibits the highest predictive utility for hospitalizations regardless of the type of input variables already in the model (i.e, Sets A and B), while both temperature and RH demonstrate no positive impact in explaining dynamics of COVID-19 related hospitalizations in Pennsylvania, except of RH yielding a contribution on the border of significance for Set A.

Finally, in Texas both AOD and RH are highly statistically significant predictors of COVID-19 related hospitalizations, regardless of the type of input variables already in the model (i.e, Sets A and B). Remarkably, RH appears to exhibit the highest predictive utility across all scenarios in Texas, except of training set A where the highest gain is yielded by AOD. In turn, as in California and Pennsylvania, temperature is likely to have no impact on COVID-19 related hospitalizations in Texas.

Overall, the contributions of AOD and RH as predictors in the testing scenarios vary across states. In particular, we find that gains of AOD are 4.8% (Set A) and 3.5% (Set B) in CA, 4.9% (Set A) and 3.7% (Set B) in PA, and 16.6% (Set A) and 12.4% (Set B) in TX. The contributions of RH on test sets are 12.9% (Set A) and 9.9% (Set B) in CA, 0.7% (Set A) in PA, and 23.1% (Set A) and 19.5% (Set B) in TX. The obtained phenomena can be partially explained by intrinsic dependency of AOD and RH on terrain variability, while differences among the results for Sets A and B are also likely to be attributed to substantial disparities in health care access across states.

4.5 Relative Contribution of Atmospheric Conditions to COVID-19 Related Hospitalizations on a County Level

We now turn to a finer scale, or county-based evaluation of our results. Table 2 shows numbers of counties in California, Pennsylvania, and Texas, where 15-day forecasts for COVID-19 related hospitalizations have been improved by adding NASA's satellite variables on a county level, while Figure 3 depicts geographical spreads of these counties. Total number of counties depends upon data availability, which are 55 (CA), 60 (PA), and 251 (TX). We find that AOD tends to improve hospitalization forecasts in about half of all counties, outperforming RH in all cases on a county level except of Set A in Texas. In turn, counties where RH enhances 15-day forecasts for COVID-19 related hospitalizations largely appear to be also the counties where AOD exhibits a predictive utility and groups of such counties tend to be more compactly located (Figure 3). These findings echo our earlier hypothesis on dependency of AOD and RH as a function of terrain.

Table 1: Root Mean Squared Errors (RMSE) for 15-day ahead forecasts of COVID-19 related hospitalizations, based on the Recurrent GNN model in three US states: (a) California (CA), (b) Pennsylvania (PA), and (c) Texas (TX), averaged over each state. Each cell contains mean±std. Numbers in bold and *italic* fonts denote the best and second-best results, respectively. Hypothesis testing among RMSEs is performed with one-sided two-sample t-test based on 10 runs; *,**,*** denote p-values of $< 0.1, 0.05, 0.01$ (i.e. significant, stat significant, highly stat significant results), respectively.

Variables	Baseline	+AOD	+Temp	+RH
Set A (train)	210.50±1.71	*196.30*±3.03***	212.60±5.28	**170.80**±5.02***
Set A (test)	492.10±2.96	*468.50*±5.55***	493.10±7.67	**428.30**±8.20***
Set B (train)	199.50±1.77	*190.40*±1.63***	211.50±3.08	**171.10**±5.12***
Set B (test)	476.00±3.09	*459.40*±2.87***	491.50±4.51	**428.90**±8.39***

(a)

Variables	Baseline	+AOD	+Temp	+RH
Set A (train)	8.04±0.17	**7.62**±0.07***	13.37±1.45	8.11±0.11
Set A (test)	98.74±1.20	**93.88**±0.68***	123.68±4.24	*98.05*±0.41*
Set B (train)	7.75±0.10	**7.42**±0.05***	12.91±0.86	7.87±0.05
Set B (test)	96.49±1.28	**92.89**±0.72***	121.86±3.31	97.46±0.37

(b)

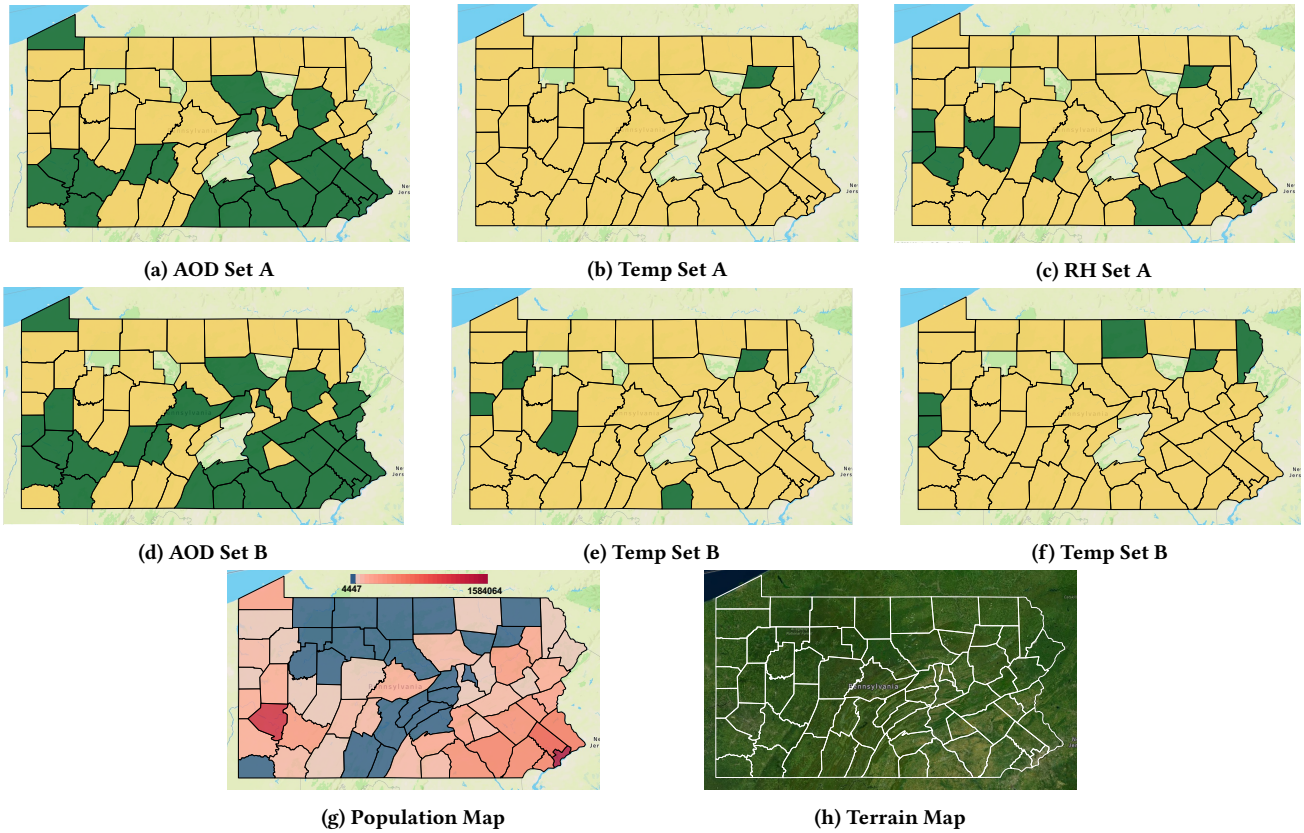
Variables	Baseline	+AOD	+Temp	+RH
Set A (train)	40.00±1.05	*34.61*±1.12***	48.90±1.30	**31.68**±0.70***
Set A (test)	90.50±2.28	*79.22*±2.67***	105.10±2.59	**69.63**±1.81***
Set B (train)	37.30±1.21	**30.89**±0.43***	48.60±2.07	*31.26*±0.89***
Set B (test)	85.10±2.47	*70.94*±1.09***	104.70±4.17	**68.54**±2.11***

(c)

Finally, temperature appears to show none or limited predictive utility in all counties of California, Pennsylvania, and Texas, regardless of the input data already in the model.

Table 2: Number of counties in three US states: California (CA), Pennsylvania (PA), and Texas (TX), where we improved 15-day forecasts for COVID-19 related hospitalizations by adding NASA's satellite variables on a county level. Total number of counties are 55 (CA), 60 (PA), and 251 (TX).

	CA		PA		TX	
Variables	Set A	Set B	Set A	Set B	Set A	Set B
+AOD	**49**	**39**	**25**	**30**	122	**132**
+Temp	21	27	1	5	124	42
+RH	36	35	13	5	**182**	93



(a) AOD Set A (b) Temp Set A (c) RH Set A

(d) AOD Set B (e) Temp Set B (f) Temp Set B

(g) Population Map (h) Terrain Map

Figure 4: Pennsylvania's counties which show an improvement (green) in forecasting by adding AOD, Temp and RH during testing phase, using variables in sets A (top) and B (medium) as baseline (see Table 2), along with population and terrain maps.

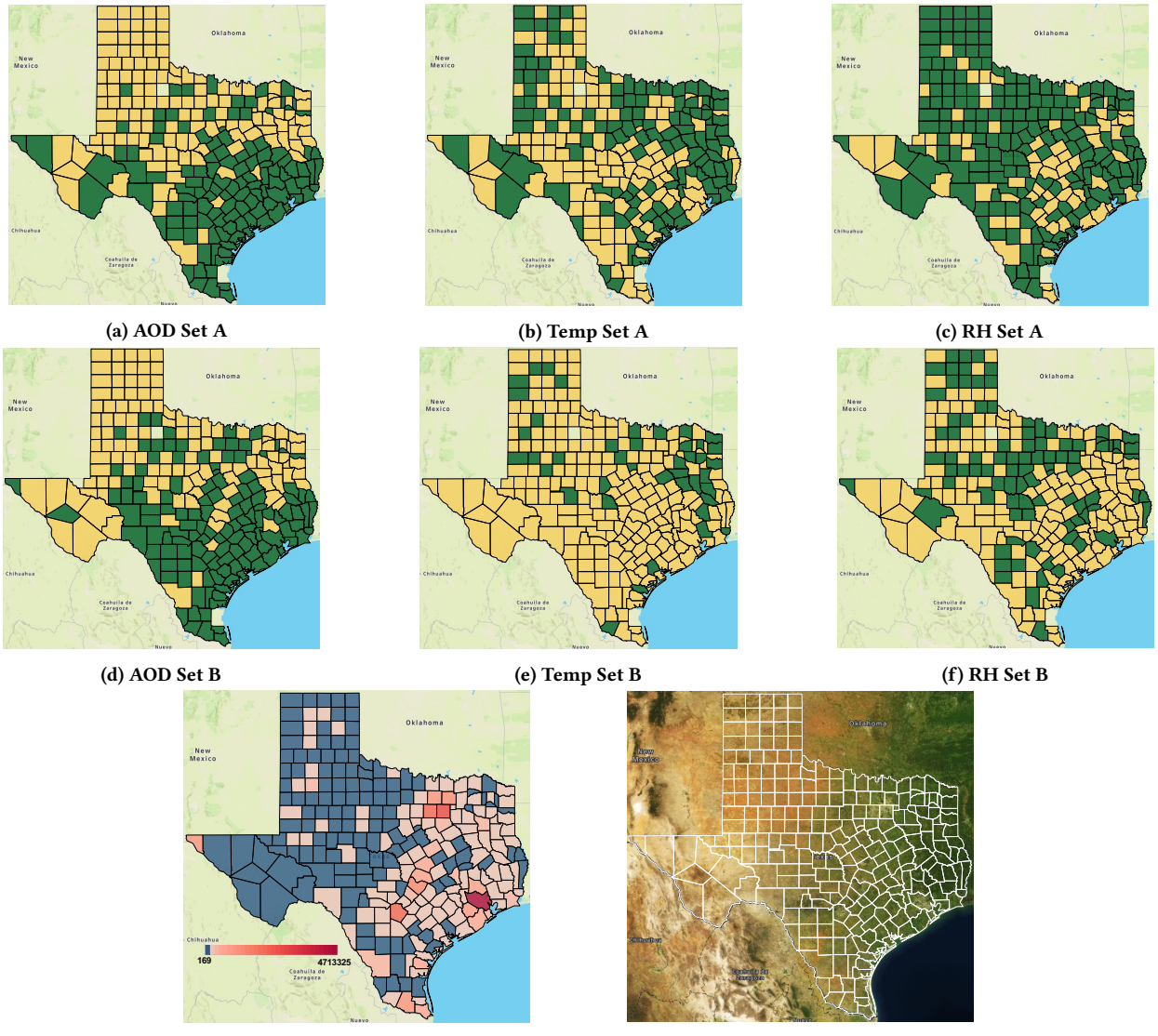
5 LESSONS LEARNED

For the past year, COVID-19 is responsible for more than 2 million deaths worldwide. Recognizing the urgent need for a comprehensive understanding of COVID-19 dynamics, the current study evaluated the utility AOD, temperature, and RH from NASA's satellites in modeling spatio-temporal clinical severity of COVID-19. The GDL model using the three variables as predictors demonstrates the value added by observations from satellites in predicting hospitalization and death due to COVID-19 15 days in advance at a county scale. Although our GDL predictions are made only in the three US States, the results have indicated that COVID-19 severity dynamics cannot be readily modeled with linear combinations of the three variables at each county. Unlike seasonal flu, the COVID-19 severity cannot be explained with temperature. The prediction skill of RH for hospitalization and mortality may be related to airborne transmission of SARS-CoV-2, but this is beyond the scope of this study. To predict the spatio-temporal dynamics of COVID-19 severity, it is important to consider chronic respiratory diseases of residents in each county. Since both AOD and RH used in this study have been consistently observed by a single instrument for the last two decades, the climatological AOD and RH maps could reflect underlying structural properties and quantify the dynamics of the topological properties in ambient air quality and associated risk of chronic respiratory diseases. As such, the predictability obtained by AOD and RH may result from the similarity of spatio-temporal variability between risk of respiratory diseases and COVID-19 clinical severity.

6 PATH TO DEPLOYMENT

To provide a highly available, globally accessible repository and searchable catalog for the GDL model, we will release Singularity images of the GDL model through the OpenNEX (Open NASA Earth eXchange; https://opennex.org) Science App Store. OpenNEX app store is a community-driven platform where scientists can securely publish their codes, application programming interface (API), workflows, containers, and machine images. The OpenNEX app store also provides intelligent recommendation engine to find images relevant to a user's research, the wish-list facility where users can call for specific customizations, and flexible options for integrating newly generated or updated images into its searchable catalog.

The forecast of COVID-19 clinical severity using the GDL model will be expanded for the entire US states by speeding up the simulations on NVIDIA DGX clusters at the NASA Center for Climate Simulation. Parallel processing capabilities and elastic scalability of the Advanced Data Analytics Platform (ADAPT) science cloud (https://www.nccs.nasa.gov/services/cloud-computing) will allow



(a) AOD Set A (b) Temp Set A (c) RH Set A

(d) AOD Set B (e) Temp Set B (f) RH Set B

(g) Population Map (h) Terrain Map

Figure 5: Texas's counties which show an improvement (green) in forecasting by adding AOD, Temp and RH during testing phase, using variables in sets A (top) and B (medium) as baseline (see Table 2), along with population and terrain maps.

us to run virtual machines (VMs) for simulating hospitalization and mortality of multiple states. The satellite datasets of AOD, temperature, and relative humidity will be published in a public data repository such as figshare (https://figshare.com).

7 CONCLUSION AND FUTURE WORK

We have explored contribution of NASA's satellite observations of AOD, temperature, and RH as potential predictors of COVID-19 related hospitalizations in three US states: California, Pennsylvania, and Texas, on a county level basis. We have found that while the impact of these atmospheric variables on COVID-19 clinical severity varies from county to county, both AOD and RH appear to deliver consistently strong predictive utilities across all states and all input data scenarios. These findings suggest that NASA's satellite observations of AOD and RH can deliver important complementary insights into modeling which geographical areas are at the highest risks of COVID-19 and that such satellite data shall be necessarily combined with more traditional epidemiological data in order to develop a reliable predictive platform for COVID-19 tracking.

ACKNOWLEDGMENTS

The project has been supported by NASA grant 20-RRNES20-0021 under the Rapid Response and Novel Research in Earth Science, the UTSystem-CONACYT ConTex program, and NSF RAPID grant

DMS 2027793. The authors are grateful to Rishabh Wagh for assistance with data curation and data visualization.

REFERENCES

[1] M. Alazab, A. Awajan, A. Mesleh, A. Abraham, V. Jatana, and S. Alhyari. 2020. COVID-19 Prediction and Detection Using Deep Learning. *Int. J. of Computer Information Systems and Industrial Management Appl.* 12 (05 2020), 168–181.

[2] S. O. Arik, C.-L. Li, J. Yoon, R. Sinha, A. Epshteyn, L. T. Le, V. Menon, S. Singh, L. Zhang, N. Yoder, et al. 2020. Interpretable sequence learning for COVID-19 forecasting. In *NeurIPS*.

[3] P. Arora, H. Kumar, and B. Panigrahi. 2020. Prediction and Analysis of COVID-19 Positive Cases using Deep Learning Models:A Descriptive Case Study of India. *Chaos Solitons & Fractals* 139 (06 2020), 110017. https://doi.org/10.1016/j.chaos.2020.110017

[4] H. H. Aumann, M. T. Chahine, C. Gautier, M. D. Goldberg, E. Kalnay, L. M. McMillin, H. Revercomb, P. W. Rosenkranz, W. L. Smith, D. H. Staelin, L. L. Strow, and J. Susskind. 2003. AIRS/AMSU/HSB on the aqua mission: Design, science objectives, data products, and processing systems. *Ieee Transactions on Geoscience and Remote Sensing* 41, 2 (2003), 253–264. https://doi.org/10.1109/Tgrs.2002.808356

[5] P. Bedi, P. Gole, N. Gupta, and V. Jindal. 2020. Projections for COVID-19 spread in India and its worst affected five states using the Modified SEIRD and LSTM models. *arXiv:2009.06457* (2020).

[6] H. Bouhamed. 2020. Covid-19 cases and recovery previsions with Deep Learning nested sequence prediction models with Long Short-Term Memory (LSTM) architecture. 8 (04 2020), 10–15.

[7] M. M. Bronstein, J. Bruna, Y. LeCun, A. Szlam, and P. Vandergheynst. 2017. Geometric deep learning: going beyond euclidean data. *IEEE Signal Processing Magazine* 34, 4 (2017), 18–42.

[8] J. Bullock, A. Luccioni, K. H. Pham, C. S. N. Lam, and M. Luengo-Oroz. 2020. Mapping the landscape of artificial intelligence applications against COVID-19. *Journal of Artificial Intelligence Research* 69 (2020), 807–845.

[9] Y. Cai, T. Huang, X. Liu, and G. Xu. 2020. The effects of "Fangcang, Huoshenshan, and Leishenshan" hospitals and environmental factors on the mortality of COVID-19. *PeerJ* 8 (2020), e9578.

[10] C. Chatfield. 2013. *The analysis of time series: theory and practice*. Springer.

[11] B. Chen, H. Liang, X. Yuan, Y. Hu, M. Xu, Y. Zhao, B. Zhang, F. Tian, and X. Zhu. 2020. Predicting the local COVID-19 outbreak around the world with meteorological conditions: a model-based qualitative study. *British Medical Journal open* 10, 11 (2020), e041397.

[12] V.K.R. Chimmula and L. Zhang. 2020. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. *Chaos, Solitons, and Fractals* 135 (2020), 109864 – 109864.

[13] M. Coccia. 2020. Factors determining the diffusion of COVID-19 and suggested strategy to prevent future accelerated viral infectivity similar to COVID. *Science of the Total Environment* 729 (2020). <GotoISI>://WOS:000537441700020

[14] M.M. Daneshvar, M. Ebrahimi, A. Sadeghi, and A. Mahmoudzadeh. 2021. Climate effects on the COVID-19 outbreak: a comparative analysis between the UAE and Switzerland. *Modeling Earth Systems and Environment* (2021), 1–14.

[15] E. Dong, H. Du, and L. Gardner. 2020. An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases* 20, 5 (1 May 2020). https://doi.org/10.1016/S1473-3099(20)30120-1

[16] J. P. Donnelly, X. Q. Wang, T. J. Iwashyna, and H. C. Prescott. 2021. Readmission and death after initial hospital discharge among patients with COVID-19 in a large multihospital system. *JAMA* 325, 3 (2021), 304–306.

[17] C. Faes, S. Abrams, D. Van Beckhoven, G. Meyfroidt, E. Vlieghe, N. Hens, et al. 2020. Time between symptom onset, hospitalisation and recovery or death: Statistical analysis of Belgian COVID-19 patients. *Int. J. of Environmental Research and Public Health* 17, 20 (2020), 7560.

[18] S. Gupta, G. S. Raghuwanshi, and A. Chanda. 2020. Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. *Science of the total environment* 728 (2020), 138860.

[19] Y. Han, J. C.K. Lam, V. O.K. Li, P. Guo, Q. Zhang, A. Wang, J. Crowcroft, S. Wang, J. Fu, Z. Gilani, and J. Downey. 2020. The Effects of Outdoor Air Pollution Concentrations and Lockdowns on Covid-19 Infections in Wuhan and Other Provincial Capitals in China. https://doi.org/10.20944/preprints202003.0364.v1

[20] N. Islam, S. Shabnam, and A. M. Erzurumluoglu. 2020. Temperature, humidity, and wind speed are associated with lower Covid-19 incidence. *MedRxiv* (2020).

[21] R. C. Levy, S. Mattoo, L. A. Munchak, L. A. Remer, A. M. Sayer, F. Patadia, and N. C. Hsu. 2013. The Collection 6 MODIS aerosol products over land and ocean. *Atmospheric Measurement Techniques* 6, 11 (2013), 2989–3034. https://doi.org/10.5194/amt-6-2989-2013

[22] Y. Li, R. Yu, C. Shahabi, and Y. Liu. 2018. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. *International Conference on Learning Representations* (2018).

[23] L. D. Martins, I. da Silva, W. V. Batista, M. de Fátima Andrade, E. D. de Freitas, and J. A. Martins. 2020. How socio-economic and atmospheric variables impact COVID-19 and influenza outbreaks in tropical and subtropical regions of Brazil. *Environmental research* 191 (2020), 110184.

[24] J. Masci, E. Rodolà, D. Boscaini, M. M. Bronstein, and H. Li. 2016. Geometric deep learning. In *SIGGRAPH ASIA 2016 Courses*. 1–50.

[25] A. Núñez-Delgado, Y. Zhou, and J. L. Domingo. 2021. Editorial of the VSI "Environmental, ecological and public health considerations regarding coronaviruses, other viruses, and other microorganisms potentially causing pandemic diseases". *Environmental Research* 192 (2021), 110322.

[26] S. Prasanth, U. Singh, A. Kumar, V. A. Tikkiwal, and P. H.J. Chong. 2021. Forecasting spread of COVID-19 using Google Trends: A hybrid GWO-Deep learning approach. *Chaos, Solitons & Fractals* 142 (2021), 110336.

[27] J. Rasheed, A. Jamil, A. A. Hameed, U. Aftab, J. Aftab, S. A. Shah, and D. Draheim. 2020. A survey on artificial intelligence approaches in supporting frontline workers and decision makers for COVID-19 pandemic. *Chaos, Solitons & Fractals* (2020), 110337.

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

[29] Copernicus Climate Change Service. 2021. Climate Data Store – Monthly Climate Explore for COVID-19. https://cds.climate.copernicus.eu/apps/c3s/app-c3s-monthly-climate-covid-19-explorer?delay=selected%20month&year_month=January%202021.

[30] L. Setti, F. Passarini, G. De Gennaro, P. Barbieri, S. Licen, M. G. Perrone, A. Piazzalunga, M. Borelli, J. Palmisani, A. Di Gilio, E. Rizzo, A. Colao, P. Piscitelli, and A. Miani. 2020. Potential role of particulate matter in the spreading of COVID-19 in Northern Italy: first observational study based on initial epidemic diffusion. *Bmj Open* 10, 9 (2020). <GotoISI>://WOS:000576641100012

[31] F. Shahid and A. Zameer. 2020. Predictions for COVID-19 with deep learning models of LSTM, GRU, and Bi-LSTM. (08 2020).

[32] M. H. Shakil, Z. H. Munim, M. Tasnia, and S. Sarowar. 2020. COVID-19 and the environment: A critical review and research agenda. *Science of the Total Environment* (2020), 141022.

[33] P. Wang, X.-Q. Zheng, G. Ai, D. Liu, and B. Zhu. 2020. Time series prediction for the epidemic trends of COVID-19 using the improved LSTM deep learning method: Case studies in Russia, Peru and Iran. *Chaos,Solitons & Fractals* 140 (08 2020), 110214. https://doi.org/10.1016/j.chaos.2020.110214

[34] X. Wu, R. C. Nethery, M. B. Sabath, D. Braun, and F. Dominici. 2020. Air pollution and COVID-19 mortality in the United States: Strengths and limitations of an ecological regression analysis. *Science Advances* 6, 45 (2020). https://doi.org/10.1126/sciadv.abd4049 arXiv:https://advances.sciencemag.org/content/6/45/eabd4049.full.pdf

[35] A. Zeroual, F. Harrou, D. Abdelkader, and Y. Sun. 2020. Deep Learning Methods for Forecasting COVID-19 Time-Series Data: A Comparative Study. *Chaos, Solitons & Fractals* 140 (07 2020), 110121. https://doi.org/10.1016/j.chaos.2020.110121

[36] A. Zeroual, F. Harrou, A. Dairi, and Y. Sun. 2020. Deep learning methods for forecasting COVID-19 time-Series data: A Comparative study. *Chaos, Solitons & Fractals* 140 (2020), 110121.