

An Environmental Data Collection for COVID-19 Pandemic Research

Qian Liu ¹, Wei Liu ^{1,2}, Dexuan Sha ¹, Shubham Kumar ³, Emily Chang ⁴, Vishakh Arora ³, Hai Lan ¹, Yun Li ¹, Zifu Wang ¹, Yadong Zhang ⁵, Zhiran Zhang ^{1,6}, Jackson T. Harris ^{1,7}, Srikar Chinala ⁸ and Chaowei Yang ^{1,*}

- ¹ NSF Spatiotemporal Innovation Center, George Mason University, Fairfax, VA 22030, USA; qliu6@gmu.edu (Q.L.); devilwei@cau.edu.cn (W.L.); dsha@gmu.edu (D.S.); hlan@terpmail.umd.edu (H.L.); yli38@gmu.edu (Y.L.); zwang31@gmu.edu (Z.W.); zhangzhiran@whu.edu.cn (Z.Z.); jackson.t.harris.20@dartmouth.edu (J.T.H.)
- ² College of Land Science and Technology, China Agricultural University, Beijing 100083, China
- ³ Dougherty Valley High School, San Ramon, CA 94582, USA; dh.skumar@students.srvusd.net (S.K.); dh.varora@students.srvusd.net (V.A.)
- ⁴ Albemarle High School, Charlottesville, VA 22901, USA; eca5@k12albemarle.org
- ⁵ School of Geographical Sciences, Nanjing University of Information Science & Technology, 219 Ningliu Road, Nanjing 210044, China; 20191210020@nuist.edu.cn
- ⁶ School of Resource and Environmental Sciences, Wuhan University, 129 Luoyu Rd., Wuhan 430079, China
- ⁷ Department of Geography, Dartmouth College, Hanover, NH 03755, USA
- ⁸ University Preparatory Academy, San Jose, CA 95125, USA; schintala@upatoday.com
- * Correspondence: cyang3@gmu.edu

Received: 4 July 2020; Accepted: 30 July 2020; Published: 3 August 2020



Abstract: The COVID-19 viral disease surfaced at the end of 2019 and quickly spread across the globe. To rapidly respond to this pandemic and offer data support for various communities (e.g., decision-makers in health departments and governments, researchers in academia, public citizens), the National Science Foundation (NSF) spatiotemporal innovation center constructed a spatiotemporal platform with various task forces including international researchers and implementation strategies. Compared to similar platforms that only offer viral and health data, this platform views virus-related environmental data collection (EDC) an important component for the geospatial analysis of the pandemic. The EDC contains environmental factors either proven or with potential to influence the spread of COVID-19 and virulence or influence the impact of the pandemic on human health (e.g., temperature, humidity, precipitation, air quality index and pollutants, nighttime light (NTL)). In this platform/framework, environmental data are processed and organized across multiple spatiotemporal scales for a variety of applications (e.g., global mapping of daily temperature, humidity, precipitation, correlation of the pandemic to the mean values of climate and weather factors by city). This paper introduces the raw input data, construction and metadata of reprocessed data, and data storage, as well as the sharing and quality control methodologies of the COVID-19 related environmental data collection.

Dataset: <https://github.com/stccenter/COVID-19-Data/tree/master/Environmental%20factors>; <https://github.com/stccenter/COVID-19/tree/master/analysis>

Dataset License: CC-BY

Keywords: COVID-19; decision support; rapid response; environmental data; spatiotemporal platform

1. Summary

At the turn of the decade (2019/2020), a highly-infectious virus, COVID-19, suddenly struck China and rapidly spread worldwide. Presently, the virus is essentially controlled in many parts of the world, but infection rates remain severe in regions like South America and the U.S. Moreover, with the gradual reopening of cities around the world, a potential second outbreak is predicted by some epidemic experts. To track, control, and prevent the pandemic from getting worse, a comprehensive virus and related data collection platform is urgently needed for studying, modeling, predicting, and validating the spatiotemporal spread of COVID-19.

A large number of studies have shown that the occurrence, development, and spread of diseases, especially infectious diseases, are closely related to meteorological conditions [1]. Since the outbreak of SARS in 2003, many studies focused on atypical pneumonia and meteorological factors and found that the propagation characteristics and spread of the SARS virus is correlated with meteorological factors such as air temperature, air pressure, cloud cover, and precipitation [2]. In most cases as humidity and wind speed increase, the prevalence of SARS decreases. Both COVID-19 and SARS viruses are coronaviruses, but COVID-19 is more infectious than SARS virus.

Recent studies mentioned that the outbreak of COVID-19 is strongly correlated with meteorological and weather factors (e.g., precipitation, humidity, heat) [3]. For example, average temperature, maximum and minimum temperature, and air quality impact infection rates in the COVID-19 pandemic [4,5]. Similarly, humidity, visibility, and wind speed affect environmental stability and the viability of viruses, while air temperature impacts virus transmission. Furthermore, absolute air temperature and humidity significantly affect COVID-19 transmission [3]. On the other hand, the pandemic has adverse impact on human daily lives, activities, and environment. Liu et al. [5] found a decrease in nighttime light usage especially within the commercial areas as well as the emission of air pollution. It is also proposed that meteorological variables can aid in predicting worldwide outbreaks of COVID-19 and help investigate viral impacts on human [6]. On the other hand, poor air quality accompanied by strong winds likely accelerates the dispersal of the virus [7], which leads to an increase of new COVID-19 infections [8]. Furthermore, environmental factors such as temperature, humidity and air quality are important input parameters for the transmission model of COVID-19 [9,10]. The transmission models are highly sensitive to the accuracy of input predictors [11]. A high-quality data collection of these factors is vital for accurately predicting the COVID-19 spread and outbreak. Therefore, environmental data is crucial for the study of COVID-19 impacts and modeling [12].

To provide integrated and convenient-utilization data sources for users in different communities, a standardized data collection with widely acceptable formats and variables needs to be produced on a stable data platform. More specifically, the COVID-19 data platform should be an integrated technology solution that allows users to access, explore, and acquire COVID-19 relevant data from a pre-processed database (s) [13]. A number of data platforms and repositories have been built by various organizations since the outbreak of COVID-19. The COVID-19 data platform could be provided and maintained by official authoritative organizations [14–18], longtail departments from the news agencies [19], research groups [20], as well as nonprofit organizations (NGOs) [21]. For example, the World Health Organization (WHO) provides international virus data, while local health and medical departments [16–18] publish state/province- and county/city-level data.

Longtail platforms integrate and summarize first- and second-hand COVID-19 data sources into a visualization dashboard and data repository for users' demands. Most platforms provide initial analysis and tracking of COVID-19 case numbers for each administrative scale and region. However, almost all existing COVID-19 data platforms only provide virus case data such as confirmed, suspected, deaths, and numbers of recovered patients. Other relevant factors, especially environmental variables, are rarely mentioned. A comprehensive and complete data collection is necessary to fill this gap. Our proposed COVID-19 related environmental data collection is associated with and distributed through the COVID-19 rapid response platform established and maintained by George Mason University's (GMU) site of the (National Science Foundation) NSF spatiotemporal innovation

center (<https://covid-19.stcenter.net/>) [22], with standardized spatiotemporal data structures in multiple spatiotemporal sales.

This paper offers a comprehensive description of the COVID-19 related environmental data collection. The paper is organized as follows: Section 2 introduces the raw data, derived values, and metadata of the collection; Section 3 describes the methodology concerning how derived values are produced, and data are processed and stored; Section 4 illustrates the data publishing method and provides downloading addresses; and finally, Section 5 introduces the data quality control method.

2. Data Description

2.1. Raw Measurements and Data Sources

2.1.1. MERRA-2 Temperature and Humidity Reanalysis

Temperature and humidity are proven to have close relationships with the spread and control of COVID-19 [23,24]. Our data collection includes reanalyzed temperature and humidity of Modern-Era Retrospective analysis for Research and Applications and Version 2 (MERRA-2) to provide historic and present casting values for the researchers and decision-makers to estimate and predict the spreading trends and patterns of the pandemic. The MERRA-2 provides data dating back to 1980 with a spatial resolution of ~50 km. It includes advances in the system that enabled assimilation of modern hyperspectral radiance and microwave observations, along with GPS-Radio Occultation datasets and additional advances in both the Goddard Earth Observing System (GEOS) model and the Gridpoint Statistical Interpolation (GSI) assimilation system [25]. The influence and spreading trend can only be accurately analyzed and predicted when the climatological factors are removed from the data. The long-term availability allows researchers to exclude these factors such as seasonal cycles and trends from COVID-19 related analyses. Detailed information is shown in Table 1.

Table 1. Raw data information.

Data Sources	Features and Information Offered	Roles in COVID-19 Related Studies	Download Address
MERRA-2	Temperature, humidity, environmental condition	Suitability of virus spread, spread range and rate	https://disc.gsfc.nasa.gov/datasets/M2T1NXSLV_5.12.4/summary?keywords=MERRA2_400.tavg1_2d_slv_Nx
IMERG	Precipitation rate	Spread range and rate	https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHHE_06/summary?keywords=IMERG
NPP VIIRS/DNB	Nighttime light radiance, human activities, community distributions, human-gathering levels	Human activities impact	https://ladsweb.modaps.eosdis.nasa.gov/
Aura-OMI	Concentration of air pollutants	Mortality rate, human activities impact	https://disc.gsfc.nasa.gov/datasets/OMNO2d_003/summary?keywords=omi
Ground-based air pollution observations	Air quality index, air pollution concentration	Mortality rate, human activities impact	http://data.cma.cn/

2.1.2. IMERG Precipitation Estimation

Precipitation is an important climate and weather factor that influences the moisture and humidity of the Earth. Although there is no study published on the relationship between COVID-19 and precipitation, precipitation plays a role in the spread of other infectious diseases [26]. Our data collection introduces Integrated Multi-satellitE Retrievals for GPM (IMERG) as a potential related data source for the study of COVID-19. The IMERG precipitation estimation is a satellite-observation-based rainfall measurement that provides global coverage and spatial resolution of 10 km and temporal

resolution of 30 min [27,28]. It also provides historic datasets since 2014 for the researchers to investigate seasonal trends and conduct a more accurate study between precipitation and COVID-19. Detailed information is shown in Table 1.

2.1.3. NPP/VIIRS Nighttime Light radiance

The nighttime light reflects human activities and economic conditions. Therefore, the impact of COVID-19 on humans can be detected through the investigation and analysis of the radiance values of nighttime light images [5]. Our data sharing platform collects and processes NASA's Suomi-NPP VIIRS-DNB (VNP46A1), archived at NASA's LAADS DAAC data center (<https://ladsweb.modaps.eosdis.nasa.gov/>), before, during, and after the quarantine policies affecting specific regions. The data have a spatial resolution of 500 m and daily temporal resolution, which is preprocessed using NASA Black Marble algorithm [29]. The nighttime light data provide an even broader domain such as remote sensing, economic, humanity, urban planning, medical, etc., to accomplish their specific study related to COVID-19. Detailed information is shown in Table 1.

2.1.4. Aura-OMI Air Pollution Observation

Air pollutants such as NO₂ are important indicators of economic [30] and influence the mortality of COVID-19 [8]. The collection of air pollution is crucial for both the study of economic impact by COVID-19 and spread of the virus. The Ozone Monitoring Instrument (OMI) flies on the National Aeronautics and Space Administration's Earth Observing System Aura satellite launched in July 2004 [31]. The spatial resolution of OMI is 25 km, covering the globe once a day. The OMI measures criteria pollutants such as NO, O₃, NO₂ and SO₂. The US Environmental Protection Agency (EPA) has designated these atmospheric constituents as posing serious threats to human health and agricultural productivity. Many countries take these pollutants into account in the pollution index for the evaluation of air quality. These measurements track industrial pollution and biomass burning, and hence can be used to evaluate pollution levels and emissions changes on large scales, such as global and country-wide. The outbreak of COVID-19 has forced many countries to lock down industrial activities. Therefore, the amounts of various types of pollutants released to the environment were significantly reduced [32]. The spatiotemporal distribution of OMI data will dynamically change with COVID-19 spreads. Detailed information is shown in Table 1.

2.1.5. Ground-Based Air Quality Data

Ground-based air quality data are derived from national and regional meteorological and environmental protection departments all over the world. Right now, the data collection includes air quality released by the China Environmental Monitoring Centre (<http://www.cnemc.cn/sss/>) and the United States Environmental Protection Agency (<https://www.epa.gov/outdoor-air-quality-data>). Ground-based air quality data are generally published in the form of a daily report, across an ambient air quality monitoring network covering four scales: country, province/state, city, and/or county. Concentrations and Air Quality Indices (AQI) of O₃, NO₂, SO₂, PM₁₀, CO and PM_{2.5} can be obtained for the Chinese and American data sources. We will continue to acquire ground-based air quality data from other countries, e.g., UK, EU, Canada, Australia, etc. Detailed information is shown in Table 1.

2.2. Derived Product and Metadata

2.2.1. Daily/Monthly Global Environmental Factors Reprocessing

(1) Daily Reanalyzed 2-m Specific Humidity

Frequency: daily mean value

Spatial Grid: 2D, single-level, full horizontal resolution

Granule size: ~836 k

Dimensions: longitude = 576, latitude = 361

Acquisition Method Description: All the hourly 2-m specific humidity, which is “QV2M” variable in the original dataset, for each day are averaged and stored in “daily_QV2M” variable in the derived daily 2-m specific humidity dataset.

(2) Daily Reanalyzed 2-m Air Temperature

Frequency: daily mean value

Spatial Grid: 2D, single-level, full horizontal resolution

Granule size: ~836 k

Dimensions: longitude = 576, latitude = 361

Acquisition Method Description: All the hourly 2-m air temperature, which is “T2M” variable in the original dataset, for each day are averaged and stored in “daily_T2M” variable in the derived daily 2-m air temperature data.

(3) Daily Precipitation

Frequency: daily mean value

Spatial Grid: 2D, single-level, full horizontal resolution

Granule size: ~25.9 MB

Dimensions: longitude = 3600, latitude = 1800

Acquisition Method Description: All the half-hourly calibrated precipitation, which is “precipitationCal” variable in the original dataset, for each day are averaged and stored in “daily_precipitation” variable in the derived daily precipitation data.

(4) Monthly Nighttime Light Radiance

Frequency: monthly mean value

Spatial Grid: 1D

Granule size: varies according to spatial coverage, China: 1.5G

Dimensions: number of pixels

Acquisition Method Description: All the cloudless daily nighttime light radiance over the target region, which is “DNB_At_Sensor_Radiance_500m” variable in the original dataset, for each day are averaged and stored in “monthly_mean_radiance” variable in the derived daily precipitation data.

(5) Metadata

The metadata of daily/monthly global environmental factors are listed in Table 2.

Table 2. Daily/monthly global environmental factors.

Dataset	Attribute Name	Dimension	Description	Units
Daily Reanalyzed Specific Humidity	daily_QV2M	xy	2-m specific humidity	kg kg ⁻¹
	nlat	y	latitude	degree north
	nlon	x	longitude	degree east
Daily Reanalyzed Near Surface Temperature	daily_T2M	xy	2-m temperature	K
	nlat	y	latitude	degree north
	nlon	x	longitude	degree east
Daily Precipitation	daily_precipitation	xy	daily precipitation	mm/hour
	nlat	y	latitude	degree north
	nlon	x	longitude	degree east
Monthly Nighttime Light Radiance	monthly_mean_radiance	number of pixels	monthly mean radiance	nW/(cm ² sr)
	nlat	number of pixels	latitude	degree north
	nlon	number of pixels	longitude	degree east

2.2.2. Environmental Factors of Multiple Administration Levels

(1) Environmental Factors of City-level

Name: City-level daily statistics for Temperature/Humidity/Precipitation/ NO_2 tropospheric vertical column density (TVCD)

Format: CSV File

Contains information: City code (GID_2), Max value, Mean value, Min value

Acquisition Method Description: The city-level daily statistics data for Temperature/Humidity/Precipitation/ NO_2 tropospheric vertical column density (TVCD) are obtained through the following steps: Firstly, the reprocessed data (e.g., temperature, humidity, precipitation from Section 2.2.1) are converted from NetCDF to GeoTIFF. Secondly, the vector boundary of each city is obtained through linking to the “GID_2” field in the administrative boundary map. Thirdly, all the pixel values within the vector boundary of each city are used as a statistical array. Fourthly, we calculate the maximum, minimum, and average values of Temperature/Humidity/Precipitation/ NO_2 TVCD from the obtained statistical array of each city, and export the results to a CSV file with the variable names as “Max”, “Min” and “Mean”.

Name: Province-/State-level daily statistics for Temperature/Humidity/Precipitation

Format: CSV File

Contains information: Province/State code (GID_1), Max value, Mean value, Min value

Acquisition Method Description: The province-/state-level daily statistics data for Temperature/Humidity/Precipitation are obtained through similar procedure as city-level, except for the vector boundary of each province/state is obtained by linking to the “GID_1” field in the administrative boundary map.

(3) Environmental Factors of Country-levels

Name: Country-level daily statistics for Temperature/Humidity/Precipitation

Format: CSV File

Contains information: Country code (GID_0), Max value, Mean value, Min value

Acquisition Method Description: The country-level daily statistics data for Temperature/Humidity/Precipitation are obtained through similar procedure as city-level, except for the vector boundary of each country is obtained by linking to the “GID_0” field in the administrative boundary map.

(4) Metadata

The metadata of multiple administration levels’ environmental factors are listed in Table 3.

Table 3. Metadata of environmental factors of multiple administration levels.

	Attribute Name	Description	Format	Units
Environmental Factors of City-level	GID_2	Used to uniquely identify the city	String	None
	Max	Maximum Temperature/Humidity/Precipitation/ NO_2 TVCD	Float	$\text{K}/(\text{kg kg}^{-1})/(\text{mm}/\text{hour})/(\text{molec}/\text{cm}^2)$
	Mean	Average Temperature/Humidity/Precipitation/ NO_2 TVCD	Float	$\text{K}/(\text{kg kg}^{-1})/(\text{mm}/\text{hour})/(\text{molec}/\text{cm}^2)$
	Min	Minimum Temperature/Humidity/Precipitation/ NO_2 TVCD	Float	$\text{K}/(\text{kg kg}^{-1})/(\text{mm}/\text{hour})/(\text{molec}/\text{cm}^2)$

Table 3. Cont.

	Attribute Name	Description	Format	Units
Environmental Factors of Province-level/State-level	GID_1	Used to uniquely identify the province/state	String	None
	Max	Maximum Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)
	Mean	Average Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)
	Min	Minimum Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)
Environmental Factors of Country-levels	GID_0	Used to uniquely identify the country	String	None
	Max	Maximum Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)
	Mean	Average Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)
	Min	Minimum Temperature/Humidity/Precipitation/NO ₂ TVCD	Float	K/(kg kg ⁻¹)/(mm/hour)/(molec/cm ²)

3. Methods

3.1. Spatiotemporal Aggregation and Collocation

Focusing on the reprocessed environmental data (e.g., temperature, humidity, precipitation), it is necessary to establish the relationship between the data in time and space. This study proposes to statistically analyze the environmental characteristics on daily and monthly scales and different administrative levels based on vector boundaries.

As shown in Figure 1, global maps of daily average factors are generated by aggregating the hourly and half-hourly data in temporal dimension for each spatial location with the means output to NetCDF format files.

The collocation with different administration levels are realized based on the Python programming language. The open-source libraries of GDAL and netCDF4 are used to convert the reprocessed data (e.g., temperature, humidity, precipitation) from NetCDF to GeoTIFF. By using open-source libraries such as “geopandas”, “shapely” and “rasterio”, the vector boundaries of different administrative levels (country, province/state and county/city) are used to obtain the GeoTIFF pixels covered by the mask as a statistical array. For the obtained pixel array, this is accomplished by setting the calculation conditions, using the NumPy scientific calculation library to extract the statistical characteristics (maximum, minimum, and average), and finally exporting array to a CSV file for storage. The specific procedure is shown in Figure 1.

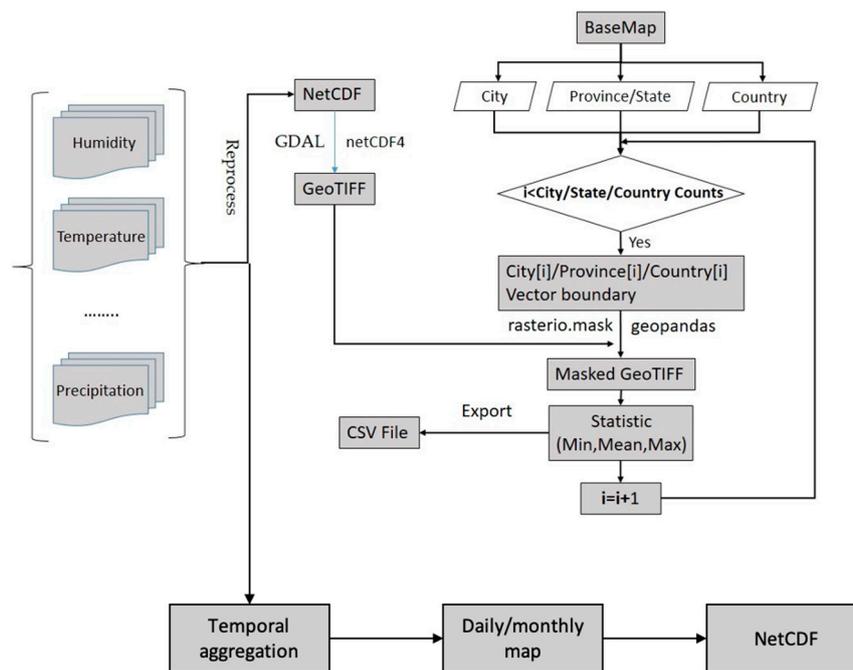


Figure 1. Workflow of spatiotemporal aggregation and collocation.

3.2. Collocating Environmental Factors with COVID-19 Case Data

The proposed environmental data collection is integrated and published together with GMU STC Center data cube to associate with COVID-19 cases data. The data cube structure is established and utilized to represent factors and values from a spatiotemporal perspective. Due to the multiple scales of target regions, the dataset is divided by country and region at the first level, and the administration scales are archived and shared under distinct regional folders. Daily report and time-series summary reports are processed and published in each country and administrative level. For example, the United States folder includes administrative 1 for the state level dataset and administrative 2 for the county level dataset. Under USA administrative 1 folder, a group of csv files keep a one-day timestamp of all extracted and processed environmental values for each state, defined as the daily report dataset. The summary report only keeps the latest updated files divided by factors to record the time-series value of each state.

3.3. Data Computing and Storage on AWS Cloud Platform

Cloud computing is becoming the standard approach to handling large scale and remotely sensed (RS) imagery dataset processing, storage, access, and management. There are many cloud platform providers that provide users a “pay as you go” service to support customized computing needs. For example, Amazon web services (AWS), Microsoft Azure, and Google’s Compute Engine provide IaaS (Infrastructure as a Service), PaaS (Platform as a Service), or SaaS (Software as a Service). In this study, AWS was adopted as the cloud to support elastic storage and processing tasks for processing Nighttime Light Radiance, Temperatures, Humidity, Pollutants, and Precipitations dataset. With automatic data scraping from multiple RS data portals, those data were stored in a virtual storage optimized instance and were published to AWS S3 distributed storage. By exploiting computing capacity with over one-hundred computing cores and two-hundred gigabytes of memory, a multi-tasked python-based processing was deployed to mine those datasets and produced covid-19 related results from the perspective of RS observation. Ongoing distributed computing approach will be developed to accommodate global scale multi-sourced RS data processing in a single run with a reasonable processing time.

4. Data Sharing

There are two main types of datasets. One is remote sensing imageries, and the other is CSV tables. The remote sensing data are often too large to be shared directly through GitHub. Therefore, they are first uploaded to AWS S3 Bucket which is a public cloud storage service provided by Amazon, and a corresponding downloadable link is created for each image. Then, the links are integrated with tables on GitHub for users to access. The Simple Storage Service is known as S3, which efficiently stores files, folders, and objects. The S3 consists of data and its descriptive metadata. All data can be accessed with permission. On the other hand, the size of CSV data is relatively small and is directly uploaded to the GitHub for downloading. Detailed resolution information and downloading path for each dataset can be found in Table 4.

Table 4. Data sharing for the COVID-19 related environmental data.

	Derived Values	Spatial Resolution	Temporal Resolution	Downloading Path
Mean value statistics corresponding with regions	Nighttime light radiance	Country/state (province)	daily	https://github.com/stccenter/COVID-19/tree/master/analysis/nightlight
	Air Quality	Country/state (province)/county (city)	daily	https://github.com/stccenter/COVID-19/tree/master/analysis/nightlight
	Precipitation	Global/country/ state (province)/ county (city)	daily	https://github.com/stccenter/COVID-19-Data/tree/master/Environmental%20factors
	Temperature	Global/country/ state (province)/ county (city)	daily	https://github.com/stccenter/COVID-19-Data/tree/master/Environmental%20factors
	Humidity	Global/country/ state (province)/ county (city)	daily	https://github.com/stccenter/COVID-19-Data/tree/master/Environmental%20factors
	NO ₂ TVCD	Global/country/ state (province)/ county (city)	daily	https://github.com/stccenter/COVID-19-Data/tree/master/Environmental%20factors
Global Distribution maps	Nighttime light radiance	Country/state (province)	monthly	https://covid19datagmu.s3.us-east-2.amazonaws.com/NightTimeLight/NightTimeLight.zip
	Precipitation	Global	daily	https://covid19datagmu.s3.us-east-2.amazonaws.com/precipitation/daily/daily_precipitation_JAN_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/precipitation/daily/daily_precipitation_FEB_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/precipitation/daily/daily_precipitation_MAR_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/precipitation/daily/daily_precipitation_APR_2020.zip
	Temperature	Global	daily	https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/Temperature/daily_MEAN_TEMP_JAN_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/Temperature/daily_MEAN_TEMP_FEB_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/Temperature/daily_MEAN_TEMP_MAR_2020.zip

Table 4. Cont.

	Derived Values	Spatial Resolution	Temporal Resolution	Downloading Path
Global Distribution maps	Humidity	Global	daily	https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/+humidity/daily_MEAN_JAN_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/+humidity/daily_MEAN_FEB_2020.zip https://covid19datagmu.s3.us-east-2.amazonaws.com/temperature_humidity/daily/+humidity/daily_MEAN_MAR_2020.zip

5. Quality Control

To provide reliable environmental data sources to the geospatial and covid-19 community, populated data are evaluated in three dimensions including data integrity, consistency, and validity to ensure high quality data publishing.

Raw data selection, cleaning and qualification: The first and crucial step to create a high-quality COVID-19 related environmental data collection is to select proper input raw data. To guarantee this, we firstly review as many literatures on COVID-19 related research as we can to decide on what environmental factors should be included in the final collection and what data sources are researchers usually dealing with. Then we sift among the potential data sources and choose the one that is most frequent-adopted, stable and authoritative for each environmental factor. In the data processing step, we filter all the invalid and unreasonable values as well as variables that are not related to COVID-19.

Data integrity: This means that populated data should be comprehensive. A thorough check is applied to time-series data, making sure the data contain all historical data stored in data sources. In addition, since daily grid environmental data are mapped to an administrative level shapefile to provide regional environmental data, integrity check ensures the generated data are provided at each unit (e.g., counties in US) at a certain administrative level if the data is available in the source files.

Data consistency: This requires that data in our repository are consistent with other sources. On one hand, extracted data should be consistent with values from data sources; on the other hand, regional derived values (e.g., country-level monthly mean temperature) should be consistent with global and temporal distributions. For example, mean temperature in a location in winter is lower than mean temperature there in summer. Precipitation value are relatively larger and frequent in the Inter Tropical Convergence Zone (ITCZ) and South Pacific Convergence Zone (SPCZ).

Data validity: This dimension estimates the data reliability. Data sources should be provided with the populated data, thereby making data sources available to data consumers to ensure data consumers can investigate the data sources for validity.

6. Conclusions

Our proposed data collection encompasses COVID-19 related environmental datasets that serves as a data basis and reference for users in broader communities (e.g., governmental and urban planning departments, meteorological and climatological scientists, medical and disease control researchers). This is an alternative to other data collection efforts that are virus-case-only platforms. The proposed collection is associated with the COVID-19 gateway of GMU's NSF Spatiotemporal Center and is stored on a stable and highly available AWS server to provide multiple-scale spatiotemporal data at high acquisition speed [33]. The collection includes various data types and features including temperature, humidity, air quality, nighttime light and precipitation.

The raw datasets are automatically downloaded from the data sources using Python programs, and the derived values are produced as soon as the newest raw data are released. The timeliness is guaranteed by this procedure.

The proposed framework is a growing data collection with content extended according to the needs and requirement of users and the evolution of the pandemic. For example, the team is working to automatically correspond OMI NO₂ data with administration shapefiles and to provide country and county level NO₂ information to the communities. It is proposed that these NO₂ data contribute to the Earth data aspects for big spatiotemporal data analytics in fighting against covid-19 pandemic [33,34].

Author Contributions: C.Y. and Q.L. came up with the original research idea; C.Y. advised Q.L., D.S. and W.L. on the data processing program developing and paper structure; S.K., E.C., V.A. and S.C. processed the data as daily routine; all authors wrote one part of the paper; Q.L. coordinated the data processing, sharing, and paper drafting; H.L. drafted the Section 3.3; Y.L. drafted Section 5; Z.W. drafted Section 4; Y.Z. drafted Section 2.1.5; and Z.Z. drafted Section 2.1.4; C.Y. and J.T.H. revised the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NSF (1835507, 1841520 and 2027521).

Acknowledgments: MERRA-2 data is downloaded from: https://disc.gsfc.nasa.gov/datasets/M2T1NXSLV_5.12.4/summary?keywords=MERRA2_400.tavg1_2d_slv_Nx; IMERG data is downloaded from: https://disc.gsfc.nasa.gov/datasets/GPM_3IMERGHHE_06/summary?keywords=IMERG; NPP/VIIRS is downloaded from: <https://ladsweb.modaps.eosdis.nasa.gov/search/order/3/VNP46A1--5000/2020-02-01..2020-04-03/DB/>; OMI data is downloaded from: https://disc.gsfc.nasa.gov/datasets/OMNO2d_003/summary?keywords=omi; ground-based air pollution is downloaded from: <http://www.cnemc.cn/sss/> and (<https://www.epa.gov/outdoor-air-quality-data>

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kagan, A.R.; Levi, L. Health and environment—Psychosocial stimuli: A review. *Soc. Sci. Med.* **1974**, *8*, 225–241. [[CrossRef](#)]
2. Chan, K.H.; Peiris, J.S.M.; Lam, S.Y.; Poon, L.L.; Yuen, K.-Y.; Seto, W.H. The Effects of Temperature and Relative Humidity on the Viability of the SARS Coronavirus. *Adv. Virol.* **2011**, *2011*, 1–7. [[CrossRef](#)] [[PubMed](#)]
3. Chen, B.; Liang, H.; Yuan, X.; Hu, Y.; Xu, M.; Zhao, Y.; Zhang, B.; Tian, F.; Zhu, X. Roles of meteorological conditions in COVID-19 transmission on a worldwide scale. *MedRxiv* **2020**. [[CrossRef](#)]
4. Bashir, M.F.; Ma, B.; Komal, B.; Bashir, M.A.; Tan, D.; Bashir, M. Correlation between climate indicators and COVID-19 pandemic in New York, USA. *Sci. Total Environ.* **2020**, *728*, 138835. [[CrossRef](#)] [[PubMed](#)]
5. Liu, Q.; Sha, D.; Liu, W.; Houser, P.; Zhang, L.; Hou, R.; Lan, H.; Flynn, C.; Lu, M.; Hu, T.; et al. Spatiotemporal Patterns of COVID-19 Impact on Human Activities and Environment in Mainland China Using Nighttime Light and Air Quality Data. *Remote Sens.* **2020**, *12*, 1576. [[CrossRef](#)]
6. Prata, D.N.; Rodrigues, W.; Bermejo, P.H. Temperature significantly changes COVID-19 transmission in (sub) tropical cities of Brazil. *Sci. Total Environ.* **2020**, *729*, 138862. [[CrossRef](#)]
7. Shereen, M.A.; Khan, S.; Kazmi, A.; Bashir, N.; Siddique, R. COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *J. Adv. Res.* **2020**, *24*, 91–98. [[CrossRef](#)]
8. Bashir, M.F.; Bilal, B.J.M.A.; Komal, B. Correlation between environmental pollution indicators and COVID-19 pandemic: A brief study in Californian context. *Environ. Res.* **2020**, *187*, 109652. [[CrossRef](#)]
9. Murgante, B.; Borruo, G.; Balleto, G.; Castiglia, P.; Dettori, M. Why Italy First? Health, Geographical and Planning Aspects of the COVID-19 Outbreak. *Sustainability* **2020**, *12*, 5064. [[CrossRef](#)]
10. You, H.; Wu, X.; Guo, X. Distribution of COVID-19 Morbidity Rate in Association with Social and Economic Factors in Wuhan, China: Implications for Urban Development. *Int. J. Environ. Res. Public Health* **2020**, *17*, 3417. [[CrossRef](#)]
11. Pirouz, B.; Haghshenas, S.S.; Piro, P.; Haghshenas, S.S. Investigating a Serious Challenge in the Sustainable Development Process: Analysis of Confirmed cases of COVID-19 (New Type of Coronavirus) Through a Binary Classification Using Artificial Intelligence and Regression Analysis. *Sustainability* **2020**, *12*, 2427. [[CrossRef](#)]

12. Peng, Z.; Wang, R.; Liu, L.; Wu, H. Exploring Urban Spatial Features of COVID-19 Transmission in Wuhan Based on Social Media Data. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 402. [CrossRef]
13. What Is a Data Platform? Available online: <https://looker.com/definitions/data-platform#:~:text=A%20data%20platform%20is%20an,technologies%20for%20strategic%20business%20purposes> (accessed on 14 June 2020).
14. World Health Organization. Coronavirus Disease (COVID-2019) Situation Reports. Available online: <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports> (accessed on 14 June 2020).
15. Center for Disease Control and Prevention. Get the Facts About Coronavirus. Available online: <https://www.cdc.gov/coronavirus/2019-ncov/index.html> (accessed on 14 June 2020).
16. Virginia Department of Health. COVID-19 in Virginia. Available online: <https://www.vdh.virginia.gov/coronavirus/> (accessed on 14 June 2020).
17. Maryland Department of Health. COVID-19 Outbreak. Available online: <https://phpa.health.maryland.gov/Pages/Novel-coronavirus.aspx> (accessed on 14 June 2020).
18. Government of the District of Columbia. Coronavirus Data. Available online: <https://coronavirus.dc.gov/page/coronavirus-data> (accessed on 14 June 2020).
19. The New York Times. The Coronavirus Outbreak. Available online: <https://www.nytimes.com/news-event/coronavirus> (accessed on 14 June 2020).
20. Dong, E.; Du, H.; Gardner, L.M. An interactive web-based dashboard to track COVID-19 in real time. *Lancet Infect. Dis.* **2020**, *20*, 533–534. [CrossRef]
21. 1 Point 3 Acres. Global COVID-19 Tracker & Interactive Chart. Available online: <https://coronavirus.1point3acres.com/en/> (accessed on 14 June 2020).
22. Yang, C.; Sha, D.; Liu, Q.; Li, Y.; Lan, H.; Guan, W.W.; Hu, T.; Li, Z.; Zhang, Z.; Thompson, J.H.; et al. Taking the pulse of COVID-19: A spatiotemporal perspective. *arXiv* **2020**, arXiv:2005.04224v1.
23. Wang, J.; Tang, K.; Feng, K.; Lv, W. High Temperature and High Humidity Reduce the Transmission of COVID-19. *SSRN Electron. J.* **2020**. [CrossRef]
24. Sajadi, M.M.; Habibzadeh, P.; Vintzileos, A.; Shokouhi, S.; Miralles-Wilhelm, F.; Amoroso, A. Temperature and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19. *SSRN Electron. J.* **2020**. [CrossRef]
25. Gelaro, R.; Mccarty, W.; Suárez, M.J.; Todling, R.; Molod, A.; Takacs, L.; Randles, C.; Darmenov, A.; Bosilovich, M.G.; Reichle, R.H.; et al. The Modern-Era Retrospective Analysis for Research and Applications, Version 2 (MERRA-2). *J. Clim.* **2017**, *30*, 5419–5454. [CrossRef]
26. Ruiz, M.O.; Chaves, L.F.; Hamer, G.L.; Sun, T.; Brown, W.M.; Walker, E.D.; Haramis, L.; Goldberg, T.L.; Kitron, U. Local impact of temperature and precipitation on West Nile virus infection in Culex species mosquitoes in northeast Illinois, USA. *Parasites Vectors* **2010**, *3*, 19. [CrossRef]
27. Huffman, G.J.; Bolvin, D.T.; Nelkin, E.J. Integrated Multi-satellitE Retrievals for GPM (IMERG) technical documentation. *NASA/GSFC Code* **2015**, 612, 2019.
28. Liu, Q.; Li, Y.; Yu, M.; Chiu, L.S.; Hao, X.; Duffy, D.Q.; Yang, C. Daytime Rainy Cloud Detection and Convective Precipitation Delineation Based on a Deep Neural Network Method Using GOES-16 ABI Images. *Remote Sens.* **2019**, *11*, 2555. [CrossRef]
29. Román, M.O.; Wang, Z.; Sun, Q.; Kalb, V.; Miller, S.D.; Molthan, A.; Schultz, L.; Bell, J.; Stokes, E.C.; Pandey, B.; et al. Nasa’s black marble nighttime lights product suite. *Remote Sens. Environ.* **2018**, *210*, 113–143. [CrossRef]
30. Russell, A.R.; Valin, L.C.; Cohen, R.C. Trends in OMI NO2 observations over the United States: Effects of emission control technology and the economic recession. *Atmos. Chem. Phys. Discuss.* **2012**, *12*, 12197–12209. [CrossRef]
31. Levelt, P.; Oord, G.V.D.; Dobber, M.; Malkki, A.; Visser, H.; De Vries, J.; Stammes, P.; Lundell, J.; Saari, H. The ozone monitoring instrument. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 1093–1101. [CrossRef]
32. Isaifan, R. The dramatic impact of Coronavirus outbreak on air quality: Has it saved as much as it has killed so far? *Glob. J. Environ. Sci. Manag.* **2020**, *6*, 275–288.

33. Yang, C.; Clarke, K.; Shekhar, S.; Tao, C.V. Big Spatiotemporal Data Analytics: A research and innovation frontier. *Int. J. Geogr. Inf. Sci.* **2019**, *34*, 1075–1088. [[CrossRef](#)]
34. Yang, C.; Yu, M.; Li, Y.; Hu, F.; Jiang, Y.; Liu, Q.; Sha, D.; Xu, M.; Gu, J. Big Earth data analytics: A survey. *Big Earth Data* **2019**, *3*, 83–107. [[CrossRef](#)]



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).