Data Descriptor: Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia

Amir S. Siraj, Isabel Rodriguez-Barraquer, Christopher M. Barker, Natalia Tejedor-Garavito, Dennis Harding, Christopher Lorton, Dejan Lukacevic, Gene Oates, Guido Espana, Moritz U.G. Kraemer, Carrie Manore, Michael A. Johansson, Andrew J. Tatem, Robert C. Reiner & T. Alex Perkins

Despite a long history of mosquito-borne virus epidemics in the Americas, the impact of the Zika virus (ZIKV) epidemic of 2015–2016 was unexpected. The need for scientifically informed decision-making is driving research to understand the emergence and spread of ZIKV. To support that research, we assembled a data set of key covariates for modeling ZIKV transmission dynamics in Colombia, where ZIKV transmission was widespread and the government made incidence data publically available. On a weekly basis between January 1, 2014 and October 1, 2016 at three administrative levels, we collated spatiotemporal Zika incidence data, nine environmental variables, and demographic data into a single downloadable database. These new datasets and those we identified, processed, and assembled at comparable spatial and temporal resolutions will save future researchers considerable time and effort in performing these data processing steps, enabling them to focus instead on extracting epidemiological insights from this important data set. Similar approaches could prove useful for filling data gaps to enable epidemiological analyses of future disease emergence events.

| Design Type(s)          | data integration objective ● database creation objective |
|-------------------------|--------------------------------------------------------|
| Measurement Type(s)     | transmission of virus                                  |
| Technology Type(s)      | database extract, transform, and load process          |
| Factor Type(s)          | DataTypes                                               |
| Sample Characteristic(s)| Zika virus ● Colombia ● Homo sapiens ● Aedes aegypti ● anthropogenic environment ● elevation ● humid air ● hydrological precipitation process ● temperature of air ● vegetation layer |

1Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, 345 Galvin Hall, Notre Dame, IN 46556, USA. 2Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA. 3Department of Pathology, Microbiology and Immunology, University of California, 5329 Vet Med 3A, Davis, CA 95616, USA. 4WorldPop, Department of Geography and Environment, University of Southampton, University Road, Southampton, SO17 1BJ, UK. 5Flowminder Foundation, Roslagsgatan 17, SE-11355, Stockholm, Sweden. 6Institute for Disease Modeling, Bellevue, 3150 139th Ave SE, WA 98005, USA. 7Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK. 8Harvard Medical School, 10 Shattuck Street, Boston, MA 02115, USA. 9Boston Children’s Hospital, 300 Longwood Ave, Boston, MA 02115, USA. 10Theoretical Biology and Biophysics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. 11Centers for Disease Control and Prevention, 1324 Calle Canada, San Juan, PR 00920-3860, USA. 12Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, 677 Huntington Ave., Suite 506, Boston, MA 02115, USA. 13Department of Global Health and Institute for Health Metrics and Evaluation, University of Washington, 2301 Fifth Ave., Suite 600, Seattle, WA 98121, USA. Correspondence and requests for materials should be addressed to A.S.S. (email: asiraj@nd.edu) or to T.A.P (email: taperkins@nd.edu).
Background & Summary

Zika virus (ZIKV) emerged as a pathogen of global concern in 2015 when it rapidly spread through the Americas and was associated with Guillain-Barré syndrome (GBS) in adults and congenital Zika syndrome (CZS) in fetuses and neonates. Though ZIKV had been discovered several decades earlier, recognition of severe outcomes and the explosive nature of ZIKV epidemics was only established recently. Moreover, an estimated 80% rate of asymptomatic infection and the presence of more infections with relatively mild symptoms who go unreported complicate efforts to estimate disease incidence and further make modeling the spread of ZIKV a challenging task. Despite these issues and the chronic lack of data at the appropriate spatio-temporal scales, efforts to understand the spatiotemporal dynamics of ZIKV rely heavily on access to data about its spatiotemporal drivers.

ZIKV is transmitted primarily by *Aedes aegypti* mosquitoes, which also transmit chikungunya, yellow fever, and dengue viruses. Like these other viruses, ZIKV transmission is highly dependent on the environment. Climatic conditions, for example, regulate the population dynamics of vectors, and the built environment plays an important role in human-vector interaction and in providing breeding grounds for mosquitoes. Even though the importance of these factors is widely recognized, their specific roles are more difficult to understand but can be aided by model-based analysis combining epidemiological and environmental data.

The availability of spatiotemporal incidence data is critical to both current and near-future responses and to planning for responses to emerging infectious disease outbreaks. For example, during the Ebola epidemic in 2014-2015, mathematical and statistical models using incidence data were critical to informing resource allocation and placement of new hospital beds, plans for vaccine trials, estimates of intervention effectiveness, and understanding how the outbreak started and where it spread in time and space. Similarly, spatiotemporal ZIKV data has informed efforts to estimate the number of people at risk for infection and the number of pregnant women infected. Such data are also potentially important for selecting sites for ZIKV vaccine trials.

Despite the widely recognized importance of spatiotemporal incidence data, there is often limited availability of such data sets for emerging infectious diseases. In the case of Zika, there has been some effort to broaden access to these data (e.g., the cdepi Github repository), but the data available through these settings are often not internally consistent and are not made available with important covariates, such as population and weather conditions. Colombia is one country for which data has been made available online by its Instituto Nacional de Salud and is of particular interest due to the high resolution of data there (available weekly for each of 1,122 municipalities). This data set is also of particular interest for modeling the spatio-temporal spread of ZIKV due to Colombia’s diverse landscape and because of substantial heterogeneity in the timing and intensity of ZIKV transmission there. Together, these factors offer a unique opportunity to examine the role of environmental and social influences on the spread of ZIKV.

In addition to spatiotemporal incidence data, several variables are commonly incorporated into analyses of the transmission dynamics of ZIKV and related pathogens. First, temperature plays a dominant role in ZIKV transmission due to its influence on vector and virus life traits. Because the effect of temperature on transmission depends not only on mean temperature but also on daily temperature range, we include estimates of mean, minimum, and maximum daily temperature. Second, a number of metrics related to moisture—including precipitation, humidity, and normalized difference vegetation index (NDVI)—are commonly used for modeling mosquito population dynamics due to their relevance to the immature stages of the mosquito life cycle. Third, we include spatiotemporal estimates of relative mosquito abundance, a spatial estimate of purchasing power as a proxy for the effect of socioeconomic effects on mosquito-human contact, and spatial estimates of travel time to allow for exploration of the effects of connectivity on spatiotemporal transmission dynamics. Fourth, we include demographic projections of total population and annual births to allow for quantification of the population at risk of ZIKV infection and severe outcomes such as GBS and CZS.

Here, we collated data on the aforementioned variables at three administrative scales on a weekly basis between January 1, 2014 and October 1, 2016, which spans the majority of ZIKV transmission activity in Colombia. Our hope is that this effort will increase access to this data set and reduce duplication of the considerable effort required to process data for epidemiological analyses of ZIKV transmission dynamics.

Methods

To achieve our central objective of assembling and collating multiple data sets pertaining to ZIKV transmission in Colombia, we first identified key data and then translated those data to comparable spatiotemporal resolution using a variety of methods. In some cases, this was as simple as downloading raster datasets and clipping them to shape files. In other cases, this involved statistical modelling to transform existing data products from certain scales into a single data product at some other desired scale. In all cases, our methods involved taking input data (Table 1) and generating output data (Table 2) at a weekly timescale between January 1, 2014 and October 1, 2016 for each of three administrative scales (Fig. 1). Throughout, we generated output data at the national scale, for each of 33 departments, and for each of 1,122 municipalities, as defined by GIS shapefiles from the National Geographical Information System of Colombia.
Zika case reports
The weekly number of Zika cases, by municipality, was reconstructed using two data sources. The main data source was a website of the Colombian National Institute of Health (Instituto Nacional de Salud) where the official weekly reports on the cumulative number of Zika suspected and confirmed cases for each municipality have been published since the beginning of 2016.

While the peak of the Colombian epidemic occurred in 2016, a significant number of cases were reported during 2015. In order to capture this initial portion of the epidemic, we used an additional data source, also available in the INS website. Unfortunately, the number of cases reported in the latter data source seemed to consistently underreport the total number of cases reported by the INS at the national scale. For example, while the official data source reports a cumulative number of 11,712 cases by the end of 2015, this secondary source only reports 3,875 cases for this same period. Therefore, in order to reconstruct the 2015 portion of the epidemic while accounting for the better known total number of cases, we multiplied the weekly 2015 data by a correction factor. This correction factor was calculated as the ratio between the cumulative number of cases reported by each municipality up to the first week of 2016 according to the official source and the alternative source. The raw and the corrected weekly counts for each municipality are included in the data set. To account for cases from unknown municipalities within a department, we also provide data at the departmental level.

Human demographics
We obtained gridded population data across Colombia for the year 2015 at a resolution of 3 arc seconds (~93 m) from the WorldPop website (http://worldpop.org.uk). Similarly, we obtained high-resolution (30 arc seconds) unpublished gridded data on the number of births for the year 2015 from the WorldPop project. These high-resolution products were developed to ensure consistencies with subnational data on sex and age structures, as well as subnational age-specific fertility rates, while adjustments on births were made at subnational scales using data from the government of Colombia, followed by national-level adjustments to contemporary numbers based on 2012 and 2015 United Nations Population Division data.

Spatial aggregation of covariates
Aggregation of raster data at the level of administrative units requires some assumption about how raster values should be weighted to obtain a single value for an administrative unit. Due to the fact that Zika virus transmission occurs predominantly in human-dominated areas, we used human population (WorldPop Project) as our weighting variable. We applied this weighting procedure to aggregate all covariates at municipal (e.g., as in Fig. 2), departmental (e.g., as in Fig. 3), and national levels.

Aedes aegypti abundance
We obtained one hundred posterior samples of Aedes aegypti occurrence probabilities in raster format, from the published work of Kraemer et al., which we used to derive weekly mosquito abundance measures for all 52 weeks of the year. We based our method on the assumption that mosquitoes are at time $t$, $m(t)$, can be represented by a Poisson distribution with rate parameter $\lambda = -\ln(occurrence\ probability)$, consistent with existing ZIKV transmission models. We obtained such an estimate of the relative density of mosquitoes across a 4.65 km x 4.65 km grid for each of 52 weeks. In addition, we generated aggregated values at the municipality, departmental and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).

Temperature
We downloaded meteorological readings from 30 stations across continental Colombia from National Oceanic and Atmospheric Administration (NOAA)’s Climate Data Online, an online archive of daily meteorological readings. The variables we extracted from this data set included minimum daily temperature, maximum daily temperature, mean daily temperature, and relative humidity, all on a daily basis between January 1, 2014 and October 1, 2016.

To facilitate interpolation of these climate variables across a more complete spatial coverage of the country, we downloaded a digital elevation dataset at a resolution of 30 arc seconds from the Global 30 Arc-Second Elevation (GTOPO30) product. Similarly, we downloaded the WorldClim gridded long-term average of monthly minimum temperature, maximum temperature, and precipitation at a 4.65 km x 4.65 km spatial resolution, as well as NOAA’s Climate Prediction Center (CPC) global monthly mean air temperature at 0.5 arc-degrees resolution.

To generate smooth, high-resolution surfaces of climate variables based on calibration to point readings from the 30 meteorological stations, we tested two approaches of spatial interpolation: (a) using non-parametric surface fitting with thin plate splines (TPS) with or without fixed-factor covariates; (b) using spatial models (kriging) with or without covariates. We selected the best interpolation models for each environmental variable based on leave-one-out cross validation, as described in the Technical Validation section.

The thin plate spline (TPS) follows the general form,

$$Y(x) = \mu(x) + P(x) + \varepsilon$$

(1)
Table 1. Input datasets, used to generate gridded and administrative aggregate outputs.

| Name                                                       | Acquisition year | Source                                    | Version, Publication year, License | Data Type                     | Spatial Resolution | Format/ Pixel Type & Depth | Spatial Reference | Spatial Coverage |
|-------------------------------------------------------------|-------------------|-------------------------------------------|-----------------------------------|-------------------------------|--------------------|--------------------------|------------------|-----------------|
| GTOP030 Gridded Elevation                                   | <1996             | USGS37                                   | 1996, CCO 1.0                     | Elevation, continuous raster  | ~30° (~930 m)      | Geo-tiff/flt32          | GCS WGS 1984     | Regional        |
| CPC Surface Air Temperature                                 | 2014-2016         | Fan & van den Dool38                     | 2008, CCO 1.0                     | Monthly surface air temperature, continuous raster | 1800° (~56 km)     | ESRI grid/flt32         | GCS WGS 1984     | Global          |
| Worldclim Average Temperature                              | 1960–1990         | Hijmans R.J., et al.39                   | v1, 2005, CCBY 4.0                | Average monthly temperature, continuous raster | 15° (~4.65 km)     | Geo-tiff/flt32          | GCS WGS 1984     | Global          |
| Worldclim Minimum Temperature                              | 1960–1990         | Hijmans R.J., et al.39                   | v1, 2005, CCBY 4.0                | Average monthly minimum temperature, continuous raster | 15° (~4.65 km)     | Geo-tiff/flt32          | GCS WGS 1984     | Global          |
| Worldclim Maximum Temperature                              | 1960–1990         | Hijmans R.J., et al.39                   | v1, 2005, CCBY 4.0                | Average monthly maximum temperature, continuous raster | 15° (~4.65 km)     | Geo-tiff/flt32          | GCS WGS 1984     | Global          |
| Daily Station Mean Temperature                              | 2014–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily mean temperature reading from 30 stations, continuous vector | Comparable to 1° (~30 m) | HTML/flt32            | GCS WGS 1984     | Colombia        |
| Daily Station Minimum temperature                           | 2014–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily minimum temperature reading from 30 stations, continuous vector | Comparable to 1° (~30 m) | HTML/flt32            | GCS WGS 1984     | Colombia        |
| Daily Station Maximum temperature                           | 2014–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily maximum temperature reading from 30 stations, continuous vector | Comparable to 1° (~30 m) | HTML/flt32            | GCS WGS 1984     | Colombia        |
| Daily Station Relative Humidity                             | 2014–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily relative humidity reading from 30 stations, continuous vector | Comparable to 1° (~30 m) | HTML/flt32            | GCS WGS 1984     | Colombia        |
| Daily Mean Dew Point Temperature                            | 2014–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily mean dew point temperature reading from 30 stations, continuous vector | Comparable to 1° (~30 m) | HTML/flt32            | GCS WGS 1984     | Colombia        |
| Gridded Population of the World (GPW)                      | 2005              | CIESIN38                                 | v3, 2004, CCBY 4.0                | Global Population Estimates, continuous raster | 15° (~4.65 km)     | Geo-tiff/flt32          | GCS WGS 1984     | Global          |
| Confirmed and Suspected Cumulative ZIKV Cases               | 2015–2016         | INS40                                    | 2016                            | Weekly suspected and confirmed cumulative ZIKV cases by municipality from two INS sources | NA                  | CSV/flt32              | NA              | Colombia        |
| Occurrence Probability of Aedes aegypti                     | 1960–2014         | Kraemer et al.201535                    | 2015, Author                     | Global occurrence probabilities of Aedes aegypti, continuous raster | 15° (~4.65 km)     | Geo-tiff/flt32          | GCS WGS 1984     | Global          |
| GEcon – Gross Cell Product                                  | 2005              | Nerdhaus49                               | 2006, CCBY 4.0                    | Global gridded gross cell product, continuous raster | 3600° (~111 km)    | XLS/flt32              | GCS WGS 1984     | Global          |
| WorldPop Population                                         | 2015              | WorldPop40                               | 2016, CCBY 4.0                    | Population count, continuous raster | 1° (~93 m)          | Geo-tiff/flt32          | GCS WGS 1984     | Colombia        |
| WorldPop Births                                            | 2015              | WorldPop40                               | 2016, Author                     | Count of births, continuous raster | 1° (~93 m)          | Geo-tiff/flt32          | GCS WGS 1984     | Colombia        |
| MODIS –MOD13A2 NDVI                                        | 2014–2016         | Didan K. (Data Citation 2)               | v6, 2015, CCO 1.0                  | 16-day NDVI from Terra MODIS, continuous raster | 30° (~930 m)        | HDF-EOS tiles/uint8     | Sinusoidal       | Global          |
| MODIS –MYD13A2 NDVI                                        | 2014–2016         | Didan K. (Data Citation 3)               | v6, 2015, CCO 1.0                  | 16-day NDVI from Aqua MODIS, continuous raster | 30° (~930 m)        | HDF-EOS tiles/uint8     | Sinusoidal       | Global          |
| NOAA’s Satellite Applications and Research Rainfall Estimates | 2015–2016         | NOAA48                                   | 2016, CCO 1.0                     | Daily precipitation estimates from satellites, continuous raster | 360° (~11 km)       | NetCDF /uint8           | GCS WGS 1984     | Global          |
| Travel Time to Major Cities                                 | 2000              | Nelson A.41                              | 2008, CCBY 3.0                    | Travel time, continuous raster | 30° (~930 m)        | Flt/flt32              | GCS WGS 1984     | Global          |
| MODIS 500m Global Urban Extent                             | 2002              | Schneider et al.2013.5, 2009, CCBY 3.0   | Urban extent                     | 15° (~465 m)          | Flt/flt32              | GCS WGS 1984     | Global          |
| Administrative Boundaries of Colombia                      | 2015              | SIGOT, Colombia48                        | 2015                            | Municipal administrative boundaries, vector | Comparable to 15° (~465 m) | ESRI polygon shapefile tiles | GCS WGS 1984     | Colombia        |

where $Y$ is the dependent variable evaluated at location $x$, $\mu$ is the fixed effect component of the model with optional covariates at location $x$, $P$ is the implicit spline polynomial function over the spatial coordinates, and $\epsilon$ is measurement error, assumed to be uncorrelated across sites and normally distributed with mean zero and standard deviation $\sigma$. The kriging approach follows the concept that spatial autocorrelation is dependent on distance between locations. We used the krig function in the geoR library of R with parameters chosen based on maximum-likelihood estimation. The model of a spatial process indexed by spatial locations $x_i$ follows

$$Y(x) = \mu(x) + S(x) + \epsilon$$

where $Y$ is the dependent variable evaluated at location $x$, $\mu$ is the fixed effect component of the model at location $x$, $S$ is a stationary Gaussian process with variance $\sigma^2$ (partial sill) and a correlation function parameterized by $\phi$ (range), and $\epsilon$ is the error term with its variance $\tau^2$ (nugget variance). When $\mu$ is included, the trend is implemented using lm, the regression model function in R, and $S(x)$ is fitted to the residuals of the regression model.
Due to Colombia’s proximity to the Equator, we ignored the small effect of distance distortion arising from non-projected spatial layers on both models. Because our goal is generating daily surfaces of climate variables, rather than developing a predictive model that works for days outside those to which we fitted the model, we treated every day separately and fitted a model for each day between January 1, 2014 and October 1, 2016 for which data was available. In addition to generating daily raster outputs and aggregating them at weekly time steps, we generated aggregated values at the municipality (Figs 2a–c), department (Fig. 3a) and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).

**Relative humidity**

Rather than interpolating relative humidity directly based on station readings (which showed poor estimates in preliminary results), we approached the task of estimating relative humidity indirectly. First, we spatially interpolated weather station measurements of mean dew point temperature from the 30 stations across Colombia. This was followed by calculating relative humidity across the 4.65 km x 4.65 km grid based on interpolated mean temperature and dew point temperature, using the August-Roche-Magnus approximation for the saturation vapour pressure of water in air, which follows

$$RH = \exp\left(\frac{-ab(T - T_d)}{(b + T_d)(b + T)}\right)$$

where $T$ and $T_d$ are the mean temperature and dew point temperature in °C and $a = 17.271$ and $b = 237.7$ °C. Finally, in addition to generating daily raster outputs and aggregating them at weekly time steps, we generated aggregated values at the municipality (Fig. 2d), department (Fig. 3b) and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).

**Normalized Difference Vegetation Index (NDVI)**

Satellite-based technologies have been used to capture spatial variation in environmental factors related to vector population dynamics, including a commonly used index called Normalized Difference Vegetation Index (NDVI) that captures the vegetation cover of regions. To account for spatial and temporal variation in vegetation cover that could influence habitat suitability for *Ae. aegypti*, the primary ZIKV vector, we downloaded NASA’s Moderate Resolution Image Spectro-radiometer (MODIS-Terra and Aqua version 13A2) vegetation indices at 16-day temporal and 1 km x 1 km spatial resolutions (Data...
These products have similar sensors but differ in their orbits as well as their daily hours and directions of crossing the equator. We linearly interpolated between data points (days on which data was reported) to generate a daily time series before aggregating the data back to a weekly resolution. In addition, we generated aggregated values at the municipality (Figs 2e and f), department (Fig. 3c) and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).

Precipitation

Among the climate datasets we explored, precipitation proved to be the most spatially variable, making it difficult to rely on spatial models to make accurate estimates. Our attempt of spatial interpolation of

---

**Figure 2.** Illustrative maps of municipality level weighted output variables for a single sample week. Variables include minimum temperature (a), mean temperature (b) maximum temperature (c), relative humidity (d), NDVI from Terra MODIS (e) NDVI from Aqua MODIS (f), total rainfall (g) average per capita gross cell product in 2005 US$ standard value (h) and average travel time to major cities in 2000 (i).

Citation 2, Data Citation 3). These products have similar sensors but differ in their orbits as well as their daily hours and directions of crossing the equator. We linearly interpolated between data points (days on which data was reported) to generate a daily time series before aggregating the data back to a weekly resolution. In addition, we generated aggregated values at the municipality (Figs 2e and f), department (Fig. 3c) and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).
precipitation using ordinary kriging resulted in large deviations from the observed values of the 30 stations obtained from NOAA. As an alternative, we used satellite-based data from NOAA’s Center for Satellite Applications and Research (STAR). We downloaded daily layers of the STAR rainfall estimates at ~ 4 km x 4 km resolution48. Once we download the daily products, we subset and resampled them into our standard resolution (4.65 km x 4.65 km) and spatial extent compatible with the other variables considered, before averaging across each consecutive seven days to generate weekly gridded data. In addition, we generated aggregated values at the municipality (Fig. 2g), department (Fig. 3d) and national scales after weighting the raster data values by population (see the section on Spatial aggregation of covariates).

Geographically based Economic data (G-Econ)
To account for socioeconomic differences, which are potentially associated with contact between humans and the vector, we used one-degree resolution gridded estimates of 2005 purchasing power parity (PPP) adjusted gross domestic product (GDP)49. To express the values in per capita, we divided the gridded GDP by the corresponding population, the latter obtained from the Gridded Population of the World product (v3)50 after resampling the latter to one-degree resolution. We chose this version of gridded population data for this task given that it was the one originally used to generate the 2005 gridded GDP values. Cells with missing values were imputed with the mean of the surrounding eight grid cell values. Once we obtained a complete grid layer at a resolution of one-degree (~111 km at the equator), we resampled the layer, without smoothing, to a resolution of 4.65 km x 4.65 km to match the resolution of all other gridded layers. We additionally computed aggregated results at the municipality, department and national levels after weighting them by the distribution of population (in the year 2005) within each administrative unit (see the section on Spatial aggregation of covariates).

Travel time
To account for the general accessibility of each municipality and department, we used travel time data downloaded from the European Commission’s Joint Research Center at a resolution of 30 arc seconds51.
This definition of travel time is a measure of overall accessibility rather than of frequency of travel. It is defined as the average length of time (in minutes) it takes individuals in a region to travel to the nearest location with a population greater than 50,000. Large travel time is indicative of a region whose population lives relatively far from urban centers. This gridded dataset has minutes of land-based travel time to the nearest settlement with population greater than 50,000 (as of the year 2000). The data is developed using a cost-distance model, which accounts for travel time increments based on the available transport networks and other environmental and political factors. We aggregated travel time weighted by population at the municipal level to generate estimates of travel time for each municipality and similarly for each department (see the section on Spatial aggregation of covariates).

Urban population
To identify the level of urbanization in each grid cell, we downloaded the MODIS global 2002 urban extent raster dataset, which has a binary (0 or 1) value for each 500 m x 500 m grid cell around the globe. By counting the number of high-resolution urban grid cells that fall within each standard grid cell of 4.65 km x 4.65 km, we were able to generate a gridded product of percentage of the physical grid cell that is urban. Furthermore, in combination with the population raster we obtained from WorldPop, we were able to generate a gridded estimate of urban population at each 500 m x 500 m grid cell in Colombia.

Code availability
The code used to generate all gridded datasets and aggregating at municipal, departmental, and national levels is freely available for download from GitHub at https://github.com/asiraj-nd/zika-colombia. This code utilizes the R programming language and Python version 2.7.10. Further explanation of the code is provided in a readme file in the repository on GitHub.

Data Records
All output datasets described in this article (Data Citation 1) are publicly and freely available through Dryad Digital Repository. The datasets stored in the datadryad.org Repository represent the ones produced at the time of writing, and will be preserved in their published form. Datasets of interest can be obtained by downloading the corresponding zipped archive files (Table 2 (available online only)).
Technical Validation

Most datasets obtained from other sources have already been validated by independent studies. We therefore limited our validation to the interpolated climate model outputs developed here by comparing spatial interpolation results to data from the 30 meteorological stations across Colombia. These comparisons were made for the two modeling approaches and for different combinations of covariates for each outcome: mean temperature, maximum temperature, minimum temperature, precipitation, and relative humidity.

We used three metrics to compare model performance: mean absolute error, coefficient of variation, and Pearson’s correlation coefficient (COR). Mean absolute error (MAE) is the mean absolute difference between predictions and observations over n data points:

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |\hat{y}_i - y_i|$$

We also used relative MAE (of two models), which is the ratio of the two MAEs. A relative MAE $m$ of models A and B respectively, would indicate that predictions from model A were $(1-m)%$ closer to the observed values than those from model B for an $m$ value less than 1. The coefficient of variation (CV) evaluates the extent to which large values are dispersed relative to their mean value. It is the ratio of the root mean square error (RMSE) to the mean of observed values,

$$CV = \frac{1}{\bar{y}} \left( \frac{1}{n} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \right)^{1/2}$$

Results of our comparison are described in Table 3. Overall, the ordinary kriging approach had higher accuracy for temperature (mean, maximum, and minimum) and relative humidity based on all three metrics. Model results also revealed that using other covariates, such as altitude and secondary climate data, improved interpolation results for temperature and relative humidity.

Usage Notes

This compilation of datasets can facilitate a variety of studies relevant to vector-borne disease epidemiology in Colombia. The archive provides ready to use data both in a raster format with resolution of 5km x 5km, and at administrative units of municipal, departmental, and national scales. These datasets have several limitations. First, the 30 meteorological stations used in generating climate surfaces are sparsely and unevenly distributed over Colombia, leading to uncertainty in the outputs. Moreover, some of the original gridded data we obtained had differing resolutions, including 0.1 arc-degrees (GPM), 0.5 arc-degrees (CPC), and 1 arc-degree (G-Econ). This meant that we had to resample these gridded products (GPM, CPC, GEcon) with crude estimates based on average values over a large swath of grid cells. Further, unlike all other products we used that were non-projected geographic WGS1984 raster files, the Terra and Aqua MODIS NDVI products were in sinusoidal projections, causing some distortions when re-projected to match population layers used in weighting.

In addition to spatial discrepancies, we also had to overcome the relatively poor temporal resolutions of Terra and Aqua MODIS NDVI products (which come at 16-day intervals) by linearly interpolating between two data points to fill in the 15 days in between, before aggregating the results at weekly time steps. Furthermore, daily satellite based rainfall data from NOAA assume 12:00-12:00 hour-day, which could potentially cause slight inconsistencies, despite the data finally being aggregated at weekly time steps. Other limitations include the modifiable area unit problem, which arises from disparities in the arbitrary sizes and borders of the administrative units which may bias aggregations based on these borders.

References

1. Cauchemez, S. et al. Association between Zika virus and microcephaly in French Polynesia, 2013–15: a retrospective study. Lancet 387, 2125–2132 (2016).
2. Duffy, M. et al. Zika virus outbreak on Yap Island, Federated States of Micronesia. New Engl. J. Med 360, 2536–2543 (2009).
3. Cao-Lormeau, V. et al. Guillain-Barré syndrome outbreak associated with Zika virus infection in French Polynesia: a case-control study. Lancet 387, 1531–1539 (2016).
4. Petersen, L. R., Jamieson, D. J., Powers, A. M. & Honein, M. A. Zika Virus. New Engl. J. Med 374, 1552–1563 (2016).
5. Krauer, F. et al. Zika virus infection as a cause of congenital brain abnormalities and Guillain–Barré syndrome: systematic review. PLoS Med. 14, e1002203 (2017).
6. Zhang, Q. et al. Spread of Zika virus in the Americas. Proc. Natl. Acad. Sci 114, 4334–4343 (2017).
7. Petersen, E. E. et al. Interim guidelines for pregnant women during a Zika virus outbreak - United States, 2016. Morb. Mortal. Wkly. Rep 65, 30–33 (2016).
8. Center for Disease Control and Prevention. Zika Virus. https://www.cdc.gov/zika/index.html (2017).
9. Kucharski, A. J. et al. Transmission dynamics of Zika virus in island populations: a modelling analysis of the 2013–14 French Polynesia outbreak. PLoS Negl. Trop. Dis 10, e0004726 (2016).
10. Christophers, S. R. Aedes aegypti, The Yellow Fever Mosquito: Its Life History Bionomics, and Structure (Cambridge University Press, 1960).
11. Focks, D. A., Haile, D. G., Daniels, E. & Mount, G. A. Dynamic life table model for Aedes aegypti (Dprera: Culicidae): analysis of the literature and model development. J. Med. Ent 30, 1003–1017 (1993).
12. Reiter, P. Climate change and mosquito-borne disease. Environ. Health Perspect. 109, 141–161 (2001).
13. Perkins, T. A., Metcalf, C. J., Grenfell, B. T. & Tatem, A. J. Estimating drivers of autochthonous transmission of chikungunya virus in its invasion of the Americas. PLoS Currents Outbreaks 7. doi:10.1371/currents.outbreaks.a7c66ac1e0a20b177e8c9767946d1fe, 1–21, February 10 (2015).
14. Meltzer, M. I. et al. Estimating the future number of cases in the Ebola epidemic—Liberia and Sierra Leone, 2014-2015. MMWR Suppl 63, 1–14 (2014).
15. Bellan, S. E. et al. Statistical power and validity of Ebola vaccine trials in Sierra Leone: a simulation study of trial design and analysis. Lancet 15, 703–710 (2015).
16. Gire, S. K. et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. Science 345, 1369–1372 (2014).
17. Perkins, T. A. Retracing Zika’s footsteps across the Americas with computational modelling. Proc. Nat. Acad. Sci 114, 5558–5560 (2017).
18. Chretien, J. P., Rivers, C. M. & Johansson, M. A. Make data sharing routine to prepare for public health emergencies. PLoS Med. 13, e1002109 (2016).
19. Center for Disease Control and Prevention. Data repository for publicly available Zika data https://github.com/cdczpi/zika (2017).
20. National Institute of Health. Number of confirmed and suspected zika cases by municipality. Institute National de Salud http://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/Forms/AllItems.aspx (2016).
21. Pacheco, O. et al. Zika virus disease in Colombia — preliminary report. New Engl. J. Med. doi:10.1056/NEJMoa1604037, NEJMoa1604037, 1–10 (2016).
22. Ali, S. et al. Environmental and social change drive the explosive emergence of Zika virus in the Americas. PLoS Negl. Trop. Dis 11, e0005135 (2017).
23. Reiner, J. C. et al. A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J. R. Soc. Interface 10, 20120921 (2013).
24. Kraemer, M. U. G et al. Big city, small world: density, contact rates, and transmission of dengue across Pakistan. J. R. Soc Interface 12, 201504682015.
25. Brady, O. J et al. Modelling adult Aedes aegypti and Aedes albopictus survival at different temperatures in laboratory and field settings. Parasit. Vectors 6, 351 (2013).
26. Chan, M. & Johansson, M. A. The incubation periods of dengue viruses. PLoS ONE 7, e59972 (2012).
27. Lambrechts, L. et al. Impact of daily temperature fluctuations on dengue virus transmission by Aedes aegypti. Proc. Nat. Acad. Sci 108, 7460–7465 (2011).
28. Kraemer, M. U. et al. The global distribution of the arbovirus vectors Aedes aegypti and Ae. albopictus. eLife 4, e08347 (2015).
29. Perkins, T. A., Siraj, A. S., Ruktanonchai, C. W., Kraemer, M. U. & Tatem, A. J. Model-based projections of Zika virus infections in childbearing women in the Americas. Nat. Microbiol 1, 16126 (2016).
30. WorldPop. Population – individual countries http://www.worldpop.org.uk/data/data_sources (2016).
31. Geographic Information System for Territorial Planning. Predefined thematic maps – national http://sigotn.igac.gov.co/sigotn/frames_pagina.aspx (2016).
32. National Administrative Department of Statistics. Population and fertility indicators https://www.dane.gov.co (2016).
33. National Administrative Department of Statistics. Geoportal http://geoportal.dane.gov.co/v2 (2016).
34. United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2015 Revision, DVD Edition. United Nations, (2015).
35. Parham, P. E. & Michael, E. Modeling the effects of weather and climate change on Malaria Transmission. Environ. Health Perspect. 11, 620–626 (2010).
36. National Oceanic and Atmospheric Administration, National Climate Data Center. Climate data online https://www7.ncdc.noaa.gov/CDO/cdselect.cmd (2016).
37. United States Geological Global. Global 30 arc-second elevations. https://lta.cr.usgs.gov/GTOPO30 (2016).
38. Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G. & Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. Int. J. Climat 25, 1965–1978 (2005).
39. Fan, Y. & van den Dool, H. A global monthly land surface air temperature analysis from 1948 to present. J. Geophys. Res. 113, D01103 (2008).
40. Nychka, D. F. Fields: tools for spatial data. National Center for Atmospheric Research. http://www.image.ucar.edu/GSP/Software/Fields (2016).
41. Ribeiro, J. R. & Diggle, P. J. geoR: A package for geostatistical analysis (R package). https://cran.r-project.org/doc/Rnews/Rnews_2001-2.pdf (2001).
42. R Core Team. R: a language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org (2016).
43. Banerjee, S., Carlin, B. P. & Gelfand, A. E. Hierarchical Modeling and Analysis for Spatial Data. 2nd edn, (Chapman and Hall/CRC, 2014).
44. Alduchov, O. A. & Eskridge, R. E. Improved Magnus form approximation of saturation vapor pressure. J. Appl. Meteorol. 35(601): 609 (1996).
45. Hay, S. I. Remote sensing and disease control: past, present and future. Trans. R. Soc. Trop. Med. Hyg 91, 105–106 (1997).
46. Thomson, M. C. et al. Predicting malaria infection in Gambia children from satellite data and bed net used surveys: he importance of spatial correlation in the interpretation of results. Am. J. Trop. Med. Hyg. 61, 2–8 (1999).
47. Zhang, Z. et al. Remote sensing and disease control in China: past present and future. Parasite & Vectors 6, 11 (2013).
48. National Oceanic and Atmospheric Research, Center for Satellite Applications and Research. Satellite rainfall estimates, 24 hours (12-12). https://www.star.nesdis.noaa.gov/smcd/emb/autoreadme.php (2017).
49. Nordhaus, W. Geography and macroeconomics: new data and new findings. Proc. Natl. Acad. Sci 103, 3510–3517 (2006).
50. Center for International Earth Science Information Network. Gridded population of the world, version 3. US NASA Socio-economic Data and Application Center. http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count/data-download (2017).
51. Nelson, A. Estimated travel time to the nearest city of 50,000 or more in year 2000. European Commission Joint Research Centre Global Environment Monitoring Unit. http://eboro.jrc.ec.europa.eu/products/gam (2017).
52. Schneider, A., Friedl, M. A. & Potere, D. A new map of global extent from MODIS data. Environ. Res. Lett. 4, 044003 (2009).
53. Schneider, A., Friedl, M. A. & Potere, D. Monitoring urban areas globally using MODIS 500m data: New methods and datasets based on urban ecoregions. Remote Sens. Environ. 114, 1733–1746 (2010).
54. Siraj, A. S. et al. Spatiotemporal incidence of Zika and associated environmental variables for the 2015-2016 epidemic in Colombia. https://www.github.org/asiraj-njd/zika-colombia (2017).

Data Citations
1. Siraj, A.S. et al. Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.83nj1 (2017).
Acknowledgements
This research was supported by a RAPID award from the National Science Foundation (DEB 1641130).

Author Contributions
Conceived of the research: A.S.S., T.A.P. Extracted and processed input data: A.S.S., I.R.B., C.M.B., D.H., C.L., D.L., G.O., G.E., M.U.G.K., N.T.G., A.J.T., R.C.R. Assembled data products and generated figures and tables: A.S.S. Performed technical validation: A.S.S., I.R.B., C.M.B., R.C.R. Wrote the manuscript: A.S.S., I.R.B., C.M.B., M.A.J., T.A.P. Reviewed the manuscript and contributed to revisions: D.H., C.L., D.L., G.O., G.E., M.U.G.K., A.J.T., R.C.R.

Additional Information
Table 2 is available only in the online version of this paper.

Competing interests: The authors declare no competing interests.

How to cite this article: Siraj, A. S. et al. Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia. Sci. Data 5:180073 doi: 10.1038/sdata.2018.73 (2018).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 openness

This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2018