American Journal of Epidemiology

Vol. 188, No. 5
DOI: 10.1093/aje/kwz010
Advance Access publication:
January 25, 2019

Practice of Epidemiology

Spatiotemporal Error in Rainfall Data: Consequences for Epidemiologic Analysis of Waterborne Diseases

Morgan C. Levy, Philip A. Collender, Elizabeth J. Carlton, Howard H. Chang, Matthew J. Strickland, Joseph N. S. Eisenberg, and Justin V. Remais*

* Correspondence to Dr. Justin V. Remais, Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, 2121 Berkeley Way #5302, Berkeley, CA 94720-7360 (e-mail: jvr@berkeley.edu).

Initially submitted June 6, 2018; accepted for publication January 10, 2019.

The relationship between rainfall, especially extreme rainfall, and increases in waterborne infectious diseases is widely reported in the literature. Most of this research, however, has not formally considered the impact of exposure measurement error contributed by the limited spatiotemporal fidelity of precipitation data. Here, we evaluate bias in effect estimates associated with exposure misclassification due to precipitation data fidelity, using extreme rainfall as an example. We accomplished this via a simulation study, followed by analysis of extreme rainfall and incident diarrheal disease in an epidemiologic study in Ecuador. We found that the limited fidelity typical of spatiotemporal rainfall data sets biases effect estimates towards the null. Use of spatial interpolations of rain-gauge data or satellite data biased estimated health effects due to extreme rainfall (occurrence) and wet conditions (accumulated totals) downwards by 35%–45%. Similar biases were evident in the Ecuadorian case study analysis, where spatial incompatibility between exposed populations and rain gauges resulted in the association between extreme rainfall and diarrheal disease incidence being approximately halved. These findings suggest that investigators should pay greater attention to limitations in using spatially heterogeneous environmental data sets to assign exposures in epidemiologic research.

bias; environmental epidemiology; exposure misclassification; extreme weather; measurement error; precipitation; waterborne diseases

Abbreviations: CI, confidence interval; GC, Atlantic Gulf Coast; NW, Pacific Northwest.

A growing body of epidemiologic research has reported associations between climate features—including extreme or heavy rainfall—and increases in the incidence of waterborne infectious diseases (1–7). These studies generally use summary measures of spatiotemporal environmental data, with a growing literature focused on rainfall data, specifically (1, 2, 8–10). Different summary measures (e.g., cumulative or maximum rainfall), as well as rainfall data drawn from different sources (e.g., weather stations or satellites), are used to derive exposures for estimating associations with disease. Error in the spatiotemporal representation of these variables, which in part determines data fidelity (defined as data quality in the context of its use (11), including instrument and sampling error), can be an important source of bias and uncertainty in effect estimates (12–15). While analogous issues related to measurement error and bias have been explored in air pollution epidemiology (12, 13, 16–21), findings generalizable to the unique characteristics of meteorological exposures—particularly extreme values—are lacking.

Analysis of measurement error in environmental data across different sources is an emerging area of research in the environmental sciences (22–26) as well as in air pollution epidemiology, where exposures are commonly assessed using sophisticated models (13–15, 20, 27, 28). Air-pollution epidemiologic research has shown that spatial errors in the measurement of air pollutant data, as well as errors introduced by modeled representations of those data, can bias effect estimates towards or away from the null value, depending on data treatment decisions and epidemiologic design. Specifically, spatial incompatibilty, defined as the lack of colocated environmental and epidemiologic data—such as in cases where researchers use nonlocal environmental monitoring data to quantify local exposure—have

been shown to produce bias in effect estimates (13–15, 27). However, few studies have compared results across various exposure data sources and types (e.g., across the number of sensors or gauges (16, 29) or for in situ versus satellite sources (20)).

As epidemiologic research expands its analysis of the effects of extreme weather on waterborne diseases (1–7), researchers need a clear understanding of the impact that meteorological exposure measurement error has on understanding patterns of risk. Developing this understanding is challenging given that complex relationships between climate regimes, seasonality, and topography result in variable spatial patterning of rainfall, especially in regions with complex physical or climatological characteristics (30–32). This suggests the need for high-resolution and regionally specific characterizations of rainfall. However, given constraints in obtaining optimal exposure measurement, there is also a need for guidance on how best to minimize bias when using low-fidelity data.

To this end, we examined the implications of spatiotemporal environmental data fidelity, as quantified by exposure measurement error, for epidemiologic analyses of rainfall effects on waterborne diseases. We quantified bias in the estimated health effects of exposure to extreme rainfall using a simulation study wherein we derived exposure variables from commonly used sources of environmental data, including satellite data and interpolations of weather station data, across randomly sampled configurations of exposed community and rainfall measurement locations and time periods. We then examined the consequences of measurement error through a case study, wherein we reanalyzed associations between extreme rainfall and diarrheal disease (2) with respect to different levels of rainfall data fidelity. While rainfall data include both spatial and temporal components, we focused on the effects of spatial error within spatiotemporal rainfall data (Web Appendix 1, available at https://academic.oup.com/aje). While numerous additional sources of error can affect spatiotemporal environmental data (Web Appendix 1), we focused here on isolating and exploring the potential for insufficient spatial fidelity implicit in commonly used sources of meteorological data to bias epidemiologic effect estimates.

METHODS

Simulation study

Simulation data. In the simulation study, which relies on commonly used spatiotemporal rainfall data sets, we considered 3 different types of rainfall data: 1) reference data, from which we obtain simulated "in situ" measurements; 2) interpolated in situ data; and 3) satellite data. Each data set included daily average depths of rainfall in units of millimeters per day (mm/day) for the period of 1983–2015 (33 years), which we aggregated to weekly summaries for analysis. We obtained data for 2 study regions in the United States (Figure 1), the Atlantic Gulf Coast (GC), located at latitudes 31–35 degrees north and longitudes 84–94 degrees west (area of 414,608 km^2), and the Pacific Northwest (NW), located at latitudes 44–48 degrees north and longitudes 114–124 degrees west (area of 343,851 km^2). We selected these regions due to: 1) their high density of weather stations, which enabled optimal characterization of spatiotemporal rainfall patterns by the reference data set (see below); and 2) the regions' relatively high annual rainfall and seasonal variation, which are patterns characteristic of rainfall evaluated in epidemiologic studies. Precipitation in the GC is less spatially varied than in the NW (Figure 1), and the GC experiences a greater number of mid- to high-magnitude rainfall events relative to the NW (Web Figure 1).

Reference rainfall data (Figures 2A and 3A) refers to high-quality gridded (0.04-degree, approximately 4 km × 4 km) Parameter-elevation Relationships on Independent Slopes Model (PRISM) daily precipitation, which is produced from sophisticated, validated spatial interpolations of in situ data over the continental United States (33, 34). Satellite data (Figure 2B and 3B) refers to Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks–Climate Data Record, Version 1.1 (PERSIANN-CDR), a satellite-derived, global gridded (0.25-degree, approximately 28 km × 28 km) daily precipitation data set (35). While other satellite sources exist, we limited our evaluation to a single, state-of-the-art satellite data set, representative of satellite sources in general (Web Appendix 2). Interpolated in situ data (Figures 2C and

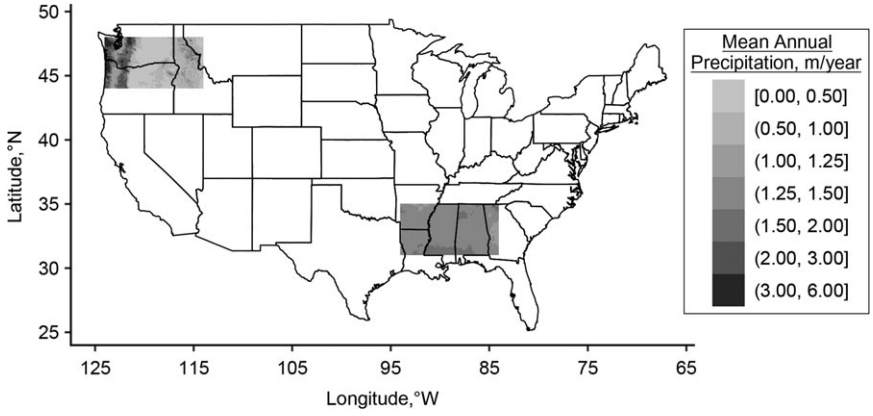

Figure 1. Regions and rainfall distributions for a simulation study. Mean annual precipitation from 1985–2013 in each of the 2 simulation study regions: the Pacific Northwest (NW, top left) and the Atlantic Gulf Coast (GC, bottom right) of the United States. Data represent daily precipitation from the Parameter-elevation Relationships on Independent Slopes Model (PRISM) (33, 34).

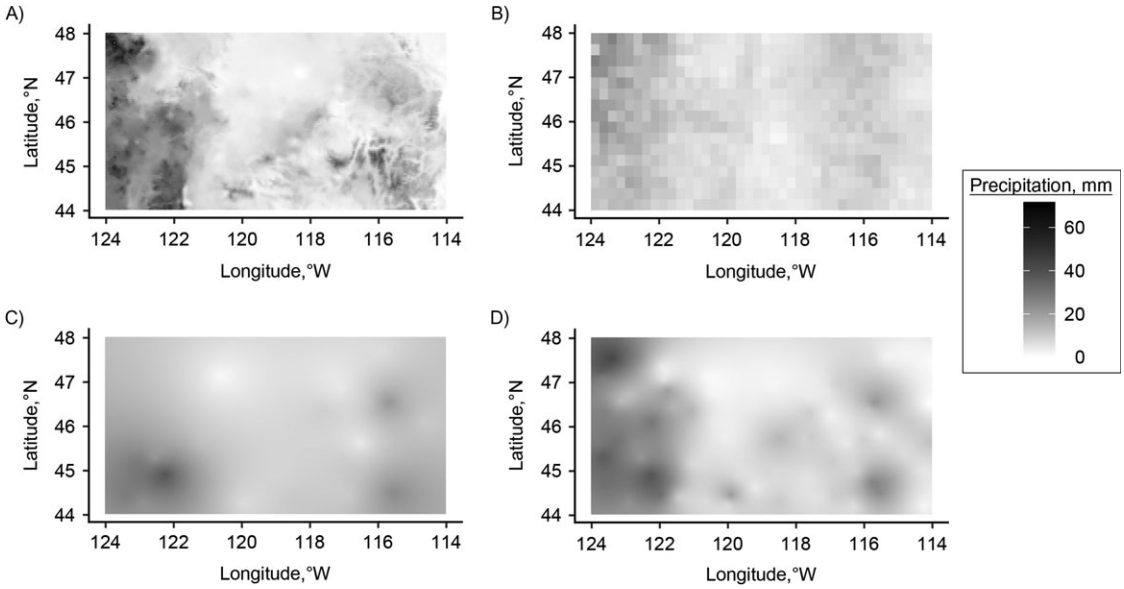

Figure 2. Rainfall data and sources in the Pacific Northwest (NW) region of the United States, 1985–2013. Rainfall data sources are reference (A), satellite (B), and interpolated in situ from 25 (C) and 100 (D) rain gauges. Reference precipitation is from the Parameter-elevation Relationships on Independent Slopes Model (PRISM) (4-km resolution) (33, 34); satellite precipitation is from the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks–Climate Data Record (PERSIANN-CDR) (28-km resolution) (35); and "rain gauges" used in interpolations are samples from PRISM, which are interpolated (shown at a 4-km resolution) using ordinary kriging.

2D and 3C and 3D) refers to spatial interpolations of rainfall from simulated rain-gauge locations, which are spatial points in the study regions from which time series of reference rainfall data are extracted. We used ordinary kriging with exponential covariance to interpolate simulated rain-gauge data at simulated exposed community locations (Web Appendix 2, Web Figure 2).

Epidemiologic model for simulation study. The simulation study implemented a simple Poisson regression model

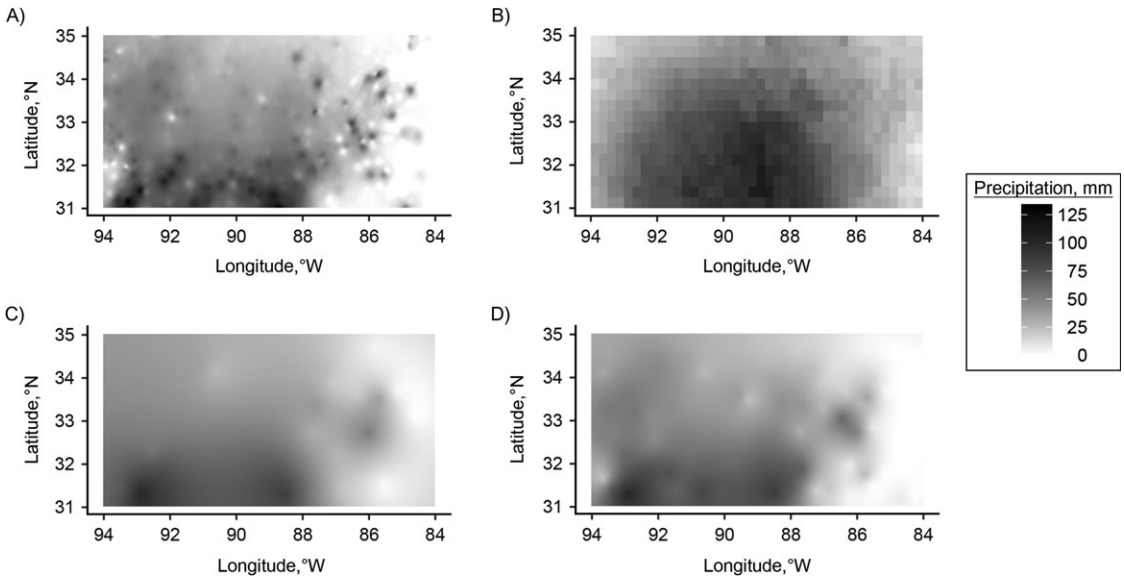

Figure 3. Rainfall data and sources in the Atlantic Gulf Coast (GC) region of the United States, 1985–2013. Rainfall data sources are reference (A), satellite (B), and interpolated in situ from 25 (C) and 100 (D) rain gauges. Reference precipitation is from the Parameter-elevation Relationships on Independent Slopes Model (PRISM) (4-km resolution) (33, 34); satellite precipitation is from the Precipitation Estimation from Remotely Sensed Information using Artificial Neural Networks–Climate Data Record (PERSIANN-CDR) (28-km resolution) (35); and "rain gauges" used in interpolations are samples from PRISM, which are interpolated (shown at a 4-km resolution) using ordinary kriging.

of disease incidence (counts) as it relates to a single rainfall exposure, summarized for a given community location (point) and week. We defined the target parameter as the exponentiated rainfall exposure coefficient ($e^{\hat{\beta}}$), or incidence rate ratio, for which we estimate bias. Based on a review of epidemiologic literature (Web Table 1), we summarized the rainfall exposure in 2 ways: 1) an indicator of exposure to extreme rainfall events, expressed as the occurrence of 1 or more days in a week with rainfall exceeding the 90th percentile of daily rainfall (hereafter, "extreme rainfall exposure"); and 2) an indicator of exposure to wet conditions, expressed as total weekly rainfall exceeding the 66th percentile of total weekly rainfall (hereafter, "wet conditions exposure"). We assumed all communities have a constant, same-size population.

Simulation procedure. We used a Monte Carlo procedure (1,000 iterations) to obtain distributions of the incidence rate ratio across simulated, randomly sampled configurations of rain-gauge locations, community locations, and time periods. We calculated bias by comparing the incidence rate ratio (mean across samples) with the true assigned incidence rate ratio. We evaluated the effect on estimation precision by comparing the standard deviation of the effect estimates and mean standard error calculated using interpolated and satellite data with those statistics calculated using reference rainfall data (Web Appendix 2). The simulation evaluated 2 connected determinants of spatial error in exposure measurement: 1) rain-gauge density (number of gauges per unit area); and 2) rain-gauge proximity (distance of communities to a rain gauge). For simulation details, see Web Appendix 2, and for example code, see Web Appendix 3. Additional code and data are available upon request to the authors.

Rain-gauge density. Within each iteration and region (GC and NW), we first sampled 25 community locations and 100 rain-gauge locations (both defined as randomly selected points within the region extent), and a 2-year time period (730 consecutive days from within the 33-year record). We refer to a single realization of the sampling in this first step as a "sample configuration." In our second step, we simulated disease counts at community locations using the true $\beta = 1.0$ and the reference rainfall at the community. In our third step, we used rainfall from the satellite data (grid cells overlapping community locations) and interpolated data (interpolated to community locations using region-specific, fixed kriging parameters) derived from a range of between 1 and 100 simulated rain gauges, to fit the Poisson model. We repeated the second and third steps 100 times, such that we simulated disease counts and fitted the Poisson model 100 times per sample configuration, and we averaged the results. Thus, we generated 1 average effect estimate ($\hat{\beta}$) and standard error specific to satellite and interpolated data sets per sample configuration. We modeled this analysis after a common scenario in epidemiologic studies carried out at large spatial scales, where rainfall measurement density varies by region.

Rain-gauge proximity: interpolation setting. Within a single iteration and region (defined above), we estimated $\hat{\beta}$ with interpolated data for individual communities located closest to (<1 km) versus farthest from (50–700 km) any gauge included in the interpolated data set. This analysis investigated whether data fidelity with respect to (1-dimensional) spatial proximity was similar to data fidelity with respect to (2-dimensional) density.

Rain-gauge proximity: individual rain-gauge setting. Within a single iteration and region, we estimated $\hat{\beta}$ for an individual community for which rainfall was obtained from a single simulated rain gauge located at increasing distances (0–100 km) from the community (Web Appendix 2); no satellite data were used. We modeled this analysis after a common scenario in epidemiologic studies in data-limited regions, where rainfall measurements might be available only at a single, potentially distant rain gauge.

In environmental epidemiologic studies, the researcher rarely selects the weather characteristics of the study period or the spatial configuration of rain-gauge locations and exposed community locations. We intended the simulation design to represent average bias across an array of real-world study designs realized from a large number of different time periods and spatial configurations. While our reference data set is available only over the continental United States, our simulation approach is generalizable to other regions and reference data sets: One could generate "true" exposures and disease data from analogous reference data sets (see Web Appendix 2).

Epidemiologic analysis of rainfall and diarrhea incidence in Ecuador

We built on and reanalyzed data from prior research on the association between rainfall and diarrheal diseases in northern coastal Ecuador (2, 36–38) in order to assess the sensitivity of this association to spatial incompatibility between epidemiologic and rainfall data. In the prior research, extreme rainfall following dry periods was found to be positively associated with diarrheal disease incidence, based on a combination of temporal imputation and spatial kriging of precipitation measurements from 3 rain gauges in the study region (2). To improve upon the original rainfall exposures, we increased the rain-gauge density by adding data from 3 additional government meteorological stations (Figure 4), and we modified the interpolation approach, using ordinary kriging with exponential variance and no temporal imputation (Web Appendix 4).

Epidemiologic model. We employed the same model used in the original study (2): a random effects Poisson regression formulated to estimate the association between extreme rainfall following wet, moderate, and dry periods and diarrheal disease (2 weeks after rainfall exposure) across 19 villages between February 18, 2004, and April 18, 2007. The outcome variable was the number of incident diarrhea cases in a given village and week, and the model included an offset for village population, village random effects, village diarrhea incidence in the prior week, and village remoteness (2, 36). The target parameter was the exponentiated coefficient (incidence rate ratio) on the interaction of an extreme rainfall indicator (occurrence of daily rainfall >90th percentile of rainfall within a week) with a dry period indicator (cumulative rainfall over the preceding eight weeks in the lower tertile).

Analysis of bias. In the prior epidemiologic analysis, the interaction of extreme rainfall and prior dry periods was significantly associated with increased diarrheal disease incidence. To explore bias in this association potentially attributable to

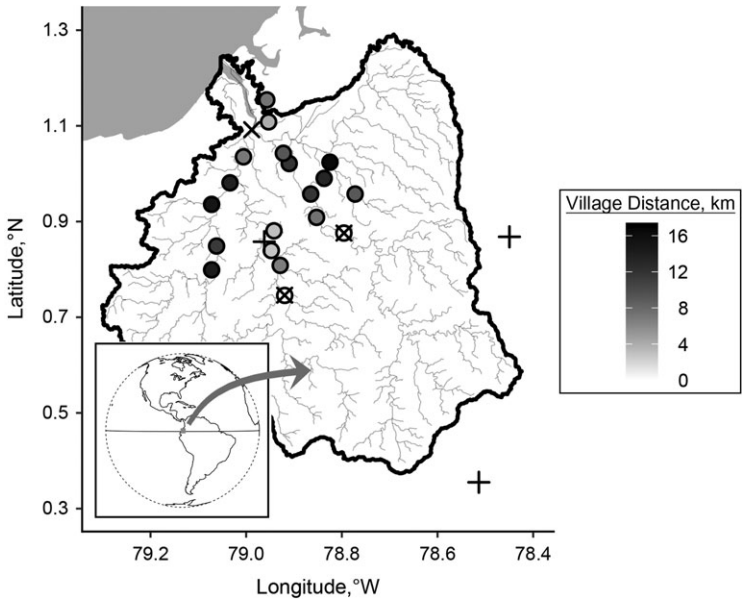

Figure 4. Relative location of rain gauges and villages in coastal Esmeraldas Province, Ecuador, 2004–2007. The map shows the location of the study region, within which there are 19 villages (points) and both original (plusses, from a previous study (2)) and new, supplementary rain-gauge data (crosses) from the same period. Shading indicates the distance of each village to the closest rain gauge; gray lines are the river network along which villages are situated; and the black outline is the coastal-draining river basin within which villages are located.

spatial incompatibility between epidemiologic and rainfall data sources, we evaluated differences between incidence rate ratios estimated on collections of villages that varied in their proximity to a local rain gauge. We estimated the incidence rate ratio on sets of villages, starting with the villages closest to a rain gauge, and progressively incorporated additional villages in order of their distance to a rain gauge. We compared the incidence rate ratios from this distance-based ordering with the incidence rate ratios made using averages from 1,000 random sequences of villages (we selected 1,000 sequences based on achieving stability in results). This comparison tested for deviations of the incidence rate ratios observable in distance-ordered sets from incidence rate ratios estimated with same-sized random groupings of villages.

RESULTS

Simulation study

Implications of rain-gauge density for bias in epidemiologic parameters. Effect estimates from satellite data and interpolations of 1–100 rain gauges exhibit significant biasing towards the null (no association) (Figure 5). For the extreme rainfall exposure, use of satellite rainfall downwardly biased incidence rate ratio estimates by 42% in the GC region and 43% in the NW region, and use of interpolated rainfall downwardly biased incidence rate ratio estimates by up to 45% (GC) and up to 40% (NW). For the wet conditions exposure, downward bias in the incidence rate ratio when using satellite rainfall was 37% (GC) and 36% (NW) and, when using interpolated rainfall, up to 43% (GC) and up to 35% (NW). Percent bias in terms of $\hat{\beta}$ was greater (Web Appendix 2, Web Figure 3). Variance of the effect estimate was relatively well-characterized by satellite and interpolated data for both exposures and across data sets, although modest bias was observed (Web Appendix 2, Web Figure 4). While bias in the NW area was slightly less negative than in the GC (Web Appendix 2), bias patterns with respect to rain-gauge density were similar across the 2 regions, despite significant differences in topography and climate. A sensitivity analysis revealed that effect estimate bias was sensitive to the magnitude of the true effect; the greater the true effect, the greater the downward bias (Web Appendix 2, Web Figure 5).

In all cases, interpolated rainfall from 5 or more rain gauges (just over 1 rain gauge per 10^5 km^2 for both regions) resulted in less bias than exposures generated from satellite data. While in all cases greater rain-gauge density reduced the magnitude of bias, there were decreasing marginal returns of increasing rain-gauge density, and substantial negative biases remained at high rain-gauge densities (Figure 5). For example, for the extreme rainfall exposure in the GC region, incidence rate ratio bias decreased by 18 percentage points (45%–27%) when increasing from 1 to 25 gauges but decreased by only 2 percentage points (21%–19%) when increasing from 75 to 100 gauges. Using 100 rain gauges (24 and 29 rain gauges per 10^5 km^2 for the GC and NW, respectively), incidence rate ratio bias ranged from 12% to 19%, depending on exposure and region. For context, these high rain-gauge densities would be considered suitable for the development of gridded rainfall data in data-scarce developing country regions (39). However, optimal density for capturing higher rainfall intensities (40) would require 8,000–9,000 rain gauges over the simulation regions. Optimally

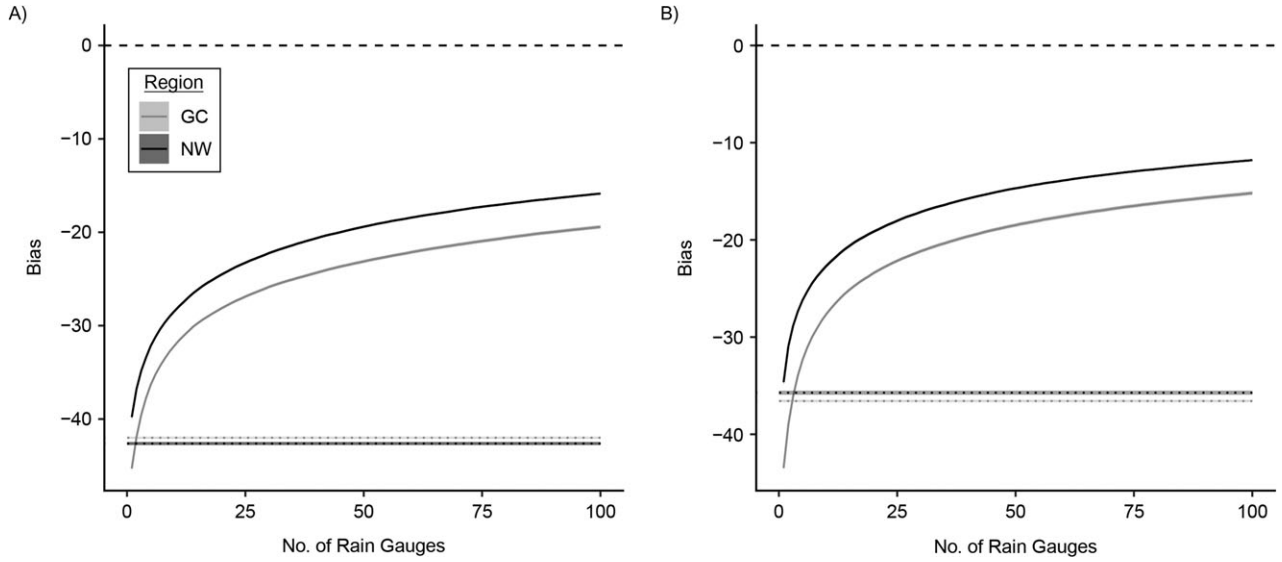

Figure 5. Bias in the estimated incidence rate ratio (IRR) measuring the effect of extreme rainfall and wet conditions on disease incidence in the simulation study based on data from the Pacific Northwest (NW) and the Atlantic Gulf Coast (GC) of the United States, 1985–2013. Bias in the extreme rainfall exposure (A) and wet conditions exposure (B) are shown with respect to region, indicated by color, and data type, where solid lines and dotted lines are results from interpolated and satellite data, respectively. The number of rain gauges (horizonal axis) does not pertain to satellite data. Bias is quantified as the percent of the true value of the IRR ($\exp(\beta) = 2.72$); negative bias indicates bias towards the null. Results are the mean of 1,000 simulation iterations; line thickness includes simulation uncertainty error.

high densities are therefore generally infeasible except at small spatial scales.

Implications of rain-gauge proximity for bias in epidemiologic parameters: interpolation setting. We evaluated the effect of proximity of individual communities to the nearest rain gauge in the case where we derived community rainfall exposures from interpolation of multiple gauges. This provided a stratified analysis of the rain-gauge density results, showing that incidence rate ratios for communities within 1 km of a rain gauge were unbiased, while incidence rate ratios for communities with distant gauges (50–700 km) were significantly biased (Web Appendix 2, Web Figures 6 and 7). This analysis demonstrated that environmental and epidemiologic spatial incompatibility governed bias associated with low rain-gauge density (Figure 5). Combined, these analyses demonstrate that elimination of bias in cases of spatial incompatibility between epidemiologic and environmental data through the increase of rain-gauge density (e.g., taking the horizontal axis of Figure 5 to its extreme) would require that densities be increased until epidemiologic and environmental data-source locations are approximately colocated.

Implications of rain-gauge proximity for bias in epidemiologic parameters: individual rain-gauge setting. We evaluated the effect of proximity of individual communities to an individual rain gauge from which rainfall estimates at the community are derived. For all exposures, the bias of the effect estimates increased with distance. For both exposures, use of a rain gauge not located "on site" (e.g., <1 km) resulted in bias towards the null (Figure 6). At the maximum distance evaluated of 100 km, estimates of the incidence rate ratio were biased downward by 32% and 35% for the extreme rainfall exposure, and 27% and 31% for the wet conditions exposure, in the NW and GC, respectively. Percent bias in terms of $\hat{\beta}$ was greater (Web Appendix 2, Web Figure 8). This result further demonstrates that the effect of spatial incompatibility realized in an interpolation setting is also present in the use of a single, distant data source used to quantify exposure. The estimated percent bias at different distances pertains both to use of an individual gauge and to the case where the distance of the closest rain gauge used in an interpolation is equivalent to those distances.

Epidemiologic analysis of rainfall and diarrhea incidence in Ecuador

When estimated with improved rainfall exposure data (see Methods), the incidence rate ratio associated with extreme rainfall events following dry periods in the Ecuador study region (incidence rate ratio = 2.05, 95% confidence interval (CI): 1.36, 3.11) was higher than the estimate from prior work (incidence rate ratio = 1.39, 95% CI: 1.03, 1.87). Subsetting the analysis to villages increasingly proximal to rain gauges increased the magnitude of the estimated incidence rate ratio (Figure 7A). For example, when the model was fitted using only the 5 villages less than 5 km from a rain gauge, the incidence rate ratio increased substantially (incidence rate ratio = 4.93, 95% CI: 2.89, 8.42). Assuming that villages close to or far from a rain gauge did not vary systematically in their sensitivity to rainfall exposures, the comparison between these incidence rate ratio estimates provides an approximation of the bias induced by spatial incompatibility (i.e., the incidence rate ratio increased above 5 when gauges and villages were colocated and decreased to less than 2.5 when including villages located up to 15 km from a rain gauge).

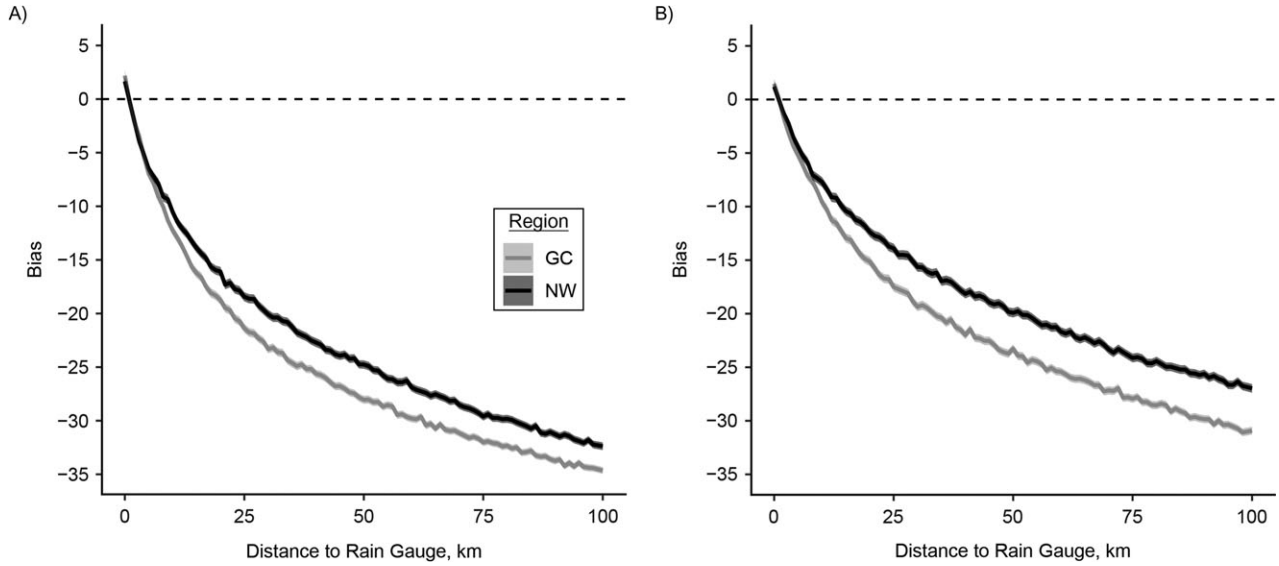

Figure 6. Bias in the estimated incidence rate ratio measuring the effect of extreme rainfall and wet conditions on disease incidence as a function of distance between an individual community and rain gauge in the simulation study based on data from the Pacific Northwest (NW) and the Atlantic Gulf Coast (GC) of the United States, 1985–2013. Bias in the extreme rainfall exposure (A) and wet conditions exposure (B) are shown with respect to region, indicated by color. Bias is quantified as the percent of the true value of the incidence rate ratio ($\exp(\beta) = 2.72$); negative bias indicates bias towards the null. Results are the mean of 5,000 simulation iterations; line thickness includes simulation uncertainty error.

There was substantial uncertainty in the estimates of incidence rate ratios for subsets of villages nearest to rain gauges due to small sample sizes (as indicated by the grey areas in Figure 7). However, comparison of the downward trend in estimates of the incidence rate ratio obtained using distance-based orderings of villages with estimates obtained using random orderings of villages (Figure 7B) strongly suggests that distance from a rain gauge explains the observed change, with distance, in the magnitude of the incidence rate ratio. The difference between the incidence rate ratios estimated for distance-based and random-ordered sets of villages are not statistically significant due to low statistical power when analyzing small subsets of villages. Nevertheless, a suggestive pattern emerged that is consistent with the simulation findings, showing an approximately 50% downward bias due to spatial incompatibility between epidemiologic and rainfall data.

DISCUSSION

In the simulation and case study settings, exposure misclassification due to spatial incompatibility between epidemiologic and precipitation data resulted in a bias towards the null. This effect attenuation stems from reduced correlation of the outcome with the modeled exposure relative to the true exposure. In the case of interpolated or satellite data, this is a result of smoothing of the rainfall signal (25, 26). A quantile-based threshold applied on smoothed data, relative to reference data, will generate the same total number of extreme exposures; however, some modeled exposures will occur on days different from those indicated by reference data. The observed attenuation of the effect measure is consistent with prior findings in air pollution epidemiology where use of exposure models tends to smooth an exposure measure, attenuating effect estimates and their standard errors (13–15, 17, 20, 27, 28). In the case of the use of a single distant rain gauge rather than interpolation, the attenuation is due to spatial correlation in rainfall naturally decreasing with distance, which occurs at different rates in different regions (41).

Two potential limitations of the simulation warrant discussion. First, we fixed interpolation parameters a priori in order to focus on error induced solely by rain-gauge density. Because interpolation specification generally depends on the number of rain gauges, our estimates of effect bias due to exposure measurement error might represent lower bound estimates. However, preliminary analyses and prior research (22) suggest that modification of the interpolation specification would have a limited impact on simulation findings, relative to the density of rain-gauge data being interpolated (see Web Appendix 2 and Web Figure 9). The magnitude of bias from the single-gauge distance analysis (Figure 6), where interpolation plays no role, further supports this view. Second, our simulation analysis design assumed that measures of rainfall determined counts of disease according to a model with no confounding. In reality, a combination of hydrological, ecological, and social processes might confound the relationship between rainfall and disease (9, 10). Nevertheless, the similarity of conclusions from the Ecuadorian case study, which examined the influence of spatial incompatibility on estimates of the incidence rate ratio using data with residual confounding, revealed similar downward biases. This supports the conclusion that spatial incompatibility between epidemiologic and environmental data can be a primary cause of bias in environmental epidemiologic studies. Relative to

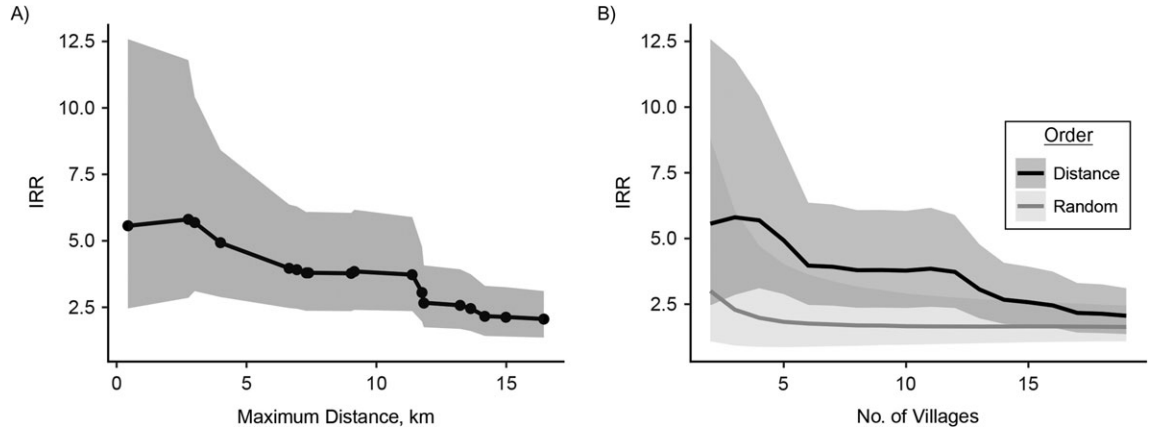

Figure 7. Changes in the incidence rate ratio (IRR) measuring the association between extreme rainfall and disease incidence according to village and rain-gauge proximity, using data from previous studies in Ecuador (2, 36–38). A) Points and lines are IRR estimates made using data from groups of villages that include those closest to (left) and those farthest from (right) a rain gauge; the horizontal axis is the maximum distance from any village in the group to a rain gauge. The points indicate values (IRR, maximum distance) for groups of villages (increasing in count from 2 to 19) that were used for regression model fitting. B) The same black line as in panel A, plotted along the number of villages (instead of distances) and alongside the mean IRR from 1,000 random sequencing of same-sized groups of villages (gray line). The shaded regions are the 95% confidence intervals.

the attention paid to confounding, the issue of spatial incompatibility receives little attention in the literature despite the significant bias it generates.

In situations where investigators have access to more than 1 environmental data source, investigators are commonly encouraged to compare sources by evaluating basic data features (e.g., spatial and temporal resolution) a priori to judge their analytical suitability (42). Yet, examining such features is often not sufficient to understand how the choice of data source might yield exposure measurement error and bias in estimated effects. Thorough understanding generally requires a study design- and site-specific evaluation of multiple environmental exposure data sets or data models (20), which might not always be possible. Unfortunately, few studies of the association between rainfall and the incidence of waterborne disease include adequately detailed information about rainfall data sources to perform such assessments post hoc, which limits their reproducibility and generalizability (43).

Given the reality of environmental data limitations, our findings indicate 3 recommendations. First, environmental epidemiologic studies in regions with limited environmental monitoring would benefit from reporting of data-source fidelity. For example, this could include reporting of average distance between epidemiologic and environmental data measurement points when in situ (e.g., weather station) or interpolated data are used; in situ measurement location density when interpolations are used; and in situ measurement location densities as reported in quality-controlled global data sets used in calibration and validation of most satellite and gridded data sets (e.g., the National Center for Atmospheric Research and Global Precipitation Climatology Centre data sets) when gridded products are used. This would indicate (qualitatively) the vulnerability of study findings to bias in regions where it is not possible to determine and select an optimal data source.

Second, studies with more than one accessible data source could include exploratory analyses that evaluate differences in effect estimates as a function of environmental data distance or density, as was demonstrated in our Ecuador case study. For example, when multiple weather stations are used in an analysis, estimates conditional on station proximity could be compared; when satellite data are used, satellite data–derived estimates could be compared with estimates made with available ground network data. Differences between results achieved at different spatial proximities, or between results achieved by satellite and ground network sources, would provide an estimate of bias attributable to data-source fidelity. The likely outcome of these efforts would be improved understanding of the true magnitude of effects of extreme climate on disease, which, according to this study, are susceptible to substantial underestimation.

While this study focused on rainfall, the findings are relevant to other environmental data types as well. The limited exploration of measurement error issues in estimating rainfall exposures for epidemiology, as well as rainfall's high spatial heterogeneity relative to other environmental features, motivated our analysis. Because other environmental features (e.g., temperature) tend to be less spatially variable, we would expect similar biases to arise for other environmental exposures but to a lesser degree. While our results are specific to indicator exposures, exploratory analyses demonstrated that similar biases arise for continuous exposures (Web Appendix 2, Web Figures 10–12). A thorough examination of continuous exposure quantifications was beyond the scope of this study, largely due to the diversity of continuous exposure formulations in the literature, and significant variation in their association with disease, relative to indicator exposures (Web Appendix 1, Web Table 1). While topographic and climate regime variation between different regions generated differences in bias,

bias trends and magnitudes were nevertheless remarkably similar. While the 2 simulation regions do not capture the complete set of topographies, climate regimes, and spatial scales represented in health studies, the results indicate that significant biasing might arise in any study where data constraints or study design prohibit close proximity between environmental and epidemiologic data.

ACKNOWLEDGMENTS

Author affiliations: School of Global Policy and Strategy, University of California, San Diego, San Diego, California (Morgan C. Levy); Division of Environmental Health Sciences, School of Public Health, University of California, Berkeley, Berkeley, California (Philip A. Collender, Justin V. Remais); Department of Environmental and Occupational Health, Colorado School of Public Health, University of Colorado, Aurora, Colorado (Elizabeth J. Carlton); Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, Georgia (Howard H. Chang); School of Community Health Sciences, University of Nevada, Reno, Reno, Nevada (Matthew J. Strickland); and Department of Epidemiology, University of Michigan, Ann Arbor, Ann Arbor, Michigan (Joseph N. S. Eisenberg).

This work was supported by National Science Foundation Water, Sustainability and Climate (grants 1360330 and 1646708), the National Science Foundation/US Department of Agriculture, National Institute of Food and Agriculture Integrated Food Energy Water Systems (INFEWS) program (grants T1–1639318/1316536), the National Institute of Allergy and Infectious Diseases (grants R01AI125842 and R01AI050038), and the National Institutes of Health Fogarty International Center (grant R01TW010286).

We thank the Ecuador Instituto Nacional de Meteorología e Hidrología (INAMHI) for providing meteorological data. M.C.L. thanks Drs. Christopher Paciorek, Sally Thompson, and Jennifer Burney, as well as the University of California, Berkeley, Remais group students and staff for their support and feedback.

Preliminary results of this research were presented at University of California, San Diego Institute for Public Health's 4th Annual Public Health Research Day, April 4, 2018, San Diego, California.

Conflict of interest: none declared.

REFERENCES

1. Bush KF, O'Neill MS, Li S, et al. Associations between extreme precipitation and gastrointestinal-related hospital admissions in Chennai, India. *Environ Health Perspect*. 2014; 122(3):249–254.
2. Carlton EJ, Eisenberg JN, Goldstick J, et al. Heavy rainfall events and diarrhea incidence: the role of social and environmental factors. *Am J Epidemiol*. 2014;179(3):344–352.
3. Chen MJ, Lin CY, Wu YT, et al. Effects of extreme precipitation to the distribution of infectious diseases in Taiwan, 1994–2008. *PLoS One*. 2012;7(6):e34651.
4. Curriero FC, Patz JA, Rose JB, et al. The association between extreme precipitation and waterborne disease outbreaks in the United States, 1948–1994. *Am J Public Health*. 2001;91(8): 1194–1199.
5. Constantin de Magny G, Murtugudde R, Sapiano MR, et al. Environmental signatures associated with cholera epidemics. *Proc Natl Acad Sci U S A*. 2008;105(46): 17676–17681.
6. Thomas KM, Charron DF, Waltner-Toews D, et al. A role of high impact weather events in waterborne disease outbreaks in Canada, 1975–2001. *Int J Environ Health Res*. 2006;16(3): 167–180.
7. Tornevi A, Axelsson G, Forsberg B. Association between precipitation upstream of a drinking water utility and nurse advice calls relating to acute gastrointestinal illnesses. *PLoS One*. 2013;8(7):e69918.
8. Jagai JS, Li Q, Wang S, et al. Extreme precipitation and emergency room visits for gastrointestinal illness in areas with and without combined sewer systems: an analysis of Massachusetts data, 2003–2007. *Environ Health Perspect*. 2015;123(9):873–879.
9. Levy K, Woster AP, Goldstein RS, et al. Untangling the impacts of climate change on waterborne diseases: a systematic review of relationships between diarrheal diseases and temperature, rainfall, flooding, and drought. *Environ Sci Technol*. 2016;50(10):4905–4922.
10. Mukabutera A, Thomson D, Murray M, et al. Rainfall variation and child health: effect of rainfall on diarrhea among under 5 children in Rwanda, 2010. *BMC Public Health*. 2016;16:731.
11. Natarajan P, Frenzel JC, Smaltz DH. *Demystifying Big Data and Machine Learning for Healthcare*. HIMSS Book Series. Boca Raton, FL: CRC Press, Taylor & Francis Group; 2017.
12. Goldman GT, Mulholland JA, Russell AG, et al. Impact of exposure measurement error in air pollution epidemiology: effect of error type in time-series studies. *Environ Health*. 2011;10:61.
13. Keller JP, Chang HH, Strickland MJ, et al. Measurement error correction for predicted spatiotemporal air pollution exposures. *Epidemiology*. 2017;28(3):338–345.
14. Szpiro AA, Paciorek CJ. Measurement error in two-stage analyses, with application to air pollution epidemiology. *Environmetrics*. 2013;24(8):501–517.
15. Szpiro AA, Paciorek CJ, Sheppard L. Does more accurate exposure prediction necessarily improve health effect estimates? *Epidemiology*. 2011;22(5):680–685.
16. Alexeeff SE, Schwartz J, Kloog I, et al. Consequences of kriging and land use regression for $PM_{2.5}$ predictions in epidemiologic analyses: insights into spatial variability using high-resolution satellite data. *J Expo Sci Environ Epidemiol*. 2015;25(2):138–144.
17. Armstrong BG. Effect of measurement error on epidemiological studies of environmental and occupational exposures. *Occup Environ Med*. 1998;55(10):651–656.
18. Baxter LK, Dionisio KL, Burke J, et al. Exposure prediction approaches used in air pollution epidemiology studies: key findings and future recommendations. *J Expo Sci Environ Epidemiol*. 2013;23(6):654–659.
19. Goldman GT, Mulholland JA, Russell AG, et al. Ambient air pollutant measurement error: characterization and impacts in a time-series epidemiologic study in Atlanta. *Environ Sci Technol*. 2010;44(19):7692–7698.
20. Jerrett M, Turner MC, Beckerman BS, et al. Comparing the health effects of ambient particulate matter estimated using ground-based versus remote sensing exposure estimates. *Environ Health Perspect*. 2017;125(4):552–559.

21. Strickland MJ, Gass KM, Goldman GT, et al. Effects of ambient air pollution measurement error on health effect estimates in time-series studies: a simulation-based analysis. *J Expo Sci Environ Epidemiol*. 2015;25(2):160–166.
22. Levy MC, Cohn A, Lopes AV, et al. Addressing rainfall data selection uncertainty using connections between rainfall and streamflow. *Sci Rep*. 2017;7(1):219.
23. Ljungqvist FC, Krusic PJ, Sundqvist HS, et al. Northern Hemisphere hydroclimate variability over the past twelve centuries. *Nature*. 2016;532(7597):94–98.
24. Massonnet F, Bellprat O, Guemas V, et al. Using climate models to estimate the quality of global observational data sets. *Science*. 2016;354(6311):452–455.
25. Hofstra N, New M, McSweeney C. The influence of interpolation and station network density on the distributions and trends of climate variables in gridded daily data. *Clim Dyn*. 2010;35(5):841–858.
26. Director H, Bornn L. Connecting point-level and gridded moments in the analysis of climate data. *J Clim*. 2015;28(9):3496–3510.
27. Gryparis A, Paciorek CJ, Zeka A, et al. Measurement error caused by spatial misalignment in environmental epidemiology. *Biostatistics*. 2009;10(2):258–274.
28. Lopiano KK, Young LJ, Gotway CA. Estimated generalized least squares in spatially misaligned regression models with Berkson error. *Biostatistics*. 2013;14(4):737–751.
29. Basagaña X, Aguilera I, Rivera M, et al. Measurement error in epidemiologic studies of air pollution based on land-use regression models. *Am J Epidemiol*. 2013;178(8):1342–1346.
30. Ebert EE, Janowiak JE, Kidd C. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bull Am Meteorol Soc*. 2007;88(1):47–64.
31. Gebregiorgis AS, Hossain F. Understanding the dependence of satellite rainfall uncertainty on topography and climate for hydrologic model simulation. *IEEE Trans Geosci Remote Sens*. 2013;51(1):704–718.
32. Mair A, Fares A. Comparison of rainfall interpolation methods in a mountainous region of a Tropical Island. *J Hydrol Eng*. 2011;16(4):371–383.
33. Hart EM, Bell K. *prism: Download Data From the Oregon prism Project*. R package version 0.0.6; 2015. http://github.com/ropensci/prism. Accessed August 9, 2017.
34. Prism Climate Group OSU. *PRISM Spatial Climate Datasets, Recent Time Series Datasets: Daily*. 2017. http://prism.oregonstate.edu/recent/. Accessed August 9, 2017.
35. Ashouri H, Hsu KL, Sorooshian S, et al. PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies. *Bull Am Meteorol Soc*. 2015;96(1):69–83.
36. Eisenberg JN, Cevallos W, Ponce K, et al. Environmental change and infectious disease: how new roads affect the transmission of diarrheal pathogens in rural Ecuador. *Proc Natl Acad Sci U S A*. 2006;103(51):19460–19465.
37. Markovitz AR, Goldstick JE, Levy K, et al. Where science meets policy: comparing longitudinal and cross-sectional designs to address diarrhoeal disease burden in the developing world. *Int J Epidemiol*. 2012;41(2):504–513.
38. Zelner JL, Trostle J, Goldstick JE, et al. Social connectedness and disease transmission: social organization, cohesion, village context, and infection risk in rural Ecuador. *Am J Public Health*. 2012;102(12):2233–2239.
39. Ali A, Amani A, Diedhiou A, et al. Rainfall estimation in the Sahel. Part II: evaluation of rain gauge networks in the CILSS countries and objective intercomparison of rainfall products. *J Appl Meteorol*. 2005;44(11):1707–1722.
40. Lopez MG, Wennerström H, Nordén LÅ, et al. Location and density of rain gauges for the estimation of spatial varying precipitation. *Geogr Ann Ser A Phys Geogr*. 2015;97(1):167–179.
41. Bacchi B, Kottegoda NT. Identification and calibration of spatial correlation patterns of rainfall. *J Hydrol*. 1995;165(1):311–348.
42. Duncan BN, Prados AI, Lamsal LN, et al. Satellite data of atmospheric pollution for US air quality applications: examples of applications, summary of data end-user resources, answers to FAQs, and common mistakes to avoid. *Atmos Environ*. 2014;94:647–662.
43. Cann KF, Thomas DR, Salmon RL, et al. Extreme water-related weather events and waterborne disease. *Epidemiol Infect*. 2013;141(4):671–686.