Title Page

Title: Spatiotemporal variation in risk of Shigella infection in childhood: a global risk mapping and prediction model using individual participant data

Author names and affiliations:
Hamada S. Badr - Department of Earth and Planetary Sciences, Johns Hopkins Krieger School of Arts and Sciences, Baltimore, 21218, MA, USA
Josh M. Colston - Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, 22903, USA
Nhat-Lan H. Nguyen – College of Arts and Sciences, University of Virginia, Virginia, USA
Yen Ting Chen - Department of Emergency Medicine, Chi-Mei Medical Center, Tainan, Taiwan
Syed Asad Ali - Department of Pediatrics and Child Health, Aga Khan University, Karachi, 74800, Pakistan
Ajit Rayamajhi - Department of Pediatrics, National Academy of Medical Sciences, Kanti Children's Hospital, Kathmandu, Nepal
Syed M. Satter - Programme for Emerging Infections, Infectious Diseases Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, 1212, Bangladesh
Nguyen Van Trang - National Institute of Hygiene and Epidemiology, Ha Noi, Vietnam
Daniel Eibach - Department of Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany
Ralf Krumkamp - Department of Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany
Jürgen May - Department of Infectious Disease Epidemiology, Bernhard Nocht Institute for Tropical Medicine (BNITM), Hamburg, Germany
Ayola Akim Adegnika - Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany
Gédéon Prince Manouana - Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany
Peter Gottfried Kremsner - Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany
Roma Chilengi - Centre for Infectious Disease Research in Zambia, Lusaka, Zambia
Luiza Hatyoka - Enteric diseases and Vaccines Unit, Centre for Infectious Disease Research in Zambia, Lusaka, Zambia
Amanda K. Debes - Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
Jerome Ateudjieu - Faculty of Medicine and Pharmaceutical Sciences, University of Dschang, Dschang, Cameroon; Department of Health Research, M.A. SANTE (Meilleur

1These authors contributed equally.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.
Access aux Soins en Santé), Cameroon; Division of Health Operations Research, Cameroon Ministry of Public Health
Abu S. G. Faruque - Centre for Nutrition & Food Security, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, 1212, Bangladesh
M. Jahangir Hossain - Medical Research Council Unit, The Gambia at the London School of Hygiene & Tropical Medicine, P. O. Box 273, Banjul, The Gambia
Suman Kanungo - National Institute of Cholera and Enteric Diseases, Kolkata, India
Karen L. Kotloff - Department of Pediatrics, University of Maryland School of Medicine, Baltimore, 21201, MD, USA
Inácio Mandomando - Centro de Investigação em Saúde de Manhiça, Manhica, Mozambique
M. Imran Nisar - Department of Pediatrics and Child Health, The Aga Khan University, Karachi, 74800, Pakistan
Richard Omore - Kenya Medical Research Institute, Center for Global Health Research, Kisumu, Nyanza, 40100, Kenya
Samba O. Sow - Centre pour le Développement des Vaccins, Mali, Bamako, Mali
Anita K. M. Zaidi - Department of Pediatrics and Child Health, The Aga Khan University, Karachi, 74800, Pakistan
Nathalie Lambrecht - Institute of Public Health, Charité - Universitätsmedizin Berlin, Corporate Member of Freie Universität Berlin and Humboldt-Universität zu Berlin, Berlin, Germany; Research Department 2, Potsdam Institute for Climate Impact Research (PIK), Member of the Leibniz Association, Potsdam, Germany
Bright Adu - Department of Immunology, Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, Ghana
Nicola Page - Centre for Enteric Diseases, National Institute for Communicable Diseases, Pretoria, South Africa
James A. Platts-Mills - Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, 22903, USA
Cesar Mavacala Freitas - Hospital Pediátrico David Bernardino, Luanda, Angola
Tuula Pelkonen - New Children’s Hospital, Pediatric Research Center and Helsinki University Hospital, Helsinki, Finland
Per Ashorn - Centre for Child, Adolescent, and Maternal Health Research, Faculty of Medicine and Health Technology, Tampere University and Tampere University Hospital, Tampere, Finland
Kenneth Maleta - College of Medicine, University of Malawi, Blantyre, Malawi
Tahmeed Ahmed - Nutrition and Clinical Services Division, International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b), Dhaka, 1212, Bangladesh
Pascal Bessong - HIV/AIDS & Global Health Research Programme, University of Venda, Thohoyandou, Limpopo, 0950, South Africa
Zulfiqar A. Bhutta - Center of Excellence in Women and Child Health, Aga Khan University, Karachi, 74800, Pakistan
Carl Mason - Department of Enteric Diseases, Armed Forces Research Institute of Medical
Sciences (AFRIMS), Bangkok, Thailand
Estomih Mduma - Haydom Global Health Institute, Haydom, P.O. Box 9000, Tanzania
Maribel P. Olortegui - Asociacion Benefica PRISMA, Iquitos, 16006, Peru
Pablo Peñataro Yori - Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, 22903, USA
Aldo A. M. Lima - Department of Physiology and Pharmacology, Faculty of Medicine, Federal University of Ceará, Fortaleza, 60020-181, Brazil
Gagandeep Kang - Department of Gastrointestinal Sciences, Christian Medical College, Vellore, 632004, India
Jean Humphrey - Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, 21205, MA, USA
Robert Ntozini - Zvitambo Institute for Maternal and Child Health Research, Harare, Zimbabwe
Andrew J. Prendergast - Blizard Institute, Queen Mary University of London, London, E1 2AT, UK
Kazuhiisa Okada - Research Institute for Microbial Diseases, Osaka University, Osaka
Warawan Wongboot - Department of Medical Sciences, National Institute of Health, Nonthaburi, Thailand
Nina Langeland - Department of Clinical Science, University of Bergen, Bergen, Norway
Sabrina J. Moyo - Department of Clinical Science, University of Bergen, Bergen, Norway
James Gaensbauer - Center for Global Health, Department of Epidemiology, Colorado School of Public Health, Aurora, CO, 80045, USA
Mario Melgar - Pediatric Infectious Diseases, Hospital Roosevelt, Guatemala City, Guatemala
Matthew Freeman - Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, 30322, GA, USA
Anna N. Chard - Gangarosa Department of Environmental Health, Rollins School of Public Health, Emory University, Atlanta, 30322, GA, USA
Vonethalom Thongpaseuth - Laboratory and Treatment Unit, Center for Malariology, Parasitology, and Entomology, Ministry of Health, Vientiane, Lao PDR
Eric Houpt - Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, 22903, USA
Benjamin F. Zaitchik - Department of Earth and Planetary Sciences, Johns Hopkins Krieger School of Arts and Sciences, Baltimore, 21218, MA, USA
Margaret N. Kosek - Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, 22903, USA

Corresponding author:
Correspondence about this manuscript can be addressed to: Dr. Benjamin F. Zaitchik, Department of Earth and Planetary Sciences, Johns Hopkins Krieger School of Arts and Sciences, Baltimore, 21218, MA, USA, zaitchik@jhu.edu, +1-410-516-4223
Abstract:

**Background:** Diarrheal disease remains a leading cause of childhood illness and mortality and *Shigella* is a major etiological contributor for which a vaccine may soon be available. This study aimed to model the spatiotemporal variation in pediatric *Shigella* infection and map its predicted prevalence across low- and middle-income countries (LMICs).

**Methods:** Independent participant data on *Shigella* positivity in stool samples collected from children aged ≤59 months were sourced from multiple LMIC-based studies. Covariates included household- and subject-level factors ascertained by study investigators and environmental and hydrometeorological variables extracted from various data products at georeferenced child locations. Multivariate models were fitted, and prevalence predictions obtained by syndrome and age stratum.

**Findings:** 20 studies from 23 countries contributed 66,563 sample results. Age, symptom status, and study design contributed most to model performance followed by temperature, wind speed, relative humidity, and soil moisture. *Shigella* probability exceeded 20% when both precipitation and soil moisture were above average and had a 43% peak in uncomplicated diarrhea cases at 33°C temperatures, above which it decreased. Improved sanitation and open defecation decreased *Shigella* odds by 19% and 18% respectively compared to unimproved sanitation.

**Interpretation:** The distribution of *Shigella* is more sensitive to climatological factors like temperature than previously recognized. Conditions in much of sub-Saharan Africa are particularly propitious for *Shigella* transmission, though hotspots also occur in South and Central America, the Ganges–Brahmaputra Delta, and New Guinea. These findings can inform prioritization of populations for future vaccine trials and campaigns.

**Funding:** NASA 16-GEO16-0047; NIH-NIAID 1R03AI151564-01; BMGF OPP1066146.

**Keywords:** *Shigella*; Diarrheal disease; Climate; Water, sanitation, and hygiene; Child health.
Introduction:

Diarrheal diseases remain a leading cause of childhood illness responsible for some 573,000 deaths in children under 5 in 2019 and caused by numerous infectious pathogens including bacterial, viral, and parasitic organisms. Moreover, many such enteropathogens have large reservoirs of asymptotically infected individuals, and even in the absence of clinical diarrheal symptoms, repeated subclinical episodes can still impair growth, cognitive development and vaccine response and may predispose children to chronic disease later in life. *Shigella*, a genus of gram-negative bacteria with 50 serotypes, infects over 163 million people annually in low- and middle-income countries (LMICs) and is responsible for over 212,000 diarrheal disease deaths. It is most commonly recognized via its manifestation as bacillary dysentery, with 64,000 such deaths occurring in under 5s, and an estimated case fatality of 2 per 1,000 in that age group. *Shigella* transmission is directly associated with temperature and is common in rural areas and where environmental risk factors such as poor water quality, lack of access to care, and inadequate sanitation are prevalent. Vaccine development for *Shigella* has been hampered by biotechnical and financial limitations, but there are multiple candidate vaccines now in late stages of development. It is therefore becoming increasingly important to geographically map global *Shigella* infection risk to guide and inform prospective rollout efforts towards high-priority populations.

Recent years have seen burgeoning in the field of infectious disease cartography facilitated by advances in computing, remote sensing, geostatistical methods and the availability of spatially referenced environmental data. High-resolution model-based
maps with continent-wide or quasi-global coverage have been generated, quantifying the spatial distribution of infectious pathogens including *Plasmodium falciparum* and *P. vivax*, lymphatic filariasis, and leishmaniasis as well as all-cause diarrhea. However, attempts to map *Shigella* have hitherto been limited to indirect estimation methods. One study used estimates of overall diarrhea-related mortality and morbidity and adjusted these by published pathogen-specific attributable fractions of diarrhea with *Shigella* etiology to arrive at national-level incidence and death rates. Another used a similar approach, but focused on 11 African countries to derive subnational, province-level estimates based partly on covariates ascertained through household surveys. Notably, these analyses did not attempt to account for intra-annual seasonal and temporal variation in transmission, or spatial variation below the administrative unit level. Also, in focusing on morbidity and mortality, they could not quantify the contribution of asymptomatic infections that act as reservoirs for transmission.

This study aims for the first time to: 1) model the spatiotemporal variation in childhood *Shigella* infection probability using covariates with quasi-global coverage and gain insights into pathogen transmission, risk, and seasonality: 2) use the resulting estimates to map the predicted prevalence of the pathogen sub-nationally across LMICs.

**Methods:**

**Objective and scope:**

The objective of this analysis was to estimate the percent prevalence of enteric *Shigella* infection in three separate age groups - 0-11, 12-23, and 24-59 months - and three
syndrome strata – asymptomatic, uncomplicated (community detected) diarrhea, and medically attended diarrhea – at all locations throughout the world’s LMICs (as defined in the supplementary appendix).

**Data sources and outcome variable:**

To compile a dataset representative of diverse geographical and climatic contexts, data were sourced and compiled from multiple completed studies identified through non-systematic, exploratory literature review and professional networks according to inclusion criteria and within an Independent Participant Data Meta-Analyses (IPD-MA) framework described previously and in the supplementary appendix. 29 studies contributed data from 23 countries with a range of latitudes spanning the tropics and sub-tropics, including locations in Central and South America, sub-Saharan Africa and South and Southeast Asia (figure 1). The outcome of interest was *Shigella* infection status ascertained by polymerase chain reaction (PCR) performed on diarrheal and surveillance (asymptomatic) stool samples collected from children aged ≤59 months and was treated as a binary variable (positive or negative). The *Shigella* stool positivity rate (the probability of PCR-detection) was modeled as an approximation of the prevalence of pediatric *Shigella* infection. Discrete infection episodes were defined as previously described. 29 Figure 1 shows the locations, number of samples and the proportion positive for *Shigella* at each study site, while supplementary table S1 summarizes key features of contributing studies.

**Covariates:**
Symptom status and study design: 13 (65%) of the included studies had health facility-based case-control or surveillance designs which actively recruited cases of gastrointestinal syndromes ranging in severity from uncomplicated to acute, often watery diarrhea and dysentery (bloody diarrhea). Results for diarrheal samples from symptomatic individuals were therefore overrepresented in the overall database relative to the frequency of diarrhea occurrence in the general population, which could lead to overestimation of risk for *Shigella*, a pathogen strongly associated with diarrhea and dysentery. To adjust for this, a categorical variable was included indicating whether the stool sample was collected while the child was asymptomatic or symptomatic experiencing a diarrheal episode of any severity and whether the sample was from a study with a community surveillance- or health facility-based design (four categories, representing each combination of the two symptom statuses and two study designs). This was to account for the assumed differential pathogen positivity rates in symptomatic samples and, especially, diarrhea for which facility care was sought. The inclusion of these terms allowed the predicted stool *Shigella*-positivity rate to be modeled separately for three symptom status categories: 1) asymptomatic; 2) community detected diarrhea; and 3) medically attended diarrhea (cases identified in health facility patients).

Age: Subject ages at sample collection were categorized into three age groups (0-11, 12-23, and 24-59 months) to adjust for the well-documented age-dependence of childhood *Shigella* risk and predict risk separately for age-groups commonly reported in studies of enteropathogen burden. Age, symptom status, and study design are hereafter referred to collectively as the “control variables”, since they were included in the model to control for
confounding and are the only variables for which separate predictions were made for each value, rather than letting their values vary spatially.

**Subject- and household-level covariates:** Most contributing studies conducted baseline and/or follow-up assessments of information relevant to *Shigella* transmission risk and vulnerability. These included children’s feeding and nutritional status—exclusively breastfed, fully weaned, stunting, underweight and wasting—and household characteristics—flooring material, household crowding, caregiver education, sanitation facility, open defecation and water source. These data were recoded to match as closely as possible to standardly used definitions of variables (supplementary table S2), and, where missing or not collected by some studies, imputed or interpolated based on household survey data according to methods described previously and in the supplementary appendix.  

**Environmental spatial covariates:** A suite of time-static environmental and sociodemographic spatial covariates (summarized in supplementary table S2) available in raster format were compiled based on their hypothesized or demonstrated associations with diarrheal disease outcomes. These included accessibility to cities, cropland areas, distance to major river, elevation, enhanced vegetation index, growing season length, human footprint index, irrigated areas, population density and urbanicity. Variable values were extracted at each child’s georeferenced location according to methods described in the supplementary appendix. All continuous spatial covariates were normalized and standardized to their overall distributions.
Time-varying hydrometeorological variables: A set of historical daily Earth Observation- and model-based re-analysis-derived estimates of hydrometeorological variables (supplementary table S2) were selected based on their demonstrated or hypothesized potential to influence enteric pathogen transmission, extracted from version 2.1 of the Global Land Data Assimilation System (GLDAS)⁴⁴, where appropriate, standardized to local distributions and summarized over a lagged period of exposure, using methods described previously and in the supplementary appendix.²⁹ These included precipitation, relative and specific humidity, soil moisture, solar radiation, surface pressure and runoff, temperature and wind speed.

Statistical analysis:

Model fitting: Generalized multivariate models were fitted to the binary infection status outcome to estimate predictive probabilities of Shigella positivity using the additive³⁵ and bayesian³⁶ R packages (developed for this analysis and extended as general-purpose software tools). An initial generalized linear model (GLM) including only the control variables was fitted to the outcome to serve as a reference for intercomparisons with subsequent models of increasing complexity in a repeated $k$-fold cross-validation analysis, namely: a GLM which included all non-hydrometeorological variables modeled as linear; a GLM which included all variables modeled as linear; and a final generalized additive model (GAM) with splines specified for the hydrometeorological variables to allow for the previously documented non-linearity of their relationships with Shigella.²⁹ The performance of the models was compared based on numerous in-sample and out-of-sample classification metrics (figure 2a.) using cross validation, and the importance of
variables was identified for all categories (figure 2b) using accumulated local effects (ALE)\textsuperscript{37} and the average expected marginal contribution (Shapley values).\textsuperscript{38} An interaction between air temperature and symptom status was specified, based on a hypothesis and evidence from exploratory analyses, that temperature differentially affects the probability of diarrheal samples being \textit{Shigella}-positive compared to samples from asymptomatic individuals. Similarly, an interaction between precipitation and soil moisture was specified based on the hypothesis that associations with rainfall on diarrhea-causing agents may be subject to effect modification by antecedent wetness conditions, such that the impact of extreme rainfall is increased following drier conditions.\textsuperscript{39}

\textbf{Model predictions:} The final model results were used to make separate predictions for each syndrome and age stratum using the coefficient estimates for the corresponding control variables at each combination of their values. Raster files of the values of each other covariate across the geographic extent of the domain of interest, LMICs, were obtained or generated as described in the \textit{supplementary appendix}. For the binary household- and subject-level covariates, coefficient estimates of the probability of \textit{Shigella} infection in the comparison compared to the reference category were adjusted by the geographically varying proportion coverage or prevalence stored in raster format. Seasonality was explored by extracting and plotting time series of daily predictions at six key illustrative locations – settlements in areas of high and low prevalence in LMICs in each of the three regions – Africa, Asia, and the Americas.
Analyses were carried out using Stata 16\textsuperscript{40}, R 4.0.3\textsuperscript{41} and ArcMap 10.8\textsuperscript{42} and PRISMA-IPD\textsuperscript{43} and GATHER\textsuperscript{44} guidelines were followed (see supplementary appendix tables 4 and 5).

**Results**

**Table 1** shows the distribution of stool samples in the pooled database with available *Shigella* spp. positivity status included in this analysis by age group and child’s symptom status. Almost three quarters of diagnostic results were from surveillance samples, collected from asymptomatic individuals, while of the remaining diarrheal samples, almost twice as many were collected from children presenting at health facilities than those recruited in communities. Roughly equal numbers of samples were collected from children in the first and the second years of life, while those from older children (≥2 years) made up under ten percent of the database.

**Figure 2** summarizes performance statistics of the final model (2a.) and two metrics of importance for each included variable – Shapley values and ALE – which are ranked by the average of the two (2b.). By every metric, the performance of the models improved with increasing complexity, and the final GAM that modeled the hydrometeorological variables as non-linear showed improved performance over the GLMs which treated them as linear. Child age at the time of sample collection was the most important contribution by both metrics, followed by the other control variable, symptom status/study design. The next four most important were all time-varying hydrometeorological variables (temperature, wind speed, relative humidity, and soil moisture), followed by three static environmental variables (irrigated and cropland areas...
and enhanced vegetation index). The most important household-level variable was the presence of an improved sanitation facility, which ranked 12\textsuperscript{th}, while the most important subject-level variable, stunting, ranked far down the list at 22\textsuperscript{nd}.

Figure 3 shows the probability of \textit{Shigella} infection predicted by the conditional effects of the most important hydrometeorological variables (and their interactions) in the final model (3c. and 3d.) and the odds ratios for the other variables (3a. and 3b.). Children aged 24 - 59 months had around 5-fold increased odds of \textit{Shigella} positivity (odds ratio [OR]=4.82, confidence interval [4.43-5.21]) compared to those aged <1 year, while the equivalent effect in the second year of life was a more than 3 times increased odds (OR=3.42 [3.24-3.61]). Temperature exhibited a non-linear, asymmetrical, inverse U-shaped association with the probability of \textit{Shigella} infection that was most marked for symptomatic children. Probability increased steadily across the first three quartiles of the temperature distribution, peaking at a value of around 34°C with a 47\% probability of \textit{Shigella} detection for uncomplicated diarrhea cases and 28\% for asymptomatic children, before decreasing above that threshold. An interaction was observed between precipitation and soil moisture, such that probability of positivity was low across the entire range of the precipitation distribution during very dry soil conditions, and moderate during periods of below-average rainfall, but when both variables were above average concurrently, \textit{Shigella} detection probability increased and exceeded 20\%. Wind speed also had an asymmetrical, inverse U-shaped association with probability of a child having \textit{Shigella} but skewed to the other side, with the peak risk occurring at speeds of 4m/s. Relative humidity and solar
radiation had a broadly direct effect on *Shigella* risk, and specific humidity a low magnitude inverse association with the probability of *Shigella* detection.

Living in peri-urban areas had the largest direct association with *Shigella* of the non-hydrometeorological variables, a statistically significant 18% increased odds (OR=1.18 [1.07-1.29]) compared to rural residents, though the equivalent estimate for fully urban areas was non-significant. Two anthropometric markers, were next with moderate-to-severely underweight or stunted children having a statistically significant approximate 17% (OR=1.17 [1.08-1.26]) and 16% (OR=1.16 [1.09-1.23]) increased odds of testing positive for *Shigella* respectively. The other subject-level variables, exclusive breastfeeding, full weaning and wasting had negligible, non-significant effects. The largest inverse associations among the non-hydrometeorological variables were for five of the household-level covariates. Living in a household that has an improved sanitation facility or practices open defecation both decreased the odds of *Shigella* detection by just under 20% (respectively OR=0.81 [0.76-0.86], and OR=0.82 [0.76-0.88]) compared to an unimproved facility, while having improved floor material (OR=0.83 [0.78-0.88]), a caregiver who completed primary education (OR=0.85 [0.81-0.89]), or an improved water source (OR=0.86 [0.79-0.93]) had slightly smaller protective effects. Of the remaining environmental variables, irrigated areas (OR=1.13 [1.08-1.18]), growing season length (OR=1.11 [1.05-1.17]), enhanced vegetation index (OR=1.11 [1.05-1.17]), and elevation (OR=1.08 [1.03-1.13]) were associated with a statistically significantly increased risk of *Shigella* infection whereas cropland areas (OR=0.90 [0.86-0.94]), the human footprint
index (OR=0.91 [0.85-0.97]), and distance to a major river (OR=0.94 [0.90-0.98]), were associated with a lower risk.

**Figure 4** shows the geographical distribution of the annual average prevalence of *Shigella* positivity predicted by the model for children aged 12 – 23 months in 2018 for each of the three symptom strata. Predicted prevalence of *Shigella* in asymptomatic individuals (**4a.**) varied from below 2.5% in areas of western China, Central Asia, and the Argentinian Andes to over 20% in small pockets of Central America (notably eastern Nicaragua), northern South America, tropical sub-Saharan Africa, Bangladesh, Southeast Asia (notably northeastern Cambodia), and Papua New Guinea. For individuals with uncomplicated diarrhea (**4b.**), the distribution was shifted upward such that there were large areas with predicted prevalence of higher than 25%, including almost all the territories of the Central African Republic, and South Sudan. Large areas of high prevalence were also predicted across the other countries around the African Great Lakes, coastal West Africa, Angola, Madagascar, and Bangladesh, with more delimited pockets discernable in Caribbean Central America, the interiors of Bolivia, Colombia and Venezuela and continental Southeast Asia. The predicted prevalence in individuals with medically attended diarrhea (**4c.**) exceeded 30% over almost the entirety of the tropics (with some exceptions such as parts of Borneo and the Andes) as well as large areas of subtropical Mexico, eastern China, and northeastern Argentina. Across all three symptom strata, the lowest risk was predicted over the greater Central Asia region (including Mongolia and western China), Saharan North Africa, the Andes, and Southern Africa.
Supplementary animation S1 shows the temporal variation in the predictions for each day of the year 2018, while supplementary figure S4 shows the annual time series of daily and smoothed predictions for asymptomatic children aged 12-23 months over the same year at six illustrative locations. Seasonal patterns varied. The high-prevalence Central African Republic location showed a boreal spring peak in early April, compared with the low-prevalence Botswana location where a December to February peak was evident. The high-prevalence Papua New Guinea location showed low amplitude, with only slightly higher values in the March-October period, while in Bishkek, Kyrgyzstan, a more marked mid-year peak of 11% was observed, subsiding to very low prevalence at the beginning and end of year. A mid-year and secondary end-of-year peak was discernable in high-prevalence Bluefields, Nicaragua, in contrast to La Paz, Bolivia where year-round low prevalence showed little seasonal variation. Seasonality metrics are mapped in supplementary figure S5.

Discussion:

Disease burden indicators for numerous high priority pathogens have been comprehensively mapped over entire endemic regions. However, pathogen-specific enteric infectious diseases have largely been excluded from such efforts, despite calls for innovative approaches to such estimation to inform future vaccine and other programs. This is in part due to a perceived lack of readily accessible, spatially referenced data on their detection, since enteropathogen cases are not routinely reported through health information systems or household surveys. Although newly developed multiplex molecular diagnostic platforms are increasingly being deployed in surveillance
studies to detect and differentiate an array of microbes in stool samples, studies of this nature have diverse designs and no single one offers a sufficiently broad range of geographical and environmental contexts. This study applied an IPD-MA approach to address this knowledge gap and is the first attempt to map Shigella prevalence at sub-unit level and across multiple continents. It is also the first to model the comparative effects of a large suite of covariates that vary on different scales, from the individual to the wider macroclimate, and, for several, include temporal variability at daily resolution. The data sources and analytic methods were selected to support spatial inference, resulting in a dataset of unparalleled scale and scope, and allowing for the first time to derive spatiotemporally complete, quasi-global predictions of Shigella infection rates as a function of climatic, environmental, and socio-demographic factors that can be extracted at specific locations. The findings reveal not only small-scale zones of potential elevated transmission risk, but also generalizable evidence about the relative influence of different drivers of Shigella transmission. Furthermore, the seasonal variability in infection risk within the same locations revealed by these findings may be relevant for diagnosis and treatment and for optimal timing of health system interventions.

The results confirm that Shigella risk has a strong association with age – the highest prevalence occurring in older children (24-59m) - and diarrhea symptom status, consistent with effects documented previously. While these control variables ranked highest in importance and effect size magnitude, they were closely followed by hydrometeorological variables including temperature, wind speed, relative humidity, and soil moisture. These associations exhibited considerable non-linearity and were consistent in shape, though
larger in magnitude, than those identified from an earlier version of this dataset.\textsuperscript{29} Atmospheric and hydrological factors likely impact the survival and dispersal of bacteria like \textit{Shigella} outside the host where warm and moist conditions may prolong their viability.\textsuperscript{29,57} The importance of temperature in \textit{Shigella} transmission also confirms previously reported findings. A meta-analysis found a relative risk (RR) of \textit{Shigella} diagnosis of 1.07 for each 1°C increase in temperature\textsuperscript{58}, while the equivalent for a 5°C increase in data pooled from several hundred sentinel sites across China was 1.31.\textsuperscript{59} However, we also demonstrate that \textit{Shigella} prevalence drops sharply in the upper temperature extreme above a value of around 33°C, and that the inverse U-shape of this association is more pronounced for symptomatic than asymptomatic \textit{Shigella}.

While some studies have found no association between rainfall and \textit{Shigella} infection or shigellosis\textsuperscript{59,60}, here not only was a direct effect of precipitation observed, but also an interaction with soil moisture such that prevalence was highest when both parameters were above their average values. Kraay and colleagues hypothesize that extreme rainfall increases diarrhea incidence to a greater extent if it follows a drier period.\textsuperscript{39} These findings suggest that such an effect, if true, is not mediated by \textit{Shigella}. This compounding of soil saturation conditions on precipitation is, however, consistent with findings from an interrupted time-series analysis of a large, La Niña-related flood in the study site in Loreto, Peru, where an increase in \textit{Shigella} infection was observed, but only in the later period of the flood.\textsuperscript{61}

While household factors ranked lower in importance, several statistically significant protective associations of these variables were observed. Living in a household that
practices open defecation (i.e., has no sanitation facility) conferred lower odds of Shigella infection by almost one fifth (18%), a protective association almost equal to that of having an improved sanitation facility (19%). This runs counter to the widely held perception that open defecation promotes transmission of infectious intestinal diseases, and carries the implication that, from the perspective of childhood Shigella prevention, having no toilet at all is preferable to an inadequate one (though open defecation may still be a risk factor for other diarrheal pathogens and adverse health and social outcomes). These effects sizes are smaller than those from a trial of pour-flush toilets with septic tanks in Maputo, Mozambique, which reduced the prevalence of Shigella in children under 2 years by half. While the further protective effects of caregiver primary education and covered floors (both ~15%) are consistent with those reported previously from an earlier version of this dataset, the 14% reduced odds of Shigella infection conferred by having an improved water source is modest compared with findings from Manhiça District Mozambique that water availability is among the most protective factors against Shigella infection in under 2s.

The direct association of two indicators of undernutrition, moderate-to-severe underweight and stunting, were of comparable magnitude to the protective effects of household factors, meaning that children with a weight- or height-for-age Z-score (WAZ and HAZ) 2 or more standard deviations below that of a healthy reference population were at increased risk of Shigella infection. This is further evidence of the well-documented bidirectional interplay between malnutrition and enteric infections, and consistent with findings from multi-site studies that both symptomatic and asymptomatic,
untreated Shigella infections in young children are associated with later adverse anthropometric outcomes. It is notable that feeding status was not statistically significantly associated with Shigella positivity, since early introduction of complementary foods has been shown to be associated with a 10% increased RR of infection with Shigella.\textsuperscript{13} However, it is consistent with a comparative analysis of the effect of full breastfeeding on multiple enteropathogens which found associations with some viruses and bacteria, but a non-significant association with Shigella.\textsuperscript{32}

This study had various limitations. It was only possible to directly estimate Shigella PCR-positivity in diarrheal samples and not the Shigella-specific attributable burden of diarrhea, which, due to the pervasiveness of coinfection with multiple diarrhea-causing agents, are far from equivalent, the former tending to overestimate the latter considerably. The prediction estimates for children with diarrhea provided in supplementary files can be adjusted downward using generalizable or context-specific Shigella-attributable fraction as these become available. Similarly, these prevalence estimates should not be conflated with incidence, which could be higher in settings where repeated diarrheal episodes are common even where point prevalence is low. Potential sources of error include the use of interpolation and imputation of missing subject- and household-level data due to inconsistency of covariate ascertainment across contributing studies, as well as the use of recruitment center coordinates as proxies for georeferencing children’s residence locations where the latter were unavailable (see supplementary materials). Furthermore, the relatively coarse resolution of the GLDAS estimates restricted the precision of the prevalence predictions to a horizontal resolution of around 25km\textsuperscript{2}. However, the
consistency of many major findings with those established in the prior literature give cause for confidence that any such error would be so minimal as to not bias the broad conclusions.

In conclusion, the spatiotemporal distribution of *Shigella* risk in LMICs appears highly sensitive to both macro-scale and inter-diurnal variation in temperature and other climatological factors. This makes conditions in sub-Saharan Africa particularly propitious for propagation of the bacterium. Bands of elevated prevalence were predicted to the immediate north of the Congo basin and Great Lakes region, just to the north of the Tropic of Capricorn, and in tropical coastal West Africa, consistent with the high rates reported in studies not included in this analysis, such as in the Central African Republic, Guinea Bissau, Mozambique and Madagascar. Zones of high predicted prevalence also occurred in South America, Caribbean Central America, the Ganges–Brahmaputra Delta, and New Guinea among others, and may reflect the secondary influence of static environmental and household-level risk factors. These findings should inform the design and selection of sites for forthcoming phase 3 *Shigella* vaccine trials, and populations living in these hotspots should be considered high priority for eventual vaccine rollout campaigns.

**Acknowledgements**

We would like to thank the many fieldworkers, research teams and participants of the included studies who dedicated their time and information to better the understanding of the transmission and impact of enteric infections in early childhood. GLDAS data is
disseminated as part of the mission of NASA's Earth Science Division and archived and
distributed by the Goddard Earth Sciences (GES) Data and Information Services Center
(DISC). The findings and conclusions of this report are those of the authors and do not
necessarily represent the official position of the authors’ affiliated institutions.

Author contributions

JMC, MNK and BFZ conceived of the study and obtained ethical approval. JMC and
MNK coordinated data sharing. JMC wrote the original draft. MNK and BFZ secured funding
for the research and provided review and editing. JMC and HB carried out data processing,
analysis, and visualization. NHN carried out literature review. YTC assisted with data
visualization. All other authors carried out investigation and provided data sharing
support, review, and editing.

Financial support

The research presented in this article was supported financially by NASA’s Group on
Earth Observations Work Programme (16-EO16-0047), the National Institutes of Health’s
National Institute of Allergy and Infectious Diseases grant 1R03AI151564-01, and the
Department of Internal Medicine and the Division of Infectious Diseases and International
Health of the University of Virginia. Further funding was obtained from the BMGF under
OPP1066146 to MNK. AJP is funded by Wellcome (108065/Z/15/Z). The funders played
no role in the design and implementation of the study or the analysis and interpretation of
the results.
Declaration of interests

Drs. Gaensbauer and Melgar report grants from PanTheryx, Inc, during the conduct of the study; Dr. Page reports grants from GlaxoSmithKline, during the conduct of the study; the remaining authors have no competing interests to disclose.

Ethical standards

Ethical approval for this study was given by the University of Virginia Institutional Review Board for Health Sciences Research (IRB-HSR #21544) and the Johns Hopkins University Homewood Institutional Review Board (Study# HIRB0011882). Each contributing study obtained ethical approval from their respective institutions, and written informed consent from participants’ caregivers for collection, storage and analysis of biological specimens as detailed elsewhere. All contributing authors and investigators gave their consent to analyze and publish the data.

Data Availability Statement

The epidemiological data used in this analysis contains identifiable human subject data, which cannot be disseminated under the terms of the IRB and data use agreements with contributing institutions. Investigators from contributing studies may be contacted with reasonable request for data access. Data from the MAL-ED and GEMS studies are available from the ClinEpiDB website. GLDAS data is disseminated as part of the mission of NASA’s Earth Science Division and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC). Other input data is available from the sources cited in the supplementary appendix. Processed versions in raster/TIFF
format are available upon reasonable request to the corresponding author as is statistical source code. Output model predictions of probabilities and standard errors are available for each of nine age groups/symptom stratum combinations at the following GitHub repository https://github.com/joshcolston/Badr_Shogella_predictions.
References:


34 Rodell M, Houser PR, Jambor U, et al. The Global Land Data Assimilation System. http://dx.doi.org/10.1101/2022.08.04.22277641; this version posted August 5, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

36  Badr HS. Bindings for Bayesian TidyModels. 2021

37  Apley DW, Zhu J. Visualizing the Effects of Predictor Variables in Black Box
    Supervised Learning Models. 2019; published online Aug 19.
    DOI:10.48550/arXiv.1612.08468.

38  Štrumbelj E, Kononenko I. Explaining prediction models and individual predictions

39  Kraay ANM, Man O, Levy MC, Levy K, Ionides E, Eisenberg JNS. Understanding the
    impact of rainfall on diarrhea: Testing the concentration-dilution hypothesis using a
    systematic review and meta-analysis. Environmental Health Perspectives 2020; 128:
    126001-1–126001–16.


    Review and Meta-Analyses of individual participant data: the PRISMA-IPD Statement.


    burden of Plasmodium vivax, 2000–17: a spatial and temporal modelling study. The

    DOI:10.1038/s41586-019-1200-9.

47  Chammartin F, Scholte RG, Guimarães LH, Tanner M, Utzinger J, Vounatsou P. Soil-
    transmitted helminth infection in South America: a systematic review and geostatistical

    Neglected Tropical Diseases 2015; 9. DOI:10.1371/journal.pntd.0003756.


63 Saleem M, Burdett T, Heaslip V. Health and social impacts of open defecation on women: a systematic review. *BMC Public Health* 2019; **19**: 158.


# Tables

**Table 1:** Number (and percent) of stool samples with available *Shigella* spp. positivity status used in this analysis by child’s age and diarrhea symptom status. (Community detected diarrhea – diarrhea cases identified in community-based surveillance. Medically attended diarrhea – diarrhea cases identified in health facility-based studies).

<table>
<thead>
<tr>
<th>Age group</th>
<th>Diarrhea symptom status</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asymptomatic</td>
<td>Diarrheal symptoms</td>
</tr>
<tr>
<td></td>
<td>Community detected</td>
<td>Medically attended</td>
</tr>
<tr>
<td>0 - 11 months</td>
<td>23,013 (34.6)</td>
<td>3,190 (4.8)</td>
</tr>
<tr>
<td>12 - 23 months</td>
<td>22,753 (34.2)</td>
<td>2,577 (3.9)</td>
</tr>
<tr>
<td>24 - 59 months</td>
<td>3,514 (5.3)</td>
<td>328 (0.5)</td>
</tr>
<tr>
<td>Total</td>
<td>49,280 (74.0)</td>
<td>6,095 (9.2)</td>
</tr>
</tbody>
</table>


Figures:

**Figure 1:** Locations of the sites and designs of the studies contributing data to this analysis, number of samples from each site included in the analysis and *Shigella*-positivity rates.
**Figure 2:** Variable importance plot for variables included in the full and final models

- **Balanced Accuracy**
  - Child's age
  - Symptom status / study design
  - Air temperature
  - Wind speed
  - Relative humidity
  - Soil moisture
  - Irrigated areas
  - Cropland areas
  - Enhanced vegetation index
  - Specific humidity
  - Growing season length
  - Sanitation facility
  - Human footprint index
  - Elevation
  - Solar radiation
  - Floor material
  - Surface runoff
  - Caregiver education
  - Open defecation
  - Urbanicity
  - Distance to major river
  - Stunting
  - Underweight
  - Accessibility to cities
  - Water source
  - Precipitation deviations
  - Exclusive breastfeeding
  - Household crowding
  - Population density
  - Weaning
  - Wasting
  - Surface pressure deviations

- **Recall**
  - Control GLM
  - Full GLM
  - Static GLM
  - Final GAM

- **ROC AUC**

- **Sensitivity**

- **Specificity**

**Metric (%)**

- 76
- 75
- 74
- 73

- 72
- 71
- 70
- 69

- 68
- 67
- 66
- 65

The figure shows the variable importance plot for variables included in the full and final models. The metrics used are balanced accuracy, recall, ROC AUC, sensitivity, and specificity. The variables are color-coded according to their category and method of analysis.
Figure 3: Probabilities and odds ratios of *Shigella* infection predicted by the conditional effects of each variable in the final model. The reference category for age is the 0 – 11-month group and the coefficients for the diarrhea categories are in comparison to asymptomatic children from community surveillance studies. Odds ratios for continuous static environmental variables are for one standard deviation increase. Precipitation deviations are relative to long term (2005-2019) site-specific averages.
**Figure 4:** Geographical distribution of the annual average prevalence of *Shigella* infection in children aged 12 – 23 months predicted by the final model in a). Asymptomatic individuals; b). Community detected diarrhea cases; c). Medically attended diarrhea cases.