Environmental Research 165 (2018) 91–109

Contents lists available at ScienceDirect

Environmental Research

journal homepage: www.elsevier.com/locate/envres

Evaluating meteorological data from weather stations, and from satellites and global models for a multi-site epidemiological study

Josh M. Colston[a,*], Tahmeed Ahmed[b], Cloupas Mahopo[c], Gagandeep Kang[d], Margaret Kosek[a], Francisco de Sousa Junior[e], Prakash Sunder Shrestha[f], Erling Svensen[g], Ali Turab[h], Benjamin Zaitchik[i], The MAL-ED Network

[a] Department of International Health, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[b] Nutrition & Clinical Services Division, International Centre for Diarrhoeal Disease Research, Bangladesh (ICDDR,B), Dhaka, Bangladesh
[c] Department of Nutrition, University of Venda, South Africa
[d] Christian Medical College, Vellore, India
[e] Universidade Federal do Ceará, Brazil
[f] Department of Child Health, Institute of Medicine of Tribhuvan University, Kathmandu, Nepal
[g] Haukeland University Hospital, Norway
[h] Research and Development, Maternal and Child Health (MCH) Program, Karachi, Pakistan
[i] Department of Earth and Planetary Sciences, Johns Hopkins Krieger School of Arts and Sciences, Baltimore, MD, USA

ARTICLE INFO

Keywords:
Meteorological data
Environmental epidemiology
Climate
Earth Observation data
Rotavirus

ABSTRACT

Background: Longitudinal and time series analyses are needed to characterize the associations between hydrometeorological parameters and health outcomes. Earth Observation (EO) climate data products derived from satellites and global model-based reanalysis have the potential to be used as surrogates in situations and locations where weather-station based observations are inadequate or incomplete. However, these products often lack direct evaluation at specific sites of epidemiological interest.

Methods: Standard evaluation metrics of correlation, agreement, bias and error were applied to a set of ten hydrometeorological variables extracted from two quasi-global, commonly used climate data products – the Global Land Data Assimilation System (GLDAS) and Climate Hazards Group InfraRed Precipitation with Stations (CHIRPS) - to evaluate their performance relative to weather-station derived estimates at the specific geographic locations of the eight sites in a multi-site cohort study. These metrics were calculated for both daily estimates and 7-day averages and for a rotavirus-peak-season subset. Then the variables from the two sources were each used as predictors in longitudinal regression models to test their association with rotavirus infection in the cohort after adjusting for covariates.

Results: The availability and completeness of station-based validation data varied depending on the variable and study site. The performance of the two gridded climate models varied considerably within the same location and for the same variable across locations, according to different evaluation criteria and for the peak-season compared to the full dataset in ways that showed no obvious pattern. They also differed in the statistical significance of their association with the rotavirus outcome. For some variables, the station-based records showed a strong association while the EO-derived estimates showed none, while for others, the opposite was true.

Conclusion: Researchers wishing to utilize publicly available climate data – whether EO-derived or station based - are advised to recognize their specific limitations both in the analysis and the interpretation of the results. Epidemiologists engaged in prospective research into environmentally driven diseases should install their own

Abbreviations: CHIRPS, Climate Hazards Group Infrared Precipitation with Stations; DD, Decimal degree; DISC, Data and Information Services Center; EID, Enteric infectious diseases; EO, Earth observation; FPR, False positive rate; GDAS, Global Data Assimilation System; GES, Goddard Earth Sciences; GLDAS, Global Land Data Assimilation System; LSM, Land Surface Model; MAL-ED, the Malnutrition & Enteric Infections: Consequences for Child Health and Development project; MBE, Mean bias error; NASA, National Aeronautics and Space Administration; NOAA, National Oceanic and Atmospheric Administration; NSE, Nash-Sutcliffe efficiency coefficient; R, Pearson's correlation coefficient; RMSE, Root mean square error; TPR, True positive rate

* Correspondence to: Department of International Health, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe Street, Baltimore, MD 21205, USA.
E-mail addresses: josh.colston@jhu.edu (J.M. Colston), tahmeed@icddrb.org (T. Ahmed), tjale.mahopo@univen.ac.za (C. Mahopo), gkang@cmcvellore.ac.in (G. Kang), mkosek@jhu.edu (M. Kosek), fjuniior@yahoo.com.br (F. de Sousa Junior), prakashsunder@hotmail.com (P.S. Shrestha), erling.svensen@helse-bergen.no (E. Svensen), turab.ali@irdresearch.org (A. Turab), zaitchik@jhu.edu (B. Zaitchik).

https://doi.org/10.1016/j.envres.2018.02.027
Received 18 October 2017; Received in revised form 7 February 2018; Accepted 19 February 2018
Available online 21 April 2018

weather monitoring stations at their study sites whenever possible, in order to circumvent the constraints of choosing between distant or incomplete station data or unverified EO estimates.

1. Introduction

Climate and weather influence population health through a number of interrelated pathways. Extreme weather events such as heatwaves, coastal floods and storm surges can both cause mortality directly and can compromise water sources and crop production, leading to widespread food and water insecurity, illness, undernutrition and other morbidities (World Health Organization, 2014). Moreover, climate is one of the primary constraints on the geographic and seasonal distribution of pollutants (Fann et al., 2016) and infectious agents (Wu et al., 2016). The growth, survival and dispersal of microorganisms and the viable range of their intermediary hosts and vectors is determined by environmental and hydrometeorological conditions (Hellberg and Chu, 2015). An increased awareness of the knowledge gaps surrounding these relationships, as well as the urgency of the climate change threat and greater understanding of its likely impact on public health has spurred calls for a research agenda to elucidate the interactions and biological mechanisms through which weather influences health (Xu et al., 2012; Rodó et al., 2013). A major barrier to this is the scarcity of empirical data linking climate and health at a sufficient level of spatiotemporal disaggregation for use in longitudinal and time series regression analyses (Kolstad and Johansson, 2011). To isolate interactions between the numerous, collinear climatic variables, quantify annual cycles and long-term trends, and incorporate lag effects, the health outcome and environmental exposure must be matched by their precise timing (Kolstad and Johansson, 2011; Hervás et al., 2014; Patel et al., 2013; Ahmed et al., 2013). Until recently, such analyses were hindered by the difficulty of accessing accurate and complete data on hydrometeorological predictors at high temporal resolution. The increased accessibility of Earth Observation (EO) climate data products – those derived from satellites and model-based reanalysis - is beginning to change this, but uptake has been slow due to a lack of interdisciplinary collaboration between the planetary sciences and public health fields (Rodó et al., 2013; Grace et al., 2015; Moore et al., 2017; Grace, 2017).

Researchers wishing to include climate variables as predictors in analyses of health outcomes generally have two options: to use either EO-derived or station-based data. The former have the advantage of completeness, both temporal and spatial. Estimates may be available at a daily or even sub-hourly resolution (Fang et al., 2009) without gaps and can be extracted for any location for which the geographical coordinates are known or a relevant geographic area can be mapped. Many also offer a larger suite of mutually consistent variables than are typically available from weather stations, and the data are often freely available to access online. Disadvantages include the wide variation in the uncertainty of the estimates (Hamm et al., 2015).

Weather conditions recorded at ground-based stations may be considered the gold standard for meteorological data, insofar as one exists, but are also subject to limitations. Lack of capacity to maintain routine record keeping may lead to significant data gaps, forcing researchers either to exclude outcome data for which no coincident exposure measures are available thus reducing statistical power, or to rely on summary measures such as moving mean values or binned aggregates, reducing variability and temporal resolution. Furthermore, weather stations are often situated in locations key to their primary uses in aviation or in monitoring weather for large population centers (i.e. cities and airports) and may be more geographically representative of some areas than others. Epidemiological surveillance sites may lie many kilometers from their nearest weather stations, distances greater than those over which localized meteorological conditions vary, introducing further error. Accessing data may be a challenge and, while the US National Oceanic and Atmospheric Administration (NOAA) offers a substantial online repository of historical data for some 9000 stations around the globe, for less well-served locations coordination with local meteorological agencies and organizations on the ground may be

Table 1
Köppen-Geiger climate classifications, precipitation and temperature patterns and other features of the locations of each MAL-ED study site (Institute for Vetinary Public Health, 2011; MAL-ED, 2015; Ahmed et al., 2014; Bessong et al., 2014; John et al., 2014; Lima et al., 2014; Mduma et al., 2014; Shrestha et al., 2014; Turab et al., 2014; Yori et al., 2014).

Site	Main Climate	Precipitation	Temperature	Topography	Geographic extent of site (km)		Altitude (m)	Distance to weather station (km)	Settlement type	Hemisphere
					North-south	East-west				
Dhaka, Bangladesh	Tropical savanna	Dry Nov - Feb, Monsoon Jun - Oct	Mar – May peak	Alluvial plain	0.8	0.6	12	4.6	Urban	Northern
Fortaleza, Brazil	Tropical monsoon	Dry Aug - Dec, rainy Jan - July	Hot year-round	Coastal	1.5	1.2	28	5.3	Urban	Southern
Vellore, India	Tropical savanna	Dry Jan - May, Monsoon Jun - Dec	Mar – Jun peak	Hilly	1.5	1.1	231	1.0	Urban	Northern
Bhaktapur, Nepal	Humid subtropical	Dry Oct- Mar, Monsoon May - Aug	Hot during Monsoon, Apr – Jun	Hilly	2.2	2.8	1317	7.5	Peri-urban	Northern
Naushero Feroze, Pakistan	Desert/Arid	Dry, short Monsoon Jul -Sep	Very hot Mar – Oct	Flat	8.4	4.6	44	21.9	Rural	Northern
Loreto, Peru	Tropical rainforest	Fully humid, year-round rain	Hot year-round	Flat	2.1	1.5	89	3.3	Rural	Southern
Venda, South Africa	Humid subtropical	Dry May - Sep, rainy Oct - Mar	Hot Sep – Feb	Hilly	3.8	12.5	657	36.9	Peri-urban	Southern
Haydom, Tanzania	Tropical savanna	Dry Jun - Sep, rainy Nov - May	Temperate year round	Hilly	3.0	5.5	1650	31.8	Rural	Southern