Epidemiology and Infection

www.cambridge.org/hyg

Short Paper

Cite this article: Brintz BJ, Colston JM, Ahmed SM, Chao DL, Zaitchik BF and Leung DT (2024). Assessment and comparison of model estimated and directly observed weather data for prediction of diarrhoea aetiology. *Epidemiology and Infection*, **152**, e122, 1–3
https://doi.org/10.1017/S0950268824001183

Received: 17 October 2023
Revised: 19 June 2024
Accepted: 07 July 2024

Keywords:
Clinical Prediction; Infectious Diarrhoea; Weather; Aetiology; Global Health

Corresponding author:
Ben J. Brintz;
Email: ben.brintz@hsc.utah.edu

Assessment and comparison of model estimated and directly observed weather data for prediction of diarrhoea aetiology

Ben J. Brintz[1], Josh M. Colston[2], Sharia M. Ahmed[3], Dennis L. Chao[4], Benjamin F. Zaitchik[5] and Daniel T. Leung[3]

[1]Division of Epidemiology, University of Utah School of Medicine, Salt Lake City, UT, USA; [2]Division of Infectious Diseases and International Health, University of Virginia School of Medicine, Charlottesville, VA, USA; [3]Division of Infectious Diseases, University of Utah School of Medicine, Salt Lake City, UT, USA; [4]Institute for Disease Modeling, Bill & Melinda Gates Foundation, Seattle, WA, USA and [5]Department of Earth and Planetary Sciences, Johns Hopkins University, Baltimore, MD, USA

Abstract

Recent advances in clinical prediction for diarrhoeal aetiology in low- and middle-income countries have revealed that the addition of weather data to clinical data improves predictive performance. However, the optimal source of weather data remains unclear. We aim to compare the use of model estimated satellite- and ground-based observational data with weather station directly observed data for the prediction of aetiology of diarrhoea. We used clinical and etiological data from a large multi-centre study of children with moderate to severe diarrhoea cases to compare their predictive performances. We show that the two sources of weather conditions perform similarly in most locations. We conclude that while model estimated data is a viable, scalable tool for public health interventions and disease prediction, given its ease of access, directly observed weather station data is likely adequate for the prediction of diarrhoeal aetiology in children in low- and middle-income countries.

Infectious diarrhoea is a significant public health concern, particularly in low-resource settings where access to clean water and sanitation is limited. The incidence of infectious diarrhoea is influenced by a variety of factors, including seasonality and weather conditions [1]. Previous studies have shown that weather conditions, such as precipitation, are strongly associated with the incidence of infectious diarrhoea in low and middle-income countries (LMICs) [2, 3]. Weather conditions in the months prior to illness have also been found to be predictive of the incidence of specific causes of diarrhoeal illness, such as rotavirus [4] and Shigella [5]. In addition, we have previously shown that directly observed (DO) data from weather stations accessed through the Global Surface Summary of the Day online repository can be used to derive the local season [6] or a moving average of recent weather conditions [7] and incorporated with clinical variables to improve prediction of the aetiology of infectious diarrhoea in children living in LMICs. These models can then be adapted to clinical decision-support tools for improved stewardship of antimicrobial and diagnostic use [8]. Gridded meteorological estimates from models using satellite- and ground-based observational data, such as that obtained from the Global Land Data Assimilation System (GLDAS), enable access to granular temperature and precipitation data over time and space that have no missing data, unlike the DO weather station data. In this analysis, we use data from a large multi-centre case–control study of children with moderate to severe diarrhoea to assess and compare the performance of the model estimated (ME) data of GLDAS to the DO weather station data of the Global Historical Climatology Network daily (GHCNd) when making predictions of individual-level diarrhoeal episode aetiology.

We tested the predictive performance of the weather data applied to the Global Enteric Multicenter Study (GEMS), which we previously used to derive and test prediction models for a viral-only aetiology versus other aetiologies of diarrhoea [6, 7, 9, 10]. Briefly, the GEMS was an observational case–control study conducted between 2007 and 2011 at healthcare facilities in 7 countries, in which 9 439 children with moderate-to-severe diarrhoea were enrolled at local healthcare centres along with one to three matched community control-children. A faecal sample was taken from each child at enrollment to identify enteropathogens and clinical information was collected. We used the quantitative real-time PCR-based (qPCR) attribution models developed by [11] in order to best characterize the cause of diarrhoea. Using, only diarrhoea cases, we defined viral aetiology as a diarrhoea episode with at least one viral pathogen with an episode-specific attributable fraction ($AFe \geq 0.5$) and no bacterial or parasitic pathogens with an episode-specific attributable fraction. Other aetiology includes any episode with at least one bacterial or

CAMBRIDGE
UNIVERSITY PRESS

parasite pathogen with an episode-specific attributable fraction. Prediction of viral attribution is clinically meaningful since it indicates that a patient would not benefit from antibiotic therapy.

We used daily aggregates of gridded meteorological estimates extracted from GLDAS, which includes 3-hourly weather information on a 25 × 25 km grid, based on the locations of GEMS hospital for each country [2, 12]. The temperature is extracted as minimum and maximum daily temperature. We averaged the daily minimum and maximum temperature in order to calculate an average daily temperature. Although, the average of the min and max is not necessarily equivalent to the daily average, prior literature has shown agreement between this approach and various other approaches used to get a daily temperature summary as well as ground-based measures [13]. We additionally extracted the total daily precipitation in millimetres.

We used the GHCNd to obtain DO average daily temperature and daily total precipitation. The GHCNd combines daily observations from over 30 different sources of climate observations, and undergoes a quality assurance process approximately weekly [14]. We utilized the *readr* package in R to directly download station data from GHCN daily during the study period. For each GEMS study site, we selected the closest weather station, based on haversine distance, that contained data during the GEMS years. For each location, missing weather data was completed using the most recent non-missing weather data prior to it from that same station. We note that there were only missing precipitation values and that missing values only followed days with no precipitation. The max number of missing precipitation values was 173 in Mali, of which the dates coincided with the dry season starting in late October.

We used logistic regression to assess the predictive performance of weather variables on viral versus other aetiology. The term t_{-d} refers to d days prior to the day of the patient visit to the health centre. We fit a separate model using the average exposures over each of the following moving windows of exposure:

$$[t_{-7}, t_{-1}], [t_{-14}, t_{-1}], [t_{-21}, t_{-1}], [t_{-28}, t_{-1}], [t_{-14}, t_{-7}],$$
$$[t_{-21}, t_{-7}], [t_{-28}, t_{-7}], [t_{-21}, t_{-14}], [t_{-28}, t_{-14}], [t_{-28}, t_{-21}].$$

Each model contained three exposure covariates: (1) temperature, (2) precipitation, and (3) their interaction.

We assessed the predictive performance of the model fits with various performance metrics using repeated cross-validation. We conducted Monte Carlo cross-validation by generating 500 random 80% training and 20% testing splits and compared model performance between GLDAS data and GHCNd data using the models with covariates calculated using the intervals described above [15]. Within each iteration, we calculated the area under the receiver operating characteristic curve (AUC), the calibration intercept and slope of the model for assessing weak calibration [16], the area under the precision-recall curve (PRAUC), and we used bootstrapping to compare the AUC between models [17]. That is, bootstrapped differences between two models' AUCs were calculated, standardized, and compared to the normal distribution. The PRAUC, though less common than the AUC, can be more pertinent when assessing the performance of predictive models for outcomes with imbalanced classes due to its emphasis on the performance when the ground truth is the positive class. In this case, viral etiologies account for approximately 1/3 of the outcomes.

With this approach, we assessed whether certain intervals of weather conditions prior to the presentation of the patient are more predictive of viral aetiology than others, including some intervals which do not start immediately prior to the day of presentation. Finally, we assessed the performance of cross-validated predictions by site. We ordered them by distance to the reporting weather station to assess whether there was a relationship between distance and the relative average AUC from models using the different data sources.

We found that when we plotted the ME and DO temperatures and the precipitation in time series by site, they were very similar. However, the Modelled estimates outperformed the directly observed data in both AUC and PRAUC at each interval explored with the best AUC from each source being 0.66 and 0.64, and the best PRAUC from each source being 0.47 and 0.44, respectively (Figure 1). The best-performing models used the averaged exposure over moving windows which include the most recent days from the health care facility visit.

When predictions from cross-validation were broken down by site, the superior predictive performance using ME over DO data was limited to The Gambia with a difference in average AUC, that is, about 0.102, or about 10%. Notably, the Gambia's clinical study site is 77 km away from the closest weather station, the fourth farthest site from its weather station. The next largest difference in average AUC between data sources by site in the 1–7 days interval was 0.038 in Mali. When we used bootstrapping within each iteration of cross-validation to compare AUCs between models driven by the two sources of data using the 1–7 days interval for all sites, we found that the median p-value in cross-validation is 0.025. This suggests that there was evidence of a difference between models fit using the two sources of data; however, when we conducted the same bootstrapping test and compared the two sources of data within sites, Gambia had a median p-value of 0.055,

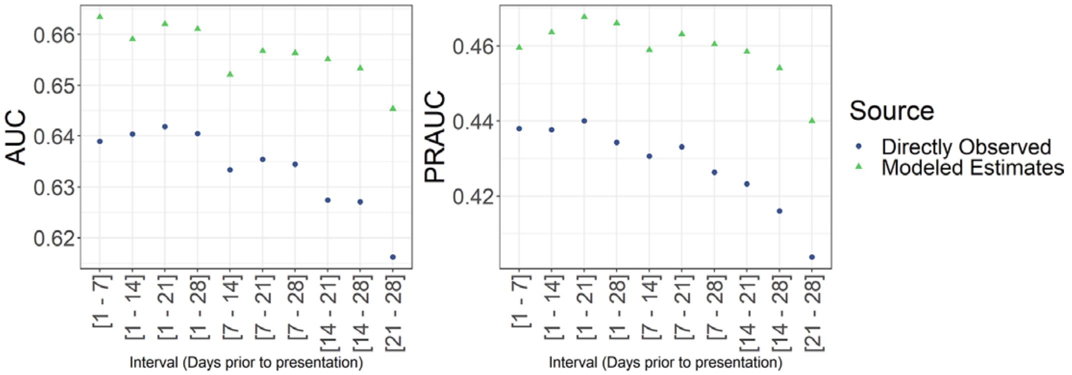

Figure 1. Average cross-validated AUC and PRAUC for model estimated and directly observed weather data averaged exposure over various moving windows.

https://doi.org/10.1017/S0950268824001183 Published online by Cambridge University Press

while the other sites had median p-values of 0.291 and above. We additionally found that all models satisfied the criteria for weak calibration, that is, on average, the models do not over- or underestimate risk and do not give overly extreme or modest risk esimates. We estimated the calibration intercepts were close to 0 with 95% CIs that contain 0 and the calibration slopes were close to 1 with 95% CIs that contain 1, satisfying the criteria.

Clinicians in low-resource settings are often required to make clinical decisions of infectious syndromes without information from laboratory diagnostics. Traditional clinical prediction rules typically focus on data from the patient at presentation, but more recently, the use of location-specific (or patient-extrinsic) data sources such as climate have shown to improve the performance of prediction rules over clinical factors alone [6, 7]. The optimal method for estimating climate and weather data for incorporation into clinical prediction rules has not been assessed. We demonstrated that on average, while DO weather station data vary in availability, there is ultimately a similar predictive performance as the ME weather data based on satellite- and ground-based observation, for the prediction of aetiology of diarrhoea in paediatric patients.

In both methods, though most prominently with ME weather, we found that the top 4 models by AUC all included temperature and precipitation data with the shortest lag length, (i.e., included the prior days' weather data in the aggregate predictors), consistent with the known short incubation periods of most enteropathogens, and previous findings observed with individual organisms [2]. The top 3 models by PRAUC also included the shortest lag.

We note some discrepancies in data preparation between the two sources of data. First, the DO data were less complete than the ME, for example, precipitation missing in periods of time where there was no precipitation. In contrast, by its nature, the model estimates had no missing data. However, when using this estimated data, advanced data storage and data cleaning will likely be required for any given application due to its granularity in raw form. Although we found evidence of an average difference between predictive performance, this difference can be attributed to differences at a single site (The Gambia). We did not find a clear linear trend in performance difference when regressed on distance from closest weather station (p-value = 0.66, Pearson correlation = 0.2). Further explorations should be made to assess if the best-performing lags from the two sources of weather data apply to other prediction outcomes, and whether our findings can be generalized to other climate factors such as humidity or environmental factors such as air pollution.

In conclusion, we found that GHCNd's DO weather station-derived data is likely adequate for the prediction of diarrhoeal aetiology in children in LMICs.

Data availability statement. R code and analytic data files are available at https://github.com/LeungLab/CompareMEandDOweather.

Author contribution. Conceptualization: D.L., J.C., B.Z.; Funding acquisition: D.L., B.B.; Methodology: D.L., J.C., B.Z., B.B.; Supervision: D.L.; Writing – review & editing: D.L., J.C., D.C., S.A., B.Z.; Resources: J.C., B.Z.; Data curation: B.B.; Writing – original draft: B.B.

Funding statement. This work was supported by the NIAID of the NIH (D.T. L., B.J.B., R01AI135114); NIH/NIAID grant K01AI168493 to J.M.C.; Translational Research: Implementation, Analysis and Design (TRIAD), with funding in part from the National Center for Advancing Translational Sciences of the National Institutes of Health (B.J.B., UM1TR004409). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

[1] **Chao DL, Roose A, Roh M, Kotloff KL and Proctor JL** (2019) The seasonality of diarrheal pathogens: A retrospective study of seven sites over three years. *PLoS Neglected Tropical Diseases* **13**, e0007211.

[2] **Colston JM, Zaitchik BF, Badr HS, Burnett E, Ali SA, Rayamajhi A**, et al. (2022) Associations between eight earth observation-derived climate variables and enteropathogen infection: An independent participant data meta-analysis of surveillance studies with broad spectrum nucleic acid diagnostics. *GeoHealth* **6**, e2021GH000452.

[3] **Horn LM, Hajat A, Sheppard L, Quinn C, Colborn J, Zermoglio MF**, et al. (2018) Association between precipitation and diarrheal disease in Mozambique. *International Journal of Environmental Research and Public Health* **15**, 709.

[4] **Colston JM, Ahmed T, Mahopo C, Kang G, Kosek M, de SJF**, et al. (2018) Evaluating meteorological data from weather stations, and from satellites and global models for a multi-site epidemiological study. *Environmental Research* **165**, 91–109.

[5] **Badr HS, Colston JM, Nguyen N-LH, Chen YT, Burnett E, Ali SA**, et al. (2023) Spatiotemporal variation in risk of shigella infection in childhood: A global risk mapping and prediction model using individual participant data. *The Lancet Global Health* **11**, e373–e384.

[6] **Brintz BJ, Howard JI, Haaland B, Platts-Mills JA, Greene T, Levine AC**, et al. (2020) Clinical predictors for etiology of acute diarrhea in children in resource-limited settings. *PLoS Neglected Tropical Diseases* **14**, e0008677.

[7] **Brintz BJ, Haaland B, Howard J, Chao DL, Proctor JL, Khan AI**, et al. (2021) A modular approach to integrating multiple data sources into real-time clinical prediction for pediatric diarrhea. *eLife* **10**, e63009. https://doi.org/10.7554/elife.63009.

[8] **Nelson EJ, Khan AI, Keita AM, Brintz BJ, Keita Y, Sanogo D**, et al. (2022) Improving antibiotic stewardship for diarrheal disease with probability-based electronic clinical decision support: A randomized crossover trial. *JAMA Pediatrics* **176**, 973–979.

[9] **Garbern SC, Nelson EJ, Nasrin S, Keita AM, Brintz BJ, Gainey M**, et al. (2022) External validation of a mobile clinical decision support system for diarrhea etiology prediction in children: A multicenter study in Bangladesh and Mali. *eLife* **11**, e72294.

[10] **Kotloff KL, Blackwelder WC, Nasrin D, Nataro JP, Farag TH, van Eijk A**, et al. (2012) The global enteric multicenter study (GEMS) of diarrheal disease in infants and young children in developing countries: Epidemiologic and clinical methods of the case/control study. *Clinical Infectious Diseases* **55**(suppl 4), S232–S245.

[11] **Liu J, Platts-Mills JA, Juma J, Kabir F, Nkeze J, Okoi C**, et al. (2016) Use of quantitative molecular diagnostic methods to identify causes of diarrhoea in children: A reanalysis of the GEMS case-control study. *The Lancet* **388**, 1291–1301.

[12] **Rodell M, Houser P, Jambor U, Gottschalck J, Mitchell K, Meng C-J**, et al. (2004) The global land data assimilation system. *Bulletin of the American Meteorological Society* **85**, 381–394.

[13] **Weiss A and Hays CJ** (2005) Calculating daily mean air temperatures by different methods: Implications from a non-linear algorithm. *Agricultural and Forest Meteorology* **128**, 57–65.

[14] **Menne MJ, Durre I, Korzeniewski B, McNeal S, Thomas K, Yin X**, et al. (2012) Global historical climatology network-daily (GHCN-daily), version 3. *NOAA National Climatic Data Center* **10**, V5D21VHZ.

[15] **Shao J** (1993) Linear model selection by cross-validation. *Journal of the American Statistical Association* **88**(422), 486–494.

[16] **Van Calster B, McLernon DJ, Van Smeden M, Wynants L and Steyerberg EW** (2019) Calibration: The achilles heel of predictive analytics. *BMC Medicine* **17**, 1–7.

[17] **DeLong ER, DeLong DM and Clarke-Pearson DL** (1988) Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach. *Biometrics* **44**, 837–845.

https://doi.org/10.1017/S0950268824001183 Published online by Cambridge University Press