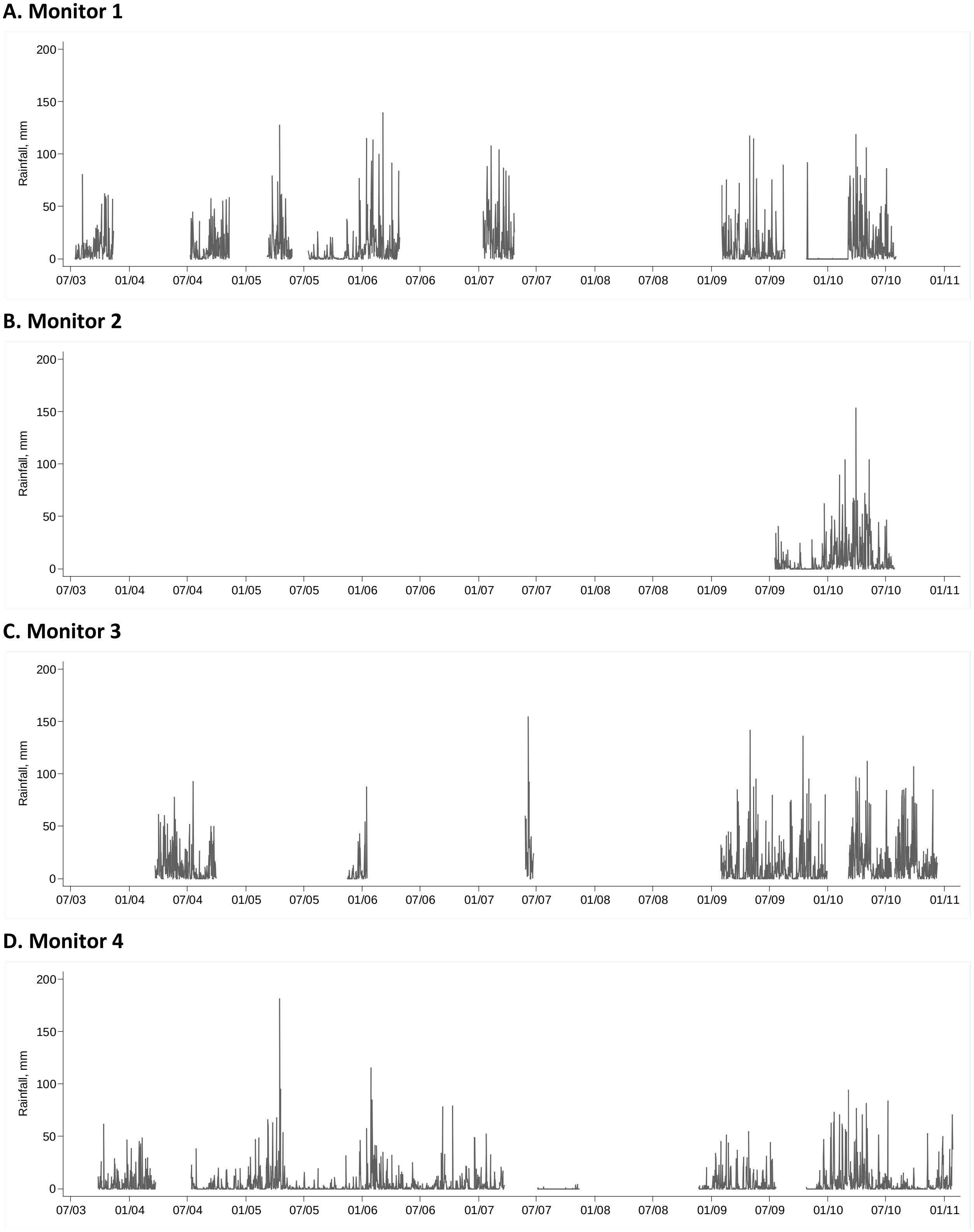
WEB APPENDIX

Social and environmental factors modify the relationship between heavy rainfall events and diarrhea incidence

Table of Contents

Web Figure 1. Daily precipitation measured using data-logging rain gages in four locations. Gaps in the line graphs indicate missing data. For a map for rain gage locations, see Figure 1 in the manuscript.



A. Monitor 1

B. Monitor 2

C. Monitor 3

D. Monitor 4

Web Appendix 1. Imputation of rainfall data

Rainfall was measured using HOBO data-logging rain gages (Onset Corporation, Borne, MA) in four locations from July 2003 through January 2011 (see Figure 1 in the main text and Web Figure 1). In order to estimate rainfall in each of the study villages each week, we used a two-step imputation procedure.

Step 1. We first used ordinary least squares regression models to estimate weekly average and weekly maximum rainfall for each weather monitoring location in weeks where rainfall measures were missing. We calculated the maximum one-day rainfall and the average rainfall at each location each week. If rainfall measurements were missing from any day during a given week, the rainfall measures for that week were treated as missing. For a given monitoring location at time t, e.g. $M_1(t)$, we estimated missing data by fitting a regression equation using available data from the other three locations, such that $M_1(t) \sim M_2(t) + M_3(t) + M_4(t)$. If data were missing from more than one location at a given time point, we use the next most complex model (e.g., $M_1(t) \sim M_2(t) + M_3(t)$). If two models were equal in terms of complexity, the model with the higher R^2 value was used. Interactions and higher order terms did not improve model fit and therefore were not included. While the study period is February 18, 2004 to April 17, 2007, data from all four loggers across the entire rainfall monitoring period were used to train the predictive model. No predictions were made for weeks when rainfall was missing at all of the four stations. This interpolation method assumes that the response variable is independent of missing status – that is, the fact that a value is missing does not indicate anything about whether it will be a “high” or “low” value.

We applied multiple regression diagnostic methods and found an approximately normal distribution of errors and a lack of correlation between predictors and residuals. These findings suggest that the linear modeling approach was appropriate. We tested the accuracy of our predictions by cross-validation (e.g. withholding some testing data where we knew the true rainfall values). We conducted 200 simulations, each simulation including a random subset of the data. The simulations showed that our linear predictions were unbiased, positively correlated with observed values (Web Table 1) and had mean squared errors (MSE) on the order of 10^{-1} for the average rainfall values and on the order of 10^{0} for maximal rainfall values (Web Table 2). Note that these MSE estimates are upper bound estimates of prediction error, since we withheld some data to make test predictions.

Web Table 1. The correlation of observed and predicted rainfall measurements at each rainfall monitoring location, estimated through cross-validation.

Rainfall Monitor	Village	Average weekly rainfall	Maximum one-day rainfall in a week
1	Village 11	0.748	0.513
2	Village 17	0.584	0.544
3	Village 21	0.585	0.514
4	Borbón	0.636	0.586

Web Table 2. The observed prediction error for each rainfall monitoring location, estimated through cross-validation.

Rainfall Monitor	Village	Average weekly rainfall	Maximum one-day rainfall in a week
1	Village 11	0.128	1.387
2	Village 17	0.122	1.109
3	Village 21	0.136	1.138
4	Borbón	0.048	0.631

The observed prediction error is the mean squared difference between predicted and observed values, calculated by withholding approximately half of the observed values for a given station and imputing the missing values.

Step 2. We used a non-parametric kriging approach to impute rainfall values in all study villages at all time points, based on the complete, imputed data from the four weather monitoring locations for the study period. We employed a spline interpolation method, because, while the formulation of spline and kriging models are different, the resulting imputed values are mathematically equivalent (1). Each rainfall observation from the four monitoring locations was viewed as a sparsely observed function that we interpolated across the study region. A separate model, with a separate smoothing parameter, was estimated for each week. In weeks where there was very little variation between stations, or very disparate values that were not related to a spatial process, the optimal level of smoothness chosen was a constant function, leading to predictions that are simply the average of the observed stations. This was the case for 61% of the predictions of average weekly rainfall.

Web Table 3. The association between heavy rainfall events and diarrhea incidence, stratified by the amount of rainfall in the past eight weeks, using three- and four-week lags.

Average 8-week rainfall	3-week lag		4-week lag	
	Adjusted[a]	Unadjusted	Adjusted[a]	Unadjusted
	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Total 8-week rainfall[b]				
Low (78 - 425 mm)	0.82 (0.56, 1.20)	0.83 (0.57, 1.22)	1.13 (0.82, 1.58)	1.12 (0.81, 1.56)
Medium (426 - 604 mm)	1.16 (0.80, 1.67)	1.15 (0.79, 1.66)	0.86 (0.57, 1.31)	0.85 (0.56, 1.30)
High (605 - 1356 mm)	0.97 (0.79, 1.20)	0.97 (0.78, 1.19)	0.82 (0.65, 1.02)	0.82 (0.65, 1.02)

A heavy rainfall event was defined as a maximum one-day rainfall in a 7-day period above the 90th percentile value (56 mm). Estimates were modeled using random-effects Poisson regression, with a random intercept for each village.

IRR – Incidence rate ratio; CI – Confidence interval

[a] Adjusted models include diarrhea incidence one-week prior and remoteness.

[b] Total rainfall during the 8-week period prior to the week used to define heavy rainfall events. Total 8-week rainfall was categorized based on the 33rd and 66th percentile values.

IRR, Incidence rate ratio; CI, Confidence interval

Web Table 4. Sensitivity analysis evaluating the association between heavy rainfall events, defined using the 80th percentile value, and diarrhea incidence.

	1-week lag		2-week lag	
	Adjusted[a]	Unadjusted	Adjusted[a]	Unadjusted
	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Total 8-week rainfall[b]				
Low (78 - 425 mm)	1.11 (0.85, 1.45)	1.09 (0.84, 1.42)	1.17 (0.91, 1.52)	1.18 (0.91, 1.53)
Moderate (426 - 604 mm)	1.01 (0.81, 1.27)	1.00 (0.80, 1.25)	0.76 (0.59, 0.98)	0.76 (0.59, 0.99)
High (605 - 1356 mm)	0.88 (0.73, 1.05)	0.86 (0.72, 1.03)	0.91 (0.77, 1.09)	0.91 (0.76, 1.08)

A heavy rainfall event was defined as a maximum one-day rainfall in a one-week period above the 80th percentile value (41.3 mm). Estimates were modeled using random-effects Poisson regression, with a random intercept for each village. Likelihood ratio test for the significance of the interaction in the adjusted models: p=0.3042 using a one-week lag and p=0.0674 using a two-week lag.

IRR – Incidence rate ratio; CI – Confidence interval.

[a] Adjusted models include diarrhea incidence one-week prior and remoteness.

[b] Total rainfall during the 8-week period prior to the week used to define heavy rainfall events. Total 8-week rainfall was categorized based on the 33rd and 66th percentile values.

Web Table 5. The association between heavy rainfall events and diarrhea incidence, at different levels of drinking water treatment and stratified by the amount of rainfall in the past eight weeks.

	Low 8-week rainfall		Moderate 8-week rainfall		High 8-week rainfall	
	IRR	(95% CI)	IRR	(95% CI)	IRR	(95% CI)
Drinking water treatment[a]						
0%	1.86	(1.27 - 2.72)	0.93	(0.56 - 1.54)	0.94	(0.70 - 1.26)
9%	1.72	(1.22 - 2.42)	0.86	(0.53 - 1.39)	0.87	(0.67 - 1.12)
26%	1.48	(1.09 - 2.01)	0.74	(0.47 - 1.17)	0.75	(0.60 - 0.94)
45%	1.26	(0.92 - 1.72)	0.63	(0.39 - 1.01)	0.63	(0.48 - 0.83)
67%	1.04	(0.70 - 1.54)	0.52	(0.30 - 0.89)	0.52	(0.36 - 0.77)

Estimates were modeled using random-effects Poisson regression, with a random intercept for each village. Models include diarrhea incidence one-week prior and remoteness.

Estimates are shown for the 10th, 25th, 50th, 75th and 90th percentile values for drinking water treatment, observed across all study villages throughout the study period.

IRR – Incidence rate ratio; CI – Confidence interval

[a] The percent of households that report treating their drinking water using filtration, boiling and/or chlorination.

Web Appendix 2. The association between rainfall and diarrhea incidence, using only observed rainfall data

We repeated our analysis of the impact of heavy rainfall events on diarrhea incidence, restricting the analysis to observed weather measurements, and the nine villages that are located within 10 km of a rainfall monitor. This analysis includes 567 village-weeks. Diarrhea incidence across all villages and weeks included in the analysis was 5.48 cases per 1,000 person-weeks.

As in the principal analysis, total rainfall in the previous eight weeks modified the association between of heavy rainfall events and diarrhea incidence using a two-week lagged model (*p*=0.0454). Extreme rainfall two weeks prior was associated with increased diarrhea incidence during periods when rainfall in the past eight weeks had been low, and decreased diarrhea incidence during periods when rainfall over the past eight weeks had been high, but there is considerable uncertainty around these estimates (Web Table 6). In contrast to the principal analysis, total rainfall in the previous eight weeks also modified the association between heavy rainfall events and diarrhea incidence using a one-week lagged model (*p*=0.0167). Heavy rainfall events were associated with reduced diarrhea incidence during periods when rainfall in the past eight weeks had been high.

Web Table 6. The association between heavy rainfall events and diarrhea incidence, stratified by the amount of rainfall in the past eight weeks, using only directly measured rainfall in villages located within 10 km of a rainfall monitor.

	1-week lag		2-week lag	
	Adjusted[a]	Unadjusted	Adjusted[a]	Unadjusted
	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)	IRR (95% CI)
Total 8-week rainfall[b]				
Low (78 - 425 mm)	1.10 (0.69, 1.76)	1.09 (0.69, 1.75)	1.45 (0.96, 2.20)	1.45 (0.96, 2.19)
Medium (426 - 604 mm)	1.08 (0.45, 2.59)	1.12 (0.47, 2.69)	0.33 (0.08, 1.38)	0.35 (0.08, 1.43)
High (605 - 1356 mm)	0.32 (0.14, 0.72)	0.32 (0.14, 0.71)	0.86 (0.47, 1.60)	0.85 (0.46, 1.56)

Estimates were modeled using random-effects Poisson regression, with a random intercept for each village.
IRR – Incidence rate ratio; CI – Confidence interval
[a] Adjusted models include diarrhea incidence one-week prior and remoteness.
[b] Total rainfall over the 8 week period starting one week before extreme rainfall was defined as low, medium and high based on the 33rd and 66th percentile values.

Reference

1. Wahba G. *Spline models for observational data*. Philadelphia: Society for Industrial and Applied Mathematics; 1990.