Investigating the impact of social and environmental extremes on cholera time varying reproduction number in Nigeria

Gina E C Charnley[1,2], Sebastian Yennan[3], Chinwe Ochu[3], Ilan Kelman[4,5,6], Katy A M Gaythorpe[1,2], Kris A Murray[1,2,7]

1. *Department of Infectious Disease Epidemiology, School of Public Health, Imperial College London, London, UK*
2. *MRC Centre for Global Infectious Disease Analysis, School of Public Health, Imperial College London, London, UK*
3. *Surveillance and Epidemiology Department/IM Cholera, Nigeria Centre for Disease Control, Abuja, Nigeria*
4. *Institute for Risk and Disaster Reduction, University College London, London, UK*
5. *Institute for Global Health, University College London, London, UK*
6. *University of Agder, Kristiansand, Norway*
7. *MRC Unit The Gambia at London School of Hygiene and Tropical Medicine, Fajara, The Gamiba*

Correspondence to: Gina E C Charnley, g.charnley19@imperial.ac.uk

https://orcid.org/0000-0003-2087-7822

Abstract

Cholera is reported as endemic in more than 50 countries, many of which are in sub-Saharan Africa. Nigeria currently reports the second highest number of cases, with several risk factors potentially contributing to this including poverty, water, sanitation and climate. Enteric pathogens have a significant global burden, especially on children and those most vulnerable. Despite this, attention is often drawn away from these diseases, most recently to Ebola and COVID-19. To address the need for more research and focus on cholera, a covariate selection process and

machine learning was used. Data for environmental (floods, droughts) and social (conflicts) extremes, along with pre-existing social vulnerabilities, were fit to time varying reproductive number in Nigeria. We analysed this both spatially and temporally and used it to create a traffic-light system for cholera transmission, highlighting potential thresholds and triggers for outbreak. Improved access to sanitation, number of monthly conflict events, Multidimensional Poverty Index and Palmers Drought Severity Index were retained in the best fit model. Varying exposure periods showed that those living in decreased poverty, with more access to sanitation were not as vulnerable to changes and offset some of the cholera risk caused by extremes. The work presented here shows the need to address these pre-existing vulnerabilities and sustainable development for disaster prevention and mitigation and improve health and quality of life.

Introduction

The seventh and continuing cholera pandemic has far reaching implications, with 3 to 5 million reported annual cases and 120,000 deaths, in more than 50 endemic countries. Many of these cases and deaths are reported in low- and middle-income countries and it is the second leading cause of mortality in children under 5[1]. The short incubation period of cholera (2 hours to 5 days) and high numbers of asymptomatic infections which can still contaminate the environment, means explosive outbreaks can occur which can spread rapidly and cause high mortality rates (>1%)[2].

Since the re-introduction of cholera into Africa in the 1970s, the disease has become endemic in several countries, with some countries reporting the highest burdens of cholera globally. Nigeria currently reports the second highest number of estimated cases in Africa[1,3] and has seen several large reported outbreaks, especially in the northeastern region[4-7]. There are several potential reasons for Nigeria's high cholera burden including its climate[8,9], access to water, sanitation and hygiene (WASH)[10,11] and the proportion of the population living in poverty (62% at <$1.25/day)[12-14]. It also has a robust reporting system, meaning more cases are likely to be reported compared to countries with weaker surveillance systems[15].

Cholera is a preventable disease through wide-spread access to safe drinking water and sanitation and is considered a disease of inequity[16]. Improved access to sanitation and water are key determinants of global health and 73% of the enteric disease burden in Nigeria is associated with inadequate WASH[17]. These pre-existing vulnerabilities can be worsened in times of environmental and social extremes, which can act as a catalyst for outbreaks. Previous research has found several links between extreme events and cholera including floods, drought and conflict[18-20], all of which have impacted Nigeria in recent decades. Risk factors leading to disease outbreaks include an inability to access routine care, fears over safety, disruption of WASH services and human displacement[21,22].

Despite accounting for a large proportion of cholera cases, Africa is a chronically understudied continent compared to the Indian subcontinent and Haiti. Other disease outbreaks have drawn attention away from cholera in Africa in recent years, including COVID-19 and Ebola[23,24]. Research linking several social and environmental extremes to diseases is a global research gap. This is important in terms of cholera transmission which is often the result of a range of risk factors and cascades[25]. We aim to build a more cohesive understanding of how these conditions result in outbreaks and understand potential risk factors for these outbreaks through a detailed sub-national analysis.

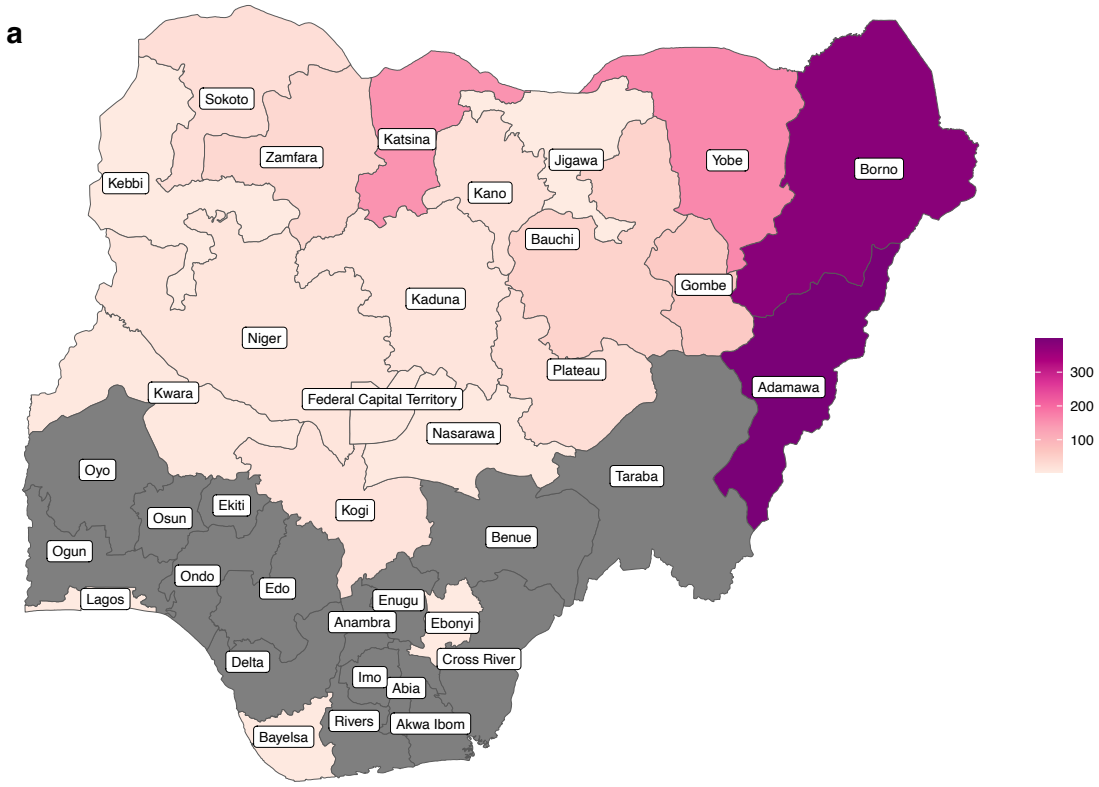
Previous research on disaster-related infectious disease outbreaks have examined disasters in isolation[19,22], while others do not include multiple pre-existing socio-economic factors into the methodology[26,27]. We aim to fill this research gap by evaluating the effects of multiple disaster types on cholera outbreaks, taking into context population vulnerability. Using data from the Nigeria Centre for Disease Control (NCDC), we evaluated a range of environmental and social covariates and fit the data using machine learning to cholera time-varying reproductive number (R). Using the model with the best predictive power, we created a traffic light system of cholera risk, in terms of multiple disasters and pre-existing vulnerabilities. No previous research has created such framework for cholera and we hope the simplicity of the traffic light approach means that our

results can be used by a wide range of scientific disciplines and organisations to reduce cholera risk in fragile settings.

Results

Incidence and R

In Nigeria, there were 837 and 564 cholera cases for 2018 and 2019, respectively. These were confirmed by either rapid diagnostic tests or culture. The confirmed cases are shown spatially in Fig. 1a and are concentrated in the northeast of the country, with Adamawa, Borno, Katsina and Yobe having the highest burden. Younger populations had much higher numbers of cases, which decreased through the age groups until age 45, over which saw a small increase. Cases by sex were relatively similar with more men affected in 2018 (51.6% male) and more women in 2019 (43.6% male) (Fig. 1b).



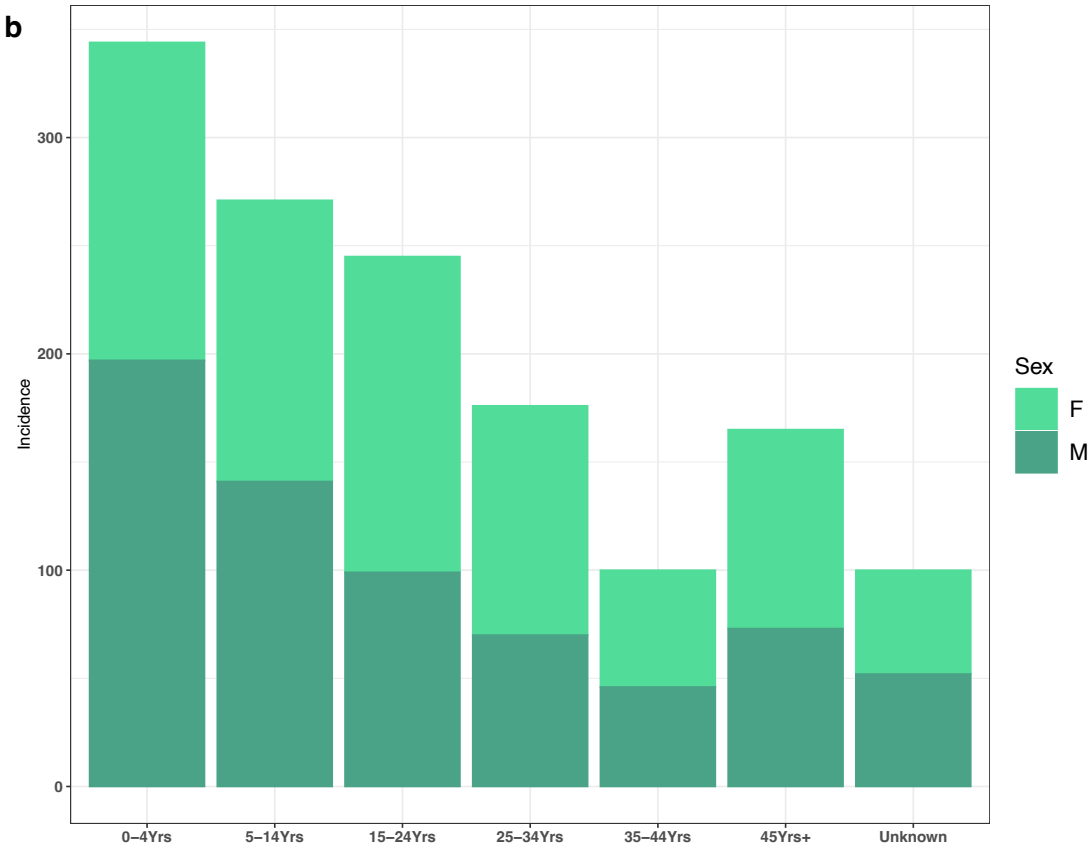


Fig. 1: Number of confirmed cholera cases. a, by state, grey indicates states that had no reported confirmed cases and **b**, by sex and age group, all for 2018 and 2019.

Six states for 2018 and two states for 2019 met the 40 case criteria to be included for R calculations, including Adamawa (2018 & 2019), Bauchi (2018), Borno (2018 & 2019), Gombe (2018), Katsina (2018) and Yobe (2018). Both the R values and the incidence used to calculate R are shown temporally in Fig. 2 for each state and year. Some states appear to have a peak in transmission around June-July, whereas others appear later during September to October.

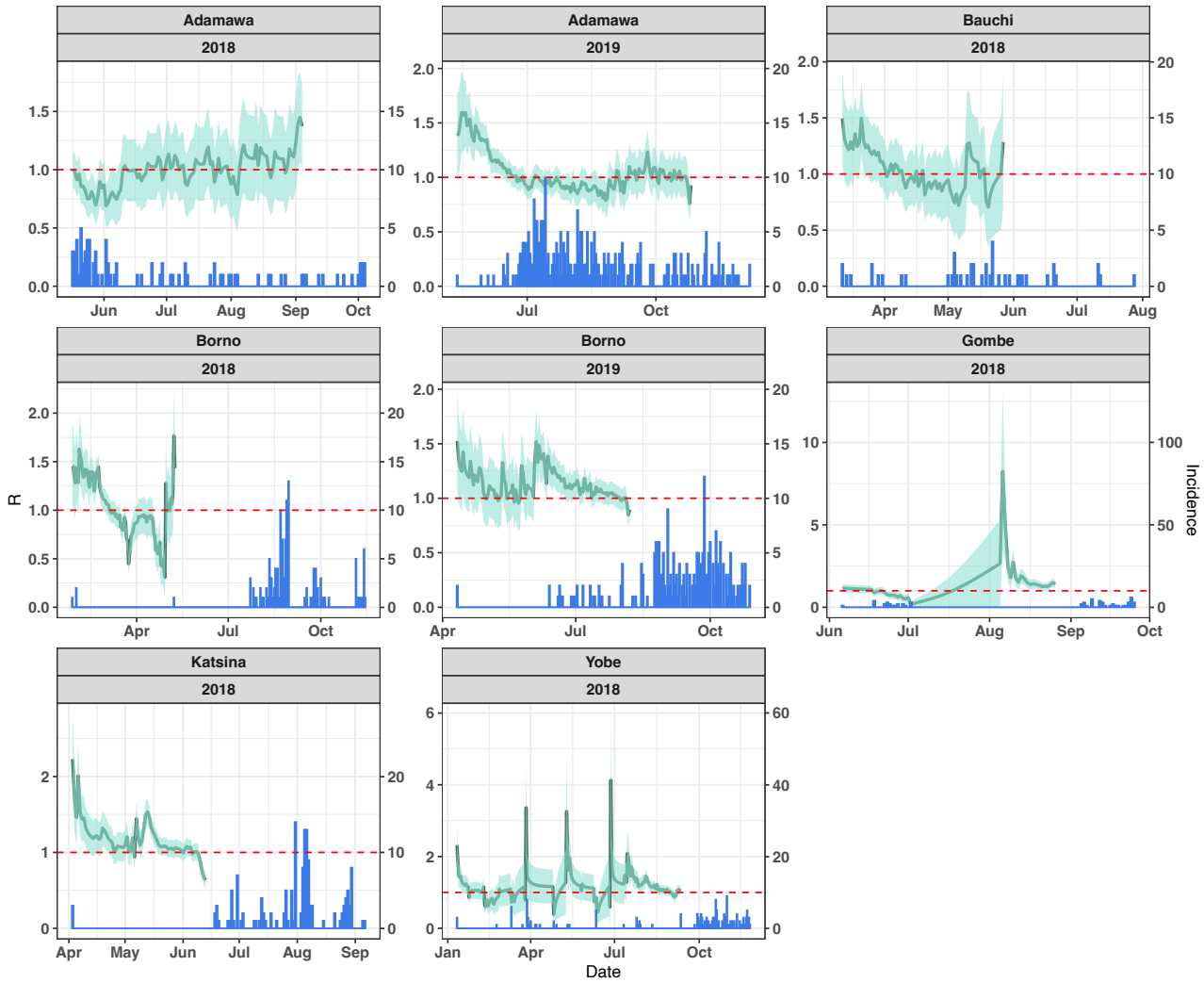


Fig. 2: R values (line) calculated from the incidence (bar) of cholera. Data is for the confirmed cholera cases for 2018 and 2019 of states which met the threshold equal to or more than 40 cases. The R values are not present for the full timescale of incidence values due to be calculated on monthly sliding windows.

Covariate Selection and Random Forest Models

Nineteen covariates were included in the covariate selection process. Five were removed, either because they were not significantly associated with the outcome variable (R) or because they were too highly correlated with other covariates (healthcare facilities, piped water, open defecation, population, internally displaced populations (IDPs). Eleven covariates remained and were grouped into five clusters, the clusters and variable importance (based on reducing node impurity) of each covariate are shown in Fig. 3.

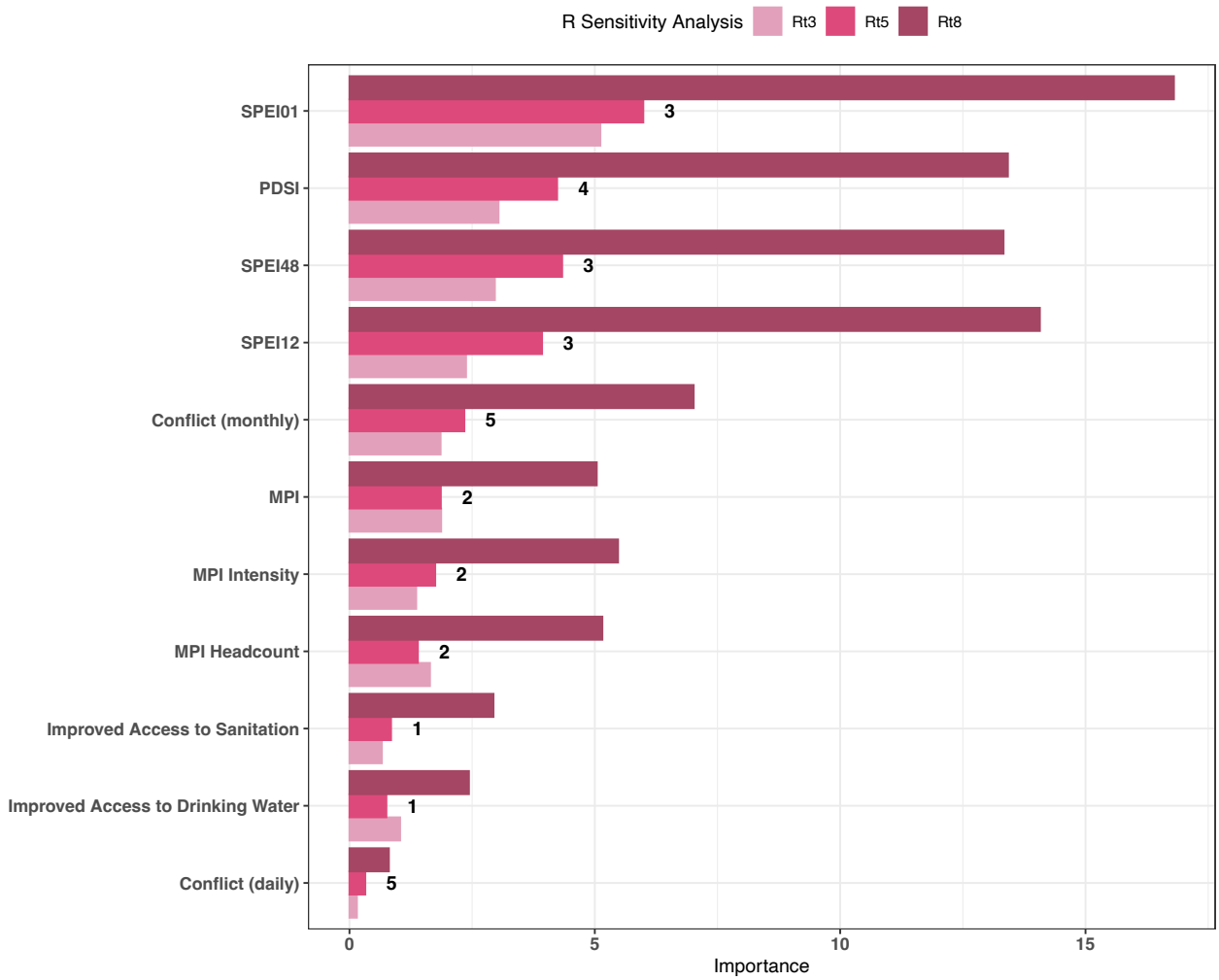


Fig. 3: The variable importance for the eleven remaining covariates after variable selection. All three serial interval values tested are shown (Rt3 - 3 days, Rt5 - 5 days, Rt8 - 8 days) and the numbers represent the clusters. SPEI01, 12, 48 - Standardised Precipitation Index calculated on 1, 12 and 48 month scale. PDSI - Palmers Drought Severity Index. MPI - Multidimensional Poverty Index.

Stepping through the model possibilities using hierarchical stepwise analysis, the best fit model was found according to mean standard error of the residuals (RMSE), R^2 and correlations between the incidence-based and covariate-based R values. The model included number of monthly conflict events, Multidimensional Poverty Index (MPI), Palmers Drought Severity Index (PDSI) and improved access to sanitation, fitted to R values with a serial interval of 5 days (standard deviation: 8 days). The fit of the incidence-based vs covariate-based R values are shown below, along with the measures of model performance used in the hierarchical analysis (Fig. 4).

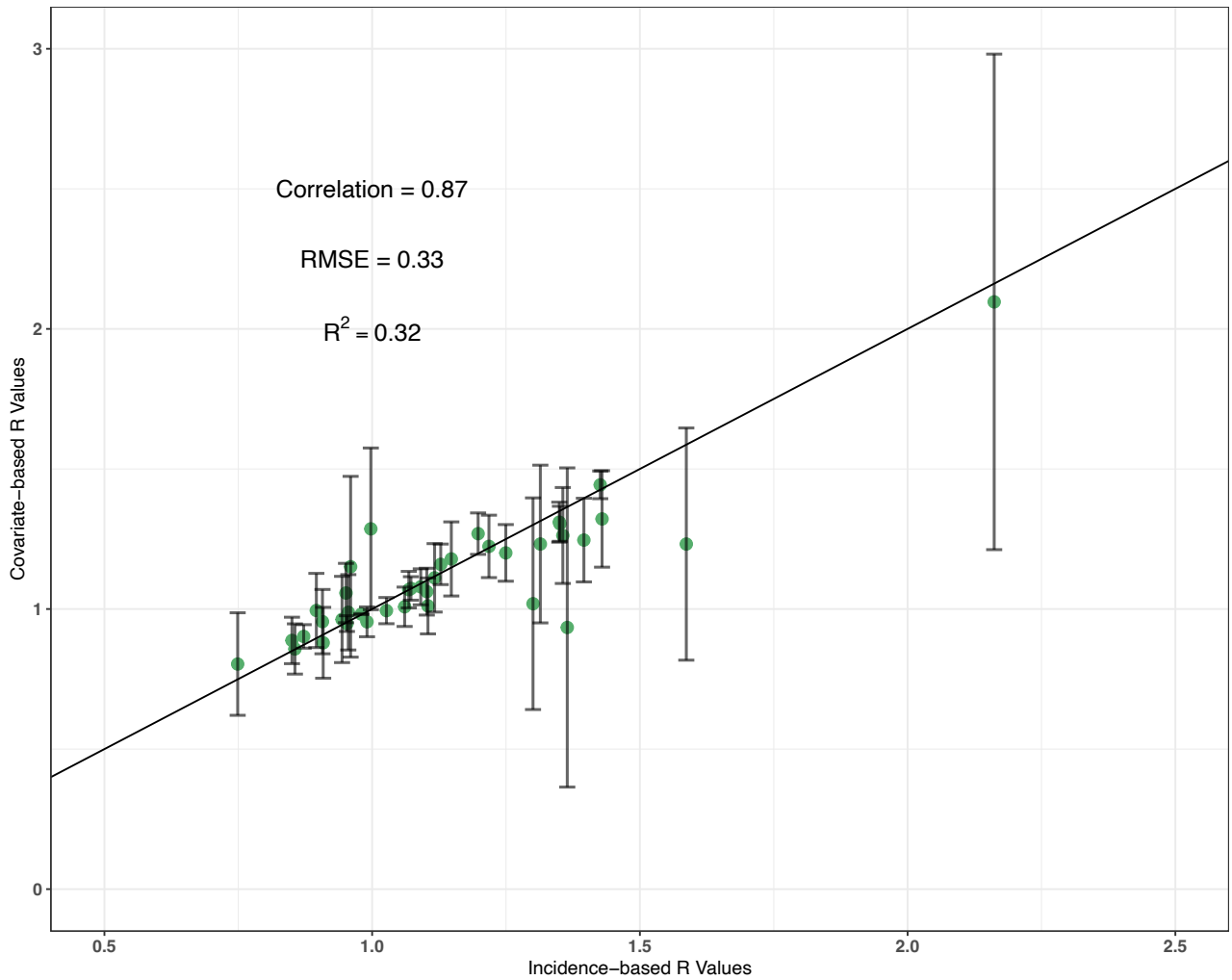


Fig. 4: Incidence-based vs covariate-based R values for the best fit model fitted to the testing dataset. The error bars show mean absolute error and the line is a linear trend line intercepting at 0.

Nowcasting

Using the best fit model, R was predicted for the remaining 31 states which did not meet the 40-case threshold and the remaining dates for the six states which were included. This created estimations of R for all 37 states on a monthly temporal scale for 2018 and 2019. The predictions suggest that the model fits well and accurately predicts R, as the higher R values are in areas with known elevated cholera burden (northern and northeastern regions) and the states which only marginally fell below the threshold for R calculations (Fig. 5).

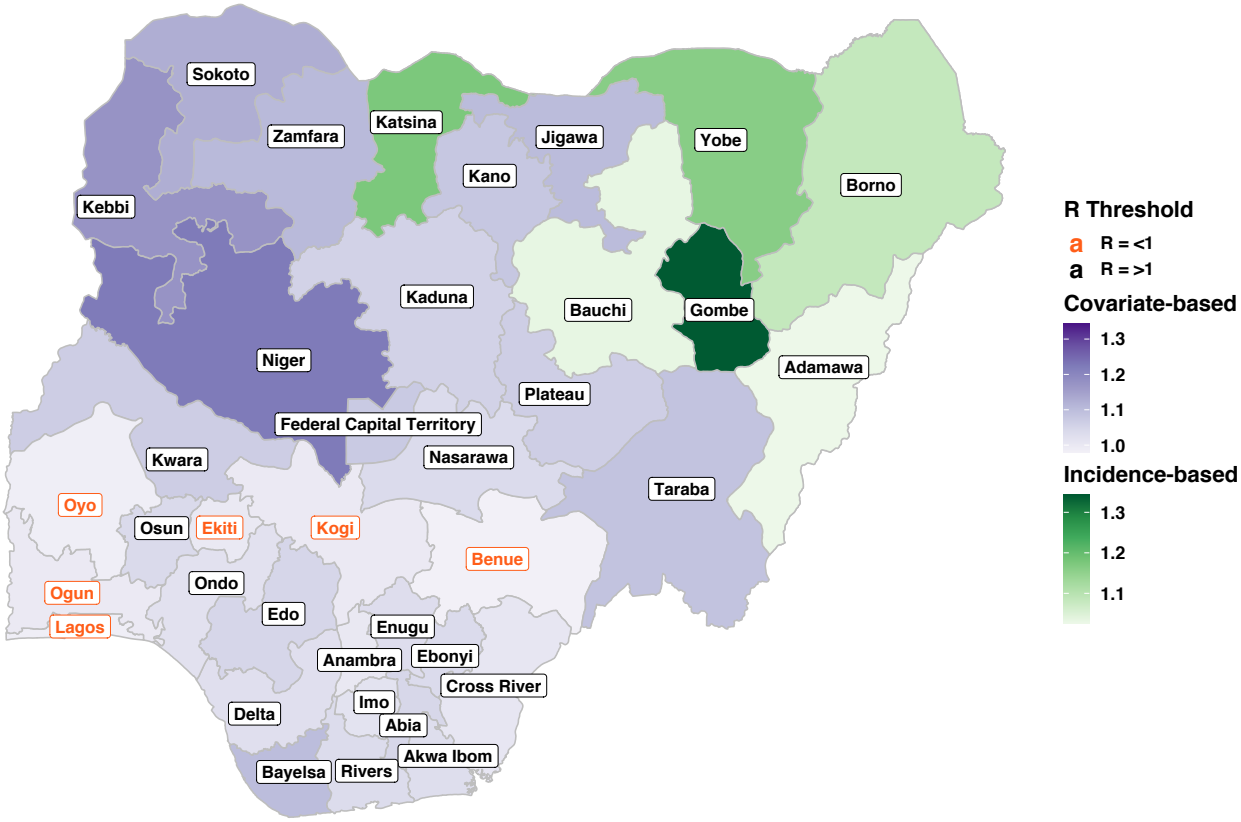


Fig. 5: Average R values for 2018 and 2019 for all 37 Nigerian states. Incidence-based (green) - the five states which met the equal to or more than 40 case thresholds. Covariate-based (purple) - the 31 states which did not meet the threshold and had R predicted using the best fit model. State label colour shows which states had an average R of R = >1 (black) and R = <1 (orange).

Historical Exposure Periods

The 48 historical exposure periods by month and R threshold are shown in Fig. 6. Sanitation and MPI have a clear relationship with the R threshold, with consistently lower MPI (less poverty) and a higher proportion of people with access to sanitation seeing lower R values. There is marginal overlap between the standard error of the two R thresholds. The average sanitation level for R = >1 was 52.8%, whereas for R = <1 it was 61.2%, for MPI the mean values were 0.27 and 0.13 for R =

>1 and R = <1, respectively. In contrast, monthly conflict events and PDSI shows a less defined relationship with some months showing either a positive or negative association.

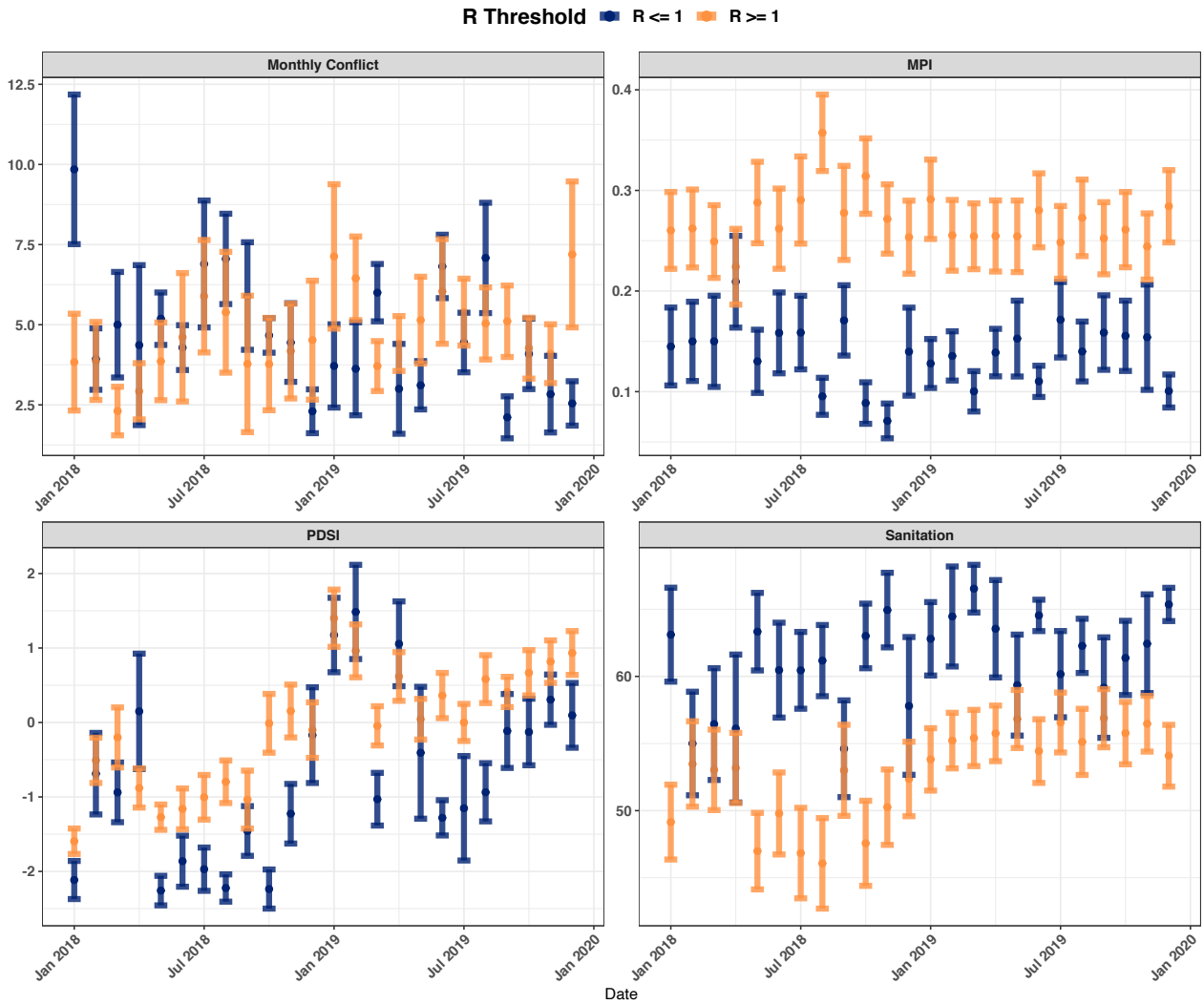


Fig. 6: Historical exposure periods. The mean and standard error for the four covariates included in the best fit model for each 48 historical exposure periods (24 months) and each R threshold (R = >1 and R = <1).

Hypothetical Exposure Periods and Spatial Heterogeneities

Based on the historical median and standard error values as a starting point for the four covariates, three hypothetical exposure periods were created: red (R = >1 values), amber (mid-point between green and red) and green (R = <1 values). R appeared to raise above 1 around 50% or lower for improved sanitation access and MPI values of above 0.32. For PDSI and conflict, R values increased above 1 at around -1.1 for PDSI and monthly conflict events of 1.6 (Fig. 7). The less

defined relationship between conflict and PDSI and R meant further investigation was needed to understand possible spatial differences.

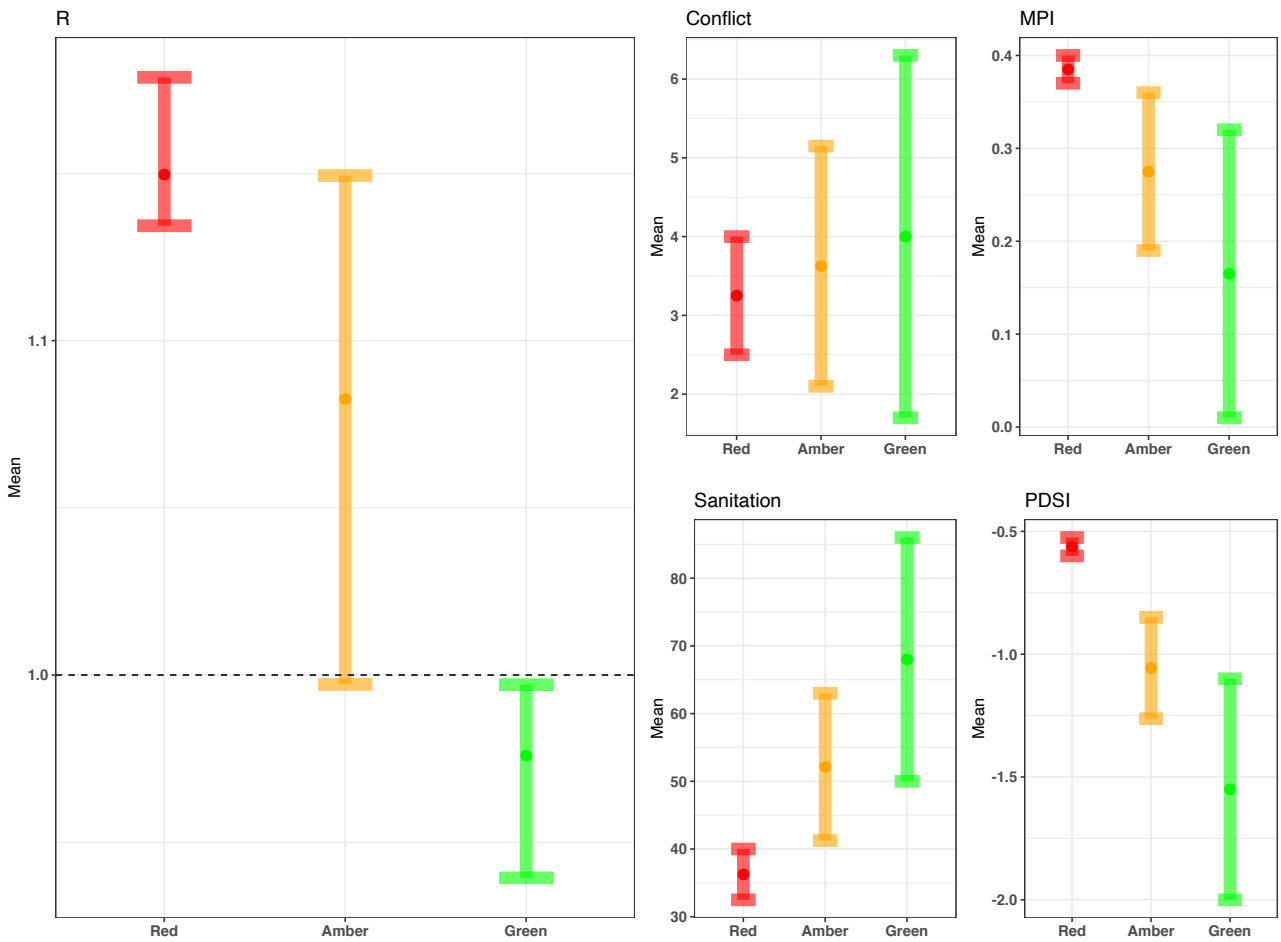


Fig. 7: Traffic-light system of cholera risk. Mean and range values for the three hypothetical exposure periods (red, amber and green) for each of the four covariates in the best fit model and the corresponding predicated R value means and range using the best fit model.

To investigate this potential spatial heterogeneity the data were split for each R threshold by state for the full dataset (2018-2019) (Fig. 8). This showed significant spatial heterogeneity among the two thresholds by state for monthly conflict events and PDSI. Borno and Kaduna were investigated further for their clear relationship with conflict (more conflict = more cholera transmission). Furthermore, Kwara and Nasarawa had additional analysis due to their relationship between extreme dryness and higher R values and Ekiti and Lagos for extreme wetness and higher R values.

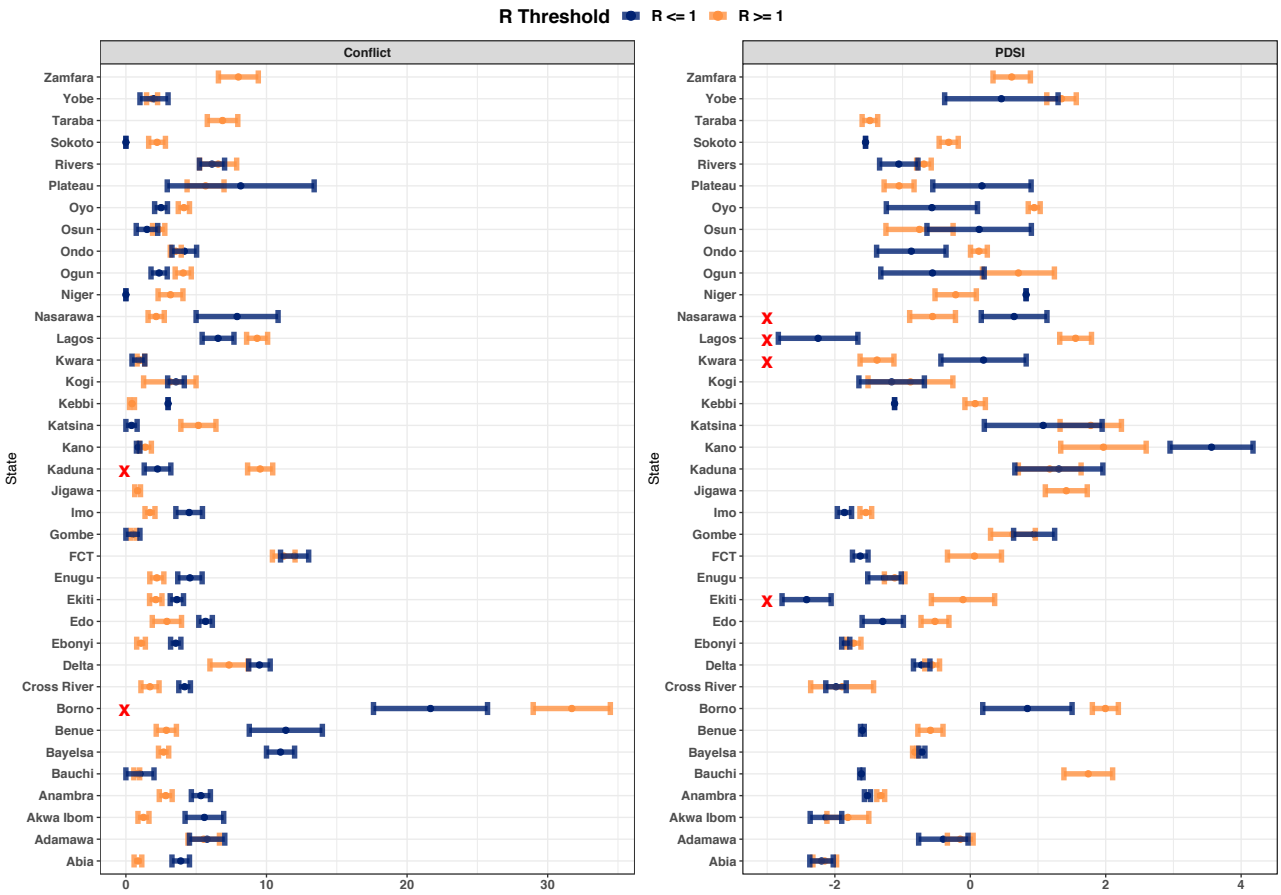


Fig. 8: Spatial differences for PDSI and conflict. The mean and standard error for PDSI and monthly conflict events by state for the two R threshold (R = >1 and R = <1). The red "x" shows the states which were included in the sub-national analysis: Conflict (Borno and Kaduna), extreme wetness (Lagos and Ekiti), extreme dryness (Nasarawa and Kwara).

For Borno and Kaduna, new exposure periods were created for conflict only. These were very small for the green exposure period, with only a low narrow range of conflict values causing R values less than 1. Both Kaduna and Borno have high levels of poverty and low access to sanitation (40-41% access). For Borno, raising monthly conflict events from 1 to 2 increased R above 1, but an increase in access to sanitation from 41-46% pushed the R value back below one. This relationship continued in a stepwise pattern and in a similarly way for MPI but to a lesser degree of magnitude. This showed that increasing sanitation and therefore decreasing vulnerability, allowed the states to adapt to increasing conflict and keep the R value below 1 (See Supplementary Figure 2).

For the four states investigating the differences between extreme wetness (Lagos and Ekiti) and extreme dryness (Nasarawa and Kwara) and R values, the analysis and subsequent PDSI hypothetical exposure periods yielded similar results. All four states found low predicted R values and higher ranges (Supplemental Figs. 3 & 4). A potential explanation for this is the high variable importance of PDSI (Fig. 3) and the high levels of sanitation and low levels of poverty in all four states contribute to overall lower levels of cholera. Therefore, the model was detecting a signal in only small changes in PDSI, that resulted in changing R values which have not been detected in other states with higher rates of poverty and lower levels of sanitation access. It also helps to highlight the multi-directionality of the relationship between PDSI and cholera transmission, with both extreme wetness and extreme dryness causing increases in R.

Discussion

The results presented here show the importance of social and environmental extremes on cholera outbreaks in Nigeria, along with the importance of underlying vulnerability and socioeconomic factors. Of the 1,401 positive cases for Nigeria in 2018 and 2019, the northeast of the country and children under 5 carried the highest burden of disease, whereas there was minimal differentiation in cases between sex. Six states were used to calculate the R values, including Adamawa, Bauchi, Borno, Gombe, Katsina and Yobe. After covariate selection, a stepwise hierarchical analysis was used to find the best fit model according to the selected model performance measures and included monthly conflict events, percentage of the population with access to sanitation, MPI and PDSI. Using the best fit model, nowcasting was used to calculate the R values for the remaining thirty-one states which did not meet the threshold.

Both historical and hypothetical exposure periods helped to shed light on the thresholds and triggers for raising R values above 1 in Nigeria. MPI and sanitation showed a well-defined relationship with R, with consistently higher access to sanitation and less poverty when R was less than 1. Thresholds which pushed R above one included decreasing access to sanitation below 50% and increasing the MPI above 0.32. Whereas the relationship between R and conflict events and PDSI

appeared to vary spatially, with some states showing a negative and some states a positive association. For these two covariates, the effect on R was largely dependent on the access to sanitation and poverty within the states, with high levels of sanitation and low poverty resulting in a decreased effect of PDSI and conflict. This showed that better sustainable development in the state acted as a buffer to social and environmental extremes and allowed people to adapt to these events better, due to less pre-existing vulnerability.

According to the World Bank[28], up to 47.3% (98 million people) of Nigeria's population live in multidimensional poverty. Poverty is a well-known risk factor for cholera, which is considered a disease of inequity[29]. Poverty can result in several risk factor cascades, which puts people at risk of not just cholera but several other diseases. Examples of these risks include poor access to WASH[10], inadequate housing[30], malnutrition[31] and overcrowding[32]. The expansion of sustainable development helps to reduce these risks and meeting or exceeding the Sustainable Development Goals would see significant gains in global health[33]. People living in poverty have fewer options and abilities to adapt to new and extreme situations, becoming trapped in the affected area or displaced to areas where their needs are not met. This provides further evidence for the need to reduce pre-existing vulnerabilities and their importance in terms of disasters adaptation and resilience[34,35].

Poverty when measured in monetary terms alone can create issues due to its impact on the risk factors stated and is an advantage of using the MPI as a poverty indicator. Nigeria's cash transfer scheme has allowed many Nigerians to meet the household income limit for poverty but there is a case for turning these funds and attention onto structural reform[36]. Nigeria's nationwide average access to sanitation is around 25%, therefore using these funds to increase access to sanitation may significantly improve health[37]. Here we show the need for expansion of sanitation to reduce cholera risks and the shocks of extremes on its transmission. In a recent review on the implementation of non-pharmaceutical cholera interventions, there was generally a high acceptance of several WASH interventions. Despite this, education was key and building community relationships is needed to achieve this, such as understanding cultural differences and barriers[38].

This is especially important in areas with conflict, where trust between the government and residents may have been lost[31].

Since 2002, Boko Haram (and Islamic State's West Africa Province) has been gaining a foothold and territory in northeastern Nigeria which has resulted in ongoing conflict, unrest and oppression of civilians[39]. Currently 5,860,200 people live in Borno state[40], where the fighting has been most concentrated. Millions of people live in conflict-affected populations and there is an increasing proportion of people living in early post conflict areas[41]. This is significant in terms of health and disease, as conflict has known risk factors for cholera along with several other diseases[20,2242] and can worsen several of the social risk factors discussed above. Here, conflict was included in the best fit model and in some states, highly influential in terms of cholera transmission. Providing services and protecting health in conflict zones is especially challenging and coordination across organisations in reporting and operations are needed to streamline resources and prevent duplication of services[43]. The traffic light system used here helps highlight what is needed in these situations to protect health and when outbreaks may occur.

PDSI and several of the other drought indices tested here showed high variable importance, this resulted in only small changes altering R values in some states. When analysing spatial differences between R and PDSI, the relationship appears to be multi-directional. Furthermore, access to sanitation and poverty were important in how PDSI impacted R, similar to the impacts of conflict. There is significant evidence to show that both droughts[19,26] and floods[27,44] can cause cholera outbreaks and elevated transmission. Mechanisms through which this can occur includes a lack of water increasing risky drinking water behaviour and floods allowing for the dispersal of the pathogen. Nigeria has a varied climate across the country and therefore both extremes are likely to be felt by those living there. Cholera outbreaks have been seen in both the rainy and the dry season and the traffic light system shows potential triggers for when extra vigilance is needed. This immediate insight is important, while continually working to offset cholera risks from extremes through sanitation and hygiene, which can take significant time and resources[45].

Despite adapting the methodology to account for this, a potential limitation may be lagged effects of the covariates on cholera[46,47]. Both long-term and short-term changes to the population may take time before changes in cholera transmission are evident. While some disasters may be considered slow-onset or rapid-onset and therefore defining their beginning is subjective. Despite this, the incubation period of cholera is short (<2 hours - 5 days) and previous research has suggested that acute shocks cause increases in cholera cases within the first week of the event[48,50]. Calculating R on monthly sliding windows and using monthly covariate data helped to reduce potential lagged effects on the R values, which would be captured if the one-week lag estimate is applicable here. Although beyond the scope of the research presented here, the impacts of different lagged periods for several of these covariates and cholera outbreaks is an essential area of future research.

Cholera is considered an under-reported disease, and the lack of symptomatic cases means that many are likely to be missed. There are also incentives not to report cholera cases, due to travel restrictions and isolations and implications for trade and tourism[51]. However, the robust reporting system in Nigeria suggests that the data used here is the best available for analysis. While during times of crisis, cholera may be over-reported or more accurately represent the cholera burden in the area. This is due to the presence of cholera treatment centres and external assistance from humanitarian aid and non-governmental organization, detecting cases that may have been missed previously[20].

The Global Task Force on Cholera Control's 2030 target of reducing cholera deaths by 90%[52] will require acceleration of current efforts and significant commitment. Increasing cholera research and data are important in achieving this and the traffic light system for cholera risk presented here sheds light on ways to reduce cholera outbreaks in fragile settings. The results highlight the importance of extreme events on cholera transmission and how reducing pre-existing vulnerability could offset the resultant cholera risk. This research is the first time several disaster types and measures of population vulnerability have been evaluated together quantitatively in terms of cholera. We hope it

shows the importance of doing so to gain a more accurate understanding of disease outbreaks in complex emergencies. Nigeria is currently working towards its ambitious goal of lifting 100 million people out of poverty by 2030[36]. If it is successful, this could significantly improve health, increase quality of life and decrease the risks of social and environmental extremes.

Methods

Datasets

Cholera data were obtained from NCDC and contained linelist data for 2018 and 2019. The data were age and sex-disaggregated, on a daily temporal scale and to administrative level 4. The data also provided information on the outcome of infection and whether the patient was hospitalised. The data were subset to only include cases which were confirmed either by rapid diagnostic tests or by laboratory culture.

A range of covariates were investigated based on previously understood cholera risk factors. Covariates included conflict (monthly, daily)[53], drought (Palmers Drought Severity Index, Standardised Precipitation Index)[54,55], IDPs (households, individuals)[56], WASH (improved drinking water, piped water, improved sanitation, open defecation)[57], healthcare (total facilities, facilities per 100,000)[53], population (total, density)[58] and poverty (MPI, headcount ratio in poverty, intensity of depravation among the poor)[53].

The covariate data were on a range of spatial and temporal scales, therefore administrative level one (state) was set as the spatial granularity (data on a finer spatial scale were attributed to the upper level) and the smallest temporal scale possible was used for covariate selection (repeating values if data were not available to the lower level). The datasets and methods used here were approved by Imperial College Research Ethics Committee and a data sharing agreement through NCDC and the authors.

Incidence and R

The 2018 and 2019 positive linelist data were used to calculate incidence. Incidence was calculated on a daily scale by taking the sum of the data entry points by state and date of onset of symptoms. This created a new dataset with a list of dates and corresponding incidence for each state. All analysis was completed in R Studio version 4.1.0. (packages "incidence"[59] & "EpiEstim"[60]).

Using the incidence data, R was calculated with the parametric standard interval method, which uses the mean and the standard deviation of the standard interval (SI). The SI for cholera is well-documented and there are several estimates in the literature[61-63]. The parametric method was used (vs the non-parametric which uses a discrete distribution), as assumptions can be made here about data distribution and parameters. SI is the time from illness onset in the primary case to onset in the secondary case and therefore impacts the evolution of the epidemic and speed of transmission. To account for several reported SI values for cholera, a sensitivity analysis was used including 3, 5 and 8 days with a standard deviation of 8 days.

Estimating R too early in the epidemic increases error, as R calculations are less accurate when there are lower incidence cases over the time window. A way to understand how much this impacts R values is to use the coefficient of variation (CV), which is a measure of how spread out the dataset values are relative to the mean. The lower the value, the lower the degree of variation in the data and a posterior coefficient of variation was set to 0.3 (or less) as standard, based on previous work[60]. To reach the CV threshold, calculation start date for each state was altered until the threshold CV was reached. States with <40 cases were removed, as states with fewer cases did not have high enough incidence across the time window to reach the CV threshold.

Additionally, the R values were calculated over monthly sliding windows, to ensure sufficient cases in the time window. Daily and two-week sliding windows did not have incidence values sufficient to reach the posterior CV threshold. Monthly data were the most common temporal data granularity

used in the model and many of the covariates could have potential lag effects on cholera. Several of the tested covariates may not have impacted cholera immediately and using monthly data and calculating R over monthly sliding windows helped to account for some of this uncertainty.

Covariate Selection and Random Forest Models

The covariates listed above (conflict, drought, IDPs, WASH, healthcare, population and poverty) were run through a covariate selection process[64,65]. The selection process removes covariates not significantly associated with the outcome variable (R) and clusters the remaining based on the degree of correction between them. The threshold for clustering was set to an absolute pairwise correlation of above 0.75 and the aim was to reduce multi-linearity in the final model. R was chosen as the outcome variable, rather than incidence (which has less implicit assumptions) as it is more descriptive, providing information on the evolution of the epidemic (e.g., R = >1, cases are increasing) and not a single time point of disaster burden.

Using the subset list of covariates, the aim was to fit a model which could accurately predict R values under changing conditions. Supervised machine learning algorithms such as decision-making algorithms, are now a widely used method. They work by choosing random data points from a training set and building a decision tree to predict the expected value given the attributes of these points. Transparency is increased by allowing the number of trees (estimators), number of features at each node split and resampling method to be specified. Random forest then combines several decision trees, combining predictions from multiple algorithms into one model, this makes them more accurate, while also dealing well with interactions and non-linear relationships[66,67].

Random forest variable importance was tested on all the covariates which were not removed from the covariate selection process. Variable importance is a measure of the cumulative decreasing mean standard error each time a variable is used as a node split in a tree. The remaining error left in predictive accuracy after a node split is known as node impurity and a variable which reduces this impurity is considered more important.

A training and testing dataset were created at a proportion of 70% training, to train the model and test the predictions. Covariates were fit to all three R values (SI 3, 5 & 8 days) on a daily temporal scale and to administrative level 1. Random forest regression models were used (vs classification) due to the continuous outcome variable. The parameters for training were set to repeated cross-validation for the resampling method, with ten resampling interactions and five complete sets of folds to complete. The model was tuned with an optional number of predictors at each split set to 2, based on the lowest out-of-bag (OOB) error rate and the evaluation metric used was RMSE (package "caret"[68]).

A hierarchical stepwise analysis was used to fit the models taking into consideration the covariate clustering and variable importance. One covariate was selected from each cluster, and different combinations tested until the best-fit model. Models were assessed against each other in terms of predictive accuracy, based upon R^2 and RMSE. Predictions were then calculated on the testing dataset to compare incidence-based vs covariate-based R values, evaluations were built on multiple metrics including correlation, R^2 and RMSE. Despite random forest models being accurate and powerful at predicting, they are easily over-fit and therefore calculating error for the predictions is important. Error was calculated using mean absolute error (MAE), where y_i is the prediction and x_i is the true value, with the total number of data points as n.

$$MAE = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

Nowcasting & Exposure Periods

The best fit model in terms of predictive power according to the metrics above, was used to predict R for the remaining states which did not have sufficient reported cases to calculate R using incidence or had missing data for certain dates. Data for the best fit model covariates were collected for the states and missing dates from the sources given above. The data for the selected covariates are shown spatially in Supplementary Figure 1.

To understand the conditions needed to raise R values from less than 1 to more than 1, the data were split into exposure periods, to investigate the relationship between the covariates and R in these different time periods. Historical periods were set by splitting the data into monthly periods for 2018 and 2019 and by the R threshold being equal to or more than 1 (R = >1) or less than 1 (R = <1). For each of the 48 historical exposure periods, the mean, standard deviation, median and standard error were calculated for each covariate. This was to understand how the mean and median covariate values changed when R was over or less than 1. The two thresholds were also compared by state to investigate any spatial differences in the historical covariate values.

Hypothetical exposure periods were then created based on these findings, using the median values and standard error for R = >1 and R = <1 as a starting point for the red exposure period and green exposure period, respectively. The amber exposure period was taken as a mid-point between the two. The best fit model was used to predict how this would impact R and the exposure periods altered as needed. This provided thresholds and triggers for outbreaks in Nigeria, creating a traffic light system for cholera risk.

To understand spatial differences, six states had additional sub-national analysis and included Borno, Kaduna, Nasarawa, Ekiti, Lagos and Kwara. These states were selected based on their relationship with conflict or PDSI and R (clear positive/clear negative relationship). Three hypothetical exposure periods (red, amber and green) for either conflict or PDSI were produced for these states, keeping all the other covariates at the R = >1 mean values. This was to understand spatial differences in these two covariates and understand the threshold needed to push R values below 1.

References

1. Ali, M., Nelson, A. R., Lopez, A. L. & Sack, D. A. Updated global burden of cholera in endemic countries. *PLoS Neglect. Trop. Dis.* **9**, e0003832 (2015).

2. King, A. A., Ionides, E. L., Pascual, M. & Bouma, M. J. Inapparent infections and cholera dynamics. *Nature* **454**, 877-880 (2008).

3. Lessler, J. et al. Mapping the burden of cholera in sub-Saharan Africa and implications for control: an analysis of data across geographical scales. *Lancet* **391**, 1908-1915 (2018).

4. Dalhat, M. M. et al. Descriptive characterization of the 2010 cholera outbreak in Nigeria. *BMC Public Health*. **14**, 1-7 (2014).

5. Ngwa, M. C. et al. The multi-sectorial emergency response to a cholera outbreak in internally displaced persons camps in Borno state, Nigeria, 2017. *BMJ Glob. Health*. **5**, e002000 (2020).

6. Sule, I. B., Yahaya, M., Aisha, A. A., Zainab, A. D., Ummulkhulthum, B. & Nguku, P. Descriptive epidemiology of a cholera outbreak in Kaduna State, Northwest Nigeria, 2014. *Pan Afr. Med. J*. **27**, (2017).

7. Adeneye, A. K. et al. Risk factors associated with cholera outbreak in Bauchi and Gombe States in North East Nigeria. *J. Public Health Epidemiol*. **8**, 286-296 (2016).

8. De Magny, G. C., Guégan, J. F., Petit, M. & Cazelles, B. Regional-scale climate-variability synchrony of cholera epidemics in West Africa. *BMC Infect. Dis.* **7**, 1-9 (2007).

9. Abdussalam, A. F. Modelling the climatic drivers of cholera dynamics in Northern Nigeria using generalised additive models. *Int. J. Geogr. Environ. Manage*. **2**, 84-97 (2016).

10. Gidado, S. et al. Cholera outbreak in a naïve rural community in Northern Nigeria: the importance of hand washing with soap, September 2010. *Pan Afr. Med. J*. **30** (2018).

11. Hutin, Y., Luby, S., Paquet, C. A large cholera outbreak in Kano City, Nigeria: the importance of hand washing with soap and the danger of street-vended water. *J. Water Health*. **1**, 45-52 (2003).

12. Dan-Nwafor, C. C. et al. A cholera outbreak in a rural north central Nigerian community: an unmatched case-control study. *BMC Public Health* **19**, 1-7 (2019).

13. Leckebusch, G. C. & Abdussalam, A. F. Climate and socioeconomic influences on interannual variability of cholera in Nigeria. *Health Place*. **34**, 107-17 (2015).

14. United Nations Statistical Division. Millennium Development Goal Indicators. https://unstats.un.org/unsd/mdg/SeriesDetail.aspx?srid=580 (2015).

15. Ali, M. et al. The global burden of cholera. *Bull. World Health Organ*. **90**, 209-218 (2012).

16. Anbarci, N., Escaleras, M. & Register, C. A. From cholera outbreaks to pandemics: the role of poverty and inequality. *Working Paper 05003* (Florida Atlantic University, FL, 2006).

17. World Bank Group. A Wake Up Call: Nigeria Water Supply, Sanitation, and Hygiene Poverty Diagnostic. World Bank; 2017 Aug.

18. Elimian, K.O. et al. Descriptive epidemiology of cholera outbreak in Nigeria, January–November, 2018: implications for the global roadmap strategy. *BMC Public Health* **19**, 1-11 (2019).

19. Charnley, G. E. C., Kelman, I., Green, N., Hinsley, W., Gaythorpe, K. A. M. & Murray, K. A. Exploring relationships between drought and epidemic cholera in Africa using generalised linear models. *BMC Infect. Dis*. **21,** 1-12 (2021).

20. Charnley, G. E. C., Jean, K., Kelman, I., Gaythorpe, K. A. M. & Murray, K. A. Using self-controlled case series to understand the relationship between conflict and cholera in Nigeria and the Democratic Republic of Congo. Preprint at https://doi.org/10.1101/2021.10.19.21265191 (2021).

21. Charnley, G. E. C., Kelman, I., Gaythorpe, K. A. M. & Murray, K. A. Traits and risk factors of post-disaster infectious disease outbreaks: a systematic review. *Sci. Rep*. **11**, 1-4 (2021).

22. Wells, C. R. et al. The exacerbation of Ebola outbreaks by conflict in the Democratic Republic of the Congo. *PNAS*. **116**, 24366-72 (2019).

23. Carter, S. E. et al. What questions we should be asking about COVID-19 in humanitarian settings: perspectives from the social sciences analysis cell in the Democratic Republic of the Congo. *BMJ Glob. Health*. **5**, e003607 (2020).

24. Musa, S. S. et al. Dual tension as Nigeria battles cholera during the COVID-19 pandemic. *Clin. Epidemiology Glob Health*. **12** (2021).

25. Elimian, K. O. et al. What are the drivers of recurrent cholera transmission in Nigeria? Evidence from a scoping review. *BMC Public Health*. **20**, 1-3 (2020).

26. Rieckmann, A., Tamason, C. C., Gurley, E. S., Rod, N. H. & Jensen, P. K. Exploring droughts and floods and their association with cholera outbreaks in sub-Saharan Africa: a register-based ecological study from 1990 to 2010. *Am. J. Trop. Med. Hyg*. **98**, 1269 (2018).
27. Jutla, A. et al. Environmental factors influencing epidemic cholera. *Am. J. Trop. Med. Hyg*. **89**, 597 (2013).
28. World Bank. Tackling poverty in multiple dimensions: A proving ground in Nigeria. https://blogs.worldbank.org/opendata/tackling-poverty-multiple-dimensions-proving-ground-nigeria (2021).
29. Talavera, A. & Perez, E. M. Is cholera disease associated with poverty?. *J. Infect. in Dev. Countr*. **3**, 408-11 (2009).
30. Penrose, K., Castro, M. C., Werema, J. & Ryan, E. T. Informal urban settlements and cholera risk in Dar es Salaam, Tanzania. *PLoS Neglect. Trop. Dis*. **4**, e631 (2010).
31. Charnley, G. E. C., Kelman, I. & Murray, K. A. Drought-related cholera outbreaks in Africa and the implications for climate change: a narrative review. *Pathog. Glob. Health*. 1-10 (2021).
32. Ververs. M. & Narra, R. Treating cholera in severely malnourished children in the Horn of Africa and Yemen. *Lancet*. **390**, 1945-6 (2017).
33. von Schirnding Y. Health and sustainable development: can we rise to the challenge?. *Lancet*. **360**, 632-7 (2002).
34. Masozera, M., Bailey, M. & Kerchner, C. Distribution of impacts of natural disasters across income groups: A case study of New Orleans. *Ecol. Econ*. **63,** 299-306 (2007).
35. Lahsen, M. & Ribot, J. Politics of attributing extreme events and disasters to climate change. *Wiley Interdiscip. Rev. Clim. Change*. **13**, e750 (2022).
36. Onyeiwu, S. *Nigeria's poverty profile is grim. It's time to move beyond handouts.* https://theconversation.com/nigerias-poverty-profile-is-grim-its-time-to-move-beyond-handouts-163302 (2021).
37. Ajisegiri, B. et al. Geo-spatial modeling of access to water and sanitation in Nigeria. *J. Water Sanit. Hyg. Dev.* **9**, 258-80 (2019).

38. Polonsky, J. A. et al. Feasibility, acceptability, and effectiveness of non-pharmaceutical interventions against infectious diseases among crisis-affected populations: a scoping review. *Infect. Dis. Poverty*. **11**, 1-9 (2022).

39. Falode, J. A. The nature of Nigeria's Boko Haram war, 2010-2015: A strategic analysis. *Perspect. Terror*. **10**, 41-52 (2016).

40. Borno State Government. Population. https://bornostate.gov.ng/population/ (2016).

41. Garfield, R. M., Polonsky, J. & Burkle, F. M. Changes in size of populations and level of conflict since World War II: implications for health and health services. *Disaster Med Public Health Prep*. **6**, 241-6 (2012).

42. Federspiel F, Ali M. The cholera outbreak in Yemen: lessons learned and way forward. BMC public health. 2018 Dec;18(1):1-8.

43. Ricau, M., Lacan, L., Ihemezue, E., Lantagne, D. & String, G. Evaluation of monitoring tools for WASH response in a cholera outbreak in northeast Nigeria. *J. Water Sanit. Hyg. Dev.* **11**, 972-82 (2021).

44. Sidley, P. Floods in southern Africa result in cholera outbreak and displacement. *BMJ* **336**, 471 (2008).

45. Onwe, F. I., Agu, A. P., Umezuruike, D. & Ogbonna, C. Factors responsible for the 2015 Cholera outbreak and spread in Ebonyi state, Nigeria. *J. Epidemiol. Soc. Nigeria*. **2**, 53-58 (2018).

46. Reyburn, R., Kim, D. R., Emch, M., Khatib, A., Von Seidlein, L. & Ali, M. Climate variability and the outbreaks of cholera in Zanzibar, East Africa: a time series analysis. *Am. J. Trop. Med. Hyg*. **84**, 862 (2011).

47. Emch, M., Feldacker, C., Yunus, M., Streatfield, P. K., DinhThiem, V. & Ali, M. Local environmental predictors of cholera in Bangladesh and Vietnam. *Am. J. Trop Med. Hyg.* **78**, 823-32 (2008).

48. Fredrick, T. et al. Cholera outbreak linked with lack of safe water supply following a tropical cyclone in Pondicherry, India, 2012. *J. Health. Popul. Nutr*. **33**, 31 (2015).

49. Jeandron, A. et al. Water supply interruptions and suspected cholera incidence: a time-series regression in the Democratic Republic of the Congo. *PLoS Med.* **12**, e1001893 (2015).
50. Bhunia, R. & Ghosh, S. Waterborne cholera outbreak following cyclone Aila in Sundarban area of West Bengal, India, 2009. *Trans R Soc Trop*. **105**, 214-219 (2011).
51. Ganesan, D., Gupta, S. S. & Legros, D. Cholera surveillance and estimation of burden of cholera. *Vaccine* **38**, A13-7 (2020).
52. Global Task Force on Cholera Control. Roadmap 2030. https://www.gtfcc.org/about-gtfcc/roadmap-2030/ (2020).
53. HDX. The Humanitarian Data Exchange. https://data.humdata.org (2021).
54. University of East Anglia. Climate Research Unit. https://www.uea.ac.uk/groups-and-centres/climatic-research-unit (2020).
55. CEDA. High resolution Standardized Precipitation Evapotranspiration Index (SPEI) dataset for Africa. https://catalogue.ceda.ac.uk/uuid/bbdfd09a04304158b366777eba0d2aeb (2019).
56. IOM. DTM Nigeria. https://displacement.iom.int/nigeria (2021).
57. JMP. Nigeria. https://washdata.org (2020).
58. WorldBank. Data Bank Subnational Population. https://databank.worldbank.org/source/subnational-population (2021).
59. Kamvar, Z. N., Cai, J., Pulliam, J. R. C., Schumacher, J. & Jombart, T. Epidemic curves made easy using the R package incidence https://doi.org/10.12688/f1000research.18002.1 (2019).
60. Cori, A. EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves. R package version 2.2-4. https://CRAN.R-project.org/package=EpiEstim (2021).
61. Azman, A. S. et al. Urban cholera transmission hotspots and their implications for reactive vaccination: evidence from Bissau city, Guinea bissau. *PLoS Neglect Trop. Dis.* **6**, e1901 (2012).
62. Azman, A. S. et al. Population-level effect of cholera vaccine on displaced populations, South Sudan, 2014. *Emerg. Infect. Dis*. **22**, 1067 (2016).
63. Kahn, R. et al. Incubation periods impact the spatial predictability of cholera and Ebola outbreaks in Sierra Leone. *PNAS*. **117**, 5067-73 (2020).

64. Garske, T. et al. Yellow fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological data. *PLoS Med.* **11**, e1001638 (2014).

65. Gaythorpe, K. A. M. et al. The global burden of yellow fever. *Elife* **10**, e64670 (2021).

66. Breiman, L. Random forests. *Mach. Learn.* **45**, 5-32 (2001).

67. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res.* **13**, 1063-95 (2012).

68. Kuhn, M. caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret (2021).

Acknowledgements

We would like to thank and acknowledgement the Nigeria Centre for Disease Control for providing the data used here and those who work for the NCDC who collected the data in the field. We would also like to thank Anwar Musah (University College London) and Kelly Elimian (Karolinska Institutet) for their guidance on cholera data for Nigeria and facilitating the partnership with NCDC. This work was supported by the Natural Environmental Research Council [NE/S007415/1], as part of the Grantham Institute for Climate Change and the Environment's (Imperial College London) Science and Solutions for a Changing Planet Doctoral Training Partnership. We also acknowledge joint Centre funding from the UK Medical Research Council and Department for International Development [MR/R0156600/1].

Author Contributions

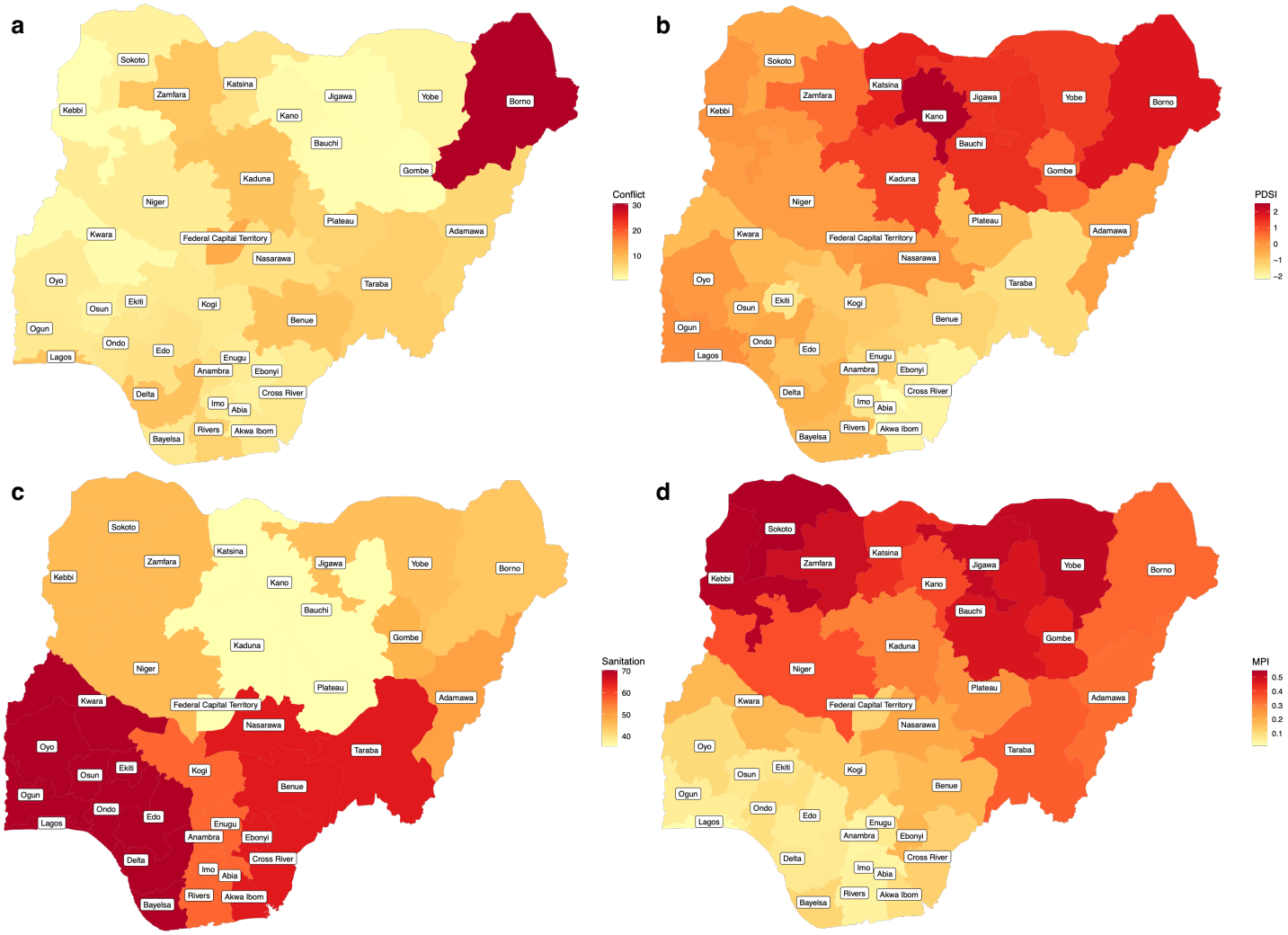
GECC was part of the study design and conceptualisation of ideas, ran the analysis, wrote and finalised the manuscript and incorporated any feedback. SY & CO provided the cholera datasets, facilitated the data sharing agreement and provided expertise on cholera in Nigeria. IK offered expertise on disasters and health, provided expertise in the methodology and revised several drafts. KAMG was part of the study design and conceptualisation of ideas, provided expertise in the methodology, provided supervision and revised several drafts. KAM was part of the study

design and conceptualisation of ideas, provided expertise in the methodology, provided supervision and revised several drafts. All authors have read and approved the manuscript.

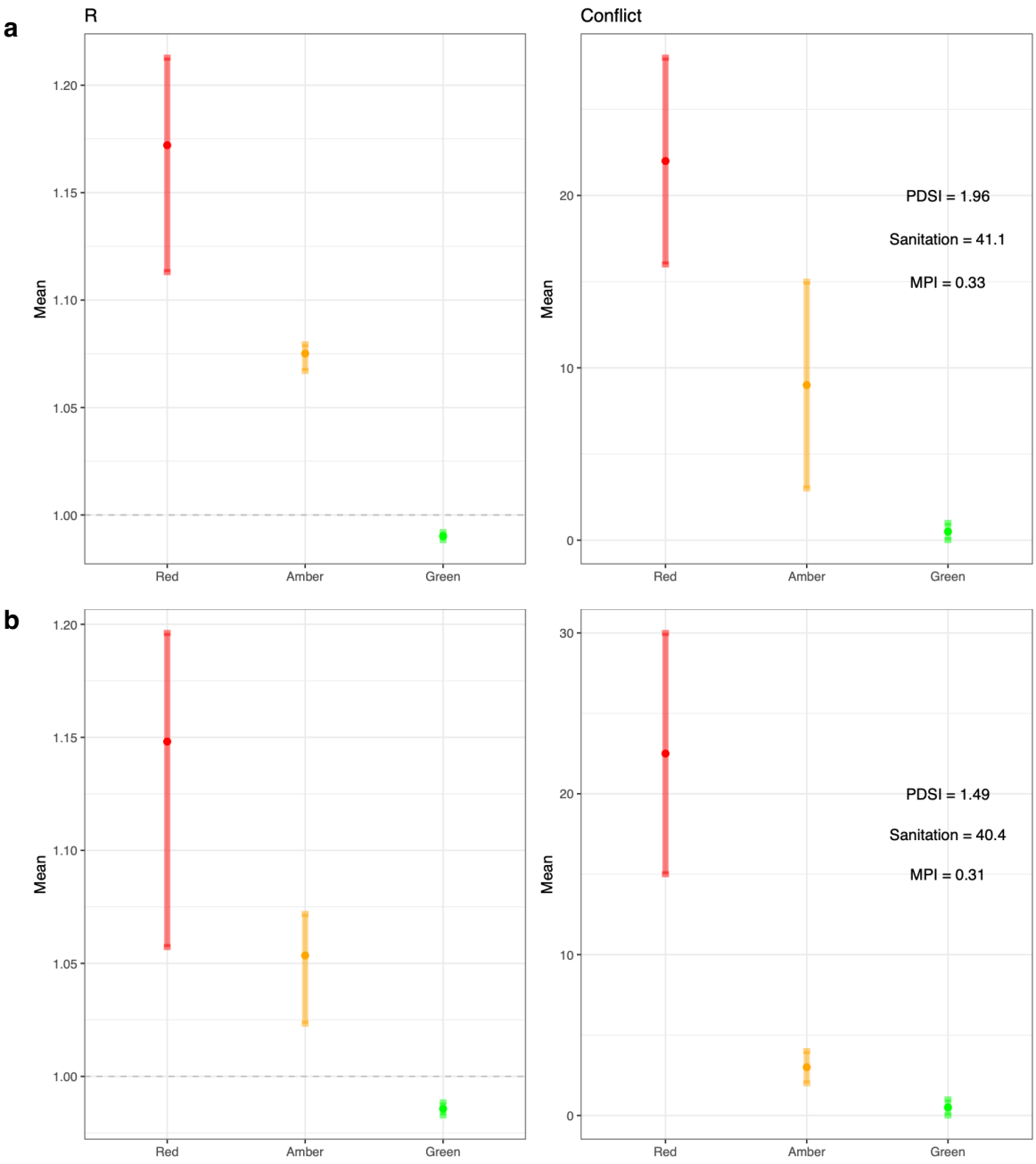
Competing interests

The authors declare no competing interests.

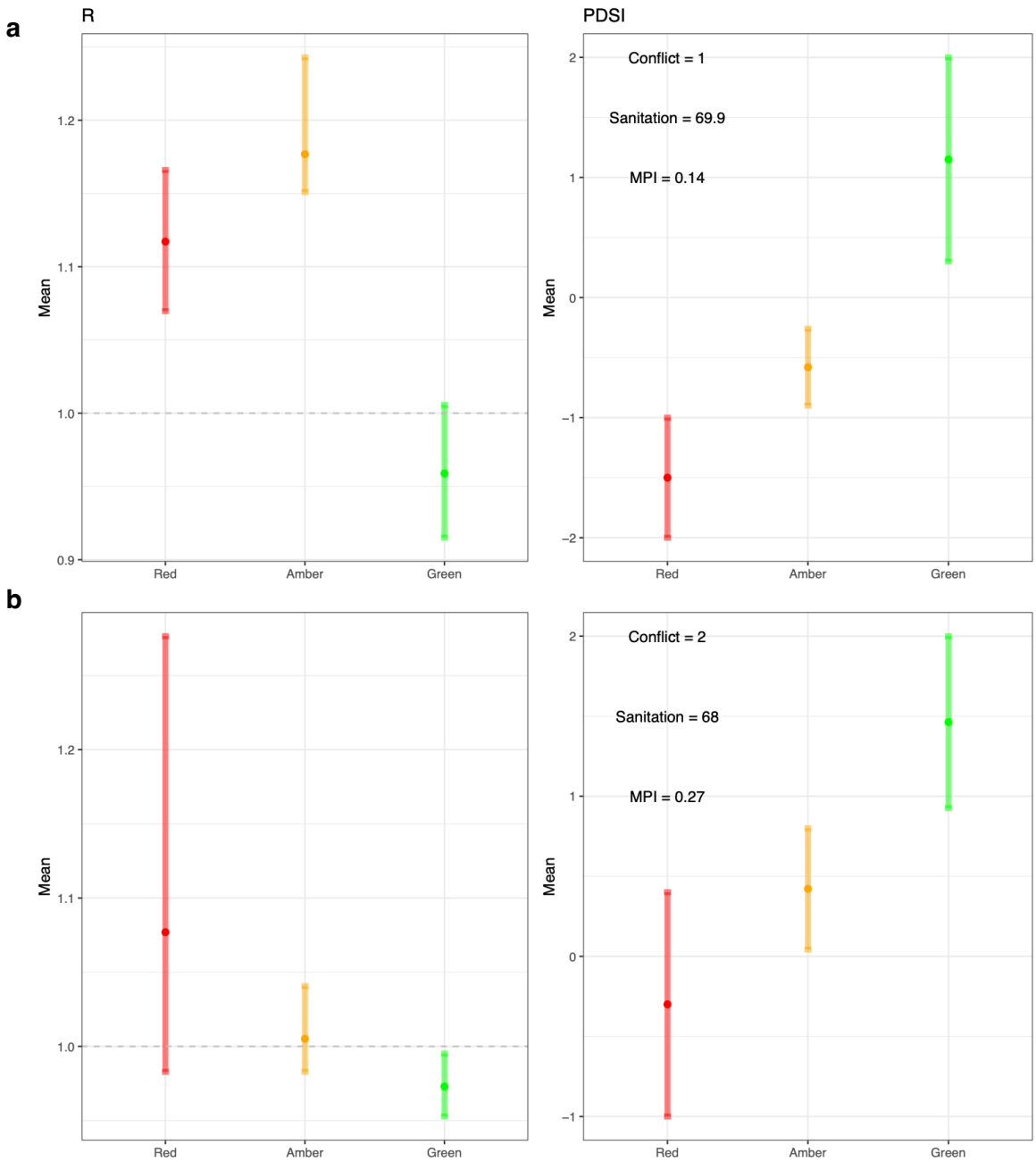
Supplementary Information



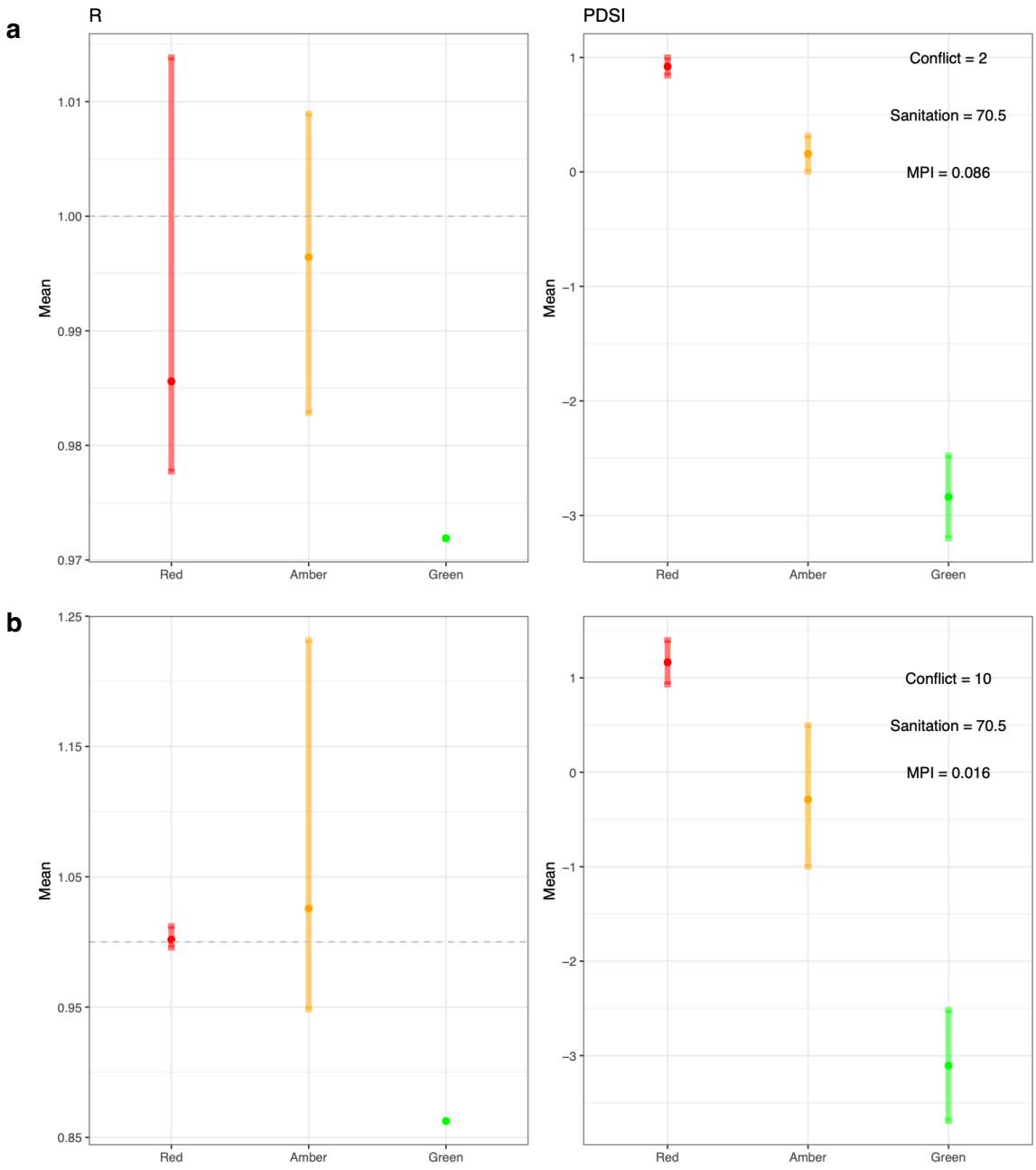
Supplementary Figure 1: Average values of the four covariates included in the best fit model. By state, covariates included: **a**, monthly conflict events, **b**, Palmers Drought Severity Index (PDSI), **c**, percentage access to sanitation and **d**, Multidimensional Poverty Index (MPI).



Supplementary Figure 2: Mean and range values for the three hypothetical conflict scenarios. The other three covariate values are kept the same at the mean value for R = >1 and the predicted R values mean and range for **a**, Borno and **b**, Kaduna.



Supplementary Figure 3: Mean and range values for the three hypothetical PDSI scenarios. The other three covariate values are kept the same at the mean value for R = >1 and the predicted R values mean and range for the states which found drier conditions caused R = >1. **a**, Kwara and **b**, Nasarawa.



Supplementary Figure 4: Mean and range values for the three hypothetical PDSI scenarios. The other three covariate values are kept the same at the mean value for R = >1 and the predicted R values mean and range for the states which found wetter conditions caused R = >1. **a**, Ekiti and **b**, Lagos.