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

Nigeria currently reports the second highest number of cholera cases in Africa, with numerous socioeconomic and environmental risk factors. Less investigated are the role of extreme events, despite recent work showing their potential importance. To address this gap, we estimated time varying reproductive number (R) from cholera incidence in Nigeria and used a machine learning approach to evaluate its association with extreme events (conflict, flood, drought) and pre-existing vulnerabilities (poverty, sanitation, healthcare). We then created a traffic-light system for cholera

outbreak risk, using three hypothetical traffic-light scenarios (Red, Amber and Green) and used this to predict R. The system highlighted potential extreme events and socioeconomic thresholds for outbreaks to occur. We found that reducing poverty and increasing access to sanitation lessened vulnerability to increased cholera risk caused by extreme events (monthly conflicts and the Palmers Drought Severity Index). The work presented here shows the need for sustainable development for disaster prevention and mitigation and to improve health and quality of life.

Introduction

Cholera was reintroduced into Africa in the 1970s during the seventh and continuing cholera pandemic. It has since caused significant mortality and morbidity, especially amongst the most vulnerable, such as children under five[1]. Despite this, other disease outbreaks have drawn attention away from cholera in Africa in recent years, including COVID-19 and Ebola[2,3]. Explosive cholera outbreaks are not uncommon due to the short incubation period (2 hours to 5 days) and high numbers of asymptomatic infections, which when contaminating the environment can sustain transmission[4]. Cholera is considered a disease of inequity and is preventable through wide-spread access to safe drinking water and sanitation[5]. However, the effect of these pre-existing vulnerabilities on disease risk can be exacerbated in times of environmental and social extremes, which can in turn act as a catalyst for, or exacerbate the impacts of, outbreaks.

Previous research has found several links between extreme events and cholera including floods, drought and conflict[6-8]. Disaster-related risk factors leading to disease outbreaks include an inability to access routine care such as vaccination, fears over safety, destruction of infrastructure, disruption of water, sanitation and hygiene (WASH) services and human displacement[9,10]. Previous research on disaster-related infectious disease outbreaks have examined extreme events in isolation[7,10], while others do not include multiple pre-existing socio-economic factors into the methodology[11,12]. Research linking several social and environmental extremes to diseases, including via risk factor cascades, is a global research gap and is important for predicting cholera transmission and mitigating outbreaks[13].

Nigeria currently reports the second highest number of estimated cholera cases in Africa[1,14] and has experienced many large outbreaks[15-18]. This is likely due to the presence of many underlying social and environmental risk factors, including a favourable climate[19,20], poor access to WASH[21,22] and a high proportion of the population living in poverty (62% at <$1.25/day)[23-25]. It also has a relatively robust reporting system which may correlate with more cases, as cholera is an under-reported disease and cases and deaths are often missed or misattributed. The country has been frequently challenged by both social and environmental extremes such as drought and floods, which may alter in intensity and frequency with climate change[13,24], along with ongoing conflict in the northeastern region due to Boko Haram (Islamic State West Africa Province)[8,13].

Here, we aim to resolve the role of extreme events in causing or contributing to cholera and increase the attention on cholera in Nigeria. In collaboration with the Nigeria Centre for Disease Control (NCDC), we evaluated by way of machine learning how a range of environmental and social covariates influence time-varying reproductive number (R) of cholera. Using the model with the best predictive power, we predicted a traffic-light system of cholera risk to illustrate how disasters and pre-existing vulnerabilities alter R and therefore the risk of cholera outbreaks. We anticipate that this novel and relatively simple framework of cholera outbreak risks could be employed by a range of professionals working in fragile settings by targeting interventions towards key disaster-related risk factors.

Results

Incidence and R

In Nigeria, there were 837 and 564 confirmed cholera cases for 2018 and 2019, respectively. The geographic distribution of confirmed cases is shown in Fig. 1a and are concentrated in the northeast of the country, with Adamawa, Borno, Katsina and Yobe having the highest burden. The number of cases declined steeply with age to a minimum in the 35-44 years category, before increasing again

over 45 years. Whereas, cases were relatively evenly split by sex overall, with slightly more males affected in 2018 (51.6% male) and more females in 2019 (43.6% male) (Fig. 1b).

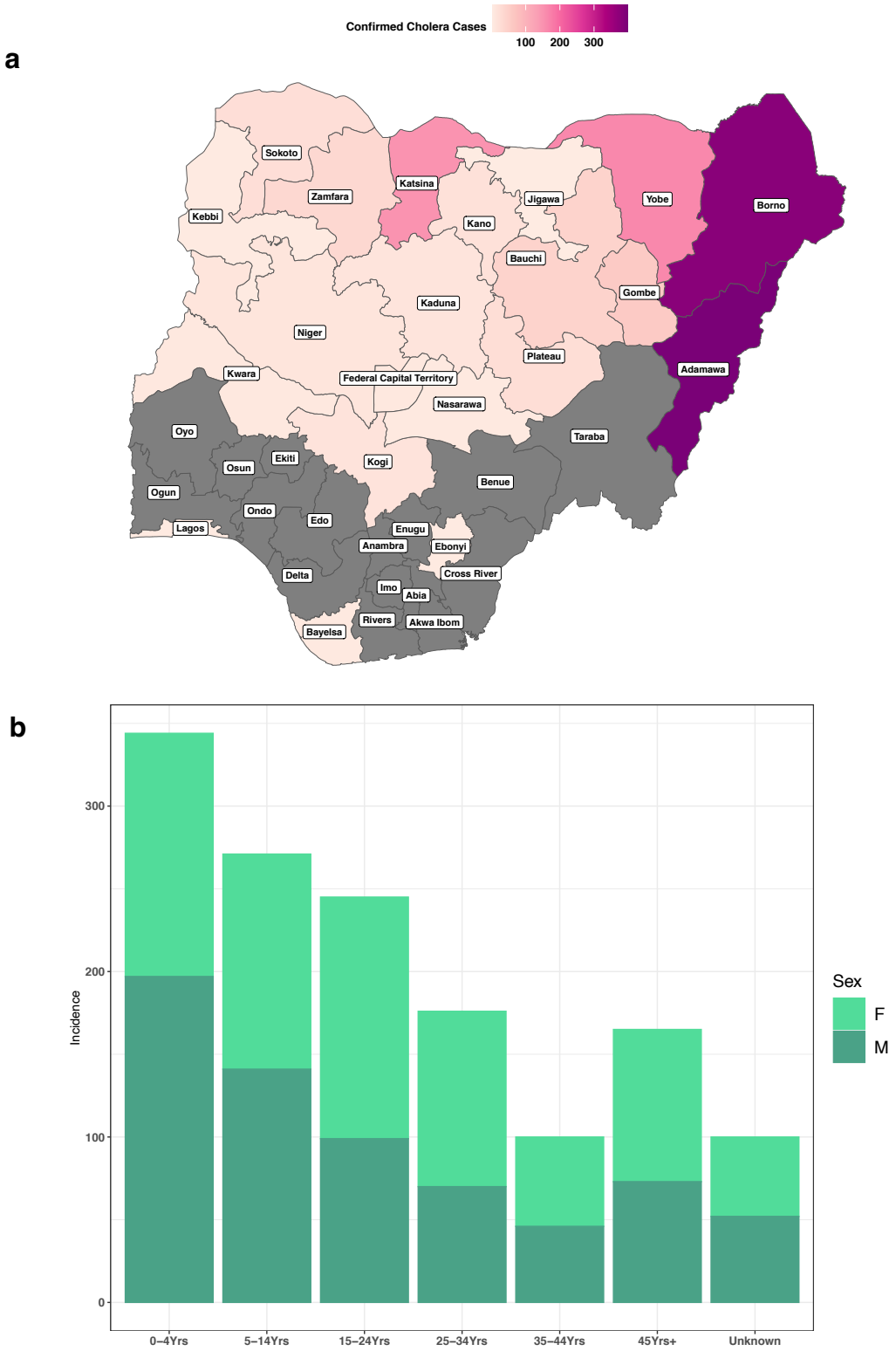


Fig. 1: Number of confirmed cholera cases. a, by state, grey indicates states that had no reported confirmed cases and **b**, by sex and age group, all for 2018 and 2019.

Six states for 2018 and two states for 2019 had sufficient cases 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 data 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.

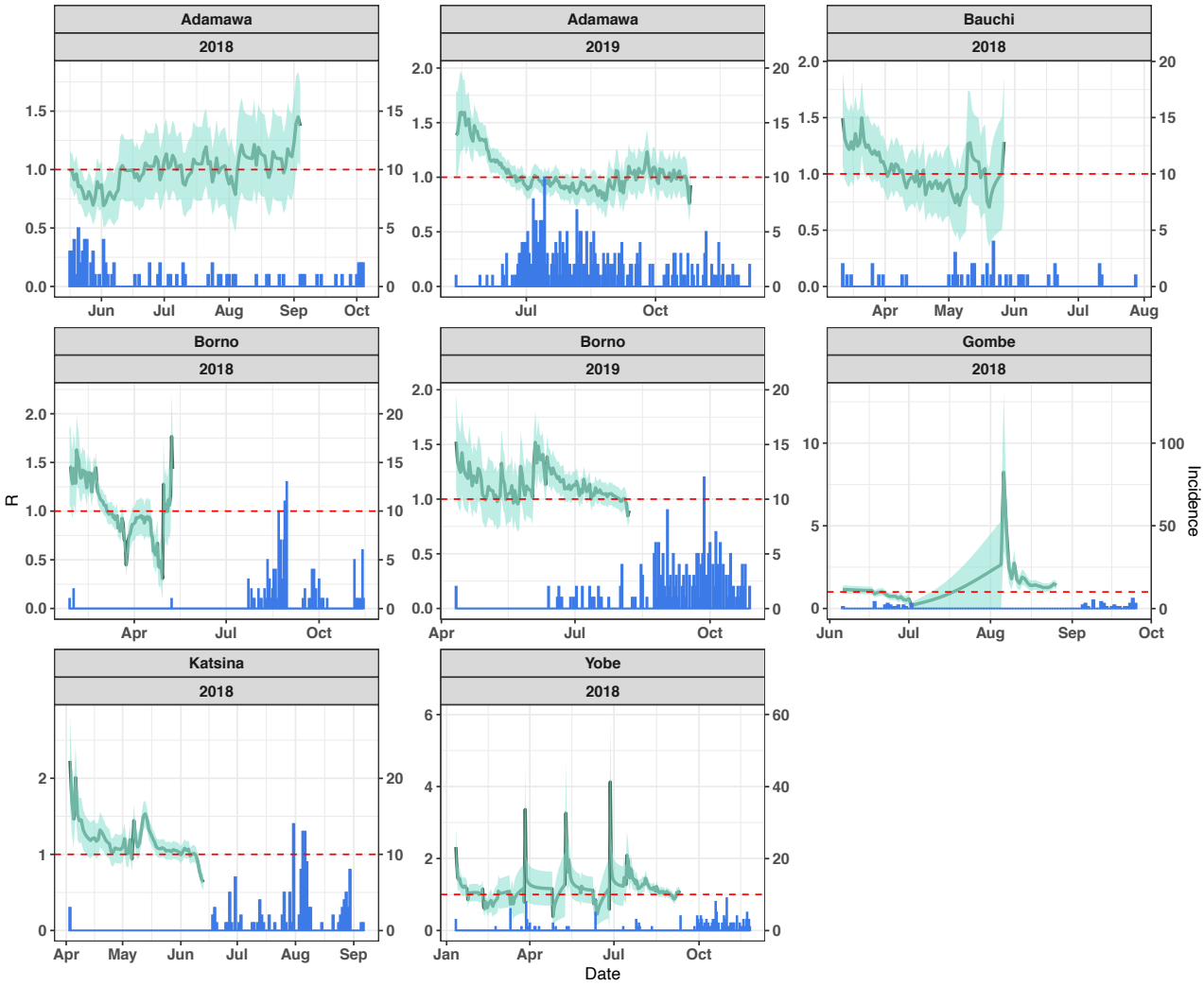


Fig. 2: R values (line) calculated from the incidence (bar) of cholera. Data is for the confirmed cholera cases for 2018 and 2019 of states which met the threshold equal to or more than 40 cases.

Covariate Selection and Random Forest Models

Twenty-one covariates were included in the clustering and variable importance analyses and were grouped into nine clusters. The clusters and variable importance (based on reducing node impurity)

of each covariate are shown in Fig. 3. Stepping through different covariate combinations, the best fit 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 (including error) are shown in Fig. 4 and had a correlation of 0.87, with the model RMSE at 0.33 and R^2 of 0.32.

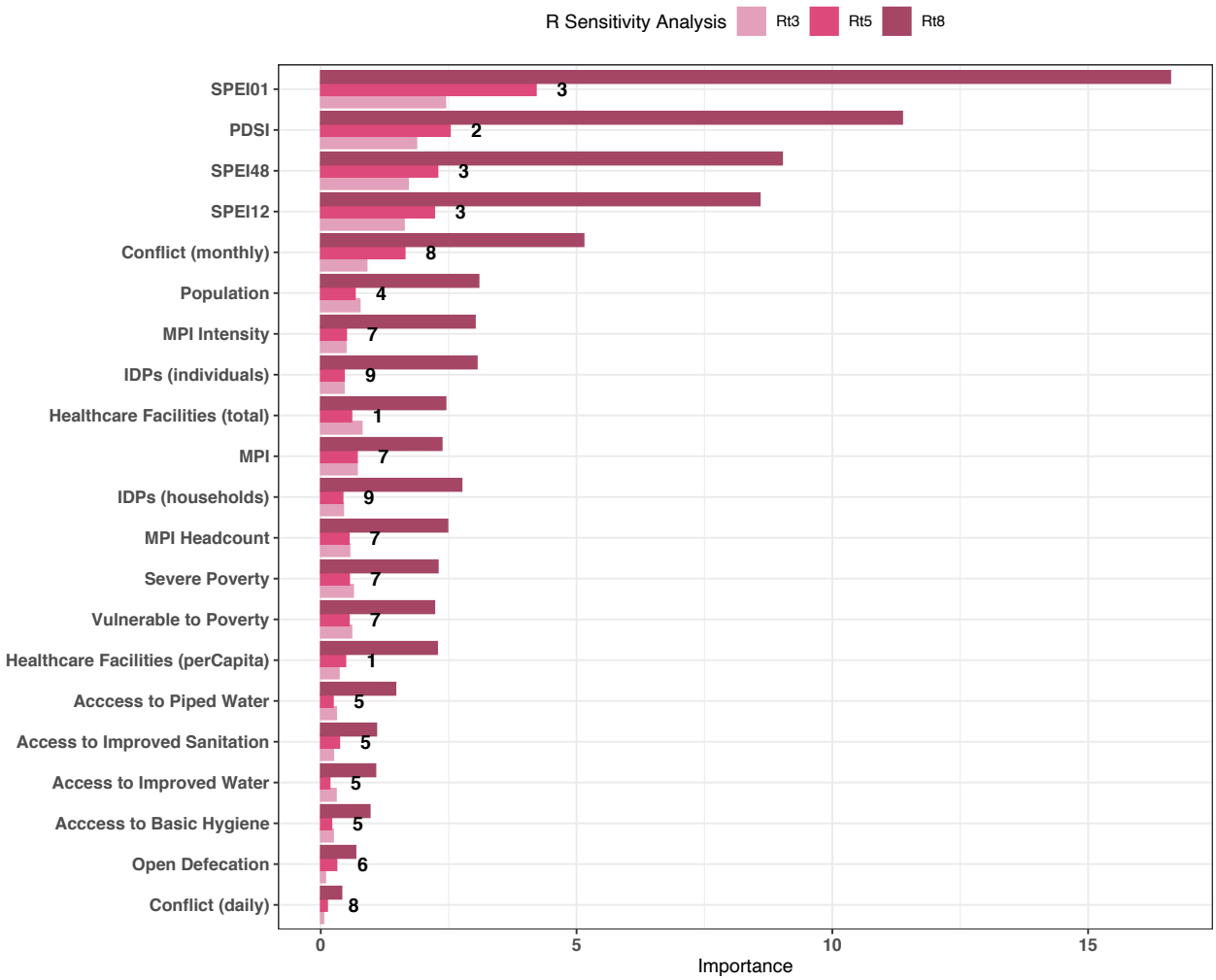


Fig. 3: The variable importance for the 21 tested for inclusion in the best fit model. All three serial interval values tested are shown (Rt3 - 3 days, Rt5 - 5 days, Rt8 - 8 days) and the numbers represent the clusters. Variable importance is measured through node impurity (see Methods for details). SPEI01, 12, 48 - Standardised Precipitation Index calculated on 1, 12 and 48 month scale. PDSI - Palmers Drought Severity Index. MPI - Multidimensional Poverty Index. IDP – Internally Displaced Persons.

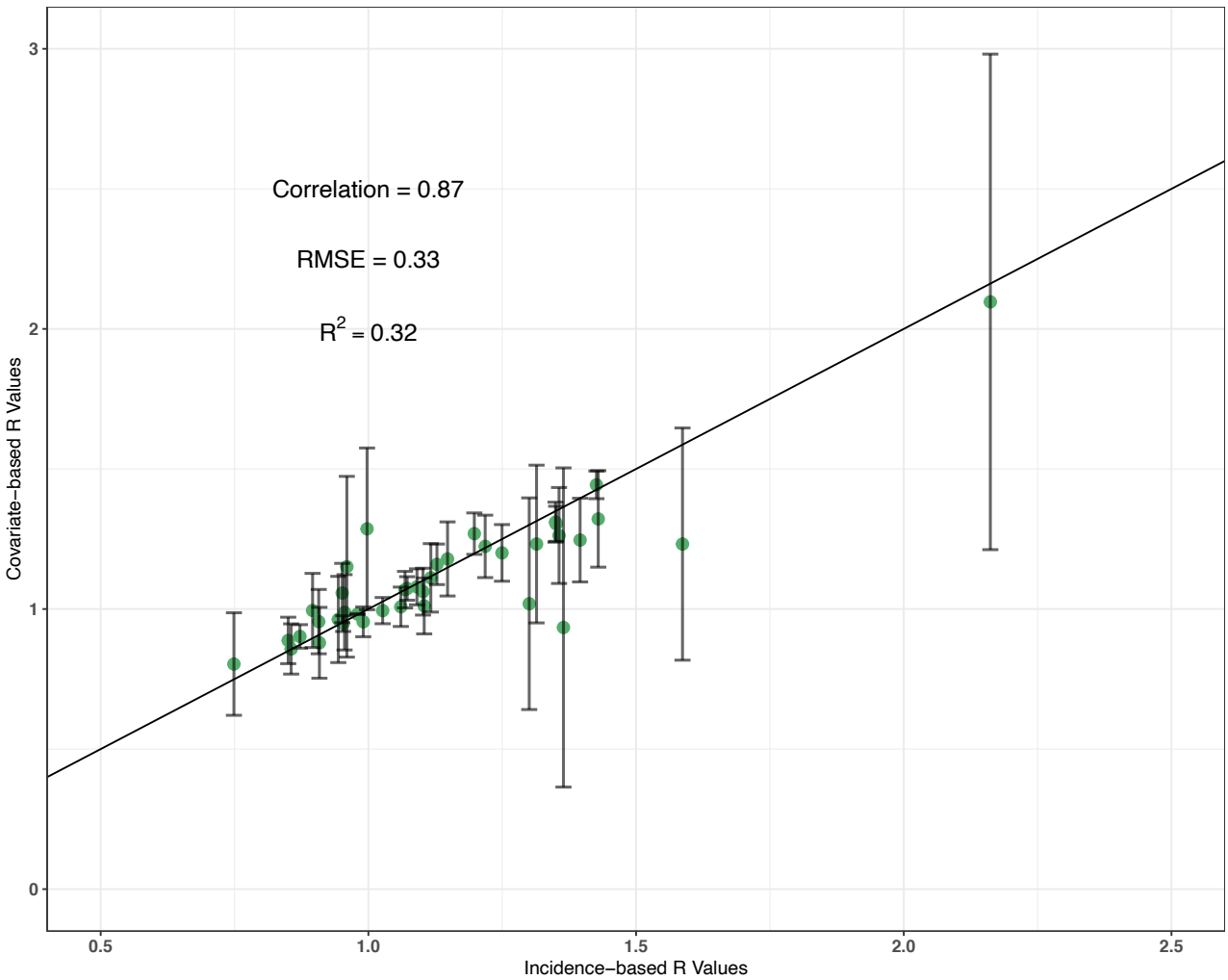


Fig. 4: Incidence-based vs covariate-based R values for the best fit model fitted to the testing dataset. The error bars show mean absolute error and the line is a linear trend line intercepting at 0.

Nowcasting

Using the best fit model, R was predicted for the remaining 31 states which did not have sufficient cases to be included in the R calculations and any missing dates for the six states which were included. This created estimates of R for all 37 states on a monthly temporal scale for 2018 and 2019. The predictions provide further evidence that the model accurately predicts R, as the higher R values were 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).

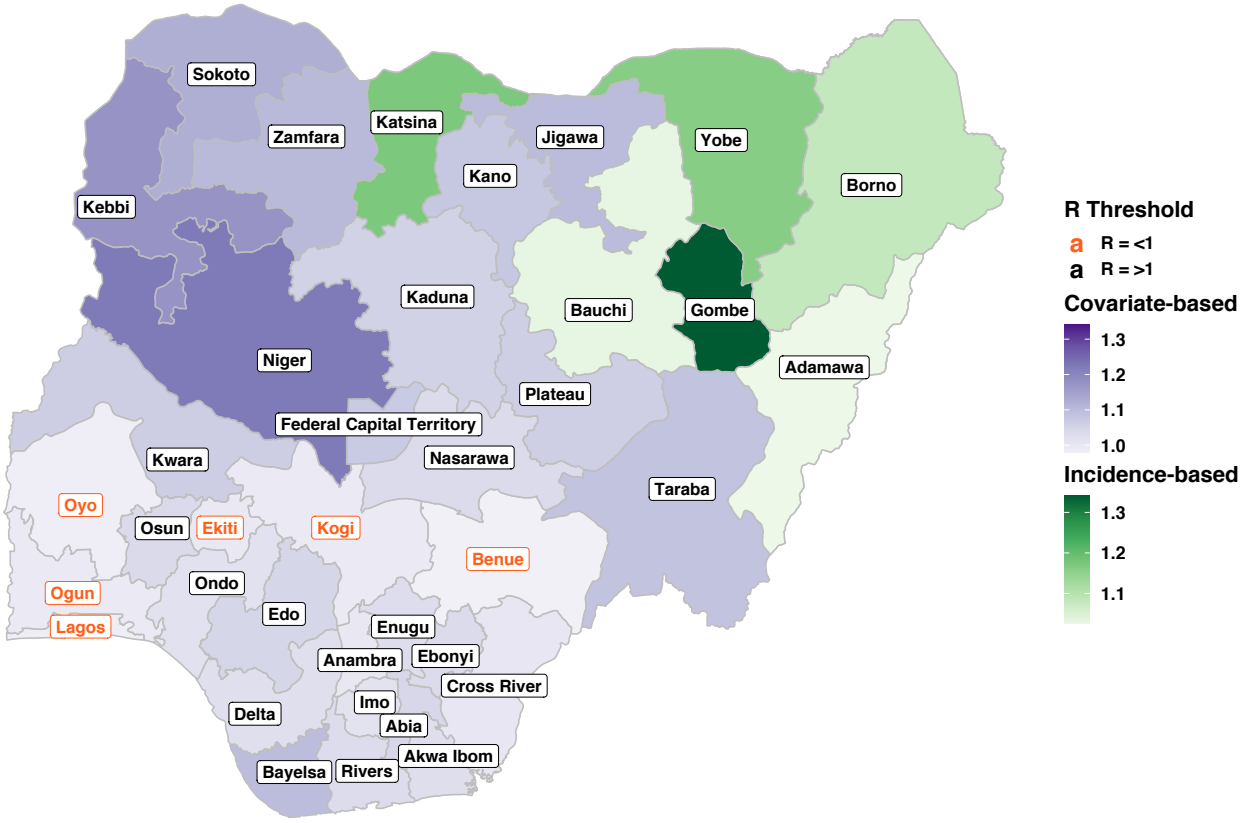


Fig. 5: Average R values for 2018 and 2019 for all 37 Nigerian states. Incidence-based (green) - the five states which met the equal to or more than 40 case thresholds. Covariate-based (purple) - the 31 states which did not meet the threshold and had R predicted using the best fit model. State label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

Traffic-Light System for Cholera Outbreak Risk

Fig. 6 shows the predicted R values for the three traffic-light scenarios (Red = R over 1, Amber = R around 1 and Green = R less than 1) of cholera outbreak risk, based on the four selected covariates. Sanitation and MPI had 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. R increased above 1 at 50% or lower for improved sanitation access and MPI values of above 0.32. The historical average sanitation level for R = >1 was 52.8% for the full dataset, 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 conflict having a wide range of values in each of the three traffic-light scenarios. For PDSI and conflict, R values increased above 1 at around -1.1 for PDSI and monthly conflict events of 1.6. The historical spatial trends for conflict and PDSI are presented in Supplementary Figure 1 and shows polarity in the relationships between the selected social and environmental extremes and R values, which differ between states.

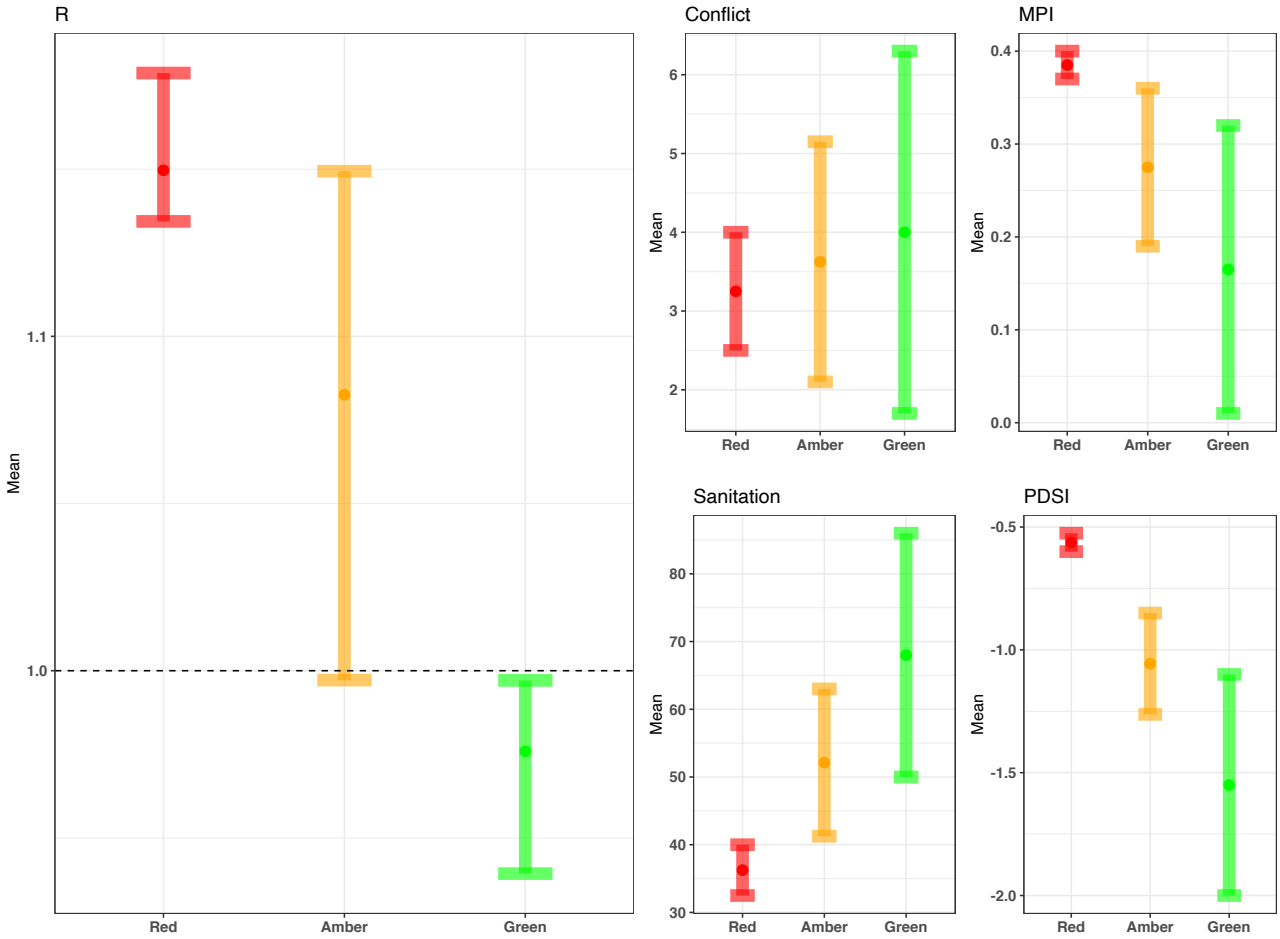


Fig. 6: Traffic-light system of cholera risk. The three traffic-light scenarios (Red = R over 1, Amber = R around 1 and Green = R less than 1) for each of the four covariates in the best fit model and the corresponding predicated R value using the best fit model.

Spatial Heterogeneities

Conflict

Borno and Kaduna were selected due to their clear positive relationship between conflict and R (increased conflict and R = >1). The three traffic-light scenarios created for conflict in these two

134 states found a consistently high cholera outbreak risk. The Green traffic-light scenario was relatively
135 small, with only a narrow range of conflict values causing R values less than 1. Both Kaduna and
136 Borno have high levels of poverty and low access to sanitation (40-41% access). For Borno, raising
137 monthly conflict events from 1 to 2 increased R above 1, but an increase in access to sanitation
138 from 41-46% pushed the R value back below one. This relationship continued in a stepwise pattern
139 and in a similar way for MPI but to a lesser degree. This showed that increasing sanitation and
140 therefore decreasing vulnerability, allowed the states to adapt to increasing conflict and keep the R
141 value below 1 (See Supplementary Figure 2).
142
143 *Drought*
144 Four states were investigated to evaluate the differences between extreme wetness (Lagos and
145 Ekiti) and extreme dryness (Nasarawa and Kwara) and R values over 1. In contrast to Borno and
146 Kaduna, all four states predicted consistently low R values (Supplementary Figs. 3 & 4), a potential
147 explanation for this is the high variable importance of PDSI (Fig. 3) and the high levels of sanitation
148 and low levels of poverty in all four states, contributing to overall lower predicted levels of cholera.
149 Therefore, the model was detecting a signal in only small changes in PDSI, that resulted in changing
150 R values which have not been detected in other states with higher rates of poverty and lower levels
151 of sanitation access. It also helps to highlight the multi-directionality of the relationship between
152 PDSI and cholera transmission, with both extreme wetness and extreme dryness causing increases
153 in R.
154
155 **Discussion**
156 The results presented here show the importance of social and environmental extremes on cholera
157 outbreaks in Nigeria, along with the importance of underlying vulnerability and socioeconomic
158 factors. Of the 1,401 positive cases for Nigeria in 2018 and 2019, the northeast of the country and
159 children under 5 carried the highest burden of disease, whereas there was minimal differentiation
160 in cases between sex. Six states were used to calculate the R values, including Adamawa, Bauchi,
161 Borno, Gombe, Katsina and Yobe. Twenty-one covariates were considered for model inclusion and

162 the best fit model according to the selected model performance measures (variable importance
163 based on node impurity, RMSE, R^2 and correlations) included monthly conflict events, percentage
164 of the population with access to sanitation, MPI and PDSI. Using the best fit model, nowcasting was
165 used to calculate the R values for the remaining thirty-one states which did not meet the threshold.
166
167 The predicted R values from the three traffic-light scenarios helped to shed light on the thresholds
168 and triggers for raising R values above 1 in Nigeria. MPI and sanitation showed a well-defined
169 relationship with R, with consistently higher access to sanitation and less poverty (lower MPI value)
170 when R was less than 1. Thresholds which pushed R above one included decreasing access to
171 sanitation below 50% and increasing the MPI above 0.32. Whereas the relationship between R and
172 conflict events and PDSI appeared to vary spatially, with some states showing a negative and some
173 states a positive association. For these two covariates, the effect on R was largely dependent on
174 the access to sanitation and poverty within the states, with high levels of sanitation and low poverty
175 resulting in a decreased effect of PDSI and conflict. This showed that better sustainable
176 development in the state acted as a buffer to social and environmental extremes and allowed people
177 to adapt to these events better, due to less pre-existing vulnerability.
178
179 According to the World Bank[26], up to 47.3% (98 million people) of Nigeria's population live in
180 multidimensional poverty. Poverty is a well-known risk factor for cholera, which is considered a
181 disease of inequity[27]. Poverty can result in several risk factor cascades, which puts people at risk of
182 not just cholera but several other diseases. Examples of these risks include poor access to WASH[21],
183 inadequate housing[28], malnutrition[29] and overcrowding[30]. The expansion of sustainable
184 development helps to reduce these risks and meeting or exceeding the Sustainable Development
185 Goals would see significant gains in global health[31]. People living in poverty have fewer options and
186 abilities to adapt to new and extreme situations, becoming trapped in the affected area or displaced
187 to areas where their needs are not met. This provides further evidence for the need to reduce pre-
188 existing vulnerabilities and to implement known techniques for reducing disasters[32,33].
189

190 Poverty when measured in monetary terms alone can create issues due to its impact on the risk
191 factors stated and is an advantage of using the MPI as a poverty indicator. Nigeria's cash transfer
192 scheme has allowed many Nigerians to meet the household income limit for poverty but there is a
193 case for turning these funds and attention onto structural reform[34]. Nigeria's nationwide average
194 access to sanitation is around 25%, therefore using these funds to increase access to sanitation
195 may significantly improve health[35]. Currently, 73% of the enteric disease burden in Nigeria is
196 associated with inadequate WASH[36] and here we show the need for expansion of sanitation to
197 reduce cholera risks and the shocks of extremes on its transmission. In a recent review on the
198 implementation of non-pharmaceutical cholera interventions, there was generally a high acceptance
199 of several WASH interventions. Despite this, education was key and building community
200 relationships is needed to achieve this, such as understanding cultural differences and barriers[37].
201 This is especially important in areas with conflict, where trust between the government and residents
202 may have been lost[29].
203
204 Since 2002, Boko Haram (and Islamic State's West Africa Province) has been gaining a foothold
205 and territory in northeastern Nigeria which has resulted in ongoing conflict, unrest and oppression
206 of civilians[38]. Currently 5,860,200 people live in Borno state[39], where the fighting has been most
207 concentrated. Millions of people comprise conflict-affected populations globally and there is an
208 increasing proportion of people living in early post conflict areas[40]. This is significant in terms of
209 health and disease, as conflict has known risk factors for cholera along with several other
210 diseases[8,10,41] and can worsen several of the social risk factors discussed above. Here, conflict was
211 included in the best fit model and in some states, highly influential in terms of cholera transmission.
212 Providing services and protecting health in conflict zones is especially challenging and coordination
213 across organisations in reporting and operations are needed to streamline resources and prevent
214 duplication of services[42]. The traffic-light system used here helps highlight what is needed in these
215 situations to protect health and when outbreaks may occur.
216

PDSI and several of the other drought indices tested here showed high variable importance but, in some states, had only marginal influence on R predictions when the PDSI values were manipulated. When analysing spatial differences between R and PDSI, the relationship appears to be multi-directional, with both extreme wetness (PDSI = +4) and extreme dryness (PDSI = -4) associated with R values above 1. 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[7,11] and floods[12,43] 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 work presented here shows potential triggers for when extra vigilance is needed, especially in certain states. 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[44].

Despite adapting the methodology to account for this, a potential limitation may be lagged effects of the covariates on cholera[45,46]. Both long-term and short-term changes to the population may take time before changes in cholera transmission are evident. While some influences 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 impacts cause increases in cholera cases within the first week of the event[47-49]. 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

245 restrictions and isolations and implications for trade and tourism[50]. However, the robust reporting
246 system in Nigeria suggests that the data used here is the best available for analysis. While during
247 times of crisis, cholera may be over-reported or more accurately represent the cholera burden in
248 the area. This is due to the presence of cholera treatment centres, increased awareness among the
249 population and external assistance from non-governmental organization, detecting cases that may
250 have been missed previously[8].

251 The Global Task Force on Cholera Control's 2030 target of reducing cholera deaths by 90%[51] will
252 require acceleration of current efforts and significant commitment. Increasing cholera research and
253 data are important in achieving this and the traffic-light system for cholera risk presented here sheds
254 light on ways to reduce cholera outbreaks in fragile settings. The results highlight the importance of
255 extreme events on cholera transmission and how reducing pre-existing vulnerability could offset the
256 resultant cholera risk. This research is the first time several disaster types and measures of
257 population vulnerability have been evaluated together quantitatively in terms of cholera. We hope it
258 shows the importance of doing so to gain a more accurate understanding of disease outbreaks in
259 complex emergencies. Nigeria is currently working towards its ambitious goal of lifting 100 million
260 people out of poverty by 2030[34]. If it is successful, this could significantly improve health, increase
261 quality of life and decrease the risks of social and environmental extremes.

262

263 **Methods**

264 *Datasets*

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

270

A range of covariates were investigated based on previously understood cholera risk factors. Covariates included factors related to conflict (monthly, daily)[52], drought (Palmers Drought Severity Index, Standardised Precipitation Index)[53,54], internally displaced persons (IDPs) (households, individuals)[55], WASH (improved drinking water, piped water, improved sanitation, open defecation, basic hygiene)[56], healthcare (total facilities, facilities per 100,000 people)[52], population (total)[57] and poverty (MPI, headcount ratio in poverty, intensity of deprivation among the poor, severe poverty and population vulnerable to poverty)[52].

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 administrative level 1) and the finest temporal scale possible (daily) was used for covariate selection (repeating values if data were not available at the daily level). The datasets and methods used here were approved by Imperial College Research Ethics Committee and a data sharing agreement between NCDC and the authors.

Incidence and R

The 2018 and 2019 laboratory confirmed linelist data were used to calculate incidence. Incidence was calculated on a daily scale by taking the sum of the cases reported by state and date of onset of symptoms. This created a new dataset with a list of dates and corresponding daily incidence for each state. All analysis was completed in R with R Studio version 4.1.0. (packages “incidence”[58] & “EpiEstim”[59]).

Rather than using incidence as the outcome variable (which has less implicit assumptions), R was calculated, as it is more descriptive providing information on epidemic evolution (e.g., R = >1, cases are increasing), instead of new reported disease cases for a single time point. R was calculated from incidence using the parametric standard interval method, which uses the mean and the standard deviation of the standard interval (SI). SI is the time from illness onset in the primary case

298 to onset in the secondary case and therefore impacts the evolution of the epidemic and speed of
299 transmission. The SI for cholera is well-documented and there are several estimates in the
300 literature[60-62]. To account for this reported variation in SI, a sensitivity analysis was conducted with
301 SI set at 3, 5 and 8 days with a standard deviation of 8 days. The parametric method was used (vs
302 the non-parametric which uses a discrete distribution), as this can be adequately modelled by
303 a normal probability distribution and has a fixed set of parameters.
304
305 Estimating R too early in an epidemic increases error, as R calculations are less accurate when
306 there is lower incidence over a time window. A way to understand how much this impacts R values
307 is to use the coefficient of variation (CV), which is a measure of how spread out the dataset values
308 are relative to the mean. The lower the value, the lower the degree of variation in the data. A
309 coefficient of variation threshold was set to 0.3 (or less) as standard, based on previous work[59]. To
310 reach the CV threshold, calculation start date for each state was altered until the threshold CV was
311 reached. States with <40 cases were removed, as states with fewer cases did not have high enough
312 incidence across the time window to reach the CV threshold. Additionally, R values were calculated
313 over monthly sliding windows, to ensure sufficient cases were available for analysis within the time
314 window.
315
316 *Covariate Selection and Random Forest Models*
317 Supervised machine learning algorithms such as decision-tree based algorithms, are now a widely
318 used method for predicting disease outcomes and risk mapping[63,64]. They work by choosing data
319 points randomly from a training set and building a decision tree to predict the expected value given
320 the attributes of these points. Transparency is increased by allowing the number of trees
321 (estimators), number of features at each node split and resampling method to be specified. Random
322 Forests (RF) then combines several decision trees into one model, which has been shown to
323 increase predictive accuracy over single tree approaches, while also dealing well with interactions
324 and non-linear relationships[65,66].
325

326 The covariates listed above (conflict, drought, IDPs, WASH, healthcare, population and poverty)
327 were first clustered to assist in the selection of covariates for model inclusion and to understand any
328 multicollinearities. Despite RF automatically reducing correlation through subsetting data and tuning
329 the number of trees and depth[64,67], the process here lends support that the final model is measuring
330 somewhat independent processes and not purely overfitting the same patterns[63]. The clustering
331 was based on the correction between the covariates meeting an absolute pairwise correlation of
332 above 0.75. A secondary covariate selection process was run during preliminary analysis and acted
333 as a method of validation. The process is detailed in Supplementary Information 1.
334
335 Random forest variable importance was used to rank all 21 clustered covariates. Variable
336 importance provided an additional method of guiding the fitting of the best fit model, by testing the
337 covariates which found the highest variable importance first. In this context, variable importance is
338 a measure of the cumulative decreasing mean standard error each time a variable is used as a
339 node split in a tree. The remaining error left in predictive accuracy after a node split is known as
340 node impurity and a variable which reduces this impurity is considered more important.
341
342 Training (70% of data) and testing (30%) datasets were created to train the model and test the
343 model's predictive performance. Random forest regression models (as opposed to classification
344 models) were used since the outcome variable (R) is continuous. The parameters for training were
345 set to repeated cross-validation for the resampling method, with ten resampling interactions and
346 five complete sets of folds to complete. The model was tuned and estimated an optimal number of
347 predictors at each split of 2, based on the lowest out-of-bag (OOB) error rate with RMSE used as
348 the evaluation metric (package "caret"[68]).
349
350 A stepwise analysis was used to fit the models under each SI condition (3, 5 & 8 days), taking into
351 consideration the covariate clustering and variable importance. One covariate was selected from
352 each cluster, and all combinations of covariates were tested until the best-fit model was found.

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 (R values calculated using the incidence data) vs covariate-based R values (R values calculated through model predictions). The terms, actual vs predicted was not used here, as all R values were modelled making the term "actual" misleading in this context. Model performance evaluations were built on multiple metrics including correlation, R^2 and RMSE.

Despite random forest models being accurate and powerful for prediction, they are easily over-fit (fitting to the testing dataset too closely or exactly) and therefore calculating error for the predictions are important. Little to no error in the predictions are an indication of over-fitting which can occur through predictions based off too small a dataset, more parameters than can be justified by the data and multicollinearity. Here, 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

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 5.

Traffic-Light System for Cholera Outbreak Risk

The best fit model was then used to predict the traffic-light system for cholera outbreak risk, by manipulating the covariates values and using these to predict R. The traffic light system was defined by:

- Red - Covariate values which pushed R over 1

379 • Amber - Covariates values with predicted R around 1

380 • Green - Covariate values which predicted R below 1.

381 By using these three traffic-light scenarios, cholera outbreak triggers were identified based on the
382 conditions of the four selected covariates. No specific R value had to be met for each traffic-light
383 scenario, to account for the complexity of the relationships and non-linearity (Supplementary Figs.
384 6 & 7). To illustrate the historical trends between the best fit model covariates and the R thresholds
385 (R = >1, R <1), the data is split both spatially (by month) and temporally (by state) in Supplementary
386 Figs. 1 & 8.

387

388 *Spatial Heterogeneities*

389 To understand spatial differences in the relationship between the selected social and environmental
390 extremes (conflict and PDSI) and cholera outbreak risk and the role pre-exiting vulnerabilities played
391 in altering these relationships, six states were selected for additional analysis. These states were
392 selected because they had either a clear positive or clear negative relationship with conflict or PDSI
393 and R (PDSI is hypothesised to increase R at either end of the scale, +4/-4) and included Borno,
394 Kaduna, Nasarawa, Ekiti, Lagos and Kwara (see Supplementary Figure 1). The processes above
395 for predicting R under the three traffic-light scenarios was repeated for the six states but only PDSI
396 and conflict values were manipulated, keeping the other three covariates at the mean value for R =
397 >1 across the full dataset for the state. The spatial analyses identified the thresholds in conflict and
398 PDSI needed to push R values below 1.

399

400 **References**

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

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

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

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

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

412 6. Elimian, K.O. et al. Descriptive epidemiology of cholera outbreak in Nigeria, January–November,
413 2018: implications for the global roadmap strategy. *BMC Public Health* **19**, 1-11 (2019).

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

417 8. Charnley, G. E. C., Jean, K., Kelman, I., Gaythorpe, K. A. M. & Murray, K. A. Using self-
418 controlled case series to understand the relationship between conflict and cholera in Nigeria and
419 the Democratic Republic of Congo. Preprint at https://doi.org/10.1101/2021.10.19.21265191
420 (2021).

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

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

425 11. Rieckmann, A., Tamason, C. C., Gurley, E. S., Rod, N. H. & Jensen, P. K. Exploring droughts
426 and floods and their association with cholera o20utbreaks in sub-Saharan Africa: a register-
427 based ecological study from 1990 to 2010. *Am. J. Trop. Med. Hyg*. **98**, 1269 (2018).

428 12. Jutla, A. et al. Environmental factors influencing epidemic cholera. *Am. J. Trop. Med. Hyg*. **89**,
429 597 (2013).

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

14. 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).

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

16. 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).

17. 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).

18. 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).

19. 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).

20. 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).

21. 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).

22. 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).

23. 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).

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

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

26. 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).
27. Talavera, A. & Perez, E. M. Is cholera disease associated with poverty?. *J. Infect. in Dev. Countr*. **3**, 408-11 (2009).
28. 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).
29. 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).
30. Ververs. M. & Narra, R. Treating cholera in severely malnourished children in the Horn of Africa and Yemen. *Lancet*. **390**, 1945-6 (2017).
31. von Schirnding Y. Health and sustainable development: can we rise to the challenge?. *Lancet*. **360**, 632-7 (2002).
32. 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).
33. Lahsen, M. & Ribot, J. Politics of attributing extreme events and disasters to climate change. *Wiley Interdiscip. Rev. Clim. Change*. **13**, e750 (2022).
34. 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).
35. Ajisegiri, B. et al. Geo-spatial modeling of access to water and sanitation in Nigeria. *J. Water Sanit. Hyg. Dev.* **9**, 258-80 (2019).
36. World Bank Group. A Wake Up Call: Nigeria Water Supply, Sanitation, and Hygiene Poverty Diagnostic. World Bank; 2017 Aug.
37. 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).

485 38. Falode, J. A. The nature of Nigeria's Boko Haram war, 2010-2015: A strategic analysis.
486 *Perspect. Terror*. **10**, 41-52 (2016).
487 39. Borno State Government. Population. https://bornostate.gov.ng/population/ (2016).
488 40. Garfield, R. M., Polonsky, J. & Burkle, F. M. Changes in size of populations and level of conflict
489 since World War II: implications for health and health services. *Disaster Med Public Health Prep*.
490 **6**, 241-6 (2012).
491 41. Federspiel F, Ali M. The cholera outbreak in Yemen: lessons learned and way forward. BMC
492 public health. 2018 Dec;18(1):1-8.
493 42. Ricau, M., Lacan, L., Ihemezue, E., Lantagne, D. & String, G. Evaluation of monitoring tools
494 for WASH response in a cholera outbreak in northeast Nigeria. *J. Water Sanit. Hyg. Dev.* **11**,
495 972-82 (2021).
496 43. Sidley, P. Floods in southern Africa result in cholera outbreak and displacement. *BMJ* **336**, 471
497 (2008).
498 44. Onwe, F. I., Agu, A. P., Umezuruike, D. & Ogbonna, C. Factors responsible for the 2015
499 Cholera outbreak and spread in Ebonyi state, Nigeria. *J. Epidemiol. Soc. Nigeria*. **2**, 53-58
500 (2018).
501 45. Reyburn, R., Kim, D. R., Emch, M., Khatib, A., Von Seidlein, L. & Ali, M. Climate variability and
502 the outbreaks of cholera in Zanzibar, East Africa: a time series analysis. *Am. J. Trop. Med. Hyg*.
503 **84**, 862 (2011).
504 46. Emch, M., Feldacker, C., Yunus, M., Streatfield, P. K., DinhThiem, V. & Ali, M. Local
505 environmental predictors of cholera in Bangladesh and Vietnam. *Am. J. Trop Med. Hyg.* **78**, 823-
506 32 (2008).
507 47. Fredrick, T. et al. Cholera outbreak linked with lack of safe water supply following a tropical
508 cyclone in Pondicherry, India, 2012. *J. Health. Popul. Nutr*. **33**, 31 (2015).
509 48. Bhunia, R. & Ghosh, S. Waterborne cholera outbreak following cyclone Aila in Sundarban area
510 of West Bengal, India, 2009. *Trans R Soc Trop*. **105**, 214-219 (2011).

511 49. Jeandron, A. et al. Water supply interruptions and suspected cholera incidence: a time-series
512 regression in the Democratic Republic of the Congo. *PLoS Med.* **12**, e1001893 (2015).
513 50. Ganesan, D., Gupta, S. S. & Legros, D. Cholera surveillance and estimation of burden of
514 cholera. *Vaccine* **38**, A13-7 (2020).
515 51. Global Task Force on Cholera Control. Roadmap 2030. https://www.gtfcc.org/about-
516 gtfcc/roadmap-2030/ (2020).
517 52. HDX. The Humanitarian Data Exchange. https://data.humdata.org (2021).
518 53. University of East Anglia. Climate Research Unit. https://www.uea.ac.uk/groups-and-
519 centres/climatic-research-unit (2020).
520 54. CEDA. High resolution Standardized Precipitation Evapotranspiration Index (SPEI) dataset for
521 Africa. https://catalogue.ceda.ac.uk/uuid/bbdfd09a04304158b366777eba0d2aeb (2019).
522 55. IOM. DTM Nigeria. https://displacement.iom.int/nigeria (2021).
523 56. JMP. Nigeria. https://washdata.org (2020).
524 57. WorldBank. Data Bank Subnational Population.
525 https://databank.worldbank.org/source/subnational-population (2021).
526 58. Kamvar, Z. N., Cai, J., Pulliam, J. R. C., Schumacher, J. & Jombart, T. Epidemic curves made
527 easy using the R package incidence https://doi.org/10.12688/f1000research.18002.1 (2019).
528 59. Cori, A. EpiEstim: Estimate Time Varying Reproduction Numbers from Epidemic Curves. R
529 package version 2.2-4. https://CRAN.R-project.org/package=EpiEstim (2021).
530 60. Azman, A. S. et al. Urban cholera transmission hotspots and their implications for reactive
531 vaccination: evidence from Bissau city, Guinea bissau. *PLoS Neglect Trop. Dis.* **6**, e1901 (2012).
532 61. Azman, A. S. et al. Population-level effect of cholera vaccine on displaced populations, South
533 Sudan, 2014. *Emerg. Infect. Dis*. **22**, 1067 (2016).
534 62. Kahn, R. et al. Incubation periods impact the spatial predictability of cholera and Ebola outbreaks
535 in Sierra Leone. *PNAS*. **117**, 5067-73 (2020).
536 63. Hamlet A, Ramos DG, Gaythorpe KA, Romano AP, Garske T, Ferguson NM. Seasonality of
537 agricultural exposure as an important predictor of seasonal yellow fever spillover in Brazil. *Nat.*
538 *Commun*. **12**, 1-1 (2021).

539 64. Kapwata T, Gebreslasie MT. Random forest variable selection in spatial malaria transmission
540 modelling in Mpumalanga Province, South Africa. *Geospat. Health*. **11**, 251-262 (2016).
541 65. Breiman, L. Random forests. *Mach. Learn*. **45**, 5-32 (2001).
542 66. Biau, G. Analysis of a random forests model. *J. Mach. Learn. Res*. **13**, 1063-95 (2012).
543 67. Genuer R, Poggi JM, Tuleau-Malot C. Variable selection using random forests. *Pattern*
544 *Recognit. Lett.* **31**, 2225-36 (2010).
545 68. Kuhn, M. caret: Classification and Regression Training. https://CRAN.R-
546 project.org/package=caret (2021).
547

548 **Acknowledgements**

549 We would like to thank and acknowledgement the Nigeria Centre for Disease Control for providing
550 the data used here and those who work for the NCDC who collected the data in the field. We would
551 also like to thank Anwar Musah (University College London) and Kelly Elimian (Karolinska Institutet)
552 for their guidance on cholera data for Nigeria and facilitating the partnership with NCDC. This work
553 was supported by the Natural Environmental Research Council [NE/S007415/1], as part of the
554 Grantham Institute for Climate Change and the Environment's (Imperial College London) Science
555 and Solutions for a Changing Planet Doctoral Training Partnership. We also acknowledge joint
556 Centre funding from the UK Medical Research Council and Department for International
557 Development [MR/R0156600/1].
558

559 **Author Contributions**

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

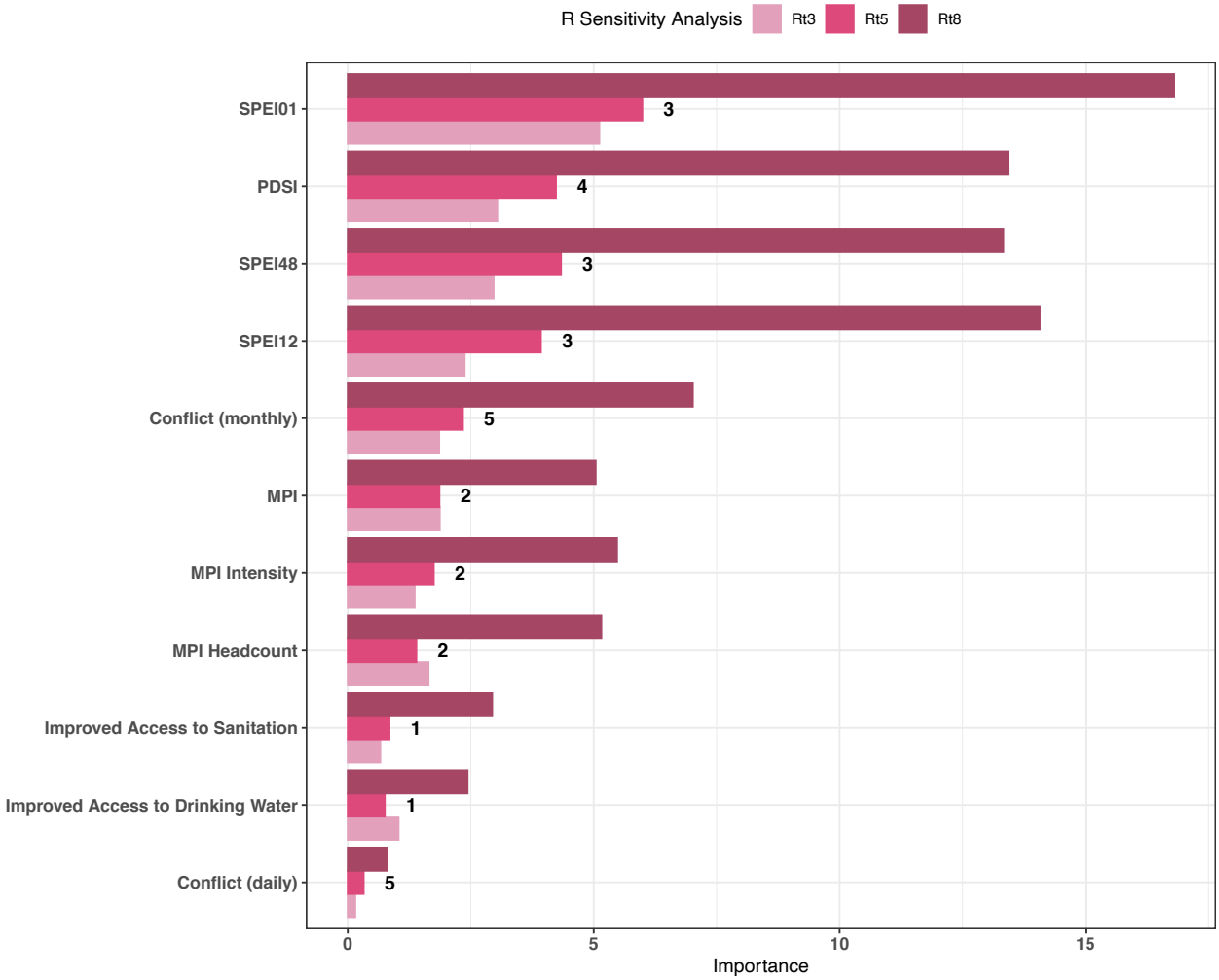
Supplementary Information 1: Additional covariate selection using linear regression

The same 21 covariates (conflict, drought IDPs, WASH, healthcare, population and poverty) analysed using variable importance were also run through an additional covariate selection process and stepwise analysis as developed by:

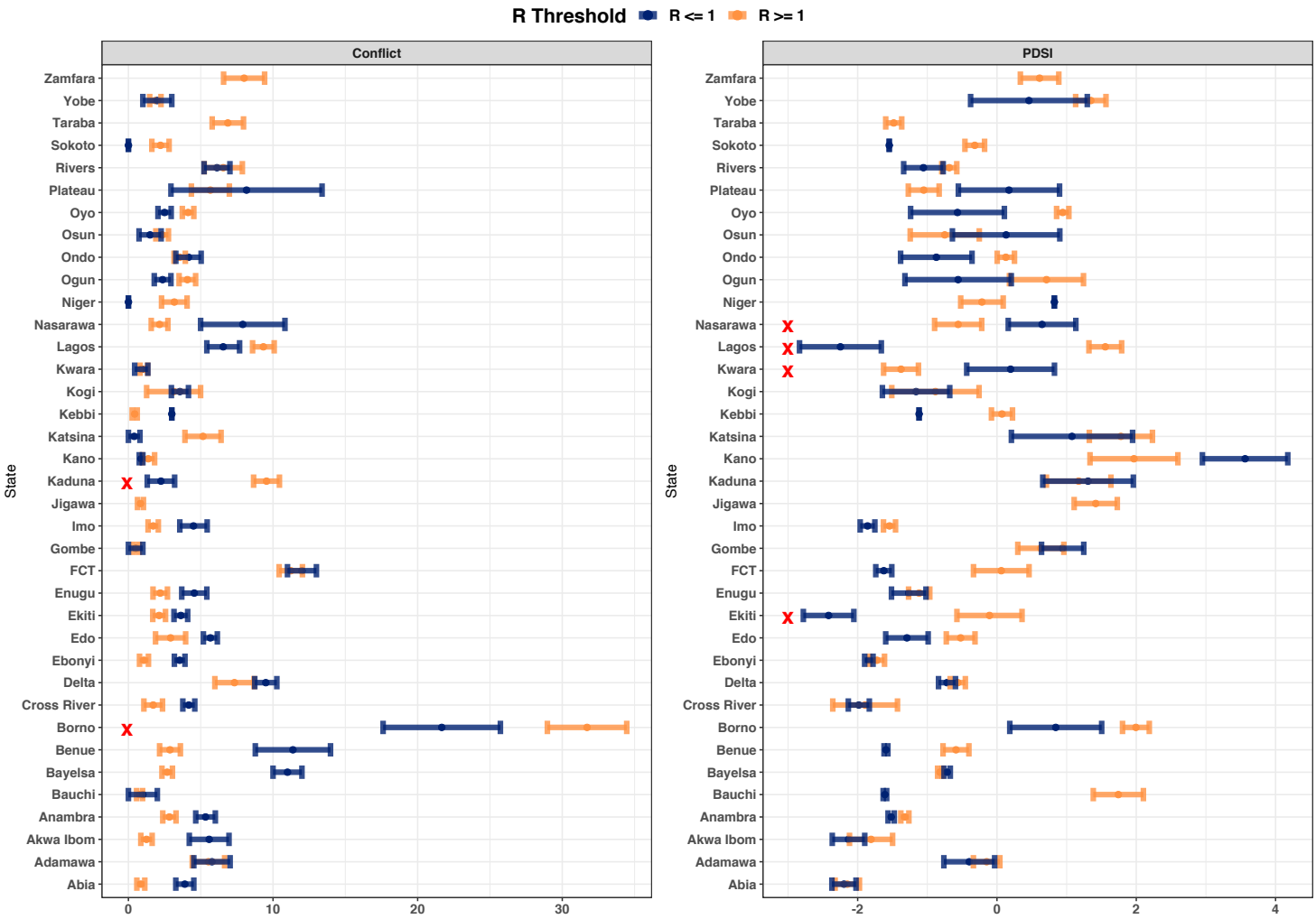
1. Garske, T. et al. Yellow fever in Africa: estimating the burden of disease and impact of mass vaccination from outbreak and serological data. *PLoS Med.* **11**, e1001638 (2014).
2. Gaythorpe, K. A. M. et al. The global burden of yellow fever. *Elife* **10**, e64670 (2021).

The selection process removes covariates that are not significantly associated with the outcome variable (Rt3, Rt5, Rt8) at p = <0.1 using linear regression. It then clusters the remaining covariates based on the correction between them at an absolute pairwise correlation of above 0.75.

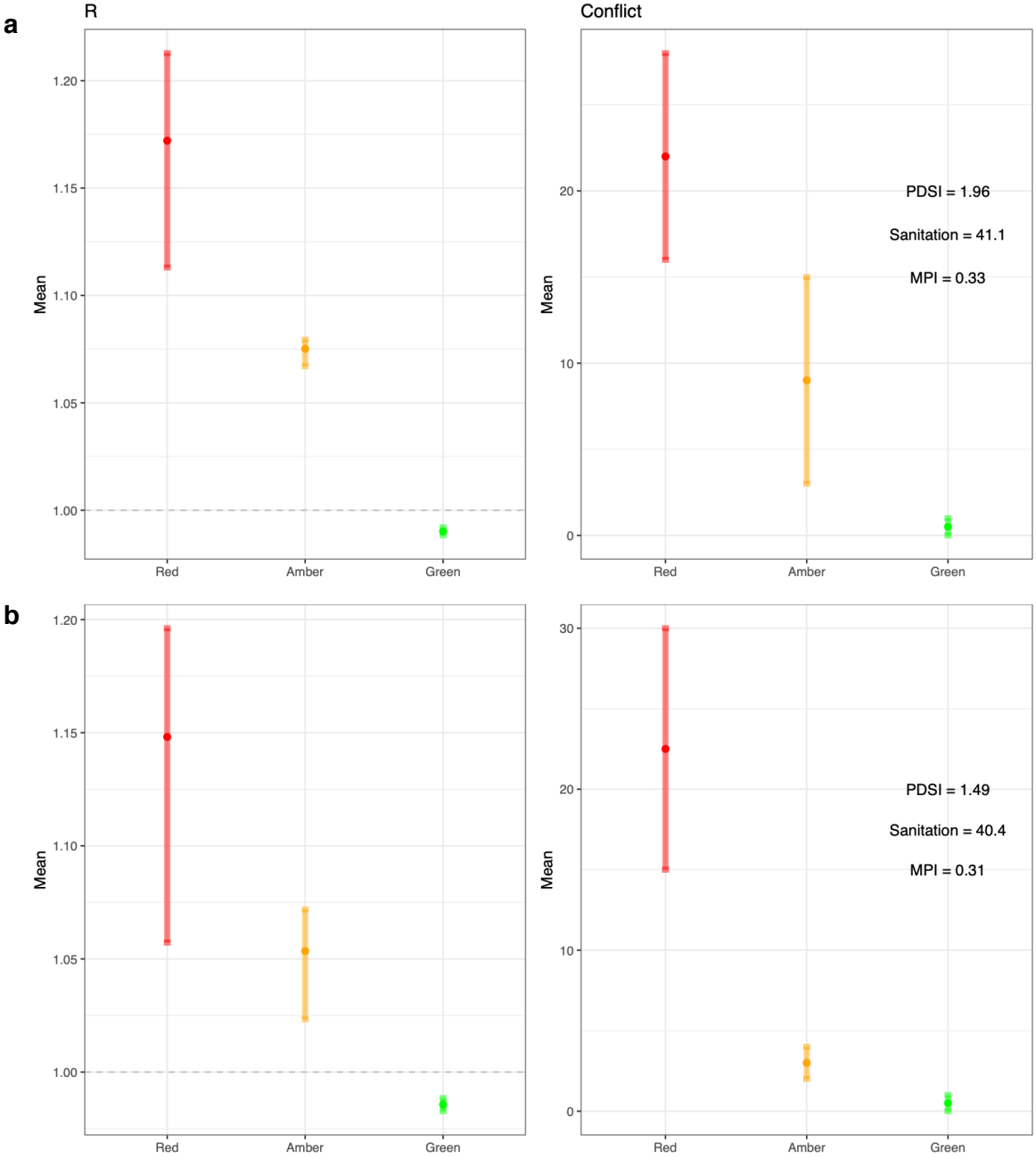
Ten were removed, either because they were not significantly associated with the outcome variable (R) or because they were too highly correlated with other covariates (healthcare facilities, piped water, open defecation, population, IDPs, severe poverty, vulnerable to poverty, basic hygiene). Eleven covariates remained and were grouped into five clusters, the clusters and variable importance of each covariate are shown below



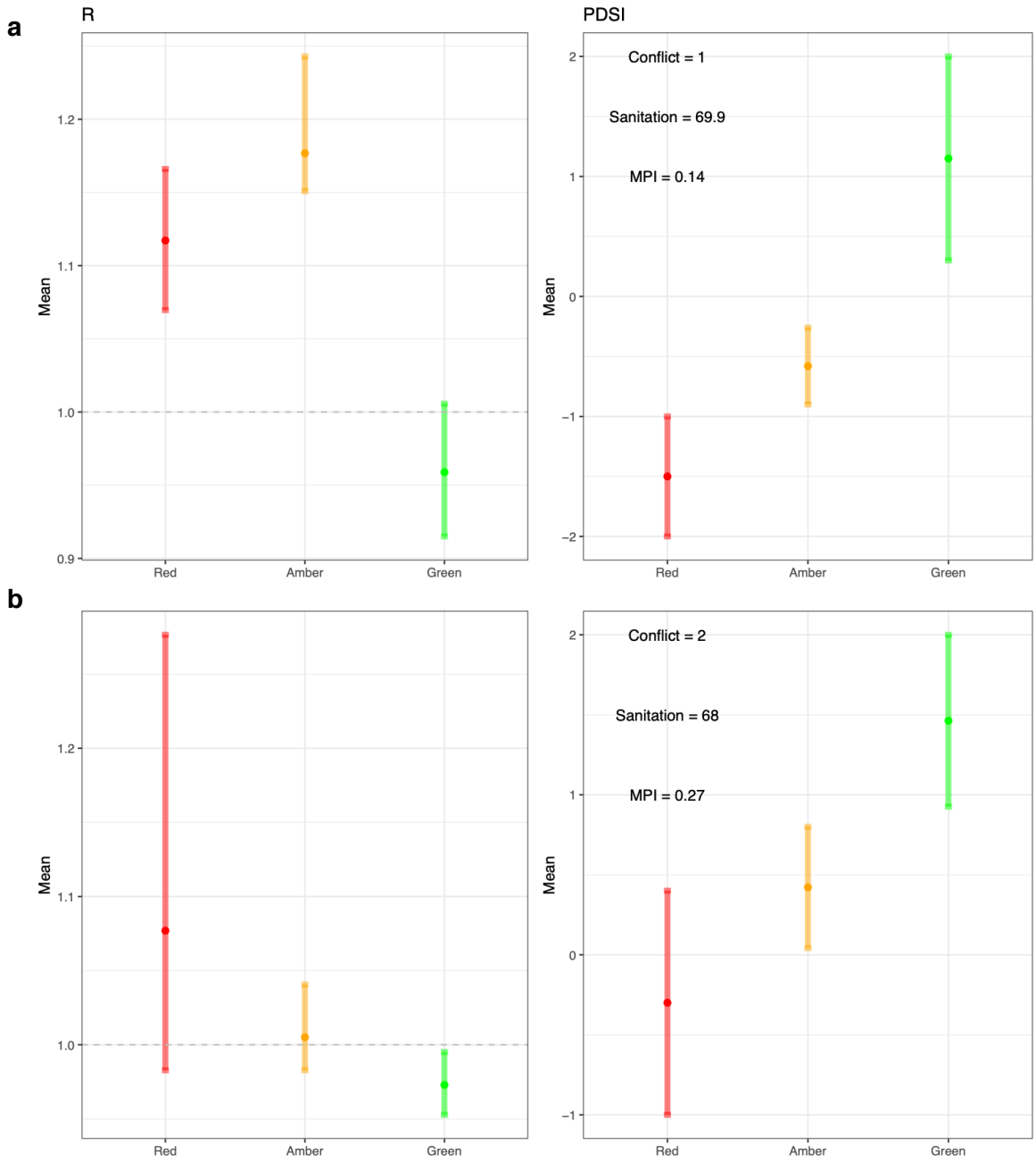
The variable importance for the eleven remaining covariates after variable selection. All three serial interval values tested are shown (Rt3 - 3 days, Rt5 - 5 days, Rt8 - 8 days) and the numbers represent the clusters. SPEI01, 12, 48 - Standardised Precipitation Index calculated on 1, 12 and 48 month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



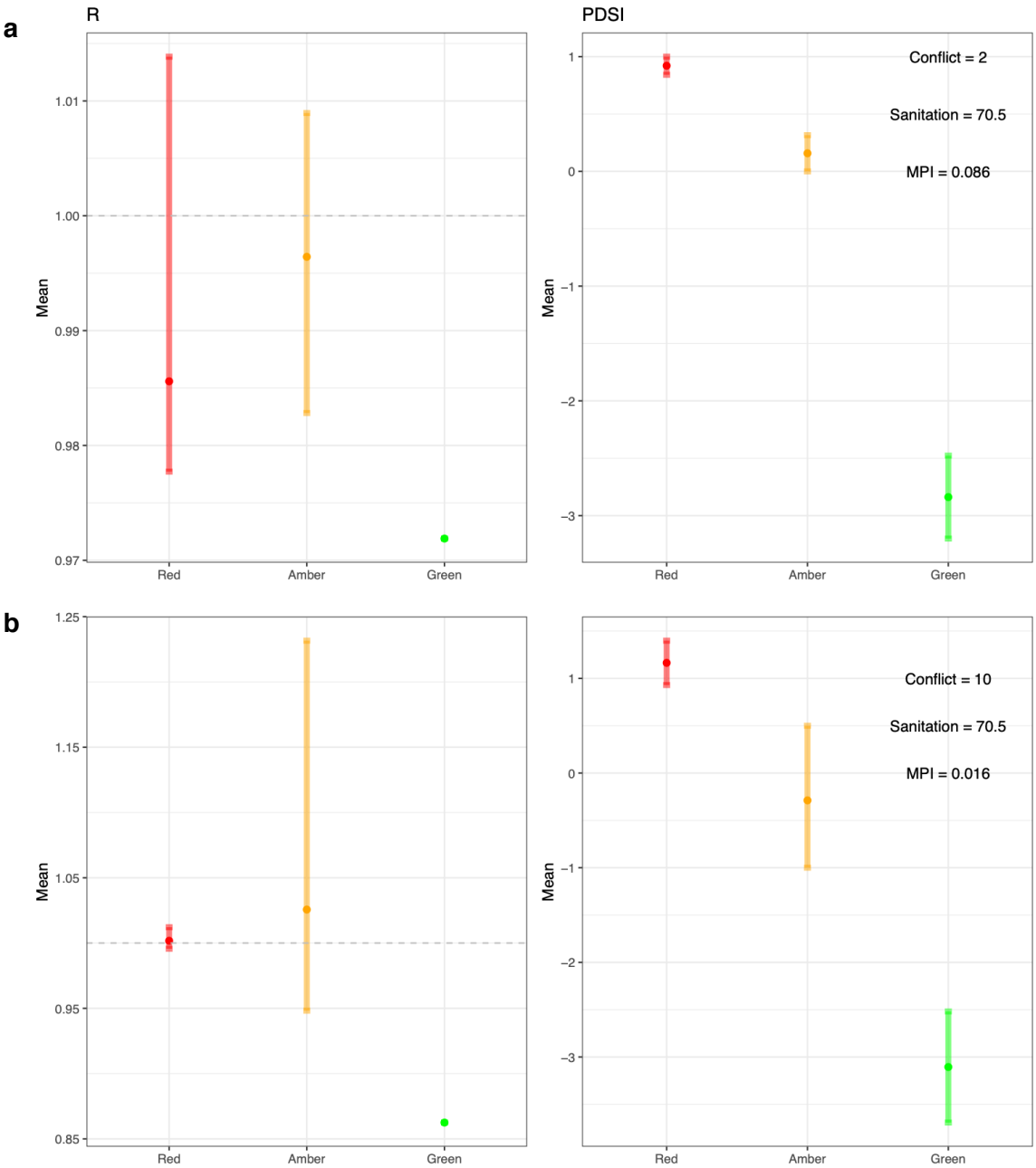
Supplementary Figure 1: Historical spatial trends between the selected social and environmental extremes (conflict and PDSI) and the R thresholds (R = >1, R <1). The mean and standard error for the two covariates for the full dataset split by state and R threshold. The red "x" shows the states which were included in the sub-national analysis: Conflict (Borno and Kaduna), extreme wetness (Lagos and Ekiti), extreme dryness (Nasarawa and Kwara).



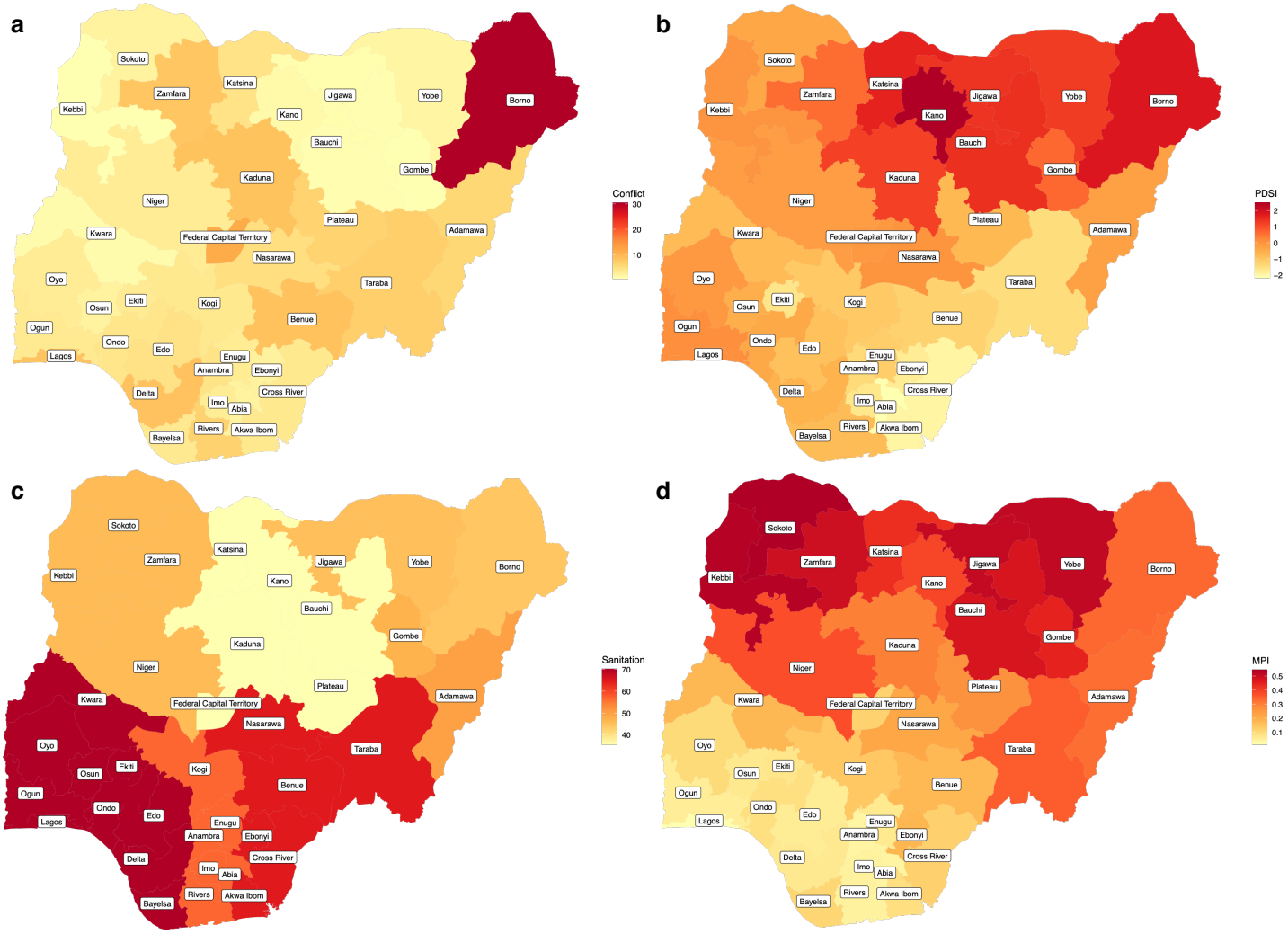
Supplementary Figure 2: Three traffic-light scenarios for conflict only and the corresponding predicted R values. The other three (PDSI, Sanitation and MPI) covariate values were retained at the mean value for R = >1 for the full dataset (values shown in the plot) for **a**, Borno and **b**, Kaduna.



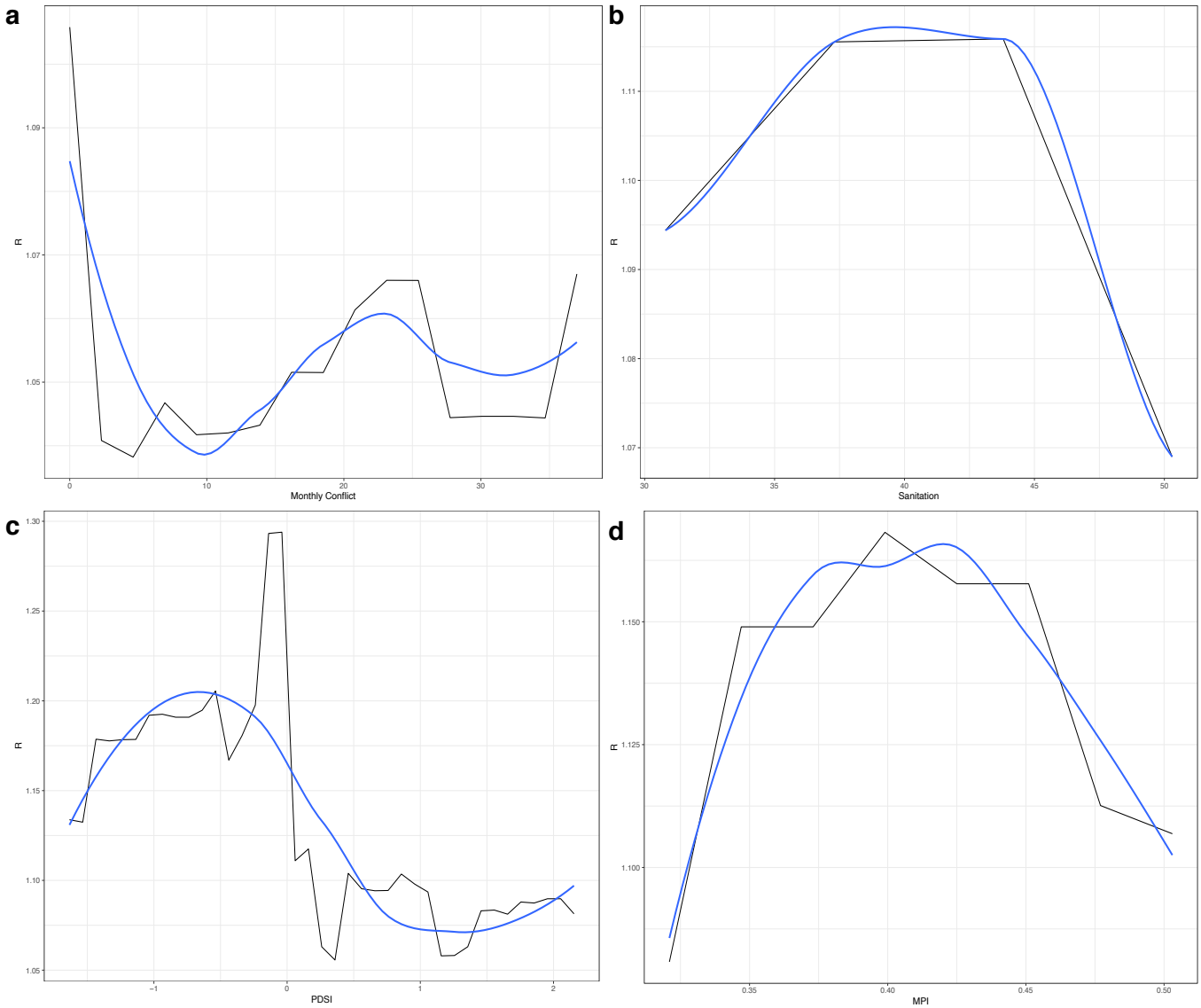
Supplementary Figure 3: Three traffic-light scenarios for PDSI (drier conditions) only and the corresponding predicted R values. The other three (Conflict, Sanitation and MPI) covariate values were retained at the mean value for R = >1 for the full dataset (values shown in the plot) for **a**, Kwara and **b**, Nasarawa.



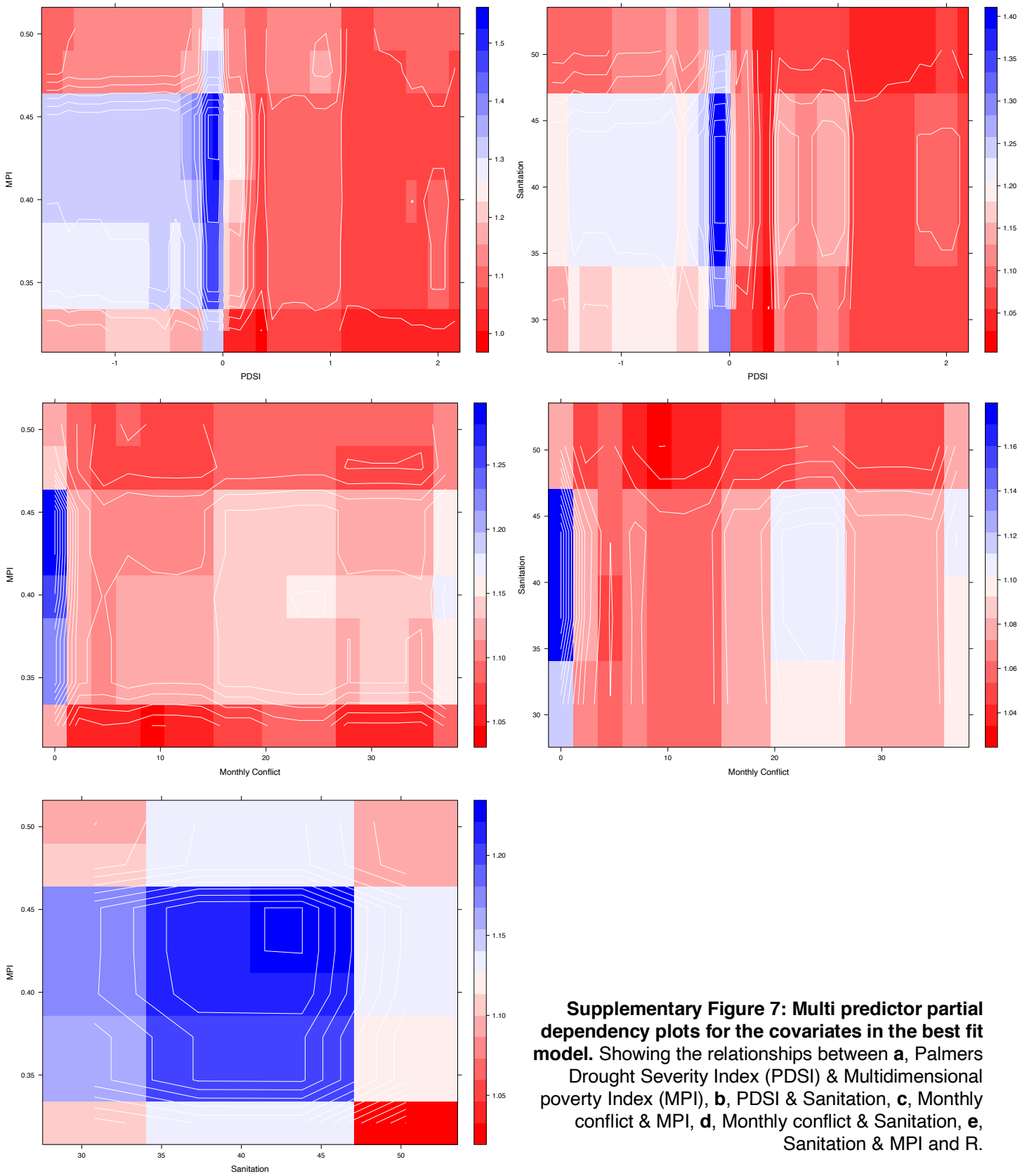
Supplementary Figure 4: Three traffic-light scenarios for PDSI (wetter conditions) only and the corresponding predicted R values. The other three (Conflict, Sanitation and MPI) covariate values were retained at the mean value for R = >1 for the full dataset (values shown in the plot) for **a**, Ekiti and **b**, Lagos.



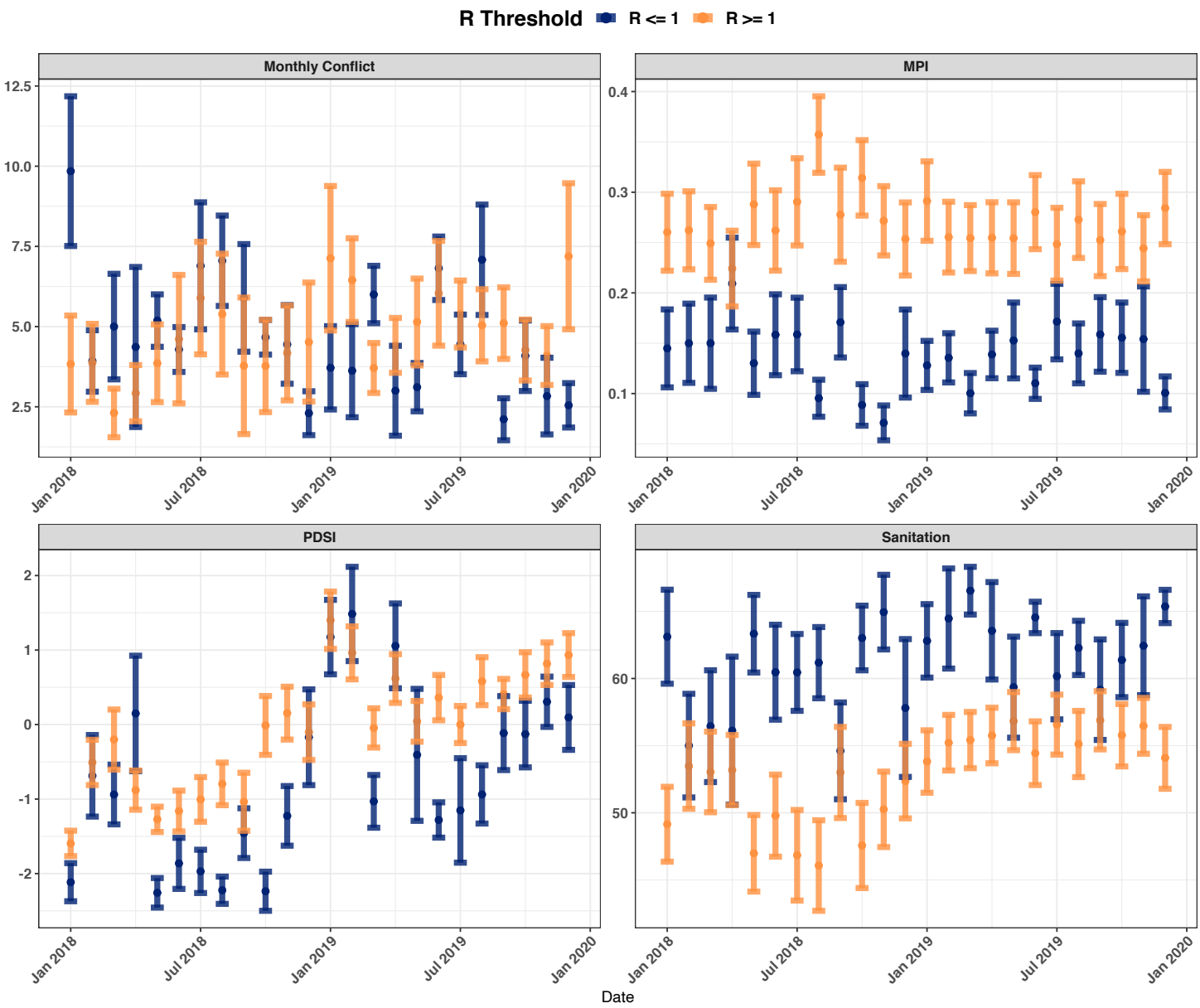
Supplementary Figure 5: Average values of the four covariates included in the best fit model. By state, covariates included: **a**, monthly conflict events, **b**, Palmers Drought Severity Index (PDSI), **c**, percentage access to sanitation and **d**, Multidimensional Poverty Index (MPI).



Supplementary Figure 6: Single predictor partial dependency plots for the covariates in the best fit model. Showing the relationships between **a**, monthly conflict events, **b**, access to sanitation, **c**, Palmers Drought Severity Index (PDSI) and **d**, Multidimensional poverty Index (MPI) and R.



Supplementary Figure 7: Multi predictor partial dependency plots for the covariates in the best fit model. Showing the relationships between **a**, Palmers Drought Severity Index (PDSI) & Multidimensional poverty Index (MPI), **b**, PDSI & Sanitation, **c**, Monthly conflict & MPI, **d**, Monthly conflict & Sanitation, **e**, Sanitation & MPI and R.



Supplementary Figure 8: Historical temporal trends between the best fit model covariates and the R thresholds (R = >1, R <1). The mean and standard error for the four covariates included in the best fit model for the full dataset split by month and R threshold.