MOSAIC: a spatial model of endemic cholera

John R Giles

Contents

Website under development. Last compiled on 2025-05-12 at 05:19 PM PDT.

Welcome

Welcome to the **Metapopulation Outbreak Simulation with Agent-based Implementation for Cholera (MOSAIC)**. The MOSAIC framework simulates the transmission dynamics of cholera in Sub-Saharan Africa (SSA) and provides tools to understand the impact of interventions, such as vaccination, as well as large-scale drivers like climate change. MOSAIC is built using the Light-agent Spatial Model for ERadication (LASER) platform, and this site serves as documentation for the model's methods and associated analyses. Please note that MOSAIC is currently under development, so content may change regularly. We are sharing it here to increase visibility and welcome feedback on any aspect of the model.

Contact

MOSAIC is developed by a team of researchers at the Institute for Disease Modeling (IDM) dedicated to developing modeling methods and software tools that help decision-makers understand and respond to infectious disease outbreaks. This website is currently maintained by John Giles (`@gilesjohnr`). For general questions, contact John Giles (john.giles@gatesfoundation.org), Jillian Gauld (jillian.gauld@gatesfoundation.org), and/or Rajiv Sodhi (rajiv.sodhi@gatesfoundation.org).

Funding

This work was developed at the Institute for Disease Modeling in support of funded research grants made by the Bill & Melinda Gates Foundation.

Chapter 1

Rationale

1.1 Background

The bacterium *Vibrio cholerae* was introduced to the African continent from Asia in 1967 and has since become endemic in many countries in Sub-Saharan Africa (SSA). While sporadic outbreaks have occurred each year over the past six decades, a significant surge in transmission has been observed since 2021, which is consistent with a global increase in cases and deaths during this time. This increase is likely driven by a combination of factors such as climate change, disruptions to municipal services due to conflict, population displacement, and past shortages of the Oral Cholera Vaccine (OCV). Therefore, a spatial model of endemic cholera that accounts for the many drivers of transmission and provides insights into the most impactful interventions will be a valuable tool to support cholera control.

1.2 OCV Stockpiles

Containing cholera transmission relies primarily on improvements to Water, Sanitation, and Hygiene (WASH) and the use of the OCV. However, implementing WASH improvements takes time, and in conflict-affected areas, poor infrastructure often hinders progress. As a result, OCV remains a critical tool for slowing cholera spread in both outbreak and endemic settings.

The depletion of OCV stockpiles in 2024 led to a shift toward single-dose reactive OCV strategies. Increased vaccine production by multiple manufacturers is expected to improve OCV availability in 2025 and beyond. A key question now is how best to allocate OCV through preventative campaigns to reduce transmission and support the goal of reducing cholera deaths by 90% by 2030, led by the WHO Global Task Force for Cholera Control (GTFCC). The MOSAIC framework is designed to maximize the impact of regional preventative OCV

strategies by assessing country prioritization, OCV dosing schedules, and overall OCV demand.

1.3 Impacts of Climate Change

Environmental factors play a crucial role in cholera outbreaks, with extreme weather events creating local conditions that foster *V. cholerae* transmission. Models incorporating climate change can provide valuable insights into future cholera dynamics by accounting for environmentally forced transmission, which enables more accurate forecasts and scenarios that will contribute to achieving the GTFCC goal. The MOSAIC framework leverages Artificial Intelligence (AI) models and global climate model projections to predict climate change's impact on cholera transmission and generate mid-term forecasts of high-risk areas across SSA.

1.4 Data and Modeling

A significant challenge in controlling cholera transmission in SSA is the lack of comprehensive data sets and dynamic models designed to support ongoing policy-making. The persistent endemic nature of cholera in SSA presents a complex quantitative challenge, requiring sophisticated models to produce meaningful inferences. Models that incorporate the necessary natural history and disease dynamics, and operate at adequate spatial and temporal scales, are crucial for providing timely and actionable information to address ongoing and future cholera outbreaks.

Although developing data and models at these scales is challenging, our goal with the MOSAIC framework is to create a landscape-scale transmission model for cholera in SSA (first at the country-level and then later at the district-level). Models are being developed with an array of historical and real-time data sources that include incidence and mortality reports, patterns of human movement, WASH, OCV campaigns, and climate variables.

Key questions we aim to address include when and where to administer a limited supply of oral cholera vaccine (OCV) and how severe weather events and climate change will impact future outbreaks. Our approach will include 5-month forecasts and long-range OCV scenarios out to 2030. Model outputs will be updated biweekly or monthly based on data availability.

Chapter 2

Data

The MOSAIC model requires a diverse set of data sources, some of which are directly used to define model parameters (e.g., birth and death rates), while others help fit models a priori and provide informative priors for the transmission model. As additional data sources become available, future versions of the model will adapt to incorporate them. For now, the following data sources represent the minimum requirements to initiate a viable first model.

2.1 Historical Incidence and Deaths

Data on historical cholera incidence and deaths are crucial for establishing baseline transmission patterns. We compiled the annual total reported cases and deaths for all AFRO region countries from January 1970 to August 2024. These data comes from several sources which include:

1. **Our World in Data (1970-2021)**: Number of Reported Cases of Cholera (1949-2021) and the Number of Reported Deaths of Cholera from (1949-2021). The Our World in Data group compiled these data from previously published annual WHO reports.
2. **WHO Annual Report 2022**: These data were manually extracted from the World Health Organization's Weekly Epidemiological Record No 38, 2023, 98, 431–452.
3. **Global Cholera and Acute Watery Diarrhea Dashboard (2023-2024)**: Unofficial tallies of reported cases and deaths for 2023 and part of 2024 are available at the WHO Global Cholera and AWD Dashboard.

2.2 Recent Incidence and Deaths

To capture recent cholera trends, we retrieved reported cases and deaths data from the WHO Global Cholera and Acute Watery Diarrhea Dashboard REST

API. These data provide weekly incidence and deaths from January 2023 to August 2024 which provides up-to-date counts at the country level.

2.3 Vaccinations

Accurate data on oral cholera vaccine (OCV) campaigns and vaccination history are vital for understanding the impact of vaccination efforts. These data come from:

- **WHO Cholera Vaccine Dashboard**: This resource (link) provides detailed information on OCV distribution and vaccination campaigns from 2016 to 2024.
- **GTFCC OCV Dashboard**: Managed by Médecins Sans Frontières, this dashboard (link) tracks OCV deployments globally, offering granular insights into vaccination efforts from 2013 to 2024.

2.4 Human Mobility Data

Human mobility patterns significantly influence cholera transmission. Relevant data include:

- **OAG Passenger Booking Data**: This dataset (link) offers insights into air passenger movements, which can be used to model the spread of cholera across regions.
- **Namibia Call Data Records**: An additional source from Giles et al. (2020) (link) provides detailed mobility data based on mobile phone records, useful for localized modeling.

2.5 Climate Data

Climate conditions, including temperature, precipitation, and extreme weather events, play a critical role in cholera dynamics. These are captured through:

- **OpenMeteo Historical Weather Data API**: This API (link) offers access to historical climate data, which is essential for modeling the environmental factors influencing cholera outbreaks.

2.5.1 Storms and Floods

Data on extreme weather events, specifically storms and floods, are obtained from:

- **EM-DAT International Disaster Database**: Maintained by the Centre for Research on the Epidemiology of Disasters (CRED) at UCLouvain, this database (link) provides comprehensive records of disasters from 2000 to the present, including those affecting African countries.

2.6 WASH (Water, Sanitation, and Hygiene)

Data on water, sanitation, and hygiene (WASH) are critical for understanding the environmental and infrastructural factors that influence cholera transmission. These data are sourced from:

- **WHO UNICEF Joint Monitoring Program (JMP) Database**: This resource (link) offers detailed information on household-level access to clean water and sanitation, which is integral to cholera prevention efforts.

2.7 Demographics

Demographic data, including population size, birth rates, and death rates, are foundational for accurate disease modeling. These data are sourced from:

- **UN World Population Prospects 2024**: This database (link) provides probabilistic projections of key demographic metrics, essential for estimating population-level impacts of cholera.

Chapter 3

Model description

Here we describe the methods of MOSAIC version 1.0. This model version provides a starting point for understanding cholera transmission in Sub-Saharan Africa, incorporating important drivers of disease dynamics such as human mobility, environmental conditions, and vaccination schedules. As MOSAIC continues to evolve, future iterations will refine model components based on available data and improved model mechanisms, which we hope will increase its applicability to real-world scenarios.

The model operates on daily time steps and will be fitted to historical incidence data, however current development is based on data from January 2023 to August 2024 and includes 40 countries in Sub-Saharan Africa (SSA), see Figure 3.1 and the Table of MOSAIC framework countries.

3.1 Transmission dynamics

The model has a metapopulation structure with familiar compartments for Susceptible, Exposed, Infected, and Recovered individuals with SEIRS dynamics. The model also contains compartments for one- and two-dose vaccination (V_1 and V_2) and Water & environment based transmission (W) which we refer to as SVEIWRS.

The SVEIWRS metapopulation model, shown in Figure 3.2, is governed by the following difference equations:

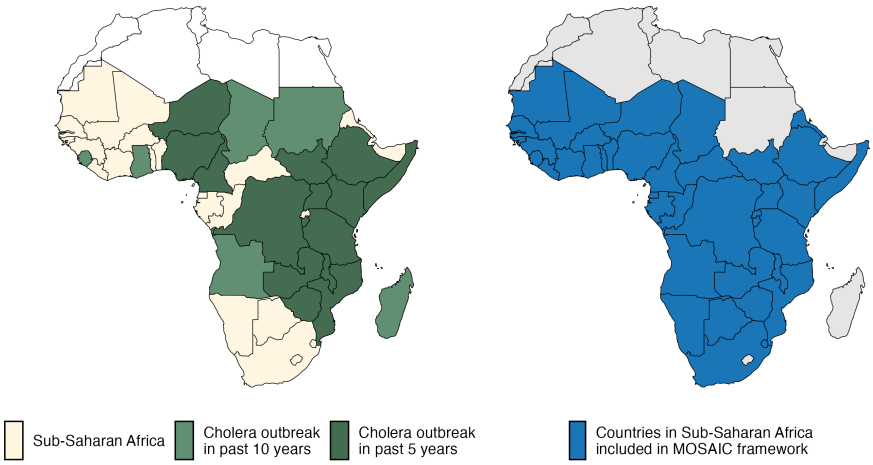

Figure 3.1: A map of Sub-Saharan Africa with countries that have experienced a cholera outbreak in the past 5 and 10 years highlighted in green. The 40 countires included in the MOSAIC modeling framework are indicated in blue.

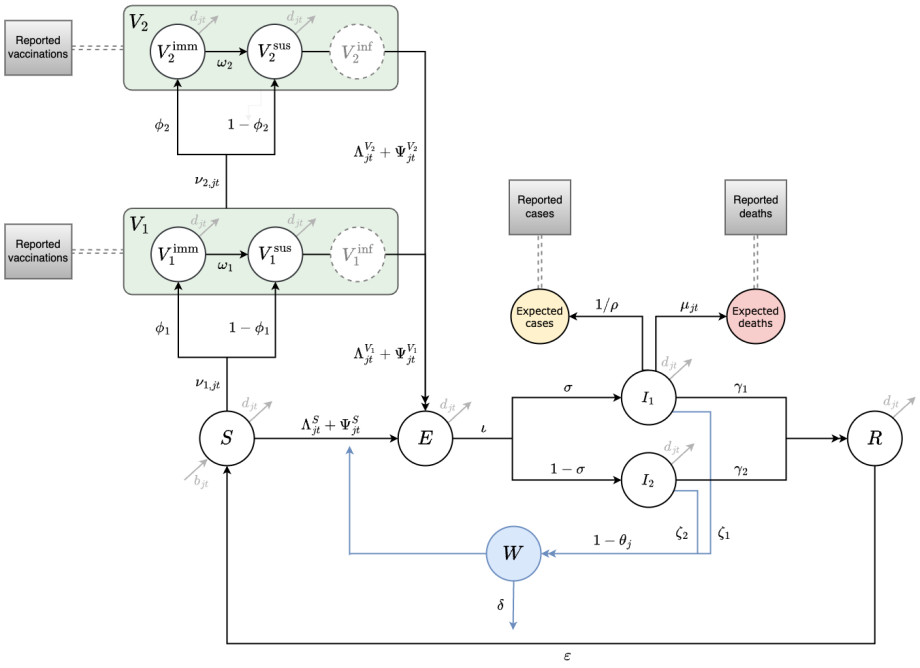

Figure 3.2: This diagram of the SVEIWRS (Susceptible-Vaccinated-Exposed-Infected-Water/environmental-Recovered-Susceptible) model shows model compartments as circles with rate parameters displayed. The primary data sources the model is fit to are shown as square nodes (Vaccination data, and reported cases and deaths).

Susceptible population:

$$S_{j,t+1} = S_{jt} + b_{jt}\, N_{jt} + \varepsilon\, R_{jt} - \frac{\nu_{1,jt}\, S_{jt}}{\left(S_{jt} + E_{jt}\right)} - \left(\Lambda^{S}_{j,t+1} + \Psi^{S}_{j,t+1}\right) - d_{jt}\, S_{jt}$$

One-dose vaccination:

$$V^{\text{imm}}_{1,j,t+1} = V^{\text{imm}}_{1,jt} + \frac{\phi_1\, \nu_{1,jt}\, S_{jt}}{\left(S_{jt} + E_{jt}\right)} - \omega_1\, V^{\text{imm}}_{1,jt} - \frac{\nu_{2,jt}\, V^{\text{imm}}_{1,jt}}{\left(V^{\text{imm}}_{1,jt} + V^{\text{sus}}_{1,jt}\right)} - d_{jt}\, V^{\text{imm}}_{1,jt}$$

$$V^{\text{sus}}_{1,j,t+1} = V^{\text{sus}}_{1,jt} + \frac{(1-\phi_1)\, \nu_{1,jt}\, S_{jt}}{\left(S_{jt} + E_{jt}\right)} + \omega_1\, V^{\text{imm}}_{1,jt} - \left(\Lambda^{V_1}_{j,t+1} + \Psi^{V_1}_{j,t+1}\right) - \frac{\nu_{2,jt}\, V^{\text{sus}}_{1,jt}}{\left(V^{\text{imm}}_{1,jt} + V^{\text{sus}}_{1,jt}\right)} - d_{jt}\, V$$

$$V^{\text{inf}}_{1,j,t+1} = V^{\text{inf}}_{1,jt} + \left(\Lambda^{V_1}_{j,t+1} + \Psi^{V_1}_{j,t+1}\right) - d_{jt}\, V^{\text{inf}}_{1,jt} \quad \text{(tracking only)}$$

Two-dose vaccination:

$$V^{\text{imm}}_{2,j,t+1} = V^{\text{imm}}_{2,jt} + \phi_2\, \nu_{2,jt} - \omega_2\, V^{\text{imm}}_{2,jt} - d_{jt}\, V^{\text{imm}}_{2,jt}$$

$$V^{\text{sus}}_{2,j,t+1} = V^{\text{sus}}_{2,jt} + (1-\phi_2)\, \nu_{2,jt} + \omega_2\, V^{\text{imm}}_{2,jt} - \left(\Lambda^{V_2}_{j,t+1} + \Psi^{V_2}_{j,t+1}\right) - d_{jt}\, V^{\text{sus}}_{2,jt}$$

$$V^{\text{inf}}_{2,j,t+1} = V^{\text{inf}}_{2,jt} + \left(\Lambda^{V_2}_{j,t+1} + \Psi^{V_2}_{j,t+1}\right) - d_{jt}\, V^{\text{inf}}_{2,jt} \quad \text{(tracking only)}$$

Infection dynamics:

$$E_{j,t+1} = E_{jt} + \left(\Lambda_{j,t+1} + \Psi_{j,t+1}\right) - \iota\, E_{jt} - d_{jt}\, E_{jt}$$

$$I_{1,j,t+1} = I_{1,jt} + \sigma\, \iota\, E_{jt} - \gamma_1\, I_{1,jt} - \mu_j\, I_{1,jt} - d_{jt}\, I_{1,jt}$$

$$I_{2,j,t+1} = I_{2,jt} + (1-\sigma)\, \iota\, E_{jt} - \gamma_2\, I_{2,jt} - d_{jt}\, I_{2,jt}$$

$$R_{j,t+1} = R_{jt} + \left(\gamma_1\, I_{1,jt} + \gamma_2\, I_{2,jt}\right) - \varepsilon\, R_{jt} - d_{jt}\, R_{jt}$$

Environment:

$$W_{j,t+1} = W_{jt} + \left(1-\theta_j\right)\left(\zeta_1\, I_{1,jt} + \zeta_2\, I_{2,jt}\right) - \delta_{jt}\, W_{jt} \tag{3.1}$$

For detailed descriptions of all parameters appearing in Equation (3.1), see the Table of model parameters. Transmission dynamics in the model are governed primarily by two distinct force-of-infection terms: the human-to-human force of infection, Λ_{jt}, and the environmental force of infection, Ψ_{jt}.

The human-to-human transmission component at time $t+1$ in location j is defined separately for susceptible (S), one-dose vaccinated (V_1), and two-dose vaccinated (V_2) individuals as:

$$\Lambda^{S}_{j,t+1} = \frac{\beta^{\text{hum}}_{jt}\,(1-\tau_j)S_{jt}\,\left[(1-\tau_j)(I_{1,jt}+I_{2,jt}) + \sum_{\forall i\neq j}\pi_{ij}\tau_i(I_{1,jt}+I_{2,jt})\right]^{\alpha_1}}{N^{\alpha_2}_{jt}},$$

$$\Lambda^{V_1}_{j,t+1} = \frac{\beta^{\text{hum}}_{jt}\,(1-\tau_j)V^{\text{sus}}_{1,jt}\,\left[(1-\tau_j)(I_{1,jt}+I_{2,jt}) + \sum_{\forall i\neq j}\pi_{ij}\tau_i(I_{1,jt}+I_{2,jt})\right]^{\alpha_1}}{N^{\alpha_2}_{jt}},$$

$$\Lambda^{V_2}_{j,t+1} = \frac{\beta^{\text{hum}}_{jt}\,(1-\tau_j)V^{\text{sus}}_{2,jt}\,\left[(1-\tau_j)(I_{1,jt}+I_{2,jt}) + \sum_{\forall i\neq j}\pi_{ij}\tau_i(I_{1,jt}+I_{2,jt})\right]^{\alpha_1}}{N^{\alpha_2}_{jt}}. \tag{3.2}$$

The total human-to-human force of infection is then the sum of these three components:

$$\Lambda_{j,t+1} = \Lambda^{S}_{j,t+1} + \Lambda^{V_1}_{j,t+1} + \Lambda^{V_2}_{j,t+1}. \tag{3.3}$$

In these equations, β^{hum}_{jt} represents the rate of human-to-human transmission. Movement within and among metapopulations is governed by the parameter τ_i, indicating the probability of departing origin location i, while π_{ij} describes the relative probability of travel from origin i to destination j (see section on spatial dynamics). The terms Λ^{S}_{jt}, $\Lambda^{V_1}_{jt}$, and $\Lambda^{V_2}_{jt}$ explicitly partition the overall human-to-human force of infection into separate contributions from susceptible, one-dose vaccinated, and two-dose vaccinated individuals, linking directly to the compartmental structure of the model described by the system of difference equations.

The environmental force of infection (Ψ_{jt}), capturing environment-to-human transmission at location j and time $t+1$, is also explicitly partitioned into susceptible (S), one-dose vaccinated (V_1), and two-dose vaccinated (V_2) compartments:

$$\Psi^{S}_{j,t+1} = \frac{\beta^{\text{env}}_{jt}\,(1-\tau_j) S_{jt}\,(1-\theta_j) W_{jt}}{\kappa + W_{jt}},$$

$$\Psi^{V_1}_{j,t+1} = \frac{\beta^{\text{env}}_{jt}\,(1-\tau_j) V^{\text{sus}}_{1,jt}\,(1-\theta_j) W_{jt}}{\kappa + W_{jt}}, \tag{3.4}$$

$$\Psi^{V_2}_{j,t+1} = \frac{\beta^{\text{env}}_{jt}\,(1-\tau_j) V^{\text{sus}}_{2,jt}\,(1-\theta_j) W_{jt}}{\kappa + W_{jt}}.$$

The total environmental force of infection is then the sum of these three components:

$$\Psi_{j,t+1} = \Psi^{S}_{j,t+1} + \Psi^{V_1}_{j,t+1} + \Psi^{V_2}_{j,t+1}. \tag{3.5}$$

Here, β^{env}_{jt} denotes the rate of environment-to-human transmission, and θ_j is the proportion of the population at location j with at least basic access to Water, Sanitation, and Hygiene (WASH). The environmental exposure is scaled by the concentration of *V. cholerae* (cells per mL) associated with a 50% probability of infection (Fung 2014). Additional details regarding environmental compartments, water reservoirs, and climatic factors influencing transmission

Note that all model processes are stochastic. Transition rates are converted to probabilities with the commonly used method based on the exponential waiting time distribution $p(t) = 1 - e^{-rt}$ (see Ross 2007). Integer quantities are thus moved between model compartments at each time step according to a binomial process similar to the recovery of infected individuals γI_{jt}:

$$\frac{\partial R}{\partial t} \sim \text{Binom}(I_{jt}, 1 - e^{-\gamma}). \tag{3.6}$$

For a detailed list of all stochastic transitions in the model, see the Table of stochastic transitions below.

3.2 Latency

An important feature of the SVEIWRS model is the inclusion of an exposed compartment (E) , which captures the latent period between exposure and the onset of infectiousness. In our model, individuals who become infected first enter the E compartment, where they remain for a period governed by the incubation period ι, before progressing to the infectious compartments I_1 (severe symptomatic infection) or I_2 (mild and/or asymptomatic infection).

A systematic review by Azman et al (2013) estimated the median incubation period for cholera to be approximately 1.4 days (1.3–1.6 95%CI). This relatively

short latency is one of the key characteristic governing cholera dynamics and is critical for accurately capturing the rapid spatial spread observed during outbreaks.

3.3 Seasonality

Cholera transmission is seasonal and is typically associated with the rainy season, so both transmission rate terms β^*_{jt} are temporally forced. For human-to-human transmission we used a sinusoidal mechanism as in Altizer et al 2006. Specifically, the function is a truncated sine-cosine form of the Fourier series with two harmonic features which has the flexibility to capture seasonal transmission dynamics driven by extended rainy seasons and/or biannual trends:

$$\beta^{\text{hum}}_{jt} = \beta^{\text{hum}}_{j0}\left(1 + a_1\cos\left(\frac{2\pi t}{p}\right) + b_1\sin\left(\frac{2\pi t}{p}\right) + a_2\cos\left(\frac{4\pi t}{p}\right) + b_2\sin\left(\frac{4\pi t}{p}\right)\right). \tag{3.7}$$

Where, β^{hum}_{j0} is the mean human-to-human transmission rate at location j over all time steps. Seasonal dynamics are determined by the parameters a_1, b_1 and a_2, b_2 which gives the amplitude of the first and second waves respectively. The periodic cycle p is 365, so the function controls the temporal variation in β^{hum}_{jt} over each day of the year.

We estimated the parameters in the Fourier series (a_1, b_1, a_2, b_2) using the Levenberg–Marquardt algorithm in the `minpack.lm` R library. Given the lack of reported cholera case data for many countries in SSA and the association between cholera transmission and the rainy season, we leveraged seasonal precipitation data to help fit the Fourier wave function to all countries. We first gathered weekly precipitation values from 1994 to 2024 for 30 uniformly distributed points within each country from the Open-Meteo Historical Weather Data API. Then we fit the Fourier series to the weekly precipitation data and used these parameters as the starting values when fitting the model to the more sparse cholera case data.

For countries with no reported case data, we inferred seasonal dynamics using the fitted wave function of a neighboring country with available case data. The selected neighbor was chosen from the same cluster of countries (grouped hierarchically into four clusters based on precipitation seasonality using Ward's method; see Figure 3.4) that had the highest correlation in seasonal precipitation with the country lacking case data. In the rare event that no country with reported case data was found within the same seasonal cluster, we expanded the search to the 10 nearest neighbors and continued expanding by adding the next nearest neighbor until a match was found.

Using the model fitting methods described above, and the cluster-based approach for inferring the seasonal Fourier series pattern in countries without

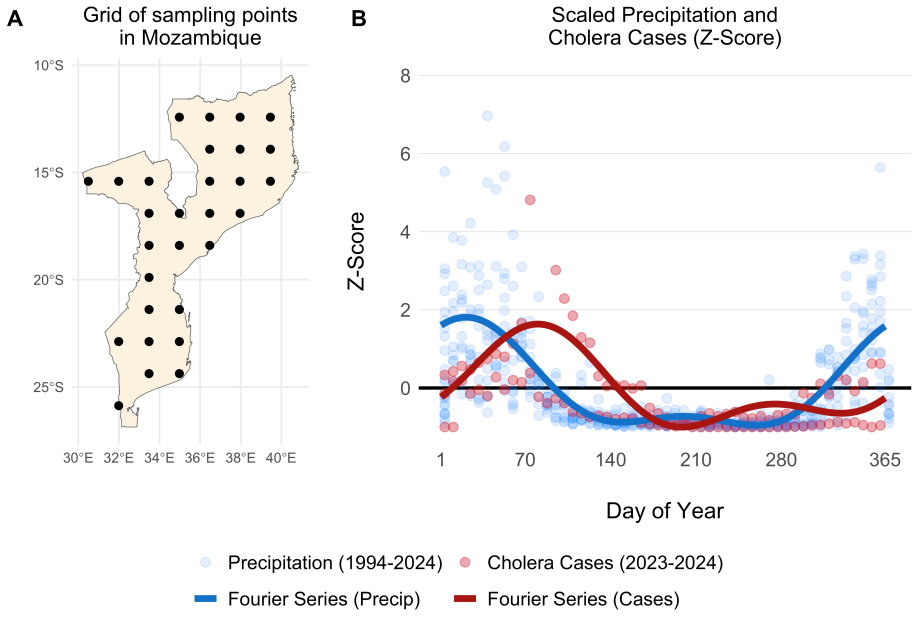

Figure 3.3: Example of a grid of 30 uniformly distributed points within Mozambique (A). The scatterplot shows weekly summed precipitation values at those 30 grid points and cholera cases plotted on the same scale of the Z-Score which shows the variance around the mean in terms of the standard deviation. Fitted Fourier series fucntions are shown as blue (fit precipitation data) and red (fit to cholera case data) lines.

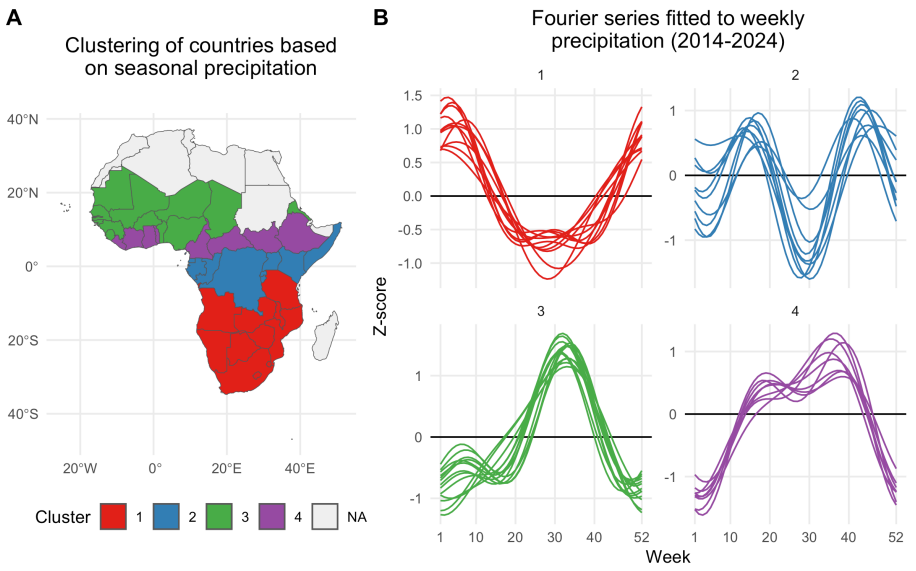

Figure 3.4: A) Map showing the clustering of African countries based on their seasonal precipitation patterns (2014-2024). Countries are colored according to their cluster assignments, identified using hierarchical clustering. B) Fourier series fitted to weekly precipitation for each country. Each line plot shows the seasonal pattern for countries within a given cluster. Clusteres are used to infer the seasonal transmission dynamics for countries where there are no reported cholera cases.

reported cholera cases, we modeled the seasonal dynamics for all 40 countries in the MOSAIC framework. These dynamics are visualized in Figure 3.5, with the corresponding Fourier model coefficients presented in Table 3.1.

Table 3.1: Estimated coefficients for the truncated Fourier model in Equation eqrefeq:beta1 fit to countries with reported cholera cases. Model fits are shown in Figure
reffig:seasonal-all.

	Fourier Coefficients			
Country	a_1	a_2	b_1	b_2
Burundi	-0.42 (-0.52 to -0.32)	-0.3 (-0.4 to -0.21)	-0.06 (-0.16 to 0.04)	-0.22
Cameroon	-0.97 (-1.15 to -0.78)	-0.08 (-0.26 to 0.1)	0.44 (0.27 to 0.62)	-0.71
DRC	0.01 (-0.03 to 0.05)	-0.08 (-0.12 to -0.04)	0.23 (0.19 to 0.27)	-0.04
Ethiopia	-0.42 (-0.47 to -0.36)	-0.12 (-0.17 to -0.06)	0.22 (0.16 to 0.27)	0.05
Ghana	-0.71 (-1.7 to 0.27)	1.31 (0.47 to 2.15)	-0.29 (-0.95 to 0.37)	-1.17
Kenya	0.12 (-0.08 to 0.31)	-0.28 (-0.48 to -0.08)	0.93 (0.73 to 1.13)	0.25
Malawi	1.29 (1.06 to 1.53)	0.29 (0.05 to 0.53)	1.11 (0.87 to 1.35)	1.23
Mozambique	0.4 (0.23 to 0.57)	-0.7 (-0.87 to -0.53)	1.2 (1.03 to 1.37)	0.19
Niger	2.95 (1.4 to 4.51)	-3.42 (-4.63 to -2.2)	1.79 (0.8 to 2.79)	1.72
Nigeria	-0.25 (-0.39 to -0.11)	-0.3 (-0.43 to -0.16)	-0.91 (-1.05 to -0.77)	0.17
Somalia	-0.22 (-0.28 to -0.17)	-0.24 (-0.29 to -0.18)	0.91 (0.86 to 0.97)	-0.37
South Africa	-2.33 (-3.43 to -1.22)	1.06 (0.06 to 2.07)	-2.72 (-3.74 to -1.71)	3.23
South Sudan	1.01 (0.73 to 1.29)	1.54 (1.3 to 1.78)	0.18 (-0.05 to 0.41)	0.02
Tanzania	0.49 (0.37 to 0.61)	-0.13 (-0.25 to -0.02)	-0.6 (-0.71 to -0.48)	-0.29
Togo	1.11 (0.77 to 1.46)	0.02 (-0.33 to 0.36)	-1 (-1.36 to -0.63)	-0.91
Uganda	-0.17 (-0.56 to 0.22)	0.52 (0.13 to 0.9)	0.6 (0.22 to 0.99)	0.42
Zambia	1.53 (1.29 to 1.77)	0.88 (0.64 to 1.12)	0.67 (0.44 to 0.91)	0.78
Zimbabwe	0.99 (0.87 to 1.11)	0.34 (0.23 to 0.46)	0.54 (0.42 to 0.65)	0.12

3.4 Environmental transmission

Environmental transmission is a critical factor in cholera spread and consists of several key components: the rate at which infected individuals shed *V. cholerae* into the environment, the pathogen's survival rate in environmental conditions, and the overall suitability of the environment for sustaining the bacteria over time.

To capture the impacts of climate-drivers on cholera transmission, we have included the parameter ψ_{jt}, which represents the current state of environmental suitability with respect to: *i*) the survival time of *V. cholerae* in the environ-

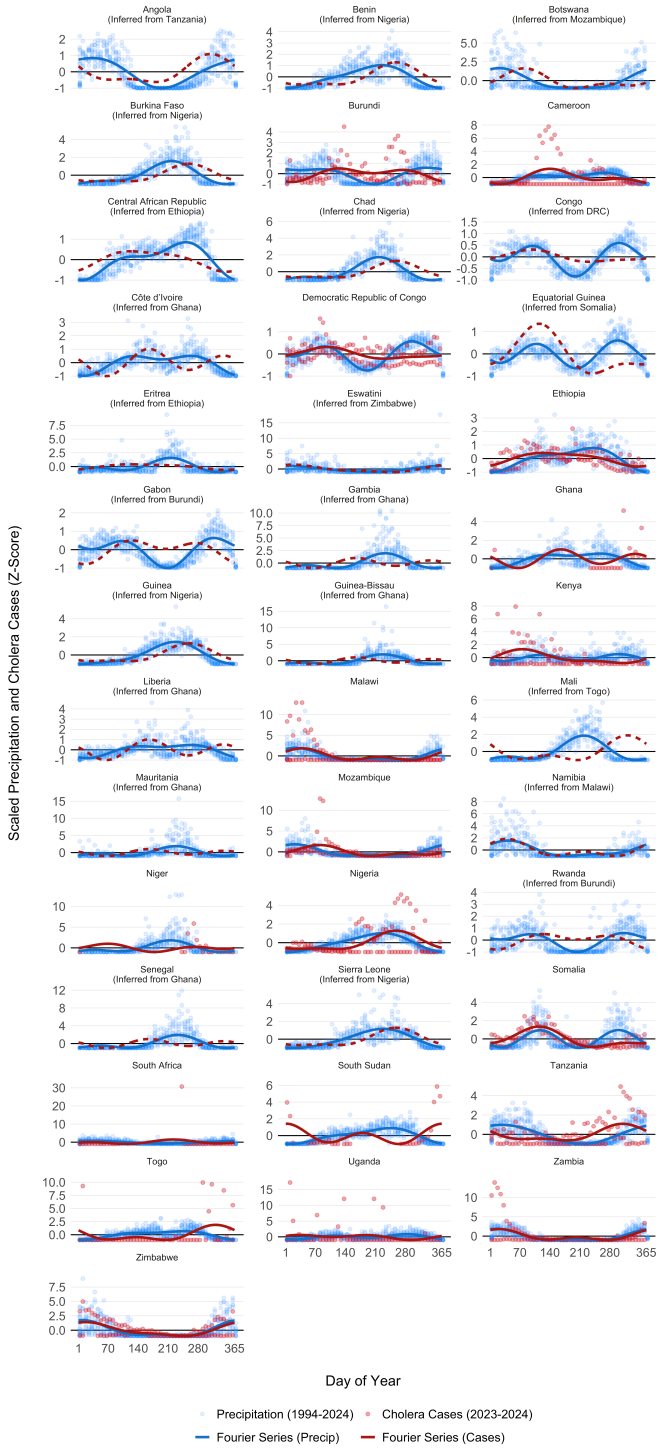

Figure 3.5: Seasonal transmission patterns for all countries modeled in MOSAIC as modeled by the truncated Fourier series in Equation eqrefeq:beta1. Blues lines give the Fourier series model fits for precipitation (1994-2024) and the red lines give models fits to reported cholera cases (2023-2024). For countries where reported case data were not available, the Fourier model was inferred by the nearest country with the most similar seasonal precipitation patterns as determined by the hierarchical clustering. Countries with inferred case data from neighboring locations are annotated in red. The X-axis represents the weeks of the year (1-52), while the Y-axis shows the Z-score of

ment and, *ii*) the rate of environment-to-human transmission which contributes to the overall force of infection.

$$\beta_{jt}^{\text{env}} = \beta_{j0}^{\text{env}} \left(1 + \frac{\psi_{jt} - \bar{\psi}_j}{\bar{\psi}_j} \right) \quad \text{and} \quad \bar{\psi}_j = \frac{1}{T} \sum_{t=1}^{T} \psi_{jt} \tag{3.8}$$

This formulation effectively scales the base environmental transmission rate β_{jt}^{env} so that it varies over time according to the climatically driven model of suitability. Note that, unlike the the cosine wave function of β_{jt}^{hum}, this temporal term can increase or decrease over time following multi-annual cycles.

3.4.1 Suitability-dependent decay rate

Suitability also influences how long *V. cholerae* survives in the environment. The decay rate δ_{jt} is modeled as the inverse of survival time, which varies with suitability. This is defined as:

$$\delta_{jt} = \frac{1}{\text{days}_{\text{short}} + f(\psi_{jt}) \cdot (\text{days}_{\text{long}} - \text{days}_{\text{short}})}.$$

Where $\text{days}_{\text{short}}$ is the shortest survival time (e.g., 3 days) and $\text{days}_{\text{long}}$ is the longest survival time (e.g., 90 days). Suitability is mapped to *V. cholerae* decay rate through a transformation function $f(\psi_{jt})$ that scales suitability values using a cumulative Beta distribution and two shape parameters s_1 and s_2: $f(\psi_{jt}) = \text{pbeta}(\psi_{jt} \mid s_1, s_2)$.

The transformation $f(\psi_{jt}) \in [0, 1]$ enables a range of functional forms, including linear, convex, concave, sigmoidal, or arcsine responses to suitability. This flexibility ensures that survival dynamics can reflect a variety of empirically plausible relationships with environmental conditions which can be seen in Figure 3.6.

3.4.2 Modeling environmental suitability

3.4.2.1 Environmental data

The mechanism for environment-to-human transmission (Equation (3.8)) and rate of decay of *V. cholerae* in the environment (Equation (3.4.1)) is driven by the parameter ψ_{jt}, which we refer to as environmental suitability. The parameter ψ_{jt} is modeled as a time series for each location using a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model and a suite of 24 covariates which include 19 historical and forecasted climate variables under the MRI-AGCM3-2-S climate model. Covariates also include 4 large-scale climate drivers such as the Indian Ocean Dipole Mode Index (DMI), and the El Niño Southern Oscillation (ENSO) from 3 different Pacific Ocean regions. We also included a location specific variable giving the mean elevation for each country.

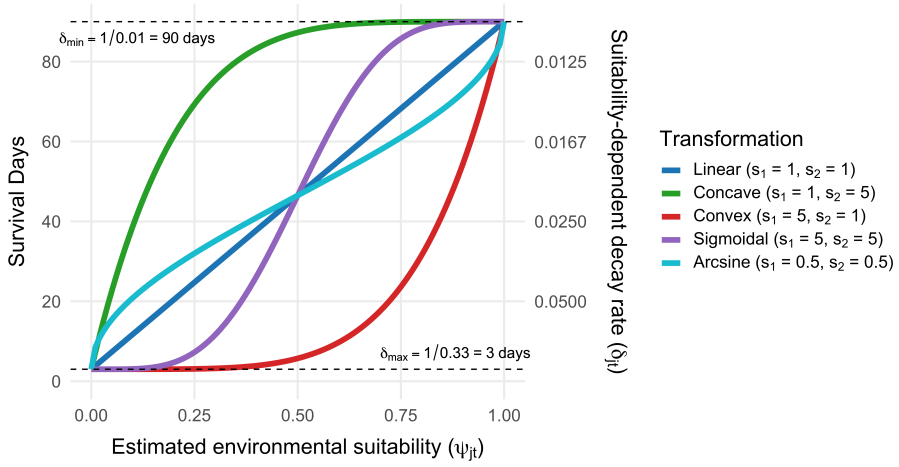

Figure 3.6: Relationship between environmental suitability (ψ_{jt}) and the survival and decay rate of *V. cholerae* in the environment (δ_{jt}). Curves represent four transformation types used to map suitability to survival time via the cumulative Beta distribution with different shape parameters (s_1, s_2). The primary y-axis shows survival time in days; the secondary y-axis shows the corresponding decay rate, defined as $\delta_{jt} = 1/\text{days}(\psi_{jt})$. Horizontal dashed lines indicate the bounds on survival time, from 3 days (low suitability) to 90 days (high suitability) in this example.

Table 3.2: A full list of covariates and their sources used in the LSTM RNN model to predict the environmental suitability of *V. cholerae* (ψ_{jt}).

Covariate	Description	Source
temperature_2m_mean	Average temperature at 2 meters	OpenMeteo [
temperature_2m_max	Maximum temperature at 2 meters	OpenMeteo [
temperature_2m_min	Minimum temperature at 2 meters	OpenMeteo [
wind_speed_10m_mean	Average wind speed at 10 meters	OpenMeteo [
wind_speed_10m_max	Maximum wind speed at 10 meters	OpenMeteo [
cloud_cover_mean	Mean cloud cover	OpenMeteo [
shortwave_radiation_sum	Total shortwave radiation	OpenMeteo [
relative_humidity_2m_mean	Mean relative humidity at 2 meters	OpenMeteo [
relative_humidity_2m_max	Maximum relative humidity at 2 meters	OpenMeteo [
relative_humidity_2m_min	Minimum relative humidity at 2 meters	OpenMeteo [
dew_point_2m_mean	Mean dew point at 2 meters	OpenMeteo [
dew_point_2m_min	Minimum dew point at 2 meters	OpenMeteo [
dew_point_2m_max	Maximum dew point at 2 meters	OpenMeteo [
precipitation_sum	Total precipitation	OpenMeteo [
pressure_msl_mean	Mean sea level pressure	OpenMeteo [
soil_moisture_0_to_10cm_mean	Mean soil moisture at 0 to 10 cm	OpenMeteo [
et0_fao_evapotranspiration_sum	Total evapotranspiration (FAO method)	OpenMeteo [
DMI	Dipole Mode Index (DMI)	[NOAA](http
ENSO3	El Niño Southern Oscillation (ENSO) - Region 3	[NOAA](http
ENSO34	ENSO - Region 3.4	[NOAA](http
ENSO4	ENSO - Region 4	[NOAA](http
elevation	Mean elevation	[Amazon Web

See example time series of climate variables from one country (Mozambique) in Figure 3.7 and DMI and ENSO variables in Figure 3.8. A list of all covariates and their sources can be seen in Table 3.2.

Note that while the 19 climate variables offer forecasts up to 2030 and beyond, the forecasts of the DMI and ENSO variables are limited to 5 months into the future. So, environmental suitability model predictions are currently limited to a 5 month time horizon but future iterations may allow for longer forecasts. Additional data sources will be integrated into subsequent versions of the suitability model. For instance, flood and cyclone data will likely be incorporated later, though not in the initial version of the model.

3.4.2.2 Deep learning neural network model

As mentioned above, we model environmental suitability ψ_{jt} using a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN) model. The LSTM model was developed using `keras` and `tensorflow` in R to predict binary outcomes. Thus the modeled quantity ψ_{jt} is a proportion implying unsuitable

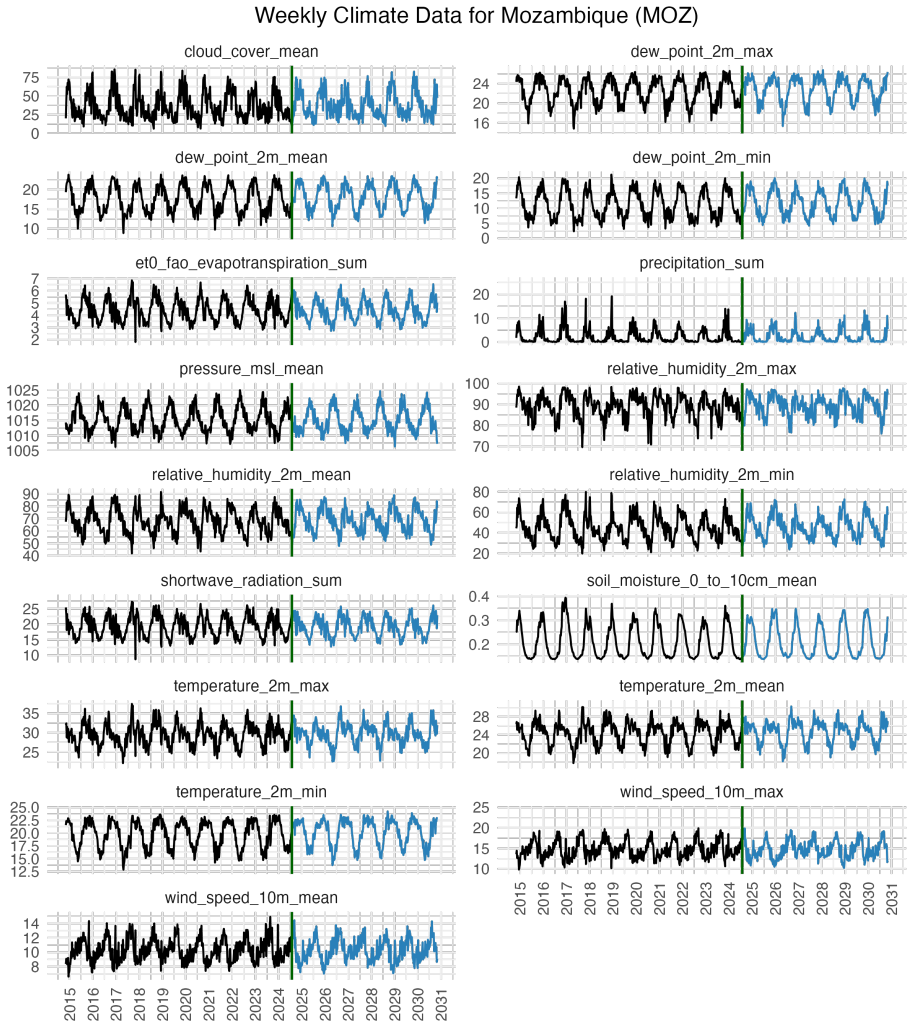

Figure 3.7: Climate data acquired from the OpenMeteo data API. Data were collected from 30 uniformly distributed points across each country and then aggregated to give weekly values of 17 climate variable from 1970 to 2030.

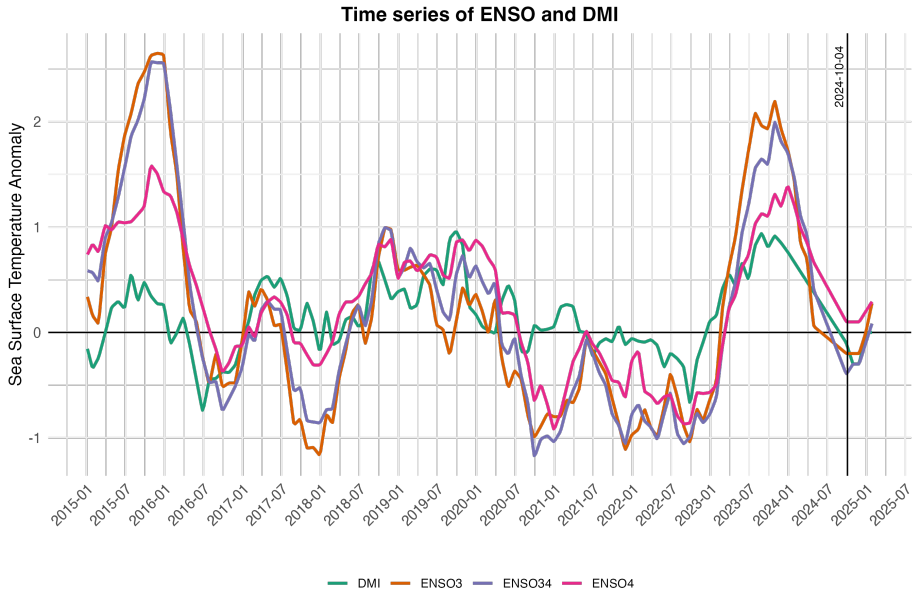

Figure 3.8: Historical and forecasted values of the Indian Ocean Dipole Mode Index (DMI) and the El Niño Southern Oscillation (ENSO) from 2015 to 2025. The ENSO values come from three different regions: Niño3 (central to eastern Pacific), Niño3.4 (central Pacific), and Niño4 (western-central Pacifi). Data are from National Oceanic and Atmospheric Administration (NOAA) and Bureau of Meteorology (BOM).

conditions at 0 and perfectly suitable conditions at 1.

The model was fitted to reported case counts that were converted to a binary variable using a threshold of 200 reported cases per week. Given delays in reporting and likely lead times for environmental suitability ahead of transmission and case reporting, we also set the preceding one week to be suitable and in cases where there were two consecutive weeks of >200 cases per week, we assumed that the preceding two weeks were also suitable. See Figure 3.9 for an example of how reported case counts are converted to a binary variable representing presumed environmental suitability for *V. cholerae*.

The model is a Long Short-Term Memory (LSTM) neural network designed for binary classification, where environmental suitability, ψ_{jt}, is modeled as a function of the hidden state h_t and hidden bias term b_h. Specifically, ψ_{jt} is defined by a sigmoid activation function applied to the linear combination of the hidden state h_t and the bias b_h which if given by the 3 layers of the LSTM model:

$$\psi_{jt} \sim \text{Sigmoid}(w_h \cdot h_t + b_h) \tag{3.9}$$

$$h_t = \text{LSTM}(\text{temperature}_{jt},\ \text{precipitation}_{jt},\ \text{ENSO}_t, \dots) \tag{3.10}$$

In this formulation, h_t represents the hidden state generated by the LSTM network based on input variables such as temperature, precipitation, and ENSO conditions, while b_h is a bias term added to the output of the hidden state transformation.

The deep learning LSTM model consists of three stacked LSTM-RNN layers. The first LSTM layer has 500 units and the second and third LSTM layers have 250 and 100 units respectively. The architecture the LSTM model is configured to pass node values to subsequent LSTM layers allowing deep learning of more the complex interactions among the climate variable over time. We enforced model sparsity for each LSTM layer using L2 regularization (penalty = 0.001) and used a dropout rate of 0.5 for each LSTM layer to further prevent overfitting on the limited amount of data. The final output layer was a dense layer with a single unit and a sigmoid activation function to produce a probability value for binary classification, i.e. a prediction of environmental suitability ψ_{jt} on a scale of 0 to 1.

To fit the LSTM model to data, we modified the learning rate by applying an exponential decay schedule that started at 0.001 and decayed by a factor of 0.9 every 10,000 steps to enable smoother convergence. The model was compiled using the Adam optimizer with this learning rate schedule, along with binary cross-entropy as the loss function and accuracy as the evaluation metric. The model was trained for a maximum of 200 epochs with a batch size of 1024. We allowed model fitting to stop early with a patience parameter of 10 which halts

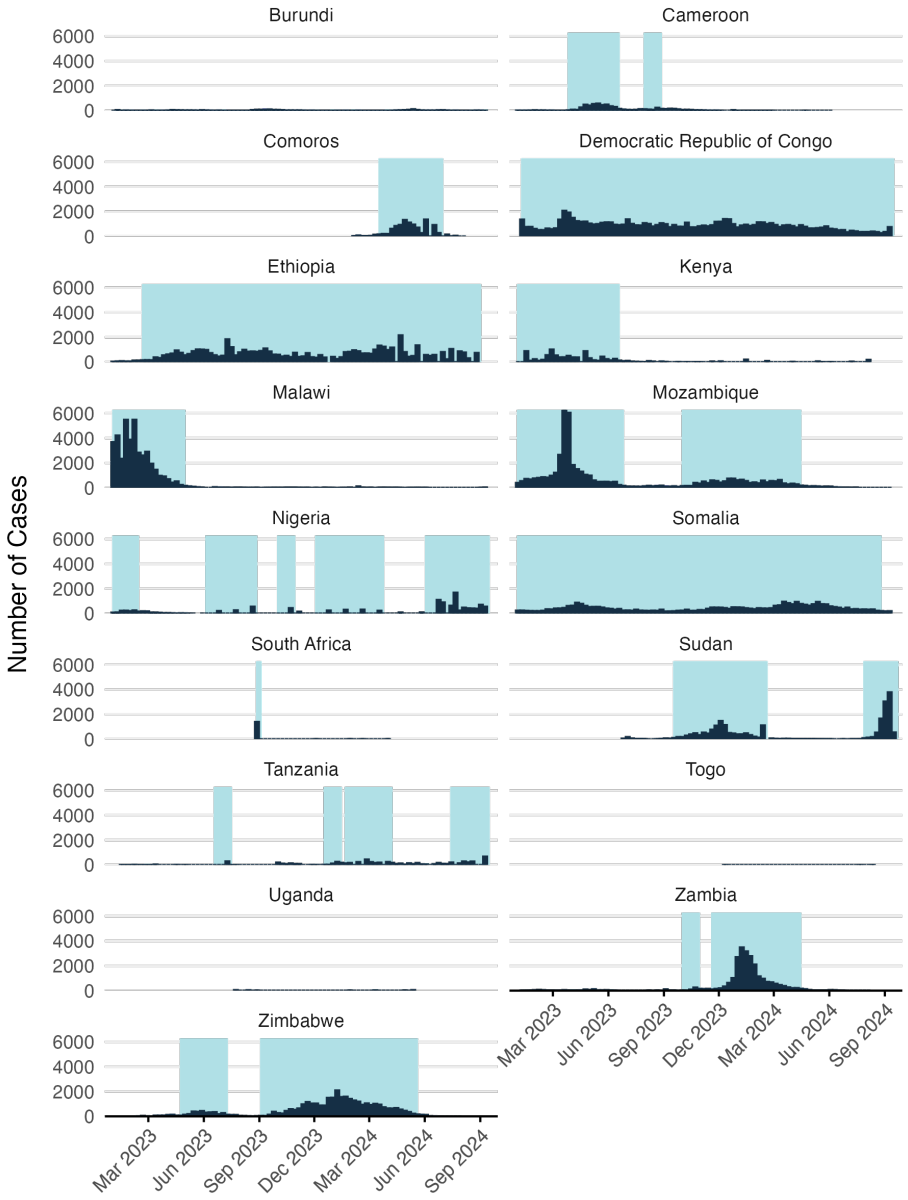

Figure 3.9: Reported cases converted to binary variable for modeling environmental suitability.

training if no improvement is observed in validation accuracy for 10 consecutive epochs. To train the model we set aside 20% of the observed data for validation and also used 20% of the training data for model fitting. The training history, including loss and accuracy, was monitored over the course of training and gave a final test accuracy of 0.73 and a final test loss of 0.56 (see Figure 3.10).

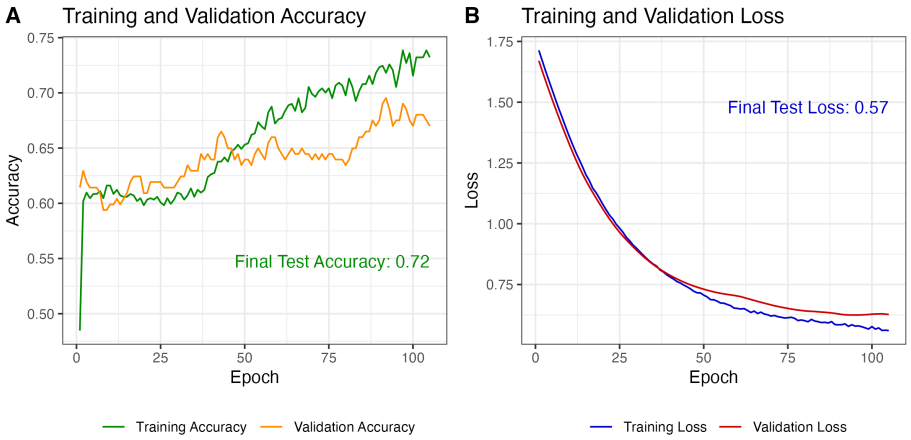

Figure 3.10: Model performance on training and validation data.

After model training was completed, we predicted the values of environmental suitability ψ_{jt} across all time steps for each location. Predictions start in January 1970 and go up to 5 months past the present date (currently February 2025). Given the amount of noise in the model predictions, we added a simple LOESS spline with logit transformation to smooth model predictions over time and give a more stable value of ψ_{jt} when incorporating it into other model features (e.g. Equations (3.8) and (3.4.1)). The resulting model predictions are shown for an example country such as Mozambique in Figure 3.11 which compares model predictions to the original case counts and the binary classification. Predicitons for all model locations are shown in a simplified view in Figure 3.12.

Also, please note that this initial version of the model is fitted to a rather small amount of data. Model hyper parameters were specifically chosen to reduce overfitting. Therefore, we recommend to not over-interpret the time series predictions of the model at this early stage since they are likely to change and improve as more historical incidence data is included in future versions.

3.4.3 Shedding of *V. cholerae*

The rate at which infected individuals shed *Vibrio cholerae* into the environment is a critical factor influencing cholera transmission dynamics. Shedding rates vary widely depending on the severity of infection, the host immune response, and environmental conditions. To reflect this heterogeneity, the model

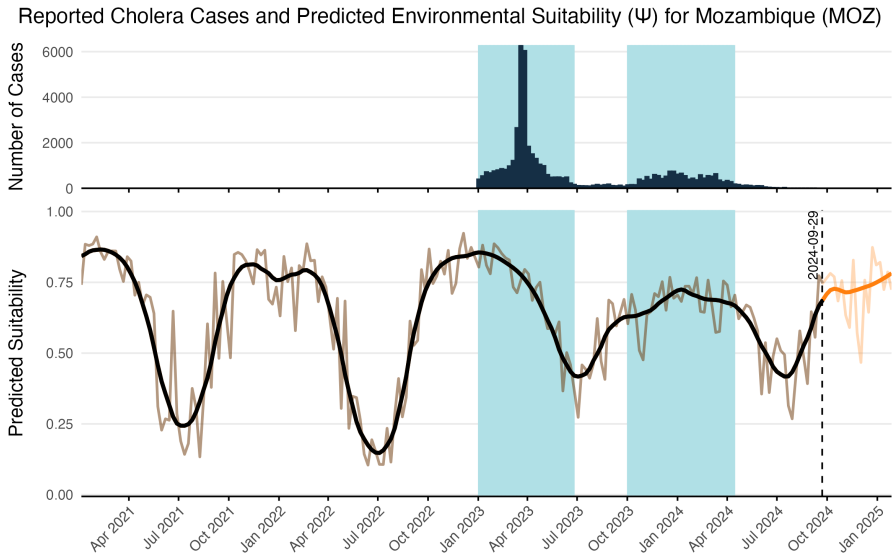

Figure 3.11: The LSTM model predictions over time and reported cases for an example country such as Mozambique. Reported cases are shown in the top panel and tje shaded areas show the binary classification used to characterize environmental suitability. Raw model predicitons are shown in the transparent brown line with the solid black line showing the LOESS smoothing. Forecasted values beyond the current time point are shown in orange and are limited to 5 month time horizon.

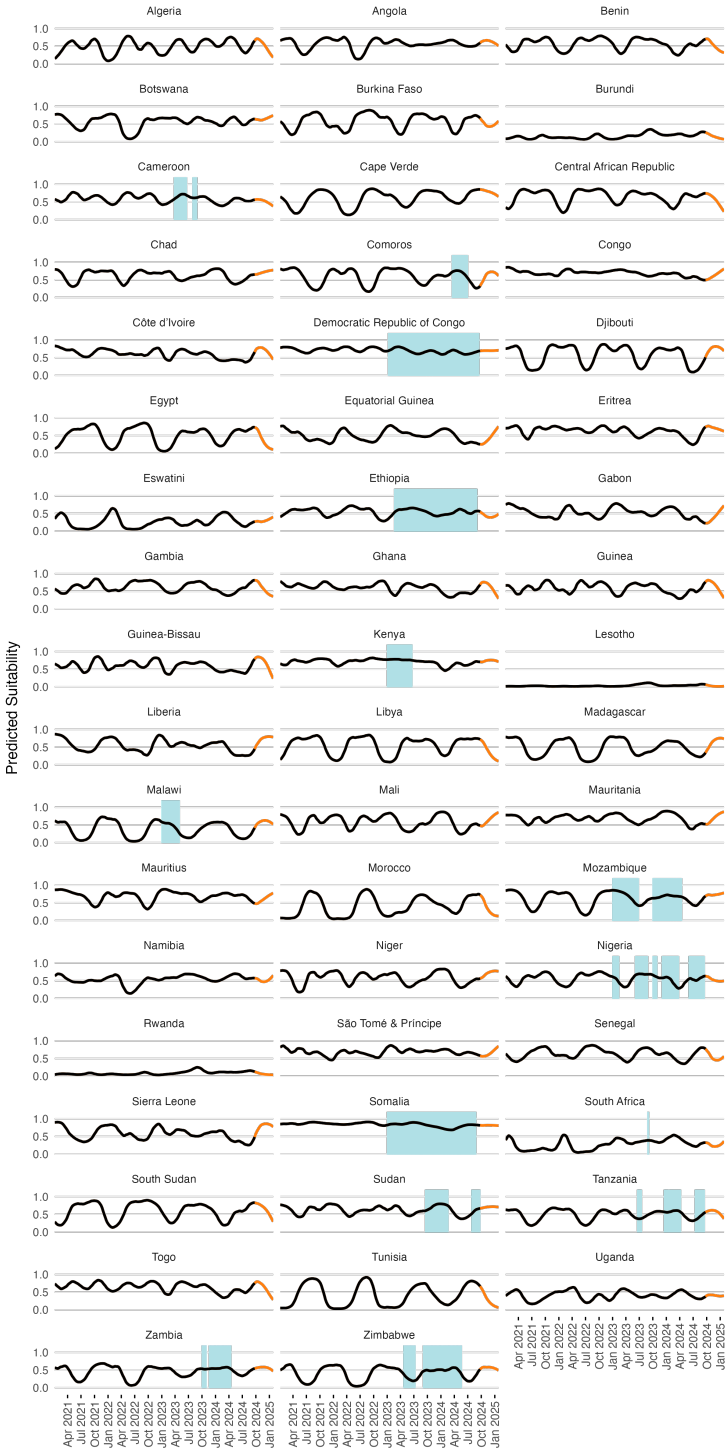

Figure 3.12: The smoothed LSTM model predictions (lines) and binary suitability classification (shaded areas) over time for all countries in the MOSAIC framework. Orange lines show forecasts beyond the current date. With ENSO and DMI covariates included in the model, forecasts are limited to 5 months.

distinguishes between two types of infected individuals:

- Symptomatic individuals (I_1), who tend to shed substantially more bacteria for longer due to more severe gastrointestinal symptoms;
- Asymptomatic individuals (I_2), who shed less per capita and for a shorter period of time, but may contribute significantly to environmental contamination due to their larger numbers.

According to the modeling study done by Fung et al. (2014), estimates of *V. cholerae* shedding across the population can range from 0.01 to 10 cells per mL per person per day. However, this estimate does not fully capture the range of possible shedding that can occur depending on the type of infection. In contrast, Nelson et al. (2009) report that individuals may shed between 10^3 cells g^{-1} stool in asymptomatic cases and up to 10^{12} cells g^{-1} stool in severe symptomatic infections. While these quantities are slightly different from the cells mL^{-1} $person^{-1}$ day^{-1} units used in cholera transmission models, it implies that symptomatic individuals may shed several orders of magnitude more bacteria into the environment per day than asymptomatic individuals.

To account for the uncertainty in levels of *V. cholerae* shedding, we collated a short list of studies that either report empirical findings or modeling analyses that set priors for shedding parameters. These sources reflect a wide range of assumptions and contexts, but nonetheless provide a spectrum of estimated *V. cholerae* shedding rates that we can use to inform our model (see the table of shedding parameters below below).

We currently assume that shedding rates are constant and drawn from independent uniform distributions in units of cells mL^{-1} $person^{-1}$ day^{-1}, which is consistent with the frequently cited sources for shedding rates of Codeço 2001 and others:

$$\begin{aligned} \zeta_1 &\sim \text{Uniform}(10^4,\ 10^8) \quad \text{(symptomatic shedding)}, \\ \zeta_2 &\sim \text{Uniform}(0.01,\ 10^3) \quad \text{(asymptomatic shedding)}. \end{aligned}$$

The definition of these priors assumes that:

1) the watery stool of infected individuals has approximately the same density as water (1kg/L), such that 10^5cells $g^{-1}day^{-1}$ $\approx$ 10^5cells $mL^{-1}person^{-1}day^{-1}$, and
2) shedding in symptomatic individuals is always greater than that of asymptomatic individuals with the potential to be many orders of magnitude greater.

These priors also reflect the observed variability in the literature while preserving identifiability in model fitting. The upper bound for symptomatic shedding (10^8) is conservative relative to extreme values (e.g., 10^{12} cells/L in rice water stool), but comfortably spans values seen in both clinical observations and theoretical models. The lower bound (10^4) ensures that small but still epidemiologically

significant shedding is captured. For asymptomatic individuals, the range of 0.01 to 10^3 cells mL^{-1} $person^{-1}$ day^{-1} captures both low empirical estimates (e.g. Mosley et al. 1968) and broader assumptions made in cholera transmission models (e.g. Codeço 2001). The range of these priors therefore provides sufficient flexibility to represent both high-intensity shedding in severe cases and low-level contributions distributed across a larger number of asymptomatic individuals.

The table below summarizes key published estimates and assumptions regarding *V. cholerae* and related bacterial shedding rates:

Value(s)	Units	Infection	Description	Source
10^3	cells g^{-1} stool	Asymptomatic	Approx. 1 day of shedding at ~10^3 vibrios per gram of stool	Mosley et al. (1968)
10^6–10^9	cells g^{-1} stool	NA	Number of fecal coliform indicator bacteria in human feces	Feachem et al. (1983)
1–100	cells mL^{-1} $person^{-1}$ day^{-1}	All	Point estimate of 10; range 1–100 used in sensitivity analysis	Codeço (2001)
10	cells mL^{-1} $person^{-1}$ day^{-1}	All	Assumed shedding rate used in epidemic model incorporating hyperinfectivity	Hartley et al. (2006)
$\leq 10^5$	cells g^{-1} stool	Asymptomatic	No symptoms; low-level shedding of vibrios	Nelson et al. (2009)
$\leq 10^8$	cells g^{-1} stool	Mild	Diarrhoea with moderate vibrios in stool	Nelson et al. (2009)
10^7–10^9	cells g^{-1} stool	Severe	Vomiting and profuse diarrhoea with high shedding	Nelson et al. (2009)
10^{10}–10^{12}	cells L^{-1} stool	Severe	Concentration in rice water stool from symptomatic individuals	Nelson et al. (2009)
0.01–10	cells mL^{-1} $person^{-1}$ day^{-1}	All	Reported as general estimate across all infections	Fung (2014)
10–100	cells mL^{-1} $person^{-1}$ day^{-1}	All	Represents shedding rates in two distinct sub-populations	Njagarah & Nyabadza (2014)

3.4.4 Recovery rates

The recovery rates in the MOSAIC model are defined as the inverse of the shedding duration for infected individuals. This reflects the period during which individuals contribute to the environmental load of *Vibrio cholerae*, regardless of the presence of clinical symptoms. The model distinguishes between:

Symptomatic individuals (γ_1):
Individuals in the I_1 compartment typically experience acute watery diarrhea and may shed large quantities of *V. cholerae* for several days. Clinical studies and reviews suggest that symptomatic patients shed vibrios for approximately **3 to 5 days**, with shedding sometimes persisting up to 14 days (Nelson et al. 2009, Harris et al. 2012). Based on these estimates, we define the recovery rate as a uniform distribution over plausible durations:

$$\gamma_1 \sim \text{Uniform}(1/7,\ 1/3) \quad \text{day}^{-1}$$

Asymptomatic individuals (γ_2):
Asymptomatic individuals in the I_2 compartment may not show clinical symptoms but can still shed *V. cholerae* for extended periods. Observational studies indicate shedding can persist for **7 to 14 days**, and potentially longer (Mosley et al. 1968, Public Health Ontario 2022). To capture this range, we define the asymptomatic recovery rate as:

$$\gamma_2 \sim \text{Uniform}(1/14,\ 1/7) \quad \text{day}^{-1}$$

In cases where point estimates are preferred or required for model fitting, we use the mean values of each distribution:

$$\gamma_1 = \frac{1}{5} = 0.2 \text{ day}^{-1}, \qquad \gamma_2 = \frac{1}{10} = 0.1 \text{ day}^{-1}$$

These parameterizations reflect the empirical difference in shedding durations by symptom status and are consistent with previous cholera transmission models (e.g., Codeço 2001). They also align with the structure of the environmental shedding process and the I_1, I_2 compartments in the model.

3.4.5 WAter, Sanitation, and Hygiene (WASH)

Since *V. cholerae* is transmitted through fecal contamination of water and other consumables, the level of exposure to contaminated substrates significantly impacts transmission rates. Interventions involving Water, Sanitation, and Hygiene (WASH) have long been a first line of defense in reducing cholera transmission, and in this context, WASH variables can serve as proxy for the rate of contact with environmental risk factors. In the MOSAIC model, WASH variables are incorporated mechanistically, allowing for intervention scenarios that

Figure 3.13: Estimated shedding duration (x-axis) for symptomatic and asymptomatic *V. cholerae* infections. Shaded bars indicate the assumed range of plausible durations; solid vertical lines mark the mean value for each group. These durations are used to derive recovery rates (γ_1 and γ_2) as the inverse of duration and parameterize the infectious period in the MOSAIC transmission model.

include changes to WASH. However, it is necessary to distill available WASH variables into a single parameter that represents the WASH-determined contact rate with contaminated substrates for each location j, which we define as θ_j.

To parameterize θ_j, we calculated a weighted mean of the 8 WASH variables in Sikder et al 2023 and originally modeled by the Local Burden of Disease WaSH Collaborators 2020. The 8 WASH variables (listed in Table 3.4) provide population-weighted measures of the proportion of the population that either: *i*) have access to WASH resources (e.g., piped water, septic or sewer sanitation), or *ii*) are exposed to risk factors (e.g. surface water, open defecation). For risk associated WASH variables, we used the complement ($1 -$ value) to give the proportion of the population *not* exposed to each risk factor. We used the `optim` function in R and the L-BFGS-B algorithm to estimate the set of optimal weights (Table 3.4) that maximize the correlation between the weighted mean of the 8 WASH variables and reported cholera incidence per 1000 population across 40 SSA countries from 2000 to 2016. The optimal weighted mean had a correlation coefficient of r = -0.33 (-0.51 to -0.09 95% CI) which was higher than the basic mean and all correlations provided by the individual WASH variables (see Figure 3.14). The weighted mean then provides a single variable between 0 and 1 that represents the overall proportion of the population that has access to WASH and/or is not exposed to environmental risk factors. Thus, the WASH-mediated contact rate with sources of environmental transmission is represented as $(1 - \theta_j)$ in the environment-to-human force of infection (Ψ_{jt}). Values of θ_j for all countries are shown in Figure 3.15.

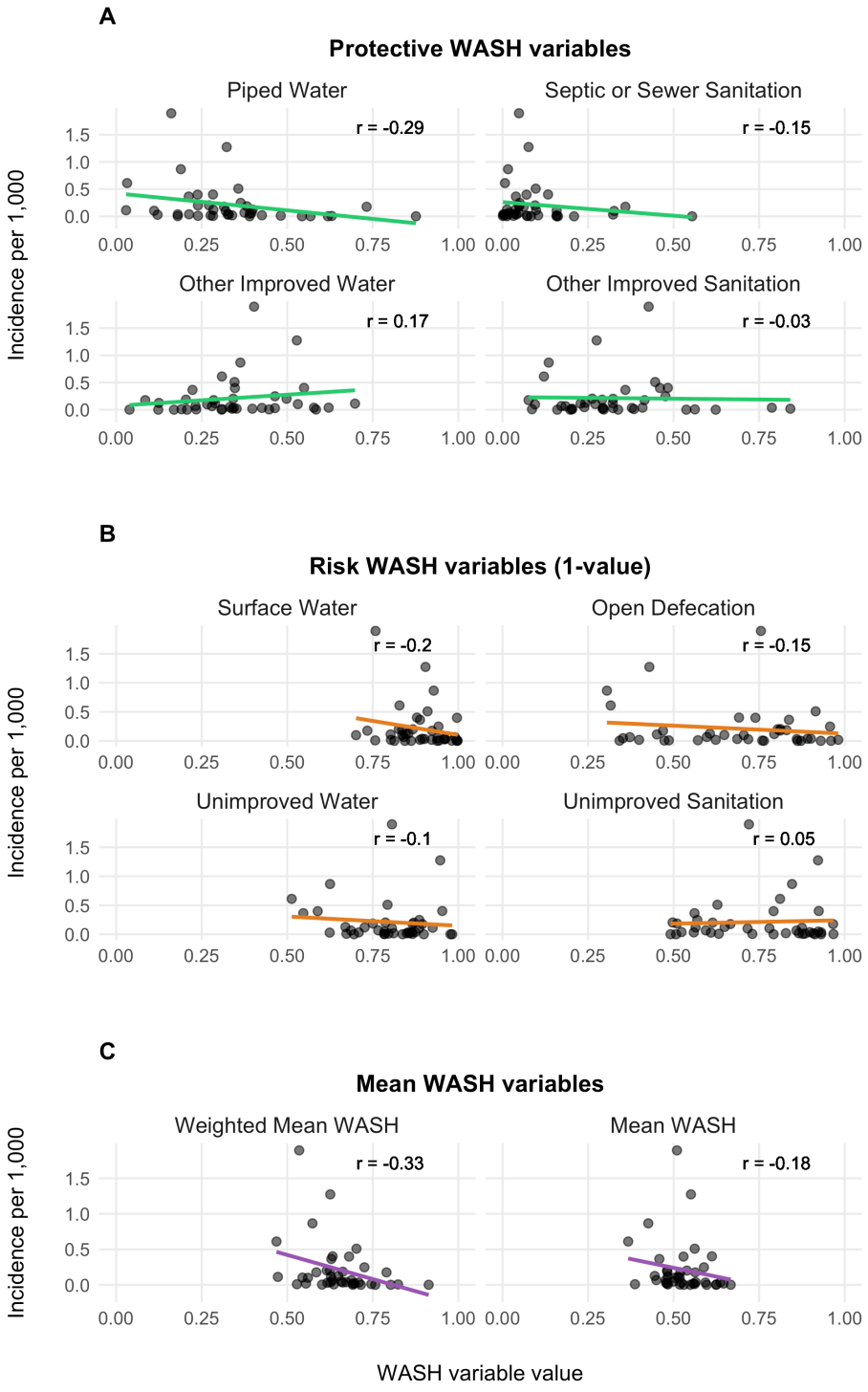

Figure 3.14: Relationship between WASH variables and cholera incidences.

Table 3.4: Table of optimized weights used to calculate the single mean WASH index for all countries.

WASH variable	Optimized weight
Piped Water	0.356
Septic or Sewer Sanitation	0.014
Other Improved Water	0.000
Other Improved Sanitation	0.000
Surface Water	0.504
Unimproved Sanitation	0.000
Unimproved Water	0.000
Open Defecation	0.126

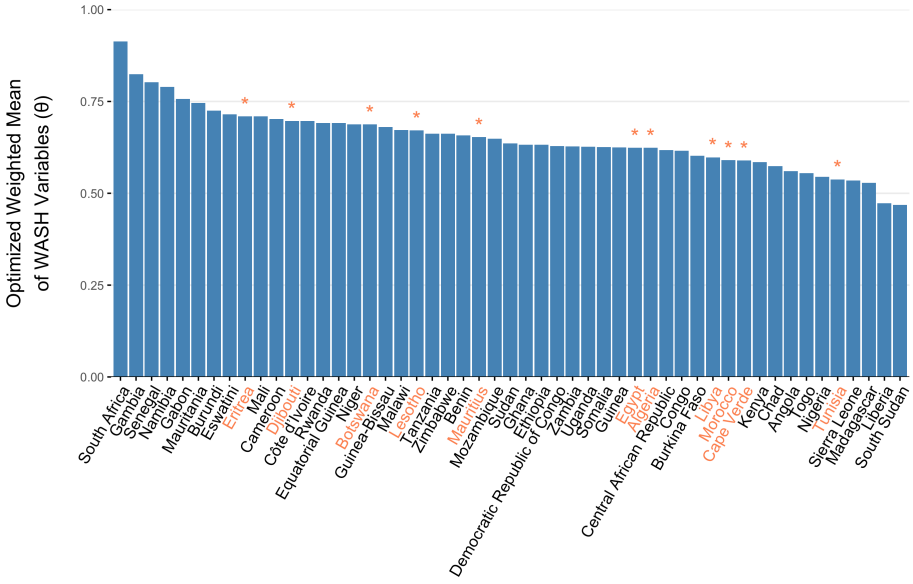

Figure 3.15: The optimized weighted mean of WASH variables for AFRO countries. Countries labeled in orange denote countries with an imputed weighted mean WASH variable. Imputed values are the weighted mean from the 3 most similar countries.

3.5 Immune dynamics

Aside from the current number of infections, population susceptibility is one of the key factors influencing the spread of cholera. Further, since immunity from both vaccination and natural infection provides long-lasting protection, it's crucial to quantify not only the incidence of cholera but also the number of past vaccinations. Additionally, we need to estimate how many individuals with immunity remain in the population at any given time step in the model.

To achieve this, we estimate the vaccination rate over time (ν_{jt}) based on historical vaccination campaigns and incorporate a model of vaccine effectiveness (ϕ) and immune decay post-vaccination (ω) to estimate the current number of individuals with vaccine-derived immunity. We also account for the immune decay rate from natural infection (ε), which is generally considered to last longer than immunity from vaccination.

3.5.1 Estimating Vaccination Rates

To estimate the past and current vaccination rates, we sourced data on reported OCV vaccinations from the WHO International Coordinating Group (ICG) Cholera vaccine dashboard. This resource lists all reactive OCV campaigns conducted from 2016 to the present, with approximately 103 million OCV doses shipped to Sub-Saharan African (SSA) countries as of October 9, 2024. However, these data only capture reactive vaccinations in emergency settings and do not include preventive campaigns organized by GAVI and in-country partners.

As a result, our current estimates of the OCV vaccination rate likely underestimate total OCV coverage. We are working to expand our data sources to better reflect the full number of OCV doses distributed in SSA and will update the results here as soon as these are available.

To translate the reported number of OCV doses into the model parameter ν_{jt}, we take the number of doses shipped and the reported start date of the vaccination campaign, distributing the doses over subsequent days according to a maximum daily vaccination rate. Therefore, the vaccination rate ν_t is not an estimated quantity, it is defined by the reported number of OCV doses administered with a assumption about the daily rate of distribution for an OCV campaign:

$$\nu_{jt} = f(\text{reported OCV doses distributed}_{jt} \mid \text{daily distribution rate}).$$

See Figure 3.16 for an example of OCV distribution using a maximum daily vaccination rate of 100,000. The resulting time series for each country is shown in Figure 3.17, with current totals based on the WHO ICG data displayed in Figure 3.18.

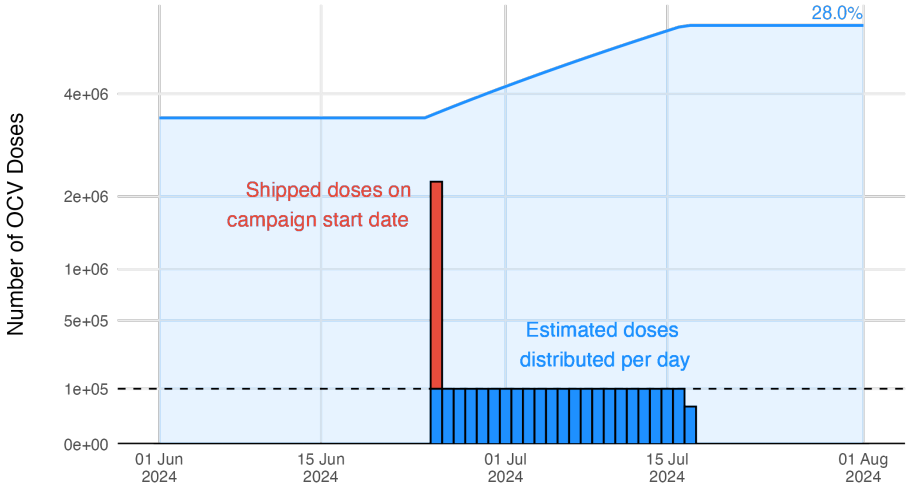

Figure 3.16: Example of the estimated vaccination rate during an OCV campaign.

Table 3.5: Summary of Effectiveness Data

Effectiveness	Upper CI	Lower CI	Day (midpoint)	Day (min)	Day (max)	Source
60.0	0.873	0.990	0.702	NA	NA	Azman et al (2016)
93.5	0.400	0.600	0.110	7	180	Qadri et al (2016)
368.5	0.390	0.520	0.230	7	730	Qadri et al (2018)
435.0	0.527	0.674	0.314	360	510	Malembaka et al (2024)
900.0	0.447	0.594	0.248	720	1080	Malembaka et al (2024)

3.5.2 Immunity from vaccination

The impacts of Oral Cholera Vaccine (OCV) campaigns is incorporated into the model through the Vaccinated compartment (V). The rate that individuals are effectively vaccinated is defined as $\phi\nu_t$, where ν_t is the number of OCV doses administered in location j at time t and ϕ is the estimated vaccine effectiveness. The vaccination rate ν_{jt} is not an estimated quantity. Rather, it is directly defined by the reported number of OCV doses administered as described above. Note that there is just one vaccinated compartment at this time, though future model versions may include V_1 an V_2 compartments to explore two dose vaccination strategies or to emulate more complex waning patterns.

The evidence for waning immunity comes from 4 cohort studies (Table 3.5) from Bangladesh (Qadri et al 2016 and 2018), South Sudan (Azman et al 2016), and Democratic Republic of Congo (Malembaka et al 2024).

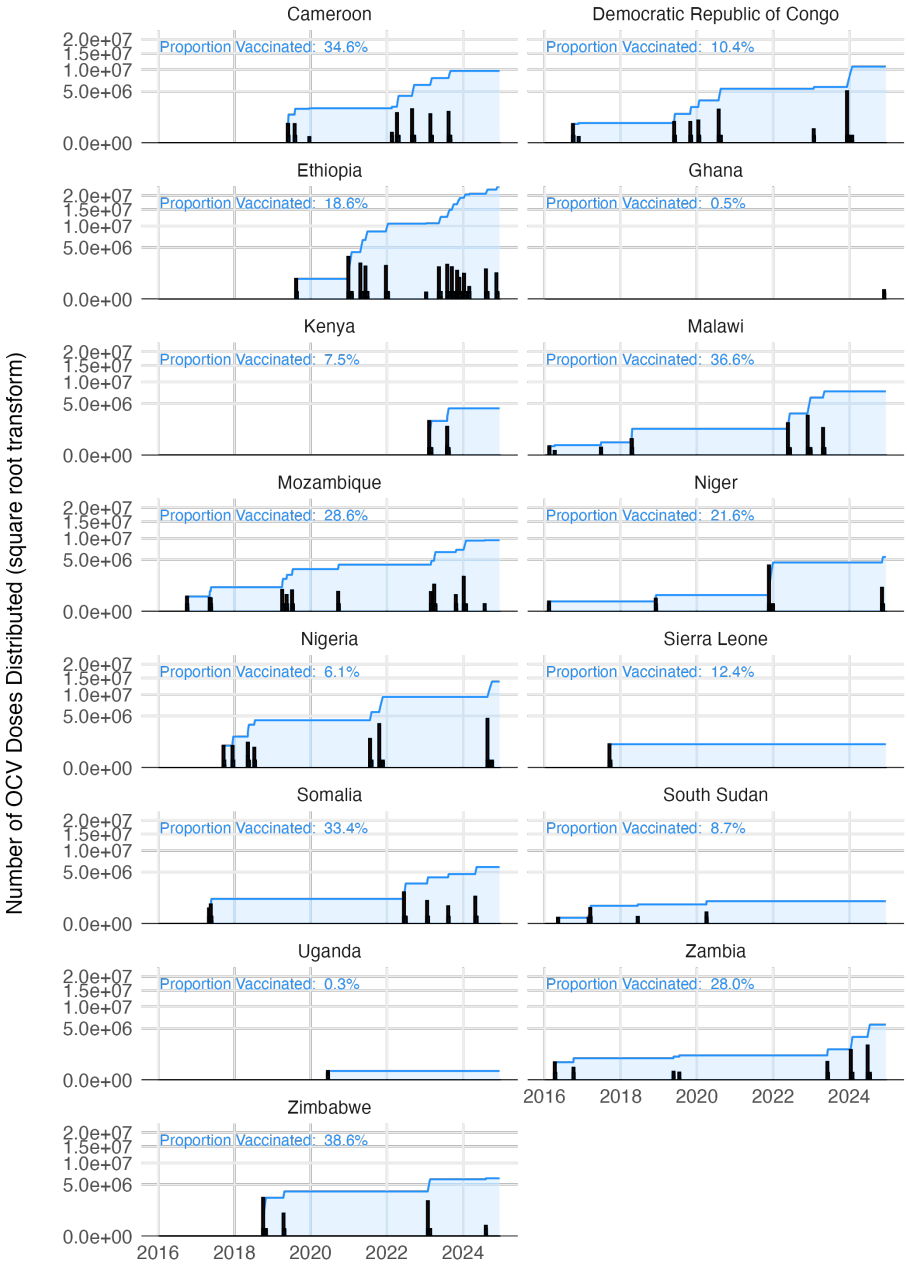

Figure 3.17: The estimated vaccination coverage across all countries with reported vaccination data one the WHO ICG dashboard.

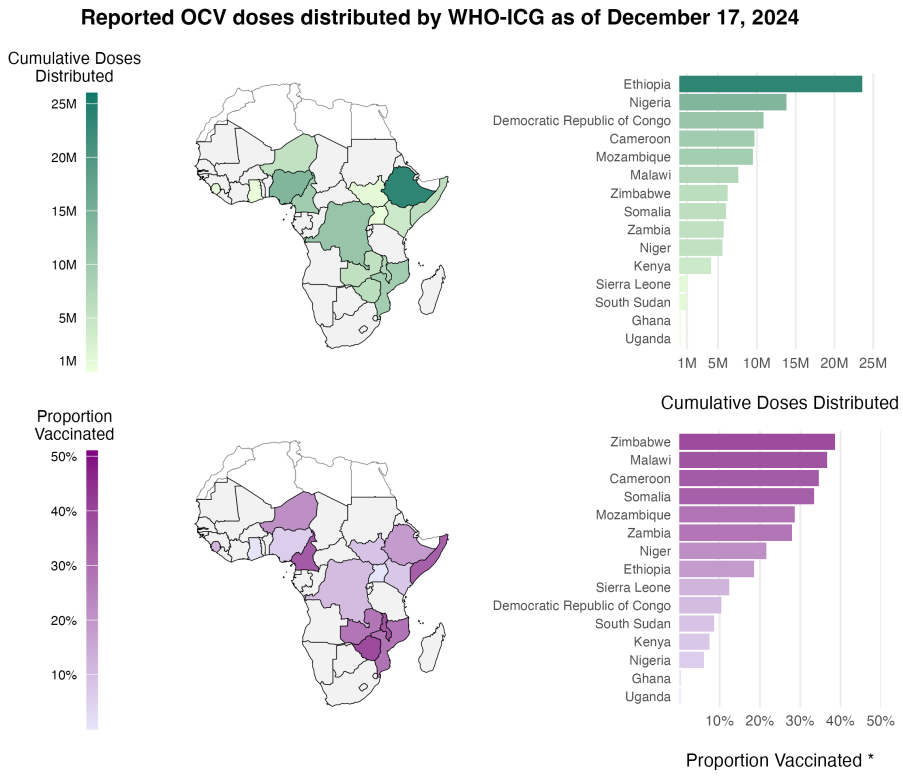

Figure 3.18: The total cumulative number of OCV doses distributed through the WHO ICG from 2016 to present day.

We estimated vaccine effectiveness and waning immunity by fitting an exponential decay model to the reported effectiveness of one dose OCV in these studies using the following formulation:

$$\text{Proportion immune } t \text{ days after vaccination} = \phi \times (1-\omega)^{t-t_{\text{vaccination}}} \quad (3.11)$$

Where ϕ is the effectiveness of one dose OCV, and the based on this specification, it is also the initial proportion immune directly after vaccination. The decay rate parameter ω is the rate at which initial vaccine derived immunity decays per day post vaccination, and t and $t_{\text{vaccination}}$ are the time (in days) the function is evaluated at and the time of vaccination respectively. When we fitted the model to the data from the cohort studies shown in Table (3.5) we found that $\omega = 0.00057$ ($0 - 0.0019$ 95% CI), which gives a mean estimate of 4.8 years for vaccine derived immune duration with unreasonably large confidence intervals (1.4 years to infinite immunity). However, the point estimate of 4.8 years is consistent with anecdotes that one dose OCV is effective for up to at least 3 years.

The wide confidence intervals are likely due to the wide range of reported estimates for proportion immune after a short duration in the 7–90 days range (Azman et al 2016 and Qadri et al 2016). Therefore, we chose to use the point estimate of ω and incorporate uncertainty based on the initial proportion immune (i.e. vaccine effectiveness ϕ) shortly after vaccination. Using the decay model in Equation (3.11) we estimated ϕ to be 0.64 ($0.32 - 0.96$ 95% CI). We then fit a Beta distribution to the quantiles of ϕ by minimizing the sums of squares using the Nelder-Mead optimization algorithm to render the following distribution (shown in Figure 3.19B):

$$\phi \sim \text{Beta}(4.57, 2.41). \quad (3.12)$$

3.5.3 Immunity from natural infection

The duration of immunity after a natural infection is likely to be longer lasting than that from vaccination with OCV (especially given the current one dose strategy). As in most SIR-type models, the rate at which individuals leave the Recovered compartment is governed by the immune decay parameter ε. We estimated the durability of immunity from natural infection based on two cohort studies and fit the following exponential decay model to estimate the rate of immunity decay over time:

$$\text{Proportion immune } t \text{ days after infection} = 0.99 \times (1-\varepsilon)^{t-t_{\text{infection}}}$$

Where we make the necessary and simplifying assumption that within 0–90 days after natural infection with *V. cholerae*, individuals are 95–99% immune. We

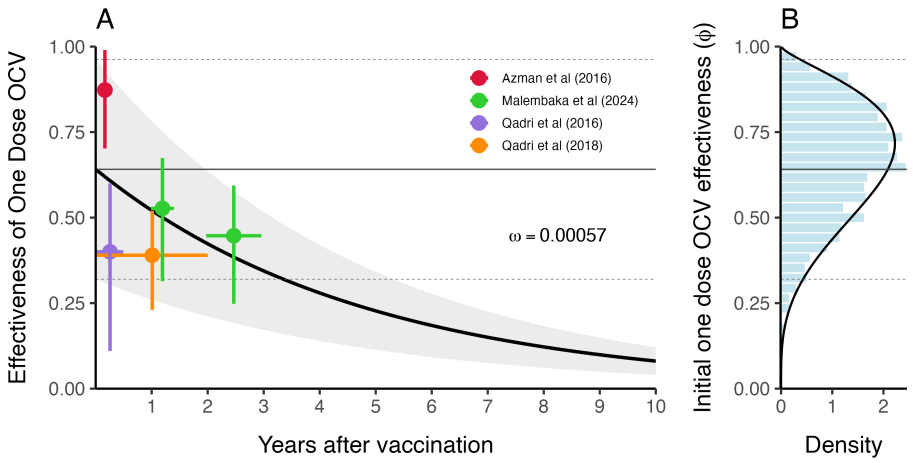

Figure 3.19: This is vaccine effectiveness

Table 3.6: Sources for the duration of immunity fro natural infection.

Day	Effectiveness	Upper CI	Lower CI	Source
90	0.95	0.95	0.95	Assumption
1080	0.65	0.81	0.37	[Ali et al (2011)](https://doi.org/10.1093/infdis/jir416)
1260	0.61	0.81	0.21	[Clemens et al (1991)](https://www.sciencedirect.com/science/artic

fit this model to reported data from Ali et al (2011) and Clemens et al (1991) (see Table 3.6).

We estimated the mean immune decay to be $\bar{\varepsilon} = 3.9 \times 10^{-4}$ ($1.7 \times 10^{-4} - 1.03 \times 10^{-3}$ 95% CI) which is equivalent to an immune duration of 7.21 years ($2.66 - 16.1$ years 95% CI) as shown in Figure 3.20A. This is slightly longer than previous modeling work estimating the duration of immunity to be ~5 years (King et al 2008). Uncertainty around ε in the model is then represented by a Log-Normal distribution as shown in Figure 3.20B:

$$\varepsilon \sim \text{Lognormal}(\bar{\varepsilon} + \frac{\sigma^2}{2}, 0.25)$$

3.6 Spatial dynamics

The parameters in the model diagram in Figure 3.2 that have a jt subscript denote the spatial structure of the model. Each country is modeled as an independent metapopulation that is connected to all others via the spatial force of infection Λ_{jt} which moves contagion among metapopulations according to the

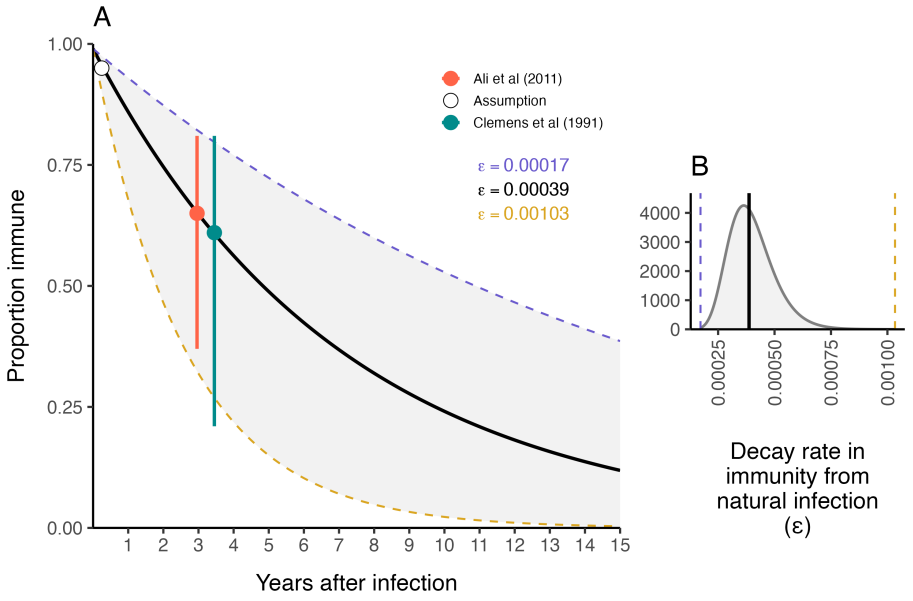

Figure 3.20: The duration of immunity after natural infection with *V. cholerae*.

connectivity provided by parameters τ_i (the probability departure) and π_{ij} (the probability of diffusion to destination j). Both parameters are estimated using the departure-diffusion model below which is fitted to average weekly air traffic volume between all of the 41 countries included in the MOSAIC framework (Figure 3.21).

3.6.1 Human mobility model

The departure-diffusion model estimates diagonal and off-diagonal elements in the mobility matrix (M) separately and combines them using conditional probability rules. The model first estimates the probability of travel outside the origin location i—the departure process—and then the distribution of travel from the origin location i by normalizing connectivity values across all j destinations—the diffusion process. The values of π_{ij} sum to unity along each row, but the diagonal is not included, indicating that this is a relative quantity. That is to say, π_{ij} gives the probability of going from i to j given that travel outside origin i occurs. Therefore, we can use basic conditional probability rules to define the travel routes in the diagonal elements (trips made within the origin i) as

$$\Pr(\neg\text{depart}_i) = 1 - \tau_i$$

and the off-diagonal elements (trips made outside origin i) as

$$\Pr(\text{depart}_i, \text{diffuse}_{i\to j}) = \Pr(\text{diffuse}_{i\to j} \mid \text{depart}_i)\Pr(\text{depart}_i) = \pi_{ij}\tau_i.$$

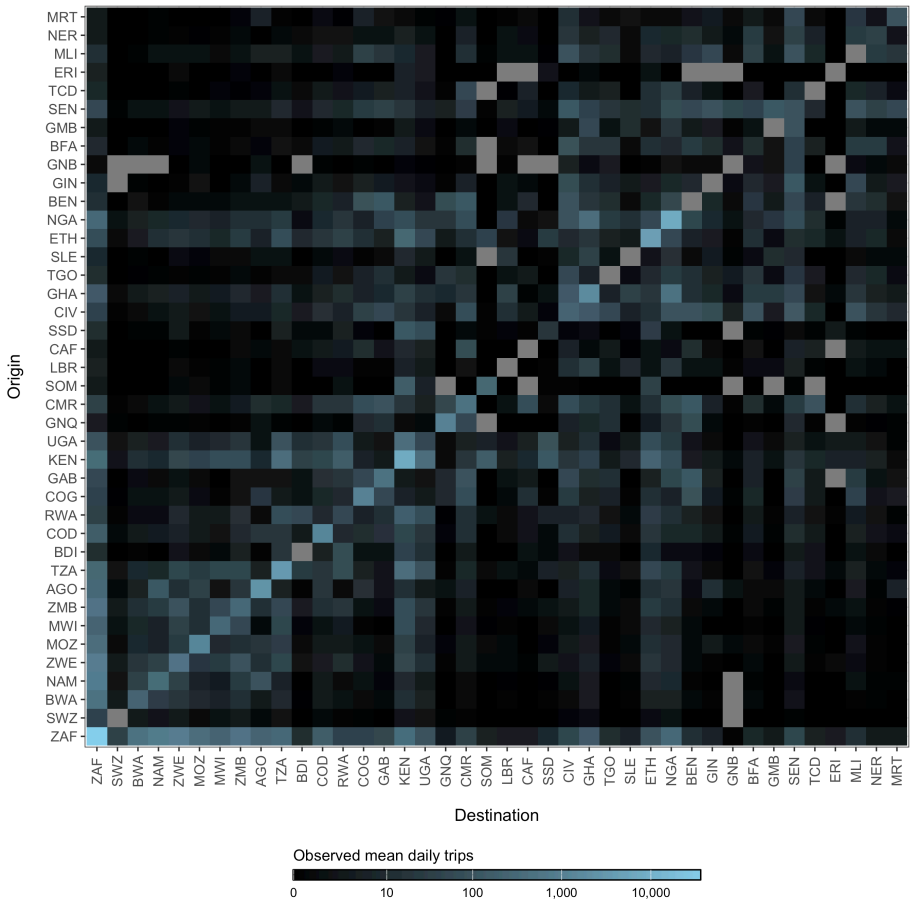

Figure 3.21: The average number of air passengers per day in 2017 among all countries.

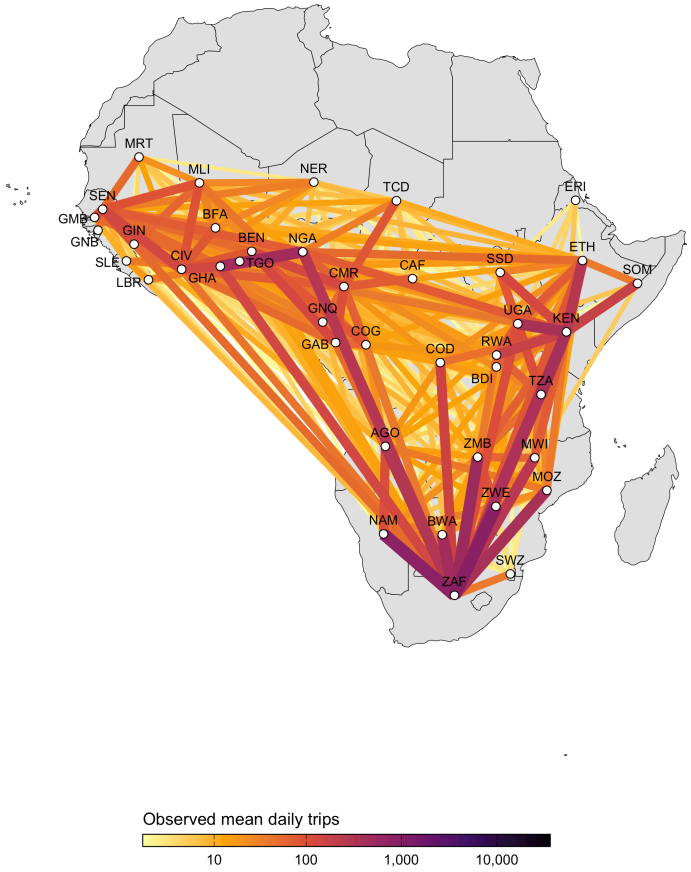

Figure 3.22: A network map showing the average number of air passengers per day in 2017.

The expected mean number of trips for route $i \rightarrow j$ is then:

$$M_{ij} = \begin{cases} \theta N_i (1 - \tau_i) & \text{if } i = j \\ \theta N_i \tau_i \pi_{ij} & \text{if } i \neq j. \end{cases} \tag{3.13}$$

Where, θ is a proportionality constant representing the overall number of trips per person in an origin population of size N_i, τ_i is the probability of leaving origin i, and π_{ij} is the probability of travel to destination j given that travel outside origin i occurs.

3.6.2 Estimating the departure process

The probability of travel outside the origin is estimated for each location i to give the location-specific departure probability τ_i.

$$\tau_i \sim \text{Beta}(1 + s, 1 + r)$$

Binomial probabilities for each origin τ_i are drawn from a Beta distributed prior with shape (s) and rate (r) parameters.

$$s \sim \text{Gamma}(0.01, 0.01)$$
$$r \sim \text{Gamma}(0.01, 0.01)$$

3.6.3 Estimating the diffusion process

We use a normalized formulation of the power law gravity model to defined the diffusion process, the probability of travelling to destination j given travel outside origin i (π_{ij}) which is defined as:

$$\pi_{ij} = \frac{N_j^{\omega} d_{ij}^{-\gamma}}{\sum\limits_{\forall j \neq i} N_j^{\omega} d_{ij}^{-\gamma}} \tag{3.14}$$

Where, ω scales the attractive force of each j destination based on its population size N_j. The kernel function $d_{ij}^{-\gamma}$ serves as a penalty on the proportion of travel from i to j based on distance. Prior distributions of diffusion model parameters are defined as:

$$\omega \sim \text{Gamma}(1, 1)$$
$$\gamma \sim \text{Gamma}(1, 1)$$

The models for τ_i and π_{ij} were fitted to air traffic data from OAG using the `mobility` R package (Giles 2020). Estimates for mobility model parameters are shown in Figures 3.23 and 3.24.

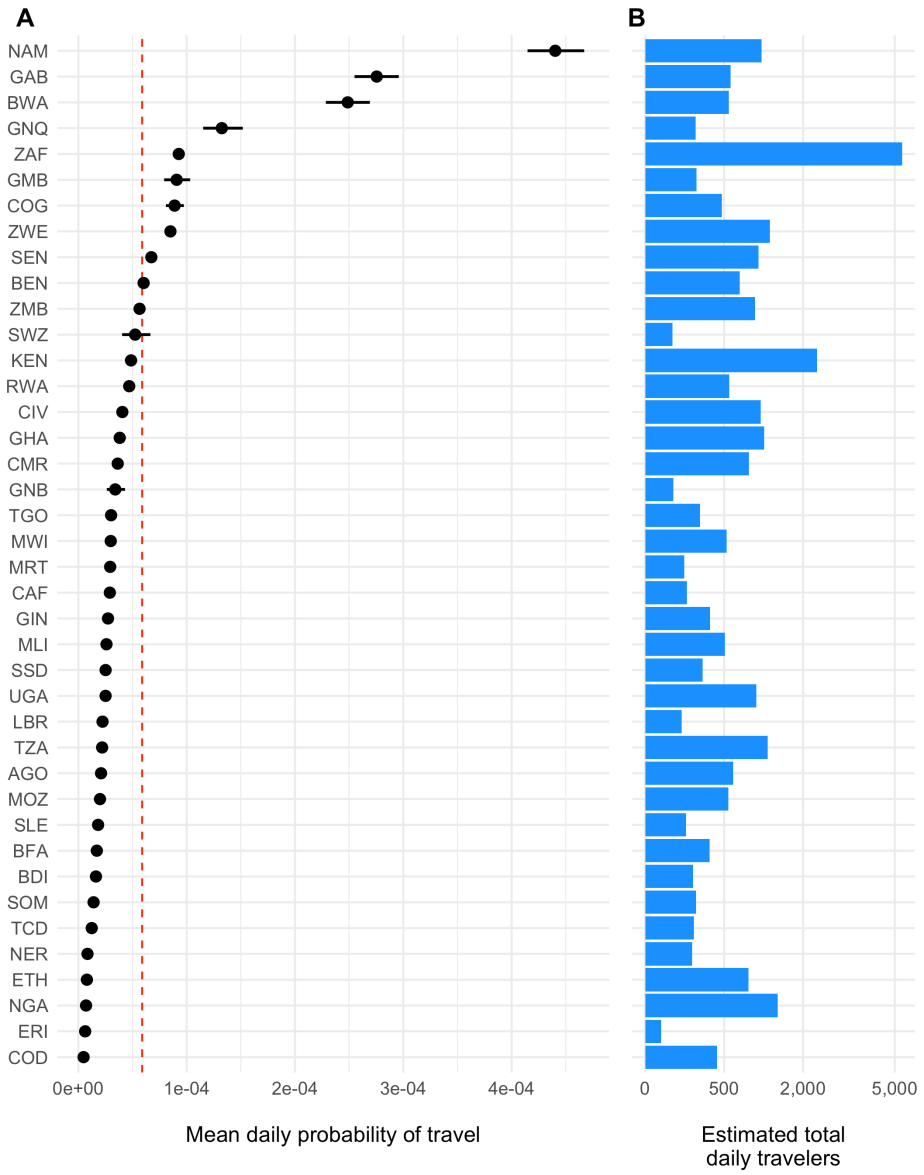

Figure 3.23: The estimated weekly probability of travel outside of each origin location τ_i and 95% confidence intervals is shown in panel A with the population mean indicated as a red dashed line. Panel B shows the estimated total number of travelers leaving origin i each day.

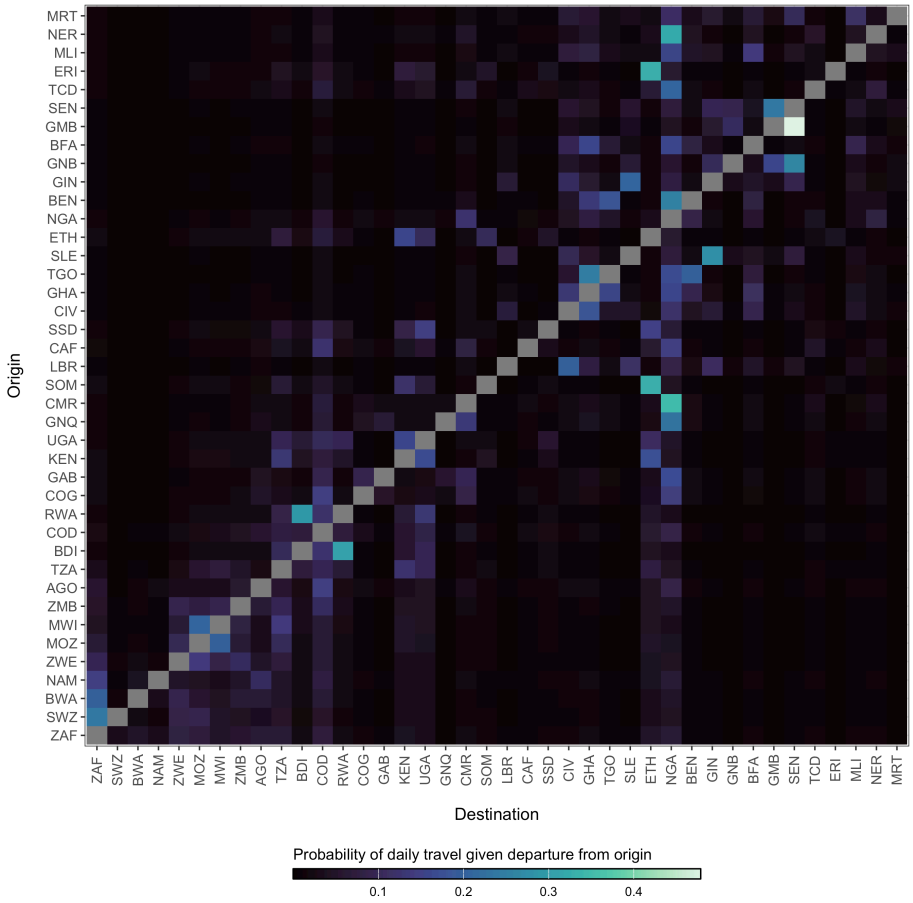

Figure 3.24: The diffusion process π_{ij} which gives the estimated probability of travel from origin i to destination j given that travel outside of origin i has occurred.

3.6.4 The probability of spatial transmission

The likelihood of introductions of cholera from disparate locations is a major concern during cholera outbreaks. However, this can be difficult to characterize given the endemic dynamics and patterns of human movement. We include a few measures of spatial heterogeneity here and the first is a simple importation probability based on connectivity and the possibility of incoming infections. The basic probability of transmission from an origin i to a particular destination j and time t is defined as:

$$p(i,j,t) = 1 - e^{-\beta_{jt}^{\text{hum}}(((1-\tau_j)S_{jt})/N_{jt})\pi_{ij}\tau_i I_{it}} \tag{3.15}$$

3.6.5 The spatial hazard

Although we are more concerned with endemic dynamics here, there are likely to be periods of time early in the rainy season where cholera cases and the rate of transmission is low enough for spatial spread to resemble epidemic dynamics for a time. During such times periods, we can estimate the arrival time of contagion for any location where cases are yet to be reported. We do this be estimating the spatial hazard of transmission:

$$h(j,t) = \frac{\beta_{jt}^{\text{hum}}\Big(1 - \exp\big(-((1-\tau_j)S_{jt}/N_{jt})\sum_{\forall i \neq j}\pi_{ij}\tau_i(I_{it}/N_{it})\big)\Big)}{1/(1+\beta_{jt}^{\text{hum}}(1-\tau_j)S_{jt})}. \tag{3.16}$$

And then normalizing to give the waiting time distribution for all locations:

$$w(j,t) = h(j,T)\prod_{t=1}^{T-1} 1 - h(j,t). \tag{3.17}$$

3.6.6 Coupling among locations

Another measure of spatial heterogeneity is to quantify the coupling of disease dynamics among metapopulations using a correlation coefficient. Here, we use the definition of spatial correlation between locations i and j as C_{ij} described in Keeling and Rohani (2002), which gives a measure of how similar infection dynamics are between locations.

$$C_{ij} = \frac{(y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)}{\sqrt{\text{var}(y_i)\text{var}(y_j)}} \tag{3.18}$$

Where $y_{it} = I_{it}/N_i$ and $y_{jt} = I_{jt}/N_j$. Mean prevalence in each location is $\bar{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$ and $\bar{y_j} = \frac{1}{T}\sum_{t=1}^{T} y_{jt}$.

Table 3.7: Summary of Studies on Cholera Immunity

Mean	Low CI	High CI	Location	Source
0.570	NA	NA	NA	[Nelson et al (2009)](https://www.ncbi.nlm.nih.gov/pmc/articles/
NA	1.000	0.250	NA	[Lueng & Matrajt (2021)](https://journals.plos.org/plosntds/artic
NA	0.600	0.200	Endemic regions	[Harris et al (2012)](https://www.sciencedirect.com/science/articl
0.238	0.250	0.227	Haiti	[Finger et al (2024)](https://www.ncbi.nlm.nih.gov/pmc/articles/
0.213	0.231	0.194	Haiti	[Jackson et al (2013)](https://www.ajtmh.org/view/journals/tpm
0.204	NA	NA	Pakistan	[Bart et al (1970)](https://doi.org/10.1093/infdis/121.Supplement
0.371	NA	NA	Pakistan	[Bart et al (1970)](https://doi.org/10.1093/infdis/121.Supplement
0.184	0.256	0.112	Bangladesh	[Harris et al (2008)](https://journals.plos.org/plosntds/article?id=
0.001	0.000	0.001	Bangladesh	[Hegde et al (2024)](https://www.nature.com/articles/s41591-024

3.7 The observation process

3.7.1 Rate of symptomatic infection

The presentation of infection with *V. cholerae* can be extremely variable. The severity of infection depends many factors such as the amount of the infectious dose, the age of the host, the level of immunity of the host either through vaccination or previous infection, and naivety to the particular strain of *V. cholerae*. Additional circumstantial factors such as nutritional status and overall pathogen burden may also impact infection severity. At the population level, the observed proportion of infections that are symptomatic is also dependent on the endemicity of cholera in the region. Highly endemic areas (e.g. parts of Bangladesh; Hegde et al 2024) may have a very low proportion of symptomatic infections due to many previous exposures. Inversely, populations that are largely naive to *V. cholerae* will exhibit a relatively higher proportion of symptomatic infections (e.g. Haiti; Finger et al 2024).

Accounting for all of these nuances in the first version of this model not possible, but we can past studies do contain some information that can help to set some sensible bounds on our definition for the proportion of infections that are symptomatic (σ). So we have compiled a short list of studies that have done sero-surveys and cohort studies to assess the likelihood of symptomatic infections in different locations and displayed those results in Table (3.7).

To provide a reasonably informed prior for the proportion of infections that are symptomatic, we calculated the combine mean and confidence intervals of all studies in Table 3.7 and fit a Beta distribution that corresponds to these quantiles using least-squares and a Nelder-Mead algorithm. The resulting prior distribution for the symptomatic proportion σ is:

$$\sigma \sim \text{Beta}(4.30, 13.51) \tag{3.19}$$

The prior distribution for σ is plotted in Figure 3.25A with the reported values of the proportion symptomatic from previous studies shown in 3.25B.

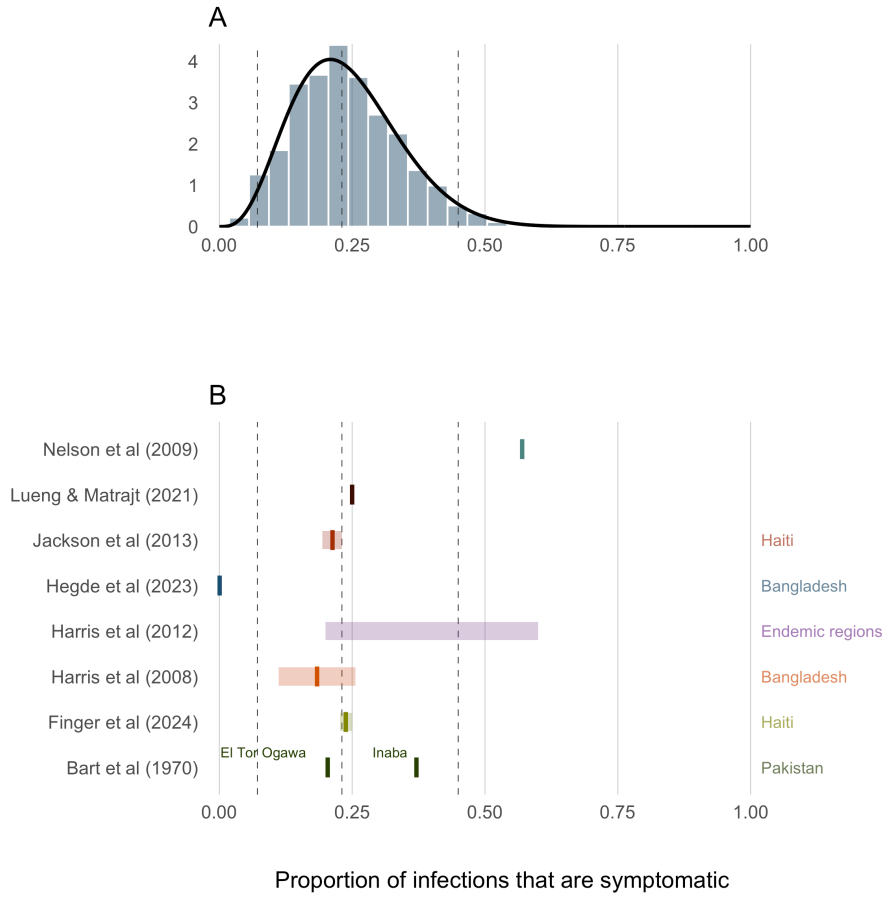

Figure 3.25: Proportion of infections that are symptomatic.

3.7.2 Suspected cases

The clinical presentation of diarrheal diseases is often similar across various pathogens, which can lead to systematic biases in the reported number of cholera cases. It is anticipated that the number of suspected cholera cases is related to the actual number of infections by a factor of $1/\rho$, where ρ represents the proportion of suspected cases that are true infections. To adjust for this bias, we use estimates from the meta-analysis by Weins et al. (2023), which suggests that suspected cholera cases outnumber true infections by approximately 2 to 1,

with a mean across studies indicating that 52% (24-80% 95% CI) of suspected cases are actual cholera infections. A higher estimate was reported for ourbreak settings (78%, 40-99% 95% CI). To account for the variability in this estimate, we fit a Beta distribution to the reported quantiles using a least squares approach and the Nelder-Mead algorithm, resulting in the prior distribution shown in Figure 3.26B:

$$\rho \sim \text{Beta}(4.79, 1.53). \tag{3.20}$$

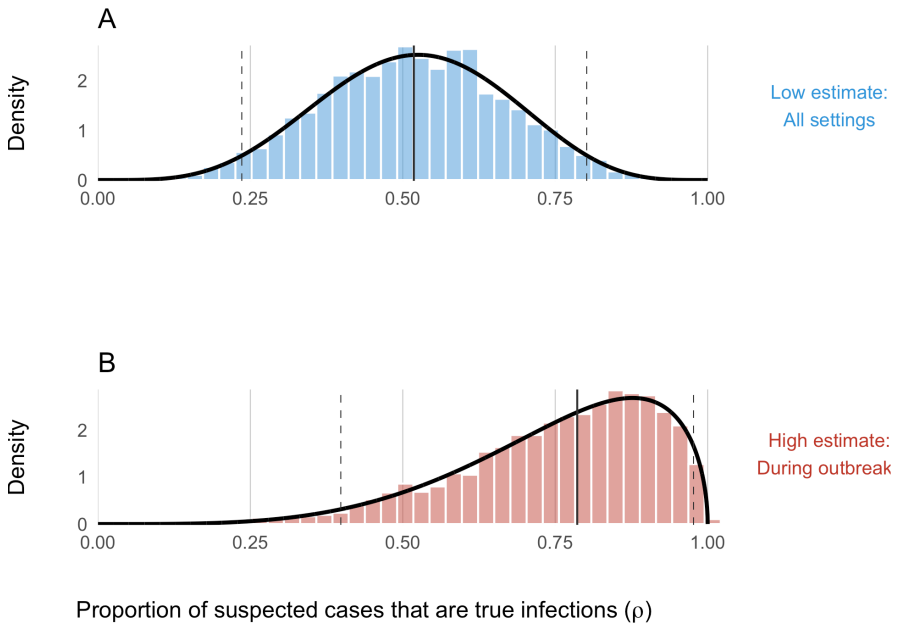

Figure 3.26: Proportion of suspected cholera cases that are true infections. Panel A shows the 'low' assumption which estimates across all settings: $\rho \sim \text{Beta}(5.43, 5.01)$. Panel B shows the 'high' assumption where the estimate reflects high-quality studies during outbreaks: $\rho \sim \text{Beta}(4.79, 1.53)$

3.7.3 Case fatality rate

The Case Fatality Rate (CFR) among symptomatic infections was calculated using reported cases and deaths data from January 2021 to August 2024. The data were collated from various issues of the WHO Weekly Epidemiological Record the Global Cholera and Acute Watery Diarrhea (AWD) Dashboard (see Data section) which provide annual aggregations of reported cholera cases and deaths. We then used the Binomial exact test (`binom.test` in R) to calculate the mean

probability for the number of deaths (successes) given the number of reported cases (sample size), and the Clopper-Pearson method for calculating the binomial confidence intervals. We then fit Beta distributions to the mean CFR and 95% confidence intervals calculated for each country using least squares and the Nelder-Mead algorithm to give the distributional uncertainty around the CFR estimate for each country (μ_j).

$$\mu_j \sim \text{Beta}(s_{1,j}, s_{2,j})$$

Where $s_{1,i}$ and $s_{2,j}$ are the two positive shape parameters of the Beta distribution estimated for destination j. By definition μ_j is the CFR for reported cases which are a subset of the total number of infections. Therefore, to infer the total number of deaths attributable to cholera infection, we assume that the CFR of observed cases is proportionally equivalent to the CFR of all cases and then calculate total deaths D as follows:

$$\begin{aligned} \text{CFR}_{\text{observed}} &= \text{CFR}_{\text{total}} \\ \frac{[\text{observed deaths}]}{[\text{observed cases}]} &= \frac{[\text{total deaths}]}{[\text{all infections}]} \\ \text{total deaths} &= \frac{[\text{observed deaths}] \times [\text{true infections}]}{[\text{observed cases}]} \\ D_{jt} &= \frac{[\sigma\rho\mu_j I_{jt}] \times [I_{jt}]}{[\sigma\rho I_{jt}]} \end{aligned} \tag{3.21}$$

3.8 Demographics

The model includes basic demographic change by using reported birth and death rates for each of the j countries, b_j and d_j respectively. These rates are static and defined by the United Nations Department of Economic and Social Affairs Population Division World Population Prospects 2024. Values for b_j and d_j are derived from crude rates and converted to birth rate per day and death rate per day (shown in Table 3.9).

3.9 The reproductive number

The reproductive number is a common metric of epidemic growth that represents the average number of secondary cases generated by a primary case at a specific time during an epidemic. We track how R changes over time by estimating

Table 3.8: CFR Values and Beta Shape Parameters for AFRO Countries

Country	Cases (2014-2024)	Deaths (2014-2024)	CFR	CFR Lower	CFR Upper	
AFRO Region	1290616	24610	0.019	0.019	0.019	
Angola	3881	122	0.031	0.026	0.037	
Burundi	5695	41	0.007	0.005	0.010	
Benin	3617	56	0.015	0.012	0.020	
Burkina Faso	7	0	0.019	0.019	0.019	
Cote d'Ivoire	446	18	0.040	0.024	0.063	
Cameroon	29978	926	0.031	0.029	0.033	
Democratic Republic of Congo	324021	5857	0.018	0.018	0.019	
Congo	144	10	0.019	0.019	0.019	
Comoros	11171	153	0.014	0.012	0.016	
Ethiopia	73920	928	0.013	0.012	0.013	
Ghana	35107	293	0.008	0.007	0.009	
Guinea	1	0	0.019	0.019	0.019	
Guinea-Bissau	11	2	0.019	0.019	0.019	
Kenya	47956	683	0.014	0.013	0.015	
Liberia	580	0	0.000	0.000	0.006	
Mali	12	4	0.019	0.019	0.019	
Mozambique	85493	335	0.004	0.004	0.004	
Malawi	62916	1859	0.030	0.028	0.031	
Namibia	485	13	0.027	0.014	0.045	
Niger	12705	357	0.028	0.025	0.031	
Nigeria	265652	7242	0.027	0.027	0.028	
Rwanda	453	0	0.000	0.000	0.008	
Sudan	362	11	0.030	0.015	0.054	
Somalia	134839	1849	0.014	0.013	0.014	
South Sudan	56108	1140	0.020	0.019	0.022	
Eswatini	2	0	0.019	0.019	0.019	
Chad	1359	90	0.066	0.054	0.081	
Togo	771	38	0.049	0.035	0.067	
Tanzania	45865	667	0.015	0.013	0.016	
Uganda	9286	182	0.020	0.017	0.023	
South Africa	1403	47	0.033	0.025	0.044	
Zambia	30686	894	0.029	0.027	0.031	
Zimbabwe	45684	793	0.017	0.016	0.019	

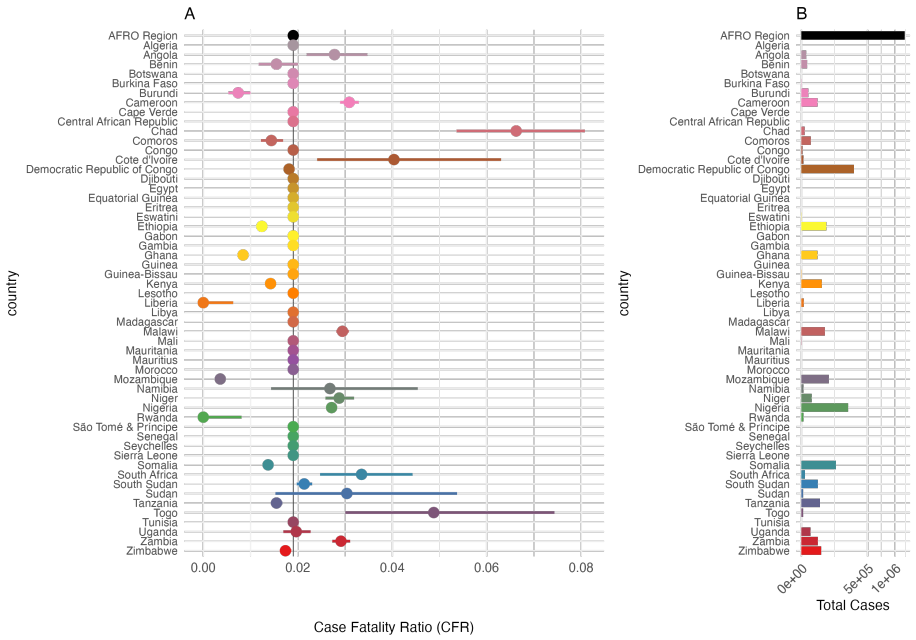

Figure 3.27: Case Fatality Rate (CFR) and Total Cases by Country in the AFRO Region from 2014 to 2024. Panel A: Case Fatality Ratio (CFR) with 95% confidence intervals. Panel B: total number of cholera cases. The AFRO Region is highlighted in black, all countries with less than $3/0.2 = 150$ total reported cases are assigned the mean CFR for AFRO.

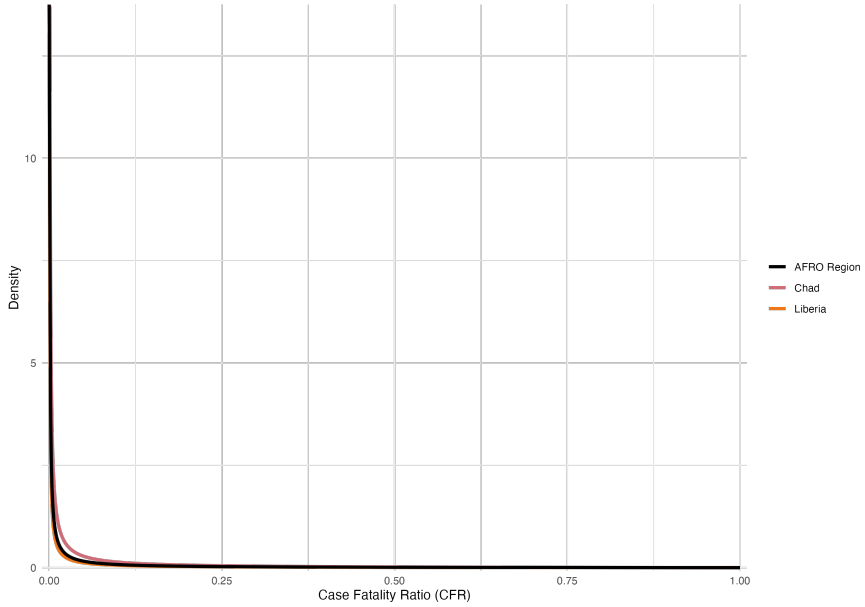

Figure 3.28: Beta distributions of the overall Case Fatality Rate (CFR) from 2014 to 2024. Examples show the overall CFR for the AFRO region (2%) in black, Congo with the highest CFR (7%) in red, and South Sudan with the lowest CFR (0.1%) in blue.

Table 3.9: Demographic for AFRO countries in 2023. Data include: total population as of January 1, 2023, daily birth rate, and daily death rate. Values are calculate from crude birth and death rates from UN World Population Prospects 2024.

Country	Population	Birth rate	Death rate
Algeria	45831343	0.0000542	1.28e-05
Angola	36186956	0.0001046	1.93e-05
Benin	13934166	0.0000940	2.44e-05
Botswana	2459937	0.0000683	1.58e-05
Burkina Faso	22765636	0.0000877	2.21e-05
Burundi	13503998	0.0000935	1.87e-05
Cameroon	27997833	0.0000937	1.99e-05
Cape Verde	521047	0.0000339	1.39e-05
Central African Republic	5064592	0.0001292	2.63e-05
Chad	18767684	0.0001196	3.11e-05
Comoros	842267	0.0000793	1.99e-05
Congo	6108142	0.0000849	1.74e-05
Côte d'Ivoire	30783520	0.0000887	2.12e-05
Democratic Republic of Congo	104063312	0.0001150	2.37e-05
Equatorial Guinea	1825480	0.0000821	2.18e-05
Eritrea	3438999	0.0000789	1.67e-05
Eswatini	1224706	0.0000663	2.12e-05
Ethiopia	127028360	0.0000886	1.65e-05
Gabon	2457715	0.0000766	1.74e-05
Gambia	2666786	0.0000843	1.74e-05
Ghana	33467371	0.0000728	1.95e-05
Guinea	14229395	0.0000939	2.53e-05
Guinea-Bissau	2129290	0.0000832	1.95e-05
Kenya	54793511	0.0000750	2.00e-05
Lesotho	2298496	0.0000664	2.93e-05
Liberia	5432670	0.0000858	2.24e-05
Madagascar	30813475	0.0000890	2.09e-05
Malawi	20832833	0.0000871	1.49e-05
Mali	23415909	0.0001113	2.40e-05
Mauritania	4948362	0.0000957	1.54e-05
Mauritius	1274659	0.0000254	2.39e-05
Mozambique	33140626	0.0001042	1.95e-05
Namibia	2928037	0.0000718	1.71e-05
Niger	25727295	0.0001167	2.47e-05
Nigeria	225494749	0.0000912	3.25e-05
Rwanda	13802596	0.0000785	1.64e-05
São Tomé & Príncipe	228558	0.0000780	1.54e-05
Senegal	17867073	0.0000816	1.55e-05
Seychelles	126694	0.0000377	2.27e-05
Sierra Leone	8368119	0.0000848	2.30e-05
Somalia	18031404	0.0001198	2.74e-05
South Africa	62796883	0.0000518	2.55e-05
South Sudan	11146895	0.0000807	2.71e-05
Tanzania	65657004	0.0000979	1.61e-05
Togo	9196283	0.0000863	2.13e-05
Uganda	47981110	0.0000978	1.35e-05
Zambia	20430382	0.0000919	1.45e-05
Zimbabwe	16203259	0.0000840	2.10e-05

the instantaneous reproductive number R_t as described in Cori et al 2013. We track R_t across all metapopulations in the model to give R_{jt} using the following formula:

$$R_{jt} = \frac{I_{jt}}{\sum_{\Delta t=1}^{t} g(\Delta t) I_{j,t-\Delta t}} \tag{3.22}$$

Where I_{jt} is the number of new infections in destination j at time t, and $g(\Delta t)$ represents the probability value from the generation time distribution of cholera. This is accomplished by using the weighed sum in the denominator which is highly influenced by the generation time distribution.

3.10 Initial conditions

Since the first version of the model begins in January 2023 (to leverage available weekly data), we must estimate the initial state of population immunity. Our approach is as follows:

1. **Reported Cases and Infections:**
   We start by using historical data to determine the total number of reported cholera cases for a given location over the previous X years. Because only symptomatic cases are reported, we multiply the reported case counts by $1/\sigma$ (where σ is the proportion of infections that are symptomatic) to approximate the total number of infections.

2. **Natural Immunity:**
   Next, we estimate the number of individuals who acquired immunity through natural infection during the past X years. This involves adjusting the total infections by accounting for immune decay, governed by the parameter ε.

3. **Vaccine-derived Immunity:**
   We also sum the total number of vaccinations administered over the past X years. This number is adjusted by the vaccine effectiveness (denoted ϕ) and the waning immunity rate (ω) to estimate the current number of individuals with vaccine-derived immunity.

4. **Deconvolution:**
   Finally, we combine these estimates using a deconvolution approach based on the estimated immune decay parameters (from both vaccination and natural infection) to set the model's initial conditions.

In total, the initial conditions reflect: - The estimated total number infected (backed out from reported cases), - The number immune due to natural infection, and - The number immune from past vaccination campaigns.

3.11 Model calibration

Model calibration is performed to fine-tune the hyperparameters and ensure the model accurately represents the observed data. Our calibration strategy involves:

- **Latin Hypercube Sampling (LHS):**
  We use LHS to explore the parameter space efficiently. This method helps generate a diverse set of hyperparameter combinations for evaluation.

- **Likelihood Fitting:**
  For each set of hyperparameters, the model likelihood is computed based on the observed incidence and death data. The calibration process searches for the hyperparameter set that maximizes the model's likelihood.

- **Data Challenges:**
  A key challenge in calibration is the incomplete or aggregated nature of the available data. To address this, we incorporate methods that allow flexible fitting even when data are sparse or reported at different temporal scales.

By combining LHS and likelihood-based calibration, we aim to identify a robust set of hyperparameters that accurately capture both the temporal and spatial dynamics of cholera transmission in Sub-Saharan Africa.

Table 3.10: Listof MOSAIC Countries with Cholera News

ISO	Country	Region	Cholera News
BDI	Burundi	Eastern Africa	[Cholera News: Burundi](https://news.google.com
ERI	Eritrea	Eastern Africa	[Cholera News: Eritrea](https://news.google.com/
ETH	Ethiopia	Eastern Africa	[Cholera News: Ethiopia](https://news.google.com
KEN	Kenya	Eastern Africa	[Cholera News: Kenya](https://news.google.com/s
MWI	Malawi	Eastern Africa	[Cholera News: Malawi](https://news.google.com/
MOZ	Mozambique	Eastern Africa	[Cholera News: Mozambique](https://news.google
RWA	Rwanda	Eastern Africa	[Cholera News: Rwanda](https://news.google.com
SOM	Somalia	Eastern Africa	[Cholera News: Somalia](https://news.google.com
SSD	South Sudan	Eastern Africa	[Cholera News: South Sudan](https://news.google
TZA	Tanzania	Eastern Africa	[Cholera News: Tanzania](https://news.google.co
UGA	Uganda	Eastern Africa	[Cholera News: Uganda](https://news.google.com
ZMB	Zambia	Eastern Africa	[Cholera News: Zambia](https://news.google.com/
ZWE	Zimbabwe	Eastern Africa	[Cholera News: Zimbabwe](https://news.google.co
AGO	Angola	Middle Africa	[Cholera News: Angola](https://news.google.com/
CMR	Cameroon	Middle Africa	[Cholera News: Cameroon](https://news.google.co
CAF	Central African Republic	Middle Africa	[Cholera News: Central African Republic](https:/
TCD	Chad	Middle Africa	[Cholera News: Chad](https://news.google.com/se
COD	Democratic Republic of the Congo	Middle Africa	[Cholera News: Democratic Republic of the Congo
GNQ	Equatorial Guinea	Middle Africa	[Cholera News: Equatorial Guinea](https://news.g
GAB	Gabon	Middle Africa	[Cholera News: Gabon](https://news.google.com/
COG	Republic of the Congo	Middle Africa	[Cholera News: Republic of the Congo](https://ne
BWA	Botswana	Southern Africa	[Cholera News: Botswana](https://news.google.co
SWZ	Kingdom of eSwatini	Southern Africa	[Cholera News: Kingdom of eSwatini](https://new
NAM	Namibia	Southern Africa	[Cholera News: Namibia](https://news.google.com
ZAF	South Africa	Southern Africa	[Cholera News: South Africa](https://news.google
BEN	Benin	Western Africa	[Cholera News: Benin](https://news.google.com/s
BFA	Burkina Faso	Western Africa	[Cholera News: Burkina Faso](https://news.google
CIV	Côte d'Ivoire	Western Africa	[Cholera News: Côte d'Ivoire](https://news.google
GHA	Ghana	Western Africa	[Cholera News: Ghana](https://news.google.com/
GIN	Guinea	Western Africa	[Cholera News: Guinea](https://news.google.com/
GNB	Guinea-Bissau	Western Africa	[Cholera News: Guinea-Bissau](https://news.goog
LBR	Liberia	Western Africa	[Cholera News: Liberia](https://news.google.com/
MLI	Mali	Western Africa	[Cholera News: Mali](https://news.google.com/sea
MRT	Mauritania	Western Africa	[Cholera News: Mauritania](https://news.google.c
NER	Niger	Western Africa	[Cholera News: Niger](https://news.google.com/se
NGA	Nigeria	Western Africa	[Cholera News: Nigeria](https://news.google.com/
SEN	Senegal	Western Africa	[Cholera News: Senegal](https://news.google.com/
SLE	Sierra Leone	Western Africa	[Cholera News: Sierra Leone](https://news.google
GMB	The Gambia	Western Africa	[Cholera News: The Gambia](https://news.google
TGO	Togo	Western Africa	[Cholera News: Togo](https://news.google.com/se

3.12 Table of MOSAIC framework countries

3.13 Table of model parameters

Parameter	Description
i	Index representing the origin metapopulation.
j	Index representing the destination metapopulation.
t	Time step (one week).
b_{jt}	Birth rate of population j.
d_{jt}	Mortality rate of population j.
N_{jt}	Population size of destination j at time t.
S_{jt}	Number of susceptible individuals in destination j at time t.
$V_{1,jt}$	Number of individuals with one-dose vaccination in destination j at time t.
$V_{2,jt}$	Number of individuals with two-dose vaccination in destination j at time t.
$I_{1,jt}$	Number of symptomatic infected individuals in destination j at time t.
$I_{2,jt}$	Number of asymptomatic infected individuals in destination j at time t.
W_{jt}	Amount of *V. cholerae* in the environment in destination j at time t.
R_{jt}	Number of recovered (immune) individuals in destination j at time t.
$\Lambda_{j,t+1}$	Human-to-human force of infection in destination j at time $t+1$.
$\Psi_{j,t+1}$	Environment-to-human force of infection in destination j at time $t+1$.
ι	The incubation period of cholera infection
ϕ_1	Vaccine effectiveness of one-dose OCV.
ϕ_2	Vaccine effectiveness of two-dose OCV.
ν_{jt}	Vaccination rate (OCV doses administered) in destination j at time t.
ω_1	Waning immunity rate of vaccinated individuals with one-dose OCV.
ω_2	Waning immunity rate of vaccinated individuals with two-dose OCV.
ε	Waning immunity rate of recovered individuals.
γ_1	Recovery rate of symptomatic infected individuals.
γ_2	Recovery rate of asymptomatic infected individuals.
μ	Mortality rate due to *V. cholerae* infection.
σ	Proportion of infections that are symptomatic.
ρ	Proportion of suspected cases that are true infections.
ζ_1	Shedding rate of *V. cholerae* by symptomatic individuals.
ζ_2	Shedding rate of *V. cholerae* by asymptomatic individuals.
δ	Environmental decay rate of *V. cholerae*.
δ_{min}	Minimum decay rate when $\psi_{jt}=0$.
δ_{max}	Maximum decay rate when $\psi_{jt}=1$.
ψ_{jt}	Environmental suitability of *V. cholerae* in destination j at time t.
β_{j0}^{hum}	Baseline human-to-human transmission rate in destination j.
β_{jt}^{hum}	Seasonal human-to-human transmission rate in destination j at time t.
β_{j0}^{env}	Baseline environment-to-human transmission rate in destination j.
β_{jt}^{env}	Environment-to-human transmission rate in destination j at time t.
a_1	First Fourier cosine coefficient for seasonality.
b_1	First Fourier sine coefficient for seasonality.
a_2	Second Fourier cosine coefficient for seasonality.
b_2	Second Fourier sine coefficient for seasonality.
p	Period of the seasonal cycle (set to days).
α_1	Exponent on infectious individuals in the force of infection numerator.
α_2	Exponent on population size in the force of infection denominator; determines den
τ_i	Probability an individual departs from origin i.
π_{ij}	Probability of travel from origin i to destination j given departure.
θ_{j}	Proportion with adequate WASH in destination j.
κ	Concentration of *V. cholerae* (cells/mL) required for 50% infection probability.

3.14 Table of stochastic transitions

Term	Description
$\mathbf{S}$ (susceptible)	
$+ b_{jt} N_{jt}$	New individuals entering the susceptible class from births.
$+ \varepsilon R_{jt}$	Loss of immunity for recovered individuals.
$+ \omega_1 V_{1,jt}$	Waning immunity from one-dose OCV.
$+ \omega_2 V_{2,jt}$	Waning immunity from two-dose OCV.
$- \nu_{1,jt}S_{jt}/(S_{jt} + E_{jt})$	Susceptible individuals receiving one-dose OCV (leaving S).
$- \Lambda^{S}_{j,t+1}$	Human-to-human force of infection on the susceptible class.
$+ \Psi^S_{j,t+1}$	Environment-to-human force of infection on the susceptible class.
$- d_{jt} S_{jt}$	Background death among susceptible individuals.
$\mathbf{V_1}$ (one-dose OCV)	
$+ \nu_{1,jt} S_{jt}/(S_{jt} + E_{jt})$	Entry of susceptible individuals into the one-dose vaccinated class.
$- \omega_1 V_{1,jt}$	Waning immunity in the one-dose vaccinated class.
$- \Lambda^{V_1}_{j,t+1}$	Human-to-human force of infection on the one-dose vaccinated class
$+ \Psi^{V_1}_{j,t+1}$	Environment-to-human force of infection on one-dose vaccinated cla
$- d_{jt} V_{1,jt}$	Background death among one-dose vaccinated individuals.
$\mathbf{V_2}$ (two-dose OCV)	
$+ \nu_{2,jt}$	Transition of one-dose vaccinated individuals to the two-dose vaccin
$- \omega_2 V_{2,jt}$	Waning immunity in the two-dose vaccinated class.
$- \Lambda^{V_2}_{j,t+1}$	Human-to-human force of infection on the two-dose vaccinated class
$+ \Psi^{V_2}_{j,t+1}$	Environment-to-human force of infection on the two-dose vaccinated
$- d_{jt} V_{2,jt}$	Background death among two-dose vaccinated individuals.
$\mathbf{E}$ (exposed)	
$+ \Lambda_{j,t+1}$	Human-to-human force of infection contributing to new exposures.
$+ \Psi_{j,t+1}$	Environment-to-human force of infection contributing to new exposu
$- \iota E_{jt}$	Progression of exposed individuals toward the infectious class.
$- d_{jt} E_{jt}$	Background death among exposed individuals.
$\mathbf{I_1}$ (symptomatic)	
$+ \sigma\,\iota\,E_{jt}$	Exposed individuals progressing to symptomatic infection.
$- \gamma_1 I_{1,jt}$	Recovery from symptomatic infection.
$- \mu_j I_{1,jt}$	Deaths due to symptomatic infection.
$- d_{jt} I_{1,jt}$	Background death among individuals with symptomatic infection.
$\mathbf{I_2}$ (asymptomatic)	
$+ (1-\sigma)\,\iota\,E_{jt}$	Exposed individuals progressing to asymptomatic infection.
$- \gamma_2 I_{2,jt}$	Recovery from asymptomatic infection.
$- d_{jt} I_{2,jt}$	Background death among individuals with asymptomatic infection.
$\mathbf{W}$ (environment)	
$+ \zeta_1 I_{1,jt}$	Amount of *V. cholerae* (cells/ml) shed into the environment by sy
$+ \zeta_2 I_{2,jt}$	Amount of *V. cholerae* (cells/ml) shed into the environment by as
$- \delta_{jt} W_{jt}$	Decay of viable *V. cholerae* in the environment.
$\mathbf{R}$ (recovered)	
$+ \gamma_1 I_{1,jt}$	Recovery of individuals with symptomatic infection.
$+ \gamma_2 I_{2,jt}$	Recovery of individuals with asymptomatic infection.
$- \varepsilon R_{jt}$	Loss of immunity for recovered individuals.
$- d_{jt} R_{jt}$	Background death among recovered individuals.

3.15 Table of vaccination model terms

Term	Population	Equation	Notes
$V^{\text{imm}}_{1,j,t+1}$	Effectively immunized one-dose recipients	$V^{\text{imm}}_{1,j,t+1} = V^{\text{imm}}_{1,jt} + \phi_1\nu_{1,jt}\cdot S_{jt}/(S_{jt}+E_{jt}) - \omega_1 V^{\text{imm}}_{1,jt} - \nu_{2,jt}\cdot V^{\text{imm}}_{1,jt}/(V^{\text{imm}}_{1,jt}+V^{\text{sus}}_{1,jt})$	+ Incoming newly vaccinated - Waning vaccine immunity (V^{sus}_1) - Second dose recipients (V_2 compartment)
$V^{\text{sus}}_{1,j,t+1}$	Still susceptible one-dose recipients	$V^{\text{sus}}_{1,j,t+1} = V^{\text{sus}}_{1,jt} + (1-\phi_1)\nu_{1,jt} + \omega_1 V^{\text{imm}}_{1,jt} - (\Lambda^{V_1}_{j,t+1}+\Psi^{V_1}_{j,t+1}) - \nu_{2,jt}\cdot V^{\text{sus}}_{1,jt}/(V^{\text{imm}}_{1,jt}+V^{\text{sus}}_{1,jt})$	+ Incoming newly vaccinated + Waning vaccine immunity - Infected ($E_{j,t}$) - Second dose recipients (V_2 compartment)
$V^{\text{inf}}_{1,j,t+1}$	Infected one-dose recipients	$V^{\text{inf}}_{1,j,t+1} = V^{\text{inf}}_{1,jt} + (\Lambda^{V_1}_{j,t+1}+\Psi^{V_1}_{j,t+1})$	+ One-dose recipients infected ($E_{j,t}$) **Compartment used for tracking only.**
$V^{\text{imm}}_{2,j,t+1}$	Effectively immunized two-dose recipients	$V^{\text{imm}}_{2,j,t+1} = V^{\text{imm}}_{2,jt} + \phi_2\nu_{2,jt} - \omega_2 V^{\text{imm}}_{2,jt}$	+ Incoming second dose recipients - Waning vaccine immunity (V^{sus}_2)
$V^{\text{sus}}_{2,j,t+1}$	Still susceptible two-dose recipients	$V^{\text{sus}}_{2,j,t+1} = V^{\text{sus}}_{2,jt} + (1-\phi_2)\nu_{2,jt} + \omega_2 V^{\text{imm}}_{2,jt} - (\Lambda^{V_2}_{j,t+1}+\Psi^{V_2}_{j,t+1})$	+ Incoming second dose recipients + Waning vaccine immunity - Infected ($E_{j,t}$)

Term	Population	Equation	Notes
$V^{\text{inf}}_{2,j,t+1}$	Infected two-dose recipients	$V^{\text{inf}}_{2,j,t+1} = V^{\text{inf}}_{2} + (\Lambda^{V_2}_{j,t+1} + \Psi^{V_2}_{j,t+1})$	+ Infected two-dose recipients ($E_{j,t}$). **Compartment used for tracking only.**
$V_{1,j,t}$	Total one-dose recipients	$V_{1,j,t} = V^{\text{imm}}_{1,j,t} + V^{\text{sus}}_{1,j,t} + V^{\text{inf}}_{1,j,t}$	Sum of all one-dose sub-compartments. Tracked only and approximately equal to reported OCV campaign data. **Compartment used for tracking only.**
$V_{2,j,t}$	Total two-dose recipients	$V_{2,j,t} = V^{\text{imm}}_{2,j,t} + V^{\text{sus}}_{2,j,t} + V^{\text{inf}}_{2,j,t}$	Sum of all two-dose sub-compartments. Tracked only and approximately equal to reported OCV campaign data. **Compartment used for tracking only.**

Chapter 4

Model calibration

4.1 Bayesian Likelihood Approach

The MOSAIC framework employs Bayesian inference to calibrate its spatial transmission model. As in many other algorithms that use Bayesian inference, the model systematically estimates parameters based on their ability to recreate the observed data, which is measured through a *likelihood function*. One major assumption of the Bayesian method is that all model parameters—and most importantly, the link between model and data—have a known probability distribution (e.g. Normal, Poisson, Uniform). Therefore, all parameters in the MOSAIC framework have a prior distribution (before model calibration) that is highly-informed by other data sources and meta-analyses (see the Model Description page).

To calibrate the MOSAIC model to observed cholera surveillance data, the algorithm updates prior beliefs about model parameters through the likelihood function $\mathcal{L}(\Theta)$. The likelihood is essentially a function of model parameters that measures how probable our particular model is given the observed data, or more formally, the posterior probability distribution of the parameter vector Θ. This parameter vector includes all quantities required for a single iteration of the model, which includes transmission rates, mobility parameters, and seasonal forcing coefficients for example (see the Table of Model Parameters).

$$\Theta = \{ \beta, \gamma, \omega, a_1, a_2, b_1, b_2, ... \} \tag{4.1}$$

During model calibration, we aim to identify the best set of model parameters that maximize the log-likelihood while sampling from the large parameter space of Θ using the a brute force sampling algorthim (more details below):

$$\hat{\Theta} = \arg\max_{\Theta} [\log \mathcal{L}(\Theta)]. \tag{4.2}$$

To specify the likelihood function, we treat $\mathcal{L}(\Theta)$ as the posterior density of Θ given the observed cholera surveillance data, and then use the Bayes' theorem to set up the model-data link.

$$\mathcal{L}(\Theta) \Longrightarrow P(\Theta \mid \text{data}) \tag{4.3}$$

Because the observed data $P(\text{data})$ does not depend on the model parameters and the prior $P(\Theta)$ is assumed to be uniform, these two terms can be treated as constants in Bayes' theorem. Consequently, the posterior density is proportional to the likelihood:

$$P(\Theta \mid \text{data}) = \frac{P(\text{data} \mid \Theta) \overbrace{P(\Theta)}^{\text{constant}}}{\underbrace{P(\text{data})}_{\text{constant}}} \propto P(\text{data} \mid \Theta), \tag{4.4}$$

therefore maximizing the posterior (or minimizing its negative log) is equivalent to maximizing the likelihood $P(\text{data} \mid \Theta)$, and we can now construct the likelihood function using the relevant probability density functions $f(y \mid \mu)$ as described below.

4.2 Total Log-likelihood for Cases and Deaths

Because the model posterior is proportional to $P(\text{data} \mid \Theta)$, we constructed the likelihood function with the appropriate distribution for each of the observed data types using common notation for a probability density function $f(y \mid \mu)$. The MOSAIC framework is a spatial model, so we also included the J spatial locations and T time points in the full likelihood, which gives the product over both indices:

$$P\left(\text{data} \mid \Theta\right) = \prod_{j=1}^{J} \prod_{t=1}^{T} f\big(y_{jt} \mid \mu_{jt}(\Theta)\big)\,, \tag{4.5}$$

where y_{jt} is the observed count (cases, deaths, etc.) for location j at time t, and $\mu_{jt}(\Theta)$ is the corresponding model-generated mean. Substituting $f(\cdot)$ with the appropriate probability distribution (Poisson, Negative Binomial, etc.) yields the explicit likelihood function used in calibration.

The total log-likelihood combines contributions from observed cases and deaths across locations and time points. This combined likelihood function quantifies the probability of the observed epidemiological data given the model parameters Θ.

$$\log \mathcal{L}(\Theta) = \sum_{j=1}^{J} w_j \left[w_{\text{cases}} \sum_{t=1}^{T} w_t \, \log P(C_{j,t}^{\text{obs}} \mid C_{j,t}^{\text{est}}(\Theta), k_{\text{cases},j}) + w_{\text{deaths}} \sum_{t=1}^{T} w_t \, \log P(D_{j,t}^{\text{obs}} \mid D_{j,t}^{\text{est}}(\Theta), k_{\text{dea}} \right. \tag{4.6}$$

Note that each log-likelihood term is weighted three times — by a location weight w_j, a time-step weight w_t, and an outcome-specific weight w_{cases} or w_{deaths} — so that contributions reflect data reliability and public-health priorities across space, time, and outcome. The choice between Poisson and Negative Binomial for the density $P(\cdot)$ is driven by the local mean-variance relationship (VMR), ensuring that the assumed error structure mirrors the dispersion actually observed in the surveillance data. The next subsection details the parameterisation of each probability distribution and how the corresponding likelihood is computed.

Parameter	Description
J	Number of locations
T	Number of time points
$C_{j,t}^{\text{obs}}$	Observed cases at location j and time t
$D_{j,t}^{\text{obs}}$	Observed deaths at location j and time t
$C_{j,t}^{\text{est}}(\Theta)$	Model-estimated mean cases at location j, time t
$D_{j,t}^{\text{est}}(\Theta)$	Model-estimated mean deaths at location j, time t
w_j	Location-specific weights (reflecting population or data confidence)
w_t	Time-specific weights (typically uniform, $w_t = 1$)
w_{cases}	Relative weight for cases
w_{deaths}	Relative weight for deaths
$k_{\text{cases},j}$	Dispersion parameter for cases at location j
$k_{\text{deaths},j}$	Dispersion parameter for deaths at location j

4.3 Distributional Assumptions for Likelihood Components

For each location j and time step t the density $f\big(y_{jt} \mid \mu_{jt}(\Theta)\big)$ in Equation (4.5) is chosen to match the observed mean–variance relationship at that location, which is calculated as $\text{VMR}_j = \text{Var}(y_{j\cdot})/\text{Mean}(y_{j\cdot})$ from the raw surveillance counts. If $\text{VMR}_j < 1.5$ the data are close to *equi-dispersion* and we adopt a *Poisson* distributed error model. Otherwise, the count data are considered to be *over-dispersed* and we use a *Negative Binomial* error model with a location-specific dispersion parameter k_j.

4.3.1 Negative Binomial density (VMR ≥ 1.5)

$$\log P_{\text{NB}}(y_{jt} \mid \mu_{jt}, k_j) = \log\Gamma(y_{jt}+k_j) - \log\Gamma(k_j) - \log\Gamma(y_{jt}+1) + k_j \log\Big[\tfrac{k_j}{k_j+\mu_{jt}}\Big] + y_{jt} \log\Big[\tfrac{\mu_{jt}}{k_j+\mu_{jt}}\Big] \tag{4.7}$$

The dispersion is estimated*per location via the method-of-moments:

$$k_j \;=\; \frac{\mu_j^2}{\mathrm{Var}(y_{j\cdot}) - \mu_j}, \tag{4.8}$$

so that $\mathrm{Var}(y_{jt}) = \mu_{jt} + \mu_{jt}^2/k_j$. As $k_j \to \infty$ the density in (4.7) collapses smoothly to the Poisson form.

4.3.2 Poisson density (VMR < 1.5)

$$\log P_{\mathrm{Pois}}(y_{jt} \mid \mu_{jt}) \;=\; y_{jt} \log \mu_{jt} - \mu_{jt} - \log(y_{jt}!). \tag{4.9}$$

The automatic Poisson/Negative-Binomial switch ensures that the error structure embedded in the likelihood replicates the empirical dispersion seen in the surveillance data, while the weighting scheme w_j, w_t, w_{cases}, w_{deaths} (introduced in Equation (4.6)) controls the relative influence of each location, time step, and outcome on the overall fit.

4.4 Algorithm for Parameter Estimation

The MOSAIC calibration relies on a *brute-force random sampling* (BFRS) workflow with importance-sampling for estimating posterior parameter distributions. The BFRS approach is deliberately simple, fully parallelisable, and maps directly onto the informative priors which have been painstakingly estimated a priori (see the Model Description page).

Unlike Markov-Chain Monte Carlo (MCMC) sampling methods, the BFRS workflow generates independent parameter draws, so there is no need to worry about convergence diagnostics, burn-in, or autocorrelation, and simulations can be distributed across hundreds of CPUs. The trade-off is efficiency: for a fixed computational budget MCMC can concentrate samples in the highest-posterior region, whereas BFRS spends many draws in moderately likely parts of the space. Although this wastes some compute, the penalty is small because the LASER modelling engine, whose fast, metapopulation implementation can evaluate each $\Theta^{(i)}$ parameter draw in milliseconds.

Relative to Latin-hypercube or Sobol sequence sampling designs, which are also intended to do broad surveys of the parameter space, BFRS keeps the exact prior shape, can be extended at any time by simply adding more draws, and feeds directly into likelihood weighting without extra transformations. In combination with LASER's speed, these features make Bayesian calibration in MOSAIC both fast and easily reproducible.

The steps below summarise how this BFRS workflow is turned into a practical calibration routine—moving from prior draws, through model simulation and likelihood evaluation, to the identification of the best-fitting parameter set.

1. *Generate reproducible parameter sample*

Draw n_{sim} independent parameter vectors $\Theta^{(i)} \sim P(\Theta)$ using predetermined random seeds, ensuring that the calibration can be rerun and audited exactly.

2. *Forward-simulate*

The transitions between most model compartments are stochastic, so for each independent sampling of the parameter space $\Theta^{(i)}$, we run the stochastic transmission model for n_{iter} internal iterations.

3. *Evaluate the fit for every draw*

For each of the $n_{\text{sim}} \times n_{\text{iter}}$ internal iterations, compute the total *negative* log-likelihood $-\log \mathcal{L}(\Theta^{(i)})$ via Equation (4.6).

4. *Post-processing end-points*

- *Posterior parameter distributions* – convert all log-likelihoods to importance weights for estimating marginal posteriors (next section).
- *Bayesian model averaging* – use the weighted ensemble to generate probabilistic forecasts.
- *Best-fit scenario set* – select $\hat{\Theta} = \arg\max_i [\log \mathcal{L}(\Theta^{(i)})]$ for deterministic scenario and counter-factual analyses.

4.5 Estimating the Posterior Distribution of Model Parameters

To transform the BFRS ensemble of samples from the parameter space and corresponding likelihood values $\{\Theta^{(i)}, \log \mathcal{L}(\Theta^{(i)})\}_{i=1}^{n_{\text{sim}}}$ into an legitimate Bayesian posterior, we used Importance Sampling (IS). The IS method a well-known method technique to estimate posterior distributions originally described by Kahn & Marshall 1953 and reviewed in a more modern context by Tokdar & Kass 2010. Thus, we calculate the IS-weights using the ΔAIC with a practical cut–off of $\Delta = 4$ and retain the IS for a subset of supported models as described the in steps below:

4.5.1 Compute ΔAIC for every draw

For any model, the Akaike Information Criterion is AIC $= 2k - 2\log \mathcal{L}$, where k is the number of estimated parameters. Because every MOSAIC simulation has the same k, the ΔAIC of draw i relative to the best draw is determined solely by the difference in log-likelihood.

$$\Delta_i \;=\; \text{AIC}_i - \text{AIC}_{\min} \;=\; -2[\log \mathcal{L}(\Theta^{(i)}) - \log \mathcal{L}_{\max}] \tag{4.10}$$

4.5.2 Assign truncated importance weights

Since the BFRS method generates a large ensemble of candidate parameter sets $\Theta^{(i)}$, we reduce the influence of poorly fitting models by truncating the importance weights using a ΔAIC cut-off. This ensures that only models with substantially better fit to the data contribute to the posterior.

$$\tilde{w}_i \;=\; \begin{cases} \exp[-\frac{1}{2}\Delta_i], & \Delta_i \le 4, \\ 0, & \Delta_i > 4, \end{cases} \qquad \text{and} \qquad \tilde{w}_i \;=\; \frac{w_i}{\sum_{j=1}^{n_{\text{sim}}} w_j} \tag{4.11}$$

The threshold of $\Delta_i \le 4$ is widely used in model selection and corresponds approximately to a likelihood ratio of $\exp(-2) \approx 0.14$, which in nested-model comparisons aligns loosely with a frequentist p-value of 0.05 (Burnham & Anderson 2002). This cut-off removes models with essentially no empirical support, while preserving relative likelihood ratios among the retained models.

4.5.3 Posterior summaries

Because the vector of truncated ΔAIC weights $\tilde{\mathbf{w}}$ are proportional to the posterior density $P(\Theta \mid \text{data})$, we estimate the true Bayesian posterior distributions of each fitted model parameter as a weighted empirical statistic. Take for example the scalar σ, which gives the proportion of infections that are symptomatic. It is an element of each $\Theta^{(i)}$ sample of the parameter space, so $\{\sigma^{(i)}\}_{i=1}^{n_{\text{sim}}}$ gives all σ values for which a likelihood has been calculated. Therefore, we derive the posterior mean and 95% credible intervals for σ as:

$$\mathbb{E}[\sigma] \;=\; \sum_{i=1}^{n_{\text{sim}}} \tilde{w}_i\, \sigma^{(i)} \qquad \text{and} \qquad 95\%\,\text{CI} \;=\; [\,Q_{0.025}^{(\tilde{w})},\; Q_{0.975}^{(\tilde{w})}]. \tag{4.12}$$

Where $Q_p^{(\tilde{w})}$ denotes the $\tilde{w}$-weighted p-th quantile of the retained ensemble, ensuring that the resulting intervals are fully consistent with the Bayesian posterior implied by the importance-sampling weights.

4.6 Model convergence

Because the MOSAIC framework uses an i.i.d. brute-force random sampling (BFRS) scheme for model fitting, traditional chain-based diagnostics such as $\hat{R}$ are not relevant. Instead we track three complementary weight–based statistics that together tell us *how many* models inform the posterior, *how strongly* they agree, and *how evenly* their support is distributed. These metrics also provide mathematical criteria that supported by previous theory and empirical studies, which is crucial for assessing model convergence. Specifically these metrics are:

1. **Effective sample size** — $\widehat{\mathrm{ESS}}$
   gauges the amount of independent posterior information retained in the subset of best fitting model runs.
2. **Agreement index** — A
   measures the level of consensus among the retained best subset models.
3. **Coefficient of variation of weights** — $\mathrm{CV}_{\tilde{w}}$
   measures the variability of the retained models and detects extremely skewed model weights.

4.6.1 Effective sample size (ESS)

Since the BFRS draws are independent of $P(\Theta)$, ESS plays the role that $\hat{R}$ does in MCMC. We employ the specification of the ESS in Elvira *et al.* 2022 using the ΔAIC-truncated model weights $\tilde{w}_i$ from Equation (4.11):

$$\widehat{\mathrm{ESS}} \;=\; \Big[\sum_{i=1}^{n_{\mathrm{sim}}} \tilde{w}_i^2\Big]^{-1}. \tag{4.13}$$

Because discarded model runs have $\tilde{w}_i = 0$, ESS reflects only the retained subset. In dynamic Bayesian models an ESS $\gtrsim$500–1000 is generally adequate for stable posterior medians and 95 % credible intervals (Gelman *et al.* 2014; Bürkner 2017).

4.6.2 Agreement index A

We quantify consensus among the retained subset of best models $\mathcal{B} = \{\, i : \Delta_i \leq 4\}$ by the normalized Shannon entropy of their model weights $\tilde{w}_i$:

$$A = \frac{H(\tilde{\mathbf{w}})}{\log|\mathcal{B}|} \qquad \text{and} \qquad H(\tilde{\mathbf{w}}) = -\sum_{i\in\mathcal{B}} \tilde{w}_i \,\log \tilde{w}_i. \tag{4.14}$$

Note that by definition $A \in \{0, 1\}$ because $0 \;\leq\; H(\tilde{\mathbf{w}}) \;\leq\; \log|\mathcal{B}|$. When $A = 1$, all weights are equal ($\tilde{w}_i = 1/|\mathcal{B}|$) and there is a high degree of consensus in the sense that all models share similar fit. When $A = 0$, a single model dominates ($\tilde{w}_i = 1$) for one model run i and zero for the rest), which gives poor consensus among the subset of best models $\mathcal{B}$.

4.6.3 Coefficient of variation of weights

Beyond the first-moment metric of model agreement, we use a second-moment check that detects extremely skewed model weights:

$$\mathrm{CV}_{\tilde{\mathbf{w}}} = \frac{\sqrt{\sum_{i\in\mathcal{B}}(\tilde{w}_i - \bar{w})^2}}{\bar{w}} \qquad \text{and} \qquad \bar{w} = \tfrac{1}{|\mathcal{B}|}. \tag{4.15}$$

Large $\mathrm{CV}_{\tilde{\mathbf{w}}}$ warns that the variance within the best fitting subset of models may be high even if ESS and A look satisfactory. A common rule-of-thumb from Kong et al. 1994 is $\mathrm{CV}_{\tilde{\mathbf{w}}} \lesssim 2$ for comfortably balanced weights.

Table 4.2: Detials on convergence diagnostics with recommended thresholds and troubleshooting guidelines.

Metric	Target range
Effective Sample Size $\left(\widehat{\text{ESS}}\right)$	$\gt 500$ $\gt 1000$
Agreement Index $\left(A\right)$	$\gt 0.7 \ \text
Weight Coefficient of Variation $\left(\mathrm{CV}_{\tilde{\mathbf w}} \right)$	$\lt 1 \ \text{c

4.6.4 Practical convergence guidelines

A calibration run that meets all three criteria indicates that the retained ensemble is *informative* (high ESS), *internally consistent* (high A), and *numerically stable* (moderate $\tilde{\mathbf{w}}$), providing confidence in the posterior summaries and subsequent MOSAIC forecasts. We also perform Posterior Predictive Checks (PPC) on calibration runs to assess model fit and to refine the definition of model priors. Out targets for successful model calibration are shown in Table 4.2 with an example log-likelihood curve for a calibration test showing the diminishing returns in model fit with the number of simulations (Figure 4.1).

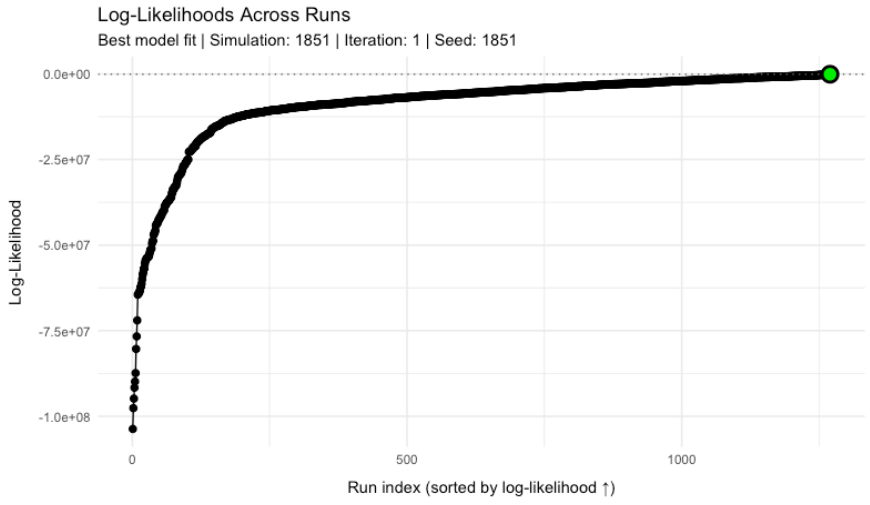

Figure 4.1: Example log-Likelihood curve for a sample of 2000 model simulations. Log-likelihood values are sorted from minumum to maximum with the model simulation giving the maximum likelihood highlighted in green.

4.7 Model Forecasting

This section is in development.

4.8 Scenarios and Counter Factuals

This section is in development.

Chapter 5

Scenarios

A key aim of the MOSAIC model is to provide near-term forecasts of cholera transmission in Sub-Saharan Africa (SSA) using the most current data available. However, MOSAIC is not just a forecasting tool; it is a dynamic model designed to explore various scenarios that influence critical factors such as vaccination, environmental conditions, and Water, Sanitation, and Hygiene (WASH) interventions.

5.1 Vaccination

5.1.1 Spatial and Temporal Strategies

Understanding the spatial and temporal distribution of cholera vaccination efforts is crucial for effective outbreak control. Key resources include:

- **Stockpile Status**: The availability of the oral cholera vaccine in emergency stockpiles can be tracked through UNICEF's Emergency Stockpile Availability.
- **WHO OCV Dashboard**: This dashboard (link) provides insights into the deployment of oral cholera vaccines (OCV) across different regions.

5.1.2 Reactive Vaccination

The timing and logistics of reactive vaccination campaigns are critical for controlling ongoing outbreaks. Relevant resources include:

- **WHO Recommended Timing**: Guidelines and recommendations for the timing of reactive OCV campaigns are available from the WHO (link).
- **Requests and Delay Time Distributions**: Information on vaccine request processes and the distribution of delays in vaccine deployment can be accessed through the GTFCC OCV Dashboard (link).

5.2 Impacts of Climate Change

5.2.1 Severe Weather Events

Projections of climate shocks, including the frequency and severity of cyclones and floods, are essential for modeling the future impacts of climate change on cholera transmission. Key references include:

- **Chen and Chavas 2020**: A study on cyclone season dynamics under climate change scenarios (link).
- **Sparks and Toumi 2024**: Research on projected flood frequencies due to climate change (link).
- **Switzer et al. 2023**: An analysis of climate shock impacts on cholera outbreaks (link).

5.2.2 Long-Term Trends

Long-term trends in weather variables under various climate change scenarios can be explored using the following resource:

- **Weather Variables Under Climate Change**: The OpenMeteo Climate API provides access to projected weather data under different climate change scenarios (link).

Chapter 6

Usage

The open-source code used to run MOSAIC is currently under development and will be presented here in the future.

Chapter 7

News

November 25, 2024 — The MOSAIC framework presented at ASMTH 2024

John Giles presented the MOSAIC modeling framework in a talk entitled "*Cholera modeling capacity at IDM: leveraging diverse data streams for scenarios and forecasting*" at the American Society of Tropical Medicine and Hygiene (ASTMH) on November 14, 2024 as part the symposium entitled "*Infectious Disease Surveillance and Modeling in LMIC's: From Data Collection to Forecasting*".

Chapter 8

References