MOSAIC: a spatial model of endemic cholera

John R Giles

Contents

Website under development. Last compiled on 2024-11-25 at 12:00 PM PST.

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

Introduction

1.1 History

Vibrio cholerae was introduced to the African continent from Asia in the 1970s and has since become endemic in many countries.

1.2 Recent Surge

Although there have been sporadic cholera outbreaks over the past five decades, there has been a significant surge in cases since 2021. This increase is likely due to factors such as climate change and disruptions to municipal services during the COVID-19 pandemic.

1.3 GTFCC Goals

The Global Task Force on Cholera Control (GTFCC) aims to reduce cholera deaths by 90% by 2030.

1.4 OCV Stockpiles

A major concern with the recent surge in cases is the depletion of oral cholera vaccine (OCV) stockpiles. In response, officials have shifted to single-dose OCV strategies. Efforts are underway to replenish stockpiles, with a key question being how best to allocate them to reduce transmission regionally and support the GTFCC's goal.

1.5 Climate Change

Environmental factors play a significant role in cholera outbreaks, with warmer and wetter conditions creating a more favorable environment for Vibrio cholerae. These conditions are likely to exacerbate the already challenging endemic and outbreak settings. Models that incorporate climatic forcing can provide insights into future cholera dynamics due to climate change and aid in achieving the GTFCC goal.

Chapter 2

Rationale

A significant challenge in controlling cholera transmission in Sub-Saharan Africa (SSA) is the lack of comprehensive datasets 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 is to iteratively create a landscape-scale transmission model for cholera in SSA that can provide weekly predictions of key epidemiological metrics. Our modeling methods will leverage a wide array of up-to-date data sources, including incidence and mortality reports, patterns of human movement, vaccination history and schedules, and environmental factors.

Key questions 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. A landscape-scale model that accounts for endemic transmission patterns will be a valuable tool in addressing these questions.

Chapter 3

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.

3.1 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.

3.2 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.

3.3 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.

3.4 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.

3.5 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.

3.5.1 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.

3.6 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.

3.7 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 4

Model description

Here we describe the methods of MOSAIC beta version 0.1. 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 weekly time steps from January 2023 to August 2024 and includes 41 countries in Sub-Saharan Africa (SSA), see Figure 4.1.

4.1 Transmission dynamics

The model has a metapopulation structure with familiar compartments for Susceptible, Infected, and Recovered individuals with SIRS dynamics. The model also contains compartments for vaccinated individuals (V) and Water & environment based transmission (W) which we refer to as SVIWRS.

The SVIWRS metapopulation model, shown in Figure 4.2, is governed by the following difference equations:

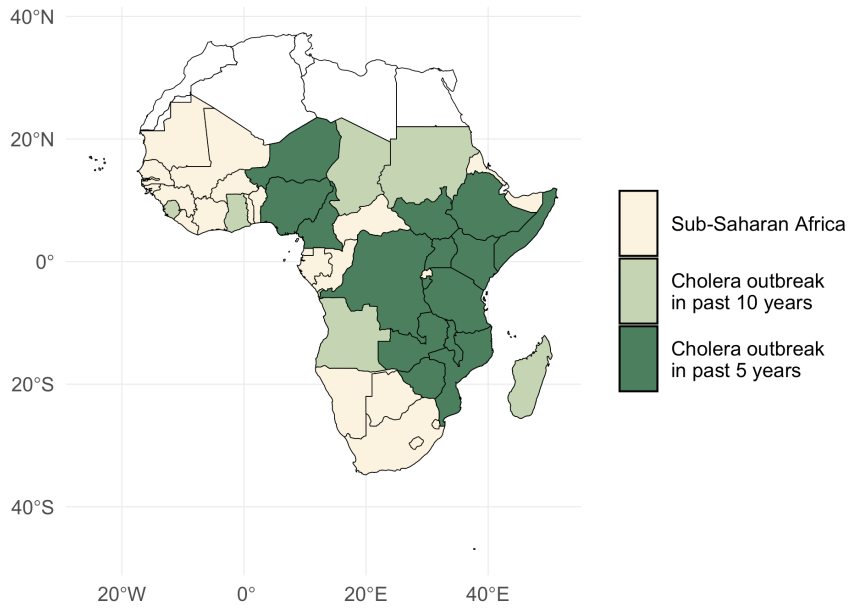

Figure 4.1: A map of Sub-Saharan Africa with countries that have experienced a cholera outbreak in the past 5 and 10 years highlighted in green.

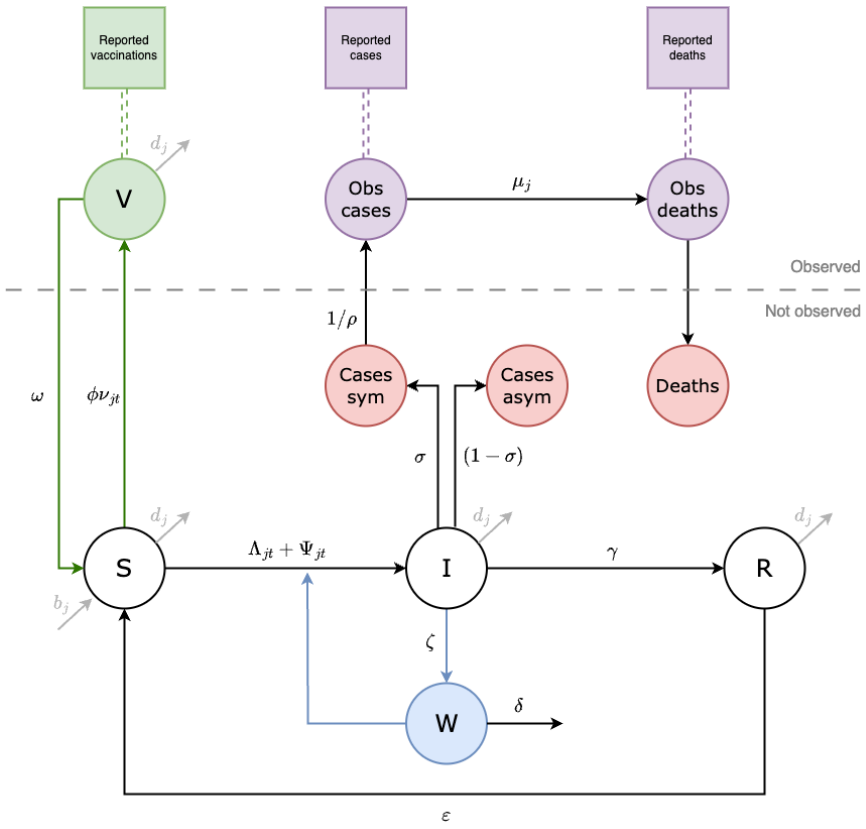

Figure 4.2: This diagram of the SVIWRS (Susceptible-Vaccinated-Infected-Water/environmental-Recovered-Susceptible) model shows model compartments as circles with rate parameters displayed. The primary data sources the model is fit to are shown as square nodes (Vaccination data, and reported cases and deaths).

$$
\begin{aligned}
S_{j,t+1} &= b_j N_{jt} + S_{jt} - \phi\nu_{jt} S_{jt} + \omega V_{jt} - \Lambda_{j,t+1} - \Psi_{j,t+1} + \varepsilon R_{jt} - d_j S_{jt} \\
V_{j,t+1} &= V_{jt} + \phi\nu_{jt} S_{jt} - \omega V_{jt} - d_j V_{jt} \\
I_{j,t+1} &= I_{jt} + \Lambda_{j,t+1} + \Psi_{j,t+1} - \gamma I_{jt} - \mu\sigma I_{jt} - d_j I_{jt} \\
W_{j,t+1} &= W_{jt} + \zeta I_{jt} - \delta_{jt} W_{jt} \\
R_{j,t+1} &= R_{jt} + \gamma I_{jt} - \varepsilon R_{jt} - d_j R_{jt}
\end{aligned}
\tag{4.1}
$$

For descriptions of all parameters in Equation (4.1), see Table (4.10). Transmission dynamics are driven by the two force of infection terms, Λ_{jt} and Ψ_{jt}. The force of infection due to human-to-human (Λ_{jt}) is:

$$
\Lambda_{j,t+1} = \frac{\beta_{jt}^{\text{hum}} \Big((S_{jt}(1-\tau_j))(I_{jt}(1-\tau_j) + \sum_{\forall i \neq j} (\pi_{ij}\tau_i I_{it})) \Big)^{\alpha}}{N_{jt}}. \tag{4.2}
$$

Where β_{jt}^{hum} is the rate of human-to-human transmission. Movement within and among metapopulations is governed by τ_i, the probability of departing origin location i, and π_{ij} is the relative probability of travel from origin i to destination j (see section on spatial dynamics). To include environmental effects, the force of infection due to environment-to-human transmission (Ψ_{jt}) is defined as:

$$
\Psi_{j,t+1} = \frac{\beta_{jt}^{\text{env}} (S_{jt}(1-\tau_j))(1-\theta_j) W_{jt}}{\kappa + W_{jt}}, \tag{4.3}
$$

where β_{jt}^{env} is the rate of environment-to-human transmission and θ_j is the proportion of the population at location j that at least basic access to Water, Sanitation, and Hygiene (WASH). The environmental compartment of the model is also scaled by the concentration (cells per mL) of *V. cholerae* that is required for a 50% probability of infection Fung 2014. See the section on environmental transmission below for more on the water/environment compartment and climatic drivers of transmission.

Note that all model processes are stochastic. Transition rates are converted to probabilities with the commonly used formula $p(t) = 1 - e^{-rt}$ (see Ross 2007), and then integer quantities are moved between model compartments at each time step according to a binomial process like the example below for the recovery of infected individuals (γI_{jt}):

$$\frac{\partial R}{\partial t} \sim \text{Binom}(I_{jt}, 1 - \exp(-\gamma)) \tag{4.4}$$

4.2 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 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:

$$\beta^{\text{hum}}_{jt} = \beta^{\text{hum}}_{j0} + a_1 \cos\left(\frac{2\pi t}{p}\right) + b_1 \sin\left(\frac{2\pi t}{p}\right) + a_2 \cos\left(\frac{4\pi t}{p}\right) + b_2 \sin\left(\frac{4\pi t}{p}\right) \tag{4.5}$$

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 52, so the function controls the temporal variation in β^{hum}_{jt} over the 52 weeks of the year.

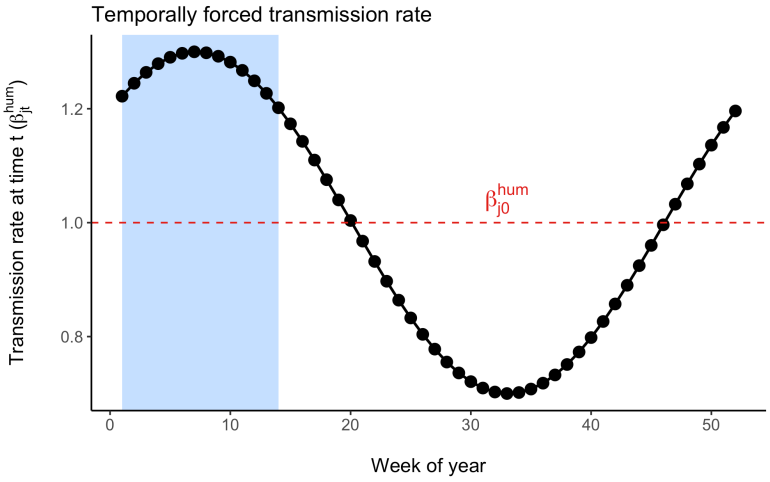

Figure 4.3: An example of the temporal distribution of the human-to-human transmission rate across each of the 52 weeks of the year given by the cosine wave function. The wave function is fitted to each country and is designed to align with the rainy season as indicated by the shaded region in this figure.

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.

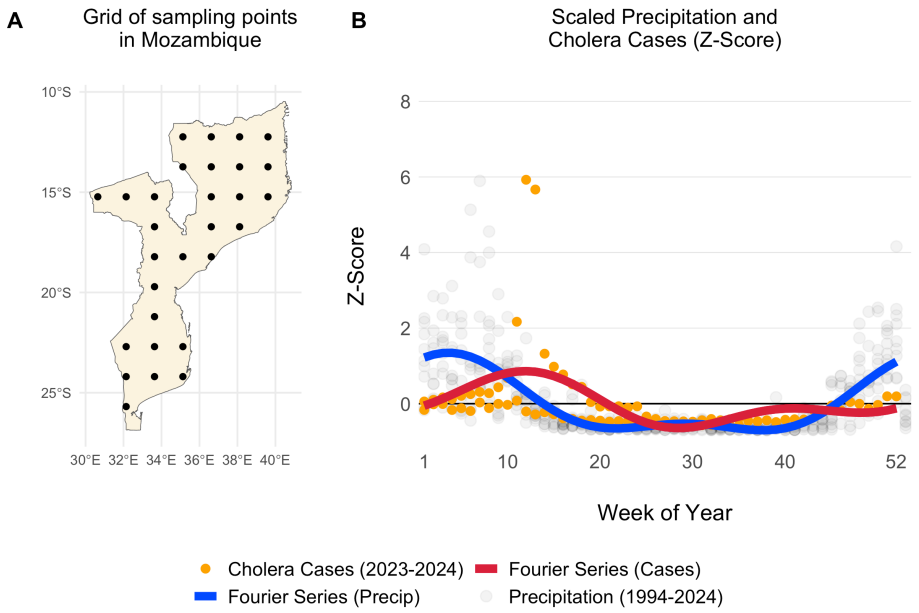

Figure 4.4: Example of a grid of 30 uniformly distributed points within Mozambique (A). The scatterplot shows weekly summed precipitation values at those 30 grid points and cholera cases plotted on the same scale of the Z-Score which shows the variance around the mean in terms of the standard deviation. Fitted Fourier series fucntions are shown as blue (fit precipitation data) and red (fit to cholera case data) lines.

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 4.5) 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.

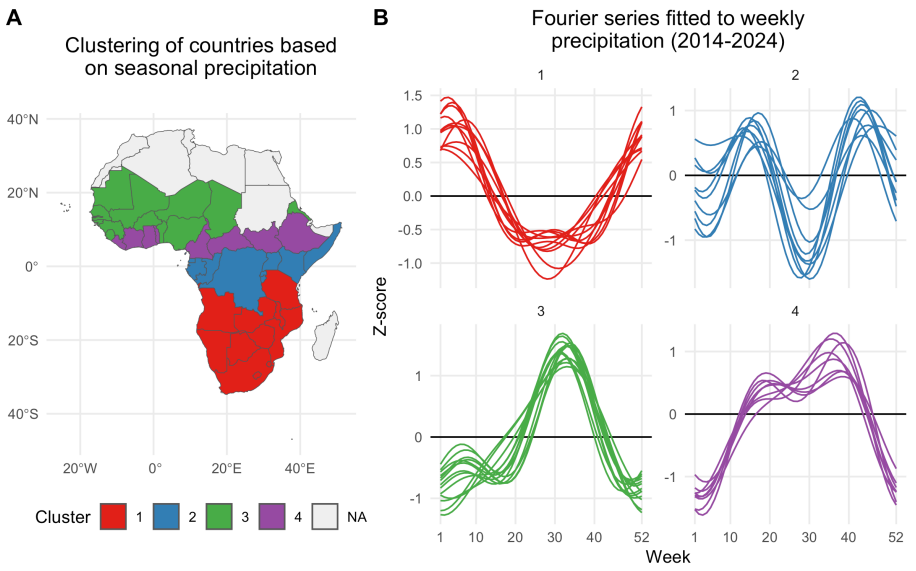

Figure 4.5: A) Map showing the clustering of African countries based on their seasonal precipitation patterns (2014-2024). Countries are colored according to their cluster assignments, identified using hierarchical clustering. B) Fourier series fitted to weekly precipitation for each country. Each line 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.

Using the model fitting methods described above, and the cluster-based approach for inferring the seasonal Fourier series pattern in countries without reported cholera cases, we modeled the seasonal dynamics for all 41 countries in the MOSAIC framework. These dynamics are visualized in Figure 4.6, with the corresponding Fourier model coefficients presented in Table 4.1.

Table 4.1: 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	-0.31 (-0.35 to -0.27)	-0.49 (-0.53 to -0.45)	-0.42 (-0.45 to -0.39)	-0.36
Cameroon	-0.64 (-0.67 to -0.62)	0.08 (0.04 to 0.11)	0.01 (-0.04 to 0.06)	-0.28
DRC	0.28 (0.26 to 0.31)	-0.22 (-0.31 to -0.13)	0.15 (0.08 to 0.23)	-0.2
Ethiopia	-0.52 (-0.59 to -0.46)	-0.38 (-0.41 to -0.34)	-0.04 (-0.12 to 0.03)	-0.03
Kenya	0.03 (-0.02 to 0.07)	-0.16 (-0.19 to -0.12)	0.5 (0.46 to 0.53)	0.12
Malawi	0.51 (0.47 to 0.54)	0.16 (0.12 to 0.2)	0.39 (0.36 to 0.43)	0.4 (
Mozambique	0.22 (0.19 to 0.26)	-0.36 (-0.39 to -0.33)	0.48 (0.44 to 0.53)	0.02
Nigeria	-0.27 (-0.3 to -0.24)	0.22 (0.15 to 0.29)	-0.16 (-0.2 to -0.11)	0.42
Somalia	-0.16 (-0.2 to -0.12)	-0.24 (-0.3 to -0.17)	0.84 (0.82 to 0.86)	-0.59
South Africa	-0.41 (-0.62 to -0.19)	0.13 (-0.07 to 0.33)	-0.54 (-0.67 to -0.41)	0.63
Tanzania	0.03 (0 to 0.06)	-0.06 (-0.12 to -0.01)	0.09 (0.06 to 0.12)	0.23
Togo	-0.17 (-0.45 to 0.12)	-0.28 (-0.43 to -0.13)	-0.63 (-0.72 to -0.54)	0.28
Uganda	-0.79 (-0.9 to -0.69)	0.39 (0.32 to 0.46)	0.12 (-0.07 to 0.31)	0.04
Zambia	0.69 (0.63 to 0.74)	0.43 (0.39 to 0.46)	0.23 (0.18 to 0.28)	0.2 (
Zimbabwe	1.12 (1.08 to 1.17)	0.51 (0.48 to 0.53)	0.3 (0.24 to 0.35)	-0.08

4.3 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.

4.3.1 Climate-driven transmission

To capture the impacts of climate-drivers on cholera transmission, we have included the parameter ψ_{jt}, which represents the current state of environmental

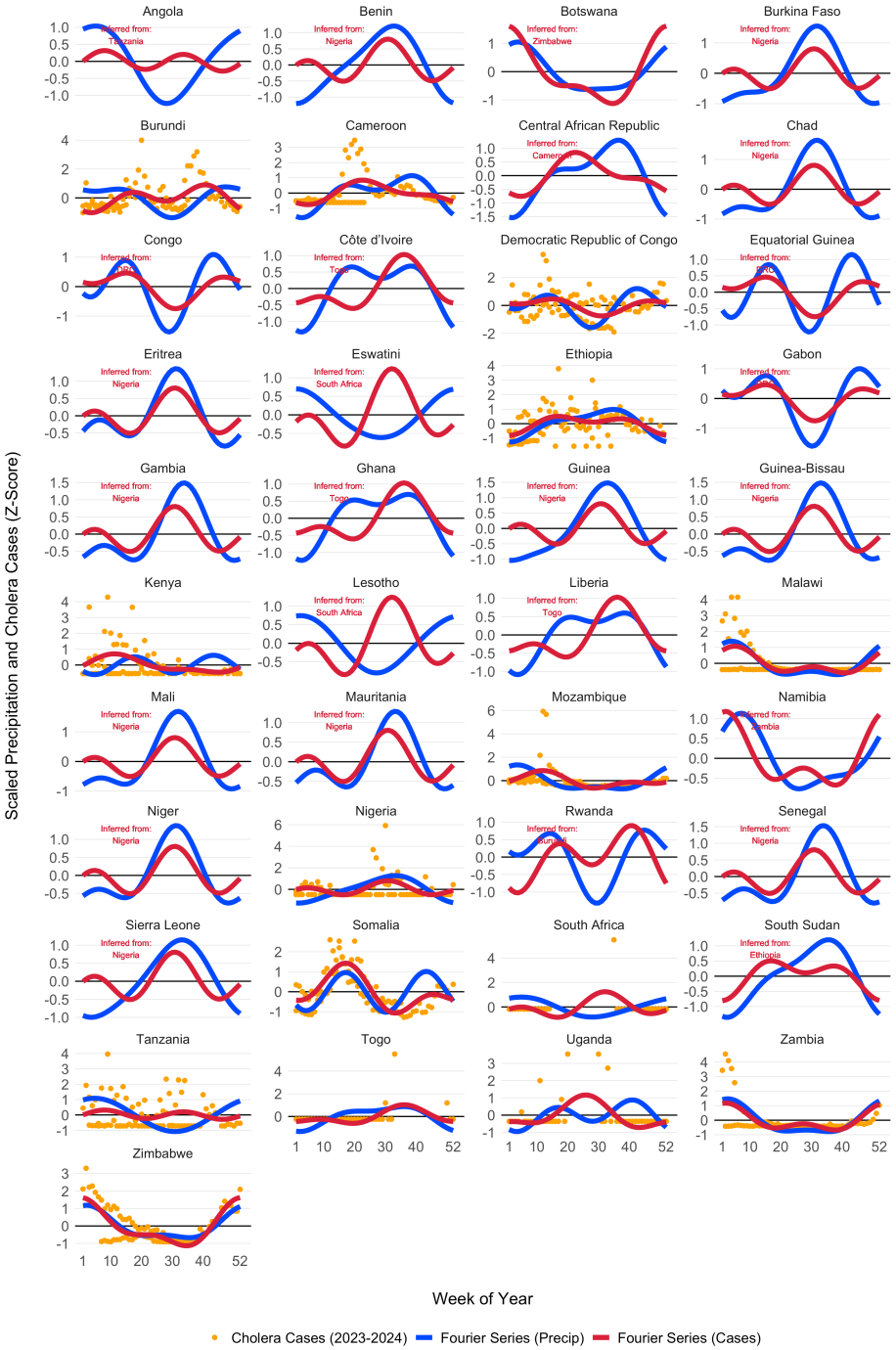

Figure 4.6: Seasonal transmission patterns for all countries modeled in MOSAIC as modeled by the truncated Fourier series in Equation eqrefeq:beta1. Blues lines give the Fourier series model fits for precipitation (1994-2024) and the red lines give models fits to reported cholera cases (2023-2024). For 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 weekly precipitation and cholera cases.

suitability with respect to: *i*) the survival time of *V. cholerae* in the environment and, *ii*) the rate of environment-to-human transmission which contributes to the overall force of infection.

$$\beta_{jt}^{\text{env}} = \beta_{j0}^{\text{env}} \left(1 + \frac{\psi_{jt} - \bar{\psi}_j}{\bar{\psi}_j} \right) \quad \text{and} \quad \bar{\psi}_j = \frac{1}{T} \sum_{t=1}^{T} \psi_{jt} \tag{4.6}$$

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.

Environmental suitability (ψ_{jt}) also impacts the survival rate of *V. cholerae* in the environment (δ_{jt}) with the form:

$$\delta_{jt} = \delta_{\min} + \psi_{jt} \times (\delta_{\max} - \delta_{\min}) \tag{4.7}$$

which normalizes the variance of the suitability parameter to be bounded within the minimum ($\delta_{\min}$) and maximum ($\delta_{\max}$) survival times of *V. cholerae*.

4.3.2 Modeling environmental suitability

4.3.2.1 Environmental data

The mechanism for environment-to-human transmission (Equation (4.6)) and rate of decay of *V. cholerae* in the environment (Equation (4.7)) 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. See example time series of climate variables from one country (Mozambique) in Figure 4.8 and DMI and ENSO variables in Figure 4.9. A list of all covariates and their sources can be seen in Table 4.2.

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.

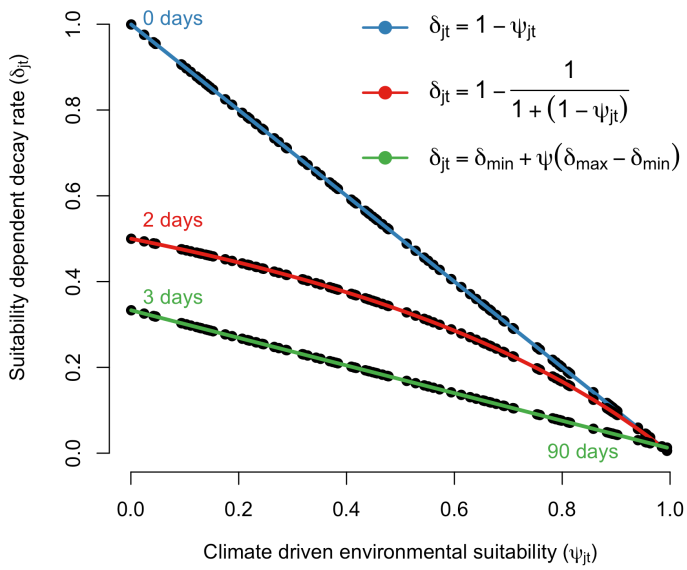

Figure 4.7: Relationship between environmental suitability (ψ_{jt}) and the rate of *V. cholerae* decay in the environment (δ_j). The green line shows the mildest penalty on *V. cholerae* survival, where survival in the environment is $1/\delta_{\min}$ = 3 days when suitability = 0 and $1/\delta_{\max}$ = 90 days when suitability = 1.

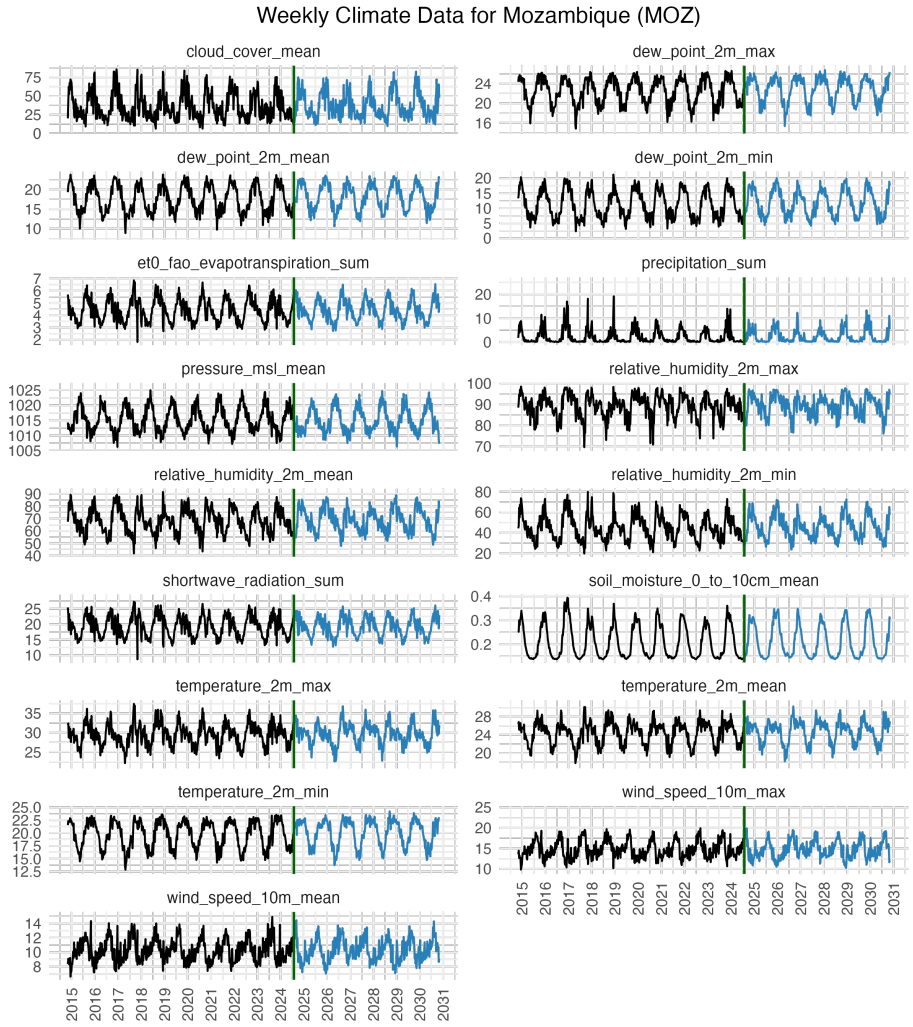

Figure 4.8: Climate data acquired from the OpenMeteo data API. Data were collected from 30 uniformly distributed points across each country and then aggregated to give weekly values of 17 climate variable from 1970 to 2030.

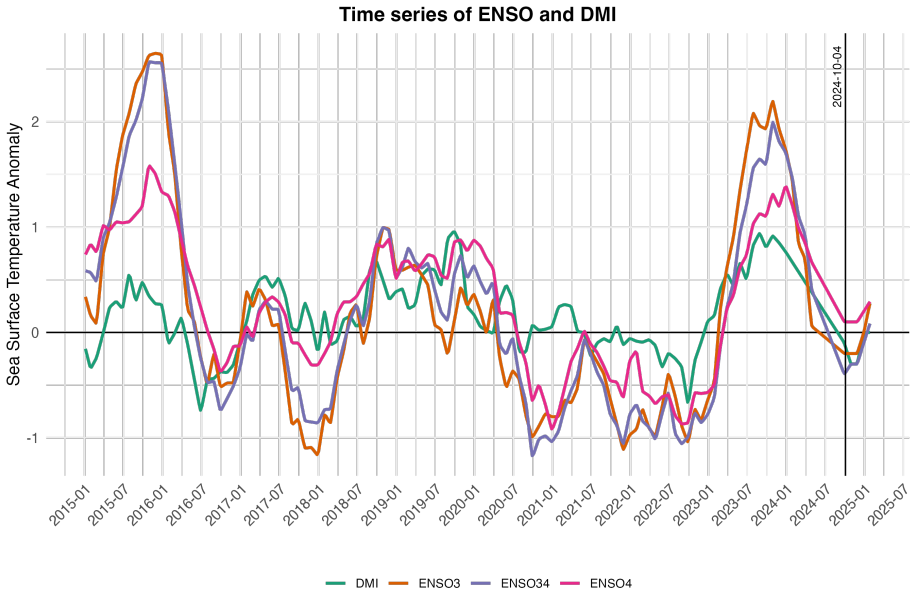

Figure 4.9: Historical and forecasted values of the Indian Ocean Dipole Mode Index (DMI) and the El Niño Southern Oscillation (ENSO) from 2015 to 2025. The ENSO values come from three different regions: Niño3 (central to eastern Pacific), Niño3.4 (central Pacific), and Niño4 (western-central Pacifi). Data are from National Oceanic and Atmospheric Administration (NOAA) and Bureau of Meteorology (BOM).

Table 4.2: 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

4.3.2.2 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 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 4.10 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:

$$\psi_{jt} \sim \text{Sigmoid}(w_h \cdot h_t + b_h) \tag{4.8}$$

$$h_t = \text{LSTM}(\text{temperature}_{jt},\ \text{precipitation}_{jt},\ \text{ENSO}_t, \dots) \tag{4.9}$$

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

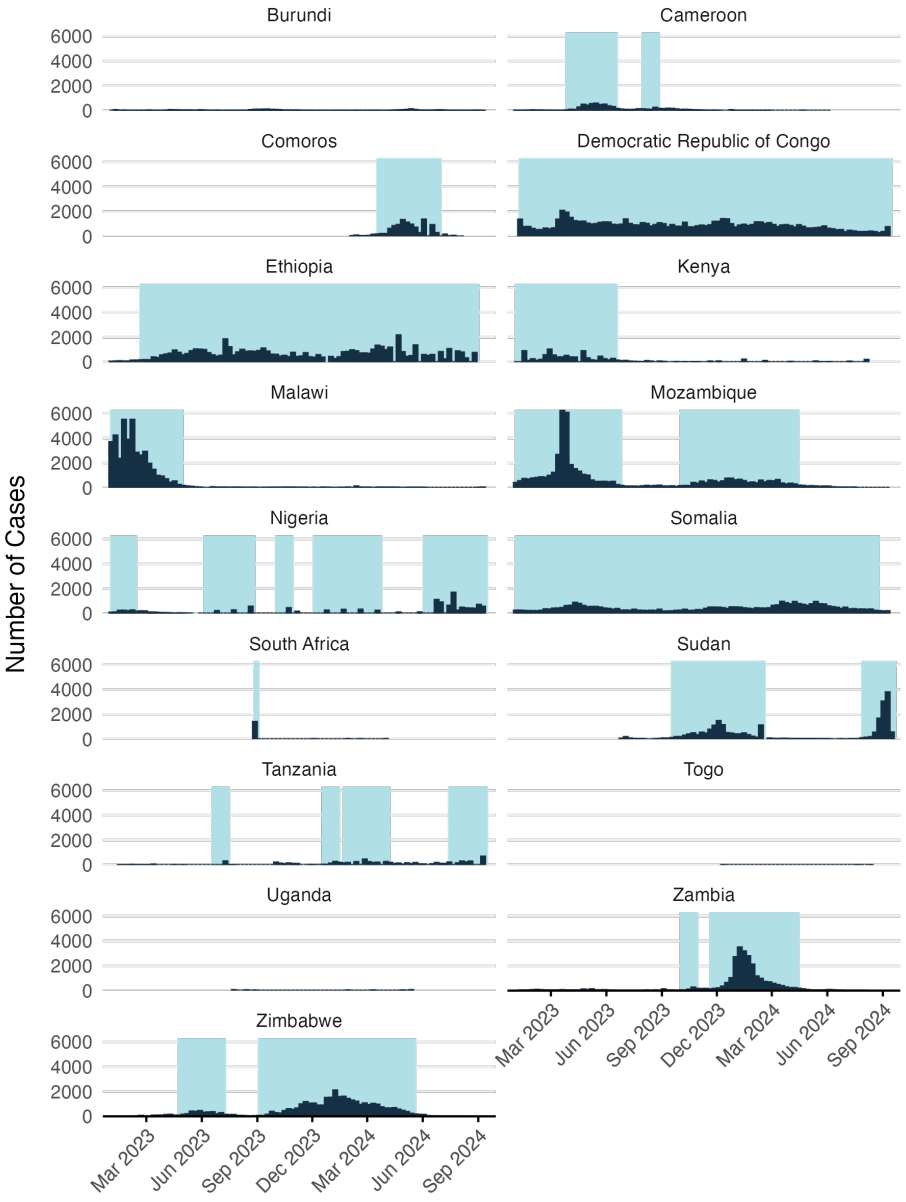

Figure 4.10: Reported cases converted to binary variable for modeling environmental suitability.

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 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 4.11).

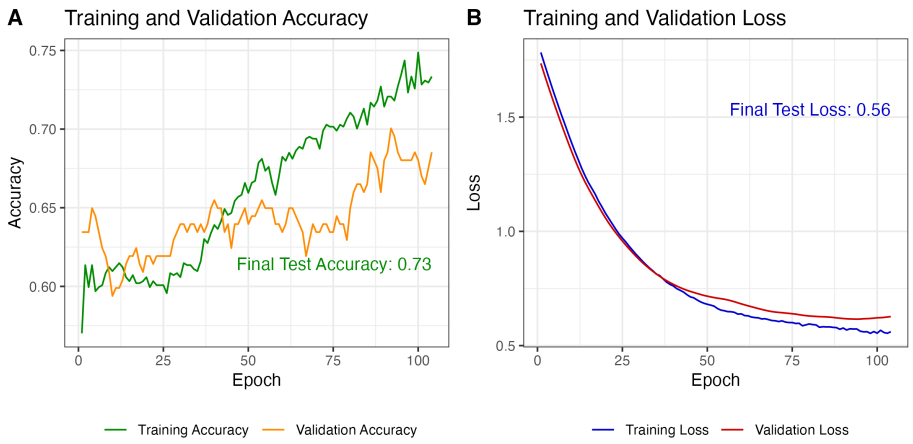

Figure 4.11: Model performance on training and validation data.

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 (4.6) and (4.7)). The resulting model predictions are shown for an example country such as Mozambique in Figure 4.12 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 4.13.

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.

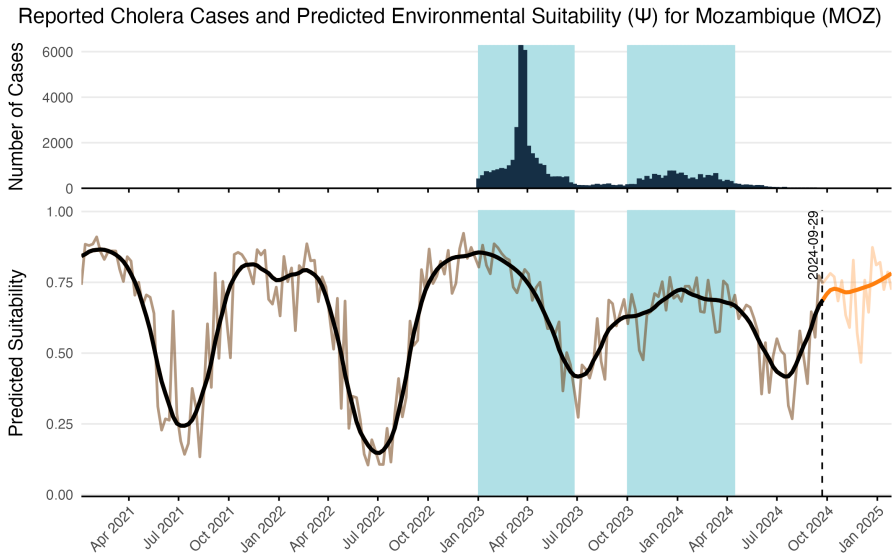

Figure 4.12: The LSTM model predictions over time and reported cases for an example country such as Mozambique. Reported cases are shown in the top panel and tje shaded areas show the binary classification used to characterize environmental suitability. Raw model predicitons are shown in the transparent 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.

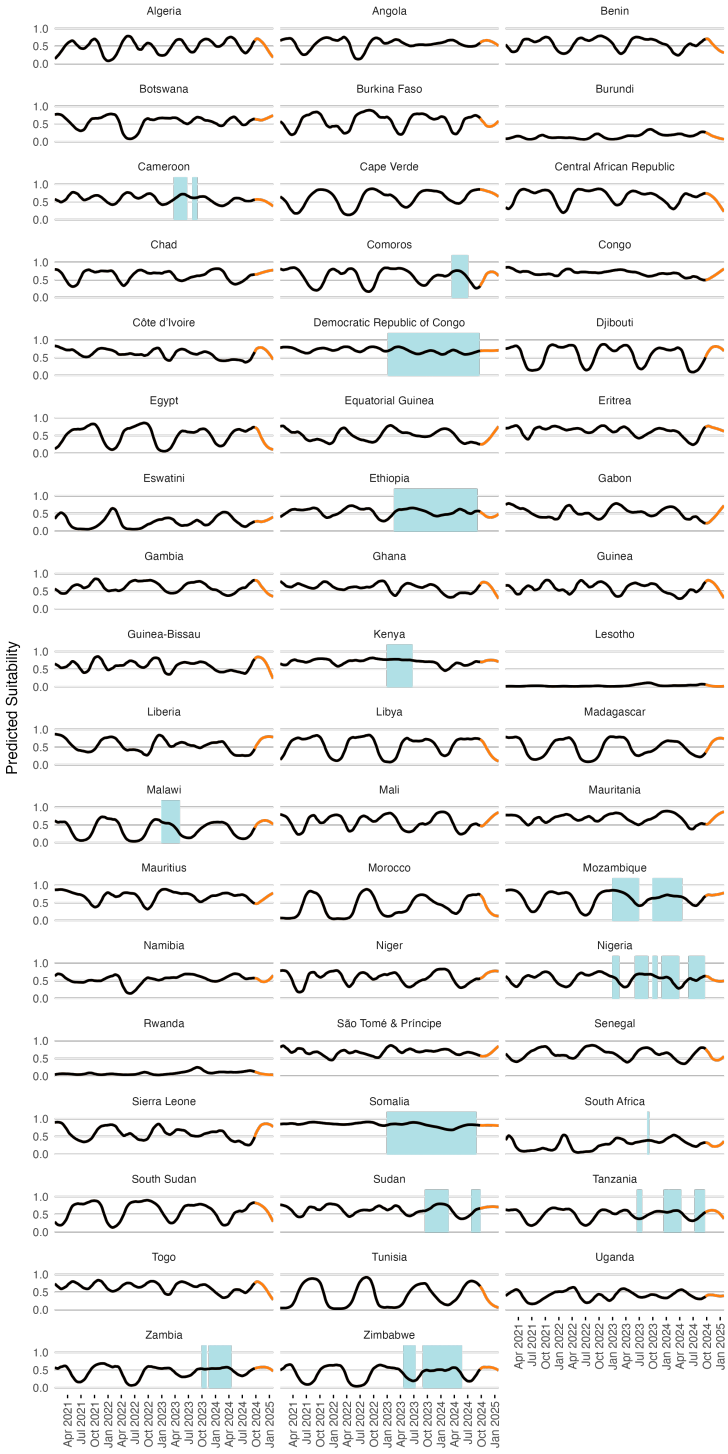

Figure 4.13: The smoothed LSTM model predictions (lines) and binary suitability classification (shaded areas) over time for all countries in the MOSAIC framework. Orange lines show forecasts beyond the current date. With ENSO and DMI covariates included in the model, forecasts are limited to 5 months.

4.3.3 Shedding

The rate at which infected individuals shed *V. cholerae* into the environment (ζ) is a critical factor influencing cholera transmission. Shedding rates can vary widely depending on the severity of the infection, the immune response of the individual, and environmental factors. According to Fung 2014, the shedding rate is estimated to range from 0.01 to 10 cells per mL per person per day.

Further studies support these findings, indicating that shedding rates can indeed fluctuate significantly. For instance, Nelson et al (2009) note that during the, depending on the phase of infection, individuals can shed 10^3 (asymptomatic cases) to 10^{12} (severe cases) *V. cholerae* cells per gram of stool. Future version of the model may attempt to capture the nuances of shedding dynamics, but here we make the simplifying assumption that shedding is constant across infected individuals and has a wide range of variability with no prior distributional assumptions:

$$\zeta \sim \text{Uniform}(0.01, 10).$$

4.3.4 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 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 4.3) 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 4.3) 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 4.14). The weighted mean then provides a single variable between 0

Table 4.3: 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

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 4.15.

4.4 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.

4.4.1 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

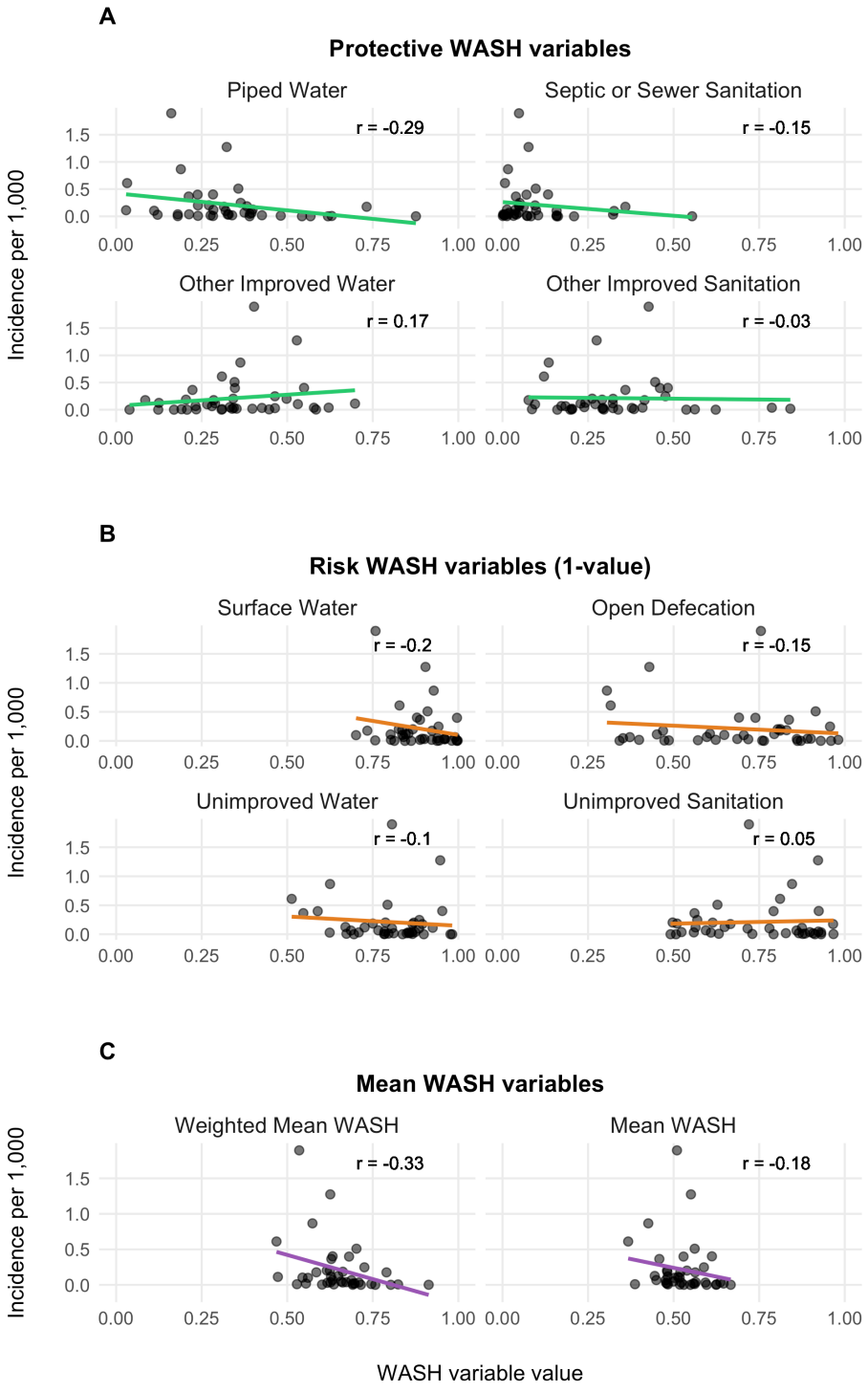

Figure 4.14: Relationship between WASH variables and cholera incidences.

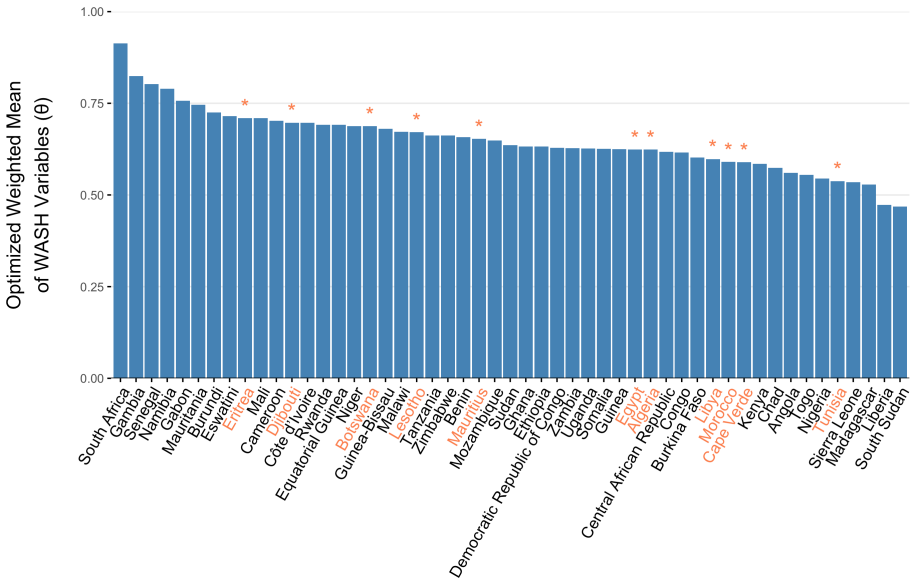

Figure 4.15: The optimized weighted mean of WASH variables for AFRO countries. Countries labeled in orange denote countries with an imputed weighted mean WASH variable. Imputed values are the weighted mean from the 3 most similar countries.

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 4.16 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 4.17, with current totals based on the WHO ICG data displayed in Figure 4.18.

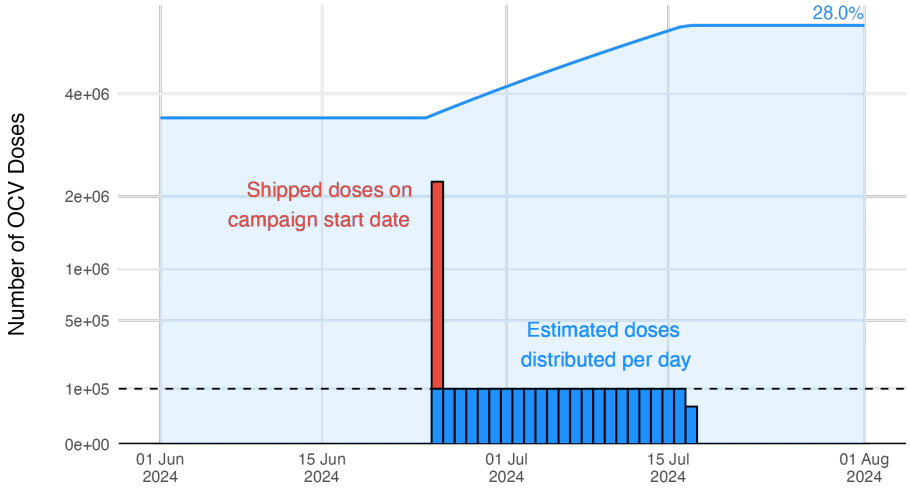

Figure 4.16: Example of the estimated vaccination rate during an OCV campaign.

4.4.2 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

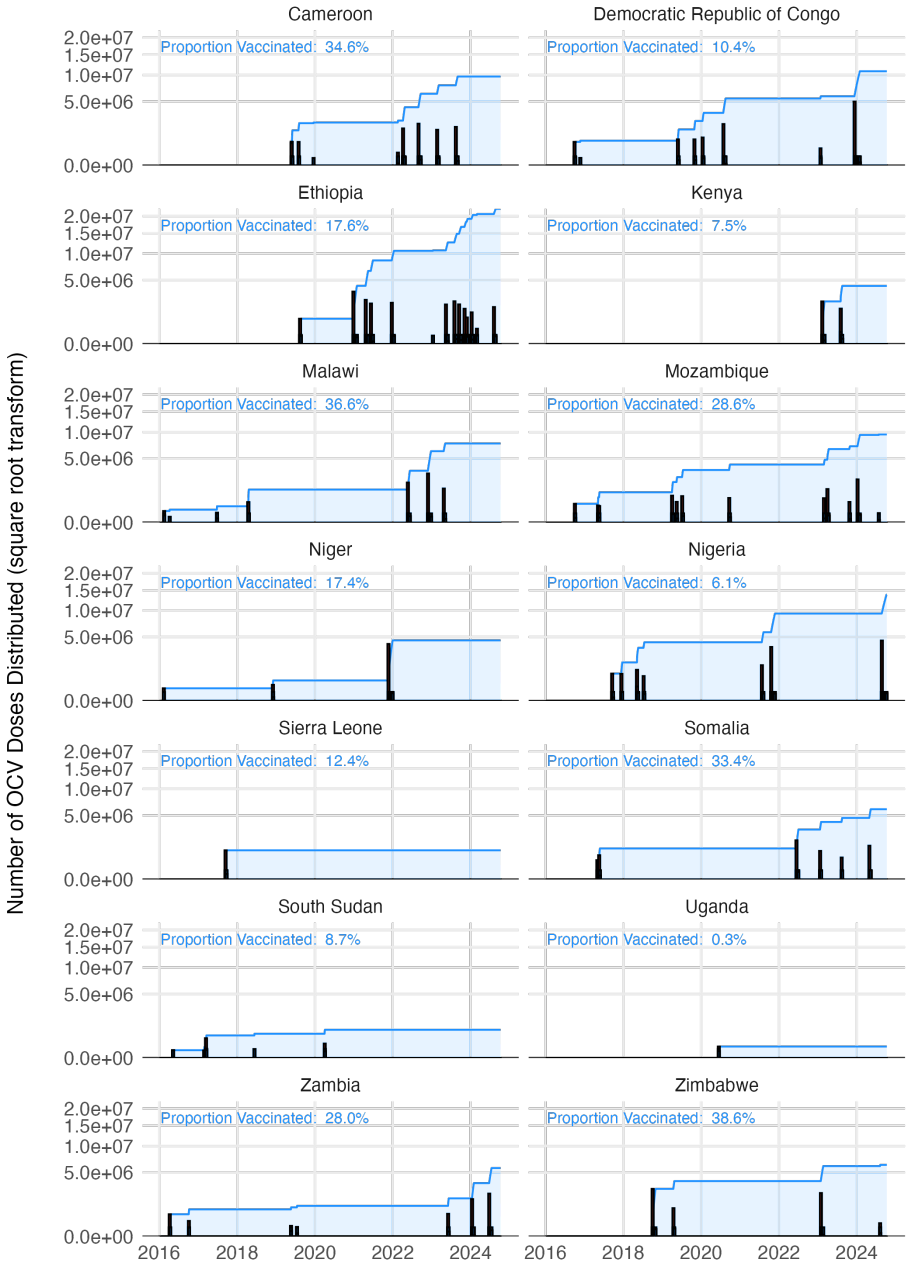

Figure 4.17: The estimated vaccination coverage across all countries with reported vaccination data one the WHO ICG dashboard.

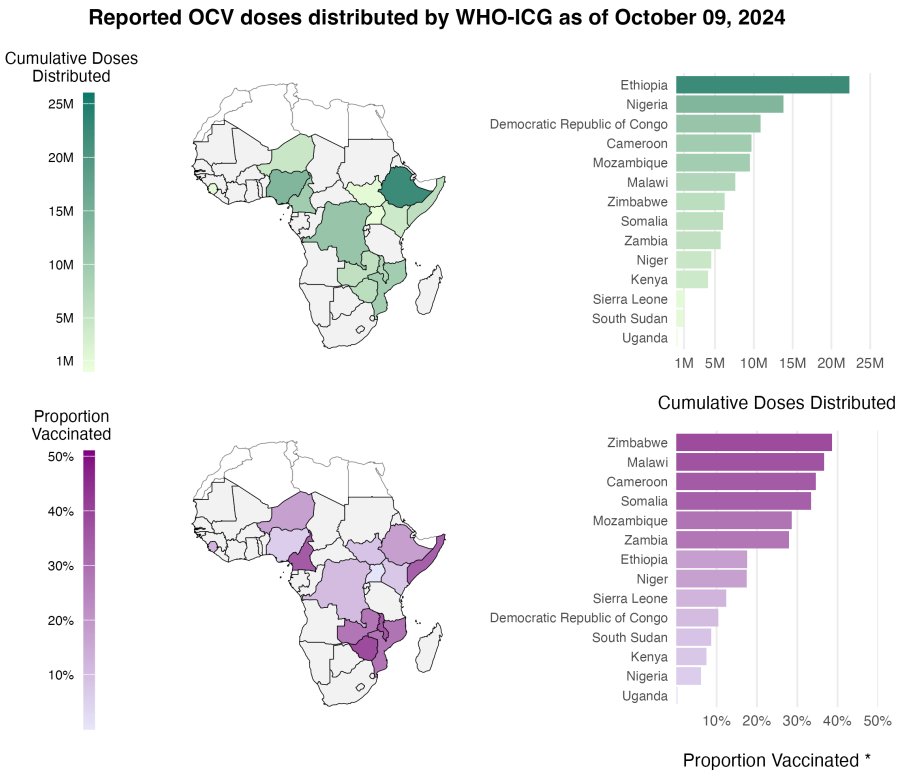

Figure 4.18: The total cumulative number of OCV doses distributed through the WHO ICG from 2016 to present day.

Table 4.4: 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)

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 4.4) from Bangladesh (Qadri et al 2016 and 2018), South Sudan (Azman et al 2016), and Democratic Republic of Congo (Malembaka et al 2024).

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:

$$\text{Proportion immune } t \text{ days after vaccination} = \phi \times (1-\omega)^{t-t_{\text{vaccination}}} \quad (4.10)$$

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 (4.4) 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 (4.10) 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 4.19B):

$$\phi \sim \text{Beta}(4.57, 2.41). \tag{4.11}$$

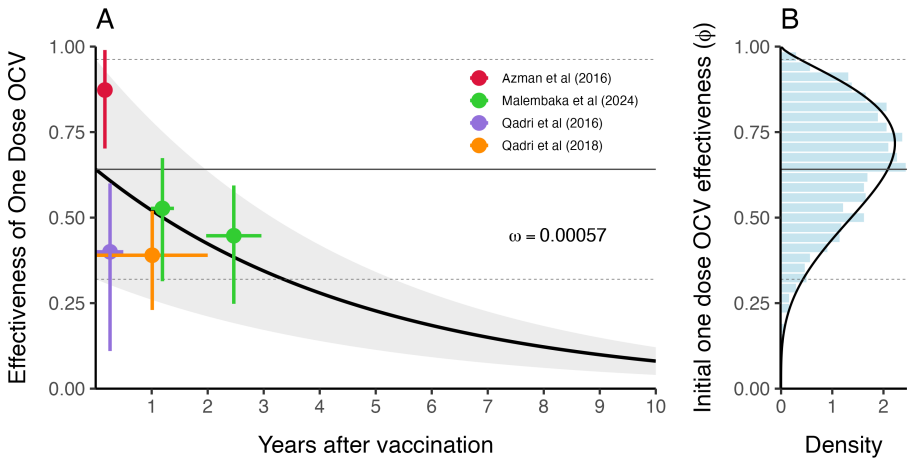

Figure 4.19: This is vaccine effectiveness

4.4.3 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 fit this model to reported data from Ali et al (2011) and Clemens et al (1991) (see Table 4.5).

We estimated the mean immune decay to be $\bar{\varepsilon} = 3.9 \times 10^{-4}$ (1.7×10^{-4} – 1.03×10^{-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 4.20A. This is slightly longer than

Table 4.5: 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

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 4.20B:

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

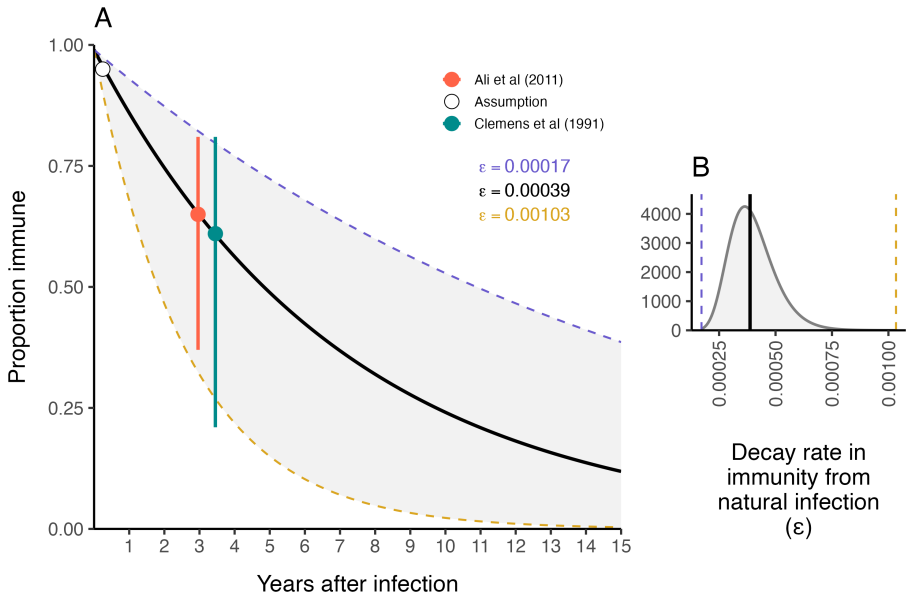

Figure 4.20: The duration of immunity after natural infection with *V. cholerae*.

4.5 Spatial dynamics

The parameters in the model diagram in Figure 4.2 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

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 4.21).

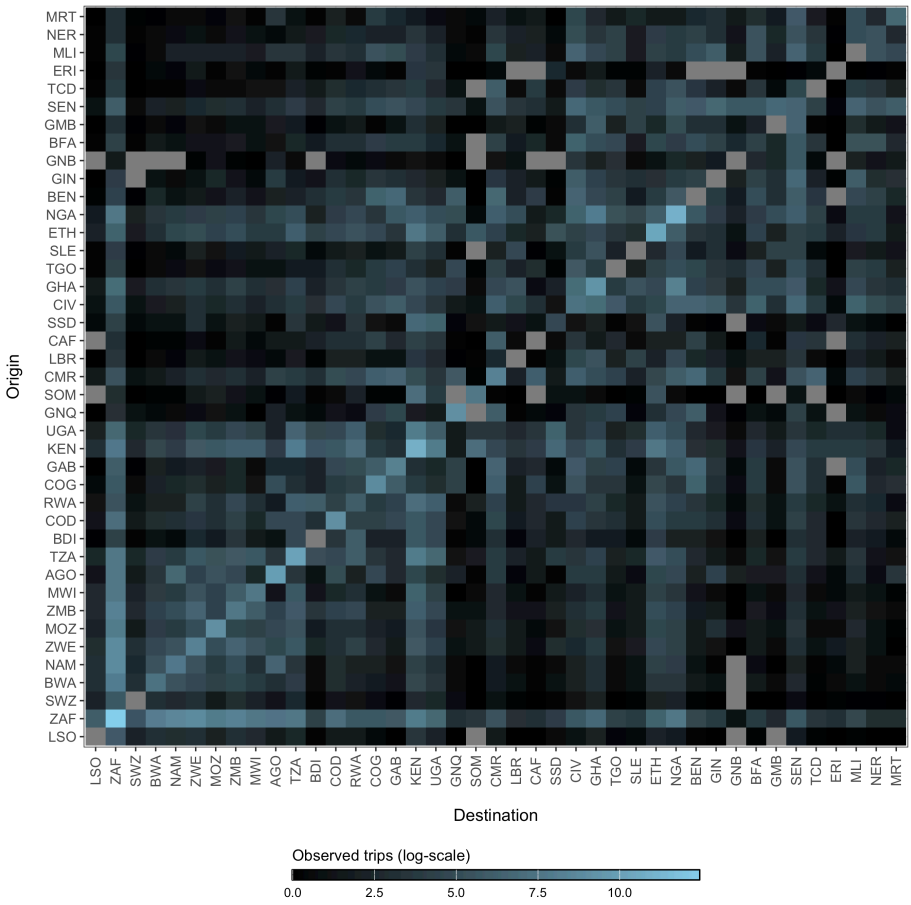

Figure 4.21: The average number of air passengers per week in 2017 among all countries.

4.5.1 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—

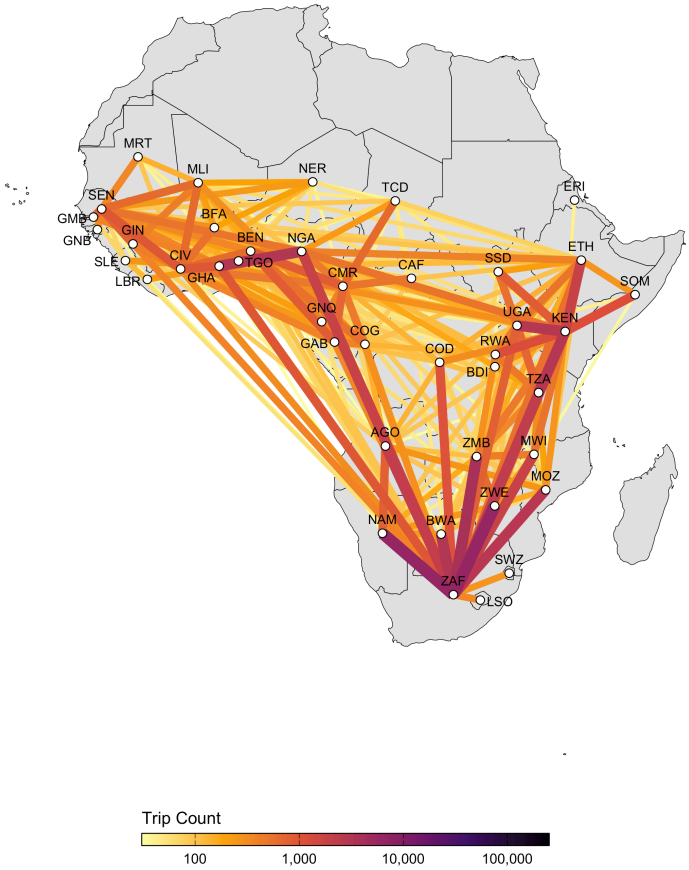

Figure 4.22: A network map showing the average number of air passengers per week in 2017.

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

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

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

$$M_{ij} = \begin{cases} \theta N_i (1 - \tau_i) & \text{if } i = j \\ \theta N_i \tau_i \pi_{ij} & \text{if } i \neq j. \end{cases} \tag{4.12}$$

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.

4.5.2 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.

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

4.5.3 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:

$$\pi_{ij} = \frac{N_j^{\omega} d_{ij}^{-\gamma}}{\sum\limits_{\forall j \neq i} N_j^{\omega} d_{ij}^{-\gamma}} \tag{4.13}$$

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 4.23 and 4.24.

4.5.4 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:

$$p(i, j, t) = 1 - e^{-\beta_{jt}^{\text{hum}}(((1-\tau_j)S_{jt})/N_{jt})\pi_{ij}\tau_i I_{it}} \tag{4.14}$$

4.5.5 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:

$$h(j, t) = \frac{\beta_{jt}^{\text{hum}}\Big(1 - \exp\big(-((1-\tau_j)S_{jt}/N_{jt})\sum_{\forall i \neq j}\pi_{ij}\tau_i(I_{it}/N_{it})\big)\Big)}{1/(1 + \beta_{jt}^{\text{hum}}(1-\tau_j)S_{jt})}. \tag{4.15}$$

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

$$w(j, t) = h(j, T)\prod_{t=1}^{T-1} 1 - h(j, t). \tag{4.16}$$

4.5.6 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

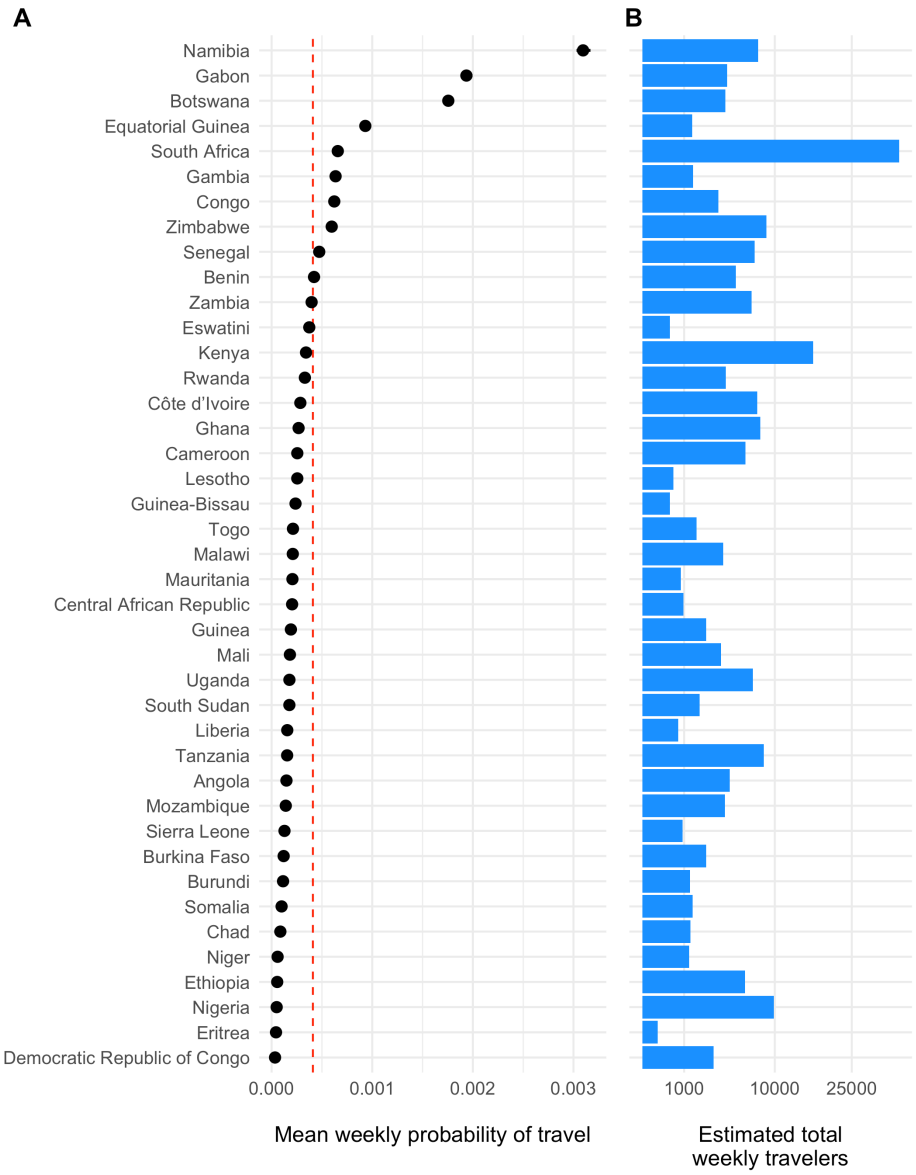

Figure 4.23: The estimated weekly probability of travel outside of each origin location τ_i and 95% confidence intervals is shown in panel A with the population mean indicated as a red dashed line. Panel B shows the estimated total number of travelers leaving origin i each week.

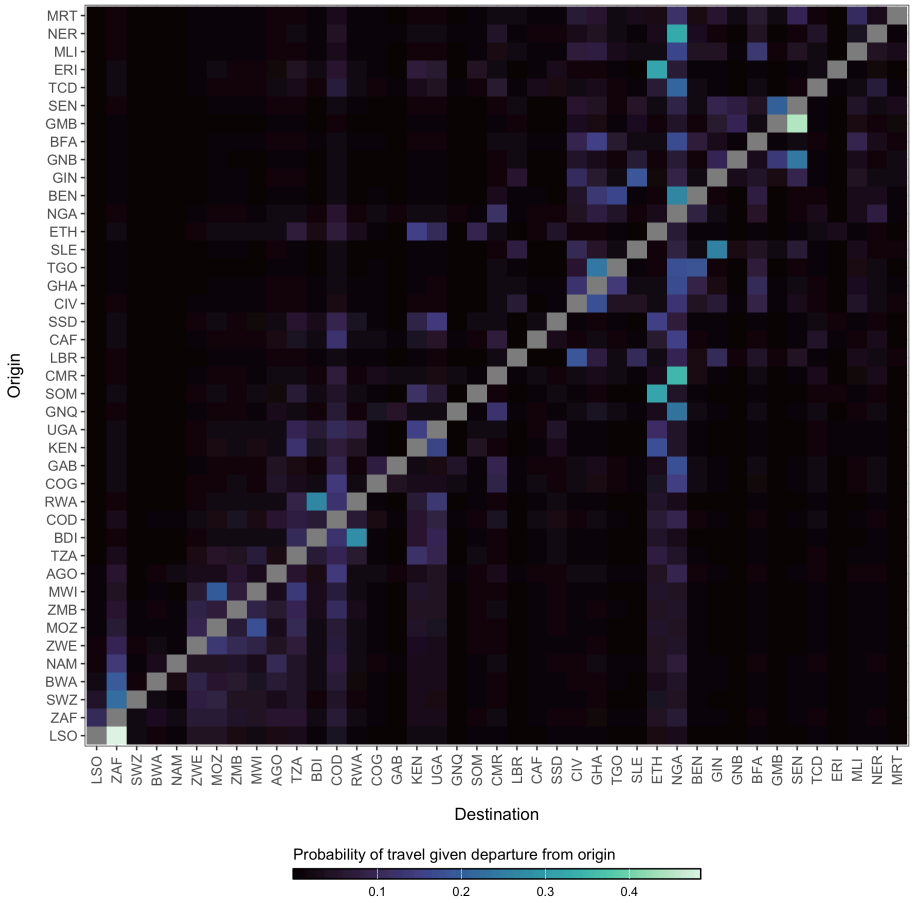

Figure 4.24: The diffusion process π_{ij} which gives the estimated probability of travel from origin i to destination j given that travel outside of origin i has occurred.

in Keeling and Rohani (2002), which gives a measure of how similar infection dynamics are between locations.

$$C_{ij} = \frac{(y_{it} - \bar{y}_i)(y_{jt} - \bar{y}_j)}{\sqrt{\text{var}(y_i)\text{var}(y_j)}} \tag{4.17}$$

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 $\bar{y_j} = \frac{1}{T}\sum_{t=1}^{T} y_{jt}$.

4.6 The observation process

4.6.1 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 (4.6).

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 4.6 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:

$$\sigma \sim \text{Beta}(4.30, 13.51) \tag{4.18}$$

The prior distribution for σ is plotted in Figure 4.25A with the reported values of the proportion symptomatic from previous studies shown in 4.25B.

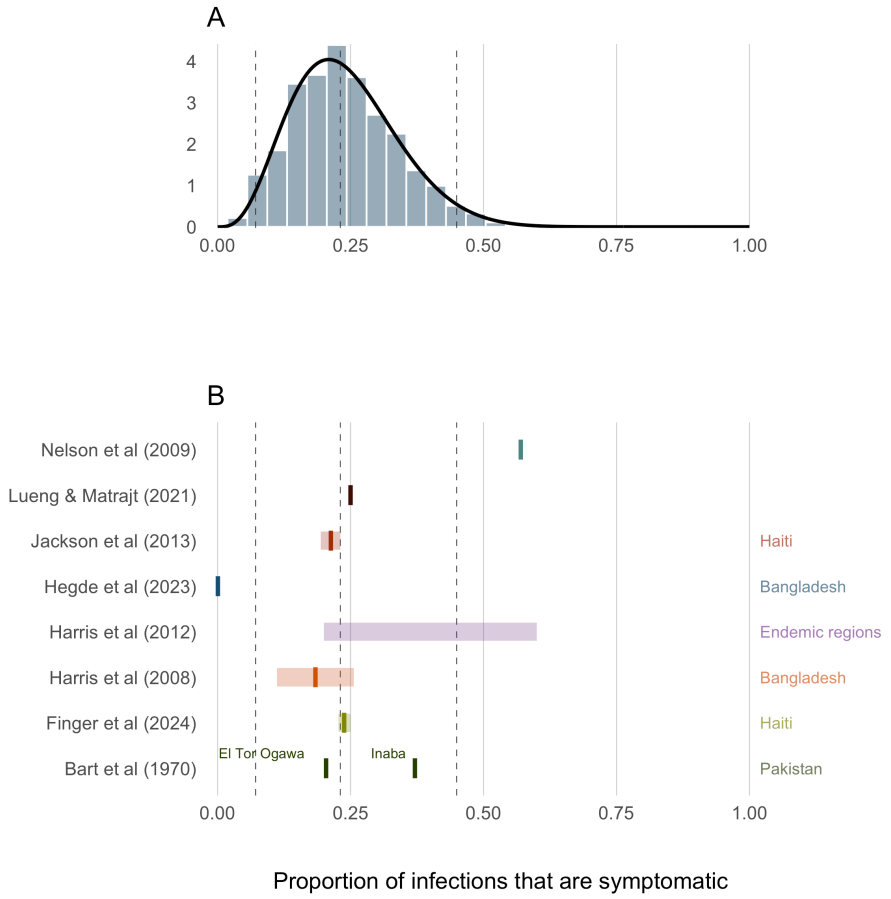

Figure 4.25: Proportion of infections that are symptomatic.

Table 4.6: 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/p
NA	1.000	0.250	NA	[Lueng & Matrajt (2021)](https://journals.plos.org/p
NA	0.600	0.200	Endemic regions	[Harris et al (2012)](https://www.sciencedirect.com/s
0.238	0.250	0.227	Haiti	[Finger et al (2024)](https://www.ncbi.nlm.nih.gov/p
0.213	0.231	0.194	Haiti	[Jackson et al (2013)](https://www.ajtmh.org/view/j
0.204	NA	NA	Pakistan	[Bart et al (1970)](https://doi.org/10.1093/infdis/121
0.371	NA	NA	Pakistan	[Bart et al (1970)](https://doi.org/10.1093/infdis/121
0.184	0.256	0.112	Bangladesh	[Harris et al (2008)](https://journals.plos.org/plosntd
0.001	0.000	0.001	Bangladesh	[Hegde et al (2024)](https://www.nature.com/articles

4.6.2 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 4.26B:

$$\rho \sim \text{Beta}(4.79, 1.53). \tag{4.19}$$

4.6.3 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

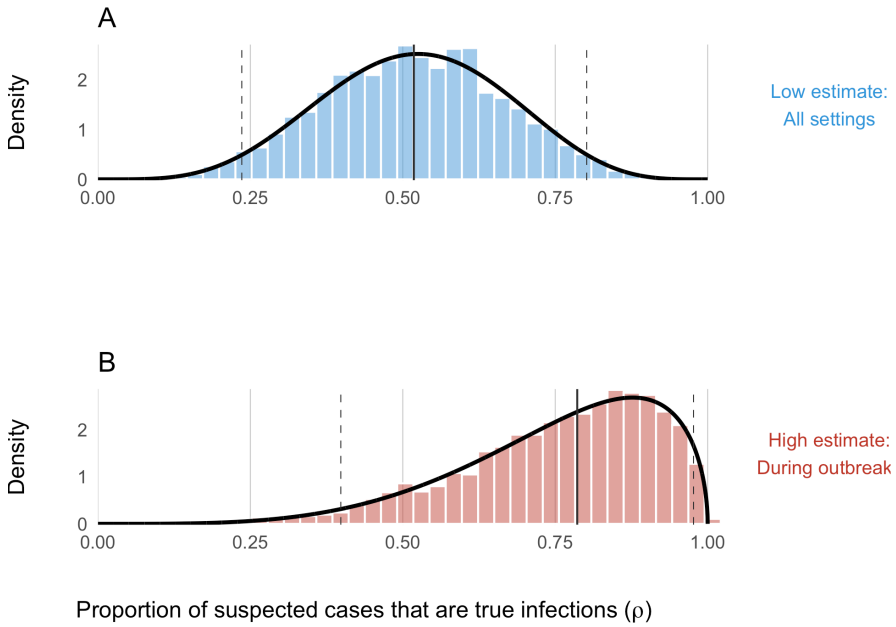

Figure 4.26: Proportion of suspected cholera cases that are true infections. Panel A shows the 'low' assumption which estimates across all settings: $\rho \sim \text{Beta}(5.43, 5.01)$. Panel B shows the 'high' assumption where the estimate reflects high-quality studies during outbreaks: $\rho \sim \text{Beta}(4.79, 1.53)$

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:

$$\begin{aligned} \text{CFR}_{\text{observed}} &= \text{CFR}_{\text{total}} \\ \frac{[\text{observed deaths}]}{[\text{observed cases}]} &= \frac{[\text{total deaths}]}{[\text{all infections}]} \\ \text{total deaths} &= \frac{[\text{observed deaths}] \times [\text{true infections}]}{[\text{observed cases}]} \\ D_{jt} &= \frac{[\sigma\rho\mu_j I_{jt}] \times [I_{jt}]}{[\sigma\rho I_{jt}]} \end{aligned} \tag{4.20}$$

4.7 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 4.8).

4.8 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 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:

Table 4.7: CFR Values and Beta Shape Parameters for AFRO Countries

Country	Cases (2014-2024)	Deaths (2014-2024)	CFR	CFR Lower	CFR Upper	
AFRO Region	1233933	23690	0.019	0.019	0.019	
Angola	2665	74	0.028	0.022	0.035	
Burundi	5581	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	29946	925	0.031	0.029	0.033	
Democratic Republic of Congo	315630	5751	0.018	0.018	0.019	
Congo	144	10	0.019	0.019	0.019	
Comoros	10549	152	0.014	0.012	0.017	
Ethiopia	72168	903	0.013	0.012	0.013	
Ghana	29825	251	0.008	0.007	0.010	
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	85191	306	0.004	0.003	0.004	
Malawi	62700	1848	0.029	0.028	0.031	
Namibia	485	13	0.027	0.014	0.045	
Niger	12666	355	0.028	0.025	0.031	
Nigeria	258048	7119	0.028	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	30566	653	0.021	0.020	0.023	
Eswatini	2	0	0.019	0.019	0.019	
Chad	1359	90	0.066	0.054	0.081	
Togo	509	29	0.057	0.038	0.081	
Tanzania	40965	636	0.016	0.014	0.017	
Uganda	9199	181	0.020	0.017	0.023	
South Africa	1403	47	0.033	0.025	0.044	
Zambia	30671	894	0.029	0.027	0.031	
Zimbabwe	45377	789	0.017	0.016	0.019	

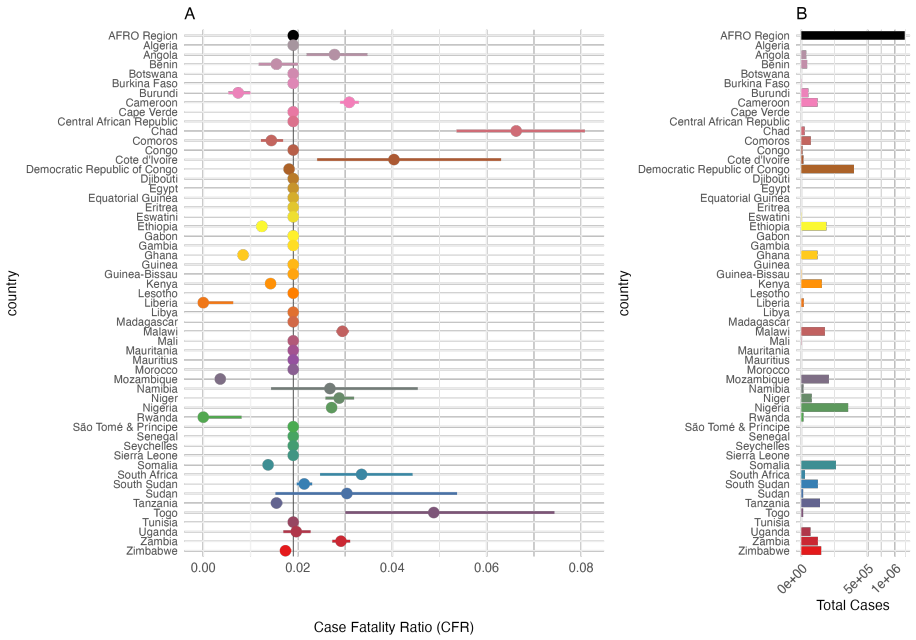

Figure 4.27: Case Fatality Rate (CFR) and Total Cases by Country in the AFRO Region from 2014 to 2024. Panel A: Case Fatality Ratio (CFR) with 95% confidence intervals. Panel B: total number of cholera cases. The AFRO Region is highlighted in black, all countries with less than $3/0.2 = 150$ total reported cases are assigned the mean CFR for AFRO.

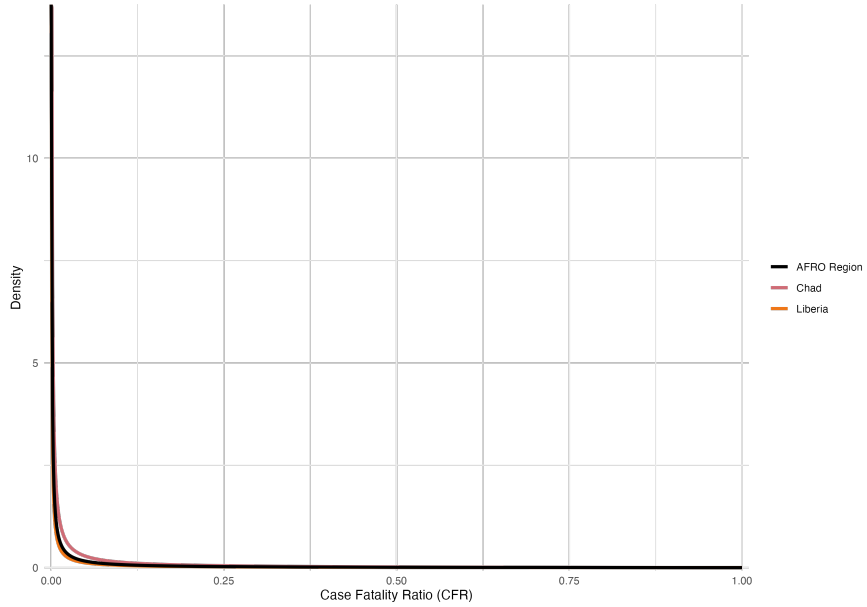

Figure 4.28: Beta distributions of the overall Case Fatality Rate (CFR) from 2014 to 2024. Examples show the overall CFR for the AFRO region (2%) in black, Congo with the highest CFR (7%) in red, and South Sudan with the lowest CFR (0.1%) in blue.

Table 4.8: 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

$$R_{jt} = \frac{I_{jt}}{\sum_{\Delta t=1}^{t} g(\Delta t) I_{j,t-\Delta t}} \tag{4.21}$$

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.

4.8.1 The generation time distribution

The generation time distribution gives the time between when an individual is infected and when they infect subsequent individuals. We parameterized this quantity using a Gamma distribution with a mean of 5 days:

$$g(\cdot) \sim \text{Gamma}(0.5, 0.1). \tag{4.22}$$

Here, shape=0.5, rate=0.1, and the mean if given by shape/rate. Previous studies use a mean of 5 days (Kahn et al 2020 and Azman 2016), however a mean of 3, 5, 7, or 10 days may be admissible (Azman 2012).

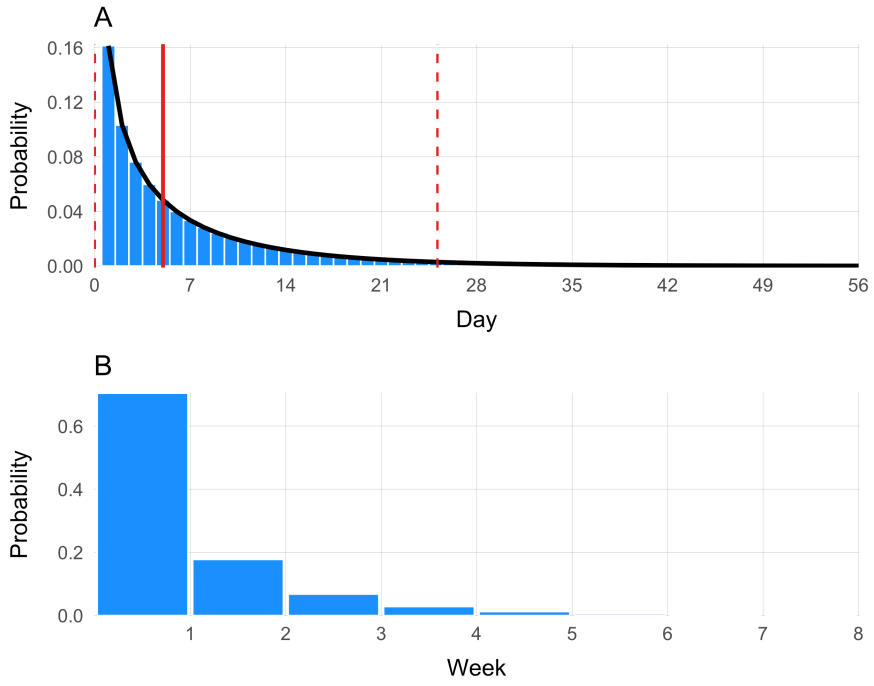

Figure 4.29: This is generation time

Table 4.9: Generation Time in Weeks

Interval	Week	Probability
[0,7]	1	0.704
(7,14]	2	0.178
(14,21]	3	0.068
(21,28]	4	0.028
(28,35]	5	0.012
(35,42]	6	0.006
(42,49]	7	0.003
(49,56]	8	0.001

4.9 Initial conditions

Since this first version of the model will begin on Jan 2023 (to take advantage of available weekly data), the initial conditions surrounding population immunity must be estimated. To set these initial conditions, we use historical data to find the total number of reported cases for a location over the previous X years, multiply by $1/\sigma$ to estimate total infections from those symptomatic cases that are reported, and then adjust based on waning immunity. We also sum the total number of vaccinations over the past X years and adjust for vaccine efficacy ϕ and waning immunity from vaccination ω.

- total number infected? From reported cases… back out symptomatic and asymptomatic
- Total number immune due to natural infections in the past X years
- total number immune due to past vaccinations in the X years

Use deconvolution based on immune decay estimated in vaccine section

4.10 Model calibration

- The model will be calibrated using Latin hypercube sampling for hyper-parameters and model likelihoods fit to incidence and deaths.
- An important challenge is flexibly fitting to data that are often missing or only available in aggregated forms.

[Fig: different spatial and temporal scales of available data]

4.11 Caveats

- Simplest model to start. Easier for initial spatial structure but with minimum additional compartments to calibrate to available data (vaccination,

cases, deaths).
- Country level aggregations. First generation data is 2023/24...
- Assumes vaccinating susceptible only individuals.
- For climate, summarizing for whole country.

4.12 Table of parameters

4.13 References

Table 4.10: Descriptions of model parameters along with prior distributions and sources where applicable.

Parameter	Description
i	Index i represents the origin metapopulation.
j	Index j represents the destination metapopulation.
t	Index t is the time step which is one week (7 days).
b_j	Birth rate of population j.
d_j	Overall mortality rate of population j.
N_{jt}	Total population size of destination j at time t.
S_{jt}	Number of susceptible individuals in destination j at time t.
V_{jt}	Number of effectively vaccinated individuals in destination j at time
I_{jt}	Number of infected individuals in destination j at time t.
W_{jt}	Total amount of V. cholerae in the environment in destination j at t
R_{jt}	Number of recovered (and therefore immune) individuals in destinatio
$\Lambda_{j,t+1}$	The force of infection due to human-to-human transmission in destina
$\Psi_{j,t+1}$	The force of infection due to environment-to-human transmission in de
ϕ	The effectiveness of Oral Cholera Vaccine (OCV).
ν_{jt}	The reported rate of vaccination with OCV in destination j at time
ω	Rate of waning immunity of vaccinated individuals.
ε	Rate of waning immunity of recovered individuals.
γ	Recovery rate of infected individuals.
μ	Mortality rate due to infection with *V. cholerae*.
σ	Proportion of *V. cholerae* infections that are symptomatic.
ρ	The proportion of suspected cholera cases that are true infections.
ζ	Rate that infected individuals shed *V. cholerae* into the environmen
δ	The environmental suitability dependent decay rate of *V. cholerae* i
δ_{min}	The minimum decay rate of *V. cholerae* in the environment obtained
δ_{max}	The maximum decay rate of *V. cholerae* in the environment obtaine
ψ_{jt}	The climatically driven environmental suitability of *V. cholerae* in d
β_{j0}^{hum}	The baseline rate of human-to-human transmission in destination j.
β_{jt}^{hum}	The seasonal rate of human-to-human transmission in destination j
β_{j0}^{env}	The baseline rate of environment-to-human transmission in destinatio
β_{jt}^{env}	The rate of environment-to-human transmission in destination j at t
α	The overall population mixing parameter.
τ_i	The probability that an individual departs origin i.
π_{ij}	The probability that an individual travels from origin i to destinatio
θ_{j}	The proportion of the population that have adequate Water, Sanitatio
κ	The concentration (number of cells per mL) of *V. cholerae* required

Chapter 5

Model versions

Table 5.1: Current and future planned model versions with brief descriptions.

Version	Description
v0.1	**Current:** Beta version of the model to establish mechanisms and links to data. Basic SIR dynamics
v1.0	**Future:** First implementation in LASER. Model definition is the same. Improvements to data sourc
v2.0	**Future:** Maximizing metapopulation approach. District level data and improvements to model fittin
v3.0	**Future:** Agent-based component with better immune dynamics.

Chapter 6

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.

6.1 Vaccination

6.1.1 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.

6.1.2 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).

6.2 Impacts of Climate Change

6.2.1 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).

6.2.2 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 7

Usage

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

Chapter 8

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 9

References