Malaria Journal

RESEARCH

Open Access

An archetypes approach to malaria intervention impact mapping: a new framework and example application

Amelia Bertozzi-Villa[1,2,3*], Caitlin A. Bever[1], Jaline Gerardin[1,4], Joshua L. Proctor[1], Meikang Wu[1], Dennis Harding[1], T. Deirdre Hollingsworth[3], Samir Bhatt[5,6] and Peter W. Gething[2,7]

Abstract

Background As both mechanistic and geospatial malaria modeling methods become more integrated into malaria policy decisions, there is increasing demand for strategies that combine these two methods. This paper introduces a novel archetypes-based methodology for generating high-resolution intervention impact maps based on mechanistic model simulations. An example configuration of the framework is described and explored.

Methods First, dimensionality reduction and clustering techniques were applied to rasterized geospatial environmental and mosquito covariates to find archetypal malaria transmission patterns. Next, mechanistic models were run on a representative site from each archetype to assess intervention impact. Finally, these mechanistic results were reprojected onto each pixel to generate full maps of intervention impact. The example configuration used ERA5 and Malaria Atlas Project covariates, singular value decomposition, k-means clustering, and the Institute for Disease Modeling's EMOD model to explore a range of three-year malaria interventions primarily focused on vector control and case management.

Results Rainfall, temperature, and mosquito abundance layers were clustered into ten transmission archetypes with distinct properties. Example intervention impact curves and maps highlighted archetype-specific variation in efficacy of vector control interventions. A sensitivity analysis showed that the procedure for selecting representative sites to simulate worked well in all but one archetype.

Conclusion This paper introduces a novel methodology which combines the richness of spatiotemporal mapping with the rigor of mechanistic modeling to create a multi-purpose infrastructure for answering a broad range of important questions in the malaria policy space. It is flexible and adaptable to a range of input covariates, mechanistic models, and mapping strategies and can be adapted to the modelers' setting of choice.

Keywords Malaria, Modeling, Mapping

*Correspondence:
Amelia Bertozzi-Villa
abertozzivilla@idmod.org
Full list of author information is available at the end of the article

Background

Malaria is one of humanity's oldest and most insidious ailments, co-evolving with mosquitoes and humans over millions of years [1]. This long history, combined with its complex transmission pathways, makes malaria a uniquely heterogeneous and environmentally sensitive disease. Understanding malaria in a given location requires not just an understanding of human behavior, movement, and demographics, but also a detailed knowledge of resident mosquito species and their behavior, local climate and hydrology, and the seasonal patterns of the landscape.

Computational modeling and malaria policy have been linked since the 1950s, from seminal pen-and-paper equations by Ronald Ross and George Macdonald to today's detailed computational models capable of simulating individual humans, mosquitoes, climates, immune responses, malaria interventions, and much more. The fundamental goal of these models is to help decision-makers craft a malaria strategy that aligns with the epidemiological, economic, and cultural circumstances at hand.

The two main classes of malaria model in use today are mechanistic transmission models, which explicitly simulate disease spread in a population, and spatiotemporal models, which utilize geospatially-referenced covariates and classical statistical methods to generate estimated maps of malaria burden, intervention coverage, and other malaria-relevant features. The Malaria Atlas Project (MAP) is known for its high-resolution maps of malaria-related variables [2–10]. Over the last 15 years a number of mechanistic modeling groups, most prolifically at Imperial College, the Swiss Tropical and Public Health Institute, Northwestern University, PATH, the Ifakara Health Institue, the MORU Tropical Health Network, and the Institute for Disease Modeling (IDM), have supported decision-making at the local, national, and global scale [11–15].

Mechanistic and spatiotemporal models serve complementary purposes in malaria. Spatiotemporal models elucidate the past and present of disease burden and related metrics, which mechanistic models can then use for insight into the future. Maps provide a rich descriptive landscape, while mechanistic models contribute causal structure and exploration of counterfactual or hypothetical scenarios. As both modeling methods have grown in popularity and demand, there is increased interest in products that combine the two approaches, allowing for a spatially and temporally detailed exploration of the consequences of different policy decisions. This paper introduces a novel methodology which combines the richness of spatiotemporal mapping with the rigor of mechanistic modeling to create a multi-purpose infrastructure for answering a broad range of important questions in the malaria policy space.

Mechanistic models draw strength from their ability to replicate detailed relationships, but suffer several limitations. First, these detailed models require a large amount of input data to configure and calibrate to a particular setting. When tasked with simulating settings without such a rich data space, mechanistic models can be a challenge to appropriately configure. Second, mechanistic model complexity carries with it an associated computational cost. A single simulation can take minutes or hours to run, even on high-performance computing systems, often limiting the number of scenarios or locations that can reasonably be simulated. In contrast to these challenges, spatiotemporal methods are designed with incomplete data as an assumption and perform well at high resolutions [16, 17]. However, these statistical methods crucially lack the explicit causal relationships that allow mechanistic models to effectively test the consequences of different policies.

In the past, mechanistic modeling analyses working at the continental scale in Africa have utilized a range of methods to address these practical and computational challenges. Griffin et al. [18] modeled a range of stylized seasonality patterns and mosquito species mixes to assess intervention strategies. While this approach showcases hypothetical scenarios, there is no explicit link between any true location and any stylized seasonal profile, making geographic trends or burden estimation impossible. Walker et al. [11] grouped rainfall, entomological, and transmission intensity spatial covariates into three, four, and 18 clusters respectively, and ran simulations on all possible combinations of these groups before re-projecting onto the pixel level. This exhaustive strategy does not take advantage of the spatial relationships between different covariates, which could reduce computational burden and generate more informative cluster properties.

The methodology presented in this paper leverages the strength of both mechanistic and spatiotemporal methods to allow computationally-feasible generation of high-resolution maps that reflect mechanistically modeled scenarios. This process generates a range of archetypal seasonal and entomological profiles that can be useful for exploratory data analysis while also generating an explicit link between these profiles and any given location across Africa. This mapping capacity provides a computationally efficient pathway for presenting results geospatially and in terms of expected burden change.

Here, high-resolution, high-dimensional spatial covariates and machine learning methods are harnessed to generate a small number of spatially-explicit "archetypes" of malaria transmission, characterized by their covariate similarity. Next, mechanistic models are

run on a representative site from each archetype, rather than on every pixel individually. Finally, spatial data and model results create a lookup table through which maps of intervention impact are generated.

Sorting many observations of high-dimensional data into groups by similarity is a common problem in machine learning. A common solution is to deploy a two-step process of first reducing the dimensionality of the dataset, and then clustering this reduced-dimensional data into groups [19–21]. This approach is appealing for its flexibility, but requires a number of specific decisions that can materially impact results. Namely:

1 Which covariates should be selected, and how should they be standardized?
2 Which algorithm should be used for dimensionality reduction?
3 Which algorithm should be used for clustering, and how should the number of clusters be determined?

When clustered results are being used as inputs for mechanistic models and subsequent generation of new maps, additional questions emerge:

4 How should representative sites be selected from each cluster?
5 Which mechanistic model should be used to assess intervention impact?
6 How should results from representative sites be reprojected back onto other members of the group?

This paper reviews each of these questions in turn, beginning with a discussion of the choices available in each space and their implications. As a demonstration, it describes the results of an eradication feasibility exercise undertaken by IDM and MAP in 2018. The original work was conducted by request of international stakeholders seeking guidance on whether eradication might be possible under highly optimistic environmental circumstances and malaria control scenarios. It harnessed spatiotemporal climate and mosquito covariates, IDM's EMOD malaria transmission model, and MAP's malaria prevalence maps to generate scenarios of malaria intervention impact at the 5km-by-5km pixel level for all of sub-Saharan Africa in 2050. The original analysis [22, 23] utilized an earlier version of the framework, and the version presented here represents a validation check to ensure that similar results arise from a more in-depth approach. A sensitivity analysis exploring the representativeness of a given "representative site" is also described.

The primary goal of this document is to introduce a novel modeling paradigm in sufficient detail to be adapted for other use cases. Throughout, the terms "archetypes strategy" and "archetypes framework" will be used to describe the general methodology of using dimensionality reduction and clustering to locate a subset of sites appropriate for mechanistic modeling. The term "example configuration" will be used to describe the specific set of choices within the broader strategy that are shown here for demonstration. The archetypes strategy is flexible and generalizable to any covariate set, mechanistic modeling platform, and geographic scale. The methods sections are subdivided into a general discussion and an explicit description of the example configuration parameters. The results section focuses on describing example configuration outcomes. Finally, the discussion section highlights lessons learned and other use cases of this archetypes framework.

Methods

In the sections below, each of the six methodological questions described in the introduction is explored in detail. General theory is described first, followed by a description of the specific choices made in the example configuration. For the example configuration, simplicity was favored, but more complex options are noted.

Covariate selection

General

The covariates used for clustering should be selected mindfully to capture the types of variation to which the transmission model is most sensitive. For malaria, these should almost always include covariates that capture the different seasonality patterns of different malaria-endemic regions, as this directly impacts disease seasonality. For models that explicitly simulate different mosquito behaviors by species, covariates describing relative species abundance are also valuable. Baseline malaria transmission intensity is another useful source of input data, but its inclusion depends on the use case. The example configuration described below intentionally excluded it from the clustering process, instead running simulations for each archetype over a range of transmission intensities. Other possible covariates of interest include intervention history, health care accessibility, population demographics, and other behavioral inputs. Covariate data must geographically cover the region of interest, and should be utilized on the same spatial resolution as the final results. Covariate data for malaria will often also include a temporal component showing either seasonality or secular time.

Covariates are collected into a "stack" of spatial or spatiotemporal input files measuring different metrics on different scales. It is important to normalize or rescale these covariates prior to performing

dimensionality reduction to avoid an artificial effect due to differently-scaled inputs. Even after normalization, however, a decision must be made regarding the relative weight of different inputs. For instance, in the example configuration below, clustering covariates included 12 rainfall layers, 12 temperature layers, and three mosquito species relative abundance layers. These covariates were not differentially weighted prior to dimensionality reduction, meaning that the rainfall and temperature layers had a stronger impact on results than the mosquito abundance layers. This choice was acceptable for the work at hand, but different circumstances might encourage different weighting.

Example configuration

Environmental covariate data on rainfall and air temperature were sourced from the ERA5 project (https://www.ecmwf.int/en/forecasts/datasets/reanalysis-datasets/era5), a global reanalysis that generates internally-consistent estimates for a wide range of climate parameters across the globe. Specifically, monthly mean total precipitation and 2-meter air temperature from 2000 to 2018 were downloaded. Monthly means were averaged across the time series to generate synoptic 12-month trends. Because ERA5 spatial resolution is 0.25°-by-0.25° (approximately 30km-by-30km at the equator), covariates were resampled down to the 5km-by-5km level and realigned to match MAP's standard spatial specifications.

Covariate data on the relative abundance of *Anopheles arabiensis*, *funestus*, and *gambiae* were obtained from Sinka et al. [6]. These values are static estimates and did not require resampling as they already met MAP spatial specifications.

All covariate rasters were masked to align with MAP estimates of *Plasmodium falciparum* transmission limits, and pixel-level values extracted into a nonspatial dataset for further analysis. Once extracted, pixel-level values of air temperature and rainfall were rescaled to fall between zero and one. Because they contained some extreme values that severely skewed the distribution, the most extreme 1% of rainfall values were reassigned to the 99th percentile value before being rescaled. Relative vector abundance values are proportions, and therefore these values already fell between zero and one and did not require rescaling. See Fig. 1 for covariate distributions before and after rescaling.

Dimensionality reduction

General

If the covariate selection process only locates a few variables of interest, a dimensionality reduction step may not be necessary. However, beyond the point that covariates can comfortably be plotted together (four or five layers at most), dimensionality reduction can be a valuable tool both for data exploration and for more effective clustering. The goal of these techniques is to collapse high-dimensional data into a lower-dimensional space in a way that preserves as much of the original data variation as possible, creating a denser and richer database from a sparser one. A wide range of dimensionality reduction techniques are available in machine learning software packages.

This analysis utilized Singular Value Decomposition (SVD), a strategy similar to Principle Components Analysis which locates the set of orthogonal vectors in a high-dimensional dataset that cover the most variance in the data, and returns those vectors in order of variance explained [24, 25]. SVD is evaluated by plotting how much variance in the original dataset is explained by the most informative singular vectors. If a small number of vectors covers a large portion of the initial variance, then only that smaller-dimensional dataset need be retained for further analysis. While straightforward to implement, SVD is not a time-series method, meaning that dimensionality reduction does not explicitly take into account causal correlations between the different monthly layers of climate data. Other strategies, such as Fourier transforms or dynamic mode decomposition, are worth consideration if simple methods such as SVD do not yield informative results, or if the input data covers long time series rather than synoptic trends.

Example configuration

The final covariate dataset was comprised of 27 dimensions: 12 layers each of rainfall and air temperature data and three layers of relative vector abundance data. SVD was applied using the `svd` command in `R` version 3.6.2. Points that are close together in this reprojected space are expected to be similar in terms of their rainfall, temperature, and relative vector proportions. Upon running SVD, the first three singular vectors explained over 95% of the variation in the full dataset, and were retained for clustering (Fig. 2).

Clustering

General

The literature and theory on clustering algorithms is vast [26, 27], including many excellent tutorials for beginners. Briefly, clustering strategies are unsupervised learning methods that group data points together based on proximity in all provided dimensions.

The example configuration utilized k-means for clustering. This algorithm benefits from simplicity and

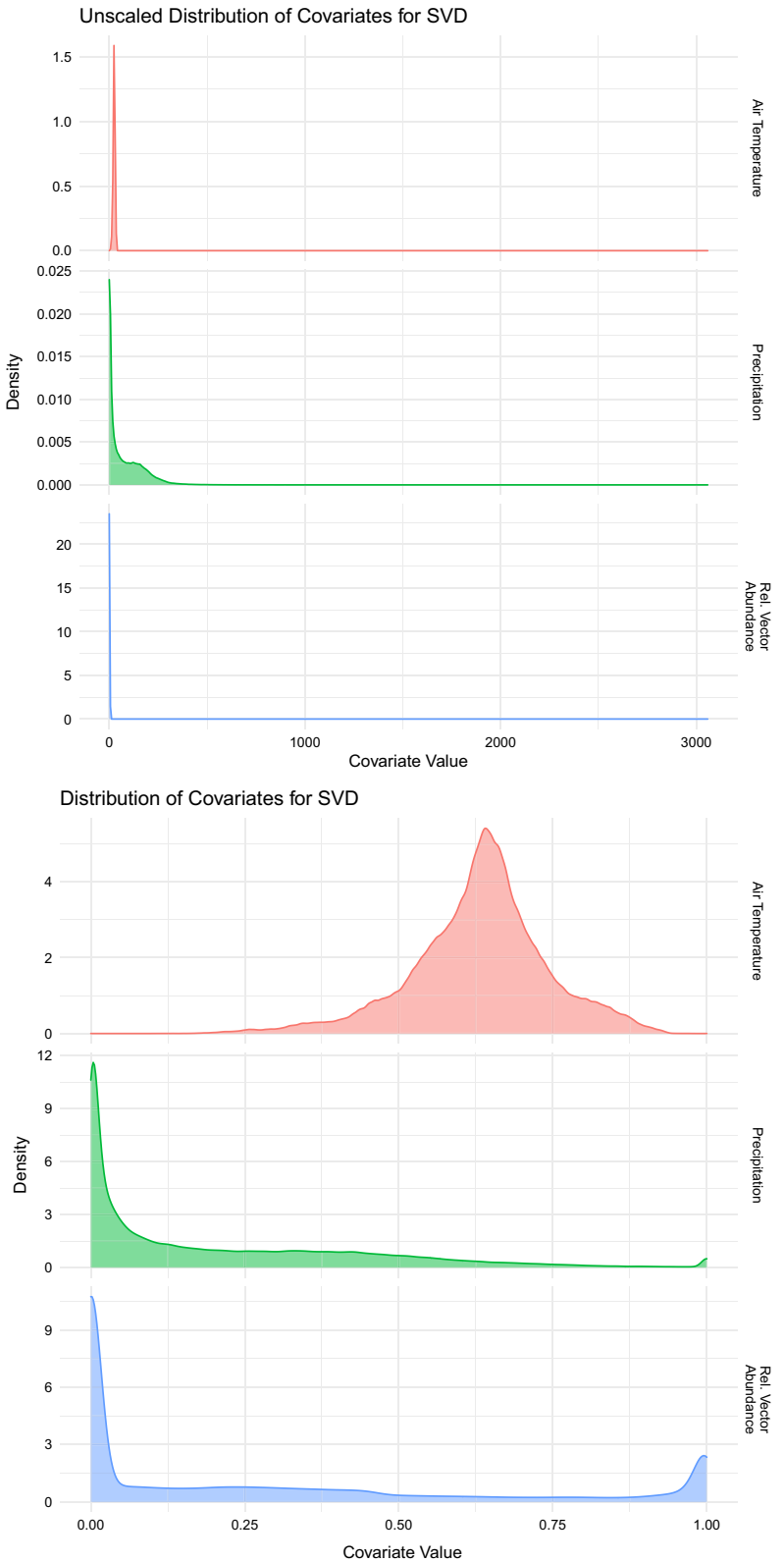

Fig. 1 Distribution of covariates for Singular Value Decomposition (SVD) across all pixels before and after rescaling. Synoptic monthly mean temperature (degrees C) and rainfall (mm/month) are from the ERA5 project. Relative vector abundance, a static metric, tracks the proportion of *Anopheles arabiensis*, *funestus*, and *gambiae* in each pixel

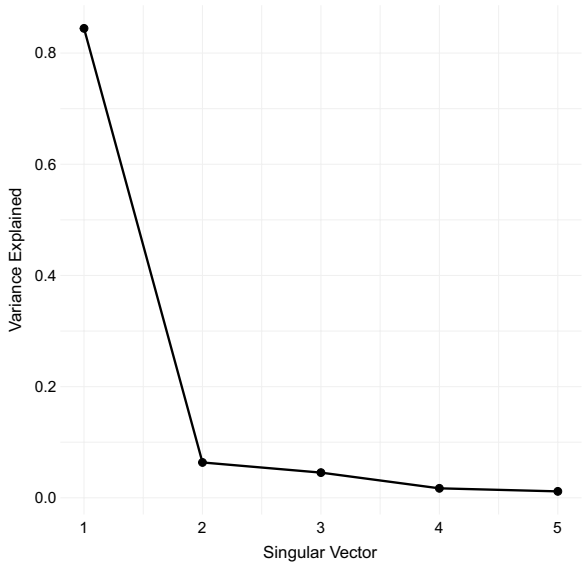

Fig. 2 Proportion of variance explained by the first five singular vectors after singular value decomposition (SVD). The first three vectors were retained for clustering analysis

relatively quick runtime using an iterative geometric search, but suffers two major setbacks. First, the choice of cluster count *k* must be specified by the user, which can lead to nonintuitive cluster groupings if an inappropriate *k* is selected. Intuitively, the goal of k-means is to minimize data variance within clusters while maximizing it between clusters. A heuristic but common solution for addressing the cluster count problem is the use of "elbow plots", in which cluster count is plotted against the ratio of between-cluster variance and total variance. This ratio is monotonically increasing, since the cluster count that maximizes this ratio is a *k* equal to the number of data points, but the goal is to find a *k* for which the plot makes an "elbow" and begins to increase more slowly. Other strategies for selecting cluster counts for k-means include the use of Jaccard, Dunn, and Silhouette indices [27].

A second weakness of k-means is its radial search methodology. Because this clustering algorithm minimizes the Euclidean distance between a cluster centroid and the data points around it, clusters that are tightly grouped in all dimensions will be captured more effectively than elliptical or other cluster shapes. Heuristically, this did not appear to affect the algorithm's ability to find distinct groups, but nonspherical methods such as CLARA or Gaussian Mixture Models (GMMs) may be more appropriate in some settings. GMMs, which utilize a Bayesian fitting procedure, have the added benefit of allowing for the calculation of information criteria, giving a more principled and less heuristic way to select cluster counts compared to k-means.

Example configuration

Having reduced the dimensionality of the dataset from 27 to three, the next step was to classify points into transmission archetypes representing unique environmental and entomological patterns of relevance to malaria transmission. To identify these distinct archetypes, k-means clustering was performed on the reduced-dimensional dataset output by SVD. K-means was run using the MacQueen algorithm [28] on the reduced-dimensional dataset over a range of possible cluster counts from 3 to 15, and elbow plots computed. Unfortunately, these plots did not show a strong elbow (Fig. 3), so a cluster count of 10 was selected as a compromise between relatively high variance ratio and a visually intuitive and distinguishable set of archetypes (Figs. 4, 5, 6). K-means was applied using the `kmeans` command in the `stats` R package version 4.0.5. See the Results section for a more in-depth description of k-means outputs.

Selecting representative sites

General

The clustering step groups the data into archetypes. Next, a set of specific covariate values must be selected for each archetype to use as inputs into the transmission model. One viable option would be to take the

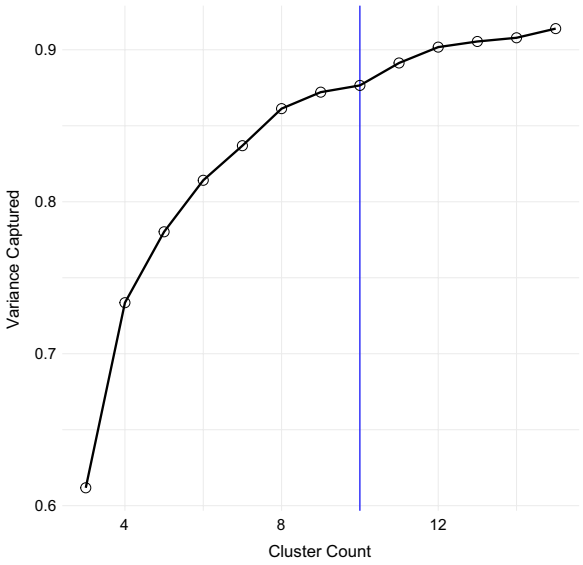

Fig. 3 K-means elbow plot for *k* between 3 and 15. The x-axis shows cluster count, while the y-axis shows the proportion of total data variance that is captured by between-cluster variance in each setting. A cluster count of ten (vertical blue line) was chosen for further analysis

Fig. 4 K-means results. Top and side view of pixel values reprojected onto the first three singular vectors, after k-means clustering with a *k* of ten. Colors refer to different clusters and match subsequent ten-cluster plots. Black dots indicate true cluster centroids, while dark grey dots show procedurally-selected representative sites

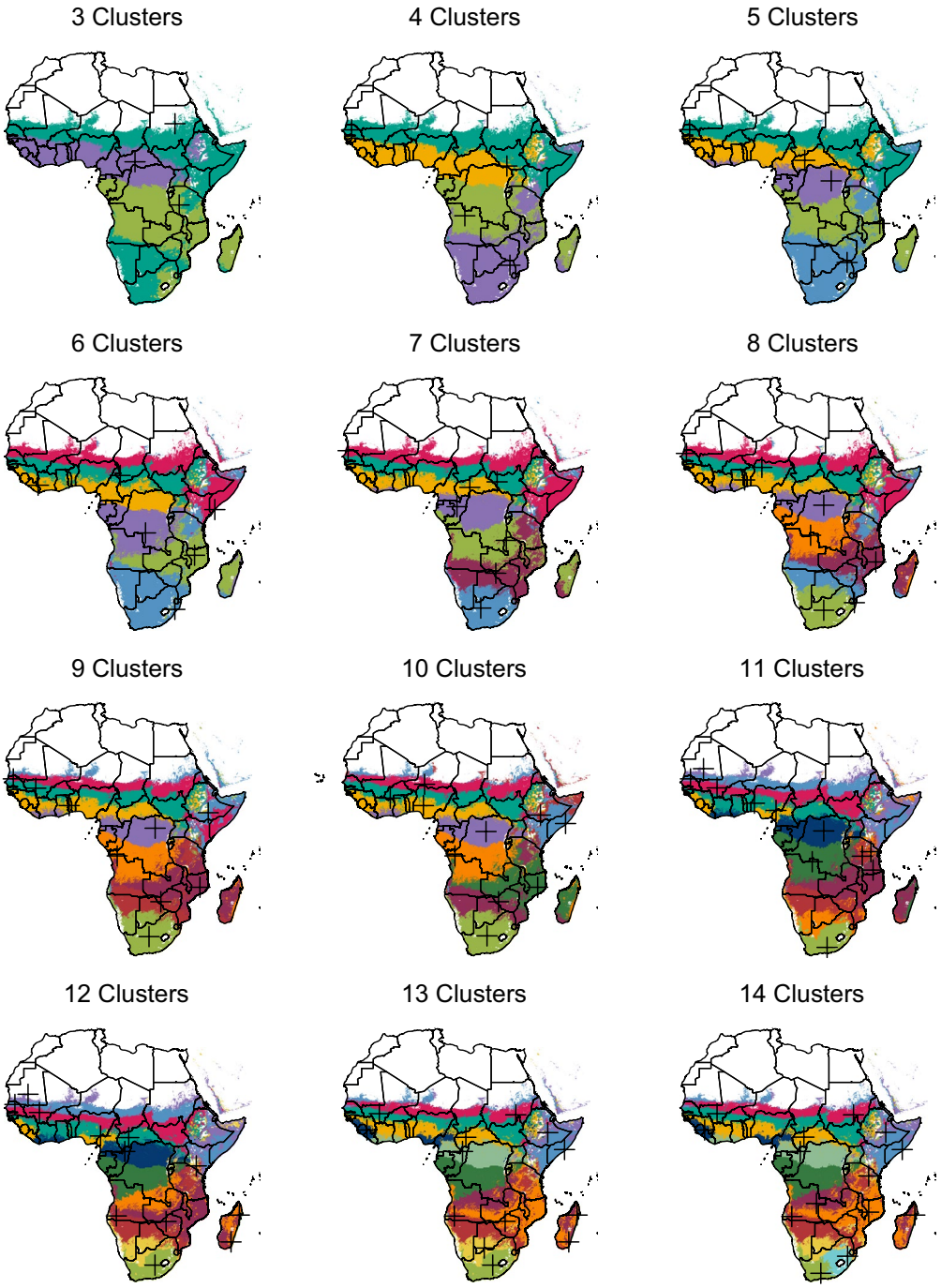

Fig. 5 Cluster maps for *k* of 3 to 14. Black crosses indicate procedurally-generated representative sites (the pixels closest to cluster centroids)

mean or median across all data points and use these as model inputs for the archetype, acknowledging that this measure of center will not reflect any true location and may be sensitive to outliers in the clustering process. Using the values at each cluster centroid would be less sensitive to outliers, but still not reflect any true physical location. A third alternative, used in the example configuration, is to select the data point closest to each k-means cluster centroid and use this as a representative site for the archetype as a whole. Visual examination of the k-means clusters (Fig. 4) shows that the third approach is nearly identical to the second in the example configuration, as there is generally a trivial distance between the true cluster centroids and

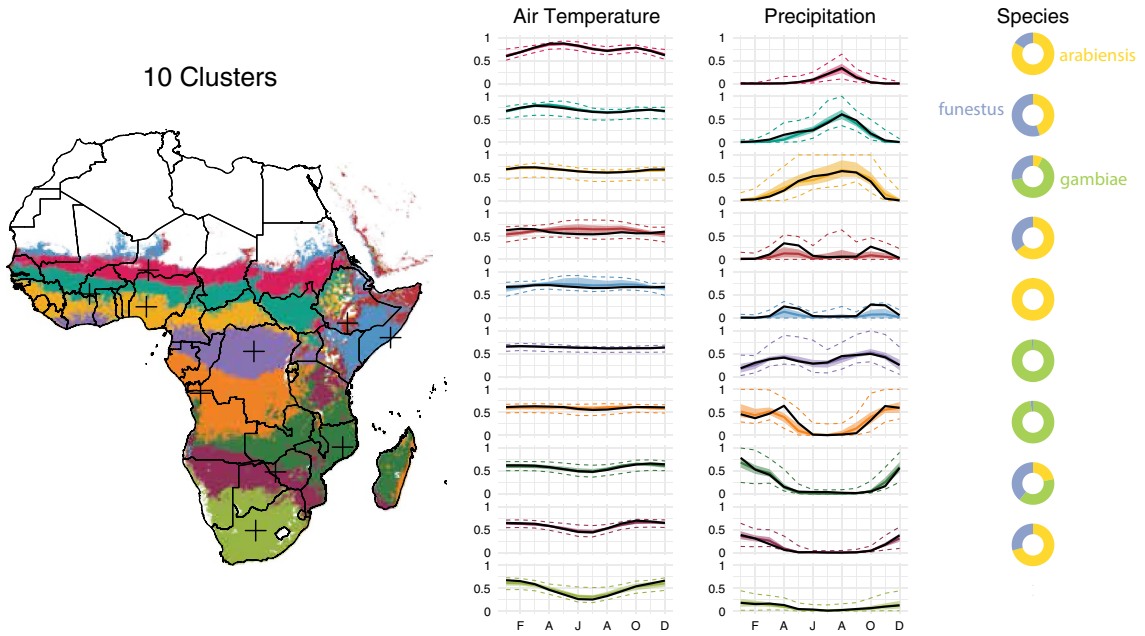

Fig. 6 Maps and time series for the ten-site setting. In the time series, solid colored lines represent the median across the archetype, shaded areas indicate the interquartile range, and colored dotted lines indicate the 95% variance interval. Solid black lines represent the procedurally-selected climate values of the representative site for each archetype, also indicated as black crosses on the map. Doughnut plots show the relative vector abundance of the representative sites

their nearest data point. A sensitivity analysis was conducted to assess the acceptability of this approach.

Example configuration

The data point closest to each cluster centroid (determined using k-Nearest Neighbors, with *k* set to one) was selected as the site whose input data would be used for simulation. K-nearest neighbors was applied using the `get.knn` command in the `FNN` R package version 1.1.3.

Site selection sensitivity analysis

In the main analysis, a single centrally-located point is used as a proxy for all other locations in that cluster. A sensitivity analysis was conducted to assess how other locations in each cluster behaved under the same intervention scenarios. Ten pixels were randomly chosen from each cluster, and all intervention scenarios were re-run on these 100 points in the same way that they were on the original ten. Results were compared to those from the cluster centroid sites alone to assess within-cluster variance.

Selecting and configuring a malaria model

General

A number of mechanistic malaria models are commonly used by groups consulting with national or international malaria policy groups. These include OpenMalaria [29], EMOD [14], the Imperial model [18], and VCOM [30]. These models vary with respect to their structure, ease of use, and public accessibility and can produce measurably different estimates under similar initial conditions [31]. Modelers should have a close familiarity with their model of choice prior to embarking on the exercises described in this document. In particular, it is necessary to understand the input parameters to which each model is most sensitive and how this sensitivity impacts covariate and archetype selection; how to appropriately parameterize each site's demographics, climate, entomology, and interventions; and how to run checks to diagnose unexpected model behavior before acting upon any conclusions from model results.

When pursuing the archetypes strategy, modelers will utilize values from the clustering covariates for some model inputs, but must choose how to initialize all variables not included in the clustering covariates. For instance, in the example configuration, the clustering analysis provided input rainfall, temperature, and entomology variables, but not demographic, intervention history, or transmission intensity variables. As described in detail below, this analysis held most other variables static across all sites, and ran simulations at a range of initial transmission intensities, but different use cases might call

for different input parameters in different representative sites.

Example configuration

This modeling analysis was conducted using the EMOD simulation software produced by the Institute for Disease Modeling. See the Appendix for an introduction to EMOD. All simulations were run with an initial population of 2000 to balance population size with computational constraints, a birth rate of 36.3 live births per 1000 people per year from the World Bank Database, and a complementary mortality rate to ensure a stable population over time. Each human in the model was assigned a unique risk of being bitten by a mosquito, such that the distribution of biting risk in the community overall was exponential [32].

Simulations were initialized by running one 39-year intervention-free simulation for each of 25 initial transmission intensities and 10 random seeds, generating a collection of 250 baseline populations upon which to test intervention impact. Transmission intensities were varied by scaling mosquito larval habitat capacity, which linearly impacts the number of adult mosquitoes and the level of malaria transmission in the absence of interventions. The intervention-free immunity establishment period should approximate the length of a human life, and is frequently set to 50 years. A 39-year period was selected because this is the duration of time for which climate data was available from ERA5. Ten is a common choice for number of random seeds to test in EMOD, and has been shown to cover the variation in most parameters well. Twenty-five transmission intensities were chosen to thoroughly cover variability in this important parameter. One hundred fifty-two different intervention packages were tested, covering a wide range of vector control, drug, and vaccine-based malaria prevention strategies. These intervention parameters were originally part of a project to test eradication feasibility in best-case scenarios, and therefore often represent coverage levels or policies more rigorous than those commonly in use today (for example mass bednet campaigns every year instead of every three years). Each intervention was run for three years, and its efficacy assessed by mean *Plasmodium falciparum* parasite rate among 2–10 year-olds ($PfPR_{2-10}$) in the final year of the intervention compared to the final year of the intervention-free simulation.

Climate inputs of air temperature, rainfall, and relative humidity were constructed from the publicly-available ERA5 climate reanalysis model. For each representative site, daily climate data from 1980 to 2018 was downloaded from ERA5. The "2m_temperature" channel was used for air temperature, the "total_precipitation" channel for rainfall, and the "2m_temperature" and "2m_dewpoint_temperature" channels used to calculate relative humidity according to a method developed by the company Vaisala [33]. The intervention-free simulations to establish population immunity were run using climate data from 1980 to 2018, and the three-year intervention simulations were run using climate data from 2016 to 2018.

Mosquito parameters can be found in Table 1, and intervention descriptions and parameters in Tables 4 and 3.

Recreating maps

General

Once model simulations have run for each archetype's representative site, these individual-site results must be reprojected back onto pixel level results. The methodology described below constructs a lookup table to convert from baseline per-pixel transmission to intervention-impacted transmission in each archetype. If transmission intensity was included as a clustering covariate, the lookup table approach would differ in its details but be similar in essence.

Example configuration

Once all EMOD simulations were complete, maps of intervention coverage were reconstructed as follows. First, a pixel-level map of *Plasmodium falciparum* prevalence among children aged 2–10 ($PfPR_{2-10}$) was selected from the Malaria Atlas Project. Because these EMOD simulations were initiated from intervention-free scenarios, the MAP estimate of $PfPR_{2-10}$ in 2000 was selected as a proxy for malaria prevalence in the absence of any interventions. Then, for each intervention scenario and each pixel p in the selected map, the following steps were taken:

1. The archetype a to which the pixel p belongs was identified;
2. A spline was generated between data points of initial and final $PfPR_{2-10}$ across all transmission intensities in archetype a;
3. This spline was used to map the initial $PfPR_{2-10}$ corresponding to that of pixel p and archetype a to the final $PfPR_{2-10}$ in that intervention scenario;

Table 1 EMOD mosquito species parameters

Name	Anthropophily (%)	Endophily (%)
arabiensis	65	50
funestus	65	85
gambiae	85	85

4 This final $PfPR_{2-10}$ was logged as the "intervention impact" of pixel p.

This allows for the reconstruction of maps hypothesizing the potential impact of different interventions across the continent.

Results

Presented here are the results of the example configuration.

Covariate scaling, dimensionality reduction, and clustering

Before rescaling, synoptic ERA5 air temperature values ranged from 2.2 to 39.0 degrees Celsius, with a median value of 24.9 degrees and an interquartile range (IQR) from 22.4 to 27.0. ERA5 rainfall ranged from 0 to 3,058 mm/month, with a median value of 40.6 and an IQR from 4.9 to 130.8. Rainfall values were capped at the 99th percentile cutoff of 367 mm. Relative mosquito abundance values ranged from 0 to 1, with continent-wide proportions of 43.0% *arabiensis*, 36.8% *gambiae*, and 20.2% *funestus* among those pixels with mosquitoes. Figure 1 shows covariate distributions before and after rescaling. Both rainfall and mosquito vector abundance show peaks near zero and one, while temperature peaks most strongly in the center of the distribution.

Figure 2 shows the variance explained by the first five singular vectors in the SVD procedure. The first three singular vectors accounted for 95.3% of all variance and were retained for the clustering analysis. Figure 3 shows the elbow plot for the k-means procedure. It shows that, while a k of ten captures over 85% of total data variance as between-cluster variance, a slighlty higher cluster count would have covered over 90% of this ratio. Figure 4 shows a snapshot of the reprojected data points in 3D space, colored according to the k-means results. Black dots represent true cluster centroids, while gray dots indicate representative sites. The shape of this data is roughly circular or toroidal, with those sites that have appreciable numbers of mosquitoes and rainfall comprising the edges of the circle and only the low-rainfall, low-mosquito-count lime green vector clustering near the origin. Visual examination shows that the ten-cluster k-means differentiates several distinct groups, but may combine some groups that are visually fairly distinct, such as the two lobes of the turquoise cluster (Cluster 4).

Archetype creation and site selection

Figure 5 shows archetype maps over a range of cluster counts from 3 to 14. Across all cluster counts, archetypes tend toward geographic contiguity. Given that the clustering covariates themselves demonstrate strong spatial autocorrelation, this result is perhaps unsurprising, but lends algorithmic credence to historical strategies of heuristically grouping regions by climate or seasonality. Also notable is the stability of archetype affiliation across different cluster counts. The three-cluster map identifies a strongly seasonal Sahelian/coastal band across west Africa, a band in central Africa with opposite seasonality to the first, and more arid regions to the north and south. Adding clusters allows separation of the northern and southern arid regions, followed by increasingly distinct latitudinal bands in west Africa and a stable boundary between central Democratic Republic of the Congo (DRC) and northern Zambia and Mozambique. Lake Victoria, Rwanda, and Burundi are also reliably grouped together and distinct from their immediate surroundings, usually matched to an archetype further south. The southern tip of Madagascar is reliably grouped with southern Mozambique, while the rest of the island joins the same archetype as northern Mozambique and Zambia. The eastern coastline of Madagascar sometimes joins the archetype dominated by southern DRC. The ten-cluster map was selected for the example configuration to strike a heuristic balance between variance explained in the elbow plot (Figure 3) and the communicative value of showing archetypes with visually distinct seasonality and mosquito mixes.

Figure 6 shows a map of the ten archetypes used in the example configuration, along with the associated temperature and rainfall time series and relative abundance of different mosquito species. In the time series, solid colored lines represent the median across the archetype, shaded areas indicate the interquartile range, and dotted lines indicate the 95% variance interval. Solid black lines represent the climate values of the representative site for each archetype, also indicated as black crosses on the map. Additional file 1 shows similar maps for all cluster counts tested. The ten-archetype setting captures a detailed range of different climate modalities on the continent. Three latitudinal bands across the Sahara to the coast show seasonal rainfall of increasing magnitude and duration (magenta, turquoise, and gold). Relative *arabiensis* proportions decline, and *gambiae* proportions increase, along the same gradient. These latitudinal bands continue down the western side of the continent, with a bimodal *gambiae*-dominated archetype in northern DRC (purple), a strongly seasonal *gambiae*-dominated archetype in southern DRC (orange), and a more gently seasonal *arabiensis* and *funestus*-dominated archetype across northern Namibia and Botswana into southern Zimbabwe and Mozambique (plum). Archetypes are least contiguous along the continental divide, with Ethiopia, Uganda, and much of western Kenya showing a mix of archetypes. This effect is likely driven by the complex and highly varied topography of the Great Rift Valley,

which introduces sharper gradients of temperature, precipitation, and mosquito species than more smoothly-varying landscapes elsewhere on the continent. The horn of Africa is classified into two bimodal archetypes dominated by *arabiensis*, one with a substantial fraction of *funestus* (maroon) and one without (blue). Zambia, northern Mozambique, southern Tanzania, and most of Madagascar comprise a strongly seasonal archetype with a near-even mix of all three vectors (dark green). A broad area of central Tanzania is associated with the lower-rainfall, *arabiensis*-dominated southern band rather than any of the archetypes it borders (plum). The lime green archetype in southern Africa is distinguished by having few to no mosquitoes, and is excluded from simulation analysis. With the exception of the two representative sites in the horn of Africa, which display considerably higher rainfall than their archetypes' median values, representative site time series line up well with the time series of the archetype medians. These two clusters have fewer members and higher variability than many of the others, which may explain this effect. Table 2 describes the geography and mosquito characteristics of each representative site.

Simulation outputs and intervention maps

Figure 7 shows example impact curves for each archetype under different intervention scenarios, as well as their translation into maps of predicted impact. The impact curves show pre-intervention malaria prevalence on the x-axis and malaria prevalence after three years of interventions on the y-axis. The line of equivalence, shown in grey, indicates no change between initial and final prevalence, while the colored curves show simulation results across the range of transmission intensities. More effective intervention packages have curves that swing farther toward the lower right corner of the plots.

The first scenario, upper left map and dot-dashed lines, does not have interventions and therefore recreates the baseline 2000 prevalence map. The second, upper right map and dotted lines, shows a complex intervention mix of annual ITN distributions with moderate coverage (40%), annual IRS campaigns with 20% coverage, and low access and use of artemether-lumefantrine case management (AL CM), with 20% of clinical cases receiving treatment. The third, lower left map and solid lines, shows three years of high antimalarial access and use (80% of clinical cases receiving treatment), but no other interventions. The fourth, lower right map and dashed lines, shows a scenario in which the only intervention is a hypothetical attractive targeted sugar bait (ATSB) with a 3% mortality rate. For more intervention details see Table 3.

In all cases, initial prevalence is the primary driver of intervention impact, but the importance of accounting for archetype is highlighted in comparing scenarios. For example, the upper and lower right maps generally produce similar results despite consisting of non-overlapping interventions. However, difference between the two are clear especially in the plum-colored archetype, in which Angola and southern Zambia have different elimination outcomes between the two exercises and southern Mozambique performs better under an ATSB approach than a complex intervention mix. The antimalarials-only map differs considerably from both of the other intervention packages. In particular, it is more effective than either of the others at high baseline prevalence levels, but often comparable or less effective at low baseline prevalence. This generates a map with a similar elimination landscape, but many fewer hotspots, than the other two. Plots and maps of all 152 interventions are available under the "10-Site Setting" label at https://institutefordiseasemodeling.github.io/archetypes-intervention-impact-idmtools/.

Each archetype, because of its unique mix of vectors, possesses a different percentage of mosquito bites that occur indoors as opposed to outdoors. It is important

Table 2 EMOD Site descriptions

Site #	Description	*arabiensis* Proportion	*funestus* Proportion	*gambiae* Proportion	Anthropophily (%)	Endophily (%)	Total Indoor Biting (%)
1	Lime, Southern Africa	0	0	0	0	0	0
2	Teal, Sahel	0.45	0.55	0	65	70	45
3	Purple, northern DRC and coastal west Africa	0	0.01	0.99	85	85	72
4	Gold, west Africa	0.07	0.28	0.65	78	82	64
5	Blue, horn of Africa	1	0	0	65	50	32
6	Magenta, Saharan border	0.84	0.16	0	65	56	36
7	Plum, southern Africa and central Tanzania	0.71	0.29	0	65	60	39
8	Orange, southern DRC to Gabon	0	0.02	0.98	85	85	72
9	Maroon, northern Somalia	0.64	0.36	0	65	62	41
10	Green, Zambia, Mozambique, Malawi, and Madagascar	0.22	0.39	0.39	73	78	56

Table 3 EMOD intervention descriptions

Intervention	Details
Insecticide-treated nets (ITNs)	**Distributed annually on January 1st;** Initial blocking 90%, decaying exponentially with a 2-year half-life; Initial killing 60%, decaying exponentially with a 4-year half-life; Nets are discarded with exponential decay and a 6-month half-life; A randomly-selected 10% of the population never receives a net; Everyone who owns a net uses it every night.
Indoor residual spraying (IRS)	**Distributed annually on January 1st;** Initial killing 60%, decaying exponentially with a 6-month half-life; No correlation between IRS and ITN coverage.
Artemether-Lumefantrine (AL) case management	**Constant throughout simulation;** "Coverage" refers to the probability of seeking care given a clinical case; Upon deciding to seek care, per-day probability of care-seeking is 0.15; No age dependence; Drugs clear infection and provide about two weeks of additional protection.
Dihydro-artemisinin-Piperaquine (DP) case management	**Constant throughout simulation;** Same as AL case management, but with 1 month of protection after use.
Pre-Erythrocytic vaccine (PEV)	**Constant throughout simulation;** 90% acquisition blocking, decaying exponentially; 6- and 12- month decay half-lives; Distributed to infants upon reaching 6 months of age.
Transmission-blocking vaccine (TBV)	**Distributed annually on January 1st;** 90% transmission blocking, decaying exponentially; 6- and 12- month decay half-lives; Distributed to adults age 15-49.
Attractive targeted sugar baits (ATSB)	**Distributed twice per year on January 1st and July 1st;** Reported in terms of per-feed mortality rate, not coverage; 6-month box duration of efficacy.
Monoclonal antibodies (mABs)	**Distributed annually on January 1st;** Simulated via a PEV (see parameters above); Distributed to adults age 15-49; 3-month box duration of protection.

for this model framework to capture the differential impact of indoor-biting-targeted interventions in settings with different indoor biting rates. Figure 8 demonstrates this capability. Here, each line represents an archetype, colored according to its indoor biting percentage as determined by its mosquito species mix. The left panel shows an intervention setting with high coverage of insecticide-treated nets and indoor residual spraying, both of which target mosquitoes indoors. This intervention package, as expected, is much more effective in areas with a higher percentage of bites occurring indoors. The right panel demonstrates the inverse property: under an intervention package that includes only anti-malarial drugs and therefore does not directly target mosquitoes

(See figure on next page.)

Fig. 7 Example intervention impact results for four of the 152 interventions tested. The line plots (top) show intervention impact curves disaggregated by archetype, with archetypes colored according to the map (top right). The four prevalence maps (bottom) show hypothetical *Plasmodium falciparum* parasite rate among children aged 2–10 ($PfPR_{2-10}$) for each scenario. These maps were created by starting from a baseline map of 2000 as a proxy for a landscape without malaria interventions. For each initial prevalence value in the baseline map, a new prevalence value was found using the lines in the top set of plots. The linetype of subplot borders in the lower plots matches the linetype of interventions in the upper plots. For more detail about interventions see Table 3. ITN: Insecticide-treated net; IRS: Indoor residual spraying; AL CM: Artemether-Lumefantrine case management; ATSB: Attractive targeted sugar bait

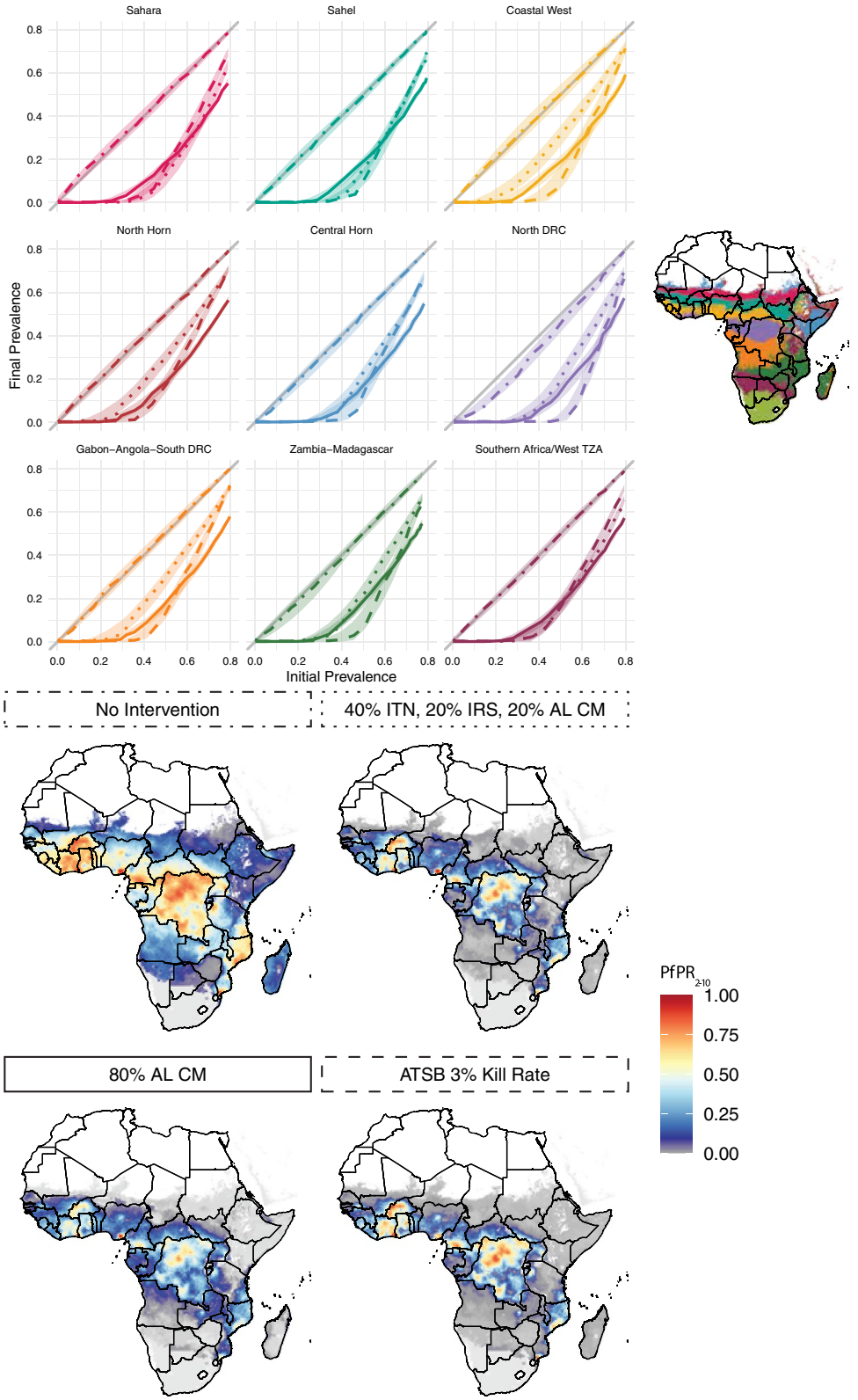

Fig. 7 (See legend on previous page.)

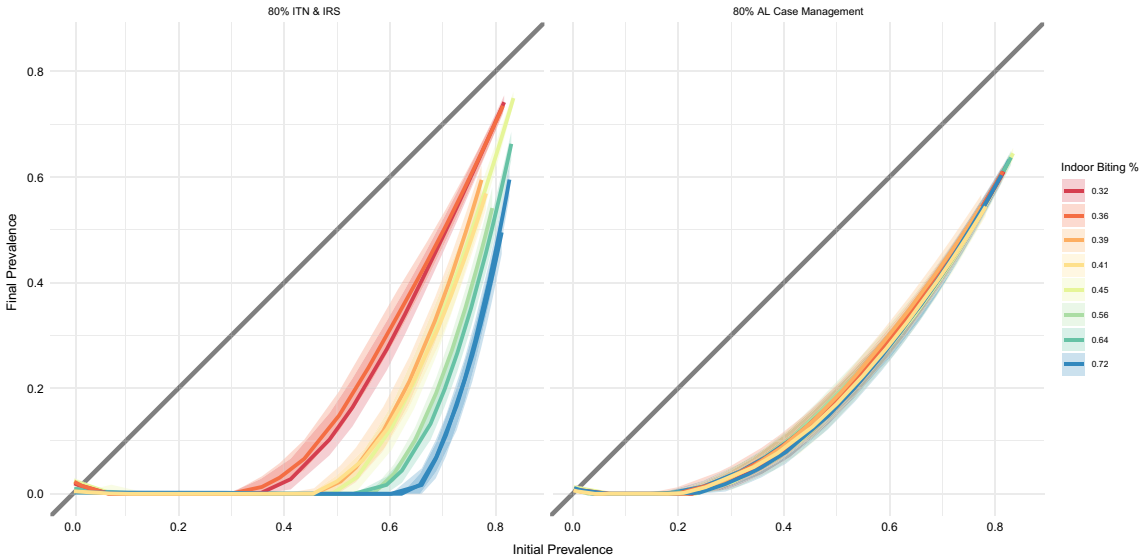

Fig. 8 Sensitivity of sites to biting intensity for vector control vs non-vector-control interventions. In both plots, each line represents an archetype's representative site, colored according to its indoor biting percentage as determined by its mosquito species mix. The left panel shows an intervention setting with 80% coverage of insecticide-treated nets and indoor residual spraying, both of which target mosquitoes indoors. This intervention package, as expected, is much more effective in areas with a higher percentage of bites occurring indoors. The right panel demonstrates the inverse property: under an intervention package that includes only anti-malarial drugs and therefore does not directly target mosquitoes at all, there is no differential impact across indoor biting intensities

at all, there is no differential impact across indoor biting intensities. Table 2 describes the mosquito characteristics of each representative site.

Site selection sensitivity analysis

Figure 9 shows the location of all 100 sensitivity sites. With one exception, the sensitivity analysis showed that the representative sites appropriately capture variation within archetypes. Figure 10 shows an example intervention package, with the impact curve of each archetype's representative site plotted in color and the curves of the ten randomly-selected sites shown in black. Shaded areas represent variation across the ten random seeds run for each site. If the representative sites were effective proxies for their archetypes, the colored lines and shaded areas would cover all or most of the black lines and shaded areas. While there are a few archetypes in which the representative sites' variation is slightly too narrow, especially the North Horn and Coastal West curves, in general the representative sites capture both the shape and the variation across the sensitivity sites. The one exception is the Sahelian archetype, in which the representative site curve overestimates intervention impact at moderate transmission intensities compared to the sensitivity analysis sites. The representative site for this archetype is in southern Burkina Faso, while the sensitivity

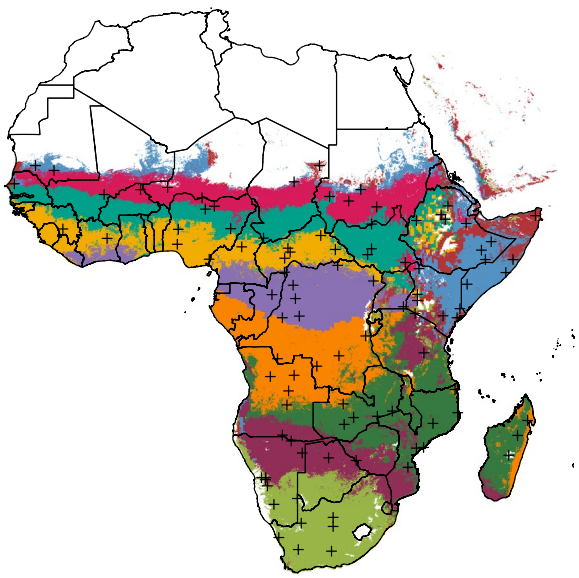

Fig. 9 Sensitivity analysis sites. Black crosses indicate the location of all 100 reqpresentative sites used for sensistivity analysis

sites are almost all situated farther east in Ethiopia, South Sudan, and northern Nigeria– perhaps too heterogeneous a landscape to be accurately captured by a single site. Future work may consider limiting the longitudinal

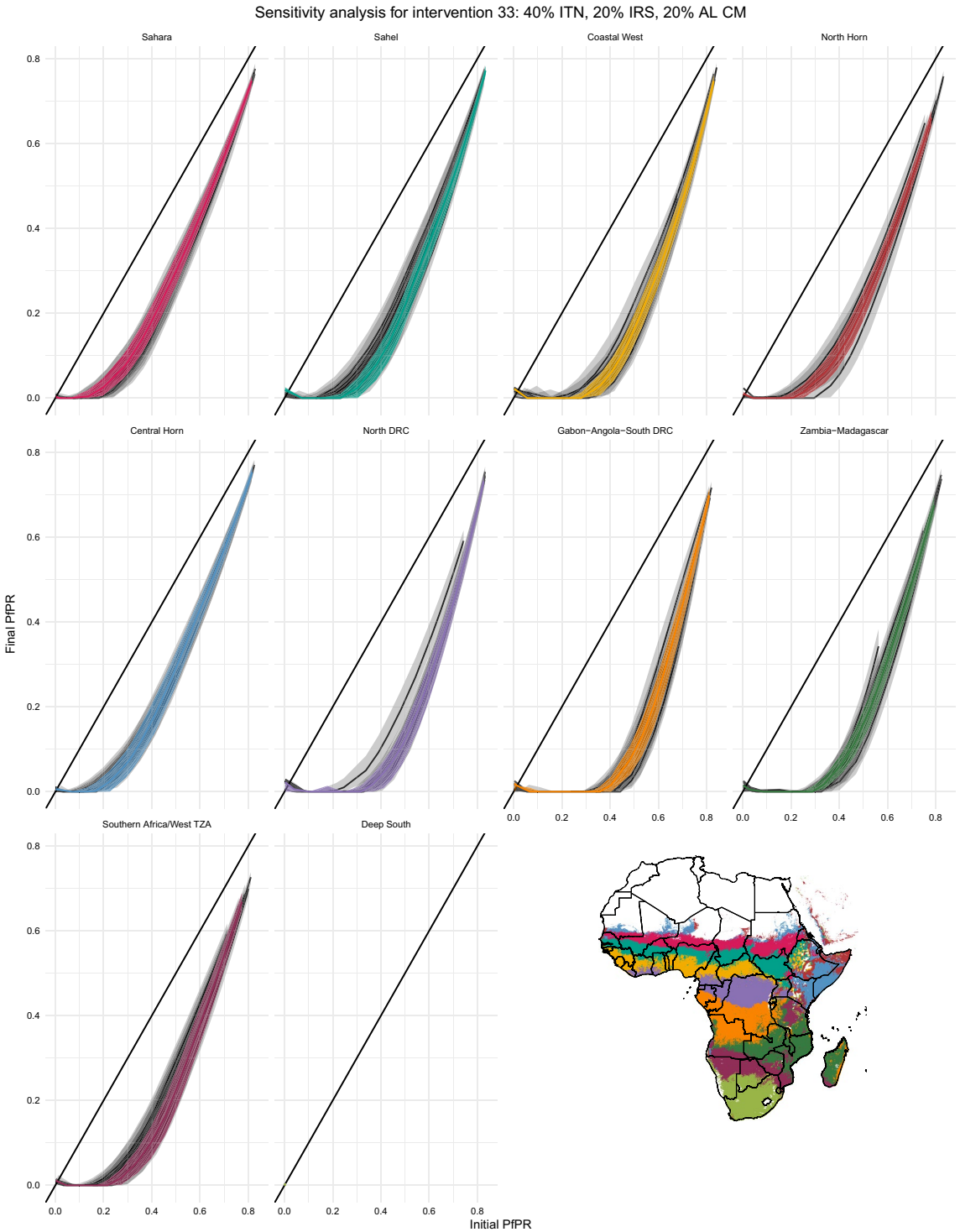

Fig. 10 Example sensitivity analysis results. Sensitivity analysis example intervention curves for an intervention package of 40% ITN coverage, 20% IRS coverage, and 20% coverage with artemether-lumefantrine case management. The impact curve of each archetype's representative site is plotted in color and the curves of the ten randomly-selected sites are shown in black. Shaded areas represent variation across the ten random seeds run for each site. If the representative sites were effective proxies for their archetypes, the colored lines and shaded areas would cover all or most of the black lines and shaded areas

or latitudinal extent allowed by a single archetype to improve the accuracy of the framework. Similar plots for all 152 intervention packages are available under the "Sensitivity Analysis" label at https://institutefordiseasemodeling.github.io/archetypes-intervention-impact-idmtools/.

Discussion

Mechanistic models are versatile tools for malaria intervention planning, but deploying them over large geographic areas at a useful spatial resolution poses both theoretical and computational challenges. This paper introduces a flexible and customizable archetypes-based framework that harnesses the richness of available spatial data and spatiotemporal modeling products to assess potential intervention impact across a range of settings, and discusses key decision points to consider when constructing such a framework. It includes a detailed example of how the framework might be configured. This example configuration demonstrates that dimensionality reduction and clustering can identify meaningfully different environmental archetypes, that the mechanistic model behaves as expected in relation to intervention effect and mosquito bionomics, that the creation of new maps showing intervention impact can highlight interventions and areas of interest, and that the representative sites selected can appropriately represent each archetype as a whole. The demonstrative analysis used the EMOD mechanistic model and malaria risk surfaces from the Malaria Atlas Project, but this methodology could just as easily be applied with different modeling software or malaria risk estimates. The choice of covariates upon which to cluster is likewise arbitrarily modifiable, noting that the mechanistic model used should have the capacity to accurately reflect any covariate variation that distinguishes archetypes from one another.

An archetypes-based approach can be useful for any project in which there is interest in the differential effects of a disease model under a range of initial conditions. Historically in malaria, such analyses either used idealized "characteristic" settings that did not directly represent real-world locations [18], or tested an extremely wide combinatoric collection of possible initial conditions [11]. The first approach is simple and straightforward, but suffers from arbitrary selection and the inability to claim that the sites shown are actually representative of other areas. The second approach, while thorough, requires an unnecessary computational burden. An archetypes-based strategy provides a useful middle ground, providing a relatively streamlined computational infrastructure for projects that require national or continental coverage, while also offering a useful suite of example settings that can easily be used for any type of exploratory simulation analysis.

The example configuration shows the utility of an archetypes framework for scenario-based analyses that are not calibrated to field data, but the framework has been extended to more data-rich use cases. In the recent High Burden to High Impact (HBHI) initiative, ten high-burden countries developed stratified intervention packages to effectively distribute malaria prevention and treatment to their communities. In Burkina Faso and Nigeria, researchers utilized a covariate-based dimensionality-reduction and clustering strategy to generate subnational archetypes, which then helped inform a rigorous admin-level calibration and stratification process to recommend targeted intervention strategies in upcoming years in these countries [34, 35]. In addition, it has repeatedly proven useful to have a diverse range of archetypal models pre-configured when stakeholders ask policy questions that are not geographically specific but may have different answers in different settings. While identifying different malaria seasonality modalities was not the primary goal of this work, the archetype selection process consistently identified spatial groupings with similar seasonality to heuristically identified seasonality profiles [11, 36]. This observation lends confidence to the effectiveness of both the archetypes framework and other seasonality detection methods.

An archetypes-based approach to intervention impact planning has several limitations. When utilized without fitting to field data, as shown in this document, this method is useful as an intuition-building tool, but should not be used to inform specific decisions in particular geographic locations. While the ability to project results from the analysis to an arbitrarily fine spatial scale may be useful for highlighting heterogeneities, such results might convey an unintended sense of confidence in the sensitivity and specificity of results. It is best to consider this framework as a strategy for obtaining informative priors, rather than as a way of generating quantitiatively rigorous model results. However, as described above in the HBHI framework, these concerns fade when the strategy is utilized as a precursor to more formal model fitting.

The example configuration specifically has additional limitations. While the selected covariates cover many environmental model sensitivities, no human-centered or malaria intervention history covariates were included. While some covariate standardization was performed, a

limited number of rescaling or standardizing approaches were conducted. Similarly, while common heuristics were utilized to select singular vector and cluster counts in SVD and k-means, a full and formal sensitivity analysis of these heuristics has yet to be conducted. These would include running k-means with more or fewer singular vectors, and repeating the clustering analysis with k-fold data holdouts. Both of these tests would check how cluster membership changes in response to these varied initial conditions, especially for pixels on the borders between archetypes. However, the robustness of cluster assignment between different k-values provides some confidence that such sensitivity checks would support the choices made in this analysis. While the EMOD modeling software has been extensively vetted and tested, it has not undergone a parameter-by-parameter sensitivity analysis to conclusively determine which variables are most robust to model changes. Because the example configuration focused on hypothetical scenarios rather than prediction or projection, uncertainty was not considered. A full uncertainty propagation strategy would begin with covariate uncertainty and reflect both mechanistic model parameter uncertainty and uncertainty in the prevalence maps used as baseline metrics.

The example configuration additionally does not consider a number of important malaria-related factors, but could easily be extended to take them into account. These include insecticide resistance, human migration, historical interventions, and site-specific population demographics. Future work will also include model comparison exercises between the archetypes framework and models that have been finely calibrated to specific locations to assess what improvements can be made in archetype-level predictions (while continuing to acknowledge that superseding site-specific analysis is not and will not be the goal). The examples shown here are in Africa, but this framework can also be extended to malaria-endemic regions of Asia and the Americas.

This paper presents a novel archetypes-based strategy for high-resolution, large-area malaria intervention impact assessment. Compared to similar approaches, it adds speed, computational efficiency, and interpretability to results, lends itself well to more detailed calibration-based approaches, and guides intuition in data-sparse settings. It has already proved a useful tool in the malaria modeling repertoire, and the authors look forward to further expanding its utility to new use cases and stakeholder needs.

Appendix

Details on the EMOD software

EMOD: an overview

EMOD (https://idmod.org/documentation) is an individual-based stochastic mechanistic modeling software developed by the Institute for Disease Modeling. While EMOD is capable of modeling a number of infectious diseases, only the malaria framework is described in detail here.

Operating on a daily time step, EMOD simultaneously tracks multiple simulation layers, including but not limited to: mosquito habitat and life cycle, mosquito behavior and movement, human population demographics, human movement, human immune responses to malaria, and malaria interventions that disrupt any of the aforementioned layers. Typically, humans are modeled as individuals while mosquitoes are modeled as cohorts, though there are options to model mosquitoes individually for extremely small-scale simulations. As an individual-based model, most input parameters are defined via probability distributions which are sampled at the appropriate points. Spatially distinct communities can be defined, with or without movement between them. While some heterogeneity in risk and behavior can be imposed within communities, under most circumstances a community behaves as a well-mixed population.

Mosquito habitat, life cycle, behavior, and movement In EMOD, species-specific mosquito population size is driven by the size of larval habitat available. In the absence of interventions, mosquito abundance in turn drives community-level malaria prevalence. Larval habitat size over time can be specified manually, if entomological data is available to inform this estimate, or can be driven by climate files indicating daily air temperature, rainfall, and relative humidity in the site being modeled. If the climate-driven option is selected, a decision must also be made about the types of larval habitat available in the model, allowing for hydrology-based conversions between climate and mosquito emergence and mortality. For more detail see Eckhoff [14]. Species-specific anthropophily and endophily values can also be specified. In the cohort-based model, mosquitoes do not travel from one community to another. See Table 1 for a list of vector parameters utilized in the example configuration.

Human population demographics Humans are modeled individually in EMOD, with each person assigned an age,

sex, and "home" community at model initialization. Age distributions and birth and death rates can be specified at model initiation. In addition, individuals can be assigned custom properties related to disease risk or exposure to better capture heterogeneity within communities. In the example configuration, 10% of individuals were assigned to a custom "No Nets" category, to indicate that they would always be skipped during insecticide-treated bednet campaigns, and each individual was assigned a unique biting risk such that the distribution of biting risk in the community overall was exponential [32]. While various types of human movement between communities can be specified in the model, this analysis did not allow for human movement to keep the effects of any one archetypal site clear.

Human immune responses to malaria and accounting for malaria population immunity Every time a person is infected with malaria in EMOD, they trigger a full intrahost infection cascade. Among the variables tracked are parasitemia, gametocytemia, and var-gene expression. After infection, waning immunity is also tracked and taken into account upon reinfection [37]. Individuals may remain asymptomatic, or develop a mild, clinical, or severe case of malaria. Due to the uncertain pathway from severe malaria to death, malaria-specific mortality was not enabled in this anlaysis.

Upon model initiation, all individuals are immunologically naive to malaria (though a few individuals are initialized as having an active infection). To generate a realistic immune profile, EMOD is typically run with no interventions for several decades, at which point the simulation state can be "frozen". From this starting point, a range of different intervention scenarios can be initialized.

Malaria interventions EMOD offers a wide range of malaria interventions, in which the user can specify intervention timing, coverage, targeting, efficacy, and waning, among many other variables. Interventions function by altering default model probabilities– for example, giving an individual an insecticide-treated net would reduce their probability of being bitten by a mosquito indoors, while receiving antimalarial drugs after infection would alter intrahost immune parameters which, in turn, would lower the likelihood of a mosquito acquiring gametocytes and increase the rate of recovery from clinical disease.

Table 3 presents brief descriptions of the malaria interventions used in this analysis. See the EMOD documentation for more.

Table of tested interventions

See Table 4.

Table 4 Intervention descriptions and populations at risk (PAR) across sub-Saharan Africa after three years of each intervention package

Intervention descriptions and population at risk		
Int #	**Description**	**PAR (Millions)**
0	Baseline (pre-intervention)	1374
1	0/0/0% ITN/IRS/AL CM	1372
2	20/0/0% ITN/IRS/AL CM	1306
3	40/0/0% ITN/IRS/AL CM	1236
4	60/0/0% ITN/IRS/AL CM	1014
5	80/0/0% ITN/IRS/AL CM	831
6	0/20/0% ITN/IRS/AL CM	1308
7	20/20/0% ITN/IRS/AL CM	1183
8	40/20/0% ITN/IRS/AL CM	964
9	60/20/0% ITN/IRS/AL CM	826
10	80/20/0% ITN/IRS/AL CM	654
11	0/40/0% ITN/IRS/AL CM	1093
12	20/40/0% ITN/IRS/AL CM	915
13	40/40/0% ITN/IRS/AL CM	771
14	60/40/0% ITN/IRS/AL CM	606
15	80/40/0% ITN/IRS/AL CM	479
16	0/60/0% ITN/IRS/AL CM	870
17	20/60/0% ITN/IRS/AL CM	713
18	40/60/0% ITN/IRS/AL CM	619
19	60/60/0% ITN/IRS/AL CM	472
20	80/60/0% ITN/IRS/AL CM	334
21	0/80/0% ITN/IRS/AL CM	600
22	20/80/0% ITN/IRS/AL CM	504
23	40/80/0% ITN/IRS/AL CM	436
24	60/80/0% ITN/IRS/AL CM	326
25	80/80/0% ITN/IRS/AL CM	234
26	0/0/20% ITN/IRS/AL CM	1316
27	20/0/20% ITN/IRS/AL CM	1253
28	40/0/20% ITN/IRS/AL CM	1065
29	60/0/20% ITN/IRS/AL CM	943
30	80/0/20% ITN/IRS/AL CM	695
31	0/20/20% ITN/IRS/AL CM	1212
32	20/20/20% ITN/IRS/AL CM	995
33	40/20/20% ITN/IRS/AL CM	834
34	60/20/20% ITN/IRS/AL CM	691
35	80/20/20% ITN/IRS/AL CM	558
36	0/40/20% ITN/IRS/AL CM	928
37	20/40/20% ITN/IRS/AL CM	823
38	40/40/20% ITN/IRS/AL CM	635
39	60/40/20% ITN/IRS/AL CM	492
40	80/40/20% ITN/IRS/AL CM	403
41	0/60/20% ITN/IRS/AL CM	692
42	20/60/20% ITN/IRS/AL CM	584
43	40/60/20% ITN/IRS/AL CM	471
44	60/60/20% ITN/IRS/AL CM	367
45	80/60/20% ITN/IRS/AL CM	259

Table 4 (continued)

Intervention descriptions and population at risk		
Int #	**Description**	**PAR (Millions)**
46	0/80/20% ITN/IRS/AL CM	506
47	20/80/20% ITN/IRS/AL CM	386
48	40/80/20% ITN/IRS/AL CM	299
49	60/80/20% ITN/IRS/AL CM	240
50	80/80/20% ITN/IRS/AL CM	206
51	0/0/40% ITN/IRS/AL CM	1243
52	20/0/40% ITN/IRS/AL CM	1096
53	40/0/40% ITN/IRS/AL CM	905
54	60/0/40% ITN/IRS/AL CM	770
55	80/0/40% ITN/IRS/AL CM	566
56	0/20/40% ITN/IRS/AL CM	1052
57	20/20/40% ITN/IRS/AL CM	891
58	40/20/40% ITN/IRS/AL CM	733
59	60/20/40% ITN/IRS/AL CM	591
60	80/20/40% ITN/IRS/AL CM	416
61	0/40/40% ITN/IRS/AL CM	823
62	20/40/40% ITN/IRS/AL CM	655
63	40/40/40% ITN/IRS/AL CM	536
64	60/40/40% ITN/IRS/AL CM	396
65	80/40/40% ITN/IRS/AL CM	274
66	0/60/40% ITN/IRS/AL CM	592
67	20/60/40% ITN/IRS/AL CM	480
68	40/60/40% ITN/IRS/AL CM	348
69	60/60/40% ITN/IRS/AL CM	296
70	80/60/40% ITN/IRS/AL CM	228
71	0/80/40% ITN/IRS/AL CM	394
72	20/80/40% ITN/IRS/AL CM	280
73	40/80/40% ITN/IRS/AL CM	259
74	60/80/40% ITN/IRS/AL CM	241
75	80/80/40% ITN/IRS/AL CM	196
76	0/0/60% ITN/IRS/AL CM	1103
77	20/0/60% ITN/IRS/AL CM	932
78	40/0/60% ITN/IRS/AL CM	795
79	60/0/60% ITN/IRS/AL CM	650
81	0/20/60% ITN/IRS/AL CM	896
82	20/20/60% ITN/IRS/AL CM	718
83	40/20/60% ITN/IRS/AL CM	575
84	60/20/60% ITN/IRS/AL CM	432
85	80/20/60% ITN/IRS/AL CM	335
86	0/40/60% ITN/IRS/AL CM	684
87	20/40/60% ITN/IRS/AL CM	549
88	40/40/60% ITN/IRS/AL CM	405
89	60/40/60% ITN/IRS/AL CM	297
90	80/40/60% ITN/IRS/AL CM	234
91	0/60/60% ITN/IRS/AL CM	449
92	20/60/60% ITN/IRS/AL CM	377
93	40/60/60% ITN/IRS/AL CM	275
94	60/60/60% ITN/IRS/AL CM	216

Table 4 (continued)

Intervention descriptions and population at risk		
Int #	**Description**	**PAR (Millions)**
95	80/60/60% ITN/IRS/AL CM	182
96	0/80/60% ITN/IRS/AL CM	306
97	20/80/60% ITN/IRS/AL CM	276
98	40/80/60% ITN/IRS/AL CM	224
99	60/80/60% ITN/IRS/AL CM	201
100	80/80/60% ITN/IRS/AL CM	160
101	0/0/80% ITN/IRS/AL CM	925
102	20/0/80% ITN/IRS/AL CM	813
103	40/0/80% ITN/IRS/AL CM	613
104	60/0/80% ITN/IRS/AL CM	470
105	80/0/80% ITN/IRS/AL CM	332
106	0/20/80% ITN/IRS/AL CM	727
107	20/20/80% ITN/IRS/AL CM	567
108	40/20/80% ITN/IRS/AL CM	464
109	60/20/80% ITN/IRS/AL CM	314
110	80/20/80% ITN/IRS/AL CM	237
111	0/40/80% ITN/IRS/AL CM	522
112	20/40/80% ITN/IRS/AL CM	403
113	40/40/80% ITN/IRS/AL CM	320
114	60/40/80% ITN/IRS/AL CM	242
115	80/40/80% ITN/IRS/AL CM	204
116	0/60/80% ITN/IRS/AL CM	331
117	20/60/80% ITN/IRS/AL CM	263
118	40/60/80% ITN/IRS/AL CM	250
119	60/60/80% ITN/IRS/AL CM	183
120	80/60/80% ITN/IRS/AL CM	160
121	0/80/80% ITN/IRS/AL CM	227
122	20/80/80% ITN/IRS/AL CM	227
123	40/80/80% ITN/IRS/AL CM	178
124	60/80/80% ITN/IRS/AL CM	167
125	80/80/80% ITN/IRS/AL CM	118
126	80/80/80% ITN/IRS/AL CM + 40% mAb	131
127	80/80/80% ITN/IRS/AL CM + 40% TBV	117
128	80/80/80% ITN/IRS/AL CM + 40% PEV	120
129	80/80/80% ITN/IRS/AL CM + 80% mAb	103
130	80/80/80% ITN/IRS/AL CM + 80% TBV	98
131	80/80/80% ITN/IRS/AL CM + 80% PEV	146
132	80/80/80% ITN/IRS/DP CM	131
133	0/0/0% ITN/IRS/AL CM + 0.15% ATSB Kill Rate	1332
134	0/0/0% ITN/IRS/AL CM + 3% ATSB Kill Rate	767
135	0/0/0% ITN/IRS/AL CM + 15% ATSB Kill Rate	56
136	0/0/0% ITN/IRS/AL CM + 25% ATSB Kill Rate	11
137	20/20/20% ITN/IRS/AL CM + 0.15% ATSB Kill Rate	999
138	20/20/20% ITN/IRS/AL CM + 3% ATSB Kill Rate	343
139	20/20/20% ITN/IRS/AL CM + 15% ATSB Kill Rate	38
140	20/20/20% ITN/IRS/AL CM + 25% ATSB Kill Rate	0
141	40/40/40% ITN/IRS/AL CM + 0.15% ATSB Kill Rate	470
142	40/40/40% ITN/IRS/AL CM + 3% ATSB Kill Rate	165

Table 4 (continued)

Intervention descriptions and population at risk		
Int #	**Description**	**PAR (Millions)**
143	40/40/40% ITN/IRS/AL CM + 15% ATSB Kill Rate	1
144	40/40/40% ITN/IRS/AL CM + 25% ATSB Kill Rate	28
145	60/60/60% ITN/IRS/AL CM + 0.15% ATSB Kill Rate	238
146	60/60/60% ITN/IRS/AL CM + 3% ATSB Kill Rate	94
147	60/60/60% ITN/IRS/AL CM + 15% ATSB Kill Rate	26
148	60/60/60% ITN/IRS/AL CM + 25% ATSB Kill Rate	0
149	80/80/80% ITN/IRS/AL CM + 0.15% ATSB Kill Rate	105
150	80/80/80% ITN/IRS/AL CM + 3% ATSB Kill Rate	79
151	80/80/80% ITN/IRS/AL CM + 15% ATSB Kill Rate	0
152	80/80/80% ITN/IRS/AL CM + 25% ATSB Kill Rate	0

Abbreviations

ALCM	Artemether-Lumefantrine case management
ATSB	Attractive targeted sugar bait
DPCM	Dihydroartemisinin-Piperaquine case management
DRC	Democratic Republic of the Congo
GMM	Gaussian mixture model
HBHI	High burden to high impact
IDM	Institute for Disease Modeling
ITN	Insecticide-treated net
IQR	Interquartile range
IRS	Indoor residual spraying
MAP	Malaria Atlas Project
mAB	Monoclonal antibody
PEV	Pre-erythrocytic vaccine
$PfPR_{2-10}$	*Plasmodium falciparum* parasite rate among children aged 2–10
SVD	Singular value decomposition
TBV	Transmission-blocking vaccine
WHO	World Health Organization

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12936-023-04535-0.

Additional file 1. Full Cluster Maps. The plots below show maps and covariate summaries for cluster counts of three to 14. In the time series, solid colored lines represent the median across the archetype, shaded areas indicate the interquartile range, and dotted lines indicate the 95% variance interval. Solid black lines represent the climate values of the representative site for each archetype, also indicated as black crosses on the map. Doughnut plots show the relative vector abundance of the representative sites. Absent doughnuts indicate no mosquitoes in that site.

Acknowledgements
The authors would like to thank all members of the WHO Strategic Advisory Group for Malaria Elimination and the Lancet Commission on Malaria Eradication for their support, commentary, and feedback on this project.

Author contributions
ABV, SJB, and PWG conceived the study. ABV, JG, CAB, SJB, and PWG designed the models. ABV wrote and ran all analytical steps. JLP contributed to the design of the machine learning infrastructure. MW and DH wrote a tool to convert ERA5 data into model input files. ABV wrote the first draft of the manuscript. ABV, CAB, JG, JLP, MW, DH, TDH, and PWG interpreted the results, contributed to writing, and approved the final version for submission. All authors read and approved the final manuscritpt.

Funding
This publication is based on research conducted by the Malaria Atlas Project and funded in whole or in part by the Bill & Melinda Gates Foundation, including models and data analysis performed by the Institute for Disease Modeling at the Bill & Melinda Gates Foundation.
This work was supported, in whole or in part, by the Bill & Melinda Gates Foundation (OPP1197730). Under the grant conditions of the Foundation, a Creative Commons Attribution 4.0 Generic License has already been assigned to the Author Accepted Manuscript version that might arise from this submission. This work was also supported by the Telethon Trust, Western Australia.

Availability of data and materials
All code, and instructions for analysis replication including data downloads, are available at https://github.com/InstituteforDiseaseModeling/archetypes-intervention-impact-idmtools. An interactive Jupyter notebook allowing users to explore k-means clustering for different covariate sets described in this analysis is available at https://deepnote.com/workspace/idm-903a8509-b110-4d4d-93be-a752fefd0d6b/project/Malaria-Archetypes-Bertozzi-Villa-65d2dc5e-e433-4869-b897-5fd659a93879/notebook/arch-4ca7fee69255442e99836f2d085e6d52.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute for Disease Modeling, Bill & Melinda Gates Foundation, Seattle, USA. [2]Malaria Atlas Project, Telethon Kids Institute, Perth, Australia. [3]Big Data Institute, Nuffield Department of Medicine, Oxford University, Oxford, UK. [4]Department of Preventive Medicine and Institute for Global Health, Northwestern University, Chicago, USA. [5]MRC Centre for Global Infectious Disease Analysis, Department of Infectious Disease Epidemiology, Imperial College, London, UK. [6]Section of Epidemiology, Department of Public Health, University of Copenhagen, Copenhagen, Denmark. [7]Curtin University, Perth, Australia.

Received: 21 November 2022 Accepted: 15 March 2023
Published online: 26 April 2023

References

1. Packard RM. The making of a tropical disease: a short history of malaria. Baltimore: Johns Hopkins University Press; 2011.
2. Weiss DJ, Lucas TCD, Nguyen M, Nandi AK, Bisanzio D, Battle KE, et al. Mapping the global prevalence, incidence, and mortality of *Plasmodium falciparum*, 2000–17: a spatial and temporal modelling study. Lancet. 2019;394:322–31.
3. Gething PW, Elyazar IR, Moyes CL, Smith DL, Battle KE, Guerra CA, Patil AP, Tatem AJ, Howes RE, Myers MF, George DB. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. PLoS Negl Trop Dis. 2012;6: e1814.
4. Battle KE, Lucas TCD, Nguyen M, Howes RE, Nandi AK, Twohig KA, et al. Mapping the global endemicity and clinical burden of *Plasmodium vivax*, 2000–17: a spatial and temporal modelling study. Lancet. 2019;394:332–43.
5. Hay SI, Sinka ME, Okara RM, Kabaria CW, Mbithi PM, Tago CC, et al. Developing global maps of the dominant anopheles vectors of human malaria. PLoS Med. 2010;7: e1000209.
6. Sinka ME, Golding N, Massey NC, Wiebe A, Huang Z, Hay SI, et al. Modelling the relative abundance of the primary African vectors of malaria before and after the implementation of indoor, insecticide-based

vector control. Malaria J. 2016;15:142. https://doi.org/10.1186/s12936-016-1187-8.
7. Bhatt S, Weiss DJ, Mappin B, Dalrymple U, Cameron E, Bisanzio D, Smith DL, Moyes CL, Tatem AJ, Lynch M, Fergus CA. Coverage and system efficiencies of insecticide-treated nets in Africa from 2000 to 2017. eLife. 2015;4:e09672.
8. Bertozzi-Villa A, Bever CA, Koenker H, Weiss DJ, Vargas-Ruiz C, et al. Maps and metrics of insecticide-treated net access. use, and nets-per-capita in Africa from 2000–2020. Nat Commun. 2021;12:1–12.
9. Rathmes G, Rumisha S, Lucas T, Twohig K, Python A, Nguyen M, et al. Global estimation of anti-malarial drug effectiveness for the treatment of uncomplicated Plasmodium falciparum malaria 1991–2019. Malaria J. 2020;19:374.
10. Weiss DJ, Nelson A, Vargas-Ruiz CA, Gligorić K, Bavadekar S, Gabrilovich E, Bertozzi-Villa A, Rozier J, Gibson HS, Shekel T, Kamath C. Global maps of travel time to healthcare facilities. Nat Med. 2020;26:1–4.
11. Walker PGT, Griffin JT, Ferguson NM, Ghani AC. Estimating the most efficient allocation of interventions to achieve reductions in P. falciparum malaria burden and transmission in Africa. Lancet Glob Health. 2015. https://doi.org/10.1016/S2214-109X(16)30073-0.
12. Crowell V, Briët OJT, Hardy D, Chitnis N, Maire N, Pasquale AD, et al. Modelling the cost-effectiveness of mass screening and treatment for reducing Plasmodium falciparum malaria burden. Malaria J. 2013;12:4.
13. Stuckey EM, Smith TA, Chitnis N. Estimating malaria transmission through mathematical models. Trend Parasitol. 2013;29:477–82. https://doi.org/10.1016/j.pt.2013.08.001.
14. Eckhoff PA. A malaria transmission-directed model of mosquito life cycle and ecology. Malaria J. 2011;10:303.
15. Smith NR, Trauer JM, Gambhir M, Richards JS, Maude RJ, Keith JM, et al. Agent-based models of malaria transmission: a systematic review. Malaria J. 2018. https://doi.org/10.1186/s12936-018-2442-y.
16. Ton JF, Flaxman S, Sejdinovic D, Bhatt S. Spatial mapping with Gaussian processes and nonstationary Fourier features. Spat Stat. 2018;28:59–78.
17. Milton P, Coupland H, Giorgi E, Bhatt S. Spatial analysis made easy with linear regression and kernels. Epidemics. 2019;29: 100362.
18. Griffin JT, Hollingsworth TD, Okell LC, Churcher TS, White M, Hinsley W, et al. Reducing Plasmodium falciparum malaria transmission in Africa: a model-based evaluation of intervention strategies. PLOS Med. 2010;7: e1000324. https://doi.org/10.1371/journal.pmed.1000324.
19. Sonnewald M, Dutkiewicz S, Hill C, Forget G. Elucidating ecological complexity: unsupervised learning determines global marine eco-provinces. Sci Adv. 2020;6:eaay4740.
20. Proctor JL, Brunton SL, Brunton BW, Kutz JN. Exploiting sparsity and equation-free architectures in complex systems. Eur Phys J Spec Top. 2014;223:2665–84.
21. Sirovich L, Kirby M. Low-dimensional procedure for the characterization of human faces. JOSA A. 1987;4:519–24.
22. Feachem RGA, Chen I, Akbari O, Bertozzi-Villa A, Bhatt S, Binka F, et al. Malaria eradication within a generation: ambitious, achievable, and necessary. Lancet. 2019;394:1056–112.
23. Malaria eradication: benefits, future scenarios & feasibility. A report of the Strategic Advisory Group on Malaria Eradication. 2020. https://www.who.int/publications/i/item/9789240003675. Accessed 30 June 2021.
24. Eckart C, Young G. The approximation of one matrix by another of lower rank. Psychometrika. 1936;1:211–8. https://doi.org/10.1007/BF02288367.
25. Alter O, Brown PO, Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA. 2000;97:10101.
26. Xu D, Tian Y. A comprehensive survey of clustering algorithms. Ann Data Sci. 2015;2:165–93. https://doi.org/10.1007/s40745-015-0040-1.
27. Rodriguez MZ, Comin CH, Casanova D, Bruno OM, Amancio DR, da Costa LF, et al. Clustering algorithms: a comparative approach. PLoS ONE. 2019;14: e0210236.
28. MacQueen J. Some: methods for classification and analysis of multivariate observations. 1967. p. 281–298.
29. Smith T, Killeen G, Maire N, Ross A, Molineaux L, Tediosi F, et al. Mathematical modeling of the impact of malaria vaccines on the clinical epidemiology and natural history of *Plasmodium falciparum* malaria: Overview. Am J Trop Med Hyg. 2006;75:1–10.
30. Kiware SS, Chitnis N, Tatarsky A, Wu S, Castellanos HMS, Gosling R, et al. Attacking the mosquito on multiple fronts: insights from the vector control optimization model (VCOM) for malaria elimination. PLoS ONE. 2017;12: e0187680. https://doi.org/10.1371/journal.pone.0187680.
31. Penny MA, Verity R, Bever CA, Sauboin C, Galactionova K, Flasche S, et al. Public health impact and cost-effectiveness of the RTS, S/AS01 malaria vaccine: a systematic comparison of predictions from four mathematical models. Lancet. 2015;6736:1–9.
32. Guelbéogo WM, Gonçalves BP, Grignard L, Bradley J, Serme SS, Hellewell J, et al. Variation in natural exposure to anopheles mosquitoes and its effects on malaria transmission. eLife. 2018. https://doi.org/10.4269/ajtmh.2006.75.2_suppl.0750001.
33. Vaisala. HUMIDITY CONVERSION FORMULAS Calculation formulas for humidity; 2013. www.vaisala.com. Accessed 30 June 2021.
34. Organization WH, to End Malaria RP. WHO—High burden to high impact: a targeted malaria response. Geneva: World Health Organization; 2018.
35. WHO. World malaria report 2020. Geneva: World Health Organization; 2020.
36. Beck HE, Zimmermann NE, McVicar TR, Vergopolan N, Berg A, Wood EF. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci Data. 2018;5: 180214.
37. McCarthy KA, Wenger EA, Huynh GH, Eckhoff PA. Calibration of an intrahost malaria model and parameter ensemble evaluation of a pre-erythrocytic vaccine. Malaria J. 2015;14:6. https://doi.org/10.1186/1475-2875-14-6.

Publisher's Note