Spatial modelling for population replacement of mosquito vectors at continental scale

Nicholas J. Beeton[a,*], Andrew Wilkins[b,*], Adrien Ickowicz[a], Keith R. Hayes[a], Geoffrey R. Hosack[a]

[a]*CSIRO, 3 Castray Esplanade, Battery Point TAS 7004, Australia*
[b]*CSIRO, 1 Technology Court, Pullenvale QLD 4069, Australia*

Abstract

Malaria is one of the deadliest vector-borne diseases in the world. Researchers are developing new genetic and conventional vector control strategies to attempt to limit its burden. Novel control strategies require detailed safety assessment to ensure responsible and successful deployments. *Anopheles gambiae* sensu stricto (s.s.) and *Anopheles coluzzii*, two closely related subspecies within the species complex *Anopheles gambiae* sensu lato (s.l.), are among the dominant malaria vectors in sub-Saharan Africa. These two subspecies readily hybridise and compete in the wild and are also known to have distinct niches, each with spatially and temporally varying carrying capacities driven by precipitation and land use factors.

We model the spread and persistence of a population-modifying gene drive system in these subspecies across sub-Saharan Africa by simulating introductions of genetically modified mosquitoes across the African mainland and its offshore islands. We explore transmission of the gene drive between the two subspecies that arise from different hybridisation mechanisms, the effects of both local dispersal and potential wind-aided migration to the spread, and the development of resistance to the gene drive. Given the best current available knowledge on the subspecies' life histories, we find that an introduced gene drive system with typical characteristics can plausibly spread from even distant offshore islands to the African mainland with the aid of wind-driven migration, with resistance taking over within a decade. Our model accounts for regional to continental scale mechanisms, and demonstrates a range of realistic dynamics including the effect of prevailing wind on spread and spatio-temporally varying carrying capacities for subspecies. As a result, it is well-placed to answer future questions relating to mosquito gene drives as important life history parameters become better understood.

1 Introduction

Contemporary malaria control interventions – insecticide treated bed nets, indoor residual spraying and artemisinin based combination therapy – have dramatically reduced the burden of malaria in Africa (Bhatt et al., 2015). Since 2017, however, the rate of progress on malaria reduction has stalled and in 2019 malaria still claimed an estimated 389,000 African lives, mainly children under 5 years of age (World Health Organisation, 2020). At least 99% of these cases are caused by *Plasmodium falciparum*, transmitted by a small number of dominant malaria vectors, most notably *Anopheles arabiensis*, *Anopheles coluzzii*, *Anopheles gambiae* sensu stricto (s.s.) and *Anopheles funestus* (Wiebe et al., 2017).

The ongoing burden of malaria, together with increasing rates of insecticide resistance in malarial vector mosquitoes (Wiebe et al., 2017), has motivated proposals to develop new genetic control strategies, including: a) self-limiting, population suppression methods that induce male sterility (Windbichler et al., 2008; Klein et al., 2012) or male bias (Galizi et al., 2014; Facchinelli et al., 2019); b) self-sustaining (gene drive), suppression methods that induce female sterility (Hammond et al., 2016; Kyrou et al., 2018); and, c) self-sustaining, population replacement methods that make vectors refractory for the malaria parasite (Gantz et al., 2015; Carballar-Lejarazú et al., 2020). Any proposal to test these genetic control strategies outside of contained laboratory settings will likely require a detailed quantitative risk assessment that predicts the potential spread and persistence of transgenic mosquitoes from release sites, and the possible introgression of a transgenic construct into closely related species through interspecific mating (James et al., 2018, 2020). Spatial models of spread and persistence are also needed to describe the dynamics of important gene drive processes such as the development of resistance to the gene drive (Price et al., 2020). Quantitative spatial models have been developed for the spread and persistence of self-limiting, population suppressing constructs (Dufourd and Dumont, 2013; Facchinelli et al., 2019), together with self-sustaining, population-

*Corresponding author
Email addresses: nick.beeton@csiro.au (Nicholas J. Beeton), andy.wilkins@csiro.au (Andrew Wilkins)

modifying (Beaghton et al., 2017; Tanaka et al., 2017; Noble et al., 2018; Wyse et al., 2018) and population-suppressing (Eckhoff et al., 2016; North and Godfray, 2018; North et al., 2019) constructs.

This paper models the spread and persistence of a population-modifying gene drive system (Alphey et al., 2020; Nolan, 2021) in *Anopheles gambiae* s.s. and *Anopheles coluzzii* across sub-Saharan Africa. These two subspecies, which are often modelled as single group, are together with *Anopheles arabiensis* and *Anopheles funestus* the dominant vectors of malaria in sub-Saharan Africa. *An. gambiae* s.s. and *An. coluzzii* are both highly anthropophilic and efficient malaria vectors. The two subspecies are closely related enough to interbreed but hybridisation rates vary in space and time (Pombi et al., 2017). They also have different larval habitat preferences (Lehmann and Diabate, 2008), and *An. coluzzii* is thought to have a superior resistance to desiccation stress (Tene Fossog et al., 2015), hence is more drought tolerant than *An. gambiae* s.s. (Sinka et al., 2016). Although they are defined as two separate species by Coetzee et al. (2013), we refer to them here and subsequently as subspecies to emphasise the lack of reproductive isolation between the two taxa (see Pombi et al., 2017; Tene Fossog et al., 2015), which is a focus of our model.

In addition to general concerns for gene drives such as the development of resistance, the following ecological hypotheses proposed in the literature are investigated: transmission of the gene drive between two hybridising subspecies of *Anopheles gambiae* sensu lato (s.l.) by vertical gene transfer (Pombi et al., 2017; Beeton et al., 2020; Selvaraj et al., 2020); possible long range dispersal or long distance migration (North and Godfray, 2018; Huestis et al., 2019); and the nature of spatially and temporally varying carrying capacities driven by precipitation and land use factors (White et al., 2011; Tene Fossog et al., 2015; North and Godfray, 2018). This model is designed to provide scenario based testing of structural hypotheses that formalise the current state of knowledge for key gene drive and population life history parameters.

Each of these structural issues are described in the following subsections:

1.1 Choice of Construct

Our focus is on the spread and persistence of a gene drive system with near-neutral fitness that incorporates an unavoidable small reproductive payload cost of expression of the nuclease (see Hammond et al. (2016) and Beaghton et al. (2017)) through a spatially and temporally dynamic population with differential gene flow across sub-Saharan Africa. This scenario thereby evaluates the behaviour of an idealised nearly fitness-neutral population replacement gene drive system at the continental scale. Our analysis focusses on three alleles: the wild–type, the genetic construct for a population replacement gene drive and a resistant allele. Together these form a minimal gene drive spatial model (see Beaghton et al., 2017). Further, we assume that the gene drive is activated in a single locus in each parent's genetic code as in Beaghton et al. (2017). However, we avoid modelling a gene drive "payload" of a nuclease or effector gene as done in their paper. That is, the nuclease and effector gene may be considered as the same unit, with the effector gene either absent or nearly fitness neutral. In practice, an effector would also be likely to exhibit a genetic load on the receiving organism (e.g. Beaghton et al., 2017). Therefore the model predictions are deliberately optimistic in terms of the magnitude of spread and persistence of the construct, and provide an indication of the spread of a population replacement drive for an idealised effector with negligible fitness cost.

1.2 Taxonomic Resolution

We model an intervention where the genetic construct has been introgressed into locally sourced, wild-type *An. gambiae* s.s. or *An. coluzzii* mosquitoes, and subsequently released back into this local population. Depending on geographic location, these subspecies of the *An. gambiae* s.l. complex may introgress with each other and potentially other subspecies such as *An. arabiensis* (*Anopheles gambiae* 1000 Genomes Consortium, 2017; Pombi et al., 2017; Clarkson et al., 2020). Studies at the scale of sub-Saharan Africa often do not discriminate between *An. gambiae* s.s. and *An. coluzzii*. For example, Sinka et al. (2012) combined these two subspecies when plotting species distribution maps due to lack of data; *An. arabiensis*, however, was plotted separately. Similarly, North and Godfray (2018) argued that currently there is a lack of available data for parametrising alternative life history strategies of *An. gambiae* s.s. and *An. coluzzii*, and so did not discriminate between these subspecies in their process model. Indeed, we are currently unaware of any analysis of genetic vector control strategies at this continental scale, with explicit spatial and temporal dynamics, that discriminates between these two co-dominant malaria vectors.

We include alternative subspecies in our model because this leads to altered population dynamics via interspecific mating and density dependence effects (Beeton et al., 2020). Here we explore two different approaches to species assignment of first generation hybrids: 1) species assignment by maternal descent, and 2) equal proportions.

1.3 Larval carrying capacity

To parameterise a spatial model that discriminates between *An. gambiae* s.s. and *An. coluzzii*, and their inter– and intra–specific density dependence at the larval stage, we require spatially explicit carrying capacity information about each subspecies, which is anticipated to depend on environmental and social covariates. As observation data is relatively sparse at the subspecies level, we approach the problem in two parts. First, we model the

larval carrying capacity of the two subspecies taken together using a functional form. Second, we use empirical relative abundance data to spatially model the relative carrying capacities between subspecies.

1.3.1 Total abundance

The carrying capacity of *Anopheles* species in Africa is often expressed as a function of rainfall. For example, White et al. (2011) found exponentially weighting the past 4 days of rainfall gave the best fit when modelling the abundance of *An. gambiae* s.s. and *An. arabiensis* in Nigeria, an approach subsequently adopted by Wu et al. (2020); whilst Magombedze et al. (2018) used a 7 day moving average of rainfall to model the carrying capacity of the aquatic population of *An. gambiae* s.s., *An. coluzzii* and *An. arabiensis* in Mali.

More complex functional forms invoke additional parameters such as the location (and sometimes length or size) of perennial, intermittent, permanent or human-associated water bodies, as in the models developed by Eckhoff (2011) and Lunde et al. (2013). The most relevant approach for our purposes, however, is that of North and Godfray (2018), who group our two proposed subspecies *An. gambiae* s.s. and *An. coluzzii* in a spatially explicit, individual-based model, across an area of West Africa that exhibits significant environmental variation. They predict local larval carrying capacity based on rainfall, as well as level of access to temporary and permanent water courses. We adapt their results for our model of total carrying capacity for the aggregate of *An. coluzzii* and *An. gambiae* s.s.

1.3.2 Relative abundance

An. coluzzii has only relatively recently been described as its own subspecies (Coetzee et al., 2013) after its earlier description as a molecular form within *An. gambiae* s.s. (della Torre et al., 2001, 2002). Despite this taxonomic difficulty, several papers have examined differences in habitat preference between *An. gambiae* s.s. and *An. coluzzii*. In particular, Lehmann and Diabate (2008) note that larval predation and competition has led to selection for temporary freshwater habitats in *An. gambiae* s.s. and conversely permanent habitats for *An. coluzzii*. They suggest that this leads to humidity and/or rainfall clines in relative abundance. Other sources provide data suggesting this is the case for rainfall (Touré et al., 1998; Wondji et al., 2005; Diabaté et al., 2005, 2008), and that a particular chromosomal arrangement in *An. coluzzii* performs well in low-rainfall environments (Touré et al., 1994). These conclusions, however, tend to be based on relatively small-scale observations or experiments. Some information on relative abundance at larger scales is available (della Torre et al., 2005; Caputo et al., 2008; Simard et al., 2009; Costantini et al., 2009), but very little modelling has been done to quantify these differences across sub-Saharan Africa. So far only occurrence information has been generally used (Sinka et al., 2010). In contrast, the relative abundance of *An. gambiae* s.s. in its former definition (including *An. coluzzii*) versus *An. arabiensis* has long since been modelled and estimated across sub-Saharan Africa (Lindsay et al., 1998).

A notable exception in this context is Tene Fossog et al. (2015), who develop a logistic regression model using abundance data of each subspecies across western sub-Saharan Africa to predict relative probability of occurrence of the two subspecies. They use model selection to select a subset of relevant spatial, climatic and land cover variables in their predictions. However, despite acknowledging the likelihood of nonlinear effects of some variables (see Figure 3b in their paper), they use only linear predictors in a logistic regression. We use their work as a starting point to develop a flexible neural network model to incorporate nonlinear relationships, along with additional predictors and newly collated records of relative abundance (VectorBase: Giraldo-Calderón et al., 2015) to extend our predictions to include the rest of sub-Saharan Africa.

1.4 Dispersal

Anopheline mosquitoes have historically been categorised as being unlikely to migrate long distances (with mean dispersal distances typically less than 1 km, maximum distances typically no greater than 5 km). Although longer range dispersal events are possible and have been linked to mosquito-borne disease outbreaks, short range dispersal is supposed to predominate life history strategies (Gillies, 1961; Service, 1997). A recent empirical study (Huestis et al., 2019), however, provides evidence for wind-driven long-range dispersal of *An. gambiae* s.s. and *An. coluzzii* mosquitoes in large numbers. These mosquitoes remain capable of reproduction and pathogen transmission (Sanogo et al., 2020), and are estimated to regularly travel much further than even a rare long-range dispersal event could achieve. Moreover, it has been recently suggested that *An. gambiae* s.l. populations in areas of low human density may also facilitate migration, gene flow or both (Epopa et al., 2020).

1.5 Aestivation

Another source of controversy is aestivation, in which mosquitoes become dormant during dry conditions in which they would not otherwise survive. While not proven to occur widely on a population scale, it is a popular hypothesis for wet season reemergence of *An. coluzzii* in the Sahel (Dao et al., 2014; Lehmann et al., 2017). Modelling studies that address this problem include Magombedze et al. (2018) and North and Godfray (2018). The latter simulation study notes that rare persistent water sources provide a competing explanation for persistence through the dry season, as does long distance migration. In this model, these latter proposed processes are accommodated by spatially and temporally varying carrying capacities (Section 1.3) and dispersal behaviour (Section 1.4); aestivation is not explicitly modelled.

1.6 Spatial Scope

The spatial scope for this analysis includes all countries within the African region as defined by the United Nations geoscheme that are within the range of *Anopheles gambiae* s.l. (United Nations regions are listed here: `https://unstats.un.org/unsd/methodology/m49/overview/`). The spatial scope includes the range of *Anopheles gambiae* s.l. on the African continent and also island countries or territories of the African region where *Anopheles gambiae* s.l. is present, such as Madagascar, Mauritius, Comoros and São Tomé and Príncipe. *Anopheles gambiae* s.l. is also found in Cabo Verde (DePina et al., 2018).

2 Methods

2.1 Spatio-temporal mosquito demographic model

We represent each combination of age class (a), sex (s), genotype (g) and subspecies of mosquito (m) as a separate scalar field $X_{a,s,g,m}(t, \mathbf{x})$ in a Partial Differential Equation (PDE) model with time t and 2D location $\mathbf{x}$. Mosquitoes are assumed to not persist in ocean regions, and we set a zero-population Dirichlet condition in those regions. We allow, however, for the possibility that mosquitoes can advect across the ocean to neighbouring islands for a maximum duration of one day.

We represent the model numerically by spatially discretising the mosquito population of sub-Saharan Africa into 5 km $\times$ 5 km grid cells using the Africa Albers Equal Area Conic projection (ESRI:102022), and temporally discretising in 1 day timesteps, using fourth-order Runge-Kutta to integrate between timesteps. This timestep was determined experimentally to be both numerically stable and accurate. Combined with the spatial resolution, the model can be run within a reasonable time frame, while still being suitable for modelling the relevant biological processes, with a mosquito home range being roughly one cell in size. Within a timestep, we model demographic processes, followed by diffusion (see Section 2.1.4), followed by advection (see Section 2.1.5). While our model is deterministic, we allow extinction to occur by making a continuous model correction: at the end of the demographic processes step in each timestep, any cell with less than one total mosquito in a subspecies is set to zero for all classes of that subspecies.

We define $N_{subspecies}$ subspecies, such that each subspecies is given an integer number from $m = 0$ to $m = N_{subspecies}$: in our main results, $N_{subspecies}$ is set to 2 to represent *An. gambiae* s.s. and *An. coluzzii* as $m = 0$ and 1 respectively. We also discretise age into N_{age} age classes in our model: in our main results, we use $N_{age} = 2$ (one newborn and one adult class), but we describe the general model for $N_{age} \geq 1$ here, as different values of N_{age} may be more relevant for different model applications. For $N_{age} = 2$ age classes, the constant transition rate b results in an exponential distribution of maturation times between larval classes. Some larvae will quickly mature, potentially making the population more resilient to intervals without rainfall. Although we focus on $N_{age} = 2$, Figure S2 shows that results for $N_{age} = 6$ closely match results for $N_{age} = 2$ in Figure S1, with resistance spreading slightly more slowly over time.

Ages vary from the "newborn" larvae age class $a = 0$ (into which all individuals are born) through to adults at $a = N_{age} - 1$, with $[1, N_{age} - 2]$ representing intermediate larval stages where these exist ($N_{age} > 2$). We here describe the PDE separately for these three stage types. Note that in our model we only track those mosquito eggs that produce viable larvae, hence our newborn class consists of larvae instead of eggs. Our numerical model is written in Cython (the Python programming language with C extensions Behnel et al., 2011) and visualisations are performed in the R programming language (R Core Team, 2020).

Variable	*Name*	*Description*	*Estimate*	*Source*
Demographics				
d_A	Adult mortality	Daily probability of mortality	$0.1\ \mathrm{d}^{-1}$	North and Godfray (2018) (or see Lunde et al. (2013), Eckhoff (2011), North et al. (2013), Wu et al. (2020))
d_J	Juvenile (larval) mortality	Daily probability of mortality	$0.05\ \mathrm{d}^{-1}$	North and Godfray (2018) (or see Lunde et al. (2013), Eckhoff (2011), North et al. (2013), Magombedze et al. (2018))
b	Larval transition rate	Number of days from an egg being laid to when it emerges as a sexually mature adult (when $N_{age} = 2$) or reaches next larval stage ($N_{age} > 2$)	$0.1\ \mathrm{d}^{-1}$	North and Godfray (2018) (or see Eckhoff (2011), North et al. (2013), Wu et al. (2020))
D	Diffusion coefficient	Typical rate of spread of population from a point source	$900\ \mathrm{m}^2\ \mathrm{d}^{-1}$	Lunde et al. (2013) (or see Ickowicz et al. (2021))
λ	Larvae per female	Expected number of larvae per female per day (wild type mosquitoes)	$9\ \mathrm{female}^{-1}\ \mathrm{d}^{-1}$	North and Godfray (2018) (or see Lunde et al. (2013), Eckhoff (2011), Wu et al. (2020))
k	Relative probability of mating between subspecies	The relative probability that a female has offspring with a male of a different subspecies to her own ($k < 1$)	0.01	Pombi et al. (2017) (or see Lee et al. (2013))
α_{ij}	Lotka-Volterra competition between subspecies	The relative effect on subspecies X of a member of a different subspecies Y taking up its resources (and thus larval carrying capacity), as compared to a conspecific	$\alpha_{11} = \alpha_{22} = 1$, $\alpha_{12} = \alpha_{21} = 0.4$	Pombi et al. (2017)
Genetics				
k_c	Probability of cleavage		0.995	Beaghton et al. (2017)
k_j	Probability of non-homologous repair		0.02	Beaghton et al. (2017)
k_n	Probability nuclease gene lost during homing		10^{-4}	Beaghton et al. (2017)
h_n	Dominance coefficient for nuclease expression		0.5	Beaghton et al. (2017)
s_n	Cost of nuclease expression		0.05	Beaghton et al. (2017)
Larval carrying capacity				
$K_m(\mathbf{x})$	Larval carrying capacity		Details in text (2.2) and below parameters	
$\alpha_0(\mathbf{x})$	Permanent larval site population		0 (no permanent sites)	North and Godfray (2018)
α_1	Contribution to breeding from rainfall		200,000	North and Godfray (2018)
α_2	Larval sites associated with rivers and lakes		200,000	North and Godfray (2018)
ϕ	Rate of carrying capacity population increase with rainfall		0.03 per mm rain per week	North and Godfray (2018)
κ	Rate of carrying capacity population increase with water bodies		0.8 per km water	North and Godfray (2018)
δ	Replenishment rate of intermittent water sites with rainfall		0.03 per mm rain per week	North and Godfray (2018)
Model details				
$X(0, \mathbf{x})$	Initial condition		Details in text (2.3.2)	
$A\,\mathbf{V}(t, \mathbf{x})$	Advection		Details in text (2.1.5,2.3.1)	Huestis et al. (2019)
t	Time domain for integration		2005-2015	

Table 1: Parameter definitions and estimates for final model ($N_{age} = 2$)

2.1.1 Adults

The PDE governing each scalar field representing adult populations $X_{N_{age}-1,s,g,m}(t,\mathbf{x})$ for $N_{age} > 1$ is given by

$$\frac{\partial X_{N_{age}-1,s,g,m}}{\partial t} = -d_A X_{N_{age}-1,s,g,m} + bX_{N_{age}-2,s,g,m} + \nabla \cdot (D\nabla X_{N_{age}-1,s,g,m} - A\mathbf{V}(t,\mathbf{x})X_{N_{age}-1,s,g,m})$$

where subscript s denotes sex (M or F), g genotype, and m mosquito subspecies ((1) *An. gambiae* s.s. and (2) *An. coluzzii*). We model three alleles: w for wild–type, c for construct (gene drive system), and r for resistant. This results in a set of six potential genotypes $G = \{ww, wc, wr, cc, cr, rr\}$. The vector field $\mathbf{V}(t,\mathbf{x})$ represents the wind experienced across the spatial domain at a specified time t and place $\mathbf{x}$. Note that our model makes some modifications to the advection process for biological reasons, specified below in Section 2.1.5. Other parameters are given in Table 1.

Here $N_{age}-1$ indicates the adult age bracket. If $N_{age} > 1$ then $N_{age} - 2$ is the eldest larvae age bracket; the special case of $N_{age} = 1$ is considered below (Section 2.1.3). The first term on the right-hand side ($-d_A X_{N_{age}-1,s,g,m}$) describes the mortality of adults, while the second term ($bX_{N_{age}-2,s,g,m}$) describes aging from the eldest larvae. The final term represents diffusion ($\nabla D \nabla X$) and advection ($-\nabla A\mathbf{V}(t,\mathbf{x})X_{N_{age}-1,s,g,m}$), with A being the probability of an adult mosquito being advected by wind.

2.1.2 Larvae

For $a \in [1, N_{age}-2]$, the populations are governed by

$$\frac{\partial X_{a,s,g,m}}{\partial t} = -d_J X_{a,s,g,m} + b\left[X_{a-1,s,g,m} - X_{a,s,g,m}\right] \tag{2.1}$$

As above, the first term on the right-hand side ($-d_J X_{a,s,g,m}$) describes the mortality of this age-bracket of larvae, while the second ($b\left[X_{a-1,s,g,m} - X_{a,s,g,m}\right]$) describes aging to/from older and younger age brackets respectively. Note that for $N_{age} \leq 2$ there are no such intermediate larvae.

2.1.3 Newborns and $N_{age} = 1$ model

The PDE governing $X_{0,s,g,m}(t,\mathbf{x})$ is given by

$$\frac{\partial X_{0,s,g,m}}{\partial t} = -d_J X_{0,s,g,m} - bX_{0,s,g,m} + B(s,g,m) \quad . \tag{2.2}$$

Where $N_{age} = 1$, there is no age structure and the full PDE takes the form

$$\frac{\partial X_{s,g,m}}{\partial t} = -d_A X_{s,g,m} + B(s,g,m) + \nabla \cdot (D\nabla X_{s,g,m} - A\mathbf{V}(t,\mathbf{x})X_{s,g,m}) \tag{2.3}$$

Note that for readability, we only use subscripts when referring to mosquito population classes (i.e. $X_{a,s,g,m}$) and refer to parameters which differ by population class as functions of the offspring classes. Note that we do not include parental classes (i.e. g_M, g_F, m_M, m_F) in the function names, also for brevity, and that these are only defined explicitly in the function definitions.

In Equation 2.2, the first term ($-d_J X_{0,s,g,m}$) describes mortality of newborns, while the second ($-bX_{0,s,g,m}$) describes aging into the youngest age-bracket of larvae (or adults for $N_{age} = 2$). The final term describes the birth of newborn larvae of a given sex s, genotype g and subspecies m, given all possibilities of the mother's and father's genotype and subspecies (g_M, g_F, m_M and m_F; described in more detail later). For brevity, we describe it as the product of functions describing the relevant biological mechanisms:

$$B(s,g,m) = \lambda \max\left(0,\ 1 - \frac{C(m)}{K_m(\mathbf{x})}\right) \times \sum_{\substack{g_M,g_F,\\ m_M,m_F}} J(m)\, O(s,g)\, R(g_F,g_M)\, X_{N_{age}-1,F,g_F,m_F} \tag{2.4}$$

where the baseline fecundity rate is λ, the expected number of larvae per clutch of eggs per female per day, assumed produced by a mating of wildtype mosquitoes. The J and O terms represent subspecies inheritance and genotype inheritance (including sex) respectively, and are described below. In our main results, we keep the relative fecundity function $R(g_M, g_F)$ constant at $R = 1$, but this can readily be varied to represent reduced fecundity due to inviability of a gene drive construct — see Supplementary Materials Figure S3 for an example.

The $\max\left(0,\ 1 - \frac{C(m)}{K_m(\mathbf{x})}\right)$ term models the effect of $K_m(\mathbf{x})$, the (spatially varying) larval carrying capacity for subspecies m, on the fecundity of that subspecies using a logistic function. The inclusion of a $\max$ term is to keep the model biologically plausible: without it, $C(m) > K_m(\mathbf{x})$ would mean that negative newborns are produced. Here $C(m)$ is the competition that a newborn of subspecies m experiences from all larval populations:

$$C(m) = \sum_{m'=0}^{N_{\text{subspecies}}-1} \alpha_{m,m'} \sum_{a=0}^{\max(0,N_{age}-2)} \sum_{s\in\{M,F\}} \sum_{g\in G} X_{a,s,g,m'} \tag{2.5}$$

where $\alpha_{m,m'}$ represents the effect of competition of subspecies m on subspecies m', and the $\max$ term is here used to ensure that where $N_{age} = 1$, the larval carrying capacity instead applies to the adult population.

The function $J(m)$ returns the probability that a male adult of genotype g_M and subspecies m_M successfully mates with a female adult of genotype g_F and subspecies m_F to produce newborns of subspecies m:

$$J(m) = \frac{H(m_M, m_F, m)\, W(g_M, m_M, m_F)}{\sum_{g'\in G} \sum_{m'=0}^{N_{\text{subspecies}}-1} W(g', m', m_F)} \quad , \tag{2.6}$$

		Female					
	Genotype	$\boldsymbol{wc}$	$\boldsymbol{ww}$	$\boldsymbol{wr}$	$\boldsymbol{cc}$	$\boldsymbol{cr}$	$\boldsymbol{rr}$
Male	$\boldsymbol{wc}$	$\boldsymbol{ww}(w^2)$, $\boldsymbol{wc}(2wc)$, $\boldsymbol{wr}(2wr)$, $\boldsymbol{cc}(c^2)$, $\boldsymbol{cr}(2cr)$, $\boldsymbol{rr}(r^2)$	*	*	*	*	*
	$\boldsymbol{ww}$	$\boldsymbol{ww}(w)$, $\boldsymbol{wc}(c)$, $\boldsymbol{wr}(r)$	$\boldsymbol{ww}(1)$	*	*	*	*
	$\boldsymbol{wr}$	$\boldsymbol{ww}\left(\frac{w}{2}\right)$, $\boldsymbol{wc}\left(\frac{c}{2}\right)$, $\boldsymbol{wr}\left(\frac{w+r}{2}\right)$, $\boldsymbol{cr}\left(\frac{c}{2}\right)$, $\boldsymbol{rr}\left(\frac{r}{2}\right)$	$\boldsymbol{ww}\left(\frac{1}{2}\right)$, $\boldsymbol{wr}\left(\frac{1}{2}\right)$	$\boldsymbol{ww}\left(\frac{1}{4}\right)$, $\boldsymbol{wr}\left(\frac{1}{2}\right)$, $\boldsymbol{rr}\left(\frac{1}{4}\right)$	*	*	*
	$\boldsymbol{cc}$	$\boldsymbol{wc}(w)$, $\boldsymbol{cc}(c)$, $\boldsymbol{cr}(r)$	$\boldsymbol{wc}(1)$	$\boldsymbol{wc}\left(\frac{1}{2}\right)$, $\boldsymbol{wr}\left(\frac{1}{2}\right)$	$\boldsymbol{cc}(1)$	*	*
	$\boldsymbol{cr}$	$\boldsymbol{wc}\left(\frac{w}{2}\right)$, $\boldsymbol{wr}\left(\frac{w}{2}\right)$, $\boldsymbol{cc}\left(\frac{c}{2}\right)$, $\boldsymbol{cr}\left(\frac{c+r}{2}\right)$, $\boldsymbol{rr}\left(\frac{r}{2}\right)$	$\boldsymbol{wc}\left(\frac{1}{2}\right)$, $\boldsymbol{wr}\left(\frac{1}{2}\right)$	$\boldsymbol{wc}\left(\frac{1}{4}\right)$, $\boldsymbol{wr}\left(\frac{1}{4}\right)$, $\boldsymbol{cr}\left(\frac{1}{4}\right)$, $\boldsymbol{rr}\left(\frac{1}{4}\right)$	$\boldsymbol{cc}\left(\frac{1}{2}\right)$, $\boldsymbol{cr}\left(\frac{1}{2}\right)$	$\boldsymbol{cc}\left(\frac{1}{4}\right)$, $\boldsymbol{cr}\left(\frac{1}{2}\right)$, $\boldsymbol{rr}\left(\frac{1}{4}\right)$	*
	$\boldsymbol{rr}$	$\boldsymbol{wr}(w)$, $\boldsymbol{cr}(c)$, $\boldsymbol{rr}(r)$	$\boldsymbol{wr}(1)$	$\boldsymbol{wr}\left(\frac{1}{2}\right)$, $\boldsymbol{rr}\left(\frac{1}{2}\right)$	$\boldsymbol{cr}(1)$	$\boldsymbol{cr}\left(\frac{1}{2}\right)$, $\boldsymbol{rr}\left(\frac{1}{2}\right)$	$\boldsymbol{rr}(1)$

Table 2: Possible values for each offspring g' in inheritance table $i(g_M, g_F, g')$, with probability of occurrence given by the number in parentheses, where allele probabilities $w = 1/2 - k_c/2$, $c = 1/2 + k_c(1-k_j)(1-k_n)/2$ and $r = k_c(k_n + k_j(1-k_n))/2$. Table is symmetric, so cells marked with * have the same values as their transposes.

where the number of available males of a subspecies and genotype is

$$W(g_M, m_M, m_F) = S(m_M, m_F)G(g_M)X_{N_{age}-1,M,g_M,m_M} \quad .$$

The numerator is the relative number of expected matings between a male of subspecies m_M (through S) and genotype g_M (through G) and the given female, while the denominator normalises the probability.

The relative probability of mating based on subspecies $S(m_M, m_F) = 1$ if $m_M = m_F$ and k otherwise. The relative fitness based on genotype is:

$$G(g) = \begin{cases} 1 & g \in \{ww, wr, rr\} \\ (1 - h_n s_n) & g \in \{wc, cr\} \\ (1 - s_n) & g = cc \end{cases} \tag{2.7}$$

where h_n and s_n are adapted from the Beaghton et al. (2017) model (see Table 1).

We compare two different scenarios for subspecies inheritance $H(m_M, m_F, m)$; maternal inheritance and equal inheritance. For maternal inheritance, the proportion of offspring born of a mating between subspecies $H(m_M, m_F, m) = 1$ if $m_F = m$ and 0 if $m_F \neq m$ (mother is always the same subspecies as her offspring). For equal inheritance, the proportion of offspring born of a mating between subspecies $H(m_M, m_F, m) = 0.5$ if $m_M \neq m_F$ (parents are different subspecies, i.e. cross-species offspring are equally split between subspecies). In both scenarios, $H(m_M, m_F, m) = 1$ if $m_M = m_F = m$ (both parents and offspring same subspecies) and 0 if $m_M \neq m$ and $m_F \neq m$ (both parents different subspecies to offspring). Where there are more than two subspecies, we also need to specify that $H = 0$ when $m_M \neq m_F$, $m_F \neq m$ and $m \neq m_M$ (i.e. parents and offspring are all of different subspecies).

The function $O(s, g)$ gives the probability of sex s and genotype g for the offspring. This probability depends on the genotypes of the parents such that

$$O(s, g) = i(g_M, g_F, g)\, p(g_M, g_F, s) \; . \tag{2.8}$$

The first term $i(\cdot)$ describes the probability of an offspring inheriting genotype g given parents of genotypes g_M and g_F. We again adapt the Beaghton et al. (2017) model to our target genotypes, including the effect of the gene drive construct, and using their parameter values (see Table 1). However, instead of assuming full random mixing of alleles as in their model, we directly model genotypes of each parent (see Table 2).

The second term $p(\cdot)$ describes the proportion of male offspring (and thus sex bias mechanisms). In the results presented in the main paper, we keep p constant at $p = 0.5$, though this can be readily varied to model constructs that induce sex bias in viable offspring (see Supplementary Materials Figure S4 for an example). The sex ratio is kept constant between subspecies (Paaijmans et al., 2009).

2.1.4 *Diffusion*

Diffusion is modelled by using the nearest-neighbour finite-difference approximation to the Laplacian. That is, with timestep size Δt and cell length Δx, the fraction of diffusing mosquitoes removed from one cell is $4D\Delta t/(\Delta x)^2$, where Δx is the cell-size. One quarter of this amount is added to each of the 4 neighbours. If the neighbours happen to be inactive (such as on the model boundary) those mosquitoes are assumed to die instantly.

2.1.5 *Wind advection*

Wind advection is expected to occur over very short time periods (overnight, as mosquitoes are not believed

to travel during daylight; see Huestis et al., 2019) and potentially very large distances (hundreds of kilometres, passively carried by the wind with negligible resistance; see Huestis et al., 2019). As such, we explicitly trace the trajectories of mosquitoes from each cell during each timestep (1 day), as advected by the wind vector field $\mathbf{V}(t, \mathbf{x})$. We use the Cross-Calibrated Multi-Platform (CCMP) Ocean Surface Wind Vector Analyses dataset (Atlas et al., 2011) to define this vector field; the data is available at the required daily timesteps over the time period required. We interpolate their vector field, given at 0.25 degree intervals (approximately 28 km at the equator), to fit our grid. Huestis et al. (2019) set up two different scenarios in their own modelling, where mosquitoes mosquitoes travel either 2 hours or 9 hours a night. We explore both scenarios here, and also a third scenario with no wind advection.

2.2 Larval carrying capacity

2.2.1 *Total abundance*

We apply the method of North and Godfray (2018) to estimate larval carrying capacity for both species combined. Specifically, we use Equation 1 from their paper:

$$K(t, \mathbf{x}) = \alpha_0(\mathbf{x}) + \alpha_1 \left(1 - e^{-\phi r(t,\mathbf{x})}\right) + \alpha_2 \left(1 - e^{-\kappa\left[W_p(\mathbf{x}) + W_n(\mathbf{x})\left(1 - e^{-\delta r(t,\mathbf{x})}\right)\right]}\right). \tag{2.9}$$

with parameters as estimated in their paper using Markov Chain Monte Carlo simulation with population data (see Table 1). The model incorporates rainfall ($r(t, \mathbf{x})$ in mm per week), as well as the availability of nearby rivers and lakes; these may be either permanent or intermittent ($W_p(\mathbf{x})$ and $W_n(\mathbf{x})$ respectively; details in their paper). It also has scope for modelling permanent larval sites ($\alpha_0(\mathbf{x})$) but we set this to be zero for our model. Whereas North and Godfray (2018) uses settlements as their sites for mosquito populations, we are calculating populations across a square grid, so we assume that each cell contains a settlement as modelled in their paper if and only if there is a human population present in the cell. An implicit assumption in this approach is that areas without human settlements yield zero carrying capacity. This assumption is explored for a sparsely populated region as follows. To calculate $K(t, \mathbf{x})$ at each cell at each required timestep using Equation 2.9, we estimate human presence or absence by using data produced by the Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University (2016) and publicly available from the Humanitarian Data Exchange (HDX; `https://data.humdata.org/`), except in Sudan, South Sudan and Somalia where these data are not available. There we assume human presence in all cells, which as an extreme case would facilitate the spatial spread of populations by local dispersal or advection. The other extreme, assuming human absence, is explored in Supplementary Materials Figure S7: it was found to have minimal effect outside of these countries, which take up a relatively small area near the edges of the species range.

We use inland water data from the Digital Chart of the World (DCW) as in their paper, and obtain monthly rainfall data from NASA's Land Data Assimilation System (`https://ldas.gsfc.nasa.gov/fldas`). We also set the mosquito population and carrying capacity to zero for each subspecies in locations outside their range as estimated by the Malaria Atlas Project (see Wiebe et al., 2017, updated by the Malaria Atlas Project, accessed 25 May 2021 from `https://malariaatlas.org/` and available for use under a Creative Commons Attribution 3.0 Unported License, `https://creativecommons.org/licenses/by/3.0/legalcode`).

2.2.2 *Relative abundance*

Once we have a measure of the combined larval carrying capacity $K(t, \mathbf{x})$, we need to separate this for two subspecies $K_1(t, \mathbf{x})$ and $K_2(t, \mathbf{x})$, which requires knowledge of the relative abundance of the two subspecies at each site. We use available field data collated by Tene Fossog et al. (2015) on the number of captures of both subspecies at various locations across Africa to flexibly predict a spatially-varying but temporally-constant relative abundance. We call this relative abundance $K_r(\mathbf{x}) = K_1(t, \mathbf{x})/K(t, \mathbf{x})$, or the proportion of *An. gambiae* s.s. at a site.

We first attempt to replicate the results of the Tene Fossog et al. (2015) logistic regression model by independently sourcing the predictors that they used in their final model: these are Latitude, Distance to Coast, Annual Mean Temperature, Mean Temperature of Wettest Quarter, Mean Elevation, Annual Normalized Vegetation Difference Index (NVDI) and Annual Variation in NVDI. We then apply a logistic regression model to the (slightly different) predictors, as they did. As well as providing an independent verification of the results in their paper, sourcing the predictors ourselves allows us to use more data sources across sub-Saharan Africa, allowing us to extrapolate and test different modelling approaches.

Using our independently sourced version of the predictors, we then apply a dense feed-forward neural network to flexibly model the relative abundance function $K_r(x)$ (see Supplementary Materials Appendix S2 for details) . We perform leave-one-out cross-validation (jackknifing) on each of these models, comparing the replicated logistic regression model and the original Tene Fossog et al. (2015) model with the neural network model, to test whether there is an increase in performance by incorporating nonlinear effects on the same set of variables.

The Tene Fossog et al. (2015) model uses only allopatric sites (those where only one subspecies was detected) to train their models, as did our replication model and initial neural network model. As we are interested specifically in subspecies overlap, we include sympatric

sites (where both subspecies were detected), including three new sites (Antonio-Nkondjio et al., 2012, 2013) publicly available from VectorBase (Giraldo-Calderón et al., 2015). We also add further predictors of potential interest to the described neural network model: mean annual precipitation (BIO12) and precipitation of wettest quarter (BIO16), which were also of interest to their model but excluded by their model selection process, along with salinity (Pombi et al., 2017) which was discussed in detail by Tene Fossog et al. (2015) but not included as a model predictor. We then apply forward selection to select the model variables in the final model (see Supplementary Materials Appendix S2 for details). To ensure convergence in the larger range of scenarios, we here use a 50:50 training-testing split on the data, but otherwise keep the same network topology and approach. Once this process is complete, we then have our final estimate of $K_r(\mathbf{x})$, which we can then apply to the previously calculated $K(t, \mathbf{x})$ to obtain results for both $K_1(t, \mathbf{x})$ and $K_2(t, \mathbf{x})$.

2.3 Parameters

Where possible and reasonable, we take parameter estimates from literature for our model (see Table 1), while checking that multiple sources give similar results.

2.3.1 Wind advection

We use estimates from Huestis et al. (2019) to indirectly estimate the probability of adult mosquitoes being advected by wind. They estimate that 6 million *An. coluzzii* mosquitoes cross a 100 km line perpendicular to the prevailing wind direction every year. Over the course of a 2 or 9 hour flight (the two night-time flight times tested in their Methods) their calculated trajectories give displacements of 3–69 km and 47–270 km respectively (means 38.6 and 154.1; 95% mean CIs 37–41 and 140–168). If we assume that each migrating individual completes just one 2 or 9 hour overnight flight each way and are only counted once, the mosquitoes that will cross the imaginary 100 km line will largely come from a rectangle bounded by this line on one side, and another 100 km line positioned either 38.6 or 154.1 km upwind on the other side. So the area over which mosquitoes will cross this line will be roughly 100 km $\times$ 38.6 km = $3,860$ km^2 or 100 km $\times$ 154.1 km = $15,410$ km^2, within which 6 million *An. coluzzii* migrating mosquitoes are estimated to cross. This gives us a density of 1.55×10^{-3} or 3.89×10^{-4} migrating mosquitoes per square metre per year: converting to relevant units gives us 106.39 and 26.65 mosquitoes per 5 km $\times$ 5 km cell per daily timestep. If we can then estimate the overall density of mosquitoes at their capture sites, we can then estimate the daily migration rate. Using the total abundance model from North and Godfray (2018), the average carrying capacity for larvae at coordinates $14°$N, $6.7°$W (located in between the capture sites in Mali) across the simulation period (2005–2015) is $41,526$. Given the other parameters $d_J = 0.05$ d^{-1}, $d_A = 0.1$ d^{-1} and $b = 0.1$ d^{-1}, at equilibrium we would expect approximately the same number of adults as larvae at carrying capacity. This gives a daily migration rate of $106.39/41,526 \approx 2.6 \times 10^{-3}$ (or 1 in 390) for 2 hour migration; and $26.65/41,526 \approx 6.4 \times 10^{-4}$ (or 1 in 1558) for 9 hour migration.

2.3.2 Initial conditions

As we expect similar numbers of larvae and adults, we set the initial conditions of the model for each age a, sex s, genotype g and mosquito subspecies m to

$$X_{a,s,g,m}(0, \mathbf{x}) = \begin{cases} \frac{1}{2}K_m(\mathbf{x}) & \text{for } g = ww \\ 0 & \text{for } g \neq ww \end{cases}$$

In cells where both subspecies exist, there will initially be competitive effects reducing numbers of one or both subspecies. In addition, advection and diffusion will affect the equilibrium. For these reasons, we run the model once from the initial condition described for the 11-year time period from 1 January 2005 to 31 December 2015 as a "burn-in" period, in order for the population to approximately reach a dynamic equilibrium that might be encountered in the wild.

Once the burn-in period is complete, we simultaneously introduce 10,000 male mosquitoes of each subspecies (*An. gambiae* s.s. and *An. coluzzii* that are heterozygous with the construct (genotype *wc*) in 15 separate locations (see Figure 4 in the Results), and run the model for another 11-year period. We choose heterozygous mosquitoes to introduce as these have less fitness cost than homozygous mosquitoes (see Equation 2.7), increasing the chance of spread. The first five of these are placed on islands off the African mainland at the nearest suitable location (see below) to assess the potential for incursion of the genetic construct onto the mainland. The island sites chosen as illustrative examples are:

1. the Bijagós islands (off Guinea-Bissau),
2. Bioko (off Cameroon),
3. Zanzibar (off Tanzania),
4. Comoros (off Mozambique), and
5. Madagascar.

The other ten sites (numbered 6–15) were chosen to be as evenly spaced as possible across the range of *An. gambiae* s.s. where at least 10,000 mosquitoes of either subspecies is available year-round (the algorithm is described in Supplementary Materials Appendix S1).

2.4 Scenario tests

As mentioned in Section 2.1.5, we use the two scenarios from Huestis et al. (2019) that mosquitoes are passively advected with the wind for either 2 hours or 9 hours a night. We also add a third scenario of no wind advection at all, to fully explore the effect of wind on mosquito movement. These three scenarios are then combined with the two scenarios of subspecies inheritance (maternal and fifty-fifty) described in Equation 2.6 to make six total scenarios modelled.

3 Results

	Original	Replication	NN
Error rate	0.1051	0.1018	0.1051
Cross-entropy	170.06	172.43	158.30
Expected error rate	0.1595	0.1610	0.1513
RMS error	0.2839	0.2839	0.2728

Table 3: Comparison of modelling approaches using four different measures of accuracy. The "Original" model is that of Tene Fossog et al. (2015), the "Replication" is our attempt to replicate their model with available data for predictors, and "NN" is our neural network model as described in this paper, but using their predictors. The models are compared with the dataset described in their paper (Tene Fossog et al., 2015) and from VectorBase (Antonio-Nkondjio et al., 2012, 2013; Giraldo-Calderón et al., 2015))

3.1 Larval carrying capacity

3.1.1 Relative abundance

To compare the effectiveness of the modelling approaches on the jackknifed mosquito data from Tene Fossog et al. (2015) and VectorBase (Antonio-Nkondjio et al., 2012, 2013; Giraldo-Calderón et al., 2015), where the model probability of the subspecies being *An. coluzzii* at site i of N sites is p_i, and the true probability is s_i (either 0 or 1 depending on subspecies present), we use four measures:

- the actual misclassification rate or "error rate", where the subspecies at site i is predicted to be *An. coluzzii* if $p_i > 0.5$, otherwise *An. gambiae* s.s.;
- the cross-entropy, calculated as $-\sum_i [s_i \log(p_i) + (1 - s_i) \log(1 - p_i)]$
- the expected number of misclassifications or errors, $\sum_i \mathrm{abs}(p_i - s_i)/N$
- and the root mean square (RMS) error, $\sqrt{\sum_i (p_i - s_i)^2/N}$

Table 3 shows that our neural network model performs as well (error rate), slightly better (expected error rate and RMS error) or much better (cross-entropy) on all of the measures. Its much better performance on cross-entropy is likely due to the fact that it is trained specifically to minimise cross-entropy, which may not directly correlate with lower error rates — for example, compared to the other measures, cross-entropy will much more harshly penalise a model for assigning a very low probability to an event which then occurs in the testing data.

Figure 1 illustrates the forward selection process. The chosen variables are Mean Annual Temperature, Latitude, Elevation and Distance to Coast — after this point, even the best chosen variable added to the model only increases the mean validation loss.

Note that the NVDI Coefficient of Variation (CV) actually had the lowest mean validation loss in the second round, but Latitude was instead selected for four reasons:

- The two had mean validation losses that were statistically indistinguishable even after 500 model runs, based on the bootstrap 95% confidence intervals,
- Comparing the results of the forward selection process where either the NVDI CV or Latitude is chosen in the second round, the latter gives much better mean validation loss results in *subsequent* rounds, demonstrating that the forward selection process is sub-optimally choosing the NVDI CV,
- Data for the NVDI CV is much harder to source, and the variable is a less directly biologically relevant predictor than Latitude, and
- The NVDI CV becomes a much less desirable predictor after Latitude is selected, suggesting that it has little to contribute to the model that is unique to it, and not already contributed by Latitude.

Figure 2 shows the final estimated relative abundance as a mean of 100 neural network model runs using the chosen variables, alongside the standard deviation of these runs. The relative abundance follows a similar pattern to that of Tene Fossog et al. (2015), with *An. gambiae* s.s. mostly dominant except in some coastal areas and towards the Sahel. As expected, the variance between model runs is mostly low around the data points in western sub-Saharan Africa upon which the model was trained, with much higher variance around central and southern Africa. Interestingly the model runs also yield low variance around Ethiopia in particular, where they predict *An. gambiae* s.s. to be dominant.

3.1.2 Total abundance

Figure 3 demonstrates the resulting estimated larval carrying capacity for each subspecies across the first year of modelling given the relative abundance model above and the North and Godfray (2018) total abundance model, combined with information on human presence and the known distributions of each subspecies.

3.2 Scenario tests

The choice of subspecies inheritance made no noticeable difference to the results for any of the three wind scenarios. Results shown use the maternal inheritance model.

Within the three wind scenarios, the zero-advection scenario experiences no noticeable mosquito transport on a continental scale, demonstrating that with the current selected parameters, advection is far more important than diffusion. We thus show figures for the 9-hour case below, contrast the 2-hour results in text, and present the full 2-hour and zero-advection results in Supplementary Materials.

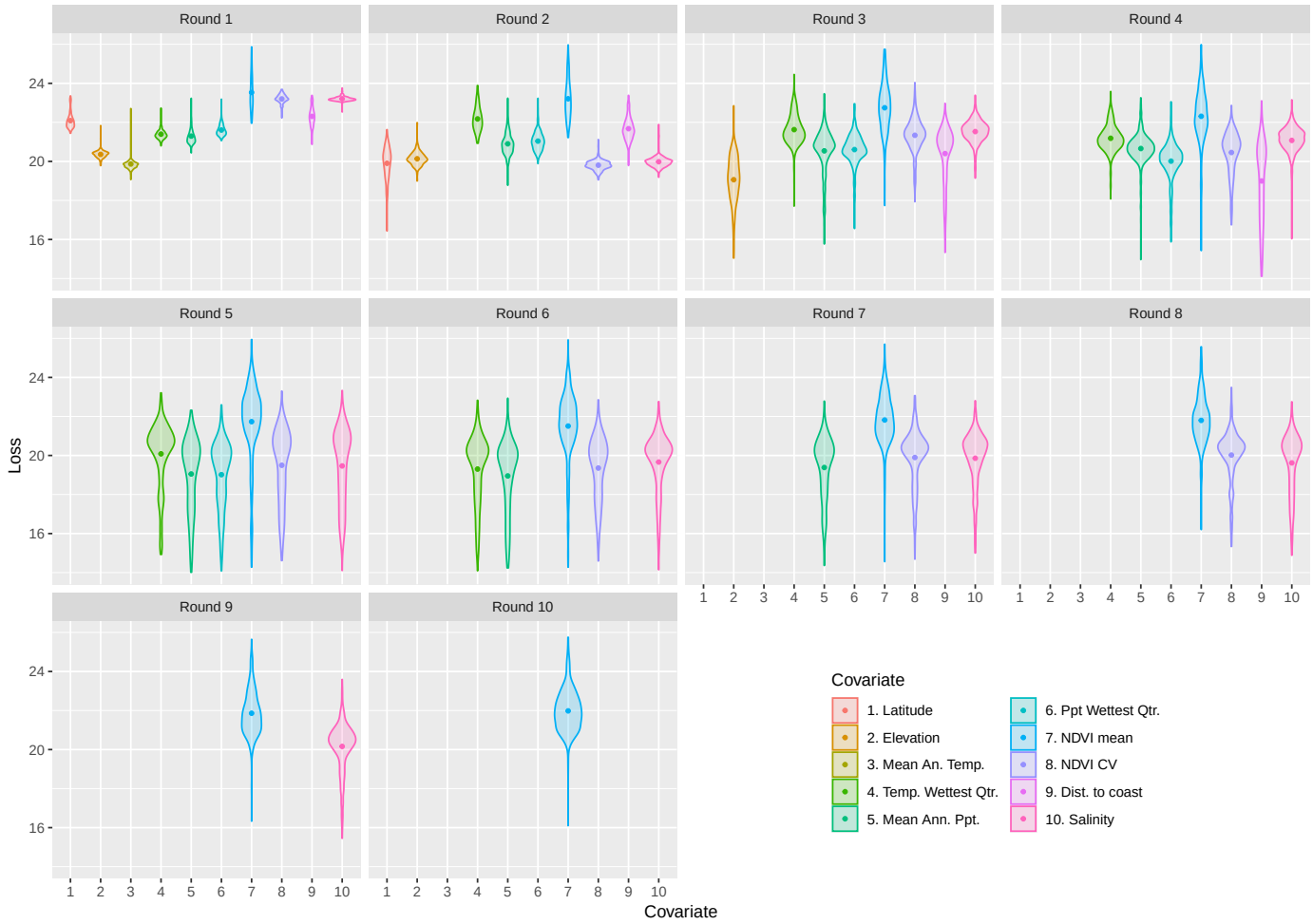


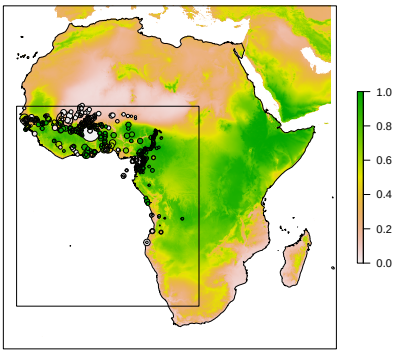
Figure 1: The forward selection process for the neural network model that estimates the relative abundance of the two subspecies. For each round of selection, the validation loss for each variable is shown, both for the 500 individual neural network runs (small dots) and the mean (large labelled dot), with a different colour. After a given round, the variable with the lowest mean validation loss is selected then the process is repeated with a model including that variable, testing all remaining variables. 95% bootstrap confidence intervals for the mean are displayed for each variable in each round.

Figure 4 shows the location and spread of the construct from the 15 introduction sites of the genetic construct (labelled 1–15) for the highest wind advection scenario (9 hours).

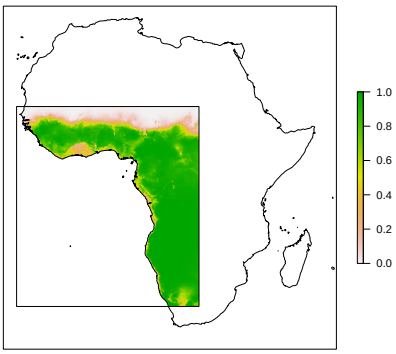
All of the islands were able to maintain a population with the genetic construct (Figure 5), with all except for Madagascar (Site 5) then invading the African mainland in subsequent years. There appears to be a barrier to dispersal in central Africa, with most introductions either remaining in western Africa (sites 1, 2, 7, 9, 14 and 15) or eastern Africa (sites 3, 4, 6, 8, 10–13). When advection was reduced to 2 hours, only the Bijagós (Site 1) and Zanzibar (Site 3) introductions were able to reach the mainland (see Supplementary Materials Figure S6).

Figure 5 shows more detail about the introductions at each site. At all sites, the construct allele completely takes over from the original wildtype allele in a matter of months. Once the construct is established in the population, resistance builds slowly but surely: the heterozygous resistant genotype cr becomes noticeable after a couple of years, and is beginning to overtake the wildtype by the end of the 11-year simulation period, with the homozygous resistant genotype (rr) starting to become noticeable in the population. All sites show a similar pattern, despite differences in scale, subspecies composition and seasonality. Only *An. gambiae* s.s. persists in Sites 3, 4, 5, 7, 8, 11 and 12, and conversely only *An. coluzzii* persists in Sites 9 and 14. The two coexist in the remaining sites (1, 2, 6, 10, 13 and 15). All sites show a regular seasonal pattern to abundance based on rainfall (the only seasonal driver in the model, other than wind), with the possible exceptions of Sites 4, 6 and 8, which show slightly more variable seasonality patterns. Why this might be is unclear, though these sites are all in central to eastern Africa.

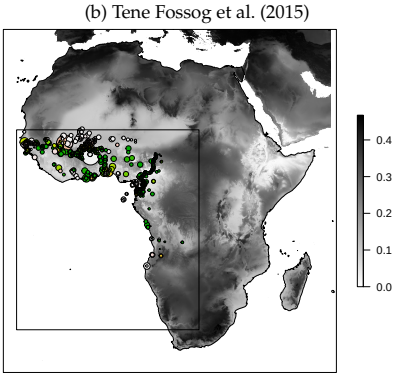
Full colour animations of all process model outputs are available in Supplementary Materials.



(a) Mean



(b) Tene Fossog et al. (2015)



(c) Standard deviation

Figure 2: Summary statistics of relative abundance based on 100 neural network model runs. Mean relative abundance (a) is given as the proportion of *An. gambiae* s.s. (as opposed to *An. coluzzii*) in the local mosquito population, with low proportions given as white and high proportions as green. Circles represent data points on which the model is trained, filled with colour representing the proportion measured at the given site. The corresponding results from Tene Fossog et al. (2015) are given for the purposes of direct comparison (b). The standard deviation of relative abundance (c) is shown in greyscale, with white as low and black as high variance between model estimates at a given site. The rectangle denotes the area of study used in Tene Fossog et al. (2015).

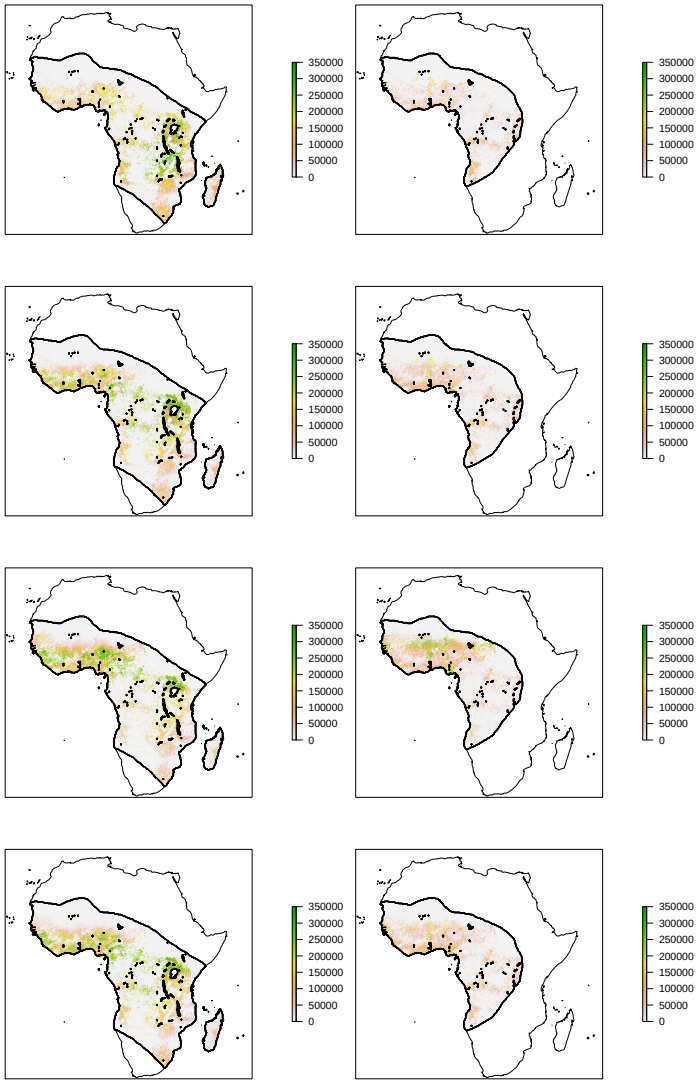


Figure 3: Estimated larval carrying capacity of *An. gambiae* s.s. (left) and *An. coluzzii* (right), for 2005 (the first year of modelling) in January (Southern Hemisphere summer), April (autumn), July (winter) and October (spring) from top to bottom.

4 Discussion

This study is the first continental scale model of population dynamics for two of the dominant malaria vector species in the *Anopheles gambiae* sensu lato species complex. The transient spread and persistence of a population replacement gene drive was predicted for the two hybridising subspecies *An. gambiae* sensu stricto and *An. coluzzii*. The two major factors that determine the spread of the gene drive at the continental scale were 1) potential wind-driven dispersal and 2) spatially and temporally varying carrying capacities for the two subspecies with both intraspecific and interspecific density dependence occurring at the larval stage. Our presented results are intended to demonstrate the plausibility of wide-scale spread of gene drive and structurally evaluate hypotheses of relevant life history strategies for these two dominant malaria vectors. The spatially explicit process model is designed to support more specific scenario-based assessments of both genetic and conventional vector control strategies.

Vertical gene transfer among the investigated subspecies is complicated by the spatially heterogeneous population structure, where introgression rates between *An. gambiae* s.s. and *An. coluzzii* vary with geographic location (*Anopheles gambiae* 1000 Genomes Consortium,

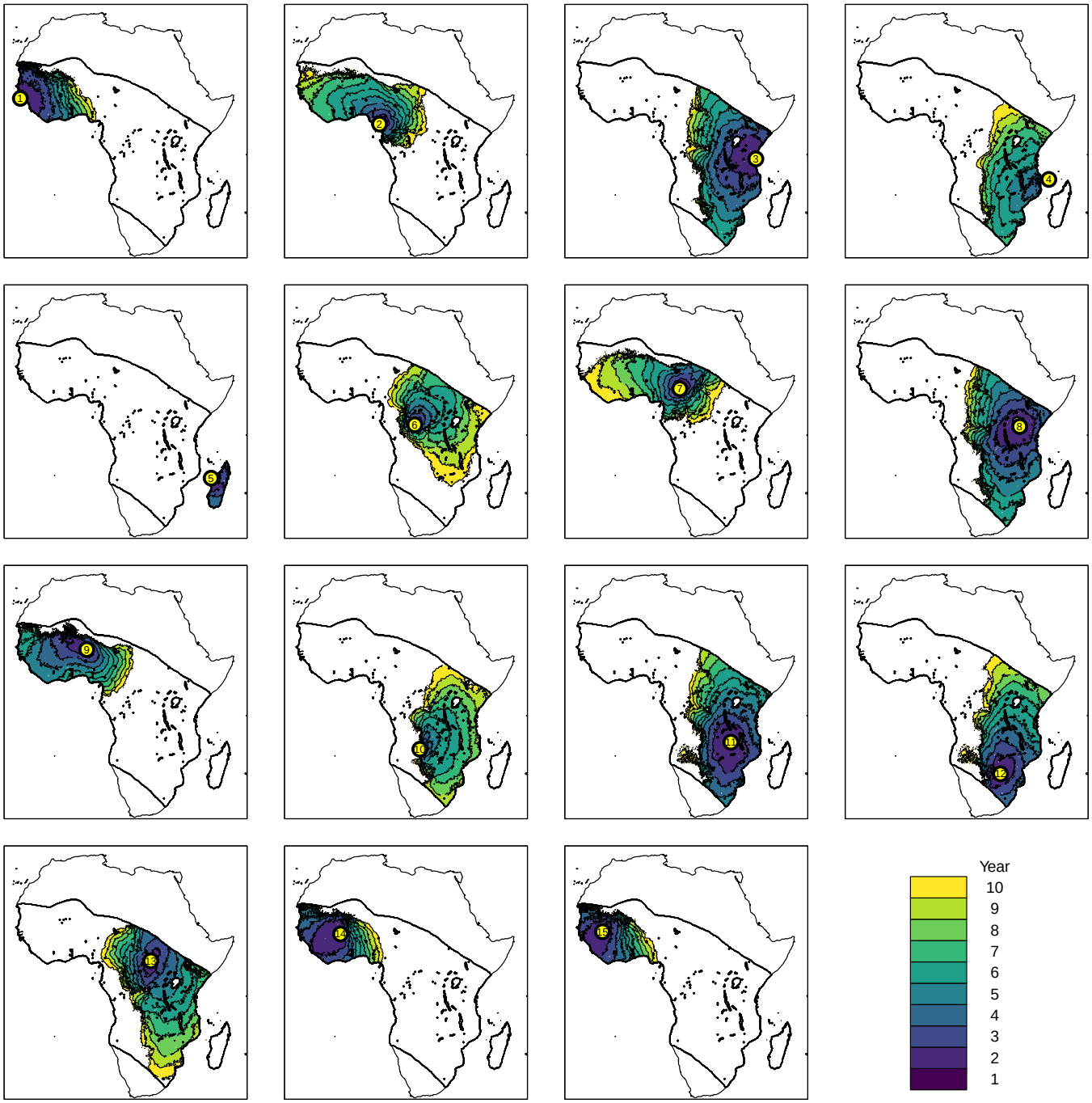


Figure 4: The invasion front of the construct (defined as having at least two alleles, e.g. one *cc* or two *wc* mosquitoes, in a cell) from a selection of starting points, with a separate colour given for each year. The island introductions are (1) the Bijagós islands (off Guinea-Bissau), (2) Bioko (off Cameroon), (3) Zanzibar (off Tanzania), (4) Comoros (off Mozambique) and (5) Madagascar.

2017; Pombi et al., 2017; Clarkson et al., 2020), and perhaps also over time (Lee et al., 2013). Pombi et al. (2017) suggests that the rate of hybridisation between these subspecies depends on their relative frequency in the population. The model therefore tracks the relative number of available males by subspecies, and two alternative choices for species assignment of the first generation hybrids was investigated. The model results were found to be robust to species assignment via maternal descent when compared to equal proportions of species among the first generation hybrids.

The home range of *An. gambiae* s.l. is typically believed to be less than 5 square kilometres, although there is evidence for long-range dispersal (Huestis et al., 2019). Spatial dispersal was modelled through two different mechanisms: a local diffusive process, which may be me-

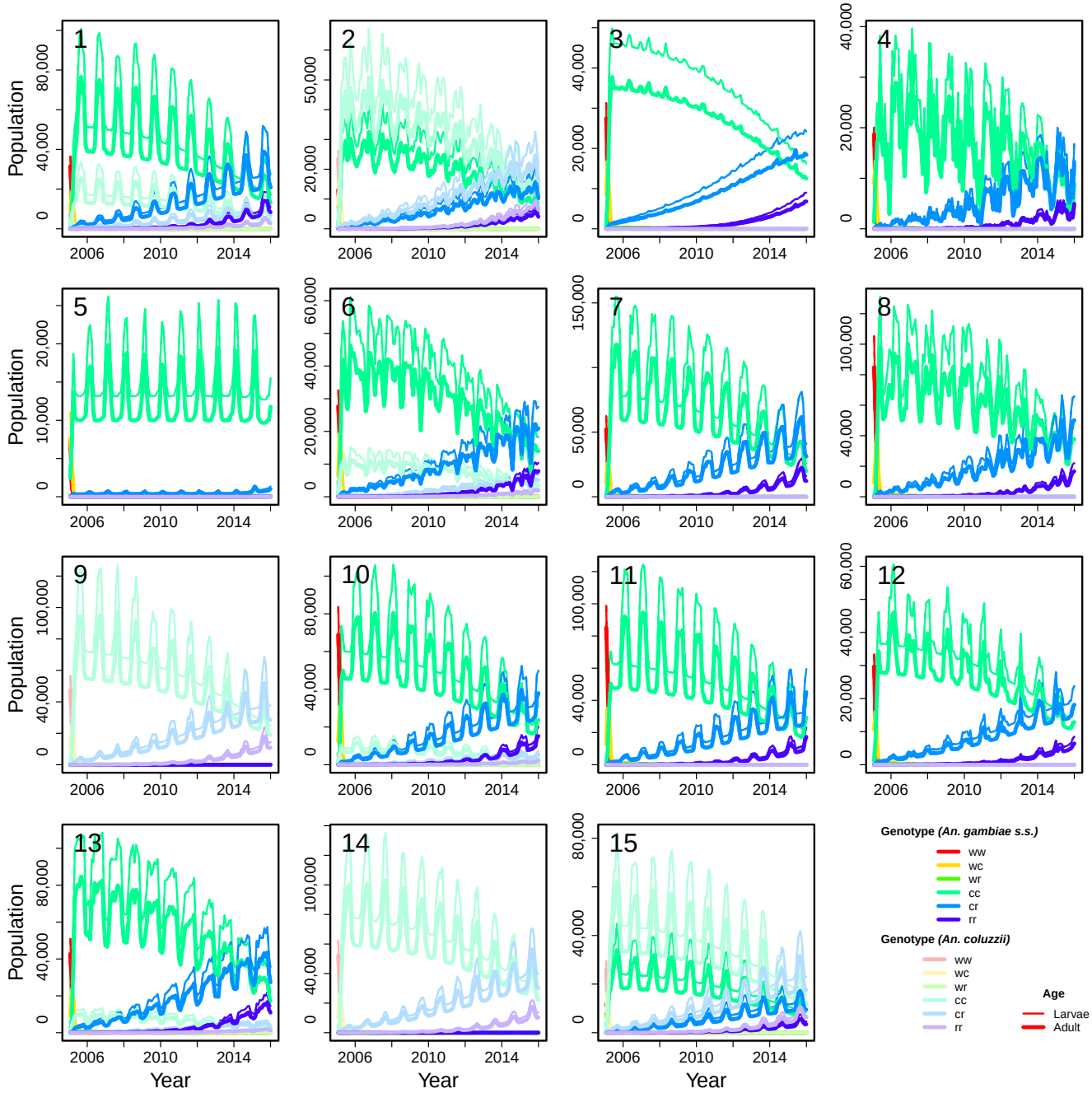


Figure 5: The time series abundance of male mosquitoes at each introduction point (the sub-figure number corresponds to the release site; see Figure 4), separated by species, genotype and age (female mosquitoes occur in identical numbers to males in this model). The colours correspond to genotype and the line thickness to age class.

diated by the presence of *An. gambiae* s.l. between human settlements (Epopa et al., 2020), versus wind-driven advection. More generally, wind-driven dispersal events have been observed for mosquitoes including over ocean, although it is not understood whether or not wind-driven dispersal is a deliberate life history strategy for some species (Service, 1997). For genetic vector control strategies, the possibility of long range dispersal events are an important consideration that should be incorporated into the selection of field sites (James et al., 2018; Lanzaro et al., 2021). Unsurprisingly our results indicate that passive wind-driven advection, if present, can greatly increase the speed and spatial footprint of the invasion front for a near-neutral population replacement gene drive (Figure 4, Section 3.2).

The importance of wind-driven advection as a dispersal strategy for *An. gambiae* s.l. will likely be a key uncertainty in future risk assessments for genetic control strate-

gies. The large difference in the dispersal between the 9 hour (Figure 4) and 2 hour results (Supplementary Materials Figures S5 and S6) have important implications for regional (trans-national) governance arrangements, the scope of stakeholder engagement activities and the degree of geographic containment that islands might provide during a staged-release strategy. In our simulated releases of the gene drive on islands, only Madagascar was sufficiently distant to prevent spread to the mainland under the 9 hour wind-assisted dispersal. Even when reducing wind-assisted dispersal to 2 hours, two of the five island releases resulted in spread to the mainland. Simulated introductions onto islands such as Bioko (32 km from the mainland) spread to the mainland, but only under the 9 hour scenario. Hence, only this result is consistent with recent genomic analysis that shows mosquito populations on this island are not isolated from the mainland (Campos et al., 2021). Although our results make the potential scale of wind-mediated spread of gene drive clear, exactly how and where this spread occurs in the relevant mosquito taxa, and the mosquito behaviour that helps drive it, is still being studied (Huestis et al., 2019; Florio et al., 2020). In addition, while data on previous wind patterns is readily available, predicting future wind patterns can only be done in very general terms, especially given the added complications and uncertainties caused by anthropogenic climate change. These issues together mean that future predictions of gene drive spread will likely involve high levels of uncertainty.

Environmental carrying capacity is another key factor that determines the spread and persistence of the simulated gene drive releases, as well as the wild-type population abundance. The introgression of the gene drive tended to initially follow seasonal patterns of carrying capacity driven by precipitation (Figure 5). Later, however, the importance of resistance gradually overcame the drive within 11 years despite the spatio-temporal variability of the population abundances and a relatively mild genetic load imposed by the construct. Resistance is a recognised challenge for gene drive systems (e.g., Hammond et al., 2017; Price et al., 2020; James et al., 2020) for which various counter-strategies have been proposed (Beaghton et al., 2017; Nash et al., 2019; Nolan, 2021; Wang et al., 2021). Our simulations assume resistance alleles arise with frequency r (Table 2) determined by the probability of cleavage (k_c), the probability of non-homologous repair (k_j) and the probability that the nuclease gene becomes non-functional due to mutation of the target site during homologous repair (k_n) (Beaghton et al., 2017). Our model, however, does not account for pre-existing resistance in wild type populations caused by sequence variation in the target locus. Hence the (nonetheless rapid) progression of resistance in hybridising and spatially heterogeneous populations shown here could be underestimated, further emphasising the importance of managing drive resistance.

Our characterisation of environmental carrying capacities and abundance for the wild-type populations can accommodate alternative functional forms and parametrisation. Moreover, the functional forms and parametrisation of carrying capacity may be expected to change with time as climate (Kelly-Hope et al., 2009; Afrane et al., 2012) and land use (Keating et al., 2004; Munga et al., 2009) changes. The current lack of quantitative, species-specific, data on mortality and dispersal within the *An. gambiae* s.l. complex, however, limits our ability to parametrise relationships such as the larval carrying capacity (Equation 2.9) in a species-specific fashion (North and Godfray, 2018), beyond a few, spatially limited, empirical studies (e.g. Magombedze et al., 2018). We anticipate that as data from entomological surveys, coupled with species differentiation through genetic methods (e.g., *Anopheles gambiae* 1000 Genomes Consortium, 2017; Clarkson et al., 2020; Giraldo-Calderón et al., 2015) is increasingly centralised, then more detailed parametrisations will become possible at the subspecies level.

A lack of data also constrains our ability to compare alternative hypotheses of how *An. gambiae* s.l. populations persist in marginal habitat zones such as the Sahel, where observed patterns seasonal abundance can be explained by either aestivation, permanence of larval microhabitats (i.e., non-zero carrying capacities during the dry season) or long range dispersal (North and Godfray, 2018). Although persistence in the Sahel can be explained without an explicit aestivation model (see Figures 3, 5), aestivation may nevertheless be an important life history strategy for some species within the *An. gambiae* s.l. complex in the Sahel (Lehmann et al., 2017). As for wind-driven dispersal, the relative importance of aestivation suffers from limited data.

The process model allows for the scenario-based testing of genetic and conventional vector control strategies. For genetic vector control, the sex-differentiated compartment model allows for both population replacement and population suppression gene drives. For the latter, the relative frequency of males and females may be an important component of the gene drive system (Kyrou et al., 2018) and also for sex-biased, self-limiting forms of genetic vector control strategies (Windbichler et al., 2008; Klein et al., 2012; Galizi et al., 2014); the model has the flexibility to accommodate sex bias from either maternal or paternal descent (Section 2.1.3). Conventional control strategies that target adult females or larval stage mosquitoes can be accommodated by increasing mortality rates in locations and times dependent on the intervention scenario (sensu Magombedze et al., 2018), or by reducing carrying capacity for interventions where larval habitat is removed.

Vector control is a key component in strategies developed to combat mosquito-borne and vector borne diseases (Wilson et al., 2020). The deployment of vector control strategies, whether genetic or conventional, into hybridising spatially heterogeneous populations will require the development of spatially explicit models. These

models should be constructed at a commensurate spatial and temporal scale of the intervention and include the possibility of resistance, which is not only a feature of genetic methods such as gene drive systems but is also an expected development for conventional strategies such as insecticide applications (Achee et al., 2019; Nolan, 2021). Numerical simulation based scenario assessments can be used to investigate alternative ecological hypotheses in concert with proposed vector control intervention strategies. The spatio-temporal model developed here shows the importance of resistance, vertical gene transfer among hybridising subspecies, long range dispersal, spatio-temporal variability in larval mosquito habitats and deployment strategies at the continental scale.

Acknowledgments

The authors gratefully acknowledge the support of CSIRO Data61, CSIRO Health and Biosecurity and the Foundation for the National Institutes of Health. We would also like to thank Roslyn Hickson and Mathieu Legros for their thoughtful comments.
Declarations of interest: none

Data availability

Illustrative animations are available: see Supplementary Materials Appendix S3.

References

Achee, N., Grieco, J. P., Vatandoost, H., Seixas, G., Pinto, J., Ching-NG, L., Martins, A. J., Juntarajumnong, W., Corbel, V., Gouagna, C., David, J.-P., Logan, J. G., Orsborne, J., Marois, E., Devine, G. J., and Vontas, J. (2019). Alternative strategies for mosquito-borne arbovirus control. *PLOS Neglected Tropical Diseases*, 13:e0006822.

Afrane, Y. A., Githeko, A. K., and Yan, G. (2012). The ecology of *Anopheles* mosquitoes under climate change: case studies from the effects of environmental changes in east Africa highlands. *Annals of the New York Academy of Sciences*, 1249:204.

Alphey, L. S., Crisanti, A., Randazzo, F., and Akbari, O. S. (2020). Opinion: Standardizing the definition of gene drive. *Proceedings of the National Academy of Sciences*, 117(49):30864–30867.

Anopheles gambiae 1000 Genomes Consortium (2017). Genetic diversity of the African malaria vector *Anopheles gambiae*. *Nature*, 552(7683):96–100.

Antonio-Nkondjio, C., Defo-Talom, B., Tagne-Fotso, R., Tene-Fossog, B., Ndo, C., Lehman, L. G., Tchuinkam, T., Kengne, P., and Awono-Ambene, P. (2012). High mosquito burden and malaria transmission in a district of the city of Douala, Cameroon. *BMC Infectious Diseases*, 12(1):275.

Antonio-Nkondjio, C., Demanou, M., Etang, J., and Bouchite, B. (2013). Impact of cyfluthrin (Solfac EW050) impregnated bed nets on malaria transmission in the city of Mbandjock : lessons for the nationwide distribution of long-lasting insecticidal nets (LLINs) in Cameroon. *Parasites & Vectors*, 6(1):10.

Atlas, R., Hoffman, R. N., Ardizzone, J., Leidner, S. M., Jusem, J. C., Smith, D. K., and Gombos, D. (2011). A cross-calibrated, multiplatform ocean surface wind velocity product for meteorological and oceanographic applications. *Bulletin of the American Meteorological Society*, 92(2):157–174.

Beaghton, A., Hammond, A., Nolan, T., Crisanti, A., Godfray, H. C. J., and Burt, A. (2017). Requirements for driving antipathogen effector genes into populations of disease vectors by homing. *Genetics*, 205(4):1587.

Beeton, N. J., Hosack, G. R., Wilkins, A., Forbes, L. K., Ickowicz, A., and Hayes, K. R. (2020). Modelling competition between hybridising subspecies. *Journal of Theoretical Biology*, 486:110072.

Behnel, S., Bradshaw, R., Citro, C., Dalcin, L., Seljebotn, D. S., and Smith, K. (2011). Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39.

Bhatt, S., Weiss, D. J., Cameron, E., Bisanzio, D., Mappin, B., Dalrymple, U., Battle, K. E., Moyes, C. L., Henry, A., Eckhoff, P. A., Wenger, E. A., Briët, O., Penny, M. A., Smith, T. A., Bennett, A., Yukich, J., Eisele, T. P., Griffin, J. T., Fergus, C. A., Lynch, M., Lindgren, F., Cohen, J. M., Murray, C. L. J., Smith, D. L., Hay, S. I., Cibulskis, R. E., and Gething, P. W. (2015). The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*, 526:207–211.

Campos, M., Hanemaaijer, M., Gripkey, H., Collier, T. C., Lee, Y., Cornel, A. J., Pinto, J., Ayala, D., Rompão, H., and Lanzaro, G. C. (2021). The origin of island populations of the African malaria mosquito, *Anopheles coluzzii*. *Communications Biology*, 4(1):630.

Caputo, B., Nwakanma, D., Jawara, M., Adiamoh, M., Dia, I., Konate, L., Petrarca, V., Conway, D. J., and della Torre, A. (2008). *Anopheles gambiae* complex along The Gambia river, with particular reference to the molecular forms of *An. gambiae* s.s. *Malaria Journal*, 7(1):182.

Carballar-Lejarazú, R., Ogaugwu, C., Tushar, T., Kelsey, A., Pham, T. B., Murphy, J., Schmidt, H., Lee, Y., Lanzaro, G. C., and James, A. A. (2020). Next-generation gene drive for population modification of the malaria vector mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*, 117(37):22805–22814.

Chollet, F., Allaire, J., et al. (2017). R Interface to Keras. `https://github.com/rstudio/keras`.

Clarkson, C. S., Miles, A., Harding, N. J., Lucas, E. R., Battey, C. J., Amaya-Romero, J. E., Cano, J., Diabate, A., Constant, E., Nwakanma, D. C., Jawara, M., Essandoh, J., Dinis, J., Le Goff, G., Robert, V., Troco, A. D., Costantini, C., Rohatgi, K. R., Elissa, N., Coulibaly, B., Midega, J., Mbogo, C., Mawejje, H. D., Stalker, J., Rockett, K. A., Drury, E., Mead, D., Jeffreys, A. E., Hubbart, C., Rowlands, K., Isaacs, A. T., Jyothi, D., Malangone, C., Kamali, M., Henrichs, C., Simpson, V., Ayala, D., Besansky, N. J., Burt, A., Caputo, B., della Torre, A., Fontaine, M., Godfray, H. C. J., Hahn, M. W., Kern, A. D., Lawniczak, M. K. N., O'Loughlin, S., Pinto, J., Riehle, M. M., Sharakhov, I., Schrider, D. R., Vernick, K. D., White, B. J., Donnelly, M. J., and Kwiatkowski, D. P. (2020). Genome variation and population structure among 1,142 mosquitoes of the African malaria vector species *Anopheles gambiae* and *Anopheles coluzzii*. *bioRxiv*.

Coetzee, M., Hunt, R. H., Wilkerson, R., Della Torre, A., Coulibaly, M. B., and Besansky, N. J. (2013). *Anopheles coluzzii* and *Anopheles amharicus*, new members of the *Anopheles gambiae* complex. *Zootaxa*, 3619(3). doi:10.11646/zootaxa.3619.3.2.

Costantini, C., Ayala, D., Guelbeogo, W. M., Pombi, M., Some, C. Y., Bassole, I. H. N., Ose, K., Fotsing, J.-M., Sagnon, N., Fontenille, D., Besansky, N. J., and Simard, F. (2009). Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. *BMC Ecology*, 9(1):16.

Dao, A., Yaro, A., Diallo, M., Timbine, S., Huestis, D., Kassogue, Y., Traoré, A., Sanogo, Z., Samaké, D., and Lehmann, T. (2014). Signatures of aestivation and migration in Sahelian malaria mosquito populations. *Nature*, 516(7531):387–390.

della Torre, A., Costantini, C., Besansky, N. J., Caccone, A., Petrarca, V., Powell, J. R., and Coluzzi, M. (2002). Speciation within *Anopheles gambiae* – the glass is half full. *Science*, 298(5591):115–117.

della Torre, A., Fanello, C., Akogbeto, M., Dossou-yovo, J., Favia, G., Petrarca, V., and Coluzzi, M. (2001). Molecular evidence of incipient speciation within *Anopheles gambiae* s.s. in West Africa. *Insect Molecular Biology*, 10(1):9–18.

della Torre, A., Tu, Z., and Petrarca, V. (2005). On the distribution and genetic differentiation of *Anopheles gambiae* s.s. molecular forms. *Insect Biochemistry and Molecular Biology*, 35(7):755 – 769. Genetic manipulation of insects.

DePina, A., Namountougou, M., Leal, S., Varela, I., Monteiro, D.,

Ramos de Sousa, C., Goncalves, A., Moreira, A., and Faye, O. (2018). Susceptibility of *Anopheles gambiae s.l.* to the insecticides in Praia, Cape Verde: A country in the pre-elimination of malaria. *Vector Biology Journal*, 3:2.

Diabaté, A., Dabiré, R. K., Heidenberger, K., Crawford, J., Lamp, W. O., Culler, L. E., and Lehmann, T. (2008). Evidence for divergent selection between the molecular forms of *Anopheles gambiae*: role of predation. *BMC Evolutionary Biology*, 8(1):5.

Diabaté, A., Dabiré, R. K., Kim, E. H., Dalton, R., Millogo, N., Baldet, T., Simard, F., Gimnig, J. E., Hawley, W. A., and Lehmann, T. (2005). Larval development of the molecular forms of *Anopheles gambiae* (Diptera: Culicidae) in different habitats: A transplantation experiment . *Journal of Medical Entomology*, 42(4):548–553.

Dufourd, C. and Dumont, Y. (2013). Impact of environmental factors on mosquito dispersal in the prospect of sterile insect technique control. *Computers & Mathematics with Applications*, 66(9):1695 – 1715. BioMath 2012.

Eckhoff, P. A. (2011). A malaria transmission-directed model of mosquito life cycle and ecology. *Malaria Journal*, 10(21999664):303–303.

Eckhoff, P. A., Wenger, E. A., Godfray, C. J., and Burt, A. (2016). Impact of mosquito gene drive on malaria elimination in a computational model with explicit spatial and temporal dynamics. *Proceedings of the National Academy of Sciences*, 114(2):E255–E264.

Epopa, P. S., Millogo, A. A., Collins, C. M., North, A. R., Benedict, M. Q., Tripet, F., O'Loughlin, S., Dabiré, R. K., Ouédraogo, G. A., and Diabaté, A. (2020). *Anopheles gambiae* (sl) is found where few are looking: assessing mosquito diversity and density outside inhabited areas using diverse sampling methods. *Parasites & Vectors*, 13(1):1–11.

Facchinelli, L., North, A. R., Collins, C. M., Menichelli, M., Persampieri, T., Bucci, A., Spaccapelo, R., Crisanti, A., and Benedict, M. Q. (2019). Large-cage assessment of a transgenic sex-ratio distortion strain on populations of an African malaria vector. *Parasites & Vectors*, 12:70–70.

Facebook Connectivity Lab and Center for International Earth Science Information Network - CIESIN - Columbia University (2016). High Resolution Settlement Layer (HRSL). Source imagery for HRSL © 2016 DigitalGlobe. Accessed 31 August 2020.

Florio, J., Verú, L. M., Dao, A., Yaro, A. S., Diallo, M., Sanogo, Z. L., Samaké, D., Huestis, D. L., Yossi, O., Talamas, E., Chamorro, M. L., Frank, J. H., Biondi, M., Morkel, C., Bartlett, C., Linton, Y.-M., Strobach, E., Chapman, J. W., Reynolds, D. R., Faiman, R., Krajacich, B. J., Smith, C. S., and Lehmann, T. (2020). Diversity, dynamics, direction, and magnitude of high-altitude migrating insects in the Sahel. *Scientific Reports*, 10(1):20523.

Galizi, R., Doyle, L. A., Menichelli, M., Bernardini, F., Deredec, A., Burt, A., Stoddard, B. L., Windbichler, N., and Crisanti, A. (2014). A synthetic sex ratio distortion system for the control of the human malaria mosquito. *Nature Communications*, 5(1):3977.

Gantz, V. M., Jasinskiene, N., Tatarenkova, O., Fazekas, A., Macias, V. M., Bier, E., and James, A. A. (2015). Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito *Anopheles stephensi*. *Proceedings of the National Academy of Sciences*, 112(49):E6736–E6743.

Gillies, M. (1961). Studies on the dispersion and survival of *Anopheles gambiae* (Giles) in East Africa, by means of marking and release experiments. *Bulletin Entomological Research*, 52:99–127.

Giraldo-Calderón, G. I., Emrich, S. J., MacCallum, R. M., Maslen, G., Dialynas, E., Topalis, P., Ho, N., Gesing, S., VectorBase Consortium, Madey, G., Collins, F. H., and Lawson, D. (2015). VectorBase: an updated bioinformatics resource for invertebrate vectors and other organisms related with human diseases. *Nucleic Acids Research*, 43:D707–D713.

Hammond, A., Galizi, R., Kyrou, K., Simoni, A., Siniscalchi, C., Katsanos, D., Gribble, M., Baker, D., Marois, E., Russell, S., Burt, A., Windbichler, N., Crisanti, A., and Nolan, T. (2016). A CRISPR-Cas9 gene drive system targeting female reproduction in the malaria mosquito vector *Anopheles gambiae*. *Nature biotechnology*, 34(26641531):78–83.

Hammond, A. M., Kyrou, K., Bruttini, M., North, A., Galizi, R., Karlsson, X., Kranjc, N., Carpi, F. M., D'Aurizio, R., Crisanti, A., and Nolan, T. (2017). The creation and selection of mutations resistant to a gene drive over multiple generations in the malaria mosquito. *PLOS Genetics*, 13(10):1–16.

Huestis, D. L., Dao, A., Diallo, M., Sanogo, Z. L., Samake, D., Yaro, A. S., Ousman, Y., Linton, Y.-M., Krishna, A., Veru, L., et al. (2019). Windborne long-distance migration of malaria mosquitoes in the Sahel. *Nature*, 574(7778):404–408.

Ickowicz, A., Foster, S. D., Hosack, G. R., and Hayes, K. R. (2021). Predicting the spread and persistence of genetically modified dominant sterile male mosquitoes. *Parasites & Vectors*, 14(1):480.

James, S., Collins, F. H., Welkhoff, P. A., Emerson, C., Godfray, H. C. J., Gottlieb, M., Greenwood, B., Lindsay, S. W., Mbogo, C. M., Okumu, F. O., Quemada, H., Savadogo, M., Singh, J. A., Tountas, K. H., and Touré, Y. T. (2018). Pathway to deployment of gene drive mosquitoes as a potential biocontrol tool for elimination of malaria in Sub-Saharan Africa: Recommendations of a scientific working group. *The American Journal of Tropical Medicine and Hygiene*, 98(6):1 – 49.

James, S. L., Marshall, J. M., Christophides, G. K., Okumu, F. O., and Nolan, T. (2020). Toward the definition of efficacy and safety criteria for advancing gene drive-modified mosquitoes to field testing. *Vector-Borne and Zoonotic Diseases*, 20(4). PMID: 32155390.

Keating, J., Macintyre, K., Mbogo, C. M., Githure, J. I., and Beier, J. C. (2004). Characterization of potential larval habitats for *Anopheles* mosquitoes in relation to urban land-use in Malindi, Kenya. *International Journal of Health Geographics*, 3(1):1–13.

Kelly-Hope, L. A., Hemingway, J., and McKenzie, F. E. (2009). Environmental factors associated with the malaria vectors *Anopheles gambiae* and *Anopheles funestus* in Kenya. *Malaria journal*, 8(1):1–8.

Klein, T. A., Windbichler, N., Deredec, A., Burt, A., and Benedict, M. Q. (2012). Infertility resulting from transgenic IpoI male *Anopheles gambiae* in large cage trials. *Pathogens and Global Health*, 106:20–31.

Kyrou, K., Hammond, A. M., Galizi, R., Kranjc, N., Burt, A., Beaghton, A. K., Nolan, T., and Crisanti, A. (2018). A CRISPR–Cas9 gene drive targeting *doublesex* causes complete population suppression in caged *Anopheles gambiae* mosquitoes. *Nature Biotechnology*, 36:1062–1066.

Lanzaro, G. C., Campos, M., Crepeau, M., Cornel, A., Estrada, A., Gripkey, H., Haddad, Z., Kormos, A., Palomares, S., and Sharpee, W. (2021). Selection of sites for field trials of genetically engineered mosquitoes with gene drive. *bioRxiv*.

Lee, Y., Marsden, C. D., Norris, L. C., Collier, T. C., Main, B. J., Fofana, A., Cornel, A. J., and Lanzaro, G. C. (2013). Spatiotemporal dynamics of gene flow and hybrid fitness between the M and S forms of the malaria mosquito, *Anopheles gambiae*. *Proceedings of the National Academy of Sciences*, 110(49):19854–19859.

Lehmann, T. and Diabate, A. (2008). The molecular forms of *Anopheles gambiae*: A phenotypic perspective. *Infection, Genetics and Evolution*, 8(5):737 – 746.

Lehmann, T., Weetman, D., Huestis, D. L., Yaro, A. S., Kassogue, Y., Diallo, M., Donnelly, M. J., and Dao, A. (2017). Tracing the origin of the early wet-season *Anopheles coluzzii* in the Sahel. *Evolutionary Applications*, 10:704–717.

Lindsay, S. W., Parson, L., and Thomas, C. J. (1998). Mapping the range and relative abundance of the two principal African malaria vectors, *Anopheles gambiae* sensu stricto and *An. arabiensis*, using climate data. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1399):847–854.

Lunde, T. M., Korecha, D., Loha, E., Sorteberg, A., and Lindtjørn, B. (2013). A dynamic model of some malaria-transmitting anopheline mosquitoes of the Afrotropical region. I. Model description and sensitivity analysis. *Malaria Journal*, 12(23342980):28–28.

Magombedze, G., Ferguson, N. M., and Ghani, A. C. (2018). A trade-off between dry season survival longevity and wet season high net reproduction can explain the persistenceof *Anopheles* mosquitoes. *Parasites & Vectors*, 11:576.

Munga, S., Yakob, L., Mushinzimana, E., Zhou, G., Ouna, T., Minakawa, N., Githeko, A., and Yan, G. (2009). Land use and land cover changes and spatiotemporal dynamics of anopheline larval habitats during a four-year period in a highland community of Africa. *The American journal of tropical medicine and hygiene*, 81(6):1079.

Nash, A., Urdaneta, G. M., Beaghton, A. K., Hoermann, A., Papathanos, P. A., Christophides, G. K., and Windbichler, N. (2019). Integral gene

drives for population replacement. *Biology Open*, 8(1):bio037762.

Noble, C., Adlam, B., Church, G. M., Esvelt, K. M., and Nowak, M. A. (2018). Current CRISPR gene drive systems are likely to be highly invasive in wild populations. *eLife*, 7:e33423.

Nolan, T. (2021). Control of malaria-transmitting mosquitoes using gene drives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 376(1818):20190803.

North, A., Burt, A., and Godfray, H. C. J. (2013). Modelling the spatial spread of a homing endonuclease gene in a mosquito population. *Journal of Applied Ecology*, 50(5):1216–1225.

North, A. R., Burt, A., and Godfray, H. C. J. (2019). Modelling the potential of genetic control of malaria mosquitoes at national scale. *BMC Biology*, 17(1):26.

North, A. R. and Godfray, H. C. J. (2018). Modelling the persistence of mosquito vectors of malaria in Burkina Faso. *Malaria Journal*, 17(1):140.

Paaijmans, K. P., Huijben, S., Githeko, A. K., and Takken, W. (2009). Competitive interactions between larvae of the malaria mosquitoes *Anopheles arabiensis* and *Anopheles gambiae* under semi-field conditions in western Kenya. *Acta Tropica*, 109(2):124–130.

Pombi, M., Kengne, P., Gimonneau, G., Tene-Fossog, B., Ayala, D., Kamdem, C., Santolamazza, F., Guelbeogo, W. M., Sagnon, N., Petrarca, V., et al. (2017). Dissecting functional components of reproductive isolation among closely related sympatric species of the *Anopheles gambiae* complex. *Evolutionary Applications*, 10(10):1102–1120.

Price, T. A., Windbichler, N., Unckless, R. L., Sutter, A., Runge, J.-N., Ross, P. A., Pomiankowski, A., Nuckolls, N. L., Montchamp-Moreau, C., Mideo, N., et al. (2020). Resistance to natural and synthetic gene drive systems. *Journal of Evolutionary Biology*.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sanogo, Z. L., Yaro, A. S., Dao, A., Diallo, M., Yossi, O., Samaké, D., Krajacich, B. J., Faiman, R., and Lehmann, T. (2020). The effects of high-altitude windborne migration on survival, oviposition, and blood-feeding of the African malaria mosquito, *Anopheles gambiae* s.l. (Diptera: Culicidae). *Journal of Medical Entomology*, 58(1):343–349.

Selvaraj, P., Wenger, E. A., Bridenbecker, D., Windbichler, N., Russell, J. R., Gerardin, J., Bever, C. A., and Nikolov, M. (2020). Vector genetics, insecticide resistance and gene drives: An agent-based modeling approach to evaluate malaria transmission and elimination. *PLOS Computational Biology*, 16(8):e1008121.

Service, M. (1997). Mosquito (diptera: Culicidae) dispersal—the long and short of it. *Journal of medical entomology*, 34(6):579–588.

Simard, F., Ayala, D., Kamdem, G. C., Pombi, M., Etouna, J., Ose, K., Fotsing, J.-M., Fontenille, D., Besansky, N. J., and Costantini, C. (2009). Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. *BMC Ecology*, 9(1):17.

Sinka, M. E., Bangs, M. J., Manguin, S., Coetzee, M., Mbogo, C. M., Hemingway, J., Patil, A. P., Temperley, W. H., Gething, P. W., Kabaria, C. W., Okara, R. M., Van Boeckel, T., Godfray, H. C. J., Harbach, R. E., and Hay, S. I. (2010). The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. *Parasites & Vectors*, 3(1):117.

Sinka, M. E., Bangs, M. J., Manguin, S., Rubio-Palis, Y., Chareonviriyaphap, T., Coetzee, M., Mbogo, C. M., Hemingway, J., Patil, A. P., Temperley, W. H., Gething, P. W., Kabaria, C. W., Burkot, T. R., Harbach, R. E., and Hay, S. I. (2012). A global map of dominant malaria vectors. *Parasites & Vectors*, 5(1):69.

Sinka, M. E., Golding, N., Massey, N. C., Wiebe, A., Huang, Z., Hay, S. I., and Moyes, C. L. (2016). Modelling the relative abundance of the primary african vectors of malaria before and after the implementation of indoor, insecticide-based vector control. *Malaria Journal*, 15(1):142.

Snee, R. (1977). Validation of regression models: Methods and examples. *Technometrics*.

Tanaka, H., Stone, H. A., and Nelson, D. R. (2017). Spatial gene drives and pushed genetic waves. *Proceedings of the National Academy of Sciences*, 114(32):8452–8457.

Tene Fossog, B., Ayala, D., Acevedo, P., Kengne, P., Ngomo Abeso Mebuy, I., Makanga, B., Magnus, J., Awono-Ambene, P., Njiokou, F., Pombi, M., Antonio-Nkondjio, C., Paupy, C., Besansky, N. J., and Costantini, C. (2015). Habitat segregation and ecological character displacement in cryptic African malaria mosquitoes. *Evol Appl*, 8(4):326–345.

Touré, Y., Petrarca, V., Traoré, S., Coulibaly, A., Maiga, H., Sankaré, O., Sow, M., Di Deco, M., and Coluzzi, M. (1998). The distribution and inversion polymorphism of chromosomally recognized taxa of the *Anopheles gambiae* complex in Mali, West Africa. *Parassitologia*, 40(4):477—511.

Touré, Y. T., Petrarca, V., Traoré, S. F., Coulibaly, A., Maïga, H. M., Sankaré, O., Sow, M., Di Deco, M. A., and Coluzzi, M. (1994). Ecological genetic studies in the chromosomal form Mopti of *Anopheles gambiae* s.str. in Mali, West Africa. *Genetica*, 94(2):213–223.

Wang, G.-H., Gamez, S., Raban, R. R., Marshall, J. M., Alphey, L., Li, M., Rasgon, J. L., and Akbari, O. S. (2021). Combating mosquito-borne diseases using genetic control technologies. *Nature Communications*, 12(1):4388.

White, M. T., Griffin, J. T., Churcher, T. S., Ferguson, N. M., Basáñez, M.-G., and Ghani, A. C. (2011). Modelling the impact of vector control interventions on *Anopheles gambiae* population dynamics. *Parasites & Vectors*, 4(1):153.

Wiebe, A., Longbottom, J., Gleave, K., Shearer, F. M., Sinka, M. E., Massey, N. C., Cameron, E., Bhatt, S., Gething, P. W., Hemingway, J., Smith, D. L., Coleman, M., and Moyes, C. L. (2017). Geographical distributions of African malaria vector sibling species and evidence for insecticide resistance. *Malaria Journal*, 16(1):85.

Wilson, A. L., Courtenay, O., Kelly-Hope, L. A., Scott, T. W., Takken, W., Torr, S. J., and Lindsay, S. W. (2020). The importance of vector control for the control and elimination of vector-borne diseases. *PLOS Neglected Tropical Diseases*, 14:e0007831.

Windbichler, N., Papathanos, P. A., and Crisanti, A. (2008). Targeting the X chromosome during spermatogenesis induces Y chromosome transmission ratio distortion and early dominant embryo lethality in *Anopheles gambiae*. *PLOS Genetics*, 4:e1000291.

Wondji, C., Simard, F., Petrarca, V., Etang, J., Santolamazza, F., Torre, A. D., and Fontenille, D. (2005). Species and populations of the *Anopheles gambiae* complex in Cameroon with special emphasis on chromosomal and molecular forms of *Anopheles gambiae* s.s. *Journal of Medical Entomology*, 42(6):998–1005.

World Health Organisation (2020). World Malaria Report 2020. Technical report, World Health Organisation, Geneva, Switzerland, 299 pp.

Wu, S. L., Bennett, J. B., Sánchez C., H. M., Dolgert, A. J., León, T. M., and Marshall, J. M. (2020). MGDrivE 2: A simulation framework for gene drive systems incorporating seasonality and epidemiological dynamics. *bioRxiv*, page 2020.10.16.343376.

Wyse, A. P., Santos, A. J. B., Azevedo, J. S., Lima, J. S., and Faria, J. R. (2018). Modeling the spreading and interaction between wild and transgenic mosquitoes with a random dispersal. *PLOS One*, 13(10):e0205879.

Supplementary Materials

S1: Algorithm for placing introduction sites

As mentioned in the main text, we first select the cells in which at least 10,000 of either subspecies are available year-round on the African mainland. We then randomly select a set Y consisting of 5,000 of those cells as site candidates for ease of computation, discarding the others from consideration. We adapt an algorithm that maximises the probability of animals being captured by a set of traps spread across a domain using gradient descent.

We define a goodness of fit $G(X)$ of a candidate set of 10 release sites X by:

- defining a kernel function $f(d) = \exp(-d/1000)$ where d is distance in km (analogous to probability of detection), then
- for each candidate site $y \in Y$, calculate $p = 1 - \prod_{d(X,y)} (1 - f(d))$ where $d(X, y)$ is the set of distances between each point in X and the candidate y (analogous to the probability of detection by at least one site in X),
- taking $G(X)$ as the sum of p over all the candidate sites (analogous to the expected number of candidate cells in Y detected by the sites in X).

For 20 iterations, we:

- initially select X to be 10 sites selected at random from Y, and
- calculate $G(X)$ for this initial X.
- For each of the currently selected release sites $x \in X$:
  - pick the 20 spatially closest candidates to x and recalculate $G(X)$ for each, temporarily replacing x with the nearby candidate in the set X; and
  - permanently replace x with the candidate that gives the highest $G(X)$ in the set X, if higher than the current value.
- If the set X has changed after going through all of the release sites, then repeat the process.
- Otherwise, stop, outputting the final set X and final goodness of fit $G(X)$.

Of the 20 output sets X, we then use the set that gave the highest value for $G(X)$.

S2: Neural network details

The inputs of the neural network were normalised by the mean and standard deviation of each predictor. The network was given three hidden layers of size 40, 20 and 10 (arrived at by experimentation; these sizes gave near-optimal validated errors: see Table 3); a rectified linear (ReLu) activation function for the hidden layers, and a sigmoid activation function for the output neuron to produce a probability of the predictors producing a mosquito of *An. gambiae* s.s. as opposed to *An. coluzzii*. We use a cross-entropy loss function, such that the neural network effectively finds the maximum likelihood of the neuron weights given the observed subspecies at each site. The neural network is implemented using the Keras package (Chollet et al., 2017) in the R programming language (R Core Team, 2020). We use a 75:25 training-testing split on the data, using the DUPLEX algorithm (Snee, 1977) to ensure a spread of geographic locations for both training and testing sets. We then run the neural network for 200 epochs, and use the minimum validated loss to determine the number of epochs to subsequently run the model for.

For each tested set of predictors in the forward selection process, we train the neural network 500 times, running for 200 epochs each time. We calculate the validation loss after each epoch, take the minimum as our measure of goodness-of-fit for that trained network, and take the mean of the 500 validation losses obtained in this way. This is to account for variations in neural network effectiveness caused by the stochastic gradient descent process. Starting with an empty

set of predictors, the predictor whose inclusion gives the best mean validation loss is accepted, until accepting new predictors no longer reliably decreases the validation loss.

We then train a neural network on the final selected set of predictors 100 times, taking the pointwise mean of the 100 runs as our final estimate of relative abundance at each location. This model averaging is performed in order to ensure that our final ensemble estimate is representative of the neural network modelling process as a whole: as the optimisation process is stochastic, a single network may differ substantially from another, especially where little data is available.

S3: Illustrative animations

GIF animations of the simultaneous introduction of the construct at all sites are available in Supplementary Files: `wind9hF.ag.1.gif`, `wind9hF.ac.1.gif`, `wind2hF.ag.1.gif`, `wind2hF.ac.1.gif` show the animations for 9 hour wind advection for *Anopheles gambiae* s.s. and *Anopheles coluzzii*, and the 2 hour equivalents for both subspecies, respectively. The release points are marked as pink squares. Cells are colour coded: the amount of red in a cell represents the relative number of mosquitoes of the given subspecies with a wildtype w allele. Green and blue similarly represent construct c and resistance r respectively.

Figures

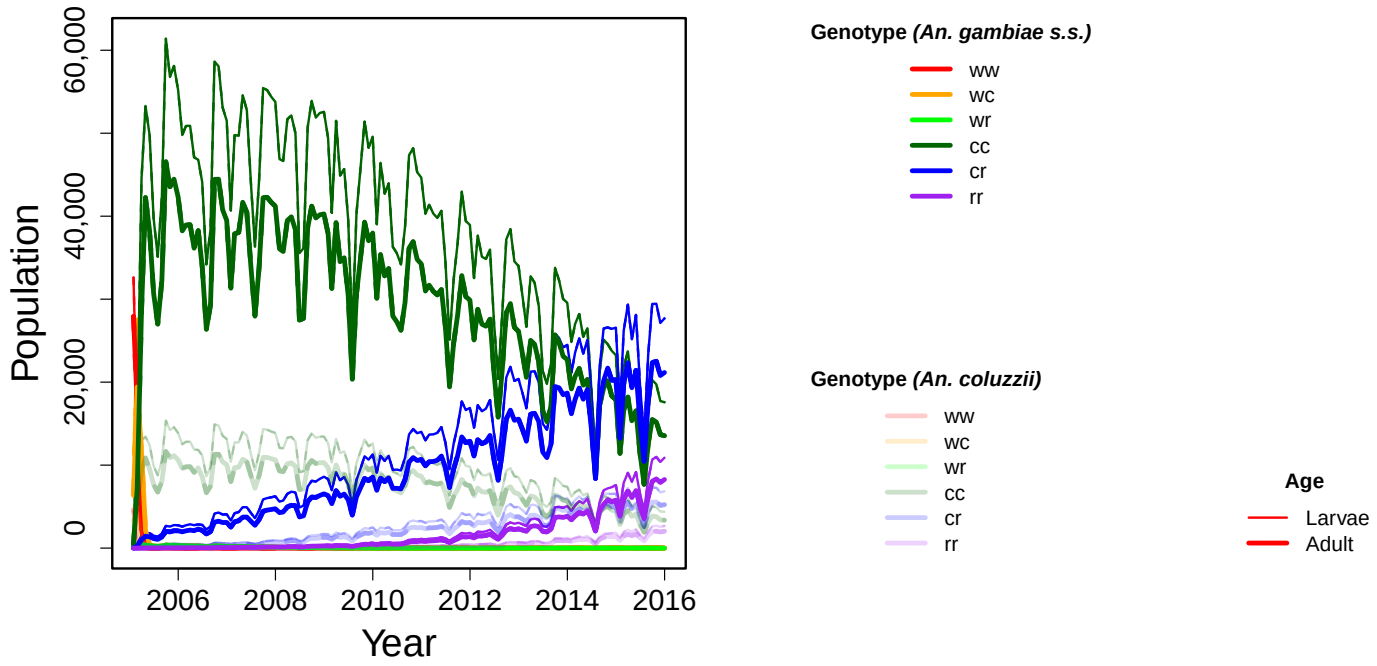


Figure S1: Time series plot of Site 6 **(as in Figure 5)** but with a closed population, i.e. no diffusion or advection. The colours correspond to genotype and the line thickness to age class.

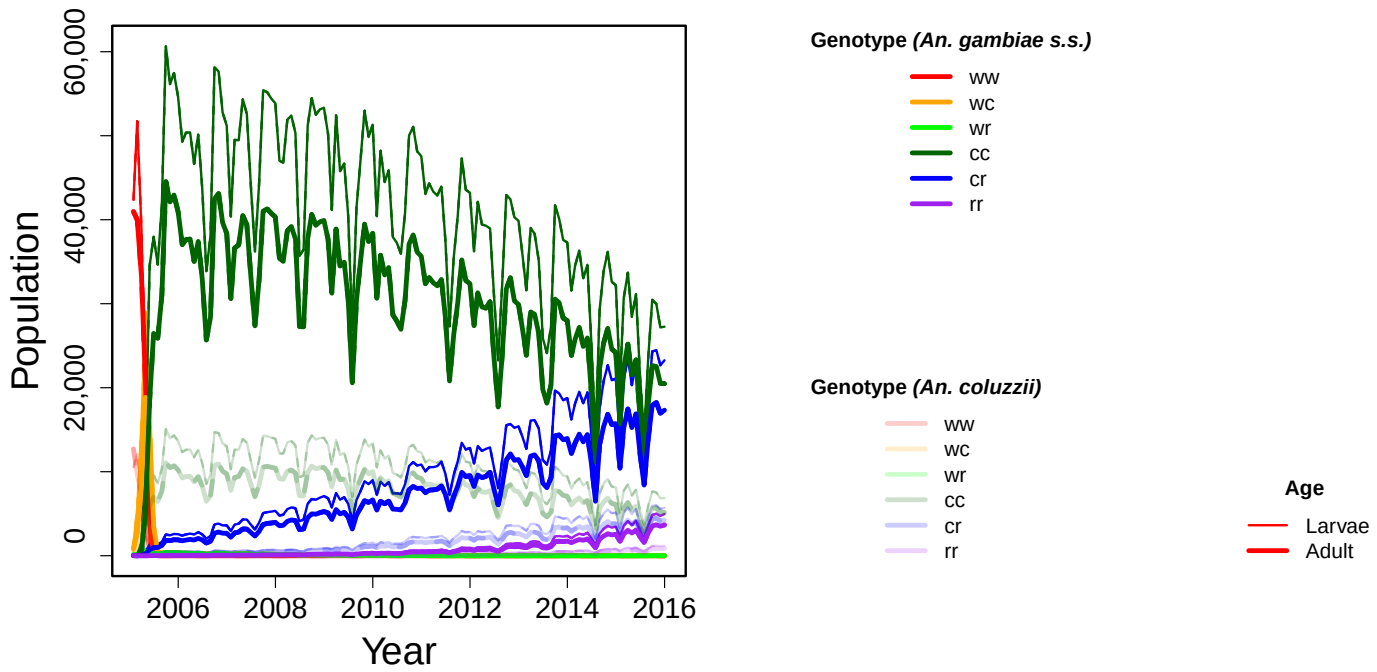


Figure S2: Time series plot of Site 6 **as in Figure S1** but with 6 age classes, keeping overall larval emergence period and mortality constant. This result is more closely analogous to a fixed larval emergence time than to the exponentially-distributed version in the main model. The first five age classes combined are plotted here under "larvae". The colours correspond to genotype and the line thickness to age class.

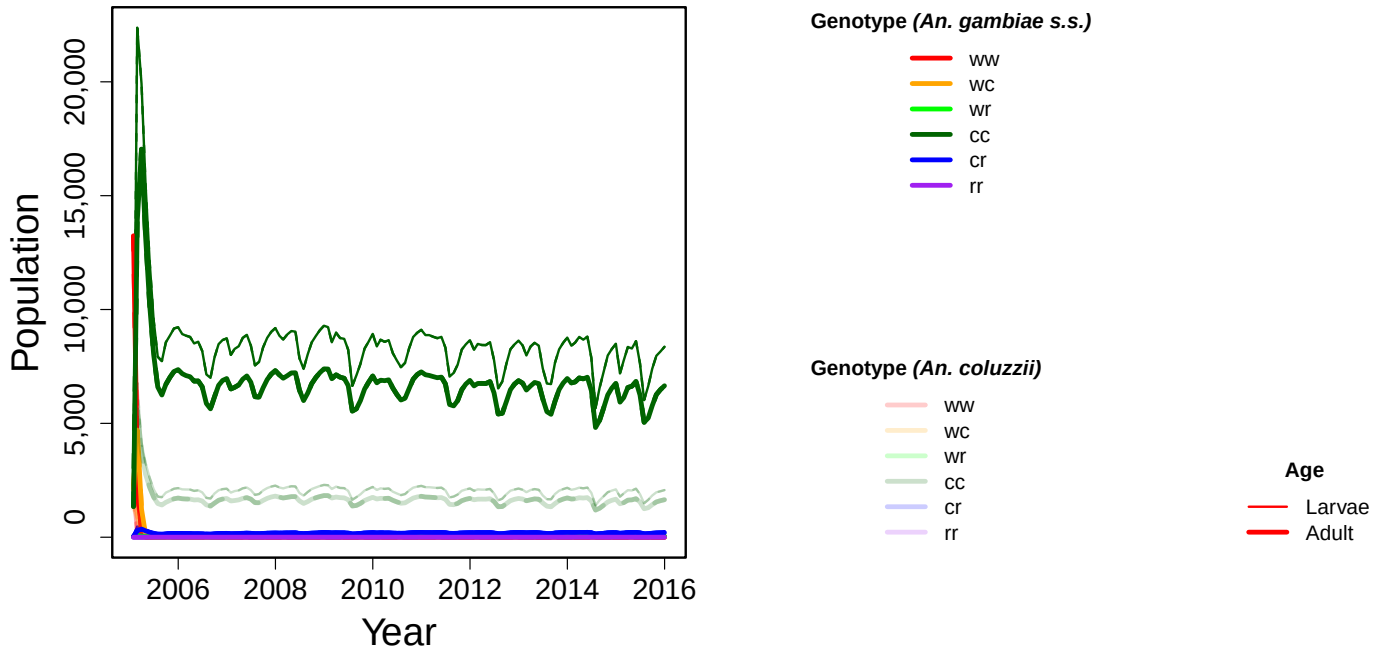


Figure S3: Time series plot of Site 6 **as in Figure S1** but with $R(g_M, g_F) = 0.05$ where $g_M \neq ww$ or $g_F \neq ww$. The colours correspond to genotype and the line thickness to age class.

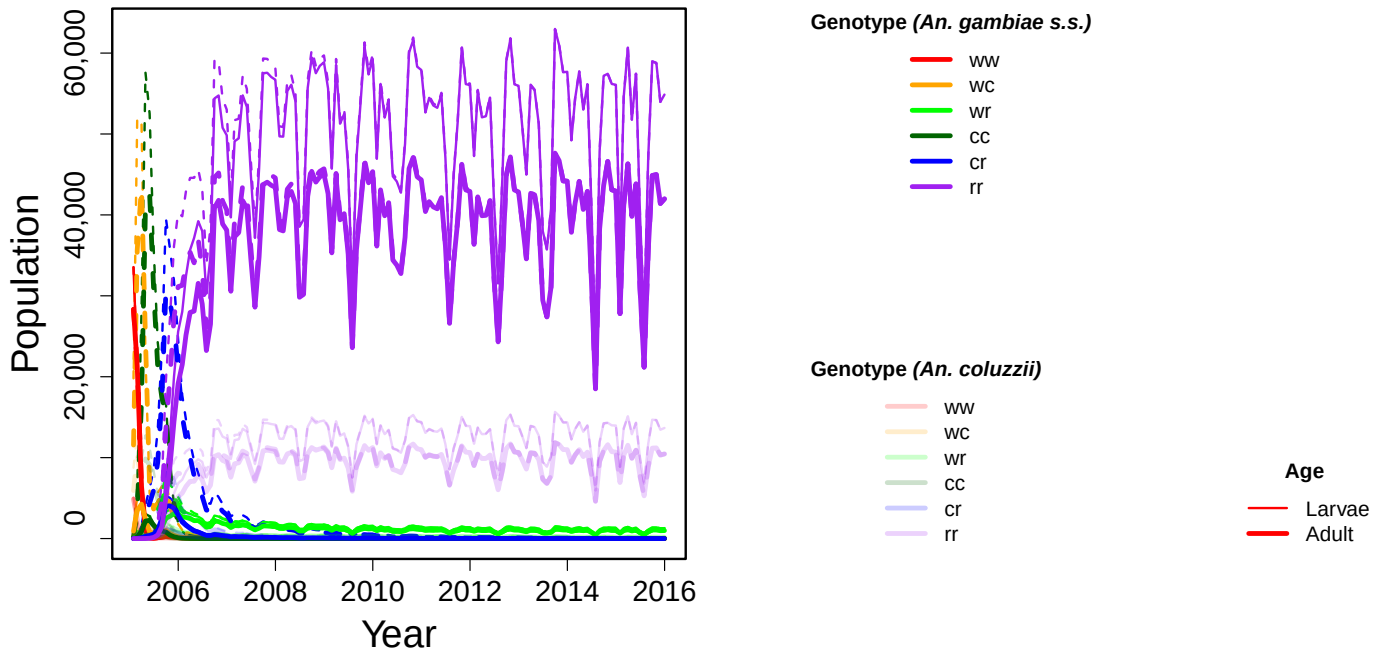


Figure S4: Time series plot of Site 6 **as in Figure S1** but with $p(g_M, g_F, s)$ modified so that any presence of the genetic construct in the father (wc, cc or cr) results in a 95% male sex bias in offspring. Dashed lines represent males, solid lines females. The colours correspond to genotype and the line thickness to age class.

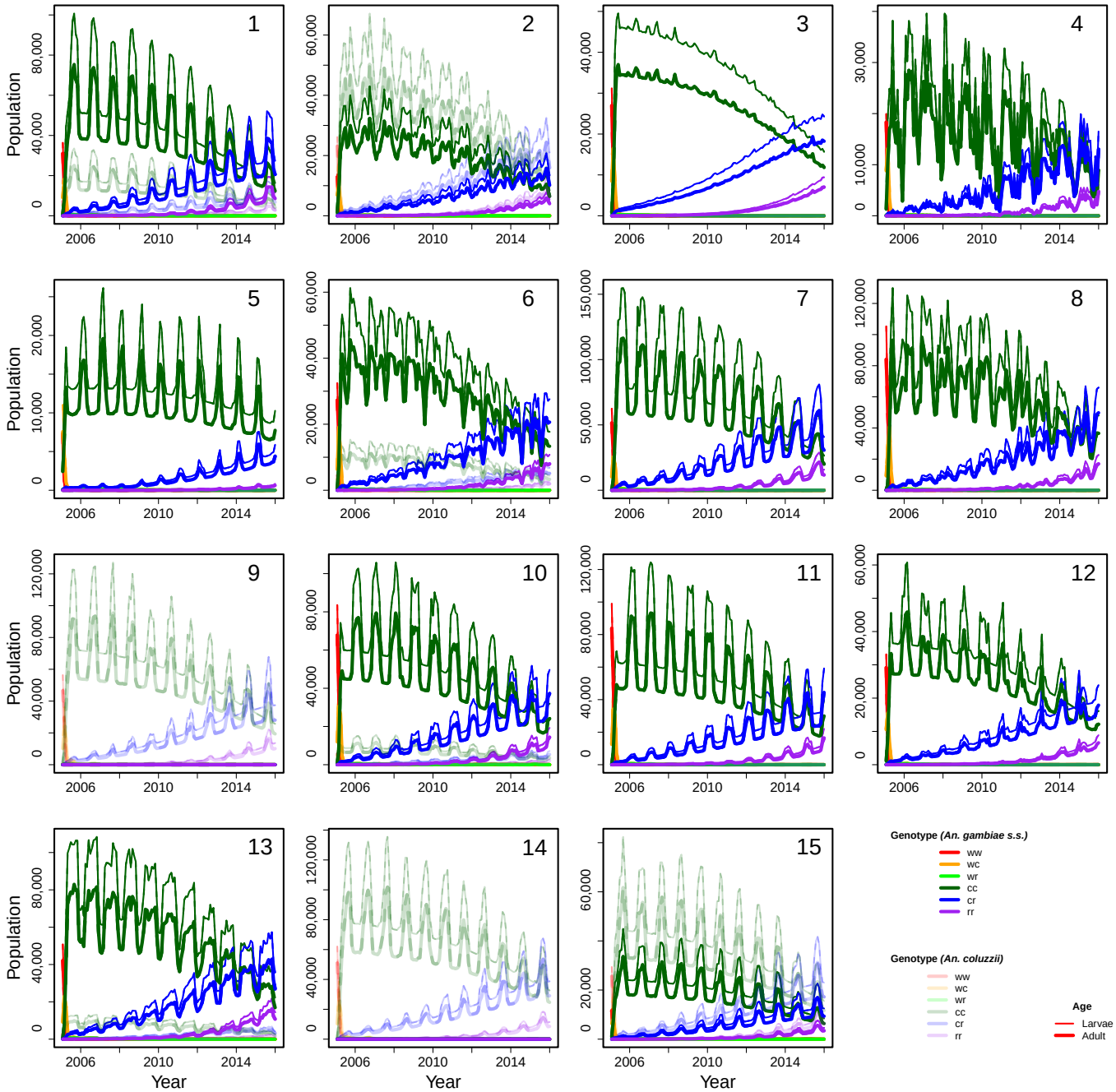


Figure S5: The time series abundance of male mosquitoes at each introduction point **as in Figure 5, but using 2 hours instead of 9 hours advection**, separated by species, genotype and age (female mosquitoes occur in identical numbers to males in this model). The colours correspond to genotype and the line thickness to age class.

.

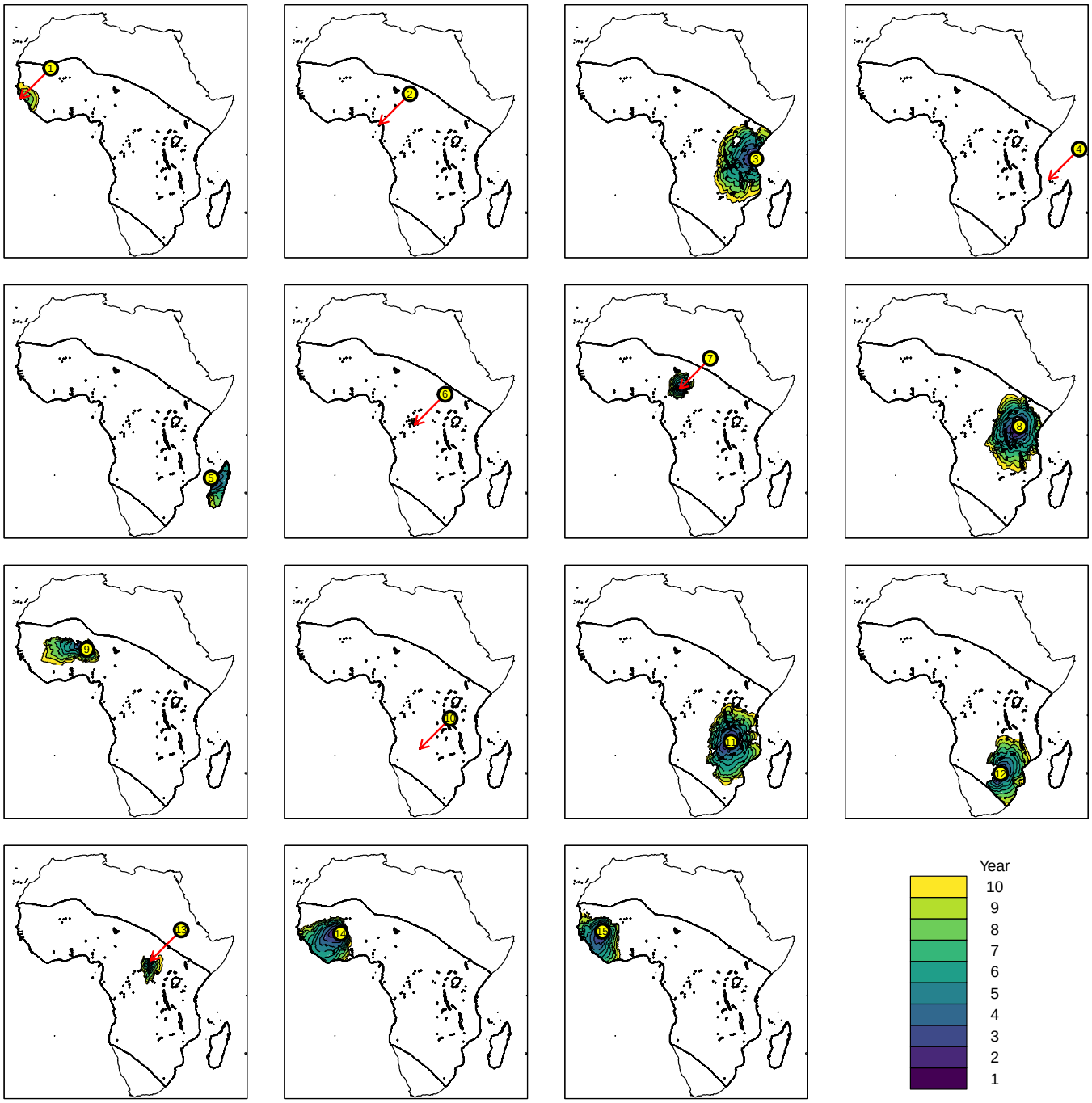


Figure S6: The invasion front of the construct at each introduction point **as in Figure 4, but using 2 hours instead of 9 hours advection**. A separate colour is given for each year. The island introductions are (1) the Bijagós islands (off Guinea-Bissau), (2) Bioko (off Cameroon), (3) Zanzibar (off Tanzania), (4) Comoros (off Mozambique) and (5) Madagascar.

.

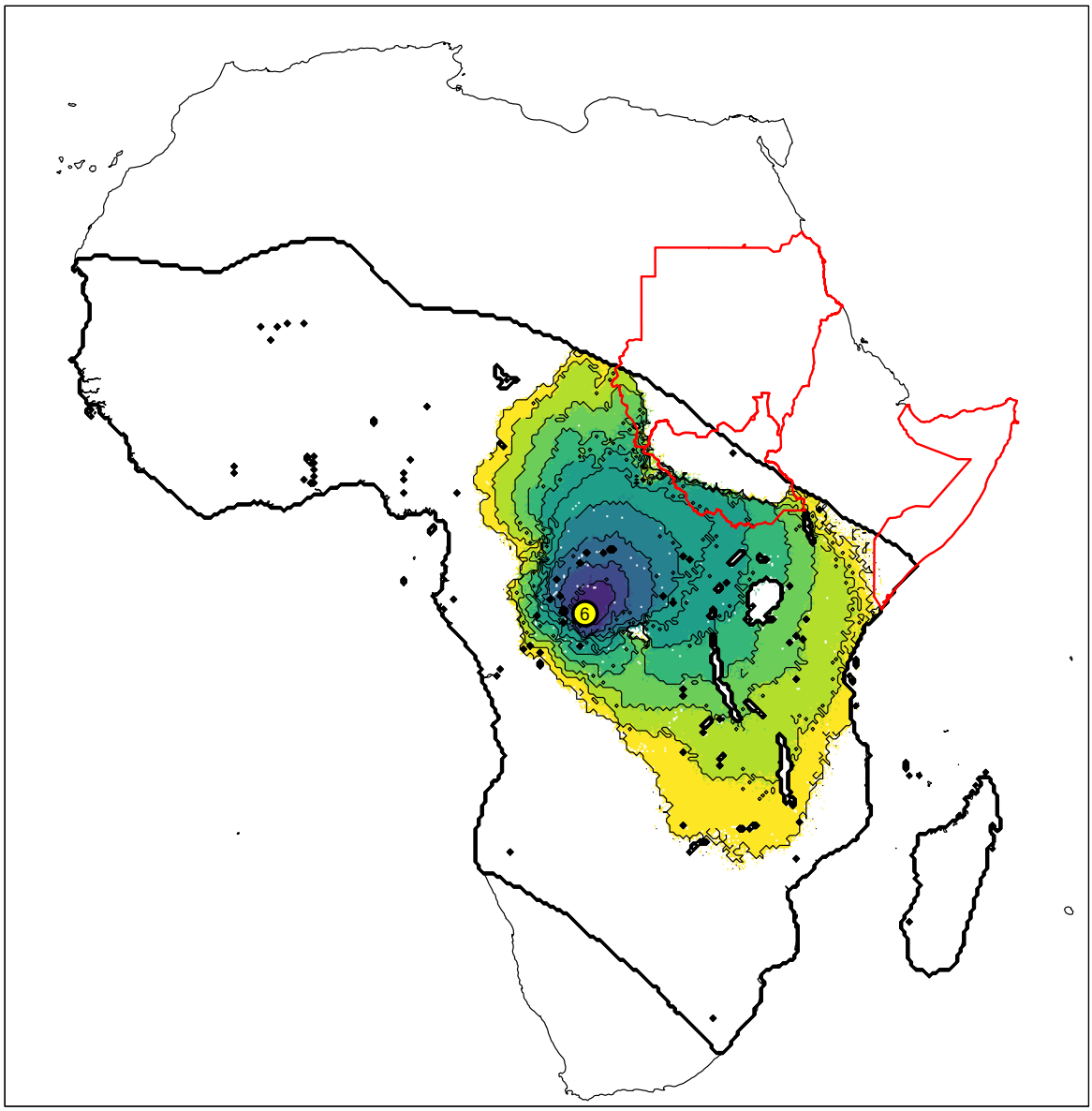


Figure S7: The invasion front of the construct as in Figure 4 for Site 6, but assuming human (and thus mosquito) absence in Sudan, South Sudan and Somalia (national boundaries given in red, as defined by the UN Office for the Coordination of Humanitarian Affairs).

.