scientific reports

Check for updates

OPEN

Developing a predictive model for an emerging epidemic on cassava in sub-Saharan Africa

David Godding[1]✉, Richard O. J. H. Stutt[1], Titus Alicai[2], Phillip Abidrabo[2], Geoffrey Okao-Okuja[2] & Christopher A. Gilligan[1]

The agricultural productivity of smallholder farmers in sub-Saharan Africa (SSA) is severely constrained by pests and pathogens, impacting economic stability and food security. An epidemic of cassava brown streak disease, causing significant yield loss, is spreading rapidly from Uganda into surrounding countries. Based on sparse surveillance data, the epidemic front is reported to be as far west as central DRC, the world's highest per capita consumer, and as far south as Zambia. Future spread threatens production in West Africa including Nigeria, the world's largest producer of cassava. Using innovative methods we develop, parameterise and validate a landscape-scale, stochastic epidemic model capturing the spread of the disease throughout Uganda. The model incorporates real-world management interventions and can be readily extended to make predictions for all 32 major cassava producing countries of SSA, with relevant data, and lays the foundations for a tool capable of informing policy decisions at a national and regional scale.

A principal challenge in dealing with emerging epidemics and pest infestations of agricultural crops is to estimate the current extent of infection and the rates of spread across heterogeneous landscapes. The challenges are particularly acute for epidemics that impact smallholder agriculture in sub-Saharan Africa (SSA), where many staple crops are under threat from emerging pests and pathogens. Examples include maize lethal necrosis[1], fall armyworm[2], banana bunchy top disease[3], wheat rusts[4,5], cassava viruses[6], and desert locust[7].

Cassava is the second most important source of calories in SSA after maize[8]. Cassava brown streak disease (CBSD) is caused by cassava brown streak ipomoviruses (CBSIs); ssRNA *Ipomoviruses* of the family *Potyviridae* that are epidemiologically equivalent. The disease poses one of the most significant threats to cassava production in SSA, causing necrosis of the edible root tissue[9,10]. The CBSIs are spread by an insect vector, the whitefly *Bemisia tabaci*[11], with additional spread by trade movement of virus-infected cuttings used for planting[12].

Following an initial report of CBSD in Uganda in 2004[13], the disease rapidly spread throughout Uganda[14] to surrounding countries, including Rwanda[15], Burundi[16], western Kenya[17], lake-zone Tanzania[18], eastern DRC[19], and Zambia[20]. The disease has more recently been reported as far west as the northcentral province of Tshopo, DRC[21]. Continued westward spread and the risk of direct introduction via the movement of planting material poses a major threat to food security and economic stability in Central and West African countries.

An initial step in predicting the onward spread of the pathogen is to estimate transmission and dispersal parameters at landscape scales. We do this by fitting and validating a stochastic, spatially explicit metapopulation epidemic model of CBSD spread in Uganda, at a 1 km^2 resolution, using a unique multi-year country-wide surveillance dataset[14] that documents the progressive spread of CBSD throughout Uganda from a few isolated initial infected sites near Kampala. The model takes account of the spatial distribution and connectedness of the cassava crop and variability in the abundance of the insect vector throughout Uganda. Estimating parameters from sparse spatiotemporal data is extremely challenging and an area of active research. We use approximate Bayesian computation (ABC), which does not require the explicit definition of the likelihood and is well adapted to dealing with unobserved data[22] when inferring sequences of infection spread across a heterogeneous landscape. However, a central challenge is specifying summary statistics that capture as much information as possible about the dynamics of the system in the simplest possible form[22–24].

Specifically, we address the following questions considering epidemic spread of CBSD:

- Can a simple epidemiological model structure capture the fundamental dynamics of the CBSD epidemic?

[1]Epidemiology and Modelling Group, Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, UK. [2]National Crops Resources Research Institute, P. O. Box 7084, Kampala, Uganda. ✉email: dsg38@cam.ac.uk

- Can transmission rates and dispersal parameters be estimated from disjoint snapshots of annual surveillance data?
- Can the parameterised model predict the future spread of the virus in successive years within Uganda?

Results

Incorporating data-driven model layers. A spatially explicit, stochastic SI metapopulation epidemic model provided the framework for the spread of infection, and by implication disease, within and between rasterised cells in the landscape. A data-driven host landscape layer was generated to account for the spatial heterogeneity of cassava production. The host landscape layer was derived by converting the CassavaMap model[25] from production volume in tonnes per km^2 to the number of fields per km^2 (Fig. 1A).

Through an iterative process of model development (Supplementary Methods S1.1), the model was extended to incorporate an additional epidemiologically important data-driven spatial layer accounting for the variation in the abundance of the vector, *B. tabaci*, across the Ugandan landscape (Fig. 1B). The rasterised vector abundance layer was generated from the *B. tabaci* count data collected as part of cassava field surveys[14].

Parameterising the spatiotemporal epidemic dynamics. The Ugandan surveillance data, d_{real}, were divided into two distinct datasets. The training dataset, d_{real}^{fit} consisting of data from 2004 to 2010 inclusive, and validation dataset, d_{real}^{val}, consisting of the remaining data from 2011 to 2017. We applied ABC rejection to estimate three model parameters using d_{real}^{fit}: a dispersal kernel exponent, α, a transmission rate, β, and the proportion of dispersed inoculum that remains in the source cell, p. The posterior probability distribution for these parameters was calculated as the number of simulations generating simulated surveillance data, d_{sim}^{fit} sufficiently close to the real-world surveillance data, d_{real}^{fit}, normalised relative to the sampling density (Supplementary Methods S1.2)[24–26].

In order to quantify the distance between the real-world training data, d_{real}^{fit}, and the simulated surveillance data covering the same time period, d_{sim}^{fit}, we constructed three epidemiologically informed summary statistics (see "Parameter estimation" in "Methods"). The three summary statistics capture different aspects of the spatiotemporal characteristics of the epidemic: S_{nat} and S_{cen} are based on calculations of the proportion of survey points in different regions that are reported as positive each year. A third statistic, S_{grid}, explicitly tracks the spatial expansion of the epidemic throughout Uganda in terms of the year of first CBSD detection in each cell of a regular grid covering the full extent of Uganda.

The statistic, S_{cen}, is designed to capture the local bulk-up dynamics of the epidemic in a small, densely sampled area in central Uganda surrounding Kampala, with dimensions: $x_{min} = 32.20$, $x_{max} = 33.21$, $y_{min} = 0.09$, $y_{max} = 1.16$ in the WGS 84 coordinate system. The statistic, S_{nat}, covers the remaining non-overlapping spatial extent of Uganda and captures the regional bulk up rate (Fig. 2A). For convenience, we refer to the combination of summary statistics, S_{cen}, S_{nat} and tolerances, ε_{cen}, ε_{nat} as S_{inf} and ε_{inf}, respectively.

The third statistic, S_{grid}, is derived by dividing the latitude/longitude extent of Uganda into a 5×5 grid of quadrats. For a given simulation, the statistic is scored as the proportion of the quadrats where infected fields are detected in either the same year as the real-world surveillance data or ± 1 year either side, or in the case where

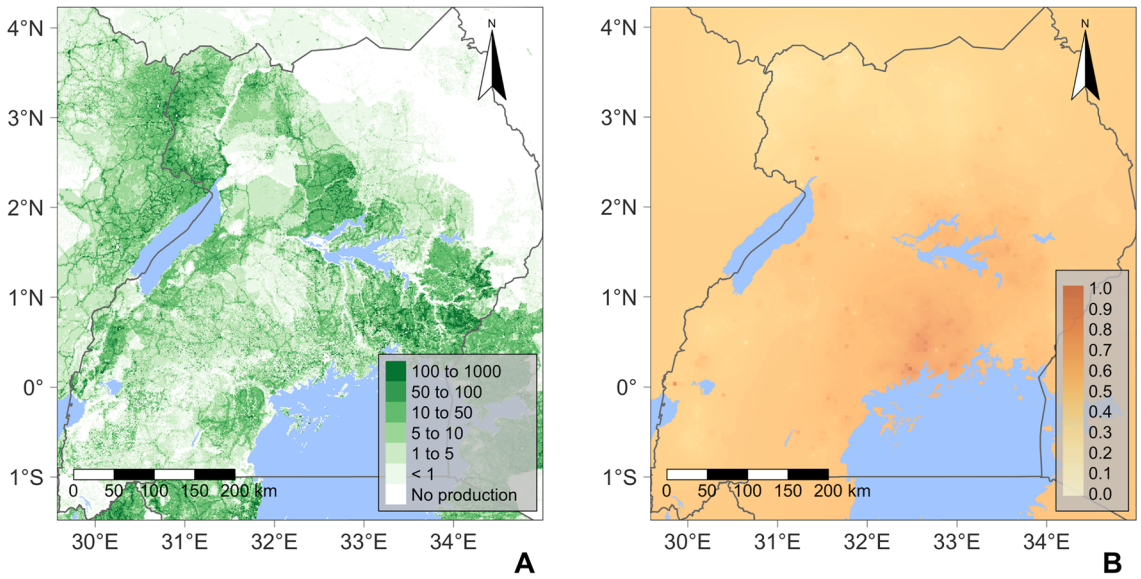

Figure 1. Maps representing the rasterised data-driven model layers: (**A**) the model host landscape, representing the number of cassava fields at a 1 km^2 resolution derived from CassavaMap[25]. (**B**) The vector abundance layer represents the relative abundance of *B. tabaci* at a 5 km resolution derived from the Ugandan CBSD field surveys[14].

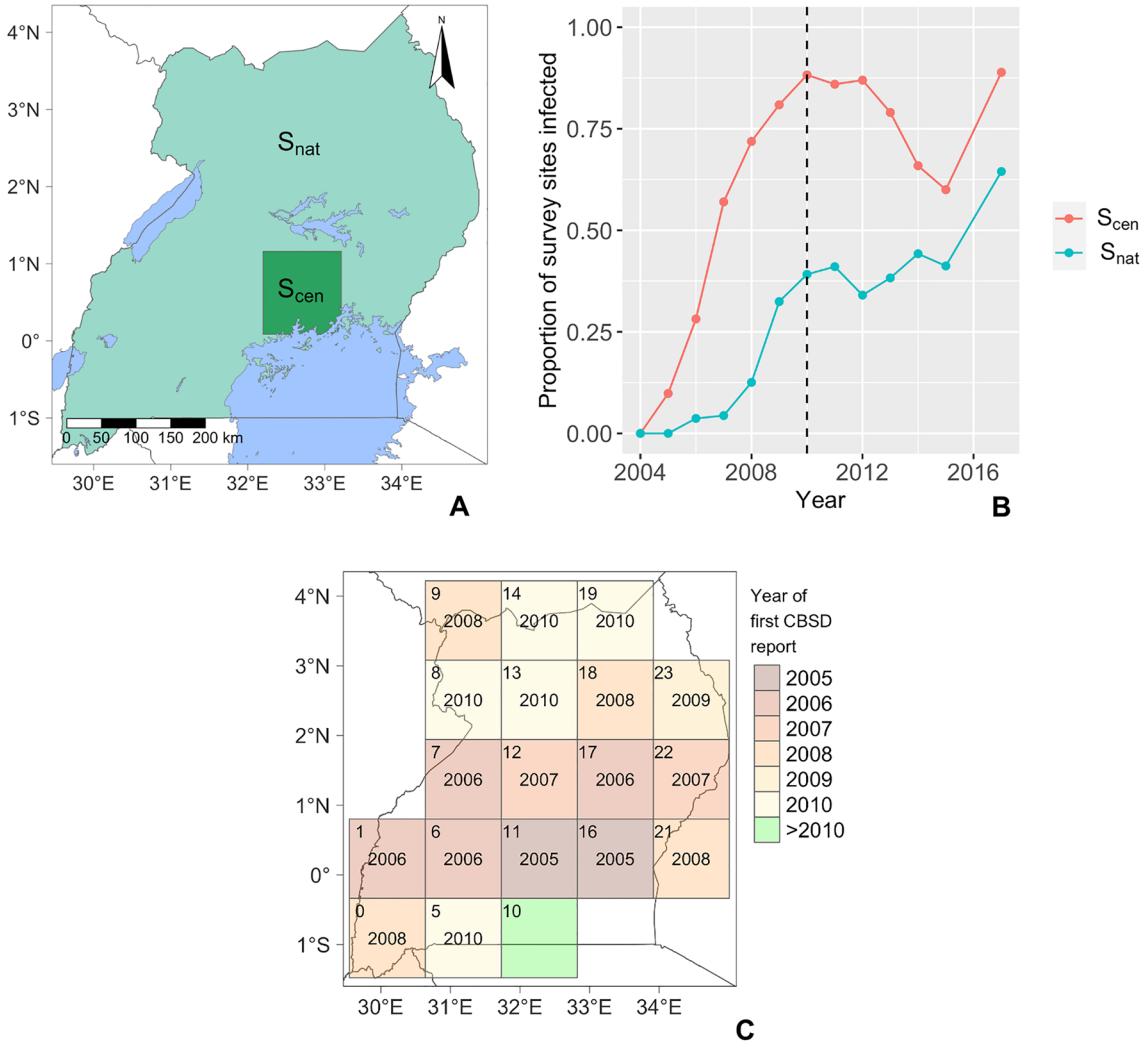

Figure 2. Overview of the three summary statistics used for ABC parameter estimation. Both S_{cen} and S_{nat} statistics were derived by calculating the proportion of survey points in a given year that were reported as positive within a given region: (**A**) represents the two non-overlapping areas of Uganda covered by S_{cen} and S_{nat} and (**B**) summarises the values derived when applying S_{cen} and S_{nat} to the Ugandan national survey data covering the period 2004 to 2017. The dotted black line indicates the divide between the fitting data, d^{fit}_{real}, and the validation data, d^{val}_{real}. (**C**) Overview of the summary statistic S_{grid} highlighting the survey year, up to and including 2010, in which a CBSD infected field was first detected in a given quadrat. If no positive surveys were reported prior to 2010, as in the case of quadrat 10, the quadrat is shaded green. If no surveys were carried out prior to 2010, the quadrats have been excluded from the plot. Quadrat indices are shown in the top left corner of each quadrat.

all surveys in a given quadrat were negative for CBSD, the simulation must remain negative in all simulated surveys (Fig. 2C).

The ability of the summary statistics to recover known parameter values was first tested using synthetic data for the spread of CBSD across the Ugandan landscape. Preliminary analyses also showed that the statistics were best used in combination, and guided the selection of appropriate tolerances for each statistic (Supplementary Methods S1.3).

The posterior distribution is derived from 1440 simulations that passed the fitting criteria (i.e., tolerances applied to the three summary statistics) from a total of 233,600 fitting simulations (Fig. 3). For the kernel scale parameter, α, and log of transmission rate, ln(β), the posterior distribution covers a clear and distinct region of highest posterior density, with a correlation between shorter dispersal distances requiring higher transmission rates and vice versa. Within the credible ranges of α and ln(β), the third parameter, p, governing the amount of inoculum that remains in the source cell, lies almost exclusively below 0.8 and is concentrated around 0.12.

Simulating the impact of a disease management programme. The summary statistic S_{cen} highlights a clear but temporary reduction in the proportion of field surveys reporting CBSD from 2013 to 2015

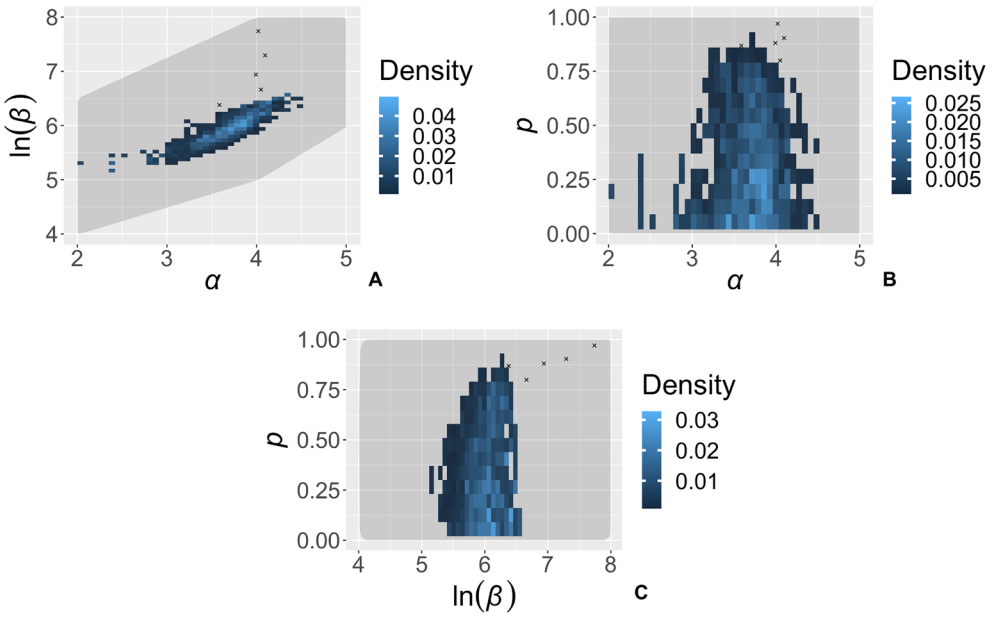

Figure 3. Posterior distribution of the three parameters estimated using the fitting data from 2004 to 2010, d_{real}^{fit}. The parameters are the transmission rate, β, the kernel exponent, α, and the proportion of dispersed inoculum that remains in the source cell, p. The posterior is composed of 1440 fitting simulations that met the fitting criteria out of a trial of 233,600 fitting simulations. Five outliers, indicated by black crosses, were excluded from sparsely sampled parameter space (Supplementary Fig. 7).

in the small central area surrounding Kampala (Fig. 2B). The reduction in the intensity of the epidemic in this region was the result of a number of projects that disseminated a total of 40 million virus-free cassava cuttings, focusing on four Ugandan districts: Luwero, Mukono, Nakasongola, Wakiso (Fig. 4A), with surveillance in each district capturing the same characteristic pattern of temporary decline (Fig. 4B)[27,28]. Beyond this high-level information, specific details on precise location and timing of the different programmes that disseminated clean planting material are not available.

We implemented a process equivalent to the clean seed programmes in the model via three discrete rounds of $I \rightarrow S$ replacement at the start of the 2013, 2014, and 2015 growing seasons. The parameter, r_{clean}, defines the proportion of CBSI infected cassava fields across the four districts that should be replaced by virus-free planting material in each of the three rounds, with the exact fields being selected at random. A value for r_{clean} of 0.15 was selected based on a parameter sweep to identify the value that best fitted the observed impact of the clean seed

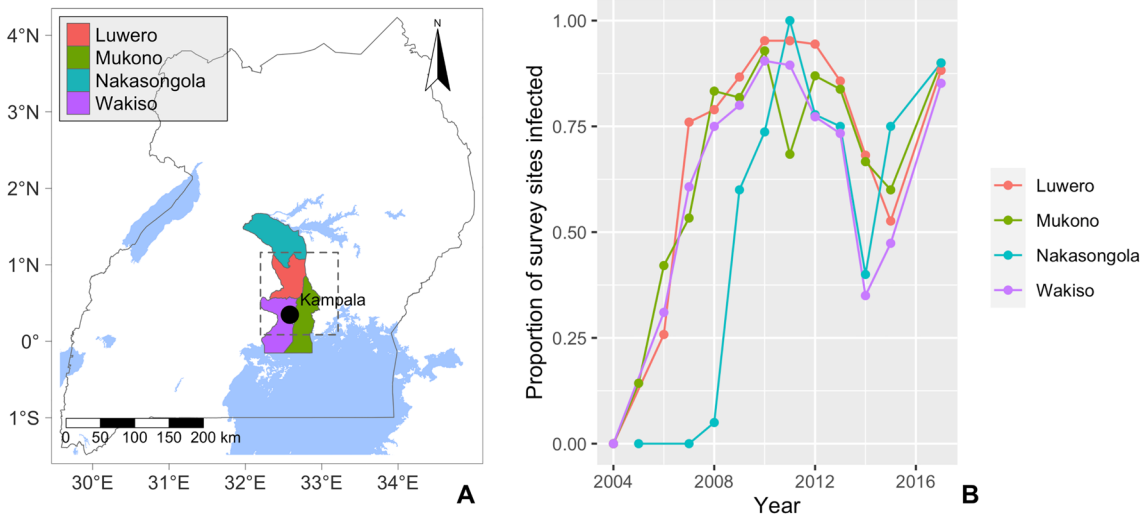

Figure 4. Overview of the districts where clean seed dissemination programmes were primarily carried out: (**A**) shows the location of the districts in Uganda as well as the region covered by S_{cen} (dotted line) and (**B**) summarises the proportion of field surveys that reported CBSD in a given year from each district.

programmes in S_{cen} (Supplementary Methods S1.1.4). The clear correspondence of the simulated clean seed programme to the real-world observations, viewed through the lens of the summary statistic, S_{cen}, is illustrated during model validation.

Validating predictions of epidemic spread. An ensemble of 10,000 simulations were run from 2004 to 2010 by sampling from the posterior parameter distribution to generate initial conditions for the validation simulations. Of these 10,000 simulations, 65 met the fitting criteria, which were then used as initial conditions for the system state at the start of 2011. These simulations were then run for the validation period of 2011–2017 and their correspondence to the validation data, d_{real}^{val}, was assessed by applying the validation criteria (i.e., tolerances applied to the summary statistics, S_{inf}, during the validation period). All 65 validation simulations passed the validation criteria, resulting in a validation score of 100% (Fig. 5).

Figure 6 illustrates the spatial structure of the simulated survey data from a single validation simulation, illustrating the strong yearly correspondence between simulated and real-world surveillance from 2005 to 2017 in terms of the spatial distribution of fields reported as present/absent for CBSD and the local density of CBSD positive surveys.

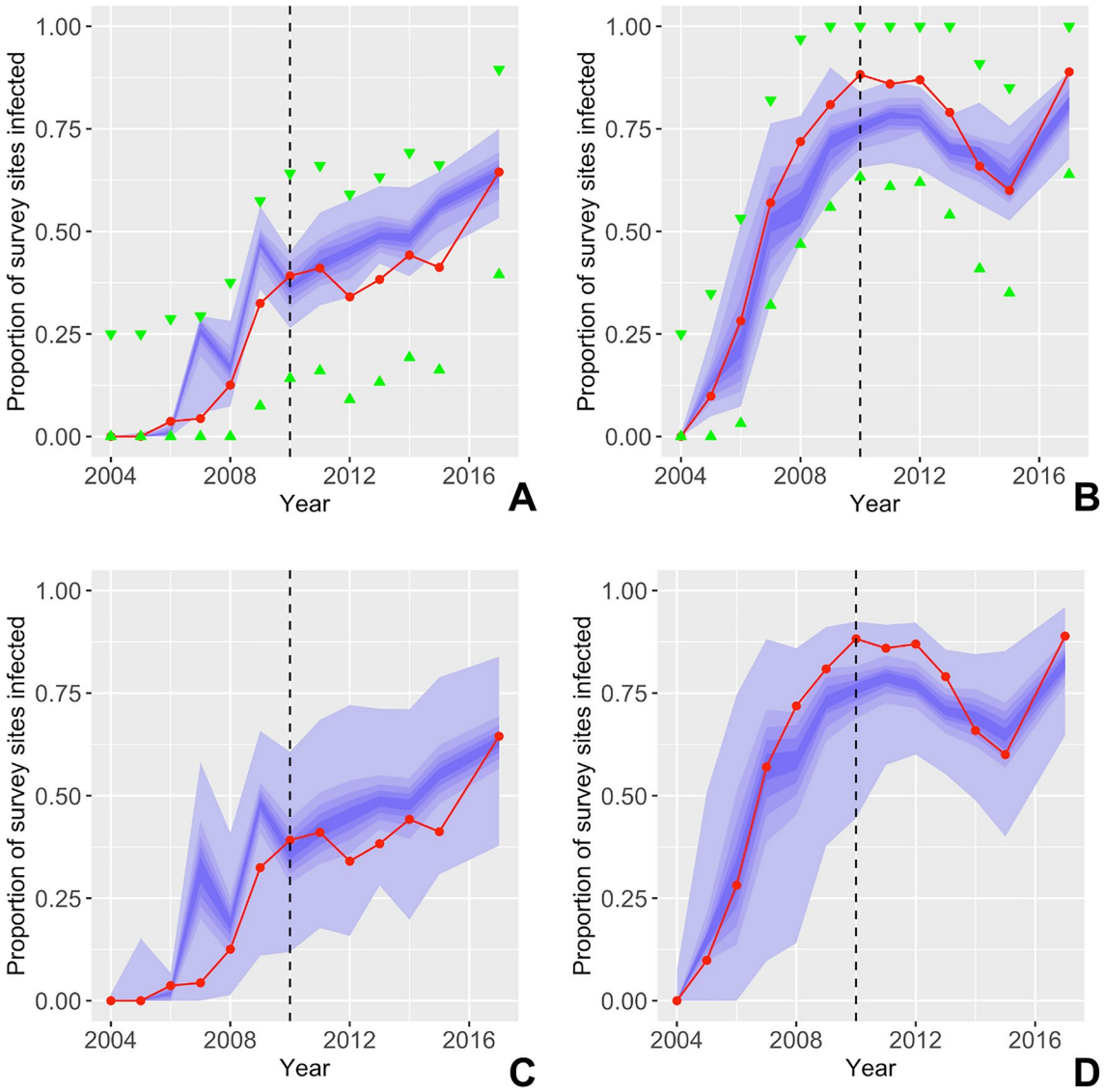

Figure 5. Time series probability distributions for the statistics (**A**) S_{nat} and (**B**) S_{cen} for the subset of validation simulations that pass within the tolerances of the fitting and validation criteria. (**C**,**D**) The same statistics but without applying any tolerances to illustrate the unconstrained behaviour of the parameterised model. The red line indicates the target value of each statistic derived from surveillance data. Tolerances are indicated by green arrows. The central blue band is the median ± 10% and each gradation beyond is a further ± 10% from the median. The dotted black line indicates divide between d_{real}^{fit} and d_{real}^{val}.

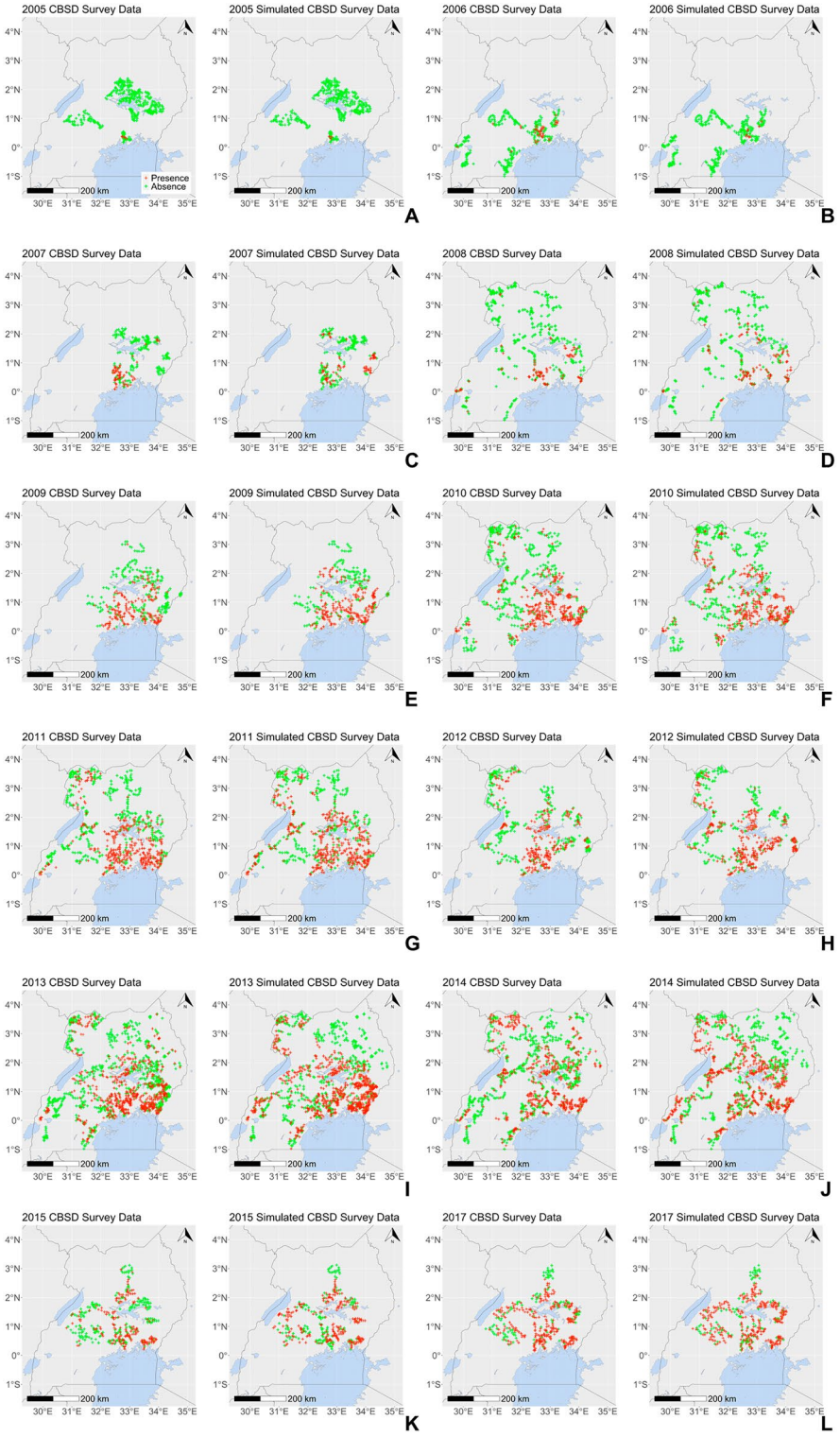

Figure 6. Comparison of a single validation simulation that passes the fitting and validation criteria with the real-world surveillance data from 2005 to 2017. We define the fitting criteria as the values selected for the three tolerances for parameter estimation using Ugandan survey data from 2004 to 2010: $\varepsilon_{cen}^{fit} = 0.25$, $\varepsilon_{nat}^{fit} = 0.25$, $\varepsilon_{grid}^{fit} = 0.48$ and the validation criteria as the comparison of the simulated surveillance data to the Ugandan survey data covering the validation period from 2011 to 2017, d_{real}^{val}, using the two S_{inf} statistics with tolerances $\varepsilon_{inf}^{val} = 0.25$. Red crosses indicate an observation of CBSD at the field-level. Green crosses indicate no CBSD observed.

Discussion

It is essential that policy makers across the cassava producing countries of sub-Saharan Africa have a clear understanding of the current state of the CBSD epidemic and the likely future spread to assist in deciding when and how to mitigate the impact of the CBSD epidemic and minimise future spread. However, surveillance data for the ongoing CBSD epidemic are extremely sparse, especially further west of the now endemic regions in East Africa due to complex geographic, political and financial constraints. Moreover, until now, no landscape-scale spatial epidemic models of CBSD existed to extrapolate beyond surveillance data and provide a shared quantitative framework to assist policy formulation. To this end, we have presented the development, parameterisation and validation of a landscape-scale stochastic model of the CBSD epidemic in Uganda. The fitted model shows strong correspondence to the validation dataset, d_{real}^{val} (Figs. 5, 6). Importantly, whilst this study focuses on Uganda, the data-driven host and vector input layers can be readily extended throughout sub-Saharan Africa, dependent on the availability of relevant data. Moreover, the success of the model in simulating the impact of the localised deployment of virus-free planting material (clean seed programme) in the region around Kampala between 2013 and 2015 provides initial evidence for the flexibility of the model to predict and analyse the impacts of management scenarios (Fig. 5). Notably, our analysis of the clean seed programme provides a rare landscape-scale view of the temporary impact of a large-scale management programme involving approximately 40 million cuttings over 3 years. The empirical data along with the model analysis indicated that clean seed programmes had a notable impact on regional disease prevalence, even under conditions of high disease pressure. However, the impact on disease levels was temporary and the epidemic returned to rapid rates of spread and high levels of intensity, when the clean-seed programme was stopped. Although beyond the scope of the current paper, further work is underway to compare the immediate and longer-term effectiveness of a range of management scenarios to reduce the impact and spread of CBSV.

The available data on the CBSD pathosystem are both spatially and temporally sparse. The pathogen has two dispersal mechanisms: vector-borne spread and human-mediated movement of infected planting material. In the absence of targeted data collection to distinguish between the two mechanisms, we have parameterised a single dispersal kernel that represents the net effect of both underlying forms of dispersal across the Ugandan cassava landscape. The parameterised model proved sufficient to characterise the spread of the pathogen in Uganda, with a marked correspondence between simulated and real-world surveillance data for both the data fitting period, d_{real}^{fit} (2004–2010), and validation period, d_{real}^{val} (2011–2017) (Fig. 6). We recommend, however, that future work should focus on disentangling the two dispersal mechanisms.

Due to an absence of cheap, reliable in-field diagnostics for CBSD, the survey protocol was based on above ground foliar symptoms, which likely leads to an underestimation of the true prevalence at the plant level[29]. The impact of this underestimation is mitigated by modelling spread at the field level with data being aggregated across the 30-plant sample to a field level presence/absence record. Moreover, we allow for false negative biases in the model structure to account for the limitations of the protocol, such as the limited number of plants surveyed per field and biases in cultivar selection (Supplementary Methods S1.1.1). Nonetheless, the false negative rate does conflate a number of underlying factors that contribute to the probability of surveillance not detecting the presence of disease in an infected field, including asymptomatic or mild symptoms. Further experimental and modelling studies would be valuable to quantify heterogeneity in the distribution of cassava cultivars across SSA, the variability in the cultivar specific symptom expression, and the associated efficacy of different surveillance protocols.

The model is modulated by data for the spatial distribution of the host crop and the insect vector across the Ugandan landscape, thereby improving model fit (Supplementary Methods S1.1.3). The host landscape layer provides the best available model of cassava production at a 1 km^2 resolution across sub-Saharan Africa[25], providing a single temporal snapshot of the amount of cassava being grown. We therefore assume a constant distribution of cassava production across Uganda over time. We also assume that all cassava is equally susceptible to infection by CBSIs. The evidence to date indicates all African cassava varieties are susceptible to CBSIs[30,31], however, there is a need for further studies to systematically assess variation in susceptibility and transmission across cultivars along with improving our understanding of the landscape-scale distribution of these cultivars. Moreover, it is critical to agree shared terminology in relation to tolerance and resistance[32]. Similarly, evidence to date does not indicate major differences in either the yield impact or geographic distribution of the two CBSIs[33]. Hence, for the purposes of this study, we did not distinguish between the two.

The vector abundance layer is derived from the interpolation of values from the Ugandan cassava field surveys that quantified *B. tabaci* abundance. It is important to note that the uncertainty in vector abundance layer values is higher in regions with lower spatiotemporal surveillance density. As with the host landscape, the vector abundance layer provides a single atemporal snapshot, therefore assuming the vector abundance remains stable over time and the different species in the *B. tabaci* complex are equivalently capable of transmitting CBSIs. There is an ongoing debate over the extent to which there is an interaction between the cassava epidemics of CBSD and cassava mosaic disease (CMD) and changes in local vector abundance or the specific abundance of species within the *B. tabaci* complex[34,35]. Extensive experimental and modelling work would be necessary to improve our understanding of the significance of the local composition and abundance of the *B. tabaci* species complex on the spread of the epidemic. Despite this complexity, it is clear that the incorporation of the vector abundance layer improved the predictive power of the model (Supplementary Methods S1.1).

The model presented in this study represents a significant advancement in our ability to predict the spread of the CBSD epidemic and simulate disease management scenarios. Importantly, the model has the potential to act as an overarching quantitative framework to assist in addressing a number of vital questions: how can CBSD endemic countries minimise the impact and reduce the prevalence of CBSD; when will CBSIs spread to currently

unaffected countries in West Africa; and how can these not yet affected countries optimise surveillance for early detection and prepare to control an outbreak.

Methods

We developed, parameterised and validated a stochastic metapopulation epidemic model for CBSD in Uganda at a 1 km^2 resolution via an iterative process of model development (Supplementary Methods S1.1). The model integrates a host landscape of cassava production and the relative spatial abundance of the insect vector, *B. tabaci* (Fig. 1). We simulated the spread of CBSD across the host landscape as a spatially explicit, SI (Susceptible-Infected) epidemic via a discrete event, continuous time stochastic process using an optimised Gillespie algorithm[36,37]. An SI compartmental structure was selected as cassava is a vegetatively propagated crop, so infection persists from one harvest to the next planting[38]. The model was parameterised and validated using annual surveillance data for the spread of CBSD in Uganda[14] (Supplementary Fig. 1).

CBSD surveillance data. Surveillance of the CBSD epidemic in Uganda was carried out in annual field surveys since the start of the epidemic in 2004 through to 2017, with the exception of 2016[14]. In a given year, surveyors visited between 253 and 1250 fields, with a mean of 587. The spatial distribution of surveys was not uniform. In some years, surveys were carried out in relatively small regions, whereas in other years surveys were more evenly distributed throughout the country. The same fields were not revisited across multiple years. The survey protocol for a given field involved surveyors randomly selecting 30 plants of the dominant cultivar across two diagonal transects and recording the severity or absence of visual CBSD foliar symptoms, along with the number of individual *B. tabaci* on the upper five leaves.

From the perspective of assessing disease presence at the field level, two factors likely resulted in a degree of systematic under-reporting of disease. Firstly, by sampling only the dominant cultivar surveyors did not record disease on non-dominant varieties. Secondly, a sample size of 30 plants is small relative to a total of approximately 1000 plants in a field of 0.1 ha. However, the dataset contained additional information for a subset of fields that allowed us to estimate a false negative reporting rate. For data collected from 2009 to 2014, surveyors reported whether CBSD disease symptoms were observed anywhere in the field, as opposed to just on the dominant variety or the 30-plant sample. Based on an analysis of these records, the average false negative under-estimation rate was estimated to be approximately 0.15 (Supplementary Methods S1.1.1).

Model structure. For the host landscape layer, we assume an average per field cassava yield of 10 tonnes per hectare[39] and an average field size of 0.1 ha[40–42]. The CassavaMap model used two forms of input data: human population data and regional cassava production statistics. For each region, the total production volume was allocated in proportion to the number of inhabitants per km^2 with the exception of spatial locations with populations greater than 5000 inhabitants per km^2, which were excluded to avoid the allocation of production to urban areas. The model caps production at 1000 tonnes per km^2[25].

For the vector abundance layer, field-level vector abundance mean values were collapsed across all survey years to create a single atemporal dataset. Field-level means were then capped to a maximum credible mean value of 100. Inverse distance weighted (IDW) interpolation was applied with a power value of 1.0, generating a rasterised layer with a 5 km resolution. A linear relationship between *B. tabaci* count and field-level infectiousness and susceptibility is used, reaching saturation at 20 *B. tabaci*[43]. Therefore, raster values above 20 post-IDW were set to 20, and the resultant raster was normalised with a maximum value of 1.

The instantaneous state of the model is defined by the number of susceptible and infectious fields in each raster cell. The model is updated via a discrete event, continuous time stochastic process using an optimised Gillespie algorithm[36,37]. Spatial coupling between infected and susceptible cells is governed by an isotropic discrete power law dispersal kernel, K, where the distance in km between the centroids of two raster cells i and j is d_{ij} and α is the exponent thus:

$$K\left(d_{ij}\right) = Ad_{ij}^{-\alpha} \tag{1}$$

Due to an absence of data independently quantifying the two dispersal mechanisms, the model integrates both the net effects of dispersal by *B. tabaci* and the movement of planting material on the spread of infection into the single kernel.

An additional parameter, p, defines the kernel value at $d = 0$, where p is the proportion of dispersed inoculum that remains within the source cell. A kernel cut-off distance, D_{max}= 500 km, sets the maximum distance from the source cell that the kernel covers. For the finite set of cell centroids in the range $0 < d \leq D_{max}$, values are calculated based on the kernel function. A normalisation factor, A, is applied such that the sum of kernel values for $d > 0$ is equal to the value of $1 - p$. Therefore, the sum of the kernel is 1.

The force of infection at location i, φ_i, incorporates the kernel function, K, the transmission rate, β, the vector abundance parameter at location j, w_i, and the current number of hosts in the infectious state, I_j, where j represents all locations in the rasterised landscape, including i (Eq. 2). The instantaneous rate of infection at a given raster cell, i, from all locations, j, is ψ_i, that incorporates the vector abundance parameter, w_i, and the number of susceptible hosts, S_i, at location i (Eq. 3). The effect of an infection event at location i is given by Eq. (4). We assume a linear relationship between vector abundance and its effect on infection and susceptibility up to a field-level mean of 20 *B. tabaci*[43]. Exploratory analyses on alternative model structures are outlined in Supplementary Methods S1.1.3.

$$\text{Force of infection at location } i : \varphi_i = \sum_j \beta w_j I_j K(d_{ij}) \tag{2}$$

$$\text{Infection rate at location } i : \psi_i = \varphi_i w_i S_i \tag{3}$$

$$\text{Effect of infection event at location } i : S_i \rightarrow S_i - 1, I_i \rightarrow I_i + 1 \tag{4}$$

The model implements a surveillance scheme that replicates the real-world surveillance structure and intensity. For all years that surveillance was carried out in Uganda[14], we perform one instantaneous survey at the end of the simulation year. For example, we assume that all surveys that are carried out in 2005 are representative of the state on 31st December 2005. For each raster cell in the model landscape, we summed the number of fields that were surveyed in the Ugandan national survey within the bounds of the 1 km^2 cell for a given survey year. We then randomly sampled the equivalent number of fields in each model cell and the numbers of sampled fields that were in each system state of susceptible and infectious were recorded allowing for the false negative survey detection rate of 0.15 (CBSD surveillance data in Methods).

The first reported observations of CBSD epidemic in Uganda occurred in November 2004[13]. However, the study did not provide exact coordinate locations for the CBSD positive fields in November 2004. The dataset includes survey data from January 2005, reporting infected fields in the same region. Therefore, we assume the small number of CBSD infected fields reported during the January 2005 surveys is representative of the state of the epidemic on 1st January 2004, which we take as the simulation start time.

We have made a number of underlying assumptions in the formulation of the model which for convenience we summarise here. We address the potential limitations of these assumptions in the discussion. The model simulates only one type of host, vector and virus. Hence, we assume that all cassava fields in the model are equivalently susceptible to CBSIs and, if infected, have the same probability of being detected if surveyed. Moreover, we therefore implicitly assume that all species in the *B. tabaci* complex vector CBSIs with equivalent efficiency. Similarly, we implicitly assume that all CBSIs are equivalently infectious. We simulate dispersal using a single dispersal kernel that represents the net effect of both vector and human-mediated forms of dispersal. Whilst both the host landscape and vector abundance layer vary spatially, we assume that both are static in time.

Parameter estimation. In the case of the Ugandan CBSD survey data, the model is unlikely to reproduce the spatiotemporal pattern of over 7600 records of CBSD presence/absence exactly. Therefore, given finite computational resources, the ABC methodology involved accepting parameter values from simulations that generated simulated survey data, d_{sim}, that were considered sufficiently close to the real-world data, d_{real}. Summary statistics, S, were used to simplify the comparison between simulated and real data, along with the selection of a distance measure, ρ, and a maximum allowed distance (tolerance), ε, between $S(d_{sim})$ and $S(d_{real})$ defined:

$$\rho(S(d_{sim}), S(d_{real})) \leq \varepsilon. \tag{5}$$

The target data for S_{cen} and S_{nat} were derived by applying the statistics to the real-world survey data, $S_{cen}(d_{real})$ and $S_{nat}(d_{real})$, resulting in the yearly proportion of survey sites within their geographical regions (the central area surrounding Kampala and the remaining area of Uganda not covered by the central area respectively) that were reported as positive for CBSD symptoms (Fig. 2). The S_{inf} distance measure, ρ_{inf}, is calculated by taking the maximum of the absolute annual differences between simulated and real survey proportions of infected survey sites (Eq. 6). A tolerance, ε_{inf}, then governs the maximum allowed value of ρ_{inf} (Eq. 7).

$$\rho_{inf}\left(S_{inf}(d_{real}), S_{inf}(d_{sim})\right) = \max_{year} \left| S_{inf}\left(d_{real}^{year}\right) - S_{inf}\left(d_{sim}^{year}\right) \right| \tag{6}$$

$$\rho_{inf}\left(S_{inf}(d_{real}), S_{inf}(d_{sim})\right) \leq \varepsilon_{inf} \tag{7}$$

The statistic S_{grid} applied to the real Ugandan survey data, d_{real}, has a perfect score of 1 (Eq. 8). The distance measure, ρ_{grid} is calculated by subtracting the statistic as applied to simulated data, d_{sim}, from 1 (Eq. 9) and the tolerance governs the maximum allowed value of the distance measure (Eq. 10). A deviation of ± 1 year for a given quadrat in the year in which CBSD was first detected in the Ugandan surveillance data is only allowed for quadrats with surveys both 1 year earlier and 1 year later than the target infection year, otherwise no deviation is allowed. Supplementary Fig. 6 summarises the target data for each quadrat and highlights the quadrats with surveys both years either side of the target first year of infection.

$$S_{grid}(d_{real}) = 1 \tag{8}$$

$$\rho_{grid}\left(S_{grid}(d_{real}), S_{grid}(d_{sim})\right) = 1 - S_{grid}(d_{sim}) \tag{9}$$

$$\rho_{grid}\left(S_{grid}(d_{real}), S_{grid}(d_{sim})\right) \leq \varepsilon_{grid} \tag{10}$$

The summary statistic assessment methodology using synthetic data allowed an exploration of the convergence of the posterior distribution to the known parameter values as the tolerances are reduced, whilst retaining enough simulations to enable a smooth posterior distribution given the finite number of fitting simulations (Supplementary Methods S1.3). Based on these analyses, we use all three statistics in combination and define the

fitting criteria as the following values for the three tolerances for each of the statistics: $\varepsilon_{cen}^{fit} = 0.25$, $\varepsilon_{nat}^{fit} = 0.25$, $\varepsilon_{grid}^{fit} = 0.48$.

For the prior distribution of ρ, we sampled from a uniform distribution between 0 and 1. For α and β, we carried out multiple batches of simulations, updating the search space at each iteration to sufficiently explore parameter space in order to identify the regions of highest likelihood density (Supplementary Methods S1.2). The total number of fitting simulations was 233,600.

Model validation. We assess the ability of the parameterised model to predict the validation dataset, d_{real}^{val}, which spans 2011–2017. We run 10,000 validation simulations starting in 2004 using the same initial conditions as during parameter estimation and isolate the subset of simulations that meet the fitting criteria. We take the subset of simulations that pass the fitting criteria as representative of the system state at the start of 2011, then for the validation period calculate the summary statistics, S_{inf}, and score model performance according to the percentage of simulations that satisfy $\varepsilon_{inf}^{val} = 0.25$. In addition, we present the full dynamics of $S_{inf}(d_{real})$ and spatial comparisons of a simulated survey with the real-world survey.

Data availability

A script to automatically download the datasets analysed during this study from their published sources is shared as part of the code repository.

Code availability

The code is available at https://github.com/camepidem/cbsd_model_development.

Received: 5 July 2022; Accepted: 15 July 2023
Published online: 03 August 2023

References

1. Redinbaugh, M. G. & Stewart, L. R. Maize lethal necrosis: An emerging, synergistic viral disease. *Annu. Rev. Virol.* **5**, 301–322 (2018).
2. Day, R. *et al.* Fall armyworm: Impacts and implications for Africa. *Outlooks Pest Manage.* **28**, 196–201 (2017).
3. Kumar, P. L., Selvarajan, R., Iskra-Caruana, M.-L., Chabannes, M. & Hanna, R. Chapter seven—biology, etiology, and control of virus diseases of banana and plantain. In *Advances in Virus Research* Vol. 91 (eds Loebenstein, G. & Katis, N. I.) 229–269 (Academic Press, 2015).
4. Tadesse, W., Bishaw, Z. & Assefa, S. Wheat production and breeding in sub-Saharan Africa: Challenges and opportunities in the face of climate change. *Int. J. Clim. Change Strat. Manage.* **11**, 696–715 (2018).
5. Allen-Sader, C. *et al.* An early warning system to predict and mitigate wheat rust diseases in Ethiopia. *Environ. Res. Lett.* **14**, 115004 (2019).
6. Jacobson, A. L., Duffy, S. & Sseruwagi, P. Whitefly-transmitted viruses threatening cassava production in Africa. *Curr. Opin. Virol.* **33**, 167–176 (2018).
7. FAO. *Desert Locust Upsurge: Progress Report on the Response in the Greater Horn of Africa and Yemen, January–April 2021* (FAO, 2021).
8. Nweke, F. I., Lynam, J. K. & Spencer, D. S. C. *The Cassava Transformation: Africa's Best-Kept Secret* (Michigan State University Press, 2002).
9. Hillocks, R. & Jennings, D. Cassava brown streak disease: A review of present knowledge and research needs. *Int. J. Pest Manage.* **49**, 225–234 (2003).
10. Legg, J. P. P. *et al.* Comparing the regional epidemiology of the cassava mosaic and cassava brown streak virus pandemics in Africa. *Virus Res.* **159**, 161–170 (2011).
11. Maruthi, M. N. *et al.* Transmission of Cassava brown streak virus by *Bemisia Tabaci* (Gennadius). *J. Phytopathol.* **153**, 307–312 (2005).
12. Legg, J. P. *et al.* Cassava virus diseases: Biology, epidemiology, and management. *Adv. Virus Res.* **91**, 85–142 (2015).
13. Alicai, T. *et al.* Re-emergence of cassava brown streak disease in Uganda. *Plant Dis.* **91**, 24–29 (2007).
14. Alicai, T. *et al.* Expansion of the cassava brown streak pandemic in Uganda revealed by annual field survey data for 2004 to 2017. *Sci. Data* **6**, 1–8 (2019).
15. Munganyinka, E. *et al.* Cassava brown streak disease in Rwanda, the associated viruses and disease phenotypes. *Plant Pathol.* **67**, 377–387 (2018).
16. Bigirimana, S., Barumbanze, P., Ndayihanzamaso, P., Shirima, R. & Legg, J. P. First report of cassava brown streak disease and associated Ugandan cassava brown streak virus in Burundi. *New Dis. Rep.* **24**, 2044–2588 (2011).
17. Mware, B. O., Ateka, E. M. & Songa, J. M. Transmission and distribution of cassava brown streak virus disease in cassava growing areas of Kenya. *J. Appl. Biosci.* **20**, 864–870 (2009).
18. Mbanzibwa, D. R. *et al.* Simultaneous virus-specific detection of the two cassava brown streak-associated viruses by RT-PCR reveals wide distribution in East Africa, mixed infections, and infections in *Manihot glaziovii*. *J. Virol. Methods* **171**, 394–400 (2011).
19. Mulimbi, W. *et al.* First report of Ugandan cassava brown streak virus on cassava in Democratic Republic of Congo. *New Dis. Rep.* **26**, 11–11 (2012).
20. Mulenga, R. M. *et al.* Cassava brown streak disease and ugandan cassava brown streak virus reported for the first time in Zambia. *Plant Dis.* **102**, 1410–1418 (2018).
21. Muhindo, H. *et al.* Optimum time for harvesting cassava tubers to reduce losses due to cassava brown streak disease in Northeastern DRC. *J. Agric. Sci.* **12**, 70 (2020).
22. Csilléry, K., Blum, M. G. B., Gaggiotti, O. E. & François, O. Approximate Bayesian computation (ABC) in practice. *Trends Ecol. Evol.* **25**, 410–418 (2010).
23. McKinley, T. J., Cook, A. R. & Deardon, R. Inference in epidemic models without likelihoods. *Int. J. Biostat.* **5**, 25 (2009).
24. Beaumont, M. A. Approximate Bayesian computation. *Annu. Rev. Stat. Its Appl.* **6**, 379–403 (2019).
25. Szyniszewska, A. M. CassavaMap, a fine-resolution disaggregation of cassava production and harvested area in Africa in 2014. *Sci. Data* **7**, 1–5 (2020).
26. Pritchard, J. K., Seielstad, M. T., Perez-Lezaun, A. & Feldman, M. W. Population growth of human Y chromosomes: A study of Y chromosome microsatellites. *Mol. Biol. Evol.* **16**, 1791–1798 (1999).
27. Carroll, R. *Eastern Africa Agricultural Productivity Project Implementation completion and Results Report* (2016).

28. AfrII. *Commercializing Quality Cassava Planting Material Delivery System in Uganda (css)* (Africa Innovations Institute, 2017).
29. Kawuki, R. S. *et al.* Alternative approaches for assessing cassava brown streak root necrosis to guide resistance breeding and selection. *Front. Plant Sci.* **10**, 25 (2019).
30. Sheat, S., Fuerholzner, B., Stein, B. & Winter, S. Resistance against cassava brown streak viruses from Africa in cassava germplasm from South America. *Front. Plant Sci.* **10**, 567 (2019).
31. Elegba, W., Gruissem, W. & Vanderschuren, H. Screening for resistance in farmer-preferred cassava cultivars from Ghana to a mixed infection of CBSV and UCBSV. *Plants* **9**, 1026 (2020).
32. Jeger, M. J. Tolerance of plant virus disease: Its genetic, physiological, and epidemiological significance. *Food Energy Secur.* **14**, e440 (2022).
33. Ndunguru, J. *et al.* Analyses of twelve new whole genome sequences of cassava brown streak viruses and Ugandan cassava brown streak viruses from east Africa: Diversity, supercomputing and evidence for further speciation. *PLoS One* **10**, e0139321 (2015).
34. Mugerwa, H., Wang, H.-L., Sseruwagi, P., Seal, S. & Colvin, J. Whole-genome single nucleotide polymorphism and mating compatibility studies reveal the presence of distinct species in sub-Saharan Africa *Bemisia Tabaci* whiteflies. *Insect Sci.* **28**, 1553–1566 (2021).
35. Donnelly, R. & Gilligan, C. A. The role of pathogen-mediated insect superabundance in the East African emergence of a plant virus. *J. Ecol.* **110**, 1113–1124 (2022).
36. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361 (1977).
37. Stutt, R. O. J. H. *Large Scale Epidemiological Modelling of Invading Plant Pathogens* (University of Cambridge, 2015).
38. Hillocks, R. J. & Thresh, J. M. Cassava mosaic and cassava brown streak virus diseases in Africa: A comparative guide to symptoms and aetiologies. *ResearchGate* **7**, 25 (2000).
39. FAOSTAT. Food and Agriculture Data (2022).
40. Lose, S. J., Hilger, T. H., Leihner, D. E. & Kroschel, J. Cassava, maize and tree root development as affected by various agroforestry and cropping systems in Bénin, West Africa. *Agric. Ecosyst. Environ.* **100**, 137–151 (2003).
41. Night, G. *et al.* Occurrence and distribution of cassava pests and diseases in Rwanda. *Agric. Ecosyst. Environ.* **140**, 492–497 (2011).
42. Owusu, V. & Owusu-Sekyere, E. Assessing the determinants of adoption of improved cassava varieties among farmers in the Ashanti region of Ghana. *Afr. Dev. Resour. Res. Inst. ADRRI J.* **5**, 92–104 (2014).
43. Ferris, A. C., Stutt, R. O. J. H., Godding, D. & Gilligan, C. A. Computational models to improve surveillance for cassava brown streak disease and minimize yield loss. *PLoS Comput. Biol.* **16**, e1007823 (2020).

Acknowledgements

The authors gratefully acknowledge financial support from the Bill & Melinda Gates Foundation (Grant Number: INV-010472), the UK Foreign, Commonwealth and Development Office and the BBSRC. We also acknowledge many helpful discussions and support from members of the Epidemiology & Modelling Group in Cambridge, especially Anna Szyniszewska and Renata Retkute, in addition to partners from the Cassava Diagnostics Project (CDP) and the Central and West African Virus Epidemiology (WAVE) project.

Author contributions

C.A.G., R.O.J.H.S. and D.G. formulated the problem and modelling approach. D.G. and R.O.J.H.S. developed, tested and implemented the modelling framework and performed parameter estimation and validation. T.A., P.A. and G.O. provided the surveillance data underpinning the analysis and gave expert insight into the clean seed programmes. D.G., C.A.G. and R.O.J.H.S. planned and D.G. wrote the manuscript and created the figures in collaboration with C.A.G. and R.O.J.H.S. C.A.G. supervised the project.

Funding

This article was funded by Bill & Melinda Gates Foundation (Grant Number: INV-010472), Foreign, Commonwealth and Development Office and Biotechnology and Biological Sciences Research Council.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-38819-x.

Correspondence and requests for materials should be addressed to D.G.