Assessing the Multi-pathway Threat from an Invasive Agricultural Pest: *Tuta absoluta* in Asia

Joseph McNitt[1], Young Yun Chungbaek[2], Henning Mortveit[2], Madhav Marathe[2], Mateus R Campos[3], Nicolas Desneux[3], Thierry Brévault[4,5,6], Rangaswamy Muniappan[7], and Abhijin Adiga[2]

[1]*Epic Systems Corporation, United States,* [2]*Biocomplexity Institute & Initiative, University of Virginia,* [3]*French National Institute for Agricultural Research,* [4]*BIOPASS, CIRAD-IRD-ISRA-UCAD, Dakar, Senegal,* [5]*CIRAD, UPR AIDA, F-34398 Montpellier, France,* [6]*Université de Montpellier, CIRAD, Montpellier, France,* [7]*Feed the Future Integrated Pest Management Innovation Lab, Virginia Tech*

Abstract

Modern food systems facilitate rapid dispersal of pests and pathogens through multiple pathways. The complexity of spread dynamics and data inadequacy make it challenging to model the phenomenon and also to prepare for emerging invasions. We present a generic framework to study the spatiotemporal spread of invasive species as a multi-scale propagation process over a time-varying network accounting for climate, biology, seasonal production, trade, and demographic information. Machine learning techniques are used in a novel manner to capture model variability and analyse parameter sensitivity. We applied the framework to understand the spread of a devastating pest of tomato, *Tuta absoluta*, in South and Southeast Asia, a region at the frontier of its current range. Analysis with respect to historical invasion records suggests that even with modest self-mediated spread capabilities, the pest can quickly expand its range through domestic city-to-city vegetable trade. Our models forecast that within five to seven years, *T. absoluta* will invade all major vegetable growing areas of Mainland Southeast Asia assuming unmitigated spread. Monitoring high consumption areas can help in early detection, and targeted interventions at major production areas can effectively reduce the rate of spread.

1 Introduction

As the intensity of trade and human mobility increase, so does the rate of exotic species invasions [18]. Climate change and the detrimental impact of intensive agriculture on natural resources further aggravate this problem [12]. Understanding the dynamics of invasive species spread is imperative for achieving zero hunger, no poverty, good health and well being, which are among the sustainable development goals drafted by the United Nations [25]. Models play an important role in predicting the spatiotemporal spread, identifying roles of different pathways, assessing efficacy of control strategies, and exposing gaps in the understanding of the phenomenon [9, 13]. However, impending invasions of agricultural pests present difficult challenges. Accounting for multiple drivers of dispersal invariably makes the model complex. At the same time, data inadequacy makes it nearly impossible to calibrate and validate these models. Despite these limitations, a natural goal for a modeller is to provide useful insights into the mechanisms of spread, and thus help design effective policies for its prevention and mitigation.

Network propagation models have been widely used to study ~~large interacting biological, social and technical systems. Some examples include infectious disease~~ phenomena as diverse as infectious disease and invasive species spread, online social networks, and cascading failures in infrastructure networks [3]. ~~Increasingly, they are being applied to study invasive species dynamics [7, 11, 26]~~Douma et al. [11] survey the invasive species literature categorising various efforts into flow-based pathway models and agent-based models. Network representations and analysis are being increasingly applied to capture human-mediated pathways of spread [7, 26], multi-scale spread [34] monitoring [22] and mitigation [29]. Unlike pest risk maps generated by species distribution models [28], the resulting dynamics of such a validated model yields a causal description of the underlying complex system. ~~Here, we~~

We present a multi-pathway propagation model to study the spread of invasive agricultural pests. ~~The model accounts for both self-mediated and human-mediated spread and effectively encapsulate the spatial heterogeneity, temporal variations and multi-scale nature of the propagation mechanisms.~~

~~We applied this framework~~ We applied it to study the spread of the South American tomato leafminer or *Tuta absoluta*, a pest of the tomato crop and representative of recent biological invasions that have significantly perturbed global food production. Indigenous to South America, *T. absoluta* was accidentally introduced to Spain in 2006, and since then has rapidly spread throughout Europe, Africa, Western and Central Asia, the Indian subcontinent, and parts of Central America [4, 10]. It is well accepted that trade played a critical role in

T. absoluta's rapid spread. On multiple occasions it has been discovered in packaging stations and its spread pattern is correlated with prime trade routes [19]. Our study region is South and Southeast Asia– a region at the frontier of its current range – comprising of 10 countries: members of the Association of Southeast Asian Nations (ASEAN) and Bangladesh. In recent years, there has been a thrust to improve vegetable production in all the countries of this region. With the pest having already spread to major tomato producing areas in Bangladesh, there is a high chance that it will be introduced to the remaining countries in the near future. Such invasions can have devastating effect on the economy and livelihood of farmers. Moreover, the invasion in Mainland Southeast Asia in particular is a serious threat to China, the largest producer of tomato, and Australasian neighbours. To our knowledge, this is the first study that explicitly considers multiple pathways of introduction and spread of *T. absoluta*. Earlier modelling efforts have only accounted for ecological aspects and self-mediated spread [10, 16, 31]. A precursor to this work [33] modelled the seasonal production and trade of tomato in Nepal to study the role of trade in the spread of *T.* absoluta using a gravity model and network dynamics.

Our ~~contributions are multifold. We provide a generic data-driven approach to build complex multi-pathway models. We~~ model accounts for both self-mediated and human-mediated spread and encapsulates the spatial heterogeneity, temporal variations and multi-scale nature of the propagation mechanisms. To construct this model we identified, analysed, and fused disparate datasets corresponding to biology, climate, production, and agricultural commodity flow. ~~We estimated the spatiotemporal distribution of host crops by integrating information from research articles, annual reports, and models. Trade flows were estimated using a gravity model approach. Machine learning techniques such as clustering and decision-tree algorithmswere used for parameterization, to capture model variability and study parameter sensitivity~~With *T.* absoluta being an emerging invasion in the focus region, some of the pertinent questions are (a) what are the possible explanations for the observed spread (b) what are the possible patterns of future spread, and (c) what steps could be taken to mitigate it. We develop a framework to parameterize and analyse the multi-pathway model with respect to ground truth by a novel application of popular supervised and unsupervised machine learning algorithms. Our approaches are motivated by recent research on machine learning surrogates for agent-based models [21] and interpretable artificial intelligence [15]. The analysis provides valuable insights into the dynamics of *T. absoluta* spread and its control, particularly from the perspective of human-mediated spread.

2 Methods

Data. The global datasets used in the model and for analysis are described in Table S1 of the supplement. Country specific data on seasonal production, consumption, processing, and trade was obtained from websites of agriculture ministries, research articles and technical reports (Table S2 in the supplement). Almost all the datasets used are openly available. The details of *T. absoluta* biology can be found in the supplement.

Multi-pathway spread model. We developed a stochastic multi-scale propagation model to simulate the multi-pathway spread of *T. absoluta*. Key concepts are illustrated in Figure 1a. The model parameters and their values are summarised in Table 1. A discussion on the choice of model structure and assumptions is presented in Section S3.1 of the supplement. The study region is divided into cells – the smallest spatial units – by overlaying a grid ($0.25° \times 0.25°$). Each cell is in one of the three states: susceptible (S) denoting pest free state, exposed (E) denoting that the pest has been introduced but the population has not yet built up to influence other cells, and infectious (I) denoting that the pest has established and the cell can influence its neighbours. The cell states are updated in discrete time steps, ~~where the interval between successive time steps corresponds to a~~ each corresponding to one month. The probability that a cell v transitions from state S to E is determined by (i) suitability ~~of the cell for *T. absoluta*~~ to establish at that time step $\epsilon(v, t)$ and (ii) influence of ~~"neighbouring" cells~~ neighbours in state I depending on the pathway. An exposed cell transitions to state I after a latency period of ℓ time steps. This is the time required for the excess population to build up to infect other cells. Once the pest has established in a cell, the cell remains infected forever, a fair assumption considering that, historically, eradication of *T. absoluta* has not been successful (the only exception being United Kingdom). The infectiousness of a cell $\rho(v, t)$ is modelled as a linear function of host ~~presence~~ density at time t, for which we use the weighted sum of production volume of tomato, eggplant, and potato in that cell at time t. The ~~weights correspond to relative carrying capacity of *T. absoluta* on the three hosts [30]~~details are in the supplement.

There are three pathways by which a cell can become infected: short-distance dispersal, local human-mediated dispersal and long-distance dispersal. Short-distance dispersal captures the spread through natural means; from an infested cell to cells in its Moore neighbourhood of range r_{M}. The probability that a

(a) Multi-pathway model (b) Model construction pipeline

(c) ~~An illustration~~ Outline of the ~~multi-pathway model (a) and the~~ process ~~of constructing~~ used for analysing the ~~spatiotemporal network of cells~~multi-pathway spread.

Figure 1: Multi-pathway model concept, construction and analysis.

susceptible cell gets exposed (E) at time step t through short-distance spread is as follows:

$$p_s(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_s \sum_{v' \in \mathrm{M}_v(r_\mathrm{M})} \rho(v',t)\Big)\Big). \tag{1}$$

The probability depends on the suitability of the cell $\epsilon(v,t)$, infestation level of each neighbouring cell in the Moore neighbourhood with range r_M, $\rho(v',t)$, and the scaling factor, α_s, which is the transmission rate for this pathway. The function form is explained in Section ~~??~~S3.1.

For human-assisted spread we identified large urban areas in the region which we refer to as *localities* (Figure 1a) and considered interactions within and between localities. These areas have significant trade flows due to high consumption or production. Each *locality* consists of all grid cells which are within a certain distance (determined by *locality radius*) from its corresponding centre. Local human-mediated dispersal is modelled as the spread between cells belonging to a locality. Every cell v is influenced by cells in its

Table 1: Model parameters, their values and notes on parameter choices and ranges.

Parameter	Description	Value/range
r_{M}	Range of Moore neighbourhood	$\{1, 2, 3\}$ corresponding to spread per month of approximately 25km, 50km and 75km respectively ~~[16, 23]~~.
ℓ	Latency period to transition from E to I	$\{1, 2, 3\}$ months based on the time for the pest to complete life cycle (*T. absoluta* biology in ~~Methods~~Section S2).
season	Disaggregation of annual production to monthly values	*Uniform* throughout the year or *seasonal* based on regression analysis (Methods).
β	Gravity model distance function exponent	$\{0, 1, 2\}$
κ	Gravity model distance function cut-off	Between 4 to 16 hours of travel time.
seed	Location and time of initial infestation	Scenarios based on countries (see Table S5)
locality radius	Determines cells assigned to a locality	100km (See Section S3.2 in the supplement for locality construction and analysis)
t_s	Time of initial infestation during parameterization	$\{3, 4, 5\}$ corresponding to March, April and May respectively based on first report in Bangladesh ~~[17]~~ (Table S3).
$\alpha_s, \alpha_\ell, \alpha_\ell$	Pathway scaling factors	In the interval $[0, 500]$.

locality L based on their infectiousness. The expression is similar to that in equation (1), but with cells in the locality instead of the Moore neighbourhood

$$p_\ell(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_\ell \sum_{v' \in L} \rho(v',t)\Big)\Big), \tag{2}$$

where α_ℓ is the scaling factor. The details of locality construction are provided in Section S3.2 of the supplement.

Long-distance human-mediated dispersal corresponds to spread through trade between localities. For this purpose, we considered only tomato trade as there is not much evidence of *T. absoluta* spreading through trade of other hosts. We modelled domestic trade using a gravity model approach accounting for tomato production, processing, imports, and exports in each locality, and the travel time between localities. The probability of spread is directly proportional to the trade flow F_{ij} from one locality (i) to another (j). Suppose cell v belongs to locality i. Then, the probability of cell v transitioning from S to E due to long-distance human-mediated dispersal is given by:

$$p_{\ell d}(v,t) = \epsilon(v,t)\Big(1 - \exp\Big(-\alpha_{\ell d} \sum_{j \neq i} \sum_{v' \in L(j)} F_{ji}\rho(v',t)\Big)\Big), \tag{3}$$

where $\alpha_{\ell d}$ is the pathway scaling factor. ~~The model parameters and their values are summarised in Table 1.~~

Network construction. Figure 1b provides a schematic of the network construction. The first step was to estimate monthly production volume of tomato, eggplant, and potato for each cell. We estimated annual production in each cell followed by disaggregation to monthly production. The annual production was estimated using the vegetable production available at the highest resolution for each country (at the level of province to just one value for the entire country) and a synthetic dataset called Spatial Production Allocation Model [36]. For monthly production, we used linear regression to model the production rate as a function of precipitation, temperature, and elevation. Seasonal tomato and eggplant production data for different regions of Philippines was used. For most of the other countries only qualitative information on seasonal production is available (Table S2). The regression function was applied to locations of these countries where this information is available and visually compared available data. The details are in Section S3.3 of the supplement.

To model locality-to-locality trade, we applied the approach of Venkatramanan et al. [33] with some modifications. We modelled the flow of fresh tomato crop between markets based on the following assumptions: (i) the total outflow from a city depends on the amount of produce in its surrounding regions and imports from countries outside the focus region at time t, and (ii) the total inflow depends on total consumption, processing demand, and exports from the city to countries outside the focus region. The details are in Section S3.4. Trade between countries of the focus region was not modelled as there is no adequate information on ports of entry or monthly flow volumes. But, it was accounted for while analyzing the possible routes of introduction.

Parameterization and experiment design. The goodness of fit of a parameter instance was determined by comparing the simulation output with *T. absoluta* incidence reports (Figure 2a and Table S3 for Bangladesh). The spread was simulated with infestation starting from the location of first report. For each cell~~, empirical probability~~ v and model configuration C, the empirical probability $p(C, v, t)$ that it is in state I at time t was computed (averaged over 100 repetitions). The output was compared to ground truth using a similarity function adapted from [7]. Let v be a reporting cell and t_v denote the month of the actual report of pest presence. To account for uncertainty in reporting, we consider a time window $U_\tau = [t_v - \tau, t_v + \tau]$ during comparison, where τ is the uncertainty parameter. We set $\tau = 2$, that is, error within ± 2 months is tolerated. Supposing $\mathcal{C}_R$ is the set of cells corresponding to ground truth, ~~and $p(C, v, t)$ is the empirical probability~~

~~that cell v is infected at time t for the input configuration C,~~ then the similarity $\mathcal{S}$ is given by

$$\mathcal{S}(C) = \frac{1}{|\mathcal{C}_R|} \sum_{v \in \mathcal{C}_R} \Big(\sum_{t \in U_\tau} p(C, v, t) + \sum_{t \notin U_\tau} (1 - p(C, v, t)) \Big) . \tag{4}$$

~~Note that $\mathcal{S}(C)$ attains a maximum value of one when the simulation output is a perfect match (within the error tolerance limit).~~ For parameter space exploration, we were motivated by a recent approach of using machine learning surrogates [21]. In our iterative "go with the winners" process [1] the subspace under consideration is sampled uniformly (the first part of Figure 1c). Then, with model parameters as features and the similarity score as the dependent variable, we use Classification and Regression Trees (CART) approach to identify parts of the subspace for which the similarity score is high and reject the remaining. In the following iteration, these subspaces are sampled uniformly, and the process continues. The approach is very similar to the *reverse engineering approach* used to build an interpretable learner for a black box model [15, Fig. 10]. In the interpretable machine learning framework, the black box is a machine learning algorithm like a neural network or tree ensemble, while in our case, the black box is an agent-based model. Simulations were performed for more than $500,000$ parameter combinations using a high performance computing cluster. Configurations with similarity score $\mathcal{S}(C) \geq 0.75$ were chosen for further analysis.

Analysis of spread pattern. The objective here is to analyse the variability in the simulation outcomes within the set of best fit configurations. We leverage well-known machine learning techniques in a novel way to address this question. The methodology is outlined in Figure ~~?? in the supplement~~1c. First, we cluster the simulation outputs (time and cell indexed empirical probabilities) of selected configurations from the parameterization phase. This step captures the variability in outcomes; simulation outputs belonging to different clusters can be considered to be significantly different from one another. In the second step, we attempt to infer relationships between model parameters and cluster membership. To this end, our approach is to cast this as a classification problem using CART with model parameters as the features and cluster index as the label. The relationships are inferred from the decision tree that resulted from the algorithm. To avoid any bias introduced by the clustering algorithm, we also apply more than one method – hierarchical agglomerative clustering and the k-means algorithm. In both cases, we use the Euclidean distance as the distance measure to compare two simulation outputs. The analysis is repeated for different values of k, the number of clusters. More details are provided in Section S5 of the supplement.

To assess the relative importance of model parameters, we adopted the approach of Lamperti et al. [21]. We use the random forest algorithm [6] to assess the importance with respect to (i) similarity score ($\mathcal{S}$) and (ii) spread pattern, which in our case, is akin to cluster membership. The setup is similar to the parameterization and cluster analysis case, with CART replaced by random forest algorithm. Details are in the supplement (Section S5). To evaluate parameter importance with respect to the similarity score, we used *mean increase in node purity* as the criterion (as it is a regression problem), and for cluster membership, we used *decrease in accuracy* (classification problem).

3 Results

Variability in spread pattern. The clustering analysis of the configurations selected during the parameterization phase (approximately 8000 of them) reveals two distinct spread patterns primarily determined by the pathway parameters. The first class of models (Figure 2a), referred to as Class A, is characterised by the absence of long-distance human-mediated spread ($\alpha_{\ell d}$ negligible) and brisk spread between geographically adjacent cells, driven by the latency period ℓ, the Moore range r_{M}, and the short-distance scaling factor α_s. In contrast, for Class B models (Figure 2b), the long-distance pathway ($\alpha_{\ell d}$) plays a significant role and there is relatively slow spread between geographically adjacent neighbours. Both hierarchical clustering and k-means clustering (Figures S5(b) and S6(a)) are consistent in this regard.

The Class A spread pattern does not capture the gap between the time of first report (Panchagarh) and the report in Gaibandha district (Figure 2a). Even though the distance between the two locations is only 185km, the latter reported the presence only after 10 months of first report, suggesting that self-mediated spread might have been much slower. In the model output on the other hand, the corresponding cell gets infected between the second and fourth months. In Class B, this location is infected much later in comparison. However, the eastward spread towards the location Jaintiapur is slower than what was observed (Figure 2b). Even though Panchagarh is quite far from this location, pest presence was reported by February 2017, just nine months after the first report. As a baseline, we also simulated the spread using the cellular automata model developed by Guimapi [16] for Bangladesh. The spread pattern is similar to Class A as the model does not account for long-distance hops. However, the predicted rate of range expansion is much higher than our models (see Section S6.3 for model details and results).

~~The relative importance of model parameters was assessed using random forest algorithm with regard~~

(a) Class A (b) Class B (c) Parameter importance

Figure 2: ~~Possible spread patterns in Bangladesh.~~ Explaining observed spread pattern in Bangladesh. The contour plots show the spread starting from the location of first report in Panchagarh district for a simulation time of 12 months. Here, the time of infection for a cell is the minimum time step t such that the empirical probability that the cell is infected by time t is ≥ 0.8. Also highlighted are the eight monitored locations and the localities applied in the model. The colours of the monitored locations correspond to the time of infection relative to the first report (Panchagarh). Two distinct spread patterns emerged from the cluster analysis. (a) and (b) show representative spreads observed for each class. The similarity in each case was $\mathcal{S} > 0.8$. Importance of model parameters with respect to (i) similarity score $\mathcal{S}$ and (ii) cluster membership based on random forest method. The latter plot shows how the results vary with an increase in the number of clusters for hierarchical clustering algorithm. More results are presented in Figure S9 in the supplement. Videos depicting the spatiotemporal spread for each class are provided of the supplement.

~~to their influence on (i) similarity score ($\mathcal{S}$) and (ii) spread pattern, which in our case, is akin to cluster membership.~~ In the case of spread pattern, the importance was derived for each k (number of clusters) and clustering algorithm. Some results are presented in Figure 2c. We note that the long-distance scaling factor ($\alpha_{\ell d}$) is among the top three important parameters. The start month (t_s) is also important for two reasons. Firstly, the distance between two time shifted simulation outputs can be large. Secondly, outputs are sensitive to seasonal variations or temporality of the network. Latency period (ℓ) and Moore range (r_{M}) together control the extent of radial spread in a time step. Typically for Class A models, r_{M} is high and ℓ is low and the other way round in the case of Class B models. Analysis of trade flows and seasonality is presented in Section S6.2 in the supplement.

Scenarios of pest introduction to countries in Southeast Asia. To identify routes of introduction to other countries in the region, we applied both Class A and Class B models. The starting point of the spread corresponds to the Panchagarh district (Figure 2a). Both model classes strongly indicate that *T. absoluta* is

already present in parts of Myanmar (curves corresponding to time step 24, or two years from first report). Also, the pest is likely to enter Thailand from Myanmar, and subsequently move to Laos, Cambodia, and Vietnam as it spreads eastwards, and to China when it spreads northwards. From Thailand, spreading southward, it will enter Malaysia and subsequently enter Indonesia (~~Figure ??~~Figures 3a and 3b).

We also analysed the international tomato trade network (Section S6.1 of supplement) to assess the risk due to imports from *T. absoluta* infested countries outside this region. Malaysia and Singapore are important hubs with tomato imports from *T. absoluta* infested regions. There is a possibility that *T. absoluta* is directly introduced to these regions. However, in both cases, the import volume is very low. Also, the introduction risk depends on the preventive measures taken by the exporting countries. With respect to both trade and natural pathways, there is a low chance that the pest will be introduced into Philippines from neighbouring countries, as there are no shared borders with any country in the region nor evidence of tomato trade. However, human mobility is a possible pathway. For example, the Middle East is the top destination for Filipino workers.

Predicted spread is model and region dependent. In the case of Class A models, the eastward spread is faster than southward spread (Figure ~~?? in the supplement~~3a). This is mainly because the Moore neighbourhood is smaller at the narrow region in the south of Myanmar and Thailand bordering Malaysia. However, in the case of Class B (Figure ~~??~~3b), the spread is much faster in the same region aided by domestic trade flows from northern and central Thailand to the southern region. The Class A spread pattern predicts that within the next 4-5 years, much of the northern part of Mainland Southeast Asia will be invaded. The Class B spread pattern predicts that in the same period, *T. absoluta* will spread all over Malaysia and Singapore. However, the rate of spread observed is slower than that observed in Bangladesh for both classes. Also, even though the models exhibited a similar rate of spread for Bangladesh, we observed high variance in intensity of infestation as well as range expansion for the rest of the region. The results are in Figure S11 in the supplement. The reason for slow spread is as follows. Bangladesh has the highest tomato volume per country surface area (≈ 2.5tonnes/km^2). The next country is Vietnam (≈ 1.5tonnes/km^2). Therefore, in the case of Bangladesh, not only is the extent of infestation in a cell $\rho(\cdot)$ typically high, but also since it is a densely populated country, most cells have vegetable production. Hence, the rate of spread is much higher for relatively lower values of pathway parameters and Moore range. Also, we observed a strong dependence on Moore range (Figure S11b). In geographically larger countries, the production is scattered. Therefore, the lower the Moore range, the slower the spread.

Influence of domestic trade on spread pattern and rate. Here, we focus on long-distance dispersal and therefore restrict our discussion to Class B models. For the country-specific studies, the starting location was decided based on our analysis of possible entry points through different pathways (Section S4.1 in the supplement). We observed the following common spread pattern. When the invasive species is introduced to a country, dispersal is slow until the invasion front reaches a production source. Once it establishes at a source, the spread is very fast. Depending on the country, within 12 to 24 time steps (or 1-2 years), it spreads to almost all major localities of the country (see Figure 3d for example). Production areas which are very close to high-consumption localities (large urban areas) are particularly vulnerable. Since local production typically does not satisfy demands of such localities, they have high inflows from other production areas and possibly from other countries. As a result, these localities are quickly infected. Once introduced to such localities, farmer–market interactions (local human-mediated dispersal) can facilitate the introduction of the pest to nearby production regions where it can establish and proliferate. Similar observations have been made in a number of works [20, 27].

Given that monitoring and quarantining are both resource intensive and potentially disruptive, developing strategies that involve few locations yet provide near-optimal control is a goal for modellers. Market-level phytosanitary measures in terms of import restrictions have been undertaken by countries [32]. Here, we evaluated a simple strategy for containing the spread through the trade pathway. Localities associated with high annual outflows were identified (at most four in each country). As discussed earlier, pest establishment in these areas can potentially lead to rapid range expansion. The outflow from the targeted localities was cut off to mimic control at the trade/market level. In the strictest sense, this can be implemented by restricting trade of host crops. But, it is possible that phytosanitary measures have the same effect. Figure 3 shows results for two countries. More results are present in Figure S12 in the supplement. Consistently, across countries, we observed a significant reduction in range expansion as well as intensity of spread. Besides, as seen in Figure 3d, stifling these flows localises the spread that resembles those of Class A models, but with much less intensity.

4 Discussion

The variability in the spread patterns that explain the incidence reports exposes the lack of understanding of the pathways of spread. Nevertheless, the analysis does strongly indicate the role of human-assisted spread of

(a) (b) (c) ~~Thailand without intervention~~ (d) ~~Thailand with intervention~~ (e) ~~Thailand: range of expansion (all Class B models)~~

Figure 3: ~~**Rate and pattern of spread with and without intervention.** Representative~~ **Predicting rate and pattern of spread.** The contour plots show the spatiotemporal spread. The colours indicate the time interval at which there is at least a 50% chance that a location will be infected. **Spread in Southeast Asia.** Figures (a) and (b) correspond to the spread starting from northern Myanmar for 120 time steps or 10 years. (a) corresponds to representative simulation output for Class A models and (b) corresponds to Class B models. **Domestic spread with and without intervention.** In (c) and (d), representative spread dynamics of Class B models ($r_{\mathrm{M}} = 1, \ell = 3$) are shown for ~~two countries~~the country of Thailand. More plots are ~~shown~~ in Figure S12. In each case, a cell close to a high production region was seeded ~~. The first column corresponds to spread~~ and simulation run for 48 ~~months after introduction. The colours indicate the~~ time ~~interval at which there is at least a 50% chance that a location will be infected~~steps. ~~The second column~~ (c) corresponds to unmitigated spread and (d) corresponds to spread after cutting off flows from chosen localities. ~~The third column~~ (e) shows average spread with respect to origin of infection for all Class B models. The cells are binned based on their distance from the origin of infection. Given time step t (48), let $\Pr_{\leq t}(v)$ be the probability that cell v is in state I by time t. For each configuration, we computed the "total infection" for every bin at time t by aggregating $\Pr_{\leq t}(v)$ for each v in the bin. The red points referred to as "max" correspond to the total number of cells in each bin, which is also the maximum possible accumulated probability for that bin.

T. absoluta. The pest was reported in May 2016 in the northwestern part of Bangladesh bordering India. The region is among the top three tomato producers in the country. By the beginning of the next production season, *T. absoluta* was found in almost every major urban region. Similar correlation between tomato trade and *T. absoluta* spread was observed in Nepal [33]. Studies on self-mediated spread (flying capability or by wind) can definitely help estimate more accurately the rate of self-mediated spread. It is also important to consider alternate scenarios of introduction. We recall that the far eastern part of Bangladesh (locality Jaintiapur in Figure 2b) reported pest presence nine months after the first report. This place happens to be close to an important trade route connecting Northeastern Bangladesh to Meghalaya in India, where *T. absoluta* was officially reported in January 2017. Therefore, it is possible that multiple incursions took place.

Historically, international trade has played a strong role in the spread of *T. absoluta* between countries. For example, the pest was first reported by India in 2014. By early 2016, it was discovered in the Kathmandu

area of Nepal and in the northern part of Bangladesh in May 2016. Both countries import significant volumes of tomatoes from India. However, there has been no report from Pakistan, another neighbour which does not import tomato from India. There are similar examples outside the region such as its slow advance from South America to Central America, or the fact that it is not reported in China despite being present in neighbouring Central Asian countries since 2015. We recall the discussion on slow predicted rate of spread in Mainland Southeast Asia compared to the observed rate in Bangladesh. One reason for this could be the unaccounted trade flows between countries. International trade within this region is not documented well. It is critical to address the data gaps concerning international trade, particularly considering that production and trade between countries in this region have been increasing over the years (details are in Section S6.1 of supplement).

While several integrated pest management strategies have been suggested for managing *T. absoluta*, hardly any work has been done in designing effective interventions at the trade level. ~~Designing phytosanitary measures targeted towards markets and vehicles of transportation for preventing introductions (or reintroductions) is therefore a promising research direction.~~ Some countries have already taken measures in this regard. ~~Our results also indicate that monitoring markets with inflows from many regions is important.~~ In the United States~~for example~~, the Animal and Plant Health Inspection Service of the Department of Agriculture (USDA-APHIS) has instituted quarantine regulations for imports from regions where the pest is present [32]. Identifying ~~the optimal set of nodes in a network to reduce infectious disease spread is a widely studied topic. There are very few works that apply such techniques to invasive species spread (Nopsa et al. [26] for example) . As the world moves towards concentrated and specialised agricultural production, focusing on this aspect becomes increasingly important.~~ important locations to mitigate an epidemic ([2, 20, 27]) or monitor are problems being increasingly studied with the lens of network analysis. While good algorithms exit for undirected networks, solutions for these problems on directed weighted networks are few and far in between.

~~**Literature survey.** Multi-pathway models are being increasingly used to study the role of invasive species dispersal. Douma et al. [11] survey the literature categorising various efforts into flow-based pathway models and agent-based models. Carrasco et al. [7] combine spatially explicit models of human-mediated spread with a phenology model to incorporate population dynamics of the western corn rootworm. Nopsa et al. [26] use a network science approach to study the role of transport and storage infrastructure in the spread of pests~~

~~and pathogens of wheat. Our model is in part motivated by the hybrid approaches used in the study of infectious diseases of humans and livestock [5, 35]. Although there is a general consensus that vegetable and seedling trade is a primary driver of *T.* absoluta spread, previous modelling efforts have exclusively focused on ecological aspects. Several studies [10, 31] provide risk maps using CLIMEX and take additional factors into account. Guimapi et al. [16] used a cellular automata approach to capture the global spread of the pest by factoring in temporal variations and spatial distribution of vegetation, temperature, and tomato production. A precursor to this work [33] modelled the seasonal production and trade of tomato in Nepal to study the role of trade in the spread of *T.* absoluta using a gravity model and network dynamics.~~

Emulators –based on Gaussian processes for example [14] – and machine learning surrogates [21] are emerging as solutions to overcome computational challenges, parameterization, and sensitivity analysis of complex agent-based models. ~~The parameter importance study and parameterization was partly motivated by the work of Lamperti et al. [21].~~ Our approaches were motivated by these works. We are not aware of any previous work that analyses the dynamics of simulation systems using unsupervised learning as presented in this paper. However, clustering has been considered in the context of multi-resolution simulation models as an interfacing component between simulators with different resolutions [8]. ~~We cast the problem of deriving relationship between model parameters and cluster index as a classification problem. CART was our choice of algorithm since the learned model is a decision tree that can be interpreted . In principle, any such algorithm which provides such an explanation (such as multinomial regression for example) can be employed.~~ Our use of CART to explain the clustering is motivated by recent work in interpretable aritifical intelligence [15], where deep learning models are interpreted using decision-tree proxy models.

Challenges and limitations. Modelling emerging invasions is particularly challenging. Limited data on incidence and understanding of the underlying dynamics makes it nearly impossible to calibrate and validate the models. We have had to simplify or ignore some of the processes that might significantly influence the spread. For example, our model uses monthly production as a surrogate for infectiousness of a cell. Complex phenology models can be used instead (as in Carrasco et al. [7]), but would add to the complexity of the model. Since our focus region spans multiple countries, identifying and collecting data for each country was a lengthy process. For many countries, data had to be collected (or even inferred) from several publications and reports (Table S2). Further, these datasets were misaligned in time and spatial resolution. It is important to account for heterogeneity in production, consumption, awareness, cultural factors, etc. both within and

between countries. Some countries are technologically more advanced than others, which manifests as differences in yield, crop loss, trade infrastructure, pest awareness, and preparation for invasion [12].

In particular, it is hard to model human-assisted spread owing to lack of seasonal trade data. To determine outflows and inflows for each locality, we had to identify major ports for import(s) and export(s) as well as estimate the fraction of production which was used for processing and was available only for a few countries. The farm–market-consumer interactions (local human-mediated spread) involves various actors such as farmers, wholesalers, retailers, wet markets, supermarkets, and so on. Modelling this is a challenge in itself. If data on actual flows of vegetables is provided, the gravity model can be improved or replaced by more sophisticated approaches. Also, the relationship between long-distance invasion risk and trade volume is hard to determine. While a direct relationship between volume and risk is plausible, whether the relation is linear (as assumed by our model) is unclear.

Conclusion. Traditionally, in developing countries, crops such as tomato are seasonal. However, over the past decade, due to rising demand and opportunities to export, there has been a thrust towards year-round production using protected cultivation methods and resilient varieties. An increase in urban population, short shelf life of vegetables, and the advantages of short marketing chains have encouraged urban agriculture in developing countries [24]. Our results indicate that such urban and peri-urban agriculture is particularly vulnerable to invasive species attacks. In particular, in Southeast Asia, vegetable production and internal trade have steadily increased. In comparison, the export of tomato outside of the focus region has risen steeply in recent years (after 2011), while the imports generally indicate a downward trend. Therefore, invasions from pests such as *T. absoluta* can have a huge negative impact on the socioeconomic fabric of this region. The modelling and analysis framework presented here is generic and applicable to other invasive species. The methodology is modular and leverages popular learning algorithms to analyse complex models under data scarcity. Other potential applications for this work include studies of natural or human-initiated disasters, climate change, and optimisation of food flows.

Data availability. The authors declare that the data supporting the findings of this study are available within the paper and its Supplementary Information file, or from the authors upon reasonable request.

Acknowledgements This work was supported in part by the United States Agency for International Development under the Cooperative Agreement NO. AID-OAA-L-15-00001 Feed the Future Innovation

Lab for Integrated Pest Management, DTRA CNIMS Contract HDTRA1-11-D-0016-0001, NSF BIG DATA Grant IIS-1633028, NSF DIBBS Grant ACI-1443054, NIH Grant 1R01GM109718 and NSF NRT-DESE Grant DGE-154362. We are grateful to Yousuf Mian, Nguyen Van Hoa, and Kimhian Seng for their help with obtaining country-specific information on production, trade, and pest incidence. We thank Richard Beckman, Irene Eckstrand, Srinivasan Venkatramanan, Stephen Eubank and Erin Raymond for useful discussions on model design and paper organisation.

Author contributions. AA defined the scope of the research. AA, JM, TB, MRC collected and interpreted data. AA, MM conceived and designed the experiments. JM, AA and YYC performed the analysis. HM, ND, TB and RM provided assistance in interpreting the results. AA and JM wrote the paper with significant inputs from MRC and YYC. AA supervised the research. All authors discussed the results and commented on the manuscript.

References

[1] D. Aldous and U. Vazirani. "Go with the winners" algorithms. In *Proceedings 35th Annual Symposium on Foundations of Computer Science*, pages 492–501. IEEE, 1994.

[2] N. C. Banks, D. R. Paini, K. L. Bayliss, and M. Hodda. The role of global trade and transport network topology in the human-mediated dispersal of alien species. *Ecology letters*, 18(2):188–199, 2015.

[3] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes on complex networks*. Cambridge university press, 2008.

[4] A. Biondi, R. Guedes, and F. Wan. Ecology, worldwide spread, and management of the invasive South American tomato pinworm, *Tuta absoluta*: Past, present, and future. *Annual Review of Entomology*, (63):239–258, 2017.

[5] R. A. Bradhurst, S. E. Roche, I. J. East, P. Kwan, and M. G. Garner. A hybrid modeling approach to simulating foot-and-mouth disease outbreaks in Australian livestock. *Frontiers in Environmental Science*, 3:17, 2015.

[6] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[7] L. Carrasco, J. Mumford, A. MacLeod, T. Harwood, G. Grabenweger, A. Leach, J. Knight, and R. Baker. Unveiling human-assisted dispersal mechanisms in invasive alien insects: integration of spatial stochastic simulation and phenology models. *Ecological Modelling*, 221(17):2068–2075, 2010.

[8] C. G. Cassandras, C. G. Panayiotou, G. Diehl, W. Gong, Z. Liu, and C. Zou. Clustering methods for multiresolution simulation modeling. In *Enabling Technology for Simulation Science IV*, volume 4026, pages 37–49. International Society for Optics and Photonics, 2000.

[9] N. J. Cunniffe, B. Koskella, C. J. E. Metcalf, S. Parnell, T. R. Gottwald, and C. A. Gilligan. Thirteen challenges in modelling plant diseases. *Epidemics*, 10:6–10, 2015.

[10] N. Desneux, E. Wajnberg, K. A. Wyckhuys, G. Burgio, S. Arpaia, C. A. Narváez-Vasquez, J. González-Cabrera, D. C. Ruescas, E. Tabone, J. Frandon, et al. Biological invasion of European tomato crops by *Tuta absoluta*, ecology, geographic expansion and prospects for biological control. *Journal of Pest Science*, 83(3):197–215, 2010.

[11] J. Douma, M. Pautasso, R. Venette, C. Robinet, L. Hemerik, M. Mourits, J. Schans, and W. van der Werf. Pathway models for analysing and managing the introduction of alien plant pests an overview and categorization. *Ecological Modelling*, 339:58–67, 2016.

[12] R. Early, B. A. Bradley, J. S. Dukes, J. J. Lawler, J. D. Olden, D. M. Blumenthal, P. Gonzalez, E. D. Grosholz, I. Ibañez, L. P. Miller, et al. Global threats from invasive alien species in the twenty-first century and national response capacities. *Nature Communications*, 7, 2016.

[13] J. M. Epstein. Why model? *Journal of Artificial Societies and Social Simulation*, 11(4):12, 2008.

[14] A. Fadikar, D. Higdon, J. Chen, B. Lewis, S. Venkatramanan, and M. Marathe. Calibrating a stochastic, agent-based model using quantile-based emulation. *SIAM/ASA Journal on Uncertainty Quantification*, 6(4):1685–1706, 2018.

[15] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):93, 2019.

[16] R. Y. Guimapi, S. A. Mohamed, G. O. Okeyo, F. T. Ndjomatchoua, S. Ekesi, and H. E. Tonnang. Modeling the risk of invasion and spread of *Tuta absoluta* in Africa. *Ecological Complexity*, 28:77–93, 2016.

[17] M. S. Hossain, M. Y. Mian, and R. Muniappan. First record of *Tuta absoluta* (Lepidoptera: Gelechiidae) from Bangladesh. *Journal of Agricultural and Urban Entomology*, 32(1):101–105, 2016.

[18] P. E. Hulme. Trade, transport and trouble: managing invasive species pathways in an era of globalization. *Journal of Applied Ecology*, 46(1):10–18, 2009.

[19] O. Karadjova, Z. Ilieva, V. Krumov, E. Petrova, V. Ventsislavov, et al. *Tuta absoluta* (Meyrick) (Lepidoptera: Gelechiidae): Potential for entry, establishment and spread in Bulgaria. *Bulgarian Journal of Agricultural Science*, 19(3):563–571, 2013.

[20] I. Z. Kiss, D. M. Green, and R. R. Kao. The network of sheep movements within great britain: network properties and their implications for infectious disease spread. *Journal of the Royal Society Interface*, 3(10):669–677, 2006.

[21] F. Lamperti, A. Roventini, and A. Sani. Agent-based model calibration using machine learning surrogates. *Journal of Economic Dynamics and Control*, 90:366–389, 2018.

[22] D. Martinetti and S. Soubeyrand. Identifying lookouts for epidemio-surveillance: application to the emergence of *Xylella fastidiosa* in France. *Phytopathology*, 109(2):265–276, 2019.

[23] J. C. Martins, M. C. Picanço, R. S. Silva, A. H. Gonring, T. V. Galdino, and R. N. Guedes. Assessing the spatial distribution of *Tuta absoluta* (Lepidoptera: Gelechiidae) eggs in open-field tomato cultivation through geostatistical analysis. *Pest management science*, 74(1):30–36, 2018.

[24] P. Moustier and H. Renting. Urban agriculture and short chain food marketing in developing countries. *Cities and Agriculture—Developing Resilient Urban Food Systems; de Zeeuw, H., Drechsel, P., Eds*, pages 121–138, 2015.

[25] U. Nations. Sustainable development goals, Apr. 2019.

[26] J. F. H. Nopsa, G. J. Daglish, D. W. Hagstrum, J. F. Leslie, T. W. Phillips, C. Scoglio, S. Thomas-Sharma, G. H. Walter, and K. A. Garrett. Ecological networks in stored grain: Key postharvest nodes for emerging pests, pathogens, and mycotoxins. *BioScience*, page biv122, 2015.

[27] M. Pautasso, X. Xu, M. J. Jeger, T. D. Harwood, M. Moslonka-Lefebvre, and L. Pellis. Disease spread

in small-size directed trade networks: the role of hierarchical categories. *Journal of Applied Ecology*, 47(6):1300–1309, 2010.

[28] R. G. Pearson. Species' distribution modeling for conservation educators and practitioners. *Synthesis. American Museum of Natural History*, 50:54–89, 2007.

[29] G. Strona, C. J. Carstens, and P. S. Beck. Network analysis reveals why *Xylella fastidiosa* will persist in Europe. *Scientific reports*, 7(1):71, 2017.

[30] S. E. Sylla, T. Brévault, L. Monticelli, K. Diarra, and N. Desneux. Geographic variation of host preference by the invasive tomato leafminer *Tuta absoluta*: implications for host range expansion. *Journal of Pest Science, Accepted*, 2018.

[31] H. E. Tonnang, S. F. Mohamed, F. Khamis, and S. Ekesi. Identification and risk assessment for worldwide invasion and spread of *Tuta absoluta* with a focus on Sub-Saharan Africa: implications for phytosanitary measures and management. *PloS one*, 10(8):e0135283, 2015.

[32] USDA. New Pest Response Guidelines: Tomato Leafminer (*Tuta absoluta*). *Animal and Plant Health Inspection Service, Plant Protection and Quarantine*, 2012.

[33] S. Venkatramanan, S. Wu, B. Shi, A. Marathe, M. Marathe, S. Eubank, L. Sah, A. Giri, L. Colavito, K. Nitin, et al. Modeling commodity flow in the context of invasive species spread: Study of *Tuta absoluta* in Nepal. *Crop Protection*, 2019.

[34] M. Wildemeersch, O. Franklin, R. Seidl, J. Rogelj, I. Moorthy, and S. Thurner. Modelling the multi-scaled nature of pest outbreaks. *Ecological Modelling*, 409:108745, 2019.

[35] W. Yang, D. R. Olson, and J. Shaman. Forecasting influenza outbreaks in boroughs and neighborhoods of New York city. *PLoS Computational Biology*, (11):1–19.

[36] L. You, U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. Spatial Production Allocation Model (SPAM) 2005 v3.2. http://mapspam.info, 2017.

Supplementary material: Assessing the Multi-pathway Threat from an Invasive Agricultural Pest: *Tuta absoluta* in Asia

Joseph McNitt[1], Young Yun Chungbaek[2], Henning Mortveit[2], Madhav Marathe[2], Mateus R Campos[3], Nicolas Desneux[3], Thierry Brévault[4,5,6], Rangaswamy Muniappan[7], and Abhijin Adiga[2]

[1]Epic Systems Corporation, United States

[2]Biocomplexity Institute & Initiative, University of Virginia

[3]French National Institute for Agricultural Research

[4]BIOPASS, CIRAD-IRD-ISRA-UCAD, Dakar, Senegal

[5]CIRAD, UPR AIDA, F-34398 Montpellier, France

[6]Université de Montpellier, CIRAD, Montpellier, France

[7]Feed the Future Integrated Pest Management Innovation Lab, Virginia Tech

S1 Data

Table S1 lists all publicly available datasets incorporated in our model and analysis. Table S2 provides country-specific data sources, literature and reports used in the model and for validation. They are further tagged based on the information they provide for different aspects of the model. Table S3 provides incidence reports from Bangladesh. Code is available in Github [2].

S2 *T. absoluta* biology.

The tomato leafminer exhibits a short life cycle of about 24–38 days (temperature at $25 \pm 3°$C), from egg to adult, as it is a multivoltine species with overlapping generations in the field. It causes serious damage to numerous solanaceous

Table S1: Global data sets used for model construction.

Data	Source	Spatial or temporal resolution
Temperature	WorldClim [26]	2.5 arc min., monthly.
Precipitation	GPCC [14]	2.5 arc deg., monthly.
Humidity	NASA [47]	1 arc deg., monthly.
NDVI	NEO [46]	0.1 arc deg., monthly.
Country Production	FAO [16]	country, yearly.
Shape files of study region	ADC Worldmap [4]	country/province/district.
Cell Production	Mapspam [66]	5 arc min., yearly. Gives vegetable and potato data.
Consumption	FAO [18]	country, yearly.
Population	Landscan[37]	5 arc min.
Cities	MAXMIND [40]	city resolution
Distance between cities	Google API [20]	city resolution

Table S2: Country-specific data obtained from reports, peer-reviewed articles and experts' inputs.

Country	Seasons	Production	Consumption	Domestic trade		Processing	International trade
				Markets	Flows		
Bangladesh	[6]	[6]	[17]	[6]	–	[6, 64]	[15]
Cambodia	[8, 19, 53]	[19, 53]	[53]	[53]	[53]	–	[45, 53]
Indonesia	[5, 21]	[5, 21]	[17]	–	–	–	[17]
Laos	[34]	[34]	[34]	[34]	[34]	–	[34, 45]
Malaysia	–	[23, 42]	[17]	–	–	–	[17]
Myanmar	–	[36, 51]	[44]		[36]		
Philippines	[7, 50]	[50]	[11]	[12]			ignored [12]
Singapore	–	[17]	imp.-exp.	city	ignored		
Thailand	[30]	[30, 43, 61]	[61]	[30, 61]	[30]	[43]	
Vietnam	[28, 63, 65]	[65, Table 16] (2003)	[65, Table 23]	[9, 45, 65]	[9, 45, 65]	[3, 65]	[45, 65]

Table S3: Data of *T. absoluta* infestation for Bangladesh (source: Bangladesh Agricultural Research Institute) and neighbouring region of India.

Location	Coordinates	Month of first detection of *T. absoluta*
Chaklarhat, Panchagarh district	26.19° N, 88.43° E	May 2016 [27]
Umiam [52], BARI Research Field, Gazipur district	23.99° N, 90.42° E	January 2017
RARS, Akbarput, Moulvibazar district	24.25° N, 91.45° E	February 2017
Lalakhal, Jointiapur, Sylhet district	25.07° N, 92.09° E	February 2017
Palashbari, Gaibandha district	25.15° N, 89.23° E	March 2017
Battara, Bogra district	24.48° N, 89.25° E	February 2017
Barura, Comilla district	23.22° N, 91.06° E	March 2017
Patengali, Jessore district	23.09° N, 89.10° E	March 2017
Umiam, Meghalaya India		January 2017 [52]

crops such as eggplants, potatoes, and especially tomato crops ~~[59]~~[57]. It penetrates into tomato leaves, stems, or fruits, wherein it feeds and develops by creating conspicuous mines as well as galleries. Considering the warm weather throughout the year, particularly in the dry season, the study region presents ideal conditions for rapid development

and spread of *T. absoluta*. Pest risk analysis [60] shows that the Ecoclimatic Index for this region is above 50 (highly suitable). Spatial distribution assessment survey of *T. absoluta* eggs has shown its high dispersive capacity in tomato producing areas [39]. The dispersion in a tomato cultivation starts mainly at the periphery and the pest is able to migrate between tomato farms to generate egg aggregation at the crop boundaries. The pest spread behaviour among seasonal crop resources is often non-random and directional [39]. Sylla et al. ~~[59]~~ [57] analysed host preference of *T. absoluta* in France and Senegal. While the highest preference is for tomato, it can survive well on eggplant and potato, which happen to be major vegetable crops in the study region. However, since *T. absoluta* primarily attacks leaves of eggplant and potato, the chance of the pest spreading through trade of these crops seems to be low.

S3 Multi-pathway Model

S3.1 Model structure, parameter ranges and assumptions

Sylla et al. [58] show that relative area of host crops (particularly tomato and eggplant) is positively correlated with number of trapped moths/trap/day. The authors also show that higher the host density, the greater the proportion of infected plants. They use data from open field experiments and generalized linear mixed model (GLMM) to establish these relationships. Hence, we model the infectiousness of a cell $\rho(v,t)$ as a linear function of host density at time t, for which we use the weighted sum of production volume of tomato, eggplant, and potato in that cell at time t. The weights correspond to relative carrying capacity of *T.* absoluta on the three hosts. [57].

Transmission probability. At any given time t, for a cell v and its neighbour v' corresponding to the considered pathway, let $\rho(v',t)$ be the strength of infestation at v'. Let, α be the transmission rate corresponding to the pathway. Then, the probability that the pest will be introduced to v from v' is given by $1-\exp\big(-\alpha\rho(v',t)\big)$. The function form is similar to that in [41, 48], but with contact time between two nodes replaced with strength of infestation. Also, α can be interpreted as the transmission probability per unit infestation. Since in our case production in the cell is used as a surrogate for $\rho(v',t)$, it can be interpreted as the probability of infection per unit production (weighted by carrying capacity). This form of the infection probability function has been seen before in epidemiological models [13, 24, 25], both continuous and discrete time models. Diekmann et al. [13] consider a continuous time SI model, and define the probability that an individual will not be infected with a function which relates to the form above, taking the exponent as an integral of the "force of infection". more complexity, but we also see some from the "proportional hazards model". continuous time model and an exponential remaining susceptible. In [24], we see a discrete time version of this function form, with the probability of still being exposed after t time steps as: $P(t)=e^{-\epsilon t}$, where ϵ

relates to the "transfer rate". This function form was derived when considering traditional SIR models mathematically in [25].

For long-distance dispersal we use the same function form (equation 3 in the main paper), with the infectiousness of a cell replaced by infectiousness of a locality times the flow from that locality to the current cell's locality. In the literature, there is no clear consenus on the function form. For example, Stanaway et al. [54] model the colonisation time as an exponential distribution with mean being the flow from source to target cell. Stansbury et al. [55] calculate the probability of at least one establishment (E) using a binomial function: $E = 1 - (1 - p_t p_c)^I$, where, I is the number of successful entries in a year, p_t is the proportion of contaminated ports, p_c is the proportion of those ports that result in establishment. Caton et al. [10] apply Central Limit Theorem to model the number of insects arriving on contaminated flights by region.

Following Guimapi et al. [22], the Moore range r_{M} was varied between 1 and 3 corresponding to spread per month of approximately 25km, 50km and 75km respectively.

S3.2 Locality construction

In the model, *localities* are centres of consumption or production. From the perspective of consumption, we selected cities with population greater than a certain population threshold (as per [40]) in the entire study region. The number and size of localities is controlled by population threshold and locality radius parameters. We considered a range of threshold values, and chose $250,000$ as the threshold for the model with the main criterion for the choice being coverage of population and knowledge of major wholesale markets (Figure S1). Then, we added major production centres if their population did not meet the threshold. The locality radius was chosen to be 100km since local production in an urban area was within 50–60km from the city centre for several countries in this region [8, 53, 65]. To obtain long distance trade flows, the travel times between pairs of cities were computed using Google API [1].

S3.3 Seasonal production

We estimated monthly production volume of tomato, eggplant and potato for each cell using production data available at the country/state/province level and the Spatial Production Allocation Model [66]. The latter uses a generalised cross-entropy approach to allocate crop production in administrative units into individual pixels, through judicious interpretation of all accessible evidence such as production statistics, farming systems, satellite image, crop biophysical suitability, crop price, local market access and prior knowledge. The monthly production volume was estimated in two steps: (i) Estimation of cell-level annual production, and (ii) disaggregation of annual production to monthly production.

(a) (b)

Figure S1: **Locality construction.** The number of localities and the locality area is controlled by population threshold and locality radius. We analysed the effect of these parameters on the amount of production and population captured in the region. Setting a threshold of 250,000 and radius of 100km, 80% of the production and population is covered by localities.

Annual production. From SPAM [66], we obtained annual production estimates for each cell. However, there are several issues with directly using this data. Firstly, these are estimates for the year 2005, and secondly, tomato and eggplant production estimates are not available. Instead, total vegetable production volume is available. Also, for countries where data was available, we did not find any correlation between reported tomato (eggplant) production and total SPAM vegetable production for that region. Therefore, for each country, we obtained the most recent production data available (2013 or later) at the highest spatial resolution (region/province/country) (Table S1). The production of a particular vegetable type at a cell was computed as follows:

$$\frac{\text{Total production in the region}}{\text{Total SPAM production for cells in the region}} \times \text{SPAM production in the cell} .$$

For tomato and eggplant, we used vegetable production as the surrogate. There were also cases where no data was available (Cambodia, Myanmar and Laos for example). In such cases, for potato, we used SPAM data as is. For tomato and eggplant, the SPAM value for vegetables was scaled by a scaling factor which was determined as follows. For countries where data was available, we computed the ratio of total tomato production and total SPAM vegetable production for the country. The median value (≈ 0.05) was used as the scaling factor. The same procedure was used for eggplant.

Monthly production. In order to estimate the seasonal production rates of tomatoes, we considered quarterly production data from Philippines [50] for the 16 regions of the country and studied its relationship between precipitation,

elevation and temperature. We first assigned average monthly precipitation, temperature and elevation data to the cells by latitude and longitude. Next, for each region, we obtained the seasonal relative tomato production (or production rate). This was obtained by dividing each of the quarterly production volume by total annual production volume. We conducted a linear regression with the product rate as a dependent variable and the precipitation as an independent variable in SPSS 24.0. Since the dependent variable was highly skewed, we implemented a log-transformation for it. To control elevation, we classified the elevations into two groups, high and low, using K means in SPSS 24.0. Due to the small sample size, we excluded the samples in the high-elevation group and conducted a linear regression analysis for the group of low elevation (< 235).

The regression results showed that precipitation was a statistically significant predictor ($p < 0.001$, $R^2 = 0.54$). When we accounted for temperature along with precipitation, the R value increased to 0.58 with temperature exhibiting weak correlation with production. Though the results with both variables included gave a slightly stronger correlation, in the validation step, the regression function corresponding to precipitation was a better match for the rest of the study region. Thus we decided to use only precipitation as a predictor for seasonality of tomato production. Table S4 provides information on these regression results.

Table S4: Linear regression results for seasonal production as a function of precipitation.

	B	Std. Error	Beta	t	Sig.
Intercept	-0.208	0.229		-0.908	0.368
Precipitation	-0.008	0.001	-0.734	-7.935	0.000

Validation. With the exception of data from Philippines on tomato and eggplant, there is no regional seasonal production information for the three crops considered for other countries. However, there are several reports that provide qualitative information. See Table S2, column "Seasons". In some cases seasonal production data is available at the country level. For example, for Bangladesh, eggplant production volume is available by season. Similar information is available for Cambodia. In Figure S2, we have compared seasonal production predicted by the regression model with available qualitative and quantitative information. We note that the precipitation based regression function captures the general trend of seasonal production in the different regions considered.

S3.4 Network construction

For each country, the domestic flow was estimated using a doubly constrained gravity model [32, 62]. For a city i, let O_i and I_i denote total outflow and total inflow respectively, and $\mathrm{L}(i)$ denotes all cells which are assigned to it. The flow F_{ij} from city i to city j is given by $F_{ij}(t) = a_i(t)b_j(t)O_i(t)I_j(t)f(d_{ij})$, where, d_{ij} is the time to travel

(a) Bangladesh tomato production compared with growing season information [6]

(b) Bangladesh potato production compared with growing season information [6]

(c) Bangladesh eggplant production compared with growing season information [6]

(d) Indonesia tomato production compared with growing season information [5]

(e) Cambodia tomato production compared with seasonal production information [19]

(f) Laos tomato production compared with growing season information [34]

(g) Philippines tomato production compared with growing season information [50] (North)

(h) Philippines tomato production compared with growing season information [50] (Central and South)

(i) Philippines tomato production compared with growing season information [50]: Regions with mismatch.

Figure S2: Comparing relative production as per the regression model with quantitative/qualitative reports for different regions. For the plots corresponding to the regression model we chose representative cells for each administrative region.

from i to j, and $f(\cdot)$ is the *distance deterrence function*: $d_{ij}^{-\beta}\exp(-d_{ij}/\kappa)$, where β and κ are tunable parameters. The coefficients a_i and b_j are computed through an iterative process such that the total outflow and total inflow at each node agree with the input values [32]. Overall, we have 12 networks representing flows for each month. The outflows

and inflows are calculated as follows:

$$O_i(t) = \mathrm{Prod}(i,t) + \mathrm{Import}(i,t) - \mathrm{Export}(i,t) - \mathrm{Proc}(i,t), \tag{1}$$

$$I_i(t) = \mathrm{Pop}(i)\,. \tag{2}$$

Here, $\mathrm{Prod}(i,t)$ is the monthly production and $\mathrm{Pop}(i)$ is the population of the locality as a surrogate to consumption. The latter is the sum total of population in every cell that belong to the locality. Export and Import are the monthly total export to and import from outside the country respectively. $\mathrm{Proc}(i,t)$ is the tomato produced for processing. Since for this purpose tomato is typically cultivated and consumed locally, we subtract this volume from the outflow. Country specific details of how locality attributes were estimated is in Section S3.5.

S3.5 Locality production, consumption, imports, exports and processing

Monthly tomato production (consumption) at a locality was obtained by aggregating production (population) at all cells that belong to it. Population data was obtained from Landscan (Table S1). For most countries, imports and exports are a small fraction of domestic production. The main exceptions were the significant tomato imports from India to Bangladesh and trade between Malaysia and Singapore. We identified major routes of trade from India to Bangladesh [15] and the total imports from India (FAOSTAT) was distributed uniformly between three cities close to the border with India along these routes. Finally, these imports were evenly distributed for the later half of the year since Bangladesh imports mostly during the rainy season. To capture the significant trade between Singapore and Malaysia, we included Singapore in the domestic flow network of Malaysia as there is high interaction between the two countries. The resulting flow from the gravity model from Malaysia to Singapore was obtained by aggregating network flows across months and across edges with Singapore as destination. This flow was comparable to the annual imports from Malaysia to Singapore. With the exception for Thailand [43], Vietnam [65] and Cambodia (pers. comm.), there is no information on the amount of tomato production consumed by the processing industry even though there is evidence of processing industry in Malaysia and Indonesia. For each locality in Thailand, we scaled the monthly production by the ratio annual production of fresh tomatoes over total annual production (fresh and processed).

For consumption, we used country-level production [18], which was available for half of the countries. For Singapore, we estimated it as the difference between total inflow (production and imports) and total outflow (exports) based on FAO data. For Vietnam, Wijk et al. [65] provides this information. For Myanmar, Cambodia and Laos, we found no information. We used median consumption for the region. We also analysed consumption with respect to per capita gross domestic product (GDP). However, we did not find any correlation between GDP and consumption both globally

as well as restricted to the study region. In Venkatramanan et al. [62], consumption was modelled as a function of population and Gross Domestic Product (GDP). In the study region, analysing FAO data on consumption [18] we did not find any relation between GDP and consumption.

S3.6 Validation of domestic trade flows

Since no market-to-market flow data is available for the region, we searched the literature for evidence of tomato trade between cities or regions for each country. Even this information is hardly available. The next step was to further generalise the search by considering the flow of vegetables. The results of our comparison of the networks obtained using gravity flow model with literature are shown Figure S3. Our general observations as follows: (i) To obtain good representation of trade flow, at minimum, region/state level data of production is required; and (ii) densely populated urban pockets are a good representation of consumption centres. Some country specific details follow.

Bangladesh. There is evidence of vegetable flow from Rangpur region to Dhaka particularly during winter [29, 56]. The flow from Southwest Bangladesh to Dhaka [33] is captured by the edge from Khulna to Dhaka (Figure S3a). We assigned import of tomato from India to cities in the border based on information on important trade routes [15]. Since, there is no data on volume, we distributed it evenly.

Vietnam. There are primarily two consumption centres: Ho Chi Minh City in the south supplied by Central Highlands and the Mekong River Delta, and Hanoi-Haiphong urban centres in the north supplied by Red River Delta. For Ho Chi Minh City, Lam Dong is the main district that supplies vegetables [9]. In our model (Figure S3d), this flow is represented by Buon Me Thuot–a city in Central Highlands–to Ho Chi Minh City. Buon Me Thuot is less than 100km away from the Lam Dong province thus covering the production in the area. In the north, much of the production surplus is assigned to the Haiphong locality, and therefore, it acts as a big source for almost all localities in the north. Our network also has some long distance edges from north to south. This is mostly due to the large population in the south. There is not much evidence of this flow and it could be a result of not accounting for heterogeneity in consumption.

Malaysia and Singapore. We combined the locality corresponding to Singapore with Malaysia, so that the flow between these two countries is treated as a domestic flow as there is free movement of people and goods between these countries. The gravity model flows from Malaysia to Singapore is between 50,000–60,000 tonnes compared to $\approx 30,000$ tonnes according to FAOSTAT. However, our network does not capture the flow from Cameron Highlands,

(a) Bangladesh (b) Myanmar (c) Thailand

(d) Vietnam (e) Malaysia and Singapore (f) Indonesia

(g) Philippines

Index	Citation
1	Daily Sun 2018 [56]
2	Karim et al. 2016 [33]
3	Cadilhon et al. 2006 [9]
4	Wijk et al. 2007 [65]
5	Ledesma 2011 [38]
6	Johnson et al. 2008 [31]

(h) References corresponding to trade flow evidence.

Figure S3: **Long-distance annual trade flows resulting from the gravity model.** We combined monthly flows and to avoid clutter, we have displayed only major flows (> 500). Some of the edges are labelled based on information from literature. The references corresponding to the labels are listed in (h). For Laos, Cambodia and Brunei, there are no flows as there is only one locality in each country.

major tomato producer to cities of Malaysia. This shortcoming is mainly due to unavailability of state/province level tomato production information.

Philippines. The major production centres are in Cagayan de Oro and regions north of Manila. These are captured by our network (Figure S3g).

S3.7 ~~Transmission probability~~

~~At any given time t, for a cell v and its neighbour v' corresponding to the considered pathway, let $\rho(v', t)$ be the strength of infestation at v'. Let, α be the transmission rate corresponding to the pathway. Then, the probability that the pest will be introduced to v from v' is given by $1 - \exp\big(-\alpha\rho(v', t)\big)$. The function form is similar to that in [41, 48], but with contact time between two nodes replaced with strength of infestation. Also, α can be interpreted as the transmission probability per unit infestation. Since in our case production in the cell is used as a surrogate for $\rho(v', t)$, it can be interpreted as the probability of infection per unit production.~~

~~This form of the infection probability function has been seen before in epidemiological models [13, 24, 25], both continuous and discrete time models. Dickmann et al. [13] consider a continuous time SI model, and define the probability that an individual will not be infected with a function which relates to the form above, taking the exponent as an integral of the "force of infection". In [24], we see a discrete time version of this function form, with the probability of still being exposed after t time steps as: $P(t) = e^{-\epsilon t}$, where ϵ relates to the "transfer rate". This function form was derived when considering traditional SIR models mathematically in [25].~~

S4 Parameterization and simulation

For each parameter setting, we evaluated the model output by comparing it with historical invasion data from Bangladesh using a similarity score as defined in the main document (equation (4)). Simulations were run with 100 repetitions. From the output we computed the empirical probability that a cell is in state I at time t, which in turn was used to compute similarity scores. We applied Classification and Regression Trees (CART) to guide parameter space exploration. Initially, the parameter space was coarsely sampled. The parameter values are listed in Table 1 in the main document. For each of these samples, simulations were run and the similarity scores computed. With the model parameters as independent variables and similarity score as the dependent variable, we analysed the results using CART (see Figure S4). We chose parameter subspaces with high similarity score (0.7 or more) and rejected those with lower values. For each chosen parameter subspace, the process was repeated.

We recall that the emergent outcome of the model can be classified into two classes: Class A and Class B. In Figure S4, we show the results of CART analysis on the parameter sets which yielded a similarity score of ≥ 0.7. The chosen set is partitioned by Moore range r_{M} and latency period ℓ. For (r_{M}, ℓ) values that permit rapid range expansion through short-distance spread over the contiguous landscape, including long-distance spread leads to faster spread than observed leading to lower score. Hence, for those regimes, we observe only Class A spread. The corresponding (r_{M}, ℓ) values are $(1, 1)$, $(2, 1)$, $(2, 2)$ and $(3, 1)$. For the regime $(3, 1)$ we did not observe any instance with score > 5.5.

(a) $r_{\mathrm{M}} = 1$, $\ell = 1$

(b) $r_{\mathrm{M}} = 1$, $\ell = 2$

(c) $r_{\mathrm{M}} = 1$, $\ell = 3$

(d) $r_{\mathrm{M}} = 2$, $\ell = 1$

(e) $r_{\mathrm{M}} = 2$, $\ell = 2$

(f) $r_{\mathrm{M}} = 2$, $\ell = 3$

(g) $r_{\mathrm{M}} = 3$, $\ell = 2$

(h) $r_{\mathrm{M}} = 3$, $\ell = 3$

Figure S4: CART analysis of the parameter space: The models with similarity score ≥ 0.7 were partitioned based on the Moore range r_{M} and latency period ℓ tuple and analysed. For the case $(r_{\mathrm{M}} = 3, \ell = 1)$, there were no instances with similarity score ≥ 0.7 among the sampled instances. Minimum number of observations at a node for splitting was set to 50. Minimum number of observations at each leaf node was set to 20.

S4.1 Seeding scenarios

During the model evaluation phase, to account for both spatial and temporal observation noise, we considered six seeding scenarios: The locations Bangladesh 1 and Bangladesh 2 (see Table S5) and start times March, April or May (the month *T. absoluta* was reported). For the prediction phase, we constructed three different seeding scenarios reflecting three different possible introductions to the region. The first corresponds to spread from Bangladesh through

northeast Myanmar. This region represents a possible pathway of *T. absoluta* should it spread radially through the border with Bangladesh. The second scenario we considered is the region of Johor in Malaysia - including Singapore. This region contains major ports in Malaysia including Tanjung Pelepas [35], and Singapore is also a major trading hub for Bangladesh. We considered two scenarios of introduction for Philippines. The first is motivated by the fact that there is a large migrant Filipino worker population travelling to and from infested countries. We seeded the port of Masinloc in this scenario as it processes many of these travellers. The second was the introduction to Northern Mindanao region, one of the major tomato production areas in the country. We conducted country-specific experiments in the case of Thailand and Vietnam. The seed locations were determined by spread observed from both Class A and Class B models.

Table S5: Seeding scenarios

Region seeded	Description
Bangladesh 1	Parameterization phase: The cell corresponding to the location of first report was seeded.
Bangladesh 2	Parameterization phase: The cells corresponding to the location of first report and its adjacent cells (Moore neighbourhood 1) were seeded.
Bangladesh and Northeastern Myanmar	This seeding is the estimated current range of *T. absoluta* in the study region. It is used to study the spread in the rest of the region.
Singapore and southern (mainland) Malaysia	This reflects possible introduction through trade, as Singapore is a major port.
Philippines 1	This reflects the possibility of *T. absoluta* being introduced through workers travelling to and from India, and other countries, which have reported the pest. The port of Masinloc was seeded.
Philippines 2	Northern Mindanao region
Thailand	Northwestern region: informed by observed spread in Mainland Southeast Asia.
Vietnam	Northwester region: Informed by observed spread in Mainland Southeast Asia.

S4.2 Software and Computational aspects.

The model was implemented in Python 2.7. For data management and processing, we used PostGreSQL and SQlite. Statistical analysis was done using Python, R 3.4.3 and SPSS 24.0. The experiments were run using Discovery, a high performance computing cluster with 232 nodes (16-core Sandy Bridge-EP E5-2670 2.60GHz (3.30GHz Turbo) Dual Processor (8 Cores per Processor) nodes with 32 GB of Memory and 500 GB Internal Hard Drive) in the Biocomplexity Institute of Virginia Tech.

S5 Clustering analysis of the spread pattern

The analysis process is illustrated in Figure ~~??~~ 1c We recall that each simulation output has the following format. Suppose, the simulations were run for T time steps. For every v, $p(v,t)$ is the empirical probability that cell v is infected at time t. Each simulation output can be viewed as a matrix of size $n \times T$, where n is the number of cells. To normalise each matrix, we introduced an extra index $T+1$ which carries the residual probability, i.e., $p(v, T+1) = \sum_{t=1}^{T} p(v,t)$. The simulation outputs for models with similarity $\mathcal{S} \geq 0.75$ were thus processed and clustered using two algorithms: hierarchical agglomerative clustering (SPSS 24.0) and k-means algorithm (using Pyclustering [49]). In each case, the number of clusters considered ranged from $k = 1$ to 10. ~~Outline of the process used for analysing the multi-pathway spread.~~

To discover relationships between model parameters and cluster membership, we cast it as a classification problem. With configurations as features and cluster index as labels, we applied CART. The resulting decision tree was used to interpret the relationship between parameters and cluster membership. The results are in Figures S5 (hierarchical) and S6 (k-means).

S6 More results

S6.1 International trade

Figure S7a shows the annual tomato trade for the study region. The network was constructed using data from FAOSTAT Trade matrix (Table S1). The trade within the focus region is represented by green edges. Imports from *T. absoluta* affected countries categorised by region is represented by red edges, and the blue edges correspond to exports from the focus region to countries yet to report the presence of the pest. The edge thickness is a function of the trade volume. We note that with the exception of Philippines, there is considerable trade activity among countries in the focus region. However, the majority of trading happens between neighbouring countries and except for a few flows, the volume is negligible compared to production within. Most significant flow is from Malaysia to Singapore ($\approx 30,000$ tonnes), India to Bangladesh ($\approx 20,000$ tonnes), and Thailand and Vietnam to Malaysia (between 1000 to $2,000$ tonnes). Philippines does not report any fresh tomato trade with other countries. We also analysed how the network structure evolved across years. While we did not observe much variation in the network structure, there is some change in the countries importing to and exporting from this region

General trends. Figure S7b shows the evolution of trade and production volume over time. The total tomato production in the study region obtained by summing the production for all countries is plotted for each year between 2004

(a) Hierarchy

(b) $k = 2$

(c) $k = 3$

(d) $k = 4$

(e) $k = 5$

(f) $k = 6$

(g) $k = 7$

(h) $k = 8$

(i) $k = 9$

(j) $k = 10$

Figure S5: **Agglomerative clustering of simulation outputs.** The 9-level hierarchy of the clusters is shown in (a). At each level, we applied CART to understand which parameter was significantly influencing the split. Here, `a_long` corresponds to $\alpha_{\ell d}$, `latency_period` is ℓ and `moore` is r_{M}.

(a) $k = 2$

(b) $k = 5$

(c) $k = 10$

Figure S6: k**-means clustering of simulation outputs.** For each k, the number of clusters, we applied CART to understand which parameter was significantly influencing the split. Here, `a_long` corresponds to $\alpha_{\ell d}$, `latency_period` is ℓ and `moore` is r_{M}.

(a) Tomato trade in the focus region (2013).

(b) General trends of overall production and trade in the region over a decade based on FAOSTAT 2013.

(c) Evolution over years 2011, 2012, 2013.

to 2013 (blue). We computed the total tomato import volume per year (green) from outside the study region using FAOSTAT Trade Matrix (Table S1). This was obtained by aggregating imports where the importing country belongs to the study region and the exporting country is outside the region. Similarly export volume per year to outside the study region (red) is shown. Finally, the amount of tomato traded within the study region was obtained by aggregating tomato trade between countries in the region for each year (orange). In order to compare these plots, each plot was normalised by dividing by the corresponding maximum value across all years. There is a steady increase in the production and amount of internal trade, with comparable rate of change for both quantities. In comparison, the export of tomato to outside of the focus region (plot "Exports") has risen steeply in the recent years (after 2011), while the imports generally indicate a downward trend (plot "Imports"). Recent efforts to increase production and trade infrastructure in these countries support these trends [8, 28, 45]. Under these circumstances, *T. absoluta*'s invasion can have a high negative impact on the economy and livelihood of the people in this region.

S6.2 Localities, production and trade

We identified 102 localities representing urban and production centres. For the purpose of this discussion, a *major source* (*major sink*) is a locality with net outflow (inflow) at least 10 Kilo tonnes. Among these, less than 20% of the localities have net outflow of at least 10 Kilo tonnes in the gravity model derived trade network, indicating that there are few major sources (Figure S8a). We also note that there are a few high sinks which have production comparable to some of the major sources. These are regions not only vulnerable to attacks, but also potentially sustain a huge negative impact due to loss in production.

To predict seasonal production, we used linear regression to model the production rate as a function of precipitation, temperature, and elevation using seasonal production data from different regions of Philippines. The results showed that precipitation was a statistically significant predictor ($p < 0.001$). More details are in Section S3.3. To study the effect of seasonality on monthly flows, we accumulated the net flow per month by directionality. For a locality i, let $F_i(m)$ denote the net flow in month m (positive means outflow, else inflow). The accumulated outflow is $F_i^+ = \sum_{m=1}^{12} I\big(F_i(m) > 0\big)F_i(m)$, where $I(\cdot)$ is the indicator function. Corresponding definition of accumulated inflow is $F_i^- = \sum_{m=1}^{12} I\big(F_i(m) < 0\big)F_i(m)$. The results in Figure S8b show that there are very few localities which switch from the role of source to sink and vice versa. This is unlike what is observed in Nepal by Venkatramanan et al. [62]. The flow between high altitude regions and low altitude regions switches direction according to season. Here, while the intensity of flows vary between seasons, we do not see such a switch in roles.

(a) Net annual production and outflow

(b) Effect of seasonality on monthly flows

Figure S8: **Production, inflow and outflow at localities.** These are results for $\beta = 2$ and $\kappa = 500$ (i)

(a) Similarity score ($\mathcal{S}$) based evaluation.

(b) Agglomerative clustering

(c) k-means clustering

Figure S9: **Parameter importance.** Continued from Figure 2c

S6.3 Cellular automata model from Guimapi et al

Guimapi et al. [22] use a cellular automata approach to model the spread of *T. absoluta* through the study region of Africa, Spain, and Portugal. Here, we implemented and applied the model to study the spread in the focus region.

Methodology. Each cell in the automata is of size $25\text{km} \times 25\text{km}$ induced by a grid overlayed on the study region. Both square and hexagonal cell configurations are considered, but square cells (Moore neighbourhood) were chosen in order to cover a larger area per cell. Each cell can be in one of three states: S, E and I with very similar definitions as in our model. The spread is modelled as follows: For an infected cell to spread to another, each cell within its moore neighbourhood of respective distance is considered in each time step (one month). If this neighbour cell is suitable, it will change states. The suitability of a cell for pest establishment depends on temperature, humidity, NDVI, and tomato production in that cell at the given time step. Two Moore neighbourhoods were considered, one with a range of 50 km and another with a range of 75 km.

Given a cell within the Moore neighbourhood of an infected cell, its state in the next time step depends on the following rules.

- If NDVI $\geq$ NDVI threshold 1, then state becomes ”Exposed”.
- If NDVI $\geq$ NDVI threshold 1 and temperature $>=$ temperature threshold and humidity $>=$ humidity threshold, then state becomes ”Invaded”.
- If NDVI $\geq$ NDVI threshold 2 and tomato production high or very high, then state becomes ”Invaded”.
- Otherwise the state remains unchanged

Temperature, humidity, and NDVI data is at the level of the cell, while production data is at the country-level. The temperature threshold is 22 degrees Celsius and humidity threshold is 55%. The NDVI threshold 1 is 0.1, and threshold 2 is 0.3. The production yield threshold is 10 tons/hectare to be considered high or very high. Thus if a cell is within a country with high or very high yield, that cell would be classified at high or very high yield.

Results. We applied the seeding scenarios described in Table S5. Each scenario was simulated with Moore neighbourhoods of 2 and 3. The results are shown in Figure S10. The model predicts a rapid radial spread in this region as vegetation and temperature are generally favourable throughout this region. The tomato production yield in each country in the focus region is in the high or very high categories, and thus did not limit the spread. Generally, the predicted range expansion is much faster compared to our models, particularly in the case where Moore range is 3.

(a) Moore neighbourhood of 2

(b) Moore neighbourhood of 3, baseline model

Figure S10: Guimapi et al. [22] model results: Here the region of Chin in northeast Myanmar was seeded.

S6.4 ~~Spread patterns and rate~~

~~Class A —— Class B~~ ~~**Spread pattern in Mainland Southeast Asia without accounting for trade between countries.**~~

~~The contour plots show the simulated spread starting from northern Myanmar for 120 time steps or 10 years. Representative simulation outputs for Class A and Class B models are shown. With Class A, spread is faster eastward than southward and the other way around in the case of Class B.~~

(a) Class A

(b) Class B

Figure S11: **Spread in Mainland Southeast Asia with respect to distance from origin.** The cells are binned based on their distance from the origin of infection (Northern Myanmar). Given time step t (60 or five years from time of start in this case), let $\Pr(v, \leq t)$ be the probability that cell v is in state I by time t. For each parameter instance, we computed the "total infection" for every bin at time t by aggregating $\Pr(v, \leq t)$ for each v in the bin. Configurations were grouped by model class and Moore range (r_{M}). The average total infection for each group are plotted. The red points referred to as "max" correspond to the total number of cells in each bin, which is also the maximum possible accumulated probability for that bin. We observe that even though the models exhibit similar spread for Bangladesh, there is high variance in spread rate in both classes for spread in in the case of he spread rate when applied to the rest of the region. Also, the range of expansion is influenced by Moore range r_{M}.

(a) Bangladesh: without intervention

(b) Thailand: without intervention

(c) Vietnam: without intervention

(d) Bangladesh: with intervention (100%)

(e) Thailand: with intervention (100%)

(f) Vietnam: with intervention

(g) Bangladesh: spread intensity and distance with respect to origin

(h) Thailand: spread intensity and distance with respect to origin

(i) Vietnam: spread intensity and distance with respect to origin

Figure S12: **Interventions.** This is a continuation of Figure 3 from the main document.

References

[1] Google API. Retrieved March, 2018.

[2] Abhijin Adiga. Spread multi-pathway model. https://github.com/rmuniappan/SPREAD_model.git, 2019.

[3] Japan International Cooperation Agency. Data collection survey on selecting the processed food to be focused and promoting foreign direct investment in food business in Vietnam.

[4] American Digital Cartography, Inc. Adc worldmap v. 7.4, 2018. Last accessed: June 2018.

[5] I. W. Arsanti and M. H. Bohme. Assessing vegetable cropping patterns in upland areas of Indonesia. *Acta Horticulturae*, 1103:37–42, 2015.

[6] Bangladesh Bureau of Statistics. Yearbook of Agricultural Statistics. http://bbs.portal.gov.bd, 2017.

[7] P. J. Batt, M. T. Lopez, J. T. Axalan, L. A.T. Hualda, M. O. Montiflor, and S. B. Concepcion. Exploring the institutional market for fresh vegetables in the Southern Philippines. *Acta Horticulturae*, 895:59–68, 2011.

[8] B. Buntong, V. Srilaong, T. Wasusri, S. Kanlayanarat, and A. L. Acedo. Reducing postharvest losses of tomato in traditional and modern supply chains in Cambodia. *International Food Research Journal*, 20(1):233–238, 2013.

[9] Jean-Joseph Cadilhon, Paule Moustier, Nigel D Poole, Thi Giac Tam Phan, and Andrew P Fearne. Traditional vs. modern food systems? Insights from vegetable supply chains to Ho Chi Minh City (Vietnam). *Development Policy Review*, 24(1):31–49, 2006.

[10] Barney P Caton, Thomas T Dobbs, and Charles F Brodel. Arrivals of hitchhiking insect pests on international cargo aircraft at miami international airport. *Biological Invasions*, 8(4):765–785, 2006.

[11] Sylvia B Concepcion. Consumer market segments in the Philippine vegetable industry. *IFAMA Forum and Symposium*, pages 1–8, 2009.

[12] Sylvia B Concepcion. Institutional markets for fresh vegetables Trends. pages 1–30, 2011.

[13] Odo Diekmann and Johan Andre Peter Heesterbeek. *Mathematical epidemiology of infectious diseases: model building, analysis and interpretation*, volume 5. John Wiley & Sons, 2000.

[14] Earth System Research Laboratory. CPC Merged Analysis of Precipitation (CMAP). https://www.esrl.noaa.gov/psd/data/gridded/data.cmap.html.

[15] Export-Import Bank of India. Bangladesh: A Study of India's Trade and Investment Potential. 2015.

[16] FAO. Food supply quantity. http://www.fao.org/faostat/en/#data/CC, 2016.

[17] FAO. Production and trade. http://www.fao.org/faostat/en/#data, 2016.

[18] FAO. Per capita tomato consumption. http://www.helgilibrary.com/indicators/tomato-consumption-per-capita/world/, Accessed in February 2018.

[19] Christian Genova II, Katinka Weinberger, Srun Sokhom, Mong Vanndy, and En Chan Yarith. *Postharvest Loss in the Supply Chain for Vegetables–The Case of Tomato, Yardlong Bean, Cucumber and Chinese Kale in Cambodia*. AVRDC-WorldVegetableCenter, 2006.

[20] Google. Distance Matrix API. https://developers.google.com/maps/documentation/distance-matrix/, 2017.

[21] G J H Grubben. Timing of vegetable production in Indonesia. In *VI Symposium on the Timing of Field Production of Vegetables*, pages 261–270, 1989.

[22] Ritter YA Guimapi, Samira A Mohamed, George O Okeyo, Frank T Ndjomatchoua, Sunday Ekesi, and Henri EZ Tonnang. Modeling the risk of invasion and spread of *Tuta absoluta* in Africa. *Ecological Complexity*, 28:77–93, 2016.

[23] S H Hengky. Competition Gaps of Tomatoes' Industries in Cameron Highland, Malaysia. *Business and Economic Research*, 6(2):368, 2016.

[24] Herbert W Hethcote. A thousand and one epidemic models. In *Frontiers in mathematical biology*, pages 504–515. Springer, 1994.

[25] Herbert W Hethcote and Simon A Levin. Periodicity in epidemiological models. In *Applied mathematical ecology*, pages 193–211. Springer, 1989.

[26] Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International journal of climatology*, 25(15):1965–1978, 2005.

[27] M. S. Hossain, M. Y. Mian, and R. Muniappan. First record of *Tuta absoluta* (Lepidoptera: Gelechiidae) from Bangladesh. *Journal of Agricultural and Urban Entomology*, 32(1):101–105, 2016.

[28] Pham Thi Thu Huong, A. P. Everaarts, J. J. Neeteson, and P. C. Struik. Vegetable production in the Red River Delta of Vietnam. I. Opportunities and constraints. *NJAS - Wageningen Journal of Life Sciences*, 67:27–36, 2013.

[29] The Independent. Rangpur region to produce 7.02 lakh tonnes winter vegetables. http://www.theindependentbd.com/printversion/details/68010.

[30] K Itharattana. Market prospects for upland crops in Thailand. (November), 1996.

[31] Greg I. Johnson, Katinka Weinberger, and Mei-huey Wu. *The vegetable industry in tropical asia: Thailand. An overview of production and trade*. Number January 2008. 2008.

[32] Pablo Kaluza, Andrea Kölzsch, Michael T Gastner, and Bernd Blasius. The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103, 2010.

[33] Rubayet Karim and Jony Biswas. Value Stream Analysis of Vegetable Supply Chain in Bangladesh: A Case Study. *International Journal of Managing Value and Supply Chains*, 7(2):41–60, 2016.

[34] Somsack Kethonga, Khamtanh Thadavong, and Paule Moustier. Vegetable Marketing in Vientiane. (November):56, 2004.

[35] Nazery Khalid. The development of ports and shipping sector in malaysia. *Maritime Institute of Malaysia*, pages 1–15, 2005.

[36] Frauke Kraas, Hartmut Gaese, and Mi Mi Kyi. *Megacity Yangon: Transformation Processes and Modern Developments: Second German-Myanmar Workshop in Yangon, Myanmar, 2005*. Lit Verlag, 2006.

[37] LandScan Geographic Information Science & Technology. LandScan 2016. https://landscan.ornl.gov/.

[38] Merlyn P Ledesma. Northern Mindanao vegetables. Technical report, Department of Agriculture RFU- 10, 2011.

[39] Júlio C Martins, Marcelo C Picanço, Ricardo S Silva, Alfredo HR Gonring, Tarcísio VS Galdino, and Raul NC Guedes. Assessing the spatial distribution of *Tuta absoluta* (Lepidoptera: Gelechiidae) eggs in open-field tomato cultivation through geostatistical analysis. *Pest management science*, 74(1):30–36, 2018.

[40] MaxMind. Cities database. https://www.maxmind.com/en/free-world-cities-database.

[41] Lauren Ancel Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society*, 44(1):63–86, 2007.

[42] Ministry of Agriculture, Malaysia. Vegetable production by state. http://www.moa.gov.my/, Accessed April 2018.

[43] Ministry of Information and Communication Technology, Thailand. Agricultural census. http://web.nso.go.th/en/census/agricult/cen_agri03.htm, 2013.

[44] Zaw Aye Moe. Overview of food accessibility situation in Myanmar Zaw. *Thematic papers on Myanmar census of agriculture*, (May), 2013.

[45] Paule Moustier. Final summary report of SUSPER. *Sustainable Development of Peri-Urban Agriculture in South-East Asia*, (April):152, 2007.

[46] NASA Earth Observations (NEO). Normalized Difference Vegetation Index (NDVI). https://neo.sci.gsfc.nasa.gov.

[47] NASA Langley Research Center Atmospheric Science Data Center Surface meteorological and Solar Energy (SSE) web portal. Relative humidity. https://eosweb.larc.nasa.gov/sse/.

[48] Mark EJ Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[49] Andrei Novikov. annoviko/pyclustering: pyclustering 0.8.2 release, November 2018.

[50] Philippines Statistics Authority. Major vegetables and rootcrops quarterly bulletin. https://psa.gov.ph/content/major-vegetables-and-rootcrops-quarterly-bulletin, 2017.

[51] San-San-Yi, S. A. Jatoi, T. Fujimura, S. Yamanaka, J. Watanabe, and K. N. Watanabe. Potential loss of unique genetic diversity in tomato landraces by genetic colonization of modern cultivars at a non-center of origin. *Plant Breeding*, 127(2):189–196, 2008.

[52] E. Sankarganesh, D.M. Firake, B. Sharma, V.K. Verma, and G.T. Behere. Invasion of the South American Tomato Pinworm, *Tuta absoluta*, in northeastern India: a new challenge and biosecurity concerns. *Entomologia Generalis*, 36(4):335–345, 2017.

[53] Chhean Sokhen, Diep Kanika, and Paule Moustier. Vegetable market flows and chains in phnom penh. 2004.

[54] MA Stanaway, Robert Reeves, and KL Mengersen. Hierarchical bayesian modelling of plant pest invasions with human-mediated dispersal. *Ecological modelling*, 222(19):3531–3540, 2011.

[55] Chris D Stansbury, Simon J McKirdy, Art J Diggle, and Ian T Riley. Modeling the risk of entry, establishment, spread, containment, and economic impact of tilletia indica, the cause of karnal bunt of wheat, using an australian context. *Phytopathology*, 92(3):321–331, 2002.

[56] Daily Sun. Bumper winter vegetable production likely in rangpur region. http://www.daily-sun.com/post/282672/Bumper-winter-vegetable-production-likely-in-Rangpur-region.

[57] Serigne Sylla, Thierry Brévault, Lucie S Monticelli, Karamoko Diarra, and Nicolas Desneux. Geographic variation of host preference by the invasive tomato leaf miner tuta absoluta: implications for host range expansion. *Journal of Pest Science*, pages 1–10, 2019.

[58] Serigne Sylla, Oumar Seydi, Karamoko Diarra, and Thierry Brévault. Seasonal decline of the tomato leafminer, *Tuta absoluta*, in the shifting landscape of a vegetable-growing area. *Entomologia Experimentalis et Applicata*, 166(8):638–647, 2018.

[59] Serigne Elhadji Sylla, Thierry Brévault, Lucie Monticelli, Karamoko Diarra, and Nicolas Desneux. Geographic variation of host preference by the invasive tomato leafminer *Tuta absoluta*: implications for host range expansion. *Journal of Pest Science, Accepted*, 2018.

[60] Henri EZ Tonnang, Samira F Mohamed, Fathiya Khamis, and Sunday Ekesi. Identification and risk assessment for worldwide invasion and spread of *Tuta absoluta* with a focus on Sub-Saharan Africa: implications for phytosanitary measures and management. *PloS one*, 10(8):e0135283, 2015.

[61] Chuthaporn Vanit-Anunchai. Possibilities and constraints of marketing environmentally friendly produced vegetables in Thailand. *Faculty of Economics and Management*, 2006.

[62] S Venkatramanan, S Wu, B Shi, A Marathe, M Marathe, S Eubank, LP Sah, AP Giri, LA Colavito, KS Nitin, et al. Modeling commodity flow in the context of invasive species spread: Study of *Tuta absoluta* in Nepal. *Crop Protection*, 2019.

[63] Tran Duc Vien. Overview on tomato production and tomato varieties in Vietnam. (March):1–10, 2003.

[64] Katinka Weinberger and Christian A Genova II. *Vegetable production in Bangladesh: commercialization and rural livelihoods*. AVRDC-WorldVegetableCenter, 2005.

[65] Siebe Van Wijk and A. P. Everaarts. The market for vegetables in North Vietnam. Technical report, Wageningen University, 2007.

[66] L. You, U. Wood-Sichra, S. Fritz, Z. Guo, L. See, and J. Koo. Spatial Production Allocation Model (SPAM) 2005 v3.2. http://mapspam.info, 2017.