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 [15]. Climate change and the detrimental impact of intensive agriculture on natural resources further aggravate this problem [9]. 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 [21]. 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 [6, 10]. 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 phenomena as diverse as infectious disease and invasive species spread, online social networks, and cascading failures in infrastructure networks [2]. Douma et al. [8] 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 [4, 22], multi-scale spread [29] monitoring [18] and mitigation [24]. Unlike pest risk maps generated by species distribution models [23], the resulting dynamics of such a validated model yields a causal description of the underlying complex system.

We present a multi-pathway propagation model to study the spread of invasive agricultural pests. 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 [3, 7]. 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 [16]. 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

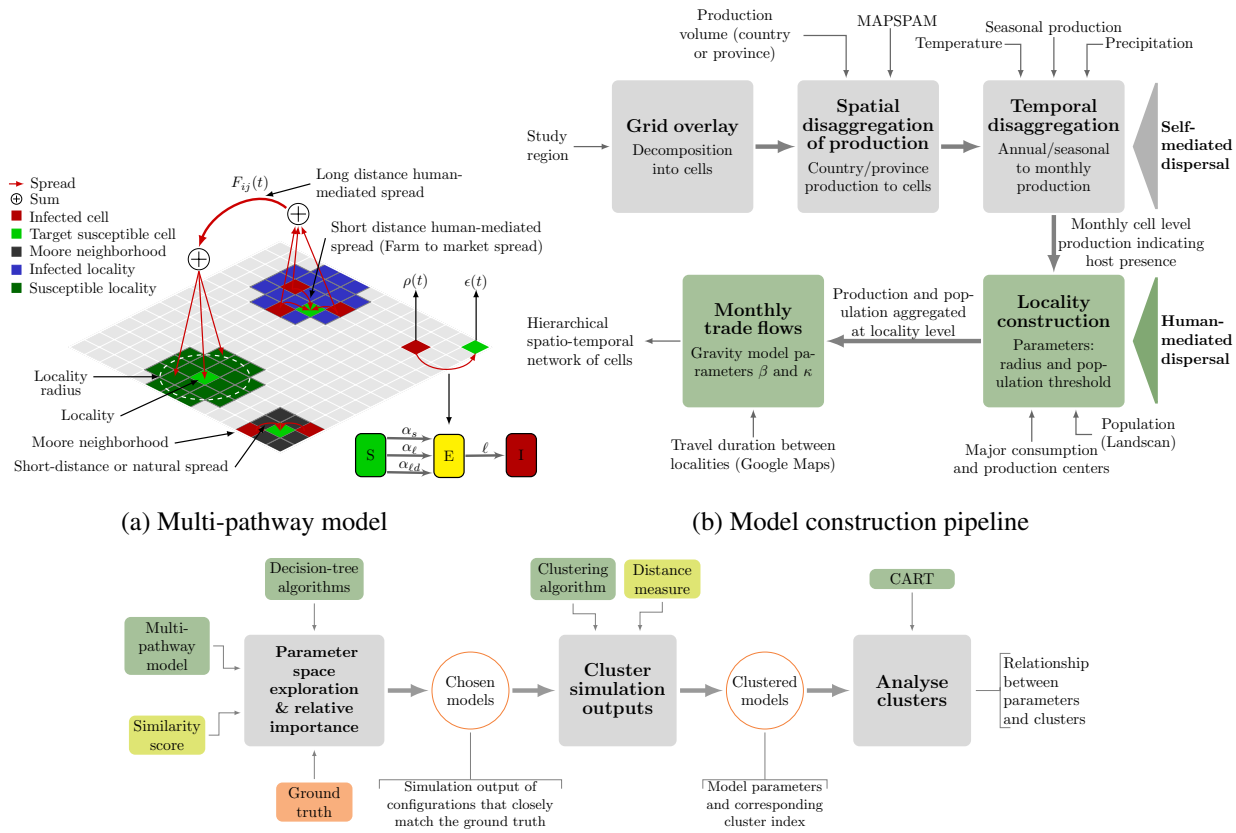
this region. With the pest having already spread to major tomato producing areas in Bangladesh, there is a high chance that it will be introduced to the remaining countries in the near future. Such invasions can have devastating effect on the economy and livelihood of farmers. Moreover, the invasion 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 [7, 13, 26]. A precursor to this work [28] 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 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. To construct it we identified, analysed, and fused disparate datasets corresponding to biology, climate, production, and agricultural commodity flow. 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 spread (b) what are the possible patterns of 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 [17] and interpretable artificial intelligence [12]. 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 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



(a) Multi-pathway model

(b) Model construction pipeline

(c) Outline of the process used for analysing the multi-pathway spread.

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

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, each corresponding to one month. The probability that a cell v transitions from state S to E is determined by (i) suitability to establish at that time step $\epsilon(v,t)$ and (ii) influence of 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 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 [25].

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 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 S3.6.

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 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 S3.1 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

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 [13, 18].
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).
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 S3.1 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 [14].
$\alpha_s, \alpha_\ell, \alpha_\ell$	Pathway scaling factors	In the interval $[0, 500]$.

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 [30]. 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 S3.2 of the supplement.

To model locality-to-locality trade, we applied the approach of Venkatramanan et al. [28] 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 S3.3. 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 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 [4]. 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, then the similarity $\mathcal{S}$ is given by

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

For parameter space exploration, we were motivated by a recent approach of using machine learning surrogates [17]. In our iterative "go with the winners" process [1] the subspace under consideration is sampled uniformly. 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. 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.

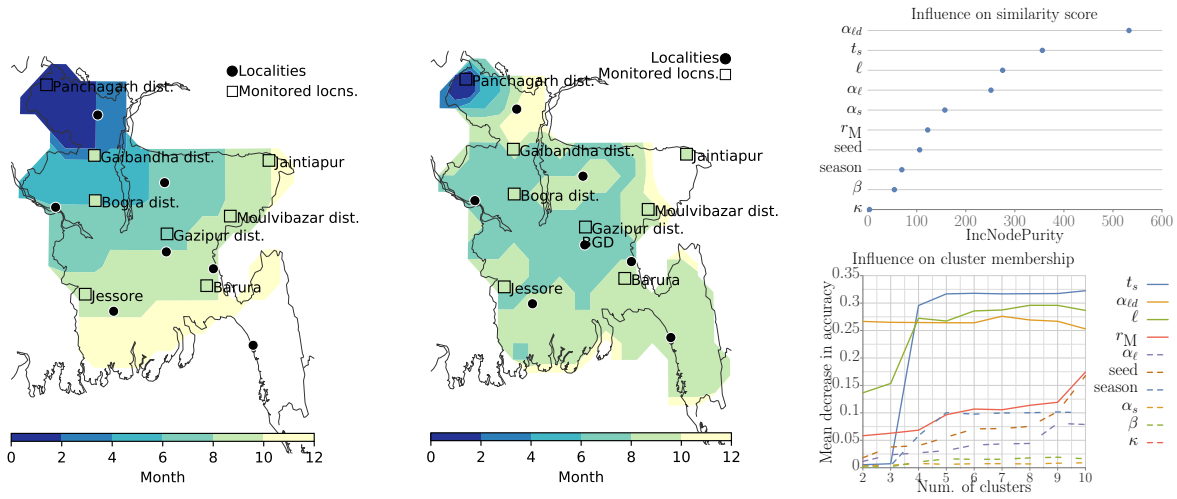
Analysis of spread pattern. The objective here is to analyse the variability in the simulation outcomes within the set of best fit configurations. We leverage well-known machine learning techniques in a novel way to address this question. The methodology is outlined in Figure 1c. First, we cluster the 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.

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 [13] 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).



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

Figure 2: **Possible spread patterns 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.

The relative importance of model parameters was assessed using random forest algorithm with regard 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 S11).

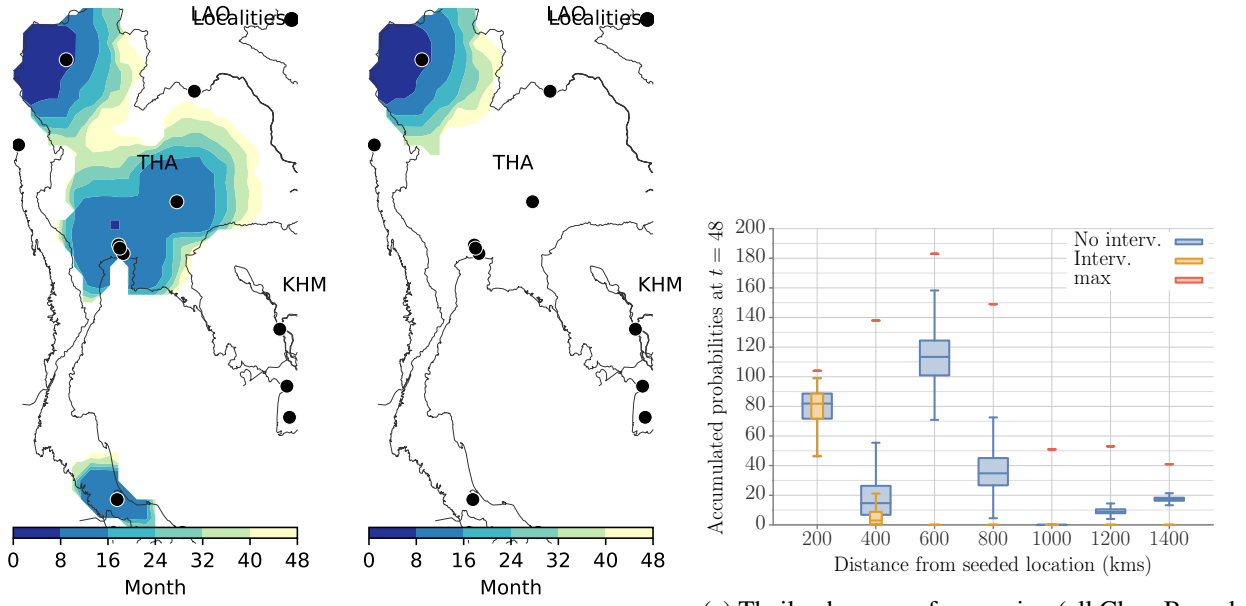
We also analysed the international tomato trade network (Section S6.1 of supplement) to assess the risk due to imports from *T. absoluta* infested countries outside this region. Malaysia and Singapore are important hubs with tomato imports from *T. absoluta* infested regions. There is a possibility that *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 S11a in the supplement). 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 S11b), 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 S12. 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 S12b). 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 3b 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.

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 [27]. 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 S13 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 3b, stifling these flows localises the spread that resembles those of Class A models, but with much less intensity.



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

Figure 3: **Rate and pattern of spread with and without intervention.** Representative spread dynamics of Class B models ($r_{\mathrm{M}} = 1, \ell = 3$) for two countries. More plots are shown in Figure S13. In each case, a cell close to a high production region was seeded. The first column corresponds to spread 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.The second column corresponds to spread after cutting off flows from chosen localities. The third column 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.

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 *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 [28]. 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. Some countries have already taken measures in this regard. In the United States, 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 [27]. Identifying the optimal set of nodes to mitigate an epidemic on a network is

well-studied (e.g., [22]). As the world moves towards concentrated and specialised agricultural production, focusing on this aspect becomes increasingly important.

Emulators –based on Gaussian processes for example [11] – and machine learning surrogates [17] are emerging as solutions to overcome computational challenges, parameterization, and sensitivity analysis of complex agent-based models. 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 [5]. 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. This is similar to approaches in the emerging field of interpretable machine learning where decision-tree algorithms are used as surrogates for various machine learning algorithms [19].

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. [4]), 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 [9].

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] C. Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[20] 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.

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

[22] 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.

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

[24] 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.

[25] 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.

[26] 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.

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

[28] 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.

[29] 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.

[30] 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.