Title: Many unreported crop pests and pathogens are probably already present

Running title: Crop pest and pathogen distributions

Daniel P. Bebber[1*], Elsa Field[2], Gui Heng[3], Peter Mortimer[3], Timothy Holmes[4], Sarah J. Gurr[1]

[1] Department of Biosciences, University of Exeter, Stocker Road, Exeter, EX4 4QD, UK
[2] Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, UK
[3] Kunming Institute of Botany, Chinese Academy of Sciences, 132 Lanhei Road, Kunming 650201, People's Republic of China
[4] CABI, Nosworthy Way, Wallingford, OX10 8DE, UK
[*] Corresponding author, email d.bebber@exeter.ac.uk, tel. +44 1392 725851

Summary

Biotic invasions threaten global biodiversity and ecosystem function, and present challenges to agriculture where invasive pest species require major economic investment in control and can cause significant production losses. Pest Risk Analysis (PRA) is key to prioritizing agricultural biosecurity efforts, but is hampered by incomplete knowledge of current crop pest and pathogen distributions. Here we develop predictive models of current pest distributions and test these models using new observations at sub-national resolution. We apply generalized linear models (GLM) to estimate presence probabilities for 1901 crop pests in the CABI pest distribution database. We test model predictions for 100 unobserved pests in the People's Republic of China (PRC), against observations of these pests abstracted from the Chinese literature which has hitherto been omitted from databases on global pest distributions. Finally, we predict occurrences of all unobserved pests globally. Presence probability increases with host presence, presence in neighbouring regions, and global prevalence, and decreases with mean distance from coast, per capita GDP, and host number. The models are good predictors of pest presence in Provinces of the PRC, with AUC values of 0.76 – 0.80. Large numbers of currently unobserved, but probably present pests, are predicted in China, India, southern Brazil and some countries of the former USSR. GLMs can predict presences of pseudo-absent pests at sub-national resolution. Controlling for countries' scientific capacity improves model fit. The Chinese scientific literature has been largely inaccessible to Western academia but contains important information that can support PRA. Prior studies have often assumed that unreported pests in a global distribution database is a true absence. Our analysis provides a method for quantifying pseudo-absences to enable improved PRA and species distribution modelling.

Keywords

biogeography, crop pathogens, crop pests, Generalized Linear Model, observational bias, Pest Risk Analysis, pseudo-absence, species distribution model.

Introduction

The spread of invasive species is homogenizing the biosphere, with wide-ranging implications for natural ecosystems (Baiser *et al.*, 2012; Santini *et al.*, 2013) and agriculture (Fisher *et al.*, 2012; Bebber *et al.*, 2014a; Bebber, 2015). The number of first observations of crop pests and pathogens (CPPs) has accelerated in recent years, driven primarily by global trade (Ding *et al.*, 2008; Bacon *et al.*, 2013), but also potentially by climate change and our improving ability to monitor and identify threats (Bebber *et al.*, 2014a; Bebber, 2015). Emerging CPPs can be extremely damaging to agricultural production and the economy, through both pre-harvest and post-harvest losses (Bebber & Gurr, 2015; Paini *et al.*, 2016; Savary *et al.*, 2017). Recently, for example, sub-Saharan Africa has suffered from the virulent Ug99 strain of the wheat stem rust fungus (*Puccinia graminis tritici*) (Patpour *et al.*, 2015), the newly-evolved Maize Lethal Necrosis viral syndrome (Wangai *et al.*, 2012), and the appearance of Tropical Race 4 of *Fusarium oxysporum f. sp. cubense* attacking Cavendish bananas (Ordonez *et al.*, 2015). Central America, Europe, East Africa and Australia have been identified as hotspots of new CPP invasions, with maize, bananas, citrus and potato as the crops most likely to be affected (Bebber, 2015). Outbreaks of resident pests due to weather, virulence evolution, or management factors, add to the burden on farmers. For example, a major outbreak of coffee leaf rust (*Hemileia vastatrix*) in Latin America, likely to have been triggered by a failure in disease management, is reported to have caused large-scale unemployment and social upheaval in recent years (Avelino *et al.*, 2015).

Despite the expanding ranges of many CPPs, complete occupation of their potential ranges has not yet occurred (Bebber *et al.*, 2014a) and so there remains a strong impetus for biosecurity measures at international borders (Fears *et al.*, 2014; Flood & Day, 2016; MacLeod *et al.*, 2016). Control of spread within countries is extremely difficult because of largely unhindered transport of plants and soils (Ward, 2016), and biosecurity measures focus largely on quarantine and inspections at borders (MacLeod *et al.*, 2016). A key component of international phytosanitary action is Pest Risk Analysis (PRA), a suite of methods that allow countries to prioritize protective measures against those pests most likely to arrive and cause serious economic damage (Robinet *et al.*, 2012; Baker *et al.*, 2014). PRA involves assessment of the likelihood of CPP arrival, the likelihood of establishment, the potential economic impact if uncontrolled, and the likelihood of successful control or eradication (Baker *et al.*, 2014). To date, PRA has largely been based upon expert opinion regarding the likelihood of arrival and potential impact of individual pests. For example, the UK's recently-established Plant Health Risk Register (PHRR) (Baker *et al.*, 2014) employs simple climate-matching (based on known pest distributions) and host availability to assign qualitative risks of invasion and impact, but not quantitative predictive models. Examples of registered CPPs include the Oleander aphid *Aphis nerii* which has been assigned very low likelihoods of arrival and establishment, and would cause negligible damage if it did, whereas the zebra chip phytoplasma *Candidatus Liberibacter solanacearum* is thought moderately likely to arrive but would have a very serious impact if it did (DEFRA, 2018)

The rarity of quantitative PRA modelling in international phytosanitary legislation and practice contrasts with the long and vibrant history of research in predictive species distribution modelling (SDM) for CPPs (Elith & Leathwick, 2009; Sutherst, 2014). The geographic distributions of species are non-random, determined by their biotic environment (e.g. hosts or prey), the abiotic environment (e.g. climate, edaphic factors), and migration (dispersal to suitable habitat) (Soberón & Peterson, 2005; Soberón, 2007; Soberón & Nakamura, 2009). Thus, pest invasion risk is, in theory, quantifiable. Numerous modelling approaches are now available to predict the likely distributions and impacts of CPPs (Elith & Leathwick, 2009; Venette *et al.*, 2010; Robinet *et al.*, 2012), ranging from process-based, or mechanistic models, to statistical, or correlative approaches (Dormann *et al.*, 2012). Regional and global databases on known pest distributions are commonly used to parameterize these models, either providing direct estimates of pests' ecological niches (Venette *et al.*, 2010; Kriticos, 2012), or indirectly via shared geographic ranges (Paini *et al.*, 2010, 2016; Eschen *et al.*, 2014).

One seldom-acknowledged issue with pest distribution data in global databases is geographic bias in the likelihood that a pest will be detected, correctly identified, reported and recorded (Pyšek *et al.*, 2008). Analysis of the CABI pest distribution database (CABI, 2017), one of the most commonly used global pest distribution databases, suggests that hundreds of pests already present in many developing countries have not been reported (Bebber *et al.*, 2014b). The total number of observed CPPs in an administrative area (country, or administrative division for larger countries) can be largely explained by scientific capacity and agricultural production. Under a scenario of globally high scientific and technical capacity (i.e. where all countries have US-level per capita GDP and research expenditure), analysis predicts that many countries across the developing world would report hundreds more pests. This suggests that a large fraction of the *current* agricultural pest burden is unreported and unknown, and that even the best global databases suffer from severe observational bias, with potentially serious consequences for both plant biosecurity activities and for research based upon these databases. This observational bias may have implications for SDM methods that infer environmental tolerances from observed distributions. Scientific capacity, economic development, and the ability to detect, identify and report pests, are strongly correlated with latitude, as is climate (Bebber *et al.*, 2014b). Under-reporting of pests at low latitudes will therefore bias estimation of climate tolerances, as occurrence is underreported in warmer regions. Reducing this observational bias by strengthening pest identification efforts in the developing world is therefore critical in improving scientific understanding of pest distributions, and in PRA.

The People's Republic of China (henceforth referred to as China) has been predicted to harbour the largest number of CPPs (Bebber *et al.*, 2014b). China produces the largest quantity of crops, and has the greatest diversity of production – both factors are strong determinants of recorded pest numbers (Bebber *et al.*, 2014b). Yet, the actual recorded number of pests in China is much smaller than expected (Bebber *et al.*, 2014b). For many countries, under-reporting of agricultural pests is likely to be purely a function of the lack of institutional capacity to detect, identify, and report incidences in the scientific and 'grey' literature used by CABI to populate the distribution database. For China, there is potentially an interesting alternative. The Chinese literature was, until the reforms of 1978, largely inaccessible to Western academia. Even post-reform and the opening of China instigated by Deng Xiaoping, Chinese-language publications are not commonly accessed by English-speaking researchers. A famous of translation of the Chinese literature is the reporting of the anti-malarial compound artemisinin (Klayman, 1985). The Chinese research literature, having developed largely independently of the Western literature, therefore provides a largely independent data source for testing models of pest distributions. Here, we test statistical models of pest presence using a global database of known pest occurrence and confront the predictions of pest presence in China's Provinces with observations from the Chinese literature. In addition, we develop models where observational bias

Materials and Methods

We obtained pest distribution data from CABI distribution database in January 2014 with permission. Briefly, the database comprised 91,030 records of the observed distributions of 1901 agricultural pests by administrative division of each country, e.g. US States, Chinese Provinces. These pests comprise 419 species and pathovars of Fungi, 219 Coleoptera, 252 Lepidoptera, 236 Hemiptera, 230 viruses, 126 Bacteria, 110 Diptera, 104 Nematoda, 59 Oomycota, and smaller numbers of Acari, Gastropoda, and various other insect and microbial taxa.

We developed a statistical model for the presence of pests in global administrative regions (countries, and sub-national divisions for Brazil, Canada, China, India, Russia and the USA). We constructed Generalized Linear Models, using the *glm* function (*MASS* package) for R v.3.4.0 (R Development Core Team, 2017), for the presence or (pseudo-) absence of each pest in each administrative region. Predictors were log-transformed per capita GDP for the country as a whole in 2016 (World Bank data, http://data.worldbank.org/), log-transformed total number of known hosts for the pest (CABI, obtained with permission), log-transformed area of neighbouring regions which have reported the pest as present (set to zero if no neighbours have reported the pest), and log-transformed total fraction of regions globally that have reported the pest. Host crop spatial distributions were obtained from the EarthStat database (http://www.earthstat.org/; Monfreda *et al.*, 2008), and used to estimate mean distance of host areas to coastline. Briefly, the rationale for these predictors was that GDP is a proxy for historical trade (Pyšek *et al.*, 2010) and observational capacity (Bebber *et al.*, 2014b), host number indicates the degree of biotic generalism of the CPP, neighbouring-region presence indicates the potential for spread across a border, fraction of regions reporting presence indicates global ubiquity and environmental generalism, and distance to coast indicates proximity to international shipping ports (Chapman *et al.*, 2017).

We developed two pest distribution models. The 'unweighted' model included geographical and bioclimatic predictors and treated all unobserved pests as absent from a region. The 'weighted' model treated unobserved pests as potentially pseudo-absent, using a function of the scientific and technical capacity of each country (Bebber *et al.*, 2014b). Presences were taken as being correct and unambiguous, and given a weighting of unity. Absences were weighted by the logarithm of the agricultural and biological sciences publication output of each country from 1996 – 2016 (Scimago Lab, 2017), normalized to the logarithm of the output of the USA (the world's most scientifically productive country), such that the absence weight $w_0 = \log(s)/\log(s_{USA})$. Thus, pests unreported from scientifically advanced nations were assumed not to be present (or, present at undetectable population density), while pests unreported from developing nations were less informative of absence. China, with the second largest research output, had $w_0 = 0.93$, suggesting that non-reporting of a pest should be relatively strong evidence of its physical absence. However, we hypothesized that non-reporting in the CABI databases could be due to lack of translation from the Chinese literature, therefore we set w_0 to zero for China, effectively omitting these pseudo-absences from the analysis. The models were compared with a null model assuming constant presence probability using Likelihood Ratio Tests.

To validate the models we predicted the probability of presence for a random sample of 100 as-yet unobserved pests in all Chinese Provinces, but excluding Taiwan. The Chinese literature was searched for observations of these unobserved pests in China. We used the text mining methodology designed by CABI for their Plantwise Knowledge Bank. The following rules were followed to locate pest records in the Chinese literature:

- Include only papers that are primarily about distribution data, not those where distribution is mentioned, but something else is the primary focus. If this is unclear do not process the paper.
- Mine only the primary literature (including Masters and Doctoral theses), not meta-analyses, reviews, or non-peer reviewed ("grey") literature.
- Pest and host names must be preferred scientific names, following the CAB Thesaurus (www.cabi.org/cabthesaurus/) and the Plant List (http://www.theplantlist.org/).
- Record country and location information given in the paper, including latitude/longitude. CABI uses five levels for location, from the largest scale (i.e., provincial) to the smallest (i.e., village/town).
- Record date of observation/collection (entering each year separately) and date of publication. Can be left blank if not given, or use the date of receipt in the diagnostic laboratory as a surrogate for date of collection.
- Record pest status – present/not found. Only record absence if pest absence is specifically stated in the paper.
- Record pest status using only the status terms defined by CABI, and only if used in the paper e.g. "widespread", "restricted" "soil only" "greenhouse only" (see CABI guidelines for complete list).
- Record if the paper was a first record of that pest or not and details of this (e.g. "first record in <country/location>", "first record on <host species name>")

- Only enter data where the pest/pathogen has been clearly identified, not just symptoms seen.
- Record only natural infections, not artificial inoculants.

Combinations of pests and locations were submitted to several search engines. The priority of search engines was: Baidu (www.baidu.com), China National Knowledge Infrastructure (CNKI, http://www.cnki.net), Chongqing VIP Information Company (CQVIP, http://lib.cqvip.com/), and Wangfang Data (http://www.wanfangdata.com.cn). Baidu is the most popular Chinese internet search engine. CNKI is led by Tsinghua University, and supported by ministries of the Chinese Government. CQVIP, formerly known as Database Research Center under the Chongqing Branch of the Institute of Scientific & Technical Information of China (CB-ISTIC), was China's first Chinese journal database research institution. Wanfang Data is an affiliate of the Chinese Ministry of Science & Technology, and provides access to a wide range of database resources.

Publication titles were searched first, then full text. The first 50 search results were scanned before dismissing a search term. The first search term combination was pest name and location (Province). If this yielded no result, then pest name and various distribution terms were tried. These distribution terms were: "catalogues" OR "checklists" OR "distribution" OR "inventories" OR "new records" OR "surveys" OR "geographical distribution" OR "new geographic records" OR "new host records". Searches included local names in Chinese where these were known or could be identified from the literature, preferred scientific names, and non-preferred scientific names from CAB Thesaurus (https://www.cabi.org/cabthesaurus/).

Searches continued until one piece of literature was found for that pest in that region, that fitted all of the requirements for CABI text mining. If a pest was not found from any of these searches, it was assumed to be absent from the literature. We then compared our probability of presence predictions with the observed presence-absence data for our Chinese sample data using logistic regressions (*glm* function for R) and ROC curves (*pROC* library for R). The logistic regression coefficients c and m determined the probability of pest presence $P(\text{present}) = e^{(c+mx)} / (1+ e^{(c+mx)})$.

Results

Globally, the probability of pest presence within a geographical area increased significantly with presence in neighbouring regions, the area of host crops, and the global prevalence of the pest, in both models (Tables 1). Presence probability declined with mean distance from the coast, per capita GDP, and known host species number per pest. Presence probability increased with GDP in the unweighted model but declined with GDP in the weighted model. The weighted model explained a larger fraction of the deviance than the unweighted model (Table 1), while both models had very similar ROC curves with AUC around 84 per cent (Figure 1). Predicted probabilities were always higher for the weighted model, because absences were down-weighted (i.e. fewer true zeros). The models indicated greater overall presence probabilities for viruses and Hemiptera, and lower probabilities for nematodes, compared with other CPP groupings. Presence probabilities for both models, for all CPPs and all regions, are provided in Supplementary Online Material.

Table 1. GLMs for global pest presence. The unweighted model treated unobserved pests as true absences. The weighted model weighted pseudo-absences as a function of country scientific capacity. The unweighted model had AIC = 268448, Nagelkerke R^2 = 0.26, McFadden R^2 = 0.34. The weighted model had AIC = 220886, Nagelkerke R^2 = 0.49, McFadden R^2 = 0.39.

	Unweighted model				Weighted model			
	Mean	SE	Z	Pr(>|Z|)	Mean	SE	Z	Pr(>|Z|)
Acari (Intercept)	-1.56	0.056	-27.7	0.000	1.309	0.061	21.5	0.000
+ Coleoptera	0.061	0.033	1.8	0.068	0.061	0.036	1.7	0.089
+ Diptera	0.077	0.037	2.1	0.038	0.070	0.040	1.7	0.082
+ Hemiptera	0.147	0.031	4.7	0.000	0.118	0.034	3.5	0.001
+ Lepidoptera	0.096	0.032	3.0	0.003	0.058	0.035	1.7	0.098
+ Bacteria	-0.045	0.035	-1.3	0.187	-0.036	0.038	-0.9	0.343
+ Fungi	0.074	0.031	2.4	0.016	0.074	0.034	2.2	0.028
+ Nematoda	-0.155	0.035	-4.5	0.000	-0.163	0.038	-4.3	0.000
+ Oomycota	0.056	0.038	1.5	0.137	0.076	0.041	1.8	0.064
+ Virus	0.128	0.033	3.8	0.000	0.143	0.036	3.9	0.000
log(coastdist + 1)	-0.097	0.004	-25.5	0.000	-0.190	0.004	-45.3	0.000
log(GDP + 1)	0.135	0.004	34.3	0.000	-0.065	0.004	-15.4	0.000
log(hosts + 1)	-0.124	0.005	-28.3	0.000	-0.114	0.005	-23.9	0.000
log(hostarea + 1)	0.064	0.001	50.8	0.000	0.056	0.001	42.1	0.000
log(nbarea + 1)	0.125	0.001	147.4	0.000	0.130	0.001	141.1	0.000
log(prev)	0.882	0.008	115.5	0.000	0.904	0.008	111.8	0.000

We defined a 'probably present pest' (PPP) as one unreported from a region, but with high (> 0.75) predicted presence probability (using the weighted model). Overall, only 1.0 per cent of all unreported CPP-region combinations were highly likely. The number of PPPs per CPP category was greatest for Fungi (2.2 per cent) and Hemiptera (1.6 per cent), with fewer than 1.0 per cent of unreported CPP-region combinations being highly likely (> 0.75 predicted presence probability) for other taxonomic groups. Overall, 80 per cent of unreported CPP-region combinations were predicted to be highly unlikely (presence probability < 0.25). China, India and Eastern Europe had the largest numbers of predicted PPPs, along with other parts of East Asia and Southern Brazil (Figure 2). The top ten PPPs by number of global regions were *Cochliobolus heterostrophus* (Ascomycota: Pleosporales, a pathogen of maize), *Rhopalosiphum maidis* (Arthropoda: Hemiptera, pest of maize and other crops), *Cochliobolus sativus* (cereal pathogen), *Aphis spiraecola*, *Nezara viridula* (Arthropoda: Hemiptera, legume pest), *Setosphaeria turcica* (Ascomycota: Pleosporales, maize pathogen), *Schizaphis graminum* (Arthropoda: Hemiptera, pest of Poaceae cereals), *Delia platura* (Arthropoda: Diptera, pest of legumes), *Rhopalosiphum padi* (Arthropoda: Hemiptera, cereal pest), and *Gibberella fujikuroi* (Ascomycota: Hypocreales, rice pathogen).

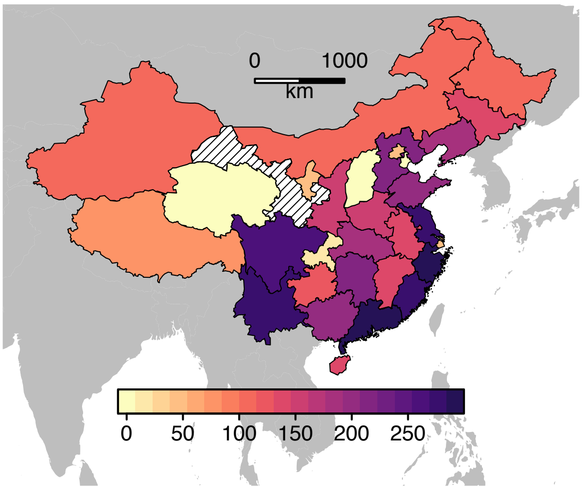

Figure 1. Total number of pests recorded in the CABI pest distribution database by China Province (excluding Taiwan). Hatched region is Gansu (804 recorded pests), see text for details.

Total numbers of recorded pests in China's Provinces and municipalities increased from northern and central regions to southern and coast regions (Figure 1), except for the central province of Gansu which had 804 reported pests. There is no obvious reason why numbers would be so large in Gansu. For example, agricultural production is moderate, and there are no particular academic centres which could account for observational bias. Hence, the Gansu values appear to be an artefact of the CABI database. The smallest numbers of recorded pests were in the mountainous province of Qinghai (0), in the central provinces of Shanxi (0) and Ningxia (46), and the municipalities of Chongqing (24), Tianjin (4) and Shanghai (48). Total numbers were largest in the coastal provinces of Guangdong (292), Zeijiang (286), Jiangsu (277), Fujian (266), and also in the southern provinces of Yunnan (275) and Sichuan (256).

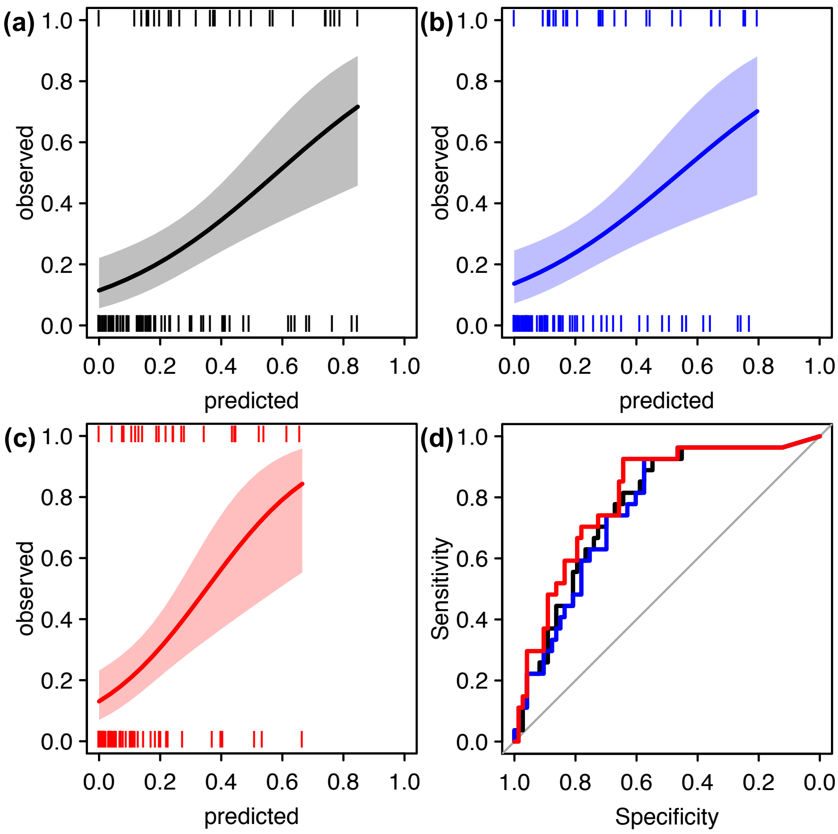

Figure 2. Model prediction tests. Observed presence/absence of 100 pest x Province combinations vs. predicted presence probability from a) unweighted model, b) weighted model. Curves show mean 95% CI for logistic regression fits. c) ROC curves for unweighted (blue), and weighted (red) models. AUC was 0.76 (0.66 – 0.86, 95% CI) for the unweighted model and 0.75 (0.65 – 0.85) for the weighted model.

We validated our models using published CPP observations from the Chinese literature. Both models were significant predictors of pest presence/absence for 100 randomly-sampled CPP-Province combinations, of which 25 were found to be present (Figure 2, Table S1). For the unweighted model, the coefficients of the logistic function were c = -1.73 ± 0.34 and m = 3.52 ± 1.25 (likelihood ratio test vs null model, p = 0.004). For the weighted model, the coefficients were -1.90 ± 0.38 and 3.10 ± 1.03 (p = 0.002). The predictive power of the models was also tested using ROC curves, demonstrating significant discriminant ability with AUC of 0.76 (95 per cent Confidence Interval 0.66 – 0.86) for the unweighted model, and AUC 0.75 (0.67 – 0.85) for the weighted model (Figure 2d). Our analysis revealed gaps in the CABI database, which is commonly used for analyses of global pest distributions. Taking one important potato pest, *Phytophthora infestans* (Oomycota), as an example, very high presence probabilities (> 0.80) were predicted for Guangdong, Hainan, Shandong, and Zheijiang, and Jiangxi, while a search of the Chinese literature found references to *P. infestans* in all of these but Guangdong. However, this pathogen has been reported present throughout the potato-growing regions of China, including Guangdong (Guo *et al.*, 2010).

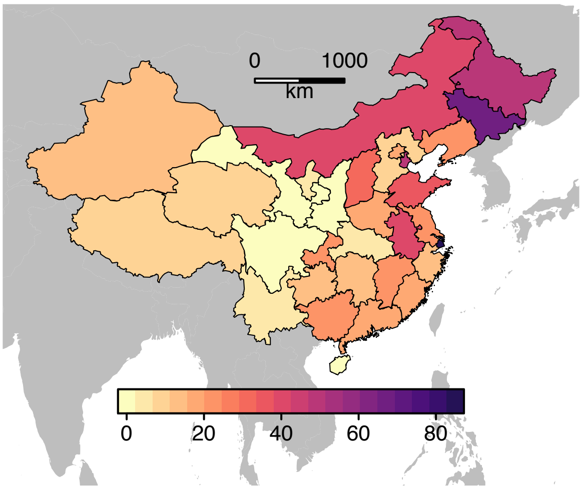

Figure 3. Total number of probably present pests (PPP) in China Provinces (excluding Taiwan). We defined a PPP as one unreported from a region, but with high (> 0.75) predicted presence probability using the weighted model.

The total number of PPPs (including all those in the CABI database), was greatest in the North Western provinces of Jilin (64), Heilongjiang (51), and Inner Mongolia (42), as well as the ports of Shanghai (85) and Tianjin (51), while Central provinces had the lowest numbers (Figure 3). Fungi (282) and Hemiptera (176) were the most commonly reported PPPs for China. The top ten most-common PPPs in China were (in decreasing order) *Gibberella fujikuroi* (Ascomycota: Hypocreales, rice pathogen), *Aphis spiraecola* (Arthropoda: Hemiptera, generalist), *Delia platura* (Arthropoda: Diptera, pest of legumes), *Rhopalosiphum maidis* (Arthropoda: Hemiptera, pest of maize and other crops), *Athelia rolfsii* (Basidiomycota: Atheliales, generalist facultative pathogen), *Rhopalosiphum padi* (Arthropoda: Hemiptera, cereal pest), *Agrotis ipsilon* (Arthropoda: Lepidoptera, generalist pest), *Cochliobolus lunatus* (Ascomycota: Pleosporales, pathogen of rice and sorghum), *Sitobion avenae* (Arthropoda: Hemiptera, cereal pest), and *Lasiodiplodia theobromae* (Ascomycota: Botryosphaeriales, generalist pathogen).

Extending the analysis globally, the regions with the largest numbers of PPPs were China, India, and Eastern Europe, along with other parts of East Asia and Southern Brazil (Figure 4). The top ten PPPs by number of global regions were *Cochliobolus heterostrophus* (maize pathogen), *Rhopalosiphum maidis*, *Cochliobolus sativus* (cereal pathogen), *Aphis spiraecola*, *Nezara viridula* (Arthropoda: Hemiptera, legume pest), *Setosphaeria turcica* (Ascomycota: Pleosporales, maize pathogen), *Schizaphis graminum* (Arthropoda: Hemiptera, pest of Poaceae cereals), *Delia platura*, *Rhopalosiphum padi*, and *Gibberella fujikuroi*. Hence, many of the global PPPs were also likely to be present, but unreported, in China.

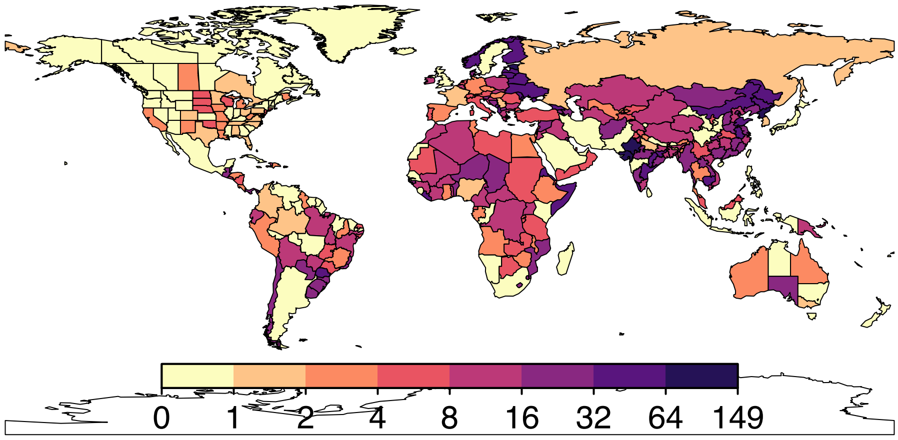

Figure 4. Total number of probably present pests (PPP) in all countries and sub-national regions.

Discussion

The Chinese literature provided strong and significant support for the predictions of pest distribution models based upon host distribution, pest prevalence, and other socioeconomic factors. China's growing economy is expected to lead to large influxes of invasive species, including CPPs, in coming years (Ding *et al.*, 2008). Analysis of temporal trends in CABI CPP observations show a relatively smooth increase in pests from 1950-2000, but the pattern for China is more complex, with a slow increase from 1950 until the late 1970s, a step increase, and then a more rapid growth in pest numbers from 1980 onwards (Bebber *et al.*, 2014a). One potential determinant of this sudden acceleration is the strong support for science and technology given by Deng Xiaoping in 1978, which lead to an increase in funding and academic freedom following the anti-intellectualism of the Cultural Revolution. China now ranks second only to the USA in annual R&D expenditure (IMF, 2013) and scientific output (Scimago Lab, 2017).

We identified a number of CPPs that were very likely to be present, and the majority of these PPPs were globally distributed and had wide host ranges. Their distributions commonly spanned wide latitudinal ranges, indicating broad climatic tolerances. *C. heterostrophus*, or Southern Leaf Spot, is primarily known as a pathogen of maize but has a wide host range. It has a wide geographic distribution both latitudinally and across continents, resulting in a high likelihood of occurrence in other regions where hosts are present. For example, *C. heterostrophus* is currently recorded only in eastern regions of North America, where most maize is grown. The lack of reported observations in the western regions of North America may be due to the fact that maize, the major host, is uncommon, and hence the disease currently has little impact. *C. sativus*, causing root and foot rot, also has a very wide geographic distribution, but an even wider host range. It is reported from Texas, Oklahoma, Mississippi, Illinois and Tennessee, but not from neighbouring Arkansas or Missouri. Hence, the high presence probability in these States. A similar pattern is seen for the maize pathogen *S. turcica*. *R. maidis*, the green corn aphid, is another global pest species. It is reported across Europe and in Russia, but, like many other pests, not from the former Soviet states of Ukraine, Belarus, Lithuania, Latvia and Estonia. It is plausible that reporting from these nations was less likely when they were part of the USSR. This lack of observations in former USSR border states is also seen in *Gibberella fujikuroi*, cause of bakanae disease of rice.

Predictors like host distribution, presence in neighbouring territories and global prevalence were expected to have positive relations with presence probability. The negative relation with distance from coast is likely to be related to import via shipping ports (Huang *et al.*, 2012; Liebhold *et al.*, 2013), and supports the observation that islands report more pests than countries with land borders (Bebber *et al.*, 2014b). Inclusion of climatic factors as predictors did not markedly improve model performance, which was unsurprising as we only modelled overall climatic suitability, rather than for individual pests. Detailed modelling of individual pest climate responses (Bregaglio *et al.*, 2012; Kriticos *et al.*, 2013) for such a large number of pests was beyond the scope of this study. Implicitly, we can assume that the presence of the host crop indicates that the climate is suitable for the pest, though we acknowledge that this is not necessarily the case (Berzitis *et al.*, 2014). For the practical purposes of PRA, our models provide reliable probability estimates for the presence of unreported pests at subnational resolution, and we have provided a global list of the unreported pests whose presence is most likely (Table S2).

We addressed the issue of pseudo-absences in the CABI data by statistically weighting missing pest observations in proportion to the scientific output of the reporting nation, since scientific output had been confirmed as a strong determinant of total reported pest numbers (Bebber *et al.*, 2014b). Often, unreported pests are treated as true absences in pest risk analyses (Paini *et al.*, 2016). The positive relation of GDP with presence probability in the unweighted model, but negative relation in the weighted model, supports our hypothesis that wealthy countries are more likely to detect and report pests (Bebber *et al.*, 2014b). Once observational bias is controlled for using scientific capacity-based weighting, per capita GDP reduces presence likelihood, perhaps because wealthier countries are better able to prevent pests from arriving and establishing. Our weighted model had improved explanatory power compared with our unweighted model when considering the entire dataset, but there was no appreciable difference in model performance when tested against the Chinese literature. Nevertheless, the issue of observational biases related to country-level socioeconomic variation has been raised several times for various classes of organism (Jones *et al.*, 2008; Pyšek *et al.*, 2008; Westphal *et al.*, 2008; Boakes *et al.*, 2010; Bebber *et al.*, 2013, 2014b), and we therefore recommend the application of appropriate statistical controls when analysing datasets produced from reports of species presences (as opposed to distributional datasets derived from rigorous sampling protocols).

Our SDM was statistical, fitting response functions for various predictors to the probability of pest presence. Many SDM approaches exist, from highly mechanistic models based on pest biology and ecology (Bregaglio *et al.*, 2012; Skelsey *et al.*, 2016) to purely statistical models that utilize only patterns in known distributions (Paini *et al.*, 2010). The lack of quantitative model input into PRAs is partly due to the scarcity of empirical data available on pest biology and epidemiology required to parameterize mechanistic models, and so key biological parameters are often inferred from known distributions (Robinet *et al.*, 2012). This is particularly the case for newly emergent pathogens for which experimental investigations have not yet been conducted. Epidemiological parameters can be poorly constrained even for long-established pests. For example, coffee leaf rust fungus (*Hemileia vastatrix*) has affected coffee production for more than a century, but a recent infection model relied upon temperature response functions derived from the single available study published three decades previously (Bebber *et al.*, 2016). Initiatives such as the EU-funded PRATIQUE project (2008-11) have attempted to fill this knowledge gap and enable modelling by collating available ecophysiological data for insect pests (Baker, 2012). While the advantages and disadvantages of the many different pest distribution and impact models continue to be researched and debated (Venette *et al.*, 2010; Dormann *et al.*, 2012; Robinet *et al.*, 2012; Sutherst, 2014), it is clear that practical application of these methods in PRA remains limited.

SDM for CPPs has direct policy implications for PRA and plant biosecurity. PRA is guided by International Standards for Phytosanitary Measures (ISPM), which are part of the International Plant Protection Convention (IPPC) (MacLeod *et al.*, 2010). ISPMs tend to rely on expert judgement for PRA, rather than quantitative modelling to support decision making. ISPM No. 21 "Pest Risk Analysis for Regulated Non-Quarantine Pests", endorsed in 2004, mentions use of pest and host life-cycle and epidemiological information, but not quantitative modelling (FAO, 2004). Individual PRAs similarly employ a qualitative approach. For example, the Australian Government's PRA for *Drosophila suzukii* references only a single unpublished report on SDM for this species, conducted for North America. Probabilities of *D. suzukii* spread within Australia are qualitatively assessed by comparison with observations in other countries (Department of Agriculture, Fisheries and Forestry, 2013). The European and Mediterranean Plant Protection Organization (EPPO) PRAs occasionally include model results. For example, a climate matching for the fungal pathogen *Xanthomonas axonopodis* pv. *allii* was undertaken using the CLIMEX model, to identify areas at risk within the EPPO region (EPPO, 2008). Our results contribute to the quantification of risk within PRA by providing probabilistic estimates for the presence of hundreds of unreported CPPs around the world, thereby improving understanding of the threats to global agriculture.

Acknowledgements

We thank the British Society for Plant Pathology for an Undergraduate Vacation Bursary for Elsa Field. We thank Le Yei and Li Huili (Kunming Institute of Botany) for assistance in text mining the Chinese literature and Ho Wai Yim for translations. CABI is an international intergovernmental organisation and we gratefully acknowledge the core financial support from our member countries (and lead agencies) including the United Kingdom (Department for International Development), China (Chinese

Ministry of Agriculture), Australia (Australian Centre for International Agricultural Research), Canada (Agriculture and Agri-Food Canada), Netherlands (Directorate-General for International Cooperation), and Switzerland (Swiss Agency for Development and Cooperation). See https://www.cabi.org/about-cabi/who-we-work-with/key-donors/ for full details.

Author contribution

DB conducted the analyses and wrote the manuscript. EF and GH searched the Chinese literature. TH assisted with CABI data acquisition. All authors contributed ideas and edited the manuscript.

Data accessibility

Pest distribution data are available with permission from CABI, Nosworthy Way, Wallingford, OX10 8DE, UK

Sources for other datasets used in the analysis are given in the text.

References

Avelino, J., Cristancho, M., Georgiou, S., Imbach, P., Aguilar, L., Bornemann, G., Läderach, P., Anzueto, F., Hruska, A.J. & Morales, C. (2015) The coffee rust crises in Colombia and Central America (2008–2013): impacts, plausible causes and proposed solutions. *Food Security*, **7**, 303–321.

Bacon, S.J., Aebi, A., Calanca, P. & Bacher, S. (2013) Quarantine arthropod invasions in Europe: the role of climate, hosts and propagule pressure. *Diversity and Distributions*, n/a–n/a.

Baiser, B., Olden, J.D., Record, S., Lockwood, J.L. & McKinney, M.L. (2012) Pattern and process of biotic homogenization in the New Pangaea. *Proceedings of the Royal Society B: Biological Sciences*, **279**, 4772–4777.

Baker, R.H.A. (2012) An introduction to the PRATIQUE Research Project. *EPPO Bulletin*, **42**, 1–2.

Baker, R.H.A., Anderson, H., Bishop, S., MacLeod, A., Parkinson, N. & Tuffen, M.G. (2014) The UK Plant Health Risk Register: a tool for prioritizing actions. *EPPO Bulletin*, **44**, 187–194.

Bebber, D.P. (2015) Range-Expanding Pests and Pathogens in a Warming World. *Annual Review of Phytopathology*, **53**, 335–356.

Bebber, D.P., Castillo, Á.D. & Gurr, S.J. (2016) Modelling coffee leaf rust risk in Colombia with climate reanalysis data. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **371**, 20150458.

Bebber, D.P. & Gurr, S.J. (2015) Crop-destroying fungal and oomycete pathogens challenge food security. *Fungal Genetics and Biology*, **74**, 62–64.

Bebber, D.P., Holmes, T. & Gurr, S.J. (2014a) The global spread of crop pests and pathogens. *Global Ecology and Biogeography*, **23**, 1398–1407.

Bebber, D.P., Holmes, T., Smith, D. & Gurr, S.J. (2014b) Economic and physical determinants of the global distributions of crop pests and pathogens. *New Phytologist*, **202**, 901–910.

Bebber, D.P., Ramotowski, M.A.T. & Gurr, S.J. (2013) Crop pests and pathogens move polewards in a warming world. *Nature Climate Change*, **3**, 985–988.

Berzitis, E.A., Minigan, J.N., Hallett, R.H. & Newman, J.A. (2014) Climate and host plant availability impact the future distribution of the bean leaf beetle (Cerotoma trifurcata). *Global Change Biology*, **20**, 2778–2792.

Boakes, E.H., McGowan, P.J.K., Fuller, R.A., Chang-qing, D., Clark, N.E., O'Connor, K. & Mace, G.M. (2010) Distorted views of biodiversity: spatial and temporal bias in species occurrence data. *PLoS Biol*, **8**, e1000385.

Bregaglio, S., Cappelli, G. & Donatelli, M. (2012) Evaluating the suitability of a generic fungal infection model for pest risk assessment studies. *Ecological Modelling*, **247**, 58–63.

CABI (2017) Crop Protection Compendium.

Chapman, D., Purse, B.V., Roy, H.E. & Bullock, J.M. (2017) Global trade networks determine the distribution of invasive non-native species. *Global Ecology and Biogeography*, **26**, 907–917.

DEFRA (2018) UK Plant Health Risk Register.

Department of Agriculture, Fisheries and Forestry (2013) *Final pest risk analysis report for Drosophila suzukii*, Australian Government, Canberra, Australia.

Ding, J., Mack, R.N., Lu, P., Ren, M. & Huang, H. (2008) China's booming economy is sparking and accelerating biological invasions. *BioScience*, **58**, 317.

Dormann, C.F., Schymanski, S.J., Cabral, J., Chuine, I., Graham, C., Hartig, F., Kearney, M., Morin, X., Römermann, C., Schröder, B. & Singer, A. (2012) Correlation and process in species distribution models: bridging a dichotomy. *Journal of Biogeography*, **39**, 2119–2131.

Elith, J. & Leathwick, J.R. (2009) Species Distribution Models: Ecological Explanation and Prediction Across Space and Time. *Annual Review of Ecology, Evolution, and Systematics*, **40**, 677–697.

EPPO (2008) *Report of a Pest Risk Analysis for Xanthomonas axonopodis pv. allii*, European and Mediterranean Plant Protection Organization, Paris.

Eschen, R., Holmes, T., Smith, D., Roques, A., Santini, A. & Kenis, M. (2014) Likelihood of establishment of tree pests and diseases based on their worldwide occurrence as determined by hierarchical cluster analysis. *Forest Ecology and Management*, **315**, 103–111.

FAO (2004) *ISPM 21. Pest risk analysis for regulated non-quarantine pests*, IPPC, FAO, Rome.

Fears, R., Aro, E.-M., Pais, M.S. & ter Meulen, V. (2014) How should we tackle the global risks to plant health? *Trends in Plant Science*, **19**, 206–208.

Fisher, M.C., Henk, D.A., Briggs, C.J., Brownstein, J.S., Madoff, L.C., McCraw, S.L. & Gurr, S.J. (2012) Emerging fungal threats to animal, plant and ecosystem health. *Nature*, **484**, 186–194.

Flood, J. & Day, R. (2016) Managing risks from pests in global commodity networks – policy perspectives. *Food Security*, **8**, 89–101.

Guo, L., Zhu, X.-Q., Hu, C.-H. & Ristaino, J.B. (2010) Genetic Structure of Phytophthora infestans Populations in China Indicates Multiple Migration Events. *Phytopathology*, **100**, 997–1006.

Huang, D., Zhang, R., Kim, K.C. & Suarez, A.V. (2012) Spatial Pattern and Determinants of the First Detection Locations of Invasive Alien Species in Mainland China. *Plos One*, **7**, e31734.

IMF (2013) IMF Data and Statistics.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L. & Daszak, P. (2008) Global trends in emerging infectious diseases. *Nature*, **451**, 990–993.

Klayman, D.L. (1985) Qinghaosu (Artemisinin): An Antimalarial Drug from China. *Science*, **228**, 1049–1055.

Kriticos, D.J. (2012) Regional climate-matching to estimate current and future sources of biosecurity threats. *Biological Invasions*, **14**, 1533–1544.

Kriticos, D.J., Morin, L., Leriche, A., Anderson, R.C. & Caley, P. (2013) Combining a Climatic Niche Model of an Invasive Fungus with Its Host Species Distributions to Identify Risks to Natural Assets: Puccinia psidii Sensu Lato in Australia. *PLoS ONE*, **8**, e64479.

Liebhold, A.M., McCullough, D.G., Blackburn, L.M., Frankel, S.J., Von Holle, B. & Aukema, J.E. (2013) A highly aggregated geographical distribution of forest pest invasions in the USA. *Diversity and Distributions*, **19**, 1208–1216.

MacLeod, A., Jones, G.D., Anderson, H.M. & Mumford, R.A. (2016) Plant health and food security, linking science, economics, policy and industry. *Food Security*, **8**, 17–25.

MacLeod, A., Pautasso, M., Jeger, M.J. & Haines-Young, R. (2010) Evolution of the international regulation of plant pests and challenges for future plant health. *Food Security*, **2**, 49–70.

Monfreda, C., Ramankutty, N. & Foley, J.A. (2008) Farming the planet: 2. Geographic distribution of crop areas, yields, physiological types, and net primary production in the year 2000. *Global Biogeochemical Cycles*, **22**, GB1022.

Ordonez, N., Seidl, M.F., Waalwijk, C., Drenth, A., Kilian, A., Thomma, B.P.H.J., Ploetz, R.C. & Kema, G.H.J. (2015) Worse Comes to Worst: Bananas and Panama Disease—When Plant and Pathogen Clones Meet. *PLOS Pathogens*, **11**, e1005197.

Paini, D.R., Sheppard, A.W., Cook, D.C., Barro, P.J.D., Worner, S.P. & Thomas, M.B. (2016) Global threat to agriculture from invasive species. *Proceedings of the National Academy of Sciences*, **113**, 7575–7579.

Paini, D.R., Worner, S.P., Cook, D.C., De Barro, P.J. & Thomas, M.B. (2010) Threat of invasive pests from within national borders. *Nature Communications*, **1**, 115.

Patpour, M., Hovmøller, M.S., Shahin, A.A., Newcomb, M., Olivera, P., Jin, Y., Luster, D., Hodson, D., Nazari, K. & Azab, M. (2015) First Report of the Ug99 Race Group of Wheat Stem Rust, Puccinia graminis f. sp. tritici, in Egypt in 2014. *Plant Disease*, **100**, 863–863.

Pyšek, P., Jarošík, V., Hulme, P.E., Kühn, I., Wild, J., Arianoutsou, M., Bacher, S., Chiron, F., Didžiulis, V., Essl, F., Genovesi, P., Gherardi, F., Hejda, M., Kark, S., Lambdon, P.W., Desprez-Loustau, M.-L., Nentwig, W., Pergl, J., Poboljšaj, K., Rabitsch, W., Roques, A., Roy, D.B., Shirley, S., Solarz, W., Vilà, M. & Winter, M. (2010) Disentangling the role of environmental and human pressures on biological invasions across Europe. *Proceedings of the National Academy of Sciences*, **107**, 12157–12162.

Pyšek, P., Richardson, D.M., Pergl, J., Jarošík, V., Sixtová, Z. & Weber, E. (2008) Geographical and taxonomic biases in invasion ecology. *Trends in Ecology & Evolution*, **23**, 237–244.

R Development Core Team (2017) *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna.

Robinet, C., Kehlenbeck, H., Kriticos, D.J., Baker, R.H.A., Battisti, A., Brunel, S., Dupin, M., Eyre, D., Faccoli, M., Ilieva, Z., Kenis, M., Knight, J., Reynaud, P., Yart, A. & van der Werf, W. (2012) A Suite of Models to Support the Quantitative Assessment of Spread in Pest Risk Analysis. *PLoS ONE*, **7**, e43366.

Santini, A., Ghelardini, L., De Pace, C., Desprez-Loustau, M.L., Capretti, P., Chandelier, A., Cech, T., Chira, D., Diamandis, S., Gaitniekis, T., Hantula, J., Holdenrieder, O., Jankovsky, L., Jung, T., Jurc, D., Kirisits, T., Kunca, A., Lygis, V., Malecka, M., Marcais, B., Schmitz, S., Schumacher, J., Solheim, H., Solla, A., Szabò, I., Tsopelas, P., Vannini, A., Vettraino, A.M., Webber, J., Woodward, S. & Stenlid, J. (2013) Biogeographical patterns and determinants of invasion by forest pathogens in Europe. *New Phytologist*, **197**, 238–250.

Savary, S., Bregaglio, S., Willocquet, L., Gustafson, D., D'Croz, D.M., Sparks, A., Castilla, N., Djurle, A., Allinne, C., Sharma, M., Rossi, V., Amorim, L., Bergamin, A., Yuen, J., Esker, P., McRoberts, N., Avelino, J., Duveiller, E., Koo, J. & Garrett, K. (2017) Crop health and its global impacts on the components of food security. *Food Security*, **9**, 311–327.

Scimago Lab (2017) Scimago Journal & Country Rank.

Skelsey, P., Cooke, D.E.L., Lynott, J.S. & Lees, A.K. (2016) Crop connectivity under climate change: future environmental and geographic risks of potato late blight in Scotland. *Global Change Biology*, n/a-n/a.

Soberón, J. (2007) Grinnellian and Eltonian niches and geographic distributions of species. *Ecology Letters*, **10**, 1115–1123.

Soberón, J. & Nakamura, M. (2009) Niches and distributional areas: concepts, methods, and assumptions. *Proceedings of the National Academy of Sciences*, **106**, 19644–19650.

Soberón, J. & Peterson, A.T. (2005) Interpretation of models of fundamental ecological niches and species' distributional areas. *Biodiversity Informatics*, **2**, 1–10.

Sutherst, R.W. (2014) Pest species distribution modelling: origins and lessons from history. *Biological Invasions*, **16**, 239–256.

Venette, R.C., Kriticos, D.J., Magarey, R.D., Koch, F.H., Baker, R.H.A., Worner, S.P., Gómez Raboteaux, N.N., McKenney, D.W., Dobesberger, E.J., Yemshanov, D., De Barro, P.J., Hutchison, W.D., Fowler, G., Kalaris, T.M. & Pedlar, J. (2010) Pest Risk Maps for Invasive Alien Species: A Roadmap for Improvement. *BioScience*, **60**, 349–362.

Wangai, A.W., Redinbaugh, M.G., Kinyua, Z.M., Miano, D.W., Leley, P.K., Kasina, M., Mahuku, G., Scheets, K. & Jeffers, D. (2012) First report of maize chlorotic mottle virus and Maize Lethal Necrosis in Kenya. *Plant Disease*, **96**, 1582–1582.

Ward, M. (2016) Action against pest spread—the case for retrospective analysis with a focus on timing. *Food Security*, **8**, 77–81.

Westphal, M.I., Browne, M., MacKinnon, K. & Noble, I. (2008) The link between international trade and the global distribution of invasive alien species. *Biological Invasions*, **10**, 391–398.