Ecological Informatics 61 (2021) 101196

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf

Validation and uncertainty analysis of the match climates regional algorithm (CLIMEX) for Pest risk analysis

Mariona Roigé [a,b,*], Craig B. Phillips [a,b]

[a] AgResearch Ltd, Lincoln, New Zealand
[b] Better Border Biosecurity, New Zealand

ARTICLE INFO

Keywords:
Climate matching
Invasive species
Environmental suitability
Species distribution model
Biosecurity
Validation
Pest risk analysis

ABSTRACT

Pest risk analysis (PRA) is conducted by plant health authorities to identify and implement phytosanitary measures to limit the accidental introduction and spread of species that are, or may become, pests. The 'Match Climates Regional' (MCR) algorithm is part of the CLIMEX species distribution modelling package and is a simple climate matching method that has often been applied in PRA to estimate the potential geographic distributions of weeds, insects and mites. However, there is a lack of studies that address its predictive validity and sensitivity to the input data. Here we investigated the validity of the MCR algorithm by comparing its predictions to empirical observations of the distributions of 30 species of biosecurity concern. We also evaluated MCR's sensitivity to its inputs, using data for a further 30 species, by creating numerous models for each species using different sized subsets of their distribution data, then comparing the results. We found that the algorithm showed overall high accuracy and its outputs were relatively insensitive to sample size. Thus, the results generally supported use of MCR for PRA.

1. Introduction

Trade and tourism are facilitating the global spread of alien invasive species, which can harm both natural and modified ecosystems. To reduce the international spread of such species, many nations are signatory to the Agreement on the Application of Sanitary and Phytosanitary Measures (SPS Agreement) under the umbrella of the International Plant Protection Convention (International Plant Protection Convention (IPPC), 2007).

National phytosanitary agencies typically conduct Pest Risk Analyses (PRA) to identify the alien invasive species that present risks to geographical areas (PRA areas) within their jurisdictions. PRA methods were reviewed by Leung et al. (2012) and often involve both quantitative and qualitative approaches. PRA generally have three main stages: The first identifies the species that could be transported on the entry pathway(s) of interest; The second evaluates the species' potential to establish and persist in the PRA area, often employing various measures of habitat and climate suitability; And the third assesses the species' potential impacts in the PRA area. Based on the information obtained, the species are assigned to risk categories, and methods for preventing the entry and establishment of the most important risks are developed.

Species distribution models (SDMs) are often used to assess the potential environmental suitability of PRA areas for invasive species. SDMs are statistical models that use species' occurrence locations and associated environmental data to predict the species' potential future geographical distributions. A plethora of methods can be used to develop SDMs (reviewed by Elith and Leathwick (2009)) ranging from simple generalised linear models to machine learning algorithms. The method chosen must be appropriate to the available data, otherwise its predictions will be biased (Guillera-Arroita et al., 2015).

In PRA, species distribution models must often be constructed when only minimal information about the species' geographical distribution and ecology is available: Records of locations where species are absent are particularly rare (Phillips et al., 2006). Species occurrence data are often gathered from sources such as the Global Biodiversity Information Facility (GBIF) and the CABI Crop Pest Compendium (CPC). Such databases do not contain absence records, and presence records are often biased by differences between geographic regions in accessibility, political stability, scientific infrastructure and socioeconomic factors (Meyer et al., 2015; Meyer et al., 2016). Moreover, the databases' taxonomic coverage is influenced by factors such as species' conspicuousness and their degree of interest to observers (Meyer et al., 2016).

* Corresponding author at: AgResearch Ltd, Lincoln, New Zealand.
E-mail address: mariona.roige.valiente@gmail.com (M. Roigé).

https://doi.org/10.1016/j.ecoinf.2020.101196
Received 26 April 2020; Received in revised form 14 October 2020; Accepted 31 October 2020
Available online 9 November 2020

Risk analysts employed by national phytosanitary agencies often must work at the 'speed of commerce' to evaluate risks from long lists of potentially hazardous species in short time intervals. Thus, choice of method used to construct SDMs must consider the information that is available to create the model, the modeler's level of expertise, and the time required to complete the modelling (Venette et al., 2010). In a guide to selecting models to support biosecurity decision making, Froese (2012) regarded the time required to build the model as a prime consideration and identified climate matching approaches such as Climatch (Crombie et al., 2008) and CLIMEX Match Climates Regional (Sutherst and Maywald, 1985) as particularly suitable. Similarly, Magarey et al. (2018) noted that speed, simplicity and reproducibility were key requirements and suggested that information gathering and modelling should take less than 2 h per species. Magarey et al. (2018) also considered climate matching approaches such as CLIMEX Match Climates Regional to be well suited to PRA.

Climate matching (climate envelope) models are a subset of species distribution models that solely use climatic variables to make spatial predictions of environmental suitability for a species (Watling et al., 2013). Match Climates Regional (MCR) is a climate matching algorithm from CLIMEX software (Kriticos et al., 2015) that estimates the climatic suitability of a PRA area for a species by comparing climatic data from locations where the species is present to climatic data in the PRA area.

CLIMEX MCR has often been applied in PRA to estimate the potential geographic distributions of weeds, insects and mites. Its algorithm calculates a single value called the Composite Match Index (CMI) to represent the degree of climatic similarity between two locations. There are two ways to use MCR for PRA. The first compares climatic data collected at locations where the species is present to climatic data collected from the PRA area. Each presence location is compared to every location in the PRA area and assigned the highest CMI found. The potential climatic suitability of the PRA area for a species can then be evaluated by calculating, for example, the proportion of presence locations with CMIs that exceed some threshold value. This method is quick and simple because results from a single large MCR model that has compared climatic data collected from throughout the entire PRA area to data collected from throughout all other regions (potentially the rest of the world, e.g. b3nz.shinyapps.io/CMI-maps-csv) can be used to help evaluate risks from numerous species with widely differing geographical distributions. An evaluation of this method's utility for PRA (Phillips et al., 2018) concluded that it was helpful for predicting whether climatic factors in a PRA area could limit species' potential to establish there, though some caveats were identified and recommended for further study.

The second way to use MCR is the reverse of the first: Each location in the PRA area is compared to every location where the species of interest has been recorded then assigned the highest CMI calculated. Thus, this approach produces a climatic suitability map of the PRA area that may be used in conjunction with information on habitat availability (e.g. host plant presence) to help predict the species' potential geographic distribution in the PRA area. This method is slower and more complex than the first because a separate MCR model must be developed for each species of interest. To our knowledge, no published studies have evaluated the performance of this second approach for PRA.

Here, we evaluate the validity of predictions of species' potential distributions based on climate envelopes calculated by CLIMEX MCR and perform an uncertainty analysis to investigate the effects of the input data on MCR's results. We use two sets of species of biosecurity concern; one for validation and the other for uncertainty analysis. For validation, the potential distribution of each species (based solely on climatic variables) is estimated using MCR and the results are compared to each species' observed distribution. In the uncertainty analysis, we measure the response of MCR's composite match index (CMI) to the number of climate locations used to build the model.

2. Methods

2.1. CLIMEX match climates regional algorithm

CLIMEX's match climates regional algorithm (MCR) compares meteorological data from two locations to calculate an index of climatic similarity, the composite match index (CMI), which ranges from 0 (poorly matched) to 1 (perfectly matched) (Kriticos, 2012; Sutherst and Maywald, 1985). By default, MCR calculates CMIs using records of each location's monthly rainfall, and minimum and maximum monthly temperature, with all three variables weighted at one. There are options to adjust the weight and include data for soil moisture and relative humidity. Before calculating the CMIs, MCR interpolates the monthly climate records to weekly. Depending on the options chosen, MCR calculates up to six separate indices of climatic similarity (match) between two locations (maximum temperature, minimum temperature, total rainfall, rainfall pattern, humidity, and soil moisture). It then combines the indices to produce the CMI.

2.2. The match climates regional algorithm as a climate envelope

CLIMEX MCR defines the region for which the CMIs will be calculated as the HOME region, and other areas to which the HOME region will be compared as the AWAY region. The steps to build a climate suitability projection for a HOME region using MCR are: 1. Obtain AWAY presence points where the species of interest has been recorded. 2. Divide the AWAY area into a grid of climatic cells, with the centre of each cell corresponding to a location for which climatic data are available. 3. Select the AWAY cells where the species has been recorded one or more times. 4. Divide HOME into a grid of cells with their corresponding climatic data 5. Compare the climate of each HOME cell with that of every AWAY cell where the species has been recorded and assign each HOME cell the highest CMI found.

This process creates potential for many HOME cells to find their best matches with the same AWAY cell, whereas other AWAY cells may fail to produce the best match for any HOME cells. We call AWAY cells that become matched with at least one HOME cell 'influential', and investigate relationships between the number of AWAY cells, the number of influential AWAY cells, and the HOME cells' CMIs.

2.3. Climate data used for validation and uncertainty analysis

New Zealand was treated as the HOME region, for which we used climate data recorded between 1960 and 2004 for 11,471 locations, which corresponds to a spatial resolution of 0.05° (Tait et al., 2006). The remainder of the world was treated as the AWAY region, for which we used climate data that were provided with CLIMEX for 61,076 locations, which corresponds to a resolution of 0.5°. The CLIMEX data were recorded between 1964 and 1990, and were derived from a global splined grid produced by the Climate Research Unit at Norwich, UK (Kriticos et al., 2015). MCR was used with its default variables (rainfall, and minimum and maximum temperature) at their default weights of one.

2.4. Validation

2.4.1. Test species for validation

Thirty species representing a range of taxonomic groups (Table 1) were chosen based on two main criteria: They were established in the HOME region (New Zealand); and had reliable distribution data (sources described below).

2.4.2. Sources of HOME species distribution data for validation

Distribution data for the HOME region were required only for the validation study. For plants ($n = 16$ species), distribution data were obtained from the *Australasian virtual herbarium* (2019); For ants ($n = 3$),

Table 1
List of species used for model validation. 'Difference' is the increase in the number of suitable HOME cells (CMI ≥ 0.7) between using 10 AWAY cells to create the models to using all available AWAY cells (total HOME cells is 11,471).

Species	Common name	Taxon	Max number climate cells	Difference
Acacia longifolia	golden wattle	plant	330	3354
Amblyopone australis	southern michelin ant	ant	113	69
Carduus nutans	musk thistle	plant	1480	3453
Cirsium vulgare	spear thistle	plant	11	1359
Citysius scoparis	Scotch broom	plant	1344	1660
Conium maculatum	poison hemlock	plant	1759	3586
Galium aparine	cleavers	plant	2508	1734
Heteronychus arator	African black beetle	beetle	30	1357
Latrodectus hasselti	redback spider	spider	382	723
Mecinus pascuorum		weevil	192	2682
Mentha pulegium	pennyroyal	plant	816	2408
Naupactus cervinus	Fuller's rose beetle	weevil	151	2117
Naupactus leucoloma	white fringed weevil	weevil	129	1194
Ochetellus glaber	black house ant	ant	113	3676
Oenanthe pimpinelloides	corky-fruited water-dropwort	plant	216	1918
Otiorhynchus ovatus	strawberry root weevil	weevil	564	6080
Otiorhynchus rugosostriatus	rough strawberry weevil	weevil	159	2277
Otiorhynchus sulcatus	vine weevil	weevil	481	2481
Pennisetum clandestinum	kikuyu grass	plant	280	1419
Persicaria maculosa	redshank	plant	2139	3932
Phlyctinus callosus		weevil	28	104
Pilosella officinarum	mouse-ear hawkweed	plant	1579	3703
Pinus pinea	stone fruit pine	plant	340	2175
Pinus radiata	Monterey pine	plant	380	1130
Sitona discoideus	lucerne weevil	weevil	121	1537
Sitona obsoletus	clover root weevil	weevil	378	4567
Sonchus oleraceus	sowthistle	plant	3549	3649
Tetramorium grassii	pavement ant	ant	113	2415
Tradescantia fluminensis	small leaf spiderwort	plant	356	2401
Ulex europaeus	gorse	plant	827	439

data were obtained from a New Zealand ant distribution database (Manaaki Whenua-Landcare); For the beetle *Heteronychus arator*, unpublished data were provided by a local expert (W. King, AgResearch, pers. comm.); Published data were used for the spider *Latrodectus hasselti* (Bryan et al., 2015) and the weevil *Sitona obsoletus* (Ferguson et al., 2012; Hardwick et al., 2016); And for the remaining eight weevils, distribution data obtained from the New Zealand Arthropod Collection (D. Ward, Landcare Research, personal communication).

2.4.3. Sources of AWAY species distribution data for validation

Species distribution data for the AWAY cells for all the species were obtained from the Global Biodiversity Information Facility (GBIF, https://gbif.org) online platform. The records were then 'cleaned' using the R package CoordinateCleaner (Zizka et al., 2019) which performed a series of data quality tests on the occurrence data such as: remove duplicated records, records found in biodiversity facilities (such like museums and herbariums) or at GBIF headquarters, records that point to the centroid of the country, and records with latitude or longitude values of zero. Further processing verified that coordinates matched the country specified by GBIF, and removed records with high spatial uncertainty.

2.4.4. Sub-sampling of AWAY climate cells for validation

Two HOME climatic suitability projections were created for each species using two subsets of each species' AWAY climate cells. The first subset comprised ten randomly selected AWAY cells, and the second comprised all the available AWAY cells.

2.4.5. Validation method

Model performance was assessed by comparing each species' climatic suitability projection to its actual distribution and calculating truth tables. A truth table is a summary table that shows, for a given suitable/unsuitable threshold CMI, the number of True Presences (TP), False Presences (FP), True Absences (TA) and False Absences (FA) the model has produced.

We focused on the ability of the model to correctly identify TP and FA. A TP is an instance where the species was present in a HOME climate cell which the model predicted as suitable, and a FA occurs when the model deems a HOME climatic cell as unsuitable, but the species is nonetheless present. In contrast, we did not evaluate FP (model incorrectly predicts a presence) and TA (model correctly predicts an absence) because the comprehensive species presence/absence data required to do so were unavailable. With FP, for example, there are numerous possible non-climatic reasons why a species may be absent from a climate cell that the model predicts is climatically suitable (e.g. absence of host plants, success of preventive control measures, etc.), but that does not make the climate suitability prediction wrong. Similarly, robust absence data for any species are seldom available, thus confirming TA would require intensive sampling of each climatically unsuitable cell, which is impractical. We only report TP using the metric Proportion of True Presences (PTP), which is TP/Total Presences, and thus Proportion of False Absences (PFA) is implied since TP + FA = Total Presences.

For the thirty validation species, we evaluated the truth tables after assuming that CMI ≥ 0.7 were climatically suitable and CMI < 0.7 were unsuitable because this empirically derived threshold has been widely adopted in previous studies (Kriticos, 2012).

2.5. Uncertainty analysis

2.5.1. Test species for uncertainty analysis

We chose thirty species with AWAY geographic distributions that exhibited a range of characteristics which we considered relevant to our analysis (Table 2). We reasoned that species with large geographic distributions spanning many climate cells would be probabilistically more likely to generate many climatically matched locations in the HOME region than species with smaller AWAY distributions, thus we included species with relatively large and small AWAY distributions.

2.5.2. Sources of AWAY species distribution data for uncertainty analysis

All AWAY species distribution data were obtained from the Global Biodiversity Information Facility (GBIF, https://gbif.org). The records were cleaned programmatically following a series of tests which included; omitting the records without coordinates, delete records with high spatial uncertainty, delete fossil records and delete records further away than 10 km from its country border.

2.5.3. Sub-sampling of AWAY climate cells for uncertainty analysis

For the uncertainty analysis, different sized randomly chosen (with replacement) subsets of each species' AWAY climate cells were used to create numerous HOME climatic suitability projections for each species. The subsets ranged from a minimum of ten AWAY climate cells to the maximum number of cells available for a species in increments of ten. Thus, the total number of HOME projections produced per species varied with its total number of available climate cells.

2.5.4. Uncertainty analysis method

For each of thirty species (Table 2), the uncertainty analysis (Makowski (2013)) investigated the relationship between the number of

Table 2
Species and corresponding taxonomic group used for uncertainty analysis with their maximum number of AWAY climate cells, and the mean CMI of the HOME region obtained when all available AWAY cells were used in the CLIMEX MCR model. 'Intercept', 'b' and 'R2' contain the values of the linear-logarithmic regression between the number of climate cells used in building the model and the value of CMI.

Species	Taxon	Max number climate cells	Mean CMI	Intercept	β	R^2
Acacia longifolia	plant	330	0.79	0.66	0.025	0.711
Aedes aegypti	insect	2670	0.72	0.74	0.032	0.802
Aphis gossypii	insect	270	0.75	0.63	0.02	0.804
Bactrocera tryoni	insect	40	0.68	0.57	0.032	0.783
Bemisia tabaci	insect	220	0.66	0.33	0.065	0.791
Bromus lanceolatus	plant	270	0.75	0.63	0.021	0.579
Capusa senilis	insect	20	0.76	0.75	0.006	1
Cardiocondyla minutior	insect	80	0.65	0.23	0.098	0.818
Carduus nutans	plant	1480	0.8	0.68	0.016	0.867
Clidemia hirta	plant	820	0.67	0.4	0.042	0.818
Drosophila suzukii	insect	90	0.74	0.65	0.021	0.926
Gabriola dyari	insect	40	0.7	0.6	0.027	0.938
Halyomorpha halys	insect	570	0.77	0.67	0.017	0.877
Homalodisca vitripennis	insect	90	0.65	0.6	0.01	0.71
Lymantria dispar	insect	880	0.79	0.67	0.018	0.886
Mentha aquatica	plant	1310	0.8	0.7	0.015	0.896
Miconia calvescens	plant	280	0.68	0.51	0.03	0.682
Monomorium floricola	insect	150	0.63	0.3	0.065	0.603
Myzus persicae	insect	100	0.73	0.65	0.017	0.655
Oecophylla smaragdina	insect	240	0.48	0.36	0.023	0.654
Paratrechina longicornis	insect	380	0.69	0.41	0.047	0.642
Pennisetum clandestinum	plant	280	0.77	0.63	0.025	0.691
Pinus pinea	plant	340	0.78	0.66	0.021	0.914
Pinus radiata	plant	380	0.8	0.71	0.016	0.894
Popillia japonica	insect	500	0.72	0.64	0.014	0.758
Solenopsis geminata	insect	280	0.67	0.4	0.05	0.776
Spodoptera frugiperda	insect	240	0.69	0.59	0.019	0.837
Ulex europaeus	plant	1000	0.82	0.71	0.016	0.915
Vanessa cardui	insect	2990	0.8	0.68	0.015	0.893
Wasmannia auropunctata	insect	260	0.64	0.37	0.049	0.665

AWAY climate cells used to construct the projection and the CMIs of the HOME cells. The analysis comprised three steps: 1. Selecting a set of thirty GBIF distribution records of species of biosecurity concern. Although Makowski (2013) suggested random sampling, we used the criteria specified in 'Test species for uncertainty analysis' to better simulate real PRA scenarios. Choosing to study 30 species was a compromise that balanced computation time with statistical power. 2. CLIMEX MCR was used to create separate HOME climate suitability models for each species and sample size. To reduce computation time, we wrote and used a C++ version of MCR on a high performance computing cluster. We also recorded the AWAY cell that was best matched with each HOME cell, thus enabling the AWAY cells that influenced each model to be identified. 3. The results were used to investigate the relationship between the number of AWAY cells and the values of the HOME CMIs.

2.6. Choice of variable

We hypothesized that CMIs would tend to increase with number of AWAY cells because HOME cells would have higher probabilities of finding closely matched AWAY cells when the latter were numerous. We used the mean CMI of all cells in each HOME projection as our main indicator of the predicted climatic suitability of the HOME region for a species, and modeled the relationship between mean CMI and number of AWAY cells as a log-linear relationship. We also investigated an alternative indicator of climatic suitability, proportion of HOME cells with CMI $\geq$ 0.7 (Annex 2), and found it was closely correlated with mean CMI (Annex 3), thus we only report the latter in the main text.

2.7. Influential AWAY cells

We hypothesized that any increase in CMI with number of AWAY cells would approach an asymptote as AWAY samples became increasingly saturated with climate cells closely matched to the HOME cells. For each subsample of each species' AWAY climate cells, we recorded the identities of the AWAY cells that had the highest climatic matches with one or more HOME cells and termed these 'influential AWAY cells'. Then, for each species and subsample, the number of influential AWAY cells was divided by the total number of AWAY cells and termed the 'proportion influential'. According to our hypothesis, the proportion influential should decline with increasing total AWAY cells. The relationship between the proportion influential and the number of AWAY cells was modeled using log-linear regression. Additionally, we investigated the relationship between the total number (rather than the proportion) of influential AWAY cells and the number of AWAY cells in the sample and the frequencies at which those influential AWAY cells became influential in all the subsamples for a species.

3. Results

3.1. Model validity

The assessment of the proportions of true presences (PTP) and false absences (FA) generally indicated that CLIMEX MCR predicted species distributions with reasonable accuracy irrespective of the number of AWAY cells used to create the model, though model accuracy improved with number of AWAY cells. When just ten AWAY cells were used, the mean proportion of TP for the thirty species was 80%, and this increased to 96% when all available AWAY cells were used (see Fig. 1 and Fig. 2). The increase in proportion TP with number of AWAY cells was evident for all but one of the thirty species (Fig. 2).

3.2. Uncertainty analysis

3.2.1. Relationship between mean CMI and number of climate cells in the sample

The mean of the 11,471 CMIs calculated for the HOME area increased at a declining rate with the number of climate cells in the AWAY area, but the magnitude of the increase was small (e.g. Fig. 3, Annex 1). This pattern was evident in all species except *Capusa senilis*, *Gabriola dyari* and *Bactrocera tryoni* (Annex 1). For *Aedes aegyptii* (Fig. 3), which was representative of many species (Annex 1), the predicted increase in CMI associated with a 10% increase in number of AWAY cells was 0.00304%. All 30 species had a similar beta parameter in their logarithmic regression, which ranged from 0.006 (*Capusa senilis)* to 0.098 (*Cardiocondyla minuitor*) (Table 3).

Fig. 4 shows an example of how HOME CMIs changed spatially with increasing AWAY cells. With just ten AWAY cells, the HOME CMIs were similar, though often marginally lower, than the CMIs obtained with more AWAY climate cells. Thus, the MCR algorithm showed relatively low sensitivity to the number of AWAY climate cells.

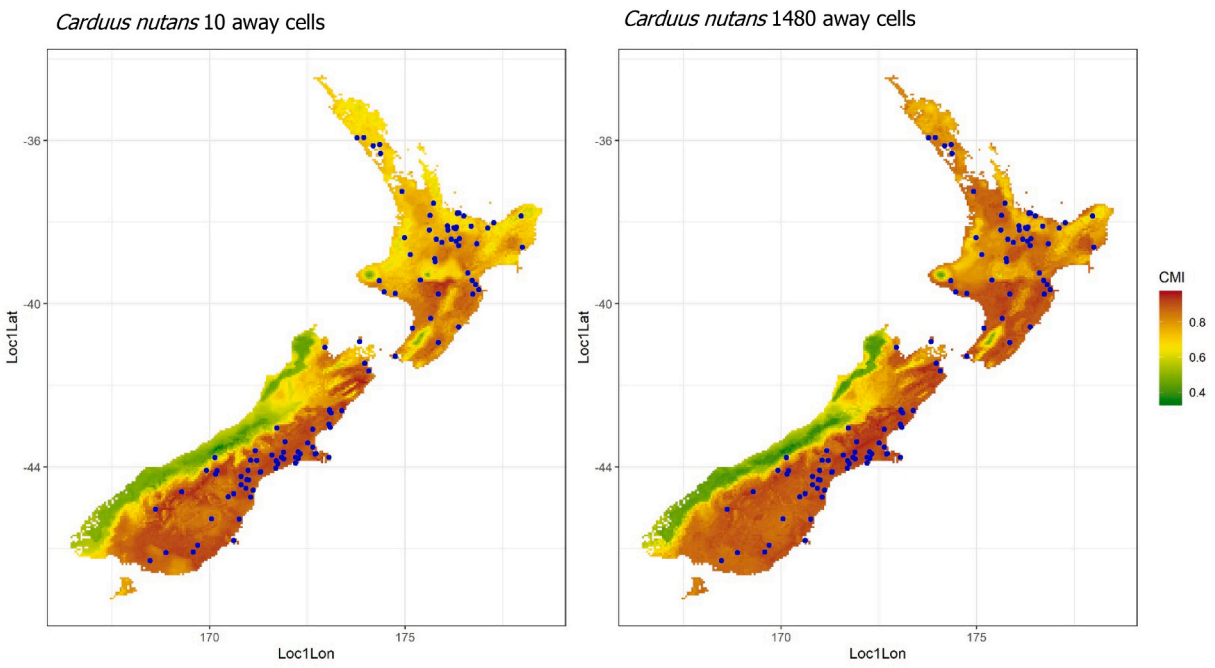


Fig. 1. Example validation for Carduus nutans. Green-red fill in left map shows HOME CMI values calculated using 10 AWAY cells, and blue points shows locations where the species has been observed in the HOME region. The right map shows the same data after calculating HOME CMI values using all available AWAY cells (n = 1480). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

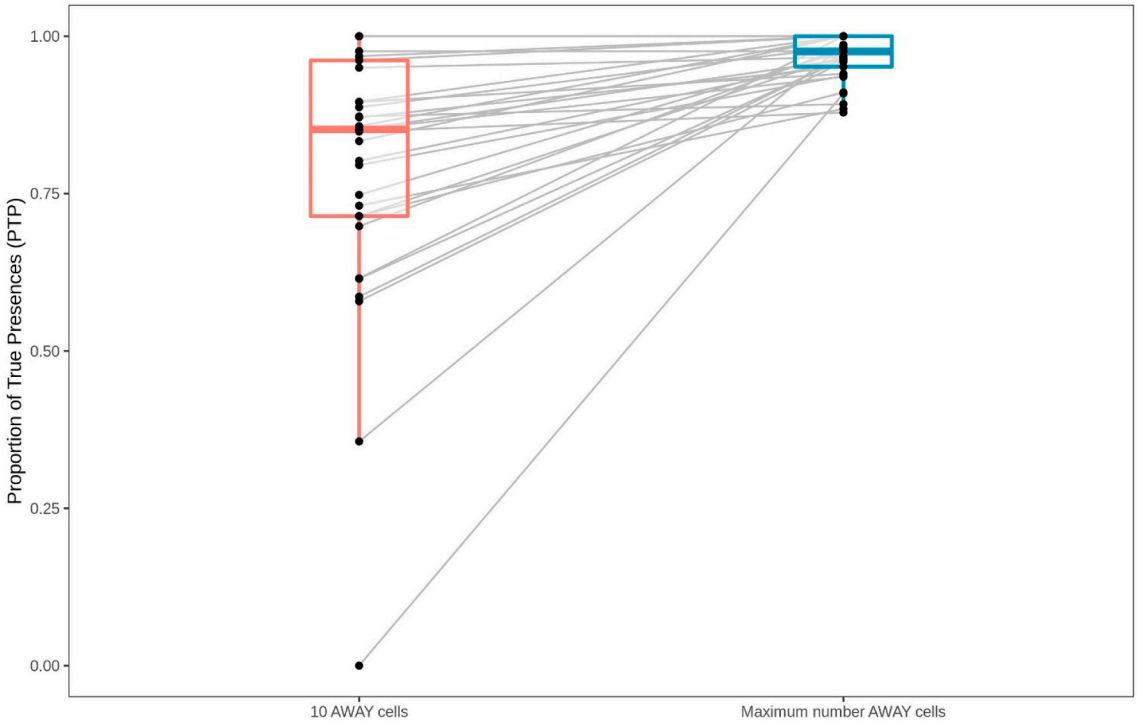


Fig. 2. Proportion of true presences obtained for each of 30 species when ten randomly sampled AWAY cells were used per species (left) compared to when all available AWAY cells were used (right). Lines connect results for the same species.

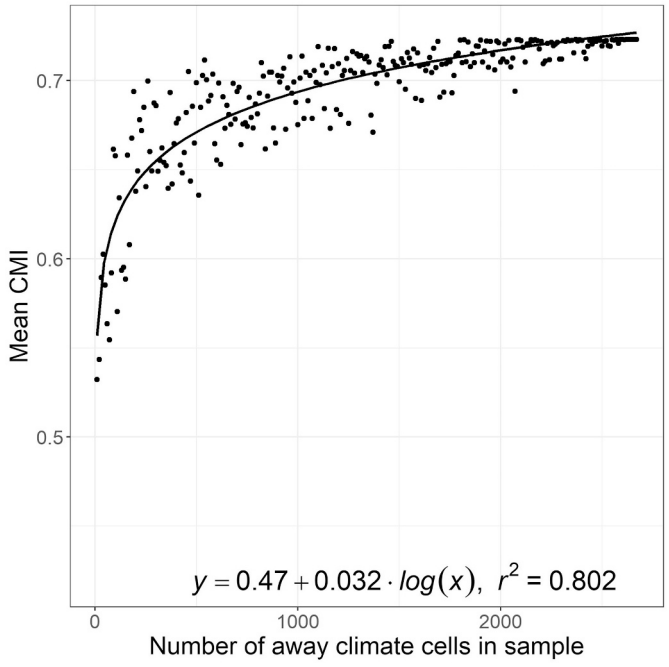


Fig. 3. Mean CMI increase for Aedes aegypti with number of climate cells in the sample.

3.3. Proportions of AWAY climate cells that were influential

In each subsample of each species' AWAY climate cells, we recorded the identity of the AWAY cell that had the highest climatic match with each HOME cell and termed these 'influential AWAY cells'. The proportions of influential cells varied with species and subsample (Annex 4) though, as expected, they generally declined with increasing number of AWAY cells. For each species, when all available AWAY cells were used the proportion influential ranged from 1% for *Aedes aegyptii* ($n = 2670$ AWAY cells) to 73% for *Gabriola dyarii* ($n = 40$ AWAY cells) (Annex 6). The shape of the decrease in proportion influential was logarithmic and similar for all species, thus, *C. nutans* is presented as an example (Fig. 5), whereas results for other species are given in Annex 4. The only exceptions were *Bactrocera tyroni*, *Capusa senilis* and *Gabriola dyari* which showed more linear responses probably due to the small number of AWAY cells available to perform the analysis (Annex 4). For all species, there were some AWAY cells that became influential much more frequently than others (Annex 5b).

4. Discussion

4.1. Model validity

Our results generally supported the value of CLIMEX MCR for pest risk analysis as a quick reproducible method for obtaining first estimates of insects' and plants' potential geographic distributions in short time

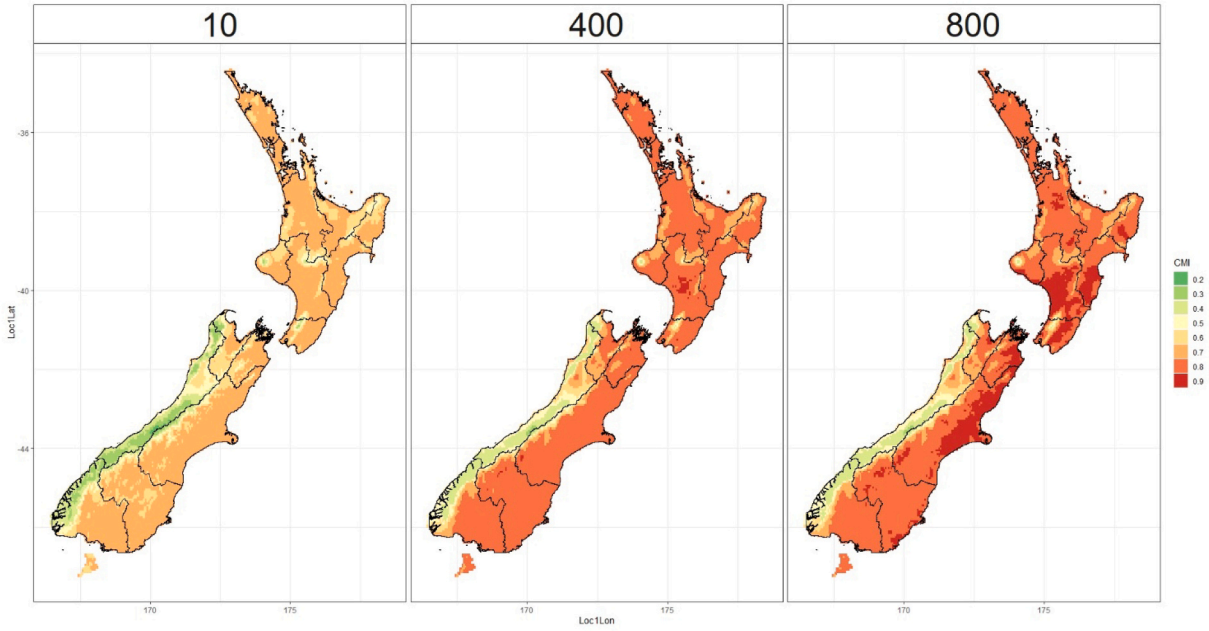


Fig. 4. HOME CMIs obtained for Aedes aegypti using 10, 400 and 800 randomly sampled AWAY climate cells.

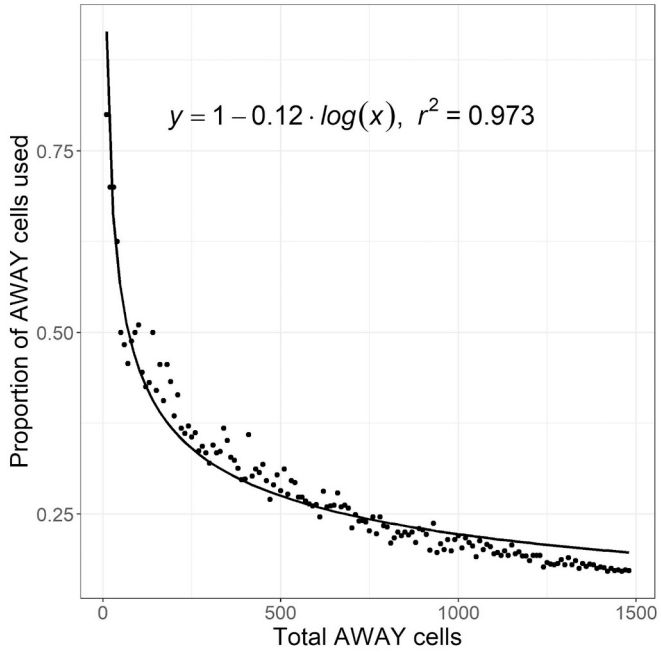


Fig. 5. Proportion of influential AWAY cells versus total AWAY cells for Carduus nutans.

intervals using relatively little ecological information. Even when the models used only 10 AWAY cells, the proportion of true presences was high for most species and only two species had proportions <50%. And when all available AWAY cells were used, models for all species had proportions of true presences >80%.

Climate envelope models such as CLIMEX MCR have been accused of overestimating distributions (Sutherst, 2013). Such overestimates, however, are difficult to discern due to the general insufficiency of species distribution data, which means locations incorrectly deemed as climatically suitable by dint of overestimation may be confounded with climatically suitable areas where the species is present but unrecorded. Moreover, species may be absent from climatically suitable areas due to non-climatic factors such as habitat availability. Nevertheless, in our study the regular occurrence of low proportions of false absences (i.e. species occurrences in locations classified as climatically unsuitable) in most of the models developed for each of the 30 test species may suggest that overestimation was modest. It has also been argued that some overestimation is useful in PRA because it is risk-averse and may reflect invasive species' tendency to expand their range (Jiménez-Valverde et al., 2011).

4.2. Model uncertainty with respect to the number of AWAY climate locations

As expected, mean HOME CMI increased with number of AWAY climate cells, but only slightly. One reason for the modest increase was the small subset of AWAY cells that influenced each prediction; once most influential cells were included in a subsample, the response of mean CMI to the number of AWAY cells declined. We suggest that it would be valuable when using the CLIMEX MCR algorithm to develop methods for identifying influential AWAY cells, as we did in our study, because it is the species records within these influential cells that are most important to scrutinise for validity.

The rates of CMI increase varied between species, but we did not identify any species characteristics that might contribute to this variation. One possibility we considered was that more geographically widespread species would produce higher mean CMIs, since they should occur in a greater variety of climates. We evaluated this idea by calculating the number of biogeographical realms each species occurred in and examining its relationship with mean CMI. However, the results showed no relationship, thus we did not include them in the paper.

With the aim of providing recommendations for the use of MCR in PRA, we tried to find a minimum number of AWAY climate cells that would ensure reliable projections. We found that as few as ten AWAY cells provided useful results, though model accuracy improved slightly when more AWAY cells were used.

4.3. Proportion of influential AWAY cells

A possible reason why the CMI values did not present high increases in value could because of Influential AWAY cells that were 'very influential', that is, they had the climatic characteristics that allow them to become Best Matches to a large number of HOME cells. Since the sampling was random, those very influential AWAY cells had smaller chance to be included in the sample when the sample size was small but had increasing probability to be included when the sampling size increased. Once those very influential locations were included in the sample, the proportion of AWAY locations that become influential decreased, since the very influential locations become Best Matches to more HOME locations.

4.4. The use of a suitability threshold

Through repeated use of CLIMEX MCR it has become common practice to apply CMI ≥ 0.7 as the threshold for climatic suitability. We challenged this idea by exploring how changes in the number of AWAY climate cells influenced mean CMI. If mean CMI had been highly sensitive to number of AWAY cells, then the suitability threshold applied would need to be adjusted with the number of AWAY cells. However, our results showed that mean CMI increased only marginally with number of AWAY cells, thus it seems pragmatic to continue using just one threshold irrespective of number of AWAY cells.

4.5. Model results uncertainty

We have provided measures of model performance and uncertainty for the use of MCR in the context of PRA, as recommended in Venette et al. (2010).

Due to their fairly common climatic characteristics, some HOME locations could have higher probability of finding a match with AWAY locations, whereas some particularly exceptional areas could struggle more. In New Zealand, areas such as Fiordland, the Southern Alps, and Egmont National Park have unique patterns of rainfall and temperature that could maybe make them harder to find a match. It would be expected that for these types of 'rare' climates, the variability of the CMI

amongst data subsamples would be higher. However, that would also depend on what AWAY locations they are being compared to, that is, on the distribution of the particular species being assessed. Further research is needed to understand whether relatively unique HOME cells are less likely to find close matches in AWAY regions, and whether those 'unique' HOME cells are more sensitive to the number of AWAY cells used in the study. Practical approaches to reduce model uncertainty include cleaning the data before running the model and identifying those influential AWAY cells that match with more than one HOME cell. If a very influential AWAY cell (ie. matches with many HOME locations) corresponds to an erroneous record then its effects on model outputs are higher than those of a less influential AWAY cell (ie, matches with few or no HOME locations).

MCR is a simple algorithm and shares most of the limitations with other climate matching tools. Its major limitation is that many additional non-climatic factors that are not included in the model—such as dispersal, food availability, natural enemies and competition—also help to determine species distributions. The method is also susceptible to sampling biases in species occurrence records, extrapolation errors (although less so than correlative SDM due to its focus on ecological parameters (climatic characteristics) rather than statistical correlations and does not factor in spatial autocorrelation. Other practical limitations include limited validation of the results other than visual because there is no statistical measure of model fit (Froese, 2012). Also, climate modelling at the spatial grain MCR was applied (0.5 degrees for the world, 0.05 for NZ), cannot account for microclimates. Although resolution is a choice of the modeler or limited by data availability, it represents a trade-off between accuracy and computation time. Despite all that, this study showed that the predictions of climatic suitability produced by MCR are informative in the context of PRA, when the time and knowledge of the species biology is limited.

4.6. Recommendations

We can recommend using the MCR algorithm to construct projections of climatic suitability for the HOME location even when the number of available AWAY climate locations is low. Ten climatic locations have shown to produce good results in terms of discrimination of True Presences and False absences.

The more climatic locations available the better in terms of accuracy. The model will be able to better differentiate true presences and false absences.

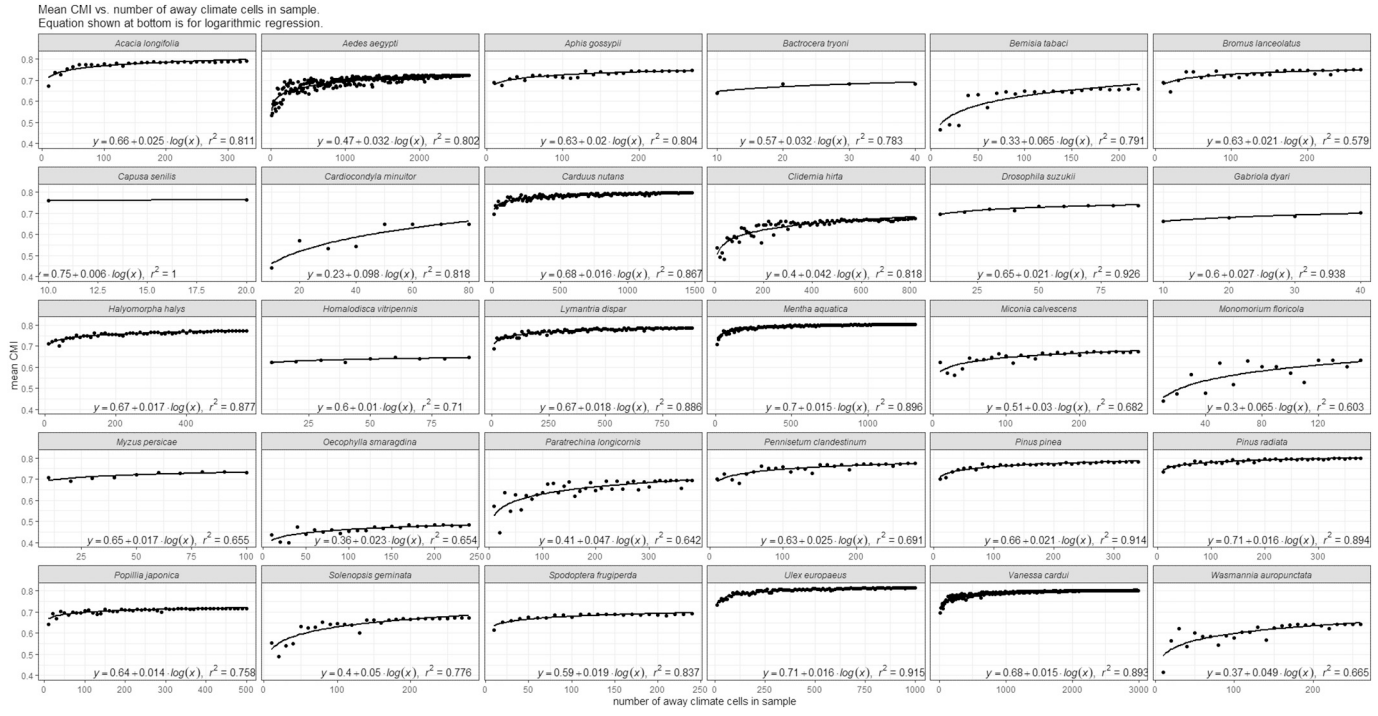
The use of the CMI ≥ 0.7 as a threshold for suitability was supported by both the accurate results obtained from validation results and the finding CMIs increase only slowly as the number of AWAY climate cells increases.

Declaration of Competing Interest

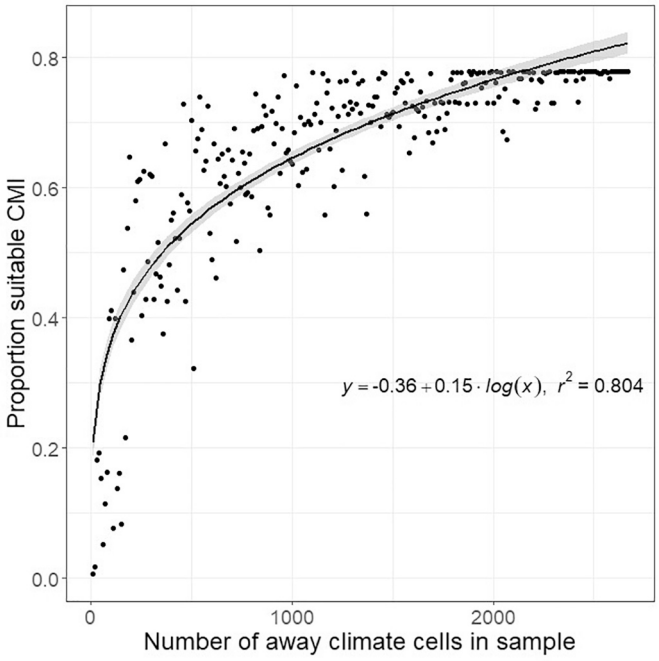
None.

Acknowledgments

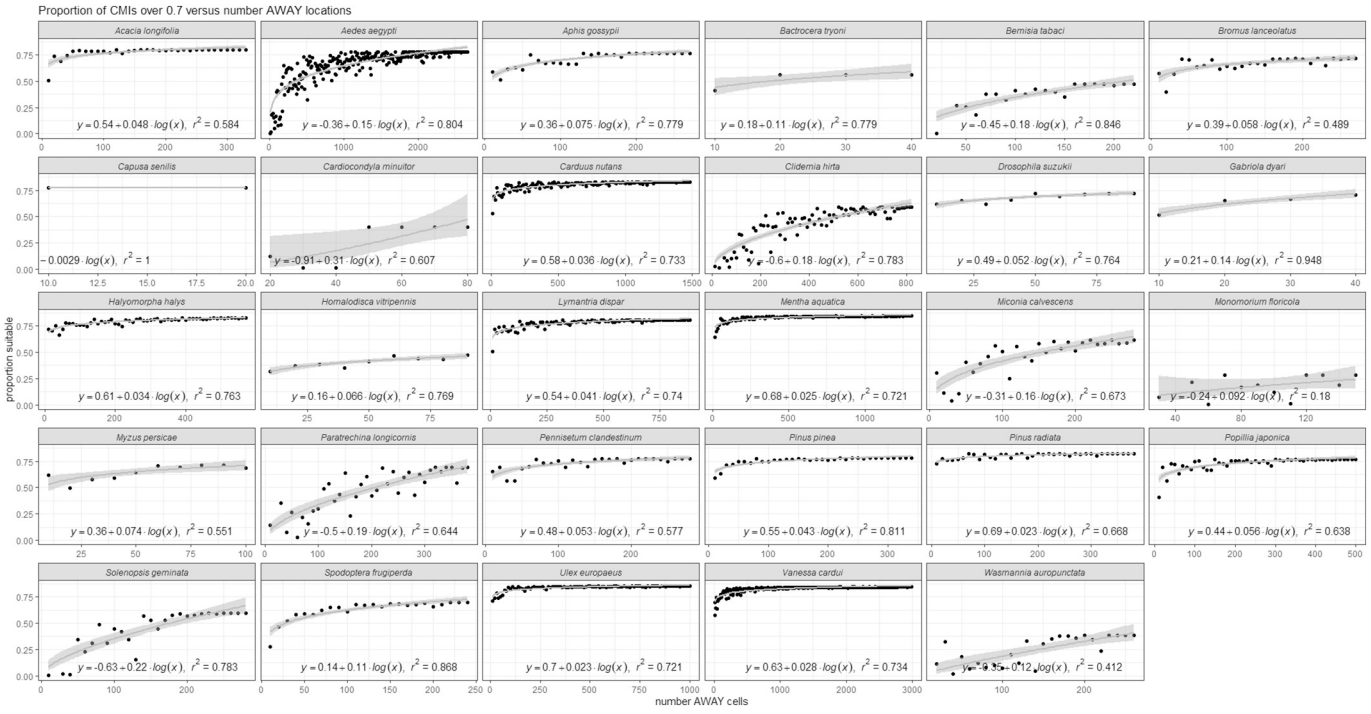
We thank Warren King (AgResearch, Ruakura) for unpublished data on the distribution of *Heteronychus arator* in New Zealand, and Darren Ward (Manaaki-Whenua Landcare Research, Auckland) for data from the New Zealand Arthropod Collection on the New Zealand distributions of eight weevil species. We also thank the following people from AgResearch at Lincoln: John Kean, Russel McAuliffe and Dan Sun for assistance writing MCR in C++ and implementing it in a high performance computing cluster, and Chikako Van Koten for statistical advice. We wish to acknowledge the Agricultural and Marketing Research and Development Trust (AGMARDT) postdoctoral fellowship P18002 (to MR) and AgResearch via its contribution to the Better Border Biosecurity research collaboration (www.b3nz.org).



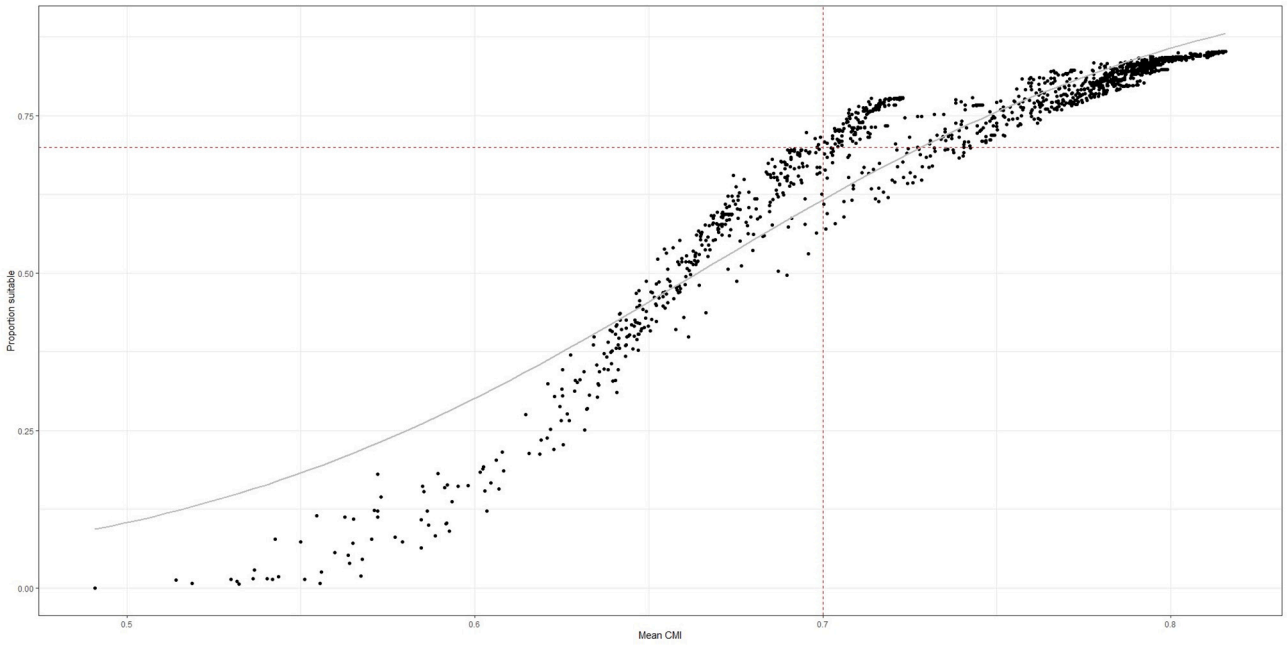
Annex 1. Relationship between mean CMI and number of AWAY cells (all species).



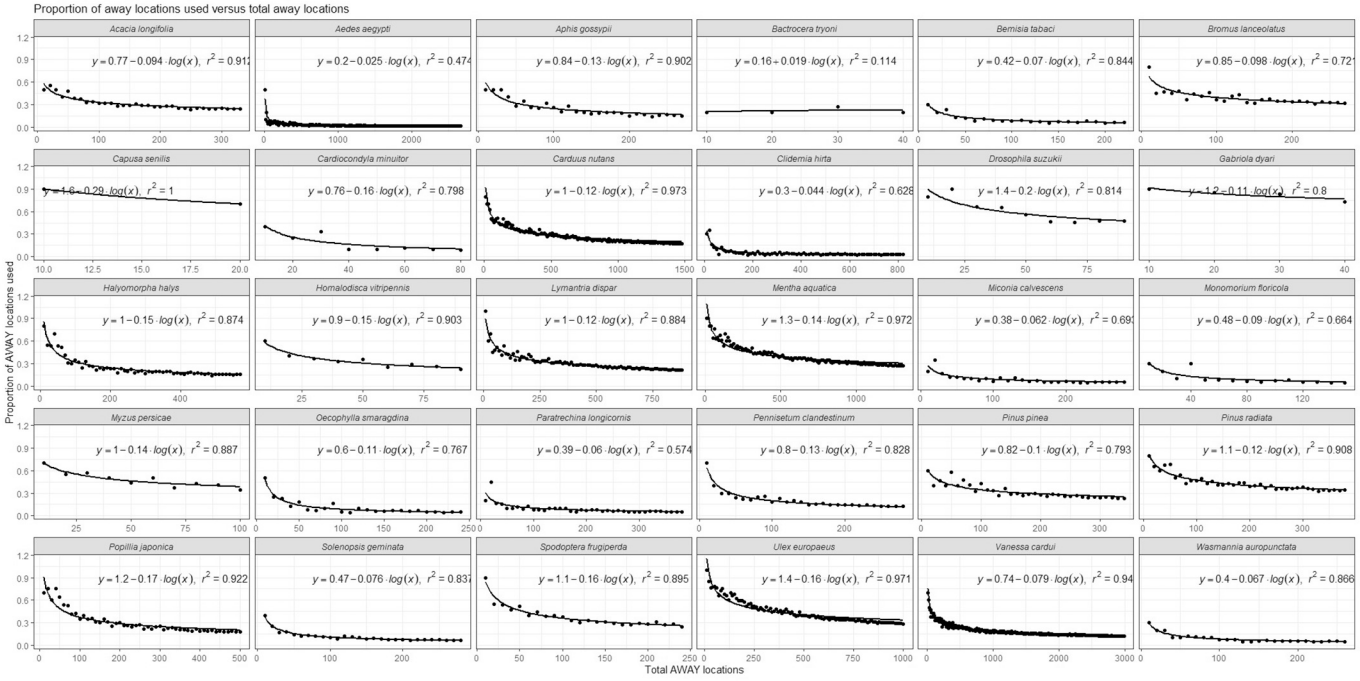
Annex 2a. Proportion of HOME cells with CMI ≥ 0.7 (*Aedes aegypti*).



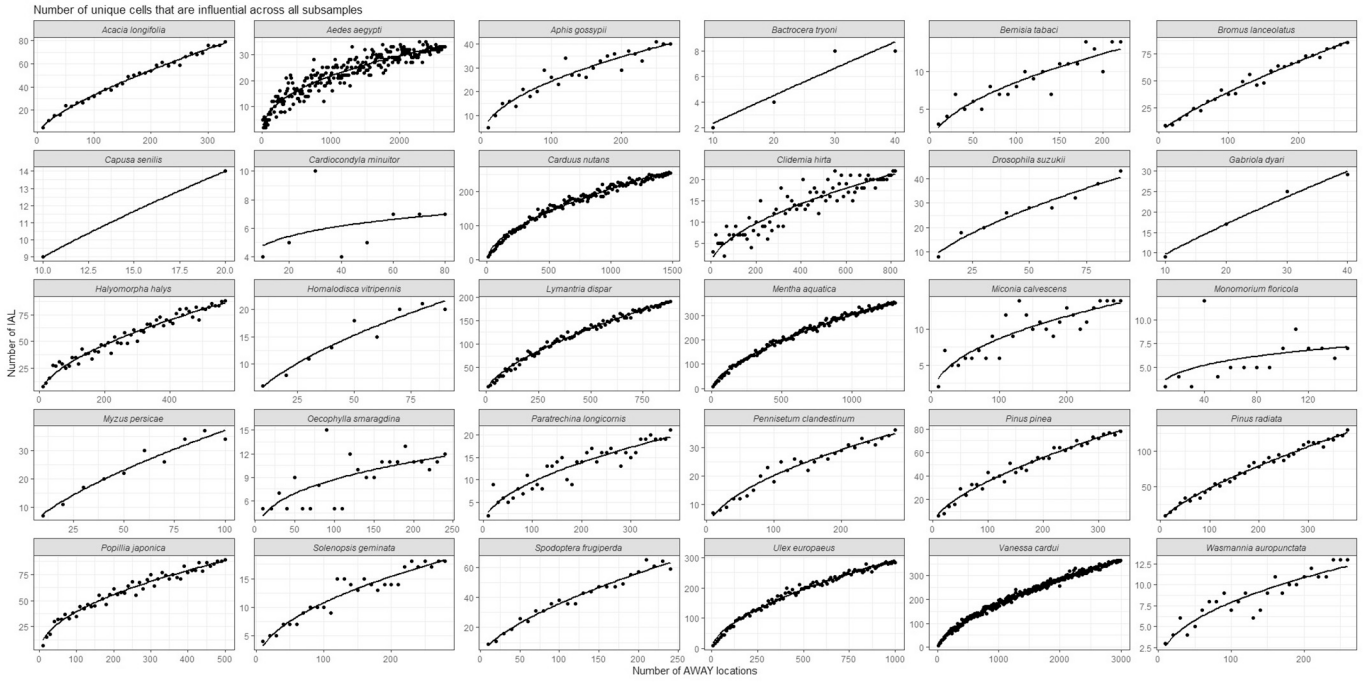
Annex 2b. Proportion of HOME cells with CMI ≥ 0.7 (all species).



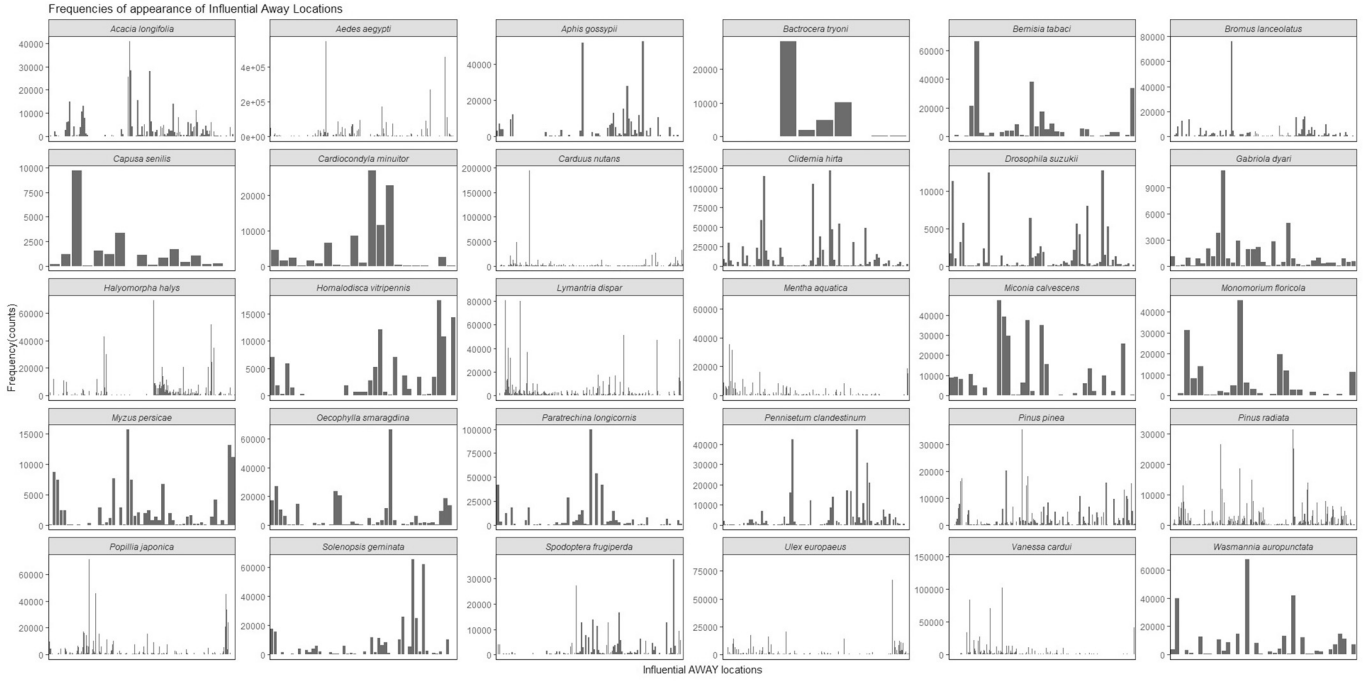
Annex 3. Relationship between variables 'mean CMI' and 'proportion of HOME cells $\geq$ 0.7', all species.



Annex 4. Relationship between proportion influential AWAY cells and number of AWAY cells (all species).



Annex 5.a. Numbers of influential AWAY locations and their increase with number of AWAY locations in the sample.



Annex 5.b. Frequencies at which AWAY locations were influential over all samples for all species. For each species, each bar shows the number of times that AWAY climate cell was influential. Some counts exceed the theoretical maximum of 11,471 per model because they are the sums of all models created for each species.

Species	min_p_used
Acacia longifolia	24%
Aedes aegypti	1%
Aphis gossypii	14%
Bactrocera tryoni	20%
Bemisia tabaci	5%
Bromus lanceolatus	31%
Capusa senilis	70%
Cardiocondyla minutior	9%
Carduus nutans	17%
Clidemia hirta	2%
Drosophila suzukii	46%
Gabriola dyari	73%
Halyomorpha halys	14%
Homalodisca vitripennis	22%
Lymantria dispar	22%
Mentha aquatica	27%
Miconia calvescens	5%
Monomorium floricola	4%
Myzus persicae	34%
Oecophylla smaragdina	5%
Paratrechina longicornis	5%
Pennisetum clandestinum	12%
Pinus pinea	23%
Pinus radiata	32%
Popillia japonica	18%
Solenopsis geminata	6%
Spodoptera frugiperda	25%
Ulex europaeus	28%
Vanessa cardui	12%
Wasmannia auropunctata	5%

Annex 6. Minimum percentage of climate locations that were influential for each species.

Table A6
Percentage of climate locations that became influential when all AWAY cells were used.

Species	min_p_used
Acacia longifolia	24%
Aedes aegypti	1%
Aphis gossypii	14%
Bactrocera tryoni	20%
Bemisia tabaci	5%
Bromus lanceolatus	31%
Capusa senilis	70%
Cardiocondyla minutior	9%
Carduus nutans	17%
Clidemia hirta	2%
Drosophila suzukii	46%
Gabriola dyari	73%
Halyomorpha halys	14%
Homalodisca vitripennis	22%
Lymantria dispar	22%
Mentha aquatica	27%
Miconia calvescens	5%
Monomorium floricola	4%
Myzus persicae	34%
Oecophylla smaragdina	5%
Paratrechina longicornis	5%
Pennisetum clandestinum	12%
Pinus pinea	23%
Pinus radiata	32%
Popillia japonica	18%
Solenopsis geminata	6%
Spodoptera frugiperda	25%
Ulex europaeus	28%
Vanessa cardui	12%
Wasmannia auropunctata	5%

References

Australasian virtual herbarium. Available at: https://avh.chah.org.au/.

Bryan, S.A., et al., 2015. Invasive redback spiders (Latrodectus hasseltii) threaten an endangered, endemic New Zealand beetle (Prodontria lewisii). J. Insect Conserv. 19 (5), 1021–1027. https://doi.org/10.1007/s10841-015-9818-x. Springer International Publishing.

Crombie, J., et al., 2008. 'Climatch user manual', Australian Government. Bureau Rural Sci. 16.

Elith, J., Leathwick, J.R., 2009. Species distribution models: ecological explanation and prediction across space and time. Annu. Rev. Ecol. Evol. Syst. 40 (1), 677–697. https://doi.org/10.1146/annurev.ecolsys.110308.120159.

Ferguson, C.M., et al., 2012. Status of clover root weevil and its biocontrol agent in the South Island after six years. Proc. N. Z. Grassland Assoc. 74, 171–176.

Froese, J., 2012. A Guide to Selecting Species Distribution Models to Support Biosecurity Decision-Making. https://doi.org/10.13140/RG.2.1.4956.0409.

Guillera-Arroita, G., et al., 2015. Is my species distribution model fit for purpose? Matching data and models to applications. Glob. Ecol. Biogeogr. 24 (3), 276–292. https://doi.org/10.1111/geb.12268.

Hardwick, S., et al., 2016. Response to clover root weevil outbreaks in South Canterbury, Otago and Southland; the agricultural sector and government working together. J. N. Z. Grasslands 2010, 117–122. Available at: https://www.grassland.org.nz/publications/nzgrassland_publication_2823.pdf.

International Plant Protection Convention (IPPC), 2007. ISPM 2 Framework for Pest Risk Analysis. Rome. Available at: https://www.ippc.int/fr/publications/592/.

Jiménez-Valverde, A., et al., 2011. Use of niche models in invasive species risk assessments. Biol. Invasions 13 (12), 2785–2797. https://doi.org/10.1007/s10530-011-9963-4.

Kriticos, D.J., 2012. Regional climate-matching to estimate current and future sources of biosecurity threats. Biol. Invasions 14 (8), 1533–1544. https://doi.org/10.1007/s10530-011-0033-8.

Kriticos, D.J., et al., 2015. CLIMEX Verion 4: Exploring the Effects of Climate on Plants, Animals and Diseases. CSIRO, Canberra, ACT.

Leung, B., et al., 2012. Teasing apart alien species risk assessments: a framework for best practices. Ecol. Lett. 15 (12), 1475–1493. https://doi.org/10.1111/ele.12003.

Magarey, R., et al., 2018. Comparison of four modeling tools for the prediction of potential distribution for non-indigenous weeds in the United States. In: Biological Invasions, 20. Springer International Publishing, pp. 679–694. https://doi.org/10.1007/s10530-017-1567-1 (3).

Makowski, D., 2013. Uncertainty and sensitivity analysis in quantitative pest risk assessments; practical rules for risk assessors. NeoBiota 18, 157–171. https://doi.org/10.3897/neobiota.18.3993.

Meyer, C., et al., 2015. Global priorities for an effective information basis of biodiversity distributions. In: Nature Communications, 6. Nature Publishing Group, pp. 1–8. https://doi.org/10.1038/ncomms9221.

Meyer, C., et al., 2016. Range geometry and socio-economics dominate species-level biases in occurrence information. Glob. Ecol. Biogeogr. 25 (10), 1181–1193. https://doi.org/10.1111/geb.12483.

Phillips, C.B., et al., 2018. Utility of the CLIMEX "match climates regional" algorithm for pest risk analysis: an evaluation with non-native ants in New Zealand. In: Biological Invasions, 20. Springer International Publishing, pp. 777–791. https://doi.org/10.1007/s10530-017-1574-2 (3).

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distributions. Ecol. Model. 6 (2–3), 231–259.

Sutherst, R.W., 2013. Pest species distribution modelling: origins and lessons from history. Biol. Invasions. https://doi.org/10.1007/s10530-013-0523-y.

Sutherst, R.W., Maywald, G.F., 1985. A computerised system for matching climates in ecology. Agric. Ecosyst. Environ. 13 (3–4), 281–299. https://doi.org/10.1016/0167-8809(85)90016-7.

Tait, A., et al., 2006. Thin plate smoothing slipe interpolation of daily rainfall for New Zealand using a climatological rainfall surface. Int. J. Climatol. 26 (3), 2097–2115. https://doi.org/10.1002/joc.

Venette, R.C., et al., 2010. Pest risk maps for invasive alien species. A roadmap for improvement. Biosicence 60, 349–362.

Watling, J.I., et al., 2013. Use and Interpretation of Climate Envelope Models: A Practical Guide. University of Florida, pp. 1–43. https://doi.org/10.1136/acupmed-2012-010169.

Zizka, A., et al., 2019. CoordinateCleaner: standardized cleaning of occurrence records from biological collection databases. Methods Ecol. Evol. 10 (5), 744–751. https://doi.org/10.1111/2041-210X.13152.