

**Mang T., Essl F., Moser D. & Dullinger S. (2018): Climate warming drives invasion history of *Ambrosia artemisiifolia* in central Europe. – *Preslia* 90: 59–81.**

**Electronic Appendix 1. – *Species records collection and date information of records***

We collected *Ambrosia artemisiifolia* L. records for the countries studied in central Europe until (inclusively) 2010 (see also Mang et al. (2018)). The records were gathered from many different sources. For Austria and Germany, we built an updated version of a pre-existing database (Essl et al. 2009, Richter et al. 2013). For the other countries, we searched the floristic literature, databases of the national floristic mapping projects, major herbaria and gathered unpublished records from colleagues. Such, a total of 11,856 records were collected. Since naturalization of this species had not occurred before the 20th century we did not consider earlier records, and this provided a total of 11,800 records for the modelling period 1900–2010 (see Table 1 in the main text).

Most of the 11,800 records stem from recent years: only 134 records refer to years up to 1950, 252 up to 1960, 512 up to 1970, and 990 up to 1980. There were 1,868 records up to 1990, 3,508 up to 2000, 4,788 up to 2005, and finally, 11,800 up to 2010. These 11,800 records mapped to a total of 3,598 floristic mapping grid cells (see Fig. 1 in the main text). Most records (75%) could be unambiguously assigned to a single year. For the remaining 25% of the records, we only had information on the interval within which this species had been recorded. The general kind of temporal information used in this study was therefore interval-censored dates (with date intervals varying in length across records). Records were condensed to time of first detection in each grid cell by using the date of the earliest record for each cell; when this record referred to a date interval of two or more years, the information of all records from that particular cell was used to minimize the interval length. The hierarchical model accounted for the residual date uncertainty of interval-censored dates (see Electronic Appendix 2). In the simplified non-hierarchical model, for each record with an interval-censored date a year was randomly drawn from a uniform distribution on the record's date interval to obtain a reference year. These reference years were also used (i) for date summaries of records (e.g. the chronology of the cumulative number of records stated above); and (ii) in model validation, for unique assignment of records either to the fitting data set or validation data set in the case that a date interval referred to years of both the fitting period and the validation period.

**Electronic Appendix 2.** – *Hierarchical and non-hierarchical Bayesian models, generalized waiting-time distribution and likelihood function*

*Hierarchical Bayesian model*

In the hierarchical model, the invasion process layer models the time of first invasion in each grid cell by the species, and the observation process layer models the time of first detection of the species' occurrence in each invaded grid cell (Mang et al. 2017). The invasion process layer and the observation process layer are linked by a cell's invasion time, which is the outcome of the invasion process and concurrently the start time of the observation process. Species records document the outcome of the observation process whereas the actual invasion times remain unknown. Moreover, for records with interval-censored dates the precise years of detection are also unknown.

Bayesian inference approaches (Gelman et al. 2004) with data augmentation are particularly well suited for fitting models with a large number of unknown quantities. In our model, the actual invasion times of cells are quantities of interest and therefore regarded as additional parameters of the posterior distribution. By contrast, for interval-censored dates the precise years of first detection (within the date intervals specified by the records) are regarded as nuisance parameters and therefore integrated out. Let  $\mathbf{x}$  be the actual invasion times of all cells,  $\mathbf{y}$  be the actual (precise) times of first detection in invaded cells,  $\mathbf{z}$  be the interval-censored dates of first detection in invaded cells stated by the records (i.e. the data),  $\boldsymbol{\theta}$  be all model parameters used in the invasion process, and  $\boldsymbol{\delta}$  be all model parameters used in the observation process; then the Bayesian posterior distribution of the hierarchical model is given by

$$p(\boldsymbol{\theta}, \boldsymbol{\delta}, \mathbf{x} | \mathbf{z}) = \frac{\int p(\boldsymbol{\theta}, \boldsymbol{\delta}) p(\mathbf{x} | \boldsymbol{\theta}) p(\mathbf{y} | \mathbf{x}, \boldsymbol{\delta}) p(\mathbf{z} | \mathbf{y}) d\mathbf{y}}{p(\mathbf{z})}, \quad (\text{E2.1})$$

where  $p(\boldsymbol{\theta}, \boldsymbol{\delta})$  is the (joint) prior distribution of  $\boldsymbol{\theta}$  and  $\boldsymbol{\delta}$ ,  $p(\mathbf{x} | \boldsymbol{\theta})$  is the invasion process likelihood,  $p(\mathbf{y} | \mathbf{x}, \boldsymbol{\delta})$  is the observation process likelihood,  $p(\mathbf{z} | \mathbf{y})$  is the detection interval likelihood, and  $p(\mathbf{z})$  is the marginal distribution of  $\mathbf{z}$  (a normalizing constant).

*Non-hierarchical Bayesian model*

In the non-hierarchical model, it is assumed that the documented spread accurately reflects the actual spread of the species. In this model, observation and detection issues are thus ignored. Effectively, under these assumption  $\mathbf{x}$ ,  $\mathbf{y}$  and  $\mathbf{z}$  from the hierarchical model are equivalent and, consequently, in the non-hierarchical model record dates are directly used as the cells' invasion times and only the invasion process layer is used (as outlined in Electronic Appendix 1, in our study for each record with an interval-censored date a year was randomly drawn from a uniform distribution on the record's date interval to obtain a reference year). Let  $\mathbf{x}$  be the invasion times of all cells that are, according to the model's assumptions, accurately stated by the records (i.e. the data), and  $\boldsymbol{\theta}$  be all model parameters used in the invasion process; then the Bayesian posterior distribution of the non-hierarchical model is given by

$$p(\boldsymbol{\theta} | \mathbf{x}) = \frac{p(\boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta})}{p(\mathbf{x})}, \quad (\text{E2.2})$$

where  $p(\boldsymbol{\theta})$  is the prior distribution of  $\boldsymbol{\theta}$ ,  $p(\mathbf{x} | \boldsymbol{\theta})$  is the invasion process likelihood, and  $p(\mathbf{x})$  is the marginal distribution of  $\mathbf{x}$  (a normalizing constant). The structure of the non-hierarchical model given by equation (E2.2) corresponds also to the structure of the models used in Mang et al. (2018).

*Generalized waiting-time distribution*

Invasion times and detection times can be equivalently expressed as waiting-times since model start time until invasion and since first invasion into a grid cell until first detection, respectively (Mang et al. 2017). The invasion risk function given by equation (2) in the main text and the detectability function given by equation (3) in the main text quantify the respective waiting-times by parameterizing the generalized waiting-time distribution (Arens et al. 2009). This distribution is central to survival analysis where objects become exposed to some hazard, and the random variable of interest is the residual time from initial hazard exposure until failure occurs. The magnitude of this hazard imposed at a given time is quantified by the hazard function (also called

hazard rate). In our model, grid cells are initially (at the model start time) unoccupied but exposed to an invasion risk; the invasion risk function is therefore the hazard function for a cell's invasion time distribution. Where this invasion occurs, species occurrences in the respective cells are subsequently detectable; the detectability function is therefore the hazard function for a cell's detection time distribution.

Let  $W$  be a general waiting-time random variable,  $b(t)$  be its associated hazard function, and  $t_0$  be the start time of hazard exposure; then the cumulative distribution function (CDF) of a continuous random variable  $W$  is given by

$$F_W(w; b(t), t_0) = P(W \leq w; b(t), t_0) = 1 - e^{-\int_{t_0}^w b(t) dt} \quad (\text{E2.3})$$

(for  $w \geq t_0$ ). Conveniently, the hazard function  $b(t)$  only needs to be  $\geq 0$ , otherwise no constraints are imposed onto its shape. Specifically, at any time it can be increasing, decreasing, or constant. For  $b(t) = 0$  no hazard is imposed by definition and hence the event of interest cannot occur at time  $t$ . Higher values of  $b(t)$  correspond to earlier expected failure time (i.e. in our model earlier invasion time or detection time).

In survival analysis, the complementary cumulative distribution function (CCDF) is referred to as the survival function as it states the probability that an object could retain its original state until time  $w$ , and is given by

$$S_W(w; b(t), t_0) = P(W > w; b(t), t_0) = e^{-\int_{t_0}^w b(t) dt} \quad (\text{E2.4})$$

(for  $w \geq t_0$ ). Evaluated for the model end time  $t_e$ , that is  $w = t_e$ , the survival function hence states the probability that a cell remained unoccupied (invasion process) or the probability that an invaded cell remained undetected (observation process) during the modelling period.

Equations (E2.3 & E2.4) use continuous time. In practice, however, invasion times and detection times may frequently be measured in discrete units (e.g. years, as in our study). Let  $t_{s_k}$  and  $t_{e_k}$  denote the start time and end time of the  $k$ -th discrete modelling sub-period (here a given year), respectively, then the probability mass function (PMF) of a discrete random variable  $W$  is given by

$$f_W(w_k; b(t), t_0) = P(W = w_k) = e^{-\int_{t_0}^{t_{s_k}} b(t) dt} - e^{-\int_{t_0}^{t_{e_k}} b(t) dt}. \quad (\text{E2.5})$$

As equation (E2.5) uses exclusively the integral of  $b(t)$ , it makes no difference whether the hazard function is specified as a truly continuous function or as a step function stating an averaged value per sub-period. Using the step function approach can greatly facilitate the alignment with environmental data and/or facilitate the computing implementation, and so this was used in our study (i.e. the invasion risk function and detectability function changed values at annual intervals and thus effectively had a discrete resolution).

For more statistical properties of the waiting-times until first invasion in grid cells and until first detection in invaded grid cells see also Mang et al. (2017) and Mang et al. (2018).

### *Invasion process likelihood*

The invasion process likelihood considers the invasion times of all grid cells (Mang et al. 2017). Let  $\Psi_m$  be the set of cells which became invaded during the modelling period; then for each cell in  $\Psi_m$  the likelihood assesses the cell's specific invasion time,  $x_i$ , by using the probability mass function given by equation (E2.5). Conversely, let  $\Psi_e$  be the complementary set of cells still unoccupied at the model end time; then for each cell in  $\Psi_e$  the likelihood assesses the probability that invasion has not occurred yet by using the survival function given by equation (E2.4). Let  $t_s$  be the model start time,  $t_e$  be the model end time, and let each cell's invasion risk function,  $g_i(t)$ , be dependent on all invasion process parameters  $\theta$  as given by equation (2) in the main text, then the invasion process likelihood is therefore:

$$p(\mathbf{x} | \theta) = \prod_{i \in \Psi_m} f_{X_i}(x_i; g_i(t), t_s) \times \prod_{i \in \Psi_e} S_{X_i}(t_e; g_i(t), t_s). \quad (\text{E2.6})$$

### *Observation process likelihood*

The observation process likelihood considers the times of first detection of all cells which were invaded until the model end time (Mang et al. 2017). Let  $\Phi_d$  be the subset of these cells for which the species has also been recorded during the modelling period; then for each cell in  $\Phi_d$  the likelihood assesses the cell's specific

detection time,  $y_i$ , by using the probability mass function given by equation (E2.5). Conversely, let  $\Phi_u$  be the complementary subset of cells which were invaded but for which no record is available by the model end time; then for each cell in  $\Phi_u$  the likelihood assesses the probability that detection has not occurred yet by using the survival function given by equation (E2.4). Let each invaded cell's detectability function,  $h_i(t)$ , be dependent on all observation process parameters  $\delta$  as given by equation (3) in the main text, then the observation process likelihood is therefore:

$$p(\mathbf{y} | \mathbf{x}, \delta) = \prod_{i \in \Phi_d} f_{Y_i}(y_i; h_i(t), x_i) \times \prod_{i \in \Phi_u} S_{Y_i}(t_e; h_i(t), x_i). \quad (\text{E2.7})$$

#### *Detection interval likelihood*

For all cells in which the species has been recorded during the modelling period, i.e. the cells in  $\Phi_d$ , the detection interval likelihood considers the actual (precise) times of first detection within the respective date intervals. Let  $\mathbf{z}$  be the interval-censored dates of first detection in invaded cells stated by the records (i.e. the data), where for each cell the information of all records from that particular cell was used to minimize the interval length. In our model, the actual (precise) times of first detection within the respective date intervals were assessed using uniform distributions on each cell's date interval (i.e. equal probability for each year), and the detection interval likelihood is therefore:

$$p(\mathbf{z} | \mathbf{y}) = \prod_{i \in \Phi_d} \begin{cases} \frac{1}{z_{b_i} - z_{a_i} + 1} & \text{for } y_i \in [z_{a_i}, z_{b_i}] \\ 0 & \text{otherwise} \end{cases}, \quad (\text{E2.8})$$

where  $z_{a_i}$  and  $z_{b_i}$  are the (inclusive) lower and upper endpoints (years), respectively, of the date interval of cell  $i$ . If both endpoints of a cell's date interval refer to the same year then in the context of an annual modelling resolution for this cell a precise date of first detection is available.

### **Electronic Appendix 3. – GIS data sources and data processing**

#### *Lattice system representing the study area*

The study area was represented by a lattice of grid cells of size 5'×3' (~ 6×6 km<sup>2</sup>) corresponding to the Central European Floristic Mapping Project (Niklfeld 1998), comprising a total of 22,451 cells (see Fig. 1 in the main text; and see also Mang et al. (2018)). Terrestrial cell area was calculated by subtracting freshwater and marine area as documented by the Corine 2000 data set (Büttner et al. 2002) from total cell area. For edge cells, only the cells' area within the countries studied was considered.

#### *Spatio-temporal variables*

Variables used to model the invasion process (mean temperature and total precipitation during the growing season; the proportion of cropland area and of urban area; and the length of motorway and railway networks) and the observation process (human population density in and around a grid cell; and intensified sampling for *Ambrosia artemisiifolia* in particular regions and years) were compiled on a spatio-temporal basis (per grid cell and year) from the following sources (see also Mang et al. (2017) and Mang et al. (2018)).

#### *Climate data*

Climate data were taken from the CRU TS 1.2 data set (Mitchell et al. 2004), providing a resolution of 10', for the period 1901–1949 (for 1900 we used the mean from 1901–1920); and from the E-OBS 5.0 data set (Haylock et al. 2008), providing a resolution of 0.25°, for the period 1950–2010. Annual temperature and precipitation values were calculated by averaging and summing, respectively, the monthly values from the growing season April–October for each cell and year.

#### *Land use data*

Data on the proportion of cropland area and of urban area were taken from the History Database of the Global Environment (HYDE), Version 3.1 (Klein Goldewijk et al. 2010, Klein Goldewijk et al. 2011) and the Corine 2000 data set (Büttner et al. 2002). HYDE data provide a spatial resolution of 5' and a historical time series at decennial intervals. On the contrary, Corine data provide a much finer spatial resolution of 100 m with a minimum mapping area of 25 ha, but lack historical data for the 20th century. We therefore combined these two data sets by adding the per-cell anomaly between HYDE and Corine data for the year 2000 (this year was available in both data sets) to HYDE values. Cropland and urban area values for individual years (within the decennial intervals) were linearly interpolated. To obtain the proportion as an area-independent measure of environmental suitability for invasion, absolute cropland and urban area values were divided by total terrestrial area of the individual grid cell.

Spatial motorway and railway vector data were taken from OpenStreetMap (OpenStreetMap contributors 2011). For each cell we calculated the total length of motorway and railway networks. To create a time series, we used pooled historical trade data from the study area (Bolt & van Zanden 2014). For years not documented in these data the values were linearly interpolated, and the time series values were scaled such that a value of 1 corresponded to the year 1900 (= first year of the modelling period). For motorways we overlaid the spatial data directly with this time series, whereas for railways we assumed that the network has remained, on average, constant. We acknowledge that this approach only approximates the historical development of motorway and railway networks; accuracy is however higher for the most recent decades when *Ambrosia artemisiifolia* was spreading most intensively. To obtain the proportion as an area-independent measure of environmental suitability for invasion, absolute motorway and railway length values were divided by the square-root (for dimensionality matching) of total terrestrial area of the individual grid cell.

#### *Intensified sampling for *Ambrosia artemisiifolia**

Intensified sampling for *Ambrosia artemisiifolia* was classified as a spatio-temporal binary indicator variable (1 for intensified sampling, and 0 otherwise). For Austria, this classification was performed at the level of political districts, where intensified sampling occurred in the following districts since 1996: Braunau am Inn, Ried im Innkreis and Schärding; and in the following districts since 2005: Baden, Bregenz, Eisenstadt,

Eisenstadt-Umgebung, Gänserndorf, Graz, Graz-Umgebung, Hartberg-Fürstenfeld, Innsbruck, Innsbruck-Land, Korneuburg, Krems an der Donau, Krems-Land, Leibnitz, Linz, Linz-Land, Mödling, Neusiedl am See, Salzburg, Salzburg-Umgebung, Südoststeiermark, Wien, Wien-Umgebung, Wiener Neustadt and Wiener Neustadt-Land. For the other seven countries of the study area, the information was only available at the country level, where intensified sampling occurred in Hungary since 1995 and in Switzerland since 2005.

#### *Human population density*

Data on human population in and around a grid cell were also based on the History Database of the Global Environment (HYDE), Version 3.1 (Klein Goldewijk et al. 2010, Klein Goldewijk et al. 2011) with a spatial resolution of 5' and a historical time series at decennial intervals. Human population numbers for individual years (within the decennial intervals) were linearly interpolated. To account for human mobility, we used a distance-weighted accumulation of human population numbers from grid cells within a country. We assumed that in one-dimensional space the frequency of human movement is inversely proportional to distance. This relationship was then projected into two-dimensional space (grid cells as areal units) analogously to the two-dimensional kernel function for dispersal described in Electronic Appendix 4. As a result, weighting was proportional to the inverse of the squared distance between two cells. To obtain human population density, absolute human population numbers were divided by total terrestrial area of the individual grid cell.

#### *Scaling*

Climate, land use and human population raster data were matched to the resolution of the lattice system as follows.

For climate data we used a statistical downscaling approach (e.g. Zimmermann et al. 2007, Randin et al. 2009) against the WorldClim data set (<http://www.worldclim.org>) which provides long-term monthly mean values of temperature and precipitation for the period 1950–2000 at a fine-grained spatial resolution of 30" (ca.  $1 \times 1 \text{ km}^2$ ). In a first step, for both the 10' CRU and the 0.25° E-OBS data set we calculated long-term monthly mean values of temperature and precipitation for the 1950–2000 reference period of the WorldClim data set (the original CRU data are actually until 2000). Next, for every year we calculated monthly temperature and precipitation anomalies from these long-term monthly mean values, using absolute differences for temperature and relative differences for precipitation. These anomalies were then interpolated to the 30" resolution of the WorldClim data set using a two-dimensional minimum curvature spline technique as implemented in ESRI® ArcGIS 10.1 (Esri 2012), and added to (temperature) or multiplied with (precipitation) the 30" WorldClim values. Finally, we overlaid the downscaled 30" climate data with the 5'×3' raster layer of our lattice system and averaged the downscaled data for every 5'×3' grid cell and month.

For HYDE data, we first transformed the 5'×5' data resolution to a 1'×1' resolution (dividing values of cropland area, urban area and human population numbers by 25), and subsequently re-accumulated at the 5'×3' resolution of our lattice system.

GIS operations were conducted using ESRI® ArcGIS 10.1 (Esri 2012).

#### *Acknowledgments due to data licenses*

We acknowledge the E-OBS dataset from the EU-FP6 project ENSEMBLES (<http://ensembles-eu.metoffice.com>) and the data providers in the ECA&D project (<http://www.ecad.eu>).

#### Electronic Appendix 4. – Dispersal in two-dimensional space

Dispersal from invaded source cells to unoccupied recipient cells was modelled as isotropic dispersal in two-dimensional space dependent on the geographical distance between grid cells (see also Mang et al. (2017) and Mang et al. (2018)). To derive the kernel function for dispersal, we started with a leptokurtic, one-dimensional kernel function from the power-law family of form

$$f_{1D}(d_{i,j}) = d_{i,j}^{-\alpha} \quad (\text{E4.1})$$

as a base, where  $d_{i,j}$  is the distance between the centroids of cells  $i$  and  $j$ , and  $\alpha$  is a shape parameter (Portnoy & Willson 1993). Distances were calculated using the R (R Development Core Team 2011) function *geoDist* from package ‘SoDA’ (Chambers 2008) and measured in kilometres, at a precision of one decimal place. To match the dimensionality of the lattice system (dispersal to recipient cells as areal units), this base function was then projected into two-dimensional space and normalized to give

$$f_{2D}(d_{i,j}) = \int_{d_{i,j}-m}^{d_{i,j}+m} f_{1D}(x) dx \frac{1}{c} \frac{m}{4 d_{i,j}}, \quad (\text{E4.2})$$

where  $m$  is the average cell radius under a circular cell shape approximation (keeping cell area constant), and  $c$  is an (approximate) normalizing constant given by  $c = \int_{d_{\min}}^{d_{\max}} f_{1D}(x) dx$  with finite lower and upper dispersal bounds  $d_{\min}$  and  $d_{\max}$ , respectively. Due to computing constraints, dispersal was restricted to cell-pairs with  $d_{i,j} \leq 150$  km. For the normalization we hence used  $d_{\min} = m$  and  $d_{\max} = m + 150$ .

## Electronic Appendix 5. – Prior distributions and Markov chain Monte Carlo (MCMC)

Table E5.1. – Marginal prior distributions of invasion process parameters,  $\theta$ , and observation process parameters,  $\delta$ , and transformations used for Markov chain Monte Carlo (MCMC) sampling of these parameters. Prior distributions refer to the untransformed parameter scale and are assumed to be independent, hence the joint prior distributions of the hierarchical model and non-hierarchical model,  $p(\theta, \delta)$  and  $p(\theta)$ , respectively, are given by their product. For normally-distributed prior distributions the provided parameterizations state the mean and standard deviation, and for Gamma-distributed prior distributions the shape and scale.

Parameter	Prior distribution	Transformation
Invasion process:		
Environmental suitability		
$\beta_{\text{temperature}}$	N(0, 5)	identity
$\beta_{\text{precipitation}}$	N(0, 5)	identity
$\beta_{\text{cropland area}}$	N(0, 5)	identity
$\beta_{\text{urban area}}$	N(0, 5)	identity
$\beta_{\text{motorways}}$	N(0, 5)	identity
$\beta_{\text{railways}}$	N(0, 5)	identity
Dispersal, $\alpha$	Gamma(1.5, 100)	log
Source cells propagule production rate, $\eta$	Gamma(1.5, 100)	log
Background introduction rate, $\lambda$	Gamma(1.5, 100)	log
Background introduction start boost, $\lambda_b$	Gamma(1.5, 100)	log
Observation process:		
Detection rate, $\gamma$	Gamma(1.5, 100)	log
Detection dependence on invasion level, $\varphi$	Gamma(1.5, 100)	log
Sampling intensity		
$\rho_{\text{human population density}}$	N(0, 5)	identity
$\rho_{\text{pre-1970}}$	N(0, 5)	identity
$\rho_{\text{intensified sampling in Austria, Hungary \& Switzerland}}$	N(0, 5)	identity
$\rho_{\text{Germany}}$	N(0, 5)	identity
$\rho_{\text{Czech Republic, Slovakia \& Slovenia}}$	N(0, 5)	identity

For model fitting we used a Bayesian inference approach with Markov chain Monte Carlo (MCMC) (Gelman et al. 2004, Brooks et al. 2011) for parameter estimation (see also Mang et al. (2017) and Mang et al. (2018)). MCMC is a family of numerical methods that derives a sample of the posterior distribution by evolving a Markov chain. We used a Metropolis-Hastings sampling scheme in which a new parameter state is proposed based on the current parameter state, and subsequently either accepted (replacing the current parameter state) or rejected (retaining the current parameter state). In general notation, let  $\zeta^c$  be the current parameter state and  $\zeta^*$  be the proposed state, then the acceptance probability is:

$$A(\zeta^c \rightarrow \zeta^*) = \min\left(1, \frac{p(\zeta^* | \psi) q(\zeta^c | \zeta^*)}{p(\zeta^c | \psi) q(\zeta^* | \zeta^c)}\right), \quad (\text{E5.1})$$

where  $p(\zeta^* | \psi)$  and  $p(\zeta^c | \psi)$  are the posterior densities under the proposed parameter state and current parameter state, respectively (each proportional to an unknown normalizing constant which cancels out in their ratio),  $q(\zeta^* | \zeta^c)$  is the proposal density for the proposed parameter state given the current parameter state, and  $q(\zeta^c | \zeta^*)$  is the proposal density for the current parameter state given the proposed parameter state. Acceptance occurs unconditionally if the ratio on the right-hand side is  $\geq 1$ , and otherwise if it is greater than a drawn  $U(0,1)$  random variate (a uniform random variate). The Markov chain is evolved by repeating this process iteratively many times (MCMC iterations). Efficient proposal schemes have intermediate acceptance rates. For normally distributed proposal and posterior distributions it can be shown that an acceptance rate of approximately 44% is optimal for scalar parameter updates and 23.4% for multivariate updates of more than five

parameters, but sampling still remains efficient unless acceptance rates deviate considerably from these figures (Gelman et al. 2004, Brooks et al. 2011). The Metropolis-Hastings sampling scheme used in this study builds on the sampling scheme used in Mang et al. (2017) and was implemented as follows.

For sampling invasion process parameters,  $\theta$ , and observation process parameters,  $\delta$ , parameters were transformed to the real line (see Table E5.1 for transformation functions) and the posterior density was transformed via the applicable Jacobian term. Each parameter was updated as scalar, with the proposed value given by drawing from a (univariate) normal distribution centred at the current parameter value and the standard deviation chosen empirically to yield the desirable intermediate acceptance rate.

In the hierarchical model, the actual invasion times of grid cells,  $\mathbf{x}$ , are regarded as additional parameters of the posterior distribution; by contrast, for interval-censored dates the actual (precise) times of first detection in invaded cells,  $\mathbf{y}$ , are merely regarded as nuisance parameters. Consequently, both  $\mathbf{x}$  and  $\mathbf{y}$  were sampled during MCMC but values were stored permanently only for  $\mathbf{x}$ . In the following descriptions of sampling  $\mathbf{x}$  and  $\mathbf{y}$ , the model end time,  $t_e$ , refers to the last year used for MCMC fitting (2010; or, for models fitted for model validation, 2005) and only records up to  $t_e$  were considered. For sampling  $\mathbf{x}$ , each cell's invasion time was updated as scalar with the new time proposed as follows to yield good chain mixing: candidate years for proposal were all years up to 50 less than and greater than the current value (= year) of  $x_i$  subject to the constraints that (i) the current year  $x_i$  was not a candidate; (ii) no year before the model start time (= 1900) was proposed; (iii) for cells with a record available by  $t_e$  no year after the current value (= year) of  $y_i$  was proposed; (iv) for cells without such a record available any year greater than  $t_e$  was implicitly considered as the 'unoccupied at the model end time' state; and (v) if the current value of  $x_i$  represented this 'unoccupied at the model end time' state then only years less or equal than  $t_e$  were candidates. From all candidate years a new invasion time was then proposed by drawing from a uniform distribution. For sampling  $\mathbf{y}$ , updating was only necessary for cells for which the respective date intervals had a length of two or more years (i.e. cells with residual uncertainty regarding the year of first detection). For each of these cells, the actual (precise) time of first detection was updated as scalar with the new time proposed as follows to yield good chain mixing: candidate years for proposal were all years up to 50 less than and greater than the current value (= year) of  $y_i$  subject to the constraints that (i) the current year  $y_i$  was not a candidate; (ii) no year outside the cell's date interval,  $[z_{a_i}, z_{b_i}]$ , was proposed; and (iii) no year before the current value (= year) of  $x_i$  was proposed. From all candidate years a new actual (precise) time of first detection was then proposed by drawing from a uniform distribution.

For fitting the hierarchical model, all invasion process parameters, observation process parameters and parameters of cells' actual invasion times and actual (precise) times of first detection were updated once per iteration, with the update order newly randomized with each iteration. The non-hierarchical model used only the invasion process parameters and sampling of these parameters followed the same scheme as was described for the hierarchical model.

**Electronic Appendix 6.** – *Parameter estimates for models with different representations of the variation in climate over time*

Table E6.1. – Parameter estimates of the invasion process and observation process for modelling the spread of *Ambrosia artemisiifolia* in central Europe. In the hierarchical model, the observation process accounts for lagged and incomplete records of occurrence. The non-hierarchical model assumes that records accurately reflect the species' actual spread and hence does not use the observation process. Models in (A) use a linear trend in climate warming to represent the variation in climate over time, and models in (B) use the long-term climate average; models using an annual time series of temperature and precipitation values are presented in Table 2 in the main text. Estimates are stated as median (top row) and 95% (central) credible interval (bottom row) of the marginal posterior distributions. Significance tests apply only to parameters of environmental suitability and sampling intensity, with significant results marked by \*.

**A** – Linear trend in climate warming

Parameter	Hierarchical model	Non-hierarchical model
Invasion process:		
Environmental suitability		
$\beta_{\text{temperature}}$	0.55* (0.49, 0.61)	0.83* (0.79, 0.87)
$\beta_{\text{precipitation}}$	0.36* (0.32, 0.40)	0.20* (0.17, 0.22)
$\beta_{\text{cropland area}}$	0.01 (-0.03, 0.05)	-0.02 (-0.06, 0.03)
$\beta_{\text{urban area}}$	0.11* (0.07, 0.14)	0.11* (0.08, 0.15)
$\beta_{\text{motorways}}$	0.01 (-0.01, 0.04)	0.03* (0.01, 0.05)
$\beta_{\text{railways}}$	0.10* (0.07, 0.13)	0.10* (0.07, 0.14)
Dispersal, $\alpha$	1.78 (1.66, 1.91)	1.04 (0.97, 1.10)
Source cells propagule production rate, $\eta$	0.022 (0.019, 0.025)	0.018 (0.016, 0.019)
Background introduction rate, $\lambda$	0.00017 (0.00011, 0.00023)	0.00010 (0.00009, 0.00012)
Background introduction start boost, $\lambda_b$	0.0060 (0.0035, 0.0096)	0.000086 (0.000009, 0.000285)
Observation process:		
Detection rate, $\gamma$	0.0070 (0.0060, 0.0082)	
Detection dependence on invasion level, $\varphi$	0.54 (0.39, 0.69)	
Sampling intensity		
$\rho_{\text{human population density}}$	0.48* (0.39, 0.58)	
$\rho_{\text{pre-1970}}$	-0.74* (-0.96, -0.53)	
$\rho_{\text{intensified sampling in Austria, Hungary \& Switzerland}}$	4.14* (3.91, 4.39)	
$\rho_{\text{Germany}}$	0.41* (0.07, 0.77)	
$\rho_{\text{Czech Republic, Slovakia \& Slovenia}}$	0.76* (0.52, 0.98)	

**B** – Long-term climate average

Parameter	Hierarchical model	Non-hierarchical model
Invasion process:		
Environmental suitability		
$\beta_{\text{temperature}}$	0.54* (0.48, 0.60)	0.82* (0.78, 0.86)
$\beta_{\text{precipitation}}$	0.27* (0.23, 0.31)	0.23* (0.21, 0.26)
$\beta_{\text{cropland area}}$	0.00 (-0.04, 0.04)	0.03 (-0.02, 0.07)
$\beta_{\text{urban area}}$	0.11* (0.07, 0.15)	0.16* (0.12, 0.19)
$\beta_{\text{motorways}}$	0.03 (-0.00, 0.05)	0.02 (-0.00, 0.05)
$\beta_{\text{railways}}$	0.09* (0.06, 0.13)	0.08* (0.05, 0.11)
Dispersal, $\alpha$	1.75 (1.63, 1.88)	0.93 (0.85, 0.99)
Source cells propagule production rate, $\eta$	0.026 (0.022, 0.030)	0.023 (0.021, 0.025)
Background introduction rate, $\lambda$	0.00015 (0.00010, 0.00021)	0.000076 (0.000062, 0.000092)
Background introduction start boost, $\lambda_b$	0.0033 (0.0015, 0.0056)	0.000076 (0.000009, 0.000248)
Observation process:		
Detection rate, $\gamma$	0.0082 (0.0071, 0.0097)	
Detection dependence on invasion level, $\varphi$	0.19 (0.04, 0.35)	
Sampling intensity		
$\rho_{\text{human population density}}$	0.49* (0.38, 0.59)	
$\rho_{\text{pre-1970}}$	-0.62* (-0.84, -0.38)	
$\rho_{\text{intensified sampling in Austria, Hungary \& Switzerland}}$	4.02* (3.83, 4.21)	
$\rho_{\text{Germany}}$	0.13 (-0.25, 0.54)	
$\rho_{\text{Czech Republic, Slovakia \& Slovenia}}$	0.95* (0.70, 1.21)	

**Electronic Appendix 7.** – *Kernel function for dispersal according to the non-hierarchical model*

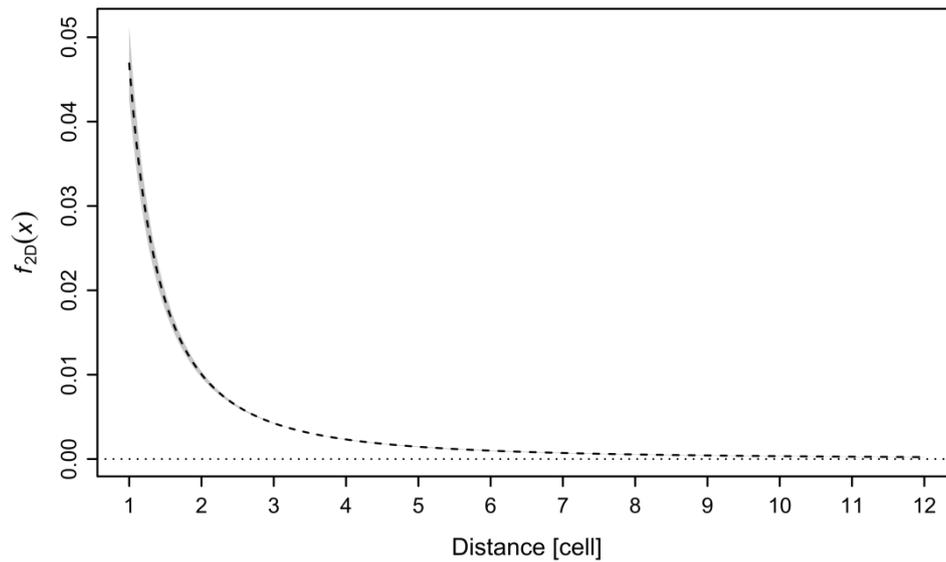


Fig. E7.1. – The kernel function for dispersal from invaded source cells to unoccupied recipient cells in a two-dimensional lattice system according to the non-hierarchical model of the spread of *Ambrosia artemisiifolia* in central Europe. The dashed line shows the function for the median of the marginal posterior distribution of the shape parameter  $\alpha$  (see Table 2 in the main text). The curve is very similar for the 95% credible interval of  $\alpha$ , which is shown as the shaded area in the background of the dashed line.

## References

- Arens T., Hettlich F., Karpfinger C., Kockelkorn U., Lichtenegger K. & Stachel H. (2009): Ergänzungen und Vertiefungen zu Arens et al., Mathematik. – Spektrum Akademischer Verlag, Heidelberg.
- Bolt J. & van Zanden J. L. (2014): The Maddison Project: collaborative research on historical national accounts. – *Econ. Hist. Rev.* 67: 627–651.
- Brooks S., Gelman A., Jones G. L. & Meng X.-L. (2011): Handbook of Markov chain Monte Carlo. – Chapman & Hall/CRC, Boca Raton.
- Büttner G., Feranec J. & Jaffrain G. (2002): Corine land cover update 2000: technical guidelines. – European Environment Agency, Copenhagen.
- Chambers J. M. (2008): SoDA: functions and examples for "Software for Data Analysis", version 1.0-3. – R Foundation for Statistical Computing, Vienna.
- Esri (2012): ArcGIS Desktop 10.1, version 10.1. – Esri, Redlands.
- Essl F., Dullinger S. & Kleinbauer I. (2009): Changes in the spatio-temporal patterns and habitat preferences of *Ambrosia artemisiifolia* during its invasion of Austria. – *Preslia* 81: 119–133.
- Gelman A., Carlin J. B., Stern H. S. & Rubin D. B. (2004): Bayesian data analysis. Ed. 2. – Chapman and Hall/CRC, Boca Raton.
- Haylock M. R., Hofstra N., Klein Tank A. M. G., Klok E. J., Jones P. D. & New M. (2008): A European daily high-resolution gridded data set of surface temperature and precipitation for 1950–2006. – *J. Geophys. Res. Atmos.* 113: 1–12.
- Klein Goldewijk K., Beusen A. & Janssen P. (2010): Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. – *Holocene* 20: 565–573.
- Klein Goldewijk K., Beusen A., van Drecht G. & de Vos M. (2011): The HYDE 3.1 spatially explicit database of human-induced global land-use change over the past 12,000 years. – *Glob. Ecol. Biogeogr.* 20: 73–86.
- Mang T., Essl F., Moser D., Karrer G., Kleinbauer I. & Dullinger S. (2017): Accounting for imperfect observation and estimating true species distributions in modelling biological invasions. – *Ecography* 40: 1187–1197.
- Mang T., Essl F., Moser D., Kleinbauer I. & Dullinger S. (2018): An integrated, spatio-temporal modelling framework for analysing biological invasions. – *Divers. Distrib.* (in press, doi: 10.1111/ddi.12707).
- Mitchell T. D., Carter T. R., Jones P. D., Hulme M. & New M. (2004): A comprehensive set of high-resolution grids of monthly climate for Europe and the globe: the observed record (1901-2000) and 16 scenarios (2001-2100). – Tyndall Centre for Climate Change Research.
- Niklfeld H. (1998): Mapping the flora of Austria and the eastern Alps. – *Revue Valdôtaine d'Histoire Naturelle* 51: 53–62.
- OpenStreetMap contributors (2011): OpenStreetMap. – URL: <http://download.geofabrik.de>.
- Portnoy S. & Willson M. F. (1993): Seed dispersal curves: behavior of the tail of the distribution. – *Evol. Ecol.* 7: 25–44.
- R Development Core Team (2011): R: a language and environment for statistical computing, version 2.14.1. – R Foundation for Statistical Computing, Vienna.
- Randin C. F., Engler R., Normand S., Zappa M., Zimmermann N. E., Pearman P. B., Vittoz P., Thuiller W. & Guisan A. (2009): Climate change and plant distribution: local models predict high-elevation persistence. – *Glob. Change Biol.* 15: 1557–1569.
- Richter R., Berger U. E., Dullinger S., Essl F., Leitner M., Smith M. & Vogl G. (2013): Spread of invasive ragweed: climate change, management and how to reduce allergy costs. – *J. Appl. Ecol.* 50: 1422–1430.
- Zimmermann N. E., Edwards Jr T. C., Moisen G. G., Frescino T. S. & Blackard J. A. (2007): Remote sensing-based predictors improve distribution models of rare, early successional and broadleaf tree species in Utah. – *J. Appl. Ecol.* 44: 1057–1067.