Supplementary Material S1: Theory for "A theoretical framework for upscaling species distribution models"

Christine N. Meynard, Cyril Piou, David M. Kaplan

Contents

S1.1 Baseline theory for upscaling effects

Building off of the theory developed in Meynard et al. (2019), we want to understand how probability of species occurrence changes as a function of how we aggregate sampling (and environmental covariate) data. Suppose that we have a set of sites, $\{i\}$, at which fieldwork is carried out to assess species presence-absence. We assume that species presence at a given site i is a binomial stochastic process with probability

of occurrence (i.e., presence), p_i, being a function of one or more environmental covariates. Implicitly in this formulation we assume that site i being occupied has no impact on the occurrence in nearby site j. This means that individual sites are far enough apart (or big enough) that we are beyond the scale of nearest-neighbor colonization events so that occurrence at one site is independent of occurrence at all other sites (*note:* given this assumption, probability of occurrence *can* have spatial auto-correlation due to spatial auto-correlation in environmental covariates, it is just occurrence itself that must be a perfect stochastic process of the environmentally-determined probability of occurrence). Later we reflect on what would happen if there was correlation in occurrence between sites due to colonization events.

Now suppose that we want to group presence-absence observations at several such sites, e.g., to match the spatial scales of environmental data. There are at least 3 approaches to doing this (also see Fig. 1 in the paper):

1) One could randomly sample one of the sub-sites within an aggregate and declare the aggregate to be a presence if that randomly sampled site was a presence (Fig. 1a).
    - This is equivalent to what would happen if one had a fixed amount of sampling effort to determine if a species is present in a small subset of all possible sites. The set of presences and absences that one observes could be associated with (environmental) covariate information on different spatial scales, but increasing or decreasing this scale will not change one's declaration of the corresponding site as a presence or an absence because there is essentially never more than one site sampled per aggregate.
    - An aggregate being considered a presence would then just mean that the species was present at the (smaller) sampled site. Nevertheless, one would model as if the species was present throughout the grid cell. If environmental covariates represent an average over the entire aggregate and environment is not constant throughout the aggregated sites, then differences between average environmental conditions and the specific conditions at the sampled site within an aggregate could produce variability or bias in our estimates of species environment-occurrence relationships.
2) Another possibility is to declare an aggregate a presence if the species is present in any sub-site within the aggregate (Fig. 1b).
    - This is what would happen if one was aggregating species observations over space (or time), so that sampling effort increases proportional to the size of the aggregate.
    - This approach is exactly equivalent to declaring an absence if and only if the species is absent from all sub-sites within an aggregate.

3) A final possibility is to declare an aggregate as a presence if and only if all sites within the aggregate are occupied (Fig. 1c).

- This unlikely choice is essentially the inverse of approach #2, and, therefore, it has the same theoretical behavior (but in an inverse sense) as approach #2. As such, we will present primarily approach #2 and only briefly mention approach #3.

If there is very strong spatial autocorrelation among observations of occurrence due, e.g., to nearest-neighbor colonization and extinction processes, then species presence in any one site in an aggregate will guarantee presence in all sites of an aggregate. In this case, assessments of species presence based on a single site (i.e., approach #1) are no different than assessments based on all sites in the aggregate and approach #2 reduces to approach #1. Therefore, approaches #1 and #2 essentially bound the possibilities of what will happen if local species nearest-neighbor colonization and extinction processes are important or negligible, respectively, on the scale of individual sites.

For all three approaches described above, we would like to derive the probability of "occurrence" in an aggregate based on the probabilities of occurrence of the sites (or observations) that are inside the aggregate. Initially we will assume that environmental conditions are constant within each aggregate (or, in an approximate sense, that the scale of aggregation is considerably inferior to the spatial autocorrelation scales of environmental variability). For this case, one can develop a relatively simple set of analytic equations describing what will happen to probability of occurrence as a function of level of aggregation. Once we understand this theory, we will then turn our attention to how environmental variability within aggregates will impact bias and variability in estimated environment-occurrence relationships.

Approach #1 is relatively simple to analyze, so we will develop it first before proceeding to the more complex Approach #2.

S1.1.1 Approach #1

Approach #1 will achieve on an average an unbiased estimate of probability of occurrence within an aggregate. If single sites are randomly drawn from within an aggregate many times, the average percent occurrence of the sites drawn (i.e., the average of the zeros representing absences and ones representing presences) will just reflect the percentage of presences within the aggregate. As the percentage of presences within an aggregate is simply a reflection of the mean (environmentally-determined) probability of occurrence of all sites within the aggregate, on average over many such upscalings the perceived probability of occurrence will

simply be the mean of the probabilities of occurrence of the sites that compose the aggregate. If we assume that environmental conditions are the same for all sites within the aggregate, then probability of occurrence will be the same for any site within the aggregate and the probability of occurrence of the aggregate based on approach #1 will simply be the probability of occurrence for any individual site within the aggregate. As such, the environment-occurrence relationship will be unchanged by the aggregation process.

However, the observed relationship between average environmental conditions over the aggregate and average probability of occurrence over the aggregate may differ from the underlying (site-level) environment-occurrence relationship if environment varies within an aggregate. Differences will potentially be important if the environment-occurrence relationship is a non-linear function of environment over the range of conditions within an aggregate. This possibility will be considered in more detail later on in the document.

One might think that upscaling would also increase the variance (i.e., RMS difference) between the observed aggregate presence-absence map and the aggregated probability of occurrence because randomly sampling a single cell may in some cases very poorly reflect the average probability of occurrence (e.g., by randomly selecting the one cell in an aggregate that is actually occupied), but this will not on average be the case as every site is an equally valid potential presence and the probability of occurrence of that site contributes equally to the average probability of occurrence over the aggregate. An increase in discrepancy between the presence-absence map and predicted probability of occurrence would be entirely due in this case to the reduced number of "sites" after the aggregation process.

S1.1.2 Approach #2

For approach #2, the upscaled probability of occurrence in the aggregate is the probability that at least one of the sampled sites within the aggregate is a presence. This is the same as one minus the probability that none of the sites in an aggregate are occupied. As the probability that none of the sites in an aggregate is occupied is the probability that site 1 is an absence *AND* site 2 is an absence *AND* site 3 is an absence..., this can be calculated as the product of the probabilities of absence in each site:

$$\tilde{p} = 1 - \prod_{i=1}^{\nu} (1 - p_i) \tag{1}$$

where ν is the number of sites in the aggregate. Note that this reduces to $\tilde{p} = p$ for the case $\nu = 1$, as one would expect.

If environment is constant over all sites within an aggregate (i.e., $p_i = p$), then the aggregate probability of

occurrence, $\tilde{p}$, becomes:

$$\tilde{p} = 1 - (1 - p)^{\nu} \tag{2}$$

Note that this equation can be inverted to give the probability of occurrence in a single site as a function of the probability of occurrence of the aggregate:

$$p = 1 - (1 - \tilde{p})^{1/\nu} \tag{3}$$

S1.1.2.1 Case of a logistic probability of occurrence

One special case of the logic above for constant environment within aggregates for which a number of intuitive analytic results can be obtained is when probability of occurrence is a logistic function of a single environmental variable, x:

$$p(x) = \frac{1}{1 + e^{(x-\beta)/\alpha}} \tag{4}$$

This formulation is reasonably representative of any one-sided environmental gradient in probability of occurrence. If $\alpha > 0$, then probability of occurrence will approach one for $x << 0$ and it will approach zero for $x >> 0$. The inflection point of the logistic curve occurs when $x = \beta$ and $p = \frac{1}{2}$. At this inflection point, the slope of the environment-occurrence relationship is $-\frac{1}{4\alpha}$.

We would like to understand the behavior of this type of environment-occurrence relationship as we aggregate probability of occurrence over multiple sites following approach #2. In particular, we would like to understand at least three things:

1) How does the environmental point at which $\tilde{p} = \frac{1}{2}$ change as a function of the level of aggregation?
2) How does the inflection point change as a function of level of aggregation?
3) How does the steepness at the inflection point change as a function of level of aggregation? Does the environment-occurrence relationship become more threshold-like as level of aggregation is increased?

In order to develop analytic relationships that answer these questions, it is useful to first note that the logistic function can be inverted to get the value of the environment x corresponding to a probability of occurrence p:

$$x = \beta + \alpha \left[\log(1-p) - \log(p)\right] \tag{5}$$

This, in combination with Eq. 3 can be used to get the environmental values corresponding to a value of the theoretical upscaled probability of occurrence:

$$x = \beta + \alpha \left[\frac{1}{\nu}\log(1-\tilde{p}) - \log\left(1-(1-\tilde{p})^{1/\nu}\right)\right] \tag{6}$$

This can also be written:

$$x = \beta - \alpha \log\left((1-\tilde{p})^{-1/\nu} - 1\right)$$

This equation can be useful to look at what environmental conditions produce a given value of aggregated probability of occurrence as a function of ν. In particular, we can look at what value of environment produces $\tilde{p} = \frac{1}{2}$. Substituting this into Eq. 6, we have:

$$x_{\nu,0.5} = \beta - \alpha \log\left(2^{1/\nu} - 1\right) \tag{7}$$

First, note that as $1 \geq 2^{1/(\nu-1)} - 1 > 2^{1/(\nu)} - 1 > \cdots > 0$ for all $\nu > 1$, we have that $x_{\nu,0.5} > x_{\nu-1,0.5} \geq \beta$, so the environmental value at which probability of occurrence is 50% moves continuously to the right as the level of aggregation (i.e., ν) increases. For large ν, we can use the exponential expansion to estimate this point:

$$x_{\nu,0.5} \approx \beta - \alpha \log\left(1 + \frac{1}{\nu}\log 2 - 1\right) = \beta + \alpha \log\left(\frac{\nu}{\log 2}\right)$$

So the position of this point moves rightward approximately at the rate of $\alpha \log \nu$.

Next, to examine the inflection point and the steepness at the inflection point, we need to look at the derivatives of $\tilde{p}(x)$. The inflection of this function is defined as the point at which the second derivative with respect to x is zero. We begin with derivatives of the unaggregated logistic function:

$$\frac{dp}{dx} = \frac{-1}{\alpha\left(1+e^{(x-\beta)/\alpha}\right)^2} e^{(x-\beta)/\alpha} = -\frac{p^2}{\alpha} e^{(x-\beta)/\alpha} = -\frac{1}{\alpha} p(1-p) \tag{8}$$

and

$$\frac{d^2p}{dx^2} = -\frac{1}{\alpha}\left[(1-p)\frac{dp}{dx} - p\frac{dp}{dx}\right] = \frac{1}{\alpha^2}p(1-p)(1-2p) \tag{9}$$

From this final equation, it is easy to see that the inflection point of the logistic occurs when $p = \frac{1}{2}$, which occurs when $x = \beta$.

We can also look at the derivatives of $\tilde{p}$ to identify it's inflection points:

$$\frac{d\tilde{p}}{dx} = \nu(1-p)^{\nu-1}\frac{dp}{dx}$$

For a logistic probability of occurrence, this gives:

$$\frac{d\tilde{p}}{dx} = -\frac{\nu}{\alpha}p(1-p)^{\nu} \tag{10}$$

The second derivative of $\tilde{p}$ is:

$$\frac{d^2\tilde{p}}{dx^2} = \nu(1-p)^{\nu-1}\frac{d^2p}{dx^2} - \nu(\nu-1)(1-p)^{\nu-2}\left(\frac{dp}{dx}\right)^2$$

We can substitute the derivatives of the logistic (Eqs. 8 & 9) into this formula to get a precise location for the inflection point for this specific case:

$$\frac{d^2\tilde{p}}{dx^2} = \frac{\nu}{\alpha^2}p(1-p)^{\nu}(1-2p) - \frac{\nu}{\alpha^2}(\nu-1)p^2(1-p)^{\nu}$$

$$\frac{d^2\tilde{p}}{dx^2} = \frac{\nu}{\alpha^2}p(1-p)^{\nu}\left[1-2p-(\nu-1)p\right] = \frac{\nu}{\alpha^2}p(1-p)^{\nu}\left[1-(\nu+1)p\right]$$

From this we see that the inflection point given by $\frac{d^2\tilde{p}}{dx^2} = 0$ occurs when:

$$p_{\nu,\text{infl}} = \frac{1}{\nu+1} \tag{11}$$

Note that this is the probability of occurrence of the original, unaggregated logistic function. The probability of occurrence in the aggregated environment-occurrence relationship can be derived from this using Eq. 2:

$$\tilde{p}_{\nu,\text{infl}} = 1 - \left(1 - \frac{1}{\nu+1}\right)^{\nu} = 1 - \left(\frac{\nu}{\nu+1}\right)^{\nu} = 1 - \frac{1}{\left(1+\frac{1}{\nu}\right)^{\nu}} \tag{12}$$

Using the well-known limit equation:

$$\lim_{\nu\to\infty}\left(1+\frac{1}{\nu}\right)^{\nu} = e \tag{13}$$

we find that for large ν, Eq. 12 approaches:

$$\tilde{p}_{\nu,\text{infl}} \approx 1 - \frac{1}{e} \approx 0.632 \tag{14}$$

Therefore, the probability of occurrence at which the inflection point occurs asymptotes to 0.632.

This gives the location of the inflection point as a function of the original, unaggregated logistic probability of occurrence. Substituting this into Eq. 5 gives the value of the environmental variable at which the inflection occurs:

$$x_{\nu,\text{infl}} = \beta + \alpha\left[\log\left(\frac{\nu}{\nu+1}\right) - \log\left(\frac{1}{\nu+1}\right)\right] = \beta + \alpha\log(\nu) \tag{15}$$

Eq. 15 indicates that the position of the inflection point moves rightward at precisely the rate $\alpha\log(\nu)$. For large ν, this is approximately the same rate at which the point of $\tilde{p} = \frac{1}{2}$ (Eq. 7) moves to the right, so these two remain at an approximately fixed distance from each other in environmental variable space.

One can also evaluate the steepness at the inflection point using Eq. 10:

$$\left.\frac{d\tilde{p}}{dx}\right|_{x_{\nu,\text{infl}}} = -\frac{\nu}{\alpha}\frac{1}{\nu+1}\left(\frac{\nu}{\nu+1}\right)^{\nu} = -\frac{1}{\alpha}\left(\frac{\nu}{\nu+1}\right)^{\nu+1} \tag{16}$$

Again using Eq. 13, we find that for large ν, Eq. 16 approaches:

$$\left.\frac{d\tilde{p}}{dx}\right|_{x_{\nu,\text{infl}}} \approx -\frac{1}{e\alpha} \tag{17}$$

Therefore, the steepness of the aggregate environment-occurrence relationship does not increase indefinitely, but rather asymptotes toward $\frac{4}{e} \approx 1.472$ times the steepness of the original, unaggregated logistic species-environment relationship.

Eqs. 7, 11, 15, 12, 16 and 17 form the core theoretical results of this analysis regarding the effects of aggregating probabilities of occurrence for sets of sites with similar environmental conditions. They indicate that aggregation moves the center of the transition zone of the probability of occurrence distribution towards more and more marginal habitat areas. This process continues indefinitely as the log of the level of aggregation (in environmental variable space; in real space, movement of the "presence" frontier will depend on the relationship between physical space and environmental conditions; Fig. 2). The steepness of the transition does not, however, increase indefinitely towards a threshold species-environment relationship, but rather asymptotes towards a fixed value no matter what the level of aggregation.

S1.1.2.2 Impact of aggregation on model performance statistics

One particularly important question regarding the impact of spatial aggregation on SDMs is how indicators of model performance (e.g., AUC, TSS, proportion of correctly classified data, sensitivity and specificity) will vary as a function of spatial scale, with it generally being assumed that higher resolution models will have better performance (Mertes & Jetz 2018). When aggregation is carried out following approach #2, whether or not model performance indicators will increase or decrease as a function of the spatial scale of aggregation depends on a number of factors. One is the particular modeling approach used. For example, if one uses logistic regression, then this will fit less well the aggregated data than the unaggregated data as the aggregated data no longer has a true logistic environment-occurrence relationship. However, as we will see when we examine below numerically the theory developed above, divergence from a logistic relationship is generally relatively minor. Furthermore, true environment-occurrence relationships in nature will never be exactly logistic regardless of spatial scale, and this is issue is easily addressed by using a more flexible, non-linear modeling approach, such as general additive models (GAMs) and random forests.

Likely a more important effect is the shape of the environment-occurrence relationship itself. Model performance is essentially determined by areas for which environmental conditions lead to probabilities of occurrence that are far from zero or one. In all other areas, models will have little problem correctly determining occurrence as the species will be (almost) always present or always absent. As aggregation using approach #2 increases the slope of the environment-occurrence relationship, this will have a tendency to reduce the area over which probabilities of occurrence are intermediate, thereby increasing model performance as level of aggregation is increased. However, our analysis shows that the slope (in environmental space) does not increase indefinitely, but rather asymptotes towards a maximum value that is not strongly different from the initial slope. As such, the effect of increased aggregation may be relatively minor.

Perhaps the most important impact of aggregation on model performance indicators is via the range of

environmental conditions for which probabilities of occurrence are intermediate. As aggregation is increased, the environmental conditions over which probability of occurrence is rapidly varying will shift (to the right in our logistic formulation provided that $\alpha > 0$). Depending on how prevalent these environmental conditions are over space, this could lead to significant changes in model performance. For example, if the range of environmental conditions for which the unaggregated data has intermediate occurrence are far more prevalent in space than the range of environmental conditions for which the aggregated data has intermediate occurrence then model performance will likely increase as data are aggregated. On the other hand, if the reverse is true, model performance will likely decrease as data are aggregated.

S1.1.3 Approach #3

As previously mentioned, approach #3 is conceptually similar to approach #2, but in an opposite sense. As such, all equations developed in the previous section are valid for approach #3 so long as one replaces p by $1 - p$ and $\tilde{p}$ by $1 - \tilde{p}$. As a result, aggregation continually lowers the probability of occurrence, moving the transition zone of the environment-occurrence relationship to the left with the same rate and properties (e.g., slope) as developed in the previous section.

S1.2 Visualization of theory

Here we visually explore the analytic results developed above for approach #2. The other two approaches will not be given further consideration as approach #3 is conceptually equivalent to (but opposite) approach #2, and approach #1 does not modify the underlying environment-occurrence relationship. This visualization will be based purely on the analytic equations developed above and the relationship between physical space and environment will be ignored (i.e., we assume that environmental conditions do not vary within aggregates at the scales of aggregation examined). Numerical tests of the theory involving real space-environment relationships will be carried out in Section S1.3 and in Supplementary Material S2.

S1.2.1 Logistic probability of occurence

We begin by examining a logistic functional relationship between an environmental variable, in this case taken without loss of generality to be (mean annual) temperature in °C, and the probability of occurrence of a virtual species. The logistic curve representing species probability of occurrence as a function of environment follows Eq. (4) with $\beta = 7.5$ and $\alpha = 0.3$. As mentioned above, the relationship between environment and

physical, habitat space is not relevant as we assume that environment does not vary within aggregates at the scales of aggregation considered. We will consider aggregation in a two-dimensional space at the following scales: 1x1, 2x2, 4x4, 8x8, 16x16 (corresponding to $\nu = 1, 4, 16, 64, 256$).

For these parameter values, the unaggregated probability of occurrence is very close to one in areas with temperatures $< 6°C$ and very close to zero for temperatures $> 9°C$ with intermediate probabilities of occurrence between these two temperatures (Fig. S1.1). As level of aggregation is increased, the inflection point of the probability of occurrence curve displaces to the right (Fig. S1.1), as does the range of temperatures for which probability of occurrence is significantly different from zero and one (Fig. S1.2), at a rate proportional to $log(\nu) = log(N^2)$ (Fig. S1.3). Though the range of temperatures for which probabilities of occurrence are intermediate shrinks as data are aggregated, this shrinkage more or less reaches the asymptote for aggregation above a scale of 8x8 (Fig. S1.4).

S1.2.2 Normally distributed probability of occurence

Though a logistic curve is representative of many observed gradients in species occurrence, it is a one-sided gradient that is ultimately unrealistic for species that have lower *and* upper bounds on suitable environmental conditions. To explore the impact of aggregation on a two-sided environment-occurrence relationship, we consider a normally distributed probability of occurrence:

$$p(x) = Ce^{-(x-\mu)^2/\sigma^2} \tag{18}$$

where the maximum probability of occurrence, C, can be chosen to ensure a certain overall prevalence level over the habitat domain.

Although analytic results may be possible for species occurrence following a normal distribution, for simplicity we choose to numerically examine changes in the environment-occurrence relationship as a result of aggregation based on the combination of Eq. (18) and Eq. (2). We consider two virtual species with normally distributed environment-occurrence relationships differing in their level of overall habitat occurrence. For both species, $\mu = 0$ and $\sigma = 1$, but for the high prevalence species $C = 1$, whereas for the low occurrence species $C = 0.5$

For both the high (Fig. S1.5) and the low (Fig. S1.6) prevalence virtual species, aggregation has approximately the effect that one would expect from a two-sided logistic curve. On both sides of the environment-occurrence curve, aggregation has the effect of displacing the perceived core area of occurrence outward at

a decreasing, approximately-logarithmic rate. The range of environmental conditions over which probability of occurrence is intermediate decreases as aggregation increases, but this decrease appears to stabalize after a certain level of aggregation (e.g., the maximum and minimum slopes on each side of the environment-occurrence distribution is approximately the same for the blue and cyan curves in Fig. S1.5) as it would for a logistic environment-occurrence relationship.

These results suggest that analytical results for a logistic environment-occurrence relationship are approximately applicable to many other functional forms for the distribution of occurrence in environmental space.

S1.3 Numerical exploration of hypothetical case study and model performance indicators

Whereas the visualization above provides a basic demonstration of the theory, to understand how it might work in a real scenario and to examine the impact of aggregation on SDM model performance indicators, it is useful to consider hypothetical case studies based on virtual species distributed in a two-dimensional space. The habitat area is taken to be a square with 160 units on each side, for a total of 25,600 unaggregated grid cells. The spatial units of the grid cells (i.e., their physical size) have no impact on results, but so as to talk in concrete terms we can take them to be km.

We consider three different potential relationships between space and environmental conditions, here taken without loss of generality to be mean annual temperature in each grid cell (Fig. S1.7). In all three cases, mean annual temperature in the grid cells varies in the "latitudinal" (i.e., y or vertical) dimension from 4.5°C at the bottom of the domain to 10.5°C at the top of the domain. For the first relationship, temperature increases linearly (i.e., at a fixed rate) over the spatial domain (red curve in Fig. S1.7). In the other two cases, temperature increases at varying rates: the green curve in Fig. S1.7 has temperature values approximately constant and around the inflection point of the virtual species (i.e., 7.5°C) for a wide range of y values (i.e., from $y \approx 50$ to $y \approx 100$), whereas the blue curve in Fig. S1.7 changes quickly on the y axis for temperature values close to the inflection point of the virtual species. These differences in rates of change in the temperature gradient for the three types of landscapes mean that for the slow rate of change scenario of the green curve, the environments for which the probability of occurrence is intermediate are relatively common, whereas intermediate probabilities of occurrence are rare in the high-rate of change landscape given by the blue curve. As such, the space-environment relationship given by the green curve will at times be referred to as the "common scenario" and the space-environment relationship given by the blue curve will

at times be referred to as the "rare scenario" (the red curve being referred to as the "linear scenario).

In all cases, the environment-occurrence relationship is take to be a logistic function of temperature with the same form as that used in the previous section on visualization of analytic results: $\beta = 7.5$ and $\alpha = 0.3$. The combination of the space-environment relationships (e.g., the top panel in Fig. S1.8 presents the linear space-environment relationship) and the environment-occurrence relationship yields the spatial distributions of probability of occurrence (e.g., the middle panel in Fig. S1.8 shows that for the linear space-environment relationship). These distributions can be used to simulate potential distributions of presence-absence for the species by drawing one random number between zero and one for each grid cell of the model domain and comparing it with the probability of occurrence for that grid cell (Meynard et al. 2019), one example of which is shown in the bottom panel of Fig. S1.8. 80 such potential distributions of presence-absence were randomly drawn. These were used as the basis for developing SDMs for each of the 80 presence-absence maps after aggregation using approach #2 at a variety of spatial scales ranging from 1x1 (i.e., no aggregation) to 16x16. SDMs were estimated using binomial GAM models for which probability of occurrence was assumed to be a simple smooth of mean annual temperature. From the GAM SDMs, deviance explained was extracted, predictions of probability of occurrence were calculated and these were used to estimate the area under the ROC curve (i.e., AUC).

Results from this process demonstrate how SDM performance and model estimated probability of occurrence behave as a function of scale of aggregation and the spatial distribution of environmental conditions. For the linear space-environment relationship (top panel in Fig. S1.10), aggregation initially improves AUC before it stabilizes and begins to decrease slightly on average. As scale of aggregation increases, so does variability in performance between realizations of presence-absence (i.e., large size of boxplots in Fig. S1.10).

The initial increase in presence-absence classification rates (as measured by AUC) as a function of aggregation can be explained by the decrease in the range of temperatures yielding intermediate probabilities of occurrence as a result of aggregation (Fig. S1.4). The increased variability in performance as a function of scale of aggregation can be explained by the decreasing number of grid cells post aggregation (Fig. S1.9) leading to larger random swings in the clarity of the relationship between presence-absence and temperature.

Perhaps the most intriguing pattern is the decrease in AUC for larger scales of aggregation. This can be explained by the fact that increasingly large aggregates encompass a wider range of temperatures, thereby leading to mean environmental conditions over aggregates being a relatively poor indicator of the aggregate probability of occurrence (i.e., Eq. (1)). This can be seen by considering that Eq. (1) can be rewritten as:

$$\tilde{p} = 1 - G^{\nu}$$

where G is the geometric mean of the individual probabilities of *absence* (i.e., one minus probability of occurrence) that compose the aggregate:

$$G = \left[\prod_{i=1}^{\nu}(1 - p_i)\right]^{1/\nu}$$

The discrepancy between the arithmetic mean (of environmental conditions that then are used to calculate a probability of occurrence at the mean environmental conditions) and the geometric mean (of the true probabilities of occurrence) grows as a function of both the range and the number of values within the mean, leading to decreased model performance for large aggregates encompassing a wide range of temperatures. Indeed, at the 16x16 scale of aggregation, the temperature range within an aggregate is 0.6°C, which is not small with respect to the range of temperatures leading to intermediate probabilities of occurrence (Fig. S1.4). This theory based on geometric means and arithmetic means also explains why model-estimated and Eq. (1)-estimated probabilities of occurrence generally exceed probabilities of occurrence based on theory ignoring environmental heterogeneity within aggregates (Fig. S1.11) because the geometric mean is always inferior to the arithmetic mean, meaning that:

$$\tilde{p} = 1 - G^{\nu} > 1 - (1 - \bar{p})^{\nu} = \tilde{\bar{p}}$$

where $\bar{p}$ is the probability of occurrence for the mean environmental conditions of the aggregate.

When temperatures corresponding to intermediate probabilities of occurrence are relatively common in space (green curve in Fig. S1.7), unaggregated model performance indicators (left-hand side of middle panel in Fig. S1.10) are far lower than those for a linear space-environment relationship. Aggregation at the 4x4 scale increases performance indicators to approximately the level of those for the linear space-environment relationship and performance indicators at the 16x16 scale exceed those of the linear space-environment relationship. The opposite tendencies are true when temperatures for which probability of occurrence are intermediate are relatively rare over space (bottom panel in Fig. S1.10); in this case, performance indicators begin higher than those for the linear relationship, increase less with aggregation and decrease, in a relative sense, more at large scales of aggregation. These results can be explained by the change in temperatures that produce intermediate probabilities of occurrence as a function of scale of aggregation (Fig. S1.3). For the

space-environment relationship given by the green curve in Fig. S1.7, aggregation moves the temperatures that produce intermediate probabilities of occurrence from those that are relatively common over space to those that are relatively rare over space, causing a relative increase in performance indices (middle panel of Fig. S1.10). The opposite occurs for the space-environment relationship given by the blue curve in Fig. S1.7 (bottom panel of Fig. S1.10).

Theory ignoring heterogeneity in unaggregated grid cell probability of occurrence driven by variability in environmental conditions within aggregates represents reasonably well SDM estimates of probability of occurrence in all cases, however, discrepancies increase as scale of aggregation increases (Fig. S1.11). This can be explained by the aforementioned differences between arithmetic and geometric means and the increasing "pixelization" of the model system. This effect is particularly visible at the 16x16 scale (bottom panel of Fig. S1.11), for which the small number of grid cells leads to regular bumps in mean model-estimated probabilities of occurrence as a function of how many cells are randomly classified as presences in the model domain. This discretization also explains discrepancies between the application of Eq. (1) to the underlying probability of occurrence (dots in Fig. S1.11) and GAM model predictions (colored curves in Fig. S1.11) as the finite number of cells only permits certain particular levels of species prevalence (for a given latitudinal level).

S1.4 Limits to the applicability of theoretical results

There are a number of limits to the applicability of the theoretical results for approach #2 developed above. One was explored in the previous section, namely the impact of environmental heterogeneity within spatial aggregates on the applicability of theoretical results assuming constant environmental conditions. However, there are at least two other limits that are worth considering. Key to the results above for aggregation approach #2 are the infinite tales of the probability distribution. For a logistic or normal probability of occurrence, there are no areas for which probability of occurrence is truly zero. In real cases, there may, however, be situations for which environmental conditions are truly inhospitable to a species and the probability of occurrence is identically zero (e.g., alligators in the Arctic). More generally, as spatial scale of aggregation grows, one may reach scales for which the environmental, biological and ecosystemic processes driving occurrence are no longer the same, thereby leading to divergences between occurrence gradients that are reasonable at one scale and those that are reasonable at another scale.

Regarding the first of these limits, if probability of occurrence is identically zero for some environmental conditions, then the movement of the inflection point in the aggregated probability of occurrence distribution towards areas that had low probability of occurrence before aggregation will be blocked at these fundamen-

tal barriers to presence. In this case, the aggregated probability of occurrence will approach a threshold distribution as scale of aggregation is increased.

Given this context, the results above exploring the effects of aggregation should be thought of as applicable for intermediate spatial scales, above those for which temporal and spatial autocorrelation in presence-absence are important (i.e., the scale at which nearest-neighbor colonization and extinction events are important), but below those over which fundamental barriers to life or major changes in what processes determine occurrence are likely to occur. For example, the scale of a single forest track would likely be too small for many species (unless the species has very short dispersal potential), and the scale of an entire continent would likely be too large for many species for this theory to apply without some modification. Intermediate scales of countries or regions would likely be ideal for applying the theory developed above.

Another limit relates to the SDM development process itself. In our virtual species analyses, we are fortunate enough to know exactly what environmental variable is driving changes in probability of occurrence. However, in real situations, one rarely has access to the precise environmental variables that directly drive occurrence. Instead, one uses proxies that presumably are more or less correlated with the true drivers of occurrence. In this case, the theory and simulations developed above should be approximately applicable, though one would expect that model performance indicators would be significantly lower than those seen in our simulations. One can imagine, however, cases where the process of averaging environmental variables over larger scales may produce variables that are more or less correlated with the true drivers of occurrence (i.e., occurrence and environment are related on specific scales of spatial variability, but not others). This could cause changes in model performance indices that are not purely driven by aggregation itself, but rather by the processes underlying occurrence.

References

Mertes K, Jetz W (2018) Disentangling scale dependencies in species environmental niches and distributions. *Ecography* **41**:1604–1615. doi:10.1111/ecog.02871

Meynard CN, Leroy B, Kaplan DM (2019) Testing methods in species distribution modelling using virtual species: What have we learnt and what are we missing? *Ecography* **42**:2021–2036. doi:10.1111/ecog.04385

List of Figures

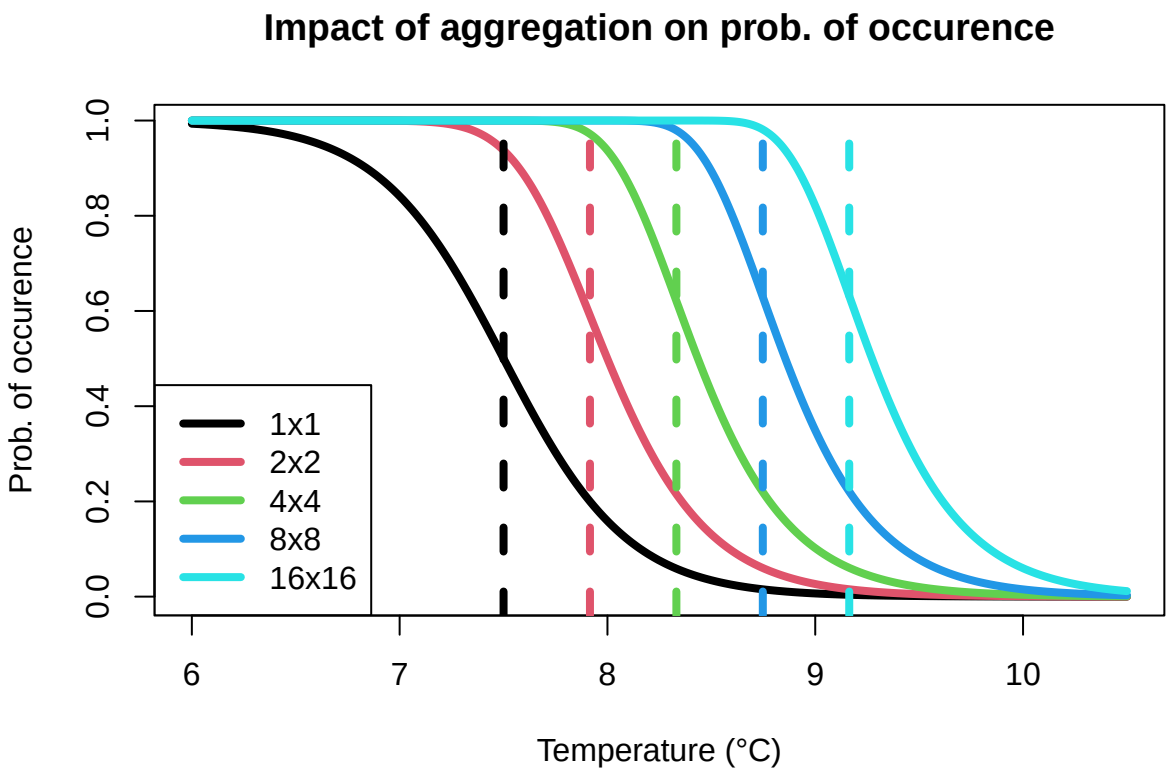

Figure S1.1: Probability of occurrence for various scales of spatial aggregation. Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed logistic Eq. 4 with $\beta = 7.5$ and $\alpha = 0.3$. Vertical dashed lines indicate positions of inflection points.

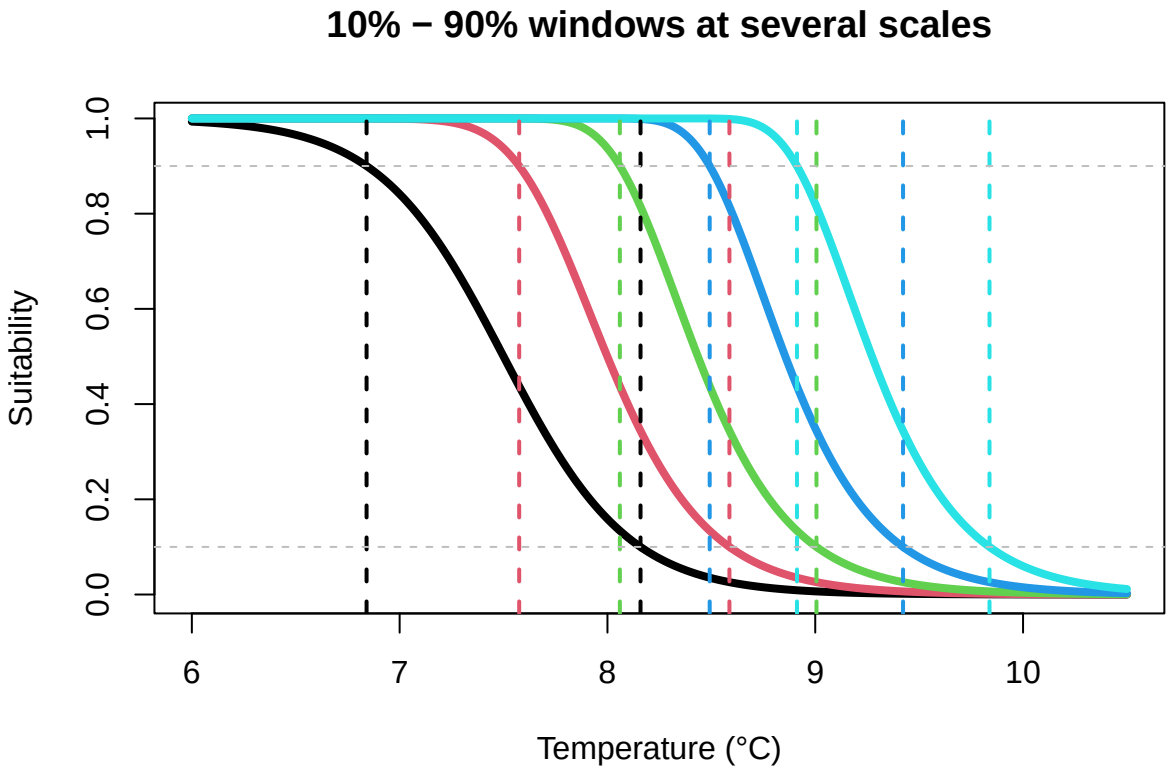

Figure S1.2: Span of area representing 10% and 90% probability of occurrence for various scales of spatial aggregation. Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed logistic Eq. 4 with $\beta = 7.5$ and $\alpha = 0.3$. Vertical dashed lines indicate environmental conditions for which aggregate probability of occurrence is 10% or 90%.

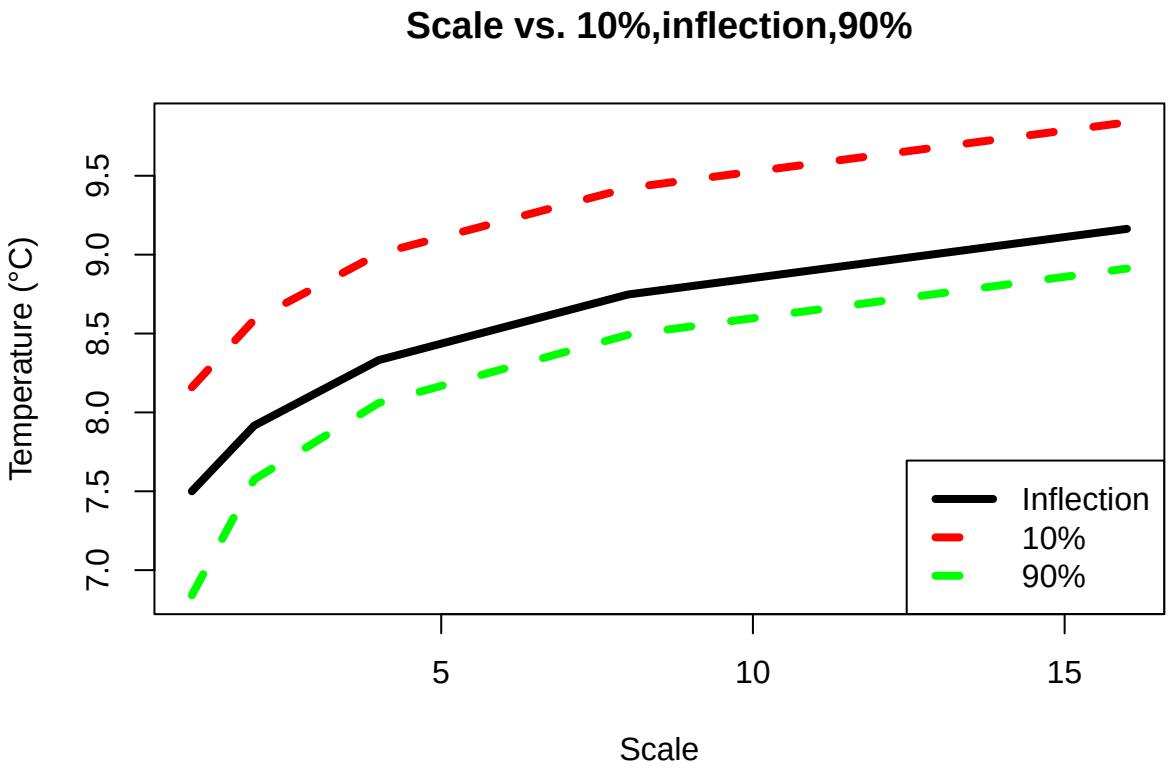

Figure S1.3: Temperature value corresponding to 10% and 90% probability of occurrence for various scales of spatial aggregation. Black central curve shows the temperature of the inflection point as a function of scale. Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed logistic Eq. 4 with $\beta = 7.5$ and $\alpha = 0.3$.

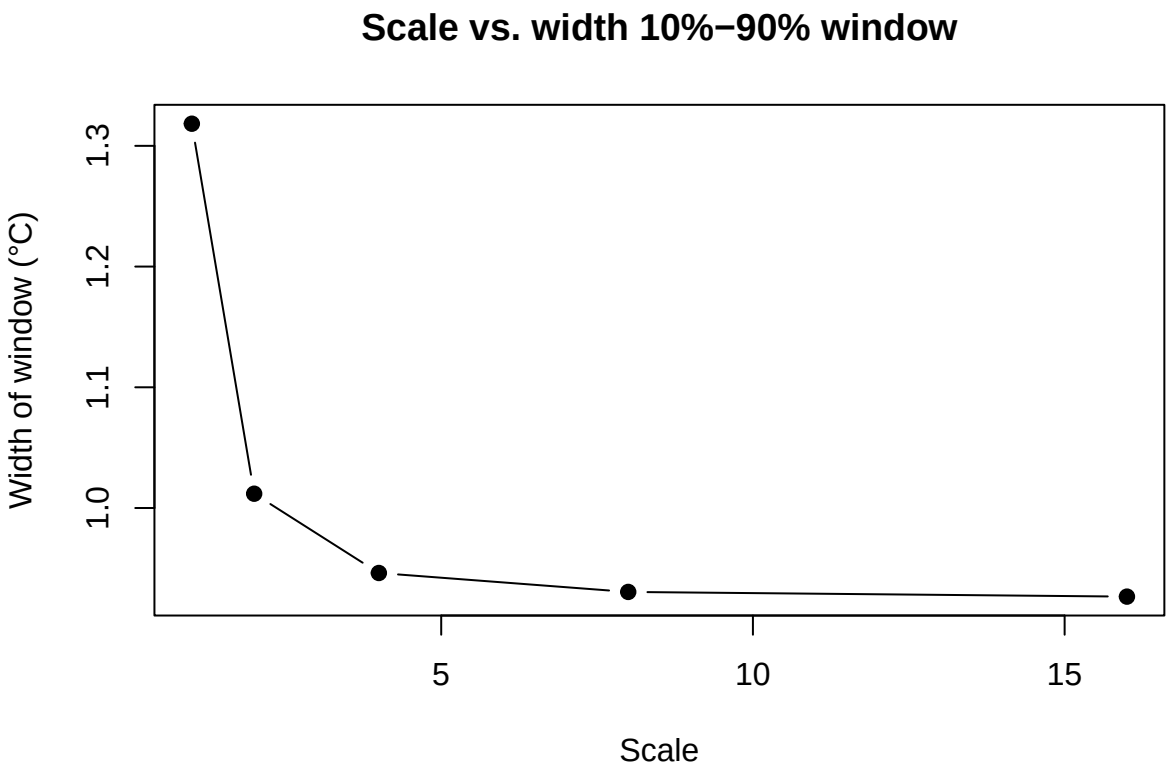

Figure S1.4: Width of 10%-90% window for probability of occurrence for various scales of spatial aggregation (i.e., width of spans shown in Fig. S1.2). Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed logistic Eq. 4 with $\beta = 7.5$ and $\alpha = 0.3$.

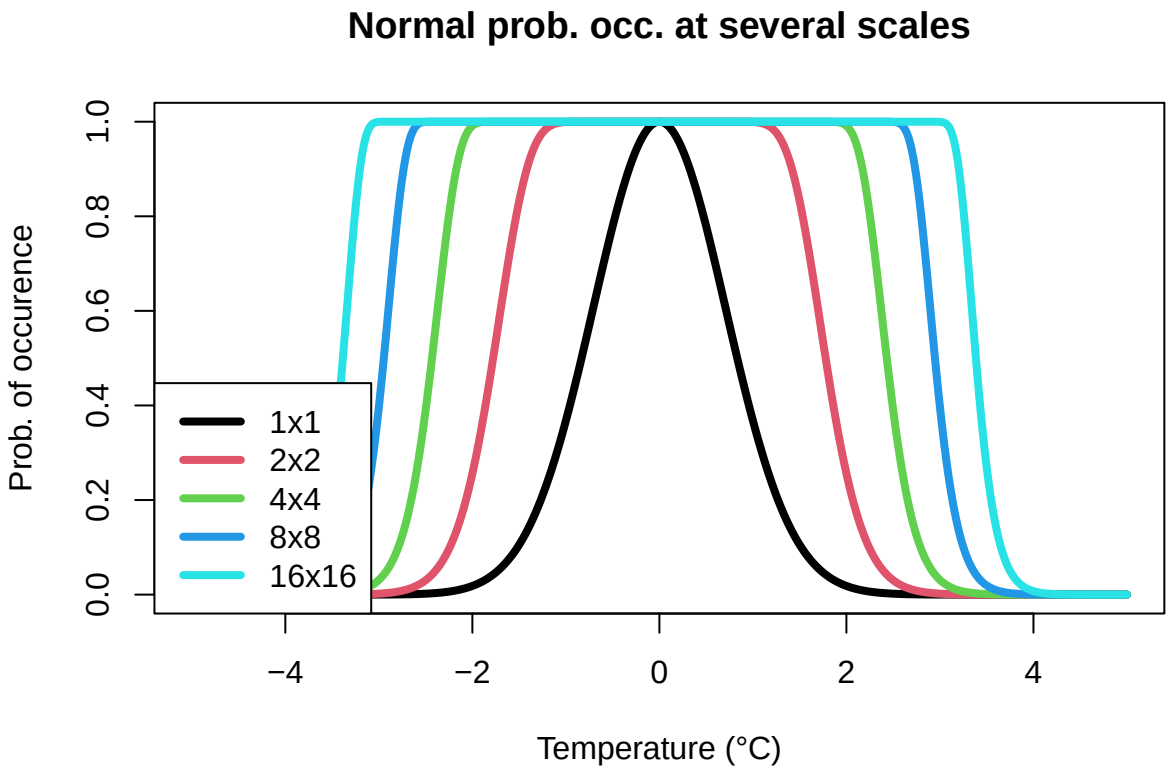

Figure S1.5: Probability of occurrence for various scales of spatial aggregation. Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed a normal distribution (i.e., Eq. (18)) with $\mu = 0$, $\sigma = 1$. C, the maximum probability value, was equal to 1.

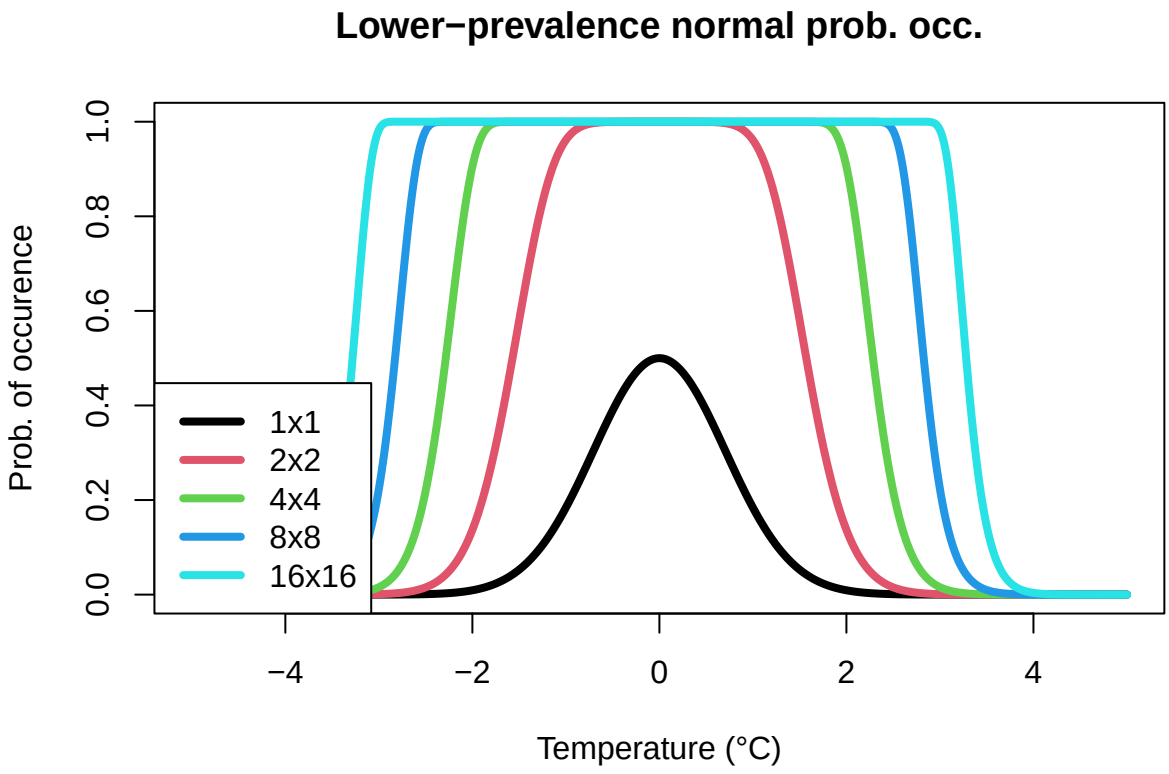

Figure S1.6: Probability of occurrence for various scales of spatial aggregation. Aggregates are assumed to be square grids of $N \times N$ cells (i.e., $\nu = N^2$). The underlying logistic curve representing species occurrence as a function of environment followed a normal distribution (i.e., Eq. 18) with $\mu = 0$, $\sigma = 1$. C, the maximum probability value, was equal to 0.5.

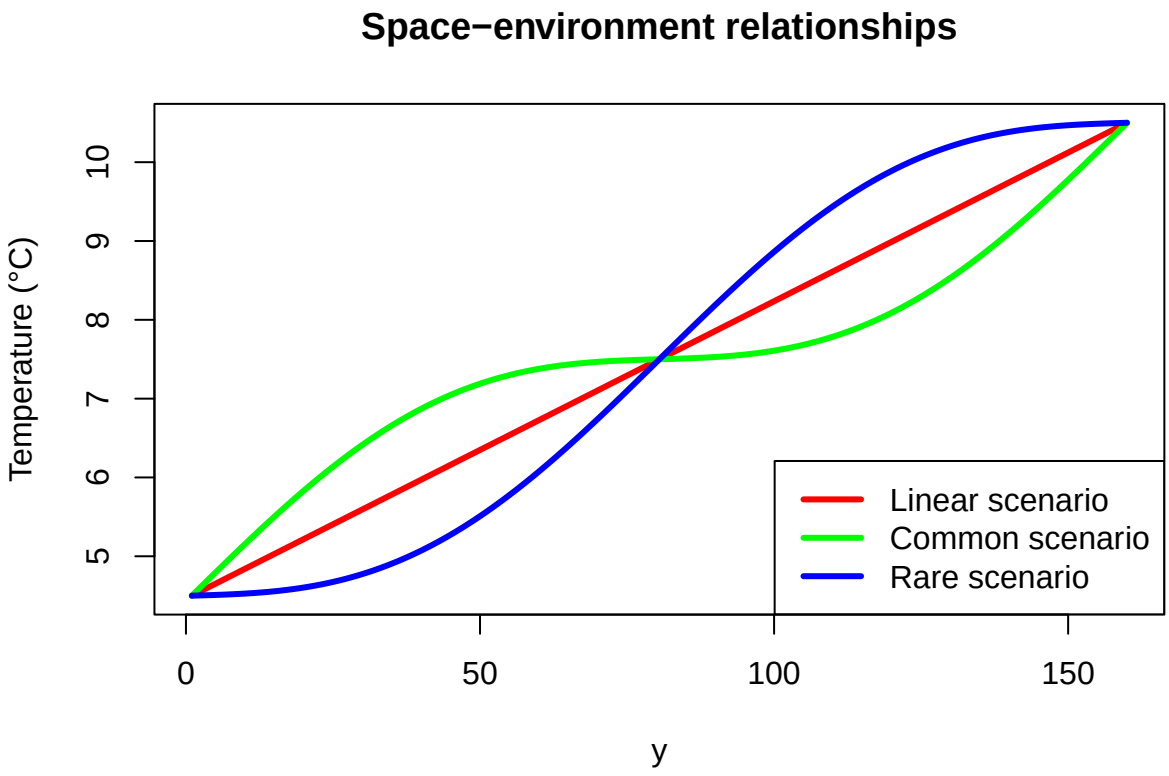

Figure S1.7: Assumed relationships between spatial coordinates in the "latitudinal" direction (i.e., y direction) and mean annual temperature. The red line shows a linear relationship (i.e., constant rate of change of temperature across space), the green line shows a relationship for which temperatures around 7.5°C, the inflection point of the unaggregated environment-occurrence relationship, are relatively common over space, and the blue line shows a relationship for which temperatures around 7.5°C are relatively rare.

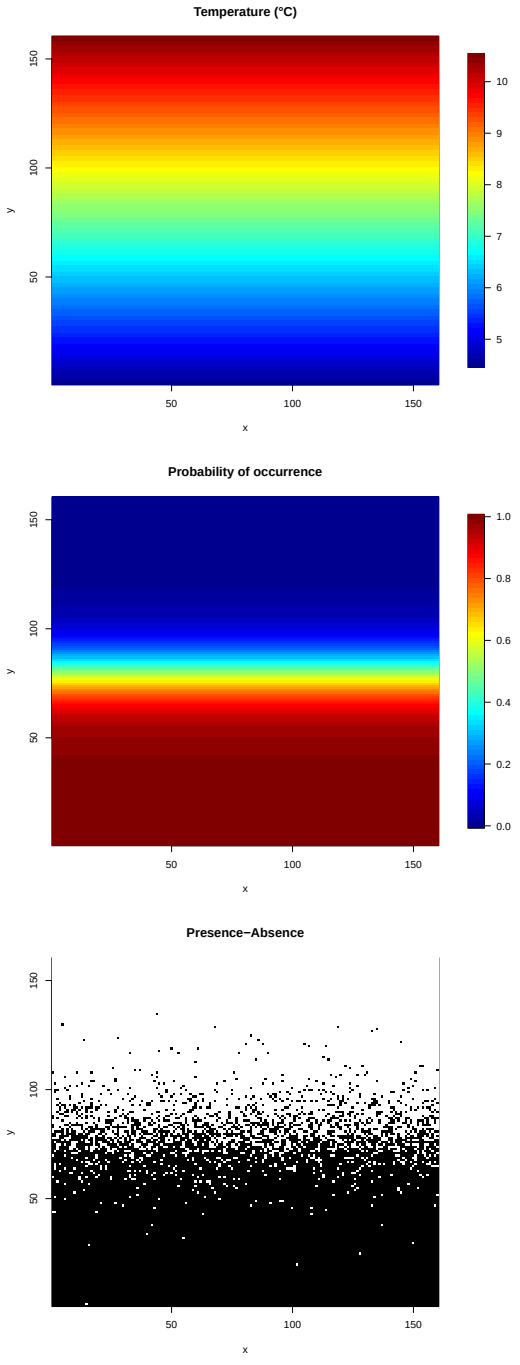

Figure S1.8: Temperature (top), probability of occurrence (middle) and one example of presence-absence based on the probability of occurrence (bottom) for the case when temperature has a linear gradient over the y dimension. For the bottom panel, white and black grid cells indicate species absence and presence, respectively.

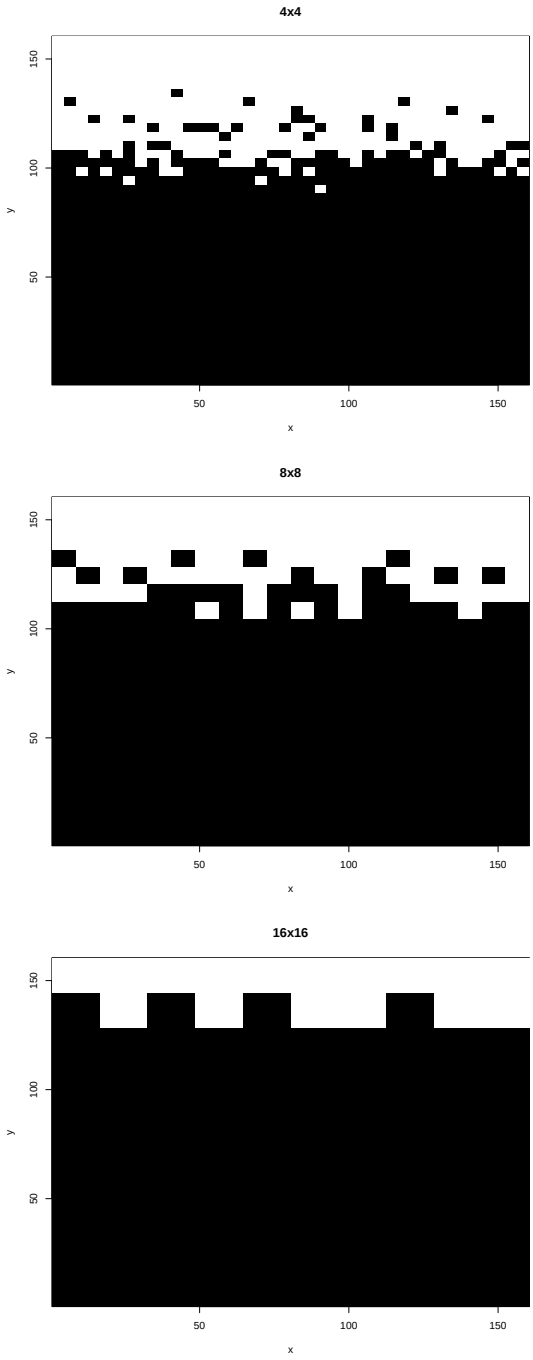

Figure S1.9: Examples of the impact of aggregation following approach #2 on perceived presence-absence. The underlying small-scale presence-absence distribution is as in the bottom panel of Fig. S1.8. The panels show aggregation at the 4x4 (top), 8x8 (middle) and 16x16 (bottom) scales. In all panels, white and black grid cells indicate assessments of species absence and presence, respectively, according to approach #2 to aggregating presence observations.

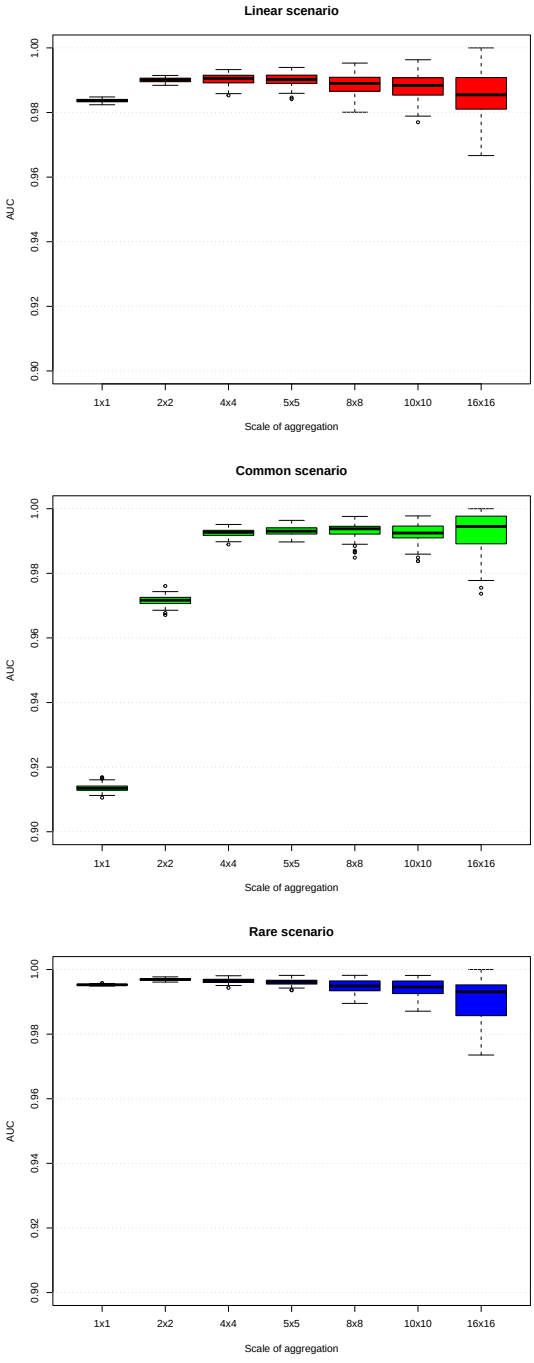

Figure S1.10: Area under the ROC curve (AUC) of GAM model predictions as a function of scale of spatial aggregation and the functional form of the space-environment relationship. Box colors and panel titles correspond to the space-environment functional forms shown in Fig. S1.7.

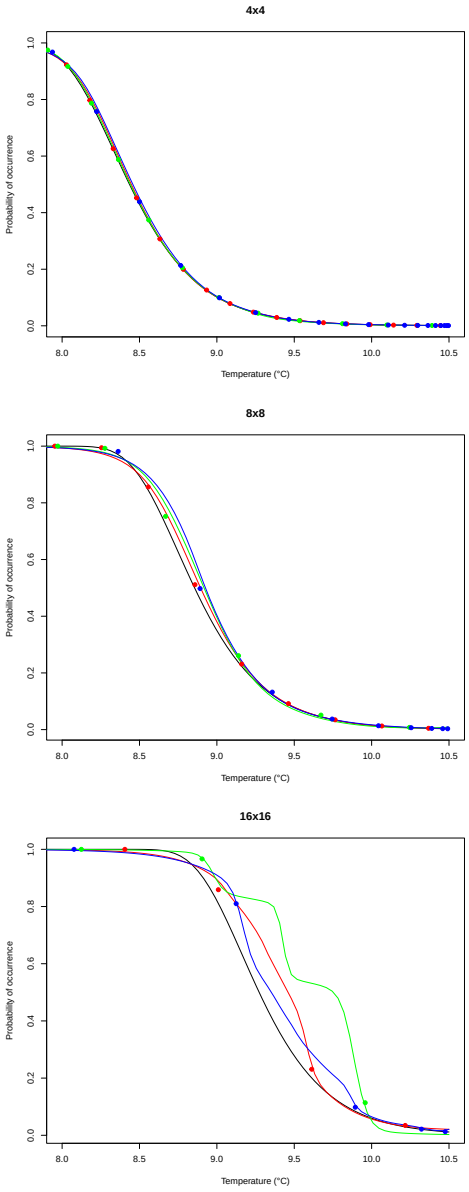

Figure S1.11: Theoretical and model estimated environment-occurrence relationships at 4x4 (top), 8x8 (middle) and 16x16 (bottom) scales of aggregation. In each panel, the black curve indicates the theoretical probability of occurrence ignoring environmental heterogeneity within aggregate as shown in Fig. S1.1. The red, green and blue curves indicate GAM model predictions for the environment-occurrence relationship based on the space-environment relationship of the corresponding color in Fig. S1.7. Probabilities have been averaged over 80 random realizations of species occurrence over the model domain (one example of which is shown in Fig. S1.9). The red, green and blue dots correspond to theoretical predictions of probability of occurrence based on application of Eq. (1) to the underlying probability of species occurrence over the model domain (e.g., for the red dots, this probability of occurrence is shown in the middle panel of Fig. S1.8). These theoretical predictions are only shown for discrete values of temperature because these are the finite set of averages over the aggregates in the model system.