Models of presence-absence estimate abundance as well as (or even better than) models of abundance: the case of the butterfly *Parnassius apollo*

David Gutiérrez · Jennifer Harcourt · Sonia B. Díez · Javier Gutiérrez Illán· Robert J. Wilson

D. Gutiérrez (Corresponding author) · S.B. Díez

Área de Biodiversidad y Conservación, Escuela Superior de Ciencias Experimentales y

Tecnología, Universidad Rey Juan Carlos, Tulipán s/n, Móstoles, Madrid, ES-28933, Spain

e-mail: david.gutierrez@urjc.es

phone: +34 914888160/ fax: +34 916647490

J. Harcourt

Evolutionary Ecology Group, Department of Zoology, University of Cambridge, Downing

Street, Cambridge CB2 3EJ, UK

J. Gutiérrez Illán

Department of Forest Ecosystems and Society, Oregon State University, Corvallis, Oregon

97331, USA

R.J. Wilson

Centre for Ecology and Conservation, University of Exeter, Cornwall Campus, Penryn TR10

9EZ, UK

Abstract (max 250 words)

Models relating species distribution records to environmental variables are increasingly applied to biodiversity conservation. Such techniques could be valuable to predict the distribution, abundance or habitat requirements of species that are rare or otherwise difficult to survey. However, despite widely-documented positive intraspecific relationships between occupancy and abundance, few studies have demonstrated convincing associations between models of habitat suitability based on species occurrence, and observed measures of habitat quality such as abundance. Here we compared models based on field-derived abundance and distribution (presence-absence) data for a rare mountain butterfly in 2006-08. Both model types selected consistent effects of environmental variables, which corresponded to known ecological associations of the species, suggesting that abundance and distribution may be a function of similar factors. However, the models based on occurrence data identified stronger effects of a smaller number of environmental variables, indicating less uncertainty in the factors controlling distribution. Furthermore, cross-validation of the models using observed abundance data from different years, or averaged across years, suggested a marginally stronger ability of models based on occurrence data to predict observed abundance. The results suggest that, for some species, distribution models could be efficient tools for estimating habitat quality in conservation planning or management, when information on abundance or habitat requirements is costly or impractical to obtain.

Keywords (max 10 words)

Introduction

Understanding patterns and mechanisms of species distribution and abundance is a central issue in ecology and conservation (Brown 1984). Spatial patterns in the distribution of individuals may have important effects not only on the occupancy and abundance of single species (Lawton 1993), but also on interspecific occupancy-abundance relationships (He and Gaston 2000, Holt et al. 2002) and community species-abundance distributions (McGill et al. 2007). In practical terms, conservation planning requires the identification of sites whose environmental characteristics allow them to support sustainable populations of focal species (e.g. Araújo and Williams 2000). However, for many species and parts of the world, information on distribution, abundance, and the environmental determinants of these, is far from comprehensive.

Species distribution models could help to overcome the problem of incomplete information on species distributions and abundance, because they relate occurrence or abundance data with the environmental attributes of known locations, and use the relationships to estimate occurrence or abundance more widely (Guisan and Zimmermann 2000; Guisan and Thuiller 2005; Elith and Leathwick 2009; Schröder et al. 2009). With rapid recent advances in large data-base management, statistical techniques, physical geography and geographic information systems, species distribution models are now widely used for explaining and predicting occurrences (and to a much lesser degree, abundances) for many biological groups, over a wide range of spatial scales and environments (Elith and Leathwick 2009).

Due to data availability, species distribution models are most commonly based on occurrence data (presence-absence or presence-only), and therefore estimates of habitat suitability often consist of predicted probabilities of occurrence (Guisan and Zimmermann 2000). Predicted habitat suitability may subsequently be used for evaluating the impact of environmental change on species distributions (e.g. Schweiger et al. 2008), supporting

management plans for species recovery and reintroduction (e.g. Willis et al. 2009), selecting

reserves (e.g. Araújo and Williams 2000; Cabeza et al. 2010) and assessing species invasion

(e.g. Peterson and Vieglais 2001). All these approaches assume that habitat suitability from

models is positively correlated with habitat quality, but testing this assumption is by no means

5   straightforward. Empirical estimation of habitat quality (sensu Van Horne 1983) involves

detailed information on density, mean individual survival, and mean expectation of future

offspring, and these demographic parameters may be prohibitively intensive and costly to

collect over a large number of sites at broad spatial scales (e.g. Elmendorf and Moore 2008;

St-Louis et al. 2010). As an alternative, population density might be assumed to be positively

10   correlated with habitat quality, but there are limitations to this approach, such as in temporally

variable environments where abundance varies greatly from year to year (Van Horne 1983).

Nevertheless, positive intraspecific occupancy-abundance relationships suggest that

common environmental factors may indeed govern both the distribution and abundance of

individual species (Gaston et al. 2000; Holt et al. 2002). Analyses based on time-series data

15   suggest that this relationship is stronger for species showing positive or negative trends in

distribution and abundance than for species whose populations are fluctuating in response to

interannual stochasticity (Gaston et al. 1998, 2000; Holt et al. 2002). There is also less

evidence for positive intraspecific occupancy-abundance relationships using spatial data for

single time periods (but see Venier and Fahrig 1998), even though relationships of abundance

20   and occupancy over spatial environmental gradients have important implications for the

structure of species' geographic ranges (e.g. Sagarin and Gaines 2002) and species responses

to environmental change (e.g. Maclean et al. 2011). In fact, despite the potential importance

of a positive correlation between abundance and habitat suitability, few distribution-

modelling studies have validated the relationship (Jiménez-Valverde 2011). Furthermore, for

25   a substantial number of species, predicted habitat suitability does not appear to be

significantly correlated with observed abundance, particularly when restricting analyses to occupied, potentially habitable sites (Pearce and Ferrier 2001; Nielsen et al. 2005; Jiménez-Valverde et al. 2009; Duff et al. 2011; Guarino et al. 2012; but see VanDerWal et al. 2009; Oliver et al. 2012). Limitations to these habitat suitability models may result from inadequate

5   survey techniques, or inappropriate choice of survey scale and available environmental variables (Pearce and Ferrier 2001). For instance, data pooled from surveys conducted over a number of years may obscure the effects of environmental variables on abundance by combining variation in space and time (e.g. Pearce and Ferrier 2001; Jiménez-Valverde et al. 2009). Furthermore, refinement of habitat predictors from the broad abiotic and vegetation

10   information used in some studies, to variables of clear importance to focal species, could also improve abundance-habitat suitability correlations (e.g. Pearce and Ferrier 2001; Jiménez-Valverde et al. 2009; Oliver et al. 2012).

In this paper, we develop models of abundance and distribution for adults of the apollo butterfly, *Parnassius apollo*, in a mountain area of central Spain. The system provides

15   considerable environmental variation over space (elevation, topography and vegetation structure), and time (interannual weather variability) (Gutiérrez Illán et al. 2010). *P. apollo* is appropriate for the research because there are reliable methods for estimating abundance and distribution (Pollard and Yates 1993), it is an easily detectable species in the field (Sánchez-Rodríguez and Baz 1996), and relevant environmental variables can be deduced from larval

20   habitat requirements at local scales (Ashton et al. 2009). For this system, we (1) compared the consistency of variables selected in abundance and distribution models based on empirical data for a single year, under comparable environmental conditions; and (2) evaluated the models' predictive power using observed abundances collected in the same year and in two subsequent years, allowing us to account for spatial and temporal environmental variation in

25   the region.

Methods

Study system

Parnassius apollo (L.) is a predominantly mountain species, whose larvae feed on *Sedum* spp. and sometimes other Crassulaceae (Deschamps-Cottin et al. 1997). It has one adult generation

5   per year (June-August), and hibernates as a small larva in the eggshell (Tolman and Lewington 1997). Its European population is estimated to have declined by almost 30% since 2000, with greatest declines at low elevations (van Swaay et al. 2010). Climate warming is thought to be implicated in the loss of low-elevation populations (Descimon et al. 2005), but land use change and pollutants have also been linked to its decline (Gomariz Cerezo 1993;

10   Sánchez-Rodríguez and Baz 1996; Nieminen et al. 2001, but see Fred and Brommer 2005).

The Sierra de Guadarrama (central Spain) is an approximately 100 x 30 km mountain range located at 40°45' N 4°00' W. The mountain range includes 25 separate 10 km grid squares in which *P. apollo* has been recorded historically, in a population network that is geographically separated from all other records of the species in Spain (García-Barros et al.

15   2004). The mountain range is bordered by plains with elevations of *c.* 700 m (to the north) and *c.* 500 m (to the south) and reaches a maximum elevation of 2428 m (Fig. 1). The main regional host plant reported for *P. apollo* is *Sedum amplexicaule* (Sánchez-Rodríguez and Baz 1996), although larvae have also been observed feeding on *S. brevifolium*, *S. forsterianum* and *S. album* (Ashton et al. 2009). Recent phylogeographic analyses have shown that

20   southwestern European populations retain a large fraction of genetic variation of *P. apollo*, highlighting their conservation value (Todisco et al. 2010).

#Fig. 1 approximately here#

Abundance and distribution of *P. apollo*

In 2006, butterflies (including *P. apollo*, if present) were counted on standardised 500 m long by 5 m wide transects (Pollard and Yates 1993) every two weeks at 43 sites (random sites henceforth; elevation range 550-2250 m), of which 40 were also sampled in 2007 and 2008

5 (in total, 246 visits over the *P. apollo* flight period). However, *P. apollo* is rare in the study system (only a maximum of 5 occupied sites from the 43 sites visited in 2006), so it would have been necessary to sample very many random sites to achieve 20 or more presences, an appropriate minimum number for abundance and distribution models (e.g. Wisz et al. 2008). Therefore, in 2006 we visited 47 additional locations to increase our sample of *P. apollo*

10 presences, selected using (1) *P. apollo* records from butterfly surveys in 2004 and 2005 (Gutiérrez Illán et al. 2010), or (2) *S. amplexicaule* records from 2005. Our 90 (43 + 47) sample sites were located in 29 UTM 10 km grid squares in total, including 17 of the 10 km grid squares where *P. apollo* has ever been recorded (García-Barros et al. 2004).

At each additional site, we walked the 500 m transect twice (usually 1-2 weeks apart,

15 weather permitting) around the *P. apollo* peak flight period expected for the elevation based on preliminary data from 2005 and by walking weekly transects at four, nine and seven sites in 2006, 2007 and 2008, respectively, between early June and mid-August (Ashton et al. 2009). We sampled for *P. apollo* at all 90 sites in 2006, 62 sites in 2007, and 59 in 2008 (40 of the 43 random sites plus 22 or 19 additional sites, respectively). Because *P. apollo*

20 frequently occurs in low density populations (but is easily visually detected), it was considered present where one or more individuals were counted (including records before or after the transect count in a few cases), and absent where no individuals at all were observed.

Spatial autocorrelation can influence the reliability of biogeographic analyses, because it potentially inflates Type I errors in null hypothesis significance testing and generates longer

25 models in information theoretic approaches (e.g. Diniz-Filho et al. 2008). We ensured that

survey sites were selected to be located in separate 1 km grid squares, corresponding to a distance travelled by fewer than 10% of adult butterflies in another study (Brommer and Fred 1999). In addition, to test formally for spatial autocorrelation, we generated all-directional correlograms (Legendre and Legendre 1998) for abundance data in 2006 by plotting values of

5   Geary's c coefficient (recommended for variables departing from normality) against Euclidean distances between sites. Geary's c calculation and significance testing were performed using 4999 Monte Carlo permutations in Excel add-in Rookcase (Sawada 1999). No correlogram was globally significant, indicating that spatial autocorrelation in *P. apollo* abundance data was negligible.

10   **Environmental variables**

Universal Transverse Mercator (UTM) coordinates were recorded every 100 m along transects using a handheld Garmin GPS unit, and were used to plot transects in a geographic information system (ArcGIS) (ESRI 2001). The average elevation of 100 m cells intercepted by transects was determined using a digital elevation model (Farr et al. 2007).

15   We estimated insolation as the total direct solar radiation per 100 m grid cell during the whole year using the Solar Analyst 1.0 extension for ArcView GIS (Fu and Rich 2000), based on latitude, slope, aspect, and elevations of surrounding cells in a 110 km x 155 km area. Insolation variables were estimated as the mean for 100 m grid cells intercepted by each transect.

20   In the study area, elevation is related to climate parameters (annual mean temperature: 5.8-5.9°C/km decrease; annual rainfall 683-767 mm/km increase; $R^2 = 0.94$ in both cases; Wilson et al. 2005), but these gradients are based on relatively few meteorological stations (10-11). Hence, we use elevation and modelled insolation intensity instead of estimated temperature and rainfall in our models.

We used twenty 0.25 m^2 quadrats (50 x 50 cm) per transect at 25 m intervals to estimate percentage cover of each *Sedum* species in 2006. Vegetation height at the centre of each quadrat (in 2006), and bare ground and shrub cover (in 2008) were also recorded. A site average was taken for each variable (*n* = 20 quadrats). We also estimated *Sedum* frequency, the proportion of quadrats occupied by each species (range 0-1), as a measure of the host plant distribution over each transect site. As an estimate of the total host plant resource available for *P. apollo* at each site, we calculated percentage cover and frequency for the four *Sedum* species known to be eaten by larvae (see above). All measured variables in our study have biological significance (Ashton et al. 2009), as detailed in Table 1.

#Table 1 approximately here#

Abundance and distribution models

To analyse *P. apollo* abundance, we used GLMs applying a quasi-likelihood estimation of regression coefficients using a log-link and setting the variance equal to mean (quasi-Poisson regression, McCullagh and Nelder 1989; Ver Hoef and Boveng 2007). For *P. apollo* distribution, we performed GLMs with logit-link and binomial error (logistic regression). Sample size was *n* = 90 sites in both cases. We included six candidate variables for *P. apollo* abundance and distribution models (Table 1). We selected only one (host plant *Sedum* frequency) from the two potential host plant variables in Table 1 because univariate analyses showed stronger relationships of *P. apollo* abundance and distribution with that variable than with host plant *Sedum* cover (results not shown) and they were highly correlated (r_s = 0.92, *P* < 0.001). Only one pair-wise correlation between the remaining independent variables had absolute values higher than 0.7 (the most commonly applied threshold, Dormann et al. 2012), elevation-vegetation height (r_s = -0.75, *P* < 0.001). However, we did not exclude these variables from analyses because they had potentially different biological significance (see Dormann et al. 2008).

We used the information-theoretic approach (Burnham and Anderson, 2002) to model

abundance and distribution of *P. apollo*. We included linear and quadratic terms for the 5

condition variables and only linear terms for the resource variable (Table 1). For each

response variable, we fitted all possible combinations of linear and quadratic terms (subject in

5   the last case to the condition that the corresponding linear term was included in the model),

with no interactions, and used the Akaike Information Criterion, adjusted for small sample

size ($QAIC_c$ for abundance and AIC_c for distribution; Burnham and Anderson 2002) to rank

models. To obtain our model confidence sets, we selected models that were within six

$\Delta(Q)AIC_c$ units of the top-ranked model (Richards 2005), excluding more complex models

10   that do not have a $\Delta(Q)AIC_c$ which is lower than all the simpler models within which they are

nested (Richards 2008). This procedure guards against the selection of over-parameterised

models whilst maintaining a high probability of selecting the true best model (Richards 2008).

The adequacy of quasi-Poisson regression for modelling abundance data was examined using

estimated and empirical variance-mean plots for the full model (Ver Hoef and Boveng 2007).

15   Following model selection, we used model-averaging to obtain model coefficients based

on the confidence sets. Doing so incorporates model selection uncertainty whilst weighting

the influence of each model by the strength of its supporting evidence (Burnham and

Anderson 2002). Model–averaged coefficients were derived by weighting using Akaike

weights and averaging coefficients over all models in the confidence set. Averaging over all

20   models means that in those cases in which a variable was not in a particular model, its

coefficient value was set to zero. This serves to ameliorate much of the model selection bias

of coefficients (Burnham and Anderson 2002). We also estimated relative variable importance

by summing the Akaike weights across all models in the confidence set that contain that

variable. This parameter lies in the range 0-1 and provides evidence for the importance of

25   each variable relative to the other variables in the context of the set of models considered.

Model selection and model averaging were performed with "MuMIn" package version 1.6.6

(R Development Core Team 2011; Bartoń 2012).

Model evaluation

Abundance and distribution models were evaluated in two ways, verification and cross-

5   validation (Araújo and Guisan 2005). For verification, we calculated Spearman's rank

correlation coefficients (r_s) for predicted abundance or probability of occurrence (from model-

averaged coefficients) against observed abundance values in 2006, 2007, 2008 and averaged

for 2006-2008 (Guisan and Zimmermann 2000; Potts and Elith 2006). Correlations for 2006-

2008 average abundance were calculated for testing the effect of interannual variability in

10   abundances on model predictions: we would expect larger correlations with averaged than

with individual annual abundances. We used rank correlations between predicted and

observed values because our transect counts were relative estimates of local abundance rather

than absolute densities or population sizes (Pearce and Ferrier 2001).

Given that there were insufficient sites to have separate calibration and evaluation data

15   sets, we used a Jackknife procedure for cross-validation (e.g. Elmendorf and Moore 2008;

Jiménez-Valverde et al. 2009). This method consisted of generating N confidence set models,

sequentially omitting one site, where N is the number of sites. We then calculated model-

averaged coefficients for each confidence set. Based on those coefficients, we calculated r_s for

predicted abundance and probability of occurrence against observed abundance values for

20   each omitted site. We examined the relationships between observed abundances and predicted

values (abundance or probability of occurrence) using two tests (e.g. Pearce and Ferrier 2001;

Nielsen et al. 2005): (1) observed abundance with all samples (including absences) against

predicted values; and (2) observed abundance-where-present (omitting absences) against

predicted values. Both tests represent the model contribution to explaining abundance, but

25   only the first one includes the discrimination of absent locations (Nielsen et al. 2005). r_s

coefficients were calculated with "pspearman" package version 0.2-5 (Savicky 2009; R

Development Core Team 2011).

Results

Abundance and distribution models

5 In 2006, a total of 231 *P. apollo* were counted in 26 of the combined sample of 43 random

sites and 47 targeted sites, with a maximum local abundance of 36 individuals. The lowest

elevation presence was at 1287 m. In 2007 and 2008, we counted 184 (21 out of 62 sites) and

98 (21 out of 59 sites) *P. apollo* butterflies, with maximum local abundances of 43 and 18

individuals, respectively. The species was observed in 15 separate 10 km grid squares.

10 Abundance and distribution models were based on multi-model inference with $(Q)AIC_c$

(Table 2). The dispersion parameter for the full quasi-Poisson model was 3.20, indicating that

the data were not excessively over-dispersed (values above 4 would suggest that model

structure could be inadequate; Burnham and Anderson 2002). Estimated and empirical

variance-mean plots for the full model suggested that quasi-Poisson regression was

15 appropriate for this data set (results not shown).

#Table 2 approximately here#

For abundance, the confidence set consisted of 13 models. The final model included

quadratic relationships with elevation, bare ground cover, shrub cover, vegetation height (all

with positive linear and negative quadratic coefficients) and insolation intensity (with

20 negative linear and positive quadratic coefficients), and a positive linear relationship with host

plant *Sedum* frequency. Relative variable importance was higher for elevation, shrub cover

and their corresponding quadratic terms, and host plant *Sedum* frequency (values ≥ 0.90;

Table 2).

For distribution, the confidence set consisted of four models. The final model included quadratic relationships with elevation and shrub cover (with positive linear and negative quadratic coefficients) and a positive linear relationship with host plant *Sedum* frequency. Relative variable importance was higher for elevation and its quadratic term, shrub cover and host plant *Sedum* frequency (values ≥ 0.90 in all cases; Table 2).

Model evaluation

The performance of abundance and distribution models was evaluated by testing the correlation of predicted abundances and probabilities of occurrence against observed abundances (Table 3). For the complete data set (90 sites sampled in 2006), predicted abundance from the final model was significantly positively correlated with observed abundance, with smaller values for cross-validation than for verification (Fig. 2). Including only those sites in which *P. apollo* was present, produced smaller correlation coefficients between predicted and observed abundance values (significant for verification and non-significant for cross-validation, Fig. 2). The pattern was very similar for the correlations between predicted probability of occurrence and observed abundance, but in this case all coefficients were significant. The scatter plot suggests larger variability in observed abundance values for higher predicted probabilities of occurrence (Fig. 2).

#Table 3 approximately here#

#Fig. 2 approximately here#

For the reduced data sets (59 sites sampled in 2006-2008), all correlations were on average higher than their corresponding correlations performed with the complete data set (90 sites in 2006). This was due probably to the fact that, in the reduced data sets, a relatively large number of sites with high probability of occurrence but unoccupied by *P. apollo* were excluded from analyses (Fig. 2). Apart from this, the pattern shown by correlation values was also similar to that for the complete data set, with no apparent trend over the sampling years.

The only non-significant correlation coefficients were those between predicted and observed abundances for cross-validations for occupied sites for 2006, 2007 and 2006-2008 average.

Discussion

Ecological significance of abundance and distribution models

Most studies to date concerning correlations between abundance and modelled habitat suitability have provided no details of comparisons between abundance and distribution models (Pearce and Ferrier 2001), or have performed no abundance models at all (Jiménez-Valverde et al. 2009; VanDerWal et al. 2009; Oliver et al. 2012). The exceptions are more specific studies involving a few species, which suggest that environmental factors influencing abundance may differ from those limiting distribution at least in some cases (Nielsen et al. 2005; Duff et al. 2011). In this study, there was high concordance in the variables selected by the two different approaches using count and presence-absence data, suggesting that abundance and distribution of the butterfly *Parnassius apollo* were associated with similar environmental factors. This supports the idea that common environmental factors govern both the abundance and distribution of individual species, which may result in positive intraspecific occupancy-abundance relationships (Gaston et al. 2000), contributing to positive interspecific occupancy-abundance relationships (Holt et al. 2002).

Models from our study consistently identified quadratic relationships for *P. apollo* abundance and distribution with elevation and shrub cover, and linear positive relationships with host plant *Sedum* cover (with coefficients of similar magnitude – in the link scale – and large relative variable importance). In the case of abundance, there were also quadratic relationships with the remaining environmental variables (insolation intensity, bare ground and vegetation height) but with smaller relative variable importance.

The modelled relationships with environmental variables are supported by our knowledge

of the ecology of *P. apollo*. The association of *P. apollo* with intermediate elevations suggests

a restriction to relatively cold sites in the region, possibly because of direct effects of

temperature on *P. apollo* individuals. Larvae appear to select for microhabitats with

5 temperatures in the range 20-28°C, and occupy those that are cooler than ambient above 27°C

(Ashton et al. 2009). For adults, the information is much sparser, but they may be quite

vulnerable to dehydration under warm temperatures (Baz 2002). An alternative explanation is

that *P. apollo* requires cold sites through indirect effects of temperature and humidity on host

plant phenology, because its main host *S. amplexicaule* senesces in spring-early summer.

10 *P. apollo* abundance and distribution were also associated with intermediate cover of

shrubs. In sites where *S. amplexicaule* is the main host plant, shrubs may be important egg

substrates (shrubs received 32% of eggs; S. Ronca, unpublished data from female tracking, *n*

= 71) because eggs laid on the host plant itself might be displaced on senescent tissue away

from the following year's growth (Fordyce and Nice 2003). In addition, larvae appear to use

15 shrubs to provide shade when the ambient temperature is high, and shelter during cold

conditions (Ashton et al. 2009). Larvae bask on bare ground during cold but sunny weather,

so excessively dense shrub could be detrimental for *P. apollo* larvae (Ashton et al. 2009), and

dense shrub could also be unsuitable because the host plant species (particularly *Sedum*

amplexicaule, *S. brevifolium* and *S. album*) are generally associated with open areas.

20 Local site frequency of host plants was more important than its overall abundance,

suggesting that sites with widespread but low density plants were more favourable than those

with relatively few high density patches of plants. This result could reflect oviposition and

larval behaviour in the species, since females do not lay eggs on host plants (see above;

Gomariz Cerezo 1993; Deschamps-Cottin et al. 1997; Fred and Brommer 2003). Although we

25 do not know the dispersal ability of newly hatched larvae, it seems unlikely that they could

successfully locate host plants more than *c*. 0.5 m away (Fred and Brommer 2010). Older

larvae move between different microhabitats related to ambient temperature (Ashton et al.

2009). Hence, host plants which are both widely distributed (for newly hatched larvae) and

growing in a range of microhabitats (for later instars) may be important.

Estimating abundance from abundance and distribution models

A key finding from this study was that, for *P. apollo*, predicted probabilities of occurrence

from logistic regression models performed as consistently when considering all test data, or

even better when considering presence data only, than indices of abundance derived from

quasi-Poisson regression models. Encouragingly, the Spearman's correlation coefficients

obtained from our study were on average larger than those previously shown in comparable

studies (e.g. Pearce and Ferrier 2001; Elmendorf and Moore 2008; Guarino et al. 2012). Thus,

a model of probability of occurrence based on presence-absence data might serve as a

surrogate for estimates of *P. apollo* abundance. Guarino et al. (2012) suggested that sample

size might influence the detection of relationships between abundance and predicted

probability of occurrence when comparing complete data sets with those with absences

excluded, but this does not appear to be the case here because correlations were of similar

magnitudes and significance (Table 2).

The smaller predictive ability of the abundance model relative to the presence-absence

model could be due to larger model uncertainty in the first case. *P. apollo* is an annual species

which shows marked yearly fluctuations in abundance. Although insect populations show

synchronic dynamics over regional scales, local dynamics are likely to depend on habitat

differences (Powney et al. 2010). Variables measured for habitat models are frequently

assumed constant in ecological time and are consequently unable to explain yearly

fluctuations in abundance, which can be the result of weather variability or demographic

factors (Jiménez-Valverde et al. 2009). The larger uncertainty in modelling abundance was

reflected in the size of the confidence set and model Akaike weights relative to that for presence-absence models (Table 2). Nevertheless, it is encouraging that despite the large population fluctuations shown by *P. apollo*, observed abundance was still correlated with predicted occurrence, in contrast with the observation from comparative analyses that positive

5   occupancy-abundance relationships may be masked by interannual population variability (Gaston et al. 1998, 2000).

It is worth noting that abundance variability was larger for those sites with higher probability of occurrence (Fig. 2), suggesting that habitat suitability could indicate the upper limit of abundance, rather than average abundance (VanDerWal et al. 2009). Hence, when

10   habitat suitability is low, abundance is consistently low. However, when habitat suitability and potential abundance are high, other environmental factors (e.g. adult resources) or unmeasured constraints (e.g. biotic interactions such as parasitism, dispersal limitations –see below-) may limit abundance in some sites (VanDerWal et al. 2009; Oliver et al. 2012). Our results (Fig. 2) tally with the polygonal distribution of points over the space defined by

15   abundance and predicted habitat suitability found by VanDerWal et al. (2009).

The positive relationship between abundance and predicted probability of occurrence suggests the possibility of using predicted distributions for ranking habitat quality. Sampling presence-absence may often be easier than sampling abundance, such as in our system for a species with many populations located in remote mountain sites. Abundance surveys are

20   extremely time limited because the flight period of *P. apollo* is shorter than one month in many populations; and for transects to be comparable they must be walked during a limited period around the peak of the flight season, during the limited time of day when temperatures are warm enough for butterfly activity. In contrast, occurrence can be sampled on the basis of adult data collected during the whole flight period, or immature stage data (in the case of *P.*

apollo, mostly larvae) recorded during spring. This approach could also be applicable to other annual species in which the realistic time period available for sampling is limited.

Landscape-scale persistence and conservation

Apart from methodological issues, the positive relationships between abundance and predicted probability of occurrence suggest some important points concerning the *P. apollo* distribution. Firstly, based on a minimized difference threshold value of 0.366 (Sing et al. 2005; Jiménez-Valverde and Lobo 2007), there were 6 occupied sites for which the distribution model predicted *P. apollo* to be absent (Fig. 2). All these sites showed relatively low *P. apollo* abundance, suggesting lower habitat quality, and that presence might partly depend on immigration. Secondly, using the same threshold, there were 14 unoccupied sites for which the distribution model predicted *P. apollo* to be present (Fig. 2). Although we cannot entirely rule out the possibility that some important habitat variables were missing from the model, this could also suggest that *P. apollo* was absent from some suitable habitat. *P. apollo* almost certainly inhabits a discontinuous patch network in the Sierra de Guadarrama, in which metapopulation processes may be important for persistence (e.g. Hoyle and James 2005). Nevertheless, to evaluate the importance of such processes for *P. apollo* persistence would require further data on dispersal between different areas (e.g. Brommer and Fred 1999). In this context, the role of metapopulation dynamics on the relationship between observed abundance and predicted probability of occurrence is an additional issue that remains to be examined (e.g. Hanski et al. 1993).

Our results suggest that distribution models can produce estimated probabilities of occurrence that are reasonable predictors of abundance rankings. This is encouraging because occurrence models are widely used in conservation planning, in which their outputs are used to estimate persistence (e.g. Araújo and Williams 2000; Cabeza et al. 2010). Nevertheless, we focused only on abundance, which is just one component of persistence. Other factors such as

population stability are known to influence persistence in other animal populations (Oliver et al. 2012), and may indeed be important to local dynamics in environmentally variable areas such as mountains. We conclude that in this case distribution models may be useful for predicting habitat suitability in a rare insect, but that their wider use may require further

5 validation in terms of abundance and population variability.

Acknowledgements

Table 1. List of environmental variables included in the present study, classified by their biological significance (Ashton et al. 2009). Variables included in the final multivariate analyses are those with code.

Environmental variable (units)	Code	Mean (min-max)
Macroclimate conditions: adult thermoregulation and larval development		
Elevation (km)	Elev	1.43 (0.56-2.25)
Insolation intensity (kWh m^{-2} per day)	Insol	2.81 (2.04-3.38)
Microclimate conditions: larval thermoregulation		
Bare ground cover (percentage cover)	Bgr	21.33 (1.05-72.65)
Shrub cover (percentage cover)	Shr	15.91 (0-60.80)
Vegetation height (cm)	Heig	9.96 (2.00-30.90)
Resources: larval host plants		
Host plant *Sedum* cover (percentage cover)		5.35 (0-38.20)
Host plant *Sedum* frequency (proportion occupied)	Sedf	0.27 (0-0.95)

Table 2. Confidence set GLM models for *P. apollo* (a) abundance (quasi-Poisson error and log-link) and (b) distribution (binomial error and logit-link) in 2006 ($n = 90$ in both cases). The table indicates the variables included in the model and the direction of their coefficients (+/-; codes in Table 1); number of parameters (K, including one extra parameter for over-dispersion factor in $QAIC_c$, see below); Akaike Information Criterion for small sample size [$(Q)AIC_c$ (corrected for over-dispersed count data)]; difference in $(Q)AIC_c$ between current and best model [$\Delta(Q)AIC_c$]. Confidence sets are within six $(Q)AIC_c$ units of the top ranked model, excluding models which are higher-ranking nested variants of simpler models with lower $\Delta(Q)AIC_c$ (Richards 2008). Relative importance (Imp), model-averaged coefficients (Coef) and unconditional standard errors (SE) for each variable are also shown.

a)

Rank	Elev	Elev^2	Insol	Insol^2	Bgr	Bgr^2	Shr	Shr^2	Heig	Heig^2	Sedf	K	QAIC$_c$	ΔQAIC$_c$	QAIC$_c$w
1	+	-	-	+			+	-	+	-	+	11	129.94	0.00	0.16
2	+	-	-	+			+	-	+		+	10	130.06	0.12	0.15
3	+	-	-		+		+	-	+		+	10	130.37	0.43	0.13
4	+	-	-				+	-	+	-	+	10	130.73	0.80	0.11
5	+	-	-				+	-	+		+	9	130.90	0.96	0.10
6	+	-	-	+			+	-			+	9	131.03	1.09	0.09
7	+	-			+	-	+	-			+	9	131.93	2.00	0.06

	1	2	3	4	5	6	7	8	9	10	11				
8	+	-			+	-	+				+	8	132.19	2.26	0.05
9	+	-	-		+		+	-			+	9	132.19	2.26	0.05
10	+	-					+	-			+	7	132.20	2.26	0.05
11	+	-			+		+				+	7	133.64	3.70	0.03
12	+		-		+		+	-	+	-	+	10	135.21	5.27	0.01
13	+	-					+				+	6	135.73	5.79	0.01
Imp	1	0.99	0.80	0.40	0.33	0.11	1	0.91	0.66	0.28	1				
Coef	31.68	-8.61	-3.74	0.58	0.02	-0.0001	0.12	-0.002	0.17	-0.005	4.10				
SE	11.72	3.24	4.92	0.79	0.05	0.0007	0.05	0.0007	0.24	0.01	0.81				

Intercept (SE) = -25.89 (11.13)

b)

Rank	Elev	Elev2	Shr	Shr2	Sedf	K	AIC$_c$	ΔAIC$_c$	AIC$_c$w
1	+	-	+	-	+	6	74.37	0	0.61
2	+	-	+		+	5	76.56	2.19	0.20
3	+	-			+	4	77.94	3.58	0.10
4	+		+	-	+	5	78.35	3.98	0.08
Imp	1	0.92	0.90	0.69	1				
Coef	31.17	-8.83	0.14	-0.002	3.92				
SE	17.13	4.53	0.09	0.001	1.41				

Intercept (SE) = -30.15 (14.16)

Table 3. Spearman's rank correlation coefficients between abundance observed in 2006, 2007, 2008 and 2006-2008 average, and (a) abundance predicted by quasi-Poisson GLM and (b) probability of occurrence predicted by binomial GLM (further details in Table 2). Results are presented for all sites in 2006 and for sites sampled in 2006-2008 (separate and averaged).

5   V: verification; CV: cross-validation. ns: $P > 0.1$; +: $P < 0.1$; *: $P < 0.05$; **: $P < 0.01$; ***: $P < 0.001$.

a) Predicted abundance

Data set	All sites			Occupied sites		
	V	CV	n	V	CV	n
2006 all sites	0.660***	0.580***	90	0.619***	0.283ns	26
2006	0.759***	0.671***	59	0.695***	0.383+	24
2007	0.669***	0.571***	59	0.589**	0.409+	20
2008	0.674***	0.581***	59	0.643**	0.553*	21
2006-2008 average	0.756***	0.669***	59	0.658***	0.361+	24

b) Predicted probability of occurrence

Data set	All sites			Occupied sites		
	V	CV	n	V	CV	n
2006 all sites	0.654***	0.571***	90	0.548**	0.561**	26
2006	0.762***	0.711***	59	0.639**	0.669***	24
2007	0.691***	0.659***	59	0.655**	0.625**	20
2008	0.702***	0.666***	59	0.680***	0.650**	21
2006-2008 average	0.766***	0.714***	59	0.679***	0.691***	24

10

Figure captions

Figure 1. Site distribution for *P. apollo* in 2006-08. Squares show 2006-2008 random sites (*n* = 43) and circles additional 2006 sites (*n* = 47) for modelling *P. apollo* distribution. Filled symbols are sites where *P. apollo* was observed, open symbols where absent. Elevation bands

5 are shown as 0.25 km increments from < 0.75 km (pale grey) to > 2 km (black). The inset map shows the geographical context of the study area in Spain. Georeferencing units are in UTM (30T).

Figure 2. Relationships between observed abundance in 2006 and predicted abundance (a:

10 verification; b: cross-validation), and predicted probability of occurrence (c: verification; d: cross-validation) (see methods for details). Spearman correlation coefficients are shown in Table 3. Empty symbols: unoccupied sites; filled symbols: occupied sites; circles: sites sampled in 2006-2008; triangles: sites sampled in 2006 or 2006-2007 only.
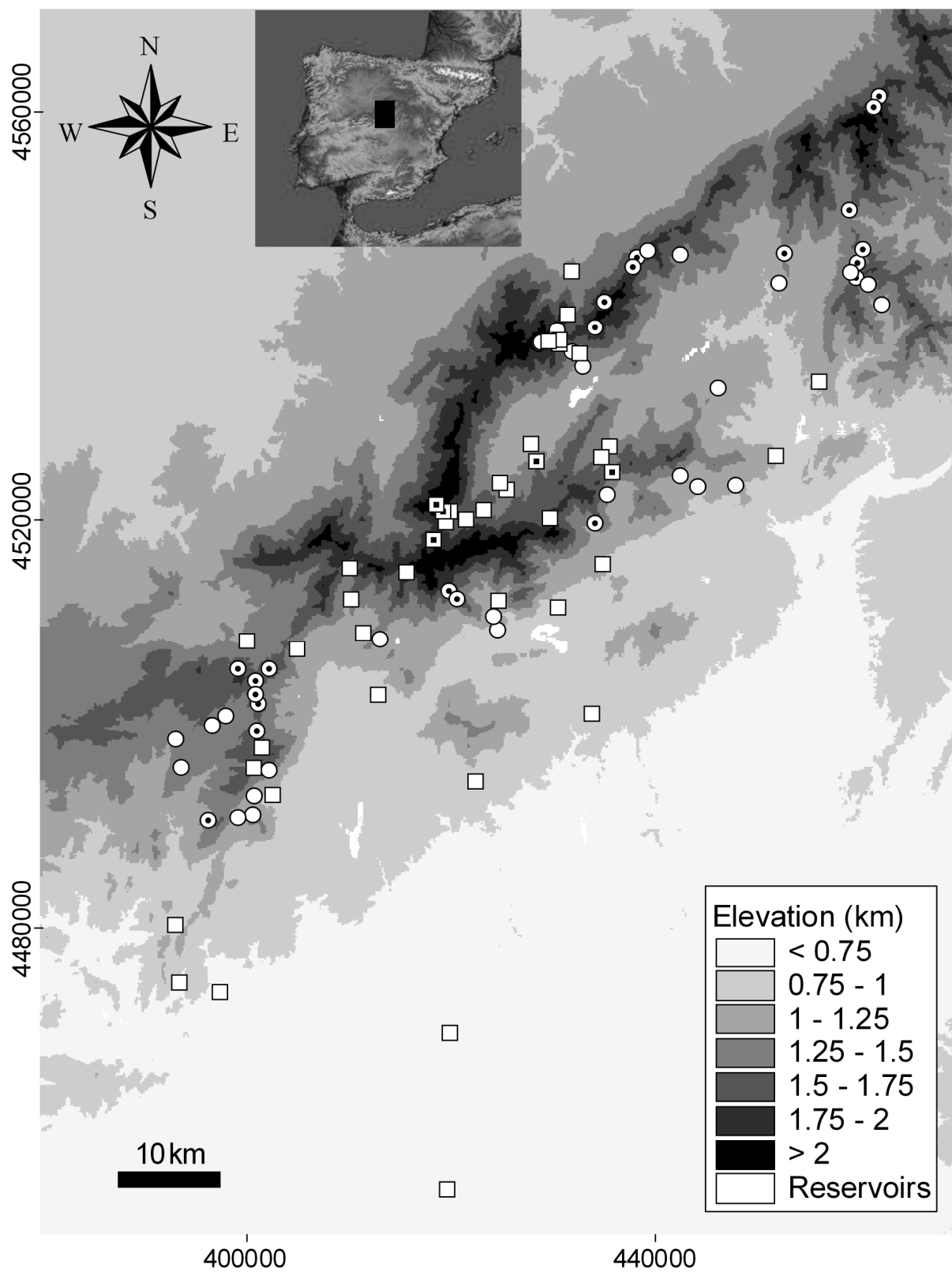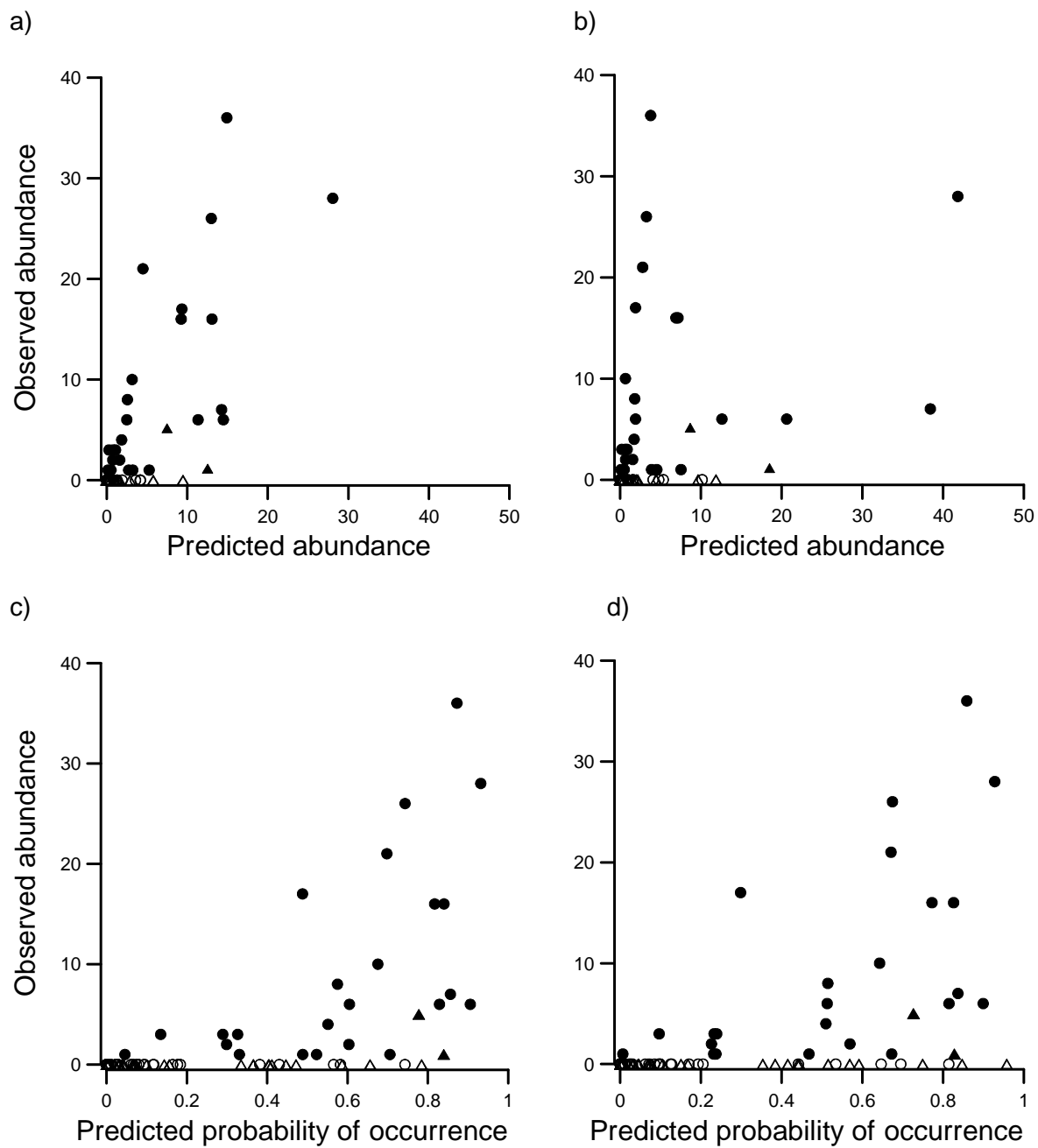
Fig. 1

Elevation (km)

- < 0.75
- 0.75 - 1
- 1 - 1.25
- 1.25 - 1.5
- 1.5 - 1.75
- 1.75 - 2
- > 2
- Reservoirs

10 km

Fig. 2

a)

b)

c)

d)

5

References

Araújo MB, Guisan A (2005) Five (or so) challenges for species distribution modelling. J Biogeogr 33: 1677-1688

Araújo MB, Williams PH (2000) Selecting areas for species persistence using occurrence data. Biol Conserv 96: 331-345

Ashton S, Gutiérrez D, Wilson RJ (2009) Effects of temperature and elevation on habitat use by a rare mountain butterfly: implications for species responses to climate change. Ecol Entomol 34: 437-446

Bartoń K. (2012) MuMIn: Multi-model inference. R package version 1.6.6. Available at http://CRAN.R-project.org/package=MuMIn

Baz A. (2002) Nectar plants for the threatened Apollo butterfly (*Parnassius apollo* L. 1758) in populations of central Spain. Biol Conserv 103: 277-282

Brommer JE, Fred MS (1999) Movement of the Apollo butterfly *Parnassius apollo* related to host plant and nectar plant patches. Ecol Entomol 24: 125-131

Brown, JH (1984) On the relationship between abundance and distribution of species. Am Nat 124: 255-279

Burnham KP, Anderson DR (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edition. Springer, New York

Cabeza M, Arponen A, Jäättelä L, Kujala H, van Teeffelen A, Hanski I (2010) Conservation planning with insects at three different spatial scales. Ecography 33: 54-63

Deschamps-Cottin M, Roux M, Descimon H (1997) Valeur trophique des plantes nourricières et préférence de ponte chez *Parnassius apollo* L. (Lepidoptera, Papilionidae). CR Acad Sci Paris, Sciences de la vie 320 : 399-406

Descimon, H., Bachelard, P., Boitier, E. and Pierrat, V. (2005) Decline and extinction of *Parnassius apollo* populations in France - continued. In: Kühn E, Feldmann R, Thomas

JA, Settele J (ed) Studies on the Ecology and Conservation of Butterflies in Europe Vol. 1: General Concepts and Case Studies, PENSOFT Publishers, Sofia, pp 114-115

Diniz-Filho JAF, Rangel TFLVB, Bini LM (2008) Model selection and information theory in geographical ecology. Global Ecol Biogeogr 17: 479-488

5   Dormann C, Purschke O, García Marquez JR, Lautenbach S, Schröder B (2008) Components of uncertainty in species distribution analysis: a case study of the great grey shrike. Ecology 89: 3371-3386

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, García Marquéz JR, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B,

10   Schröder B, Skidmore AK, Zurell D, Lautenbach S (2012) Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 35: 1-20-EV

Duff TJ, Bell TL, York A (2011) Patterns of plant abundances in natural systems: is there value in modelling both species abundance and distribution? Aust J Bot 59: 719-733

15   Elith J, Leathwick JR (2009) Species distribution models: ecological explanation and prediction across space and time. Annu Rev Ecol Evol Syst 40: 677-697

Elmendorf SC, Moore KA (2008) Use of community-composition data to predict the fecundity and abundance of species. Conserv Biol 22: 1523-1532

ESRI (2001) ArcGIS 8.1. Environmental Systems Research Institute Inc., Redlands, CA

20   Farr TG, Rosen PA, Caro E, Crippen R, Duren R, Hensley S, Kobrick M, Paller M, Rodriguez E, Roth L, Seal D, Shaffer S, Shimada J, Umland J, Werner M, Oskin M, Burbank D, Alsdorf D (2007) The Shuttle Radar Topography Mission. Rev Geophys 45: RG2004

Fordyce JA, Nice CC (2003) Variation in butterfly egg adhesion: adaptation to local host plant senescence characteristics? Ecol Lett 6: 23-27

Fred MS, Brommer JE (2003) Influence of habitat quality and patch size on occupancy and persistence in two populations of the Apollo butterfly (*Parnassius apollo*). J Insect Conserv 7: 85-98

Fred MS, Brommer JE (2005) The decline and current distribution of *Parnassius apollo* (Linnaeus) in Finland: the role of Cd. Ann Zool Fennici 42: 69-79

Fred MS, Brommer JE (2010) Olfaction and vision in host plant location by *Parnassius apollo* larvae: consequences for survival and dynamics. Anim Behav 79: 313-320

Fu P, Rich PM (2000) The solar analyst 1.0 user manual. Helios Environmental Modelling Institute, LLC, Lawrence, KS. Available a: http://www.hemisoft.com/

García-Barros E, Munguira ML, Cano JM, Romo H, Garcia-Pereira P, Maravalhas ES (2004) *Atlas of the butterflies of the Iberian Peninsula and Balearic Islands (Lepidoptera: Papilinoidea & Hesperioidea)*. Sociedad Entomológica Aragonesa, Zaragoza (Spain).

Gaston K, Blackburn TM, Gregory RD (1998) Interspecific differences in intraspecific abundance-range size relationships of British breeding birds. Ecography 21: 149-158

Gaston KJ, Blackburn TM, Grenwood JJD, Gregory RD, Quinn RM, Lawton JH (2000) Abundance-occupancy relationships. J Appl Ecol 37 (Suppl. 1): 39-59

Gomariz Cerezo G (1993) Aportación al conocimiento de la distribución y abundancia de *Parnassius apollo* (Linnaeus, 1758) en Sierra Nevada (España meridional) (*Lepidoptera: Papilionidae*). SHILAP Revta lepid 21: 71-79

Guarino ESG, Barbosa, AM, Waechter, JL (2012) Occurrence and abundance models of threatened plant species: Applications to mitigate the impact of hydroelectric power dams. Ecol Model 230: 22-33.

Guisan A, Thuiller W (2005) Predicting species distribution: offering more than simple habitat models. Ecol Lett 8: 993-1009

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecol Model 135: 147-186

Gutiérrez Illán J, Gutiérrez D, Wilson RJ (2010) The contributions of topoclimate and land cover to species distributions and abundance: fine resolution tests for a mountain butterfly fauna. Global Ecol Biogeogr 19: 159-173

Hanski I, Kouki J, Halkka A (1993) Three explanations of the positive relationship between distribution and abundance of species. In: Ricklefs R, Schlueter D (ed) Species Diversity in Ecological Communities: Historical and Geographical Perspectives, University of Chicago Press, Chicago, pp 108-116

He F, Gaston KJ (2000) Occupancy-abundance relationships and sampling scales. Ecography 23: 503-511

Holt AR, Gaston KJ, He F (2002) Occupancy-abundance relationships and spatial distribution: A review. Basic Appl Ecol 3: 1-13

Hoyle M, James H (2005) Global warming, human population pressure, and viability of the World's smallest butterfly. Conserv Biol 19: 1113-1124

Jiménez-Valverde A (2011) Relationship between local population density and environmental suitability estimated from occurrence data. Front Biogeogr 3.2: 59-61

Jiménez-Valverde A, Lobo JM (2007) Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecol 31: 361-369

Jiménez-Valverde A, Diniz F, Azevedo EB, Borges PAV. (2009) Species distribution models do not account for abundance: the case of arthropods on Terceira Island. Ann Zool Fennici 46: 451-464

Lawton JH (1993) Range, population abundance and conservation. Trends Ecol Evol 8: 409-413

Legendre P, Legendre L (1998) Numerical Ecology, 2nd English edn. Elsevier Science B.V., Amsterdam

Maclean, IMD, Wilson RJ, Hassall M (2011) Predicting changes in the abundance of African wetland birds by incorporating abundance-occupancy relationships into habitat association models. Divers Distrib 17: 480-490

McCullagh P, Nelder JA (1989) Generalized linear models, 2nd edn. Chapman & Hall/CRC, Boca Raton

McGill BJ, Etienne RS, Gray JS, Alonso D, Anderson MJ, Benecha HK, Dornelas M, Enquist BJ, Green JL, He F, Hurlbert AH, Magurran AE, Marquet PA, Maurer BA, Ostling A., Soykan CU, Ugland KI, White EP (2007) Species abundance distributions: moving beyond single prediction theories to integration within an ecological framework. Ecol Lett 10: 995-1015

Nielsen SE, Johnson CJ, Heard DC, Boyce MS (2005) Can models of presence-absence be used to scale abundance? Two case studies considering extremes in life history. Ecography 28: 197-208

Nieminen M, Nuorteva P, Tulisalo E (2001) The effect of metals on the mortality of *Parnassius apollo* larvae (Lepidoptera: Papilionidae). J Insect Conserv 5: 1-7

Oliver TH, Gillings, S., Girardello M, Rapacciuolo G, Brereton TM, Siriwardena GM, Roy DB, Pywell R, Fuller RJ (2012) Population density but not stability can be predicted from species distribution models. J Appl Ecol 49: 581-590

Pearce J, Ferrier S (2001) The practical value of modelling relative abundance of species for regional conservation planning: a case study. Biol Conserv 98: 33-43

Peterson AT, Vieglais DA (2001) Predicting species invasions using ecological niche modelling: new approaches from bioinformatics attack a pressing problem. Bioscience 51: 363-371

Pollard E, Yates TJ (1993) Monitoring butterflies for ecology and conservation. Chapman & Hall, London

Potts JM, Elith J (2006) Comparing species abundance models. Ecol Model 199: 153-163

Powney GD, Roy DB, Chapman D, Oliver TH (2010) Synchrony of butterfly populations across species ' geographic ranges. Oikos 119: 1690-1696

R Development Core Team (2011) R: a language and environment for statistical computing. R foundation for Statistical Computing. R Foundation for Statistical Computing, Vienna. Available at: http://www.R-project.org/

Richards SA (2005) Testing ecological theory using the information-theoretic approach: examples and cautionary results. Ecology 86: 2805-2814

Richards SA (2008) Dealing with overdispersed count data in applied ecology. J Appl Ecol 45: 218-227

Sagarin RD, Gaines SD (2002) The 'abundant centre' distribution: to what extent is it a biogeographical rule? Ecol Lett 5: 137-147

Sánchez-Rodríguez JF, Baz A (1996) Decline of *Parnassius apollo* in the Sierra de Guadarrama, Central Spain (Lepidoptera: Papilionidae). Holarct Lepid 3: 31-36

Sawada M (1999) Rookcase: an Excel 97/2000 Visual Basic (VB) add-in for exploring global and local spatial autocorrelation. Bull Ecol Soc Am 80: 231–234.

Sawicky P (2009) pspearman: Spearman's rank correlation test. R package version 0.2-5. Available at http://CRAN.R-project.org/package=pspearman

Schröder B, Strauss B, Biedermann R, Binzenhöfer B, Settele J. (2009) Predictive species distribution modelling in butterflies. In: Settele J., Shreeve T, Konvička M, Van Dyck H (ed) Ecology of Butterflies in Europe, Cambridge University Press, Cambridge, pp 62-77

Schweiger O, Settele J, Kudrna O, Klotz S, Kühn I (2008) Climate change can cause spatial mismatch of trophically interacting species. Ecology 89: 3472-3479

Sing T, Sander O, Beerenwinkel N, Lengauer T (2005) ROCR: visualizing classifier performance in R. Bioinformatics 21: 3940-3941

St-Louis V, Pidgeon AM, Clayton MK, Locke BA, Bash D, Radeloff VC (2010) Habitat variables explain Loggerhead Shrike occurrence in the northern Chihuahuan Desert, but are poor correlates of fitness measures. Landscape Ecol 25: 643-654

Todisco V, Gratton P, Cesaroni D, Sbordoni V (2010) Phylogeography of *Parnassius apollo*: hints on taxonomy and conservation of a vulnerable glacial butterfly invader. Biol J Linn Soc 101: 169-183

Tolman T, Lewington R. (1997) Butterflies of Britain and Europe. HarperCollins Publishers, London

Van Horne B (1983) Density as a misleading indicator of habitat quality. J Wild Manag 47: 893-901

Van Sway C, Cuttlelod A, Collins S, Maes D, López Munguira M, Šašić M, Settele J, Verovnic R, Verstrael T, Warrwn M, Wiemers M, Wynhof I (2010) European Red List of butterflies. Publications Office of the European Union, Luxembourg

VanDerWal J, Shoo LP, Johnson CN, Williams SE (2009) Abundance and the environmental niche: environmental suitability estimated from niche models predicts the upper limit of local abundance. Am Nat 174: 282-291

Venier LA, Fahrig L (1998) Intra-specific abundance-distribution relationships. Oikos 82: 483-490

Ver Hoef JM, Boveng PL (2007) Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data? Ecology 88: 2766-2772

Willis SG, Hill JK, Thomas CD, Roy DB, Fox R, Blakeley DS, Huntley B (2009) Assisted colonization in a changing climate: a test-study using two U.K. butterflies. Conserv Lett 2: 45-51

Wilson RJ, Gutiérrez D, Gutiérrez J, Martínez D, Agudo R, Monserrat VJ (2005) Changes to the elevational limits and extent of species ranges associated with climate change. Ecology Letters 8: 1138-1146

Wisz MS, Hijmans RJ, Li J, Peterson AT, Graham CH, Guisan A, NCEAS Predicting Species Distribution Working Group (2008) Effects of sample size on the performance of species distribution models. Divers Distrib 14: 763-773

5