# Comparison of single- and multi-scale models for the prediction of the *Culicoides* biting midge distribution in Germany

**Renke Lühken,[1,2] Jörn Martin Gethmann,[3] Petra Kranz,[3] Pia Steffenhagen,[4] Christoph Staubach,[3] Franz J. Conraths,[3] Ellen Kiel[2]**

***[1]Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research, Hamburg; [2]Research Group Aquatic Ecology and Nature Conservation, Department of Biology and Environmental Sciences, Carl von Ossietzky University of Oldenburg, Oldenburg; [3]Friedrich-Loeffler-Institut, Greifswald-Insel, Riems; [4]Institute of Environmental Planning, Leibniz University of Hannover, Hannover, Germany***

## Abstract

This study analysed *Culicoides* presence-absence data from 46 sampling sites in Germany, where monitoring was carried out from April 2007 until May 2008. *Culicoides* presence-absence data were analysed in relation to land cover data, in order to study whether the prevalence of biting midges is correlated to land cover data with respect to the trapping sites. We differentiated eight scales, *i.e.* buffer zones with radii of 0.5, 1, 2, 3, 4, 5, 7.5 and 10 km, around each site, and chose several land cover variables. For each species, we built eight single-scale models (*i.e.* predictor variables from one of the eight scales for each model) based on averaged, generalised linear models and two multi-scale models (*i.e.* predictor variables from all of the eight scales) based on averaged, generalised linear models and generalised linear models with random forest variable selection. There were no significant differences between performance indicators of models built with land cover data from different buffer zones around the trapping sites. However, the overall performance of multi-scale models was higher than the alternatives. Furthermore, these models mostly achieved the best performance for the different species using the index area under the receiver operating characteristic curve. However, as also presented in this study, the relevance of the different variables could significantly differ between various scales, including the number of species affected and the positive or negative direction. This is an even more severe problem if multi-scale models are concerned, in which one model can have the same variable at different scales but with different directions, *i.e.* negative and positive direction of the same variable at different scales. However, multi-scale modelling is a promising approach to model the distribution of *Culicoides* species, accounting much more for the ecology of biting midges, which uses different resources (breeding sites, hosts, *etc.*) at different scales.

Correspondence: Renke Lühken, Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Reference and Research, Bernhard-Nocht-Straße 74, 20359 Hamburg, Germany.
Tel: +49.040.42818959 - Fax: +49.04042818959.
E-mail: renkeluhken@gmail.com

Key words: Ceratopogonidae; *Culicoides*; Species distribution model; Multi-scale model.

Acknowledgements: this work was financially supported as an internal call project (IC 6.7 BT-DYNVECT) by the EU network of Excellence (EPIZONE, contract nr FOOD-CT-2006 016236). We want to thank Dr. G. Liebisch (ZeckLab), Prof. G.A. Schaub (University of Bochum), Dr. M. Geier and Dr. T. Hörbrand (Biogents), who permitted us to analyse *Culicoides* data they sampled during the German BTD-monitoring, which was funded by the German Federal Ministry of Food and Agriculture (BMEL). We like to express our sincere gratitude to Prof. Dr. H.-J. Bätza (BMEL), who initiated the German BTD-monitoring. Finally, many thanks to Dr. Ute Bradter (University of Leeds), who provided the R code for the multi-scale variable selection and the random forest variable selection as published in Bradter *et al.* (2013).

Received for publication: 6 August 2015.
Revision received: 20 October 2015.
Accepted for publication: 21 October 2015.


*Geospatial Health 2016; 11:405*
*doi:10.4081/gh.2016.405*



## Introduction

Bluetongue disease (BTD) is a reportable, non-contagious viral infection of ruminants, which occurred in Germany for the first time in late summer 2006 (Conraths *et al*., 2012). Several species of the genus *Culicoides* (Diptera: Ceratopogonidae) are considered to be potential vectors of the bluetongue virus (BTV) (Meiswinkel *et al*., 2007; Dijkstra *et al*., 2008), while the concrete vector competence of the different species is still unresolved. In 2006 and 2007, a massive spread of BTD was observed in Germany and, at the end of 2007, nearly all federal states were affected. Until spring 2008, more than 17,000 cattle, sheep, and goat died from this disease, resulting in total costs of approximately 250 million Euros. Hence, Germany decided to start a compulsory vaccination program in 2008. The recent epidemic of the Schmallenberg virus in Europe again highlights the importance of

*Culicoides* species as capable vectors (Beer *et al*., 2013), as these are also here considered to be the main vectors (De Regge *et al*., 2012; Rasmussen *et al*., 2012). Although there is huge lack of knowledge about the causal connection between environmental variables and the distribution of biting midges, several studies modelled biting midge distribution and phenology using different sets of environmental data (*e.g*. Purse *et al*., 2004, 2011; Calvete *et al*., 2008; Rigot *et al*., 2012; Kluiters *et al*., 2013). These modelling approaches used environmental data from various scales, *e.g*. all environmental data at one scale (*e.g*. 1 km, Kluiters *et al*., 2013) or at different scales (*e.g*. between 1 and 8 km, Calvete *et al*., 2008; Purse *et al*., 2011). However, as previously shown (Hamer and Hill, 2000), the selection of the spatial scales affects the outcome of the modelling, *e.g*. decrease the variance explained or bias regression coefficients, which might result in wrong conclusions and interpretations (reviewed by Bradter *et al*., 2013). Therefore, a selection of the appropriate scale is important to allow accurate species distribution modelling. Furthermore, as generally described by Bradter *et al*. (2013), species distribution can also be affected by land cover variables at multiple scales, *e.g*. if breeding sites, resting sites, and hosts of *Culicoides* biting midges are distributed over several scales. Land cover changes with distance to farm buildings, where sampling of biting midges commonly takes place (*e.g*. Kiel *et al*., 2009). The environment is generally modified most intensively around the main buildings in order to optimise farm management. The percentage of other, natural land cover variables (*e.g*. forest) increase with increasing scales around the trapping sites. This should, depending on breeding and resting sites, host preferences or species-specific flight range of *Culicoides*, result into different scale-specific variables, which are useful for the prediction of biting midge distribution. Therefore, an impact of multiple spatial scales relative to the trapping sites might be expected. A land cover variable can be a predictor at several scales. For example, when focusing on grassland as breeding site at the local scale, hosts are scattered on the grassland at a medium scale, while resting sites are on the edge of the grassland were the vegetation might be higher at the largest scale. Moreover, different variables can be predictors at different scales, *e.g*., a species breeds in the forest at a large distance, the hosts are present in direct vicinity of the trapping sites, and the resting sites are at a medium distance.

In this study, we investigated the performance of single- and multi-scale models to predict the distribution of *Culicoides* species on farms in western Germany with land cover variables at different scales. The objectives were: i) the evaluation of the spatial scales giving the best predictions for the species distribution of different biting midge species; ii) the evaluation if multi-scale models increase predictive ability; and iii) the determination of the most important landscapes variables for the prediction of *Culicoides* species distribution at the different scales.

## Materials and Methods

### *Culicoides* and landscape data

In this study, we analysed a dataset from 46 trapping sites (for trapping site information see Werner, 2010; Hoffmann *et al*., 2009), covering a gradient from northwest to southwest Germany. At every site, adult *Culicoides* were sampled for 14 months (April 2007 until May 2008). Sampling was conducted during the first seven consecutive days of each month, using the BG-sentinel trap with black light following a standardised sampling protocol (Mehlhorn *et al*., 2009). All traps were placed in the immediate vicinity of the predominant residences of cattle. The main objective of this monitoring was to document the distribution and spread of BTV, but did not concern the distribution and abundance of the biting midge vectors. Therefore, most *Culicoides* samples were sorted at the group level only. Species identification was restricted to aliquots and based on morphological characters (Werner, 2010). These aliquots were restricted to a maximum of 10% of the total *Culicoides* sample. During the monitoring, the total number of trapped *Culicoides* ranged from zero to several thousands (Mehlhorn *et al*., 2009). Therefore, the number of *Culicoides* with respect to species differed strongly between the study months and trapping sites. Thus, only aggregated presence-absence data over a time-span of 14 month were analysed in this study. Furthermore, species with a prevalence of less than 10% or more than 90% were excluded. Biting midge data were obtained from light-trap sampling and landscape variables from the Authoritative Topographic-Cartographic Information System (*Amtliches Topographisch-Kartographische Informationssystem*, ATKIS®). As little is known about the ecology and flight range of *Culicoides* biting midges, an a priori selection of the appropriate scale for the modelling of species distribution was not possible. Therefore, we extracted the same landscape variables at eight different spatial scales (radii of 0.5, 1, 2, 3, 4, 5, 7.5, 10 km), which were used separately for single-scale models or all together for multi-scale modelling approaches for the prediction of species distribution. In order to analyse the land cover of each trapping site, we referred to a selection of land cover attributes provided by ATKIS®, assumed to be important for *Culicoides* biting midges. ATKIS® provides linear and polygon vector data with a resolution of 1:5000 +/- 2.5 m positional accuracy. The same 14 landscape attributes were measured for all scales (Table 1). We extracted the percentage of

**Table 1. The ATKIS® land cover variables used for *Culicoides* species distribution modelling.**

| | | Abbrevation | Data type |
|---|---|---|---|
| Forested areas/woodland | Deciduous forest/coniferous forest (undifferentiated) | Deco | Polygon |
| | Deciduous forest | Deci | Polygon |
| | Coniferous forest | Coni | Polygon |
| | Other forest (unspecified) | Othf | Polygon |
| | Forest (sum of all forest) | Fore | Polygon |
| | Other vegetation (unspecified) | Othe | Polygon |
| Agricultural and urban | Arable land | Acre | Polygon |
| | Grassland | Gras | Polygon |
| | Garden | Gard | Polygon |
| | Fallow land | Fall | Polygon |
| | Settlement | Sett | Polygon |
| Water bodies | Ditch length | Ditc | Line |
| | Stream length | Stre | Line |
| | Water | Wate | Polygon |

surface per circular zone for each variable provided as polygon vector data and the line length per circular zone for each variable provided as linear vector data. This data collection was carried out using ArcGIS9.2 (ESRI, Redlands, CA, USA).

## Statistical analyses

### Selection of scales for the variables included in the multi-scale models

Selection of variables for the multi-scale models was applied as proposed by Bradter *et al*. (2013). This preceding reduction of variables was selected to prevent inclusion of several variables at neighbouring scales that are often highly correlated with each other (Figure 1). Another advantage of such exclusion of variables is a significant reduction of computation time.

We used univariate binomial logistic regression models for presence or absence data of each *Culicoides* species for each variable and all eight scales. Due to the small sample size (n=46), we used the corrected form of the Akaike information criterion (AICc), which indicates the best compromise between model complexity and likelihood for each model. The predictors of the different variables at the eight different scales were selected if i) the AICc was at least two lower than the AICc of the null model (intercept only); ii) the AICc was less than the next smaller or larger scale; and iii) the AICc was less than the AICc at the second smaller or larger scale (not applicable for the smallest and largest scale). With this method, we selected all local minima of the AICc, which had at least a difference of two compared to the null model for each predictor and each scale.

### Single- and multi-scale models built with model-averaging

According to methods for species distribution modelling applied in other studies (*e.g*. Kattwinkel *et al*., 2009; Gray *et al*., 2010), single-scale and multi-scale generalised linear models were built in four steps. First, highly correlated variables (Spearman's rho≥ 0.7) were excluded. For each highly correlated pair, the variable with the largest mean correlation with other variables were dropped. Second, univariate binomial logistic regression models were calculated for all variables on all scales for each species. Variables with P≥0.15 were excluded from further analysis, as they were not regarded as statistically significant. Third, multivariate binomial logistic regression models were built with every combination of the remaining variables from the previous two steps. We considered all possible models and did not use a stepwise model selection strategy, which are often criticised, *e.g*. because the results of this methods depend on the order in which variables enter the model (Burnham and Anderson, 2002; Whittingham *et al*., 2006). The large number of variables in different buffer zones per species results in a very large number of possible models. Such a brute force method might therefore not be the optimal approach (Burnham and Anderson, 2002). However, we had not enough information about the ecology of biting midges (*e.g*. breeding site preferences or resting sites) for an a priori exclusion of variables or restriction to a subset of possible models. Fourth, if several models were obtained for one species, model averaging was conducted (Burnham and Anderson, 2002). Model averaging approaches are considered to overcome problems such as overfitting or variable selection, which are found in modelling approaches aiming for a single best model (Burnham and Anderson, 2002). The Akaike weight, using AICc for calculation, can be interpreted as a measure of the strength of evidence for each model. We selected a 95% confidence set of models by sequentially summing until 0.95 was reached. According to Burnham and Anderson (2002), this set of models can be interpreted as having a 95% confidence that the best approximating model is included. The final averaged models were built by multiplication of estimated model coefficients with corresponding (see Strauss and Biedermann, 2006 for an example). Weighted coefficients were summed for each variable including all models per species on each buffer zone (single-scale models) or all buffer zones (multi-scale models).

### Multi-scale models with random forest variable selection

Although the modelling approach with averaging of multiple, generalised linear models are considered to be relative robust against overfitting (Burnham and Anderson, 2002), the large number of potential land cover variables included in the multiple models might cause such problems. Additionally, the exclusion of highly correlated variables can lead to the inclusion of variables at not meaningful, spatial scales. Therefore, but for multi-scale models only, we used a second modelling approach based on random forests for the variable selection, which was found to be robust even if the number of response data is generally small in comparison to the number of predictors (Strobl *et al*., 2007). This variable selection method was applied as described in detail by Bradter *et al*. (2013). In the random forest approach, several classification or regression trees are built from random subsets of the dataset (Breiman, 2001; Liaw and Wiener, 2002). The procedure uses a selection based on the unscaled permutation importance (Genuer *et al*., 2010). Each predictor is permutated in turn and the prediction error, *i.e*. the *out of bag* (OOB) error, before and after permutations is used as a measure of variable importance (Liaw and Wiener, 2002; Strobl *et al*., 2008). A training set is created by sampling 2/3 of the data set (with

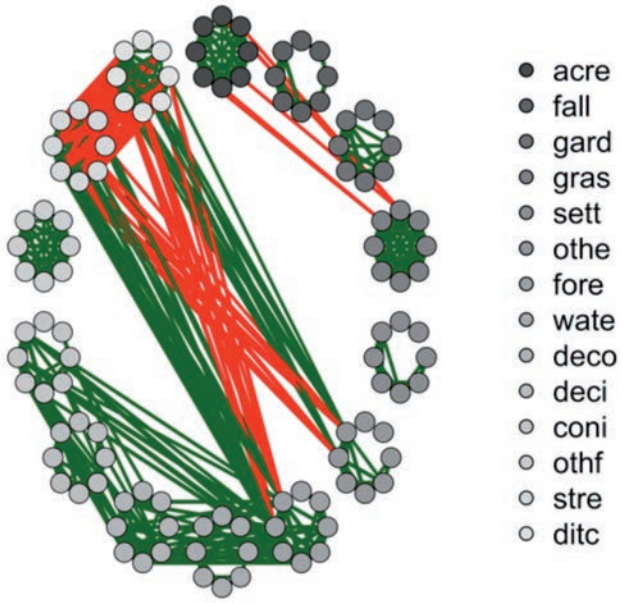


**Figure 1. Correlation network of all one hundred and fourteen predictors. The eight different scales of the fourteen variables are grouped. All correlations with a Spearman rho≥0.7 are indicated by a connection (red=negative correlation, green=positive correlation). See Table 1 for the abbreviation of the coefficients. Starting with acre on the 12 o'clock position and continuing clockwise as indicated in the legend.**

replacement) for each classification tree, which is then used to predict the remaining 1/3 of the data. The proportion of false classified classes over trees is the OOB error (Breiman, 2001; Liaw and Wiener, 2002).

There were five steps to identify the number of predictors suitable for the model interpretation (Genuer *et al*., 2010): i) all predictors were ranked by the unscaled permutation importance (average value over 50 repetitions); ii) a regression tree was fitted to the curve of the plot of standard deviations of the importance measures ordered by their mean importance, with variables with a mean importance of less than the smallest predicted value of the regression tree model discarded; iii) the OOB errors for the models (average over 50 repetitions) were computed by starting with the most important variables and adding the other predictors in order of their ranking; iv) the model with the smallest OOB error, augmented with the standard deviation of the 50 repetitions, was selected; and finally v) the nested model with OOB error smaller than this with fewer predictors was selected. Parameters which have to be specified in the random forest were used as proposed by Genuer *et al*. (2010): number of trees built in the forest $n_{tree}$=2000, the number of predictors available at each node split $m_{try}$=p/2 with p denoting the number of predictors, and error default values were used for the calculation of the OOB.

### Spatial autocorrelation

For all models built with model averaging, Moran eigenvector filtering was applied for the full model without highly correlated and non-significant variables (Powney *et al*., 2010). If significant, these eigenvectors were added to the model and included in the model averaging procedure. Furthermore, as recommended by Bradter *et al*. (2013), we applied Moran eigenvector filtering for all multi-scale models selected with random forest variable selection (Dray *et al*., 2006; Griffith and Peres-Neto, 2006). Spatial eigenvectors were added until residual spatial autocorrelation was no longer significant at the P=0.05 level.

### Performance assessment

Nagelkerke's R squared ($R^2_N$) was used as a measure of model calibration (Hosmer and Lemeshow, 2000). Area under the receiver operating characteristic curve (AUC) was used to compare prediction performance (Fielding and Bell, 1997). AUC thresholds were interpreted as proposed by Hosmer and Lemeshow (2000): 0.7-0.8 is considered an *acceptable* prediction; 0.8-0.9 is *excellent* and >0.9 is *outstanding*. Although this index is criticised as unreliable by some authors (Lobo *et al*., 2008), we predominantly referred to AUC, because it is the most commonly used performance indicator for species distribution models. However, as recommended by Lobo *et al*. (2008), we present further accuracy indices: root mean square error (RMSE), overall correct classification rate (CCR), sensitivity (SENS), specificity (SPEC), positive predictive value (PPV), negative predictive value (NPV), true skill statistic (TSS), and Cohen's kappa (KAPPA) (Liu *et al*., 2009b, for accuracy indices formulae). For threshold dependent indices (*e.g*. CCR or KAPPA) and prevalence prediction, requiring binary results, presence and absence were differentiated using a threshold value set to achieve the observed prevalence in the training data set (Freeman and Moisen, 2008).

We used bootstrapping 95% percentile confidence intervals to evaluate the statistical differences between the model performances on different scales (Pearman *et al*., 2008; Liu *et al*., 2009a). We generated 1000 bootstrap data sets (with replacement) for each species on each scale (single- and multi-scale models). Models were refitted with the

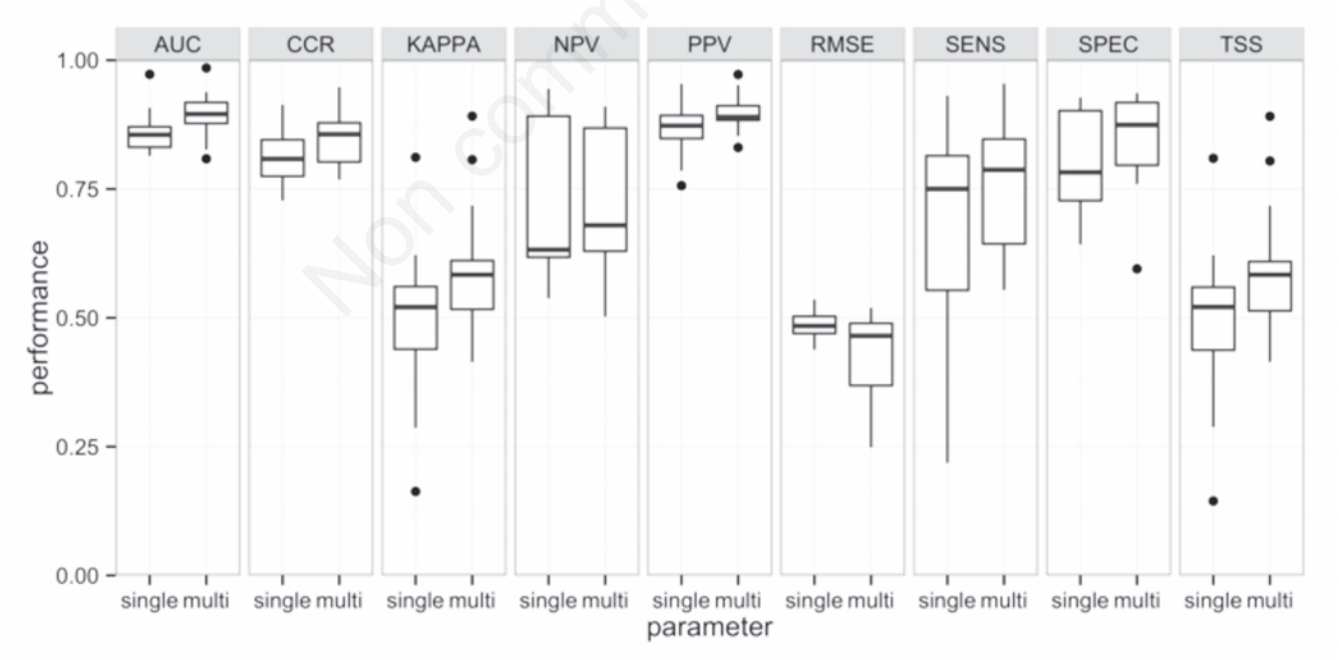


**Figure 2. Performance of the *Culicoides* species models with area under the receiver operating characteristic curve (AUC) values ≥0.7. For each performance criterion [AUC, overall correct classification rate (CCR), Cohen's kappa (KAPPA), negative predictive value (NPV), positive predictive value (PPV), root mean square error (RMSE), sensitivity (SENS), specificity (SPEC), true skill statistics (TSS)], the range and distribution of values for all models is shown. For each criterion, the left boxplot represents the single-scale model (single) and the right boxplot the multi-scale (multi) model.**

bootstrap data set. 95% confidence intervals (upper and lower 2.5% quartiles of the distribution) were calculated for each accuracy index. Non-overlapping confidence intervals were interpreted as significant differences between the scales. A threshold of 0.7 for the lower 2.5% quantile of the AUC, *i.e.* $AUC_{2.5}$, was used to select acceptable models.

### Software used

Data visualisation and statistical analyses were conducted with R (R Core Team, 2014) using functions from the packages ggplot2 (Wickham, 2009), plyr (Wickham, 2011), qgraph (Epskamp *et al.*, 2012), randomForest (Liaw and Wiener, 2002), spdep (Bivand, 2014) and the built-in function glm for logistic regression models.

## Results

Eighteen species of the 26 species in the available dataset had prevalence higher than 10% and lower 90%, and thus were used in this modelling study. From these, 57 models for thirteen species (*C. albicans*, *C.*

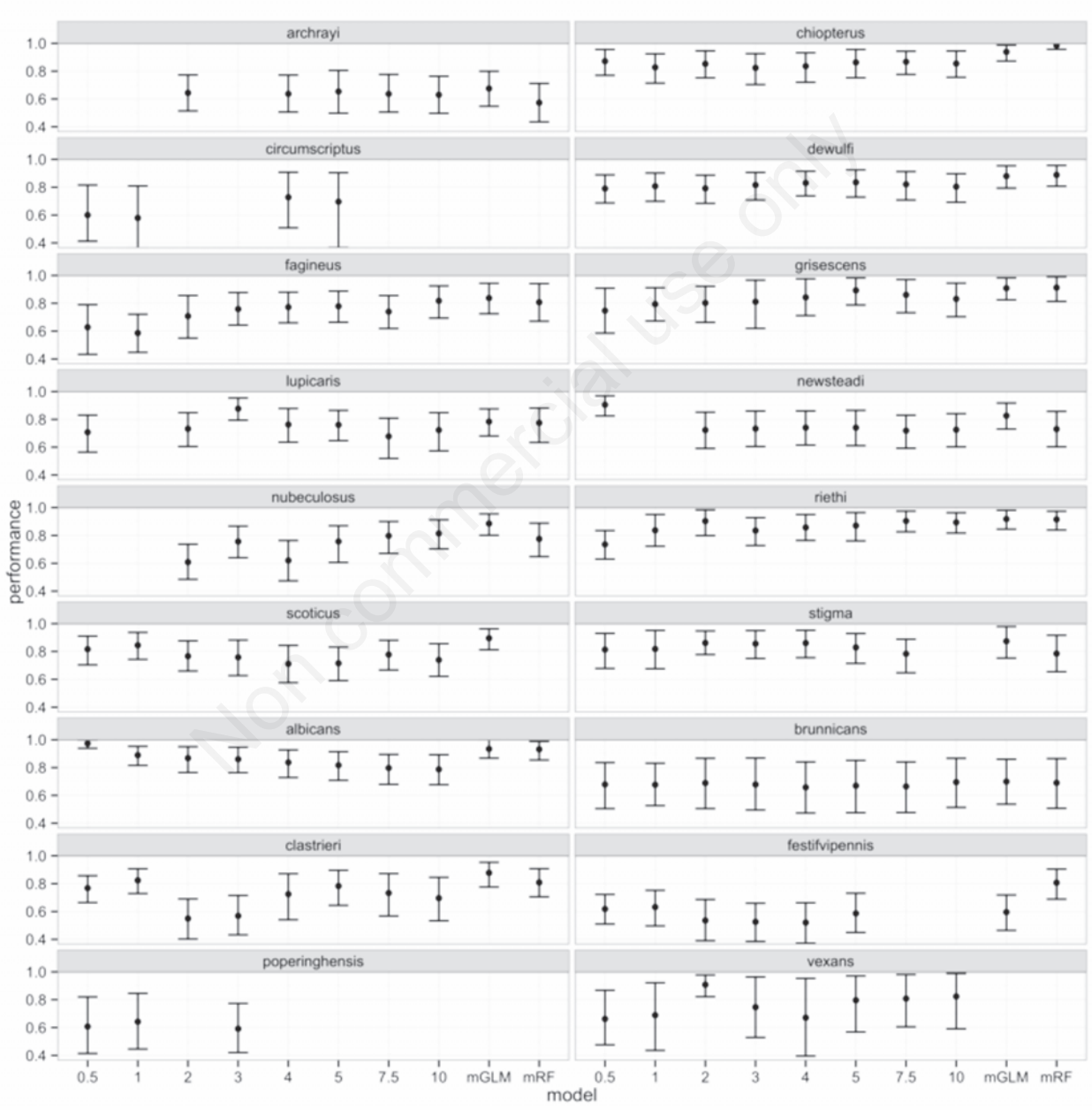




**Figure 3. Area under the receiver operating characteristic curve values with 95% bootstrapped confidence intervals. Upper and lower 2.5% quantiles of the distribution for all *Culicoides* species with prevalence between 10 and 90% and the different models are shown. Single-scale models at eight different scales and multi-scale models [multi-scale model built with model-averaging (mGLM) and random forest variable selection (mRF)] are shown.**

*chiopterus*, *C. clastrieri*, *C. dewulfi*, *C. fagineus*, *C. grisescens*, *C. lupicaris*, *C. newsteadi*, *C. nubeculosous*, *C. riethi*, *C. stigma*, *C. scoticus*, and *C. vexans*) fulfilled our performance criteria, *i.e.* at least one single- or multi-scale model with $AUC_{2.5} \geq 0.7$. Only seven of these models provided a better model fit with spatial eigenvectors, *i.e.* three for *C. albicans* (0.5 km, 2 km, and multi-scale models built with model-averaging), one for *C. lupicaris* (3 km), one for *C. newsteadi* (0.5 km) and two for *C. riethi* (1 and 2 km). This result indicates that spatial autocorrelation has little or no influence on the presence-absence at the other scales. $R^2_N$ ranged from 0.2 to 0.5, which can be considered to be good for logistic regression models (Hosmer and Lemeshow, 2000; Kattwinkel *et al*., 2009). Moreover, according to the other accuracy indices, the performance of these models was satisfactory and indicated a better prediction than occurrence by chance (Figure 2). Most of the *Culicoides* species studied here had a relative high prevalence resulting in a higher specificity and positive predictive value compared

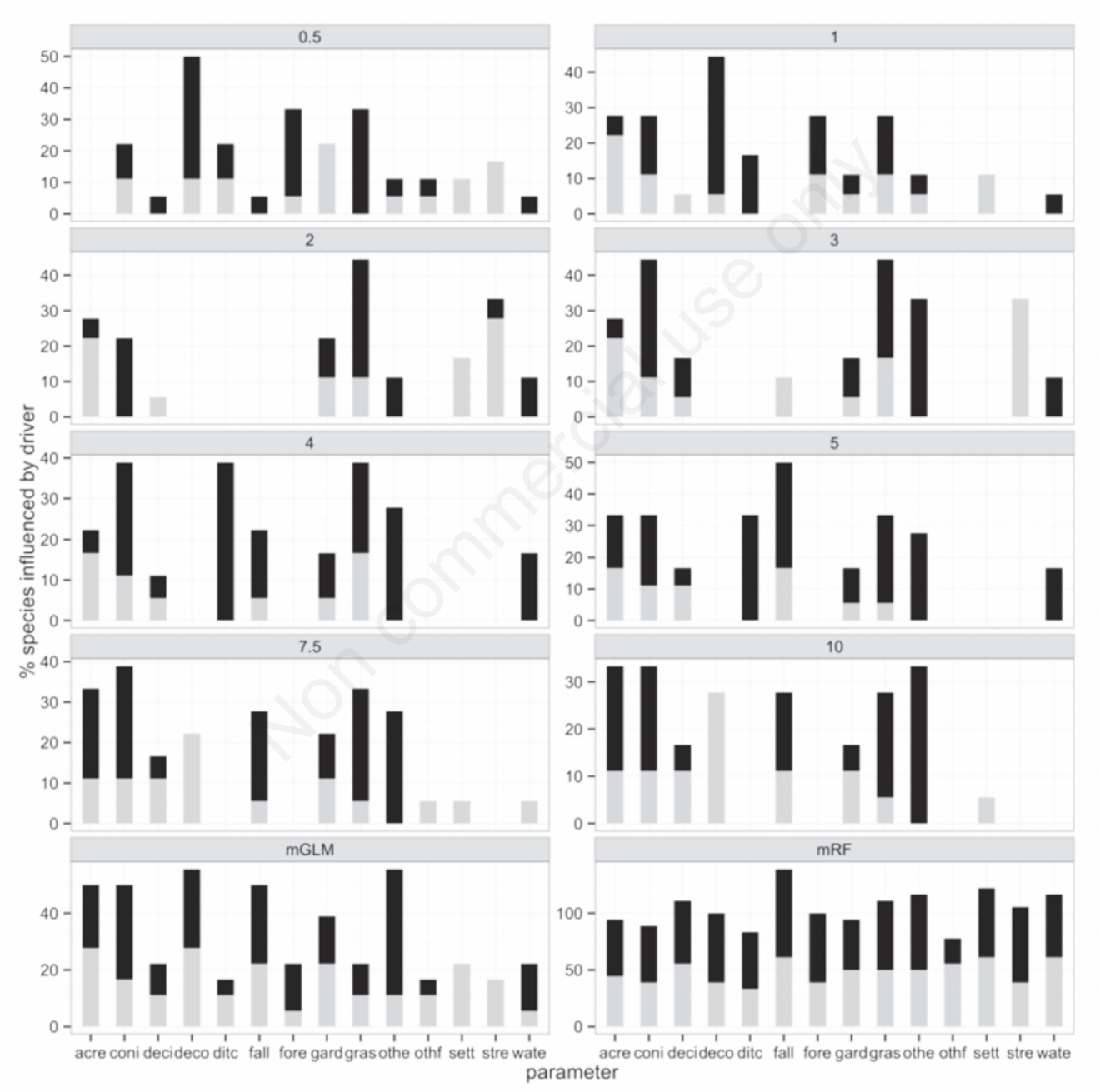


**Figure 4. Percentage of *Culicoides* species influenced by each variable in the different models. Single-scale models on the eight different scales and multi-scale models [multi-scale model built with model-averaging (mGLM) and random forest variable selection (mRF)] are shown (gray=positive coefficient, black=negative coefficient). For the abbreviations of the coefficients see Table 1.**

to sensitivity and negative predictive value.

In general, the accuracy indices did not show statistically significant differences, *i.e.* they had overlapping confidence intervals for the different species and scales (Figure 3). Nevertheless, the mean accuracy indices were overall slightly higher for the multi-scale models compared to the single-scale ones (Figures 2 and 3).

The summary of the results with the multi- and single-scale models tells us that nearly all of the species studied were influenced by *agricultural/urban* and *forest* variables, while around 50% of the species were also influenced by water-related variables. However, looking into more detail, the percentage of species showing correlations with the different land cover variables could strongly differ between the different models and scales (Figure 4), while the weights of the different variables in the models built with model averaging did not show this (Figure 5). For most of these species (9 out of 13), multi-scale models showed the best performance, *i.e.* the highest $AUC_{2.5}$ value per species. According to the mean AUC, seven of these models were characterised by excellent, another six by outstanding performance. These models were exemplarily applied for three different artificial landscapes for evaluation of the impact of the different land cover variables on the distribution of *Culicoides* species: i) increasing grassland-related and decreasing forest-related variables (Figure 6); ii) increasing arable-land type variables and decreasing forest-related ones (Figure 6); and

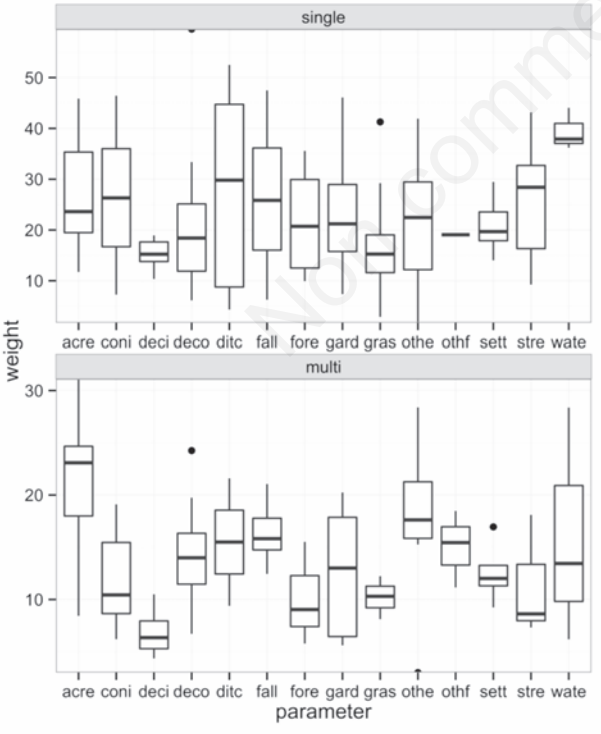


**Figure 5. Range and distribution of factor weights for *Culicoides* single- and multi-scale models. Models built with model-averaging separately shown for multi-scale and single-scale models. For the abbreviations of the coefficients see Table 1.**

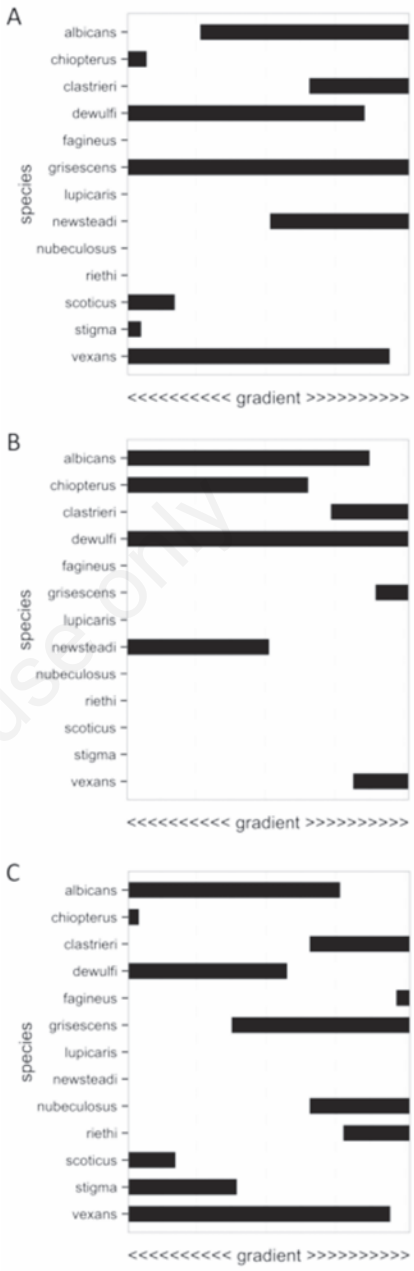


**Figure 6. The *Culicoides* species occurrence in relation to variables illustrated by single-species habitat models. Black bars signify occurrence of *Culicoides* species. The gradient refers to the 95%-percentile of the data distribution: from low (5%-percentile) to high (95%-percentile) or vice versa. A) The gradient from left to right at all scales runs from low to high values of grassland (variable grass) and from high to low values of all forest variables (coni, deci, deco, fore, othf). Water variables (ditc, stre, and wate) are fixed to high values. All other values are fixed to mean values. B) The gradient from left to right at all scales runs from low to high values of arable land (variable acre), and from high to low values of all forest variables (coni, deci, deco, fore, and othf). Water variables (ditc, stre, and wate) are fixed to high values. All other values are fixed to mean values. C) The gradient from left to right in all scales runs from low to high values of all water variables (ditc, stre, and wate). All other values are fixed to mean values. For the abbreviations of the coefficients see Table 1.**

iii) increasing water-related variables (Figure 6). Some of the *Culicoides* species responded with a wide range under these scenarios: *e.g*. *C. grisescens* in the scenario increasing grassland type of variables and decreasing forest ones (Figure 6) or *C. dewulfi* in the scenario with increasing water-related variables (Figure 6). In contrast, *C. lupicaris* did not occur under the three applied scenarios (Figure 6). In the best model, the species had a negative association with the proportion of fallow land at the 3-km scale that was not studied in the three scenarios. However, the other species showed a distinct response under at least one of the scenarios, *e.g*. *C. chiopterus*, *C. scoticus* and *C. stigma* were more restricted to the left of the gradient for the forest type of variables (low grassland/low arable-land, and high forest variables), while *C. clastrieri* were more restricted to the right end of the gradient (Figure 6).

Under the same landscape scenarios, the presence-absence predictions changed in dependence of the applied single-scale models on the different scales (Figure 7). In addition, the landscape context could be important. For example, the distribution of *C. chiopterus* is affected by grassland, but the prevalence predictions differ depending on the scale choosen (Figure 7).

## Discussion

Species distribution modelling is the most important method to predict species distribution in general, including *Culicoides* biting midges. Since the availability of digital datasets of land cover, temperature, or potential hosts is continuously increasing, several studies have also used this kind of data to predict the prevalence of biting midge species, *e.g*. the normalised difference vegetation index (NDVI) (Purse *et al*., 2004; Calvete *et al*., 2008; Kluiters *et al*., 2013) or the CORINE land cover data (Kirkeby *et al*., 2009; Purse *et al*., 2011). These data are available or used on different scales raising the question, which spatial scale, or scales, should be chosen to reach the best possible predictions for different biting midge species. At the same time, there are huge knowledge gaps on the ecology of *Culicoides* species, which would allow choosing the appropriate scale of predictors, *e.g*. missing information on the flight range or resting sites. Therefore, *a priori* selections of appropriate scaling of the variables used for *Culicoides* distribution modelling is not possible.

Active dispersal of *Culicoides* is generally expected to be limited. Kettle (1995) identifies the zone of about 500 m around the farmyard to be the most important and a substantial reduction of the number of adult *C. molestus* and *C. subimmaculatus* was achieved by measures targeting breeding sites within this radius. Furthermore, *Culicoides* abundance was found to decrease with increasing distance to potential hosts or breeding sites (Kettle, 1995; Lühken and Kiel, 2012; Rigot *et al*., 2012; Kirkeby *et al*., 2013a). Moreover, it has been proposed that the direct surroundings of farms provide a huge number of potential breeding sites (Zimmer *et al*., 2008, 2014; Foxi and Delrio, 2010; González *et al*., 2013).

In this study, all traps were placed in immediate vicinity to the predominant residences of the cattle directly on the farms, thus, it might be expected that *Culicoides* species are captured near their breeding sites and land cover information at the smaller scales around the light-traps should have the highest predictive performance. However, in the majority of cases, the model performance did not differ significantly between the models based on variables from different buffer zones. This matches the study by Kirkeby *et al*. (2013a), where the covariate *distance to the breeding site* also did not explain differences in

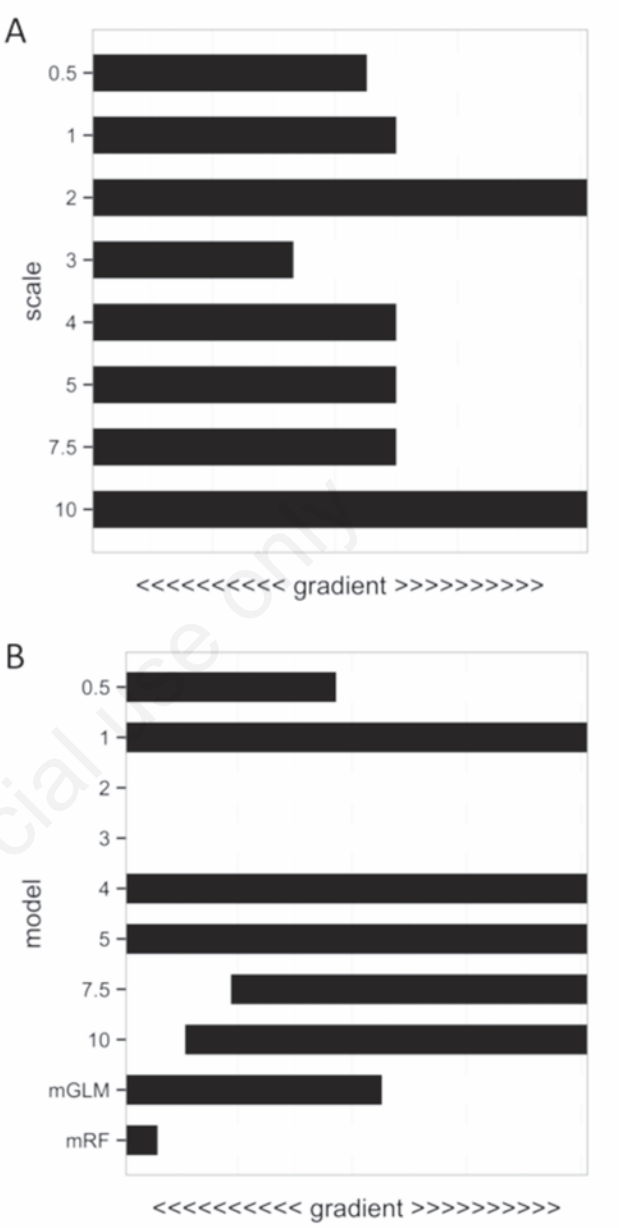


**Figure 7. Occurrence of *Culicoides chiopterus* illustrated by different habitat models at different scales (A) and illustrated by different single-scale models at different scales and multi-scale models (B). Occurrence of *Culicoides chiopterus* is signified by black bars. The gradient refers to the 95%-percentile of the data distribution: from low (5%-percentile) to high (95%-percentile) or vice versa. The multi-scale model shown in B was built with model-averaging (mGLM) and random forest variable selection (mRF). A) For each scale, the gradient from left to right runs from low to high values of grassland (variable gras), and from low to high values of water variables (ditc, stre, and wate) in the same scale are fixed to high values. Furthermore, the gradient over all scales runs from high to low values for all forest variables (coni, deci, deco, fore, and othf). All other values are fixed to mean values. B) The gradient from left to right runs from low to high values of grassland (variable gras), and from high to low values of all forest variables (coni, deci, deco, fore, othf). Water variables (ditc, stre, and water) are fixed to high values. All other values are fixed to mean values. For the abbreviations of the coefficients see Table 1.**

*Culicoides* trapping. One explanation for this result might be that the dispersal of *Culicoides* is much higher in general than generally expected. Indeed, the small number of data available from mark-release-recapture studies, indicate dispersal distances between two and six km (reviewed by Kirkeby *et al*., 2013b). Another explanation for the lack of higher performance of models at smaller scales could be the underlying data for *Culicoides*. They represent aggregated presence-absence data from a sampling conducted over several months. Therefore, the probability to trap rare *Culicoides* species might have been high. Furthermore, at this point, it must also be taken into consideration that the analysed dataset was relative small (N=46), which might result in reduced predictive accuracy (Stockwell and Peterson, 2002). It would perhaps be more appropriate to interpret the modelling results as descriptive rather than predictive (Williams and Hero, 2001; Stockwell and Peterson, 2002).

A comprehensive interpretation of our modelling results is hampered by different circumstances. According to our data, the same variables, *e.g*. the forest-related ones, had a significant correlation with each other at different scales or with other variables. At the same time, highly correlated variables should not be included in the same statistical regression models, because then small changes in the model or data can result in strong changes of the coefficient estimates (reviewed by Dormann *et al*., 2013). Therefore, as conducted in this study, it might be recommended to conduct a threshold-based pre-selection to exclude highly correlated variables. However, a preliminary exclusion of variables can result in problems regarding the interpretation of final models and omitted variables have to be considered in the conclusions to be drawn (Dormann *et al*., 2013). Furthermore, as presented in this study, several species were influenced by different land cover variables at different scales or the same variables have a different algebraic sign (positive or negative) at different scales, *e.g*. a negative correlation with forest variables in the model at the local scale and a positive correlation with forest variables in the model at a larger scale. This causes problems for the interpretation, which even increases in multi-scale models where one final model can include the same variable at different scales with different algebraic signs, *e.g*. a negative and positive correlation with the forest variable at different scales in the same model.

Our analysis was restricted to *Culicoides* presence-absence data from 46 sampling sites, as part of a wide-meshed monthly monitoring over 14 months in Germany and not primarily focused on entomological data, but virus detection in biting midges. However, additional data on species abundance or data covering longer time periods with shorter sampling intervals do not exist at present. Nevertheless, the available data give a first impression on land cover variables explaining the distribution of the German *Culicoides* fauna. Moreover, the German land cover data ATKIS® were successfully used to develop species distribution models for thirteen *Culicoides* species, including *C. chiopterus*, *C. dewulfi*, and *C. scoticus* as potential vectors of the BTV (Meiswinkel *et al*., 2007; Dijkstra *et al*., 2008) and Schmallenberg virus (Meiswinkel *et al*., 2007; Dijkstra *et al*., 2008; De Regge *et al*., 2012; Rasmussen *et al*., 2012). Furthermore, our study showed that multi-scale modelling is a promising approach to model the distribution of *Culicoides* species. Although multi-scale models did often not show significant differences compared to single-scale models, the overall performance of these models was higher. Furthermore, multi-scale models principally fulfilled the best performance for the different species using the AUC values. A multi-scale approach offers the opportunity to include a diverse set of variables at different scales. This is especially important for haematophagous insects, *e.g*. when breeding sites, resting sites or host density have to be taken into account for modelling generally distributed across several scales.

## Conclusions

Although several studies have increased our knowledge on the breeding sites and their colonisation by different *Culicoides* species (Foxi and Delrio, 2010; González *et al*., 2013; Harrup *et al*., 2013; Zimmer *et al*., 2014), the causal connections with environmental parameters mostly remain unknown. Therefore, besides the evaluation of different modelling techniques and the implementation of further environmental parameters, there is an urgent need for experimental studies on these relationships.

## References

Beer M, Conraths FJ, Van Der Poel WHM, 2013. 'Schmallenberg virus': a novel orthobunyavirus emerging in Europe. Epidemiol Infect 141:1-8.

Bivand R, 2014. Spdep: spatial dependence: weighting schemes, statistics and models. Available from: https://cran.r-project.org/web/packages/spdep/index.html

Bradter U, Kunin WE, Altringham JD, Thom TJ, Benton TG, 2013. Identifying appropriate spatial scales of predictors in species distribution models with the random forest algorithm. Methods Ecol Evol 4:167-74.

Breiman L, 2001. Random forests. Mach Learn 45:5-32.

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

Calvete C, Estrada R, Miranda MA, Borrás D, Calvo JH, Lucientes J, 2008. Modelling the distributions and spatial coincidence of bluetongue vectors *Culicoides imicola* and the *Culicoides obsoletus* group throughout the Iberian Peninsula. Med Vet Entomol 22:124-34.

Conraths FJ, Eschbaumer M, Freuling C, Gethmann J, Hoffmann B, Kramer M, Probst C, Staubach C, Beer M, 2012. Bluetongue disease: an analysis of the epidemic in Germany 2006-2009. In: Mehlhorn H, ed. Arthropods as vectors of emerging diseases. Springer, Heidelberg, Germany, pp 103-35.

De Regge N, Deblauwe I, De Deken R, Vantieghem P, Madder M, Geysen D, Smeets F, Losson B, van den Berg T, Cay AB, 2012. Detection of Schmallenberg virus in different *Culicoides* spp. by real-time RT-PCR. Transbound Emerg Dis 59:471-5.

Dijkstra E, van der Ven IJK, Meiswinkel R, Holzel DR, Van Rijn PA, Meiswinkel R, 2008. *Culicoides chiopterus* as a potential vector of bluetongue virus in Europe. Vet Rec 162:422.

Dormann CF, Elith J, Bacher S, Buchmann C, Carl G, Carré G, Marquéz JRG, Gruber B, Lafourcade B, Leitão PJ, Münkemüller T, McClean C, Osborne PE, Reineking B, Schröder B, Skidmore AK, Zurell D, Lautenbach S, 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. Ecography 36:27-46.

Dray S, Legendre P, Peres-Neto PR, 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbour matrices (PCNM). Ecol Model 196:483-93.

Epskamp S, Cramer AOJ, Waldorp LJ, Schmittmann VD, Borsboom D, 2012. qgraph: network visualizations of relationships in psychometric data. J Stat Softw 48:1-18.

Fielding AH, Bell JF, 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models.

Environ Conserv 24:38-49.
Foxi C, Delrio G, 2010. Larval habitats and seasonal abundance of *Culicoides* biting midges found in association with sheep in northern Sardinia, Italy. Med Vet Entomol 24:199-209.
Freeman EA, Moisen GG, 2008. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. Ecol Model 217:48-58.
Genuer R, Poggi J-M, Tuleau-Malot C, 2010. Variable selection using random forests. Pattern Recogn Lett 31:2225-36.
González M, López S, Mullens BA, Baldet T, Goldarazena A, 2013. A survey of *Culicoides* developmental sites on a farm in northern Spain, with a brief review of immature habitats of European species. Vet Parasitol 191:81-93.
Gray TNE, Phan C, Long B, 2010. Modelling species distribution at multiple spatial scales: gibbon habitat preferences in a fragmented landscape: gibbon distribution in fragmented landscapes. Animal Conserv 13:324-32.
Griffith DA, Peres-Neto PR, 2006. Spatial modeling in ecology: the flexibility of eigenfunction spatial analyses. Ecology 87:2603-13.
Hamer KC, Hill JK, 2000. Scale-dependent effects of habitat disturbance on species richness in tropical forests. Conserv Biol 14:1435-40.
Harrup LE, Purse BV, Golding N, Mellor PS, Carpenter S, 2013. Larval development and emergence sites of farm-associated *Culicoides* in the United Kingdom. Med Vet Entomol 27:441-9.
Hoffmann B, Bauer B, Bätza H-J, Beer M, Clausen PH, Geier M, Gethmann JM, Kiel E, Liebisch G, Liebisch A, Mehlhorn H, Schaub G, Werner D, Conraths FJ, 2009. Monitoring of putative vectors of bluetongue virus serotype 8, Germany. Emerg Infect Dis 15:1481-4.
Hosmer DW, Lemeshow S, 2000. Applied logistic regression. Wiley, New York, NY, USA.
Kattwinkel M, Strauss B, Biedermann R, Kleyer M, 2009. Modelling multi-species response to landscape dynamics: mosaic cycles support urban biodiversity. Landscape Ecol 24:929-41.
Kettle DS, 1995. Ceratopogonidae (biting midges). In: Kettle DS, ed. Medical and veterinary entomology. CAB International, Wallingford, UK, pp 152-76.
Kiel E, Liebisch G, Focke R, Liebisch A, Werner D, 2009. Monitoring of *Culicoides* at 20 locations in northwest Germany. Parasitol Res 105:351-7.
Kirkeby C, Bødker R, Stockmarr A, Enøe C, 2009. Association between land cover and *Culicoides* (Diptera: Ceratopogonidae) breeding sites on four Danish cattlefarms. Entomol Fenn 20:228-32.
Kirkeby C, Bødker R, Stockmarr A, Lind P, 2013a. Spatial abundance and clustering of *Culicoides* (Diptera: Ceratopogonidae) on a local scale. Parasite Vector 6:43.
Kirkeby C, Bødker R, Stockmarr A, Lind P, Heegaard PMH, 2013b. Quantifying dispersal of European *Culicoides* (Diptera: Ceratopogonidae) vectors between farms using a novel mark-release-recapture technique. PLoS One 8:e61269.
Kluiters G, Sugden D, Guis H, McIntyre KM, Labuschagne K, Vilar MJ, Baylis M, 2013. Modelling the spatial distribution of *Culicoides* biting midges at the local scale. J Appl Ecol 50:232-42.
Liaw A, Wiener M, 2002. Classification and regression by randomForest. R News 2:18-22.
Liu C, White M, Newell G, 2009a. Assessing the accuracy of species distribution models more thoroughly. Available from: www.mssanz.org.au/modsim09/J1/liu_c_J1a.pdf
Liu C, White M, Newell G, 2009b. Measuring the accuracy of species distribution models: a review. Available from: www.mssanz.org.au/modsim09/J1/liu_c_J1b.pdf
Lobo JM, Jiménez-Valverde A, Real R, 2008. AUC: a misleading measure of the performance of predictive distribution models. Global Ecol Biogeogr 17:145-51.
Lühken R, Kiel E, 2012. Distance from the stable affects trapping of biting midges (Diptera, Ceratopogonidae). J Vector Ecol 37:453-7.
Mehlhorn H, Walldorf V, Klimpel S, Schaub G, Kiel E, Focke R, Liebisch G, Liebisch A, Werner D, Bauer C, Clausen H, Bauer B, Geier M, Hörbrand T, Bätza H-J, Conraths FJ, Hoffmann B, Beer M, 2009. Bluetongue disease in Germany (2007-2008): monitoring of entomological aspects. Parasitol Res 105:313-9.
Meiswinkel R, van Rijn P, Leijs P, Goffredo M, 2007. Potential new *Culicoides* vector of bluetongue virus in northern Europe. Vet Rec 161:564-5.
Pearman PB, Randin CF, Broennimann O, Vittoz P, van der Knaap WO, Engler R, Lay GL, Zimmermann NE, Guisan A, 2008. Prediction of plant species distributions across six millennia. Ecol Lett 11:357-69.
Powney GD, Grenyer R, Orme CDL, Owens IPF, Meiri S, 2010. Hot, dry and different: Australian lizard richness is unlike that of mammals, amphibians and birds: hot, dry and different. Global Ecol Biogeogr 19:386-96.
Purse BV, Falconer D, Sullivan MJ, Carpenter S, Mellor PS, Piertney SB, Mordue Luntz AJ, Albon S, Gunn GJ, Blackwell A, 2011. Impacts of climate, host and landscape factors on *Culicoides* species in Scotland. Med Vet Entomol 26:168-77.
Purse BV, Tatem AJ, Caracappa S, Rogers DJ, Mellor PS, Baylis M, Torina A, 2004. Modelling the distributions of *Culicoides* bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. Med Vet Entomol 18:90-101.
Rasmussen LD, Kristensen B, Kirkeby C, Rasmussen TB, Belsham GJ, Bødker R, Bøtner A, 2012. *Culicoides* as vectors of Schmallenberg virus. Emerg Infect Dis 18:1204-5.
R Core Team 2014. R: a language and environment for statistical computing. R foundation for statistical computing, Vienna, Austria.
Rigot T, Conte A, Goffredo M, Ducheyne E, Hendrickx G, Gilbert M, 2012. Predicting the spatio-temporal distribution of *Culicoides imicola* in Sardinia using a discrete-time population model. Parasite Vector 5:270.
Rigot T, Vercauteren Drubbel M, Delécolle J-C, Gilbert M, 2012. Farms, pastures and woodlands: the fine-scale distribution of Palearctic *Culicoides* spp. biting midges along an agro-ecological gradient. Med Vet Entomol 27:29-38.
Stockwell DR, Peterson AT, 2002. Effects of sample size on accuracy of species distribution models. Ecol Model 148:1-13.
Strauss B, Biedermann R, 2006. Urban brownfields as temporary habitats: driving forces for the diversity of phytophagous insects. Ecography 29:928-40.
Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A, 2008. Conditional variable importance for random forests. Bioinformatics 9:307.
Strobl C, Boulesteix A-L, Zeileis A, Hothorn T, 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. Bioinformatics 8:25.
Werner D, 2010. Forschungsvorhaben 2808HS007. Available from: http://download.ble.de/08HS007.pdf
Whittingham MJ, Stephens PA, Bradbury RB, Freckleton RP, 2006. Why do we still use stepwise modelling in ecology and behaviour? J Anim Ecol 75:1182-9.
Wickham H, 2009. Ggplot2: elegant graphics for data analysis. Springer, New York, NY, USA.
Wickham H, 2011. The split-apply-combine strategy for data analysis. J

Stat Softw 40:1-29.
Williams SE, Hero J-M, 2001. Multiple determinants of Australian tropical frog biodiversity. Biol Conserv 98:1-10.
Zimmer J-Y, Brostaux Y, Haubruge E, Francis F, 2014. Larval development sites of the main *Culicoides* species (Diptera: Ceratopogonidae) in northern Europe and distribution of coprophilic species larvae in Belgian pastures. Vet Parasitol 205:676-86.
Zimmer J-Y, Haubruge E, Francis F, Bortels J, Simonon G, Losson B, Mignon B, Paternostre J, De Deken R, De Deken G, Deblauwe I, Fassotte C, Cors R, Defrance T, 2008. Breeding sites of bluetongue vectors in northern Europe. Vet Rec 162:131.