Monthly variation in the probability of presence of adult Culicoides populations in nine European countries and the implications for targeted surveillance

Ana Carolina Cuéllar 1*, Lene Jung Kjær 1, Andreas Baum 2, Anders Stockmarr 2, Henrik Skovgard 3, Søren Achim Nielsen 4, Mats Gunnar Andersson 5, Anders Lindström 5, Jan Chirico 6, Renke Lühken 6, Sonja Steinke 7, Ellen Kiel 7, Jörn Gethmann 8, Franz J. Conraths 8, Magdalena Larska 9, Marcin Smreczak 9, Anna Orłowska 9, Inger Hamnes 10, Ståle Sviland 10, Petter Hopp 10, Katharina Brugger 11, Franz Rubel 11, Thomas Balenghien 12, Claire Garros 12, Ignace Rakotoarivony 12, Xavier Allène 12, Jonathan Lhoi 12, David Chavencarc 12, Jean-Claude Delécolle 13, Bruno Mathieu 13, Delphine Delécolle 13, Marie-Laure Setier-Rio 14, Roger Venail 14, Bethsabée Scheid 14, Miguel Ángel Miranda Chueca 15, Carlos Barceló 15, Javier Lucientes 16, Rosa Estrada 16, Alexander Mathis 17, Wesley Tack 18 and René Bødker 1

Abstract

Background: Biting midges of the genus Culicoides (Diptera: Ceratopogonidae) are small hematophagous insects responsible for the transmission of bluetongue virus, Schmallenberg virus and African horse sickness virus to wild and domestic ruminants and equids. Outbreaks of these viruses have caused economic damage within the European Union. The spatio-temporal distribution of biting midges is a key factor in identifying areas with the potential for disease spread. The aim of this study was to identify and map areas of neglectable adult activity for each month in an average year. Average monthly risk maps can be used as a tool when allocating resources for surveillance and control programs within Europe.

Methods: We modelled the occurrence of C. imicola and the Obsoletus and Pulicaris ensembles using existing entomological surveillance data from Spain, France, Germany, Switzerland, Austria, Denmark, Sweden, Norway and Poland. The monthly probability of each vector species and ensembles being present in Europe based on climatic and environmental input variables was estimated with the machine learning technique Random Forest. Subsequently, the monthly probability was classified into three classes: Absence, Presence and Uncertain status. These three classes are useful for mapping areas of no risk, areas of high-risk targeted for animal movement restrictions, and areas with an uncertain status that need active entomological surveillance to determine whether or not vectors are present.

Results: The distribution of Culicoides species ensembles were in agreement with their previously reported distribution in Europe. The Random Forest models were very accurate in predicting the probability of presence for C. imicola (mean AUC = 0.95), less accurate for the Obsoletus ensemble (mean AUC = 0.84), while the lowest accuracy was found for the Pulicaris ensemble (mean AUC = 0.71). The most important environmental variables in the models were related to temperature and precipitation for all three groups.

(Continued on next page)
Background

Culicoides (Diptera: Ceratopogonidae) biting midges are small blood-sucking insects responsible for the transmission of viruses causing the European outbreaks of bluetongue (BT) and Schmallenberg diseases in wild and domestic ruminant livestock [1, 2], and for African horse sickness in equids [1, 3]. BTV historically made sporadic incursions into some countries of the Mediterranean Basin (Portugal, Spain, the Greek islands close to Turkey and Cyprus) but from 1998 onwards the situation worsened when five other serotypes spread within France (Corsica), Italy, Greece and countries in the Balkans region [4]. BT was never reported in northern Europe until August 2006, when an unprecedented bluetongue virus (BTV) serotype 8 outbreak started in the border region of Germany, Belgium and the Netherlands and, over the next two years, it spread further over central and northern Europe [5–8]. This epidemic had a significant economic impact within the European Union, as a consequence of the restriction of animal movements and the large amount of financial resources invested in vaccination campaigns and vector surveillance programs [9–11]. In northern Europe, the Afro-Asian vector Culicoides imicola Kieffer is absent and therefore, the vector species incriminated in the transmission of BTV were the Palaearctic species belonging to the Obsoletus ensemble Culicoides obsoletus (Meigen)/Culicoides scoticus Downes & Kettle [12, 13], Culicoides chiopterus (Meigen) [14, 15] and Culicoides dewulfi Goetghebuer [16].

Many factors contribute to the transmission of vector-borne diseases, including the presence of infected hosts, competent vectors and suitable environmental temperatures for the pathogen to replicate inside the vector [17]. In the absence of ongoing entomological surveillance, a temporal map of the potential distribution of the vectors is key for health authorities to quickly delineate possible areas and time periods of risk for disease transmission in the case of an outbreak of a known or emerging vector-borne disease [18–20]. The spatial distribution and phenology of vectors can be predicted from climate and environmental variables such as temperature, precipitation and land cover [18]. Temporal occurrence data (the presence or absence of a species at a specific time) in non-sampled areas or periods can be modelled using statistical techniques. This methodology is used to generate species distribution maps depicting the probability of the species being present at a given time [21], thus identifying areas with low or null adult activity and therefore, periods during which animal movements are safe.

Since the start of the BT outbreaks, European authorities have established a series of regulations for BT surveillance including vector monitoring to analyse the seasonal fluctuation of the vector populations and determine the seasonal vector-free periods (SVFP) for different regions [22, 23]. The EU defines SVFP by using a threshold on the abundance of female specimens, considering the parity stage of the Culicoides caught in the traps. This approach has been used to estimate the SVFP in Scotland for species of the Obsoletus group [24]. The authors estimated phenological events for each species such as the start and end of the SVFP. Brugger et al. [23] estimated vector-free periods in Austria using an approach based on the European Commission definition but without considering parity stage of female specimens. In the present study, we identified months where adult activity is null or very low, based on the monthly mean abundance for each farm, without considering the parity of the specimens collected as previously proposed by the EU legislation. Our definition of adult activity is different but comparable to the vector-free season defined by this legislation and, therefore, we keep the term “vector-free” season or period to refer to a period of the year with negligible adult activity.

The SVFP during the winter was not ubiquitous across all European countries. Austria [23], Switzerland [25] and Sweden [26] reported the existence of a SVFP, while other countries such as Germany, France, Belgium and the Netherlands reported that a SVFP might not exist in these countries [16, 27–29]. Imposing restrictions of animal movement in areas where the vector is not present has a negative economic impact as the restriction is unnecessary. On the other hand, allowing animal movement in areas where the vector is present poses a risk of spreading infections to new areas, if environmental conditions are suitable for the virus to develop inside the vector. Being able to define vector-free areas and periods is not only useful for BT management, but also for emerging
**Culicoides**-borne diseases in the future. For instance, Schmallenberg virus appeared suddenly in 2011 in Germany, and spread throughout 29 European countries [30], causing economic losses for sheep and cattle farmers [31]. In addition, the spread of African horse sickness has previously been reported in horses in Spain in 1966 and Spain and Portugal from 1987 to 1990 [32]. Knowing the geographical distribution of vectors allows veterinary authorities to focus control measurements in those areas at a specific time of year.

In this study, we used entomological data of *C. imicola*, Obsoletus ensemble and Pulicaris ensemble collected from nine European countries over a seven-year period. This entomological dataset was used previously to analyse the temporal fluctuation at different latitude bands for Europe, to analyse the start of the season at the geographical NUTS level and to interpolate the observed *Culicoides* abundance spatially [32]. In this work, we use the machine learning algorithm “Random Forest” (RF) to model the average monthly presence/absence observed and predict the probability of presence of *C. imicola*, Obsoletus ensemble and Pulicaris ensemble in unsampled areas, using climatic and environmental variables as predictors. The aim of this work was to predict areas and months likely to be free of biting midges or likely to have vectors as well as areas of uncertain status that need to be targeted for entomological surveillance in case of an outbreak. The resulting maps represent the first spatial distribution model for a transect comprising nine European countries from southern Spain to northern Sweden. The maps are useful tools as inputs for decision making by veterinary authorities to detect areas with adult activity and use this information to focus financial resources for active entomological surveillance programs.

**Methods**

**Culicoides data**

We used entomological data collected in farms from Spain, France, Germany, Switzerland, Austria, Denmark, Sweden, Norway and Poland between 2007 and 2013 as part of national surveillance programs or research projects [33]. For each trap site, observations consisted of the number of *C. imicola*, Obsoletus ensemble [C. obsoletus, *C. scoticus*, *Culicoides montanus* Shakirzjanova, *Culicoides chiopterus* (Meigen) and *C. dewulfi*] and Pulicaris ensemble [Culicoides pulicaris (Linnaeus) and *Culicoides punctatus* (Meigen)]. *Culicoides* biting midges were sampled from a total of 904 livestock farms comprising 31,429 trap collections. Onderstepoort traps were used for sampling biting midges, except for Germany (Biogents Sentinel traps) and in Spain (mini CDC traps). For these two countries, we multiplied the number of *Culicoides* for each observation by a conversion factor to make the number of specimens comparable between the different trapping methods. Details of both the sample protocols and the conversion factors used have been published previously [33].

For *C. imicola* and each of the *Culicoides* ensembles, we split the observation data set into 12 subsets according to month of the year. For each 12 monthly dataset, we calculated the average abundance on each farm for each year sampled. This resulted in 12 datasets with farms containing one monthly average abundance per year sampled. Then, we classified each monthly average each year into Presence or Absence according to the average abundance of the vector. Based on the European Union regulation [22] for the definition of the SVFP, in which an abundance threshold of biting midges is proposed to define Presence or Absence, we considered each monthly average for each year as Presence when it was above or equal to an abundance threshold of five midges for the Obsoletus and Pulicaris ensembles, and one specimen for *C. imicola*. Even though the European Union definition of Presence is based on the catch of five parous specimens per observation, we here considered the number of midges without differentiating females into their gonotrophic stage because this information was missing for some of the countries. This will result in a more conservative definition of SVFP. Our approach also differed from the approach used by the EU commission as for each farm we only classified the monthly average each year into Presence or Absence, and not each of the individual observations (when there were several observations per month).

We constructed preliminary Random Forest (RF) models using occurrence data from January and February. The data collected in this period did not include any farms from northern Scandinavia. The resulting models predicted the occurrence of biting midges in January and February in this region (data not shown). However, earlier studies have reported an absence of biting midges in the Scandinavian peninsula during winter [26, 34]. Therefore, it was useful to provide pseudo-absence points to the models in order to increase their accuracy for predicting absences in the area. For January and February, we created 11 random pseudo-absence points above 60 degrees latitude in the highlands in Norway, central and northern Sweden and Finland and were added by hand using ArcMap 10.1 (ESRI, Redlands, CA, USA) (Fig. 1).

**Predictor variables**

We used raster files (images) of 112 environmental and climatic variables, land cover and livestock density, each with a 1 km² spatial resolution.

The environmental predictors included Mid-infrared (MIR), daytime Land Surface Temperature (dLST), night-time Land Surface Temperature (nLST), Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) as predictor variables. Each variable was
derived from a MODIS temporal series from 2001 to 2012, and subjected to Temporal Fourier Analysis (FTA) [35]. For each environmental variable, the resulting products of FTA were the 14 images described in Table 1. This dataset was originally created by the TALA research group at the Department of Zoology at Oxford University, and was provided through the EDENext project [36].

We also included WORLDCLIM altitude data (digital elevation model) and bioclimatic variables as climatic predictors for Culicoides distribution. BIOCLIM images were obtained from the WORLDCLIM database [37, 38] (Table 2).

We used a Corine Land Cover (CLC) map with 250 m pixel resolution to extract information on 16 relevant land cover classes (Table 2). For each class, we created a binary image with pixel values of 1 and 0 according the presence or absence of the class. Due to the higher spatial resolution of the CLC map compared to the other predictors, we resampled each of the binary class images to a resolution of 1 km². This was done by overlaying a grid with cells of 1 km² resolution. To each of these cells, we assigned the sum of all pixels with a value of 1 within them. Each 1 km² cell of the grid was made up of 16 (4 × 4) pixels of the original CLC map. This resulted in new images for each land cover class with a pixel resolution of 1 km², representing the frequency of each of the 16 different classes found in every 1 km² area (pixel) on a scale of 0–16. CLC map was obtained from the European Environment Agency website [39].

We obtained livestock density data for cattle, goats, sheep, small ruminants and chickens from the Food and Agriculture Organization repository “GeoNetwork”. This dataset consisted of a series of raster files with information regarding livestock density at a global scale (“The gridded livestock of the world”) [40].

Modelling the probability of presence

Combining our Culicoides data with the predictors, we explored modelling approaches using VECMAP© software, v.2.0.16350.2473. For the final modelling of each month and each species, we used the Random Forest (RF) machine learning technique [41, 42] in R v.3.4.2 [43] (packages caret [44] and randomForest [45]) to model the probability of presence (PP) in the nine European countries using the Presence/Absence observations calculated at each farm. For each month we obtained a map showing the PP at the same resolution as the predictors (1 km²). The RF algorithm consists of an ensemble of decision trees used to predict the probability of class membership where the response variable is categorical (e.g. classification into presence and absence). An advantage of RF is the model’s capability of detecting nonlinear relationships between the response and the predictor variables [46] and that RF can handle a large number of predictor variables [46]. In addition, RF can produce a list of the most important predictors and scale them from 0–100 according to their importance as calculated by permuting each predictor and measuring the prediction error after the permutation [44].

The number of farms sampled varied from month to month. As expected, during summer more farms were sampled compared to winter, as in many countries of northern Europe entomological surveillance is not...
carried out during the cold winter months. For each monthly dataset, we used a stratified random split to divide the data into two subsets: one included 70% of the farms containing at least one year classified as presence together with the farms with only absence observations (training set). The second subset contained the remaining 30% of the farms as a test set to evaluate model performance [42, 47, 48]. We conducted a stratified random split based on farm ID in order to avoid having observations belonging to the same farm in both the training and the evaluation datasets (Table 3).

The number of Culicoides caught per farm highly varied between the different years. In this work, we considered each farm’s monthly classification into Presence or Absence for each year and included them in the training set as independent observations. Therefore, a farm might contain Presence and Absence observations from different years depending on the variation in mean monthly abundance between the different years.

The monthly Presence/Absence data were highly imbalanced, meaning that it contained a high proportion of one of the classes (Presence or Absence), i.e., the majority class. We investigated and compared five different balancing methods (no balancing, down-sampling, oversampling, ROSE [49], SMOTE [50], Tomek [50]) to cope with the imbalance and to improve model performance. We ran cross-validation (CV) for each balancing method 10 times with different random seeds and the best method was chosen according to highest AUC (data not shown). The balancing method chosen to balance the training set was oversampling, which entails duplicating the observations for the minority class in order to reach the same number of observations as the majority class [42]. We used the balanced training set of each month to train the RF model, and used the test sets to calculate the receiver operating characteristics (ROC) curve [42, 51, 52] and the area under this curve (AUC). We used the AUC as a measurement of model performance. AUC values close to 0.5 indicate that the model is not able to classify new samples better than random, values between 0.7 and 0.8 indicate acceptable model performance, values from 0.8 to 0.9 indicate excellent performance and values above 0.9 are considered outstanding [53]. For each month, we performed 5-fold CV to optimize the model parameter “mtry” (i.e., number of predictors used at each split). The “ntrees” parameter (number of trees of the forest) was set to 1000 trees in all cases.

For C. imicola, after the test set was created, we removed all the observations from farms not belonging to Spain or France, as the vector was not found in the seven remaining countries [33]. This reduced the large amount of Absence observations in the test set, which have an influence in the distribution of the classes.

### Classification

Classification of predicted probabilities into Presence/Absence classes can be determined using a predetermined threshold (in ecology studies, normally the default is a PP of 0.5 [54]). Here, we were interested in defining a data-dependent threshold, as a predefined threshold of 0.5 might not be optimal [54]. The monthly PP maps obtained from our RF models were classified into three categories. We calculated a lower and upper threshold and all areas with a PP below the lower threshold were considered to be in the Absence class, while the areas with a PP above the upper threshold were classified as Presence areas. Regions with a PP between the two thresholds could not be classified as either Absence or Presence class, and were therefore classified as an Uncertain status category that may be targeted for active vector surveillance. The Absence and Presence classes refer here to the occurrence of adult activity and not to the ecological establishment of the vector, as in the classical species distribution modelling.

Lower and upper thresholds were calculated using the density function for the PP predicted by the model for each test set class (true presence/absence). To define the two thresholds for each month, we derived two gain functions $G_{\text{presence}}$, $G_{\text{absence}}$ for 100 possible thresholds from 0 to 1, based on the area under the density function for Presence and Absence, respectively. We calculated $G_{\text{presence}}$ as the probability of a true presence and subtracted the probability of a misclassified presence multiplied by a parameter δ, which indicates the cost of a misclassified presence relative to a true presence. Similarly, we calculated $G_{\text{absence}}$ as the probability of a corrected classified absence (true absence)
Table 2 MODIS Fourier-transformed, BIOCLIM and Corine Land Cover predictors used to model the probability of *Culicoides* presence

<table>
<thead>
<tr>
<th>Source</th>
<th>Code</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODIS (Fourier transformed) 2001–2012</td>
<td>MIR</td>
<td>Mid-infrared</td>
</tr>
<tr>
<td></td>
<td>dlST</td>
<td>Daytime land surface temperature</td>
</tr>
<tr>
<td></td>
<td>nLST</td>
<td>Nighttime land surface temperature</td>
</tr>
<tr>
<td></td>
<td>NDVI</td>
<td>Normalized difference vegetation index</td>
</tr>
<tr>
<td></td>
<td>EVI</td>
<td>Enhanced vegetation index</td>
</tr>
<tr>
<td>BIOCLIM 1960–1990</td>
<td>BIO 1</td>
<td>Annual mean temperature</td>
</tr>
<tr>
<td></td>
<td>BIO 2</td>
<td>Mean diurnal range: mean of monthly (max. temp - min. temp)</td>
</tr>
<tr>
<td></td>
<td>BIO 3</td>
<td>Isothermality (BIO2/BIO7) ÷100</td>
</tr>
<tr>
<td></td>
<td>BIO 4</td>
<td>Temperature seasonality (standard deviation × 100)</td>
</tr>
<tr>
<td></td>
<td>BIO 5</td>
<td>Maximum temperature of warmest month</td>
</tr>
<tr>
<td></td>
<td>BIO 6</td>
<td>Minimum temperature of coldest month</td>
</tr>
<tr>
<td></td>
<td>BIO 7</td>
<td>Temperature annual range (BIO5-BIO6)</td>
</tr>
<tr>
<td></td>
<td>BIO 8</td>
<td>Mean temperature of wettest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 9</td>
<td>Mean temperature of driest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 10</td>
<td>Mean temperature of warmest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 11</td>
<td>Mean temperature of coldest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 12</td>
<td>Annual precipitation</td>
</tr>
<tr>
<td></td>
<td>BIO 13</td>
<td>Precipitation of wettest month</td>
</tr>
<tr>
<td></td>
<td>BIO 14</td>
<td>Precipitation of driest month</td>
</tr>
<tr>
<td></td>
<td>BIO 15</td>
<td>Precipitation seasonality (coefficient of variation)</td>
</tr>
<tr>
<td></td>
<td>BIO 16</td>
<td>Precipitation of wettest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 17</td>
<td>Precipitation of driest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 18</td>
<td>Precipitation of warmest quarter</td>
</tr>
<tr>
<td></td>
<td>BIO 19</td>
<td>Precipitation of coldest quarter</td>
</tr>
<tr>
<td></td>
<td>Altitude</td>
<td>Digital elevation model (DEM)</td>
</tr>
<tr>
<td>Corine Land Cover&lt;sup&gt;a&lt;/sup&gt;</td>
<td>CLC 12</td>
<td>Non-irrigated arable land</td>
</tr>
<tr>
<td></td>
<td>CLC 13</td>
<td>Permanently irrigated land</td>
</tr>
<tr>
<td></td>
<td>CLC 15–17</td>
<td>Vineyards, fruit trees and berry plantations, olive groves</td>
</tr>
<tr>
<td></td>
<td>CLC 18</td>
<td>Pastures</td>
</tr>
<tr>
<td></td>
<td>CLC 19</td>
<td>Annual crops associated with permanent crops</td>
</tr>
<tr>
<td></td>
<td>CLC 20</td>
<td>Complex cultivation patterns</td>
</tr>
<tr>
<td></td>
<td>CLC 21</td>
<td>Land principally occupied by agriculture with significant areas of natural vegetation</td>
</tr>
<tr>
<td></td>
<td>CLC 22</td>
<td>Agro-forestry areas</td>
</tr>
<tr>
<td></td>
<td>CLC 23</td>
<td>Broad-leaved forest</td>
</tr>
<tr>
<td></td>
<td>CLC 24</td>
<td>Coniferous forest</td>
</tr>
<tr>
<td></td>
<td>CLC 25</td>
<td>Mixed forest</td>
</tr>
<tr>
<td></td>
<td>CLC 26</td>
<td>Natural grasslands</td>
</tr>
<tr>
<td></td>
<td>CLC 29</td>
<td>Transitional woodland-shrub</td>
</tr>
<tr>
<td></td>
<td>CLC 35</td>
<td>Inland marshes</td>
</tr>
<tr>
<td></td>
<td>CLC 40</td>
<td>Water courses</td>
</tr>
<tr>
<td></td>
<td>CLC 41</td>
<td>Water bodies</td>
</tr>
</tbody>
</table>

<sup>a</sup>CLC plus the number refers to the CORINE land cover class used for modelling.
and subtracted the probability of misclassified absence multiplied by parameter \( \gamma \), which indicates the cost of a misclassified absence relative to a true absence. Setting \( \delta = 2 \), for example, means that the cost of a false positive classification is twice the gain of a true positive classification. The gain value can be considered in terms of timely initiation of countermeasures and a lower probability of an epidemic and trade restrictions, while the loss value would be the cost to the farm and society of incorrectly applied countermeasures. Similarly, for the interpretation of \( \gamma \), the gain of a true negative classification and the loss from a false negative classification can be likened to being declared free from disease, with the cost to both farmer and society of a subsequent discovery of the disease. Similar considerations can be used to relate \( \delta \) and \( \gamma \) to each other. If, for example, we set \( \delta = \rho \times \gamma \) in Eq. 1, the cost of misclassifying a presence is \( \rho \) times the cost of misclassifying an absence. We assign \( \delta = 2 \times \gamma \) in order to assign twice the importance to the Presence misclassifications compared to Absence misclassifications and we set \( \gamma = 2 \) to still give some importance to the Absences misclassifications.

The equations for \( G_{\text{presence}} \) and \( G_{\text{absence}} \) were:

\[
G_{\text{presence}} (q) = \int_{q}^{1} \text{Presence} (x) \, dx - \delta \int_{q}^{1} \text{Absence} (x) \, dx
\]

(1)

\[
G_{\text{absence}} (q) = \int_{0}^{q} \text{Absence} (x) \, dx - \gamma \int_{0}^{q} \text{Presence} (x) \, dx
\]

(2)

where \( q \) represents the possible threshold value between 0 to 1, and where \( \delta \) and \( \gamma \) are loss parameters.

To calculate the lower threshold, we used Eq. 1 to find the optimal upper threshold when assuming a loss parameter of \( \delta = 4 \) by optimizing the gain \( G_{\text{presence}} \). Similarly, Eq. 2 was used to find the optimal lower threshold, assuming a loss parameter \( \gamma = 2 \). The upper and lower thresholds depend on the predictive power of the model, being more separated when the overlapping between classes is large. If the model performance is good, the overlapping between classes will be less and the two thresholds will be closer together.

In order to evaluate the sensitivity of the thresholds to the distribution of different test sets, we divided each monthly test set into ten equally sized folds (10 subsets) and calculated the density functions using nine out of the ten folds. This procedure was repeated for all the different folds (10 times), excluding a different fold each time, and plotted the new lower and upper thresholds together in the same graph. We applied this 10-fold cross-validation scheme to compare the threshold calculated with different subsets of the test set versus the thresholds calculated using all the observations of the test set.

We classified the monthly probability maps into the three classes: “Absence”, “Uncertain” and “Presence” using the thresholds calculated from all the observations of the test set.

### Results

#### Obsoletus ensemble

The 12 models were shown to perform well for the Obsoletus ensemble, with an AUC ranging from 0.76 in June and December to 0.91 in November (mean AUC = 0.84) (Fig. 2).

The majority class shifted from Absence in December-March, to Presence in April-November, and the models generally had good predictive power when predicting the majority class. However, the models...
performed less well when predicting the minority class. For January and February, the model predicted the Presence class relatively poorly, with a relatively flat density function (Fig. 3). The additional thresholds calculated using 10-fold CV were similar to the main threshold, indicating that the distribution of classes in the test set were robust when subtracting 10% of the data. The lower thresholds showed more variation compared to the variation of the upper thresholds (Fig. 3).

Classifications did not result in clearly delineated geographical zones for the three classes (Presence, Absence and Uncertain), although spatial patterns were observed (Fig. 4). In January, the Obsoletus ensemble was predicted present in areas within the western part of France, northern coast of Spain and in scattered areas of Germany, and it was predicted absent from northern and central Scandinavia, eastern France and parts of Germany. The Uncertain class area was present in southern Scandinavia, eastern Germany and Poland. In February the Presence area in western France and the northern coast of Spain appeared clearly segregated while more dispersed patches appeared in Germany and Poland. The Uncertain class area was reduced to patches in Germany, Poland and a small portion of southern Sweden. During March, the Presence area extended further west into France, while the Absence area was clearly concentrated in the eastern part of Europe and
Scandinavia. The Uncertain area was a more coherent intermediate region between these two areas, found in eastern France, Belgium, and the Netherlands. In April, the Presence class expanded from western France occupying most of France while the eastern part of the study area and Scandinavia remained in the Uncertain area. From May onwards, the general pattern showed the Obsoletus ensemble to be widely distributed in France, Germany, Austria, Switzerland, Poland and southern Scandinavia. The Absence class areas were located in southern Spain during this period. In November, Scandinavia was classified as an Absence class area together with Spain (except the northern coast of Spain, that was included in Presence area). Finally, in December the Presence class was clustered in western France and some patches in northern Germany while the remaining areas, with exception of southern Spain, appeared classified as Uncertain areas, including the Scandinavian peninsula (Fig. 4).

Pulicaris ensemble

The RF models performed less well in predicting the PP for the Pulicaris ensemble. The mean AUC was 0.81, ranging from 0.69 in April to 0.92 in December (Fig. 5).

For January, the test set contained only three Presence observations from a single farm and the density function and thresholds could not be calculated. Therefore, the PP map could not be classified into the three classes. For February, the PP predicted for the observed Presences were completely included within the range of the PP predicted for the Absence class, meaning that the model was incapable of distinguishing the Presence class. Nevertheless, because both density functions were computed, the lower and upper thresholds were still calculated. The distribution of predicted Presence and Absence areas for the Pulicaris ensemble test set contained larger overlapping areas between both distributions than for the Obsoletus ensemble, resulting in poorer predictive power for distinguishing between the classes. For the months of April, May and June, the distribution of both classes overlapped so much that the lower threshold was calculated as close to 0 to avoid false negative classifications (Fig. 6). For the Pulicaris ensemble, the additional thresholds calculated using 10-fold CV, were similar to the main threshold for all the months, meaning that the distribution of classes in the test set were robust when subtracting 10% of the data. Both lower and upper thresholds seemed to be robust for the different test sets (Fig. 6).

Due to the lack of Presence observations in January, we could not define thresholds for classifying the PP map. In February, because PP of the observed Presence observations were completely included in the range of the PP of the Absence class, we decided not to classify the map as the model was incapable of distinguishing the Presence class and would lead to an incorrect interpretation of the classification. In March, the Pulicaris ensemble was predicted to be present on the west coast of France, northern coast of Spain and in central and northern Scandinavia, while the Absence class was distributed in eastern France, Germany and Poland. The Uncertain area was located between the Presence and Absence class. During April, May and June, the model was able to predict the Presence class
but it was incapable of distinguishing the Absence class, resulting in classification only for the Presence and Uncertain class. From July to October, the Presence class extended towards the eastern part of the study area while the Uncertain class occupied northern Scandinavia. During September, the Uncertain class was additionally found in France. In November, the Presence areas were located mostly in Germany and some patches in France while Scandinavia was classified into the Uncertain class. The Absence class was predicted in Denmark and southern Spain. During December, the Absence class was localized in Spain, France and northern Scandinavia while the Presence class remained in some patches in Germany (Fig. 7).

**Culicoides imicola**
The RF models for *C. imicola* had a very high accuracy for distinguishing the Presence and Absence classes. The models had a mean AUC of 0.95, ranging from 0.92 in January to 0.97 in August (Fig. 8).

The RF models predicted the *C. imicola* Absence class very well. Absence constituted the majority class for all months as the species was only found in Spain and southern France. The Presence class was less well predicted, as reflected in a flatter distribution. Nevertheless, the model was able to distinguish both classes, resulting in a narrow area of uncertainty between the lower and upper thresholds (Fig. 9). The additional thresholds calculated using
10-fold CV, were similar to the main threshold, indicating that the distribution of classes in the test set were robust when subtracting 10% of the data. The upper thresholds showed more variation compared to the variation in the lower thresholds. Particularly April, July and November seemed to have upper thresholds sensitive to the class distribution of the test set (Fig. 9).

Compared to the models for the Obsoletus ensemble, the models for *C. imicola* resulted in a clearer geographical division into three separate coherent zones. *Culicoides imicola* was found to be present in January and February in some areas in southern Spain, the Balearic Islands and Corsica. Uncertain areas were identified in central Spain, while the Absence regions were located in northern Spain and most of France with the exception of the southern coast. From March onwards, the Presence region extended northwards, occupying the southern and central regions of Spain until October, when it retracted back to the southern coast of Spain during late autumn. On Corsica, the Presence areas were located around the coast, with the vector being absent inland. The Uncertain area was always clearly located between the Presence and Absence areas and was generally small due to the high accuracy of the model in distinguishing between Presence and Absence classes (Fig. 10).

**Fig. 5** Predicted monthly probability of presence of Pulicaris ensemble. Monthly model performance is shown as the AUC value.
Important predictors
The most important predictors driving the distribution of the Obsoletus ensemble, Pulicaris ensemble and C. imicola were related to temperature and precipitation for most months (dLST_MN, nLST_A0, nLST_MX, BIO 10, BIO 18, BIO 5). EVI- and NDVI-derived variables were the most important for some months and for some of the taxa, but with lesser importance compared to temperature and precipitation. Corine land cover classes were not selected as important variables and only one class (CLC 12: non irrigated arable land) was selected for Pulicaris during August. A similar situation occurred for the animal density variables, in which the only variable appearing in the top 5 most important variables was sheep density for the Pulicaris ensemble. Altitude was selected as an important variable only for the Obsoletus and Pulicaris ensembles, for the month of December (Additional file 1).

Discussion
This study was based on the most extensive Culicoides dataset created to date. For these prediction maps, we used 31,429 Culicoides trap catches from nine European countries from 2007 to 2013 [33]. The objectives of this work were to predict the monthly probability of Culicoides presence and to demarcate regions of Europe into three presence classes, each for C. imicola and the Obsoletus and Pulicaris ensembles. We also identified areas and periods when the model was not able to predict with reasonable certainty. In these areas, targeted entomological surveillance programs implemented by the CVO’s of European Union member states are needed to clarify the present entomological status in case of an outbreak. The maps presented here can be used to determine vector-free areas (Absence areas) and areas where the vector can be found. The Absence and Presence areas were delimited to minimize misclassification errors, making these classes more accurate in terms of the occurrence of Culicoides.

The models generated for the Obsoletus ensemble performed well for all months, and we were able to detect a spatial pattern in the three classes. However, the Absence and Presence classes were not completely separated by the model, and some geographical areas with Uncertain status were found among the Presence or Absence areas. For some of the months, our RF models were not able to clearly distinguish the minority class from the majority class, resulting in the threshold from the gain function being moved to the extremes to avoid misclassifications. This, in turn, resulted in a large Uncertain area that should potentially be targeted for costly entomological surveillance. This was the case for the Obsoletus ensemble during August, when the vector was indeed present in most of Europe but where our models classified the status as Uncertain in many smaller areas. For instance, in December the model predicted a large Uncertain status area that occupied most of the Scandinavia peninsula while the cold winter conditions make it unlikely that specimens will be found in northern Scandinavia. The Uncertain status areas should be interpreted with care and expert knowledge must be considered.
when making decisions regarding implementation of surveillance programs. The maps presented here are merely intended as tools and inputs to decision makers for long-term planning and in case of outbreaks in areas without ongoing entomological surveillance. The presented maps are based on a given gain function, but the gain function should reflect the severity of the vector borne diseases with an increasing emphasis on sensitivity as the severity of a disease increases.

In our models, the most important variables for the Obsoletus ensemble were the minimum daytime land surface temperature in January and February, and temperature- and precipitation-related variables (BIO 5 and BIO 14) throughout the rest of the year. Our results are in agreement with the findings of Calvete et al. [55] and Ducheyne et al. [56] who stated that temperature-related variables were the most important for the Obsoletus group distribution in Spain. Additionally, Purse et al. [57] found that temperature had an effect in the occurrence of *C. obsoletus* in Italy. The Obsoletus ensemble are Palaearctic species requiring relatively low temperatures and humid climates for optimal development and survival [58, 59]. Temperature plays an important role in *Culicoides* ecology as it determines the seasonal fluctuation of the vector populations [60, 61], while humidity has been reported to create the optimal conditions for *C. obsoletus* breeding sites (e.g. dung heaps) [62].

Fig. 7 Classification of the predicted probability of presence of Pulicaris ensemble into Absence, Presence and Uncertain areas at a 1 km² resolution.
To date, maps showing the PP and distribution of the Obsoletus ensemble for the entire Europe are scarce and incomplete. EFSA developed a website displaying distribution maps of Culicoides spp. On this site, a map of C. obsoletus/C. scoticus shows the distribution of this species [63] but the map is lacking information from some countries in Europe. At country level, some studies predicted the probability of Obsoletus group presence based on entomological data collected [56, 64–66]. Therefore, there is a need for predictions on a continental scale summarizing historical surveillance data to allow CVO’s of EU Member States to make rapid decisions in case of a future outbreak, as it would provide them with information on which areas and which time periods are likely to be vulnerable, which are likely to be safe and where the resources for surveillance should be allocated.

The RF models for the Pulicaris ensemble had poorer predictive power compared to Obsoletus ensemble and C. imicola. The abundance of the Pulicaris ensemble was ten-fold less than the abundance of the Obsoletus ensemble [33]. This lead to a lower number of Presence farms and, therefore, when the data were split into training and test sets, only a few Presence points were present in the test set. This resulted into heavily imbalanced monthly datasets e.g. February only included three farms with Presence observations in the test set. It is not...
recommended to assess model performance based only on a couple of observations from a certain class because it might lead to results with high variability. *Culicoides pulicaris* (*sensu stricto*) has been implicated in BTV transmission [67], but the Pulicaris ensemble species is not thought to have played a significant role in the 2006 BT outbreak in northern Europe [16]. Nevertheless, species of this ensemble might play a role in future outbreaks of emerging infections.

The model performance for *C. imicola* was highly accurate, with high AUC values for all months, indicating that this species has particular environmental requirements that can be detected through satellite imagery. This is likely to be related to hot and dry summers with low seasonal variation [64]: characteristic of the Mediterranean basin. The three classes were clearly distinguishable in the maps, and Presence and Uncertain areas could be delimited to the Iberian Peninsula. *Culicoides imicola* maps can be used directly to allocate resources for surveillance programs or to determine appropriate animal movement restrictions.

In our models, the most important explanatory variables for classification of areas for the Presence/Absence affecting *C. imicola* distribution were related to temperature and precipitation. We found that during winter, the mean temperature of the coldest quarter was the variable driving the presence of *C. imicola*, while variables related to precipitation were the most predominant drivers during the warmer months. This is in accordance with the results of previous studies [56, 64, 68].

The distribution of *C. imicola* has previously been modelled at continental level using classical statistical models fitted to data collected from single European countries [57, 64, 69]. In our maps, *C. imicola* appeared to be present all year round, as it can be found on the southern coast of Spain during January and February. This agrees with previous analyses of the start of the vector season in Europe, where *C. imicola* was found to be present during the winter months in southern Spain and central and southern Portugal [65]. The predicted probability of presence shown in our maps are in agreement to the distribution models made for Spain by Ducheyne et al. [56], Calvete et al. [55] and Peters et al. [70], and for France, where the Presence areas for the species are mainly located in the coastal regions of Corsica and VAR department [15].

In our study, we used *Culicoides* data aggregated into groups, namely the Obsoletus and Pulicaris ensembles. Aggregating species into a single group, or ensemble, might represent a challenge for ecological modelling, as the different species might require different environmental conditions and phenology differ between them. This has been studied by Searle et al. [24], who estimated the start and end date of the vector season and length of the vector-free period for four species of the Obsoletus ensemble. They observed that there were differences in phenology among the species. The lower model performance obtained for Obsoletus and Pulicaris ensembles compared to *C. imicola* may reflect that different species within each ensemble have different phenology and different environmental drivers. It
would therefore be useful to identify *Culicoides* specimens to the species level. Molecular techniques, such as high-throughput real-time RT-PCR assays, can be used in a fast way for species identification. More accurate results could be expected if modelling is carried out on individual species data.

In practice, maps based on the classifications made for each 1 km² pixel might be difficult to use for decision making, as it becomes challenging to define classes for larger areas in which pixels from different classes are found. For practical use, predicted pixel values may therefore be summarized by area, such as at NUTS level (nomenclature of territorial units for statistics) defined by Eurostat (2013). This would facilitate the implementation of control and surveillance programs by European veterinary authorities.

Random Forest is a machine learning technique that has previously been used for ecological species modelling [19, 56, 70–75]. This technique has been proven to perform better compared to other applications of classical statistical methods such as Non-Linear Discriminant Analysis and Generalized Linear Models [19, 71], as well as Linear Discriminant Analysis, logistic regression [70, 74] and Additive Logistic Regression [75]. In this work, the monthly predicted probability of *Culicoides* presence had medium-high accuracy, but it is
important to keep in mind that there might be other variables that cannot be captured by satellite imagery and that may have an influence on the occurrence of these species on a local scale, such as soil conditions (affecting breeding sites) and farming practices. Nevertheless, for some months, our models performed slightly better than other RF models used for predicting the occurrence of biting midges and mosquitoes [70, 71]. This highlights the challenges faced in predicting the occurrence of insect vectors using remote sensing data, as vectors are highly influenced by local microenvironments [76] and these data are difficult to obtain from satellite images without high spatial resolution.

Conclusions
We present here maps as a risk assessment tool that can be used in the future to predict potential risk areas and risk seasons for Culicoides-borne disease outbreaks. They are particularly useful for European veterinary authorities, who can classify both areas likely to have vectors and likely to be vector-free in advance and during a sudden outbreak in areas without active entomological surveillance. Predicting areas of uncertain status allows focusing costly active entomological surveillance to limited areas. The developed gain functions used to delimit the areas for targeted active surveillance can easily be adjusted to new diseases where the cost of concluding false presence or false absence may be different than suggested here.

Additional file

Additional file 1: Table S1. The top five of the most important variables by species group for each month. The variable importance is scaled from 0 to 100. Within each month (columns), the most important variable has a value of 100. (XLSX 67 kb)

Abbreviations
Abs: Absence class; AUC: Area under the ROC curve; BT: Bluetongue disease; BTV: Bluetongue virus; CLC: Corine Land Cover; CV: Cross-validation; CVO: Chief Veterinary Officer; dLST: Daytime land surface temperature; MIIR: Mid-infrared; NDVI: Normalized difference vegetation index; nLST: Nighttime land surface temperature; NUTS: Nomenclature of territorial units for statistics; PP: Probability of presence; Pres: Presence class; RF: Random Forest; ROC: Receiver operating characteristics curve; SVFP: Seasonal vector-free period; TFA: Temporal Fourier analysis

Acknowledgments
We would like to thank the Direction Générale de l’Alimentation from the French Ministry in charge of agriculture for the funding, and the Directions départementales de la protection des populations for their support in collecting the biting midges during the survey. We also thank the Ministerio de Agricultura y Pesca, Alimentación y Medio Ambiente for providing data on the national surveillance of Culicoides in Spain.

Funding
This study was funded by the EMIDA ERA-NET-supported project VICE (Vector-borne Infections: Risk-based and Cost-Effective Surveillance Systems). Culicoides data from Germany were partly collected within the German part of the VICE project funded by EMIDA ERA-NET through the Federal Office for Agriculture and Food (grant no. 314-06.01-2811ERA248). The Swiss Food Safety and Veterinary Office and the Vet-Austria project provided financial support to the Swiss and Austrian partners, respectively.

Availability of data and materials
The national surveillance and research data that support the findings of this study are available from the following people: Spain, Miguel Ángel Miranda Chueca; France, Thomas Balenghien; Germany, Jörn Gethmann; Denmark, Rene Badier; Sweden, Anders Lindström; Norway, Petter Hopp; Poland, Magdalena Lanska; Austria, Katharina Brugger; Switzerland, Alexander Mathis. Restrictions apply to the availability of these data, which were used under license for the current study and are not publicly available. Data are, however, available from the authors upon reasonable request and with permission from the national surveillance programmes of each country.

Authors’ contributions
ACC analysed the data and drafted the manuscript. RB planned the original study and contributed to the analysis and drafting of the manuscript. LIK contributed to the analysis and drafting of the manuscript. AB and AS contributed to the analysis and interpretation of the results. HS, SAN, MGA, AL, JC, RL, SS, EK, JG, FC, ML, MS, AO, IH, SS, PH, KB, FR, TB, CG, IR, XX, JL, JCD, CM, OD, MSL, RV, BS, MAMC, CB, JL, RE, AM and WT discussed and identified preliminary national data, negotiated data access with national surveillance authorities and research projects, jointly discussed the taxonomic and spatial resolution for analysis and identified, selected and extracted the final data and relevant variables that met the criteria for the joint database. They also wrote the protocol summaries and commented on the analysis results and edited the manuscript. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher’s Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
1Division for Diagnostics and Scientific Advice, National Veterinary Institute, Technical University of Denmark (DTU), Lyngby, Denmark. 2Department of Applied Mathematics and Computer Science, Technical University of Denmark (DTU), Lyngby, Denmark. 3Department of Agroecology - Entomology and Plant Pathology, Aarhus University, Aarhus, Denmark. 4Department of Science and Environment, Roskilde University, Roskilde, Denmark. 5National Veterinary Institute (SVA), Uppsala, Sweden. 6Bernhard Nocht Institute for Tropical Medicine, WHO Collaborating Centre for Arbovirus and Hemorrhagic Fever Research and Reference National Centre for Tropical Infectious Diseases, Hamburg, Germany. 7Department of Biology and Environmental Sciences, Carl von Ossietzky University, Oldenburg, Germany. 8Institute of Epidemiology, Friedrich Loeffler Institute, Greifswald, Germany. 9Department of Virology, National Veterinary Research Institute, Puslawy, Poland. 10Norwegian Veterinary Institute, Oslo, Norway. 11Institute for Veterinary Public Health, Vetmeduni, Vienna, Austria. 12CIRED, UMR ASTRE, F-34398 Montpellier, France. 13Institute of Parasitology and Tropical Pathology of Strasbourg, EA7292, Université de Strasbourg, Strasbourg, France. 14ÉID Méditerranée, Montpellier, France. 15Laboratory of Zoology, University of the Balearic Islands, Palma, Spain. 16Department of Animal Pathology, University of Zaragoza, Zaragoza, Spain. 17Institute of Parasitology, University of Zürich, Zürich, Switzerland. 18Avia-QIS NV, Zoersel, Belgium.
References