Remote Sens. **2014**, *6*, 6604-6619; doi:10.3390/rs6076604

OPEN ACCESS

remote sensing

ISSN 2072-4292

www.mdpi.com/journal/remotesensing

Article

Modelling the Spatial Distribution of *Culicoides imicola*: Climatic *versus* Remote Sensing Data

Jasper Van Doninck [1], Bernard De Baets [2], Jan Peters [2], Guy Hendrickx [3], Els Ducheyne [3] and Niko E.C. Verhoest [1,*]

[1] Laboratory of Hydrology and Water Management, Ghent University, Coupure links 653, 9000 Ghent, Belgium; E-Mail: Niko.Verhoest@UGent.be

[2] Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Coupure links 653, 9000 Ghent, Belgium; E-Mails: Bernard.DeBaets@UGent.be (B.D.B.); jan.peters@vito.be (J.P.)

[2] Avia-GIS, Risschotlei 33, 2980 Zoersel, Belgium; E-Mails: ghendrickx@avia-gis.com (G.H.); educheyne@avia-gis.com (E.D.)

* Author to whom correspondence should be addressed; E-Mail: Niko.Verhoest@UGent.be; Tel.: +32-926-461-38; Fax:+32-926-462-36.

Received: 17 February 2014; ; in revised form: 24 June 2014 / Accepted: 14 July 2014 / Published: 18 July 2014

Abstract: *Culicoides imicola* is the main vector of the bluetongue virus in the Mediterranean Basin. Spatial distribution models for this species traditionally employ either climatic data or remotely sensed data, or a combination of both. Until now, however, no studies compared the accuracies of *C. imicola* distribution models based on climatic *versus* remote sensing data, even though remotely sensed datasets may offer advantages over climatic datasets with respect to spatial and temporal resolution. This study performs such an analysis for datasets over the peninsula of Calabria, Italy. Spatial distribution modelling based on climatic data using the random forests machine learning technique resulted in a percentage of correctly classified *C. imicola* trapping sites of nearly 88%, thereby outperforming the linear discriminant analysis and logistic regression modelling techniques. When replacing climatic data by remote sensing data, random forests modelling accuracies decreased only slightly. Assessment of the different variables' importance showed that precipitation during late spring was the most important amongst 48 climatic variables. The dominant remotely sensed variables could be linked to climatic variables. Notwithstanding the slight decrease in predictive performance in this study, remotely sensed datasets could be preferred over climatic datasets for the modelling of *C. imicola*. Unlike climatic observations, remote

sensing provides an equally high spatial resolution globally. Additionally, its high temporal resolution allows for investigating changes in species' presence and changing environment.

Keywords: species distribution modelling; bluetongue; MODIS; WorldClim; random forests; variable importance

1. Introduction

Bluetongue is a disease, classified as a listed (notifiable) disease by the World Organisation for Animal Health, caused by the bluetongue virus (BTV), a species of the *Orbivirus* genus [1]. BTV is capable of infecting any type of ruminant, but mainly affects populations of sheep. It is a vector-borne disease, spread by female adults of several species of biting midges of the genus *Culicoides* (Diptera: Ceratopogonidae). In the Mediterranean Basin, a single species, *C. imicola*, is considered as the main vector of BTV. *C. imicola* is confined to old world regions with a Mediterranean climate, where it may also vector African horse sickness virus, equine encephalosis virus, bovine ephemeral fever virus and akabane virus [2]. The recent northward expansion of the area affected by BTV has been linked to climate change, allowing *C. imicola* populations to settle in regions where the species was previously absent, while indigenous European *Culicoides* species are responsible for outbreaks in Western and Central Europe [3].

Numerous studies tried to model the spatial distribution of the occurrence and/or abundance of *C. imicola* at different study sites and over different spatial scales. In an early study, Baylis *et al.* [4] compared *C. imicola* abundances at 28 trapping sites in Morocco with data from meteorological stations installed at the trapping sites (wind speed, humidity, air temperature and soil temperature) and remotely sensed data (NDVI). Linear correlations between the different climatic variables and the abundance of the midge were highest for wind speed and for average annual minimum NDVI. A model operationally applicable was developed by Wittmann *et al.* [5], where *C. imicola* of 30 trapping sites in the Iberian peninsula were used to train a logistic regression model based on altitude and 10 climatic variables for the period 1931–1960. The trained model resulted in a percentage of correctly classified trapping sites of 85% in an internal validation, meaning that the validation dataset was the same as the training dataset. The model parameters were then used to extrapolate the results over the entire Mediterranean Basin. Calistri *et al.* [6], however, validated these parameters for Italy, and Calabria in particular, and found that the trained model was unable to classify *C. imicola* presence and absence sites.

An alternative method was introduced by Baylis *et al.* [7], who used a combination of 40 remote sensing variables derived from temporal Fourier processing of AVHRR data, topographic data and vapour pressure deficit in a model based on discriminant analysis. The model, trained on 44 *C. imicola* trapping sites in Portugal, Spain and Morocco resulted in 93.2% correctly classified sites in an internal validation. Similar models, but based on remote sensing data only, were applied to 87 trapping sites in Portugal [8] and 248 sites in Sicily [9], resulting in 95.4% and 87% correctly classified pixels, respectively, again in an internal validation.

Most *C. imicola* distribution models are based on either discriminant analysis or logistic regression, although the data sources used may vary strongly. Guis *et al.* [10], for example, included land cover variables, as well as landscape metrics derived from high resolution imagery in a logistic regression model for Corsica. Conte *et al.* [11] added the presence of water in rivers to meteorological and topographic data for a model for Italy based on 546 trapping sites, resulting in 75% correctly classified sites in an internal validation. Acevedo *et al.* [12] included the availability of host species, as well as land cover, climatic and pedological variables, in an analysis based on trappings during the period 2005–2008 in Spain.

Few studies have investigated the difference in performance between different modelling techniques. In a study by Peters *et al.* [13], linear discriminant analysis and logistic regression were compared with the random forests ensemble learning technique. In a study over the Iberian peninsula using trapping data from 2004 to 2006 and both climatic and remote sensing data, the random forests model was found to be superior to the other models when no preprocessing of the trapping data, consisting of a reduction of false absences, was performed.

While the studies mentioned here employ either climatic data or remote sensing data or both, no studies compared the accuracies of the predicted distributions based solely on climatic or remote sensing data. Yet, the use of remote sensing data can be assumed to offer some advantages over climatic data. First of all, space-borne remote sensing offers global coverage at a fixed spatial resolution, while area-covering climatic data must be interpolated between meteorological stations. The accuracy of the interpolated values is therefore dependent on the density of the meteorological network. Furthermore, climatic datasets, e.g., those developed by Hijmans *et al.* [14], provide monthly values based on averaging over long time spans of up to 50 years. This prevents assessment of the influence of climate change during this period, rendering meteorological observations from dynamic into static variables. The long records of remote sensing products allow the assessment of changing climate and land cover conditions, while providing temporal resolutions higher than the monthly resolutions of climatic datasets. It should be noted, however, that many studies undo the advantages of remote sensing by using multi-year averages. Calvete *et al.* [15], for example, tried to model the distribution of *Culicoides* from trapping data from 2004 to 2006, while using monthly averaged NDVI data from 1981 to 2003. Similarly, Baylis *et al*. [7] used remote sensing data acquired between 1982 and 1994 in combination with trapping data between 1993 and 1995, and Tatem *et al.* [8] used trapping data from 2000 and 2001 with remote sensing data from 1992 to 1996.

Clearly, the use of solely remote sensing data entails some disadvantages. Some environmental variables that may restrict the spatial distribution of *C. imicola*, such as air temperature, air humidity or precipitation, are much easier to measure from *in situ* stations than from remote sensing, if measurable at all. This paper aims at comparing the accuracies of *C. imicola* distribution modelling based on solely climatic data records *versus* solely routinely produced remote sensing products. Modelling is performed using random forests, a technique that also allows for assessing the importance of the different climatic or remote sensing variables. The accuracies obtained using random forests are compared with those obtained using linear discriminant analysis and multiple logistic regression, two more traditional modelling techniques.

2. Materials and Methods

2.1. Culicoides Life Cycle

More than 1400 species of the genus *Culicoides* have been identified worldwide [2], all measuring from 1 mm to 3 mm in size. The vast majority of *Culicoides* species are blood sucking, attacking mammals and birds, in order to allow maturation of the eggs. The life cycle of *Culicoides* includes three immature stages, egg, larva and pupa, and a mature or imago stage, each characterized by species-specific environmental demands. Several laboratory and field studies have been conducted in order to determine the habitat characteristics of the different species. Additionally, modelling studies tried to infer these habitat characteristics using data from of *in situ Culicoides* trapping sites and meteorological, remote sensing and other datasets.

Air temperature is considered as one of the major limiting factors of *C. imicola* distribution on global and continental scale, although the precise biological basis of this dependency is unclear. For several *Culicoides* species, correlations between air temperature and species activity, larval survivorship and adult mortality were observed in trapping or laboratory studies [2]. In a laboratory study, Veronesi *et al*. [16] observed that the period required for blood-feeding *C. imicola* females to produce adult progeny occupied 34–56 days at 20 °C, 15–21 days at 25 °C and 11–16 days at 28 °C. Additionally, freezing temperatures are known to kill adult midges, thus reducing catch abundances at sites affected by frost [17]. This dependency on air temperature was confirmed in several *C. imicola* modelling studies using climatic data. Purse *et al.* [18] found *C. imicola* in the Mediterranean Basin to occur in warm (annual mean 12–20 °C) regions with low seasonal variations. A model developed by Wittmann *et al.* [5] identified three temperature variables (minimum of the monthly minimum temperatures, maximum of the monthly maximum temperatures and number of months per year with a mean temperature above 12.5 °C) as significant determinants for the Iberian peninsula.

A second important variable is soil moisture, since a large part of the life cycle of *Culicoides* species (the development from egg to larva and pupa) is completed in the upper soil layer, with highest concentrations of immature *Culicoides* in the first 5 cm [19,20]. *C. imicola* has been observed to prefer semi-moist breeding sites, and has been found in drainage canals and puddles created by leakage from water pipes [21,22]. Foxi and Delrio [23] state that *C. imicola* was found to breed preferentially in mud 20 cm above the pond shoreline, where soils are not subject to flooding. While Delrio *et al.* [24] observed *C. imicola* larvae in saturated soils, [25] observed that pupae of *C. imicola* drown on immersion in water.

Land use and vegetation cover conditions under which *C. imicola* preferably lives and breeds are poorly understood. From modelling studies, Conte *et al.* [26] state that *C. imicola* can be classified as heliophilic, favouring less vegetated shrub and grassland. This is in accordance with a number of observations of breeding sites in moist grasslands [21,22]. The preferred land use or land cover for adults is, however, less documented.

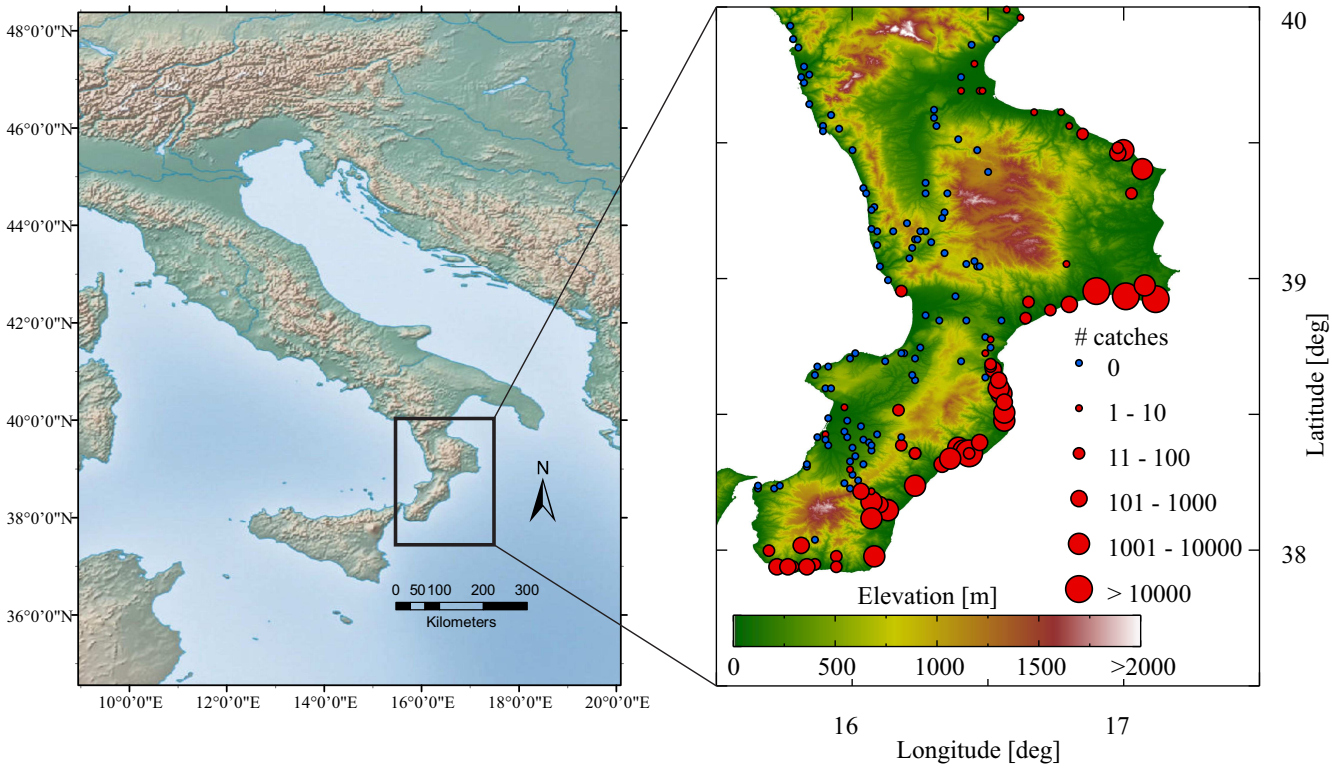
Other factors found to determine the species' distribution include topography (affecting temperature and soil drainage), wind velocity, soil properties (soil organic matter, soil texture) and the availability of hosts. Wind speed affects adult *Culicoides* activity, by suppressing activity above certain velocities [2]. Soil properties determine the suitability of breeding sites. The importance of soil texture can be related to

the soil moisture factor, where clayey soils can retain water for longer periods than more sandy soils. Soil organic matter is essential for larval growth [27], and ruminant hosts are required for the blood-feeding of adult females and the production of progeny [2].

2.2. C. imicola in situ Data

One region characterized by a peculiar pattern of *C. imicola* occurrence, probably caused by varying environmental conditions, is the Italian region of Calabria, situated in the southwestern tip of mainland Italy (Figure 1). Here, the insect is present at the eastern Ionian seaboard and absent at the western Tyrrhenian coastline. This pattern has previously been explained as a result of the difference in soil texture, which is, in general, finer at the eastern coastline. These finer textures would allow soils to remain sufficiently wet during summer for the immature insect stages to develop [28,29].

Figure 1. Topography of the study site (Shuttle Radar Topography Mission digital elevation model) and *C. imicola* catch abundances for the years 2000 and 2001 (Entomological National Surveillance Programme).



Since 2000, the Entomological National Surveillance Programme monitors the spatial and temporal dynamics of the *C. imicola* population in Italy using Onderstepoort-type blacklight traps in accordance with standardized surveillance procedures [30]. Captures are examined to determine the total number of insects, the total number of *Culicoides* and the total number of *C. imicola*. For this study, the dataset consists of the total number of *C. imicola* collected during the months of highest abundance (August–October) of the years 2000 and 2001, at 168 trapping sites throughout Calabria (Figure 1). *C. imicola* abundances were transformed to absences (zero catches) and presences (non-zero catches), resulting in 102 absence records and 66 presence records.

2.3. Climatic and Remote Sensing Datasets

The WorldClim dataset [14] is a set of raster layers of climatic variables at 30″ (approximately 1 km) resolution, covering all land masses except Antarctica. The available layers contain monthly values of minimum (T_{min}), maximum (T_{max}) and mean (T_{mean}) temperature and total precipitation (P), representative for the period 1950–2000, derived through interpolation of monthly values recorded at meteorological stations. The data accuracy is restricted by the density of the meteorological stations, which is especially low in parts of Asia, Africa and South-America [14]. For Europe, a relatively dense network of stations is used to produce the climatic variables. The 48 climatic data layers for Calabria are freely available for academic and other non-commercial use, and downloadable through the WorldClim website. No further preprocessing on the climatic variables was performed.

Routinely generated remote sensing products used in this study are the Aqua MODIS 8-day daytime ($T_{\mathrm{S,day}}$) and nighttime ($T_{\mathrm{S,night}}$) land surface temperature and the monthly NDVI at 1 km resolution. Images from 2002 to 2010 were acquired through the USGS Land Processes Distributed Active Archive Center and combined into representative monthly averages for this period of daytime and nighttime surface temperature and NDVI, resulting in a total of 36 data layers. Remotely sensed data representative for the period 2002–2010 can, for Calabria, be used in combination with *C. imicola* trapping data from 2000 and 2001, since no expansion of the geographical range of *C. imicola* was detected in the region during the first decade of this century [31]. Multi-year monthly averaging of MODIS imagery is performed in order to transform the remotely sensed data to a format comparable with that of the WorldClim dataset, even though this implies a loss of information. A compositing period of eight years was chosen to capture between-year variability, without creating a large time gap between *in situ* data collection and remote sensing image acquisition. MODIS Aqua datasets were preferred over those of Terra, since Aqua's equatorial crossing times are at 1:30 a.m. and 1:30 p.m., while Terra's are at 10:30 a.m. and 10:30 p.m. Daytime and nighttime surface temperatures observed by Aqua will thus be closer to maximum and minimum temperatures, respectively, which are relevant in the *C. imicola* life cycle [2,16,17].

2.4. Modelling Techniques

Random forests (RF, [32]) is a data-driven modelling technique, classifying observations with unknown class membership based on a model trained using observations with known class membership. This machine learning technique generates many classification trees, each of which is grown using a randomly drawn subset of the original dataset. The nodes of the different classification trees are grown using the best split variable selected out of a randomly selected subset of predictive variables [33]. The number of trees grown and the number of predictive variables used to split the nodes are two user-defined parameters, here set to 300 and 3, respectively, following Peters *et al.* [13]. Once all classification trees of the random forest are trained, observations with unknown class membership are classified by each classification tree, resulting in a unique class label (absence or presence) for each tree. The proportion of trees assigning a presence label is an indicator of the similarity to the training locations where *C. imicola* was observed, but could also be seen as a proxy for the probability of occurrence [34]. Random forests

have been applied successfully in ecological distribution modelling [13,35,36], but also in, e.g., land cover classification [34,37] or yield prediction [38].

An additional feature of random forests is the assessment of the predictive variables' importance, where the effect of a random permutation of a variable on the classifier performance is investigated. The decrease in classifier performance can be interpreted as a measure of the variable's importance. The permutation of informative variables will thus result in a strong decrease in the classifier's performance, while non-informative variables will cause a minor change in performance when permuted.

The random forests classifier will be compared with the linear discriminant analysis (LDA) and multiple logistic regression (MLR) classifiers [39], which are traditionally applied in species distribution modelling. As for random forests, both LDA and MLR require training observations with known membership to compute a probability of class membership. A threshold set on this probability then allows for classifying observations with unknown class membership. All three classifiers are implemented in the statistical software environment `R`.

2.5. Model Validation

Due to the difficulties encountered in collecting large datasets of *C. imicola* trapping data, many modelling studies only validated model outputs internally, *i.e.*, the validation data equals the training dataset. This type of validation increases the risk of overfitting and provides an over-optimistic accuracy estimate. In this study, the model performance is assessed by a 3-fold cross validation, where the dataset is split randomly in three disjoint subsets of equal size, and each subset is used to validate the model trained on the remaining two subsets. As a result, each record in the original dataset will be assigned a predicted probability of *C. imicola* occurrence and an absence/presence label, and this for the RF, LDA and MLR classifiers.

Two accuracy measures are used to mutually validate the different modelling techniques: the percentage correctly classified sites (PCC) and the area under the Receiver Operating Characteristic (ROC) curve (AUC). While the PCC is a threshold-dependent accuracy measure, with the threshold here set at a probability of occurrence of 0.5, the AUC is threshold-independent [40]. In ROC graphs, the true positive rate (fraction of observed presences that are predicted correctly) is plotted *versus* the false positive rate (fractions of observed absences that are predicted incorrectly), for all possible threshold values between zero and one. The AUC, which ranges between zero and one, thus describes the likelihood that a presence site is assigned a higher modelled probability than an absence site, with a value higher than 0.5 when the model performs better than random guessing.

Because of the relatively limited size of the *in situ* dataset (168 sites), the accuracy measures can be strongly affected by the random selection of the training and validation folds. The 3-fold cross validation, hence computation of the accuracy statistics, is therefore performed for 100 runs, each with a random selection of the training and validation folds, in order to minimize chance effects introduced by this random selection. This results in 100 values of PCC and AUC for each modelling technique. The average and standard deviation over these 100 runs provide an estimate of the overall performance and stability of the different techniques. Additionally, since each trapping site will be assigned a predicted absence or presence during each run, the spatial distribution and number of misclassifications can be displayed.

Finally, the probability of *C. imicola* occurrence for the entire Calabrian peninsula can be predicted from the area-covering climatic or remote sensing data layers.

3. Results and Discussion

3.1. Modelling Accuracies

Table 1 gives the average value and standard deviation of PCC and AUC over the 100 model runs for the different modelling techniques based on the WorldClim dataset. Random forests modelling performs significantly better ($p < 0.0001$) than linear discriminant analysis for both performance measures, and multiple logistic regression performs much poorer than both of these methods. Nearly 88% of the sites are correctly classified using RF, compared with 84% with LDA and only 78% with MLR. These accuracies are close to the 87.5% obtained by Conte *et al.* [26] over the same study site with an LDA model trained using meteorological data, topography, land cover, NDVI and aridity and *Culicoides* trapping data from 2000 to 2004. Figure 2 shows that the random forests model based on climatic data captures the general east–west distribution of *C. imicola*, with the percentage of trees assigning presence close to 100% along the eastern coast and low values at the western coastline and the central mountain ranges. Misclassification of absence and presence sites are located in regions with mixed responses of the individual trees and often in regions where presence and absence locations are found over short distances, e.g., near the city of Rosarno at the southern half of the western coast.

Table 1. Accuracy measures of *C. imicola* distribution models for different modelling techniques and input datasets.

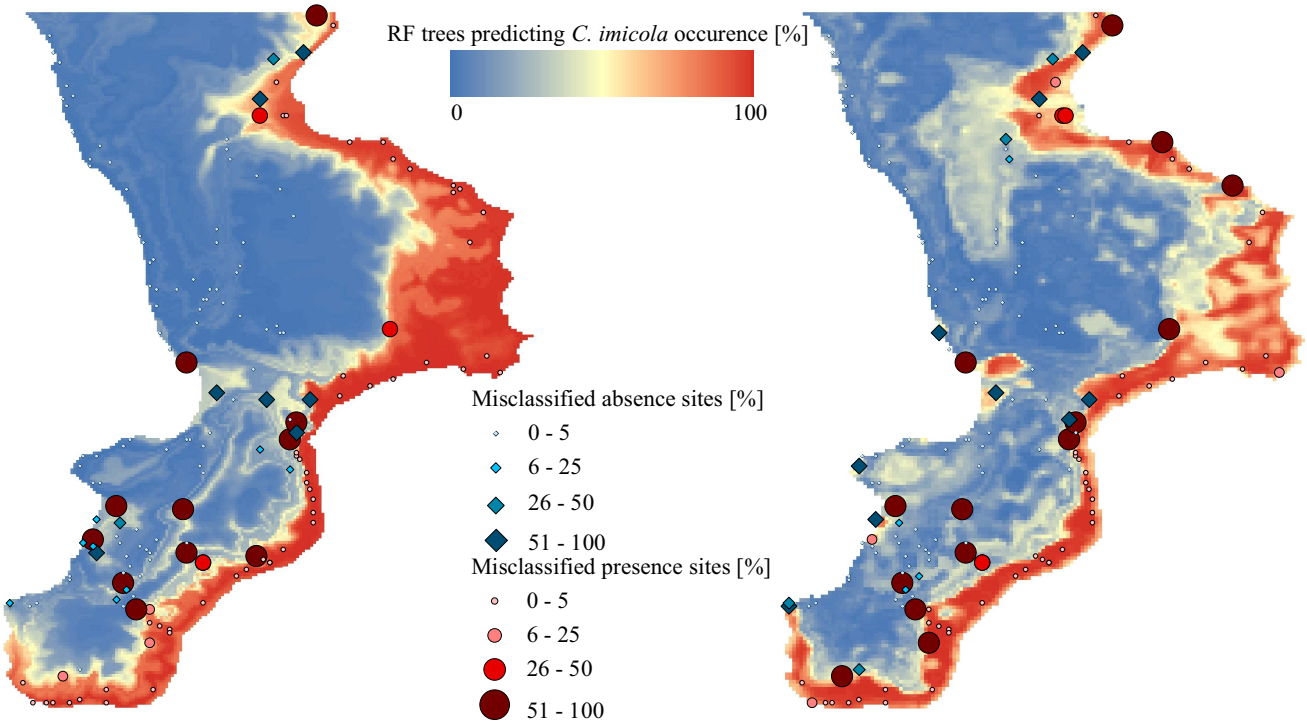
	Climatic Data		**Remote Sensing Data**	
	PCC (St. Dev.) (%)	**AUC (St. Dev.) (%)**	**PCC (St. Dev.) (%)**	**AUC (St. Dev.) (%)**
RF	87.7 (1.4)	92.5 (1.2)	85.9 (1.6)	91.2 (1.4)
LDA	84.3 (2.2)	89.2 (1.8)	86.3 (1.8)	89.9 (1.3)
MLR	76.7 (2.4)	80.7 (2.7)	75.7 (2.8)	77.6 (2.9)

When comparing the misclassifications in Figure 2 with the catch abundances in Figure 1, it is observed that many of the misclassified presence sites are characterized by low (<10) *C. imicola* catch abundances, with sometimes as few as one single catch over the two-year period. This can indicate a very small population of *C. imicola* in these regions, and might explain the failure of trapping the species at neighbouring sites, even though it might also have been present there. In this case, the absence sites can represent false absences. Alternatively, these presence sites might represent false presences, where the trapped individual was transported to this region by the wind, but was unable to establish a population. Finally, the possibility of a misclassification of a different species of *Culicoides* as *C. imicola* cannot be excluded.

For the models trained on MODIS data, the accuracy measures (Table 1) indicate no significant difference between the performances of the RF and LDA models in terms of percentage correctly classified sites. In terms of AUC, RF performs significantly ($p < 0.0001$) better than LDA. Again, MLR

accuracies are inferior to those of the other techniques. When comparing accuracies of the models based on remotely sensed input data *versus* those using climatic data, it is observed that the RF predictions using climatic data are significantly better ($p < 0.0001$ for both accuracy measures) than these using remote sensing data. The percentage correctly classified sites decreased by approximately two percent when replacing climatic variables by remotely sensed variables. The opposite is observed for the LDA models, with the PCC increasing by two percent when using remotely sensed data, and the AUC also increasing significantly ($p = 0.0019$).

Figure 2. Percentage of RF trees assigning *C. imicola* occurrence and percentage of misclassifications at absence and presence sites, based on climatic data (**left**) and remotely sensed data (**right**).



The predicted *C. imicola* distribution map (Figure 2) obtained from the RF model using remotely sensed data strongly resembles the one obtained from the WorldClim dataset. The misclassified trapping sites also occur in the same regions. Climatic datasets can thus be replaced by remotely sensed datasets for *C. imicola* distribution modelling without compromising prediction accuracies. In certain situations, the use of remotely sensed data may be preferred in modelling studies, e.g., in regions where the meteorological stations between which the climatic data are interpolated are sparse, in which case uncertainties on these data may be high. Sensors such as AVHRR or MODIS provide global coverage at a fixed spatial resolution of approximately 1 km, which is appropriate for most species distribution modelling studies. Another important advantage of remote sensing data is their high temporal resolution. Climatic datasets generally consist of monthly layers, averaged over several years or decades. This makes them of limited use in the study of dynamic phenomena such as emergent vector-borne diseases. Remote sensing data, on the other hand, are provided at a temporal resolution of up to one day, or combined into 8-day, 16-day or monthly products, with time series of multiple decades. This offers possibilities for

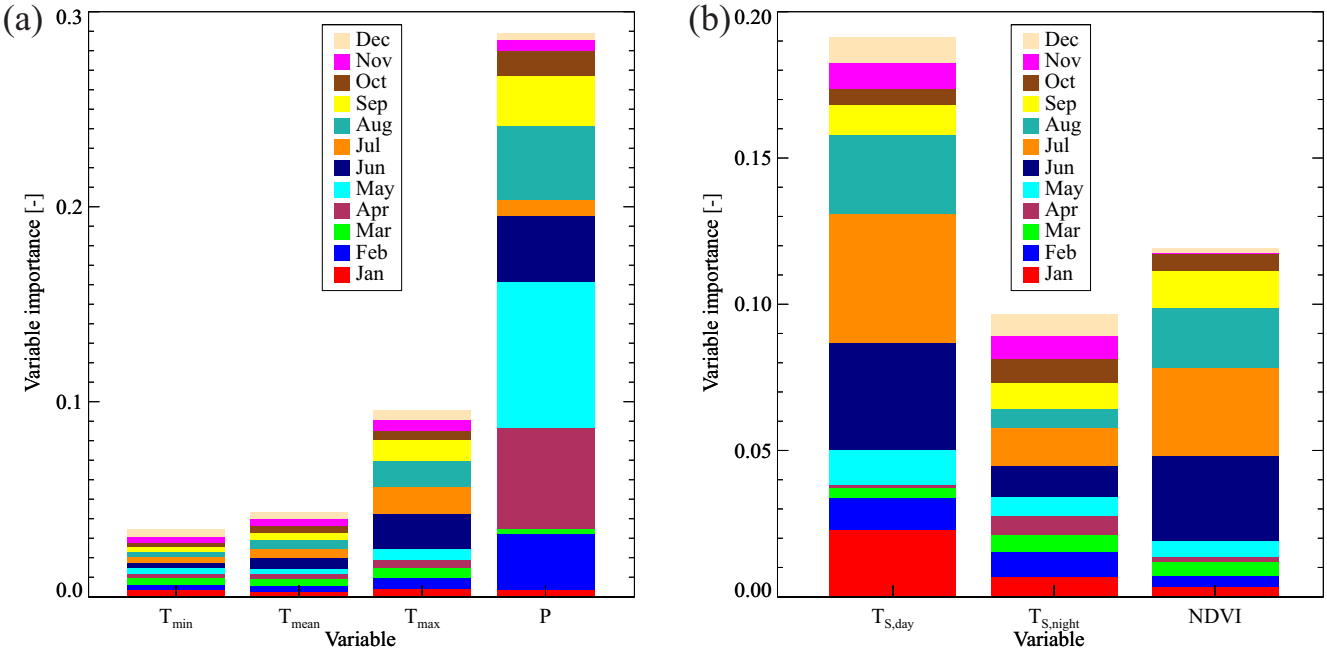
the study of the interaction between dynamic environmental conditions (such as land cover or climate change) and species distribution ranges. Peters *et al.* [41] presented such a study, where the spatial distribution of *C. imicola* for a given year was predicted using remotely sensed data of that same year. One important remaining challenge is how to transform the high-dimensional remotely sensed datasets into variables useful in species distribution modelling.

This study deliberately did not take advantage of the higher temporal resolution of the remote sensing data, but instead degraded it to that of the WorldClim dataset. This allowed for a fairer comparison of the relative model importance of the different variables.

3.2. Variable Importance

The importance of the different climatic variables, as identified by the RF model, is given in Figure 3a. The variable importance of the different precipitation variables is much higher than the temperature variables. These numbers should, however, be treated with consideration, since the different temperature variables can be expected to be strongly correlated, both between T_{min}, T_{mean} and T_{max} mutually as between the different months. Strongly correlated variables are known to result in lower variable importance [42]. Nevertheless, the total amount of precipitation in the months of May and April seems to influence the distribution of *C. imicola*.

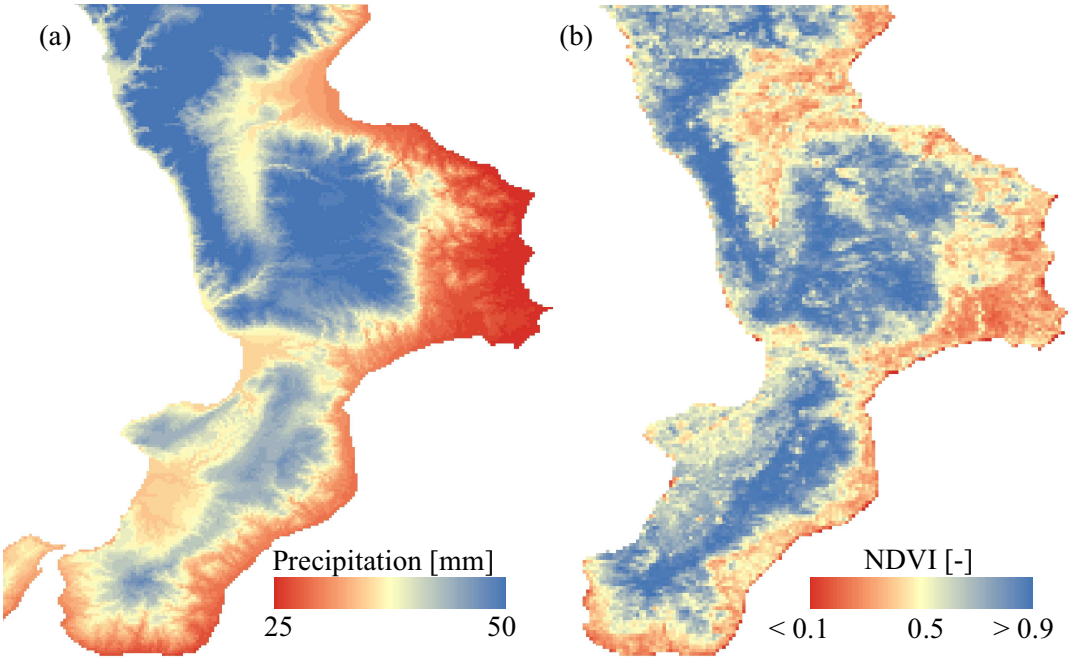
Figure 3. Relative importance of the different variables in the random forest models (**a**) based on climatic data and (**b**) based on standard Aqua MODIS products.



For the model based on the remotely sensed variables, the daytime surface temperatures during the months of June, July and August are the most important (Figure 3b). This is in agreement with the variable importance obtained for the climatic variables, where the maximum temperatures during these three months were identified as the most important temperature variables. The months of June, July and August also yield the highest importance among the different NDVI variables.

When considering the spatial pattern of the most important climatic variable, *i.e.*, the total precipitation for the month of May (Figure 4a), it is observed that this pattern corresponds strongly to the predicted probability of *C. imicola* presence (Figure 2). Absence sites correspond to regions with rich precipitation, while presence sites receive much less rainfall. A similar pattern is found for the month of April, albeit with much higher absolute precipitation totals. The east–west gradient of rainfall can be explained by the predominant westerly winds in this part of the Mediterranean [43], which causes precipitation to fall at the western side of the mountain range dissecting the peninsula.

Figure 4. (**a**) Total precipitation for the month of May in the WorldClim [14] dataset. (**b**) Multiyear averaged MODIS NDVI for the month of July.



The high amount of precipitation at *C. imicola* absence sites might support the hypothesis that these sites are unsuitable because the species' pupae will drown due to the high rainfall, thus interrupting the life cycle of *C. imicola* [25]. At the eastern coast, *C. imicola* might survive due to dryer conditions in spring. This would complement previous assumptions [28] stating that *C. imicola* survives where fine-textured soils are capable of retaining sufficient moisture during summer, thus avoiding desiccation. The rainfall-based assumption, however, can explain the absence of the species in the northwestern part of the study site where fine-textured soils can be found, which should support populations of *C. imicola* according to the texture-based assumption [29].

The overall importance of the 36 temperature variables is low, even though temperature is known to restrict *C. imicola* distribution at the global scale and is in Calabria probably largely responsible for its absence at higher elevations. As mentioned earlier, this low importance is partly due to the strong correlations between the different variables. A reduction or transformation of the variables (e.g., through temporal Fourier analyses as performed by Baylis *et al.* [7]) will possibly result in an entirely different view.

One of the most important remotely sensed variables, the multiyear averaged NDVI for the month of July (Figure 4b), partly reflects the general land cover and topography of Calabria. The NDVI during summer, however, also shows a clear correspondence to the total rainfall during the month of May (Figure 4a). This is not surprising, since abundant rainfall during late spring and early summer allows vegetation to grow throughout the dryer summer season. The existence of a relationship between NDVI or other vegetation indices and rainfall or soil moisture has been demonstrated earlier for semi-arid or Mediterranean conditions [44,45].

High NDVI values here correspond to *C. imicola* absence sites. This contradicts the findings of Baylis *et al.* [4], who linked the species' abundance in Morocco to soil moisture through NDVI proposing that "areas in Morocco with high levels of soil moisture in late summer or autumn provide more, larger and/or more enduring breeding sites for *C. imicola*, as well as supporting more photosynthetically active vegetation and hence having higher NDVI". This link has subsequently been used to extrapolate trained models to Mediterranean-wide predictions [7]. The present study, however, demonstrates that conclusions drawn for one region of the Mediterranean Basin cannot simply be extrapolated to others.

4. Conclusions

This study aimed at comparing *C. imicola* distribution modelling accuracies using climatic data *versus* standard remote sensing data. Three modelling techniques were employed: the established linear discriminant analysis and multiple logistic regression, and random forests, a more recent ensemble learning technique. The importance of the different input variables was assessed using the random forests technique.

For the models based on climatic data, RF outperformed both other techniques for the investigated accuracy measures. Analysis of the variable importance revealed that the *C. imicola* distribution is largely determined by the amount of precipitation during spring. It is unclear whether this dependence is causal, where intense rainfall might interrupt the species' life cycle by drowning during one of its immature stages, or merely coincidental or linked through other processes.

Replacing the climatic variables by standard MODIS products resulted in a significant, although limited, reduction of the predictive capability of the RF model, while slightly increasing the accuracies of the LDA model. The most important remotely sensed variables could be linked to climatic variables. Given the restrictions of climatic data with respect to temporal and spatial resolution, the use of remotely sensed datasets in *C. imicola* distribution modelling is advisable.

Acknowledgments

This research was funded by the Belgian Federal Science Policy under the Research Programme for Earth Observation Stereo II as part of the EPIDEMOIST project (contract nr. SR/02/124), and by the Special Research Fund of Ghent University.

Author Contributions

Jasper Van Doninck and Jan Peters conducted the research under supervision of Niko E.C. Verhoest and Bernard De Baets. Guy Hendrickx and Els Ducheyne provided *C. imicola* trapping data. The manuscript was written with inputs from all authors.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Wilson, A.J.; Mellor, P.S. Bluetongue in Europe: Past, present and future. *Philos. Trans. R. Soc. B* **2009**, *364*, 2669–2681.
2. Mellor, P.S.; Boorman, J.; Baylis, M. *Culicoides* biting midges: Their role as arbovirus vectors. *Annu. Rev. Entomol.* **2000**, *45*, 307–340.
3. Saegerman, C.; Berkvens, D.; Mellor, P.S. Bluetongue epidemiology in the European Union. *Emerg. Infect. Dis.* **2008**, *14*, 539–544.
4. Baylis, M.; Bouayoune, H.; Touti, J.; El Hasnaoui, H. Use of climate data and satellite imagery to model the abundance of *Culicoides imicola*, the vector of African horse sickness virus, in Morocco. *Med. Vet. Entomol.* **1998**, *12*, 255–266.
5. Wittmann, E.J.; Mellor, P.S.; Baylis, M. Using climate data to map the potential distribution of *Culicoides imicola* (Diptera: Ceratopogonidae) in Europe. *Rev. Sci. Tech. Off. Int. Epiz.* **2001**, *20*, 731–740.
6. Calistri, P.; Goffredo, M.; Caporale, V.; Meiswinkel, R. The distribution of *Culicoides imicola* in Italy: Application and evaluation of current Mediterranean models based on climate. *J. Vet. Med.* **2003**, *B50*, 132–138.
7. Baylis, M.; Mellor, P.S.; Wittmann, E.J.; Rogers, D.J. Prediction of areas around the Mediterranean at risk of bluetongue by modelling the distribution of its vector using satellite imaging. *Vet. Rec.* **2001**, *149*, 639–643.
8. Tatem, A.J.; Baylis, M.; Mellor, P.S.; Purse, B.V.; Capela, R.; Pena, I.; Rogers, D.J. Prediction of bluetongue vector distribution in Europe and North Africa using satellite imagery. *Vet. Microbiol.* **2003**, *97*, 13–29.
9. Purse, B.V.; Tatem, A.J.; Caracappa, S.; Rogers, D.J.; Mellor, P.S.; Baylis, M.; Torina, A. Modelling the distribution of *Culicoides* bluetongue virus vectors in Sicily in relation to satellite-derived climate variables. *Med. Vet. Entomol.* **2004**, *18*, 90–101.
10. Guis, H.; Tran, A.; de La Rocque, S.; Baldet, T.; Gerbier, G.; Barragué, B.; Biteau-Coroller, F.; Roger, F.; Viel, J.F.; Mauny, F. Use of high spatial resolution satellite imagery to characterize landscapes at risk for bluetongue. *Vet. Res.* **2007**, *38*, 669–683.
11. Conte, A.; Giovannini, A.; Savini, L.; Goffredo, M.; Calistri, P.; Meiswinkel, R. The effect of climate on the presence of *Culicoides imicola* in Italy. *J. Vet. Med.* **2003**, *B50*, 139–147.

12. Acevedo, P.; Ruiz-Fons, F.; Estrada, R.; Márques, A.L.; Miranda, M.A.; Gortázar, C.; Lucientes, J. A broad assessment of factors determining *Culicoides imicola* abundance: Modelling the present and forecasting its future in climate change scenarios. *PLoS One* **2010**, *5*, e14236.
13. Peters, J.; De Baets, B.; Van doninck, J.; Calvete, C.; Lucientes, J.; De Clercq, E.M.; Ducheyne, E.; Verhoest, N.E.C. Absence reduction in entomological surveillance data to improve niche-based distribution models for *Culicoides imicola*. *Prev. Vet. Med.* **2011**, *100*, 15–28.
14. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **2005**, *25*, 1965–1978.
15. Calvete, C.; Estrada, R.; Miranda, M.A.; Borrás, D.; Calvo, J.H.; Lucientes, J. Modelling the distributions and spatial coincidence of bluetongue vectors *Culicoides imicola* and the *Culicoides obsoletus* group throughout the Iberian Peninsula. *Med. Vet. Entomol.* **2008**, *22*, 124–134.
16. Veronesi, E.; Venter, G.J.; Labuschagne, K.; Mellor, P.S.; Carpenter, S. Life history of *Culicoides (Avaritia) imicola* Kieffer in the laboratory at different rearing temperatures. *Vet. Parasitol.* **2009**, *163*, 370–373.
17. Venter, G.J.; Nevil, E.M.; van der Linde, T.C.D. Seasonal abundance and parity of stock-associated *Culicoides* species (Diptera : Ceratopogonidae) in different climatic regions in southern Africa in relation to their viral vector potential. *Onderstepoort J. Vet. Res.* **1997**, *64*, 25–271.
18. Purse, B.V.; McCormick, B.J.J.; Mellor, P.S.; Baylis, M.; Boorman, J.P.T.; Borras, D.; Burgu, I.; Capela, R.; Caracappa, S.; Collantes, F.; *et al*. Incriminating bluetonge virus vectors with climate envelope models. *J. Appl. Ecol.* **2007**, *44*, 1231–1242.
19. Blackwell, A.; King, F.C. The vertical distribution of *Culicoides impunctatus* larvae. *Med. Vet. Entomol.* **1997**, *11*, 46–48.
20. Uslu, U.; Dik, B. Vertical distribution of *Culicoides* larvae and pupae. *Med. Vet. Entomol.* **2006**, *20*, 350–352.
21. Braverman, Y.; Galun, R.; Ziv, M. Breeding sites of some *Culicoides* spp. (Diptera: Ceratopogonidae) in Israel. *Mosq. News* **1974**, *34*, 303–308.
22. Mellor, P.S.; Pizolis, G. Observations on breeding sites and light-trap collections of *Culicoides* during an outbreak of bluetongue in Cyprus. *Bull. Entomol. Res.* **1979**, *69*, 229–234.
23. Foxi, C.; Delrio, G. Larval habitats and seasonal abundance of *Culicoides* biting midges found in association with sheep in northern Sardinia, Italy. *Med. Vet. Entomol.* **2010**, *24*, 199–209.
24. Delrio, G.; Deliperi, S.; Foxi, S.; Pantaleoni, R.A.; Piras, S. Osservazioni in Sardegna sulla dinamica di popolazione di *Culicoides imicola* Kiefer vettore della bluetongue. In Proceedings of the 2002 Atti XIX Congresso Nazionale Italiano di Entomologia, Catania, Italy, 10–15 June 2002; pp. 1089–1094.
25. Nevill, E.M. Biological Studies on Some South African *Culicoides* Species (Diptera: Ceratopogonidae) and the Morphology of Their Immature Stages. Master's Thesis, University of Pretoria, Pretoria, South Africa, 1967.
26. Conte, A.; Goffredo, M.; Ippoliti, C.; Meiswinkel, R. Influence of biotic and abiotic factors on the distribution and abundance of *Culicoides imicola* and the *Obsoletus Complex* in Italy. *Vet. Parasitol.* **2007**, *150*, 333–344.

27. Kettle, D.S. Biology and bionomics of bloodsuckung Ceratopogonids. *Annu. Rev. Entomol.* **1977**, *22*, 33–51.
28. Conte, A.; Ippoliti, C.; Savini, L.; Goffredo, M.; Meiswinkel, R. Novel environmental factors influencing the distribution and abundance of *Culicoides imicola* and the Obsoletus Complex in Italy. *Vet. Italiana* **2007**, *43*, 571–580.
29. Peters, J.; Conte, A.; Van doninck, J.; Verhoest, N.E.C.; De Clercq, E.M.; Goffredo, M.; De Baets, B.; Hendrickx, G.; Ducheyne, E. On the relation between soil moisture dynamics and the geographical distribution of *Culicoides imicola*. *Ecohydrology* **2014**, *7*, 622–632.
30. Goffredo, M.; Meiswinkel, R. Entomological surveillance of bluetongue in Italy: Methods of capture, catch analysis and identification of *Culicoides* biting midges. *Vet. Italiana* **2004**, *40*, 260–265.
31. Conte, A.; Gilbert, M.; Goffredo, M. Eight years of entomological surveillance in Italy show no evidence of *Culicoides imicola* geographical range expansion. *J. Appl. Ecol.* **2009**, *46*, 1332–1339.
32. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32.
33. Liaw, A.; Wiener, M. Classification and regression by random forest. *R News* **2002**, *2*, 18–22.
34. Loosvelt, L.; Peters, J.; Skriver, H.; Lievens, H.; Van Coillie, F.M.B.; De Baets, B.; Verhoest, N.E.C. Random Forests as a tool for estimating uncertainty at pixel-level in SAR image classification. *Int. J. Appl. Earth Observ. Geoinf.* **2012**, *19*, 173–184.
35. Peters, J.; De Baets, B.; Verhoest, N.E.C.; Samson, R.; Degroeve, S.; De Becker, P.; Huybrechts, W. Random forests as a tool for ecohydrological distribution modelling. *Ecol. Model.* **2007**, *207*, 304–318.
36. Sehgal, R.N.M.; Buermann, W.; Harrigan, R.J.; Bonneaud, C.; Loiseau, C.; Chasar, A.; Sepil, I.; Valkiũnas, G.; Iezhova, T.; Saatchi, S.; *et al*. Spatially explicit predictions of blood parasites in a widely distributed African rainforest bird. *Proc. R. Soc. B* **2011**, *278*, 1025–1033.
37. Loosvelt, L.; Peters, J.; Skriver, H.; De Baets, B.; Verhoest, N.E.C. Impact of reducing polarimetric SAR input on the uncertainty of crop classifications based on the Random Forests algorithm. *IEEE Trans. Geosci. Remote Sens.* **2012**, *50*, 4185–4200.
38. Vincenzi, S.; Zucchetta, M.; Franzoi, P.; Pellizzato, M.; Pranovi, F.; de Leo, G.A.; Rorricelli, P. Application of a Random Forest algorithm to predict spatial distribution of the potential yield of *Ruditapes philippinarum* in the Venice Lagoon, Italy. *Ecol. Model.* **2011**, *222*, 1471–1478.
39. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: Berlin, Germany, 2002.
40. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874.
41. Peters, J.; Waegeman, W.; Van doninck, J.; Ducheyne, E.; Calvete, C.; Lucientes, J.; Verhoest, N.E.C.; De Baets, B. Predicting spatio-temporal *Culicoides imicola* distribution in Spain based on environmental habitat characteristics and species dispersal. *Ecol. Inform.* **2014**, *22*, 69–80.
42. Genuer, R.; Poggi, J.; Tuleau-Malot, C. Variable selection using Random Forests. *Pattern Recognit. Lett.* **2010**, *31*, 2225–2236.
43. Abulafia, D. *The Great Sea*; Penguin Books: London, UK, 2012.
44. Al-Bakri, J.; Suleiman, A. NDVI response to rainfall in different ecological zones in Jordan. *Int. J. Remote Sens.* **2004**, *25*, 3897–3912.

45. Gu, Y.; Hunt, E.; Wardlow, B.; Basara, J.B.; Brown, J.F.; Verdin, J.P. Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data. *Geophys. Res. Lett.* **2008**, *35*, L22401.