


RESEARCH

Open Access



# Data-driven and interpretable machine-learning modeling to explore the fine-scale environmental determinants of malaria vectors biting rates in rural Burkina Faso

Paul Taconet<sup>1,2\*</sup> , Angélique Porciani<sup>1</sup>, Dieudonné Diloma Soma<sup>1,2,3</sup>, Karine Mouline<sup>1</sup>, Frédéric Simard<sup>1</sup>, Alphonsine Amanan Koffi<sup>4</sup>, Cedric Pennetier<sup>1,2</sup>, Roch Kounbo Dabiré<sup>2</sup>, Morgan Mangeas<sup>5</sup> and Nicolas Moiroux<sup>1,2</sup>

## Abstract

**Background:** Improving the knowledge and understanding of the environmental determinants of malaria vector abundance at fine spatiotemporal scales is essential to design locally tailored vector control intervention. This work is aimed at exploring the environmental tenets of human-biting activity in the main malaria vectors (*Anopheles gambiae* s.s., *Anopheles coluzzii* and *Anopheles funestus*) in the health district of Diébougou, rural Burkina Faso.

**Methods:** *Anopheles* human-biting activity was monitored in 27 villages during 15 months (in 2017–2018), and environmental variables (meteorological and landscape) were extracted from high-resolution satellite imagery. A two-step data-driven modeling study was then carried out. Correlation coefficients between the biting rates of each vector species and the environmental variables taken at various temporal lags and spatial distances from the biting events were first calculated. Then, multivariate machine-learning models were generated and interpreted to (i) pinpoint primary and secondary environmental drivers of variation in the biting rates of each species and (ii) identify complex associations between the environmental conditions and the biting rates.

**Results:** Meteorological and landscape variables were often significantly correlated with the vectors' biting rates. Many nonlinear associations and thresholds were unveiled by the multivariate models, for both meteorological and landscape variables. From these results, several aspects of the bio-ecology of the main malaria vectors were identified or hypothesized for the Diébougou area, including breeding site typologies, development and survival rates in relation to weather, flight ranges from breeding sites and dispersal related to landscape openness.

**Conclusions:** Using high-resolution data in an interpretable machine-learning modeling framework proved to be an efficient way to enhance the knowledge of the complex links between the environment and the malaria vectors at a local scale. More broadly, the emerging field of interpretable machine learning has significant potential to help improve our understanding of the complex processes leading to malaria transmission, and to aid in developing

\*Correspondence: paul.taconet@ird.fr

<sup>1</sup> MIVEGEC, Université de Montpellier, CNRS, IRD, Montpellier, France  
Full list of author information is available at the end of the article



© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

operational tools to support the fight against the disease (e.g. vector control intervention plans, seasonal maps of predicted biting rates, early warning systems).

**Keywords:** Malaria, Anopheles, Biting behavior, Abundance, Ecological niche, Earth observation data, Statistical modeling, Cross-correlation maps, Random forest, Interpretable machine learning, Africa

## Background

Malaria is a vector-borne disease transmitted by *Anopheles* mosquitoes still affecting 229 million people and causing more than 400,000 deaths worldwide annually [1]. Malaria control efforts, mainly through the massive use of long-lasting insecticidal nets [2], led to a sustained decrease of the disease burden between 2000 and 2015 [1]. However, malaria cases have plateaued in the past 5 years or even increased in certain areas [1]. Vector resistance to insecticides, population growth and environmental changes are involved in such worrying trends [3, 4]. For effective and sustainable vector control (VC), locally tailored interventions, built on a thorough knowledge of the local determinants of malaria transmission, are needed [3–5]. To do so, it is of particular importance to decipher with vector bio-ecology at fine and operational spatiotemporal scales [3, 4, 6].

To develop efficient (i.e. species-, place- and time-specific) vector control strategies, important features of malaria vector ecology such as breeding site typologies, development and survival rates, flight ranges, or dispersal have to be considered. Meteorological conditions (temperature, precipitation) and land cover are major environmental factors frequently used to define the ecological niche of malaria mosquitoes [5] in complex, sometimes hardly hypothesizable, ways. Temperature affects the mosquito life history traits, nonlinearly, at each stage of its life cycle (e.g. larval growth, adult survival, biting rate). For example, the daily mortality rate of several adult *Anopheles* species follows a unimodal relationship with air temperature, with an optimal adult survival rate at around 25 °C [7–9]. Rainfall generates additional mosquito breeding sites and is therefore an important factor explaining the seasonality in species abundance. However, excessive rainfall can destroy developing larvae by flushing them out of their aquatic habitat [10, 11]. Land cover may affect mosquito population dynamics by creating breeding sites in hydromorphic areas or altering the dispersal ability of mosquitoes. A modification of land cover/use may therefore either increase or decrease vector abundance relative to species ecological preferences. As an example, deforestation can increase larval breeding sites of malaria vectors growing in sunny puddles, whereas it destroys habitats of some deep-forest *Anopheles* species [5, 12]. Moreover, even when found together, *Anopheles* species often exhibit specific ecological

preferences [13, 14]. As an example, *Anopheles gambiae* s.s. was more frequently observed in temporary, rainfall-dependent breeding sites [15–17], whereas *Anopheles coluzzii* showed a preference for more permanent breeding sites [17–19]. Altogether, these examples illustrate that vector ecology is finely tuned with the environment. Using large-scale environmental indicators could therefore jeopardize the characterization of ecological niches of malaria vectors and consequently lead to suboptimal or even inappropriate VC intervention at a smaller scale [5]. To overcome this issue, we propose the use of high-resolution Earth observation (EO) data and develop novel statistical modeling approaches.

Indeed, in malariometric statistical modeling studies, “data” models [20] like linear or logistic regression are traditionally used with environmental variables extracted from EO data [21–23]. These models are well suited for testing pre-established hypotheses about theoretical constructs (e.g. to answer questions like “how much higher is mosquito abundance for each additional millimeter of rainfall?”); however, to explore hypotheses and extract knowledge in complex systems, machine-learning (ML) “algorithmic” models might be more suitable [20, 24]. In fact, these models are inherently able to capture complex patterns (such as nonlinear relationships and complex interactions between variables) contained in data. After a good predictive algorithm is fitted to a dataset, post hoc interpretation methods may uncover the complex relationships contained in the data and learned by the model, which in turn can be carefully linked to prior knowledge to identify meaningful—possibly unforeseen—cause-effect relationships, valuable thresholds or interactions [25–27]. This predict-then-explain modeling workflow is being increasingly used to generate knowledge from complex datasets [24, 26] and is commonly referred to as “interpretable machine learning” (IML) [25, 26].

The main objective of this study was to improve our overall understanding of the ecological niche and determinants of the biting rates of the main malaria vectors in a rural area of southwestern Burkina Faso. To do so, we used entomological collections and environmental variables extracted from high-resolution EO data, in a data-driven and IML modeling framework. In this research article, after a presentation of the methods and results, we discuss the environmental (landscape and meteorological) drivers of the human-biting activity of the main malaria

vectors in the Diébougou area. We also briefly present some potential practical uses of our results to support the conceptualization and deployment of locally tailored VC interventions. We conclude with methodological insight regarding the use of algorithmic models and IML for knowledge-building in the field of landscape entomology.

## Methods

### Data collection and preparation

#### Entomological data

*Anopheles* human-biting activity was monitored as part of a study carried out in the Diébougou rural health district located in southwest Burkina Faso [28]. Twenty-seven villages in this 2500 km<sup>2</sup> wide area were selected according to the following criteria: accessibility during the rainy season, 200–500 inhabitants per village, and distance between two villages greater than 2 km. Seven rounds of mosquito collection were conducted in each village between January 2017 and March 2018. The periods of the surveys span some of the typical climatic conditions of this tropical area (three surveys in the “dry-cold” season, two in the “dry-hot” season, one at each extremum of the rainy season) (see Additional file 1: Summary of the meteorological conditions around the sampling points).

Mosquitoes were collected using the human landing catch (HLC) technique from 17:00 to 09:00 both indoors and outdoors at four sites per village for one night during each survey. The procedure for conducting HLC was for a person to sit on a stool, and mosquitoes to alight on his exposed legs, where they were then collected using a hemolysis tube. Collectors were rotated hourly between collection sites and/or position (indoor/outdoor). Independent staff supervised rotations and regularly checked the quality of mosquito collections. Malaria vectors were identified using morphological keys [29, 30]. Individuals belonging to the *Anopheles gambiae* complex and the *Anopheles funestus* group were identified to species by PCR [31–33]. Mosquito collection design for this study has been described extensively elsewhere [28].

HLC enabled us to measure the presence and abundance of aggressive malaria vectors in time and space. In fact, landing on human legs is the behavioral event preceding the biting event. To avoid exposing mosquito collectors to infectious bites, we used landing as a proxy for biting, and in turn biting probability/rate as a proxy for the overall presence/abundance of aggressive vectors at the time and place of collection.

#### Landscape data

A land cover map of the study area was produced by carrying out a geographic object-based image analysis (GEOBIA) [34] using multisource very-high- and high-resolution satellite-derived products. The GEOBIA

involved the following main steps: acquisition/collation of the satellite products (Satellite Pour l'Observation de la Terre (SPOT)-6 image acquired on 2017-10-11, Sentinel-2 image acquired on 2018-11-16, and a digital elevation model (DEM) from the Shuttle Radar Topography Mission [35]), acquisition of a ground-truth dataset composed of 420 known land cover samples by both fieldwork (held in November 2018) and photo-interpretation of satellite images, and classification of the land cover over the whole study area using a random forest (RF) algorithm [36]. The definitions of land cover classes were those proposed by the Permanent Interstate Committee for Drought Control in the Sahel [37]. The resulting dataset was a georeferenced raster image, where each 1.5 × 1.5 m pixel was assigned a land cover class. The confusion matrix was generated using the internal RF validation procedure based on the out-of-bag observations, and the quality of the final classification was assessed by calculating the overall accuracy from the confusion matrix [38].

Spatial buffers were then defined to characterize the environmental conditions at the neighborhood of each HLC collection point. Four buffer radii were considered: 250 m, 500 m, 1 km, 2 km. The distance of 2 km was chosen as the largest radius to minimize overlaps among buffers coming from different villages and because local dispersal of *Anopheles* beyond this distance can be considered negligible [29, 39–41]. We calculated the percentage of landscape occupied by each land cover class in each buffer zone around each collection site.

Additional indices related to the presence of water were calculated. The theoretical stream network was produced for the study area by first generating a flow accumulation raster dataset from the DEM and then applying a threshold value to select cells with accumulated flow greater than 1000 [42]. The quality of the product was assessed visually by overlaying it on the SPOT-6 satellite image. We then derived two indices for each collection site: the length of streams in each buffer zone, and the shortest distance to the streams.

In order to describe attractiveness and penetrability of households for malaria vectors, the geographical location of the households in the villages were recorded and two indices were computed. First, the Clark and Evans aggregation index [43] was calculated to describe the degree of clustering of the households in each village, as it has been suggested that scattered habitations in a village might increase the attractiveness for some vector species [44]. Second, we calculated the distance from each collection point to the edge of the village (defined as the convex hull polygon of each village—i.e. the minimum polygon that encompasses all the locations of the households), as it has been suggested elsewhere that living on the edge can increase biting rates [45].

### **Meteorological data**

Daily rainfall estimates were extracted from the Global Precipitation Measurement (GPM) Integrated Multi-satellite Retrievals for GPM (IMERG) Final products [46]. The raw satellite products were resampled from their original 10-km spatial resolution to a 1-km resolution using a bilinear interpolation method.

Daily diurnal and nocturnal temperatures were derived from the Moderate Resolution Imaging Spectroradiometer (MODIS) Land Surface Temperature (LST) Terra and Aqua products [47, 48]. Terra and Aqua daily products were first combined, keeping the highest (or lowest) available pixel values for the diurnal (nocturnal) temperature. Missing values in pixels (mostly due to cloud presence) were then filled by temporally interpolating the values of the closest preceding and following available dates.

These meteorological data (daily rainfall, daily diurnal temperatures, daily nocturnal temperatures) were collected up to 42 days (i.e. 6 weeks) preceding each mosquito collection, so as to encompass largely the whole duration of the *Anopheles* life cycle in the field (including aquatic and aerial stages) [49]. They were then aggregated pixel-by-pixel on a weekly scale (cumulative 7-day rainfall and average 7-day diurnal and nocturnal LST). This temporal granularity represents a reasonable trade-off between the raw, daily information—which might overfit in the statistical models—and larger scales, which might prevent them from capturing fine-scale temporal relationships. Next, we calculated the cumulative rainfall and average temperatures for all possible intervals of time available in the data (e.g. b/w 0 and 1 week before the dates of collection, b/w 0 and 2 weeks, b/w 1 and 2 weeks, etc.). The data were finally averaged in the 2-km buffer zone only (considering the 1-km spatial resolution of the source data).

### **Statistical analyses**

#### **Overall approach**

We used a two-step statistical modeling approach to study the relationships between the biting rates of each vector species and the environmental conditions. We first calculated correlation coefficients between the biting rates and the environmental variables at the various buffer sizes/time lags considered. The objectives of this bivariate analysis were twofold: (i) to better apprehend several aspects of the ecology of the vectors in the study area, and (ii) to screen out variables for the multivariate analysis. In a second stage, we integrated selected variables in multivariate algorithmic models that we further analyzed using interpretable machine-learning tools, to search for potential complex links (nonlinear

relationships, relevant thresholds) between the environmental factors and the biting rates.

We ran the whole modeling framework separately for each species, as they might exhibit different ecological preferences.

From a statistical point of view, most algorithmic machine-learning models, although nonparametric, have difficulty coping with zero-inflated negative binomial response variables [50, 51], which are typically found in insect count data such as mosquito biting rates [52]. An alternative approach to model such data is the hurdle model that considers the data responding to two processes: one causing zero versus nonzero and the second process explaining the nonzero counts [53]. The hurdle methodology in the frame of a widely used algorithmic model (random forest) was proposed elsewhere to deal with such distributions of data [51]. Besides, this separation is biologically pertinent since it has been shown that the drivers of the presence might differ from those of the abundance [17, 44, 54]. Lastly, separate modeling of presence and abundance might enable us to identify distinct targets for vector control answering to, respectively, eradication (absence of bites) and control (reduction of the number of bites) [17].

We therefore separately modeled the probability of human–vector contact (called “presence” models in the rest of this article) and the positive counts of human–vector contact (called “abundance” models). Given that HLC data are used as a proxy for human-biting rate, presence models analyzed the probability of at least one individual biting a human during a night, while abundance models analyzed the number of bites received by one human in one night conditional on their presence (i.e. zero-truncated data). Hence, in our presence models, the dependent variable was the presence/absence of vectors (binarized as 1/0) collected during 1512 nights of HLC (27 villages  $\times$  4 collection sites  $\times$  2 places (indoors and outdoors)  $\times$  7 surveys), while in the abundance models, the dependent variable was the number of bites per human during the positive catch sessions—i.e. the sessions with at least one bite.

#### **Bivariate analysis using correlation coefficients**

The bivariate relationship between the presence/abundance of each vector species and the environmental variables was assessed using multilevel Spearman correlation coefficients [55] with the village entered as a random effect. Multilevel correlations, contrary to simple correlation, account for non-independency between observations in a dataset, by introducing a factor as a random effect in the correlation (on the same principle as random effects in mixed linear regressions).



**Landscape variables:** The correlation coefficient was calculated for each landscape variable (i.e. percentage of landscape occupied by each land cover class in each buffer zone).

**Meteorological variables:** Past weather is likely to influence the size of the sampled mosquito generation with varied delays. For example, (i) past weather in the week preceding the collection may influence adult survival rates of the collected generation, (ii) weather during 1 or 2 weeks preceding the collection may influence the development rates of the collected generation during the larval stages, and (iii) weather beyond the third week preceding the collection date may influence the development rates of parent generations. For the meteorological variables, cross-correlation maps (CCM) were hence computed [56] to assess the relationships between the biting rates and the precipitation and temperatures preceding the dates of collection. A CCM enables one to study the influence of environmental conditions during time intervals (instead of single time points) prior to the collection event. CCMs hence allowed us to account for the effects of cumulative precipitation and average temperature on the collected mosquito generation over intervals of weeks preceding the bites (e.g. average diurnal temperature between 1 and 3 weeks preceding the bite), instead of single weeks (e.g. average diurnal temperature during the third week preceding the bite).

#### ***Multivariate analysis using random forests and interpretable machine learning***

We used the results of the bivariate analysis to select the environmental variables to include as predictors in the multivariate analysis. We first excluded variables that were poorly correlated with the response variable (i.e. correlation coefficients less than 0.1 or  $p$ -values greater than 0.2 at all time associations or buffer radii considered), except for variables related to the presence of water—i.e. possible breeding sites—which were all retained whatever their correlation. Then, for each meteorological (or landscape) variable, we retained the time lag interval (buffer radius) showing the higher absolute correlation coefficient value (see Additional file 6: Feature selection for the multivariate models). Because the entomological data used in this study were part of a trial, different vector control strategies were implemented in the villages of the study after the third survey. The implemented VC strategies and the place of collection (interior/exterior) were therefore introduced as adjustment variables in our models, but their effect on the biting rate was found to be negligible in our analysis (see “Results” section), and these results will not be discussed further.

Random forest classifiers were then trained for each species and response variable (presence and abundance models). Random forests are an ensemble machine-learning method that generates a multitude of random decision trees that are then aggregated to compute a classification or a regression [36]. They are known for their good predictive capacity, which is mainly due to their ability to inherently capture complex associations between the variables [20]. Binary classification RFs were generated for the presence models and regression RFs were generated for the abundance models. The modeling process involved the following steps:

- **Feature collinearity:** Collinear covariates (i.e. Pearson correlation coefficient > 0.7) were checked for and removed based on empirical knowledge.
- **Feature engineering:** In the classification models, data were up-sampled within the model resampling procedure to account for the imbalanced structure of the response variable [57]. In the regression models, the response variables were log-transformed prior to the model resampling procedure in order to reduce their overdispersion.
- **Model training, tuning, selection:** Model hyperparameters were optimized using a random 10-combination grid search [58]. For each set of hyperparameters tested, a leave-village-out cross-validation (LVO-CV) resampling method was used. The resampling method involved training the model using in turn the data from 26 of the 27 sampled villages, validating with the data from the remaining village using a predictive performance metric [for the presence model: the precision–recall area under the curve (PR-AUC); for the abundance model: the mean absolute error (MAE)] and averaging the metric across all hold-out predictions at the end of the procedure. The model retained was the one leading to the highest overall PR-AUC (lowest MAE). The retained model was then fit to all the observations and further used for the interpretation phase.
- **Model evaluation:** The predictive power of each model was assessed by LVO-CV. We hence evaluated the ability of the models to predict the presence or abundance of vectors on unseen nights of HLC, whilst excluding from the training sets all the observations belonging to the village of the evaluated observation. Doing so enables us to limit overfitting and over-optimistic performance metrics due to spatial autocorrelation [59]. For the presence models, precision–recall plots were then generated from the observed and predicted values, and the PR-AUC was calculated and compared to the baseline of PR curve (i.e. the PR-AUC of a random-guess classifier

for the dataset). The PR-AUC is a measure of predictive accuracy of a binary classification model particularly suitable for imbalanced classification problems [60]. It makes sense when compared to the baseline PR-AUC, which is the rate of “presence” observations in the dataset. For example, a model with a baseline of 0.01 and a PR-AUC of 0.2 performs  $0.2/0.01 = 20$  times better than a random-guess, or no-skill, classifier. We also calculated sensitivity and specificity at the optimal probability threshold (i.e. the one maximizing the AUC). For the abundance models, a visual evaluation (i.e. graphical comparison between observed and predicted values) was preferred to a numerical one because performance metrics were expected to be low given the overdispersion of the response data and the type of model used [51]. Evaluation plots for the abundance models included (i) the distribution of MAEs and (ii) observed versus predicted values for each out-of-sample village.

To interpret the models, we further generated permutation-based variable importance plots (VIPs) [36] and partial dependence plots (PDPs) [61] including standard deviation bands (that can be interpreted as confidence intervals). These plots, part of the interpretable machine-learning toolbox, enable us to study the effects of one predictor on the response variable while accounting for the effect of the other predictors in the model [25]. Variable importance measures a feature’s importance by calculating the degradation of the predictive accuracy of the model after randomly permuting the values of the feature: the higher a variable’s importance, the more that variable contributes to the prediction. Partial dependence plots, on their side, show the marginal effect that one feature has on the predicted outcome [25]. PDPs hence help visualize the relationship, learned by a model, between a feature and the response. A PDP is likely to reveal complex (nonlinear, nonmonotonic) effects when a model has learned such relationships. Importantly, the information provided by these tools should be trusted only if the underlying model has good predictive power [27].

We finally identified primary and secondary predictors for each model, according to the following criteria. Primary predictors were the top three most important predictors of the VIP. Secondary predictors were variables either presenting marked variations in their PDP (e.g. thresholds, significant slopes) or known to influence the bio-ecology of the vector.

#### Software used

The software packages used in this work were all free and open-source. The R programming language [62] and the RStudio environment [63] were used as the main

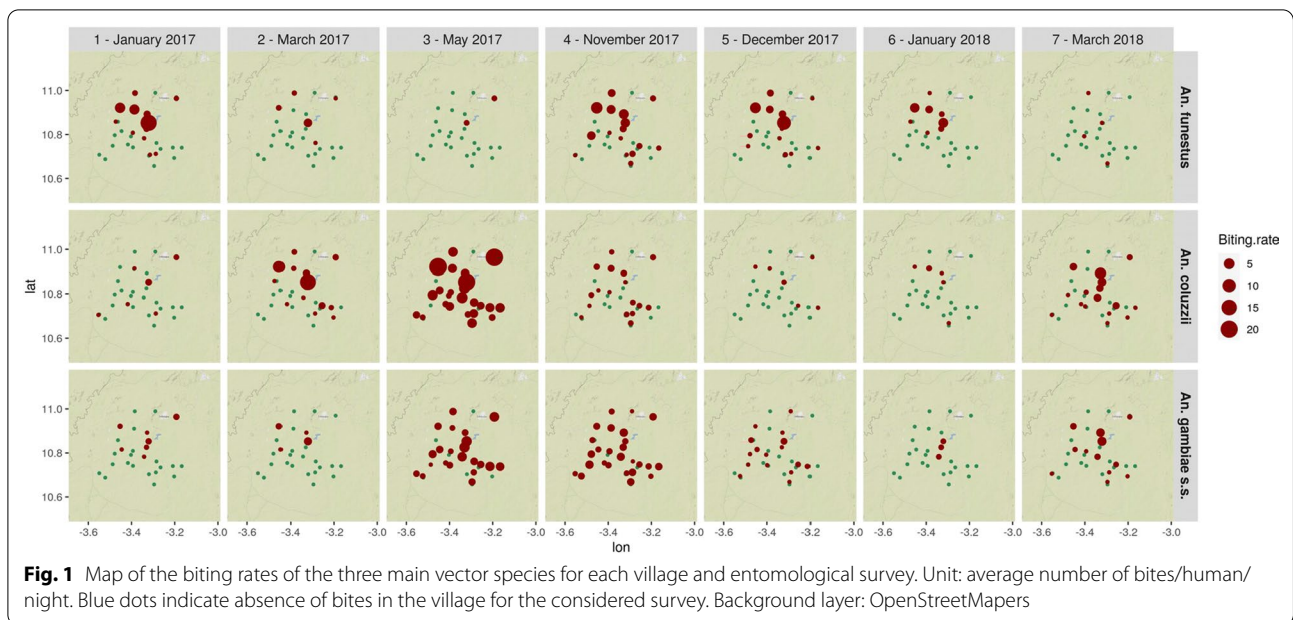
programming tools. An R package was developed [64] to extract the NASA meteorological data (MODIS and GPM). The land cover layer was generated using the following R packages: “RSAGA” [65], “rgrass7” [66], “raster” [67], “sf” [68], “rgdal” [69] and “randomForest” [70]. The “spatstat” [71] package was used to compute the Clark and Evans aggregation index. The QGIS software [72] was used to create the map of the study area. The “landscapemetrics” package [73] was used to calculate the percentage of landscape occupied by each land cover class in the buffer areas. The “correlation” [55] package was used for the correlation analysis. The “caret” [74] and “ranger” [75] packages were used to fit the random forest models in the statistical analysis. The “CAST” [76] package was used to create the temporal folds for cross-validation. The “MLmetrics” [77] package was used to calculate the model evaluation metrics. The “iml” [78] and “pdp” [79] packages were used to generate the partial dependence plots. The “patchwork” [80] package was used to create various plot compositions. The “ggmap” [81] package was used to generate the map of the vector biting rates. The “precrec” [82] package was used to generate the precision–recall plots for the presence models. The “tidyverse” meta-package [83] was used throughout the entire analysis.

## Results

### Entomological data

A total of 1512 nights of HLC were conducted among the 27 villages during the seven entomological surveys. Altogether 3056 vectors belonging to the *Anopheles* genus were collected: 1322 *An. coluzzii*, 708 *An. funestus*, 616 *An. gambiae s.s.* and 410 from other species. *An. funestus* was present in 12% of the nights of HLC (182 times), while both *An. coluzzii* and *An. gambiae s.s.* appeared on 20% of the nights of HLC (respectively 297 and 302 times). The distribution of the biting rates in the positive sessions (i.e. sessions with at least one bite) was highly left-skewed (for *An. funestus*: median = 2, SD = 4.7, max. = 36; for *An. gambiae s.s.*: median = 2, SD = 1.5, max. = 10; for *An. coluzzii*: median = 2, SD = 6.8, max. = 50).

Figure 1 shows the distribution of the biting rates of the three main vector species by village and survey. Overall, the map reveals heterogeneous spatiotemporal patterns of biting rates for the three main species. *An. funestus* was found in a few villages only, mainly at the end of the rainy season (November) and in the dry-cold season (December, January) (see Additional file 1: Summary of the meteorological conditions around the sampling point). It almost disappeared during the dry-hot season (March) and the beginning of the rainy season (May). *An. gambiae s.s.* and *An. coluzzii* were found



in almost all the villages at the beginning and the end of the rainy season (May, November). They were also found year-round in some villages, in particular, those located close to dams or to the Bougouriba River. *An. coluzzii* was particularly abundant at the beginning of the rainy season (May), while *An. gambiae s.s.* was found in similar abundance at the beginning and the end of the rainy season (May, November).

#### Land cover map

Eleven land cover classes were discriminated: ligneous savanna (52% of the total surface), crop (25%), grassland (7%), marsh (5%), riparian forest (4%), woodland (3%), rice (1%), settlements (0.5%), bare soil (0.5%), main roads (0.3%) and permanent water bodies (0.3%). In the buffer areas considered for the modeling study (250 m, 500 m, 1 km, 2 km radii), similar trends were observed regarding the percentage of area occupied by each land cover class (see Additional file 2: Summary of the landscape conditions around the sampling points). Ligneous savanna included shrub savanna, tree savanna and wooded savanna. Grassland included herbaceous savanna and Sahelian short grass savanna. Permanent water bodies included dams and the Bougouriba River. Marshlands included wetland–floodplain and agriculture in shallows and recessions. The overall accuracy of the classification was 0.84. The resulting land cover map of the study area, including the geographical position of the study villages, is presented in Fig. 2. Pictures representative of the main land cover classes are

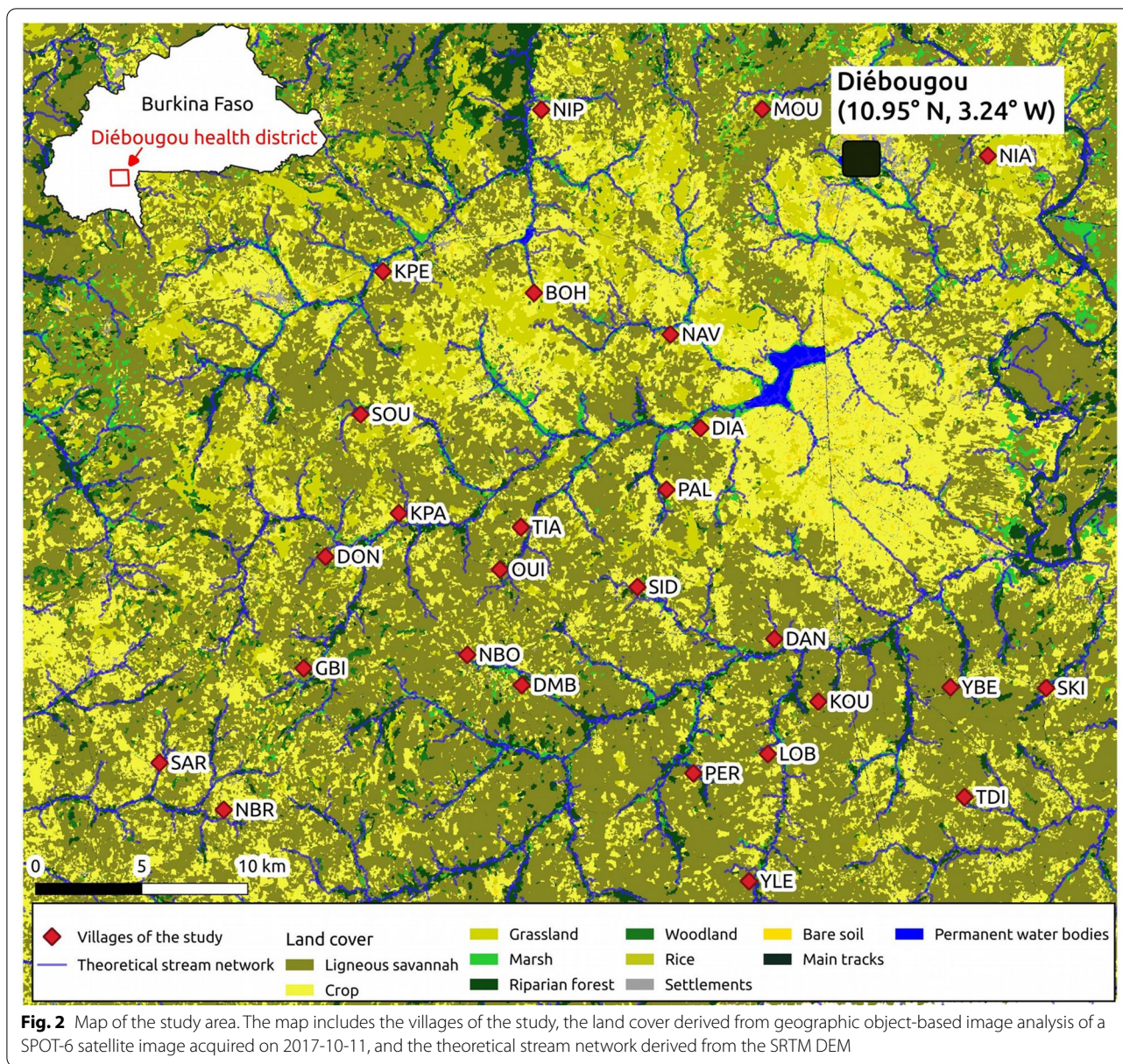
provided in Additional file 3: Pictures representative of the main land cover classes in the Diébougou area.

#### Bivariate analysis

Figure 3 shows the landscape variables that were significantly correlated [multilevel Spearman's correlation coefficient ( $cc$ ) > 0.1 and  $p$ -value < 0.2] with the presence or abundance of each of the studied vector species. The presence or abundance of *An. funestus* was correlated to two to six landscape variables depending to the buffer radius considered, one to five variables for *An. gambiae s.s.* and two to five for *An. coluzzii*. Overall, among the three species, the highest correlation coefficients with the landscape variables were observed for *An. funestus*.

Both the presence and the abundance of *An. funestus* were positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. They were also positively correlated with the % of surface occupied by marshlands in all buffer zones with radius  $\geq 500$  m, with increasing correlation coefficients as the buffer zone radii increased. The presence and the abundance of *An. funestus* were also positively correlated with the % of surface occupied by grasslands, and negatively correlated with the % of surface occupied by ligneous savannas, for all buffer radii. The correlation between the abundance and the % of surface occupied by grasslands and ligneous savannas increased (both in absolute value and significance) with smaller buffer radii. The presence of *An. funestus* was positively correlated with the % of surface occupied by crops in all buffer zones





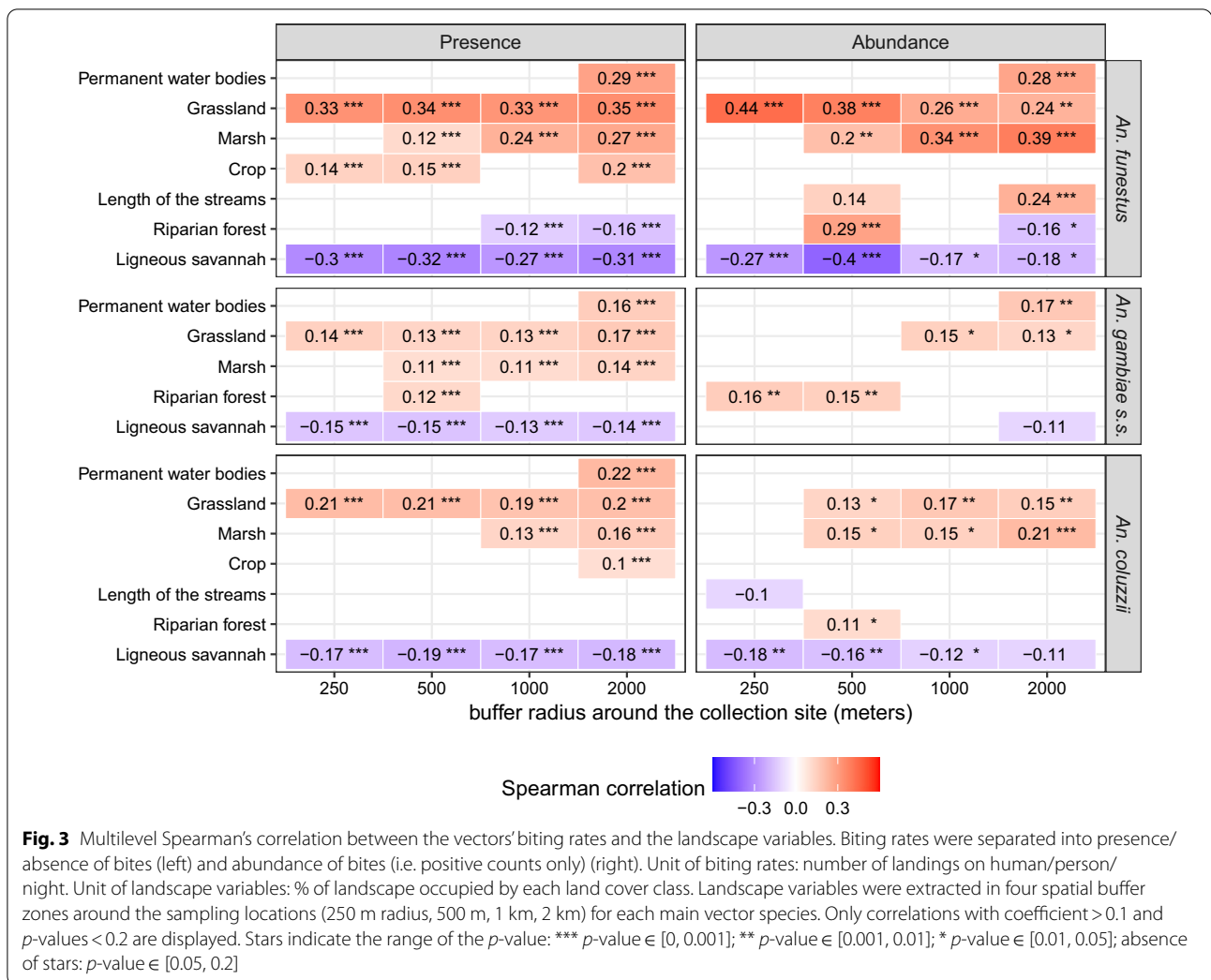
except for the 1-km radius. The abundance of *An. funestus* was positively correlated with the length of streams in the 250-m and the 2-km radii buffer zones.

Both the presence and the abundance of *An. gambiae s.s.* were positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. Its presence was also positively correlated with the % of surface occupied by marshlands in all buffer zones with radius  $\geq 500$  m. The presence and abundance of *An. gambiae s.s.* were also positively correlated with the % of surface occupied by grasslands (in all buffer zones for the presence, and in the buffer zones with radius  $\geq 1$  km for the abundance), and negatively correlated with the % of

surface occupied by ligneous savannas (in all buffer zones for the presence, and in the 2-km radius buffer zone for the abundance). The presence and abundance of *An. gambiae s.s.* were also positively correlated with the % of surface occupied by riparian forests, only in the 500-m radius buffer zone for the presence, and in buffer zones with radius  $\leq 500$  m for the abundance.

The presence of *An. coluzzii* was positively correlated with the % of surface occupied by permanent water bodies in the 2-km radius buffer zone. The presence and the abundance of that species were positively correlated with the % of surface occupied by marshlands in all buffer zones with radius  $\geq 1$  km for the presence, and in



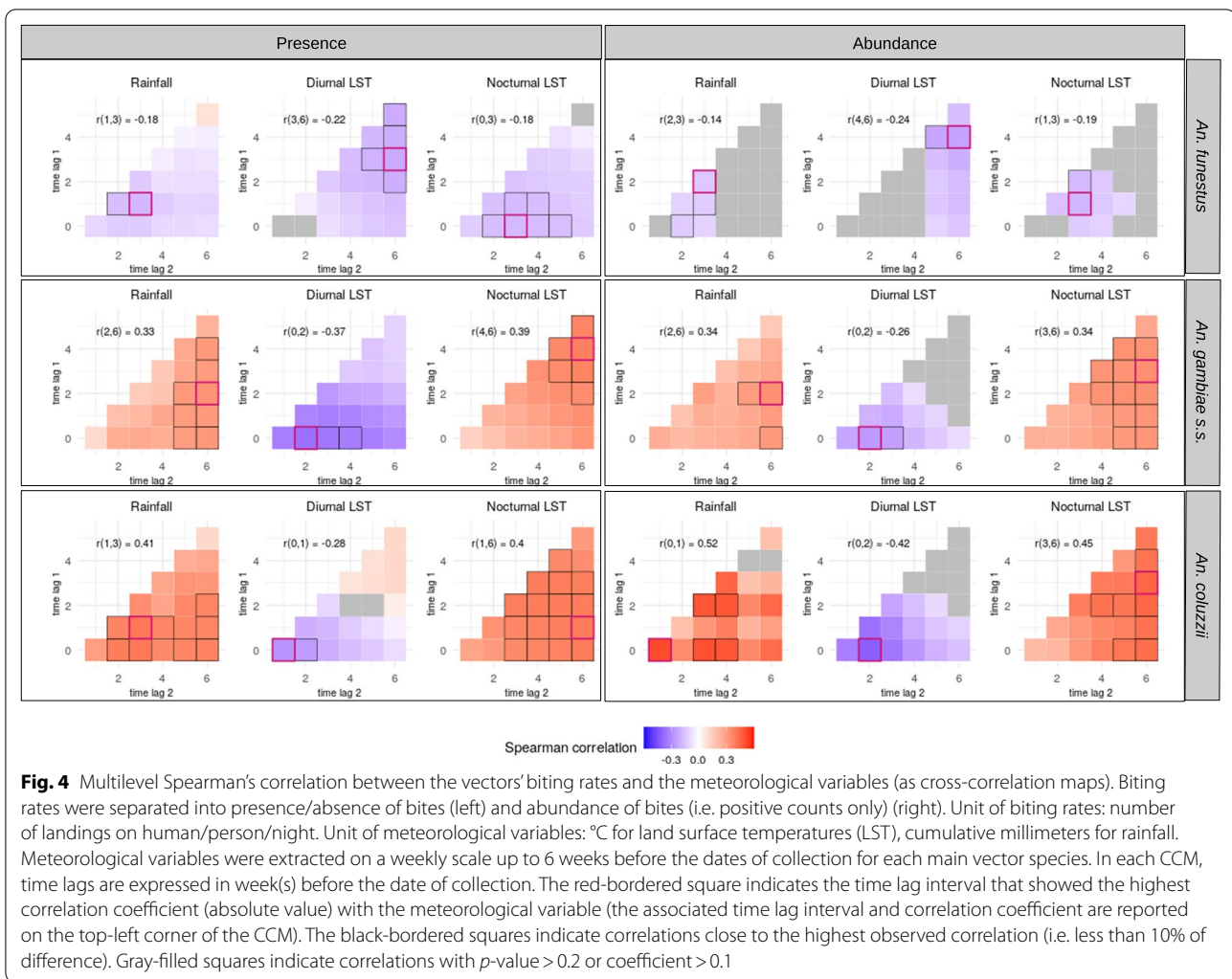


all buffer zones with radius  $\geq 500$  m for the abundance. Presence and abundance of *An. coluzzii* were also positively correlated with the % of surface occupied by grasslands, and negatively correlated with the % of surface occupied by ligneous savannas, in all the buffer zones (except in the 250-m radius buffer zone for the abundance). The correlation between the abundance of *An. coluzzii* and the % of surface occupied by ligneous savannas increased (in both absolute value and significance) with smaller buffer zones.

Figure 4 shows the meteorological variables that were significantly correlated [multilevel Spearman correlation coefficient (cc) > 0.1 and p-value < 0.2] with the presence or abundance of bites for each of the studied vector species (in the form of cross-correlation maps). Overall, among the three species, the highest correlation coefficients with the meteorological variables were observed for *An. coluzzii*, closely followed by *An. gambiae s.s.*

The presence and abundance of *An. funestus* showed quite weak correlations with the meteorological variables (Spearman's correlation coefficient always < 0.25, with all the meteorological variables at all time frames) when compared to *An. gambiae s.s.* or *An. coluzzii*. Correlations between both response variables (presence and abundance) and the three meteorological variables (cumulative rainfall, diurnal LST, nocturnal LST), when significant, were negative. The maximum correlation coefficients between each meteorological variable and both the presence and abundance were found for the following: cumulative rainfall recorded b/w 1–2 and 3 weeks before the date of collection, diurnal temperatures recorded b/w 3–4 and 6 weeks before the date of collection, and nocturnal temperatures recorded b/w 0–1 and 3 weeks before the date of collection.

The presence and abundance of *An. gambiae s.s.* were positively correlated with cumulative rainfall, at all time lags. The maximum correlation coefficients for both



presence and abundance were found for cumulative rainfall recorded b/w 2 and 6 weeks before the date of collection. The presence and abundance of *An. gambiae s.s.* were also positively correlated with nocturnal temperatures at all time lags, and the maximum correlation coefficients with both response variables were found for temperatures recorded b/w 3–4 and 6 weeks before the date of collection. The presence and the abundance of *An. gambiae s.s.* were negatively correlated with diurnal temperatures preceding the date of collection at almost all time lags. The maximum correlation coefficient with both response variables was found for temperatures recorded b/w 0 and 2 weeks before the date of collection.

The correlations between meteorological variables and both the presence and abundance of *An. coluzzii* exhibited similar trends as *An. gambiae s.s.*, with few notable differences. The presence and abundance of *An. coluzzii* were positively correlated with cumulative rainfall preceding the date of collection at all time lags. The

maximum correlation coefficient with cumulative rainfall was found b/w 1 and 3 weeks before the date of collection for presence, and b/w 0 and 1 week before the date of collection for abundance (the correlation coefficient b/w 1 and 3 weeks was also among the highest for the abundance). The presence and abundance of *An. coluzzii* were positively correlated with nocturnal temperatures at all time lags with maximum correlation coefficients found b/w 1–3 and 6 weeks before the date of collection for both response variables. The presence and abundance of *An. coluzzii* were, overall, negatively correlated with diurnal temperatures preceding the date of collection. The maximum correlation coefficient between diurnal temperatures and both response variables was found b/w 0 and 1–2 weeks before the date of collection.

**Multivariate analysis**

The PR-AUC of the presence models were 0.56 (baseline = 0.12), 0.46 (baseline = 0.20) and 0.60

(See figure on next page.)

**Fig. 5** Interpretation plots of the random forest models for *An. funestus*. Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function  $\pm$  one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

(baseline = 0.20) for *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, respectively. The specificity and sensitivity of the models at the optimal probability thresholds were respectively 80% and 73% for *An. funestus*, 75% and 76% for *An. gambiae s.s.*, and 79% and 75% for *An. coluzzii*. Overall, these results indicate good predictive accuracy of the presence models. The abundance models reflected the trends well for the three species, although they often underestimated high counts. The model evaluation plots are available in Additional file 4: Model evaluation plots for the presence models and Additional file 5: Model evaluation plots for the abundance models (for the presence models: precision–recall plots and observed versus predicted values for each out-of-sample village; for the abundance models: distribution of MAE and observed versus predicted values for each out-of-sample village). Figures 5, 6 and 7 show the model interpretation plots (variable importance plot and partial dependence plots) for *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, respectively.

The most important predictors of the presence and abundance of *An. funestus* were landscape-based, including % of surface occupied by marshlands (in the 2-km radius buffer zone), grasslands (in the buffer zone radii  $\leq$  500 m) and ligneous savannas (in the 500-m radius buffer zone). The probability of the presence of *An. funestus* increased linearly with surface occupied by marshlands in the range available in the data (0–10%), while the abundance was constant in the range of 0–3%, increased approximately linearly from 3 to 6%, and finally stabilized in the range of 6–10%. Both the probability of presence and the abundance increased linearly with surface occupied by grasslands in the range of 0–20%, and stabilized above that threshold. Conversely, they decreased linearly with surface occupied by ligneous savannas in the range of 0–50%, and above that threshold stabilized for abundance and tended to diminish (with a lower trend though) for presence.

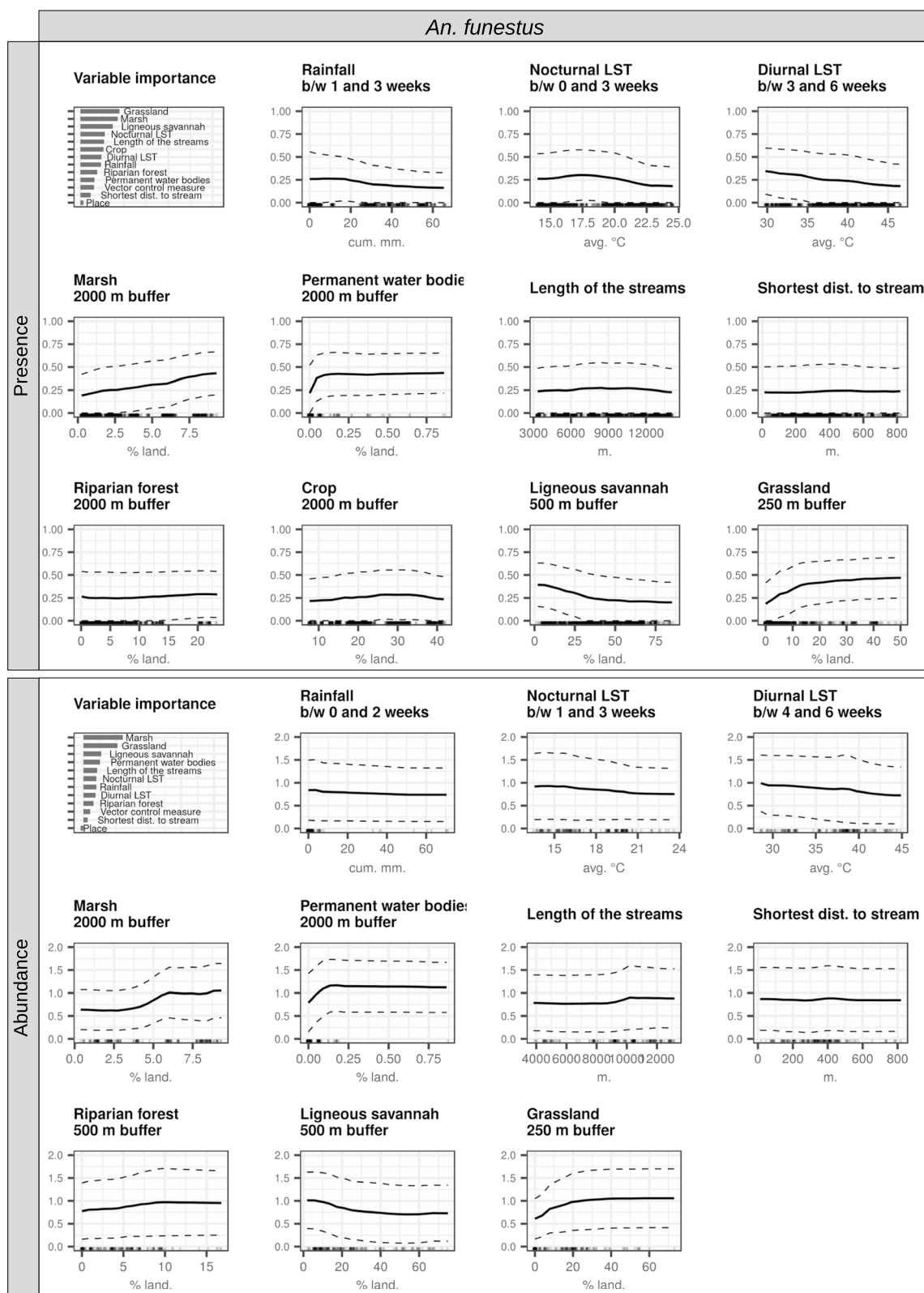
Secondary predictors of the presence and abundance of *An. funestus* were as follows: % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range

0.1–1%), diurnal LST b/w 3 to 4 weeks and 6 weeks before the date of collection (negative, approximately linear, association), and nocturnal LST b/w 0 to 1 and 3 weeks before the date of collection (stable in the range 14–19 °C, decrease in the range 19–25 °C).

The most important predictors of the presence of *An. gambiae s.s.* were meteorological variables, namely nocturnal LST (b/w 4 and 6 weeks, i.e. 28 and 42 days, before the date of collection), rainfall (b/w 2 and 6 weeks before the date of collection), and diurnal LST (b/w 0 and 2 weeks before the date of collection). The probability of presence of *An. gambiae s.s.* increased slowly for nocturnal LSTs in the range of 14–20 °C and more rapidly above that threshold. It increased linearly with the cumulative rainfall in the range of 0–10 mm, and was stable above that threshold. It decreased for diurnal LSTs in the range of 35–41 °C, and stabilized above that threshold. Secondary predictors of the presence of *An. gambiae s.s.* were as follows: % of surface occupied by grasslands in the 2-km radius buffer zone (increase in the range 0–15%, stable above), % of surface occupied by ligneous savannas in the 500-m radius buffer zone (negative linear association), and % of surface occupied by marshlands in the 2-km radius buffer zone (positive linear association).

When *An. gambiae s.s.* was present, cumulative rainfall (b/w 2 and 6 weeks before the date of collection) was, by far, the most important predictor of its abundance. Other primary predictors were diurnal temperatures (b/w 0 and 2 weeks before the date of collection) and % of surface occupied by marshlands (in the 2-km radius buffer zone). The abundance of *An. gambiae s.s.* increased linearly in the range of 0–50 mm cumulative rainfall and stabilized above that threshold (range 50–80 mm). It slowly increased with the % of surface occupied by marshlands. Secondary predictors of the abundance of *An. gambiae s.s.* included the following: % of surface occupied by ligneous savannas in the 2-km radius buffer zone (negative linear association), % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range 0.1–1%), % of surface occupied by riparian forests in the 250-m radius buffer zone (positive linear association), and % of surface





**Fig. 5** (See legend on previous page.)

(See figure on next page.)

**Fig. 6** Interpretation plots of the random forest models for *An. gambiae* s.s. Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function  $\pm$  one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

occupied by grasslands in the 1-km radius buffer zone (positive linear association).

The most important predictors of the presence of *An. coluzzii* were as follows: cumulative rainfall (b/w 1 and 3 weeks before the date of collection), nocturnal LST (b/w 1 and 6 weeks before the date of collection), and % of surface occupied by marshlands (in the 2-km radius buffer zone). A total of approximately 40 mm of rainfall b/w 1 and 3 weeks before the date of collection was enough to double the probability of presence of *An. coluzzii* (from an average 0.25 without rainfall to 0.55). Beyond that amount of rainfall, the probability of presence tended to diminish (range 50–65 mm). The probability of presence of *An. coluzzii* increased linearly with nocturnal LSTs and % of surface occupied by marshlands. Secondary predictors of the presence of *An. coluzzii* included the following: diurnal LST b/w 0 and 1 week before the date of collection (decrease in the range 34 °C–40%, stable in the range 40–50 °C), % of surface occupied by ligneous savannas in the 500-m radius buffer zone (decrease in the range 0–40%, stable above), % of surface occupied by grasslands in the 250-m radius buffer zone (increase in the range 0–30%, stable above), and % of surface occupied by permanent water bodies in the 2-km radius buffer zone (increase in the range 0–0.1%, stable in the range 0.1–1%).

When *An. coluzzii* was present, primary predictors of its abundance were nocturnal LST (b/w 3 and 6 weeks before the date of collection), cumulative rainfall (b/w 0 and 1 week before the date of collection), and % of surface occupied by grasslands (in the 1-km radius buffer zone). The abundance of *An. coluzzii* was constant for nocturnal LSTs under 22 °C and strongly increased above, until 23 °C. The association between abundance and cumulative rainfall was quite weak, but overall positive. The abundance increased linearly with the surface of grasslands in the range of 0–20%, and stabilized above that threshold. Secondary predictors of the abundance of *An. coluzzii* were as follows: % of surface occupied by marshlands in the 2-km radius buffer zone (positive linear association), % of surface occupied by riparian forests in the 500-m radius buffer zone (positive linear association), % of surface occupied by ligneous savannas in the 250-m

radius buffer zone (decrease in the range 0–40%, stable above), and distance to the closest stream (decrease in the range 0–100 m, stable above).

Notably, the confidence intervals of the partial dependence functions were overall high for all species and variables and, with a few exceptions, no variable emerged as much more predictive than others (in the VIPs) nor had signals outstandingly strong (in the PDPs).

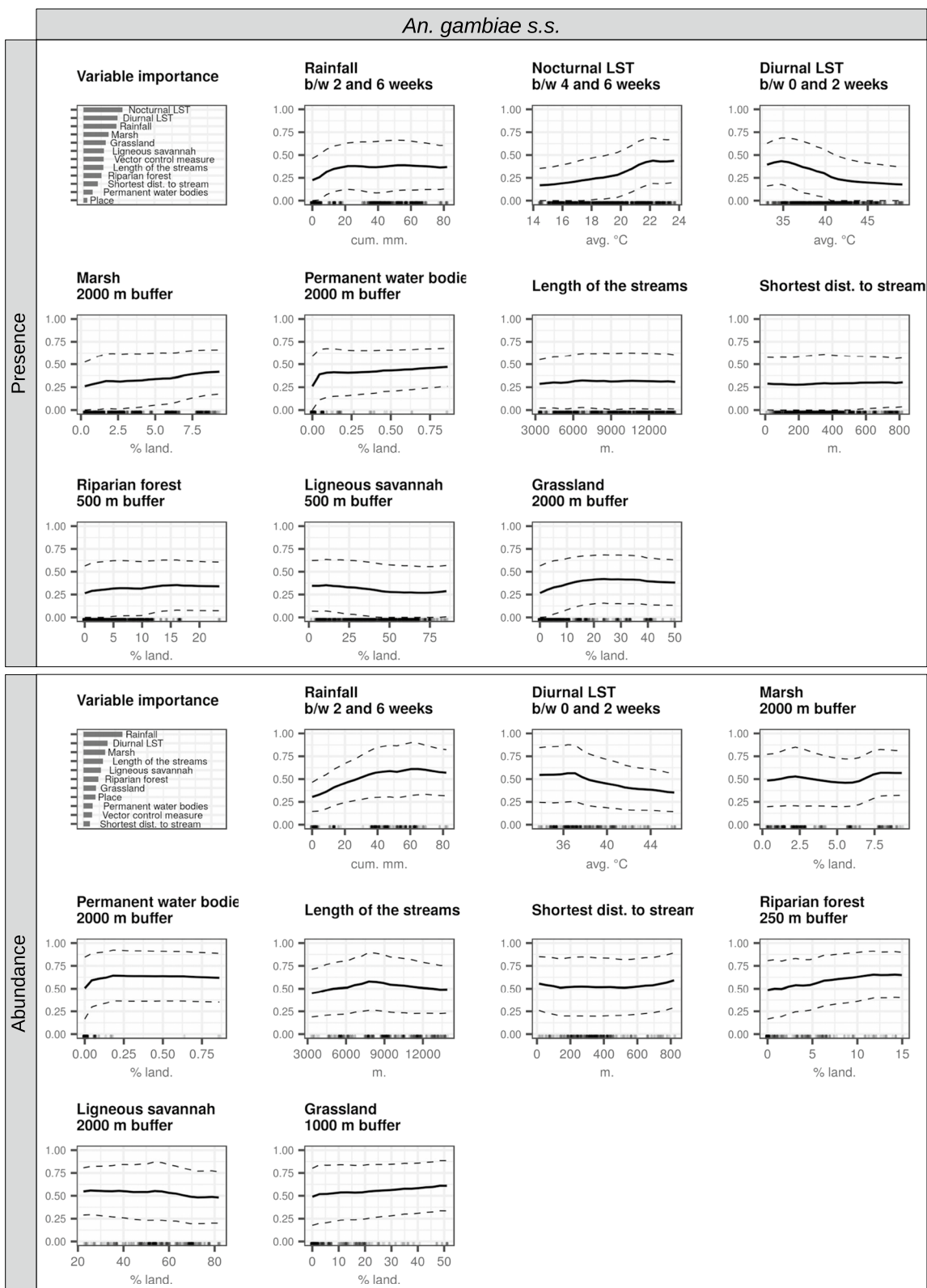
## Discussion

In this modeling study, we linked the biting rates of three major malaria vector species with environmental conditions at vicinities of places and periods of time of biting events to better understand their bio-ecology at fine spatiotemporal scales and identify important factors leading to increased biting risk. First, we correlated the biting rates of the vector species with (i) each meteorological variable at various time lags before the mosquito collection (using cross-correlation maps) and (ii) each landscape variable in various buffer zones around the HLC locations. Then, for selected time lags or spatial radii (the ones with the highest correlation coefficients), we generated multivariate models to study (i) the contribution of each environmental variable in predicting the biting rates and (ii) the nature of the relationship between each environmental variable and the biting rates (all other environmental conditions considered).

In this section, we first discuss the relationships between the biting rates of the malaria vectors and the meteorological and landscape conditions in the Diébou-gou area, and link them to the bio-ecology of the species. We then discuss how the results of our study could concretely support the conceptualization and deployment of locally tailored VC interventions. Next, we briefly summarize some of the advantages of the modeling method used for knowledge generation in the field of landscape entomology. We conclude the discussion with some limitations of this study and directions for future research.

### Effects of meteorological variables

The cross-correlation maps enable us to study how meteorological conditions affect the various stages of the mosquito life cycle [56]. Here, we found that weather



**Fig. 6** (See legend on previous page.)



(See figure on next page.)

**Fig. 7** Interpretation plots of the random forest models for *An. coluzzii*. Biting rates were separated into presence/absence of bites and abundance of bites (i.e. positive counts only), and two models were therefore generated [presence (top) and abundance (bottom)]. For each model, the top-left corner plot is the variable importance plot. The other plots are partial dependence plots (PDPs) for each variable included in the models (1 plot/variable). The y-axis in the PDPs represents: in the presence models, the probability of at least one individual biting a human during a night; in the abundance models, the log-transformed number of bites received by one human in one night conditional on their presence. The dashed lines represent the partial dependence function  $\pm$  one standard deviation (i.e. variability estimates). The range of values in the x-axis represents the range of values available in the data for the considered variable. The rugs above the x-axis represent the actual values available in the data for the variable. LST = land surface temperature, b/w = between

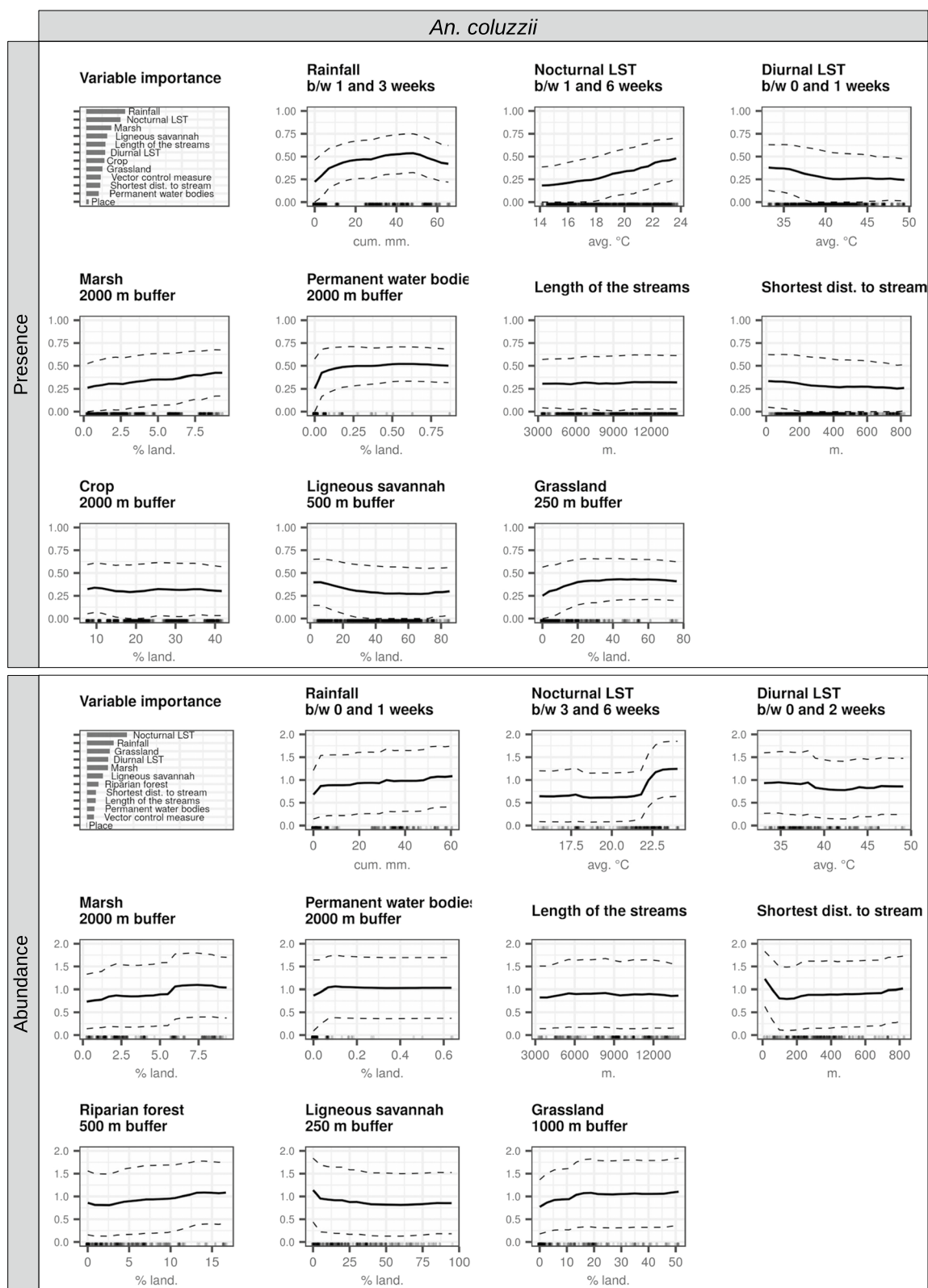
conditions (rainfall, nocturnal LST and diurnal LST) were significantly correlated with, and almost always primary predictors of, the presence and abundance of the species of the *Anopheles gambiae* complex. Stronger effects of these meteorological variables were found at various time lags in the studied range (from 0 to 6 weeks before collections). As discussed by Lebl and colleagues [84], weather-dependent life expectancy and development rates make it difficult to link time lags (of weather recordings) influencing mosquito abundance to different development stages. Given the mean life span and larval development duration of the *Anopheles* species collected in our area [49, 85, 86], weather during the first week (i.e. b/w 0 and 1 week) before collection was expected to influence the adult lifetime of collected mosquitoes, and weather during weeks 1–3 (i.e. b/w 1 and 2–3) before collection was expected to influence the larval lifetime of collected mosquitoes. Weather during preceding weeks (i.e. beyond 3 weeks before the date of collection) might affect observed densities by influencing the survival and development rates of (i) parent generations through mechanical effects on the population dynamic [84], (ii) the current/sampled generation through maternal/paternal effects [87, 88], or (iii) the current generation by preparing different biotic and abiotic conditions (for instance, by filling suitable larval development sites with water or by enabling the development of food sources, competitors or predators of *Anopheles* larvae).

In the spatiotemporal frame of our study, nocturnal LST ranged from 14 to 24 °C, and diurnal LST from 33 to 50 °C. Both nocturnal and diurnal LST were important predictors of the presence and abundance of *An. gambiae s.s.* and *An. coluzzii*, often with marked thresholds. Indeed, for *An. gambiae s.s.* we were able to identify a threshold of minimal LST over which the probability started to increase (20 °C), and for both species we also identified a threshold of maximum LST over which the probability reached a minimum (40 °C). For both species, diurnal temperature had the strongest effect during the 2 weeks preceding the dates of collection. This indicates that increasing diurnal temperatures probably reduced adult survival and larval development rates of the sampled generation of mosquitoes, leading to lower observed

abundance. Indeed, high temperatures are known to inhibit development of anopheline larvae [89] and to reduce adult survival [90]. Regarding nocturnal temperatures, the time period with the strongest effect on the presence and abundance of both *An. gambiae s.s.* and *An. coluzzii* was between 3 and 6 weeks before collection. This indicates that nocturnal (i.e. minimal) temperatures had their strongest impact by either affecting previous generations (low temperatures are known to reduce adult survival and inhibit larval development [89, 90]) or modifying habitats (with a delay) for the collected generation (for instance, low temperatures may inhibit the development of algae [91], whose biomass has been found to be associated with larval densities [92, 93]).

The high correlation coefficients between cumulative rainfall and both the presence and abundance of *An. gambiae s.s.* and *An. coluzzii*, and the fact that rainfall was systematically an important predictor of these species, might indicate that in our area *An. gambiae s.s.* and *An. coluzzii* are preferably attached to rainfall-dependent breeding sites, confirming the results of other studies [19, 94, 95] and explaining their seasonality. The time period with the strongest effects of rainfall on the presence and abundance of *An. gambiae s.s.* was between 2 and 6 weeks before collection, suggesting an effect on parental generations (as observed by Lebl and colleagues [84] for other mosquito species) or by modifying habitats (abiotic and/or biotic conditions) for the collected generation of these species. Conversely, rainfall had one of its highest correlation coefficients with the presence and abundance of *An. coluzzii* during weeks 1–3 before the dates of collection, indicating that rainfall might have had the greatest influence on the larval stages of the sampled generation of mosquitoes.

Different amounts of rainfall were needed for *An. coluzzii* and *An. gambiae s.s.* to be present or abundant, suggesting different breeding habitat preferences. Minimal rainfall was needed for *An. gambiae s.s.* to increase its probability of being present; additionally, rainfall was by far the most important predictor of its abundance (with a strong positive and approximately linear association). This could indicate that *An. gambiae s.s.* was more attached to breeding sites that quickly appear (presence)



**Fig. 7** (See legend on previous page.)

and abound (abundance) when limited rain falls and disappear after it stops, i.e. temporary breeding sites like puddles. *An. coluzzii* needed more rainfall to be present, suggesting preferences for breeding sites that require more water to be flooded, i.e. semipermanent surface water collections like marshlands or streams, which are usually filled in by rainfall throughout the rainy season and shortly after. Indeed, the % of surface occupied by marshlands was significantly correlated with, and the fourth most important predictor of, the abundance of *An. coluzzii*. Overall, these hypotheses about the preferred breeding habitats of *An. gambiae s.s.* and *An. coluzzii* confirm the literature reports [15–19].

### Effects of landscape variables

The biting rates of the three species were significantly correlated with several landscape variables, at varying distances from the collection points, and with fluctuating correlation coefficients. In addition, primary predictors of the presence and abundance of *An. funestus* were systematically landscape-based, and some were also primary predictors of the abundance of *An. gambiae s.s.* and *An. coluzzii*. Overall, this indicates that local landscape conditions are important drivers of the bio-ecology of the malaria vectors in rural areas, confirming the literature [5, 12, 17].

The mere presence of permanent water bodies (irrespective of the surface that they occupied) was sufficient to increase (even moderately) the probability of presence and the abundance of the three species. Permanent water bodies, where available, are likely to form breeding habitats for the *Anopheles* species [17, 44, 96–98], and explain why the few study villages located close to the dams and the main river are exposed to year-round bites of high densities of the three species [28]. The % of surface occupied by marshlands at the vicinities of the biting sites was the most important predictor of the presence and abundance models of *An. funestus*. In our study area, marshlands, a semipermanent aquatic environment, hence seemed to be one of the preferred breeding habitats of *An. funestus*, as it has been observed in other places [96, 97]. Notably, the correlation coefficients between the presence/abundance of bites and the % of surface occupied by breeding habitat land cover types (marshlands and permanent water bodies) increased as buffer sizes increased. This might indicate that mosquitoes are able to fly over quite large distances to reach their biting site from these breeding habitats ( $\geq 2$  km), as observed elsewhere in similar landscapes [99, 100]. Proximity to the streams ( $< 100$  m) and % of landscape occupied by the riparian forests ( $\leq 500$  m) were secondary predictors of the presence and/or abundance of *An. gambiae s.s.* and *An. coluzzii*. Streams and riparian forests (which are

spatially interrelated, i.e. streams flow under riparian forests) might hence form secondary, semipermanent breeding sites for the species of the *Anopheles gambiae* complex in the Diébougou area.

Grasslands—a very “open” landscape—and ligneous savannas—the most “closed” landscape in our study area—were alone or together significantly correlated with, and important predictors of, the presence and/or the abundance of the three malaria vectors studied. Increasing surfaces of grassland areas were associated with increasing probabilities of presence or abundance, while increasing surfaces of savannas were associated with decreasing probabilities of presence or abundance. With some rare exceptions, these landscape indicators were most highly correlated in small-radii ( $\leq 500$  m) buffer areas around the collection sites. Although grassland may provide suitable breeding sites for, at least, *An. gambiae s.s.* and *An. coluzzii* [101], these results seem to indicate that the degree of openness of the landscape some hectometers around villages had a great impact on malaria mosquito biting rates. Our observations are supported by the hypothesis of the lower dispersal of *Anopheles* mosquitoes in closed landscape (in comparison to open landscape) [102] leading to shorter gonotrophic cycle durations and therefore increased biting frequencies and higher biting rates [103]. A similar observation was previously made with *An. coluzzii* in Benin [17]. In the Diébougou area, ligneous savannas seemed to act as natural protective barriers against the malaria vectors and, conversely, grasslands as an aggravating biting risk factor. In a country which is increasingly replacing its closed landscapes (savannas) with opened ones [37], this observation is worrying for malaria transmission. Removal of savannas may significantly increase biting densities. This concern may however be mitigated for our study area, as savannas are usually mainly replaced by crops [37], which themselves did not seem to be an aggravating risk factor (i.e. crops did not emerge as an important variable in our models).

### Back to the field: how can these models and knowledge concretely support the fight against malaria transmission at the local scale?

An important question is how these results can ultimately help build locally tailored VC interventions (i.e. deploy the right VC tool at the right time and the right place) to support prevention and reduction of malaria transmission. The knowledge and models generated in this study could support at least three actions: (i) conceptualization of tailored vector control intervention plans, (ii) decisions regarding the places and times where recurrent (long-term) interventions should be deployed, in the form of seasonal maps of predicted biting rates, and



(iii) decisions regarding the places and times where occasional (short-term) interventions should be deployed, in the form of an early warning system.

#### **Support conceptualization of tailored vector control intervention plans**

The scientific knowledge confirmed, clarified, or gained through the interpretation of the models could help conceptualize tailored (i.e. species-, time- and place-specific) VC intervention plans. For example, management of temporary breeding sites (through e.g. larval control, or information education, and communication) during the rainy season is likely to be impactful, given the biting densities of *An. coluzzii* and *An. gambiae s.s.* in these seasons and their preferred breeding habitats. Similarly, larval source management in semi-temporary breeding sites (marshlands) would be an interesting option to reduce the presence and abundance of *An. funestus* during the dry-cold season, and if done, should cover quite large buffer zones (at least 2 km) around the households. Beyond these few examples, efficacious and cost-effective VC action plans could be designed on the basis of our characterization of the vectors' local bio-ecology. Importantly, however, several important traits of the vectors (e.g. physiological resistance, behavioral resistance) remain to be characterized in order to design highly efficient action plans.

#### **Support decisions regarding the places and times where recurrent interventions should be deployed through seasonal maps of predicted distribution of biting rates**

Once VC interventions have been conceptualized, one must choose the places and times they should be deployed. Here, the multivariate models could be used to generate maps of the predicted distribution of biting rates for each species over the whole study area, at fine spatial resolutions (village or household), and spanning the typical meteorological conditions in the area (for instance, three maps could be generated for each species: one for the dry-cold, one for the dry-hot, and one for the rainy season). These maps could help target and possibly prioritize the places and times for the deployment of recurrent, long-term VC interventions [4, 17, 21].

#### **Support decisions regarding the places and times where occasional interventions should be deployed through an early warning system**

A limitation of these maps is that they would only consider “typical,” i.e. average, meteorological conditions within a given season. However, different-than-expected events, such as rainfall episodes in the dry season or longer/shorter rainy season, could possibly lead to higher-than-expected biting rates and consequently

peaks of infectiousness and transmission of malaria. Our study has shown that meteorological conditions several weeks prior to the mosquito collections can accurately predict future biting rates. To help identify these potential “hot spots” of transmission in a timely manner, an early warning system (EWS) based on these predictive models could be built. Such an EWS, in the form of an automated algorithm, would routinely extract the up-to-date meteorological data and use the models to generate high-resolution maps of short-term forecasts (1 week ahead, 2 weeks ahead, etc.) of the biting rates. The potential hot spots of malaria transmission identified in the maps could then benefit from special interventions that remain to be conceptualized (e.g. increased vector control, special prevention or curation actions).

#### **Methodological bonus: on the use of algorithmic models and interpretable machine learning in landscape entomology**

In our study, we have shown how complex statistical models and IML can be used to enhance the fundamental knowledge and understanding of the complex links between the environment and the malaria vectors. Advantages of this modeling workflow over more traditional modeling methods (e.g. linear or logistic regressions) include the ability to (i) inherently capture and unveil complex patterns such as nonlinear or nonmonotonic relationships (e.g. effect of temperature and rainfall) and interactions and (ii) easily include more variables [20] and hence capture small—yet relevant—effects (e.g. effects from riparian forests or distance to streams). Necessary conditions to perform causal interpretation from “black-box” models are to (i) generate a good predictive model and (ii) have some prior domain knowledge about the causal structure of the system under study [27]. Both conditions were met in our work.

Using machine-learning black-box models for scientific discovery, i.e. to generate new knowledge from data, is an emerging trend in many disciplines [26, 104] that has been made possible by the recent development of both IML tools [78] and the know-how to interpret these complex models [25–27, 104, 105]. ML models enable us to integrate knowledge from existing theory in a less formal way than “data” models, and as such can be useful for theory development, provided that a careful linkage to existing knowledge is made [104, 106]. New theoretical insights generated from data and models may then in turn lead to unforeseen experimental research questions. We believe that the fields of landscape epidemiology and entomology still need to fully embrace the potential offered by these methods, in support not only of prediction or forecasting, but also explanation, i.e. to improve

our understanding of the complex processes leading to malaria transmission.

#### Limitations and directions for future research

An important limitation of our work is linked to the spatiotemporal sampling distribution of mosquito collection. First, no collection was conducted during the high rainy season (July to October) at the known mosquito abundance and malaria transmission peaks. Second, all the collection points were less than 800 m away from the theoretical hydrographic network (which is spatially interrelated with many breeding habitats such as marshlands, streams, riparian forests), meaning that our study could not identify potential differences in the drivers of vector abundance for households further than this distance. Year-round longitudinal collections, including sites further away from permanent or semipermanent breeding habitats, may enable a better understanding of the overall malaria mosquito spatiotemporal dynamics in the area. Meanwhile, these limitations must be accounted for if our models are used to generate predictive maps of the spatiotemporal distribution of vector abundance in the study area.

Similarly, predicting vector abundance outside the study area using the models generated in this study would be of high interest, to extend the operational tools previously mentioned (maps, EWS) to other areas than the Diébougou health district. However, careful attention should be given because of well-known problems linked to predicting beyond the model sampling locations or range of values (e.g. overfitting) [107, 108]. The scalability of our models to places with similar landscape and weather dynamics as the Diébougou area could be tested by collecting similar entomological data in another health district and comparing these ground-truth data with predictions generated by the models. In any case, these models should not be used to predict in remote ecoregions or urban settings, or at higher/lower spatial resolution.

Another limitation is the nature and diversity of the variables introduced in the models. Very fine-scale potential important drivers of mosquito abundance, such as the presence of alternative sources of blood meal (e.g. cattle), of domestic breeding sites, market gardening, or the micro-climatic conditions on the night of collection, have not been investigated. These variables were significant drivers elsewhere in West African rural settings [17, 44, 109]. Yet, the good predictive accuracy of the models suggest that, most probably, the most important drivers of vector abundance in our study area have been identified.

The absence of a strong signal from single variables in the models and the large confidence intervals in the PDPs suggest that the models might have learned important

interactions between variables. IML tools such as the *H*-statistic [110] might help reveal such interactions, and others like the two-variable PDP [78] might help explore their effect on malaria vectors abundance. Other tools can be used to analyze individual, or a target set of, predictions made by the models (these tools are called local interpretation methods, e.g. LIME [111] or Shapley values [112]). Local interpretation could be useful, for example, to precisely determine the environmental drivers of vector presence/abundance for a village of interest, or to better identify the drivers of the spatial heterogeneity of biting rates within a season of interest. Altogether, these IML tools might enable us to dig deeper into the models and, hence, the complexity of the ecological niche of malaria vectors.

This work has revealed how landscape can influence the biting rates of vectors, either directly (by impacting vector dispersal) or indirectly (by providing suitable breeding sites and hence increasing vector densities and consequently biting rates). Further investigations on the role played by the level of openness of the landscape are needed to confirm the various hypotheses that we have previously enumerated. A finer-grained land cover classification (e.g. discriminating shrub, tree and wooded savanna) could help test some of our hypotheses: for instance, is there a correlation between the gradient of closedness of the savannas and the abundance of vectors in our study area? Moreover, additional fieldwork could help identify the potential cause–effect relationship between surface of grasslands and malaria vector presence/abundance (i.e. breeding habitat or open landscape favoring dispersal).

Lastly, as stated previously, the scientific knowledge confirmed or acquired in this study and the good predictive accuracy of our models lay the ground for the development of operational tools to support vector control and improve forecasting of epidemic outbreaks at the local scale (e.g. locally tailored VC intervention plans, seasonal maps of the spatiotemporal distribution of vector abundance, EWS).

#### Conclusion

In this study, several aspects of the bio-ecology of the main malaria vectors in the Diébougou area were explored using field mosquito collections and high-resolution EO data (reflecting both meteorological and landscape local conditions) in a state-of-the-art statistical modeling framework. Overall, the spatiotemporal distributions of biting rates of *An. coluzzii* and *An. gambiae s.s.* were closely associated with meteorological conditions (temperature, precipitation), while those of *An. funestus* were more closely linked to landscape conditions. Meteorological conditions (temperatures, rainfall) putatively

affected all developmental stages of the mosquitoes (larval, adult) at varying levels according to the species and the meteorological variable. Weather occasionally had an even greater impact on time periods preceding the life span of the sampled generation. Primary and possible secondary breeding habitats of each vector species were proposed: *An. funestus*, *An. coluzzii* and *An. gambiae s.s.* seemed to be distributed along a gradient of persistence of the breeding sites, from permanent to temporary, confirming the literature reports. The rate of openness of the landscape seemed to play a major role in the biting rates, which could represent a major concern in a context of progressive shrinkage of the savanna and forest surfaces in Burkina Faso. This work lays the foundation for the development of operational tools to enhance and optimize the fight against malaria transmission at the local scale, such as vector control action plans, seasonal maps of predicted distribution of biting rates or early warning systems for the detection of malaria outbreaks.

#### Abbreviations

IML: Interpretable machine learning; ML: Machine learning; HLC: Human landing catch; VC: Vector control; RF: Random forest; cc: Correlation coefficient; EO: Earth observation; GEOBIA: Geographic object-based image analysis; SPOT: Satellite Pour l'Observation de la Terre; SRTM: Shuttle Radar Topography Mission; DEM: Digital elevation model; LST: Land surface temperature; CCM: Cross-correlation map; PCR: Polymerase chain reaction; LVO-CV: Leave-village-out cross-validation; PR-AUC: Precision–recall area under the curve; MAE: Mean absolute error; VIP(s): Variable importance plot(s); PDP(s): Partial dependence plot(s); b/w: Between; SD: Standard deviation; LIME: Local interpretable model-agnostic explanations; EWS: Early warning system.

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13071-021-04851-x>.

**Additional file 1: Figure S1.** Summary of the meteorological conditions around the sampling points. Average meteorological conditions in a 2 km radius buffer zone around the collection points (weekly aggregation). Vertical red lines indicate the dates of the entomological surveys. Ribbons indicate the mean  $\pm$  one standard deviation considering all the sampling points for the week.

**Additional file 2: Figure S2.** Summary of the landscape conditions around the sampling points. Average percentage of surface occupied by each land cover class in the various buffer zones (250 m, 500 m, 1 km, 2 km radii) around the collection points. Error bars indicate the mean  $\pm$  one standard deviation considering all the sampling points.

**Additional file 3: Figure S3.** Pictures representative of the main land cover classes in the Diébougou area. Pictures were taken in November 2018.

**Additional file 4: Figure S4.** Model evaluation plots for the presence models. A1, A2, A3 are precision–recall curves for the presence models of respectively *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*. Precision–recall curves show the precision and the recall of the models for different probability thresholds of the “presence” class. Precision is the proportion of presence identifications that was actually correct, while recall is the proportion of actual presence observations that were identified correctly. The horizontal dashed line represents the baseline (i.e. random or no-skill classifier). A precision–recall curve above the horizontal line indicates a better-than-no-skill classifier. The higher the area between the precision–recall

curve and the horizontal line, the better the classifier. Plots B1, B2, B3 are observed vs. predicted presence probabilities for each out-of-sample village. The y-axis represents the sum over the 8 sampling points/village/survey (4 points by village \* 2 places (interior and exterior)). Overall, the plots A1, A2, A3 show that the models had good predictive accuracies (precision–recall curves are higher than the baseline curve, particularly for *An. funestus* and *An. coluzzii*). The plots B1, B2, B3 show that the models predicted well the spatiotemporal trends of presence/absence of bites (lines of predicted presence probabilities are generally close to lines of observed probabilities), although they usually slightly overestimated the probabilities of being bitten (predicted presence probability > observed presence probability).

**Additional file 5: Figure S5.** Model evaluation plots for the abundance models. A1, A2, A3 are violin plots of the distribution of the residuals for the abundance models of respectively *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*, by observed counts of bites (4 classes: 1 bite, 2–3 bites, 4–10 bites, > 10 bites). Black dots indicate the median value. B1, B2, B3 are observed vs. predicted number of bites/village/entomological surveys. The y-axis represents the sum of bites over the 8 sampling points/village/survey (4 points by village \* 2 places (interior and exterior)) on a logarithmic scale. The absence of a dot indicates that no vector was collected. MAE = mean absolute error;  $n$  = number of observations. Overall, the plots A1, A2, A3 show that the models predicted well small observed counts of bites (1 bite, 2–3 bites) (cf. small MAEs, small residuals), which represent the vast majority of observations (high  $n$ ). Larger counts (4–10 bites, > 10 bites) tended to be underestimated by the models, especially for *An. funestus* and *An. gambiae s.s.* However, large counts (> 10 bites) represented few observations (small  $n$ ). The plots B1, B2, B3 confirm these observations, and additionally show that general trends of biting rates over time were well predicted by the models (lines of predicted abundance are generally close to lines of observed abundance).

**Additional file 6: Figure S6.** Feature selection for the multivariate models. The figure shows the Spearman correlation coefficient between the explanatory variables and each response variable (presence and abundance of *An. funestus*, *An. gambiae s.s.* and *An. coluzzii*). Based on these results, variables were retained for the multivariate models according to the following criteria: we first excluded variables that were poorly correlated with the response variable (i.e. correlation coefficients less than 0.1 or  $p$ -values greater than 0.2 at all time associations or buffer radii considered), except for variables related to the presence of water—i.e. possible breeding sites—that were all retained whatever their correlation. Then, for each meteorological (resp. landscape) variable, we retained the time lag interval (resp. buffer radius) showing the higher absolute correlation coefficient value. We finally excluded collinear variables (i.e. Pearson correlation coefficient between the variables > 0.7) based on empirical knowledge.

#### Acknowledgements

We thank populations of the villages for their kind support and collaboration. We also thank all the field and laboratory staff for their strong commitment to the REACT project. We thank all the anonymous persons that make data-science web forums alive by asking and answering technical questions; these forums were extensively used to elaborate this modeling work.

Map data copyrighted by OpenStreetMap contributors and available from <https://www.openstreetmap.org>. This work contains modified Copernicus Sentinel data (2018).

#### Authors' contributions

PT, CP and NM conceived and designed the study. DDS and PT collected and prepared the data. NM, KM and RKD supervised fieldwork. PT analyzed the data, helped by AP and MM. PT and NM drafted the manuscript. KM, CP, DDS, AAK, RKD, MM and AP reviewed the manuscript. All authors read and approved the final manuscript.

#### Funding

This work was part of the REACT project, funded by the French Initiative 5%—Expertise France (no. 15SANIN213). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the



manuscript. PT was supported by the French Institute of Research for Sustainable Development (IRD) through an international volunteer fellowship and the French National Research Agency (ANR) through the ANORHYTHM project (ANR-16-CE35-008). This work was supported by public funds received in the framework of GEOSUD, a project (ANR-10-EQPX-20) of the program "Investissements d'Avenir" managed by the French National Research Agency.

#### Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

#### Declarations

##### Ethics approval and consent to participate

The protocol of this study was reviewed and approved by the Institutional Ethics Committee of the Institut de Recherche en Sciences de la Santé (IEC-IRSS) and registered as N°A06/2016/CEIRES. Mosquito collectors and supervisors gave their written informed consent. They received a vaccine against yellow fever as a prophylactic measure. Collectors were treated free of charge for malaria according to WHO recommendations.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>MIVEGEC, Université de Montpellier, CNRS, IRD, Montpellier, France. <sup>2</sup>Institut de Recherche en Sciences de La Santé (IRSS), Bobo-Dioulasso, Burkina Faso. <sup>3</sup>Université Nazi Boni, Bobo-Dioulasso, Burkina Faso. <sup>4</sup>Institut Pierre Richet (IPR), Bouaké, Côte d'Ivoire. <sup>5</sup>ESPACE-DEV, Université Montpellier, IRD, Université Antilles, Université Guyane, Université Réunion, Montpellier, France.

Received: 14 April 2021 Accepted: 12 June 2021

Published online: 29 June 2021

#### References

- WHO. World malaria report 2020: 20 years of global progress and challenges. Licence: CC BY-NC-SA 3.0 IGO; 2020.
- Bhatt S, Weiss DJ, Cameron E, Bisanzio D, Mappin B, Dalrymple U, et al. The effect of malaria control on *Plasmodium falciparum* in Africa between 2000 and 2015. *Nature*. 2015;526(7572):207–11.
- Wilson AL, Courtenay O, Kelly-Hope LA, Scott TW, Takken W, Torr SJ, et al. The importance of vector control for the control and elimination of vector-borne diseases. *PLoS Negl Trop Dis*. 2020;14(1):e0007831.
- WHO. Global vector control response 2017–2030. Geneva: World Health Organization; 2017. Licence: CC BY-NC-SA 3.0 IGO; 2017.
- Stresman GH. Beyond temperature and precipitation: Ecological risk factors that modify malaria transmission. *Acta Trop*. 2010;116(3):167–72.
- Ferguson HM, Dornhaus A, Beeche A, Borgemeister C, Gottlieb M, Mulla MS, et al. Ecology: a prerequisite for malaria elimination and eradication. *PLoS Med*. 2010;7(8):e1000303.
- Beck-Johnson LM, Nelson WA, Paaijmans KP, Read AF, Thomas MB, Bjørnstad ON. The effect of temperature on *Anopheles* mosquito population dynamics and the potential for malaria transmission. *PLoS ONE*. 2013;8(11):e79276.
- Mordecai EA, Paaijmans KP, Johnson LR, Balzer C, Ben-Horin T, de Moor E, et al. Optimal temperature for malaria transmission is dramatically lower than previously predicted. *Ecol Lett*. 2013;16(1):22–30.
- Shapiro LLM, Whitehead SA, Thomas MB. Quantifying the effects of temperature on mosquito and parasite traits that determine the transmission potential of human malaria. *PLoS Biol*. 2017;15(10):e2003489.
- Shaman J, Stieglitz M, Stark C, Le Blancq S, Cane M. Using a dynamic hydrology model to predict mosquito abundances in flood and swamp water. *Emerg Infect Dis*. 2002;8(1):6–13.
- Paaijmans KP, Takken W, Githeko AK, Jacobs AFG. The effect of water turbidity on the near-surface water temperature of larval habitats of the malaria mosquito *Anopheles gambiae*. *Int J Biometeorol*. 2008;52(8):747–53.
- Fornace KM, Diaz AV, Lines J, Drakeley CJ. Achieving global malaria eradication in changing landscapes. *Malar J*. 2021;20(1):69.
- Hamon Jacques, Mouchet Jean. (1961). Les vecteurs secondaires du paludisme humain en Afrique. In : Etudes sur le paludisme en Afrique. *Médecine Tropicale*, 21 (No spécial), p. 643–60. ISSN 0025-682X
- Delmont J. Paludisme et variations climatiques saisonnières en savane soudanienne d'Afrique de l'Ouest. *Cah DÉtudes Afr*. 1982;22(85):117–33.
- Simard F, Ayala D, Kamdem G, Pombi M, Etouna J, Ose K, et al. Ecological niche partitioning between *Anopheles gambiae* molecular forms in Cameroon: the ecological side of speciation. *BMC Ecol*. 2009;9(1):17.
- Costantini C, Ayala D, Guelbeogo WM, Pombi M, Some CY, Bassole IH, et al. Living at the edge: biogeographic patterns of habitat segregation conform to speciation by niche expansion in *Anopheles gambiae*. *BMC Ecol*. 2009;9(1):16.
- Moiroux N, Djènontin A, Bio-Bangana AS, Chandre F, Corbel V, Guis H. Spatio-temporal analysis of abundances of three malaria vector species in southern Benin using zero-truncated models. *Parasit Vectors*. 2014;7(1):103.
- Diabaté A, Dabiré RK, Heidenberger K, Crawford J, Lamp WO, Culler LE, et al. Evidence for divergent selection between the molecular forms of *Anopheles gambiae*: role of predation. *BMC Evol Biol*. 2008;8(1):5.
- Diabaté A, Dabiré RK, Kim EH, Dalton R, Millogo N, Baldet T, et al. Larval development of the molecular forms of *Anopheles gambiae* (Diptera: Culicidae) in different habitats: a transplantation experiment. *J Med Entomol*. 2005;42(4):548–53.
- Breiman L. Statistical modeling: the two cultures. *Stat Sci*. 2001;16(3):199–215.
- Machault V, Vignolles C, Borchi F, Vounatsou P, Pages F, Briolant S, et al. The use of remotely sensed environmental data in the study of malaria. *Geospat Health*. 2011;5:151–68.
- Ebhuma O, Gebreslasie M. Remote sensing-driven climatic/environmental variables for modelling malaria transmission in sub-Saharan Africa. *Int J Environ Res Public Health*. 2016;13(6):584.
- Parselia E, Kontoes C, Tsouni A, Hadjichristodoulou C, Kioutsoukis I, Magiorkinis G, et al. Satellite earth observation data in epidemiological modeling of malaria, dengue and west Nile virus: a scoping review. *Remote Sens*. 2019;11(16):1862.
- Fayyad U, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery in Databases. *AIMag*. 1996;17(3):37.
- Molnar, Christoph. "Interpretable machine learning. A Guide for Making Black Box Models Explainable", 2019. <https://christophm.github.io/interpretable-ml-book/>.
- Murdoch WJ, Singh C, Kumbier K, Abbasi-Asl R, Yu B. Definitions, methods, and applications in interpretable machine learning. *Proc Natl Acad Sci*. 2019;116(44):22071–80.
- Zhao Q, Hastie T. Causal interpretations of black-box models. *J Bus Econ Stat*. 2021;39(1):272–81.
- Soma DD, Zogo BM, Somé A, Tchikoi BN, de Hien DFS, Pooda HS, et al. *Anopheles* bionomics, insecticide resistance and malaria transmission in southwest Burkina Faso: a pre-intervention study. *PLoS ONE*. 2020;15(8):e0236920.
- Gillies MT, De Meillon. The Anophelinae of Africa south of the Sahara (Ethiopian Zoogeographical Region). Publications of the South African Institute for Medical Research. 1968;54.
- Gillies MT, Coetzee M. A supplement to the Anophelinae of Africa South of the Sahara. *Publ Afr Inst Med Res*. 1987;55:1–143.
- Koekemoer LL, Kamau L, Hunt RH, Coetzee M. A cocktail polymerase chain reaction assay to identify members of the *Anopheles funestus* (Diptera: Culicidae) group. *Am J Trop Med Hyg*. 2002;66(6):804–11.
- Cohuet A, Simard F, Berthomieu A, Raymond M, Fontenille D, Weill M. Isolation and characterization of microsatellite DNA markers in the malaria vector *Anopheles funestus*. *Mol Ecol Notes*. 2002;2(4):498–500.
- Santolamazza F, Mancini E, Simard F, Qi Y, Tu Z, della Torre A. Insertion polymorphisms of SINE200 retrotransposons within speciation islands of *Anopheles gambiae* molecular forms. *Malar J*. 2008;7(1):1–10.
- Hay GJ, Castilla G. Geographic Object-Based Image Analysis (GEOBIA): A new name for a new discipline. In: Blaschke T, Lang S, Hay GJ, editors. Object-based image analysis. Lecture notes in geoinformation

- and cartography. Berlin: Springer; 2008. [https://doi.org/10.1007/978-3-540-77058-9\\_4](https://doi.org/10.1007/978-3-540-77058-9_4).
35. NASA JPL. NASA shuttle radar topography mission global 1 arc second. NASA EOSDIS land processes DAAC; 2013. <https://lpdaac.usgs.gov/products/srtmgl1v003/>. Accessed 12 Apr 2021
  36. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
  37. CILSS, 2016. Landscapes of West Africa—A window on a changing world: Ouagadougou, Burkina Faso, CILSS, 219 p. (Comité Permanent Inter-états de Lutte contre la Sécheresse dans le Sahel)
  38. Cohen J. A coefficient of agreement for nominal scales. *Educ Psychol Meas*. 1960;20(1):37–46.
  39. Takken W, Charlwood JD, Billingsley PF, Gort G. Dispersal and survival of *Anopheles funestus* and *A. gambiae* s.l. (Diptera: Culicidae) during the rainy season in southeast Tanzania. *Bull Entomol Res*. 1998;88(5):561–6.
  40. Service MW. Mosquito (Diptera: Culicidae) dispersal—the long and short of it. *J Med Entomol*. 1997;34(6):579–88.
  41. Clarke SE, Bøgh C, Brown RC, Walraven GEL, Thomas CJ, Lindsay SW. Risk of malaria attacks in Gambian children is greater away from malaria vector breeding sites. *Trans R Soc Trop Med Hyg*. 2002;96(5):499–506.
  42. Jensen SK, Domingue JO. Extracting topographic structure from digital elevation data for geographic information-system analysis. *Photogramm Eng Remote Sens*. 1988;54(11):1593–600.
  43. Clark PJ, Evans FC. Distance to nearest neighbor as a measure of spatial relationships in populations. *Ecology*. 1954;35(4):445–53.
  44. Moiroux N, Bio-Bangana AS, Djènontin A, Chandre F, Corbel V, Guis H. Modelling the risk of being bitten by malaria vectors in a vector control area in southern Benin, west Africa. *Parasit Vectors*. 2013;6(1):71.
  45. Debebe Y, Hill SR, Tekie H, Dugassa S, Hopkins RJ, Ignell R. Malaria hot-spots explained from the perspective of ecological theory underlying insect foraging. *Sci Rep*. 2020;10(1):21449.
  46. NASA Goddard Earth Sciences Data And Information Services Center. GPM IMERG final precipitation L3 1 day 0.1 degree x 0.1 degree V06. NASA Goddard Earth Sciences Data and Information Services Center; 2019. [https://disc.gsfc.nasa.gov/datacollection/GPM\\_3IMERGDF\\_06.html](https://disc.gsfc.nasa.gov/datacollection/GPM_3IMERGDF_06.html). Accessed 11 Feb 2021.
  47. Wan, Zhengming, Hook, Simon, Hulley, Glynn. MOD11A1 MODIS/terra land surface temperature/emissivity daily L3 global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC; 2015. <https://lpdaac.usgs.gov/products/mod11a1v006/>. Accessed 11 Feb 2021.
  48. Wan, Zhengming, Hook, Simon, Hulley, Glynn. MYD11A1 MODIS/aqua land surface temperature/emissivity daily L3 global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC; 2015. <https://lpdaac.usgs.gov/products/myd11a1v006/>. Accessed 11 Feb 2021.
  49. Holstein M. Biologie d'*Anopheles gambiae* : recherches en Afrique-Occidentale Française. Genève: OMS; 1952. (Monographies - OMS). <http://www.documentation.ird.fr/hor/fdi:42581>. Accessed 2 Dec 2020.
  50. Lee S-K, Jin S. Decision tree approaches for zero-inflated count data. *J Appl Stat*. 2006;33(8):853–65.
  51. Mathlouthi W, Larocque D, Fredette M. Random forests for homogeneous and non-homogeneous Poisson processes with excess zeros. *Stat Methods Med Res*. 2020;29(8):2217–37.
  52. Bousari O, Moiroux N, Iwaz J, Djènontin A, Bio-Bangana S, Corbel V, et al. Use of a mixture statistical model in studying malaria vectors density. *PLoS ONE*. 2012;7(11):e50452.
  53. Cragg JG. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica*. 1971;39(5):829.
  54. Gadiaga L, Machault V, Pagès F, Gaye A, Jarjaval F, Godefroy L, et al. Conditions of malaria transmission in Dakar from 2007 to 2010. *Malar J*. 2011;10(1):312.
  55. Makowski D, Ben-Shachar MS, Patil I, Lüdtke D. Methods for correlation analysis. CRAN; 2020. <https://github.com/easystats/correlation>.
  56. Curriero FC, Shone SM, Glass GE. Cross correlation maps: a tool for visualizing and modeling time lagged associations. *Vector Borne Zoonotic Dis Larchmt N*. 2005;5(3):267–75.
  57. Tyagi S, Mittal S. Sampling approaches for imbalanced data classification problem in machine learning. In: Singh P, Kar A, Singh Y, Kolekar M, Tanwar S, editors. Proceedings of ICRIC 2019. Lecture notes in electrical engineering, vol. 597. Cham: Springer; 2020. [https://doi.org/10.1007/978-3-030-29407-6\\_17](https://doi.org/10.1007/978-3-030-29407-6_17).
  58. Ten CD. quick tips for machine learning in computational biology. *BioData Mining*. 2017;10(1):35.
  59. Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environ Model Softw*. 2018;101:1–9.
  60. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*. 2015;10(3):e0118432.
  61. Friedman JH. *Machine. Ann Stat*. 2001;29(5):1189–232.
  62. R Core Team. R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing; 2018. <https://www.R-project.org/>. Accessed 15 Oct 2020.
  63. RStudio Team. RStudio: integrated development environment for R. Boston: RStudio, PBC; 2020. <http://www.rstudio.com/>. Accessed 15 Oct 2020.
  64. opendapr. Fast download of many earth observation data in R using the OPeNDAP Capacities. <https://github.com/ptaconet/opendapr>. Accessed 01 Apr 2021
  65. Brenning A, Bangs D, Becker M. RSAGA: SAGA geoprocessing and terrain analysis; 2018. <https://CRAN.R-project.org/package=RSAGA>. Accessed 15 Oct 2020.
  66. Bivand R. rgrass7: interface between GRASS 7 geographical information system and R; 2018. <https://CRAN.R-project.org/package=rgrass7>. Accessed 15 Oct 2020.
  67. Hijmans RJ. raster: geographic data analysis and modeling; 2020. <https://CRAN.R-project.org/package=raster>. Accessed 15 Oct 2020.
  68. Pebesma E. Simple features for R: standardized support for spatial vector data. *R J*. 2018;10(1):439–46.
  69. Bivand R, Keitt T, Rowlingson B. rgdal: bindings for the “Geospatial” data abstraction library; 2019. <https://CRAN.R-project.org/package=rgdal>. Accessed 15 Oct 2020.
  70. Liaw A, Wiener M. Classification and regression by randomForest. *R News*. 2002;2(3):18–22.
  71. Baddeley A, Rubak E, Turner R. Spatial point patterns: methodology and applications with R. London: Chapman and Hall/CRC Press; 2015. <http://www.crcpress.com/Spatial-Point-Patterns-Methodology-and-Applications-with-R/Baddeley-Rubak-Turner/9781482210200/>. Accessed 15 Oct 2020.
  72. QGIS Development Team. QGIS geographic information system. QGIS Association; 2021. <https://www.qgis.org>. Accessed 15 Oct 2020.
  73. Hesselbarth MHK, Sciaini M, With KA, Wiegand K, Nowosad J. land-scapemetrics: an open-source R tool to calculate landscape metrics. *Ecography*. 2019;42:1–10.
  74. Wing MKC from J, Weston S, Williams A, Keefer C, Engelhardt A, Cooper T, et al. caret: classification and regression training; 2018. <https://CRAN.R-project.org/package=caret>. Accessed 15 Oct 2020.
  75. Wright MN, Ziegler A. ranger: a fast implementation of random forests for high dimensional data in C++ and R. *J Stat Softw*. 2017;77(1):1–17.
  76. Meyer H. CAST: “caret” applications for spatial-temporal models; 2020. <https://CRAN.R-project.org/package=CAST>. Accessed 15 Oct 2020.
  77. Yan Y. MLmetrics: machine learning evaluation metrics; 2016. <https://CRAN.R-project.org/package=MLmetrics>. Accessed 15 Oct 2020.
  78. Molnar C, Bischl B, Casalicchio G. iml: an R package for interpretable machine learning. *JOSS*. 2018;3(26):786.
  79. Greenwell BM. pdp: an R package for constructing partial dependence plots. *R J*. 2017;9(1):421–36.
  80. Pedersen TL. patchwork: the composer of plots; 2019. <https://CRAN.R-project.org/package=patchwork>. Accessed 15 Oct 2020.
  81. Kahle D, Wickham H. ggmap: spatial visualization with ggplot2. *R J*. 2013;5(1):144–61.
  82. Saito T, Rehmsmeier M. Precrec: fast and accurate precision-recall and ROC curve calculations in R. *Bioinformatics*. 2017;33(1):145–7.
  83. Wickham H. tidyverse: Easily Install and Load the “Tidyverse”; 2017. <https://CRAN.R-project.org/package=tidyverse>. Accessed 15 Oct 2020.
  84. Lebl K, Brugger K, Rubel F. Predicting *Culex pipiens/restuans* population dynamics by interval lagged weather data. *Parasit Vectors*. 2013;6(1):129.
  85. Townson H. The biology of mosquitoes. Volume 1. Development, nutrition and reproduction. By A.N. Clements. (London: Chapman

- & Hall, 1992).viii 509 pp. ISBN 0-412-40180-0. Bull Entomol Res. 1993;83(2):307–308.
86. Carnevale P, Robert V, Manguin S, Corbel V, Fontenille D, Garros C, et al. Les anophèles : biologie, transmission du Plasmodium et lutte antivectorielle. IRD; 2009. 391 p. (Didactiques). <http://www.documentation.ird.fr/hor/fdi:010047862>
  87. Zirbel K, Eastmond B, Alto BW. Parental and offspring larval diets interact to influence life-history traits and infection with dengue virus in *Aedes aegypti*. R Soc Open Sci. 2018;5(7):180539.
  88. Zirbel KE, Alto BW. Maternal and paternal nutrition in a mosquito influences offspring life histories but not infection with an arbovirus. Ecosphere. 2018;9(10):e02469.
  89. Lyons CL, Coetzee M, Chown SL. Stable and fluctuating temperature effects on the development rate and survival of two malaria vectors, *Anopheles arabiensis* and *Anopheles funestus*. Parasit Vectors. 2013;6(1):104.
  90. Lyons CL, Coetzee M, Terblanche JS, Chown SL. Desiccation tolerance as a function of age, sex, humidity and temperature in adults of the African malaria vectors *Anopheles arabiensis* and *Anopheles funestus*. J Exp Biol. 2014;217(21):3823–33.
  91. Raven JA, Geider RJ. Temperature and algal growth. New Phytol. 1988;110(4):441–61.
  92. Kweka EJ, Zhou G, Munga S, Lee M-C, Atili HE, Nyindo M, et al. Anopheline larval habitats seasonality and species distribution: a prerequisite for effective targeted larval habitats control programmes. PLoS ONE. 2012;7(12):e52084.
  93. Kaufman MG, Wanja E, Maknoja S, Bayoh MN, Vulule JM, Walker ED. Importance of algal biomass to growth and development of *Anopheles gambiae* larvae. J Med Entomol. 2006;43(4):669–76.
  94. Gimonneau G, Pombi M, Choisy M, Morand S, Dabiré RK, Simard F. Larval habitat segregation between the molecular forms of the mosquito *Anopheles gambiae* in a rice field area of Burkina Faso, West Africa. Med Vet Entomol. 2012;26(1):9–17.
  95. Gimnig JE, Ombok M, Kamau L, Hawley WA. Characteristics of larval anopheline (*Diptera: Culicidae*) habitats in Western Kenya. J Med Entomol. 2001;38(2):282–8.
  96. Pages F, Orlandipradines E, Corbel V. Vecteurs du paludisme: biologie, diversité, contrôle et protection individuelle. Médecine Mal Infect. 2007;37(3):153–61.
  97. Sinka ME, Bangs MJ, Manguin S, Coetzee M, Mbogo CM, Hemingway J, et al. The dominant *Anopheles* vectors of human malaria in Africa, Europe and the Middle East: occurrence data, distribution maps and bionomic précis. Parasit Vectors. 2010;3(1):117.
  98. Zogo B, Koffi AA, Alou LPA, Fournet F, Dahounto A, Dabiré RK, et al. Identification and characterization of *Anopheles* spp. breeding habitats in the Korhogo area in northern Côte d'Ivoire: a study prior to a Bti-based larviciding intervention. Parasit Vectors. 2019;12(1):146.
  99. Verdonschot PFM, Besse-Lototskaya AA. Flight distance of mosquitoes (*Culicidae*): A metadata analysis to support the management of barrier zones around rewetted and newly constructed wetlands. Limnologia. 2014;45:69–79.
  100. Thomas CJ, Cross DE, Bøgh C. Landscape movements of *Anopheles gambiae* Malaria vector mosquitoes in rural Gambia. PLoS ONE. 2013;8(7):e68679.
  101. Fillinger U, Majambere S, Lindsay SW, Green C, Sayer DR. Spatial distribution of mosquito larvae and the potential for targeted larval control in the Gambia. Am J Trop Med Hyg. 2008;79(1):19–27.
  102. Le Goff G, Carneval P, Robert V. Low dispersion of anopheline malaria vectors in the African equatorial forest. Parasite. 1997;4(2):187–9.
  103. Afrane YA, Lawson BW, Githeko AK, Yan G. Effects of microclimatic changes caused by land use and land cover on duration of gonotrophic cycles of *Anopheles gambiae* (*Diptera: Culicidae*) in western Kenya highlands. J Med Entomol. 2005;42(6):974–80.
  104. Karpatne A, Atluri G, Faghmous JH, Steinbach M, Banerjee A, Ganguly A, et al. Theory-guided data science: a new paradigm for scientific discovery from data. IEEE Trans Knowl Data Eng. 2017;29(10):2318–31.
  105. Shmueli G. To explain or to predict? Stat Sci. 2010;25(3):289–310.
  106. Shmueli G, Koppius O. Predictive Analytics in Information Systems Research. SSRN Electron J. 2010. <http://www.ssrn.com/abstract=1606674>. Accessed 18 Dec 2020.
  107. Wardrop NA, Geary M, Osborne PE, Atkinson PM. Interpreting predictive maps of disease: highlighting the pitfalls of distribution models in epidemiology. Geospat Health. 2014;9:237–46.
  108. Meyer H, Pebesma E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. Methods in ecology and evolution; 2021. <https://doi.org/10.1111/2041-210X.13650>. Accessed 11 Jun 2021.
  109. Ngowo HS, Kaindoa EW, Matthiopoulos J, Ferguson HM, Okumu FO. Variations in household microclimate affect outdoor-biting behaviour of malaria vectors. Wellcome Open Res. 2017;2:102.
  110. Friedman JH, Popescu BE. Predictive learning via rule ensembles. Ann Appl Stat. 2008;2(3):916–54.
  111. Ribeiro MT, Singh S, Guestrin C. “Why Should I Trust You?”: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. San Francisco California USA: ACM; 2016. p. 1135–44. <https://doi.org/10.1145/2939672.2939778>. Accessed 12 Apr 2021.
  112. Strumbelj KI. Explaining prediction models and individual predictions with feature contributions. Knowl Inf Syst. 2013;41:647–65.

## Publisher's Note

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

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

