

Forecasting invasive mosquito abundance in the Basque Country, Spain using machine learning techniques

Vanessa Steindorf^{1*}, Hamna Mariyam K. B.¹,
Nico Stollenwerk¹, Aitor Cevidane³, Jesús F. Barandika³,
Patricia Vazquez³, Ana L. García-Pérez³, Maíra Aguiar^{1,2}

¹Basque Center for Applied Mathematics, Bilbao, Spain

²Ikerbasque, Basque Foundation for Science, Bilbao, Spain

³Animal Health Department, NEIKER-Basque Institute for Agricultural Research and Development, Basque Research and Technology Alliance (BRTA), Derio, Spain

January 2, 2025

Abstract

Mosquito-borne diseases cause millions of deaths each year and are increasingly spreading from tropical and subtropical regions into temperate zones, creating significant public health risks. The establishment of mosquito species in new areas increases the risk of local transmission (autochthonous cases), driven by both rising mosquito populations and viremic imported cases, infected travelers who can spark local transmission. Such developments present new challenges for public health systems in non-endemic regions.

In Spain, in the Basque Country region, the spread of mosquitoes, driven by changing climatic conditions, has enhanced mosquito adaptation alongside an increase in imported cases of dengue, Zika, and chikungunya. By employing a model that captures the complexities of the mosquito life cycle driven by the interaction with weather variables, including temperature, precipitation, and humidity, and leveraging machine learning techniques, this study aims to predict *Aedes* invasive mosquito abundance in provinces of the Basque Country, using egg count as a proxy and the weather features as key independent variables.

Statistical analyses explored the impact of temperature, precipitation, and humidity on mosquito egg abundance. Using lagged climate variables and ovitrap egg counts, models were evaluated using Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics. The Random Forest (RF) model demonstrated the highest accuracy, followed by the Seasonal Autoregressive Integrated Moving Average (SARIMAX) model. Lastly, the best models were implemented to forecast *Aedes* invasive mosquito abundance in the Basque Country provinces. This forecasting tool aids vector control strategies in regions with expanding mosquito populations, highlighting the need for ongoing entomological surveillance to improve mosquito spread assessments.

Keywords: mosquito eggs; dengue; *Aedes albopictus*; machine learning; vector-borne diseases; entomological surveillance.

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

*Corresponding Author. Email: vsteindorf@bcamath.org. ORCID: 0000-0002-0707-9511

26 1 Introduction

27 Vector-borne diseases, particularly those transmitted by mosquitoes, have become a significant
28 global concern. The expansion of mosquitoes and the increase in transmitted diseases are escalating
29 worldwide [18]. In the Americas, dengue cases alone surpassed 7 million by May 2024, exceeding
30 the total annual of 4.6 million cases reported in the previous year [44]. Traditionally, these diseases
31 primarily affected tropical and subtropical regions [18, 21, 44]. However, climate change and global
32 warming are facilitating the spread, adaptation, and establishment of competent mosquitoes into
33 temperate zones previously unaffected by such diseases, such as Europe [11]. Additionally, increased
34 human mobility also plays a critical role, as travelers returning from endemic areas to non-endemic
35 regions may introduce infections (imported cases), potentially sparking local transmission in areas
36 with competent vectors and susceptible populations. Recently, countries like France, Italy, and
37 Spain have experienced a significant rise in dengue imported cases. In France, from the beginning
38 of 2024 up to June of the same year, the imported cases overpasses the 200 cases recorded over
39 the whole previously year (in 2023) [35]. And, around 500 imported cases were registered in Italy.
40 Additionally, there has been a marked increase in autochthonous cases, with 85 reported in France
41 and 207 in Italy [18].

42 In the Basque Country, an autonomous community in northern Spain, no autochthonous cases
43 of *Aedes* mosquito-borne diseases have been recorded to date. However, with the lifting of mobility
44 restrictions after the SARS-CoV-2 pandemic, the Public Health Epidemiological Unit in the Basque
45 Country has registered an increase of dengue, chikungunya, and Zika imported cases [32]. On the
46 other hand, entomological surveillance in various localities has shown an increase in the abundance
47 of *Aedes albopictus* eggs, and the establishment of *Aedes japonicus* populations [9]. These devel-
48 opments highlight the critical importance of maintaining robust surveillance systems, as effective
49 monitoring is essential for preventing and controlling the spread of arboviruses.

50 The mosquitoes undergo to three life stages before becoming adults: egg, larva, and pupa.
51 Female mosquitoes search for human blood since it provides the essential nutrients required for egg
52 development. After feeding, the female typically rests while her eggs mature, and then lays them
53 in small batches in areas with stagnant water, such as containers, tire ruts, or tree holes. A female
54 mosquito can lay an average of 200 to 400 eggs at a time [17, 31]. Most eggs hatch into larvae within
55 48 hours if still water is available. However, they can survive several days, from 300 to 400 days,
56 without coming into contact with water [17, 31]. This reproductive process is strongly influenced
57 by environmental factors such as temperature, humidity, and rainfall, which affect the availability
58 of suitable breeding sites and ultimately the success of egg development.

59 The worst conditions for *Aedes albopictus* eggs are high temperatures and low relative humidity
60 [25]. Egg mortality decreases with increasing relative humidity and median temperatures of 24-26°C.
61 Conversely, the optimum temperature for females to lay eggs is between 25-30°C. At temperatures
62 of 20°C and 34°C, mosquitoes lay significantly fewer eggs [20, 25]. The optimal temperature for the
63 development and survival of *Aedes albopictus* occurs at summer temperatures of 25-30°C. While a
64 mean winter temperature of more than 0°C allows egg survival, a mean annual temperature of more
65 than 11°C is required for adult activity [18]. At least 500 mm of annual rainfall is required for the
66 breeding habitat, although mosquito populations have been established in areas with lower rainfall
67 [1]. In contrast, periods of high precipitation temporarily reduce the number of females actively
68 searching for a host. The reproductive season is influenced by increasing temperatures in spring
69 and the onset of egg diapause in autumn, triggered by daylight hours below 13-14 hours [1, 18].

70 The association between climate factors and the prediction of dengue outbreaks has been widely
71 studied [5, 7, 10, 12, 14, 26, 33]. By employing machine learning approaches, particularly those
72 applied in endemic regions, have shown promise in enhancing the accuracy of dengue outbreak
73 forecasts. On the other hand, some studies have incorporated vector data, such as adult mosquito
74 populations, as proxies [12, 16, 27, 41], or larvae abundance [40]. Moreover, the role of *Aedes*
75 *aegypti* abundance, climatic factors, and disease surveillance has been also evaluated in regions
76 where autochthonous dengue transmission was recently introduced, such as in southern Brazil [12].

77 However, one of the key challenges in fitting and validating predictive models is the necessity
78 of local incidence data on mosquito-borne diseases cases and vector surveillance information. This
79 data serves as a critical predictor variable for outbreak forecasting, but it is typically only available
80 in endemic regions, where autochthonous cases is a persistent public health concern. Unfortunately,
81 such data is often limited or spatially restricted due to various factors, primarily the high costs
82 associated with collecting and maintaining accurate, up-to-date surveillance systems, making it
83 difficult to obtain comprehensive data for non-endemic or under-resourced regions [15].

84 Despite these challenges, numerous studies have successfully used climate variables and also his-
85 torical data on mosquito adult abundance as proxies to forecast mosquito abundance. For example,
86 mosquito abundance has been predicted using artificial neural network (ANN) models [27, 28], with
87 some studies using adult mosquito populations as predictors [27], while others employed mechanistic
88 models [38]. Another study used an ordinary differential equation (ODE) model to predict mosquito
89 abundance, considering temperature, rainfall, egg diapause, and population dynamics of mosquitoes
90 in southern France [43]. Nonetheless, this study did not include humidity as a climate factor, and
91 prior hypotheses based on vector-related parameters were necessary, drawn from existing literature.

92 Finally, only a few studies have considered mosquito eggs as predictors for temporal forecasting
93 [5, 8, 10]. A more recent study employed spatio-temporal forecasting using stacked machine learning
94 techniques [13]. Most studies that have used egg counts for forecasting have linked them with climate
95 changes and ovitrap data to predict dengue outbreaks in endemic regions. However, in non-endemic
96 areas like the Basque Country, where there is no local *Aedes* mosquito-borne diseases transmission
97 and adult mosquito populations are not systematically monitored, predicting mosquito abundance
98 becomes crucial for controlling the spread of the disease and informing surveillance and intervention
99 strategies.

100 In this study, we aim to estimate *Aedes* invasive mosquito abundance in a region where au-
101 tochthonous mosquito-borne diseases transmitted by *Aedes albopictus* (such as dengue) have not yet
102 been recorded, such as the Basque Country. By using the available data from the Basque Country's
103 provinces, we use machine learning techniques to model the relationship between recorded mosquito
104 ovitrap egg counts and key environmental factors, including temperature, humidity, and precipi-
105 tation. In Section 2, the relationship between climate variables and the abundance of mosquito
106 eggs is analyzed within the context of a maritime climate, as the Basque Country, at the provincial
107 and municipality levels. We explore and compare different machine learning models, considering
108 variations such as including and excluding lagged versions of egg counts as a predictor, in Section
109 3. Notably, incorporating lagged versions of both independent and dependent variables consistently
110 improves the performance of most models, demonstrating the importance of temporal dependencies
111 in mosquito abundance forecasting. Further, in Section 3, fitting the best-performing models to the
112 available data on recorded egg counts in ovitraps allow us to produce more accurate predictions of
113 invasive mosquito abundance.

114 2 Materials and Methods

115 2.1 Entomological and meteorological data

116 Data on *Aedes* mosquito egg counts from 2013 to 2023 in the Basque Country were obtained using
117 ovitraps as described in [9, 22]. Following the European Centre for Disease Prevention and Control
118 (ECDC) recommended guidelines [18], the ovitraps were distributed across the three provinces,
119 covering 63 municipalities, as shown in Figure 1(b).

120 The number of ovitraps varies by municipality, with two sampling areas selected in most cases.
121 Each sampling area typically contains five ovitraps, which are positioned in sheltered spots away
122 from direct sunlight and wind, often hidden within vegetation. Therefore, up to 10 ovitraps per
123 municipality were placed in most cases. Each ovitrap contains water and a wooden stick (or tablex)
124 that serves as a substrate for mosquito egg-laying. Every 14 days (on average), these paddles are
125 removed, and new ones are put in their place. Thus, each municipality and area is sampled roughly
126 10 to 12 times per year, from June through November [9].

127 Meteorological data for the Basque Country were collected from the Basque Meteorological
128 Agency (Euskalmet) across several weather stations (see Figure 1(b))¹, covering the period from
129 2016 to 2023. The data, obtained from the OpenData Euskadi website [29], include precipitation
130 (recorded as cumulative precipitation in millimeters (*mm*) or liters per square meter (l/m^2)), tem-
131 perature (measured in degrees Celsius ($^{\circ}C$)), and humidity (relative air humidity as a percentage
132 (%)). Weather observations were recorded every 10 minutes at each station. For this study, we
133 calculated daily averages of temperature and humidity and daily cumulative precipitation for each
134 meteorological station.

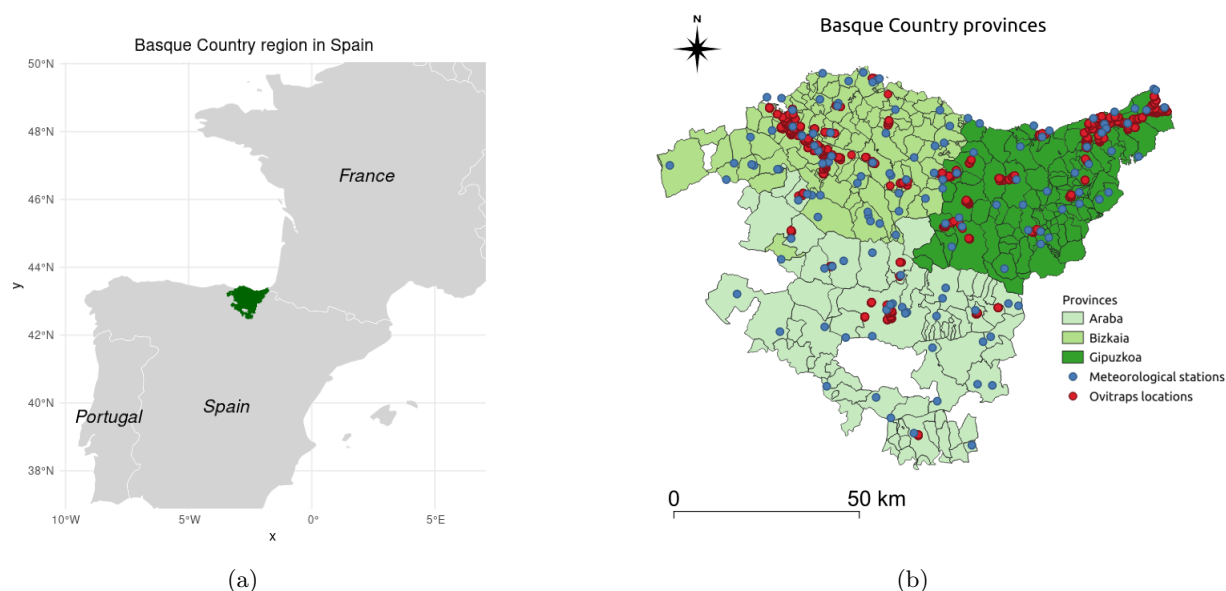


Figure 1: (a) Basque Country region in Spain (location in the European map). (b) Meteorological stations and ovitraps locations in the Basque Country provinces during the intersection study period (2016 to 2023).

¹Note that not all meteorological stations displayed on the map contain records of all environmental features selected for this study.

135 2.1.1 Study area and data per provinces

136 The Basque Country, located in northern Spain, is divided into three administrative provinces:
137 Araba (Álava), Bizkaia (Biscay), and Gipuzkoa (see Figure 1). With a total area of 7234 km^2 and
138 a population of approximately 2.18 million [19], the region is characterized by diverse landscapes
139 and a maritime climate, with temperate conditions and high annual precipitation, particularly
140 in the coastal areas. Araba, the southernmost province, has a more continental influence in its
141 climate, with drier and slightly colder conditions than the coastal provinces of Bizkaia and Gipuzkoa.
142 Bizkaia and Gipuzkoa, bordered by the Cantabrian Sea, experience milder temperatures and higher
143 humidity. These climatic differences across the provinces influence the mosquito abundance patterns,
144 which this study aims to capture and analyze through the environmental data collected.

145 For this study, we analyzed ovitrap mosquito egg counts collected in various locations across all
146 three provinces. The data was pre-processed by averaging the 20 highest egg counts per province
147 over a 14-day interval, considering that each municipality had a maximum of 10 ovitraps. This
148 approach was necessary to address inconsistencies in the number of monitored ovitraps over the
149 studied period and to avoid skewing the results with prevalent zero counts. By selecting the 20
150 largest egg counts, the data reflects meaningful mosquito activity (in at least two distinct locations),
151 effectively filtering out areas with consistently low or zero activity.

152 Meteorological data, specifically daily precipitation (cumulative precipitation in millimeters
153 (mm), air temperature (in degrees Celsius ($^{\circ}C$)), and relative humidity (percentage (%)), were
154 obtained by averaging daily values from all available meteorological stations in each province.
155 These features were then aggregated over the previous 14 days to maintain consistent time in-
156 tervals between the entomological and meteorological datasets. The average annual temperature
157 and accumulated precipitation in each province align with environmental conditions favorable for
158 *Aedes albopictus* survival, approximately 11.5 $^{\circ}C$ and 878 mm in Araba, 13.8 $^{\circ}C$ and 1278 mm in
159 Bizkaia, and 13.4 $^{\circ}C$ and 1610 mm in Gipuzkoa [9], which are consistent with the survival thresholds
160 discussed in the literature for this species [1, 18].

161 The time series of the average egg counts, temperature, humidity, and cumulative precipitation
162 for each province in the Basque Country are shown in Figure 2.

163 Mosquito eggs are typically found during the summer months, from June to October, when
164 the combination of higher temperatures and favorable humidity conditions promotes their activity
165 and reproduction. As shown in Figure 2(a), the egg count in the entire Gipuzkoa province has
166 significantly increased over the last years of collected data, although this trend may vary between
167 municipalities. For example, in the city of Irun (see Supplementary Material B), the second most
168 populated city in Gipuzkoa, located on the border with France, where variability is present without
169 a clear increasing trend.

170 In Gipuzkoa, temperature exhibited a clear seasonal annual pattern, while accumulated rainfall
171 showed no apparent trend. Humidity, however, decreased during the winter and followed a quasi-
172 periodic structure (see Figure 2(b)). The winter of 2019, right after the expected period of higher
173 egg presence, was exceptionally rainy compared to other winters in the province. Combined with low
174 humidity (below 75%), this may have contributed to the lower egg counts observed in the following
175 summer season (2020). In contrast, the dry summer of 2022, accompanied by higher humidity levels
176 (above 75%), may explain the increased egg counts observed that year.

177 In Bizkaia, the time series of egg counts has displayed a consistent upward trend over the years,
178 with positive egg traps first recorded in 2017 (see Figure 2(c)). Notably, the average mosquito egg

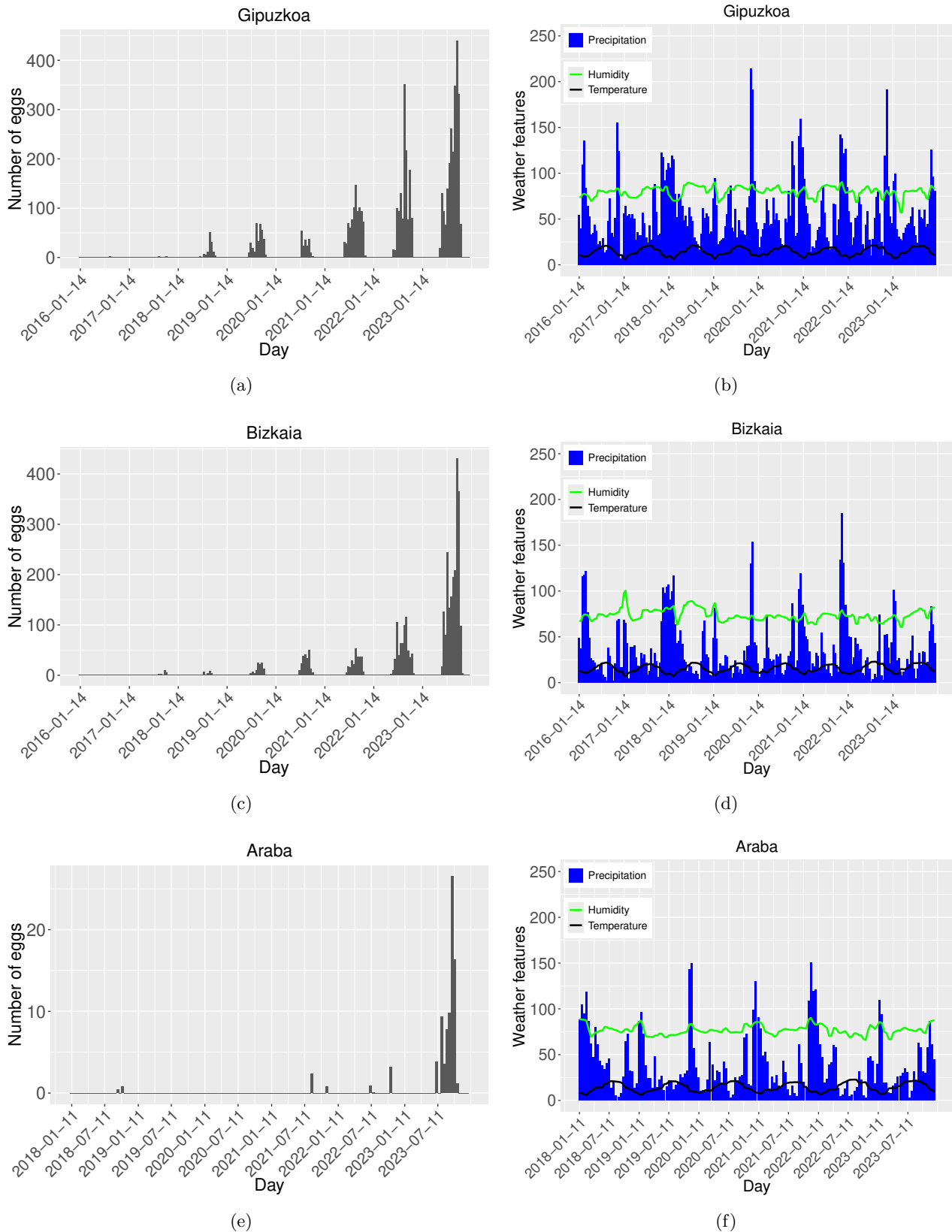


Figure 2: Number of mosquitoes eggs collected in (a), (c), (e). And average temperature ($^{\circ}C$), relative air humidity (%), and cumulative precipitation (mm) in (b), (d), (f). Data gathered biweekly for Gipuzkoa, Bizkaia, and Araba, respectively.

179 count in Bizkaia during 2023 serves as a good proxy for the province-wide average, as shown by the
 180 time series for Bilbao, the capital of Bizkaia (see Figure 13(c) in the Supplementary Material B).

181 The temperature in Bizkaia followed a clear seasonal pattern, while accumulated rainfall showed
182 no apparent trend, with significant cumulative precipitation occurring later in 2021. In contrast to
183 Gipuzkoa, however, humidity in Bizkaia exhibited periodic increases approximately every two years,
184 with higher levels typically observed during winter months (see Figure 2(d)).

185 Moreover, average precipitation in Bizkaia was slightly lower than in Gipuzkoa. Temperature
186 fluctuations in Bizkaia were more pronounced, as indicated by the steeper slope of its temperature
187 curve compared to Gipuzkoa, potentially explaining the lower average egg counts in the region.
188 Additionally, the dry summer of 2021, followed by a rainy winter, may have contributed to the
189 consistent egg count trend observed.

190 Furthermore, although ovitraps have been distributed and data collected in the province of Araba
191 since 2013, positive egg traps were not recorded until 2018, with no positive ovitraps observed in
192 2019 or 2020 (see Figure 2(e)). In Laudio, the second most populated municipality in Araba, positive
193 ovitraps were only recorded in 2021 (see Figure 13(e) in Supplementary Material B).

194 The average temperature in Araba exhibits annual seasonality, while precipitation lacks a clear
195 trend, though cumulative rainfall is typically higher during winter. On the other hand, humidity
196 also tends to increase alongside precipitation (see Figure 2(f)). The lower average temperature in
197 this province may contribute to the reduced presence of mosquito eggs.

198 Given the dispersed nature of data in Araba, with many zero values in egg counts (Figure 2(e)),
199 there is insufficient information to develop a reliable training dataset for model fitting. Therefore,
200 this province is excluded from further analysis. Smaller spatial units, such as individual municipal-
201 ities, are similarly excluded, with the focus of this study being the two Basque Country provinces,
202 Gipuzkoa and Bizkaia. Nonetheless, descriptive statistics and detailed analyses at the municipal
203 level for Irun and Bilbao, which have adequate data, are provided in the Supplementary Material
204 B.

205 **2.2 Methodological approach**

206 **2.2.1 Data processing**

207 After gathering data, pre-processing is a crucial initial step before model training, forecasting,
208 and evaluation. In this study, data pre-processing included the following steps. First, we ensured a
209 consistent interval for both the independent and dependent variables, selecting a biweekly interval
210 for the entomological data based on the average 14-day period in which egg counts were collected.

211 Next, we addressed missing values through imputation, filling gaps with zero values. This choice
212 is scientifically justified within the context of this dataset, as institutional data indicated that, for
213 months without data collection, ovitrap counts would have likely been zero [9]. This assumption
214 was based on data from four sentinel points (two in Gipuzkoa and two in Bizkaia) monitored over
215 a year to determine the start and end of *Aedes* mosquito activity in regions with recorded presence
216 in the previous year.

217 Additionally, we included only the 20 highest egg counts at the provincial level to account
218 for variations in the number of monitored ovitraps over time, helping to reduce dataset skewness.
219 Outliers were then removed using a central moving average as a smoothing method, commonly
220 applied to mitigate white noise, random fluctuations, and extreme values [39].

221 For the meteorological data, no imputation was required as daily weather data was available
222 for the entire study period. In this case, outliers were retained as they could signal significant

223 events associated with the presence or absence of mosquito eggs. Basic exploratory analysis was
224 then conducted using descriptive statistics and correlation tests, incorporating both the original
225 and lagged versions of the meteorological data.

226 Finally, we split the data into training and testing sets, with the training data comprising
227 85.71% and 83.33% for Gipuzkoa and Bizkaia, respectively. The remaining 26 data points (one year
228 of biweekly data) were allocated for testing.

229 2.2.2 Models

230 In this study, we applied different models including and excluding the lagged version of eggs count
231 as a proxy and the lagged version of the independent environmental variables. To appropriately
232 handle the discrete and non-negative nature of counts, we restrict our choices and applications of
233 the models presented here. For instance, the statistical methods such as the Poisson Regression
234 and Negative Binomial Regression are foundational models for count data [36]. However, while the
235 first one assume that the time series follows a Poisson distribution, the second one can be useful
236 when time series presents more variability and over-dispersion (i.e., the variance is greater than the
237 mean) (as it is the case). Both models are an extension of the Generalized Linear model (GLM)
238 with a log link function.

239 The GLM is a flexible extension of ordinary linear regression that accommodates response vari-
240 ables with error distributions other than the normal distribution. This model often outperforms
241 others when applied directly to the original data, compared to the transformed data such as using
242 logarithmic scale [2, 30]. As such, we initially avoided any normalization or transformation of the
243 data. When we applied the GLM to this dataset, it performed better on the smoothed data (using
244 a three-point central moving average) than on the original, unprocessed data. And, a GLM with
245 the canonical link function was used, assuming a Gaussian distribution for the response variable. In
246 other words, the response variable follows a Gaussian exponential family distribution. This allows
247 for more flexibility in modeling, as it does not impose the strict relationship between mean and
248 variance required by models such as when using the Poisson distribution [23].

249 The GLM with a Gaussian family assumes a linear relationship between the predictors and the
250 response variable \mathbf{Y} , using the identity link function. That is, the conditional mean $\boldsymbol{\mu}$ is a linear
251 combination of unknown parameters $\boldsymbol{\beta}$ via the link function g :

$$E(\mathbf{Y} | \mathbf{X}) = \boldsymbol{\mu} = g^{-1}(\mathbf{X}\boldsymbol{\beta}) = \beta_0 + \beta_1x_1 + \cdots + \beta_px_p,$$

252 where $E(\mathbf{Y} | \mathbf{X})$ is the expected value of \mathbf{Y} conditional to \mathbf{X} , and g is the link function, which
253 in this case, is the identity function [23]. The model predicts the mean of the response variable
254 based on the input variables, and estimates the coefficients $\boldsymbol{\beta}$ by maximizing the likelihood function,
255 assuming that the residuals are normally distributed.

256 Moreover, not only the dataset is over-dispersed but the response variables contains a lot of zeros,
257 due to two main reasons: one the absence of positive eggs count outside during winter season (the so-
258 called true negative) and the absence of more samples in more localities (the so-called false negative).
259 In this case, zero-inflated models can handle excess zeros effectively. Zero-Inflated Negative Binomial
260 (ZINB) can be effective in this case assuming that the data come from a mixture of two processes:
261 one generating zeros and another generated by a negative binomial distribution [37]. Other models
262 that can handle over-dispersion and the zero counts is the Generalized Additive Models for Location,

263 Scale, and Shape (GAMLSS) which is flexible in modeling different distributions, not just the mean
264 but also the variance [45]. However, after applying these models to the dataset, we observed that
265 they were prone to over-fitting, indicating that the model might learned the noise in the training
266 data rather than generalizing the unseen data.

267 On the other hand, our predictors are temporal series mostly exhibiting seasonal trends. Time se-
268 ries models like Seasonal Autoregressive Integrated Moving Average (SARIMA) are commonly used
269 for forecasting since it is suitable for temporal count data and can handle seasonality. SARIMA has
270 been extended to the Seasonal Autoregressive Integrated Moving Average with Exogenous variables
271 (SARIMAX) which can include exogenous variables giving more accurate outcomes. SARIMAX
272 combines differencing, autoregression, moving averages, and seasonal components, incorporating
273 as well exogenous predictors [24]. Unlike models such as the GLM, this model assumes that the
274 response variable depends on its past values Y_t . Also, of its past forecast errors ϵ_t , and external
275 predictors, capturing temporal effects. This relation reads:

$$Y_t = \phi_1 Y_{t-1} + \dots + \phi_p Y_{t-p} + \theta_1 \epsilon_{t-1} + \dots + \theta_q \epsilon_{t-q} + \mathbf{X}\boldsymbol{\beta} + \epsilon_t,$$

276 where Y_t is the response variable at time t , ϕ_i are the the autoregressive (AR) parameters, θ_i the
277 moving average (MA) parameters, \mathbf{X} is the predictors (exogenous variables), and $\boldsymbol{\beta}$ is the vector of
278 coefficients.

279 The model depend, as well, on the order of the AR terms p , representing the number of lagged
280 values of the series used in the model; the degree of differencing d , which removes trends and makes
281 the series stationary; the order of the MA terms q , representing the number of lagged forecast errors.
282 And, on the seasonal component, P , D , and Q that are the seasonal terms for the parameters p , d ,
283 and q , respectively, and s is the length of the seasonal cycle s [24]. We implemented SARIMAX in
284 the R compute language by using the `auto.arima()` function that automatically selects the best
285 seasonal and non-seasonal parameters p, d, q, P, D, Q , and s based on the data.

286 On the other side, machine learning techniques such as Random Forest (RF) [6], Conditional
287 Inference Trees (CTree), and Artificial Neural Networks (ANNs) can be also used for forecasting
288 count data. However, in this study, ANNs is the least performing machine learning model, a finding
289 corroborated by previous research [37] which do not advises using ANNs for count data with over-
290 dispersion.

291 RF builds decision trees using bootstrap samples and random feature subsets, and combines
292 the predictions from all trees. Each tree is developed using a subset of features, as chosen by the
293 `mtry` parameter [6]. The `mtry` parameter determines the number of predictor variables considered
294 at each split, playing a crucial role in controlling over-fitting. The `ntree` parameter refers to the
295 number of trees to be generated in the RF. Increasing the number of trees generally enhances model
296 stability and robustness, although beyond a certain threshold number of trees, the additional trees
297 yield to insignificant improvements in the terms of model performance. The advantage of RF lies
298 in its ability to handle complex data and it is designed to mitigate over-fitting [14, 37].

299 The final RF predictions for the conditional mean of \mathbf{Y} , given the predictor \mathbf{X} , is based on the
300 average or weighted average of all the individual trees' predictions. Thus, the RF model can be
301 expressed as:

$$\hat{\mathbf{E}}(\mathbf{Y} | \mathbf{X}) = \frac{1}{K} \sum_{k=1}^K \omega_k h_k(\mathbf{X})$$

302 where $h_k(\mathbf{X})$ is the prediction of the k -th tree, and K is the total number of trees [6]. Each tree is
303 built using a bootstrap sample of the original data and selects features at random from the `mtry`
304 subset.

305 On the other hand, the CTree is a non-linear method to model the relationships between predic-
306 tor variables and a response variable. The CTree algorithm recursively partition the dataset based
307 on the values of the predictors, using statistical tests to determine the significance of potential splits.
308 The splits are chosen by testing the association between each predictor and the response, and the
309 predictor with the strongest association (lowest p -value) is selected for each split.

310 The conditional distribution of \mathbf{Y} , given the predictor \mathbf{X} , is estimated as:

$$\hat{\mathbf{E}}(\mathbf{Y} | \mathbf{X}) = \sum_{j=1}^J \hat{Y}_j w_j(\mathbf{X})$$

311 where \hat{Y}_j is the predicted value for the j -th terminal node, $w_j(\mathbf{X})$ is the weight indicating whether
312 observation j falls into the same terminal node as \mathbf{X} [42].

313 While RF and CTree both rely on decision tree methodologies, they differ in their approaches.
314 RF employs random feature selection to create an ensemble of trees, which enhances generalization
315 but sacrifices interpretability. Conversely, CTree focuses on unbiased variable selection, offering
316 better interpretability. RF generally offers better predictive performance on large and complex
317 datasets, while the structural differences in the partitions can highlight the unique advantage of
318 CTree.

319 We implement the GLM, SARIMAX, RF and CTree models (and other models discussed in
320 this section) in the R computing language (R version 3.6.3) using the packages `MASS`, `forecast`,
321 `randomForest` and `party`, respectively. Nevertheless, only the four models discussed earlier will be
322 presented in this study because, as previously mentioned, some models exhibit over-fitting, others
323 demonstrate under-fitting (as is the case with the ANNs model), and some fail to capture any
324 significant features of the data.

325 2.2.3 Stationary analysis

326 We applied the augmented Dickey-Fuller (ADF) test, a commonly used method for testing the
327 presence of a unit root in time series data, to assess whether the time series is non-stationary. Non-
328 stationarity in a time series often presents means, variances, and covariances that change over time,
329 making the series unpredictable and challenging to model or forecast. Although some models, such
330 as SARIMAX, can handle non-stationarity, stationary time series often yield more reliable results.

331 The null hypothesis of the ADF test states that the series contains a unit root, indicating non-
332 stationarity, while the alternative hypothesis suggests that the series is stationary. To test the null
333 hypothesis, we computed the p -value. A p -value less than 0.05 leads us to reject the null hypothesis,
334 implying stationarity.

335 We conducted the ADF test using the `tseries` package in R. For both datasets, Gipuzkoa and
336 Bizkaia, the ADF test on the predictor variable yielded a p -value of approximately $0.01 < 0.05$,
337 indicating that the datasets are stationary.

338 2.2.4 Evaluation metrics

339 To compare the performance of statistical and machine learning models, three widely used
340 evaluation metrics were employed: the Mean Absolute Error (MAE), the Root Mean Squared Error
341 (RMSE), and the R-squared (R^2) score.

342 The MAE is calculated as:

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (1)$$

343 where y_i and \hat{y}_i represent the observed and predicted values, respectively, and $|\cdot|$ denotes the
344 absolute value [37]. MAE measures the average magnitude of the errors in a set of predictions,
345 without considering their direction.

346 The RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (2)$$

347 where y_i and \hat{y}_i are the observed and predicted values, respectively. RMSE gives a higher weight to
348 large errors compared to MAE and is sensitive to outliers.

349 The R^2 score, also known as the coefficient of determination, is calculated as:

$$R^2 = 1 - \frac{S_r}{S_t} = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (3)$$

350 where, S_r is the Residual Sum of Squares, representing the sum of squared differences between the
351 observed values (y_i) and the predicted values (\hat{y}_i); and S_t is the Total Sum of Squares, calculated
352 as the sum of squared differences between the observed values (y_i) and their mean (\bar{y}). An R^2 score
353 of 1 indicates that the model explains all the variability of the response variable, while a score of 0
354 indicates no explanatory power.

355 The selection of the best model is based on achieving the lowest MAE or RMSE values, or an
356 R^2 score closest to 1. In this study, the MAE is chosen as the primary evaluation metric due to its
357 suitability for machine learning models [37].

358 3 Results

359 3.1 Exploratory statistical analysis

360 Basic exploratory statistical analysis was performed, starting with descriptive statistics for both
361 the response and predictor variables (after pre-processing and smoothing). All variables in the
362 dataset were found to be skewed and over-dispersed. The null hypothesis of normal distribution
363 was rejected based on the results of the Kolmogorov-Smirnov test and the Shapiro-Wilk test, both
364 of which yielded p -values < 0.05 , indicating significant deviation from normality for all variables.

365 For the Gipuzkoa dataset:

- 366 - Eggs count had a mean of 25 and a median of 0.
- 367 - Temperature (in °C) had a mean of 14.6 and a median of 14.2.
- 368 - Relative air humidity (in %) had a mean of 79.9 and a median of 80.8.
- 369 - Precipitation (in mm) had a mean of 57.6 and a median of 49.4.

370 For the Bizkaia dataset:

- 371 - Eggs count had a mean of 16 and a median of 0.
- 372 - Temperature (in °C) had a mean of 15.5 and a median of 14.8.
- 373 - Relative air humidity (in %) had a mean of 73.6 and a median of 72.8.
- 374 - Precipitation (in mm) had a mean of 37.4 and a median of 27.1.

375 Additional details are provided in the Supplementary Material (see Figure 10), as well for the
376 province of Araba.

377 The relationship between meteorological variables and the number of mosquito eggs was explored
378 using scatter plots (see Figures 3(a)-(c) for Gipuzkoa and 3(d)-(f) for Bizkaia). No linear relationship
379 was confirmed, as indicated by Pearson's correlation index. Nevertheless, it is well-known that the
380 combination of high temperatures (22 °C to 27 °C) and high humidity increases oviposition rates
381 (egg-laying) in adult female mosquitoes [20, 25]. This association is reflected in Figures 3(a) and
382 (b) for Gipuzkoa, and Figures 3(d) and (e) for Bizkaia.

383 As described in Section 2.2.1, the time series data were smoothed using a central moving average
384 to reduce short-term fluctuations and noise. This preprocessing step helped mitigate spurious
385 short-term correlations and revealed underlying long-term relationships between variables, thereby
386 increasing correlation indexes.

387 On the other hand, Spearman's correlation analysis confirmed a strong monotonic relationship
388 between the number of eggs and temperature, with a correlation index ≥ 0.72 (see Figures 4(a) for
389 Gipuzkoa and 4(b) for Bizkaia). Although no significant correlations were found between egg counts
390 and the other climate variables, the direction and strength of these relationships are displayed in
391 Figures 4(a) and 4(b). Specifically, As humidity increases, the number of eggs increases, showing
392 an intermediate correlation. In contrast, as accumulated precipitation increases, the number of eggs
393 decreases, albeit with very low or negligible correlation.

394 In addition, we have created lagged time series for all the meteorological variables (see Fig-
395 ure 5(a)-(c) and 5(d)-(f)), and we evaluate the monotonic correlation using Spearman correlation,
396 highlighting the time lag at which the highest index value occurs (see Figure 11(a) and 12(a) in
397 the Supplementary Material A). At the time lag at which the highest correlation value occurs, the
398 lagged time series will be used as predictor variables (see Figure 11(b) and 12(b) for Gipuzkoa and
399 Bizkaia, respectively, in the Supplementary Material A).

400 For instance, Figure 5(a) shows that the highest correlation value between egg counts and
401 temperature series, in Gipuzkoa, occurs at lag -1. This could imply that the egg production series is
402 most strongly correlated with the temperature series 2 weeks (1 period) earlier. Therefore, changes
403 in temperature might have a leading effect on egg production, where temperature changes influence
404 egg production with a delay of 1 period (2 weeks). For humidity (Figure 5(b)), the highest correlation
405 occurs at 0 units (0 weeks) with a (low) positive correlation, while for precipitation, (Figure 5(c)),
406 the highest correlation occurs at lag -5 units (10 weeks) with a negative (low) correlation.

407 For Bizkaia, Figure 5(d) shows that the highest correlation value between egg counts and tem-
408 perature series occurs at lag -1. This could imply that the egg production series is most strongly
409 correlated with the temperature series 2 weeks (1 period) earlier. Therefore, changes in temperature
410 might have a leading effect on egg production, where temperature changes influence egg production
411 with a delay of 1 period (2 weeks). For humidity, (Figure 5(e)), the highest correlation occurs at

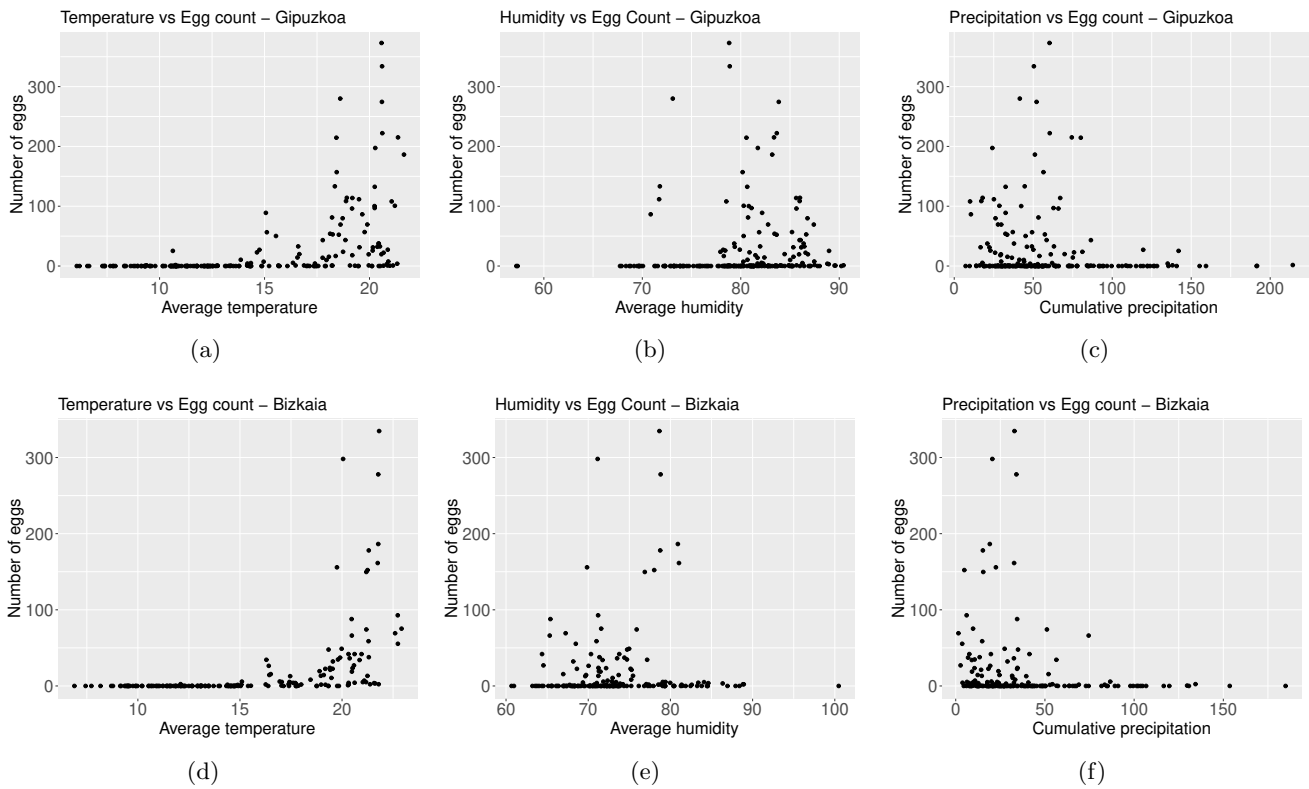


Figure 3: Average temperature (in °C) versus the number of collected mosquito eggs, in (a), (d). Average relative air humidity (in %) versus the number of collected mosquito eggs, in (b), (e). Accumulated precipitation (in mm) versus the number of collected mosquito eggs, (c), (f). Data gather biweekly, in Gipuzkoa and Bizkaia, respectively.

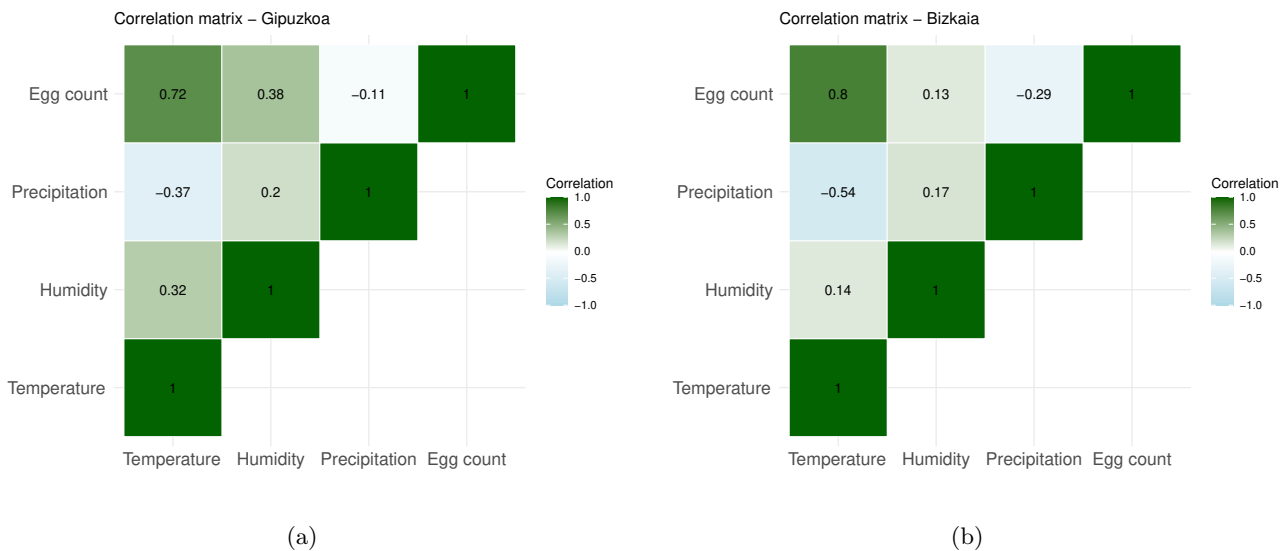


Figure 4: Spearman correlation matrix between weather features and the number of mosquito eggs. The matrix shows a high correlation between the number of eggs and temperature (index = 0.72 for Gipuzkoa and index = 0.8 for Bizkaia), but no significant correlation with the other features.

412 lag -2 units (4 weeks) with a positive correlation, while for precipitation, (Figure 5(f)), the highest
 413 correlation occurs at lag -5 units (10 weeks) with a negative correlation.

414 High correlation is shown between egg counts and temperature at lag -1, with an index value of

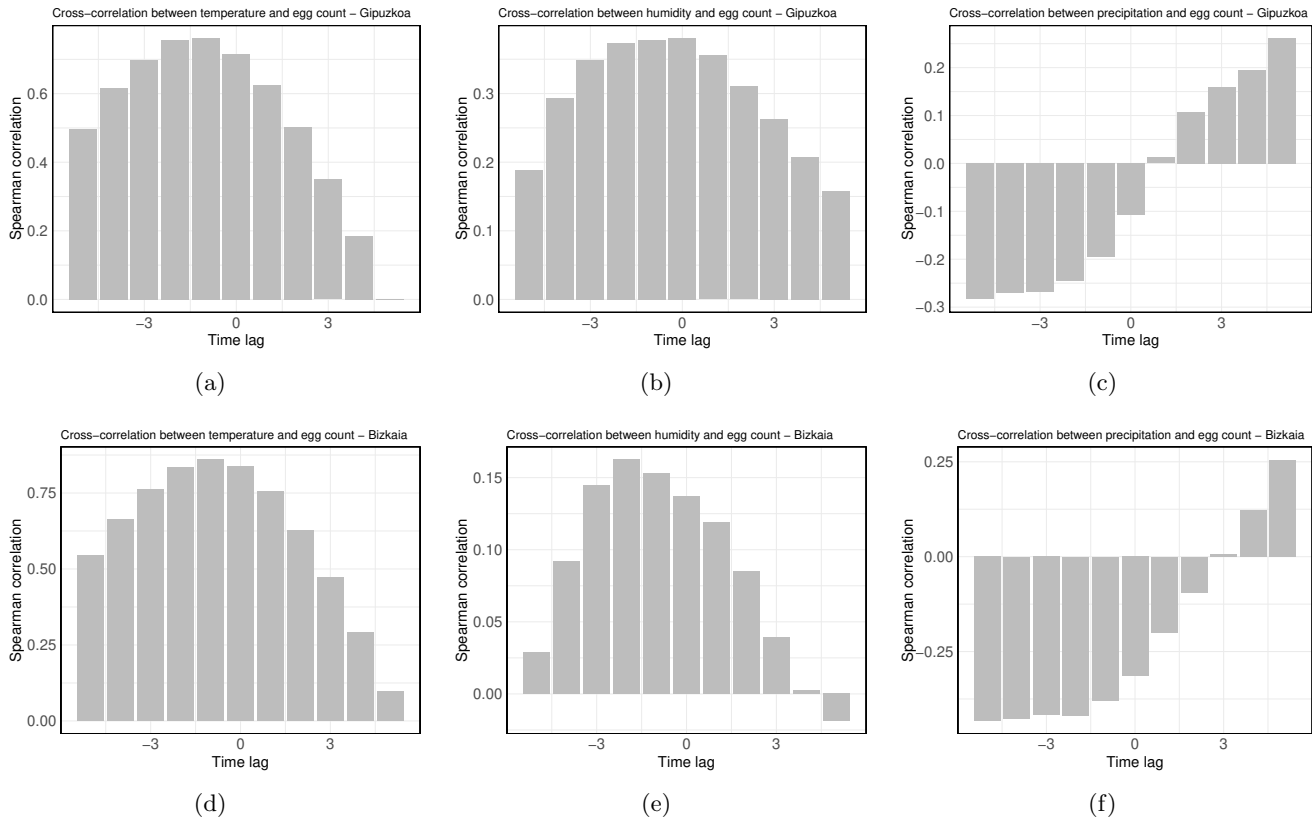


Figure 5: Spearman correlation between the lagged time series of weather features and the number of mosquito eggs, with a time lag of 1 unit (2 weeks). For temperature, the maximum correlation occurs at a lag of -1 unit, in (a), (c). For humidity, the maximum correlation occurs at a lag of 0 units and -2 units, for Gipuzkoa and Bizkaia, in (b), (e) respectively. For precipitation, the maximum correlation occurs at a lag of -5 units, in (c), (f).

415 0.76 and 0.83 for Gipuzkoa and Bizkaia, respectively (see Figures 11(b) and 12(b) in the Supple-
 416 mentary Material A). While intermediate to low correlation appears to be positive and correlated
 417 between humidity and egg counts, Figures 3(b) and 3(e) show that the highest egg count occurs
 418 when humidity percentages are between 70% and 80%.

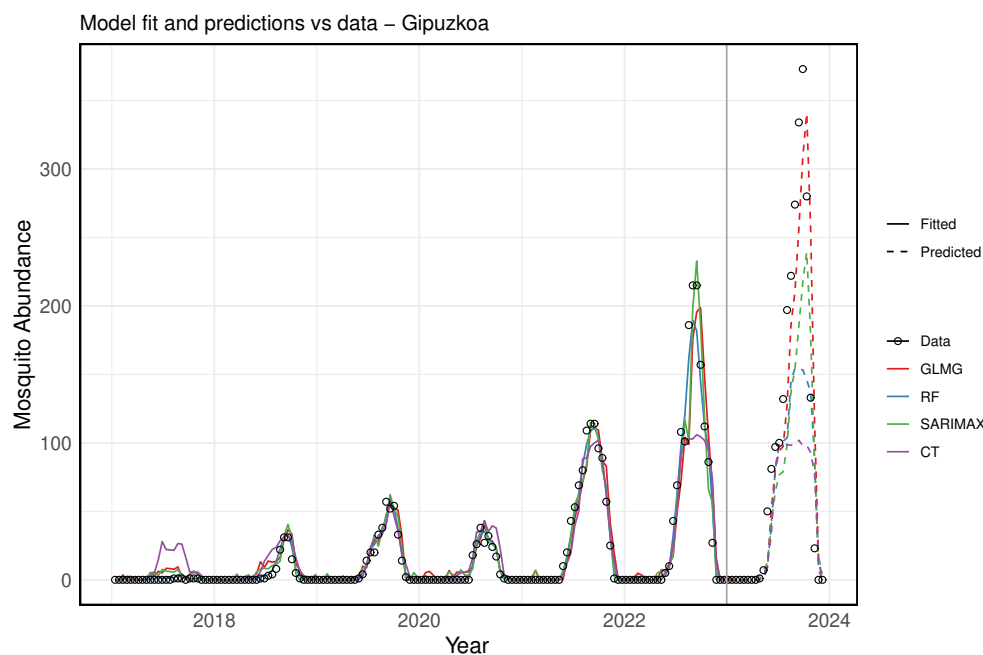
419 Moreover, a low negative correlation between precipitation and egg counts was found (see Figures
 420 5(c) and 5(f)). Although the strength of the correlation is considered low, the opposite direction in
 421 the correlation for precipitation approximately 10 weeks prior to egg collection (almost 3 months
 422 earlier) can be explained by the fact that periods of high precipitation temporarily reduce the
 423 number of females actively searching for a host and, therefore, laying eggs [18]. On the other
 424 hand, drier periods occurring 10 weeks before the collection, increase the egg counts. This can be
 425 attributed to the fact that mosquito eggs are extremely resistant. They can remain viable in a dry
 426 state within a container for 300 to 400 days without direct water contact, allowing them to stay in
 427 ovitraps for extended periods without hatching [31].

428 3.2 Fitting and error analysis

429 Prior to model fitting, the dataset was divided into training and testing sets. For Gipuzkoa,
 430 the training dataset includes data from 2017 to 2022 (85.71%), while the test dataset consists of
 431 data points from the year 2023 (see Figure 6). In contrast, for Bizkaia, the training dataset covers
 432 the period from 2018 to 2022 (83.33%), with 2023 as the test dataset (see Figure 7). The choice of

433 yearly data is related to the frequency of data availability, while the starting point corresponds to
434 the need for cleaning the missing values (NA not available) due to the lagged versions of variables.

435 We train several models on the training dataset, considering, lagged version of the independent
436 variables, as well as including and excluding lagged version of the eggs count variable. The major-
437 ity of models performed better including the proxy lagged version. Here we include only models
438 with the best performances. Which are: the Random Forest (RF) model, the Generalized Linear
439 Model (GLM) with Gaussian distribution (here abbreviated by GLMG), the Seasonal Autoregres-
440 sive Integrated Moving Average with Exogenous variables (SARIMAX) model and, the Conditional
441 Inference Trees (CTree) (here abbreviated by CT).



Model	Evaluation Metrics					
	MAE Train	MAE Test	RMSE Train	RMSE Test	R ² Train	R ² Test
RF	2.59	41.67	5.73	73.94	0.98	0.60
SARIMAX	4.77	38.73	9.85	61.03	0.94	0.73
GLM	6.09	29.37	12.32	45.71	0.90	0.85
CT	7.62	53.51	17.47	95.49	0.81	0.34

Figure 6: Comparison of actual data with the fitted and test values for Gipuzkoa. The actual data is represented by open black circles, while the fitted values are shown as solid lines and the test values as dashed lines. The models are represented as follows: in blue, the Random Forest (RF) model ($n_{tree} = 600$, $m_{try} = 5$); in red, the Generalized Linear Model (GLMG); in green, the SARIMAX model; and in purple, the Conditional Inference Trees (CT) model ($n_{tree} = 500$, $m_{try} = 3$). The vertical gray line separates the training dataset (2017–2022) from the test dataset (2023). The table shows the error metrics for each chosen model, on the training and test datasets for Gipuzkoa.

442 We implement the GLMG, SARIMAX, RF and CT models in the R computing language (R
443 version 3.6.3) using the `glm()`, `auto.arima()`, `randomForest()` and `cforest()` function, respec-
444 tively. We employed and compare the models on the training dataset and on the testing dataset.
445 Later, we evaluate each models performance in the datasets using MAE, RMSE and R^2 metrics.

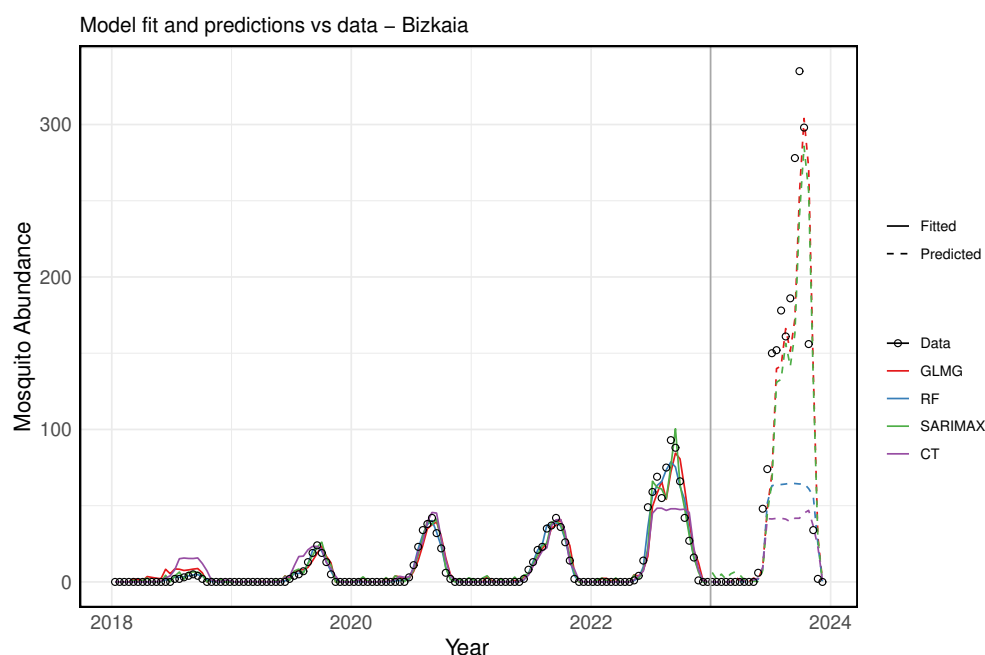
446 Based on the evaluation metrics the best performance on the training dataset for Gipuzkoa is the
447 RF model. The model could explain 98% of the variance in the data, according to the R^2 evaluation
448 While in the test dataset 60%. For the test dataset GLMG was the model that performed better,

449 explaining 85% of the variance in the dataset, followed by the SARIMAX model (see Figure 6).

450 Although the RF model performed best during training, its predictions ranked in third place,
 451 which might suggest an over-fitting. On the other hand, even though GLMG did not top the
 452 training performance, it gave better predictions, making it a more reliable model overall. This
 453 suggests that the simplicity of GLMG helped it generalize better to the unseen data, while RF may
 454 have captured the noise from the training set, which could reduce the predictive accuracy.

455 In the case of Bizkaia, the RF model performed best on the training dataset, explaining 98%
 456 of the variance, as indicated by the R^2 value. However, on the test dataset, it explained only 15%
 457 of the variance. For the test dataset, the GLMG model performed the best, explaining 81% of the
 458 variance, closely followed by the SARIMAX model with 80% (see Figure 7).

459 Among the four models evaluated, the CT model performed the worst, based on all error metrics
 460 for both the training and test datasets. Additionally, the CT model was unable to explain the test
 dataset for Bizkaia.



Model	Evaluation Metrics					
	MAE Train	MAE Test	RMSE Train	RMSE Test	R^2 Train	R^2 Test
RF	1.16	57.28	2.84	97.54	0.98	0.15
SARIMAX	2.34	30.23	4.43	47.83	0.95	0.80
GLM	3.31	27.43	5.94	46.43	0.90	0.81
CT	3.97	64.68	8.28	109.37	0.81	-0.07

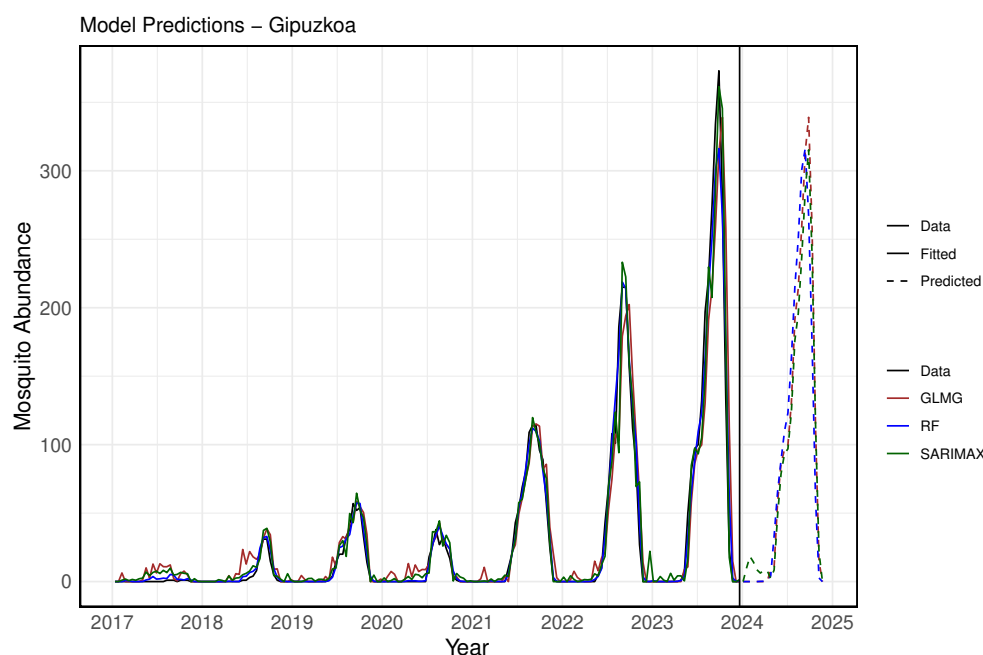
Figure 7: Actual data versus the fitted and test values of the models for Bizkaia. The actual data is represented by open black circles, while the fitted values of each model are shown with solid lines, and the test values with dashed lines. In blue, the RF model ($n_{tree} = 600$, $m_{try} = 5$); in red, the GLMG; in green, the SARIMAX model; and in purple, the CT model ($n_{tree} = 500$, $m_{try} = 3$). The vertical gray line delineates the training dataset (2018–2022) from the test dataset (2023). The table shows the error metrics for each chosen model, on the training and test datasets for Bizkaia.

461
 462 The poor performance of the models on the Bizkaia test set, as shown in the time series, can
 463 be attributed to differences in the characteristics of the training and test data. Notably, the mean
 464 value of the training data is significantly lower than that of the test data, with egg counts in 2023
 465 being unusually high. This discrepancy between the training and test datasets likely contributed to

466 the models' suboptimal performance for Bizkaia.

467 After training, testing, and evaluating each model, we used the models with the best performance
 468 to predict future *Aedes* invasive mosquito abundance. For this, we included 2023 data points in
 469 the training dataset and, using the historical time series data along with lagged versions of the
 470 variables, we forecast values based on the last observations.

471 Figure 8 shows the fitted values and predictions for mosquito abundance in Gipuzkoa for 2024,
 472 while Figure 9 presents the same for Bizkaia. Only the three models with the best performance are
 473 displayed. It is noteworthy and expected that extending the training dataset length improved the
 474 performance of all models. This highlights the importance of maintaining entomological surveillance
 475 for more accurate future predictions.



Evaluation Metrics in the training dataset			
Model	MAE	RMSE	R ²
RF	3.72	8.70	0.98
SARIMAX	6.79	14.02	0.95
GLM	9.60	20.06	0.90

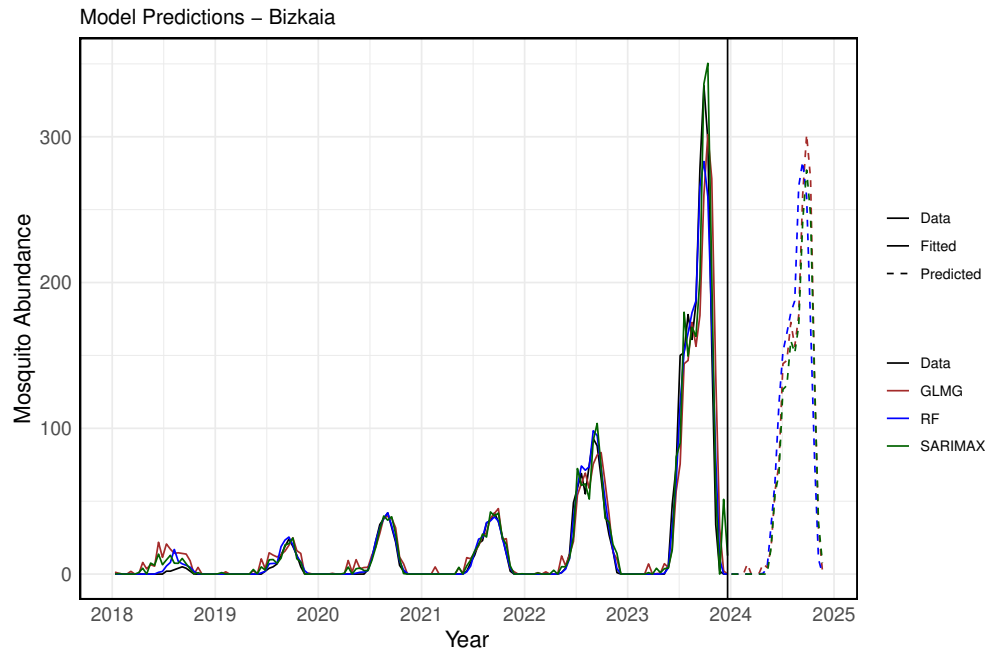
Figure 8: The actual data versus the fitted and predicted values for Gipuzkoa. The actual data is represented by a solid black line. The fitted values for each model are shown as solid colored lines, and the predicted values are displayed as dashed lines. In blue, the RF model ($n_{tree} = 600$, $m_{try} = 5$); in brown, the GLMG model; and in green, the SARIMAX model. The vertical black line delineates the training dataset (from 2017 to 2023) from the forecasted period for the year 2024. The table shows Evaluation of error metrics in the training dataset (for Gipuzkoa) showing the best model performance.

476 The error analysis for the training dataset is shown in Figure 8 and Figure 9. The results
 477 indicate that, across different metrics, the RF model provided the best fit, explaining 98% of the
 478 variance in both Gipuzkoa and Bizkaia, making it suitable for forecasting. The SARIMAX model
 479 also performed well, explaining 95% and 94% of the variance in Gipuzkoa and Bizkaia, respectively.

480 At the municipal level (see more details in Supplementary Material B), the RF model performed
 481 best for Irun (in Gipuzkoa), while for Bilbao (in Bizkaia), the SARIMAX model provided the best
 482 fit, explaining 97% of the variance in the training dataset.

483 Furthermore, we estimate and expect that mosquito abundance in 2024 will be lower compared

484 to the previous year, both at the provincial and municipal levels. This reduction may be due to
485 various factors, such as changes in optimal environmental conditions and potential variations in
486 weather patterns.



Evaluation Metrics in the training dataset			
Model	MAE	RMSE	R ²
RF	2.81	7.91	0.98
SARIMAX	5.27	12.44	0.94
GLM	7.47	18.82	0.87

Figure 9: The actual data versus the fitted and predicted values of the model for Bizkaia. The actual data is represented by a solid black line, while the fitted values of each model are shown as solid colored lines, and the predicted values as dashed colored lines. In blue, the RF model ($n_{tree}=600$, $m_{try}=5$); in brown, the GLMG model; and in green, the SARIMAX model. The vertical black line delineates the training dataset (from 2018 to 2023) from the forecasted period for the year 2024. The table shows the evaluation of error metrics in the training dataset (for Bizkaia) showing the best model performance.

487 4 Discussions and conclusions

488 The Basque Country, an autonomous community in northern Spain, has experienced an increase
489 in imported cases of mosquito-borne diseases, along with the establishment and expansion of *Aedes*
490 *albopictus* and *Aedes japonicus* mosquitoes. This study uses egg count data retrieved from ovitraps
491 monitored by the regional surveillance program conducted by the Department of Public Health of
492 the Basque Government and the public agency NEIKER at various locations across the Basque
493 provinces. We employ statistical models and machine learning techniques to model the relation-
494 ship between the recorded mosquito ovitrap egg counts and climate factors such as temperature,
495 humidity, and precipitation.

496 Before selecting the model, statistical analysis was conducted on the dataset to examine the
497 influence of environmental factors on the predictor variables. We compared different models, in-
498 cluding versions with and without lagged egg counts as a proxy. Importantly, incorporating lagged
499 versions of all independent and dependent variables improved the performance of most models.

500 We found that forecasting mosquito abundance is particularly challenging in non-endemic areas,
501 where no local mosquito-borne cases have been reported. While environmental factors are the
502 primary drivers of mosquito abundance and distribution, the time series data are not always linearly
503 correlated, which hinder the improvement of forecasting efforts. Nevertheless, temperature shows
504 to be the most important climate feature, while precipitation had less influence. As previously
505 stated, the availability of human water sources appears to have a greater impact on the breeding
506 of invasive *Aedes* mosquitoes than natural rainfall, as these mosquitoes often rely on artificial
507 containers near human habitats [34]. Although heavy rainfall can disrupt larval development by
508 washing out breeding sites, the connection between precipitation and mosquito populations varies
509 depending on local climate conditions [34].

510 Additionally, the inclusion of egg abundance proved to be a key predictor. Our findings confirm
511 that incorporating mosquito-related data improves the fitting and forecasting of predictive models.
512 Consequently, continuous monitoring of mosquitoes and egg abundance by public health systems is
513 essential for more accurate forecasting and effective control measures.

514 Furthermore, selecting the appropriate lagged variables and ovitrap egg counts, we validated
515 the models using different evaluation metrics. Based on metrics such as Root Mean Squared Error
516 (RMSE) and Mean Absolute Error (MAE), the Random Forest (RF) model outperformed the oth-
517 ers, followed by the Seasonal Autoregressive Integrated Moving Average with Exogenous variables
518 (SARIMAX) model. Among the models evaluated, RF performed best on the training data, while
519 the Generalized Linear Model (GLM) performed best on the testing data, with SARIMAX in second
520 place.

521 The poor performance of the models on the Bizkaia test set can be attributed to differences in
522 the characteristics of the training and test data. Notably, the mean value of the training data is
523 significantly lower than that of the test data, with egg counts in 2023 being unusually high. This
524 discrepancy between the training and test datasets likely contributed to the models' suboptimal
525 performance for Bizkaia. Nevertheless, for predicting egg abundance in the municipality of Bilbao
526 (Bizkaia), SARIMAX demonstrated superior performance.

527 Finally, we applied the best-performing models to estimate *Aedes* invasive mosquito abundance
528 in the Basque Country provinces for the upcoming year. By analyzing mosquito egg counts and
529 environmental factors, this study improves and contributes the understanding of seasonal influences
530 on mosquito abundance in a non-endemic region with a maritime climate, characterized by cooler
531 temperatures, rainy weather, and the presence of competent mosquito vectors. These predictions
532 could be used to inform public health strategies and mosquito control efforts, thereby helping to
533 prevent the spread of mosquito-borne diseases in non-endemic regions.

534 These findings provide valuable insights for future research on assessing the risk of arboviro-
535 sis outbreaks in non-endemic regions like the Basque Country. By considering factors such as imported
536 cases, mosquito abundance, and seasonal variations, risk evaluations for mosquito-borne diseases can
537 be refined. Nevertheless, limitations remain in generalizing these results across the diverse areas
538 within each province. For example, Bizkaia, which houses the largest human population in the
539 Basque Country, includes regions with distinct micro-climates that may influence invasive mosquito
540 abundance differently.

541 Moreover, by considering shorter temporal intervals, such as weekly data collection (depend-
542 ing on vector monitoring schedules and data availability), would improve the precision of vector
543 control strategies and strengthen the assessment of mosquito-borne disease risks. However, this

544 would depend mostly on the vector population monitoring intervals and the availability of data.
545 Furthermore, future improvements in this research should consider a deeper analysis of the meth-
546 ods for partitioning the dataset into training and testing sets, which might enhance the model's
547 performance.

548 This research aims to offer an estimate of mosquito population abundance and contribute to
549 the development of vector control strategies, thus mitigating the risks of mosquito-borne infections,
550 particularly considering the region's specific environmental conditions. Furthermore, this study
551 highlights the critical need for ongoing, localized surveillance to better understand and address the
552 expanding threat of mosquito-borne diseases.

553 References

- 554 [1] Alto BW, Juliano SA. (2001) Precipitation and Temperature Effects on Populations of *Aedes*
555 *albopictus* (Diptera: Culicidae): Implications for Range Expansion. *Journal of Medical Entomology*. 38(5):646–656. <https://doi.org/10.1603/0022-2585-38.5.646>
- 557 [2] Akram M, Cerin E, Lamb KE, et al. (2023) Modelling count, bounded and skewed continuous
558 outcomes in physical activity research: beyond linear regression models. *Int J Behav Nutr Phys Act*. 20(57). <https://doi.org/10.1186/s12966-023-01460-y>
- 560 [3] Aguiar, M., Van-Dierdonck, J.B., Mar, J. et al. Critical fluctuations in epidemic models explain
561 COVID-19 post-lockdown dynamics. *Sci Rep* 11, 13839 (2021). <https://doi.org/10.1038/s41598-021-93366-7>
- 563 [4] Stollenwerk, N., van Noort, S., Martins, J., Aguiar, M., Hilker, F., Pinto, A., Gomes,
564 G. (2010). A spatially stochastic epidemic model with partial immunization shows in mean
565 field approximation the reinfection threshold. *Journal of Biological Dynamics*, 4(6), 634–649.
566 <https://doi.org/10.1080/17513758.2010.487159>
- 567 [5] Betanzos-Reyes ÁF, Rodríguez MH, Romero-Martínez M, et al. (2018) Association of dengue
568 fever with *Aedes* spp. abundance and climatological effects. *Salud Publica Mex*. 60(1):12–20.
- 569 [6] Breiman L. (2001) Random Forests. *Machine Learning*. 45:5–32. <https://doi.org/10.1023/A:1010933404324>
- 571 [7] Carvajal TM, Viacrusis KM, Hernandez LFT, et al. (2018) Machine learning methods reveal
572 the temporal pattern of dengue incidence using meteorological factors in metropolitan Manila,
573 Philippines. *BMC Infect Dis*. 18:183. <https://doi.org/10.1186/s12879-018-3066-0>
- 574 [8] Ceia-Hasse A, Sousa CA, Gouveia BR, Capinha C. (2023) Forecasting the abundance of disease
575 vectors with deep learning. *Ecological Informatics*. 78:102272. <https://doi.org/10.1016/j.ecoinf.2023.102272>
- 577 [9] Cevidanes A, Goiri F, Barandika JF, et al. (2023) Invasive *Aedes* mosquitoes in an urban—peri-
578 urban gradient in northern Spain: evidence of the wide distribution of *Aedes japonicus*. *Parasites Vectors*. 16:234. <https://doi.org/10.1186/s13071-023-05862-6>
- 580 [10] Chaves LF, Valerín Cordero JA, Delgado G, et al. (2021) Modeling the association between
581 *Aedes aegypti* ovitrap egg counts, multi-scale remotely sensed environmental data and arboviral

- 582 cases at Puntarenas, Costa Rica (2017-2018). *Curr Res Parasitol Vector Borne Dis.* 1:100014.
583 <https://doi.org/10.1016/j.crvbd.2021.100014>
- 584 [11] Cochet A, Calba C, Jourdain F, et al. (2022) Autochthonous dengue in mainland France,
585 2022: geographical extension and incidence increase. *Euro Surveill.* 27(44):pii=2200818. <https://doi.org/10.2807/1560-7917.ES.2022.27.44.2200818>
586
- 587 [12] da Cruz Ferreira DA, Degener CM, de Almeida Marques-Toledo C, et al. (2017) Meteorological
588 variables and mosquito monitoring are good predictors for infestation trends of *Aedes aegypti*,
589 the vector of dengue, chikungunya and Zika. *Parasites Vectors.* 10:78. <https://doi.org/10.1186/s13071-017-2025-8>
590
- 591 [13] da Re D Marini G, Bonannella C, et al. (2023) Inferring the seasonal dynamics and abundance
592 of an invasive species using a spatio-temporal stacked machine learning model. <https://doi.org/10.32942/X2NG70>
593
- 594 [14] da Silva ST, Gabrick EC, Protachevicz PR, et al. (2024) When climate variables improve the
595 dengue forecasting: a machine learning approach. *ArXiv:2404.05266*. <https://arxiv.org/abs/2404.05266>
596
- 597 [15] de Jesus CP, Dias FBS, Villela DMA, Maciel-de-Freitas R. (2020) Ovitrap traps Provide a Reliable
598 Estimate of *Wolbachia* Frequency during wMelBr Strain Deployment in a Geographically Iso-
599 lated *Aedes aegypti* Population. *Insects.* 11(2):92. <https://doi.org/10.3390/insects11020092>
600
- 601 [16] de Lima CL, da Silva CC, da Silva ACG, et al. (2022) Prediction of *Aedes aegypti* breeding
602 distribution through spatiotemporal analysis and machine learning: A case study in Recife,
603 Pernambuco. Available at Research Square. <https://doi.org/10.21203/rs.3.rs-1200442/v1>
604
- 605 [17] Diniz DFA, Romão TP, Helvécio E, de Carvalho-Leandro D, Xavier MDN, Peixoto CA, de Melo
606 Neto OP, Melo-Santos MAV, Ayres CFJ. (2022) A comparative analysis of *Aedes albopictus*
607 and *Aedes aegypti* subjected to diapause-inducing conditions reveals conserved and divergent
608 aspects associated with diapause, as well as novel genes associated with its onset. *Curr Res*
609 *Insect Sci.* 2:100047. <https://doi.org/10.1016/j.cris.2022.100047>
- 610 [18] ECDC. European Centre for Disease Prevention and Control. Accessed on June, 2024. [https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-albop](https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-albopictus)
611 [ictus](https://www.ecdc.europa.eu/en/disease-vectors/facts/mosquito-factsheets/aedes-albopictus).
612
- 613 [19] Eustat. Euskal Estatistika Erakundea Instituto Vasco de Estadística. Accessed on June, 2024.
614 [https://en.eustat.eus/estad/id_268/ti_Population%20and%20housing%20census.%20P](https://en.eustat.eus/estad/id_268/ti_Population%20and%20housing%20census.%20Population%20structure/latest-press-release.html)
615 [opulation%20structure/latest-press-release.html](https://en.eustat.eus/estad/id_268/ti_Population%20and%20housing%20census.%20Population%20structure/latest-press-release.html).
- 616 [20] Ezeakacha NF, Yee DA. (2019) The role of temperature in affecting carry-over effects and larval
617 competition in the globally invasive mosquito *Aedes albopictus*. *Parasites Vectors.* 12(1):123.
618 <https://doi.org/10.1186/s13071-019-3391-1>

- 619 [21] Goiri F, González MA, Cevidanes A, et al. (2024) Mosquitoes in urban green spaces and
620 cemeteries in northern Spain. *Parasites Vectors*. 17:168. <https://doi.org/10.1186/s13071-024-06263-z>
621
- 622 [22] Goiri F, González MA, Goikolea J, Oribe M, Castro V, Delacour S, Lucientes J, Ortega-
623 Araiztegi I, Barandika JF, García-Pérez AL. (2020) Progressive Invasion of *Aedes albopictus*
624 in Northern Spain in The Period 2013-2018 and A Possible Association with the Increase in
625 Insect Bites. *Int J Environ Res Public Health*. 17(5):1678. <https://doi.org/10.3390/ijerph17051678>
626
- 627 [23] Hardin J, Hilbe J. (2003) *Generalized Estimating Equations*. London, England: Chapman and
628 Hall/CRC. ISBN 1-58488-307-3.
- 629 [24] Hyndman RJ, Khandakar Y. (2008) Automatic Time Series Forecasting: The forecast Package
630 for R. *Journal of Statistical Software*. 27(3):1–22. <https://doi.org/10.18637/jss.v027.i03>
- 631 [25] Juliano SA, O’Meara GF, Morrill JR, Cutwa MM. (2002) Desiccation and thermal tolerance
632 of eggs and the coexistence of competing mosquitoes. *Oecologia*. 130(3):458–469. <https://doi.org/10.1007/s004420100811>
633
- 634 [26] Leung XY, Islam RM, Adhami M, Ilic D, McDonald L, Palawaththa S, Diug B, Munshi SU,
635 Karim MN. (2023) A systematic review of dengue outbreak prediction models: Current scenario
636 and future directions. *PLoS Negl Trop Dis*. 17(2):e0010631. <https://doi.org/10.1371/journal.pntd.0010631>
637
- 638 [27] Keun Young Lee, Namil Chung, Suntae Hwang. (2016) Application of an artificial neural net-
639 work (ANN) model for predicting mosquito abundances in urban areas. *Ecological Informatics*.
640 36:172–180. <https://doi.org/10.1016/j.ecoinf.2015.08.011>
- 641 [28] Kinney AC, Current S, Lega J. (2021) *Aedes-AI*: Neural network models of mosquito abun-
642 dance. *PLOS Computational Biology*. 17(11):e1009467. <https://doi.org/10.1371/journal.pcbi.1009467>
643
- 644 [29] OpenData Euskadi. Estaciones meteorológicas: lecturas recogidas en el 2023. Accessed on May,
645 2024. <https://opendata.euskadi.eus/catalogo/-/estaciones-meteorologicas-lecturas-recogidas-en-2023/>
646
- 647 [30] O’Hara R, Kotze J. (2010) Do not log-transform count data. *Nat Prec*. <https://doi.org/10.1038/npre.2010.4136.1>
648
- 649 [31] Prasad A, Sreedharan S, Bakthavachalu B, Laxman S. (2023) Eggs of the mosquito *Aedes*
650 *aegypti* survive desiccation by rewiring their polyamine and lipid metabolism. *PLoS Biol*.
651 21(10):e3002342. <https://doi.org/10.1371/journal.pbio.3002342>
- 652 [32] Portal oficial de la Asociación Española de pediatría sobre vacunas y Inmunizaciones. Comité
653 Asesor de Vacunas e Inmunizaciones de la Asociación Española de Pediatría (CAV-AEP).
654 Accessed on November, 2024. <https://vacunasaep.org/profesionales/noticias/dengue-zika-y-chikunguna-en-espana-2023>
655

- 656 [33] Raizada S, Mala S, Shankar A. (2021). Vector-Borne Disease Outbreak Prediction Using Ma-
657 chine Learning Techniques. In: Advanced Deep Learning for Engineers and Scientists. (Eds)
658 Prakash KB, Kannan R, Alexander S, Kanagachidambaresan GR. EAI/Springer Innovations
659 in Communication and Computing. Springer, Cham. [https://doi.org/10.1007/978-3-030-](https://doi.org/10.1007/978-3-030-66519-7_9)
660 [-66519-7_9](https://doi.org/10.1007/978-3-030-66519-7_9)
- 661 [34] Roiz D, Rosà R, Arnoldi D, Rizzoli A. (2010) Effects of Temperature and Rainfall on the
662 Activity and Dynamics of Host-Seeking *Aedes albopictus* Females in Northern Italy. Vector-
663 Borne and Zoonotic Diseases. 10(8):811–816. <https://doi.org/10.1089/vbz.2009.0098>
- 664 [35] Sante Publique France. Accessed on June, 2024. [https://www.santepubliquefrance.fr/m](https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-a-transmission-vectorielle/chikungunya/articles/donnees-en-france-metropolitaine/chikungunya-dengue-et-zika-donnees-de-la-surveillance-renforcee-en-france-hexagonale-2024)
665 [aladies-et-traumatismes/maladies-a-transmission-vectorielle/chikungunya/artic](https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-a-transmission-vectorielle/chikungunya/articles/donnees-en-france-metropolitaine/chikungunya-dengue-et-zika-donnees-de-la-surveillance-renforcee-en-france-hexagonale-2024)
666 [les/donnees-en-france-metropolitaine/chikungunya-dengue-et-zika-donnees-de-l](https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-a-transmission-vectorielle/chikungunya/articles/donnees-en-france-metropolitaine/chikungunya-dengue-et-zika-donnees-de-la-surveillance-renforcee-en-france-hexagonale-2024)
667 [a-surveillance-renforcee-en-france-hexagonale-2024](https://www.santepubliquefrance.fr/maladies-et-traumatismes/maladies-a-transmission-vectorielle/chikungunya/articles/donnees-en-france-metropolitaine/chikungunya-dengue-et-zika-donnees-de-la-surveillance-renforcee-en-france-hexagonale-2024).
- 668 [36] Sarker IH. (2021) Machine Learning: Algorithms, Real-World Applications and Research Di-
669 rections. SN COMPUT SCI. 2:160. <https://doi.org/10.1007/s42979-021-00592-x>
- 670 [37] Sidumo B, Sonono E, Takaidza I. (2024) Count Regression and Machine Learning Techniques
671 for Zero-Inflated Overdispersed Count Data: Application to Ecological Data. Ann Data Sci.
672 11:803–817. <https://doi.org/10.1007/s40745-023-00464-6>
- 673 [38] Shutt DP, Goodsman DW, Hemez ZJL, et al. (2021) A Process-Based Model with Tempera-
674 ture, Water, and Lab-derived Data Improves Predictions of Daily Mosquito Density BioRxiv
675 2021.09.08.458905. <https://doi.org/10.1101/2021.09.08.458905>
- 676 [39] Shumway RH, Stoffer DS, Stoffer DS. (2000) Time series analysis and its applications. Vol 3,
677 p 4. New York: Springer.
- 678 [40] Sun C, Nimbalkar J, Bedi R. (2022) Predicting Future Mosquito Larval Habitats Using Time
679 Series Climate Forecasting and Deep Learning. IEEE MIT Undergraduate Research Technology
680 Conference (URTC). Cambridge, MA, USA. 2022 pp. 1-5. [https://doi.org/10.1109/URTC](https://doi.org/10.1109/URTC56832.2022.10002240)
681 [56832.2022.10002240](https://doi.org/10.1109/URTC56832.2022.10002240)
- 682 [41] Torina A, La Russa F, Blanda V, Peralbo-Moreno A, et al. (2023) Modelling time-series
683 *Aedes albopictus* abundance as a forecasting tool in urban environments. Ecological Indica-
684 tors. 150:110232. <https://doi.org/10.1016/j.ecolind.2023.110232>
- 685 [42] Hothorn T, Hornik K, Zeileis A. (2015) ctree: Conditional inference trees. The comprehensive
686 R archive network. 8:1–34. Accessed on September, 2024. [https://cran.r-project.org/web](https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf)
687 [/packages/partykit/vignettes/ctree.pdf](https://cran.r-project.org/web/packages/partykit/vignettes/ctree.pdf)
- 688 [43] Tran A, L’Ambert G, Lacour G, Benoît R, Demarchi M, Cros M, Cailly P, Aubry-Kientz M,
689 Balenghien T, Ezanno P. (2013) A rainfall- and temperature-driven abundance model for *Aedes*
690 *albopictus* populations. Int J Environ Res Public Health. 10(5):1698–719. [https://doi.org/](https://doi.org/doi:10.3390/ijerph10051698)
691 [doi:10.3390/ijerph10051698](https://doi.org/doi:10.3390/ijerph10051698)
- 692 [44] WHO. World Health Organization. Accessed on June, 2024. [https://www.who.int/emergenc](https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON518)
693 [ies/disease-outbreak-news/item/2024-DON518](https://www.who.int/emergencies/disease-outbreak-news/item/2024-DON518).

- 694 [45] Wood S. (2006) Generalized Additive Models: An Introduction with R. Chapman & Hall/CRC.
695 ISBN 1-58488-474-6.

696 **Declarations**

697 **Acknowledgments**

698 We thank NEIKER, the Basque Institute for Agricultural Research and Development, the Public
699 Health Epidemiological Unit, and the municipalities direction of Basque Country for providing the
700 mosquito eggs count monitoring data. Hamna Mariyam K.B. acknowledge the School of Data
701 Analytics Mahatma Gandhi University in Kottayam, Kerala, India for the support.

702 **Funding**

703 This research is supported by the Basque Government through the “Mathematical Modeling
704 Applied to Health” Project, BERC 2022-2025 program and by the Spanish Ministry of Sciences,
705 Innovation and Universities: BCAM Severo Ochoa accreditation CEX2021-001142-S / MICIN / AEI
706 / 10.13039/501100011033. This work is also supported by the ARBOSKADI project for monitoring
707 vector-borne diseases in the Basque Country, Euskadi.

708 The collection of the data was funded by the Department of Food, Rural Development, Agri-
709 culture and Fisheries, and the Department of Health of the Basque Government, the Ministry
710 of Health, Social Policy, and Equality of the Government of Spain and the project EU-LIFE 18
711 IPC/ES/000001 (Urban Klima 2050). Maíra Aguiar and Aitor Cevitanes acknowledges the finan-
712 cial support by the Ministerio de Ciencia e Innovacion (MICINN) of the Spanish Government and
713 European Union Next Generation EU/PRTR through the Ramon y Cajal grants RYC2021-031380-I
714 and RYC2021-033084-I, respectively.

715 **Ethical approval**

716 Not applicable.

717 **Competing interests**

718 The authors declare that they have no known competing financial interests or personal relation-
719 ships that could influence the work reported in this paper.

720 **Data availability**

721 The environmental data used in this study were retrieved from several meteorological stations
722 managed by Euskalmet, the Basque Agency of Meteorology. This data is openly available through
723 the OpenData Euskadi platform [29].

724 The mosquito egg counts, collected using ovitraps, were provided by NEIKER, the Basque Insti-
725 tute for Agricultural Research and Development (for details, see [9]). Due to ethical considerations
726 and commercial sensitivity, these data are not publicly available.

727 Supplementary Material

728 A. Dataset summary

729 A.1. Climatic variables per province

730 Figure 10 shows the distribution of climatic data, including temperature, humidity, and precip-
731 itation, across the provinces of Gipuzkoa, Bizkaia, and Araba. The graphs summarize the weather
732 variables, highlighting outliers (represented as single points or circular dots) in the dataset. The
733 horizontal line dividing the box in two represents the median value of the time series for each climatic
734 variable.

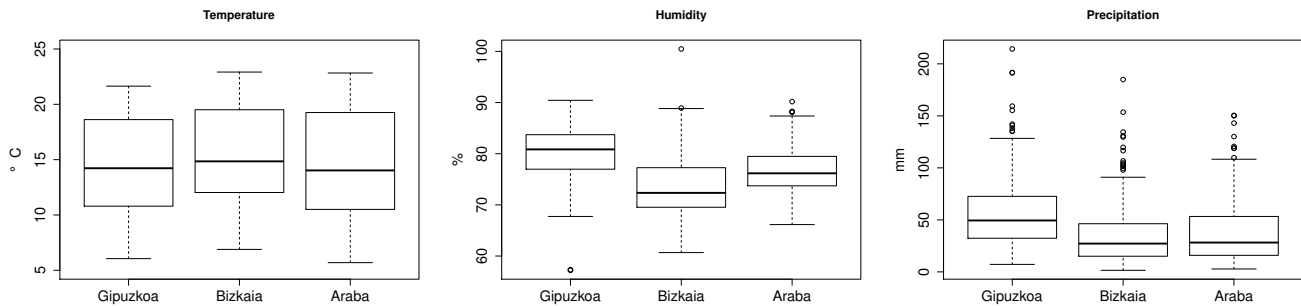


Figure 10: Distribution of time series values for average temperature ($^{\circ}\text{C}$), relative air humidity (%), and cumulative precipitation (mm) over an interval of 14 days, for all three provinces of the Basque Country.

735 A.2. Lagged time series for Gipuzkoa

736 Figure 11 (a) shows the monotonic correlation, using Spearman correlation, between the climatic
737 variables and the egg count time series for Gipuzkoa. Figure 11 (b) highlights the time lag at which
738 the highest correlation occurs. At this time lag, the lagged time series will be used as predictors.

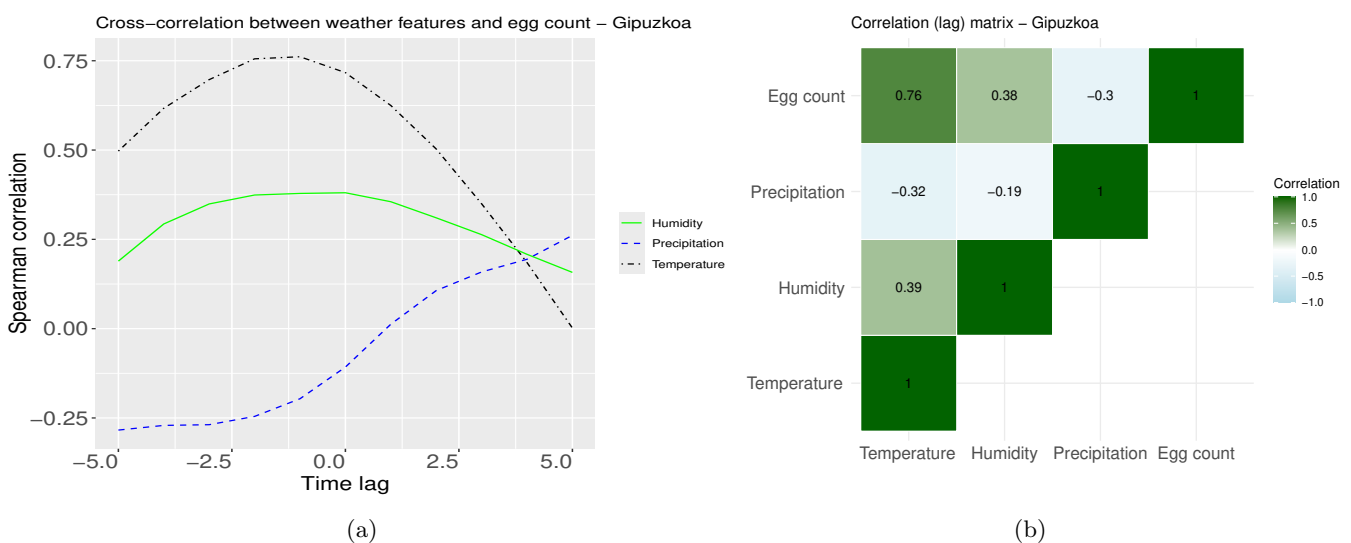


Figure 11: (a) Spearman correlation indices for the time lag between temperature, humidity, precipitation, and the number of eggs. (b) The highest Spearman correlation values between the lagged time series.

739 A.3. Lagged time series for Bizkaia

740 Figure 12 (a) shows the monotonic correlation using Spearman correlation between the climatic
 741 variables and the egg count time series for Gipuzkoa. Figure 12 (b) highlights the time lag at which
 742 the highest correlation value occurs. The lagged time series corresponding to this highest correlation
 743 will be used as predictor variables.

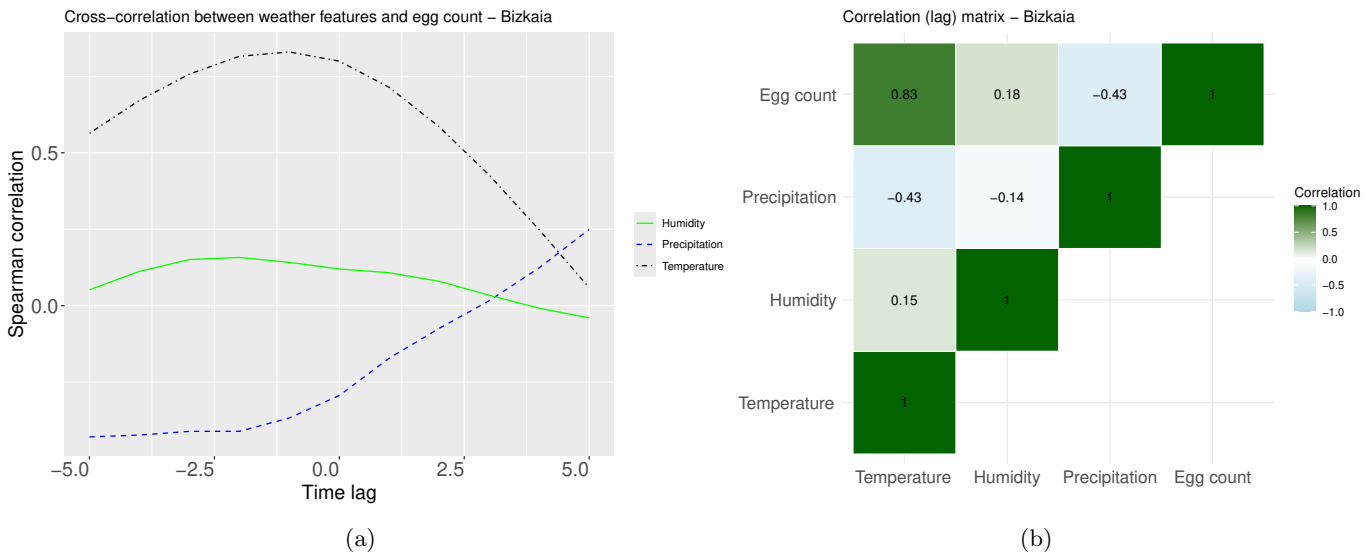


Figure 12: (a) Spearman correlation indices for time lags between temperature, humidity, and precipitation with the number of eggs. (b) The highest Spearman correlation values for the lagged time series.

744 B. Data selection per municipalities

745 Due to the dispersed data with many zero values for egg counts and the lack of sufficient
 746 information to create a reliable training dataset at the municipality level, we conducted the analysis
 747 at the provincial level. However, we selected one municipality from each province to present the
 748 results.

749 For Gipuzkoa, we selected Irun, a municipality of interest due to its proximity to the French
 750 border and the frequent movement of travelers. Irun also had more positive ovitraps during the
 751 analysis period compared to the capital, Donostia/San Sebastián. The C084 weather station in
 752 Irun was selected over C083, as the latter's dataset lacked temperature, precipitation, and humidity
 753 data.

754 For Bizkaia, we chose Bilbao as the municipality, rather than larger municipalities like Barakaldo
 755 or Basauri, since both lack meteorological stations within their boundaries. For Bilbao, station C0B0
 756 had no data on precipitation or temperature, while station C039, located in Deusto, provided data
 757 from 2016 to 2021, and station C03A began recording data in December 2021. To cover the entire
 758 study period, data from both C039 and C03A were used in the initial analysis.

759 For Araba, we chose the municipality of Laudio/Llodio. The primary reason for selecting this
 760 municipality is that no egg counts were recorded in the ovitraps in the capital, Vitoria. For weather
 761 data, two meteorological stations in Laudio/Llodio were listed in the database (see more details in
 762 [29]): station C067, which was selected, and station C027, which was not included due to missing
 763 data for the chosen period.

764 Egg count data were collected at the municipality level, disregarding specific ovitrap locations.
765 The dataset was constructed by averaging the highest three egg counts from the ovitraps for each
766 municipality every 14 days (bi-weekly). This approach was necessary because the number of mon-
767 itored ovitraps varied throughout the study period. We selected the top three counts since, on
768 average, no more than 10 ovitraps were placed in each municipality every 14 days. Weather data
769 were aggregated by municipality on a daily basis. A dataset was then constructed containing the
average temperature, average humidity, and cumulative precipitation for the previous 14 days.

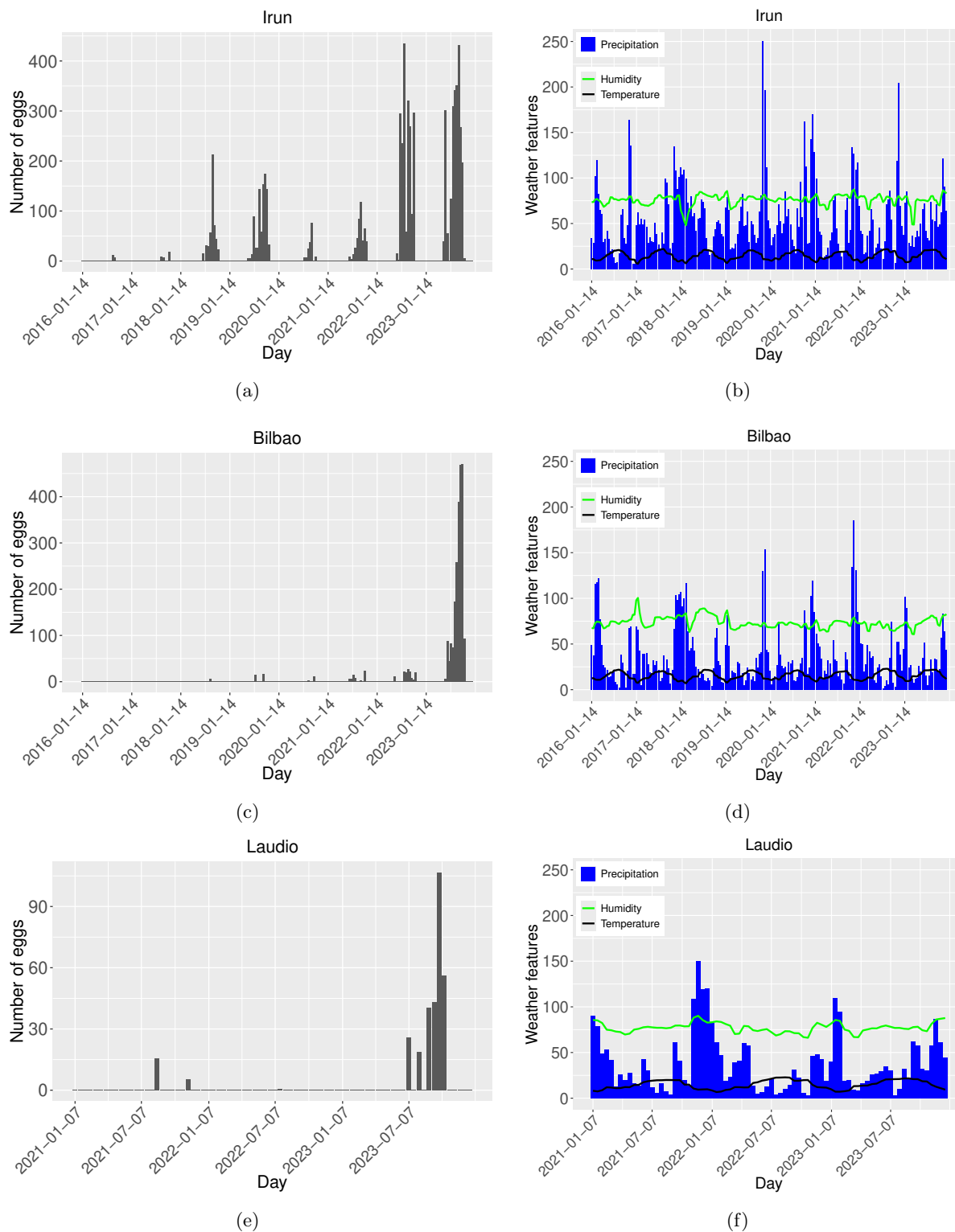


Figure 13: Number of mosquito eggs collected, in (a), (c), (f), and average temperature ($^{\circ}C$), relative air humidity (%), and cumulative precipitation (mm), in (b), (d), (f). For Irun (in Gipuzkoa), Bilbao (in Bizkaia) and Ludio (in Araba).

770 This information was combined into a single dataset for each municipality. The time series of
771 egg counts, average temperature, humidity, and cumulative precipitation are presented in Figures
772 13(a)-(b), 13(c)-(d), and 13(e)-(f).

773 For further analysis, we will focus on the municipalities of Irun and Bilbao, since Laudio has
774 only recorded positive ovttraps from 2021 onward. The same methodology and analysis applied at
775 the provincial scale will now be carried out at the municipality scale.

776 B.1. Irun

777 Statistical analysis

778 For Irun, Figure 14 shows the relationship between meteorological variables and number of
779 mosquito eggs count using scatter plots. While Figure 15 shows the monotonic correlation using
780 Spearman correlation between eggs counts and the time lagged versions of the climate features. For
781 temperature, maximum correlation occurs at -1 units (2 weeks). For humidity, maximum occurs at
782 -2 units. And for precipitation, maximum correlation occurs at -5 units, with negative correlation.

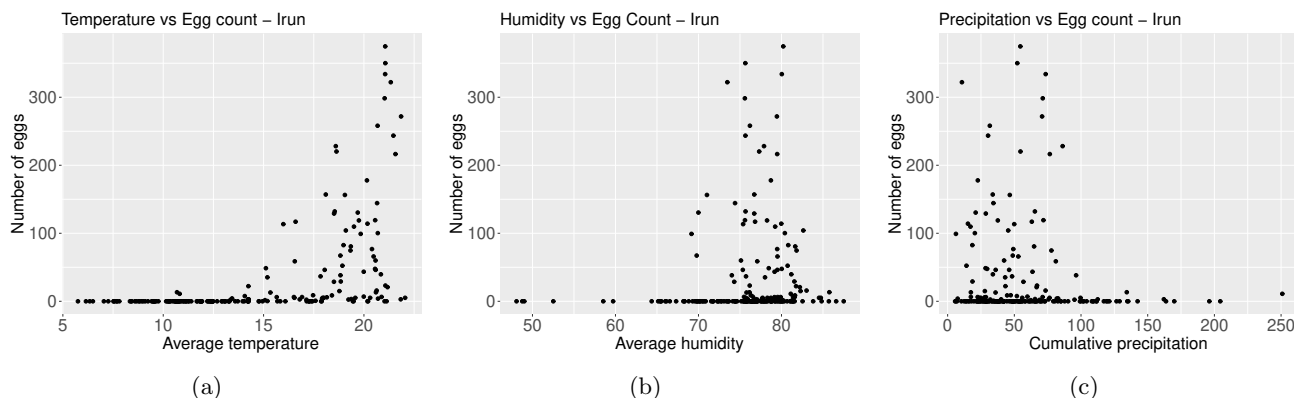


Figure 14: (a) Average temperature (in $^{\circ}C$) versus number of collected mosquitoes eggs. As temperature increases, the number of eggs increases. (b) Average air relative humidity (in %) versus number of collected mosquitoes eggs. As humidity increases, the number of eggs increases. (c) Accumulated precipitation (in mm) versus number of collected mosquitoes eggs. As precipitation increases, the number of eggs decreases.

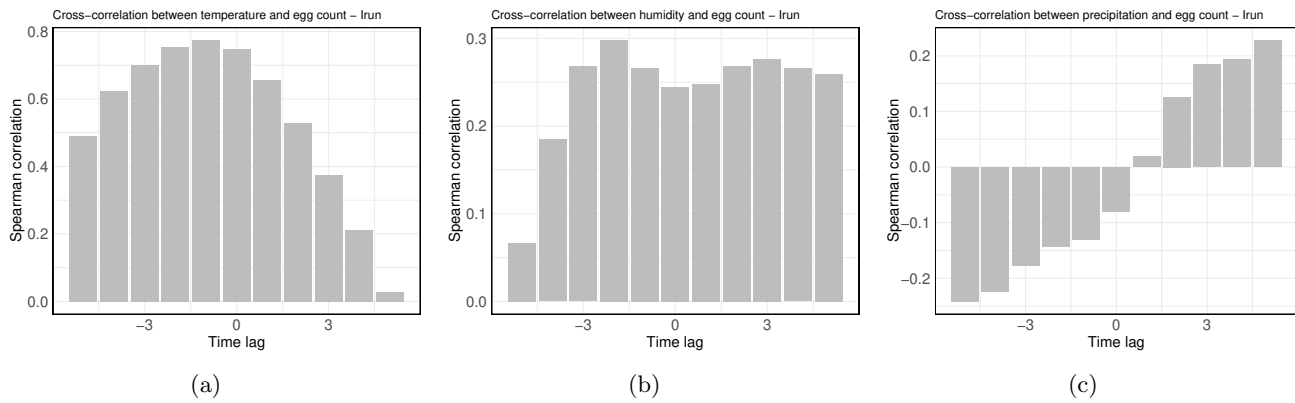


Figure 15: Spearman correlation between the weather feature time series and the number of mosquito eggs, with a time lag of 1 unit (2 weeks). (a) For temperature, the maximum correlation occurs at a lag of -1 unit. (b) For humidity, the maximum correlation occurs at a lag of -2 units. (c) For precipitation, the maximum negative correlation occurs at a lag of -5 units.

783 Fitting

784 We implemented the GLMG, SARIMAX, RF, and CT models using the R programming language
 785 for the Irun dataset. The training dataset spans from 2017 to 2022, while the test dataset consists
 786 of data from 2023, as shown in Figure 16. We compared the models' performance on both the
 787 training and testing datasets, evaluating them using the Mean Absolute Error (MAE), Root Mean
 788 Squared Error (RMSE), and R-squared score (R^2), as detailed in Table 1.

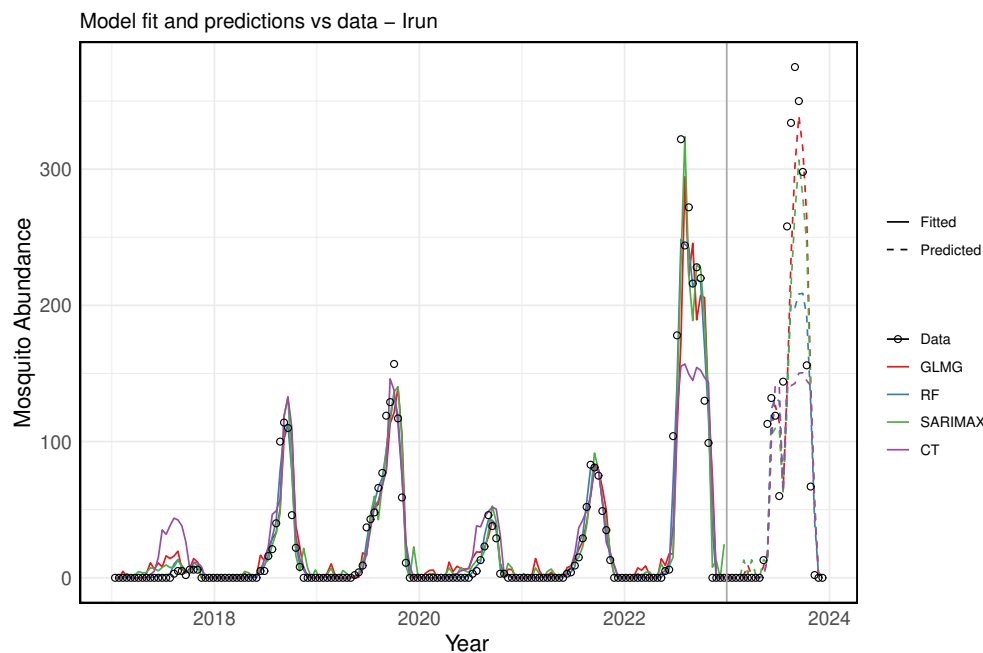


Figure 16: Comparison of actual data versus fitted and predicted values for the Irun dataset. The actual data is represented by open black circles, while the fitted values for each model are shown as solid lines and the predicted (test) values as dashed lines. The models are colored as follows: RF model ($n_{tree} = 600$, $m_{try} = 5$) in blue, GLMG model in red, SARIMAX model in green, and CT model ($n_{tree} = 500$, $m_{try} = 3$) in purple. The vertical gray line separates the training dataset (2017–2022) from the testing dataset (2023).

Model	Evaluation Metrics					
	MAE Train	MAE Test	RMSE Train	RMSE Test	R ² Train	R ² Test
RF	4.00	43.27	10.22	68.84	0.97	0.70
SARIMAX	8.45	37.55	17.32	57.35	0.91	0.79
GLM	10.10	33.51	21.57	52.87	0.86	0.82
CT	11.77	54.21	25.86	89.16	0.80	0.49

Table 1: Evaluation of error metrics for each model on the training and testing datasets (for Irun). MAE represents the Mean Absolute Error, RMSE the Root Mean Squared Error, and R^2 the R-squared score.

789 Forecasting

790 Subsequently, we used the best-performing trained models to forecast future *Aedes* invasive
791 mosquito abundance in Irun. To do this, we included 2023 data points as part of the training
792 dataset. Using the historical time series data and their lagged versions, we predicted future values
793 based on the most recent observations (see Figure 17). The error analysis for the training dataset
794 is presented in Table 2, which shows that the Random Forest (RF) model performed the best,
795 explaining 97% of the variance in the training dataset.

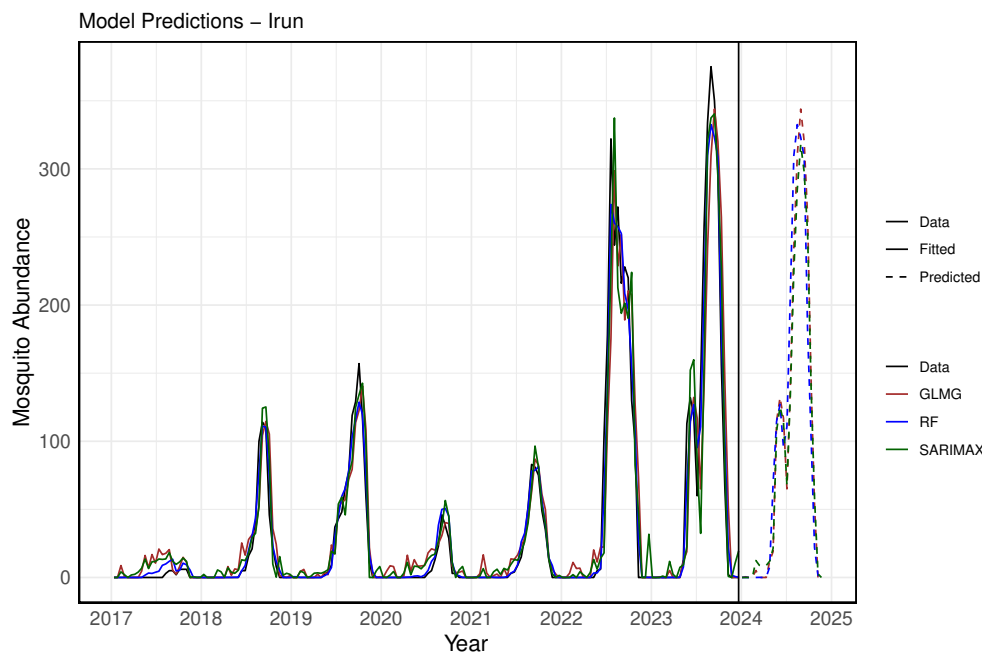


Figure 17: The actual data versus the fitted and predicted values of the model for Irun. The data is represented by a solid black line, with fitted values shown as solid lines in color and predicted values as dashed lines. In blue, the RF model ($n_{tree} = 600$, $m_{try} = 5$); in brown, the GLMG model; and in green, the SARIMAX model. The vertical black line separates the training dataset (from 2017 to 2023) from the forecasted period for the year 2024.

Evaluation Metrics in the training dataset			
Model	MAE	RMSE	R ²
RF	5.90	12.17	0.97
SARIMAX	11.89	23.56	0.90
GLM	13.68	27.71	0.86

Table 2: Evaluation of error metrics in the training dataset (for Irun), highlighting the best model performance. MAE refers to the Mean Absolute Error, RMSE stands for the Root Mean Squared Error, and R^2 represents the R-squared score.

796 **B.1. Bilbao**

797 **Statistical Analysis**

798 For Bilbao, Figure 18 presents the relationship between meteorological variables and mosquito
 799 egg counts using scatter plots. Figure 19 illustrates the monotonic correlation, calculated using
 800 Spearman’s correlation, between egg counts and the time-lagged versions of the climate features.

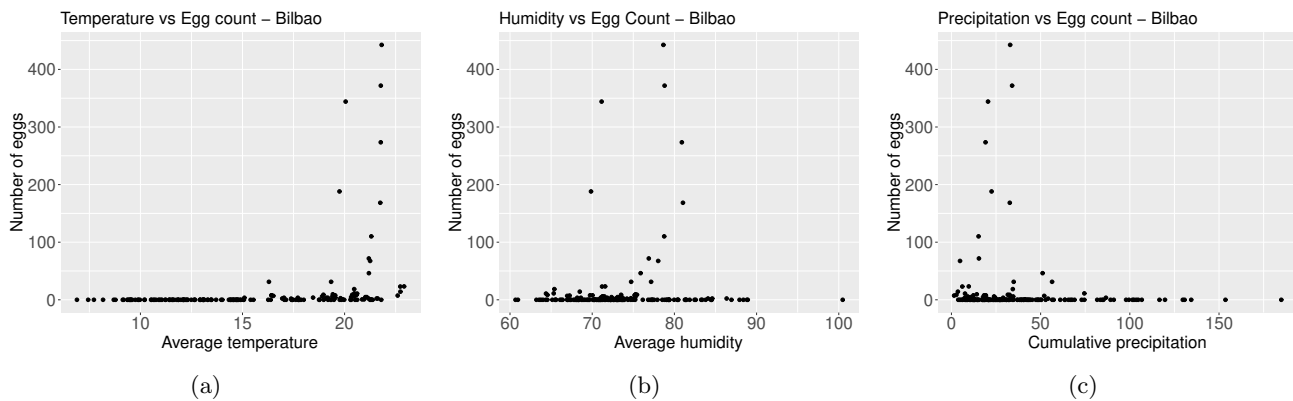


Figure 18: (a) Average temperature (in $^{\circ}C$) versus number of collected mosquito eggs. As temperature increases, the number of eggs increases. (b) Average relative humidity (in %) versus number of collected mosquito eggs. As humidity increases, the number of eggs increases. (c) Accumulated precipitation (in mm) versus number of collected mosquito eggs. As precipitation increases, the number of eggs decreases.

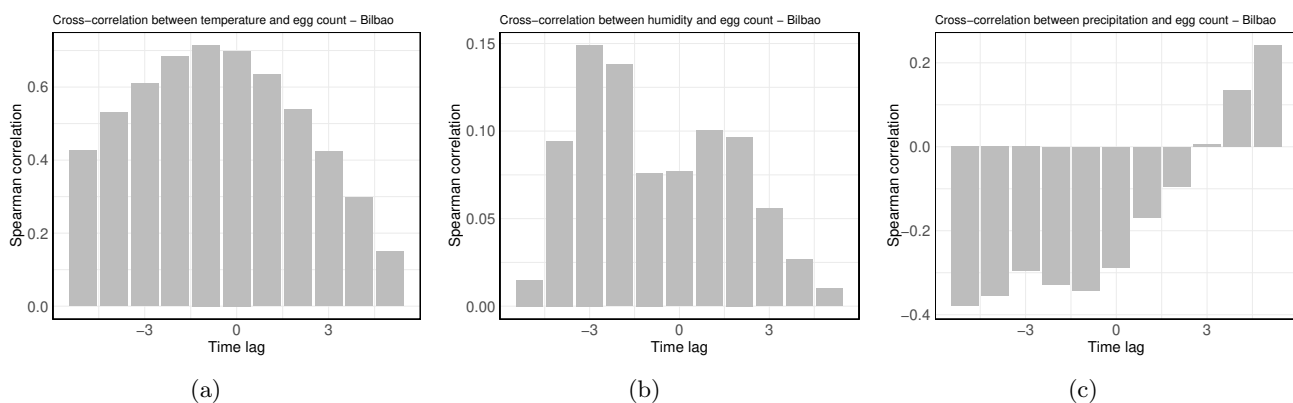


Figure 19: Spearman correlation between the weather features time series and number of mosquito eggs with a time lag of 1 unit (2 weeks). For (a) temperature, maximum correlation occurs at -1 units. For (b) humidity, maximum correlation occurs at -3 units. For (c) precipitation, maximum correlation occurs at -5 units.

801 **Fitting**

802 We also implemented the GLMG, SARIMAX, RF, and CT models for the Bilbao dataset. The
 803 training dataset consists of data from the year 2019 to 2022, while the test dataset consists of data
 804 points from 2023, as shown in Figure 20. We compare the models on both the training and testing
 805 datasets, evaluating each model’s performance using the MAE, RMSE, and R^2 metrics, as detailed
 806 in Table 3.

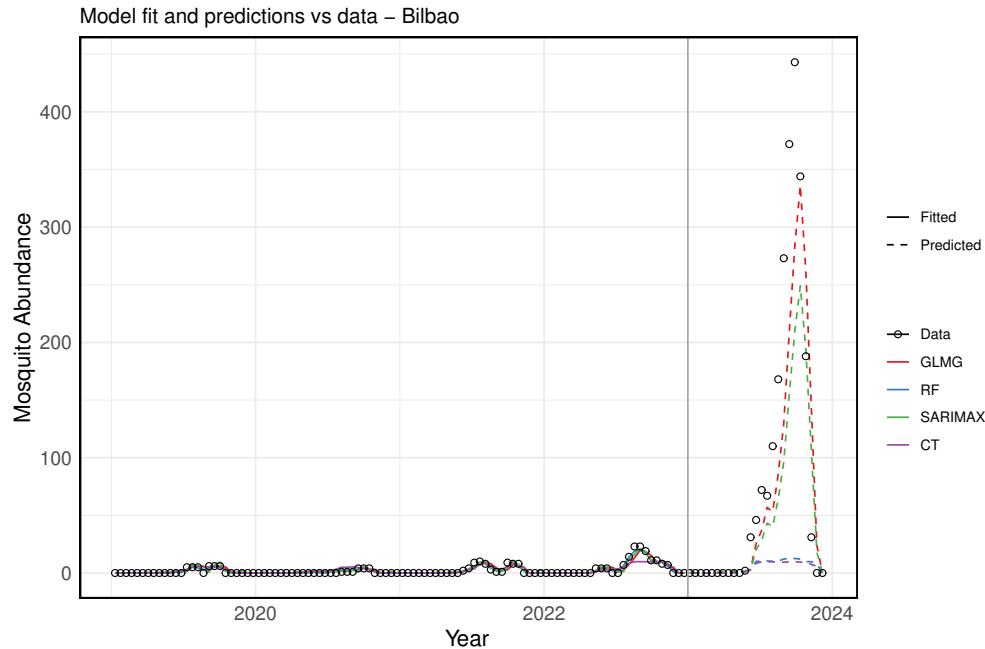


Figure 20: The actual data versus the fitted and tested values of the model for Irun. The data is represented by open black circles, while the fitted values of each model are shown in solid lines and the predicted (tested) values in dashed lines. In blue, the RF model ($n_{tree} = 600$, $m_{try} = 5$); in red, the GLMG model; in green, the SARIMAX model; and in purple, the CT model ($n_{tree} = 500$, $m_{try} = 3$). The vertical gray line delineates the training dataset (from 2017 to 2022) from the testing dataset (2023).

Model	Evaluation Metrics					
	MAE	MAE	RMSE	RMSE	R^2	R^2 Test
	Train	Test	Train	Test	Train	
RF	0.57	81.38	0.99	150.86	0.95	-0.32
SARIMAX	1.10	45.02	1.75	81.82	0.86	0.61
GLM	1.25	36.78	2.25	64.42	0.76	0.76
CT	1.41	82.00	2.72	152.01	0.65	-0.34

Table 3: Different evaluation error metrics in the train and in the test dataset (for Bilbao) of each model chosen. MAE for the Mean Absolute Error, RMSE for the Root Mean Squared Error, and R^2 representing the R-squared score.

807 **Forecasting**

808 The best-trained model was then used to predict future *Aedes* invasive mosquito abundance in
 809 Bilbao using the historical time series data and lagged versions, as shown in Figure 21. The error
 810 analysis for the training dataset is presented in Table 4, indicating that the model with the best
 811 performance is the SARIMAX model, which explains 97% of the variability in the training dataset.

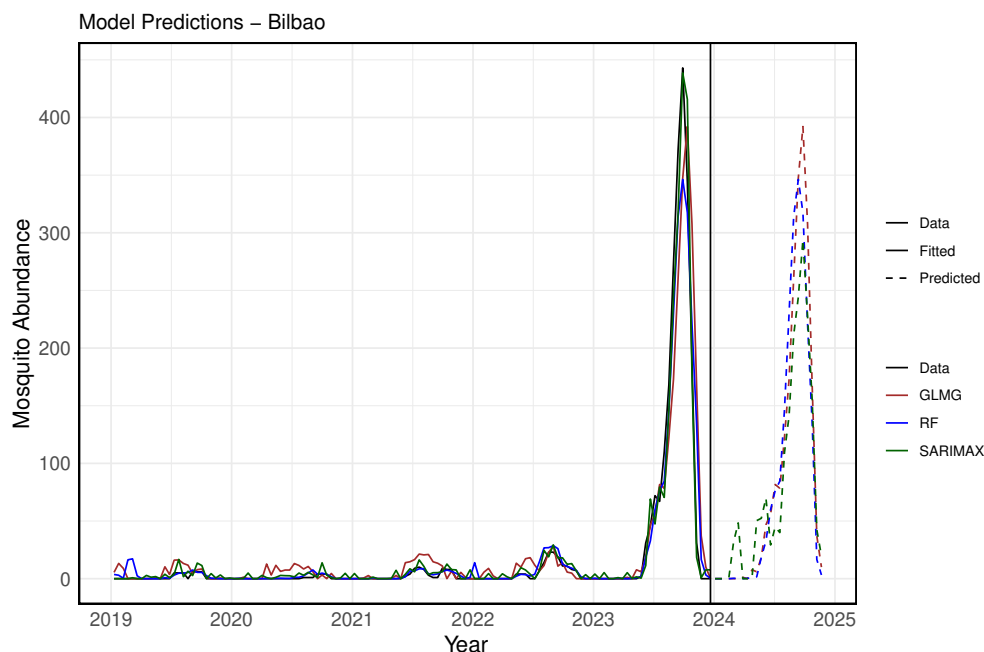


Figure 21: Actual data versus fitted and predicted values for Bilbao. The actual data is represented by a solid black line. The fitted values for each model are shown in solid colored lines, while predicted values are in dashed lines. The RF model is shown in blue ($n_{tree} = 600$, $m_{try} = 5$), the GLMG model in brown, and the SARIMAX model in green. The vertical black line separates the training dataset (from 2017 to 2023) from the forecasted period for the year 2024.

Evaluation Metrics in the training dataset			
Model	MAE	RMSE	R^2
RF	4.98	15.28	0.95
SARIMAX	4.63	10.57	0.97
GLM	9.45	24.23	0.87

Table 4: Evaluation error metrics for the training dataset (for Bilbao), showing the best model performance. MAE represents the Mean Absolute Error, RMSE is the Root Mean Squared Error, and R^2 denotes the R-squared score.